Online Retail Customer Segmentation

By,

Shafiq Abubacker

Introduction

- The primary objective of this project is to identify major customer segments within a transnational dataset covering the period from 01/12/2010 to 09/12/2011.
- The dataset contains 541,909 rows with 8 columns.
- The dataset pertains to a UK-based and registered non-store online retail company specializing in unique all-occasion gifts.
- Understanding and segmenting customers based on their behavior and preferences is crucial for tailoring marketing strategies, enhancing customer satisfaction, and optimizing business operations.

Project Workflow

- Data Preprocessing
- Exploratory Data Analysis
- Selection of Clustering Algorithms
- Optimization of Clustering Parameters
- Interpretation of Cluster Results
- Actionable Insights and Recommendations

Data Overview

- **InvoiceNo:** A unique identifier for each transaction or invoice.
- **StockCode:** A unique identifier for each product or item.
- **Description:** A textual description of the product.
- **Quantity:** The quantity of each product sold in the
- **InvoiceDate:** The date and time when the transaction occurred.
- **UnitPrice:** The unit price of each product.
- **CustomerID:** A unique identifier for each customer.
- **Country:** The country where the transaction took place.

Data Types:

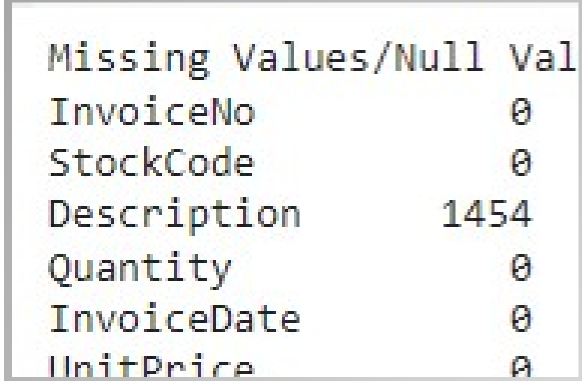
Integer: Quantity

Float: UnitPrice, CustomerID

Object: InvoiceNo, StockCode, Description, Country

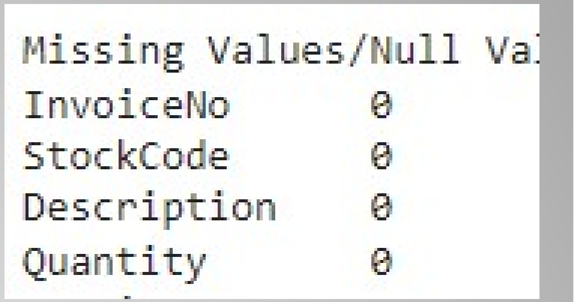
DateTime: InvoiceDate

Data Wrangling



```
Missing Values/Null Val
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
```

Removed all rows with missing values in the CustomerID column using the dropna() function.

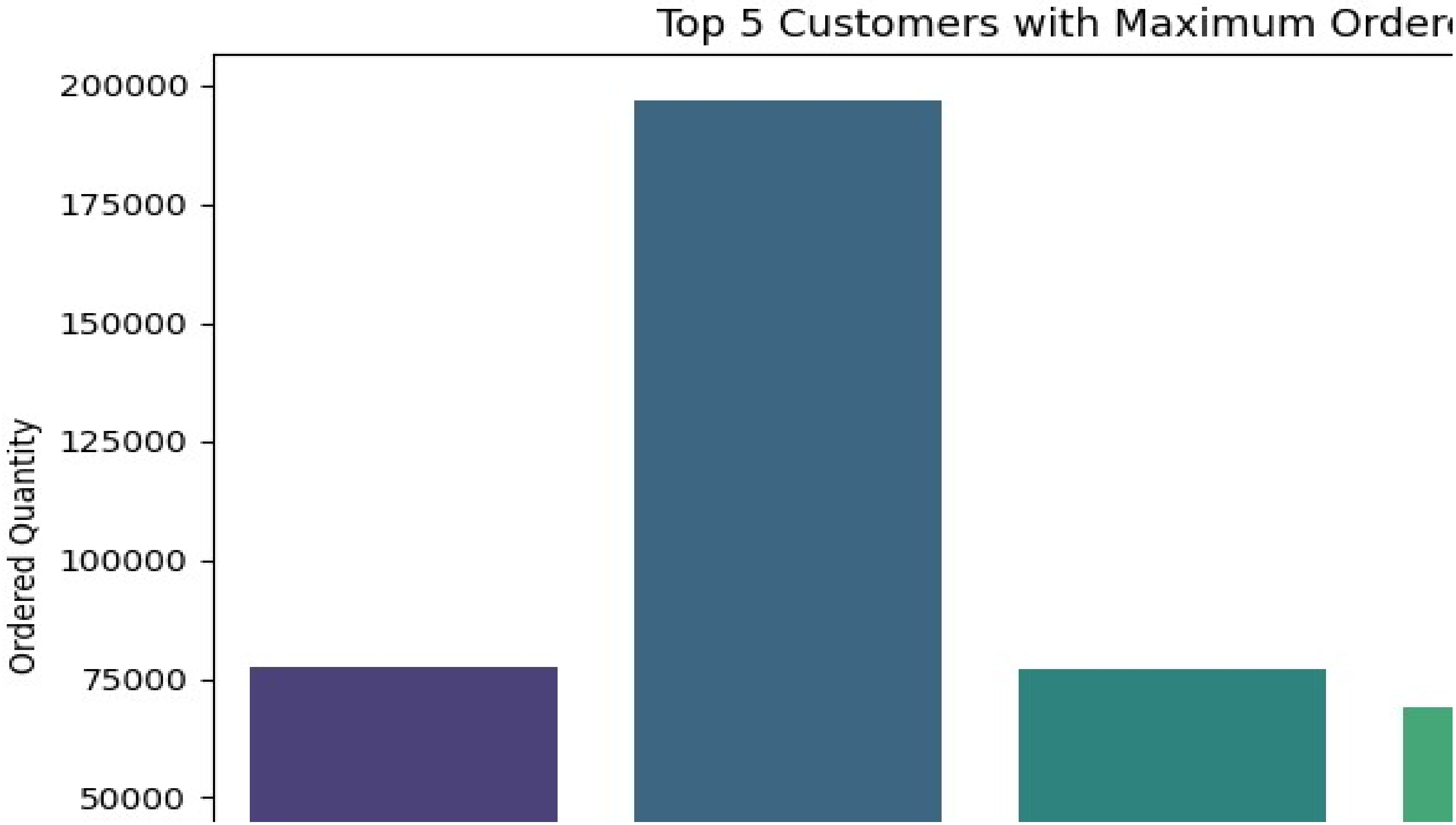


```
Missing Values/Null Va
InvoiceNo       0
StockCode       0
Description     0
Quantity        0
```

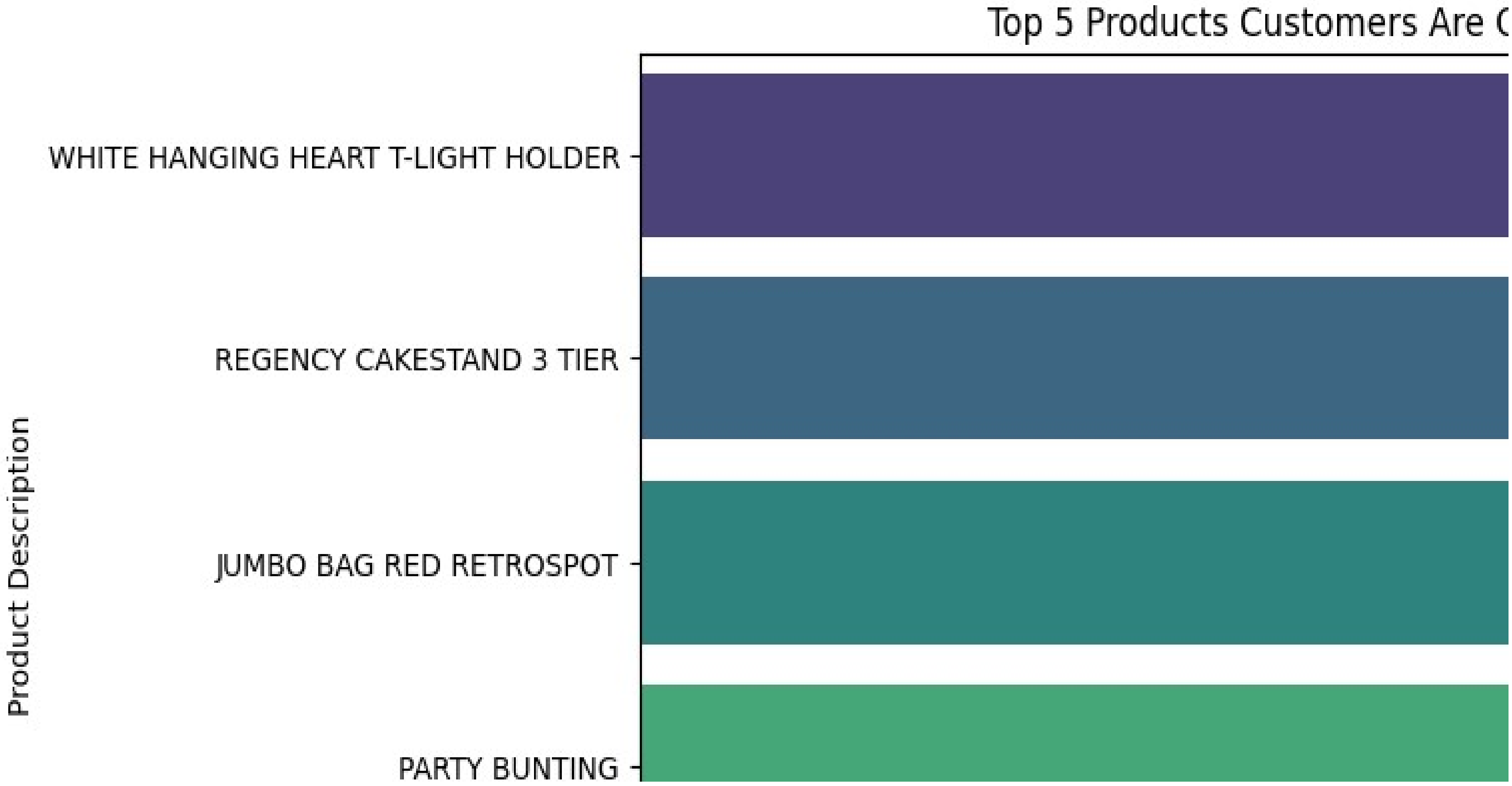
- There are 5268 duplicate values in the dataset which has been removed from the dataset using the drop_duplicates() function.

