TU WIEN Informatics

Visuelle Analytik für die Erforschung von Kulturmodellen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Payam Chini Foroushan
Matrikelnummer 01429718

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Assist.Prof.Dr. Renata Georgia Raidou

Wien, 14. April 2021

Payam Chini Foroushan

Renata Georgia Raidou

TU WIEN Informatics

Visual Analytics for the Exploration of Cultural Models

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Wirtschaftsinformatik

by

Payam Chini Foroushan
Registration Number 01429718

to the Faculty of Informatics

at the TU Wien

Advisor: Assist.Prof.Dr. Renata Georgia Raidou

Vienna, 14th April, 2021

Payam Chini Foroushan

Renata Georgia Raidou

Erklärung zur Verfassung der Arbeit

Payam Chini Foroushan

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 14. April 2021

Payam Chini Foroushan

Acknowledgements

I would like to express my deep gratitude to my supervisor, Univ.Ass. Dr. Renata Georgia Raidou for the enthusiastic encouragement, advice, excellent assistance, and guidance during the process of this project. Throughout this Master's thesis journey, you have invested enormous time during our weekly meetings to answer my questions, which I highly appreciate and will not forget.

Special thanks go to my partners, brother, and soon to be wife Rosa Nimmrichter, who have always supported me throughout my studies.

Kurzfassung

Ziel dieser Masterarbeit ist es, aktuelle Visual-Analytics-Methoden zur Erforschung und Analyse kultureller Modelle zu überprüfen und zu verstehen. Anschließend entwarfen und entwickelten wir einen Rahmen für die interaktive Darstellung von Kulturmodellen. Dieser Rahmen kann für die Generierung neuen Wissens über Kulturmodelle und für die Bestätigung neuer Hypothesen im Bereich der Kulturwissenschaften verwendet werden. Unsere Methodik basiert auf dem Forschungsrahmen „three-cycle design science research framework“.

Bei der Literaturrecherche stellten wir fest, dass die wenigen bereits vorhandenen Visual-Analytics-Methoden im Bereich der Kulturwissenschaften nicht flexibel genug sind. Ausgehend davon entwarfen und implementierten wir ein Programm, das eine Kombination aus Python- und JavaScript- Programmbibliotheken verwendet. Dies bietet Benutzern die Flexibilität, die Visualisierungen nach Bedarf zu ändern und anzupassen.

Nachdem wir Anforderungen basierend auf relevanter Literatur im Bereich der Kulturwissenschaften festgelegt hatten, teilten wir die Anforderungen auf vier Aufgaben auf. Um ihren Zweck zu erfüllen, wählten wir für jede Aufgabe geeignete Visualisierungs- und Interaktionsmethoden aus.

Um das implementierte Visualisierungsprogramm zu evaluieren, überprüften wir drei verschiedene Fallstudien aus dem Bereich der Kulturwissenschaften. Wir versuchten, die Ergebnisse der Fallstudien zu reproduzieren, indem wir deren Methodik in unser Programm einsetzten. Wir verglichen die Ergebnisse der Fallstudie mit den Ergebnissen unseres Programms unter Verwendung unseres Visualisierungs-Frameworks. Schließlich prüften wir, ob alle definierten Aufgaben erfüllt werden konnten.

Nach Abschluss der Evaluierung stellten wir unseren Prototyp mit Docker fertig, sodass andere Forscher unser Visualisierungstool wiederverwenden und unser Ergebnis reproduzieren können. Schließlich wird unser Ansatz in einem Rahmen zusammengefasst, mit dem das aktuelle Visualisierungswerkzeug angepasst und geändert oder völlig neue Ansätze für andere Kulturmodelle erstellt werden können.

Abstract

The purpose of this master thesis is to review and comprehend current Visual Analytics methods and tools for the exploration and analysis of cultural models. Subsequently, we design and develop a framework for the interactive representation of cultural models, for the generation of new knowledge about cultural models, and for the confirmation of new hypotheses within the domain of cultural science. We based our methodology on the three-cycle design science research framework.

Reviewing the existing literature, we identified that the few, already existing Visual Analytics methods used in the cultural science domain lack flexibility. Using this literature gap as a motivation, we aim to design and implement a framework, using a combination of Python and JavaScript libraries, which gives users the flexibility to change and adapt the visualizations on demand.

After establishing requirements based on related literature in the cultural science domain, we break down the requirements into four medium-level tasks and select appropriate visualization and interaction methods to fulfill each task.

To evaluate the implemented visualization framework, we review three different case studies from the cultural science domain. We attempt to reproduce previous results by following their methodology in our developed framework. We compare the original results, and the results achieved, using our visualization framework. Eventually, we examine if all the defined tasks can be fulfilled.

After the evaluation is finished, we finalize our prototype using Docker, enabling other researchers to reuse our visualization tool and reproduce our result. Finally, our approach is summarized into a framework that can be used to adapt and change the current visualization tool or to create completely new approaches for other cultural models.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Definition

Understanding culture, norms, and cultural differences in society is essential for many reasons. It aids scientists in understanding more about human behavior and uncovers facts related to a society from a socioeconomic perspective. In order to understand more about culture, researchers have tried to come up with methods and models to measure different aspects of culture. Geert Hofstede has defined a six-dimensional cultural model, which summarizes characteristics of different cultures [1]. Other examples of cultural models include the GLOBE project [2], which sets nine dimensions (different from those of Hofstede's model) to describe a culture. These cultural models allow researchers to quantify the definition of culture, making the comparison of differences within countries more effortless and straightforward.

By finding correlations between cultural dimensions, as well as between social attributes, behaviors, and monetary or societal measures of different countries, we find answers to phenomena in various aspects of a society that may be beneficial for marketing and economics. There might be a correlation between characteristics of a culture and other attributes; for instance, a country's educational level, its national average income, or its population. The exploration of these correlations leads to a profound explanation of a society's purchase behavior, which may be used for marketing purposes. Papers such as De Mooij and Hofstede [3], used the Hofstede Cultural model and applied it to advertising research. Shackleton and Ali [4] examined managers in different corporations to understand their work-related values and culture based on the Hofstede model, and Erman and Medeiros [5] tried to examine if there is a correlation between cultural characteristics (via the Hofstede model) and COVID-19 death rates in different countries. Therefore, culture may heavily influence other socioeconomic aspects.

A visual representation of cultural models has been attempted by researchers within the domain of cultural science. However, current tools for the visualization of cultural models

make use of rudimentary representations, such as bar charts, boxplots, and scatterplots [6]. All current implementations are mainly static and inflexible. Conjointly, they do not provide the necessary insights on several aspects of the employed cultural models, such as insights on correlations, patterns, or cultural partitions. These visualizations have so far only represented a small subset of the data and are, therefore, missing valuable information. This in-depth knowledge is necessary to gain a better and more complete understanding of the multitude of available cultural data and of how the cultural models have developed. At the same time, it can support the application of the gained knowledge to various professional fields and business-related applications.

Often researchers have to cope with missing data—also within the exploration of cultural models. Acquired datasets can contain missing data, which reduce the "representativeness" of the sample and can, therefore, distort inferences about the population, leading potentially to wrong conclusions. Therefore, it is necessary to handle missing data in an appropriate way. Like any other domain, the cultural science domain is not an exception in coping with missing data. For this reason, we also investigate methods and approaches of handling missing data and include it in our framework.

Without adequate visualization tools, obtaining such an in-depth insight into the complex phenomena that determine cultures and societies is a challenging task. The domain of Visual Analytics [7] is able to provide suitable alternative solutions for the exploration and analysis of multi-variate heterogeneous data [8], which describe characteristics of cultural models and of society, by combining the strengths of visualization with automatized analysis processes and with other disciplines, such as statistics, in highly interactive and expressive environments.

1.2 Aim of the Work

Our research question is: *How can Visual Analytic strategies aid the exploration and knowledge discovery in cultural and societal models?* The purpose of this master's thesis is to review and comprehend current Visual Analytics methods and tools for the exploration and analysis of cultural models. Subsequently, we aim to design and develop a framework for the interactive representation of cultural models, for the generation of new knowledge about cultural models and societal links, and for the confirmation of new hypotheses within the domain of cultural sciences. The outcome of this research is the development of a new framework, which supports the flexible manipulation (addition or removal) of dimensions to a cultural model, facilitates the analysis of the underlying data and exciting phenomena in them (e.g., correlations, patterns, cultural partitions) and provides a flexible, interactive framework for new knowledge discovery or hypothesis confirmation/generation.

1.3 Methodological Approach

The principal intention of this master's thesis is to create a new framework, which visualizes information retrieved from cultural models for analytical purposes. The end product of this research is a framework that can visually represent data. The design research paradigm is concerned with creating and designing new frameworks that meet the requirements and have an additional research value [9]. It is suitable to be used as our methodical framework in this master's thesis since we aim to create a framework. This thesis is oriented on the three-cycle research framework introduced by Hevner [10], which is based on the already existing Information Systems (IS) research framework created by himself a couple of years before in 2004 [11]. It overlays there cycles (described in Section 1.3.1, 1.3.2 and 1.3.3) on the preexisting IS research framework. Figure 1.1 gives an overview of the different stages.

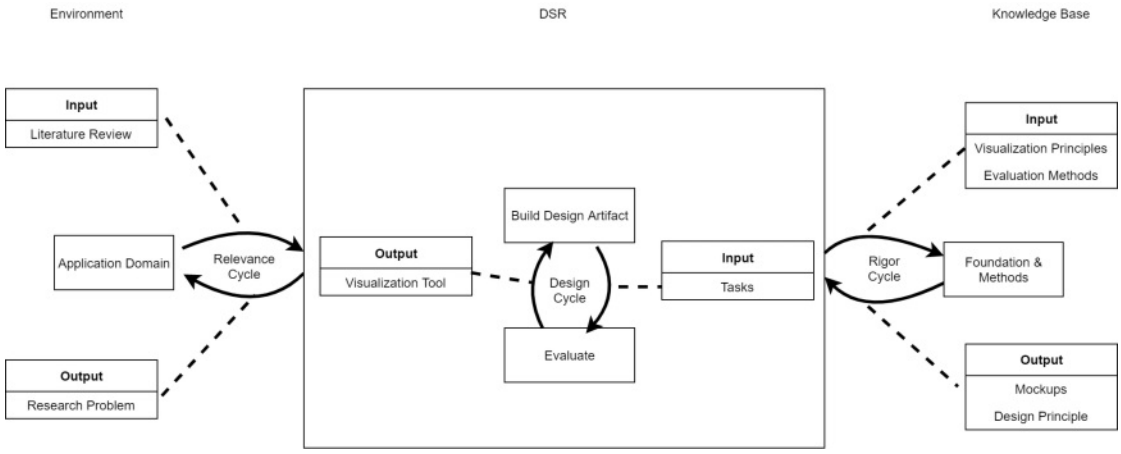


Figure 1.1: Three cycle methodology of design science research adapted from [10] - Details of relevance, design, and rigor cycle used in this master's thesis.

1.3.1 Relevance Cycle

Hevner's research model primarily starts with identifying a problem and the desire to enhance and improve processes. This desire can be seen as the basic problem description, where the incremental improvements are generated in an environment that is being interpreted as "artifacts" by Hevner. Here, as an artifact, we denote our framework. This thesis starts by identifying a demand for a framework by reviewing existing literature from both the visualization and cultural science domain. We first identify the existing methods of knowledge discovery in the cultural science domain. Furthermore, we study what the Visual Analytics domain can offer to discover knowledge in the cultural science domain. We focus on the literature gap between what exists and what can be offered by the Visual Analytics domain. The literature review is an input into the cycle, and the output is the research problem with the specific domain challenges to be addressed and

the tasks to be fulfilled by the final product. We discuss more the components of the relevance cycle in Chapter 2.

1.3.2 Design Cycle

The design cycle includes the construction, i.e., design and evaluation of the developed framework in different iterations until the goal is achieved. This is usually when the problem described in the relevance cycle is solved, the identified challenges are solved, and all tasks are fulfilled. In this cycle, we have the tasks, i.e., well-defined items that our visualization framework needs to perform to solve the research problem as input. The tasks are influenced by the rigor cycle (discussed in the following subsection), which gets its input from guidelines, i.e., established visualization principles and evaluation methods. Using these guidelines, our framework is designed, developed, and evaluated. If the framework's components successfully address our research problem, then the design cycle can be stopped. In order to ensure that the designed framework answers the research question, an evaluation is also required. The evaluation methods are explored and defined in the rigor cycle, as discussed in the next subsection. We discuss more the different aspects of the design cycle in Chapters 3–5.

1.3.3 Rigor Cycle

The rigor cycle is being described as the source of knowledge gathered from related research papers, experiences, experts, and theories. Since this thesis is mainly related to the visualization of cultural data, the scientific source of the rigor cycle are guidelines, principles, and frameworks introduced by other research in the visualization domain. Additionally, it is crucial to find suitable evaluation methods which are employed to approve or reject the designed framework. As a consequence, each iteration in this cycle might be unique and evaluated differently, as it is an ongoing process and does not have a definitive end. The output of the rigor cycle is design principles and mock-ups of the system, which are based on visualization principles. These mock-ups are then be used in the design cycle as a reference to ensure that the framework complies with the appropriate principles of visualization while also satisfying the previously determined tasks. The rigor cycle is further addressed in Chapters 4–5.

1.3.4 Evaluation

Numerous authors have used design science research in their visualization research [12, 13, 14, 15, 16]—each conducting different ways of evaluation. Reviewing the evaluation methods for visualization, we determine two of them are suitable for evaluating the outcome of this master's thesis since they are both designed to evaluate research conducted in the information visualization domain. First, Lam et al. [17] reviewed a broad literature survey of more than 800 visualization papers and derived seven guiding scenarios describing evaluation practices and common scenarios in information visualiza-

tion. Second, Munzner [18] presented a nested model for the validation of visualization strategies in four layers. We discuss more the evaluation of our outcomes in Chapter 6.

1.4 Contributions

There are two significant contributions in this master's thesis. First, this work contributes to an *interactive and flexible visualization framework that can be used in the cultural science domain* to explore and discover new knowledge. With the aid of Visual Analytics strategies, this framework provides the user with dynamic plots that can be interactively explored to discover new knowledge or confirm existing hypotheses from the cultural science domain. Our work's second contribution is *a study on the steps that should be followed to create such a visualization framework*. In the process of this master's thesis, we convey our journey in the form of a reproducible methodology, which other researchers can use to create similar visualization frameworks for the domain of cultural science.

1.5 Structure of the Work

This master's thesis has the following structure:

Chapter 2: State of the Art This chapter reviews the current literature on cultural models and defines the Hofstede cultural model in detail. We review existing approaches of visualization of cultural models.

Chapter 3: Principles of Visualization Since the main focus of this master's thesis is visualization, we explain in this chapter in detail basic concepts from the domains of information visualization and Visual Analytics. We also describe the process of Visual Analytics and explore how to visualize multidimensional data and missingness, i.e., uncertainty. Lastly, we review existing methods of evaluation in visualization.

Chapter 4: Visualization Design In this chapter, we define the requirements of our visualization framework. We break down the requirements into tasks that the framework needs to support and create UI mock-ups based on them.

Chapter 5: Implementation Here we explain how we have implemented the visualization framework and what technologies have been used in the process.

Chapter 6: Results This chapter follows the methodology of three different case studies and attempts to re-produce their research result. We then evaluate our visualization framework and create its final version. We also create a methodological framework out of our entire process.

Chapter 7: Conclusion and Future Work Finally, in this chapter, we briefly summarize the process of the master's thesis and provide ideas for future research directions.

CHAPTER 2

Cultural Models Background and State of the Art

2.1 Cultural Models

Cultural models aim to ease the process of comparing countries with each other by introducing dimensions for culture and assigning values for different countries based on certain data [19]. As mentioned in section 1.1, there are different cultural models, and in this section, the two most common models are discussed.

2.1.1 Hofstede Model

Geert Hofstede was a Dutch psychologist and researcher. He believed that many societal variables could potentially be explained with four major dimensions [20]. Hofstede created in 1970 a four-dimensional cultural model based on a survey questionnaire conducted with 116,000 employees in IBM corporation, located in 72 different countries. This approach was well received and reviewed positively by leading psychologists and sociologists such as Eysenck [21] and Triandis [22]. The initial four dimensions of the model, as published in his book [23] are the following:

2.1.1.1 Dimension 1: Power Distance (PDI)

This dimension is related to social inequality. Hofstede [23] defines this dimension as "The extent to which the less powerful members of institutions and organizations within a country expect and accept that." This dimension can be reflected in role pairs such as boss–subordinate, parent–child, and student–teacher. Table 2.1 shows differences between a society with a small power distance as compared to a society with a large power distance. For example, countries such as Russia, Romania, Malaysia, and China

Small Power Distance	Large Power Distance
Minimized inequality	Inequality is expected
Parents have a respectful relationship towards their children	Parents' relationship towards their children is based on obedience from children
Learning is based on communication between students, and teachers	Learning is merely based on the excellence of the teacher
Strict hierarchical structure in companies	A rather flat structure in companies where superiors and subordinates are treated equally

Table 2.1: Difference between small and large power distance in a society.

Collectivist	Individualist
The word "I" is not being used	Regular use of the word "I"
The larger portion of tax goes into public healthcare	A small portion of tax goes into public healthcare
More people are introverts	More people are extroverts

Table 2.2: Difference between a collectivist and individualist society.

have high values in the PDI dimension, whereas countries such as Austria, Denmark, Ireland, and Israel have low values in this dimension.

2.1.1.2 Dimension 2: Collectivism vs. Individualism (IDV)

This dimension is defined by Hofstede [1] as "individualism on the one side, versus its opposite Collectivism, as a societal, not an individual characteristic, is the degree to which people in a society are integrated into groups." In an individualist society, every single person puts themselves before everyone else, and they look after themselves instead of looking for each other within the society. On the other hand, in a collectivist society, people feel integrated into a group and have a strong feeling of loyalty towards the remainder of the society [24]. Countries such as the United States of America, Australia, Canada, and the Netherlands are individualist countries (have a high score in this dimension), while countries such as Colombia, Costa Rica, Indonesia, and South Korea are collectivist, meaning that they a have low score. Table 2.2 shows the key differences in values and norms between a collectivist and individualist society.

2.1.1.3 Dimension 3: Femininity vs. Masculinity (MAS)

The definition of femininity vs. masculinity as a dimension of the Hofstede model is that "masculinity, versus its opposite femininity, again as a societal, not as an individual characteristic, refers to the distribution of values between the genders which is another fundamental issue for any society, to which a range of solutions can be found." In a masculine society, roles between the genders are clearly distinguished. Men are supposed

Feminine	Masculine
No clear roles between men and women, both are equally contributing towards family and career	Clear split in roles between men and women in the society
Both male and female are modest	Men should be assertive, goal-oriented, and tough
Family-oriented society, where relationships are important	Career, success, income, and recognition are important

Table 2.3: Difference between a feminine and masculine society.

to be focused on careers, should be tough, and show no emotions. Women are supposed to be concerned with tender roles, such as taking care of the home and children in the family [19]. A feminine society does not have predetermined roles for its members based on their gender. The Republic of Slovakia, Japan, Hungary, and Austria is rather masculine countries—meaning that they have a high score in this dimension. While Sweden, Norway, Latvia, the Netherlands, and Denmark have low scores in this dimension and are, therefore, feminine countries. Table 2.3 shows the key differences in values and norms between feminine and masculine society.

2.1.1.4 Dimension 4: Uncertainty Avoidance (UAI)

Hofstede [1] defines this dimension as: "Uncertainty Avoidance is not the same as risk avoidance; it deals with a society's tolerance for ambiguity. It indicates to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations." In a country with high uncertainty avoidance, people tend to avoid uncertain situations due to reasons such as anxiety and fear. These countries feel threatened by ambiguous situations and are not keen on experiencing new things. On the other hand, countries with low uncertainty avoidance are open to new challenges, are curious to explore new and unknown things, and do not feel threatened by them [24]. Countries such as Greece, Portugal, Malta, and Uruguay have high values in this dimension, meaning that they are not tolerant when it comes to uncertain situations. While Singapore, Denmark, China, Sweden, Vietnam are countries, which are more open towards risks and uncertain situations, meaning that they have a low score in this dimension. Table 2.4 represents a comparison of values and norms between a society with high and low uncertainty avoidance.

2.1.1.5 Dimension 5: Long Term vs. Short Term Orientation (LTO)

In 1991, Hofstede et al. [25] added an extra dimension to the initial four dimensions described above. This was due to a study conducted on 23 different countries using a Chinese Value Survey (CVS)[26]. This study published a four-dimensional model, three of which were correlated with three dimensions from the Hofstede model. However, there was one dimension that did not have any overlap. For this purpose, Hofstede

Weak Uncertainty Avoidance	High Uncertainty Avoidance
Curious to explore unknown	Conservative to unknown situations
Low anxiety and no stress to new situations	High anxiety and stress to new situations
People are more likely to do risky investments	People stay away from risky investments

Table 2.4: Difference between high and low uncertainty avoidance in a society.

Short Term Oriented	Long Term Oriented
Spending is encouraged	Sparing resources is important
More effort into quick result	Effort into slow and stable results
Traditions are very important and strict	Traditions are flexible and can be adapted

Table 2.5: Difference between a short term and a long term oriented society.

added a new dimension to its model named Long Term Vs. Short Term Orientation [20], and defined it as: "Long term pole were perseverance, thrift, ordering relationships by status, and having a sense of shame; values at the opposite, short term pole were reciprocating social obligations, respect for tradition, protecting one's 'face', and personal steadiness and stability." When a country is more long-term oriented, it means that the country is more concerned about stable growth towards long-term goals. Short-term oriented countries are rather focused on the past and have a tendency towards quick results. Japan, China, South Korea, and Taiwan have high scores in this dimension and are long-term-oriented countries. On the other hand, Colombia, Iran, Morocco, and Venezuela are short-term-oriented countries since they have low values in this dimension. Table 2.5 shows the key differences in values and norms in long-term versus short-term societies.

2.1.1.6 Dimension 6: Indulgence vs. Restraint (IVR)

The sixth and last dimension was added by Minkov [27]. This dimension is the outcome of his latest research and describes to what extend people in a society can enjoy happiness in their life. Minkov states that "indulgence stands for a tendency to allow relatively free gratification of basic and natural human desires related to enjoying life and having fun. Its opposite pole, restraint, reflects a conviction that such gratification needs to be curbed and regulated by strict social norms". Countries such as Colombia, El Salvador, Mexico, and Venezuela have a high score in this dimension and are indulgent societies. Countries such as Bulgaria, Estonia, Latvia, Lithuania, and Pakistan have a low score in this dimension and, consequently, are restrained societies. Table 2.6 shows a comparison between the characteristics of indulgent and restrained societies.

Restrained	Indulgent
The majority of people are happy	Majority of the population unhappy and unsatisfied
Leisure is very important	Leisure is not important, there is no time for it
Likely to remember some positive memories	Likely to remember negative thoughts

Table 2.6: Difference between indulgent and restrained society.

2.1.2 GLOBE's Project

Another example of a cultural model is coming from the GLOBE (Global Leadership and Organizational Behaviour Effectiveness) project. In a period of the three years between 1994 and 1997, a study on roughly 17,000 managers working in 1,000 different organizations was conducted. The outcome of this study was published by House et al. [28]. Here, nine dimensions were used to describe a cultural model. GLOBE's nine dimensions were based on Hofstede's dimensions:

- PDI and UAI were maintained (not necessarily with the same interpretation).
- MAS was split into two dimensions: Assertiveness and Gender Egalitarian.
- IDV was split into two dimensions: In-group and Institutional Collectivism.
- LTO changed name: Future Orientation.
- New dimension Human Orientation was added.
- New dimension Performance Orientation was added.

As the Hofstede model is more popular in the cultural sciences than the GLOBE model, we decided to focus on the former. Yet, theoretically, it should be possible to apply our method to any other cultural models with a finite amount of dimensions.

2.2 Visualization of Cultural Models

There is no dedicated method for the visualization of cultural models. Different researchers have used different techniques to represent cultural models. In this section, we discuss related visualization strategies and their limitations.

A world map for each different dimension was used by House et al. [28]. This can also be found on Hofstede's website [29] and was used by Zhang [30] in their research to compare the dimensions between a set of countries. Figure 2.1 shows the world map for all six dimensions of the Hofstede model beside each other. A scale on the bottom left side

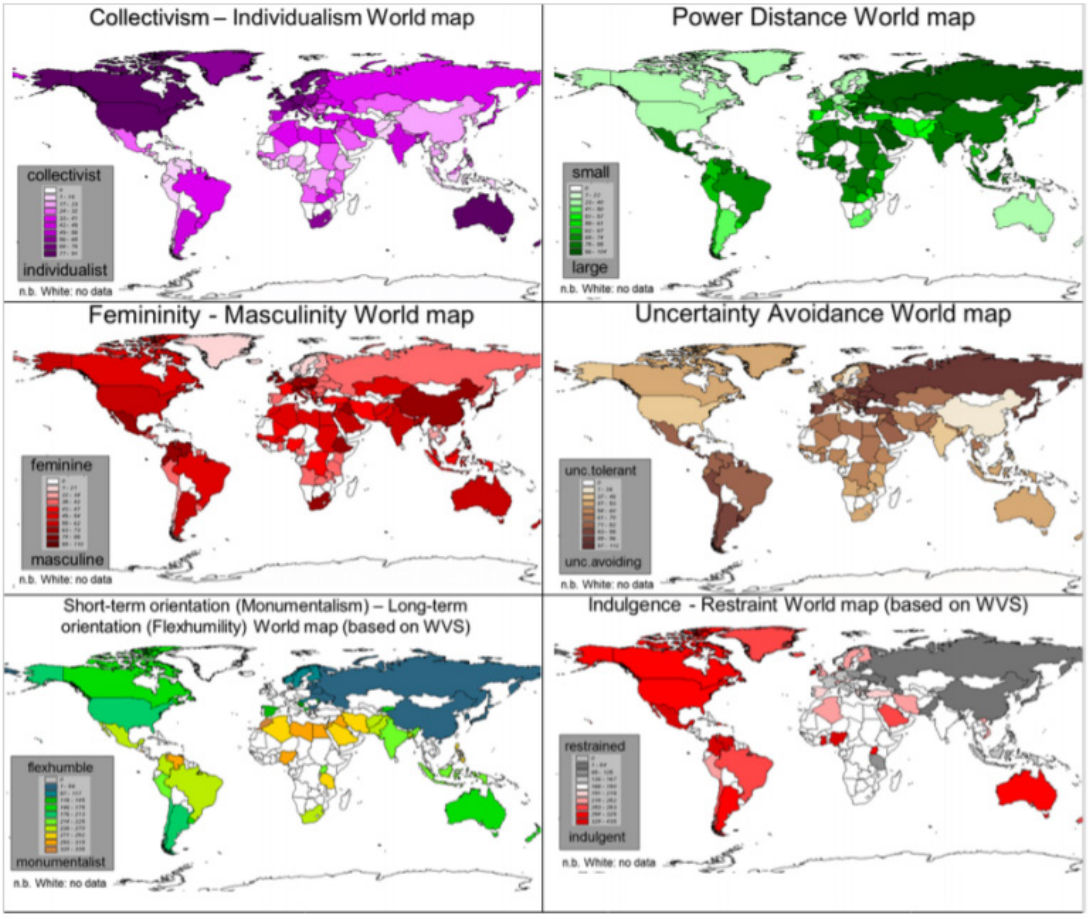


Figure 2.1: Visualization of six dimensions of the Hofstede Model using a world map [27]. Each dimension is represented in a separate panel; hence, six different panels are required for this visualization.

shows a range of values corresponding to the saturation of the color on the map. For example, saturated purple indicates more individualist countries. One limitation of this method is that this visualization lacks accuracy in the comparison between countries, i.e., it is possible to conclude that the United States has a higher value in individualism compared to Russia; however, it is not possible to specify the exact difference in the value, without additional interactions. Also, for the comparison of six dimensions, six maps are needed across six different panels, which require a higher memory load from the user.

Bar charts have also been used to compare different values of Hofstede's model across different countries. An online tool named "Hofstede Insights" exists for this purpose,

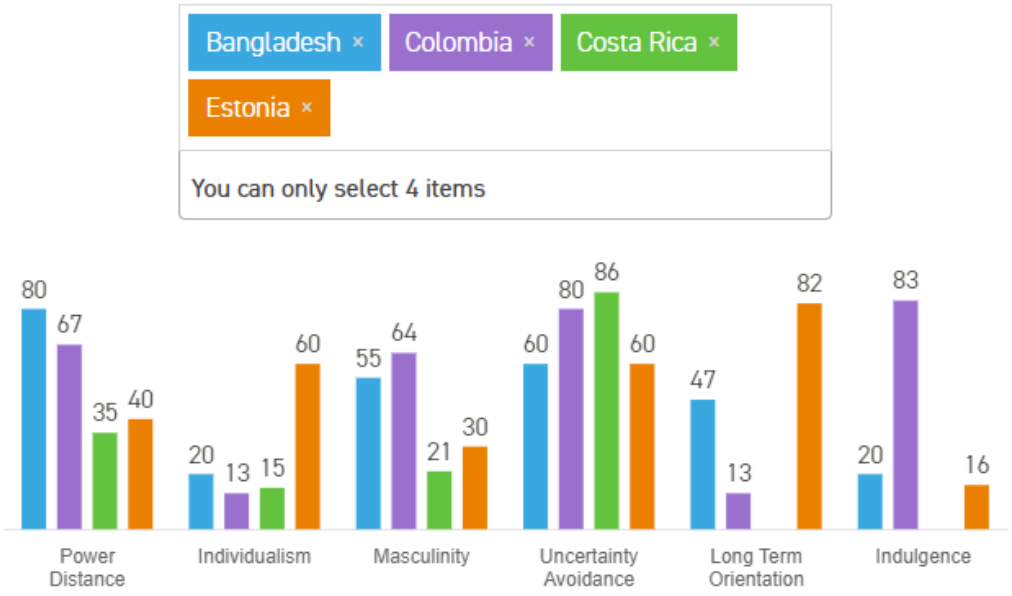


Figure 2.2: Visualization of the Hofstede Model using bar chart Insights [31]. Comparing the six dimensions of the Hofstede model for Bangladesh, Colombia, Costa Rica, and Estonia. Missing values are left blank, while the visualization supports by default the comparison of up to four countries.

which is able to visualize the data of different countries [31]. Each country is represented by a bar chart color, and the data for each dimension is displayed on the Y-axis. The limitation is that only a limited amount of countries can be added to this type of chart, as the Hofstede Insights tool can only support up to four countries. Figure 2.2 shows a visualization example of a comparison between Bangladesh, Colombia, Costa Rica, and Estonia for all the dimensions of the Hofstede model. Here we see that, for instance, Bangladesh has the highest value (80) in the Power Distance dimension, while Costa Rica has the lowest value (35). Additionally, there is no specific encoding for missing values. As seen in the dimensions of Long Term Orientation and Indulgence, Costa Rica's values are missing, and there is only a blank space in the bar chart to indicate this.

Other related work for the visualization of Hofstede's cultural model does not go beyond simple, static visualizations, which are mainly used for the dissemination of the results of cultural studies as illustrations in manuscripts. Rana et al. [32] used a bar chart to visualize a comparison between Brazil and the United States. Xiumei and Jinying [33] attempted to compare China and the United States. Yap [34] visualized the six dimensions of the Hofstede model using a bar chart to show a comparison between Malaysia and Australia's dimensions. Bar charts have been used in other works as well to show a comparison between the values of Hofstede's model within a set of countries

[35, 36, 37]. Other approaches used scatterplots. This approach has been limited to four dimensions only. Each country is represented by a dot on the chart. The individual values for each dimension can be read from a country's position with respect to the axis representing the dimension [38, 39]. An example of a scatterplot can be seen in Figure 2.3a. In other previous work, Hofstede's dimensions for different countries are represented with tables to enable comparison [40, 41, 42]. An example of the tabular comparison can be seen in Figure 2.3b.

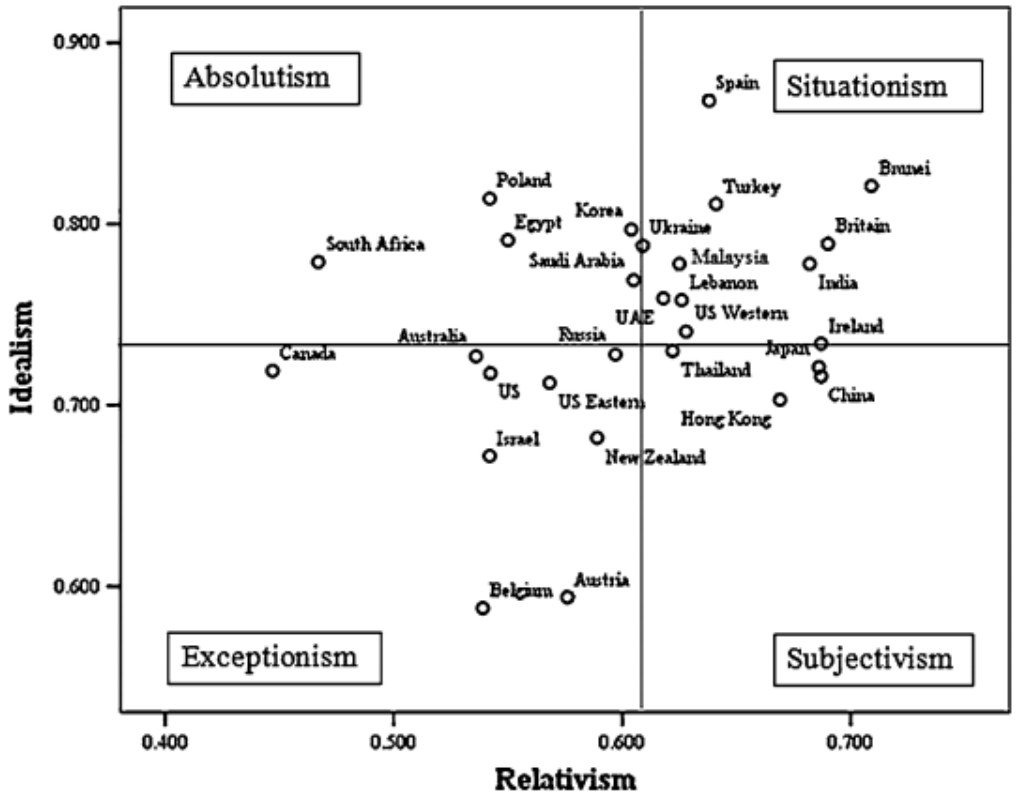
The study result of the GLOBE project is shown in a more interactive way on their website [43]. First, there is a world map view where a country can be selected. After selecting the country, the data is displayed. Second, for each dimension, a boxplot is created, on which the highest and the lowest value, as well as the mean value and the selected countries position, is marked, as shown in Figure 2.4a. House et al. [28], however, uses a radar chart (spider chart) to represent the values of each dimension for a specific country, as seen in Figure 2.4b. In this radar chart, we see all the dimensions of the GLOBE project.

All of the aforementioned visualization methods serve only the comparison across different countries, but they do not support an interactive analytical process for knowledge discovery. The visual representations (with the exception of the GLOBE project) are not interactive, and each of the figures shown above needs to be re-drawn if some other country needs to be represented. Also, the visual comparison of countries is tedious. For instance, in order to compare eight countries together, eight different boxplots, scatterplots, or bar charts need to be drawn. Other tasks, such as correlation, pattern, cultural partitions, and outlier identification, are not supported, while data missingness is addressed as a simple missing visual attribute in the representations.

Reviewing existing visualization tools in the field, we came across a visualization tool implemented by Bayat [44] to be useful as a basis for our purposes. The visualization tool displays the six dimensions of the Hofstede model on a world map, heatmap, and hive-plot. The dataset is a pre-loaded CSV file obtained from Hofstede's website [45] containing the data of 121 countries. The selection is done via drop-down boxes. A user can select a specific dimension and country by choosing it, the world map then highlights the select country. A color scale on the right side is a reference to indicate how high or low a chosen dimension is. An example can be seen in Figure 2.5, where the chosen dimension is UAI for Greece. In the world map, Greece is marked as yellow and on the heatmap, the value of UAI which is 112 can be seen.

2.3 Cluster Analysis in Cultural Models

The process of creating sub-groups (clusters) of data, where the data belong to a specific cluster with similar attributes, is named cluster analysis [46]. It is widely used in computer science for the purpose of grouping data and finding similar patterns within the clusters [47]. Cluster analysis is also used in cross-cultural researches to identify patterns across different cultures. Clustering of countries enables the prediction of attitudes [48, 49].

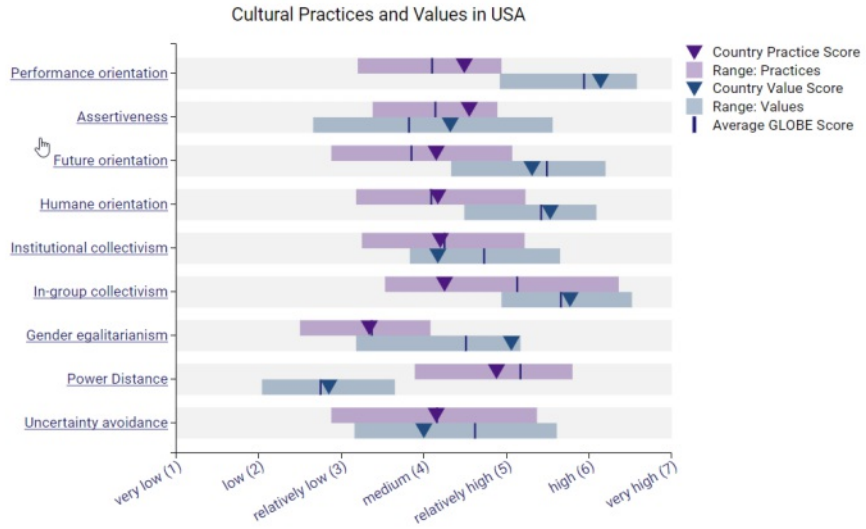


(a) Visualization of Hofstede's IDV dimension using a scatterplot [38].

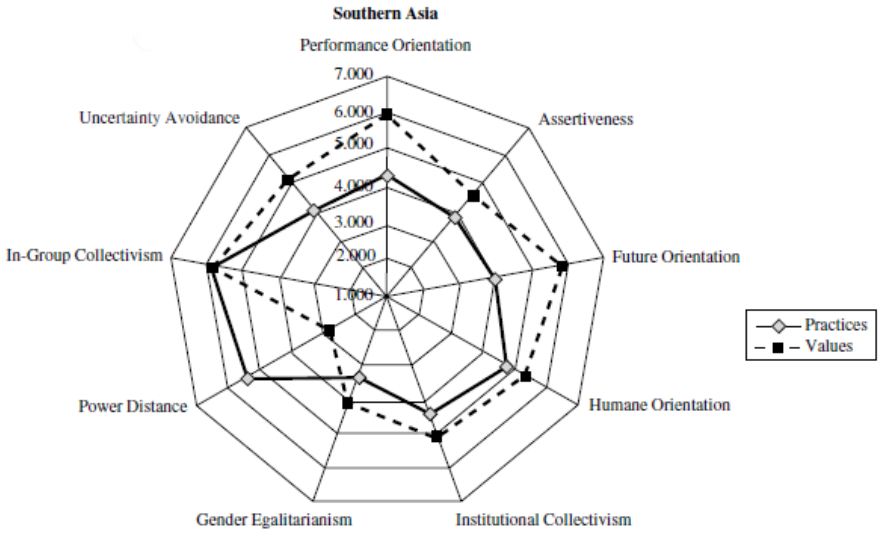
Cultural dimension	Russia	United States
Power distance	93	40
Individualism	39	91
Masculinity/femininity	36	62
Uncertainty avoidance	95	46
Long-term orientation	81	26
Indulgence	20	68

(b) Comparing dimensions of Hofstede's model for Russia and the United States, using a tabular representation [40].

Figure 2.3: Different approaches for the visualization of the Hofstede model, using a scatterplot and a tabular representation.



(a) Visualization of the GLOBE model's dimensions for the USA using boxplots [43].



(b) Visualization of the GLOBE models' dimensions for Southern Asia using a radar chart [28].

Figure 2.4: Radar chart and boxplot visualization method - two different approaches of visualization for GLOBE's model.

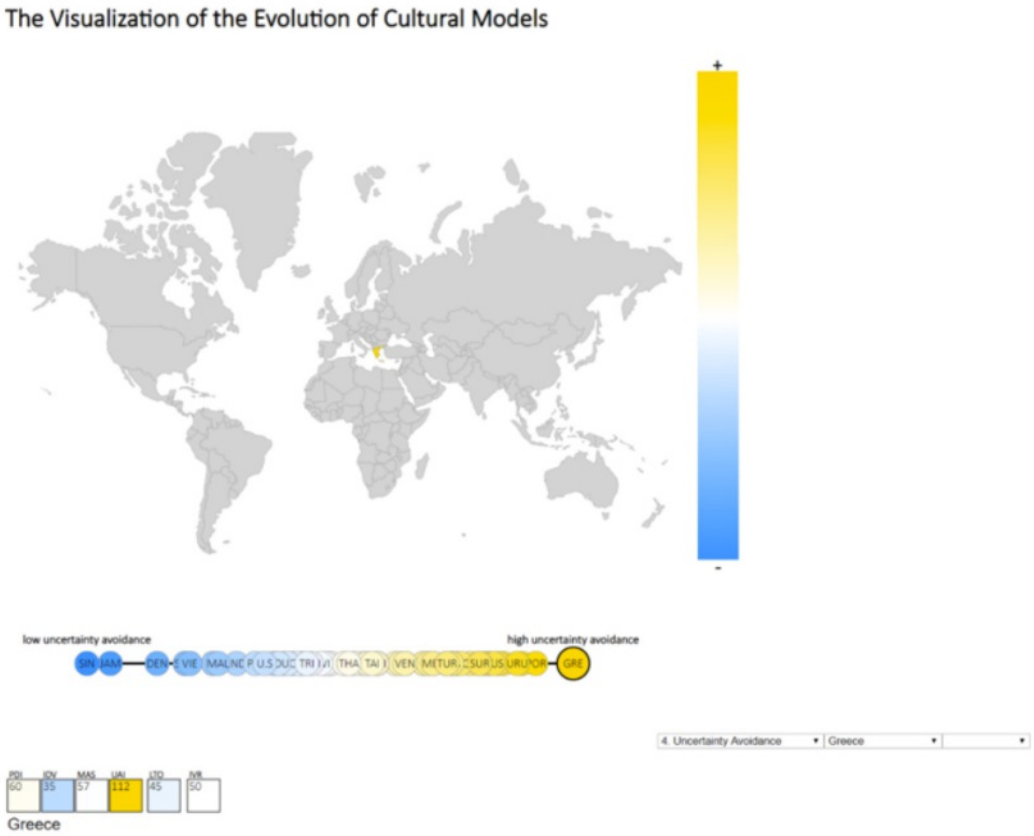


Figure 2.5: Visualization tool implemented by Bayat [44]. The world map, hive-plot, and heatmap show the value of the UAI dimension for Greece, which is selected in the drop-down list.

Clustering can be performed using different techniques, and below, we briefly discuss some of them. Then, we discuss approaches specifically applied to the cultural domain.

2.3.1 The k-means Clustering Method

This clustering method is used for clustering a set of observations into a user-defined amount of k groups. Starting with a random set of k center-points (μ), every observation x is assigned to the nearest center-point as described in the equation 2.1). If more than one observation has the same distance to the center-point, a random one is chosen.

$$S_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \leq \|x_p - \mu_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\} \tag{2.1}$$

After the first step, the mean of the assigned observation x is getting used to re-calculate the center-points; see equation (2.2). This update step gets repeated until all the observations are assigned to center-points.

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.2}$$

The k-means algorithm attempts to optimize the target function shown in equation (2.3). As there is solely a finite number of potential assignments for the number of centroids and observations available, and every iteration must end in an improved solution, the algorithm continually ends in a local minimum.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{2.3}$$

2.3.2 The Hierarchical Clustering Method

Maimon and Rokach [50] explains that in this clustering method, the clusters are created either from a top-down or bottom-up fashion and can be sub-divided into Agglomerative or Divisive hierarchical clustering. In Agglomerative methods, each observation initially is in its own cluster. Then, the clusters are merged until the desired cluster structure is created. Whereas in Divisive hierarchical clustering methods, the observations are initially in one cluster and eventually get divided into sub-clusters till the desired clusters are produced. The result of the hierarchical clustering method is represented as a dendrogram [51], which can be cut at different heights, i.e., at the desired similarity level, to retrieve different groupings.

The hierarchical clustering methods can be further sub-divided into how the similarity is calculated. Different methods of calculation are listed below:

- Single linkage: considers the distance between two clusters to be equal with the distance of the two closest members
- Complete linkage: distance between two clusters is equal to the distance of the longest distance from any member of one cluster to the other
- Average linkage: distance equals the average distance of any point in one cluster to any point of the other

2.3.3 Distance Calculation

In Sections 2.3.1 and 2.3.2, we briefly described how different clustering methods. In both methods, we used the term *distance* numerous times. The distance between two points can be calculated in different ways. Some commonly used metrics as defined by Kaufman and Rousseeuw [52] are shown in Table 2.7. In this master's thesis, we always use euclidean distance in all our calculations for clustering, since it is the most commonly used metric [53].

Method Name	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Manhattan distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maximum distance	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$

Table 2.7: Formula of different calculation methods for distance—Euclidean, Manhattan, and Maximum [52].

2.3.4 Performance of Clustering

Once the clustering is finished, it is possible to measure how well the clustering algorithm has performed. A good clustering method has minimal within-cluster distance and maximal inter-cluster distance. Once again, the distance between the points can be calculated with any of the distance methods discussed in Section 2.3.3.

The Silhouette Coefficient [54] uses two scores of a and b to evaluate the performance of a clustering algorithm. The parameter a refers to the mean distance between a sample and all other points in the same class, and the parameter b is the mean distance between an observation and all the other points in the next nearest cluster. The Silhouette Coefficient for a sample can be calculated using Equation (2.4). A high score in this coefficient means that the density of each cluster is high and the distance to the other clusters is well separated, which relates to the standard definition of a cluster.

$$s = \frac{b - a}{max(a, b)} \tag{2.4}$$

2.3.5 Use of Cluster Analysis in the Cultural Domain

Russett [55] clustered countries in his book into Afro-Asian, Latin American, Western, and Eastern Europe using data from his research. Hofstede [56] then used Russet's approach as an inspiration, and clustered the 53 countries and regions of the IBM survey into 12 different clusters using a hierarchical clustering method with the statistical program named SPSS, and eventually visualized the clusters using a dendrogram seen in Figure 2.6.

Other researchers clustered Hofstede's model alongside additional dimensions in order to find a correlation within the existing dimensions of Hofstede's model, and newly added dimensions [57, 58, 59]. For example, Kökalan [60] made a cross-country comparison of EU countries in terms of female participation in entrepreneurial activity. In this study, five economic variables were used: Women's unemployment rate, Gross Domestic Product (GDP), foreign direct investment, government expenditure, and women's education level. These were combined with the six dimensions of the Hofstede model to understand

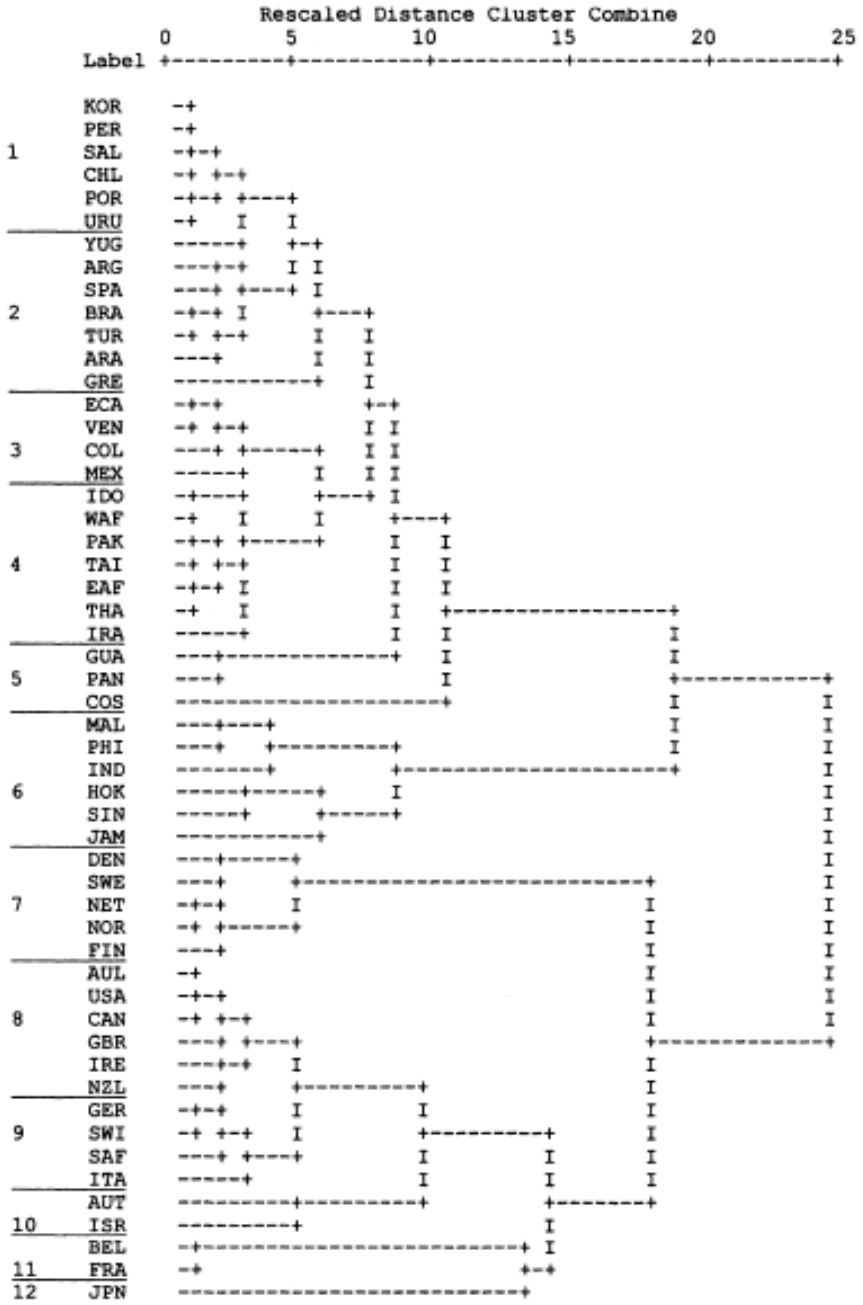


Figure 2.6: Visualization of 12 clusters created in a dataset with 53 countries—identified by Hofstede [56] and shown in a form of a dendrogram. The author divided this dendrogram into 12 arbitrary clusters, the split can be seen on the left side marked with numbers.

