pr5-mall-gs

November 2, 2024

[ ]:
```
pip install pandas
```

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
df = pd.read_csv('Mall_Customers.csv')
df
```

[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 x 5 columns]

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

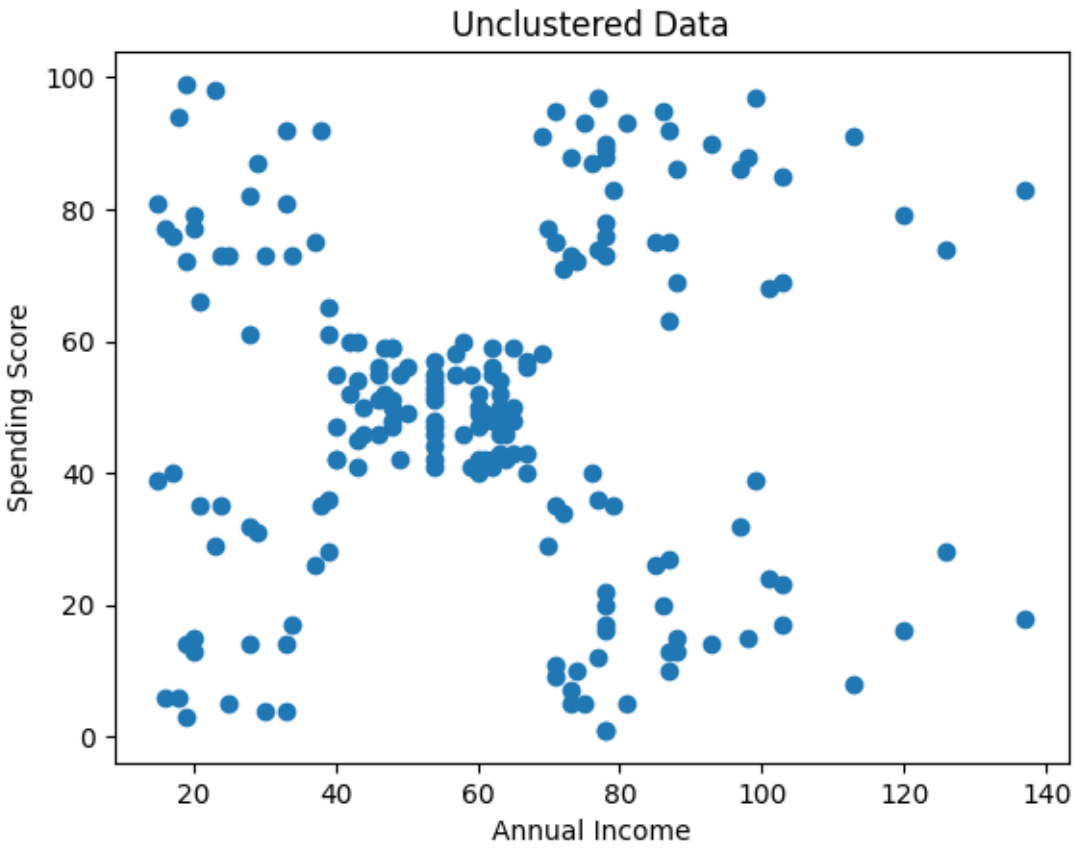
[3]:

	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 x 2 columns]
```

[6]:
```
plt.title('Unclustered Data')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'],x['Spending Score (1-100)']);
```



plt.title(‘Unclustered Data’)

sns.scatterplot(x=x[‘Annual Income (k$)’],y=x[‘Spending Score (1-100)’])

[7]:
```
from sklearn.cluster import KMeans, AgglomerativeClustering
```

[18]:
```
km = KMeans(n_clusters=3)
```

[19]:
```
x.shape
```

[19]: (200, 2)

[20]:
```
km.fit_predict(x)
```

[20]:
```
array([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, 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, 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, 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=int32)
```

[21]:
```
km.inertia_
```

[21]: 106348.37306211119

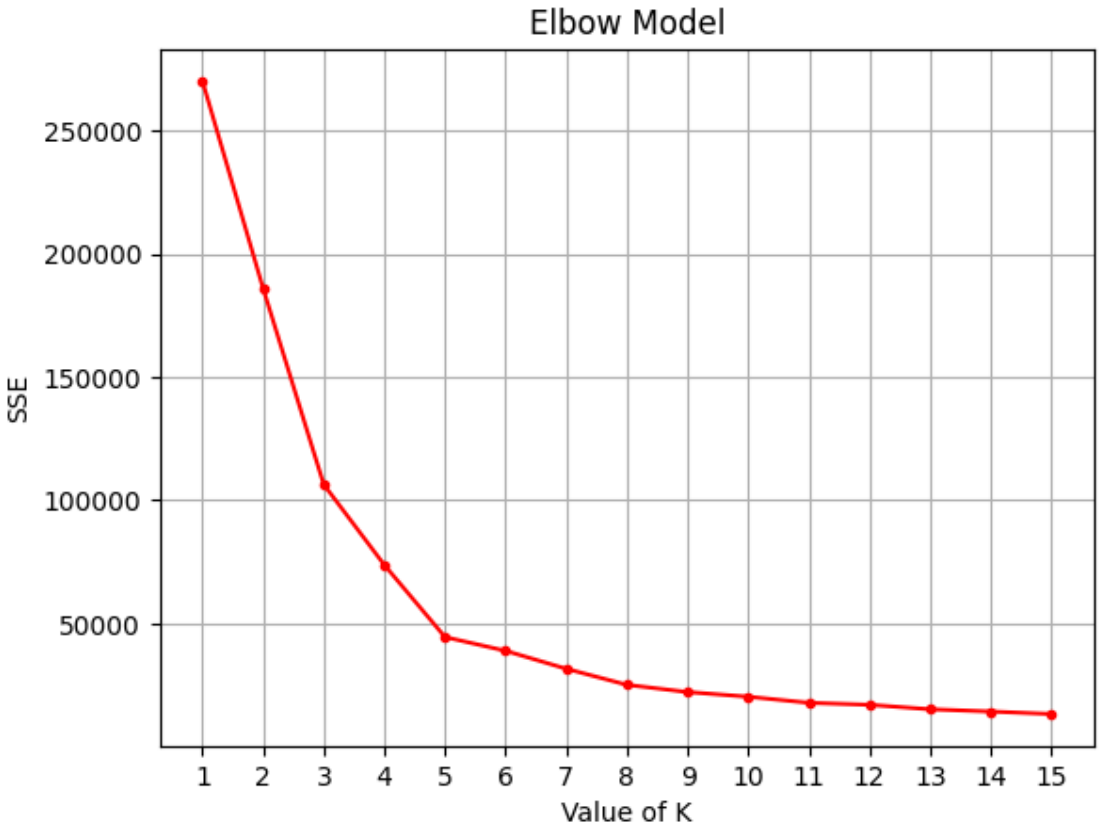
[22]:
```
sse=[]
for k in range(1,16):
    km = KMeans(n_clusters=k)
    km.fit_predict(x)
    sse.append(km.inertia_)
```

[23]:
```
sse
```

[23]:
```
[269981.28000000014,
 185917.1425392853,
 106348.37306211119,
 73679.78903948837,
 44448.45544793369,
 38858.959975143895,
 31577.72635585145,
 25061.304119069322,
 22143.222100767438,
 20248.174398469746,
 17775.977525252525,
 16968.805409786288,
 15156.2326923077,
 14240.611282073049,
 13234.875418100422]
```

[24]:
```
plt.title('Elbow Model')
plt.xlabel('Value of K')
plt.ylabel('SSE')
plt.grid()
```

```
plt.xticks(range(1,16))
plt.plot(range(1,16),sse,marker='.',color='red');
```



K at 5

[25]:
```
from sklearn.metrics import silhouette_score
```

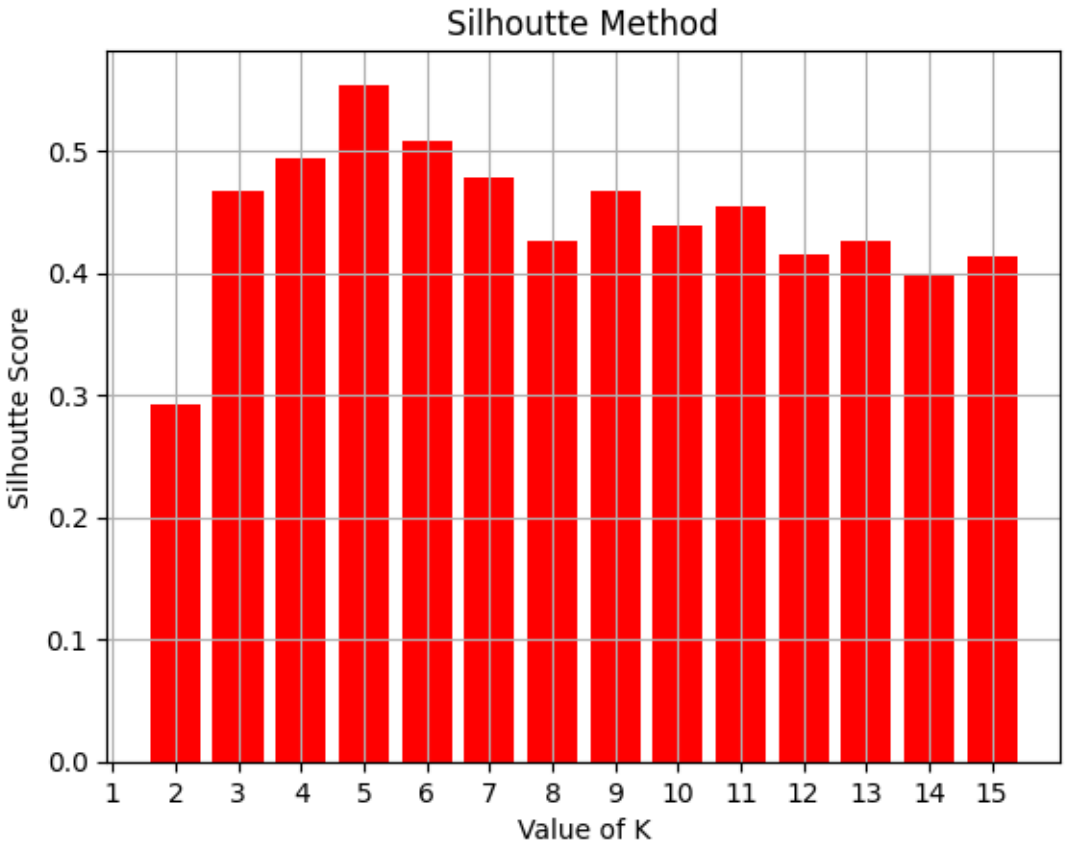
[26]:
```
silh=[]
for k in range(2,16):
    km = KMeans(n_clusters=k)
    labels = km.fit_predict(x)
    score = silhouette_score(x,labels)
    silh.append(score)
```

[27]:
```
silh
```

[27]:
```
[0.2918426367691145,
 0.46761358158775435,
 0.4931963109249047,
 0.553931997444648,
 0.5082526725498011,
```

```
0.47852679446095336,
0.42638821874961397,
0.4675793019403562,
0.43865010075435323,
0.45456539753534914,
0.4144394119208787,
0.4263243388723275,
0.39864355057622886,
0.413695146519944]
```

[28]:
```
plt.title('Silhoutte Method')
plt.xlabel('Value of K')
plt.ylabel('Silhoutte Score')
plt.grid()
plt.xticks(range(1,16))
plt.bar(range(2,16),silh,color='red');
```



Again 5

[31]:
```
km = KMeans(n_clusters=5,random_state=0)
labels = km.fit_predict(x)
```

[32]:
```
labels
```

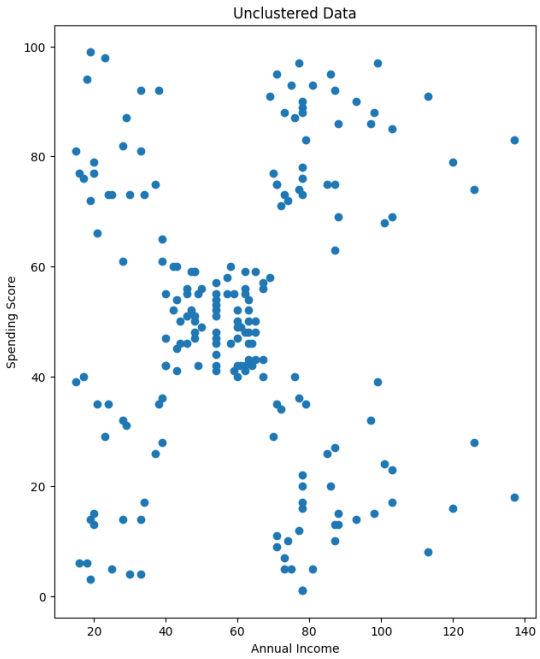
[32]:
```
array([3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4,
       3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 4, 3, 0,
       3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 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, 0, 1, 2, 1, 2, 1,
       0, 1, 2, 1, 2, 1, 2, 1, 2, 1, 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], dtype=int32)
```

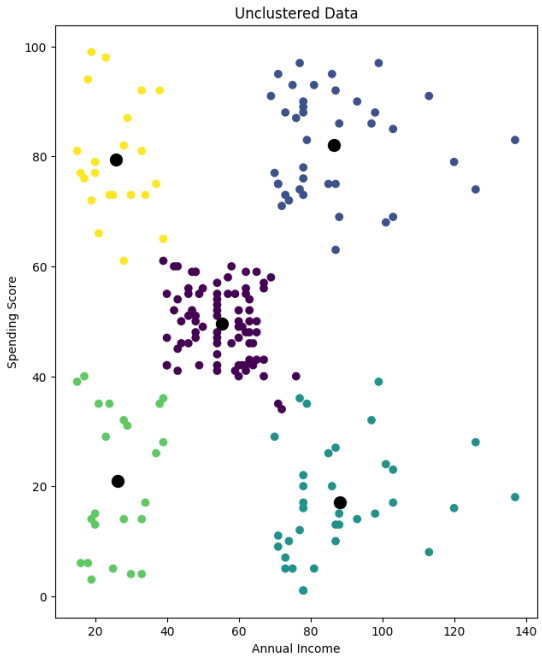
[33]:
```
cent = km.cluster_centers_
cent
```

[33]:
```
array([[55.2962963 , 49.51851852],
       [86.53846154, 82.12820513],
       [88.2       , 17.11428571],
       [26.30434783, 20.91304348],
       [25.72727273, 79.36363636]])
```

[34]:
```
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('Unclustered 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=100,color='k');
```





[35]:
```
km.inertia_
```

[35]:
```
44448.45544793369
```

[38]:
```
km.labels_
```

[38]:
```
array([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, 3, 0,
       3, 4, 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, 0, 1, 2, 1, 2, 1,
       0, 1, 2, 1, 2, 1, 2, 1, 2, 1, 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], dtype=int32)
```

[39]:
```
df[labels==4]
```

[39]:

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	2	Male	21	15	81
3	4	Female	23	16	77
5	6	Female	22	17	76
7	8	Female	23	18	94
9	10	Female	30	19	72
11	12	Female	35	19	99

```
13          14  Female   24                  20                  77
15          16    Male   22                  20                  79
17          18    Male   20                  21                  66
19          20  Female   35                  23                  98
21          22    Male   25                  24                  73
23          24    Male   31                  25                  73
25          26    Male   29                  28                  82
27          28    Male   35                  28                  61
29          30  Female   23                  29                  87
31          32  Female   21                  30                  73
33          34    Male   18                  33                  92
35          36  Female   21                  33                  81
37          38  Female   30                  34                  73
39          40  Female   20                  37                  75
41          42    Male   24                  38                  92
45          46  Female   24                  39                  65
```

[40]:
```
four = df[labels==4]
```

[41]:
```
four.to_csv('demo.csv')
```

[44]:
```
km.predict([[56,61]])
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/base.py:493: UserWarning: X does
not have valid feature names, but KMeans was fitted with feature names
  warnings.warn(
```

[44]:
```
array([0], dtype=int32)
```

[45]:
```
agl = AgglomerativeClustering(n_clusters=5)
```

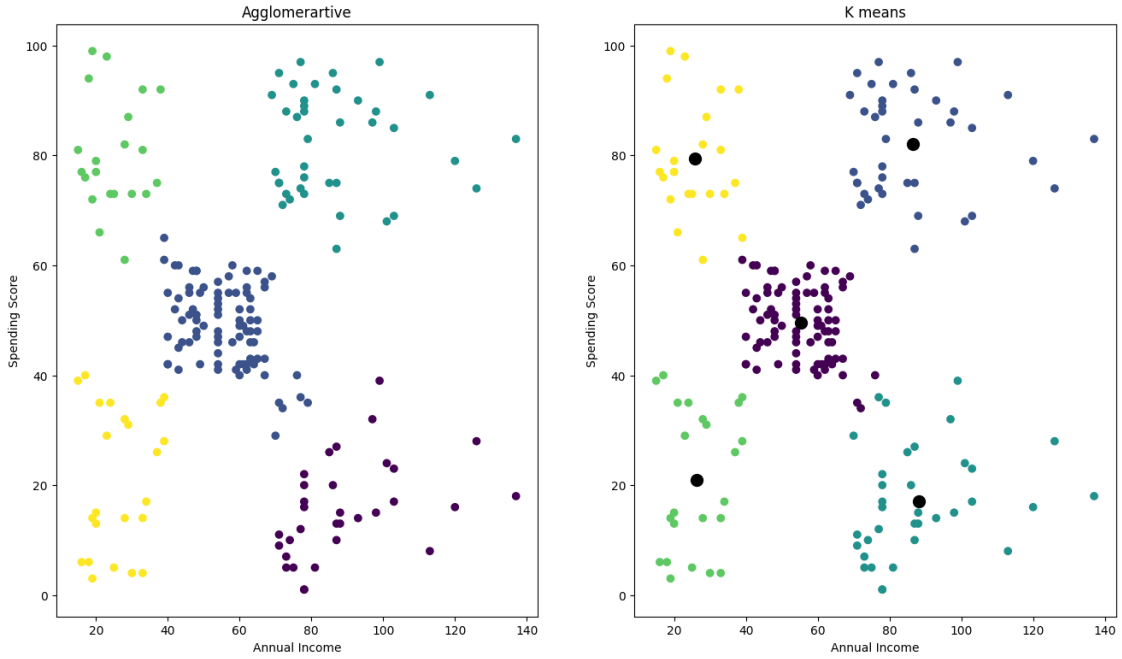
[46]:
```
alabels=agl.fit_predict(x)
alabels
```

[46]:
```
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, 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])
```

[48]:
```
plt.figure(figsize=(16,9))
plt.subplot(1,2,1)
```

```
plt.title('Agglomerartive')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.scatter(x['Annual Income (k$)'],x['Spending Score (1-100)'],c=alabels);

plt.subplot(1,2,2)
plt.title('K means')
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');
```



[ ]: