Clustering & PCA Assignment

Choosing Nations that are in direst need of aid

DATA EXPLORATION

This dataset has 2 files as explained below:

1. ‘country-data.csv' contains all the attributes of countries which helps in deriving the socio-economic and health factors that determine the overall development status of each country **.**

2. ‘data-dictionary.csv' is data dictionary which describes each attribute in country-data set.

PROBLEM STATEMENT

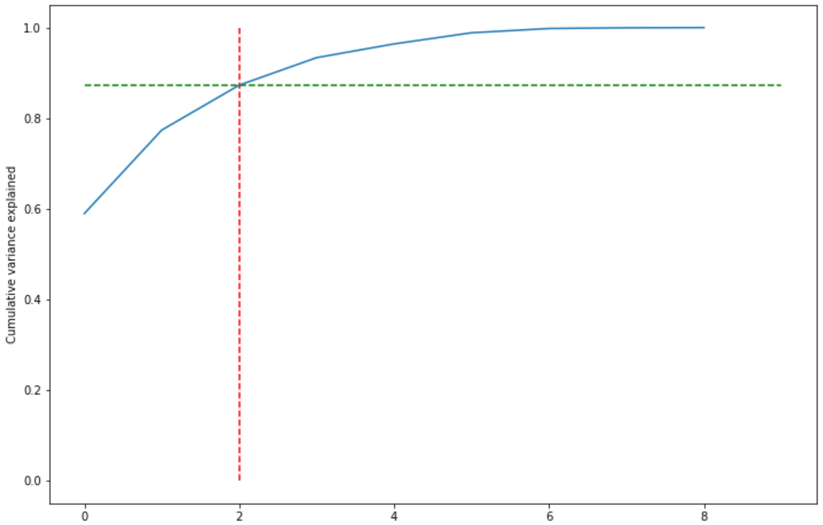
PROBLEM STATEMENT

- 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. 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 the CEO of the NGO needs 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.

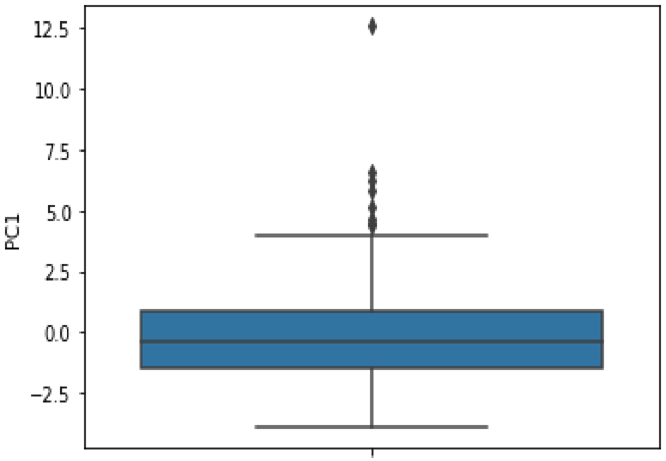
Analysis Approach

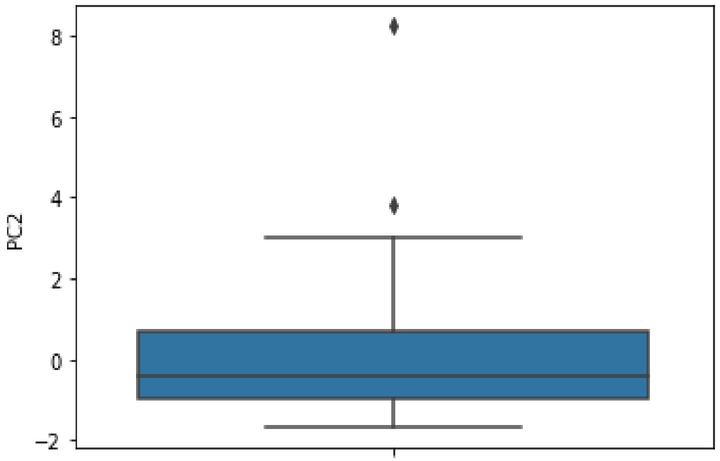
- Importing the Data (Country-data.csv) file.
- Understanding the Data set
  - Converting the columns Exports, health and imports variables which are given in % of GDP to Absolute values.
  - Check for any missing values in data set
  - Dropping the country column from data set before proceeding to Scaling step.
- Scaling the data set
  - Perform the standard scaling before proceeding for PCA.
  - Go ahead for PCA step by dropping the country column
- Dimensionality Reduction
  - Perform PCA step on the scaled dataset.
  - Get the cumulative sum of the explained variance ratio after performing the PCA step.
  - Plot Scree plot and it clearly tells us almost 90 % of cumm explained variance is covered by 3 principal components in the data set.

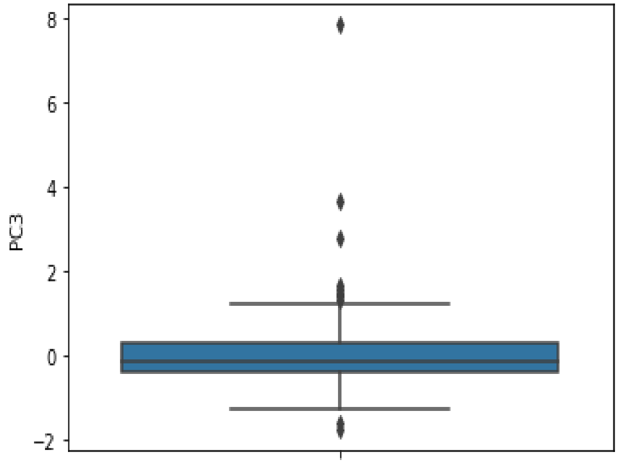


Analysis Approach

  - Perform the PCA with 3 Components
- Outlier Treatment
  - We have identified the outliers on all the 3 components





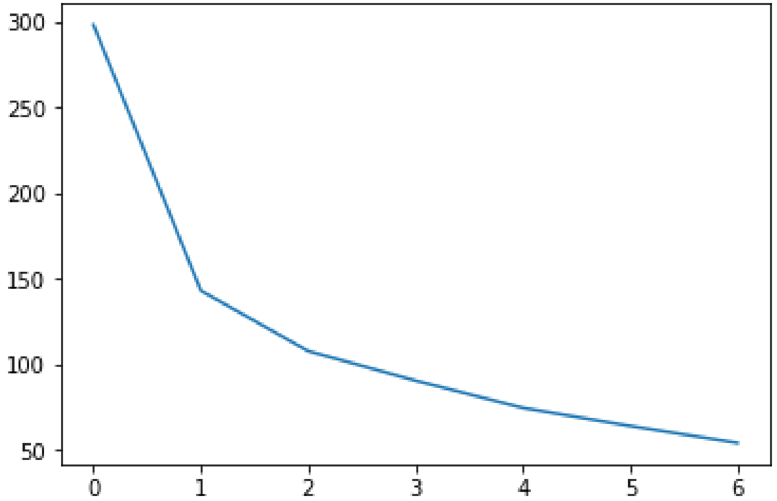


  - Removed the outliers on all the 3 components based on Statistical Technique

  - Proceed further to Clustering after removing outliers from the dataset.

Determine Optimal value for K in K-Means Clustering

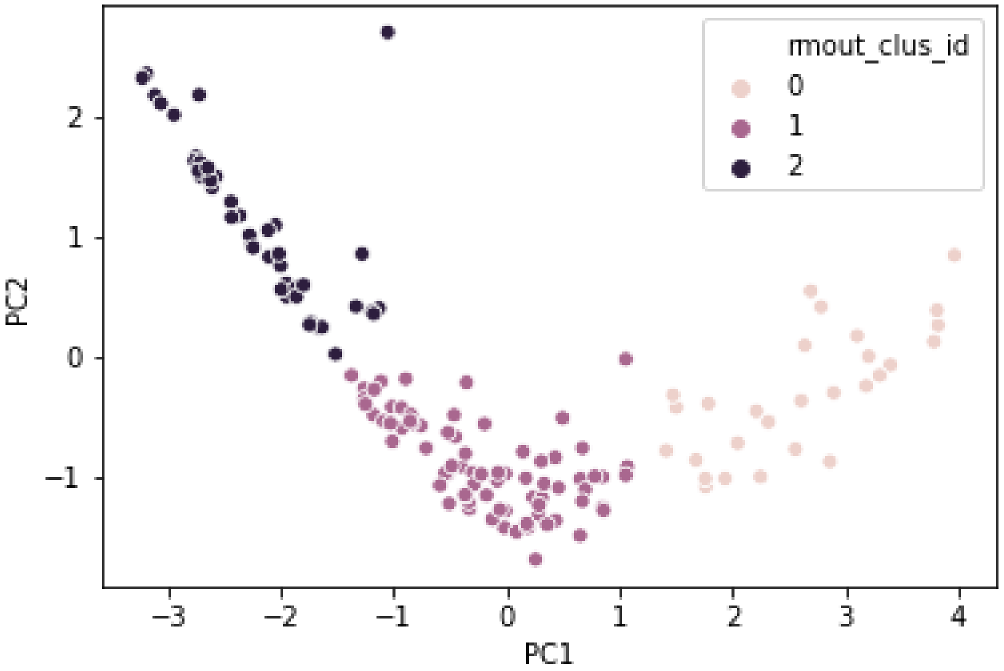
- Determine Optimal value of K for K-means Clustering
  - Check for Hopkins measure (0.81) which is good to perform clustering on Data set.
  - Plot elbow curve to determine the optimal value of K to perform the K-Means Clustering Algorithm



    - Plot clearly shows that K=3 is the optimal value
  - Lets check the silhouette analysis on different clusters.
    - For n_clusters=2, the silhouette score is 0.45946256207584163
    - For n_clusters=3, the silhouette score is 0.5171734664552139
    - For n_clusters=4, the silhouette score is 0.4374615104958549
    - It clearly shows that k=3 is the Optimal value

K-Means Clstering Approach

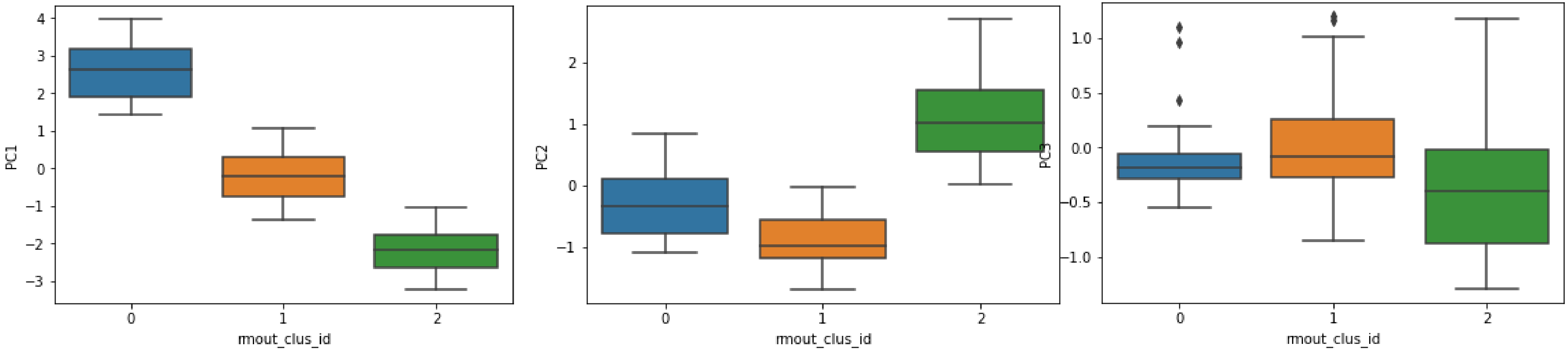
- Perform K-means Clustering with K=3 on the data set and check for results.



- Scatter Plot clearly shows the clear division of clusters (0,1,2) on PCA components.
- Lets see some more insights on data set to have a clear picture

K – Means Clustering Observations and Results

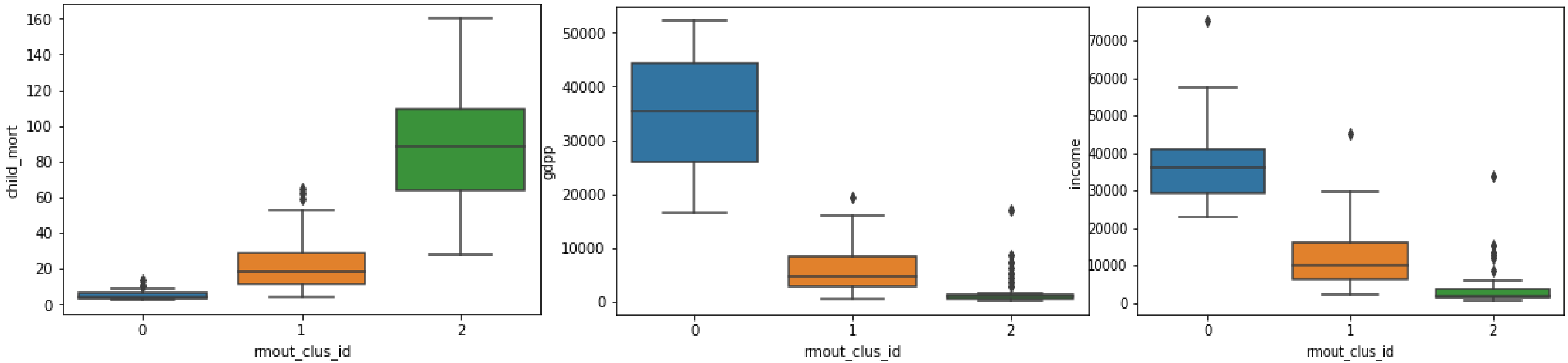
- Perform K-means Clustering with K=3 on the data set and check for results.



