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

course_data=pd.read_csv('/content/online_course_engagement_data.csv')

course_data.head()
```


```
{"summary":"{\n  \"name\": \"course_data\",\n  \"rows\": 9000,\n
\"fields\": [\n    {\n      \"column\": \"UserID\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
2596,\n        \"min\": 1,\n        \"max\": 9000,\n
\"num_unique_values\": 8123,\n        \"samples\": [\n          7442,\
n          6420,\n          4414\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"CourseCategory\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"Arts\",\n          \"Business\",\n          \"Science\"\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"TimeSpentOnCourse\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
28.491749674819843,\n        \"min\": 1.005229866297383,\n
\"max\": 99.99255785648448,\n        \"num_unique_values\": 8123,\n
\"samples\": [\n          54.05766331977805,\n
92.11859331850364,\n          42.005854876159695\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"NumberOfVideosWatched\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
6,\n        \"min\": 0,\n        \"max\": 20,\n
\"num_unique_values\": 21,\n        \"samples\": [\n          17,\n
19,\n          5\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"NumberOfQuizzesTaken\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 3,\n        \"min\": 0,\n
\"max\": 10,\n        \"num_unique_values\": 11,\n        \"samples\":
[\n          7,\n          3,\n          8\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"QuizScores\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
14.378382907872654,\n        \"min\": 50.00511862629234,\n
\"max\": 99.99498421511456,\n        \"num_unique_values\": 8123,\n
\"samples\": [\n          63.96710820778133,\n
60.3213419557592,\n          80.84957603361214\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"CompletionRate\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
28.950976960764276,\n        \"min\": 0.0093268021242876,\n
```

\"max\": 99.97971128119624,\n   \"num_unique_values\": 8123,\n
\"samples\": [\n   39.02351686669386,\n
92.43269630426978,\n   59.23725919173376\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"DeviceType\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
0,\n   \"min\": 0,\n   \"max\": 1,\n
\"num_unique_values\": 2,\n   \"samples\": [\n   0,\n
1\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"CourseCompletion\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 0,\n   \"min\": 0,\n
\"max\": 1,\n   \"num_unique_values\": 2,\n   \"samples\":
[\n   1,\n   0\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   }\n]\
n}","type":"dataframe","variable_name":"course_data"}

```
course_data.shape
```

```
(9000, 9)
```

```
course_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9000 entries, 0 to 8999
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   UserID                 9000 non-null   int64  
 1   CourseCategory         9000 non-null   object 
 2   TimeSpentOnCourse      9000 non-null   float64
 3   NumberOfVideosWatched  9000 non-null   int64  
 4   NumberOfQuizzesTaken   9000 non-null   int64  
 5   QuizScores             9000 non-null   float64
 6   CompletionRate         9000 non-null   float64
 7   DeviceType             9000 non-null   int64  
 8   CourseCompletion       9000 non-null   int64  
dtypes: float64(3), int64(5), object(1)
memory usage: 632.9+ KB
```

```
course_data.isnull().sum()
```

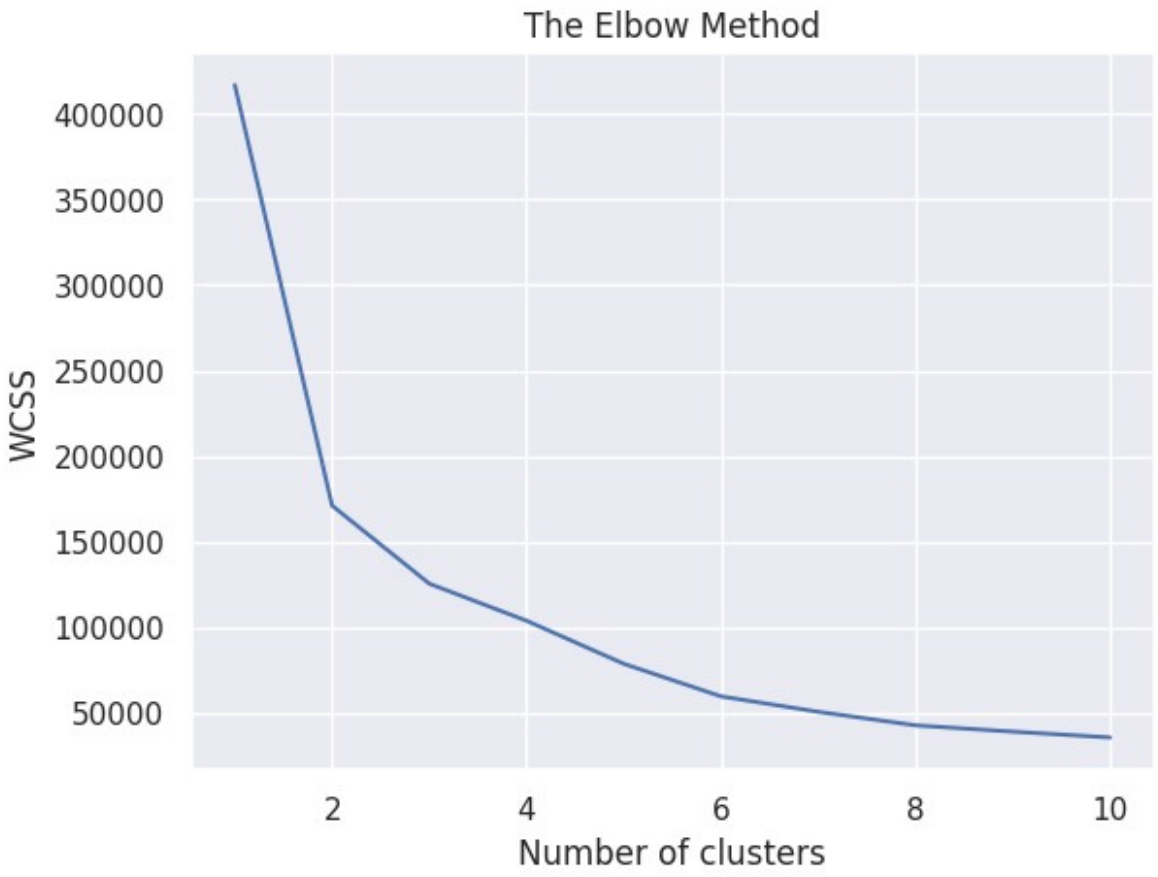
```
UserID                   0
CourseCategory           0
TimeSpentOnCourse        0
NumberOfVideosWatched    0
NumberOfQuizzesTaken     0
QuizScores               0
CompletionRate           0
DeviceType               0
```

```
CourseCompletion          0
dtype: int64
```

```
MD=course_data.iloc[:,[3,4]].values
print(MD)
```

```
[[17  3]
 [ 1  5]
 [14  2]
 ...
 [ 3  3]
 [13 10]
 [ 7  5]]
```

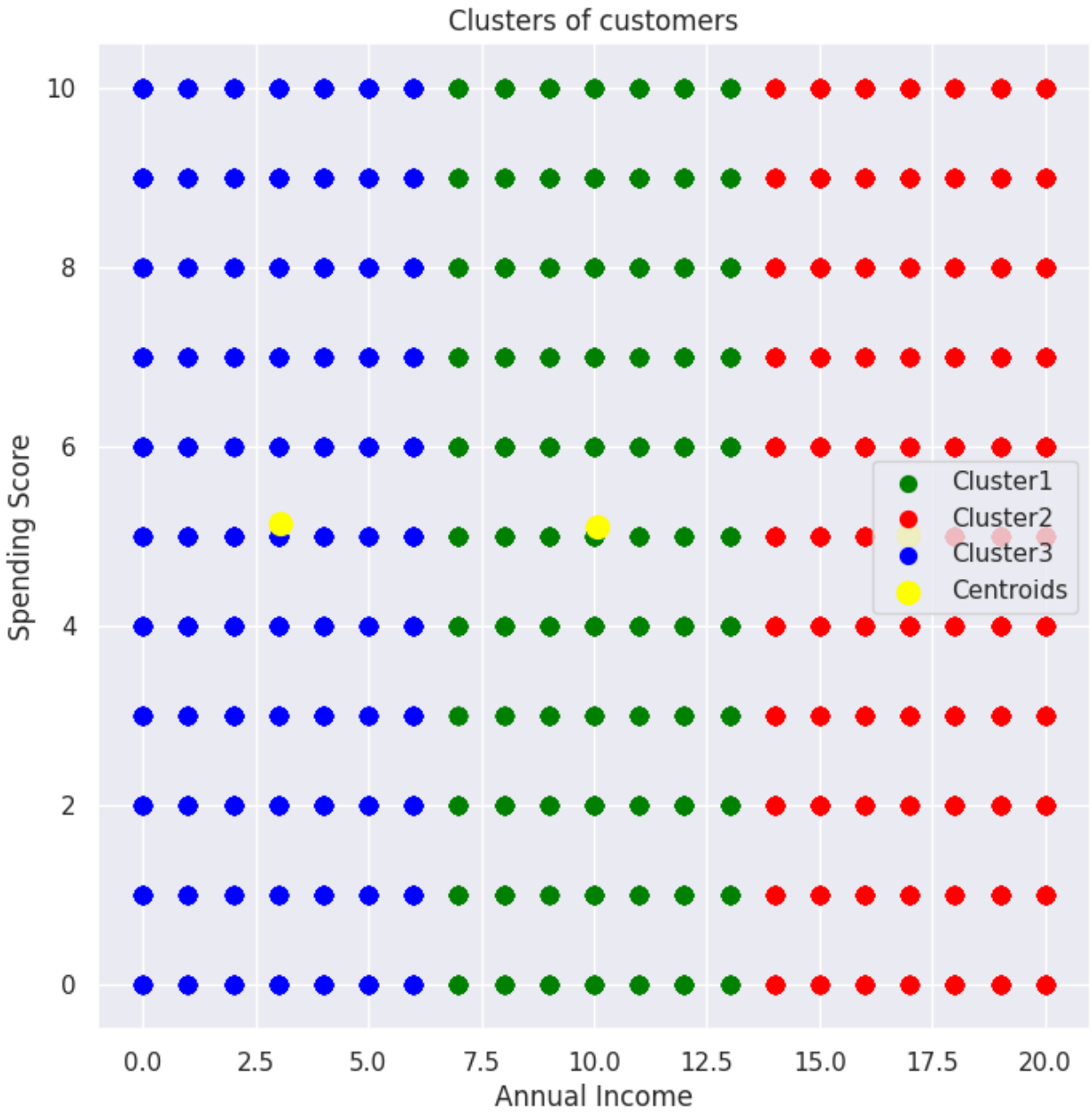
```
wcss=[]
for i in range(1,11):
  kmeans=KMeans(n_clusters=i,init='k-means++',random_state=42)
  kmeans.fit(MD)
  wcss.append(kmeans.inertia_)

sns.set()
plt.plot(range(1,11),wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



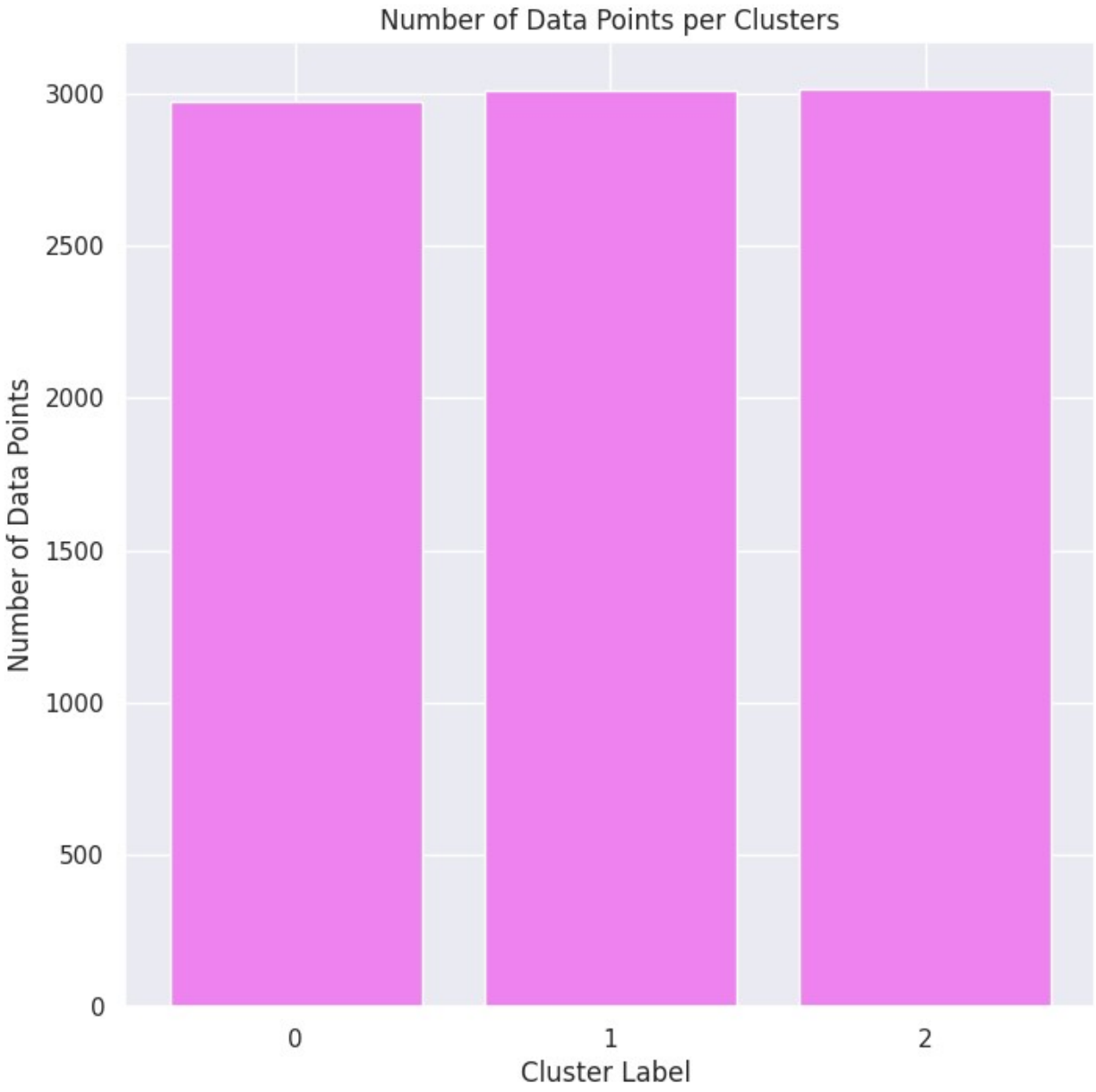

```
kmeans=KMeans(n_clusters=3,init='k-means++',random_state=42)
y=kmeans.fit_predict(MD)
print(y)
```

```
[1 2 1 ... 2 0 0]
```

```
plt.figure(figsize=(8,8))
plt.scatter(MD[y==0,0],MD[y==0,1],s=50,c='green',label='Cluster1')
plt.scatter(MD[y==1,0],MD[y==1,1],s=50,c='red',label='Cluster2')
plt.scatter(MD[y==2,0],MD[y==2,1],s=50,c='blue',label='Cluster3')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],
s=100,c='yellow',label='Centroids')
plt.title('Clusters of customers')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.legend()
plt.show()
```




```
cluster_counts=pd.Series(y).value_counts().sort_index()
plt.figure(figsize=(8,8))
plt.bar(cluster_counts.index,cluster_counts.values,color="violet")
plt.title("Number of Data Points per Clusters")
plt.xlabel("Cluster Label ")
plt.ylabel("Number of Data Points")
plt.xticks(cluster_counts.index)
plt.show()
```


Number of Data Points per Clusters
Number of Data Points
3000
2500
2000
1500
1000
500
0
0
1
2
Cluster Label