VIET NAM NATIONAL UNIVERSITY, HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEM

FINAL PROJECT: DATA ANALYTICS IN ELECTRONIC COMMERCE COURSE

TOPIC: Customer Segmentation analysis in Retail Transaction

Instructor: Assoc. Prof. Ho Trung Thanh, Ph.D.

Class: *235MI7101*

Group: *04*

Ho Chi Minh City, July 31st, 2024

Members of Group 04

No.	Full name	Student ID	Task	Contribution
1	Hoàng Minh Thương - Leader	K214110869	**Divide Task** **Chapter 0: Overview** - Structure of project - Tools and Programming languages **Chapter 1: Theoretical Background** **Chapter 2: Data Preparation** - Data understanding - EDA **Chapter 3: Experimental Method/Model and Evaluation** - Overview chapter 3 **Chapter 4: Experimental results and Discussion** - Chapter overview	100%

			- CLV Prediction - Dashboard - Implementation **Chapter 5: Conclusion** **References**	
2	Lê Ngọc Gia Thịnh - Member	K214111329	**Acknowledgments** **Slide** **Gantt chart** **Chapter 0: Overview** - Objectives **Chapter 1: Theoretical Background** - Chapter overview - RFM model **Chapter 2: Data Preparation** - Data description **Chapter 3: Experimental Method/Model and Evaluation** - Cluster evaluation using Silhouette	100%

			Chapter 4:Experimental results and Discussion - Implementation **Chapter 5: Conclusion** **References**	
3	Lê Ngọc Gia Yên - Member	K214111330	**Word Contribution** **Commitment** **Slide** **Chapter 0: Overview** - Objects and scopes **Chapter 1: Theoretical Background** - Customer Behavior - Churn Rate **Chapter 2: Data Preparation** - Data collection **Chapter 3: Experimental Method/Model and Evaluation** - Labeling	100%

			Chapter 4: Experimental results and Discussion - Implementation **References**	
4	Lê Việt Dương - Member	K214111969	**Check all code** **Chapter 0: Overview** - Experimental/Research method **Chapter 1: Theoretical Background** - CLV – customer retention **Chapter 2: Data Preparation** - EDA **Chapter 3: Experimental Method/Model and Evaluation** - RFM Calculation - Data transformation - Data scaling	100%

			Chapter 4: Experimental results and Discussion - Implementation **References**	
5	Nguyễn Ngọc Bảo Trân - Member	K214110871	**Word** **Chapter 0: Overview** - Business Objectives - Business Question **Chapter 1: Theoretical Background** - Customer Segmentation - Elbow method - Silhouette method **Chapter 2: Data Preparation** **Chapter 3: Experimental Method/Model and Evaluation** - The optimal number of clusters by Elbow method	100%

			Chapter 4: Experimental results and Discussion - Cohort Analysis - Dashboard **References**	
6	Trần Nguyễn Thiên Vũ - Member	K214111980	**Check all code** **Chapter 0: Overview** - Business problems/Challenges **Chapter 1: Theoretical Background** - K-Means Clustering **Chapter 2: Data Preparation** - Data model **Chapter 3: Experimental Method/Model and Evaluation** - Segmentation using K-means **Chapter 4: Experimental results and Discussion**	100%

			- Implementation **References**	

Acknowledgments

We would like to express our sincere gratitude to Mr. Ho Trung Thanh, our distinguished supervisor, for his tremendous help and direction during the course of this project. His steadfast dedication and astute counsel were crucial in making this project a success. The accomplishment of this endeavor would not have been possible without his help.

Furthermore, we have been greatly motivated by Mr. Thanh's lectures, which have stoked our enthusiasm for the subject while simultaneously dispensing knowledge. His passion and commitment have served as a continuing source of inspiration for us.

We realize that mistakes may occur and that there is always room for improvement, even with our greatest efforts to complete the project successfully. We sincerely appreciate any helpful criticism and recommendations for additional improvement.

July *31st* 2024, Ho Chi Minh City

Regards,

Group 04

Commitment

The authors hereby certify that the topic: "Customer Segmentation analysis in Retail Transaction" is conducted publicly with the efforts of the authors and the enthusiastic support and guidance from Assoc. Prof. Ho Trung Thanh, Ph.D.

In addition, all data and research results are completely honest, only serve research purposes and there is no copying or using the results of other similar research topics. At the same time, all theories and supporting information for the construction of the theoretical basis and theoretical model for the research paper are fully cited, clearly stated and permitted to be published.

Ho Chi Minh City, July *30st* , 2024

Group 4

Table of Content

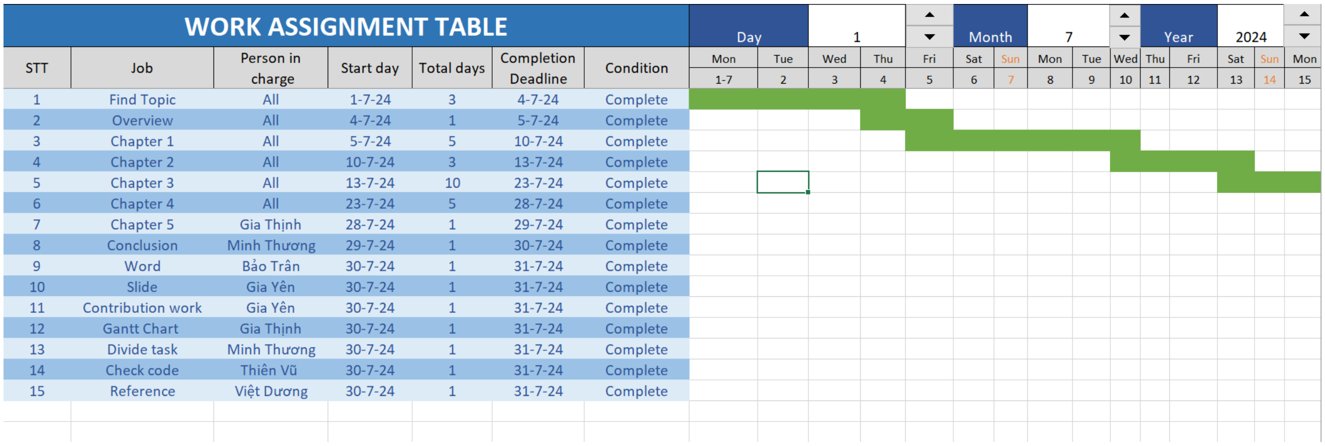
List of Tables

List of Figures

List of Abbreviations

DB	Digital Business
EC	Electronic Commerce
RFM	Recency Frequency Monetary
CS	Customer Segmentation
EDA	Exploratory Data Analysis
CRM	Customer Relationship Management
CLV	Customer lifetime value
SSE	Sum of Squared Error

GANTT CHART



WORK ASSIGNMENT TABLE							Day		1		▲▼	Month		7		▲▼	Year		2024		▲▼
STT	Job	Person in charge	Start day	Total days	Completion Deadline	Condition	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon
							1-7	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Find Topic	All	1-7-24	3	4-7-24	Complete	🟩	🟩	🟩	🟩											
2	Overview	All	4-7-24	1	5-7-24	Complete				🟩	🟩										
3	Chapter 1	All	5-7-24	5	10-7-24	Complete					🟩	🟩	🟩	🟩	🟩	🟩					
4	Chapter 2	All	10-7-24	3	13-7-24	Complete										🟩	🟩	🟩	🟩		
5	Chapter 3	All	13-7-24	10	23-7-24	Complete													🟩	🟩	🟩
6	Chapter 4	All	23-7-24	5	28-7-24	Complete															
7	Chapter 5	Gia Thịnh	28-7-24	1	29-7-24	Complete															
8	Conclusion	Minh Thương	29-7-24	1	30-7-24	Complete															
9	Word	Bảo Trân	30-7-24	1	31-7-24	Complete															
10	Slide	Gia Yên	30-7-24	1	31-7-24	Complete															
11	Contribution work	Gia Yên	30-7-24	1	31-7-24	Complete															
12	Gantt Chart	Gia Thịnh	30-7-24	1	31-7-24	Complete															
13	Divide task	Minh Thương	30-7-24	1	31-7-24	Complete															
14	Check code	Thiên Vũ	30-7-24	1	31-7-24	Complete															
15	Reference	Việt Dương	30-7-24	1	31-7-24	Complete															

ABSTRACT

The home decor, electronics, and book markets have demonstrated remarkable resilience over recent years despite economic fluctuations and are projected to flourish in the long term. To excel in this increasingly competitive landscape, businesses have adopted numerous strategies to attract more customers and boost revenue. One of the key strategies in marketing analytics is customer segmentation, which categorizes customers based on various factors to enhance business performance. RFM (Recency, Frequency, and Monetary) Analysis is a prominent method for segmenting customers based on their past behavior. With the increasing availability of historical transactional data, RFM Analysis is effectively utilized for consumer segmentation, aiding data-driven decision-making in businesses. Additionally, technological advancements have spurred growth across many fields, particularly in data science, where machine learning techniques have significantly impacted business development. In this study, leveraging a dataset from an online homeware and gift retailer, our team employs the K-means algorithm—a machine learning model—to segment customers according to the RFM method. Based on our findings, we will offer recommendations to retailers for enhancing business performance.

ABSTRACT

Thị trường trang trí nhà cửa, điện tử và sách đã cho thấy sự kiên cường đáng kể trong những năm gần đây dù có sự biến động kinh tế và dự đoán sẽ phát triển mạnh mẽ trong thời gian dài. Để vượt trội trong thị trường cạnh tranh ngày càng gia tăng này, các doanh nghiệp đã áp dụng nhiều chiến lược để thu hút thêm khách hàng và tăng doanh thu. Một trong những chiến lược quan trọng trong phân tích tiếp thị là phân khúc khách hàng, phân loại khách hàng dựa trên các yếu tố khác nhau để nâng cao hiệu suất kinh doanh. Phân tích RFM (Tần suất, Độ gần và Giá trị tiền tệ) là một phương pháp nổi bật để phân khúc khách hàng dựa trên hành vi trong quá khứ của họ. Với sự gia tăng của dữ liệu giao dịch lịch sử, phân tích RFM được sử dụng hiệu quả trong phân khúc người tiêu dùng, hỗ trợ việc ra quyết định dựa trên dữ liệu của doanh nghiệp. Bên cạnh đó, các tiến bộ công nghệ đã thúc đẩy sự phát triển trên nhiều lĩnh vực, đặc biệt là khoa học dữ liệu, nơi các kỹ thuật máy học đã ảnh hưởng đáng kể đến sự phát triển kinh doanh. Trong nghiên cứu này, tận dụng bộ dữ liệu từ một nhà bán lẻ trực tuyến về đồ gia dụng và quà tặng, nhóm của chúng tôi sử dụng thuật toán K-means - một mô hình máy học để phân khúc khách hàng theo phương pháp RFM. Dựa trên các phát hiện của chúng tôi, chúng tôi sẽ đưa ra các đề xuất cho các nhà bán lẻ nhằm cải thiện hiệu suất kinh doanh.

PROJECT OVERVIEW

Business problems/Challenges

In the contemporary competitive business environment, businesses are faced with the complex problem of identifying and prioritizing high-value clients that make a substantial contribution to overall revenue in the current competitive business climate. The complexity arises from the vast and diverse customer base, which requires sophisticated analytical tools and methodologies to discern patterns and identify those customers whose purchasing behaviors and interactions drive the most value. This challenge is further exacerbated by a persistently high customer churn rate, where a substantial portion of customers discontinue their relationship with the company, leading to considerable revenue losses and increased acquisition costs to replace these customers.

Furthermore, the inability to deliver personalized customer experiences remains a critical impediment to achieving high customer satisfaction and loyalty. Generic, one-size-fits-all methods fail to address specific client preferences and expectations, leading in lower engagement and retention. The absence of customization weakens efforts to develop strong, long-term customer connections, which are critical for maintaining a competitive edge. Therefore, addressing these interconnected issues through the implementation of advanced data analytics and customer relationship management strategies is imperative for enhancing customer value, reducing churn, and ultimately driving sustainable business growth.

Business Objectives

A customer segmentation model allows organizations to target specific groups of customers, allowing for more effective marketing resource allocation and the maximization of cross- and up-selling capability (Tabianan, 2022). Our project is carried out with the main goal of maximizing the potential of the RFM model ((Recency, Frequency, and Monetary) to enable businesses to handle three challenges through steps from collecting data to achieve optimal results in retail activities. Specifically, the project focuses on three main objectives:

Firstly, by analyzing and segmenting customers based on the RFM model, the result will not only identify and focus on high-value customers but it also identifies those who have not made recent purchases, meaning those customers have low recency scores and implement targeted retention campaigns to re-engage them.

Secondly, a solid data analysis based on the RFM model will help create personalized shopping experiences based on RFM segments, catering to the unique preferences and behaviors of different customer groups.

Finally, to attract existing customers with low retail demand, it is necessary to research and revise retail policies.

Business Questions

- *How to increase repeat purchases from high-value customer groups in the next 3 months?*

By using the RFM model to identify high-value and frequent customer groups and analyze purchase history to better understand customer habits and preferences.

=> Outcomes: Understand the purchasing habits and behavior of the target customer group, create or optimize a loyalty program to motivate repeat transactions, apply personalization tactics in marketing and customer care to enhance engagement and create a better shopping experience. Continuously monitor the effectiveness of these strategies and make timely adjustments to ensure the goal of increasing repeat purchase rates is achieved.

- *How to personalize customer experience based on RFM segments?*

Based on the RFM model, divide customers into different segments. Then, apply personalized strategies for each segment to provide customer care that is more suitable for each specific customer group.

=> Outcomes: Revenue and profit growth, reduce marketing costs and customer retention costs instead of focusing on finding new customers, improve customer loyalty and satisfaction. Providing personalized experiences helps brands differentiate and stand out fro*m their competitors.*

- *What behavioral characteristics of customers in the current market are applicable to the new market?*

Analyze current transaction data to create RFM profiles for each customer in each region. Then divide customers into groups based on RFM values. Customer groups with a high RFM index can be the target customer group for similar marketing campaigns in new markets.

=> Outcomes: Better understand customer behavior, reduce risks when expanding into new markets with well-founded strategies to increase the likelihood of success, increase efficiency and optimize marketing costs while contributing to increased customer experience.

Objectives

The results of this project will provide an interdisciplinary research model based on the method of exploiting the RFM model, K-means and through customer segmentation analysis data while optimizing business strategy. marketing campaigns, services, promotions, offering new products, services, and appropriate solutions for each customer group.

Objects and scopes

- Objects:

  Customer segmentation through customer shopping behavior in retail.

- Scopes:

  Time scope: Marketing market research from 1/7/2024 to 02/08/2024.

  Space scope: Customer priority for new products in target retail marketing.

Experimental/Research method

- Theoretical analysis, synthesis methods: Books, articles, and research papers about the RFM model and the machine learning methods that enable customer clustering. Identify the most effective machine learning model for grouping customers efficiently.

- Quantitative research method: Gathering data by using certain characteristics: Customer ID a consumer may only own an ID. we may ascertain the Recency by combining the time of the customer's purchase with the model execution time. Every order that a consumer order has a OrderID, which allows the goods to be further distinguished throughout a transaction. We are able to determine a customer's buy frequency as Frequency by using the CustomerID and OrderDate.

Tools and Programming languages

Tools: Colab, PowerBI, Excel, Jupiter

Programming language: Python

Structure of project

The topic is broken up into the following four sections, in addition to the introduction, conclusion, acronym catalog, table and chart catalog, and bibliography of references:

- *Chapter 1: Theoretical Background*

The theoretical framework that we use and refer to in order to carry out this research will be outlined in this chapter. In addition, this chapter shows how data preparation is done, including how we interpreted the provided data set that we chose to employ for the study.

- *Chapter 2: Data preparation*

This chapter explains the steps involved in preparing data, such as how we interpreted the provided data set, which data (or columns) we selected to use, how we conducted data exploration, and how we preprocessed the data for the study.

- *Chapter 3: Experimental method/model and Evaluation*

The model/research/experimental procedure, the machine learning model/method, the model parameters, the model experimental findings, and the model evaluation using metrics are all covered in this chapter.

- *Chapter 4: CLV Prediction and Visualization and Data storytelling*

This chapter offers workable ideas to maximize corporate growth strategies by fusing data insights with business analysis and CLV prediction.

- *Chapter 5: Conclusions & Further orientation*

In this chapter, we will provide specific conclusions and future directions that can be applied, helping companies increase customer lifetime value, increase revenue and improve operations.

CHAPTER 1. THEORETICAL BACKGROUND

Chapter overview

The theoretical underpinning that we have carefully considered and chosen for our team to use in carrying out this research will be presented in detail and methodically in this chapter. To create a strong foundation for our research, our team will first investigate the major theories and theories that are related to them. These theories serve as the foundation for the team's analysis and interpretation of the data they collect, in addition to helping to precisely guide the research methodology.

Furthermore, this chapter will provide a detailed overview of the data preparation process, starting with the initial comprehension and familiarization with the dataset and ending with the particular procedures we take to process and clean the data. We will go into great depth about the strategies and tactics employed to guarantee the data's dependability and correctness, as well as the crucial choices made when choosing and utilizing it for the study. Achieving worthwhile and trustworthy research outcomes requires a thorough understanding of and meticulous preparation of data.

1.1. Customer Segmentation

Customer segmentation plays an important role in the first step of CRM (Customer Relationship Management) which is Customer Identification (Tavakoli et al., 2018). The process of customer segmentation involves dividing customers into different groups based on demographics, customer purchasing behavior, and more. The most common method for analyzing this process is the RFM method. Customer segmentation is a specific application of data mining, using data mining techniques to divide customers into groups with common characteristics. Data mining helps optimize and improve the accuracy of the customer segmentation process by applying complex algorithms and models to find hidden patterns and relationships in data. (Ngai, Xiu, & Chau, 2009)

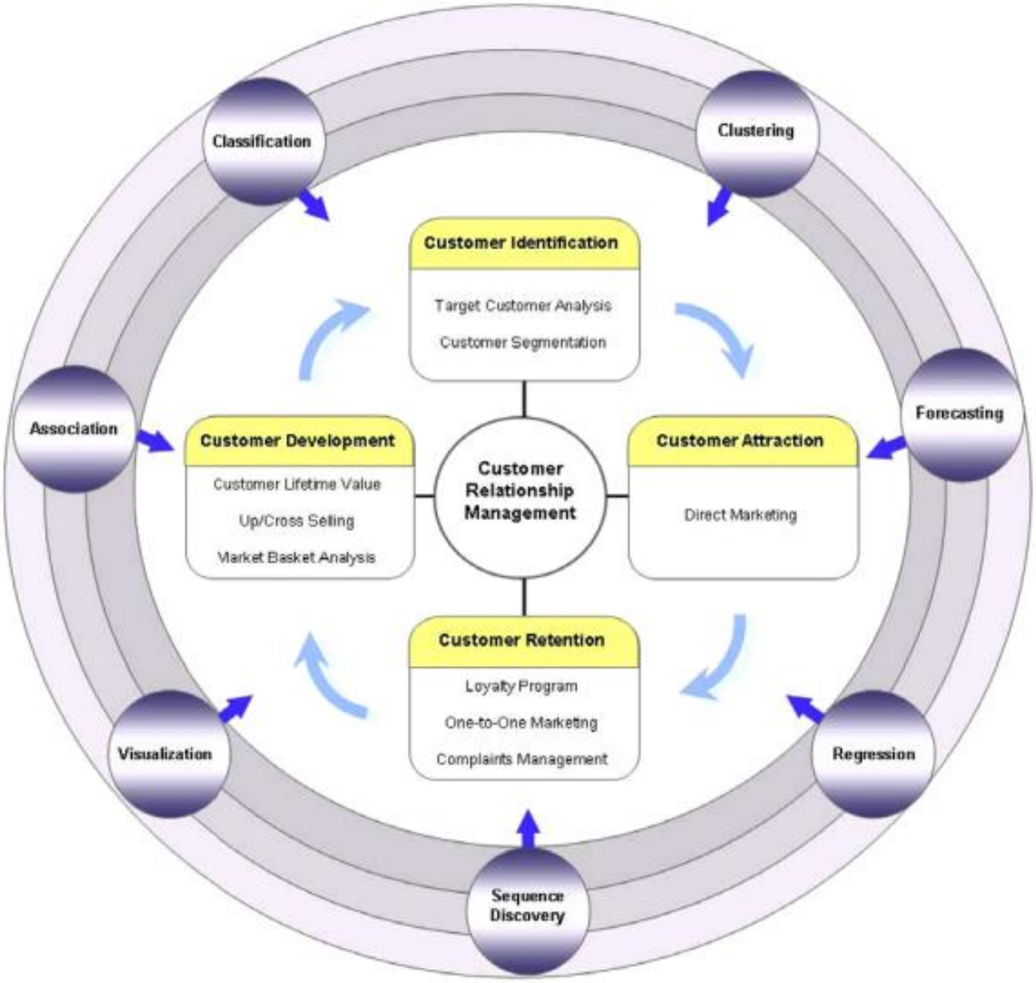


Figure 1: Classification framework for data mining techniques in CRM (Ngai, Xiu, & Chau, 2009)

According to Chen, Zhang, Hu, & Wang, (2006), advantages of the customer segmentation model based on data mining included improving promotion effect, analyzing customer value and customer loyalty, analyzing credit risk, instructing new products R&D, confirming target market.

1.2. RFM model

RFM (recency, frequency and monetary) model is a behavior-based model used to analyze the behavior of a customer and then forecast future actions based on behavior recorded in the database (Hughes, 1996; Yeh et al., 2009). Furthermore, according to Wang (2010), recency is the amount of time that has passed since the last transaction, frequency is the number of purchases made within a certain time period, and monetary is the total amount of money spent during this same time period.

The RFM model is the most commonly used segmentation technique, consisting of three measures: recency, frequency, and monetary. These measures are combined into a three-digit RFM cell code, spanning five equal quintiles. Reviewing the RFM model is essential, as it provides valuable insights for researchers and decision-makers. In fact, the RFM model has proven to be highly successful across various practical applications. Consequently, RFM can help identify valuable customers and develop effective marketing strategies for both profit-driven organizations (such as those in the marketing, banking, insurance, telecommunications, travel, and online industries) and non-profit organizations, as well as government agencies. The RFM model consists of three factors, with scores for each indicator ranging from 1 to 5. However, it is common practice to use only two out of these three factors. Examples include combinations such as RF, RM, and FM (Wei, J. T., Lin, S. Y., & Wu, H. H. ,2010).

R (Recency): This measures the time elapsed since the customer's last purchase. A lower value indicates a higher likelihood of the customer making a repeat purchase.

F (Frequency): This measures the number of purchases made within a specific period. A higher frequency signifies greater customer loyalty.

M (Monetary): This measures the amount of money spent by the customer within a certain period. A higher value suggests that the company should prioritize this customer.

There are 5 steps to performing RFM analysis:

- Step 1: Collect, collate data, and specify recency, frequency, and currency values:

The first step in building an RFM model is to assign Recency, Frequency, and Currency values to each customer. Analyzing a customer's transaction history is part of the RFM concept. Obtaining the RFM data for every customer in ascending order is the first step.

- Step 2: RFM Configuration

To properly segment their consumer base, businesses must develop bespoke filters. This is a crucial component that will change depending on their line of work.

- Step 3: Divide customers and assign scores

The customer list is split into tier groups for each of the three dimensions (R, F, and M) in the third stage. Based on the table, you may now grade each customer individually. By doing this, you're using RFM to break up transactions into groups of related transactions rather than their absolute values.

- Step 4: Labeling customer groups

The labels that are used will depend on the various attributes of the three grades that the clients have been given. Based on the RFM segments in which they exist, the client groups to which particular kinds of messages should be directed.

- Step 5: Creating customized strategies/tactics for relevant segments

Businesses may guarantee personalization in all of their messaging once they have divided and identified each customer. Loyal customers might feel more appreciated by receiving greater service, while at-risk clients can be targeted with offers, discounts, or freebies. RFM marketing enables more effective communication between marketers and customers by concentrating on the behavioral patterns of specific groups.

In the study of Miglautsch (2000), the client quintile approach is the name given to the RFM rating system. Sorting consumers using the customer quintile method involves going from best to worst. RFM score aims to predict future behavior, which will lead to better segmentation decisions. It's critical to convert client behavior into figures that can be used throughout time in order to enable projection.

1.3. Customer Behavior

Customer behavior is the study of all the behaviors and emotions that customers display while they shop, investigate, and assess things to meet their requirements. The field of consumer behavior studies how individuals, groups, and organizations select, buy, use, and dispose of goods, services, ideas, or experiences to satisfy their needs and desires (Durmaz, Y., & Diyarbakırlıoğlu, 2011). Understanding consumer behavior is never simple, because customers' behaviors are changeable.

1.4. CLV – customer retention

Customer lifetime value (CLV) is a business measure used to identify the amount of money customers will spend on companies's products or service over time. CLV is vital in evaluating the potential worth of consumers and motivating organizations to find more about the patterns of individuals or groups of customers. It is an inward-looking view of the consumer and is based on the insight that views customers in terms of on-going long term relationships and not just short-term transactions. It is a predictive tool that provides forward-looking information on customer relationship performance and resource allocation. Estimating the lifetime value of a customer involves predictions of both revenues and customer retention probabilities (Kaul, 2017).

Mahboubeh Khajvand et al (2011) have suggested 2 strategis. The first strategy uses the RFM marketing analysis tool to categorize customers. Moreover, the second strategy suggests expanding the RFM analysis method by adding a new parameter called Count.

The comparative results demonstrate that changing parameters has no effect on the research findings. Customer Lifetime Values (CLV) are therefore determined using the combined weighted RFM technique K-Means.

1.5. K-Means Clustering

K-means is a well-known clustering algorithm in data mining, extensively utilized for organizing large datasets into clusters. Initially proposed by MacQueen in 1967, the k-means algorithm is regarded as one of the simplest unsupervised learning algorithms and has been applied to address the widely recognized problem of clustering (Sun, 2008). The k-means algorithm segments the input dataset into k clusters, each represented by a centroid that is dynamically updated. These centroids begin from initial values referred to as seed points. The algorithm calculates the squared distances between the input data points and the centroids, subsequently assigning each data point to the closest centroid (MacQueen, J, 1967).

This process takes place in turn in 2 stages as follows (Kantardzic, 2003):

- Stage 1: Initialize K elements representing $\{\mu h\}h\ K=1$ for K clusters, each μj plays a role as center for cluster Cj.

- Stage 2: Repeat the process of repositioning the data object to the cluster with the nearest center and recomputing the center until the clusters do not change. This process is essentially local minimization of the objective function (the distance index here is the Euclidean index) as follows:

$$E = \sum K\ h=1 \sum xi \in Xh\ ||xi - \mu h||\text{^}2$$

1.6. Elbow method

Elbow searches for the best number of clusters according to the K-means method which the K-means method cannot do on its own (Syakur, Khotimah, Rochman & Satoto, 2018). This technique plots the explained variation as a function of the number of clusters and picks the elbow of the curve as the number of clusters to use. The first clusters will add a lot of information but at some point the marginal gain will decrease significantly and create a corner on the graph (Nainggolan, et al 2014). According to Yusuf Bakhtiar, et al (2016), the elbow method is expressed by Sum of Squared Error (SSE), is one of the statistical methods used to measure the total difference from the actual value of the achieved value:

$$\text{SSE} = \sum_{K=1}^{K} \sum_{x_i \in S_K} || X_i - C_K ||_2^2$$

With k = many clusters formed = the i-th cluster, x = the data present in each cluster.

1.7. Silhouette method

Silhouette is used to evaluate the quality of clustering. According to Rousseeuw (1987), each cluster is represented by something called a Silhouette. This silhouette is based on a comparison of the compactness and separation of groups. It shows which objects belong to which clusters and which objects lie in between clusters. To build a Silhouette chart, two main elements are needed: the partition resulting from the clustering process and the set of distances between data points.

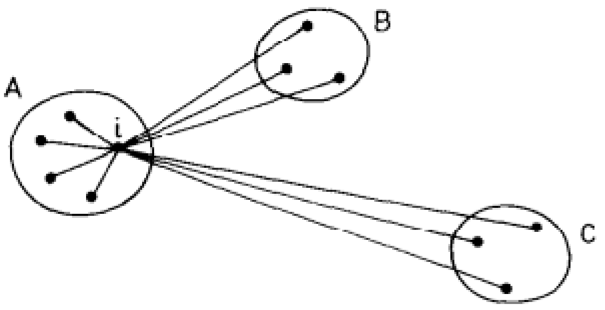


Figure 2: An example illustration of the elements involved in the computation of s(i), where the object i belongs to cluster A. (Source: Rousseeuw, 1987)

Business problems:

- Difficulty in identifying and prioritizing high-value customers who contribute significantly to revenue. High customer churn rate leading to loss of revenue.

- Generic customer experiences leading to low customer satisfaction and loyalty.

- Uncertainty about customer behavior in new geographic markets.

2.1 Data understanding

This data comes from the company's Sales and Marketing Department, and the precise information in the dataset helps with operations like monitoring sales patterns, assessing the efficiency of marketing campaigns, and to maximize business strategy.

2.2. Data collection

An essential first step in every data analysis endeavor is data collection. For our customer segmentation project, we obtained historical sales transaction data from various sources. This dataset covers a period from April 29th, 2023 to April 28th, 2024. This timeframe was chosen to ensure a comprehensive analysis of customer behavior over multiple years, capturing seasonal trends and changes in purchasing patterns.

The primary sources of our data include the Sales Department and Marketing Department. The Sales Department maintains comprehensive records of all customer transactions, providing a foundational dataset for our analysis. From the CRM System, we extracted customer-specific data such as Customer ID and Transaction Details, ensuring we have detailed information about our customer base. Additionally, the Inventory Management System provided essential product-related information, including Product ID and Product Category, allowing us to have a thorough understanding of our inventory.

About Data Fields: We have 10 entities

- CustomerID: Unique identifier for each customer.
- ProductID: Unique identifier for each product sold.
- Quantity: Quantity of the product sold.
- Price: Price of the product sold.
- TransactionDate: The date when the transaction occurred.
- PaymentMethod: Method of payment used for the transaction.
- StoreLocation: Location of the store where the transaction occurred.
- ProductCategory: Category of the product sold.
- DiscountApplied(%): Discount percentage applied to the transaction.
- TotalAmount: Total amount for the transaction after discount.

The dataset comprises 10000 rows and 10 columns, providing a robust basis for analysis. It includes both categorical and numerical attributes, essential for performing RFM analysis and subsequent clustering. The data underwent thorough validation and cleaning to ensure high quality, reliability, and accuracy for analysis.

This thorough approach to data collection guarantees that we have a rich dataset for consumer behavior analysis, efficient customer segmentation, and the extraction of useful business insights. To accomplish the goals of our research, the next phases will involve developing a model, performing exploratory data analysis (EDA), and preparing the d**ata.**

2.3. Data description

The team obtained a set of data, including comprehensive details about sales transactions, from the sales department during the course of implementing the data analysis project. The following characteristics are included in this data, which is compiled from a worldwide retail footprint:

CustomerID: A special number that only each customer has. beneficial for behavior analysis and client segmentation.

ProductID: A special number assigned to every product. aids in monitoring the performance of products.

Quantity: The total number of goods bought. crucial for managing inventories and computing total sales.

Cost: The cost of a single unit of the good. required in order to calculate revenue.

Date of Transaction: The day the transaction took place is beneficial for studies on seasonality and trend analysis.

Payment Method: The way money is paid (cash, credit card, etc.). aids in comprehending payment inclinations.

StoreLocation: The store's address where the transaction was completed. helpful for analyzing sales data geographically.

Product Category: The category that the item falls under. aids in the analysis of category performance.

Discount applied (%): The product's percentage discount. crucial to comprehending the impact of promotion.

By gathering and examining this data, the team is better able to have a broad understanding of the business operations of the shop and use that understanding to make strategic decisions that will improve productivity and the sales process. Boost revenue and enhance the consumer experience.

Our problem is how to use the RFM model in conjunction to identify high-value consumers, lower the customer churn rate, and tailor the customer experience. Model K-means that in addition to our business queries, we define the following datasets to satisfy these requirements:

a) The data set needs to be sufficiently big to enable K-means machine learning to be used for clustering.

b) To facilitate identification and clustering, the dataset needs to include identifying information (such as customer ID, product ID, etc.).

c) Variables such as customer transaction code, customer purchase date and customer purchase amount must be present in the data collection to calculate the elements (TransactionDate, Price, etc.)

Because of this, we made the decision to choose and modify the dataset in order to fulfill the aforementioned specifications and include the variables required to run the K-means model in conjunction with RFM. With the variables CustomerID to ascertain customer transaction information, TransactionDate to ascertain recency value, Price and Quantity to ascertain monetary value, and ProductID and CustomerID to ascertain frequency value for each using Group By algorithm, the data set satisfies the requirements. Furthermore, the 10000 line dataset satisfies the requisite data size for application in machine learning models and process implementation.

The dataset contains historical sales transaction data from April 29th, 2023 to April 28th, 2024 with 100,000 rows and 10 columns. The variables include both numeric and categorical data types. Below are the specifics about the data types and descriptions of every variable:

Variable	Types of data	Description
CustomerID	Categorical	A unique identifier for each customer
ProductID	Categorical	A unique identifier for each product
Quantity	Numeric	Number of items purchased
Price	Numeric	Price of a single unit of the product
TransactionDate	Categorical	Method of payment used (e.g., credit card, cash)
PaymentMethod	Categorical	A unique identifier for each customer.
StoreLocation	Categorical	Location of the store where the purchase was made
ProductCategory	Categorical	Category to which the product belongs
DiscountApplied(%)	Numeric	Percentage discount applied to the product

TotalAmount	Numeric	Total amount spent on the transaction after discounts.

Table 1: Data types and descriptions

The ProductID column data does not treat a row as a product transaction code since the majority of the transactions in the data set are orders for the purchase of significant quantities of goods. All things considered, the data set meets the requirements needed to use the RFM approach to assist in client segmentation. Aspects like ProductName, TransactionDate, which include product details and the customer's product purchase date to facilitate computations, and CustomerID, which is used to identify consumers, are among the attributes in the data. Additionally, frequency data like the number of purchases made each week, month, year, etc., can be computed using these variables. We can determine the order value that the consumer paid for using the Quantity and Price characteristics, and then we can utilize that information to determine the Monetary Value (Purchase Value for each customer). Null values, duplicate values, and mismatched values can all occur in datasets. In addition, there may be correlations between the attributes in the data set. In Chapter 3, the data collection will be cleaned in preparation for the study model. Statistical techniques and data visualization will be used to understand the relationship between variables.

2.4. EDA

- Histogram:

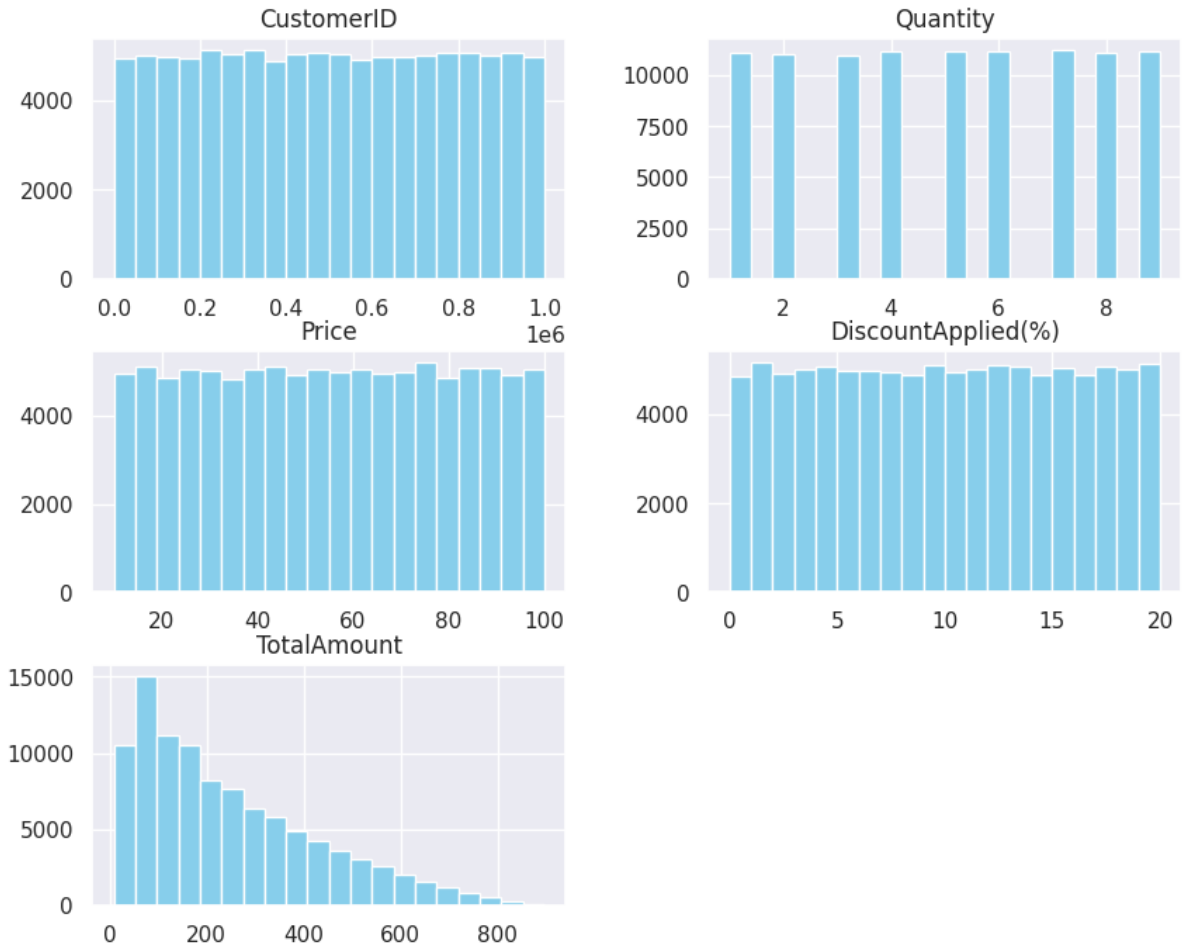


Figure 3: Histogram of factors in dataset

These histograms demonstrate the equal distribution of CustomerID, Quantity, Price, and Discount Applied (%), while the TotalAmount has a tilt towards lower values, suggesting that the majority of purchases are for smaller sums. This can assist in finding trends in the dataset and in analyzing the purchase behavior of customers.

- Total sales quantity by product category:

Product category	Total amount

Books	126047
Electronics	125347
Clothing	125044
Home Decor	124491

Table 2: The total sales quantity by product category

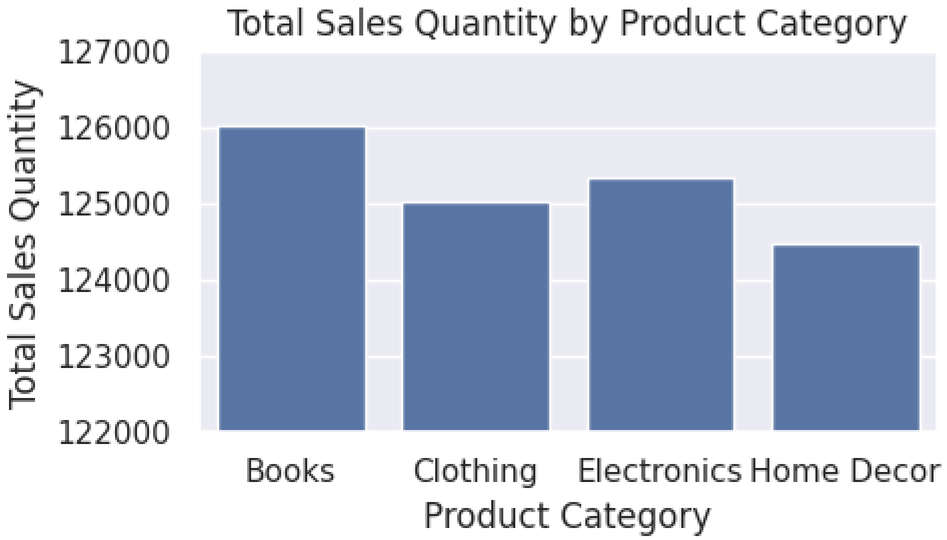


Figure 4: The total sales quantity by product category

This bar chart shows how sales amounts for several product categories are clearly compared; books have the highest sales, followed by electronics, clothing, and home decor. Marketing plans, product demand patterns, and inventory management can all benefit from this knowledge.

- Percentage of total revenue by product category:

Product category	Percentage of revenue
Books	25.2%

Electronics	24.99%
Clothing	24.95%
Home Decor	24.86%

Table 3: Percentage of total revenue by product category

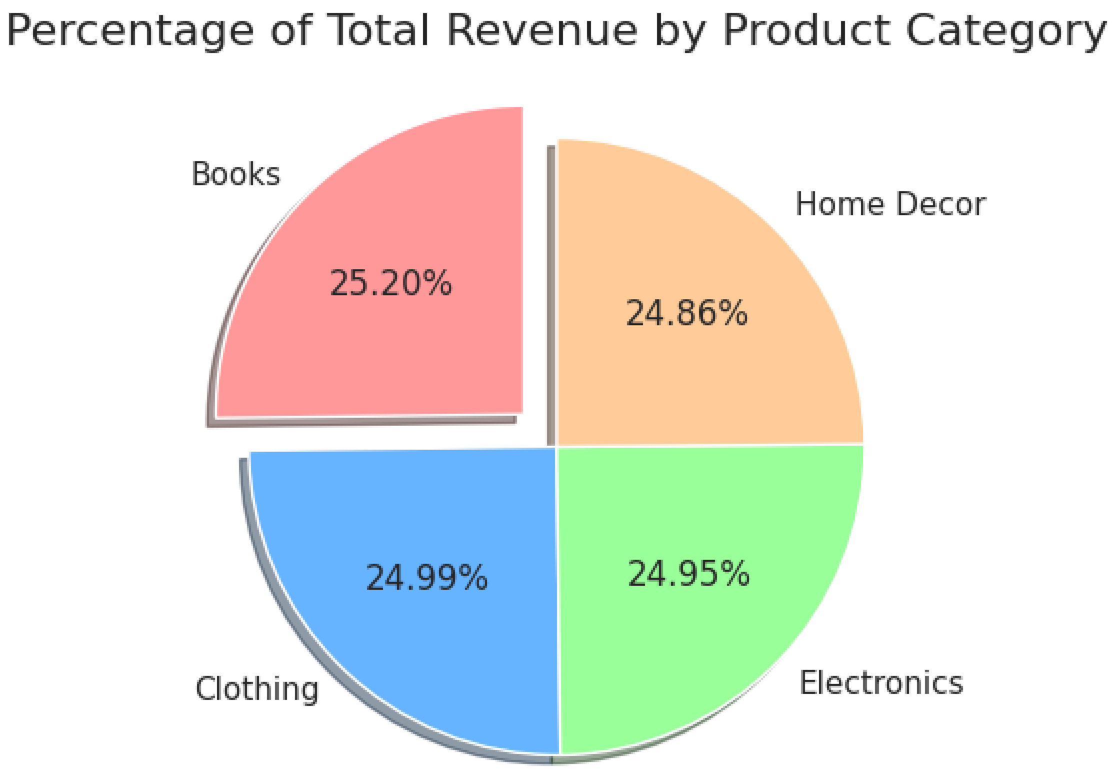


Figure 5: Percentage of total revenue by product category

Each of the four categories contributes between 24.86% and 25.20% of the overall income, with the pie chart showing a reasonably equitable distribution of money among them. All product categories appear to be doing well and making a good amount of money, while books seem to be doing better than the others, with 25.20%. Using this data to influence decisions about product focus and marketing initiatives might be beneficial.

- Total revenue by month:

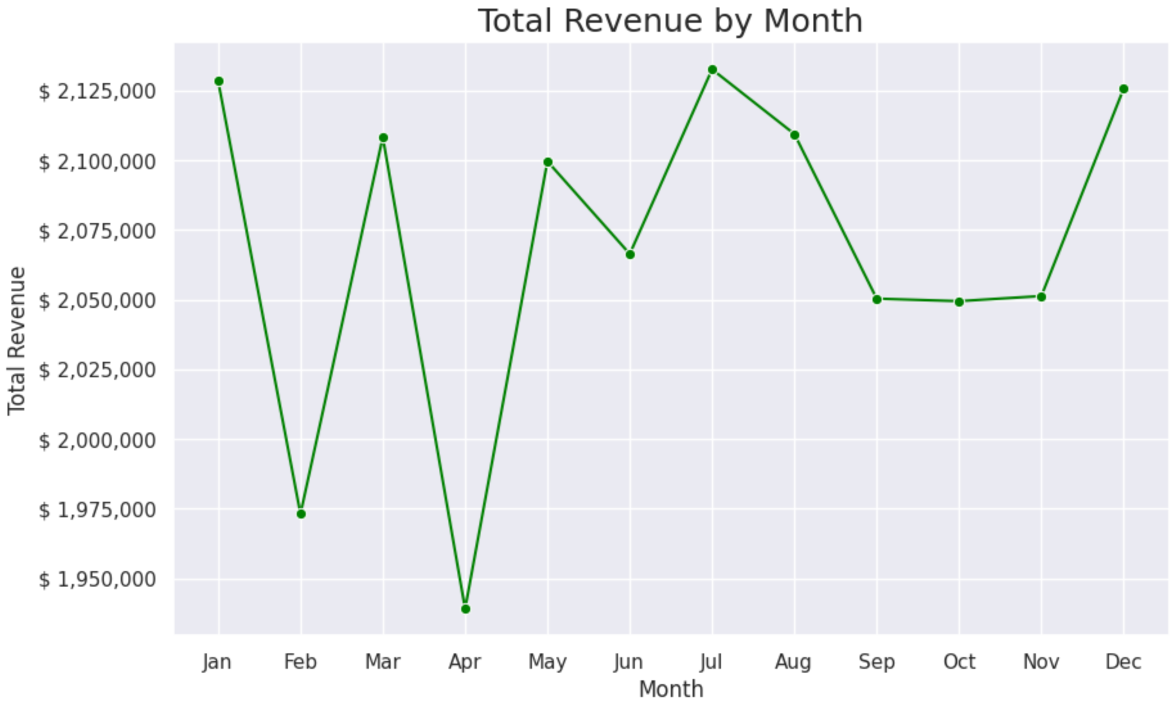


Figure 6: Total revenue by month

This line graph illustrates the year-over-year variations in overall income, with January, March, and July being the highest points and April and the late summer, early fall months representing the lowest. This pattern can point to seasonal fluctuations or other elements influencing sales results at various points throughout the year.

- Average daily sales by product category:

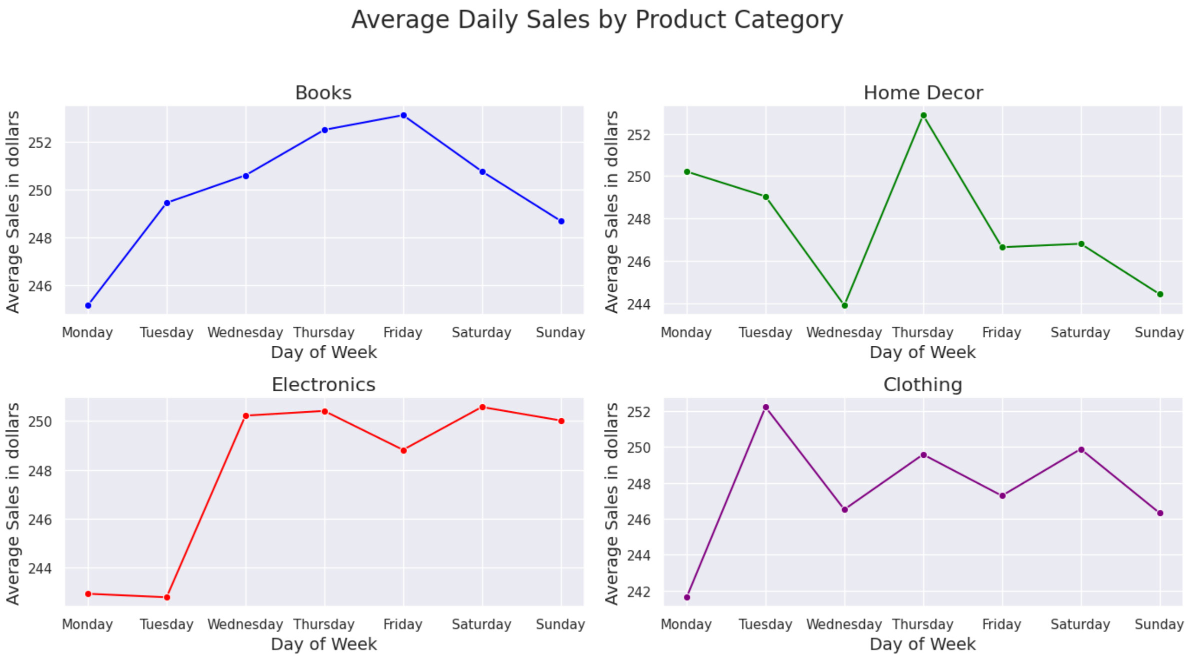


Figure 7: Average daily sales by product category

These line graphs illustrate the average daily sales for each product category in clear patterns. Books indicate a consistent rise during the workday and a little fall during the weekend. Significant sales of home decor occur during the week, and those purchases sharply decrease during the weekend. Electronics rise in the middle of the week and then often decline throughout the weekend. Sales of clothing fluctuate the greatest over the week, with noticeable increases and decreases. These patterns, which show the ideal days to promote each category, can offer insightful information for inventory control and sales tactics.

2.5. Clear data

Describe quartile description of columns in dataset:

	CustomerID	Quantity	Price	TotalAmount

count	100000	100000	100000	100000
mean	500463.98	5.00	55.06	248.33
min	14.00	1.00	10	8.27
25%	250693.75	3.00	32.55	95.16
50%	499679	5.00	55.11	200.36
75%	751104.75	7.00	77.46	362.00
max	999997	9.00	100.00	896.14
std	288460.91	2.58	25.97	184.55

Table 4: Quartile description of columns in dataset

Modifying StoreLocation:

To enhance the effectiveness of our data analysis, it is essential to modify the "StoreLocation" column. Currently, this column contains detailed descriptions of locations, which can complicate the analysis process. By refining this column to contain only the state name, we can streamline our data, making it more uniform and easier to analyze. This modification will facilitate more accurate insights and comparisons across different states, thereby improving the overall quality of our data analysis.

Input:

```
df['State'] = df['State'].replace(state_mapping)

df['State'].head()
```

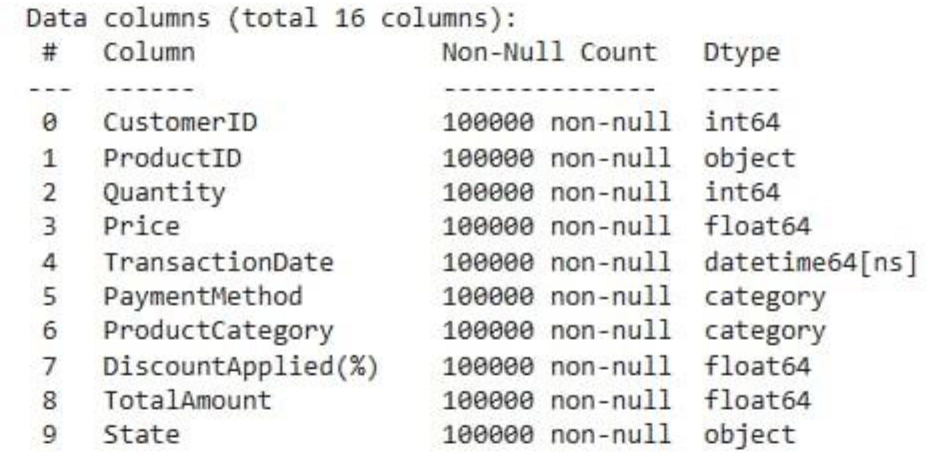
Output:

```
State:AE    3681 AA    3594 AP    3531
```

Redefine the State:

state_mapping = {'AL': 'Alabama', 'AK': 'Alaska', 'AS': 'American Samoa', 'AZ': 'Arizona', 'AR':'Arkansas','CA': 'California', 'CO': 'Colorado', 'CT': 'Connecticut', 'DE': 'Delaware', 'DC': 'District of Columbia', 'FL': 'Florida', 'GA': 'Georgia', 'GU': 'Guam', 'HI': 'Hawaii', 'ID': 'Idaho', 'IL': 'Illinois','IN': 'Indiana', 'IA': 'Iowa', 'KS': 'Kansas', 'KY': 'Kentucky', 'LA': 'Louisiana', 'ME': 'Maine', 'MD': 'Maryland', 'MA': 'Massachusetts','MI': 'Michigan', 'MN': 'Minnesota', 'MS': 'Mississippi', 'MO': 'Missouri', 'MT': 'Montana', 'NE': 'Nebraska', 'NV': 'Nevada', 'NH': 'New Hampshire', 'NJ': 'New Jersey', 'NM': 'New Mexico', 'NY': 'New York', 'NC': 'North Carolina','ND': 'North Dakota', 'MP': 'Northern ariana Islands', 'OH': 'Ohio', 'OK': 'Oklahoma', 'OR': 'Oregon', 'PW': 'Palau','PA': 'Pennsylvania', 'PR': 'Puerto Rico','RI': 'Rhode Island', 'SC': 'South Carolina', 'SD': 'South Dakota', 'TN': 'Tennessee', 'TX': 'Texas','UT': 'Utah', 'VT': 'Vermont', 'VI': 'U.S. Virgin Islands', 'VA': 'Virginia','WA': 'Washington', 'WV': 'West Virginia','WI': 'Wisconsin','WY': 'Wyoming', 'AE': 'Armed Forces Europe', 'AA': 'Armed Forces Americas', 'AP': 'Armed Forces Pacific', 'FM': 'Federated States of Micronesia', 'MH': 'Marshall Islands'}

After that, drop the StoreLocation and create the State:



```
Data columns (total 16 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   CustomerID          100000 non-null  int64
 1   ProductID           100000 non-null  object
 2   Quantity            100000 non-null  int64
 3   Price               100000 non-null  float64
 4   TransactionDate     100000 non-null  datetime64[ns]
 5   PaymentMethod       100000 non-null  category
 6   ProductCategory     100000 non-null  category
 7   DiscountApplied(%)  100000 non-null  float64
 8   TotalAmount         100000 non-null  float64
 9   State               100000 non-null  object
```

Figure 8: Output

2.6. Data model

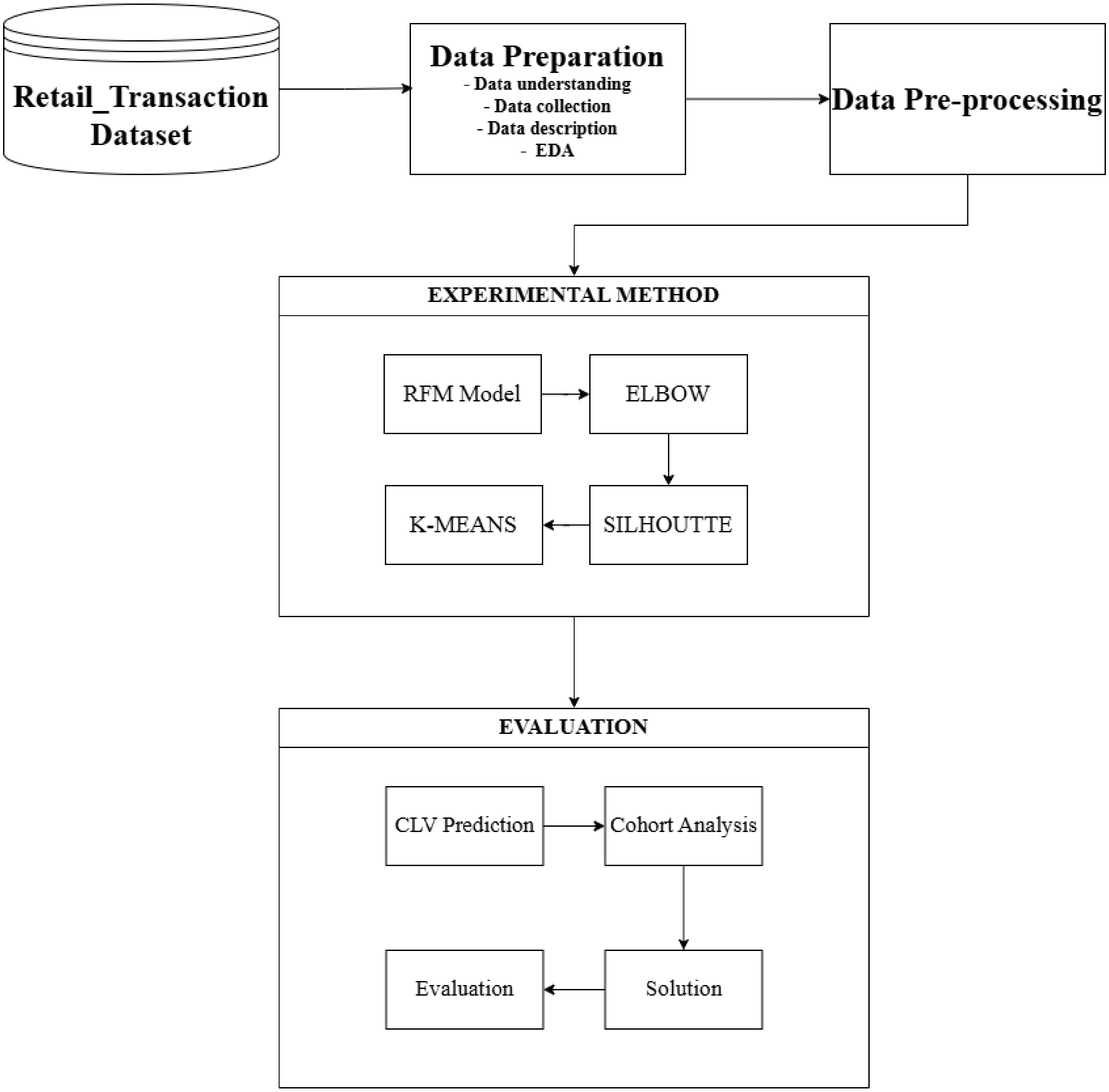


Figure 9: Data model

CHAPTER 3. EXPERIMENTAL METHOD/MODEL AND EVALUATION

Chapter overview

Following data preparation, we performed cluster analysis by computing RFM (Recency, Frequency, Monetary) scores for each client. Next, we used the Elbow and Silhouette methods to identify the best number of clusters (K). The K-means clustering technique was then used to group clients based on their RFM scores. Finally, labels were applied to each cluster to facilitate identification and analysis, resulting in comprehensive findings and recommendations.

3.1. RFM Calculation

RFM (Recency, Frequency, Monetary) is a popular customer segmentation model based on three main factors.

Three factors are used to analyze consumer behavior in RFM: Recency (time since last purchase), Frequency (frequency of purchase) and Monetary (purchasing value). The Pareto principle states that 20% of customers will bring 80% of revenue for a business (Backhaus, J., 1980). The pareto principle Analyze & Kritik, 2(2), 146-171). This theory's goal is to assess a customer's worth to the company and divide them into clusters so that the suitable marketing and customer service tactics could be used. With using the three input indices mentioned above, the K-means algorithm could be used to cluster the data.

For each order of customer, this study would focus on some specific factors: CustomerID (one order just belongs to one customer), TransactionDate and TotalAmount. 3 main factors (Recency - Frequency - Monetary) would be use to identify the value of a customer in RFM model. The TransactionDate attribute is used to calculate the Recency value and Frequency value. Meanwhile, the TotalAmount attribute is used to calculate the Monetary value.

Syntax:

```
presence = dt.datetime(2024, 4, 29)

# Chuyển đổi kiểu dữ liệu của cột 'TransactionDate' sang đối tượng datetime
df['TransactionDate'] = pd.to_datetime(df['TransactionDate'])

# Tính toán các giá trị RFM cho từng khách hàng
grouped_data = df.groupby('CustomerID')

# Tính toán các giá trị RFM
recency = grouped_data['TransactionDate'].apply(lambda x: (presence -
x.max()).days)
frequency = grouped_data['TransactionDate'].count()
monetary_value = grouped_data['TotalAmount'].sum()

# Tạo DataFrame mới chứa các giá trị RFM
rfm = pd.DataFrame({'recency': recency, 'frequency': frequency,
'monetary_value': monetary_value})

# Đảm bảo cột 'recency' có kiểu dữ liệu int
rfm['recency'] = rfm['recency'].astype(int)

# Kiểm tra kết quả
print(rfm.head())
```

CustomerID	Recency	Frequency	Monetary
14	266	1	256.232791
42	345	1	502.656523
49	328	1	21.399047
59	27	1	249.492696
65	315	1	548.006625

Table 5: Final data frame for RFM model

- FRM quartile

	Count	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Recency	95215	178.873544	105.262883	0	87	177	270	365
Frequency	95212	1.050255	0.226357	1	1	1	1	4
Monetary	95215	260.81495	197.112734	7.274825	99.493069	210.700964	178.62633	2002.072661

Table 6: RFM quartile description

The recency metric indicates how recently customers made a purchase. The mean recency is 178.87 days, with a standard deviation of 105.26 days, suggesting a wide distribution of recency values. The first quartile (25%) is 87 days, and the third quartile (75%) is 270 days, indicating that 50% of customers made a purchase between 87 and 270 days ago.

The frequency metric measures how often customers make purchases. The mean frequency is 1.05, with a standard deviation of 0.23, indicating that most customers made only one purchase. The quartiles (25%, 50%, 75%) are all 1, showing that the majority of customers made a single purchase. Only a small number of customers made more than one purchase, with the maximum being four purchases.

The monetary value metric indicates the total amount of money spent by customers. The mean monetary value is $260.81, with a standard deviation of $197.11, indicating variability in spending amounts. The first quartile (25%) is $99.49, and the third quartile (75%) is $378.63, showing that 50% of customers spent between $99.49 and $378.63. The maximum amount spent is $2,002.07.

- RFM Boxplot

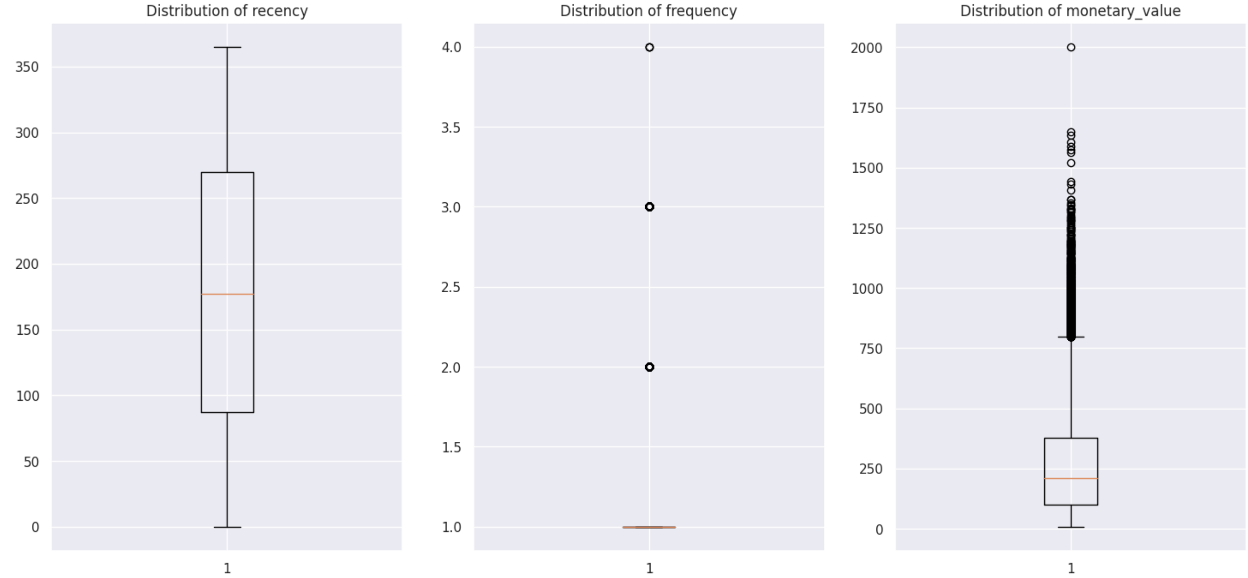


Figure 10: RFM boxplot

The RFM dataset's boxplot analysis provides important insights into customer behavior. With a median recency of around 175 days, most customers make their second purchase between 50 and 250 days following their first. In addition, the data on frequency indicates that most customers make 1 to 2 purchases, with some customers exceeding three purchases. Finally, the majority of customers spend between 0 and 500, according to monetary values, although there are several outliers that suggest some customers pay up to 2000. This result could have some negative effect on clustering results. Therefore, transforming data could be used before implementing

3.2. Data Preprocessing after RFM

There are several strange data formats that we will come across while working with raw data, making analysis and insight discovery occasionally challenging. To restore the data to a normal distribution, one straightforward method is to transform the data. The model learns more effectively and produces more accurate predictions when having the normally distributed data.

There are several methods to transform data:

- Box-Cox transformation

George Box and Sir David Roxbee Cox devised the statistical process known as the "Box-Cox transformation," which converts data that is not normally distributed into one that is.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Figure 11: The Box-Cox transformation formula

- Log transformation

Log transformation is a technique for transforming data based on exponential change. It substitutes a log(x) function for each x. This method's drawback is that negative numbers cannot be calculated using the logarithm function.

$$x = \log(x)$$

- Cube transformation

Cube transformation is a further technique for transforming data based on an exponential. But unlike log transformation, cube root transformation substitutes x^1/3 for x. It is less potent than log transformation. However, it could be applied to both values 0 and <0

	Recency	**Frequency**	**Monetary**
Original data	0.04	4.58	1.04
Log	-1.71	4.31	-0.58
Cube	-0.87	4.37	0.02
Box-Cox	-0.28	4.2	-0.06

Table 7: Data transformation result

After the testing of the normalization methods on the R, F, and M variables. We made the decision to use the box-cox transformation to transfer the currency, frequency, and recency. Following normalization, having the result:

Recency	**Frequency**	**Monetary**
55.084460	0.000000	13.303111
65.413549	0.000000	16.806716
63.268759	0.000000	4.849155
11.421534	0.034335	13.177638
61.601828	0.000000	17.305566

Table 8: RFM after normalization

- Data Scaling

Using measurements of the distances between data points, we will scale the data to fit inside a certain range in order to prepare it for the K-means method.

In this study, using the z-score method (the normalization method) to put the data within a distribution range.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- z is the standardized value.

Scaling is used on R, F, and M because they have significantly different values while these characteristics that appear at comparable scales work better for machine learning algorithms. Following scaling, having the result:

Recency	**Frequency**	**Monetary**
0.831600	-0.225849	0.258973
1.403344	-0.225849	1.186024
1.284624	-0.225849	-1.977939
-1.585268	4.427736	0.225773
1.192354	-0.225849	1.318020

Table 9: RFM scaling result

3.3. The optimal number of clusters by Elbow method

In customer segmentation, the Elbow method plays an important role in determining the optimal number of customer groups. However, it is not possible to evaluate the accuracy of K. Distortion score is the sum of the squares of the distance from each point to its cluster center, indicating how tight the clusters are. Our team implemented the Elbow method with the number of experimental clusters from 2 to 9 on the RFM model and obtained the following results:

Syntax:

```
from yellowbrick.cluster import KElbowVisualizer
k_means = KMeans(random_state=1)
elbow = KElbowVisualizer(k_means, k=(2, 10))
elbow.fit(df_scaled)
elbow.show()
```

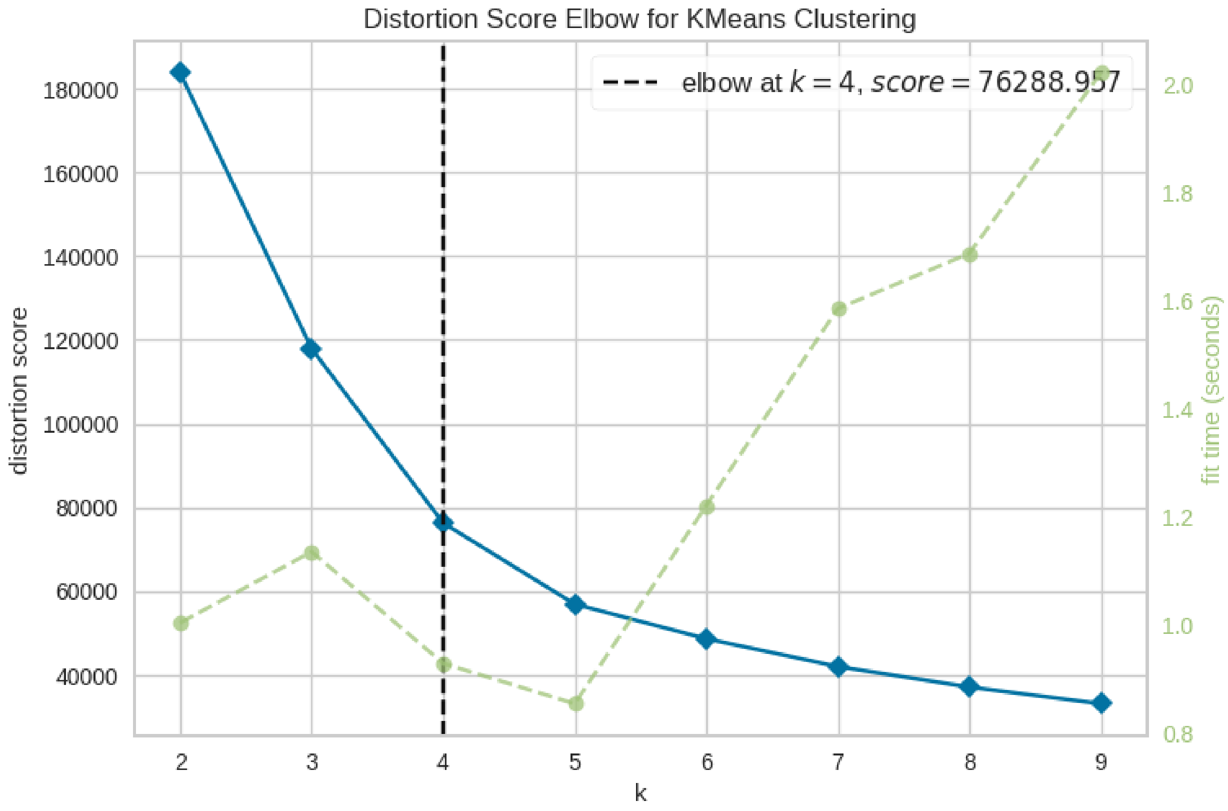


Figure 12: Distortion Score Elbow for K Means Clustering

With the SSE line resembling an elbow, we have a folding point elbow with K = 4 where the distortion score drops significantly before slowing down The elbow point at k = 4 is marked, where the distortion score drops significantly before slowing down, showing that 4 is the optimal number of clusters with an approximate score of 76288.957.

The green line on the chart represents the fit time for each k value, showing that as the number of clusters increases, the fit time will increase but there are fluctuations. The increase in distortion score from K = 2 to K = 4 shows that increasing the number of clusters significantly improves the clustering quality. From K = 4 onwards, cluster division is almost not beneficial. This further contributes to the implication that choosing K = 4 will result in more accurate results.

=> It can be considered that 4 will be the appropriate number of clusters for balancing between model complexity and compactness of the clusters.

3.4. Cluster evaluation using Silhouette:

Our team conducted an in-depth analysis to determine the optimal number of clusters for our client segmentation by calculating the Silhouette index for different values of k. Specifically, we focused on evaluating the case when the number of clusters (nclustersn_{\text{clusters}}nclusters) was set to 4. This assessment was crucial to verify whether the elbow method's suggestion of 4 client groups was indeed the most appropriate choice. The findings of this evaluation are presented below, highlighting an average Silhouette score of approximately 0.43. This score indicates the extent to which each point within a cluster is similar to other points in the same cluster, as well as how distinct the clusters are from one another.

According to the results of the previous Elbow technique analysis, there could be an ideal number of clusters of 2, 4, or 5. We determined the average Silhouette scores for these cluster numbers and made some significant findings in order to improve our clustering strategy. Using only two clusters for our dataset appeared unsuitable and maybe biased, even though the Silhouette score was maximum when the number of clusters (k) was set to 2. In spite of having a high Silhouette score, k=3's customer distribution was uneven among clusters, rendering it unsuitable for efficient segmentation.

Syntax:

```
         Calculate Silhouette Scores
        candidates = [i for i in range(2, 7)]
        sc = []

        for n_clusters in candidates:
            km = KMeans(n_clusters=n_clusters, random_state=0)
            km_labels = km.fit_predict(df_scaled)
            del km
            score = silhouette_score(df_scaled, km_labels)
            sc.append(score)
            print(f'Number  of  Clusters:  {n_clusters},  Silhouette  Score:
{score}')
```

Number of Clusters	2	Silhouette Score	0.6579864476155927
Number of Clusters	3	Silhouette Score	0.6579864476155927
Number of Clusters	4	Silhouette Score	0.37284909354995593
Number of Clusters	5	Silhouette Score	0.39103853290265056
Number of Clusters	6	Silhouette Score	0.39103853290265056

Table 10: Number of clusters and silhouette score

After calculating k = 4, we discovered that this value offered the optimal average Silhouette score for our needs. This shows that four clusters maximize the separation between each data point and the cluster center. Moreover, the utilization of four clusters guarantees that no singular cluster is unduly impacted by extreme values, specifically concerning monetary value. As a result, we discovered that segmenting our information into four clusters provides the best accurate and balanced segmentation.

Syntax:

```
        # Creating a dataframe for the sum of silhouette score
        df_sc  =  pd.DataFrame({'Number  of  Clusters':  candidates,  'Silhouette
Score': sc})
```

```
      # Plotting sum of squared error results
      plt.figure(figsize=(15, 8))
      sns.lineplot(data=df_sc,   x='Number   of   Clusters',   y='Silhouette
Score', color='blue')
      sns.scatterplot(data=df_sc,  x='Number  of  Clusters',  y='Silhouette
Score', marker='X', s=150, color='blue')

      # Customizing the plot
      plt.xticks([2, 3, 4, 5, 6])
      plt.title('Silhouette Plot', fontweight='bold', fontsize=20)
      plt.show()
```

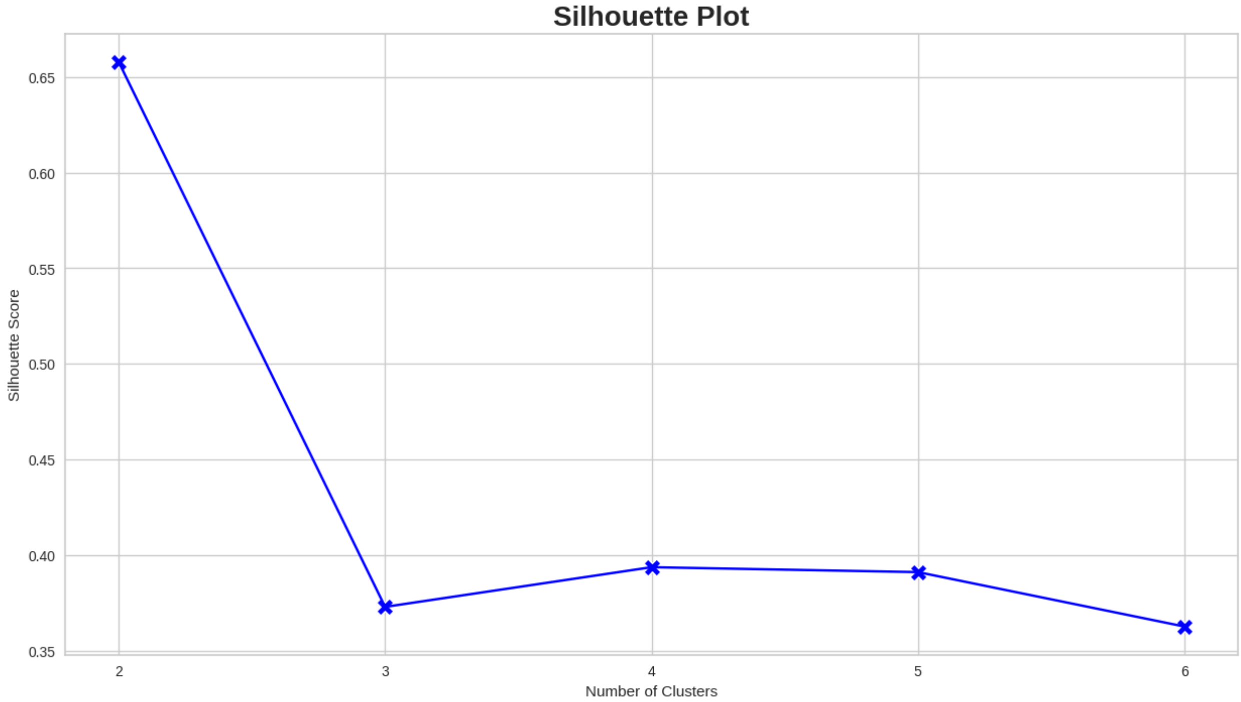


Figure 13: Silhouette Analysis for Optimal Number of Clusters

The average Silhouette score for varying numbers of clusters—from two to six—is displayed in the above graph. This is a helpful method for figuring out how many clusters is the right amount in K-means clustering: the higher the silhouette score, the better the grouping. As the Silhouette score rises, cluster discrimination becomes more lucid and efficient.

Given that there are two clusters on the graph and the Silhouette score is roughly 0.65, it is evident that cluster discrimination is rather robust. The score drastically decreases to about 0.35 when the number of clusters is increased to three, showing a decline in cluster discrimination.

The most noteworthy finding is that the Silhouette score approaches 0.40 when there are three clusters, suggesting that the degree of cluster discrimination is marginally higher than when there are only two clusters. The Silhouette score fluctuates from 0.35 to 0.40 when the number of clusters is raised to four, five, and six, suggesting that there is no discernible improvement in cluster discriminating.

In conclusion, as the Silhouette score is the greatest (0.65) and indicates the best discrimination across clusters, two clusters is the optimal number of clusters in this instance. The quality of grouping declines while the ability to distinguish between clusters remains unchanged whenever there are more than two clusters. The ideal option is to select two clusters based on the Silhouette score.

3.5. Segmentation using K-means

The dataset was segmented with the K-means clustering technique. The Elbow approach demonstrates that the reduction in SSE becomes less significant when the number of clusters k=4. The Silhouette method indicates that the highest average silhouette scores are achieved with k values of 2, 3, and 4. These scores suggest that the cluster cohesion and separation are optimal at these values, with k = 4 being a favorable choice as it balances both cohesion and separation effectively. Consequently, choosing k = 4 is determined to offer an effective and productive clustering conclusion.

Syntax:

```
# 3D scatter plot
fig = plt.figure(figsize=(12, 8))
ax = fig.add_subplot(111, projection='3d')
```

```
      scatter    =    ax.scatter(df_rfm['recency'],    df_rfm['frequency'],
df_rfm['monetary_value'], c=df_rfm['Cluster'], cmap='viridis')
      ax.set_title('3D Scatter Plot of RFM Clusters')
      ax.set_xlabel('Recency')
      ax.set_ylabel('Frequency')
      ax.set_zlabel('Monetary Value')
      legend1 = ax.legend(*scatter.legend_elements(), title="Clusters")
      ax.add_artist(legend1)
      plt.show()
```

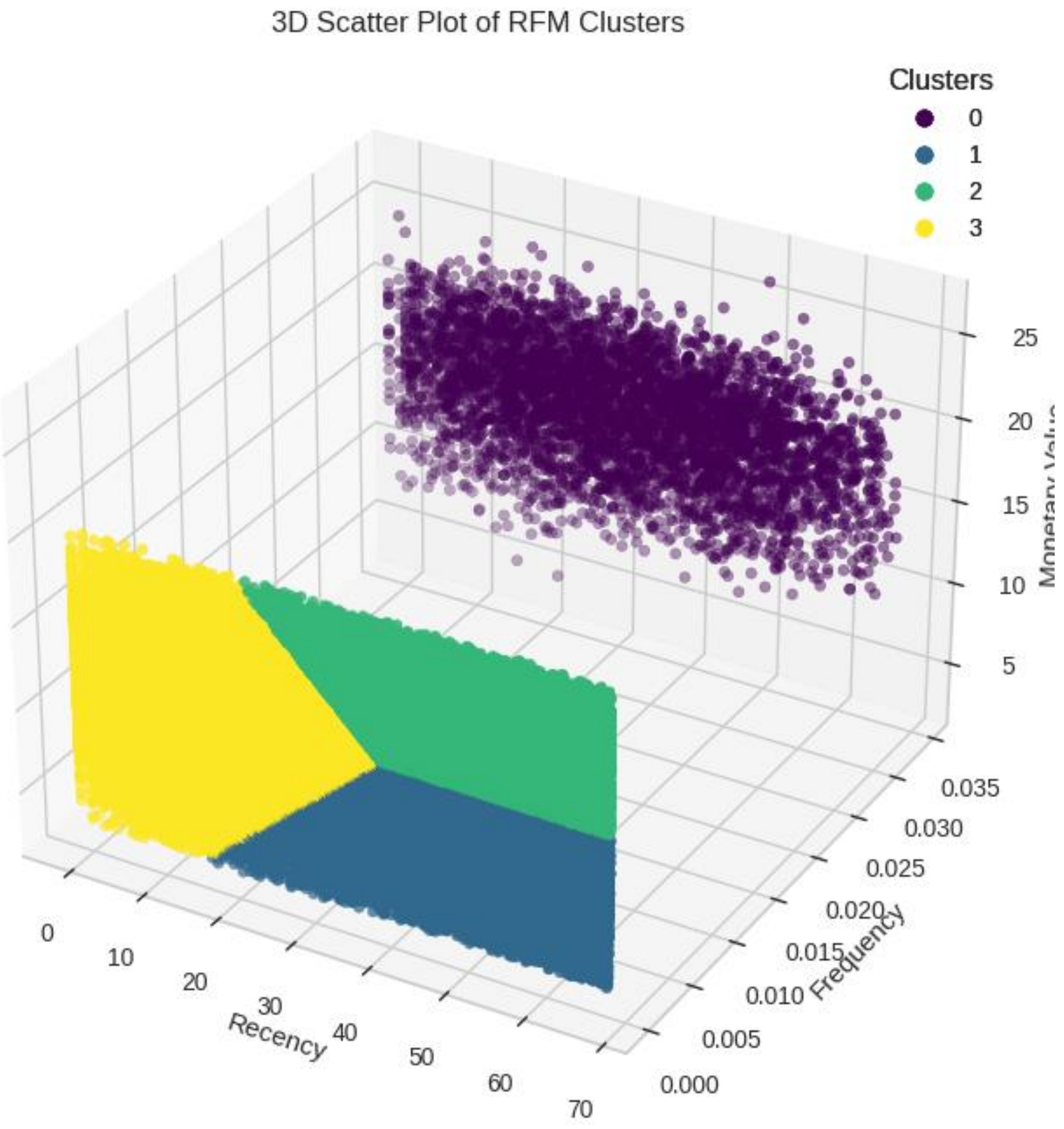


Figure 14: A 3D visualization of the data

3.6. Labeling

3.6.1 Labeling clusters

Cluster	Recency	Frequency	Monetary	Number of Customers	Percent of Customer
0	29.77	0.034	16.20	4621	4.85
1	50.27	0.0	8.70	30133	31.6
2	51.12	0.0	15.33	31350	32.92
3	19.20	0.0	12.21	29111	30.57

Table 11: Mean RFM values of cluster

The customer group is divided into 4 clusters using the K-means algorithm. According to the results, we will describe the mean values including Recency, Frequency, and Monetary to give an overview of the characteristics of each cluster.

- **Regarding Recency value - the customer's last purchase**:

Cluster 3 has an R value of 19.20, meaning the last transaction of this group is more recent compared to the other clusters.

Cluster 0 has an R value of 29.77.

Cluster 1 with an R value of 50.27.

Cluster 2 has an R value of 51.12, indicating that this cluster has the furthest last purchase from the study time.

- **When it comes to Frequency - the number of times a customer makes a purchase in a specific period**:

Cluster 0 has an F value of 0.03, much higher than the other clusters, meaning the frequency of them returning to the store to transact is very high.

The remaining clusters have relatively low F values. Clusters 1,2 and 3 have F values of 0.0.

- **With regard to Monetary value - the total amount that customers have paid for their transactions**:

Cluster 0 is at the top with an M value of 16.20, superior to the rest of the customer groups.

In second place is Cluster 2 with an M value of 15.33, followed by Cluster 3 with an M value of 12.21, and finally Cluster 1 with an M value of 8.70.

After analyzing the results, we decided to segment our customers as follows

Cluster 0: VIP customers. These customers have the most recent purchase, the highest frequency of returning to the store, and the highest amount spent on transactions compared to the other groups. This is the group that brings the greatest value and cannot be lost.

Cluster 2: Potential customers. This group has the highest number of customers surveyed. Similar to cluster 3, the Frequency of this cluster is also 0. What we need to do with them is get them to switch to using our products more often, which would convert them into VIP customers.

Cluster 3: General customers. This group has a lower Frequency than group 0, meaning they are also using products from other businesses and stores. The revenue from this group may be much less than from group 2, but it is the most potential for businesses to utilize and maximize their revenue.

Cluster 1: Infrequent customers. This group has a large number of surveyed customers, second only to group 2, but they spend the lowest amount of money. The last time they transacted was quite far away, and the frequency of purchases is also 0. That means they use the product but not very often. With this group, we need to find a way to entice them to come to the business and use the products more.

3.6.2 Cluster Analysis

- **Cluster 0: VIP Customer**

This Cluster may be examined and given the name "***VIP Customers***" based on the features derived from the indicators, namely:

	Recency	**Frequency**	**Monetary**	**Cluster**
Count	4621	4.621000e+03	4621	4621
Mean	29.77	0.0333468	16.20	0.0
Std	16.23	0.0	3.14	0.0
Min	-1.55	0.0333468	4.95	0.0
25%	16.73	0.0333468	14.02	0.0
50%	28.99	0.0333468	16.42	0.0
75%	42.17	0.0333468	18.49	0.0
Max	67.64	0.0333468	26.43	0.0

Table 12: RFM quartile description of cluster 0

According to the quartile description, this group has 4,621 consumers, which accounts for 4.85% of the company's total customers. This client category has the greatest average purchase value (Monetary), at 16.20, with a standard deviation of 0.0333468. This index has a minimum value of 4.95 and a maximum value of 26.43, showing that this client segment spends significantly and consistently. The average Recency Index for the group is 29.77 days, with a standard deviation of 16.23. The 25%, 50%, and 75% percentiles are 16.73, 28.99, and 42.17 days, showing that this group's recent purchase frequency is rather consistent. Although this group of "VIP Customers" accounts for a small fraction of overall consumers, it contributes significantly to the company's income. To maintain this set of critical consumers, the company might implement new sales practices such as increased incentives or the development of additional product features.

- **Cluster 1: Infrequent customers**

This Cluster may be examined and given the name "***Infrequent customers***" based on the features derived from the indicators, namely:

	Recency	**Frequency**	**Monetary**	**Cluster**
Count	30133	30133	30133	30133
Mean	50.27	0	8.70	1.0
Std	11	0	2.10	0.0
Min	16.21	0	2.88	1.0
25%	41.69	0	7.25	1.0
50%	51.16	0	8.83	1.0
75%	59.51	0	10.48	1.0
Max	67.88	0	12.06	1.0

Table 13: RFM quartile description of cluster 1

Some notable characteristics of this group of customers, compared to the rest, are shown in the table above. This group consists of 31,350 customers with an average Recency value of 51.12 days, indicating that customers made purchases approximately 51 days ago. The standard deviation of this index is 10.57, with a minimum value of 21.52 days and a maximum value of 67.88 days. The Frequency index of this group is 0, indicating that customers in this group do not have a high purchase frequency. The average Monetary value is 15.33. The 25%, 50%, and 75% percentiles are 13.63, 15.18, and 16.88, respectively. This group of customers has a fairly high average purchase value and tends to make purchases not too recently. This suggests that they have the potential to become important customers if properly approached and encouraged to purchase more frequently.

- **Cluster 2: Potential customers**

This Cluster may be examined and given the name "***Potential customers***" based on the features derived from the indicators, namely:

	Recency	**Frequency**	**Monetary**	**Cluster**
Count	31350	31350	31350	31350
Mean	51.12	0	15.33	2.0
Std	10.57	0	2.01	0.0
Min	21.52	0	11.94	2.0
25%	42.80	0	13.63	2.0
50%	52.02	0	15.18	2.0
75%	60.17	0	16.88	2.0
Max	67.88	0	20.38	2.0

Table 14: RFM quartile description of cluster 2

This is the group with the largest number of total customers of the business. This cluster consists of 31,350 customers with a mean Recency of 51.12 days, meaning that the customers made a purchase approximately 51 days ago. The standard deviation of Recency is 10.57, with a minimum of 21.52 days and a maximum of 67.88 days. The frequency of purchase (Frequency) of this group is 0, indicating that they do not make frequent purchases. The mean value of Monetary is 15.33, with a standard deviation of 2.01. The minimum Monetary value is 11.94 and the maximum Monetary value is 20.38. The 25%, 50%, and 75% percentiles of Monetary are 13.63, 15.18, and 16.88, respectively. With this group of customers, businesses should come up with strategies to promote them into general customers.

- **Cluster 3: General customers**

This Cluster may be examined and given the name "***General customers***" based on the features derived from the indicators, namely:

	Recency	**Frequency**	**Monetary**	**Cluster**
Count	29111	29111	29111	29111
Mean	19.20	0	12.21	3.0
Std	9.40	0	3.23	0.0
Min	-1.55	0	2.99	3.0
25%	12.03	0	9.91	3.0
50%	19.91	0	12.22	3.0
75%	26.62	0	14.57	3.0
Max	38.77	0	20.38	3.0

Table 15: RFM quartile description of cluster 3

Some characteristics of this group of customers that are notable compared to the rest are shown in the table above. This cluster includes 29,111 clients with an average Recency value of 19.20 days, indicating that they made a purchase around 19 days ago. This group's purchasing frequency is zero, suggesting that they seldom make purchases. The Monetary index has a mean value of 12.21 and a standard deviation of 3.23. The least monetary value is 2.99, while the maximum is 20.38. The Monetary Index's 25%, 50%, and 75% percentiles are 9.91, 12.22, and 14.57. This customer category has the least value contribution; however, they account for 30.57% of the enterprise's total number of customers.

This chapter continues visualizing the dataset after having cluster customers from K-means, the K-means provides 4 clusters including Vip, General, Normal and Lowest customers. This chapter will visualize more dashboards to get insights to provide some recommendations for our three problems.

4.1. CLV Prediction

Before computing CLV, the project calculates T, which reflects the customer's age in the period of choice units (in this case, weeks). It is determined as the time between the client's initial purchase and the end of the period under consideration.

Syntax:

```
      # Define observation period end
      observation_period_end = pd.Timestamp('2024-04-29')

      # Calculate the age of the customer in the dataset (T)
      df['T']             =             (observation_period_end             -
df.groupby('CustomerID')['TransactionDate'].transform('min')).dt.days

      # Sort the dataset by CustomerID
      sorted_df = df.sort_values(by='CustomerID')

      # Select the relevant columns: R, F, M, and T
      rfm_t_df    =    sorted_df[['CustomerID',    'recency',    'frequency',
'monetary_value', 'T']]

      # Display the R, F, M, and T columns sorted by CustomerID
      rfm_t_df.head()
```

CustomerID	Recency	Frequency	Monetary	T
14	266	1	256.232791	266
42	345	1	502.656523	345

49	328	1	21.399047	328
59	27	2	249.492696	253
59	27	2	249.492696	253

Table 16: R, F, M, and T

- Customer Lifetime Value Prediction:

Syntax:

```
# Identify the top 10 customers by CLV
top_10_customers = merged_df.nlargest(10, 'CLV')

# Predict purchases for the next 10, 30, 60, and 90 days for top 10
customers
time_periods = [10, 30, 60, 90]
predictions = {time: bgf.predict(time, top_10_customers['frequency'],
top_10_customers['recency'], top_10_customers['T']) for time in
time_periods}

# Add the predictions to the top 10 customers DataFrame
for time in time_periods:
    top_10_customers[f'predicted_purchases_{time}_days'] =
predictions[time]

# Display the results
print(top_10_customers[['CustomerID', 'Cluster','recency', 'frequency',
'monetary_value', 'T', 'CLV', 'predicted_purchases_10_days',
'predicted_purchases_30_days', 'predicted_purchases_60_days',
'predicted_purchases_90_days']])
```

CustomerNo	**Frequency**	**Recency**	**T**	**Monetary value**	**Predict purch 10**	**Predict purch 30**	**Predict purch 60**	**Predict purch 90**
32895	3	231	301	2002.072661	0.053929	0.161031	0.319825	0.476431
980072	3	263	276	1636.196111	0.056017	0.167265	0.332200	0.494855
796166	3	53	148	1647.617063	0.053753	0.160493	0.318717	0.474722

903169	2	29	44	1563.851160	0.056115	0.167511	0.332550	0.495176
340516	4	1	205	1604.858454	0.048385	0.144490	0.287009	0.427598
780013	2	191	233	1519.193458	0.054042	0.161338	0.320347	0.477078
823783	3	127	262	14410329882	0.050995	0.152267	0.302410	0.450475
967685	3	155	272	1405.556678	0.252034	0.155371	0.308576	0.459663
805930	2	66	85	1333.753391	0.055724	0.166347	0.330251	0.491769
887487	2	29	82	1352.743062	0.054556	0.162860	0.323328	0.481459

Table 17: Customer Lifetime Value Prediction

The CLV data analysis shows that client 823783 is the most valuable, spending a total of 14,410,329 VND. Although customer 340516 did not spend as much, they're the most devoted customer, having made four purchases, the most recent of which was just one day ago. This means they could become a high-value customer in the future. Meanwhile, client 967685 is expected to have the highest chance of making a purchase within the next 10 days, with a likelihood of 0.252034. This reflects the customer's short-term sales potential.

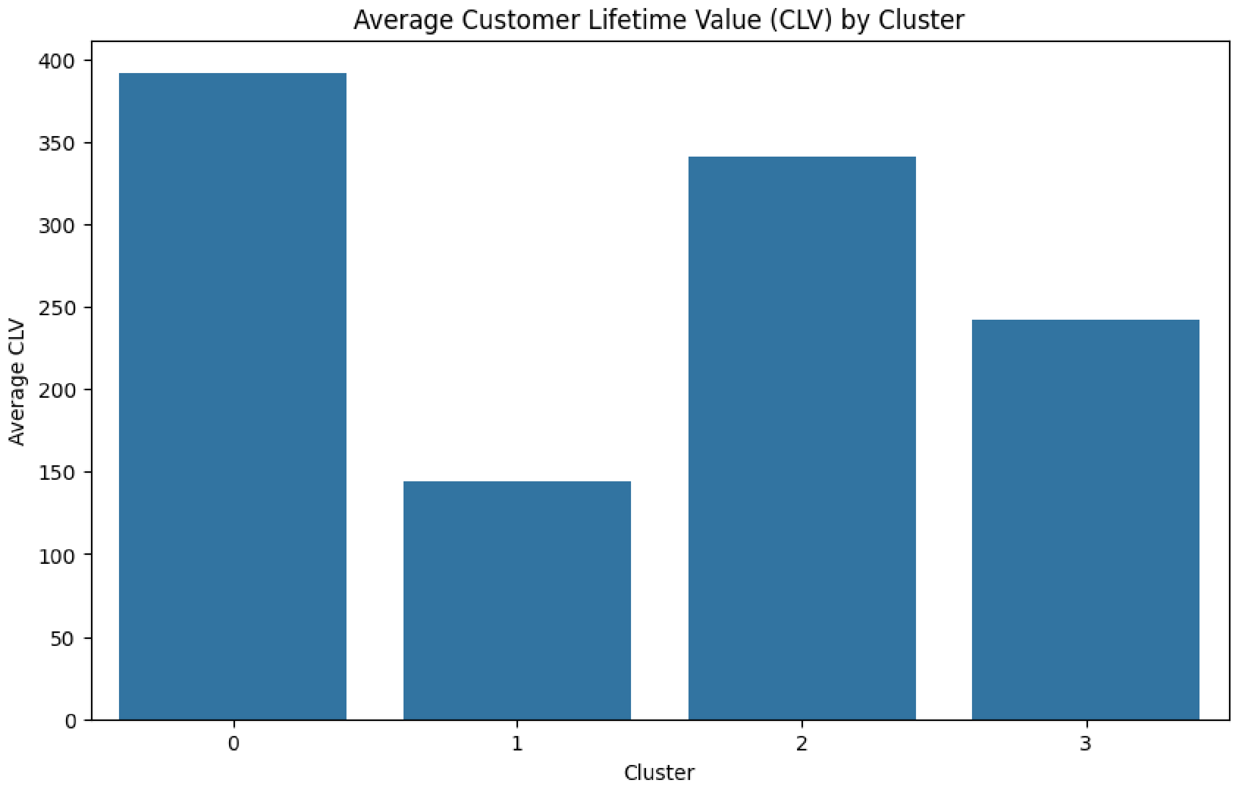
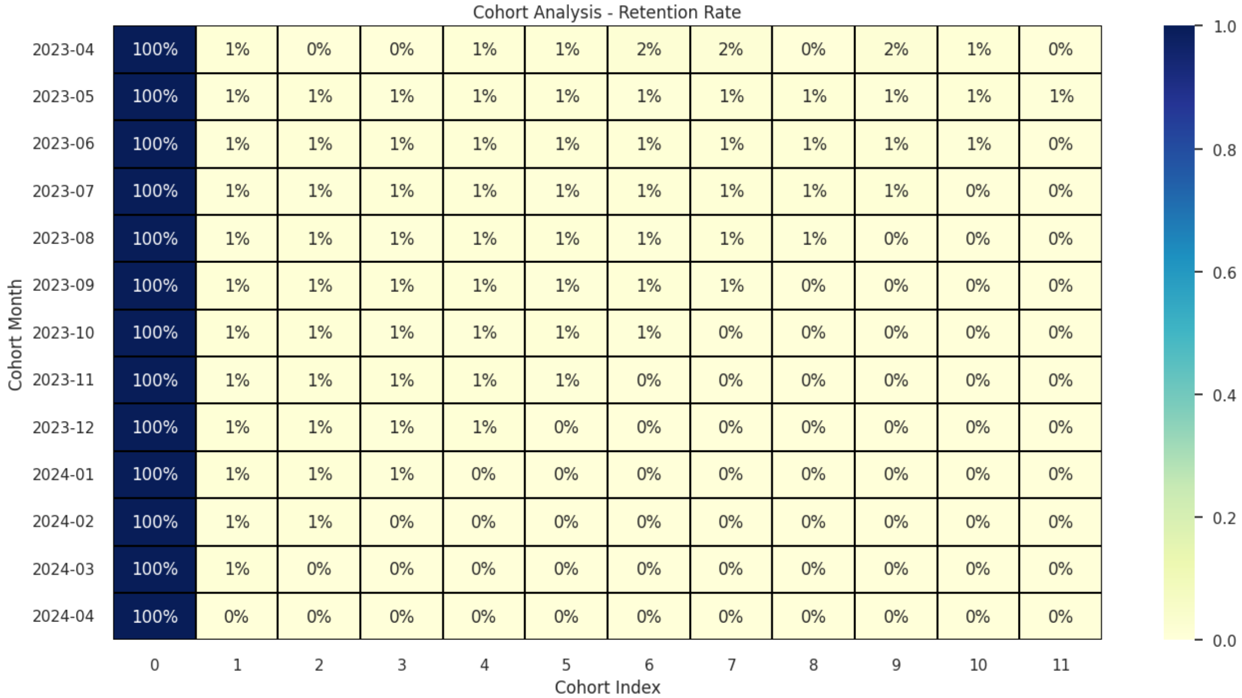


Figure 15: Avagare CLV by clusters

4.2. Cohort Analysis

Cohort analysis heat map visualizes customer retention over time, segmented by the month the customer made their first purchase (CohortMonth). Each row represents a cohort, and each column represents that cohort's retention rate in subsequent months.



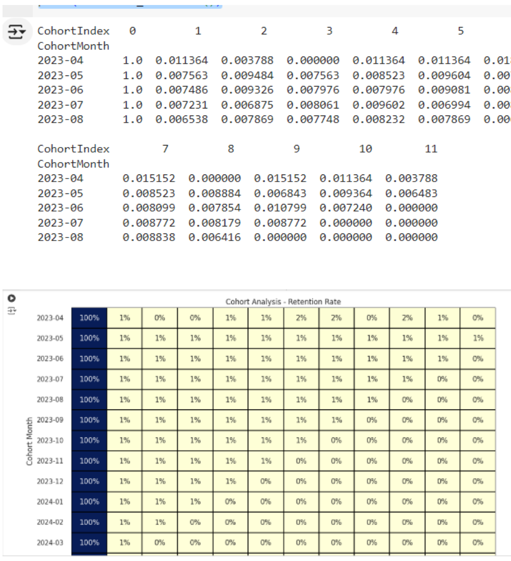
	0	1	2	3	4	5	6	7	8	9	10	11
2023-04	100%	1%	0%	0%	1%	1%	2%	2%	0%	2%	1%	0%
2023-05	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
2023-06	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	0%
2023-07	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	0%	0%
2023-08	100%	1%	1%	1%	1%	1%	1%	1%	1%	0%	0%	0%
2023-09	100%	1%	1%	1%	1%	1%	1%	1%	0%	0%	0%	0%
2023-10	100%	1%	1%	1%	1%	1%	1%	0%	0%	0%	0%	0%
2023-11	100%	1%	1%	1%	1%	1%	0%	0%	0%	0%	0%	0%
2023-12	100%	1%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%
2024-01	100%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
2024-02	100%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2024-03	100%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2024-04	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figure 16: Cohort analysis about retention rate

The retention rates drop significantly after the first month for all cohorts. The majority of retention rates in subsequent months are around 1%, indicating that very few customers make repeat purchases in the following months. Taking the first month in the data set as a benchmark (April 2023), after 1 month of sales, about 1% of customers return to buy. Cohort has 0% retention in CohortIndex 3 and 8, meaning no customers from this cohort made purchases in July 2023 and December 2023. A few cohorts show slightly higher retention rates in specific months (e.g., 2% for 2023-04 cohort in CohortIndex 5 and 6). These could correspond to specific promotions, seasonal effects, or other factors driving repeat purchases.

Code syntax	**Output**

```python
from operator import attrgetter
# Extract cohort month
df['CohortMonth'] =
df['FirstPurchaseDate'].dt.to_period('M')

# Calculate the difference in months between the
transaction date and the cohort start
df['CohortIndex'] = ((df['TransactionYearMonth'] -
df['CohortMonth']).apply(attrgetter('n')))

# Group by CohortMonth and CohortIndex to get the
count of active customers
cohort_data = df.groupby(['CohortMonth',
'CohortIndex'])['CustomerID'].nunique().reset_index()

# Create a pivot table to see the cohorts over time
cohort_pivot =
cohort_data.pivot_table(index='CohortMonth',
columns='CohortIndex', values='CustomerID')

# Fill missing values with 0
cohort_pivot = cohort_pivot.fillna(0)

# Calculate the retention rate
cohort_size = cohort_pivot.iloc[:, 0]

retention_rate = cohort_pivot.divide(cohort_size,
axis=0)

# Display the retention rate DataFrame
print(retention_rate.head())

# Visualize the retention rate
plt.figure(figsize=(16, 8))
sns.heatmap(retention_rate, annot=True, fmt='.0%',
cmap='YlGnBu', linewidths=0.3, linecolor='black')
plt.title('Cohort Analysis - Retention Rate')
plt.xlabel('Cohort Index')
plt.ylabel('Cohort Month')
plt.show()
```



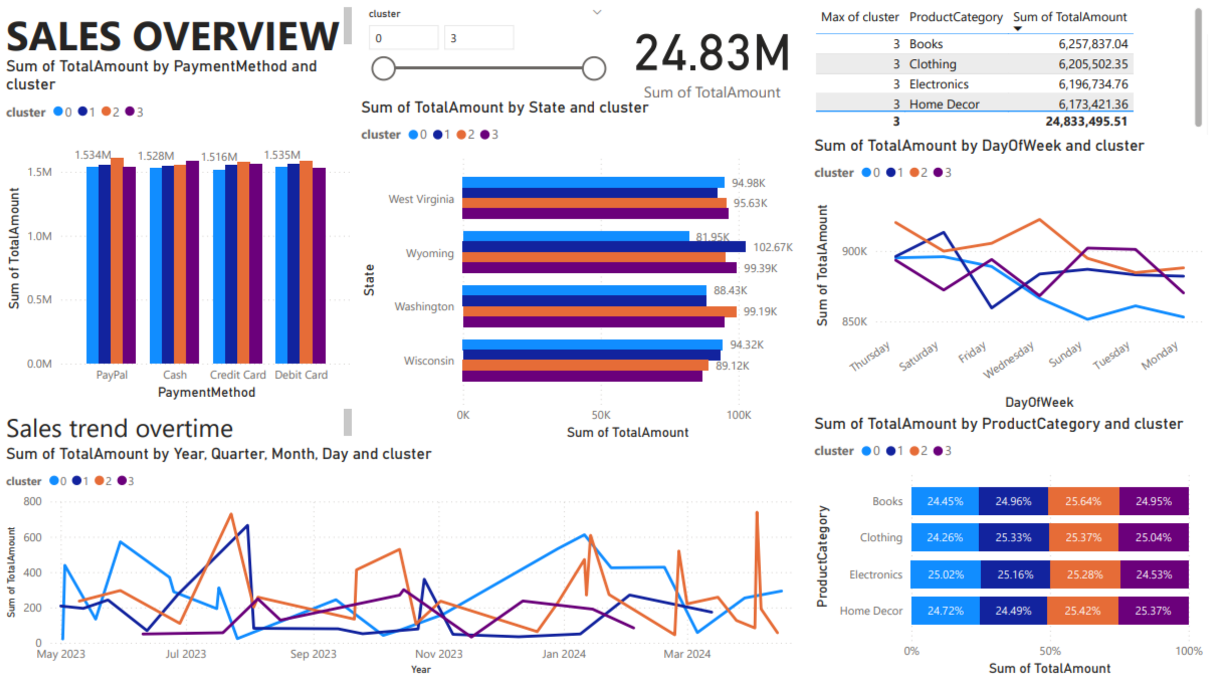
```
CohortIndex  0         1         2         3         4         5
CohortMonth
2023-04     1.0  0.011364  0.003788  0.000000  0.011364  0.011364  0.01
2023-05     1.0  0.007563  0.009484  0.007563  0.008523  0.009604  0.00
2023-06     1.0  0.007486  0.009326  0.007976  0.007976  0.009081  0.00
2023-07     1.0  0.007231  0.006875  0.008061  0.009602  0.006994  0.00
2023-08     1.0  0.006538  0.007869  0.007748  0.008232  0.007869  0.00

CohortIndex        7         8         9        10        11
CohortMonth
2023-04     0.015152  0.000000  0.015152  0.011364  0.003788
2023-05     0.008523  0.008884  0.006843  0.009364  0.006483
2023-06     0.008099  0.007854  0.010799  0.007240  0.000000
2023-07     0.008772  0.008179  0.008772  0.000000  0.000000
2023-08     0.008838  0.006416  0.000000  0.000000  0.000000
```

2023-04	100%	1%	0%	0%	1%	1%	2%	2%	0%	2%	1%	0%
2023-05	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
2023-06	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	0%
2023-07	100%	1%	1%	1%	1%	1%	1%	1%	1%	1%	0%	0%
2023-08	100%	1%	1%	1%	1%	1%	1%	1%	1%	0%	0%	0%
2023-09	100%	1%	1%	1%	1%	1%	1%	1%	0%	0%	0%	0%
2023-10	100%	1%	1%	1%	1%	1%	1%	0%	0%	0%	0%	0%
2023-11	100%	1%	1%	1%	1%	1%	0%	0%	0%	0%	0%	0%
2023-12	100%	1%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%
2024-01	100%	1%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
2024-02	100%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2024-03	100%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 18: Code of Cohort

4.3. Dashboard

4.3.1. Sales dashboard



Max of cluster	ProductCategory	Sum of TotalAmount
3	Books	6,257,837.04
3	Clothing	6,205,502.35
3	Electronics	6,196,734.76
3	Home Decor	6,173,421.36
3		**24,833,495.51**

Figure 17: Sales overview dashboard

This sales dashboard offers a complete perspective of the company's sales performance across several parameters, including clusters, payment methods, states, days of the week, and product categories. The entire sales amount comes to $24.83 million, with a very uniform distribution throughout the clusters, ranging from $6.1 million to $6.4 million. No cluster dominates, demonstrating equal revenue distribution and effective business strategies across all clusters. The balance shows that no cluster is overly outperforming or lagging behind, allowing for a stable revenue stream across the board. There is also no big difference in payment methods. Payment methods such as PayPal, cash, credit card, and debit card all had comparable total sales quantities of around $2 million apiece, showing no substantial difference across clusters.

Geographically, states like Texas, Tennessee, and West Virginia have the greatest overall sales quantities, approaching $100.000 apiece, showcasing them as critical markets. These are areas that the company should continue to exploit and focus on investing to maintain and grow sales. At the same time, find different strategies to improve the remaining areas.

Sales change throughout the week, with Thursday and Friday having the largest sales, while Monday and Tuesday had the lowest, indicating a tendency of higher sales towards the end of the week. Customers tend to shop more on weekends, and understanding this pattern allows the company to strategically plan promotions and special offers to capitalize on these high-traffic days, potentially smoothing out the sales dips seen earlier in the week.

The sales trend over time, from May 2023 to April 2024, shows high volatility, with major peaks in July 2023 and January 2024, which might represent seasonal purchasing trends or special promotional events. Product categories such as Books, Clothing, Electronics, and Home Decor had virtually identical total sales amounts, ranging from $6.1 million to $6.2 million, with Books and Electronics dominating sales. The pie chart also shows a reasonably balanced distribution of product categories, with each category contributing for around 6% of total sales.

This dashboard gives a clear and complete perspective of the company's business performance, making it easier to evaluate and strategize for the future. The dashboard findings highlight the necessity of refining sales methods across several aspects to improve overall business success.

4.3.2. Customer dashboard

- Data storytelling

The "Sum of Quantity by Cluster and Product Category" graphic displays the number of items purchased by customer groups (0, 1, 2, 3) and categories of goods (Books, Clothing, Electronics, Home Decor). Notably, all client groups have comparable expenditures across all product categories, with each category averaging around 31,000. This indicates that there are no substantial disparities in purchasing habits among consumer groups, displaying diversity and balance in customer demands.

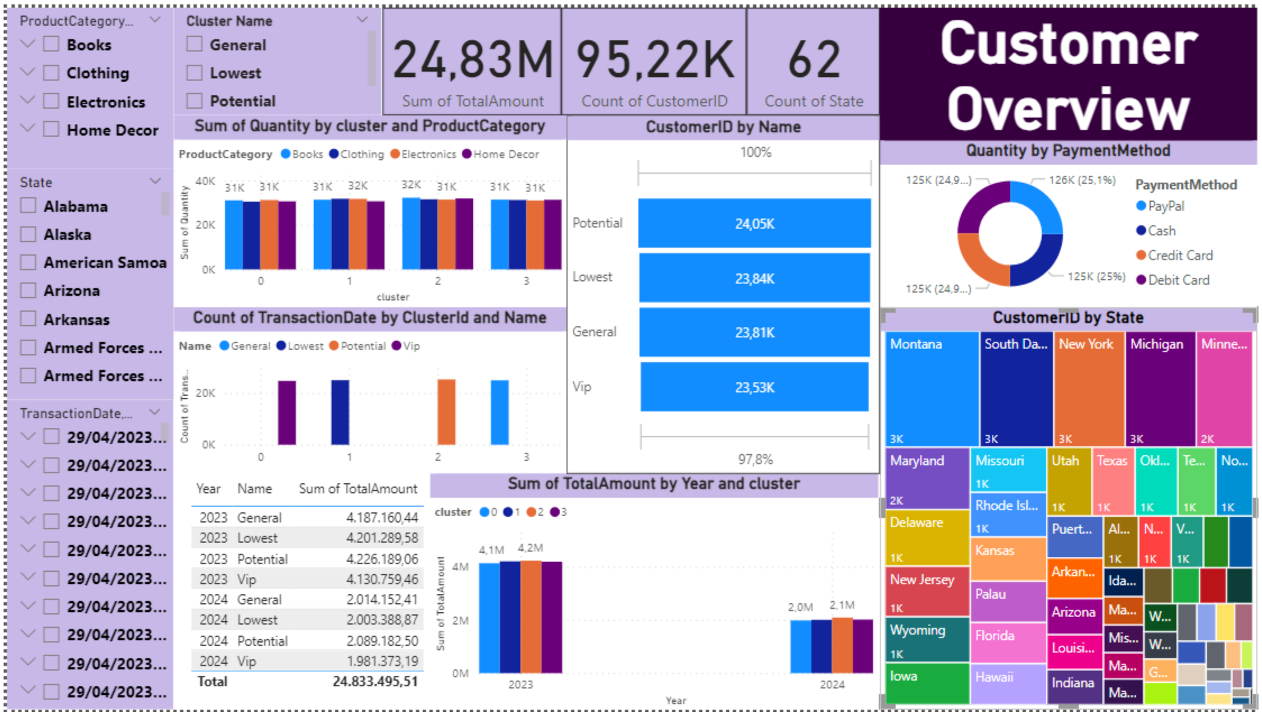
The "Customer ID by Name" study classifies customer data into four categories: potential (24.05K), lowest (23.84K), general (23.81K), and VIP (23.53K). The minor variation in customer numbers between these groups suggests a highly balanced customer base, without any of them dominating.

The "Count of Transaction Date by Cluster ID and Name" graphic displays the number of transactions for each client group and kind (General, Lowest, Potential, VIP). The Potential and VIP groups have more transactions than other groups, showing that their clients are more engaged in their shopping habits.

The "Sum of Total Amount by Year and Cluster" graph compares total sales by year (2023, 2024) and customer segment. 2023 sales are much greater, with particular figures: general (4.18 million), lowest (4.20 million), potential (4.22 million), and VIP (4.15 million). However, sales in 2024 fell significantly: General (2.01 million), Lowest (2.09 million), Potential (2.18 million), and VIP (1.98 million). The reduction in revenue in 2024 requires further investigation to understand the source and solutions.

The "Quantity by Payment Method" pie chart depicts the number of transactions completed using various payment methods, with a relatively even distribution: PayPal accounts for 25.1%, Cash 24.9%, Credit Card 25%, and Debit Card 25%. This shows that buyers have a choice of payment options.

Finally, the "Customer ID by State" treemap displays the number of clients by country, with the most consumers in Montana, South Dakota, New York, Michigan, and Minnesota. This shows that these places have significant market development potential and require targeted marketing tactics to capitalize on opportunities.



Year	Name	Sum of TotalAmount
2023	General	4.187.160,44
2023	Lowest	4.201.289,58
2023	Potential	4.226.189,06
2023	Vip	4.130.759,46
2024	General	2.014.152,41
2024	Lowest	2.003.388,87
2024	Potential	2.089.182,50
2024	Vip	1.981.373,19
Total		24.833.495,51

Figure 18: Customer overview dashboard

- Some insights from dashboard
    - Product Demand:

Customers have different and balanced shopping needs across product categories like books, clothing, electronics, and home decor. Each category sells roughly 31K products across all consumer groups (0, 1, 2, 3). This suggests that there are no substantial disparities in customer shopping preferences, implying that product demand is uniform. To maximize profitability, firms should keep a diversified inventory and make sure that products in these categories are available.

- Customer Activity:

Potential and VIP customers shop more frequently, as shown by a higher number of transactions than other groups. This shows that customers in these groups shop more frequently and at greater prices. Businesses should focus on retaining and increasing the loyalty of these clients with special offers, specialized customer service, and loyalty programs. Retaining and pleasing VIP and potential clients will boost sales and provide a consistent source of cash.

- Sales:

Sales declined significantly between 2023 and 2024. In 2023, total sales for the following client groups were quite high: general (4.18 million), lowest (4.20 million), potential (4.22 million), and VIP (4.15 million). However, in 2024, sales fell to: General (2.01 million), Lowest (2.09 million), Potential (2.18 million), and VIP (1.98 million). The cause of this deterioration must be thoroughly studied. It could be due to changes in market demand, competitive competition, or internal corporate strategy concerns. Businesses must evaluate data more thoroughly and change their business tactics to solve this circumstance, such as improving products, increasing promotion, or adjusting prices to attract customers.

- Payment Method:

Customers can pay using a variety of methods, including PayPal (25.1%), cash (24.9%), credit card (25%), and debit card (25%). This demonstrates that clients are adaptable and trust the payment methods provided by the firm. Businesses should continue to maintain and improve payment services in order to better fulfill the needs of their customers, while also contemplating the addition of new payment methods or more secure and easy features.

- Geographic Distribution:

The states having the most clients are Montana, South Dakota, New York, Michigan, and Minnesota. These are locations with a high potential for market growth. To capitalize on prospects, businesses in these states should prioritize marketing techniques and customer service. Local advertising, event planning, and collaboration with local partners are all examples of marketing actions that can improve brand presence. Simultaneously, enhancing customer service in areas such as speedy delivery, convenient return policies, and customer care would assist to increase customer satisfaction and engagement.

4.4. Implementation

4.4.1. Identifying High-Value Customers and reducing customer churn

In these problems, our project focuses on enhancing the relationship between 2 clusters is cluster 3 and cluster 0 and business.

- With high-value customers - VIP customer (cluster 0):

Solution proposing for cluster 0 (***VIP Customer***): This is a customer group (consisting of 4621 people, accounting for 4.85% of total customers) with R = 29.77, F = 0.034, and M =16.20, these figures are significantly higher than the other 3 customer groups. It can be said that this is a high-value customer group, as this is a segment of customers who purchase frequently, have a high monetary value, and have made recent purchases.

Due to the highest profit potential that this customer group brings, according to the Pareto Principle, it can be believed that 80% of the company's profit comes from 20% of these customers. This is the customer segment that generates the highest revenue, so the company must always create a close relationship with this group.

For this VIP customer segment, the company needs to focus on strategies to enhance the customer experience by personalizing each customer's journey. Additionally, the company should also organize special promotions to express gratitude to the loyal customer segment, nurturing and encouraging them to continue shopping or using the company's services. Special programs organized exclusively for this VIP customer segment aim to make them feel satisfied and valued. Our project has a few detailed suggestions as follows:

Businesses can send emails expressing gratitude to customers for their continued support and use of the company's products. This action helps customers feel valued, cared for, and appreciated. These thank-you emails can be accompanied by small gifts that might interest the customers. This gesture serves the dual purpose of showing appreciation through gifts and providing an opportunity for customers to experience other products. These emails can be sent on occasions such as customer birthdays, company events, holidays, ….

One of the most effective customer retention programs is when companies apply discount policies for specific customer groups. Based on the revenue generated by the VIP customer group, businesses should create VIP discount codes exclusively for this group. These discount codes can be applied to the entire store or to specific product categories that VIP customers often purchase. This action creates customer satisfaction towards the company.

In addition to VIP vouchers, the company can send surprise gifts or new product samples in VIP customers' orders.

As a business, it's important to be flexible in the product list, always introducing new products to keep up with modern trends. Businesses can create a list of products that

VIP customers might be interested in and allow them to purchase before the new products are launched or during special promotions.

The company should also have loyalty programs, such as earning points through the number of purchases or a percentage of the order value to redeem gifts and upgrade membership levels. This allows customers to receive special offers on their birthdays or anniversaries.

The company can also offer special benefits exclusively for VIP customers, as a way to show appreciation and acknowledge their valuable support. First and foremost, one of the prominent benefits is that customers will be given priority access to special events such as new product launches, workshops, seminars, and customer appreciation parties. These events not only provide opportunities to experience new products and services but also create a platform for customers to meet and interact with industry experts, thereby enhancing their knowledge and expanding their network.

Furthermore, the company should pay special attention to the private shopping experience of VIP customers. Organizing private shopping sessions at the store or online with professional advice from the sales staff will provide customers with the best suggestions and support, ensuring that every purchase decision brings maximum satisfaction. The dedication and professionalism in this consulting service will create a luxurious and comfortable shopping experience for the buyer.

Additionally, the business should also expand its network of cooperation with other reputable brands to create special offers exclusively for VIP customers when shopping at their partners. This not only brings added value to customers but also expands their choices and diversifies their shopping experience. Collaborating with partner brands helps diversify the products and services offered, meeting all customer needs and preferences.

These benefits are not only intended to show appreciation for the support of VIP customers but also aim to provide unique and luxurious shopping experiences, helping customers feel valued and cared for by the business.

Creating a loyal customer community is the pivotal way for businesses to keep strong relationships and increase sales. To reach this goal, businesses can do different things that help customers meet and share with each other. One of the best ways is to create an online place or a special group where loyal customers can easily talk to each other, share their experiences with the product, exchange ideas, and learn from one another. Here, they can get special deals, the newest news about products and services, and join fun contests and mini-games. Also, making private groups on social media just for VIP customers is a great way for businesses to connect directly with their most important customers. In these groups, businesses can tell about special deals, new products, and plan fun activities and mini-games. This helps VIP customers to meet and share with one another. Businesses can also plan in-person events, like meetups, seminars, workshops, or special trips just for their VIP customers, not just online activities. These activities let customers try out products and services directly. They also give them a chance to meet and connect with each other and with company staff, helping to create strong and lasting relationships. By creating a loyal group of customers, businesses can connect with people who care about them. This helps them get useful feedback to make their products and services better and improves how customers see their brand.

Email newsletters are a powerful tool for organizations, serving several important functions. By regularly sending out newsletters to customers, companies can strengthen customer connections and demonstrate their commitment to keeping customers informed and engaged. Additionally, newsletters improve brand awareness by consistently reaffirming the company's presence in the minds of consumers. This continuous interaction keeps the brand top-of-mind and fosters a sense of familiarity and trust. Using specialized software like Mailchimp, Constant Contact, or CRM systems with integrated email marketing features, companies should send newsletters about new product launches, holidays, or special occasions, including discount vouchers or special offers to customers whose email addresses or phone numbers the company already has. To avoid boring customers, companies should send newsletters every two weeks and at reasonable times like 12 noon or after 8 pm so that customers can easily see the emails. The company must always carry out campaigns to increase brand recognition among customers.

- With lowest customers (cluster 1):

Solution proposing for cluster 1 (***Lowest customers***): To improve consumer retention and create ongoing relationships with high-churn consumers, the implementation team provides a number of unique and effective techniques for boosting customer experience and brand value.

In the beginning, the organization must perform surveys and solicit feedback from clients with quick online questionnaires, direct contacts, or emails in order to determine the reasons for their leave. This will allow the organization to discover particular challenges that clients are experiencing and adapt their strategy accordingly. Additionally, using data analysis tools such as Google Analytics or CRM systems will assist them in identifying behavioral patterns of clients with high turnover. This enables companies to adjust their strategy and customize the shopping experience based on clients' purchasing history and browsing habits, resulting in adapted product suggestions and offers by email and smartphone alerts.

The corporation could also develop a stratified loyalty program with membership grades such as Silver, Gold, and Platinum, allowing customers to collect points for each purchase and return them for discounts or free items. At the same time, customer service needs to be improved, including timely and effective replies to all customer requests and concerns, as well as post-purchase follow-ups to assure satisfaction and give required support. Furthermore, providing regular and event-based incentives, such as discounts for birthdays or special holidays, would encourage customers to return for more frequent purchases.

The company could also set up a system for referrals, rewarding clients with extra points or discount coupons for each successful recommendation. Advanced technologies such as chatbots and virtual assistants will be used to give 24/7 customer service, product recommendations, and answers to commonly asked questions. The organization's mobile app should also include personalized features, special discounts, and order monitoring to ensure clients receive timely promotional alerts.

To effectively serve clients, the company should diversify its sales outlets. Developing online sales channels such as websites, mobile applications, and social media platforms in addition to traditional stores will allow people to shop at any time and from any location. Additionally, providing speedy delivery services, free shipping, and doorstep delivery would maximize client convenience. To provide clients with peace of mind while shopping, a flexible and easy return policy should be implemented. The organization can also develop a product trial program, which allows customers to try things before deciding on a purchase, lowering risks and building trust.

Furthermore, the team presents creative ideas such as a Virtual Reality Product Experience program, which allows clients to discover goods in a virtual environment prior to determining a purchase choice, resulting in increased excitement and greater knowledge of the product. Customers will be able to adjust products based on their unique preferences. Shopping challenges will generate interest and motivate customers to visit the store more frequently. Organizing livestream sessions with product usage instructions and direct guidance from professionals will make customers feel more supported and trustworthy in the company's offerings.

Finally, in order to provide clients with a pleasurable and engaging shopping experience, the company must offer events and activities such as workshops, beauty consultations, and new product demonstrations. Organizing livestream sessions featuring product usage instructions and direct guidance from professionals will also help clients feel supported and confident in the company's offerings. All of these techniques are designed to preserve positive client connections, increase contentment, and drive repeat purchases.

4.4.2. Personalizing Customer Experiences

In this section, our team will focus on two main customer segments are general customers and potential customers to enhance personalizing customer experiences:

- With general customers (cluster 3):

Solution proposing for cluster 3 (***General customers***): The data indicates that the level is in the average to good range, the transaction time is somewhat lengthy, and there aren't many buyers, according to the general customer group spending data report. The issue that the group is trying to solve is how to improve the overall customer experience without raising customer happiness or loyalty levels.

Thus, the team suggests tailoring the customer experience, raising the standard of customer care, developing loyalty programs, strengthening brand awareness, and gathering and addressing client feedback for this particular set of clients. Businesses must also provide transaction rewards to these clients in order to urge them to make purchases right away and enhance their purchasing power.

- *In terms of personalizing the customer experience:* businesses can employ customer relationship management (CRM) techniques to store and analyze customer data in order to recommend relevant products or services for this customer group. They can also use data gathered from prior interactions to gain a deeper understanding of customer preferences and behaviors. For instance: product recommendations based on past purchases to assist clients access a range of products, or birthday wishes with exclusive discount codes for customers via personalized email marketing. Customers will feel valued and cared for as a result of this sense of customization and attention.

- *Enhance the quality of customer service:* Businesses should concentrate on cultivating connections with elite groups and provide staff with additional training in communication and problem-solving techniques to guarantee that every customer group receives the best support possible, which includes attentive listening, professional handling of unfavorable feedback, and prompt and efficient resolution of issues. Ensure that clients feel they are receiving the individual attention they are due. Improve your in-person and online customer service channels at the same time. Use AI chatbots, for instance, to offer round-the-clock customer service, quickly respond to often asked queries, and establish direct lines of communication, such as hotlines or live chats, for more complicated problems.

- *Create loyalty programs:* Businesses can promote and integrate promotions and marketing activities through multiple channels such as social media, fanpages, emails in the form of mini games, gifts, etc. to encourage customers to interact with the business. At the same time, businesses can develop additional points and rewards programs to encourage customers to return to shop. This program can include accumulating points for each purchase, then redeeming points for special offers or gifts. To make these activities more effective, businesses can conduct multi-platform marketing, run ads to reach the target audience more effectively and accurately, and create incentives specifically for loyal customers, such as special discounts, priority participation in events or new product launches. This not only retains existing customers but also motivates new customers to become loyal customers.

- With potential customers (cluster 2):

Solution proposing for cluster 2 (***potential customers***): This customer segment is small in number, has relatively recent transactions (recency), but not high frequency. Therefore, for this group, our company needs to satisfy them with their initial transactions to retain them and encourage them to make more purchases with larger basket values.

To enhance the experience of this customer segment, we suggest that we should have a well-organized and highly personalized care process, which will be crucial in building an initial relationship with them. These processes are similar to the customer care recommendations proposed for cluster VIP customers. We can nurture and encourage these customers to come back again and again by offering a loyalty program with highly personalized and differentiated values. The main goal of businesses with this customer segment is to keep them happy and coming back for more purchases. Our team suggests some strategies for this segment, as follows:

- *Personalized First Purchase Experience:* Businesses should send thank you emails and solicit feedback from customers on their initial purchase experience to make them feel unique and valued. At the same time, give them a welcome voucher for their next purchase. This not only motivates consumers to return to shop, but it also makes them feel cared for and valued by the company. As a result, consumers feel unique and cared for, and they are more likely to purchase again.

- *Utilizing Transactional Rewards:* Offer quick transactional benefits with each purchase, such as discount certificates and bonus gifts. Organize promotional programs and minigames on social media to enhance client involvement. These activities will draw customers' attention and encourage them to purchase more regularly. Offer enticing incentives to boost potential consumers' purchase frequency and retention.

- *Personal Shopping Advisor and Customer Service:* Contact potential clients to establish a relationship from browsing to purchase, serving as a personal shopping adviser and offering advice and assistance as needed. This improves client satisfaction with the company's offerings and provides a better shopping experience for them. To improve client happiness and provide a better purchasing experience.

- *Developing long-term relationships:* Interact with clients on a regular basis via social media and email, offering excellent customer service in order to foster trust and long-term connections. This helps to increase consumer engagement and loyalty. The objective is to boost customer engagement and loyalty while also developing long-term connections with them in order to convert them into VIP customers in the future.

In conclusion, by employing these tactics, businesses may effectively engage potential consumers and persuade them to become VIP clients. Personalized first-purchase experiences and rapid transactional benefits provide a pleasant and gratifying first engagement. A well-structured loyalty program, when paired with spending-related rewards and referral programs, may encourage repeat purchases and boost cart values. Businesses may enhance the customer experience by developing long-term connections and providing personalized shopping help, ensuring that potential consumers feel appreciated and driven to continue buying. Businesses may use these thorough efforts to convert potential consumers into loyal, high-value clients, resulting in long-term revenue development and market success.

CHAPTER 5. CONCLUSION & FURTHER ORIENTATION

Nowadays, customer segmentation is a vital tool for organizations, particularly those in the retail sector. By classifying consumers according to shared requirements, preferences, and actions, businesses can gain deeper insights into their intended market. These are useful tools that businesses can use to learn more about their customers. In order to improve the segmentation findings, we employed a variety of techniques, including the Elbow and Silhouette methods, in conjunction with the RFM model and the K-means clustering algorithm in this work. This technique finds outliers in the data set and guarantees the precision and dependability of the clusters that have been found. Furthermore, exploratory data analysis (EDA) was carried out in order to obtain a deeper understanding of the consumer groups and business landscape. This combination strategy has a lot of promise, particularly for the retail sector. It gives companies a more complete picture of their clientele, enabling them to create revenue optimization plans that work for all of their clientele. Additionally, this approach can handle big historical data sets from companies, enabling them to get the most out of the information they have gathered over time.

In order to carry out this research, we want to put in place a number of procedures to assess the outcomes and create marketing plans that are specific to each target market in light of our findings. To be more precise, we can start segment-specific campaigns with promos, monitor development and outcomes, and assess how effective these tactics were both before and after the campaigns. These projects' outcomes will be recorded in our upcoming research. We hope that the outcomes of our research will be useful and applicable, enabling companies to enhance client lifetime value, boost revenue, and improve operations. Furthermore, our goal is to improve the model's performance across various business scenarios and situations. There is a lot of promise in this integrated approach, particularly for the retail industry. It enables companies to get a thorough grasp of their clientele, which makes it easier to create revenue optimization plans that are customized for various clientele segments. Additionally, the approach shows a great deal of promise for handling massive company historical data sets.

Following the completion of this experiment, we discovered many interesting avenues for further research:

Using a more diverse dataset would widen the scope of the research and provide more real-world experience. This would allow for the examination of more intricate correlations between variables, as well as testing the model's performance on other forms of data.

Apply Alternative Clustering Methods: For this project, we used a specific grouping method. Experimenting with alternative algorithms, such as DBSCAN or hierarchical clustering, could produce fascinating findings and make it easier to compare the performance of different methods on the same dataset.

Analyze the Framework With More Metrics: Due to time limitations, we were unable to employ a large number of model evaluation metrics. In the future, adding metrics like the Calinski-Harabasz Index or Davies-Bouldin Index could give a more complete assessment of the model's performance and clustering capabilities.

Combining clustering with other machine learning approaches, such as classification, regression, or principal component analysis, may result in the development of more robust predictive models and deeper insights into the data.

Apply to Other sectors: The clustering and data analysis techniques employed in this study can be applied in a variety of sectors, including healthcare, finance, marketing, and social sciences. Expanding the use of these strategies would aid in the resolution of real-world issues and offer value to society.

REFERENCES

AppsFlyer. (2024, March 28). Churn rate. AppsFlyer. https://www.appsflyer.com/glossary/churn-rate/

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications, 105*(9).

B. Yusuf Bakhtiar, A. Bima Murti Wijaya, and H. Dwi Cahyono, "PENGEMBANGAN SISTEM ANALISIS AKADEMIS MENGGUNAKAN OLAP DAN DATA CLUSTERING STUDI KASUS : AKADEMIK UNIVERSITAS SEBELAS MARET SURAKARTA," J. Teknol. Inf. ITSmart, vol. 4, no. 1, p. 01, Sep. 2016.

Chen, Y., Zhang, G., Hu, D., & Wang, S. (2006, June). Customer segmentation in customer relationship management based on data mining. In *International Conference on Programming Languages for Manufacturing* (pp. 288-293). Boston, MA: Springer US.

Durmaz, Y., & Diyarbakırlıoğlu, I. (2011). A Theoritical Approach to the Strength of Motivation in Customer Behavior. *Global Journal of Human Social Science, 11*(10), 36-42. https://tinyurl.com/47vrwcez

Huang SC, Chang EC, Wu HH (2009). A case study of applying data mining techniques in an outfitter's customer value analysis. Expert Syst.Appl., 36: 5909-5915.

J.R. Bult and T. Wansbeek, Optimal selection for direct mail, Marketing Science 14 (1995) 378-395.https://doi.org/10.1287/mksc.14.4.378

Kaul, D. (2017). Customer relationship management (CRM), customer satisfaction and customer lifetime value (CLV) in retail. *Review of professional management, 15*(2).

Miglautsch JR (2000). Thoughts on RFM scoring. J. *Database Mark., 8*(1): *67-72*.https://academicjournals.org/article/article1380555001_Wei%20et%20al.pdf

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of Sth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20,* 53-65.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *ICIC Express Letters, Part B: Applications, 9*(2), 91-96.

Sun, J. (2008). Clustering Algorithms Research. *Ruanjian Xuebao, 19*(1), 48–61. https://doi.org/10.3724/sp.j.1001.2008.00048

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018, October). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study. In

2018 IEEE 15th International Conference on e-Business Engineering (ICEBE) (pp. 119-126). IEEE.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), 7243. https://doi.org/10.3390/su14127243

Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. *African journal of business management*, *4*(19), 4199. https://academicjournals.org/article/article1380555001_Wei%20et%20al.pdf

Wang, C. H. (2010). Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. *Expert systems with applications*, *37*(12), 8395-8400.