IE6400 Foundations Data Analytics Engineering

Fall Semester 2023

Customer Segmentation using RFM Analysis

Final Project Report 2

Group 6

Giridhar Babu

Hemant Manohar Deshmukh

Mallika Gaikwad

Sharvari Pravin Deshpande

Medhavi Uday Pande

Index

ABSTRACT

This project delves into the realm of Customer Segmentation using the RFM (Recency, Frequency, Monetary) analysis method applied to an extensive eCommerce dataset. The dataset serves as the foundation for unveiling insightful patterns in customer purchasing behavior. RFM segmentation emerges as a robust technique, enabling businesses to categorize customers based on their recent transaction recency, purchase frequency, and monetary value.

The overarching goal is to construct a comprehensive Customer Segmentation model that empowers businesses with targeted marketing and customer engagement strategies. By dissecting the dataset and deriving meaningful RFM scores, the resultant customer segments become invaluable for crafting tailored approaches in marketing and customer retention.

Throughout the project, a meticulous exploration of the dataset is conducted, encompassing data cleaning, preprocessing, and statistical analysis. The RFM methodology is then applied to distill crucial information about customer behavior, leading to the formation of distinct customer segments. These segments, elucidated through the amalgamation of recency, frequency, and monetary values, offer a nuanced understanding of customer groups. Analysis also include Customer segmentation understanding customer’s purchasing and spending habits, customer behavior, time analysis and also payment analysis.

The implications of the segmented customer groups are discussed, emphasizing their role in guiding marketing strategies, and fostering customer retention initiatives. As a result, the project not only showcases the technical implementation of RFM analysis but also underscores its practical significance in the dynamic landscape of eCommerce. The outcomes of this study can significantly contribute to the refinement of marketing strategies, optimizing resource allocation, and ultimately enhancing customer satisfaction and loyalty.

INTRODUCTION

In the rapidly evolving landscape of eCommerce, understanding and responding to customer behavior is paramount for businesses striving to stay ahead in a competitive market. The vast amounts of data generated in this digital realm present an opportunity for businesses to extract meaningful insights that can drive targeted strategies for marketing and customer retention. One such potent methodology is the RFM (Recency, Frequency, Monetary) analysis model. This project embarks on a journey to harness the power of RFM analysis using an expansive eCommerce dataset. Sourced from Kaggle, this dataset encapsulates a treasure trove of transactional data, providing a rich canvas for unraveling the intricacies of customer purchasing behavior. The overarching goal is to construct a robust Customer Segmentation model, leveraging the nuanced dimensions of recency, frequency, and monetary value.

The significance of this project lies in its potential to revolutionize how businesses engage with their customers. By homing in on the recency of purchases, the frequency of transactions, and the monetary value attached to each customer, the RFM model enables the creation of distinct customer segments. These segments serve as windows into the diverse behaviors exhibited by customers, paving the way for targeted and personalized marketing strategies.

In the dynamic eCommerce landscape, where customer preferences can swiftly shift, the need for adaptive and data-driven approaches is more pronounced than ever. RFM analysis emerges as a beacon, guiding businesses through the labyrinth of customer data and illuminating pathways to optimize marketing efforts. This project not only seeks to implement the technical intricacies of RFM analysis but, more crucially, aims to unravel actionable insights that can shape the trajectory of eCommerce businesses.

The primary objective is clear — to perform RFM analysis on the provided dataset and distill it into distinct customer segments. These segments will not only act as mirrors reflecting customer behaviors but will also serve as the foundation for tailoring marketing and customer retention strategies. As we delve into the depths of the eCommerce dataset, our exploration will encompass meticulous data cleaning, preprocessing, and a methodical application of the RFM methodology. The subsequent unveiling of customer segments will not merely be a technical achievement but a strategic imperative, empowering businesses to navigate the complex terrain of customer engagement with precision.

In the pages that follow, we embark on this analytical journey, unraveling the stories hidden within the data and, in doing so, arming businesses with the insights needed to thrive in the dynamic realm of eCommerce.

DATA SOURCES

The dataset employed in this analysis originates from the UCI Machine Learning Repository, encompassing transactional records spanning from December 1st, 2010 to December 9th, 2011. These transactions represent a UK-based non-store online retail business specializing in unique all-occasion gifts. It is important to note that this dataset captures actual transactional data, reflecting purchases and interactions of customers, including a considerable clientele of wholesalers.

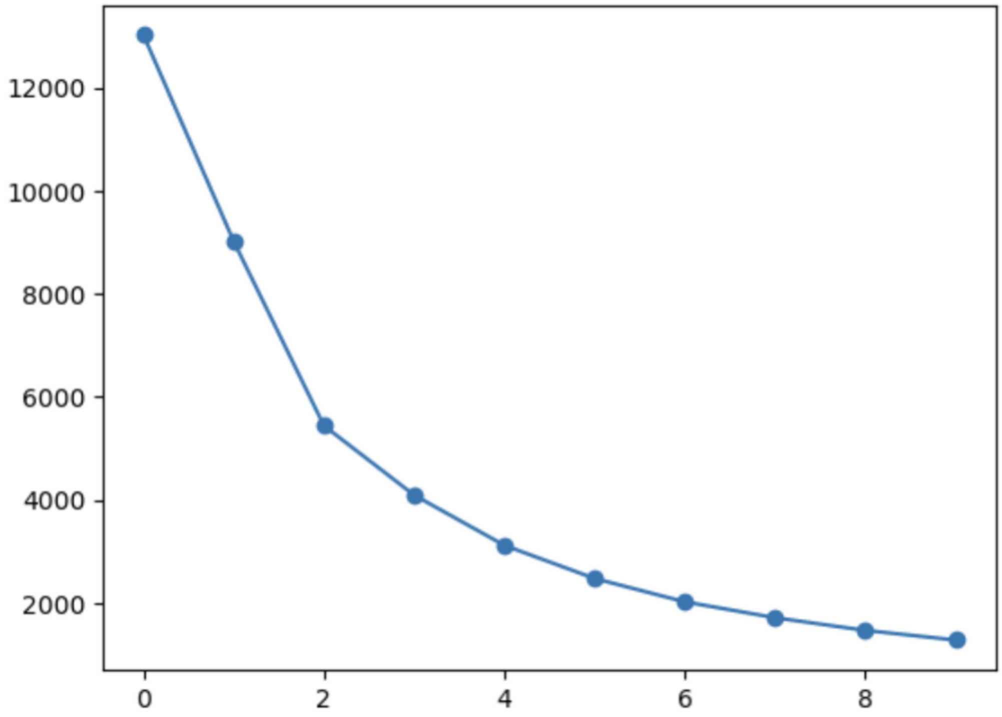
Acknowledging its source from the UCI Machine Learning Repository, the data was made accessible by Dr. Daqing Chen, Director of the Public Analytics group, School of Engineering, London South Bank University, UK (chend '@' lsbu.ac.uk). The dataset's content and the timeframe it covers offer a unique opportunity for analyses spanning various dimensions, such as time series, clustering, classification, and more.

In aligning with best practices for data preprocessing and ensuring its reliability for analytical purposes, meticulous steps were undertaken. These encompassed initial data acquisition, thorough examination, and rigorous cleaning procedures. During this process, redundant or irrelevant columns unrelated to RFM analysis, such as references and non-pertinent date records, were identified and subsequently removed, optimizing the dataset for subsequent analysis. Additionally, the focus on feature extraction techniques was maintained to derive valuable insights from the data, reinforcing its suitability for RFM segmentation and other pertinent analyses.

It is imperative to recognize that while the dataset provides a comprehensive overview of transactions and customer interactions, the accuracy and efficacy of analyses conducted are contingent upon the quality of the original records maintained in the database. Any concerns regarding data quality or specific data points are duly acknowledged and can be addressed through comments or inquiries.

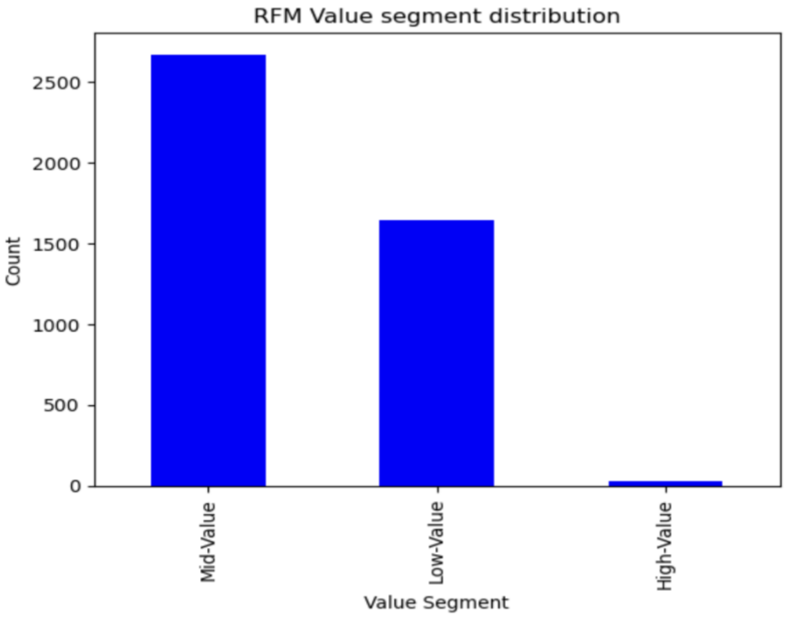
RESULT AND METHODS

1. Customer Segmentation



The resulting plot is known as the elbow curve. The goal is to visually identify the point on the curve where the rate of decrease in inertia starts to slow down, forming an elbow. This point is considered the optimal number of clusters for the K-Means algorithm. In summary, this code helps in determining the optimal number of clusters for a dataset by running the K-Means algorithm with different values of k, calculating the inertia for each, and then plotting the results to find the elbow where further clustering doesn't significantly improve the model.

