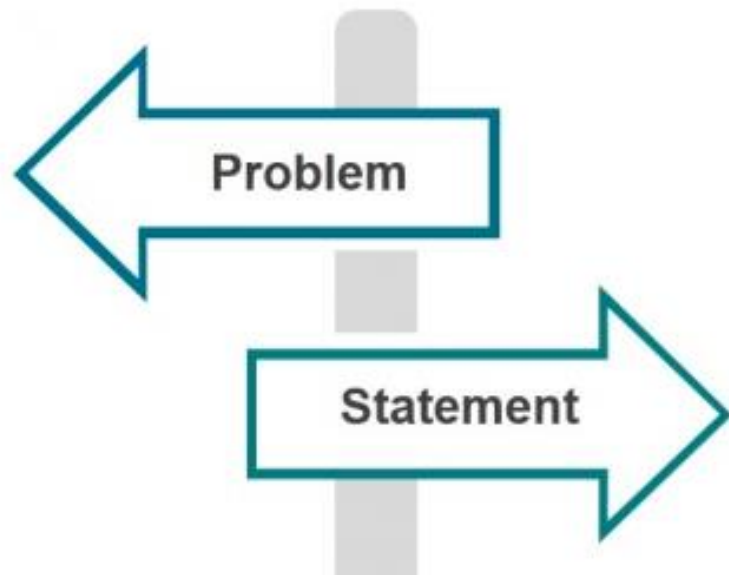


Assignment: Clustering and PCA





- **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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.**
- **After the recent funding programmes, they have been able to raise around \$ 10 million. Now we need to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.**

The steps are below followed for find the list of Country direst need of aid.

- ✓ **Step 1: Reading and Understanding the Data**
- ✓ **Step 2: Data Preparation**
- ✓ **Step 3:EDA(Exploratory Data Analytics)**
- ✓ **Step 4: Data Preparation**
- ✓ **Step 5: PCA(Principal Components Analysis)**
- ✓ **Step 6: Hopkins Statistics Test**
- ✓ **Step 7: Model Building - KMeans Clustering**
- ✓ **Step 8: Model Building - Hierarchial Clustering**
- ✓ **Step 9: Final Analysis – List out the country**
- ✓ **Step 10: Closing Statement and list of country direst need of aid via filtering on the basis of gdpp, child_mort and income.**

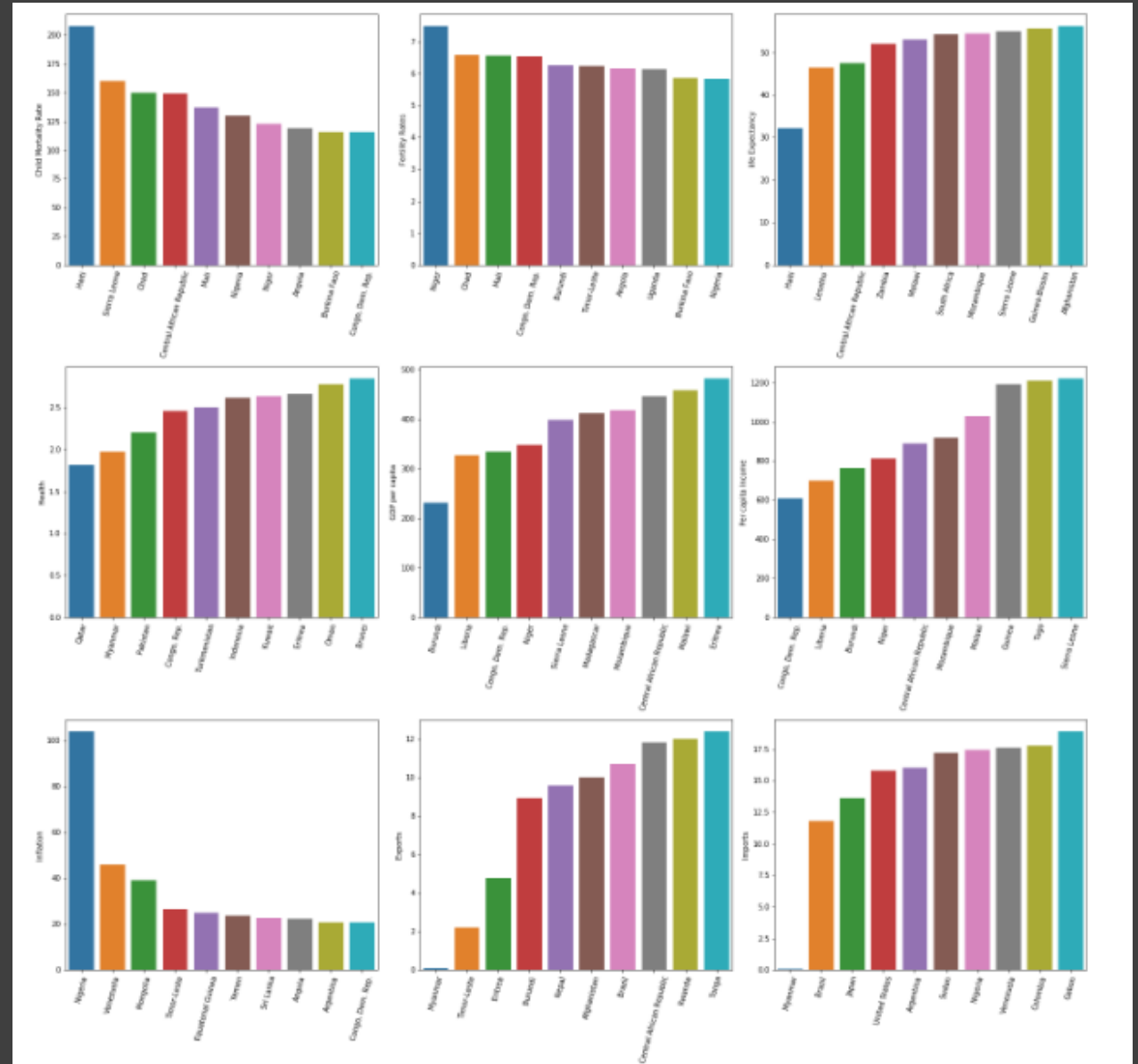
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 100...
2	exports	Exports of goods and services. Given as %age o...
3	health	Total health spending as %age of Total GDP
4	imports	Imports of goods and services. Given as %age o...
5	Income	Net income per person
6	Inflation	The measurement of the annual growth rate of t...
7	life_expec	The average number of years a new born child w...
8	total_fer	The number of children that would be born to e...
9	gdpp	The GDP per capita. Calculated as the Total GD...

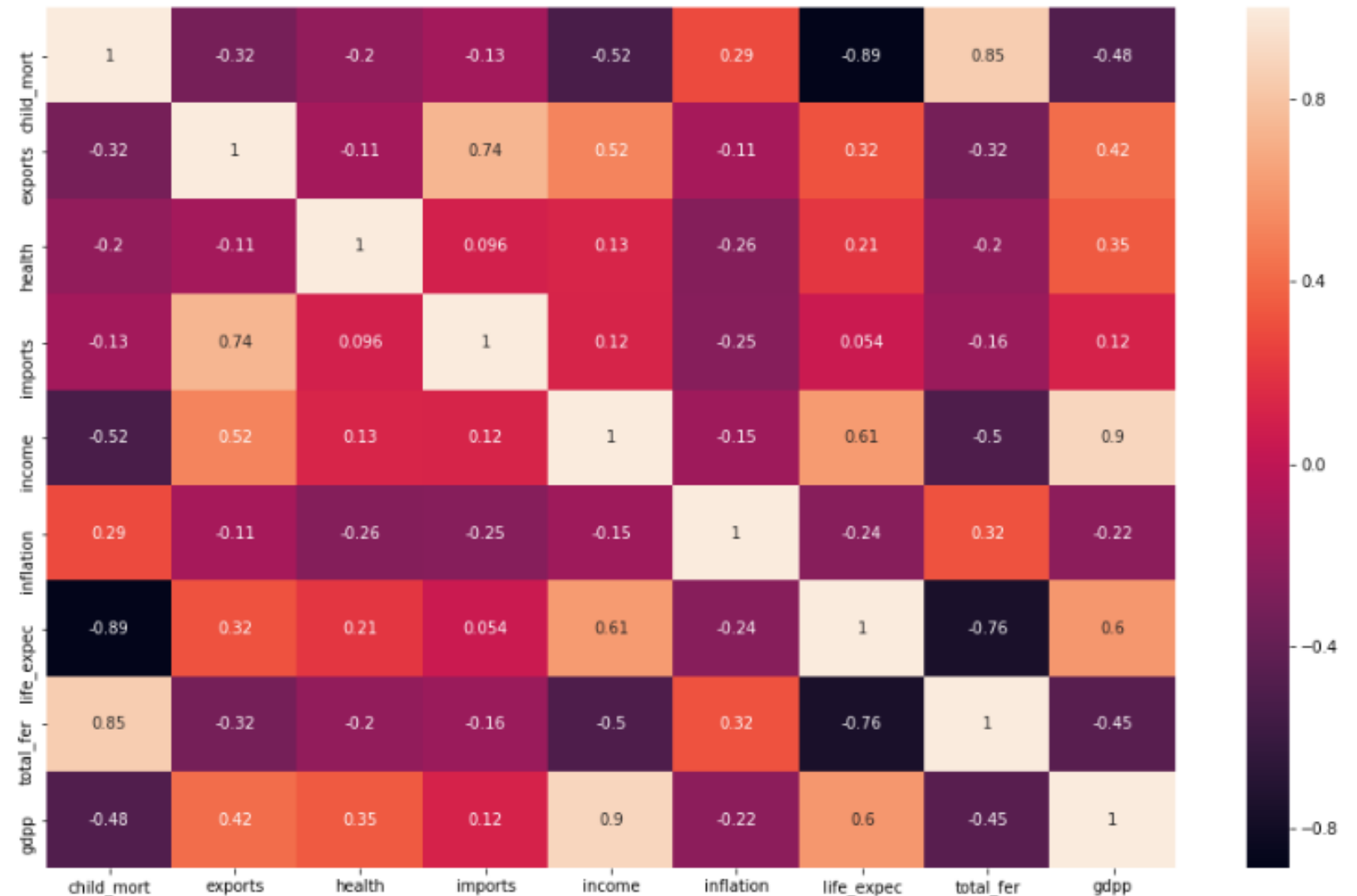
The datasets containing those socio-economic factors and the corresponding data dictionary are provided.

