In [1]:
```
import os
os.getcwd()
```

Out[1]:
```
'/home/sapatevaibhav/Documents/ML'
```

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('Mall_Customers.csv')
```

In [2]:
```
df
```

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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [3]:
```
x = df.iloc[:,3:]
x
```

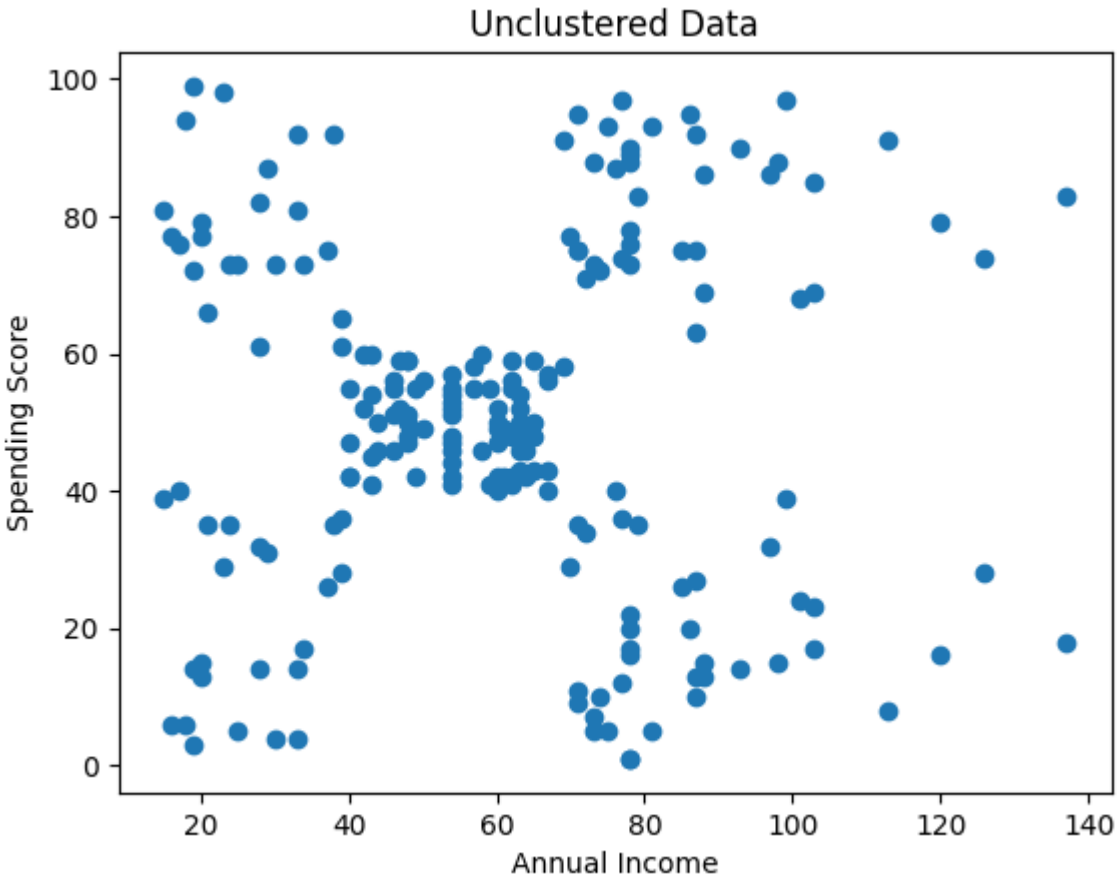
Out[3]:

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40
...	...	...
195	120	79
196	126	28
197	126	74
198	137	18
199	137	83

200 rows × 2 columns

In [4]:
```
plt.title('Unclustered Data')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'], x['Spending Score (1-100)'])
```

Out[4]: <matplotlib.collections.PathCollection at 0x7f69db1bf1c0>



In [5]:
```
from sklearn.cluster import KMeans, AgglomerativeClustering
km = KMeans(n_clusters = 3)
km.fit_predict(x)
```

Out[5]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1], dtype=int32)
```

In [8]:
```
km.inertia_
```

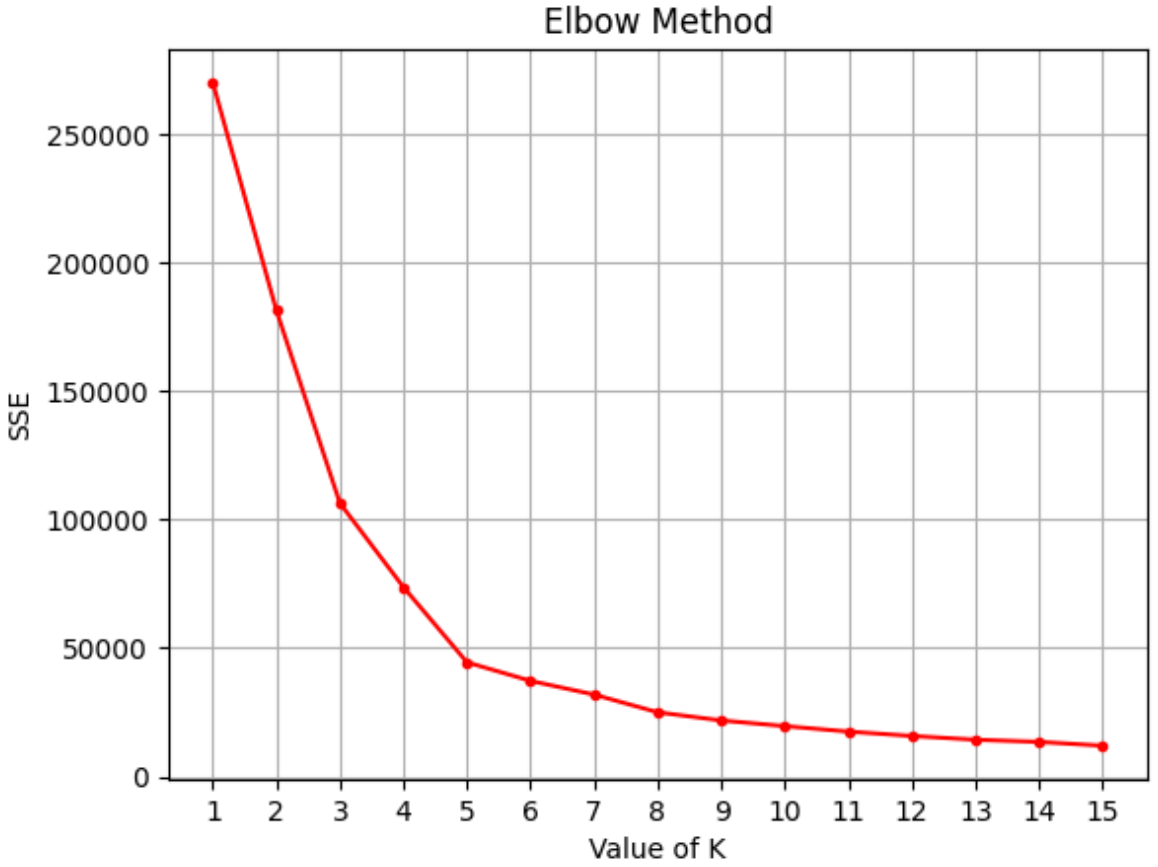
Out[8]:
```
106348.37306211119
```

In [9]:
```
sse = []
for k in range(1,16):
    km = KMeans(n_clusters = k)
    km.fit_predict(x)
    sse.append(km.inertia_)
sse
```

Out[9]:
```
[269981.28,
 181363.59595959596,
 106348.3730621112,
 73679.78903948834,
 44448.45544793371,
 37233.81451071001,
 31902.833868166716,
 25018.576334776335,
 21841.97825674864,
 19646.482018947238,
 17515.872164701657,
 15810.343426539726,
 14350.256180094417,
 13500.42225818557,
 12031.35235407008]
```

In [11]:
```
plt.title('Elbow Method')
plt.xlabel('Value of K')
plt.ylabel('SSE')
plt.grid()
plt.xticks(range(1,16))
plt.plot(range(1,16), sse, marker = '.', color = 'red')
```

Out[11]:
```
[<matplotlib.lines.Line2D at 0x7f69c7cd0b20>]
```



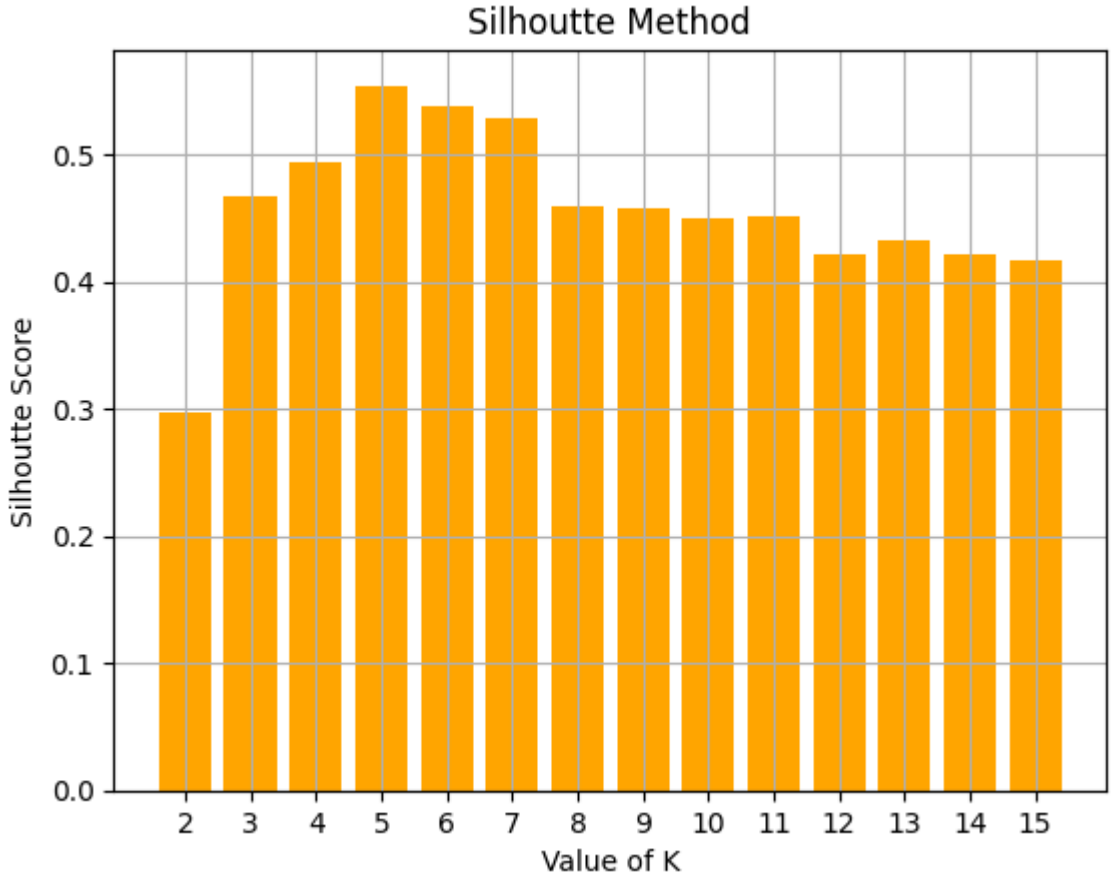
In [24]:
```
from sklearn.metrics import silhouette_score
silh = []
for k in range(2,16):
    km = KMeans(n_clusters = k)
    labels = km.fit_predict(x)
    score = silhouette_score(x, labels)
    silh.append(score)
```

In [13]:
```
silh
```

Out[13]:
```
[0.2968969162503008,
 0.46761358158775435,
 0.4931963109249047,
 0.553931997444648,
 0.5367558067779578,
 0.5264283703685728,
 0.45407359550241166,
 0.452881771814681,
 0.4483975689310094,
 0.44598673856977716,
 0.45899899549134743,
 0.4465256352099726,
 0.4389172146034429,
 0.4235485715125881,
 0.42471704088877155,
 0.42509060245793906,
 0.42937879224702896,
 0.4112114809629708,
 0.42098163607234385,
 0.4117995991918727,
 0.4209256801422762,
 0.4123843490957628,
 0.400817402463384,
 0.4243629093189331]
```

In [26]:
```
plt.title('Silhoutte Method')
plt.xlabel('Value of K')
plt.ylabel('Silhoutte Score')
plt.grid()
plt.xticks(range(2,16))
plt.bar(range(2, 16), silh, color='orange')
```

Out[26]:
```
<BarContainer object of 14 artists>
```

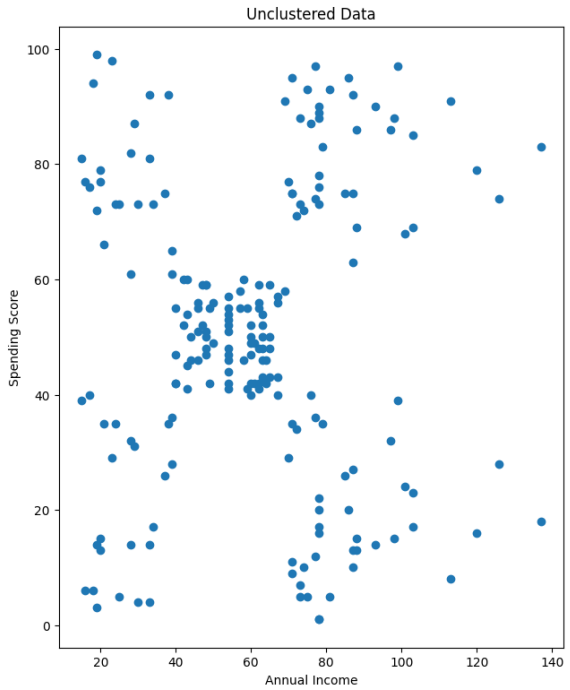


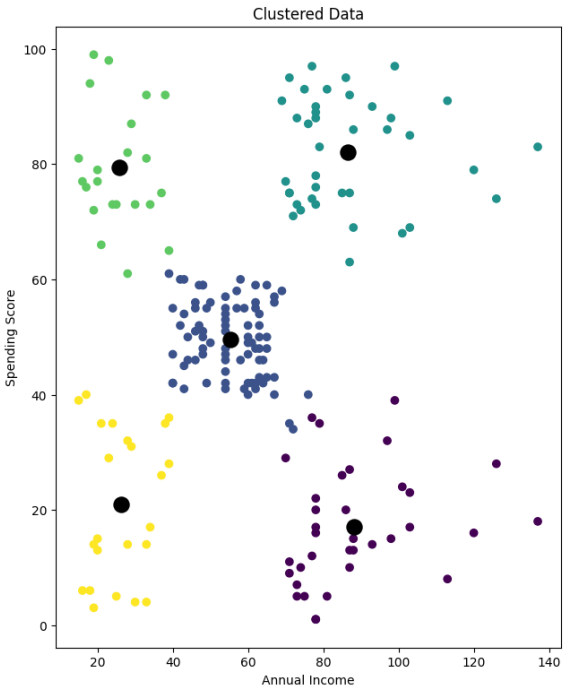
In [38]:
```
km = KMeans(n_clusters = 5, random_state = 0)
labels = km.fit_predict(x)
labels
cent = km.cluster_centers_
```

In [40]:
```
plt.figure(figsize = (16, 9))
plt.subplot(1,2,1)
plt.title('Unclustered Data')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'], x['Spending Score (1-100)'])

plt.subplot(1,2,2)
plt.title('Clustered Data')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'], x['Spending Score (1-100)'], c = labels)
plt.scatter(cent[:,0], cent[:,1], s = 150, color = 'k')
```

Out[40]: <matplotlib.collections.PathCollection at 0x7f69c499f310>





In [42]:
```
km.inertia_
```

Out[42]:
```
44448.45544793371
```

In [44]:
```
four = df[labels==4]
```

In [45]:
```
four.to_csv('mydata.csv')
```

In [46]:
```
km.predict([[56, 61]])
```

```
/home/sapatevaibhav/.local/lib/python3.10/site-packages/sklearn/base.py:450: User
Warning: X does not have valid feature names, but KMeans was fitted with feature
names
  warnings.warn(
```

Out[46]:
```
array([1], dtype=int32)
```

In [47]:
```
agl = AgglomerativeClustering(n_clusters = 5)
alabels = agl.fit_predict(x)
alabels
```

Out[47]:
```
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, 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, 2, 1, 2, 1, 2, 0, 2, 0, 2,
       1, 2, 0, 2, 0, 2, 0, 2, 0, 2, 1, 2, 0, 2, 1, 2, 0, 2, 0, 2, 0, 2,
       0, 2, 0, 2, 0, 2, 1, 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, 2])
```

In [50]:
```
plt.figure(figsize = (16,9))
plt.subplot(1,2,1)
plt.title('Agglomerative')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'], x['Spending Score (1-100)'], c = alabels)
```

```
plt.scatter(cent[:,0], cent[:,1], s = 100, color = 'k')

plt.subplot(1,2,2)
plt.title('KMeans')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'], x['Spending Score (1-100)'], c = labels)
plt.scatter(cent[:,0], cent[:,1], s = 100, color = 'k')
```

Out[50]: <matplotlib.collections.PathCollection at 0x7f69c4c00160>