Project Report

On

RFM Based Customer Segmentation using K-Means

Submitted

In partial fulfilment

For the award of the Degree of

PG-Diploma in Big Data Analytics

(C-DAC, ACTS (Pune))

Guided By:

Mr. Prakash Sinha

Submitted By:

Devashish Revadkar (240340125021)

Shruti Bharat (240340125046)

Apeksha Wankhede (240340125012)

Shreeyash Kumthekar (240340125044)

Rahul Gupta (240340125035)

Centre for Development of Advanced Computing

(C-DAC), ACTS (Pune- 411008)

ACKNOWLEDGEMENT

This is to acknowledge our indebtedness to our Project Guide, **Mr. Prakash Sinha**, C-DAC ACTS, Pune for her constant guidance and helpful suggestion for preparing this project **RFM Based Customer Segmentation Using K-Means**. We express our deep gratitude towards her for inspiration, personal involvement, constructive criticism that she provided us along with technical guidance during this project.

We take this opportunity to thank Head of the department **Mr. Gaur Sunder** for providing us such a great infrastructure and environment for our overall development.

We express sincere thanks to **Mrs. Namrata Ailawar (Process Owner)** for their kind cooperation and extendible support towards the completion of our project.

It is our great pleasure in expressing sincere and deep gratitude towards **Mrs. Risha P R (Program Head)** and **Ms. Pratiksha Gacche (Course Coordinator, PG-DBDA)** for their valuable guidance and constant support throughout this work and help to pursue additional studies.

Also, our warm thanks to **C-DAC ACTS Pune**, which provided us this opportunity to carry out, this prestigious Project and enhance our learning in various technical fields.

Devashish Revadkar (240340125021)
Shruti Bharat (240340125046)
Apeksha Wankhede (240340125012)
Shreeyash Kumthekar (240340125044)
Rahul Gupta (240340125035)

ABSTRACT

In today's competitive business landscape, understanding customer behavior is pivotal for driving success and growth. This project explores the application of data-driven techniques, specifically customer segmentation and RFM (Recency, Frequency, Monetary) analysis, to gain deeper insights into customer preferences and purchasing patterns. The project begins by gathering and preprocessing transaction data, which is then subjected to RFM analysis to evaluate the recency, frequency, and monetary value of customer purchases.

Through the application of K-Means clustering, distinct customer segments are identified, each representing a unique combination of purchasing behaviors. These segments are subsequently analyzed to extract actionable insights, enabling businesses to tailor their marketing efforts more effectively. The insights derived from this analysis have the potential to enhance customer retention, optimize marketing campaigns, and ultimately increase profitability. The results of this project demonstrate the effectiveness of RFM analysis and customer segmentation as powerful tools for understanding customer behavior and driving strategic decision-making in marketing.

TABLE OF CONTENTS

Chapter 1
INTRODUCTION

1.1 Introduction

In today's highly competitive business landscape, understanding and catering to customers' needs and preferences are crucial for a company's success. Customer segmentation is a powerful marketing technique that helps companies gain insights into their customers and create tailored marketing strategies to meet their specific needs. By dividing customers into smaller groups based on shared characteristics such as demographics, psychographics, or behavior, businesses can identify unique patterns and behaviors within their customer base.

Customer segmentation has become increasingly important in recent years, as advances in technology and data analytics have made it easier for companies to collect and analyze customer data. With the rise of e commerce and social media, businesses can track customers' browsing and purchasing behavior, as well as their likes, interests, and social connections, to gain a better understanding of their preferences and needs.

Customer Segmentation is way of organization of customers with respect to the various features. In recent years there has been a huge boom in opposition between companies to stay in the field. The income of the organization may be stepped forward through a patron segmentation model. According to the Pareto principle (Srivastava, 2016), 20% of the clients contribute greater to the sales of the organization than the relaxation Customer segmentation is the exercise of dividing an organization’s clients into agencies that mirror similarity amongst clients in every group. The intention of segmenting clients is to determine how to narrate to clients in every section that allows you to maximize the fee of every patron to the business. Customer segmentation can permit entrepreneurs to cope with every patron withinside the simplest way. Using the massive quantity of statistics to be had on clients (and ability clients), a patron segmentation evaluation lets in entrepreneurs to identify discrete agencies of clients with an excessive diploma of accuracy primarily based totally on demographic, behavioral and different indicators.

Evaluation of RFM (Recency, Frequency, and Monetary) is a famed approach is worn for comparing the clients primarily based totally on their shopping for behavior. Scoring method was developed to test Recent, Frequency, and Finance ratings. Finally, ratings of all three variables are strengthened as RFM ratings from different ranges (Haiying and Yu, 2010) which are compiled to anticipate recants trends for studying existing and higher sponsor transactions history. Next step is defined as the remaining time the consumer buys. The latest currency is the type of days the sponsor takes between purchases. The latest small payment means that the sponsor visits the organization frequently in a timely manner. Similarly, extra money means that the sponsor is less likely to go to the organization soon. Frequency is described because the variety of transaction a patron makes in a selected period. The better the fee of frequency the greater unswerving are the clients of the organization. Cash is defined as the amount spent by the investor over a period in a favorable period. The improvement in the amount of money spent by the large sales they provide to the organization. Each sponsor is given 3 different ratings of the latest, frequency, and economic volatility. Score points are used within a range from five to one. The core quintile is given a 5-point scale, while the others are given 4, 3, 2 and 1. In recent years, there has been a significant increase in the number of opposition groups among companies in care within the arena. Customer retention is more important than purchasing the latest customers. Customer segregation allows people’s messages to speak more to target audiences.

In this project, behavioral segmentation was used to group customers based on their purchasing behavior. This approach is particularly effective because it provides insights into how customers interact with a company's products or services.

RFM (Recency, Frequency, Monetary) analysis is a popular technique for customer segmentation that helps businesses identify their most valuable customers based on their purchasing behavior. It involves analyzing three key metrics: how recently a customer has made a purchase (recency), how frequently they make purchases (frequency), and how much money they spend (monetary). By analyzing these metrics, businesses can identify their high-value customers and tailor their marketing strategies to meet their specific needs.

To perform RFM analysis, the following steps were taken:

Data Collection: Data on customer purchases and interactions with the company's products or services were collected.

Data Preparation: The data was cleaned and normalized to ensure accuracy and completeness.

Analysis: RFM scores were calculated for each customer based on their recency, frequency, and monetary value.

Segmentation: Customers were segmented into distinct groups based on their RFM scores.

1.2 Objective

The objectives of the project work are as -

- To Understand and Analyze Customer Behavior.
- To Leverage Advanced Data Analytics Tools and Implementing Effective Customer Segmentation.
- To Enhance Data Visualization and Interpretation.
- To Perform RFM Analysis for Customer Value Identification.
- To have Segment Customers for Strategic Marketing.

The study aims to emphasize the importance of customer segmentation in developing effective marketing strategies by utilizing RFM (Recency, Frequency, Monetary) analysis to identify high-value customer segments based on purchasing behavior. Through the integration of advanced data analytics and visualization tools, the study seeks to provide actionable insights that will enable businesses to tailor their marketing efforts, enhance customer engagement, and ultimately drive profitability.

Chapter 2
LITERATURE REVIEW

Jiang and Tuzhilin (2009) identified that both customer segmentation and buyer targeting are necessary to improve the marketing performances. These two tasks are integrated into a step-by-step approach, but the problem faced is unified optimization. To solve the problem, the author proposed the K-Classifiers Segmentation algorithm. This approach focuses on distributing more resources to those customers who give more returns to the company. A sizable number of authors had written about different methods for segmenting the customers.

Cho and Moon (2013) proposed a customized recommendation system using weighted frequent pattern mining. Customer profiling is performed to find the potential customers using the RFM model. The author has defined varied weights for each transaction to generate weighted association rules through mining. Using the RFM model will provide a more accurate recommendation to the customer which in turn increases the profit of the firm.

Zahrotun (2017) used the customer data from online to identify the finest customer using Customer Relationship Management (CRM). By applying the CRM concept for online shopping, the author identifies the potential customers by segmenting them which helps us in increasing the profits for the company. So, to perform customer segmentation and marketing to customers in an accurate way the Fuzzy C-Means Clustering Method is used. Thus, this helps the customers to get special facilities in more than one category in the appropriate marking strategies according to their needs.

Sheshasaayee and Logeshwari (2017) designed a new integrated approach by segmentation with the RFM and LTV (Life Time Value) methods. They used a two-phase approach with the first phase being the statistical approach and the second phase is to perform clustering. They aim to perform K-means clustering after the two-phase model and then use a neural network to enhance their segmentation.

Chapter 3
METHODOLOGY AND TECHNIQUES

3.1 Methodology:

RFM Analysis Recency, frequency and monetary (RFM) analysis is a powerful and recognized technique in database marketing. It is widely used to rank the customers based on their prior purchasing history. RFM analysis finds use in a wide range of applications involving many customers such as online purchase, retailing, etc. This method groups the customers based on three dimensions, recency(R), frequency (F), and monetary (M).

Recency– When was the last time the customer made a purchase? Recency value is the number of days a customer takes between two purchases. A smaller value of recency implies that the customer visits the company repeatedly in a short period. Similarly, a greater value implies that the customer is less likely to visit the company shortly.

Frequency– How many times did the customer purchase? Frequency is defined as the number of purchases a customer makes in a specific period. The higher the value of frequency the more loyal are the customers of the company.

Monetary– How much money did the customer spend? Monetary is defined as the amount of money spent by the customer during a certain period. The higher the amount of money spent the more revenue they give to the company.

Each customer is assigned with three different scores for recency, frequency, and monetary variables. Scoring is done in the scale from 5 to 1. The top quintile is given a score of 5, and the others are given 4, 3, 2 and 1. The scores can be assumed to have unique characteristics.

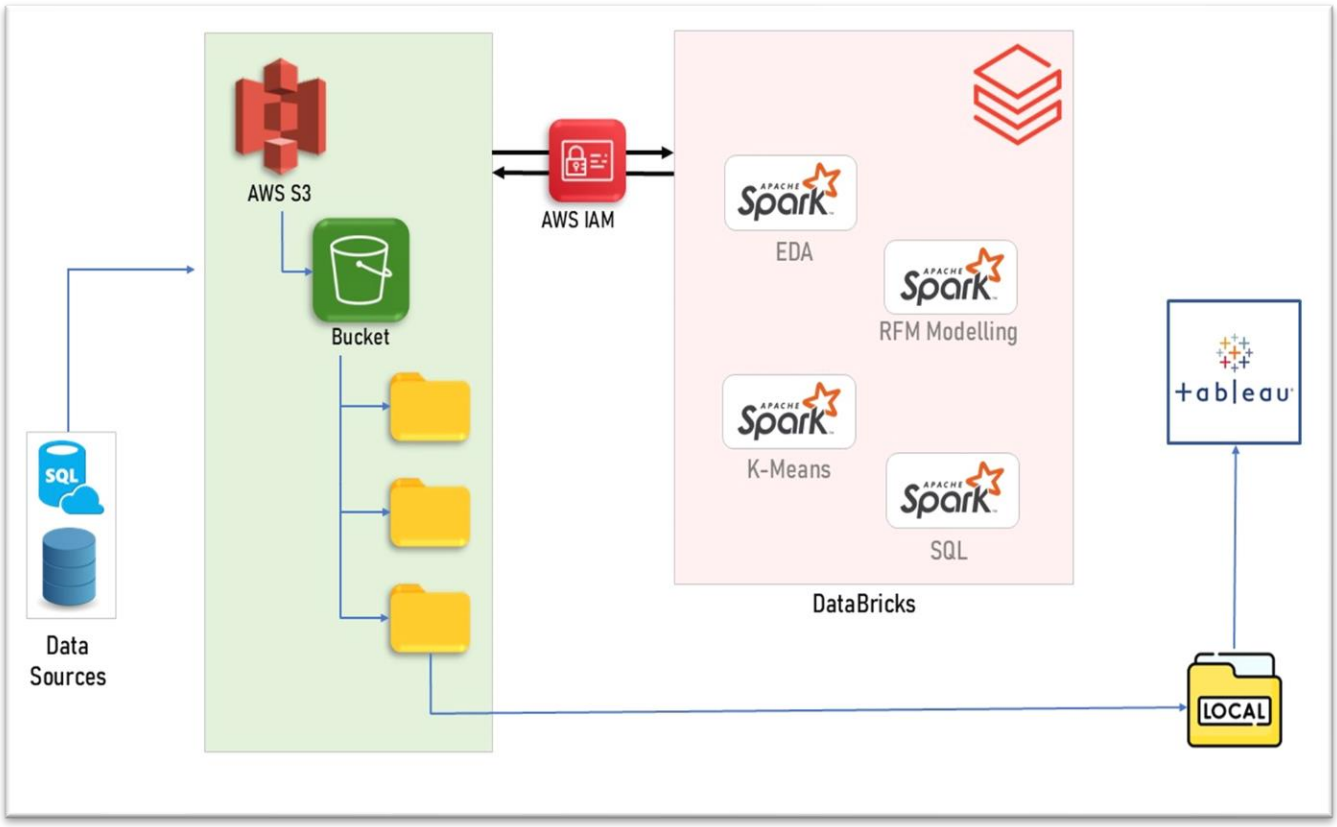


Fig.1 Architecture

3.2 Dataset

This dataset is composed of several columns, each of which represents different aspects of retail transactions. The dataset is a comprehensive collection of retail transactions, recording each purchase's essential details such as product information, quantity, pricing, customer identity, and geographic location. This type of data is invaluable for analysing. The dataset is a comprehensive collection of retail transactions, recording each purchase's essential details such as product information, quantity, pricing, customer identity, and geographic location. This type of data is invaluable for analysing customer purchasing behaviour, sales trends, and inventory management.

In this project, the dataset can be utilized to:

- **InvoiceNo**: A unique identifier for each transaction (invoice). Multiple entries with the same InvoiceNo indicate different products sold within the same transaction.
- **StockCode**: This is a unique code assigned to each product. It serves as an identifier within the inventory system.
- **Description**: A textual representation of the product sold. This column gives more context to the StockCode by describing the item.
- **Quantity**: This represents the number of units of a particular product sold in the transaction. A higher quantity indicates that multiple units of the item were purchased.
- **InvoiceDate**: The date and time when the transaction occurred. This field is crucial for time-based analysis, such as identifying peak purchasing times.
- **UnitPrice**: The price per unit of the product. This column allows for the calculation of total revenue per transaction when combined with Quantity.
- **CustomerID**: A unique identifier assigned to each customer. This field is essential for analysing customer purchasing patterns and behaviours.
- **Country**: This indicates the country where the customer is located.

3.3 Model Description

Preprocessing-

In this project, the data preprocessing phase was meticulously carried out to ensure that the dataset was ready for analysis. The raw transaction data underwent a thorough cleaning process, which involved handling missing values, removing duplicates, and standardizing the data types to ensure consistency across all records. Following the cleaning process, the Recency, Frequency, and Monetary (RFM) values for each customer were calculated. These RFM values were then normalized to bring all the variables onto a similar scale, thus preventing any single feature from disproportionately influencing the clustering process.

Initially, outliers were identified and removed to prevent them from skewing the results. The data was then filtered to focus specifically on transactions from the United Kingdom, ensuring that the analysis was geographically relevant. Dates were carefully handled to accurately calculate the Recency metric, which required converting transaction dates into a uniform format and calculating the time elapsed since the last purchase for each customer.

After these preprocessing steps, the Recency, Frequency, and Monetary (RFM) values for each customer were calculated and subsequently normalized. This normalization ensured that each RFM variable contributed equally to the clustering process.

K-means clustering:

Once the preprocessing was completed, K-Means clustering was employed to segment the customers based on their RFM scores. The Elbow Method was utilized to determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against various cluster numbers.

The point at which the WCSS began to level off indicated the ideal number of clusters. These clusters represented different customer segments, each with distinct purchasing behaviors. This segmentation allowed for a more targeted analysis, enabling the development of tailored strategies to engage and retain customers based on their specific behavior patterns.

Chapter 4
IMPLEMENTATION

Tools and Technologies:

Python, Pandas, PySpark, K-Means Clustering (from Scikit-Learn), Matplotlib/Seaborn, AWS S3, AWS IAM, Databricks, Tableau.

Hardware and Software Requirements:

While the project primarily utilized cloud-based tools and platforms, the following hardware and software were used locally:

- **Hardware**: Laptop: Windows 11, 8 GB RAM, 256 GB SSD, Ryzen 3 Processor.

This hardware configuration was sufficient for local development, testing, and managing cloud resources.

- **Software:**

The implementation of this project was carried out through a structured and iterative process, leveraging various tools and platforms to handle large-scale data processing, modeling, and visualization.

1. **Uploading Raw Data to AWS S3**: The process began with the collection of raw transactional data, which was then uploaded to an AWS S3 bucket. This cloud-based storage solution provided a scalable and secure environment for storing large datasets.
2. **Fetching and Cleaning Data on Databricks Using PySpark**: The raw data was fetched from AWS S3 into Databricks, where PySpark was used for data cleaning and preprocessing. This step involved removing duplicates, handling missing values, filtering for relevant geographical data (focusing on the United Kingdom), and dealing with outliers. Dates were also standardized to ensure consistency across the dataset.
3. **Saving Cleaned Data Back to S3**: After cleaning, the processed data was saved back into a separate folder in the AWS S3 bucket. This organized the data flow and allowed for easy retrieval in subsequent steps.

4. **Fetching Cleaned Data for RFM Modeling**: The cleaned data was then fetched again from the S3 bucket to perform RFM (Recency, Frequency, Monetary) analysis. PySpark was utilized to calculate RFM scores for each customer, providing a comprehensive view of customer behaviour.
5. **Saving RFM Results Back to S3**: The RFM results, which quantified customer purchasing patterns, were saved back to S3. This ensured that the RFM scores were preserved and could be accessed for further analysis.
6. **Fetching RFM Data for K-Means Clustering**: The next step involved fetching the RFM data from S3 for K-Means clustering. This clustering technique, implemented using Scikit-Learn, grouped customers into distinct segments based on their RFM scores.
7. **Running SQL Queries on Clustered Data**: SQL queries were run on the clustered data within Databricks to perform additional analyses and derive insights. This allowed for the exploration of customer segments in greater detail.
8. **Saving Final Data Back to S3**: The final dataset, enriched with cluster labels and additional insights, was saved back to S3. This comprehensive dataset was now ready for visualization.
9. **Creating Visualizations in Tableau**: Tableau was used to connect to the final data stored in S3 and create interactive visualizations and dashboards. These visualizations provided a clear and actionable view of customer segments, enabling data-driven decision-making.

Tableau:

Tableau is a powerful data visualization tool that allows users to create interactive and shareable dashboards. It is widely used for business intelligence and data analysis because of its intuitive interface and robust capabilities for visualizing complex data.

Tableau is a versatile tool that can greatly enhance RFM analysis by providing visual and interactive means to explore customer segments.

Chapter 5

RESULTS

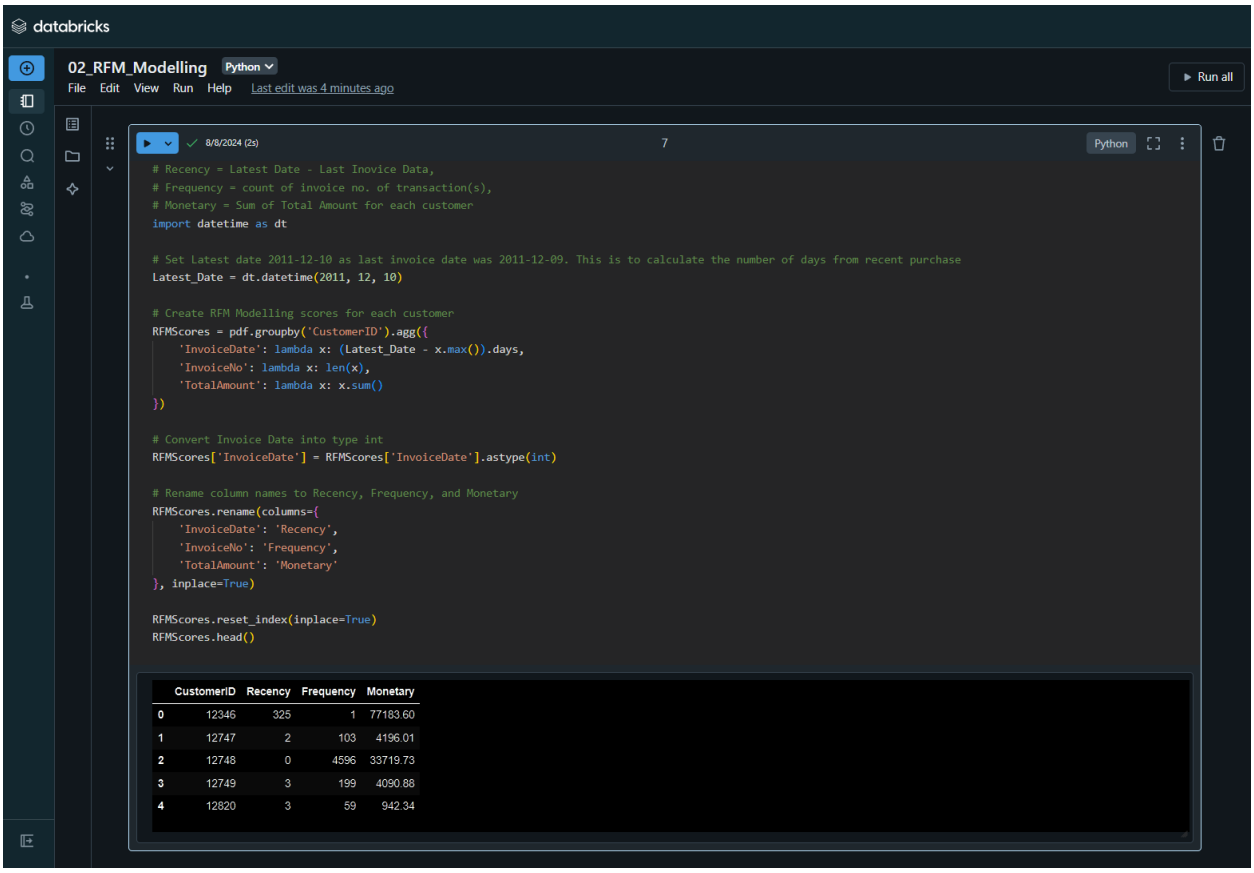
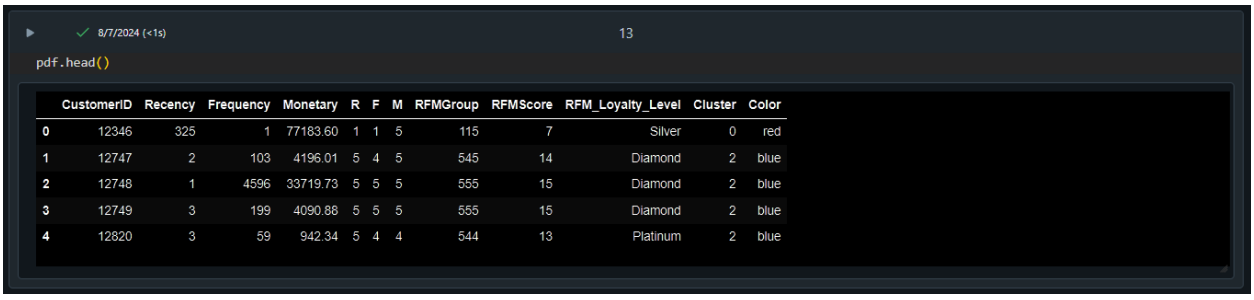
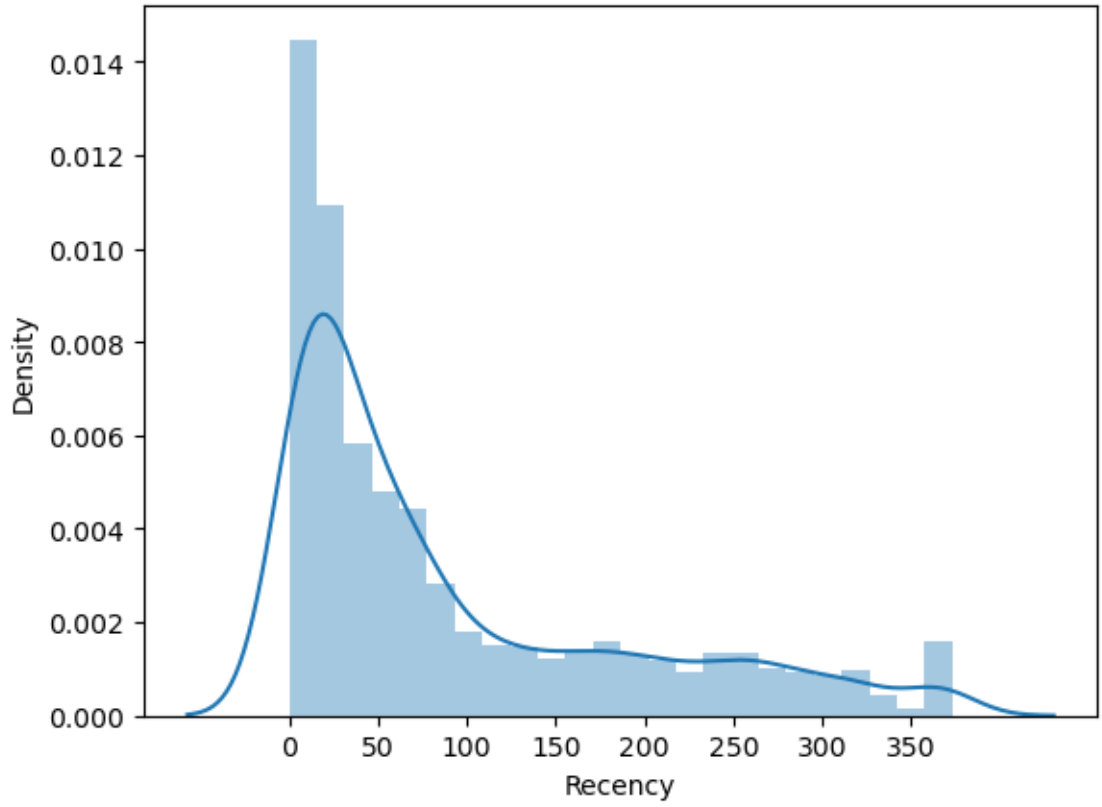


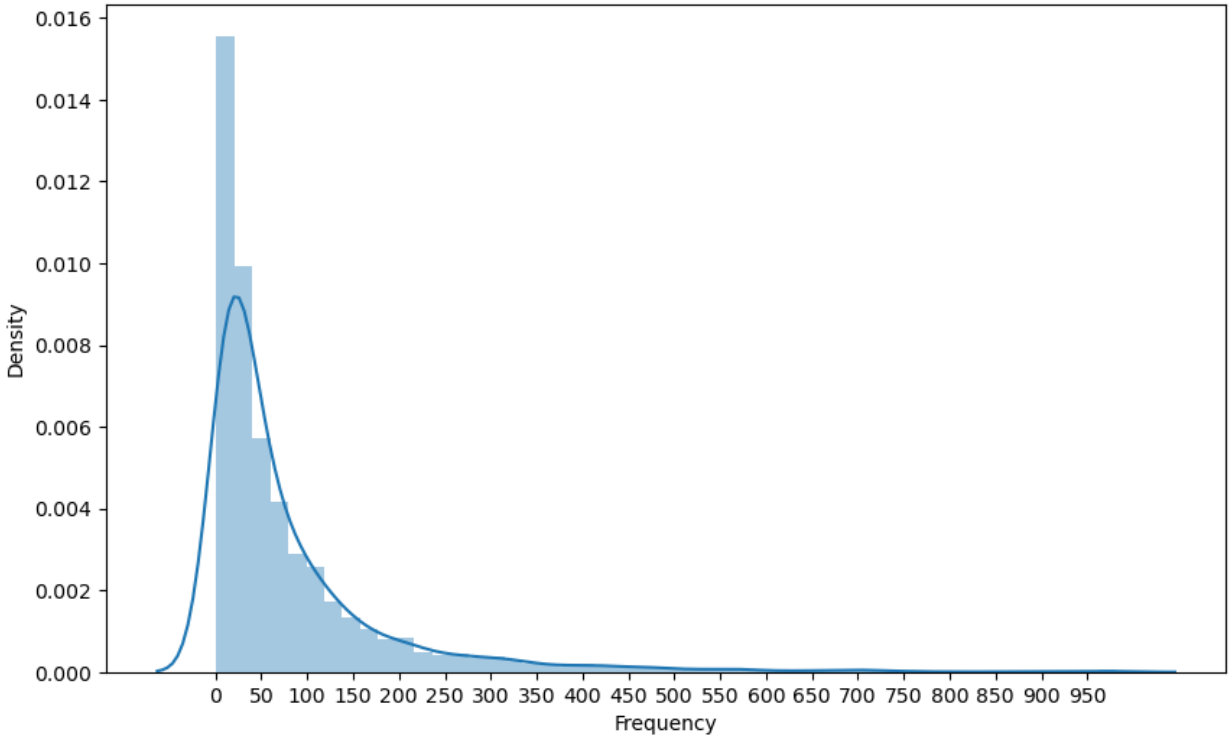
Fig. Python code for RFM Modelling

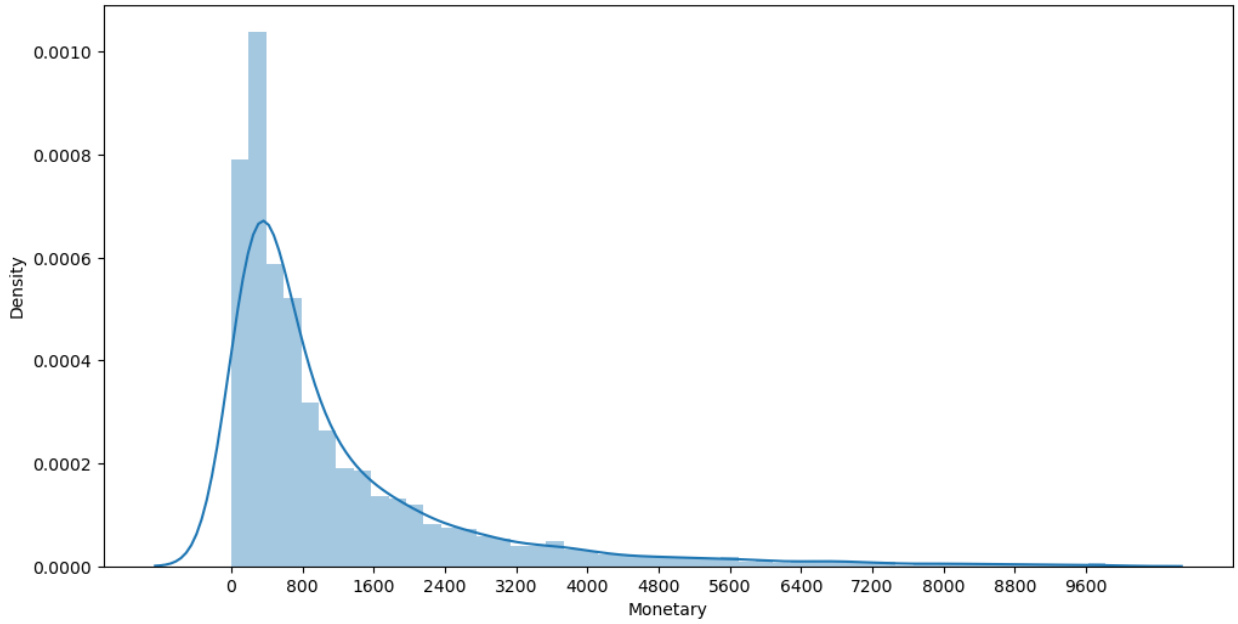


	CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level	Cluster	Color
0	12346	325	1	77183.60	1	1	5	115	7	Silver	0	red
1	12747	2	103	4196.01	5	4	5	545	14	Diamond	2	blue
2	12748	1	4596	33719.73	5	5	5	555	15	Diamond	2	blue
3	12749	3	199	4090.88	5	5	5	555	15	Diamond	2	blue
4	12820	3	59	942.34	5	4	4	544	13	Platinum	2	blue

Fig. Results after RFM and K-Means


0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
50
100
150
200
250
300
350
Recency


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950
Frequency



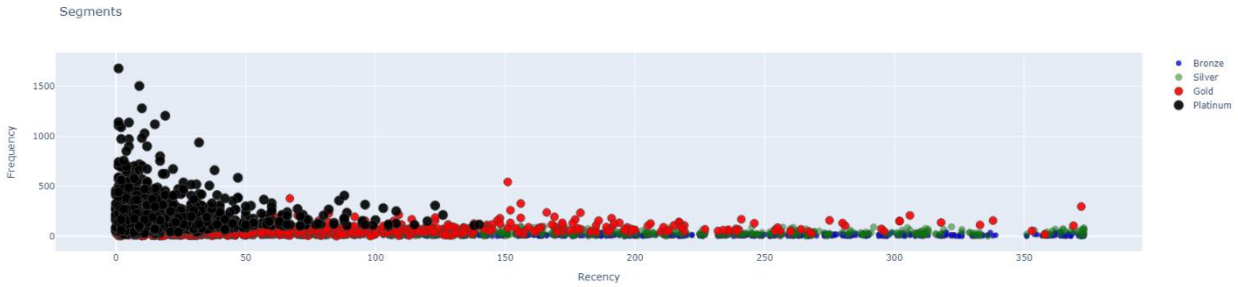


Fig. Recency to Frequency

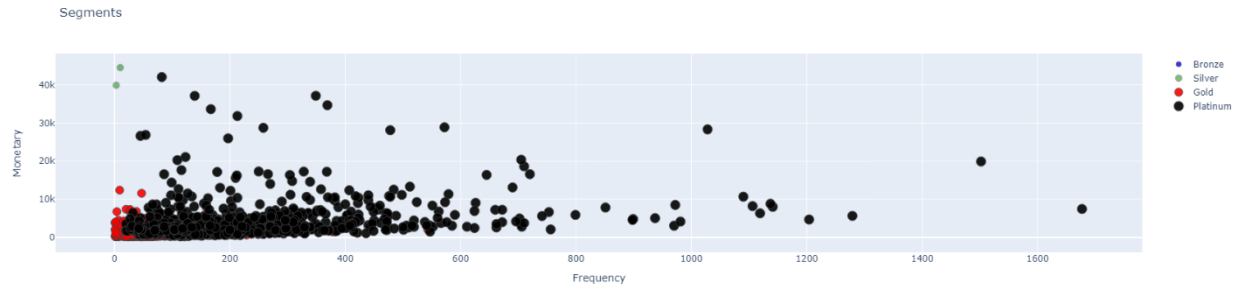
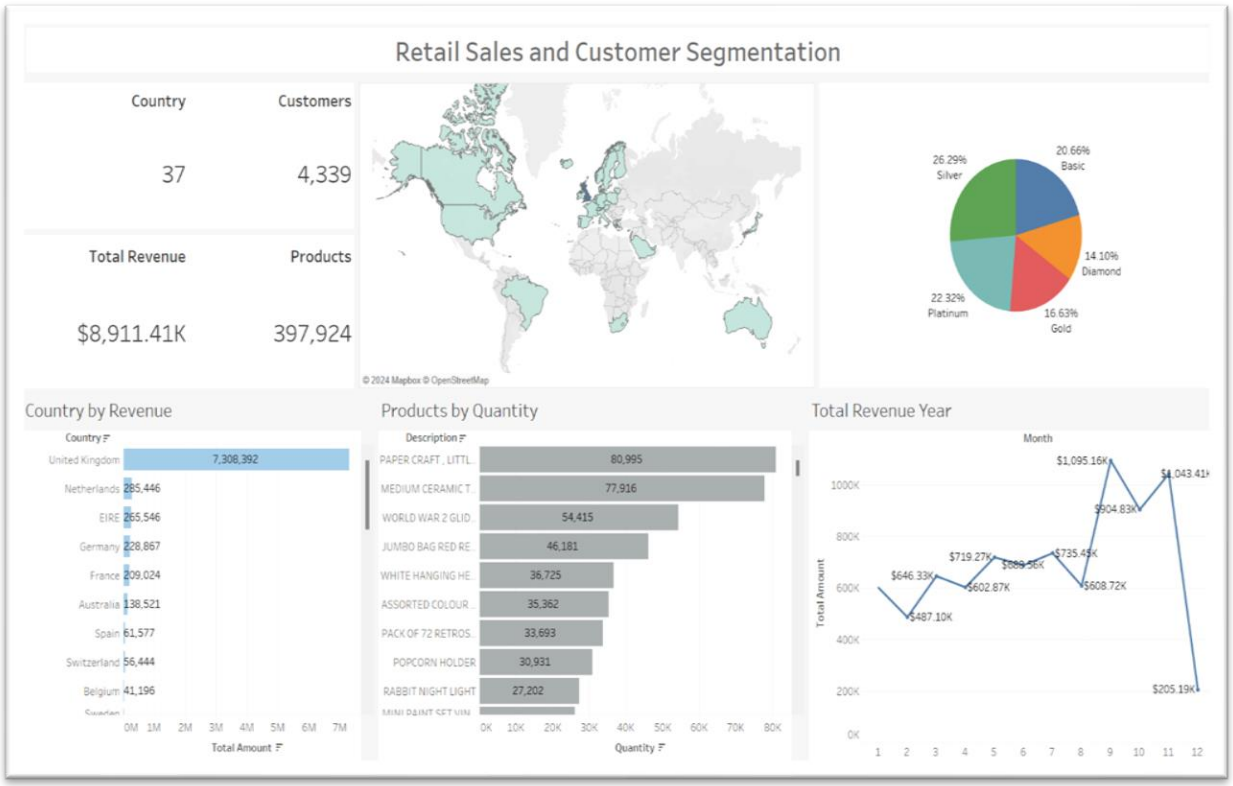


Fig. Frequency to Monetary

Fig. Recency to Monetary

Tableau Dashboard -



Chapter 6

CONCLUSION

6.1 Conclusion

The project successfully implemented customer segmentation through RFM (Recency, Frequency, Monetary) analysis and further refined the segments using K-Means clustering. By leveraging cloud-based platforms like AWS S3 and Databricks, the project efficiently processed large volumes of customer data.

The combination of RFM modeling and K-Means clustering provided a robust framework for understanding customer behavior, enabling the identification of distinct customer segments. These insights allow for more targeted marketing strategies, enhancing customer engagement and retention. The visualizations created in Tableau further facilitated the interpretation of the results, making it easier to communicate findings to stakeholders.

6.2 Future Enhancement –

Future work could focus on integrating additional data sources, such as customer feedback or social media interactions, to enrich the segmentation process. Incorporating machine learning models could also improve the predictive capabilities of the analysis, enabling more dynamic customer segmentation. Additionally, real-time data processing could be introduced to allow for more immediate and responsive marketing strategies.

Expanding the scope of the project to include personalization techniques, such as recommendation systems, could further enhance the customer experience and drive business growth.

Chapter 7

REFERENCES

[1] Haiying, M. and Yu, G. 2010. Customer segmentation study of college students based on the rfm.
3860–3863.

[2] He, X. and Li, C. 2016. The research and application of customer segmentation on e-commerce websites. 203–208.

[3] Jiang, T. and Tuzhilin, A. 2009. Improving personalization solutions through optimal segmentation
of customer bases. IEEE Trans. Knowl. Data Eng. 21, 305–320.

[4] Lee, D.-H. and Memon, K. 2016. Generalised fuzzy c-means clustering algorithm with local information. IET Image Processing 11.

[5] Liu, C., Chu, S.-W., Chan, Y.-K., and Yu, S. 2014. A modified k-means algorithm - two-layer k-
means algorithm. 447–450.

[6] Lu, N., Lin, H., Lu, J., and Zhang, G. 2014. A customer churn prediction model in telecom industry
using boosting. Industrial Informatics, IEEE Transactions on 10, 1659–1665.

[7] Shah, S. and Singh, M. 2012. Comparison of a time efficient modified k-mean algorithm with k-
mean and k-medoid algorithm.

[8] Sheshasaayee, A. and Logeshwari, L. 2017. An efficiency analysis on the tpa clustering methods
for intelligent customer segmentation. 784–788.

[9] Srivastava, R. 2016. Identification of customer clusters using rfm model: A case of diverse purchaser classification.

[10] Zahrotun, L. 2017. Implementation of data mining technique for customer relationship management (crm) on online shop tokodiapers.com with fuzzy c-means clustering. 299–303.