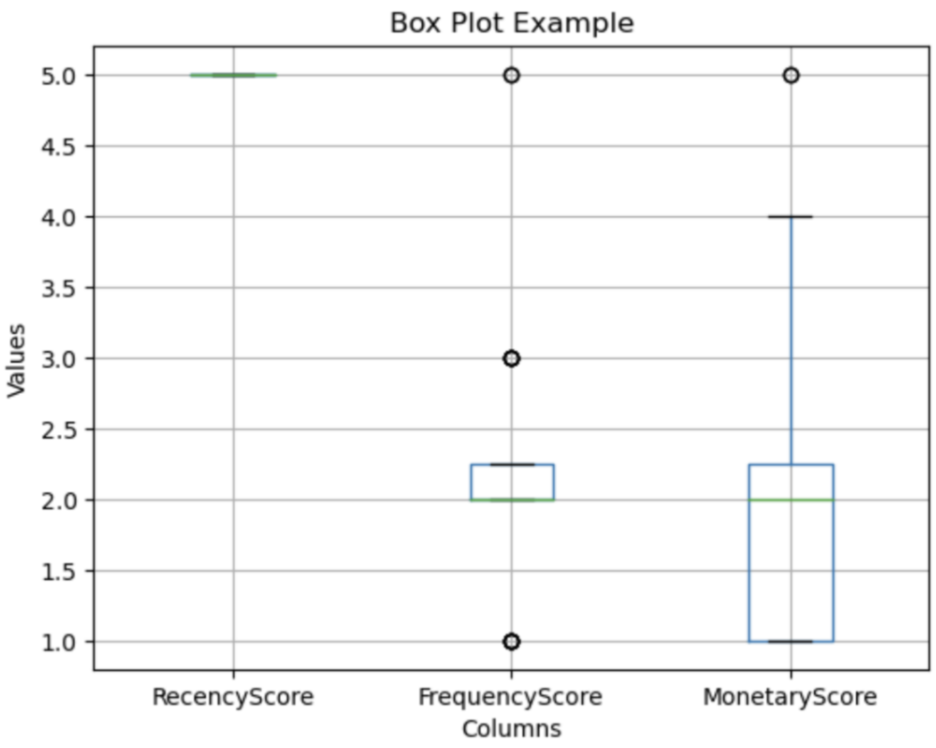
2. Segment Profiling



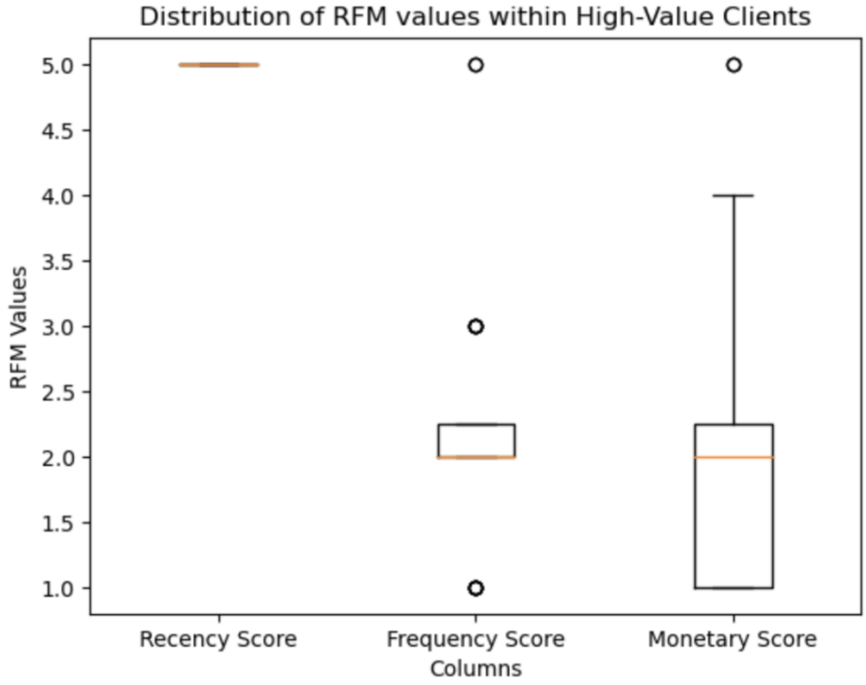
In terms of visual output, the resulting bar graph provides a clear depiction of the distribution of customers across different RFM value segments. Each bar represents a specific segment, and its height indicates the count or frequency of customers falling within that segment. This visual representation is instrumental in understanding the composition of the customer base, showcasing which value segments have higher or lower customer representation. Consequently, the graph serves as a valuable tool for marketers and analysts to identify patterns in customer behavior and tailor strategies accordingly.

The labels on the x-axis and y-axis, along with the graph's title, contribute to the overall interpretability of the visualization. The x-axis denotes the 'Value Segment,' the y-axis represents the 'Count' of customers in each segment, and the title, 'RFM Value segment distribution,' encapsulates the essence of the plotted information. This type of visualization aids stakeholders in gaining insights into the customer segmentation, facilitating informed decision-making in marketing and engagement strategies based on the distribution of customers across RFM value segments.

