K - Means Clustering
Algorithm

Aim :

To implement K - Means Clustering algorithm using Python to group data points into distinct clusters based on their similarity

Algorithm :

1. Start

2. Import required libraries (numpy, pandas, sklearn, matplotlib).

3. Load the dataset (Mall_Customers.csv) file.

4. Select features - Annual Income and Spending Score.

5. Use the Elbow method to find the optimal number of clusters (K).

6. Initialize the K Means model with chosen number of clusters.

7. Fit the model and predict the cluster labels using fit-predict().

8. Add cluster labels to the dataset.

9. Visualize the clusters using a scatter plot.

10. Stop.

Program :

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans

df = pd.read_csv('Mall_Customers.csv')
print("\n Dataset Info: \n")
print(df.info())
print("\n First 5 Rows : \n")
print(df.head())
sns.pairplot(df)
features = df.iloc[:, [3,4]].values
model = KMeans(n_clusters = 5, init = 'kmeans++',
                 random_state = 42)
model.fit(features)
Final = df.iloc[:, [3,4]]
Final['label'] = model.predict(features)
print("\n Clustered Data (Sample): \n")
print(Final.head())
```

```python
sns.set_style("whitegrid")
sns.FacetGrid(Final, hue="label", height=8).map(
        plt.scatter, "Annual Income (k$)",
        "Spending Score (1-100)").add_legend()
plt.title('Customer Clusters based on Income and
            Spending Score')

plt.show()

features_el = df.iloc[:, [2,3,4]].values

wcss = []

for i in range(1,10):
    model = KMeans(n_clusters=i, init='kmeans++',
                    random_state=42)
    model.fit(features_el)
    wcss.append(model.inertia_)
plt.plot(range(1,10), wcss, marker='o')
plt.title('The Elbow Method')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('wcss')
plt.show()
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
Range Index: 200 entities, 0 to 199
Data columns (total 5 columns):
```
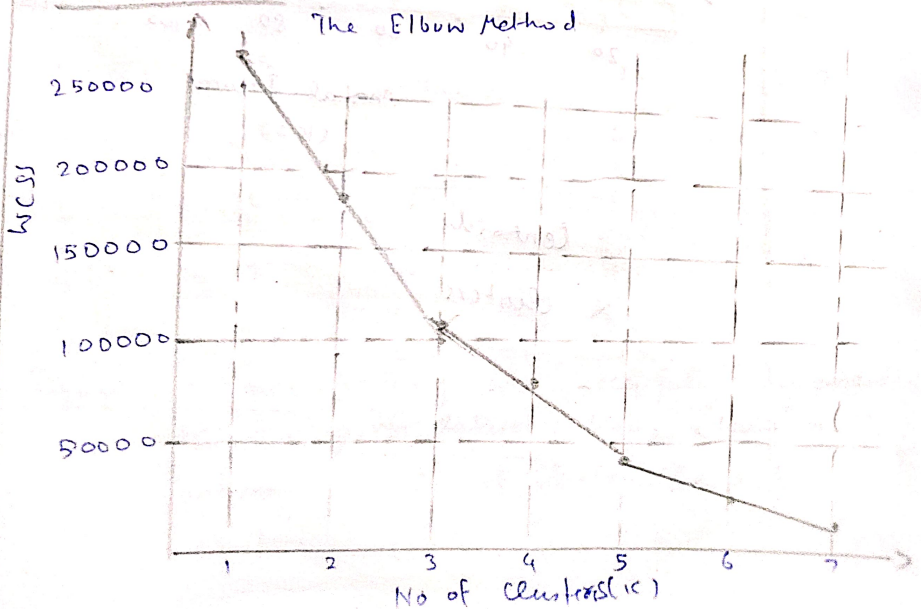
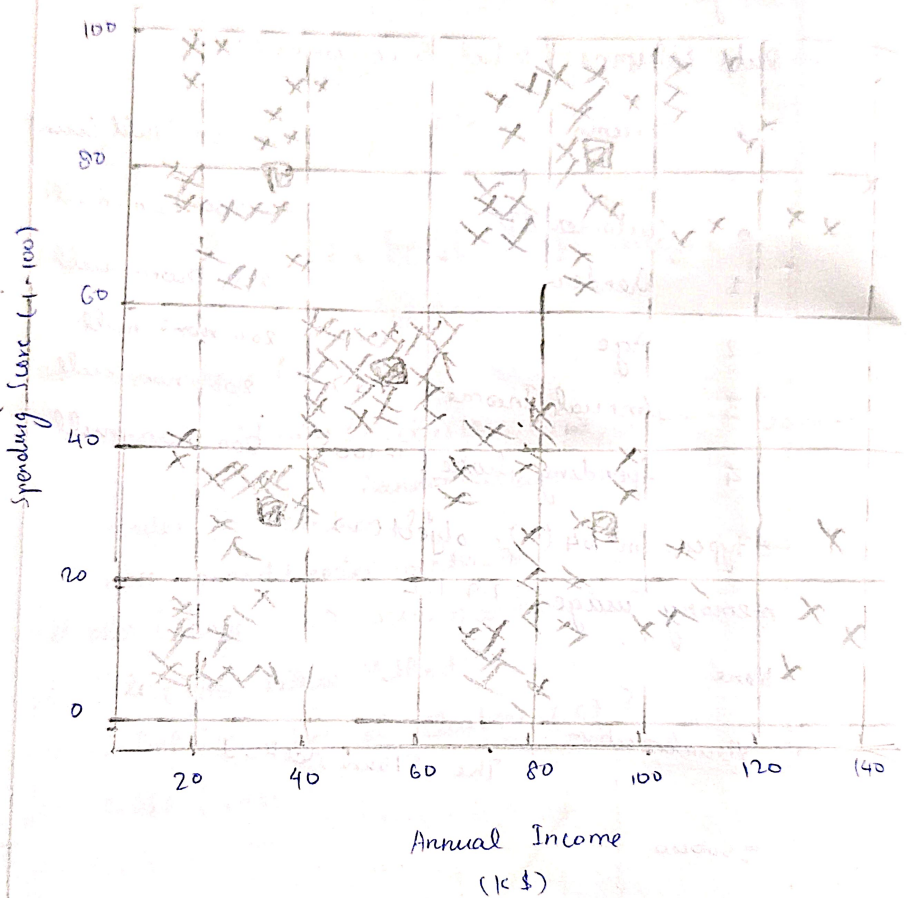| #	Column	Non-Null Count	Dtype
0	Customer ID	200 non-null	int 64
1	Gender	200 non-null	object
2	Age	200 non-null	int 64
3	Annual Income (K$)	200 non-null	int 64
4	Spending Score (1-100)	200 non-null	int 64

d types: int 64 (4), object (1)

memory usage: 7.9 KB

None

Visualization:

The Elbow Method

WCSS vs No of Clusters(K)

Visualization:

Customers Clusters based on Income and spending is.

Annual Income
(k $)

▨	Centroids
✗	clusters

Sample Rows:

Customer ID	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

Clustered Output (After Applying k-Means)

	Annual Income (k $)	Spending Score (1-100)	label
0	15	39	4
1	15	81	2
2	16	6	4
3	16	77	2
4	17	40	4

Result : k-Means clustering was successfully implemented on the given dataset. Using Elbow method optimal number of clusters found be k = 5. clusters.