- Box plots clearly shows the variations how each Principal Components (PC1,PC2,PC3) varies against each cluster in the dataset (0,1,2)

K – Means Clustering Observations and Results

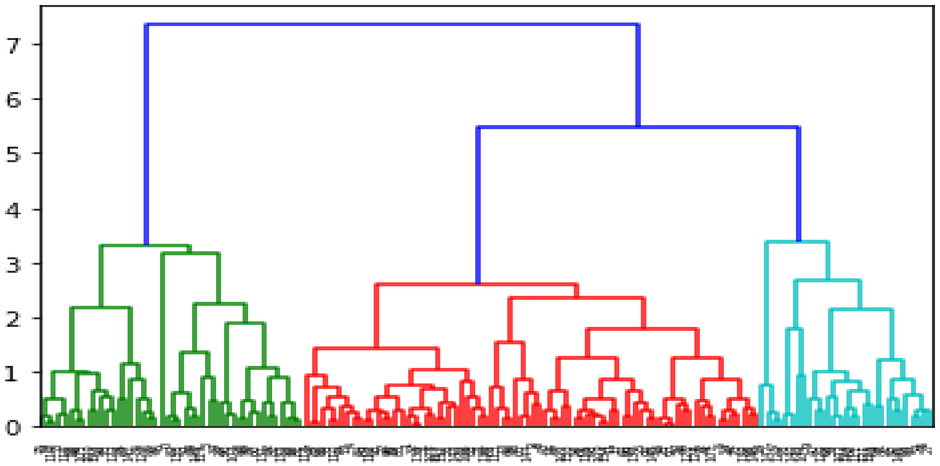
- Perform K-means Clustering with K=3 on the data set and check for results.

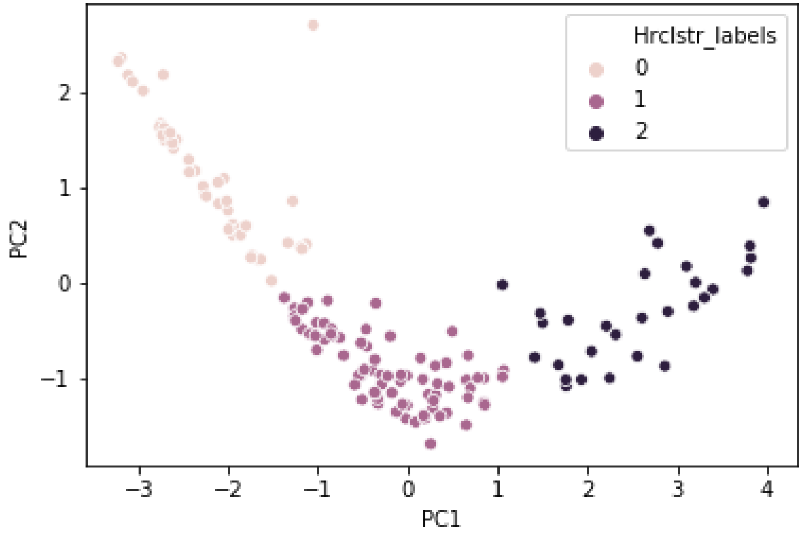


- Developed Countries:
- The countries under Cluster id ‘0’ is having low child_mort and High GDPP and High Income.
- Developing Countries:
- The countries under Cluster id ‘1’ is having medium child_mort,medium GDPP and medium Income.
- Under Developed Countries:
- The countries under Cluster id ‘2’ is having high child_mort, low GDPP and low Income.

Hierarchical Clustering Approach

- Perform Hierarchical Clustering with complete linkage and plot dendrogram.

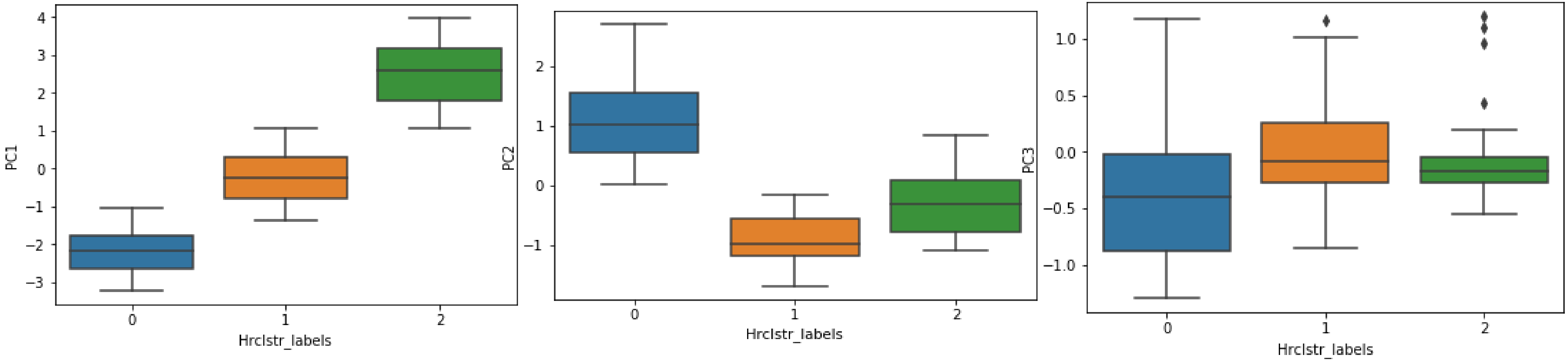




- Cut the tree with 3 clusters and assign the labels.
- Scatter Plot clearly shows the clear division of clusters (0,1,2) on PCA components.
- Lets see some more insights on data set to have a clear picture

Hierarchical Results & Observations

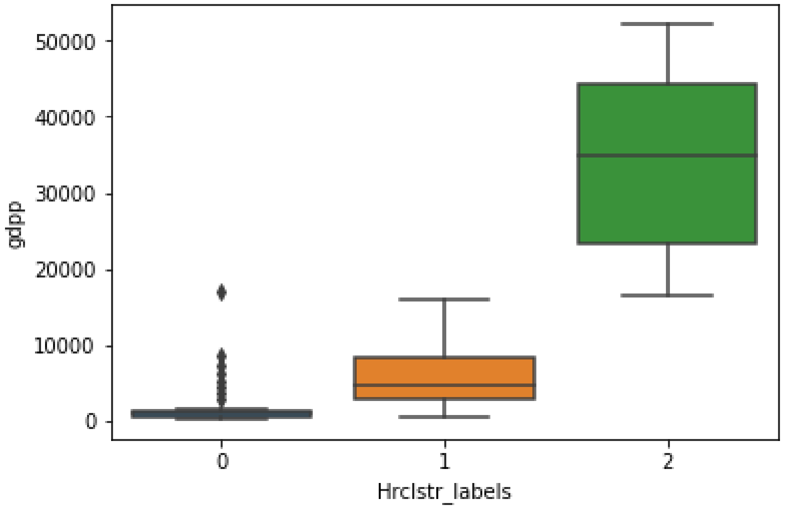
- Perform Hierarchical clustering with 3 clusters on the data set and check for results.

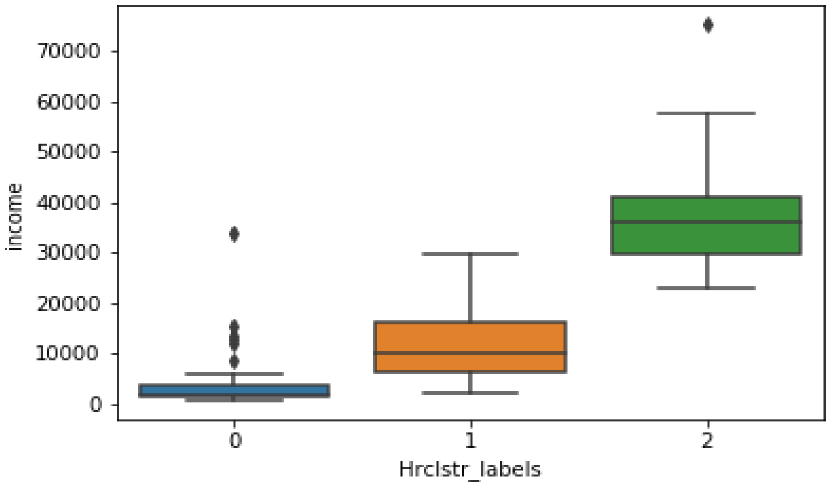


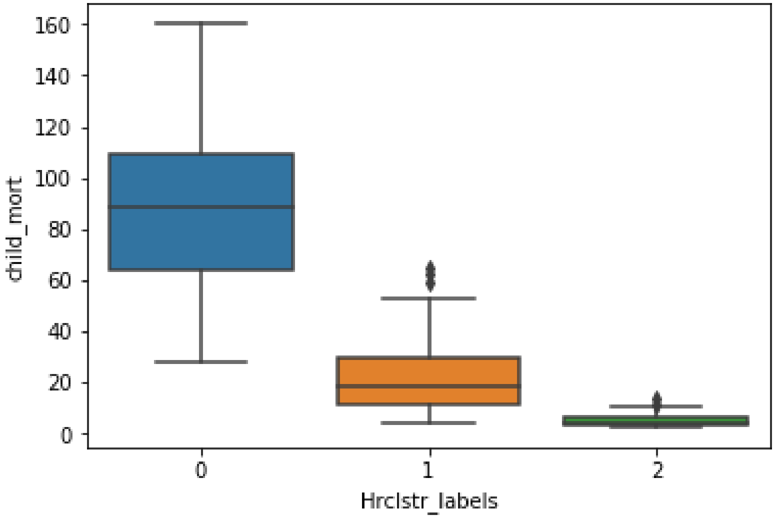
- Box plots clearly shows the variations how each Principal Components (PC1,PC2,PC3) varies against each cluster in the dataset (0,1,2)

Hierarchical Results & Observations

- Perform Hierarchical clustering with 3 clusters on the data set and check for results.



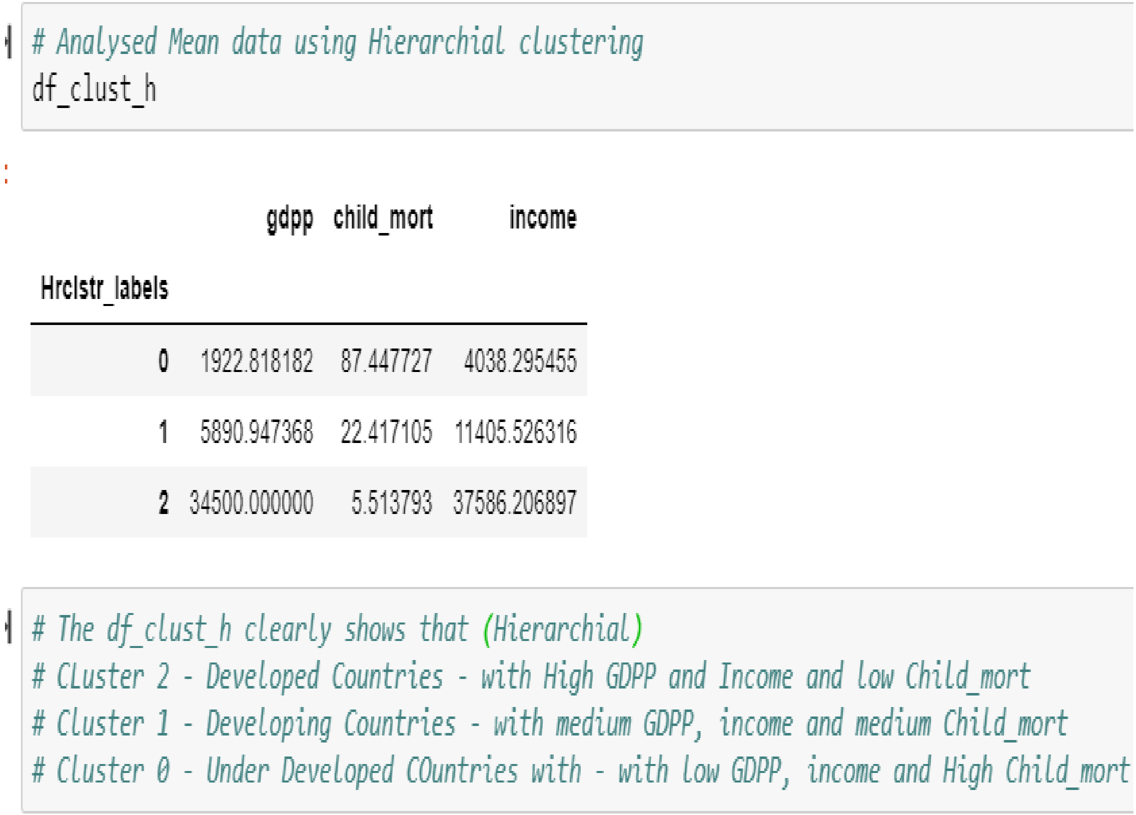




- Developed Countries:
- The countries under Cluster id ‘2’ is having low child_mort and High GDPP and High Income.
- Developing Countries:
- The countries under Cluster id ‘1’ is having medium child_mort,medium GDPP and medium Income.
- Under Developed Countries:
- The countries under Cluster id ‘0’ is having high child_mort, low GDPP and low Income.

Analysis and Observations

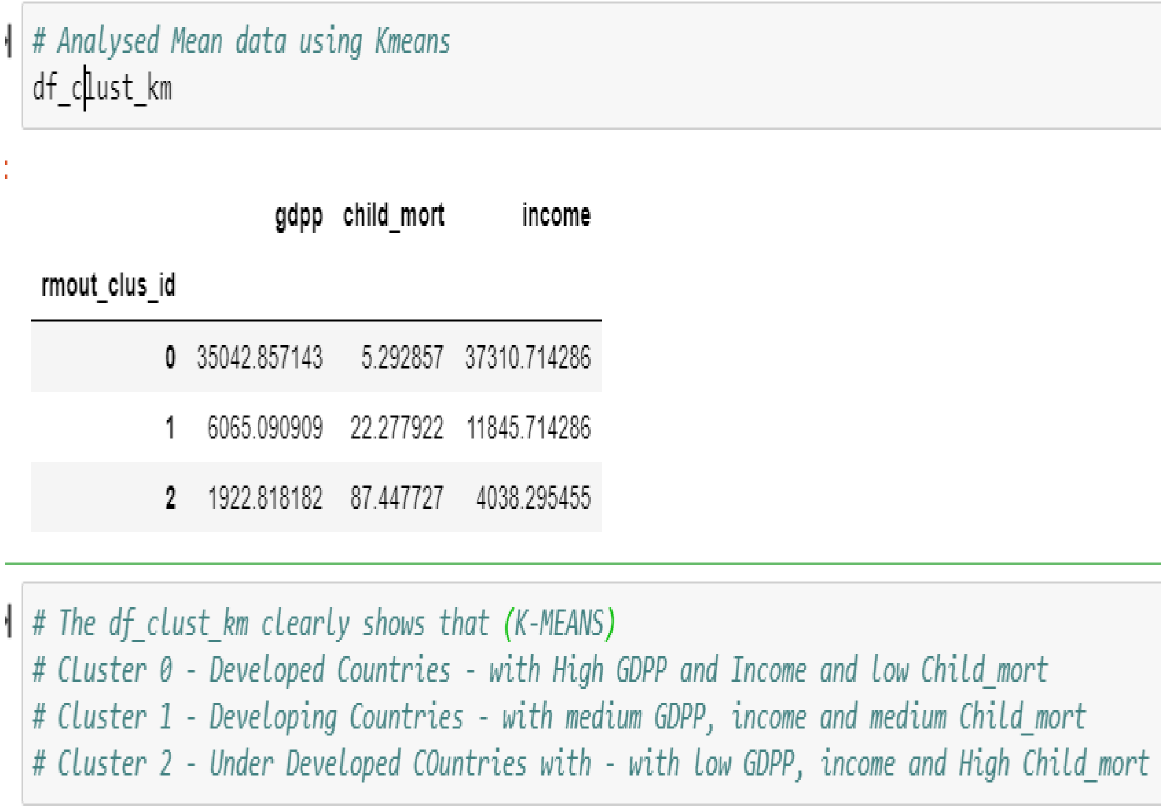
- Below Table shows the mean values of gdpp,child_mort and income after group by cluster_id using K-means and Hierarchal clustering.



```
# Analysed Mean data using Hierarchial clustering
df_clust_h
```

```
                    gdpp  child_mort        income
Hrclstr_labels
             0   1922.818182   87.447727   4038.295455
             1   5890.947368   22.417105  11405.526316
             2  34500.000000    5.513793  37586.206897
```

```
# The df_clust_h clearly shows that (Hierarchial)
# Cluster 2 - Developed Countries - with High GDPP and Income and low Child_mort
# Cluster 1 - Developing Countries - with medium GDPP, income and medium Child_mort
# Cluster 0 - Under Developed COuntries with - with low GDPP, income and High Child_mort
```



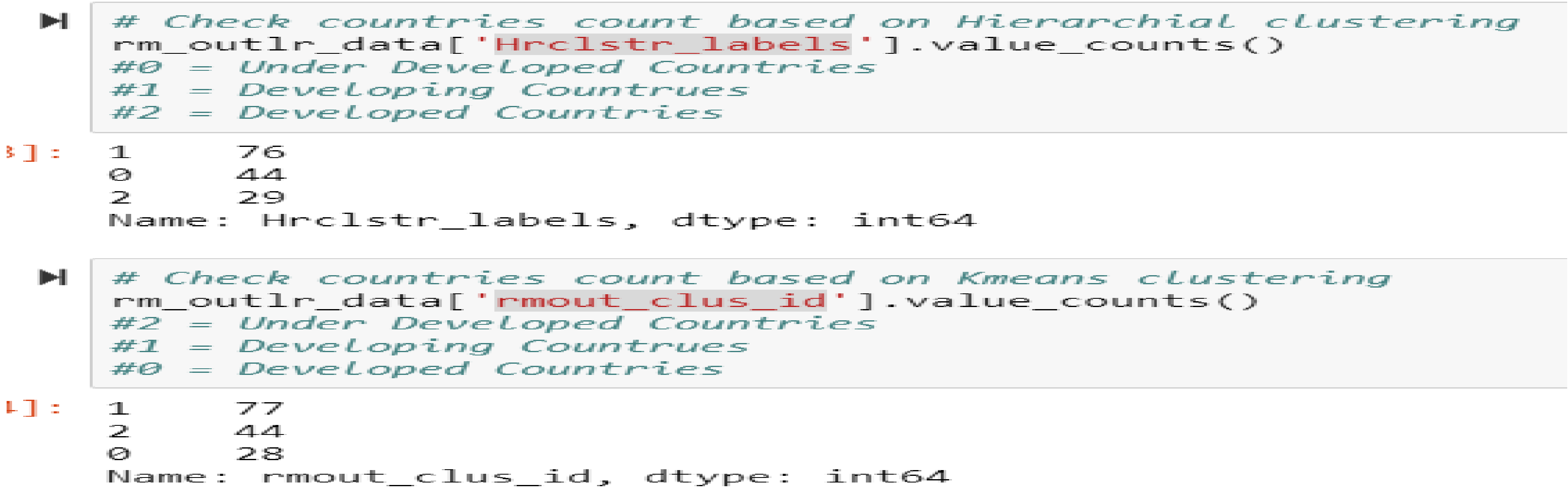
```
# Analysed Mean data using Kmeans
df_clust_km
```

```
                    gdpp  child_mort        income
rmout_clus_id
             0  35042.857143    5.292857  37310.714286
             1   6065.090909   22.277922  11845.714286
             2   1922.818182   87.447727   4038.295455
```

```
# The df_clust_km clearly shows that (K-MEANS)
# Cluster 0 - Developed Countries - with High GDPP and Income and low Child_mort
# Cluster 1 - Developing Countries - with medium GDPP, income and medium Child_mort
# Cluster 2 - Under Developed COuntries with - with low GDPP, income and High Child_mort
```

Analysis and Observations

- Below Table shows the countries count after group by cluster_id using K-means and Hierarchal clustering.



```
# Check countries count based on Hierarchial clustering
rm_outlr_data['Hrclstr_labels'].value_counts()
#0 = Under Developed Countries
#1 = Developing Countrues
#2 = Developed Countries
```

```
3]:  1      76
     0      44
     2      29
     Name: Hrclstr_labels, dtype: int64
```

```
# Check countries count based on Kmeans clustering
rm_outlr_data['rmout_clus_id'].value_counts()
#2 = Under Developed Countries
#1 = Developing Countrues
#0 = Developed Countries
```

```
4]:  1      77
     2      44
     0      28
     Name: rmout_clus_id, dtype: int64
```

- I am going with K-means as both the clustering algorithms showing the same count for the under developed countries which are in direst aid which is 44.

Analysis and Observations

- The countries which are in dire need of aid based on K- Means clustering are.
- 'Afghanistan',
- 'Angola',
- 'Benin',
- 'Botswana',
- 'Burkina Faso',
- 'Burundi',
- 'Cameroon',
- 'Chad',
- 'Comoros',
- 'Congo, Dem. Rep.',
- 'Congo, Rep.',
- "Cote d'Ivoire",
- 'Equatorial Guinea',
- 'Eritrea',
- 'Gabon',
- 'Gambia',
- 'Ghana',
- 'Guinea',
- 'Guinea-Bissau',
- 'Iraq',
- 'Kenya',
- 'Kiribati',
- 'Lao',
- 'Lesotho',
- 'Liberia',
- 'Madagascar',
- 'Malawi',
- 'Mali',
- 'Mauritania',
- 'Mozambique',
- 'Namibia',
- 'Niger',
- 'Pakistan',
- 'Rwanda',
- 'Senegal',
- 'Sierra Leone',
- 'Solomon Islands',
- 'South Africa',
- 'Sudan',
- 'Tanzania',
- 'Togo',
- 'Uganda',
- 'Yemen',
- 'Zambia'

Thank You

Prepared by

- Dayanand Sagar Kukkala