BÀI TUẦN 11: HỌC KHÔNG GIÁM SÁT THUẬT TOÁN K – MEAN CLUSTERING

1. Thông tin sinh viên

- **Họ và Tên: Dương Minh Lượng**
- **MSSV: 18521071**

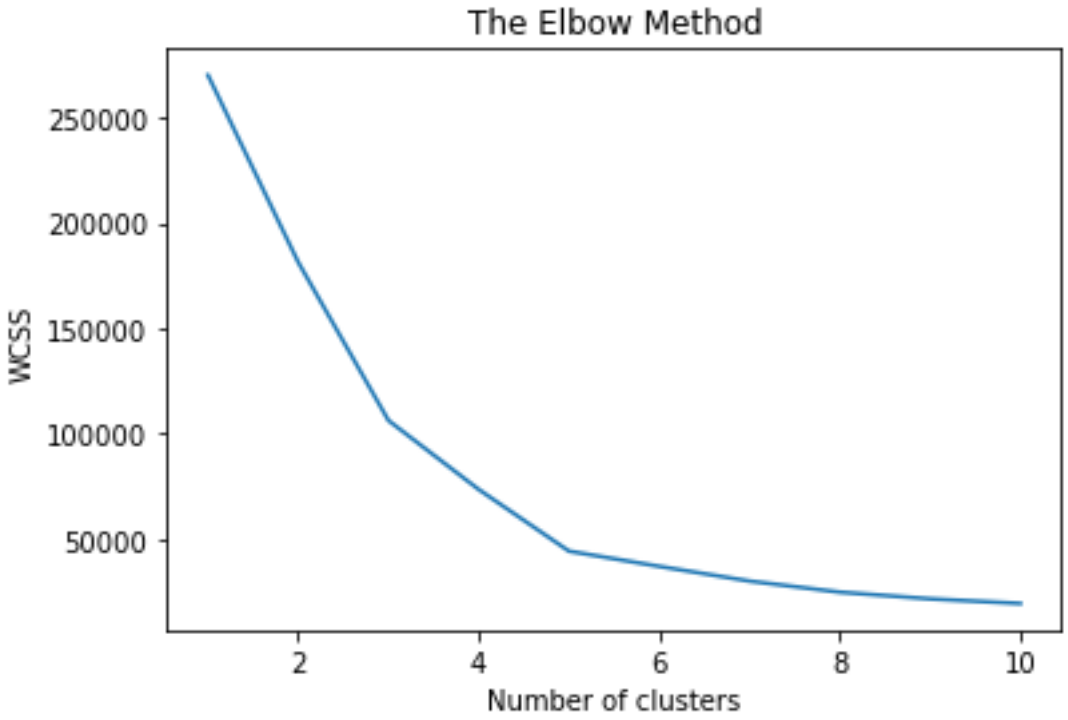
2. Source

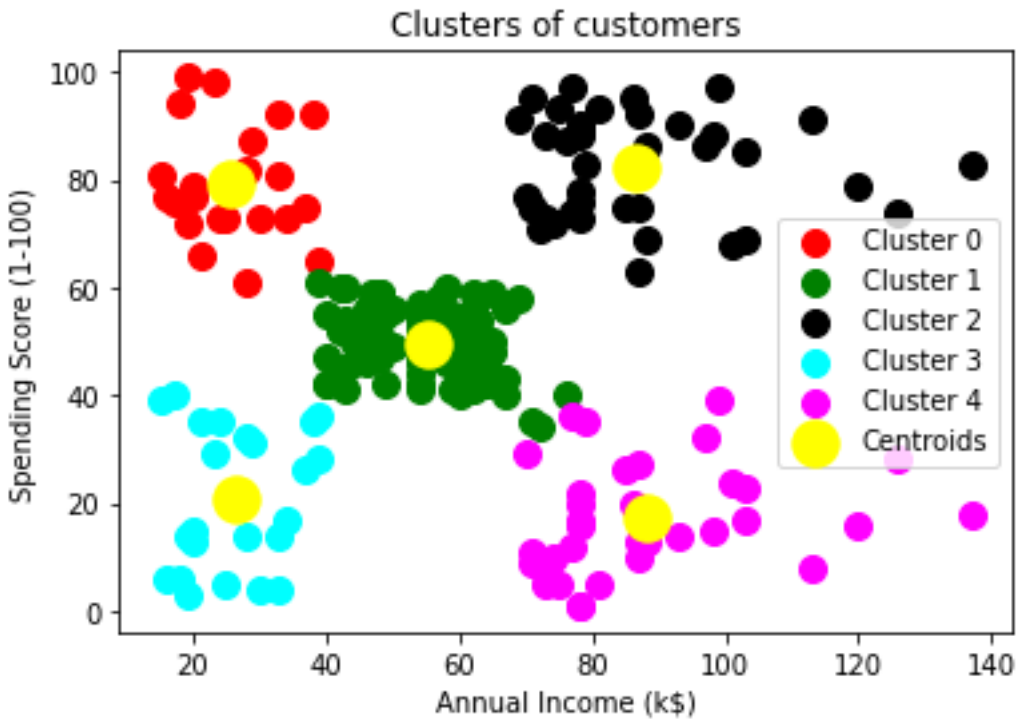
```
1. import numpy as np
2. import matplotlib.pyplot as plt
3. import pandas as pd
4.
5. # Importing the dataset
6. dataset = pd.read_csv('Mall_Customers.csv')
7. X = dataset.iloc[:, [3, 4]].values
8. from sklearn.cluster import KMeans
9. wcss = []
10.     for i in range(1, 11):
11.         kmeans = KMeans(n_clusters = i, init = 'k-
   means++', random_state = 0)
12.         kmeans.fit(X)
13.         wcss.append(kmeans.inertia_)
14.     plt.plot(range(1, 11), wcss)
15.     plt.title('The Elbow Method')
16.     plt.xlabel('Number of clusters')
17.     plt.ylabel('WCSS')
18.     plt.show()
19.     kmeans = KMeans(n_clusters = 5, init = 'k-
   means++', random_state = 42)
20.     y_kmeans = kmeans.fit_predict(X)
21.     from matplotlib.colors import ListedColormap
22.     raw_colors = ("red", "green", "black", "cyan",
   "magenta")
23.     colors = ListedColormap(raw_colors)
24.     for i in range(5):
25.         plt.scatter(X[y_kmeans == i, 0], X[y_kmeans
   == i, 1], s= 100, c = colors(i),label = 'Cluster
   '+str(i))
26.     X_clusters = kmeans.cluster_centers_[:, 0]
27.     Y_clusters = kmeans.cluster_centers_[:, 1]
28.     plt.scatter(X_clusters, Y_clusters, s = 300, c
   = 'yellow', label = 'Centroids')
```

```
29.     plt.title('Clusters of customers')
30.     plt.xlabel('Annual Income (k$)')
31.     plt.ylabel('Spending Score (1-100)')
32.     plt.legend()
33.     plt.show()
```

3. Kết quả

Trực quan hóa giá trị lỗi của kmeans theo phương pháp khuỷu tay.





Nhận xét:

- Giá trị hàm lỗi giảm mạnh từ k = 1 đến k = 5.
- Giá trị hàm lỗi giảm nhẹ dần từ k = 6 trở đi. Do đó, ta thấy k = 5 là số lượng cụm hợp lý.
- Thuật toán đảm bảo hội tụ.
- Thuật toán chạy không tốt khi dữ liệu có số chiều lớn.