Prekshita Vasudeo patil
20MAI0073
assignment - clustering

In [1]:
```
import pandas as pd
from warnings import filterwarnings
filterwarnings("ignore")
read= pd.read_csv("Salary_Data.csv")
read.head()
```

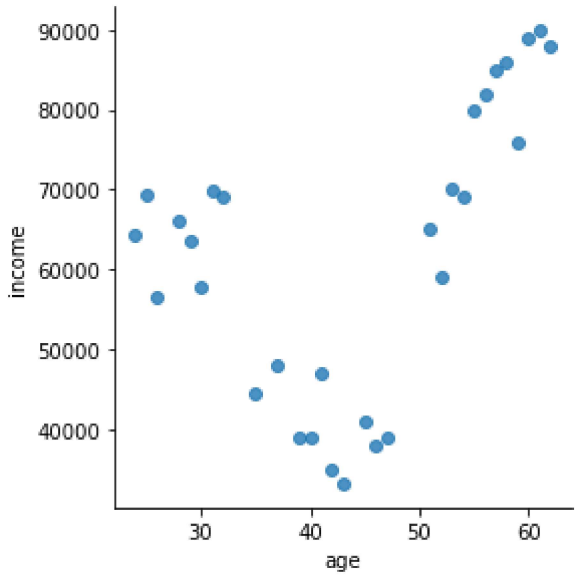
Out[1]:

	age	income
0	52	59000
1	53	70000
2	57	85000
3	29	63525
4	31	69891

In [2]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [3]:
```
sns.lmplot('age','income',data=read,fit_reg=False,size=4)
```

Out[3]:
```
<seaborn.axisgrid.FacetGrid at 0x176a81ff8b0>
```



In [4]:
```
from sklearn.cluster import KMeans
```

In [5]:
```
kmeans = KMeans(3)
kmeans.fit(read)
```

Out[5]:
```
KMeans(n_clusters=3)
```

In [6]:
```
# output variable is clusters.label_ (will have id)
read['Cluster_id'] = kmeans.labels_
```

In [7]:
```
read.head()
```

Out[7]:

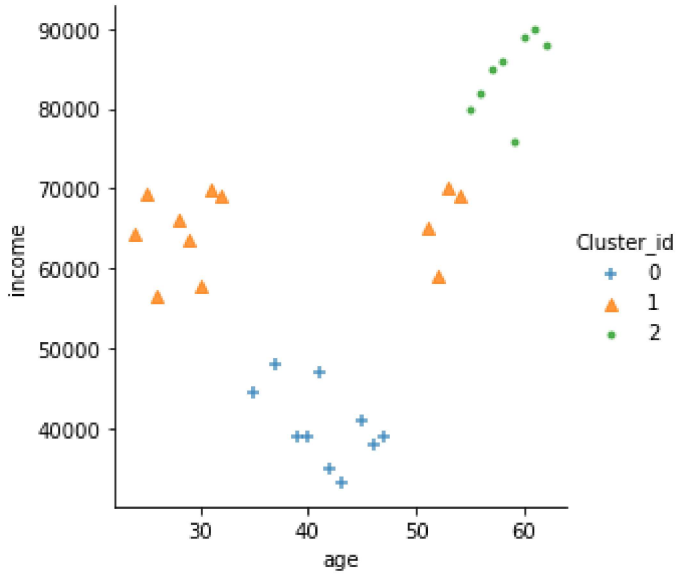
	age	income	Cluster_id
0	52	59000	1
1	53	70000	1
2	57	85000	2
3	29	63525	1
4	31	69891	1

In [8]:
```
read
```

Out[8]:

	age	income	Cluster_id
0	52	59000	1
1	53	70000	1
2	57	85000	2
3	29	63525	1
4	31	69891	1
5	26	56642	1
6	32	69150	1
7	24	64445	1
8	35	44555	0
9	37	47900	0
10	54	69000	1
11	55	80000	2
12	56	82000	2
13	39	38900	0
14	25	69343	1
15	28	66205	1
16	30	57731	1
17	40	39000	0
18	41	46899	0
19	42	35000	0
20	45	41111	0
21	59	76000	2
22	60	89000	2
23	61	90000	2
24	46	38000	0
25	47	39000	0
26	43	33088	0
27	51	65000	1
28	58	86000	2
29	62	88000	2

In [9]:
```
markers = ['+','^','.']
sns.lmplot("age","income",data=read,hue="Cluster_id",fit_reg=False,size=4,markers
plt.show()
```



In [10]:
```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
scaled_read = ss.fit_transform(read[["age","income"]])
scaled_read
```

Out[10]:
```
array([[ 0.67442184, -0.16756351],
       [ 0.75802786,  0.45106331],
       [ 1.09245191,  1.29464532],
       [-1.24851647,  0.08691707],
       [-1.08130444,  0.44493328],
       [-1.49933451, -0.3001746 ],
       [-0.99769843,  0.40326032],
       [-1.66654654,  0.13865676],
       [-0.74688039, -0.97993299],
       [-0.57966836, -0.7918142 ],
       [ 0.84163387,  0.3948245 ],
       [ 0.92523988,  1.01345132],
       [ 1.0088459 ,  1.12592892],
       [-0.41245633, -1.29796341],
       [-1.58294052,  0.41411441],
       [-1.33212248,  0.23763705],
       [-1.16491046, -0.23893055],
       [-0.32885032, -1.29233953],
       [-0.24524431, -0.84810924],
       [-0.16163829, -1.51729474],
       [ 0.08917975, -1.17361942],
       [ 1.25966394,  0.78849611],
       [ 1.34326995,  1.51960053],
       [ 1.42687597,  1.57583933],
       [ 0.17278576, -1.34857834],
       [ 0.25639178, -1.29233953],
       [-0.07803228, -1.62482333],
       [ 0.59081583,  0.1698693 ],
       [ 1.17605793,  1.35088413],
       [ 1.51048198,  1.46336173]])
```

In [11]:
```
new_kmeans = KMeans(3)
new_kmeans.fit(scaled_read)
read["Scaled_Clusters"] = new_kmeans.labels_
```

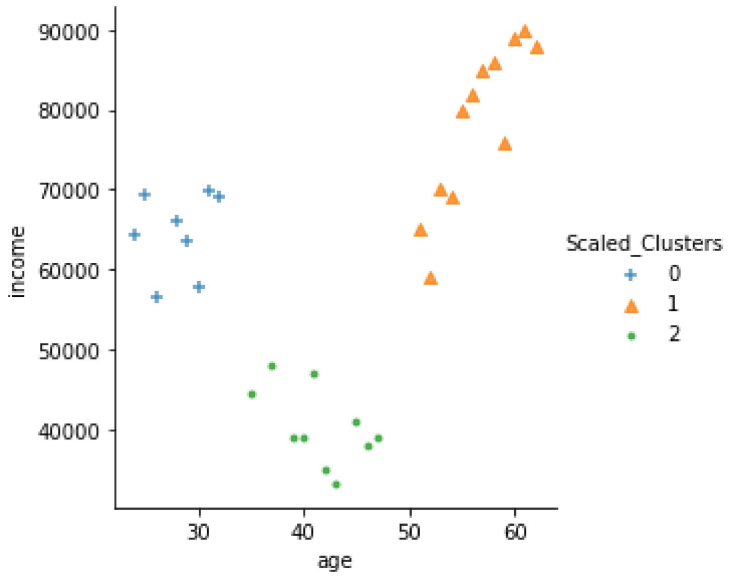
In [12]:
```
read
```

Out[12]:

	age	income	Cluster_id	Scaled_Clusters
0	52	59000	1	1
1	53	70000	1	1
2	57	85000	2	1
3	29	63525	1	0
4	31	69891	1	0
5	26	56642	1	0
6	32	69150	1	0
7	24	64445	1	0
8	35	44555	0	2
9	37	47900	0	2
10	54	69000	1	1
11	55	80000	2	1
12	56	82000	2	1
13	39	38900	0	2
14	25	69343	1	0
15	28	66205	1	0
16	30	57731	1	0
17	40	39000	0	2
18	41	46899	0	2
19	42	35000	0	2
20	45	41111	0	2
21	59	76000	2	1
22	60	89000	2	1
23	61	90000	2	1
24	46	38000	0	2
25	47	39000	0	2
26	43	33088	0	2
27	51	65000	1	1
28	58	86000	2	1
29	62	88000	2	1

In [13]:
```
plt.figure(figsize=(16,10))
markers = ['+','^','.']
sns.lmplot("age","income",data=read,hue="Scaled_Clusters",fit_reg=False,size=4,ma
plt.show()
```

```
<Figure size 1152x720 with 0 Axes>
```



In [14]:
```
from sklearn.cluster import AgglomerativeClustering
```

PROBLEM STATEMENT: USE bev.csv - Apply Clustering (KMeans and Agglomerative) - Compare the clusters created by both the techniques

In [15]:
```
coffee = pd.read_excel("bev.xlsx",sheet_name="Sheet1")
```

In [16]:
```
coffee.head()
```

Out[16]:

	Name	Potassium	Sodium	Caffeine	Cost
0	new_england_coffee	144	15	4.7	0.43
1	post_alley_blend	151	19	4.9	0.43
2	stumpdown_coffee	157	15	0.9	0.48
3	bizzy_organic_coffee	170	7	5.2	0.73
4	indian_bean	152	11	5.0	0.77

In [17]:
```
scaled_coffee = ss.fit_transform(coffee[["Potassium",'Sodium',"Caffeine","Cost"]]
scaled_coffee
```

Out[17]:
```
array([[ 0.38791334,  0.00779468,  0.43380786, -0.45682969],
       [ 0.6250656 ,  0.63136906,  0.62241997, -0.45682969],
       [ 0.82833896,  0.00779468, -3.14982226, -0.10269815],
       [ 1.26876459, -1.23935408,  0.90533814,  1.66795955],
       [ 0.65894449, -0.6157797 ,  0.71672602,  1.95126478],
       [ 0.42179223,  1.25494344,  0.3395018 , -1.5192243 ],
       [ 1.43815906,  1.41083704,  1.1882563 , -0.66930861],
       [ 0.55730781,  1.87851782,  0.43380786, -0.52765599],
       [-1.1366369 , -0.7716733 ,  0.05658363, -0.45682969],
       [-0.66233238, -1.08346049, -0.5092527 , -0.66930861],
       [ 0.25239776,  0.47547547,  0.3395018 , -0.38600338],
       [-1.03500022,  0.00779468, -0.13202848, -0.24435076],
       [ 0.08300329, -0.6157797 , -0.03772242,  0.03895447],
       [ 0.59118671,  0.63136906,  0.43380786,  1.88043848],
       [ 0.55730781, -1.39524768,  0.71672602,  2.0929174 ],
       [-2.18688263,  0.00779468, -1.82953748, -0.81096123],
       [ 0.21851887,  0.63136906,  0.15088969, -0.45682969],
       [ 0.38791334,  1.41083704,  0.62241997, -0.45682969],
       [-2.05136705, -1.39524768, -1.26370115, -0.24435076],
       [-1.20439469, -1.23935408, -0.03772242, -0.17352445]])
```

In [18]:
```
km = KMeans(n_clusters=3)
km.fit(scaled_coffee)
coffee["clusterid"] = km.labels_
```

In [19]:
```
coffee
```

Out[19]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid
0	new_england_coffee	144	15	4.7	0.43	1
1	post_alley_blend	151	19	4.9	0.43	1
2	stumpdown_coffee	157	15	0.9	0.48	0
3	bizzy_organic_coffee	170	7	5.2	0.73	2
4	indian_bean	152	11	5.0	0.77	2
5	jacobs_coffee	145	23	4.6	0.28	1
6	grounds_hounds_coffee	175	24	5.5	0.40	1
7	la_columbe_corisca	149	27	4.7	0.42	1
8	lavazza_super_crema	99	10	4.3	0.43	0
9	mount_hagen	113	8	3.7	0.40	0
10	red_bay_coffee	140	18	4.6	0.44	1
11	peerless_wholebean	102	15	4.1	0.46	0
12	stone_street_coffee	135	11	4.2	0.50	0
13	green_mountain_coffee	150	19	4.7	0.76	2
14	koffee_cuit	149	6	5.0	0.79	2
15	caribou_coffee	68	15	2.3	0.38	0
16	irish_hazelnut_coffee	139	19	4.4	0.43	1
17	cremoso_coffee	144	24	4.9	0.43	1
18	davidoff_coffee	72	6	2.9	0.46	0
19	js_coffee	97	7	4.2	0.47	0

In [20]:
```
coffee[coffee["clusterid"]==2]
```

Out[20]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid
3	bizzy_organic_coffee	170	7	5.2	0.73	2
4	indian_bean	152	11	5.0	0.77	2
13	green_mountain_coffee	150	19	4.7	0.76	2
14	koffee_cuit	149	6	5.0	0.79	2

In [21]:
```
# plt.figure(figsize=(16,10))
# markers = ['+','^','.']
# sns.lmplot("Potassium","Sodium",data=coffee,hue="clusterid",fit_reg=False,size=
# plt.grid("darkgrid",color="pink")
# plt.show()
```

In [22]:
```
# plt.figure(figsize=(10,10))
# markers = ['+','^','.']
# sns.lmplot("Potassium","Cost",data=coffee,hue="clusterid",fit_reg=False,size=4,
# plt.show()
```

In [23]:
```
from sklearn.cluster import AgglomerativeClustering
agl = AgglomerativeClustering(n_clusters=3)
agl.fit(scaled_coffee)
coffee["Agl Cluster_id"] = agl.labels_
```

In [24]:
```
coffee
```

Out[24]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid	Agl Cluster_id
0	new_england_coffee	144	15	4.7	0.43	1	1
1	post_alley_blend	151	19	4.9	0.43	1	1
2	stumpdown_coffee	157	15	0.9	0.48	0	0
3	bizzy_organic_coffee	170	7	5.2	0.73	2	2
4	indian_bean	152	11	5.0	0.77	2	2
5	jacobs_coffee	145	23	4.6	0.28	1	1
6	grounds_hounds_coffee	175	24	5.5	0.40	1	1
7	la_columbe_corisca	149	27	4.7	0.42	1	1
8	lavazza_super_crema	99	10	4.3	0.43	0	0
9	mount_hagen	113	8	3.7	0.40	0	0
10	red_bay_coffee	140	18	4.6	0.44	1	1
11	peerless_wholebean	102	15	4.1	0.46	0	0
12	stone_street_coffee	135	11	4.2	0.50	0	0
13	green_mountain_coffee	150	19	4.7	0.76	2	2
14	koffee_cuit	149	6	5.0	0.79	2	2
15	caribou_coffee	68	15	2.3	0.38	0	0
16	irish_hazelnut_coffee	139	19	4.4	0.43	1	1
17	cremoso_coffee	144	24	4.9	0.43	1	1
18	davidoff_coffee	72	6	2.9	0.46	0	0
19	js_coffee	97	7	4.2	0.47	0	0

In [25]:
```
coffee.shape
```

Out[25]:
```
(20, 7)
```

In [26]:
```
coffee[coffee["clusterid"] !=coffee["Agl Cluster_id"] ]
```

Out[26]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid	Agl Cluster_id

In [27]:
```
print("Difference in clusters is",coffee[coffee["clusterid"] !=coffee["Agl Cluste
```

```
Difference in clusters is 0 rows
```

In [28]:
```
coffee[coffee["clusterid"] ==coffee["Agl Cluster_id"] ]
```

Out[28]:

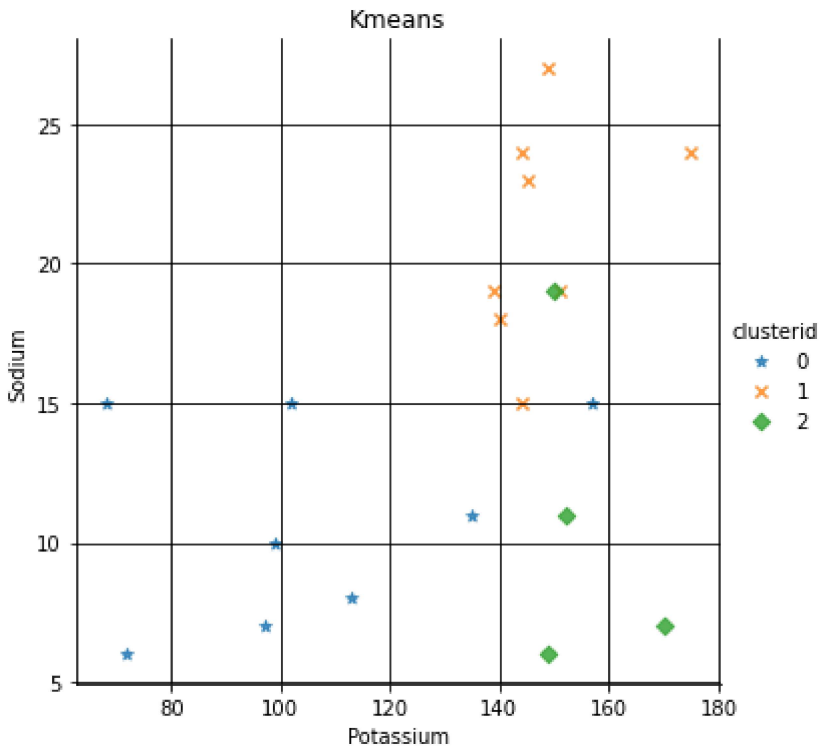
	Name	Potassium	Sodium	Caffeine	Cost	clusterid	Agl Cluster_id
0	new_england_coffee	144	15	4.7	0.43	1	1
1	post_alley_blend	151	19	4.9	0.43	1	1
2	stumpdown_coffee	157	15	0.9	0.48	0	0
3	bizzy_organic_coffee	170	7	5.2	0.73	2	2
4	indian_bean	152	11	5.0	0.77	2	2
5	jacobs_coffee	145	23	4.6	0.28	1	1
6	grounds_hounds_coffee	175	24	5.5	0.40	1	1
7	la_columbe_corisca	149	27	4.7	0.42	1	1
8	lavazza_super_crema	99	10	4.3	0.43	0	0
9	mount_hagen	113	8	3.7	0.40	0	0
10	red_bay_coffee	140	18	4.6	0.44	1	1
11	peerless_wholebean	102	15	4.1	0.46	0	0
12	stone_street_coffee	135	11	4.2	0.50	0	0
13	green_mountain_coffee	150	19	4.7	0.76	2	2
14	koffee_cuit	149	6	5.0	0.79	2	2
15	caribou_coffee	68	15	2.3	0.38	0	0
16	irish_hazelnut_coffee	139	19	4.4	0.43	1	1
17	cremoso_coffee	144	24	4.9	0.43	1	1
18	davidoff_coffee	72	6	2.9	0.46	0	0
19	js_coffee	97	7	4.2	0.47	0	0

In [29]:
```
print("Same clusters rows are :- ",coffee[coffee["clusterid"] ==coffee["Agl Clust
```

```
Same clusters rows are :-  20
```

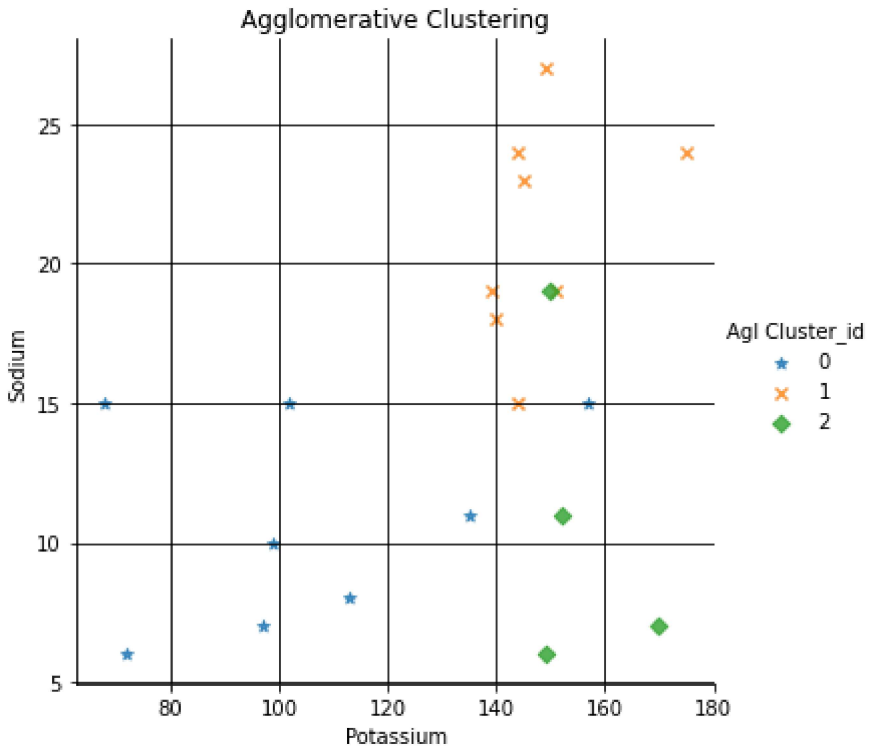
In [30]:
```
plt.figure(figsize=(16,10))
sns.lmplot("Potassium",'Sodium',hue="clusterid",data=coffee,markers=['*','x','D']
plt.title("Kmeans")
plt.grid("whitegrid",color="black")
plt.show()
```

```
<Figure size 1152x720 with 0 Axes>
```



In [31]:
```
plt.figure(figsize=(16,10))
sns.lmplot("Potassium",'Sodium',hue="Agl Cluster_id",data=coffee,markers=['*','x'
plt.title('Agglomerative Clustering')
plt.grid("whitegrid",color="black")
plt.show()
```

```
<Figure size 1152x720 with 0 Axes>
```



Conclusion

- There seems to be no difference in the predictions of clusters both clusters results the same cluster number.