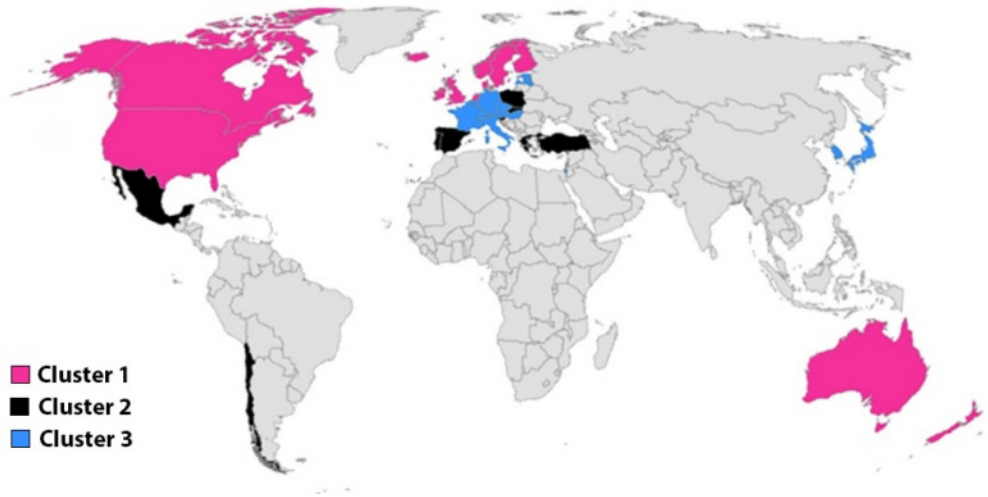


Figure 2.7: Visualization of clusters on a world map by Braithwaite et al. [61]. Countries marked with the same color belong to the same cluster.

the correlation between economic and social parameters. Hierarchical clustering with squared Euclidean distance as a classification method and average linkage as the linkage method was used in this analysis. The result was represented with deprograms and a correlation table for economic variables and cultural variables. Braithwaite et al. [61] used the same approach to compare Hofstede's six cultural dimensions against health system performance in 35 countries, which are a member of Organisation for Economic Co-operation and Development (OECD). The clustering approach used in this study is the agglomeration method of complete-linkage. Additional to the dendrogram, Braithwaite et al. [61] uses a world map to visualize the different clusters with different colors on the map, as shown in Figure 2.7.

2.4 Missing Data in Cultural Models

Dealing with cultural models and social science-related data, we often encounter data missingness. Missing data is a frequent issue, which reduces the available information. For this purpose, various techniques have been identified to fill in the missing values [62, 63], which we discuss in this section.

2.4.1 Categorization of Missing Data

Little and Rubin [64] categorized missing data into Missing data At Random (MAR), Missing data Not at Random (MNAR), and Missing Completely At Random (MCAT), where each of them has different approaches to overcome.

The probability of an observation missing in MCAR is independent of observed and not

observed data, meaning that a missing value has no relation to any other observed and non-observed data. As an example, if we take a random sample from a population, each member has the same probability to be taken into the sample. The members who are not taken (non-observed) are MCAR. If the probability of missing is only the same within the observed data, then we are dealing with MAR; as an instance, taking samples from a population is depending on a known property. Lastly, if neither MCAR nor MAR holds, then MNAR holds. For instance, there might be missing data in a survey due to a reason, and managers tend to not share their salary range, or elderly participants not wanting to share their age [65].

2.4.2 Coping With Missing Data

There are two main approaches to handle missing data, deletion and imputation. In the deletion approach, records of the data which are incomplete are dismissed. Deletion can be list-wised, where only records with all existing variables are used, or pair-wised, where a new sample of complete cases for each variable is used. Depending on the pattern of missingness, deletion can cause problems in the data analysis such as biased estimates [66, 67].

Imputation replaces the missing data with a new value based on an arithmetic calculation rather than dismissing the data completely. Imputation can be single or multiple, wherein single imputation all missing values in a feature are filled with one value [68]. Methods used for single imputation are the following:

- Mean imputation: Missing values to be replaced with the mean value of the observed data.
- Median imputation: Missing values to be replaced with the median value of the observed data.
- Most frequent imputation: Replaces missing data with the most frequent value among each column of observed data.

In order to overcome the issues with single imputation, Rubin [63] introduced a novel method of imputation named multiple imputation (MI). MI uses both Bayesian and classical statistical techniques to predict the missing value. This method aims to handle missing data by keeping existing relationships and reflect uncertainty [69]. The main concept behind MI is to fill in missing values multiple times, analyze by standard techniques and eventually combine them to a single best result. In contrast to ad-hoc methods of imputation (deletion, mean, median, and most frequent imputation), this method considers the uncertainty of the data [70, 71].

Multiple imputation by chained equations (MICE) [72] is a MI method, which operates under the assumption that all missing values in the variables are MAR, then the data are imputed in four major steps.

1. A single imputation method (ex. mean imputation) is performed for every missing value. The mean imputations are referred to as M.
2. M is set to missing for one entry (referred to as E).
3. E is regressed, meaning E is the dependent variable in a regression model and all other variables are independent.
4. The missing value that was created for E is replaced with the prediction(imputation) from the regression model. When E subsequently is used as an independent variable in the regression model, both observed and imputed values are used.

2.4.3 Evaluation of Imputation

An imputation method's goal is to complete incomplete data; thus, the quality of this imputation method shall be evaluated based on this goal. The evaluation is challenging since when a dataset has missing entries, the truth values of those missing entries are unknown. A comparison in the performance of the imputation method can only be achieved by artificially introducing missing data points in the dataset [73].

Once there is an artificially created dataset with missing entries, the performance of imputation can be measured by first imputing this dataset and then comparing it with the original values. A widely used method to evaluate the performance is the Root Mean Square Error (RMSE) score [74, 75, 76, 77, 78]. Equation (2.5) shows how the RSME value is being calculated, where n is the size of the sample, $\hat{y}_i$ is the predicted (imputed) value, and y_i is the original (truth) value. The smaller the RMSE is, the better effective was the imputation method, since the predicted values are closer to the original values.

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2.5}$$

2.4.4 Handling Missing Data in the Cultural Domain

The imputation techniques described above are used frequently in the cultural science domain. Yoon [79] tried to predict the voice behavior of employees based on the PDI dimension of the Hofstede model. This research coped with missing data by using the list-wise deletion method. Huang and Crotts [80] focused on finding a correlation between tourist satisfaction and the six dimensions of the Hofstede model. The main dataset used in this study had a total of 39,959 records which 15,997 of them were related to the scope of the study. Using the list-wise deletion method, 14,892 values out of the total available records were used in the study. Several other studies related to the Hofstede model or cultural models, in general, can be found which use list-wise or pair-wise deletion methods [81, 82, 83, 84, 85], single imputation methods [86, 87, 88, 89] or multiple imputation methods [90, 91, 92, 93, 94], in order to purify their data.

CHAPTER 3

Principles of Visualization

In this chapter, we take a closer look into visualization principles, visual design principles, and evaluation methods related to creating an analytical tool for exploring multidimensional cultural models. This, in addition to the related work discussed in Chapter 2, builds up the knowledge basis described in Figure 1.1. The knowledge acquired from Chapters 2 and 3 will serve as an input for our first framework design, which is discussed in the upcoming chapters. Two key terms—information visualization and Visual Analytics—are described below to give us a better understanding of how to meaningfully design a visualization framework. At the end of this chapter, the framework's basic requirements among the required representation methods and evaluation methods should be discovered.

3.1 Information Visualization

Card [95] describes information visualization as "the use of computer-supported, interactive, visual representations of abstract data in order to amplify cognition." Information visualization is the art of transforming data into a geometric representation that enables humans to achieve a rapid understanding of abstract information [96, 97]. The concept of information visualization is also used in the computer science domain to denote the field that enables users to benefit from visual data exploration and data analysis. The process of visual data exploration is seen as a method that allows users to better understand the data by exploring and interacting directly with the data. It could also be used for knowledge discovery, confirmation of hypotheses, decision making, or presentation of results [98].

3.1.1 Process and Methods of Information Visualization

The process of information visualization consists of four main components: data, representation, presentation, and interaction [97]. In the representation stage, methods of data

representation are identified. Methods such as point representation, scatterplots, star plots, or parallel coordinate plots exist, which depending on the following three principal as described by Spence [97] are differently suitable for a purpose:

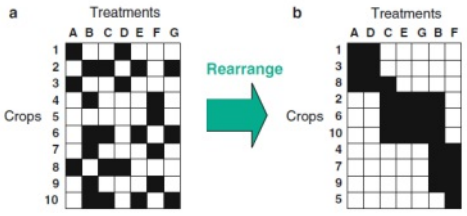
- Type: Depending on the data type, which can be numeric (i.e., price of real estate), categorical (i.e., gender), or relational (i.e., a family tree).
- Dimension: The representation method is strongly influenced by the number of dimensions in the data.
- User: The person who interprets and explores the data has an influence on the method of representation user. As an example, if an information visualization tool is mainly used by children, complex numerical representation methods would not be suitable to use.

The above principles are crucial through all stages of information visualization. In the presentation stage, the main goal is to identify the most suitable methods to present the previously identified data representations, i.e., to show it on display. Spence [97] argues that a single-page presentation with adequate size, with a minimal amount of references to any other page, shall be enough in the majority of information visualization cases. He also mentions that a computer-based display has the ability to provide features (such as filtering and interaction) for data presentation, which is difficult or impossible for paper-based documents to match.

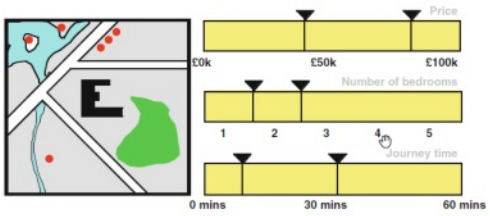
The aim in the last stage of Spence's information visualization process is to use techniques such as reordering data, filtering, and brushing to enable users to interact and change the presented visualizations. Reordering of data requires rearranging the data to show the result more understandably. Filtering, however, does not rearrange the data, as it only removes irrelevant data by filtering them out. Brushing was introduced by Becker and Cleveland [99]. This is a technique that enables the user to select a subset of the data in space and to view their corresponding points in another space. An example of each method can be found in Figure 3.1.

Figure 3.1a shows two tables representing the results of the same experiment. The rows are a collection of crops, and the columns represent different treatment methods. Black indicates a successful treatment, and white indicates an unsuccessful treatment. After rearranging the table, a pattern is visible, raising questions about a particular type of treatment and crops.

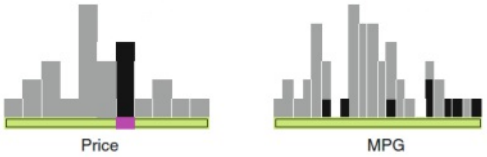
Williamson and Shneiderman [100] created a Dynamic Home Finder tool to make the exploration of data easier. In this tool, sliders provide the user with a dynamic user interface that allows the user to filter properties based on price, number of bedrooms, and commute time to their workplace. An illustration of the tool is in Figure3.1b. In this example, the user applied a price filter ranging between roughly £45,000 to £80,000,



(a) Reordering data [97] used as an interaction method.



(b) Filtering data [100] used as an interaction method.



(c) Brushing data [101, 102] as an interaction method.

Figure 3.1: Example of different interaction methods such as reordering, filtering, and brushing in information visualization [97].

several bedrooms between 1.5 and 2.5, and distance to their workplace between 10 to 30 minutes. The red dots show available properties which meet these criteria.

An example of the brushing method can be found in a tool named Attribute Explorer, which uses brushing in a histogram diagram that [101, 102]. The data belongs to a specific collection of 50 car prices and their Miles-per-Gallon (MPG) rating. Figure 3.1c shows the relation between a category of price and its corresponding MPG, where the selection of price is brushed on the second histogram.

The above-defined interaction methods are used as an inspiration in our visualization framework to enable users to interact and change the presented visualizations.

3.2 Visual Analytics

As described in Section 3.1, information visualization is mainly concerned with how to represent non-spatial data. This, however, is not sufficient for the purpose of this master's thesis since, in addition to the representation of data, a user should be able to generate hypotheses by analyzing the data. For this purpose, Keim et al. [7] has introduced Visual Analytics, which is a field that to combines statistics, machine learning, or data mining methods with interactive visualizations to enable users for a better understanding and decision making.

The fundamentals and methods used in Visual Analytics are an adapted version of the information-seeking mantra used for information visualization, which broadly describes

the process of having an *overview of the data first—zooming and filtering—and then obtaining details on demand*. The adapted version for Visual Analytics is slightly different; *analyze–show the important–filter, zoom, and analyze further–details on demand* [7]. The main difference is in analyzing the data first before showing it. Following are some concepts in Visual Analytics which are used in this thesis:

- *Multiple views* enable users to see the data in different representations, which gives the possibility to comprehend the relationship between data by brushing and linking.
- *Brushing and linking*, as also described in Section 3.1.1, allows the user to select data on one view and see its corresponding highlighted data in another view.

3.2.1 Process of Visual Analytics

Keim et al. [7] introduced a sense-making loop framework for Visual Analytics initially introduced by Van Wijk [103] represented in Figure 3.2. An initial analysis needs to be applied to the data followed by visualization; the user then can discover new knowledge and ultimately confirm hypotheses. The user has control over the process and has the ability to interact with the visualizations to fulfill their tasks.

By reviewing the above Visual Analytics process and the information visualization process introduced by Spence [97] it is possible to apply them to this master's thesis. Starting with the three principles described in Section 3.1.1, we can define the principles as follows:

- Type: The data required for this visualization framework is acquired from Hofstede's website [45]. The dataset contains 121 rows (countries), with numerical data. The data consist of integer values between 0 and 100, 1 being the minimum and 100 being the maximum value. The data might contain missing values which need to be deleted or imputed. Any additional data added two the model should comply with the two specifications of being an integer between 0 and 100.
- Dimensions: Each country has a maximum of six dimensions by default and can be extended by the user if it required to explore the correlation between the newly added dimension and the default six dimensions. Thus, multidimensional visualization methods should be considered for a proper representation since the data has more than a singular dimension.
- User: The users are researchers with academic backgrounds in the cultural science domain who would like to discover knowledge by exploring the data related to different dimensions of Hofstede's model.

Knowing the type and dimension, it is possible to choose appropriate representation charts for the visualization framework. Some of the potentially suitable charts have been

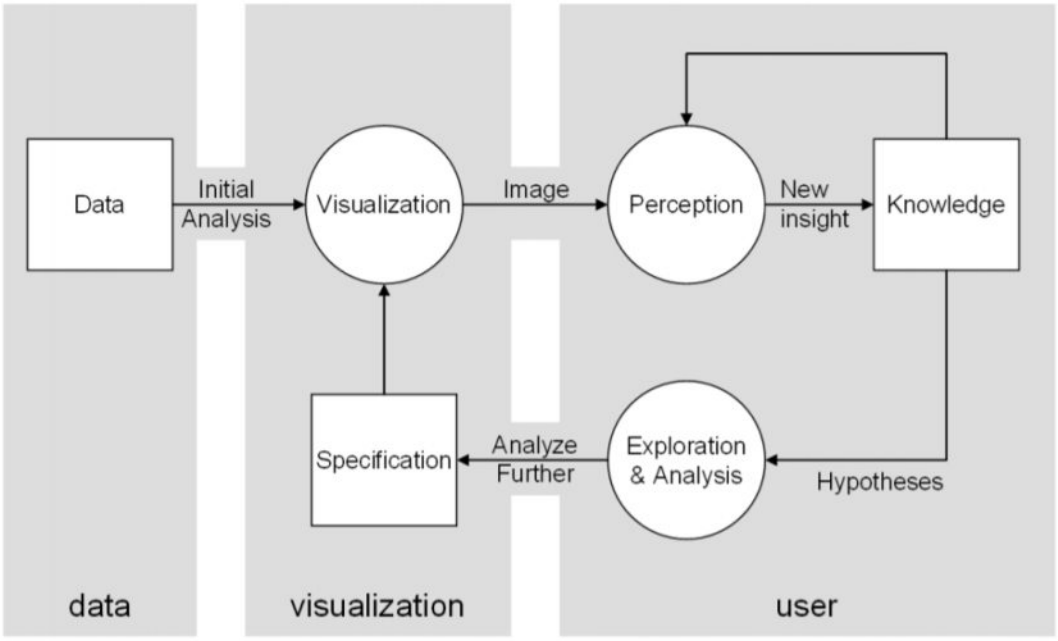


Figure 3.2: Process of Visual Analytics based on sense-making loop introduced by Van Wijk [103]. The user has control over the process and can fulfill tasks with interactions on the visualization tool.

described in Section 2.2, and are representation methods already used in the literature that have the ability to show some aspects of the multidimensional data. In the upcoming section, the required visualization methods and techniques are analyzed as to whether they introduce a possibility for knowledge discovery in the visualization framework.

3.3 Multidimensional Visualization

Numerous researches have explored the visualization of multidimensional data [98]. In this section, we explore some of them which fit the type, dimension, and user of our visualization as described in Section 3.2.

While exploring different multidimensional visualizations, we discovered that the scatterplot matrix and parallel coordinates are two suitable, basic methods and have been heavily used by other visualization tools for representing multidimensional data.

Scatterplots [104, 105] are one of the most widely used visualization methods for multidimensional data, due to their simplicity and yet flexible nature [106]. Scatterplots have a two-dimensional grid. Each entry in the data set is rendered as a point on the grid representing the value in the two-dimensional Cartesian space defined by the axes. In order to visualize multiple dimensions, a different and unique graphical property such as

shape, color or size can be assigned to each dimension. This method was applied to tools such as XmdvTool[107] and GGobi [108], which use scatterplots in their visualizations. Only two or three dimensions can be fit into a single scatterplot. For this purpose, some scatterplot visualizations allow to dynamically select the dimensions which a user wants to see [106]. Alternatively, multiple scatterplots can be set beside each other in a matrix-like configuration, where each dimension has its own dedicated scatterplot, as seen in Figure 3.3b. This is called a scatterplot matrix.

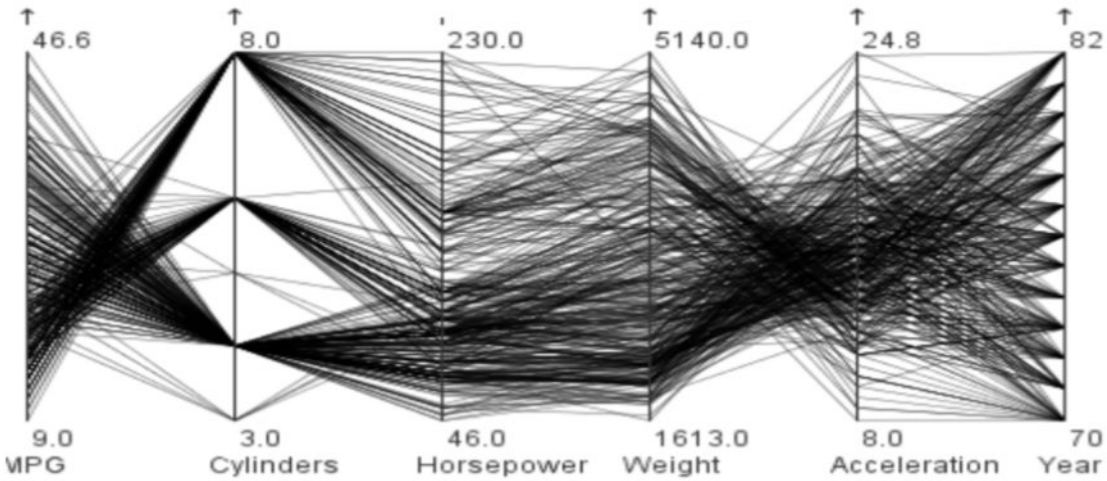
Parallel coordinates [109] are also a widely used representation for multidimensional data. They have the ability to represent N-dimensional data by polyline crossing axes at a given value. Each axis represents a dimension and is parallel to all other axes. Using brushing as a method to interact with parallel coordinates, users are able to understand relations and patterns in the data [110, 111]. Hauser et al. [112] used linking and brushing in order to explore highly dense data in parallel coordinates, as the user is only interest in part of the data and not the whole dataset. An example of parallel coordinates can be seen in Figure 3.3a.

A radar chart or also named spider chart, is a graphical approach to represent multidimensional data in the form of a two-dimensional figure. The name "radar" comes from the similarity to a radar screen [114]. Radar charts have axes integrated into a radial figure; each property's value is presented as a dot on each axis [115]. All the dots are connected via lines creating a polygon or circle-like shape depending on the chart's design. An example of a radar chart can be seen in Figure 2.4b where nine dimensions of the GLOBE project are visualized using a radar chart.

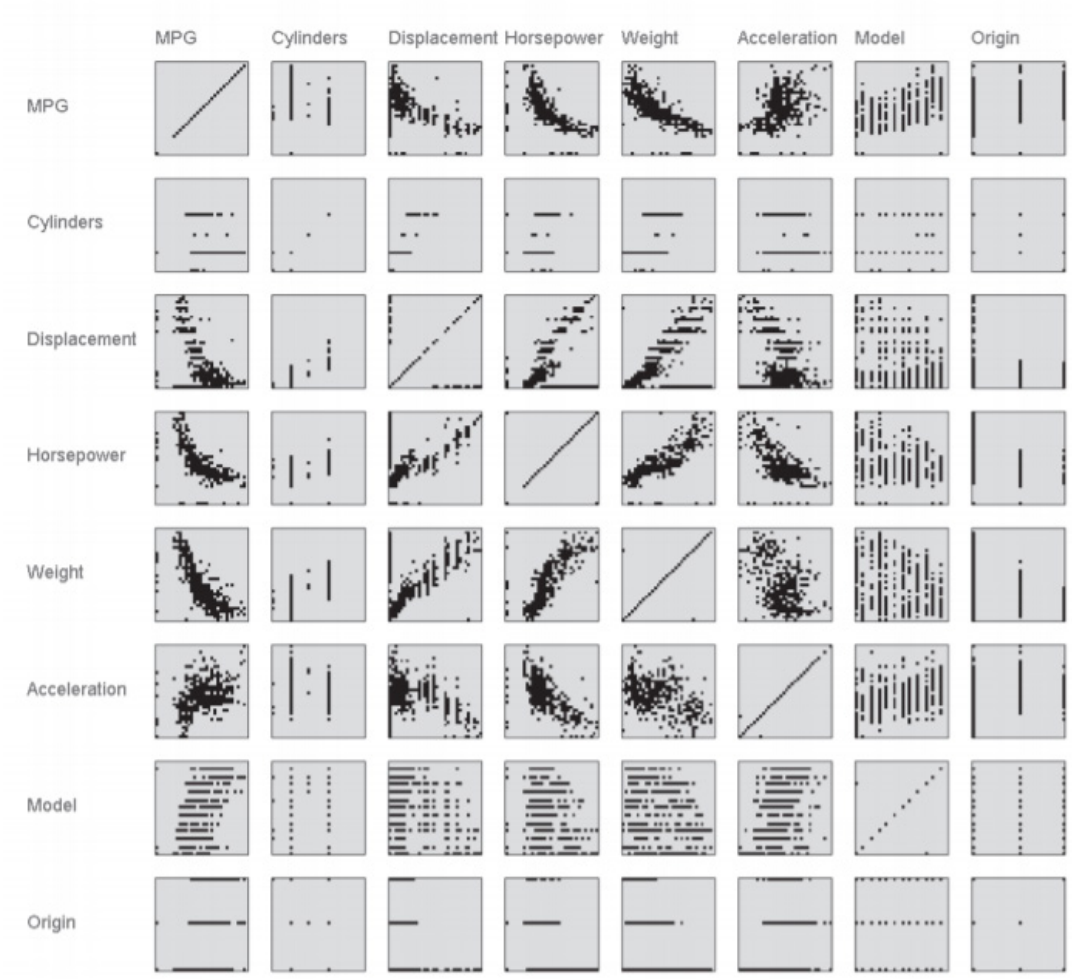
3.4 Visualization of Uncertainty

Uncertainty an often debated term [116]. Uncertainty can be defined as an error, difference to the truth value, missing data, or statistical variation [117]. Uncertainty is an often occurring phenomenon in data derivation, transformation, and imputation. If these kinds of uncertainty are not considered in the visualization, wrong assumptions can be made by the user [117]. Visualization of uncertainty has been ignored in the past by most of the research conducted in the visualization domain [118]. However, the importance of its influence was brought up multiple times [119, 120].

Uncertainty can be visualized in different ways. Size, shape, color, texture, or fuzziness can be used to show uncertainty in a visualization that [121, 122, 117]. The main challenge is to prevent information overflow since uncertainty is yet another channel of information that needs to be interpreted by the user and often tends to occlude the underlying data. The visualization of uncertain data should not pollute the visualization of specific data and not burden the understanding of the user. In an attempt to achieve that, Cedilnik and Rheingans [123] used procedural generated annotation to show uncertainty. As seen in Figure 3.4, the annotation lines of uncertain data have a softer-edged, while places with low uncertainty data have sharp and bright lines.



(a) Parallel coordinates representing the car dataset [106]. Each axis represents a dimension in the dataset, meaning that this dataset has six dimensions. The value ranges of each axis are within the minimum and maximum value of that particular dimension.



(b) Scatterplot matrix representing the car dataset [113]. Each dimension has its own dedicated scatterplot, meaning the dataset consists of eight dimensions.

Figure 3.3: Scatterplot matrix vs. Parallel coordinate plot, representing the same multidimensional data.

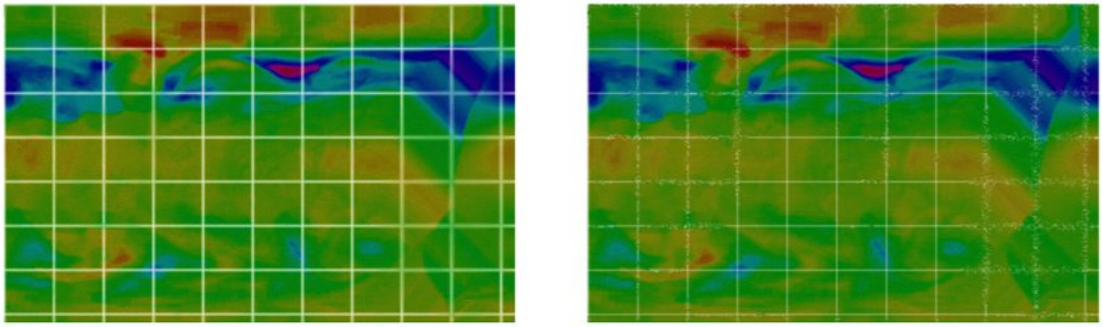


Figure 3.4: Uncertainty visualization via annotation [123]: uncertain data have a softer-edged overlaid grid.

3.5 Evaluation of Visualization

Several papers have been published which cover the subject of evaluation in the information visualization domain [124, 125, 17, 126, 127, 18]. The evaluation in this master's thesis, however, is based on one of the seven guidelines introduced by Lam et al. [17]:

- *Understanding environments and work practices (UWP)*: relies on quantitative data such as interviews from domain experts; the questions are focused on the set of functions and features a visualization tool should have.
- *Evaluating visual data analysis and reasoning (VDAR)*: to evaluate if the visualization tool has the ability to perform the required analysis, the evaluation can be based on quantifiable metrics such as the number of insights or qualitative data based on the user experience, experiments, or case studies.
- *Evaluating communication through visualization (CTV)*: In aspects such as learning and teaching, CTV can evaluate how effective communication is supported by visualization. This is usually performed by controlled experiments or observation in the field.
- *Evaluating collaborative data analysis (CDA)*: Weather a tool supports collaboration or collaborative decision making can be measured using CDA. Only a few papers exist which have performed an evaluation for collaborative information visualization system; however, this can be done by reviewing user feedback or interviews with a domain expert.
- *Evaluating user performance (UP)*: Time accuracy and task quality are the two usual metrics measured in a user performance evaluation. The numerical values are analyzed using statistical methods.
- *Evaluating user experience (UE)*: User's verbal or written feedback is evaluated to understand how people react to the visualization in the short or long term.

- *Evaluating visualization algorithms (VA)*: measures the performance of the algorithm used for visualization.

Our visualization framework mainly aids the user for knowledge discovery and hypothesis generation. For this category of visualization frameworks, the VDAR evaluation method is usually being used [17, 128]. This evaluation method's main goal is to assess the ability of the visualization framework to support visual analysis and reasoning about data. Although some studies might collect numeric metrics, the main aim is to understand how the whole framework supports the analytic process. The VDAR evaluation method can be conducted via case studies, in which the framework is evaluated by answering a set of questions and evaluating if the required tasks are being fulfilled. A detailed evaluation of our visualization framework can be found in Section 6.4.

CHAPTER 4

Visualization Design

In Chapter 2, we reviewed the literature related to existing visualizations of the Hofstede model. The existing methods in the cultural domain mentioned discussed in Section 2.2 and the visualization methods in the domain of multidimensional visualization introduced in Section 3.3 can be combined to identify literature gaps and to create a basis for a visualization framework for knowledge discovery in cultural models. The goal of this chapter is to present a set of requirements for our framework (as resulting from our literature search), and a detailed task analysis designed to fulfill the previously set requirements.

After reviewing the existing cultural visualization methods in Section 2.2, we identified that these methods lack dynamic interaction. All of them are static visualization methods, which do not give the user the capability to use filtering or brushing. The dynamic interactions give the user the freedom to explore and digest the data better, which aids the knowledge discovery process, which is currently missing. Additionally, none of them accounted for cluster visualization, missing data, and uncertainty in their methods, resulting in a loss of information. This all leads to the demand for a visualization framework that covers this gap and satisfies the relevance cycle of the three-cycle research framework introduced in Section 1.3. In order to create such a visualization framework, it is crucial to understand the process of designing and implementing a visualization framework. The framework design is discussed in this chapter and the implementation in the upcoming Chapter 5.

In Chapter 3, we introduced methods such as parallel coordinates and scatterplot matrices as basic methods for the visualization of the data coming from cultural models, which are inherently multidimensional and with a varying number of dimensions. We then discussed the process of Visual Analytics in Section 3.2, which serves as a guideline when we create our first prototype. Finally, we introduced the seven guidelines established by Lam et al. [17] as our evaluation method, which is used to evaluate the prototype in the design cycle. The guideline and evaluation method establishes our knowledge base

and is the input in our rigor cycle as discussed in Section 1.3. This chapter's focus is on exploring what is required to build the first artifact (prototype) of the visualization framework. This is the final input required in the design cycle before creating the first prototype.

Based on the identified gaps between the current visualization methods found in the literature and our goal, we formulate the following design requirements:

1. **R1—Loading Data**: Users need a way to represent the data from the Hofstede models. These data are multidimensional and with a varying number of dimensions.
2. **R2—Handling Missing Data**: The Hofstede model with six dimensions has missing data. The user must be able to deal with the missing data.
3. **R3—Cultural Profiling**: User must be able to choose a clustering method and apply it. They should be able to determine countries with similar cultural profiles, compare them, and find interesting insights and patterns regarding these countries' cultural backgrounds.
4. **R4—Knowledge Discovery**: The user must be able to interact with the visual representations of the Hofstede model data in a flexible manner, to discover additional knowledge about the cultural background of different countries.

For each of these requirements, we discuss the appropriate design for tackling them in the upcoming sections. This is done based on the typology for task analysis by Brehmer and Munzner [129].

4.1 Task Analysis and Abstraction

The term task can be interpreted in different ways. To avoid this issue, we rely on our task description based on a taxonomy introduced by Brehmer and Munzner [129] (refer to Figure 4.1). Tasks can be described at a high-level or in a low-level of detail. For instance, if we would define a task as "a user wants to explore Hofstede's cultural data", it would be too general. On the other side, defining a task as "a user wants to investigate the PDI dimension of Hofstede's model and filter out only three clusters" would be a low-level task.

The high-level task gives the designer of the framework only an overview of the user's actual intent; the low-level task only gives precise information on a specific task without understanding the user's actual motivation. This is the "gap" as described by Brehmer and Munzner [129]. They try to solve this gap by introducing medium-level tasks, which gives the designer an understanding of what specific task a user wants to achieve and the factual background and motivation. This multi-level typology targets to remove this gap between high and low-level tasks by answering three questions: *why* a task is performed, *what* are the input and outputs, and *how* this task is performed.

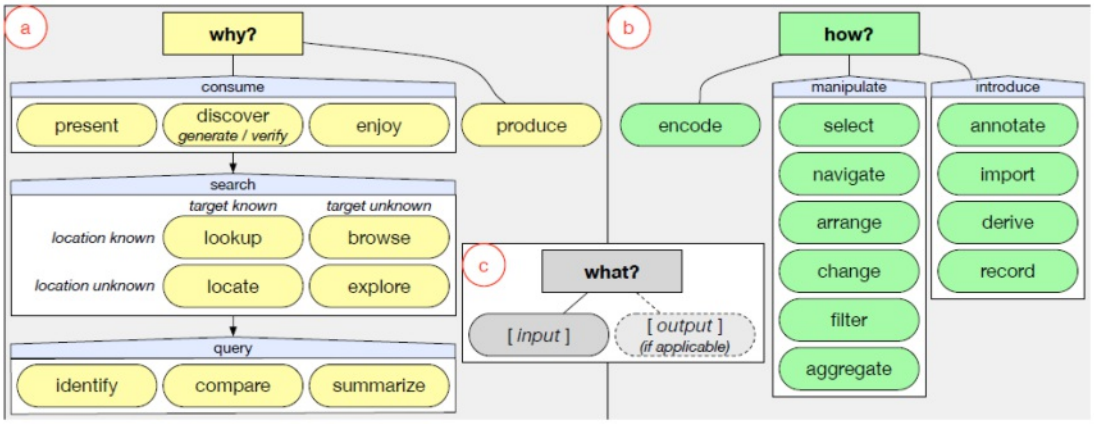


Figure 4.1: Multi-level typology introduced by Brehmer and Munzner [129]. High-level tasks can be drilled down to medium-level tasks by answering *why*, *what*, and *how*.

Looking at Figure 4.1, we see the three terms *why*, *how*, and *what* consecutively shown in a, b and c, respectively. Three sub-sections divide the *why* section:

- **Produce**: Here, the intention is to *produce* a new artifact such as an annotation, recorded visualization, visualization interaction.
- **Consume**: Here, the user uses the visualization to consume information. This can be achieved by *presenting* the information in a storytelling way, by *discovering* hypotheses and confirming them, or by casually interacting with the visualization framework and *enjoying* it.
- **Search**: Here, *lookup* or *locate* are used for searching known targets, whereas *browse* and *explore* are used when a user is searching for a target matching particular (or even unknown) characteristics. For example, if a user is familiar with the geography, they *locate* a country on the map. If they are unfamiliar they need to *explore* the map.
- **Query**: Once an element or a set of elements is found after conducting a search, the user can *compare*, *identify* or *summarize* these elements.

The *how* part of the typology shown in Figure 4.1 is classified into three sections:

- **Encode**: *How* the data is encoded into a visualization.
- **Manipulate**: Users can *manipulate* existing elements by interacting with the elements in the visualization. The term *select* refers to directly clicking or lassoing elements. *Navigate* includes methods such as zooming or rotating, which is changing

the user's viewpoint. *Arrange* means organizing and arranging the elements spatially, for instance, to re-order axes of parallel coordinates. *Change* refers to changes in the visual encoding, such as altering the size and transparency of points. *Filter* is when we add exclusion and inclusion methods in the visualization. *Aggregate* is to change the granularity of visualization. For example, a user aggregates some daily values into monthly values by changing the granularity of a continuous time scale in a time graph.

- **Introduce**: The term *introduce* is creating new elements. *Annotation* is when an additional text or graphical annotation is added into a visualization element. *Import* keyword is used when a new element is added into the visualization by importing it. *Derive* is used when new elements are being created using an existing element. A user can *derive* new elements using some algorithm in the visualization. *Record* keyword is used to save and record a visualization state, mostly via a screenshot.

The typology finally defines the term *what* in a flexible manner with a "bring your own *what*" mentality. The only requirement is that it should define the *input* and *output* explicitly.

As a basis for our cultural model visualization prototype, we employ the typology discussed above. Below, we formulate the four tasks that our prototype can support following the typology conventions.

Task 1: Visualizing and Comparing Cultural Dimensions

Task 1 addresses the visualization of the multidimensional data of the cultural model of Hofstede and the comparison of different countries. This relates directly to requirement R1 (Loading Data), and indirectly to R2 (Handling Missing Data).

- *Why:* To cover this task for the user, it should be possible to *present* values of different dimensions for the cultural model of Hofstede. The user needs to *look up* and find countries and be able to *compare* different dimensions to each other.
- *How:* First of all, the data are *encoded* in visual representations (radar chart, world map). It should be also possible to *select* specific countries and *change* them if the user requires to.
- *What:* As input, we consider the data from the model and as output, a set of countries to be compared.

Task 2: Discovering New Knowledge
The user should be able to confirm or reject hypotheses, such as "the higher the PDI is in the model, the lower is the GDP of the country", which fulfills R4 (Knowledge Discovery).

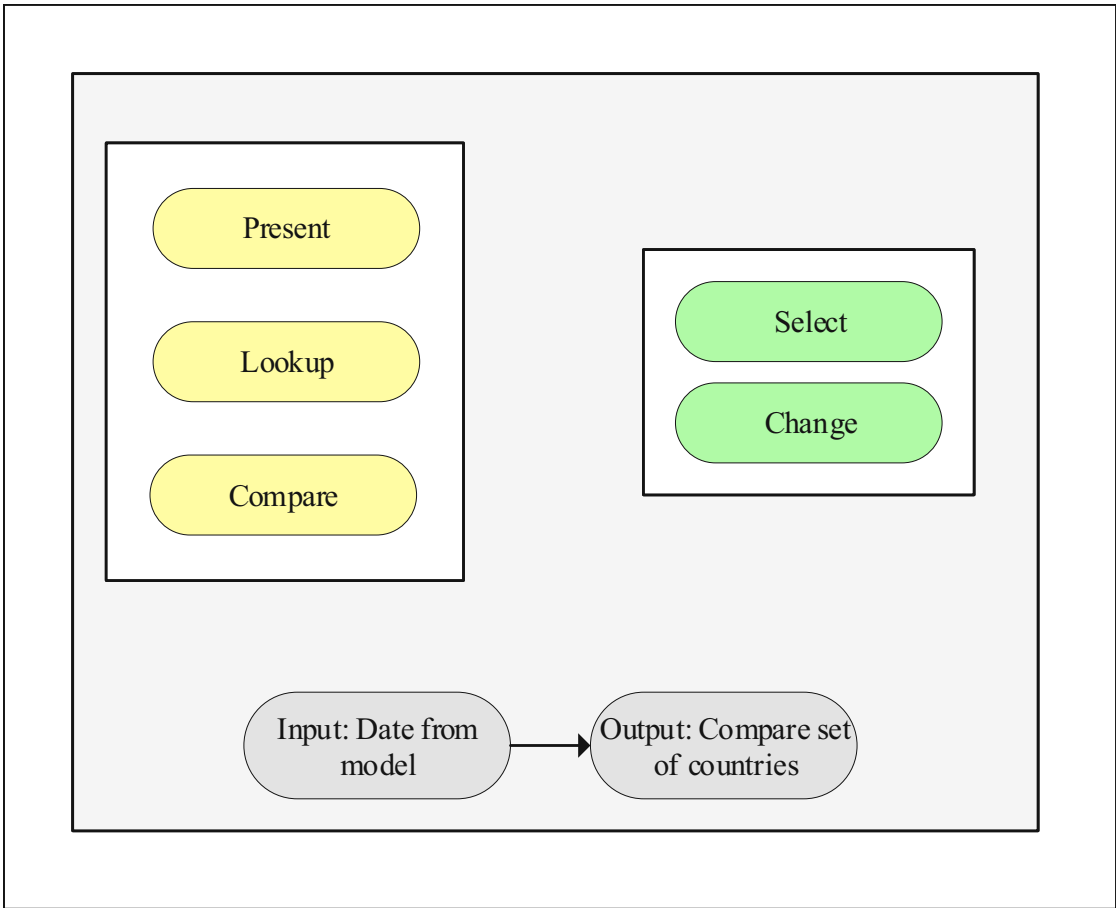


Figure 4.2: Typological diagram for Task 1: Visualizing and comparing cultural dimensions of different countries.

- *Why:* This task goes along with *discovering* knowledge using our visualization framework. The goal of the user is to *explore* the countries and to *identify* a correlation between the dimensions.
- *How:* The user should be able to *select* a specific set of countries or to *filter* countries based on criteria, while it should be also possible to *re-arrange* the data.
- *What:* The input is the standard six dimensional Hofstede model, and additional socio-economic dimensions, as added by the user. The output is the discovered knowledge to confirm or reject a pre-determined hypothesis.

Task 3: Cultural Profiling of the World
As seen in Figure 2.6 and 2.1, visualization of clusters exists in the current literature; the clustering is being used to identify similar characters or cultural clusters in the data. Task 3 is designed to fulfill R3 (Cultural Profiling).

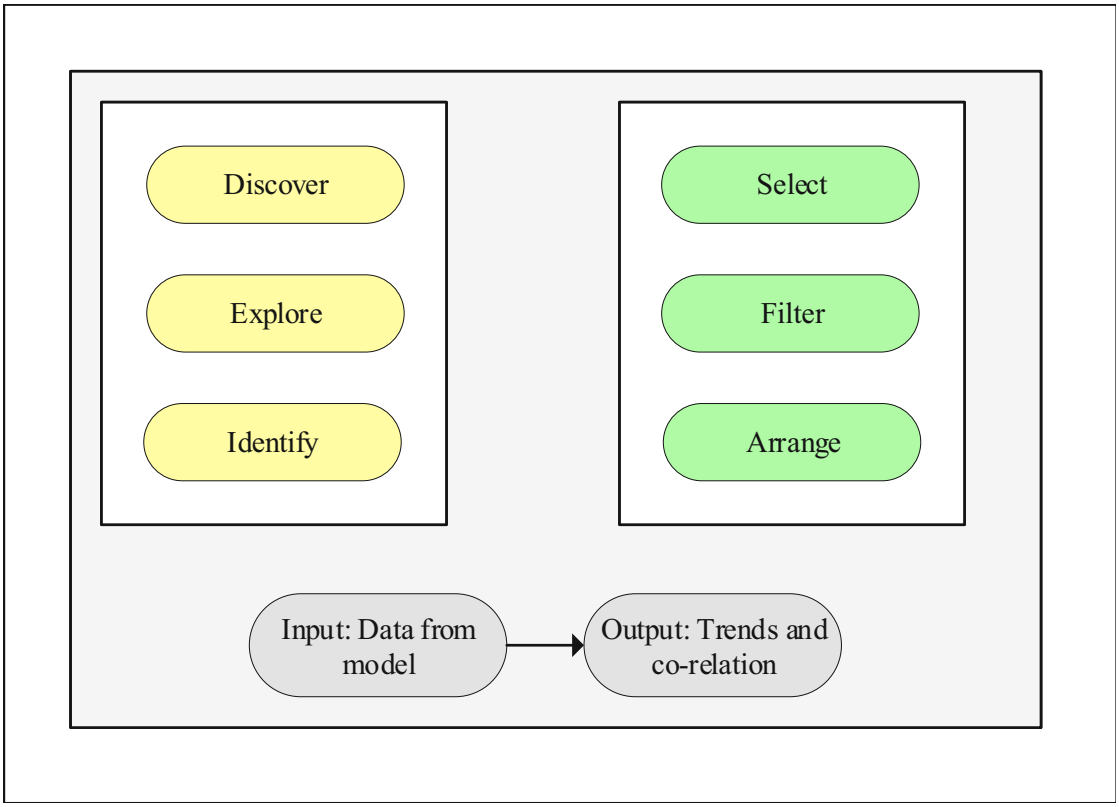


Figure 4.3: Typological diagram for Task 2: Discovering knowledge in the cultural model.

- *Why:* To *discover* which countries belong to similar cultural backgrounds, or what countries can be grouped with each other.
- *How:* By *looking up* the country and *identifying* if there is a special characterization in the cluster to which they belong. Users should be able to *select* different clusters and *change* the clustering method.
- *What:* As input, we consider the data from the model and a clustering method which results in a visualization that we use, in order to understand a specific pattern within the clusters and to see if the countries in the cluster have similar characterization.

Task 4: Visualizing Uncertainty (Data Missingness)
As described in Section 3.4, the uncertainty of a visualization can be described as the error or difference to the truth. In our framework, we impute missing data. These imputation methods have an impact on the values, and the user needs to be able to distinguish between imputed and non-imputed values. This task relates directly to R2 (Handling Missing Data), and indirectly to R4 (Knowledge Discovery).

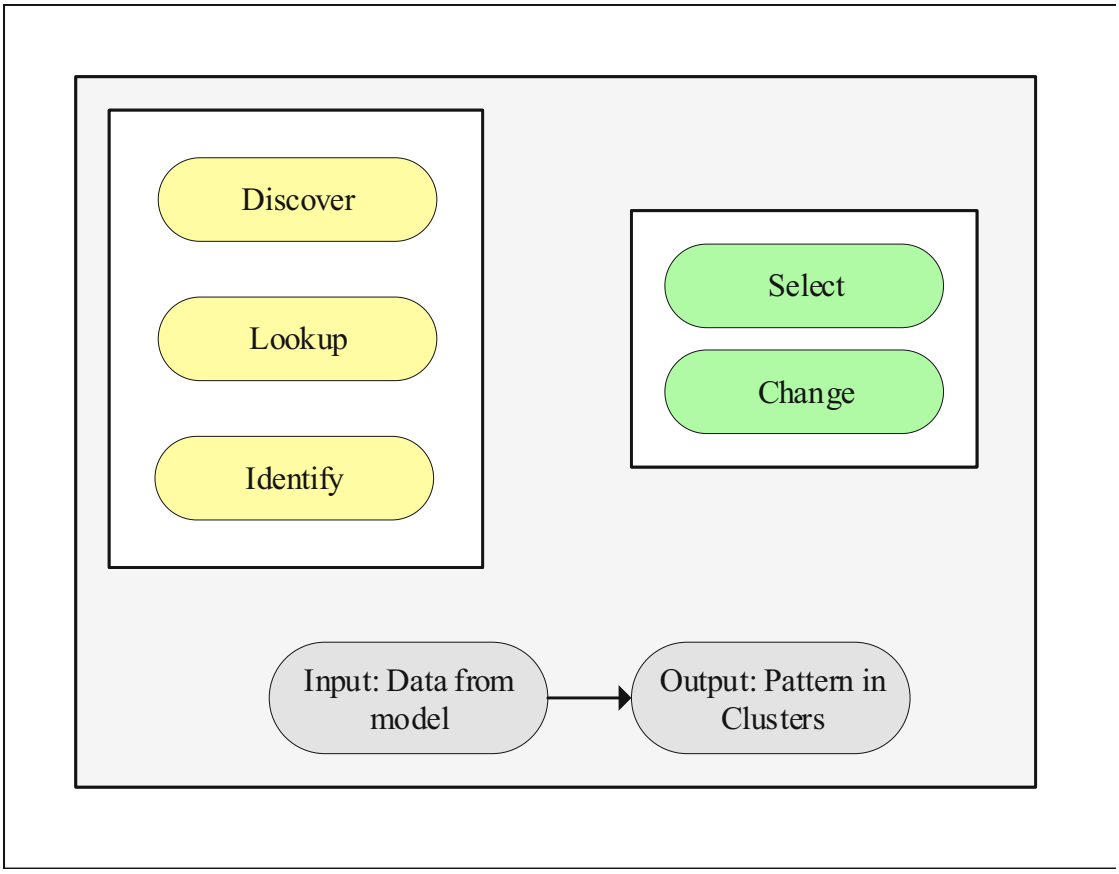


Figure 4.4: Typological diagram for Task 3: Cultural profiling of the world.

