### Importing required modules

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

### Reading csv file

In [2]:

```
df=pd.read_csv('Mall_Customers.csv')
df.head()
```

Out[2]:

| | 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 [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

**There will be 200 records and 5 columns not having null values.**

In [4]:

```
df.describe()
```

Out[4]:

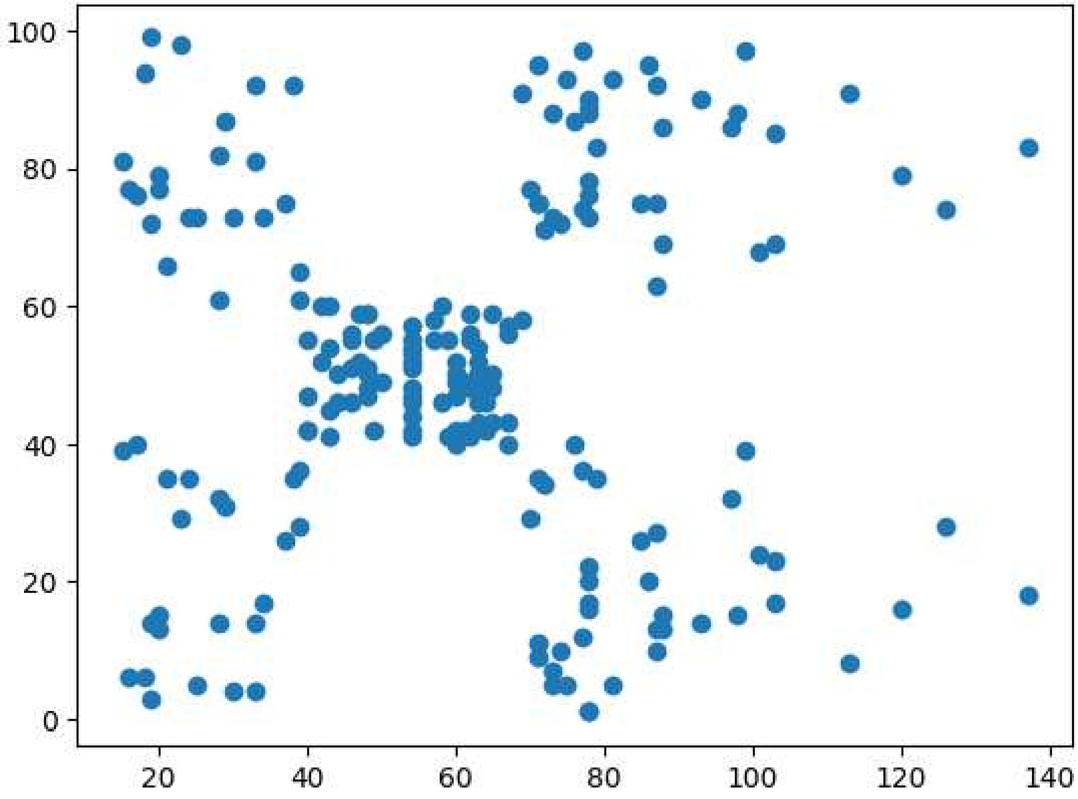
| | CustomerID | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| **count** | 200.000000 | 200.000000 | 200.000000 | 200.000000 |
| **mean** | 100.500000 | 38.850000 | 60.560000 | 50.200000 |
| **std** | 57.879185 | 13.969007 | 26.264721 | 25.823522 |
| **min** | 1.000000 | 18.000000 | 15.000000 | 1.000000 |
| **25%** | 50.750000 | 28.750000 | 41.500000 | 34.750000 |
| **50%** | 100.500000 | 36.000000 | 61.500000 | 50.000000 |
| **75%** | 150.250000 | 49.000000 | 78.000000 | 73.000000 |
| **max** | 200.000000 | 70.000000 | 137.000000 | 99.000000 |

### We will performing clustering by taking two features 'Annual Income (k$)' & 'Spending Score (1-100)'

### Scatterplot graph of 'Annual Income (k$)' & 'Spending Score (1-100)'

In [5]:

```
plt.scatter(df['Annual Income (k$)'],df['Spending Score (1-100)'])
plt.show()
```



So, we want to form clusters according similarities and difference.

### Feature Scaling

In [6]:

```python
from sklearn.preprocessing import MinMaxScaler
mm = MinMaxScaler()
df['Annual Income (k$)'] = mm.fit_transform(df[['Annual Income (k$)']])
df['Spending Score (1-100)'] = mm.fit_transform(df[['Spending Score (1-100)']])
```

In [7]:

```python
df.head()
```

Out[7]:

| | CustomerID | Genre | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 0.000000 | 0.387755 |
| **1** | 2 | Male | 21 | 0.000000 | 0.816327 |
| **2** | 3 | Female | 20 | 0.008197 | 0.051020 |
| **3** | 4 | Female | 23 | 0.008197 | 0.775510 |
| **4** | 5 | Female | 31 | 0.016393 | 0.397959 |

In [8]:

```python
x = df[['Annual Income (k$)','Spending Score (1-100)']]
x
```

Out[8]:

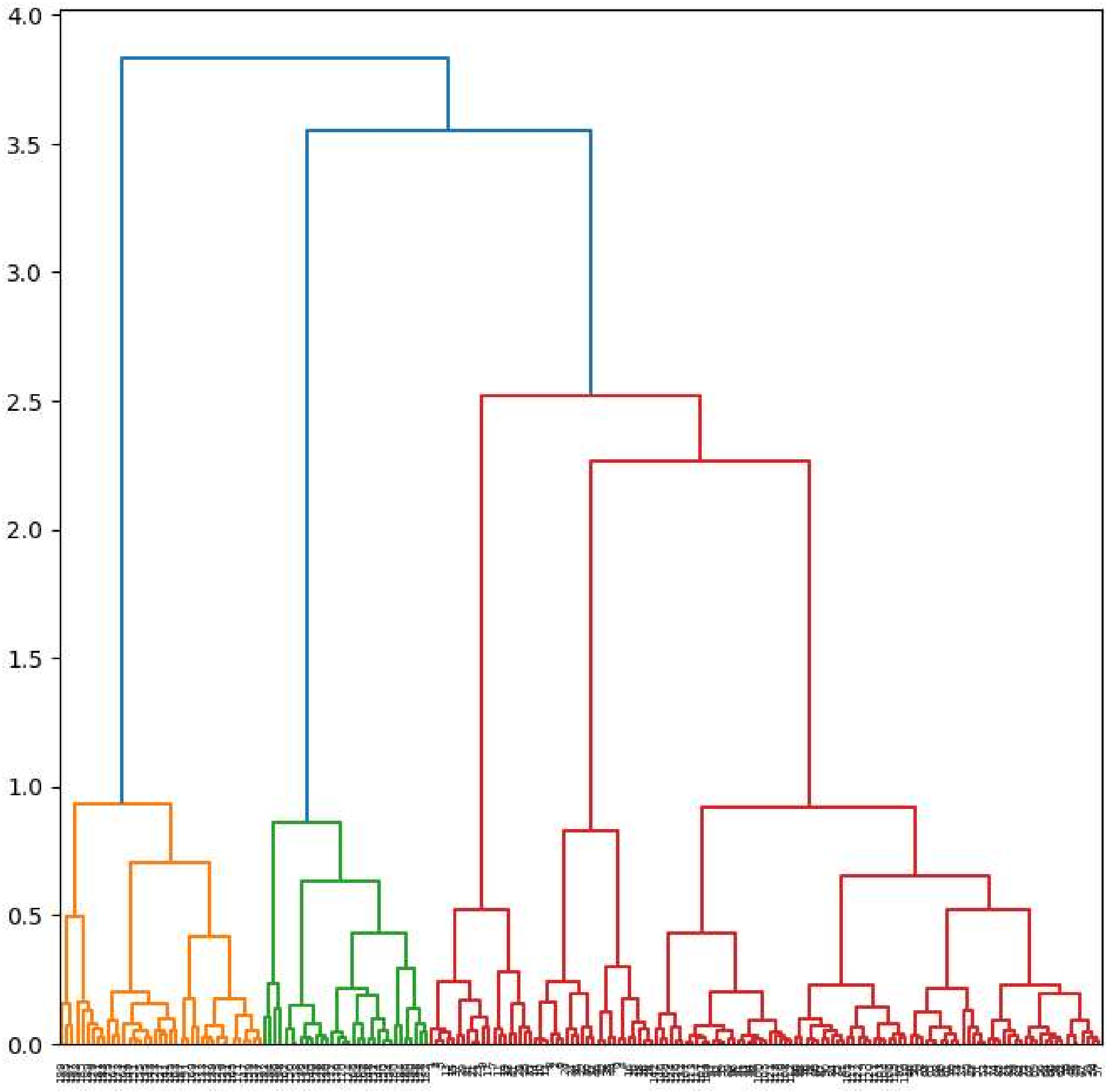
| | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|
| **0** | 0.000000 | 0.387755 |
| **1** | 0.000000 | 0.816327 |
| **2** | 0.008197 | 0.051020 |
| **3** | 0.008197 | 0.775510 |
| **4** | 0.016393 | 0.397959 |
| **...** | ... | ... |
| **195** | 0.860656 | 0.795918 |
| **196** | 0.909836 | 0.275510 |
| **197** | 0.909836 | 0.744898 |
| **198** | 1.000000 | 0.173469 |
| **199** | 1.000000 | 0.836735 |

200 rows × 2 columns

### Hierarchical graph of x

In [9]:

```
import scipy.cluster.hierarchy as sch
plt.figure(figsize=(8,8))
lk = sch.linkage(x,method = 'ward')
ddg = sch.dendrogram(lk)
```



**By observation, In Hierarchical graph there will be 5 vertical big lines observed. So we can take no. of clusters is 5**

In [10]:

```
from sklearn.cluster import AgglomerativeClustering
ac = AgglomerativeClustering(n_clusters = 5)
yhc = ac.fit_predict(x)
yhc
```

Out[10]:

```
array([4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3,
       4, 3, 4, 3, 4, 1, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 1,
       4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 2, 0, 2, 0,
       1, 0, 2, 0, 2, 0, 2, 0, 2, 0, 1, 0, 2, 0, 1, 0, 2, 0, 2, 0, 2, 0,
       2, 0, 2, 0, 2, 0, 1, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0,
       2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0,
       2, 0], dtype=int64)
```

### We will creating new feature to form clusters

In [11]:

```
df['HLables'] = yhc
```

In [12]:

```
df.head()
```

Out[12]:

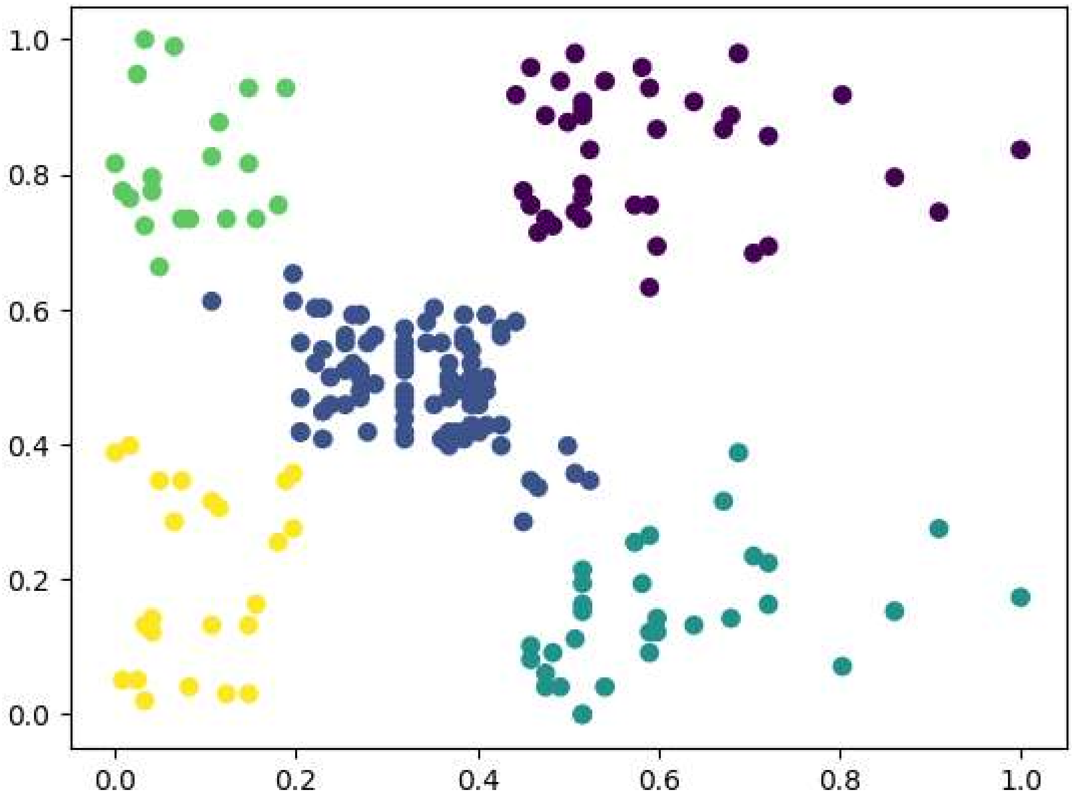
| | CustomerID | Genre | Age | Annual Income (k$) | Spending Score (1-100) | HLables |
|---|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 0.000000 | 0.387755 | 4 |
| **1** | 2 | Male | 21 | 0.000000 | 0.816327 | 3 |
| **2** | 3 | Female | 20 | 0.008197 | 0.051020 | 4 |
| **3** | 4 | Female | 23 | 0.008197 | 0.775510 | 3 |
| **4** | 5 | Female | 31 | 0.016393 | 0.397959 | 4 |

In [13]:

```
plt.scatter(df['Annual Income (k$)'],df['Spending Score (1-100)'], c = ac.labels_)
```

Out[13]:

```
<matplotlib.collections.PathCollection at 0x277b402ecd0>
```



**So there will be 5 clusters form properly**