NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

ChatGPT as a Data Scientist: Can AI Handle Clustering Better Than a Human?

Alexandru Lemesev

Master Thesis

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CHATGPT AS A DATA SCIENTIST: CAN AI HANDLE CLUSTERING BETTER THAN A HUMAN?

by

Alexandru Lemesev

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Supervisor: José Henrique Pereira São Mamede, PhD in Information Systems and Technologies, NOVA IMS

Supervisor: Vítor Manuel Duarte dos Santos, PhD in Science and Information and Technology Systems, NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisboa, 15 de julho de 2024

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisors, Henrique Mamede and Vitor Santos, for their invaluable guidance and advice throughout this entire process. Their expertise and constant availability to help were crucial in the completion of this thesis.

To my girlfriend, thank you for being a pillar of support during this journey. Your constant motivation and encouragement kept me going, even during the toughest times.

To my closest friends, your emotional support and endless humor have been indispensable. Whether through jokes or just being there when I needed a break, you have all played a significant role in helping me through this process.

I am immensely grateful to my family for the education they provided me, which has lined the way for me to reach where I am today. Their values and lessons have been my guiding light.

I would also like to extend my thanks to the data science community in Portugal. Your valuable insights and readiness to share your experience and knowledge have been instrumental in my research.

ABSTRACT

Artificial intelligence (AI) has rapidly advanced in recent years, becoming a significant trend in various fields, including data science. This thesis explores how ChatGPT, an AI language model, performs as a data scientist compared to a human counterpart. Using both quantitative and qualitative methods, the study evaluates ChatGPT's ability to handle essential data science tasks like code generation, code completion, and clustering analysis. Insights from a survey completed by data scientists in Portugal shed light on their experiences with AI tools and their views on AI's role in their work. The research also involved a hands-on comparison between ChatGPT and a human data scientist handling a clustering problem. The results show that while ChatGPT is efficient and capable in managing routine data science tasks, it struggles with more complex analyses and often requires human oversight. Ethical concerns, such as biases in AI outputs, were also addressed. The findings suggest that AI and human expertise complement each other in the field of data science. Although AI tools like ChatGPT can boost productivity and efficiency, they cannot entirely replace the nuanced decision-making and creative problem-solving skills of human data scientists. This research emphasizes the importance of continued development to overcome these limitations and better integrate AI into data science workflows.

KEYWORDS

ChatGPT; Data Science; Artificial Intelligence; Clustering; Python

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
CRISP-DM	Cross Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
PICOC	Population, Intervention, Comparison, Outcome, Context
RQ	Research Question
SLR	Systematic Literature Review
SSE	Sum of Square Errors
t-SNE	t-Distributed Stochastic Neighbor Embedding
WCSS	Within Cluster Sum of Squares

1 INTRODUCTION

In today's digital era, the importance of data science has grown tremendously, owing to the wealth of data available, advancements in computing power, and also machine learning. As such, navigating the vast and complex landscape of data from various sources causes a significant challenge for data scientists, especially with the rise of trends like machine learning, Big Data, and artificial intelligence (AI). Within this constantly evolving landscape, conversational AI interfaces, such as ChatGPT, have captured a lot of attention for their potential to democratize data science. These interfaces have shifted the focus from complex programming tasks to more streamlined problem-solving activities, making data science accessible to a wider audience (Lingo, 2023).

The work of Mahadi Hassan et al. (2023) introduces the revolutionary idea of ChatGPT as a "personal data scientist.". By envisioning an intelligent agent capable of conducting automated machine learning (AutoML) tasks through simple, natural conversations, their pioneering framework utilizes advanced large language models (LLMs) like ChatGPT to bridge the gap between users and machine learning models. This results in a seamless and effortless interaction between users and the system.

Furthermore, Feng et al. (2023) explore the code generation performance of ChatGPT through crowdsourcing social data, shedding light on its insights and patterns in code generation across several programming languages. They reveal the diverse applications of ChatGPT in code debugging, interview preparation, and academic assignment solving, while also highlighting prevalent emotions such as fear associated with its code generation capabilities.

Hassani & Silva (2023) discuss the role of ChatGPT in data science, emphasizing its potential to enhance productivity and accuracy in data science workflows by studying ChatGPT's capabilities in automating several aspects of the data science process, like data cleaning, model training, and result interpretation, while addressing concerns regarding bias and plagiarism.

Additionally, Hong (2023) provides a perspective on the potential applications of ChatGPT in computational materials science, examining its ability to assist with tasks such as building structures, writing codes for scientific software, and preparing data visualization scripts.

Despite the impressive advancements in data science, a clear void remains in the integration of ChatGPT into daily data science tasks (Hassani & Silva, 2023). This dissertation aims to fill this gap by offering an operational framework that delineates ChatGPT's function in data science, examining various practical implementations, and emerging into complex matters such as bias, contextual tracking, and interpretation of ChatGPT-generated results (Ray, 2023).

The methodology combines a rigorous literature review and an Action-Research methodology. Including insights gathered from interviews with domain experts and data scientists, that will provide a well-rounded perspective on the practical implications of ChatGPT in data science.

Notable findings highlight ChatGPT's role in natural language interfaces, semantic parsing, and text-to-SQL, highlighting its versatility in handling diverse linguistic tasks (Zhang et al., 2023).

Furthermore, recent studies examining the impact of ChatGPT on problem-solving skills and student-centered instruction offer valuable insights into its educational applications. These studies highlight the potential for ChatGPT to enhance pedagogical approaches and improve learning outcomes (Tsai, 2023). Notably, research in the field of education continues to explore the effectiveness of ChatGPT in a range of contexts, including its use in beginner-level Python courses and its integration into financial data science courses. This promising research showcases ChatGPT's ability to facilitate learning in various domains (Liu et al., 2023; Postari, 2023).

Current research has emerged into the capabilities of ChatGPT in advanced data analysis plugins for hydrological investigations, shedding light on both the advantages and limitations of such uses (Irvine et al., 2023). Further studies have delved into the realm of automatic code summarization and have shown ChatGPT's effectiveness in enhancing code comprehension tasks (Sun et al., 2023).

In opposition, other studies have focused on leveraging ChatGPT for sentiment analysis on Twitter data, revealing predominantly positive sentiments despite acknowledged shortcomings and paving the way for future enhancements (Erfina & Nurul, 2023). Notably, ChatGPT's potential application in computational materials science has also been extensively explored, underscoring its inherent strengths and possible applications in this field.

Latest investigations have delved into the function of ChatGPT as a personal data scientist and have uncovered both acknowledged limitations and potential advantages. These findings emphasize the importance of understanding the tool's capabilities in depth (Mahadi Hassan et al., 2023). Additionally, (Ray, 2023) an in-depth review provides a comprehensive overview of ChatGPT, covering its background, applications, key challenges, biases, ethics, limitations, and future potential.

Further exploration of ChatGPT's potential and limitations in science and engineering problem-solving has led to the examination of collaborative human-AI efforts. This has shed light on the complex dynamics of human-AI collaboration (Wang et al., 2023). However, despite these efforts, challenges remain, as evidenced by the research on ChatGPT's performance in the IEEEXtreme competition (Koubaa et al., 2023).

The inclusion of ChatGPT in the field of data science presents numerous avenues for research. Collaborations across disciplines are expanding, leveraging ChatGPT's conversational functionalities to facilitate communication and knowledge exchange across various domains.

Ethical implementation of ChatGPT is a key focus, with ongoing investigations into ethical frameworks and potential biases. Progress in natural language understanding is expected to enhance ChatGPT's ability to understand context and nuanced language, resulting in increased accuracy and dependability.

Integrating visual data is an emerging area, with the goal of enhancing ChatGPT's analytical capabilities. As ChatGPT becomes more essential to data science, prioritizing user-centric design and usability is crucial, requiring interfaces tailored specifically to its usage.

Each of these areas represents a significant step towards a more integrated, robust, and ethically conscious future in data science.

This study revolves around understanding how the problem-solving capabilities of a data scientist compare to those of an AI model, specifically ChatGPT, in the context of Data Science. The main objective is to provide a comprehensive understanding of ChatGPT's role in data science and to explore practical implementations that bridge theoretical concepts with real-world applications. With this in mind, the key Research Question (RQ) of this Dissertation is: "Is ChatGPT capable of outperforming a Data Scientist in his own domain?"

To answer this question, the following objectives were set:

1. Conduct a comparative analysis of a data scientist and ChatGPT in resolving a Data Science problem, considering the advancements in Machine Learning, Big Data, and Artificial Intelligence.
2. Evaluate the effectiveness, strengths, and limitations of both human and AI approaches.
3. Analyze the implications of the findings for the integration of ChatGPT in the daily data science tasks.

2 LITERATURE REVIEW

2.1 INTRODUCTION

A systematic literature review (SLR) gathers, evaluates, and synthesizes all available literature regarding a specific topic in a complete, unbiased, and continuous way by considering a specific review protocol. The guidelines provided by Kitchenham (2004) will be followed to make an SLR regarding Formal Literature. Initially, these guidelines were created to perform research in the medical field and posteriorly have been adapted for use in the field of Software Engineering. The protocol for SLR in Software Engineering will be used which is composed of essential key components for this review. An approach that is structured will be used to find the relevant studies using specific search terms and resources. The criteria for inclusion/exclusion of studies will be developed and pilot-tested to ensure consistency and repeatability. Checklists will be developed to assess the quality of the individual studies from the meta-analysis. A defined process of obtaining and validating data from each study will be used to provide certainty and dependability. A plan to align the retrieved data for drawing meaningful inferences will be developed. A plan will be developed for disseminating the results of the systematic review to relevant stakeholders. A schedule will be established for the conduct of the review, which clearly defines the key milestones and deadlines.

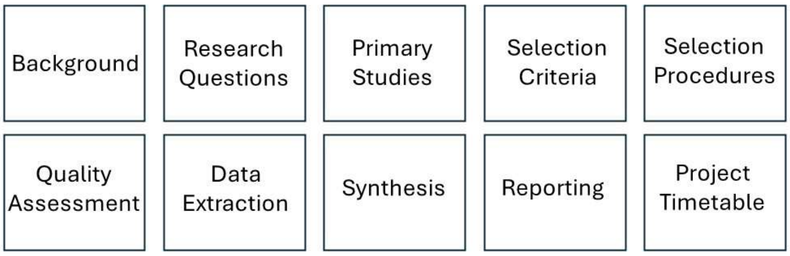

Figure 1 - Protocol Components

The systematic literature review (SLR) consists of three distinct phases: Planning, Conducting, and Reporting. During the first phase, Planning, the review protocol is meticulously explored, employing elements such as the research questions, the search string, the PICOC framework (Population, Intervention, Comparison, Outcome, Context), targeted sources, selection criteria, quality assessment checklist, and data extraction form. After establishing the protocol, the Conducting phase starts, which is related to the execution of the search string to identify sources, followed by the collection and filtering of results based on predefined selection criteria. The quality of selected studies is then judged using predetermined quality assessment questions, with some potentially disregarded if they fail to meet a predefined cut-off score. Subsequently, the content of retained studies is thoroughly reviewed and analyzed. Finally, the Reporting phase involves presenting the findings organized by the research question, ensuring a logical and coherent presentation of the review's results.

2.2 FORMAL LITERATURE REVIEW

As the field of data science continues to evolve, the integration of artificial intelligence (AI) tools such as ChatGPT has become increasingly prevalent. This section aims to provide a comprehensive formal literature review of the existing literature pertaining to ChatGPT's role as a data scientist.

2.3 PLANNING

Before the literature review itself, a research strategy was developed. Using the PICOC framework, the key elements, research questions, and the criteria for selecting studies were defined. This planning ensures that relevant literature is gathered and analyzed in an unbiased way and is in line with the research goals, ultimately ensuring a comprehensive and systematic review.

2.3.1 Research Questions

The purpose of this comprehensive systematic literature review is to examine and dissect the diverse responsibilities of ChatGPT as a data scientist. Special attention will be paid to its predispositions, capabilities in programming tasks, and applicability in different fields. In order to fully accomplish this objective and fill in the knowledge gaps outlined in the Research Background; the following research questions (RQ) have been formulated:

RQ1: What are the common biases and ethical considerations associated with ChatGPT in data analysis, and how do they impact the reliability and fairness of the generated insights?

RQ2: How effective is ChatGPT as a tool for assisting in programming tasks, such as code generation, code completion, and debugging, compared to traditional programming methods?

RQ3: How does ChatGPT demonstrate utility and effectiveness across various domains and what are the key factors influencing its applicability in these areas?

2.3.2 Data Sources and Search Strategy

The review process involved three distinct phases: Planning, Conducting, and Reporting. During the Planning phase, PICOC (Population, Intervention, Comparison, Outcome, and Context) framework was used as a helpful guide for my study. As shown in Table 1, this framework allowed me to carefully define the parameters of my research in an organized manner.

Furthermore, research questions were established, a search string, sources, selection criteria, and a quality assessment checklist, to ensure a thorough and methodical review.

To fully capture the scope of relevant literature, the research began with an exploratory search on SCOPUS with a considerable number of results (1117).

To further enrich the findings, The search was expanded to include additional sources such as Web of Science (294) and IEEE (262).

This attention to detail in the Planning phase served as a strong foundation for the following phases of the review, facilitating a structured and rigorous approach to data collection and analysis.

2.3.3 Study Selection Criteria

The criteria used to select studies for this research were carefully crafted to target the main topic at hand. This helps to ensure an unbiased selection process. The criteria are decided at the start of the research plan and can be adjusted as needed during the search phase.

Table 1 - PICOC Criteria

Population	Data Scientists and users of ChatGPT
Intervention	ChatGPT
Comparison	Traditional Data Analytics methods
Outcome	Bias and ethical considerations associated with ChatGPT; Effectiveness of ChatGPT in programming tasks; Utility and effectiveness of ChatGPT across various domains
Context	Data Science

The formal literature search process implicated systematically searching and collecting papers from various relevant bibliographic databases. Taking into account the PICOC criteria and the Research Questions, a standardized search string was created ("ChatGPT" OR "GPT" OR " Artificial Intelligence") AND ("Data Science" OR "Bias" OR "Ethic" OR "Programming") across all selected databases to ensure consistency and thoroughness in the search process.

During the literature review process, thorough attention was allocated to examining the title and abstract of each paper, and by applying predetermined criteria for inclusion and exclusion, the papers' pertinence was assessed to the study. These specific criteria were created to promoted consistent interpretation and precise identification of relevant studies, supporting in selecting those best suited to address the research questions (Kitchenham, 2004).

Table 2 - Study Selection Criteria

Inclusion Criteria	Exclusion Criteria
Studies that discuss the impact of ChatGPT on data science	Studies not directly related to ChatGPT or its applications in data science and programming tasks.
Studies exploring the utility and effectiveness of ChatGPT across various domains.	Not in Portuguese nor in English
Studies related with Research Questions	Studies published before 2022
	Article not peer-reviewed
	Article not published on scientific journal or conference proceedings

2.3.4 Quality Assessment

The quality of each study was rigorously evaluated against a set of specific criteria, as outlined above. A thorough examination of the study's content was conducted to determine its overall value and relevance to my research objectives.

Each criterion was accurately assessed and scored, considering the degree to which it was met. These scores were then combined to generate a weighted sum for each study, allowing for a comprehensive quantitative analysis of its quality.

Table 3 - Quality Assessment Criteria

Criteria	Weight
Does the study primarily focus on the application or usage of ChatGPT within the context of data science?	1
Literature Review was performed?	0.5
Does the study discuss potential bias and ethical problems related with ChatGPT?	1
Does the study add something new?	1
Does the study directly address or contribute to answering the research questions?	0.5
Is the methodology utilized in the study clearly described?	0.5
Are the limitations of the study acknowledged and discussed?	0.5
Does the study offer practical implications or recommendations for utilizing ChatGPT in data science applications?	1
Cut-off score: 3.5	

2.4 CONDUCTING

The Conducting phase of the review involves a structured process of selecting, extracting, assessing quality, and synthesizing information from the identified studies. This phase ensures adherence to the predefined inclusion and exclusion criteria while systematically handling the gathered literature.

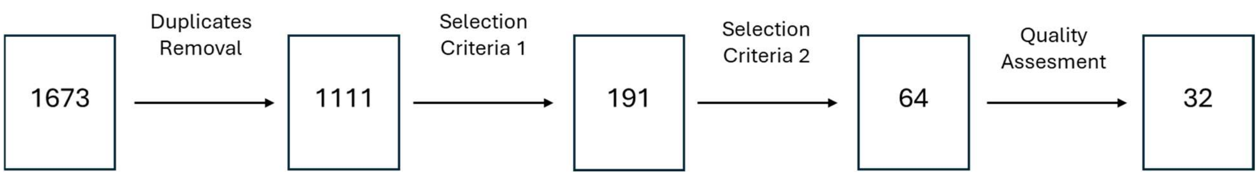

Figure 2 - SLR Flowchart

Table 4 - Number of included and excluded papers

Inclusion Criteria		Exclusion Criteria	
24	Studies that discuss the impact of ChatGPT on data science	Studies not directly related to ChatGPT or its applications in data science and programming tasks.	127
19	Studies exploring the utility and effectiveness of ChatGPT across various domains.	Not in Portuguese nor in English	143
21	Studies related with Research Questions	Studies published before 2022	337
		Article not peer-reviewed	302
		Article not published on scientific journal or conference proceedings	138

The Selection Criteria were split into two phases considering the number of articles that were found. In the first phase, direct filters were applied like the year of the study or the languages. During the second phase, more qualitative criteria were applied which required filtering the papers by their title, abstract, and sometimes the conclusion.

Table 5 - Journal Articles

Publication	Number of Publications
ACM Computing Surveys	1
Association for Computing Machinery	2
Automated Software Engineering, Springer	1
Big Data & Society	1
Big Data and Cognitive Computing, MDPI	1
Computers & Operations Research, Elsevier	1
Data & Metadata	1
Energy Material Advances	1
Expert Systems with Applications, Elsevier	1
Finance Research Letters, Elsevier	1
Heliyon 9, CellPress	1
HUMANITIES AND SOCIAL SCIENCES COMMUNICATIONS	2
International Journal of Information and Learning Technology	1
International Journal of Information Technology and Computer Engineering (IJITC)	1
Internet of Things and Cyber-Physical Systems, KeAi	1
Journal of Internet Services and Information Security (JISIS)	1
Machine Learning with Applications, Elsevier	1
Multidisciplinary Digital Publishing Institute	1
Quality & Quantity, Springer	1
TechTrends, Springer	1
The International Journal of Management Education, Elsevier	2
Wiley	1

Table 6 - Conferences

Conference	Number of Publications
2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)	1
Annual International Conference on Digital Government Research 2023	1
Conversational User Interfaces Conference 2023	2
Innovation and Technology in Computer Science Education 2023	1
International Conference of Innovative Technologies and Learning (ICITL) 2023	1
SBES '23: Proceedings of the XXXVII Brazilian Symposium on Software Engineering	1
Special Interest Group on Information Retrieval - Asia Pacific 2023	1

After the selection criteria and quality assessment, a total of 32 papers were identified for inclusion in the review.

2.5 Reporting the Review

The final phase of the systematic review involves articulating the results and circulating them to relevant stakeholders. Effective communication of the review findings is crucial, requiring a well-defined dissemination strategy established either during the commissioning stage or when preparing the review protocol.

In this section, the key findings from the systematic review are synthesized and analyzed. Each subsection corresponds to a specific research question posed in the literature review, presenting primary findings relevant to each question alongside corresponding references.

2.5.1 Systematic Literature Review Research Question 1

SLRQ 1: What are the common biases and ethical considerations associated with ChatGPT in data analysis, and how do they impact the reliability and fairness of the generated insights?

The integration of artificial intelligence (AI), particularly exemplified by technologies like ChatGPT, into various facets of society, presents both opportunities and challenges, often entangled with ethical considerations and biases. A comprehensive understanding of prevalent biases and ethical considerations is crucial for assessing the reliability and fairness of insights generated by AI systems like ChatGPT.

Bakiner (2023) emphasizes the need to harness the human rights framework to mitigate the negative impact of AI, including bias, privacy infringement, and threats to civil-political rights. This highlights the importance of regulatory oversight and ethical guidance, as exemplified by the EU's efforts in legislating digital and AI regulation (Polyportis & Pahos, 2024). However, there exists a gap in AI-related human rights legislation, necessitating coordinated efforts to address emerging challenges (Bakiner, 2023).

Moreover, responsible innovation and stakeholder engagement are essential for navigating the risks of AI, as discussed by Polyportis & Pahos (2024).

They advocate for a multi-stakeholder approach to AI development and emphasize the integration of values and ethics to ensure societal alignment. In their work, Richardson et al. (2024) emphasize the critical role of responsible innovation, ethical technological advancement, and community-building in shaping the future of our society amid the rapid advancements in technology, particularly in the realm of AI.

Stahl & Eke (2024) explore the ethical issues surrounding ChatGPT, highlighting concerns related to social justice, individual needs, culture, identity, and environmental impacts.

They focus on the need for proactive measures to address these concerns and promote ethical AI development.

One crucial aspect that must be addressed is the occurrence of hallucinations. Zuccon et al. (2023) explain that these are attributed to perceptions that appear plausible but have no actual basis in reality. When discussing ChatGPT, this poses a significant concern, as their research shows that it has the potential to generate false information, thus undermining its reliability. This can have serious consequences, such as receiving erroneous legal advice, which may lead the user to unintentionally break the law (Zhan et al., 2023).

Additionally, Shahbazi et al. (2023) discuss techniques for identifying and resolving representation bias in data, a critical aspect in ensuring fairness and reliability in AI-generated insights. They emphasize the importance of addressing representation bias across different data types and the development of comprehensive metrics for measuring bias.