- *Why:* To *produce* visualizations, which show the effect of uncertainty (missingness) in the cultural models. Additionally, to *compare* the impact of different imputation method and their uncertainty.
- *How:* By *changing* parameters of imputation and by *producing* different visualization.
- *What:* As input, we consider the data from the model and a clustering method, which results in a visualization. As output, a heatmap is represented which annotates the missing values.

4.2 Designing the First Prototype

We first identified gaps in the current visualization methods of the cultural models, and based on these gaps we identified requirements and described middle-level tasks following Brehmer's multi-layer task typology. It is now possible to design the first prototype before

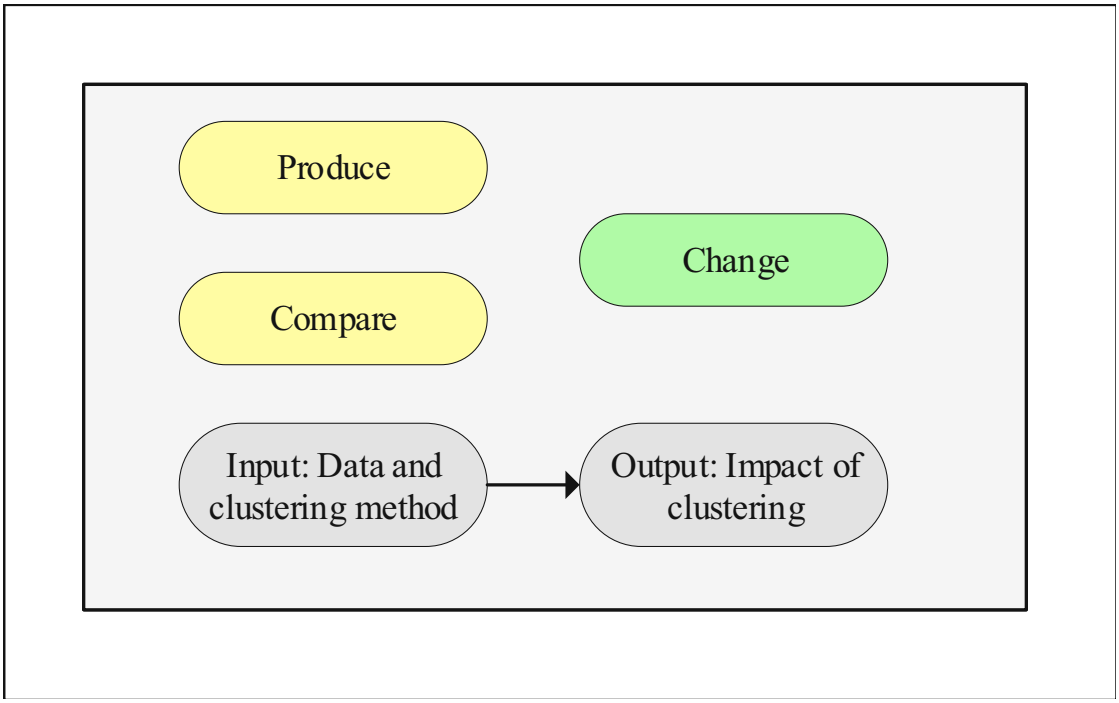


Figure 4.5: Typological diagram for Task 4: Visualizing the uncertainty (missingness) of cultural models.

implementation. As illustrated in Figure 4.6, the initial UML activity diagram shows the flow of how the visualization framework works. Starting from the top, the user loads the data into the framework, which satisfies requirement R1. The user then can choose imputation and clustering methods, modify their parameters, and apply them to the data, which is required to fulfill requirements 2 and 3. Finally, to fulfill requirement R4, the user can view a dashboard that includes all the visualization methods. Which exact visualization method is shown is discussed in the next Section refch4:choosevismethod.

4.2.1 Choosing Visualization Methods

After our detailed task analysis, we can make an informed choice on the most appropriate visualization methods for fulfilling our previously discussed requirements and tasks. We, hereby, discuss the general motivation and rationale behind our choices. The specific visualization encodings used in each one of the selected visual representations are discussed in the upcoming section.

As discussed in Section 2.2 and illustrated in Figure 2.4b, to compare two or more countries with each other, the current literature used bar charts and radar charts. Taking this as an inspiration, we choose radar charts as the visualization method to fulfill Task 1. We adapt it by giving the user the possibility to interact and dynamically change the radar chart. Radar charts are a suitable visualization method for *comparing* one or more

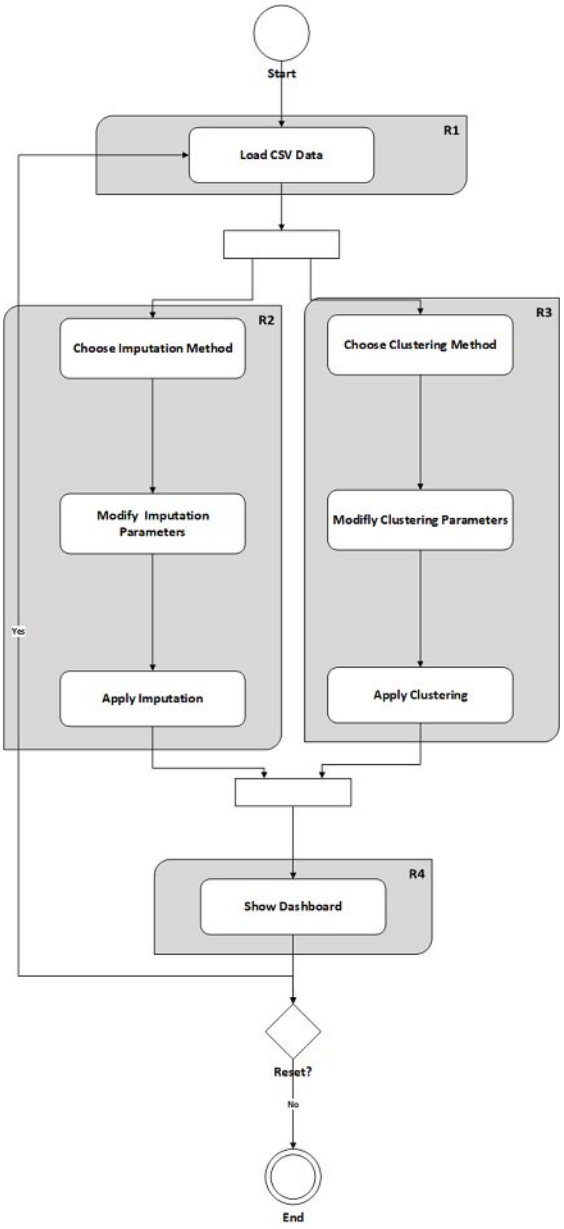


Figure 4.6: Activity diagram showing the flow of the visualization framework and how it fulfills requirements R1–4.

countries. A bar chart such as the one in Figure 2.2 is too crowded, if more than five countries are selected. Providing the user the ability to *select* and *change* the countries which are intended to be *compared* can be done via checkboxes. All the countries are listed, and the user can decide which to pick to *compare*. The visualization projected in a radar chart adapts after the user's selection and changes. This goes along with what is required in Task 1. *Presenting* a visualization, where different countries can be *looked up* and *compared* to each other by *selecting* different countries.

Task 2 requires a visualization method that has the possibility to *filter* and *re-arrange* dimensions, so that it is possible to detect co-relations. Based on Andrienko and Andrienko [130] parallel coordinates give the user the ability to compare characteristics of individual objects, understand the correlation between objects, change the order of axes and do a pairwise comparison. Parallel coordinates must have the possibility to *re-arrange* the axes and *filter* the countries via brushing. This ensures *select*, *filter*, and *arrange* is fulfilled and complies with what is required in Task 2.

Task 3 requires a visualization method that shows all the countries in their geographical position and their cluster membership. This is possible to achieve using a world map similar to how it was used in the current literature shown in Figure 2.1 and 2.7. The world map gives the user an understanding of the geographical location of the clusters. Additional visualizations, such as a clustermap [131], aid the user to understand which countries belong to the same cluster. The clustermap plots a matrix dataset as a hierarchically-clustered heatmap with an additional dendrogram [132] and illustrates which group of data belongs to one cluster in a tree diagram. Clustermaps are used to visualize and explore the impact of clustering on the data [133, 134, 135]. Visualizations such as elbow plots [136, 137] and dendrograms can help a user to select a suitable clustering method (refer to Section 6.3.1 for a detailed explanation). Using a combination of the world map and the clustermap, users can *discover* new knowledge and *identify* patterns by *looking up* countries belonging to the same cluster. For example, they can *identify* that all values of a particular dimension, belonging to a specific geographical cluster, have a high or low value.

As described in Section 3.4, the visualization of uncertainty in Task 4 can be achieved with the use of different methods. The uncertainty, which we thrive on visualizing, is the bias introduced in the data by our imputation methods. We visualize this uncertainty by using annotations for the imputed values on the clustermap. The *produced* visualization, gives the user the ability to *compare* the values of imputed data on a heatmap and view the pattern of branching of the clusters. *changing* the input parameters of imputation the imputation or clustering method affects the visualization.

A summary of the selected visualization methods and their purposes has been summarized in Table 4.1.

Visualization Method	Fulfills	Explanation
Radar Chart	Task 1	*Presenting* a visualization where users can *look up* countries and *change* the visualization by *selecting* different values. The main goal is to *compare* countries with each other.
Parallel Coordinates	Task 2	*Discovering*, *exploring* knowledge and confirming hypotheses by *rearranging* axis and *filtering* data to *select* different countries with different specifications. This also enables the user to *identify* corelation within the data.
World Map	Task 3	*Discover* and *identify* geographically related trends by *looking up* different countries or clusters on the world map representation which can be *changed* by *selecting* different countries or clusters.
Clustermap	Task 3, Task 4	*Producing* a visualization that can be used to *compare* the impact of *changing* different hierarchical clustering parameters. Also used to *discover* and *identify* the uncertainty of different imputation methods.

Table 4.1: selected visualization methods to fulfill the Tasks 1–4, as resulting from the use of Brehmer's multi-layer task typology.

4.2.2 Description of Visualization Methods and Their Encoding

Under this section, we discuss in detail each implemented visualization method chosen to fulfill Tasks 1–4. The reasoning behind each choice is shown in Table 4.1 and has already been discussed in Section 4.2.1. A detailed explanation of the implementation for each visualization method is discussed in Chapter 5.

4.2.2.1 Radar Chart

As defined in Section 3.3, radar charts can be used to represent multidimensional data in two dimensions. We use a radar chart to represent the dimensions of Hofstede model (see Figure 4.7). Additionally, the radar chart represents the lower and upper boundary

of the 50% central region which is calculated with the following math Equation (4.1), to generate a functional boxplot [138]:

$$C_{0.5} = \left\{ (t, y(t)) : \min_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\ldots,\lceil n/2 \rceil} y_r(t) \right\} \tag{4.1}$$

Figure 4.7a shows a visualization when only two countries are selected. In this visualization, we can see a comparison between Australia and Austria. Hovering over the vertices shows information about the dimension value and country. The red line in Figure 4.7b, shows the central value of the functional boxplot. The orange line shows the upper limit of the functional boxplot, which is 1.5 times the 50% central value. On each axis, the values can be found written behind the polygons.

4.2.2.2 Parallel Coordinates

As defined in Section 3.3, parallel coordinates [109] are also a widely used representation for multidimensional data. In this visualization method, we show each dimension of Hofstede's model on an axis. The user can filter the values on each axis by brushing and selecting a specific range. The filtered countries are additionally shown in a table located below the parallel coordinates. An example can be seen in Figure 4.8, where the values are filtered between scores ranging from 40 to 100. The coloring of the coordinates is based on the *Z-score* [139]. The Z-score is the number of standard deviations in which the value of the observation is below or above the mean and is calculated using Equation (4.2) where μ is the mean, and σ is the standard deviation of the population.

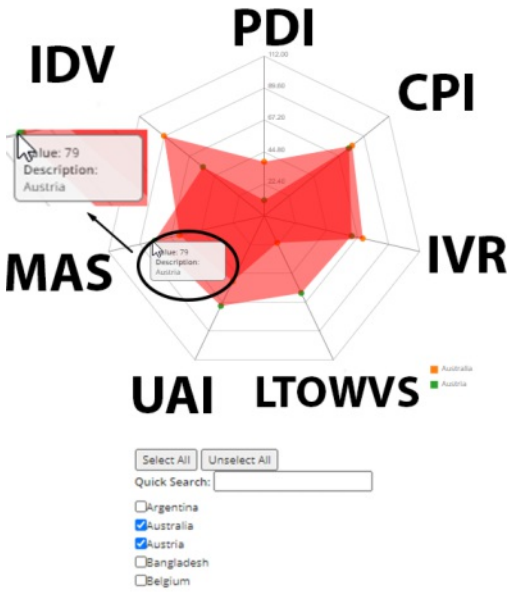
$$z = \frac{X - \mu}{\sigma} \tag{4.2}$$

Z-scores above the mean is colored blue, whereas Z-scores below the mean are red (see Figure 4.8). The coloring is always based on a single dimension that is dynamically changeable by selecting each axis's title. If the data is clustered, the coordinates' coloring is grouped based on the cluster, meaning each unique cluster has a unique coloring as seen in Figure 4.9.

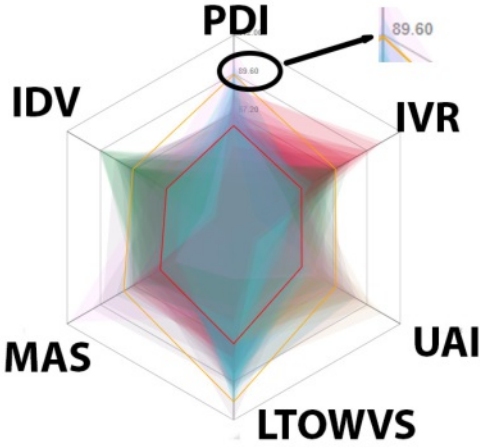
The values on each axis can also be inverted by double-clicking on the same axis, and this makes the comparison of negatively correlated data more manageable, which is discussed in detail in Chapter 6.

4.2.2.3 World Map

The world map represents a two-dimensional map where the values of a selected dimension are shown for each country. A similar visualization has been used by Minkov [27] which we illustrated in Figure 2.1. However, in our world map, the user can dynamically change the dimensions in the world map. The coloring of the world map is the same as the



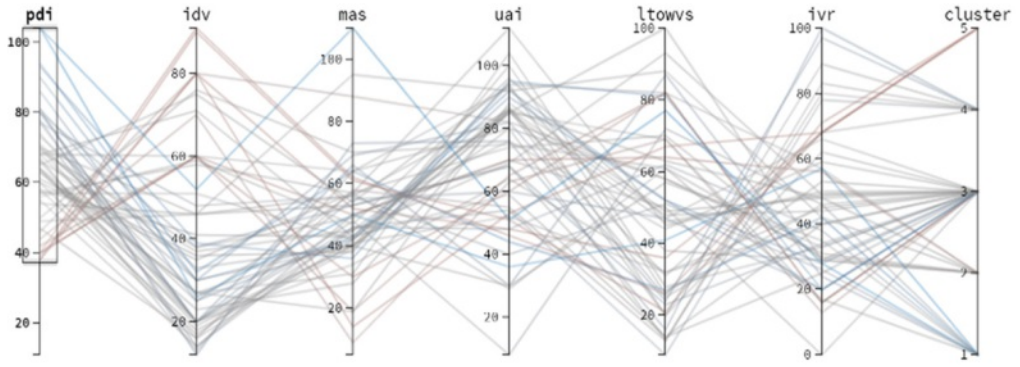
(a) Comparing Austria and Australia via our implemented radar chart.



(b) Upper limit of functional boxplot.

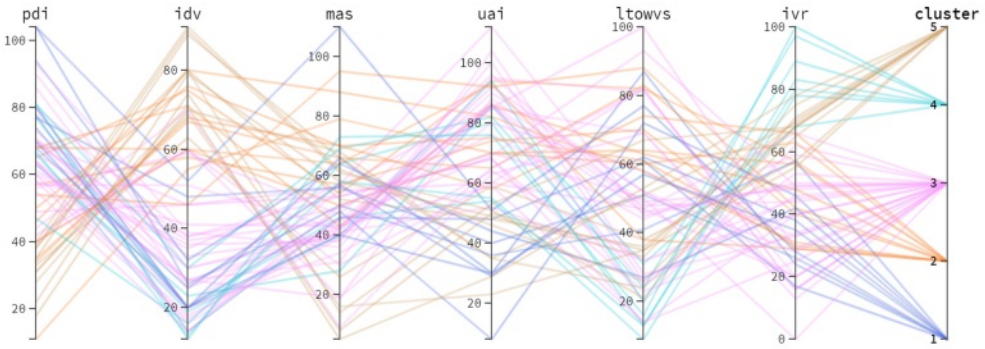
Figure 4.7: One-to-one and many-to-many comparisons via radar chart.

parallel coordinates, based on Z-score where negative Z-scores are colored red, positive Z-scores are colored blue, and countries with no entries are colored as grey. A legend on the right side of the world map shows the coloring scale. Optionally, if the data is clustered, the coloring can be grouped based on clustering where countries belonging to the same cluster have the same color. The colors of the clusters in the world map are the same colors chosen for the clusters in the parallel coordinates shown in Figure 4.9. An example of the world map, where the Hofstede model's PDI dimension is selected, can be seen in Figure 4.10. The tool-tip shows Russia's values for each dimension of the



country	pdi	idv	mas	uai	ltowvs	ivr
Africa East	64	27	41	52	32	40
Africa West	77	20	46	54	9	78
Arab countries	80	38	53	68	23	34
Argentina	49	46	56	86	20	62

Figure 4.8: Parallel coordinates plot showing 6 dimensions of the Hofstede model. Coloring is based on the PDI dimension and the values of PDI are filtered between 40 and 100.



country	pdi	idv	mas	uai	ltowvs	ivr
Africa East	64	27	41	52	32	40
Africa West	77	20	46	54	9	78

Figure 4.9: Parallel coordinates plot showing 6 dimensions of the Hofstede model. Coloring is based on the detected clusters.

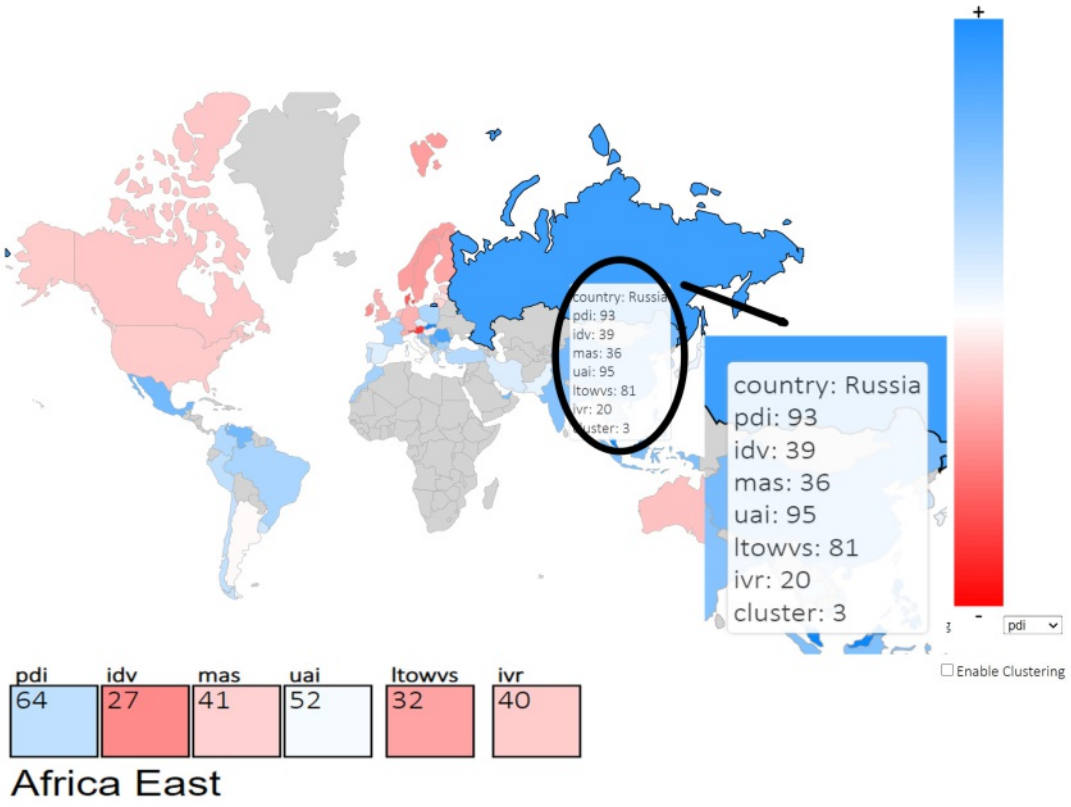


Figure 4.10: World map showing the values of PDI in Hofstede model. Coloring is based on the PDI dimension and their Z-score values. Hovering the mouse over a country shows full description of each dimension and the values. A small heatmap in the bottom of the world map shows the values of all the other dimensions.

Hofstede model.

Heatmaps located below the world map represent the actual values of all the dimensions for each country, the coloring of this heatmap is also based on the Z-score of each dimension. Alternatively, the values can be observed using a tool-tip, i.e., by hoovering the mouse over any country.

Figure 4.11 shows an example where the coloring of the world map is based on clusters. Each cluster has its unique coloring on the map. Detailed explanation on how the world map is used for knowledge discovery is discussed in Chapter 6.

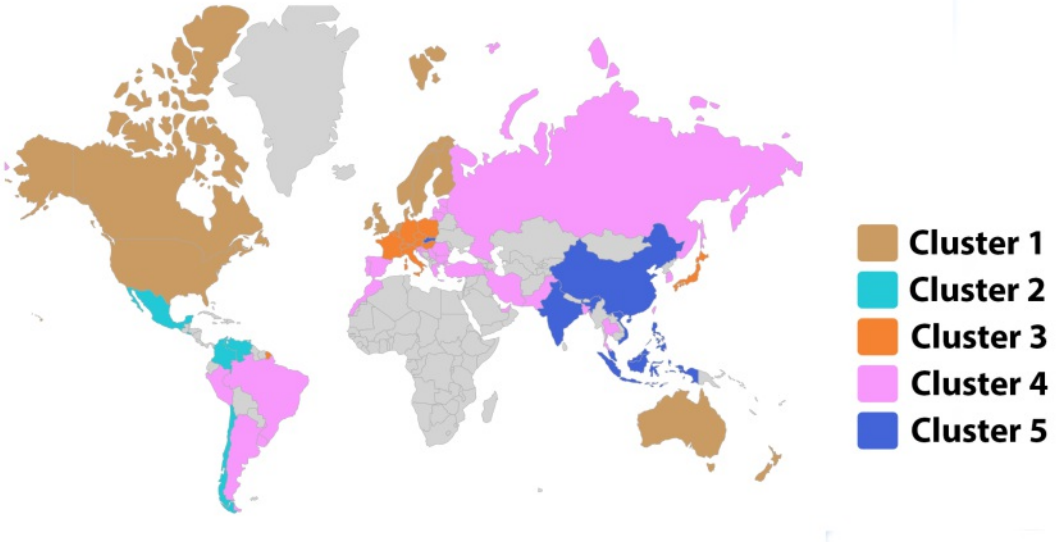


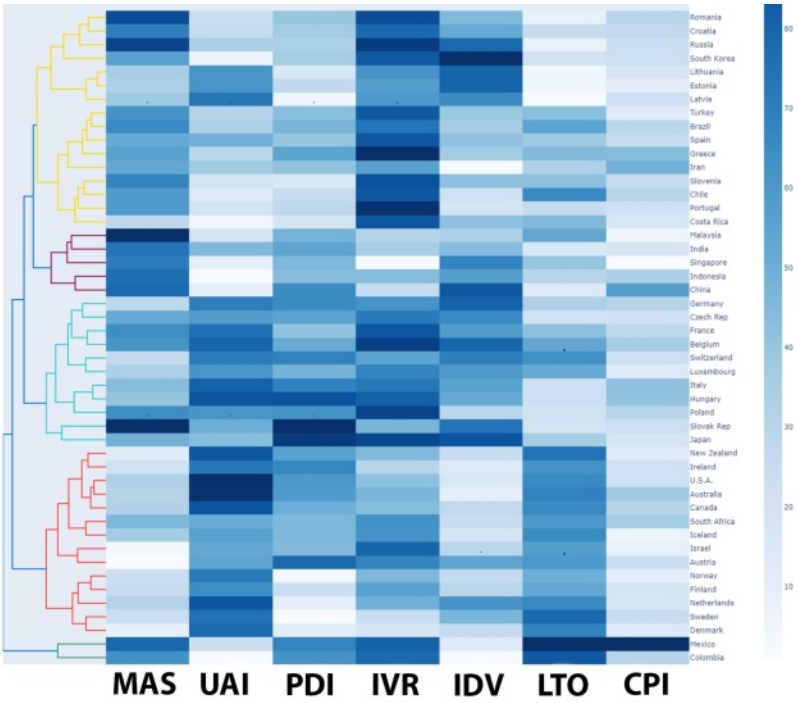
Figure 4.11: A world map representation of countries that have values in the Hofstede model. Coloring is based on the detected clusters.

4.2.2.4 Clustermap

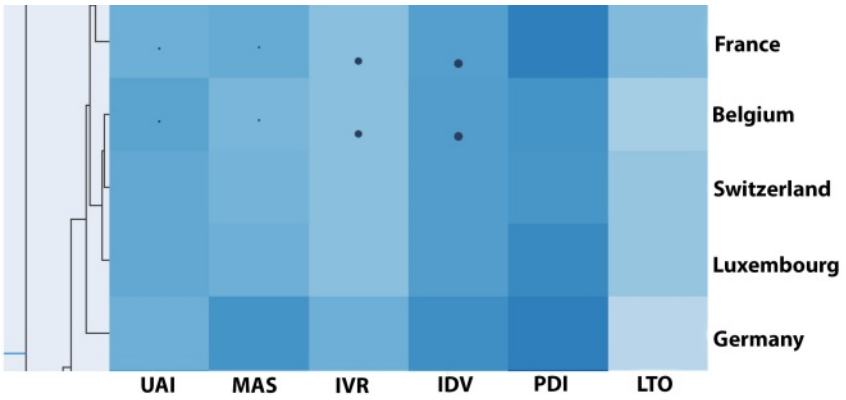
A clustermap is a combination of heatmap and dendrogram, which is used to identify the hierarchical structure (like a taxonomy) [140].

Figure 4.12a shows a full view of the clustermap, where in the middle, there is a heatmap representing each data point. The coloring is based on the scale shown on the right side, where 0 is set to white. The color gets darker and closer to blue when the value of the observation increases. There are dendrograms attached on the left and the top side of the heatmap showing the pattern of the hierarchical clustering. On the heatmap, imputed data are annotated with a *dot* symbol. The size of this symbol depends on the uncertainty of the imputation, the bigger the uncertainty is, the bigger is the size of the dot.

A zoomed-in example of the clustermap is shown in Figure 4.12b. In this clustermap, six dimensions from the Hofstede model containing data from Bosnia, Austria, Chad, Czech Republic, and Kyrgyzstan is represented. PDI and IDV dimensions had no missing values. Thus, no imputation is applied. However, for Bosnia and Austria, there are missing values in the IDV, IVR, MAS, and UAI dimensions (marked with the *dot* symbol). Comparing the size of the *dots*, it is evident that the imputation method's uncertainty was highest in IDV and lowest in UAI dimension.



(a) Clustermap (full view): dendrogram and annotated heatmap with uncertainty. Dots are representing imputed data, the size is depending on the uncertainty: the higher uncertainty—the bigger the dots. A dendrogram on the left shows the pattern of agglomerative clustering.



(b) Clustermap (zoomed-in view): dendrogram and annotated heatmap with uncertainty. Dots are representing imputed data, the size is depending on the uncertainty: the higher uncertainty—the bigger the dots. Bosnia and Austria have both imputed values in IDV, IVR, MAS and UAI dimensions. IDV's imputation has the highest uncertainty and UAI has the loweset uncertainty.

Figure 4.12: Clustermap visualization method: full and zoomed-in view. A dendrogram on the side shows the pattern of clustering in the zoomed-out view. Dots represent imputed data and the uncertainty of the imputation.

4.2.3 Choosing Imputation and Clustering Methods

As discussed in Section 2.4, in the current literature, a mixture of simple and iterative imputation methods have been applied to handle missing data in the domain of cultural science and in some cases, records with missing data were just deleted. For this purpose, the system should offer the users the possibility to do simple imputations such as mean, median, and most frequent and additionally apply iterative imputations, such as MICE and KNN. In Section 2.3.5, we have analyzed papers that applied clustering into the cultural science domain. All of them used a hierarchical clustering method. Thus, the system needs to support hierarchical clustering due to prior familiarity with the method.

4.2.4 User Interface Design

Knowing the user requirements and understanding the specific tasks the system needs to fulfill makes it possible to design a user interface. The flow in which the system shall behave can be understood by reviewing the UML activity diagram illustrated in Figure 4.6. Additionally, it is known which clustering and imputation methods need to be supported.

Statistical studies have shown that the use of user interface mock-ups eases comprehension and reduces the effort of development time [141]. For this purpose, we used mock-ups to get quick feedback and ease the supervision process. The first prototype was created using an online application named Moqups [142], which an online tool for creating UI prototypes. Moqups has a user-friendly UI with drag and drops features, which enables users to create interactive and clickable UIs quickly; for this reason, the tool got popular among researchers who intend to design applications [143, 144, 145, 146].

4.2.4.1 Mock 1: File Upload

To fulfill requirement 1, the file upload page was designed. The user first selects an imputation and clustering method, as seen in Section B and C of Figure 4.13. Then, after selecting a CSV file, the user uploads the file. On the left side of the screen, Section A shows a placeholder for a menu, which serves for navigation purposes.

4.2.4.2 Mock 2: Dashboard

After the user successfully loads a file and selects an imputation and clustering method. The result is shown in a dashboard. Section A shown in Figure 4.14 changes, and all the navigation options to other pages are available for the user. Section B, C, and D are placeholders for the visualization methods which fulfill Task 1, 2, and 3 as shown in Table 2.1. The dendrogram and clustermap diagrams are conditionally populated visualizations, and these two diagrams are only visible to the user if the hierarchical clustering method is used.

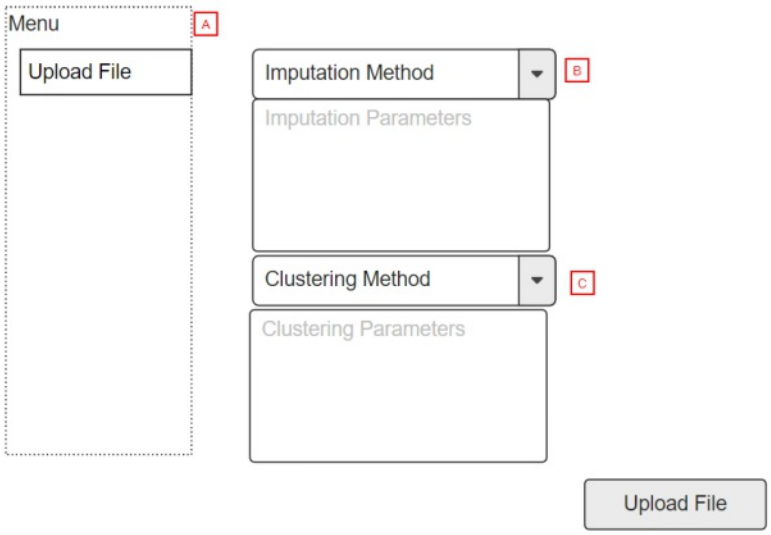


Figure 4.13: Mock 1: Upload screen. Here, the user can select the imputation and clustering method and upload a CSV file.

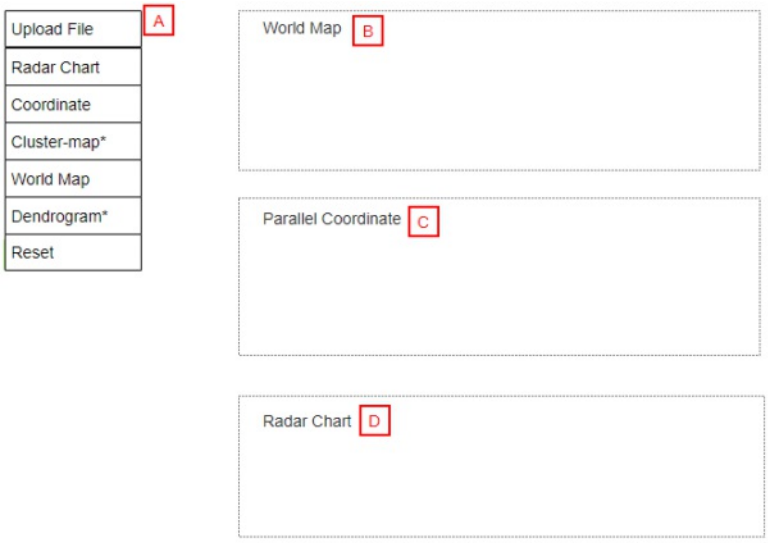


Figure 4.14: Mock 2: Dashboard. This is the screen where the world map, parallel coordinates, and radar chart visualization gets populated after the user uploads a file.

4.3 Designing the Second Prototype

After presenting the mock-ups and reviewing the process of Visual Analytics, as described in Section 3.2, we decided to change the application flow, since the loop between perception and knowledge discovery is missing in the first prototype. The user cannot see and comprehend the impact of imputation and clustering, although choosing the correct method has an enormous impact on every other visualization. Thus, a user should be able to preview the impact of the chosen imputation and clustering method before applying it. This change requires an adaption to the activity diagram illustrated in Figure 4.6 and the file upload mock-up shown in Figure 4.13.

The newly adapted activity diagram illustrated is designed to be closer to the Visual Analytics process. The user can first preview the pattern of missingness in the data. This gives the user a general overview of what imputation method would be more appropriate. For example, choosing the deletion method for a dataset with a majority of its data missing in a single dimension is not the best solution.

Once the imputation method is chosen, the user is able to select a clustering method, with the difference that now the user is able to preview the impact of the chosen clustering method via dendrogram and elbow plot. It is possible to modify the input parameters or change the clustering method and preview these plots again; this goes in accordance with the Visual Analytics feedback loop between perception and knowledge shown in Figure 3.2.

4.3.1 Additional Visualization Methods Added for the Second Prototype

After revising the conceptual design of the first prototype in Figure 4.15, it is required to add new visualization methods in order to be able to preview the impact of clustering and imputation. A dendrogram and an elbow plot are two additions to the visualization methods for the purpose of previewing the impact of the hierarchical and k-means clustering method. A heatmap is added as well to present the pattern of missing data in the dataset. These additional visualization methods are described below.

4.3.1.1 Dendrogram

As defined in Section 2.3.2, a dendrogram shows the clusters' arrangement in the hierarchical clustering method. The potential clusters are indicated with the different colors. The representation is in the form of a tree with branches, where each leaf of a branch contains an observation in the dataset. The number of clusters depends on the user's input parameters, which sets the height of cutting the tree. Figure 4.16 shows that the height is set to 105, which is drawn as a straight horizontal line. The number of clusters is equal to the number of times the horizontal line crosses with the dendrogram, which are five in this example.

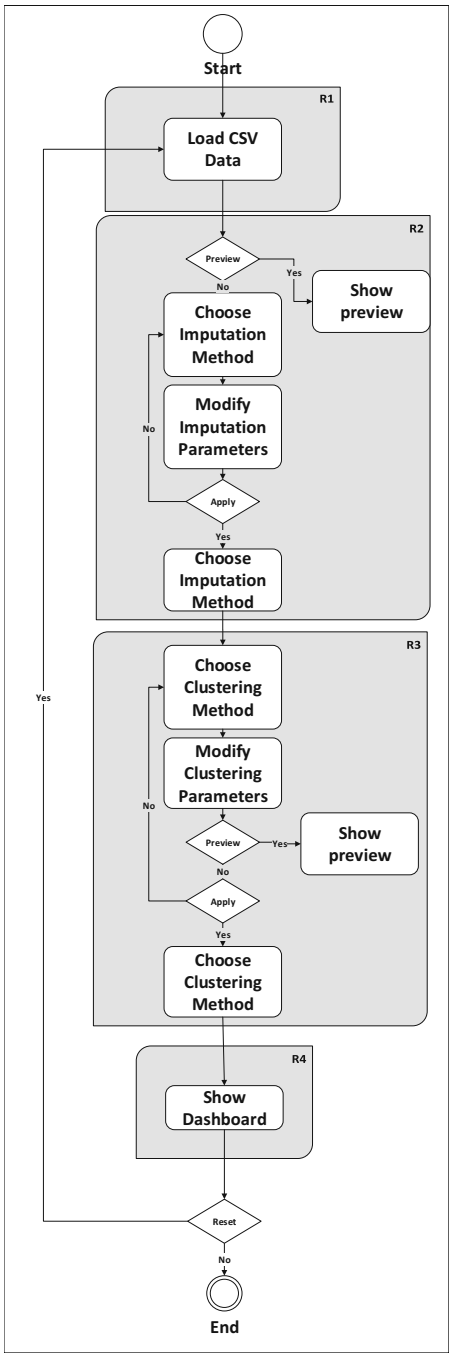


Figure 4.15: Adapted activity diagram designed in a way that is closer to the Visual Analytics process. The user has the option to preview diagrams before applying imputation and clustering.

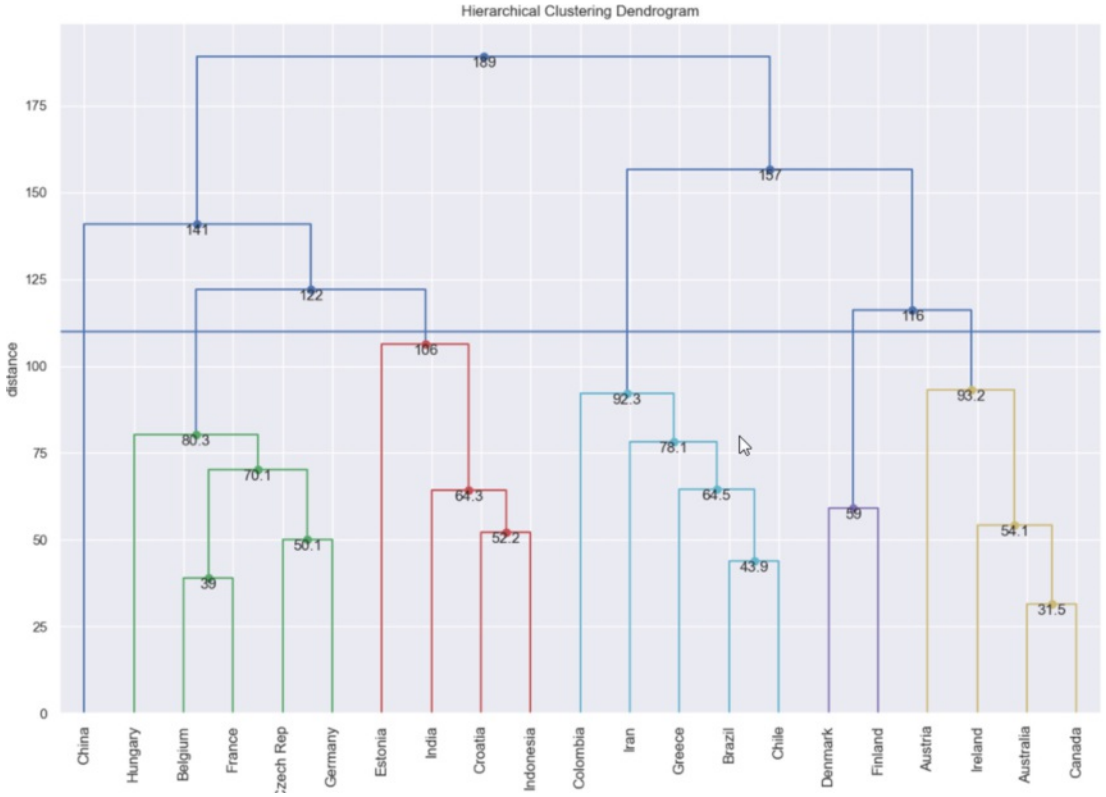


Figure 4.16: Dendrogram visualizing clusters in a dataset, where the number of clusters are cut to five by setting the cutting line's height to 105.

4.3.1.2 Elbow Plot

The elbow plot [136, 137] is used to preview the impact of a different number of clusters in k-means clustering method. The elbow plot shows the sum of squared error (SSE) on the y axis against the number of clusters in the x axis. The SSE can be calculated using Equation (4.3), where X_i is a single observation and $\bar{X}$ is the mean.

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{4.3}$$

The optimal number of clusters can be selected by determining at which point the elbow plot has the quickest increase (i.e., identification of the "elbow") [147]. The goal is to have a small SSE, meaning the error between the data point and center of the cluster is minimal. Logically, the SSE is 0 when the number of clusters equal to the number of data points (each entry is its own cluster). However, our aim is to choose the smallest possible value of k having a low SSE value. Figure 4.17 shows an example of the elbow plot. In this example, the "elbow" can be seen at the value of 3, meaning that the optimal k is set to three.

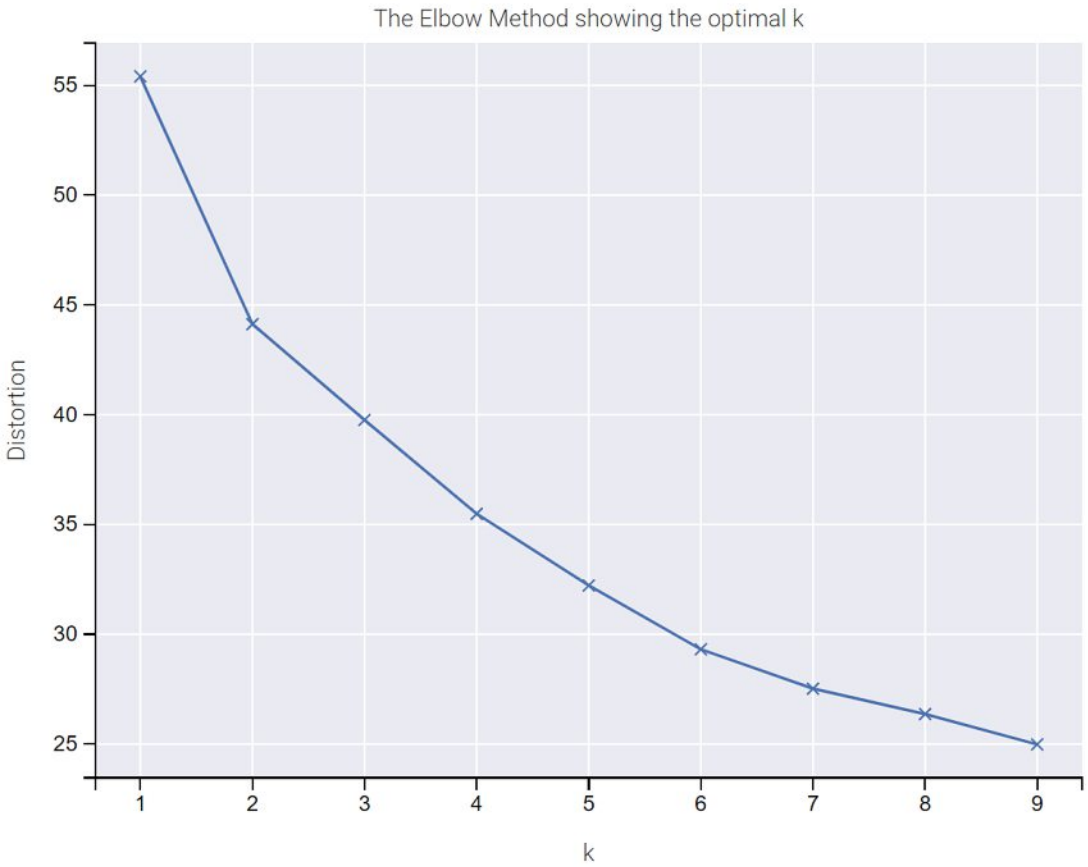


Figure 4.17: Elbow plot for finding the optimal number of clusters in k-means clustering algorithm. The "elbow" is seen when the k value is set to three.

4.3.2 Missingness Heatmap

A heatmap can be used to visualize the patterns of missingness in the data. This is to fulfill the "preview" feature for the imputation step. Before choosing an imputation, the user can preview the number of missing entries in every dimension. We use an abstracted tabular view, where each row is a country and each column is a cultural dimension. Each white line in Figure 4.18 represents a missing entry, and each black line indicates a complete entry. The vertical line chart on the right of the visualization shows how much missing data is in each row. This can be seen as an abstracted histogram, where the longer the line towards the left, the more data is missing in that particular row.

4.3.3 Pearson's Correlation Matrix

To aid users in determining correlations among dimensions of the data, we use a heatmap to encode the value of the Pearson's Correlation Coefficient (PCC, also referred to with the Greek letter ρ) [148]. The formula to calculate the PCC between two random variables X

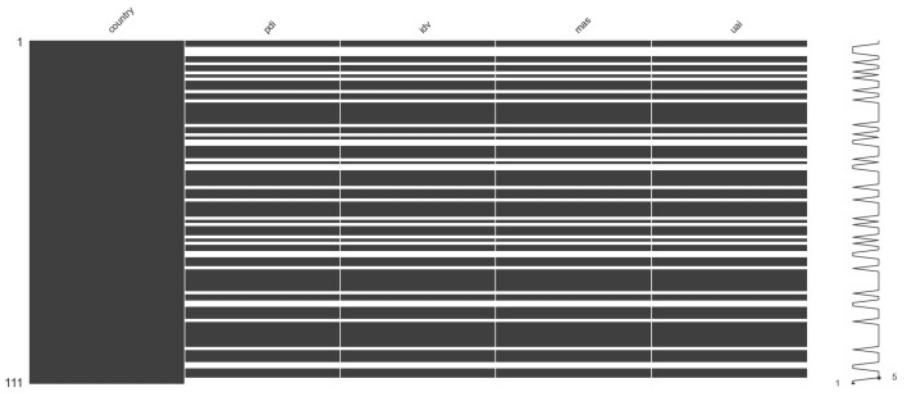


Figure 4.18: Heatmap showing missing data. White lines indicate that the data in this row are missing.The histogram in the right indicates the quantity of data missingness.

and Y can be found in Equation (4.4). Cov is the covariance, σ is the standard deviation. The PCC has a value ranging between -1 and 1. The value of 1 represents a perfect positive relationship, -1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables.

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{4.4}$$

Figure 4.19 shows an example of how the correlation matrix would look like. The heatmap is annotated with the result of ρ. It is based on the Z-score coloring, where negative values are red and positive values are blue.

The "*" symbol beside the values (as also indicated on the top left) shows the significance level of the correlation where "*" denotes 0.05, "**" is 0.01, and "***" means 0.001, similar to common practice of denoting the significance level.

4.3.4 Revised User Interface Design

In the second prototype, the focus is on enabling the user to preview changes. For this purpose, the design shown in Figure 4.13 is adapted to Figure 4.20. In the revised UI, the user is able to select an imputation method in Section A, to choose a CSV file by clicking on a button marked in Section B, to preview the pattern of missing data by clicking on the button in Section C, and finally to apply the changes. The preview plots appear under Section E and are specifically generated based on the selected imputation method and provided parameters.

After the user applies the imputation method, another similar screen is shown to choose the appropriate clustering method. This screen, as shown in Figure 4.21 as well as the

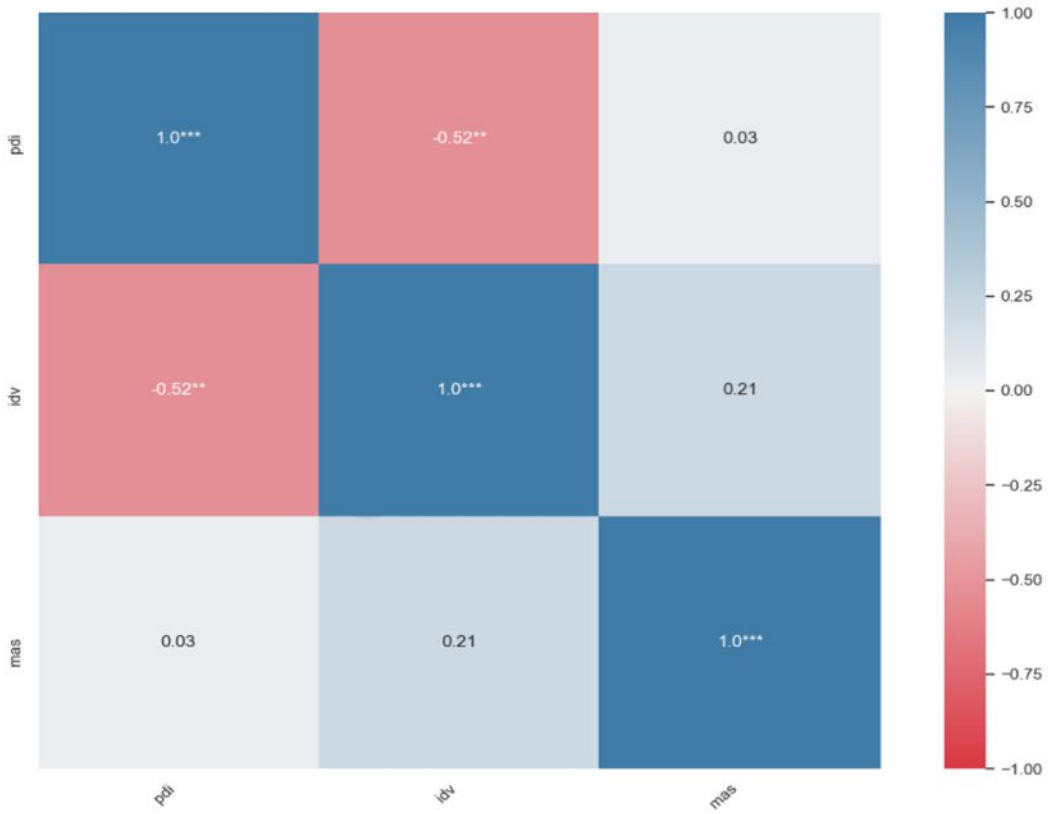


Figure 4.19: Pearson's correlation matrix of 3 dimensions showing that PDI and IDV have a positive correlation (ρ=0.52), statistically significant at 0.01.

option to preview the impact of the clustering method (marked as Section D). In Section D, plots such as a dendrogram and an elbow plot are shown, which can be changed based on the clustering method and parameters.

All the other mock-ups shown for the first prototype are unchanged for the second prototype. No user requirement or system task was changed. The revised UIs still fulfill all the requirements and tasks, with the only significant difference being according to the Visual Analytics process [149]. Our final interface is shown in the upcoming chapter, see Figure 5.2, Figure 5.4 and Figure 5.3.

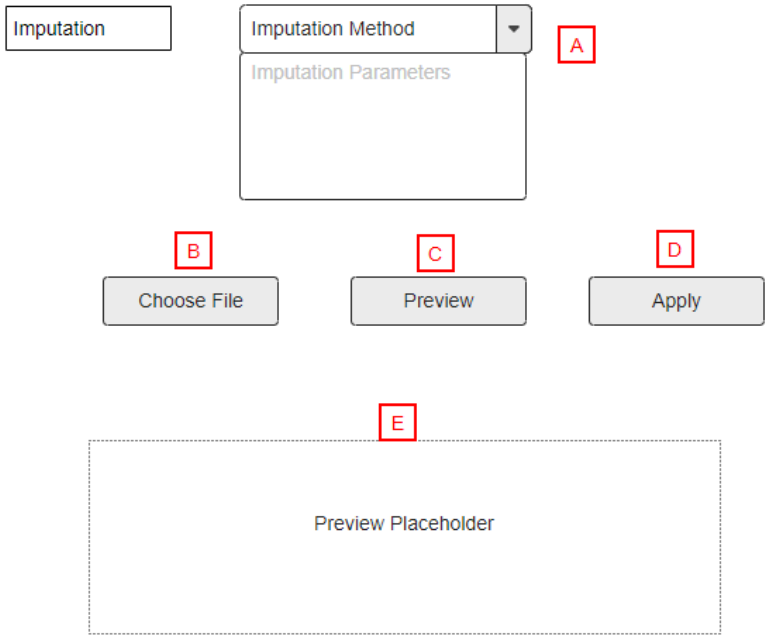


Figure 4.20: Revised Mock-up: Imputation page. First, the user is able to select an imputation method and preview the impact.

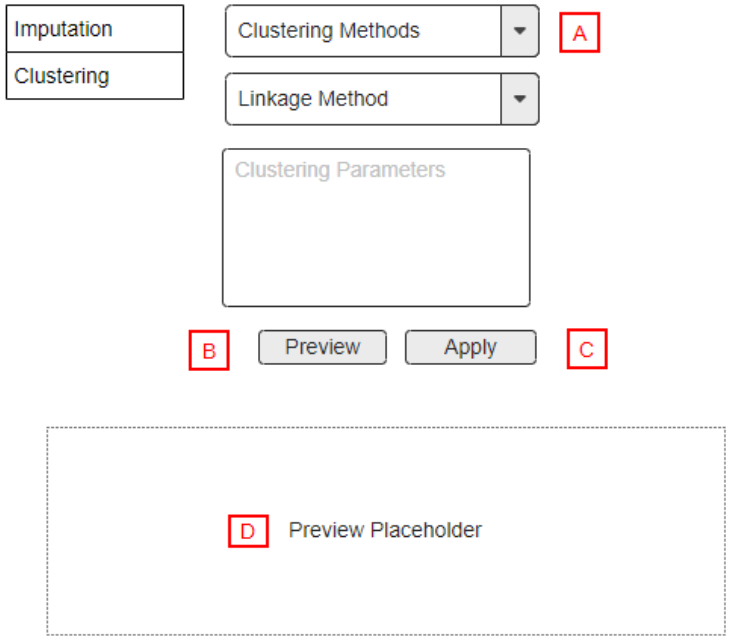


Figure 4.21: Revised Mock-up: Clustering page. The user is able to select a clustering method and preview the impact.

CHAPTER 5

Implementation

In Chapter 4 we defined the user requirements, and we designed mid-level tasks based on the gaps found in the current research field of cultural sciences. Then, a prototype was designed using mock-ups, which were evaluated and re-designed, and revised to a second version. Using these mock-ups, it is possible to choose the appropriate technology stack to build a visualization framework that satisfies all the requirements and fulfills the tasks. Before starting to write code, the final step is to analyze existing programs and libraries that have to support the required visualization methods mentioned in Table 4.1.

5.1 Choosing Appropriate Tools and Technologies

As mentioned in Section 2.2, Bayat [44] has implemented a visualization tool for visualization the dimensions of the Hofstede model that enables researchers to explore the dimensions on a world map. The visualization tool is written in HTML, JavaScript, and D3.js [150] library, which gives a user the flexibility to interact with the tool. There are, however, some limitations in this tool that need to be reworked. Although it is possible to select multiple countries consecutively by choosing them in the drop-down box, a filtering option is not given in this tool. The dataset loaded into the tool has missing values that are untreated and ignored, which does not fulfill the purpose of Tasks 2, 3, and 4. The tool needs to be extended in order to be able to discover knowledge, find correlations, confirm hypotheses, perform a grouping and profiling of the countries, and visualize uncertainty. Finally, it is not possible to revert a selection or update the hive-plot or histogram. Once a country is selected, these plots remain unchanged.

We assessed other alternative technologies such as Tableau Desktop [151], Google-Charts [152], Qlick Sense [153], or Visual.ly [154]. For our visualization purposes, we found that D3.js dominates all the other options. For the purpose of fulfilling our requirements and tasks, we require a technology that is open source, able to be customized, and support interactive visualization methods. Out of all the alternatives, only D3.js and

Google-Charts are free to use; however, Google-Charts has a limited number of charts available and does not give the user the ability to create interactive custom visualizations by coding.

D3 has the ability to bind arbitrary data to Domain Objective Model (DOM) and gives the possibility to do data-driven transformations on the document [155]. The visualizations are created with HTML, CSS, and SVG, which can be represented using a web browser; it has the ability to handle CSV and JSON as input. The fact that it is free and open-source, and the existing community support, is the cause of the increasing use of this library in the visualization domain. D3.js (Data-Driven Documents) has been extensively used to create visualization tools such as DonVis [156], INSVis [157] and libraries like StArE.js [158], networkD3 [159], GenomeD3Plot [160], which prove its usability for visualization purposes.

While D3 is only a client-side scripting library, it has its limitations when it comes to statistical analysis and machine learning features. At the same time, it is a great library to be used to visualize diagrams on a web browser. To reassure that our visualization framework can perform imputation and clustering, which is required for Task 2 and Task 3, another programming language such as Python is helpful.

Python [161] is an interpreted language with an expressive syntax that is freely accessible for programmers. It is superior to other competitors such as R and MATLAB since it is easy to read and it has various libraries and modules available for machine learning and statistical purposes. Guo [162] claims that Python is one of the top coding languages and is superior to its competitors due to its commonplace. Using Python in the visualization framework gives us greater access to public libraries such as *NumPy* [163], *Pandas* [164], *Scipy* [165], and *Sklearn* [166] which can be used for imputation and cluster analysis. Additionally, *Plotly* [167] library on Python has a wide range of access to visualization methods that can be used. Figure 5.1 shows the visualization framework's architecture design, and the API establishes the connection between the server and client-side.

5.2 Implementation of the Visualization Methods

Since we have chosen JavaScript and Python as our technology stack, we have a wide range of options to choose from to base our visualizations on. Some of the visualizations are simple enough to create via the D3.js library. Others can be created faster and easier by using the Plotly library in Python. This section describes how the world map, parallel coordinates, radar chart, dendrogram, elbow plot, heatmap, and clustermap have been implemented.

5.2.1 Parallel coordinates

The application's core relies on this visualization method since it connects all the other components to each other for knowledge discovery. For this purpose, we started our implementation journey with the implementation of this visualization.

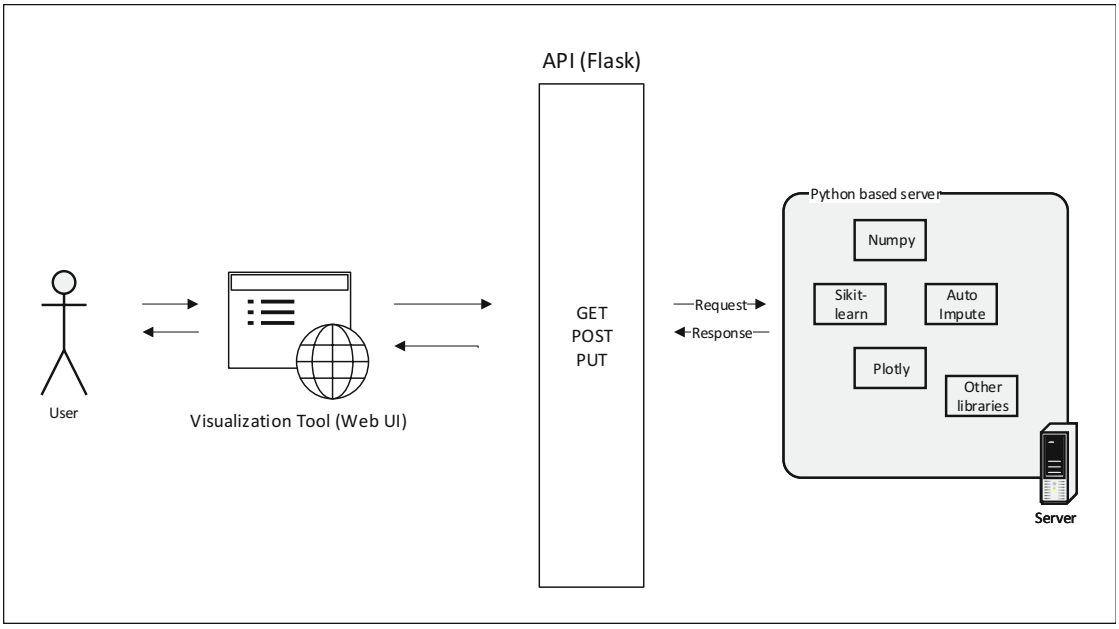


Figure 5.1: Client-server architecture design of the visualization framework. A web-based client communicates with the server using an API. The user interacts with the server using a WebUI through an API implemented in Python.

A D3 library named *d3.parcoords.js* [168] can be used to create visualizations for parallel coordinates in d3. The two methods of *reorderable()* and *brushMode("1D-axes")* written in Listing 5.1 make the parallel coordinates axis re-orderable and brushable.

```
1    d3.parcoords()("#Parallel-coordinates")
2        .data(data)
3        .render()
4        .reorderable() //make it re-ordable
5        .brushMode("1D-axes"); //make it brushable
```

Listing 5.1: Creating a re-orderable and brushable parallel coordinate plot using D3.js library.

5.2.2 World Map

As mentioned in Section 5.1, we based our implementation on the existing visualization framework for dimensions of the Hofstede model named Cultural Model Visualization (CMV), which already had the world map implemented via D3. For our purposes, we had to modify this visualization by first removing the country's dropdown list. Our filtering is based on a connection between the world map and the parallel coordinates. Hence, this dropdown list is redundant. Secondly, we removed the hive-plots since they were not identified as part of the requirements or tasks. Lastly, the code was improved to be more

dynamic so that in the future, we could add or remove more dimensions to the data set if required.

The world map was created using the *d3.geoMercator()* function to create the Mercator projection from a given TopoJSON file [169]. A TopoJSON file is a GeoGSON [170] that encodes topology. It contains several types of JSON objects in a way that represents data about geographical features.

```
1
2 //Getting the mercator projection
3 var merecator = d3.geoMercator()
4     .scale(130)
5     .rotate([352, 0, 0])
6     .translate([width / 2, height / 1.5]);
7
8 projection = d3.geoPath().projection(merecator);
9
10 //Going through all the entries of the TopoJson file and appending
11 //the projection
12 d3.json("110m.json", function (error, world) {
13         mapSVG.selectAll("path")
14             .data(function (d) {
15              return topojson
16                 .feature(world, world.countries)
17                 .features})
18             .enter().append("projection")
19             .attr("d", projection)
20 }
```

Listing 5.2: Creating the world map using textitd3.geoMercator() function in D3.js.

5.2.3 Radar Chart

The radar chart is an extended version of the radar chart implemented by Zhou [171]. We have extended the visualization by adding a filtering option. The user is able to search for a country and show or hide the represented axes and vertices on the chart. This enables the user to be able to compare the countries in a pairwise or in a many-to-many manner.

5.2.4 Dendrogram

The dendrogram is implemented using the *dendrogram* function in *scipy.cluster.hierarchy* package. This visualization is created and converted into JSON on the server-side using the Mpld3 library [172]. Then it return a JSON result to the client side, where the JSON is converted into a visualization using Mpld3's JavaScript library by simply using *mpld3.draw_figure* function. The server-side code for the generation of a dendrogram is seen in Listing 5.3.

```
1
2 #importing required packages
```

```
import pandas as pd
import scipy as sc
from matplotlib import pyplot as plt

def plot_dendrogram(data, linkage_method, p=None, max_d=None):

    #Get the data frame
    df = pd.DataFrame.from_records(data)

    #Get the linkage method
    linked = sc.linkage(df_numeric, linkage_method)

    #Create dendrogram
    sc.dendrogram(
        linked,
        labels=labels.to_numpy(),
        max_d
    )

    #Add labels and title to the plot
    plt.title('Hierarchical Clustering Dendrogram')
    plt.xlabel('sample index or (cluster size)')
    plt.ylabel('distance')

    #Draw the max_d line (where we "cut" the tree)
    if max_d:
        plt.axhline(y=max_d)

    return json.dumps(mpld3.fig_to_dict(fig))
```

Listing 5.3: Creating a dendrogram using scipy on the server-side; Returning the result as JSON using the Mpld3 library.

5.2.5 Elbow plot

The same method as for the dendrogram (see Section 5.2.4), the elbow plot is created on the server-side and sent to the client as a JSON formatted result. The *matplotlib* package is used to generate the plot and *KMeans* function in *scipy* package is used to perform k-mean clustering to calculate the value of k for each iteration as shown in Listing 5.4.

```

#importing required packages
import matplotlib.pyplot as plt
import mpld3
import pandas as pd
from sklearn.cluster import KMeans
import numpy as np
from scipy.spatial.distance import cdist
import json

def elbow_plot_json(data):
    #Get the data frame
```

```
    df = pd.DataFrame.from_records(data)
    distortions = []
    #Loop 20 times to range between 1 to 20 number of clusters
    K = range(1, 20)
    for k in K:
        #compute the SEE for each iteration and store value to distortions
    array
        kmeanModel = KMeans(n_clusters=k).fit(df)
        kmeanModel.fit(df)
        distortions.append(sum(np.min(cdist(df, kmeanModel.cluster_centers_),
     axis=1)) / df_numeric.shape[0])

    # Plot the elbow plot and return result as JSON format
    fig, ax = plt.subplots()
    ax.plot(K, distortions, 'bx-')
    ax.set_xlabel('k')
    ax.set_ylabel('Distortion')
    ax.set_title('The Elbow Method showing the optimal k')
    return json.dumps(mpld3.fig_to_dict(fig))
```

Listing 5.4: Creating a elbow plot using scipy on the server-side; Returning the result as JSON using the Mpld3 library.

5.2.6 Heatmap to Show Pattern of Missing Data

Using *missingno* [173] library, we implemented a heatmap that can be used to visualize the pattern of missing data. This is to fulfill the 'preview' feature for the imputation step. Before choosing an imputation, the user can preview how many entries in each data dimension are missing. Listing 5.5 shows the implementation details of this heatmap. The result is returned in JSON format using the Mpld3 library.

```
#importing required packages
import pandas as pd
import missingno as msno
from matplotlib import pyplot as plt
import numpy as np
import json


def plot_msno(data):
    #create a pandas DataFrame
    df = pd.DataFrame.from_records(data)
    df.replace('', np.nan, inplace=True)
    #create figure
    plt.figure(figsize=(15, 10))
    msno.matrix(df)

    #return as json
    return json.dumps(mpld3.fig_to_dict(fig))
```

Listing 5.5: Creating a heatmap to show the pattern of missing data using Missingno library.

5.2.7 Pearson's Correlation Matrix

We use the Seaborn package [174] to create the correlation matrix as seen in Listing 5.6. The calculation of ρ is based on *.corr*() function provided in Pandas package. To add the "*" symbol, we iterate through the result and add it accordingly as a suffix to the result. Lastly, we annotate the heatmap using our modified labels (value of ρ and the start symbols) and return it in a JSON format.

```
#importing required packages
import matplotlib
import searborn as sns
import pandas as pd
import numpy as np
import json

def correlation_matrix(data):
    #Create Pandas data frame
    df = pd.DataFrame.from_records(data)
    #Calculate value of PCC
    pval = df.corr(method=lambda x, y: pearsonr(x, y)[1])
    #Add the * for significance level
    p = pval.applymap(lambda x: ''.join(['*' for t in [0.001,0.01,0.05] if x
    <= t]))
    annotation = rho.round(2).astype(str) + p

    ax = sns.heatmap(
        rho,
        annot=annotation,
    )

    #return as json
    return json.dumps(mpld3.fig_to_dict(fig))
```

Listing 5.6: Creating a Pearson's correlation matrix using Searborn.

5.2.8 Clustermap

We created a clustermap using Python's *Plotly* [175] library. Alternatives such as Seaborn [176], or DashBio [177] clustermap libraries were found not to be useful. We were facing a bug in Seaborn, where zooming between the dendrogram and heatmap was not in sync, meaning that zooming into a particular section in the dendrogram would not represent the correct position in the heatmap. DasBio uses, Plotly's Dash library [178] to create a bioinformatics oriented suite of components that make visualizations for bioinformatic purposes easier. Thus, for our purpose, *Plotly* is the best option since it gives us the highest flexibility.

First, we create a dendrogram using the *figure factory* module in *Pltotly*. Then, an annotated heatmap is being created. Finally, they both get combined by adding the traces to a figure. The code behind this logic can be found in Listing 5.7.

```

```

```
import plotly.figure_factory as ff
import pandas as pd
import numpy as np

    #Create the side dendrogram
    fig = ff.create_dendrogram(data_array, orientation='right')
    for i in range(len(dendro_side['data'])):
        fig['data'][i]['xaxis'] = 'x2'
.
.
.
    # Add Side Dendrogram Data to Figure
    for data in fig['data']:
        fig.add_trace(data)

        heatmap = ff.create_annotated_heatmap(
        x=dendro_leaves,
        y=dendro_side['layout']['yaxis']['tickvals'],
        z=heat_data,
        annotation_text=missing_df.values,
        pbias=pbias_array,
        colorscale='Blues',
        showscale=True,
        colorbar={"xpad": 100}
    )
.
.
.
    # Add Heatmap Data to Figure
    for data in heatmap['data']:
        fig.add_trace(data)
    return json.dumps(mpld3.fig_to_dict(fig))
```

Listing 5.7: Creating a clustermap using Plotly's figure factory package.

We change the *create_annotated_heatmap()* function in *Plotly* library to visualize the uncertainty of the data by marking the imputed data entries in the heatmap with a dot. The bigger the dot is, the higher is the uncertainty of the imputed entry as explained in Section 4.2.2.4. Figure 4.12 illustrates the final visualization, with the dendrogram on the right side, the heatmap in the middle with marking imputed entries via a dot. This visualization is interactive in a way that when the user zooms into the heatmap or the dendrogram, the visualization is changed.

5.3 Implementation of the Imputation and Automated Algorithms for Data Mining

To ensure that our visualization framework can perform Task 2 and Task 3, it is required to implement the imputation and clustering algorithms. The *Scikit-learn* [179] library, which has been purely written in Python, gives us a wide range of options to use for

imputation and clustering purposes. In this section, we discuss each of the implemented methods in detail.

In order to perform simple imputations such as mean, median, and most frequent imputation, the *sklearn.impute.SimpleImputer* [180] can be used. As seen in Listing 5.8, by calling the function *simple_imputation_df*, putting a data frame with missing data, and one of the strategies defined in the documentation as a string, a simple imputation can be performed.

```
from sklearn.impute import SimpleImputer

missing_df #A data frame with missing data
mean_imp_result_df = simple_imputation_df(missing_df, "mean")
median_imp_result_df = simple_imputation_df(missing_df, "median")
mf_imp_result_df = simple_imputation_df(missing_df, "most_frequent")

def simple_imputation_df(data_missing_value, strategy):
    imputer = SimpleImputer(missing_values=np.nan, strategy=strategy)
    imputer = imputer.fit(data_missing_value)
    return imputer.transform(data_missing_value)
```

Listing 5.8: Simple imputation algorithm based on *sklearn.impute.SimpleImputer* library.

The k-nearest neighbor imputation (KNN) was implemented using the *KNNImputer* class in the *Skit-learn* library. The Multivariate Imputation by Chained Equations (MICE) method was implemented by using the *IterativeImputer*. The details of their implementation can be seen in Listing 5.9.

```
from sklearn.impute import KNNImputer
from sklearn.impute import IterativeImputer

def knn_post(json_data, neighbors):
    df = json_to_df(json_data)
    impute = KNNImputer(n_neighbors=cast_val(neighbors, int))
    result = impute_values(df, impute)
    return result


def iterative_post(json_data, max_iter):
    df = json_to_df(json_data)
    impute = IterativeImputer(max_iter=cast_val(max_iter, int), verbose=0)
    result = impute_values(df, impute)
    return result

def impute_values(df, imp):
    #Get bias of imputation method
    bias = calculate_bias(df, imp)

    #indicate which values are missing, gives a matrix of missing values only
    indicator = MissingIndicator(missing_values=np.nan, features="all")
    mask_missing_values_only = pd.DataFrame.from_records(indicator.
    fit_transform(df))

```

```
    #Impute missing values
    df_imputed = pd.DataFrame(imp.fit_transform(df),
                                            columns=list(df.columns.values))

    #Return a JSON with imputed data with
    # a matrix which marks missing values in the dataset and the bias
    result = jsonify(data=df_imputed.to_dict(orient='records'),
                     missing=mask_missing_values_only.to_dict(orient='records
    '),
                     pbias=bias)

    return result
```

Listing 5.9: Implementation of KNN and MICE imputation.

Lastly, Listing 5.10 shows how the *k*-means, meanshift and a hierarchical clustering is implemented using the *sk-learn* package. All parameters having a *_val* suffix are parameters in which the value needs to be set by the user. These values are sent as input from the client.

```
from sklearn.cluster import KMeans, MeanShift, AgglomerativeClustering

    #k-means clustering
    km = KMeans(
        n_clusters=n_clusters_val,
        n_init=n_init_val,
        tol=tol_val,
        random_state=random_state_val
    )
    km_result = km.fit_predict(df._get_numeric_data())

    #meanShift clustering
    mean_shift = MeanShift(
        bandwidth=bandwidth_val,
        max_iter=max_iter_val
    )
    mean_shift_result = mean_shift.fit_predict(df._get_numeric_data())

    #Spectral clustering
    spectral = SpectralClustering(
        n_clusters=n_clusters_val,
        assign_labels=assign_labels_val,
        random_state=assign_labels_val
    )
    spectral_result = spectral.fit_predict(df._get_numeric_data())

    #Agglomerative or hierarchical clustering
    agglomerative = (
    linkage=linkage_val
    )
    agglomerative_result = agglomerative.fit_predict(df._get_numeric_data())
```

Listing 5.10: Algorithms for clustering.

5.4 Implementation of API

Application Programming Interface (API) is referred to as a computing interface that defines the interaction between software intermediates. We created a web API using Flask [181] library, which provides functionalities and features of a web framework enabling developers to build web applications. Any software client can communicate with the API via the defined paths on a running server. For instance, as listed in the Listing 5.11, sending a post request to *http://SERVER_ADDRES:PORT/clustering/kmean* with a body containing JSON values of data returns the dataset as a JSON format, but the data is clustered using the k-means clustering algorithm. Any output coming from the API (the result of imputation or clustering or plots) has a JSON format. The client-side needs to be able to interpret JSON objects and use Mpld3's library to draw figures. Complete documentation on the API describing all the possible input parameters can be found in a form of Postman documentation [182] (see Section 6.5).

```
from flask import Flask

#Routing path
@app.route('/clustering/kmeans', methods=["POST"])
def clustering_kmean():
    #Getting the JSON data from request
    json_data = request.get_json()

    #Passing the parameters defined in the request to the k-means clustering
    method
    clusters_result = cluster_kmeans(json_data['data'],
                        json_data['n_cluster'],
                        json_data['init'],
                        json_data['n_init'],
                        json_data['max_iter'],
                        json_data['tol'],
                        json_data['random_state'])
    #convert to json, if it is not already the case
    return json.dumps(clusters_result)
```

Listing 5.11: API implementation via Flask library.

5.5 Implementation of the User Interface

Below, we present different UIs of the application and describe their functionality. All of the UIs described below are based on the mock-ups we have created in Section 4.2.4. The bootstrap framework [183] was used during the UI implementation process.

5.5.1 Dashboard User Interface

Figure 5.2 demonstrates the implemented dashboard, which contains four major components. Section A shows the menu. In this section, the user can navigate through the different visualizations and explore them separately. Section B shows the parallel

coordinates, Section C the adapted world map from Bayat [44], and in Section D we show a radar chart. The three sections are connected to each other through parallel coordinates, where we can select countries and filter them.

5.5.2 User Interface for the Imputation

The user interface of the imputation screen is a replica of the mock-ups screen with some minor design improvements. Figure 5.3 illustrates the imputation screen which is the first screen the user sees in the system. Section A serves as a navigation bar, in Section B the user can choose between the available imputation methods which are as following:

- Deletion
- Simple imputations
  - Mean
  - Median
  - Most Frequent
- Iterative imputations
  - MICE
  - KNN
- No imputation

Section C serves for choosing a file. The system supports only CSV files, and upon choosing other files, it throws an unsupported error. The preview button in Section C generates a heatmap on the bottom side of the screen, which shows the pattern of missing data in a heatmap.

The "apply" button in Section E applies the chosen imputation method and goes to the clustering screen described in the next section. The user has the ability not to choose any imputation method. In this case, a case deletion is applied, and every row that includes missing data is dropped.

5.5.3 User Interface for Partitioning

After the imputation is applied to the dataset and there is no missing data anymore, the system goes to the clustering screen as demonstrated in Figure 5.4. In this screen, the user has the option to choose one of the following clustering methods:

- K-Means
- Mean Shift

- Spectral
- Hierarchical
- No clustering

The preview button acts differently based on the chosen clustering method. For k-means clustering, an elbow plot is generated. For Hierarchical clustering, a dendrogram is shown and gives the user an overview of where to cut the dendrogram. A detailed description of how these plots should be interpreted can be found in Section 6.3.1. In addition to the dendrogram in the preview section, once the user chooses the hierarchical clustering method, a clustermap is populated. A detailed overview of the visualizations and how they get populated are summarized in Table 5.1. After choosing the appropriate clustering method, the system populates the visualizations by pressing the apply button.

Visualization method	Populated by
Heatmap for missing data	"Preview" button in imputation screen.
Elbow plot	"Preview" button in partitioning screen. Only for k-means clustering method.
Dendrogram	"Preview" button in partitioning screen. Only for the agglomerative clustering method.
Pearson correlation matrix	Navigation to the dedicated screen through the menu bar.
Radar chart Parallel coordinates World map	Upon loading the dashboard or navigation to its dedicated screen.
Clustermap	Navigation to the dedicated screen. Only for the agglomerative clustering method.

Table 5.1: Overview of all the visualization methods supported in the visualization framework and how to populate the visualizations.

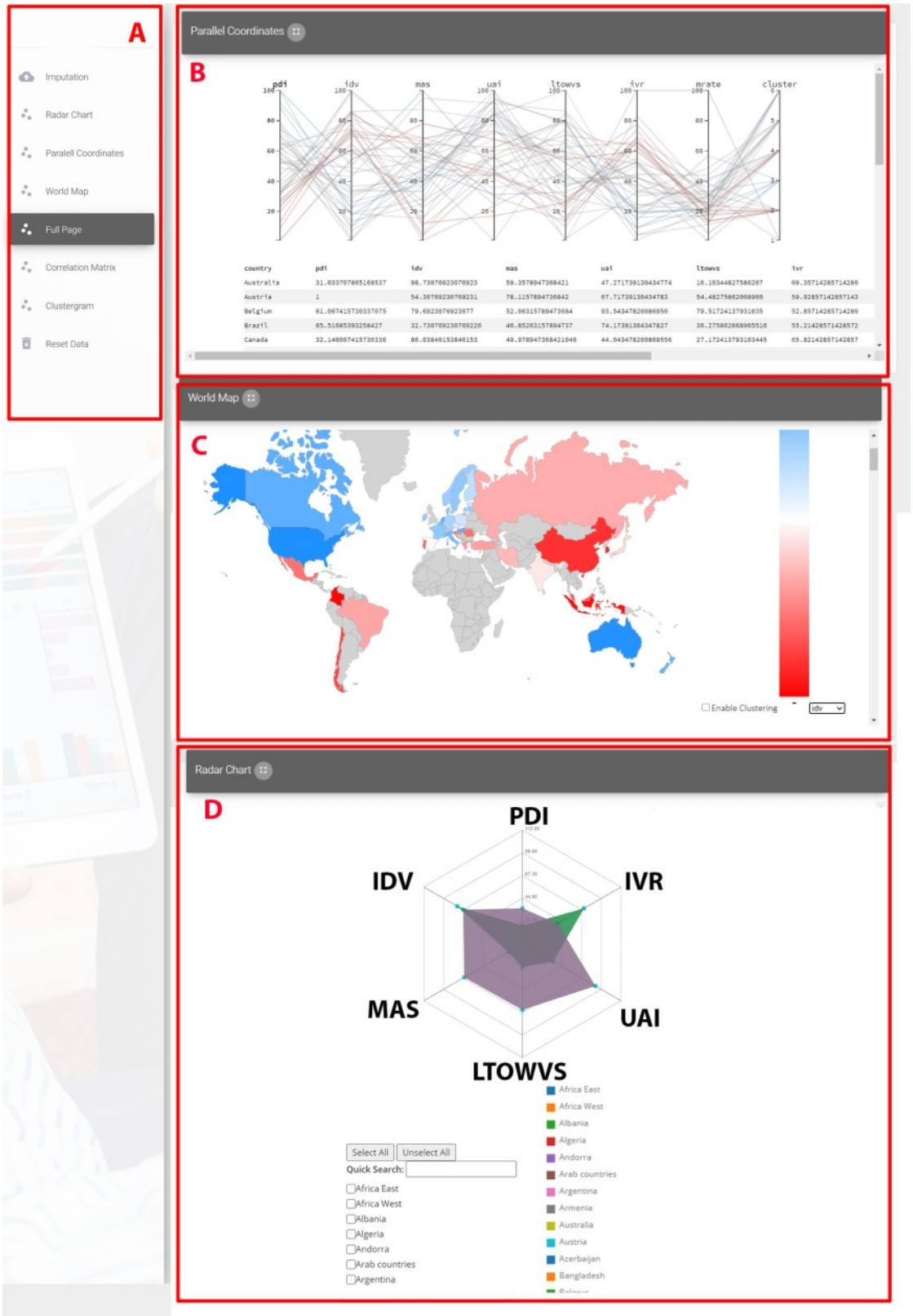


Figure 5.2: Implemented User Interface for the Dashboard. The UI is created based on the mock-ups, showing the menu, parallel coordinates, world map, and radar chart.

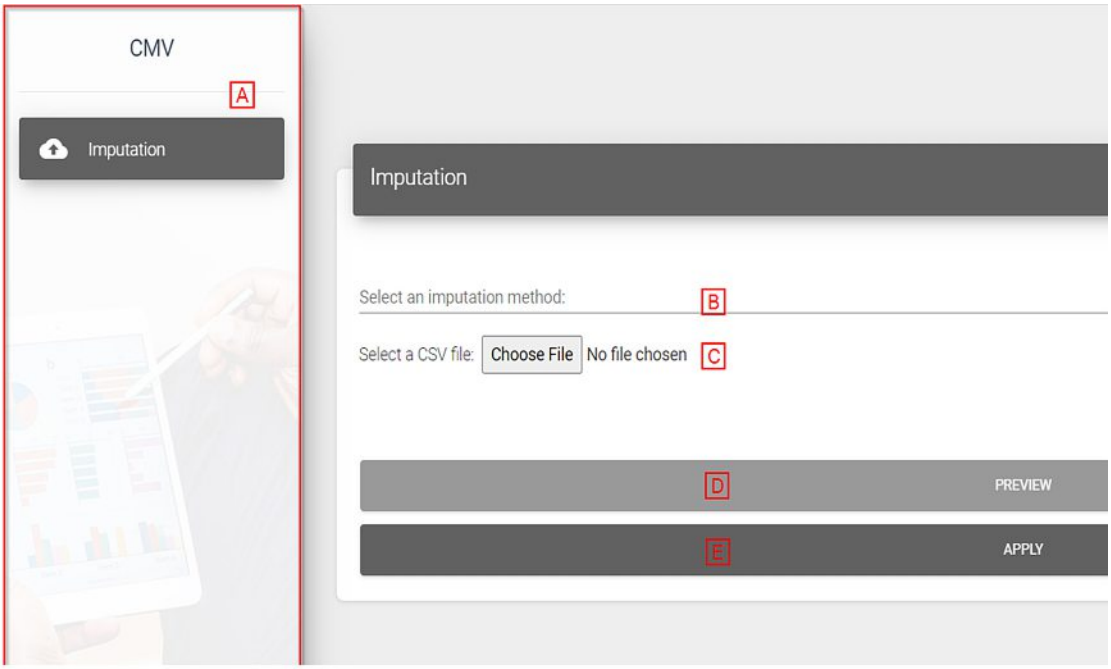


Figure 5.3: Implemented screen for imputation. The UI is created based on the mock-ups, giving the possibility for the user to select a file, choose imputation method, preview the impact and apply.

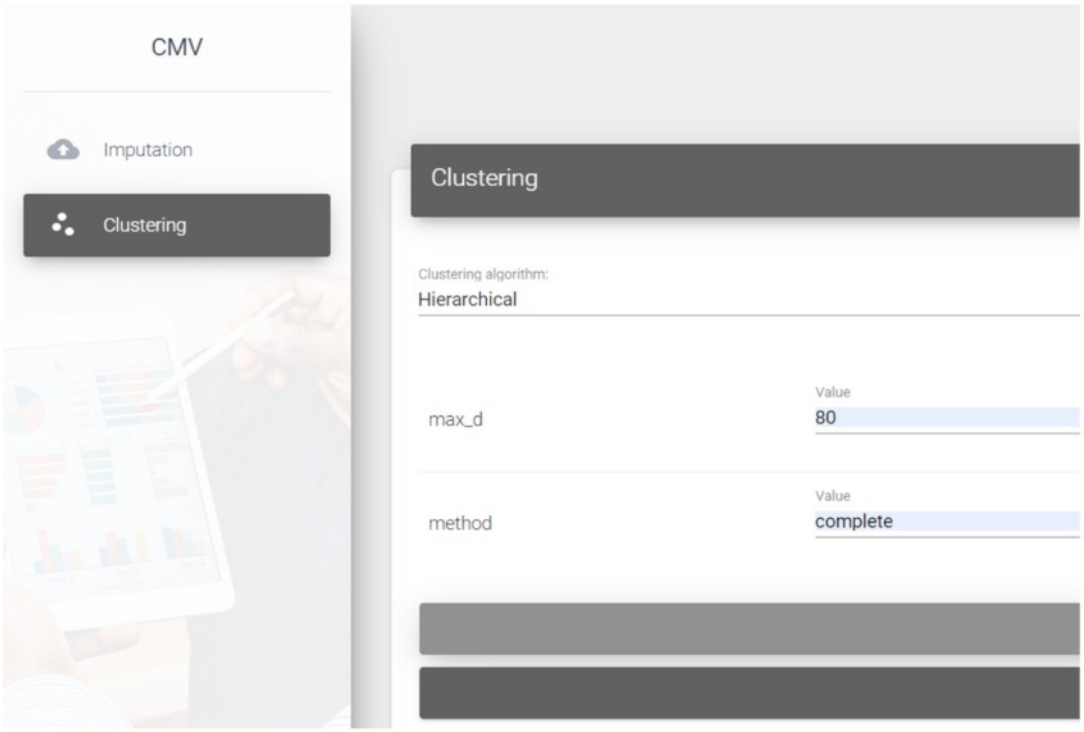


Figure 5.4: Implemented screen for selecting and previewing clustering methods. The UI is created based on the mock-ups; enabling the user to select clustering method, enter parameters, preview the impact and apply.

CHAPTER 6

Results

After the implementation of the visualization framework is finished as described in Chapter 5, we look at the finished artifact and evaluate based on the evaluation method described in Section 3.5. This evaluation is necessary to complete the design cycle of Henver's three-cycle methodology approach described in Section 1.3. Our evaluation is based on case studies, by reviewing existing literature and attempting to reproduce their scenarios using our newly implemented visualization framework and trying to confirm the same hypotheses while assessing our approach's usability.

6.1 Case Study 1: Cultural Dimension of Corruption

Achim [184] examines if culture plays a role in determining corruption levels within a country. This study is conducted using data from 98 countries, with the six dimensions of Hofstede and combining them with the Corruption Perception Index (CPI) acquired from Transparency International's annual report [185]. The CPI report provides an index of 175 countries; the score ranges between 0 (highly corrupt) and 100 (no corruption); however, Achim [184] invert each value, meaning 0 being low and 100 highly corrupted countries. We assume this is done to make the hypothesis tests more understandable. Within this case, we do not invert the values and keep the scoring as in the original report.

As claimed by the author, the hypotheses are tested on an initial sample of 98 countries that had both data of Hofstede's dimension and corruption level available. The methodology of the study of Achim [184] is first to define one main and six sub-hypotheses (see list below). Statistical methods such as ordinary least squares (OLS) analysis and ANOVA are used to identify the correlation coefficients. In the following sections, each of the hypotheses and the corresponding results found by Achim [184] are described and compared to the results found with our visualization framework's support. We provide the six hypotheses, which were tested within our case study:

- H1.1: The higher the power distance, the higher the level of corruption
- H1.2: The less individualistic (more collectivist) a society is, the higher the level of corruption.
- H1.3: The greater the masculinity of a society, the higher the level of corruption.
- H1.4: The greater the level of uncertainty avoidance, the higher the level of corruption.
- H1.5: The shorter the term of orientation, the higher the level of corruption.
- H1.6: The more indulgent the society is, the lower the level of corruption.

Achim [184] uses three tables to show their result. The table shown in Figure 6.1 represents a Pearson's correlation matrix between the dimensions of the Hofstede model and CPI. Figure 6.2 shows the result of simple regression analysis and Figure 6.3 demonstrates the result of running a multivariate regression analysis. All the hypotheses in the study of Achim [184] were assessed with our proposed framework's help, using these three tables and visualizing the correlation in a scatterplot.

Table 1 Pearson correlation

	Corruption	PD	IDV	MAS	UAI	LTO	IND
Corruption	1						
PD	0.585**	1					
IDV	−0.613**	−0.656**	1				
MAS	0.162	0.110	0.051	1			
UAI	0.048	.148	−.119	.047	1		
LTO	−0.344**	−0.121	0.269*	0.081	0.100	1	
IND	−0.165	−0.246*	0.092	−0.094	−0.198	−0.463**	1

Figure 6.1: Pearson's correlation matrix shown by Achim [184] including six dimensions of the Hofstede model and CPI index.[1]

In order to compare the results, we use our newly implemented visualization framework and upload the same data set used in Achim [184]'s methodology. First, it is required to combine the data from the six dimensions of the Hofstede model with the CPI values, which was done by simply adding a new dimension to the Hofstede model data (6 dimensions + CPI). Thus, our final dataset contains seven dimensions. The result is an

[1] **Correlation is significant at the 0.01 level (2-tailed); *Correlation is significant at the 0.05 level (2-tailed).

[2] *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 2 Descriptive statistics and results of the simple regression analysis

Variables	Mean	Standard Deviation	Regression coefficient	Standard errors	t-stat	P
Dependent variable						
Corruption	78.01	47.371				
Independent variables						
Culture						
PD	64.06	20.828	1.291	0.182	0.798	0.000
Adjusted R Square = 0.335 F = 50.38; Prob. = 0.000 $N = 98$						
IDV	39.22	22.048	−1.285	0.168	−7.649	0.000
Adjusted R Square = 0.370 F = 58.50; Prob. = 0.000 $N = 98$						
MAS	47.65	18.647	0.399	0.246	1.619	0.109
Adjusted R Square = 0.016 F = 2.62; Prob. = 0.109 $N = 98$						
UAI	63.86	21.417	0.103	0.218	.473	0.637
Adjusted R Square = −0.008 F = 0.224; Prob. = 0.637 $N = 98$						
LTO	41.75	22.897	−0.702	0.210	−3.339	0.001
Adjusted R Square = 0.108 F = 11.147; Prob. = 0.001 $N = 84$						
IND	48.22	22.907	−0.335	0.229	−1.462	0.148
Adjusted R Square = 0.015 F = 2.136; Prob. = 0.148 $N = 77$						

Figure 6.2: Result of simple regression analysis conducted by Achim [184], where the CPI index is the dependent variable and the six dimensions of the Hofstede model are the independent variables.

initial sample of 63 countries, as seen in the Appendix A.1, which is lower than what was used by the author. Unfortunately, it is not clear which countries are missing since Achim [184] does not explicitly mention which countries are used in their sample.

6.1.1 Testing for H1.1: *The higher the power distance, the higher the level of corruption*

Achim [184] shows that there is a positive and medium correlation between PDI (power distance) and corruption. This hypothesis is accepted, after examining the correlation coefficient between PDI and CPI (ρ=0.585), as shown in Figure 6.1. This figure indicates a positive and medium correlation with a 1% level of significance. The result remains significant for the initial regression (Figure 6.2) and controlled for another cultural variable in the multiple regression analysis (Figure 6.3). This positive correlation is also represented in a scatterplot, as illustrated in Figure 6.4.

We examine the correlation between CPI and PDI with the support of our implemented visualization framework. For this purpose, we use the parallel coordinates, which in their

Table 3 Models of corruption as a function of culture

Variables	Model 1	Model 2	Model 3
PD	0.792**	0.900***	0.962***
IDV	−0.551*	−0.465*	−0.554*
MAS	0.304		
UAI	−0.058		
LTO	−0.655***	−0.648***	−0.452**
IND	−0.406*	−0.401*	
Adjusted R Square	0.52	0.51	0.49
Prob.	0.000	0.000	0.000
F	14.889	21.558	28.085
N	77	77	77

Figure 6.3: Result of multivariate regression analysis conducted by Achim [184].[2]

initial state is shown in Figure 6.5.

It is not possible to spot correlations within dimensions quickly. The user needs to interact with the visualization and modify it according to the task to inspect this. In this case, we first *arrange* the axes so that PDI and CPI are adjacent. The adjustment is possible by simply dragging the PDI dimension with the mouse and re-positioning it beside CPI, as shown in Figure 6.6. An additional change in this visualization is that the CPI axis is inverted since the corruption level is higher in countries with a lower CPI score. Thus, it is easier to examine the correlation by inverting the axis. The axis of any dimension can be inverted by simply double-clicking on the title.

The parallel coordinate plot's coloring reflects the Z-score of any *selected* dimension, where blue is high values, and red is low values of Z. The coloring can be changed by clicking on the title of any axis in the parallel coordinates. Figure 6.6, the coloring is based on the PDI dimension; hence, the axis's title is in bold text format.

A correlation between PDI and CPI is already visible in Figure 6.6, since high PDI values are mapped to low values of CPI and vice versa. However, this correlation is even more visible if the user *filters* high and low PDI values by using the available brushing method. The visualization can be filtered based on each axis in the parallel coordinates by dragging and selecting the required value on the dimension's axis. Figure 6.7 shows the parallel coordinates in a state where the user has filtered only countries with a PDI level over 90 by brushing the PDI axis. While on the contrary, Figure 6.8 shows the visualization when the user filters the PDI level below 35. The result can be seen in the table below the parallel coordinates.

Finally, we review Pearson's correlation matrix in the visualization framework (shown in

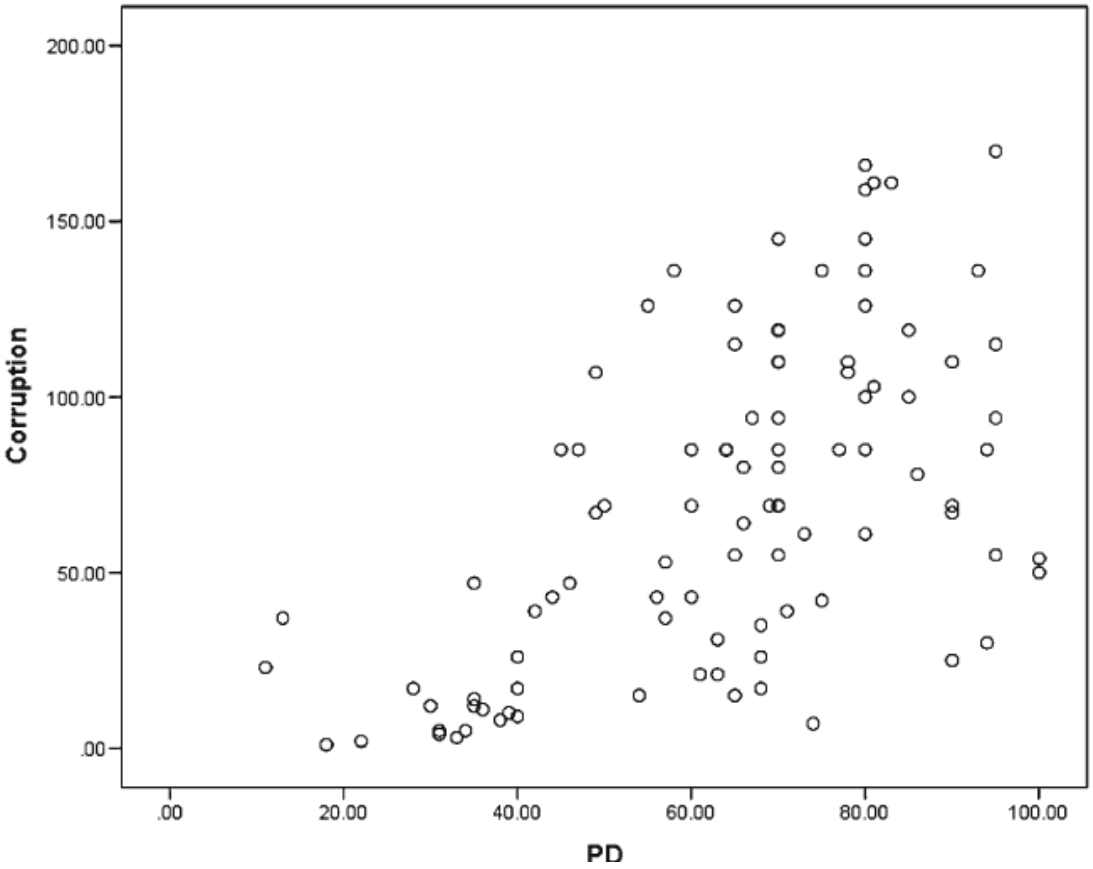


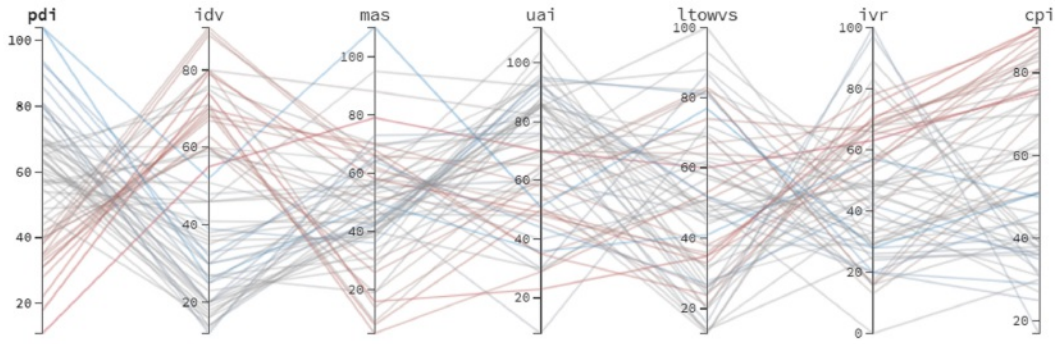
Figure 6.4: Scatterplot visualization of CPI and PDI created by Achim [184], where the y axis is showing the corruption level and the x axis representing the PDI score.

Figure 6.9). Similar to the parallel coordinates, low values are colored with red and high values are blue. Based on Pearson's correlation matrix, it is evident that the correlation coefficient between PDI and CPI is -0.67 with a significance level of 1%. Meaning, the higher the PDI score, the lower the CPI level (higher corruption in the country), which confirms H1.1.

H1.1, as proposed by Achim [184], is confirmed with the support of our framework. PDI has a negative correlation with CPI.

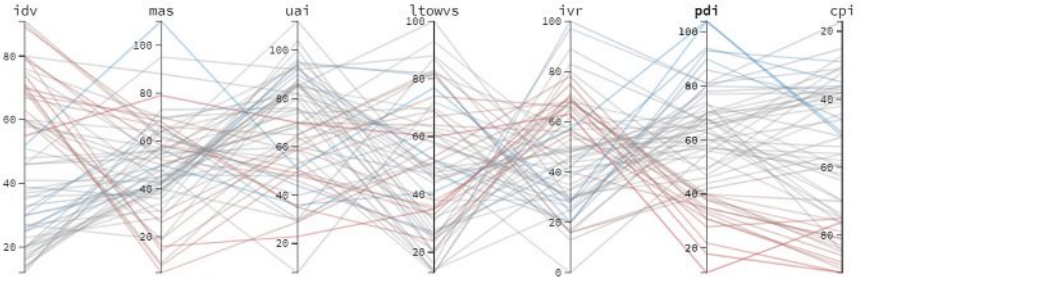
6.1.2 Testing H1.2: *The less individualistic (more collectivist) a society is, the higher the level of corruption*

The result of testing H1.2 by Achim [184] reveals that based on Figure 6.1, there is a negative and medium Pearson correlation between IDV (individualism vs. collectivism) and corruption with a correlation coefficient of -0.613 at 1% of significance. Both Figure 6.2 and Figure 6.2 show a negative correlation with a significance level of ($p < 0.001$ and



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76

Figure 6.5: Initial state of parallel coordinates for Hofstede's six dimensional model and CPI used for case study 1 (data source is Table A.1). There are seven axes in the parallel coordinate plot, and each axis is representing one cultural dimension or CPI. The coloring is based on the Z-score.

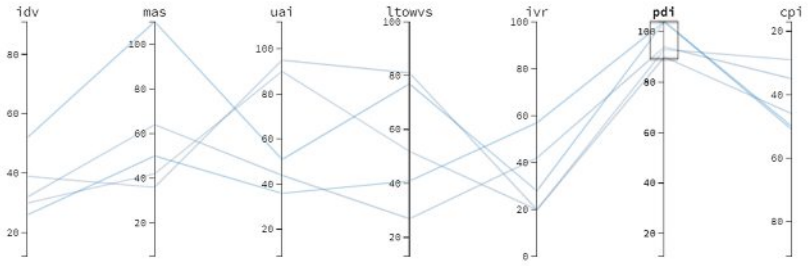


country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76
Bangladesh	80	20	55	60	47	20	25

Figure 6.6: Visualization of parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Here we changed the position of the PDI axis to be beside CPI. Changing the arrangement of coordinates leads to a better overview and comparison between dimensions.

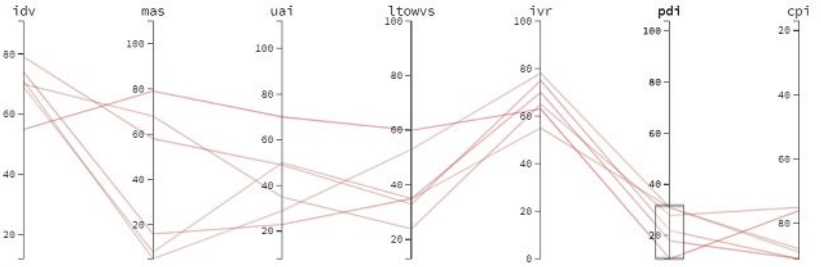
$p < 0.05$) between IDV and CPI, meaning that the less individualistic society is, the higher the level of corruption, which confirms H1.2. Figure 6.10 shows the scatterplot which represents the correlation between these two dimensions.

Using the same approach as in H1.1 from Section 6.1.1, we use the same parallel coordinates without loading the data again. *Re-arranging* the axis and putting IDV beside CPI enables easy comparison between these two dimensions. Also, the color of



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Malaysia	104	26	50	36	41	57	50
Philippines	94	32	64	44	27	42	35
Romania	90	30	42	90	52	20	46
Russia	93	39	36	95	81	20	29
Slovak Rep	104	52	110	51	77	28	51

Figure 6.7: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering PDI to values over 95 to investigate countries that have a high value of PDI, using the brushing method on the PDI axis.



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Austria	11	55	79	70	60	63	76
Denmark	18	74	16	23	35	70	91
Ireland	28	70	68	35	24	65	75
New Zealand	22	79	58	49	33	75	91
Norway	31	69	8	50	35	55	88
Sweden	31	71	5	29	53	78	89

Figure 6.8: Parallel coordinates of Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). We filtered here PDI to values below 35, using the brushing method on the PDI axis to investigate countries which have a low value of PDI.

the parallel coordinates polylines should be changed to reflect the value of IDV. Double-clicking IDV's title can achieve this. Figure 6.11 shows the visualization of the parallel coordinates after the re-arrangement is applied. A positive correlation between these two dimensions appears. However, some outliers are visible, as well. Figure 6.12 shows countries with high IDV (such as Australia, Great Britain, and the U.S.A.) also having high CPI, meaning that corruption in these countries is lower. Figure 6.13 shows countries with low values of IDV (such as Colombia, Indonesia, Venezuela, Peru) have low CPI

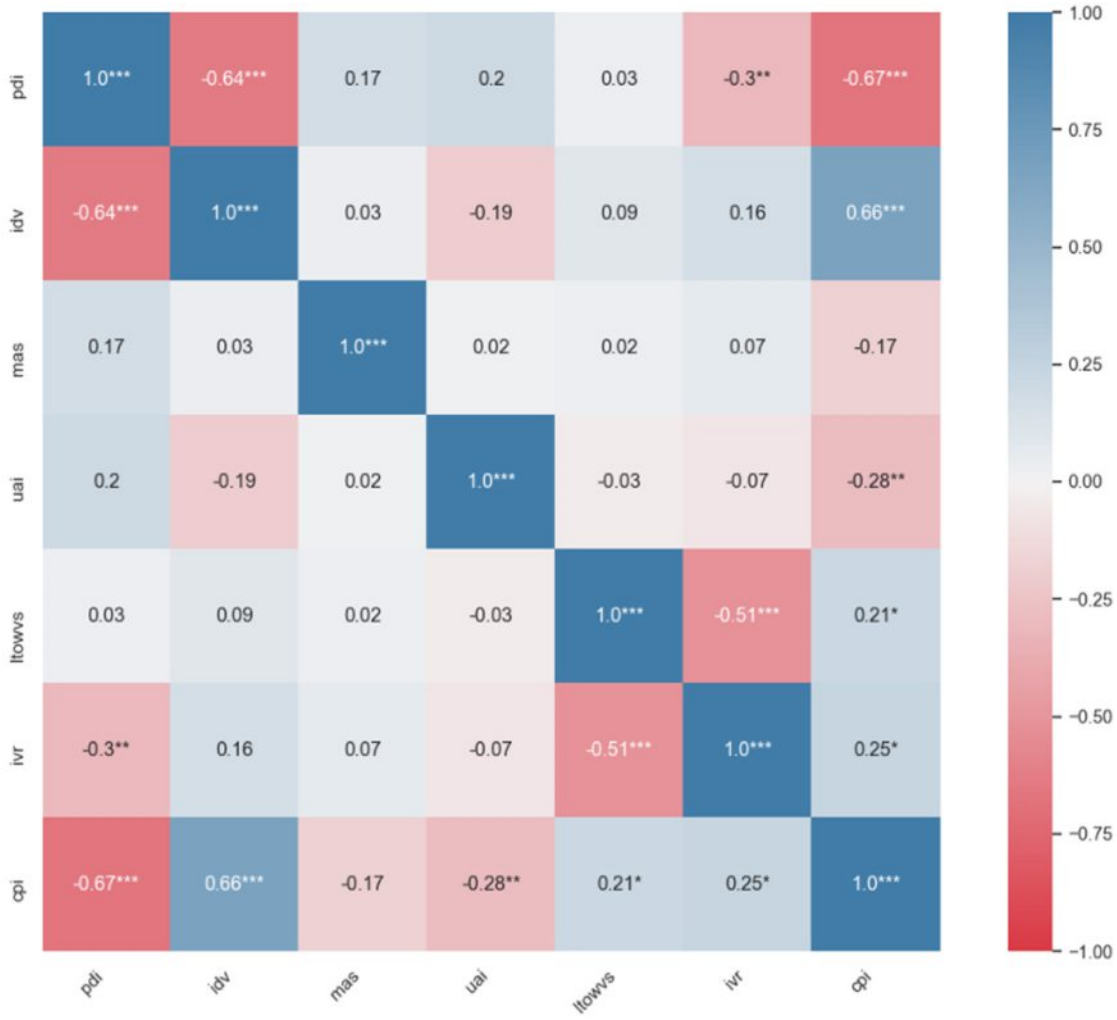


Figure 6.9: Pearson's correlation matrix generated using our visualization framework for the six dimensions of the Hofstede model and the CPI index Used for case study 1 (data source is Table A.1).

values, meaning that the corruption in these countries is higher. Referring to Pearson's correlation matrix shown in Figure 6.9, a correlation coefficient of 0.66 with a significance level of 1% can be detected. Therefore, a positive correlation between the two dimensions of IDV and CPI is confirmed. Hence, the less individualistic (or more collectivist) a society is, the higher the corruption level, which accepts H1.2.

H1.2, as proposed by Achim [184], is confirmed with the support of our framework. IDV has a positive correlation with CPI.

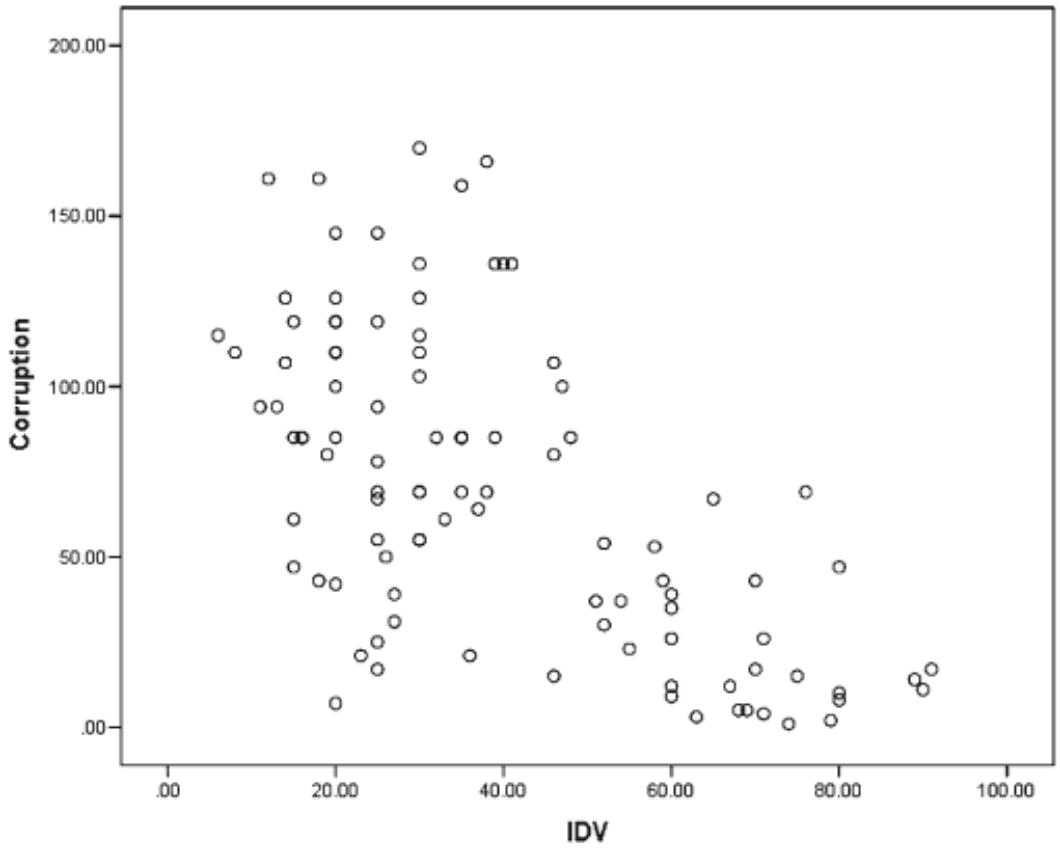


Figure 6.10: Scatterplot of CPI and IDV as indicated by Achim [184]. The y axis showing the corruption level and the x axis representing the IDV score.

6.1.3 Testing for H1.3: *The greater the masculinity of a society, the higher the level of corruption*

H1.3 investigates if there is a correlation between MAS (masculinity vs. femininity) and CPI. Achim [184] rejects this hypothesis based on the reasoning that the Pearson's correlation between these two dimensions is weak (ρ=0.162) and not statistically significant as seen in Figure 6.1. Figure 6.2 and Figure 6.3 show a positive correlation between MAS and CPI (ρ=0.304), but statistically not significant.

The same result can be derived using the visualization framework. Again, the axes of the parallel coordinates need to be *re-arranged* in a way that MAS and CPI are beside each other. The coloring needs to be changed by double-clicking on the MAS axis title as seen in Figure 6.14. Looking at Figure 6.15, it is not easy to detect a particular pattern between high values of MAS and CPI. Only eight countries are having a MAS value higher than 70. However, these country's CPI value ranges between 17 (Venezuela) and 86 (Switzerland) seen in Figure 6.16, indicate a clear pattern. This means that countries

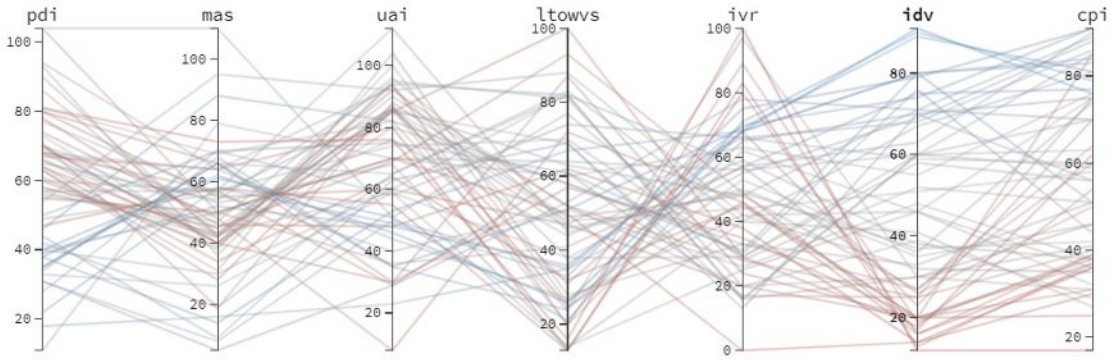
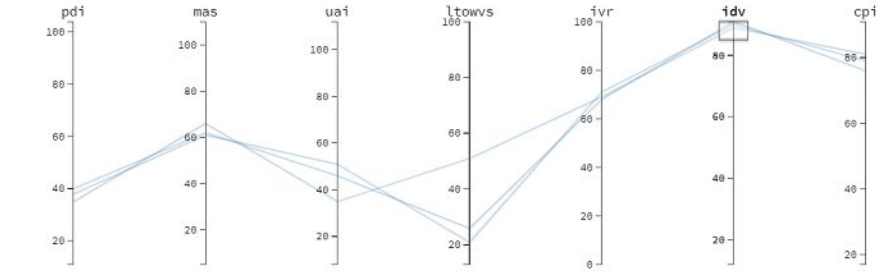


Figure 6.11: Visualization of parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). After re-arrangement of the axis: changing the position of the IDV axis to set it beside CPI. This makes the comparison between the two axes easier.

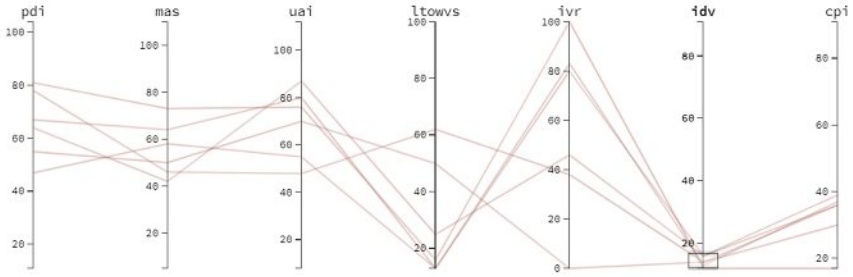


country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Australia	38	90	61	51	21	71	79
Great Britain	35	89	66	35	51	69	81
U.S.A.	40	91	62	46	26	68	76

Figure 6.12: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1): filtering high values of IDV using the brushing method on the IDV axis.

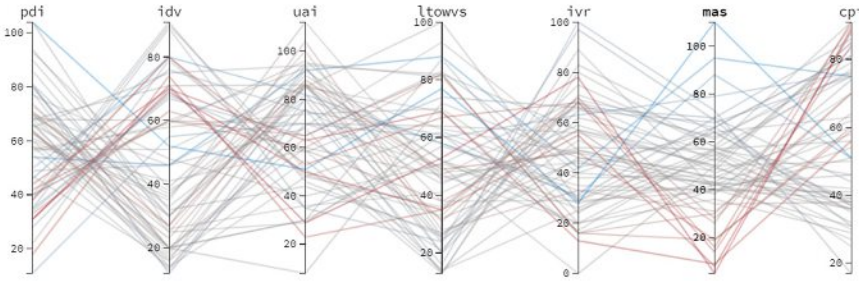
with low values of MAS, such as Denmark, Sweden, and the Netherlands, are less corrupt. However, this is not sufficient to accept the hypothesis since both of the statements do not hold. Additionally, Pearson's correlation matrix illustrated in Figure 6.9 shows a low correlation coefficient score (ρ=-0.17) without any statistical significance. Therefore, with evidence provided by our visualization framework, H1.3 is rejected.

H1.3, as proposed by Achim [184], is also rejected with the support of our framework. MAS and CPI do not correlate with each other.



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Colombia	67	13	64	80	13	83	37
Indonesia	78	14	46	48	62	38	36
Pakistan	55	14	50	70	50	0	30
Peru	64	16	42	87	25	46	36
Trinidad and Tobag	47	16	58	55	13	80	39
Venezuela	81	12	73	76	16	100	17

Figure 6.13: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1): filtering low values of IDV using the brushing method on the IDV axis.

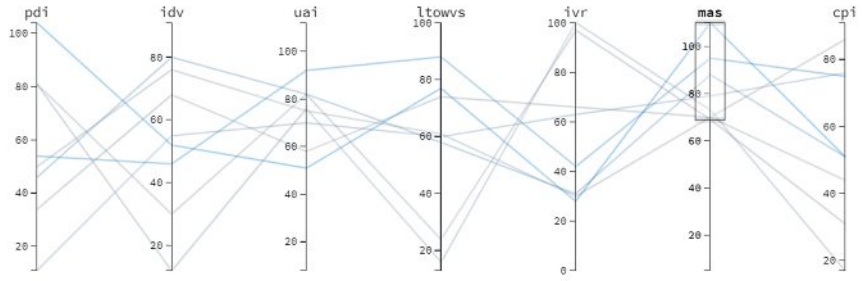


country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76
Bangladesh	80	20	55	60	47	20	25

Figure 6.14: Visualization of parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Re-arranging and changing the position of MAS to set it beside CPI dimension or easier comparison.

6.1.4 Testing for H1.4: *The greater the level of uncertainty avoidance, the higher the level of corruption*

This hypothesis examines if the level of UAI (uncertainty avoidance index) explains the level of corruption. Based on Figure 6.1, it is understandable that there is a positive and low correlation (ρ=0.0048) between UAI and CPI dimension, and not statistically significant. As the result of the simple regression analysis shown in Figure 6.2 and the multivariate regression analysis shown in Figure 6.3, both are statistically not significant.



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Austria	11	55	79	70	60	63	76
Hungary	46	80	88	82	58	31	51
Italy	50	76	70	75	61	30	44
Japan	54	46	95	92	88	42	75
Mexico	81	30	69	82	24	97	31
Slovak Rep	104	52	110	51	77	28	51
Switzerland	34	68	70	58	74	66	86
Venezuela	81	12	73	76	16	100	17

Figure 6.15: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). We filter high values of MAS, using the brushing method on the MAS axis.

Achim [184] rejects H1.4, meaning that there is no correlation between CPI and UAI.

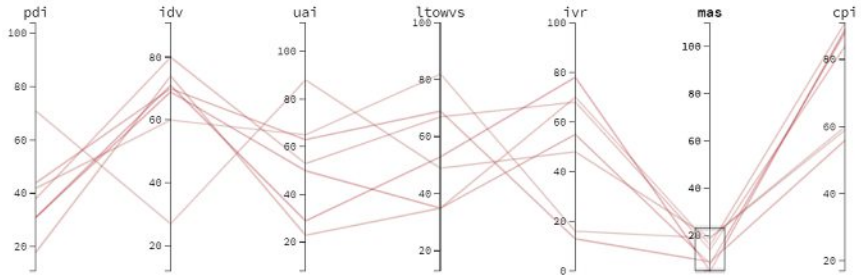
Figure 6.17 represents the parallel coordinates after *re-arranging* its axis to place UAI beside CPI and changing the coloring based on the values of UAI. Looking at the bottom side of the CPI axis, it is visible that the coloring of the polylines is red. This indicates that the countries with a low level of corruption tend to have a low level of UAI.

Upon a closer look at Figure 6.18, we discover that countries with a high level of UAI tend to have a lower rate of corruption with some exceptions, such as Russia, El Salvador, and Greece. On the other hand, referring to Figure 6.19, it is impossible to detect a correlation with low values of UAI easily. Countries such as Indonesia, India, China, and Vietnam seem to have a high level of corruption even though their uncertainty avoidance level is low. The correlation coefficient shown in the Pearson's correlation matrix of the visualization framework (Figure 6.9) indicates a negative correlation (ρ=-0.28) with a statistically significance level of 5%, which is a low level of correlation. For this reason, with the support of our visualization framework, we reject H1.4.

H1.4, as proposed by Achim [184], is rejected with the support of our framework. There is no correlation between UAI and CPI.

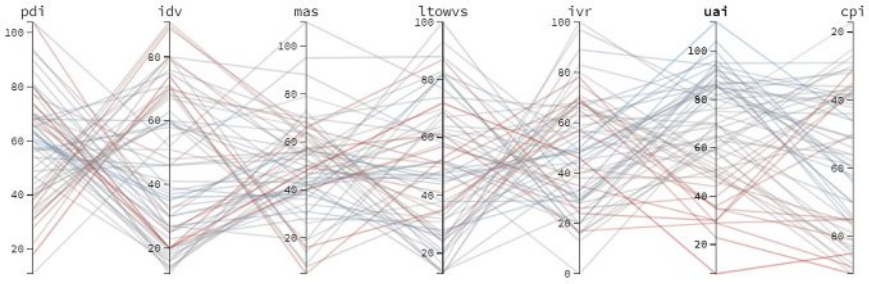
6.1.5 Testing for H1.5: *The shorter the term of orientation, the higher the level of corruption*

This hypothesis examines if LTO (short-term vs. long-term orientation) can explain the level of corruption in a country. The result presented by Achim [184] in Figure 6.1,



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Denmark	18	74	16	23	35	70	91
Latvia	44	70	9	63	69	13	56
Lithuania	42	60	19	65	82	16	59
Netherlands	38	80	14	53	67	68	84
Norway	31	69	8	50	35	55	88
Slovenia	71	27	19	88	49	48	60
Sweden	31	71	5	29	53	78	89

Figure 6.16: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). We filter low values of MAS, using the brushing method on the MAS axis.

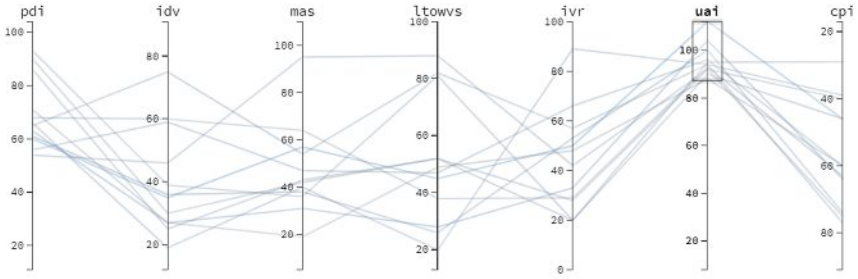


country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76

Figure 6.17: Visualization of parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Re-arranging the position of the UAI axis to set it beside CPI for easier comparison between the two dimensions.

Figure 6.2 and Figure 6.3 show a negative correlation between LTO and CPI (ρ=-0.344) which is statistically significant at a 1% level. The result of the simple linear regression and the multiple regression also revealed a negative correlation with a significance level of 1%. Thus, H1.5 is accepted and the correlation between LTO and CPI is illustrated in Figure 6.20.

The dataset loaded into our visualization framework labeled the LTO dimension of the Hofstede model as LTOWVS. Both of these terms refer to the short-term versus



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Belgium	65	75	54	94	82	57	77
El Salvador	66	19	40	94	20	89	39
Greece	60	35	57	112	45	50	46
Japan	54	46	95	92	88	42	75
Malta	56	59	47	96	47	66	60
Poland	68	60	64	93	38	29	63
Portugal	63	27	31	104	28	33	64
Romania	90	30	42	90	52	20	46
Russia	93	39	36	95	81	20	29
Serbia	86	25	43	92	52	28	40
Slovenia	71	27	19	88	49	48	60
Uruguay	61	36	38	100	26	53	74

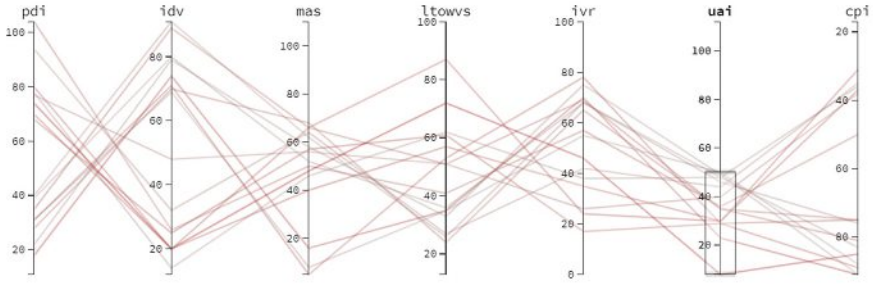
Figure 6.18: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering high values of UAI, using the brushing method on the UAI axis.

long-term orientation levels of a country. Referring to Pearson's correlation matrix of the visualization framework in Figure 6.9, the calculated correlation coefficient is 0.21, indicating a low level of positive correlation between the dimensions with a 10% level of statistical significance.

Upon closer investigation of Figure 6.21, it is impossible to spot any positive or negative correlation between the two axes of LTOWVS and CPI. Some countries such as Russia, China, and South Korea have a high level of LTOWVS and a low level of CPI. Even if the filtering of the countries is changed in the parallel coordinates by *filtering* countries with a high level of CPI, a correlation cannot be detected as seen in Figure 6.22. Countries such as New Zealand, Norway, Finland, and Denmark have a low level of LTOWVS but a high CPI level.

Figure 6.23 shows the parallel coordinates, where the *filter* is applied to countries with a low level of LTOWVS. Countries such as Argentina, Colombia, Iran, and Mexico have a low CPI level; in contrast, countries such as Australia, the U.S.A, and Uruguay have a high CPI level. Based on the visualization, neither a positive or negative relationship can be identified as the result is not conclusive. Figure 6.24 shows the parallel coordinates *filtered* with a high level of CPI. The LTOWVS value of countries with a high CPI value ranges between 33 and 83, and a pattern cannot be identified.

None of the visualizations show any correlation between LTOWVS and CPI. Thus, H1.5



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Canada	39	80	52	48	36	68	83
China	80	20	66	30	87	24	37
Denmark	18	74	16	23	35	70	91
Great Britain	35	89	66	35	51	69	81
Hong Kong	68	25	57	29	61	17	75
India	77	48	56	40	51	26	38

Figure 6.19: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering low values of UAI, using the brushing method on the UAI axis.

is rejected. This contradicts the findings of Achim [184]. However, this result is in accordance with the finding of another study by Tong [186], who found that long-term orientation (e.g., maintaining a relationship) may relate to corrupt activities.

H1.5, in contrast to what is proposed by Achim [184], is rejected. There is no correlation between LTO (LTWVS) and CPI.

6.1.6 Testing for H1.6: *The more indulgent the society is, the lower the level of corruption*

The last hypothesis in the study of Achim [184] examines the correlation between indulgence and corruption level. The results shown in Figure 6.1 indicate a negative low and statistically not significant correlation coefficient (ρ=-0.165) between IND and CPI levels. Additionally, the significance level of the linear regression was not significant (Figure 6.2). When evaluating the other cultural variables, the influence becomes statistically significant, as seen in Figure 6.3. Given these contradicting and mixed results, the author rejected H1.6.

A similar result is achieved with our visualization framework. Figure 6.25 shows the adapted visualization of the parallel coordinates, where the axes are *re-arranged* to position the IVR dimension beside CPI, and the coloring is based on the values of IVR. We also invert the CPI axis to have a better overview of the relation. Note to the author: our visualization framework labeled the indulgence dimension as IVR, instead of IND, which was used by Achim [184]. At first glance, no relation or trend can be found between these two dimensions by only looking at this visualization.

Figure 6.26 and 6.27, show the *filtered* dimensions IVR and CPI, for high values. As

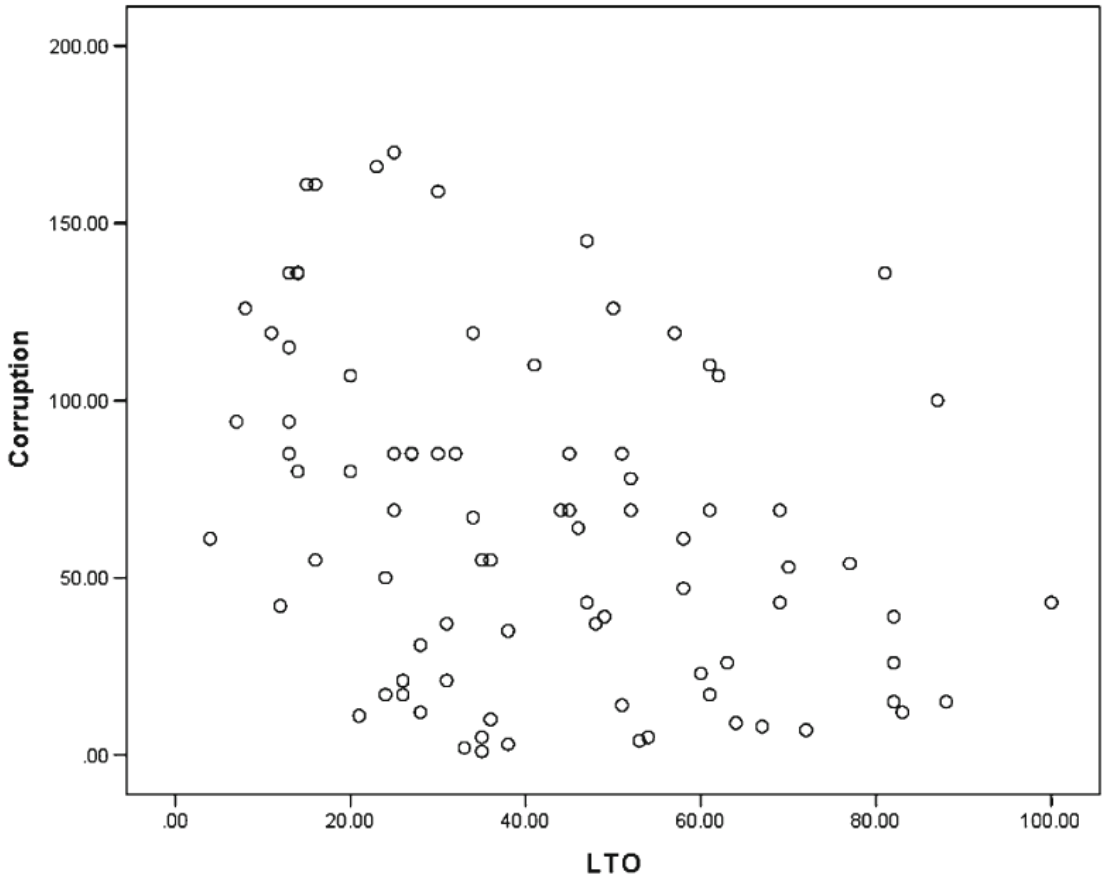
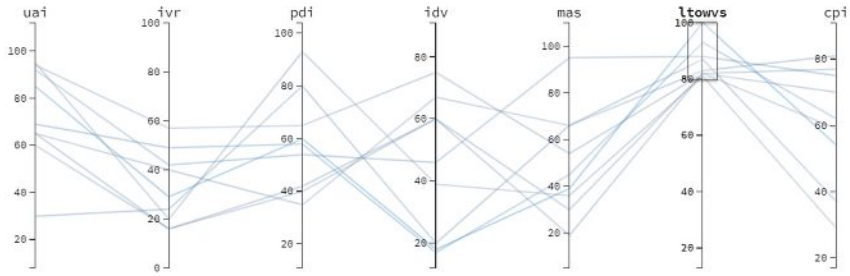


Figure 6.20: Scatterplot of CPI and LTO as indicated by Achim [184]. The y axis is showing the corruption level and the x axis is representing the LTO score.

seen in Figure 6.26, countries having high value of IVR, with the exception of Venezuela, El Salvador, Mexico, Colombia, and Trinidad and Tobago have high values of CPI (meaning low corruption rate). If we filter countries based on their low value of CPI (high corruption), as shown in Figure 6.27, we do not see a clear pattern with regard to the IVR.

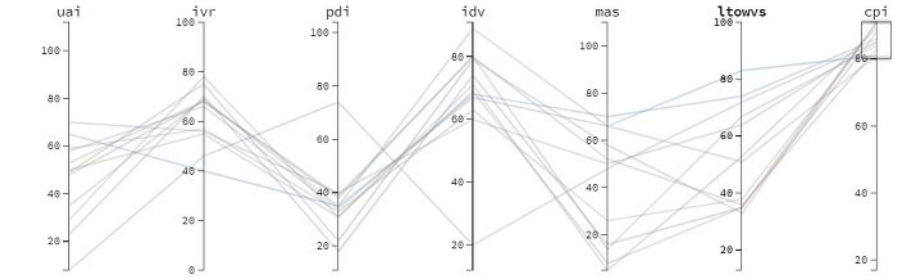
Figure 6.28 shows the parallel coordinates when we filter the low values of IVR. Figure 6.29 shows when the filter is applied to high values of CPI. Both of these representations do not show a specific pattern, which can be described as a relationship or correlation between these two dimensions. The Pearson's correlation coefficient shown in Figure 6.9 indicates a low and positive correlation coefficient of 0.25 with a statistically significance level of 10%. The low correlation coefficient and the lack of a pattern in the representations lead to a rejection of H1.6.

H1.6, as proposed by Achim [184], is rejected with the support of our framework. There is no correlation between IVR and CPI.



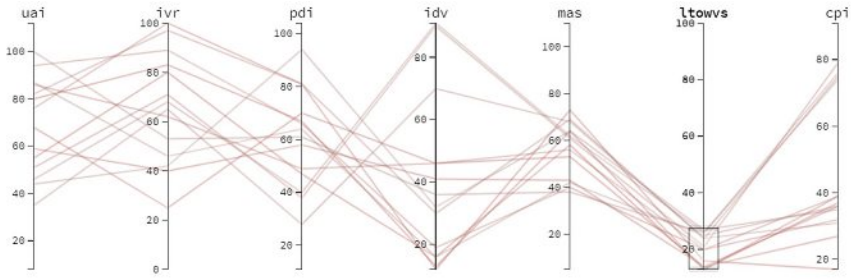
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Belgium	65	75	54	94	82	57	77
China	80	20	66	30	87	24	37
Estonia	40	60	30	60	82	16	70
Germany	35	67	66	65	83	40	81
Japan	54	46	95	92	88	42	75
Korea South	60	18	39	85	100	29	54
Lithuania	42	60	19	65	82	16	59
Russia	93	39	36	95	81	20	29
Taiwan	58	17	45	69	93	49	62

Figure 6.21: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering high values of LTOWVS, using the brushing method on the LTOWVS axis.



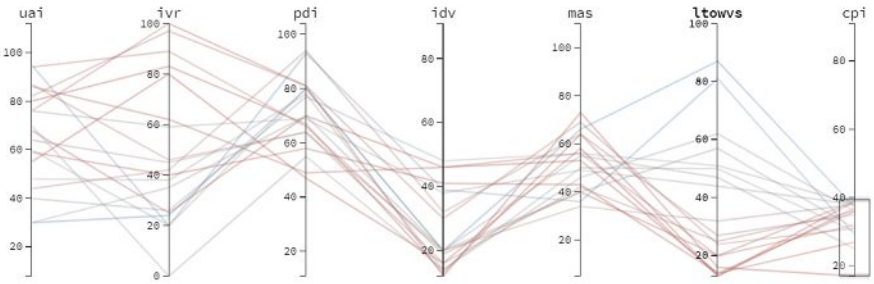
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Canada	39	80	52	48	36	68	83
Denmark	18	74	16	23	35	70	91
Finland	33	63	26	59	38	57	90
Germany	35	67	66	65	83	40	81
Great Britain	35	89	66	35	51	69	81

Figure 6.22: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering high values of CPI, using the brushing method on the CPI axis.



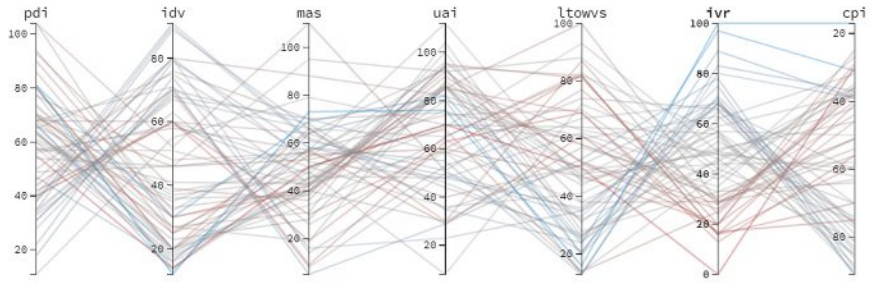
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Colombia	67	13	64	80	13	83	37
El Salvador	66	19	40	94	20	89	39

Figure 6.23: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering low values of LTOWVS, using the brushing method on the LTOWVS axis.



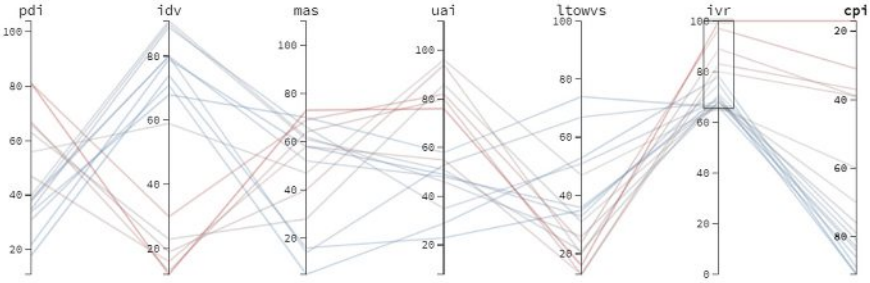
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Bangladesh	80	20	55	60	47	20	25
Brazil	69	38	49	76	44	59	38
China	80	20	66	30	87	24	37
Colombia	67	13	64	80	13	83	37

Figure 6.24: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering low values of CPI, using the brushing method on the CPI axis.



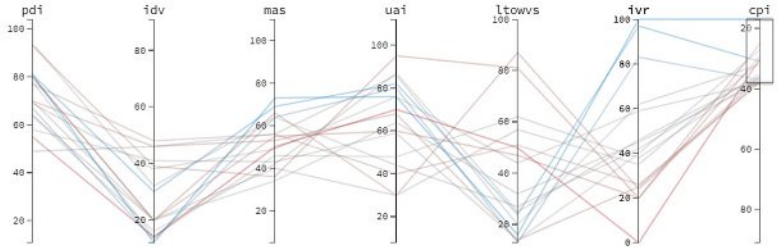
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Argentina	49	46	56	86	20	62	32
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76
Bangladesh	80	20	55	60	47	20	25

Figure 6.25: Visualization of parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). After re-arrangement: changing the position of the IVR axis to set it beside CPI or better comparison.



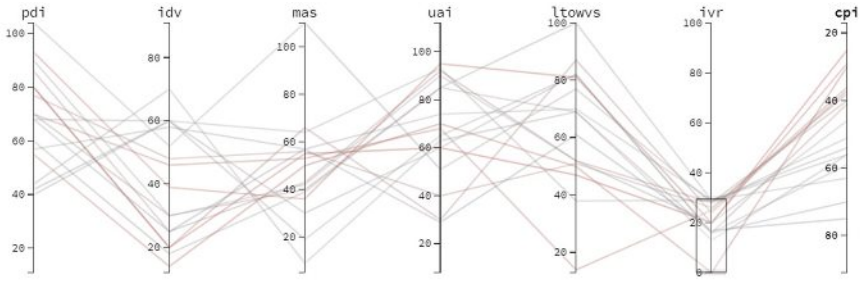
country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Australia	38	90	61	51	21	71	79
Canada	39	80	52	48	36	68	83
Chile	63	23	28	86	31	68	70
Colombia	67	13	64	80	13	83	37
Denmark	18	74	16	23	35	70	91
El Salvador	66	19	40	94	20	89	39

Figure 6.26: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering high values of IVR, using the brushing method on the IVR axis.



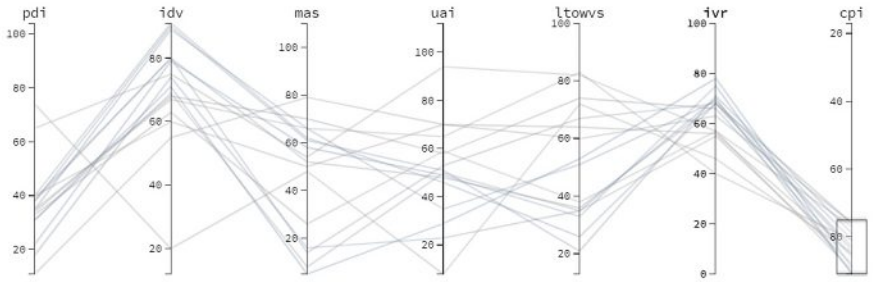
country	pdi	idv	mas	uai	ltowvs	ivr	cpi	no
Argentina	49	46	56	86	20	62	32	1
Bangladesh	80	20	55	60	47	20	25	2
Brazil	69	38	49	76	44	59	38	3
China	80	20	66	30	87	24	37	4
Colombia	67	13	64	80	13	83	37	5

Figure 6.27: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering high values of CPI, using the brushing method on the CPI axis.



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Bangladesh	80	20	55	60	47	20	25
Bulgaria	70	30	40	85	69	16	41
China	80	20	66	30	87	24	37
Czech Rep	57	58	57	74	70	29	56
Estonia	40	60	30	60	82	16	70

Figure 6.28: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering low values of IVR, using the brushing method on the IVR axis.



country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76
Belgium	65	75	54	94	82	57	77
Canada	39	80	52	48	36	68	83
Denmark	18	74	16	23	35	70	91

Figure 6.29: Parallel coordinates in Hofstede's six dimensional model and CPI, used for case study 1 (data source is Table A.1). Filtering low values of CPI, using the brushing method on the CPI axis.

6.2 Case Study 2: Partitioning and Clustering Data

In Hofstede's published book [56], 53 countries are clustered into 12 groups based on four dimensions. Hofstede's goal in clustering the data was to form an empirical typology based on the dimensions. He argues that using this typology the academic research on cultures is easier. In principle, using clustering analysis, Hofstede tried to summarize the cases into a smaller numbers of clusters on the basis of PDI, UAI, IDV and MAS.

The clustering is done using a hierarchical clustering method with average linkage using Statistical Package for the Social Sciences (SPSS) [187] program. The result is represented in a dendrogram (Figure 2.6) and a Pearson's correlation matrix (Figure 6.30). The full details on each cluster member can be found in Table 6.1.

	Product-Moment Correlations Across			
Indexes	53 Countries and Regions	40 Countries	31 Poorer Countries and Regions[4]	22 Wealthier Countries
PDI x UAI	.23	.28*	-.04	.63**
PDI x IDV	-.68***	-.67***	-.24	-.19
PDI x MAS	.06	.10	.23	.14
UAI x IDV	-.33**	-.35*	-.30	_ 59***
UAI x MAS	-.03	.12	-.47**	.37*
IDV x MAS	.08	.00	.21	-.15

a. 1970 GNP/capita < $ 1,000.

*p = .05; **p = .01 ;***/? = .001.

Figure 6.30: Pearson's Correlation matrix of the four dimensions of the Hofstede model created by Hofstede [56] in his book.

This case study aims to replicate the study results in Hofstede's book with the support of our visualization framework and to compare the achieved results. First, in Section 6.2.1, we perform the clustering on the same dataset provided by Hofstede in his book [56], where only four dimensions of the data exist. Then, in Section 6.2.2, we perform the clustering on the six dimensional data and compare it to the four dimensions.

6.2.1 Clustering Performed on Four Dimensions

The dataset provided by Hofstede [56] can be found in Table A.2. The dataset has no missing value in any of the observations. This can be seen in Figure 6.31, in our missingness heatmap. Since there are no white lines found in any row, it means that this dataset does not require any imputation method.

We generate a dendrogram (Figure 6.32) using the hierarchical clustering with an average linkage method similar to what Hofstede used in his book. Differences can be spotted when comparing branches of our dendrogram with what was presented by Hofstede

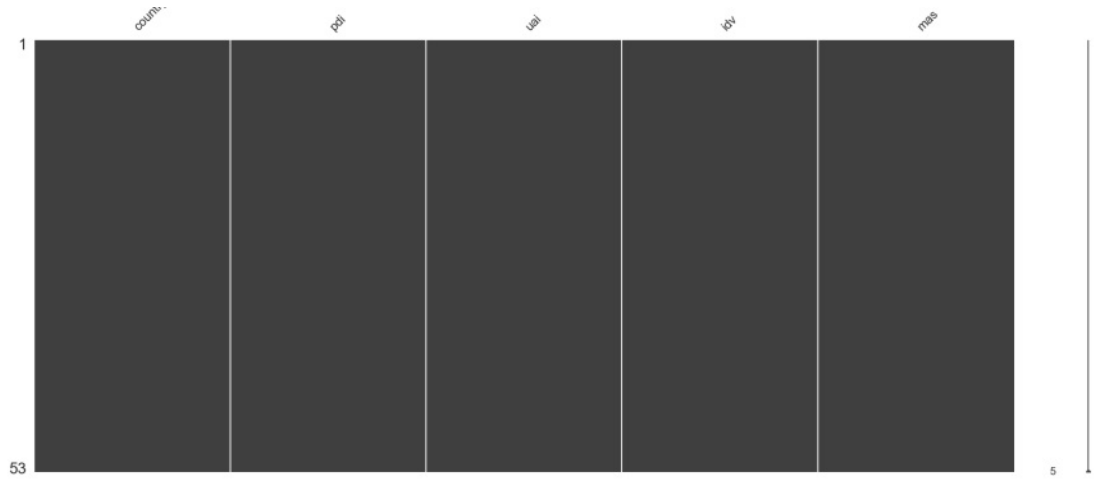


Figure 6.31: Heatmap showing the pattern of missing data. The source of the data comes from Table A.2 and used in Case study 2. Since there are no white lines in the heatmap, we conclude that no data is missing.

(Figure 2.6). The 12 clusters created by Hofstede are arbitrary, as they do not represent a cut at a specific height. To get the closest possible result, we cut the dendrogram at the height of 36.85, resulting in a silhouette coefficient of 0.33 and 12 clusters.

Table 6.1 shows a comparison between the result of our and Hofstede's clustering. Cluster 1 to 6 are the same, both defined by Hofstede and our visualization framework. The differences are in some branches' splits. Hofstede's cluster 7 is split into two clusters: 7 and 8 in our case. Cluster 9 is split into our clusters 9 and 10, while clusters 9 and 10 are merged into our cluster 11. Lastly, clusters 11 and 12 are merged into cluster 12 in our visualization framework. We visualized the result of our clustering on a world map in Figure 6.33 and using parallel coordinates in Figure 6.34. Reviewing the clusters created by our visualization framework, we can identify geographical similarities. Cluster 1 shows a Nordic cluster (Denmark, Sweden, Netherlands, Norway, and Finland). Cluster 3 shows a Latin American cluster (Ecuador, Venezuela, Colombia, and Mexico). Cluster 5 shows a geographical cluster between Belgium and France. Cluster 8 is an East-Asian cluster (Malaysia, Philippines, and India). Lastly, cluster 9 shows a Central American cluster (Guatemala and Panama).

Reviewing the world map together with the parallel coordinates, we can *identify* the following *knowledge*:

Knowledge 1: After applying agglomerative hierarchical clustering on the four dimensions (PDI, UAI, IDV, and MAS) of the Hofstede model, we grouped countries into 12 clusters. A geographical and/or linguistic area is reflected in Nordic countries, Latin America, East-Asia, Belgium and France; and Guatemala and Panama.