Before any process just want to see top and bottom 10 country of each column feature it will give some information regarding Country.

Top and Lowest 10 countries for each factor.



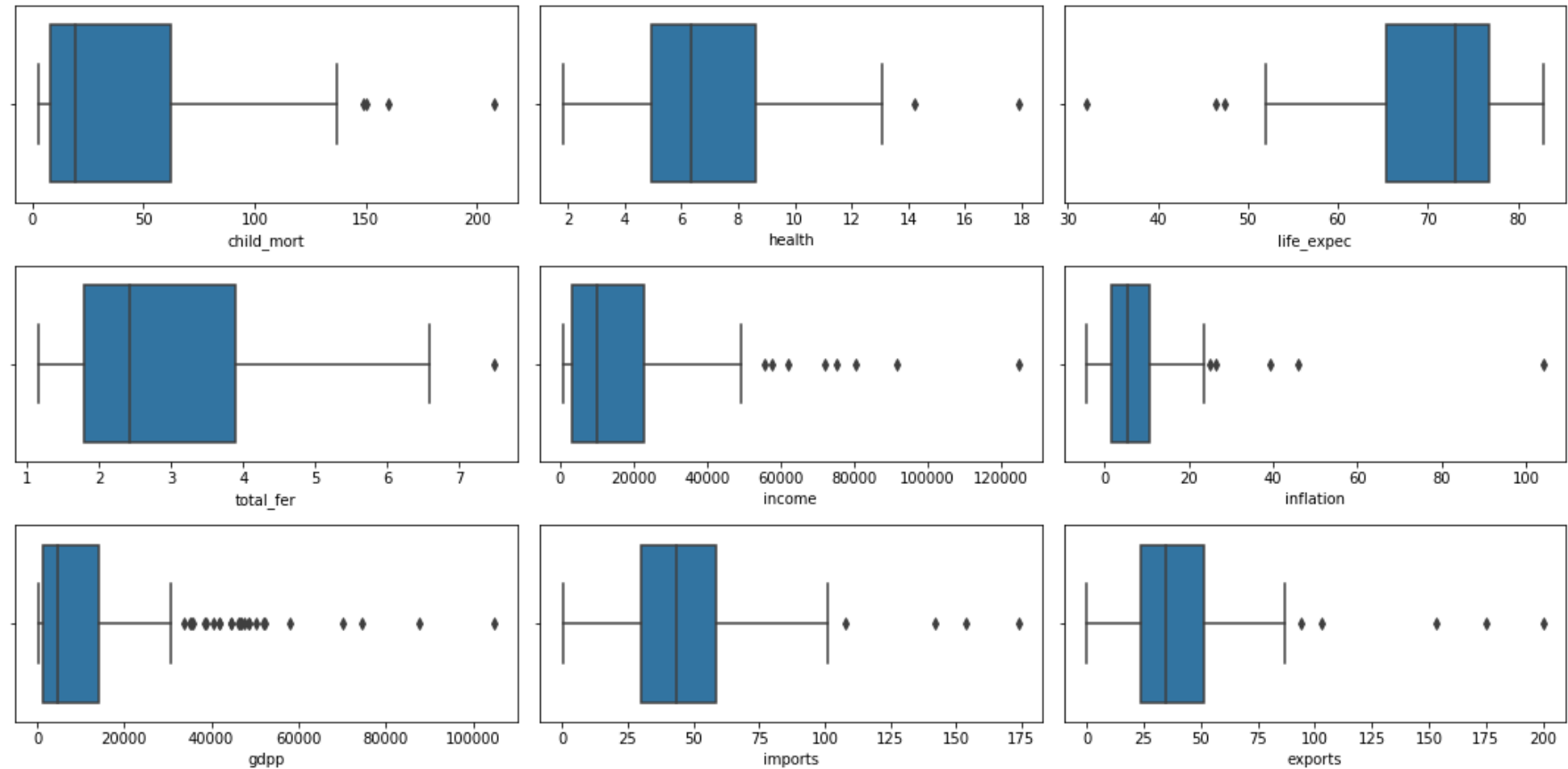
Heatmap:
Correlation
coefficients of
Country and see
how the feature
columns are
related



Inference ➡

- *Income and gdpp are positive correlation*
- *Imports and Exports are positive correlation*
- *Child_mort and total_fer are positive correlation*

Outlier: We have some outliers in data



- Income, imports, Exports and gdpp higher Outliers

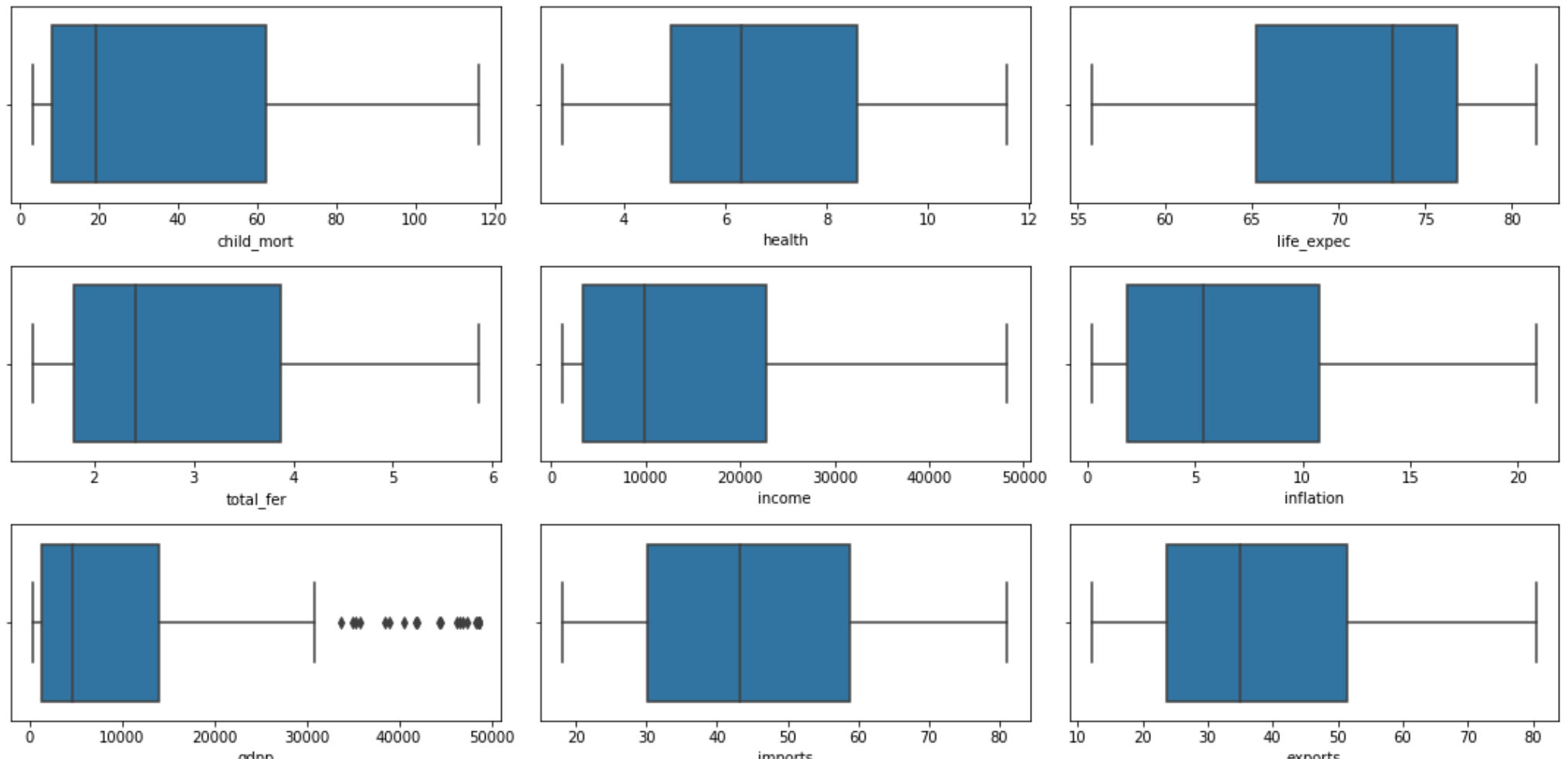
Inference ➡

- Here, will do Outlier treatment via capping the value so that will not loose the Country as every Country is important for analysis and lets do this and see the result in next slide

Outlier Treatment- We have cap the data so that we will not remove any country as country are important

Inference →

- As, we can see below the after outlier treatment outlier got removed and now we are ready to process the PCA and Clustering.

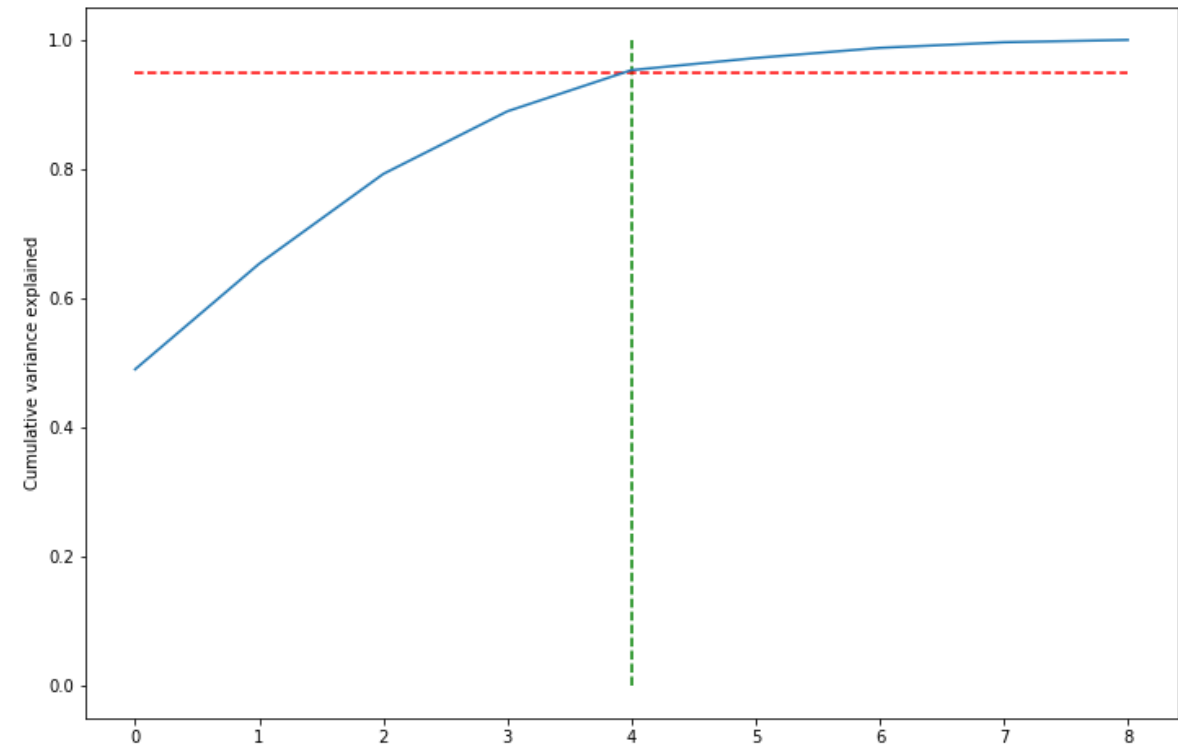


PCA (Principal Components Analysis)

Given a collection of points in two, three, or higher dimensional space, a "best fitting" line can be defined as one that minimizes the average squared distance from a point to the line. The next best-fitting line can be similarly chosen from directions perpendicular to the first. Repeating this process yields an orthogonal basis in which different individual dimensions of the data are uncorrelated. These basis vectors are called principal components, and several related procedures principal component analysis (PCA).

Inference ➡

- We have performed the PCA and got that 95% data is explained by 4 principle components so will perform the final PCA with 4 components



Hopkins Statistics Test

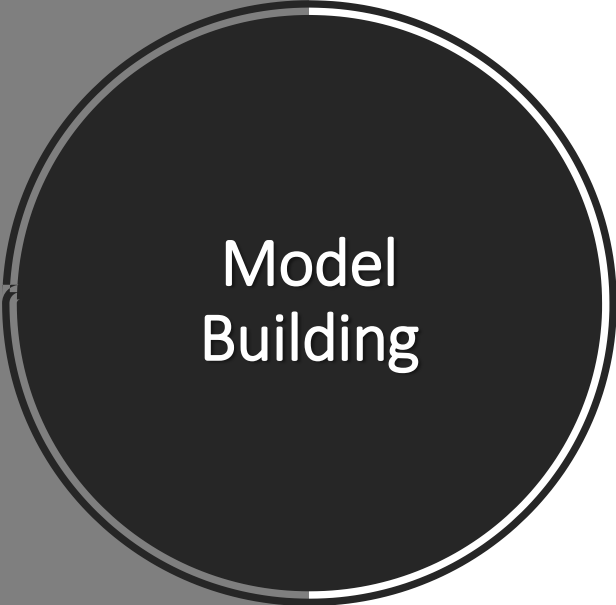
The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

- If the value is between $\{0.01, \dots, 0.3\}$, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between $\{0.7, \dots, 0.99\}$, it has a high tendency to cluster.

We got Hopkins Statistics value approx. .7 , which is good for applying the cluster

Inference ➡

- After performing the PCA and before Clustering need to check the Hopkins value to check clustering is applicable or not.
- Here, we got the 0.7 value which says that high tendency to cluster



Model Building

K- means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

- First we initialize k points, called means, randomly.
- We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
- We repeat the process for a given number of iterations and at the end, we have our clusters.

We decide the cluster on basis of:

1. *Silhouette Score*
2. *Elbow curve*

Silhouette Analysis

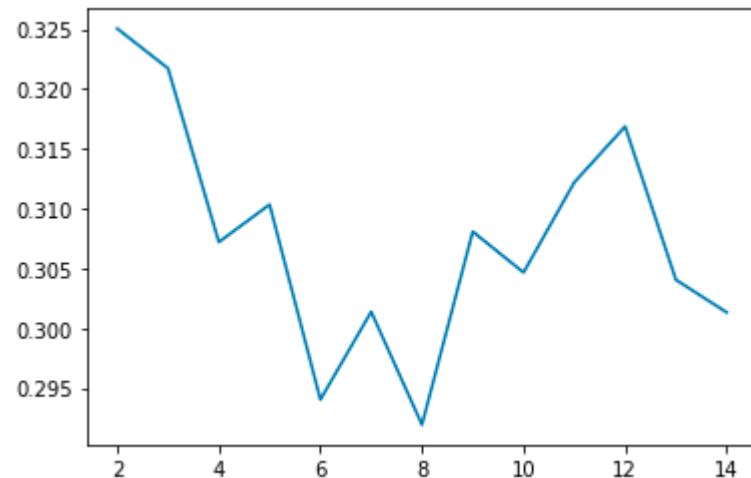
Silhouette Analysis

$$\text{silhouette score} = \frac{p - q}{\max(p, q)}$$

p is the mean distance to the points in the nearest cluster that the data point is not a part of

q is the mean intra-cluster distance to all the points in its own cluster.

- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.



```
For n_clusters=2, the silhouette score is 0.3250564127058999
For n_clusters=3, the silhouette score is 0.32172840937043035
For n_clusters=4, the silhouette score is 0.30719758172454154
For n_clusters=5, the silhouette score is 0.3118543434814181
For n_clusters=6, the silhouette score is 0.29647521740441957
For n_clusters=7, the silhouette score is 0.313714758575628
For n_clusters=8, the silhouette score is 0.304648962590413
```

Inference ➡

- Calculated Silhouette score for cluster and now we will go to Elbow Curve and see what curve is giving the result and then we will decide the cluster number

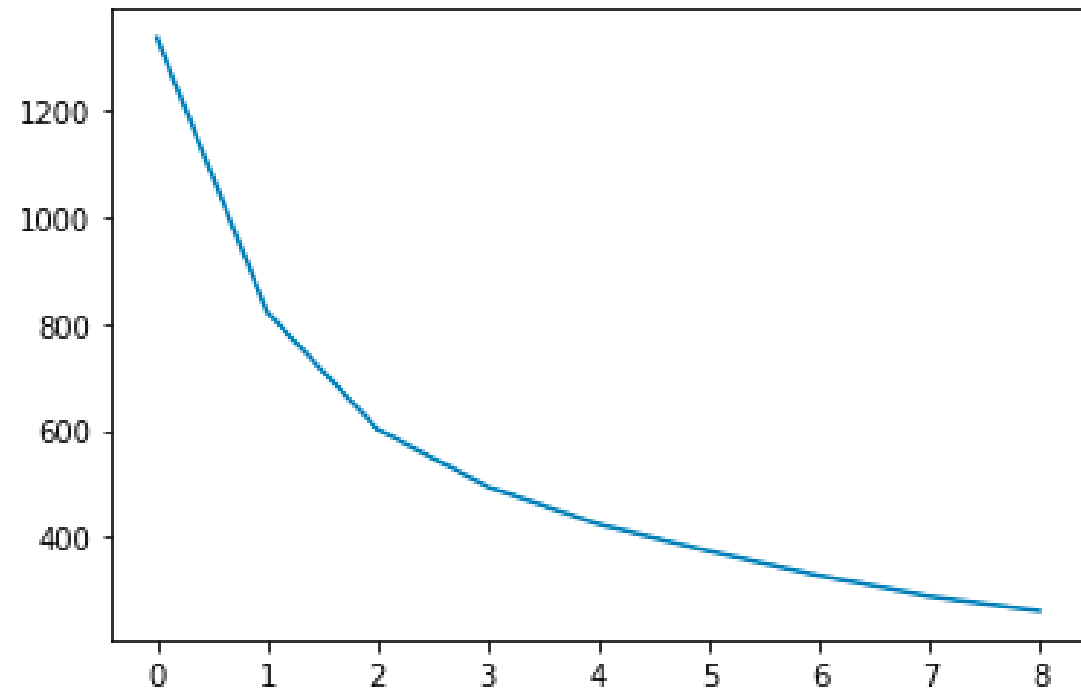
Sum of Squared Distances

Elbow Curve to get the right number of Clusters

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k .

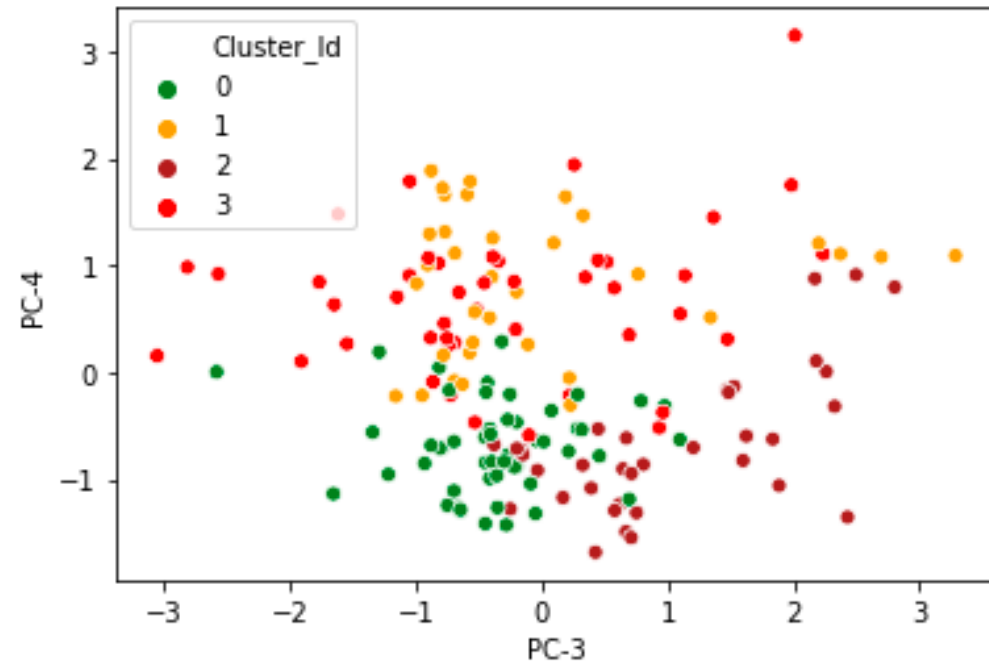
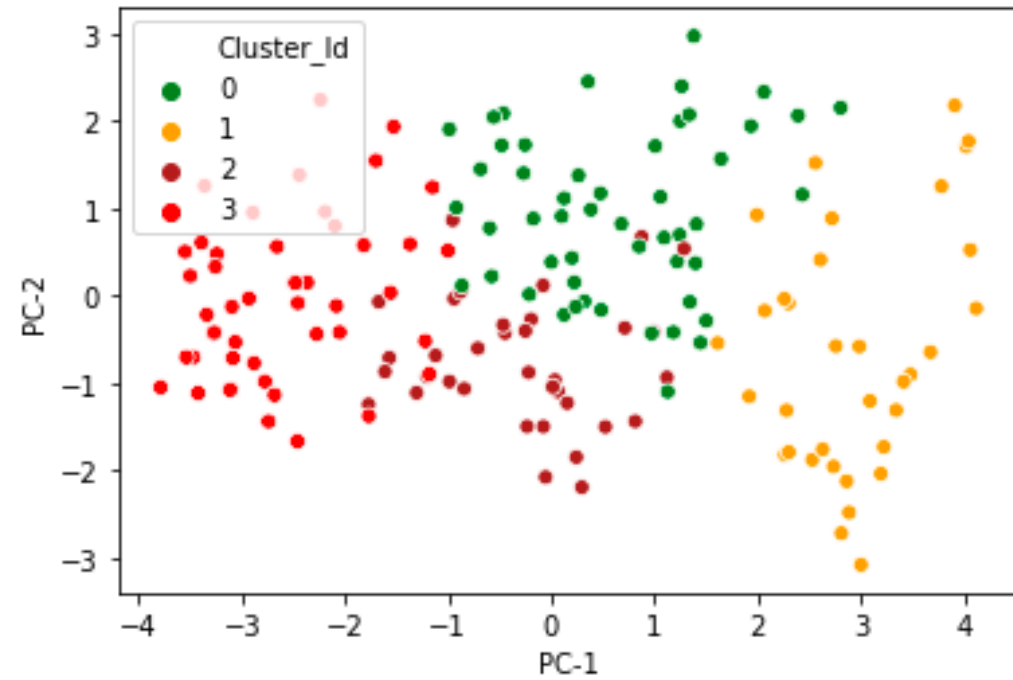
Inference ➡

- Here, we can see that at point 2 (cluster is 4) giving the Elbow curve for this plot.
- Previously score also giving the good result at cluster so we will go with 4 cluster



