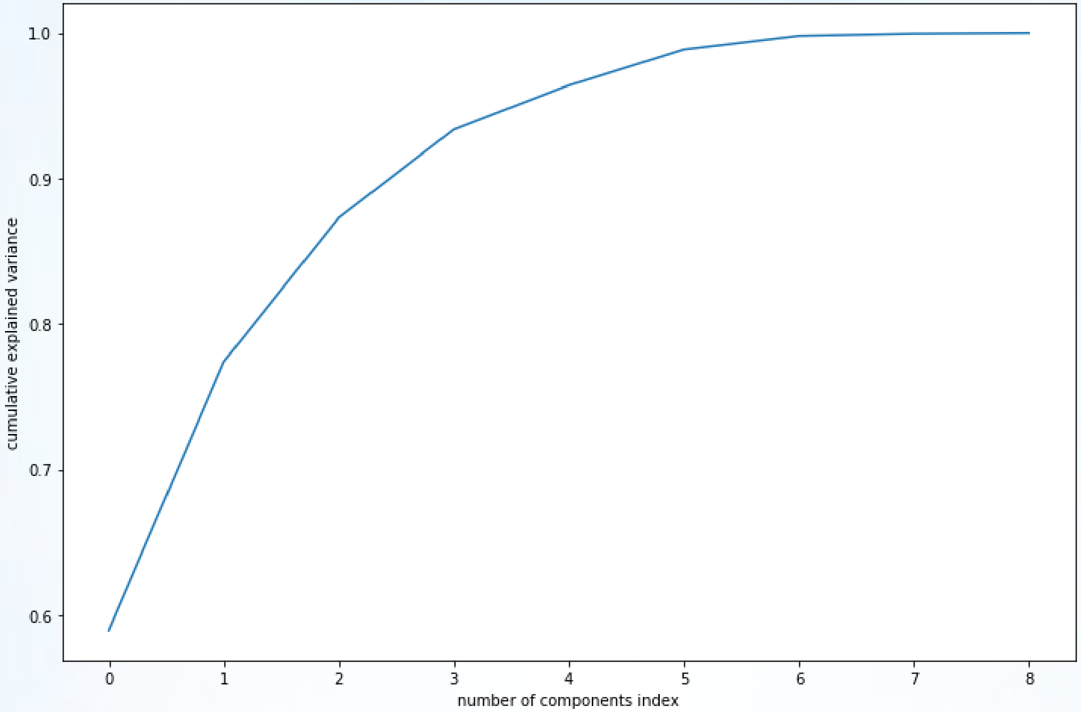
Clustering & PCA Assignment

Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

Categorizing the countries using some socio-economic and health factors that determine the overall development of the country. Then suggesting the countries which needs to be focus on the most.

Scree Plot

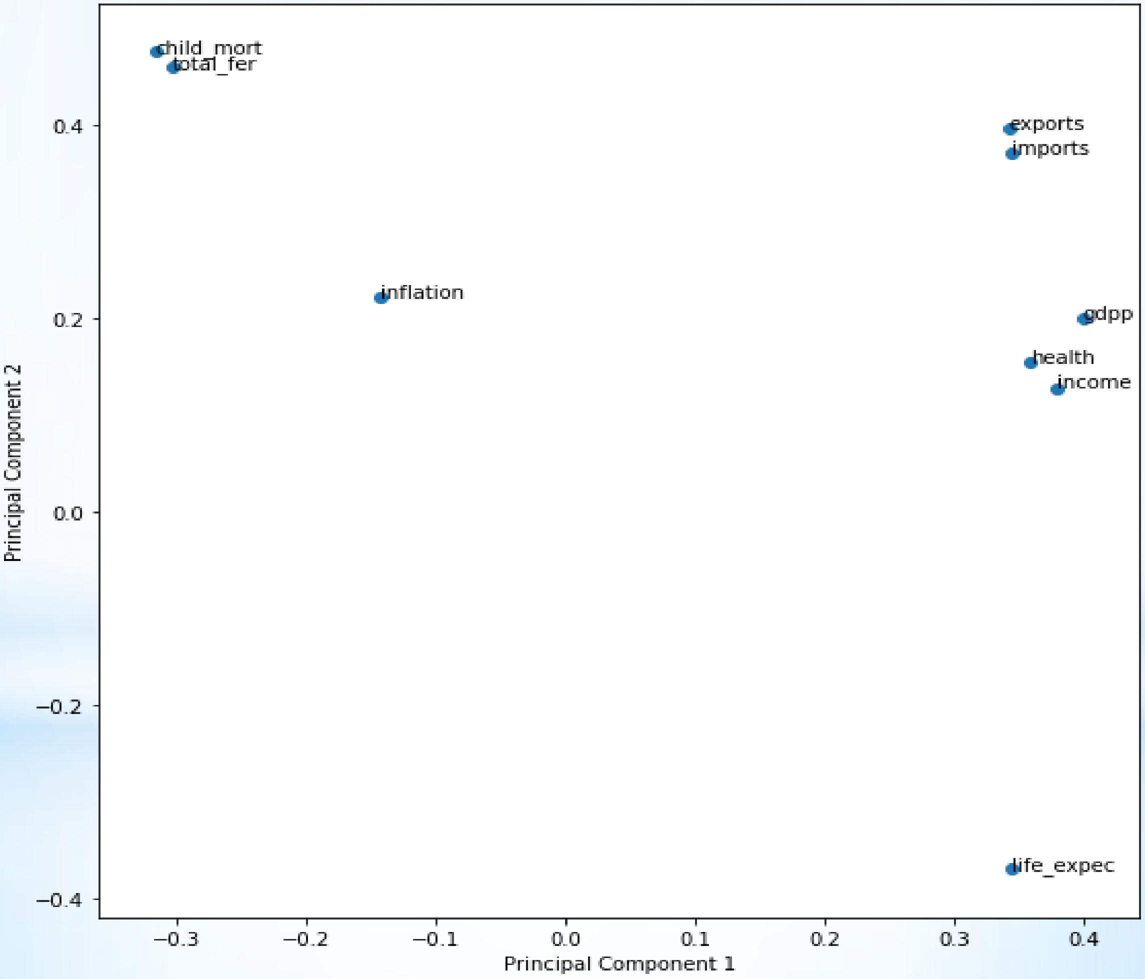


- As you can see 4 principle components sufficiently explained 93% of the variance.
- Equivalently this means that 4 themes are sufficient in explaining the dataset.

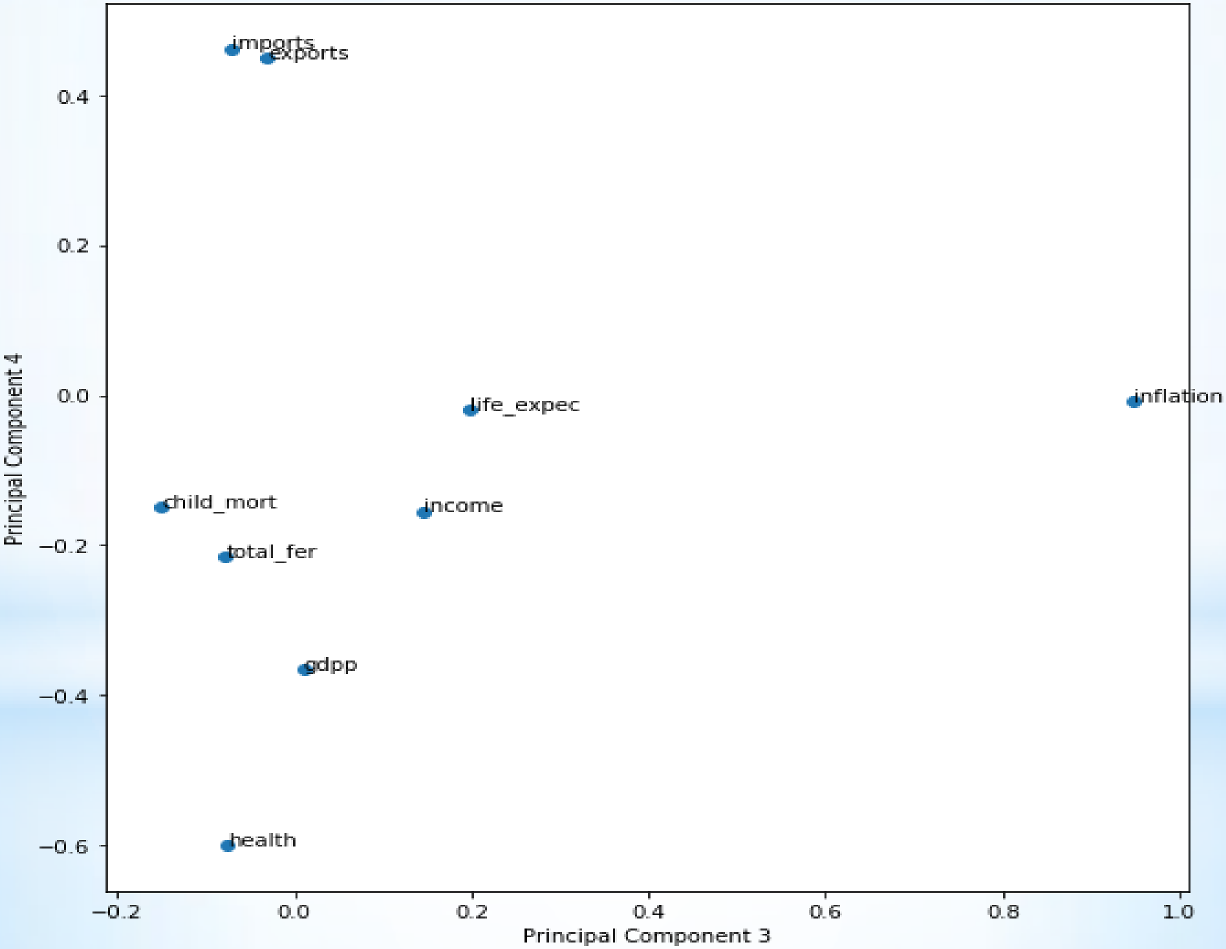
Explanation of original variable with the Principle Components

	Feature	PC1	PC2	PC3	PC4
0	child_mort	-0.316392	0.476267	-0.150012	-0.148052
1	exports	0.342887	0.397311	-0.030574	0.449425
2	health	0.358535	0.155053	-0.075703	-0.599712
3	imports	0.344865	0.370781	-0.072174	0.461798
4	income	0.380041	0.128384	0.145764	-0.154806
5	inflation	-0.143085	0.221261	0.948419	-0.007628
6	life_expec	0.343857	-0.369820	0.196752	-0.018395
7	total_fer	-0.302842	0.459715	-0.077834	-0.213928
8	gdpp	0.399988	0.200624	0.010339	-0.364772

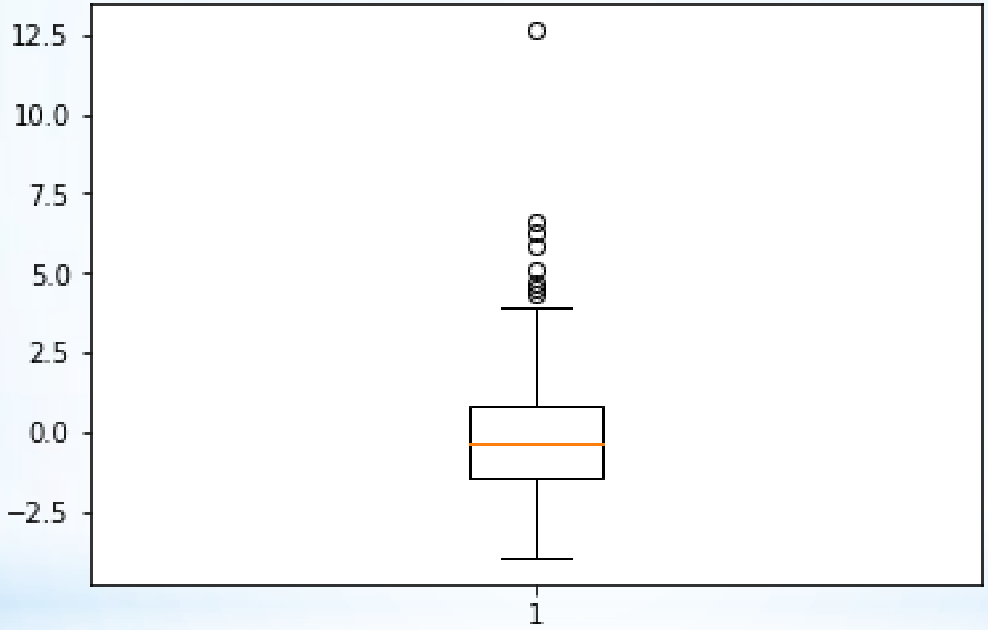
Plot of PC1 and PC2 to visualise the features



Plot of PC3 and PC4 to visualise the features

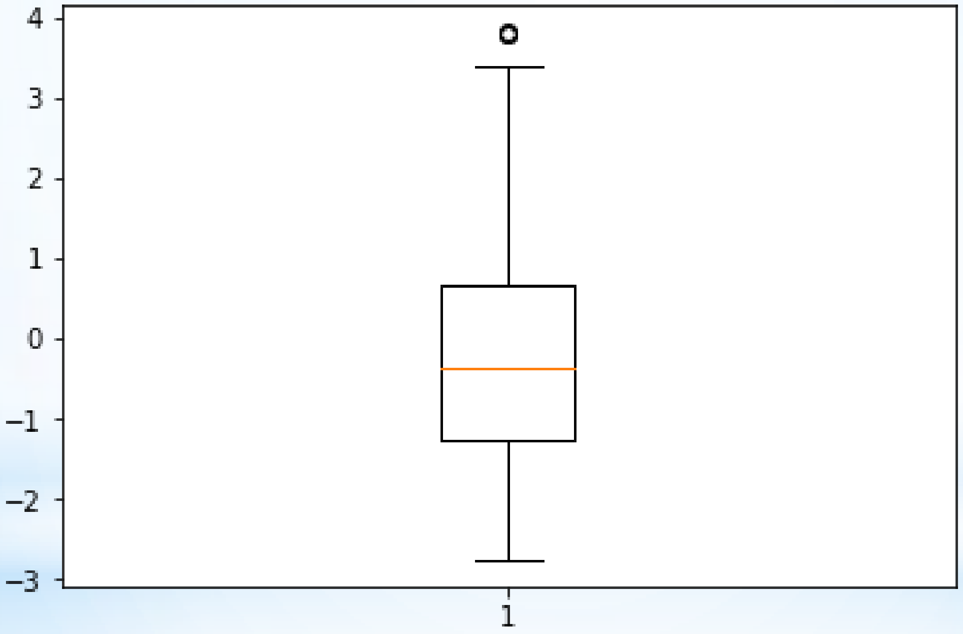


Boxplot of Principle Component 1

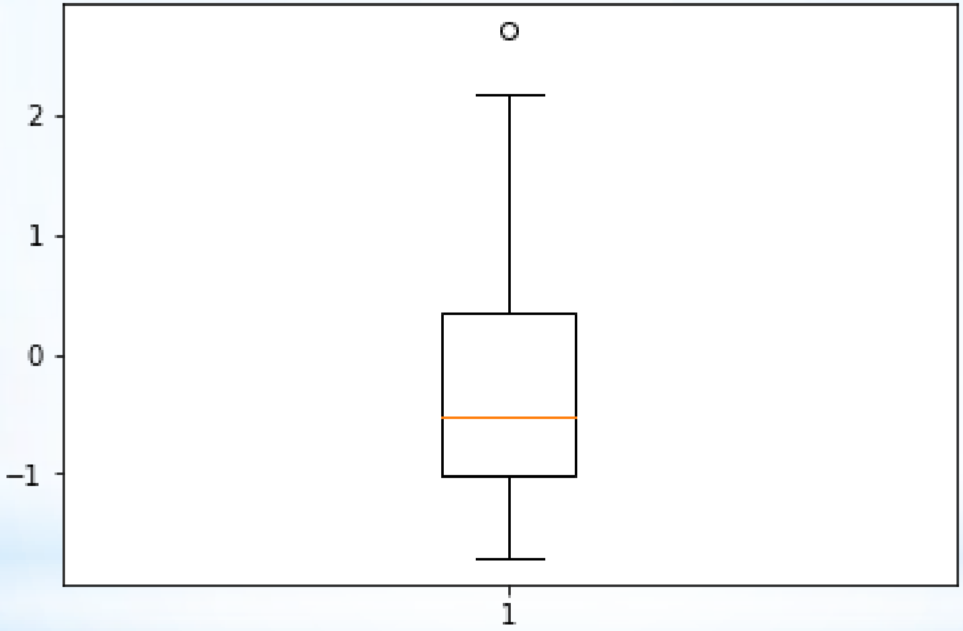


- There are lots of outliers so we have to remove them.

Box plot of PC1 after outlier treatment

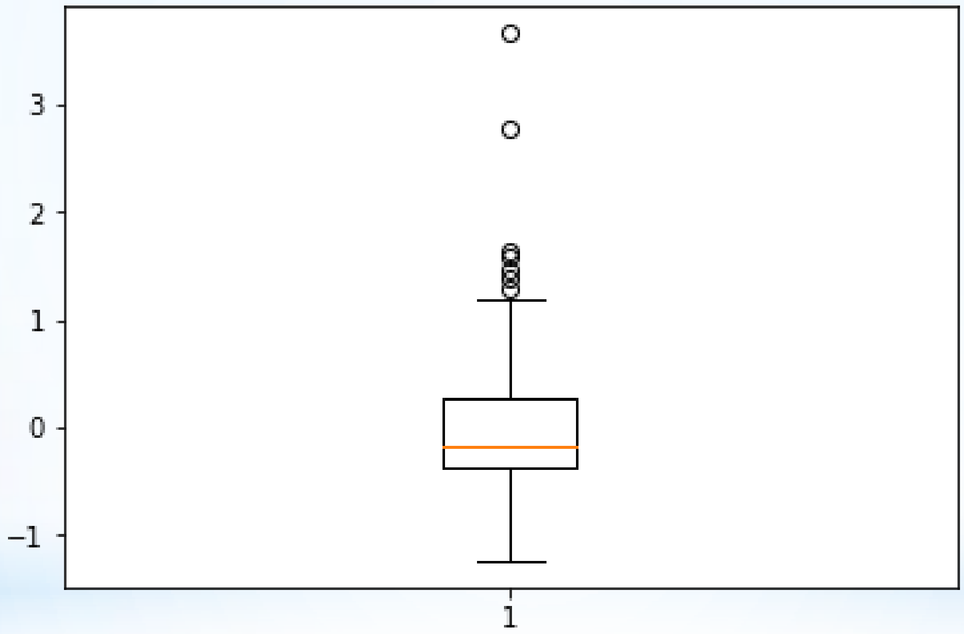


Box plot of Principle Component 2



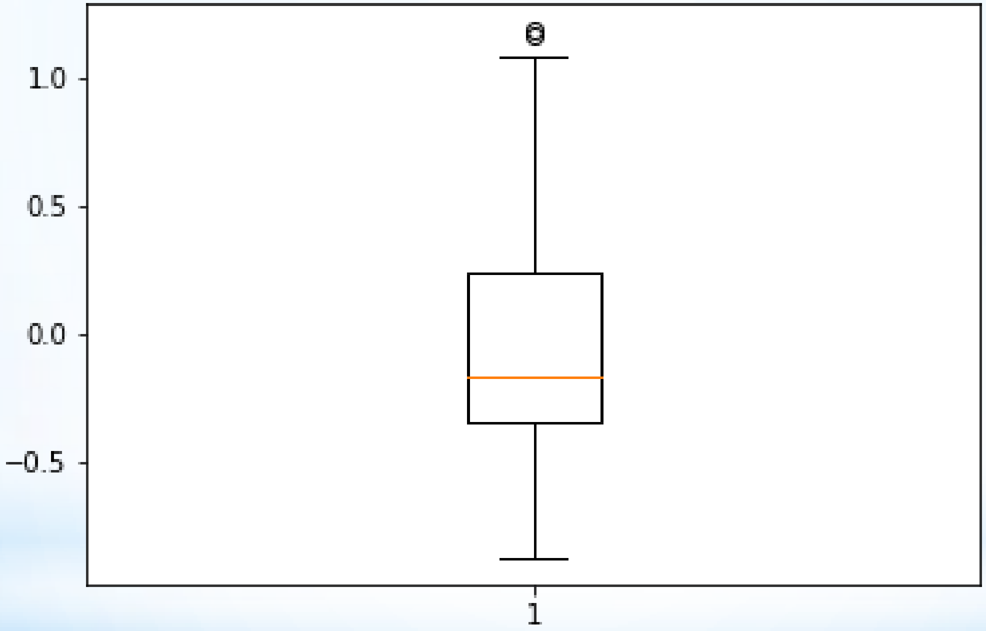
- ➢ There isn't much values outside, so we don't remove the outliers of this components.

Box plot of Principle Component 3

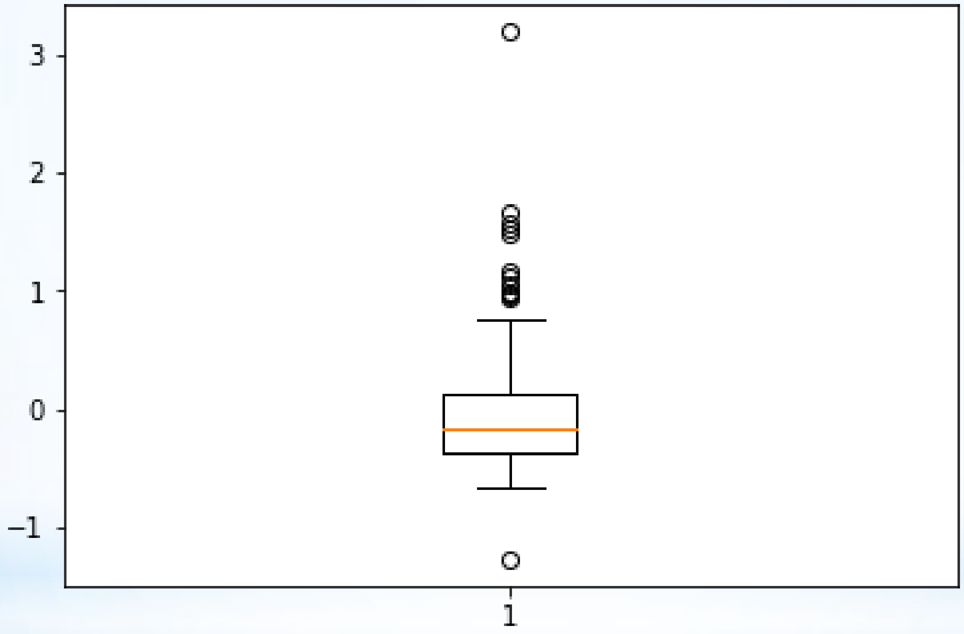


- There are lots of outliers so we have to remove them.

Box plot of PC3 after outlier treatment

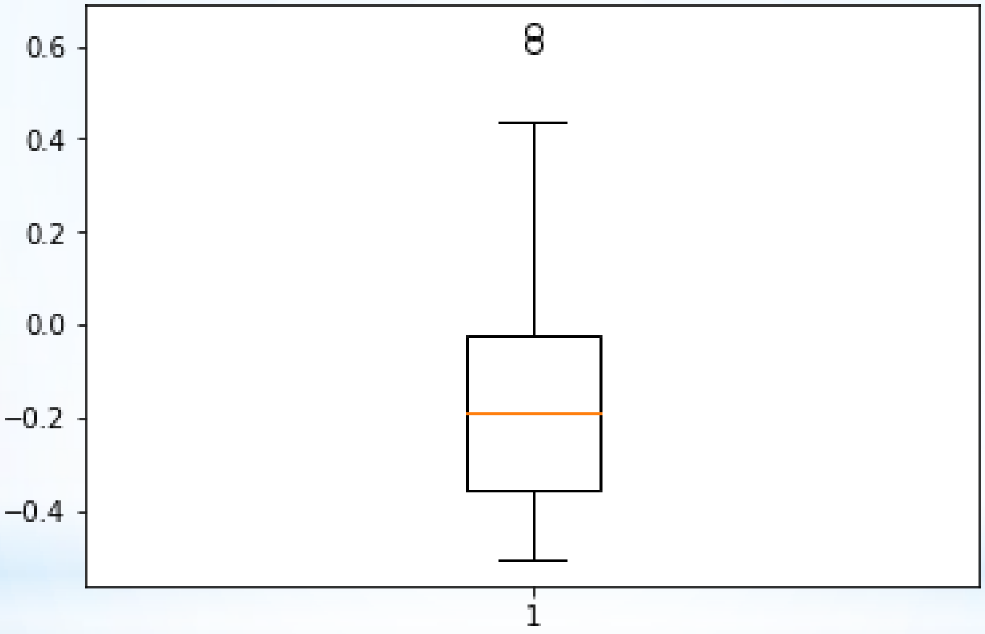


Box plot of Principle Component 4

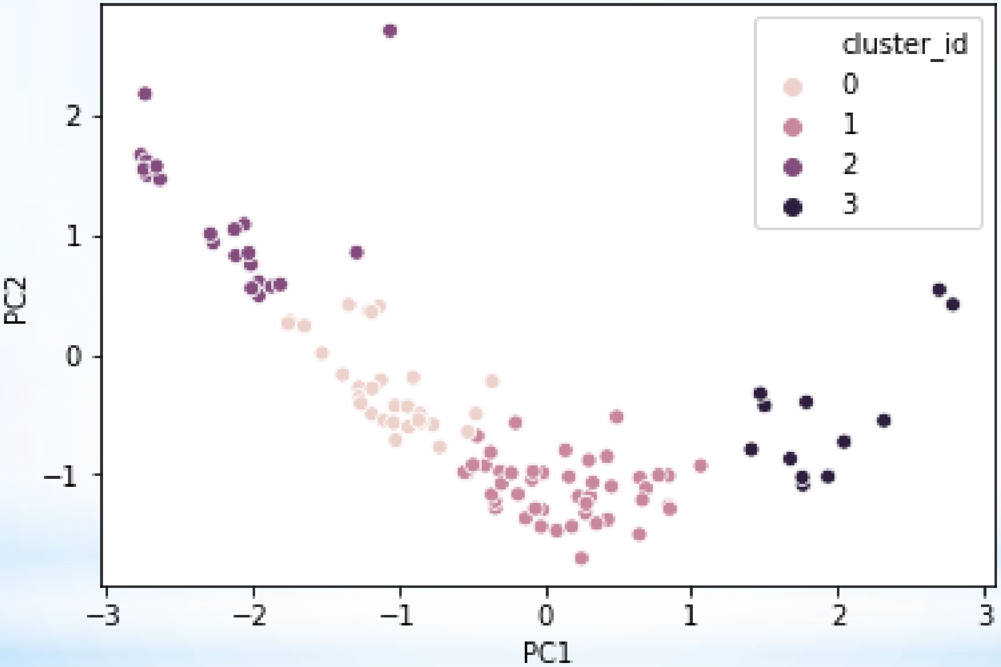


- There are lots of outliers so we have to remove them.

Box plot of PC4 after outlier treatment

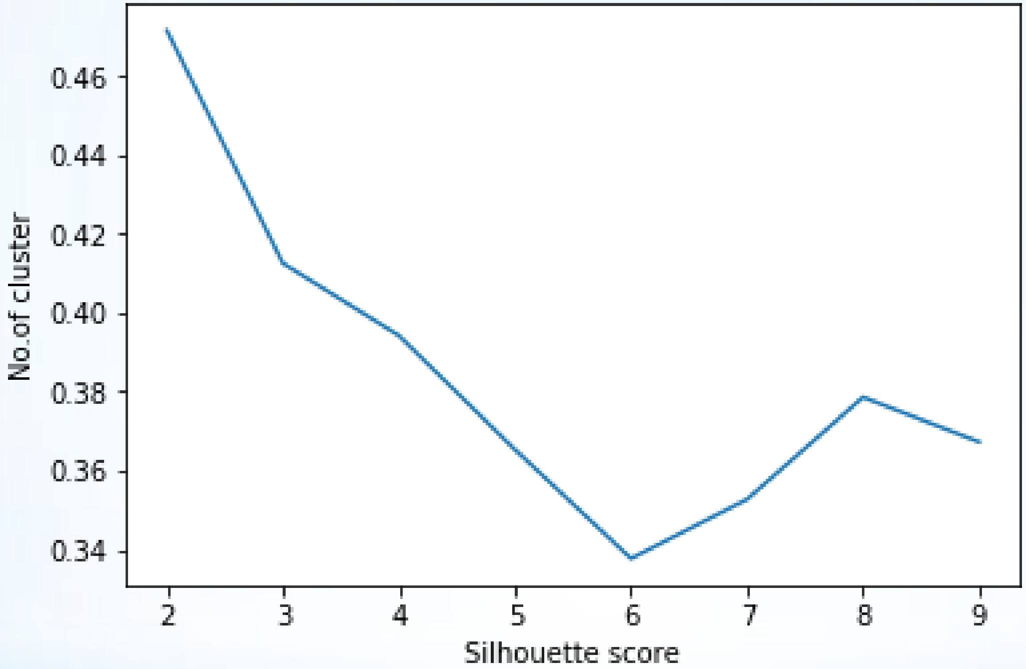


K-Means Clustering for K=4



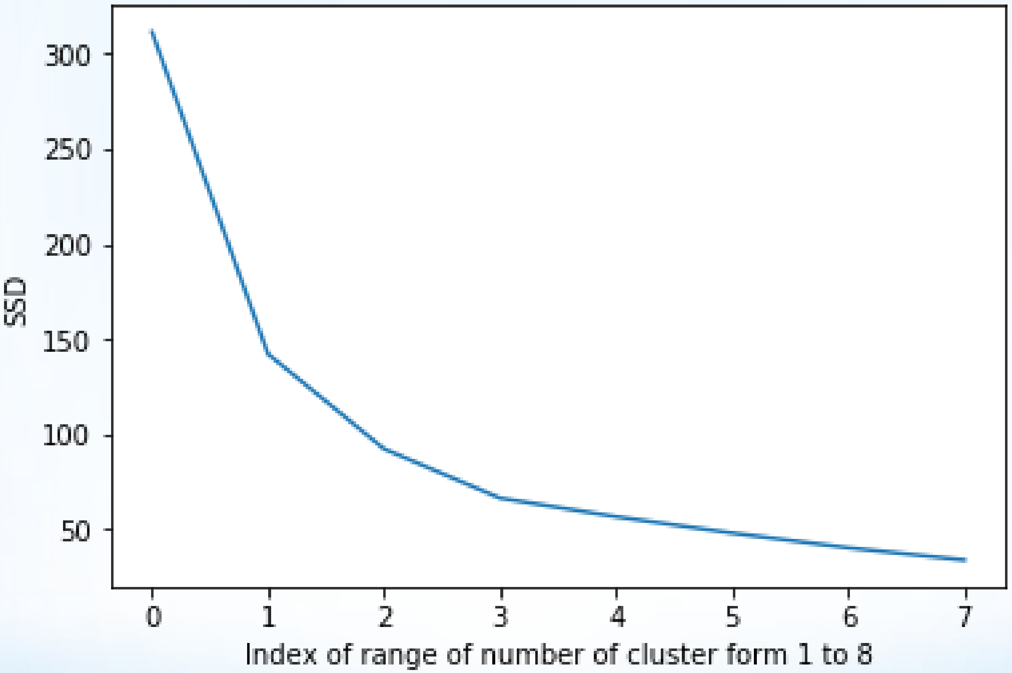
- ➢ Here you could see the 4 distinct clusters formed in PC1 and PC2 scatter plot.
- ➢ But we don’t know weather it’s ideal number of cluster.

Silhouette Score



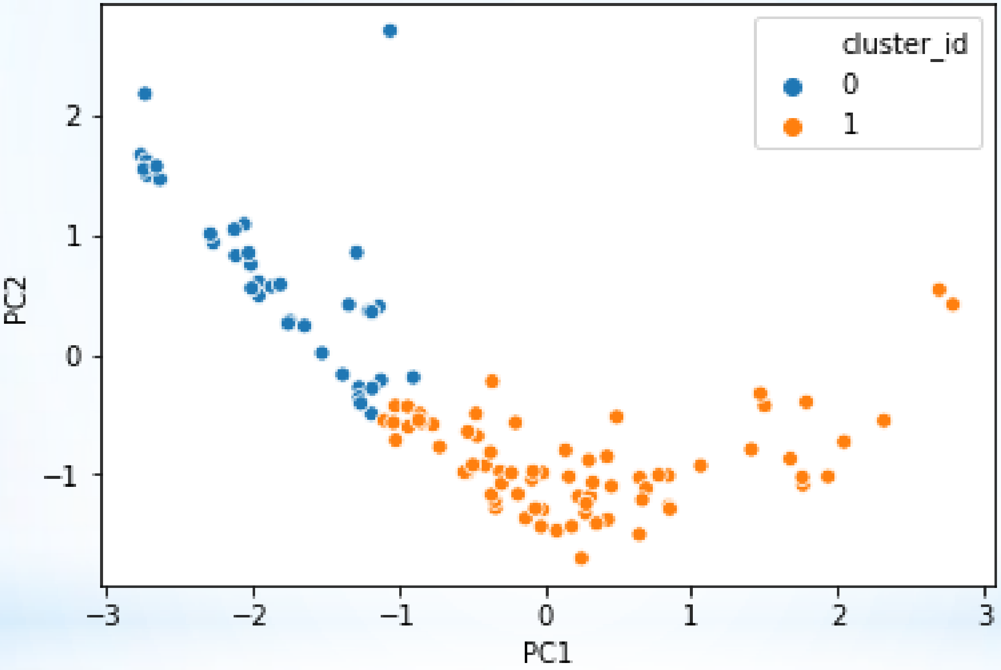
- From silhouette score method, optimum number of cluster = 2.

Elbow Curve



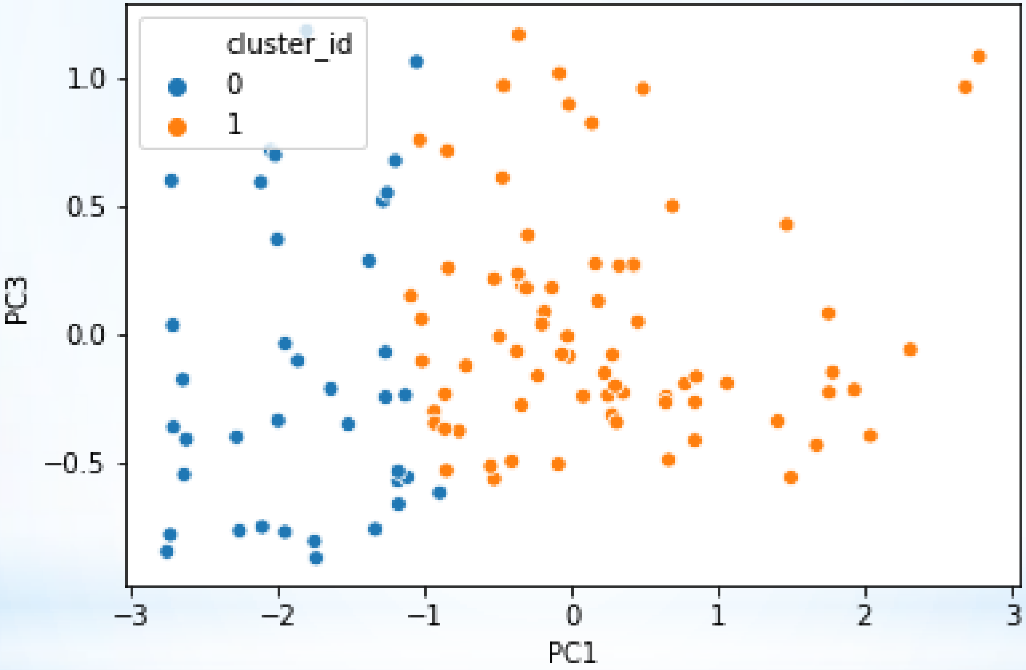
- From elbow curve method optimum number of cluster = 2.

K-Means Clustering for K=2



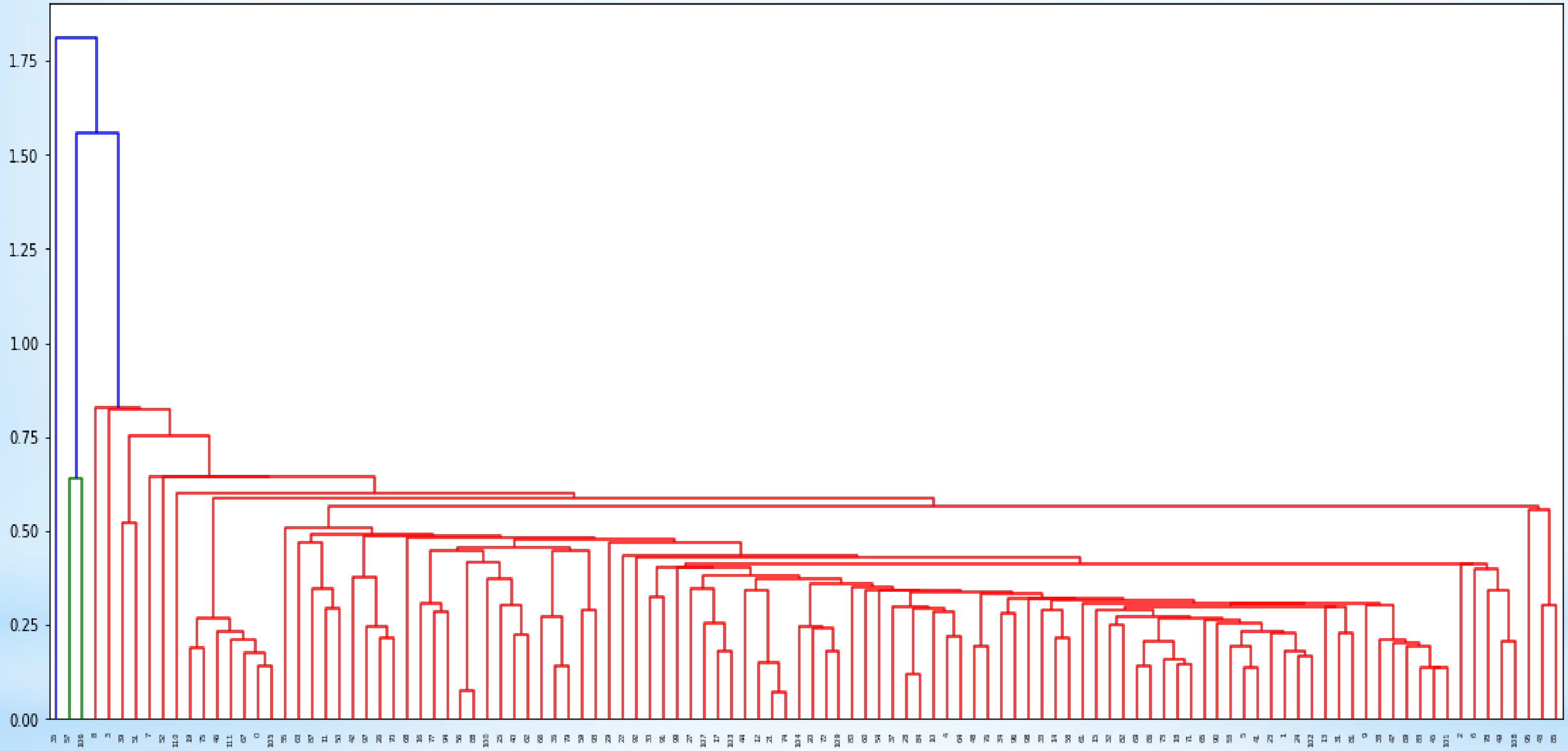
- Here you could see 2 distinct clusters formed in PC1 and PC2 plot.

K-Means Clustering for K=2



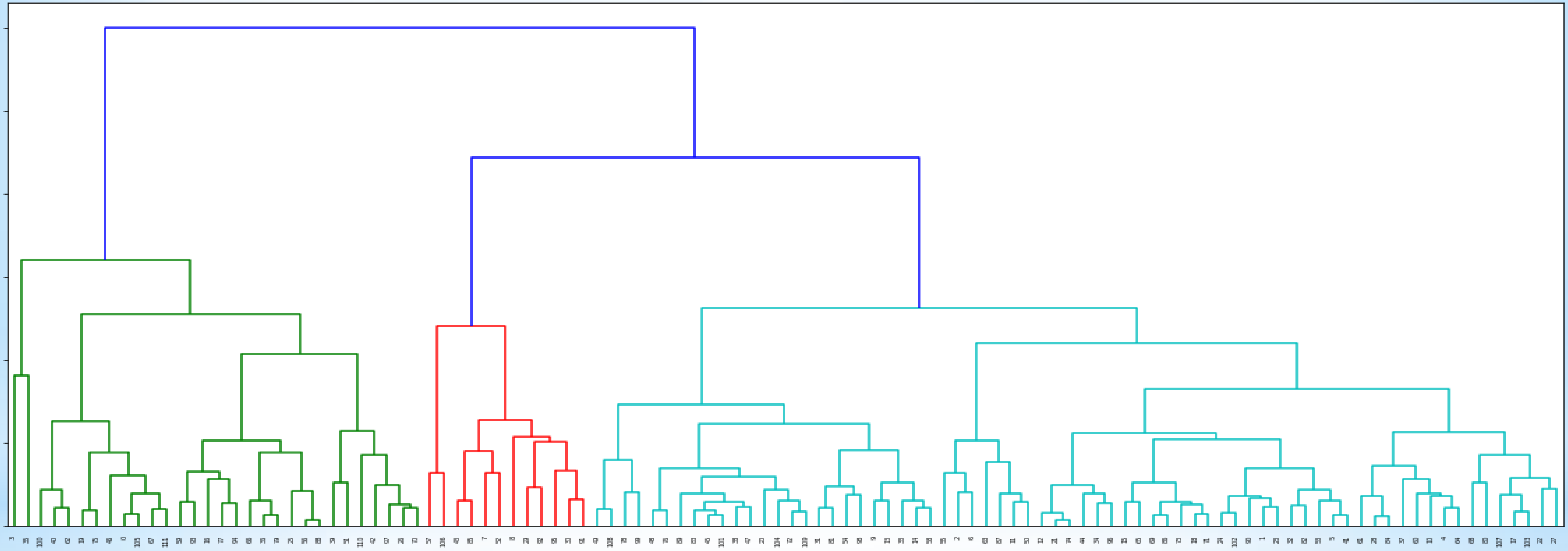
- Here you could see 2 distinct clusters formed in PC1 and PC3 plot.

Hierarchical Clustering Dendrogram (Single Linkage)



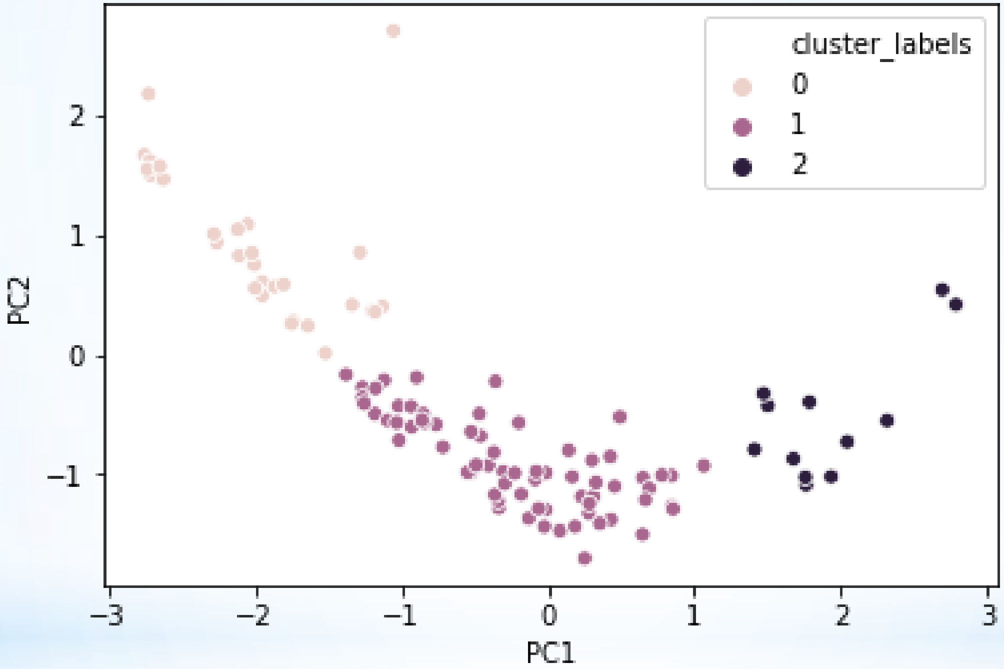
- ➢ Here you could see the dendrogram for single linkage.
- ➢ It’s very untidy, so we will see how complete linkage would look.

Hierarchical Clustering Dendrogram(Complete Linkage)



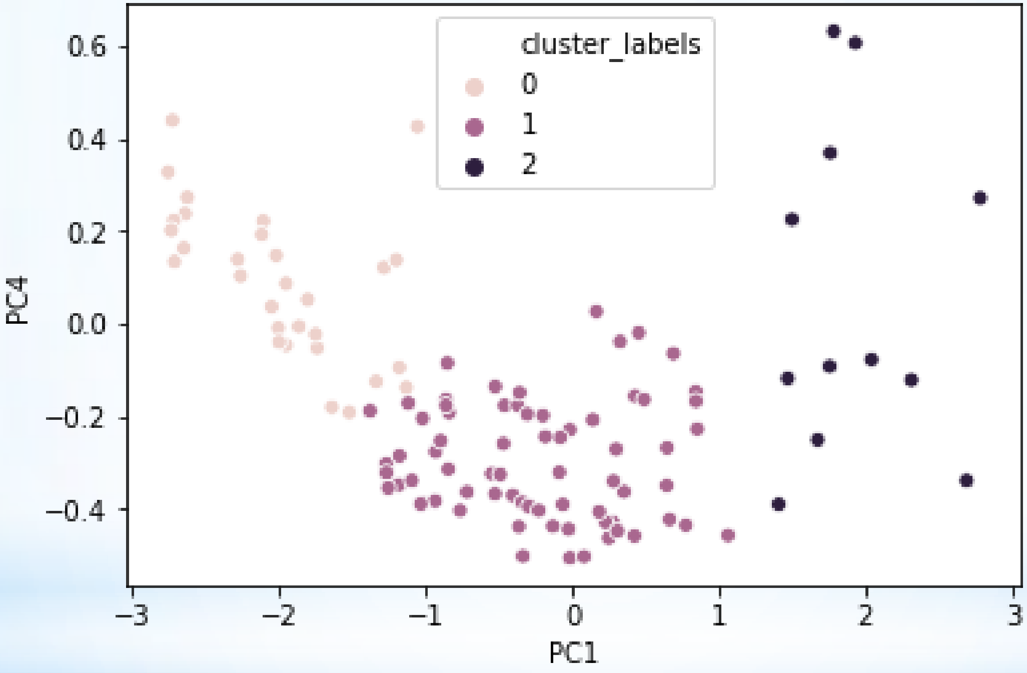
- ➢ Here you could see the dendrogram fro complete linkage.
- ➢ Here the clusters are somewhat uniformly distributed, so we will proceed using this.
- ➢ From both the linkage graph its clear to have 3 clusters.

Hierarchical Clustering with no of clusters = 3



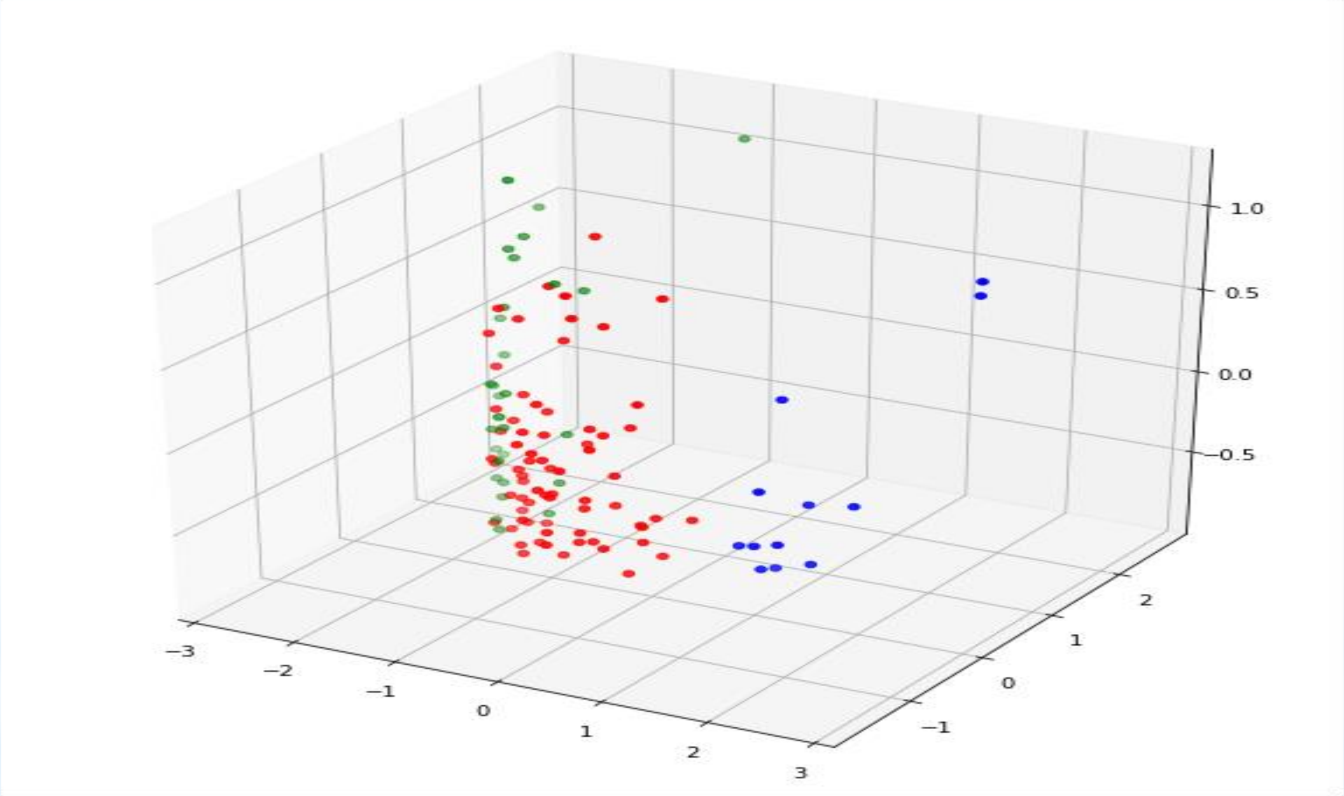
- ➢ Here you could see 3 distinct clusters formed in PC1 and PC2 plot.

Hierarchical Clustering with no of clusters = 3



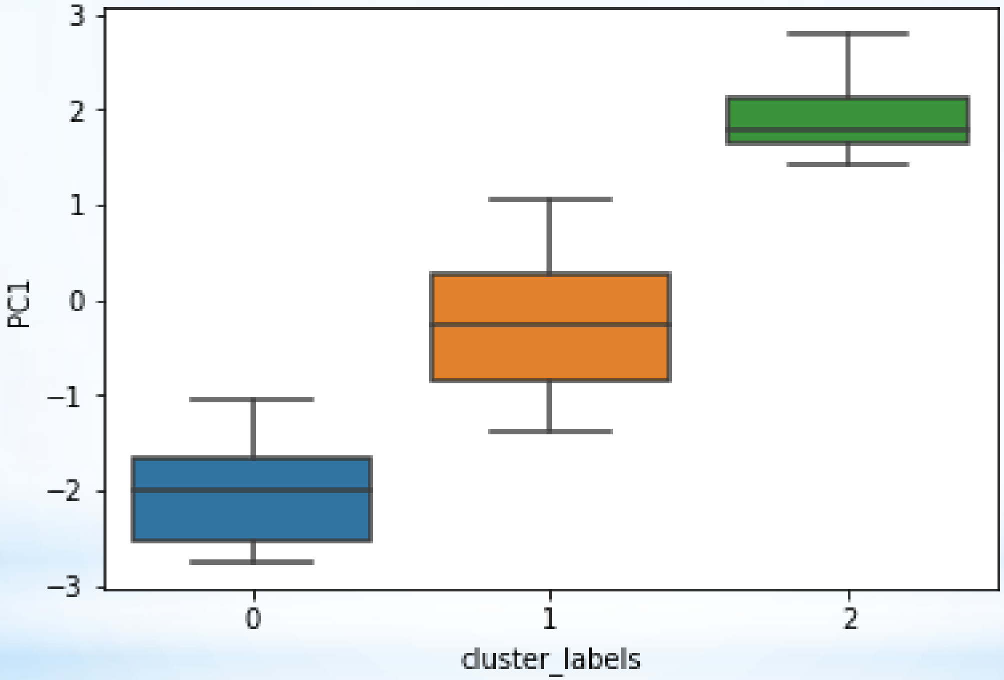
- Here you could see 3 distinct clusters formed in PC1 and PC4 plot.

Hierarchical Clustering with no of clusters = 3



- Here you could see 3 distinct clusters in PC1,PC2 and PC3 plot.

Variation of PC1 in respective clusters



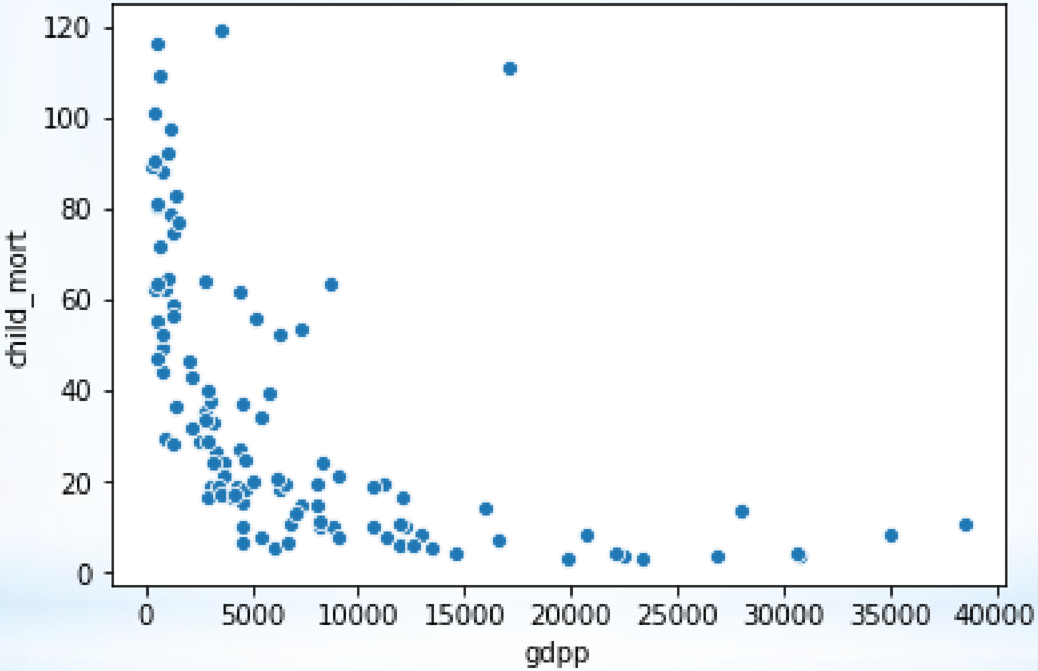
- Here you could see that PC1 has lowest average value in cluster 0.
- Cluster 2 has highest average value of PC1.

Table of average values of in each cluster

	Cluster_labels	gdpp	Child_mort	income
0	0	2456.800000	76.820000	5147.266667
1	1	5897.457143	23.088571	11348.571429
2	2	26241.666667	6.291667	35733.333333

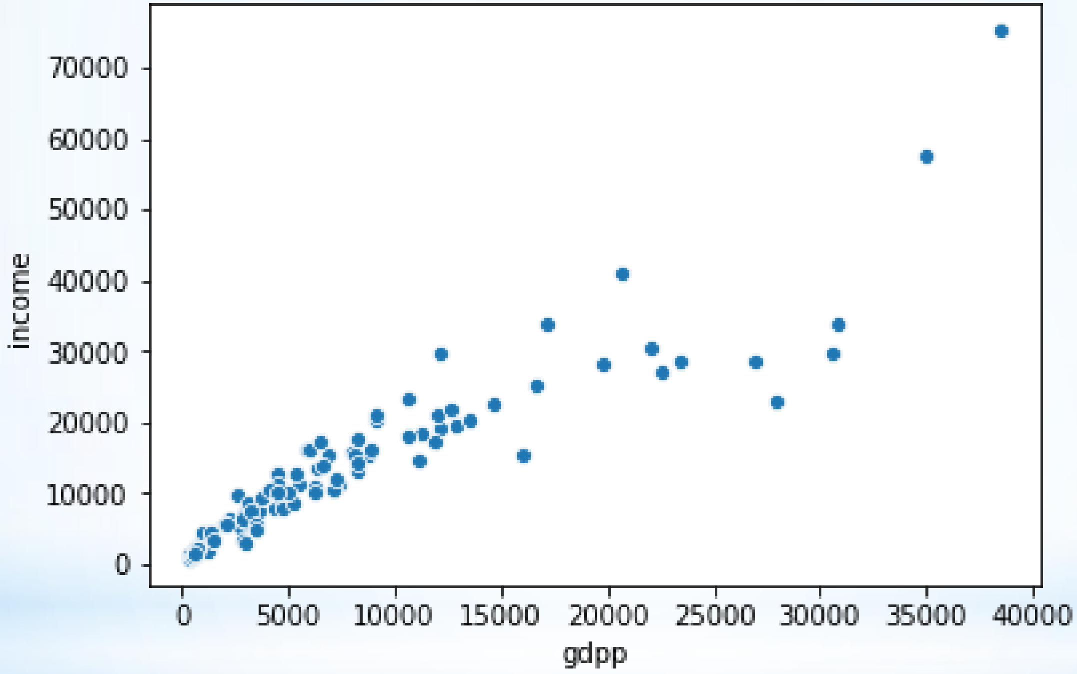
- Here you could see that the cluster '0' contains the countries having low gdpp, low income and very high child_mort.
- Cluster labels = 0 are the countries which are underdeveloped.
- Cluster labels = 1 are the countries which are developing.
- Cluster labels = 2 are the countries which are developed.

Scatter Plot of child_mort vs gdpp



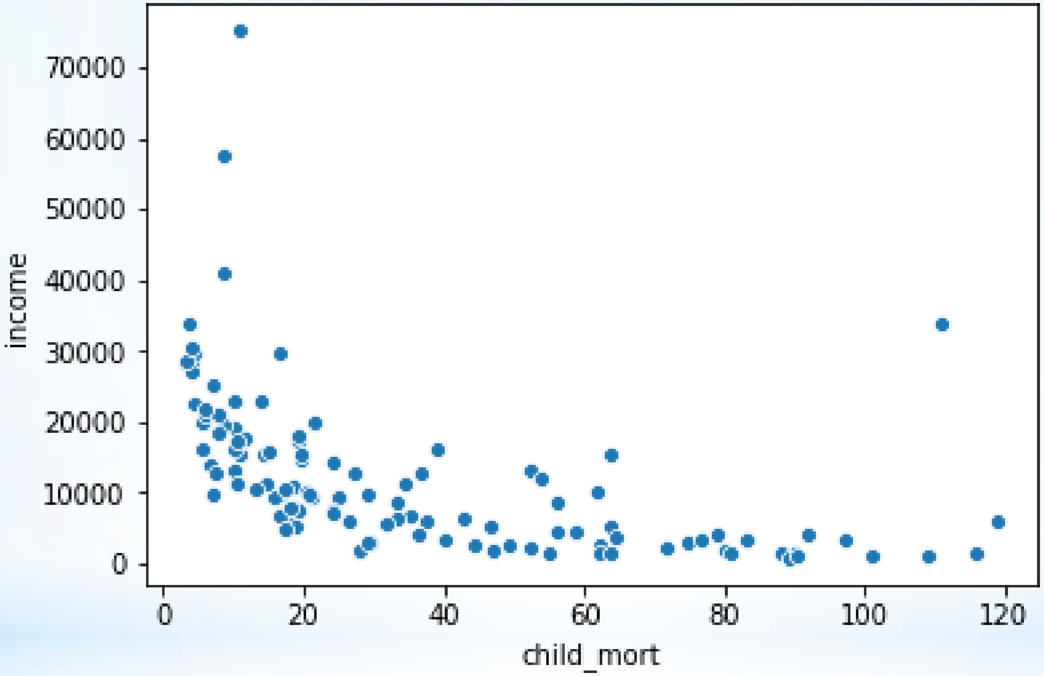
- Here you could see that the countries having high child_mort are also have very low gdpp.
- Countries having high gdpp has a very low child_mort.

Scatter Plot of income vs gdpp



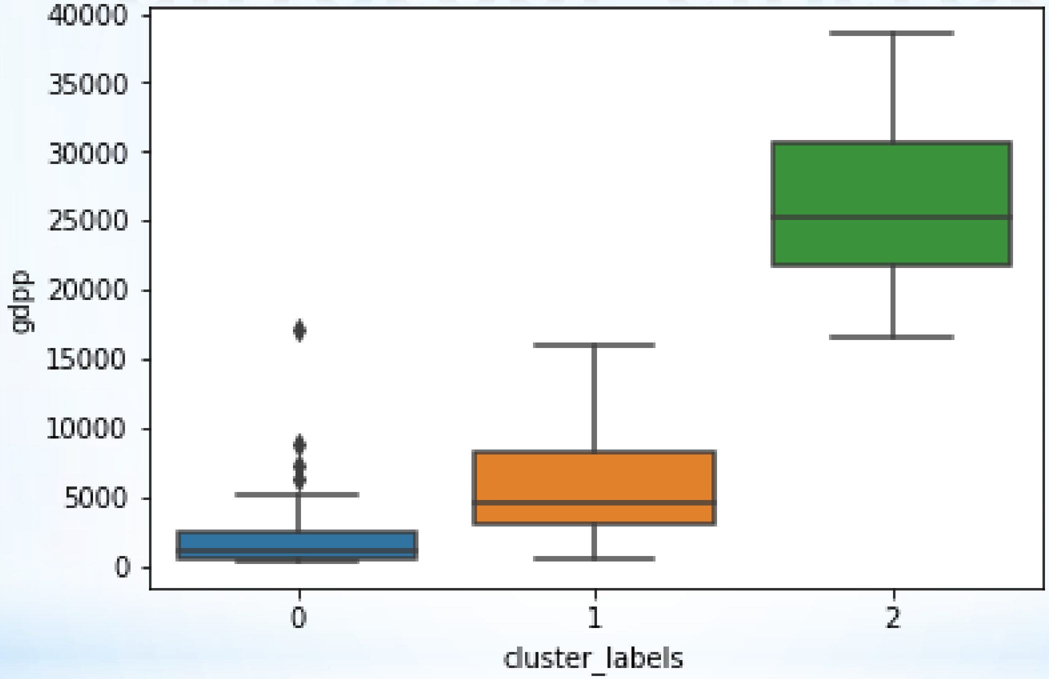
- ➢ Here you could see that the countries gdpp is directly proportional to the average income of the people of that country.

Scatter Plot of income vs child_mort



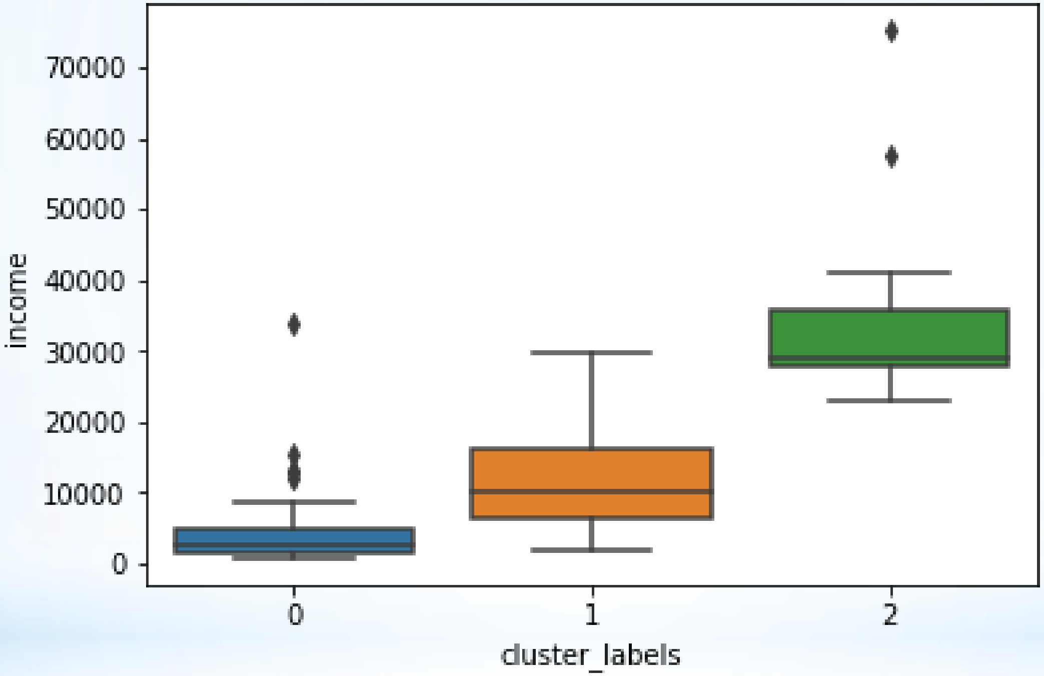
- Here you could see that countries having high child mortality is also the country where people's average income is low.
- Country where people's average income is high the child mortality is also low.

Box plot of gdpp for different clusters



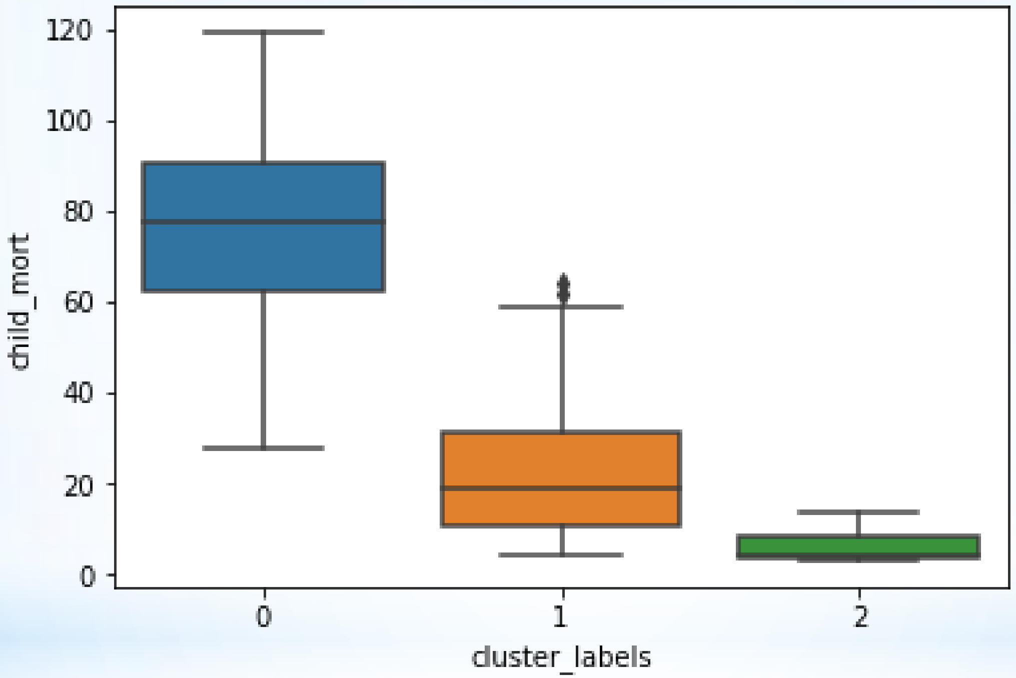
- Here you could easily see that for underdeveloped countries the average gdpp is lowest.
- Developed countries has average gdpp heighest.

Box Plot of income for different clusters



- Here you could see that the income follow somewhat similar pattern to that of gdpp graph.
- Underdeveloped countries ha0s the lowest income and developed has the heighest.

Box Plot of child_mort for different clusters



- ➢ Here you could see that the child mortality is highest in the underdeveloped countries.
- ➢ Child mortality is lowest for the developed countries.

Result

Taking into account all the variables, the countries that needed most attention are:

- Liberia
- Mozambique
- Malawi
- Afghanistan
- Burkina Faso
- Guinea
- Uganda