FDA PROJECT-2

IE6400 - Foundations Data Analytics Engineering

Final Report

Group Number 21

Group Members

Saiteja Reddy Gajula (002872000)

Pooja Arumugam (002872003)

Bhakti Paithankar (002833722)

Nihal Mallikarjun (0036010381)

Sathvik Ramappa (002772175)

PART 1: INTRODUCTION

E-commerce has transformed the way people conduct business and has grown to be an essential component of modern life. Comprehending the workings of this virtual marketplace is essential to both adjusting to the changing economic environment and guaranteeing the effectiveness and safety of online transactions. Market trends, buying patterns, and customer behaviour can all be learned using real-world datasets that contain e-commerce data.

It is impossible to overestimate the importance of e-commerce data analysis in this age of data-driven decision-making. The integrity of e-commerce databases, which cover a broad range of goods and services, is the main goal of this research. Assuring cybersecurity and improving the online shopping experience are the main objectives, and they will be achieved through data organization, cleaning, and smart analysis that may guide company strategy and policy decisions.

Data visualization is essential in communicating patterns, trends, and anomalies that may be less obvious from raw data. By acting as a bridge between raw data and useful insights, data visualization helps to create a more secure and effective e-commerce environment.

This project employs a noteworthy methodology in the form of RFM (Recency, Frequency, Monetary) segmentation. Through RFM segmentation, the analysis examines consumer transaction data to determine the frequency, recentness, and monetary value of transactional activity. Through the creation of tailored marketing strategies, the identification of high-value clients, and the optimization of overall business performance, this nuanced approach improves the understanding of customer categories.

The initiative helps to create a more secure and effective e-commerce environment by graphically illustrating patterns, trends, and anomalies that might not be immediately apparent in raw data. The ultimate objective is to use information to understand how e-commerce functions and develop a digital marketplace that is better, more secure, and specifically designed to meet the demands of businesses and customers.

PART 2: SUMMARY OF RESULTS

Through the examination of an online retail store's ecommerce dataset, we were able to derive a substantial number of insightful conclusions. RFM (Recency, Frequency, Monetary) analysis was used to investigate customer behavior, looking at things like the distribution of buy times throughout the day, the recentness of transactions, and the frequency of customers with respect to the monetary worth of their purchases. In order to shed light on patterns in customer interactions, we were able to translate the raw data throughout the project into understandable data visualization techniques. By integrating the visualization, customer loyalty trends and characteristics were effectively depicted.

K-means clustering was used to segment the client base based on RFM ratings, revealing distinct and significant clusters within the consumer group. The report aimed to optimize income and engagement by grouping customers and offering specific recommendations for marketing strategies for each group. The results aid in comprehending the dynamics of customers and provide businesses with strategic recommendations to maximize client relationships, retention, and profitability.

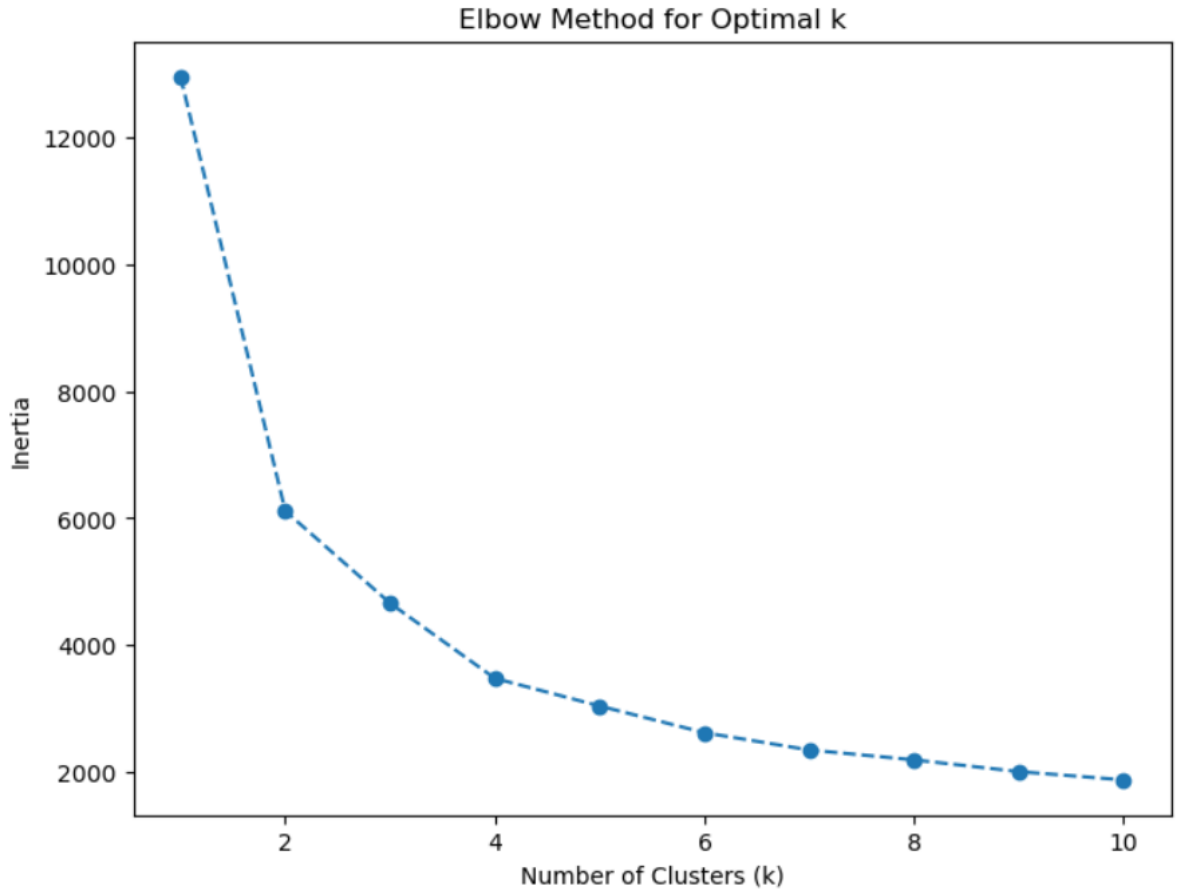
In summary, a knowledge of customer behaviour and its trends and characteristics was made possible by the integration of RFM analysis, k-means clustering, and data visualization tools. By attaining consistent expansion and boosting their competitive market environment, these findings will improve the retail store's functionality and assist them in optimizing their entire strategy.

PART 3: DATA SOURCES

The main portion of data used in this analysis was sourced from https://www.kaggle.com/datasets/carrie1/ecommerce-data

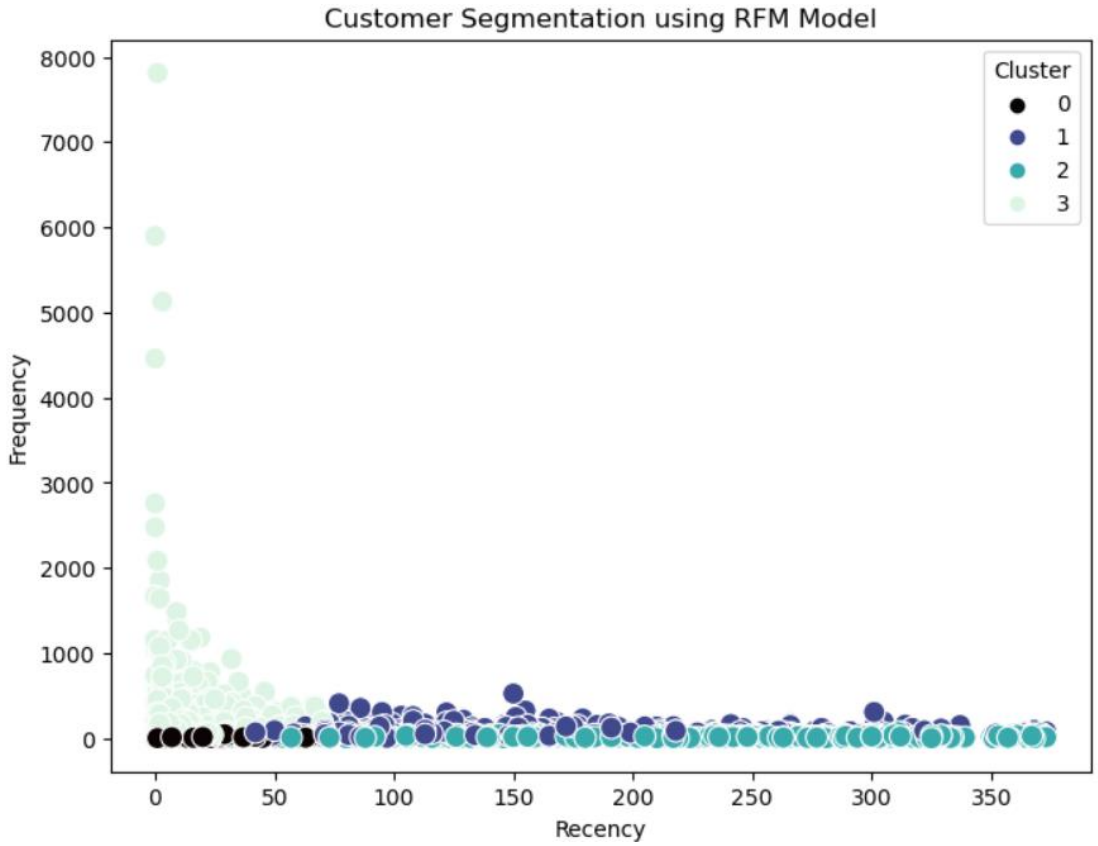
PART 4: RESULTS AND METHODS

The Elbow Method, a method for figuring out the ideal number of clusters (k) in a clustering algorithm, is depicted in the plot. The number of clusters is represented by the x-axis, and the within-cluster sum of squares (WCSS), a gauge of the compactness of clusters, is represented by the y-axis. In general, the WCSS value falls as the number of clusters rises. The Elbow Method suggests that adding more clusters beyond this point delivers diminishing gains in terms of enhancing cluster cohesion. It identifies a "elbow" point on the plot when the pace of WCSS reduction slows down. The plot acts as a visual guide to help determine the ideal number of clusters for best clustering results. See where the WCSS begins to decline more slowly in the ensuing plot.



The Elbow Method uses a scatter plot to analyse RFM (Recency, Frequency, Monetary) clustering data. In this case, the monetary value is shown on the y-axis, while the recentness of transactions is shown on the x-axis. To determine the ideal clustering configuration, the Elbow Method plots the within-cluster sum of squares (WCSS) versus the total number of clusters.

Different RFM clusters are represented by distinct data points in the scatter plot, and their placements signify both monetary and recency values. The Elbow Method looks for the location on the plot where the WCSS decreases to a point that resembles a "elbow." This is the ideal number of clusters that minimize within-cluster variability while preventing undue fragmentation.



.

1. Recency Distribution Plot: This type of plot shows how a dataset's recency values are distributed. The number of occurrences is shown on the y-axis, and various recency intervals are indicated on the x-axis. The plot gives an overview of the transactions and activities that customers have participated in recently. Plot features such as peaks and patterns highlight trends in recency and provide information about how frequently recent interactions occur in the dataset under analysis.

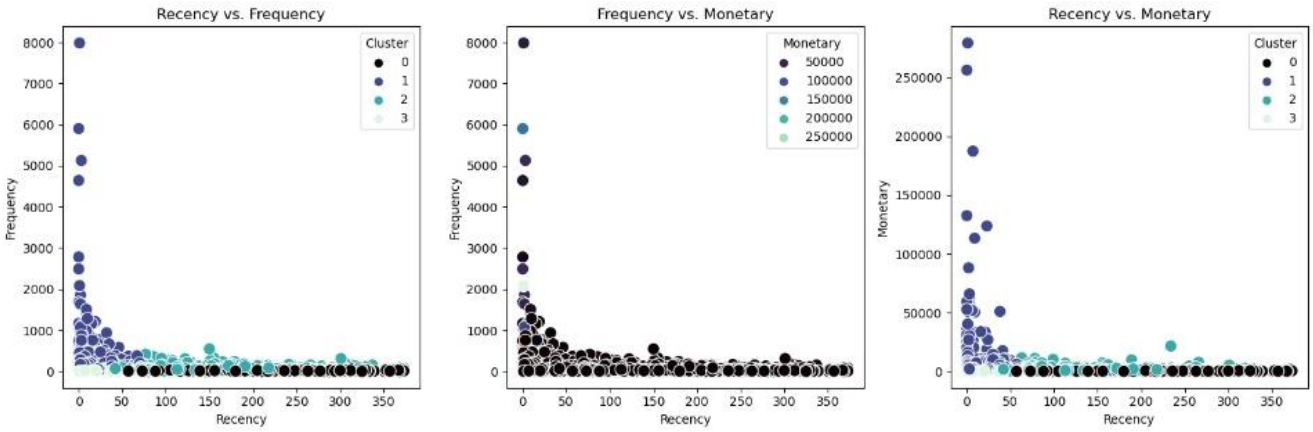
2. Frequency Distribution Plot: This graphic shows how transaction frequencies are distributed over a dataset. The x-axis shows various frequency intervals, while the y-axis shows the total number of occurrences. Plot peaks and varies show trends in the behaviour of consumer transactions. Understanding how frequently clients interact with a platform, service, or product using this visualization can assist identify customer categories based on transaction frequency.

3. Monetary Distribution Plot: The distribution of monetary values in a dataset is shown visually by the monetary distribution plot. The x-axis displays various monetary intervals, and the y-axis displays the number of occurrences. Plot features such as peaks and patterns shed light on how customers' spending varies. Understanding the volatility in customer expenditure and locating high-value dataset parts are two areas in which this visualization is especially helpful.

1. Recency vs. Frequency Plot: This plot juxtaposes recency against frequency, with the y-axis representing the frequency of transactions and the x-axis indicating the recency of those transactions. Each point on the plot represents a unique combination of recency and frequency for a particular set of data. This visualization provides insights into the relationship between how recently customers have engaged and how often they do so. Patterns in the plot help identify segments of customers based on both recency and frequency, aiding in targeted marketing and engagement strategies.

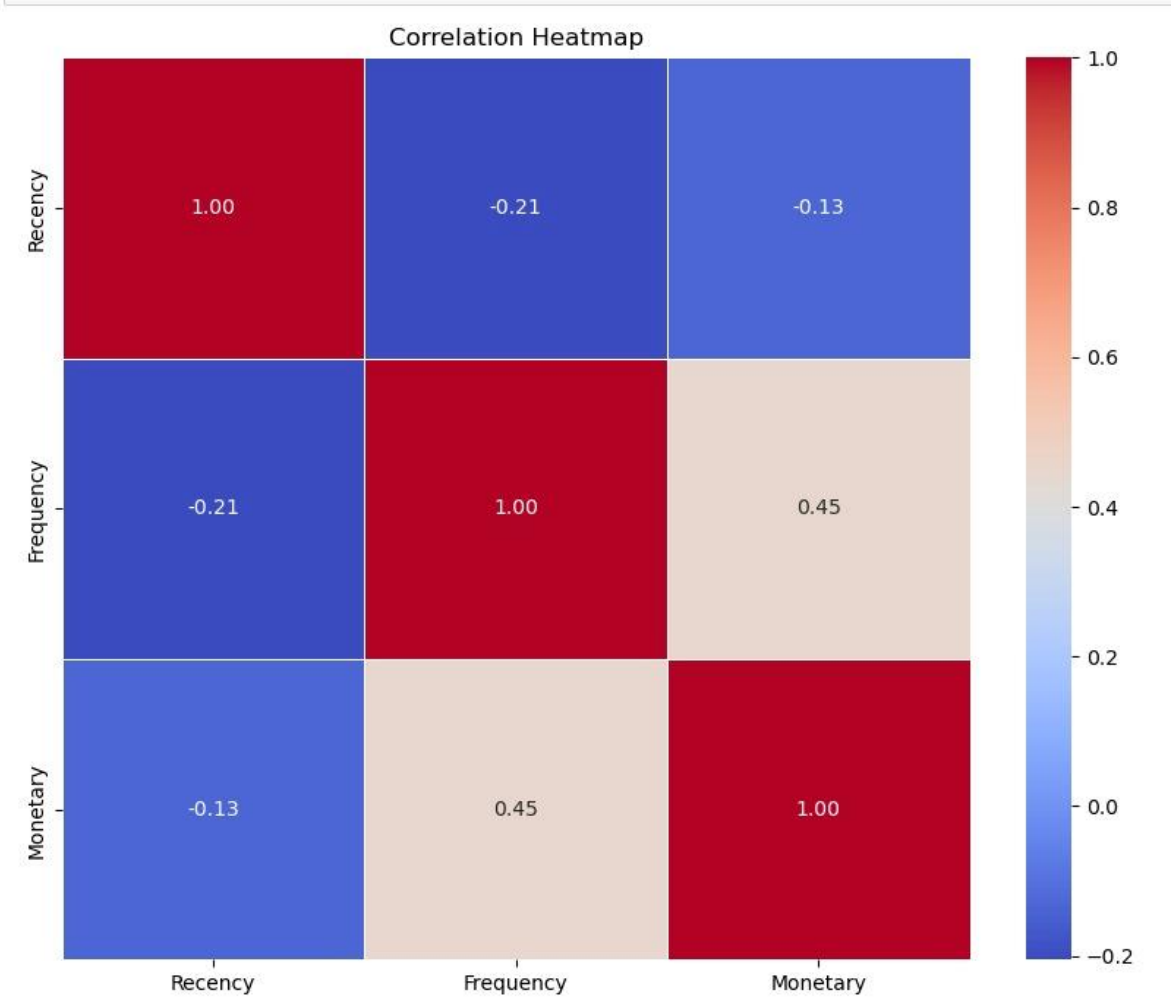
2. Frequency vs. Monetary Plot: The frequency vs. monetary plot compares the frequency of transactions (y-axis) with the monetary value of those transactions (x-axis). Each point on the plot represents a specific combination of transaction frequency and monetary expenditure. This visualization is valuable for understanding the correlation between how often customers make purchases and the monetary value associated with those transactions. Patterns in the plot can reveal distinct customer segments, guiding strategies for customer engagement and revenue optimization.

3. Recency vs. Monetary Distribution Plot: This plot illustrates the distribution of monetary values (y-axis) concerning the recency of transactions (x-axis). Each point on the plot represents a specific combination of recency and monetary value. The visualization provides a comprehensive view of how recent transactions correlate with their respective monetary amounts. Peaks and patterns in the plot offer insights into customer spending behavior over time, guiding decisions related to personalized marketing or loyalty programs based on recency and monetary considerations.

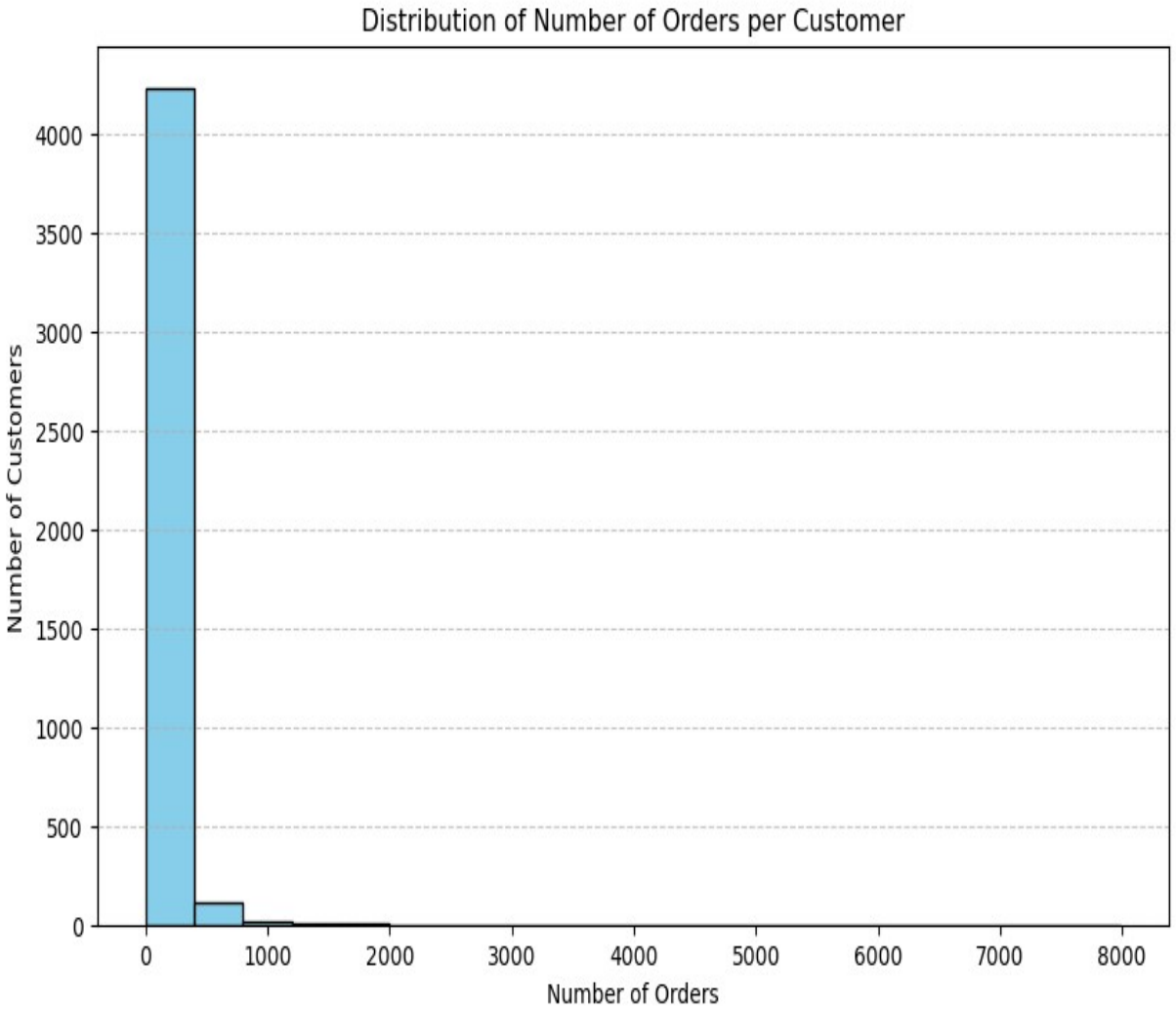


K-means clustering is a machine learning algorithm used for partitioning a dataset into distinct, non-overlapping groups or clusters based on similarity. It iteratively assigns data points to clusters by minimizing the sum of squared distances between data points and the centroid of their assigned cluster. K-means clustering is a powerful technique that can be applied to categorize customers based on RFM analysis features, utilizing the customer count as the X-axis and clusters as the Y-axis. The RFM features—Recency (R), Frequency (F), and Monetary Value (M)— serve as crucial dimensions for clustering. The algorithm iteratively groups customers into clusters by minimizing the squared distances between their RFM values and the centroid of their assigned cluster. The resulting visualization on the X-axis provides a comprehensive view of RFM distribution, while

the Y-axis showcases distinct clusters representing groups of customers with similar purchasing patterns. This graphical representation enables businesses to identify and target specific customer segments effectively, tailoring marketing strategies to address the unique needs and behaviours of each cluster.

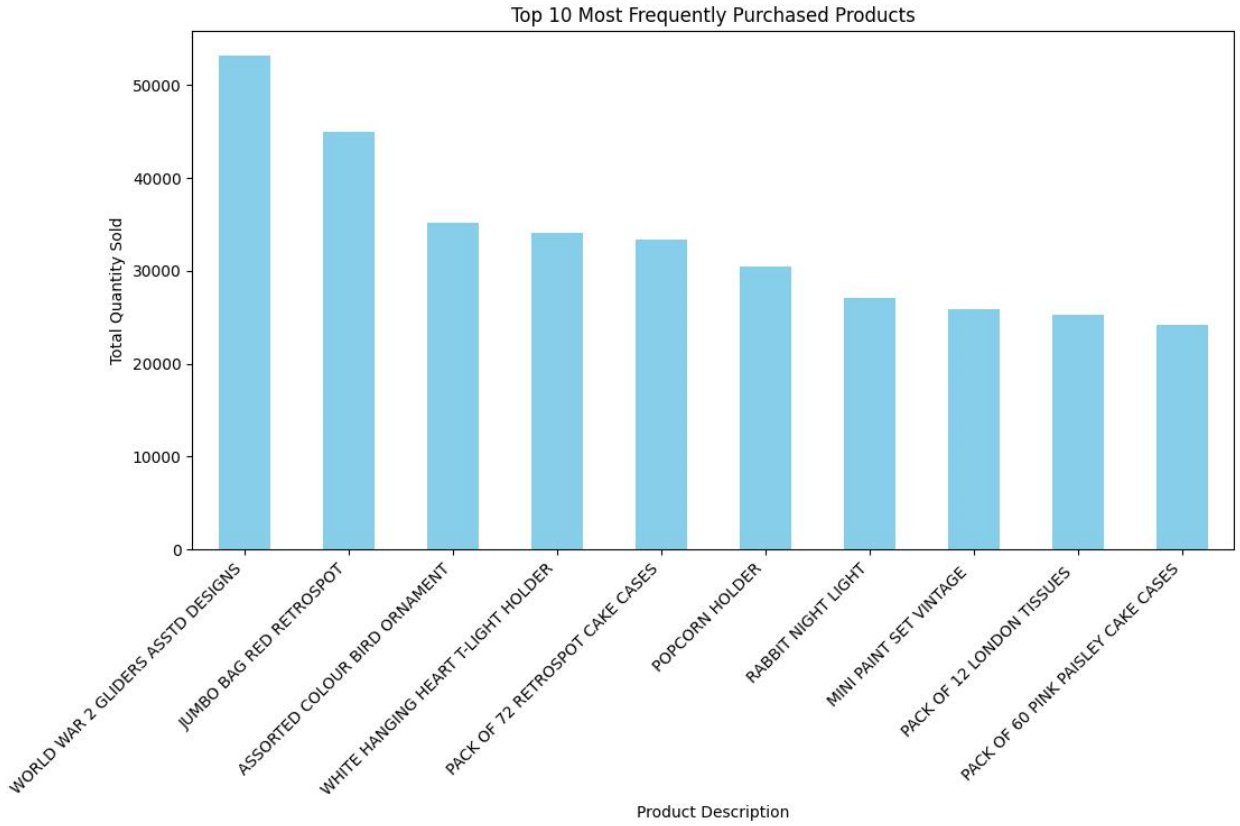


A histogram is constructed to illustrate the distribution of orders per customer. The X-axis represents the number of orders, grouped by unique CustomerID and Invoice Number combinations, while the Y-axis indicates the frequency of customers falling into each order count category. This histogram provides a visual representation of customer purchasing behavior, highlighting the distribution of customers based on the frequency of their orders. Peaks in the histogram indicate concentrations of customers with specific order counts, offering insights into purchasing patterns and potential segments within the customer base. This visualization facilitates a nuanced understanding of customer engagement, aiding businesses in tailoring their strategies to cater to various customer segments based on their ordering frequency.

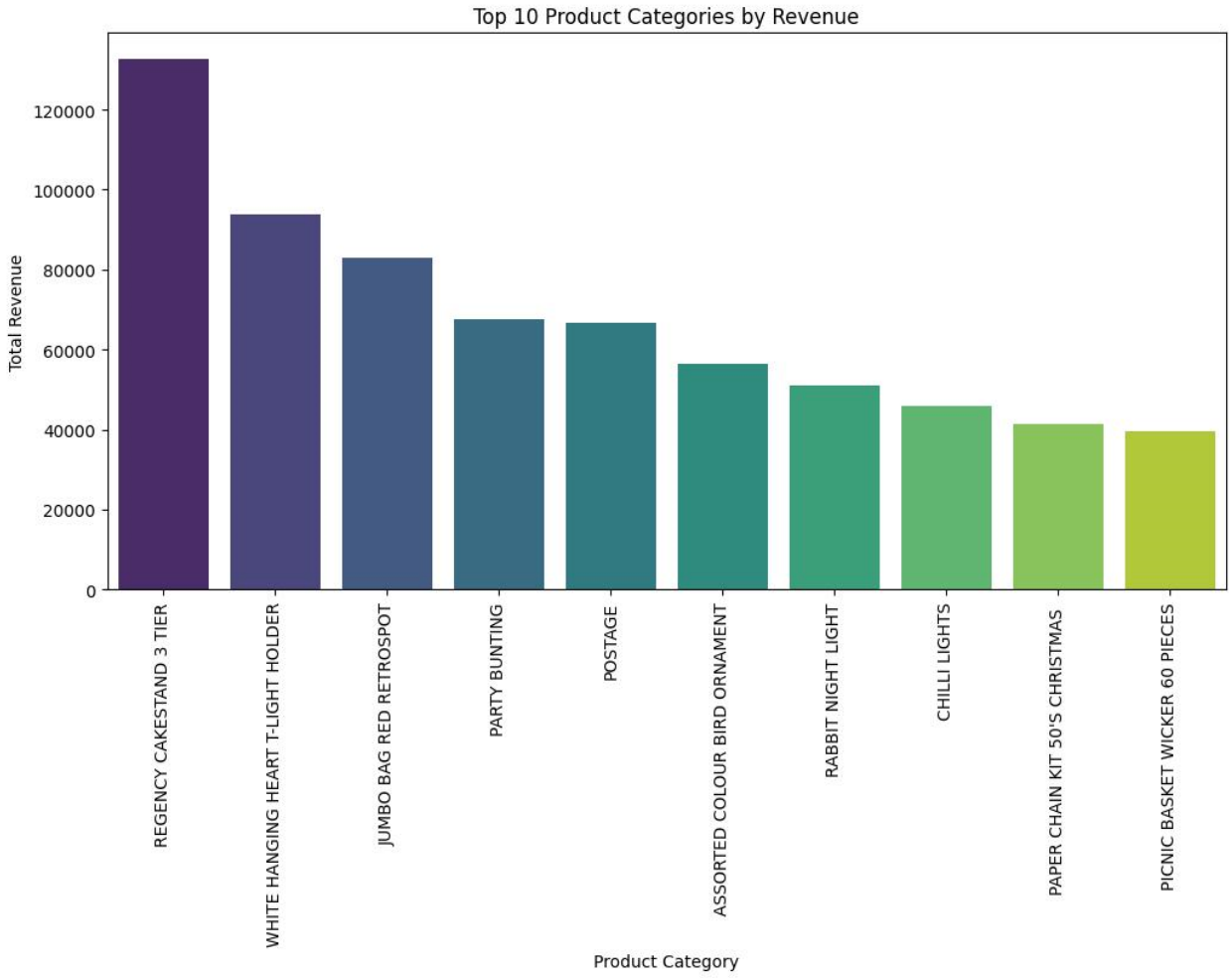


The following bar graph highlights the Top 10 Most Frequently Purchased Products. The X-axis of the graph represents the product descriptions, and the Y-axis depicts the total quantity sold for each product. By examining the graph, one can easily discern the product with the highest sales, such as the WW2 Gliders design.

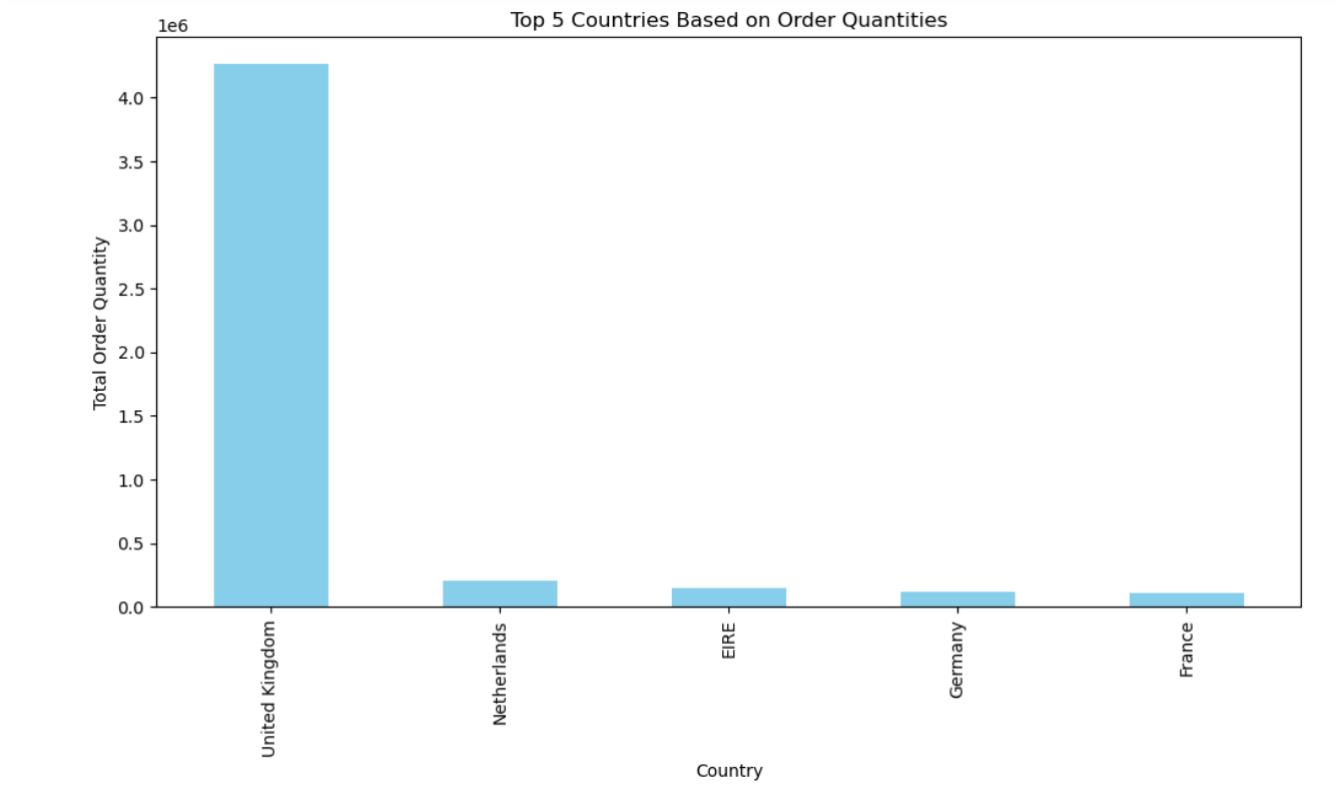
The bar graph visually showcases the relative popularity of different products, offering a clear comparison of their sales performance. The length of each bar corresponds to the total quantity of a particular product sold, making it evident that the WW2 Gliders design has garnered the highest sales among the top 10 products. This insightful visualization aids businesses in identifying best-selling items, informing inventory management, and guiding marketing strategies to capitalize on the popularity of specific products.


Top 10 Most Frequently Purchased Products
Total Quantity Sold
0
10000
20000
30000
40000
50000
WORLD WAR 2 GLIDERS ASSTD DESIGNS
JUMBO BAG RED RETROSPOT
ASSORTED COLOUR BIRD ORNAMENT
WHITE HANGING HEART T-LIGHT HOLDER
PACK OF 72 RETROSPOT CAKE CASES
POPCORN HOLDER
RABBIT NIGHT LIGHT
MINI PAINT SET VINTAGE
PACK OF 12 LONDON TISSUES
PACK OF 60 PINK PAISLEY CAKE CASES
Product Description

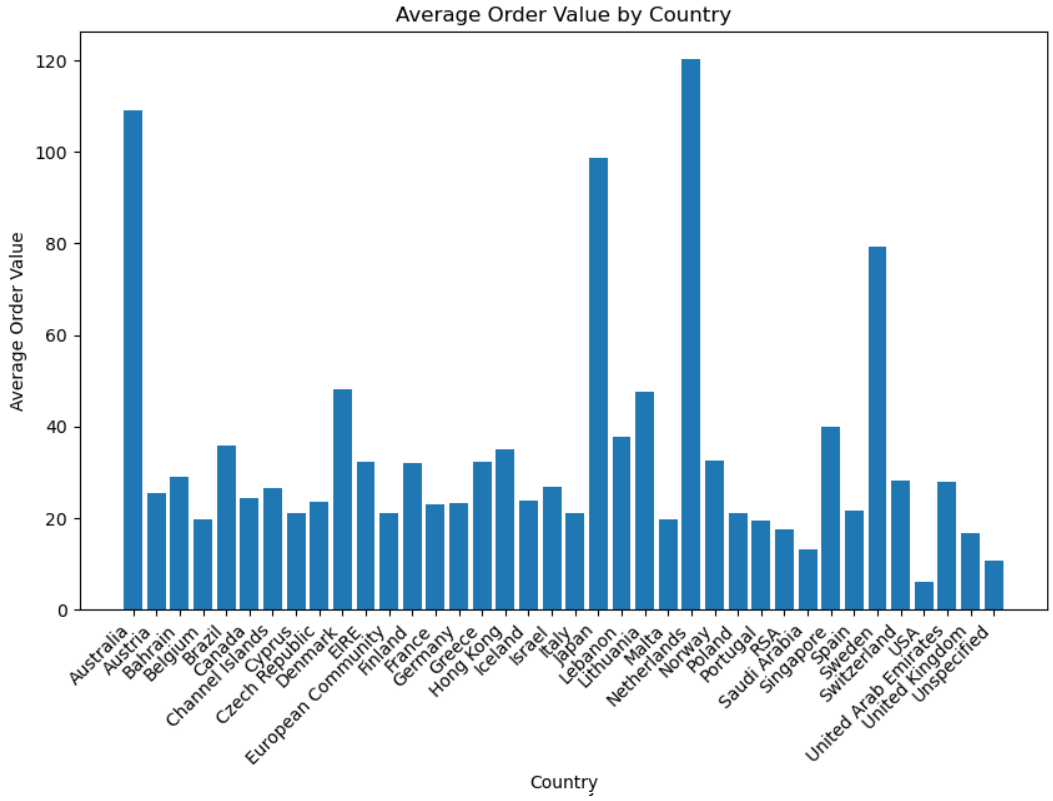
This bar graph depicts the Top 10 product categories by revenue The X-axis displays the total revenue, obtained through the summation of prices for each product category, and the Y-axis represents the product categories. Upon examination, it becomes evident that the Regency Cakestand 3 Tier has achieved the highest total revenue among the top 10 product categories. This bar graph serves as a powerful visual representation of revenue distribution across different product categories. The length of each bar corresponds to the total revenue generated by a specific product category, allowing businesses to quickly identify and focus on high-performing categories.



The bar graph illustrating the Top 5 countries with the highest number of orders provides a succinct overview of the distribution of orders across different nations. With the Y-axis denoting the number of orders and the X-axis representing countries, it becomes evident that the United Kingdom has secured the highest position, signifying the largest volume of orders among the top five countries. This visual representation not only highlights the significance of the United Kingdom in terms of order frequency but also serves as a valuable tool for businesses to identify key markets and allocate resources strategically.



The bar graph depicting the correlation between countries and average order value, with the X-axis indicating countries and the Y-axis representing the average order value, reveals insightful patterns. In this analysis, Australia emerges as having the highest correlation, signifying a strong relationship between the country and the average order value rate. This finding suggests that customers in Australia tend to generate higher average order values compared to other countries in the dataset. The graph provides a clear visual representation of the varying economic contributions of different countries to the business, allowing for strategic decision-making.



PART 5: CONCLUSION

In conclusion, this project focuses on the field of E-commerce, recognizing its importance in today's evolving economic market and landscape. Through a thorough analysis of real-world datasets of an online retail store, focusing on data cleaning, organization and insightful analysis, the project aims to contribute to a safer and more efficient online marketplace.

The integration of RFM analysis, K-means clustering, and data visualization techniques proves to be a powerful resource in unraveling the complexities of customer behavior, trends, and characteristics. By segmenting customers based on RFM scores, actionable recommendations for targeted marketing strategies are presented, providing a workflow for businesses to maximize engagement and revenue. The findings offer a comprehensive understanding of customer dynamics, guiding businesses towards relationships, retention, and profitability.

The successful translation of raw data into visual representations not only illuminates customer loyalty trend but also gives a deeper understanding of their purchasing patterns and the factors associated with them. To sum up, this project serves as a valuable resource for the online retail store, empowering them to navigate the competitive marketplace with sustained growth and an enhanced overall approach to create more secure and customer-centric online shopping experiences.