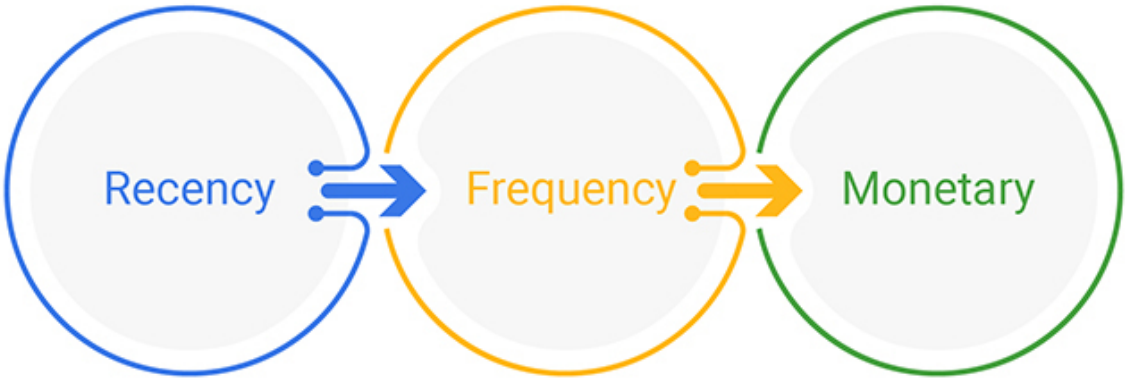

Improving Customer Segmentation With RFM Marketing Analysis

Customer Segmentation using RFM Analysis

IE6400 - Foundation of Data Analytics Engineering

Final Report

Group Number: 28

Hrishik Bhaven Parmar (002815908)

Aishwarya Belavakadi Subrahmanya (002820128)

Karan Manoj Dalal (002836524)

Batta Aditya Yadav (002874554)

1. Introduction

1.1 Background

Understanding and effective engagement with customers is essential for success in today's changing business environment. A unique opportunity exists for businesses to apply advanced methods of analyzing customer segments in an increasing amount of data generated by the online commerce sector. This project is focused on the segmentation of customers through RFM Analysis, a powerful methodology that uses correlation, frequency, and money metrics to analyze customer behavior according to their buying patterns. RFM analysis is becoming a powerful tool, providing insights that are beyond the typical segmentation of demographics, as companies attempt to integrate their marketing and retention strategies into each customer's needs. This project will seek to decipher patterns in customer data and provide actionable recommendations for businesses that want to improve their marketing effectiveness and satisfaction through the use of an existing eCommerce database derived from Kaggle.

1.2 Objective

This project's main goal is to utilize RFM (Recency, Frequency, Monetary) analysis's capability for consumer segmentation in the framework of an eCommerce dataset. The aim is to extract valuable insights from the complexities of consumer behavior so that companies may divide their clientele into discrete groups. This study seeks to reveal trends that go beyond traditional demographic analysis by calculating RFM indicators and then segmenting the data to provide a detailed insight into customers' recent transaction history, frequency of purchases, and monetary worth. The ultimate goal is to equip companies with actionable intelligence so they can efficiently customize marketing and retention plans for every recognized consumer segment. By doing this, the initiative hopes to advance the fields of customer relationship management and data-driven decision-making in the eCommerce industry.

2. Data Preprocessing

2.1 Dataset Overview

The UCI Machine Learning Repository's "Online Retail" dataset offers a rare look into the online sales of a UK-based non-store online retailer that specializes in unusual gifts for every occasion. With great generosity, Dr. Daqing Chen, the director of the Public Analytics division at the London South Bank University School of Engineering, contributed to this dataset of transactions from December 1, 2010, to September 9, 2011.

The dataset, which consists of 541,909 entries and 8 columns, includes information such as the invoice number, stock code, quantity, description, invoice date, unit price, customer ID, and country. It is a perfect resource for our RFM study because of its wide range of content, which includes wholesalers as important clients. This information makes it easier to segment customers and provides opportunities for a number of other studies, including time series, clustering, and classification.

2.2 Data Cleaning

A preliminary examination of the dataset during the data cleaning stage found null values in the "Description" and "CustomerID" columns. To fix this and guarantee completeness in this property, the label 'Unknown' was provided in the 'Description' column where it was null. To ensure data integrity, the 'CustomerID' column was handled by forward-filling the null values.

Furthermore, careful examination was given to the formatting of the 'InvoiceDate' column. After 'Date' and 'Time' were taken out of 'InvoiceDate,' further investigation showed that 'Time' didn't need to be changed. However, the Pandas library was used to convert the 'Date' column from the original MM/DD/YY format to the conventional YYYY-MM-DD format in order to improve consistency and compatibility.

Following these thorough data pretreatment procedures, the final dataset is prepared for additional analysis. The coherence of the dataset is enhanced by the way null values are handled and the date formats are standardized, which paves the way for later RFM analysis and customer segmentation.

3. RFM Analysis

3.1 RFM Calculation

To aid in later segmentation, each customer's purchase activity is methodically quantified throughout the RFM (Recency, Frequency, Monetary) computation phase. For every consumer, the computation of recency entails figuring out how many days have passed since their previous transaction. The recency measure for every transaction is represented by the 'Recency' column, which is created by deducting the transaction date from the dataset's maximum date.

In order to calculate frequency, the dataset is grouped according to distinct customer IDs, and the number of unique invoices linked to each client is counted. One important part of the RFM analysis is the resulting 'Frequency' data frame, which contains the count of invoices for each customer.

Finding the overall monetary value of each customer's transactions is a key component of monetary analysis. This is accomplished by dividing the total number of products bought by the unit price of each item, and then adding up the results for each customer ID. The total price for every customer is recorded in the 'Monetary' data frame.

A thorough summary of each customer's recency, frequency, and monetary scores is then provided by merging the Recency, Frequency, and Monetary data frames into a single RFM data frame. Based on these important variables, a more sophisticated understanding of customer behavior can be achieved by using the RFM information as the basis for further customer segmentation and analysis.

3.3 RFM Segmentation

Quartiles are used in the RFM segmentation process to group clients according to their frequency, monetary ratings, and recentness. The metrics are given quartile cut labels to form the 'RecencyScore,' 'FrequencyScore,' and 'MonetaryScore' columns. For additional study, these scores are then transformed to integer format.

Each customer's unique ratings for recency, frequency, and monetary indicators are added together to provide a full RFM score. Each customer's position is reflected in a composite score across all three dimensions in the ensuing 'RFMScore' column.

The customer ID, associated frequency, monetary, and recency scores, as well as the computed quartile-based scores for recency, frequency, monetary, hence the total RFM score, are all contained in the final RFM data frame. A detailed grasp of each customer's engagement and value is made possible by this segmentation, which paves the way for focused marketing campaigns and client retention initiatives.

3.4 Customer Segmentation

The RFM data frame is ready for clustering during the customer segmentation step by choosing the pertinent columns ('RecencyScore,' 'FrequencyScore,' and 'MonetaryScore') and putting them into a new data frame called 'X.' The Elbow Method determines the ideal number of clusters for the K-Means algorithm. Six clusters appear to be the best option when the number of clusters is plotted against the inertia (within-cluster sum of squares), which shows an inflection point.

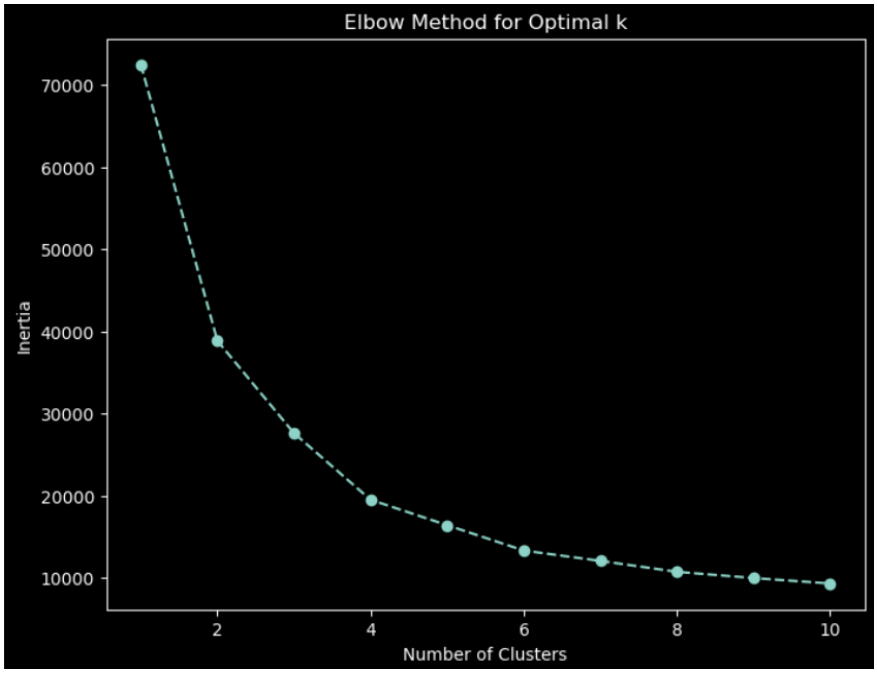

Fig 3.4.1 *Elbow method for finding optimal clusters.*

The 'X' data frame is then subjected to a six-cluster K-Means clustering technique. The 'Cluster' column in the original RFM data frame is given the cluster labels that are produced. The segmentation group assigned to each client is identified by this 'Cluster' field.

The 'Cluster' column, which offers insights into the segmentation of each client based on their frequency, monetary scores, and recency, is included to the final RFM data frame. Businesses may better target their marketing efforts and client retention campaigns by customising their tactics for each cluster as a result of this segmentation.

3.5 Segment Profiling

We want to investigate the distinctive characteristics of every client cluster found by the K-Means algorithm during the segment profiling stage. The method generates mean values for each segment in the'segment_profiles' data frame by aggregating three crucial metrics: recency, frequency, and monetary (RFM). The data frame offers a brief summary of the mean Recency, Frequency, and Monetary values for each segment after changing the 'Cluster' column to 'Segment' for clarity.

- **Segment 0**: This segment, which consists of very active and high-spending clients, has an average recency of around 69 days, a high frequency of 44 orders, and a significant monetary value of almost $21,287. These clients are regularly involved and make significant, regular purchases.
- **Segment 1**: Customers in this group are less engaged; they spend, on average, $608 and have a recency of around 61 days, as well as a lower frequency of three orders. This category consists of less involved and thriftier consumers.
- **Segment 2**: Customers with intermediate involvement and expenditure are included in this section. They have a higher average spend of $3,033, moderate frequency of orders (11), and a more recent order history (around 258 days). Spending and involvement are reasonable, although recent activity is lower.
- **Segment 3**: Moderately engaged clients who spend around $2,333, with an average order frequency of eight, and an average order recency of about 65 days. Customers in this sector tend to be somewhat engaged and have modest purchasing habits.
- **Segment 4**: Customers with intermediate involvement and expenditure are included in this section. They have a higher average spend of $3,033, moderate frequency of orders (11), and a more recent order history (around 258 days). Spending and involvement are reasonable, although recent activity is lower.
- **Segment 5**: Moderately engaged clients who spend around $2,333, with an average order frequency of eight, and an average order recency of about 65 days. Customers in this sector tend to be somewhat engaged and have modest purchasing habits.

This segmentation and profiling approach allows businesses to tailor marketing strategies and retention efforts based on the specific needs and behaviors of each customer segment.

3.6 Marketing Recommendations

Based on the client profiles that have been segmented by RFM analysis and subsequent clustering, the marketing recommendations that follow have been carefully customized to correspond with the distinct features displayed by each of the identified consumer segments. To maximize revenue production and foster steadfast loyalty, a strategy that emphasizes VIP treatment, exclusive benefits, and personalized offers has been developed for Segment 0, which is characterized by Engaged High-Spending Customers.

On the other hand, recommendations focusing on re-engagement efforts and improvements to the entire customer experience are concentrated in Segment 1, which is defined by Low-Engagement and Low-Spending. These programs are deliberately crafted to reignite curiosity, encourage return visits, and improve overall contentment.

Segments 2 and 3, categorized as Moderately Engaged with Moderate Spending, are liable to customized advertising and focused reactivation efforts. These initiatives are carefully designed to support higher levels of engagement and encourage more expenditure, in line with the particular characteristics of every market group.

Win-back efforts, incentive purchases, and referral programs are aimed at customers in Segment 4, who are classified as Low-Engagement and Low-Spending. These programs aim to rekindle curiosity, encourage more frequent interactions, and maximize the power of client recommendations.

Finally, recommendations focusing on loyalty rewards, individualized engagement techniques, and exclusive access are given to Engaged High-Frequency, High-Spending Customers in Segment 5. This all-encompassing strategy is designed to maintain and improve their high levels of involvement and expenditure.

3.7 Visualization

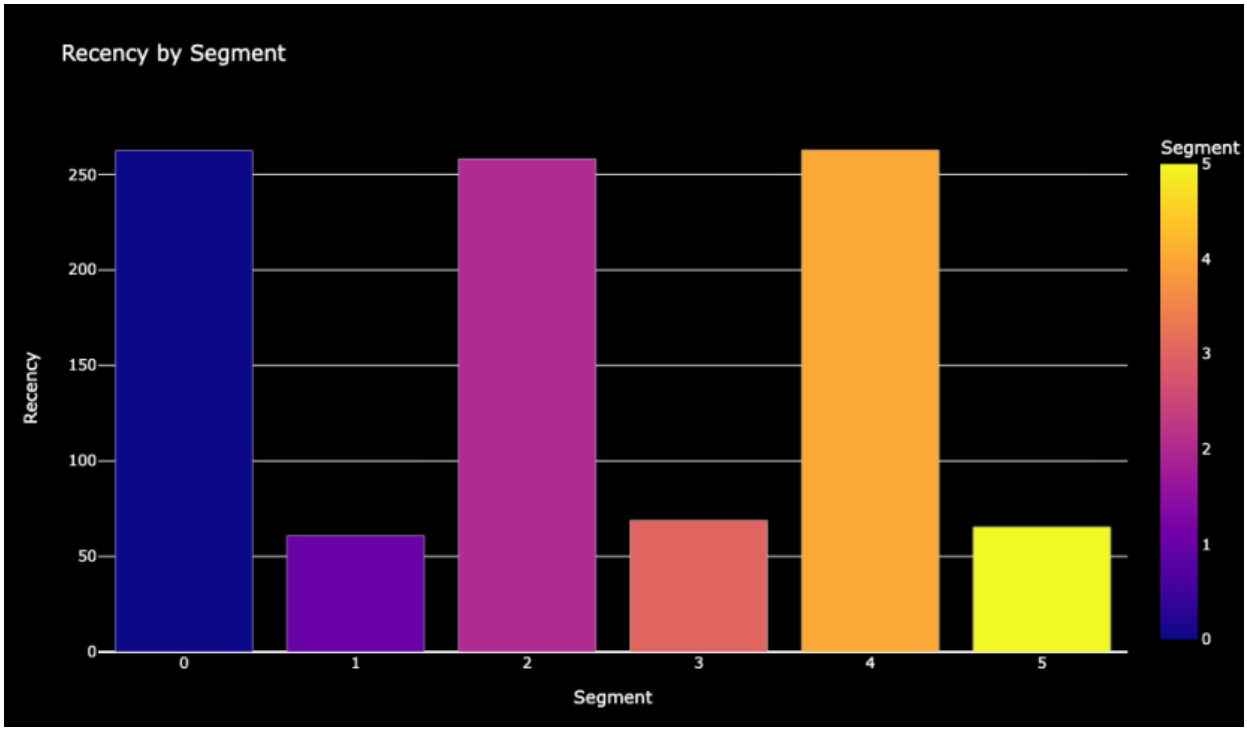

Fig 3.7.1 *Recency by Segmentation*

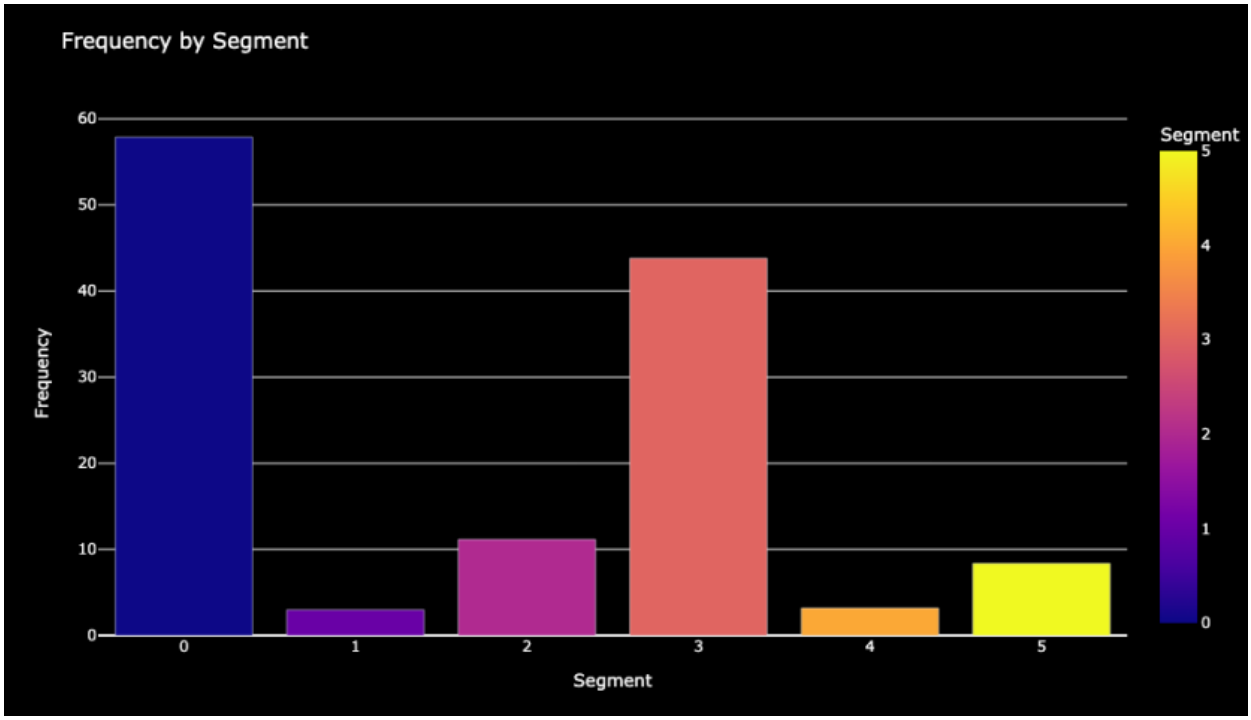

Fig 3.7.2 *Frequency by Segmentation*

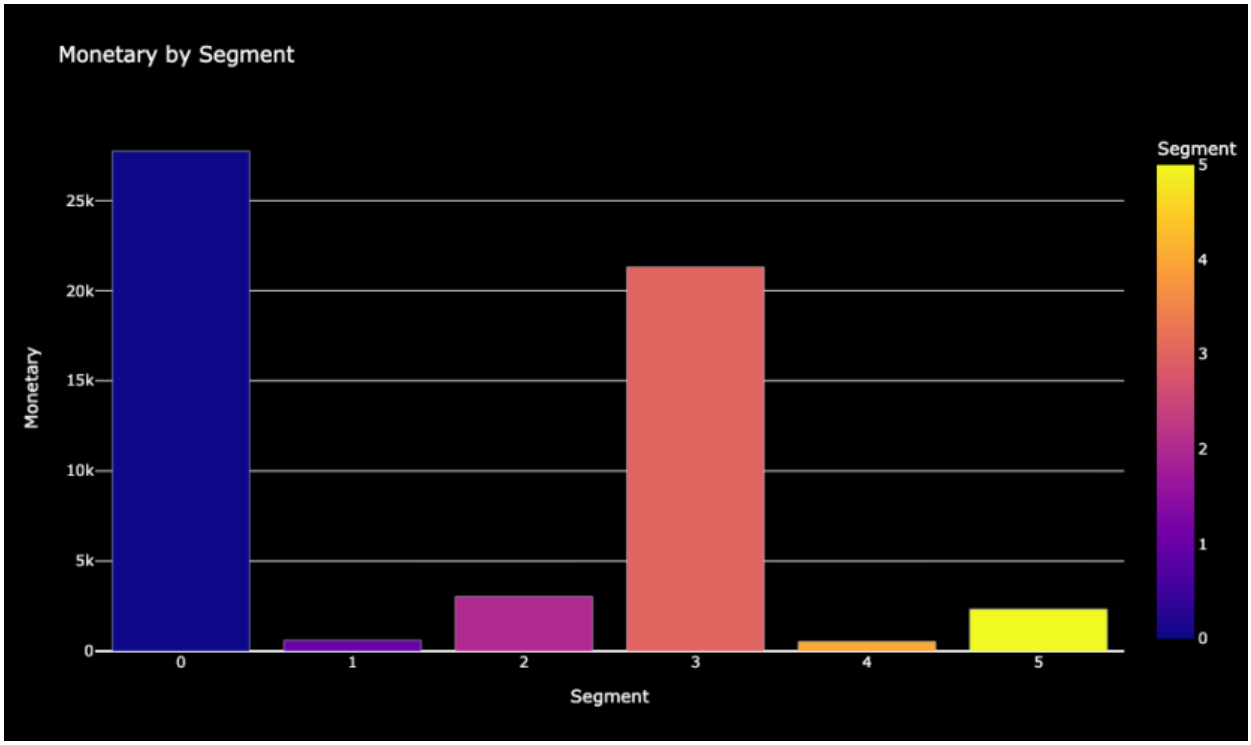

Fig 3.7.3 *Monetary by Segmentation*

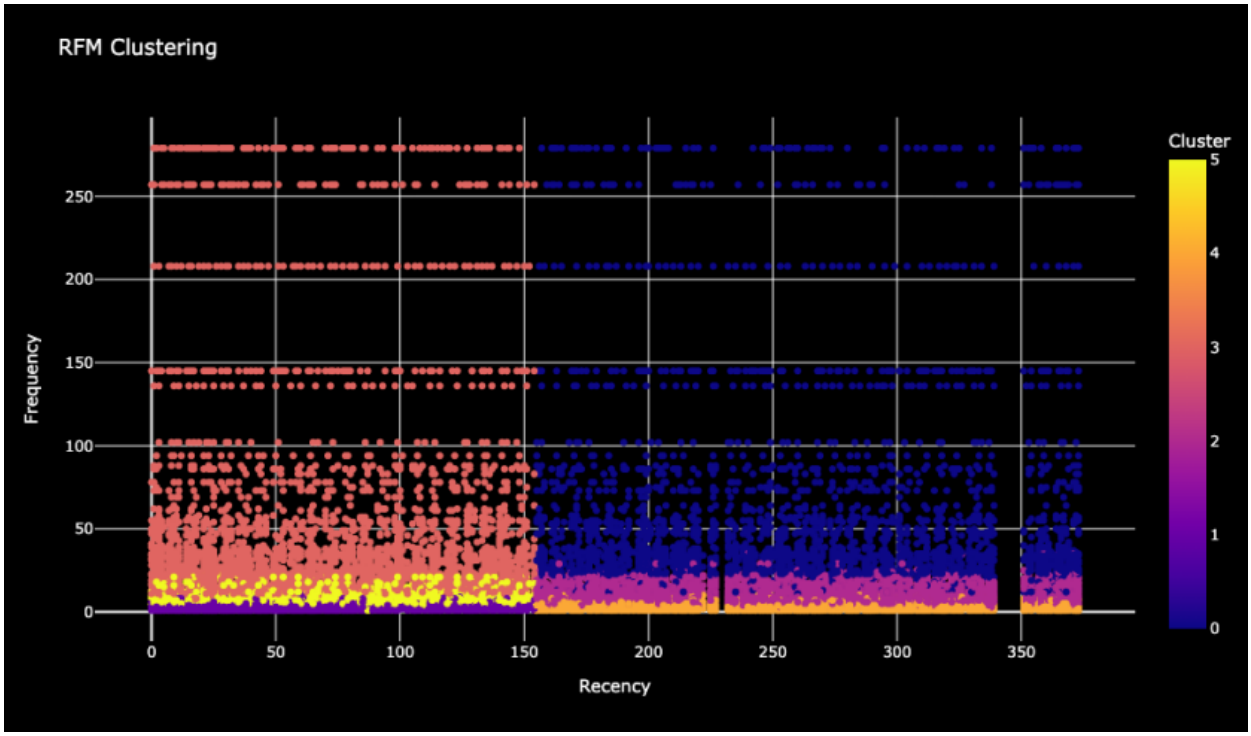

Fig 3.7.4 *RFM Clustering Frequency/Recency*

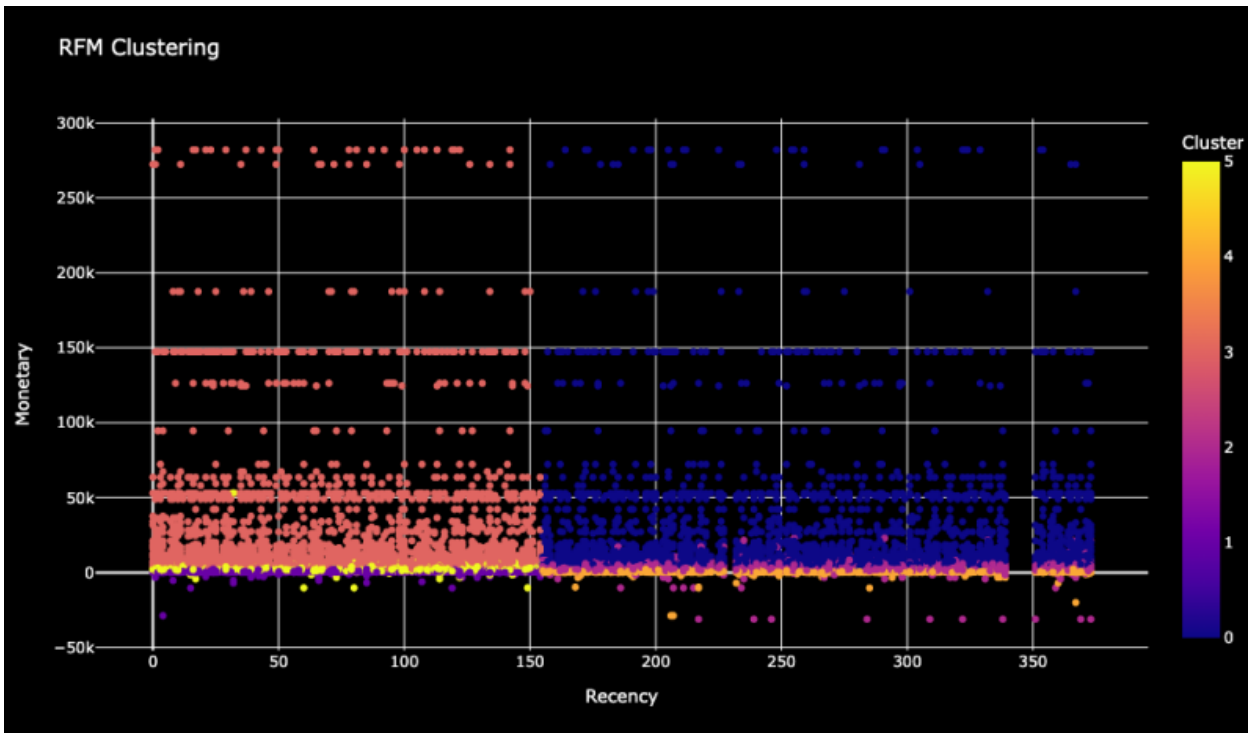

Fig 3.7.5 *RFM Clustering Monetary/Recency*

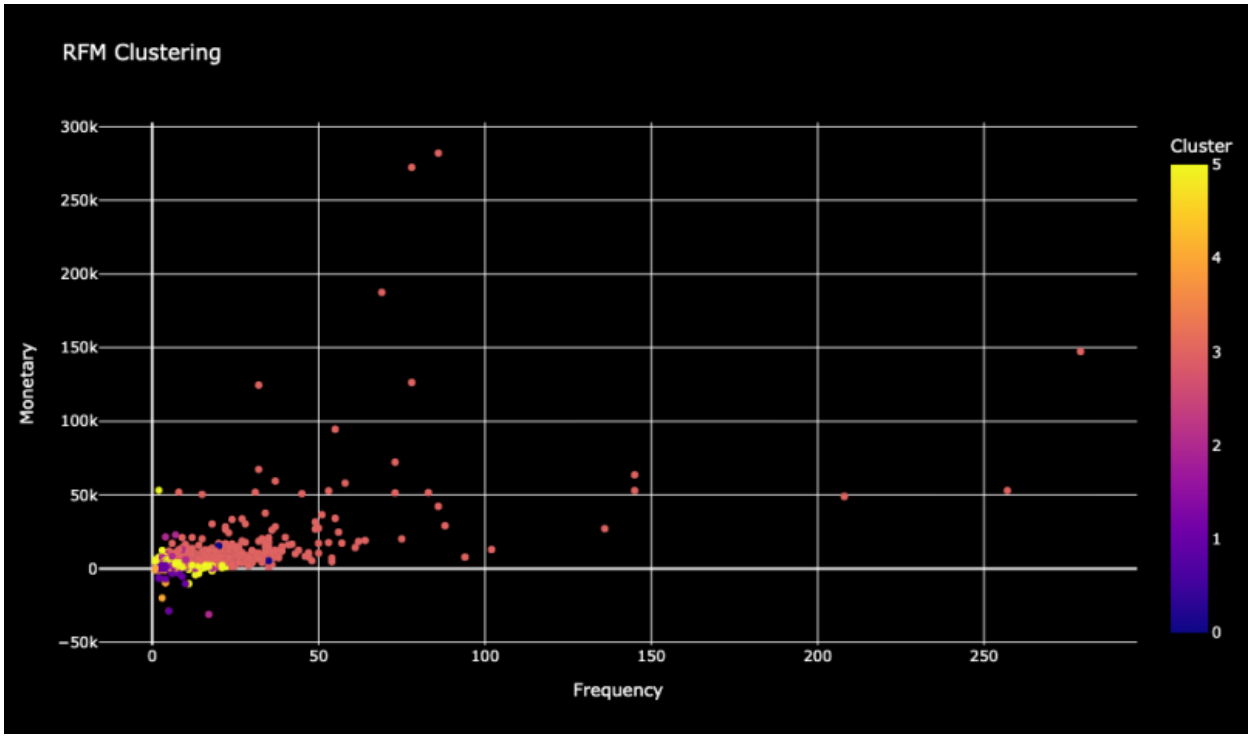

Fig 3.7.6 *RFM Clustering Frequency/Monetary*

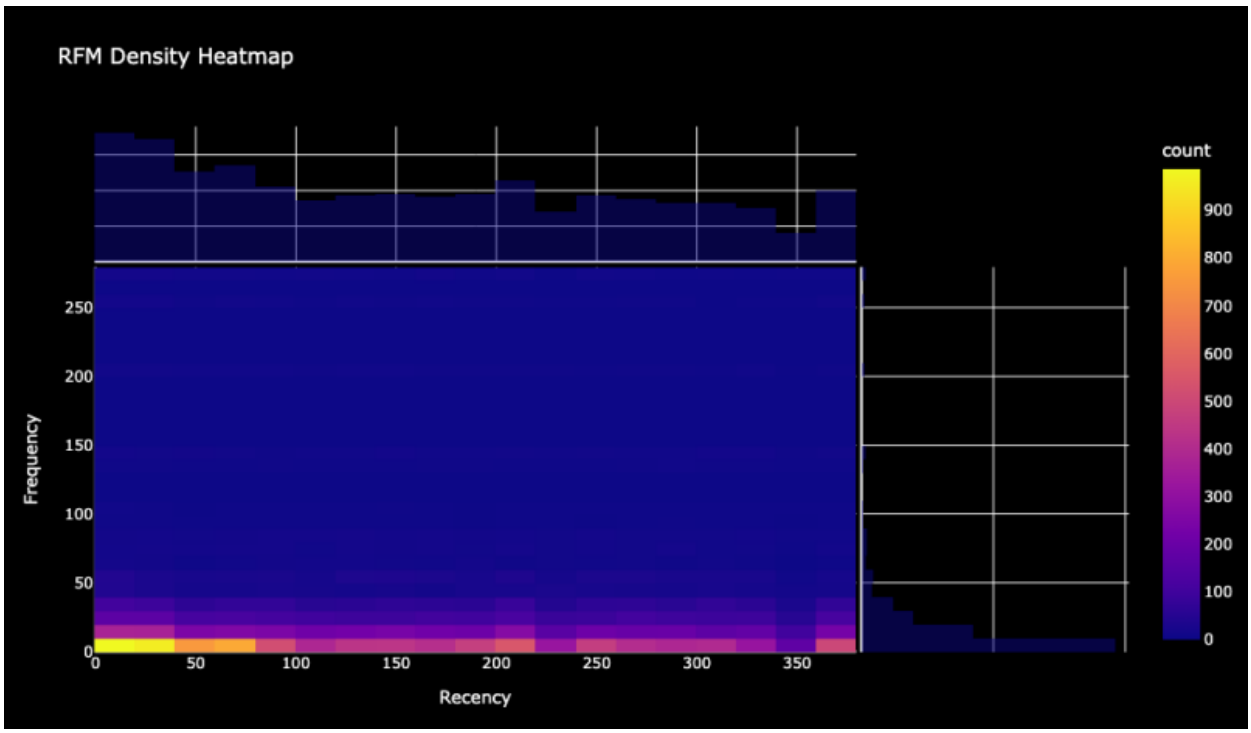

Fig 3.7.7 *RFM Density Heatmap F/R*

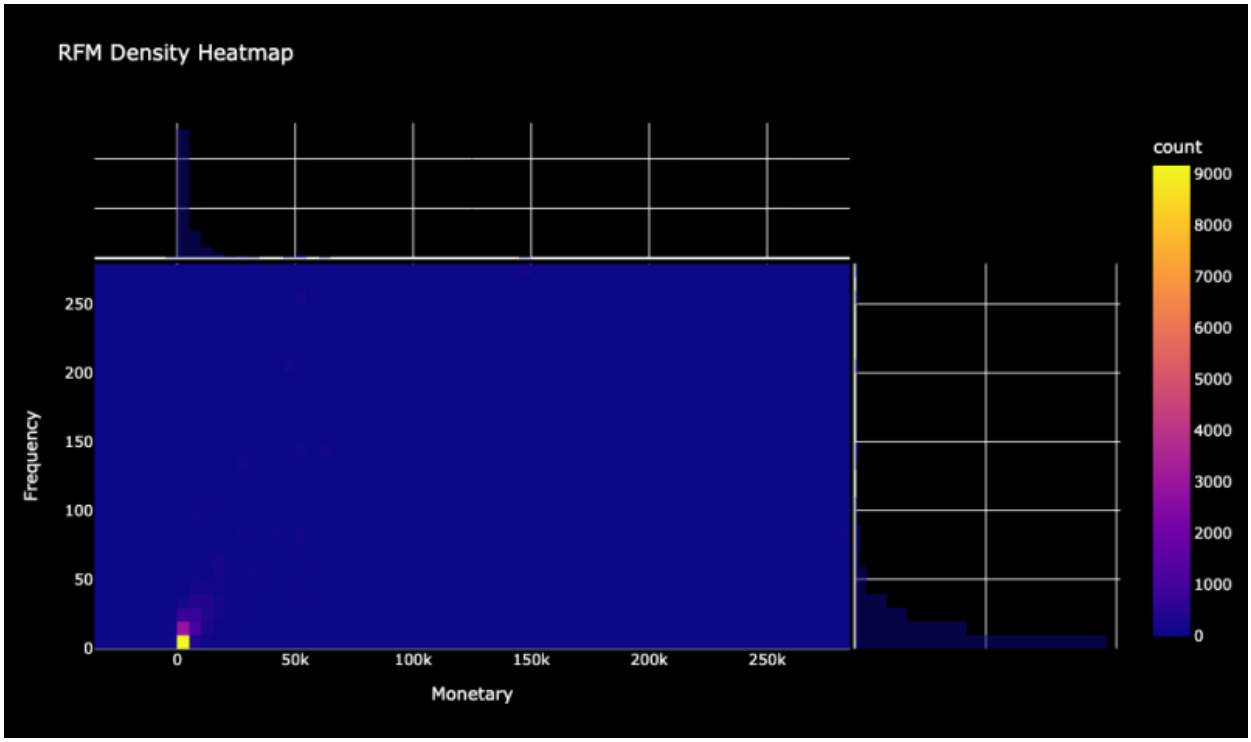

Fig 3.7.8 *RFM Density Heatmap F/M*

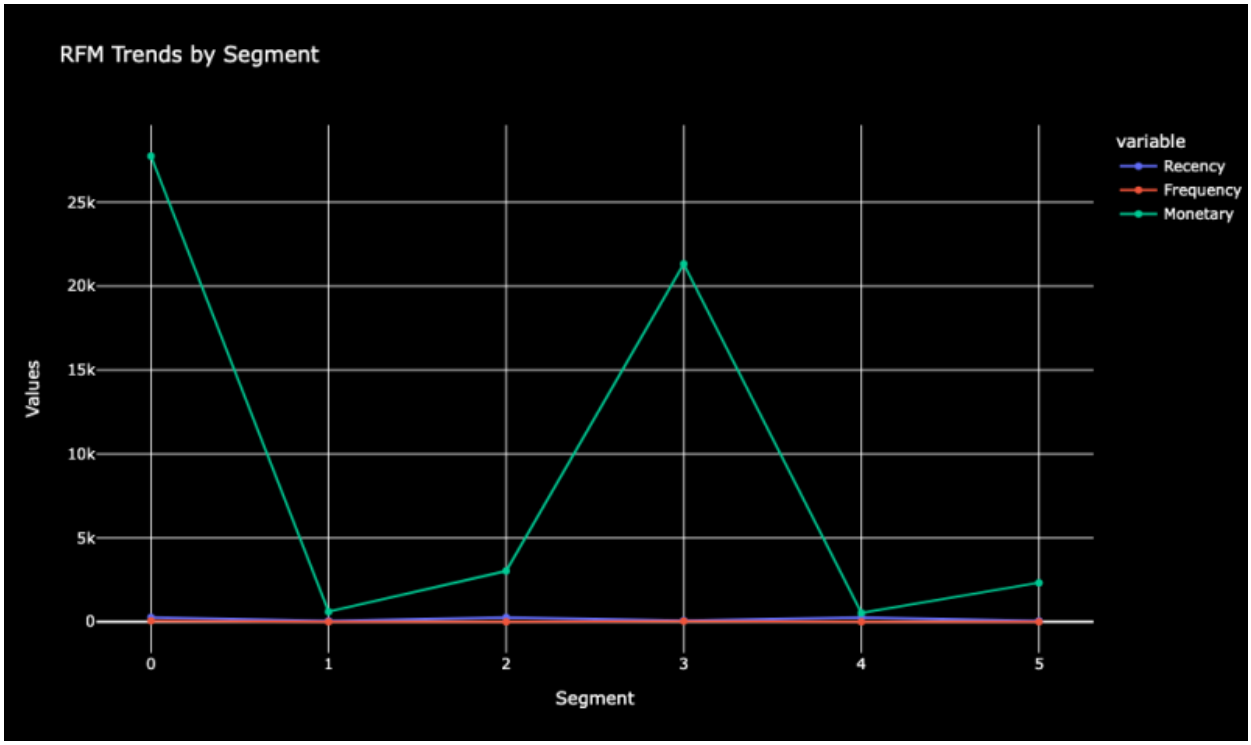

Fig 3.7.9 *RFM Trends by Segment*

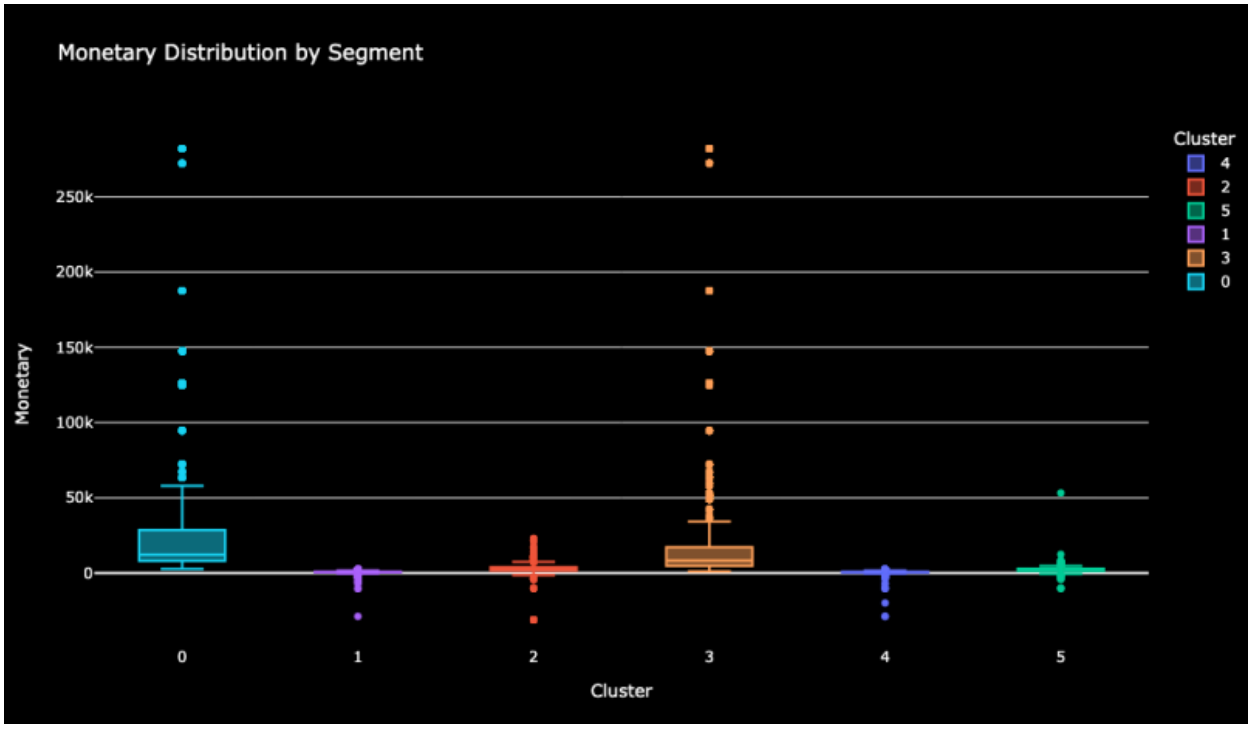

Fig 3.7.10 *Monetary Distribution by Segment*

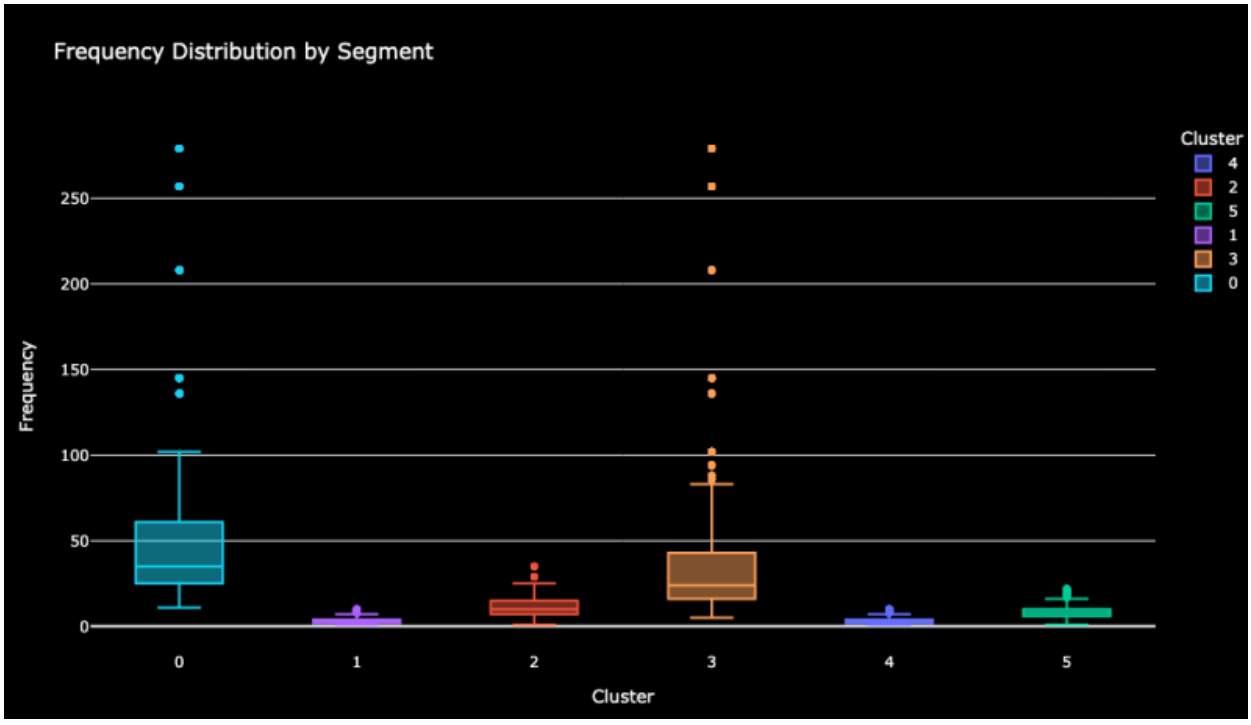

Fig 3.7.11 *Frequency Distribution by Segment*

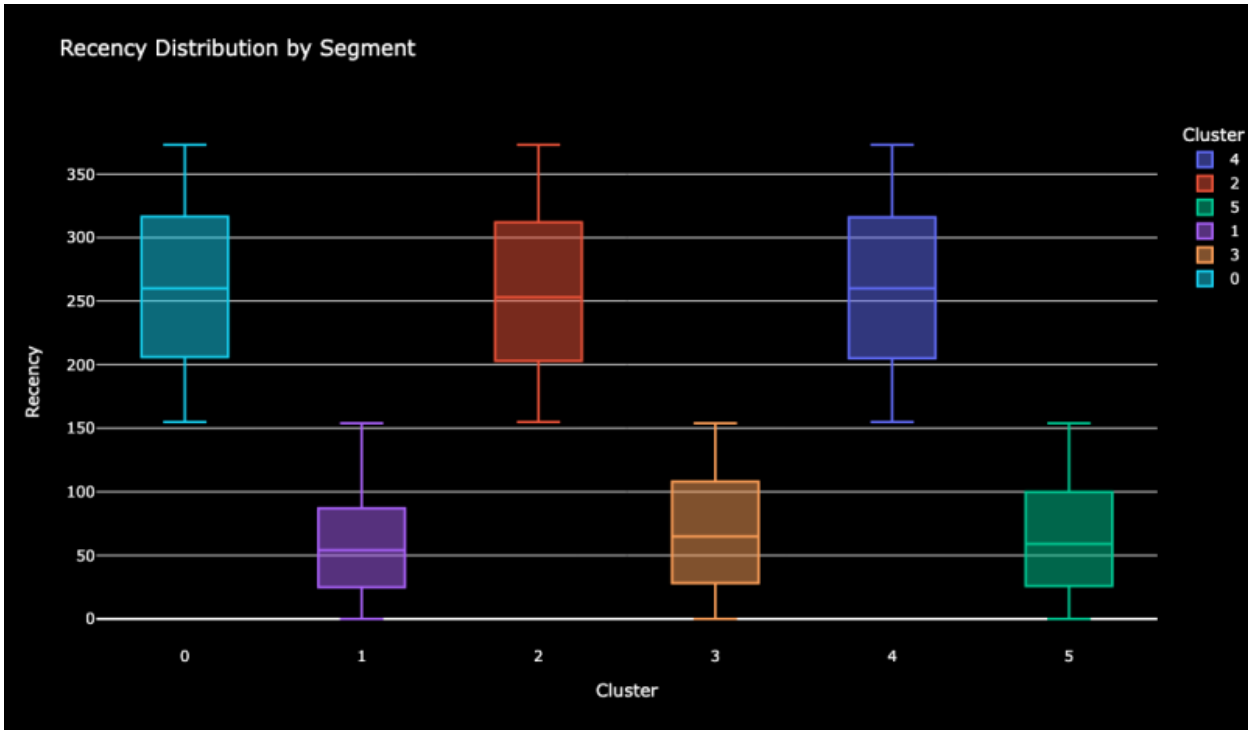

Fig 3.7.12 *Recency Distribution by Segment*

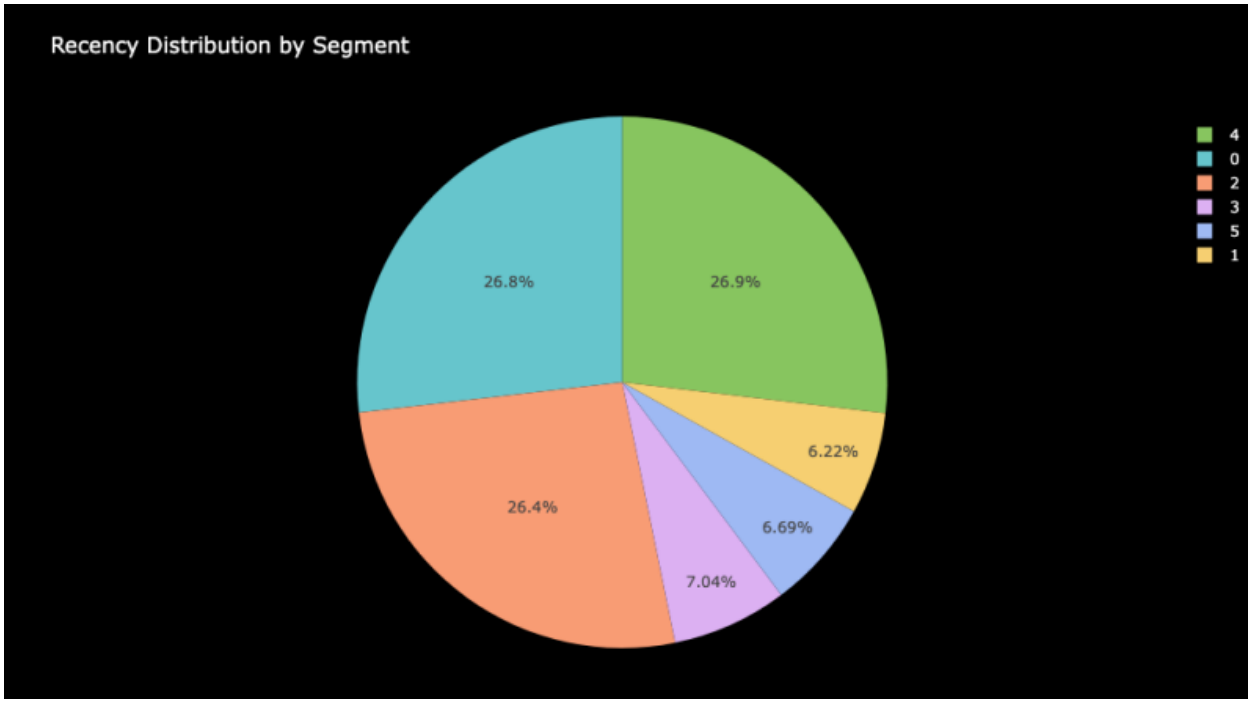

Fig 3.7.13 *Recency Distribution by Segment*

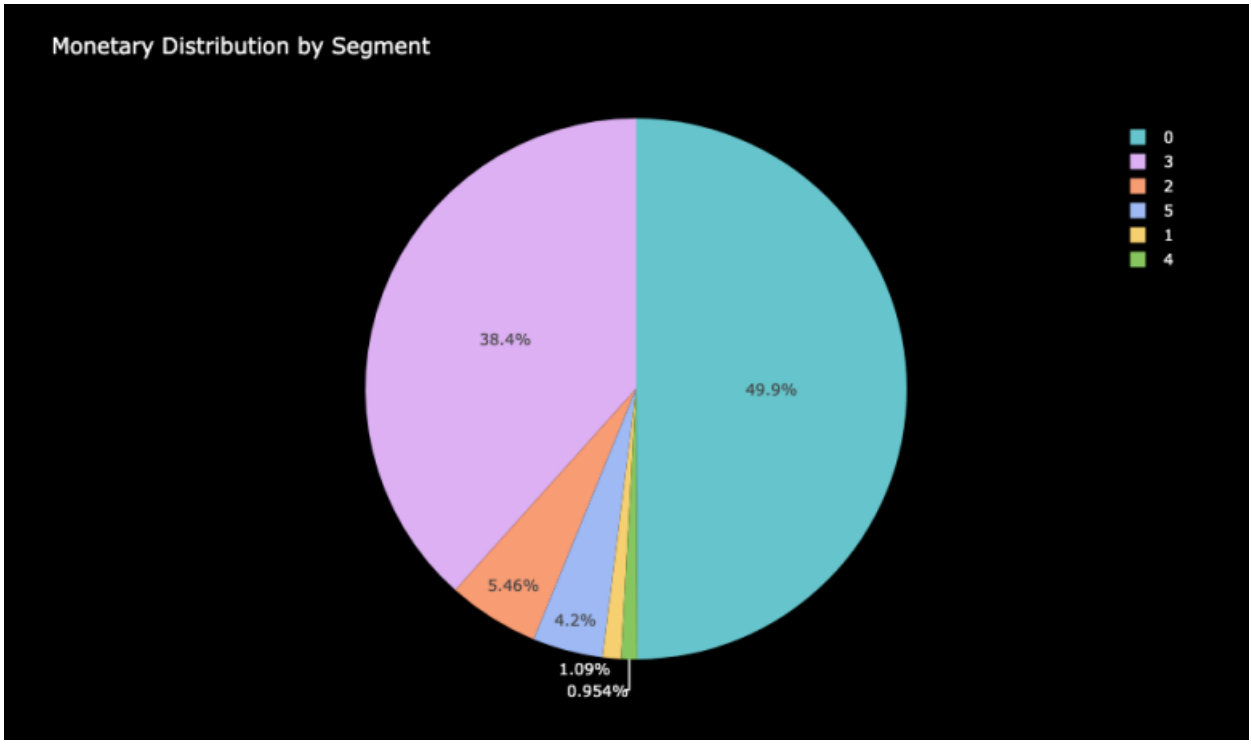

Fig 3.7.13 *Monetary Distribution by Segment*

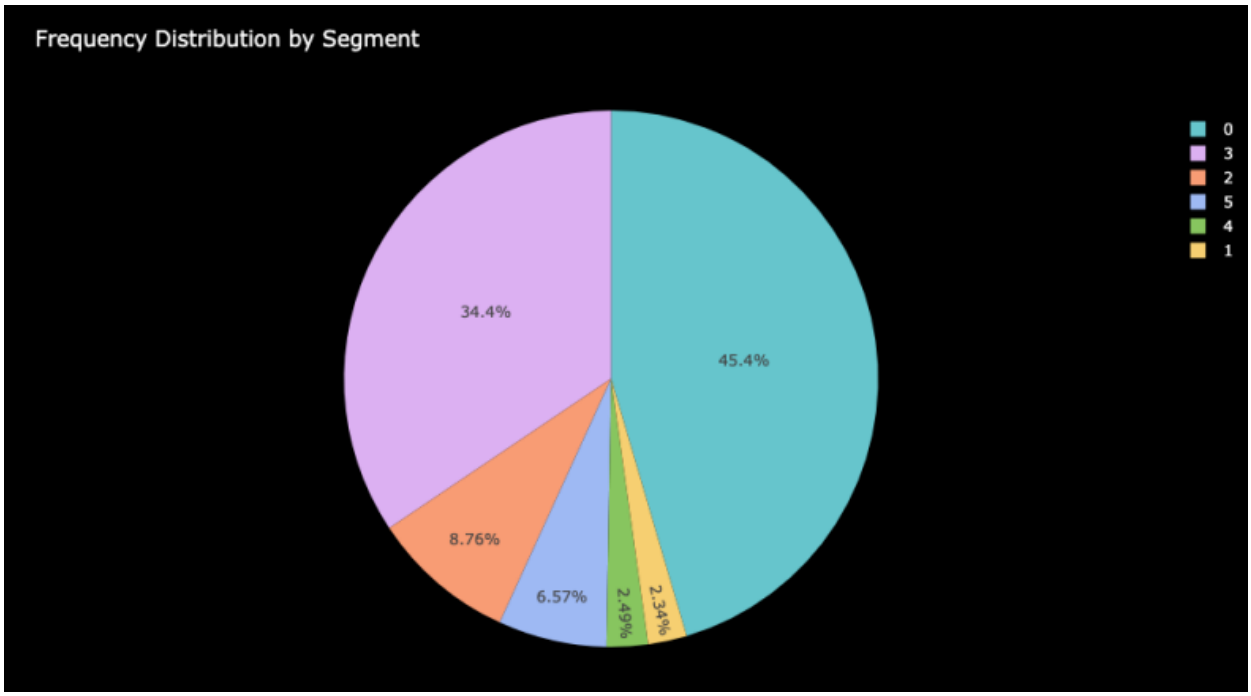

Fig 3.7.14 *Frequency Distribution by Segment*

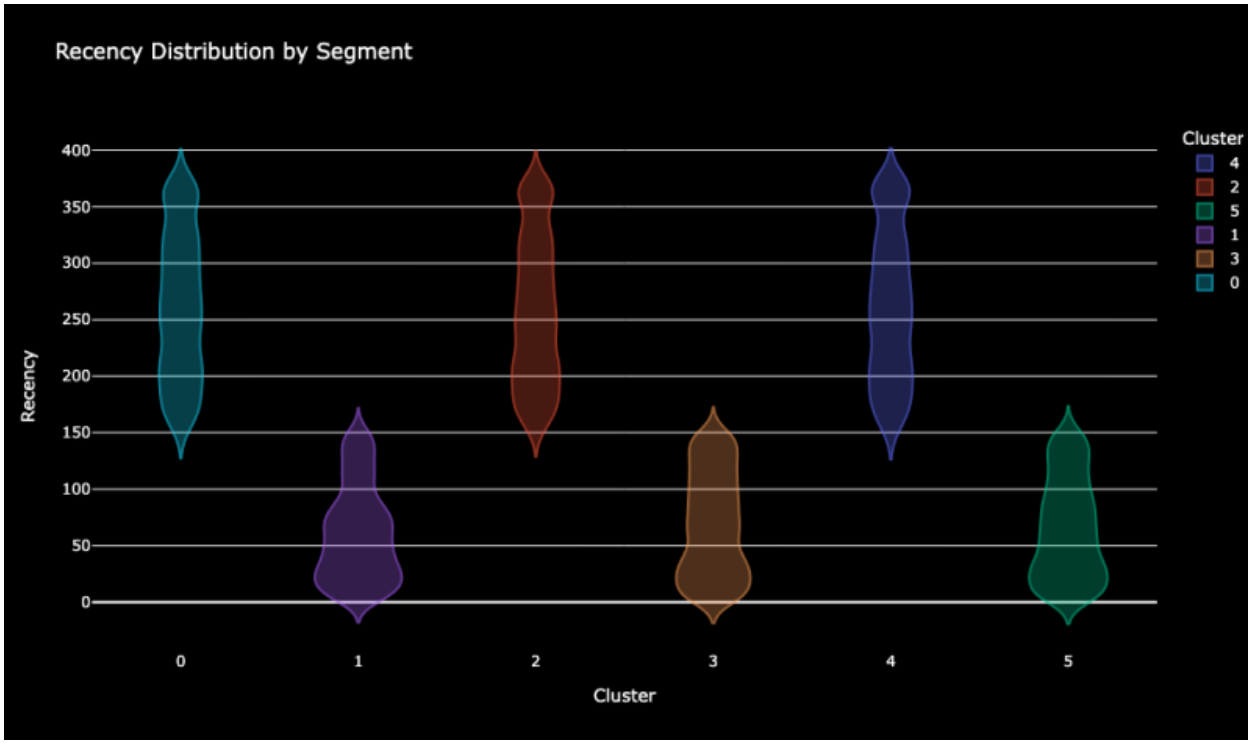

Fig 3.7.15 *RecencyDistribution by Segment*

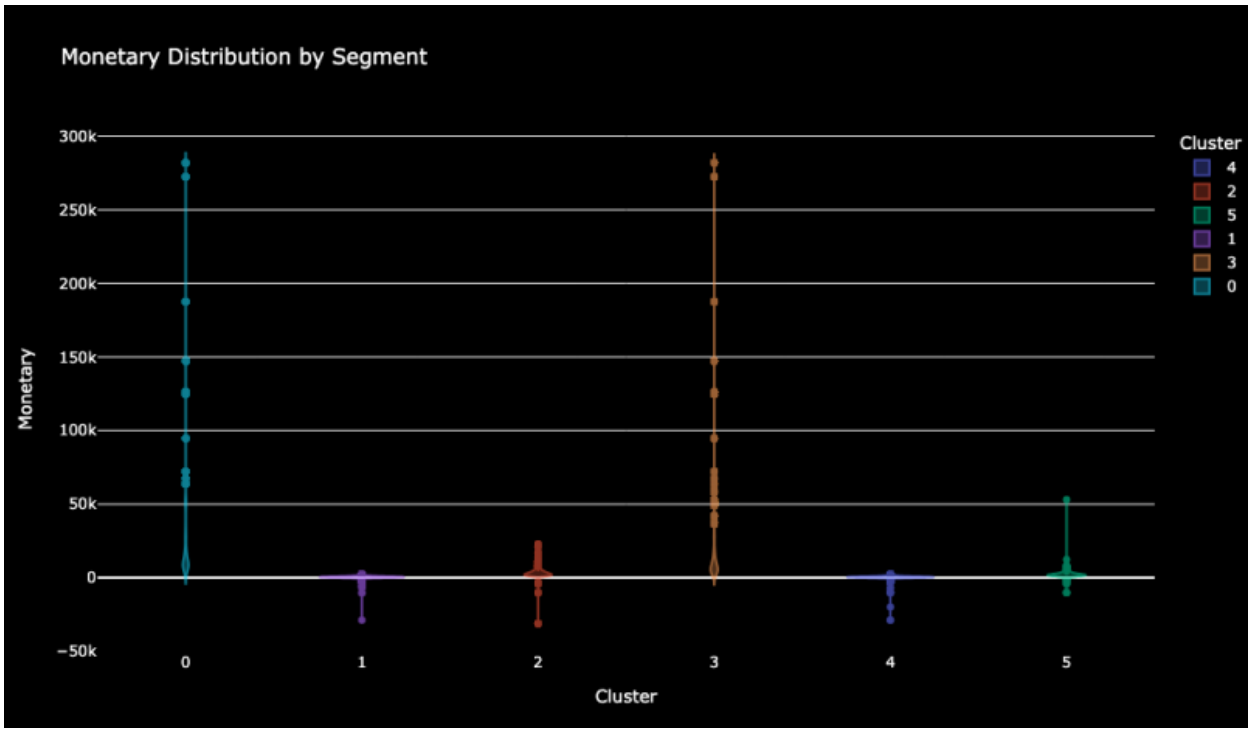

Fig 3.7.16 *Monetary Distribution by Segment*

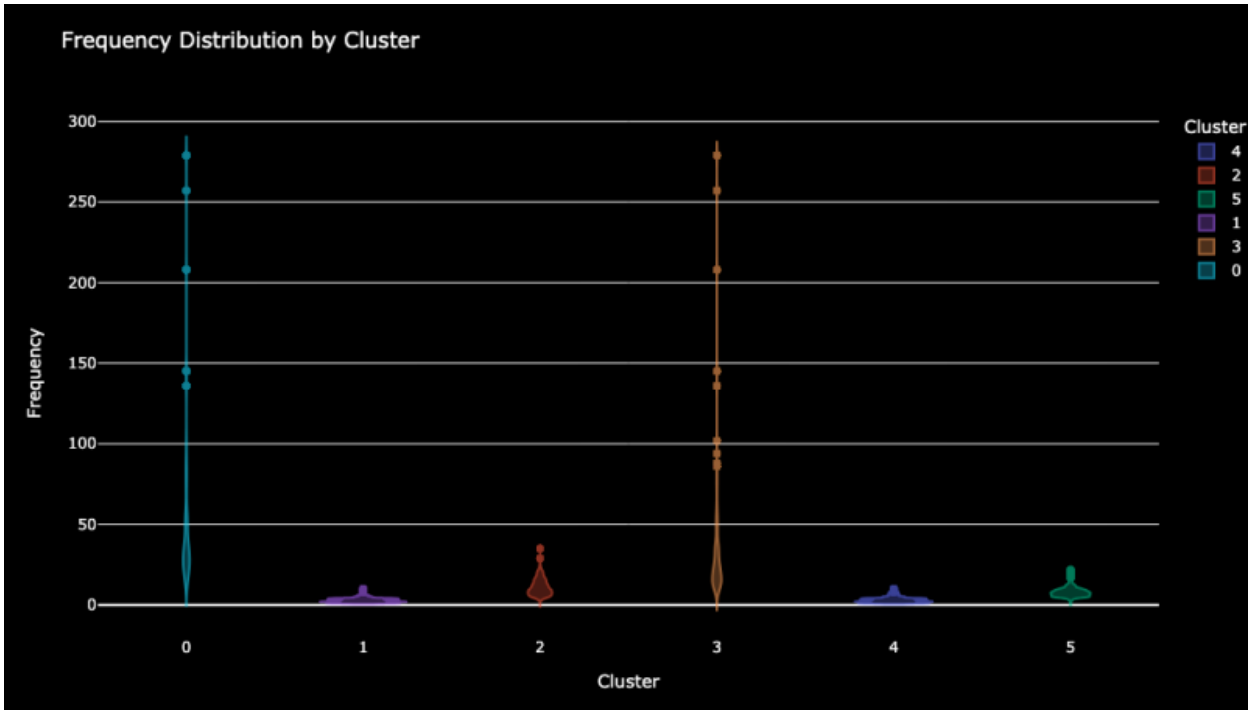

Fig 3.7.17 *Frequency Distribution by Cluster*

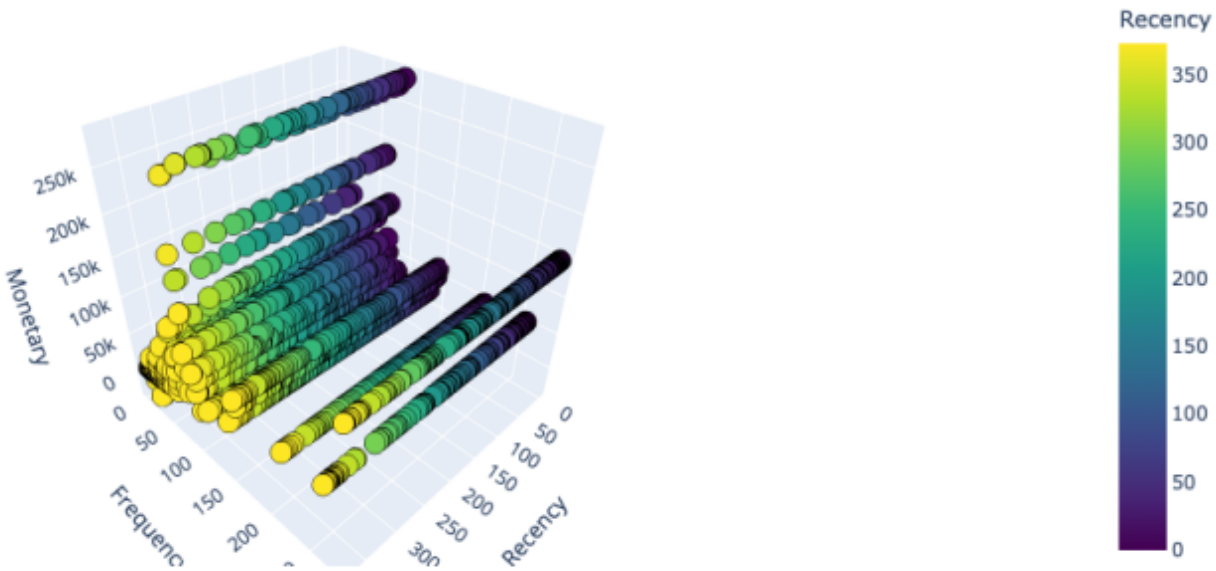

Fig 3.7.18 *RFM Segments 3D*

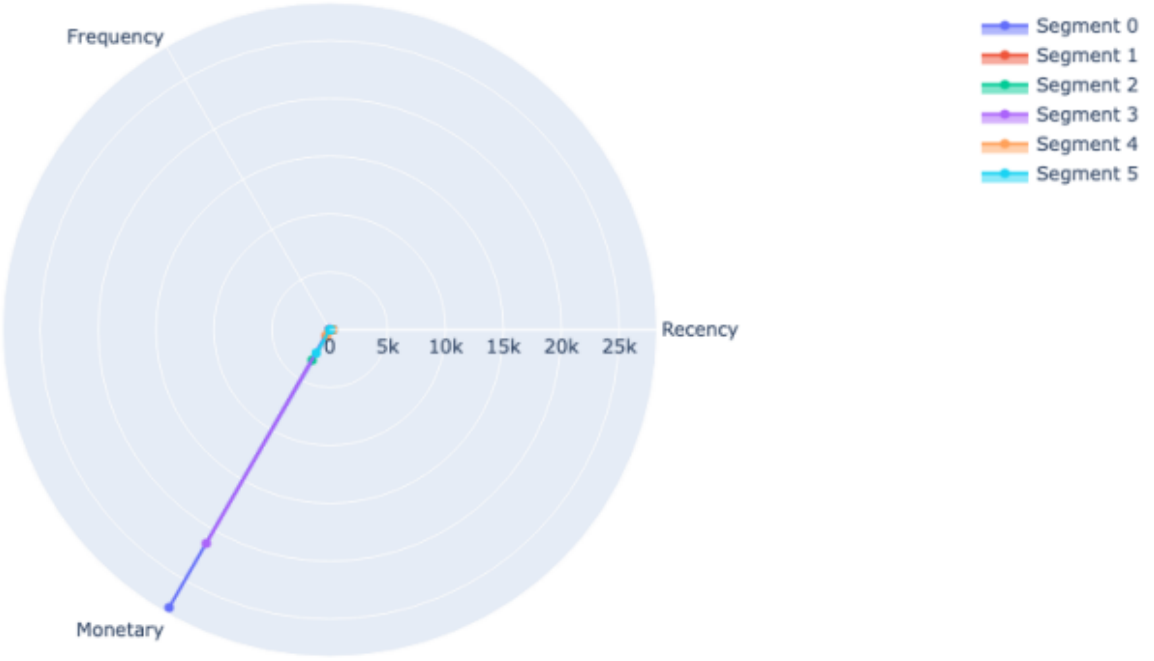

Fig 3.7.19 *RFM Segments 2D*

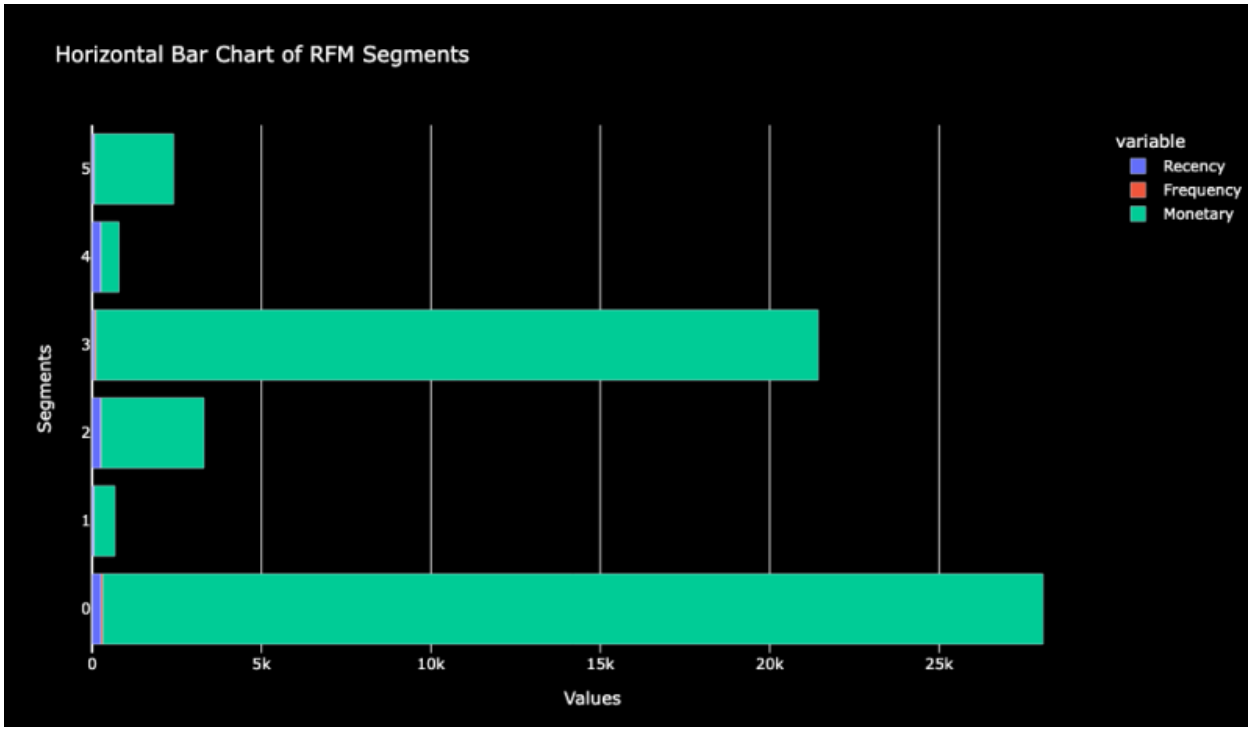

Fig 3.7.20 *Horizontal Bar RFM Segments*

4. Customer Analysis

When performing a thorough consumer analysis, the investigation included measuring the number of distinct customers as well as looking at their ordering patterns. There are 4,372 distinct customers in the dataset, all of whom may be found by using their unique CustomerID. Then, in order to identify trends in the purchasing behavior of the customers, the distribution of orders per customer was carefully examined.

The research revealed the frequency distribution of the number of orders per customer using a histogram. The bulk of clients display a particular range of order frequencies, which gives insights into the diversity of purchasing patterns.

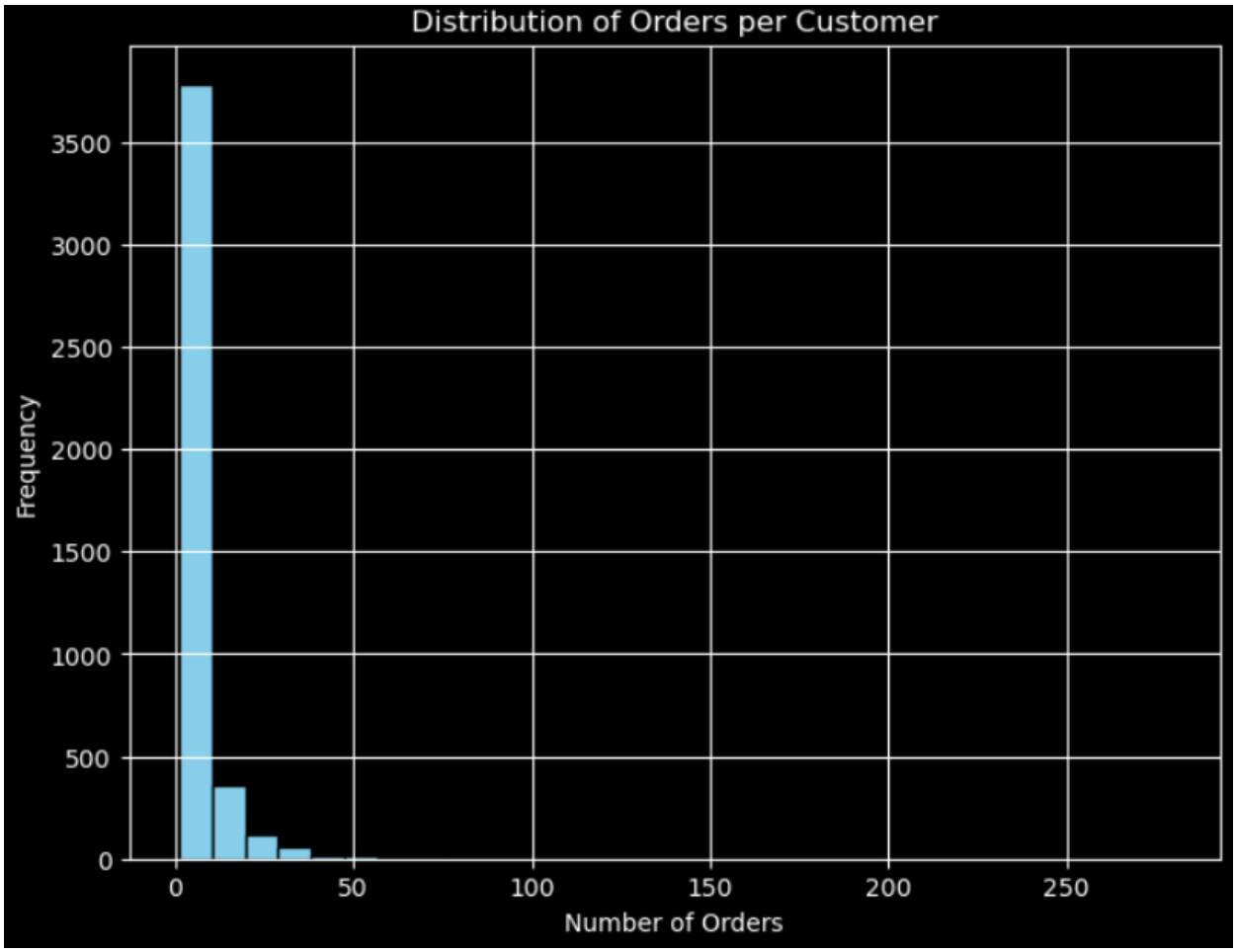

Fig 4.1 *Orders/Customer*

Identifying the top clients according to order count also provides insight into the major contributors to the total amount of transactions. The top 5 customers were determined as follows: 14911.0, 12748.0, 17841.0, 13089.0, and 15311.0. Each of these customer IDs was highlighted. Each of these consumers made a substantial contribution to the overall number of orders, exhibiting exceptionally high levels of involvement.

A bar chart was used to visually display the top 5 customers and provide a clear and straightforward picture of their relative order numbers. The concentration of purchasing activity among these high-achieving clients is highlighted by this graphical representation, which offers insightful information for tactical decision-making in marketing and customer interaction campaigns.

Fig 4.2 *Top 5 Customers*

5. Product Analysis

The product analysis explores important facets of the dataset, offering information on the most popular products, the average cost of each item, and the product category with the biggest profit margin. The most popular products were found to be the "WHITE HANGING HEART T-LIGHT HOLDER," "REGENCY CAKESTAND 3 TIER," and "JUMBO BAG RED RETROSPOT." These were the top 10 often purchased items.

The study incorporates the average product price, which is $4.61, to provide a thorough grasp of the pricing trends. A boxplot of the product price distribution offers an additional perspective of the dataset's key tendencies and variability.

Furthermore, the product category that was making the most money was shown to be "DOTCOM POSTAGE." This crucial realization is helpful for managing inventories and making strategic decisions. Bar charts were used to illustrate the top 10 most commonly purchased products and the top 10 product categories by revenue in order to improve the visual representation of these findings. Finding trends, patterns, and chances for marketing plans and inventory optimization is made easier with the help of these visualizations.

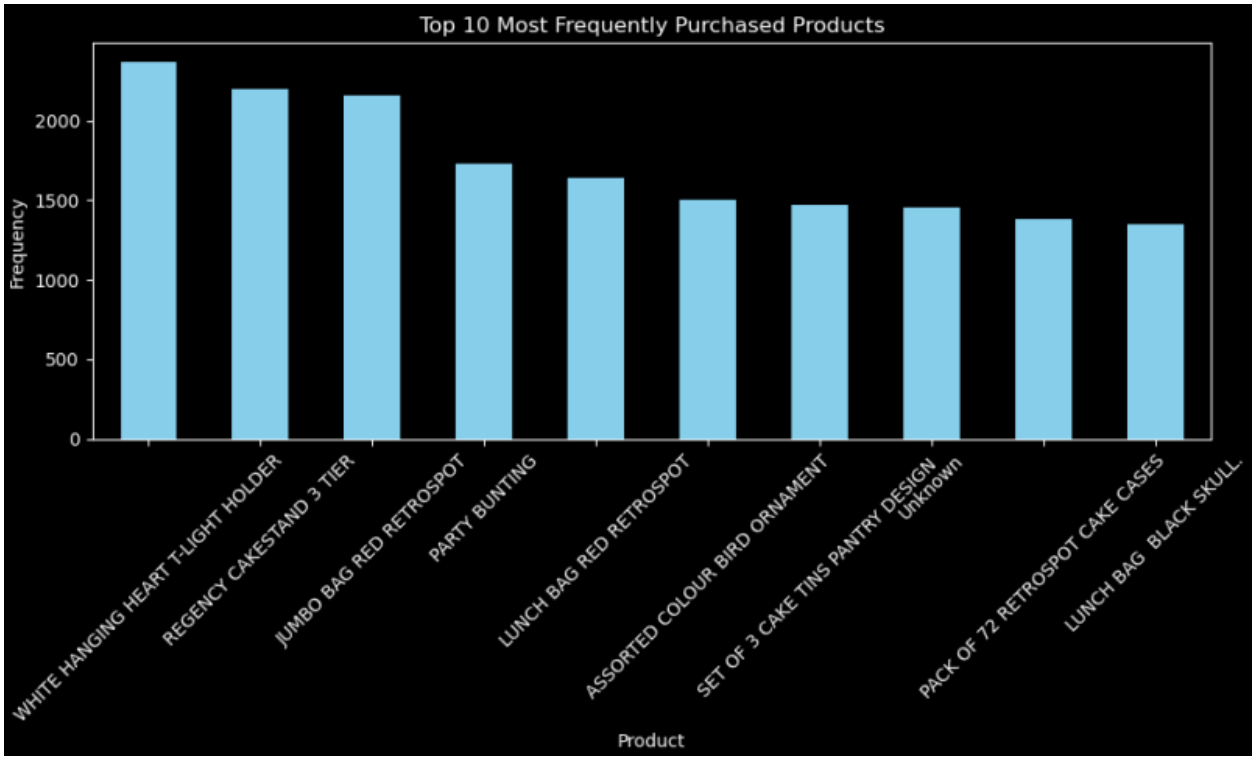

Fig 5.1 *Top 10 Purchase*

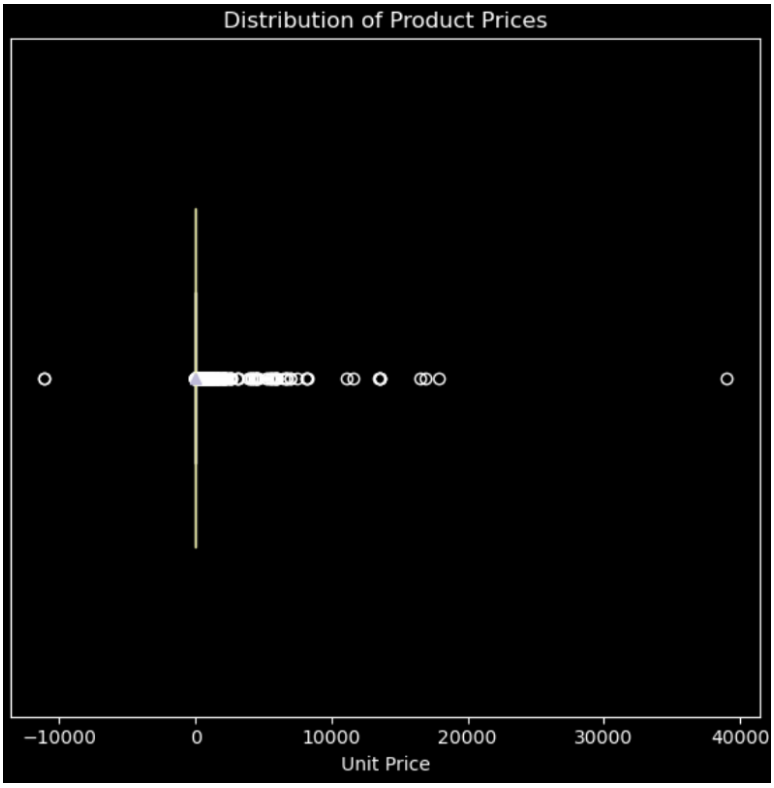

Fig 5.2 *Product Price Distribution*

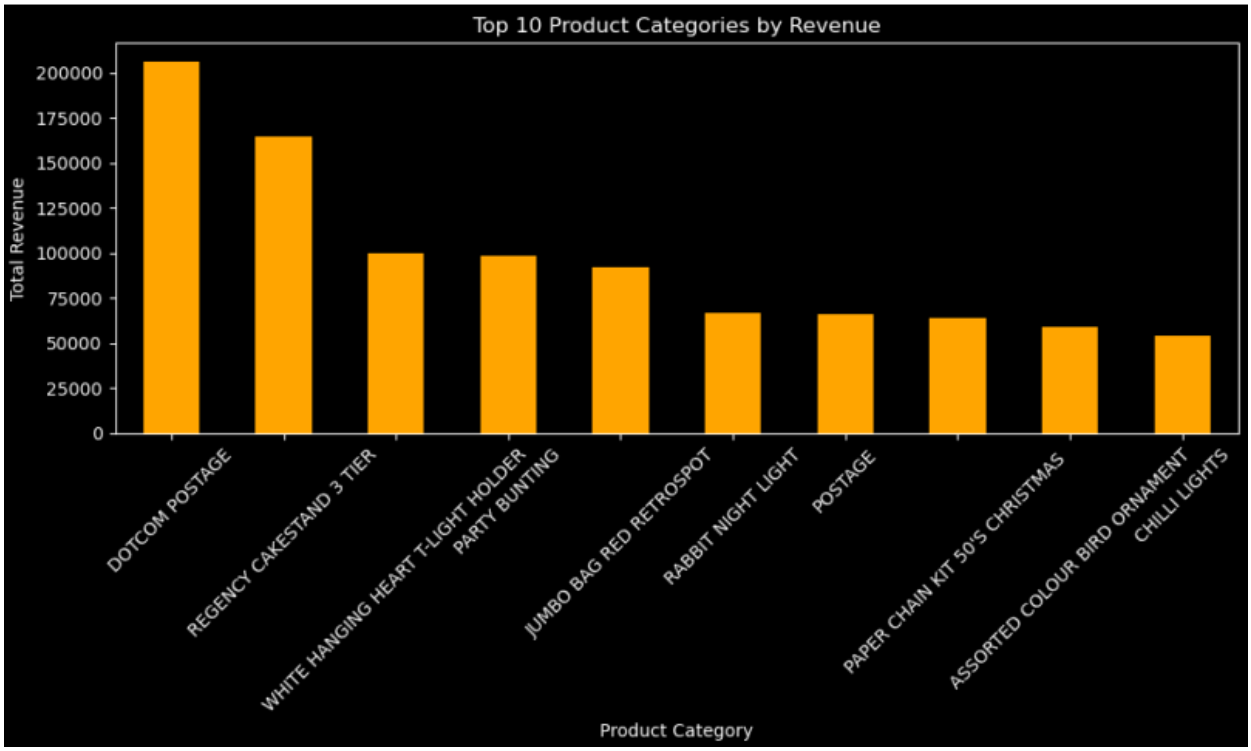

Fig 5.3 *Top 10 Product Categories/Revenue*

6. Time Analysis

The dataset's time analysis offers insightful information on seasonal trends, order processing times, and temporal patterns. The 'InvoiceDate' column was converted to a DateTime format so that different features of time-related data could be investigated. According to the analysis, most orders are placed on Thursdays, and most orders are placed at noon on those days of the week. These results provide useful information for strategically allocating resources and staffing during periods of high demand. Additionally, the average order processing time which is computed as 0 days indicates that the dataset's order fulfillment procedure is effective.

Monthly trends are included in the temporal exploration, which shows changes in order counts over the course of the dataset. A line plot was used to illustrate seasonal trends and highlight how order quantities varied from month to month. These visuals can help with inventory management, operational planning, marketing strategy, and the analysis of customer behavior patterns over time.

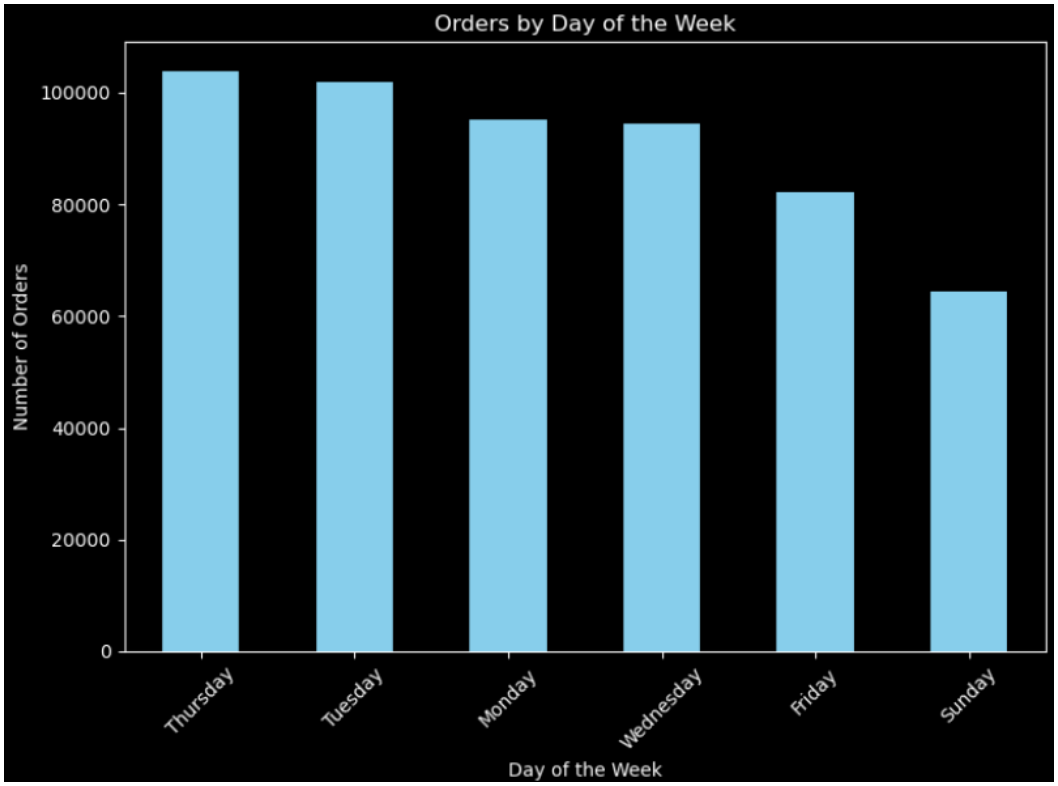

Fig 6.1 *Order/Day*

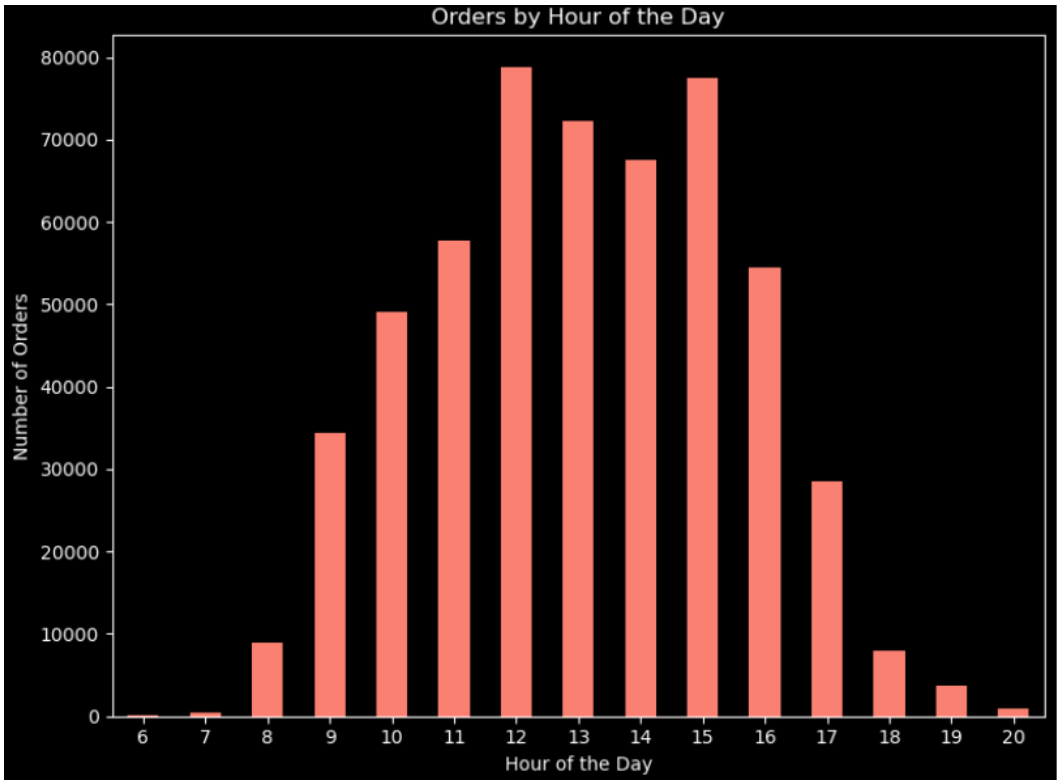

Fig 6.2 *Order/Hour*

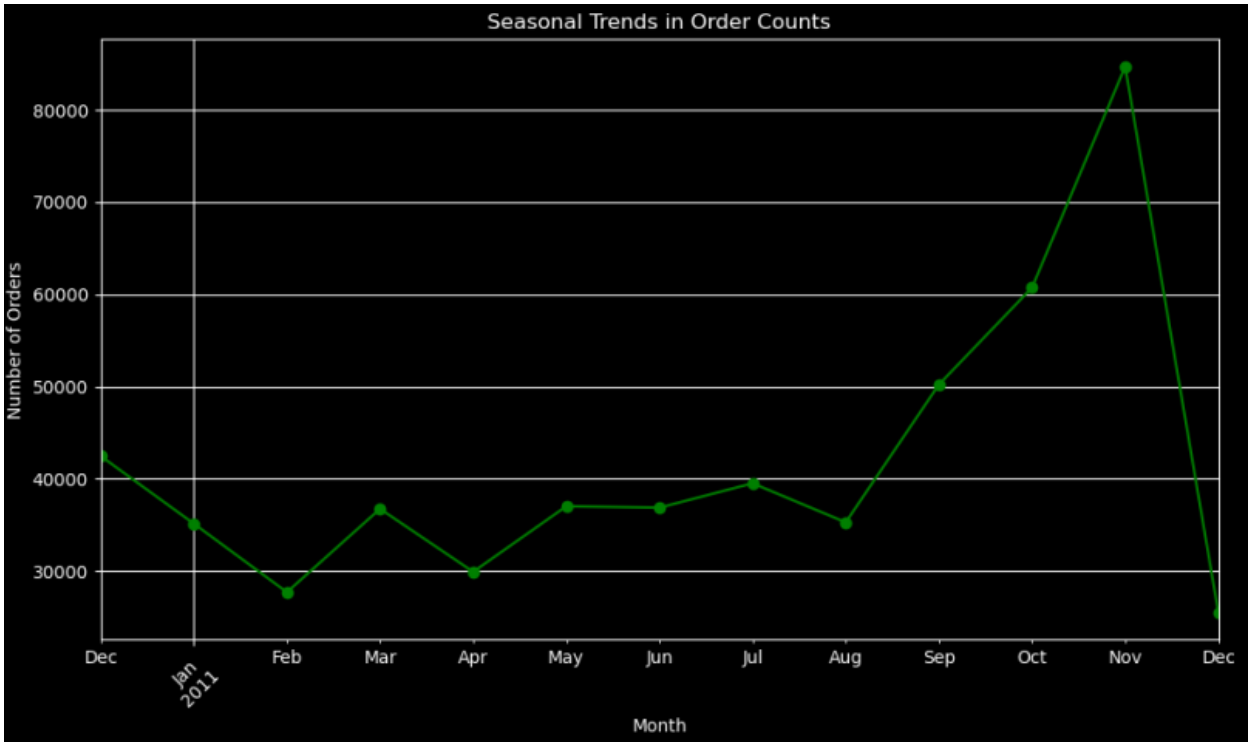

Fig 6.3 *Seasonal Trends*

7. Geographical Analysis

The dataset's geographical analysis looks at possible relationships between the average order value and the country of the client, as well as how customers are distributed throughout other nations.

The United Kingdom dominates the order count, with Germany, France, Spain, EIRE, and the United Kingdom making up the top five countries.

The ANOVA test was used in a study to look at the relationship between the average order value and the customer's country. The test produced a 3.58 p-value, which is incredibly low and indicates significant statistical significance. This finding implies that the average order values of various nations vary significantly.

The results of the ANOVA test indicate that there is a relationship between the average order value and the customer's nationality. The customer's geographic location has an influence on the observed variation in average order values among countries; this variation is not random. This discovery bears significance for pricing optimization, marketing tactics, and customization of business methods according to the varied spending patterns or buying tendencies displayed by clients in various geographic areas.

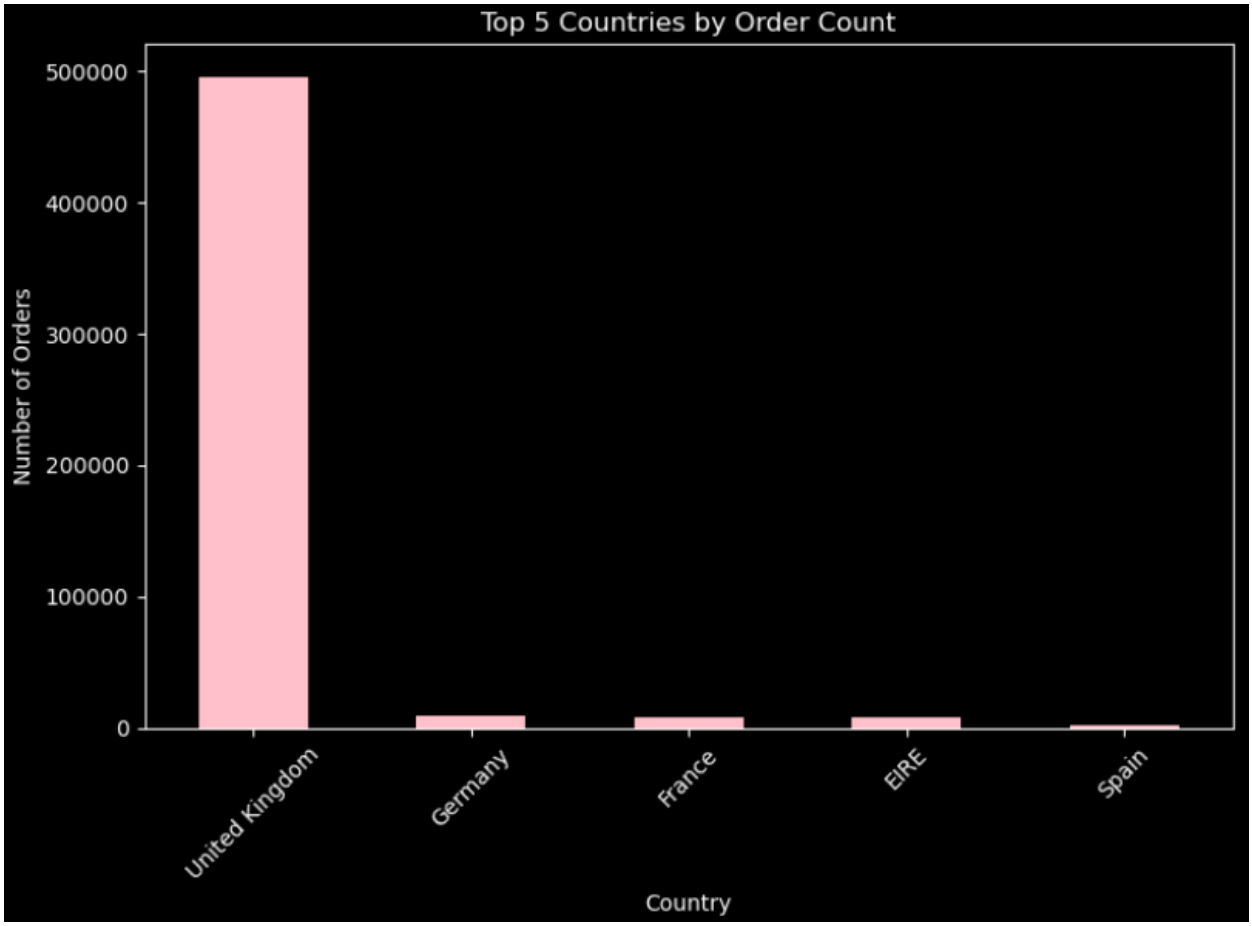

Fig 7.1 *Top 5 Countries*

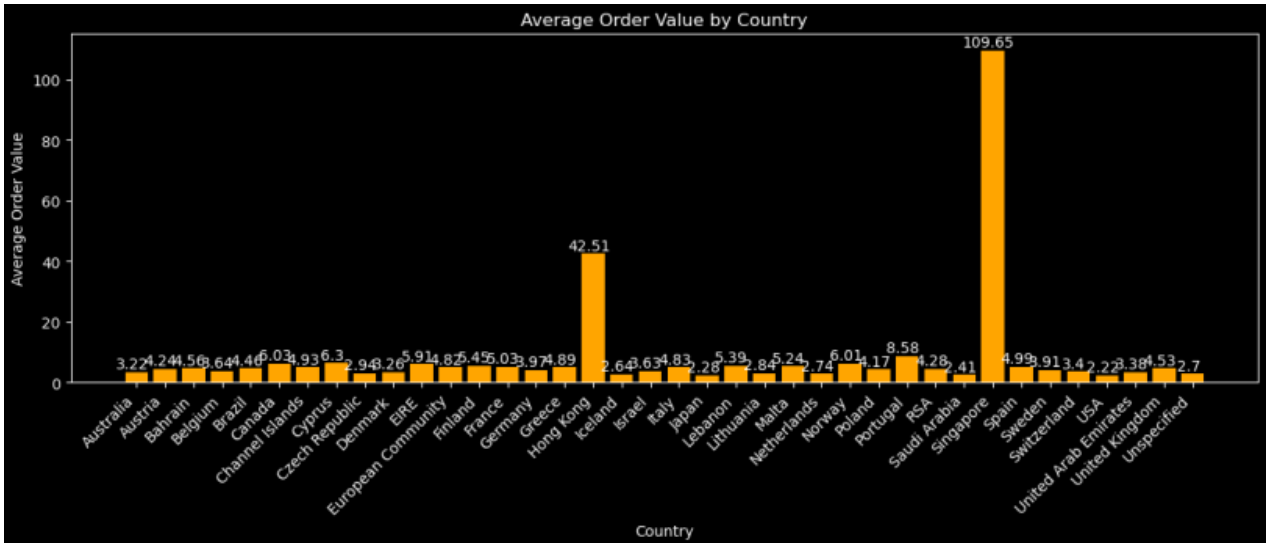

Fig 7.2 *Average Order Value/Country*

8. Payment Analysis

The analysis of payment methods reveals the most popular payment choices made by clients and determines whether a clear correlation exists between the order amounts and the selected payment method. A new column was added because the original dataset did not include payment method information. Different payment methods, including cash, cards, cryptocurrency, mobile wallets, and bank transfers, were randomly assigned to each transaction. With differing frequencies across transactions, the most popular payment methods were discovered to be Cash, Bank Transfer, Mobile Wallet, Cryptocurrency, and Card.

The mean order amount for each payment method was determined in order to look at possible connections between payment methods and order amounts. The ensuing ANOVA test produced a p-value of 0.12, suggesting that there was insufficient data to strongly refute the null hypothesis. This implies that depending on the selected payment option, there might not be any appreciable variations in the mean order amounts. Additional information about the distribution of payment method frequencies and the possible relationship between payment methods and order quantities can be gained from the data's visual representation in bar charts and box plots. Even though some payment methods might be more common, the statistical analysis suggests that, based on the data at hand, the payment method selected might not have a significant impact on the order amount.

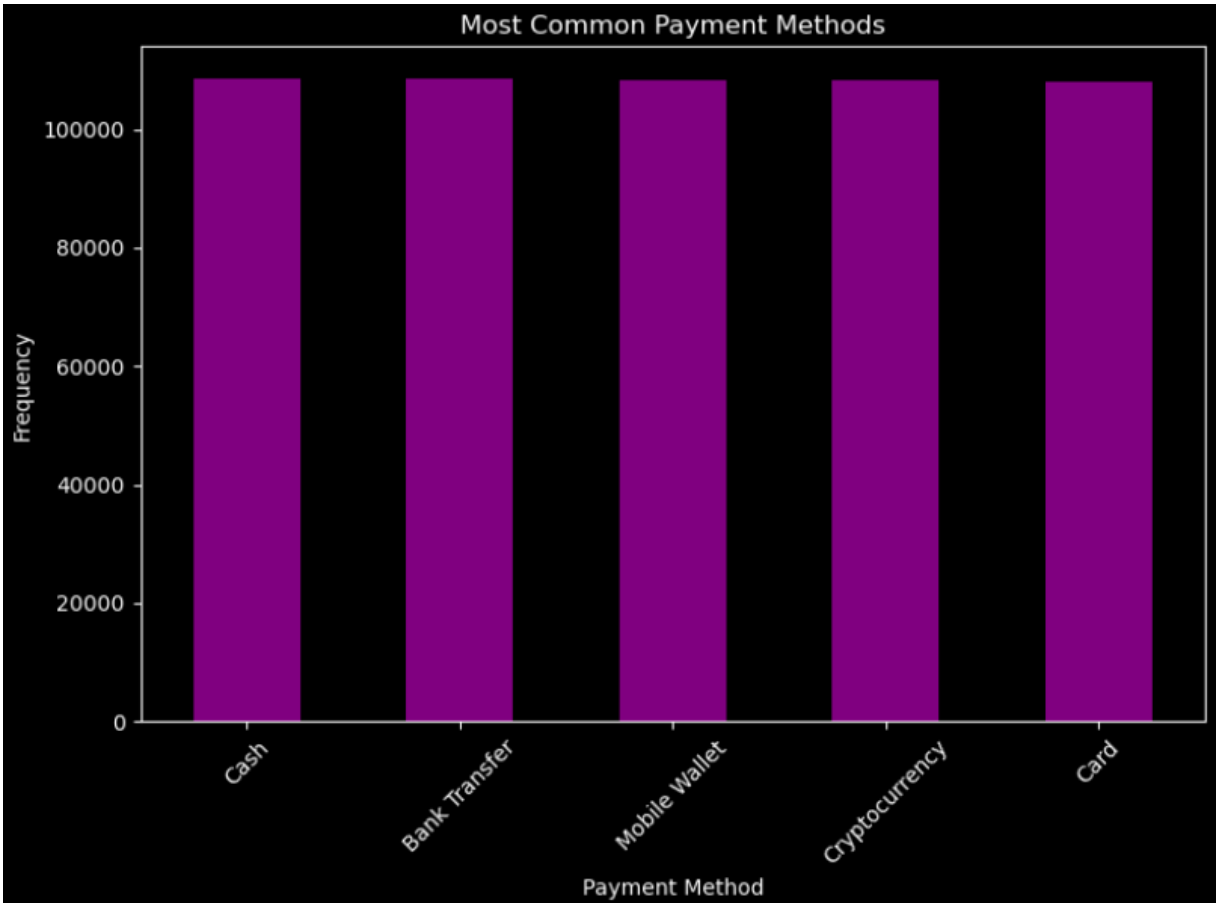

Fig 8.1 *Payment Methods*

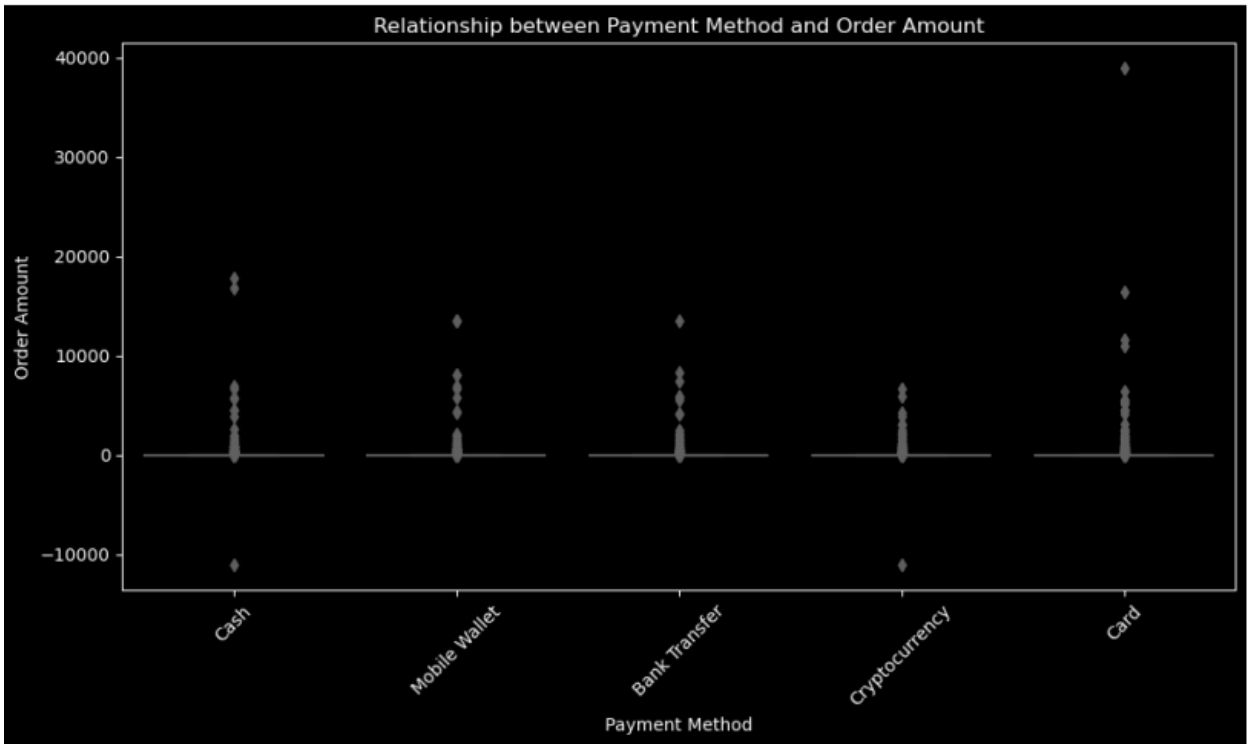

Fig 8.2 *Payment Method vs Order Count*

9. Customer Behavior

The goal of the customer behavior analysis was to determine the average length of time that consumers stay active, which was determined by measuring the period between their first and last transactions. To determine the activity duration in days, the dataset was sorted by client and the minimum and maximum invoice dates were extracted.

It was discovered that clients stay active for an average of about 133.39 days. This statistic helps businesses understand the normal period of customer activity by offering insightful information about the total engagement and retention of customers.

A histogram showing the frequency of customers throughout various activity length ranges was created to visualize the distribution of client activity durations. The distribution of client activity durations and patterns in consumer behavior can be better understood with the help of this depiction.

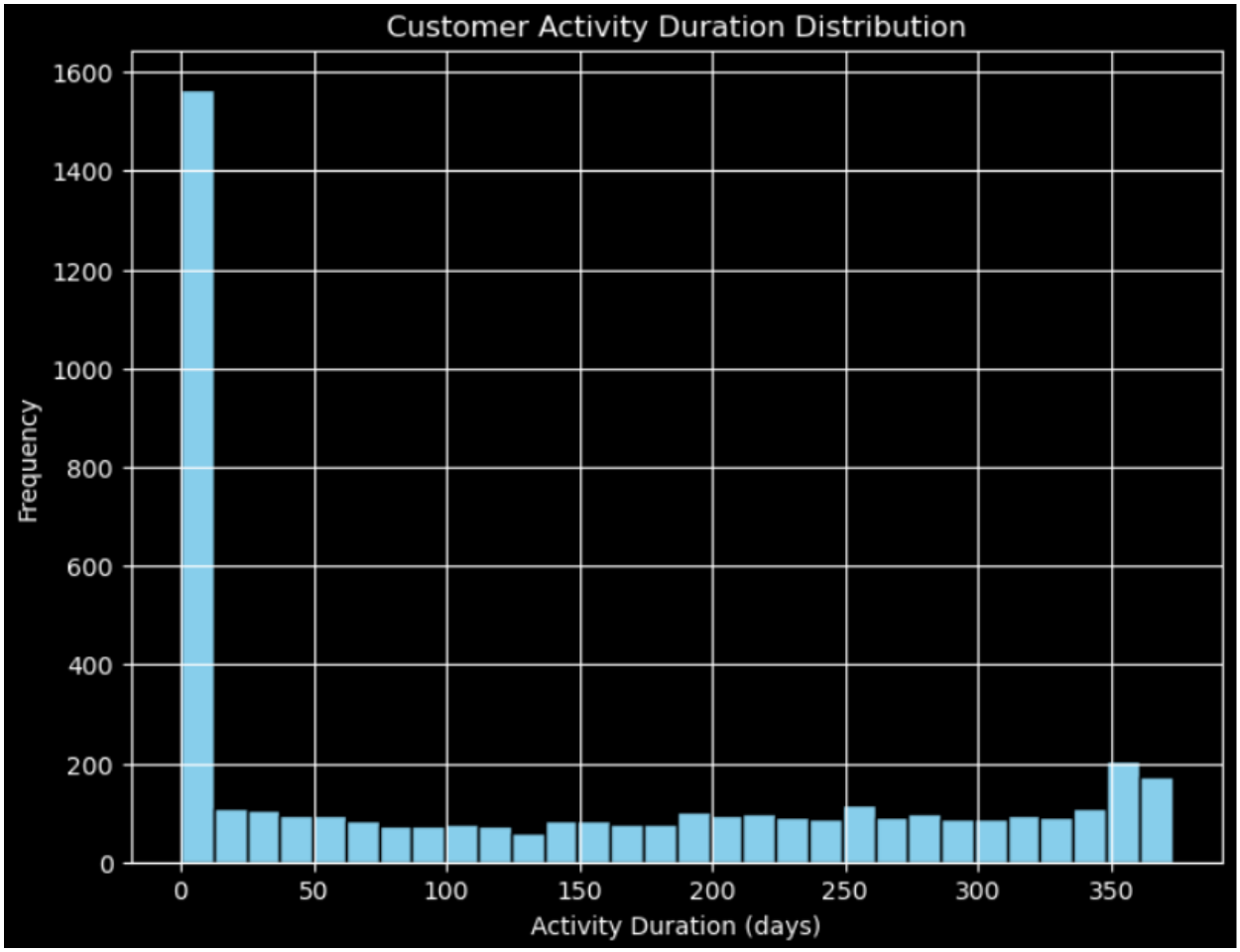

Fig 9.1 *Customer Activity*

10. Returns and Refunds

Two new columns, "ReturnRefund" and "ProductDescription," were added to the dataset to aid in the study of returns and refunds by making it easier to look at client interactions. Simulated data was added to the dataset to improve it, including details on product classifications and return/refund status.

Based on the results, it was found that 33.23% of the dataset's orders had refunds or returns. For the purpose of controlling customer happiness and streamlining operational procedures, this statistic offers information on how frequently customers express dissatisfaction or encounter problems with products.

A chi-square test was used to investigate the connection between product categories and the possibility of returns. The percentage of returns and refunds for each product category was shown in the contingency table. The chi-square test did, however, produce a p-value that was very near to 1, suggesting that there was no meaningful correlation between the types of products and the chance of returns. As a result, it doesn't seem that different product categories significantly affect the chance of returns, according to the data that is currently available.

The distribution of transactions involving returns or refunds is further highlighted by the return/refund % depiction. The percentage of various transaction types is displayed in a bar chart, giving a clear picture of how common returns and refunds are in the dataset.

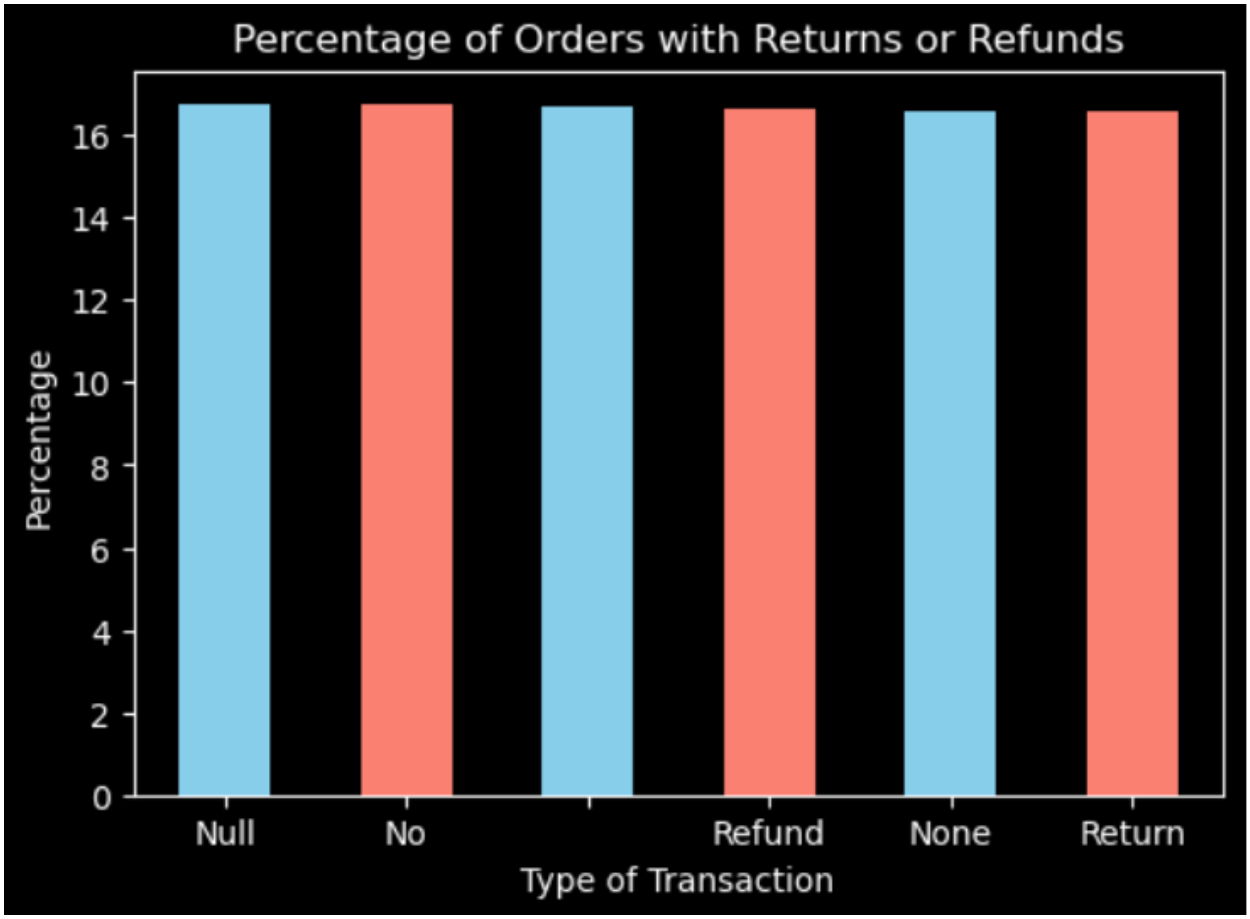

Fig 10.1 *Refunds/Returns*

11. Profitability Analysis

A thorough analysis of the business's financial performance reveals a total profit for the specified period of $9,747,747.93. This all-inclusive statistic is a critical tool that provides information on the general health and performance of the company's business activities.

The computation of profit margins provides useful data for a more thorough knowledge of profitability at the product level. The profit margin, which is expressed as a percentage of profit about the total cost, sheds light on how profitable each product is. The five products with the strongest profit margins, as determined by the average profit margin for each product, are noteworthy. Profit margins for these exceptional products range from 57.94% to 58.09%. They are the Vintage clock, Designer chair, Modern lamp, Stylish backpack, and bottle.

These highlight goods that make a substantial financial contribution to the business. As such, it becomes imperative to take a focused approach to improving sales and marketing tactics for

these high-margin products. With this strategy, the business can improve its financial results and solidify its position in the market.

12. Customer Satisfaction

Since the database did not contain explicit consumer feedback or ratings, a simulated analysis was performed utilizing extra remarks to determine sentiment. In order to simulate possible consumer feedback scenarios, these comments—which represent a range of client experiences—were randomly assigned to dataset items.

A significant amount of positive sentiments—303,482 in total—were found in the sentiment analysis, suggesting that customer contentment is common. On the other hand, negative sentiments representing regions of discontent or worries were present in 119,429 cases. Furthermore, 118,998 cases had neutral sentiments assigned to them, indicating an ambivalent or neutral position.

Pie charts are a useful tool for visualizing various sentiments since they offer a thorough picture of the overall sentiment distribution. Positive sentiments predominate, which helps to raise the total sentiment score. This study provides insightful information about consumer sentiment that helps the business pinpoint areas for development and efficiently handle customer complaints.

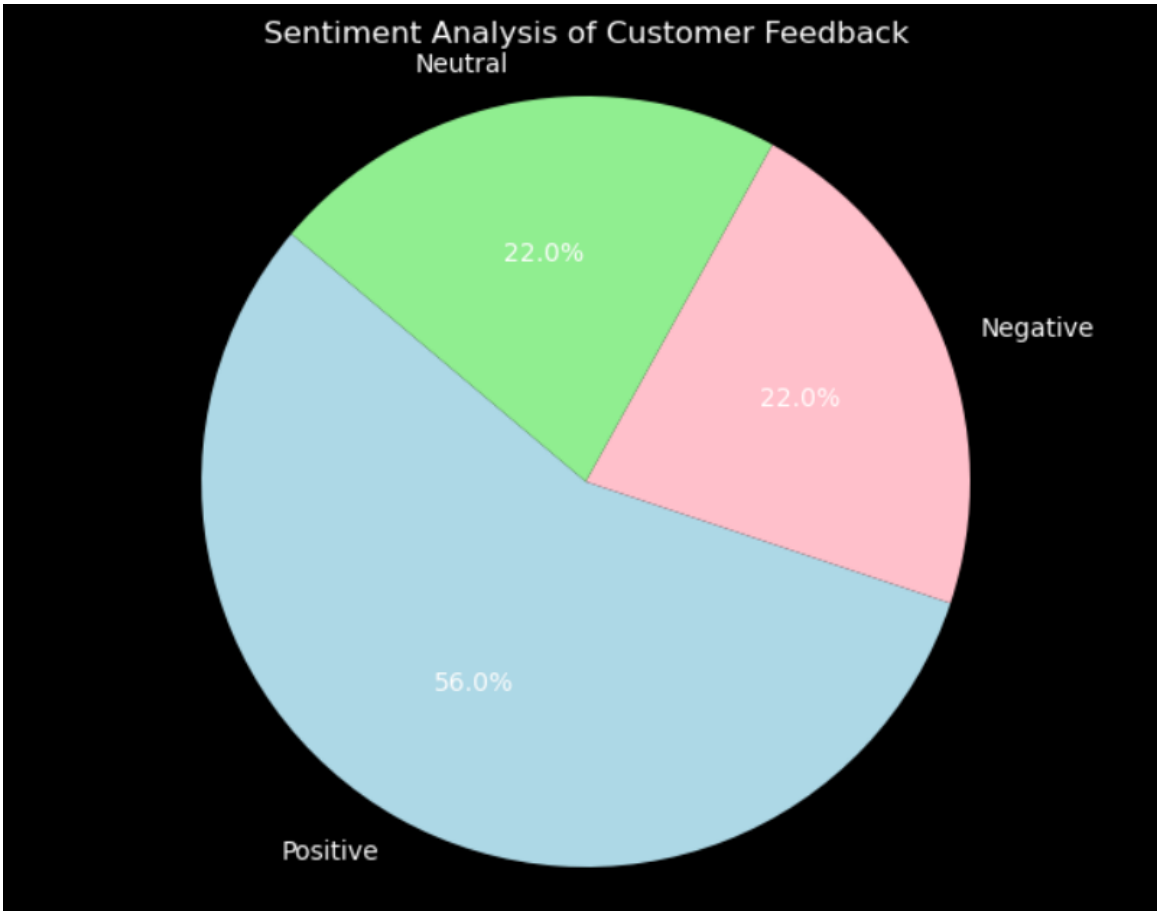

Fig 12.1 *Sentiment Analysis Customer Feedback*

13. Conclusion

This study involved a thorough examination of a retail dataset, producing insightful results in a number of areas. Key insights include product insights, temporal patterns, geographic trends, payment method preferences, profitability indicators, customer satisfaction sentiments, segmentation of the consumer base, and targeted marketing recommendations.

Targeted plans are made possible by the segmentation study, and marketing advice offer doable actions for engaging customers. Operational choices are influenced by information on product preferences, temporal trends, and geographic patterns. The analysis of returns and the mode of payment help to provide a comprehensive picture of consumer behavior.

A financial viewpoint is provided by profitability measurements, while consumer satisfaction is revealed through sentiment research. Positive feeling in general portends good for brand loyalty and perception.

In conclusion, this project gives the business useful information for making strategic decisions, improving customer satisfaction, streamlining processes, and promoting long-term success in the cutthroat retail market. The company's position for sustained success will be strengthened by further data-driven studies.