Table 7 - Publications used for RQ1

Year	Author(s)	Article/Conference
2023	Bakiner, Onur	The promises and challenges of addressing artificial intelligence with human rights
2024	Elsadig, M.A.	ChatGPT and Cybersecurity: Risk Knocking the Door
2022	Hauer, Tomas	Incompleteness of moral choice and evolution towards fully autonomous AI
2024	Polyportis, A.; Pahos, N.	Navigating the perils of artificial intelligence: a focused review on ChatGPT and responsible research and innovation
2023	Ray, Partha Pratim	ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope
2024	Richardson, C.; Oster, N.; Henriksen, D.; Mishra, P.	Artificial Intelligence, Responsible Innovation, and the Future of Humanity with Andrew Maynard
2023	Zhan, Xiao; Xu, Yifan; Sarkadi, Stefan	Deceptive AI Ecosystems: The Case of ChatGPT
2023	Zuccon, Guido; Koopman, Bevan; Shaik, Razia	ChatGPT Hallucinates when Attributing Answers
2023	Shahbazi, Nima; Lin, Yin; Asudeh, Abolfazl; Jagadish, H. V.	Representation Bias in Data: A Survey on Identification and Resolution Techniques
2024	Stahl, Bernd Carsten; Eke, Damian	The ethics of ChatGPT – Exploring the ethical issues of an emerging technology
2023	Tsai, Y.-C.	Empowering Learner-Centered Instruction: Integrating ChatGPT Python API and Tinker Learning for Enhanced Creativity and Problem-Solving Skills.

In sum, it is crucial to address the general bias and ethical concerns surrounding ChatGPT and other AI systems in order to sustain their dependability, impartiality, and alignment with society's values. This demands collective involvement from various parties, regulatory measures based on principles of human rights, and responsible advancement practices guided by ethical principles. Additionally, continual research and progress are critical in tackling emerging obstacles and promoting ethical AI practices for the future (Shahbazi et al., 2023).

Table 8 - Contributions for RQ1

Year	Author(s)	Concepts/ Ideas/ Contributes
2023	Bakiner, Onur	Representative bias from minorities, the legal framework for AI
2024	Elsadig, M.A.	Concern on ChatGPT regarding Ciber Security; Bias; Social Engineering
2022	Hauer, Tomas	Ethical Concern regarding the Weaponization of ChatGPT
2024	Polyportis, A.; Pahos, N.	Negative impacts of ChatGPT
2023	Ray, Partha Pratim	Generative AI definition; ChatGPT definition; Issues of ChatGPT; Controversial Stories; ChatGPT in computer ethics; Biases and limitations
2024	Richardson, C.; Oster, N.; Henriksen, D.; Mishra, P.	Future of Humanity with AI
2023	Zhan, Xiao; Xu, Yifan; Sarkadi, Stefan	Human-centered computing
2023	Zuccon, Guido; Koopman, Bevan; Shaik, Razia	Hallucinations
2023	Shahbazi, Nima; Lin, Yin; Asudeh, Abolfazl; Jagadish, H. V.	Bias in NLP
2024	Stahl, Bernd Carsten; Eke, Damian	ChatGPT Ethical Issues
2023	Tsai, Y.-C.	ChatGPT enhances students’ sense of accomplishment and promotes deeper thinking through qualitative analysis

2.5.2 Systematic Literature Review Research Question 2

SLRQ 2: How effective is ChatGPT as a tool for assisting in programming tasks, such as code generation, code completion, and debugging, compared to traditional programming methods?

This Research Question aims to evaluate the effectiveness of ChatGPT as a tool for assisting in programming tasks compared to traditional programming methods, the goal is to delve into various aspects including code generation, code completion, and debugging.

- **Effectiveness of ChatGPT in Programming Tasks**

Code Generation:

The experimental results presented by Calvo-Pardo et al. (2023) indicate that ChatGPT, specifically ChatGPT-4, demonstrates proficient abilities for code generation tasks. The study evaluated the performance of different large language models (LLMs) in generating Python code solutions to a set of certified problems, revealing that GPT-4 exhibited the highest success rate among the models tested.

Additionally, the study highlighted that GPT-based models tend to produce shorter and more concise code compared to non GPT-based models, resulting in more efficient solutions.

Code Completion:

In their research, Bucaioni et al. (2024) delved into the potential of ChatGPT to tackle programming bugs. Their findings showed how this tool can assist in code completion and bug prediction. ChatGPT's impressive ability to comprehend and interpret code snippets, coupled with its natural language generation capabilities, make it a promising tool for supporting developers in their coding endeavors. The results indicate that ChatGPT can provide prompt and effective bug explanations and predictions, potentially serving as a valuable resource for developers seeking to complete code segments and resolve bugs.

Debugging:

Surameery & Shakor (2023) investigated the use of ChatGPT for solving programming bugs, highlighting its potential role in debugging assistance. Its capability to understand code snippets and provide bug explanations can be advantageous in identifying and fixing bugs in new code snippets. However, it's important to note that while ChatGPT can offer debugging assistance, it may not be a perfect solution and should be used in conjunction with other debugging tools and techniques to ensure the best possible results.

- **Comparison with Traditional Programming Methods**

Cost and Speed:

Compared to traditional debugging tools and techniques, ChatGPT is often provided as a cloud-based service with a more flexible pricing model (Surameery & Shakor, 2023). Additionally, ChatGPT can provide quick and efficient bug explanations and predictions, potentially offering faster results compared to traditional debugging tools that may take longer to run.

Accuracy and Customizability:

Traditional debugging tools may offer a higher degree of accuracy in bug predictions but require a deeper understanding of the code (Surameery & Shakor, 2023). On the other hand, ChatGPT's bug predictions and explanations may be impacted by the quality of its training data but offer ease of use and natural language generation capabilities.

Integration and Scalability:

Traditional debugging tools can integrate with other tools and systems, whereas ChatGPT may not offer the same level of integration (Surameery & Shakor, 2023). However, ChatGPT can be used to debug code at scale, potentially offering scalability benefits compared to traditional debugging tools that may struggle with large and complex codebases.

In conclusion, ChatGPT shows promise as a tool for assisting in programming tasks such as code generation, code completion, and debugging. While it offers advantages in terms of cost, speed, and ease of use, it may not match the accuracy and customizability of traditional programming methods.

Ultimately, understanding the unique strengths of both human and automated approaches in software engineering tasks will facilitate more effective collaboration and task allocation processes (Nathalia et al., 2023; Surameery & Shakor, 2023).

Table 9 - Publications used for RQ2

Year	Author(s)	Article/Conference
2024	Bucaioni, Alessio; Ekedahl, Hampus; Helander, Vilma; Nguyen, Phuong T.	Programming with ChatGPT: How far can we go?
2024	Coello, Carlos Eduardo Andino; Alimam,	Effectiveness of ChatGPT in Coding: A Comparative Analysis of Popular Large Language Models

	Mohammed Nazeh; Kouatly, Rand	
2023	Diaz-de-Arcaya, Josu; Torre-Bastida, Ana I.; Zárate, Gorka; Miñón, Raúl; Almeida, Aitor	A Joint Study of the Challenges, Opportunities, and Roadmap of MLOps and AIOps: A Systematic Survey
2023	Ouh, Eng Lieh; Gan, Benjamin Kok Siew; Jin Shim, Kyong; Wlodkowski, Swavek	ChatGPT, Can You Generate Solutions for my Coding Exercises? An Evaluation on its Effectiveness in an undergraduate Java Programming Course.
2023	Surameery, Nigar M. Shafiq; Shakor, Mohammed Y.	Use Chat GPT to Solve Programming Bugs

Table 10 - Contributions for RQ2

Year	Author(s)	Concepts/Ideas/ Contributes
2024	Bucaioni, Alessio; Ekedahl, Hampus; Helander, Vilma; Nguyen, Phuong T.	Definition of Natural Language Processing (NLP), Hability of ChatGPT in writing code
2024	Coello, Carlos Eduardo Andino; Alimam, Mohammed Nazeh; Kouatly, Rand	Generative AI History, Comparison on code generation between different LLMs, GPT 4 has the best results
2023	Diaz-de-Arcaya, Josu; Torre-Bastida, Ana I.; Zárate, Gorka; Miñón, Raúl; Almeida, Aitor	AIOps (best practices)
2023	Ouh, Eng Lieh; Gan, Benjamin Kok Siew; Jin Shim, Kyong; Wlodkowski, Swavek	ChatGPT generate java code;
2023	Surameery, Nigar M. Shafiq; Shakor, Mohammed Y.	Advantages of ChatGPT in debugging

2.5.3 Systematic Literature Review Research Question 3

SLRQ 3: How does ChatGPT demonstrate utility and effectiveness across various domains and what are the key factors influencing its applicability in these areas?

- **Coding**

Regarding hydrological sciences, ChatGPT demonstrates utility by generating Python scripts for data analysis tasks, offering opportunities for rapid script setup and problem-solving abilities (Irvine et al., 2023). Its effectiveness in coding tasks lies in its ability to understand and generate code snippets based on natural language prompts, thereby simplifying the coding process for users with varying levels of programming expertise.

- **Education**

ChatGPT is a powerful tool in educational environments, aiding learners with explanations, summaries, and answers for a wide range of subjects (Erfina & Nurul, 2023). Its impact in education lies in its capacity to produce responses that resemble human ones, elevating the learning process. Through advanced natural language processing, ChatGPT enriches individual learning journeys, enabling students to readily access educational materials and knowledge (Rejeb et al., 2024; Zeb et al., 2024).

- **Finance**

The financial world is constantly evolving, making it challenging for investors to keep up with the rapidly changing landscape. This is where ChatGPT comes in, offering valuable assistance in the form of real-time investment advice and evaluation of financial information. (Pelster & Val, 2024) have proven its effectiveness through a strong correlation between ChatGPT ratings and future stock returns.

With its ability to analyze enormous amounts of financial data and news articles, ChatGPT empowers investors to make informed decisions, ultimately enhancing their investment strategies and optimizing portfolio performance.

- **Healthcare**

The potential of ChatGPT in healthcare is both exciting and promising. Through its efficient analysis of patient data and medical information, it can assist healthcare professionals in making preliminary diagnoses based on symptoms (Bano et al., 2024; Caruccio et al., 2024). Its advanced natural language processing abilities allow it to interpret medical terminology and generate proactive responses, making it a valuable tool in the diagnostic process within healthcare settings. Overall, ChatGPT has the capability to significantly streamline the identification of potential medical conditions and the recommendation of appropriate courses of action.

- **Customer Service**

In customer service contexts, ChatGPT engages with users and provides satisfactory interactions, as evidenced by positive sentiments on platforms like Twitter (Erfina & Nurul, 2023).

Its effectiveness is demonstrated through its ability to understand and respond to user queries promptly and accurately. ChatGPT's conversational abilities and natural language understanding enhance customer experiences by providing timely assistance and addressing inquiries effectively.

- **Key Factors Influencing Applicability**

In various fields, important considerations for ChatGPT's relevance include its capacity to generate accurate responses, adapt to specific tasks or prompts, and proficiently comprehend natural language input (Bano et al., 2024; Pelster & Val, 2024). Beyond these, the availability of data, training procedures, and user interface design (Skjuve et al., 2023) also significantly impact ChatGPT's practicality and efficacy across domains.

Table 11 - Publications for RQ3

Year	Author(s)	Article/Conference
2023	Albonico, Michel; Varela, Paulo Júnior	A Report on the Use of ChatGPT in Software Engineering and Systems Analysis Courses
2024	Bano, M.; Hoda, R.; Zowghi, D.; Treude, C.	Large language models for qualitative research in software engineering: exploring opportunities and challenges
2024	Biju, Ajitha Kumari Vijayappan Nair; Thomas, Ann Susan; Thasneem, J.	Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere—a bibliometric analysis
2024	Caruccio, Loredana; Cirillo, Stefano; Polese, Giuseppe; Solimando, Giandomenico; Sundaramurthy, Shanmugam; Tortora, Genoveffa	Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot
2023	Erfina, Adhitia; Nurul, Muhamad	Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language
2023	Irvine, Dylan J.; Halloran, Landon J. S.; Brunner, Philip	Opportunities and limitations of the ChatGPT Advanced Data Analysis plugin for hydrological analyses
2024	Pelster, Matthias; Val, Joel	Can ChatGPT assist in picking stocks?
2023	Skjuve, Marita; Følstad, Asbjørn; Brandtzaeg, Petter Bae	The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users
2024	Rejeb, A.; Rejeb, K.; Appolloni, A.; Treiblmaier, H.; Iranmanesh, M.	Exploring the impact of ChatGPT on education: A web mining and machine learning approach
2024	Zeb, A.; Ullah, R.; Karim, R.	Exploring the role of ChatGPT in higher education: opportunities, challenges, and ethical considerations

Table 12 - Contributions for RQ3

Year	Author(s)	Concepts/Ideas/ Contributes
2023	Albonico, Michel; Varela, Paulo Júnior	ChatGPT Definition
2024	Bano, M.; Hoda, R.; Zowghi, D.; Treude, C.	Ethical Issues; Hallucinations; Model Drift; Prompt Engineering; Importance of Humanity for AI
2024	Biju, Ajitha Kumari Vijayappan Nair; Thomas, Ann Susan; Thasneem, J.	Roots of AI; AI definition; ML definition; application of AI and ML by financial institutions
2024	Caruccio, Loredana; Cirillo, Stefano; Polese, Giuseppe; Solimando, Giandomenico; Sundaramurthy, Shanmugam; Tortora, Genoveffa	Bard outperformed ChatGPT in medical diagnosis
2023	Erfina, Adhitia; Nurul, Muhamad	52% have a positive sentiment regarding ChatGPT
2023	Irvine, Dylan J.; Halloran, Landon J. S.; Brunner, Philip	Advanced Data Analysis plugin
2024	Pelster, Matthias; Val, Joel	ChatGPT definition; LLM applications in Finance; GPT 4 has 99% in financial literacy test; WebChatGPT definition
2023	Skjuve, Marita; Følstad, Asbjørn; Brandtzaeg, Petter Bae	Positive psychological impact; Irrelevant and useless output was the higher poor user experience
2024	Rejeb, A.; Rejeb, K.; Appolloni, A.; Treiblmaier, H.; Iranmanesh, M.	Applied vs Generative AI; ChatGPT definition; Proximal Policy Optimization; Bias
2024	Zeb, A.; Ullah, R.; Karim, R.	History of OpenAI; AGI definition

2.6 DISCUSSION

In this section the key findings of the systematic literature review are summarized, focusing on ChatGPT's role as a data scientist and its implications across different domains.

After exploring ChatGPT's impact on data science, many concerns regarding biases and ethics were uncovered. These range from potential violations of privacy to threats against civil and political rights, as well as larger issues of social justice. It's evident that strict regulatory oversight and ethical guidelines are crucial in guiding the responsible development and implementation of AI-powered conversational agents like ChatGPT.

When it comes to its efficiency in programming tasks, ChatGPT showcases impressive proficiency in various areas, such as code generation, completion, and debugging. Research suggests that it has the potential to offer cost-effectiveness, speed, and user-friendliness compared to traditional programming methods. However, it falls short in comparison to human programmers.

ChatGPT offers promise in a variety of domains such as education, finance, healthcare, and customer service, going beyond just programming. Its potential applications include generating scripts for hydrological sciences, providing personalized educational support, offering financial guidance, aiding in healthcare diagnosis, and improving user interaction in customer service settings. The effectiveness of ChatGPT in each domain depends on factors like performance, adaptability, and user interface design. Through analyzing different research questions and domains, common themes and challenges emerge. ChatGPT's strengths lie in its ability to generate accurate responses, cater to specific tasks, and effectively process natural language. However, it also faces challenges such as mitigating biases, addressing ethical concerns, and ensuring fair and equal treatment for all users.

In conclusion, the existing research has been crucial in establishing ChatGPT as a valuable tool in various fields, particularly in education. However, further exploration and development of ChatGPT's capabilities are necessary, especially in the specialized field of data science. It is clear that there is a lack of substantial studies in this area, primarily due to their narrow focus on specific issues.

Thus, there exists a pronounced lacuna in the comprehensive exploration of ChatGPT's potential within the sophisticated landscape of data science. Consequently, this underscores the pressing need for further concerted efforts to elucidate and advance ChatGPT's capabilities within this domain, thereby fostering its efficacy and applicability in addressing the multifaceted challenges inherent to data-driven disciplines. Such efforts not only contribute to the refinement of ChatGPT's utility but also hold promise for augmenting the broader landscape of artificial intelligence applications in data science and related fields.

3 METHODOLOGY

For this research, a mixed methodology will be conducted. Starting with the quantitative research, a survey will be performed in order to understand better the relationship between Data Scientists and Artificial Intelligence. After that, the Action-Research Methodology will be used, as proposed by Santos et al. (2013), to systematically compare the performance of a data scientist and an AI model, ChatGPT, in solving a Data Science problem. Action-Research can be described as a family of research methodologies that incorporate simultaneous action (or change) and investigation (or understanding), utilizing a cyclical or spiral process that alternates between action and critical reflection. This approach allows for intervention in the research entity and the analysis of results, facilitating an open approach to the research field and enabling the capture of information that may not be predetermined.

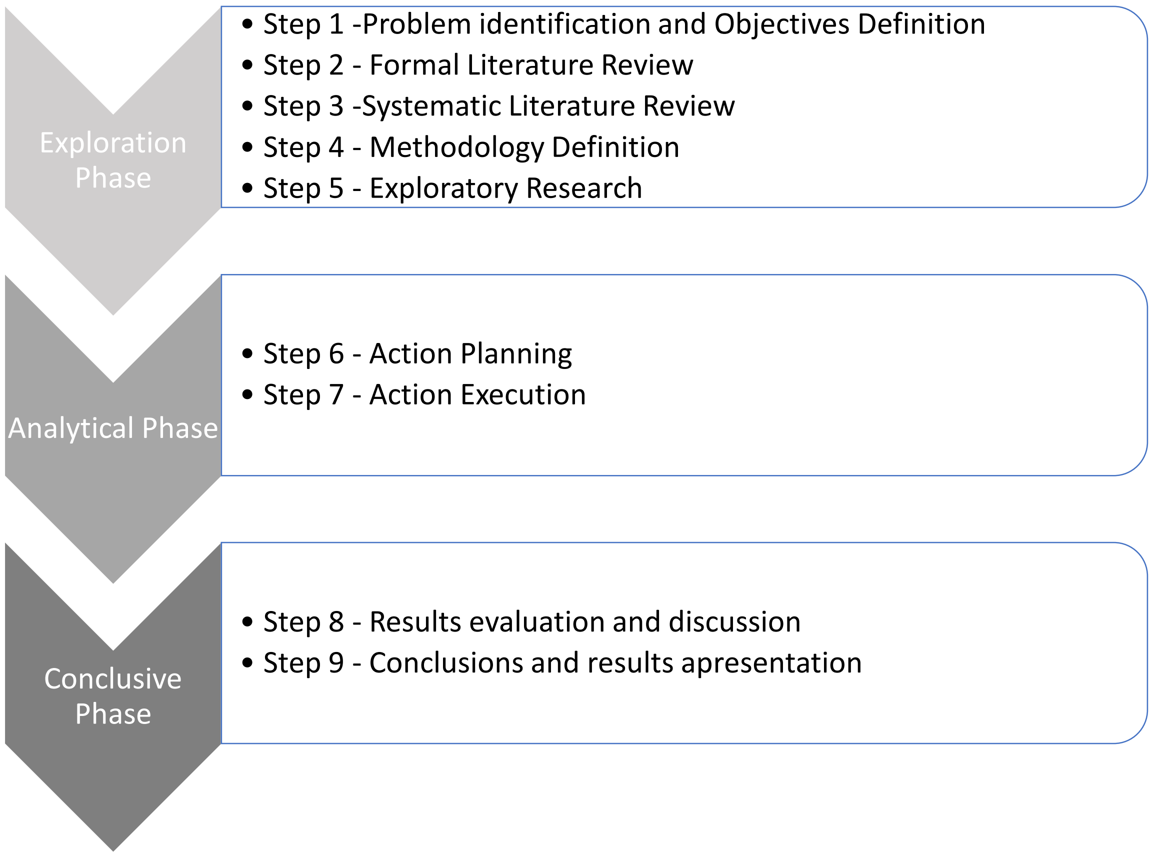

Figure 3 - Study Structure

The action research methodology consists of several iterative cycles, typically encompassing the following phases: diagnosis, action planning, action execution, evaluation, and specific learning/documentation. Each cycle involves a systematic progression through these phases, with the opportunity for refinement and adjustment based on the insights gained from previous cycles.

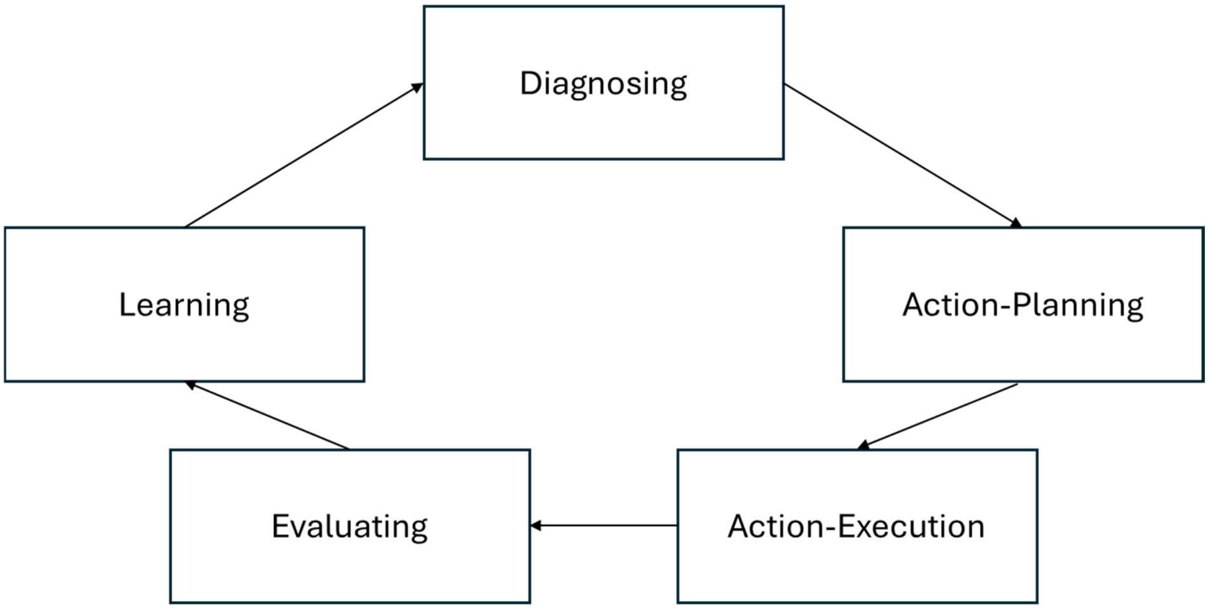

Figure 4 - Action Research Process (Adapted from Santos et al., 2013)

During the diagnosis phase, the research will aim to gain a comprehensive understanding of the problem at hand: the effectiveness of ChatGPT versus a data scientist in data science problem-solving. Regarding the following two phases, a cycle will be performed individually for ChatGPT and another one for the Data Scientist.

In the action execution phase, the data scientist will first solve the data science problem using Python, the chosen data set, and the defined evaluation metrics. Subsequently, ChatGPT will solve the same problem using the same tools, data set, and evaluation metrics. Following the action phase, the evaluation phase aims to assess the effectiveness of both approaches in solving the data science problem. Critical analysis will be conducted to evaluate the performance of ChatGPT and the data scientist based on the defined evaluation metrics. Also, during this phase, qualitative research will be done, by doing interviews with Data Scientists in order to evaluate both performances.

Finally, in the specific learning/documentation phase, the research will identify and document the conclusions resulting from the comparison between ChatGPT and the data scientist. The subsequent sections of this research will follow the principles and framework of the Action-Research methodology, guiding the exploration, analysis, and interpretation of findings within each phase.

The Action-Research methodology provides a structured approach for investigating and improving practices through iterative cycles of planning, action, observation, and reflection (Santos et al., 2013). This methodology will guide the research in systematically comparing the performance of a data scientist and ChatGPT in solving a data science problem. Each phase of the Action-Research process will be detailed, illustrating how the methodology enables comprehensive analysis and continuous learning, in the rest of this document.

4 STUDY

This chapter follows the Action Research methodology, starting with exploratory research and diagnosis, moving through action planning and execution, and ending with evaluation, to compare how ChatGPT and a data scientist solve a clustering problem.

4.1 Exploratory Research

As part of the research, a survey was directed to data scientists to gain deeper insights into their current relationship with artificial intelligence (AI). A study conducted by Eurostat in 2024 on the use of AI in European companies in 2023, positioned Portugal (7.9%) slightly below the European average (8.0%), inspiring the need to understand whether data scientists utilize AI in Portugal. This was the motive for focusing the survey on this demographic to gain a more detailed understanding of their engagement with AI technologies. Before distribution, the survey was validated by 21 data scientists, who provided valuable feedback that originated several changes and improvements, ensuring the survey was clear and comprehensive. A total of 263 responses were collected.

The survey began by gathering demographic information about the data scientists to obtain insights into their gender, age, field of study, and years of experience in their area. The gender distribution showed a predominance of male respondents (67%), followed by females (31%), and a small percentage identifying as non-binary or third gender (2%). The age range of the participants varied, with the majority (58%) being between 25 and 34 years old, followed by 21% aged 35 to 44, and 17% aged 18 to 24. The geographical distribution revealed a higher concentration of respondents in Lisboa (58%), with Setúbal (17%) and Porto (6%) also represented. Regarding educational background, a large proportion of respondents had a Master's degree (63%), followed by those with a Bachelor's degree (24%) and a Ph.D. or higher (9%). In terms of the relationship with AI, 49% stated that they consider themselves AI users, 10% as developers of AI models, and 40% considered both users and developers.

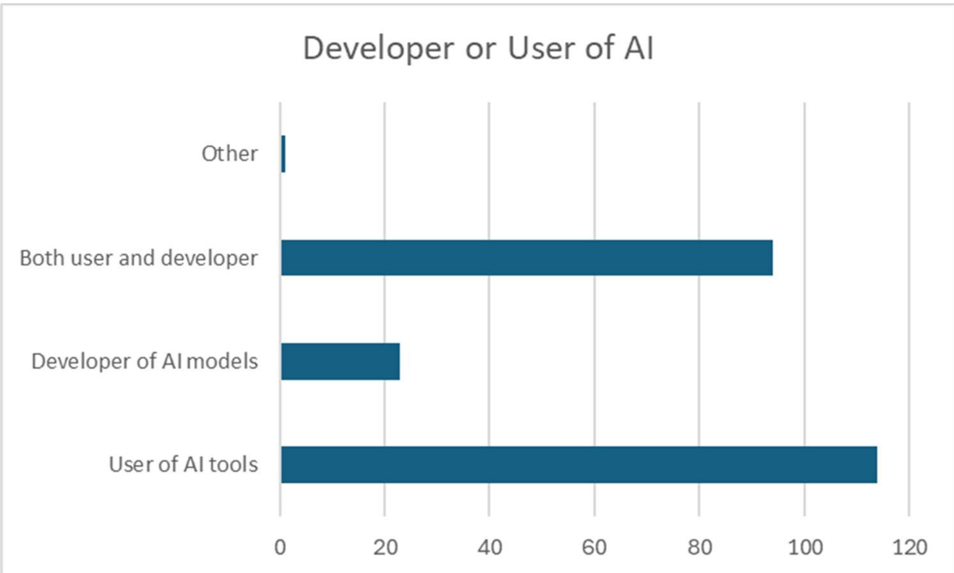

Figure 5 - Developer or User of AI

Given the low percentage of companies in Portugal using AI, the second section focused on this topic, and for those companies not currently using AI, it inquired whether they intended to start using artificial intelligence soon. Of the total respondents, 88% reported using AI tools, reducing the sample to 232 respondents for the remaining questions. Among those not using AI, 72% indicated that their companies were considering introducing AI technologies soon.

After this, the focus passed to the perception that they had regarding AI. As highlighted by Hassani & Silva (2023), the numerous applications of AI, including ChatGPT, can simplify the current work of data scientists and enable them to focus on new challenges. On the other hand, Ray (2023) lists a wide range of concerns, such as bias, ethical considerations, or hallucinations Zuccon (2023). The survey explored the freedom to use AI-powered conversational agents, with 90.34% of respondents affirming they had the freedom to use such tools. Regarding the perception of AI's role, 40.64% viewed AI as augmenting their current work, while 54.79% saw it as an opportunity for new challenges and only a small fraction (0.91%) perceived AI as a threat. Regarding the interaction with AI technologies, most data scientists used more AI for data analysis (36%), followed by task automation (32%) and task generation (27%).

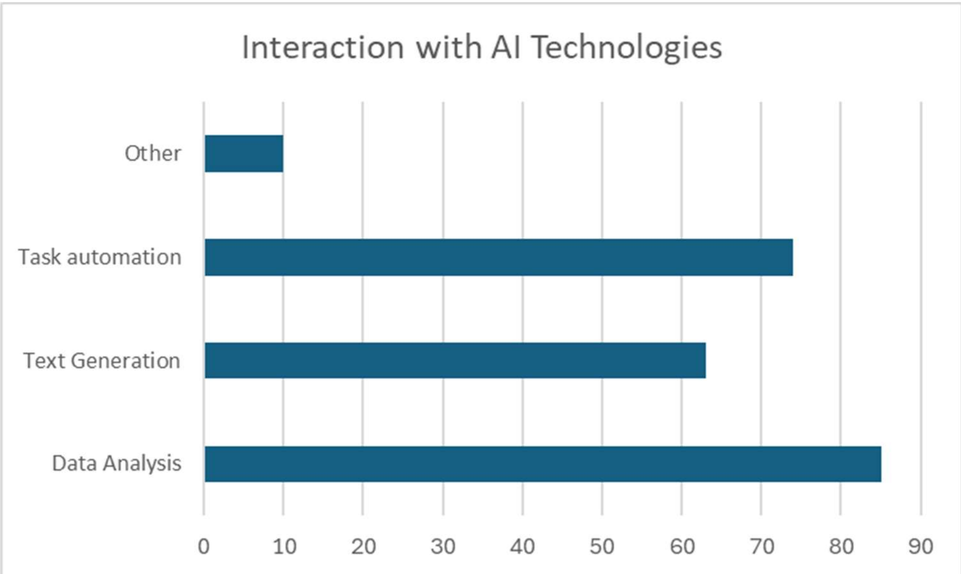

Figure 6 - Interaction with AI Technologies

The next section, related to the direct impact on work, aimed to understand the purposes for which workers use AI, like text generation and task automation (Hassani & Silva, 2023), or data analysis as mentioned by Irvine et al. (2023). When asked about the impact of AI technologies on their work processes, 57% reported increased productivity, 35% noted improved efficiency, a small number (6%) indicated no significant impact and only one respondent reported decreased productivity.

On the other hand, the topic of challenges and limitations experienced by data scientists was addressed, such as issues in data quality or lack of interpretability (Irvine et al., 2023). Respondents mentioned several concerns about using AI in their work.

The most significant concerns included bias (38%), hallucinations (27%), and ethical considerations (23%). Job displacement was a concern for only 8% of respondents.

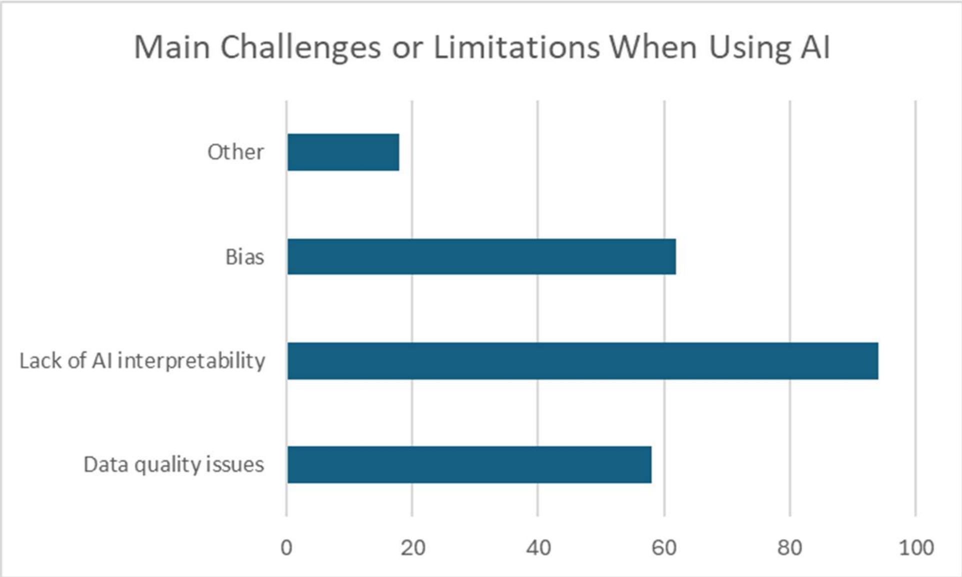

Figure 7 - Challenges and Limitations

Since some of these limitations can be improved with training, an attempt was made to understand if companies have provided any type of training and if data scientists feel they have sufficient support. Regarding formal training or guidance on effectively using AI, 62% of respondents had not received any, while 38% had taken some form of training, but despite this, 70% felt adequately supported in integrating AI technologies into their work processes.

Finally, participants were asked about their future expectations for this technology. Looking to the future, 73% of respondents expected AI to have a more prominent role in their work, while 24% expected a similar role, and only 3% anticipated a less prominent role. Additionally, when asked about specific features or capabilities desired in future AI technologies, 23% expressed interest in more advanced features such as access to real-time data and the creation of a general AI, whereas 77% did not indicate any particular needs.

4.2 Diagnosing

As the first step of this methodology, the diagnosing phase aims to gain a comprehensive understanding of the problem: the effectiveness of ChatGPT in solving a Data Science problem.

For this comparison, a clustering problem will be used, since this type of analysis is a critical task in data science that has the main goal of separating a dataset into disjoint sets of data points with similar characteristics (Tiwari et al., 2023).

The dataset for this clustering analysis was chosen from Kaggle, a platform for data scientists and machine learning enthusiasts, that from many other things like competitions, notebooks and courses, also has available free datasets that can be used by anyone.

To help ChatGPT understand the data, the search was conducted by trying to find a dataset with high usability adapted for clustering analysis. In this way, the chosen dataset was "Customer Personality Analysis", which is composed of 18 attributes and 2240 records.

In order to evaluate and compare the performance of each resolution, besides the qualitative analysis of the Data Scientists, it will also be used the Silhouette Score, a method that shows how close is an object to its cluster in comparison to the remaining clusters. The score range goes from -1 to +1 and the higher the score, the closer is the object to its cluster (Januzaj et al., 2023).

4.3 Action Planning

After Diagnosing, the next step is Action Planning, which is crucial to define the phases to solve the Data Science problem. This includes specifying the steps each will take, ensuring the use of the same tools and dataset, and defining how their performance will be evaluated. By establishing a clear and comparable framework, this phase ensures that the later execution and evaluation are aligned with the research objectives (Santos et al., 2013).

Besides the metrics mentioned before, both data scientist and ChatGPT will use CRISP-DM (CRoss Industry Standard Process for Data Mining) process in order to achieve a homogenous output that will be easier to compare. The CRISP-DM process is a structured framework for data mining projects, consisting of six phases (Schröer et al., 2021).

The first phase, Business Understanding, focuses on comprehending the project objectives and requirements from a business perspective, which involves defining the problem, identifying goals, and determining Business goals to ensure that the data mining results are aligned with the business needs. The second phase, Data Understanding, includes collecting and exploring data to become familiar with it and includes gathering initial data, describing it, exploring it, and verifying its quality. The third phase, Data Preparation, involves constructing the final dataset from the initial raw data (Andrade & Pinho, 2022). This includes data cleaning, transformation, and integration. Tasks may involve selecting relevant data, handling missing values, and creating new variables to prepare the dataset for modeling. The fourth phase, Modeling, consists of the selection and application of different models to the prepared dataset by choosing appropriate algorithms and building models to optimize performance. The fifth phase, Evaluation, assesses the models to ensure they meet the business objectives, and performance criteria and involves reviewing the modeling results, comparing model performance, and validating the models against the success criteria to ensure the model is both accurate and actionable. The final phase, Deployment, implements the models in a real-world environment. This may include deploying the model to a production system, generating reports, or integrating the findings into business processes (Peker & Kart, 2023).

Table 13 - CRISP-DM Description (Adapted from Schröer et al., 2021)

Phase	Description
Business Understanding	The business situation should be assessed to gain an overview of the available and required resources. Determining the data mining goal is a crucial aspect of this phase. First, the type of data mining (e.g., classification) should be explained, along with the success criteria (such as precision).
Data Understanding	Collecting data from various sources, exploring and describing it, and checking its quality are essential tasks in this phase. To make it more concrete, the user guide suggests performing the data description task using statistical analysis and determining attributes and their relationships.
Data Preparation	Data selection should be conducted by defining inclusion and exclusion criteria. Poor data quality can be addressed through data cleaning. Depending on the model chosen in the first phase, derived attributes may need to be constructed.
Modeling	The data modeling phase involves selecting the modeling technique, building the test case, and developing the model. Any data mining technique can be utilized, depending on the business problem and the data. Explaining the rationale behind the choice is crucial. When building the model, specific parameters need to be set. For model assessment, it is important to evaluate the model against established criteria and select the best-performing ones.
Evaluation	In this phase, the results are checked against the defined business objectives and interpreted to define further actions. The process should be reviewed to identify areas for improvement.
Deployment	The deployment phase is generally described in the user guide and may result in a final report or a software component. The user guide outlines that this phase involves planning the deployment, as well as its monitoring and maintenance.

4.3.1 Action Planning: Data Scientist

While I am not a professional data scientist, for the purposes of this study, I will adopt the role of a data scientist, which will allow a structured and methodical comparison, utilizing standard data science practices and tools. Additionally, I will document each step of the process, in order to provide a clear and transparent account of the methods and decisions made. This documentation will serve as a basis for comparison with the approaches taken by ChatGPT.

The first step involves understanding and exploring the data. The "Customer Personality Analysis" dataset will be examined to gain an understanding of its structure and contents.

Initial exploratory data analysis (EDA) will be performed to identify patterns, anomalies, and insights. This includes generating summary statistics, distribution plots, and correlation matrices to visualize and understand the relationships within the data (Otero-Escobar & Velasco-Ramírez, 2023).

Following data understanding, the next step is data preparation, which will be handled using appropriate techniques, depending on the nature and extent of the missing data. The data will then be normalized and scaled to ensure that all features contribute equally to the clustering algorithm. Feature engineering may also be undertaken to create new features or transform existing ones to improve clustering performance (Fan et al., 2021). Techniques such as dimensionality reduction as Principal Component Analysis, which aims to condense a large number of variables at the same time that retain most of the information (Bharadiya, 2023), will be considered if necessary.

During the modeling phase, appropriate clustering algorithms will be chosen based on the characteristics of the data and the goals of the analysis. Once the algorithms are selected, the clustering implementation phase begins. The selected clustering algorithm(s) will be applied to the preprocessed dataset, training the models and generating clustering results (Andrade & Pinho, 2022). The results will be evaluated using predefined evaluation metrics, including the Silhouette Score to assess the quality of the clusters.

Detailed documentation of each step is essential. This includes documenting data preprocessing steps and the logic behind them, selected algorithms, clustering results, and evaluation metrics. Observations and insights gained during the process will also be recorded. Visualizations will be created to understand and present the clustering results.

In terms of tools and technologies, Python will be used as the primary programming language, with various libraries and frameworks supporting data manipulation and analysis, exploratory data analysis, clustering algorithms, dimensionality reduction, and evaluation metrics. The entire process will be documented using Google Colab to consolidate code, outputs, and visualizations in one place.

By following this structured approach, the data scientist’s methodology will be transparent, reproducible, and comparable to ChatGPT's performance in solving the same data science problem.

4.3.2 Action Planning: ChatGPT

In this phase, ChatGPT will assume the role of a data scientist to systematically address the data science problem using the specified prompt. The primary goal is to achieve the best silhouette score for evaluating the model's performance. Additionally, ChatGPT will use t-SNE for cluster visualization.

The process will be interactive, where ChatGPT will provide the code step-by-step, and I will execute the code and provide the results for further guidance in order to have the least possible human intervention. ChatGPT-4o will be used to leverage its advanced capabilities for this task, which is the most recent version which was launched on May 13, 2024, has numerous improvements when compared with the older versions with extended capabilities in text, image recognition, vision and audio (Pang et al., 2024).

ChatGPT will approach the problem with a structured methodology, similar to that of a human data scientist, but with the freedom to use any techniques it deems necessary to optimize the clustering results. The specific techniques and methods ChatGPT uses will be determined dynamically based on the dataset and the problem requirements. The steps will be as follows:

Initially, ChatGPT will focus on understanding and exploring the data. I will provide ChatGPT with the dataset, the description of the features, and the code used to load the data into Google Colab, ensuring that any readability issues due to the dataset being tab-limited instead of comma-limited are resolved. This step ensures that both the data scientist and ChatGPT start from the same point. Following this, ChatGPT will guide the generation of summary statistics and data visualizations to understand the relationships within the data.

Next, in the data preparation phase, ChatGPT will handle data preprocessing. This involves cleaning the data, dealing with missing values, normalizing and scaling features, and possibly performing feature engineering to enhance the dataset's quality for clustering. During the modeling phase, ChatGPT will select and apply appropriate clustering algorithms. The choice of algorithms and the rationale behind them will be documented, after which ChatGPT will build the clustering models and optimize their performance.

In the evaluation phase, the clustering models will be assessed using the silhouette score. ChatGPT might also use other evaluation metrics if deemed necessary to ensure a comprehensive assessment of the model's performance. To visualize the clusters, ChatGPT will provide code to use t-SNE, a technique for dimensionality reduction that can generate 2-dimensional visualizations by identifying the relative locations of all mapped points (Cieslak et al., 2020). The t-SNE plots will help interpret the clustering results, allowing for an assessment of the quality and characteristics of the clusters.

Throughout the process, ChatGPT will ensure detailed documentation of each step, including preprocessing, algorithm selection, model training, evaluation, and visualization. Observations and insights will be recorded for a comprehensive overview. Reflection and iteration will be integral, with ChatGPT suggesting refinements based on the results to ensure continuous improvement, python will be used as the primary programming language, and the process will be documented in Google Colab.

4.4 Action Execution

The action execution phase involves implementing the plans developed in the previous sections (Santos et al., 2013), where both the data scientist and ChatGPT will solve the data science problem using the specified methodology. This phase includes data preparation, model training, evaluation, and visualization.

4.4.1 Action Execution: Data Scientist

As the designated data scientist, I systematically approached the clustering analysis, ensuring each step was meticulously documented to maintain transparency and reproducibility. The key steps taken during this execution phase are described below:

The process began with loading the "Customer Personality Analysis" dataset into Google Colab. The dataset was imported using the Pandas library and since the dataset was tab-delimited and Google Colab was treating it as comma-delimited initial exploratory data analysis (EDA) was performed to gain a comprehensive understanding of the dataset's structure and contents. This involved generating summary statistics, visualizing distributions, and examining correlations between features. Such preliminary analyses are crucial for identifying patterns, anomalies, and potential areas that require further preprocessing.

```
# Load csv file into a dataframe
data_path = "/content/drive/MyDrive/thesis/marketing_campaign.csv"

# Since colab assumed that the file is comma delimited the following correction was applied
data = pd.read_csv(data_path, delimiter='\t') #tab delimiter
```

Figure 8 - Correction of tab delimiter

Next, I created a copy of the original dataset to preserve the raw data for reference. An initial inspection of the dataset was conducted to check for missing values and to understand the data types of each column. As observed in the figure below, there are 24 missing values for Income, and the data type for 'Dt_Customer' is set to object. These problems will need to be addressed.

```
#Looking at the missing values and data types
print("\nData types and missing value counts:")
print(data.info())
```

```
Data types and missing value counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
None
```

Figure 9 - Missing values and data types

Additionally, the date format for the 'Dt_Customer' column was corrected to ensure proper datetime handling in future analyses.

```
# Date format for Dt_Customer needs to be corrected
data['Dt_Customer'] = pd.to_datetime(data['Dt_Customer'], format='%d-%m-%Y')
```

Figure 10 - Correction of 'Dt_Customer' date format

Following this, I conducted an exploratory data analysis (EDA) to understand the dataset's structure and characteristics. This involved generating descriptive statistics and visualizations for the numeric features. As can be observed below, some features, as Income, have outliers that should be treated properly to have a positive impact in the modeling phase (Schindler et al., 2023).

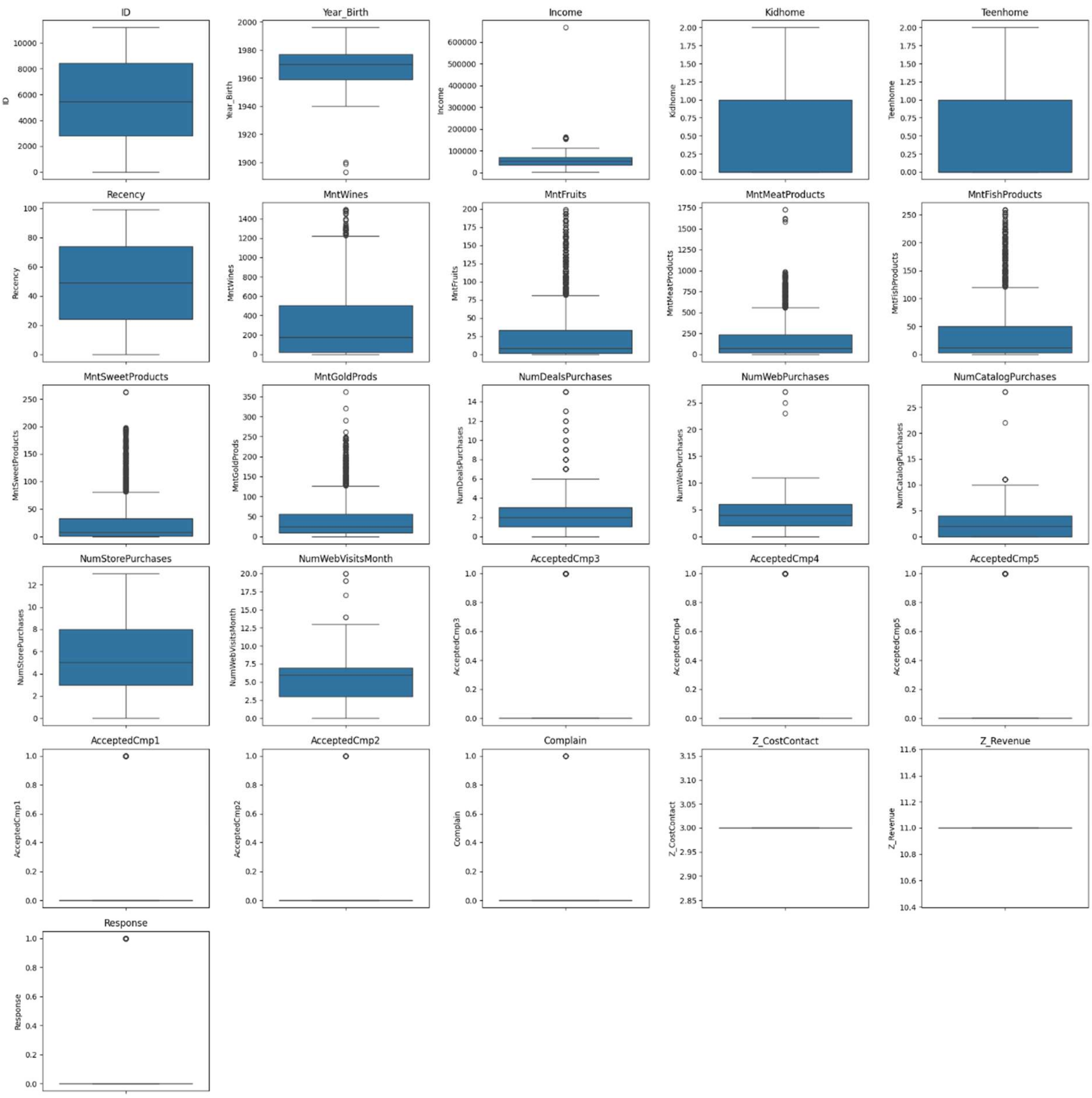

Figure 11 - Box plots for numeric features

Moving to data preparation, missing values were handled for 'Income', by being replaced with the median. Furthermore, duplicate records were addressed but there were no duplicates in this dataset.

```python
# Handle missing values for Income using median
data['Income'].fillna(data['Income'].median(), inplace=True)
```

Figure 12 - Handling missing values

```python
# Check for duplicates
duplicate_rows = data[data.duplicated()]
print(f"Number of duplicate rows: {duplicate_rows.shape[0]}")
```

```
Number of duplicate rows: 0
```

Figure 13 - Looking for duplicates

After this, outliers were handled. Outliers are observations that diverge from the norms in a specific content (Schindler et al., 2023) and need to be addressed to improve the model's performance (Sewwandi et al., 2024). In this case, by observing the box plots, various outliers were identified and removed manually as demonstrated below. Once the outliers were removed, 91.47% of the data was kept.

```python
# Manually filtering outliers for all specified features
filter_conditions = (
    (cleaned_data['Income'] <= 200000) &
    (cleaned_data['MntWines'] <= 1500) &
    (cleaned_data['MntGoldProds'] <= 300) &
    (cleaned_data['NumWebVisitsMonth'] <= 20) &
    (cleaned_data['Year_Birth'] >= 1930) &
    (cleaned_data['MntMeatProducts'] <= 1000) &
    (cleaned_data['NumWebPurchases'] <= 15) &
    (cleaned_data['MntSweetProducts'] <= 100)
)
```

Figure 14 - Manually filtering outliers

Moving forward, I went to feature engineering, which is the process of producing new variables from the ones that already exist, in order to obtain more insights from them (Chicco et al., 2022). In this case, and by looking at the description of the variables, 'Spent' was created by aggregating the spent amount on all product categories. For 'Marital_Status', to simplify the interpretability of the feature, the status Married and Together were replaced by Partner, and the remaining ones were replaced by Alone, which originated the new feature 'Living_With'. The feature 'Children' was created as a result of the aggregation of 'Kidhome' and 'Teenhome'. By subtracting the 'Year_Birth' from the current date, the feature 'Age' was created, 'Customer_Duration' was created from 'Dt_Customer' in order to understand the duration of customer relationships in years, and finally, 'AcceptedAnyCampaign' came as an aggregation of all 'AcceptedCmp' features.

```python
# Create a copy of the filtered data before feature engineering
feature_data = filtered_data.copy()

#Total amount spent
feature_data["Spent"] = (
    feature_data["MntWines"] +
    feature_data["MntFruits"] +
    feature_data["MntMeatProducts"] +
    feature_data["MntFishProducts"] +
    feature_data["MntSweetProducts"] +
    feature_data["MntGoldProds"]
)

#Create a feature to easier understand marital status
feature_data["Living_With"] = feature_data["Marital_Status"].replace({
    "Married": "Partner",
    "Together": "Partner",
    "Absurd": "Alone",
    "Widow": "Alone",
    "YOLO": "Alone",
    "Divorced": "Alone",
    "Single": "Alone",
})

#See if the customer has children
feature_data["Children"] = feature_data["Kidhome"] + feature_data["Teenhome"]

#Understand the age of the customer based on Year_Birth
feature_data["Age"] = datetime.now().year - feature_data["Year_Birth"]

# Calculate the duration of customer relationships in years
feature_data['Dt_Customer'] = pd.to_datetime(feature_data['Dt_Customer'])
current_date = pd.to_datetime('today')
feature_data['Customer_Duration'] = (current_date - feature_data['Dt_Customer']).dt.days / 365

# Creating a feature to see if the customer accepted any campaign 'AcceptedAnyCampaign'
campaign_columns = ['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5']
feature_data['AcceptedAnyCampaign'] = feature_data[campaign_columns].sum(axis=1).apply(lambda x: 1 if x > 0 else 0)
```

Figure 15 - Feature Engineering

After the creation of new features, some unnecessary columns were removed, such as “ID” since it is not relevant to the clustering analysis, “Z_CostContract” and “Z_Revenue” that only have one value, and “Year_Birth”, “Dt_Customer”, “Marital_Status”, 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4' and 'AcceptedCmp5' since I created previously other features that better represent that data.

```python
# Remove unnecessary columns
columns_to_drop = ["ID", "Year_Birth", "Z_CostContact", "Z_Revenue", "Dt_Customer", "Marital_Status"] + campaign_columns
feature_data.drop(columns=columns_to_drop, inplace=True)
```

Figure 16 - Removing Unnecessary Columns

Moving forward, One-Hot encoding was applied to categorical features, where each different state value was extended to a new column, which originated a feature matrix where all values have the same distance to each other (Yu et al., 2022).

```python
# One-hot encoding for categorical variables
categorical_features = ['Education', 'Living_With']
feature_data = pd.get_dummies(feature_data, columns=categorical_features, drop_first=True)
```

Figure 17 - One-hot encoding for categorical features

For numeric features, StandardScaler was used to recalibrate the mean of each feature and scale it to its unit variance, so features with a higher range of values don't have a disproportional impact when modeling (Kedam et al., 2024).

With the features scaled, a correlation matrix was shown so I could observe the correlation between variables. Since some of the features were highly correlated with "Spent"; "MntWines", "MntMeatProducts" and "NumCatalogPurchases" were removed.

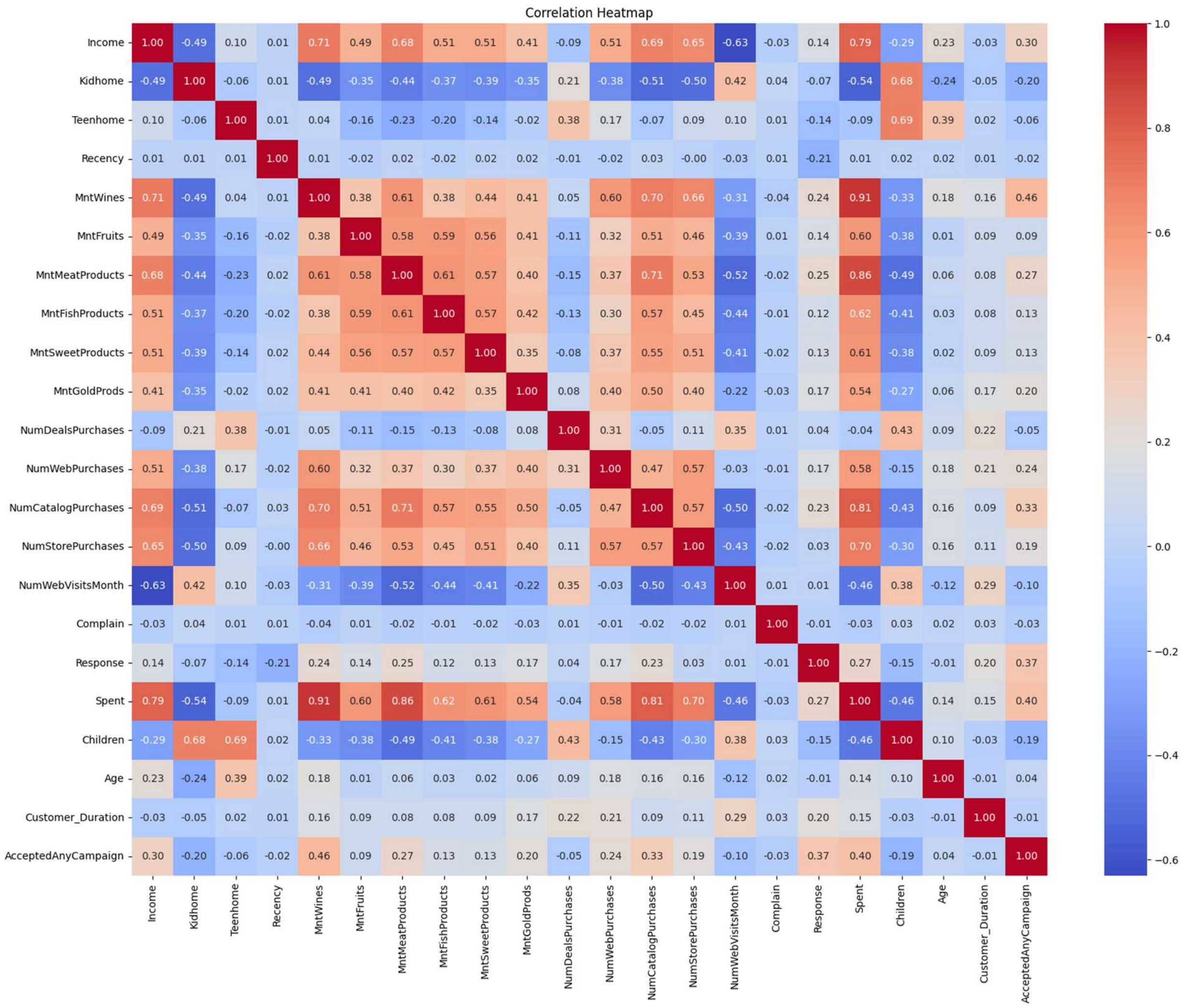

Figure 18 - Correlation Matrix

In order to proceed to dimensionality reduction, Principal Component Analysis (PCA) was used to condense the huge set of variables into a smaller number while keeping the majority of its information (Bharadiya, 2023). Five principal components were chosen since they capture the most significant patterns in the data. Moving to the modeling phase, I started by using the Elbow Method approach to discover the inflection point of the curve with the assistance of Within Cluster Sum of Squares (WCSS), which is a common technique in identifying the optimal number of clusters (Prastyabudi et al., 2024). As can be observed, the best number of clusters would be somewhere between 3 and 4 as we can see below.

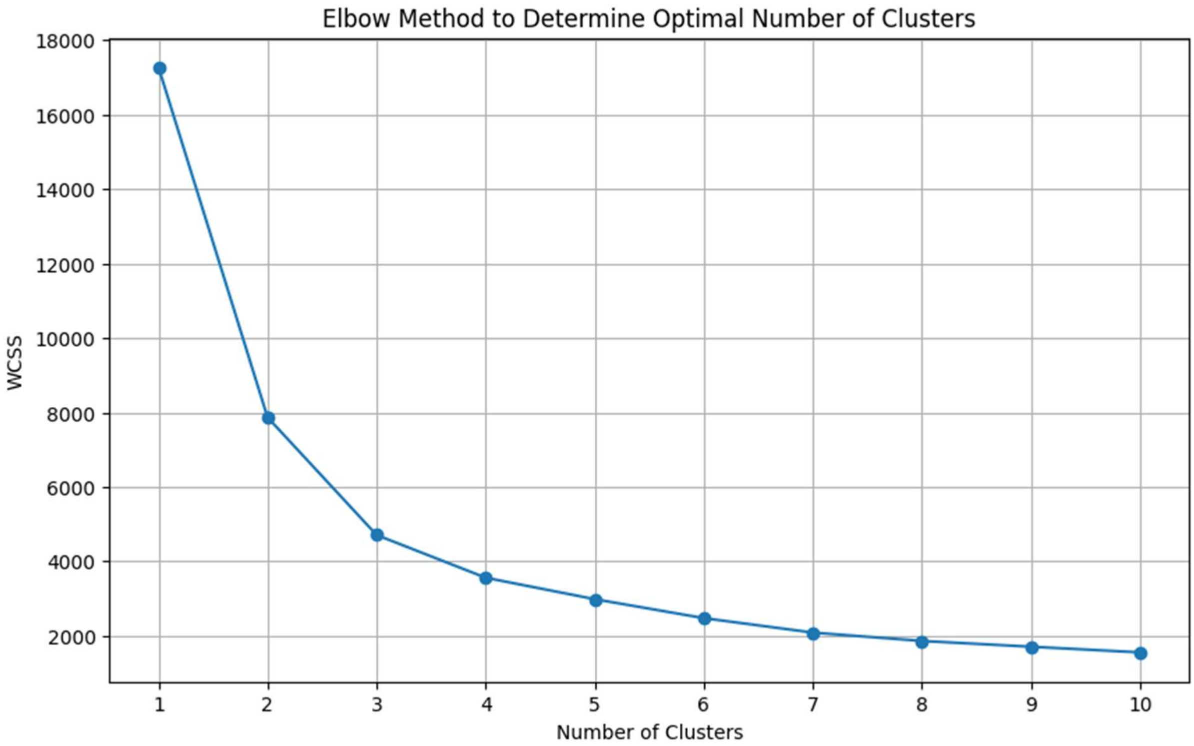

Figure 19 - Elbow Method

To perform the modeling, two algorithms were chosen, K-Means and Agglomerative Clustering. K-Means clustering is one of the simplest and most widely utilized techniques in unsupervised machine learning. The algorithm begins by randomly selecting an initial number of clusters and their corresponding centroid values. It then iteratively reduces the total distance between each data point and its respective cluster centroid, thereby effectively grouping the data points into clusters (Prastyabudi et al., 2024).

Agglomerative clustering is a widely recognized clustering method recognized for its ability to generate an informative hierarchical structure of clusters. The process begins with a multitude of small, initial clusters, each containing a single data point. The algorithm then iteratively merges the pair of clusters that exhibit the highest affinity based on a predefined criterion, such as minimum distance or maximum similarity. This merging process continues until a specified stopping condition is met, which could be a predefined number of clusters or a threshold distance. The hierarchical nature of agglomerative clustering provides valuable insights into the grouping relationships within the data, making it a powerful tool for exploratory data analysis and pattern recognition (Cai et al., 2020).

To compare the performance of K-Means and Agglomerative Clustering, I used the Silhouette Score which measures how close an object is to its cluster compared to other clusters. The score ranges from -1 to +1, with higher values indicating that the object is well-matched to its cluster and poorly matched to neighboring clusters. An average Silhouette score helps determine the optimal number of clusters. This method uses Euclidean distance to calculate how well each object aligns with its cluster (Januzaj et al., 2023).

```
[ ] # Perform KMeans with 3 clusters
    kmeans_3 = KMeans(n_clusters=3, random_state=42)
    kmeans_labels_3 = kmeans_3.fit_predict(pca_results)

    # Perform Agglomerative Clustering with 3 clusters
    agglo_3 = AgglomerativeClustering(n_clusters=3)
    agglo_labels_3 = agglo_3.fit_predict(pca_results)

    # Calculate silhouette scores
    silhouette_kmeans_3 = silhouette_score(pca_results, kmeans_labels_3)
    silhouette_agglo_3 = silhouette_score(pca_results, agglo_labels_3)

    print(f'Silhouette Score for KMeans with 3 clusters: {silhouette_kmeans_3}')
    print(f'Silhouette Score for Agglomerative Clustering with 3 clusters: {silhouette_agglo_3}')

    Silhouette Score for KMeans with 3 clusters: 0.28717749640234475
    Silhouette Score for Agglomerative Clustering with 3 clusters: 0.26188135350905334
```

```
[ ] # Perform KMeans with 4 clusters
    kmeans_4 = KMeans(n_clusters=4, random_state=42)
    kmeans_labels_4 = kmeans_4.fit_predict(pca_results)

    # Perform Agglomerative Clustering with 4 clusters
    agglo_4 = AgglomerativeClustering(n_clusters=4)
    agglo_labels_4 = agglo_4.fit_predict(pca_results)

    # Calculate silhouette scores
    silhouette_kmeans_4 = silhouette_score(pca_results, kmeans_labels_4)
    silhouette_agglo_4 = silhouette_score(pca_results, agglo_labels_4)

    print(f'Silhouette Score for KMeans with 4 clusters: {silhouette_kmeans_4}')
    print(f'Silhouette Score for Agglomerative Clustering with 4 clusters: {silhouette_agglo_4}')

    Silhouette Score for KMeans with 4 clusters: 0.24959107744754477
    Silhouette Score for Agglomerative Clustering with 4 clusters: 0.22572315775375168
```

Figure 20 - Silhouette Score Comparison

Since the higher Silhouette Score was reached by using K-Means, that was chosen as the optimal clustering method, achieving a score of 0.29~.

For the visualization of clusters, I used t-SNE (t-Distributed Stochastic Neighbor Embedding). Like other dimensionality reduction techniques, t-SNE created 2-dimensional visualizations that revealed the relationships between samples. It ensured that similar samples were mapped close together, while dissimilar samples were placed further apart. Due to its nonlinear nature and its ability to balance local and global relationships, t-SNE often produced more visually appealing clusters compared to other methods. This made it particularly useful for visualizing high-dimensional datasets, such as transcriptomic data (Cieslak et al., 2020).

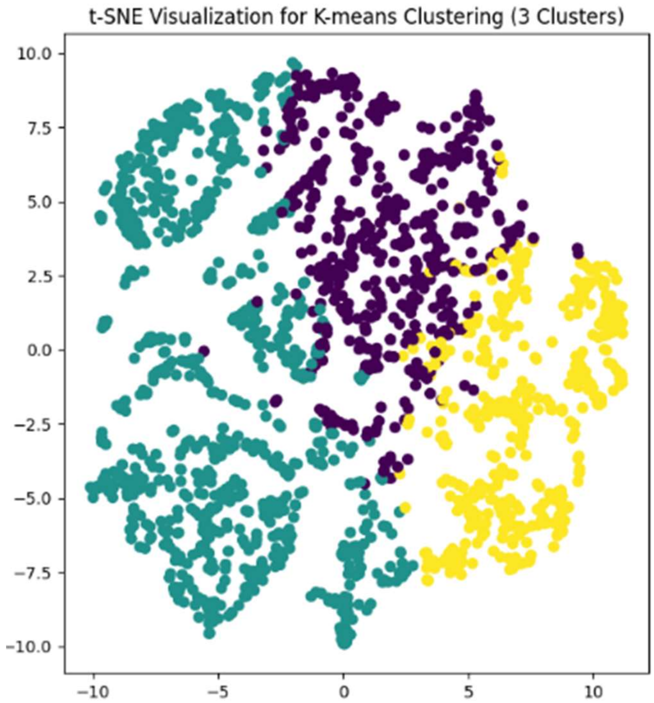

Figure 21 - t-SNE for cluster visualization

To finish, the analysis of the clusters was performed on the original data. The radar chart illustrated in Figure 18 provided a comprehensive visual representation of the profiles of the three distinct customer clusters based on standardized feature values. Each axis of the radar chart represented a different attribute, including demographic details, purchasing behaviors, and promotional engagement metrics. The distance from the center of the chart indicated the mean standardized value of each attribute for the respective clusters, facilitating a comparative analysis of customer characteristics.

Cluster 0 contained customers who had the highest average number of teenagers in their households. This cluster exhibited moderate values across most purchasing and engagement metrics. Customers in this group demonstrated moderate spending on wines, fruits, meats, fish, sweets, and gold products. Their engagement with promotional campaigns was moderate, with the highest acceptance rates in campaign 4. The demographic profile of Cluster 0 suggested a middle-aged customer base with family-oriented households.

Cluster 1 emerged as the segment with the highest average number of children in their households. Customers in this cluster exhibited the lowest values across nearly all attributes, including income, spending on various products, and engagement with promotional campaigns. They displayed minimal spending on wines, fruits, meats, fish, sweets, and gold products. Furthermore, this cluster showed the lowest acceptance rates for promotional campaigns, indicating a lower responsiveness to marketing efforts. The demographic profile of Cluster 1 suggested a younger customer base with moderate household sizes.

Cluster 2 included a middle-aged demographic with the highest average income and the lowest number of teenagers in the household compared to the other clusters. This cluster showed the highest spending across all product categories and exhibited the most significant engagement with promotional campaigns. Despite being the smallest segment in terms of customer count, Cluster 2 had higher interaction levels with the company's offerings, as evidenced by their high standardized values in purchasing behaviors and promotional engagements. The demographic profile of Cluster 2 suggested a high-income customer base with frequent purchasing habits.

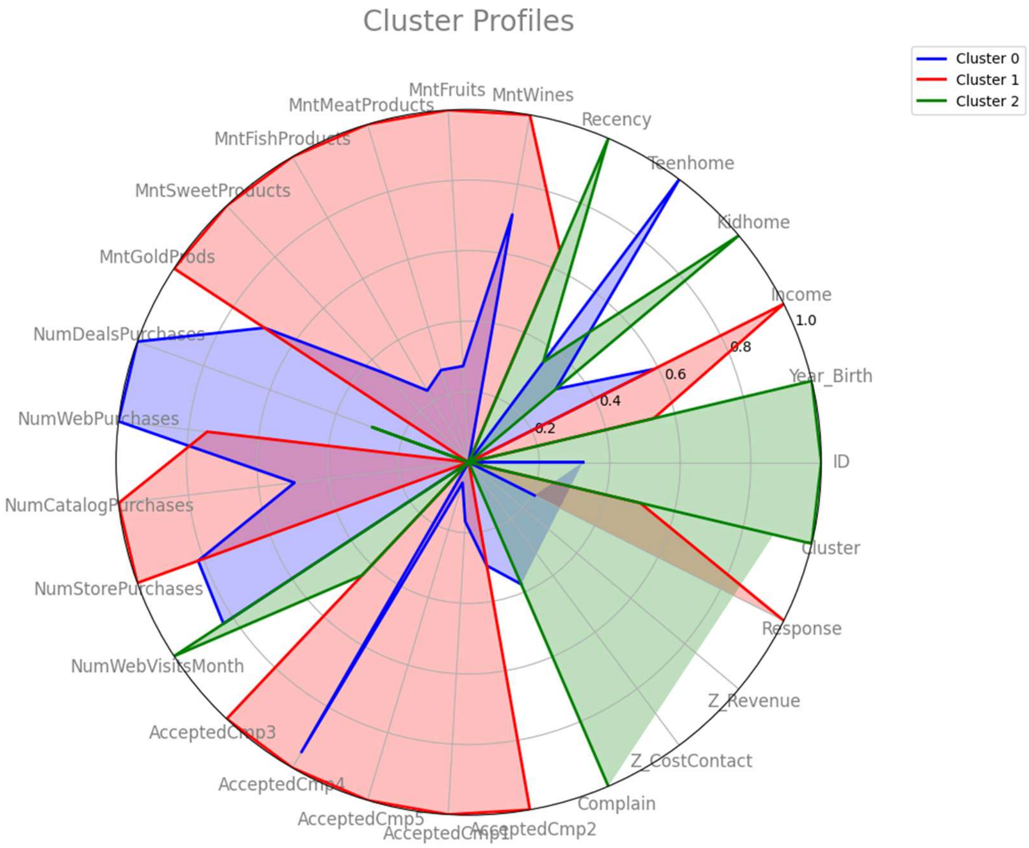

Figure 22 - Radar Chart

4.4.2 Action Execution: ChatGPT

In this section, ChatGPT 4o took on the role of a data scientist, performing a clustering analysis on the "Customer Personality Analysis" dataset from Kaggle. To start this analysis, the following prompt was utilized:

“Act as a Data Scientist in order to perform a cluster analysis by using Python. Use the CRISP-DM methodology for your code. The goal is to have the best silhouette score to evaluate your model's performance. You will also need to use to use t-sne for cluster visualization.

Before starting, I will provide you: -the dataset -the description of the features and -the code used to load the data to Google Colab where I'll be performing your code.

After this, you will give me the code, step by step, and after each step I will show you the result so we can move further."

This prompt assured that both versions of the cluster analysis will be comparable by using the same process of loading the data into Google Colab, the equal CRISP-DM methodology, Silhouette Score for evaluation and t-SNE for visualization. Another important aspect of this prompt is that ChatGPT will have the freedom to process the cluster analysis in a way that he will consider proper for this case.

Afterwards providing all required data and information and loading the data into Google Colab, ChatGPT started by exploring the data with the Data Understanding phase of CRISP-DM.

```
# Display the first few rows of the dataframe
data.head()
```

```
# Display basic information about the dataframe
data.info()
```

```
# Describe the dataframe to get a summary of the statistics for each column
data.describe(include='all')
```

Figure 23 - Code provided from ChatGPT for Data Understanding

After providing the results of his code, ChatGPT found some issues that he decided to address like converting categorical variables and scaling numerical variables, but before that, he decided to check for missing values. As is shown below, there are 24 missing values for "Income".

```
# Checking for missing values
data.isnull().sum()
```

```
ID                      0
Year_Birth              0
Education               0
Marital_Status          0
Income                 24
Kidhome                 0
Teenhome                0
Dt_Customer             0
Recency                 0
MntWines                0
MntFruits               0
MntMeatProducts         0
MntFishProducts         0
MntSweetProducts        0
MntGoldProds            0
NumDealsPurchases       0
NumWebPurchases         0
NumCatalogPurchases     0
NumStorePurchases       0
NumWebVisitsMonth       0
AcceptedCmp3            0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Z_CostContact           0
Z_Revenue               0
Response                0
dtype: int64
```

Figure 24 - Output of Missing Values

In the same answer, ChatGPT gave me the code to identify categorical and numerical columns, handle missing values, encode the categorical ones, and scale the numerical ones. The StandardScaler was used for scaling numerical features to ensure they had a mean of zero and a standard deviation of one. The OneHotEncoder was employed to convert categorical features into a one-hot numeric array, a necessary step for algorithms that cannot handle categorical data directly. ColumnTransformer allowed the application of different preprocessing steps to various subsets of features. The Pipeline facilitated the sequential application of a list of transforms and a final estimator. Lastly, the SimpleImputer handled missing values by imputing them with specific strategies.

ChatGPT's next step involved identifying which columns in the dataset were categorical and which were numerical. This was achieved by explicitly specifying the categorical columns, in this case, 'Education' and 'Marital_Status'. Numerical columns were then identified by excluding the categorical columns and any additional columns that should not be processed numerically, such as 'ID' and 'Dt_Customer'.

ChatGPT provided a preprocessing pipeline for numerical data that included two steps: an imputer and a scaler. The imputer filled any missing values in the numerical columns with the median value of the respective column, ensuring that missing data did not skew the dataset. Following the imputer, the scaler standardized the features by removing the mean and scaling to unit variance. Similarly, ChatGPT constructed a preprocessing pipeline for categorical data which also included two steps: an imputer and a one-hot encoder.

The imputer filled missing values in the categorical columns with the most frequent value (mode) of the respective column, ensuring that missing categorical data was replaced with the most common category. The one-hot encoder then converted the categorical values into a one-hot numeric array.

The numerical and categorical preprocessing pipelines were then combined using a ColumnTransformer, which applied the numerical preprocessing pipeline to the numerical columns and the categorical preprocessing pipeline to the categorical columns. The ColumnTransformer ensured that each type of data was processed appropriately without manual intervention for each column. The combined preprocessor on the dataset was performed by using the fit_transform method, which fit the preprocessor to the data and then transformed the data accordingly. The resulting preprocessed data was stored in a variable named data_preprocessed.

To facilitate further analysis and ensure ease of use, ChatGPT's code converted the preprocessed data back into a DataFrame. This conversion allowed the use of familiar DataFrame operations and methods. The DataFrame was created using pd.DataFrame(data_preprocessed).

```
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer

# Identifying categorical and numerical columns
categorical_cols = ['Education', 'Marital_Status']
numerical_cols = [col for col in data.columns if col not in categorical_cols + ['ID', 'Dt_Customer']]

# Preprocessing pipeline for numerical data
numerical_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

# Preprocessing pipeline for categorical data
categorical_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

# Combining numerical and categorical pipelines
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_pipeline, numerical_cols),
        ('cat', categorical_pipeline, categorical_cols)
    ])

# Applying transformations to the dataset
data_preprocessed = preprocessor.fit_transform(data)

# Convert the preprocessed data back to a DataFrame for convenience
data_preprocessed_df = pd.DataFrame(data_preprocessed)

# Displaying the first few rows of the preprocessed dataframe
data_preprocessed_df.head()
```

Figure 25 - Code given by ChatGPT for Handling Missing Values, Encoding and Scaling

Finally, ChatGPT suggested displaying the first few rows of the preprocessed DataFrame using the head() method which provided a quick preview of the transformed data, enabling to verify that the preprocessing steps had been applied correctly and that the data was in the expected format.

	0	1	2	3	4	5	6	7	8	9	...	28	29	30	31	32	33	34	35	36	37
0	-0.985345	0.235696	-0.825218	-0.929894	0.307039	0.983781	1.551577	1.679702	2.462147	1.476500	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1	-1.235733	-0.235454	1.032559	0.906934	-0.383664	-0.870479	-0.636301	-0.713225	-0.650449	-0.631503	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
2	-0.317643	0.773999	-0.825218	-0.929894	-0.798086	0.362723	0.570804	-0.177032	1.345274	-0.146905	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	1.268149	-1.022355	1.032559	-0.929894	-0.798086	-0.870479	-0.560857	-0.651187	-0.503974	-0.583043	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
4	1.017761	0.241888	1.032559	-0.929894	1.550305	-0.389085	0.419916	-0.216914	0.155164	-0.001525	...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

5 rows × 38 columns

Figure 26 - Preview of Preprocessed Data

Continuing with the clustering analysis, ChatGPT provided the next steps involving the determination of the optimal number of clusters using the Elbow method.

ChatGPT recommended using the KMeans clustering algorithm and provided the necessary code to implement the Elbow method. The process started by calculating the sum of squared errors (SSE) for a range of cluster numbers, specifically from 1 to 10. For each number of clusters, he fitted the KMeans algorithm to the preprocessed data and recorded the SSE, which measures the variance within the clusters in order to determine the optimal number of clusters, by creating an Elbow curve.

The provided code was executed as follows:

```python
# Determine the optimal number of clusters using the elbow method
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_preprocessed)
    sse.append(kmeans.inertia_)
```

Figure 27 - Determining the optimal number of clusters with KMeans

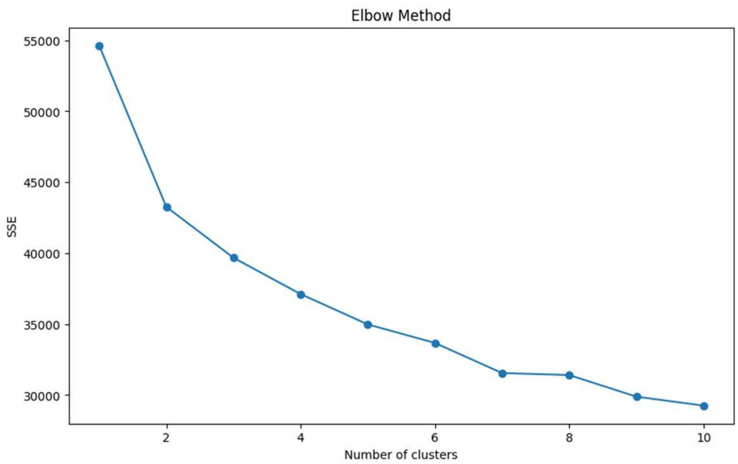

Figure 28 - Elbow Method

After the elbow method was applied, I provided the output to ChatGPT to continue the analysis and take his own conclusions to which he gave me the following answer: "*The elbow plot indicates that the optimal number of clusters is likely around 4 or 5, where the SSE starts to level off. Let's choose 5 clusters and proceed with the clustering.*".

By choosing 5 clusters, ChatGPT gave me the code to perform K-Means with 5 clusters and subsequently, calculate the Silhouette Score for this clustering.

```
# Evaluating the silhouette score
silhouette_avg = silhouette_score(data_preprocessed, cluster_labels)
print(f'Silhouette Score for {n_clusters} clusters: {silhouette_avg}')
```

```
Silhouette Score for 5 clusters: 0.17994524786451188
```

Figure 29 - Silhouette Score

When showing the Silhouette Score to ChatGPT, he gave the following answer: "*The silhouette score for 5 clusters is 0.1799, which is relatively low. This suggests that the clusters might not be well-separated. Let's try to visualize the clusters using t-SNE to better understand the clustering.*".

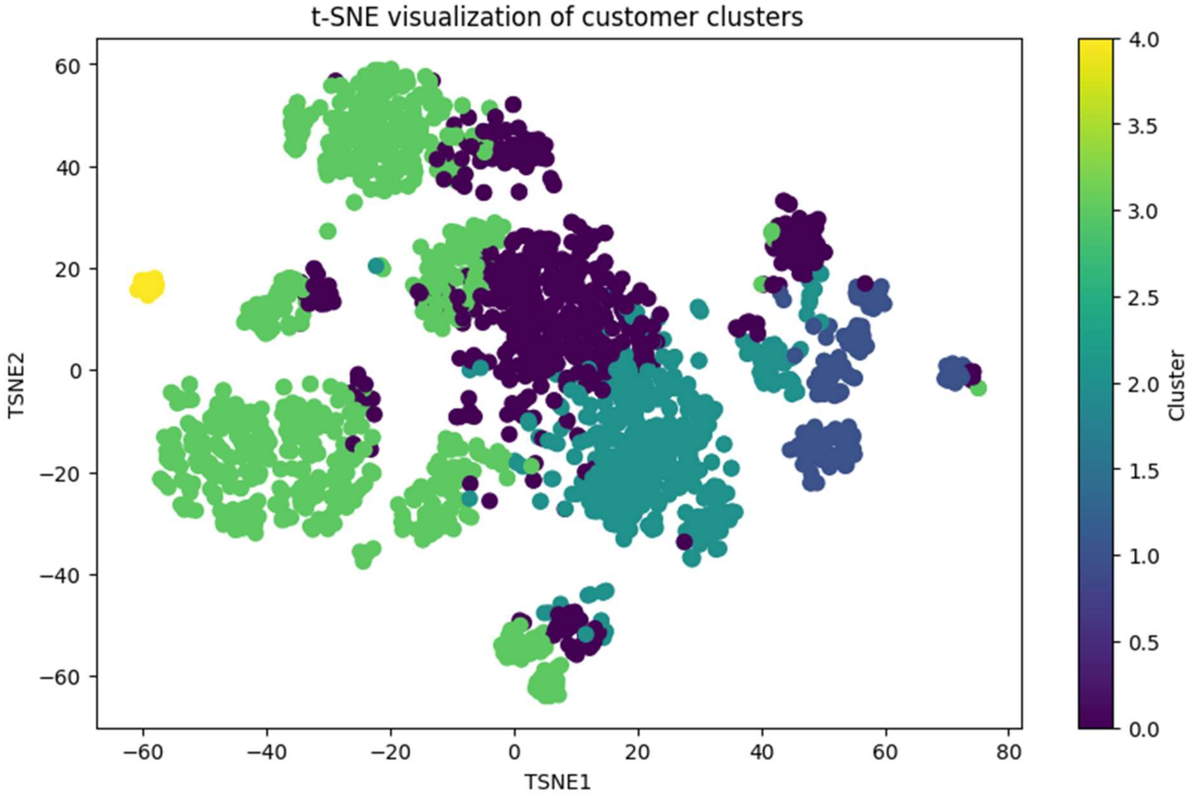

Figure 30 - Cluster Visualization using t-SNE

ChatGPT had access to the t-SNE cluster visualization and gave the following answer: *"The t-SNE visualization shows the distribution of customer clusters. It appears that some clusters are well-defined, while others are more mixed, which corresponds to the relatively low silhouette score.", followed by some suggestions:*

1. **Re-evaluate Cluster Numbers**: We can re-evaluate the number of clusters and try a different number, such as 4 or 6, to see if it improves the silhouette score and the t-SNE visualization.
2. **Feature Engineering**: We can create additional features or transform existing ones to better capture the patterns in the data.
3. **Different Clustering Algorithms**: We can try other clustering algorithms like DBSCAN or hierarchical clustering to see if they provide better results.
4. **Domain Knowledge**: Collaborate with domain experts to interpret the clusters and ensure they make business sense.

Figure 31 - Suggestions given by ChatGPT

ChatGPT also asked if I wanted him to proceed with any of these steps, I reminded him that I was only running his code and that he would make his own decisions, following what he decided to calculate the Silhouette Score for 4 and 6 clusters.

```
# Evaluating the silhouette score for 4 clusters
silhouette_avg_4 = silhouette_score(data_preprocessed, cluster_labels_4)
print(f'Silhouette Score for {n_clusters_4} clusters: {silhouette_avg_4}')
```

```
Silhouette Score for 4 clusters: 0.1719536422272451
```

Figure 32 - Silhouette Score for 4 clusters

Since K-Means with 6 clusters had slightly better Silhouette Score, ChaGPT decided to proceed with Cluster Interpretation in order to analyze them to understand their characteristics and potential business implications.

```
# Evaluating the silhouette score for 6 clusters
silhouette_avg_6 = silhouette_score(data_preprocessed, cluster_labels_6)
print(f'Silhouette Score for {n_clusters_6} clusters: {silhouette_avg_6}')
```

```
Silhouette Score for 6 clusters: 0.18157980900532764
```

Figure 33 - Silhouette Score for 6 clusters

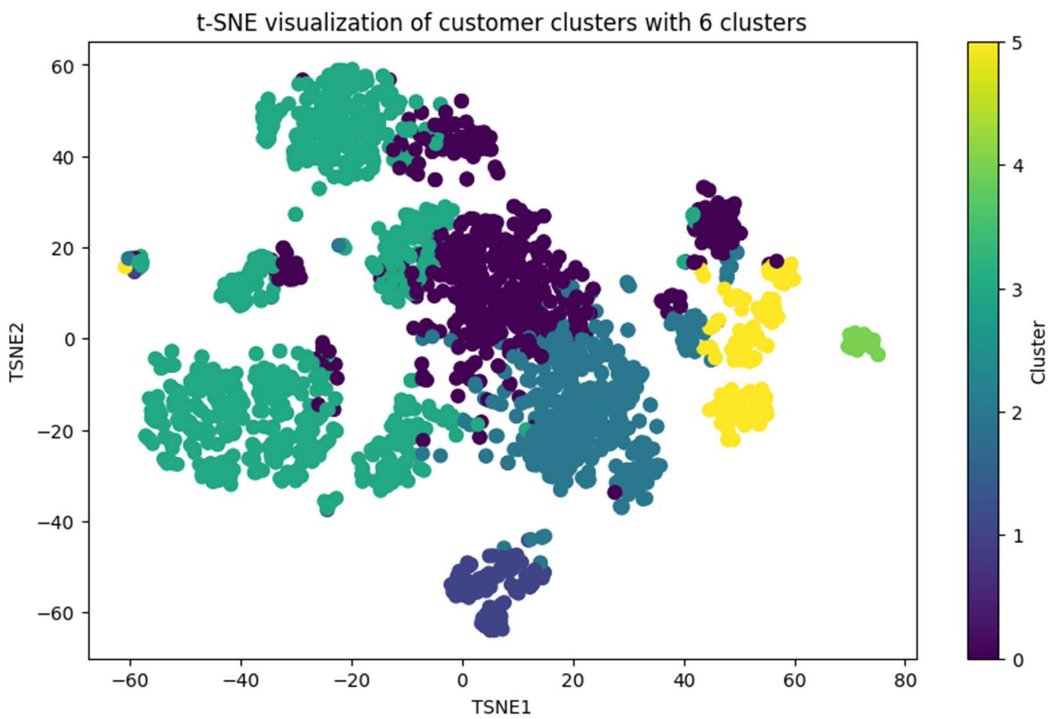

Figure 34 - t-SNE for 6 clusters

ChatGPT gave me the code to calculate the mean values of features for each cluster, but when I tried to run the code, I obtained an error since the dataset contained categorical columns that were not compatible with numerical aggregation functions. After that, ChatGPT tried to select only numerical features and calculate the mean, but when tried to import “ace_tools”, the module was not found.

```
# Selecting only the numerical columns for aggregation
numerical_data = data.select_dtypes(include=['int64', 'float64'])

# Adding cluster labels to the numerical dataframe
numerical_data['Cluster'] = cluster_labels_6

# Calculating the mean values of numerical features for each cluster
cluster_means = numerical_data.groupby('Cluster').mean()

import ace_tools as tools; tools.display_dataframe_to_user(name="Cluster Means", dataframe=cluster_means)
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-40-f370fcb87c66> in <cell line: 10>()
      8 cluster_means = numerical_data.groupby('Cluster').mean()
      9
---> 10 import ace_tools as tools; tools.display_dataframe_to_user(name="Cluster Means", dataframe=cluster_means)

ModuleNotFoundError: No module named 'ace_tools'
```

Figure 35 – ‘ace_tools’ not found

ChatGPT solved this issue by displaying the features while using standard Pandas functions, by using the code presented below.

```python
# Selecting only the numerical columns for aggregation
numerical_data = data.select_dtypes(include=['int64', 'float64'])

# Adding cluster labels to the numerical dataframe
numerical_data['Cluster'] = cluster_labels_6

# Calculating the mean values of numerical features for each cluster
cluster_means = numerical_data.groupby('Cluster').mean()

# Displaying the cluster means
cluster_means
```

Figure 36 - Using Pandas to display features means

Cluster	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
0	5503.327338	1964.088129	57672.125683	0.250000	0.958633	49.172662	441.275180	20.273381	128.107914	27.199640
1	5172.302521	1970.722689	40862.957983	0.605042	0.495798	42.210084	235.008403	11.579832	77.831933	16.941176
2	5721.777262	1968.141531	74143.932084	0.048724	0.232019	50.194896	480.116009	72.366589	429.118329	104.651972
3	5644.822906	1971.589608	34741.718884	0.798515	0.436903	49.678685	39.306469	4.955461	22.634146	7.341463
4	5172.566667	1968.133333	71054.833333	0.066667	0.433333	48.666667	898.666667	22.966667	250.300000	38.733333
5	5632.006211	1969.279503	82294.553459	0.049689	0.105590	47.832298	848.105590	60.341615	494.875776	85.267081

Figure 37 - Snapshot of the output

After observing each cluster's characteristics, ChatGPT made a description of each Cluster:

"The cluster means provide insights into the characteristics of each customer segment. Here are some key observations for each cluster:

1. ***Cluster 0**:*

    - *Generally older customers (average birth year: 1964).*
    - *Relatively high income (average: 57,672).*
    - *Low number of children and teenagers at home.*
    - *High spending on wines and meat products.*
    - *Low response rate to campaigns.*

2. ***Cluster 1**:*

    - *Slightly younger customers (average birth year: 1970).*
    - *Moderate income (average: 40,863).*
    - *Moderate spending on all product categories.*
    - *Higher number of web visits and web purchases.*
    - *Moderate response rate to campaigns.*

3. ***Cluster 2**:*

    - *Similar age to Cluster 0 (average birth year: 1968).*
    - *Higher income (average: 74,143).*
    - *Low number of children and teenagers at home.*
    - *High spending on meat products.*
    - *Low number of web visits and purchases.*
    - *Low response rate to campaigns.*

4. ***Cluster 3**:*

    - *Younger customers (average birth year: 1971).*
    - *Lower income (average: 34,741).*
    - *Low spending on most product categories.*
    - *Higher number of web visits.*
    - *Low response rate to campaigns.*

5. ***Cluster 4**:*

    - *Similar age to Cluster 1 (average birth year: 1968).*
    - *Higher income (average: 71,054).*
    - *High spending on meat products.*
    - *Higher response rate to campaigns.*

6. ***Cluster 5**:*

    - *Younger customers (average birth year: 1969).*
    - *Highest income (average: 82,294).*
    - *Highest spending on most product categories.*
    - *Moderate response rate to campaigns."*

With the description of the clusters concluded, ChatGPT asked me again, If I would like to proceed with any further analysis or actions based on the insights previously described, and I reminded him again that he should take his own decisions.

Following this, he provided a code to visualize the spending patterns across the different clusters with the objective of understanding which products are favored by each cluster.

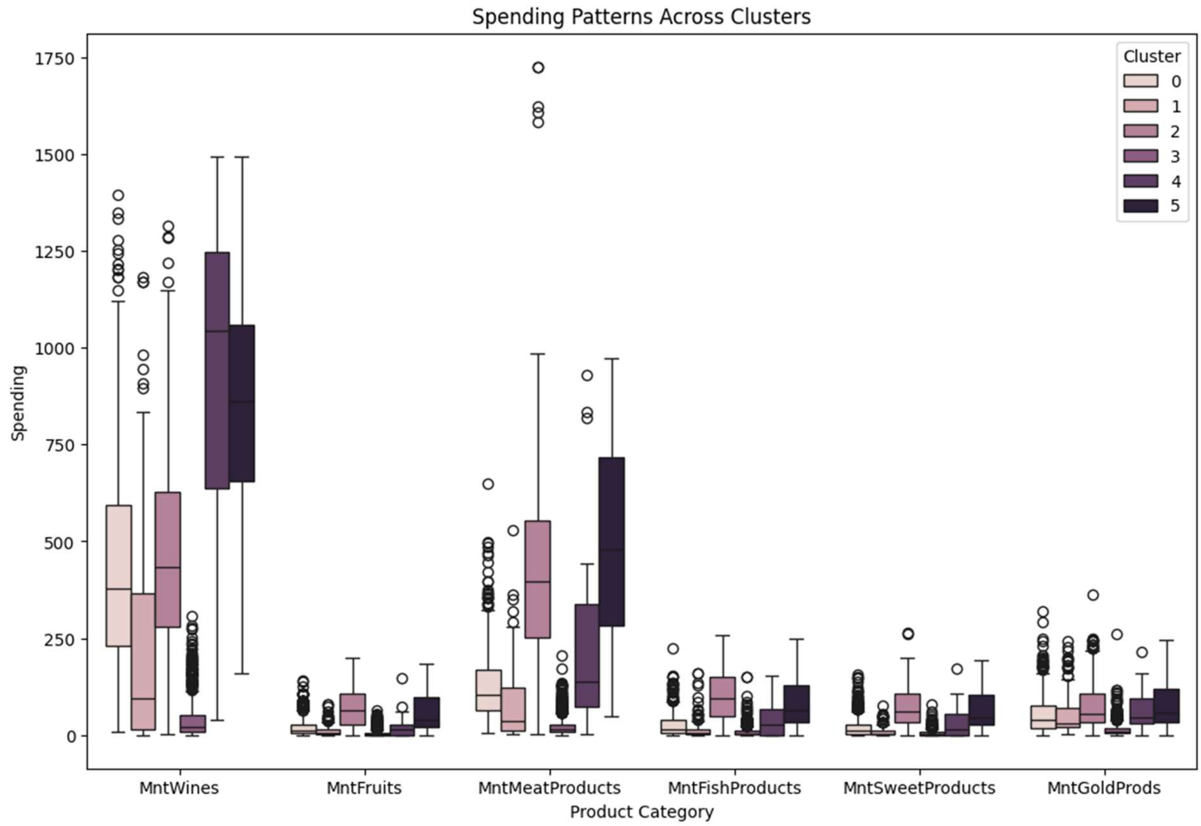

Figure 38 - Box Plot with Spending Patterns

ChatGPT proceeded to analyze the box plot and answered the following:

"The boxplot visualization provides clear insights into the spending patterns across different clusters for various product categories:

MntWines*:*

- *Cluster 5 has the highest spending on wines, followed by Cluster 2.*
- *Clusters 1, 3, and 4 have moderate spending, while Cluster 0 has the lowest spending.*

MntFruits*:*

- *Cluster 0 has the highest spending on fruits, followed by Cluster 1.*
- *Other clusters have relatively low spending on fruits.*

MntMeatProducts*:*

- *Cluster 5 spends the most on meat products, followed by Cluster 2.*
- *Cluster 4 also has significant spending, while other clusters spend less.*

MntFishProducts:

- *Cluster 5 leads in spending on fish products.*
- *Clusters 2 and 4 also spend moderately on fish products, while others spend less.*

MntSweetProducts:

- *Cluster 0 spends the most on sweet products, with Cluster 4 following.*
- *Other clusters have lower spending on sweets.*

MntGoldProds:

- *Spending on gold products is relatively low across all clusters, with Cluster 5 having slightly higher spending."*

As the final request, ChatGPT gave me a code to analyze campaign responses in order to evaluate which ones were more successful in each cluster.

Cluster	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	Response
0	0.032374	0.0	0.000000	0.136691	0.005396	0.082734
1	0.008403	0.0	1.000000	0.000000	0.000000	0.428571
2	0.083527	0.0	0.030162	0.023202	0.000000	0.146172
3	0.001060	0.0	0.000000	0.007423	0.000000	0.066808
4	0.433333	1.0	0.233333	0.733333	0.566667	0.666667
5	0.465839	0.0	0.149068	0.322981	0.888199	0.565217

Figure 39 - Mean of each accepted campaign by cluster

After analyzing the provided data, ChatGPT gave the following answer:

"The campaign response means provide valuable insights into how different clusters responded to various campaigns:

Cluster 0:

- *Low acceptance rates for all campaigns.*
- *Overall low response rate (0.0827).*

Cluster 1:

- *Low acceptance rates except for Campaign 3.*
- *Moderate overall response rate (0.4286).*

Cluster 2:

- *Moderate acceptance rates for Campaigns 3 and 4.*
- *Low overall response rate (0.1462).*

Cluster 3:

- *Very low acceptance rates for all campaigns.*
- *Lowest overall response rate (0.0668).*

Cluster 4:

- *High acceptance rates for Campaign 4 and 5.*
- *Highest overall response rate (0.6667).*

Cluster 5:

- *High acceptance rates for all campaigns, particularly Campaign 5.*
- *High overall response rate (0.5652).”*

4.5 EVALUATION

The evaluation phase in the action research methodology is a critical step where the effectiveness, impact, and outcomes of the implemented actions are systematically assessed. During the evaluation phase, data collected from the action steps are analyzed to assess the success and areas for improvement which includes both quantitative and qualitative data, providing a comprehensive understanding of the results (Santos et al., 2013).

In the context of this research, the evaluation phase includes qualitative interviews with experienced data scientists to provide insights into the clustering analyses performed by a traditional data scientist and ChatGPT. With expert feedback and detailed comparisons, the evaluation aims to draw conclusions about the methodologies' effectiveness and suggest potential enhancements and this phase is also essential for validating the research outcomes and informing future actions and decisions.

4.5.1 Interviews Description

Both cluster analyses were presented to the Data Science experts, with a focus on the most evident differences between the two approaches. After this, four qualitative questions were presented to the interviewees to get a proper evaluation and gather insights. It must be mentioned that each data scientist was interviewed separately.

Table 14 - Description of the Questions

Scope	Question
Preprocessing	How do you evaluate the preprocessing steps taken in both cluster analyses? Were there any notable strengths or weaknesses in the approaches?
Modeling	How do you compare the clustering algorithms and evaluation metrics used in both analyses? Was the choice of K-Means appropriate, and were there other algorithms or metrics that should have been considered?
Profiling	What are your thoughts on the t-SNE visualizations and the overall interpretability of the cluster results from both analyses? How do these visualizations help in understanding the clustering outcomes?
Conclusion	In your opinion, could ChatGPT replace a data scientist in performing clustering analyses, given the strengths and weaknesses you've discussed?

Table 15 – Description of the Interviewees

Interviewee Designation	Industry/Field of Work	Job Title/Position	Years of Experience as a Data Scientist
Interviewee 1	Consulting	Data Scientist - Associate Consultant	4
Interviewee 2	Healthcare	Data Scientist	6
Interviewee 3	Finance	Data Scientist - Credit Risk Modeller	4

4.5.2 Interviews Discussion

In this section, the findings from interviews conducted with three experienced data scientists will be discussed. These interviews aimed to evaluate and compare the effectiveness of two clustering analyses: one conducted by a data scientist and the other by ChatGPT, an AI model. The discussion will cover various aspects such as preprocessing steps, algorithm selection, evaluation metrics, the interpretability of the results and additionally, the potential of ChatGPT to replace human data scientists in performing complex clustering tasks will also be explored.

How do you evaluate the preprocessing steps taken in both cluster analyses? Were there any notable strengths or weaknesses in the approaches?

Interviewee 1 highlighted the attention to detail of the data scientist analysis, particularly emphasizing the detailed exploratory data analysis (EDA) that was conducted. The interviewee noted that this comprehensive EDA provided a deep understanding of the dataset, which is crucial for identifying key patterns and potential issues before proceeding with clustering. This preparatory work helped in crafting a more informed and effective preprocessing strategy. The interviewee also appreciated the creative feature engineering, such as the creation of new features like 'Spent' and 'Customer_Duration', which added significant value to the analysis by offering more dimensions for the clustering algorithm to work with.

Interviewee 2 agreed with the importance of the EDA in the data scientist analysis, adding that the manual handling of outliers, although effective, might not be scalable for larger datasets. The interviewee appreciated the accurate approach but suggested that automated methods for outlier detection and handling could improve scalability and consistency and also praised the use of pipelines in ChatGPT's analysis for preprocessing, as this ensured a structured and reproducible workflow. However, the interviewee noted that ChatGPT’s analysis lacked the depth of initial data exploration seen in the first analysis, which could mean missing out on important data patterns that influence the clustering results.

Interviewee 3 mentioned creative feature engineering in the first analysis. The interviewee pointed out that creating features tailored to the specific context of the data can greatly enhance the clustering outcome. The interviewee also mentioned that ChatGPT’s approach, while efficient and streamlined, could have benefited from a more detailed EDA. The absence of outlier handling and in-depth feature correlation analysis in ChatGPT’s preprocessing was seen as a potential weakness. Incorporating these steps could have provided a more robust foundation for the clustering analysis, potentially improving the quality of the results.

How do you compare the clustering algorithms and evaluation metrics used in both analyses? Was the choice of K-Means appropriate, and were there other algorithms or metrics that should have been considered?

Interviewee 1 noted that the first analysis utilized both K-Means and Agglomerative Clustering, ultimately selecting K-Means based on a higher Silhouette Score. This comparison of algorithms demonstrated a comprehensive evaluation process, ensuring that the chosen method was well-suited to the data. The interviewee pointed out that the Elbow Method was correctly used in both analyses to determine the optimal number of clusters, which is a widely accepted practice in clustering analysis.

Interviewee 2 appreciated the detailed algorithm evaluation in the data scientist analysis, emphasizing that the higher Silhouette Score of 0.29 indicated better cluster separation and cohesion compared to the score in ChatGPT’s analysis. The interviewee suggested that while K-Means is a solid choice, the inclusion of Agglomerative Clustering provided a valuable comparative perspective. The interviewee also proposed that using additional evaluation metrics could have offered a more comprehensive assessment of the clustering performance. The interviewee also mentioned that applying different clustering approaches based on selected groups of features, such as demographics and purchasing behavior, could have provided more detailed insights and potentially improved the clustering outcomes.

Interviewee 3 agreed that the data scientist's higher Silhouette Score demonstrated superior clustering quality. The interviewee commended the exploration of multiple algorithms, which added robustness to the analysis. However, the interviewee noted that relying solely on the Silhouette Score might limit the evaluation's depth. The interviewee suggested incorporating other metrics like the Davies-Bouldin Index and examining cluster stability across different runs to provide a more nuanced understanding of the clustering performance.

What are your thoughts on the t-SNE visualizations and the overall interpretability of the cluster results from both analyses? How do these visualizations help in understanding the clustering outcomes?

Interviewee 1 found the t-SNE visualizations from both analyses to help interpret high-dimensional data in a more comprehensible two-dimensional space. The interviewee noted that the clusters in the data scientist's t-SNE plot appeared more distinct and well-defined, which aligned with its higher Silhouette Score. This clear separation in the visualization made it easier to understand the cluster assignments and identify meaningful customer segments.

Interviewee 2 emphasized the value of t-SNE visualizations in providing a visual representation of the clustering structure. The interviewee observed that the data scientist's clusters were more clearly separated, reflecting better-defined groups. In contrast, ChatGPT's t-SNE visualization showed more overlap and mixed clusters, which corresponded with its lower Silhouette Score. The interviewee suggested that ChatGPT's analysis could benefit from incorporating outlier handling and feature correlation analysis, which could result in clearer and more distinct clusters. The interviewee also mentioned that by looking at the t-SNE output, the yellow cluster has a very small size, and would be a best practice to merge it into another cluster with similar characteristics.

Interviewee 3 agreed with the assessment of the t-SNE visualizations, noting that the data scientist's analysis provided more interpretable and distinct clusters, and also highlighted that these visualizations are crucial for identifying where the clustering algorithm might be struggling, offering insights for further refinement. The interviewee added that while ChatGPT's visualization was still useful, it indicated less distinct cluster separation. The interviewee suggested that more detailed preprocessing steps, such as removing correlated features and handling outliers, could enhance the clarity and separation of clusters in the t-SNE visualization. Additionally, Interviewee 3 mentioned the potential benefits of applying different clustering methods based on specific feature groups to further improve the interpretability and usefulness of the results.

In your opinion, could ChatGPT replace a data scientist in performing clustering analyses, given the strengths and weaknesses you've discussed?

Interviewee 1 acknowledged ChatGPT's efficiency and ability to handle routine tasks, such as preprocessing and basic clustering. The interviewee appreciated how AI can modernize workflows and ensure consistency but noted that ChatGPT lacks the depth of understanding and creativity that human data scientists have, particularly in areas like EDA, feature engineering, and interpreting complex patterns. The interviewee concluded that while ChatGPT can assist and enhance the work of data scientists, it is not yet capable of fully replacing them.

Like Interviewee 1, Interviewee 2 highlighted that ChatGPT performs well in executing standard procedures and automating repetitive tasks and also mentioned AI's ability to quickly preprocess data and apply clustering algorithms is valuable.

However, the interviewee emphasized that human expertise is crucial for making nuanced decisions, such as selecting the most appropriate features, handling outliers, and interpreting results within the specific context of the data. The interviewee concluded that ChatGPT could serve as a powerful tool for data scientists but cannot entirely replace the need for human insight and experience.

Interviewee 3 agreed with the previous interviewees, stating that ChatGPT's strengths lie in its efficiency. The interviewee appreciated the AI's ability to handle datasets and perform initial analyses quickly. However, the interviewee pointed out that ChatGPT's lack of deep contextual understanding and creativity limits its effectiveness in more complex analyses. The interviewee suggested that ChatGPT could augment the capabilities of data scientists, allowing them to focus on more sophisticated aspects of the analysis, but it cannot replace the human element entirely.

5 CONCLUSION

This thesis explored the capabilities and limitations of ChatGPT within data science, specifically focusing on its application in solving clustering analysis problems—a critical task in this field. A mixed methodology was used, combining quantitative insights from a comprehensive survey with qualitative analysis through the Action-Research methodology. The qualitative analysis was conducted through interviews that served as the evaluation phase, comparing the clustering analyses performed by the data scientist and ChatGPT.

The survey provided valuable insights into how data scientists currently interact with artificial intelligence. Conducted among 263 data scientists, it revealed a demographic skew towards individuals aged 25 to 34, with many holding a Master’s degree. Notably, while 88% of respondents reported using AI tools, 72% of those not currently using AI expressed their companies have intentions to adopt these technologies soon. One key finding from the survey was the positive perception of AI’s role in the workplace. A significant majority (54.79%) viewed AI as offering new challenges, while 40.64% saw it as augmenting their current tasks. This suggests a broad recognition of AI’s potential to enhance productivity and efficiency in data science.

Regarding AI’s impact on work processes, most respondents noted increased productivity (57%) and improved efficiency (35%) due to AI. However, there were concerns about bias (38%), hallucinations (27%), and ethical considerations (23%). These findings underscore the need for careful integration and oversight when deploying AI tools. Training and support were identified as crucial for effective AI integration, while 62% of respondents had not received formal training on using AI, 70% felt adequately supported in incorporating AI into their work which highlights the importance of comprehensive training programs to maximize AI’s benefits. Looking ahead, a majority of respondents (73%) expected AI to play a more significant role in their work, indicating strong confidence in its future relevance. There was also interest in more advanced AI features, such as real-time data access and the development of general AI.

The Action-Research methodology further clarified ChatGPT's practical applications and limitations and involved iterative cycles of diagnosing, action planning, action execution and evaluation. Both a data scientist and ChatGPT faced a clustering problem using the "Customer Personality Analysis" dataset. While ChatGPT showed efficiency in preprocessing and applying clustering algorithms, the human data scientist’s detailed exploratory data analysis (EDA) and creative feature engineering led to more defined clusters.

Interviews with experienced data scientists offered deeper insights, emphasizing the thoroughness and depth of the human-led analysis. The experts praised the extensive EDA and innovative feature engineering, noting these as crucial for understanding the dataset and enhancing clustering outcomes. In contrast, ChatGPT, besides being competent in preprocessing and clustering, failed to explore deeper in all phases of the CRISP-DM process, affecting the clarity and separation of clusters. The experts noted that ChatGPT's approach missed important data patterns, which could influence clustering results.

The interviews also discussed the choice of clustering algorithms and evaluation metrics. The human data scientist's use of both K-Means and Agglomerative Clustering, along with comprehensive evaluation using the Silhouette Score, resulted in a more detailed approach.

Regarding the interpretability of results, t-SNE visualizations were important for understanding the clustering structure. The human-led analysis produced more distinct and interpretable clusters, aligning with the higher Silhouette Score. ChatGPT's clusters, however, showed more overlap, reflecting its lower Silhouette Score. The experts recommended that ChatGPT's analysis could benefit from more detailed preprocessing, such as handling outliers and examining feature correlations. When considering whether ChatGPT could replace human data scientists, the interviewees acknowledged the AI's efficiency in handling routine tasks. However, they emphasized the irreplaceable value of human expertise in nuanced decision-making, creative feature engineering, and interpreting complex data patterns. The agreement was that ChatGPT could significantly augment the capabilities of data scientists but could not entirely replace them.

As AI evolves, adopting a collaborative environment where human expertise and AI capabilities complement each other will be crucial. This thesis provides a foundational understanding of ChatGPT's role in data science, showing the way for further exploration and innovation in integrating AI tools across various domains. The insights from this research highlight the need for ongoing development to ensure AI technologies like ChatGPT are effectively and ethically utilized, advancing data science and beyond.

5.1 LIMITATIONS

Several limitations were encountered during this research, which should be noted to provide context to the findings and guide future studies.

Firstly, the dataset used in this study was relatively small, chosen deliberately to ensure that ChatGPT could handle the information effectively. This dataset, sourced from Kaggle, might not fully represent the diversity and complexity found in larger, real-world datasets. As a result, the generalizability of the findings may be limited. Additionally, the role of the data scientist in performing the clustering analysis was impersonated by the researcher, who is not an expert in the field. This limitation may have affected the quality and depth of the human-led analysis.

The methodology also had its constraints, since the survey was administered solely to data scientists working in Portugal, with a significant concentration of respondents from Lisbon. This geographical focus might have skewed the results. A larger, more diverse sample size, including data scientists from different regions and countries, could have provided a more comprehensive understanding of how data scientists interact with AI tools like ChatGPT. Additionally, the Action-Research methodology focused on clustering analysis, an unsupervised learning task, which inherently poses challenges in evaluating and validating the results.

Another limitation of the methodology is that only one cycle was used since the scope of this study was limited by time constraints, which made it challenging to implement multiple cycles within the available timeframe. y conducting a single, well-defined cycle allowed for a more focused and manageable analysis

Technological constraints were another limitation, since ChatGPT-4o, while demonstrating good performance, was not capable of directly performing the clustering analysis. The process required manually inputting ChatGPT's code into Google Colab to execute and visualize the results, such as t-SNE visualizations.

This workaround highlights current limitations in ChatGPT's ability to perform end-to-end data science workflows autonomously.

The study's findings are also limited by the specific nature of the dataset and the tasks performed. While the insights gained are valuable, they may not be fully applicable to other datasets, domains, or types of data science problems. Ethical considerations present additional challenges. ChatGPT, like other AI models, is prone to biases embedded in its training data and these biases can influence the reliability of its outputs, posing ethical dilemmas.

Finally, the qualitative analysis through interviews was influenced by the subjective nature of the experts' knowledge and experience. This subjectivity might have affected the evaluation and interpretation of the results.

5.2 RECOMMENDATIONS FOR FUTURE WORK

Future studies could use a broader range of datasets since the one used in this study was customer-specific and relatively small. Exploring larger datasets from different domains like healthcare, finance, or social media would provide a more comprehensive evaluation of ChatGPT's capabilities. Additionally, using other types of data science problems, like supervised learning tasks, could produce more measurable and varied insights.

ChatGPT's performance can vary based on the prompts provided. Future work should include specialized AI tools like GitHub Copilot, and trying different coding languages, such as R, could demonstrate the versatility of these AI tools. Including data scientists from various regions and countries can offer a more diverse perspective on AI use in data science. Alternatively, future research could focus on specific geographies to gain detailed regional insights. Additionally, comparing different AI tools rather than just AI to human performance could highlight the unique strengths and weaknesses of each tool. Employing multiple cycles of Action Research Methodology should also be considered to potentially achieve deeper insights and more robust results.

Technological improvements are crucial and future studies should incorporate advanced AI models like GitHub Copilot and compare their performance against ChatGPT. Evaluating these tools by experts in the field can provide a thorough assessment of their effectiveness.

Moreover, utilizing more powerful computational resources could help overcome the limitations faced by current AI models, enabling more complex analyses.

Addressing ethical considerations and biases in AI-generated outputs is essential. Developing specific prompts and frameworks to identify and mitigate biases is critical, and finally collaborating with AI development experts can refine these strategies, ensuring more ethical AI applications.

BIBLIOGRAPHICAL REFERENCES

Andrade, R. R. R., & Pinho, A. F. de. (2022). Desenvolvimento de uma ferramenta de auxílio a tomada de decisão com base na CRISP-DM. *Revista dos Trabalhos de Iniciação Científica*. https://doi.org/10.29327/1307153.1-32

Bakiner, O. (2023). The promises and challenges of addressing artificial intelligence with human rights. *Big Data & Society*, *10*(2), 20539517231205476. https://doi.org/10.1177/20539517231205476

Bano, M., Hoda, R., Zowghi, D., & Treude, C. (2024). Large language models for qualitative research in software engineering: Exploring opportunities and challenges. *Automated Software Engineering*, *31*(1). Scopus. https://doi.org/10.1007/s10515-023-00407-8

Bharadiya, J. (2023). A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning. *International Journal of Innovative Research in Science Engineering and Technology*, *8*, 2028. https://doi.org/10.5281/zenodo.8002436

Bucaioni, A., Ekedahl, H., Helander, V., & Nguyen, P. T. (2024). Programming with ChatGPT: How far can we go? *Machine Learning with Applications*, *15*, 100526. https://doi.org/10.1016/j.mlwa.2024.100526

Cai, Z., Yang, X., Huang, T., & Zhu, W. (2020). A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering. *Information Sciences*, *508*, 173–182. https://doi.org/10.1016/j.ins.2019.08.048

Calvo-Pardo, H., Mancini, T., & Olmo, J. (2023). Optimal deep neural networks by maximization of the approximation power. *Computers & Operations Research*, *156*, 106264. https://doi.org/10.1016/j.cor.2023.106264

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., & Tortora, G. (2024). Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Systems with Applications*, *235*, 121186. https://doi.org/10.1016/j.eswa.2023.121186

Chicco, D., Oneto, L., & Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. *PLOS Computational Biology*, *18*(12), e1010718. https://doi.org/10.1371/journal.pcbi.1010718

Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., & Hartline, D. K. (2020). t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Marine Genomics*, *51*, 100723. https://doi.org/10.1016/j.margen.2019.100723

Erfina, A., & Nurul, M. (2023). Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language. *Data & Metadata*, *2*, 45. https://doi.org/10.56294/dm202345

Eurostat. (2024, May 24). Statistics explained. Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*. https://doi.org/10.3389/fenrg.2021.652801

Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., & Chen, H. (2023). Investigating Code Generation Performance of ChatGPT with Crowdsourcing Social Data. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 876–885. https://doi.org/10.1109/COMPSAC57700.2023.00117

Hassani, H., & Silva, E. S. (2023). The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing*, *7*(2), Artigo 2. https://doi.org/10.3390/bdcc7020062

Hong, Z. (2023). ChatGPT for Computational Materials Science: A Perspective. *Energy Material Advances*, *4*, 0026. https://doi.org/10.34133/energymatadv.0026

Irvine, D. J., Halloran, L. J. S., & Brunner, P. (2023). Opportunities and limitations of the ChatGPT Advanced Data Analysis plugin for hydrological analyses. *Hydrological Processes*, *37*(10), e15015. https://doi.org/10.1002/hyp.15015

Januzaj, Y., Beqiri, E., & Luma, A. (2023). Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique. *International Journal of Online and Biomedical Engineering (iJOE)*, *19*(04), Artigo 04. https://doi.org/10.3991/ijoe.v19i04.37059

Kedam, N., Tiwari, D. K., Kumar, V., Khedher, K. M., & Salem, M. A. (2024). River stream flow prediction through advanced machine learning models for enhanced accuracy. *Results in Engineering*, *22*, 102215. https://doi.org/10.1016/j.rineng.2024.102215

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele, UK, Keele Univ.*, *33*.

Koubaa, A., Qureshi, B., Ammar, A., Khan, Z., Boulila, W., & Ghouti, L. (2023). Humans are still better than ChatGPT: Case of the IEEEXtreme competition. *Heliyon*, *9*(11). https://doi.org/10.1016/j.heliyon.2023.e21624

Lingo, R. (2023). *The Role of ChatGPT in Democratizing Data Science: An Exploration of AI-facilitated Data Analysis in Telematics* (arXiv:2308.02045). arXiv. https://doi.org/10.48550/arXiv.2308.02045

Liu, Y., Miller, L. K., & Niu, X. (2023). *Incorporating ChatGPT into a Financial Data Science Course with Python Programming* (SSRN Scholarly Paper 4412371). https://doi.org/10.2139/ssrn.4412371

Mahadi Hassan, M., Knipper, A., & Kanti Karmaker Santu, S. (2023). ChatGPT as your Personal Data Scientist. Em *arXiv e-prints*. https://doi.org/10.48550/arXiv.2305.13657

Nathalia, N., Paulo, A., & Donald, C. (2023). Artificial Intelligence vs. Software Engineers: An Empirical Study on Performance and Efficiency using ChatGPT. *Proceedings of the 33rd Annual International Conference on Computer Science and Software Engineering*, 24–33. https://doi.org/10.21203/rs.3.rs-3126005/v1

Otero-Escobar, A. D., & Velasco-Ramírez, M. L. (2023). Study on Exploratory Data Analysis Applied to Education. *2023 IEEE International Conference on Engineering Veracruz (ICEV)*, 1–5. https://doi.org/10.1109/ICEV59168.2023.10329702

Pang, S., Nol, E., & Heng, K. (2024). *ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects* (SSRN Scholarly Paper 4837988). https://doi.org/10.2139/ssrn.4837988

Peker, S., & Kart, Ö. (2023). Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review. *Journal of Data, Information and Management*, *5*(1), 1–21. https://doi.org/10.1007/s42488-023-00085-x

Pelster, M., & Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, *59*, 104786. https://doi.org/10.1016/j.frl.2023.104786

Polyportis, A., & Pahos, N. (2024). Navigating the perils of artificial intelligence: A focused review on ChatGPT and responsible research and innovation. *Humanities and Social Sciences Communications*, *11*(1). Scopus. https://doi.org/10.1057/s41599-023-02464-6

Postari, A. (2023). *Analyzing How Well ChatGPT Can Complete a Beginner-Level Python Course*. https://trepo.tuni.fi/handle/10024/151838

Prastyabudi, W. A., Alifah, A. N., & Nurdin, A. (2024). Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering. *Procedia Computer Science*, *234*, 96–105. https://doi.org/10.1016/j.procs.2024.02.156

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (2024). Exploring the impact of ChatGPT on education: A web mining and machine learning approach. *International Journal of Management Education*, *22*(1). Scopus. https://doi.org/10.1016/j.ijme.2024.100932

Richardson, C., Oster, N., Henriksen, D., & Mishra, P. (2024). Artificial Intelligence, Responsible Innovation, and the Future of Humanity with Andrew Maynard. *TechTrends*, *68*(1), 5–11. Scopus. https://doi.org/10.1007/s11528-023-00921-2

Santos, V., Amaral, L., & Mamede, H. (2013). *Utilização do método Investigação-Ação na investigação em criatividade no planeamento de sistemas de informação*. https://repositorium.sdum.uminho.pt/handle/1822/63852

Schindler, T. F., Schlicht, S., & Thoben, K.-D. (2023). Towards Benchmarking for Evaluating Machine Learning Methods in Detecting Outliers in Process Datasets. *Computers*, *12*(12). Scopus. https://doi.org/10.3390/computers12120253

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Sewwandi, M. A. N. D., Li, Y., & Zhang, J. (2024). K-outlier removal based on contextual label information and cluster purity for continuous data classification. *Expert Systems with Applications*, *237*, 121347. https://doi.org/10.1016/j.eswa.2023.121347

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Computing Surveys*, *55*(13s), 293:1-293:39. https://doi.org/10.1145/3588433

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3571884.3597144

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*. Scopus. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Sun, W., Fang, C., You, Y., Miao, Y., Liu, Y., Li, Y., Deng, G., Huang, S., Chen, Y., Zhang, Q., Qian, H., Liu, Y., & Chen, Z. (2023). *Automatic Code Summarization via ChatGPT: How Far Are We?* (arXiv:2305.12865). arXiv. http://arxiv.org/abs/2305.12865

Surameery, N. M. S., & Shakor, M. Y. (2023). Use Chat GPT to Solve Programming Bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN : 2455-5290*, *3*(01), Artigo 01. https://doi.org/10.55529/ijitc.31.17.22

Tiwari, M., Kang, R., Lee, D., Thrun, S., Shomorony, I., & Zhang, M. J. (2023). BanditPAM++: Faster k-medoids Clustering. *Advances in Neural Information Processing Systems*, *36*, 73371–73382.

Tsai, Y.-C. (2023). Empowering Learner-Centered Instruction: Integrating ChatGPT Python API and Tinker Learning for Enhanced Creativity and Problem-Solving Skills. Em Y.-M. Huang & T. Rocha (Eds.), *Innovative Technologies and Learning* (Vol. 14099, pp. 531–541). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_52

Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2023). *Examining the Potential and Pitfalls of ChatGPT in Science and Engineering Problem-Solving* (arXiv:2310.08773). arXiv. https://doi.org/10.48550/arXiv.2310.08773

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? *Emerging Markets Finance and Trade*, *58*(2), 472–482. https://doi.org/10.1080/1540496X.2020.1825935

Zeb, A., Ullah, R., & Karim, R. (2024). Exploring the role of ChatGPT in higher education: Opportunities, challenges and ethical considerations. *International Journal of Information and Learning Technology*. Scopus. https://doi.org/10.1108/IJILT-04-2023-0046

Zhan, X., Xu, Y., & Sarkadi, S. (2023). Deceptive AI Ecosystems: The Case of ChatGPT. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6. https://doi.org/10.1145/3571884.3603754

Zhang, W., Wang, Y., Song, Y., Wei, V. J., Tian, Y., Qi, Y., Chan, J. H., Wong, R. C.-W., & Yang, H. (2023). *Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey* (arXiv:2310.17894). arXiv. http://arxiv.org/abs/2310.17894

Zuccon, G., Koopman, B., & Shaik, R. (2023). ChatGPT Hallucinates when Attributing Answers. *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 46–51. https://doi.org/10.1145/3624918.3625329

APPENDIX A – VALIDATION OF THE SURVEY STRUCTURE

Thank you for agreeing to participate in the validation of this survet. Your feedback is crucial for refining the survey before its public release. Your responses will be kept anonymous, and the information provided will be used solely for the purpose of improving the survey.

Demographic Information:

1. Gender:
    - A) Male
    - B) Female
    - C) Prefer not to say
2. Age:
    - Please specify your age range:
        A) 18 – 25
        B) 25 – 35
        C) 35 – 45
        D) 45 – 55
        E) > 55
3. Highest Level of Education:
    - A) Bachelor's degree
    - B) Master's degree
    - C) Ph.D. or higher
    - D) Other (Please specify): __________ (open-ended)
4. Field of Study:: __________ (open-ended)
5. Years of Experience in Data Science:
    - A) Less than 1 year
    - B) 1-3 years
    - C) 4-6 years
    - D) 7+ years
6. Industry/Field of Work: __________ (open-ended)
7. Job Title/Position: : __________ (open-ended)

8. Company Size:
    - A) Small (1-50 employees)
    - B) Medium (51-500 employees)
    - C) Large (501+ employees)
9. Geographic Region:
    - Please specify your geographic region: __________ (drop-down list)
10. How frequently do you interact with AI technologies in your current role?
    - A) Daily
    - B) Weekly
    - C) Monthly
    - D) Rarely
    - E) Never

Survey Feedback:

1. How clear and concise do you find the instructions provided at the beginning of the survey? (Scale: 1 - Not clear at all, 5 - Very clear)
2. Do you think the questions cover relevant aspects of the relationship between data scientists and Artificial Intelligence (AI)? If not, please provide suggestions for additional questions or areas to cover.
3. Are there any questions that you found confusing or difficult to answer? If yes, please specify which questions and provide suggestions for clarification.
4. Do you believe the survey adequately addresses the concerns and considerations relevant to data scientists' interactions with AI technologies? If not, please provide suggestions for improvement.
5. Are there any demographic or professional details about data scientists that you think should be included in the survey for better context? If yes, please specify which details and provide reasoning.
6. Overall, do you think the survey is suitable for gathering insights from data scientists about their experiences with AI technologies? If not, please provide suggestions for improvement.

APPENDIX B – SURVEY STRUCTURE

Welcome to the research study!

Dear Participant, Thank you for taking the time to participate in this research study. Your insights and experiences are invaluable contributions to a master's thesis in Information Management aimed at exploring the relationship between data scientists in Portugal and Artificial Intelligence (AI). The study should take you around 5 minutes to complete. Your participation in this research is voluntary, and your responses will be kept anonymous and confidential. You have the right to withdraw at any point during the study.

By clicking the button below, you acknowledge:

- Your participation in the study is voluntary.

- You are 18 years of age.

- You are a Data Scientist that is working in Portugal

- You are aware that you may choose to terminate your participation at any time for any reason.

- I consent
- I do not consent, I do not wish to participate

Demographic Information

Gender:

- Male
- Female
- Non-binary / third gender
- Prefer not to say

Age

- 18 – 24
- 25 – 34
- 35 – 44
- 45 – 54
- 55 +

Geographical Region

- (Drop-down list of all districts in Portugal)

Highest Level of Education

- Bachelor's degree
- Master's degree
- Ph.D or higher
- Other

Field of Study

- Open answer

Years of Experience as a Data Scientist:

- Less than 1 year
- 1 – 3 years
- 4 – 6 years
- 7+ years

Industry/Field of Work

- Open answer

Job Title/Position

- Open answer

Company Size

- Small (<50 employees)
- Medium (51-500 employees)
- Large (501+ employees)

Usage of AI at Work

1. Are you currently using any AI-powered tools or models in your work as a data scientist?

- Yes
- No

1.1. If yes, please specify which AI-powered tools are you using.

- Open answer

1.1. If no, is your company considering introducing AI technologies soon?

- Yes

- No
- Unsure

2. Do you have the freedom to use AI-powered conversational agents (e.g., ChatGPT) in your work?

- Yes
- No

Perception of AI

3. How do you perceive the role of AI in your work as a data scientist? (Select the one that applies better)

- Augmentation of current work
- Threat
- Opportunity for new challenges
- Other

4. What benefits do you see in using AI in your work? (Select all that apply)

- Increased efficiency
- Improved accuracy
- Time-saving
- Other

5. What concerns do you have about using AI in your work? (Select all that apply)

- Ethical considerations
- Job displacement
- Hallucinations
- Bias

Impact on Work Process

6. How has the introduction of AI technologies affected your work processes and workflows? (Select the one that applies better)

- Improved efficiency
- Increased Productivity
- No significant impact
- Decreased productivity

- Other

Interaction with AI

7. Do you consider yourself more a developer or a user of AI? (Select the one that applies better)

- User of AI tools
- Developer of AI models
- Both user and developer
- Other

8. How do you typically interact with AI technologies in your work? (Select all that apply)

- Data Analysis
- Text Generation
- Task Automation
- Other

Challenges and Limitations

9. What are the main challenges or limitations you encounter when using AI technologies in your work? (Select all that apply)

- Data quality issues
- Lack of AI interpretability
- Bias
- Other

Training and Support

10. Have you received any formal training or guidance on how to effectively use AI technologies in your work?

- Yes
- No

11. Do you feel adequately supported in integrating AI technologies into your work processes?

- Yes
- No

Future Expectations

12. How do you foresee the role of AI evolving in your work as a data scientist in the future? (Select the one that applies better)

- More prominent role
- Similar Role
- Less prominent role

13. Are there any specific features or capabilities you would like to see in future AI technologies to better meet your needs? (If yes, please specify)

- Yes
- No

APPENDIX C – DATA SCIENTIST'S CODE

```
# Importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.metrics import silhouette_score
from sklearn.manifold import TSNE
import warnings
warnings.filterwarnings('ignore')

# Allow Colab to see Google Drive files

from google.colab import drive
drive.mount('/content/drive')

# Load csv file into a dataframe
data_path = "/content/drive/MyDrive/thesis/marketing_campaign.csv"

# Since colab assumed that the file is comma delimited the following
correction was applied
data = pd.read_csv(data_path, delimiter='\t') #tab delimiter

new_csv_file = '_marketing_campaign.csv'
data.to_csv(new_csv_file, index=False) #new csv file made

print(f"New CSV file saved to {new_csv_file}")

# Create a copy of the original data
original_data = data.copy()

data = pd.read_csv('_marketing_campaign.csv')
print(data.head())

#Looking at the missing values and data types
print("\nData types and missing value counts:")
print(data.info())

# Date format for Dt_Customer needs to be corrected
data['Dt_Customer'] = pd.to_datetime(data['Dt_Customer'], format='%d-
%m-%Y')

# Visualize histogram for numeric features
```

```
plt.figure(figsize=(20, 20))
data[numeric_features].hist(bins=10, layout=(6, 5), figsize=(20, 20),
grid=False, alpha=0.7)
plt.show()

#Visualize boxplots for numeric features
plt.figure(figsize=(20, 20))
for i, feature in enumerate(numeric_features):
    plt.subplot(6, 5, i + 1)
    sns.boxplot(y=data[feature])
    plt.title(feature)
    plt.tight_layout()

plt.show()

# Correct Education hierarchy
education_order = ["Basic", "2n Cycle", "Graduation", "Master", "PhD"]
data['Education'] = pd.Categorical(data['Education'],
categories=education_order, ordered=True)

# Define categorical features
categorical_features =
data.select_dtypes(exclude=[np.number]).columns.tolist()

# Plot bar charts for categorical features Education and Marital_Status
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
data['Education'].value_counts().plot(kind='bar')
plt.title('Education')
plt.xlabel('Education')
plt.ylabel('Count')

plt.subplot(1, 2, 2)
data['Marital_Status'].value_counts().plot(kind='bar')
plt.title('Marital_Status')
plt.xlabel('Marital Status')
plt.ylabel('Count')

plt.tight_layout()
plt.show()

# Plot box plot for Dt_Customer
plt.figure(figsize=(10, 5))
sns.boxplot(y=data['Dt_Customer'])
plt.title('Box plot of Dt_Customer')
plt.ylabel('Dt_Customer')
plt.show()
```

```
# Handle missing values for Income using median
data['Income'].fillna(data['Income'].median(), inplace=True)

# Check for duplicates
duplicate_rows = data[data.duplicated()]
print(f"Number of duplicate rows: {duplicate_rows.shape[0]}")

# Create a copy of the cleaned data before removing outliers
cleaned_data = data.copy()

# Manually filtering outliers for all specified features
filter_conditions = (
    (cleaned_data['Income'] <= 200000) &
    (cleaned_data['MntWines'] <= 1500) &
    (cleaned_data['MntGoldProds'] <= 300) &
    (cleaned_data['NumWebVisitsMonth'] <= 20) &
    (cleaned_data['Year_Birth'] >= 1930) &
    (cleaned_data['MntMeatProducts'] <= 1000) &
    (cleaned_data['NumWebPurchases'] <= 15) &
    (cleaned_data['MntSweetProducts'] <= 100)
)

# Apply filter conditions
filtered_data = cleaned_data[filter_conditions]

# Print the percentage of data kept after manually removing outliers
print('% of data kept after manually removing outliers:',
      np.round(filtered_data.shape[0] / original_data.shape[0], 5))

# Plot box plots again to confirm removal of outliers
plt.figure(figsize=(20, 20))
for i, feature in enumerate(numeric_features):
    plt.subplot(6, 5, i + 1)
    sns.boxplot(y=filtered_data[feature])
    plt.title(feature)
plt.tight_layout()
plt.show()

# Feature engineering

# Create a copy of the filtered data before feature engineering
feature_data = filtered_data.copy()

#Total amount spent
feature_data["Spent"] = (
    feature_data["MntWines"] +
    feature_data["MntFruits"] +
    feature_data["MntMeatProducts"] +
    feature_data["MntFishProducts"] +
```

```
    feature_data["MntSweetProducts"] +
    feature_data["MntGoldProds"]
)

#Create a feature to easier understand marital status
feature_data["Living_With"] = feature_data["Marital_Status"].replace({
    "Married": "Partner",
    "Together": "Partner",
    "Absurd": "Alone",
    "Widow": "Alone",
    "YOLO": "Alone",
    "Divorced": "Alone",
    "Single": "Alone",
})

#See if the customer has children
feature_data["Children"] = feature_data["Kidhome"] +
feature_data["Teenhome"]

#Understand the age of the customer based on Year_Birth
feature_data["Age"] = datetime.now().year - feature_data["Year_Birth"]

# Calculate the duration of customer relationships in years
feature_data['Dt_Customer'] =
pd.to_datetime(feature_data['Dt_Customer'])
current_date = pd.to_datetime('today')
feature_data['Customer_Duration'] = (current_date -
feature_data['Dt_Customer']).dt.days / 365

# Creating a feature to see if the customer accepted any campaign
'AcceptedAnyCampaign'
campaign_columns = ['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3',
'AcceptedCmp4', 'AcceptedCmp5']
feature_data['AcceptedAnyCampaign'] =
feature_data[campaign_columns].sum(axis=1).apply(lambda x: 1 if x > 0
else 0)

# Remove unnecessary columns
columns_to_drop = ["ID", "Year_Birth", "Z_CostContact", "Z_Revenue",
"Dt_Customer", "Marital_Status"] + campaign_columns
feature_data.drop(columns=columns_to_drop, inplace=True)

# Display the first few rows of the updated DataFrame
print(feature_data.head())

# One-hot encoding for categorical variables
categorical_features = ['Education', 'Living_With']
feature_data = pd.get_dummies(feature_data,
columns=categorical_features, drop_first=True)
```

```
# Standardize the numeric features
numeric_features =
feature_data.select_dtypes(include=[np.number]).columns.tolist()
scaler = StandardScaler()
feature_data[numeric_features] =
scaler.fit_transform(feature_data[numeric_features])

# Display the first few rows of the updated DataFrame
print(feature_data.head())

# Correlation heatmap for numerical features
plt.figure(figsize=(20, 15))
corr_matrix = feature_data[numeric_features].corr()
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()

# Drop highly correlated features
features_to_drop = ['MntWines', 'MntMeatProducts',
'NumCatalogPurchases']
feature_data.drop(columns=features_to_drop, inplace=True)

# Scree plot
pca = PCA().fit(feature_data)
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(pca.explained_variance_)+1),
pca.explained_variance_, marker='o')
plt.xlabel('Number of Components')
plt.ylabel('Eigenvalue')
plt.title('Scree Plot')
plt.grid(True)
plt.show()

# Apply PCA
pca = PCA(n_components=5)
pca_results = pca.fit_transform(feature_data[numeric_features])

print(f'Explained Variance Ratio: {pca.explained_variance_ratio_}')

# Elbow Method to determine the number of clusters
wcss = []
max_clusters = 10
for i in range(1, max_clusters + 1):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(pca_results)
    wcss.append(kmeans.inertia_)

# Plotting the Elbow Method results
```

```
plt.figure(figsize=(10, 6))
plt.plot(range(1, max_clusters + 1), wcss, marker='o')
plt.title('Elbow Method to Determine Optimal Number of Clusters')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.xticks(range(1, max_clusters + 1))
plt.grid(True)
plt.show()

# Perform KMeans with 3 clusters
kmeans_3 = KMeans(n_clusters=3, random_state=42)
kmeans_labels_3 = kmeans_3.fit_predict(pca_results)

# Perform Agglomerative Clustering with 3 clusters
agglo_3 = AgglomerativeClustering(n_clusters=3)
agglo_labels_3 = agglo_3.fit_predict(pca_results)

# Calculate silhouette scores
silhouette_kmeans_3 = silhouette_score(pca_results, kmeans_labels_3)
silhouette_agglo_3 = silhouette_score(pca_results, agglo_labels_3)

print(f'Silhouette Score for KMeans with 3 clusters:
{silhouette_kmeans_3}')
print(f'Silhouette Score for Agglomerative Clustering with 3 clusters:
{silhouette_agglo_3}')

# Perform KMeans with 4 clusters
kmeans_4 = KMeans(n_clusters=4, random_state=42)
kmeans_labels_4 = kmeans_4.fit_predict(pca_results)

# Perform Agglomerative Clustering with 4 clusters
agglo_4 = AgglomerativeClustering(n_clusters=4)
agglo_labels_4 = agglo_4.fit_predict(pca_results)

# Calculate silhouette scores
silhouette_kmeans_4 = silhouette_score(pca_results, kmeans_labels_4)
silhouette_agglo_4 = silhouette_score(pca_results, agglo_labels_4)

print(f'Silhouette Score for KMeans with 4 clusters:
{silhouette_kmeans_4}')
print(f'Silhouette Score for Agglomerative Clustering with 4 clusters:
{silhouette_agglo_4}')

# t-SNE Visualization for 3 clusters
tsne_results_3 = TSNE(n_components=2, perplexity=30,
n_iter=300).fit_transform(pca_results)

plt.figure(figsize=(14, 7))
plt.subplot(1, 2, 1)
```

```
plt.scatter(tsne_results_3[:, 0], tsne_results_3[:, 1],
c=kmeans_labels_3, cmap='viridis')
plt.title('t-SNE Visualization for K-means Clustering (3 Clusters)')

plt.subplot(1, 2, 2)
plt.scatter(tsne_results_3[:, 0], tsne_results_3[:, 1],
c=agglo_labels_3, cmap='viridis')
plt.title('t-SNE Visualization for Agglomerative Clustering (3
Clusters)')
plt.show()

# t-SNE Visualization for 4 clusters
tsne_results_4 = TSNE(n_components=2, perplexity=30,
n_iter=300).fit_transform(pca_results)

plt.figure(figsize=(14, 7))
plt.subplot(1, 2, 1)
plt.scatter(tsne_results_4[:, 0], tsne_results_4[:, 1],
c=kmeans_labels_4, cmap='viridis')
plt.title('t-SNE Visualization for K-means Clustering (4 Clusters)')

plt.subplot(1, 2, 2)
plt.scatter(tsne_results_4[:, 0], tsne_results_4[:, 1],
c=agglo_labels_4, cmap='viridis')
plt.title('t-SNE Visualization for Agglomerative Clustering (4
Clusters)')
plt.show()

filtered_data['original_index'] = filtered_data.index

# Perform clustering (use pca_results from the previous step)
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans_labels_3 = kmeans.fit_predict(pca_results)

# Add the cluster labels to filtered_data
filtered_data['Cluster'] = kmeans_labels_3

# Merge the cluster labels back to the original data using the
original_index
original_data = original_data.merge(filtered_data[['original_index',
'Cluster']], left_index=True, right_on='original_index', how='left')

# Drop the temporary original_index column
original_data.drop(columns=['original_index'], inplace=True)

# Select only numeric columns for the cluster analysis
numeric_columns =
original_data.select_dtypes(include=[np.number]).columns.tolist()
```

```
# Perform analysis of the clusters
cluster_analysis =
original_data.groupby('Cluster')[numeric_columns].mean()
print(cluster_analysis)

# Visualize the distribution of the clusters
plt.figure(figsize=(12, 6))
sns.countplot(x='Cluster', data=original_data)
plt.title('Distribution of Clusters')
plt.show()

# Ensure you are only working with numeric features for clustering
analysis
numeric_columns =
original_data.select_dtypes(include=[np.number]).columns.tolist()

# Normalize the cluster means for radar plot
cluster_means =
original_data.groupby('Cluster')[numeric_columns].mean()
cluster_means = cluster_means.apply(lambda x: (x - x.min()) / (x.max()
- x.min()), axis=0)

import numpy as np
import matplotlib.pyplot as plt

def radar_factory(num_vars, frame='circle'):
    theta = np.linspace(0, 2 * np.pi, num_vars,
endpoint=False).tolist()
    theta += theta[:1]
    return theta

def plot_radar(labels, values, title, colors):
    num_vars = len(labels)
    theta = radar_factory(num_vars)

    fig, ax = plt.subplots(figsize=(9, 9),
subplot_kw=dict(projection='polar'))
    plt.xticks(theta[:-1], labels, color='grey', size=12)

    for i, (label, value, color) in enumerate(zip(labels, values,
colors)):
        ax.plot(theta, np.append(value, value[0]), color=color,
linewidth=2, label=f'Cluster {i}')
        ax.fill(theta, np.append(value, value[0]), color=color,
alpha=0.25)

    ax.set_ylim(0, 1)
    plt.title(title, size=20, color='grey', y=1.1)
    plt.legend(loc='upper right', bbox_to_anchor=(1.3, 1.1))
```

```
    plt.show()

# Define labels and values for radar chart
labels = cluster_means.columns
values = cluster_means.values

# Define colors for each cluster
colors = ['b', 'r', 'g', 'y']

# Plot radar chart
plot_radar(labels, values, 'Cluster Profiles', colors)

# Standardize the data
scaler = StandardScaler()
standardized_data =
scaler.fit_transform(original_data[numeric_columns])

# Create a new DataFrame for the standardized data
standardized_df = pd.DataFrame(standardized_data,
columns=numeric_columns)
standardized_df['Cluster'] =
original_data['Cluster'].reset_index(drop=True)

# Calculate the mean of each feature for each cluster
cluster_means_standardized = standardized_df.groupby('Cluster').mean()

# Plot the heatmap
plt.figure(figsize=(15, 10))
sns.heatmap(cluster_means_standardized.T, cmap='coolwarm', annot=True)
plt.title('Heatmap of Standardized Cluster Feature Means')
plt.show()
```

APPENDIX D - CHATGPT'S CODE

```
#Importing
import pandas as pd

## Allow Colab to see Google Drive files

from google.colab import drive
drive.mount('/content/drive')

## Load csv file into a dataframe
data_path = "/content/drive/MyDrive/thesis/marketing_campaign.csv"

data = pd.read_csv(data_path, delimiter='\t') #tab delimiter

# Display the first few rows of the dataframe
data.head()

# Display basic information about the dataframe
data.info()

# Describe the dataframe to get a summary of the statistics for each
column
data.describe(include='all')

# Checking for missing values
data.isnull().sum()

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer

# Identifying categorical and numerical columns
categorical_cols = ['Education', 'Marital_Status']
numerical_cols = [col for col in data.columns if col not in
categorical_cols + ['ID', 'Dt_Customer']]

# Preprocessing pipeline for numerical data
numerical_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

# Preprocessing pipeline for categorical data
categorical_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
```

```
])

# Combining numerical and categorical pipelines
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_pipeline, numerical_cols),
        ('cat', categorical_pipeline, categorical_cols)
    ])

# Applying transformations to the dataset
data_preprocessed = preprocessor.fit_transform(data)

# Convert the preprocessed data back to a DataFrame for convenience
data_preprocessed_df = pd.DataFrame(data_preprocessed)


# Displaying the first few rows of the preprocessed dataframe
data_preprocessed_df.head()

from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Determine the optimal number of clusters using the elbow method
sse = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_preprocessed)
    sse.append(kmeans.inertia_)

# Plot the elbow curve
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), sse, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('SSE')
plt.title('Elbow Method')
plt.show()

from sklearn.metrics import silhouette_score

# Applying K-Means clustering with the chosen number of clusters
n_clusters = 5
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
cluster_labels = kmeans.fit_predict(data_preprocessed)

# Evaluating the silhouette score
silhouette_avg = silhouette_score(data_preprocessed, cluster_labels)
print(f'Silhouette Score for {n_clusters} clusters: {silhouette_avg}')

from sklearn.manifold import TSNE
```

```
# Applying t-SNE to the preprocessed data
tsne = TSNE(n_components=2, random_state=42)
tsne_results = tsne.fit_transform(data_preprocessed)

# Creating a DataFrame for the t-SNE results
tsne_df = pd.DataFrame(tsne_results, columns=['TSNE1', 'TSNE2'])
tsne_df['Cluster'] = cluster_labels

# Plotting the t-SNE results
plt.figure(figsize=(10, 6))
plt.scatter(tsne_df['TSNE1'], tsne_df['TSNE2'], c=tsne_df['Cluster'],
cmap='viridis', s=50)
plt.title('t-SNE visualization of customer clusters')
plt.xlabel('TSNE1')
plt.ylabel('TSNE2')
plt.colorbar(label='Cluster')
plt.show()

# Applying K-Means clustering with 4 clusters
n_clusters_4 = 4
kmeans_4 = KMeans(n_clusters=n_clusters_4, random_state=42)
cluster_labels_4 = kmeans_4.fit_predict(data_preprocessed)

# Evaluating the silhouette score for 4 clusters
silhouette_avg_4 = silhouette_score(data_preprocessed,
cluster_labels_4)
print(f'Silhouette Score for {n_clusters_4} clusters:
{silhouette_avg_4}')

# Visualizing with t-SNE
tsne_4 = TSNE(n_components=2, random_state=42)
tsne_results_4 = tsne_4.fit_transform(data_preprocessed)

tsne_df_4 = pd.DataFrame(tsne_results_4, columns=['TSNE1', 'TSNE2'])
tsne_df_4['Cluster'] = cluster_labels_4

# Plotting the t-SNE results for 4 clusters
plt.figure(figsize=(10, 6))
plt.scatter(tsne_df_4['TSNE1'], tsne_df_4['TSNE2'],
c=tsne_df_4['Cluster'], cmap='viridis', s=50)
plt.title('t-SNE visualization of customer clusters with 4 clusters')
plt.xlabel('TSNE1')
plt.ylabel('TSNE2')
plt.colorbar(label='Cluster')
plt.show()

# Applying K-Means clustering with 6 clusters
n_clusters_6 = 6
```

```
kmeans_6 = KMeans(n_clusters=n_clusters_6, random_state=42)
cluster_labels_6 = kmeans_6.fit_predict(data_preprocessed)

# Evaluating the silhouette score for 6 clusters
silhouette_avg_6 = silhouette_score(data_preprocessed,
cluster_labels_6)
print(f'Silhouette Score for {n_clusters_6} clusters:
{silhouette_avg_6}')

# Visualizing with t-SNE
tsne_6 = TSNE(n_components=2, random_state=42)
tsne_results_6 = tsne_6.fit_transform(data_preprocessed)

tsne_df_6 = pd.DataFrame(tsne_results_6, columns=['TSNE1', 'TSNE2'])
tsne_df_6['Cluster'] = cluster_labels_6

# Plotting the t-SNE results for 6 clusters
plt.figure(figsize=(10, 6))
plt.scatter(tsne_df_6['TSNE1'], tsne_df_6['TSNE2'],
c=tsne_df_6['Cluster'], cmap='viridis', s=50)
plt.title('t-SNE visualization of customer clusters with 6 clusters')
plt.xlabel('TSNE1')
plt.ylabel('TSNE2')
plt.colorbar(label='Cluster')
plt.show()

# Adding cluster labels to the original dataframe
data['Cluster'] = cluster_labels_6

# Calculating the mean values of features for each cluster (with
errors)
cluster_means = data.groupby('Cluster').mean()

# Selecting only the numerical columns for aggregation
numerical_data = data.select_dtypes(include=['int64', 'float64'])

# Adding cluster labels to the numerical dataframe
numerical_data['Cluster'] = cluster_labels_6

# Calculating the mean values of numerical features for each cluster
cluster_means = numerical_data.groupby('Cluster').mean()

#  (with errors)
import ace_tools as tools;
tools.display_dataframe_to_user(name="Cluster Means",
dataframe=cluster_means)

# Selecting only the numerical columns for aggregation
numerical_data = data.select_dtypes(include=['int64', 'float64'])
```

```python
# Adding cluster labels to the numerical dataframe
numerical_data['Cluster'] = cluster_labels_6

# Calculating the mean values of numerical features for each cluster
cluster_means = numerical_data.groupby('Cluster').mean()

# Displaying the cluster means
cluster_means

import seaborn as sns
import matplotlib.pyplot as plt

# Melting the data for better visualization
spending_columns = ['MntWines', 'MntFruits', 'MntMeatProducts',
'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']
melted_data = pd.melt(data, id_vars=['Cluster'],
value_vars=spending_columns, var_name='Product', value_name='Spending')

# Plotting the spending patterns
plt.figure(figsize=(12, 8))
sns.boxplot(x='Product', y='Spending', hue='Cluster', data=melted_data)
plt.title('Spending Patterns Across Clusters')
plt.xlabel('Product Category')
plt.ylabel('Spending')
plt.legend(title='Cluster')
plt.show()

# Analyzing campaign responses across clusters
campaign_columns = ['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3',
'AcceptedCmp4', 'AcceptedCmp5', 'Response']
campaign_means =
numerical_data.groupby('Cluster')[campaign_columns].mean()

# Displaying the campaign response means
campaign_means
```

APPENDIX E – ETHICS COMMITTEE REPORT

NOVA IMS Information Management School

This is to certify that
Project No.: **OTHER2024-7-148461**
Project Title: **ChatGPT as a Data Scientist: Can AI Handle Clustering Better Than a Human?**
Principal Researcher: **Alexandru Lemesev**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 7/14/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of
- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 7/14/2024

NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

ANNEXES

Table with all attributes of the dataset.

Attribute	Description
ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Customer's education level
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise
MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
NumDealsPurchases	Number of purchases made with a discount
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response	1 if customer accepted the offer in the last campaign, 0 otherwise
NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month

NOVA IMS
Information Management School

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa