# Customer Segmentation Using K-Means Clustering

**Team VeryBlueberry**

**Authors:**

**- Boddu, Tharun,** [**boddutha@msu.edu**](mailto:boddutha@msu.edu)

**- Chung, Yoensuk,** [**chungyoe@msu.edu**](mailto:chungyoe@msu.edu)

**- Depree, Jack,** [**depreeja@msu.edu**](mailto:depreeja@msu.edu)

**- Krpete Thimmegowda, Shrinidhi,** [**krpetesh@msu.edu**](mailto:krpetesh@msu.edu)

Technology Used: Python

Callout: Chatgpt4 was used for code improvement and content formatting and structuring

## 1. Introduction

Customer segmentation is a critical task in understanding and categorizing customers based on their behaviors and transactions. This project employs K-Means clustering to segment customers into distinct groups for improved marketing and engagement strategies. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are utilized to optimize the dataset while preserving the variance. The analysis ensures actionable insights for personalized marketing and business growth.

## 2. Dataset Details

The dataset used for this project contains transactional data of customer purchases, with the following features:

* • InvoiceNo: Unique identifier for each transaction.
* • StockCode: Unique product code.
* • Description: Description of the purchased product.
* • Quantity: Quantity of the product purchased.
* • InvoiceDate: Date and time of the transaction.
* • UnitPrice: Price per unit of the product.
* • CustomerID: Unique identifier for each customer.
* • Country: Country where the transaction occurred.

After preprocessing, the dataset contains 4,067 observations. It provides a comprehensive view of customer behaviors across various metrics.

## 3. Data Cleaning

Data cleaning was performed to ensure the dataset was free from inconsistencies and errors, improving the reliability of the analysis. The following steps were taken:

* • Removed rows with missing critical values, such as CustomerID and Description.
* • Eliminated duplicate entries to avoid redundant data influencing results.
* • Filtered out non-product transactions, such as 'POST' and 'BANK CHARGES', identified via StockCode.
* • Identified and excluded outliers using Isolation Forest, marking 5% of the data as anomalies.

These steps ensured a clean and consistent dataset, ready for feature engineering and analysis.

## 4. Feature Engineering

Feature engineering was conducted to derive additional insights from the data and enhance clustering performance. The following features were created:

* • Total Spend: Calculated as UnitPrice × Quantity, representing the total value of a purchase.
* • Days Since Last Purchase: The number of days since the customer's most recent transaction.
* • Average Transaction Value: Average spending per transaction for each customer.
* • Cancellation Rate: Ratio of canceled transactions to total transactions.
* • Unique Products Purchased: Number of distinct products purchased by a customer.

These engineered features provided a richer dataset for clustering analysis.

## 5. Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce the dataset's dimensionality while retaining most of the variance. The following steps were undertaken:

* • Standardized all numerical features to ensure comparability.
* • Determined the optimal number of components using the explained variance plot. Six components were chosen, capturing 96% of the total variance.

A graph with a curve

Description automatically generated

## 6. Clustering Methodology

K-Means clustering was applied to segment customers into groups based on their behavior. The optimal number of clusters was determined using the following methods:

* • Elbow Method: Identified k=3 as the optimal number of clusters.
* • Silhouette Analysis: Validated the choice of k=3 with a score of 0.23.A graph of a number of clusters

  Description automatically generated

A screenshot of a graph

Description automatically generated

## 7. Cluster Evaluation

The clustering results were evaluated using the following metrics:

* Silhouette Score: 0.23, indicating fair separation between clusters.
* Calinski-Harabasz Score: 1216.78, suggesting compact and well-separated clusters.
* Davies-Bouldin Score: 1.41, indicating moderate similarity between clusters.

## 8. Customer Segmentation Insights

The identified clusters can be characterized as follows:

* Cluster 0: High-value customers with frequent and high-spending behavior.
* Cluster 1: Low-value, disengaged customers with minimal activity and high cancellations.
* Cluster 2: Moderate-value customers with steady engagement and spending.

A close-up of a map

Description automatically generated

## 9. Conclusion

Successfully segmented customers into three meaningful groups, providing actionable insights for business strategies. The combination of PCA and K-Means clustering allowed for effective dimensionality reduction and clustering. These insights can help businesses tailor marketing strategies, improve customer retention, and optimize resources.