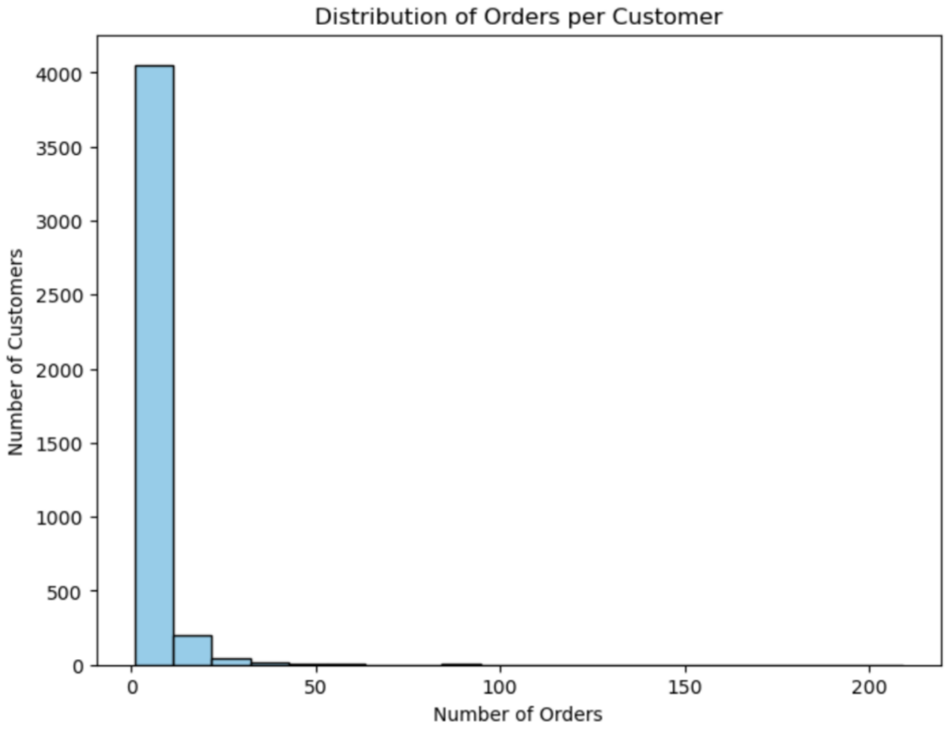
3. RFM Distribution Visualization



The expected output is a box plot with three boxes side by side, each corresponding to one of the specified score columns. The box plot will display the distribution of values within each score, including the median, quartiles, and potential outliers. This visualization is particularly useful for understanding the spread and central tendencies of these scores, allowing for a quick comparison across the three different aspects (recency, frequency, and monetary) within the high-value client segment.



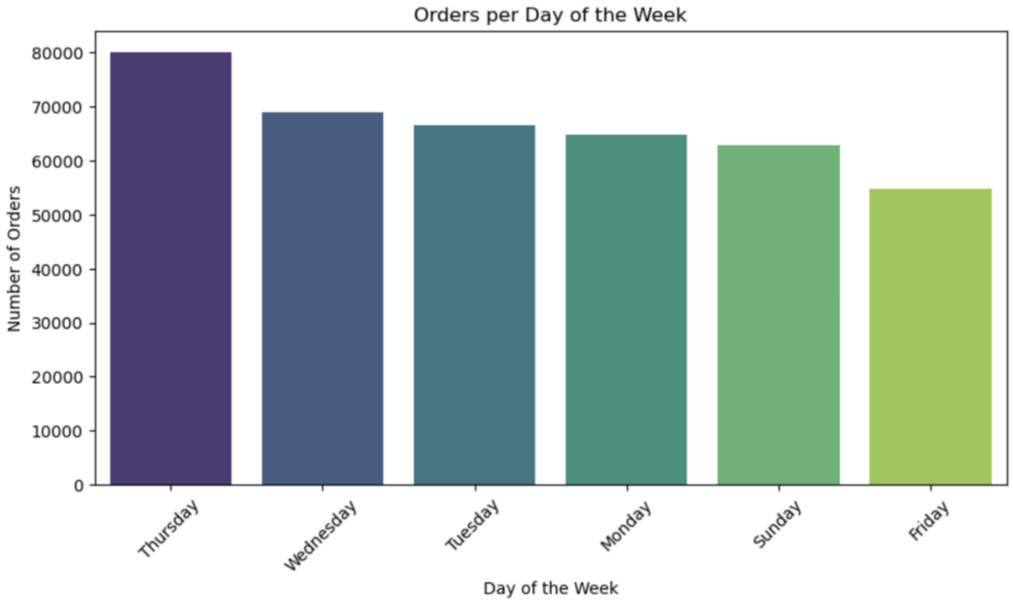
So, overall, the code is creating a box plot to visually represent the distribution of Recency, Frequency, and Monetary scores within the high-value clients dataset. The x-axis represents the three different scores, and the y-axis represents the values of these scores. The plot allows you to observe the spread, central tendency, and any potential outliers in the data for each score.



We have used Matplotlib to create a histogram to visualize the distribution of the number of orders per customer. The plt.hist function is used to create the histogram, where orders_per_customer is the data to be plotted. The histogram is divided into 20 bins (bins=20), and it is colored in 'skyblue' with black edges. The x-axis represents the number of orders, the y-axis represents the number of customers, and the title is set to 'Distribution of Orders per Customer'. Finally, plt.show() is used to display the plot.

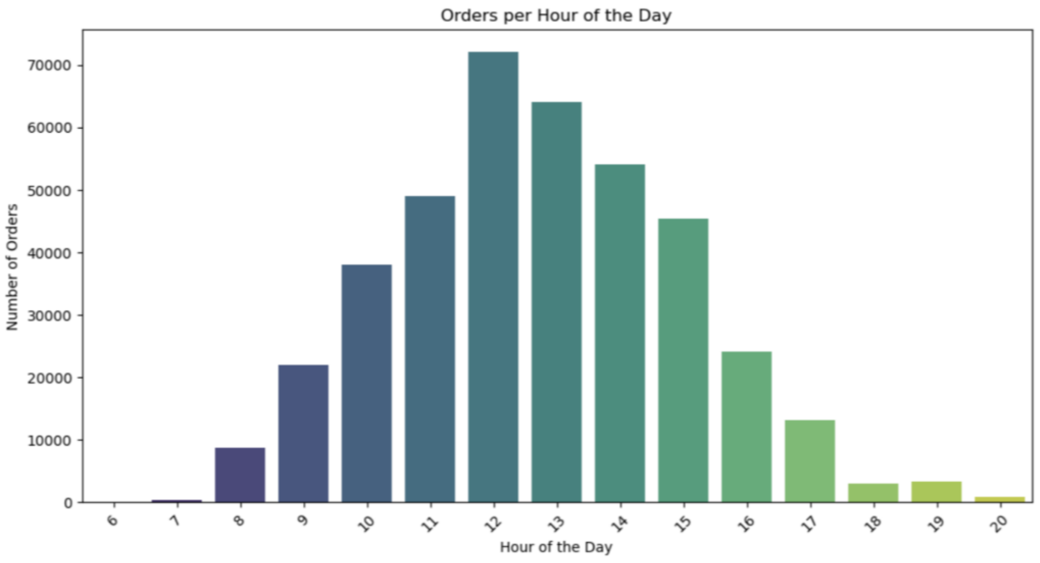
In summary, this graph provides insights into the distribution of the number of orders per customer in the given dataset. The histogram helps visualize how many customers have a certain number of orders, providing an overview of customer purchasing patterns.

4. Time Analysis

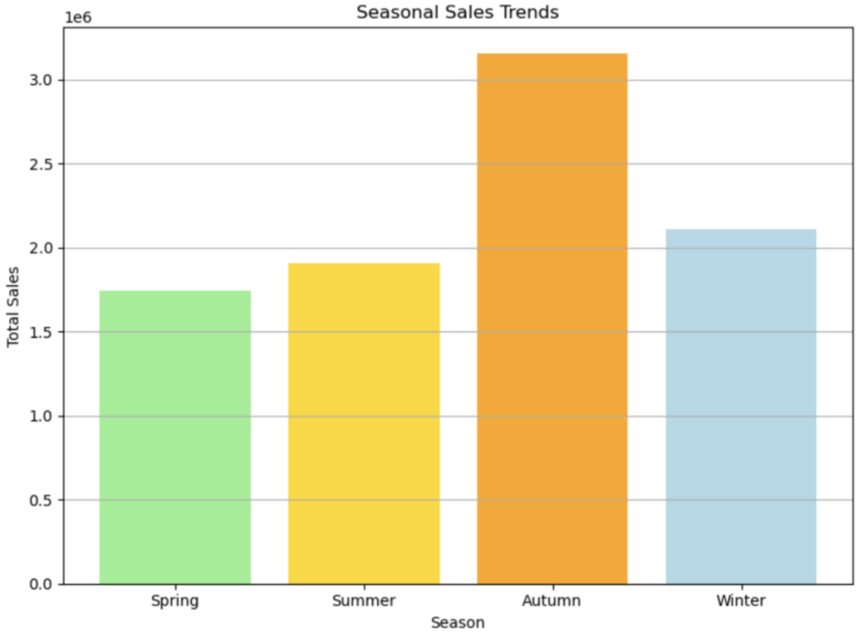


This graph is constructed using seaborn to create a bar plot that visualizes the number of orders for each day of the week. The x-axis represents the days of the week, the y-axis represents the number of orders, and each bar corresponds to a specific day. The plot is titled 'Orders per Day of the Week'.

It is a bar plot showing the distribution of orders across different days of the week, providing insights into the patterns of customer orders on specific days. The rotation=45 argument in plt.xticks (rotation=45) is used to rotate the x-axis labels for better readability. The color palette 'viridis' is applied to the bars for better visualization.

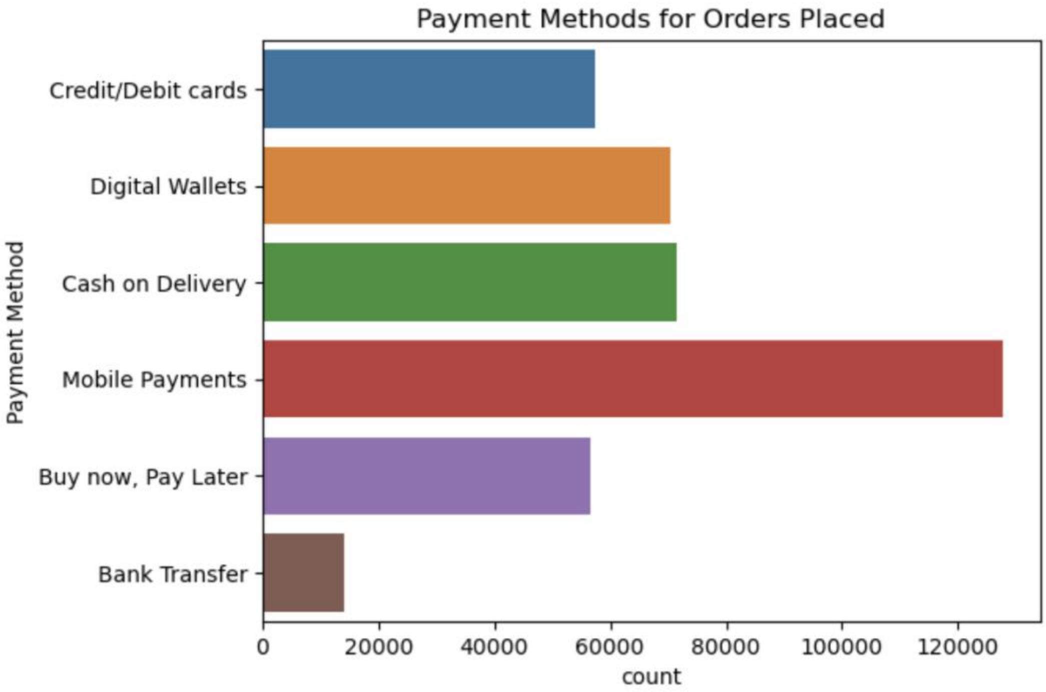


This graph shows a bar plot to visualize the distribution of orders throughout the hours of the day. In this graph, the x-axis represents the hours of the day, the y-axis represents the corresponding number of orders, and each bar corresponds to a specific hour. The plot is titled 'Orders per Hour of the Day', and the x-axis labels are rotated by 45 degrees for better readability. Overall, this visualization allows for an examination of the hourly patterns in customer order frequency, providing insights into peak and off-peak hours of activity.

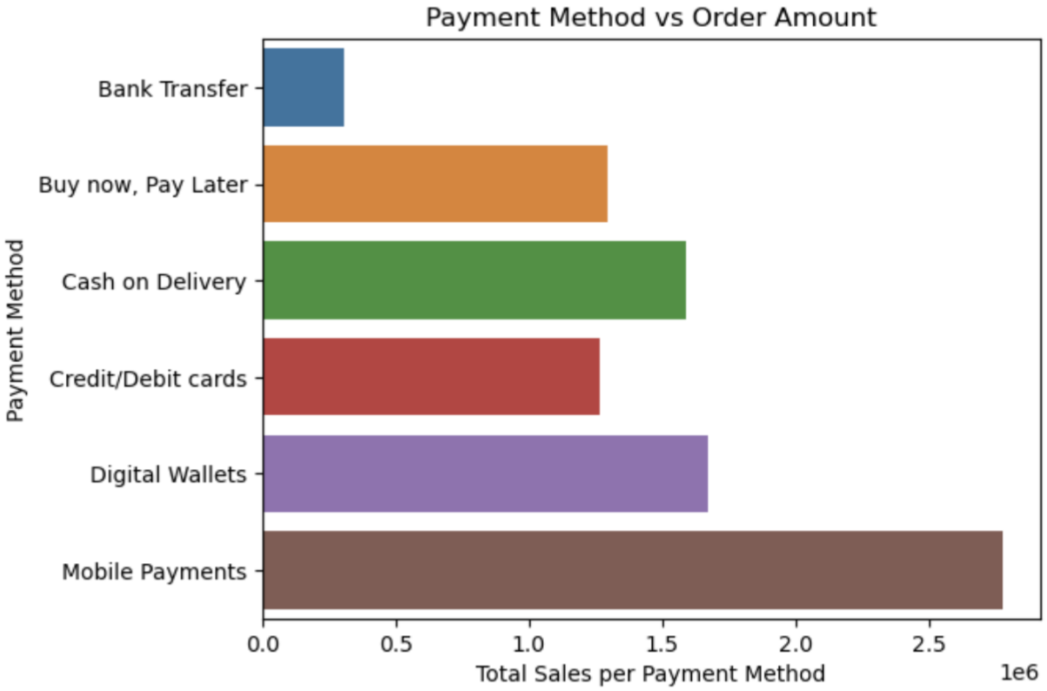


This graph represents the seasonal trends among the four seasons, Spring, Summer, Autumn, and Winter. The Sales is the highest in Autumn season as seen from the graph, following the Winter season, then the summer season. The spring has the lowest sales.

5. Payment Analysis

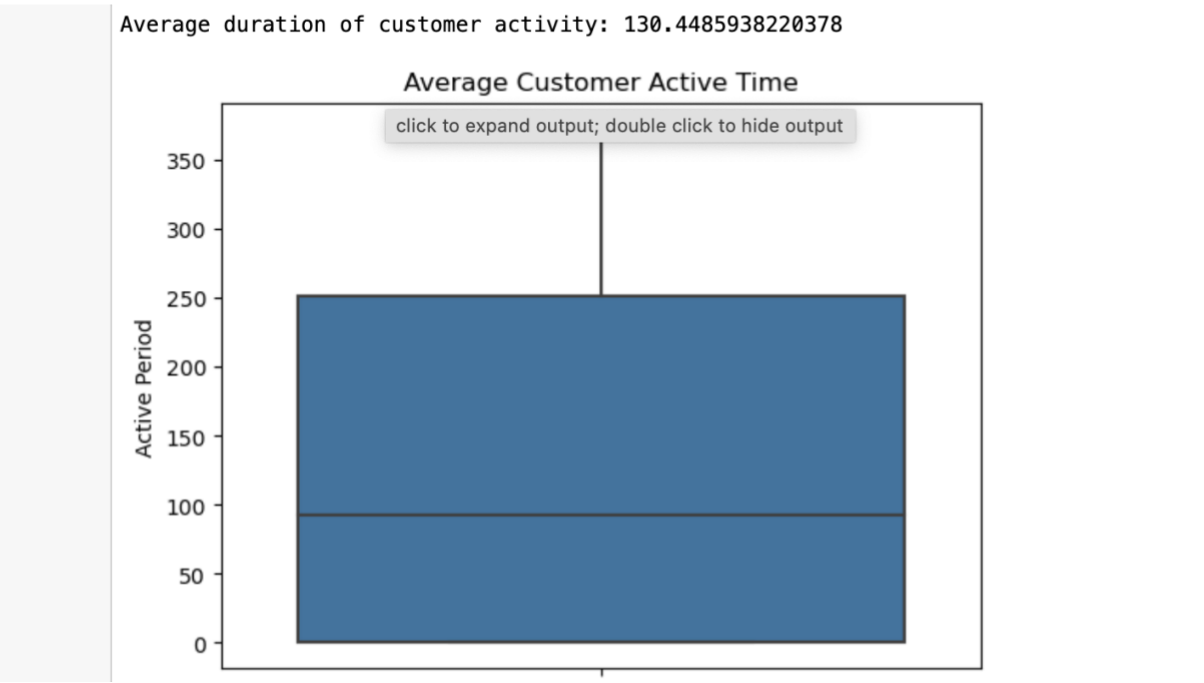


This graph shows payment analysis being performed on the customer data, it shows different types of payment methods, like Credit/Debit Cards, Digital Wallets, Cash on Delivery, Mobile Payments, Buy now- pay later, Bank Transfer. The most used payment method used/ preferred by customers is Mobile Payments. Mobile Payments is the most preferred by customers due to its convenience. Cash On Delivery is the second most used payment by customers. The third most used as per the data is the digital wallets method, followed by Credit/debit cards. The lowest preferred method is Bank Transfer.



The above graph indicates Payment Method VS Order Amount. The most orders has been placed by the Mobile Payments- payment method. It is followed by Digital Wallets being the most used while ordering. The third payment method is Cash on Delivery, the followed by Buy now-Pay Later.

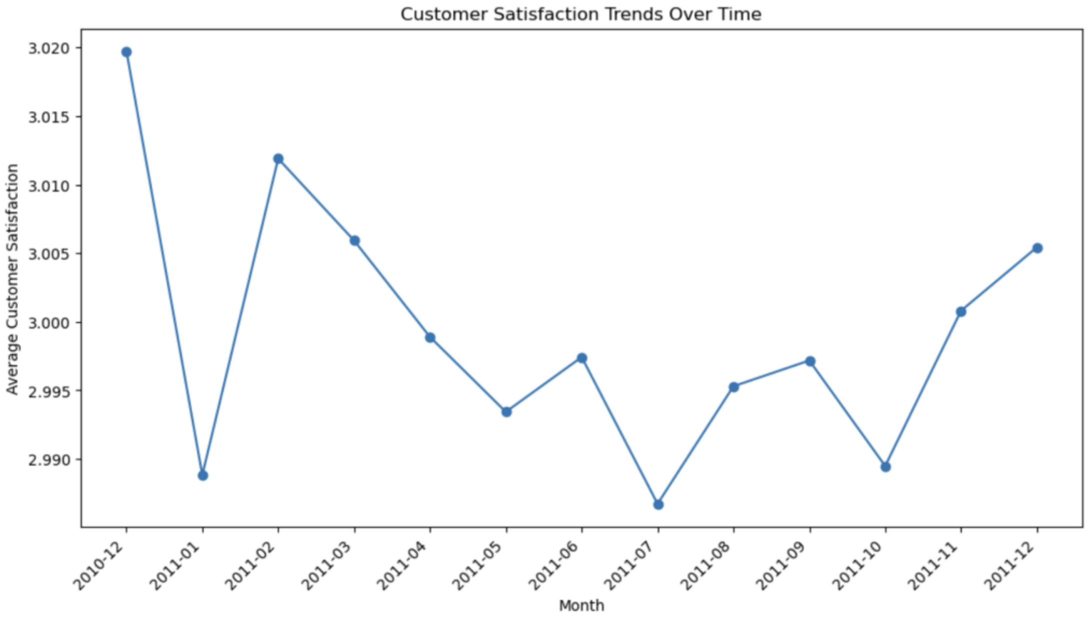
6. **Customer Behavior**



```
Average duration of customer activity: 130.4485938220378
```

This graph uses weaborn to create a box plot for the distribution of customer active time. The y-axis represents the duration of customer activity in days, and the plot is titled "Average Customer Active Time". The box plot provides a visual summary of the central tendency and spread of the customer activity durations, including any potential outliers. In summary, the code calculates and visualizes the distribution of customer activity durations, providing insights into how long, on average, customers remain active based on the time span between their first and last transactions.

7. **Customer Satisfaction**



The x-axis represents the months, and the y-axis represents the corresponding average customer satisfaction values. Each data point is marked with a circular marker ('o'). The plot is titled 'Customer Satisfaction Trends Over Time', and the x-axis labels are rotated for better readability. The resulting plot provides a visual representation of how average customer satisfaction has changed over the months, helping to identify trends or patterns in customer satisfaction over time.

SUMMARY

This project focuses on employing RFM (Recency, Frequency, Monetary) analysis, a widely used method in the field of customer relationship management, to segment customers in an eCommerce dataset. By leveraging the dataset, which can be accessed on Kaggle. The analysis aims to categorize customers based on their recent purchasing behavior, purchase frequency, and monetary value. The resulting RFM scores enable the creation of distinct customer segments, offering valuable insights for targeted marketing and customer engagement strategies. By identifying and understanding these segments, businesses can tailor their approach to customer retention, satisfaction, and personalized marketing initiatives. The project's objective is to provide actionable insights derived from RFM analysis, facilitating effective decision-making for enhancing customer relationships and optimizing business outcomes in the eCommerce domain.

LIMITATIONS

While RFM analysis is a powerful technique for customer segmentation, it's important to be aware of potential limitations that may impact the accuracy and effectiveness of the segmentation model:

1. Data Completeness and Quality:

- Missing Data: Incomplete or missing data in the dataset can lead to inaccurate RFM scores, affecting the segmentation results.

- Data Accuracy: Inaccuracies in transaction data or errors in recording customer information may distort the RFM analysis.

2. Equal Weighting of RFM Components:

- Assumption of Equal Importance: RFM analysis assumes equal importance of recency, frequency, and monetary metrics. In reality, different businesses may prioritize these components differently based on their industry and goals.

3. Static Analysis:

- Assumption of Stability: RFM analysis assumes that customer behavior remains relatively stable over time. Sudden changes in customer behavior patterns, such as during a promotion or economic event, may not be accurately captured.

4. Lack of Context:

- External Factors: RFM analysis may not consider external factors influencing customer behavior, such as market trends, seasonality, or macroeconomic conditions, which can impact the validity of segments.

5. Single-Dimensional View:

- Limited Customer Insights: RFM focuses on transactional data and may not capture other important aspects of customer behavior, such as preferences, feedback, or interactions with the brand.

6. Homogeneity Within Segments:

- Assumption of Homogeneous Segments: RFM segments assume homogeneity within each group. In reality, customer preferences and behaviors within a segment can still vary significantly.

7. Threshold Selection:

- Subjectivity in Thresholds: Choosing appropriate thresholds for recency, frequency, and monetary values is subjective and may vary based on business objectives or industry norms.

8. Overlooking Customer Lifecycle:

- Customer Lifecycle Stages: RFM may not account for different stages of the customer lifecycle, such as acquisition, retention, or churn, which could impact segmentation strategies.

Despite these limitations, RFM analysis remains a valuable tool for customer segmentation. To mitigate these challenges, it's crucial to complement RFM insights with additional data sources and to periodically review and update segmentation models based on evolving business dynamics and customer behaviors.

FUTURE SCOPE

The future scope for customer segmentation using RFM analysis in the eCommerce domain is broad and can encompass various advancements and integrations. Here are some potential areas of future development:

1. **Advanced Analytics and Machine Learning:** Integration of advanced analytics techniques and machine learning algorithms to enhance the accuracy and predictive power of customer segmentation models.

2. **Real-Time Segmentation:** Development of real-time customer segmentation models that can adapt to dynamic changes in customer behavior, allowing businesses to respond promptly to emerging trends.

3. **Multichannel Integration:** Integration of data from multiple channels (online and offline) to create a holistic view of customer interactions, enabling more comprehensive and accurate segmentation.

4. **Personalization and Hyper-Personalization:** Utilization of RFM insights for personalized marketing campaigns and hyper-personalization, tailoring product recommendations, offers, and content based on individual customer preferences.

5. **Customer Journey Mapping:** Integration of RFM analysis into comprehensive customer journey mapping, considering touchpoints and interactions across various stages of the customer lifecycle.

The future scope for customer segmentation using RFM analysis involves continuous innovation and adaptation to technological advancements, evolving customer expectations, and changes in the eCommerce landscape. It presents opportunities for businesses to gain deeper insights into customer behavior and enhance their strategies for customer engagement and retention.