# DIFFERENCES IN ATTITUDES TOWARD THE GREEN TRANSITION ACROSS EUROPE

Lappeenranta–Lahti University of Technology LUT

School of Engineering Science

Software Engineering

Master's Programme in Software Engineering and Digital Transformation

2024

**Aminul Islam**

Supervisors: Associate Professor Annika Wolff

Ajesh Kumar, Junior researcher

Examiners: Associate Professor Annika Wolff

Ajesh Kumar, Junior researcher

ABSTRACT

Lappeenranta–Lahti University of Technology LUT
School of Engineering Science
Software Engineering
Master's Programme in Software Engineering and Digital Transformation

**Aminul Islam**

# Differences in Attitudes Towards the Green Transition Across Europe

Master’s Thesis

2024

81 pages, 39 figures, 1 table, 7 appendices

Examiners: Associate Professor Annika Wolff

Ajesh Kumar, Junior researcher

**Keywords**: Energy attitude, Green Energy Attitude, Clustering Energy Attitude, Energy Citizenship, Energy Behavior

Global warming is a critical problem, and research shows that 60% of energy consumption and more than 70% of carbon emissions are from cities. In the past three decades, CO2 emission levels increased proportionately with the increase in economic growth and energy consumption. Therefore, the imperative shift towards green energy becomes increasingly vital. In response to this urgency, the European Union (EU) has embarked on proactive measures based on renewable technologies and launched a variety of initiatives and projects to meet energy targets. In addition, energy technologies ought to be both sustainable and economically viable for the benefit of the citizens. To achieve this objective, an improvement in the management of the energy market and the assurance of sustainable energy production calls for a thorough energy transition, which requires citizens’ active participation. To ensure the sustainable active participation of citizens, stakeholders need to understand the attitude of citizens toward green transition. This research has developed a paradigm to explore and identify differences in attitudes among energy citizens as a part of understanding energy citizen’s actions and interactions in the context of energy behavior. This research followed a mixed method approach, incorporating both qualitative and quantitative methodologies, along with clustering methods that frame the different attitude components (Affect, Behavior, and Cognition) in a mathematical paradigm. Each resultant cluster illustrates a different attitude mode toward green transition.

**ACKNOWLEDGEMENTS**

I would like to express most profound gratitude and admiration to my supervisor Annika Wolff, throughout this endeavor, whose constant guidance and great insights were crucial. For the successful completion, her guidance and insightful comments served as a beacon, guiding me through the project and ensuring. I am grateful to her, and I deeply appreciate for imparting her knowledge and guiding me through the experience.

Syed Bilal Haider Naqvi provided valuable assistance in enhancing the quality of the document by offering insightful feedback and constructive comments. His expert opinion helped me to the improvement of the overall presentation and readability of the document.

I also want to thank Ajesh Kumar, for helping me with the comments and feedback in the absence of Annika. Furthermore, I had interesting discussions with him about the projects and related things, which helped me to think more and put details about the topic.

Finally, I would like to give thanks to my other colleagues, staff, and LUT University who were indirectly involved in this process.

Aminul Islam

Lappeenranta 30.01.2024

SYMBOLS AND ABBREVIATIONS

Abbreviations

| | |
|---|---|
| ABC | Affect, Behavior, Cognition |
| CTPs | Community Transition Pathways |
| DBCLAs | Density-Based Clustering Algorithms |
| DER | Distributed Energy Resources |
| DSM | Demand Side Management |
| EDA | Explanatory Data Analysis |
| GRETA | GReen Energy Transition Actions |
| GIS | Geographic Information System |
| GHGs | Greenhouse Gases |
| ML | Machine Learning |
| MNS | Multinational Survey |
| NIMBYism | Not In My Backyard |
| PCA | Principal Component Analysis |
| RES | Renewable Energy Sources |
| VVP | Virtual Power Plants |

**Table of contents**

Appendices

Appendix 1.

1. GitHub repository links of implemented code figure chart
2. Filtered dataset by predefined criteria

Appendix 2.

1. Column names that are more than twenty-five percent missing
2. Unique values of each variable

Appendix 3.

1. Columns with all values are numeric
2. Columns with all values are numeric except for one
3. Columns with values are text and need to encode

Appendix 4.

1. Box plot of each variable
2. Correlation Matrix

Appendix 5.

1. K-means, deviation for each cluster
2. K-means, skewness for each cluster

Appendix 6.

1. K-modes, centroids for each cluster
2. K-modes, cluster profile for each cluster, and variables

Appendix 7.

1. Category frequency bar chart of every variable

# 1 Introduction

## 1.1 Background

During the Stone Age, an estimated 4000 calories of energy supply were accessed and consumed by an individual, on the other hand, today the average American uses 230,000 calories per day, which is around 60 times higher than in the Stone Age (Johan. E. et al., 2023). The global energy demand has increased and already raised concerns over the possible confinements of energy supply, and energy reservoir reduction, as well as the intense and uncertain environmental consequences. The International Energy Agency has compiled alarming statistics that depict the trends in energy consumption.

Over the last two decades, primary energy has grown about 49% with an annual 2% growth rate, while CO2 emission has grown by 43% with an annual rate of 1.8%. According to the scientific community, in the future, this growing trend will continue, where only in China energy consumption has doubled in the past twenty years with a striking rate of 3.7% (Luis Pérez-Lombard et al. 2008).

The EU wants to achieve carbon neutrality by 2050, and they developed their policies according to that goal. Over the years, the EU has had a vision to include citizens in an active role in the green transition. The involvement of the public and mass people would decentralize and democratize the energy decision-making process, which will accelerate renewable energy production and energy technology. This process aims to engage energy citizens more deeply in the energy system, fostering a dynamic interaction among energy production, consumption, and responsible decision-making (Madeleine and Jenny, 2022).

Until recently, Energy was a commodity traded in the market, and it may have any form such as oil, gas, electricity, etc. The supply and demand of the traditional market follow economic principles that affect pricing, availability, and distribution of energy. Energy prices are easily influenced by market dynamics, which in turn may have a great impact on industries, economies, and individual consumers. In addition to being a tradable commodity, energy is also a critical system underpinning modern life. Thus, energy has a dual nature, the first one is a commodity subject to market dynamics and the second one is an essential system for daily life. This dual nature of this energy places it within the realms of human rights and the duties of public entities (Antti and Govert, 2023; Urry, 2014).

## 1.2 Deregulation of Energy Market

In the context of energy transition and public policies, enhancing market competition and deregulating energy supply sectors are concerning issues, especially for an efficient economy, which are prime concerns in developed countries (Severin et al., 2000; Rossella et al., 2022). The energy sector in Europe experiencing a concentration of businesses in the energy sector, which may lead to less competition and higher prices at the consumer level. For the larger consumer, liberalization brought lower prices, in countries like the UK, Germany, and Nordic countries, but only a minor amount of reduction on household electricity prices (Rossella et al., 2022). However, in Europe, deregulation has already started, and notable progress has been made in the last decade. Generation of electricity in large power plants and distribution to rural areas could be argued as a natural monopoly because of long-distance transmission loss, but now there is a counterargument that the basic economics of this dynamics has changed in terms of the economics of transmission and distribution (Elisabeth, 1996; Williamson 1965).

Throughout the 1990s, Europe produced sufficient electricity and, in some cases, exceeded the requirements. However, according to researchers' prediction, this situation going to reverse in the upcoming decade. Nordic countries are already facing the issues of deficiency in capacity during dry years, notably with Norwegian and Swedish hydropower operating below the required capacity levels (Elisabeth, 1996; Rossella et al. 2022). [47, 48]

Global warming is a critical environmental concern and becoming worse, because of the emission of greenhouse gases (GHGs). The continuous combustion of fossil fuels, GHGs, poses a significant threat to the environment. Addressing this issue and overcoming the damage requires transitioning from fossil fuels to energy sources that do not exacerbate global warming. In reality, this term can entail a transition towards renewable energy sources (RES) such as solar, wind, wave, and biomass. Furthermore, nuclear power is already omitted as a solution due to several technical threats and unresolved societal issues that pose threats (Elisabeth, 1996; Rossella et al. 2022).

In addition, from the analysis (Elisabeth, 1996), it has been predicted that oil production will be at a peak within the next two decades, while projections indicate a continual rise in oil consumption. This pattern and tendency escalated oil prices and international conflicts over oil resources. The world is industrialized and depends heavily on stable oil and natural gas supplies, so executing required changes to the energy sector specifically for transportation may take several decades.

To address and resolve all the issues above, the Green Energy Transition Actions (GRETA) project has been introduced, which will be discussed in the next section (1.3).

## 1.3 Introducing GRETA Analytics

Citizens are now fundamental actors in decarbonization strategies due to their diverse roles as political actors, users, producers, consumers, and owners (Antti and Govert, 2023). In addition, citizens are also considered active participants in solving several other issues related to energy deregulation (described in section 1.2). Furthermore, there are issues and limitations related to technical, cultural, environmental, and knowledge barriers and also lack of awareness. To pave the way, citizens' engagement needs to strengthen and limitations need to be overcome (Dumitru et al. 2023). Although an individual citizen directly doesn't have legislative power, the issues with green transition explained in section 1.2 require citizens' active participation (such as environmental issues) and decisions now depend also on individuals. One obvious advantage of citizens' engagement is the quality and legitimacy of public decision-making related to green transition. Thus, this green energy transition introduces a democratic system involving citizens to shape a sustainable and democratic energy system from a broader perspective- a concept represented by the notion of 'energy citizenship' that ensures the active participation of citizens ((Devine-Wright, 2007), detail in section 1.4). Evolving around active and engaged citizens, this concept also has a scientific and practical perspective that shapes the energy system democratically and collaboratively. The urgent demand for sustainability and transitions has been emphasized because of the complex challenges posed by climate change, environmental degradation, and the decline of biodiversity. The energy citizenship concept strengthens the relationship between 'being an energy citizen' and 'endorsing low-carbon transitions'. Another important aspect of energy citizenship is to focus on the broader perspective of energy democracy and active energy participation (Horsbøl et al. 2018, Dumitru et al. 2023).

In addition to the issues, challenges, and limitations mentioned above, energy transition also requires citizen's active participation in policymaking and successful implementation. To achieve that, the GRETA (Green Energy Transition Actions) project aims for decarbonization to enhance energy citizenship, this project closely works with citizens and the findings and results will help policymakers. This multinational project provides a platform to understand energy consumption and sustainable energy planning (GRETA 2023a).

Thus, energy citizenship engagement has multiple levels, these engagement concepts are based on the research conducted by the GRETA project, which illustrates that citizens may transit with the following engagement ladder of participation (Dumitru et al. 2023):

1. From unaware to aware energy citizens: Citizens who know about energy-related issues and become aware;
2. From aware to involved energy citizens: This group of citizens is adopting energy-saving measures and started acting within the energy system;

3. From involved to active energy citizens: This group of citizens, for example, decides to join and collaborate with an energy community;
4. From active to advocate, energy citizens: Citizens encourage other citizens to join the community or take some sort of action for green transition.

Figure 1 depicts different citizens' engagement ladders of participation:

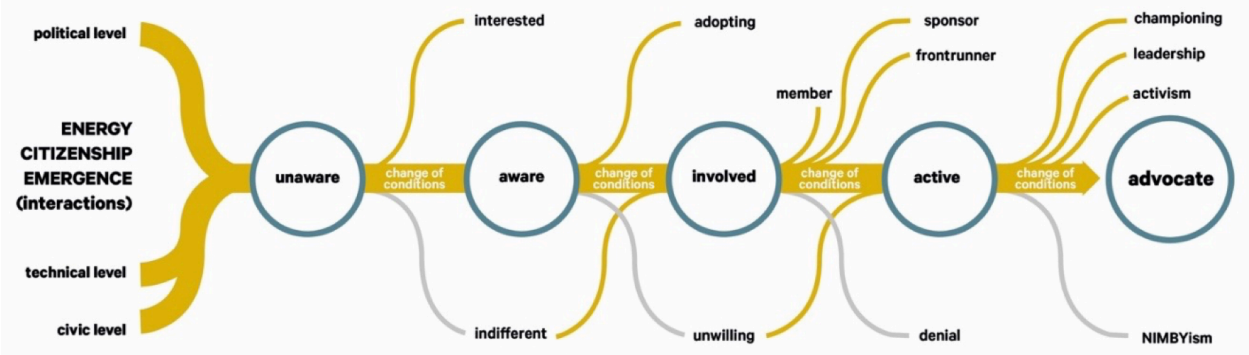


*Figure 1. Different Citizens' Engagement Towards Green Transition (ladder of participation) [61]*

Citizen's engagement in the energy system and energy transition involved stakeholders such as the establishment and management of an energy community are strictly tied to the engagement initiatives (Dumitru et al. 2023). For example, by using smart meters, energy consumers can gain greater awareness (Ajesh et al., 2023). This technology unveils detailed energy consumption patterns, empowering people to understand dynamic energy pricing. It has an economic perspective that influences the business to be customized according to the customer's needs. From a technical point of view, citizen's involvement with this process influences the energy consumption and monitoring system (Dumitru et al. 2023).

Here at this point, the research objective is strictly related and has a significant contribution to understanding the citizens' cognition or knowledge, their emotional engagement, and actions toward green transition with the attitude model and existing research. Regarding this, the subsequent sections 1.4 and 1.5 will illustrate more.

## 1.4 Framing The Research Objective

As discussed in sections 1.2 (Deregulation of Energy Market) and 1.3 (Introducing GRETA Analytics), researchers addressed the issues related to green transition and the view has changed regarding the energy users and consumers. Devine-Wright (2007) suggests the idea called ‘energy citizenship’ which assumes that people already have prior knowledge at least to some extent about energy and energy transition. This concept is for long-term sustainable energy, where energy users are also considered as energy participants in the decision-making process. However, this concept is not limited to that, it’s going even further and includes multifaceted dimensions. This decision-making process includes economic, social, and political perspectives too, where attitude is considered as energy behavior. From the research (Wilson, Charlie, 2007) learned that residential energy efficiency comprises a wide range of behaviors including low-cost efficiency improvements, capital investments, ambient temperature setting, and minor curtailments. Another significant finding (Wilson et al., 2007) was that individuals act as independent decision-makers, but are still susceptible to external influences. To understand the energy efficiency gap, the social dimension of residential energy is required, which has four characteristics. One of them is ‘embeddedness’, encompassing everyday habits like cleanliness, cooking, transportation, childcare, and entertainment, all integrating energy consumption into daily routines. The second one is ‘constraints on choice’, such as the availability of technologies within the supply chain, an individual's skill set and knowledge, and the inclinations of tradesmen and contractors, which collectively limit individual choices. Psychologists and social science researchers tried to find and explore more on the energy citizenship barrier and behavior, mentioned in sections 4.3 and 4.4. However, there are still limitations and a lack of new perspectives to learn and explore, some of those mentioned in sections 1.2, and 1.3, and others will be described in subsequent sections.

In the context of energy transition, public opinion and attitude can be very dynamic depending on the technology and usability of a technical product. Recent research (Gordon Walker, 1995) shows the importance of public opinion toward the green transition. From the socio-cultural perspective, on an individual level, energy consumers need to make decisions regarding energy technology where cognitive and affective processes shape energy behavior (Marianne et al., 2018). At a national policymaking level, the green transition aims to transform energy consumers into active contributors, marking a pivotal shift in energy consumption paradigms. The evidence (Devine-Wright, 2012) suggests that to achieve a long-term green transition, we must focus on issues that matter to the public. Because in the political arena, citizens need to contribute to the scientific and technological future. Material participation of energy users can accelerate the process of developing techno-scientific objects (energy technologies).

Previous paragraphs articulated the research gaps and explained the purposes, necessity, challenges, and importance of further exploration in understanding the public attitude toward green transition. On the way, here are the prime objectives of this research:

- ➢ Find different attitudes of energy citizens with the help of the attitude component, in the context of the energy behavior (sections 1.3, 1.5, and 10)
- ➢ Establish a process to cluster the multinational survey (GRETA) responses within an attitude paradigm (section 5, and the entire research)
- ➢ Find an appropriate and robust clustering algorithm(s) to get meaningful insight from an energy transition survey dataset (section 8, 9 and 10.1)
- ➢ Understanding the underlying characteristics of each cluster and how attitude in one cluster differs from another cluster (section 10.2)
- ➢ Analyze the clustering results in an attitude model as a paradigm (section 10.3)
- ➢ Examining findings with the existing scientific publication and research (section 10.4)

According to these objectives, *the main research questions are mentioned in section 3*. Research methods (section 5) explain details about the approach and the process of scientific solutions that lead to answers to all the research questions. *The discussion section will elaborate and explain the findings using a descriptive approach (section 10)*.

## 1.5 Significance and Contribution of the Research

To overcome the challenges (mentioned in previous sections) and to accelerate the energy transition, the core focus of this research methodology revolved around understanding an individual's mindset, actions, and attitude, as well as how this varies in terms of affect, behavioral, and cognitive aspects.

This research contributes in three ways: firstly, to establish a data annotation or labeling process that categorizes the columns or variables according to the attitude model (ABC-Affect (A), Behavior (B), and, Cognition (C)). This annotation process takes expert opinion and identifies the variables with the respective component of the particular type of attitude. These annotated variables were later used to analyze and infer the citizenship behavior from each cluster of energy citizenship. Secondly, based on the labeled or annotated data, this research applied state-of-the-art clustering algorithms which established a benchmarking process for clustering energy citizenship data. This will help for future data clustering for energy citizenship behavior. Thirdly, this research explored and explained clustering results with the attitude components and reasoning behind each outcome according to the existing research. This will help the scientific community and policymakers to get insight into human behavior toward green transition.

The contribution mentioned above has clearly explained the research methodology and for that reason, this thesis methodology has also three parts (explained in section 5.1). These three parts (or stages) all together established a framework for energy citizenship attitude analysis. This research also revealed the alternative hypothesis, counterarguments, and limitations on the way of energy citizenship attitude analysis in every stage. As a result of the thesis, made the dataset, code, associated statistics, tables, charts, and technical documentation available for future work, all of which are detailed in the Appendices.

One existing research (Divya et al., 2022) was conducted with undergraduate college students based on a survey method, which suggests a strong correlation between energy knowledge and a person's energy attitude. The research was based on a survey using Google Forms that asked questions to the students using the Likert scale method. Although this survey research has similarities with the GRETA analytics survey, the GRETA project is not limited to one single country, rather it covers the entire European Union which is much more diversified. Moreover, questions and responses are much more comprehensive and comprise more social aspects in the context of green transition (GRETA 2023a).

## 1.6 Technical Jargon for the Reader

Some technical jargon will be used for this research which may have different meanings in different contexts, here illustrating it for the readers:

- The word variable, column, and attributes are interchangeable, meaning this represents the same thing
- Attitudes always refer to human attitude and behavior is only a component of attitude there are two more components cognition and affect
- Principal Component Analysis (PCA) and attitude components have completely different meanings and context
- High-dimensional dataset, meaning the number of variables (or columns) is high
- The word clustering has a general meaning, but for this research, this will represent data clustering using ML models
- Data point means a single unit of data, this can be compared with a single cell of an Excel sheet or a CSV file
- Frequency of a value or category means how many times that particular value or category appeared in a particular variable
- In this research, each data point or cell of the GRETA dataset represents a response of a particular person to a particular question
- Missing value means a particular data point or cell is void or null in the context
- The term model will represent machine learning models

# 2 Research Motivation

During the past two decades, a drastic change has been made in the energy sector and attention has shifted in many countries towards renewable energy (Kaldellis et al., 2012). The European Union initiated multiple projects and an array of actions and support measures. This research centers on energy citizenship, specifically exploring diverse attitudes toward the green energy transition. Its outcome aims to provide insights into community engagement, individual perspectives, and behaviors concerning cognitive knowledge and involvement in this shift. This multidisciplinary problem delves into the societal and behavioral aspects of the energy domain. While rooted in the realm of energy, the specific challenge of discerning varied attitudes is inherently social and behavioral, and this research approach involves leveraging advanced technology to address this complexity. The outcome of this research will help us to understand how communities and people are engaged in energy transition, their opinions, and behavior based on cognitive knowledge and engagement.

# 3 Research Questions

The main research question is:

1. What differences in attitudes towards the green transition exist across Europe?

From a technical perspective, this main question leads (or can be broken down) to sub-questions:

1. What clusters can be found from data (from a multinational survey) when focusing on answers to affective, behavioral, and cognitive aspects?
2. How does one cluster differ from another cluster, or what different types of knowledge are found - how/why are these different?

# 4 Related Work

## 4.1 Exploring Energy Citizenship

Energy citizenship typically highlights the importance of altering behavior and engaging individuals in energy systems, frequently spotlighting individuals as catalysts for transformation (Madeleine and Jenny, 2022). According to a study, energy citizenship encompasses both rights and obligations, anchored by sustainability principles of participation, local action, equity, justice, and the alleviation of poverty (Baron et al., 1984).

The notion of energy citizenship serves as a fundamental framework to explicate how community-based energy initiatives can play a pivotal role in fostering individual comprehension of energy and sustainability. Furthermore, it allows individuals to actively engage in shaping broader energy policy landscapes (Madeleine and Jenny, 2022; Lennon and Myles, 2017).

Recent research suggests (Lennon et al. 2019; Gordon Walker, 1995) that the concept of citizenship should be revisited because that concept historically pointed to a specific group of people from society. Women were often confined to domestic spaces, and these historical divisions still influence our current understanding of citizenship. Now this raises questions about the relevance of traditional citizenship concepts to the energy transition because energy consumption occurs in traditionally female-dominated domestic areas. It's proposed that the idea of citizenship needs to be expanded, framing energy not merely as a commodity but as a necessity under a rights-based model.

The phrase ‘energy citizenship’ is often used superficially, but to understand its practical meaning, one needs to draw insights from related concepts like ecological or environmental citizenship and discussions on participation in sustainable development (Gordon Walker, 1995). Chilvers and Longhurst (2016) stated that going beyond a normative understanding of ‘deliberative versus individualist’ and ‘citizen versus consumer’ can help define more inclusive interpretations of energy citizenship. This process may challenge the current paradigm of existing inequalities embedded in current conceptualizations of citizenship and public engagement (Lennon et al. 2019).

Energy citizens’ social and psychological perspectives were usually neglected, but they are important and need to be addressed in how emergent technological innovation might contribute to sustainable development such as environmental, economic, and social policy goals (Devine-Wright, 2012). This is where the concept ‘sustainable energy system’ arises. For an environmentally significant behavior, three aspects need to be considered, the private, the public, and the corporate or institutional context. By considering the social-psychological theory of knowledge, social representation theory (Moscovici, 1984) can be a useful framework to study the belief around the ‘common sense’ of energy citizenship.

## 4.2 Evolving Role of Energy Consumers

Until recently, energy users were treated merely as customers passive participants within the market dynamics, or to some extent individuals receiving technology within the periphery of centralized systems. Regarding the latter function, public involvement has been elucidated in the context of NIMBYism (Not In My Backyard) and knowledge gaps. Research shows that emerging technologies will create material participation of energy users (Ryghaug et al. 2018).

Energy users, often refer to and produce ‘active consumers’ with the government policy which is managed under the banner of demand side management (DSM) focusing on confirming better utility and needs of the energy system. This strategy and policy have been debated and criticized, only predefined options were the only choices for the consumers. Thus, in the traditional system, the public has been seen as an impediment to progress, either due to reluctance to embrace new technologies or expressing self-centered opposition to novel advancements (Ryghaug et al. 2018).

In the energy domain, the concept of prosumer means a consumer who also produces. The significance of prosumers is that they not only add value for themselves but also other related parties such as their neighbors, other relevant energy industry actors, utilities, and the large social community. Some researchers (Laura et al. 2007; Kirsi et al., 2018) defined the energy prosumer from a different perspective based on how they are involved in the energy field. Prosumers may create energy communities or even virtual power plants (VVP), and in this platform they share or trade energy, which increases the importance of distributed energy resources (DER). Prosumers can contribute to innovation or can be considered as co-creators of innovations by giving feedback, testing products, and as a stakeholder participating in co-development.

The role and contribution of prosumers also need to be considered from the long-term sustainability perspective. Recently, studies on grassroots initiatives also research on community energy became popular, which demonstrates people are becoming more self-organized to make energy production and consumption more sustainable. Researchers (Laura et al. 2007; Kirsi et al., 2018) already addressing the issue that energy research requires a broader approach, encompassing studies that view individuals not merely as passive consumers but as active agents within their culture and society. Thus, this research based on GRETA data will help to accelerate the process of recognizing the differences in approach and attitude that lie within the society.

## 4.3 Unveiling Barriers to Energy Citizenship

To support the energy democracy agenda, some researchers have identified certain barriers that need to be addressed. The necessity of challenging prevailing notions of energy to promote a democratic agenda was emphasized by Lennon (Lennon and Myles, 2017). Instead of viewing energy merely as a neutral natural resource, he argued that it's crucial to recognize the historical and colonial context that has shaped its structure and continues to influence access to energy for different groups. However, from the research, it has been found that barriers to energy citizenship are multifaceted and encompass various dimensions including individual, economic, social, and technical factors (Tobler et al., 2012).

Research suggests (Linda Steg et al. 2015) that individual barriers are largely around a lack of knowledge and energy behavior about energy issues, green energy, and energy transition. The transition towards green energy is relying on renewable energy sources such as solar energy, wind energy, etc. Thus, to which extent and under which condition an individual is willing to accept the new system, needs to understand from various perspectives. Only fifty percent of the population knows that even if today's greenhouse gas emissions were stabilized, the climate would continue to warm for a minimum of another century (Tobler et al., 2012). It's worth noting that, knowledge of energy issues and the transition is higher among educated people, however, the correlations in this regard were not particularly strong. A survey (Marvin Olsen, 1981) has been conducted and asked respondents about the existence and seriousness of the energy crisis. In answer, about 40 to 60 percent of people agreed to a considerable extent that the world is going to face a long-term energy crisis.

Another aspect of individual barriers is economic. To adopt the new energy system and reduce energy use and cost, individuals need to invest in energy efficiency (Tobler et al., 2012). For example, adopting new technologies related to renewable energy, new devices and equipment, renovating existing houses, and adopting green energy appliances. However, research shows (Hugo et al. 2021) that renewable energy systems are not only feasible but already demonstrated economic viability, while costs are reducing every year.

Moreover, existing research (Nouri et al., 2022) demonstrates, that legacy systems and renovation are a big problem for energy transition because, for the renovation, extensive resources are required, which may lead to huge financial costs and longer periods. In some cases, after renovation, the energy efficiency we reach is valued less than the renovation cost. The discrepancy between predicted and actual cost (and efficiency) creates trust issues in the efficiency of energy projects. In addition, there is a legal and institutional framework that creates barriers, for example, urban planning for the building infrastructure may require additional budget which will impact individuals too.

At this juncture, it becomes evident that both social and technical perspectives around the individuals are essential and should be considered for the green transition. For example, collaborative efforts can pave the way for a smoother and more expeditious journey. Thus, individual attitude toward green energy has significance, which has been studied in GRETA for community transition (GRETA 2023a). Detail has been described in section 1.3 (Introducing GRETA analytics).

## 4.4 Behavioral Dynamics of Energy Citizens

In the context of residential energy use decision-making, many models have originated, including traditional and behavioral economics, social and environmental psychology, and

attitude-driven models (Hugo et al. 2021). One attitude model named the ABC Model suggests that attitude comprises three elements- Affect, Behavior, and Cognition (Vishal, 2014). This research work directly employed this ABC attitude model for data modeling and elaborated in the methodology section.

Another research (Yang et al., 2020) explains that, as an integral part of the energy attitude, energy behavior generally refers to two types of behavior, habitual energy-saving behavior and purchasing energy-saving behavior. While the first one helps to reduce energy consumption, the latter one refers to adopting new energy technology or equipment that directly leads to reduced energy consumption without changing behavior. Habitual energy-saving behavior is subjective and depends on the individual. On the other hand, emerging energy technology and technological advancement can produce revolutionary products and equipment that may impact purchasing energy-saving behavior.

To understand the attitude model and its components, need to explore the interrelation and engagements between these components. Emotional engagement (A) is the expression of reactions to a certain event, in terms of energy transition it can be to adopting a new energy technology or to energy-saving behavior (Joseph Murphy, 2007). Emotional response can be again two types, the first one is a positive response and the second one is a negative response. A positive response can promote or support the green transition energy technology, while one negative response will do the opposite. These responses can be considered as a significant signal for the policymakers to decide which technology is publicly accepted or useful for the current situation. The greater decision or impact may have on the future, because as explained earlier, how energy consumers become the energy participant depends on the citizenship attitude towards the green transition (Goda et al., 2018).

Cognitive biases can lead to poor decision-making in a broad range of situations, even under normal conditions. From the research (Yang et al., 2020), it has been found that if a significant amount of money and resources have already been spent on a project, then it’s very likely that people will continue spending more money and resources. This decision is not rational because whether more resources should be allocated or not, this decision should be based on the future potential and outcome not been done before. For the energy of decision-making, this cognitive bias can lead to serious misleading consequences. For example, as discussed above the purchasing energy-saving behavior related to technological advancement and availability, the cost of the product or a project, product durability, etc. in this situation, cognitive bias can be collectively (democratically) or individually, as Johan mentioned. Et al. (February 2023), this behavior is called sunk cost fallacy. On the individual level, this can also negatively impact habitual energy-saving behavior, which means these cognitive biases can lead to negative responses to energy transition from an individual’s habitual facts.

# 5 Research Methodology

## 5.1 Research Method

The main research question is, *what differences in attitudes towards the green transition exist across Europe?* To understand the differences in attitudes, this research follows the existing attitude model, the ABC model described in section 4.4. The research inquiry (or problem) is rooted in the realm of social science, prompting the adoption of a mixed-methods approach for this study (Creswell and John, 1999). The mixed method is one of the latest approaches for social science problems where it needs to combine qualitative and quantitative work (Öhlén and Joakim, 2011). This approach acknowledges the nuanced nature of social phenomena, recognizing that a comprehensive understanding often requires insights from both qualitative exploration and quantitative analysis. However, this is also worth noting that there are three parts of this research (as described in section 1.5, there are three major contributions of this work) and the work is diverse since this is multidisciplinary (explained in section 2).

As presented and articulated in the research methodology, Part 1 (section 5.1.1 annotating with attitude components) is an unnatural science that follows the qualitative method and Part 2 (section 5.1.2 clustering GRETA high-dimensional dataset) is quantitative. But Part 3 (section 5.1.3 framing flustering results with attitude components) itself is a mixed method, research shows that multidisciplinary studies, especially social science studies, where data collection or annotation and analysis are both involved by nature, require a mix of both qualitative and quantitative approaches, which is called the mixed method. This is because of the complex phenomena of our social world (Wang et al., 2014).

At this juncture, it needs to be emphasized the relationship between the three parts. Among these three parts, there are distinctions between research methods, but Part 1, Part 2, and Part 3 are a sequential process that are strongly related and interdependent. This relationship is not only because of the sequential process but also in a mathematical paradigm. This underlying relationship will be visible while eliciting the cluster results for the attitude component in the discussion (section 10).

### 5.1.1 Part 1: Annotating with Attitude Components

GRETA multinational survey (described in 1.3) questions that were asked to people for answers are not defined (or framed) in the attitude components by default. In terms of energy attitude, this is a limitation of the dataset that needs to be followed as an extra step to label (or annotate) the questions with attitude components. Labeling and annotation of variables (columns) according to the attitude components is a qualitative process and an unnatural science, where human

annotation and perception are required to categorize and label the variables in terms of affect, behavior, and cognitive aspects. This part of the study is unnatural because it analyzes and annotates human responses as a human. Details of the execution process of this part have been explained in section 5.2 steps of the research method.

### 5.1.2 Part 2: Clustering GRETA High-Dimensional Dataset

Since GRETA is a multinational survey (details in section 1.3 Introducing GRETA Analytics) and has a large data file with a high-dimensional dataset across Europe, need to cluster the human responses to group together similar data points. Each resultant cluster will represent attitudes that are more similar within that particular cluster and, at the same time, will show dissimilarity or (less similar attitudes) with the other cluster(s). For that purpose, a quantitative approach is a convenient and reasonable method for a robust scientific outcome for the clustering As a method, this research used Machine Learning-based clustering algorithms as a deliberate preference, motivated by the need to extract meaningful patterns and structures from the extensive dataset. Thus, depending on the rigorous explanatory analysis there will be subsequent analysis regarding algorithms, and how algorithms perform on this high dimensional dataset. Machine Learning algorithms will effectively uncover hidden relationships and distinctions within the data, thereby creating a comprehensive analysis and understanding of diverse human attitudes. Notably, this part will answer one of the research questions, *what clusters can be found from data (from a multinational survey) when focusing on answers to affective, behavioral, and cognitive aspects?*

### 5.1.3 Part 3: Framing Clustering Results with Attitude Components

The cluster results from Part 2 need further analysis for framing energy citizenship behavior. Depending on the dominant features (or variable contribution) of each cluster and analysis of clustering results, the attitude components and characteristics from each cluster have been elicited. Post-analysis of cluster results will be quantitative, which will include centroid analysis of each cluster and count (or find) dominant features in each cluster according to the attitude components. However, depending on the quantitative results, further analysis of attitude based on the existing research on energy behavior is again qualitative. At this point, this research will answer one research sub-question which is, *how does one cluster differ from another cluster, or what different types of knowledge are found - how/why are these different?*

As explained above this research methodology involved multiple research methods that included data annotation (or data collection), model development for clustering, and analysis of results (interpreting data and insights), this research particularly follows the mixed-method method. The steps of the research method will be explained in the following section (5.2).

## 5.2 Steps of Research Method

The three parts of the research have been explained in the previous section, those are interlinked and interrelated. This research will follow a scientific and structured process to get the proper outcome.

*The first step* of this thesis work was entirely hands-on. This thesis started the work by manually selecting columns from the dataset according to the attitude component as described above in Part 1.

One of the best-cited attitude models is the ABC model which suggests three different elements of attitude i.e. Affect, Behavior, and Cognition (Vishal, 2014). Thus, from the attitude model, an understanding of relevant information from the dataset has been found. In the context of the GRETA energy dataset, this research work had to apply a manual sorting process of the existing dataset and in the result, each column is a unit piece of information regarding attitude. Where each column in the dataset represents a unit response (answer to a single question) and each row represents the entire response to all questions from a single person or user. For this research, each variable has been meticulously assessed for its significance and pertinence. A three-person team conducted this manual evaluation, and the team's decision to include a variable in the modeling was contingent on unanimous agreement. Furthermore, the team categorized the variables into one of six classes. These categories are A = Affect, B = Behavior, C = Cognitive, E = Explanatory factor, T = Target for prediction, _BLANK = Irrelevant. After careful selection and annotation, the team found that there are 1179 variables (columns) that are irrelevant (E, T, and _BLANK) to our clustering work. After removing those, the resultant dataset has 344 variables altogether (A, B, C). This categorization and selection of variables are based on the prior knowledge and intuitive understanding of the team members. Our knowledge and intuitive understanding were mostly from two different perspectives- 1) the team's empirical and prior knowledge about energy citizens, green transition, and energy personas, and 2) knowledge of Machine learning-based data clustering models. There were some reasons behind this manual selection: 1) not all variables are related to the attitude and our clustering goal is focused on attitude towards green transition 2) useless features in a clustering model negatively impact the clustering (Sewell et al., 2005). The resultant dataset is open, and the download link can be found in Appendix 1.

*The second step* was to work on data pretreatment, this process was required for data cleaning and preprocessing. Missing and nosey information removal is a basic requirement for machine learning analysis (Emmanuel et al., 2021). In addition, the dataset contains text information that needs to be encoded according to the ordinal (or hierarchical) structure. This research followed a systematic way of data pretreatment, which has been explained in section 7 (data preprocessing).

*The third step* was to find the appropriate clustering model. This was challenging for several reasons. Even though this work went through a heavy pretreatment process, the number of input features in the final dataset was still higher. High-dimensional dataset reduces the model performance and reliability (Aremu et al., 2019). To overcome this, an explanatory analysis was done, identified the variable types, also dominant features of the dataset, and eventually, three types of clustering algorithms were applied: Partition-based clustering, Hierarchical clustering, and Density-based clustering. How this research work selects appropriate algorithms is described in section 8 (clustering with ML models).

*Fourth step*, this research analyzed the results using spatial analysis aligned with the particular algorithm (or method) and established the relationship for each cluster with the attitude model and its components in the context of energy behavior and energy citizenship. In addition, manual checking and verification have been done with the results from each clustering algorithm. More details are described in section 9 (results) and section 10 (discussion).

*In the fifth and final step*, this research continued to find the relationship of the results with the previous research theories and findings. Analyzing this research results with the existing scientific publications gives more insight into the context of energy attitude, (details in section 10.4)

The following Figure 2 shows the complete breakdown of the steps of the research methodology:

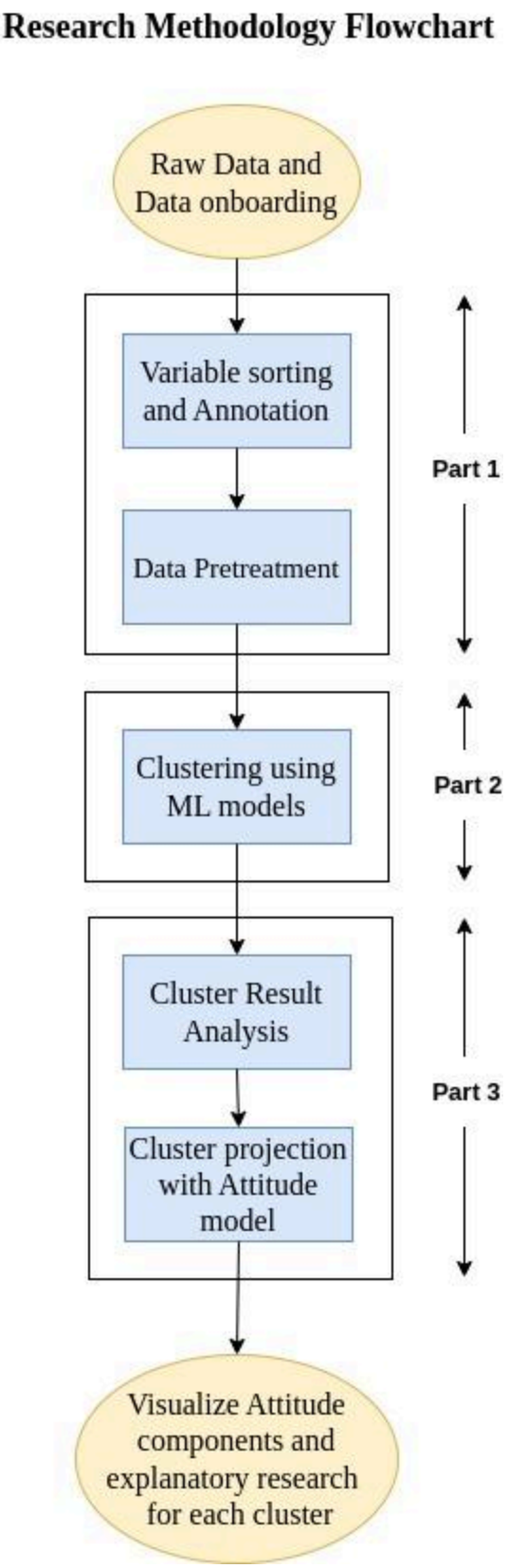


*Figure 2. Steps of Research Method*

# 6 Technology Used

As described in the research methodology (section 5) this research work was based on technical implementation of the theoretical concept and the methodology evolved around the research objective and questions.

The primary programming language was Python, and its libraries such as pandas for reading and navigating through the data and NumPy for faster data processing. Plotly, matplotlib and seaborn are used for visualization and generating charts, figures, etc. For machine learning models and algorithms, sklearn library has been used. In addition, this research leveraged engineering techniques for data loading, streamlining preprocessing, and expediting the training of ML models, thereby simplifying the development process. A laptop equipped with a basic four-core processor and 16GB of RAM proved sufficient for training the machine-learning models due to the manageable size of the dataset after preprocessing.

Processed dataset and the entire implementation with results (charts, figures, tables, etc.) released through the GitHub repository, which can be found through the link on reference (Clustering-GRETA, 2023) and in Appendix 1.

# 7 Data Preprocessing

For the clustering purpose, we followed the standard process of data pretreatment. Most of the reasoning and steps are theoretically explained in section 5.2 (steps of research method) with the appropriate research method. Here, the implementation of the preprocessing steps will be described with proper reasoning and substantial findings.

## 7.1 Basic Stat of the Dataset

GRETA data is multinational survey data (introduced in section 1.3), gathered from a survey conducted across 16 European countries.

Here are the main features and structure of the dataset:

- There are 1503 columns and, 10488 rows in the initial raw dataset
- Each row (all cells together) represents each person's responses to all questions
- Each column represents the answer to each question, and the headings of the columns represent questions
- There are missing data points explained in section 7.3

From the statistics above, this dataset has numerous attributes or columns, GRETA data is high dimensional, which has specific challenges for the clustering process and algorithms because it creates computational problems while increasing dimensionality and research shows that in some cases some algorithms may become ineffective (Ira Assent, 2012). Thus, this research had to follow preprocessing steps and some of the reasons (and purposes) already mentioned in the methodology (section 5.2). In addition, there were missing data points, and those needed to deal with standard processes to adopt the clustering algorithms. Since this was a high dimensional dataset, it was essential to find the variable (or column in the dataset) type (categorical and continuous) this information is useful for selecting the clustering algorithms. For example, in the case of continuous data K-means works better, on the other hand, in mixed data (categorical and continuous both in a single dataset) K-mode may work better (Manisha et al., 2017). The following sections describe the preprocessing steps to align the entire implementation according to the research objective and as mentioned in the methodology, and Figure 2 demonstrates the flow of the implemented work.

Here are more descriptions of the pre-treatment steps followed:

## 7.2 Step 1: Sort data according to pre-defined criteria

Research methodology (section 5) describes the manual sorting and annotation of the variables of the dataset with the attitude components. Predefined categories and criteria are explained in section 5.2 and based on those, the dataset has been filtered using Python programming which contains 344 variables. The dataset can be found in Appendix 1.

## 7.3 Step 2: Handling missing values row-wise

This dataset's row-wise records are responses from a specific person, each row represents all the answers to questions (each cell one answer) from a single person. Since row-wise missing are from a single response/person, if most of the values are missing in a single row (in this case more than 70%) then those have been removed. A total of 1074 rows are more than 70% missing, which means more than 70% of questions were not answered by 1074 users or people (10.24% = 1074 / 10488). After removing 1074 rows now the dataset has 9414.

## 7.4 Step 3: Handling missing values column-wise

Previously step 2 counted missing, row-wise, or horizontally, now at this stage, count missing vertically or column-wise. For that, checked the missing values (or null values) and counted those programmatically. And removed columns that are more than 25% missing. Found that a total of 266 columns and each of them has more than 25% missing values. After removing those, the dataset has 58 columns or variables remaining. The resultant file can be found in Appendix 2.

## 7.5 Step 4: Sanity Checking of the Dataset

This work conducted a thorough sanity check and once again verified the presence of any missing values in the dataset. At this point, it's worth noting that the dataset is free of any missing values, which is a positive development, sparing the task of imputing any data. In total, the dataset comprises 9,414 rows and 58 columns. Building on the information from Step 1, we've identified three distinct question types represented by the columns labeled 'A,' 'B,' and 'C,' with 25, 13, and 20 occurrences, respectively.

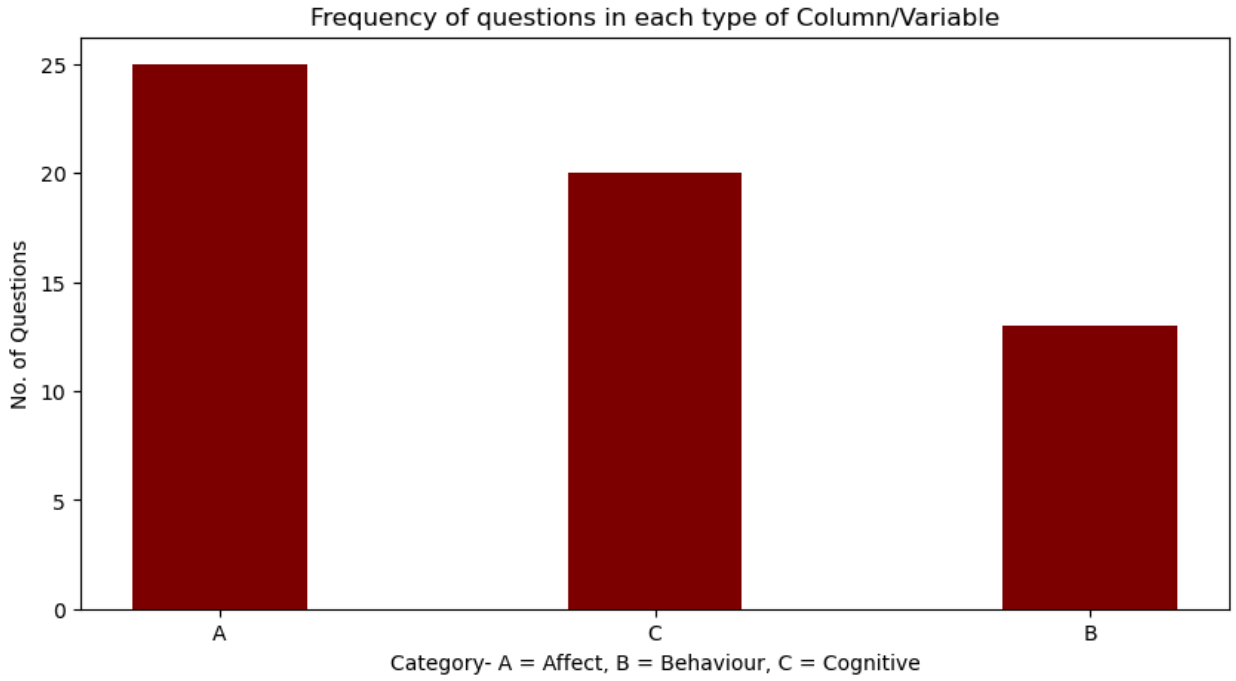


*Figure 3. Number of questions in each category (lowest number of questions related to behavior)*

## 7.6 Step 5: Identifying the Types of Variables

This research used heuristic algorithms and programming to identify categorical and continuous variables and verified results by manual checking. Heuristic algorithms were an effective choice because of their nature and implementation simplicity, for example, if the number of unique values in a variable or column is less than 10, the variable is categorical in the context of around ten thousand records. This method works because it determines near-optimal solutions to an optimization problem (Anmol Singh, 2020). However, the finding is that all the variables are categorical variables, with no continuous variable in the dataset. This is understandable because

the responses (dataset variables) or GRETA data collected from people are mostly categorical, the interview questionnaire was designed that way.

However, this work has also used heuristic programming and manual inspection through the dataset for different types of categorical values. Here are the types of values for columns-

- All values are numeric - the number of columns is 8
- All values are numeric except one (not prefer to express an opinion) - the number of columns is 27
- All values are text and need variable encoding - the number of columns is 23

After summing up, the total number of variables is 58 (=8+27+23).

Each type of variable and corresponding unique values can be found in Appendix 3.

## 7.7 Step 6: Variable Encoding

From the previous step 5, it has been known that in the dataset, there are three types of variables. The third type has all text data are total of 23 variables. These texts need to be encoded using some numerical values.

Here is the encoding applied (list of the text and corresponding values) (Table 1):

| No. | Text | Encoded Values |
|---|---|---|
| 1 | ['Fully trust','Tend to trust','Tend not to trust','Fully distrust', 'Undecided'] | [3, 2, 1, 0, -1] |
| 2 | ['Owner, no outstanding mortgage or housing loan','Owner, with mortgage or loan','Tenant, rent at market price','Tenant, rent at reduced price or free','Other, please specify:', 'Do not know / prefer not to say'] | [4, 3, 2, 1, 0, -1] |
| 3 | ['Always', 'Often', 'Occasionally', 'Rarely', 'Never', 'Prefer not to say'] | [4, 3, 2, 1, 0, -1] |
| 4 | ['I use this to follow energy related information', 'I use this but not for energy related information', 'I do not use this at all'] | [2, 1, 0] |

*Table 1. Unique text and corresponding encoding*

Following the application of encoding techniques, proceeded to explore the variable values by generating a box plot within the programming environment. This approach afforded a valuable

visual perspective on various aspects of the data, including the distribution of variable values, their central tendencies (such as the mean), and the presence of extreme values or outliers. The box plot, as an essential data visualization tool, serves as a bridge between raw data and actionable insights, enhancing the understanding of the dataset's characteristics and aiding in the identification of potential anomalies. Detailed images and information can be found in Appendix 4. Here is the shorter version of the box plot with 10 variables:

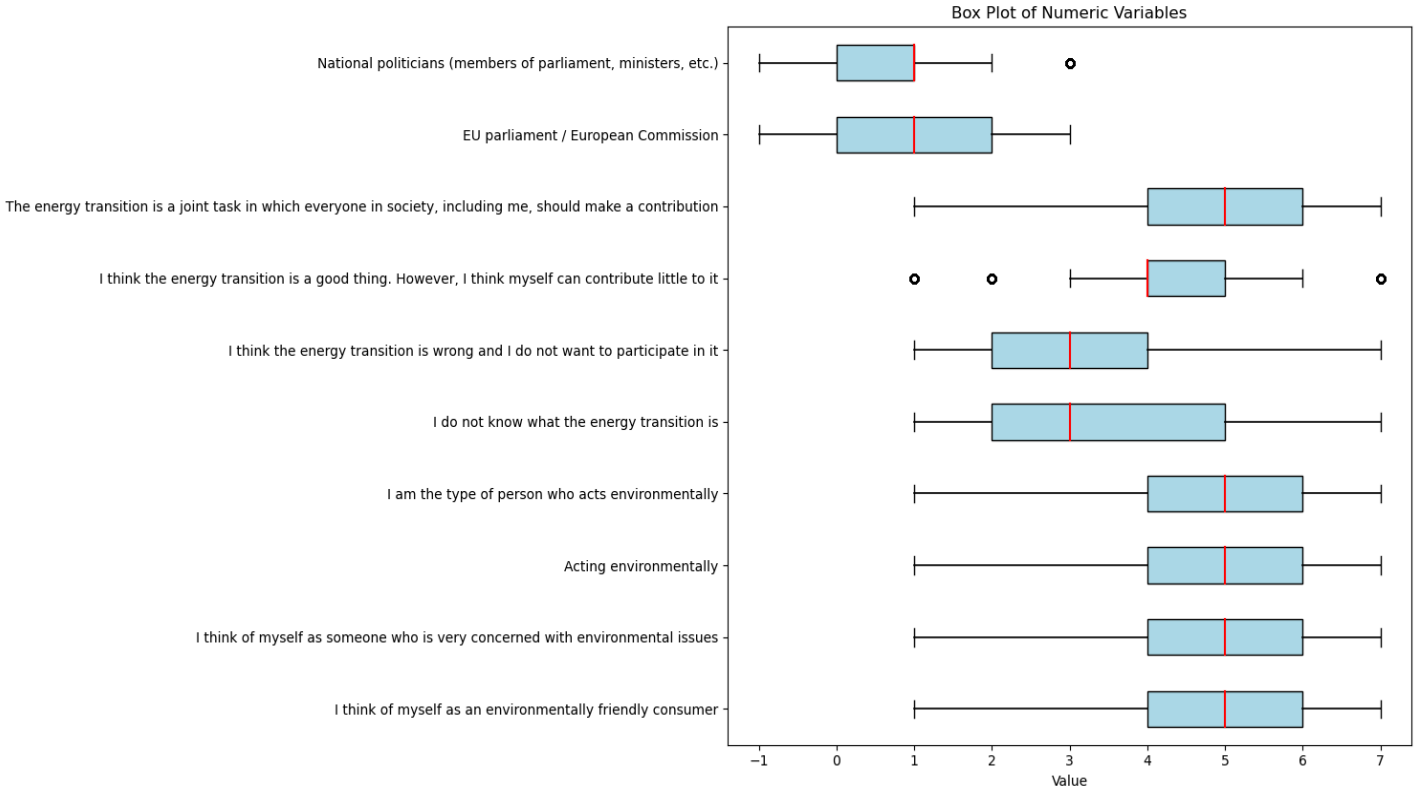


*Figure 4. Box plot of each variable (detail - Appendix 4)*

From the box plot, get the impression that the processed dataset (for all variables or columns) has a variable value range from -1 to 7. Here note that each data point represents a single response from a person for a particular question of the GRETA survey, and each question represents the heading of the column (detail of the data is in section 7.1).

Counted the frequency of each value in the entire dataset, and here are the bar charts:

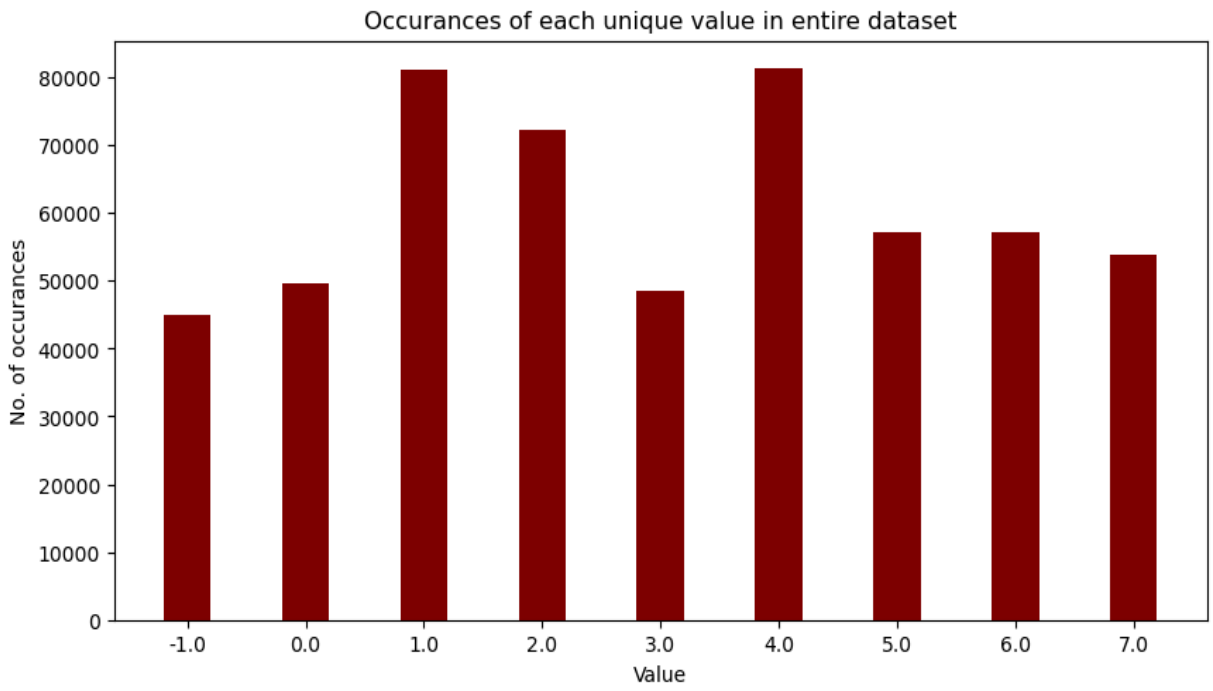


*Figure 5. Number of occurrences of each unique value in the entire dataset*

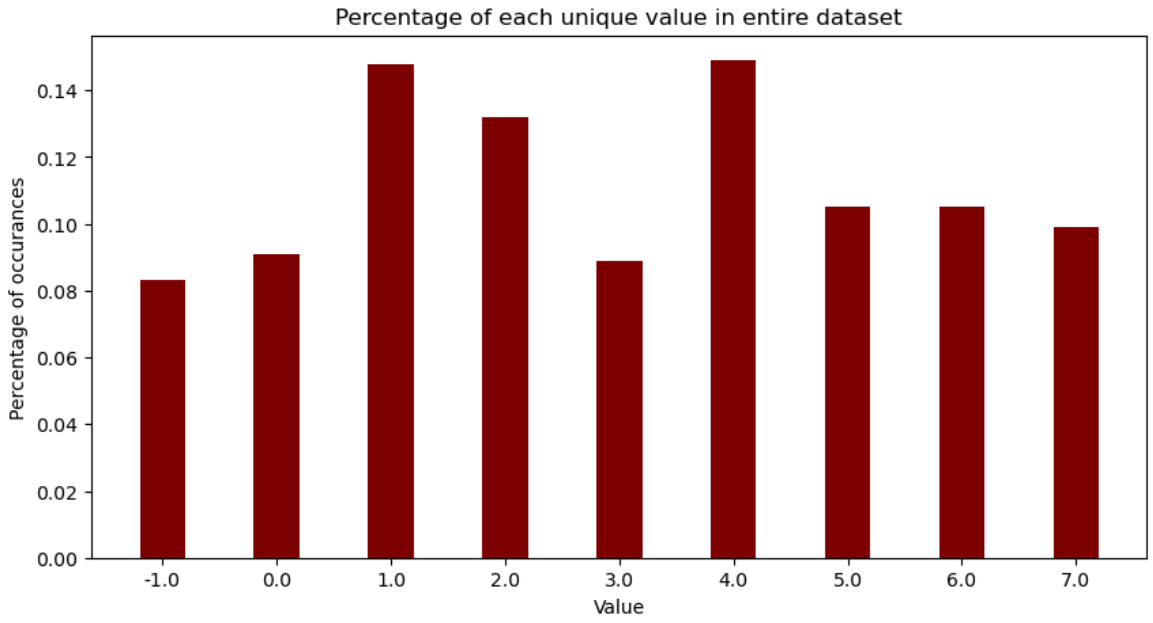


*Figure 6. The relative appearance of each unique value*

From the bar chart, it's observed that values 1 and 4 have the highest frequency.

Table 1, Figures 4, 5, and 6 all together give an impression of the shape of the data, the variable value ranges, and the frequency of each categorical value in the entire dataset. This information will be useful when selecting the clustering algorithm (section 8) because the shape of the data and its values have an impact on the nature of algorithms with fundamental mathematical equations and algorithmic processes that are applied to a clustering algorithm (Manisha et al., 2017).

## 7.8 Reasoning behind the Ordinal Encoding

Data collection of the GRETA survey was based on the order or category. The user had to select an option from multiple options for a single question. Additionally, there were 35 questions whose values are already numeric but represent ordinal variables. By nature or by design, the survey was based on selecting a value from multiple options or sometimes rating an answer by selecting an ordinal value. Unique values of each variable can be found in Appendix 2.

## 7.9 Step 7: Correlation and Dimensionality

A correlation is a table that represents a matrix, displaying correlation coefficients between variables of a dataset, to find the matrix of correlation among all 58 variables selected in the previous steps. This correlation matrix illustrates how strongly pairs of variables are related to each other. The coefficients range between -1 and 1, which indicates the strength and direction of the relationship between variables:

- A coefficient of 1 illustrates a perfect positive correlation (as one variable increases, the other increases proportionally).
- A coefficient of -1 demonstrates a perfect negative correlation (as one variable increases, the other decreases proportionally).
- A coefficient of 0 suggests no linear relationship between the variables.

The correlation matrix allows for a quick assessment of relationships between multiple variables and helps identify potential patterns or associations within the data.

Here are the strongly correlated variables:

```
'I think of myself as an environmentally friendly consumer - To what extent do you agree with the following statements?',

'I think of myself as someone who is very concerned with environmental issues - To what extent do you agree with the following statements?',
```

'Acting environmentally-friendly is an important part of who I am - To what extent do you agree with the following statements?',

'I am the type of person who acts environmentally-friendly - To what extent do you agree with the following statements?',

Here is the correlation coefficient matrix based on the encoding above:

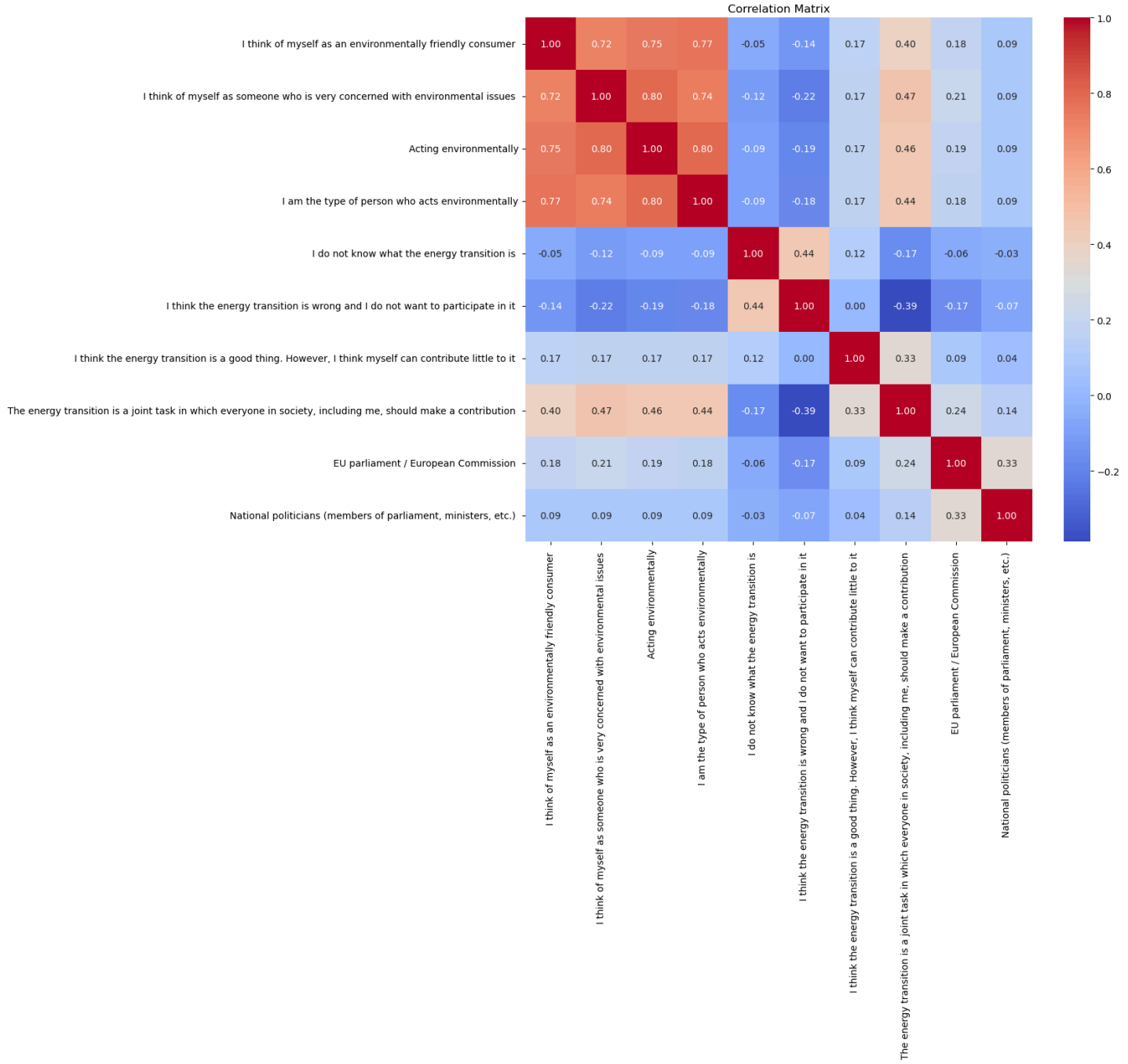


*Figure 7. Correlation Matrix (detail - Appendix 4)*

**Chi-Square Test matrix**

Understanding the heatmap:

- Dark Red: Strong association or dependency.
- Light Red: Moderate association.
- Light Blue: Weak association or independence.
- Dark Blue: Strong independence.

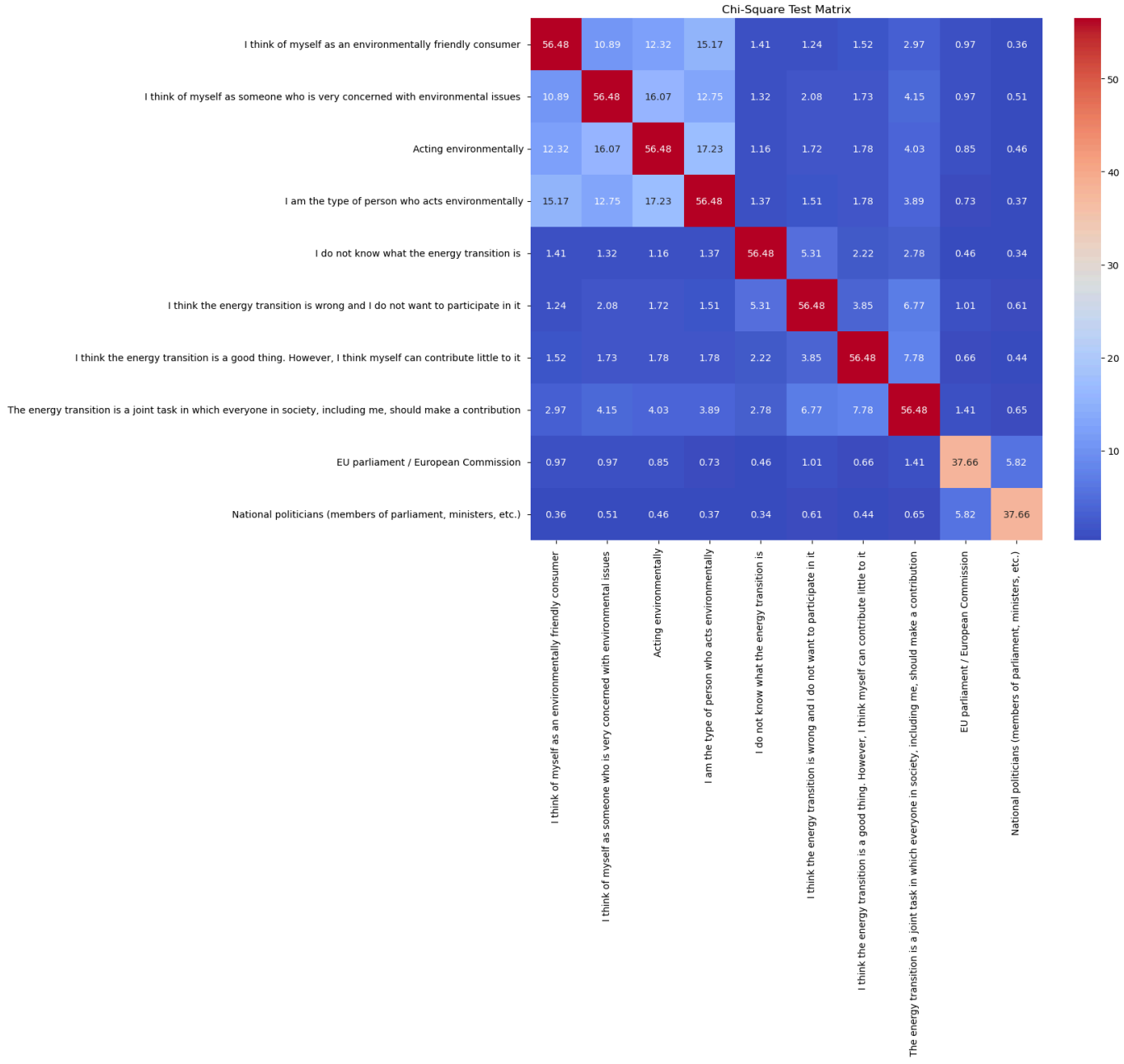


*Figure 8. Chi-Square Test Matrix (detail - Appendix 4)*

In the above Chi-square contingency matrix (Figure 8), the diagonal elements represent the count of observations or the records where the same variable analyzes itself, and don't indicate the strength of association or independence between variables. However, the off-diagonal elements, determine the level of association or independence between different categorical variables along with the statistical test (Chi-square). Here it's different from the correlation matrix because, in the correlation matrix, the diagonal elements are always 1 as they show the correlation of a variable with itself, which is perfect (a variable is perfectly correlated with itself).

From the Chi-square test matrix, a similar understanding or impression has been found that is illustrated by the correlation matrix. If the variable has a strong relationship or correlation, that means it has more dependency, too.

## 7.10 The Curse of Dimensionality

From the correlation coefficients, the relationships were acceptable because, the maximum correlation is 0.80, and didn't remove any variables. Thus, the dataset has 58 variables in the resultant dataset which will be used for further processing and clustering.

This research already conducted a heavy pretreatment process explained by previous preprocessing steps that reduced the number of variables to 58 (at the beginning it was, 1503). Yet, the dataset has a significant number of variables that are challenging for clustering with Machine Learning (ML) algorithms.

Variable values that will be used for the ML models range from 0 to 7 (Figures 5 and 6). Since the number of variables is 58 and the value range is only between 0 and 7, it concludes that the differences between objects are less or delicate. Thus, finding similarity or dissimilarity is difficult, in other words, the ML model might be confused, and it might be prone to consider objects alike rather than dissimilar. In addition, as the number of variables increases, it creates or increases complexity exponentially because of the combination of variables, processing steps, and related issues with space complexity (Ira Assent, 2012).

## 7.11 Step 8: Prioritization of Variable

Prioritizing variables or weight variables based on their importance is required because the model will be more influenced by those variables. The variable's importance in a model can be identified by the variance of a variable, the more the variance is, the more importance it has for a model.

PCA is most relevant here to understand the data variability or statistical variance because the nature of the initial data collection and inherent data structure is organized in that way. To be more specific, each row represents the response of a human to all the questions. This row-wise alignment of input data allows PCA to capture the variances according to the hierarchical order. In PCA, the first component captures the first most variance, the second component captures the second most variance, and so on (Svante et al., 1987).

The mathematical representation of PCA is,

PCA = Scores * Transposed loadings + Residuals (error and noise)

Cumulative variance represents the accumulation of variance explained by each successive principal component in a dimensionality reduction technique such as PCA.

In PCA, the components are ordered based on the amount of variance they explain in the original dataset. The cumulative variance represents the total amount of variance explained by a certain number of principal components, starting from the first one and continuing through each subsequent component.

**Cumulative variance of 58 components**

Figure 9. Shows the cumulative variance of 58 variables. Observation is that the top 20 components comprise more than 70% variance, which has a separate projection in Figure 10.

