K-Means clustering algorithm

October 25, 2024

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

[5]:
```
# Load the customer dataset
data = pd.read_csv('Mall_Customers.csv')
data
```

[5]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
..	…	…	…	…	…
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

[200 rows x 5 columns]

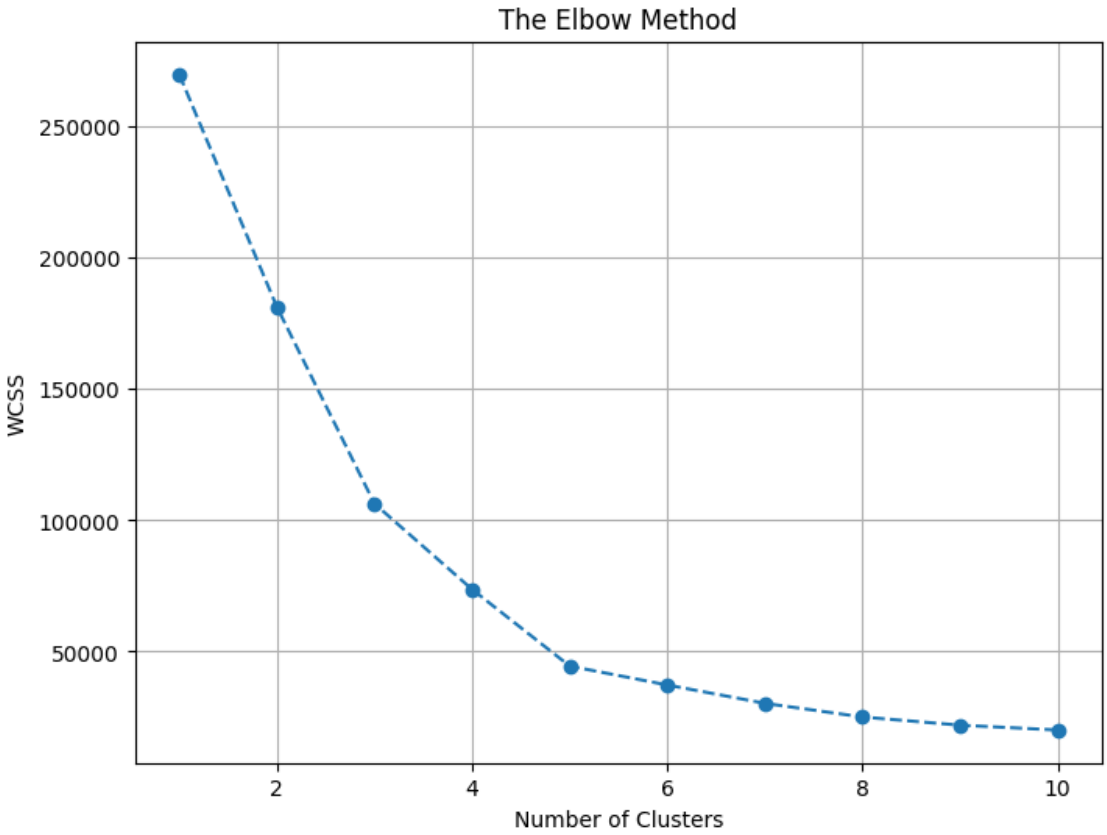
[6]:
```
data.columns.tolist()
```

[6]:
```
['CustomerID', 'Gender', 'Age', 'Annual Income (k$)', 'Spending Score (1-100)']
```

[7]:
```
X = data[['Annual Income (k$)', 'Spending Score (1-100)']].values
```

[8]:
```
wcss = []  # Within-cluster sum of squares
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10,␣
 ↪random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

[9]:
```
# Plot the Elbow method graph
plt.figure(figsize=(8,6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.title('The Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')  # Within-cluster sum of squares
plt.grid(True)
plt.show()
```

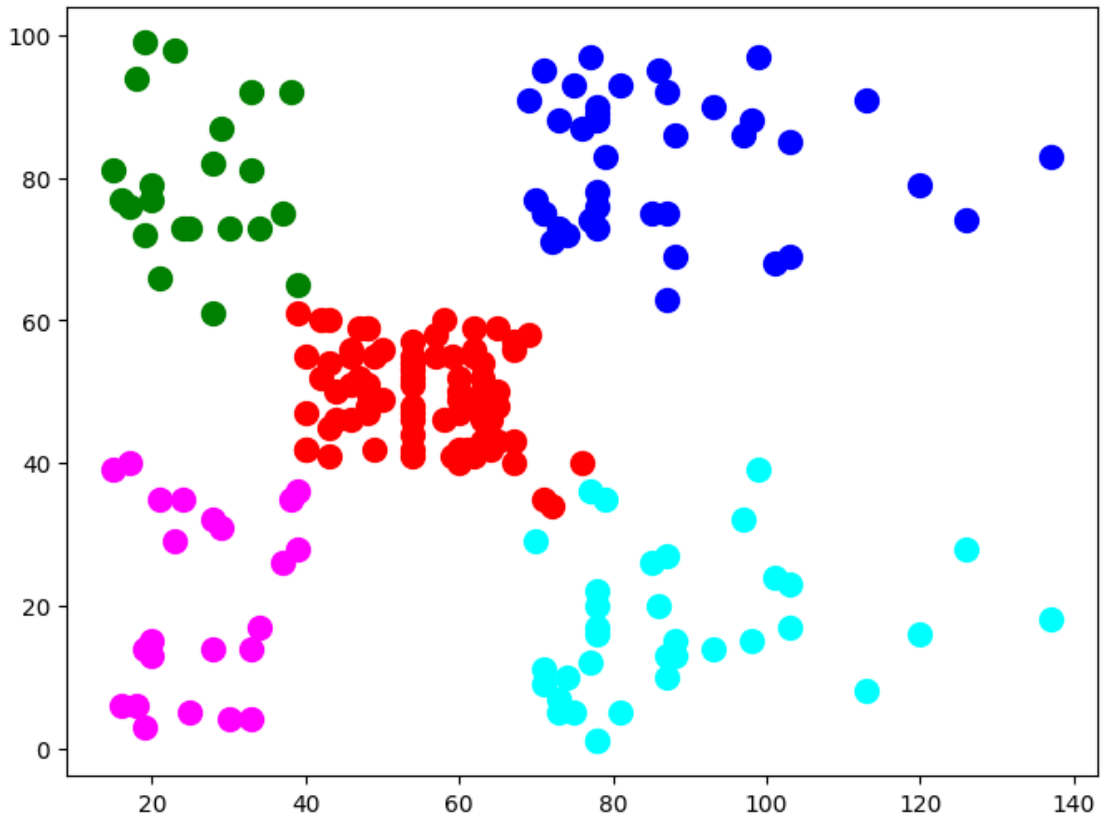


[10]:
```
optimal_clusters = 5
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300,␣
 ↪n_init=10, random_state=42)
y_kmeans = kmeans.fit_predict(X)
```

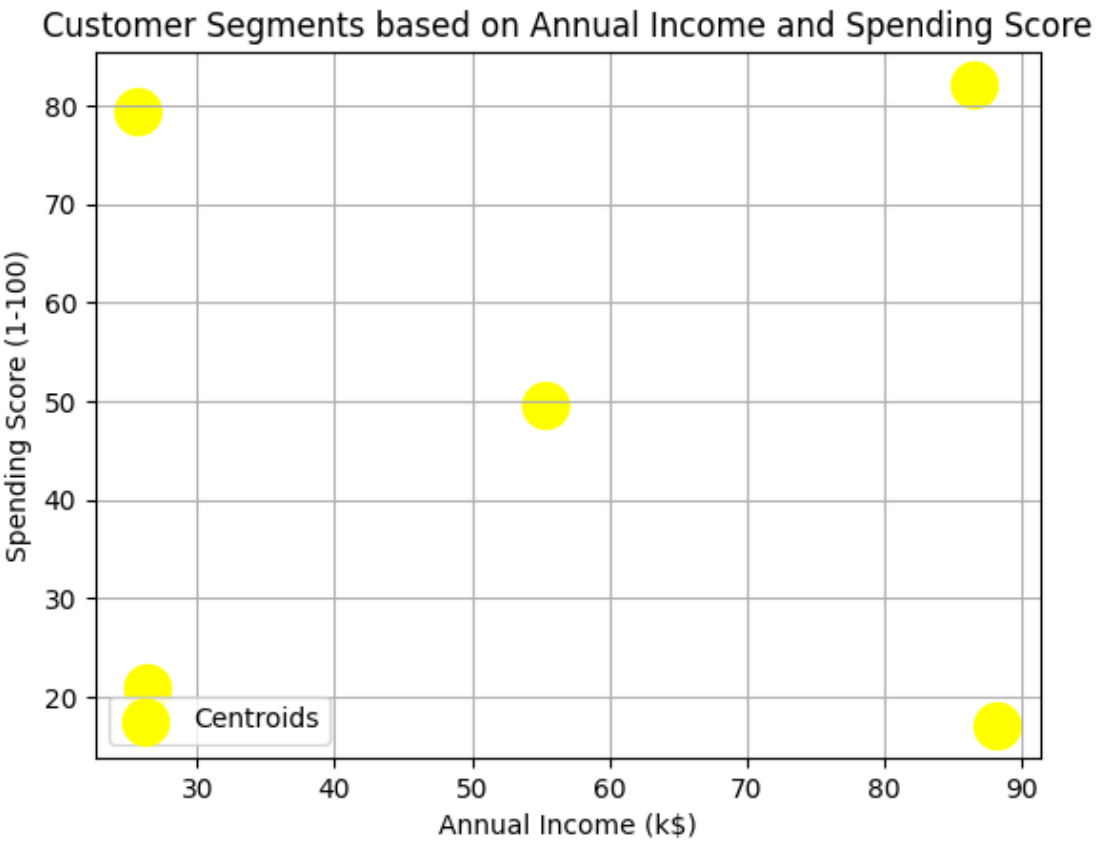
[11]:
```
plt.figure(figsize=(8,6))
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s=100, c='red',␣
 ↪label='Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s=100, c='blue',␣
 ↪label='Cluster 2')
```

```
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s=100, c='green',␣
 ↪label='Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s=100, c='cyan',␣
 ↪label='Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s=100, c='magenta',␣
 ↪label='Cluster 5')
```

[11]: <matplotlib.collections.PathCollection at 0x25633f08730>



[12]:
```
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],␣
 ↪s=300, c='yellow', label='Centroids')
plt.title('Customer Segments based on Annual Income and Spending Score')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.grid(True)
plt.show()
```


Customer Segments based on Annual Income and Spending Score
Spending Score (1-100)
Annual Income (k$)
Centroids
20
30
40
50
60
70
80
30
40
50
60
70
80
90