Customer Behavior Analytics

***Abstract*— This project will utilize comprehensive clustering analysis to using techniques such as PCA reduction, K-means clustering, and utilizing statistical scores to evaluate the final clusters. The primary objective of this project is to uncover valuable insights that can be leveraged for tailoring marketing strategies, identifying customer preferences, and enhancing customer engagement.**

I. Introduction

In today's data-driven business landscape, understanding customer behavior and preferences is important for the success of any organization. The primary objective of this project was to uncover valuable insights that can be leveraged for tailoring marketing strategies, identifying customer preferences, and enhancing customer engagement. The approach involved data exploration, visualization, and clustering techniques, which ultimately led to the division of the dataset into three distinct clusters. To evaluate the quality of the clustering, we employed the following statistical validation scores: Silhouette Score, Calinski-Harabasz and Davies Bouldin Score. The insights gained from this analysis can guide businesses in crafting strategies that can be tuned to cater to the diverse needs and preferences of their customer base. This in turn aids in improving customer satisfaction and enhanced business performance.

II. Data Processing

A. Data Set

The project's raw dataset is sourced from the E-Commerce Data Set found on Kaggle.com which encompasses all transactions conducted between 01/12/2010 and 09/12/2011 by a UK-based and registered non-store online retail company specializing in unique all-occasion gifts. Notably, a significant portion of the company's customer base consists of wholesalers. The dataset comprises eight variables: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. These dataset attributes serve as the foundation for conducting comprehensive clustering analysis and deriving insights into customer segmentation and behavioral patterns.

B. Data Cleaning

For the process of data cleaning, several steps were taken to prepare the dataset for analysis. Initially, all null values were removed to ensure data integrity. Additionally, duplicates were identified and dropped from the dataset to prevent redundancy and maintain accuracy in the analysis. In addition, records with a unit price of zero were eliminated, as they were considered irrelevant to the analysis. The dataset was also refined by removing any unusual or irregular stock codes, which could potentially skew the results. Lastly, the description feature was adjusted to exclude color words, enhancing the dataset's consistency, and reducing potential noise in subsequent analysis.

III. Feature Creation

In the process of feature creation, a new dataset was crafted from the original dataset's features, with the intention of using this refined dataset for subsequent clustering analysis. Several key customer-centric features were generated, including "Days Since Last Purchase," "Total Purchases," "Total Spent," "Average Value/Purchase," "Total Products Bought," "Cancellation Frequency," "Trend," and "Average Monthly Spending."

The newly generated features in our dataset will help in better understanding customer behavior and aid in eventual segmentation. "Days Since Last Purchase" reveals the recency of customer engagement, while "Total Purchases" and "Total Spent" provide insights into transaction frequency and total monetary contributions. "Average Value per Purchase" characterizes the typical transaction amount per purchase, and "Total Products Bought" quantifies the amount a customer has bought. "Cancellation Frequency" shows the frequency of canceled transactions per customer, potentially highlighting customer satisfaction or order fulfillment issues. The "Trend" feature indicates the direction of a customer's spending behavior over time, with a positive value suggesting growing loyalty and a negative value pointing to a decreasing interest. Lastly, "Average Monthly Spending" computes the average monthly spending, giving a glimpse into consistent spending habits.

IV. Cluster Creation

The first step to segmenting the new dataset was to standardize the data to have zero mean and unit variance across features and extract the "CustomerID" column, which is retained for identification purposes. This is crucial for the eventual usage of K-means clustering and PCA, which are sensitive to the scale of the data. The dataset is scaled using the StandardScaler, ensuring that each feature is on a similar scale. After scaling, the covariance matrix was computed, the eigenvalues and eigenvectors where calculated, and I performed cross-validation to identify the number of principal components to retain and plotted the cumulative explained variance against the number of components to visualize the "elbow point," in order to determine the optimal number of components to keep.

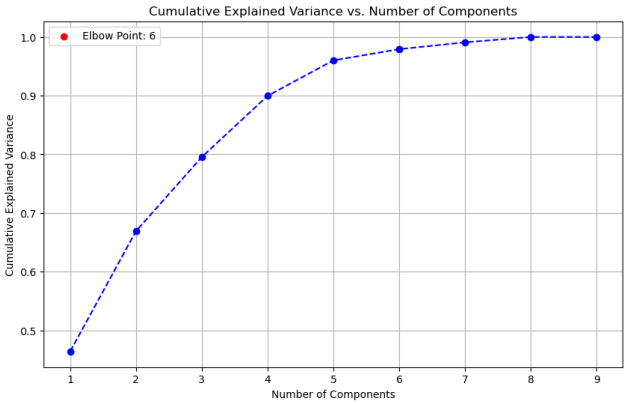

Fig. 1. Finding the Elbow Point

After identifying the elbow point, I then apply PCA with the chosen number of components (6 in this case, as shown in Figure 1) and create a new data frame with the reduced dimensions. The K-means clustering algorithm is then used to segment the data into clusters, and various clustering quality metrics such as inertia, silhouette scores, and average silhouette scores are calculated and displayed to determine the optimal number of clusters. In this case, three clusters are found to be the most suitable as shown in Figure 2.

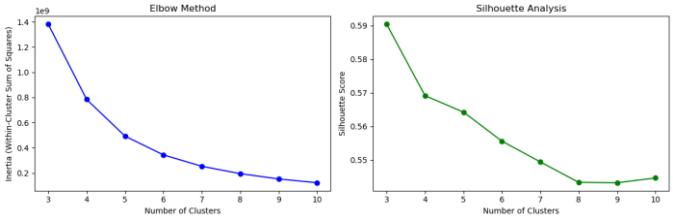

Fig. 2. Finding the Ideal Number of Clusters

Finally, the data is relabeled based on the clustering results, and a "Cluster" column is added to the data frame, which signifies the cluster assignment for each customer.

V. Cluster Evaluation

After obtaining the clustering results, I then assess the quality of the clusters using specific statistical metrics, which provided valuable insights into the clustering performance. The results are shown in Table 1 below.

TABLE I. Cluster Results

Evaluation Metric	***Score***
Silhouette Score	0.5904
Calinski Harabasz Score	16752.93
Davies Bouldin Score	0.4994

The Silhouette Score indicates that the clusters were reasonably well-separated and internally cohesive. Additionally, the Calinski Harabasz Score suggests that the clusters were distinct and well-defined. Finally, the Davies Bouldin Score indicates minimal overlap and good cohesion among the clusters. These scores collectively indicate that the clustering analysis successfully partitioned the data into three meaningful and distinct clusters. Following the comprehensive evaluation of the clusters, I proceeded to visualize the clusters, which are showcased in Figure 3.

Fig. 3. Clusters Visualized

These visualizations offer an insightful representation of the clustered data, and provide a clear and concise means of showcasing the following patterns and structures within the dataset:

A. Cluster 0

The characteristics of Cluster 0, such as the highest average value per purchase and monthly spending with a moderate total spending level, suggest a customer segment that values quality and is willing to spend more on individual items. This group might consist of high-end or luxury product enthusiasts who value exclusive and premium offerings. They are likely to appreciate personalized and high-value recommendations and could be a target for loyalty programs aimed at retaining their business and encouraging continued high-value purchases.

B. Cluster 1

Cluster 1 stands out for having the highest trend, total products bought, and total purchases among all clusters. This cluster may represent a segment of highly engaged and prolific customers. The high trend indicates growing customer loyalty and satisfaction. They could be a valuable focus for maintaining and expanding the customer base for a company.

C. Cluster 2

Cluster 2 represents customers with the highest cancellation frequency, very high trend, and the highest number of days since the last purchase. This cluster's behavior could signify customers who have previously shown interest in the products but have encountered some issues or reservations. The high trend suggests that they might be open to re-engagement efforts. With moderate monthly spending and an average value per purchase, they may appreciate customized offers, improved customer service, and incentives to return. Addressing the cancellation frequency and resolving any underlying concerns could help rekindle their engagement.