Project Report

Customer Segmentation using K-Means Clustering

Objective:

The main objective of this project is to segment customers into different groups based on their purchasing behavior. Specifically, the project focuses on clustering customers using their **Annual Income** and **Spending Score**, which helps businesses understand the buying patterns of customers and target them with specific strategies or offers. By identifying distinct customer types, marketing can become more personalized and efficient.

Tools and Technologies Used:

This project was implemented using Python in a Jupyter Notebook environment. We used several libraries:

- **Pandas** for data loading and preprocessing.
- **Matplotlib** and **Seaborn** for data visualization.
- **scikit-learn** (sklearn) for applying the K-Means clustering algorithm and evaluating the model.

Dataset Description:

We used the popular **Mall Customer Segmentation** dataset which contains demographic data for 200 customers. The columns include CustomerID, Gender, Age, Annual Income (in thousands), and Spending Score (1–100). For this clustering task, only **Annual Income** and **Spending Score** were used to identify groups of similar customers.

Project Steps and Implementation:

1. **Data Preprocessing**
   The dataset was imported using pandas. We selected only the two relevant numerical columns: Annual Income and Spending Score. These values were then

scaled using StandardScaler to ensure both features are on the same scale, which is important for distance-based models like K-Means.

2. **Elbow Method to Find Optimal Number of Clusters**
   To determine the best number of clusters, we used the Elbow Method. We plotted the “inertia” (which measures how tightly grouped the data points in each cluster are) against different values of k. The point where the graph starts to bend (elbow point) was at k=5, so we chose 5 clusters for our model.

3. **Model Training with K-Means Algorithm**
   We trained a K-Means clustering model with n_clusters=5. The algorithm grouped the customers into 5 clusters based on similarity. It also calculated the center point (centroid) of each cluster.

4. **Visualization**
   The clustered customers were plotted using a scatter plot. Each cluster was colored differently, and the centroids were marked as large 'X' symbols. This visualization clearly showed how the customer data was grouped.

5. **Model Evaluation using Silhouette Score**
   To evaluate how well the data was clustered, we used the Silhouette Score, which measures how similar a data point is to its own cluster compared to others. Our model scored approximately **0.466**, which indicates that the clustering is reasonably good, with some overlap but distinct group separation.

What is K-Means Clustering?

K-Means is an **unsupervised machine learning algorithm** used for clustering data. It doesn’t rely on labeled data but instead tries to divide the dataset into k groups by minimizing the distance between data points and the center of their group (called a **centroid**). Each data point is assigned to the nearest centroid, and the centroids are recalculated until the best grouping is found.

What is Clustering?

Clustering is a method of unsupervised learning where the goal is to discover natural groupings in the data. It helps identify hidden patterns or segments without prior labeling. In this project, clustering helped us discover different types of customers based on how much they earn and how much they spend.

Key Findings:

- The optimal number of clusters for this dataset is 5.
- Each cluster represents a different type of customer based on income and spending habits.
- For example, some clusters may represent high-income low-spending customers, or low-income high-spending ones.
- The visualization shows a clear separation between different types of customers.
- This segmentation can help businesses in marketing, promotions, and personalized targeting.

Why Scatter Plot Was Used Instead of a Bar Chart:

Scatter plots are best for showing the relationship between two continuous numerical features — in this case, income and spending. Bar charts, on the other hand, are better for comparing categories. Since we needed to observe patterns and clusters in numeric data, a scatter plot was more appropriate and informative.

Difference from Previous Project (Sales Analysis):

Unlike your earlier project where Power BI was used for building dashboards and summarizing sales metrics, this project focuses on applying **machine learning** for **behavioral grouping**. The sales project involved reporting and descriptive analysis, while this clustering project involves predictive modeling and unsupervised learning. Also, in this project we used Python and machine learning libraries, while the sales project used Power BI with SQL and Excel.

Conclusion:

This project successfully demonstrated how K-Means clustering can be applied to real-world customer data to gain actionable insights. By identifying distinct customer segments, businesses can adopt more targeted marketing strategies. The project used Python and essential data science libraries, and achieved a moderate silhouette score of 0.466, indicating that the clusters are reasonably well-formed.