DIGITAL EMPOWERMENT PAKISTAN FOUNDATION

DEP

Customer Segmentation for a Retail Store

Ayesha Iqbal

DATA SCIENCE
TASK : 01

Introduction

In today's competitive retail environment, understanding customer behavior is essential for developing effective marketing strategies and improving customer retention. With consumers being more informed and selective than ever, retailers must leverage detailed insights into customer preferences and habits. One powerful method to achieve this understanding is through customer segmentation, which involves dividing a customer base into distinct groups based on specific characteristics or behaviors. Customer segmentation allows businesses to identify and analyze these distinct groups, providing a deeper understanding of customer needs and preferences. By utilizing segmentation, retailers can tailor their marketing efforts to target specific segments more effectively. This targeted approach can lead to more personalized promotions, which not only meet the unique needs of each customer group but also enhance customer satisfaction and loyalty. This report presents a comprehensive analysis of customer segmentation for a retail store using historical sales transaction data. Through this segmentation, the retail store can create more relevant and appealing promotions, fostering a stronger connection with their customers and ultimately driving higher retention rates.

Objectives

- Segment customers based on purchasing behavior using transaction data.
- Devise targeted marketing strategies.
- Offer personalized promotions.
- Improve customer retention.

1. Data Preparation

The dataset used for this analysis is loaded from a CSV file named *'Daily Household Transactions.csv'*. The initial few rows of the data are displayed to understand its structure.

	Date	Mode	Category	Subcategory	Note	Amount	Income/Expense	Currency
0	20/09/2018 12:04:08	Cash	Transportation	Train	2 Place 5 to Place 0	30.0	Expense	INR
1	20/09/2018 12:03:15	Cash	Food	snacks	Idli medu Vada mix 2 plates	60.0	Expense	INR
2	19/09/2018	Saving Bank account 1	subscription	Netflix	1 month subscription	199.0	Expense	INR
3	17/09/2018 23:41:17	Saving Bank account 1	subscription	Mobile Service Provider	Data booster pack	19.0	Expense	INR
4	16/09/2018 17:15:08	Cash	Festivals	Ganesh Pujan	Ganesh idol	251.0	Expense	INR

2. Data Cleaning and Preprocessing

- **Date Conversion:** The Date column is converted to datetime format to facilitate time-based analysis.
- **Missing Values:** Missing values in the Subcategory and Note columns are filled with 'Unknown'.
- **Column Selection:** Relevant columns (Date, Mode, Category, Subcategory, Amount, Income/Expense) are selected for further analysis.

3. Exploratory Data Analysis (EDA)

To understand the data better, several visualizations were created:

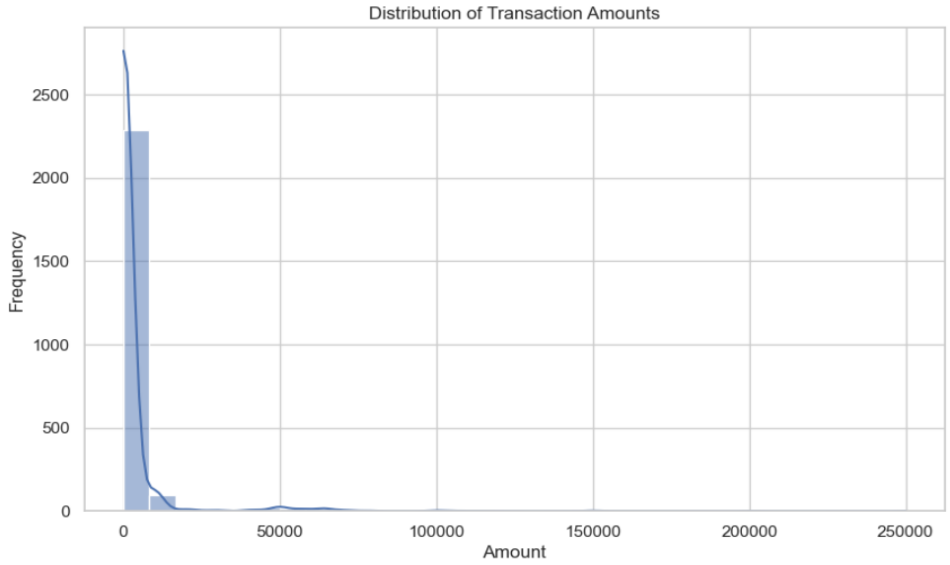


Figure 1: **A histogram illustrating the distribution of transaction amounts.**

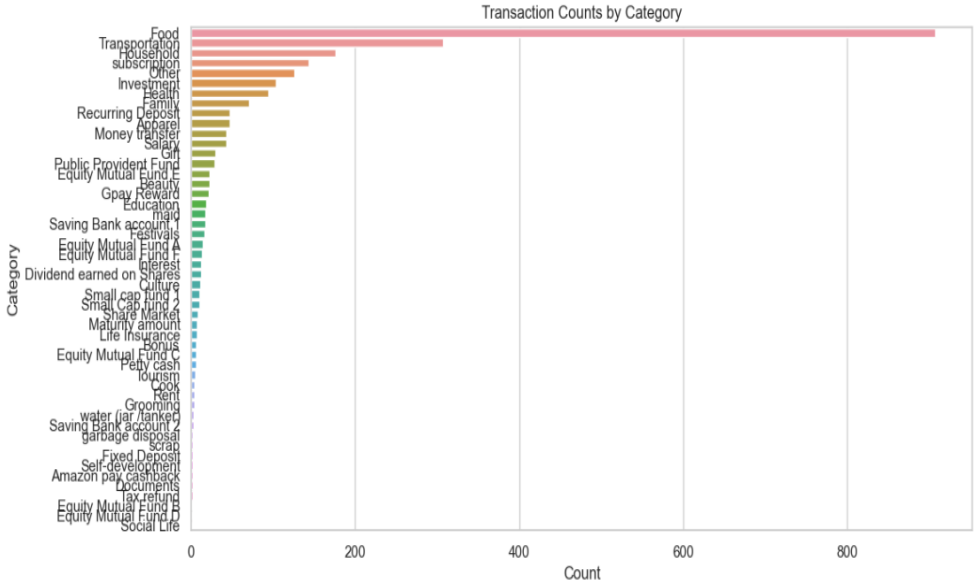


Figure 2: **A bar plot displaying the count of transactions for each category.**

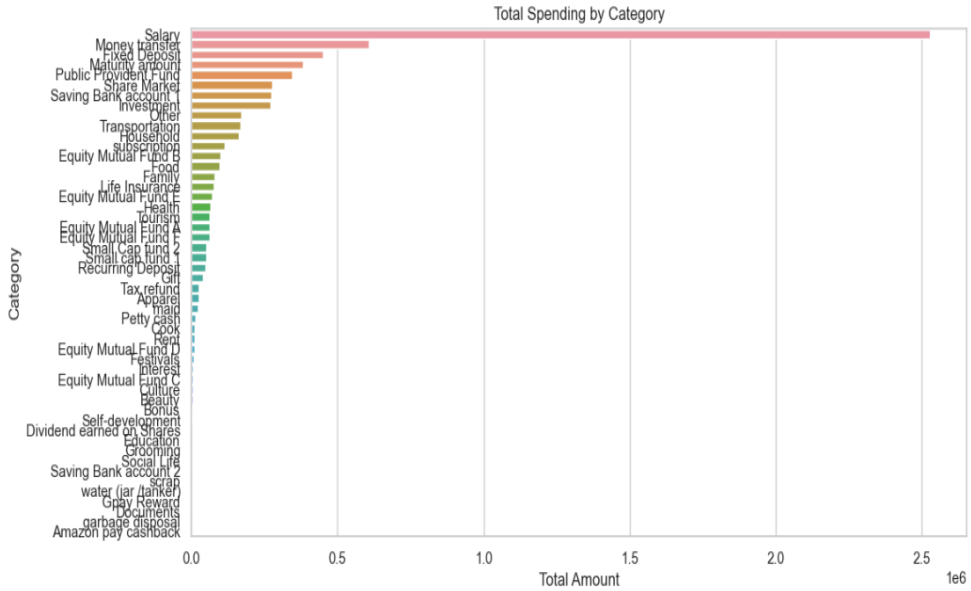


Figure 3: A bar plot illustrating the total spending for each category.

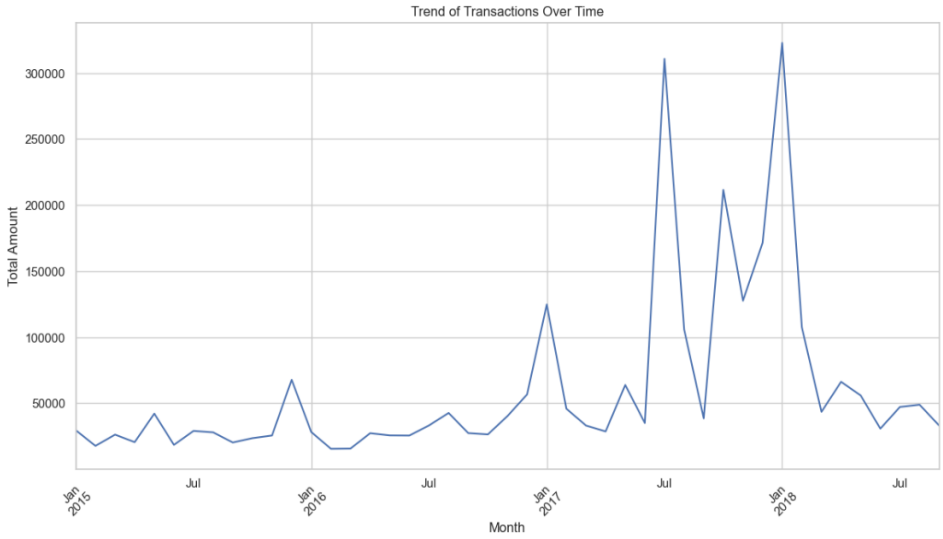


Figure 4: A time series plot showing the trend of transactions over time.

These visualizations provided valuable insights into customer purchasing behaviors, highlighting key patterns and trends. By analyzing these graphical representations, informed decisions could be made for further segmentation and strategic planning.

4. Data Transformation

Before applying clustering algorithms, categorical variables were encoded, and numerical variables were standardized.

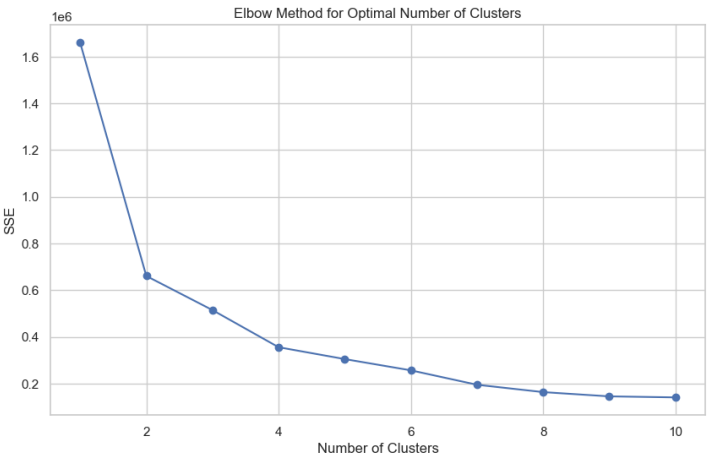
- **Encoding Categorical Variables:** Label encoding was used to convert categorical variables into numerical values.
- **Standardizing the Amount:** The Amount column was normalized using standard scaling.
- **Dropping Unnecessary Columns:** The Month column was dropped as it was no longer needed.

	Date	Mode	Category	Subcategory	Amount	Income/Expense
0	NaT	0	44	64	-0.217395	0
1	NaT	0	18	88	-0.214998	0
2	2018-09-19	9	48	45	-0.203893	0
3	NaT	9	48	41	-0.218274	0
4	NaT	0	16	20	-0.199739	0

5.Clustering

Optimal Number of Clusters

The optimal number of clusters is determined using the Elbow method, where the sum of squared errors (SSE) is plotted against the number of clusters. This method helps in identifying the point where the SSE starts to diminish at a slower rate, indicating the optimal number of clusters.



K-means Clustering

Based on the Elbow method, an optimal number of clusters (e.g., k=4) is chosen for K-means clustering. The K-means algorithm partitions the data into k clusters, where each data point belongs to the cluster with the nearest mean.

	Date	Mode	Category	Subcategory	Amount	Income/Expense	Cluster
0	NaT	0	44	64	-0.217395	0	0
1	NaT	0	18	88	-0.214998	0	3
2	2018-09-19	9	48	45	-0.203893	0	0
3	NaT	9	48	41	-0.218274	0	0
4	NaT	0	16	20	-0.199739	0	1

Cluster Visualization

To visualize the clusters, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data to 2D. This allows for a clear visual representation of the clusters, aiding in the interpretation of the segmentation results.

Conclusion

The customer segmentation analysis conducted by the team revealed distinct groups of customers based on their purchasing behavior, which facilitated targeted marketing strategies, personalized promotions, and improved customer retention efforts. The process involved several key steps. First, the data was meticulously cleaned and preprocessed to ensure accuracy and relevance for the analysis. Exploratory data analysis was then performed to understand the distribution and trends within the data, providing valuable insights into transaction patterns. Following this, K-means clustering was applied, which identified meaningful customer segments by grouping similar purchasing behaviors together. Finally, the results were visualized using Principal Component Analysis (PCA), allowing for a clear representation of the clusters in a two-dimensional space. These insights enabled the retail store to tailor its approaches to meet the specific needs and preferences of each segment, ultimately enhancing customer satisfaction and loyalty by providing more personalized and effective marketing strategies.