

## Ex3: Shopping Data

Cho dữ liệu `shopping_data.csv`, thực hiện việc phân nhóm dữ liệu theo KMeans Clustering theo 2 thuộc tính là Annual Income (k\$) và Spending Score (1-100)

1. Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần)
2. Trực quan hóa dữ liệu
3. Áp dụng Elbow tìm k
4. Áp dụng thuật toán K-Means để giải bài toán phân cụm theo K
5. Trực quan hóa kết quả, nhận xét

```
In [1]: # from google.colab import drive
# drive.mount("/content/gdrive", force_remount=True)
```

```
In [2]: # %cd '/content/gdrive/My Drive/LDS6_MachineLearning/practice_2023/Chapter9_Kmeans/'
```

```
In [3]: import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from sklearn import metrics
from scipy.spatial.distance import cdist
```

```
In [4]: df = pd.read_csv("shopping_data.csv")
df.head()
```

```
Out[4]:
```

	CustomerID	Genre	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

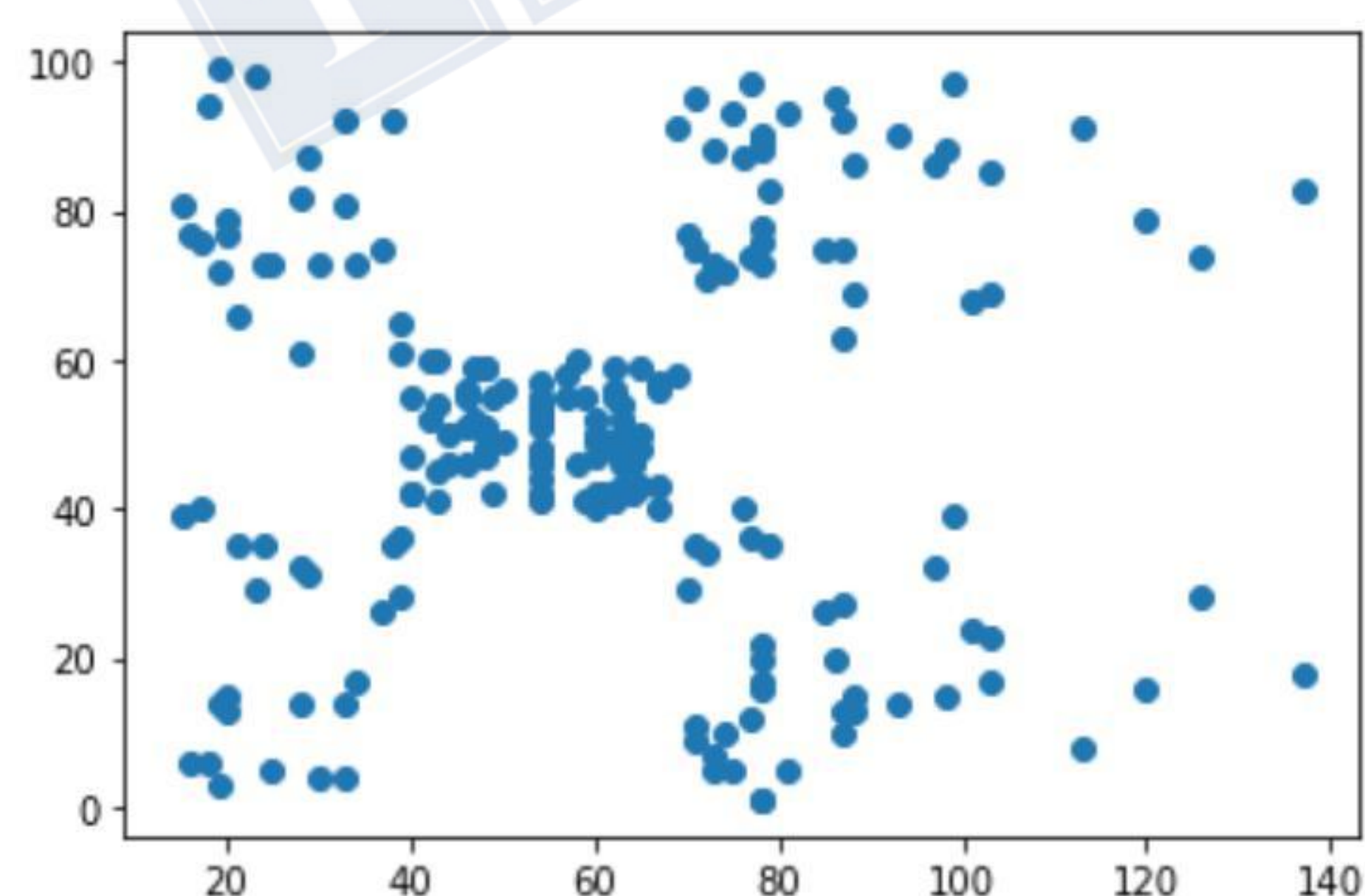
```
In [5]: df_new = df.iloc[:, 3:5]
df_new.head()
```

```
Out[5]:
```

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

```
In [6]: plt.scatter(df_new['Annual Income (k$)'], df_new['Spending Score (1-100)'])
```

```
Out[6]: <matplotlib.collections.PathCollection at 0x1974a82b630>
```

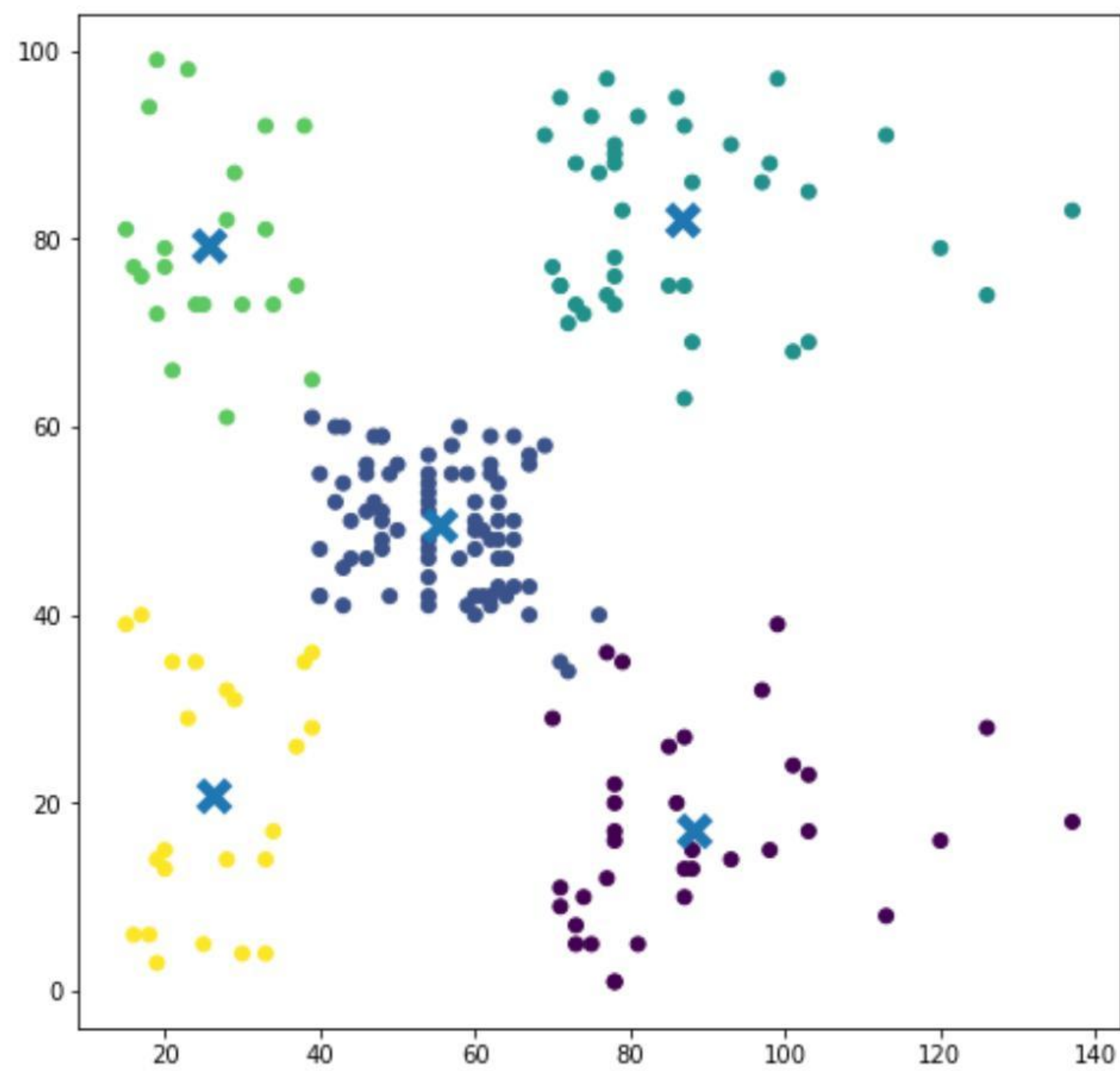








```
In [10]: plt.figure(figsize=(8,8))
plt.scatter(centroids[:, 0],centroids[:, 1],
            marker = "x", s=150, linewidths = 5, zorder = 10)
plt.scatter(df_new['Annual Income (k$)',
df_new['Spending Score (1-100)'], c = df_new.Group)
# dat x, dat x
plt.show()
```



Giải thích cụ thể từng cụm.

Nếu bây giờ phân cụm theo:

- Annual Income + Spending Score + Age => ? cụm => Giải thích
- Annual Income + Spending Score + Gender => ? cụm => Giải thích
- Annual Income + Spending Score + Age + Gender => ? cụm => Giải thích