# Import libraries

```python
In [ ]:  %matplotlib inline
         import matplotlib.pyplot as plt
         import seaborn as sns
         import warnings
         warnings.simplefilter("ignore")
         sns.set()
```

# Import Customer Data

```python
In [ ]:  import pandas as pd
         customers= pd.read_csv("Data/customers.csv")
         customers.head()
```
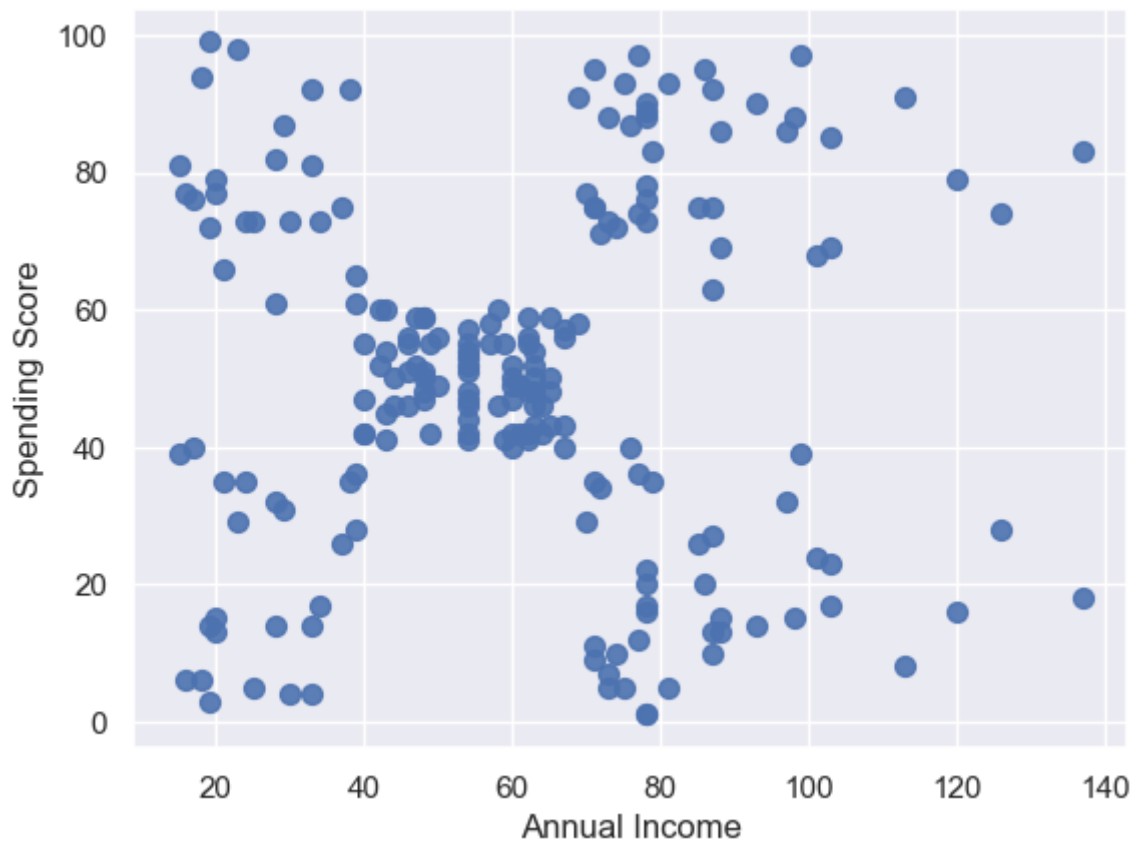
Out[ ]:

|   | CustomerID | Gender | 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 |

# Feature Selection and analysis

```python
In [ ]:  %matplotlib inline
         import matplotlib.pyplot as plt
         import seaborn as sns
         sns.set()
         points = customers.iloc[:,3:5].values
         x=points[:,0]
         y=points[:,1]
         plt.scatter(x,y,s=50,alpha=0.9)
         plt.xlabel("Annual Income")
         plt.ylabel("Spending Score")
```

Out[ ]:  Text(0, 0.5, 'Spending Score')

# Use K-Means Algorithm

In [ ]:
```python
from sklearn.cluster import KMeans
kmeans= KMeans(n_clusters=5,random_state=0)
kmeans.fit(points)
predicted_cluster_indexes=kmeans.predict(points)

plt.scatter(x,y,c=predicted_cluster_indexes,s=50,alpha=0.8,cmap="viridis")
plt.xlabel("Annual Income")
plt.ylabel("Spending Score")

centers=kmeans.cluster_centers_
plt.scatter(centers[:,0],centers[:,1],c="red",s=100)
```

Out[ ]:    <matplotlib.collections.PathCollection at 0x2a8b4c6f880>
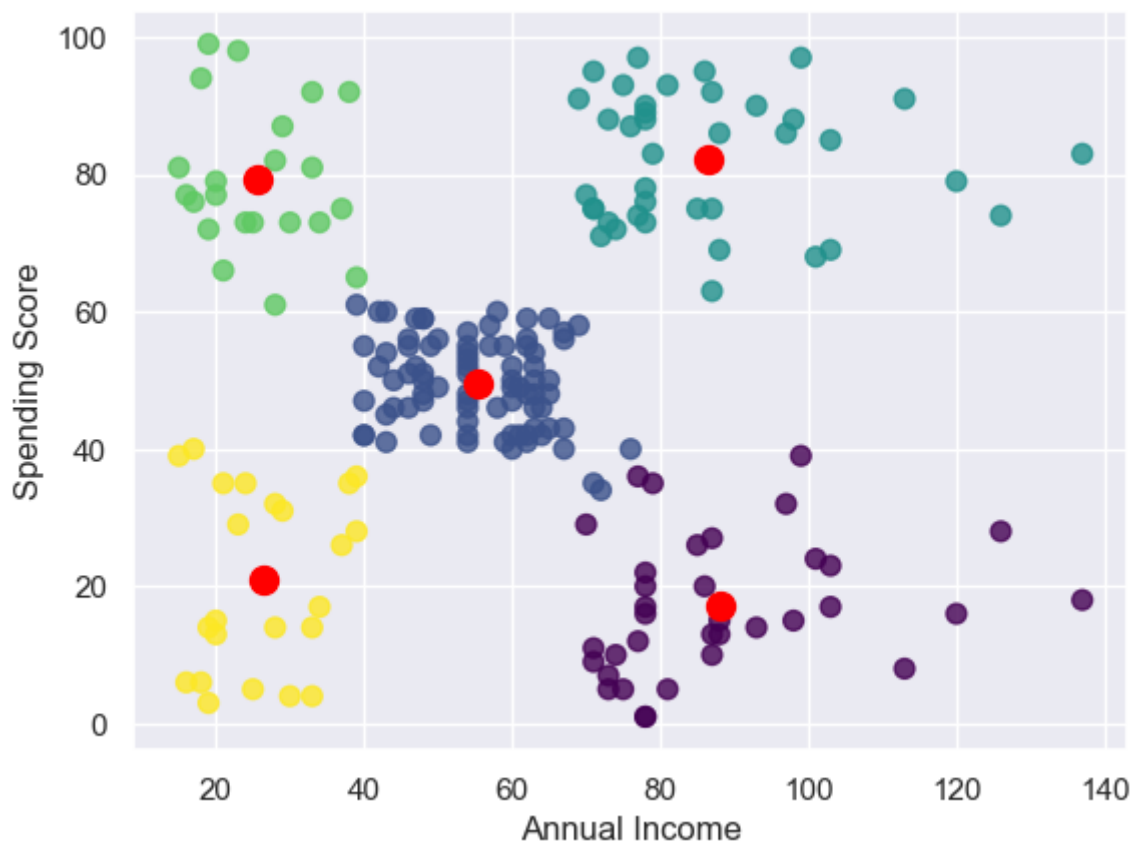
# Add a cluster column in dataset

```
In [ ]:  df= customers.copy()
         df["Cluster"]= kmeans.predict(points)
         df.head()
```

Out[ ]:

| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) | Cluster |
|---|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 15 | 39 | 4 |
| **1** | 2 | Male | 21 | 15 | 81 | 3 |
| **2** | 3 | Female | 20 | 16 | 6 | 4 |
| **3** | 4 | Female | 23 | 16 | 77 | 3 |
| **4** | 5 | Female | 31 | 17 | 40 | 4 |

```
In [ ]:  import numpy as np

         # Get the cluster index for a customer with a high income and low spending score
         cluster = kmeans.predict(np.array([[120, 20]]))[0]

         # Filter the DataFrame to include only customers in that cluster
         clustered_df = df[df['Cluster'] == cluster]

         # Show the customer IDs
         clustered_df['CustomerID'].values
```

```
Out[ ]: array([125, 129, 131, 135, 137, 139, 141, 145, 147, 149, 151, 153, 155,
               157, 159, 161, 163, 165, 167, 169, 171, 173, 175, 177, 179, 181,
               183, 185, 187, 189, 191, 193, 195, 197, 199], dtype=int64)
```

# Above array contains customers IDs with high income and low spending