- Extracted the year, month, day, and hour components from the 'InvoiceDate' column.

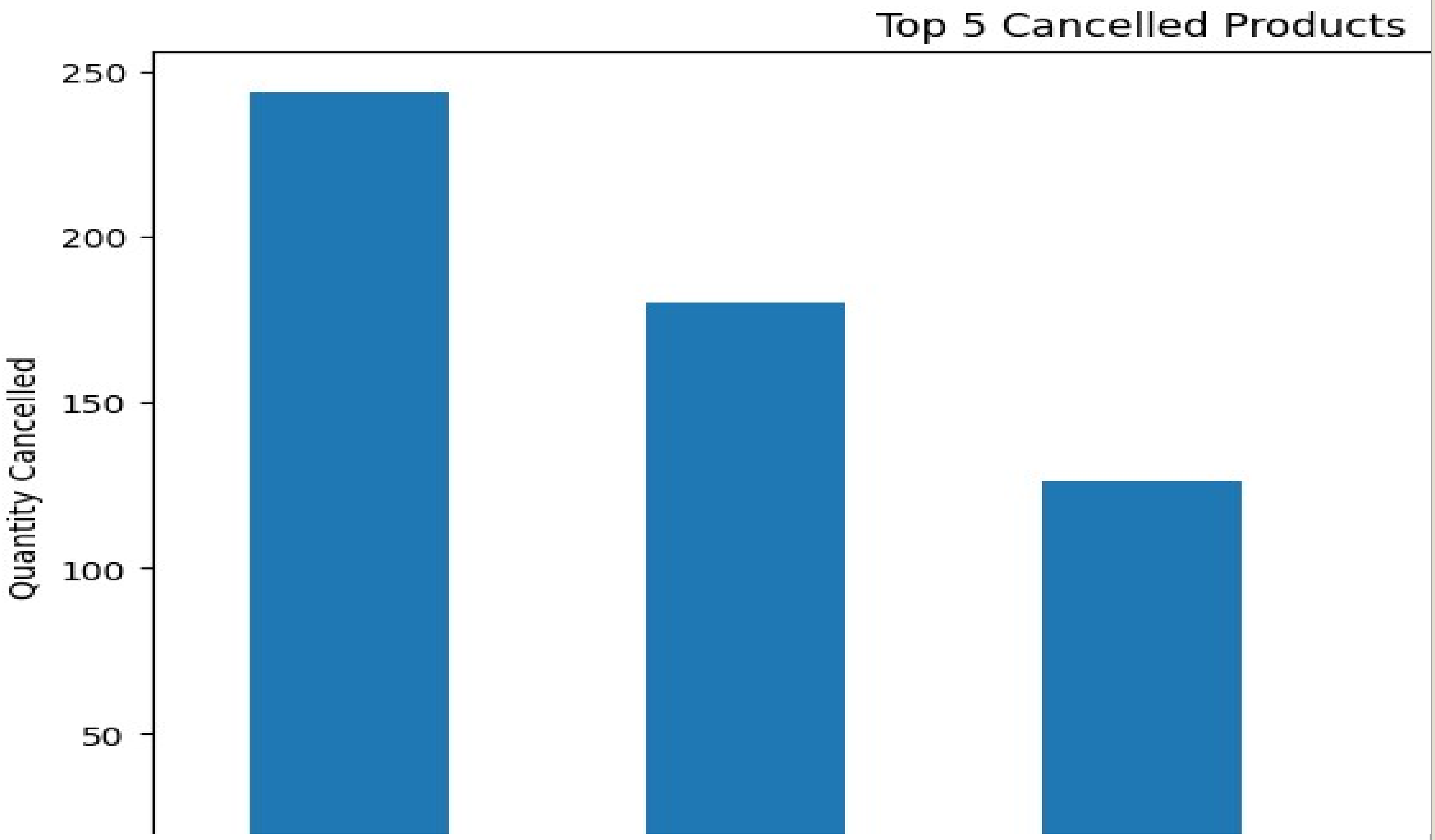
Exploratory Data Analysis



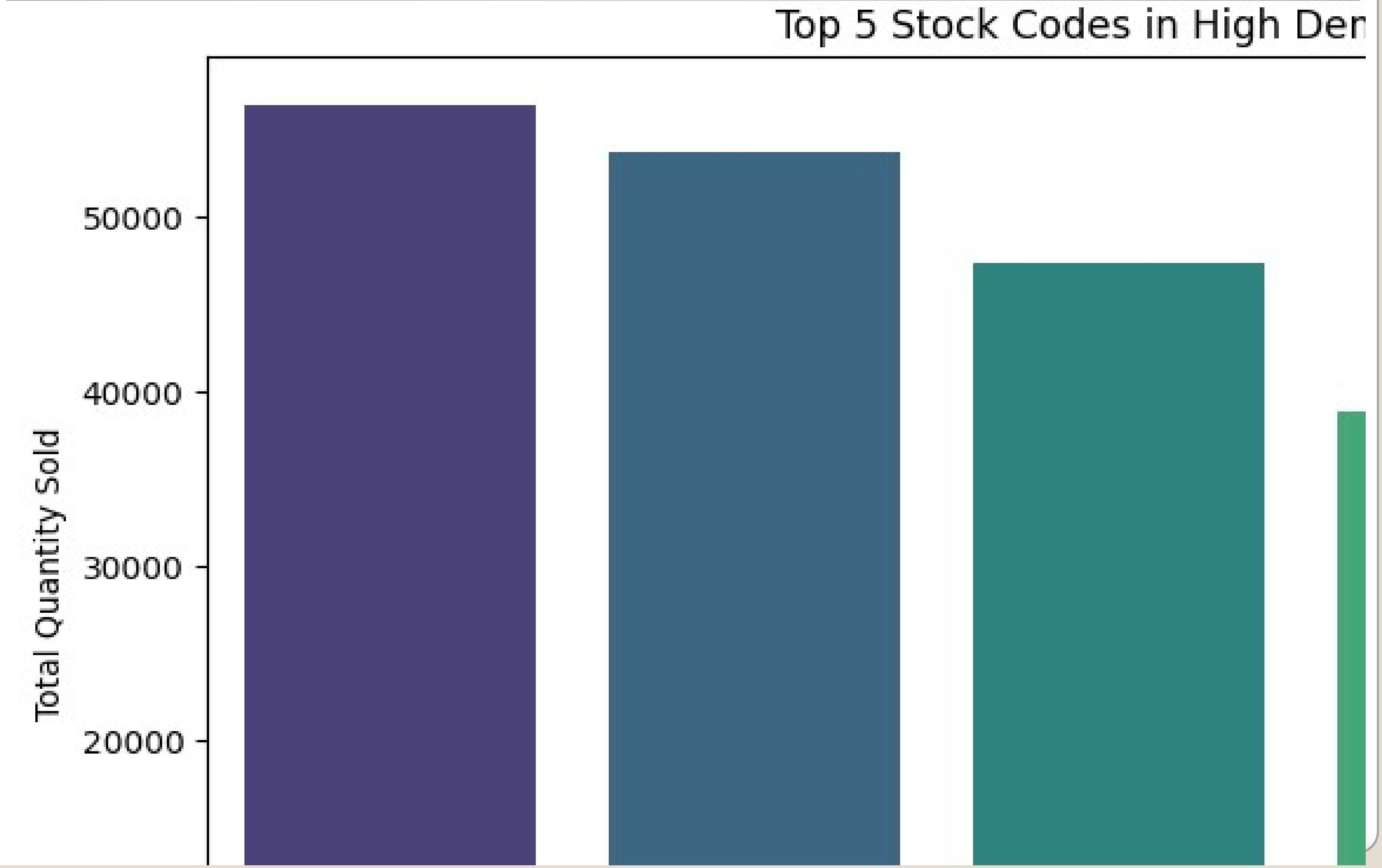
Exploratory Data Analysis



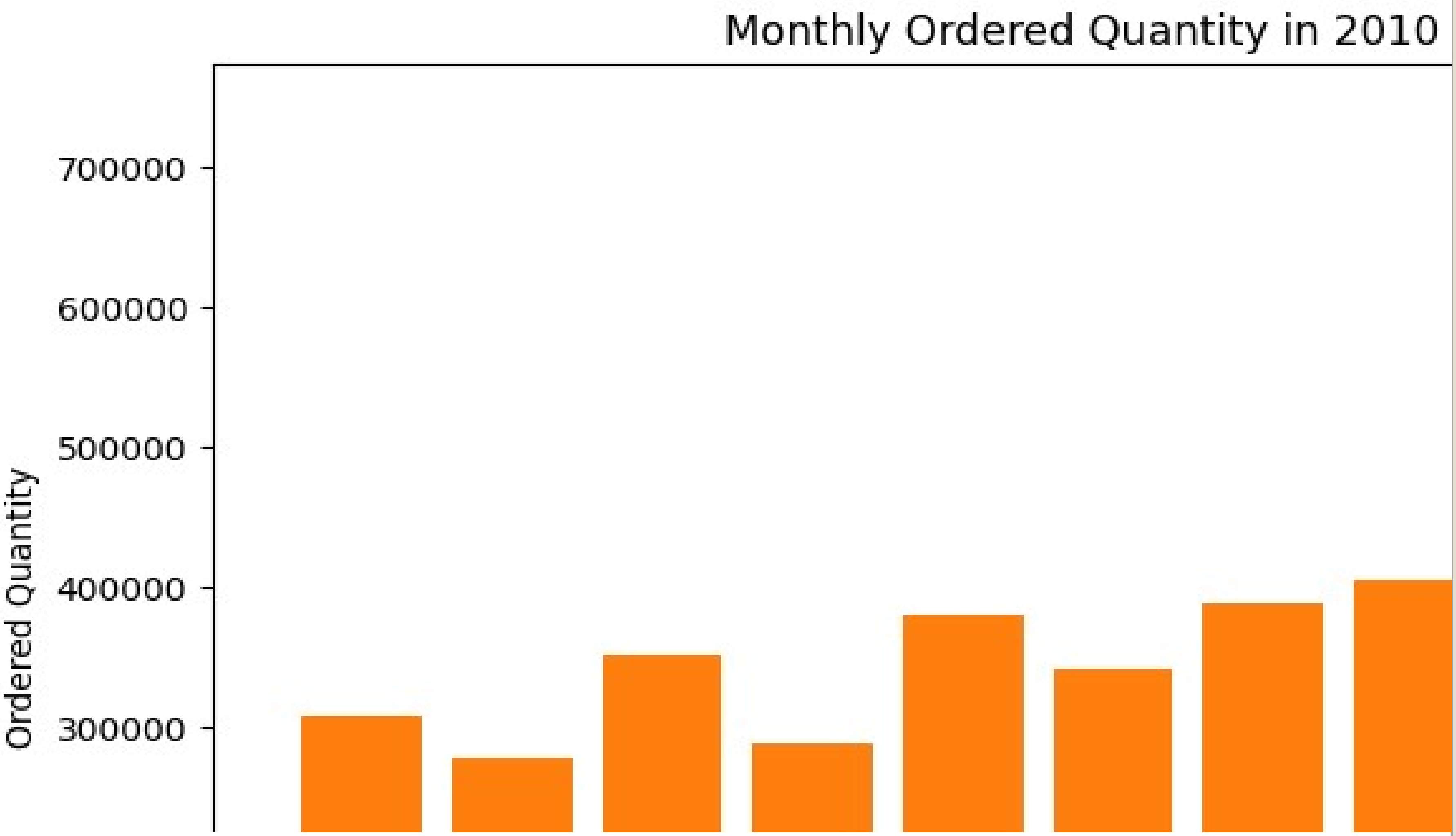
Exploratory Data Analysis



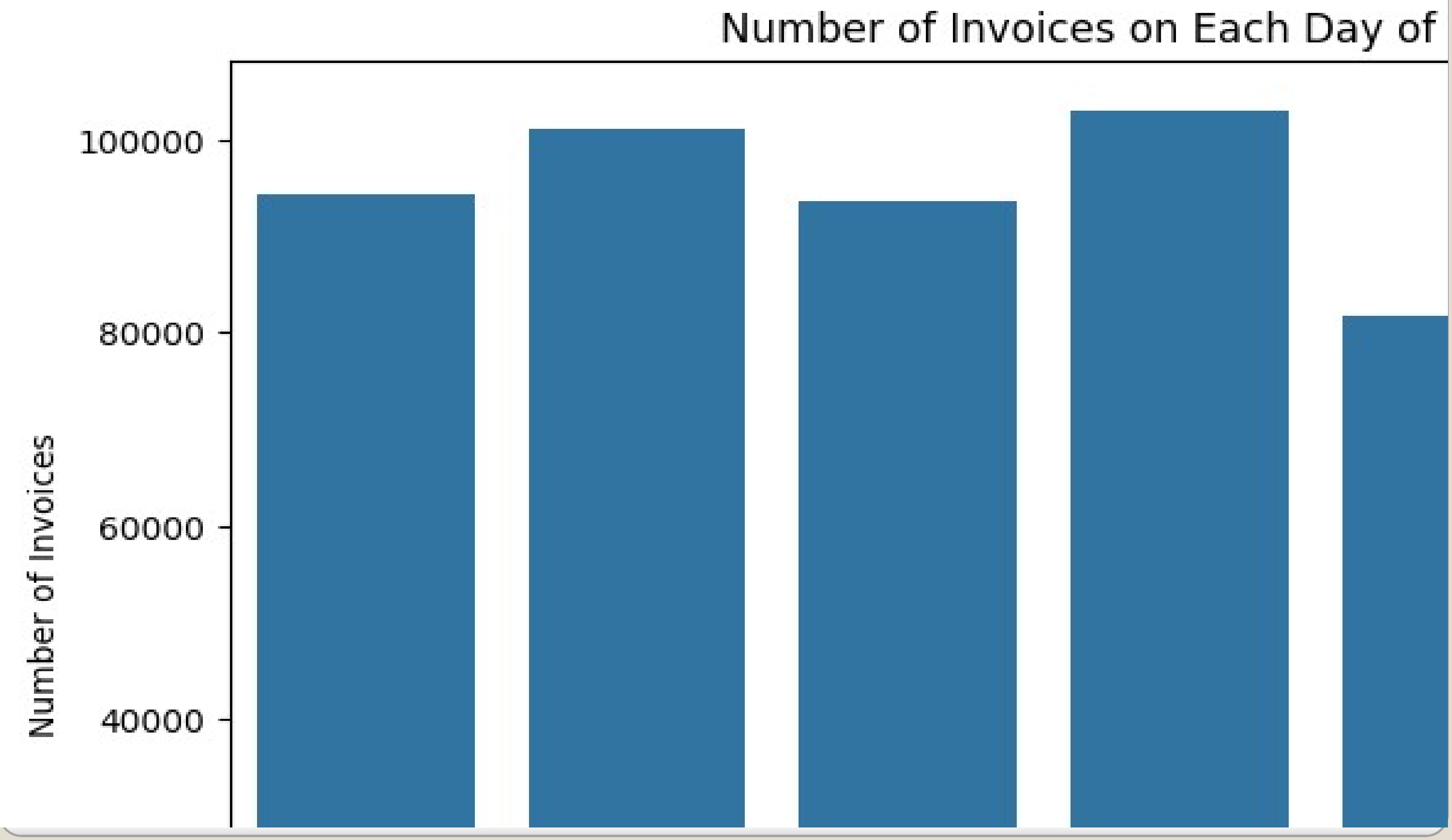
Exploratory Data Analysis



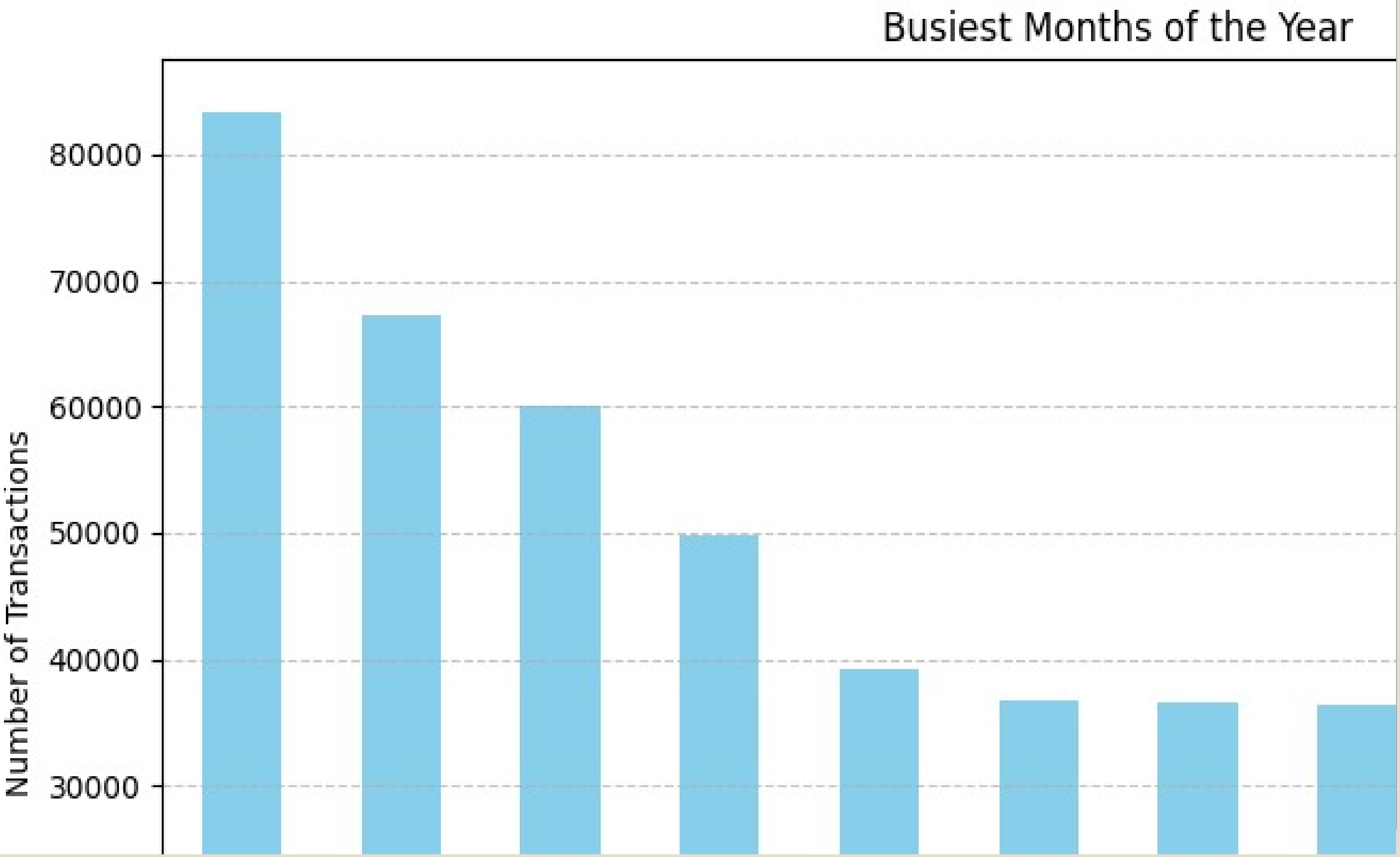
Exploratory Data Analysis



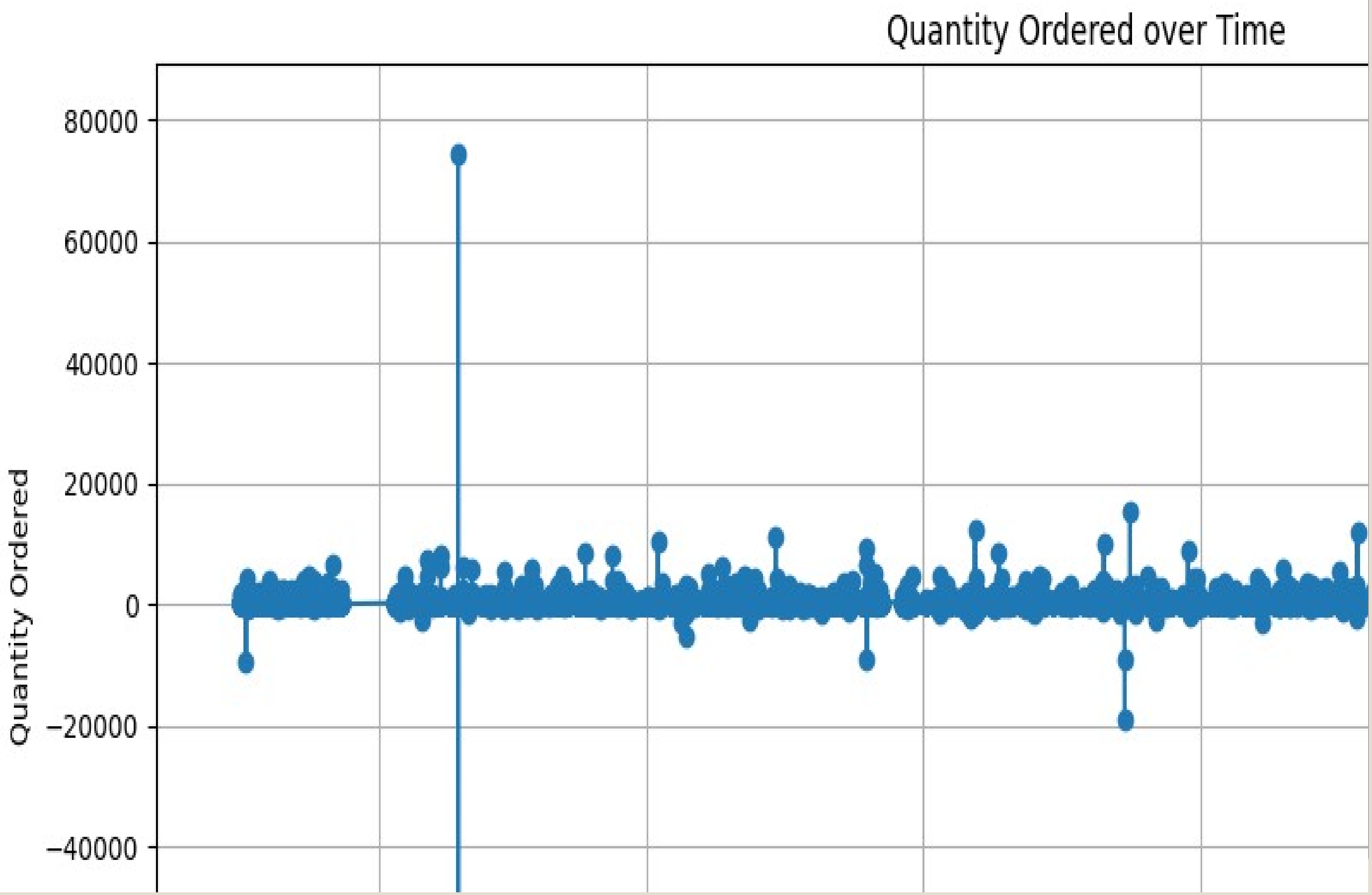
Exploratory Data Analysis



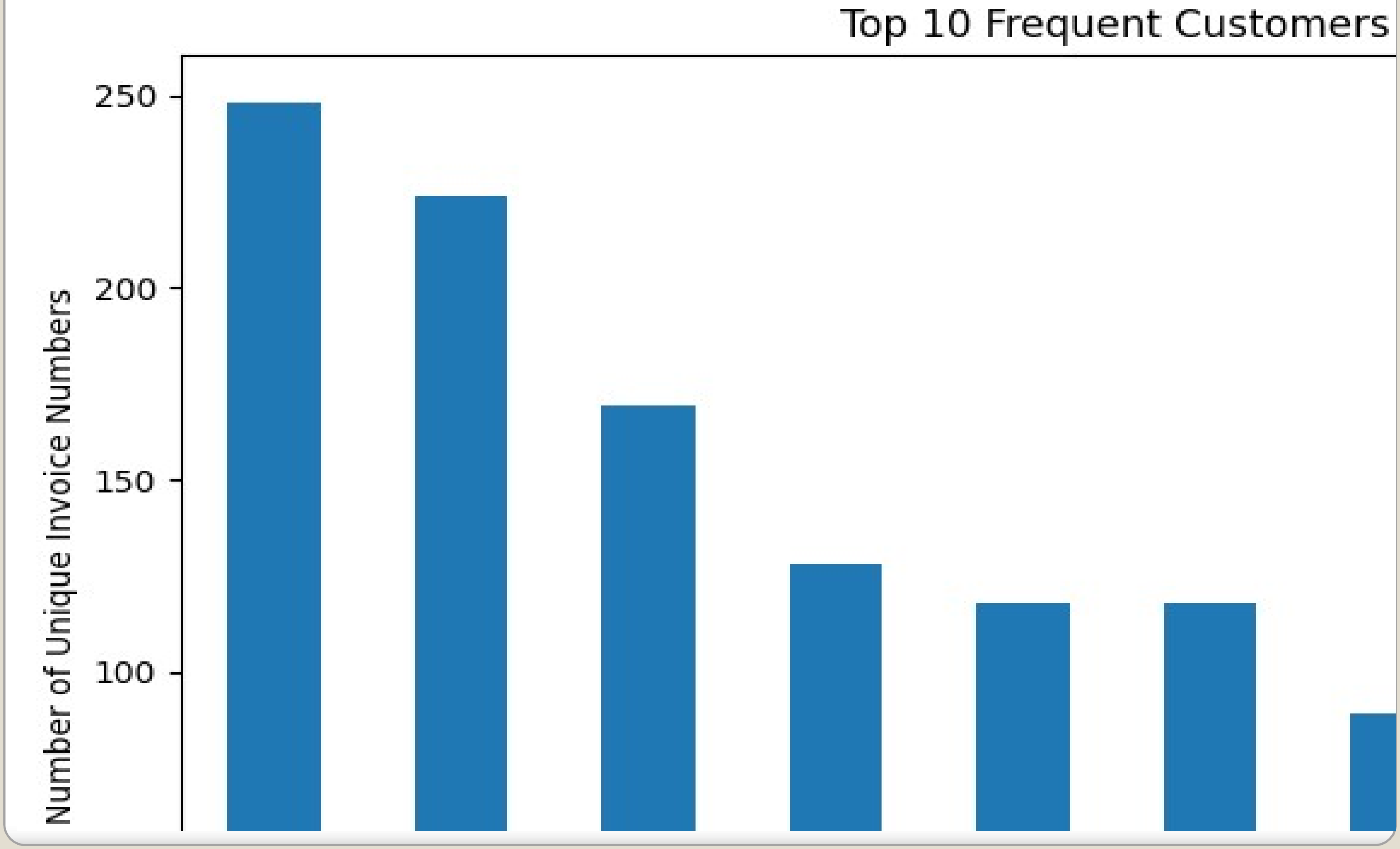
Exploratory Data Analysis



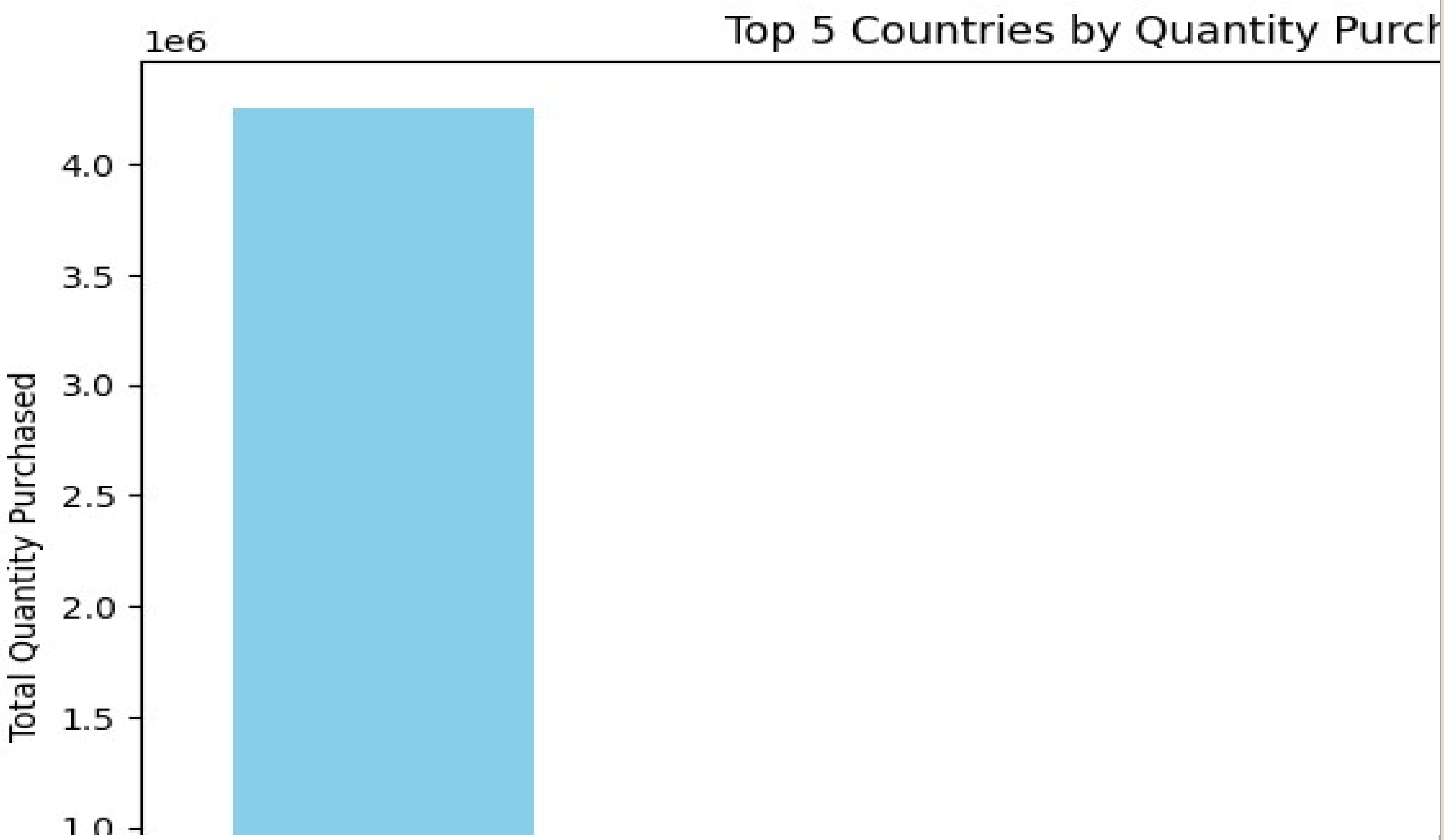
Exploratory Data Analysis



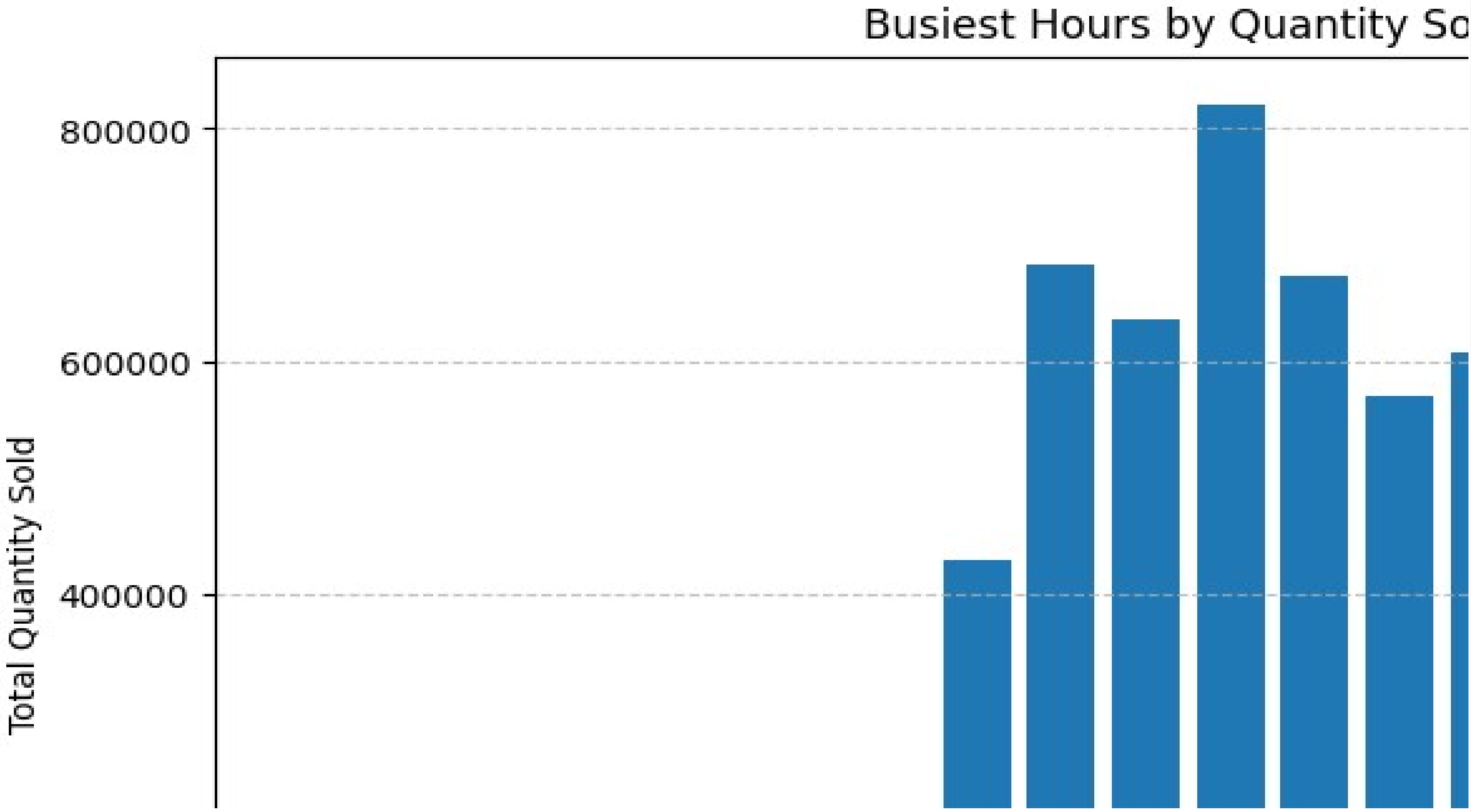
Exploratory Data Analysis



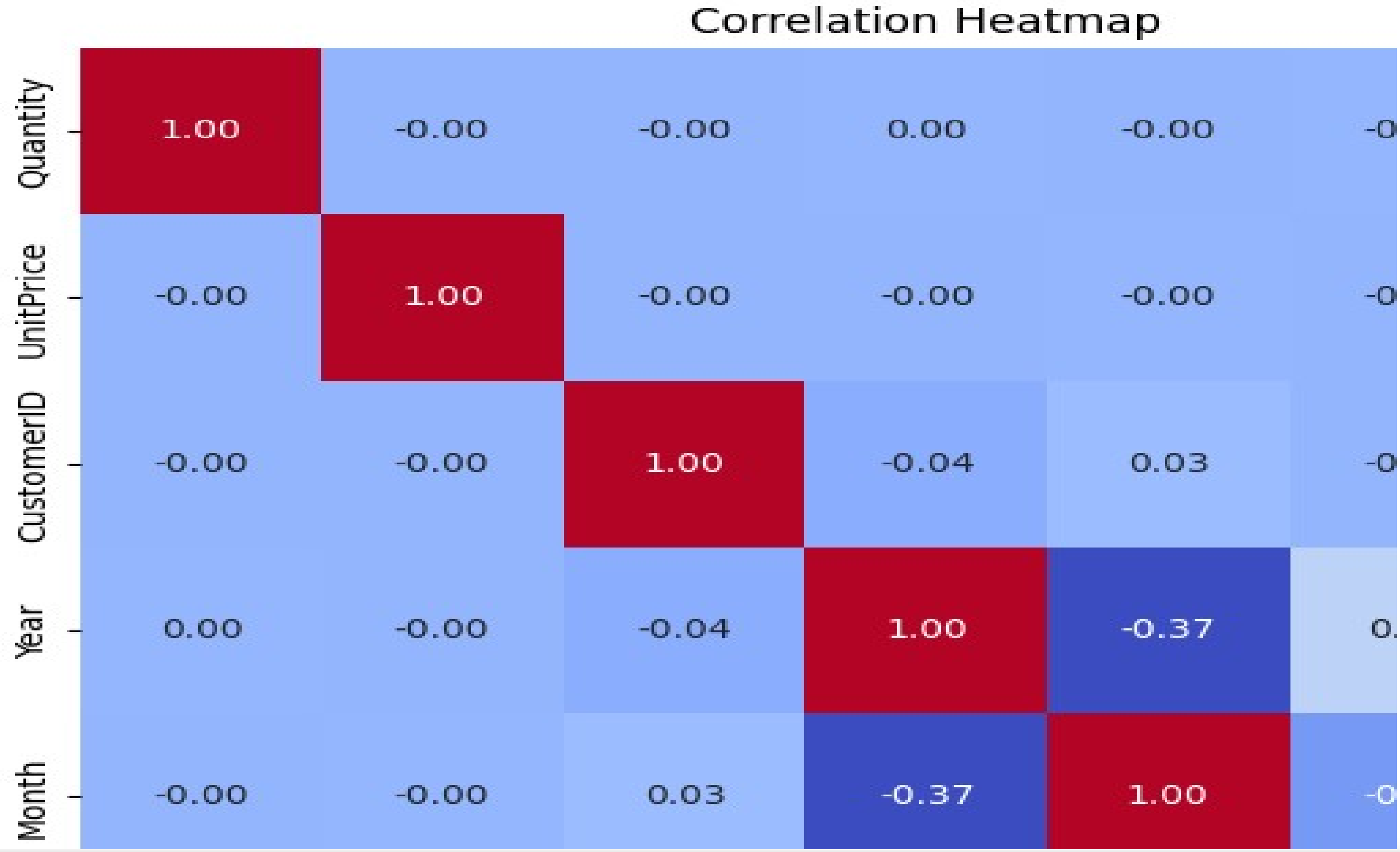
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Hypothetical Statement

Statistical Test Used: Two-sample t-test Compares the means of two independent samples to assess if there is a significant difference between them.

Hypothetical Statement 1:

Statement: Investigating if there is a difference in mean unit price between orders placed before and after noon.

Test Procedure: Conducted a two-sample t-test on unit prices of orders before noon and after noon.

Interpretation: If p-value < 0.05, we reject the null hypothesis and conclude a significant difference.

Result: We fail to reject the null hypothesis that there is no difference in mean unit price between orders placed before and after noon. (p > 0.05)

Hypothetical Statement 2:

Statement: Examining the difference in mean unit price between orders placed before and after noon.

Test Procedure: Employed a two-sample t-test on unit prices of orders before and after noon.

Interpretation: If p-value < 0.05, we reject the null hypothesis and assert a significant difference.

Result: We fail to reject the null hypothesis that there is no difference in mean unit price between orders placed before and after noon. (p > 0.05)

Feature Engineering & Data Pre-processing

Handling Missing Values

- **Identification:**
  - Initially, we identified missing values in the dataset using the isnull() function.
  - The CustomerID column had a significant portion of missing values, accounting for approximately 25.16% of the data.
- **Technique Used:**
  - The missing values in the CustomerID column were addressed by removing the entire rows where the CustomerID is missing.
  - This approach was chosen because the CustomerID is crucial for customer-specific analysis and segmentation, making rows without a CustomerID irrelevant for such analysis.
- **Implementation:**
  - After applying the removal technique, we checked the shape of the updated dataframe to ensure that missing values were properly handled.
  - The dataset's shape was significantly reduced after removing rows with missing CustomerID values, demonstrating the effectiveness of this technique.
- **Cancelled Orders Handling:**
  - Additionally, we identified and removed rows representing cancelled orders from the dataset.
  - Cancelled orders, typically denoted by negative quantities or an 'C' prefix in the InvoiceNo column, were deemed irrelevant for analysis as they may not represent actual sales transactions and could skew the analysis results.

Feature Engineering & Data Pre-processing

Handling Outliers

- **Identification:**
  - Outliers were identified by visualizing the distribution of Quantity and UnitPrice using density plots.
- **Outlier Treatment Technique:**
  - Winsorization technique was applied to address outliers.
  - Winsorization replaces extreme values in the dataset with less extreme values based on specified percentiles, thereby reducing the influence of outliers without removing them entirely.
- **Reasons for Using Winsorization:**
  - **Data Preservation:** Winsorization preserves the original data distribution by modifying extreme values, ensuring that the overall shape of the distribution remains intact.
  - **Robustness:** Winsorization is a robust technique that is less sensitive to extreme values compared to other methods, maintaining data integrity.
  - **Applicability:** Winsorization can be applied to various types of data distributions, making it a versatile technique for outlier treatment across different scenarios.
- **Summary:**

Outliers were effectively treated using the Winsorization technique, ensuring data integrity and reliability for subsequent analysis and modeling tasks.

Data Transformations

- **Feature Engineering:**
  - Utilized RFM (Recency, Frequency, Monetary) analysis to create new features that capture important customer behavior metrics.
  - **Recency:** Measures how recently a customer made a purchase, providing insights into their engagement level.
  - **Frequency:** Indicates how often a customer makes purchases, reflecting their loyalty and buying habits.
  - **Monetary Value:** Represents the total amount of money spent by a customer, indicating their value to the business.
- **RFM Calculation:**
  - Calculated recency, frequency, and monetary metrics for each customer based on their transaction history.
  - Grouped customers by their RFM metrics to segment them into different groups, enabling targeted marketing strategies.
- **Feature Selection:**
  - Chose RFM metrics as features for analysis and modeling due to their relevance in understanding customer behavior and segmentation.
  - These features provide valuable insights into customer engagement, loyalty, and purchasing patterns, facilitating effective decision-making in marketing and sales.

Data Scaling

- Utilized StandardScaler method to scale the data.
- Standardization removes the mean and scales the data to unit variance.
- Ensures that each feature has a mean of 0 and a standard deviation of 1.
- Commonly used in machine learning to improve algorithm performance, particularly for methods like K-means clustering that rely on distance calculations.

The data underwent scaling using the Standard Scaler method to standardize features and improve the performance of clustering algorithms such as K-means. This transformation ensures that features are on a comparable scale, enabling more accurate and reliable analysis and modeling.

ML Model Implementation
K-Means Clustering

- K-Means clustering is a popular unsupervised machine learning algorithm used for clustering data into distinct groups based on similarities.

- It partitions data into 'k' clusters, where each data point belongs to the cluster with the nearest mean.

- K-Means is widely used for customer segmentation, anomaly detection, and image segmentation, among other applications.

ML Model Implementation K-Means Clustering

- The Silhouette Score is employed as the evaluation metric to assess the quality of clusters generated by the KMeans algorithm. This metric quantifies the cohesion and separation of data points within clusters, ranging from -1 to 1, with higher scores indicating better-defined clusters.

- The code iterates over a range of potential cluster numbers, specifically from 2 to 15 clusters. For each number of clusters, KMeans clustering is applied, and the corresponding Silhouette Score is computed to gauge the clustering performance.

- By analyzing the Silhouette Scores obtained for different numbers of clusters, the code assists in determining the optimal number of clusters for the dataset. This aids in selecting the most suitable clustering configuration that maximizes cluster coherence and separation.

ML Model Implementation
K-Means Clustering

- **Silhouette Score Method on Recency frequency and Monetary**
  - For n_clusters = 2, silhouette score is 0.4186
  - For n_clusters = 3, silhouette score is 0.3419
  - For n_clusters = 4, silhouette score is 0.3393
  - For n_clusters = 5, silhouette score is 0.3485
  - For n_clusters = 6, silhouette score is 0.3349
  - For n_clusters = 7, silhouette score is 0.3379
  - For n_clusters = 8, silhouette score is 0.3466
  - For n_clusters = 9, silhouette score is 0.3409
  - For n_clusters = 10, silhouette score is 0.3475
  - For n_clusters = 11, silhouette score is 0.3513
  - For n_clusters = 12, silhouette score is 0.3491
  - For n_clusters = 13, silhouette score is 0.3462
  - For n_clusters = 14, silhouette score is 0.3559
  - For n_clusters = 15, silhouette score is 0.3522

•The Silhouette scores range from around 0.33 to 0.41, indicating moderate cluster quality.

•The scores suggest that while there is some clustering structure present, it may not be very distinct or well-separated.

•Customers are segmented based on their recency of purchase and frequency of transactions.

•This feature combination may not capture all dimensions of customer behavior, leading to less distinct clusters.

ML Model Implementation
K-Means Clustering

- **Silhouette Score Method on Recency and Monetary**
  - For n_clusters = 2, silhouette score is 0.4244
  - For n_clusters = 3, silhouette score is 0.3544
  - For n_clusters = 4, silhouette score is 0.3275
  - For n_clusters = 5, silhouette score is 0.3513
  - For n_clusters = 6, silhouette score is 0.3324
  - For n_clusters = 7, silhouette score is 0.3460
  - For n_clusters = 8, silhouette score is 0.3467
  - For n_clusters = 9, silhouette score is 0.3444
  - For n_clusters = 10, silhouette score is 0.3485
  - For n_clusters = 11, silhouette score is 0.3476
  - For n_clusters = 12, silhouette score is 0.3457
  - For n_clusters = 13, silhouette score is 0.3522
  - For n_clusters = 14, silhouette score is 0.3555
  - For n_clusters = 15, silhouette score is 0.3542

•The Silhouette scores range from approximately 0.32 to 0.42.

•Similar to the Recency & Frequency features, the scores suggest moderate cluster quality.

•Customers are segmented based on their recency of purchase and the monetary value of transactions.

•While there is some clustering structure, it may not be as clear as desired, indicating potential overlap or mixed behavior among segments.

ML Model Implementation
K-Means Clustering

- **Silhouette Score Method on Frequency and Monetary**
  - For n_clusters = 2, silhouette score is 0.5528
  - For n_clusters = 3, silhouette score is 0.5292
  - For n_clusters = 4, silhouette score is 0.5207
  - For n_clusters = 5, silhouette score is 0.5308
  - For n_clusters = 6, silhouette score is 0.5279
  - For n_clusters = 7, silhouette score is 0.5189
  - For n_clusters = 8, silhouette score is 0.5288
  - For n_clusters = 9, silhouette score is 0.5231
  - For n_clusters = 10, silhouette score is 0.5325
  - For n_clusters = 11, silhouette score is 0.5335
  - For n_clusters = 12, silhouette score is 0.5364
  - For n_clusters = 13, silhouette score is 0.5402
  - For n_clusters = 14, silhouette score is 0.5425
  - For n_clusters = 15, silhouette score is 0.5421

•The Silhouette scores are notably higher, ranging from approximately 0.52 to 0.55.

•These scores indicate better-defined clusters with higher cohesion and separation.

•Customers are segmented based on their purchasing frequency and monetary spending.

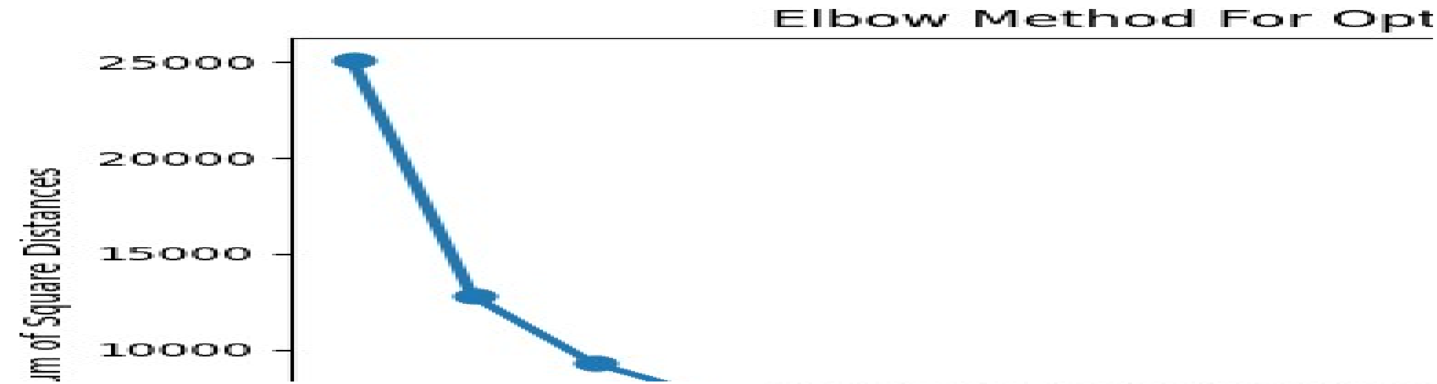
•This feature combination seems to capture more distinct patterns in customer behavior, resulting in clearer and more meaningful clusters.

ML Model Implementation K-Means Clustering

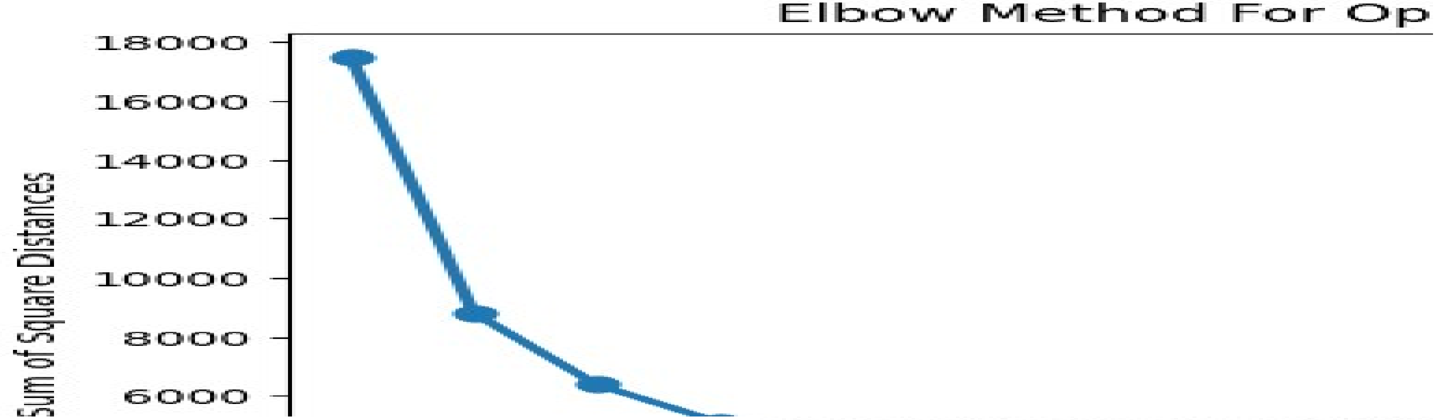
- **Elbow Method**
  - The elbow method is a heuristic used to determine the optimal number of clusters in a dataset.
  - It is based on the premise that as the number of clusters increases, the within-cluster sum of squares (WCSS) decreases.
  - In the project, the elbow method was applied to determine the optimal number of clusters for customer segmentation.
  - With more clusters in K-means, the total distance within clusters (WCSS) drops steadily, creating a smoother curve.
  - The "elbow point" is the point on the curve where the rate of decrease in WCSS slows down significantly.
  - The optimal number of clusters is often chosen at the elbow point.

Finding Optimal k

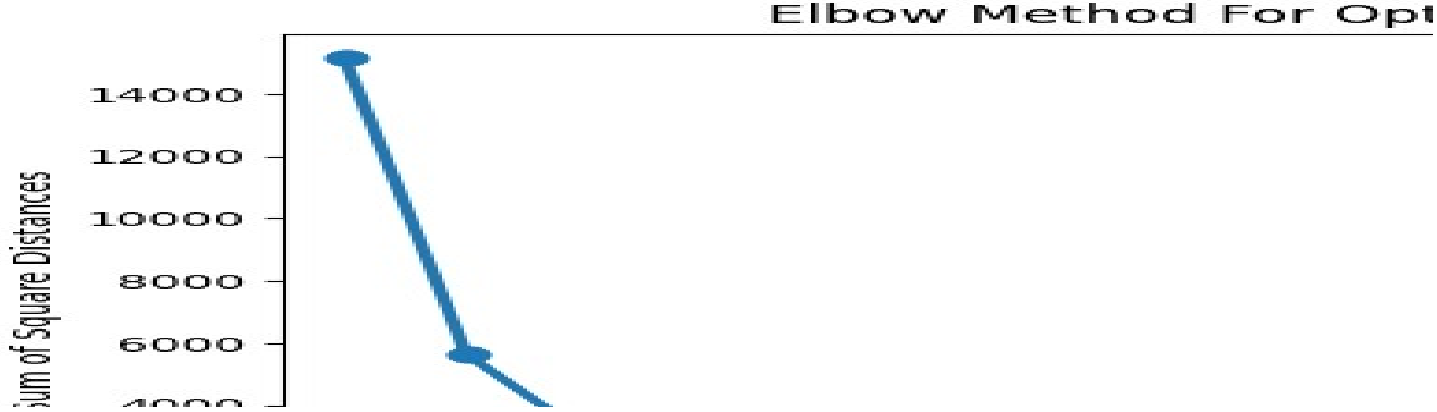
Recency & Frequency



Recency & Monetary



Frequency & Monetary



ML Model Implementation K-Means Clustering

- **Recency and Frequency:**
  - This a clear elbow at k = 3, suggesting that three clusters might be optimal for segmenting customers based on these two features.
  - A smaller number of clusters (k = 3) for recency and frequency features implies that customers can be grouped into three distinct segments based on their recent purchase behavior and frequency of purchases.
- **Recency and Monetary:**
  - The elbow occurs at k = 10, indicating that segmenting customers into ten clusters might be appropriate for this feature combination.
  - A larger number of clusters (k = 10) for recency and monetary features suggests a finer segmentation, possibly capturing variations in customers' purchasing patterns and monetary value.
- **Frequency and Monetary:**
  - There is a less distinct elbow. The elbow occurs at k = 3 or k = 9, suggesting that either three or nine clusters could be considered.
  - The less distinct elbow for frequency and monetary features indicates some ambiguity in determining the optimal number of clusters. Further analysis or domain knowledge might be needed to finalize the clustering strategy.

ML Model Implementation DBSCAN

- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm.

- Unlike K-means, DBSCAN does not require the number of clusters as input. Instead, it groups together closely packed points based on two parameters: epsilon (ε), which defines the radius of the neighborhood around a point, and min_samples, which specifies the minimum number of points required to form a dense region (cluster).

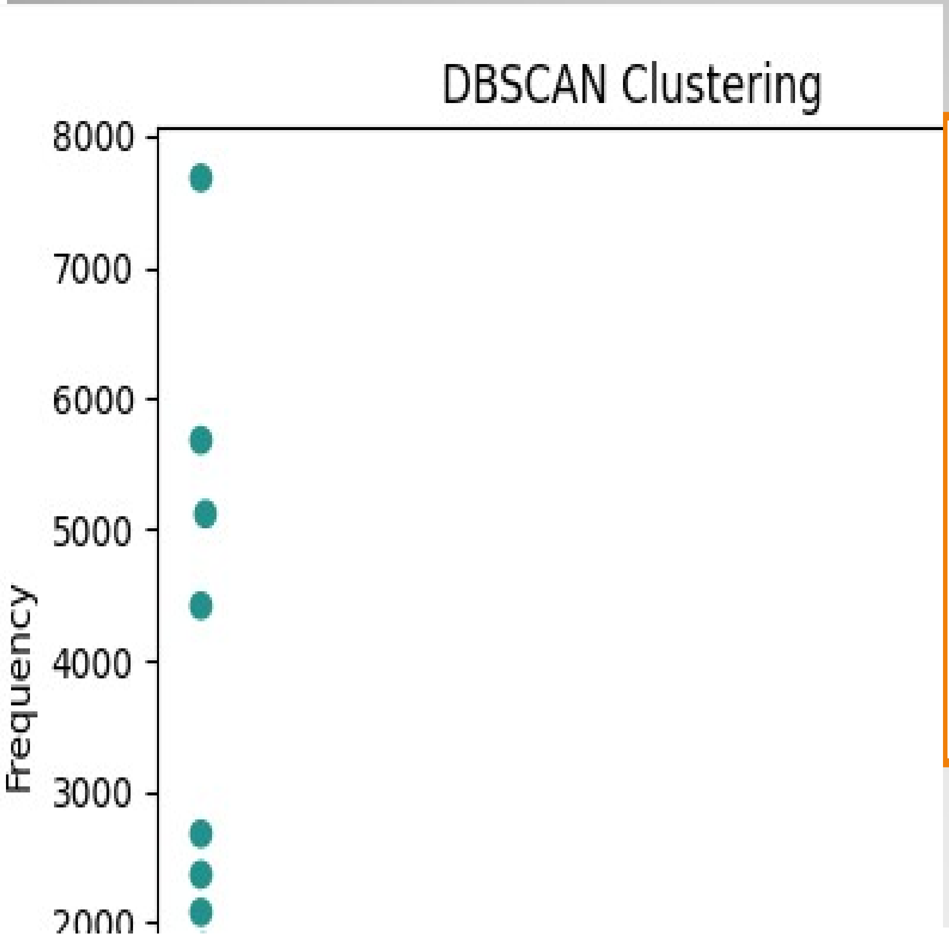
- DBSCAN is effective in identifying clusters of arbitrary shapes and sizes, making it suitable for datasets with complex geometric structures.

- DBSCAN can be used to identify groups of customers with similar purchase behaviors or preferences.

- It is particularly useful when the clusters exhibit varying densities or irregular shapes, as DBSCAN can adapt to such structures.

- By identifying dense regions of customers, DBSCAN can help businesses uncover meaningful segments for targeted marketing strategies or personalized recommendatio

ML Model Implementation DBSCAN

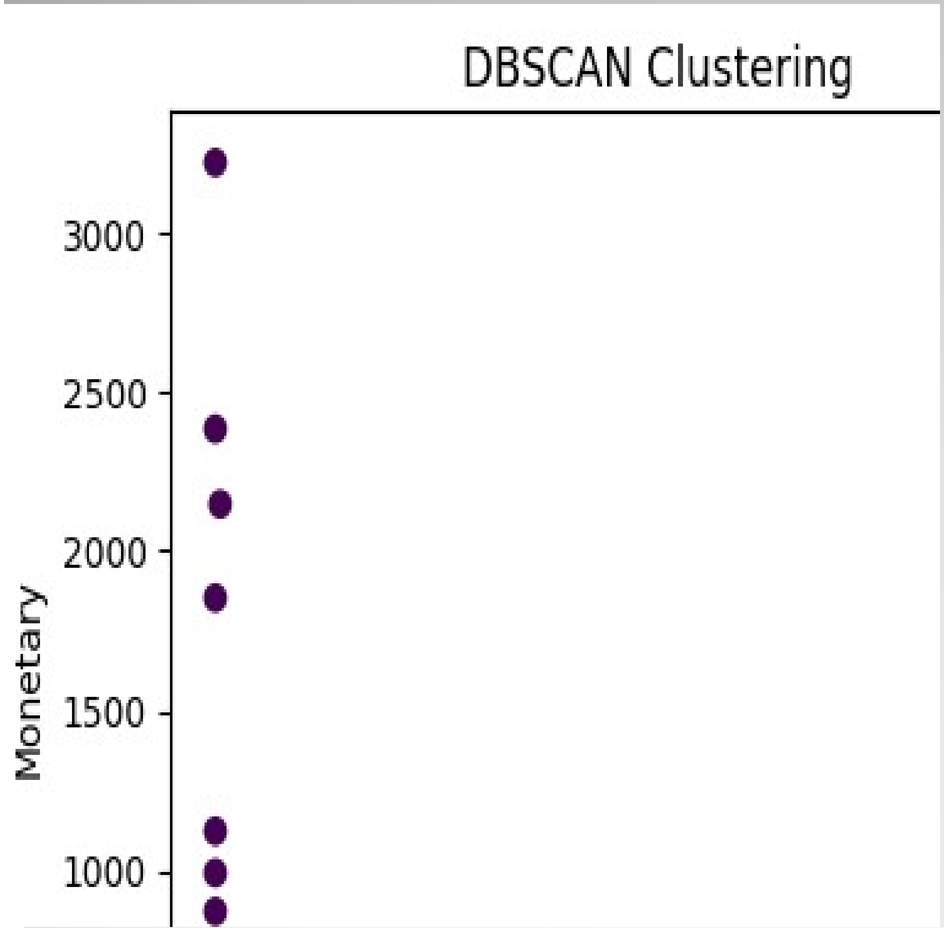


Recency VS Frequency

•The cluster primarily consists of customers with high recency (frequent purchases) and varying frequency levels.

•It potentially represents loyal customers who make purchases regularly, regardless of the purchase interval.

ML Model Implementation DBSCAN



Recency VS Monetary

•This cluster mainly consists of customers with low recency (less frequent purchases) and potentially lower frequency as well.

•It could represent infrequent or lapsed customers who rarely make purchases.

ML Model Implementation Hierarchial Clustering

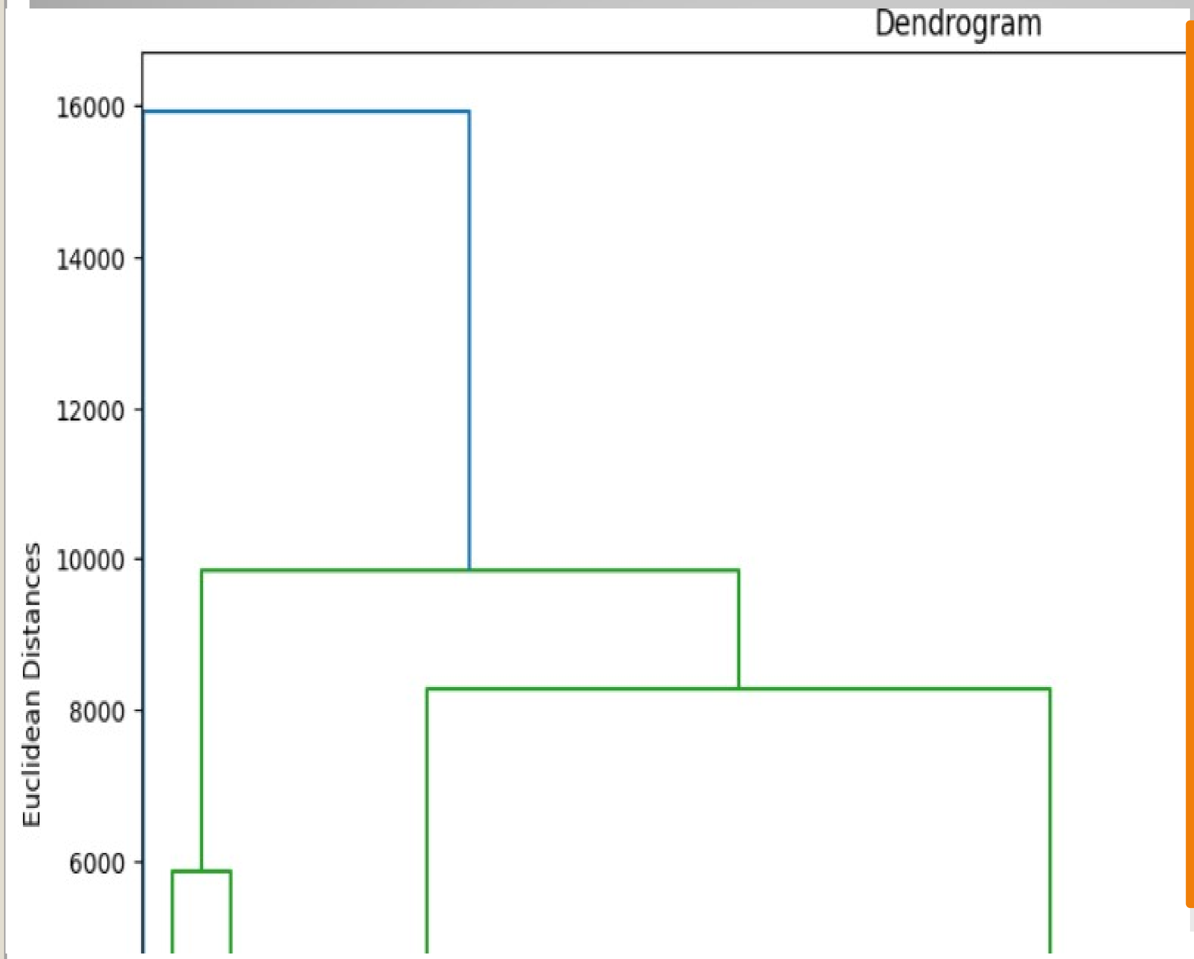
- Hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters. It seeks to group similar data points into clusters based on their distances or similarities.

- One of the key visualizations in hierarchical clustering is the dendrogram, which represents the merging process of clusters.

- The dendrogram displays the hierarchy of clusters and illustrates the order in which clusters are merged or split.

- It allows analysts to visualize the relationships between clusters and helps in determining the optimal number of clusters by observing the structure of the dendrogram.

ML Model Implementation Hierarchial Clustering



•The dendrogram displays several branches, indicating multiple potential clusterings of customers based on their recency and frequency behaviors.

•The vertical axis represents the **Euclidean distances** between clusters, with higher values signifying larger dissimilarity.

•Shorter branch lengths suggest closer relationships between clusters, while longer branches indicate more distinct clusters.

Contd…

Insights:

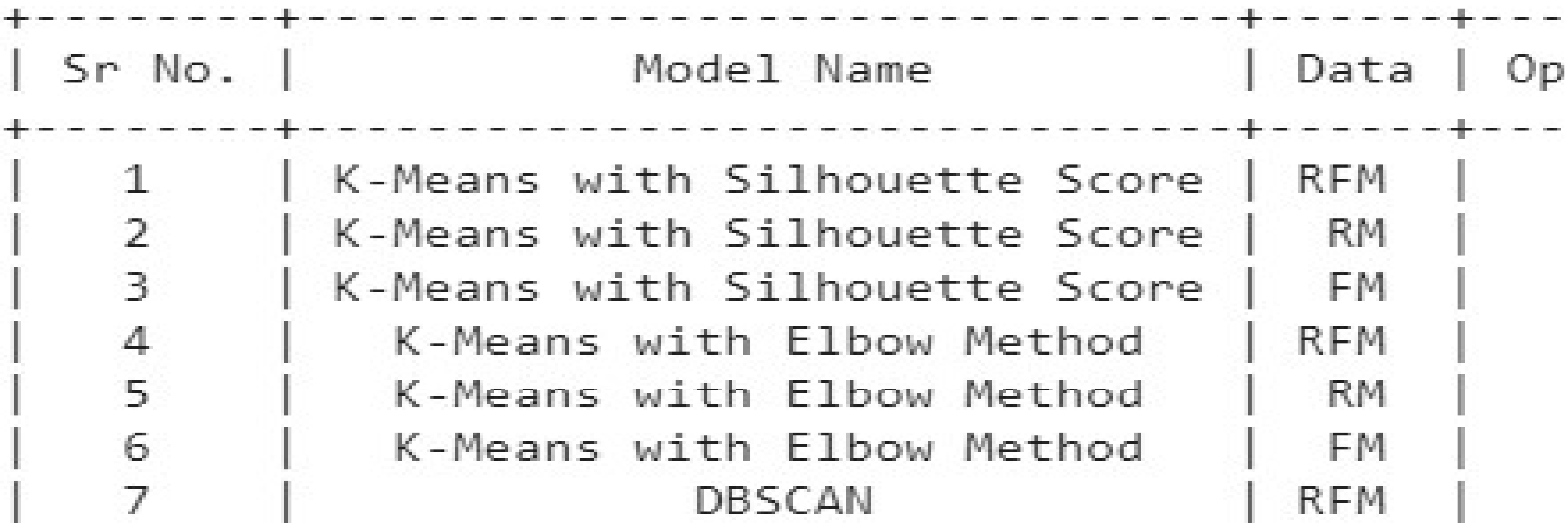
- **Two main customer groups:** The initial split separates customers into two main groups:
  - **Group 1 (left side):** This group seems to contain customers with **more similar recency and frequency patterns**, potentially representing **more frequent or regular buyers**.
  - **Group 2 (right side):** This group encompasses customers with **more diverse recency and frequency behaviors**, potentially including **infrequent or irregular buyers**.
- **Sub-clusters within groups:** Further sub-divisions within each group reveal smaller clusters with potentially more nuanced differences in recency and frequency.

For example:

Within Group 1: There might be a cluster of **highly frequent buyers** and another of **moderately frequent buyers**.

Within Group 2: There could be clusters of **infrequent buyers** with varying levels of recency.

Metric Scores



```
+---------+--------------------------------+-------+---
| Sr No. |           Model Name           | Data | Op
+---------+--------------------------------+-------+---
|   1    | K-Means with Silhouette Score | RFM  |
|   2    | K-Means with Silhouette Score |  RM  |
|   3    | K-Means with Silhouette Score |  FM  |
|   4    |   K-Means with Elbow Method   | RFM  |
|   5    |   K-Means with Elbow Method   |  RM  |
|   6    |   K-Means with Elbow Method   |  FM  |
|   7    |             DBSCAN             | RFM  |
```

Based on these metrics, the best cluster would be to use K-Means with 2 clusters obtained through either Silhouette Score or the Elbow Method, as it provides consistent results across all feature combinations and is widely used for its simplicity and effectiveness in clustering tasks.

Conclusion

- *   Customer segmentation is a crucial aspect of retail business strategy, enabling personalized marketing and enhanced customer experiences.
- *   Through exploratory data analysis (EDA), we gained insights into customer purchasing behavior, including top-selling products, customer demographics, and peak sales periods.
- *   Utilizing machine learning techniques such as K-Means clustering, we segmented customers based on their recency, frequency, and monetary value of purchases.
- *   The optimal number of clusters, determined through metrics like Silhouette Score and the Elbow Method, allowed us to effectively group customers into distinct segments.
- *   By leveraging clustering algorithms, businesses can tailor marketing campaigns, improve customer retention, and optimize product recommendations to meet the diverse needs of different customer segments.
- *   Overall, this project demonstrates the value of data-driven approaches in understanding customer behavior and driving business growth in the retail sector.

Actionable Insights and Recommendations

- **Customer Segmentation**: Utilize the identified clusters to segment customers based on their purchasing behavior, recency, and monetary value. Tailor marketing strategies, promotions, and communication channels to better meet the needs and preferences of each segment.

- **Retention Strategies**: Identify high-value customer segments with high recency and monetary value. Implement targeted retention strategies, such as loyalty programs, personalized offers, and exceptional customer service, to retain these valuable customers and encourage repeat purchases.

- **Acquisition Channels**: Analyze the characteristics of different customer segments to understand which acquisition channels are most effective for each segment. Allocate marketing budgets more efficiently by investing in channels that yield higher conversion rates and customer lifetime value.

- **Product Recommendations**: Leverage customer segmentation to provide personalized product recommendations and cross-selling opportunities. Use insights from cluster analysis to understand which products are

THANK YOU