Customer Segmentation with K-Means

Goal: Segment customers based on behavior.
Key Tasks:

- Use Mall Customer Segmentation dataset

- Normalize features (e.g., Age, Income, Spending Score)

- Apply K-Means and visualize clusters
  Skills Used: Unsupervised learning, clustering, PCA (optional)

Glimpse of code below

```python
main.py > ...
1     # customer_segmentation_kmeans.py
2
3     import pandas as pd
4     import matplotlib.pyplot as plt
5     import seaborn as sns
6     from sklearn.preprocessing import StandardScaler
7     from sklearn.cluster import KMeans
8     from sklearn.decomposition import PCA
9
10    # ----------------------------
11    # Step 1: Load Data
12    # ----------------------------
13    data = pd.read_csv('mall_customers.csv')
14    print("Dataset Sample:")
15    print(data.head())
16
17    # ----------------------------
18    # Step 2: Preprocess Data
19    # ----------------------------
20
21    # Encode Gender
22    data['Gender'] = data['Gender'].map({'Male': 0, 'Female': 1})
23
24    # Select features for clustering
25    features = data[['Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)']]
26
27    # Normalize features
28    scaler = StandardScaler()
29    scaled_features = scaler.fit_transform(features)
30
31    # ----------------------------
32    # Step 3: K-Means Clustering
33    # ----------------------------
34
35    # Find optimal number of clusters (Elbow method - optional visualization)
36    wcss = []
37    for i in range(1, 11):
38        kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
39        kmeans.fit(scaled_features)
40        wcss.append(kmeans.inertia_)
41
```

```python
41
42    # Plot the elbow
43    plt.figure(figsize=(8, 5))
44    plt.plot(range(1, 11), wcss, marker='o')
45    plt.title('Elbow Method')
46    plt.xlabel('Number of clusters')
47    plt.ylabel('WCSS')
48    plt.grid(True)
49    plt.show()
50
51    # Apply KMeans with optimal k (e.g., k=5 based on elbow)
52    k = 5
53    kmeans = KMeans(n_clusters=k, init='k-means++', random_state=42)
54    clusters = kmeans.fit_predict(scaled_features)
55
56    # Add cluster labels to original data
57    data['Cluster'] = clusters
58
59    # -----------------------------
60    # Step 4: Visualization
61    # -----------------------------
62
63    # 2D Plot using Age and Spending Score
64    plt.figure(figsize=(8, 6))
65    sns.scatterplot(
66        x=data['Age'], y=data['Spending Score (1-100)'],
67        hue=data['Cluster'], palette='Set2', s=60
68    )
69    plt.title('Customer Segments (Age vs Spending Score)')
70    plt.xlabel('Age')
71    plt.ylabel('Spending Score')
72    plt.legend(title='Cluster')
73    plt.grid(True)
74    plt.show()
75
76    # -----------------------------
77    # Step 5: PCA (Optional for 2D Visualization)
```

```
(eda_env) ubuntu@ubuntu:~/Documents/datascience/5_dcs_projs_2hrs_python/Customer_Segmentation_using_K-Means$ python3 main.py
Dataset Sample:
   CustomerID  Gender  Age  Annual Income (k$)  Spending Score (1-100)
0          1    Male   34              98560                      21
1          2  Female   26              58642                      30
2          3    Male   50              80160                      13
3          4    Male   37              32602                      34
4          5    Male   30             118320                       4
Segmented customer data saved to 'clustered_mall_customers.csv'.
```

Elbow Method

Customer Segments (Age vs Spending Score)

Customer Segments (PCA-Reduced Features)