Mall Customers Segmentation Using K-Means Clustering

Objective

This project aims to segment mall customers into distinct groups based on their spending behaviors and annual income, helping businesses to understand customer patterns and tailor their marketing strategies accordingly. K-means clustering is used for identifying clusters in customer data, allowing us to analyze different groups with varying spending scores and income levels.

Dataset

The dataset contains information on 200 customers, with the following columns:

- **CustomerID**: Unique identifier for each customer.
- **Gender**: Gender of the customer.
- **Age**: Age of the customer.
- **Annual Income (k$)**: The annual income of the customer in thousands of dollars.
- **Spending Score (1-100)**: A score assigned by the mall based on customer spending behavior and loyalty.

Steps

1. **Data Loading and Exploration**:
    - The data is loaded using pandas and initial exploration is performed with head(), info(), and describe() methods to check for missing values and to understand data structure.
    - Null values are checked with df.isnull().sum().
2. **Data Visualization**:
    - **Scatter Plots**: Three scatter plots are created to visualize relationships:
        - **Age vs. Spending Score**: Helps analyze spending patterns by age.
        - **Age vs. Annual Income**: Reveals the income distribution across different age groups.
        - **Annual Income vs. Spending Score**: Key for identifying clusters based on spending and income.
3. **Optimal Number of Clusters**:
    - To determine the optimal number of clusters, the **Elbow Method** is applied by calculating the Sum of Squared Errors (SSE) for each number of clusters (from 1 to 9).
    - A plot of SSE versus the number of clusters helps identify the "elbow point," suggesting the optimal number of clusters for segmentation.

4. **K-Means Clustering**:
    - The K-means algorithm is implemented with n_clusters=5, as determined by the Elbow Method.
    - Each customer is assigned to a cluster based on their **Annual Income** and **Spending Score**.
    - A new column, Cluster, is added to the dataset, representing each customer's cluster assignment.
5. **Visualization of Clusters**:
    - Clusters are visualized on a scatter plot of **Annual Income vs. Spending Score**, with each cluster represented by a unique color.
    - The cluster centroids are plotted using a purple star marker to show the center of each cluster.

Project Outcome

This segmentation allows businesses to group customers into clusters such as high-income/high-spending, low-income/low-spending, and other patterns based on income and spending score. The project demonstrates how K-means clustering can help reveal actionable insights in customer data, leading to targeted marketing strategies.