Customer Segmentation Analysis Using K-Means Clustering

1. Introduction

Customer segmentation is an important technique in e-commerce that enables organizations to recognize and respond to different customer categories based on their purchase patterns. Using the K-Means clustering technique, the segmentation procedure here focuses on three key statistics: "Total Spend," "Frequency," and "Recency" (time since the last purchase). Corporations can create focused marketing plans that improve consumer engagement and increase revenue with the help of these indicators.

This analysis into the segmentation process, describing procedures such as data loading, preprocessing, exploratory data analysis (EDA), K-Means clustering, and visualization of results in both 2D and 3D formats.

2. Methodology and Analysis

Step 1: Data Loading

The dataset was loaded into 'pandas' for data exploration and manipulation. It was simple to load CSV files into a DataFrame for more analysis.

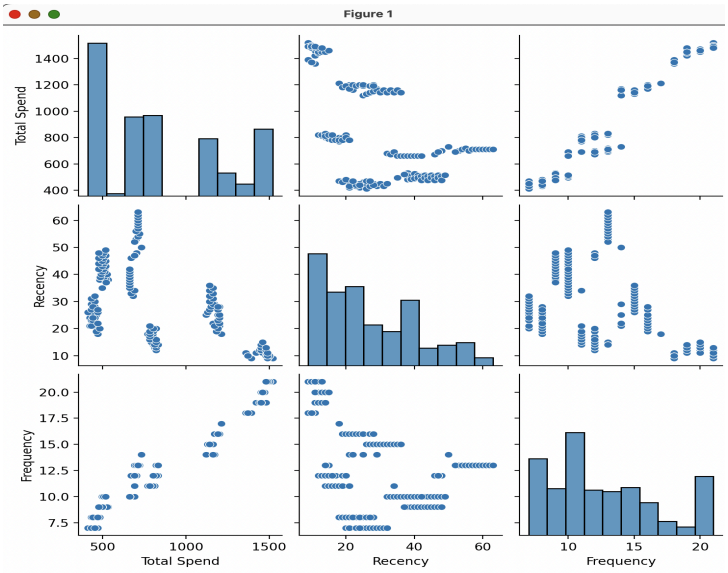
Step 2: Data Exploration

Initial data exploration revealed details about the dataset's structure, missing values, and descriptive statistics for each variable. This early stage is critical for detecting potential problems and understanding data distribution.

3. Exploratory Data Analysis (EDA)

EDA is essential for determining the underlying patterns and correlations in data. In this section, examine some of the primary visualizations created during the analysis and how they were interpreted.

Figure 1: Pair Plot Analysis



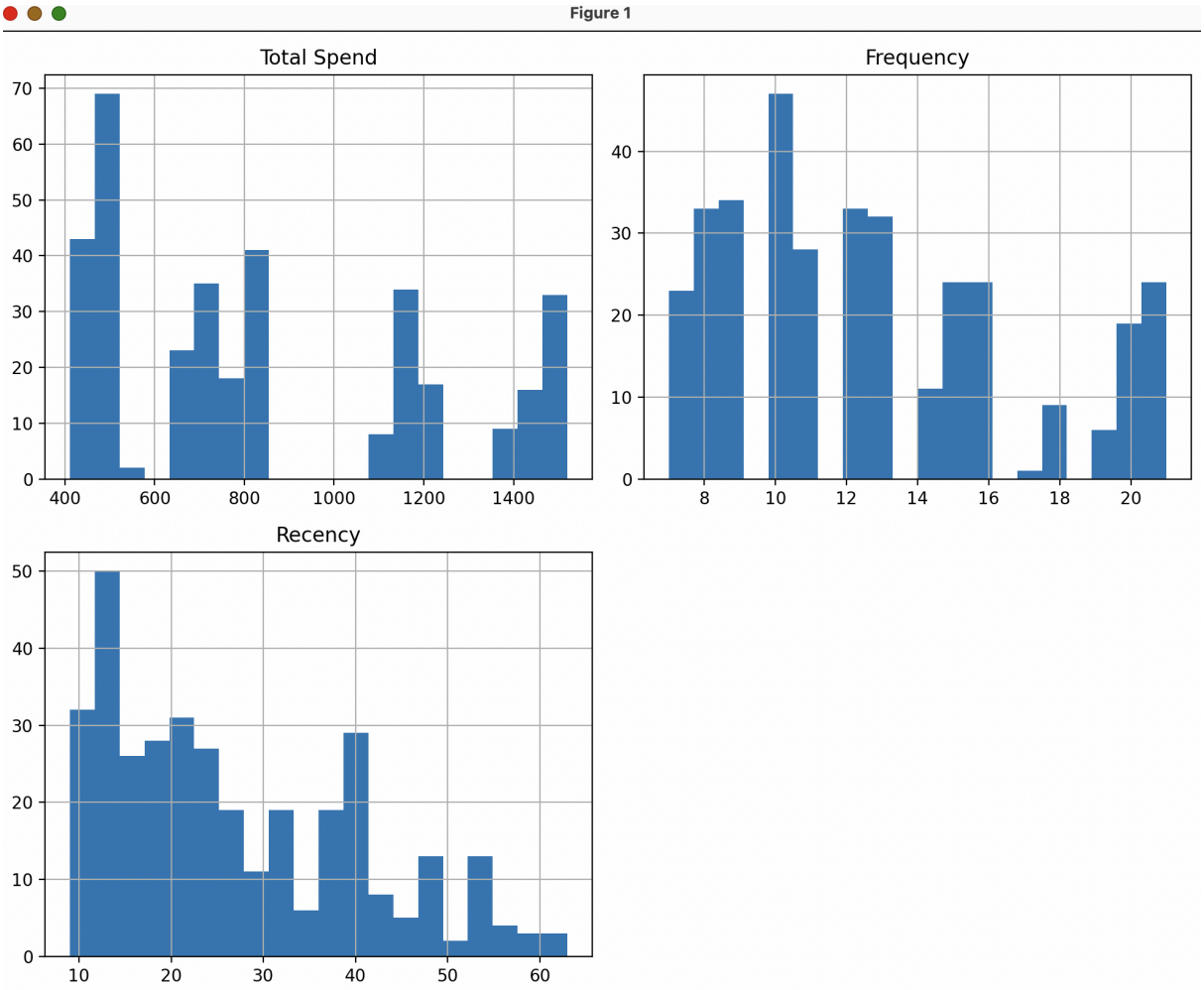
The pair plot generates a matrix of scatter plots showing the relationships between "Total Spend," "Frequency," and "Recency." Each point represents a different client, and the plots show how these attributes relate to one another.

- The pair plot shows clustering patterns, with distinct point clusters visible along particular combinations.

- The correlation between "Total Spend" and "Frequency" indicates that customers who spend more are likely to buy more frequently. In e-commerce, where regular customers typically donate more money, this beneficial correlation is in line with regular consumer behavior.

- The "Recency" feature has wider correlations with "Total Spend" and "Frequency", implying that recency alone may not predict spending or frequency as effectively.

Figure 2: Histograms of Total Spend, Frequency, and Recency



These histograms show the frequency distribution of each essential parameter (for example, "Total Spend", "Frequency", and "Recency"). Each bar represents the number of customers within particular value ranges for each feature.

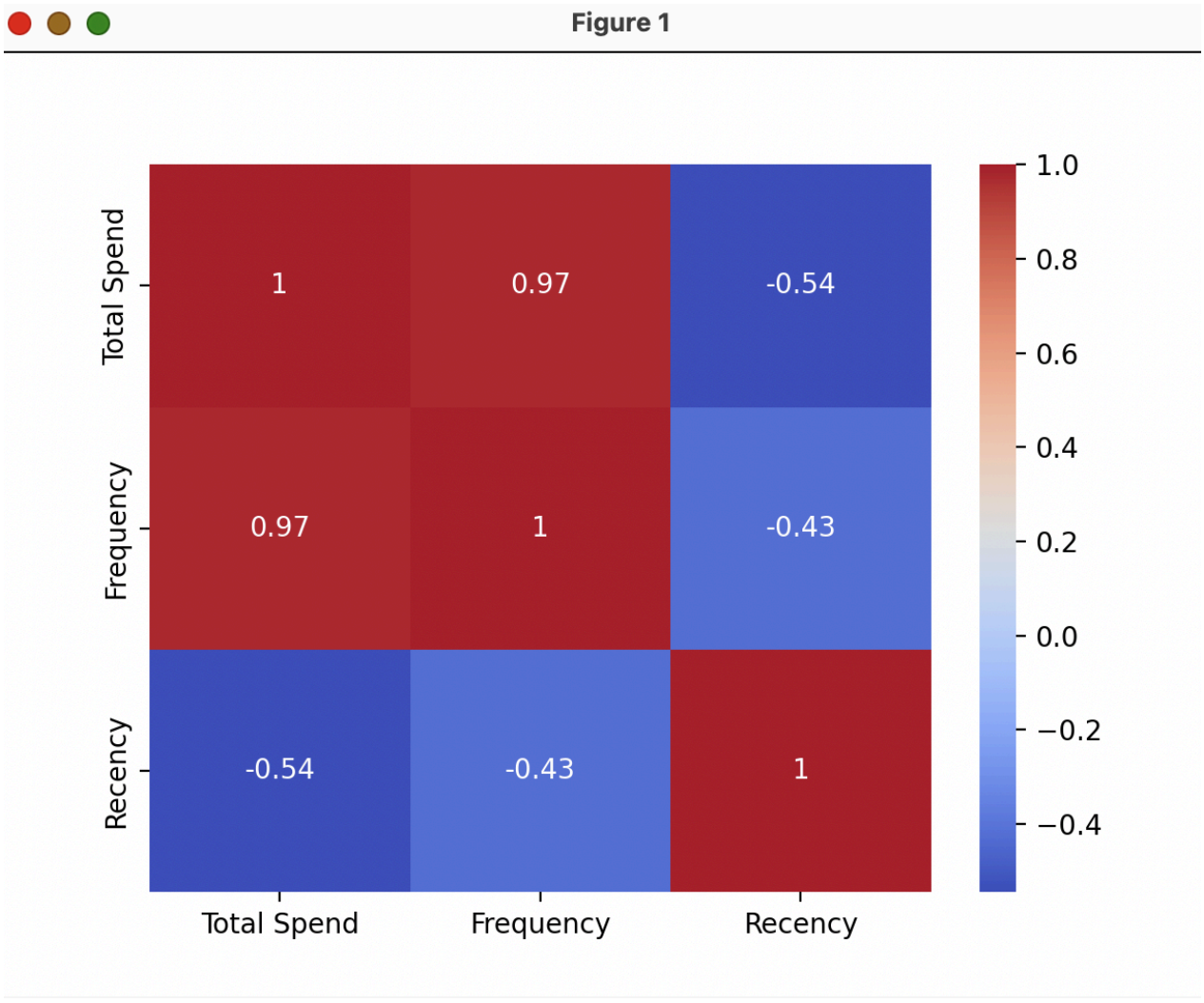
- "Total Spend": The distribution is diverse, with peaks at various spending levels, showing different groups of low, middle, and high spenders.

- "Frequency": There is a wide variation of purchase frequencies, indicating that some customers purchase regularly while others buy occasionally.

"Recency": This feature displays a distribution that has been shifted to the right, with most consumers having made recent transactions within a certain time period. This may indicate a high level of recent loyalty or its capacity to bring in new customers.

These histograms give a basic picture of how customers differ in terms of spending, frequency, and recency, which is important for clustering.

Figure 3: Correlation Heatmap



The heat map shows the correlation factors between "Total Spend", "Frequency", and "Recency", with colors indicating the intensity and pattern of relationships. Positive correlations are displayed in red, and negative correlations in blue.

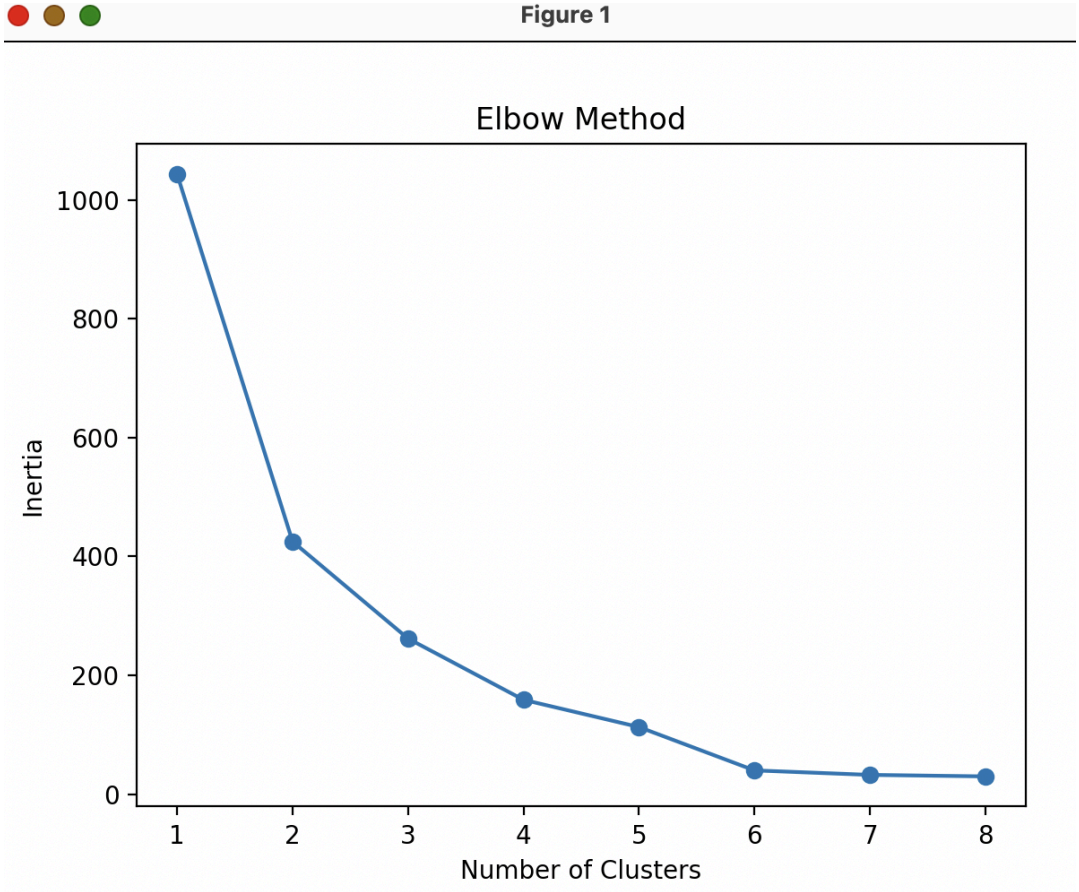
- "Total Spend" and "Frequency" have a strong positive connection (0.97), indicating that customers who spend more buy more frequently. This shows that "Total Spend" and "Frequency" are inextricably related, possibly describing one of the crucial segmentation angles.

- "Recency" shows an average negative correlation with both "Total Spend" (-0.54) and "Frequency" (-0.43), implying that customers who have just purchased were considerably more likely to be regular, significant consumers. However, this relationship is weaker than the one between "Total Spend" and "Frequency".

The heatmap confirms that "Total Spend" and "Frequency" are critical for clustering, with "Recency" adding an additional layer of differences.

4. K-Means Clustering and Model Selection

Figure 4: Elbow Method Plot



The Elbow Method graphic depicts the within-cluster sum of squares (neutrality) for various values of "K". The inertia for a specific number of clusters is represented by each point on the line.

- The ideal number of clusters is shown by the plot's "elbow" point, which is located at "K = 4". Beyond this point, adding more clusters has little effect on inertia, implying that clustering tightness is not significantly improved.

- This ideal "K = 4" indicates that segmenting customers into four unique groups provides a balanced approach for collecting variation without overfitting.

By choosing four clusters, we can effectively categorize clients based on spending, frequency, and recency without increasing model complexity.

Applying K-Means Clustering with Optimal Clusters

After determining the optimal number of clusters, we applied K-Means with “k=4”. Each customer was assigned a cluster label, allowing for further analysis of each group's characteristics.

5. Cluster Visualization

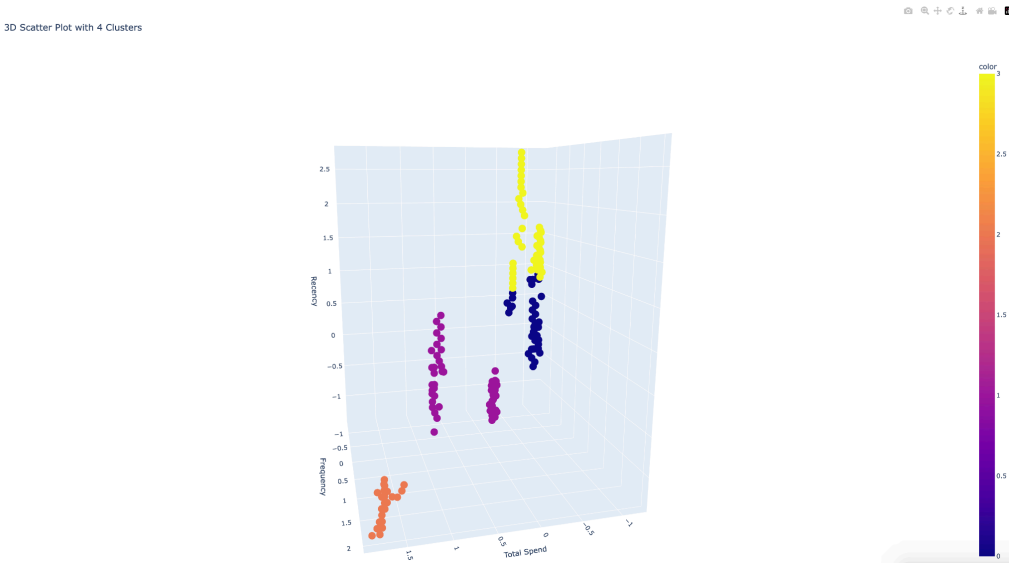
2D Scatter Plot of Clusters

The 2D scatter plot illustrates groups of consumers based on two of the specified features (for example, "Total Spend" vs "Frequency"). Each point represents a consumer, with different colors indicating the four clusters.

- The scatter plot shows the pattern of distribution and spacing of clusters in a 2D space, confirming that the clustering method successfully distinguished groups.

- Distinct color-coded clusters suggest that the algorithm has segmented customers based on significant variations in spending and how often they purchase.

Figure 5: 3D Scatter Plot of Clusters



This 3D scatter plot shows the customer groups across three key dimensions ("Total Spend", "Frequency", and "Recency"), providing a more complete picture of the segmentation.

- The 3D plot shows how each cluster utilizes various areas inside the three-dimensional feature space, showing the differences in spending, purchasing frequency, and recency between clusters.

- Interactive visualization allows cluster transformation and exploration from different points of view, providing more insight into customer behavior. For example, one cluster may contain high-frequency, high-spend consumers, whereas another may contain low-frequency, low-spend customers.

The 3D map helps to comprehend how each feature contributes to cluster formation, which validates the segmentation and provides useful information for targeted marketing.

6. Conclusion

This customer segmentation analysis successfully classified consumers based on "Total Spend", "Frequency", and "Recency", resulting in four distinct clusters. The data from each cluster allows for marketing tactics related to various customer behaviors:

- "High-Spending, Infrequent Buyers": Loyalty schemes may stimulate recurrent purchases.

- "Frequent, Low-Spending Buyers": Providing discounts may improve average spend per transaction.

- "Recent High-Spenders": Focused engagement to maintain interest and spending.

Understanding these groups allows businesses to execute specific strategies that increase customer retention and revenue. Further improvements to this analysis could include trying different clustering techniques or introducing new features for improved segmentation.