1. Partition Based Metjods - KMeans
2. Heirarchical - Agglomerative

KMeans

Step 1: It randomly selects 'k' data objects from the dataseteach of which represents a Cluster Center

Step 2: Repeat for each of the remaining data objects, an object is assigned to a cluster to which it is most similar i.e minimum distance (based on the distance between the object and cluster center)

Step 3: It then computes a new mean for each respective cluster until there is no change

In [1]:

```
#PROBLEM STATEMENT
# Use the bev.csv dataset and apply KMeans and Agglomerative clustering; Compare the cluster
```

In [2]:

```
# Importing the dataset
import numpy as np
import seaborn as sn
import matplotlib as plt
import pandas as pd
%matplotlib inline
bev_df = pd.read_csv("bev.csv")
bev_df.head()
```

Out[2]:

	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 [3]:

```
# we have already imported the libraries :-)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_bev_df = scaler.fit_transform(bev_df[["Potassium" , "Sodium" , "Caffeine" , "Cost"]])
scaled_bev_df[0:5]
```

Out[3]:

```
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]])
```

In [4]:

```
from sklearn.cluster import KMeans
#KMeans 3 -> choose 3 random centers
clusters = KMeans(3)
clusters.fit(scaled_bev_df)
bev_df["clusterid"] = clusters.labels_
```

In [5]:

```
# TO LOOK AT THE CLUSTERS
bev_df[bev_df.clusterid == 0]
```

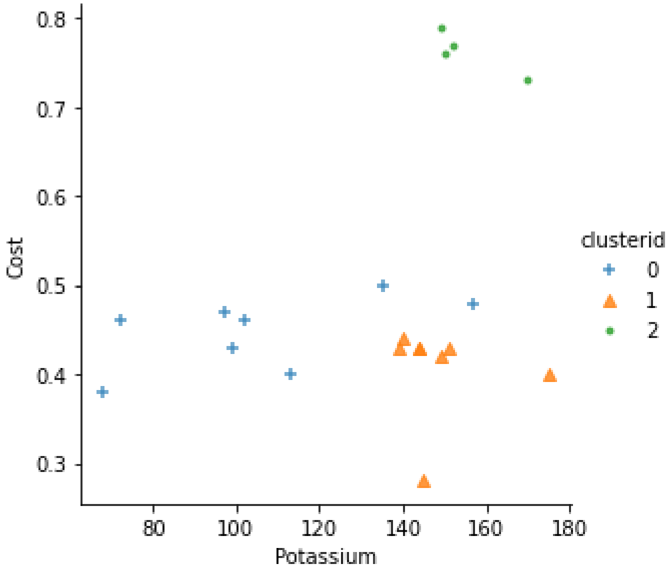
Out[5]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid
2	stumpdown_coffee	157	15	0.9	0.48	0
8	lavazza_super_crema	99	10	4.3	0.43	0
9	mount_hagen	113	8	3.7	0.40	0
11	peerless_wholebean	102	15	4.1	0.46	0
12	stone_street_coffee	135	11	4.2	0.50	0
15	caribou_coffee	68	15	2.3	0.38	0
18	davidoff_coffee	72	6	2.9	0.46	0
19	js_coffee	97	7	4.2	0.47	0

In [6]:
```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Potassium" , "Cost" , data = bev_df , hue = "clusterid" , fit_reg = False , markers = marker , size =
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[6]:
```
<seaborn.axisgrid.FacetGrid at 0x25946308550>
```



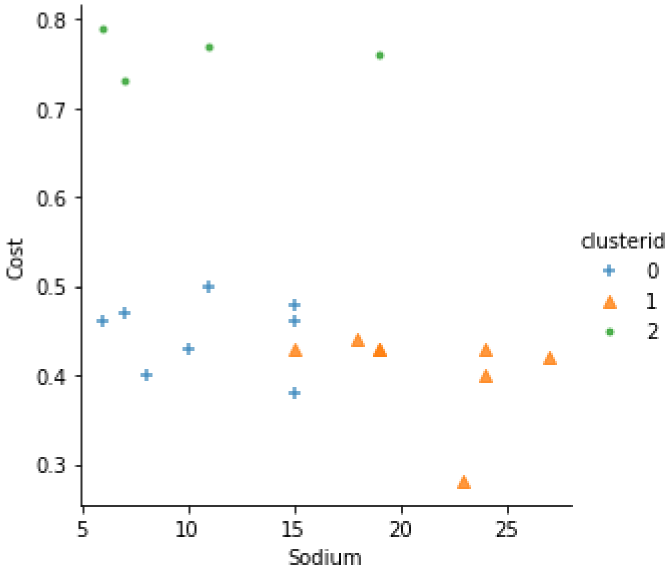
In [7]:

```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Sodium" , "Cost" , data = bev_df , hue = "clusterid" , fit_reg = False , markers = marker , size = 4)
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[7]:

```
<seaborn.axisgrid.FacetGrid at 0x259469c0fd0>
```



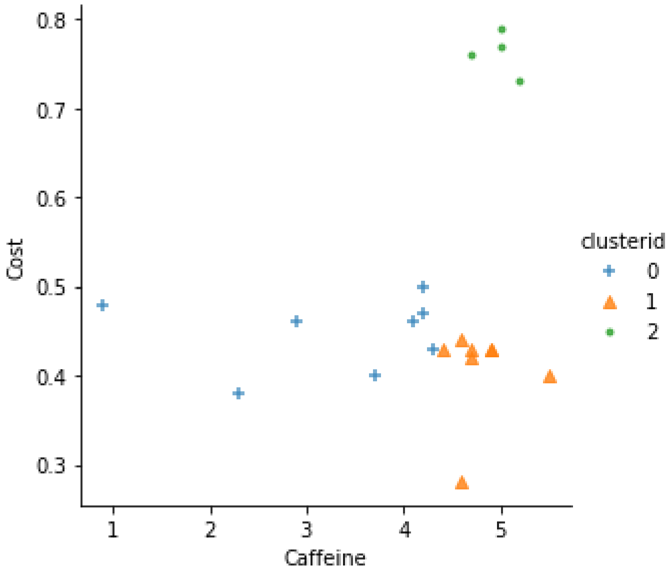
In [8]:

```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Caffeine" , "Cost" , data = bev_df , hue = "clusterid" , fit_reg = False , markers = marker , size =
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[8]:

```
<seaborn.axisgrid.FacetGrid at 0x259458c3730>
```



AGGLOMERATIVE CLUSTERING STEPS

1.Each data point is assigned as a single cluster.

2.Determine the distance measurement and calculate the distance matrix.

3.Determine the linkage criteria to merge the clusters.

4.Update the distance matrix.

5.Repeat the process until every data point become one cluster.

In [9]:

```
# AGGLOMERATIVE -> bottom up approach , initially every dataitem is considered as clusters and
# then they are merged together until their is no change
from sklearn.cluster import AgglomerativeClustering
clustering = AgglomerativeClustering(n_clusters = 3)
clustering.fit(scaled_bev_df)
bev_df["ClusteringId"]=clustering.labels_
```

In [10]:

```
# TO LOOK AT THE CLUSTERS
bev_df[bev_df.ClusteringId == 0]
```

Out[10]:

	Name	Potassium	Sodium	Caffeine	Cost	clusterid	ClusteringId
2	stumpdown_coffee	157	15	0.9	0.48	0	0
8	lavazza_super_crema	99	10	4.3	0.43	0	0
9	mount_hagen	113	8	3.7	0.40	0	0
11	peerless_wholebean	102	15	4.1	0.46	0	0
12	stone_street_coffee	135	11	4.2	0.50	0	0
15	caribou_coffee	68	15	2.3	0.38	0	0
18	davidoff_coffee	72	6	2.9	0.46	0	0
19	js_coffee	97	7	4.2	0.47	0	0

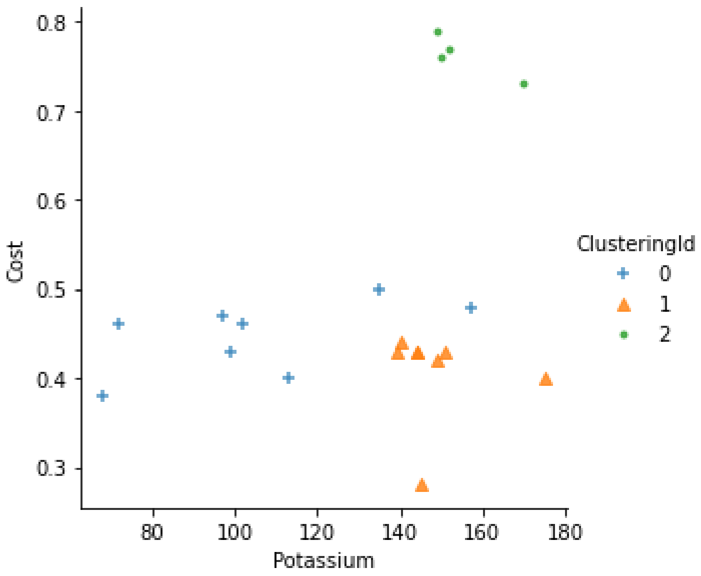
In [11]:

```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Potassium" , "Cost" , data = bev_df , hue = "ClusteringId" , fit_reg = False , markers = marker , siz
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[11]:

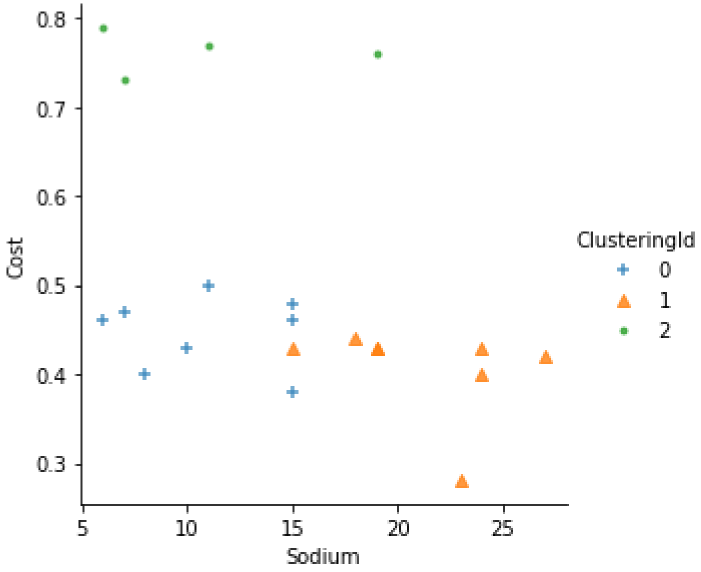
```
<seaborn.axisgrid.FacetGrid at 0x25946c05dc0>
```



In [12]:
```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Sodium" , "Cost" , data = bev_df , hue = "ClusteringId" , fit_reg = False , markers = marker , size =
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

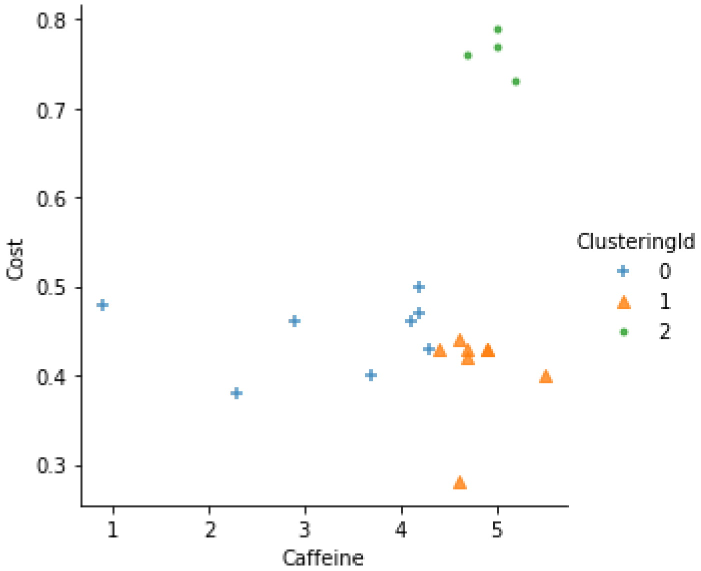
Out[12]:
```
<seaborn.axisgrid.FacetGrid at 0x25946c121c0>
```



In [13]:
```
# PLOTTING AGAIN
marker = ['+' , '^' , '.']
sn.lmplot("Caffeine" , "Cost" , data = bev_df , hue = "ClusteringId" , fit_reg = False , markers = marker , size
```

```
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following va
riables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passi
ng other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\Users\Siddharth\anaconda3\lib\site-packages\seaborn\regression.py:580: UserWarning: The `size` parameter ha
s been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

Out[13]:
```
<seaborn.axisgrid.FacetGrid at 0x25946cbf5e0>
```



In [ ]: