Document structure

CUSTOMER SEGMENTATION USING DATA SCIENCE

Introduction:

In this project, we perform clustering analysis on a dataset using the K-Means algorithm. Clustering analysis is a technique used in unsupervised learning to group similar data points together.

Dataset

We use the “Mall Customers” dataset, which contains information about customers in a mall. The dataset has the following columns:

- `CustomerID`: Unique identifier for each customer.
- `Gender`: Gender of the customer.
- `Age`: Age of the customer.
- `Annual Income (k$)`: Annual income of the customer.
- `Spending Score (1-100)`: Spending score assigned to the customer.

Code Explanation:

1. Importing Libraries:
   - We start by importing the necessary libraries: `numpy`, `pandas`, `matplotlib`, and `seaborn`.

2. Loading the Dataset:
   - We load the “Mall Customers” dataset from a CSV file using `pd.read_csv()`.

3. Exploratory Data Analysis:

- We check the first few rows of the dataset using `dataset.head()` to get an overview of the data.

- We use `dataset.shape` to find the dimensions of the dataset (rows, columns).

- `dataset.info()` provides information about the columns and data types.

- `dataset.isnull().sum()` checks for any missing values in the dataset.

4. Feature Selection:

- We select the features `Annual Income` and `Spending Score` for clustering. These two features are relevant for customer segmentation.

5. K-Means Clustering:

- We use the `KMeans` class from scikit-learn to perform clustering analysis.

- We iterate over different values of `k` (number of clusters) from 1 to 10.

- For each `k`, we calculate the Within-Cluster-Sum-of-Squares (WCSS) and store it in the `wcss` list.

- WCSS is a metric that measures the compactness of the clusters.

6. Elbow Method:

- We plot the number of clusters (`k`) against the WCSS to find the optimal number of clusters.

- The point where the WCSS starts to plateau is considered the “elbow” point, indicating the optimal number of clusters.

7. Visualization:

- We create a plot showing the “elbow” point.

Results

Based on the elbow method, the optimal number of clusters for this dataset is [insert optimal k value].