Segmenting Behavior: Using K-Means Clustering to Understand Spending Patterns

Isabella Lo | Data Science | Computer and Information Science
GitHub: github.com/beanbean9339 | LinkedIn: linkedin.com/in/igwlo

Data Source

- The dataset, sourced from **Kaggle**, contains **10,000 transaction records** from **200 unique customers**, detailing consumer spending behavior across various **categories** and **items**. It includes columns such as **Customer ID, Category, Item, Quantity, Price per Unit, Total Spent, Payment Method, Location,** and **Transaction Date**.
- The data is organized into categories like **Groceries, Shopping, Subscriptions, Housing and Utilities, Transportation, Food, Medical/Dental, Personal Hygiene, Fitness, Travel, Hobbies, Friend Activities**, and **Gifts**, with specific items listed under each category.

Problem Statement

The goal of this project is to **analyze customer spending patterns** and **segment customers into meaningful groups** based on their transaction behaviors. By examining transaction data from various **categories** and **items**, we aim to uncover insights into consumer spending habits, **identify high-value customers**, and improve business strategies related to **marketing**, **inventory management,** and **pricing**. The results of this analysis can help **retailers** target their **marketing efforts** more effectively and optimize their **sales strategies**.

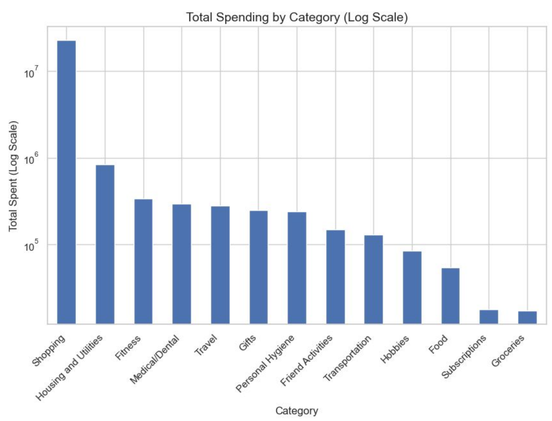
Exploratory Data Analysis (EDA)

Distribution of Key Variables:

1. **Quantity**: The average transaction involves **2.24 items** with a standard deviation of **1.48**, and most transactions include **1 or 3 items** (max 5).
2. **Price Per Unit**: The average **price per unit** is **$969.96**, with variability ranging from **$1 to $74,246.46**.
3. **Total Spent**: The average **total spent** per transaction is **$2,534.75**, with values ranging from **$1.11 to $352,230.76**. Most transactions are small, with the **25th percentile** at **$25.96**, the **median** at **$88.12**, and the **75th percentile** at **$336.38**.

Spending by Category:

A breakdown of **total spending by category** reveals the following top categories: **Shopping**: **$22,654,524.44, Housing and Utilities**: **$835,391.63, Fitness**: **$336,101.51, Medical/Dental**: **$294,709.10, Travel**: **$282,709.4.**



Feature Engineering

To enhance the dataset for analysis, several key features were engineered to provide more insight into customer behavior. The **"Recency"** feature was created to capture the time since each customer's last transaction, while **"Transaction Frequency"** was derived to measure how often customers make purchases. Additionally, **"Average Spending"** was calculated to understand typical spending patterns. **Categorical features** such as **"Payment Method"** and **"Location"** were transformed using one-hot encoding to allow for effective use in machine learning models. **Temporal features** like the **"Hour"** of transaction and the **"Weekday"** were also included to examine purchasing patterns over time. These engineered features offer a comprehensive view of customer activity, making the data more valuable for **predictive modeling** and further analysis.

Modeling

We applied **K-Means clustering** to uncover patterns in customer behavior based on spending habits, transaction frequency, and category preferences. As an unsupervised learning method, K-Means grouped customers into clusters without labeled data, enabling the identification of hidden customer segments.

Data Preprocessing

To ensure all features contributed equally, we standardized the data. We then applied **Principal Component Analysis (PCA)** for dimensionality reduction, simplifying visualization while retaining key patterns.

Optimal Number of Clusters

We used the **Elbow Method** to determine the optimal number of clusters, identifying that **4 clusters** best balanced model complexity and cohesion. The Elbow Method helps pinpoint the point where within-cluster variance decreases at a slower rate, guiding the selection of the most appropriate cluster count.

K-Means Clustering

After determining the optimal number of clusters, we applied **K-Means** to group customers. The **centroids**, or average feature values, represent the typical characteristics of each cluster.

Results and Insights

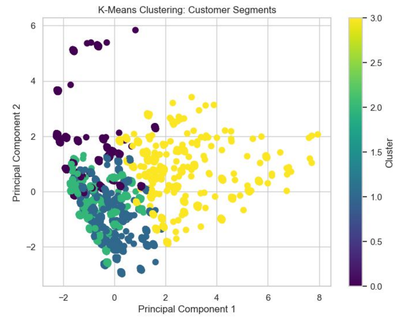
The clusters revealed distinct customer groups:

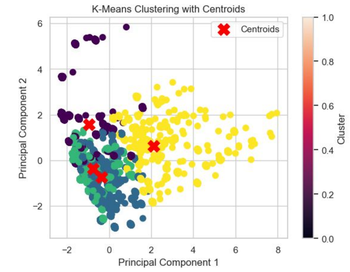
- **Cluster 0**: High spenders with a preference for Groceries and Personal Hygiene.
- **Cluster 1**: Moderate spenders, focusing on Gifts and Groceries.
- **Cluster 2**: Balanced spenders across various categories.
- **Cluster 3**: Primarily focused on Shopping and Fitness.

These insights enable **targeted marketing** and **personalized recommendations** tailored to each customer segment.

Visualization

Using PCA, we visualized the clusters in 2D, displaying the customer distribution and cluster centroids. This provided a clear representation of the distinct customer segments, allowing for further analysis and strategic development.





Discussion

The **K-Means clustering** analysis revealed distinct customer segments based on spending habits, transaction frequency, and category preferences. These insights can drive more targeted marketing and personalized recommendations. For example, **Cluster 0** (high spenders) favors **Groceries** and **Personal Hygiene**, while **Cluster 3** is more focused on **Shopping** and **Fitness**.

However, some limitations include the need for domain expertise to interpret clusters fully, the imbalance in cluster sizes, and the exclusion of other potential factors like customer demographics and seasonal trends.

Future Work

- **Enhancing Clustering**: Incorporate additional features like customer demographics and feedback, and explore alternative algorithms like DBSCAN to better capture complex customer behavior patterns.
- **Refining Marketing Strategies**: Validate clusters with external metrics (e.g., Silhouette Score) and develop dynamic, time-sensitive clustering models to create personalized and evolving marketing strategies.

References

- Datasource: https://www.kaggle.com/datasets/ahmedmohamed2003/spending-habits?resource=download

BDAA
BIG DATA & ANALYTICS ASSOCIATION