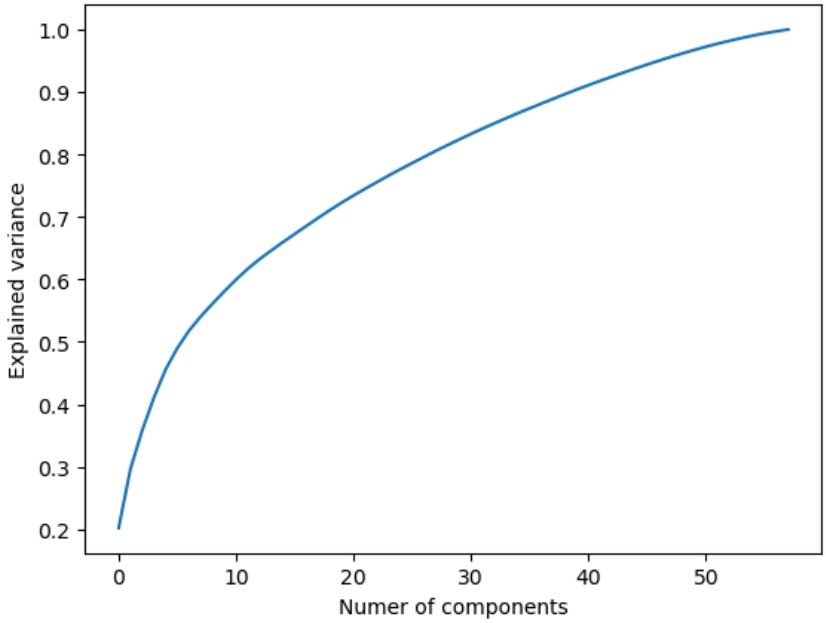


*Figure 9. Cumulative variance of 58 components*

**Cumulative variance of 20 components**

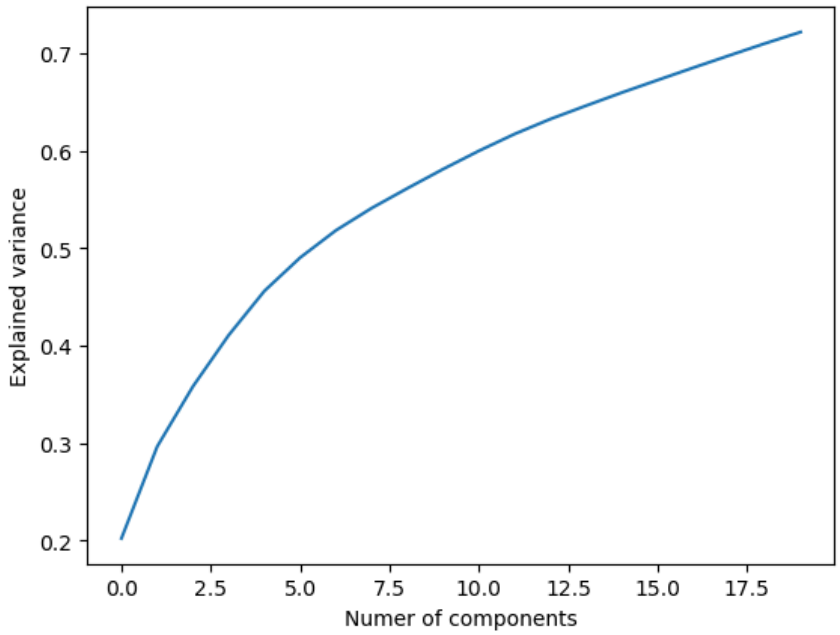


*Figure 10. Cumulative variance of 20 components*

## PCA Analysis with 58 Components

## Total Absolute Contribution (Magnitude Only)

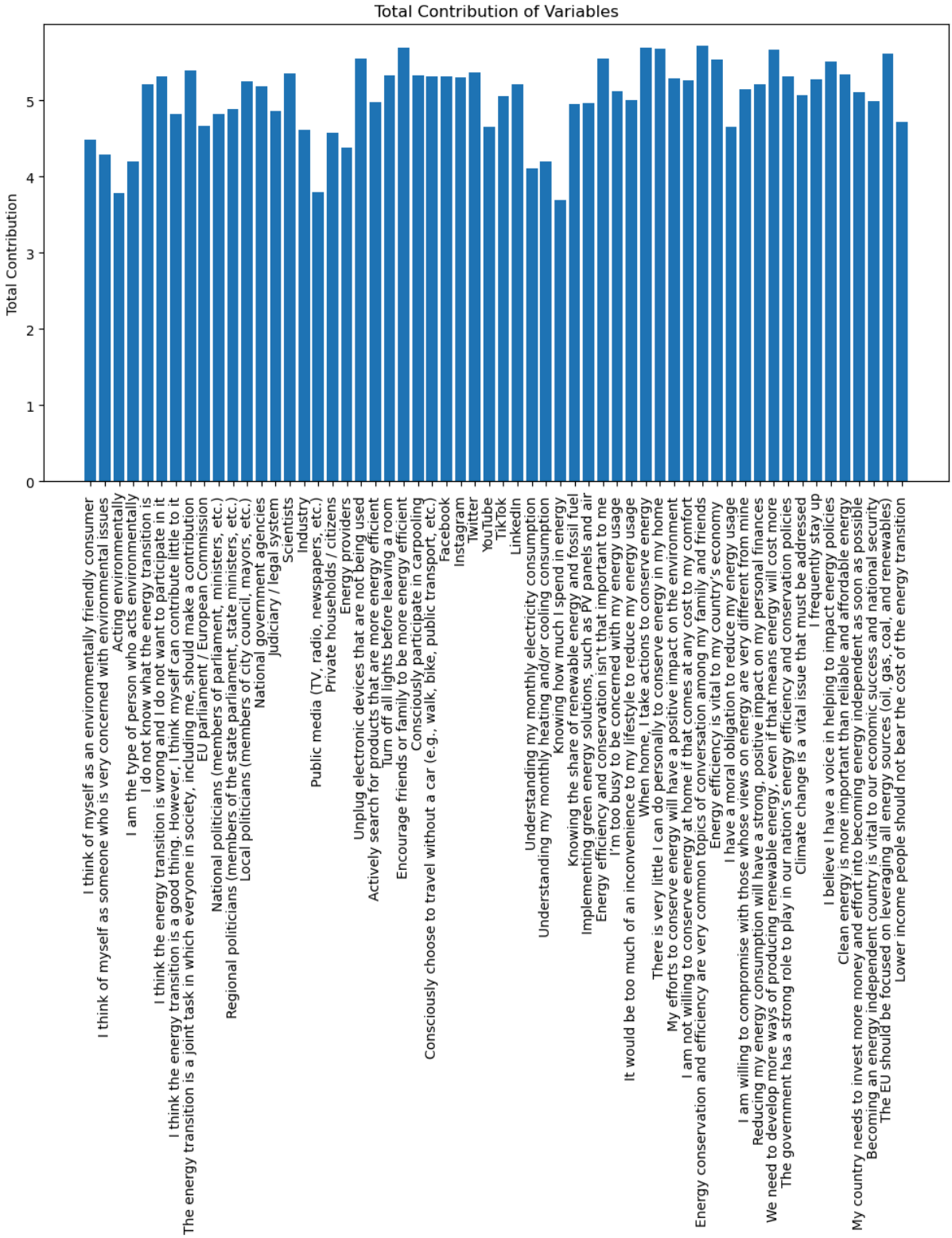


*Figure 11. Each variable total contribution in PCA (magnitude only)*

## With Magnitude and Direction

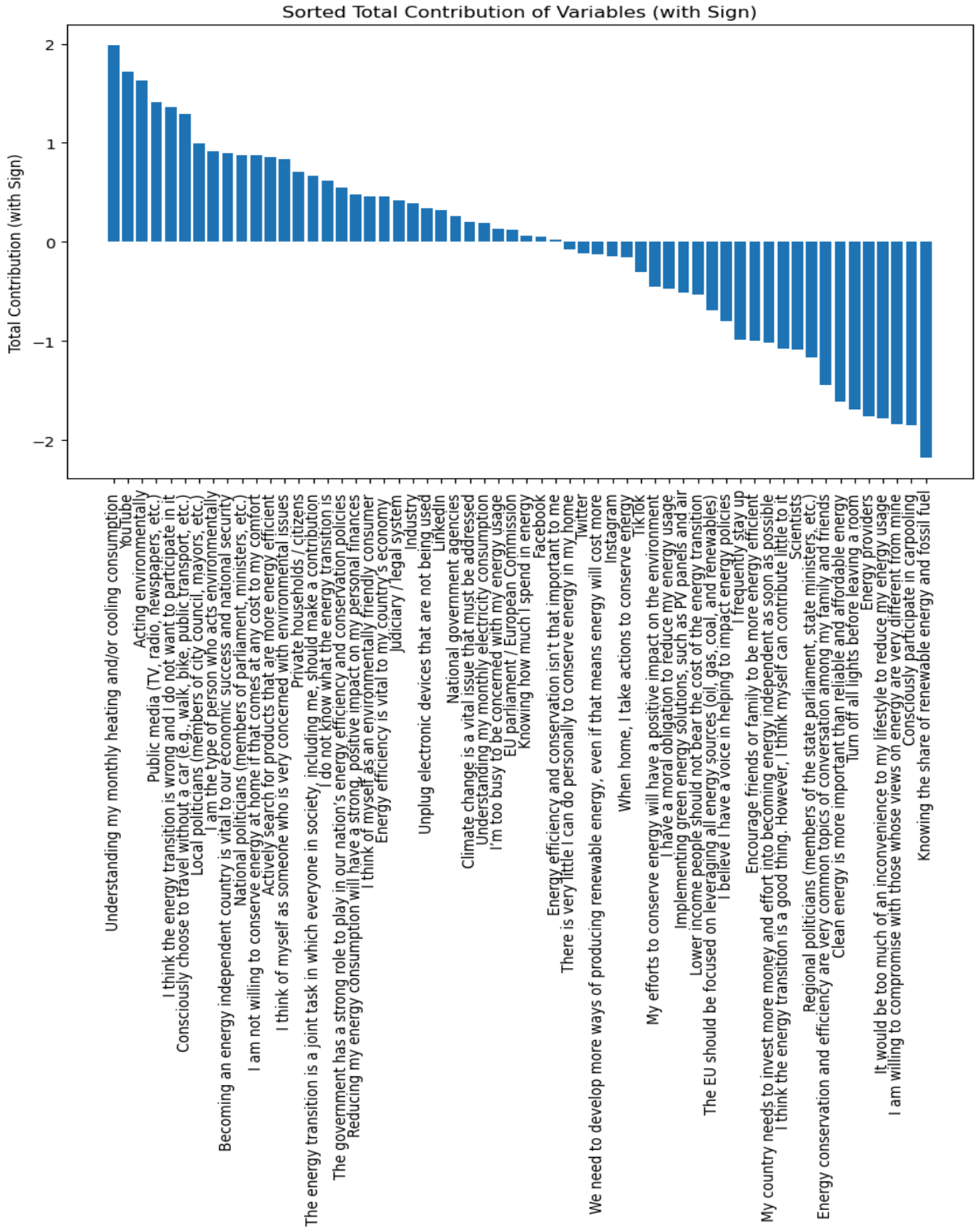


*Figure 12. Each variable contribution in PCA with the sign*

## PCA Analysis with 20 Component

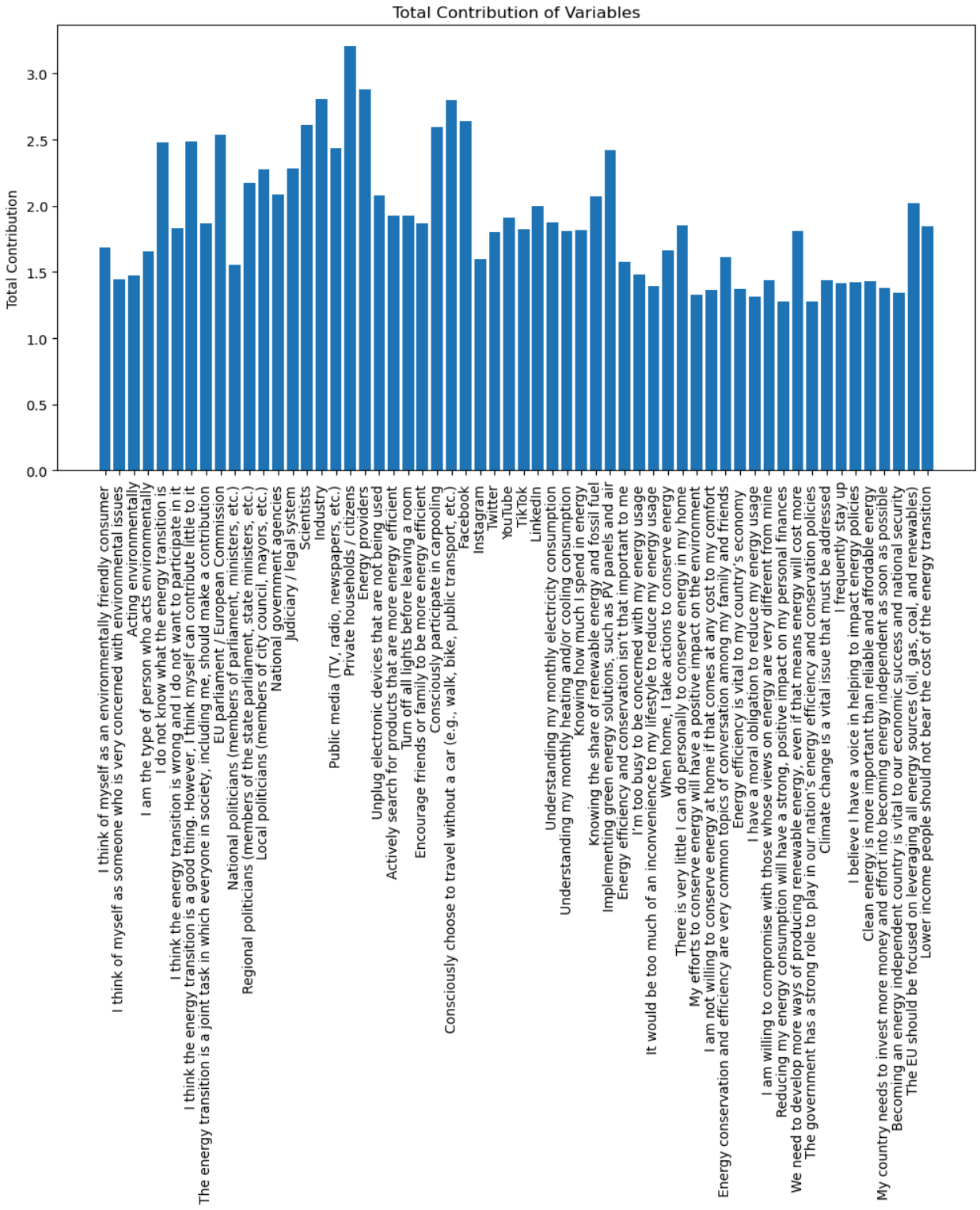


*Figure 13. Each variable total contribution in PCA (magnitude only)*

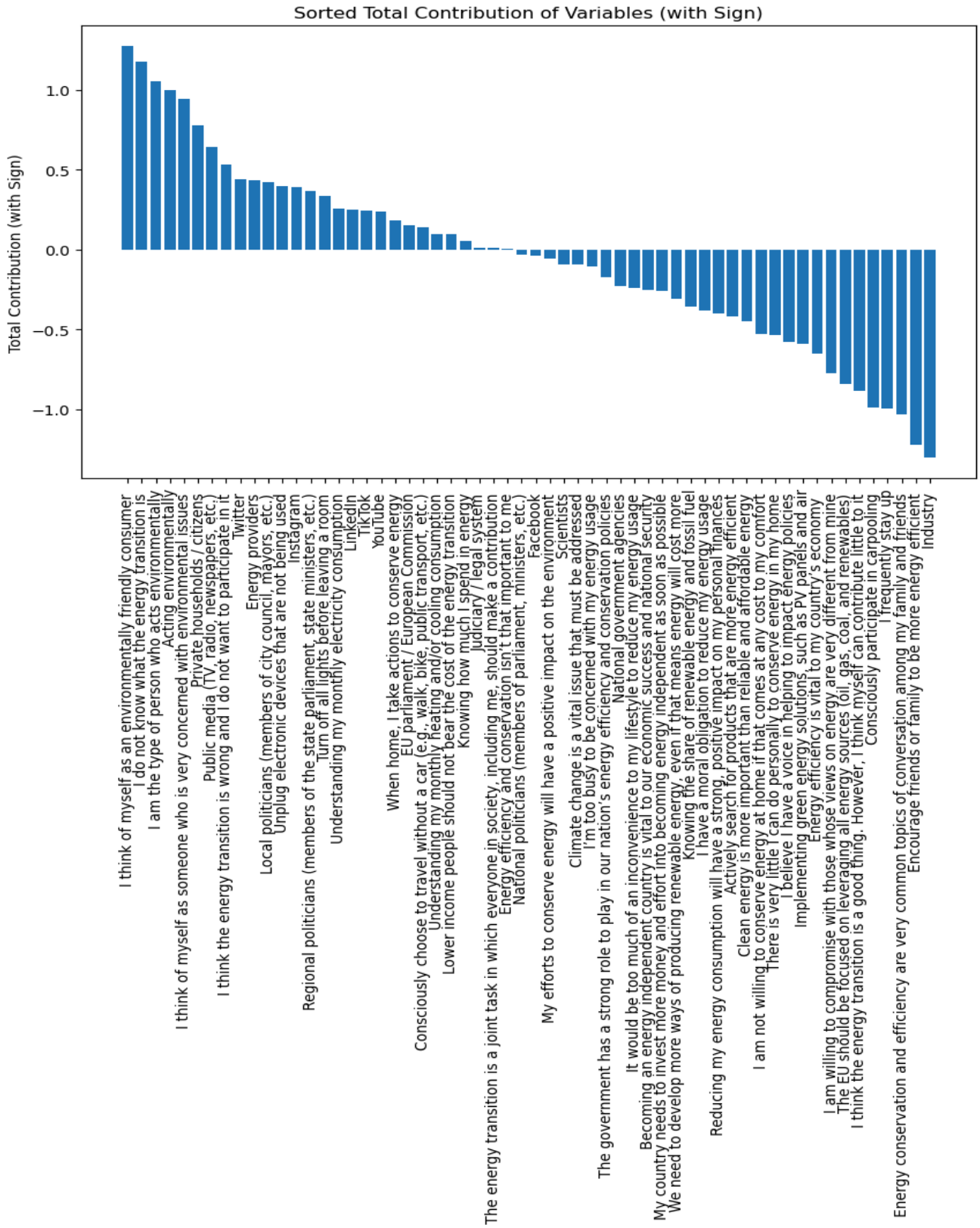


*Figure 14. Each variable contribution in PCA with sign*

### 7.11.1 Prioritization of variable: Analysis and Outcome

From the previous step (section 7.11), gradual increase and changes in cumulative variance are observable from PCA analysis, cumulative variance (Figures 9 and 10) also from the variable contribution to PCA (Figures 11, 12, 13, 14). PCA illustrates a hierarchical order of importance or variance. As the changes in the variance curve are very little, such as the first 20 components comprising 70% variance and the rest of the others comprising still 30% variance, the conclusion is that all 58 variables have importance for the model. Thus, the variable contributions are significant, and reducing components or variables may lead to the loss of valuable information. Moreover, already a significant number of variables has been removed in the preprocessing steps, further reduction may encounter inappropriate or bad results. There are some more elicitation and decisions made around the PCA and its analysis, which will be explained in the next section (7.12).

## 7.12 Normalization

For this research work and data clustering, normalization has not been applied, there are a couple of reason behind it.

Firstly, because of normalization, the model might lose the granularity of the data. From the previous steps, it has been found that almost all the variables are ordinal data ranging between -1 to 7. Ordinal data has inherent rank or order, which is important for the model. For the same reason, work on outlier removals was not required because it's already known that the range of the entire dataset is -1 to 7.

Secondly, normalization could make it harder to interpret the results of a machine learning model, as the inputs will be on a common scale, which may not align with the original scale of the data.

Thirdly, variance set to 1 means all variables are equally important because PCA centers the data around 0 and variance set to 1. But in the dataset, as explained in previous sections, data are ordinal and importance is hierarchical.

# 8 Clustering with ML Models

The primary goal of cluster analysis is to identify the inherent patterns of the data within the collection of points, and entities, in other words, this can be called natural grouping (Anil Jain, 2009).

Webster defined cluster analysis as a statistical classification technique to identify whether the subjects of a population fall into different groups based on the quantitative comparisons of multiple characteristics (Merriam, 2018; Anil Jain, 2009).

To elicit the operational definition, it can be said that, out of N number of objects finding similarities between objects and putting them into K separate groups according to their similarities. Now the definition of similarities can be debited and varied based on the predefined criteria. However, the ideal cluster can be determined as a set of points that is compact and isolated (Anil Jain, 2009).

## 8.1 Algorithms

To start with, this research work tried K-Means, a widely used, most common partition-based algorithm. For K-means, elbow analysis and silhouette analysis have been applied, which gave an impression of the possible clusters. Based on that, it has been inferred the results from the model and projected on plots to do further analysis and quality of the cluster. K-Means results are described in section 9 (results).

From the K-Means results and post-analysis, it can be inferred that this algorithm is not the most suitable algorithm for the model. To overcome that, this research applied the K-mode algorithm, which is an extension of the K-Means algorithm and is designed to deal with categorical data (Manisha et al., 2017).

To get more insight and better illustration, Hierarchical clustering and Density-based clustering have been applied, which are also included in section 9 (results).

# 9 Results

In this section, the results of data clustering models have been described for each type of clustering algorithm mentioned in section 8.1. Explanatory analysis, model selection, best models, and the outcomes will be presented with charts, figures, and supporting materials. However, further analysis and discussion will be presented in discussion section 10. Detailed implementation of algorithms and Python programming code can be found in Appendix 1.

## 9.1 Partition-Based Clustering

The partition-based clustering method starts from an initial cluster and then relocates instances from one cluster to another depending on the new calculation. For these types of algorithms, the

number of optimal clusters must be predefined by the user (Lior, 2010). A greedy heuristic iterative approach has been used for optimization purposes.

### 9.1.1 K-Means Clustering

In the research community, K-Means is the most commonly used and powerful algorithm. This partition-based algorithm required three parameters from user input, the number of clusters K, cluster initialization, and a distance metric (Lior, 2010). Despite extensive use, K-Means has some limitations, for example, random initialization of the centroids may lead to unexpected convergence (Ahmed et al., 2020).

The objective function for K-Means clustering involves minimizing the sum of squared distances between data points and their respective cluster centroids:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2$$

In this formula:

*J* represents the total dissimilarity or cost function.
*K* is the number of clusters.
$C_i$ represents the data points in cluster *i*.
*x* is a data point in a cluster.
$\mu_i$ is the mode (centroid) of cluster *i*.
$|x - \mu_i\|^2$ calculates the squared Euclidean distance between a data point *x* and the centroid $\mu_i$

#### 9.1.1.1 Identifying the Number of Clusters

The main challenge for a clustering algorithm is to determine the number of clusters or the number of model parameters, which must be determined before clustering (Trupti et al., 2013). For that, several techniques and a series of experiments with the algorithms have been applied. Two common approaches are elbow analysis and silhouette analysis, generated charts for different numbers of clusters. For this research and experimental work, the results with the figure have been visualized, which gives a visual impression regarding the data points and clustering outcome.

#### 9.1.1.1.1 The Elbow Analysis

The elbow method is a commonly used method for determining the optimal clustering algorithms, although a manual inspection is required after plotting the elbow points on a chart. Elbow analysis includes a plot comprised of the number of clusters against the cluster quality or distortion. The goal is to identify the significant change in slope called 'elbow'.

From Figure 15 the possible optimal number of clusters for the dataset can be observed as 3 or 4.

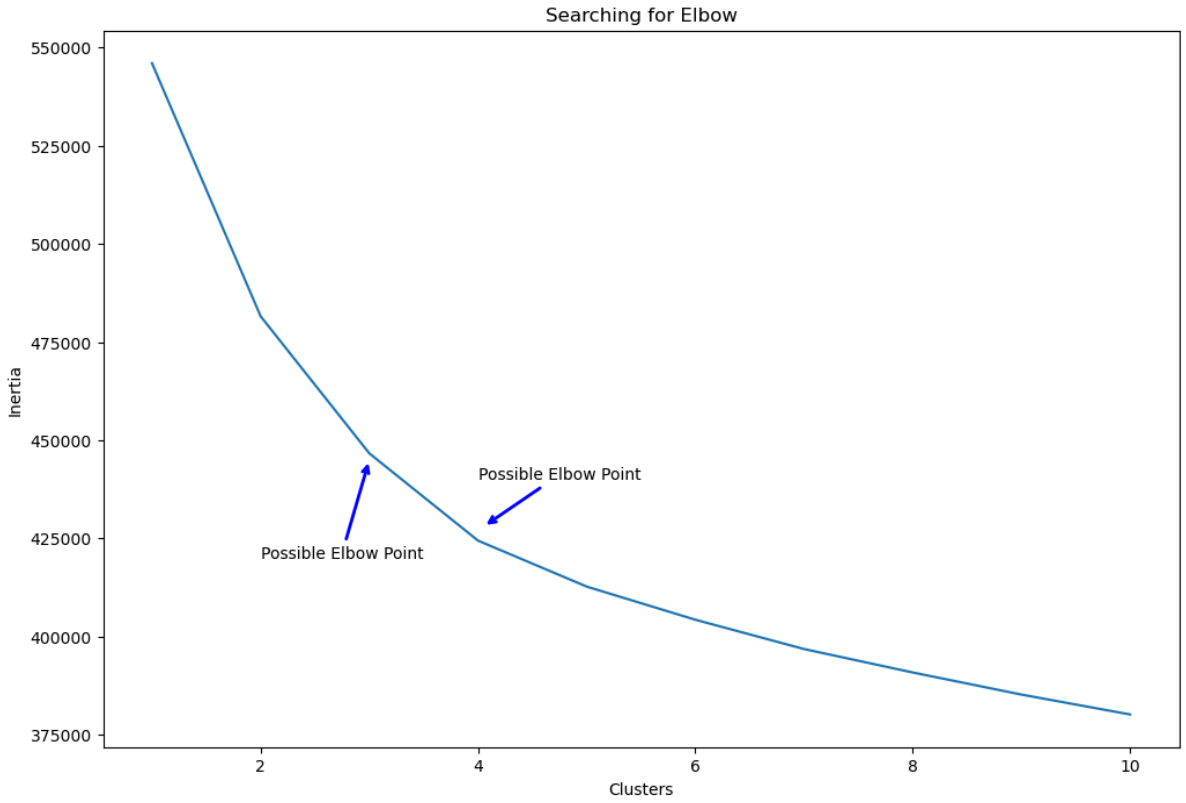


*Figure 15. Elbow Analysis- determining an optimal number of clusters for K-means*

#### 9.1.1.1.2 Silhouette Analysis

The Silhouette analysis method is used to assess the quality of a cluster. It measures the similarity between the objects of its own cluster to the other clusters. The silhouette score ranges from -1 to 1. The silhouette score is calculated on every data point and then calculated for the entire dataset, which provides the overall score for a particular cluster.

Here are the following results for silhouette analysis:

For n_clusters = 2, the average silhouette_score is 0.1338137732045716

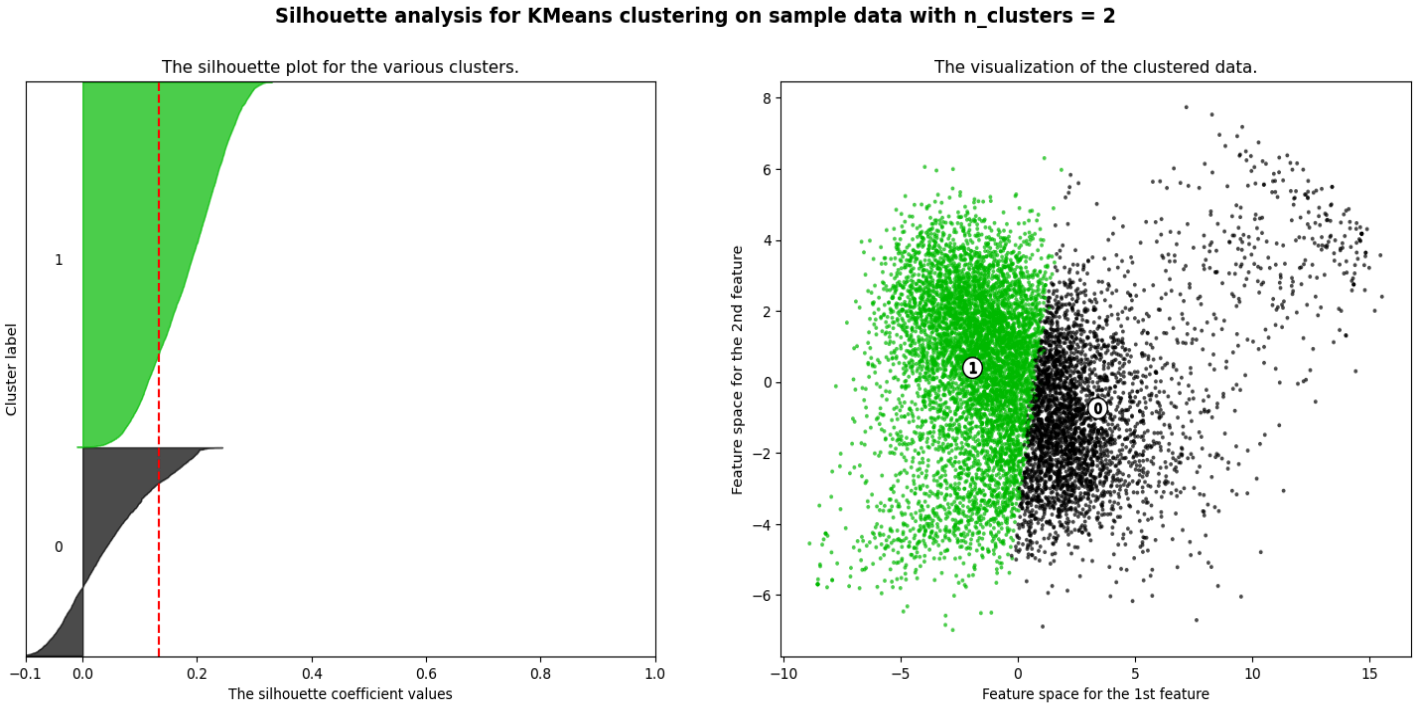


*Figure 16. Silhouette Analysis - n_clusters = 2*

For n_clusters = 3, the average silhouette_score is: 0.11003396999840832

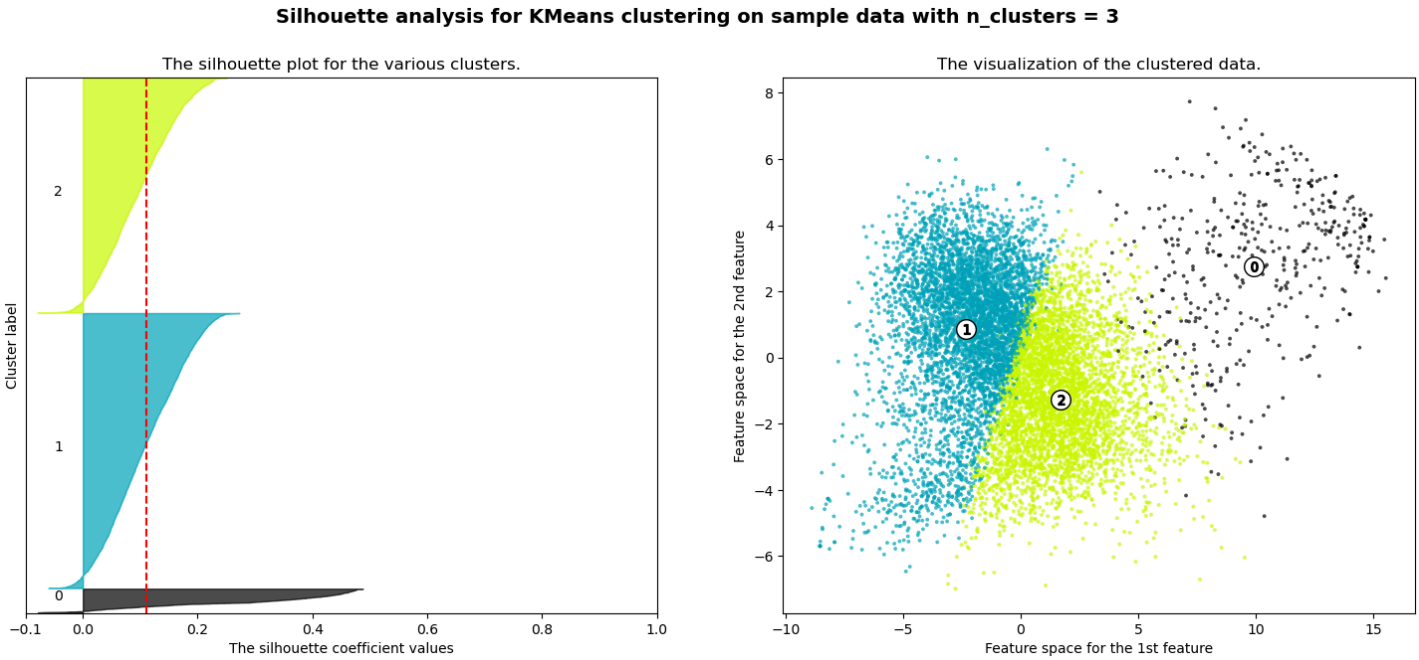


*Figure 17. Silhouette Analysis - n_clusters = 3*

For n_clusters = 4, the average silhouette_score is: 0.08907913896681384

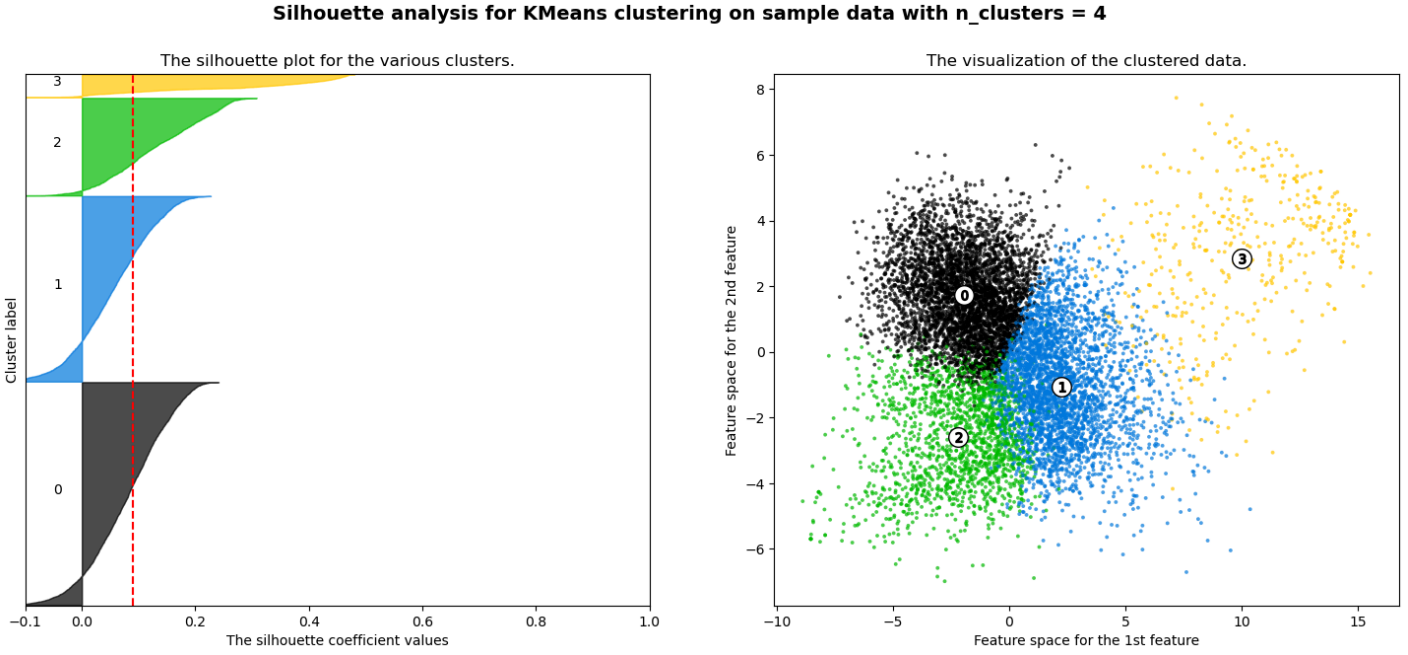


*Figure 18. Silhouette Analysis - n_clusters = 4*

For n_clusters = 5, the average silhouette_score is: 0.07229021241781215

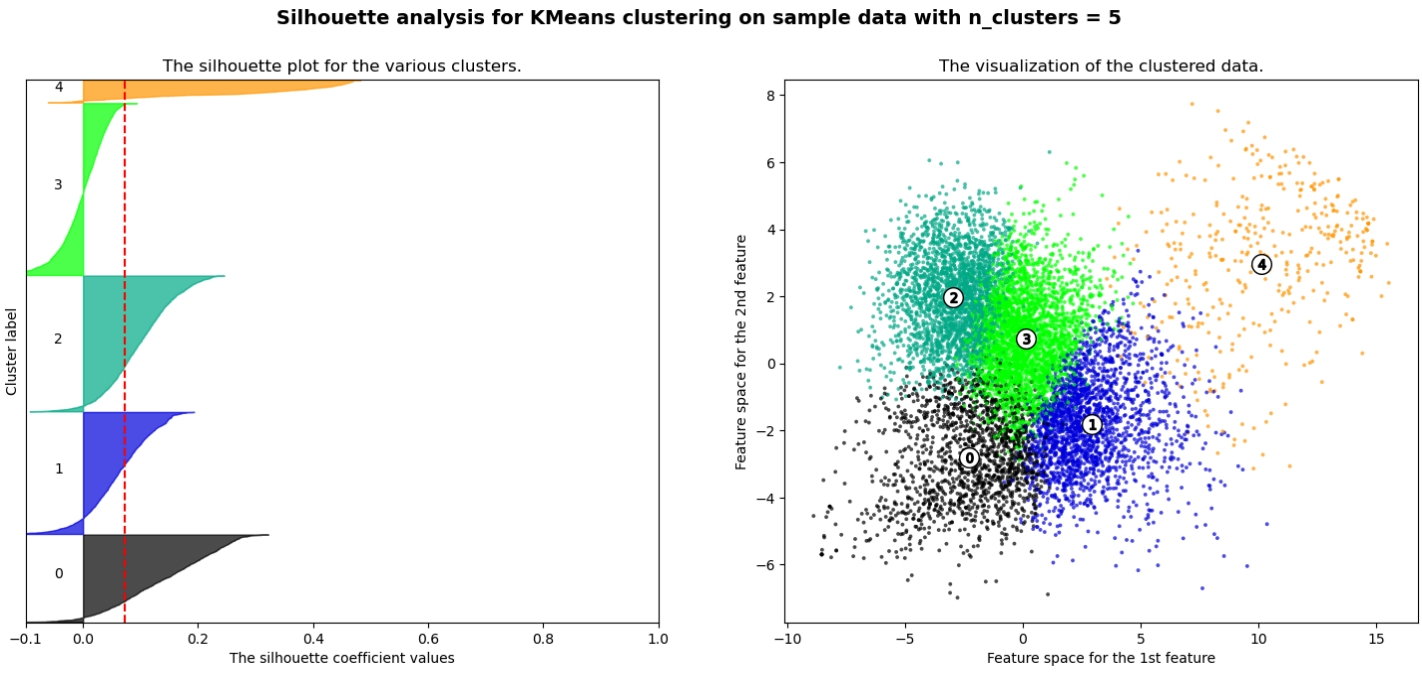


*Figure 19. Silhouette Analysis - n_clusters = 5*

### 9.1.1.1.3 Silhouette Scoring Summary

Silhouette scoring has a range between -1 to 1, where:

- 1: This means clusters are well apart from each other and distinguished.
- 0: This means clusters are indifferent, or the distance between clusters is not significant.
- -1: This means clusters are assigned in the wrong way.

For the dataset, here are the cluster's silhouette scores, starting with a number of the optimal clusters from 2 to cluster 9:

```
Silhouette Score(n=2): 0.156
Silhouette Score(n=3): 0.112
Silhouette Score(n=4): 0.098
Silhouette Score(n=5): 0.081
Silhouette Score(n=6): 0.079
Silhouette Score(n=7): 0.076
Silhouette Score(n=8): 0.076
```

```
Silhouette Score(n=9): 0.072
```

The silhouette scores range from 0.156 for 2 clusters down to 0.072 for 9 clusters. The higher the score is, the better the cluster quality implies. Based on the silhouette scores above, it can be concluded that cluster 2 is the optimum and most distinct and well-separated cluster among the tested options.

### 9.1.1.2 Cluster Visualization using PCA

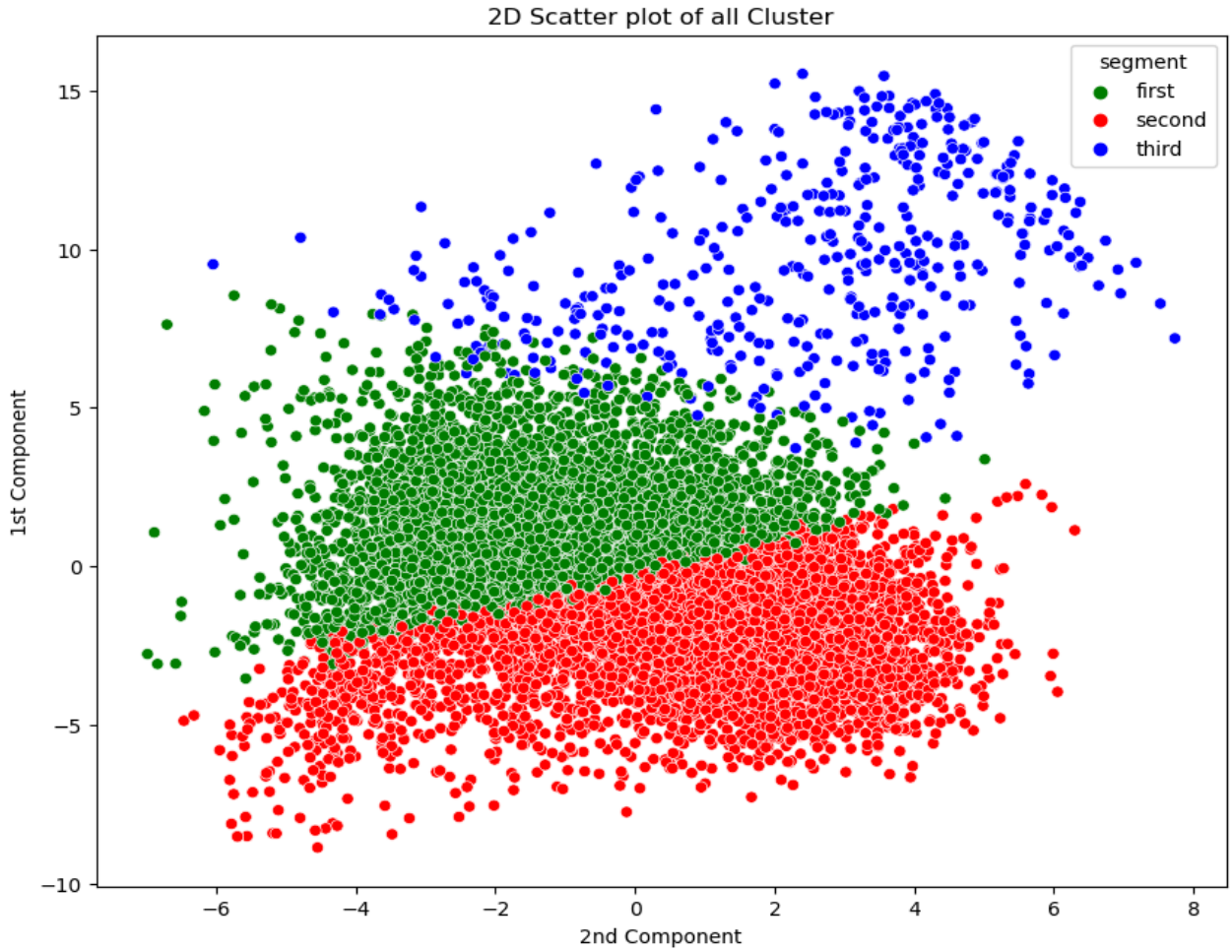


*Figure 20. PCA - Cluster visualization (2D)*

### 9.1.1.3 3D Visualization of K-Means Clustering

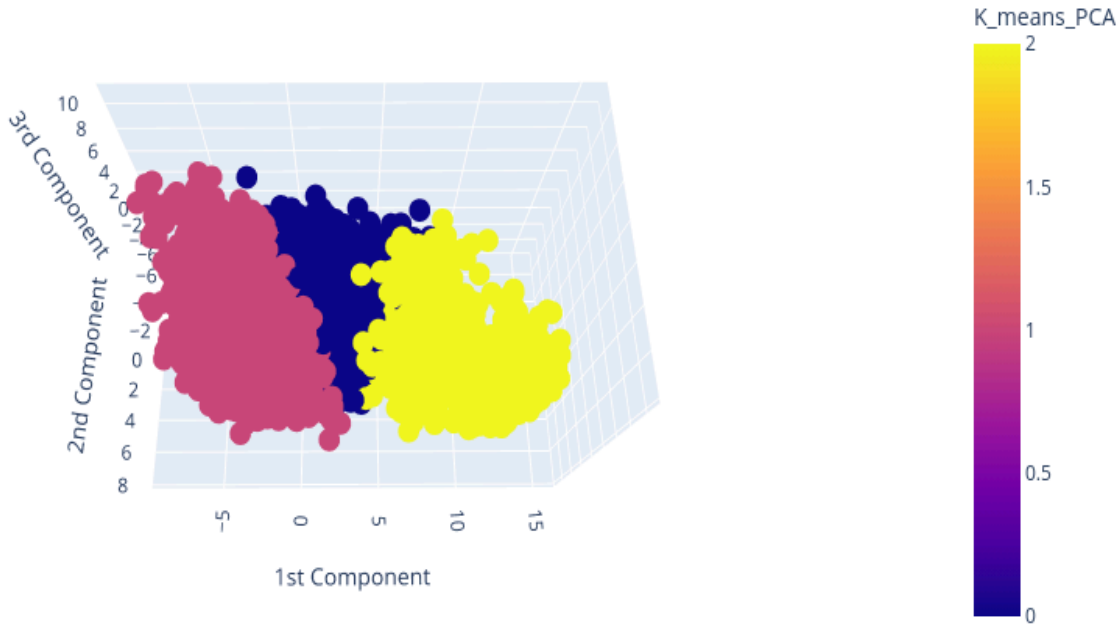


*Figure 21. PCA - Cluster visualization (3D)*

### 9.1.1.4 K-Means Cluster Centroids

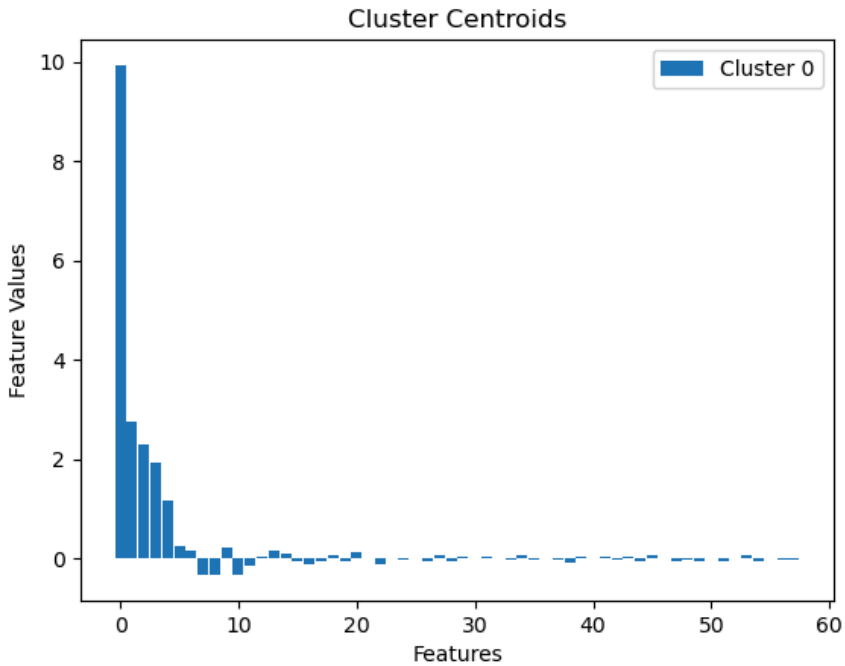


*Figure 22. Centroid feature values Cluster 0*

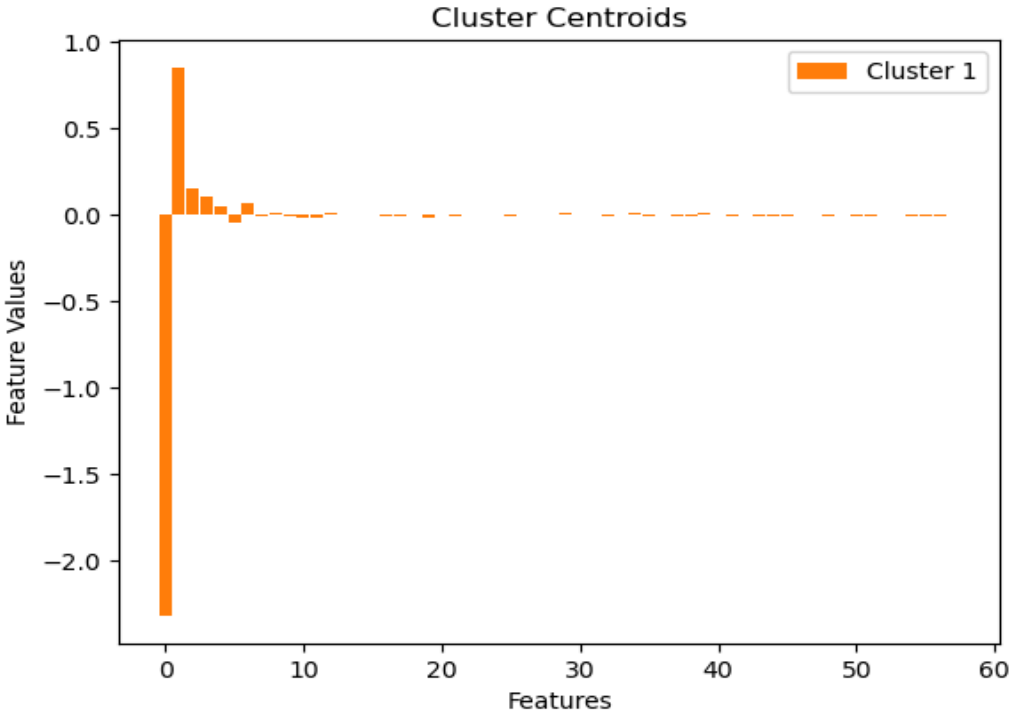


*Figure 23. Centroid feature values Cluster 1*

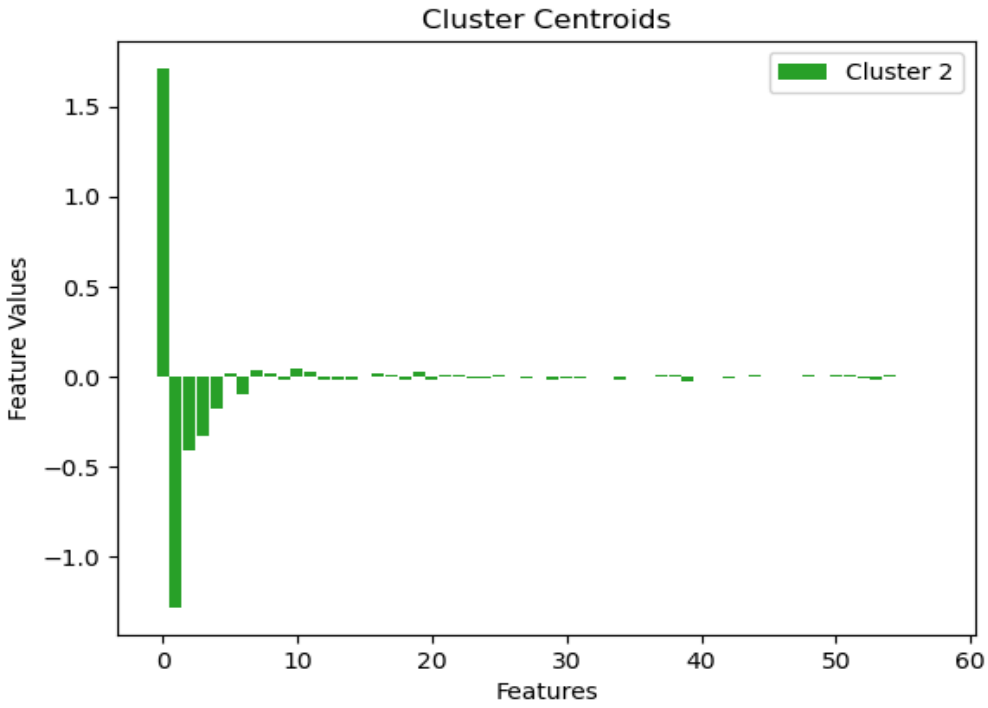


*Figure 24. Centroid feature values Cluster 2*

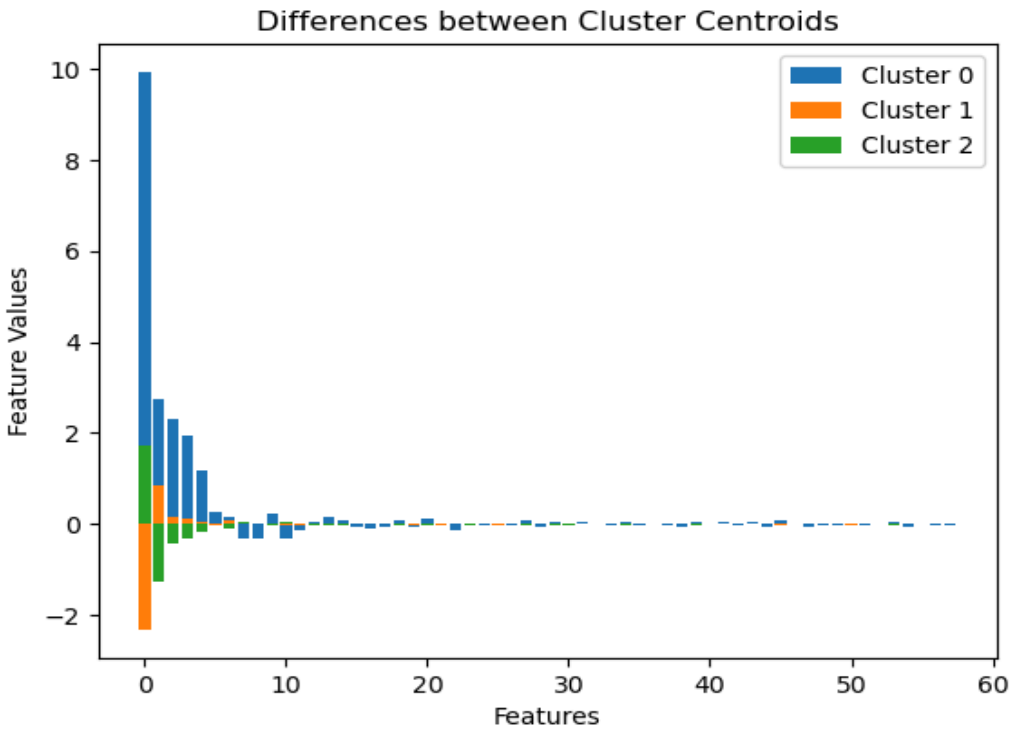


*Figure 25. Differences between cluster centroids*

### 9.1.1.5 Number of records or responses in each cluster

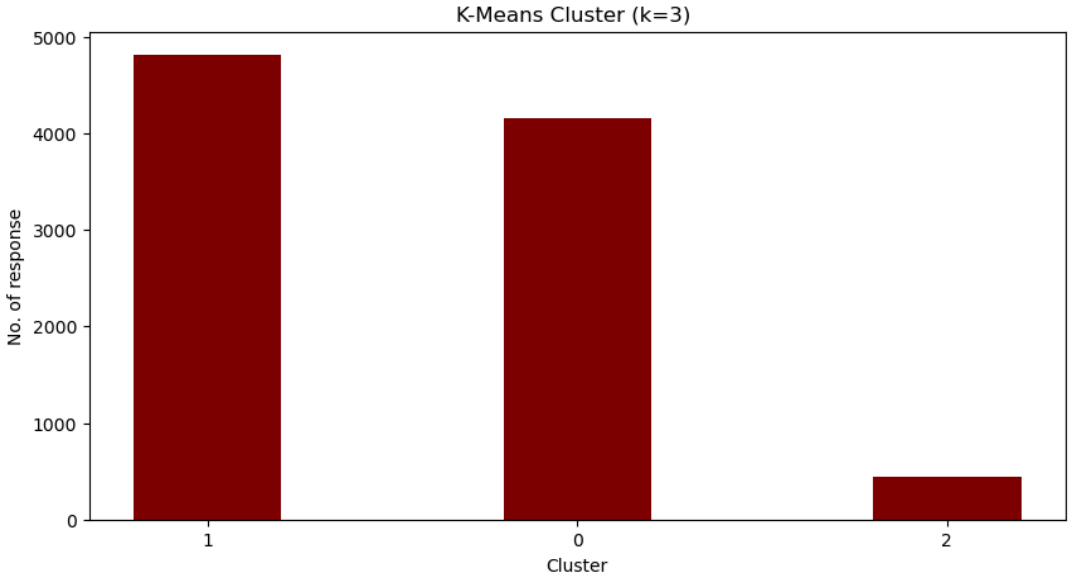


*Figure 26. Number of records in each cluster*

#### 9.1.1.6 K-means Cluster Statistics

**Standard deviation**

In the context of cluster statistics, the standard deviation measures how much the values for each variable deviate from the mean (average) value within the particular cluster. A higher standard deviation implies that the values are more spread out from the mean, indicating greater variability within that cluster for that particular variable.

Standard deviation measures the spread or dispersion of data points within a cluster. The standard deviation for each cluster and all variables can be found in Appendix 5.

**Skewness**

Within a cluster, skewness measures the asymmetry of the distribution of values for a particular variable. This can provide information regarding the shape of the distribution, such as whether it is symmetric, skewed to the left, or the right.

Within that cluster, the skewness statistics of a cluster for each variable can provide insights into the shape of the distribution. It helps to identify whether the values are concentrated towards one end of the scale or more evenly spread out. This information gives an insight and understanding of the nature and variation of attitudes or behaviors within each cluster.

Skewness measures the asymmetry of the data distribution. The Skewness for each cluster and all variables can be found in Appendix 5.

### 9.1.2 Summary of Explanatory Data Analysis (EDA)

After applying K-means and from the silhouette score above, it's observed, that there is not a clear distinction between the clusters, seems like a single cluster. Moreover, from Figure 26 it's found that the number of records per cluster is not evenly distributed, in fact in cluster 2 there are only a few observations. The impression found from the K-means cluster statistics is that, in some cases, the standard deviation (or the variance) of a variable is the same as the other cluster variance. Other algorithms have been applied to verify the findings, such as K-modes, Hierarchical clustering- Agglomerative, Density-based clustering- DBSCAN, etc. The same kind of findings have been obtained from these algorithms.

### 9.1.3 K-modes

K-means and K-modes both are partition-based algorithms, but after applying K-Means, although didn't get the expected and appropriate outcome, an impression has been found regarding an optimal number of clusters, cluster variations, and skewness. Thus, cluster statistics also aids in better understanding the data and the insight. By considering those, the K-modes algorithm has been applied to the processed data set with all 58 variables.

The objective function for K-Modes can be represented as:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}(x, \mu_i)$$

In this formula:

$J$ represents the total dissimilarity or cost function.
$K$ is the number of clusters.
$C_i$ represents the data points in cluster $i$.
$x_i$ is a data point in a cluster.
$\mu_i$ is the mode (centroid) of cluster $i$.
dist($x$, $\mu_i$) measures the dissimilarity between a data point $x$ and the centroid $\mu_i$

From the elbow method plot (figure 27) an abrupt change can be observed in the curve or the possible elbow point and the optimal number of clusters is 3.

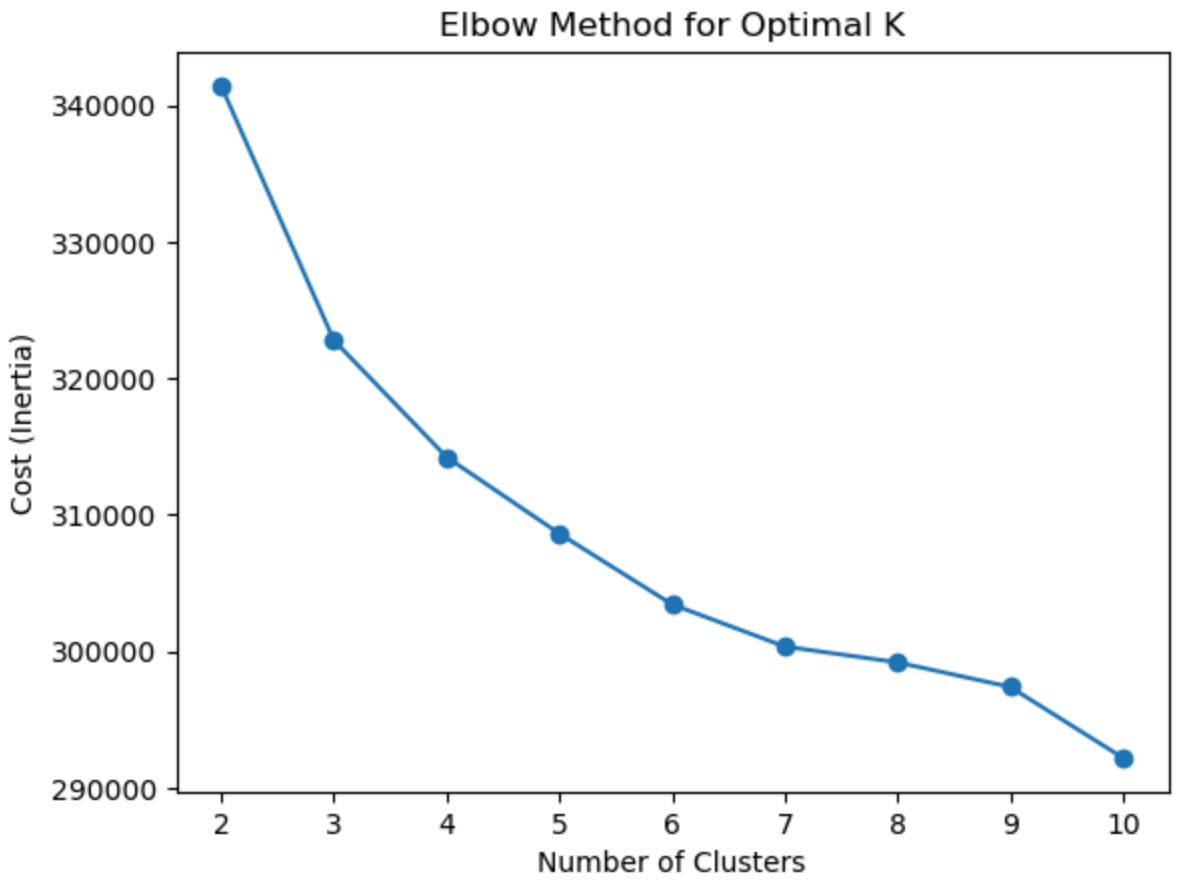


*Figure 27. Elbow Analysis- determining the optimal number of clusters for K-mode*

Silhouette method also applied to find the optimal number of clusters, here are the scores for optimal cluster numbers 2 to 10:

```
n_clusters: 2, silhouette Score: 0.05970247644052602
n_clusters: 3, silhouette Score: 0.00861303831361793
n_clusters: 4, silhouette Score: -0.023603275371850103
n_clusters: 5, silhouette Score: -0.01862640601672009
n_clusters: 6, silhouette Score: -0.0295334493508074
n_clusters: 7, silhouette Score: -0.031012025765698465
n_clusters: 8, silhouette Score: -0.03514142520766808
n_clusters: 9, silhouette Score: -0.024155333622741918
n_clusters: 10, silhouette Score: -0.02732947832000195
```

Positive scores can be observed only for clusters 2 and 3. For the other clusters, scores are negative, which means incorrect placement of data points to some clusters. Thus, from elbow analysis and silhouette score, a conclusion can be made that cluster 3 is the optimal cluster.

Cluster results for the optimal number 3 comprise a nearly equal number of observations in each cluster, which has been presented in Figure 28.

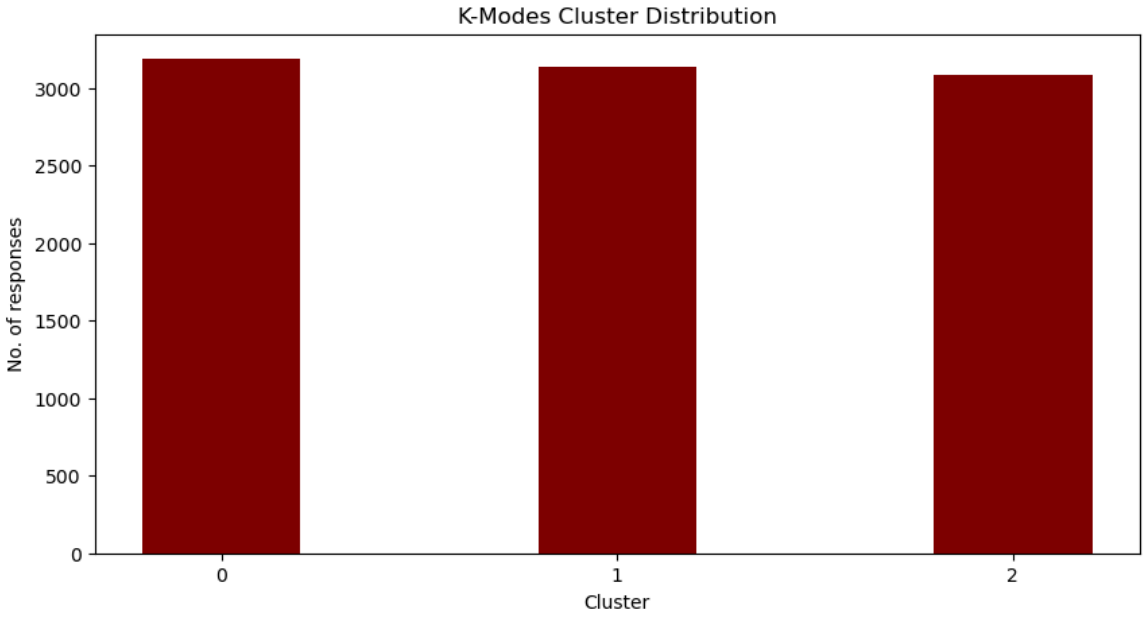


*Figure 28. Number of records in each cluster*

Since the elbow point is not clear enough, the K-mode algorithm has been applied for clusters number 2 to 10 and observed the results, which can be found on the code repository from reference (Clustering-GRETA, 2023). An optimal number of cluster 3 was the best outcome, and a post-analysis has been done to get the insight which will be explained in sections 10.1.3.1 and 10.1.3.2.

### 9.1.3.1 K-Mode Cluster Centroids

The centroid of a k-mode clustering represents the typical or central values for each categorical variable within that cluster. After performing K-mode clustering and storing cluster assignments, the centroid for each cluster has been calculated. For each attribute or variable, calculate the mode (most frequent category) within that cluster.


*Figure 29. Mode Frequency Cluster 0*

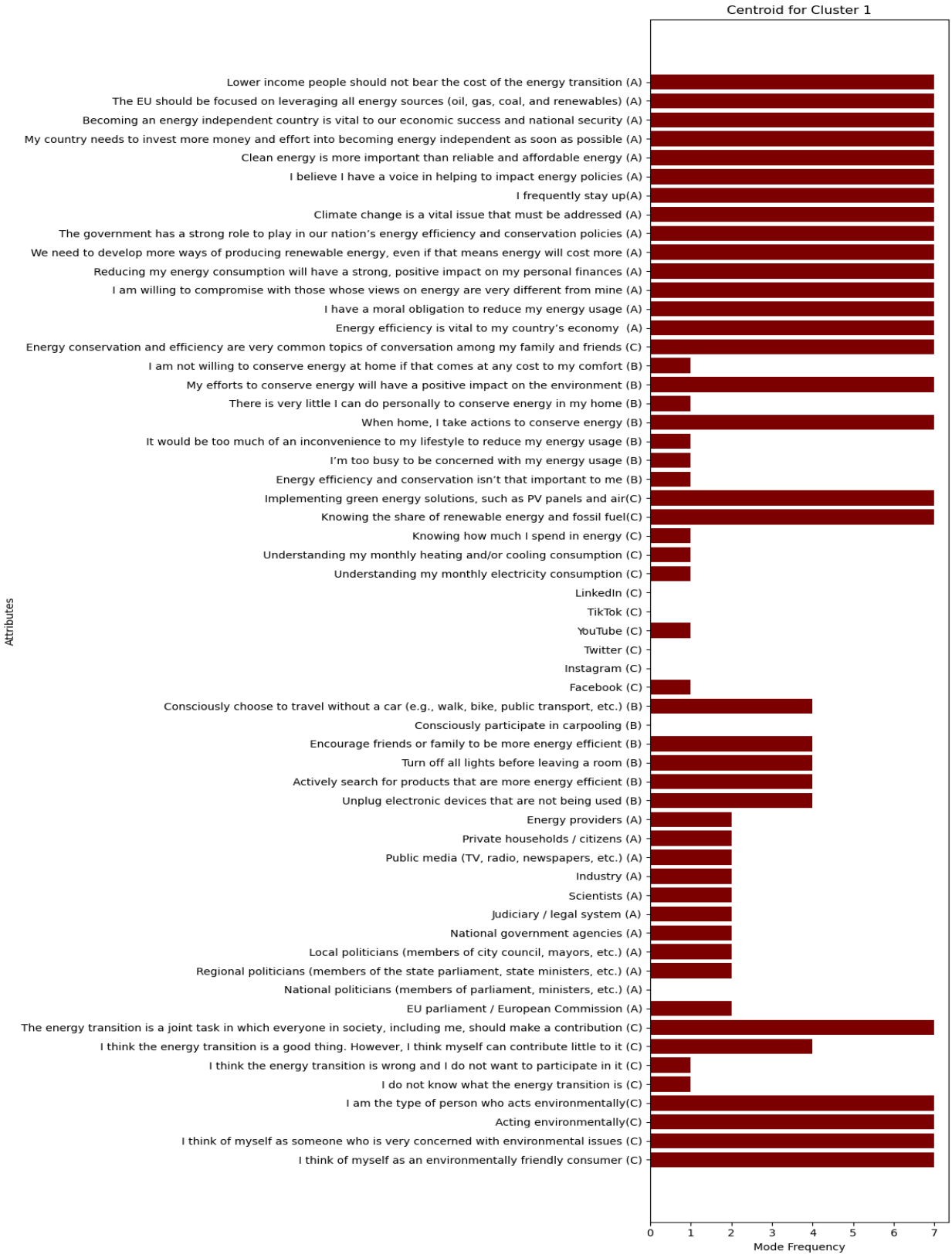


*Figure 30. Mode Frequency Cluster 1*

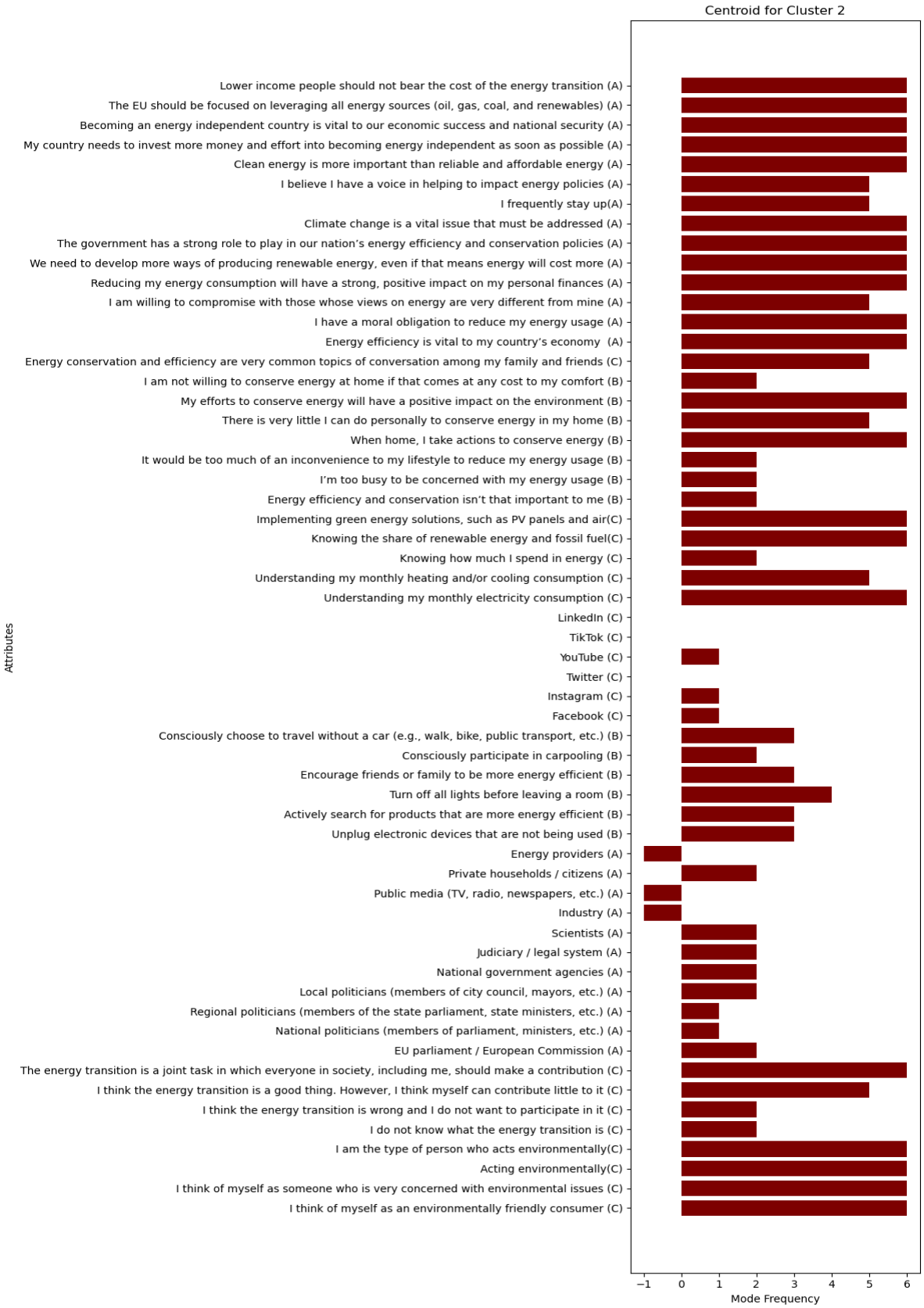


*Figure 31. Mode Frequency Cluster 2*

Centroids for each cluster can be found in Appendix 6, which allows us to see the characteristic features of each cluster.

#### 9.1.3.2 Cluster Profiling

Based on the categorical variables, a comprehensive summary or profile of each cluster is defined. This profile provides insight into the most dominant or frequent categories in each cluster. From the cluster profile, the number of records that fall under a certain category and the percentage have been counted.

To understand the cluster profile, let's consider an example:

Cluster 0 Profile:

Attribute: `I think of myself as an environmentally friendly consumer - To what extent do you agree with the following statements?`

| Category | Count | Percentage |
|---|---|---|
| 5.0 | 1392 | 38.95% |
| 6.0 | 1143 | 31.98% |
| 4.0 | 479 | 13.40% |
| 7.0 | 209 | 5.85% |
| 3.0 | 182 | 5.09% |
| 1.0 | 85 | 2.38% |
| 2.0 | 84 | 2.35% |

Here, the attribute or the variable is “I think of myself as an environmentally friendly consumer - To what extent do you agree with the following statements?” and category 5.0 appears to be the most prevalent response in Cluster 0, with 38.95% of respondents falling into this category.

After performing K-mode clustering, iterated through each cluster and performed cluster profiling. The cluster profile for each cluster and variables can be found in Appendix 6.

## 9.2 Hierarchical Clustering

Hierarchical clustering algorithms recursively partition the instances and construct the clusters in either a bottom-up or top-down fashion. This can be subdivided into:

Agglomerative hierarchical clustering: every object is represented as a cluster itself, then the unit clusters successively merge until an expected optimal cluster structure is achieved.

Divisive hierarchical clustering: all instances belong to a single cluster and then cluster division is applied, and the larger cluster is divided into smaller sub-clusters until an optimum number of clusters is achieved (Lior, 2010).

### 9.2.1 Agglomerative

Agglomerative follows a bottom-up approach, and the choice of linkage method determines how the distance between clusters is calculated during the merging process.

Average linkage computes the distance between clusters as the average of all pairwise distances between points from one cluster to another. Cluster distance is defined as the average distance between all data points in one cluster to all data points in the other cluster. This method is susceptible to noise and outliers and tends to produce more balanced clusters (Daniel, 2011).

Agglomerative clustering average linkage method results have been presented in Figure 32 which is quite imbalanced and mainly one cluster comprises almost all the records.

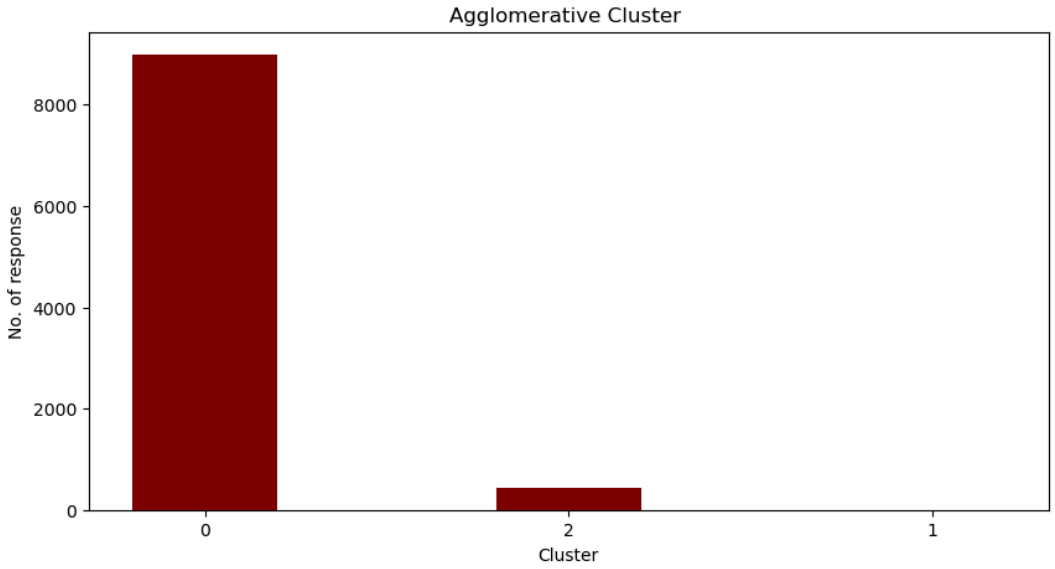


*Figure 32. Number of records in Agglomerative Clustering ('average' Linkage Method)*

The complete linkage method measures the distance between clusters by considering the farthest or maximum distance between points in different clusters. Cluster distance is defined as the maximum distance between any single data point in one cluster to any single data point in the other cluster. This method is less sensitive to outliers compared to single linkage (Daniel, 2011).

Agglomerative clustering complete linkage method results have been presented in Figure 33 where records are more spread among the clusters.

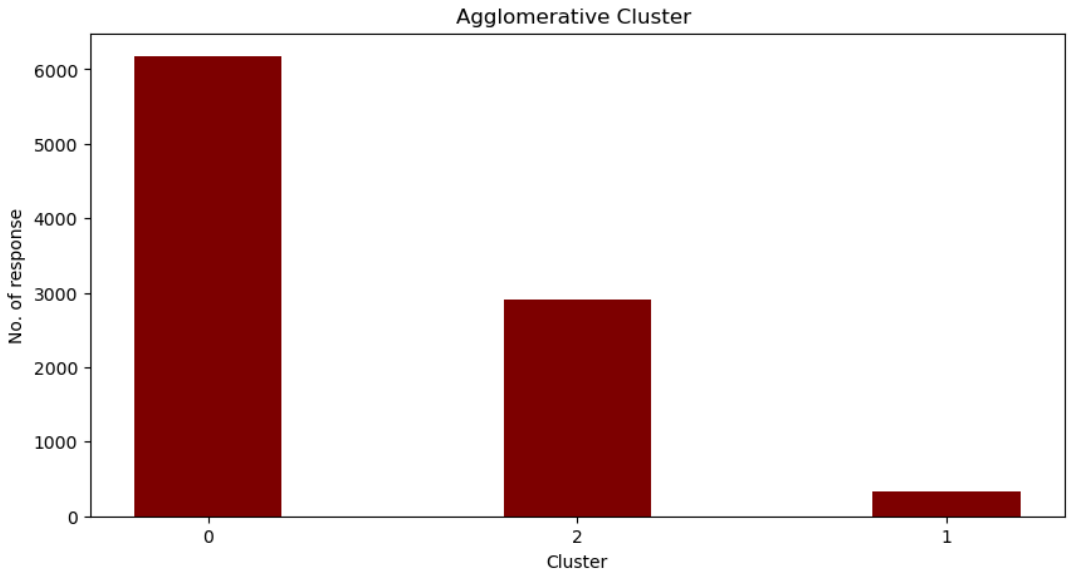


*Figure 33. Number of records in Agglomerative Clustering ('complete' Linkage Method)*

At this stage, one hot encoding has been applied to create dummy variables, where each category creates a new variable for any variable. These newly created variables all together are the input variables for the Agglomerative clustering.

Here are the properties for Agglomerative clustering on one-hot encoding:

- Applied to one-hot encoding all 58 categorical columns
- Removed dummy variables created for the value -1, since this value was a replacement of the null value and should be ignored from the original categorical variable

Results based on one-hot encoding are presented in figure 34, which shows more spread and balanced records among clusters. This strategy is effective in improving the results and generating better outcomes among the variations of Agglomerative clustering.

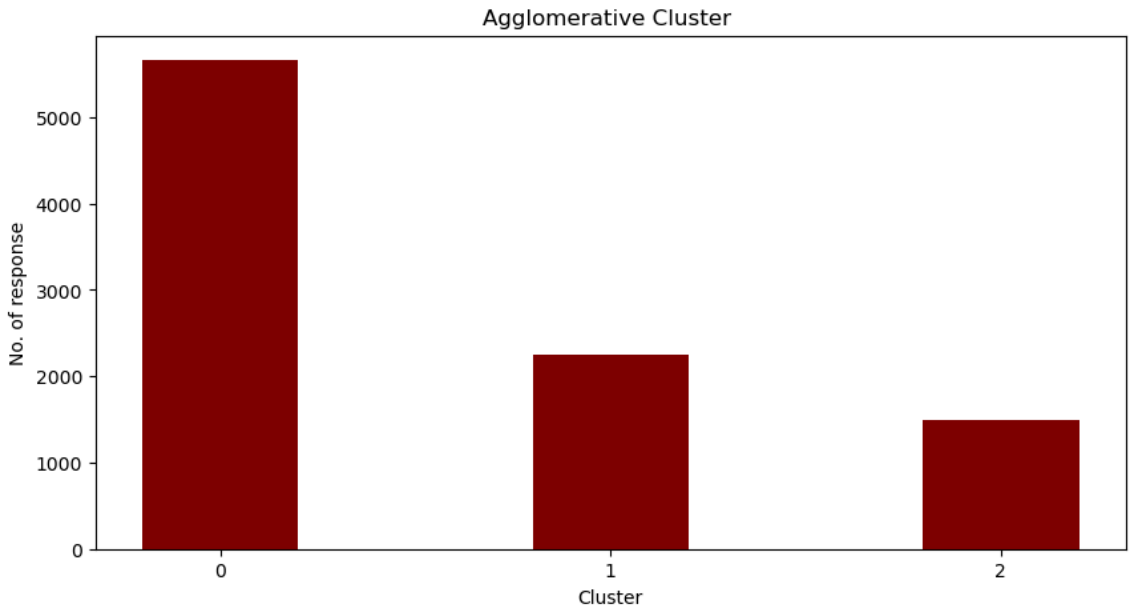


*Figure 34. Number of records in Agglomerative Clustering (One-hot variable encoding)*

## 9.3 Density-Based Clustering

Density-Based Clustering Algorithms (DBCLAs) leverage density within a dataset as a fundamental element to identify clusters that exhibit diverse shapes, sizes, and densities. This density refers to the concentration or distribution within a given area of a particular dataset (Bhattacharjee et al., 2021).

### 9.3.1 DBSCAN

DBSCAN was developed in 1996 this was the first density-based algorithm. From the research, it has been known that DBSCAN can discover clusters from large datasets with noise (Bhattacharjee et al., 2021).

However, there was only one cluster from the algorithm, and here are the properties:

- Identified only one cluster (cluster 0)
- Label -1 records are noise points
- Tried various combinations of epsilon and min_samples

Cluster result has been presented in Figure 35.

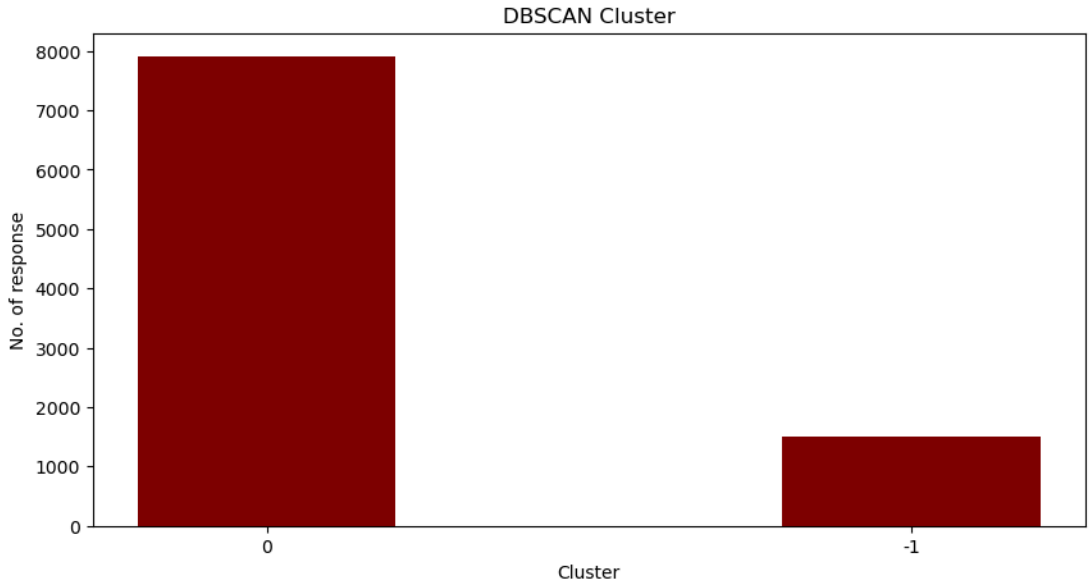


*Figure 35. Number of records in each cluster*

# 10 Discussion

From the results (section 9) of all the algorithms and their explanatory analysis, it's evident that K-Mode provides the most expected and appropriate clustering based on the categorical features. K-Mode results are evenly distributed, which is also reflected in the centroid analysis that will be explained in section 10.1. A brief description of cluster profiling has been presented in section 9.1.3.2 which shows the modes of variables for the cluster and this has a relationship with the attitude analysis (10.2).

Based on the clustering results, further analysis has been done to elicit the attitudes toward green transition, which is required for the main research question and sub-questions, which will be discussed in section 10.2.

In this context, here are the research questions from section 3:

The main research question is:

1. What differences in attitudes towards the green transition exist across Europe?

From a technical perspective, this main question leads (or can be broken down) to sub-questions:

1. What clusters can be found from data (from a multinational survey) when focusing on answers to affective, behavioral, and cognitive aspects?
2. How does one cluster differ from another cluster, or what different types of knowledge are found - how/why are these different?

The following discussion and result analysis will illustrate the underlying concept, critical reasoning, explanations, and interpretation with substantial references.

## 10.1 Analysis and Discussion on Clustering Algorithms

To illustrate one of the research questions, the following discussion provides insight into *the difference between the approach and implementation of clustering methods* based on the results:

K-means is a widely used algorithm for clustering, for the dataset and clustering model results visualized in Figure 20. It has been observed that the clusters are overlapping, and a clear distinction between clusters is difficult. From elbow analysis (Figure 15) it can be observed the optimal cluster number is 3 or 4 but the silhouette score (Figure 19) doesn't provide a clear indication of the optimal cluster number for K-means. One reason behind this is all the variables and columns are categorical, and research shows that K-means has shortcomings in differentiating attribute values among categorical values (Ahmed et al., 2020).

At this point, Agglomerative clustering has been applied for different combinations of linkage methods (average, complete, single) and demonstrated the results in Figures 32 and 33. They indicated that one cluster dominated or comprised the majority of the records, and there was a lack of clear differentiation among the clusters. Since Agglomerative clustering follows a bottom-up approach and needs user-specified parameters for tuning, hence doesn't perform well directly on categorical data (Wei et al., 2019). To overcome that, one-hot encoding was applied, and the results were improved, but still, those were dominated by a single cluster, which is shown in Figure 34.

Similar results have been found from the Density-based clustering algorithms-results were dominated by one cluster and no clear separation among clusters. One-hot encoding has also been used, which didn't improve the results for Density-Based clustering.

### 10.1.1 K-Mode: A Robust Approach for Mixed Data Clustering

K-Mode, which is a variation of K-means and specifically designed for mixed data (categorical and continuous both in the same dataset), provided satisfactory results, outperformed and surpassed other algorithms in several aspects:

- The efficacy of K-Mode is highlighted by the observed properties such as compactness and separation. Those are meticulously observed from the elbow and silhouette coefficient illustrated in section 9.1.3.
- A deeper understanding of the clustering process is unveiled by the centroid analysis and demonstrated by an explanatory analysis in section 9.1.3.1 (results). Centroid analysis depicts the connectivity of clusters, which means, shedding light on the extent to which items are grouped with their nearest neighbors in the data space. The results, which are accompanied by Figures 29, 30, and 31, depict the cluster centroid and variable contribution for each cluster.
- By considering the above insights, a pivotal strength of K-Mode results lies in its ability to ensure a well-balanced distribution of records across clusters. For instance, in clusters '0,' '1,' and '2,' the record counts stand at 3188, 3138, and 3088, respectively (detailed in section 9.1.3.1 results).

## 10.2 Relationship Between Attitude Components in Clusters

This research work continued to find the relationship and contribution among the categories (A = Affect, B = Behavior, C = Cognitive) of variables described in section 5 (research methodology) for each cluster. Here, just to note again, each category represents each attitude component of the ABC attitude model.

While analyzing the clusters of the K-modes algorithm, which has demonstrated superior performance compared to other algorithms (as discussed in section 10.1), the centroid refers to the mode of kernel density estimation, determined by the data points within each cluster. In a cluster, centroids are valid patterns that lie in high-density areas and function as representatives of their clusters and neighborhoods. For a variable within a cluster, each centroid value stands for the mode (or most frequent value) of that particular variable, among all the data points assigned to that cluster. A variable's centroid value also illustrates a cluster's central tendency. This implies that within each cluster, a centroid value has been determined for every variable. However, across clusters, the centroid value varies for a particular variable. Variables with higher centroid values within a cluster represent the higher contribution and central tendency to that particular cluster's characteristics (Williamson 1965).

At this juncture, to comprehend the interconnections among attitude components within clusters in the mathematical framework and to enhance the coherence of the discussion, it is essential to grasp the connections between different parts of research. Referring to the methodology (section 5.1) which has three parts, Part 1 describes the qualitative method for data annotation (or variable labeling) according to the attitude components, Part 2 is data clustering and quantitative approach, and Part 3 uses the mixed method where both qualitative and quantitative has been used for cluster's results analysis and attitude component elicitation. Although Part 1 is a qualitative method and an unnatural science, the resultant has a quantitative relationship with Part 2 and Part 3, within a mathematical paradigm. Thus, these three parts are strongly interrelated and dependent on each other for the outcome.

The following Figure 36 is a simple representation of the relationship between the three parts of the research method. In every part, two things are common, the first one is the variables (or columns) and the second one is the attitude components. In each part of the research, these two things, together, make an inherent relationship. Furthermore, parts of the research are sequential, interrelated, and interdependent.

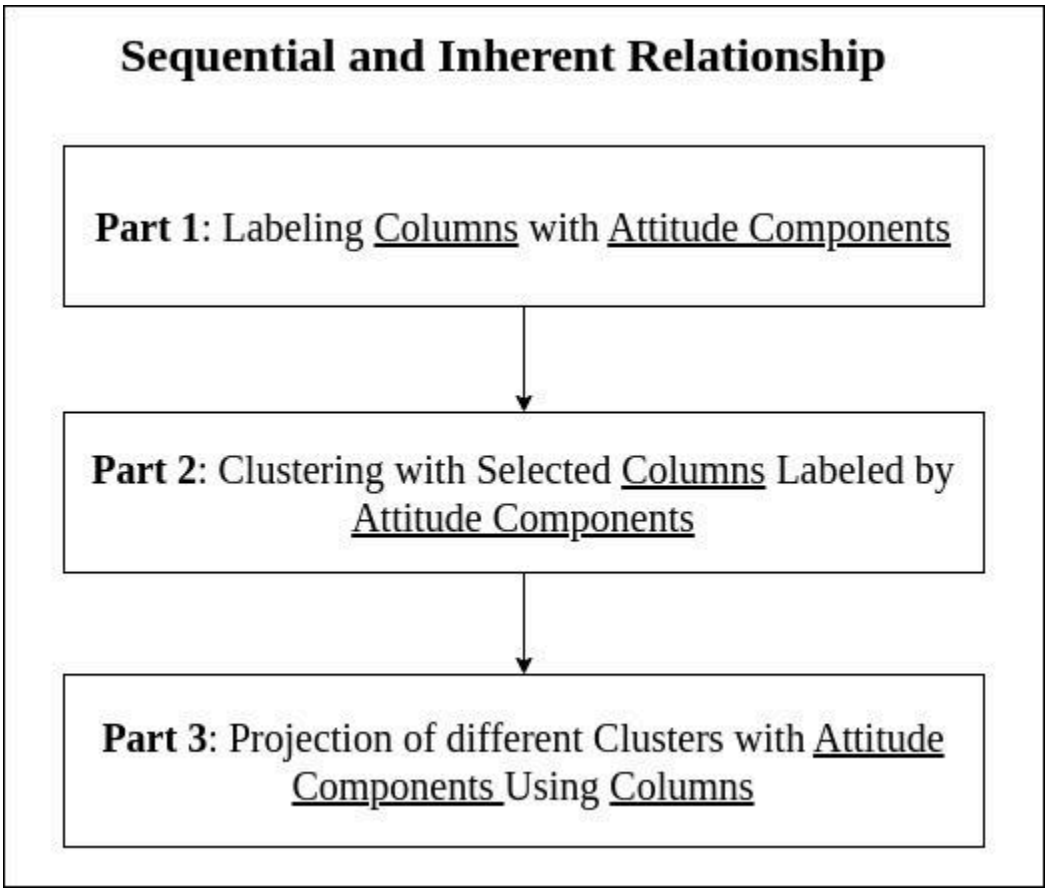


*Figure 36. Sequential and Inherent Relationship Between Parts of the Research*

Based on the postulation described above, for analyzing quantitative relationship and contribution, the mean of positive mode values of all variables in a cluster has been taken, but not in each attitude component (or category), because of cluster characteristics lies within the cluster and each attitude category is part of a particular cluster (one particular category may not necessarily exist in a single cluster). The average of the 'positive modes of variables' is calculated for an entire cluster, and the same mean value is applied uniformly on each category (of that particular cluster) to filter out the columns whose centroid value is higher than the calculated cluster mean.

For clearer articulation of the above process, the implemented steps are:

Step 1: Filtered the positive mode values of all variables and take the means of mode values for a cluster
Step 2: Count the number of variables of each attitude component (or category) whose mode value is higher than the mean value (calculated in Step 1), and visualize the statistics in figures 37, 38, 39

The mean of positive mode values (from Step 1) of all variables in a cluster are as follows:

Cluster 0: 3.86
Cluster 1: 3.25
Cluster 2: 3.72

It has been observed that the lowest mean of a cluster is 3.25 and the difference between the lowest and highest mean (3.86) is not big in the sense that in the Likert scale data structure, the mathematical difference between any two category values is at least one. Thus, even in the case of calculating the mean of these three clusters (an alternative hypothesis) and then conveying further analysis, would still give the same results.

### 10.2.1 Validating Attitude Component Relationships with an Alternative Hypothesis

There could be a limitation of the process mentioned above (in 10.2), that stems from variations in the source data, and data structure from the underlying data collection mechanism. For example, the number of categories or value ranges is not the same for all variables, most of the variable has value ranges from 0 to 7, but there are some variables whose value ranges from 0 to 4. In the case of rescaling without considering the importance of a variable in the dataset, then it impacts the entire quality or the structure of the data, moreover, rescaling would also change the granularity of categorical values. Another limitation is that this process assumes that within a variable the higher the category value is, the more importance or priority, but there are only a

few exceptions too. The specific reasons behind not applying the rescaling (or normalization) are already explained in section 7.12 (normalization).

Thus, by considering the reasoning above, the process explained in section 10.2 is valid and acceptable. Moreover, it didn't just rely on the quantitative method. Furthermore, a manual inspection has also been conducted, for example, in Figures 29, 30, and 31 the centroid value of every variable for every cluster is observable. In addition, for every variable, the frequency of each category value has been counted and verified, which can be found in Appendix 7. This process also analyzed cluster profiles in section 9.1.3.2 and detailed in Appendix 6. From those analyses, the frequencies of category values and how mode values of each categorical variable contribute to the cluster can easily be observed.

## 10.3 Characteristics of Clusters within the Attitude Paradigm

The analysis and discussion above (10.2) lead to answering other research questions with further analysis, which is the optimal cluster based on different attitudes, how clusters are different around this attitude, in other words, *which characteristics make the clusters different, and the reasoning* behind this:

For a coherent discussion and visual comprehension of the insights, this section will refer to figures and charts generated in previous sections. The plots (Figures 29, 30, 31) incorporate category labels for each variable. For each cluster, variable name, centroid value, and its category all together enhance visual comprehension. In addition, the analysis and postulation provided in section 10.2 are strongly related and reflect the following analysis.

Figure 37 illustrates the distribution of variables across different categories within Cluster 0. It's noticeable that people's cognition and affect are comparatively higher, but energy behavior is poor. To interpret, affect (or emotional engagement) with green energy is stronger, but actual energy-related behaviors (or actions) are lacking. This is because some people's beliefs and opinions are stronger on energy policies, for example, after observing the variables from Figure 29, there are some questions such as "`I believe I have a voice in helping to impact energy policies`" which related to affect and strongly contributing to cluster 0.

Cluster 0: Columns with Centroids > mean:
{'C': 12, 'B': 4, 'A': 14}

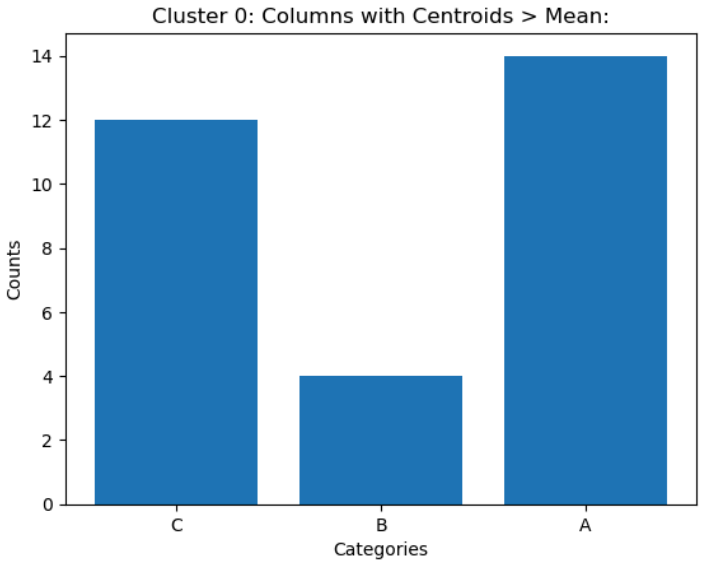


*Figure 37. Cluster 0: Columns with Centroids > Mean*

Similarly, from Figure 38 it's observed that people from Cluster 1 have higher cognition and affect but show relatively small behavior. From this finding, the conclusion is that people who have higher emotional engagement toward green transition also have strong cognition (belief or opinion in this context) (Vishal, 2014). Furthermore, another conclusion is that people's tendency to take necessary actions (or effective steps) towards green transition can be poor even though they have higher emotional engagement (or positive emotion) and strong positive opinions for the green transition.

Cluster 1: Columns with Centroids > mean:
{'C': 14, 'B': 8, 'A': 14}

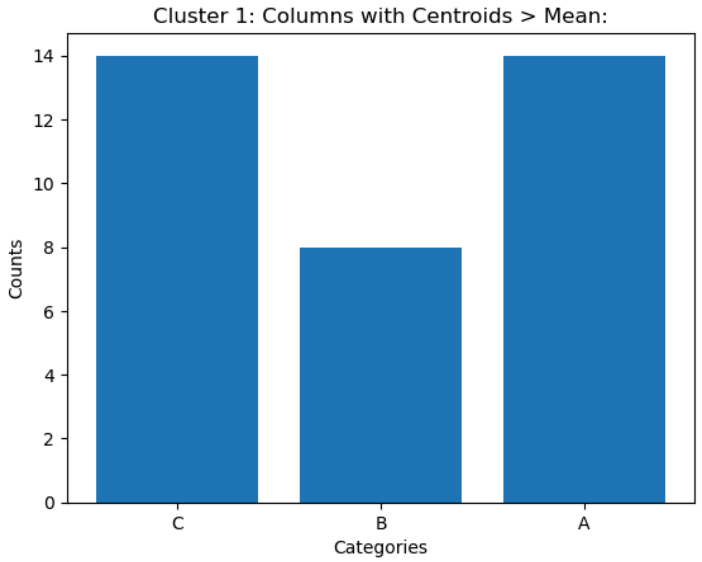


*Figure 38. Cluster 1: Columns with Centroids > Mean*

Cluster 2: Columns with Centroids > mean:
{'C': 9, 'B': 7, 'A': 13}

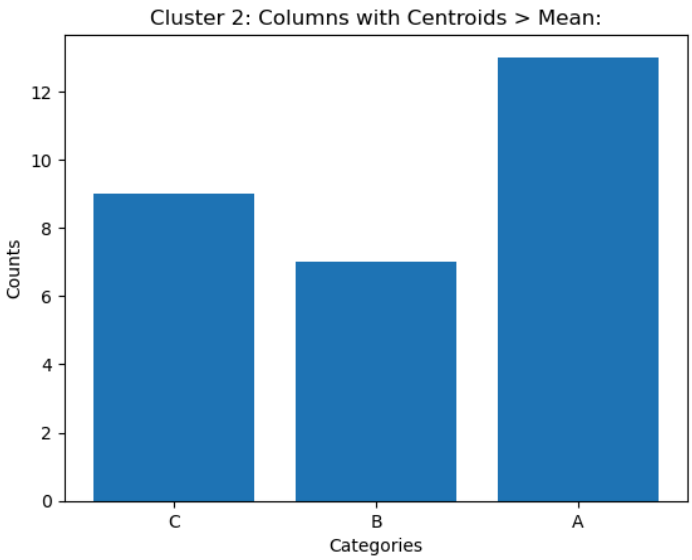


*Figure 39. Cluster 2: Columns with Centroids > Mean*

Cluster 2 (Figure 39) is interesting and a bit different compared to Cluster 0 and Cluster 1. 'Affect' is significantly high, while Cognition and Behavior are relatively smaller. In this context, Affect symbolizes the emotional response (Vishal, 2014). This means for this cluster, people may have higher emotional engagement toward green transition, compared to their knowledge and behavioral tendency.

## 10.4 Relationship and Insights with Scientific Publications and Research

This thesis work searched and explored scientific publications to get insights, explanations, and reasoning for the research findings. Research indicates that both positive and negative emotions toward an event exert significant influences on physiological reactions. Tobias Brosch and Linda Steg mentioned that people's feelings impact thinking and behavior simultaneously. This perspective contrasts with the traditional view that emotions are seen as an irrational influence that disrupts clear thinking and reasoned cognition, which is relevant to Cluster 2 (Figure 39) (Tobias and Linda, 2021).

Another research was on college students' energy-saving behavior intentions. That paper mentioned that the behavior of college students towards energy saving is having a degree of knowing and doing separation- which is knowing is easy, but doing is difficult (Yang et al., 2020) [22]. This research and facts explain the findings of this thesis work too, in every cluster, it

has been found that behavioral tendency (or actions) for green transitions, have the least contribution compared to cognition (C) or affect (A).

From the previous research, also known that people often underestimate long-term hazards and severity. This leads to cognitive bias while making decisions, such as everyone knows about global warming but does not know enough about the severity of the issue, and people have less response or tendency towards global warming (Johan. E. et al., 2023). A similar conclusion can be elicited from the research findings in the context of green transition because from the clustering (Figure 37, 38, 39) it's elicited that people have better knowledge and emotional engagement but behavioral tendency or action is poor in the context of the sustainable green energy transition.

# 11 Limitations and Future Work

As mentioned earlier, this problem is multidisciplinary and this research or clustering data were from Europe but for the social or behavioral context, the results can be varied or even entirely different to other parts of the world. For example, in Asia more specifically in developing countries, this green transition concept is still limited to the research or academic community, mass people are not engaged or aware (Fengyun et al., 2022). Governments are prioritizing other projects and issues such as poverty, natural calamities, etc.

Energy-saving behavior could be multiple types, such as habitual energy-saving behavior and purchasing energy-saving behavior (Yang et al., 2020). Future work could be based on how people's cognitive knowledge and emotional engagement impact different types of energy-saving behaviors. In addition, this research was mostly for individual energy attitude, but a collective attitude could be an interesting aspect too.

In terms of technical or engineering work, specifically for Machine Learning clustering algorithms, research shows that Agglomerative clustering doesn't perform well in the case of using categorical variables directly. Xiong et al. mentioned two problems for the categorical data, one is the lack of inherent similarity measure and the other one is clusters are prone to being embedded in different subspaces. To overcome those problems, Xiong et al. proposed a new algorithm named divisive hierarchical clustering algorithm. This approach could be applied to this research dataset since all variables are categorical and the dataset is high-dimensional at the same time (Xiong et al., 2009).

# 12 Conclusion

The main focus of this research was around the social dimensions of green energy transition based on GRETA project survey data, this research unveiled a paradigm to tackle multidisciplinary problems. Energy attitude is a social science problem, and shaping the energy data to an attitude model in the context of behavioral or social was challenging because of limited existing research and resources. However, this thesis proposed a method for the annotation of public responses to model the data according to the attitude model that reflects the different components while clustering. Comparison among different clustering algorithms was crucial because of the high dimensionality of the dataset. As described in the discussion section, K-modes gave the best outcome in terms of different attitude components that aligned with the research objective and the research questions. This research provided theoretical explanations with the existing research and publications that described the outcome of clustering algorithms. Proper analysis of cluster centroid and visualization of every outcome bolster and align the clustering results, attitude model components, and theoretical explanations. Thus, relationships among attitude components are perceivable which will help in understanding the action and interrelation of people's attitudes toward green transition from a behavioral perspective, which is one of the prime objectives of GRETA.

Moreover, this thesis delved into the technical intricacies of implementing clustering algorithms, with a particular focus on the energy domain, specifically, the concept of energy citizenship. The dataset was human responses that had been needed to leverage accurately for clustering. An initial hurdle involved categorizing columns in terms of different attitudes. In the beginning, that hurdle has been solved with the help of existing research on attitude. For employing and framing energy citizenship data to an attitude model, this thesis proposed and explained the process of categorizing or annotating the variables for each attitude based on the three components of the ABC attitude model, details are in the research methodology. Moreover, how energy citizen comprises the human persona and leveraging it out of the high-dimensional dataset was a unique and interesting work, while as a human, we are not value-free. For instance, when it comes to assessing one's environmental friendliness, individuals might provide self-ratings that could be overly optimistic or biased.

To address these complex challenges, a diverse array of clustering algorithms has been employed, including Partition-Based Clustering, Hierarchical Clustering, and Density-Based Clustering. This multipronged approach allowed us to examine the problem from various angles, gaining a more comprehensive understanding of the underlying patterns within the data. In addition to algorithmic exploration, this thesis work conducted an extensive explanatory analysis. This critical step not only aided in selecting the most suitable algorithms for the task but also served as a means to validate and refine the results. By meticulously dissecting the data

and its nuances, this thesis was able to unearth valuable insights and ensure the reliability of the clustering outcomes.

# 13 References

(GRETA 2023a) Welcome to GRETA Analytics! About GRETA. [www document], [Accessed on 15.11.2023] Available at https://projectgreta.shinyapps.io/greta-analytics/?tab=home

(Lior, 2010) Lior Rokach. A survey of Clustering Algorithms. Department of Information Systems Engineering, Ben-Gurion University of the Negev, https://link.springer.com/chapter/10.1007/978-0-387-09823-4_14

(Vishal, 2014) Vishal Jain. 3D MODEL OF ATTITUDE. International Journal of Advanced Research in Management and Social Sciences, ISSN: 2278-6236

(Sewell et al., 2005) Sewell, Grandville, and P. J. Rousseau. Finding groups in data: An introduction to cluster analysis. (2005, Page 14)

(Ira Assent, 2012) Ira Assent. Clustering high dimensional data. WIREs Data Mining Knowl Discov 2012, 2: 340–350 doi: 10.1002/widm.1062

(Madeleine and Jenny, 2022) Madeleine Wahlund, Jenny Palm; The role of energy democracy and energy citizenship for participatory energy transitions: A comprehensive review. https://doi.org/10.1016/j.erss.2021.102482

(Aremu et al., 2019) Aremu, Oluseun Omotola; Hyland-Wood, David; McAree, Peter Ross (2019). A Machine Learning Approach to Circumventing the Curse of Dimensionality in Discontinuous Time Series Machine Data. Reliability Engineering & System Safety, (), 106706–. doi:10.1016/j.ress.2019.106706

Gerard Mullally, Niall Dunphy, Paul O'Connor, Participative environmental policy integration in the Irish energy sector, https://doi.org/10.1016/j.envsci.2018.02.007

(Baron et al., 1984) Baron, R.A. & Byrne, D., 1984. Social psychology understanding human interaction, Boston: Allyn & Bacon.

Eagly, A.H. & Chaiken, S., 1998. Attitude structure and function.

D. Wuebben, J. Romero-Luis, M. Gertrudix, Citizen science and citizen energy communities: a systematic review and potential alliances for SDGs, Sustainability 12 (23) (2020) 1–24

(Lennon and Myles, 2017) Lennon, Myles (2017). Decolonizing energy: Black Lives Matter and technoscientific expertise amid solar transitions. Energy Research & Social Science, S221462961730172X–. doi:10.1016/j.erss.2017.06.002

(Linda Steg et al. 2015) Steg, Linda; Perlaviciute, Goda; van der Werff, Ellen (2015). Understanding the human dimensions of a sustainable energy transition. Frontiers in Psychology, 6(), –. doi:10.3389/fpsyg.2015.00805

(Tobler et al., 2012) Tobler, C., Visshers, V. H. M., and Siegrist, M. (2012). Consumers' knowledge about climate change. Clim. Change 114, 189–209. doi: 10.1007/s10584-011-0393-1

(Hugo et al. 2021) Hugo Lucas, Ruth Carbajo, Tomoo Machiba, Evgeny Zhukov and Luisa F. Cabeza. Improving Public Attitude towards Renewable Energy.

Marvin E. Olsen (1981). Consumers' Attitudes Toward Energy Conservation. , 37(2), 108–131. doi:10.1111/j.1540-4560.1981.tb02628.x

Gordon Walker (1995). Renewable energy and the public. , 12(1), 0–59. doi:10.1016/0264-8377(95)90074-c

(Ahmed et al., 2020) Ahmed, Mohiuddin; Seraj, Raihan; Islam, Syed Mohammed Shamsul (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics, 9(8), 1295–. doi:10.3390/electronics9081295

(Manisha et al., 2017) Manisha Goyal; Shruti Aggarwal; A Review on K-Mode Clustering Algorithm. International Journal of Advanced Research in Computer Science. Volume 8, No. 7, July – August 2017. doi: http://dx.doi.org/10.26483/ijarcs.v8i7.4301

(Wei et al., 2019) Wei, Wei; Liang, Jiye; Guo, Xinyao; Song, Peng; Sun, Yijun (2019). Hierarchical division clustering framework for categorical data. Neurocomputing, S092523121930267X–. doi:10.1016/j.neucom.2019.02.043

(Xiong et al., 2009) Xiong, Tengke; Wang, Shengrui; Mayers, André; Monga, Ernest (2009). A New MCA-Based Divisive Hierarchical Algorithm for Clustering Categorical Data. [IEEE 2009 Ninth IEEE International Conference on Data Mining (ICDM) - Miami Beach, FL, USA (2009.12.6-2009.12.9)] 2009 Ninth IEEE International Conference on Data Mining. 1058–1063. doi:10.1109/ICDM.2009.118

(Yang et al., 2020) Yang R, Yue C, Li J, Zhu J, Chen H, Wei J. The Influence of Information Intervention Cognition on College Students' Energy-Saving Behavior Intentions. Int J Environ Res Public Health. 2020 Mar 4;17(5):1659. doi: 10.3390/ijerph17051659. PMID: 32143334; PMCID: PMC7084549

(Johan. E. et al., 2023) Johan. E. (Hans) Korteling, Geerte L. Paradies and Josephine P. Sassen-van Meer (2023). Cognitive bias and how to improve sustainable decision making. TNO Netherlands Organization for Applied Scientific Research, The Hague, Netherlands.

(Tobias and Linda, 2021) Tobias Brosch and Linda Steg. Leveraging emotion for sustainable action. One Earth 4, December 17, 2021. Published by Elsevier Inc. https://doi.org/10.1016/j.oneear.2021.11.006

(Fengyun et al., 2022) Fengyun Li, Junxia Zhang, Xingmei Li. Research on supporting developing countries to achieve green development transition: Based on the perspective of renewable energy and foreign direct investment. https://doi.org/10.1016/j.jclepro.2022.133726

(Kaldellis et al., 2012) J.K. Kaldellis; M. Kapsali; Ev. Katsanou (2012). Renewable energy applications in Greece—What is the public attitude?. , 42(none), 37–48. doi:10.1016/j.enpol.2011.11.017

(Luis Pérez-Lombard et al. 2008) Luis Pérez-Lombard; José Ortiz; Christine Pout (2008). A review on buildings energy consumption information. , 40(3), 394–398. doi:10.1016/j.enbuild.2007.03.007

(Ryghaug et al. 2018) Ryghaug, Marianne; Skjølsvold, Tomas Moe; Heidenreich, Sara (2018). Creating energy citizenship through material participation. Social Studies of Science, 030631271877028–. doi:10.1177/0306312718770286

(Joseph Murphy, 2007) Energy citizenship: psychological aspects of evolution in sustainable energy technologies. Governing Technology for Sustainability By Joseph Murphy. (page 63) ISBN: 974-1-84407-345-0

(Goda et al., 2018) Goda Perlaviciute, Linda Steg, Nadja Contzen, Sabine Roeser and Nicole Huijts. Emotional Responses to Energy Projects: Insights for Responsible Decision Making in a Sustainable Energy Transition, Sustainability 2018, 10(7), 2526; https://doi.org/10.3390/su10072526

(Antti and Govert, 2023) Antti Silvast, Govert Valkenburg. Energy citizenship: A critical perspective. Energy Research & Social Science. Volume 98, April 2023, 102995. https://doi.org/10.1016/j.erss.2023.102995

Urry, J. (2014). The Problem of Energy. Theory, Culture & Society, 31(5), 3–20. doi:10.1177/0263276414536747

Strategy and policy, Green transition. [www document], [Accessed on 21.11.2023] Available at https://reform-support.ec.europa.eu/what-we-do/green-transition_en

(Clustering-GRETA, 2023) Aminul Islam (2023). Clustering-GRETA, Clustering Green Energy Transition Data (GRETA). https://github.com/aminuldidar/Clustering-GRETA. [Accessed on 21.11.2023]

(Trupti et al., 2013) Trupti M. Kodinariya, Dr. Prashant R. Makwana. Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies. Volume 1, Issue 6, November 2013, http://www.ijarcsms.com/

(Merriam, 2018) Merriam-Webster Online Dictionary, 2008. Cluster Analysis. [www document], [Accessed on 21.11.2023] Available at https://www.merriam-webster.com/dictionary/cluster%20analysis

(Anil Jain, 2009) Anil K. Jain (September 2009). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31 (2010) 651–666. http://www.elsevier.com/locate/patrec.

(Daniel, 2011) Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. arXiv:1109.2378v1 [stat.ML] 12 Sep 2011

(Bhattacharjee et al., 2021) Bhattacharjee, Panthadeep; Mitra, Pinaki (2021). A survey of density based clustering algorithms. Frontiers of Computer Science, 15(1), 151308–. doi:10.1007/s11704-019-9059-3

(Laura et al. 2007) Laura Olkkonen, Kristiina Korjonen-Kuusipuro, Iiro Grönberg. Redefining a stakeholder relation: Finnish energy “prosumers” as co-producers. https://doi.org/10.1016/j.eist.2016.10.004

(Kirsi et al., 2018) Kirsi Kotilainen, Ulla A. Saari. Policy Influence on Consumers’ Evolution into Prosumers—Empirical Findings from an Exploratory Survey in Europe. Sustainability 2018, 10, 186; doi:10.3390/su10010186

(Marianne et al., 2018) Marianne Ryghaug, Tomas Moe Skjølsvold, and Sara Heidenreich. Creating energy citizenship through material participation. Social Studies of Science, Volume 48, Issue 2, April 2018, Pages 283-303. https://doi.org/10.1177/0306312718770286

(Divya et al., 2022) Divya Chandrasenana, Reshma John Kuleenanb, Vaisakh Yesodharanc, Annie Feba Varghese. Clustering and exploring university students’ knowledge and attitude toward energy sustainability. ScienceDirect Energy Reports 8 (2022) 608–613.

(Lennon et al. 2019) Lennon, Breffní; Dunphy, Niall; Gaffney, Christine; Revez, Alexandra; Mullally, Gerard; O’Connor, Paul (2019). Citizen or consumer? Reconsidering energy citizenship. Journal of Environmental Policy & Planning, (), 1–14. doi:10.1080/1523908X.2019.1680277

(Devine-Wright, 2012) Patrick Devine-Wright. Energy citizenship: psychological aspects of evolution in sustainable energy technologies. Book, Chapter 4. Governing technology for sustainability, 2012

Meyer, Niels I. (2003). Distributed generation and the problematic deregulation of energy markets in Europe. International Journal of Sustainable Energy, 23(4), 217–221. doi:10.1080/01425910412331290724

(Elisabeth, 1996) Elisabeth Pendley. Deregulation of the Energy Industry. Land & Water Law Review, Vol. 31 [1996], Iss. 1, Art. 2. Published by Law Archive of Wyoming Scholarship, 1996.

(Rossella et al., 2022) Rossella Roversi, Andrea Boeri, Serena Pagliula and Giulia Turci. Energy Community in Action—Energy Citizenship Contract as Tool for Climate Neutrality. Department of Architecture, University of Bologna, 40136 Bologna, Italy. 2022, 5(1), 294-317; https://doi.org/10.3390/smartcities5010018

(Dumitru et al. 2023) Dumitru AC, Losada-Puente L, Peralbo M, Brenlla JC, Rebollo-Quintela N and García-Fernández M (2023) Mapping energy citizenship in the south of Europe. Front. Psychol. 14:1112457. doi 10.3389/fpsyg.2023.1112457

Creswell, John W. (1999). Handbook of Educational Policy || Mixed-Method Research. Handbook of educational policy, (), 455–472. doi:10.1016/B978-012174698-8/50045-X

(Wang et al., 2014) Wang, W., & Á., M. (2014). The Laplacian K-modes algorithm for clustering. ArXiv. /abs/1406.3895

(Williamson 1965) Williamson, O.E., 1965. Innovation and Market Structure. Journal of Political Economy 73, 67–73.

(Severin et al., 2000) Severin Borenstein, James Bushnell. Electricity Restructuring: Deregulation or Reregulation?. Is there a coherent vision for competitive electricity markets?. Volume 23, No. 2. Energy Information Administration, 2000.

(Anmol Singh, 2020) Anmol Singh (as2753) (ChemE 6800 Fall 2020). Heuristic algorithms. [www document], [Accessed on 21.11.2023] Available at https://optimization.cbe.cornell.edu/index.php?title=Heuristic_algorithms

(Horsbøl et al. 2018) Horsbøl, Anders (2018). Co-Creating Green Transition: How Municipality Employees Negotiate their Professional Identities as Agents of Citizen Involvement in a Cross-Local Setting. Environmental Communication, (), 1–14. doi:10.1080/17524032.2018.1436580

(Nouri et al., 2022) Nouri, Alireza, Shafi Khadem, Anna Mutule, Christina Papadimitriou, Rad Stanev, Mattia Cabiati, Andrew Keane, and Paula Carroll. 2022. "Identification of Gaps and Barriers in Regulations, Standards, and Network Codes to Energy Citizen Participation in the Energy Transition" Energies 15, no. 3: 856. https://doi.org/10.3390/en15030856

(Ajesh et al., 2023) Ajesh Kumar, Bilal Naqvi and Annika Wolf. Exploring the energy informatics and energy citizenship domains: a systematic literature review. Software Engineering, LENS, LUT University, 53850 Lappeenranta, Finland. https://doi.org/10.1186/s42162-023-00268-1

(Devine et al. 2007) P. Devine-Wright, in: J. Murphy (Ed.), Energy Citizenship: Psychological Aspects of Evolution in Sustainable Energy Technologies. Governing Technology for Sustainability, Earthscan, London, 2007, pp. 63–86, https://doi.org/10.4324/9781849771511

(Wilson et al., 2007) Wilson, Charlie; Dowlatabadi, Hadi . (2007). Models of Decision Making and Residential Energy Use. Annual Review of Environment and Resources, 32(1), 169–203. doi:10.1146/annurev.energy.32.053006.141137

Moscovici, Serge. (1984). The phenomenon of social representations. In: R. M. Farr, & S. Moscovici (Eds.), Social Representations. Cambridge: Cambridge University.

Massari, 2020. Ladder of Participation in the GRETA project.

Öhlén, Joakim. (2011). Janice M. Morse & Linda Niehaus (2009). mixed method design: Principles and procedures. 12.

(Emmanuel et al., 2021) Emmanuel, T., Maupong, T., Mpoeleng, D. et al. A survey on missing data in machine learning. J Big Data 8, 140 (2021). https://doi.org/10.1186/s40537-021-00516-9

(Svante et al., 1987) Svante Wold; Kim Esbensen; Paul Geladi. (1987). Principal component analysis. , 2(1-3), 37–52. doi:10.1016/0169-7439(87)80084-9

## Appendices

### Appendix 1

1. GitHub repository links of implemented code figure chart

https://github.com/aminuldidar/Clustering-GRETA

2. Filtered dataset by predefined criteria

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/filtered_by_manual_selection_and_70_percent_missing.csv

### Appendix 2

1. Column name those are more than twenty-five percent missing

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/Column%20name%20those%20are%20more%20than%2025%25%20missing.csv

2. Unique values of each variable

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/each_variable_unique_values.csv

### Appendix 3

1. Column with all values are numeric

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Documents/All%20values%20are%20numeric.pdf

2. A column with all values is numeric except one (not prefer to express an opinion) - the number of columns is 27

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Documents/All%20values%20are%20numeric%20except%20one.pdf

3. Columns with values are text and need to be encoded - the number of columns is 23

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Documents/All%20values%20are%20text.pdf

**Appendix 4**

1. Box plot of each variable

https://github.com/aminuldidar/Clustering-GRETA/blob/master/image_dep/box_plot_each_variable.png

2. Correlation Matrix
https://github.com/aminuldidar/Clustering-GRETA/blob/master/image_dep/Correlation_Matrix.png

3. Chi-Square Test Matrix
https://github.com/aminuldidar/Clustering-GRETA/blob/master/image_dep/Chi-Square%20Test%20Matrix.png

**Appendix 5**

1. K-means, deviation for each cluster

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/std_per_cluster_k_means.csv

2. K-means, skewness for each cluster

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/skewness_per_cluster_k_means.csv

**Appendix 6**

1. K-modes, centroids for each cluster

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/k_modes_cluster_centroids.txt

2. K-modes, cluster profile for each cluster and variables

https://github.com/aminuldidar/Clustering-GRETA/blob/master/Data/k_modes_cluster_profiling.txt

**Appendix 7**

1. Category frequency bar chart of every variable

https://github.com/aminuldidar/Clustering-GRETA/tree/master/image_dep/freq_chart_every_variable