Scatter plot for PC1, PC2 & Cluster ID

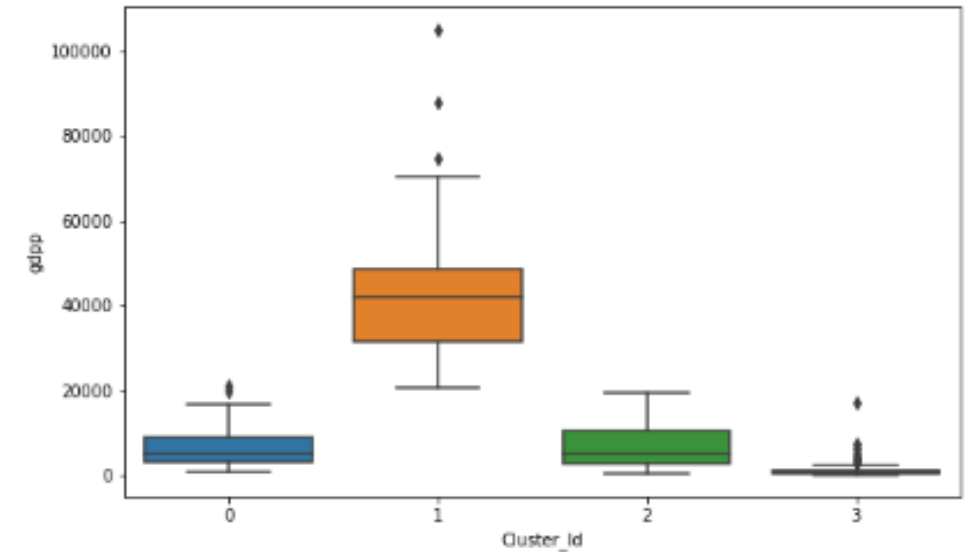
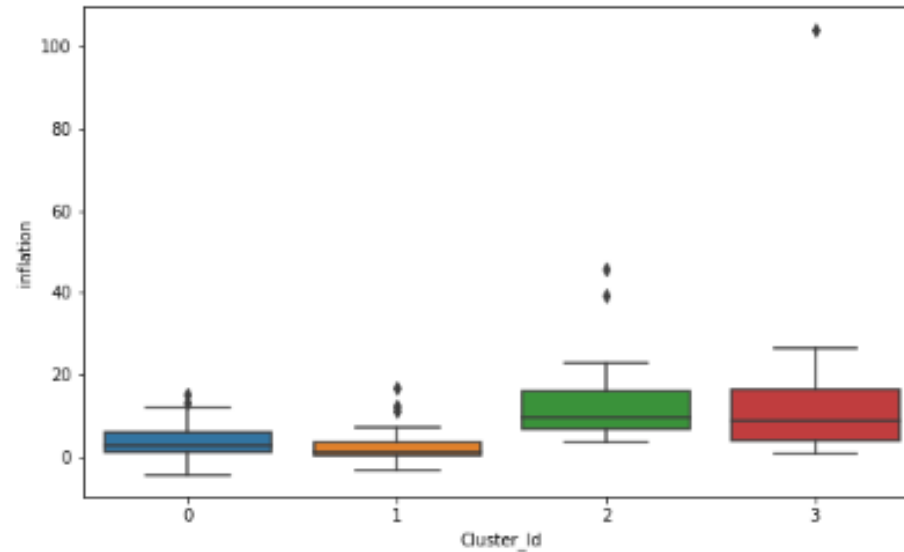
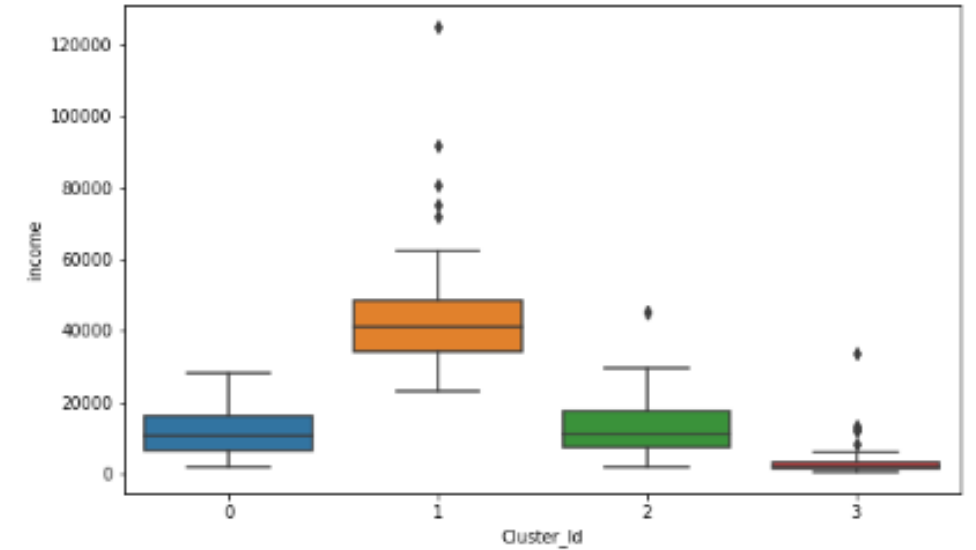
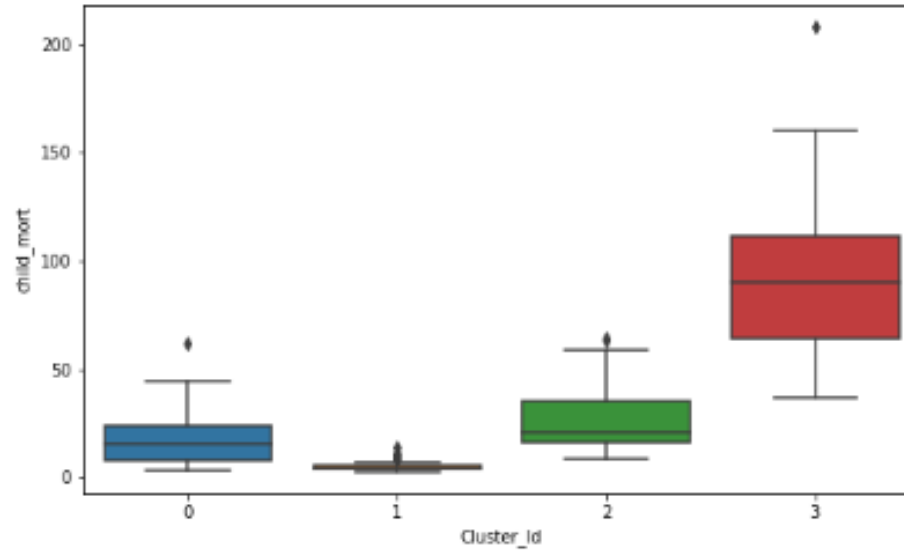
- To see the distribution of PC1, PC2, PC3, PC4 and Cluster ID

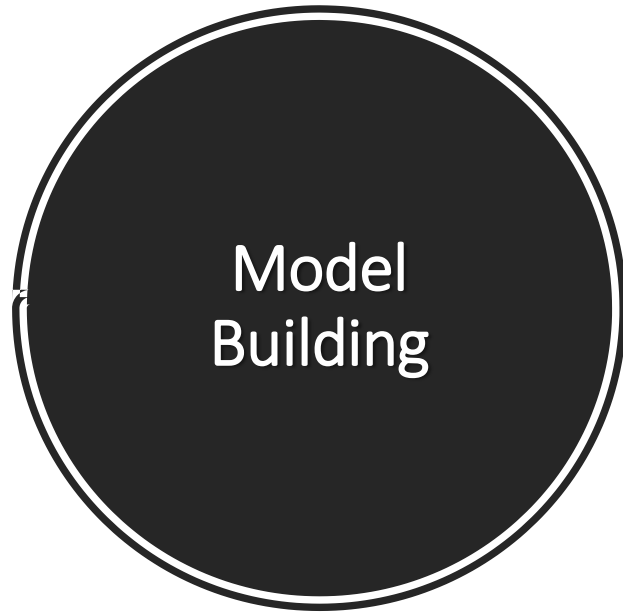


Box plot on
Original
attributes to
visualize the
spread of the
data

Inference

- Box plot of each cluster ID with respect to Child Mort, Income, gdpp and Inflation
- We can see here that Cluster ID 3 performing poor





Hierarchical Clustering:

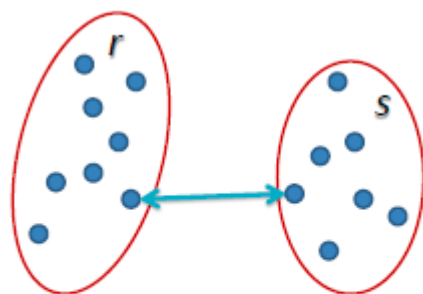
Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering

Hierarchical Clustering:

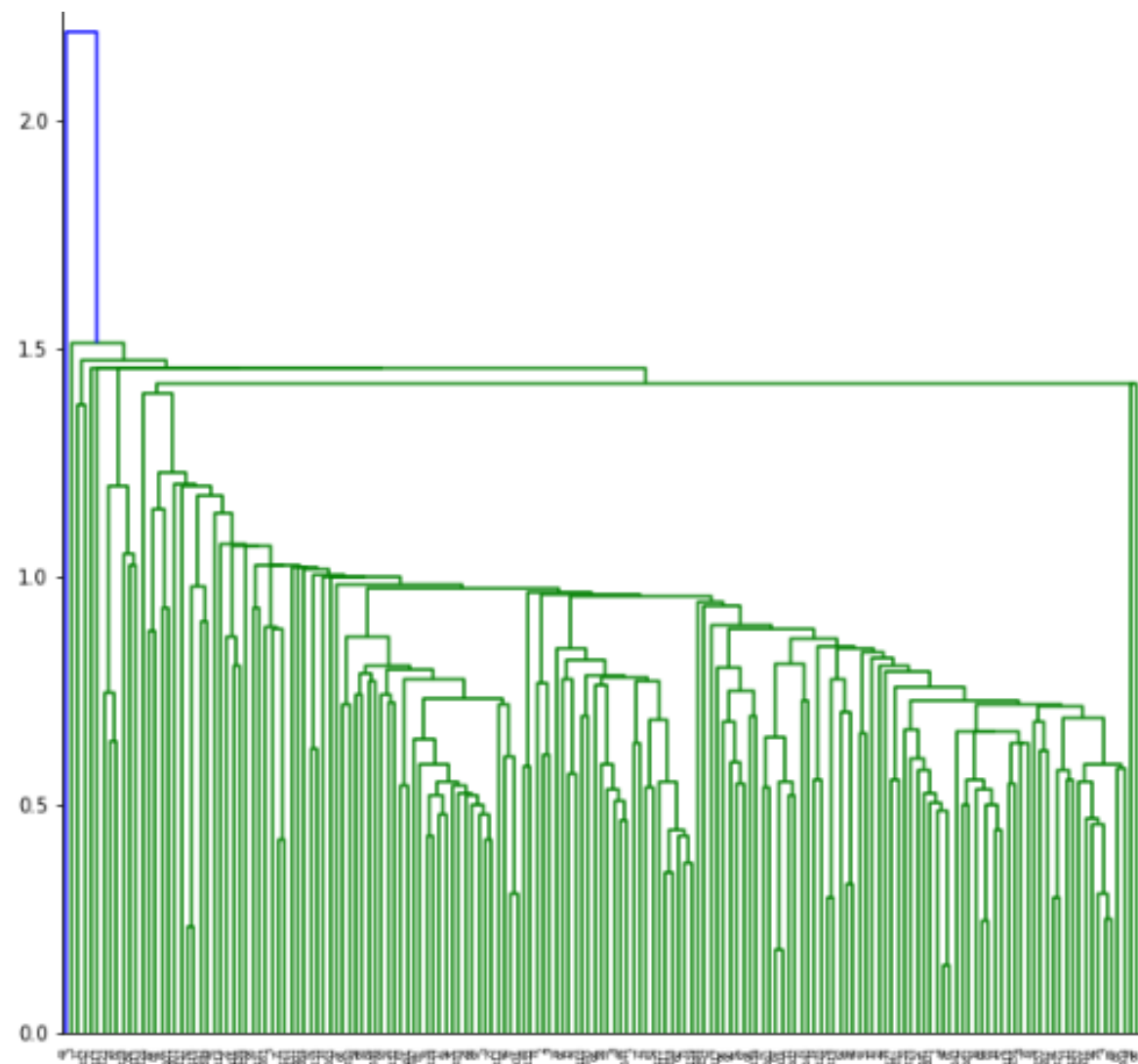
- ☐ **Single Linkage**
- ☐ **Complete Linkage**

Single Linkage:

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.

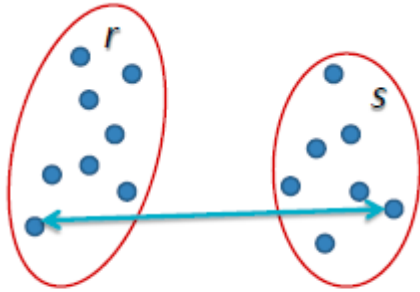


$$L(r, s) = \min(D(x_{ri}, x_{sj}))$$



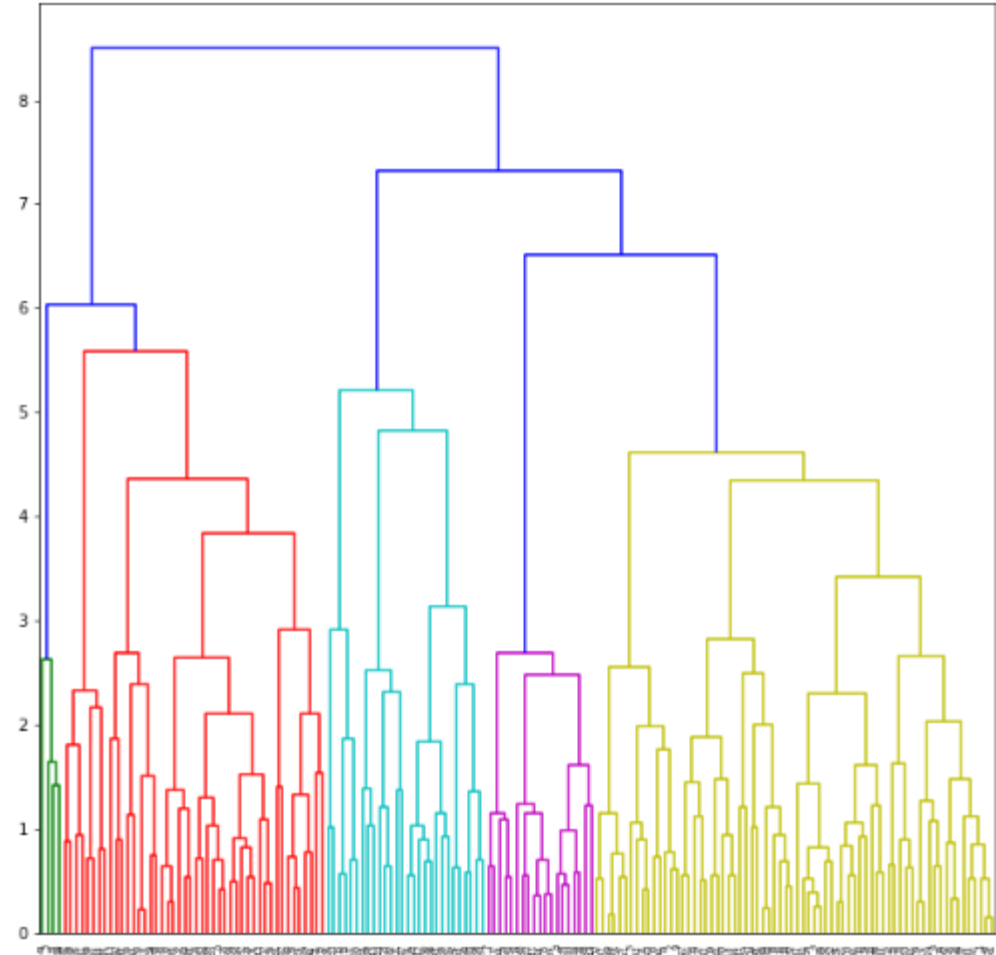
Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.

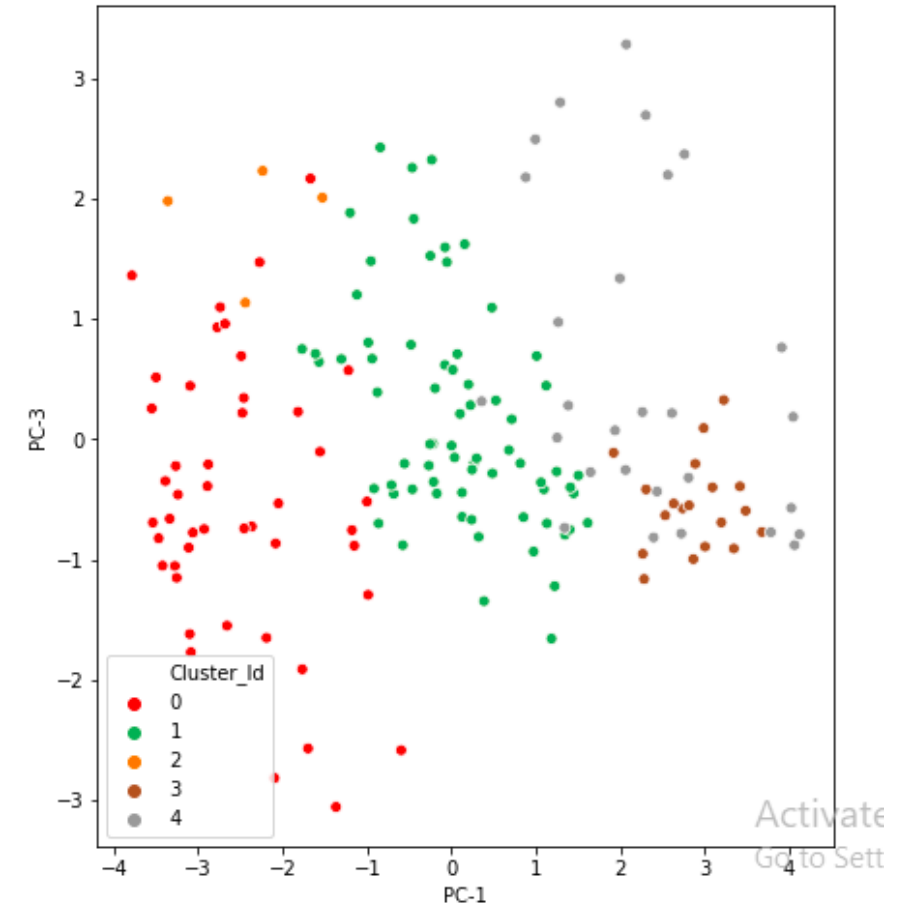
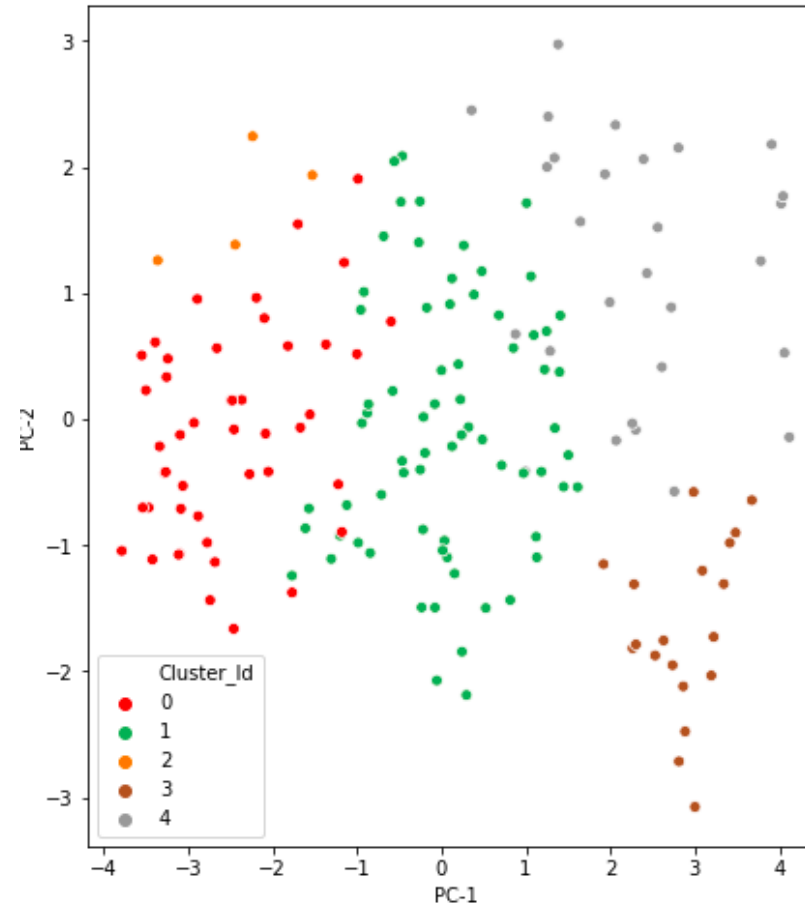


$$L(r, s) = \max(D(x_{ri}, x_{sj}))$$

- Complete linkage given better clustering as compare to Single.



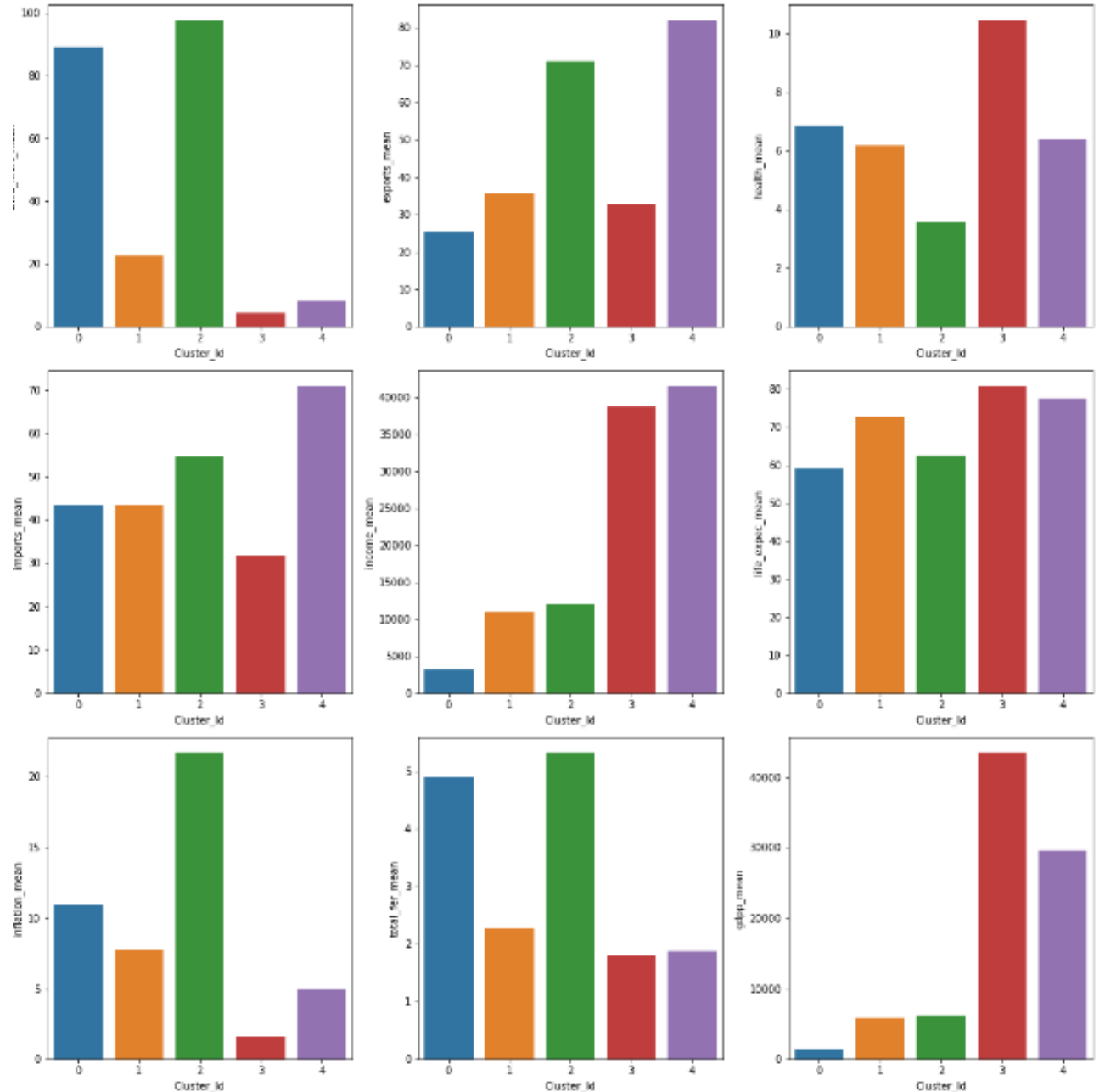
Scatter plot on
Principal
components to
visualize the
spread of the
data



Inference ➡

- Cluster 4 seems to be not properly formed, which lead to not get the good cluster formation in this clustering method.

Bar plot of the mean data for visualization



Inference →

- Here, Bar plot can show not getting significant result for the cluster as compared to K-Means clustering result.

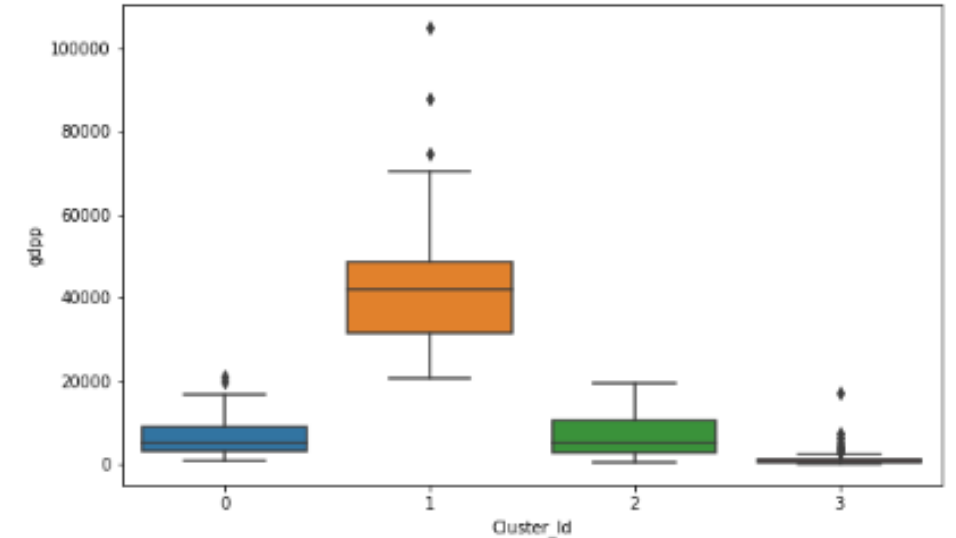
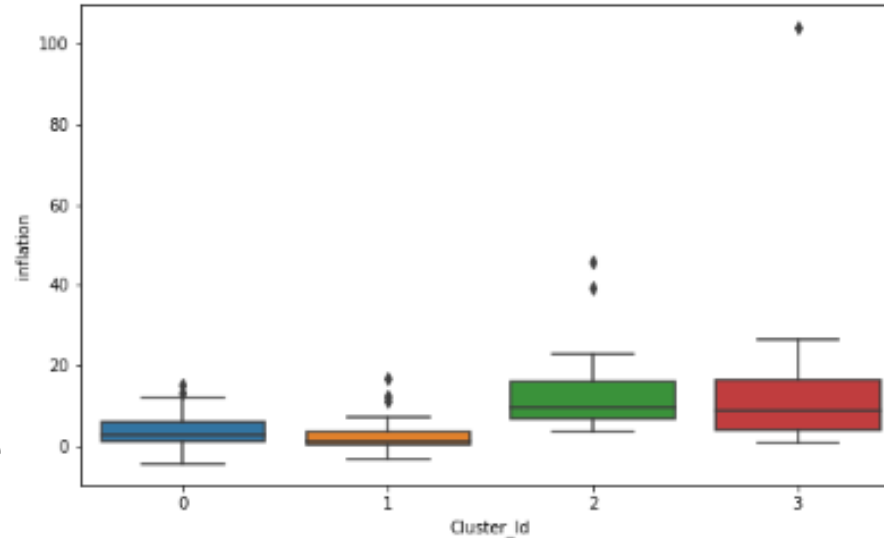
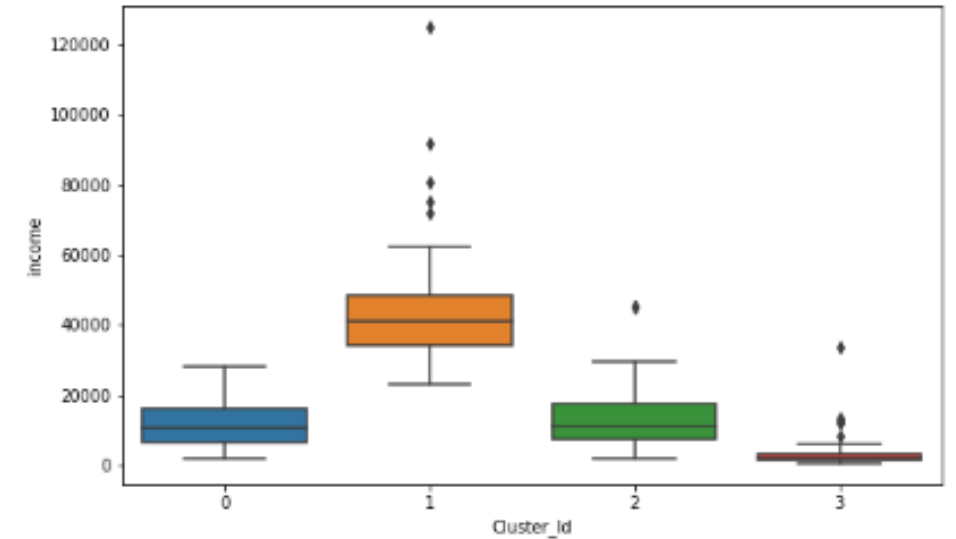
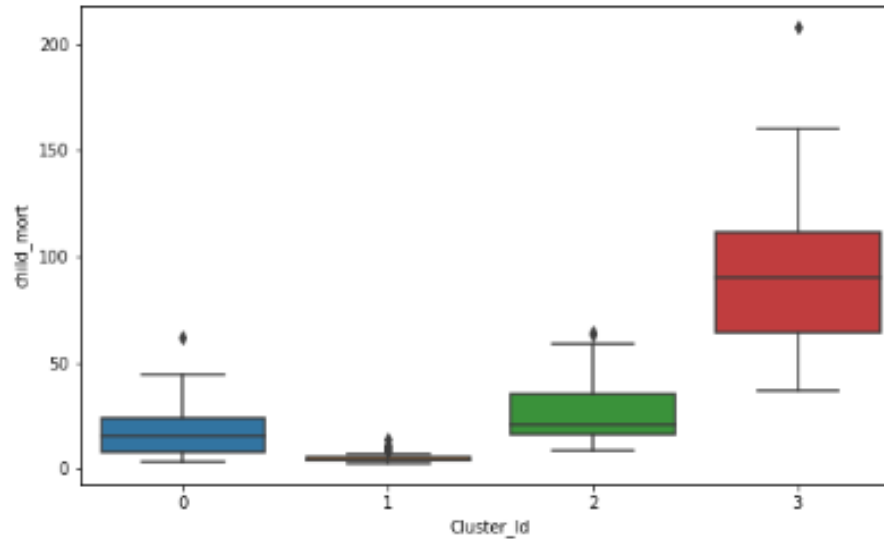


Model selection

We have analyzed both K-means and Hierarchical clustering and found clusters formed are not identical. The clusters formed in both the cases are not that great but its better in K-means as compared to Hierarchical. So, we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid.

K-Mean Clustering Result

- So, finally we have assign the cluster to Country to which they belong.
- Plotted some important feature in Box-plot and observe that Cluster 3 seem performing poor and need action.



- Finally we need to filter out the country which is performing poor in Cluster 3 on basis of Child Mort. Income, gdp and Inflation.

List of Country direst need of aid

Finally, after
applying PCA and
Clustering
technique got list
of Country which
are direst need of
aid.

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	Cluster_Id
25	Burkina Faso	116.0	19.20	29.6	6.74	1430	6.81	57.9	5.87	575	3
26	Burundi	93.6	8.92	39.2	11.60	764	12.30	57.7	6.26	231	3
31	Central African Republic	149.0	11.80	26.5	3.98	888	2.01	47.5	5.21	446	3
37	Congo, Dem. Rep.	116.0	41.10	49.6	7.91	609	20.80	57.5	6.54	334	3
63	Guinea	109.0	30.30	43.2	4.93	1190	16.10	58.0	5.34	648	3
64	Guinea-Bissau	114.0	14.90	35.2	8.50	1390	2.97	55.6	5.05	547	3
66	Haiti	208.0	15.30	64.7	6.91	1500	5.45	32.1	3.33	662	3
106	Mozambique	101.0	31.50	46.2	5.21	918	7.64	54.5	5.56	419	3
112	Niger	123.0	22.20	49.1	5.16	814	2.55	58.8	7.49	348	3
132	Sierra Leone	160.0	16.80	34.5	13.10	1220	17.20	55.0	5.20	399	3

**As per K- means clustering, the country
which are direst need of aid are:**

- Burkina Faso
- Burundi
- Central African Republic
- Congo, Dem. Rep.
- Guinea
- Guinea-Bissau
- Haiti
- Mozambique
- Niger
- Sierra Leone

**Looking at the graph we are certain that cluster 3 (i.e
this cluster above)is our cluster of concern. Because:**

- 1. It has highest child mortality**
- 2. Lowest income**
- 3. 2nd Highest Inflation**
- 4. Comparatively low life expectancy**
- 5. Highest total fertility**
- 6. Lowest gdpp**

Rohit Keshari

upGrad

Thank you