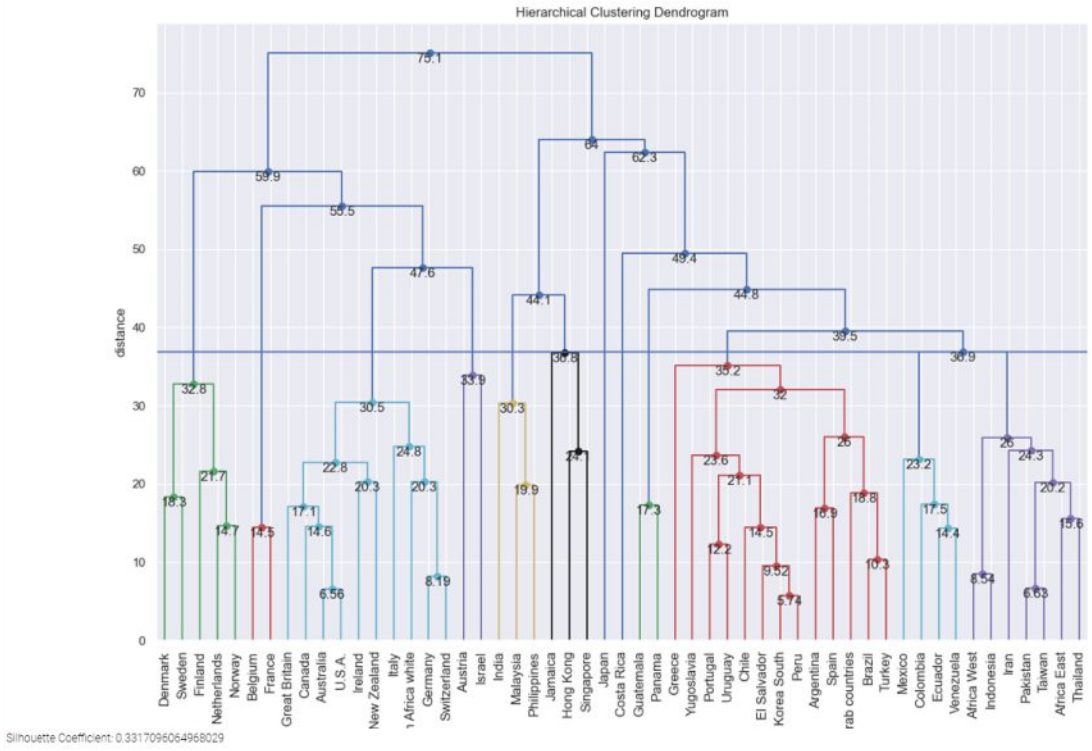


Figure 6.32: Using a dendrogram representation to visualize the effect of agglomerative hierarchical clustering. The source of the data comes from Table A.2 and used in Case study 2. We cut the tree at a height of 36.85 giving us a coefficient of 0.33 and 12 clusters.

Figure 6.35 shows the Pearson's correlation matrix plot generated by our visualization framework. All the ρ values and significance indication are the same with Pearson's correlation matrix shown by Hofstede in Figure 6.30. A strong negative correlation (ρ=-.68) exists between IDV and PDI at a 0.001 level of significance. A negative correlation between IDV and UAI (ρ=-0.33) at 0.01 level of significance is also seen. Thus, we *discover* the following knowledge:

Knowledge 2: Using the four dimensions of the Hofstede model (PDI, UAI, IDV, and MAS), we found a negative correlation between IDV and PDI; and between IDV and UAI.

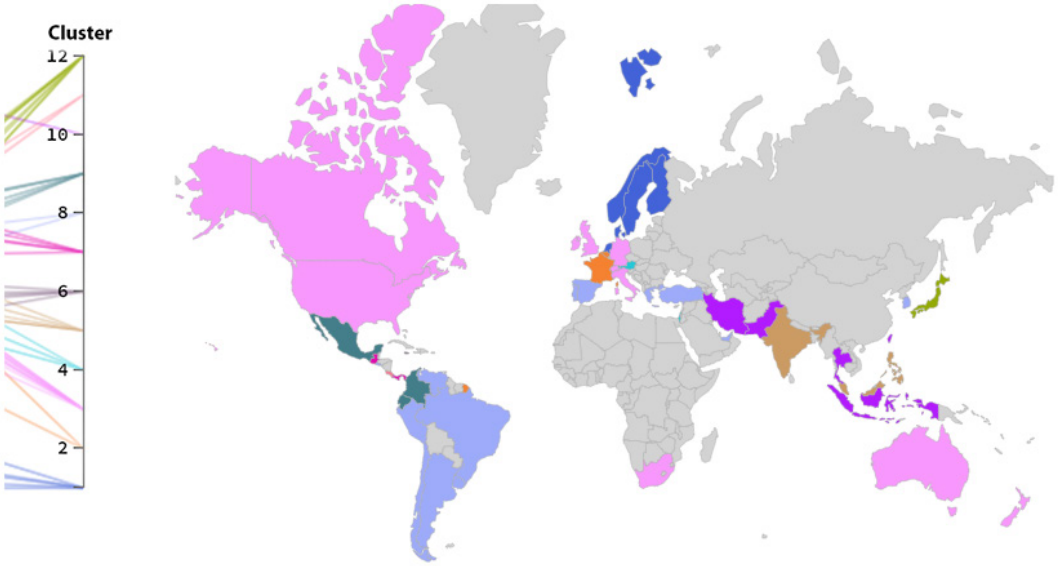
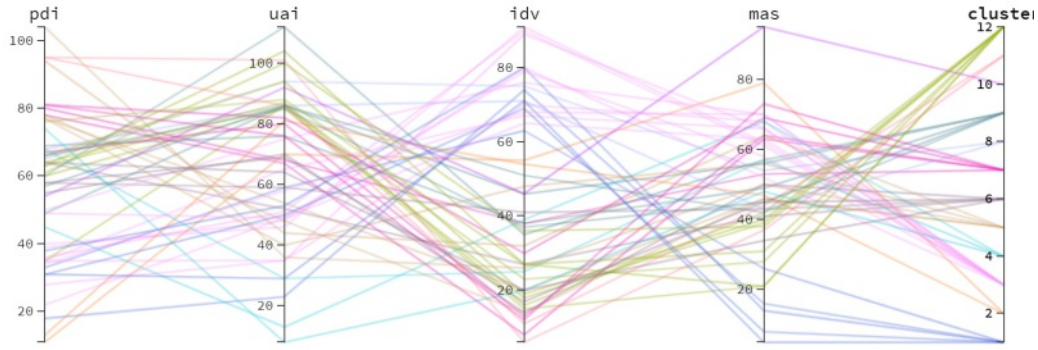


Figure 6.33: Using a world map, we show the different clusters found after applying an agglomerative hierarchical clustering analysis. The source of the data comes from Table A.2 and used in Case study 2. The colors correspond to the coloring shown in the related parallel coordinates in Figure 6.34.



country	pdi	uai	idv	mas	cluster
Africa East	64	52	27	41	6
Africa West	77	54	20	46	5

Figure 6.34: A parallel coordinate plot for Hofstede's four dimensional model. The coloring is based on the clusters, each different color represents a cluster. The source of the data comes from Table A.2 and used in Case study 2.

No.	Clusters Found by Hofstede	Clusters Found by our Visualization Framework	No.
1	Denmark, Sweden, Netherlands, Norway, and Finland	Denmark, Sweden, Netherlands, Norway, and Finland	1
2	Pakistan, Iran, Indonesia, Thailand and Taiwan, East, and West Africa	Pakistan, Iran, Indonesia, Thailand and Taiwan, East, and West Africa	2
3	Ecuador, Venezuela, Colombia, and Mexico	Ecuador, Venezuela, Colombia, and Mexico	3
4	Austria and Israel	Austria and Israel	4
5	Belgium and France	Belgium and France	5
6	Japan	Japan	6
7	Malaysia, Philippines, India, Hong Kong, Singapore, and Jamaica	Hong Kong, Singapore, and Jamaica	7
		Malaysia, Philippines, and India	8
8	Guatemala, Panama, and Costa Rica	Guatemala and Panama	9
		Costa Rica	10
9	Yugoslavia, Turkey, Arabic-speaking countries, Greece, Argentina, Spain, and Brazil	Yugoslavia, Turkey, Arabic-speaking countries, Greece, Argentina, Spain, Brazil, Korea, Peru, Salvador, Chile, Portugal, and Uruguay	11
10	Korea, Peru, Salvador, Chile, Portugal, and Uruguay		
11	Australia, United States, Canada, Great Britain, Ireland, and New Zealand	Australia, United States, Canada, Great Britain, Ireland, New Zealand, Germany, Switzerland, South Africa, Italy	12
12	Germany, Switzerland, South Africa, and Italy		

Table 6.1: Comparison between clustering groups established by our visualization framework for four dimensions of the Hofstede model vs. and Hofstede's cluster analysis on four dimensional data. We used the data in Table A.2 and applied an agglomerative clustering algorithm with the average linkage method.

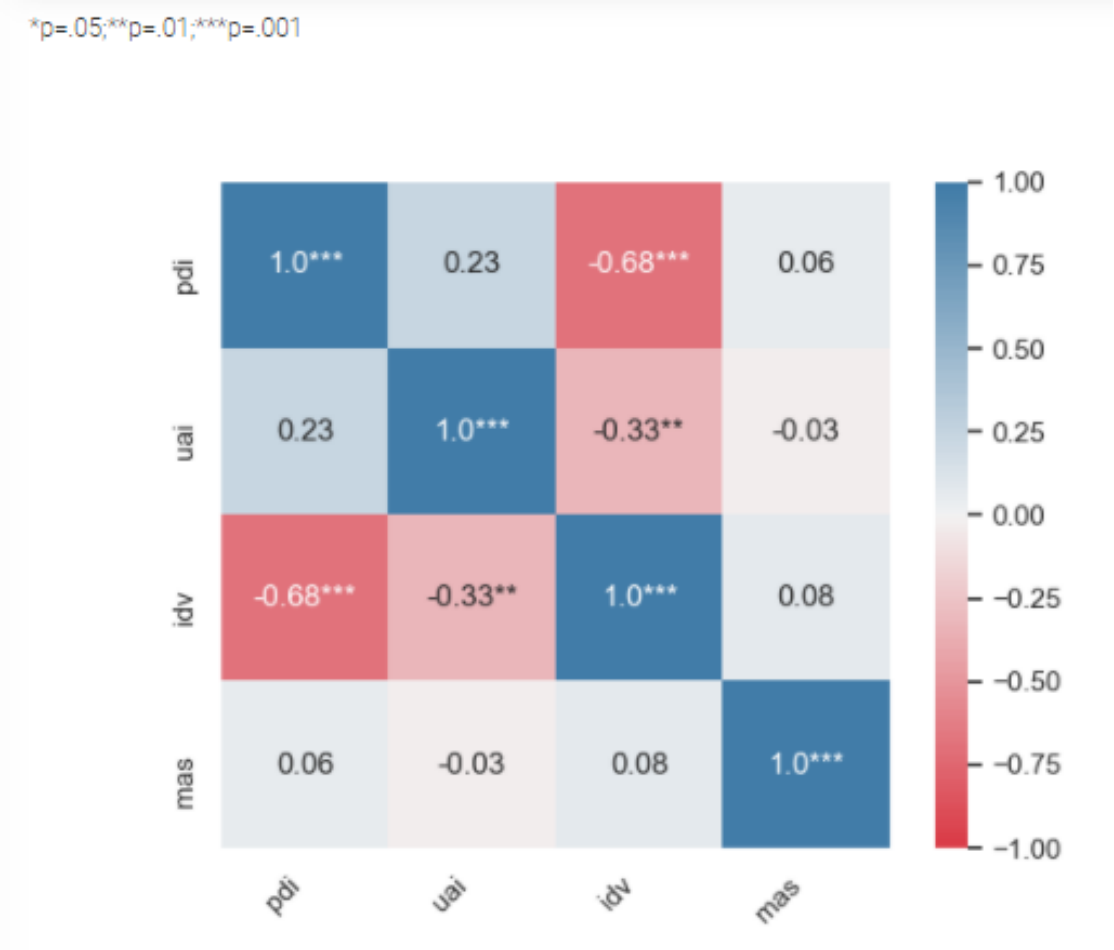


Figure 6.35: Pearson's correlation matrix for four dimensions of Hofstede model showing a negative correlation between IDV and PDI (ρ=-.68, statistically significant at 0.001), and IDV and UAI (ρ=-.33, statistically significant at 0.01). The source of the data comes from Table A.2 and used in Case study 2.

6.2.2 Clustering Performed on Six Dimensions

Using the implemented visualization framework, we first have acquired and loaded the complete dataset from Hofstede's website [45]. The full dataset, shown in Table A.3, contains scores for six dimensions for 110 countries. However, it contains missing data. A preview of the pattern of missing data can be seen in Figure 6.36, as discussed in Section 5.5.2.

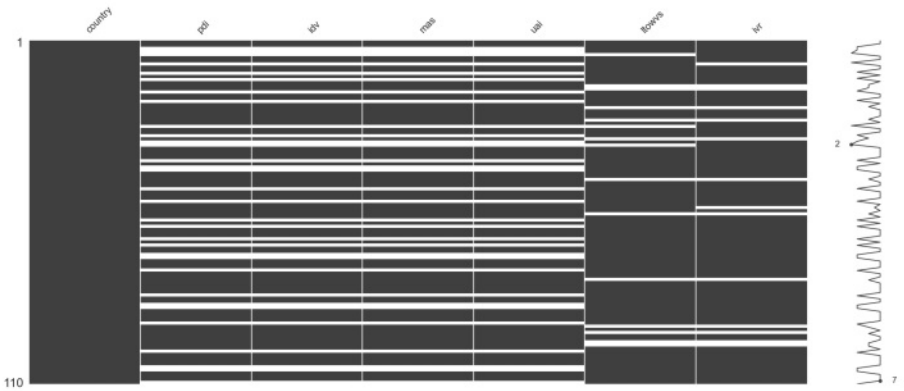


Figure 6.36: Heatmap showing the pattern of missing data in the six dimensional dataset of Hofstede. The source of the data comes from Table A.3 and used in Case study 2. Since there are white lines in the heatmap, a portion of the data is missing.

The dataset has 110 rows, with 45 of them having missing values in at least one dimension. Since Hofstede did not apply any imputation method to the data, we choose deletion as our imputation method, which deletes missing values list-wise.

There are different clustering alternatives within our visualization framework, as discussed in Section 5.5.3. To replicate the study conducted by Hofstede, we choose hierarchical clustering, using an average linkage method. Once the appropriate clustering method is chosen, it is possible to preview the clustering in a dendrogram, as illustrated in Figure 6.37.

It is possible to set a specific distance, where the dendrogram gets cut by filling the max_d parameter value. The number of clusters is reflected by the crossings of the horizontal line with the dendrogram branches. If we set the max_d parameter to 52, the result is 12 clusters as seen in Figure 6.37. The reason behind cutting the tree at this height is to replicate the 12 clusters found by Hofstede. Members of each cluster can be found in Table 6.2.

Reviewing the clusters created by our visualization framework, we can identify a clear linguistic area in cluster 1 (Australia, Canada, Great Britain, Ireland, New Zealand, and the U.S.A.) which reflects an English speaking cluster. Nordic countries are reflected in cluster 2 (Denmark, Finland, Netherlands, Norway, and Sweden), Baltic countries in

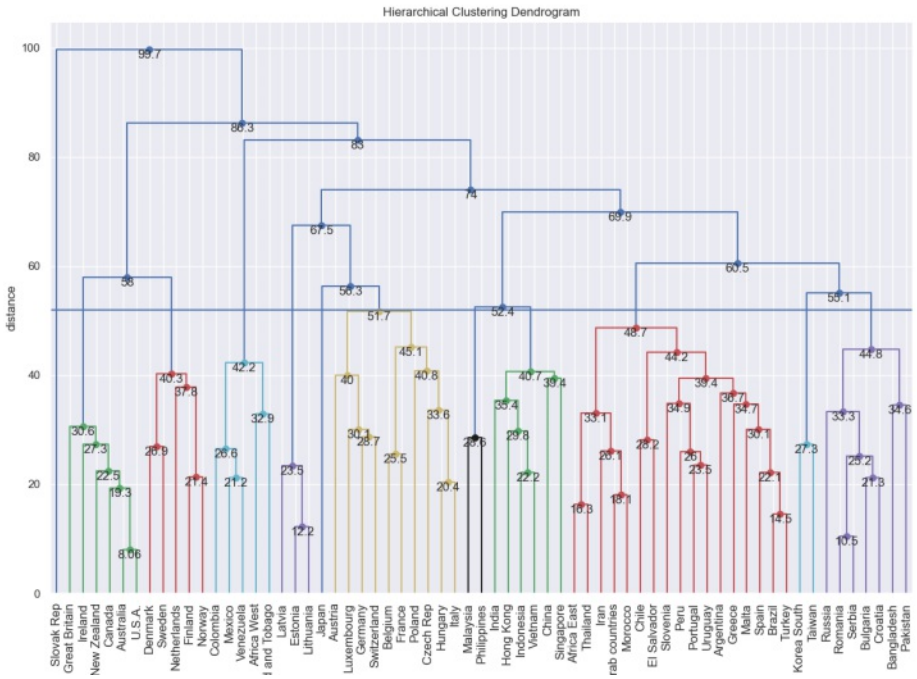


Figure 6.37: Using a dendrogram representation to visualize the effect of agglomerative hierarchical clustering in Hofstede's six dimensional model. The source of the data comes from Table A.3 and used in Case study 2: cutting the tree at height of 52, to reflect 12 clusters.

cluster 4 (Estonia, Latvia, and Lithuania) and Central European countries in cluster 5 (Austria, Belgium, Czech Republic, France, Germany, Hungary, Italy, Luxembourg, Poland, and Switzerland). An Eastern-Asian cluster appears in clusters 7 (Malaysia and Philippines), 8 (China, Hong Kong, India, Indonesia, Singapore, and Vietnam), and 10 (Korea South and Taiwan).

This pattern is also visible when visualizing the clusters on a world map, as demonstrated in Figure 6.38. The coloring of each country is according to its color when the cluster axis on the parallel coordinates is selected (Figure 6.39). Additionally, hovering over each country with the mouse displays a tooltip that includes all the country's related information. Figure 6.38, the tooltip for Canada, shows each dimension's information and the cluster where the country belongs. The missing countries are greyed out.

Figure 6.39 shows the parallel coordinates, color-coded by cluster value. The user has the ability to apply *brushing* on this axis and *filter* the values of each cluster and view the respective dimensions of the members.

Based on the clustering shown in Table 6.2, the parallel coordinates (Figure 6.39, and the

No. Cluster	Members of the Cluster
1	Australia, Canada, Great Britain, Ireland, New Zealand, U.S.A.
2	Denmark, Finland, Netherlands, Norway, Sweden
3	Africa West, Colombia, Mexico, Trinidad, and Tobago, Venezuela
4	Estonia, Latvia and Lithuania
5	Austria, Belgium, Czech Rep, France, Germany, Hungary, Italy, Luxembourg, Poland and Switzerland
6	Japan
7	Malaysia and Philippines
8	China, Hong Kong, India, Indonesia, Singapore, and Vietnam
9	Africa East, Arab countries, Argentina, Brazil, Chile, El Salvador, Greece, Iran, Malta, Morocco, Peru, Portugal, Slovenia, Spain, Thailand, Turkey, and Uruguay
10	Korea South and Taiwan
11	Bangladesh, Bulgaria, Croatia, Pakistan, Romania, Russia, and Serbia
12	Slovakia

Table 6.2: Clustering groups established by our visualization framework on Hofstede's six dimensional model. We used the data in Table A.3 and applied an agglomerative clustering algorithm with an average linkage method.

world map (Figure 6.38) we can discover following pattern after we clustered Hofstede's six dimensional model:

Knowledge 3: After applying agglomerative hierarchical clustering on six dimensions (PDI, UAI, IDV, LTO, IVR, and MAS) of the Hofstede model, we grouped countries into 12 clusters. We identified linguistic clusters in English-speaking and Nordic (Clusters 1 and 2) regions, and geographical clusters in the European region (cluster 4 and 5), and in Eastern-Asia region (clusters 7,8 and 10).

The Pearson's correlation matrix created by our visualization framework in Figure 6.40 indicates a clear negative correlation between PDI and IDV (ρ=-0.65), which is statistically significant at 0.001. This is similar to the result found by Hofstede, who found PDI and IDV to be negatively correlated (ρ=-.68) at a significance level of 0.001 as shown in Figure 6.30. The IVR dimension has a negative correlation (ρ=-0.5) with LTOWVS (LTO) dimension statistically significant at 0.001 and a negative correlation with PDI (ρ=-0.28) statistically significant at 0.01.

Comparing the Pearson's correlation matrix in this section, with the Pearson's correlation matrix from Section 6.2.1 seen in Figure 6.35, it is evident that increasing the model's dimension from four to two changed the correlation within dimensions. IDV and PDI, which are still negatively correlated, IDV and UAI do not show any correlation in the six dimensional model. Thus, we can summarize the knowledge discovery in this section as

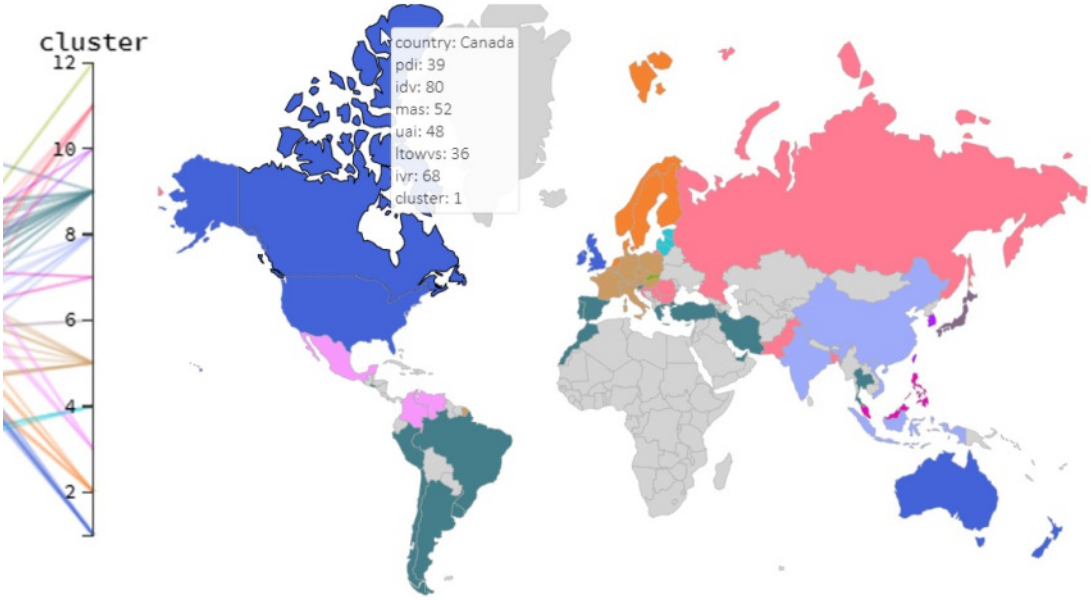
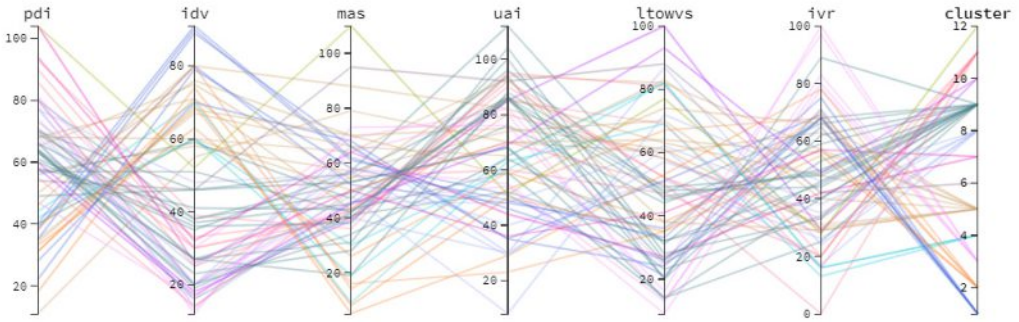


Figure 6.38: Visualization of clusters on the world map for the six dimensions of Hofstede model after agglomerative clustering with average linkage method was applied. Countries having the same color belong to the same cluster. A tooltip shows the information on each entry while hovering over the map. The source of the data comes from Table A.3 and used in Case study 2.

follows:

Knowledge 4: Adding the two dimensions of LTOWVS (LTO) and IVR into the four dimensional models of Hofstede (PDI, UAI, IDV, and MAS) does not affect the negative correlation between PDI and IDV. However, it removes the negative correlation between IDV and UAI dimension and adds a negative correlation between IVR and LTOWVS; and between IVR and PDI.



country	pdi	idv	mas	uai	ltowvs	ivr
Africa East	64	27	41	52	32	40
Africa West	77	20	46	54	9	78

Figure 6.39: Visualization of parallel coordinates for six dimensions of Hofstede model, where the cluster axis is selected. The coloring of the parallel coordinates is based on the coloring of the clusters. The source of the data comes from Table A.3 and used in Case study 2.

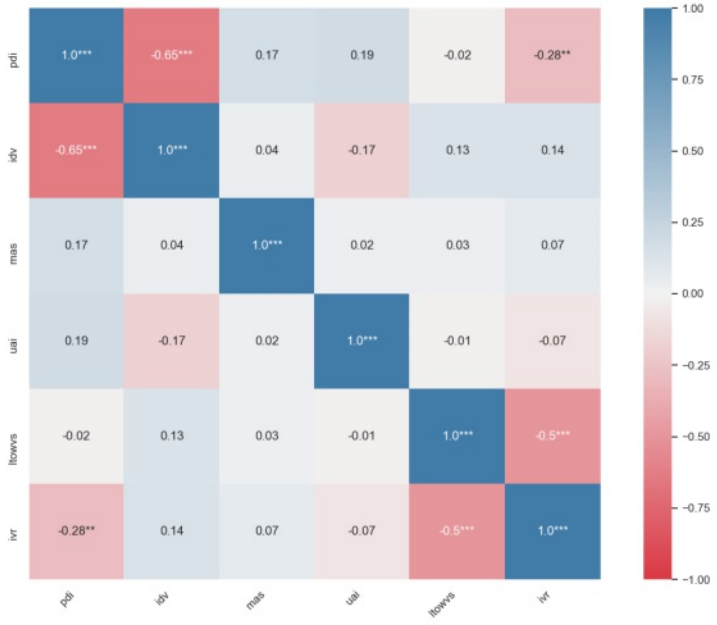


Figure 6.40: Pearson's correlation matrix, as generated for the six dimensions of the Hofstede model. The source of the data comes from Table A.3 and used in Case study 2.

6.3 Case Study 3: Knowledge Discovery and Beyond

This section aims to fully utilize every functionality available in the visualization framework for the purpose of knowledge discovery in the cultural science domain. For this free exploration and analysis, we extend the Hofstede model with additional dimensions and explore the available data, using all the offered visualization functionalities in the visualization framework.

Erman and Medeiros [5], alongside with numerous other researches [188, 189, 190], explored recently the correlation between cultural dimensions and different aspects of Covid-19, such as mortality rate and social distancing. Erman and Medeiros [5] attempted to use the six dimensions of the Hofstede model to investigate whether the mortality rate of the Covid-19 pandemic can be identified as an independent predictor of Covid-19 fatalities. The author used data from 49 countries with adequate health information and system capacity and prominent Covid-19 epidemics. The data was extracted from sources, such as the World Bank [191], the World Health Organization (WHO) [192], Organisation for Economic Co-operation and Development (OECD) [193] and Hofstede's website [45]. Any missing data were imputed using the iterative multiple imputation method, and no cluster analysis was performed in this study. Using a bootstrap variable selection approach, Erman and Medeiros [5] shows that individualism and uncertainty avoidance are independent predictors of Covid-19 fatalities. The author used case fatality rate and mortality rate per thousand population as two dependant variables to perform a meta-regression analysis on economic indicators (i.e., gross domestic product (GDP) per capita in 2018), demographics (i.e., age distribution), the extent of SARS-CoV-2 testing coverage, differential timing of the outbreak (i.e., days since first death on record), indicators of health system capacity (i.e., numbers of healthcare workers and hospital beds per 1,000 population) and pertinent cultural dimensions for each nation as defined by Hofstede.

Inspired by this idea of combining Covid-19 epidemiological data with Hofstede's cultural model, we extend the model by two extra dimensions of Covid-19 mortality rate in confirmed cases and the quantity of death per 100,000 people. We acquired the data from the Johns Hopkins Coronavirus Resource Center [194]. Our goal is to explore this new model and identify patterns or correlations and not only reproduce the exact results found in Erman and Medeiros [5]. Thus, we generate and accept new hypotheses after the data is loaded and the knowledge discovery is started.

Figure 6.41 shows a preview of the data before choosing an imputation method. The next step is to apply an iterative imputation.

6.3.1 Clustering Exploration and Analysis

Once the imputation is applied, we apply clustering to reveal groups of countries with similar characteristics regarding their cultural aspects and how the pandemic has been handled. We first preview the elbow plot of the k-nearest neighbors clustering. In this

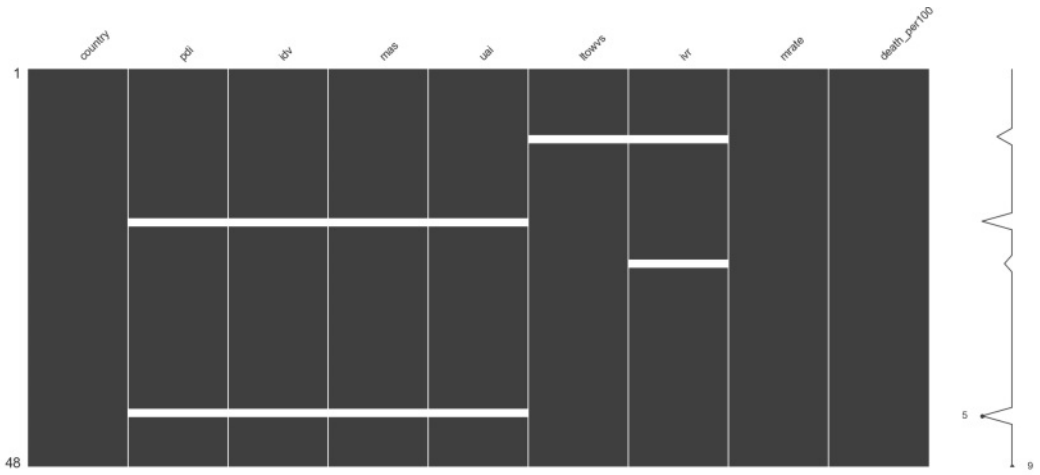


Figure 6.41: Pattern of missing data in the dataset shown in Table A.4 represented by a heatmap. Missing entries are marked with white color. Rows indicate countries, while columns indicate a cultural dimension. The dataset contains eight dimensions, six are the standard dimensions of the Hofstede mode, and two are added by us which is the Covid-19 mortality rate and death per 100,000 population.

case, the elbow plot in Figure 6.42 does not indicate any quick increase, and for this reason, the k-nearest neighbors clustering method cannot be used.

Previewing the hierarchical clustering impact with complete linkage method results in a dendrogram shown in Figure 6.43. Altering the maximum height (max_d) parameter to cut the dendrogram in different heights (and also partitions of the examined countries set) results in different values of the Silhouette Coefficient [54]. The higher the Silhouette Coefficient, the better is the split within the cluster. For example, setting the height to 98 would result in eight country clusters, with a Silhouette Coefficient rate of 0.19 (Figure 6.43), while setting its value to 105 results in seven country clusters with a Silhouette Coefficient rate of 0.23 (Figure 6.44) and setting its value to 115 results in six clusters with a Silhouette Coefficient of 0.22 (Figure 6.45). A height of 105 with the highest Silhouette Coefficient value is selected and applied to the data resulting in seven country clusters. The full detail on each cluster's content can be found in Table 6.3.

Representing the resulting clusters in the World Map indicates the geographical proximity of these clusters. The countries in clusters 1 and 7 are geographically nearby to each other—similar to cluster 5. In Figure 6.46, cluster 1 (dark blue) is mainly within Europe, cluster 7 (magenta) reflects South East Asian countries, and cluster 5 (dark beige) reflect East European countries and Russia. These colors are propagated to the parallel coordinates (Figure 6.47). The user also has the ability to hover over each country and see its full details on a tooltip box, as illustrated in Figure 6.48 for South Korea, where the country's name and the value of each dimension are shown. Grey colored countries are those for which we do not have any data in the dataset; no tooltip box is shown while

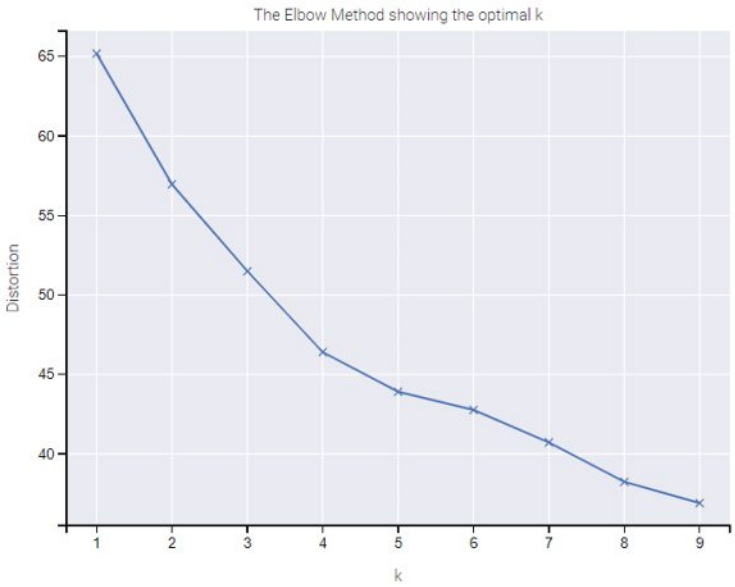


Figure 6.42: Elbow plot of k-nearest neighbors clustering method in an eight dimensional dataset (six standard dimension of the Hofstede model combined with two dimensions from Covid-19 epidemiological data) used for case study 3 (data source is Table A.4). A clear "elbow" cannot be identified in this figure; thus, this clustering algorithm is not suitable for this dataset.

hovering over those countries.

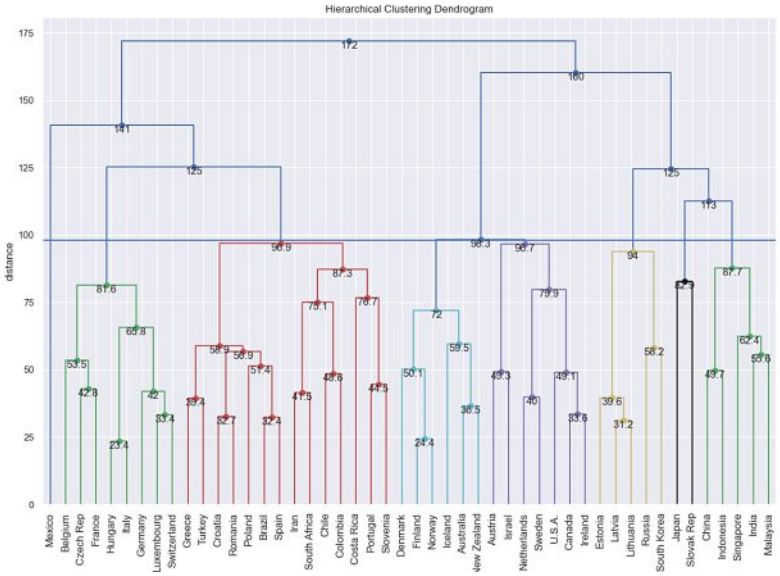


Figure 6.43: Cutting the dendrogram at a height of 98 (Silhouette Coefficient 0.19), resulting in eight clusters. This is to preview the impact of an agglomerative clustering method applied on in an eight dimensional dataset used for case study 3 (data source is Table A.4).

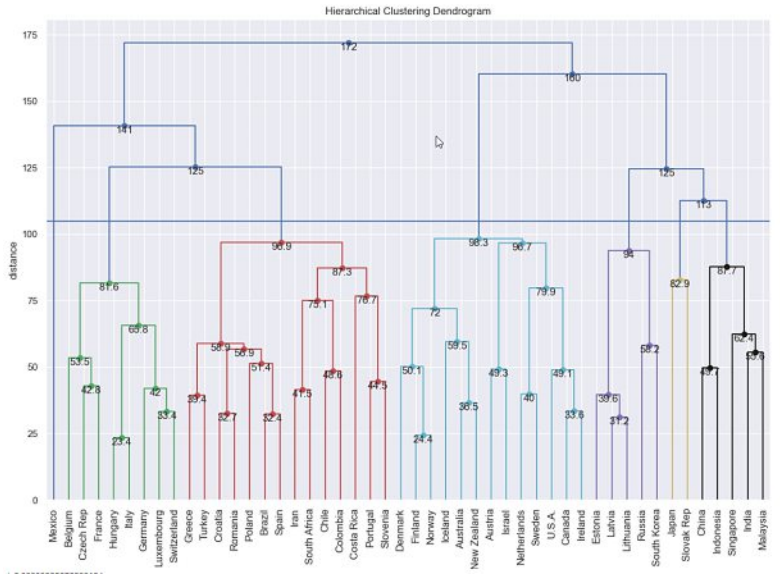


Figure 6.44: Cutting the dendrogram at a height of 105 (Silhouette Coefficient 0.23), resulting in seven clusters. This is to preview the impact of an agglomerative clustering method applied on in an eight dimensional dataset used for case study 3 (data source is Table A.4).

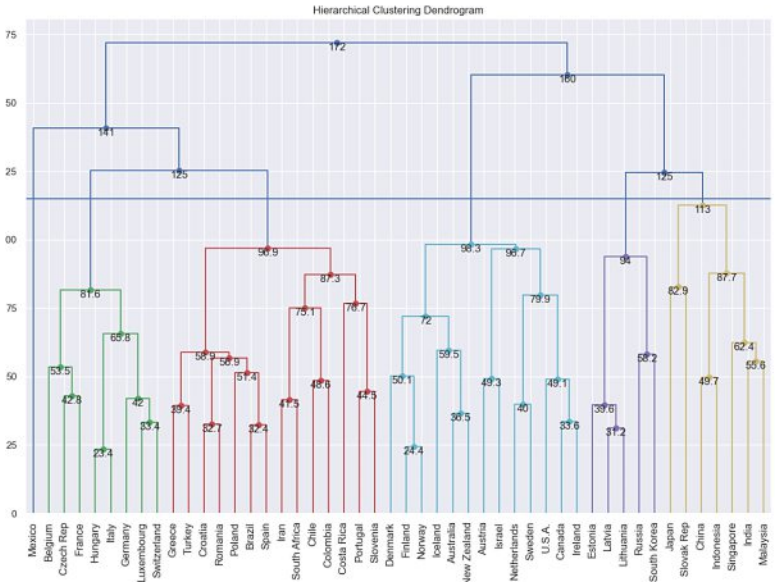


Figure 6.45: Cutting the dendrogram at a height of 115 (Silhouette Coefficient 0.22), resulting in six clusters. This is to preview the impact of an agglomerative clustering method applied on in an eight dimensional dataset used for case study 3 (data source is Table A.4).

No. Clusters	Members
1	Belgium, Czech Rep, France, Germany, Hungary, Italy, Luxembourg, Switzerland
2	Brazil, Chile, Colombia, Costa Rica, Croatia, Greece, Iran, Poland, Portugal, Romania, Slovenia, South Africa, Spain, Turkey
3	Mexico
4	Australia, Austria, Canada, Denmark, Finland, Iceland, Ireland, Israel, Netherlands, New Zealand, Norway, Sweden, U.S.A
5	Estonia, Latvia, Lithuania, Russia, South Korea
6	Japan, Slovak Rep
7	China, India, Indonesia, Malaysia, Singapore

Table 6.3: Clusters resulting from hierarchical clustering with complete linkage applied to an eight dimensional dataset (six standard dimension of the Hofstede model combined with two dimensions from Covid-19 epidemiological data) used for case study 3 (data source is Table A.4).

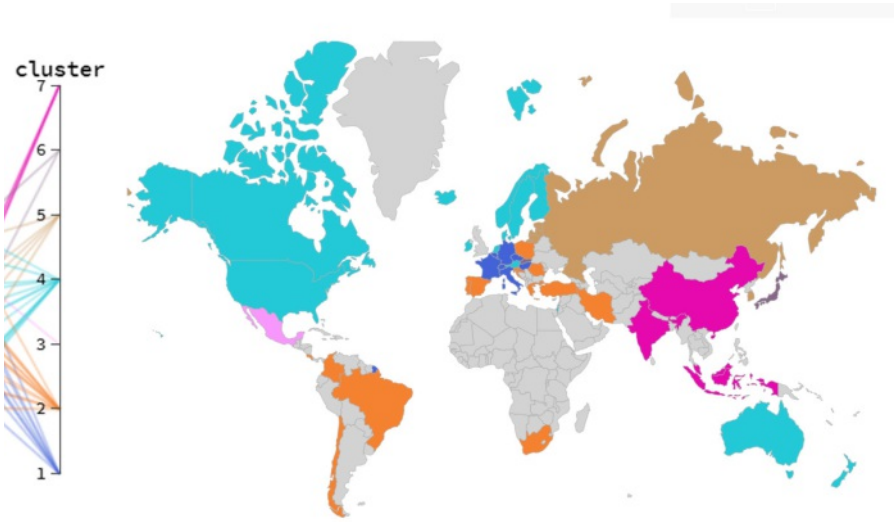


Figure 6.46: World map to reflect the clustering of countries, with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Countries with the same color belong to same cluster.

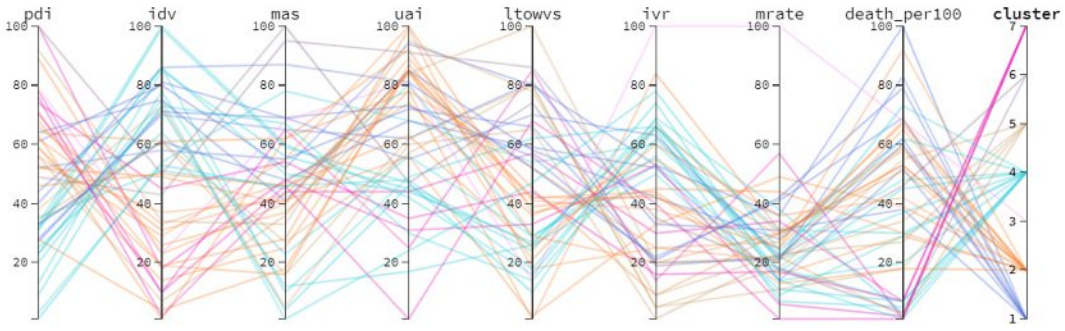


Figure 6.47: Parallel coordinates plot that reflects the clustering of countries, with regard to the Hofstede model and Covid-19 pandemic measurements used for case study 3. Countries with the same color belong to same cluster.

Figure 6.48: Demonstration of a tooltip, appearing when hovering the mouse over a country (South Korea) to reveal the dimensions of the Hofstede model and Covid-19 pandemic measurements.

6.3.2 Mortality Rate (*mrate*) Exploration and Analysis

After the data has been grouped, we investigate the two newly added dimensions of mortality rate (*mrate*) and death per 100,000 population (*death_per*100*k*). We also investigate each clustering group to identify interesting trends or patterns within each country cluster. This can be achieved using the parallel coordinates, combined with the world map and the radar chart.

Figure 6.49 shows the parallel coordinates, where the top five countries with the highest mortality rate (i.e., Mexico, China, Iran, Italy, and Greece) have been selected. The geography, culture, and language of these countries are different, and they do not show any specific trend—except for all of them being in the northern hemisphere (Figure 6.50). Except for Italy, the rest of the countries have unequally distributed power (high PDI value) and are collectivist countries (low IDV value). Except for Iran, all countries are masculine countries (high MAS), and except for China, all countries feel threatened by uncertain situations (high UAI). Therefore, there is no clear pattern within these five countries concerning the six Hofstede model dimensions.

Based on the parallel coordinates showing high mortality rate (Figure 6.49), and the observations we did within the top five countries with the highest mortality rate, we generate two new hypotheses which we investigate in this section further:

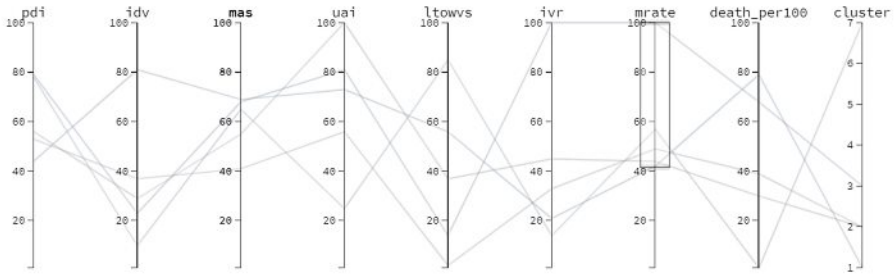
H3.1 The higher the Covid-19's mortality rate of a country is, the higher is the uncertainty avoidance in the country.
H3.2 The higher the Covid-19's mortality rate of a country is, the higher is the more masculine the country is.

Figure 6.51 shows that Singapore, Malaysia, Iceland, Israel, and Norway have the lowest mortality rate for Covid-19. All five countries are feminine (low MAS score) and are comfortable with uncertainty (low UAI score), except for Israel. Iceland and Norway, Malaysia, and Singapore are geographically close to each other. However, Israel has no graphical or cultural connection to any of the other countries, as seen in Figure 6.52.

Comparing the population of countries with the highest mortality rate, i.e., Mexico (127.6 million), China (1.398 billion), Iran (82.91 million), Italy (60.36 million), and Greece (10.72 million), to the population of countries with lowest mortality rate, i.e., Singapore (5.704 million), Malaysia (31.95 million), Iceland (356,991), Israel (9.053 million) and Norway (5.328 million), may suggest that population and mortality rate could potentially have a positive correlation. This knowledge discovery can be formed into a new hypothesis:

H3.3 The higher the population of a country, the higher is the Covid-19 mortality rate.

We investigate this newly formulated hypothesis by adding the population value to our visualization framework as a new dimension. The results are discussed in Section 6.3.5.1.



country	pdi	idv	mas	uai	ltowvs	ivr	mrate
China	78	10	65	25	85	14	57
Greece	56	29	55	100	37	45	44
Iran	53	37	41	56	2	33	49
Italy	44	81	69	73	56	21	42
Mexico	79	23	68	81	14	100	100

Figure 6.49: Parallel coordinates for the top five countries, with highest mortality rate in an eight dimensional dataset (six standard dimension of the Hofstede model combined with two dimensions from Covid-19 epidemiological data) used for case study 3 (data source is Table A.4). This is to investigate countries with high mortality rate. China, Greece, Iran, Italy and Mexico have the highest mortality rate of Covid-19.

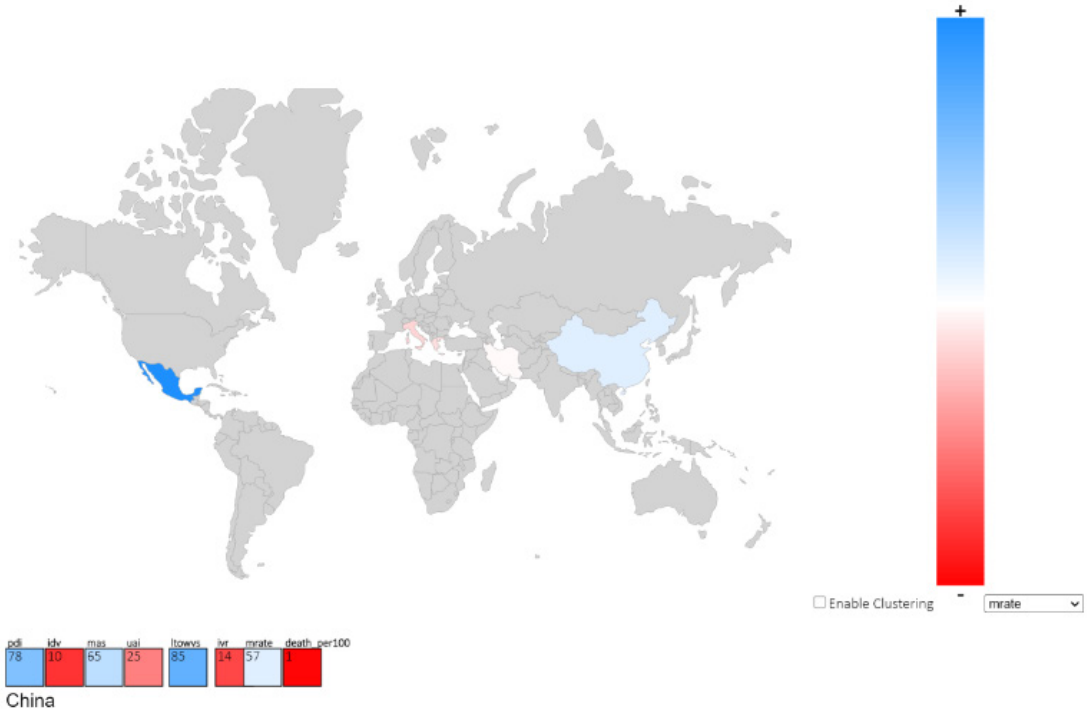
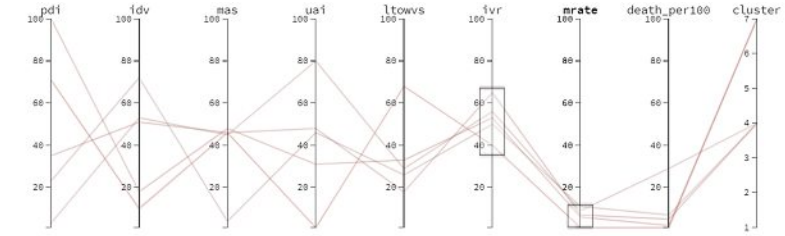


Figure 6.50: World map for the top five countries with highest mortality rate used for case study 3 (data source is Table A.4). China, Greece, Iran, Italy and Mexico jave the highest mortality rate of Covid-19, which are marked with blue color on the world map.



country	pdi	idv	mas	uai	ltowvs	ivr	mrate	death_per100
Iceland	35	51	46	48	18	65	7	5
Israel	3	53	45	80	29	56	9	29
Malaysia	100	10	48	31	33	53	6	2
Norway	23	72	4	46	26	50	11	7
Singapore	71	10	46	1	68	40	1	1

Figure 6.51: Parallel coordinates for the five countries, with least mortality rate in an eight dimensional dataset (six standard dimension of the Hofstede model combined with two dimensions from Covid-19 epidemiological data) used for case study 3 (data source is Table A.4). Iceland, Israel, Malaysia, Norway, and Singapore have the lowest mortality rate of Covid-19.

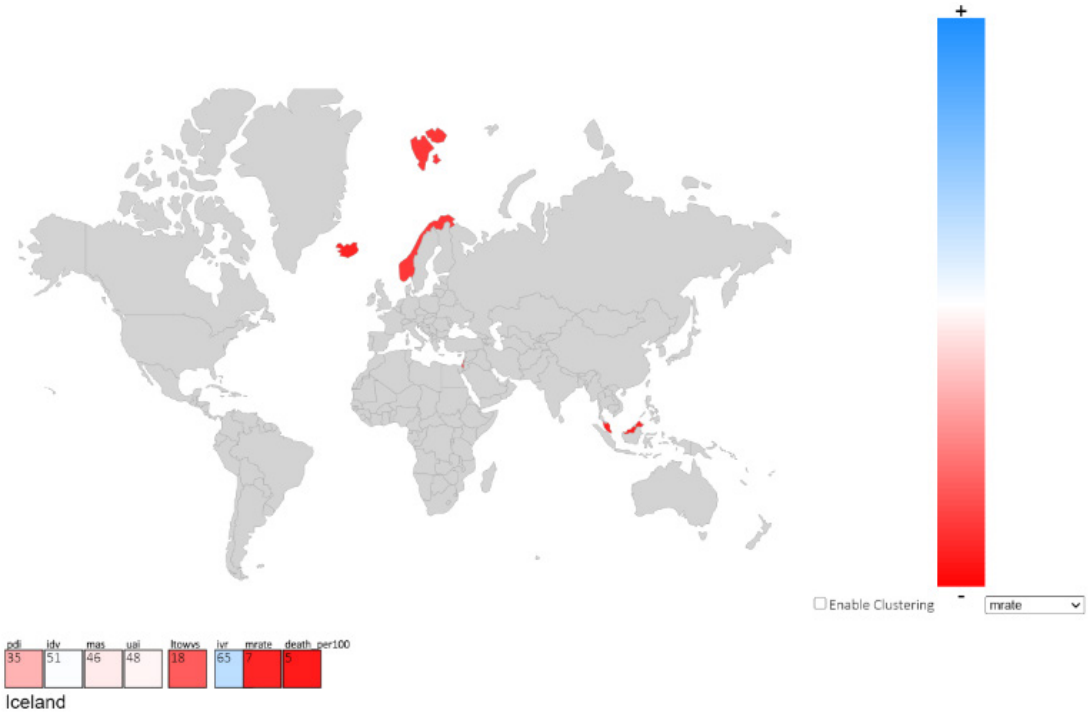
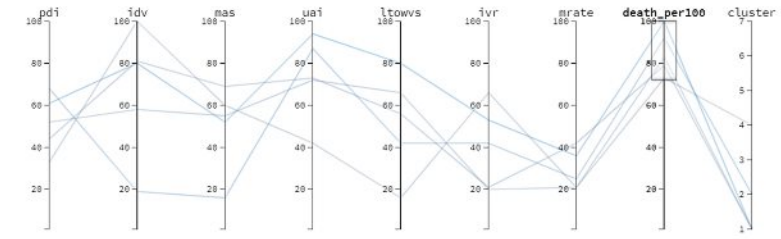


Figure 6.52: World map for the top five countries with lowest mortality rate used for case study 3 (data source is Table A.4). Iceland, Israel, Malaysia, Norway, and Singapore have the highest mortality rate of Covid-19, which are marked with blue color on the world map.

6.3.3 Death per 100,000 people (*death_per100k*) Exploration and Analysis

If we *filter* the top five countries with the highest death rate by brushing the *death_per100k* axis on the parallel coordinates (Figure 6.53), we identify Belgium, Slovenia, Czech Republic, Italy, and the U.S.A as those countries with the highest values. Except for Slovenia, the countries are masculine (high value in MAS) with an individualist attitude (high IDV score). All countries, except for the U.S.A, have a high tendency to avoid uncertain situations (high UAI value). Hence, we generate the following hypotheses to investigate further:

The world map (Figure 6.54) shows that four of these countries are centered in the European region, but there is no other particular geographical indication about them.



country	pdi	idv	mas	uai	ltowvs	ivr	mrate	death_per100
Belgium	61	88	52	94	80	53	36	180
Czech Rep	52	58	55	72	66	20	21	83
Italy	44	81	69	73	56	21	42	79
Slovenia	68	19	16	87	42	42	25	92
U.S.A.	33	100	60	42	16	66	21	73

Figure 6.53: Parallel coordinates with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Here we filter the top five countries with highest death rate per 100,000 population which are Belgium, Slovenia, Czech Republic, Italy, and the U.S.A.

As seen in Figure 6.55, after *filtering* the parallel coordinates to include only countries with low death rates, Singapore, New Zealand, China, Malaysia, and South Korea have the lowest death rates per 100,000 population. Except for New Zealand, all countries have a high PDI and low IDV value and are geographically close to each other. Except for South Korea, the countries are comfortable with uncertainty (low UAI score). We generate a new hypothesis for the UAI dimension:

H3.4 The higher the Covid-19 death rate per 100,000 of population, the higher is the uncertainty avoidance in the country.

6.3.4 Correlation Exploration and Analysis

Exploring the Pearson's correlation matrix (see Figure 6.57) to confirm our hypotheses, we see that there is a positive correlation (ρ=0.25) between mortality rate (*mrate*) and uncertainty avoidance (UAI) statistically significant at 0.05. ***Higher uncertainty***

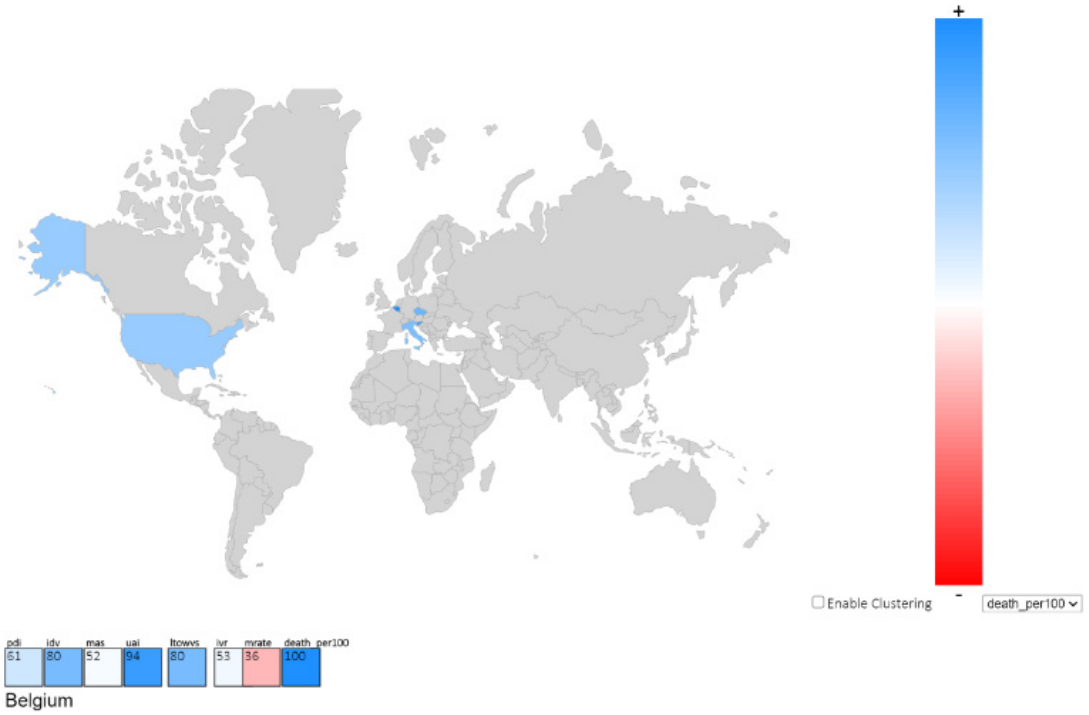
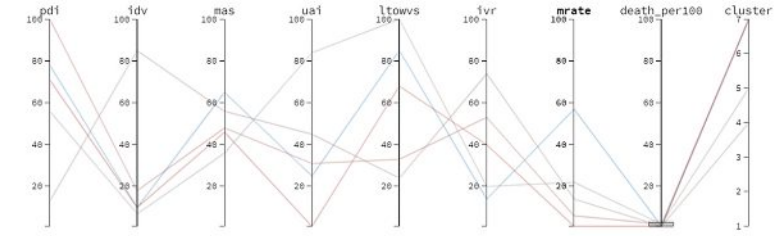


Figure 6.54: World map for the top five countries with highest death rate per 100,000 population used for case study 3 (data source is Table A.4). Belgium, Slovenia, Czech Republic, Italy, and the U.S.A have the highest death rate per 100,000 population of Covid-19, which are marked with blue color on the world map.

avoidance in a country results in a higher mortality rate for Covid-19. H3.1 is accepted. Masculinity (MAS) and mortality rate have a positive correlation (ρ=0.3) as well, statistically significant at 0.01. Additionally, a strong negative correlation (ρ=-0.59) between IVR and LTO (LTOWVS), PDI and IDV (ρ=-0.6) exists (statistically significant at 0.001). Other correlations at a significant level of 0.01 also exist between MAS and mortality rate (ρ=0.3), and between PDI and IVR (ρ=-0.36). ***The higher the masculinity level in a country is, the higher is the mortality rate for Covid-19. H3.2 is accepted.*** A strong positive correlation (ρ=0.46) between death rate per 100k population and UAI exists (statistical significant at 0.01). ***Higher uncertainty avoidance in a country results in a higher death rate per 100,000 of the population for Covid-19. H3.4 is accepted.***

A closer look at the parallel coordinates (Figure 6.58) confirms; once we *re-arrange* the axis where PDI is beside IDV (cf. A), LTOWVS beside IVR (cf. B), and UAI beside death_per100k (cf. C) that different correlation exists in the data. In A and B, there is a clear negative correlation, i.e., the high values of PDI are connected to the low values of IDV, whereas in C, there is a rather positive correlation, i.e., the majority of the high values in LTOWVS connected to the high values in IVR. By reviewing the Pearson's



country	pdi	idv	mas	uai	ltowvs	ivr	mrate	death_per100
China	78	10	65	25	85	14	57	1
Malaysia	100	18	48	31	33	53	6	2
New Zealand	13	85	56	45	24	74	14	1
Singapore	71	10	46	1	68	48	1	1
South Korea	56	7	36	84	100	28	22	2

Figure 6.55: Parallel coordinates with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Here we filter the top five countries with lowest death rate per 100,000 population which are China, Malaysia, New Zealand, Singapore, and South Korea.

correlation matrix and the parallel coordinates, following knowledge is discovered:

Knowledge 5: A positive correlation exists between the mortality rate of Covid-19 and the PDI.

6.3.5 Additional findings

So far, in Section 6.3.1 we have shown that our visualization framework is capable of supporting clustering, as well as subsequent analysis and exploration of the clustering. In Section 6.3.2 and 6.3.3 we have analyzed in detail the two newly added dimensions related to Covid-19 with regard to the Hofstede model dimensions, which resulted in some new hypothesis generation related to the geographical position of countries with Covid-19 death and mortality rate. Lastly, in Section 6.3.4, we explored the correlation across the dimensions of our extended model (Hofstede + Covid-19 dimensions) and concluded that UAI positively correlates with the death rate per 100k population. This means that countries with a high tendency to avoid uncertain situations (high values of UAI) have higher death rates per 100,000 population, and countries that are more comfortable with uncertain situations (low score of UAI) have a lower death rate per 100,000 population. While evaluating the mortality rate in Section 6.3.2, we generated a new hypothesis, that this dimension might have a positive correlation with population. In the section below we will investigate this new hypothesis.

6.3.5.1 Mortality Rate (*mrate*) and Population

We reloaded the data from Table A.5 into the visualization tool to investigate if a correlation between population and mortality rate of Covid-19 exists. Since the mortality rate is a percentage between 1 and 100, and the population is a positive integer with typically large values, a Z-score normalization was performed on the data before visualization.

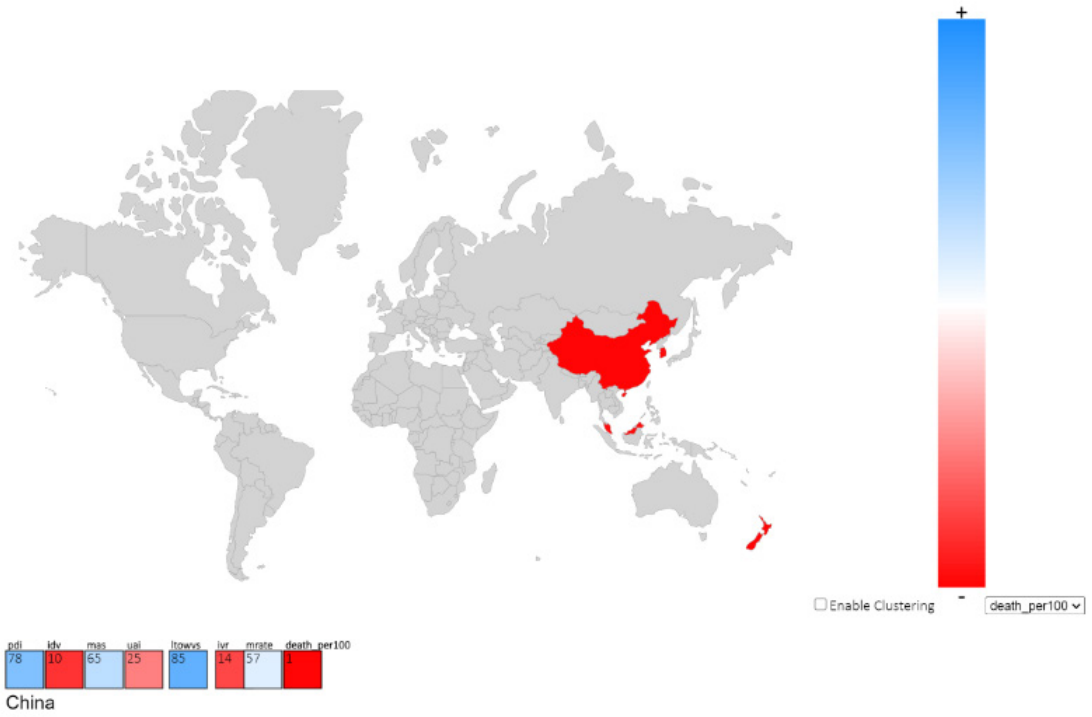


Figure 6.56: World map for the top five countries with lowest death rate per 100,000 population with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. The countries China, Malayisa, New Zealand, Singapore, and South Korea are marked with red color on the world map.

Next, we we-arranged the parallel coordinates in a way that the dimension *pop*, which stands for population, is next to mortality rate (*mrate*). This can be seen in Figure 6.59. Although there seems to be a positive correlation between these two dimensions, a closer look at the Pearson's correlation matrix in Figure 6.60, shows that this correlation is not statistically significant.

There is no correlation between mortality rate of Covid-19 and a country's population; thus H3.3 is rejected.

Furthermore, Pearson's correlation matrix shows a strong positive correlation between population with PDI (ρ=0.29) at 1% of significance level; with UAI (ρ=0.28) and with death per 100,00 population (ρ=0.26) 5% level of significance. As there was no hypothesis for any of those dimensions, we formulate them as new knowledge discoveries:

- **The more populated a country is, the higher is the power distance.**
- **The more populated a country is, the higher is the uncertainty avoidance**

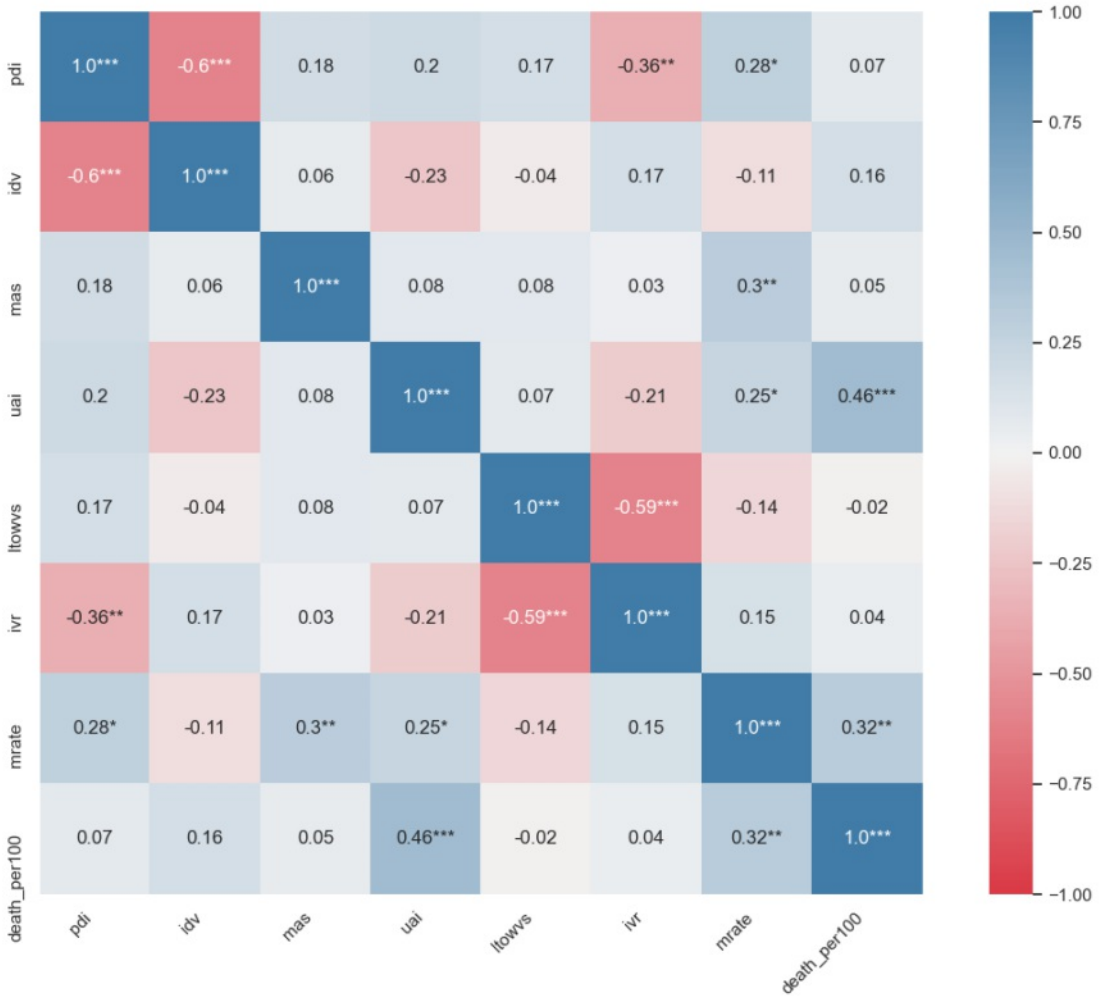


Figure 6.57: Case study 3 - Pearson's correlation matrix, after adding the two additional dimensions of mortality rate and death rate per 100,000 population to the six dimensions of the Hofstede model with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3.

- **The more populated a country is, the higher is the death ratio per 100,000 of population**

6.3.5.2 Comparing Set of Countries Using Radar Chart

Using the radar chart provided, it is possible to *lookup* a specific set of countries and *compare* the value of each of their dimensions to each other. The set of countries can either be *selected* and *changed* by using the parallel coordinates or by using the search

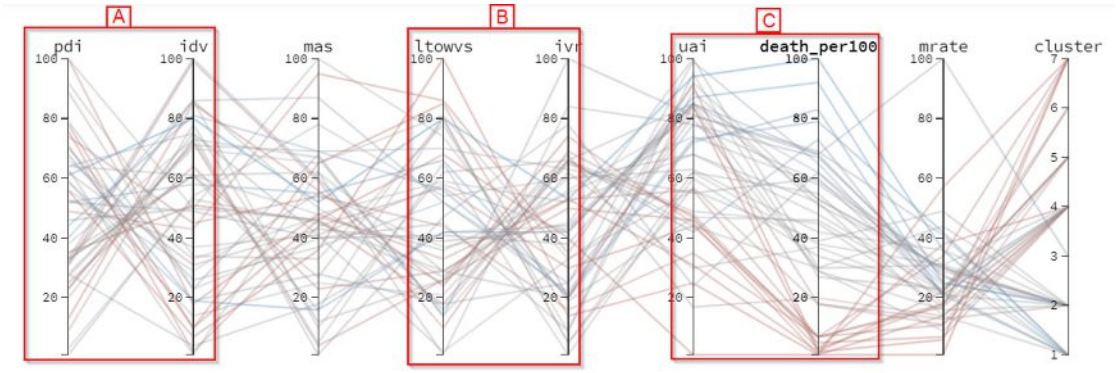


Figure 6.58: Parallel coordinates with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Rearranging the axes to investigate correlations across dimensions of the Hofstede model and the Covid-19 measurements.

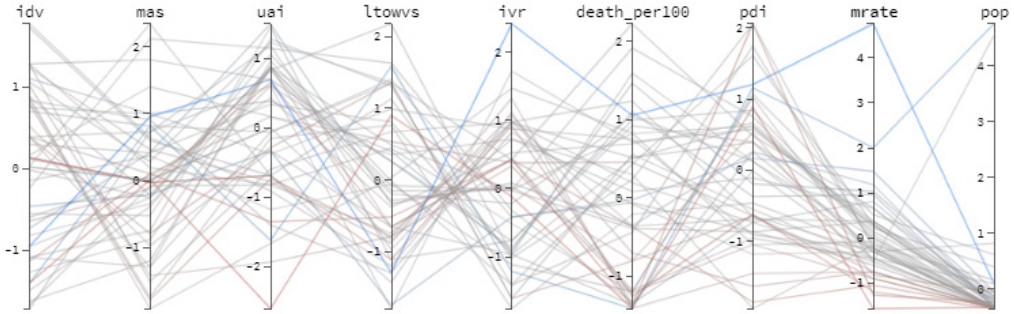


Figure 6.59: Parallel coordinates after population (*pop*) was added as a new dimension with Hodstede's standard dimensions, Covid-19 epidemiological dimension, making the model consist of 9 dimensions. The source of the data is in Table A.4 and Table A.5. We-rearranged the *mrate* axis to be beside *pop* in order to be able to investigate potential correlation between these two dimensions.

box under the radar chart. Using the brushing method on the parallel coordinates' axes causes the countries to be filtered to the selection, and these filtered countries are reflected in the radar chart. As an example, if we filter in the parallel coordinates as in Figure 6.54 the top five countries with highest death rate per 100,000 population, the result of the radar chart is as in Figure 6.61a. If we filter the parallel coordinates as in Figure 6.56, we can compare the radar chart of the top five countries with the lowest death rate per 100,000 of the population (Figure 6.61b) to the highest death rate as seen in Figure 6.61.

Alternatively, countries can be *looked up* using the search bar (cf. A) where the corresponding countries are shown on the radar chart and the legend (cf. B) if the user checks the checkbox beside the country's name. Figure 6.62 a comparison between all

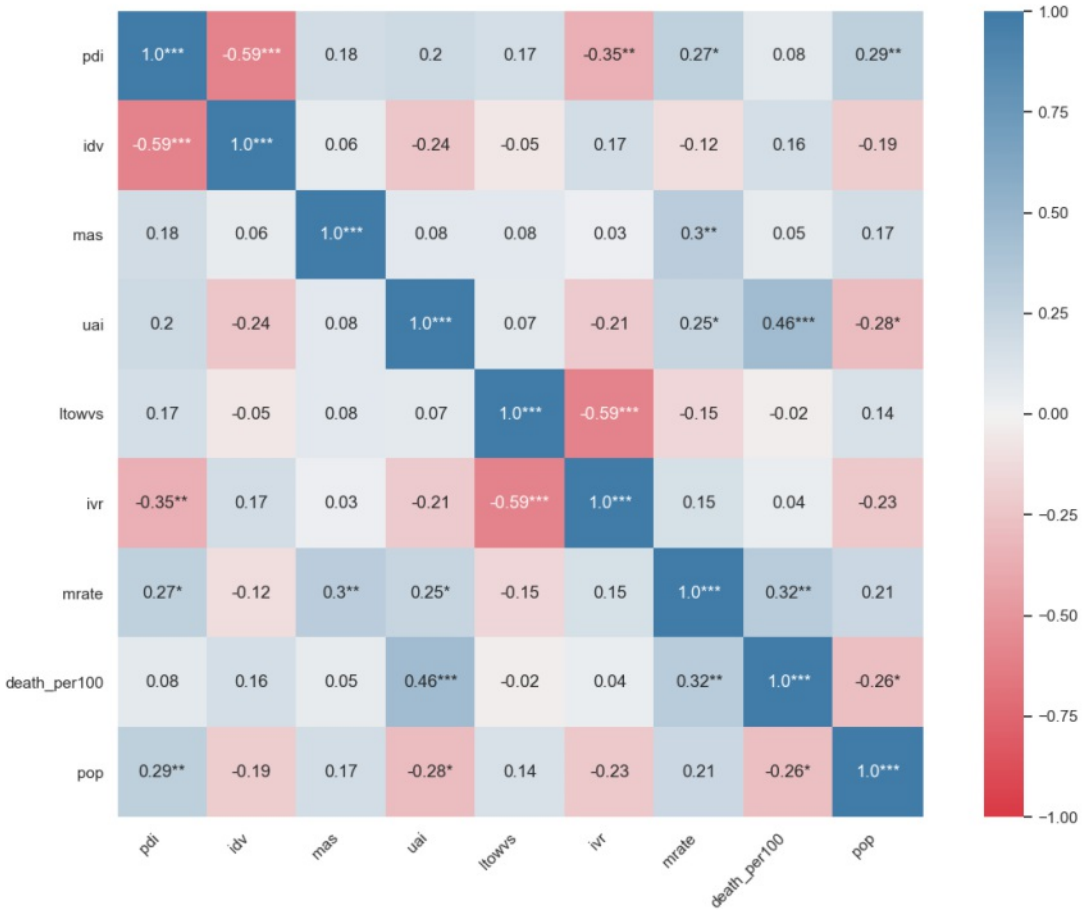


Figure 6.60: Pearson's Correlation matrix after population (*pop*) was added as a new dimension (making the model 9 dimensional). We see a positive correlation in the dimension of population with PDI, UAI and death per 100,000 of population. The source of the data is in Table A.4 and Table A.5

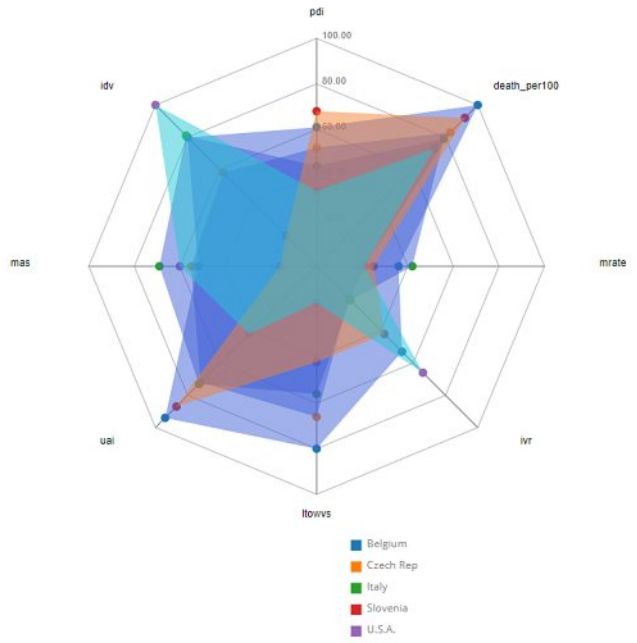
the dimensions of Australia, Austria, Belgium, Brazil, and Canada can be seen. The color of the vertices is corresponding to the colors in the legend.

6.3.5.3 Visualization of Missingness and Uncertainty

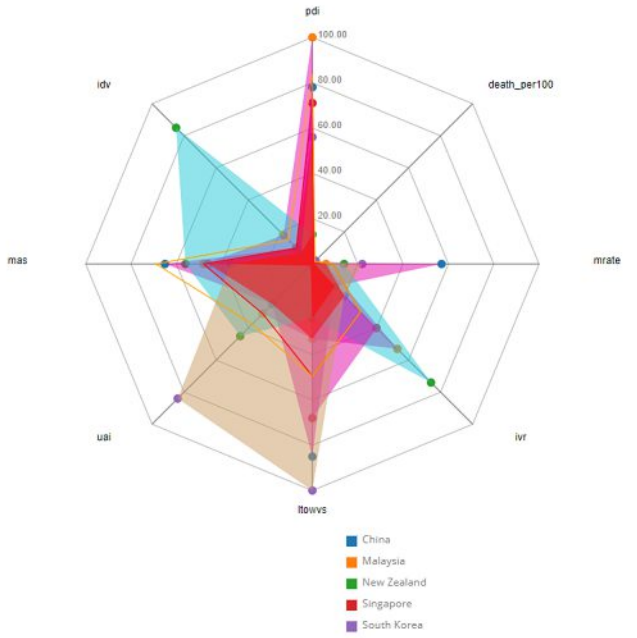
As discussed in Section 4.2.2.4, we use a clustermap to *produce* a representation. Figure 6.63 shows and eight dimensional representation of a clustermap for Hofstede's six dimension, Covid-19's mortality and death per 100,000 population. On the bottom left, the RSME value (discussed in Section 2.4.3) of each imputed dimension. This value can be used to *compare* the performance of the imputation within each dimension or between different imputation method if the user *changes* the method.

Using the iterative imputation method, the RSME of IDV is 0.73; MAS is 0.34; IVR is 1.27; PDI is 0.56, and UAI is 0.46. The remaining dimensions have no missing value. Hence the RSME will be 0 and not shown. Thus, the glyph (dot) symbol of IVR dimension is the biggest and UAI the smallest size. Additionally, the clustermap shows the dendrogram on the left side, which the user can select a range by brushing and investigate a branch closely.

Hovering the mouse over each cell in the heatmap shows the country name and value of the dimension. In Figure 6.64 we zoomed into a section of the same clustermap shown in Figure 6.63. A closer look shows that Israel has a missing value in the IVR dimension, and Iceland has missing values in MAS and PDI dimension.

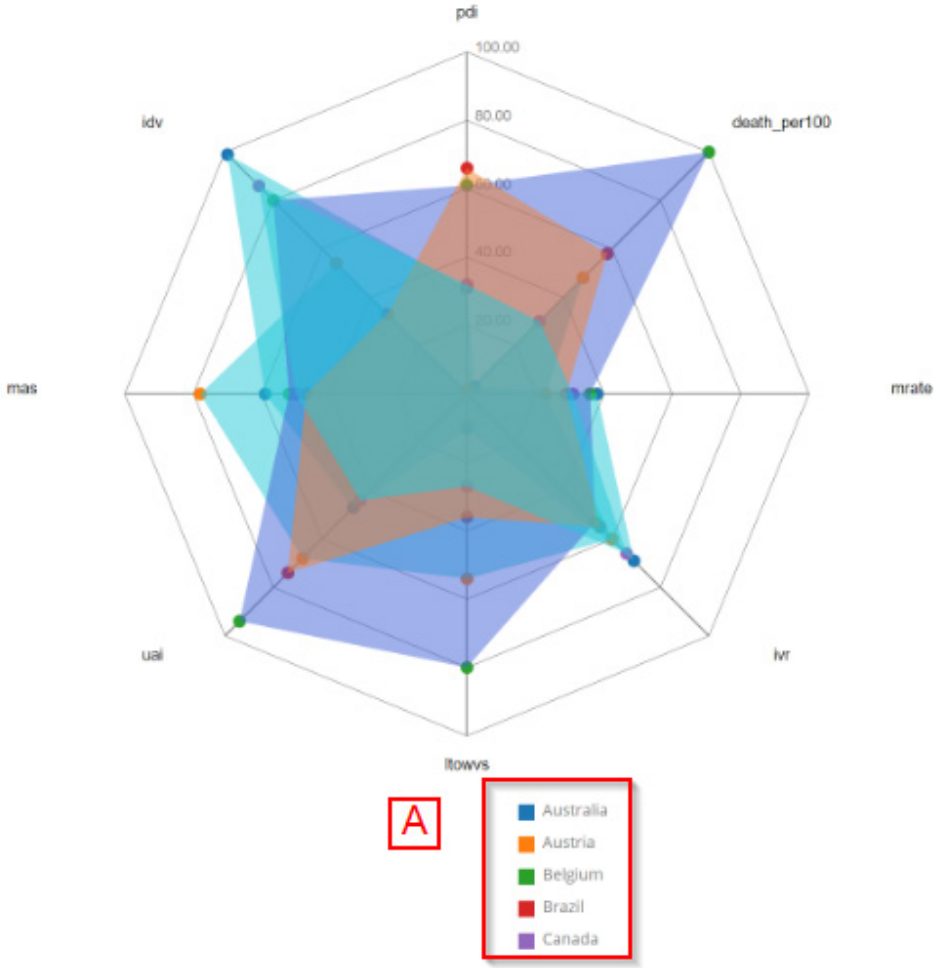


(a) Radar chart of countries with high death rate per 100,000 of population.



(b) Radar chart of countries with low death rate per 100,000 of population.

Figure 6.61: Radar chart of top five countries in the eight dimensional model (Hofstede and Covid-19 epidemiological data). with highest and top five countries with the lowest death rate per 100,000 of population. This is to explore the radar chart for case study 3. The source of the data is available in Table A.4.



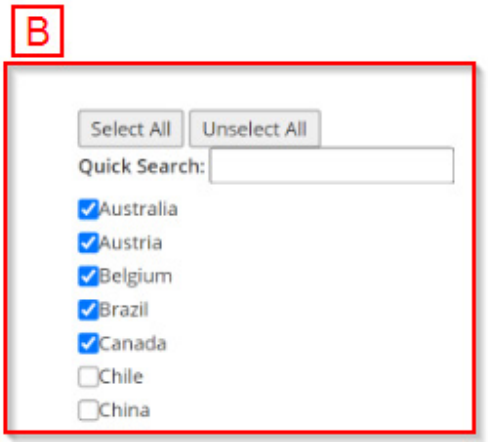


Figure 6.62: Radar chart with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Here we see a comparison between the all the dimensions of Australia, Austria, Belgium, Brazil, and Canada. This is to demonstrate the search and filtering capability of the radar chart.

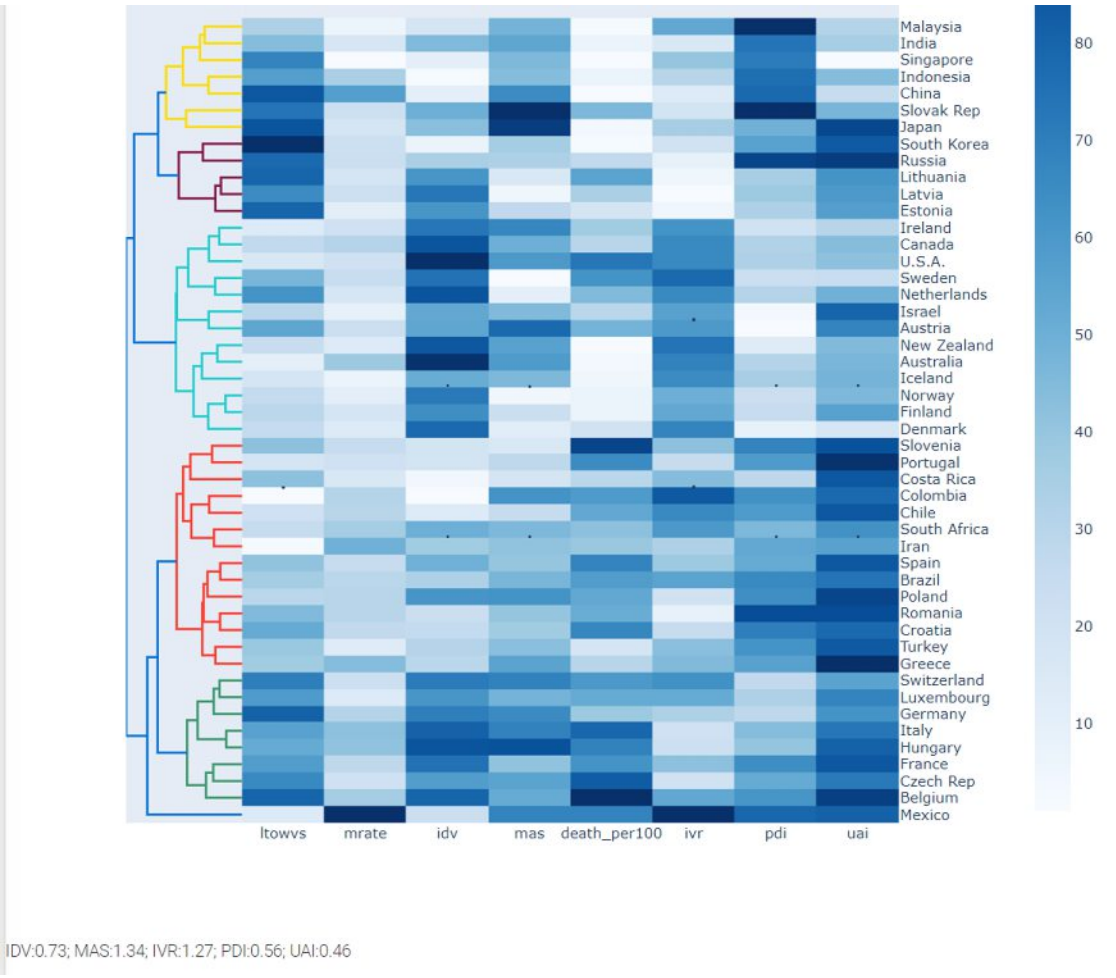


Figure 6.63: Clustermap with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Using the iterative imputation method, the RSME of IDV is 0.73; MAS is 1.34; IVR is 1.27; PDI is 0.56 and UAI is 0.46. A dendrogram on the left side shows the pattern of hierarchical clustering applied. It is possible to zoom in by selecton of an area either on the dendrogram or the heatmap.

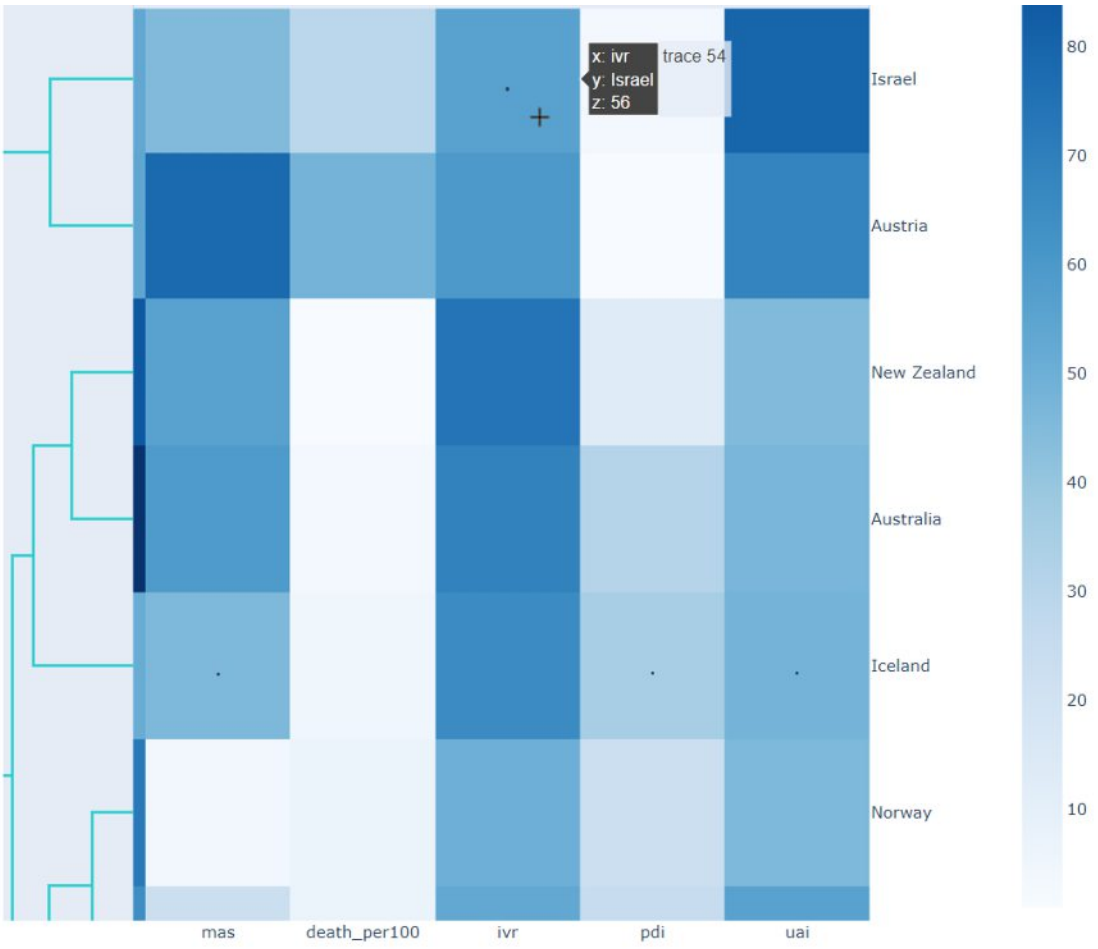


Figure 6.64: Clustermap with regard to the Hofstede model and Covid-19 pandemic measurements (data source found in Table A.4) used for case study 3. Zooming into the clustermap shown in Figure 6.63 reveals a closer look into the missing data of countries. Hovering the mouse over the cell reveals all the information i.e. Israel's PDI score is 56.

6.4 Evaluation of the Visualization Framework

We have already mentioned in Section 3.5 that we use the Visual Data Analysis and Reasoning (VDAR) approach to evaluate the visualization framework. For this reason, we have created three case studies. Section 6.1 showed how it is possible to add a new dimension to the Hofstede model and confirm or reject hypotheses. The six hypotheses used were based on Achim [184] who investigated the correlation between Hofstede's cultural model and corruption level in the countries. Each of these hypotheses was investigated using the parallel coordinates where a user could *discover* knowledge while *exploring* different axis on the parallel coordinates to *identify* correlations. *selection* and *filtering* of countries were possible by brushing the axis of each dimension. Additionally, each axis could be *re-arranged* by dragging it across the visualization with the mouse. Lastly, the coordinates' coloring can be changed by just clicking on the title of the different axis. All of these interactions fulfill what we have defined for Task 2 in Section 4.3 and 4.1.

In Section 6.2, we attempted to reproduce the clustering and grouping of the countries based on the method used in the literature written by Hofstede [56]. Using the dataset provided by Hofstede, we acquired data of 53 countries and clustered them using hierarchical clustering with the average linkage method. The result was presented in table 6.2. The difference between the clusters can be explained by Hofstede's arbitrary method of choosing clusters in the dendrogram and not cutting it at a specific height. However, if necessary, the exact result can be reproduced by not choosing any clustering method in the visualization tool. The user can create a new column in the dataset named *cluster* and manually perform the clustering in the CSV file. Alternatively, any external tool can be used to perform the clustering. The result can be imported into our visualization tool. More details can be found in the API documentation [3], where the details on communication and sample requests between the client and server are documented.

Using a combination of the world map (see Figure 6.38) and the parallel coordinates (see Figure 6.39), the user has the ability to *identify* a set of countries belonging to a specific region or cluster by locating different clusters of countries on the world map. Different clusters have different colors on the map, meaning different regional clusters which are specific to a region can be easily *discovered*.

The parallel coordinates allowed us to *select* different clusters or set of countries by *changing* the countries using a brushing method on a different axis. This enabled us to *discover* different clusters on the world map and *identify* different regional clusters which have been created due to cultural or linguistic characteristics. Using the world map, we could *lookup* for a specific country or cluster and study them more in detail. These interactions fulfill what was required in Task 3 as described in Section 4.4

In the third case study discussed in Section 6.3, we have added two additional dimensions of mortality rate and death per 100,000 population into the Hofstede model, which are

[3] https://documenter.getpostman.com/view/1113133/TzCV2j2C

related to the recent Covid-19 pandemic. This section aimed to use every feature available in the visualization framework and discover some knowledge.

Figure 6.41 *presents* a *preview* of the pattern on how missing data exists in our dataset of 49 countries using a heatmap. The white lines indicate that there is indeed missing data; hence an iterative imputation method was applied to the dataset to fill these missing data. After the imputation was applied, we investigated in Section 6.3.1 which clustering method is suitable for our dataset. k-means clustering method was not selected since the elbow plot in Figure 6.42 which *presents* the elbow plot showing the changes on the sum of the squared error (SSE), did not have a clear "elbow." In Figure 6.43, 6.44, and 6.45 we *presented* the impact of hierarchical clustering and cutting the "branch" in different heights in form a dendrogram. *Previewing* different values and using the Silhouette Coefficient, we indicated the most suitable height to cut the branches, resulting in seven clusters.

Once the chosen imputation and clustering method were applied, further visualizations such as the world map (6.46), parallel coordinates (6.47), radar chart, and clustermap get visible to the user. This process complies with the Visual Analytics process discussed in Section 3.2.1, illustrated in Figure 3.2 and eventually applied in a form of a UML activity diagram in Figure 4.15.

In Section 6.3.3 We have *identified* a pattern between Covid-19's mortality rate and a country's population by *exploring* the parallel coordinates and the world map. We *selected* different countries by *arranging* and *filtering* the parallel coordinates in a way in which we only showed the top five countries with the highest and the lowest rate of mortality. The pattern was *identified* which showed us that the countries with a higher population tend to have a higher mortality rate of Covid-19. This newly *discovered* knowledge was then transformed into a hypothesis that needs to be investigated further. The knowledge discovery falls under what we have defined for Task 2 in Section 4.3 which our interactions do align with.

The parallel coordinates, world map, heatmap, elbow plot, and dendrogram were so far sufficient to fulfill Task 2 and 3. However, to fulfill Task 1 and 4, we have to use the spider chart (radar chart) and clustermap as discussed in Section 6.3.5. Using the spider chart, we *presented* a *comparison* within the countries with the highest and lowest mortality or death rate per 100,000 of population. This comparison was also *changed* by *looking up* and *comparing* other countries used by other researchers such as Furner [40] and House et al. [28]. Finally, we *produced* a clustermap that had imputed values encoded as a dot to represent the uncertainty. Using the clusterrmap, the user is able to *compare* the performance of the imputation method in each dimension within the same model. *Changing* the parameters of the imputation allows users to *compare* the performance of different imputation methods as well. Thus, Task 4 is fulfilled by our visualization framework.

An overview of the case studies and their corresponding tasks can be found in Table 6.4. Using the three case studies, we showed that a user is able to perform all four tasks which

we have created in Chapter 4 based on our identified requirements. Thus, the evaluation process is finished, and this version of the visualization framework can be finalized.

Case Study	Fulfills	Visualization methods used
1. Hofstede and CPI	Task 2	Parallel coordinates, and Pearson's correlation matrix
2. Clustering in Hofstede	Task 3	Parallel coordinates, World map, Pearson's correlation matrix, Dendrogram, and Heatmap
3. Hofstede and Covid-19	Task 1, 2, 3, and 4	Parallel coordinates, World map, Radar chart, Clustermap, Pearson's correlation matrix, Elbow plot, Dendrogram, and Heatmap

Table 6.4: Overview of the three case studies discussed in Section 6.1, 6.2, and 6.3 and which task they evaluate.

6.5 Final Framework

We have evaluated our designed visualization framework using the VDAR evaluation method explained in Section 6.4. We showed that our visualization framework supports the four tasks defined based on our requirements in Section 4.1. Additionally, we showed that the flow of the visualization framework is according to the UML activity diagram designed in Figure 4.15 which is based on Van Wijk [103] Visual Analytic approach. The three case studies show how the four defined tasks aid knowledge discovery in the cultural domain, and the evaluation confirms that our visualization framework has the ability to perform them. Thus, we stop the iteration in the design cycle and publish the final version of the framework.

Reproducibility of research is vital in computer science [195, 196]. For this purpose, researchers suggest using methods that make re-using software easier. As explained in Chapter 5, our visualization framework has a server-client architecture and uses a mixture of Python and JavaScript libraries. These libraries can get outdated or deprecated in the future, making the reproducibility of our result challenging by other researchers. To cope with this challenge, we published our final artifact in the form of a docker image that can be accessed from a public repository via the following command:

```
docker pull payamcf/cmvserver
```

Alternatively, the source code can be accessed using following git commands:

```
git clone https://gitlab.com/payamcf/cmvserver.git
git clone https://gitlab.com/payamcf/cmv.git
```

CHAPTER 7

Conclusion and Future Work

In this chapter, we review our work which we have discussed in the previous sections. We start by describing the main goal and aim of the thesis. We review the methodology framework and the multilayered task framework that we relayed while creating our tasks. We also discuss and review an overview of the implementations and developments we did towards the creation of the visualization framework. In the end, we review the evaluation method and each case study that we discussed in the previous chapters.

7.1 Summary

The goal of this master's thesis was to determine which visual analytic methods could aid the knowledge discovery in the cultural domain. We choose Hevner [10] three-cycle research framework for our methodology approach to create a framework that fulfills this purpose.

In the relevance cycle, we have described in detail the meaning of Hofstede's cultural model and its six dimensions which can be seen in Section 2.1.1. In Section 2.2, we investigated different visualization methods applied by other researchers in the cultural domain. Visualizations such as the world map and spider chart were inspired based on the findings in this section. Furthermore, we discovered in Section 2.3.5 different clustering methods, which we used in later sections to group and cluster our data. Lastly, in Section 2.4, we categorized how missing data is observed in datasets and how to use imputation techniques to cope with them.

Upon reviewing current visualization techniques in the cultural science domain, we identified some limitations, such as a lack of dynamic changes in them. Moreover, we articulated that computer-based displays have more ability to provide information in a representation than paper-based documents since they can filter and interact with. These are the output of the relevance cycle and used as in input in the design cycle.

We also described in Chapter 3 the process of visual analytic introduced by Van Wijk [103] which is the output of our rigor cycle, and we used a basis in our design cycle in Chapter 4. The parallel coordinates were chosen after describing and reviewing the different types of multidimensional visualizations in Section 3.3. Finally, we reviewed different evaluation methods in the information visualization domain and chose the VDAR evaluation method introduced by Lam et al. [17]. These are all the basis and source of our knowledge gathered from existing research papers.

In Chapter 4, we defined requirements for our visualization framework based on the gaps identified in the existing visualization methods (output from rigor and relevance cycle). This was the primary step in the design cycle. Based on a taxonomy introduced by Brehmer and Munzner [129], we broke down the requirements into four a medium-level tasks discussed in 4.2, 4.3, 4.4 and 4.5. Based on these tasks, we have chosen the appropriate visualization methods shown in Table 4.1, which then were considered while designing the UIs using mock-ups. The medium-level tasks, UI mock-ups, and activity diagram are all outputs of the rigor cycle and input in the design cycle; based on these inputs, we then implemented the visualization framework in Chapter 5.

We chose a client-server architecture for the implementation, and this enabled us to utilize a combination of JavaScript and Python libraries. The world map was an extension to an already existing visualization framework for Hofstede's six cultural dimensions implemented by Bayat [44]. Mainly D3.js was used for the implementation of the parallel coordinates, radar chart, and world map. The rest of the visualizations such as clustermap, dendrogram, and heatmap were implemented using Plotly. Additionally, all the machine learning algorithms for imputation and clustering were implemented using Python libraries such as Scipy. In our case, the client is a web application and communicates via jQuery with the API, which is implemented using flask.

In Chapter 6, we defined three different case studies to show the features of the implemented visualization framework and verify if the framework has the ability to perform the designed tasks. In all three case studies, the focus was on using the datasets from the same data sources mentioned by the researchers and attempt to re-evaluate the hypotheses in their studies using our visualization framework.

In the first case study described in Section 6.1, the author attempted to understand the relation between corruption level and the six dimensions of the Hofstede model by adding one additional dimension named CPI (Corruption Perception Index) to the standard six dimensions of the Hofstede model. We tried to re-visit the six hypotheses in this study with our visualization framework and compared our result with the paper's result.

The second case study in 6.2 showed how it is possible to use the clustering feature in our visualization framework to group the countries. These groupings can then be visualized in the parallel coordinates and world map. Lastly, in Section 6.3 we showed how we could fully utilize our visualization framework to impute missing data, cluster data, discover knowledge, compare a set of countries and visualize uncertainty. All these were performed on a case study where Erman and Medeiros [5] attempted to combine

two new dimensions of mortality rate and death per 100,000 population to investigate a potential correlation between culture and Covid-19 pandemic. We could confirm the hypothesis that UAI has a positive correlation with death per 100,000 population and discovered additional findings of a potential correlation with the mortality rate that other researchers can investigate further.

Eventually, we reviewed all the case studies and summarized the VDAR evaluation in Section 6.4. The requirements have been identified by reviewing existing literature and building a knowledge base in the rigor cycle. Based on the requirements, we identified four tasks and showed that the visualization framework could perform all the designed tasks. After the evaluation, we created a docker image that other researchers can use to re-use our visualization framework. This whole process can be seen as a framework as illustrated in Figure 1.1, which other researchers can use in the future.

7.2 Future Work and Improvements

By reviewing the current literature and identifying four tasks for our visualization framework, we ended up with an implementation of six visualization methods in our framework. Additional visualization methods such as scatterplots, barcharts, or box plots can be added as an extension to the framework in the future if found to be helpful in the process of knowledge discovery in the cultural science domain. We suggest following the framework presented in the last section to identify the tasks and requirements, identify appropriate visualization methods, design and extend the current framework, evaluate and publish the final version of the visualization framework in a reproducible environment.

Another improvement on the framework would be to add the ability to reduce the dimensions of the data. Our current framework has a limited ability to add high dimensional (e.g., 20 dimensions) data. For this, reducing the dimensions with dimensionality reduction methods, such as Principle Component Analysis (PCA) [197] or t-Distributed Stochastic Neighbor Embedding (t-SNE) [198] would be necessary. Increasing the number of data dimensions would also have implications for the employed visualizations, which would need to be reworked to accommodate a higher data dimensionality.

Currently, the user has the ability to dynamically add or remove dimensions to the cultural model by adapting the dataset. If the visualization framework gets extended in a way where the addition or subtraction of the dimension is comparable with the old model, domain experts would have the ability to compare the impact of the dimensional changes. Studying these impacts, new and more robust cultural models can be generated.

Due to time limitations, we only had the ability to evaluate the visualization framework in a VDAR scenario with case studies. An important point for future work would be to perform additional controlled experiments with domain experts to evaluate the practical value of the designed framework and its usability in the cultural domain.

All in all, this framework is the first step towards the usage of Visual Analytics methods in the cultural science domain. We hope that in the future, more visualization methods

are used to explore cultural models in a flexible manner, to make knowledge discovery in this domain easier and faster.

APPENDIX A

Datasets Used in the Case Studies

A.1 Case Study 1: Combining Hofstede's Model with CPI

Table A.1: Used dataset in case study 1 - Hofstede's 6 dimensions + CPI

country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Africa East	64	27	41	52	32	40	
Africa West	77	20	46	54	9	78	
Albania					61	15	36
Algeria					26	32	36
Andorra						65	
Arab countries	80	38	53	68	23	34	
Argentina	49	46	56	86	20	62	32
Armenia					61		35
Australia	38	90	61	51	21	71	79
Austria	11	55	79	70	60	63	76
Azerbaijan					61	22	29
Bangladesh	80	20	55	60	47	20	25
Belarus					81	15	32
Belgium	65	75	54	94	82	57	77
Belgium French	67	72	60	93			
Belgium Netherl	61	78	43	97			
Bosnia					70	44	38
Brazil	69	38	49	76	44	59	38
Bulgaria	70	30	40	85	69	16	41
Burkina Faso					27	18	38

Continued on next page

Table A.1 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Canada	39	80	52	48	36	68	83
Canada French	54	73	45	60			83
Chile	63	23	28	86	31	68	70
China	80	20	66	30	87	24	37
Colombia	67	13	64	80	13	83	37
Costa Rica	35	15	21	86			55
Croatia	73	33	40	80	58	33	51
Cyprus						70	61
Czech Rep	57	58	57	74	70	29	56
Denmark	18	74	16	23	35	70	91
Dominican Rep					13	54	33
Ecuador	78	8	63	67			32
Egypt					7	4	36
El Salvador	66	19	40	94	20	89	39
Estonia	40	60	30	60	82	16	70
Ethiopia						46	33
Finland	33	63	26	59	38	57	90
France	68	71	43	86	63	48	70
Georgia					38	32	52
Germany	35	67	66	65	83	40	81
Germany East					78	34	
Ghana					4	72	47
Great Britain	35	89	66	35	51	69	81
Greece	60	35	57	112	45	50	46
Guatemala	95	6	37	101			28
Hong Kong	68	25	57	29	61	17	75
Hungary	46	80	88	82	58	31	51
Iceland					28	67	79
India	77	48	56	40	51	26	38
Indonesia	78	14	46	48	62	38	36
Iran	58	41	43	59	14	40	27
Iraq					25	17	16
Ireland	28	70	68	35	24	65	75
Israel	13	54	47	81	38		61
Italy	50	76	70	75	61	30	44
Jamaica	45	39	68	13			41
Japan	54	46	95	92	88	42	75
Jordan					16	43	53
Korea South	60	18	39	85	100	29	54
Kyrgyz Rep					66	39	28

Continued on next page

Table A.1 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Latvia	44	70	9	63	69	13	56
Lithuania	42	60	19	65	82	16	59
Luxembourg	40	60	50	70	64	56	85
Macedonia Rep					62	35	
Malaysia	104	26	50	36	41	57	50
Mali					20	43	35
Malta	56	59	47	96	47	66	60
Mexico	81	30	69	82	24	97	31
Moldova					71	19	33
Montenegro					75	20	44
Morocco	70	46	53	68	14	25	36
Netherlands	38	80	14	53	67	68	84
New Zealand	22	79	58	49	33	75	91
Nigeria					13	84	26
Norway	31	69	8	50	35	55	88
Pakistan	55	14	50	70	50	0	30
Panama	95	11	44	86			39
Peru	64	16	42	87	25	46	36
Philippines	94	32	64	44	27	42	35
Poland	68	60	64	93	38	29	63
Portugal	63	27	31	104	28	33	64
Puerto Rico					0	90	
Romania	90	30	42	90	52	20	46
Russia	93	39	36	95	81	20	29
Rwanda					18	37	54
Saudi Arabia					36	52	52
Serbia	86	25	43	92	52	28	40
Singapore	74	20	48	8	72	46	85
Slovak Rep	104	52	110	51	77	28	51
Slovenia	71	27	19	88	49	48	60
South Africa					34	63	44
South Africa white	49	65	63	49			
Spain	57	51	42	86	48	44	58
Suriname	85	47	37	92			36
Sweden	31	71	5	29	53	78	89
Switzerland	34	68	70	58	74	66	86
Switzerland French	70	64	58	70			
Switzerland German	26	69	72	56			
Taiwan	58	17	45	69	93	49	62
Tanzania					34	38	30

Continued on next page

Table A.1 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr	cpi
Thailand	64	20	34	64	32	45	38
Trinidad and Tobago	47	16	58	55	13	80	39
Turkey	66	37	45	85	46	49	42
U.S.A.	40	91	62	46	26	68	76
Uganda					24	52	25
Ukraine					86	14	27
Uruguay	61	36	38	100	26	53	74
Venezuela	81	12	73	76	16	100	17
Vietnam	70	20	40	30	57	35	31
Zambia					30	42	38
Zimbabwe					15	28	21

A.2 Case Study 2: Clustering the Hofstede Model

A.2.1 Hofstede's Four Dimensional Model

Table A.2: Used dataset in case study 2 - Four dimensions of the Hofstede model

country	pdi	uai	idv	mas
Africa East	64	52	27	41
Africa West	77	54	20	46
Arab countries	80	68	38	53
Argentina	49	86	46	56
Australia	36	51	90	61
Austria	11	70	55	79
Belgium	65	94	75	54
Brazil	69	76	38	49
Canada	39	48	80	52
Chile	63	86	23	28
Colombia	67	80	13	64
Costa Rica	35	86	15	21
Denmark	18	23	74	16
Ecuador	78	67	8	63
El Salvador	66	94	19	40
Finland	33	59	63	26
France	68	86	71	43
Germany	35	65	67	66
Great Britain	35	35	89	66
Greece	60	112	35	57
Guatemala	95	101	6	37
Hong Kong	68	29	25	57
India	77	40	48	56
Indonesia	78	48	14	46
Iran	58	59	41	43
Ireland	28	35	70	68
Israel	13	81	54	47
Italy	50	75	76	70
Jamaica	45	13	39	68
Japan	54	92	46	95
Korea South	60	85	18	39
Malaysia	104	36	26	50
Mexico	81	82	30	69
Netherlands	38	53	80	14
New Zealand	22	49	79	58
Norway	31	50	69	8

Continued on next page

Table A.2 – continued from previous page

country	pdi	uai	idv	mas
Pakistan	55	70	14	50
Panama	95	86	11	44
Peru	64	87	16	42
Philippines	94	44	32	64
Portugal	63	104	27	31
Singapore	74	8	20	48
South Africa white	49	49	65	63
Spain	57	86	51	42
Sweden	31	29	71	5
Switzerland	34	58	68	70
Taiwan	58	69	17	45
Thailand	64	64	20	34
Turkey	66	85	37	45
U.S.A.	40	46	91	62
Uruguay	61	100	36	38
Venezuela	81	76	12	73
Yugoslavia	76	88	27	21

A.2.2 Hofstede's Six Dimensional Model

Table A.3: Used dataset in case study 2 - Hofstede's 6 dimensions

country	pdi	idv	mas	uai	ltowvs	ivr
Africa East	64	27	41	52	32	40
Africa West	77	20	46	54	9	78
Albania					61	15
Algeria					26	32
Andorra						65
Arab countries	80	38	53	68	23	34
Argentina	49	46	56	86	20	62
Armenia					61	
Australia	38	90	61	51	21	71
Austria	11	55	79	70	60	63
Azerbaijan					61	22
Bangladesh	80	20	55	60	47	20
Belarus					81	15
Belgium	65	75	54	94	82	57
Belgium French	67	72	60	93		
Belgium Netherl	61	78	43	97		
Bosnia					70	44

Continued on next page

Table A.3 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr
Brazil	69	38	49	76	44	59
Bulgaria	70	30	40	85	69	16
Burkina Faso					27	18
Canada	39	80	52	48	36	68
Canada French	54	73	45	60		
Chile	63	23	28	86	31	68
China	80	20	66	30	87	24
Colombia	67	13	64	80	13	83
Costa Rica	35	15	21	86		
Croatia	73	33	40	80	58	33
Cyprus						70
Czech Rep	57	58	57	74	70	29
Denmark	18	74	16	23	35	70
Dominican Rep					13	54
Ecuador	78	8	63	67		
Egypt					7	4
Ethiopia						46
El Salvador	66	19	40	94	20	89
Estonia	40	60	30	60	82	16
Finland	33	63	26	59	38	57
France	68	71	43	86	63	48
Georgia					38	32
Germany	35	67	66	65	83	40
Germany East					78	34
Ghana					4	72
Great Britain	35	89	66	35	51	69
Greece	60	35	57	112	45	50
Guatemala	95	6	37	101		
Hong Kong	68	25	57	29	61	17
Hungary	46	80	88	82	58	31
Iceland					28	67
India	77	48	56	40	51	26
Indonesia	78	14	46	48	62	38
Iran	58	41	43	59	14	40
Iraq					25	17
Ireland	28	70	68	35	24	65
Israel	13	54	47	81	38	
Italy	50	76	70	75	61	30
Jamaica	45	39	68	13		
Japan	54	46	95	92	88	42

Continued on next page

Table A.3 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr
Jordan					16	43
Korea South	60	18	39	85	100	29
Kyrgyz Rep					66	39
Latvia	44	70	9	63	69	13
Lithuania	42	60	19	65	82	16
Luxembourg	40	60	50	70	64	56
Macedonia Rep					62	35
Malaysia	104	26	50	36	41	57
Mali					20	43
Malta	56	59	47	96	47	66
Mexico	81	30	69	82	24	97
Moldova					71	19
Montenegro					75	20
Morocco	70	46	53	68	14	25
Netherlands	38	80	14	53	67	68
New Zealand	22	79	58	49	33	75
Nigeria					13	84
Norway	31	69	8	50	35	55
Pakistan	55	14	50	70	50	0
Panama	95	11	44	86		
Peru	64	16	42	87	25	46
Philippines	94	32	64	44	27	42
Poland	68	60	64	93	38	29
Portugal	63	27	31	104	28	33
Puerto Rico					0	90
Romania	90	30	42	90	52	20
Russia	93	39	36	95	81	20
Rwanda					18	37
Saudi Arabia					36	52
Serbia	86	25	43	92	52	28
Singapore	74	20	48	8	72	46
Slovak Rep	104	52	110	51	77	28
Slovenia	71	27	19	88	49	48
South Africa					34	63
South Africa white	49	65	63	49		
Spain	57	51	42	86	48	44
Suriname	85	47	37	92		
Sweden	31	71	5	29	53	78
Switzerland	34	68	70	58	74	66
Switzerland French	70	64	58	70		

Continued on next page

Table A.3 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr
Switzerland German	26	69	72	56		
Taiwan	58	17	45	69	93	49
Tanzania					34	38
Thailand	64	20	34	64	32	45
Trinidad and Tobago	47	16	58	55	13	80
Turkey	66	37	45	85	46	49
U.S.A.	40	91	62	46	26	68
Uganda					24	52
Ukraine					86	14
Uruguay	61	36	38	100	26	53
Venezuela	81	12	73	76	16	100
Vietnam	70	20	40	30	57	35
Zambia					30	42
Zimbabwe					15	28

A.3 Case Study 3: Hofstede Model and Covid-19

Table A.4: Used dataset in case study 3 - Hofstede's 6 + Covid-19 dimensions

country	pdi	idv	mas	uai	ltowvs	ivr	mrate	death_per100
Australia	38	90	61	51	21	71	3.2	3.64
Austria	11	55	79	70	60	63	1.9	87.07
Belgium	65	75	54	94	82	57	3	184.43
Brazil	69	38	49	76	44	59	2.4	106.91
Canada	39	80	52	48	36	68	2.6	53.77
Chile	63	23	28	86	31	68	2.5	97.92
China	80	20	66	30	87	24	4.8	0.35
Colombia	67	13	64	80	13	83	2.6	108.06
Costa Rica	35	15	21	86			1.3	52.09
Croatia	73	33	40	80	58	33	2.2	122.22
Czech Rep	57	58	57	74	70	29	1.7	152.56
Denmark	18	74	16	23	35	70	1.1	36.34
Estonia	40	60	30	60	82	16	0.9	31.12
Finland	33	63	26	59	38	57	1.5	12.16
France	68	71	43	86	63	48	2.3	113.46
Germany	35	67	66	65	83	40	2.6	68.86
Greece	60	35	57	100	45	50	3.7	53.87
Hungary	46	80	88	82	58	31	3.4	127.58
Iceland					28	67	0.5	8.2
India	77	48	56	40	51	26	1.4	11.41
Indonesia	78	14	46	48	62	38	2.8	11.11
Iran	58	41	43	59	14	40	4.1	70.77
Ireland	28	70	68	35	24	65	1.7	67.83
Israel	13	54	47	81	38		0.7	53.33
Italy	50	76	70	75	61	30	3.5	146.08
Japan	54	46	95	92	88	42	1.5	4.5
Latvia	44	70	9	63	69	13	1.8	61.25
Lithuania	42	60	19	65	82	16	1.5	99.91
Luxembourg	40	60	50	70	64	56	1.1	94.94
Malaysia	100	26	50	36	41	57	0.4	2.37
Mexico	81	30	69	82	24	97	8.5	125.27
Netherlands	38	80	14	53	67	68	1.4	81.65
New Zealand	22	79	58	49	33	75	1.1	0.51
Norway	31	69	8	50	35	55	0.9	10.61
Poland	68	60	64	93	38	29	2.5	97.64
Portugal	63	27	31	99	28	33	1.7	118.45
Romania	90	30	42	90	52	20	2.5	93.79

Continued on next page

Table A.4 – continued from previous page

country	pdi	idv	mas	uai	ltowvs	ivr	mrate	death_per100
Russia	93	39	36	95	81	20	1.9	49.53
Singapore	74	20	48	8	72	46	0	0.51
Slovak Rep	100	52	100	51	77	28	1.8	83.81
Slovenia	71	27	19	88	49	48	2.1	168.81
South Africa					34	63	3	76.07
South Korea	60	18	39	85	100	29	1.8	2.75
Spain	57	51	42	86	48	44	2.1	124.82
Sweden	31	71	5	29	53	78	2	113.83
Switzerland	34	68	70	58	74	66	1.8	110.02
Turkey	66	37	45	85	46	49	1	31.42
U.S.A.	40	91	62	46	26	68	1.7	134.32

A.3.1 Case Study 3: Population of countries

Table A.5: Used dataset in case study 3 - Population of countries

country	population
Australia	25499884
Austria	9006398
Belgium	11589623
Brazil	212559417
Canada	37742154
Chile	19116201
China	1439323776
Colombia	50882891
Costa Rica	5094118
Croatia	4105267
Czech Rep	10708981
Denmark	5792202
Estonia	1326535
Finland	5540720
France	65273511
Germany	83783942
Greece	10423054
Hungary	9660351
Iceland	341243
India	1380004385
Indonesia	273523615
Iran	83992949
Ireland	4937786

Continued on next page

Table A.5 – continued from previous page

country	population
Israel	8655535
Italy	60461826
Japan	126476461
Latvia	1886198
Lithuania	2722289
Luxembourg	625978
Malaysia	32365999
Mexico	128932753
Netherlands	17134872
New Zealand	4822233
Norway	5421241
Poland	37846611
Portugal	10196709
Romania	19237691
Russia	145934462
Singapore	5850342
Slovak Rep	5459642
Slovenia	2078938
South Africa	59308690
South Korea	51269185
Spain	46754778
Sweden	10099265
Switzerland	8654622
Turkey	84339067
U.S.A.	331002651

List of Figures

List of Tables

Listings

Bibliography

[1] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):2307–0919, 2011.

[2] Robert House, Mansour Javidan, Paul Hanges, and Peter Dorfman. Understanding cultures and implicit leadership theories across the globe: an introduction to project globe. *Journal of world business*, 37(1):3–10, 2002.

[3] Marieke De Mooij and Geert Hofstede. The hofstede model: Applications to global branding and advertising strategy and research. *International Journal of advertising*, 29(1):85–110, 2010.

[4] Viv J Shackleton and Abbas H Ali. Work-related values of managers: A test of the hofstede model. *Journal of cross-cultural psychology*, 21(1):109–118, 1990.

[5] Aysegul Erman and Mike Medeiros. Exploring the impact of cultural variability on covid-19-related mortality: A meta-analytic approach. 2020.

[6] Colin Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[7] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[8] Mike Cammarano, Xin Dong, Bryan Chan, Jeff Klingner, Justin Talbot, Alon Halevy, and Pat Hanrahan. Visualization of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

[9] Louis Engelbrecht, Adele Botha, and Ronell Alberts. Designing the visualization of information. *International Journal of Image and Graphics*, 15(02):1540005, 2015.

[10] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

[11] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[12] Moe Thandar Wynn, Erik Poppe, Jingxin Xu, Arthur HM ter Hofstede, Ross Brown, Azzurra Pini, and Wil MP van der Aalst. Processprofiler3d: A visualisation framework for log-based process performance comparison. *Decision Support Systems*, 100:93–108, 2017.

[13] Julian J Hamm. *Technology-assisted healthcare: exploring the use of mobile 3D visualisation technology to augment home-based fall prevention assessments*. PhD thesis, Brunel University London, 2018.

[14] Louis Engelbrecht, Marna Botha, and Adele Botha. Using information visualisation to give voice to a historical community. In *13th Prato CIRN Conference*, 2016.

[15] AH Wiberg, S Løvhaug, M Mathisen, B Tschoerner, Eirik Resch, M Erdt, and Ekaterina Prasolova-Førland. Visualisation of kpis in zero emission neighbourhoods for improved stakeholder participation using virtual reality. In *IOP Conference Series: Earth and Environmental Science*, volume 323, page 012074. IOP Publishing, 2019.

[16] Michael Schelkle, Christian Karl Grund, and Lena Anja Eleonore Aurnhammer. Increasing information visualization compliance in self-service business intelligence with user assistance systems. In *Euorpean Conference on Information Systems (ECIS)-Proceedings of the Workshop on Designing User Assistance in Interactive Intelligent Systems*, pages 44–54, 2018.

[17] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2011.

[18] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.

[19] Geert H Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and organizations: Software of the mind*, volume 2. Mcgraw-hill New York, 2005.

[20] Katharina Chudzikowski, Gerhard Fink, Wolfgang Mayrhofer, Michael Minkov, and Geert Hofstede. The evolution of hofstede's doctrine. *Cross cultural management: An international journal*, 2011.

[21] HJ Eysenck. The four dimensions-review of the books culture's consequences by g. hofstede. *New Society*, 1981.

[22] Harry C Triandis. Dimensions of cultural variation as parameters of organizational theories. *International Studies of Management & Organization*, 12(4):139–169, 1982.

[23] Geert Hofstede. *Culture's consequences: International differences in work-related values*, volume 5. sage, 1984.

[24] Bader Yousef Obeidat, Rifat O Shannak, REMDT Masa'deh, and I Al-Jarrah. Toward better understanding for arabian culture: Implications based on hofstede's cultural model. *European Journal of Social Sciences*, 28(4):512–522, 2012.

[25] Geert Hofstede et al. Organizations and cultures: Software of the mind. *McGrawHill, New York*, 1991.

[26] Chinese Culture Connection. Chinese values and the search for culture-free dimensions of culture. *Journal of cross-cultural psychology*, 18(2):143–164, 1987.

[27] Michael Minkov. *What makes us different and similar: a new interpretation of the world values and other cross-cultural data.* na, 2007.

[28] Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Sage publications, 2004.

[29] Geert Hofstede. The 6-d model of national culture. URL `https://geerthofstede.com/culture-geert-hofstede-gert-jan-hofstede/6d-model-of-national-culture/`.

[30] Mingyu Zhang. *The effect of national culture on corporate policies: evidence from the US corporations.* PhD thesis, Concordia University, 2020.

[31] Hofstede Insights. Compare countries. *Retrieved from Hofstede Insights: https://www. hofstede-insights. com/product/compare-countries*, 2018.

[32] Nripendra P Rana, Emma L Slade, Ganesh P Sahu, Hatice Kizgin, Nitish Singh, Bidit Dey, Anabel Gutierrez, and Yogesh K Dwivedi. Digital and social media marketing.

[33] SHI Xiumei and WANG Jinying. Cultural distance between china and us across globe model and hofstede model. *International Business and Management*, 2(1): 11–17, 2011.

[34] Thomas SC Yap. A culture in the land down under: A malaysian manager's perspective. *Journal of the American Statistical Association (JASA)*, 2:69–72, 2007.

[35] Zhenqun Shi, Lianbo Zhu, Huini Li, and Yilei Huang. Research on the influence of cultural differences between china and japan on employee behavior based on hofstede theory. In *2019 8th International Conference on Industrial Technology and Management (ICITM)*, pages 81–84. IEEE, 2019.

[36] Johan Wiberg and Joakim Månsson. Consumers' perceptions of social media advertisements: a cross-cultural comparison among sweden, india, and japan., 2019.

[37] Mehtap Aldogan Eklund. Compensation according to culture. In *Fairness of CEO Compensation*, pages 105–112. Springer, 2019.

[38] Donelson R Forsyth, Ernest H O'boyle, and Michael A McDaniel. East meets west: A meta-analytic investigation of cultural variations in idealism and relativism. *Journal of Business Ethics*, 83(4):813–833, 2008.

[39] Sjoerd Beugelsdijk and Chris Welzel. Dimensions and dynamics of national culture: Synthesizing hofstede with inglehart. *Journal of Cross-Cultural Psychology*, 49(10): 1469–1505, 2018.

[40] Emily Furner. Cultural differences in russian and american magazine advertising. *Russian Language Journal*, 68:101–130, 2018.

[41] Sergio Barile, Rossella Canestrino, Pierpaolo Magliocca, and Francesco Caputo. The influence of cultural dimensions on corporate social responsibility: Reflections about italian firms.

[42] Florin Lucian Isac and Eugen Florin Remeș. Tradition vs. modernity in japanese management. *Studia Universitatis „Vasile Goldis" Arad–Economics Series*, 30(1): 76–90, 2020.

[43] URL `https://globeproject.com/`.

[44] Hannah Clara Bayat. The visualization of the evolution of cultural models, 2019. URL `https://www.cg.tuwien.ac.at/research/publications/2019/Bayat_2019/`.

[45] The 6 dimensions model of national culture by geert hofstede, Mar 2020. URL `https://geerthofstede.com/culture-geert-hofstede-gert-jan-hofstede/6d-model-of-national-culture/`.

[46] Roy Gelbard, Abraham Carmeli, Ran M Bittmann, and Simcha Simi Ronen. Cluster analysis using multi-algorithm voting in cross-cultural studies. *Expert Systems with Applications*, 36(7):10438–10446, 2009.

[47] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[48] Michael H Bond. Finding universal dimensions of individual variation in multicultural studies of values: The rokeach and chinese value surveys. *Journal of personality and social psychology*, 55(6):1009, 1988.

[49] Kwok Leung, Michael Harris Bond, D William Carment, Lila Krishnan, and Wim BG Liebrand. Effects of cultural femininity on preference for methods of conflict processing: A cross-cultural study. *Journal of Experimental Social Psychology*, 26(5):373–388, 1990.

[50] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. 2005.

[51] Brian Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press, Cambridge, UK; New York, 2002. ISBN 052181099X 9780521810999. URL `http://www.worldcat.org/search?qt=worldcat_org_all&q=052181099X`.

[52] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[53] Odilia Yim and Kylee T Ramdeen. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology*, 11(1):8–21, 2015.

[54] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[55] Bruce M Russett. *World handbook of political and social indicators*. Number 1. Greenwood Pub Group, 1977.

[56] Geert Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[57] Simcha Ronen and Oded Shenkar. Mapping world cultures: Cluster formation, sources and implications. *Journal of International Business Studies*, 44(9):867–897, 2013.

[58] Bert van Pinxteren. National culture and africa revisited: Ethnolinguistic group data from 35 african countries. *Cross-cultural research*, 54(1):73–91, 2020.

[59] Gyula Bakacsi, Takács Sándor, Karácsonyi András, and Imrek Viktor. Eastern european cluster: tradition and transition. *Journal of world Business*, 37(1):69–80, 2002.

[60] Gizem Sayan Kökalan. A cross-country comparison of eu countries in terms of women entrepreneurship determinants: A statistical analysis.

[61] Jeffrey Braithwaite, Yvonne Tran, Louise A Ellis, and Johanna Westbrook. Inside the black box of comparative national healthcare performance in 35 oecd countries: Issues of culture, systems performance and sustainability. *Plos one*, 15(9):e0239776, 2020.

[62] Yue Pan, George M Zinkhan, and Margy P Conchar. Investigating correlates of the subjective well-being of nations: An exploration of missing data techniques. In *American Marketing Association. Conference Proceedings*, volume 13, page 346. American Marketing Association, 2002.

[63] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[64] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[65] Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology*. CRC Press, 2014.

[66] Teresa A Myers. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication methods and measures*, 5(4):297–310, 2011.

[67] Melissa Humphries. Missing data & how to deal: An overview of missing data. *Population Research Center. University of Texas. Recuperado de: http://www. google. com/url*, pages 39–41, 2013.

[68] A Plaia and AL Bondi. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40(38):7316–7330, 2006.

[69] Patricia A Patrician. Multiple imputation for missing data. *Research in nursing & health*, 25(1):76–84, 2002.

[70] Gerko Vink. Towards a standardized evaluation of multiple imputation routines. 2016.

[71] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[72] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[73] Steffen Moritz, Alexis Sardá, Thomas Bartz-Beielstein, Martin Zaefferer, and Jörg Stork. Comparison of different methods for univariate time series imputation in r. *arXiv preprint arXiv:1510.03924*, 2015.

[74] Hakan Demirtas, Sally A Freels, and Recai M Yucel. Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1): 69–84, 2008.

[75] Sander Greenland and William D Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264, 1995.

[76] Jia Shao, Wei Meng, and Guodong Sun. Evaluation of missing value imputation methods for wireless soil datasets. *Personal and Ubiquitous Computing*, 21(1): 113–123, 2017.

[77] Bobbie-Jo M Webb-Robertson, Holli K Wiberg, Melissa M Matzke, Joseph N Brown, Jing Wang, Jason E McDermott, Richard D Smith, Karin D Rodland, Thomas O Metz, Joel G Pounds, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of proteome research*, 14(5):1993–2001, 2015.

[78] Sumanth Yenduri and S Sitharama Iyengar. Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering*, 17(01):127–152, 2007.

[79] Hea Jun Yoon. Predicting employee voice behavior: an exploration of the roles of empowering leadership, power distance, organizational learning capability, and sense of empowerment in korean organizations. 2012.

[80] Songshan Sam Huang and John Crotts. Relationships between hofstede's cultural dimensions and tourist satisfaction: A cross-country cross-sample examination. *Tourism management*, 72:232–241, 2019.

[81] Ravi Chinta and Nejat Capar. Comparative analysis of managerial values in the usa and china. *Journal of Technology Management in China*, 2007.

[82] Jing Qian, Xiaosong Lin, and George Zhen-Xiong Chen. Authentic leadership and feedback-seeking behaviour: An examination of the cultural context of mediating processes in china. *Journal of Management & Organization*, 18(3):286–299, 2012.

[83] Shirley Ye Sheng and Michael R Mullen. A hybrid model for export market opportunity analysis. *International Marketing Review*, 28(2):163–182, 2011.

[84] Mansour Javidan, Robert J House, Peter W Dorfman, Paul J Hanges, and Mary Sully De Luque. Conceptualizing and measuring cultures and their consequences: a comparative review of globe's and hofstede's approaches. *Journal of international business studies*, 37(6):897–914, 2006.

[85] Jan-Benedict EM Steenkamp and Inge Geyskens. Transaction cost economics and the roles of national culture: A test of hypotheses based on inglehart and hofstede. *Journal of the Academy of Marketing Science*, 40(2):252–270, 2012.

[86] José S Rodrigues, Alexandra R Costa, and Carlos Guillén Gestoso. Project planning and control: Does national culture influence project success? *Procedia Technology*, 16:1047–1056, 2014.

[87] Tshepiso Kgapola. *Cultural values and entrepreneurial growth motivation: a focus on township enterprises in Tshwane.* PhD thesis, 2019.

[88] Camila Lee Park and Ely Laureano Paiva. How do national cultures impact the operations strategy process? *International Journal of Operations & Production Management*, 2018.

[89] Ken Kamoche, Lisa Qixun Siebers, Aminu Mamman, Aloysius Newenham-Kahindi, Olusegun Babalola, and Nealia Sue Bruning. Examining the relationship between individual perceptions of control and contemporary career orientations. *Personnel Review*, 2015.

[90] Ian Alcock, Mathew P White, Tim Taylor, Deborah F Coldwell, Matthew O Gribble, Karl L Evans, Adam Corner, Sotiris Vardoulakis, and Lora E Fleming. 'green'on the ground but not in the air: Pro-environmental attitudes are related to household behaviours but not discretionary air travel. *Global Environmental Change*, 42:136–147, 2017.

[91] Richard A Inman, Sara MG Da Silva, Rasha R Bayoumi, and Paul HP Hanel. Cultural value orientations and alcohol consumption in 74 countries: a societal-level analysis. *Frontiers in psychology*, 8:1963, 2017.

[92] Marlene Walk, Heike Schinnenburg, and Femida Handy. What do talents want? work expectations in india, china, and germany. *German Journal of Human Resource Management*, 27(3):251–278, 2013.

[93] Ming Ming Chiu, Bonnie Wing-Yin Chow, Catherine McBride, and Stefan Thomas Mol. Students' sense of belonging at school in 41 countries: Cross-cultural variability. *Journal of Cross-Cultural Psychology*, 47(2):175–196, 2016.

[94] Silvia Bergmüller. The relationship between cultural individualism–c ollectivism and student aggression across 62 countries. *Aggressive behavior*, 39(3):182–200, 2013.

[95] Mackinlay Card. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[96] Bruce H McCormick. Visualization in scientific computing. *ACM SIGBIO Newsletter*, 10(1):15–21, 1988.

[97] Robert Spence. *Information visualization*, volume 1. Springer, 2001.

[98] Daniel A Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[99] Richard A Becker and William S Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[100] Christopher Williamson and Ben Shneiderman. The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 338–346, 1992.

[101] Robert Spence and Lisa Tweedie. The attribute explorer: information synthesis via exploration. *Interacting with Computers*, 11(2):137–146, 1998.

[102] Lisa Tweedie, Bob Spence, David Williams, and Ravinder Bhogal. The attribute explorer. In *Conference companion on Human factors in computing systems*, pages 435–436, 1994.

[103] Jarke J Van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86. IEEE, 2005.

[104] Richard A Becker, William S Cleveland, and Allan R Wilks. Dynamic graphics for data analysis. *Statistical science*, pages 355–383, 1987.

[105] Jessica M Utts. *Seeing through statistics*. Nelson Education, 2014.

[106] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE transactions on Visualization and Computer Graphics*, 14(6):1539–1148, 2008.

[107] Matthew O Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings Visualization'94*, pages 326–333. IEEE, 1994.

[108] Deborah F Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43(4):423–444, 2003.

[109] Alfred Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2): 69–91, 1985.

[110] Julian Heinrich and Daniel Weiskopf. State of the art of parallel coordinates. In *Eurographics (STARs)*, pages 95–116, 2013.

[111] Harri Siirtola and Kari-Jouko Räihä. Interacting with parallel coordinates. *Interacting with Computers*, 18(6):1278–1309, 2006.

[112] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular brushing of extended parallel coordinates. In *INFOVIS*, volume 2, page 127, 2002.

[113] Hemant Makwana, Sanjay Tanwani, and Suresh Jain. Axes re-ordering in parallel coordinate for pattern optimization. *International Journal of Computer Applications*, 40(13):43–48, 2012.

[114] Hugh Mosley and Antje Mayer. Benchmarking national labour market performance: A radar chart approach. Technical report, WZB Discussion paper, 1999.

[115] Michael M Porter and Pooya Niksiar. Multidimensional mechanics: Performance mapping of natural biological systems using permutated radar charts. *PloS one*, 13 (9):e0204309, 2018.

[116] Nadia Boukhelifa and David John Duke. Uncertainty visualization: why might it fail? In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4051–4056. 2009.

[117] Alex T Pang, Craig M Wittenbrink, Suresh K Lodha, et al. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[118] Chris R Johnson and Allen R Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, 2003.

[119] Gevorg Grigoryan and Penny Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 10(5):564–573, 2004.

[120] Ralf P Botchen, Daniel Weiskopf, and Thomas Ertl. Interactive visualization of uncertainty in flow fields using texture-based techniques. In *Proc. Intl. Symp. on Flow Visualization*, volume 2, 2006.

[121] Ken Brodlie, Rodolfo Allendes Osorio, and Adriano Lopes. A review of uncertainty in data visualization. In *Expanding the frontiers of visual analytics and visualization*, pages 81–109. Springer, 2012.

[122] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R Johnson, Manuel M Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*, pages 3–27. Springer, 2014.

[123] Andrej Cedilnik and Penny Rheingans. Procedural annotation of uncertain information. In *Proceedings Visualization 2000. VIS 2000 (Cat. No. 00CH37145)*, pages 77–84. IEEE, 2000.

[124] Geoffrey Ellis and Alan Dix. An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.

[125] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[126] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, 2004.

[127] Michael Sedlmair. Design study contributions come in different guises: Seven guiding scenarios. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 152–161, 2016.

[128] Sergi Vives, Jason Dykes, and Andrew Merryweather. Visualization for equity analysts: Using the dsm in stock picking. 2015.

[129] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19 (12):2376–2385, 2013.

[130] Gennady Andrienko and Natalia Andrienko. Constructing parallel coordinates plot for problem solving. In *1st International Symposium on Smart Graphics*, pages 9–14, 2001.

[131] Seaborn clustermap. URL `https://seaborn.pydata.org/generated/seaborn.clustermap.html`.

[132] IT Jolliffe, OB Allen, and BR Christie. Comparison of variety means using cluster analysis and dendrograms. *Experimental Agriculture*, 25(02):259–269, 1989.

[133] Xing-xu Zhang, Zhen He, Bin Feng, and Hua Shao. An epigenome-wide dna methylation study of workers with an occupational exposure to lead. *Journal of Applied Toxicology*, 39(9):1311–1319, 2019.

[134] Chi Tung Choy, Chi Hang Wong, and Stephen Lam Chan. Embedding of genes using cancer gene expression data: biological relevance and potential application on biomarker discovery. *Frontiers in genetics*, 9:682, 2019.

[135] Anisah Andini, Betty E Manurung, Marvel Sugi, Septasia Dwi Angfika, Suksmandhira Harimurti, Widyawardana Adiprawita, and Isa Anshori. Pattern recognition using machine learning for cancer classification. In *2019 Asia Pacific Conference on Research in Industrial and Systems Engineering (APCoRISE)*, pages 1–4. IEEE, 2019.

[136] Yiru Zhang, Tassadit Bouadi, and Arnaud Martin. An empirical study to determine the optimal k in ek-nnclus method. In *International Conference on Belief Functions*, pages 260–268. Springer, 2018.

[137] MA Syakur, BK Khotimah, EMS Rochman, and BD Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering*, volume 336, page 012017. IOP Publishing, 2018.

[138] Ying Sun and Marc G Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

[139] Lennie Renner, Devante Langworth IV, and London Gerlach. Advanced engineering mathematics. 1998.

[140] Leonardo Emberti Gialloreti, Roberto Enea, Valentina Di Micco, Daniele Di Giovanni, and Paolo Curatolo. Clustering analysis supports the detection of biological processes related to autism spectrum disorder. *Genes*, 11(12):1476, 2020.

[141] Filippo Ricca, Giuseppe Scanniello, Marco Torchiano, Gianna Reggio, and Egidio Astesiano. On the effectiveness of screen mockups in requirements engineering: results from an internal replication. In *Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement*, pages 1–10, 2010.

[142] moqups. wireframe and ui prototyping tool. URL `https://moqups.com/`.

[143] Haidawati Nasir, Ahmad Nazrin Aris, Adidah Lajis, Kushsairy Kadir, and Sairul I Safie. Development of android application for pest infestation early warning system. In *2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, pages 1–5. IEEE, 2018.

[144] Eric Vasey, Maryam S FakhrHosseini, Zhi Zheng, Chung-Hyuk Park, Ayanna Howard, and Myounghoon Jeon. Development and usability testing of a remote control app for an interactive robot. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 808–812. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[145] Bárbara Pimenta Caetano, Carlos Eduardo Barbosa, Melise Maria Veiga de Paula, and Jano Moreira de Souza. Wecollaborate: Citizen collaboration for government problem-solving. In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 18–23. IEEE, 2017.

[146] Jarot S Suroso, RE Tarigan, and FB Setyawan. Information systems strategic planning at startup company using design thinking method. In *Proceedings of The 4th International Conference on Computer Applications and Information Processing Technology (CAIPT 2017)*, volume 2018, pages 1–6, 2018.

[147] Sebastian Raschka. *Python machine learning*. Packt publishing ltd, 2015.

[148] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[149] Ceyhun Ozgur, Taylor Colliau, Grace Rogers, Zachariah Hughes, B Myer-Tyson, et al. Matlab vs. python vs. r. *Journal of Data Science*, 15(3):355–372, 2017.

[150] Elijah Meeks. *D3. js in Action*. Manning Shelter Island, NY, 2015.

[151] Steve Hamersky. Tableau desktop. *Mathematics and Computer Education*, 50(2): 148, 2016.

[152] Charts | google developers. URL `https://developers.google.com/chart`.

[153] Christopher Ilacqua, Henric Cronstrom, and James Richardson. *Learning Qlik Sense®: The Official Guide*. Packt Publishing Ltd, 2015.

[154] Visually: Premium content creation for better marketing. URL https://visual.ly/.

[155] Scott Murray. *Interactive data visualization for the web: an introduction to designing with D3*. " O'Reilly Media, Inc.", 2017.

[156] Arthur Sun and Tian Huaying. Donvis: An interactive tool for donation information visualization.

[157] Mahzad Zahedi and Babak Teimourpour. Insvis: an interactive visual browser for searching co-author network.

[158] Roberto González-Ibáñez, Camila Márquez, and Daniel Gacitúa. Stare. js: An extensible open source toolkit for visualizing search engine results. In *2019 38th International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–7. IEEE, 2019.

[159] JJ Allaire, Christopher Gandrud, Kenton Russell, and CJ Yetman. networkd3: D3 javascript network graphs from r. *R package version 0.4*, 2017.

[160] Matthew R Laird, Morgan GI Langille, and Fiona SL Brinkman. Genomed3plot: a library for rich, interactive visualizations of genomic data in web applications. *Bioinformatics*, 31(20):3348–3349, 2015.

[161] Python.org. URL https://www.python.org/.

[162] Philip Guo. Python is now the most popular introductory teaching language at top us universities. *BLOG@ CACM, July*, 47, 2014.

[163] Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[164] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[165] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/. [Online; accessed <today>].

[166] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[167] Plotly Technologies Inc. Collaborative data science, 2015. URL `https://plot.ly`.

[168] Kai Chang. Parallel coordinates. *A visual toolkit for multidimensional detectives: URL: http://syntagmatic. github. io/parallel-coordinates*, 2012.

[169] Mike Bostock. Topojson. *URl: https://github.com/topojson/world-atlas*, 2020.

[170] Howard Butler, Martin Daly, Allan Doyle, Sean Gillies, Tim Schaub, and Christopher Schmidt. Geojson. *Electronic. URL: http://geojson. org*, 2014.

[171] Chris Zhou. D3 radar chart. URL `https://gist.github.com/chrisrzhou/2421ac6541b68c1680f8`.

[172] Bringing matplotlib to the browser. URL `https://mpld3.github.io/`.

[173] Aleksey Bilogur. Missingno: a missing data visualization suite. *Journal of Open Source Software*, 3(22):547, 2018.

[174] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017. URL `https://doi.org/10.5281/zenodo.883859`.

[175] Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, and Pedro Despouy. plotly: Create interactive web graphics via 'plotly. js'. *R package version*, 4(1):110, 2017.

[176] M Waskom. seaborn: statistical data visualization. python 2.7 and 3.5.

[177] Shammamah Hossain, C Calloway, D Lippa, D Niederhut, and D Shupe. Visualization of bioinformatics data with dash bio. In *Proceedings of the 18th Python in Science Conference*, pages 126–133, 2019.

[178] Dash overview. URL `https://plotly.com/dash/`.

[179] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[180] sklearn.impute.simpleimputer. URL https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html.

[181] flask documentation (1.1.x). URL https://flask.palletsprojects.com/en/1.1.x/.

[182] The collaboration platform for api development. URL https://www.postman.com/.

[183] Jacob Thornton Mark Otto. Bootstrap. URL https://getbootstrap.com/.

[184] Monica Violeta Achim. Cultural dimension of corruption: A cross-country survey. *International Advances in Economic Research*, 22(3):333–345, 2016.

[185] Transparency international. URL https://www.transparency.org/en/cpi/2015.

[186] Winnie Tong. Analysis of corruption from sociocultural perspectives. *International Journal of Business and Social Science*, 5(11), 2014.

[187] Norman H Nie, Dale H Bent, and C Hadlai Hull. *SPSS: Statistical package for the social sciences*, volume 227. McGraw-Hill New York, 1975.

[188] Yong Shuai, Chunxu Jiang, Xinyi Su, Can Yuan, and Xiaoping Huang. A hybrid clustering model for analyzing covid-19 national prevention and control strategy. In *2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE)*, pages 68–71. IEEE, 2020.

[189] Viktor Stojkoski, Zoran Utkovski, Petar Jolakoski, Dragan Tevdovski, and Ljupco Kocarev. The socio-economic determinants of the coronavirus disease (covid-19) pandemic. *arXiv preprint arXiv:2004.07947*, 2020.

[190] Toan Luu Duc Huynh. Does culture matter social distancing under the covid-19 pandemic? *Safety Science*, 130:104872, 2020.

[191] Alexander Irwin, Dorina Georgieva, Shobhana Sosale, and Brian Min. World bank open data, Jan 2021. URL https://data.worldbank.org/.

[192] Gho | by theme. URL https://apps.who.int/gho/data/node.home.

[193] Oecd data. URL https://data.oecd.org/.

[194] Johns hopkins coronavirus resource center. URL https://coronavirus.jhu.edu/.

[195] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242): 1422–1425, 2015.

[196] Stuart Buck. Solving reproducibility, 2015.

[197] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26 (3):303–304, 2008.

[198] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.