UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEM

PRACTICUM REPORT

LEVERAGING CUSTOMER SEGMENTATION AND MARKET BASKET ANALYSIS TO DRIVE PERSONALIZED PRODUCT RECOMMENDATIONS IN RETAIL INDUSTRY

Instructor: MSc. Nguyen Quang Hung

Student	Student code	Email
Dinh Thi Be Nhan	K214111949	nhandtb21411@st.uel.edu.vn

Ho Chi Minh City, June 2024

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to the Faculty of Information System at the University of Economics and Law. Thank my faculty for organizing a wide array of informative programs that have helped me gain a deeper understanding of the role of Data Analyst and other key career paths within the field. The insights and guidance provided through these programs have been invaluable in shaping my academic and profession.

Secondly, I would also like to extend my heartfelt thanks to my instructor, MSc.Nguyen Quang Hung. His dedicated assistance and guidance have had a strong, positive impact on the success of this report. I am sincerely grateful for his commitment to supporting and empowering students like myself.

Thirdly, I would like to express my gratitude to MSc. Le Ba Thien for his valuable support and contributions that have also led to the success of this report. His insights and mentorship have been instrumental in shaping the final outcome.

Last but not least, I would like to sincerely thank the researchers and authors of the published works cited in this research study. Their valuable resources and related knowledge have been instrumental in the successful completion of this project.

Although I spent all my effort and dedication to complete this project, I understand that there may still be some limitations. Because I am aware of my lack of skills and knowledge when hands-on with this project myself, I welcome and greatly appreciate any feedback and suggestions from my respected instructors to help me improve and enhance the quality of this work.

COMMITMENT

As the author of this report, I, Dinh Thi Be Nhan, declare that the work presented herein is entirely my own. I can confidently confirm that this report has been written solely by me and has not been submitted in any form for another degree or diploma at this university or any other institution of higher education.

Any information or ideas drawn from the published works of others have been duly acknowledged in the text, and a comprehensive list of references has been provided. Wherever I have consulted the work of others, the source has always been clearly attributed.

I understand that any false claim regarding the originality or ownership of this work will result in disciplinary action, in strict accordance with the university's policies. I take full responsibility for the integrity of the content presented in this report.

Ho Chi Minh City, June 2024

Dinh Thi Be Nhan

EVALUATION FORM FOR MENTOR

Student: Dinh Thi Be Nhan

Student code: K214111949

Representative: MSc. Le Ba Thien

- Student has a pretty good techical skill ;
- Student shows a strong dedication on this report;
- The report demonstrates a thorough understanding of the key concepts covered in the assignment;
- The analysis provided in the report is well-researched and supported by relevant data and examples;
- The writing is clear, concise, and easy to follow. The structure of the report is logical and coherent.

……
……
……
……
……
……
……
……
……
……

Ho Chi Minh City, June 2024

Confirmation of mentor

Le Ba Thien

EVALUATION FORM FOR INSTRUCTOR

Student: Dinh Thi Be Nhan

Student code: K214111949

Instructor: MSc.Nguyen Quang Hung

No.	Evaluation Criteria	Specific Evaluation Criteria	Score	Note
1	**Report format (15%)**	Presentation (5%)		
		Report structure(5%)		
		Writing style (5%)		
2	**Report content (30%)**	Analytical skill (5%)		
		Object (10%)		
		Expertise (15%)		
3	**Student's attitude (15%)**			
4	**Business evaluation (40%)**			*Instructor convert from business evaluation*
TOTAL SCORE				

Ho Chi Minh City, June 2024

Instructor

TABLE OF CONTENTS

ACKNOWLEDGMENTS i

COMMITMENT ii

EVALUATION FORM FOR MENTOR iii

EVALUATION FORM FOR INSTRUCTOR iv

LIST OF TABLES ix

LIST OF EQUATIONS x

ABBREVIATIONS xi

CHAPTER 1:DATA ANALYST- AN INTRODUCTION TO THE PROFESSION 1

1.1 The overview of Vietnam Data Analytics Market 1

1.2 Who are data analysts? 1

1.3 What are the responsibilities of a data analyst? 1

1.4 The skill requirements for a data analyst 2

1.5 Career paths and growth opportunities for data analysts 3

CHAPTER 2: INTRODUCTION OF THE PROJECT 5

2.1 Reason for choosing the topic 5

2.2 Objectives of the project 6

2.2.1 General objectives 6

2.2.2 Specific objectives 6

2.3 Object and scope of the research 7

2.3.1 Research object 7

2.3.2 Research scope 7

2.4 Research methodology 7

2.4.1 Qualitative research methods 7

2.4.2 Quantitative research methods 8

2.5 Research contributions 8
2.5.1 Scientific significance 8
2.5.2 Practical significance 8
2.6 Structure of the project 8
CHAPTER 3: THEORETICAL BACKGROUND AND RELATED WORK 11
3.1 Market basket analysis 11
3.2 Association Rule Mining and Apriori algorithm 12
3.2.1 Association Rule Mining 12
3.2.2 Apriori algorithm 12
3.3 RFM Model 14
3.4 Overview of related research 16
CHAPTER 4: METHODOLOGY AND PROPOSED RESEARCH MODEL 17
4.1 The research process 17
4.2 Data description 18
4.2.1 Data context 18
4.2.2 Attribute description 18
CHAPTER 5: EXPERIMENTAL RESULT AND RECOMMENDATIONS 20
5.1 Preprocessing data 20
5.2 Exploratory Data Analysis - EDA 22
5.3 Customer segmentation using RFM 26
5.4 Market Basket Analysis 30
5.5 Product Recommendation 32
CHAPTER 6: CONCLUSION AND DEVELOPMENT DIRECTION 36
6.1 Conclusion 36
6.2 Limitations of the research and future development directions 36

REFERENCES..... 37

LIST OF FIGURES

Figure 1.1. The most essential nontechnical skills for data analysts 3

Figure 4.1.The research process diagram 17

Figure 5.1.Monthly Sales Revenue and Quantity 22

Figure 5.2.Weekly Sales Revenue and Quantity 23

Figure 5.3.Day of the Week Sales Revenue and Quantity 23

Figure 5.4.Percentage for Time of Day Sales Revenue and Quantity 24

Figure 5.5.Product per quantity 24

Figure 5.6.Product per revenue 25

Figure 5.7.Product by Volume Quantity 26

Figure 5.8.Product by Sales Revenue 26

Figure 5.9.Elbow method result 28

Figure 5.10.Silhouette score result 28

Figure 5.11.RFM metric distributions across customer segments 29

LIST OF TABLES

Table 4.1.Describe data attributes 18

Table 5.1.Overview of dataset before preprocessing data 20

Table 5.2.Overview of dataset after preprocessing data 21

Table 5.3.Quartiles description in RFM table 27

Table 5.4.Each cluster description 28

Table 5.5.Data reshaping for Market Basket Analysis 30

Table 5.6.Data after one-hot encoding 30

Table 5.7.Association rules 31

Table 5.8.Product Recommendations for Platinum customers 32

Table 5.9.Product Recommendations for Gold customers 34

LIST OF EQUATIONS

Equation 3.1. Support equation 13

Equation 3.2.Confidence equation 13

Equation 3.3.Lift equation 13

Equation 3.4.Within-cluster sum of squares equation 15

Equation 3.5.Silhouette equation 15

ABBREVIATIONS

CAGR	Compound Annual Growth Rate
SQL	Structured Query Language
RFM	R – Recency, F- Frequency, M – Monetary Value
MBA	Market Basket Analysis
EDA	Exploratory Data Analysis
ARM	Association Rule Mining

CHAPTER 1 :DATA ANALYST- AN INTRODUCTION TO THE PROFESSION

1.1 The overview of Vietnam Data Analytics Market

According to Vietnam Data Analytics Market Report, Vietnam data analytics market size is projected to exhibit a growth rate (CAGR) of 9.80% during 2024-2032. The growth of data generated by individuals and organizations, the rapid advancements in data analytics technologies, the increasing adoption of business intelligence tools and platforms, and the growing demand for real-time analytics are some of the factors propelling the market .

As the Vietnam data analytics market size continues to grow, there is a corresponding increase in demand for skilled professionals in the field. Data analytics has become integral to businesses across various industries, driving the need for individuals who can effectively analyze and interpret data to derive valuable insights. Among these professionals, data analysts play a crucial role in helping organizations make data-driven decisions.

1.2 Who are data analysts?

A data analyst is a person whose job is to gather and interpret data in order to solve a specific problem. Data analyst has the responsibility to turn raw data into meaningful insights. Therefore, a data analyst is a professional who collects, processes, and analyzes large sets of data to extract meaningful insights and support decision-making.

1.3 What are the responsibilities of a data analyst?

Data analysts play a crucial role in helping organizations make data-driven decisions, improve operational efficiency, and identify opportunities for growth and optimization. Therefore, the key responsibilities of a data analyst typically include:

- Gather data: Analysts often collect data themselves. This could include conducting surveys, tracking visitor characteristics on a company website, or buying datasets from data collection specialists.

- Clean data: Raw data might contain duplicates, errors, or outliers. Cleaning the data means maintaining the quality of data in a spreadsheet or through a programming language so that your interpretations won't be wrong or skewed.

- Model data: This entails creating and designing the structures of a database. Data analysts might choose what types of data to store and collect, establish how data categories are related to each other, and work through how the data actually appears.

- Interpret data: Interpreting data will involve finding patterns or trends in data that could answer the question at hand.

- Present: Communicating the results of your findings will be a key part of this job. Data analysts can do this by putting together visualizations like charts and graphs, writing reports, and presenting information to interested parties (Sammydsouza, 2022).

1.4 The skill requirements for a data analyst

According to the study of Jin Zhang et al.(2023), which collected 2500+ data-analyst job ads posted on LinkedIn and analyzed them using distribution analysis, crosstabulation analysis, and cluster analysis. Among many findings, this study identified five most essential nontechnical skills and five most essential areas of technical skills.

Of all the nontechnical skills, this study has identified analytical skills, communication skills (especially written communication skills), leadership skills, collaboration skills, and management skills as the most essential nontechnical skills for data analysts. A total of 84.8% of all the data analyst job postings required analytical skills, 74.1% required communications skills (66.1% specifically required written communication skills), 61.5% required leadership skills, 53.7% required collaboration skills, and 51.7% required management skills.

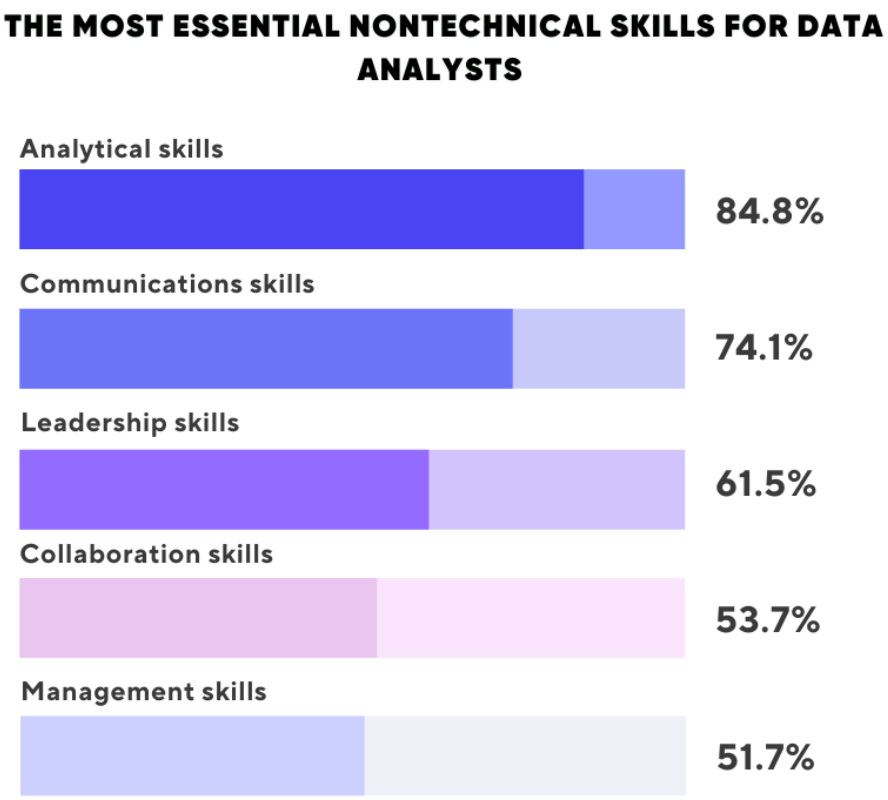

Figure 1.1. The most essential nontechnical skills for data analysts

Source: Jin Zhang et al.(2023)

In terms of technical skills, of 90+ computer programs business organizations expect data analysts to use, this study identified SQL, Microsoft Excel, Tableau, Python, and Microsoft Power BI to be the five most essential computer programs for potential data analysts to master.

In conclusion, data analyst is a profession that is a harmonious blend of technical skills and non-technical skills, exemplifying the synergy between the two.

1.5 Career paths and growth opportunities for data analysts

As organizations across Vietnam place greater emphasis on data-driven decision-making, data analysts can look forward to diverse career advancement opportunities. While the core responsibilities of a data analyst may remain consistent, experienced professionals can explore specialized roles or take on more senior-level responsibilities.

One common trajectory is to become a senior data analyst, taking on complex analytical projects, mentoring junior team members, and driving high-impact initiatives. This involves expanding one's technical skills in areas like advanced statistical modeling, machine learning, and data visualization.

Many data analysts also leverage their analytical expertise to move into adjacent roles, such as business intelligence analyst or data engineer. These positions require a deeper understanding of data architecture, database management, and translating business requirements into technical solutions.

For data analysts interested in people management, there are openings for data team leads or analytics managers. In these roles, individuals oversee a group of analysts, coordinate cross-functional data projects, and align analytical efforts with broader organizational goals.

At the executive level, chief data officers or vice presidents of data and analytics are responsible for setting the strategic data vision and driving enterprise-wide data transformation. Data analysts with strong business acumen and leadership abilities can aspire to these top-level positions.

Outside of the traditional corporate path, data analysts can also explore opportunities in consulting, freelancing, or starting their own data-focused businesses. They may leverage their analytical expertise to provide advisory services, develop data products, or drive innovation in data-intensive industries.

Regardless of the specific career direction, continuous learning and upskilling are crucial for data analysts to stay relevant and capitalize on emerging technologies, methodologies, and industry trends. Pursuing certifications, participating in professional communities, and engaging in ongoing training can all help data analysts expand their competencies and unlock new possibilities.

CHAPTER 2 : INTRODUCTION OF THE PROJECT

2.1 Reason for choosing the topic

The retail industry today is facing immense pressure to provide a personalized, engaging customer experience that drives loyalty and lifetime value. Consumers have more choices than ever before, with the explosive growth of e-commerce and the evolution of shopping behaviors accelerated by the pandemic. In this highly competitive landscape, retailers must find ways to truly understand and cater to the needs of their target customers. According to a study by McKinsey, retailers that excel at customer centricity enjoy 60% higher profitability compared to their peers. However, achieving this level of customer intimacy requires leveraging advanced data analytics techniques that can unlock critical insights about shopper preferences and behaviors.

There are many aspects that businesses can exploit to enhance customer experience, but one of the core of a truly customer-centric retail strategy lies in the ability to segment the customer base and tailor product recommendations accordingly. By applying advanced analytical techniques like RFM (Recency, Frequency, Monetary) modeling, retailers can identify their most valuable, engaged shoppers and develop targeted initiatives to drive loyalty and lifetime value. In fact, a Bain & Company study revealed that retailers excelling at customer centricity can boost customer lifetime value by 82%. But the real power comes when retailers pair customer segmentation with granular insights from basket analysis. By understanding the specific product affinities and co-purchasing behaviors within each segment, retailers can optimize their assortments and deliver hyper-personalized recommendations that resonate. By identifying complementary products that customers tend to purchase together, businesses can launch marketing strategies such as up- sell and cross-sell, or enhancing product recommendations method. As the retail landscape becomes increasingly crowded, the retailers that can harness the power of data-driven customer intelligence to power these types of targeted, segment-specific strategies will be poised to drive sustainable growth and profitability.

In recent years, market basket analysis and customer segmentation have gained significant attention in academic and industry research. Studies such as M Qisman et al.

(2021) have leveraged market basket analysis to uncover consumer trends, while extensive research on customer segmentation has been conducted, including by Jun Wu et al. (2020) and P. Anitha and Malini M. P. However, there is a lack of studies that combine these two techniques to drive personalized product recommendations.

Recognizing the potential of integrating customer segmentation and market basket analysis, a new research study was undertaken, titled "Leveraging Customer Segmentation and Market Basket Analysis to drive Personalized Product Recommendations in Retail Industry".

2.2 Objectives of the project

2.2.1 General objectives

This project leverages the powerful combination of the RFM (Recency, Frequency, Monetary Value) model and Market Basket Analysis to deliver unparalleled insights into the purchasing behavior of distinct customer segments. By meticulously examining the recency, frequency, and monetary value of each customer's transactions, the RFM model enables the identification of high-value, loyal, and potentially churning customers.

Complementing this customer-centric approach, Market Basket Analysis delves into the intricate relationships between products, uncovering the items that are frequently purchased together. This invaluable information empowers retailers to make informed recommendations, tailoring their product offerings to the specific preferences and buying patterns of each customer segment.

By integrating RFM and Market basket analysis allows retailers to provide personalized product recommendations, driving cross-selling and upselling opportunities that enhance the customer experience and boost sales.

2.2.2 Specific objectives

- Segment customers into clusters that have similar purchasing patterns. As a result, retailers can carry out distinct strategies for each segment.

- Apply the Apriori algorithm to the customer transaction data to identify frequent itemsets - groups of products that are commonly purchased together for each segment.
- Use Apriori-generate association rules for product recommendations that can propose complementary items to customers based on their customer segment's previous purchases.
- Offer an efficient and simple way to make targeted, data-driven product recommendations to customers, enhancing cross-selling opportunities and boosting sales performance.

2.3 Object and scope of the research

2.3.1 Research object

The research objects are transactions, product details, and customer information documented by an online retail company based in the United Kingdom.

2.3.2 Research scope

The scope of the data spans vastly, from granular details about each product sold to extensive customer data sets from different countries. However, this project will focus on the transaction made from the United Kingdom in 2011.

2.4 Research methodology

2.4.1 Qualitative research methods

The qualitative research method is carried out through the process of exchange, discussion, and searching for relevant literature and scientific research sources both domestically and internationally. Through the process of selection and information discovery, an objective perspective and deeper understanding of the issues being researched can be developed, as well as the synthesis of points that previous studies have and have not addressed, in order to identify a research direction and proceed with the next steps.

2.4.2 Quantitative research methods

After preprocessing the customer purchase data, the first step was to conduct an exploratory data analysis (EDA) to gain a deeper understanding of the overall purchasing patterns and behaviors. This provided valuable contextual insights that informed the subsequent RFM (Recency, Frequency, Monetary) analysis. Building on these segmentation insights, then applying association rule mining techniques to uncover frequent itemsets and product affinities within each customer group. This helps generate highly personalized product recommendations tailored to the shopping habits and interests of each RFM segment.

2.5 Research contributions

2.5.1 Scientific significance

This research makes a substantial contribution to using RFM analysis and Market Basket Analysis to evaluate customer data and improve customer centricity in the retail business. As a result, this study will pave the way for future research in this area, as well as the potential of RFM and MBA to improve customer experience and accelerate sales performance, among other things.

2.5.2 Practical significance

Through analyzing data, this project helps businesses understand their customers' behavior and the value of customer segmentation, allowing them to develop appropriate strategies to keep and retain their attachment to the businesses.

One of the other notable implications of this project is the deployment of Market Basket Analysis in conjunction with the RFM model. With this access technique, building sales plans will be easier than ever because it allows retailers to create product recommendations and cross-sells tailored to each group of customers.

2.6 Structure of the project

This project consist 6 chapters:

Chapter 1: Data analyst: an introduction to the profession

Chapter 1 provides an overview of the data analyst profession, exploring the growing importance of data analytics in Vietnam's evolving business landscape. The chapter begins by defining who data analysts are and outlining their core responsibilities, which include collecting, processing, analyzing, and interpreting data to support organizational decision-making.

Chapter 2: Introduction of the project

Chapter 2 provides an introduction to the project, outlining the rationale for the chosen topic. It presents the general and specific objectives that guide the research. The chapter then defines the research object and scope, clarifying the boundaries and focus of the project. To achieve the stated objectives, the chapter describes the application of both qualitative and quantitative research methods. Furthermore, it highlights the scientific and practical significance of the research contributions that the project aims to deliver. Finally, the chapter concludes by outlining the overall structure of the project.

Chapter 3: Theoretical background and related work

Chapter 3 provides the theoretical background and overview of related research. It introduces the concept of market basket analysis and the association rule mining technique, including details on the Apriori algorithm.The chapter also discusses the RFM (Recency, Frequency, Monetary) model, a customer segmentation approach commonly used in data analysis.

Chapter 4: Methodology and proposed research model

Chapter 4 outlines the methodology and proposed research model for the project. It begins by describing the overall research process, including the steps and procedures to be followed. The chapter then provides a detailed description of the data used in the analysis, including the data context and the attributes or variables available. This lays the groundwork for the subsequent data analysis and model development detailed in the following chapters. The comprehensive methodological approach and data details presented in this chapter set the stage for the core analysis and findings of the research.

Chapter 5: Experimental result and recommendations

It starts by detailing the data preprocessing steps taken, including cleaning, transformation, and preparation of the data for analysis. The chapter then describes the Exploratory Data Analysis (EDA) conducted, uncovering insights and patterns in the data. Next, it explains the customer segmentation analysis using the RFM (Recency, Frequency, Monetary) model, identifying distinct customer groups. Finally, the chapter covers the product recommendation system developed, leveraging association rule mining techniques to suggest relevant products to customers.

Chapter 6: Conclusion and development direction

Chapter 6 concludes the research project. It summarizes the key findings and contributions. The chapter also identifies limitations of the study and suggests future development directions to build upon the current work.

CHAPTER 3 : THEORETICAL BACKGROUND AND RELATED WORK

3.1 Market basket analysis

Market basket analysis is a popular concept which is referenced in a lot of research relating to economics and business.

R.Agrawal et al.(1993), apparently first used Market Basket Analysis who had a large collection of consumer transaction data previously collected and the association rules between items purchased were discovered. According to R.Agrawal et al. market basket analysis is a set of analytical techniques aimed to discover the associations and correlations among products by analyzing the customer's shopping baskets. The method was rapidly implemented as a standard method for a number of practical applications in the field of marketing(Chen et al., 2005). (R. Agrawal, 1993)

In the research of Herman Aguinis et al (2013), MBA, also known as association rule mining or affinity analysis, is a data-mining technique that originated in the field of marketing to identify relationships between groups of products, items, or categories.

In another research, Valle MA et al. (2018) stated that market basket analysis is a process that analyzes the habits of buyers to find relationships between various items in their shopping basket.

Therefore, Market Basket Analysis can be understood as a data mining technique used to identify associations and relationships between the items that customers purchase together in a single transaction or "market basket". The goal is to uncover patterns in customer buying behavior and understand which products are commonly bought in conjunction with one another. Through this concept, if it is known that customers who purchase one product are likely to purchase another product, it is possible for retailers to market these products together, or to make the purchasers of a target prospects for the second product. For example, If customers who purchase strawberries are likely to purchase whipping cream, they will be more likely to if whipping cream is displayed just beside a strawberries aisle.

3.2 Association Rule Mining and Apriori algorithm

3.2.1 Association Rule Mining

According to Akbar Telikani et al.(2020), Association Rule Mining (ARM) is a significant task for discovering frequent patterns in data mining. ARM aims to find close relationships between items in large datasets, which was first introduced by Agrawal et al (2016). In the context of market basket analysis, association rule mining is used to identify patterns and relationships in customer purchase data, with the goal of understanding which products are commonly purchased together. There are several key association mining algorithms that are commonly used in data analysis such as FP-Growth, Eclat, but this project will use Apriori algorithm.

3.2.2 Apriori algorithm

One of the well-known and commonly used association rule discovery data mining methods is the Apriori algorithm. The Apriori Algorithm is a basic algorithm proposed by Agrawal & Srikant in 1994 for the determination of the frequent itemset for Boolean association rules.

According to Shabtay et al. (2021), "an item-set is a set of 0 or more items, and a frequent item-set is an item-set whose support is greater than the custom minimum support count."

In the context of market basket analysis, an "item-set" is a set of products that customers buy together, such as {bread, butter} or {milk, eggs, flour}. These item-sets reflect common purchasing patterns.

However, not all item-sets are equally informative. "Frequent item-sets" are sets of products that are frequently bought together, and they hold more value in uncovering associations and patterns in the data. A "frequent item-set" is defined as an item-set whose frequency (support) in the transaction data exceeds a minimum threshold set by the user.

Identifying these frequent item-sets is a crucial step in the Apriori algorithm, as it forms the foundation for discovering valuable insights from the market basket data. There are some common evaluation criteria for frequent item-sets: support, confidence, lift.

Support

The support of an item or set of items is defined as the proportion of transactions in our data collection that contain the number of that specific item , relative to the total number of transactions. Support indicates how frequently an itemset appears in the total transactions.

Equation 3.1. Support equation

$$Support\ A = \frac{Number\ of\ transaction\ that\ contains\ A}{Total\ transaction}$$

Confidence

Confidence measures the possibility that a client who buys product A will also buy product B. A rule of association is a comment of the form (item set A) ⇒ (item set B), with A as the precedent and B as the consequence. Confidence indicates the likelihood of Consequence occurring on the cart given pre-existing antecedents.

Equation 3.2.Confidence equation

$$Confidence(A \Rightarrow B) = P(A|B) = \frac{Number\ of\ transaction\ that\ contains\ A\ and\ B}{Total\ transaction\ that\ contains\ A}$$

Lift

Lift is a measure of the strength of an association rule, specifically the degree to which the occurrence of the antecedent increases the chances of the occurrence of the consequent.

Equation 3.3.Lift equation

$$\frac{Lift(A \Rightarrow B) = \ Confidence(A \Rightarrow B)}{Support(B)}$$

Lift uses 1 as the target value to show the relationship between *A* and *B*. If the value is greater than 1, then $A \Rightarrow B$ is a valid strong association rule. Conversely, $A \Rightarrow B$ is an invalid strong association rule. When the value is equal to 1, however, there is a special case, that is, the *A* and *B* at independence, at the time Lift(A⇒B)=1 (Zhou et al., 2010).

3.3 RFM Model

Customer segmentation is the process of dividing customers into groups based on past data with the demands, characteristics, and the same functioning (Manero et al., 2018). The RFM model is a widely used customer segmentation and analysis technique in the field of marketing and customer relationship management. It is a data-driven approach that evaluates customers based on three key behavioral metrics: Recency, Frequency, and Monetary value.

- Recency (R): It refers to the number of days before the reference date when a customer made the last purchase. Lesser the value of recency, higher is the customer visit to a store.

- Frequency (F):The Frequency variable captures the number of transactions or interactions a customer has had with the business over a defined time period. Frequency can be calculated by counting the total number of purchases, visits, or other trackable interactions made by the customer within a given timeframe. Higher the value of Frequency, more is the customer visit to the company.

- Monetary (M):The Monetary (M) variable in the RFM model measures the monetary value of the transactions a customer has made with the business. This variable can be estimated by looking at the total revenue or amount of money the customer has spent over a certain time period. Higher the value, more is the profit generated to the company.

According to Sinaga et al.(2020), the k-means is one of the most popular unsupervised learning algorithms that solve the well-known clustering problem. It was first proposed by MacQueen in 1967.

Kmean is an unsupervised, non-deterministic, numerical, iterative method of clustering (Jyoti Yadav et al., 2013). Kmean algorithm selects k centroid randomly, where the value k is fixed in advance. Each data point is assigned to the cluster with the closest centroid, based on the distance metric. Then the centroids are updated by taking the mean of the points of each cluster (Aslekar et al., 2019). Repeat this process until the centroid tends to be unchangeable or we can say that until the criterion function converges.

Advantages of Kmean are its simplicity and speed (Arthur & Vassilvitskii, 2007). However, because the number of clusters in the Kmean algorithm is determined at random, the results may differ between runs. To determine the precise number of clusters, employed the Elbow approach to determine the best number of clusters, followed by the Silhouette method to re-evaluate the results. Meanwhile, Mengyao Cui (2020) exclusively uses the elbow approach to determine the optimal k value. Using only silhouette can also find the suitable k value, this method applied in the research of Shutaywi et al. (2021).

The Elbow Method is a technique used to identify the optimal number of clusters (k) in the K-means algorithm. The basic idea is to run K-means clustering on the dataset for different values of k, and then look at the total within-cluster sum of squares (WCSS) for each value of k. The better the clustering, the lower the overall WCSS(Mengyao Cui, 2020).

Equation 3.4.Within-cluster sum of squares equation

$$WCSS = \sum_{P_i \, in \, cluster \, 1} distance(P_i, C_1)^2 + \sum_{P_i \, in \, cluster \, 2} distance(P_i, C_2)^2 + \cdots + \sum_{P_i \, in \, cluster \, n} distance(P_i, C_n)^2$$

The Silhouette Score is a metric that quantifies how well each data point fits into its assigned cluster. The silhouette coefficient is calculated by taking into account the mean intra-cluster distance a and the mean nearest-cluster distance b for each data point.

Equation 3.5.Silhouette equation

$$Silhouette \; score_i = \frac{b_i - a_i}{max(a_i, \, b_i)}$$

The silhouette score ranges from -1 to 1. According to K. R. Shahapure and C. Nicholas (2020), a silhouette score with a value near + 1 means the data point is in the correct cluster, a silhouette score with a value near 0 means the data point might belong in some other cluster, a silhouette score with a value near -1 means, the data point is in (a) wrong cluster.

3.4 Overview of related research

The research of M Qisman et al. (2021) focuses on using Apriori algorithm to find consumer patterns in buying goods through transaction data. The outcome of this research is to find the combination of products that are usually bought together in a computer store.

The research of P. Anitha and Malini M. P (2022) focuses on finding customer purchase behavior through variables Rencency, Frequency and Monetary for each cluster using the Kmean algorithm.

The research of Shruthi Gurudath (2020) leverages transaction data analysis and the Apriori algorithm to uncover frequent purchase patterns and product associations, supporting inventory management and building product recommendation capabilities.

Overall, current research has consistently highlighted the importance of the RFM (Recency, Frequency, Monetary) Model and Market Basket Analysis in analyzing customer purchasing behavior. While there has been a wealth of research related to the RFM (Recency, Frequency, Monetary) Model and Market Basket Analysis, which have focused on analyzing customer buying behavior, there has been a lack of studies that leverage the advantages of these two techniques to enhance personalized product recommendations and drive a more customer-centric approach.

CHAPTER 4 : METHODOLOGY AND PROPOSED RESEARCH MODEL

4.1 The research process

Figure 1 describes methodology and proposed research model with four stages:

(1) The first stage is data preparation. Firstly, this study carried out preprocessing data to handle missing values and invalid values, etc. After that, the cleaned data will be used for Exploratory Data Analysis - EDA to find some core insights relating to the customer pattern.

(2) Stage 2 is customer segmentation analysis according to RFM. From the input data extracted from stage 1, calculate recency, frequency and monetary values for use in the RFM model. After the project will implement the standardization for input data and determining the optimal number of clusters (k) using techniques like the elbow method or silhouette analysis, the Kmean algorithm is used for customer segmentation.

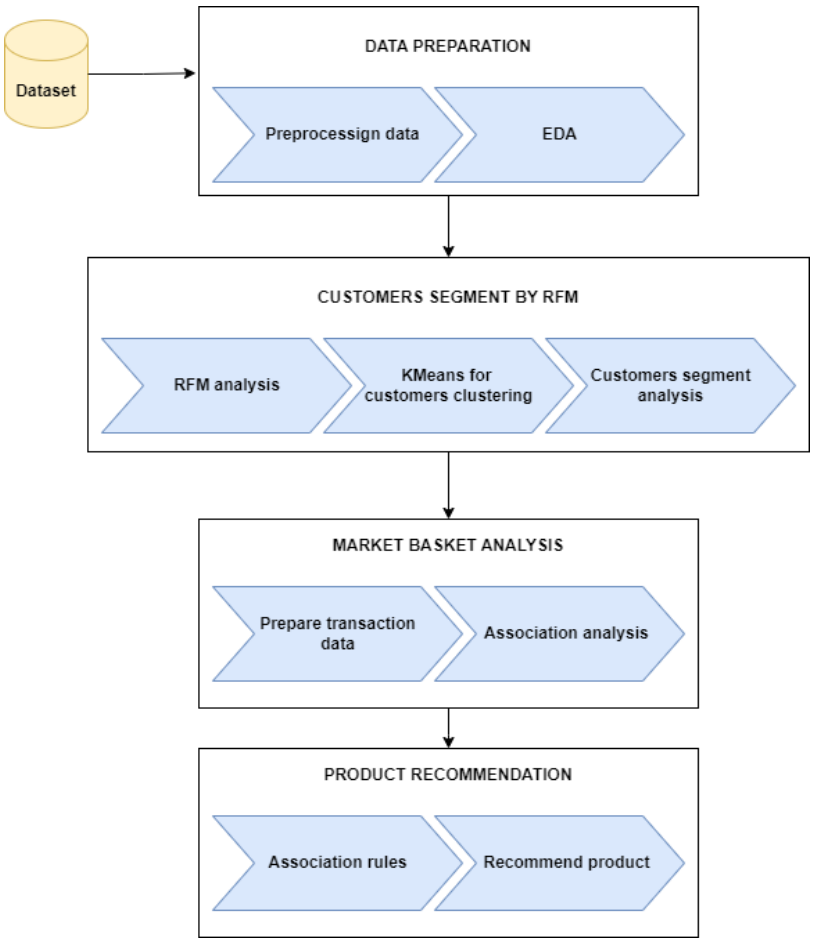

Figure 4.1.The research process diagram

(3) Stage 3 is Market Basket Analysis. The main purpose of this stage is using the Apriori Algorithm for the association rule mining/analysis. To reach this aim, the transactional data should be converted into a format suitable for analysis.

(4) Stage 4 is product recommendation. The product recommendation part of this project is going to make use of the Association Rules that were uncovered in the MBA section. Product recommendation is basically one of the advantages of Market Basket Analysis where customers can be recommended products that are in the same itemsets as their current products.

4.2 Data description

4.2.1 Data context

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

4.2.2 Attribute description

Table 4.1.Describe data attributes

Variables	Role	Type	Description
Invoice No	ID	object	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	ID	object	a 5-digit integral number uniquely assigned to each distinct product
Description	Feature	object	product name
Quantity	Feature	int64	the quantities of each product (item) per transaction
InvoiceDate	Feature	datetime64[ns]	the day and time when each transaction was generated
UnitPrice	Feature	float64	product price per unit

CustomerID	Feature	float64	a 5-digit integral number uniquely assigned to each customer
Country	Feature	object	the name of the country where each customer resides

CHAPTER 5 : EXPERIMENTAL RESULT AND RECOMMENDATIONS

5.1 Preprocessing data

This study presents a comprehensive data preprocessing and cleaning procedure applied to the Online Retail dataset, with the goal of preparing the data for in-depth analysis of customer purchasing patterns and behaviors.

Table 5.1.Overview of dataset before preprocessing data

Column	Non-Null	Count	Dtype
InvoiceNo	541909	non-null	object
StockCode	541909	non-null	object
Description	540455	non-null	object
Quantity	541909	non-null	int64
InvoiceDate	541909	non-null	datetime64[ns]
UnitPrice	541909	non-null	float64
CustomerID	406829	non-null	float64
Country	541909	non-null	object

The first step was to address the CustomerID column. The data type was converted from numeric to string, and any missing CustomerID values were filled with the string "GuestCustomer". This ensures consistency and allows for proper handling of customer identities.

Next, the Description column underwent several transformations. Leading and trailing white spaces were removed, and rows with missing descriptions were dropped. Additionally, rows where the product description had 8 or fewer characters were eliminated, as these short descriptions likely do not represent actual products. A specific row with the description '20713' was also removed, as was any row where the description was "DOTCOM POSTAGE", as this refers to shipping fees rather than a product.

For the StockCode column, leading and trailing white spaces were trimmed. Similarly, white spaces were removed from the InvoiceNo column, and the data type was changed from

numeric to string. Furthermore, rows where the InvoiceNo started with the letter 'C' were removed, as these represent canceled orders.

Regarding the Quantity column, any rows with a quantity less than or equal to 0 were dropped to ensure the analysis focuses on valid, completed transactions.

In addition to these column-specific transformations, the data underwent several other preprocessing steps. Duplicate rows with the same StockCode but different Descriptions were removed and consolidated. A new 'SaleRevenue' column was created by multiplying the Quantity and UnitPrice columns, providing a valuable metric for revenue analysis. Moreover, new columns related to the date and time of the transactions, such as year, month, day, hour, minute, and second, were generated to enable temporal insights.

Finally, to concentrate the analysis on 2011 and United Kingdoms' transactions, any rows where the transaction occurred in 2010 and other countries were removed from the dataset.

Table 5.2.Overview of dataset after preprocessing data

Column	Non-Null	Count	Dtype
InvoiceNo	445183	non-null	object
StockCode	445183	non-null	object
Description	445183	non-null	object
Quantity	445183	non-null	int64
InvoiceDate	445183	non-null	datetime64[ns]
UnitPrice	445183	non-null	float64
CustomerID	445183	non-null	float64
Country	445183	non-null	object
SaleRevenue	445183	non-null	float64
Date	445183	non-null	object
Month	445183	non-null	object

Year	445183	non-null	int32
Week of the Year	445183	non-null	UInt32
Day of Week	445183	non-null	object
Time of Date	445183	non-null	category

5.2 Exploratory Data Analysis - EDA

The initial phase of EDA focuses on sales analysis. This sales analysis dives deeper into the data, examining sales performance at the monthly, weekly, and daily levels. Figure 5.1 displays the monthly trends for Quantity of products ordered (left) and Sales Revenue (right). Both measures peaked in November, followed by October and September as the next highest months.

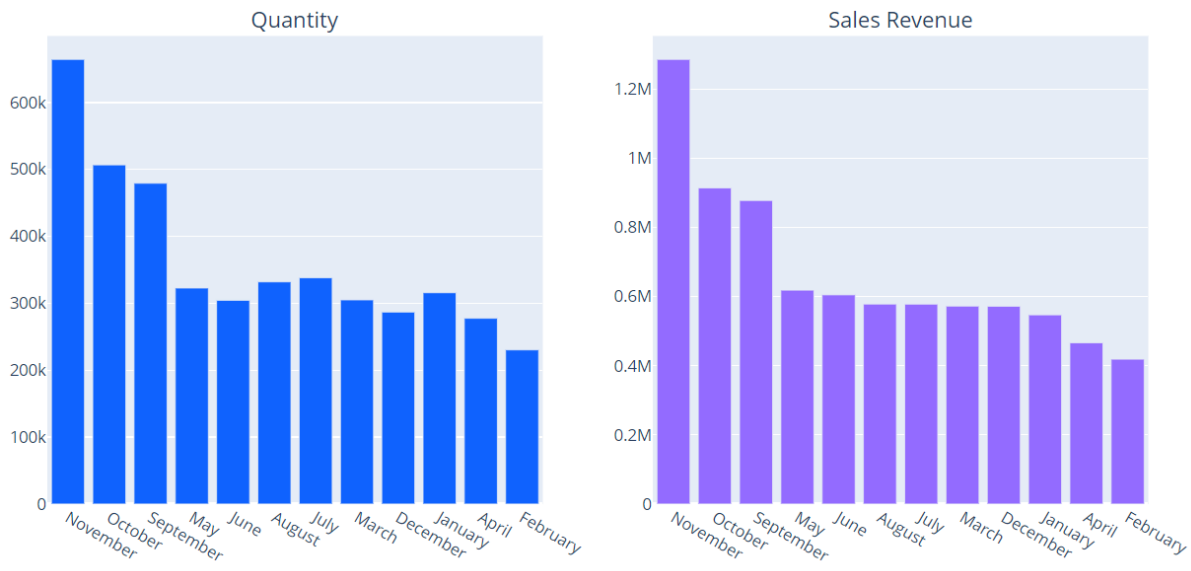

Figure 5.1.Monthly Sales Revenue and Quantity

Figure 5.1 illustrates the weekly trends in sales revenue and the quantity of products ordered. The highest peak across both measures occurred during the 49th week, which falls within the November holiday season. This suggests a surge in demand for decoration items during this period, driving both the quantity of products ordered and the corresponding sales revenue to their highest levels. The strong correlation between the quantity and sales revenue trends indicates that as the quantity of products ordered increased, the sales revenue also rose accordingly.

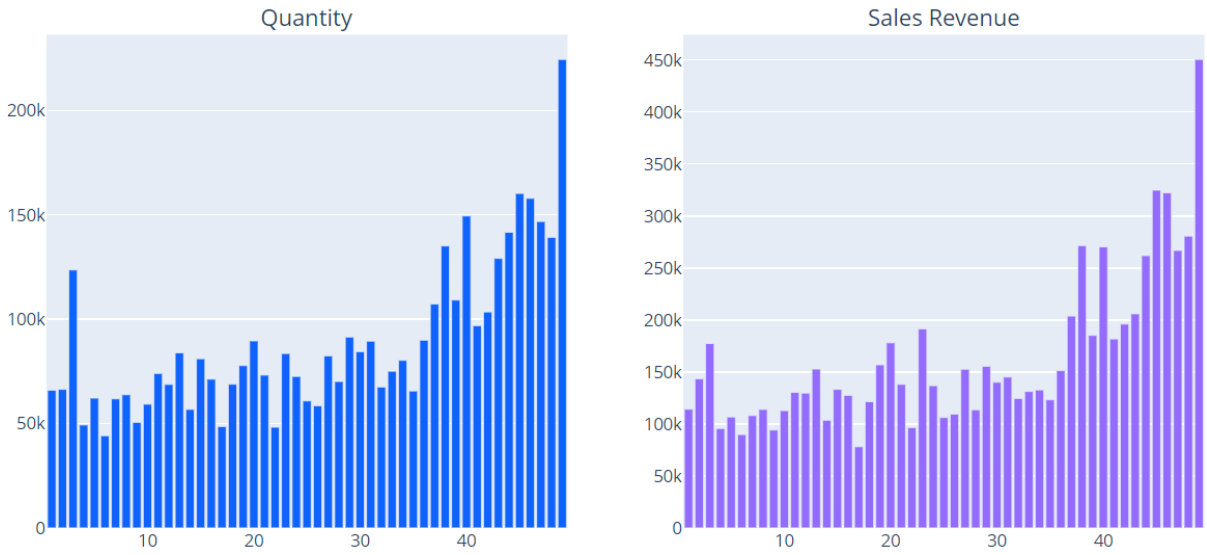

Figure 5.2.Weekly Sales Revenue and Quantity

From figure 5.2 Thursday and Tuesday are observed to generate the highest quantity of products and sales revenue.

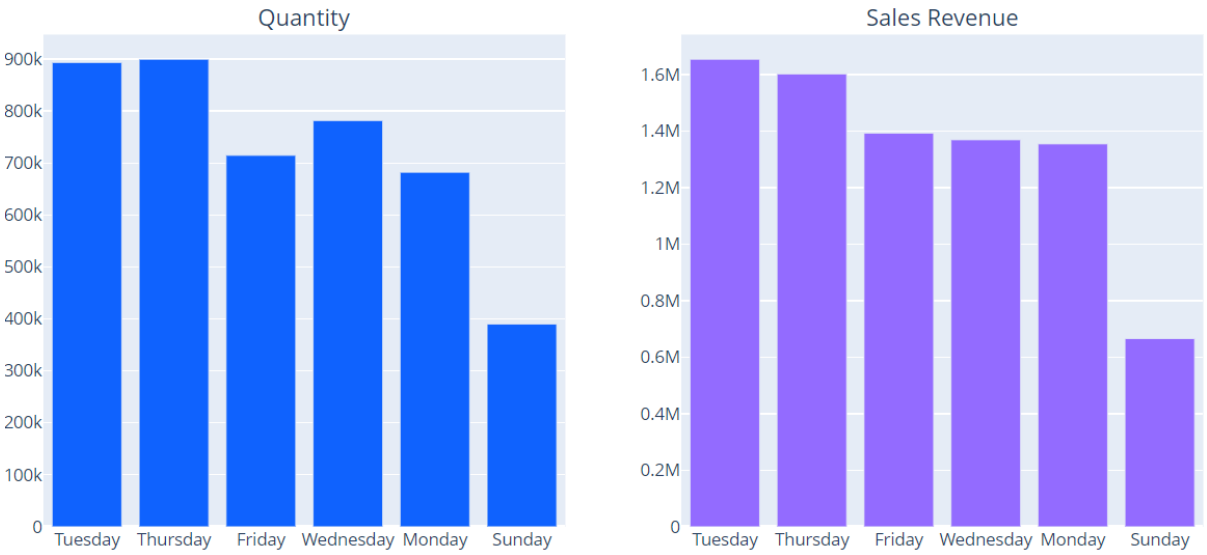

Figure 5.3.Day of the Week Sales Revenue and Quantity

Figure 5.3 presents a breakdown of Quantity of orders (left) and Sales revenue (right) by time of day. Over 99% of the orders were placed during the morning and afternoon hours, with minimal activity during evening and the late night periods. This trend is consistent across both the order quantity and sales revenue metrics, indicating a strong preference among customers to place orders during typical business hours.

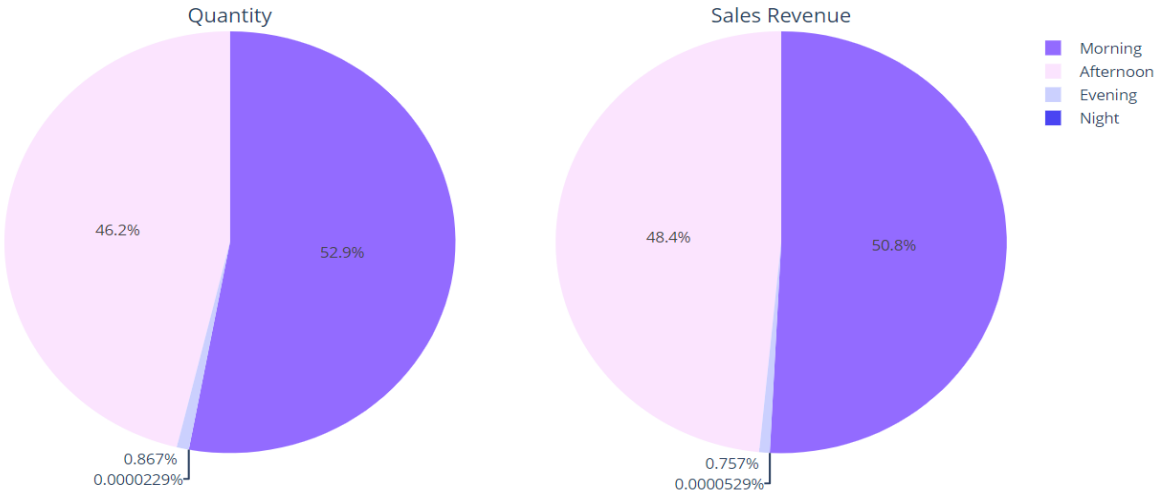

Figure 5.4.Percentage for Time of Day Sales Revenue and Quantity

The next section will focus on product analysis. This comprehensive product understanding from EDA will enable the deeper insights in the Market Basket Analysis and Product Recommendations sections.

To look at the products, which ones have high Quantity sold, or which product has high Sales Revenue the WordCloud is generated. The WordCloud in Figure 5.5 highlights the product descriptions for the items with the highest total quantity sold. This provides visibility into the top-moving products in terms of sales volume. The larger the word, the higher the total quantity sold for that product. As a result, Medium Ceramic Top Storage Jar, Paper Craft Little Birdie, Word War 2 Gilders Asstd Designs, Small Popcorn Holder, and others are among the best-selling items.

Figure 5.5.Product per quantity

Complementing the quantity analysis, Figure 5.6 presents a WordCloud focused on the product descriptions with the highest total sales revenue. This helps identify the most valuable products in the portfolio, even if their sales quantities may not be the highest.

Among the products with the biggest sales revenue are the Paper Craft Little Birdie, Cream Hanging Heart T-light Holder, Regency CakeStand 3 Tier, Medium Ceramic Top Storage Jar, and so on.

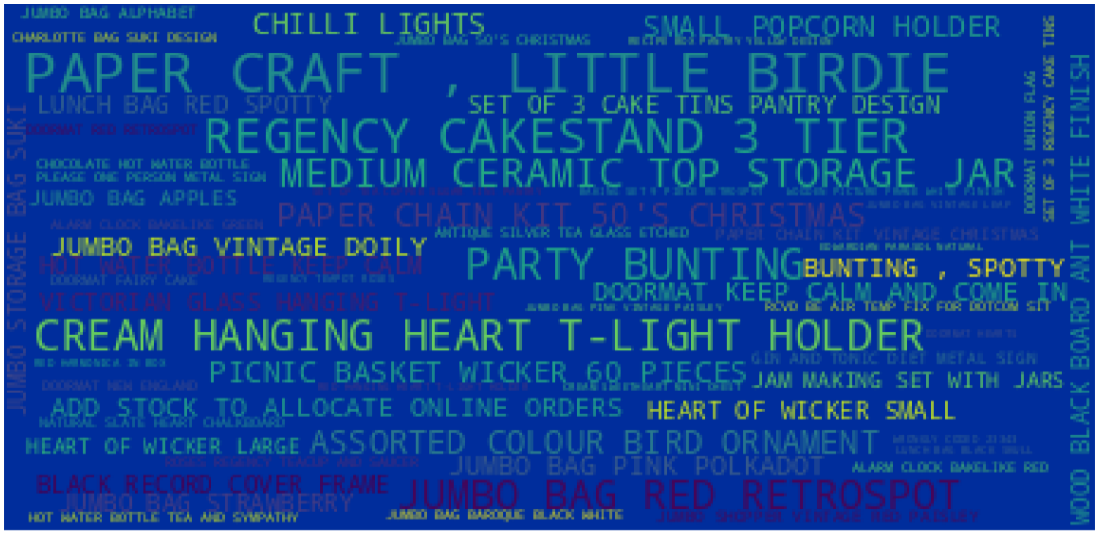

Figure 5.6.Product per revenue

Figure 5.7 and Figure 5.8 present a comparative analysis of the top 10 and bottom 10 products based on two key metrics: volume quantity and sales revenue.

The left panel of the figure displays a bar chart depicting the top 10 products with the highest volume quantities. This provides insights into the best-selling products in terms of the total units sold. The right panel, on the other hand, showcases the bottom 10 products with the lowest volume quantities, highlighting the least popular items.

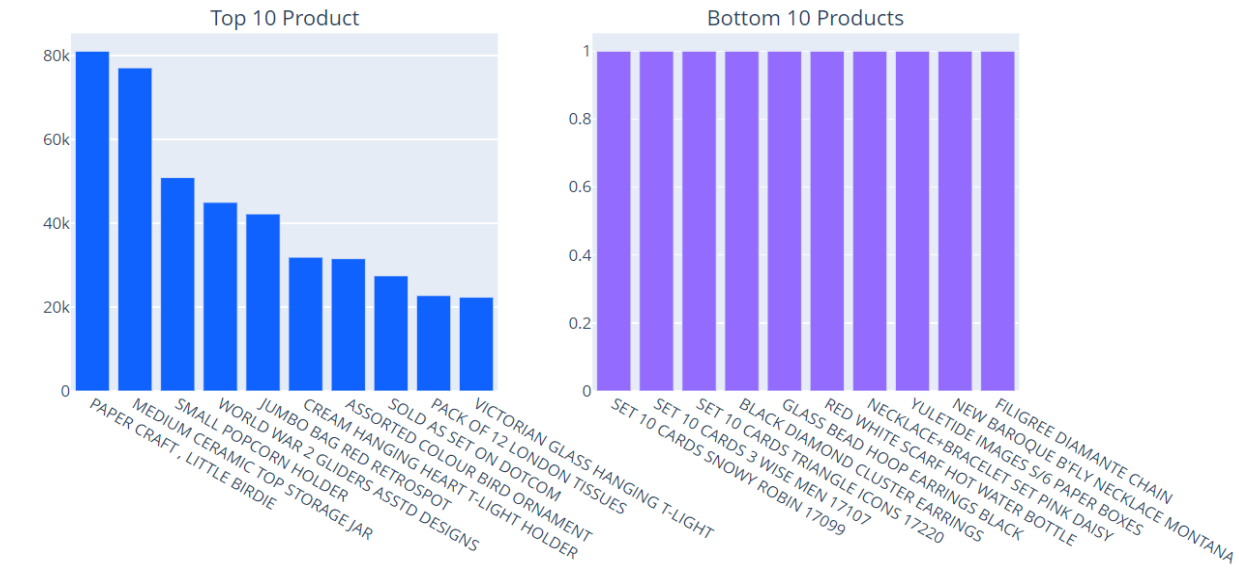

Figure 5.7.Product by Volume Quantity

Complementing the volume quantity analysis, figure 5.7 also includes a similar comparative analysis for sales revenue. The left panel displays the top 10 products with the highest sales revenue. The right panel focuses on the bottom 10 products with the lowest sales revenue, shedding light on the underperforming products.

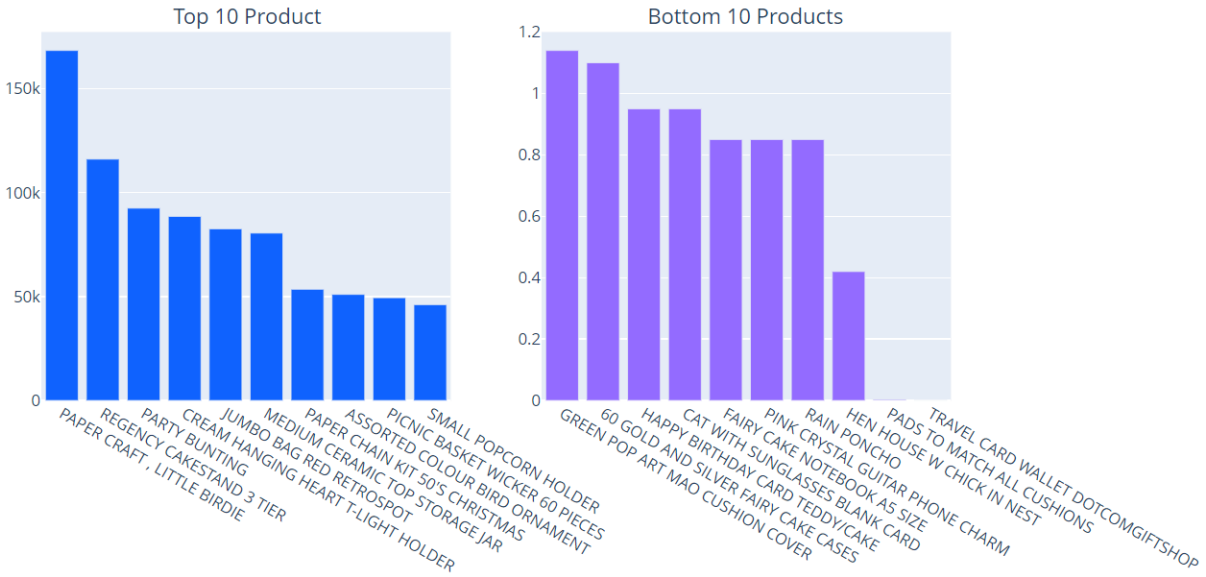

Figure 5.8.Product by Sales Revenue

5.3 Customer segmentation using RFM

Customer segmentation using the RFM (Recency, Frequency, Monetary) model provides valuable insights into customer behavior. The RFM analysis revealed that on

average, customers have made their last purchase 84 days ago, with an average of 4 orders and a total revenue of 1,775.87.

Further analysis of the RFM metrics and their quartiles provides a more granular understanding of the customer base. 25% of customers have made a purchase within the last 17 days, while 75% have purchased within the last 128 days. In terms of frequency, 25% of customers have made only 1 purchase, compared to 75% who have 4 or fewer purchases. For monetary value, 25% of customers have spent 291.76 or less, whereas 75% have a total spend of 1,534.97 or less.

Table 5.3.Quartiles description in RFM table

	Recency	**Frequency**	**Monetary**
mean	84.488	4.018	1775.870
min	1.000	1.000	0
max	340.000	171.000	231822.690
25%	17.000	1.000	291.763
50%	47.000	2.000	635.670
75%	128.000	4.000	1534.968

This study assigns numeric labels to each RFM dimension, dividing the values into quartiles and mapping them to a scale of 1-4. This allows the customers to be categorized based on their relative Recency, Frequency, and Monetary values. Finally, an overall RFM_Score is calculated by summing the individual R, F, and M labels.

This analysis uses both the Elbow method and Silhouette score to determine the optimal number of clusters in the dataset. The Elbow method suggests 4 clusters is a compelling choice, as an "elbow" forms in the plot around this point. However, the Silhouette analysis indicates 6 clusters may be even better, as this configuration achieves the highest score of 0.4632. Ultimately, after considering the insights from both techniques, the decision is made to proceed with 4 clusters. Four clusters have a silhouette score of 0.461, not too different from 6 clusters.

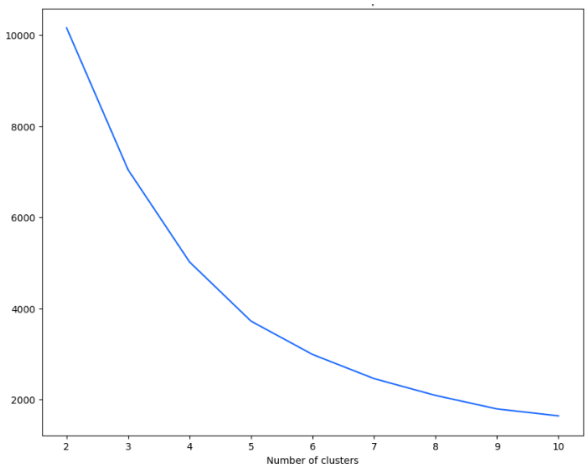

Figure 5.9.Elbow method result

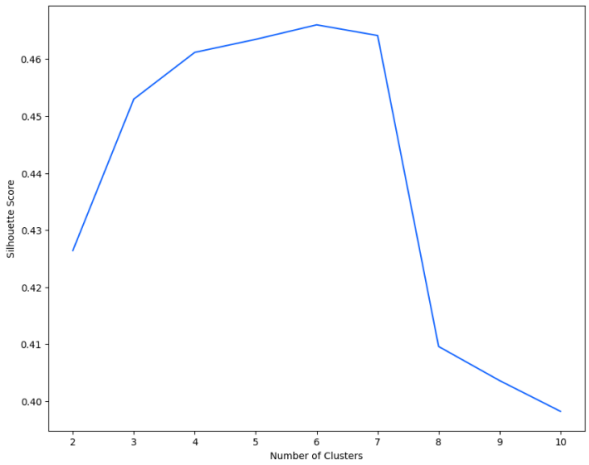

Figure 5.10.Silhouette score result

The analysis culminated in the selection of a 4-cluster solution, which the team has thoughtfully named Platinum, Gold, Silver, and Bronze. The Platinum cluster, though comprising the fewest customers, represents the segment with the highest contribution to the company's revenue. This elite group exhibits the highest average order value and most recent purchase behavior, underscoring their immense value to the business. The remaining clusters - Gold, Silver, and Bronze - demonstrate successively lower levels of revenue contribution and customer engagement. The Gold cluster still maintains a strong connection to the company, while the Silver and Bronze segments show the most gradual erosion in both monetary and recency . This tiered structure provides a clear and actionable framework for tailoring the company's targeting, retention, and growth strategies to maximize value across the diverse customer base. The table displays an in-depth examination of four clusters based on their average RFM_Score, recency, frequency, monetary value, and number of customers.

Table 5.4.Each cluster description

	RFM_Score mean	**Recency mean**	**Frequency mean**	**Monetary mean**	**Customer count**
Platinum	11.640000	18.280000	51.680000	63739.415600	25
Gold	11.031308	20.457643	8.295580	3333.405285	1086
Silver	6.911863	53.054324	2.075942	645.792263	1804
Bronze	4.277716	227.713326	1.403135	429.943036	893

The below pie charts show the percent representation by each cluster concerning the RFM metrics. The customers in the Platinum cluster have a high average monetary value and

high average frequency. The recency of customers in the Platinum cluster is very low among the 4 clusters with 5.72%. The Gold cluster contains customers with a very low monetary value of 4.89%, recency of 6.41%, and a frequency of 13.1%, this by far is the customers with poor metrics; however, they are better than 'Silver' and 'Bronze' customers as far as monetary value is concerned. These Gold customers, at a glance, are probably customers that the store has had for a long period, but starting to lose interest in the store, these customers can be prime candidates for targeted promotions. The management has to drill down on the group of customers looking at what they purchase and find the best ways to reel them back. They can even do a filter Market Basket Analysis just for these customers. The 'Bronze' customers have high recency and low monetary value and frequency, these might be lost customers, management targets promotions for this group to entice them to purchase more and frequently. The ‘Silver’ customers are somewhat similar to 'Gold' customers, but with a less monetary value of 0.947%. A filtered MBA can help uncover the buying habits of these customers, which in turn can bring up some ideas on how to improve the monetary and frequency of this group.

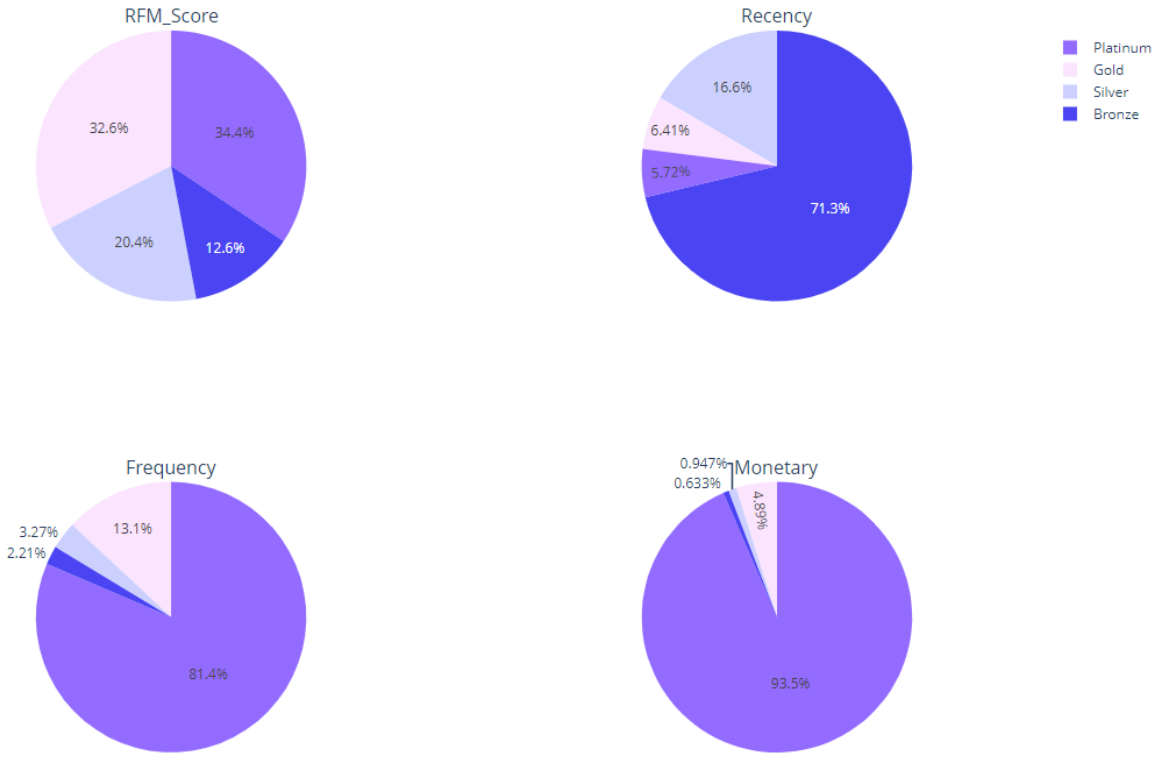

Figure 5.11.RFM metric distributions across customer segments

5.4 Market Basket Analysis

One of the most critical steps in Market Basket Analysis is reshaping the data to suit the analysis needs. In this case, the data frame was rearranged to have the 'InvoiceNo' column as the index, such that each row represents all the items purchased under the same invoice.

This transformation is essential for the subsequent association rule mining, as it allows the analysis to consider the items purchased together in each transaction.The table below shows the first five rows of the data frame after this reshaping process.

Table 5.5.Data reshaping for Market Basket Analysis

	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 DAISY PEGS IN WOOD BOX	12 EGG HOUSE PAINTED WOOD	...	12 IVORY ROSE PEG PLACE SETTINGS
540157	0	0	0	0	...	3
540163	0	0	0	0	...	24
540176	0	0	0	0	...	0
544466	0	0	0	1	...	0
547946	0	0	12	0	...	0

Then a one-hot encoding transformation was applied to the data. This process ensured that any values less than or equal to 0 were converted to 0, while all other values were set to 1.

Table 5.6.Data after one-hot encoding

	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 DAISY PEGS IN WOOD BOX	12 EGG HOUSE PAINTED WOOD	...	12 IVORY ROSE PEG PLACE SETTINGS
540157	0	0	0	0	...	1
540163	0	0	0	0	...	1
540176	0	0	0	0	...	0
544466	0	0	0	1	...	0
547946	0	0	1	0	...	0

Additionally, any transactions that contained less than 2 items were removed from the dataset. This step was taken to focus the analysis on more meaningful transactions that involved the purchase of multiple items.

To identify frequently purchased item combinations, the Apriori algorithm from the MLxtend library was utilized. This algorithm was applied to a one-hot encoded version of the DataFrame, where each column represented a unique product. The minimum support threshold was set at 10%, meaning the algorithm would only consider itemsets that appeared in at least 10% of the transactions.

The Apriori function returns a DataFrame containing the frequent itemsets and their corresponding support values. This provided valuable information about which products are often purchased together by customers.

Building upon the frequent itemsets, the association rules were extracted using the association_rules function from MLxtend, with the resulting output including the antecedent (X), consequent (Y), support, confidence, and lift for each rule.

Table 5.7.Association rules

antecedents	**consequents**	**antecedent support**	**consequent support**	**support**	**confidence**	**lift**
frozenset({'PINK PARTY BAGS'})	frozenset({'BLUE PARTY BAGS'})	0.01235	0.01146	0.01058	0.85714	74.76923
frozenset({'BLUE PARTY BAGS'})	frozenset({'PINK PARTY BAGS'})	0.01146	0.01235	0.01058	0.92308	74.76923
frozenset({'KEY FOB , FRONT DOOR'})	frozenset({'KEY FOB , BACK DOOR'})	0.01146	0.01411	0.01058	0.92308	65.42308
frozenset({'KEY FOB , BACK DOOR'})	frozenset({'KEY FOB , FRONT DOOR'})	0.01411	0.01146	0.01058	0.75000	65.42308

frozenset({'REGENCY TEA PLATE PINK'})	frozenset({'REGENCY TEA PLATE ROSES'})	0.01146	0.01587	0.01058	0.92308	58.15385
frozenset({'REGENCY TEA PLATE ROSES'})	frozenset({'REGENCY TEA PLATE PINK'})	0.01587	0.01146	0.01058	0.66667	58.15385
…	…	…	…	…	…	…

5.5 Product Recommendation

Building upon the insights from market basket analysis and customer segment, the product recommendation component can leverage association rules to deliver the most relevant suggestions to customers. By identifying the consequences with the highest lift for a given antecedent (a product or itemset), the system can recommend the products that customers are most likely to purchase in addition to their current selections.

This approach enables the retailer to provide the classic "People who bought this also bought" recommendations, but with a data-driven, personalized twist. The association rules uncover the hidden connections between products, allowing the recommendation engine to suggest complementary items that are frequently purchased together.

Although the “Platinum” cluster only has 25 customers, they are those who actively contribute to the business's performance. Consequently, investigating their consumer behavior will help the organization maintain and strengthen their engagement. Table 5.8 highlights a selection of products that are commonly purchased together by Platinum customers.

Table 5.8.Product Recommendations for Platinum customers

Product	Product Recommendations
SMALL DOLLY MIX DESIGN ORANGE BOWL	['SMALL CHOCOLATES PINK BOWL']
SMALL HANGING IVORY/RED WOOD BIRD	['SET OF 3 WOODEN SLEIGH DECORATIONS']

SMALL MARSHMALLOWS PINK BOWL	['SMALL CHOCOLATES PINK BOWL']
SMALL POPCORN HOLDER	['GARDENERS KNEELING PAD KEEP CALM']
SMALL PURPLE BABUSHKA NOTEBOOK	['PACK OF 60 DINOSAUR CAKE CASES']
SMALL RED BABUSHKA NOTEBOOK	['SMALL YELLOW BABUSHKA NOTEBOOK']
SMALL YELLOW BABUSHKA NOTEBOOK	['SMALL RED BABUSHKA NOTEBOOK']
SPACEBOY LUNCH BOX	['DOLLY GIRL LUNCH BOX']
STRAWBERRY CHARLOTTE BAG	['CHARLOTTE BAG PINK POLKADOT' , 'RED RETROSPOT CHARLOTTE BAG']
SUKI SHOULDER BAG	['SET OF 3 WOODEN HEART DECORATIONS']
TOILET METAL SIGN	['KITCHEN METAL SIGN', 'BATHROOM METAL SIGN']
WHITE SPOT BLUE CERAMIC DRAWER KNOB	['BLUE STRIPE CERAMIC DRAWER KNOB']
WHITE SPOT RED CERAMIC DRAWER KNOB	['RED STRIPE CERAMIC DRAWER KNOB']
WOODEN HAPPY BIRTHDAY GARLAND	['PAPER BUNTING RETROSPOT']
WOODEN HEART CHRISTMAS SCANDINAVIAN	['WOODEN STAR CHRISTMAS SCANDINAVIAN']
WOODEN STAR CHRISTMAS SCANDINAVIAN	['WOODEN HEART CHRISTMAS SCANDINAVIAN']
WOODLAND CHARLOTTE BAG	['STRAWBERRY CHARLOTTE BAG']
…	…

The "Gold" customers appear to be long-standing, loyal customers who may be starting to disengage with the store. This is a critical customer segment that the management should focus on retaining and re-engaging. Leveraging the insights from the targeted Market Basket Analysis, the retailer can provide personalized product recommendations to the Gold customers based on their purchase history and affinities. This could include suggesting complementary products or bundles that align with their demonstrated preferences.

The analysis also revealed that while both Platinum and Gold customers tend to purchase similar product categories, the Platinum segment tends to have a higher average order value and repeat purchase rate. This suggests they may be more responsive to premium product offerings, exclusive promotions, and personalized service, compared to the Gold segment who may be more price-sensitive. Tailoring the product recommendations, marketing communications, and customer experience accordingly can help maximize engagement and loyalty within each segment. The analysis for the remaining segments is also conducted in a similar manner as for the Platinum and Gold segments.

Table 5.9.Product Recommendations for Gold customers

Product	**Product Recommendations**
ALARM CLOCK BAKELIKE GREEN	['ALARM CLOCK BAKELIKE RED']
ALARM CLOCK BAKELIKE IVORY	['ALARM CLOCK BAKELIKE RED']
ALARM CLOCK BAKELIKE PINK	['ALARM CLOCK BAKELIKE RED']
ALARM CLOCK BAKELIKE RED	['ALARM CLOCK BAKELIKE PINK']
JUMBO BAG APPLES	['JUMBO BAG PEARS']
JUMBO BAG PEARS	['JUMBO BAG APPLES']
JUMBO BAG PINK POLKADOT	['JUMBO STORAGE BAG SUKI']
LUNCH BAG ALPHABET DESIGN	['LUNCH BAG VINTAGE LEAF DESIGN']
LUNCH BAG APPLE DESIGN	['LUNCH BAG VINTAGE LEAF DESIGN']
LUNCH BAG BLACK SKULL	['LUNCH BAG PINK POLKADOT' , 'LUNCH BAG CARS BLUE']
LUNCH BAG CARS BLUE	['LUNCH BAG PINK POLKADOT' , 'LUNCH BAG RED SPOTTY']
PAPER CHAIN KIT 50'S CHRISTMAS	['PAPER CHAIN KIT VINTAGE CHRISTMAS']
PAPER CHAIN KIT VINTAGE CHRISTMAS	["PAPER CHAIN KIT 50\'S CHRISTMAS"]
WOODEN FRAME ANTIQUE WHITE	['WOODEN PICTURE FRAME WHITE FINISH']
WOODEN HEART CHRISTMAS SCANDINAVIAN	['WOODEN STAR CHRISTMAS SCANDINAVIAN']
…	…

In general, customers exhibit a strong tendency to purchase products that are part of the same product suite or collection. Rather than buying items individually, these consumers often opt for coordinating pieces that complement one another, such as variants in different colors, sizes, or related accessories. This bundled purchasing approach allows them to curate a cohesive set of products that integrate seamlessly and maximize the value they derive from the brand. By catering to this preference for a tailored, end-to-end solution, the retailer is able to drive higher average order values and foster deeper loyalty among its most valuable clientele.

The combination of the RFM (Recency, Frequency, Monetary) Model and Market Basket Analysis (MBA) offers retailers a powerful advantage in delivering a more personalized customer experience. Rather than running MBA across the entire customer base with a massive dataset of transactions, this approach focuses the analysis on smaller, segmented groups of customers identified through the RFM Model.

By first categorizing customers into distinct segments based on their purchasing behavior, the retailer can then apply MBA to each specific group. This targeted approach yields more personalized insights and recommendations for cross-selling and up-selling opportunities that are tailored to the unique preferences and habits of each customer segment.

Retailers who implement this combined RFM-MBA method can expect to see improved results in their personalization efforts and more effective product recommendations. The system is able to precisely identify which products are most likely to appeal to a given customer segment, rather than relying on a one-size-fits-all approach. This level of personalization can significantly boost customer engagement, loyalty, and ultimately, sales.

In summary, the strategic integration of RFM and MBA enables retailers to enhance the customer experience through highly targeted, segment-specific recommendations and offers. This personalized approach is a key driver in stimulating cross-sell and up-sell opportunities and fostering stronger, more profitable relationships with their customers.

CHAPTER 6: CONCLUSION AND DEVELOPMENT DIRECTION

6.1 Conclusion

By combining customer segmentation insights with market basket analysis, retailers were able to develop a streamlined, cost-effective approach to delivering personalized product recommendations. This integrated methodology unlocked a straightforward way for retailers to leverage valuable customer and transaction data to inform targeted cross-selling strategies that resonate with different shopper segments. Through this data-driven, customer-centric framework, retailers can drive increased basket sizes, repeat purchases, and long-term brand loyalty.

6.2 Limitations of the research and future development directions

One key limitation of this study is the scope of the customer data available for analysis. While the retail transaction records provided detailed insights into customer purchasing behaviors and market basket patterns, the demographic and psychographic information about the customers was relatively limited. Incorporating a richer set of customer profile data, such as age, income, education, household composition, and personal interests, could have enabled even more refined customer segmentation and personalized product recommendations. However, this access also requires more resources than which the study proposed.

Future research should focus on developing analytical techniques to incorporate a more holistic understanding of the customer base. By blending demographic, psychographic, and behavioral data, retailers can develop an even more robust customer segmentation framework to power personalized product suggestions that drive stronger engagement, loyalty, and revenue growth.

REFERENCES

Aguinis, H. F. (2013). Using Market Basket Analysis in Management Research. *Journal of Management, vol.39,* 1799-1824.

Akbar Telikani, A. H. (2020). A survey of evolutionary computation for association rule mining. *Information Sciences,* 318-352.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,* (pp. 1027-1035). New Orleans, LA, USA.

Aslekar, A. P. (2019). Big Data Analytics for Customer Lifetime Value Prediction. *Telecom Business Review*, 12(1), 46-49.

Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Geoscience and Remote Sensing,* Vol. 3: 9-16.

Gurudath, S. (2020). *Market Basket Analysis & Recommendation System Using Association Rules.*

Jyoti Yadav, M. S. (2023). A Review of K - mean Algorithm. *International Journal of Engineering Trends and Technology (IJETT),* V4(7):2972-2976.

M Qisman, R. R. (2021). Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores). *Journal of Physics: Conference Series*.

Nanavati, A. A. (2016). Association rule mining using hybrid GA-PSO for multi-objective optimisation. *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),* (pp. 1-7). Chennai, India.

Nicholas, K. R. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA),* (pp. 747-748). Sydney, NSW, Australia.

P. Anitha, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 1785-1792.

R. Agrawal, T. a. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD international conference on Management of data*, (pp. 207-216).

Sammydsouza. (2022, 11 23). Retrieved from Course Hero: https://www.coursehero.com/file/179815467/M4-DataAnalyticsRolespptx/

Shabtay, L. F.-V. (2021). A Guided FP-Growth Algorithm for Mining Multitude-Targeted Item-Sets and Class Association Rules in Imbalanced Data. *Information Sciences*, 553, 353-375.

Shutaywi, M., & Kachouie, N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, 23, 759.

Tran, K.-G. &.-H. (2021). *Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model.*

Valle M A, R. G. (2018). Market basket analysis: Complementing association rules with minimum spanning trees. *Expert Systems with Applications, vol.97*, 146-162.

Y. L. Chen, K. T. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems, vol.40, no.2*, 339–354.

Yang, K. P.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, vol. 8, pp. 80716-80727.

Zhang, J. L. (2023). Investigation of Essential Skills for Data Analysts: An Analysis Based on LinkedIn. *Journal of Global Information Management (JGIM), vol.31(1)*, 1-21.

Zhou, J. Z. (2010). Research and Application of Data Mining Based on Web Log. *Science, Technology and Engineering*, 10, 2762-2766.