CS23334 -Fundamental Of Data Science

240701463

K-Means Clustering

Exp : 10

Date: 07-10-2025

Aim:

To apply the **K-Means Clustering** algorithm to the Mall_Customers dataset to identify natural groupings of customers based on their **Annual Income** and **Spending Score** , and to determine the optimal number of clusters using the **Elbow Method**.

Algorithm:

1. **Load Data:** Load the Mall Customers dataset and inspect its structure.

2. **Feature Selection:** Select Annual Income and Spending Score as features (X).

3. **K-Means Training:** Fit the KMeans model with an initial n_clusters=5.

4. **Labeling and Visualization:** Predict cluster labels, add them to the DataFrame, and visualize the clusters on a scatter plot.

5. **Elbow Method:** Calculate WCSS for K=1 to 9 (using Age, Income, Score) to find the optimal number of clusters.

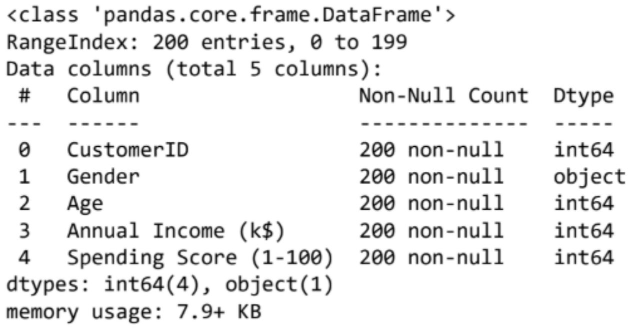
6. **Elbow Plot:** Plot WCSS versus the number of clusters.

Code:

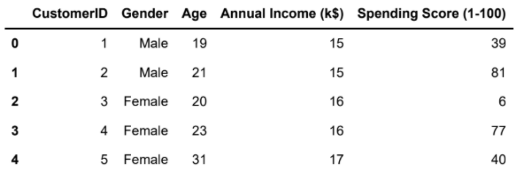
```
import numpy as np

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.cluster import KMeans
df=pd.read_csv('Mall_Customers.csv')
df.info()
df.head()
sns.pairplot(df)
features=df.iloc[:, [3,4]].values
model=KMeans(n_clusters=5, random_state=42, n_init='auto')
# Added random_state and n_init to suppress warnings
model.fit(features)
Final=df.iloc[:, [3,4]].copy() # Used .copy() to avoid SettingWithCopyWarning
Final['label']=model.predict(features)
Final.head()
sns.set_style("whitegrid")
sns.FacetGrid (Final, hue="label", height=8) \
.map(plt.scatter, "Annual Income (k$)", "Spending Score (1-100)") \
.add_legend();
plt.show()
# Elbow Method to find optimal K
features_el=df.iloc[:, [2,3,4]].values
wcss=[]
for i in range(1,10):
    model=KMeans(n_clusters=i, random_state=42, n_init='auto')
    model.fit(features_el)
    wcss.append(model.inertia_)
plt.plot(range(1,10), wcss)
plt.show()
```

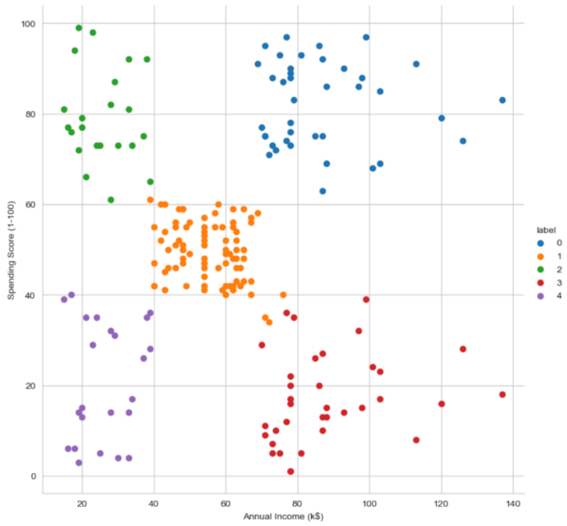
Output:

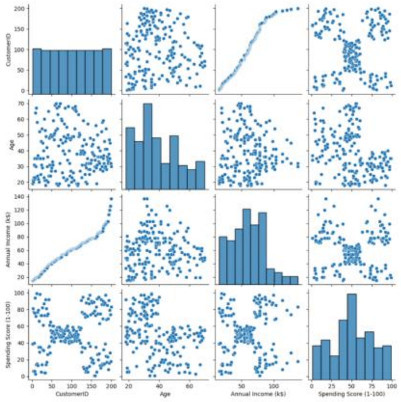


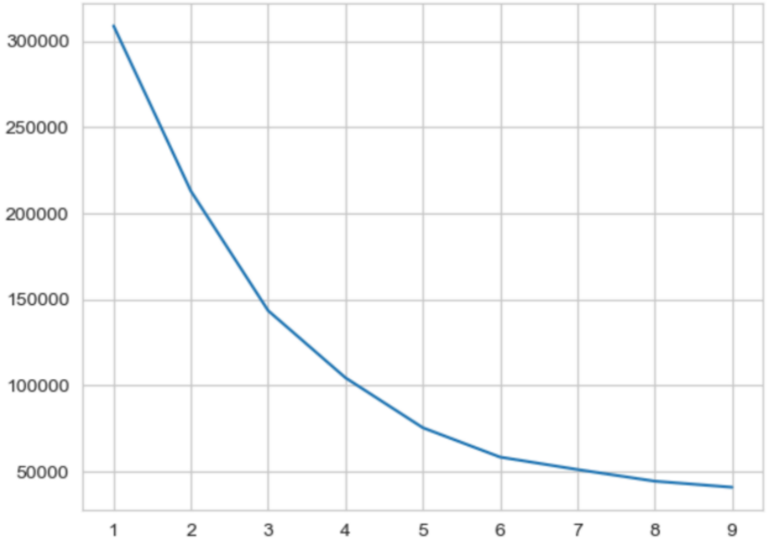
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```



	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







Result:

The K-Means clustering algorithm was successfully implemented on the Mall Customer data using **Annual Income** and **Spending Score**. The scatter plot, generated by training the model with **n_clusters=5**, clearly separates the customers into five distinct segments based on the two features. The Elbow Method plot, calculated using **Age, Annual Income, and Spending Score**, showed the most significant drop in WCSS occurring at **K=3** and **K=5**, with the curve starting to flatten significantly around K=5, which visually supports the selection of 5 clusters for the 2-feature segmentation. This confirms the method's ability to segment customers for targeted marketing strategies.