Clustering & PCA Assignment

Submitted By:

Yogita Goswami

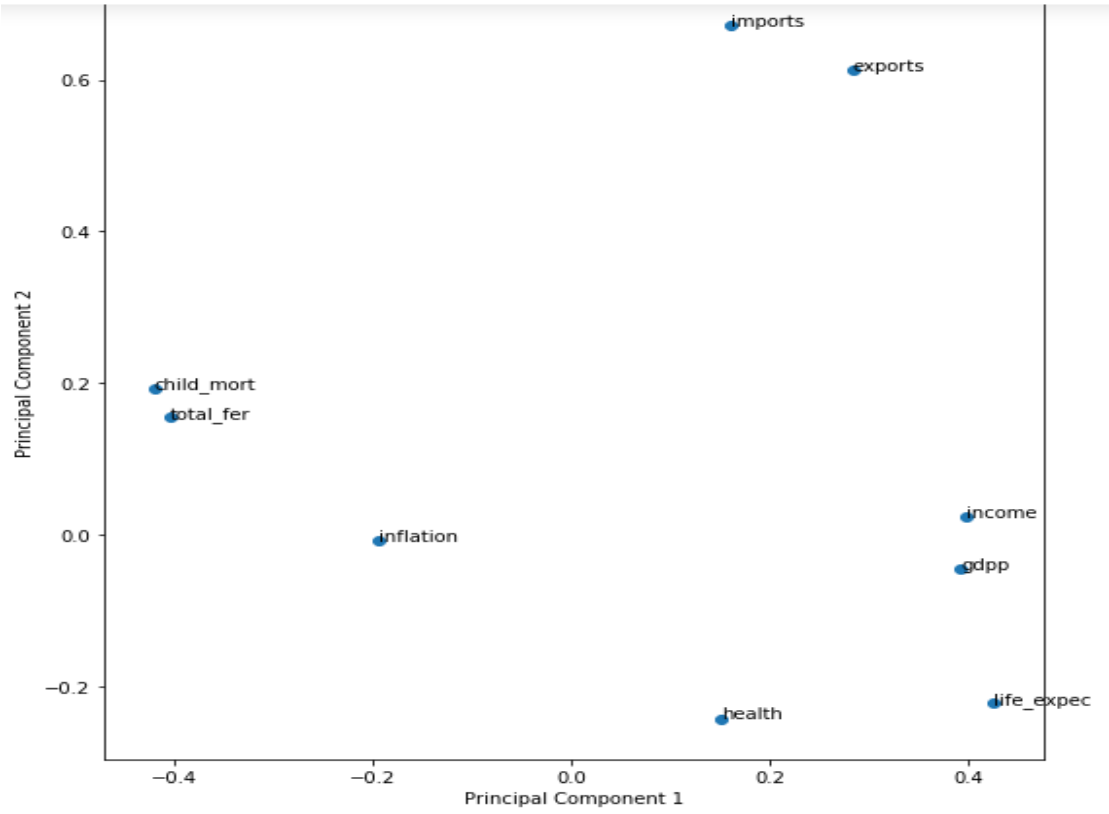
Principal Components Analysis

- Data Frame size is (167,10)
- First we will do the scaling to see whether we can explain the dataset using fewer variables.
- Then performing the PCA.
- The highest variance ratio of components is 0.4595.

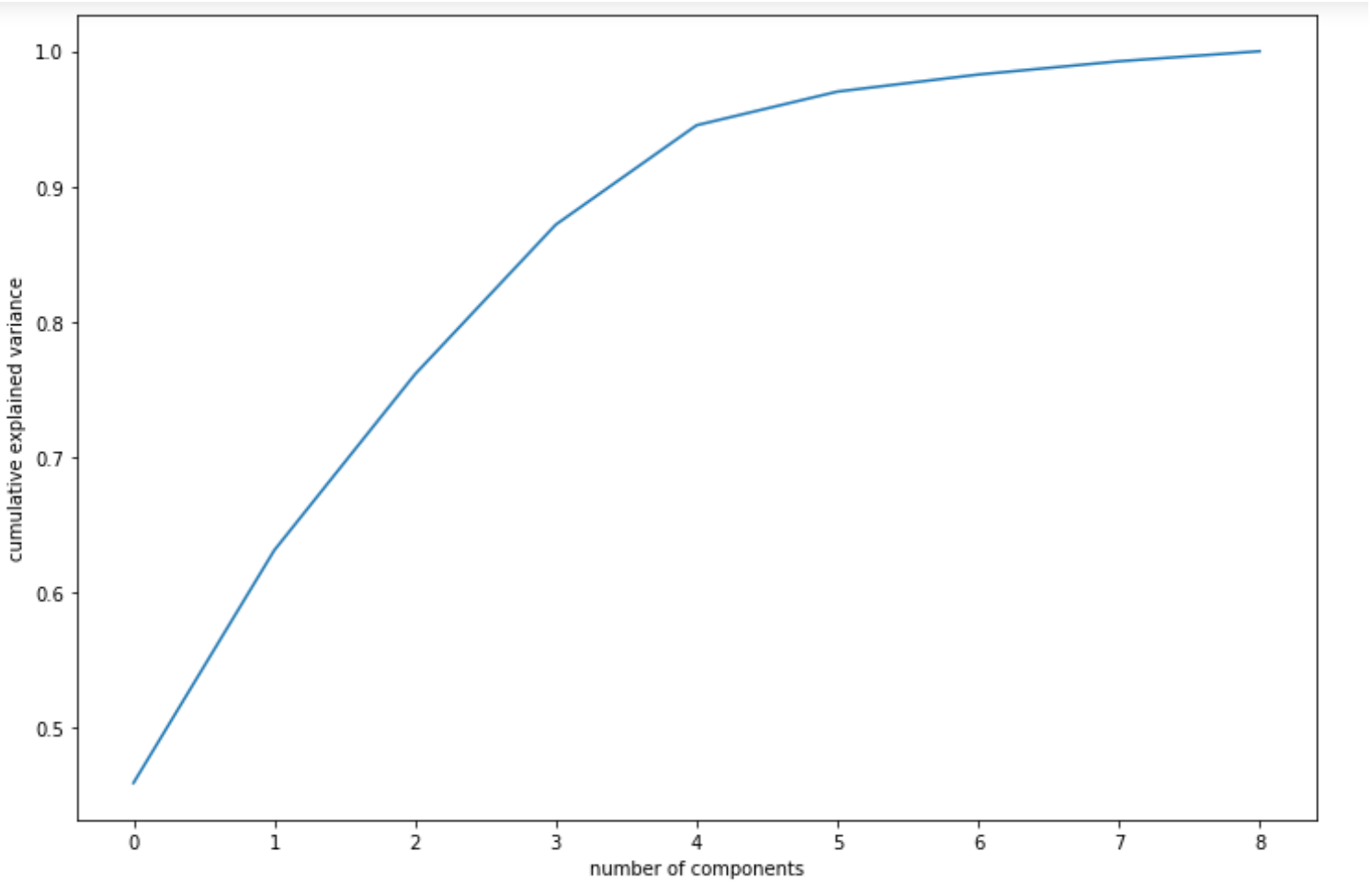
```
[0.4595174 , 0.17181626, 0.13004259, 0.11053162, 0.07340211,
 0.02484235, 0.0126043 , 0.00981282, 0.00743056])
```

- Then understanding how the original variables are loaded on the principal components.

Visualization of how these codes ae loaded



Scree Plot:

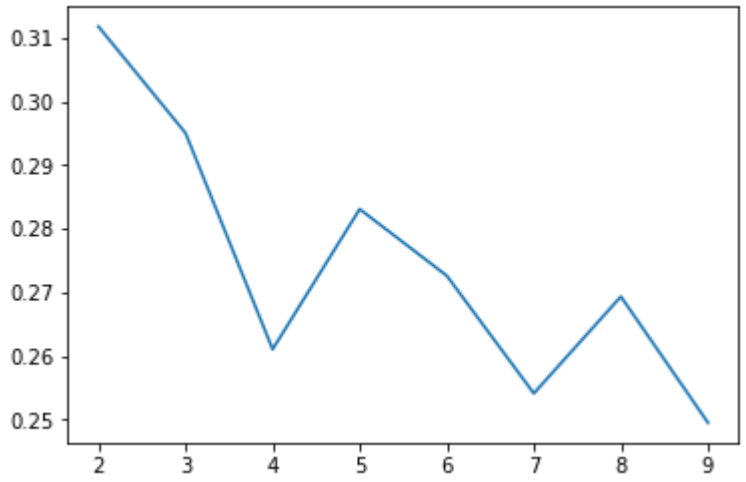


From the scree plot we got to know that around 96% of the information is being explained by 5 components.

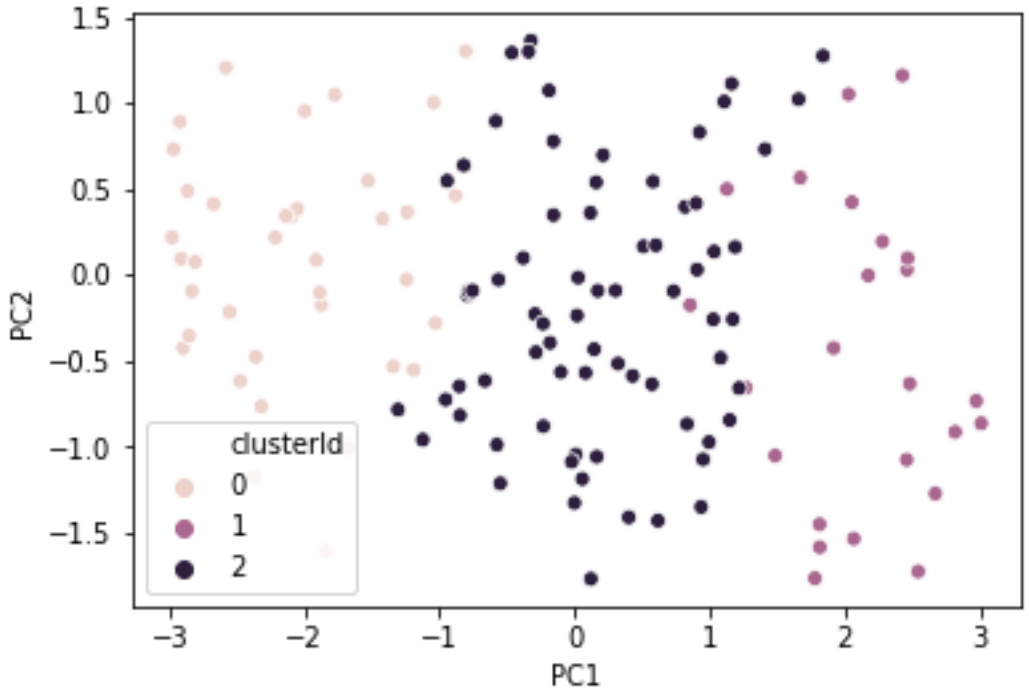
- We used IncrementalPCA to perform dimensionality reduction using the principal components.
- After checking the correlation we see that the correlation is near to 0.
- Performed the outliers for component1 and component2.
- After performing the outlier the data is of size (133,5)
- visualising the points in the pcs.
- one of the prime advantage of PCA is that you can visualise high dimensional data

Kmeans CLUSTERING

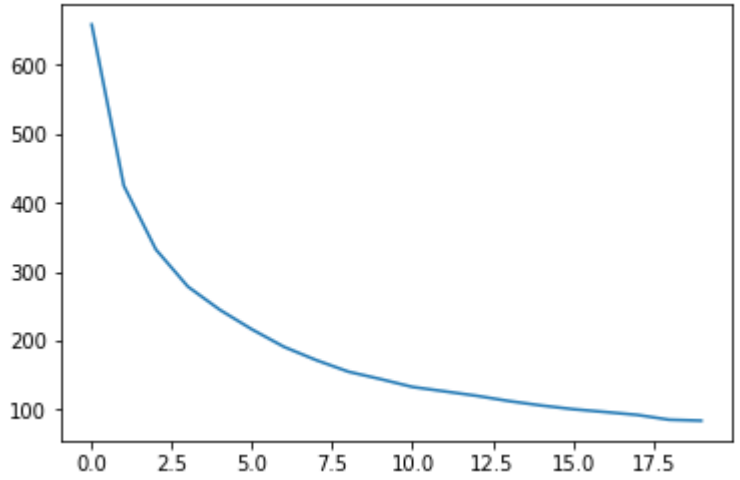
- Applying Hopkins algorithm we got the Hopkins measure of .7630.
- Hopkins measure will vary every time.
- since the data has value >.5 the given dataset has a good tendency to form clusters.
- first we will do the silhoutte score analysis to check the clusters.



Scatter plot for 3 components in Kmeans Clustering

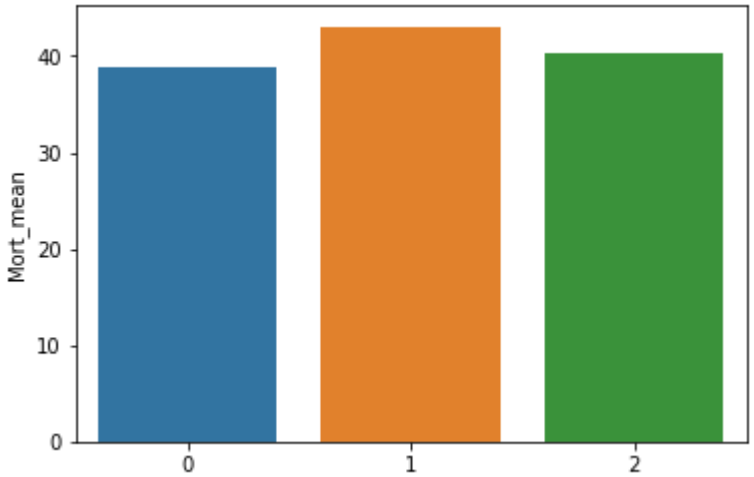


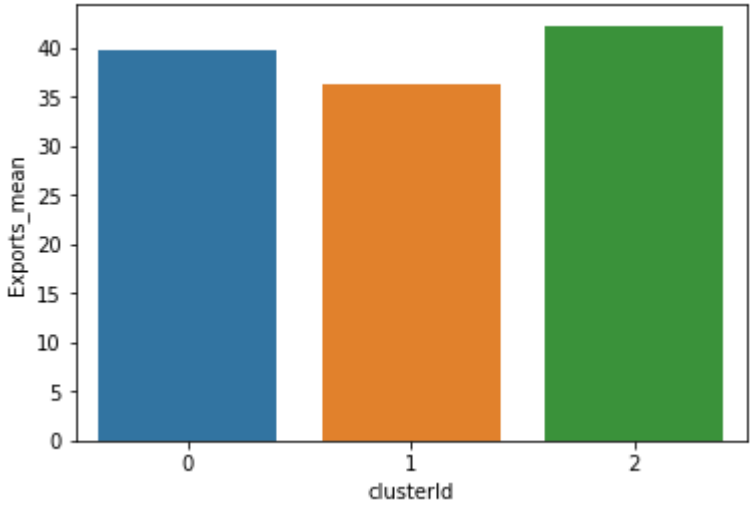
- Checking sum of squared distances.
- From elbow curve we analyzed that clusters required are 3.

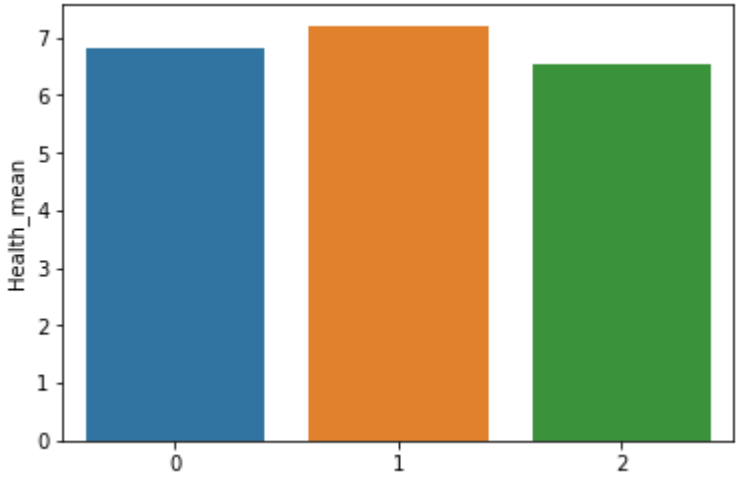


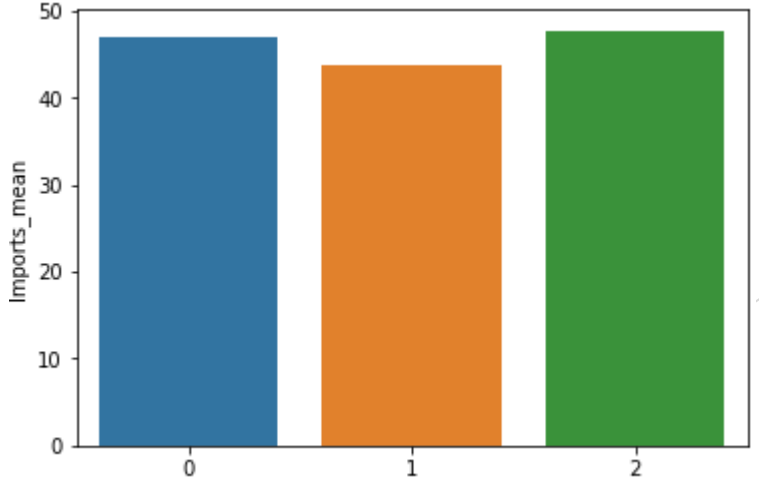
- Then we will merge the original dataset with the Kmeans dataset to check the mean.
- Then we will plot Barplot of cluster Id with each variable.
- From mean we got to know that inflation and child_mort variables are changing with huge difference so will take the inflation values more than or equal to 10.669 and child_mort values more than and equal to 43.07.
- From the analysis we got to know that poor countries belongs to cluster1.

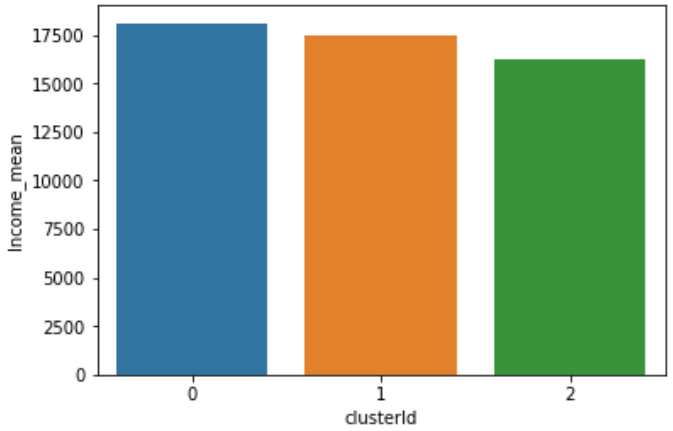
Visualization using Barplots

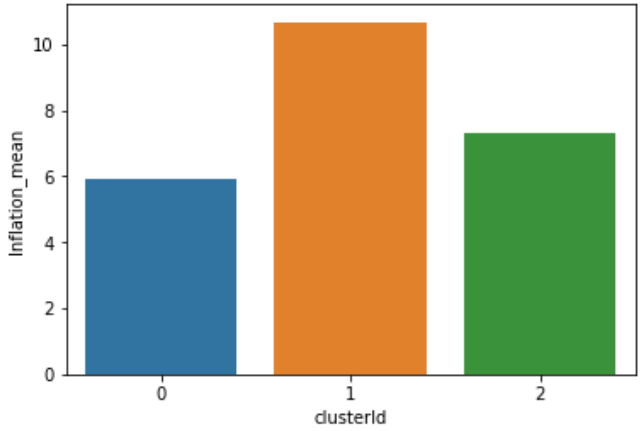


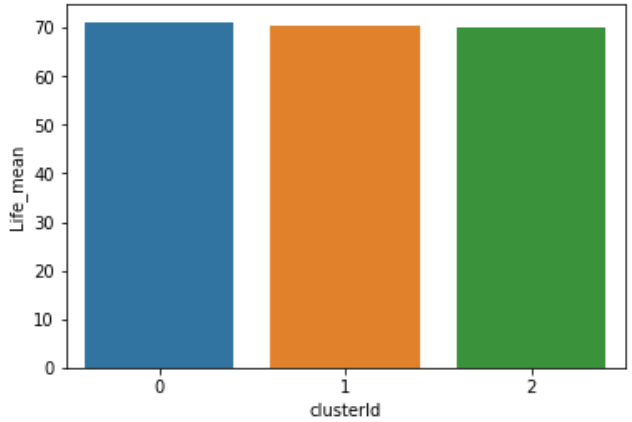


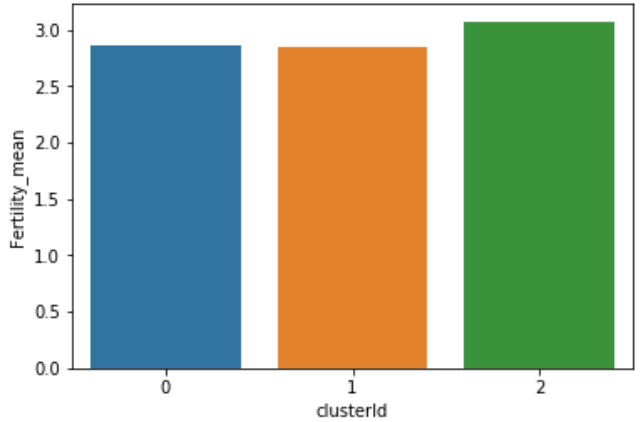


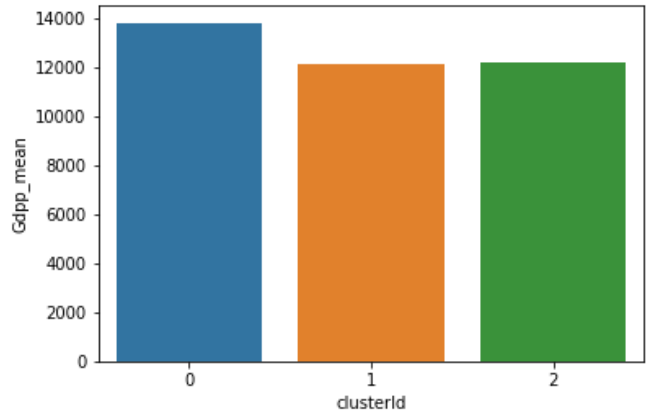



Income_mean
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
clusterId


Inflation_mean
10
8
6
4
2
0
0
1
2
clusterId


Life_mean
70
60
50
40
30
20
10
0
0
1
2
clusterId

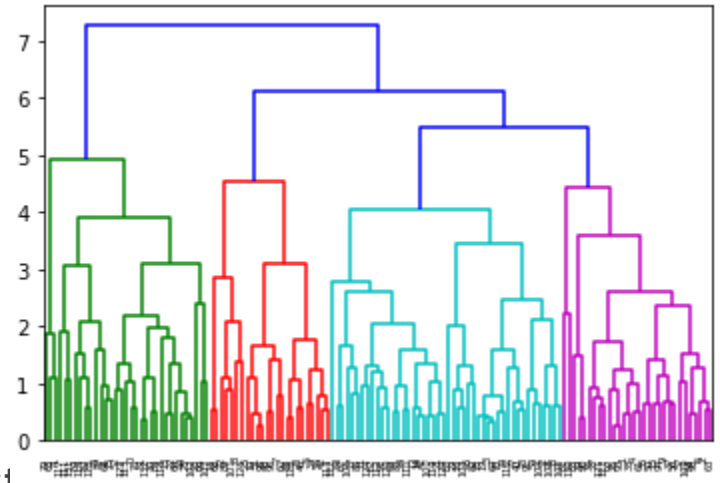

Fertility_mean
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
clusterId


Gdpp_mean
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
clusterId

- So looking at inflation, child_mort and gdpp, the top 5 countries which require financial assistance from HELP are Nigeria, Angola, Congo-Dem-Rep,Sierra Leone and Burundi. The other countries except Equatorial Guinea are also in dire need of financial aid.

Heirarchical Clustering

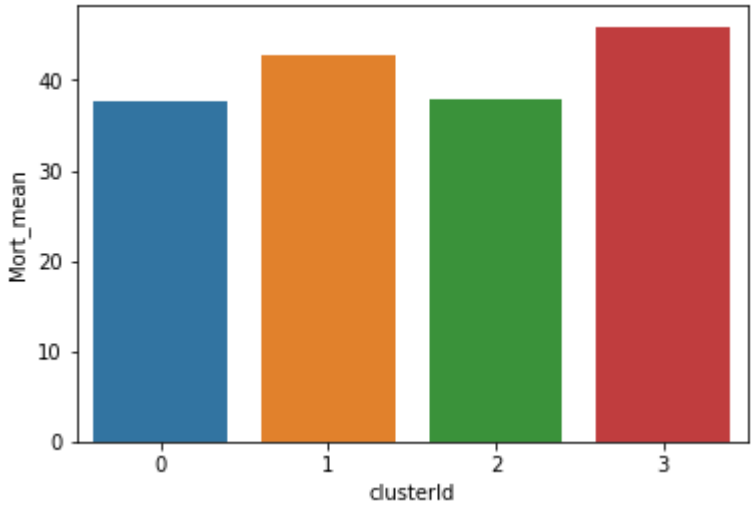
- We will use single and complete analysis method on dataset.
- Here based on the dendogram we will take 4 clusters.

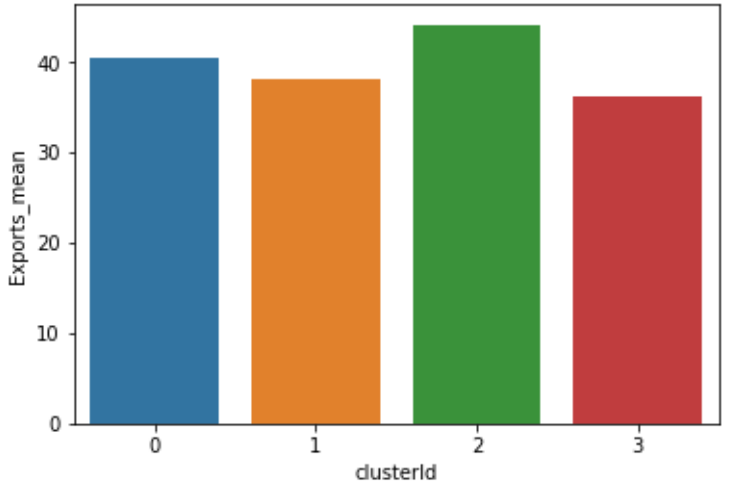


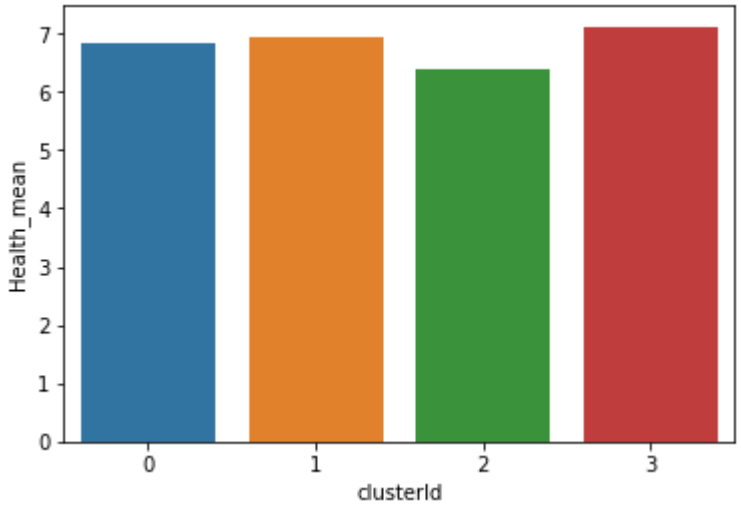
- We will count the value of clusters formed the values are:

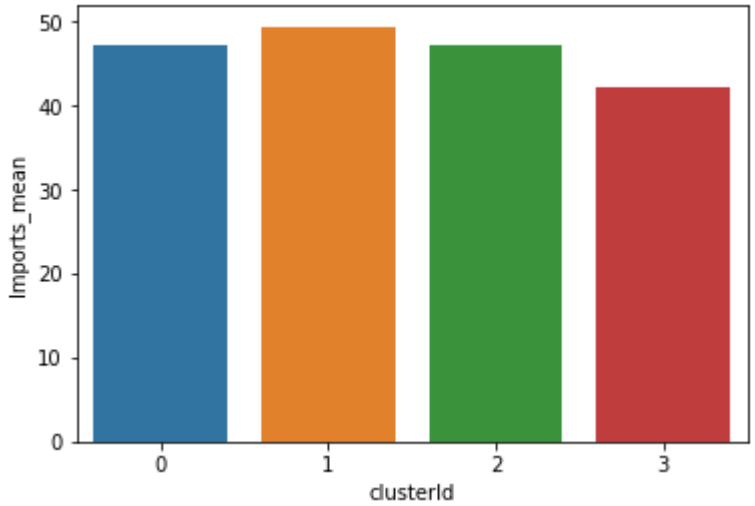
2.0	46
0.0	33
1.0	30
3.0	24

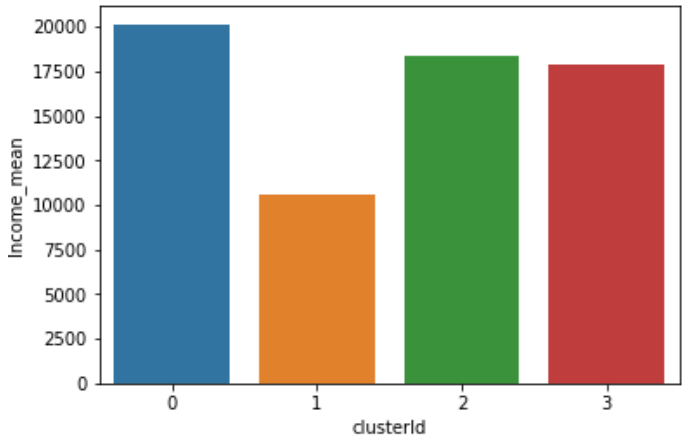
- To check the mean we will perform group by operation and then we will plot the bar plot.
- From the Barplot we got to know that the poor countries belongs to cluster1.

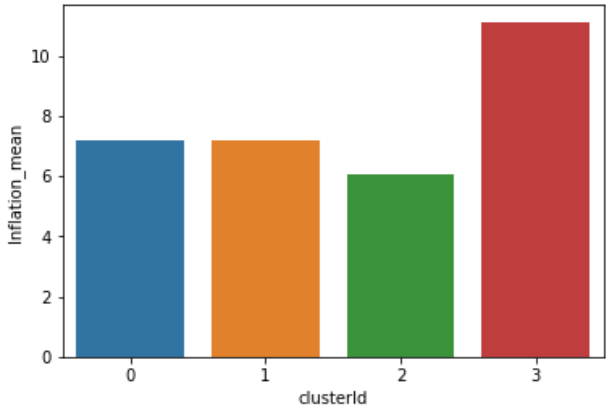


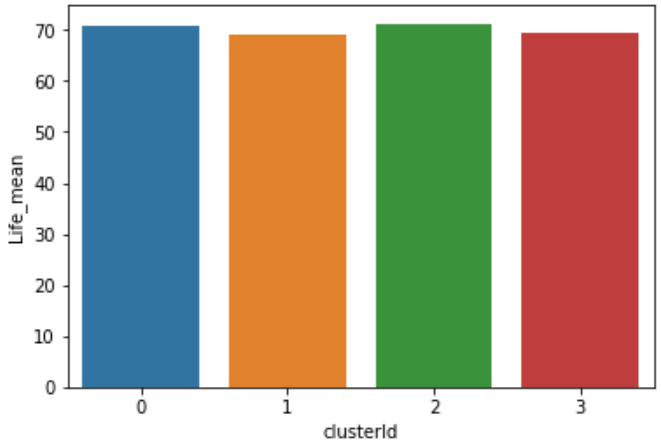


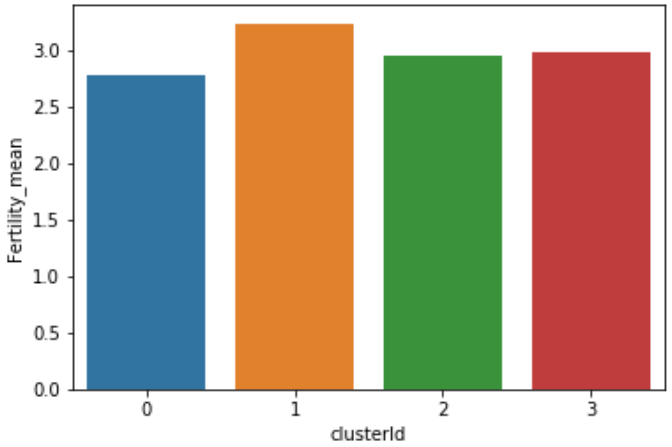


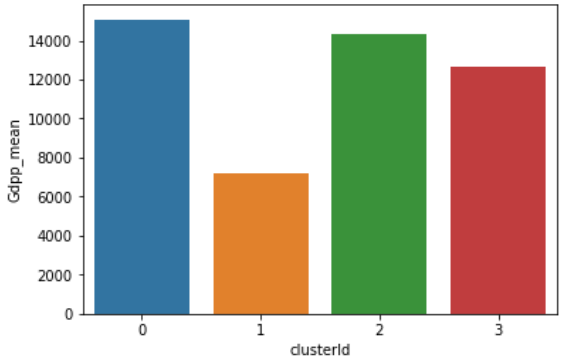



Income_mean
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
clusterId


Inflation_mean
10
8
6
4
2
0
0
1
2
3
clusterId


Life_mean
70
60
50
40
30
20
10
0
0
1
2
3
clusterId


Fertility_mean
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
clusterId


Gdpp_mean
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
clusterId

- Hierchical Clustering, by comparing gdpp & child_mort, we find Burundi, Liberia, Congo-Dem-Rep, Niger and Sierra Leone to be the Top 5 countries which need financial aid from HELP.

Thank You