BEMM466
BUSINESS PROJECT

CUSTOMER SEGMENTATION AND PRODUCT RECOMMENDATION USING DATA-DRIVEN MARKETING ANALYTICS & MACHINE LEARNING TECHNIQUES

NAY LINN (NOAH) AUNG (710082397)

Acknowledgment: I would like to extend my heartfelt appreciation to all those who have assisted me in completing this project.

First and foremost, my sincere gratitude is extended to **Mohsen Mosleh**, assistant professor at the University of Exeter Business School, for his invaluable guidance, insightful feedback and continuous encouragement. His dedication and expertise have been crucial in the development of this project's direction and success.

I am also deeply appreciative of **Dr. Stuart So**, the programme director for MSc Business Analytics at the University of Exeter Business School, for his unwavering support and for providing the resources and platform that were essential for the successful completion of this project. His dedication to excellence and leadership have significantly influenced my academic journey.

Lastly, I would like to express my gratitude to my family, friends, and colleagues for their unwavering support and motivation, which have enabled me to complete this project.

Executive Summary

Title: Customer Segmentation and Product Recommendation Using Data-Driven Marketing Analytics & Machine Learning Techniques

According to International Trade Administration (2022), the UK e-commerce sector, which is currently the third-largest in the world, is experiencing tremendous growth, which presents both opportunities and challenges for businesses. To remain competitive in these ever-changing customer expectations, businesses must implement sophisticated data analytics and machine learning (ML) algorithms. This project addresses the urgent need for UK e-commerce companies to better understand customer behaviours and personalised shopping experiences through effective customer segmentation and product recommendations.

This project aims to develop a comprehension of customer behaviour within the UK e-commerce industry. The objective is to enhance customer engagement and sales by identifying distinct customer groups and providing personalised product recommendations through the use of advanced marketing analytics and machine learning techniques. The following specific objectives are what this paper aims to achieve:

Identify Customer Segments: Use advanced data analytics to divide customers into groups based on their behaviours and preferences.
Develop Personalised Recommendations: Develop algorithms that can make product suggestions more relevant to certain customer groups, which will improve the shopping experience.
Optimise Marketing Strategies: Use insights from customer segmentation to develop marketing campaigns and strategies that more effectively and efficiently target certain consumer groups.
Increase Customer Engagement: Improve customer satisfaction and engagement by making more appropriate and relevant product recommendations.

The research employed a quantitative approach, using secondary datasets from the UK e-commerce sector to analyse customer preferences and behaviours. The following are the four phases employed in this study:

Data Collection: The collection of extensive secondary data included customer transaction records, website analytics, market research reports, customer demographics and technology usage data.

Data Preprocessing: To get high-quality input for analysis, the raw data was cleaned, integrated, reduced and transformed.

Customer Segmentation: Based on the demographic and behavioural data, K-means clustering was used to segment customers into different groups.

Product Recommendation: Three different recommendation models, content-based, collaborative, and hybrid model filterings, were developed, compared and provided suitable recommendation items for different customer groups.

For the key findings, the analysis identified three different customer groups within the UK e-commerce market. Segment 1 is a group with younger customers who have lower income levels and are highly responsive to promotions and discounts. This group prioritises essential goods and budget-friendly products. Segment 2 is filled with customers who love quality products and convenience while shopping. They are identified as time-conscious and high-income individuals who have a preference for more efficient shopping experiences. In the last segment 3, wealthy, predominantly older consumers who prefer luxury and premium products, are included. This category looks for exclusivity and brand reputation, thus personalised and high-quality products are essential for this group. When it comes to product recommendations, the study found that the most effective approach is the hybrid filtering model, which combines collaborative filtering and content-based filtering models. It effectively leverages both user behaviours and product attributes to provide personalised and precise product recommendations. The cold start problem in collaborative filtering and the narrow focus of content-based filtering were overcome by this model.

Customer segmentation and product recommendation systems have ethical implications, as with any data-driven approach. This study followed strict ethical standards to guarantee the safety of customer privacy and responsible data usage. Key ethical considerations are as follows:

Data Privacy: All of the data was anonymised to protect consumer information and identifications. Personal data was handled with utmost confidentiality and security, as a result of the strict adherence to the General Data Protection Regulation (GDPR).
Bias and Fairness: The paper acknowledged that both customer segmentation and recommendation systems might show algorithmic bias. To mitigate this risk, the algorithms were consistently reviewed and modified to guarantee that they delivered fair results for all the segments, preventing any unintentional bias.
Transparency: The methodologies and algorithms used in the research were thoroughly documented and made accessible to stakeholders to make sure they comprehend the processes and decisions associated with the data analysis.

From the findings of the study, there are several business implications for the UK e-commerce sector. By customising marketing strategies to the unique requirements and preferences of various customer segments, businesses can improve customer engagement. For example, segment 1 can be targeted with promotions that prioritise value, while segment 3 can be engaged with premium experiences and exclusive offers. Furthermore, businesses can improve customer satisfaction and loyalty by offering more relevant and diverse product recommendations through the use of hybrid recommendation model. On top of that, e-commerce platforms can more effectively allocate marketing resources by comprehending the unique characteristics of each consumer segment, focusing on the most profitable segments.

Although this study makes significant contributions to the understanding of customer segmentation and product recommendations in the UK e-commerce sector, additional research is required to address some of the areas. To offer more immediate and actionable insights, future research should explore the integration of real-time data. On top of that, further research should include external factors, including seasonal trends, cultural and technological contexts, to examine better customer behaviours and recommendation systems. By examining these factors, further research can create more adaptive and better recommendation systems that more accurately reflect customer behaviours. Finally, the ethical implications of data usage in e-commerce should be further explored through ongoing research, with particular focus on the balance between personalisation and privacy, consent and algorithmic bias (Naz & Kashif, 2024). By addressing these areas, future research can build upon the foundation created by this study, contributing to the ongoing evolution of personalised marketing in the e-commerce sector.

Table of Contents

1. Introduction

Over the past decade, the e-commerce sector in the United Kingdom has experienced significant growth, leading to a complete transformation of the retail business and influencing consumer behaviours. International Trade Administration (2022) stated that the UK became the third-largest e-commerce market in the world, surpassed only by China and the United States in 2023. Various factors fuel this growth, such as a tech-savvy population that increasingly favours the convenience of online purchasing over physical stores, advanced logistics and delivery systems and social media marketing. In 2022, e-commerce sales in the UK contributed around 30% of the overall retail sales, making a significant growth compared to prior years. Popular e-commerce products/services in the UK include fashion, electronics, beauty, health care, food & beverages with dominant market players AmazonUK, Sainsbury, Tesco and ASOS. In addition, e-commerce purchases made online using smartphones have surpassed that of tablets and computers, accounting for well over two-thirds of all e-commerce sales. The ever-changing nature of this environment presents both challenges and opportunities for companies and businesses. Although there is a huge possibility to reach a wide range of customers, the competition is intense, and customer expectations are always changing. In order to remain competitive, businesses have to utilise advanced technologies and data-driven strategies to understand and meet the needs of their customers (Ahmed, 2023).

It is crucial for every business to understand and meet the distinct preferences and requirements of its customers in the highly competitive UK e-commerce market. This can be achieved by strategically implementing two highly valuable approaches: customer segmentation and personalised product recommendations. Customer segmentation is the practice of categorising the customer group into smaller groups based on common features or traits. A simple method for segmentation involves geographic, demographic, psychographic and behavioural segmentation (Chongkolnee Rungruang et al., 2024). Furthermore, segmentation helps correlate between the specific group of consumers and the comprehension of their expectations. Customer segmentation is important in the marketing sector, as it enables decision-makers to make precise decisions and strategies in a timely manner, enhancing the marketing of specific products within the industry (Wang, 2022). For instance, younger consumers may be more responsive to influencer endorsements and social media marketing whereas the older generations may prefer loyalty programmes and email newsletters. Product recommendations are essential on all e-commerce platforms since they help users find appropriate and relevant products or services that align with their interests and previous interactions/behaviours. Recent advancements in machine learning (ML) and deep learning (DL) approaches have drastically transformed recommendation systems, resulting in more precise and customised recommendations (Nguyen et al., 2024). These techniques include dynamically displaying products or services to users on webpages, applications or emails. The process of product or service selection is based on several data points, including customer attributes, browsing behaviour and situational context. This approach aims to provide a personalised purchasing experience that increases customer satisfaction and loyalty (Monetate, n.d.). To summarise, implementing efficient customer segmentation and product recommendations not only enhances the overall customer experience but also contributes to the growth of the business by increasing conversion rates and sustaining long-term customer loyalty.

This project aims to develop a comprehension of customer behaviour within the UK e-commerce industry. The objective is to enhance customer engagement and sales by identifying distinct customer groups and providing personalised product recommendations through the use of advanced marketing analytics and machine learning techniques. The following specific objectives are what this paper aims to achieve:

Identify Customer Segments: Use advanced data analytics to divide customers into groups based on their behaviours and preferences.

Develop Personalised Recommendations: Develop algorithms that can make product suggestions more relevant to certain customer groups, which will improve the shopping experience.

Optimise Marketing Strategies: Use insights from customer segmentation to develop marketing campaigns and strategies that more effectively and efficiently target certain consumer groups.
Increase Customer Engagement: Improve customer satisfaction and engagement by making more appropriate and relevant product recommendations.

This study aims to provide e-commerce organisations with actionable insights that they can utilise to understand better and serve their customers. By achieving these objectives, online businesses can increase market growth and competitiveness. To guide this study, the following key research questions are addressed:

1. How can different customer groups be identified in the UK e-commerce market by effectively implementing customer segmentation?
2. How can personalised product recommendations be generated for customer segments to optimise engagement and sales?
3. In what ways does customer segmentation improve the accuracy and relevance of product recommendations in UK e-commerce?

These questions will be addressed through data analysis, the development of machine learning models, and practical implementation in the e-commerce context. This will result in a thorough understanding of how to improve customer engagement and business performance. This study is motivated by the need to address the challenges that e-commerce enterprises experience in understanding and meeting the different needs of their customers and users. To stay ahead in the highly competitive e-commerce market, it is crucial to implement advanced data analytics and machine learning approaches (BigCommerce, n.d.; Nguyen et al., 2024). This paper serves as a comprehensive guide, providing valuable insights and information, helping businesses achieve long-term success in a constantly evolving market setting.

2. Literature Review

2.1 Overview of Relevant Literature

The e-commerce sector has a diverse and wide range of literature regarding product recommendation systems and customer segmentation. Traditional methods of customer segmentation, including geographic, demographic and psychographic segmentation, have been the focus of numerous studies. Nevertheless, the development of big data and advanced analytics has facilitated the rise of more sophisticated approaches, such as behavioural and technographic segmentation (Ahlm et al., 2007; Nur Balqish Hassan & Noor Hazarina Hashim, 2024).

Numerous authors have written extensively on various techniques for segmenting customers. For example, Jiang & Tuzhilin (2009) determined that both customer segmentation and buyer targeting are essential for enhancing marketing performance. These two tasks are combined into a step-by-step approach, but the challenge encountered is unified optimisation. On the other hand, Yang et al. (2016) and Gomes & Meisen (2023) addressed that in marketing analytics, customer segmentation divides customers into groups with similar characteristics, while buyer-targeting aims to identify promising customers. Both these methods enhance marketing performances by allocating resources to the most profitable consumers. Researchers have conducted studies on integrating customer segmentation and buyer-targeting to make customised marketing strategies. The authors create a K-Means Segmentation algorithm, most used approach, to solve the unified optimisation issue. This method provides not only more accurate targeting but also significant purchasing decisions for each customer segment. However, Shah & Singh (2012) pointed out that the proposed K-Means algorithm reduces the cluster error score, but it does not provide a perfect solution in all cases. The author notes that the new method requires less time to execute than traditional methods since the number of clusters increases.

When it comes to personalised recommendation system, Cho & Moon (2013) suggested the use of weighted frequent pattern mining. Customer profiling is performed to identify potential customers utilising the RFM (Recency, Frequency, Monetary) approach. The author has added various weights to each transaction to get weighted association rules through the process of mining. Implementing the RFM approach will result in more precise customer recommendations, ultimately leading to higher profits for the company. Meanwhile, Afoudi et al. (2021) and Thorat et al. (2015) discussed about three recommender systems, content-based, collaborative and hybrid filtering methods. These systems are increasingly applied on e-commerce platforms, especially on the sites connected with news, music, books and different products. Initially, these systems depend on demographic data, but it is anticipated that these methods will use personal, and local data from the internet. These papers provide a comprehensive examination of each recommender systems.

Russo Spena et al. (2021) emphasised the growing role of analytics in segmentation and customer profiling. The author specifically mentioned the necessity of understanding how companies can identify the role of segmentation in the broader process of implementing consumer insights. The growing use of analytics indicates that organisations are improving their understanding of consumer preferences and are adjusting their operations and marketing strategies to accommodate them. Dynamic data and analytics tools are essential for matching the right users or customers to the right outcomes based on the level of personalisation provided. Furthermore, Granato et al. (2018) demonstrated that machine learning algorithms such as the principal component analysis (PCA) and hierarchical cluster analysis (HCA) are commonly used to examine similarities and identify hidden patterns among samples, especially when the relationships and groupings within the data are not yet evident. Meanwhile, Xue et al. (2011) also suggested that PCA is a technique used to integrate the variables of a big dataset in a way that emphasises the most important factors. By doing so, the first few variables in the reconstructed dataset account for the majority of the variability in the data.

2.2 Gaps in the Literature

Although there have been notable improvements in customer segmentation and product recommendation systems, the literature still has unresolved gaps. One notable gap is the lack of real-time data integration in user segmentation methods. Most studies and research are heavily based on past/historical data, which might not accurately represent current customer trends, behaviours, and consumer preferences (Jabbar et al., 2019). Real-time data analytic systems can provide more immediate and actionable insights, but they are difficult to implement due to technological and practical issues. Another major issue that has caused a lot of debate is the proper handling of customer data. There is a growing debate regarding the most effective way to balance the advantages of personalised recommendations with the need to protect customers' privacy and data security. While this personalised marketing can greatly improve the consumer experience, many researchers and experts argue that it also creates issues with consent, data ownership and the potential misuse of sensitive information (Edquist et al., 2022; Boerman et al., 2021). Hybrid recommendation models also cause controversy regarding their effectiveness, diversity and scalability. Although these models combine the best features of several algorithms to generate more accurate recommendations, their computational demands, complexity and vulnerability to cyber-attacks can be a challenge, particularly for smaller companies with limited resources (Thorat et al., 2015; Afoudi et al., 2021). Additionally, the majority of research focuses on immediate sales improvements, which results in a lack of understanding the long-term impacts of personalised recommendations on consumer relationships. Furthermore, the literature and research emphasise the difference in the performance of several clustering algorithms under different settings. Although K-means clustering is generally applied for its simplicity and efficiency, its performance can be greatly influenced by the initial centroid selection and the presence of outliers. The introduction of new algorithms, such as the one that combines K-means and K-medoids, is promising, but it also shows that there is no one-size-fits-all answer, demanding additional research into optimising these methods for different situations (Shah & Singh, 2012). Lastly, the subject of technographic segmentation is still developing, with few comprehensive studies exploring its full potential. The rapid evolution of technology usage patterns demands the development of more adaptable and dynamic segmentation methods that can keep up with these changes (FasterCapital, 2024). The existing literature frequently fails to provide comprehensive frameworks that address customers' ever-changing digital behaviours. To summarise, additional exploration and innovation are required in these mentioned areas.

2.3 Key Theories & Concepts

Traditional customer segmentation methods, such as demographic segmentation, divide customers by age, gender, income, education, and other basic characteristics. Although this strategy provides a general understanding, it usually fails to capture complex consumer behaviour fully. Psychographic segmentation is a more thorough study of consumers' lifestyles, values, and personality traits to provide more detailed insights. However, this approach requires significant effort to gather a lot of data (Ahlm et al., 2007; Gomes & Meisen, 2023). On the other hand, advanced data-driven methods such as behavioural segmentation analyse consumer purchasing habits, brand interactions, product usage and engagement levels to provide a more accurate representation of customer behaviours (Otebele, 2023; Susilo, 2016). Furthermore, understanding digital behaviours in today's technologically driven market depends significantly on the technographic segmentation approach, which categorises consumers based on their technology usage and preferences (Bodhe, 2023; Gulhane et al., 2024). For product recommendation systems, a wide range of methods are used to improve the shopping experience by recommending items to customers based on various factors. Content-based filtering utilises the analysis of attributes of items that a user has previously shown interest in to recommend similar products. This approach is especially efficient for specialised products. However, it has limitations as it does not consider consumer preferences beyond known items (Murel & Kavlakoglu, 2024b; Thorat et al., 2015). On the other hand, collaborative filtering includes both user-based and item-based approaches. User-based collaborative filtering is a recommendation system that recommends products by analysing the preferences of users or customers who are similar to each other. However, this approach has problems such as the cold start problem and data sparsity. Item-based collaborative filtering is a scalable approach that compares items to identify similarities but is less personalised (Evelyn, 2023; Murel & Kavlakoglu, 2024a). On the other hand, hybrid model filtering combines content-based and collaborative filtering to overcome the limitations of each method. This leads to recommendations that are more relevant and accurate (Afoudi et al., 2021). Clustering algorithms and dimensionality reduction methods are essential components of machine learning for marketing, as they enhance data analysis. K-means clustering is a widely used method for customer segmentation that effectively identifies distinct groups within extensive datasets. Hierarchical clustering, a method that makes nested clusters, is beneficial for visualising complex data structures (Yusuf, 2023). Dimensionality reduction methods, such as Principal Component Analysis (PCA), help to minimise and reduce data complexity while maintaining variance. By enhancing the effectiveness of clustering algorithms, it is possible to increase the efficiency and depth of data analysis (IBM, 2023).

2.4 Theoretical Framework

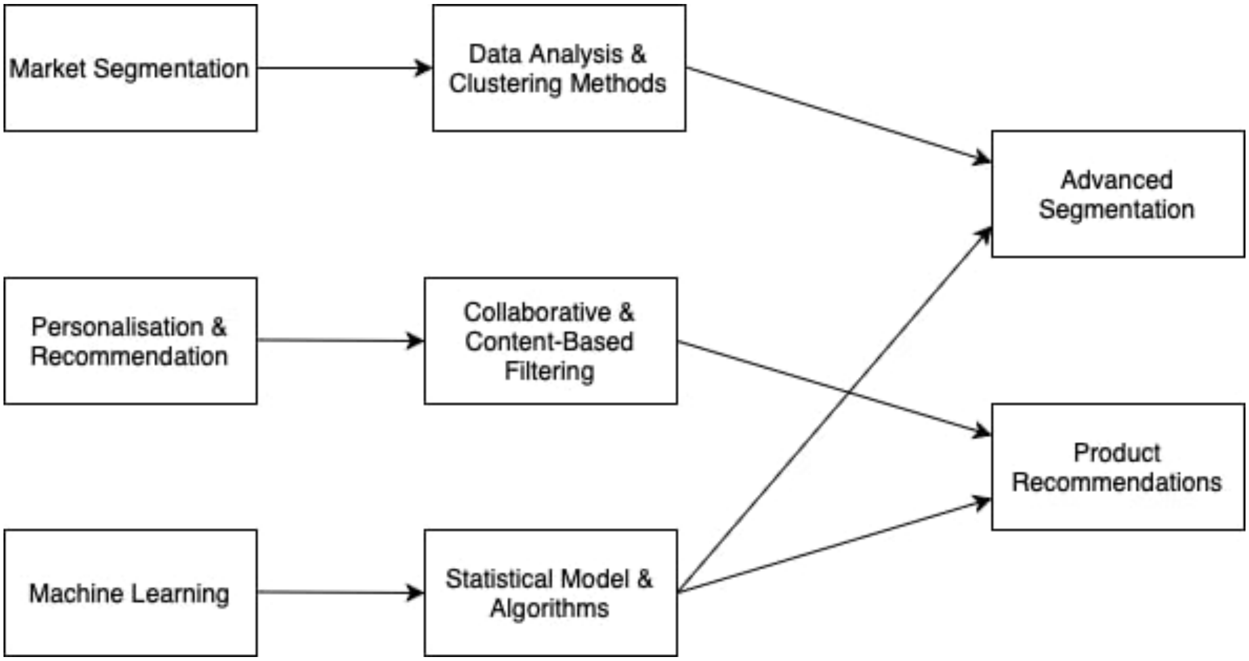

Market Segmentation
Data Analysis & Clustering Methods
Advanced Segmentation
Personalisation & Recommendation
Collaborative & Content-Based Filtering
Product Recommendations
Machine Learning
Statistical Model & Algorithms

Fig. 1. Theoretical Framework (Author's own)

The research is grounded in a theoretical framework (**figure 1**) that includes major marketing and machine learning theories and concepts. Market segmentation is the process of dividing a market into smaller segments based on customers' various needs, traits, or behaviours (Yi, 2018). This theory supports the effective segmentation obtained by using data analytics and clustering methods. Secondly, as the personalisation and recommendation theory implies that tailored recommendations based on user data can significantly increase consumer satisfaction and engagement, it encourages the use of collaborative and content-based filtering approaches in recommendation systems (Thorat et al., 2015). Finally, machine learning theory, including PCA, uses statistical models and algorithms to analyse and interpret large datasets. This lays the foundation for developing advanced segmentation and recommendation models that dynamically change and enhance performance over time (IBM, 2023). By integrating these theories, this study aims to develop a comprehensive approach to consumer segmentation and product recommendations using advanced data analytics and machine learning techniques.

3. Methodology

3.1 Research Design

This study's research design employs a comprehensive quantitative approach, utilising large datasets from the UK e-commerce sector to analyse consumer behaviours and preferences. This approach is chosen for its capacity to manage and analyse large amounts of datasets, offering unbiased and measurable insights into the effectiveness of personalised product recommendations and customer segmentation (Savela, 2018). The study focuses on the utilisation of advanced data analytics and machine learning methods to achieve these objectives and goals. The research is categorised into four major phases: data collection, reprocessing, analysis, and model evaluation. Every phase is intended to build upon the previous one, making sure that there is a seamless transition from raw data to actionable insights. The first step is to acquire a large number of secondary datasets from reliable sources which are then thoroughly preprocessed to verify data quality and sustainability for analysis. The following phases include using clustering techniques for customer segmentation and developing recommendation models to deliver personalised product recommendations. Finally, the models' performance is assessed using robust metrics to ensure their reliability and effectiveness.

3.1.1 Justification For Chosen Methods

The research objectives necessitate the analysis of extensive datasets to identify patterns and trends in customer behaviour, which justifies the utilisation of quantitative methods. This method enables the implementation of machine learning and statistical analysis, which leads to a high level of accuracy and objectivity. Moreover, the quantitative method is employed to summarise, make predictions and generalise findings to broader groups (Rana et al., 2021). Therefore, through the use of quantitative methodologies, the research has the capacity to offer generalizable insights that can be applied to the entire UK e-commerce industry. Data collection primarily focuses on secondary sources, such as customer data, product data, interacting data and existing market research reports. These sources provide extensive datasets that include customer demographics, purchasing histories, device preferences and conversion rates. The availability, reliability and richness justify the utilisation of secondary data, as it provides a cost-effective way to collect large datasets required for machine learning applications (University College London, 2022). Data preprocessing is an essential procedure that guarantees the quality and usability of data that has been collected. This process includes four primary tasks: data cleansing (which involves the removal of duplicates and the correction of errors), data integration (which includes the merging of various sources into a single dataset), dimensionality reduction (which involves the reduction of the volume of the data) and data transformation (which involves the normalisation and scaling of numerical data) (Anunaya, 2021). These steps/tasks are critical in preparing the data for accurate and efficient analysis. Ensuring high-quality data is crucial for the reliability of following analytical processes.

On the other hand, clustering which is the fundamental component of data analysis, plays an important role. On one hand, the information increase and subject intersection have resulted in the development of numerous cluster analysis tools. Conversely, it is important to note that each clustering method has unique advantages and disadvantages due to the complexity of data that obtained. For example, while hierarchical clustering offers good scalability and a more comprehensive representation of data structures, K-means clustering is particularly well-suited for its computational efficiency and low time complexity (Xu & Tian, 2015). The development of targeted marketing strategies depends on these algorithms since they assist in categorising consumers based on common characteristics. On top of that, the justification of recommendation models is based on their proven effectiveness in personalising user experiences, thereby increasing customer satisfaction and engagement. The content-based filtering, which depends on user's previous choices, heavily relies on item descriptions while collaborative filtering (both item-based and user-based) utilises the preferences of similar users to provide recommendations. The hybrid filtering model combines these two methods to overcome the limitations and increase recommendation accuracy (Thorat et al., 2015). The employment of machine learning techniques, such as Principal Component Analysis (PCA), which lowers the number of dimensions, is justified by their ability to visually represent and explore datasets with a large number of dimensions or characteristics, as well as their ability to easily detect and analyse trends, patterns, and outliers (IBM, 2023). This can increase the performance of clustering algorithms and recommendation models. Machine learning techniques are important for adapting and improving models over time, ensuring they remain effective as new data is collected.

3.1.2 Ethical Considerations

Ethical considerations are of the utmost importance in this research to ensure that data protection regulations are followed and customers' trust and privacy are protected. (Halej, 2017; Rana, Dilshad, et al., 2021) stated that the management of consumer data poses substantial privacy concerns and has the potential to result in the misuse of sensitive data, such as personal identifiers, browsing and purchase histories. In order to mitigate this, data anonymisation techniques will be implemented to eliminate any identifiable information, protecting the privacy of individuals. Furthermore, strict access control procedures will be implemented to guarantee that only authorised individuals have access to the datasets. To maintain legal compliance and protect customer confidentiality, it is crucial to comply with data protection laws and regulations, including the UK General Data Protection Regulation (GDPR). Moreover, the secondary data collection process is specifically designed to reduce bias by directly collecting data from government agencies and reputable international organisations, such as GlobalData Explorer, IBIS World, ScienceDirect, Statista, Kaggle, ECDB (E-CommerceDB) and so on. In order to prevent any potential conflicts that could impact the research purpose or the implementation of results, any personal or financial interests are strictly avoided. On top of that, this project maintains transparency about the methodologies and algorithms used for the research. As previously stated, secondary data will be collected and analysed only from reliable sources. However, by any chance, if individual data collection is necessary, the study will implement strict measures to protect privacy and confidentiality. Ethical approval and informed consent will be obtained, and data will be handled exclusively, stored and analysed for research purposes.

3.2 Data Collection

The data collection approach for this research is carefully planned to obtain thorough and relevant datasets from many reliable secondary sources. The datasets will cover several factors of customer behaviour and interactions in the UK e-commerce industry, offering a strong basis for further analysis and model building.

3.2.1 Sources of Secondary Data

The main sources of data comprise customer transaction data, website analytics (social media listening tools), market research reports, customer demographics, and technology usage data. Customer transaction data provides information on the types of products purchased, quantities, discount & return rates and payment methods. These records are important for comprehending buying trends and identifying customers’ preferences and behaviours, which can help organisations and companies develop better marketing strategies (Hayden, 2022). Website analytics obtained from tracking tools offer valuable insights on the conversion KPIs such as add-to-cart rate, cart abandonment rate and conversion rate which are essential for behavioural segmentation. On top of that, market research reports from reliable sources and businesses provide contextual information on industry trends (present & future), customer demographics, market size, structure and revenue forecast insights which supplement behavioural data to provide a thorough picture of the market landscape (Yallop et al., 2022). Additionally, customer demographics data such as age, gender, income, and online behaviours, help in dividing customers into separate groups which in turn helps organisations to make targeted marketing strategies. Finally, technology usage data regarding customers’ devices and payment methods is crucial for technographic segmentation and enhancing the user experience on various devices.

3.2.2 Data Collection Process

The process of data collection involves multiple steps aimed at ensuring the relevance and quality of the obtained data. First, the process of identifying data sources involves collaborating with e-commerce websites and companies, analytics firms and market research organisations to access relevant datasets. This guarantees that the chosen data sources are trustworthy, reliable and up-to-date. Following this, the collected datasets are checked to verify compliance with legal and ethical standards and to extract data from different platforms and databases using data export & analysis tools. Finally, data quality assurance is performed by conducting initial evaluations to determine the extent to which the data is complete and accurate. This process involves identifying and correcting any inconsistencies or gaps in the data to ensure the dataset’s quality and reliability (Kindling & Strecker, 2022).

3.3 Data Preprocessing

(Agarwal, 2015; Anunaya, 2021) demonstrated that data preprocessing is the process of converting raw data into a comprehensible format. Real-world data can often be incomplete, inconsistent, redundant, and noisy. Data preprocessing includes a range of procedures that facilitate the transformation of raw data into a refined and sensible format. The procedure consists of four main tasks: data cleaning, data integration, dimensionality reduction and data transformation (**Figure 2**).

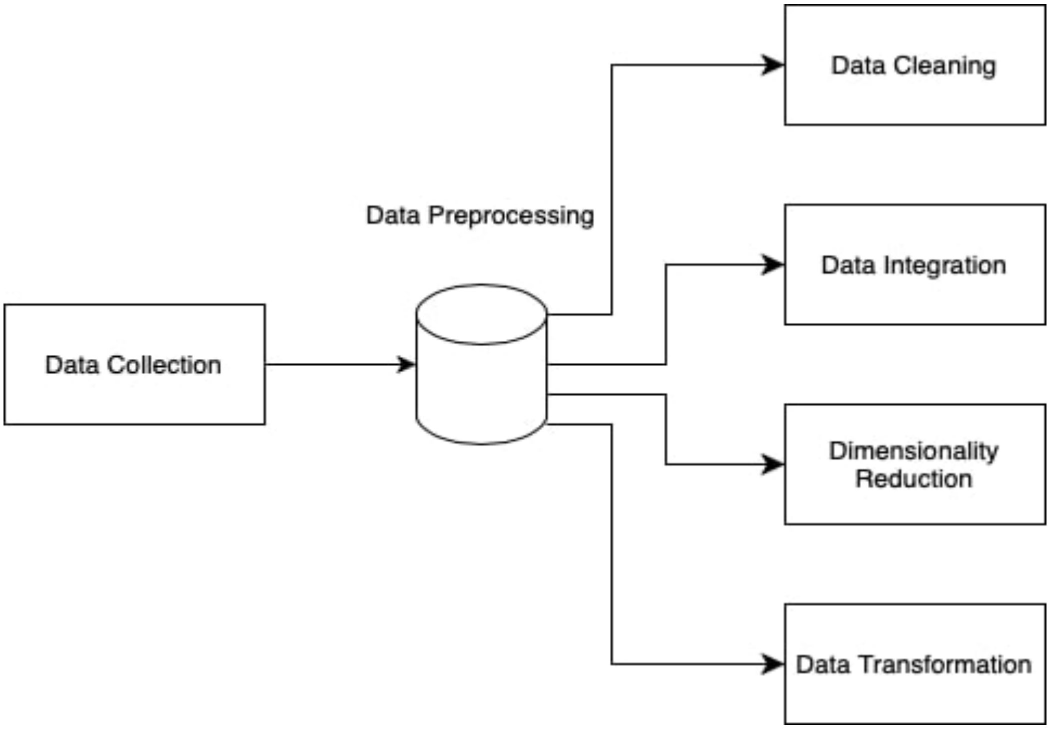


Fig.2. Data Preprocessing Process (Author’s own)

3.3.1 Data Cleaning

Maharana et al. (2022) stated that data cleaning is the first process of data preprocessing. It is used to identify inaccurate or noisy data and either correct or eliminate it from the dataset. Duplicate records, missing values and incorrect information are among the most common issues that must be addressed. Duplicate or missing data records are identified and eliminated as they have the potential to affect the results of the analysis. Suppose there is a moderate amount of missing values, basic interpolation methods are used to address this issue. The most common approach employed for dealing with it is utilising mean, median or mode values. Accuracy and consistency of data are essential at this stage to ensure the reliability of analysis results.

3.3.2 Data Integration

IBM (n.d.) argued that data integration is the process of merging and harmonising data from numerous sources into a unified and consistent format, which may then be utilised for various analytical, operational, and decision-making purposes. This step is essential to developing a full understanding of the data. Datasets retrieved from the customer transaction records, website analysis, market research reports, customer demographics and technology use data are then merged into one dataset. This process not only increases the dataset's depth but also offers a comprehensive perspective that is crucial for thorough analysis and model development.

3.3.3 Dimensionality Reduction

Dimensionality reduction methods are used to minimise the volume of data while preserving its essential information. Dimensionality reduction methods, such as Principal Component Analysis (PCA), are used to reduce the complexity of the dataset. Data reduction techniques provide benefits such as improving the efficiency of data. After the reduction is accomplished, the data becomes more easily used for artificial intelligence (AI) approaches to employ in many ways, such as sophisticated data analytics applications, hence enhancing the efficiency of clustering and recommendation algorithms (Powell & Smalley, 2024).

3.3.4 Data Transformation

Data transformation is an important part of data preprocessing, wherein raw data is turned into a unified format or structure. This process guarantees that data is compatible with systems and improves its quality and usability. Data normalisation, one of the data transformation types, adjusts the data to a uniform scale, such as numerical values from 1 to 10, while preserving the relative differences in the range of values. This is especially crucial for algorithms that depend on distance metrics like clustering. Moreover, data encoding is the process of transforming category data into numerical representation, making it easier to analyse. For example, data encoding could allocate a unique number to each data category (Hayes & Downie, 2024).

3.4 Machine Learning Techniques

In the context of customer segmentation and product recommendations, machine learning (ML) techniques are critical for assessing complicated datasets and obtaining actionable insights. This section addresses the ML approaches used in this research, focussing on clustering algorithms for customer segmentation and recommendation algorithms for personalised product recommendations.

3.4.1 Clustering for Customer Segmentation

Sharma (2019) and Xu & Tian (2015) expressed that customers/users are classified into distinct categories through the use of clustering, a fundamental approach in machine learning. It facilitates the execution of personalised services and targeted marketing strategies by classifying customers based on their shared characteristics and attributes. K-means is a widely used clustering method that divides data into k clusters based on the similarity of data points. The technique progressively improves the centroids of each cluster in order to lower the variance within each cluster. This research utilises the following stages to implement K-means:

- Employ the Elbow Method & calculate the silhouette score using sklearn to determine K.
- Choose the optimal number of clusters, K.
- Distribute clusters based on the results of the above analysis.
- Do the cluster analysis

This study may or may not employ hierarchical agglomerative clustering (HAC), also known as the bottom-up approach, if K-means clustering can’t provide the expected results. In this approach, the number of clusters does not need to be predetermined by the clustering algorithm. Bottom-up algorithms initially consider each data point as a separate cluster and then gradually combine pairs of clusters until all clusters are merged into a single cluster that includes all the data. This method is implemented as follows:

- Consider each data point a single cluster and determine the distance between it and the other clusters.
- Similar clusters are combined to create a single cluster. Assuming that cluster (B) and cluster (C) are extremely similar to each other, these clusters are combined in this stage, just as we do with clusters (D) and (E). Then, we obtain the clusters [(A), (BC), (DE), (F)]. The recalculation of the proximity based on the algorithm is employed, which in turn combines the two closest clusters. Continue this process until only one cluster remains. (Figure 4)
- Create a dendrogram to visualise the hierarchical structure of the clusters.
- Divide the dendrogram at the desired level to create k clusters (Dey, 2019).

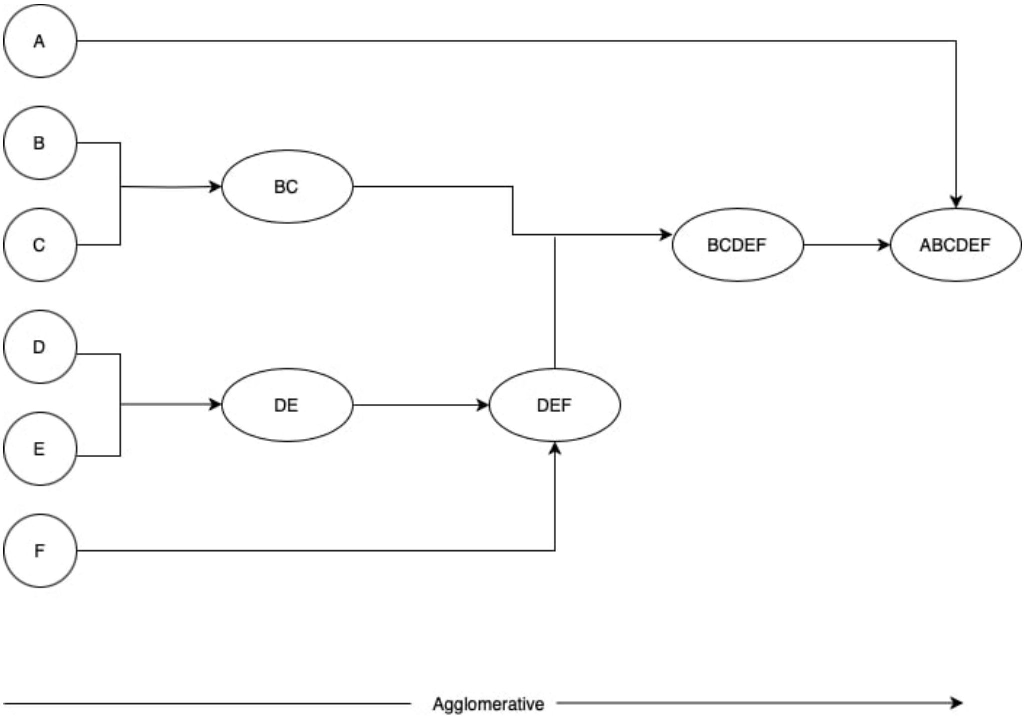


Fig.3. Hierarchical Agglomerative Clustering *This figure illustrates the process of hierarchical agglomerative clustering, where groups are progressively merged based on their similarity until all data points are combined into a single cluster.*

3.4.2 Algorithms for Product Recommendations

Kimnaruk (2022) and Murel & Kavlakoglu (2024b) claimed that recommendation systems leverage machine learning algorithms to provide users with product recommendations that are tailored to their preferences and behaviours. This is how content-based filtering approach is implemented:

- Extract relevant attributes from products, including keywords, brand and category.
- Build a user profile by utilising the characteristics of items that have been previously liked or bought.
- Analyse the similarity between newly introduced items and the user profile.
- Recommend items based on the analysis.

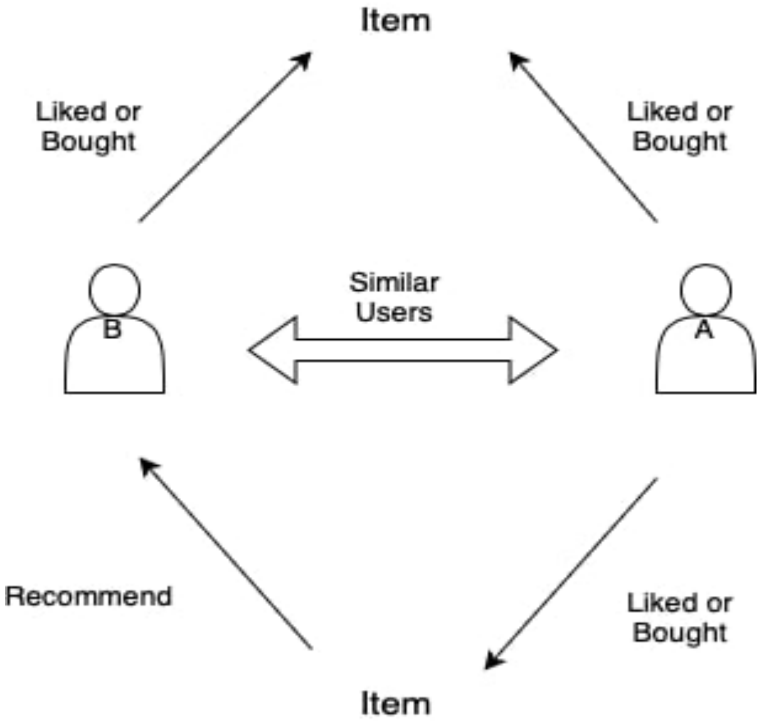


Fig.4. Content-Based Filtering *This figure shows the process of content-based filtering, emphasising the extraction of relevant characteristics from items and matching them with user profiles to generate personalised recommendations.*

Collaborative Filtering (User-Based and item-based) generates product recommendations by analysing the preferences of similar users and items. The following are the stages involved in the implementation process.

- Create a user-item/item-item interaction matrix.
- Analyse the similarity between users/items.
- Recommend users/items that have previously interacted with (Evelyn, 2023).

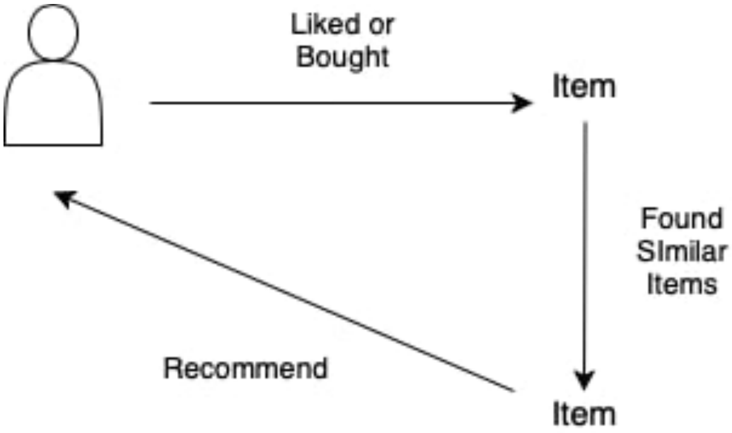


Fig.5. Collaborative Filtering *The figure illustrates the collaborative filtering method, which involves analysing user-item interaction matrices to generate suggestions by identifying similarities between both users and items.*

A hybrid recommendation, a combination of content-based and collaborative filtering techniques, is employed for a better recommendation approach.

3.5 Data Analysis Procedures

Data analysis processes are important for obtaining useful insights and ensuring the reliability of the developed models. The procedure starts with Exploratory Data Analysis (EDA), which tries to summarise the dataset’s main characteristics and identify patterns, trends, anomalies, and correlations. In order to comprehend the distribution of variables, descriptive statistics, including mean, median, mode, standard deviation, and range, will be computed. Data visualisation tools such as Python, R and Excel will be employed to generate histograms, scatter plots, box plots, and correlation matrices to figure out the trends and patterns. Furthermore, the extent and patterns of missing data will be assessed through missing value analysis to determine appropriate handling strategies (IBM, 2020; US EPA, 2015).

4. Data Analysis and Results

4.1 Overview of Collected Data

The data for this research was obtained from reliable secondary sources, which include customer transaction data, website analytics, market research studies, customer demographics, and technology usage data. This overview offers a concise examination of the data characteristics and key statistics that serve as the foundation for further research. Further analysis of results and findings will be thoroughly discussed in the subsequent section.

4.1.1 Summary of Data Characteristics, Key Insights and Initial Findings

A wide variety of consumer interactions and behaviours within the UK e-commerce sector are captured in the customer transaction data sources. Detailed insights such as quantities of products purchased, popular products among online shoppers, transactional KPIs (discount & return rates) and reasons to shop online. The data helps in the identification of customer spending habits and purchasing patterns. On average, 30% and 25% of online shoppers purchase groceries, fresh food & beverages on a weekly basis, respectively, while electronics and home goods are purchased less frequently (Statista, 2023a). The average order value (AOV) net increased from 86.2$ in 2020 to 93.4$ in 2023, whereas the discount rate rose from 10.2% to 11.0%. A return rate remains consistent at approximately 14%. On the other hand, the average returned value increased from 13.9$ to 15.2$ (**Table 1**) (ECDB, 2023).

Table 1 Transactional KPIs of the E-commerce Sector from 2020 to 2023 (UK) retrieved from ECDB

Transactional KPIs	2020	2021	2022	2023
AOV Net (in $)	86.2	89.7	87.7	93.4
Discount Rate (%)	10.2	10.8	10.9	11.0
Avg. Discount Rate (in $)	11.3	12.6	12.4	13.5
Return Rate (%)	13.9	13.7	14.0	14.0
Avg. Return Rate (in $)	13.9	14.3	14.3	15.2

The website analytics data includes metrics such as add-to-cart rate, cart abandonment rate and conversion rate. The percentage of adding items to the shopping cart increased from 11.0% in 2020 to 11.5% in 2023. Nevertheless, there was a rise in the cart abandonment rate, which went from 69.9% to 74.2%. Simultaneously, the conversion rate observed a decline from 3.3% to 3.0% (**Table 2**) (ECDB, 2023). The conversion rates of desktop users decreased by half, dropping from 20% in 2021 to 10% in 2022. The percentage of mobile phone users' conversion rate decreased from 11% to 7% in 2022 and for tablet users, the rate declined from 15% to 8% (Chevalier, 2024). On the other hand, in 2023, 77% of mobile baskets, 71% of tablet baskets and 68% of computer baskets did not result in completed orders in the UK (Statista, 2024).

Table 2 Cart Conversion KPIs of the E-commerce Sector from 2020 to 2023 (UK) retrieved from ECDB

KPIs	2020	2021	2022	2023
Add-to-cart Rate (%)	11.0	10.7	10.8	11.5
Cart Abandonment (%)	69.9	69.5	72.1	74.2
Conversion Rate (%)	3.3	3.3	3.0	3.0

Market research reports provide information on industry trends (present & future), market size and revenue forecast. AmazonUK dominated the UK e-commerce market in 2023, with a net sales volume of $16.1 billion. Tesco was ranked third with $6.4 billion, while Sainsbury's followed with $6.6 billion. Apple was rated tenth with a revenue of $2.3 billion (Coppola, 2024a). Furthermore, it is anticipated that the e-commerce sector will expand, with the number of e-commerce users in the United Kingdom predicted to rise from 61 million in 2024 to 62.1 million in 2025 (Statista, 2021). Furthermore, it is forecasted that the revenue of this sector will substantially increase from $160.6 billion in 2024 to $194.1 billion in 2027 (Statista, 2022). Data regarding customer demographics includes age, gender, income, and online behaviours. Females comprise the majority of online consumers, as indicated by (Statista, 2023c), whereas the largest age group is within the 30-49 age bracket. Looking at income, the majority of individuals have an annual household income between £22,800 and £43,200. Technology usage data offers valuable insights into the device preferences and payment methods used for shopping. In the UK, digital wallets, credit cards and debit cards are the most used payment methods in e-commerce, making up a combined 81% of all transactions (Coppola, 2024b). The younger generations, specifically Gen Z and Millennials, show a strong preference for mobile phones, pointing out the need for e-commerce platforms that are designed for mobile devices. Tablets are generally the least favoured devices among all age groups, yet Baby Boomers use them comparatively more than other generations (Statista, 2023b).

4.2 Results of Customer Segmentation

4.2.1 Optimal K Selection and Cluster Analysis

Rushirajsinh (2023) stated that Accurate client segmentation depends on optimal cluster (K) numbers. The optimal number of K clusters is acquired in this study by employing the frequently used cluster analysis technique, the elbow method. It functions by calculating the Within-Cluster Sum of Squares (WCSS), which is the total of the squared distances separating every data point from its corresponding cluster centre. By examining the chart or plot, we search for a point where the rate of decline sharply decelerates, resulting in the formation of a distinct ‘elbow’ shape. This point shows a balance between minimising inertia and avoiding overfitting by having too many clusters.

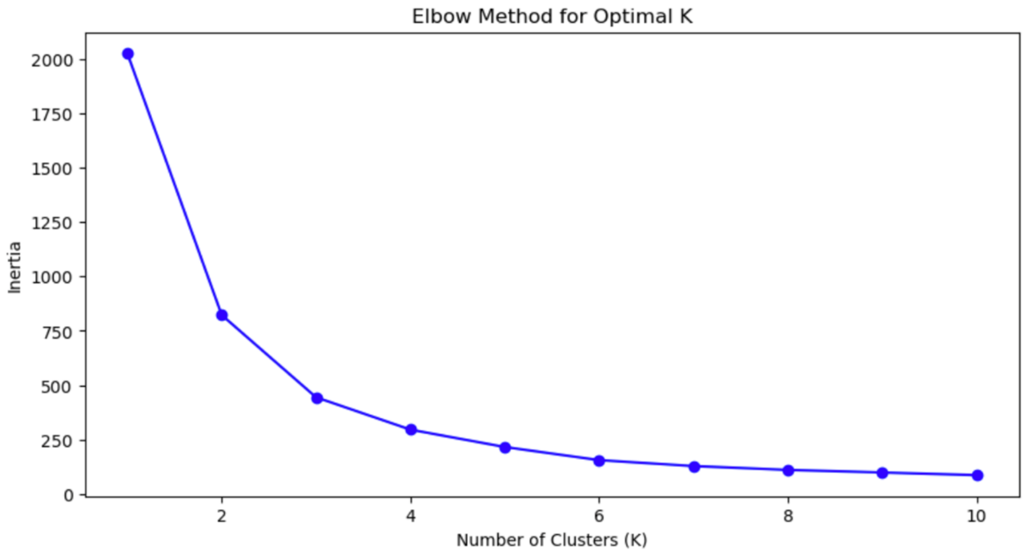


Fig.6. Elbow Method Graph for Cluster Analysis (Author’s own) *This figure shows the optimal K selection, employing Elbow Method*

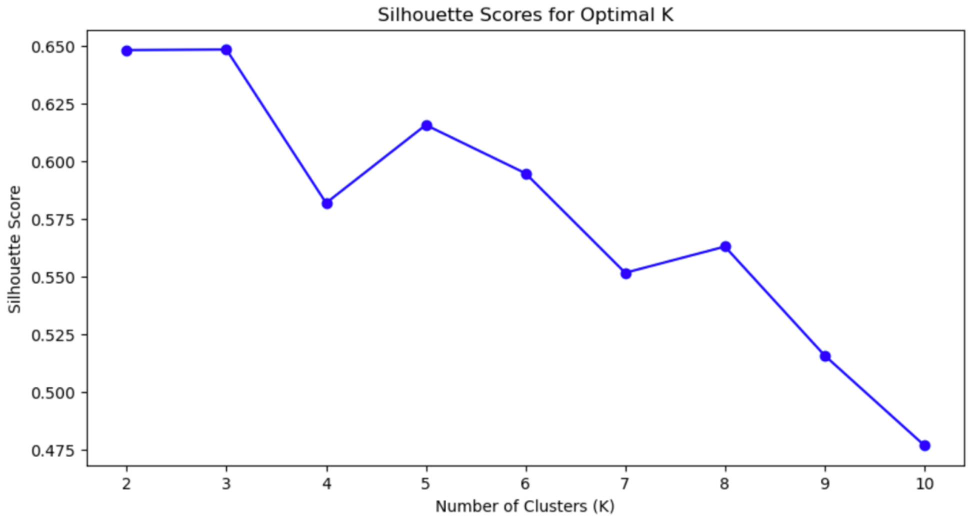


Fig.7. Silhouette Scores for Optimal K (Author’s own)

The elbow method in the analysis demonstrates a distinct bend at K=3, indicating that three clusters provide an optimal equilibrium between performance and complexity (**Figure 7**). On top of that, Gültekin (2023) examined that the silhouette scores, which measure the similarity of each data point to its respective cluster and quantify its difference from other clusters, are relatively high at K=3 (**Figure 8**). These metrics indicate that selecting K=3 is the most suitable option for segmenting the research customer dataset.

Each step of the cluster analysis using Python can be found in Appendix A. Please also note that these demographic data are extracted from Statista (2023c).

4.2.2 Description of Identified Segments

Table 3 Analysed Clusters (Gender & Age)

Clusters	Total	Female	Male	Age 18-29	Age 30-49	Age 50+
1	47.64	25.78	21.86	12.49	22.29	12.86
2	449.57	180.04	148.63	62.39	144.88	122.39
3	890.2	474.2	416	183.6	367.6	339

Table 4 Analysed Clusters (Income)

Clusters	Income <= £ 22,800	Income £ 22,800- 43,200	Income £ 43,200- 98,400	Income > £ 98,400
1	11.56	13.76	16.24	3.06
2	87.65	102.71	102.02	13.34
3	267	272.8	244.8	32

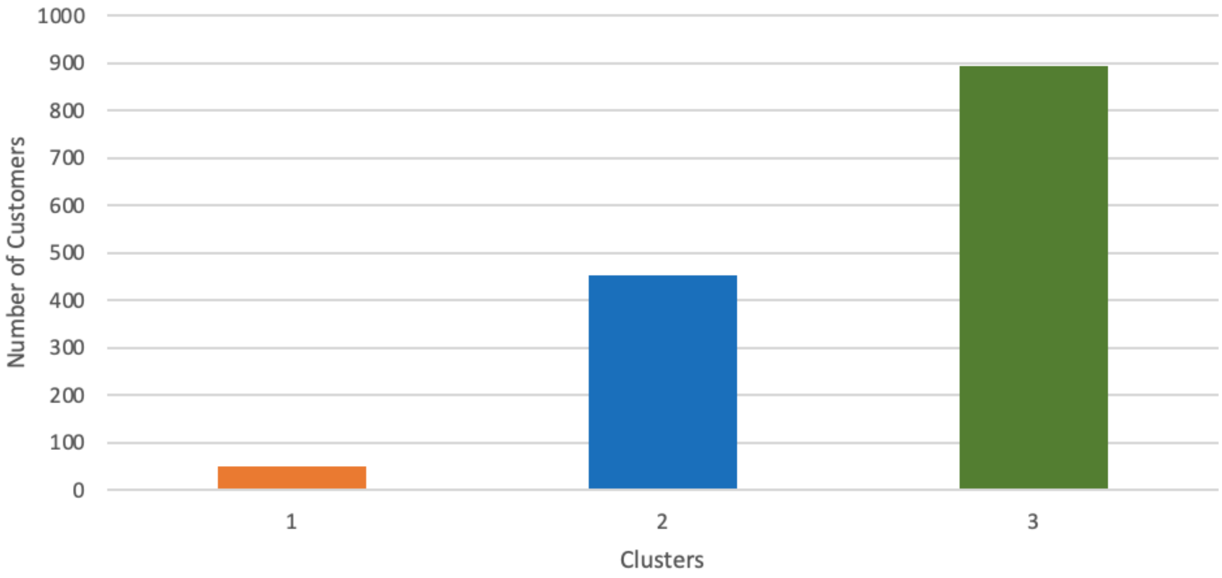


Fig.8. Distribution of Clusters (Author’s own)

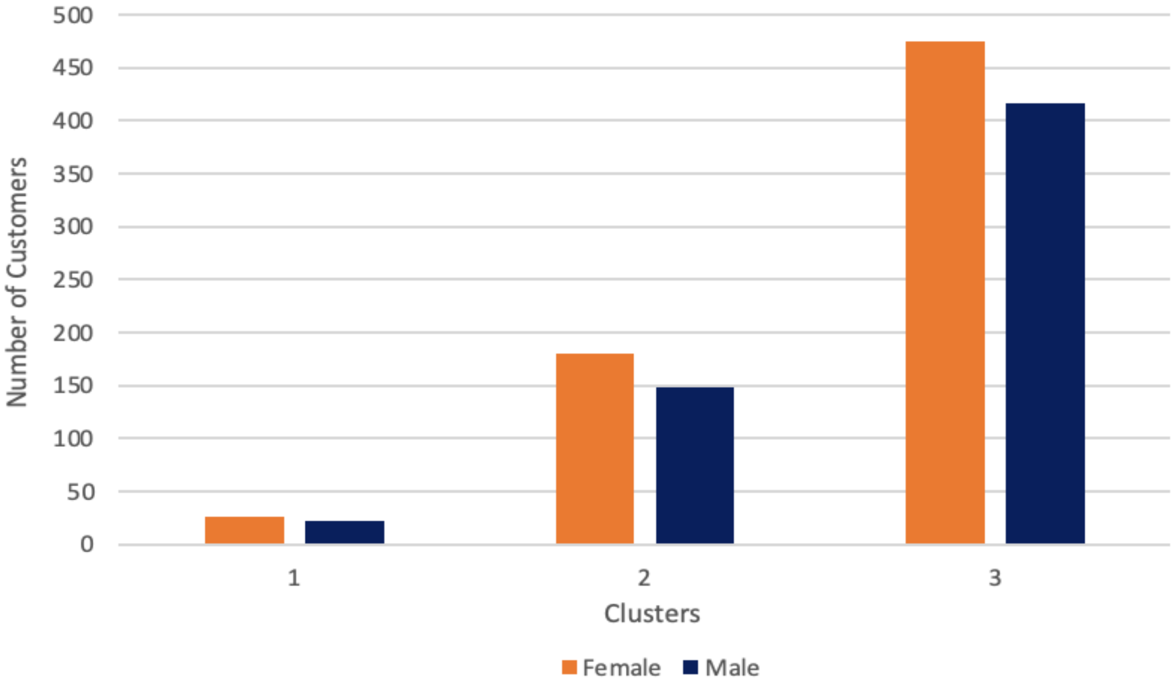


Fig.9. Gender Distribution Across Clusters (Author’s own)

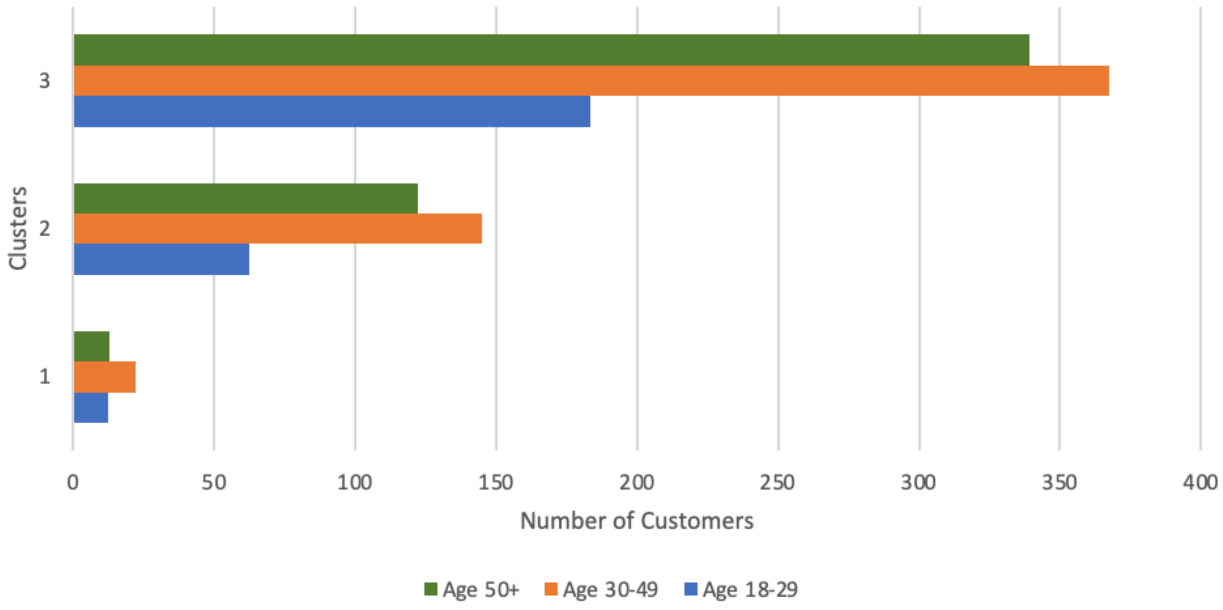


Fig.10. Age Distribution Across Clusters (Author’s own)

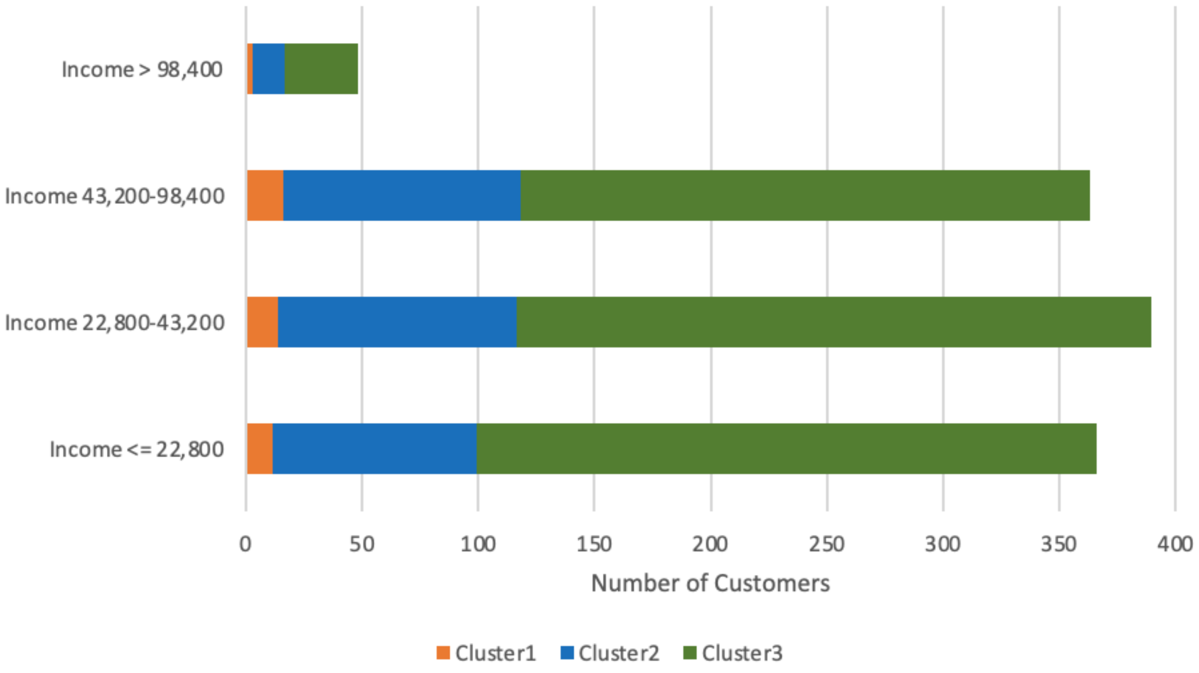


Fig.11. Income Distribution Across Clusters (Author’s own)

Cluster (Segment) 1: This cluster includes a smaller segment of customers, with an average total count of 47.6. It is composed of an average of 21.9 males and 25.8 females in terms of demographics. The age distribution suggests that the customer base is young, as demonstrated by the significant representation in the 18-29 age group (12.5) and moderate representation in the 30-49 age group (22.3) (**Table 3**). The income levels in this cluster are comparatively low, with a significant number of customers earning less than £22,800. Only a small subset of consumers can be classified as high-income (Statista, 2023c) (**Table 4**). In this cluster, there is an attraction for budget-friendly products, with a probable preference for essential goods and items of high utility. This segment can be effectively engaged through marketing strategies that include value-for-money offers, promotions, and discounts. To summarise, this cluster comprises relatively young customers with limited incomes who are drawn to products that are affordably priced.

Cluster (Segment) 2: With an average of 449.6 customers, this group comprises of a rather small population. Within this cluster, the female proportion (180.04) exceeds the male count (148.63) (**Table 3**). There is a broad age distribution, with significant representation in the 18-29 and 30-49 ranges. Consumers in this cluster have a higher income than those in cluster 1, with a significant proportion falling between £43,200 and £98,400 (Statista, 2023c) (**Table 4).** This cluster is expected to spend on luxury goods, health products, and designer clothes. Marketing strategies that emphasise loyalty programs, and high-quality content marketing can increase the engagement of this cluster/segment. On top of that, based on their behaviours and demographics, it is essential to prioritise the quality of the product and the reputation of the brand in order to attract customers in this segment.

Cluster (Segment) 3: This group is the largest, with an average of 890.2 customers in total. The cluster shows a greater proportion of females (474.2) in comparison to males (416) within the age categories of 30-49 and 50+ (**Table 3**). Customers in this cluster have the highest levels of income in comparison to other clusters, with a presentation in the income brackets of £43,200-£98,400 and >£98,400 (Statista, 2023c) (**Table 4**). This demographic is more inclined to buy luxury items such as designer clothing, real estate, and premium technology. They value exclusivity, greater quality, and the prestige associated with a brand. Furthermore, offering customised experiences and services can increase customer satisfaction and loyalty.

4.3 Results of Product Recommendation Models

In this analysis, three different recommendation models were implemented: content-based filtering, item-based collaborative filtering, and hybrid filtering. The results, implications of each model and performance comparison are detailed in the following sections. Every step of the models' implementation can be found in **Appendix B**.

4.3.1 Results & Implication of Content-Based Filtering

Content-based filtering provides suggestions based on product attributes, such as descriptions. Using TF-IDF (Term Frequency-Inverse Document Frequency), we transformed the product descriptions into a numerical matrix that shows the significance of the terms in each description (Karabiber, n.d.). Following this, the similarity between products was assessed using cosine similarity. For example, if "Set of 3 Pantry Wooden Spoons" (ProductID: 23495) is chosen, the content-based filtering model will recommend the following products:

1. Set Of 3 Cake Tins Pantry Design
2. Small Heart Measuring Spoons
3. Long Heart Measuring Spoons
4. Set 2 Pantry Design Tea Towels
5. Set of 3 Wooden Stocking Decoration
6. Set of 60 Pantry Design Cake Cases
7. Set of 3 Wooden Tree Decorations
8. Set of 4 Pantry Jelly Moulds

9. Set of 3 Wooden Sleigh Decorations
10. Wooden School Colouring Set (**Appendix B**) (Ramos, 2022)

All of these recommendations are in line with the selected product in terms of category (cooking supplies and decorations) and material (wood). This demonstrates how well the model works to reliably identify products with comparable features. Murel & Kavlakoglu (2024b) argued that content-based filtering is especially beneficial for suggesting products that share similarities with the ones the consumer has already expressed interest in. This model performs exceptionally well in areas where product features are rich and descriptive. However, this model tends to suggest very similar products, which might restrict the range of recommendations.

4.3.2 Results & Implications of Item-based Collaborative Filtering

Item-based collaborative filtering suggests products by analysing the purchasing behaviours of other users. We generated a user-item interaction matrix and applied the Nearest Neighbours algorithm to detect goods that are frequently purchased together (Latifi et al., 2022). For the same example product that we have chosen in the previous section, the following products were recommended by the item-based collaborative filtering model:

1. Vintage 2 Metre Folding Ruler
2. Bundle of 3 Retro Note Books
3. Gymkhana Treasure Book Box
4. Vintage Leaf Magnetic Notepad
5. Sketchbook Magnetic Shopping List
6. Pantry Magnetic Shopping List
7. Star Wreath Decoration With Bell
8. Jumbo Bag Apples
9. Jumbo Bag Pears (**Appendix B**) (Ramos, 2022)

The purchasing behaviours of customers who purchased the example product are seen in the above recommendations, which contain a diverse selection of household items and gifts. This method captures associations between items that might not be clear through the content alone. Item-based collaborative filtering is capable of capturing trends and co-purchase patterns due to the actual transaction data usage. It is especially advantageous in situations where user behaviour serves as an essential indicator of product relevance. However, as mentioned in the “Key Theories & Concepts” section earlier, this method may face some problems with the cold-start problem when approaching new products that do not have sufficient data.

4.3.3 Results and Implications of Hybrid Filtering

Hybrid filtering combines both content-based and collaborative filtering to capitalise on the advantages of both approaches. This approach delivers the following product recommendations for the same example product by taking into account both user behaviour and product attributes:

1. Small Heart Measuring Spoons
2. Long Heart Measuring Spoons
3. Set 2 Pantry Design Tea Towels
4. Set of 3 Wooden Stocking Decoration
5. Set of 60 Pantry Design Cake Cases
6. Set of 3 Wooden Tree Decorations
7. Set of 4 Pantry Jelly Moulds
8. Set of 3 Wooden Sleigh Decorations
9. Wooden School Colouring Set (**Appendix B**) (Ramos, 2022)

This model recommends better and more suitable products in terms of their descriptions and the overall frequency of purchase. Hybrid filtering provides a balanced approach by ensuring that recommendations are both relevant and diverse. This model has the ability to overcome the limitations of other approaches, hence offering a more comprehensive recommendation system.

5. Discussion

5.1 Interpretation of Finding

The data analysis provides valuable insights into the changing behaviours of e-commerce consumers in the UK, which will be further examined in this discussion. We will explore the potential of recommendation systems to improve business performance and customer satisfaction, as well as the effectiveness of segmentation strategies in engaging and reaching various customer segments.

5.1.1 Customer Transaction and Behaviour Insights

The rising trend in Average Order Value (AOV) over the years shows that consumers are increasingly inclined to spend more per transaction. This may be because of a variety of factors, such as the convenience of online shopping or simply due to a wider selection of products. However, this positive trend has been somewhat offset by the steady increase in the discount rates, meaning that customers are still highly motivated by the sales and discounts, despite the fact they are spending more (**Table 1**). This behaviour shows that buyers adopt a dual perspective where they prioritise quality products as well as affordable rates. As a result, e-commerce businesses are under pressure to offer the best product together with competitive pricing strategies to meet the demand (Gao, 2023).

The decline in conversion rates and the increase in cart abandonment rates are particularly informative. These trends show that a significant number of consumers are not completing their purchases, despite the fact they are exploring and adding items to their shopping carts (**Table 2**). Sondhi (2017) suggested that this might be due to factors like indecision, complicated checkout processes, or unexpected costs like shipping fees or security concerns. It highlights the importance of e-commerce platforms to make the shopping experience smoother by simplifying the checkout process, clearly displaying all costs upfront, and possibly offering perks like free shipping or easy returns to reduce the chances of customers abandoning their purchases.

5.1.2 Customer Segmentation Insights

The segmentation analysis revealed the existence of three separate customer clusters, each showing its different features and behaviours. These offer a strategic framework for personalised service offerings and targeted marketing.

The demographic of cluster/segment 1 is primarily younger, has lower income levels, and is highly responsive to promotions and discounts. The importance of consistently providing sale promotions and maintaining competitive pricing is highlighted by this segment. However, businesses should consider long-term strategies to convert these price-sensitive customers into loyal consumers. By looking at their behaviours, this can be easily achieved by providing exclusive discounts or rewards for their repeated purchases (Rane et al., 2023).

Cluster 2 represents customers who prioritise features that save time and prefer a shopping experience that is not complicated. The willingness of this segment to pay for convenience along with their preference for high-quality products show that businesses can gain their trust by improving their service offerings. This includes the implementation of intuitive mobile shopping experiences, simple return processes, and faster delivery options. Marketing strategies should focus on quality, convenience and service over price as this segment is less price-sensitive (Rosário & Raimundo, 2021; Zott et al., 2000).

Cluster 3 represents a demographic of individuals with a substantial wealth who have notable preference for premium and luxury goods. The behaviour of this segment suggests a preference for brands that provide superior quality and exclusivity. The key to engaging this segment for businesses is to emphasise the unique qualities of their products and provide personalised purchasing experiences that cater to their high spending capacity and demand for quality (Zott et al., 2000).

5.1.3 Product Recommendation Insights

The hybrid model's performance reveals vital insights into how e-commerce platforms may improve customer engagement and satisfaction. By merging content-based filtering methods, the hybrid model demonstrates the effectiveness of a unified approach to recommendations. The hybrid model's capacity to combine product qualities with user behaviour results in precise and personalised suggestions, leading to enhanced consumer satisfaction and loyalty. This demonstrates the growing importance of personalisation in the e-commerce market. Businesses that are capable of providing relevant and timely recommendations are more likely to succeed, as consumers nowadays expect customised products or services. The value of investing in advanced, integrated recommendation systems has been shown by the hybrid model’s capacity to eliminate the limitations of single-method approaches, such as the cold start problem in collaborative filtering or the limited focus of content-based filtering (Dong et al., 2017).

5.2 Answering Research Questions

5.2.1 Research Question 1

“How can different customer groups be identified in the UK e-commerce market by effectively implementing customer segmentation?”

The above analysis successfully identified customer groups in the UK e-commerce sector by using machine learning methods such as clustering techniques, specifically K-means clustering. From that, we got three different customer groups through the use of the elbow method and silhouette scores to determine the optimal number of clusters. Based on the shopping behaviours and demographics including gender, age and income, these segments had been defied. For example, as we have discussed earlier, cluster/segment 1 has younger generations who love to buy products that are at reasonable prices. Moreover, they are attracted to the promotions and discounted products. On the other hand, cluster/segment 3 has been identified as high-income groups of people who enjoy personalised products and brand exclusivity. Due to the effective implementation of customer segmentation through clustering algorithms, which were tailored to the preferences of each group/cluster, e-commerce platforms can now employ personalised marketing strategies.

5.2.2 Research Question 2

“How can personalised product recommendations help improve engagement and sales?”

To provide personalised product recommendations for e-commerce businesses to optimise engagement and sales, the study examined various recommender systems such as content-based filtering, collaborative filtering and hybrid filtering. Among these methods, hybrid filtering technique, which combines the strengths of the other two methods, is the most effective when it comes to personalised recommendations. As seen in the analysis part, the hybrid model can provide recommendations based on product attributes and customers’ purchasing behaviours. For example, the system can track the users’ engagement with certain items and recommend related products, thereby increasing not only engagement and sales but also conversion rates. On top of that, customers are more inclined to interact with e-commerce platforms that recommend related products (Deloitte, 2019).

5.2.3 Research Question 3

“In what way does customer segmentation improve the accuracy and relevance of product recommendations in the UK e-commerce?”

By providing the product recommendations to specific customer groups, customer segmentation is essential for improving the accuracy and relevance of product recommendations. This study enabled the development of customer profiles through purchasing behaviours and demographic, which in turn influenced the recommendation models. For instance, the recommendation system could prioritise products with discounts or promotions in its suggestions to customers in cluster/segment 1. Similarly, the system can recommend branded products and exclusive offers to clusters 2 and 3 which prioritise brand quality, reputation and exclusivity. By employing this approach, online businesses can leverage the benefits, of increasing conversion rate and customer satisfaction.

5.3 Comparison with Existing Literature

This study research’s findings align with most of the existing literature in several key areas. First, the efficiency and usefulness of K-means clustering as a customer segmentation method are aligned with existing literature. The results of our research and analysis show that K-means are highly effective in the identification of different customer segments based on their shopping behaviours and demographic characteristics. This finding supports the Gomes & Meisen (2023)’s discussion, which focuses on K-means as one of the most frequently employed methods for customer segmentation in marketing analytics. Furthermore, the results of our analysis on the importance of personalised product recommendations align with existing literature, which highlights the role of recommendations in improving the customer experience and increasing sales. For example, the importance of hybrid filtering that combines both content-based and collaborative filtering is highlighted by the research done by Afoudi et al. (2021) and Thorat et al. (2015). Our research clearly pointed out this specific subject when we did the analysis for product recommendations to improve sales and customer engagement. On top of that, when it comes to getting better marketing strategies through behavioural segmentation, our findings show that customisation and shopping behaviours have many impacts in this competitive e-commerce sector. Segmenting the right customer groups based on their purchasing habits and preferences can lead to better marketing strategies, which are highlighted by Jiang & Tuzhilin (2009) & Yang et al. (2016). The lack of research on the long-term effects of personalised recommendations on consumer relationships is another significant gap in the literature, as per Afoudi et al. (2021). The majority of research focuses on immediate results, including sales and conversion rates. Our research fills this gap by taking into account long-term effects including customer loyalty and retention, which are essential for the long-term success of a business.

5.4 Limitations of The Study

Even though this research provides valuable insights into customer segmentation and product recommendations, there are still limitations regarding this study. First, this paper heavily relied on secondary sources, which might affect the customer clustering analysis and product recommendation system which is more effective with up-to-date trends. The absence of this real-time data could limit the relevance of findings, especially in this ever-changing e-commerce market. Furthermore, the study focused on only the e-commerce sector of the UK, which means that this might restrict how broadly the models and insights can be applied to different regions and industries. Differences in culture, economy, and technology across other markets can more or less influence behaviours of consumers and the effectiveness of segmentation and recommendation models in this research (Gupta et al., 2024). Another challenge is related to insufficient data in certain segments where there are less customer interactions. Zhang et al. (2020) addressed that this particular challenge can impact the accuracy and reliability of the recommendation models, especially for the products that are new or have limited purchasing history information. On top of that, external factors including seasonal trends or economic changes were not considered in this analysis research, which can influence customer behaviours and buying habits (Martínez et al., 2020). Therefore, further studies need to overcome this by including primary data gathering, examining various market scenarios and taking into account how external factors can influence customer behaviours and the accuracy of product recommendations.

6. Conclusion

6.1 Summary of Key Findings

In order to remain competitive in the ever-changing UK e-commerce sector, it is not only advantageous but also necessary for businesses to implement advanced data analytics and machine learning techniques. The analysis conducted in this study provided valuable insights into customer segmentation and product recommendation strategies that can be used to increase both customer engagement and sales performance. Using K-means clustering method, the study identified three separate customer groups; young, price-sensitive customers, time-conscious, convenience-targeted customers and wealthy customers who love to enjoy luxury and premium goods. These segments highlight the significance of personalised marketing strategies by revealing the different needs and preferences within the e-commerce market. The segmentation analysis of this research showed that businesses could improve customer engagement and loyalty by customising their offerings and communication to address the requirements of each segment. On top of that, the analysis of recommendation systems showed that they are really effective in enhancing customer loyalty and satisfaction. The hybrid model effectively addressed the limitations of single-method approaches, such as the cold start problem and limited product focus, by combining content-based and collaborative filtering methods. The findings from this method highlight the importance of personalisation in the e-commerce sector, where e-commerce businesses that offer relevant and timely product recommendations are more likely to succeed.

6.2 Practical Implications

There are several implications for the e-commerce sector from this research paper's findings. First of all, e-commerce platforms can leverage the results from the three identified customer segments to create more personalised and targeted marketing strategies, which means these businesses can improve customer engagement and satisfaction by targeting the right customer groups by their shopping behaviours and demographic characteristics. For example, businesses can prioritise quality and convenience for customers who are in segment 2, while providing exclusive discounts to price-sensitive customers through targeted campaigns. Furthermore, businesses can more accurately forecast the demand for a variety of products by understanding the preferences of different customer segments, which can result in more efficient inventory management and vice versa (Knox, 2024). This can prevent overstocking or stockouts and improve the overall supply chain efficiency. Additionally, this study's data-driven approach enables businesses to make informed decisions about customer service, marketing, and product development. This reduces the dependence on intuition or traditional outdated practices and puts the businesses on the right path of profited marketing strategies. E-commerce businesses can guarantee that their strategies remain effective and relevant over time by consistently updating their customer segmentation and recommendation models with new, updated data (Guo et al., 2019).

6.3 Recommendations for Future Research and Final Thoughts

This study shows how well customer segmentation and product recommendations using machine learning techniques can be implemented in e-commerce platforms to increase customer satisfaction and engagement. However, to get more rounded and complete findings, we came up with the following recommendations based on our analysis for future research.

Anpar Research (2020) demonstrated that to capture the most recent customer behaviours and trends, future research should explore the utilisation of primary data collection methods such as surveys and interviews focus groups. Insights into ever-changing e-commerce landscapes can be gained through real-time data, particularly in response to economic conditions, technological advancements and customer behaviour shifts. This primary research can improve the precision of customer segmentation and recommendation systems by integrating the most recent market dynamics and customer preferences. Secondly, further research should include external factors, including seasonal trends, cultural and technological contexts, to examine better customer behaviours and recommendation systems. By examining these factors, further research can create more adaptive and better recommendation systems that more accurately reflect customer behaviours. Lastly, further research should look into the ethical implications of using customer data for segemtation and recommendation systems, with a particular focus on the potential for bias in algorithmic decision-making, consent and data privacy. Further research must address these concerns of data privacy and the ethical use of artificial intelligence (AI). This includes creating ethical data collection and analysis frameworks, as well as the prevention of recommendation systems that promote harmful biases or violate customer privacy (Naz & Kashif, 2024).

To conclude the research, this study emphasises the critical importance of employing advanced data analytics and machine learning techniques in this rapidly changing UK e-commerce market. Businesses can dramatically improve customer engagement, satisfaction and loyalty by providing personalised product recommendations and effectively segmenting customers. In this competitive landscape, a one-size-fits-all approach is no longer enough, as shown by the insights and findings from this study. Rather, it is important to have an advanced clustering algorithm and hybrid recommendation system that provide a complete understanding of customer preferences and behaviours in order to maintain a competitive edge and drive growth. This research has a wide range of practical applications, providing businesses with actionable strategies to improve their marketing efforts, and ultimately increase profitability. Nevertheless, the ever-changing nature of customer behaviours and technological advancements requires consistent innovation and adaptation. E-commerce platforms must maintain flexible by continuously updating their models and strategies to align with changing trends. Those who prioritise personalised customer experiences and data-driven decision-making will be the most successful in this competitive e-commerce industry.

7. Reference List

Afoudi, Y., Lazaar, M., & Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, *113*(113), 102375. https://doi.org/10.1016/j.simpat.2021.102375

Agarwal, V. (2015). Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. *International Journal of Computer Applications*, *131*(4), 30–36. https://doi.org/10.5120/ijca2015907309

Ahlm, S., Holmström, M., Stenman, V., Johansson, U., & Sneistrup, S. (2007). *International Marketing and Brand Management BUS 809 Master Thesis in International Marketing*. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1339435&fileOId=2434865

Ahmed, A. (2023, July 26). *E-commerce and Its Growing Influence on Consumer Behaviour*. Medium. https://medium.com/@wisewealthywizard./e-commerce-and-its-growing-influence-on-consumer-behaviour-8f0cd7f6bcac

Anpar Research. (2020, September 4). *Primary Research vs Secondary Research: Pros & Cons, Types*. Anpar Research Ltd. https://www.anparresearchltd.com/post/primary-research-vs-secondary-research

Anunaya, S. (2021, August 10). *Data Preprocessing in Data Mining - A Hands On Guide (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/#

BigCommerce. (n.d.). *Ecommerce Machine Learning: Business Benefits + Use Cases*. BigCommerce. https://www.bigcommerce.co.uk/articles/ecommerce/machine-learning/

Bodhe, S. (2023, March 14). *What is Technographic Segmentation and its Significance in B2B*. Vereigen Media. https://vereigenmedia.com/what-is-technographic-segmentation-and-its-significance-in-b2b/

Boerman, S. C., Kruikemeier, S., & Bol, N. (2021). When is personalized advertising crossing personal boundaries? How type of information, data sharing, and personalized pricing influence consumer perceptions of personalized advertising. *Computers in Human Behavior Reports*, *4*(4), 100144. https://doi.org/10.1016/j.chbr.2021.100144

Chevalier, S. (2024, March 15). *Great Britain: online shopping conversion rate by device 2020*. Statista. https://www.statista.com/statistics/960431/great-britain-online-shopper-conversion-rate-by-device/

Cho, Y. S., & Moon, S. C. (2013). Weighted Mining Frequent Pattern based Customer's RFM Score for Personalized u-Commerce Recommendation System. *JoC*, *4*(4), 36–40. https://www.earticle.net/Article/A215904

Chongkolnee Rungruang, Pakwan Riyapan, Arthit Intarasit, Khanchit Chuarkham, & Jirapond Muangprathub. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449–121449. https://doi.org/10.1016/j.eswa.2023.121449

Coppola, D. (2024a, April 11). *United Kingdom top online stores 2018*. Statista. https://www.statista.com/forecasts/870307/united-kingdom-top-online-stores-united-kingdom-ecommercedb

Coppola, D. (2024b, June 24). *UK: preferred e-commerce payment methods 2020*. Statista. https://www.statista.com/statistics/1031317/card-not-present-payments-in-the-united-kingdom-by-payment-method/

Deloitte. (2019). The Deloitte Consumer Review Made-to-order: The rise of mass personalisation. In *Deloitte*.

https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/consumer-business/ch-en-consumer-business-made-to-order-consumer-review.pdf

Dey, D. (2019, May 8). *Hierarchical Clustering in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/hierarchical-clustering/

Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L., & Zhang, F. (2017). A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10747

ECDB. (2023). *eCommerce Benchmark KPIs: UK*. Ecommercedb.com. https://ecommercedb.com/benchmarks/gb/all

Edquist, A., Grennan, L., Griffiths, S., & Rowshankish, K. (2022, September 23). *Data ethics: What it means and what it takes | McKinsey*. Www.mckinsey.com. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes

Evelyn. (2023, November 4). *Collaborative Filtering in Recommender System: An Overview*. Medium. https://medium.com/@evelyn.eve.9512/collaborative-filtering-in-recommender-system-an-overview-38dfa8462b61

FasterCapital. (2024, June 1). *Technographic Segmentation: The Role of Technographic Segmentation in Effective Marketing Strategies*. FasterCapital . https://fastercapital.com/content/Technographic-Segmentation--The-Role-of-Technographic-Segmentation-in-Effective-Marketing-Strategies.html

Gao, R. (2023). On the Importance of Pricing Strategy in Marketing Strategy. *Frontiers in Business, Economics and Management*, *10*(1), 158–161. https://doi.org/10.54097/fbem.v10i1.10234

Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems* *and E-Business Management*, *21*, 527–570. https://doi.org/10.1007/s10257-023-00640-4

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology*, *72*, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006

Gulhane, A., Mohite, J., Khajbage, V., Turkade, A., Pund, A., & Kanade, S. (2024). IJIRID International Journal of Ingenious Research, Invention and Development Customer Segmentation with Machine Learning. *Customer Segmentation with Machine Learning*, *3*(2). https://doi.org/10.5281/zenodo.11098286

Gültekin, H. (2023, September 7). *What is Silhouette Score?* Medium. https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a

Guo, M., Liao, X., Liu, J., & Zhang, Q. (2019). Consumer preference analysis: A data-driven multiple criteria approach integrating online information. *Omega*, *96*, 102074. https://doi.org/10.1016/j.omega.2019.05.010

Gupta, C., Jindal, P., & Madhavi Shamkuwar. (2024). Impact of Cultural Marketing on Buying Behaviour of the Consumers. *Emerald Publishing Limited EBooks*, 153–162. https://doi.org/10.1108/978-1-83753-734-120241011

Halej, J. (2017). *Ethics in Primary Research (focus groups, interviews and surveys)*. Equality Challenge Unit. https://forms.docstore.port.ac.uk/A816773.pdf

Hayden, M. (2022, August 25). *What is transactional data and how can you leverage it?* Lytics Customer Data Platform (CDP). https://www.lytics.com/blog/what-is-transactional-data-and-how-can-you-leverage-it/

Hayes, M., & Downie, A. (2024, June 19). *What is Data Transformation? | IBM*. Www.ibm.com. https://www.ibm.com/think/topics/data-transformation

IBM. (n.d.). *What is Data Integration? | IBM*. Www.ibm.com. https://www.ibm.com/topics/data-integration

IBM. (2020). *What Is Exploratory Data Analysis? | IBM*. Www.ibm.com. https://www.ibm.com/topics/exploratory-data-analysis

IBM. (2023, December 8). *What is principal component analysis? | IBM*. Www.ibm.com. https://www.ibm.com/topics/principal-component-analysis

International Trade Administration. (2022, September 12). *United Kingdom - eCommerce*. Www.trade.gov. https://www.trade.gov/country-commercial-guides/united-kingdom-ecommerce

Jabbar, A., Akhtar, P., & Dani, S. (2019). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, *90*. https://doi.org/10.1016/j.indmarman.2019.09.001

Jiang, T., & Tuzhilin, A. (2009). Improving Personalization Solutions through Optimal Segmentation of Customer Bases. *IEEE Transactions on Knowledge and Data Engineering*, *21*(3), 305–320. https://doi.org/10.1109/tkde.2008.163

Karabiber, F. (n.d.). *TF-IDF — Term Frequency-Inverse Document Frequency – LearnDataSci*. Www.learndatasci.com. https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/

Kimnaruk, Y. (2022, July 11). *Basic content-based recommendation system with Python code*. Medium. https://yannawut.medium.com/basic-content-based-recommendation-system-with-python-code-be920b412067

Kindling, M., & Strecker, D. (2022). Data Quality Assurance at Research Data Repositories. *Data Science Journal*, *21*. https://doi.org/10.5334/dsj-2022-018

Knox, T. (2024, August 12). Council Post: Why Effective And Efficient Inventory Management Is Key To Delivering Positive Customer Experiences In Retail. *Forbes*. https://www.forbes.com/councils/forbestechcouncil/2022/04/19/why-effective-

and-efficient-inventory-management-is-key-to-delivering-positive-customer-experiences-in-retail/

Latifi, S., Dietmar Jannach, & Ferraro, A. (2022). Sequential recommendation: A study on transformers, nearest neighbors and sampled metrics. *Information Sciences*, *609*, 660–678. https://doi.org/10.1016/j.ins.2022.07.079

Maharana, K., Mondal, S., & Nemade, B. (2022). A Review: Data Pre-Processing and Data Augmentation Techniques. *Global Transitions Proceedings*, *3*(1), 91–99. Sciencedirect. https://doi.org/10.1016/j.gltp.2022.04.020

Martínez, J. M. G., Martín, J. M. M., & Rey, M. del S. O. (2020). An analysis of the changes in the seasonal patterns of tourist behavior during a process of economic recovery. *Technological Forecasting and Social Change*, *161*, 120280. https://doi.org/10.1016/j.techfore.2020.120280

Monetate. (n.d.). *What are Product Recommendations?* Monetate. https://monetate.com/resources/glossary-product-recommendations/

Murel, J., & Kavlakoglu, E. (2024a, March 21). *What is collaborative filtering? | IBM*. Www.ibm.com. https://www.ibm.com/topics/collaborative-filtering

Murel, J., & Kavlakoglu, E. (2024b, March 21). *What is content-based filtering? | IBM*. Www.ibm.com. https://www.ibm.com/topics/content-based-filtering

Naz, H., & Kashif, M. (2024). Artificial intelligence and predictive marketing: an ethical framework from managers' perspective. *Spanish Journal of Marketing - ESIC*. https://doi.org/10.1108/sjme-06-2023-0154

Nguyen, D.-N., Nguyen, V.-H., Trinh, T., Ho, T., & Le, H.-S. (2024). A personalized product recommendation model in e-commerce based on retrieval strategy. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(2), 100303. https://doi.org/10.1016/j.joitmc.2024.100303

Nur Balqish Hassan, & Noor Hazarina Hashim. (2024). Technographic segmentation of smartphone usage at the Rainforest World Music Festival. *International Journal of Event and Festival Management*. https://doi.org/10.1108/ijefm-11-2023-0085

Otebele, K. (2023, August 16). *The Psychology of Purchase: How Behavioral Segmentation Influences Buying Decisions - Invesp*. Invesp. https://www.invespcro.com/blog/behavioral-segmentation/

Powell, P., & Smalley, I. (2024, March 20). *What Is Data Reduction? | IBM*. Www.ibm.com. https://www.ibm.com/topics/data-reduction

Ramos, G. (2022). *An Online Shop Business Transaction*. Www.kaggle.com. https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business

Rana, J., Dilshad, S., & Ahsan, Md. A. (2021). Ethical Issues in Research. *Global Encyclopedia of Public Administration, Public Policy, and Governance*, 1–7. https://doi.org/10.1007/978-3-319-31816-5_462-1

Rana, J., Gutierrez, P. L., & Oldroyd, J. C. (2021). Quantitative Methods. *Global Encyclopedia of Public Administration, Public Policy, and Governance*, 1–6. Springer.

Rane, N., Achari, A., & Choudhary, S. P. (2023). Enhancing customer loyalty through quality of service: Effective strategies to improve customer satisfaction, experience, relationship, and engagement. *International Research Journal of Modernization in Engineering Technology and Science*, *5*(5), 427–452. Researchgate. https://doi.org/10.56726/irjmets38104

Rosário, A., & Raimundo, R. (2021). Consumer Marketing Strategy and E-Commerce in the Last Decade: a Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(7), 3003–3024. mdpi. https://www.mdpi.com/0718-1876/16/7/164

Rushirajsinh, Z. (2023, November 4). *The Elbow Method: Finding the Optimal Number of Clusters*. Medium. https://medium.com/@zalarushirajsinh07/the-elbow-method-finding-the-optimal-number-of-clusters-d297f5aeb189

Russo Spena, T., D'Auria, A., & Bifulco, F. (2021). Customer insights and consumer profiling. *Contributions to Management Science*, 95–117. https://doi.org/10.1007/978-3-030-63376-9_5

Savela, T. (2018). The Advantages and Disadvantages of Quantitative Methods in Schoolscape Research. *Linguistics and Education*, *44*(1), 31–44. Sciencedirect. https://doi.org/10.1016/j.linged.2017.09.004

Shah, S., & Singh, M. (2012). Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. *2012 International Conference on Communication Systems and Network Technologies*. https://doi.org/10.1109/csnt.2012.100

Sharma, P. (2019, August 19). *The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#:~:text=K%2Dmeans%20clustering%2C%20originating%20from

Sondhi, N. (2017). Segmenting & profiling the deflecting customer: understanding shopping cart abandonment. *Procedia Computer Science*, *122*, 392–399. https://doi.org/10.1016/j.procs.2017.11.385

Statista. (2021, May 20). *E-commerce users in the United Kingdom 2025*. Statista. https://www.statista.com/forecasts/477128/e-commerce-users-in-the-united-kingdom

Statista. (2022, August 17). *E-commerce revenue in in the United Kingdom 2017-2025*. Statista. https://www.statista.com/forecasts/477091/e-commerce-revenue-forecast-in-the-united-kingdom

Statista. (2023a, July 25). *Goods most frequently bought online in the UK 2022*. Statista. https://www.statista.com/statistics/1322601/online-shopping-frequency-category-uk/

Statista. (2023b, July 26). *UK: online shopping device preference by generation 2023*. Statista. https://www.statista.com/statistics/1358321/online-shopper-device-preference-uk-generation/

Statista. (2023c, October). *E-Commerce in the United Kingdom (UK) 2023*. Statista. https://www.statista.com/study/105482/e-commerce-in-the-uk/

Statista. (2024, January 29). *UK: cart abandonment rate by device 2021*. Statista. https://www.statista.com/statistics/1254962/cart-abandonment-rate-in-the-uk/

Susilo, W. H. (2016). An Impact of Behavioral Segmentation to Increase Consumer Loyalty: Empirical Study in Higher Education of Postgraduate Institutions at Jakarta. *Procedia - Social and Behavioral Sciences*, *229*, 183–195. https://doi.org/10.1016/j.sbspro.2016.07.128

Thorat, P., Goudar, R., & Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. *International Journal of Computer Applications*, *110*(4), 31–36. https://doi.org/10.5120/19308-0760

University College London. (2022). Advantages of Secondary Data Analysis. In *University College London*. https://ethics.grad.ucl.ac.uk/forms/Secondary-data-analysis-file-note.pdf

US EPA, O. (2015, April 2). *Exploratory Data Analysis*. Www.epa.gov. https://www.epa.gov/caddis/exploratory-data-analysis

Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management*, *59*(6), 103085. https://doi.org/10.1016/j.ipm.2022.103085

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Xue, J., Lee, C., Wakeham, S. G., & Armstrong, R. A. (2011). Using principal components analysis (PCA) with cluster analysis to study the organic geochemistry of sinking particles in the ocean. *Organic Geochemistry*, *42*(4), 356–367. https://doi.org/10.1016/j.orggeochem.2011.01.012

Yallop, A. C., Baker, J. J., & Wardle, J. (2022). Market Research and Insight: Past, Present and Future. *International Journal of Market Research*, *64*(2), 163–168. Sagepub. https://doi.org/10.1177/14707853221080735

Yang, J., Liu, C., Teng, M., Liao, M., & Xiong, H. (2016). Buyer targeting optimization: A unified customer segmentation perspective. *IEEE Xplore*. https://doi.org/10.1109/bigdata.2016.7840730

Yi, Z. (2018). Market Segmentation, Targeting, and Positioning. *Marketing Services and Resources in Information Organizations*, 39–48. https://doi.org/10.1016/b978-0-08-100798-3.00004-0

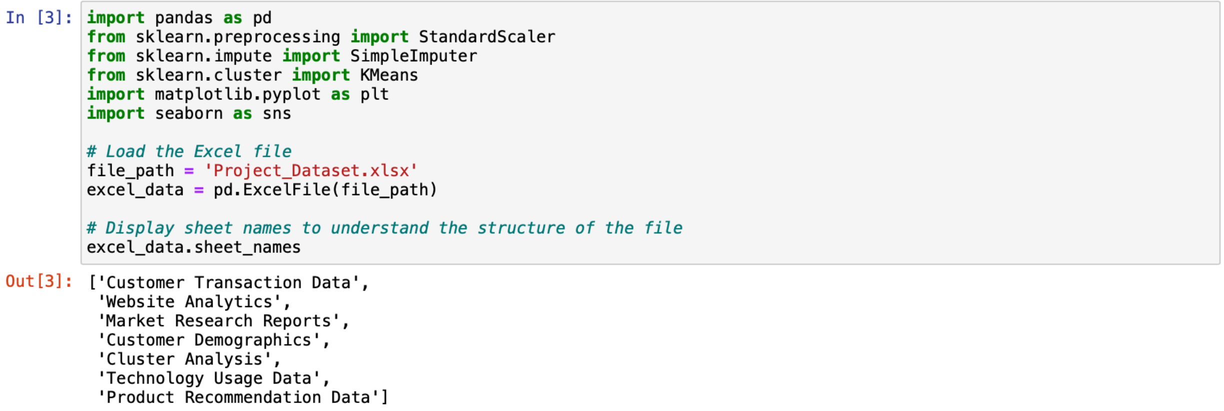
Yusuf, F. (2023, June 1). *Difference between K means and Hierarchical Clustering*. Medium. https://medium.com/@wazirphareeyda/difference-between-k-means-and-hierarchical-clustering-edfec55a34f8

Zhang, F., Qi, S., Liu, Q., Mao, M., & Zeng, A. (2020). Alleviating the data sparsity problem of recommender systems by clustering nodes in bipartite networks. *Expert Systems with Applications*, *149*, 113346. https://doi.org/10.1016/j.eswa.2020.113346

Zott, C., Amit, R., & Donlevy, J. (2000). Strategies for value creation in e-commerce: *European Management Journal*, *18*(5), 463–475. https://doi.org/10.1016/s0263-2373(00)00036-0

8. Appendix

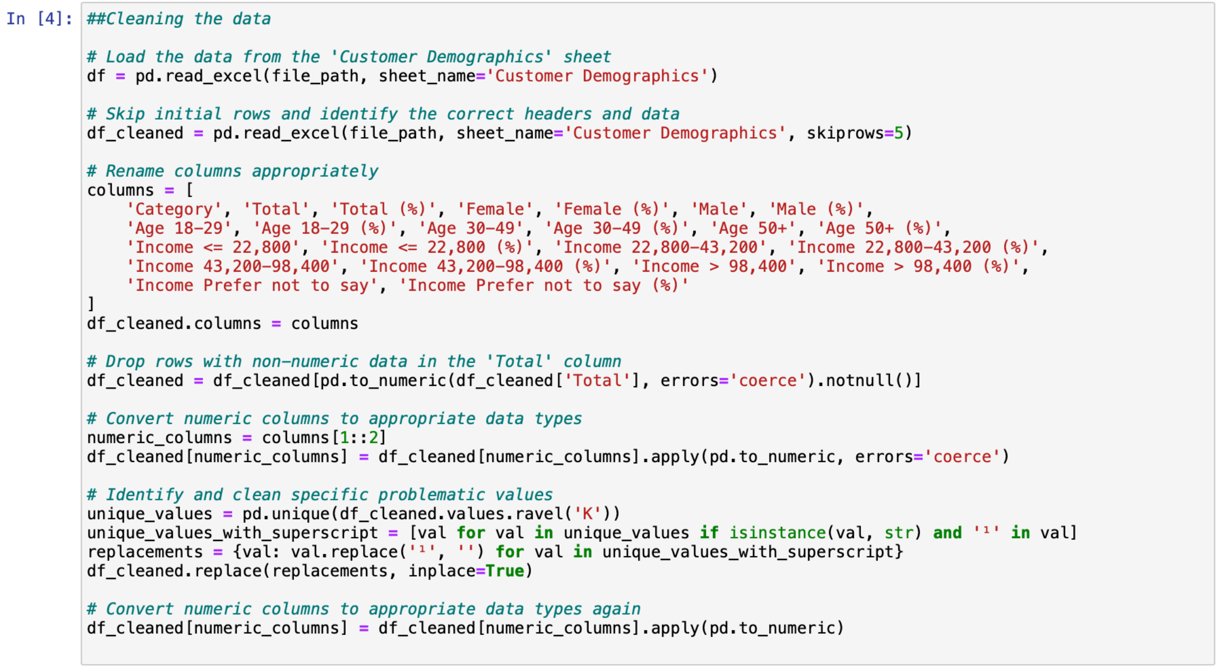
8.1 Appendix A



```
In [3]: import pandas as pd
        from sklearn.preprocessing import StandardScaler
        from sklearn.impute import SimpleImputer
        from sklearn.cluster import KMeans
        import matplotlib.pyplot as plt
        import seaborn as sns

        # Load the Excel file
        file_path = 'Project_Dataset.xlsx'
        excel_data = pd.ExcelFile(file_path)

        # Display sheet names to understand the structure of the file
        excel_data.sheet_names
```

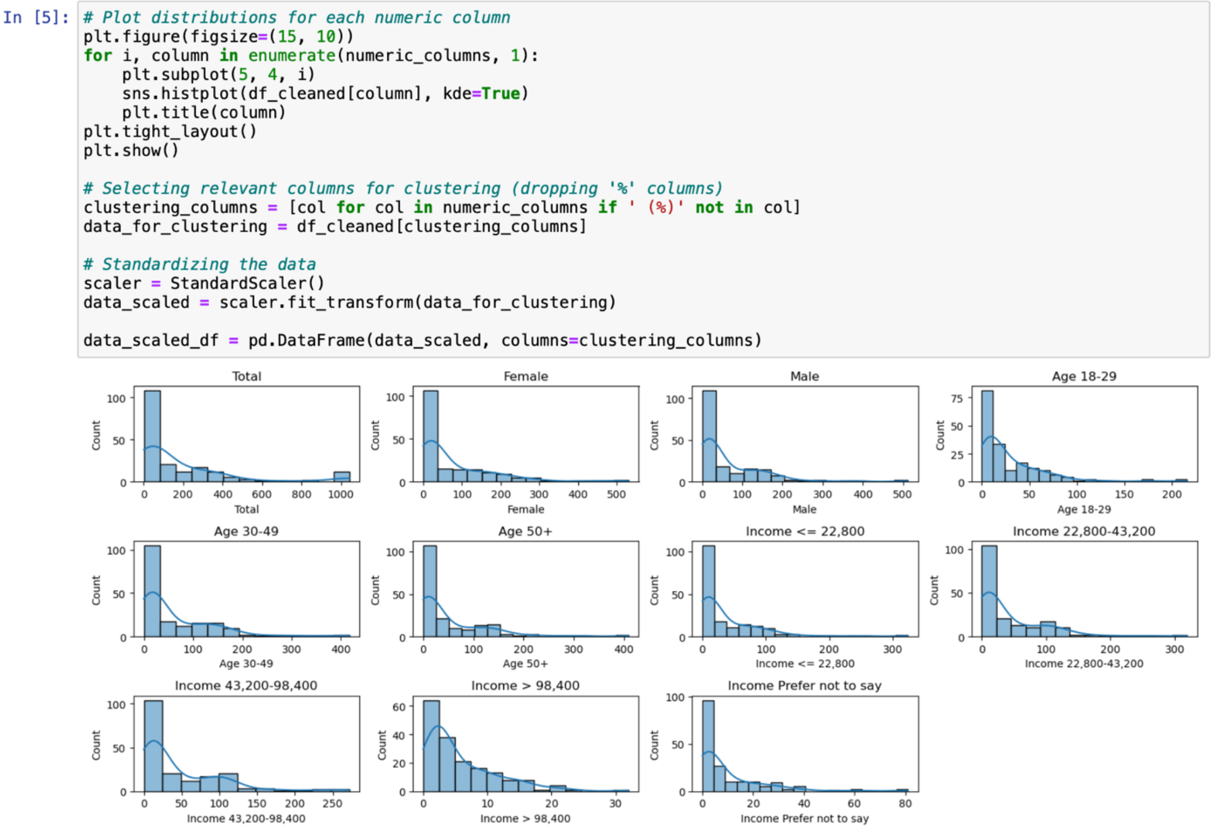
```
Out[3]: ['Customer Transaction Data',
         'Website Analytics',
         'Market Research Reports',
         'Customer Demographics',
         'Cluster Analysis',
         'Technology Usage Data',
         'Product Recommendation Data']
```

- Import necessary libraries for data manipulation, preprocessing, clustering, and visualisation.
- Load the Project Dataset Excel file named 'Project_Dataset.xlsx'.
- Display the sheet names in the Excel file to understand its structure.



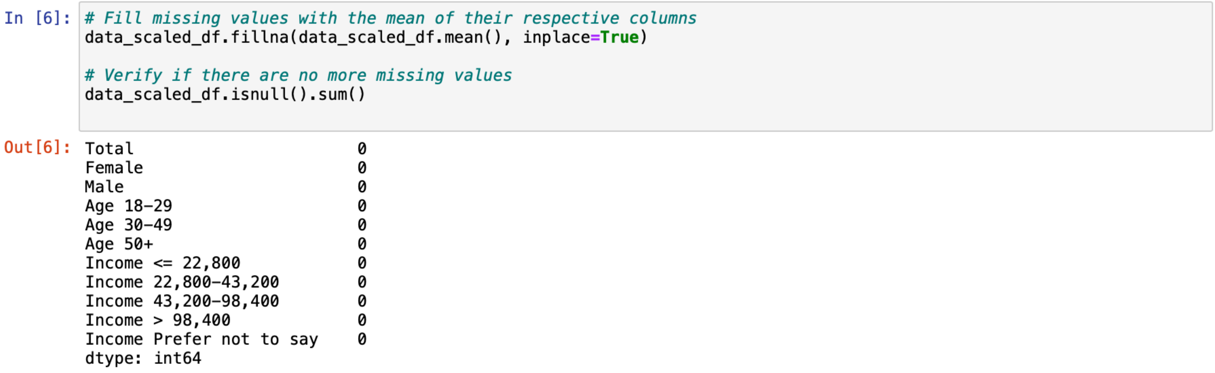
```
In [4]: ##Cleaning the data

        # Load the data from the 'Customer Demographics' sheet
        df = pd.read_excel(file_path, sheet_name='Customer Demographics')

        # Skip initial rows and identify the correct headers and data
        df_cleaned = pd.read_excel(file_path, sheet_name='Customer Demographics', skiprows=5)

        # Rename columns appropriately
        columns = [
            'Category', 'Total', 'Total (%)', 'Female', 'Female (%)', 'Male', 'Male (%)',
            'Age 18-29', 'Age 18-29 (%)', 'Age 30-49', 'Age 30-49 (%)', 'Age 50+', 'Age 50+ (%)',
            'Income <= 22,800', 'Income <= 22,800 (%)', 'Income 22,800-43,200', 'Income 22,800-43,200 (%)',
            'Income 43,200-98,400', 'Income 43,200-98,400 (%)', 'Income > 98,400', 'Income > 98,400 (%)',
            'Income Prefer not to say', 'Income Prefer not to say (%)'
        ]
        df_cleaned.columns = columns

        # Drop rows with non-numeric data in the 'Total' column
        df_cleaned = df_cleaned[pd.to_numeric(df_cleaned['Total'], errors='coerce').notnull()]

        # Convert numeric columns to appropriate data types
        numeric_columns = columns[1::2]
        df_cleaned[numeric_columns] = df_cleaned[numeric_columns].apply(pd.to_numeric, errors='coerce')

        # Identify and clean specific problematic values
        unique_values = pd.unique(df_cleaned.values.ravel('K'))
        unique_values_with_superscript = [val for val in unique_values if isinstance(val, str) and '¹' in val]
        replacements = {val: val.replace('¹', '') for val in unique_values_with_superscript}
        df_cleaned.replace(replacements, inplace=True)

        # Convert numeric columns to appropriate data types again
        df_cleaned[numeric_columns] = df_cleaned[numeric_columns].apply(pd.to_numeric)
```

- Load and clean the data from the 'Customer Demographics' sheet, skipping the initial rows to find the correct headers.
- Rename columns for clarity.
- Remove non-numeric data from the 'Total' column.
- Convert columns to appropriate numeric data types.
- Handle problematic values such as superscripts in the data.



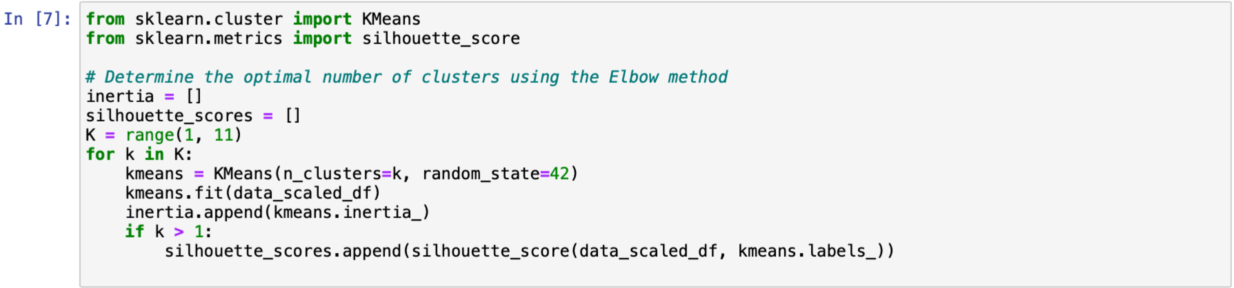
```
In [5]: # Plot distributions for each numeric column
        plt.figure(figsize=(15, 10))
        for i, column in enumerate(numeric_columns, 1):
            plt.subplot(5, 4, i)
            sns.histplot(df_cleaned[column], kde=True)
            plt.title(column)
        plt.tight_layout()
        plt.show()

        # Selecting relevant columns for clustering (dropping '%' columns)
        clustering_columns = [col for col in numeric_columns if ' (%)' not in col]
        data_for_clustering = df_cleaned[clustering_columns]

        # Standardizing the data
        scaler = StandardScaler()
        data_scaled = scaler.fit_transform(data_for_clustering)

        data_scaled_df = pd.DataFrame(data_scaled, columns=clustering_columns)
```

- Plot distributions for each numeric column.
- Select relevant columns for clustering, excluding percentage columns.
- Standardise the data using StandardScaler.



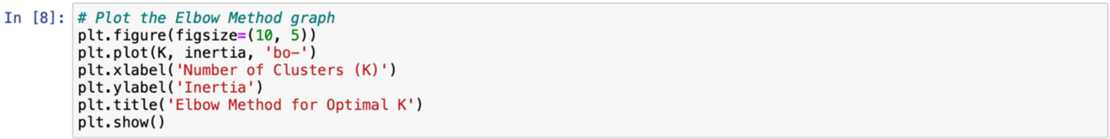
```
In [6]: # Fill missing values with the mean of their respective columns
        data_scaled_df.fillna(data_scaled_df.mean(), inplace=True)

        # Verify if there are no more missing values
        data_scaled_df.isnull().sum()

Out[6]: Total                      0
        Female                     0
        Male                       0
        Age 18-29                  0
        Age 30-49                  0
        Age 50+                    0
        Income <= 22,800           0
        Income 22,800-43,200       0
        Income 43,200-98,400       0
        Income > 98,400            0
        Income Prefer not to say   0
        dtype: int64
```

- Fill missing values with the mean of their respective columns.
- Verify that there are no more missing values.

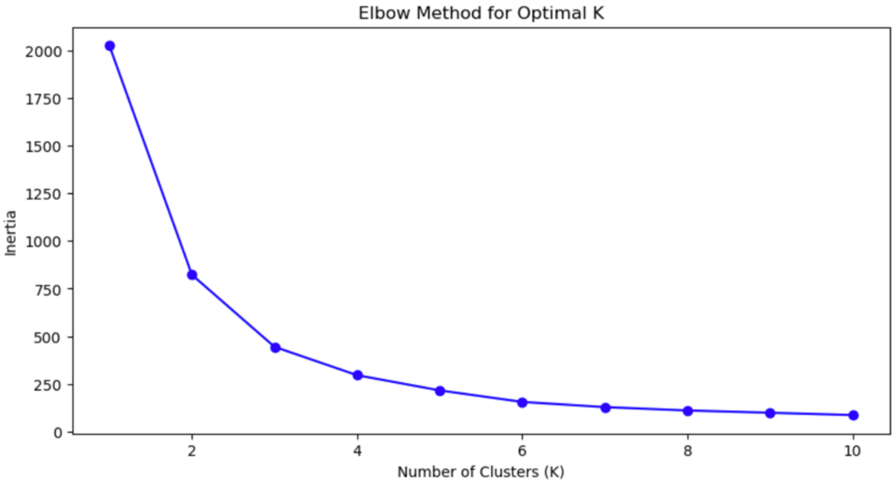


```
In [7]: from sklearn.cluster import KMeans
        from sklearn.metrics import silhouette_score

        # Determine the optimal number of clusters using the Elbow method
        inertia = []
        silhouette_scores = []
        K = range(1, 11)
        for k in K:
            kmeans = KMeans(n_clusters=k, random_state=42)
            kmeans.fit(data_scaled_df)
            inertia.append(kmeans.inertia_)
            if k > 1:
                silhouette_scores.append(silhouette_score(data_scaled_df, kmeans.labels_))
```

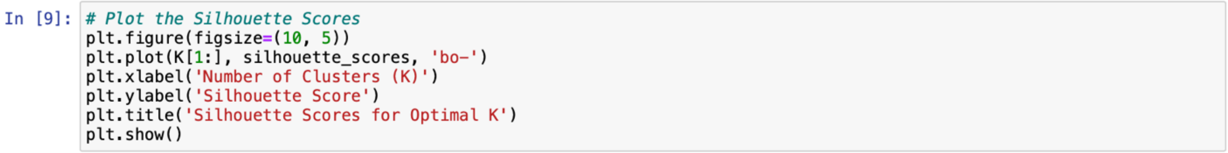
- Initialise lists for inertia and silhouette scores.
- Calculate inertia and silhouette scores for K values from 1 to 10 using K-means clustering.



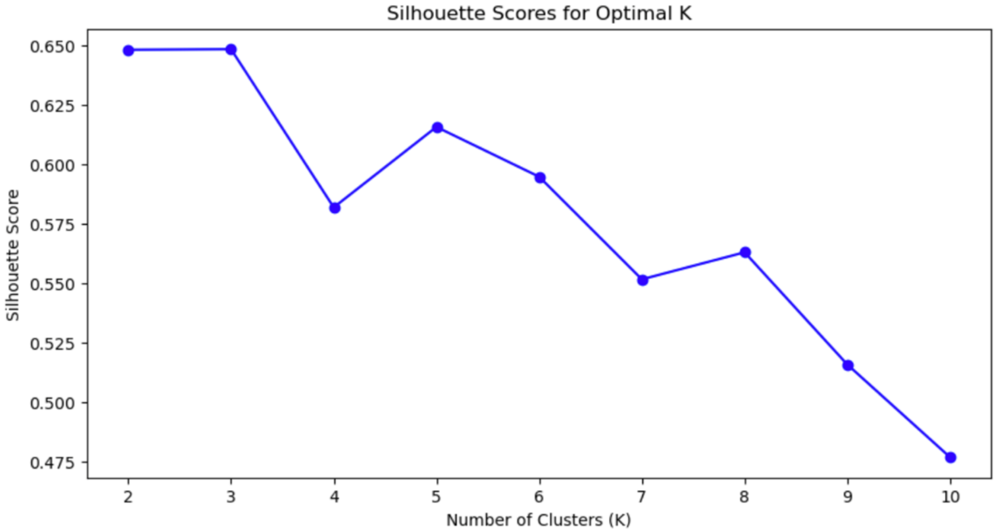
```
In [8]: # Plot the Elbow Method graph
plt.figure(figsize=(10, 5))
plt.plot(K, inertia, 'bo-')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal K')
plt.show()
```



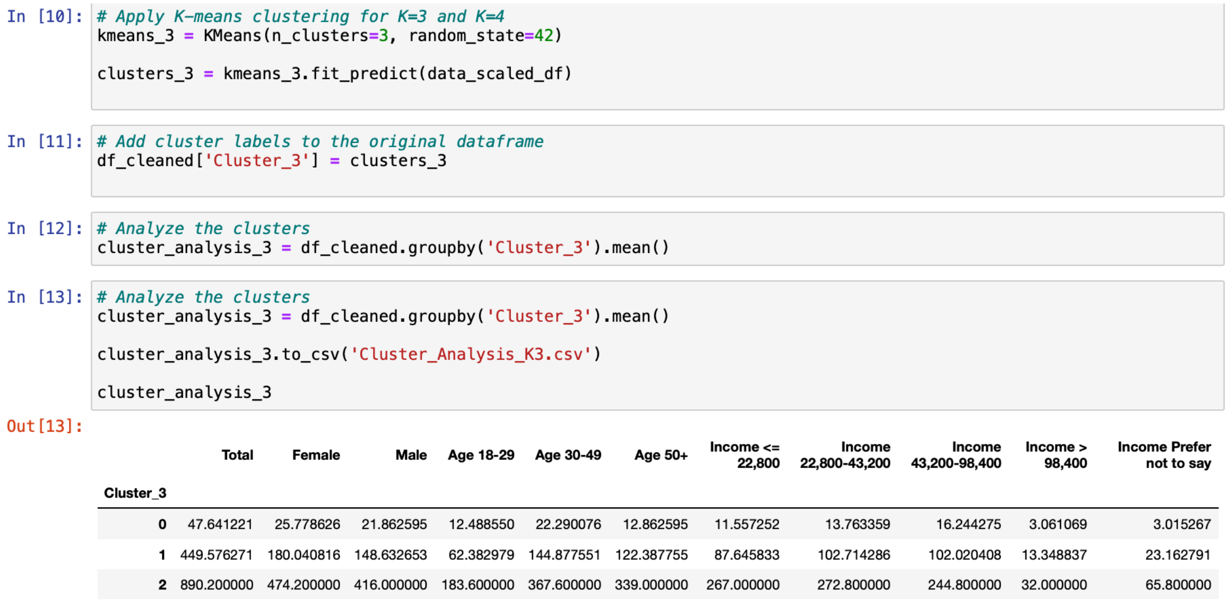
Plot the Elbow Method graph to visualise the optimal number of clusters.



```
In [9]: # Plot the Silhouette Scores
plt.figure(figsize=(10, 5))
plt.plot(K[1:], silhouette_scores, 'bo-')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Scores for Optimal K')
plt.show()
```



Plot the Silhouette Scores to further validate the optimal number of clusters.



```
In [10]: # Apply K-means clustering for K=3 and K=4
         kmeans_3 = KMeans(n_clusters=3, random_state=42)

         clusters_3 = kmeans_3.fit_predict(data_scaled_df)
```

```
In [11]: # Add cluster labels to the original dataframe
         df_cleaned['Cluster_3'] = clusters_3
```

```
In [12]: # Analyze the clusters
         cluster_analysis_3 = df_cleaned.groupby('Cluster_3').mean()
```

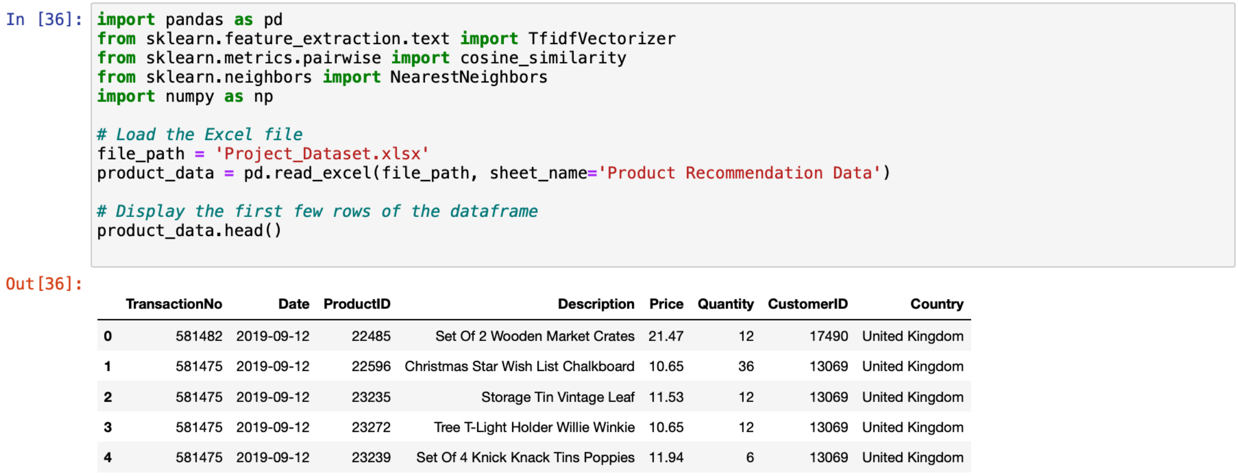
```
In [13]: # Analyze the clusters
         cluster_analysis_3 = df_cleaned.groupby('Cluster_3').mean()

         cluster_analysis_3.to_csv('Cluster_Analysis_K3.csv')

         cluster_analysis_3
```

Out[13]:

	Total	Female	Male	Age 18-29	Age 30-49	Age 50+	Income <= 22,800	Income 22,800-43,200	Income 43,200-98,400	Income > 98,400	Income Prefer not to say
Cluster_3											
0	47.641221	25.778626	21.862595	12.488550	22.290076	12.862595	11.557252	13.763359	16.244275	3.061069	3.015267
1	449.576271	180.040816	148.632653	62.382979	144.877551	122.387755	87.645833	102.714286	102.020408	13.348837	23.162791
2	890.200000	474.200000	416.000000	183.600000	367.600000	339.000000	267.000000	272.800000	244.800000	32.000000	65.800000

- Apply K-means clustering for K=3.
- Add cluster labels to the original dataframe.
- Analyse the clusters by calculating the mean for each cluster and save the result to a CSV file.

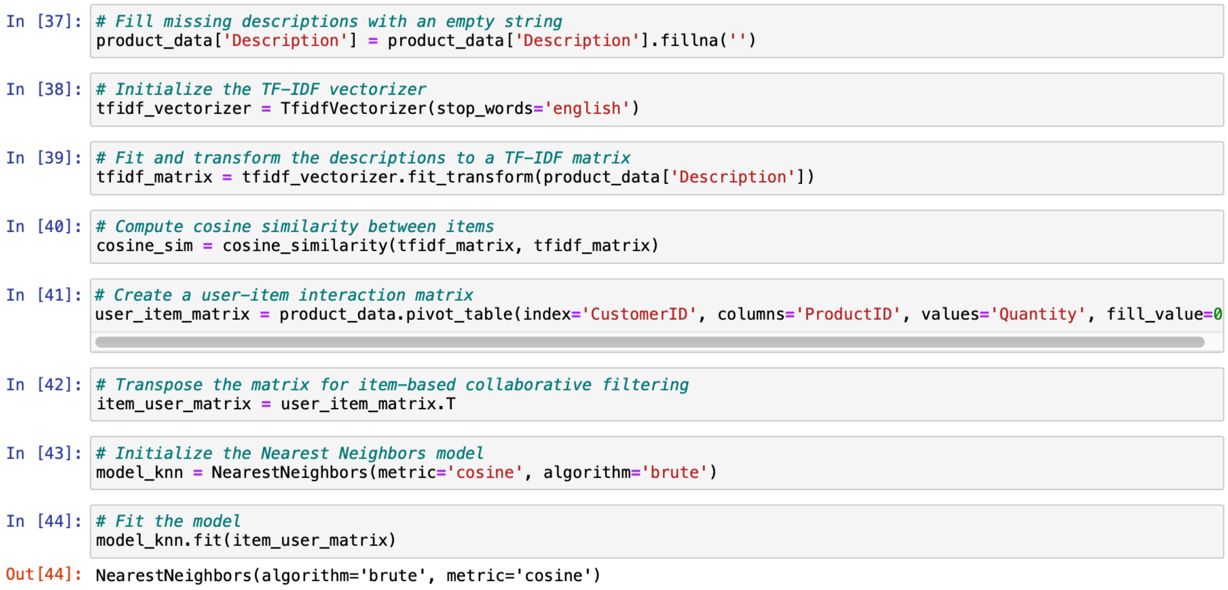
8.2 Appendix B



```
In [36]: import pandas as pd
         from sklearn.feature_extraction.text import TfidfVectorizer
         from sklearn.metrics.pairwise import cosine_similarity
         from sklearn.neighbors import NearestNeighbors
         import numpy as np

         # Load the Excel file
         file_path = 'Project_Dataset.xlsx'
         product_data = pd.read_excel(file_path, sheet_name='Product Recommendation Data')

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

Out[36]:
```

	TransactionNo	Date	ProductID	Description	Price	Quantity	CustomerID	Country
0	581482	2019-09-12	22485	Set Of 2 Wooden Market Crates	21.47	12	17490	United Kingdom
1	581475	2019-09-12	22596	Christmas Star Wish List Chalkboard	10.65	36	13069	United Kingdom
2	581475	2019-09-12	23235	Storage Tin Vintage Leaf	11.53	12	13069	United Kingdom
3	581475	2019-09-12	23272	Tree T-Light Holder Willie Winkie	10.65	12	13069	United Kingdom
4	581475	2019-09-12	23239	Set Of 4 Knick Knack Tins Poppies	11.94	6	13069	United Kingdom

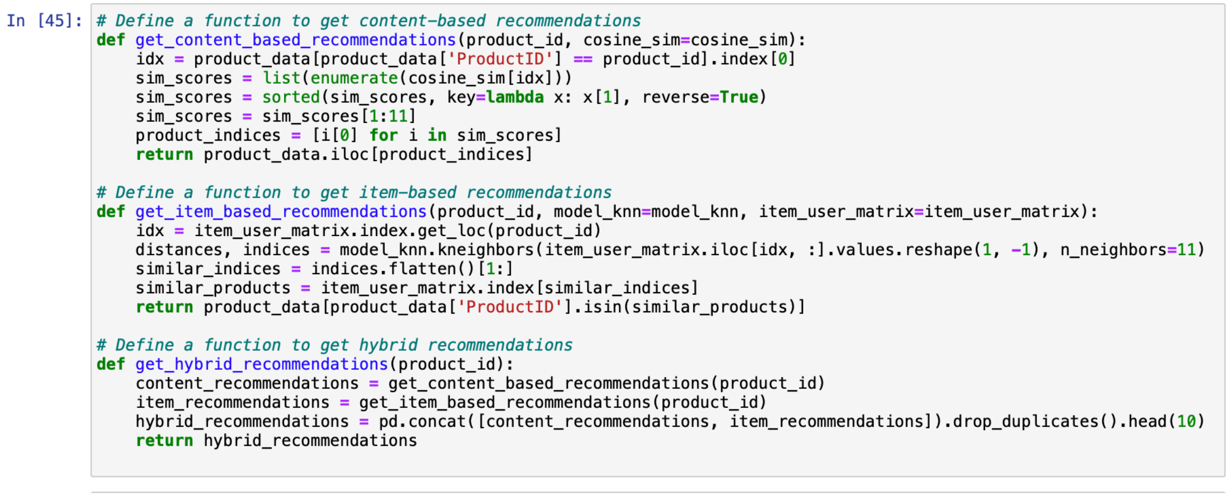
- Import necessary libraries for data manipulation, text vectorisation, similarity calculation, and nearest neighbors.
- Load the Project Dataset Excel file 'Project_Dataset.xlsx' and read the 'Product Recommendation Data' sheet.
- Display the first few rows of the dataframe.



```
In [37]: # Fill missing descriptions with an empty string
         product_data['Description'] = product_data['Description'].fillna('')

In [38]: # Initialize the TF-IDF vectorizer
         tfidf_vectorizer = TfidfVectorizer(stop_words='english')

In [39]: # Fit and transform the descriptions to a TF-IDF matrix
         tfidf_matrix = tfidf_vectorizer.fit_transform(product_data['Description'])

In [40]: # Compute cosine similarity between items
         cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)

In [41]: # Create a user-item interaction matrix
         user_item_matrix = product_data.pivot_table(index='CustomerID', columns='ProductID', values='Quantity', fill_value=0

In [42]: # Transpose the matrix for item-based collaborative filtering
         item_user_matrix = user_item_matrix.T

In [43]: # Initialize the Nearest Neighbors model
         model_knn = NearestNeighbors(metric='cosine', algorithm='brute')

In [44]: # Fit the model
         model_knn.fit(item_user_matrix)

Out[44]: NearestNeighbors(algorithm='brute', metric='cosine')
```

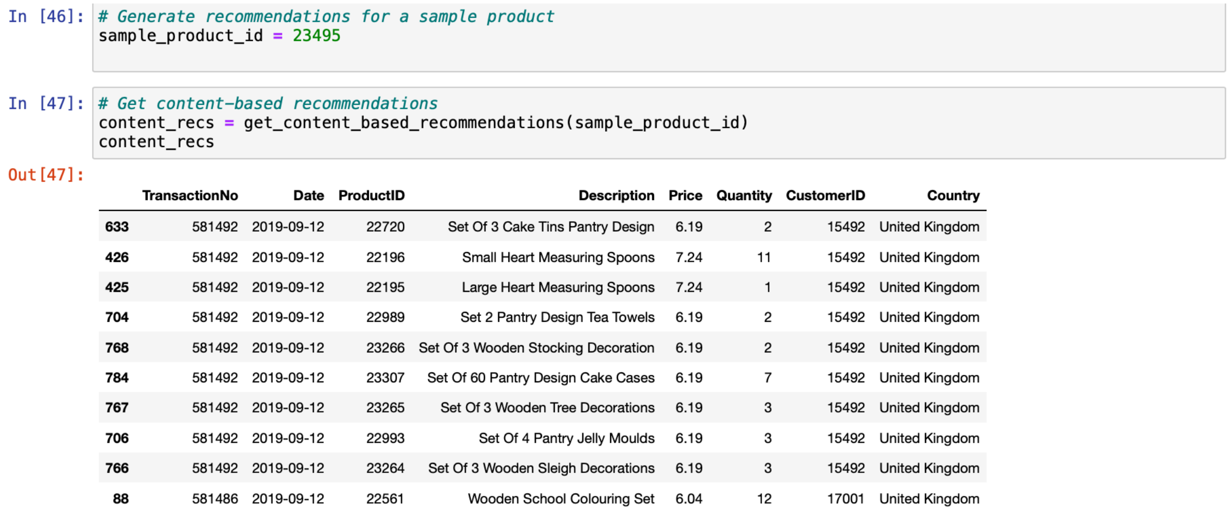
- Fill missing descriptions with empty strings.
- Initialise the TF-IDF vectoriser to ignore common English stop words.

- Transform the product descriptions into a TF-IDF matrix.
- Compute cosine similarity between items.
- Create a user-item interaction matrix using a pivot table.
- Transpose of the matrix for item-based collaborative filtering.
- Initialise and fit the Nearest Neighbors model using cosine similarity.



```
In [45]: # Define a function to get content-based recommendations
         def get_content_based_recommendations(product_id, cosine_sim=cosine_sim):
             idx = product_data[product_data['ProductID'] == product_id].index[0]
             sim_scores = list(enumerate(cosine_sim[idx]))
             sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
             sim_scores = sim_scores[1:11]
             product_indices = [i[0] for i in sim_scores]
             return product_data.iloc[product_indices]

         # Define a function to get item-based recommendations
         def get_item_based_recommendations(product_id, model_knn=model_knn, item_user_matrix=item_user_matrix):
             idx = item_user_matrix.index.get_loc(product_id)
             distances, indices = model_knn.kneighbors(item_user_matrix.iloc[idx, :].values.reshape(1, -1), n_neighbors=11)
             similar_indices = indices.flatten()[1:]
             similar_products = item_user_matrix.index[similar_indices]
             return product_data[product_data['ProductID'].isin(similar_products)]

         # Define a function to get hybrid recommendations
         def get_hybrid_recommendations(product_id):
             content_recommendations = get_content_based_recommendations(product_id)
             item_recommendations = get_item_based_recommendations(product_id)
             hybrid_recommendations = pd.concat([content_recommendations, item_recommendations]).drop_duplicates().head(10)
             return hybrid_recommendations
```

- Define a function for content-based recommendations using cosine similarity scores.
- Define a function for item-based recommendations using the Nearest Neighbors model.
- Define a function for hybrid recommendations combining content-based and item-based recommendations.



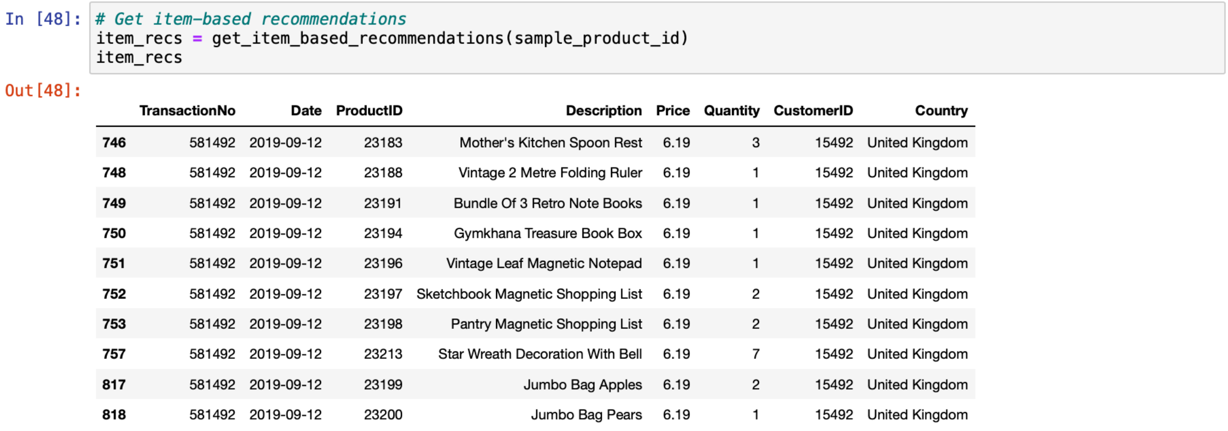
```
In [46]: # Generate recommendations for a sample product
         sample_product_id = 23495
```

```
In [47]: # Get content-based recommendations
         content_recs = get_content_based_recommendations(sample_product_id)
         content_recs
```

Out[47]:

	TransactionNo	Date	ProductID	Description	Price	Quantity	CustomerID	Country
633	581492	2019-09-12	22720	Set Of 3 Cake Tins Pantry Design	6.19	2	15492	United Kingdom
426	581492	2019-09-12	22196	Small Heart Measuring Spoons	7.24	11	15492	United Kingdom
425	581492	2019-09-12	22195	Large Heart Measuring Spoons	7.24	1	15492	United Kingdom
704	581492	2019-09-12	22989	Set 2 Pantry Design Tea Towels	6.19	2	15492	United Kingdom
768	581492	2019-09-12	23266	Set Of 3 Wooden Stocking Decoration	6.19	2	15492	United Kingdom
784	581492	2019-09-12	23307	Set Of 60 Pantry Design Cake Cases	6.19	7	15492	United Kingdom
767	581492	2019-09-12	23265	Set Of 3 Wooden Tree Decorations	6.19	3	15492	United Kingdom
706	581492	2019-09-12	22993	Set Of 4 Pantry Jelly Moulds	6.19	3	15492	United Kingdom
766	581492	2019-09-12	23264	Set Of 3 Wooden Sleigh Decorations	6.19	3	15492	United Kingdom
88	581486	2019-09-12	22561	Wooden School Colouring Set	6.04	12	17001	United Kingdom

- Set a sample product ID.
- Generate content-based recommendations for the sample product and display the results.

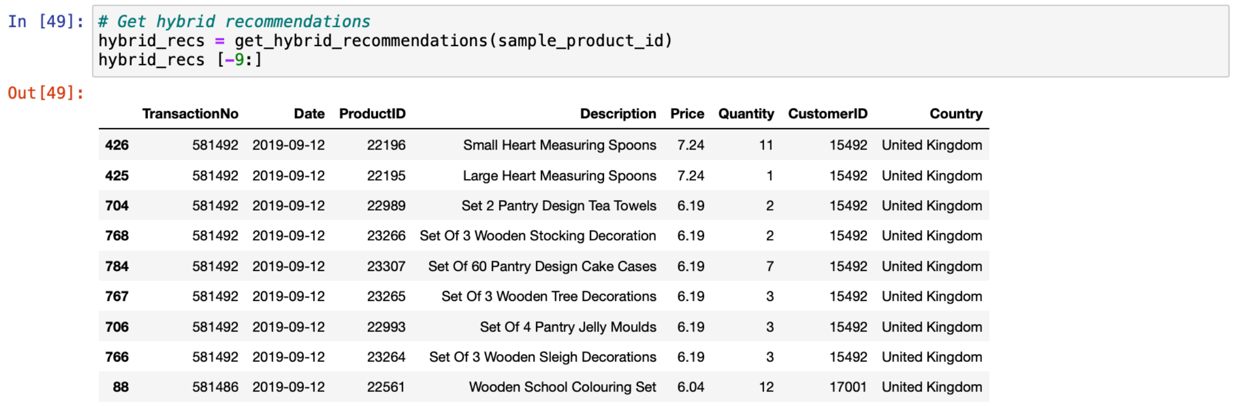


```
In [48]: # Get item-based recommendations
         item_recs = get_item_based_recommendations(sample_product_id)
         item_recs
```

Out[48]:

	TransactionNo	Date	ProductID	Description	Price	Quantity	CustomerID	Country
746	581492	2019-09-12	23183	Mother's Kitchen Spoon Rest	6.19	3	15492	United Kingdom
748	581492	2019-09-12	23188	Vintage 2 Metre Folding Ruler	6.19	1	15492	United Kingdom
749	581492	2019-09-12	23191	Bundle Of 3 Retro Note Books	6.19	1	15492	United Kingdom
750	581492	2019-09-12	23194	Gymkhana Treasure Book Box	6.19	1	15492	United Kingdom
751	581492	2019-09-12	23196	Vintage Leaf Magnetic Notepad	6.19	1	15492	United Kingdom
752	581492	2019-09-12	23197	Sketchbook Magnetic Shopping List	6.19	2	15492	United Kingdom
753	581492	2019-09-12	23198	Pantry Magnetic Shopping List	6.19	2	15492	United Kingdom
757	581492	2019-09-12	23213	Star Wreath Decoration With Bell	6.19	7	15492	United Kingdom
817	581492	2019-09-12	23199	Jumbo Bag Apples	6.19	2	15492	United Kingdom
818	581492	2019-09-12	23200	Jumbo Bag Pears	6.19	1	15492	United Kingdom

Generate item-based recommendations for the sample product and display the results.



```
In [49]: # Get hybrid recommendations
         hybrid_recs = get_hybrid_recommendations(sample_product_id)
         hybrid_recs [-9:]
```

Out[49]:

	TransactionNo	Date	ProductID	Description	Price	Quantity	CustomerID	Country
426	581492	2019-09-12	22196	Small Heart Measuring Spoons	7.24	11	15492	United Kingdom
425	581492	2019-09-12	22195	Large Heart Measuring Spoons	7.24	1	15492	United Kingdom
704	581492	2019-09-12	22989	Set 2 Pantry Design Tea Towels	6.19	2	15492	United Kingdom
768	581492	2019-09-12	23266	Set Of 3 Wooden Stocking Decoration	6.19	2	15492	United Kingdom
784	581492	2019-09-12	23307	Set Of 60 Pantry Design Cake Cases	6.19	7	15492	United Kingdom
767	581492	2019-09-12	23265	Set Of 3 Wooden Tree Decorations	6.19	3	15492	United Kingdom
706	581492	2019-09-12	22993	Set Of 4 Pantry Jelly Moulds	6.19	3	15492	United Kingdom
766	581492	2019-09-12	23264	Set Of 3 Wooden Sleigh Decorations	6.19	3	15492	United Kingdom
88	581486	2019-09-12	22561	Wooden School Colouring Set	6.04	12	17001	United Kingdom

Generate hybrid recommendations for the sample product and display the results.

8.3 Appendix C (Proposal)

1. Project Aims & Objectives

Title: Customer Segmentation & Product Recommendations using data-driven marketing analytics and machine learning techniques in the UK e-commerce sector

1.1 Aims

This project aims to develop an understanding of customer behaviour within the UK e-commerce industry. Using advanced marketing analytics and machine learning techniques, it aims to identify different customer groups and provide personalised product recommendations that can boost customer engagement and sales.

1.2 Objectives

Identify Customer Segments: Utilise advanced data analytics to divide clients into distinct segments based on their behaviours and preferences.

Develop Personalised Recommendations: Develop algorithms that customize product recommendations for certain client segments, improving the overall shopping experience.

Optimise Marketing Strategies: Apply the insights gained from customer segmentation to improve marketing campaigns and strategies, targeting specific customer groups more effectively and efficiently.

Increase Customer Engagement: Improve customer satisfaction and engagement by providing more relevant product recommendations.

1.3 Research Questions

- How can customer segmentation can be effectively implemented in the UK e-commerce industry to identify distinct customer groups?
- How can personalised product recommendations be generated for different customer segments to maximise engagement and sales?
- How does customer segmentation improve the accuracy and relevance of product recommendations in the UK e-commerce sector?

2. Problem Statement

2.1 Background and Context

International Trade Administration (2022) stated that the e-commerce industry in the UK is competitive and experiencing significant growth worldwide, being the third largest behind China and the USA. To effectively give service to their customers, e-commerce businesses must employ complex techniques to understand and meet the needs of the growing number of online shoppers and the wide range of products accessible. Traditional customer segmentation methods, which often depend mainly on demographic data, do not properly capture the complexity of customer behaviours (Marcelo, 2023). To accurately identify consumer segments and personalise product recommendations, this gap requires implementing advanced data-driven approaches.

2.2 Scope of Study

This project's research will focus on using data-driven marketing analytics and machine learning methodologies to improve consumer segmentation and product recommendations within the UK e-commerce field. The research will begin by collecting and analysing secondary data such as customer demographics and behavioural data. The study will conclude by evaluating personalised recommendation systems specifically designed for the identified customer segments. The scope will cover the whole process, including data collection and preprocessing, as well as model development, validation and evaluation. The insights from this research project will be useful for e-commerce businesses, marketers, data scientists, and academic researchers.

3. Literature Overview

USU (2020) demonstrated that when conducting a thorough study into the UK e-commerce industry's customer segmentation and product recommendations, it is crucial to draw on a variety of sources. This literature overview will cover academic articles, journals, business reports, and market research studies that will provide the fundamental knowledge and context for the research.

Academic Articles & Journals offer a strong theoretical basis and valuable knowledge of the most recent developments in research. The main areas of focus will include:

Customer Segmentation Techniques: Research exploring different methods, including clustering algorithms (K-means, hierarchical clustering) and how well they work for segmenting online customers.

Product Recommendation Systems: Primary research on recommendation models, content-based filtering, collaborative filtering and hybrid filtering.

Machine Learning in Marketing: Articles that explore the utilisation of machine learning methods in marketing analytics, with a particular emphasis on improving customer experience.

Business reports and statistics from market research organisations offer practical insights and industry trends. Notable sources include Statista, the eCommerce Database, Kaggle, Global Data Explorer and IBIS World. These platforms offer crucial information for comprehending customer segments and providing recommendations in the e-commerce industry. Relevant studies include research on consumer preferences and purchasing behaviours, specifically focusing on the factors influencing generational differences, device choices and online shopping decisions in the UK. These studies also examine e-commerce product categories and trends, emphasising the most popular categories among online shoppers and the frequency of online purchasing by category.

4. Proposed Methodology

To resolve the research questions and accomplish the project’s objectives, a structured methodology covering various key steps will be implemented. This methodology mainly uses secondary data to get comprehensive and actionable insights, leveraging advanced data analytics and machine learning techniques.

4.1 Data Collection

The study will employ secondary data sources that offer comprehensive information on market, consumer behaviours, transactions, product preferences and interaction metrics. These datasets are categorised into five types, which will be merged to enhance the analysis if necessary.

Customer Transaction Data

- Customer transaction data provides information on the types of products purchased, quantities, discount & return rates and payment methods.

Website Analytics

- Add-to-cart rate
- Cart abandonment rate
- Conversion rate

Market Research Reports

- Industry trends (present & future)
- Customer demographics
- Market Size
- Market structure
- Revenue forecast insights

Customer Demographics

- Age
- Gender
- Income
- Online Behaviours

Technology Usage Data

- Devices
- Payment methods

4.2 Data Reprocessing

Data Cleaning: Clear duplicates, deal with missing values and ensure data consistency.

Data Integration: The process of merging and harmonising data from numerous sources into a unified and consistent format, which may then be utilised for various analytical, operational, and decision-making purposes.

Data Reduction: Minimise the volume of data while preserving its essential information.

Data Transformation: An important part of data preprocessing, wherein raw data is turned into a unified format or structure.

4.3 Machine Learning Techniques

Customer Segmentation

- Employ the Elbow Method & calculate the silhouette score using sklearn to determine K.
- Choose the optimal number of clusters, K.
- Distribute clusters based on the results of the above analysis.
- Do the cluster analysis

Product Recommendations

- Collaborative Filtering: Utilise user-based and item-based collaborative filtering techniques to recommend products based on user preferences and behaviour.
- Content-Based Filtering: Use the product's distinctive features to generate customised recommendations.
- Hybrid Filtering: Combine collaborative & content-based filterings to make better recommendations (Maruti Techlabs, 2017)

4.4 Data Analysis Approach

Exploratory data analysis (EDA) will be the first step in the data analysis process. This will help understand the main insights of the datasets. EDA will involve:

- Descriptive Statistics: Conduct statistical analysis to determine the mean, median, mode, standard deviation and other relevant summary statistics to comprehend the data distribution.
- Data Visualisation: Develop visualisations including scatter plots, line charts, graphs and histograms to analyse the data for patterns and trends.
- Identify Outliers: Identify any outliers and flaws that may impact the analysis and determine suitable methods to handle them. (IBM, 2020)

4.5 Foreseeable Limitations

The **quality and accuracy of the secondary data** are critical to the analysis's reliability. Attempts will be made to clean and validate the data, however, the outcomes may be influenced by data quality limitations.

There is also a risk of introducing **biases** in the machine learning models which may result in unfair treatment of certain segments. Implementing regular checks and employing different datasets will effectively mitigate the risk.

Moreover, the findings may not be universally applicable across all e-commerce platforms as there are differences in customer bases and product offerings. In an effort to enhance the **generalisability** of the findings, representative datasets will be implemented.

5. Ethical Challenges & Risks

Halej (2017) & Rana et al. (2021) stated that managing consumer data presents significant privacy concerns and may lead to the misuse of sensitive data, including browsing behaviour, purchase history, and personal identifiers. To manage this, it is important to make sure that any personal identifiers are removed or anonymized prior to research and analysis, protecting individual privacy. Additionally, strict access control procedures will be implemented to ensure that only those with permission can access the data. To maintain legal compliance and protect customer confidentiality, adherence to data protection laws and regulations, such as the UK General Data Protection Regulation (GDPR), is important.

The secondary data collection process is specifically designed to mitigate bias by directly gathering information from government agencies and reputable international organisations such as GlobalData Explorer, IBIS World, Statista and so on to ensure that the data is representative and unbiased. In order to prevent any potential conflicts that could affect the research purpose or the implementation of results, personal or financial interests are strictly avoided.

Maintain transparency about methodologies and algorithms used in this project. As mentioned above, secondary data will be gathered only from trusted sources. However, by any chance, if an individual data collection is necessary, the study will employ strict measures to protect privacy and confidentiality. Data will be handled, stored and analysed exclusively for research purposes and ethical approval and informed consent will be obtained.

ChatGPT will be employed to gather key information from research reports and academic journals and help with data cleaning and preprocessing in python. It will help ensure coherence and clarity while drafting literature reviews and evaluating findings. By leveraging this AI tool, this research project aims to enhance the depth and effectiveness of the analysis, providing solid and useful insights for product recommendation and customer segmentation in the UK e-commerce sector.

6. Executive Summary Presentation & Gantt Chart for Project Timeline

The Executive Summary will be presented as a brief written summary. The article will summarise the main findings, insights, and recommendations from the research, ensuring the information is readily accessible and comprehensible. The written summary will ensure that the key components of the project are effectively communicated, offering an in-depth overview of the results and their implications for the UK e-commerce industry.

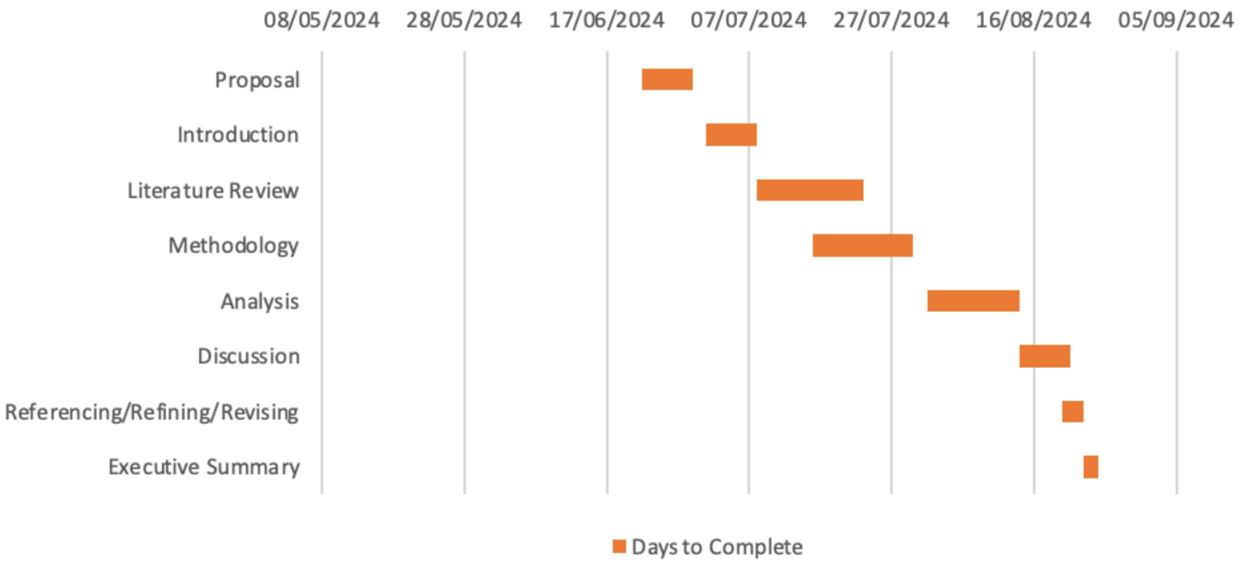


Figure: Gantt Chart for Project Timeline

7. Reference List

Halej, J. (2017). *Ethics in primary research (focus groups, interviews and surveys)*. Equality Challenge Unit. https://forms.docstore.port.ac.uk/A816773.pdf

IBM. (2020). *What Is Exploratory Data Analysis? | IBM*. Www.ibm.com. https://www.ibm.com/topics/exploratory-data-analysis

International Trade Administration. (2022, September 12). *United Kingdom - eCommerce*. Www.trade.gov. https://www.trade.gov/country-commercial-guides/united-kingdom-ecommerce

Marcelo, M. (2023, October 31). *Building a Deeper Understanding of the Customer Through a Values-Based Customer Segmentation*. Www.greenbook.org. https://www.greenbook.org/insights/focus-on-apac/building-a-deeper-understanding-of-the-customer-through-a-values-based-customer-segmentation#

Maruti Techlabs. (2017, September 28). *What are Product Recommendation Engines? And the various versions of them?*Medium; Towards Data Science. https://towardsdatascience.com/what-are-product-recommendation-engines-and-the-various-versions-of-them-9dcab4ee26d5

Rana, J., Dilshad, S., & Ahsan, Md. A. (2021). Ethical Issues in Research. *Global Encyclopedia of Public Administration, Public Policy, and Governance*, 1–7. https://doi.org/10.1007/978-3-319-31816-5_462-1

USU. (2020). *LibGuides: Variety of Sources: Home*. Libguides.usu.edu. https://libguides.usu.edu/variety