Tittle : Mall Customer Segmentation

Project Desription : Mall Customer Segmentation groups shoppers based on their buying habits using K-Means Clustering. This helps businesses understand customers better and create targeted marketing strategies.

Problem Statement : Shoppers have different spending habits some buy more but visit less, while others visit often but spend less. The goal is to group customers based on their shopping patterns to better understand and serve them.

Desire Outcome : To group mall customers based on their shopping behavior, enabling businesses to personalize marketing strategies and improve customer engagement.

Importing required Libraries

In [3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Dataset Overview

In [4]:
```
df = pd.read_csv ("C:/Users/sande/Downloads/archive (5)/Mall_Customers.csv")
```

In [5]:
```
df
```

Out[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 × 5 columns

In [6]:
```
df.head()
```

Out[6]:

	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

Data Exploration (Exploratory Data Analysis - EDA)

In [7]:
```
df.isna().sum()
```

Out[7]:
```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

In [8]:
```
df.describe().T
```

Out[8]:

	count	mean	std	min	25%	50%	75%	max
CustomerID	200.0	100.50	57.879185	1.0	50.75	100.5	150.25	200.0
Age	200.0	38.85	13.969007	18.0	28.75	36.0	49.00	70.0
Annual Income (k$)	200.0	60.56	26.264721	15.0	41.50	61.5	78.00	137.0
Spending Score (1-100)	200.0	50.20	25.823522	1.0	34.75	50.0	73.00	99.0

In [9]:
```
df.columns
```

Out[9]:
```
Index(['CustomerID', 'Gender', 'Age', 'Annual Income (k$)',
       'Spending Score (1-100)'],
      dtype='object')
```

In [10]:
```
df['Gender'].value_counts()
```

Out[10]:
```
Female    112
Male       88
Name: Gender, dtype: int64
```

In [11]:
```
features = df [['Annual Income (k$)','Spending Score (1-100)']]
```

Feature Scaling

In [12]:
```
# data scaling
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)
```

In [13]:
```
scaled_features
```

```
[ 1.23821628,  1.54509812],
[ 1.390894  , -0.7065524 ],
[ 1.390894  ,  1.38981187],
[ 1.42906343, -1.36651894],
[ 1.42906343,  1.46745499],
[ 1.46723286, -0.43480148],
[ 1.46723286,  1.81684904],
[ 1.54357172, -1.01712489],
[ 1.54357172,  0.69102378],
[ 1.61991057, -1.28887582],
[ 1.61991057,  1.35099031],
[ 1.61991057, -1.05594645],
[ 1.61991057,  0.72984534],
[ 2.00160487, -1.63826986],
[ 2.00160487,  1.58391968],
[ 2.26879087, -1.32769738],
[ 2.26879087,  1.11806095],
[ 2.49780745, -0.86183865],
[ 2.49780745,  0.92395314],
```

Applying K-means Clustering

In [14]:
```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
```

In [15]:
```
conda install -c conda-forge openblas
```

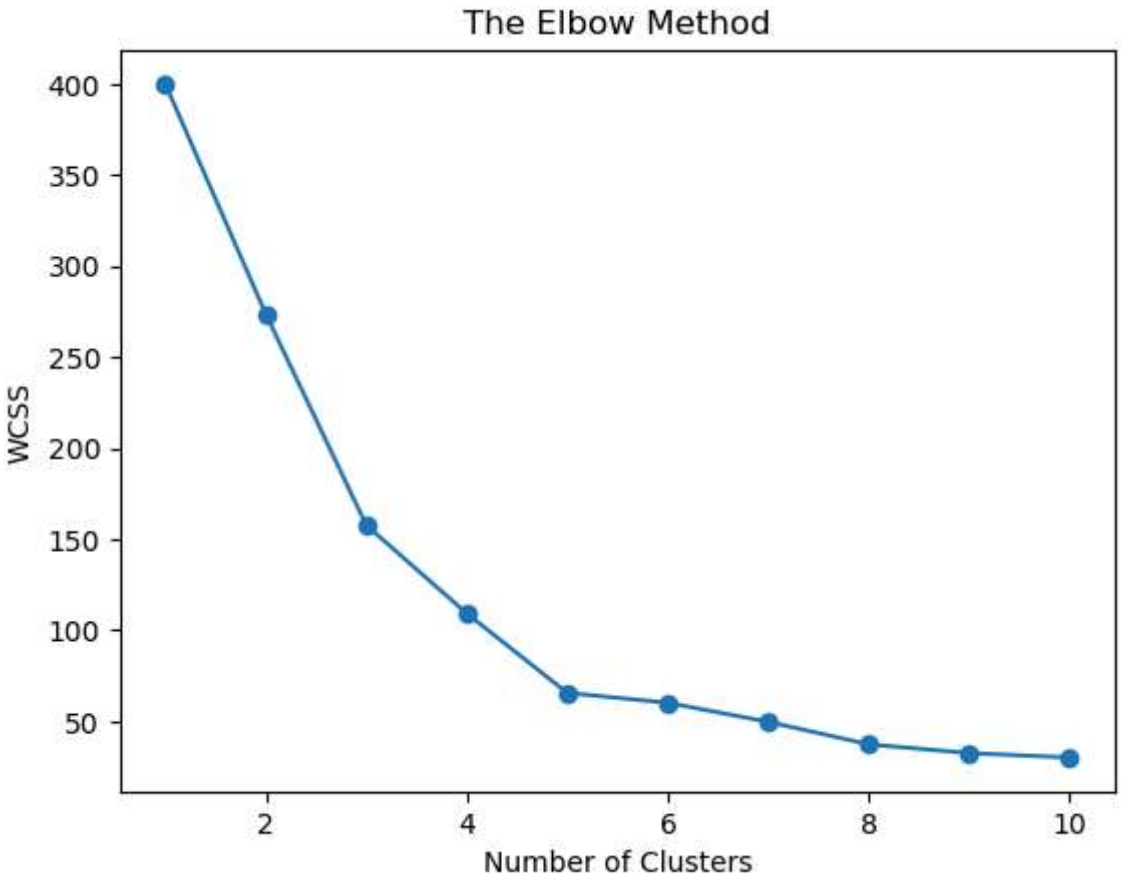
```
Note: you may need to restart the kernel to use updated packages.

The system cannot find the path specified.
```

Finding the Optimal Number of Clusters (Elbow Method)

In [16]:
```
# Elbow method to find optimal K
wcss = []  # with in cluster sum of squares
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
    kmeans.fit(scaled_features)
    wcss.append(kmeans.inertia_)
```

In [17]:
```
# plot the result
plt.plot(range(1, 11), wcss,marker = 'o')
plt.title('The Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('WCSS')
plt.show()
```



In [18]:
```
optimal_k = 5
kmeans = KMeans(n_clusters=optimal_k, init='k-means++', random_state=42)
clusters = kmeans.fit_predict(scaled_features)
```

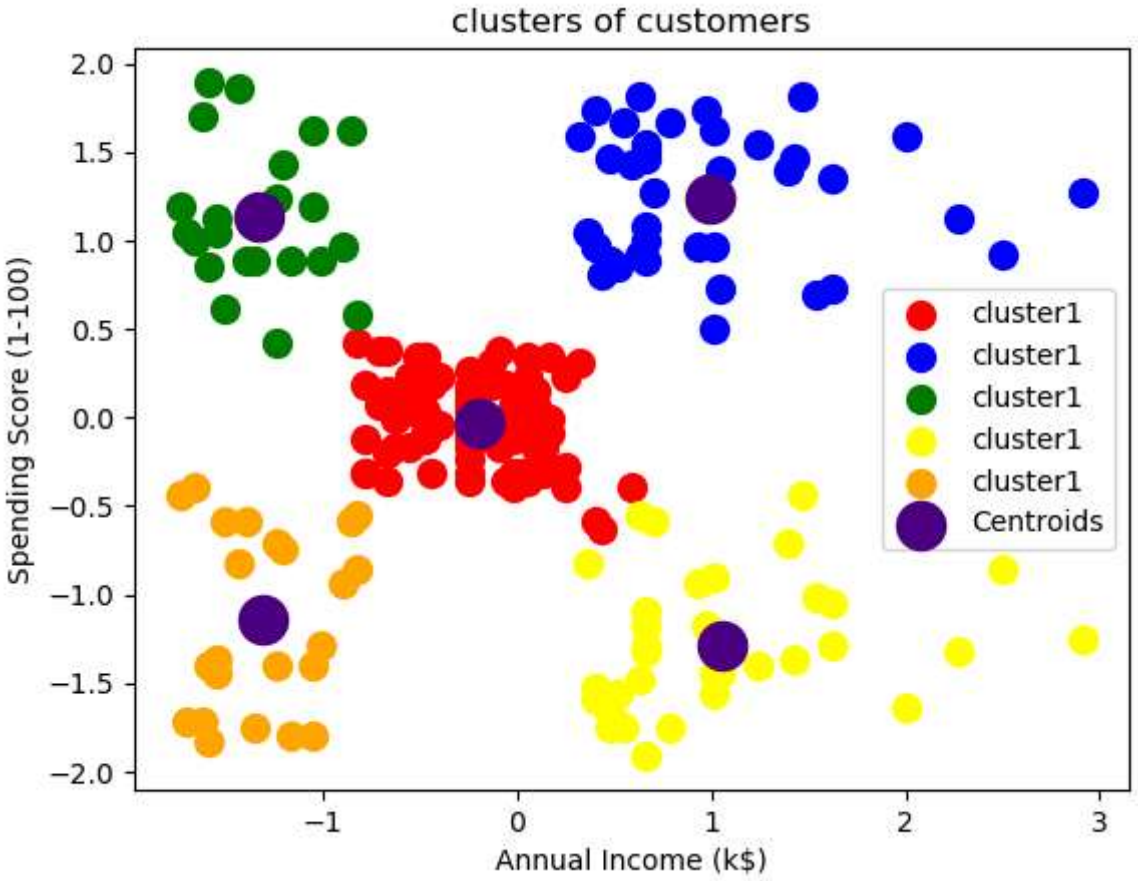
In [19]:
```
clusters
```

Out[19]:
```
array([4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2,
       4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 0,
       4, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 1, 0, 1, 3, 1, 3, 1,
       0, 1, 3, 1, 3, 1, 3, 1, 3, 1, 0, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1,
       3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1,
       3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1,
       3, 1])
```

In [20]:
```
df['Cluster'] = clusters # creating a new column name is Cluster in that colum
```

Visualizing the Customer Clusters

In [21]:
```
plt.scatter(scaled_features[clusters == 0,0], scaled_features[clusters == 0,1]
plt.scatter(scaled_features[clusters == 1,0], scaled_features[clusters == 1,1]
plt.scatter(scaled_features[clusters == 2,0], scaled_features[clusters == 2,1]
plt.scatter(scaled_features[clusters == 3,0], scaled_features[clusters == 3,1]
plt.scatter(scaled_features[clusters == 4,0], scaled_features[clusters == 4,1]
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=30
plt.title ('clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```



In [22]:
```
kmeans.cluster_centers_
```

Out[22]:
```
array([[-0.20091257, -0.02645617],
       [ 0.99158305,  1.23950275],
       [-1.32954532,  1.13217788],
       [ 1.05500302, -1.28443907],
       [-1.30751869, -1.13696536]])
```

Final Clustered Dataset

In [23]:
```
df
```

Out[23]:

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

200 rows × 6 columns

Final result :

In this project, we successfully segmented mall customers into meaningful groups based on their spending habits and demographic data using K-Means Clustering. This segmentation helps businesses understand customer behavior and enables targeted marketing strategies for different shopper types.

In [ ]:
```

```