Customer Segmentation Using K-Means Clustering

1. Introduction

Customer segmentation is a crucial aspect of data-driven marketing and business strategy. It allows businesses to categorize customers into distinct groups based on their behaviour, demographics, or purchasing patterns. The objective of this project is to implement **K-Means clustering** to segment customers efficiently and evaluate the quality of clustering using the **silhouette score**. By identifying patterns within the data, businesses can tailor their marketing strategies, improve customer experiences, and optimize resource allocation.

2. Data Collection

The dataset used for this study contains details about customers, including demographic and spending behaviour information. These details include customer ID, age, gender, annual income, and spending score. The dataset provides insights into purchasing habits, allowing for meaningful segmentation. Data sources may include:

- **Customer purchase history**
- **Survey responses**
- **E-commerce transaction records**
- **Social media interactions**

The features used for clustering are carefully selected to ensure meaningful segmentation. In this project, key features include age, annual income, and spending score.

3. Data Preprocessing

To ensure the accuracy and efficiency of the clustering model, the dataset undergoes a series of preprocessing steps:

- **Loading and inspecting the data**: Understanding the structure of the dataset, checking for missing values, and ensuring data consistency.
- **Handling missing values**: If any missing values exist, appropriate imputation techniques such as mean or median replacement are applied.
- **Feature selection**: Choosing the most relevant attributes that contribute to meaningful clusters.

4. Choosing the Optimal Number of Clusters

A critical step in K-Means clustering is selecting the optimal number of clusters (k). The following methods are used to determine the best value for k:

- **Elbow Method**: This method evaluates the Within-Cluster Sum of Squares (WCSS) for different values of k. The point where the WCSS curve starts to level off (elbow point) is chosen as the optimal k.
- **Silhouette Score**: This metric assesses how well each point fits within its cluster. A higher silhouette score indicates well-separated clusters.

- **Domain expertise**: Sometimes, understanding the business context helps determine a meaningful number of customer segments.

5. Implementing K-Means Clustering

After determining the optimal number of clusters, K-Means is applied to the dataset. The model partitions the data into clusters, assigns labels to each customer, and identifies the cluster centroids. The centroids represent the central points of each cluster and help in analyzing customer groups.

6. Visualizing Clusters

To interpret the clustering results, visualization techniques are employed:

- **Scatter plots**: Used to display customer segmentation in two or three dimensions.
- **Centroid markers**: Cluster centers are highlighted to show the representative characteristics of each segment.

Cluster visualization helps businesses understand different customer behaviours and identify key patterns that influence decision-making.

7. Evaluating Cluster Performance

Cluster quality is assessed using the **Silhouette Score**, which measures the separation and cohesion of clusters. The score ranges from -1 to 1, where:

- **A score close to 1** indicates well-defined clusters.
- **A score close to 0** suggests overlapping clusters.
- **A negative score** implies that some data points may be incorrectly assigned.

8. Conclusion

Customer segmentation using K-Means clustering allows businesses to classify customers into meaningful groups based on their purchasing behaviour. This enables personalized marketing strategies, improved customer engagement, and optimized resource allocation. The silhouette score provides a quantitative measure to validate the effectiveness of clustering.