

Pranav Nair
2019130042
TE Comps Batch – C
15th November, 2021

Experiment-5

Aim: To train a machine learning model using K-means clustering algorithm to find the optimal value of number of clusters for the annual income and spending score of Mall Customers found in the dataset.

Code:

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

df = pd.read_csv('Mall_Customers.csv')
# df = df.iloc[1400:1500]
data_for_clustering = df[["AnnualIncome", "SpendingScore"]]
data_for_clustering

x = data_for_clustering.values
x

plt.scatter(data_for_clustering.AnnualIncome.to_list() ,
data_for_clustering.SpendingScore.to_list())
plt.title("Mall customers Spending Score")
plt.xlabel("Annual Income")
plt.ylabel("Spending Score")
plt.show()

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

    return wcss_list

wcss = get_wcss(x)
print(wcss)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method Graph')
plt.xlabel('Number of clusters(k)')
plt.ylabel('wcss_list')
plt.show()
```

```

def clustering_kmeans(X,k):
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state= 42)
    y= kmeans.fit_predict(X)
    return kmeans,y

# Using the Elbow method the optimal value of cluster(k) is 5 for the given
dataset
k_means, y = clustering_kmeans(x, 5)

plt.scatter(x[y == 0, 0], x[y == 0, 1], s = 100, c = 'blue', label = 'Cluster
1') #for first cluster
plt.scatter(x[y == 1, 0], x[y == 1, 1], s = 100, c = 'green', label = 'Cluster
2') #for second cluster
plt.scatter(x[y== 2, 0], x[y == 2, 1], s = 100, c = 'red', label = 'Cluster
3') #for third cluster
plt.scatter(x[y == 3, 0], x[y == 3, 1], s = 100, c = 'cyan', label = 'Cluster
4') #for fourth cluster
plt.scatter(x[y == 4, 0], x[y == 4, 1], s = 100, c = 'magenta', label =
'Cluster 5') #for fifth cluster
plt.scatter(k_means.cluster_centers_[ :, 0], k_means.cluster_centers_[ :, 1], s
= 300, c = 'yellow', label = 'Centroid')
plt.title('Customers Cluster')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.legend()
plt.show()

```

Observation:

Task 1:

Below is the WCSS score for cluster values from 1-10:

```

[1, 269981.280000000014]
[2, 181363.59595959607]
[3, 106348.37306211119]
[4, 73679.78903948837]
[5, 44448.45544793369]
[6, 37233.81451071002]
[7, 30259.657207285458]
[8, 25011.839349156595]
[9, 21850.16528258562]
[10, 19672.07284901432]

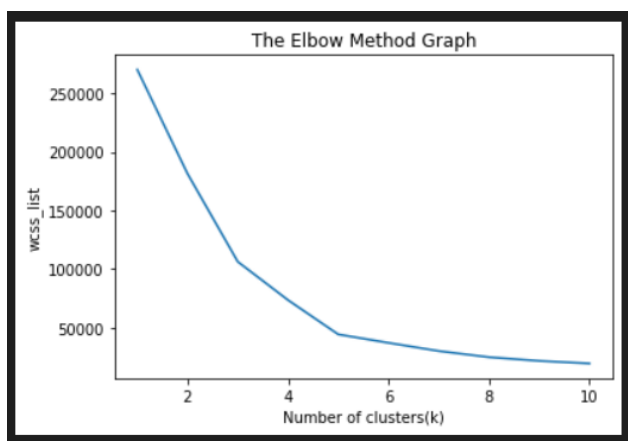
```

The data points and cluster centroids for cluster k=5:

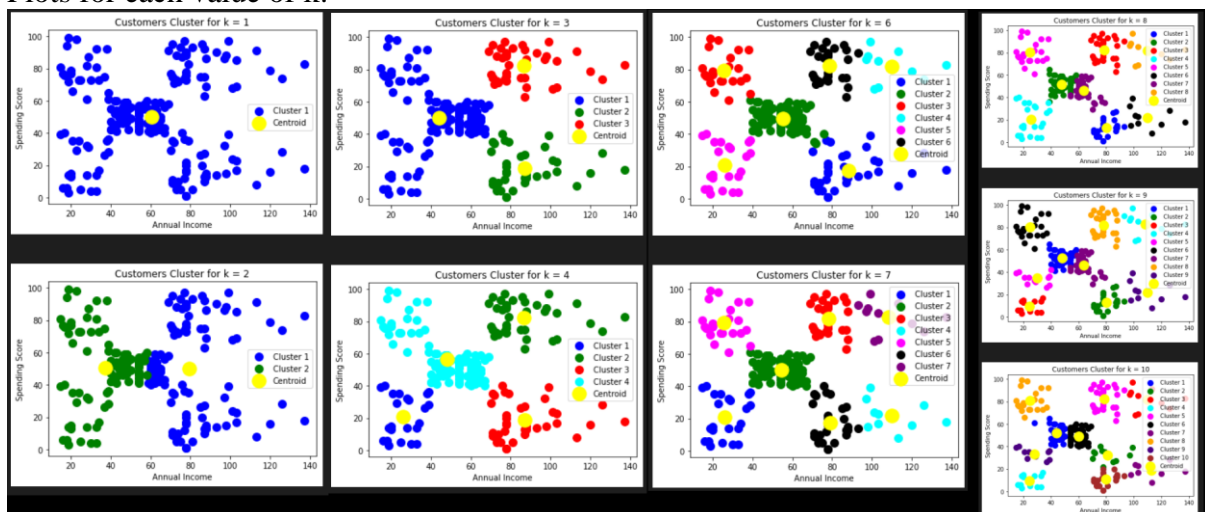


Task 2:

From the below Elbow Method graph which is wcss values plotted against each cluster value k we observe that the optimal value of k is 5 and the plot for k=5 is below.



Plots for each value of k:



Conclusion:

GitHub: <https://github.com/pranav567/AI-ML-Lab/tree/master/experiment-5>

DataSet: <https://www.kaggle.com/shwetabh123/mall-customers>