K-Means Clustering on Mall Customers Dataset

Code Overview

This code applies K-Means clustering to the Mall Customers dataset to identify distinct customer segments based on their demographic and purchasing behavior. The goal is to understand different types of customers that shop at the mall, which can help in targeted marketing and personalized customer experiences.

Steps Performed

1. **Load the Dataset:**
   - The dataset is loaded from a CSV file into a pandas DataFrame.
   - The dataset contains 200 entries with the following features:
     CustomerID: Unique identifier for each customer.
     Gender: Gender of the customer.
     Age: Age of the customer.
     Annual Income (k$): Annual income of the customer in thousands of dollars.
     Spending Score (1-100): Score assigned by the mall based on customer behavior and spending nature.

2. **Data Preprocessing:**
   - Only the relevant features ('Age', 'Annual Income (k$)', 'Spending Score (1-100)') are selected for clustering.
   - The selected features are standardized using 'StandardScaler' to ensure all features contribute equally to the distance calculations during clustering.

3. **K-Means Clustering:**
   - The K-Means algorithm from 'scikit-learn' is used to cluster the customers.
   - Clustering is performed for 3, 4, and 5 clusters to explore different segmentation scenarios.
   - The optimal number of clusters can be determined by analyzing the cluster centers and the distribution of customers in each cluster.

4. **Cluster Analysis:**
   - Cluster centers for each scenario (3, 4, and 5 clusters) are computed and inverse transformed to original feature scales.
   - Value counts for each cluster are displayed to understand the distribution of customers among clusters.

- Summary statistics of the dataset, including minimum, maximum, and average values for age, annual income, and spending score, are calculated.

5. **Visualization:**
- Scatter plots are generated to visualize the clusters in 2D space.
- Each plot shows the standardized values of 'Age' vs. 'Annual Income (k$)', colored by cluster labels.
- These visualizations help in interpreting the characteristics of each cluster.

Files

"kmeans_mall_customers.ipynb": The Jupyter notebook containing the implementation of K-Means clustering.

"Mall_Customers.csv": The dataset used for clustering.