In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
import os for dirname, _, filenames in
os.walk('/kaggle/input'):     for filename in
filenames:         print(os.path.join(dirname,
filename))
```

In [3]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [8]:
```
df = pd.read_csv(r'C:\Users\admin\Downloads\archive (1)\Mall_Customers.csv' )
```

In [9]:
```
df
```

Out[9]:

	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 [10]:
```
x = df.iloc[:,3:]
x
```

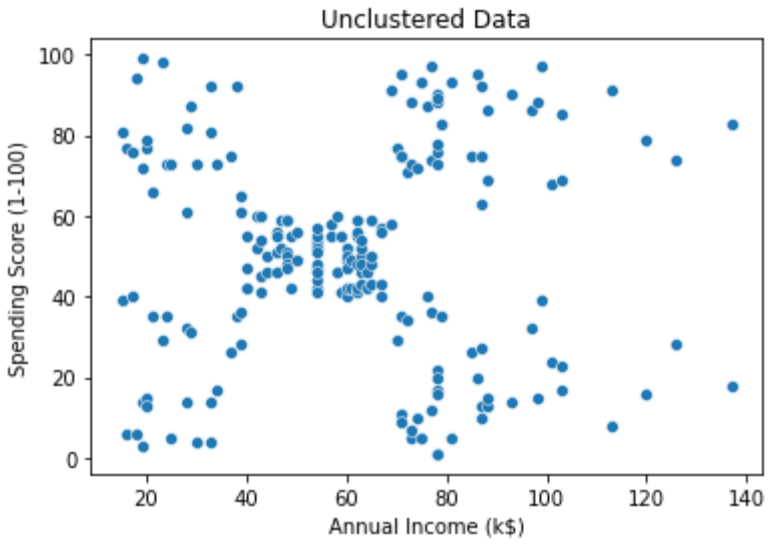
Out[10]:

	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 [11]:
```
plt.title('Unclustered Data')
sns.scatterplot(x=x['Annual Income (k$)'],y=x['Spending Score (1-100)'])
```

Out[11]:
```
<AxesSubplot:title={'center':'Unclustered Data'}, xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'>
```



In [12]:
```
from sklearn.cluster import KMeans, AgglomerativeClustering
```

In [13]:
```
km = KMeans(n_clusters=4)
```

In [14]:
```
km.fit_predict(x)
```

```
C:\Users\admin\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1429: UserWarning: KMeans is known to have a memory leak on W
indows with MKL, when there are less chunks than available threads. You can avoid it by setting the environment variable OMP_NUM_TH
READS=1.
  warnings.warn( array([3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3,
0, 3, 0, 3, 0,
```

Out[14]:
```
       3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 0, 3, 3,
       3, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 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])
```

In [17]:
```
#sse
km.inertia_
```

Out[17]:
```
12152.549592074593
```

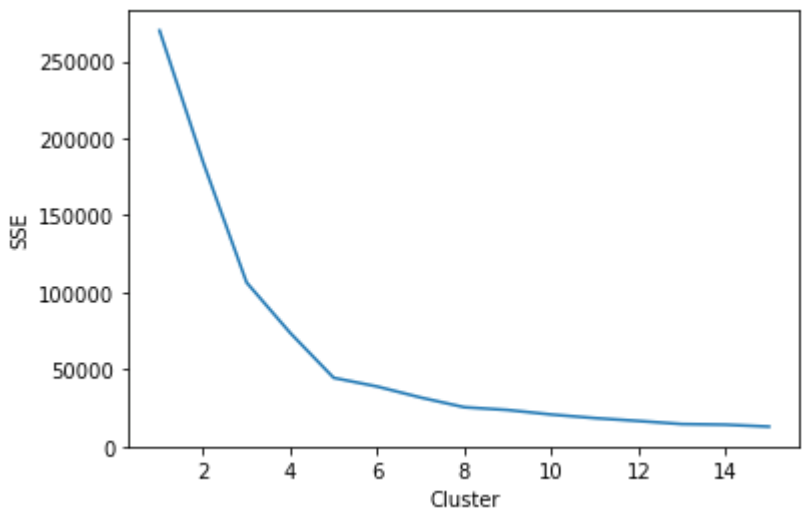
In [18]:
```
sse =[] for k in range(1,16):
 km = KMeans(n_clusters=k)
 km.fit_predict(x)
 sse.append(km.inertia_)
```

In [19]:
```
sse
```

Out[19]:
```
[269981.28,
 184609.98434090617,
 106348.37306211118,
 73679.78903948836,
 44448.45544793371,
38797.9027638142,
 31644.31903792021,
 25354.360937251156,
 23543.475418928974,
20614.671734467087,
 18283.650928500927,
 16429.94737981317,
 14416.328255078253,
14011.958208458209,
 12746.851332869057]
```

In [20]:
```
sns.lineplot(range(1,16),y = sse)
plt.xlabel('Cluster')
plt.ylabel('SSE')
```

```
Text(0, 0.5, 'SSE')
```



In

In [21]:
```
from sklearn.metrics import silhouette_score
```

In [22]:
```
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)
```

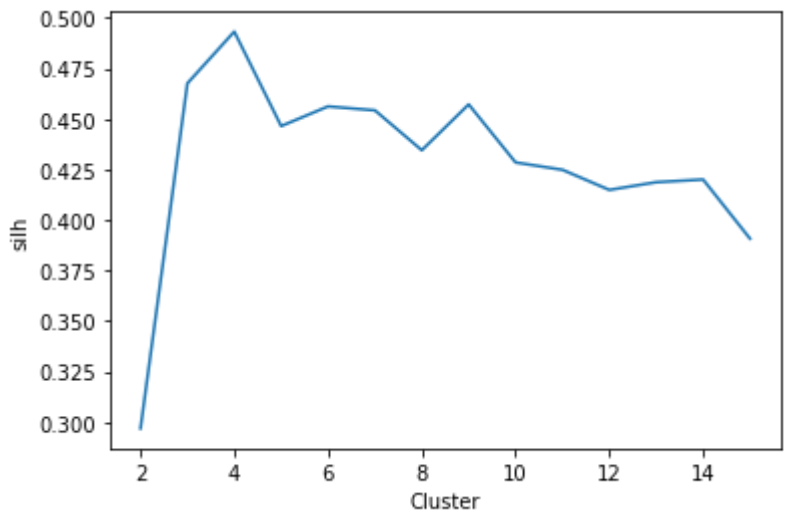
In [23]:
```
silh
```

Out[23]:
```
[0.2968969162503008,
 0.46761358158775435,
0.4931963109249047,
 0.44647974211285657,
0.4561972992633143,
0.45425910793220675,
  0.4345270427308732,
 0.4571957283008456,
 0.4284833830885232,
0.42483696591624037,
0.4148939785943696,
 0.41870926222608434,
 0.42004018597903964,
 0.39079438087316754]
```

In [24]:
```
sns.lineplot(range(2,16),y = silh)
plt.xlabel('Cluster')
plt.ylabel('silh')
```

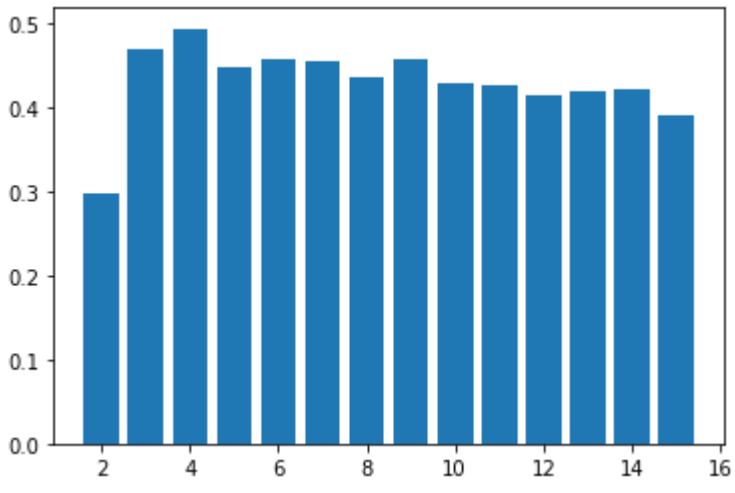
```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg:
x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword wi ll
result in an error or misinterpretation.
  warnings.warn(
Text(0, 0.5, 'silh')
```

Out[24]:



In [25]:
```
plt.bar(range(2,16,1),silh)
```

<BarContainer object of 14 artists >



In [25]:
```
km = KMeans(n_clusters=5,random_state=1)
```

Out[25]:

In [26]:

In [27]:
```
labels = km.fit_predict(x)
```

```
C:\Users\admin\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1429: UserWarning: KMeans is known to have a memory leak on W
indows with MKL, when there are less chunks than available threads. You can avoid it by setting the environment variable OMP_NUM_TH
READS=1.
warnings.warn(
```

In [28]:
```
km.labels_
```

Out[28]:
```
array([4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2,
       4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 0,
4, 2, 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, 3, 1, 3, 0, 3, 1, 3, 1, 3,
0, 3, 1, 3, 1, 3, 1, 3, 1, 3, 0, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3,
       1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3,
1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3,
       1, 3])
```

In [29]:
```
cent = km.cluster_centers_
```

In [30]:
```
plt.title('Clustered Data')
sns.scatterplot(x=x['Annual Income (k$)'],y=x['Spending Score (1-
100)'],c=labels ) sns.scatterplot(cent[:,0],cent[:,1], s=200, color='red')
```

```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x,
y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
<AxesSubplot:title={'center':'Clustered Data'}, xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'> Out[30]:
```

In [31]:


```
df[labels==0]
```

Out[31]:

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
43	44	Female	31	39	61
46	47	Female	50	40	55
47	48	Female	27	40	47
48	49	Female	29	40	42
49	50	Female	31	40	42
	CustomerID	**Genre**	**Age**	**Annual Income (k$)**	**Spending Score (1-100)**
...	...	...	...	...	...
121	122	Female	38	67	40
122	123	Female	40	69	58
126	127	Male	43	71	35
132	133	Female	25	72	34
142	143	Female	28	76	40

81 rows × 5 columns

In [32]:
```
agl = AgglomerativeClustering (n_clusters=5)
```

In [33]:
```
alabels = agl.fit_predict(x)
```

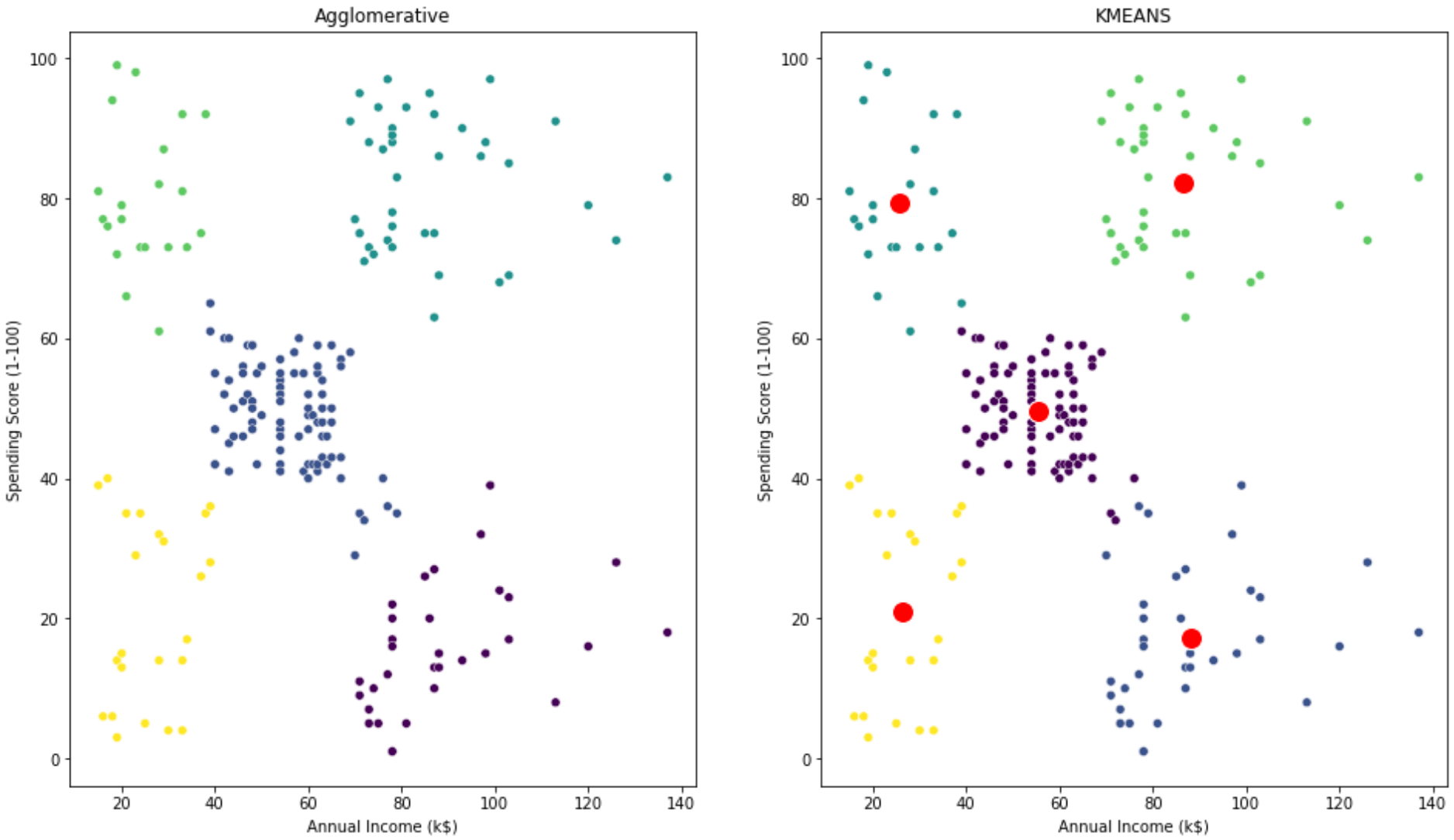
In [34]:
```
alabels
```

Out[34]:
```
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], dtype=int64)
```

In [35]:
```
plt.figure(figsize=(16,9)) plt.subplot(1,2,1)
plt.title('Agglomerative')
sns.scatterplot(x=x['Annual Income (k$)'],y=x['Spending Score (1-100)'], c= alabels)

plt.subplot(1,2,2)
plt.title('KMEANS')
sns.scatterplot(x=x['Annual Income (k$)'],y=x['Spending Score (1-100)'],c=labels ) sns.scatterplot(cent[:,0],cent[:,1],
s=200, color='red')
```

```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x,
y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
<AxesSubplot:title={'center':'KMEANS'}, xlabel='Annual Income (k$)', ylabel='Spending Score (1-100)'> Out[35]:
```


Agglomerative
KMEANS
Spending Score (1-100)
Annual Income (k$)
0
20
40
60
80
100
120
140

In [ ]: