CLUSTERING & PCA ASSIGNMENT

DHARINI P

MAY 2019 COHORT

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.

GOAL

- To cluster the countries using some socio-economic and health factors that determine the overall development of the country
- To give suggestion to CEO about the countries which are in dire need of aid

APPROACH

DATA UNDERSTANDING, CLEANING & EDA:

- Identify the data quality issues and clean the data for analysis
- Datatype inspection of all the variables in the data
- Outlier analysis
- Exploratory data analysis

APPROACH

Scaling & PCA:

- Scaling of the numeric variables
- Principle component analysis to reduce dimensions and avoid multi collinearity by preventing information loss
- Choosing the number of PC which explains maximum variance

APPROACH

MODELLING on PCA modified dataset:

- Hopkins statistic – to check whether the dataset is suitable for clustering
- K-means
  - Determine optimal no of clusters based on elbow method, silhouette and business case
  - Perform k-means with selected no of cluster
  - Cluster analysis

APPROACH

MODELLING:

- Hierarchical
  - Create dendrogram using single & complete linkage
  - Find the optimal no of clusters from dendrogram
  - Perform k-means with selected no of cluster
  - Cluster analysis
- Compare k means and hierarchical and choose one to report final list of countries

APPROACH

RECOMMENDATIONS:

- Analyze the clusters, make visualisations
- Find the cluster which needs aid based on socio economic factors
- Report the countries from that cluster which are in dire need of help which can benefit from the funding

DATA UNDERSTANDING & CLEANING

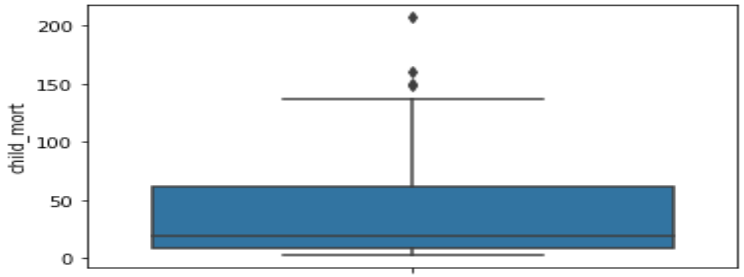
- The dataset contains details of countries and their socio economic and heath factors
- Dimension of dataset is 167*10 – Details of 167 countries & their 9 socio economic factors like gdp, income per person, child mortality rate etc..
- NO MISSING VALUES in the data
- NO DUPLICATE ENTRIES as well
- **Outliers are present which needs to be treated**
- NO DATA QUALITY ISSUES OTHER THAN OUTLIERS FOUND IN THE DATA
- New metrics are not derived as i feel the given variables are sufficient for analysis

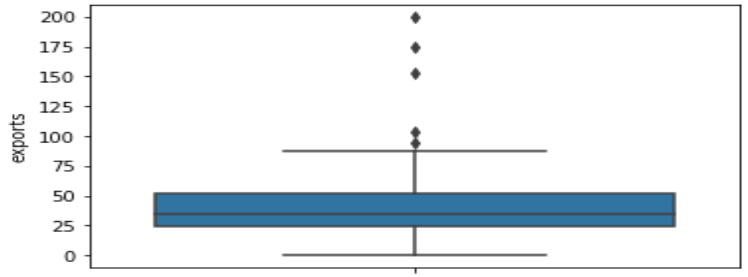
DATA UNDERSTANDING & CLEANING

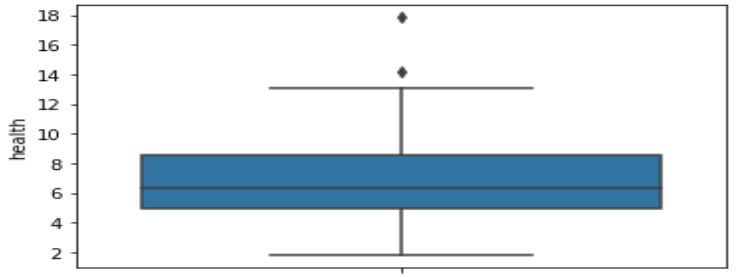
- **CATEGORISATION OF VARIABLES:**
  - Country variable is of object datatype, other variables are of type int & float
  - Datatype of variable is not changed as we feel its not necessary
  - Column conversion
- **COLUMN CONVERSION:**
  - Variables like exports, imports, health & inflation are given as percentage of GDP
  - I choose not to convert them to actual values as we are more interested in clustering.
  - Hence converting or not - is not going to affect the clustering

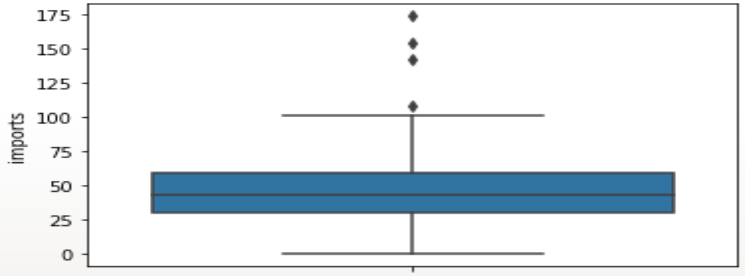
EDA

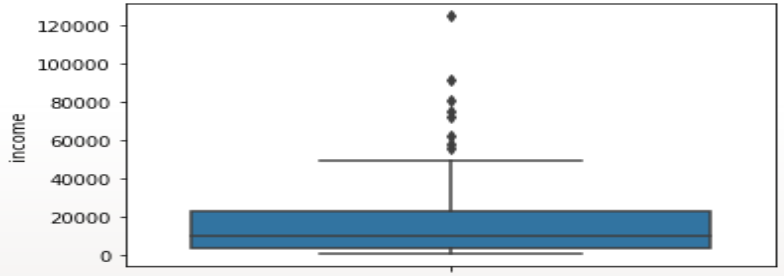
- DISTRIBUTION OF EACH VARIABLE IN THE DATASET

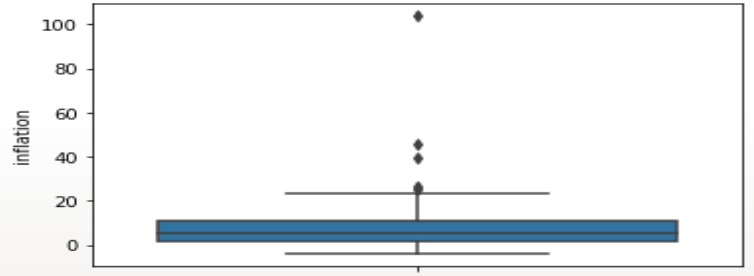


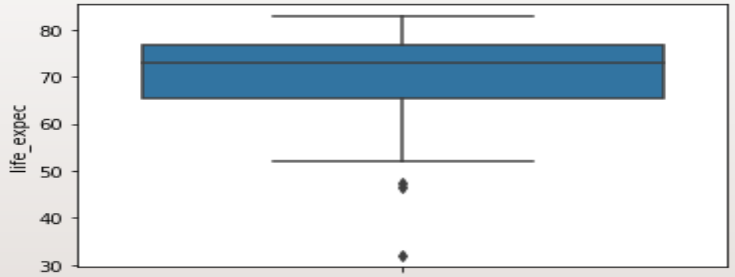


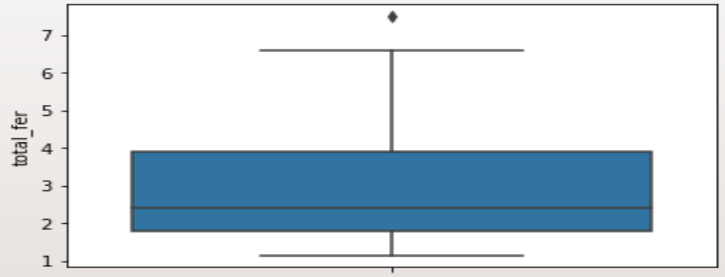


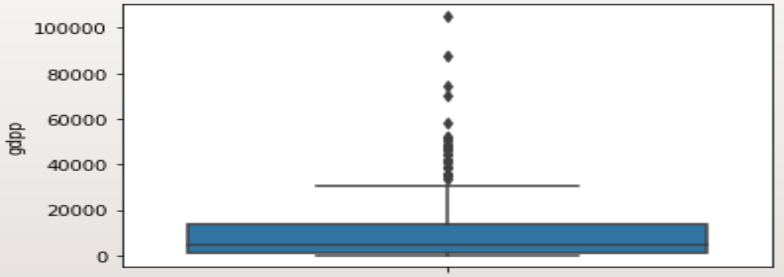










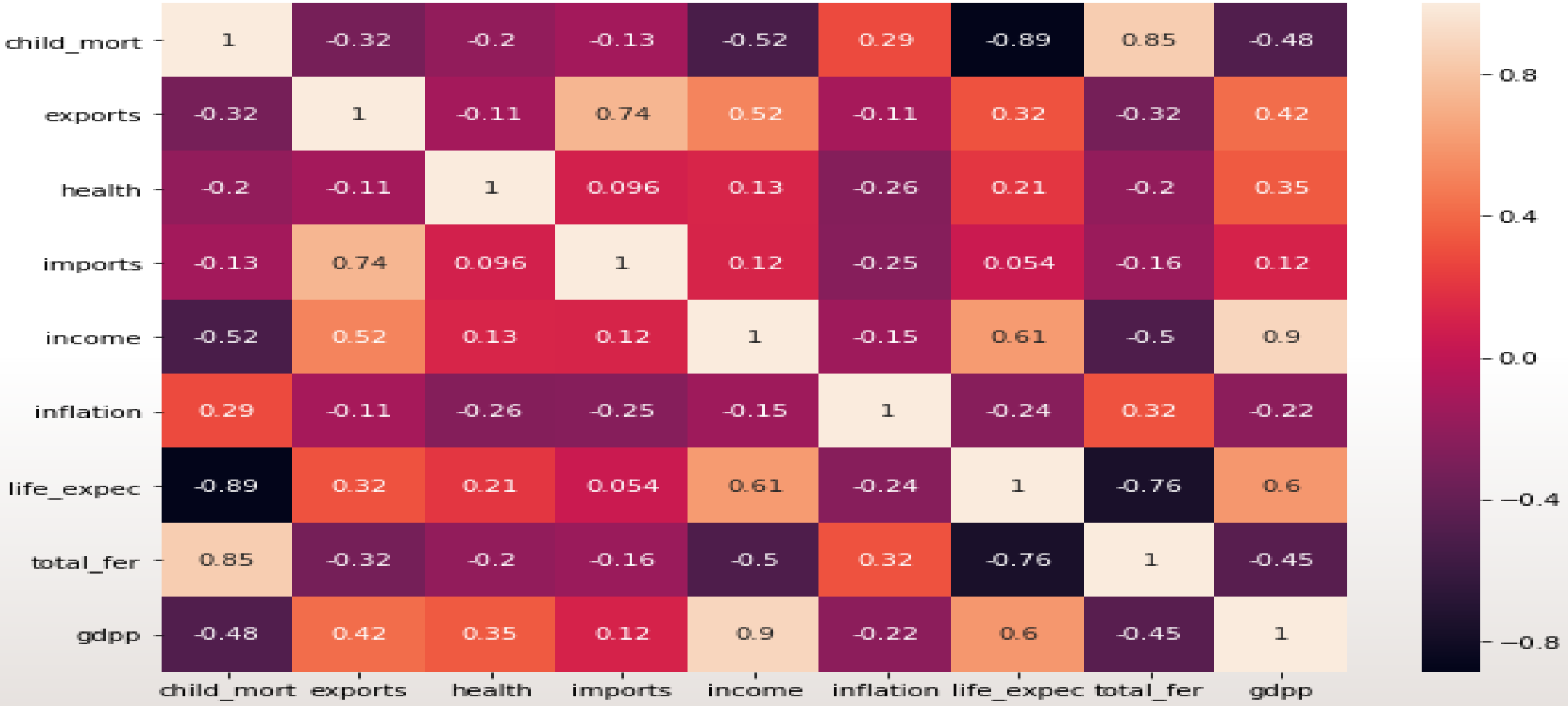


EDA

- From the variable distributions, we can infer that:
  - gdpp – gdp of most of the countries are on the lower end. Spread is minimal. Some countries have high gdp, which are beyond whiskers
  - Income – Income is averagely spread, but the median is very low.
  - Child_mort – child mortality rate is high for some countries. Rest shows mortality on lower side
- Also, its evident from the box plots that **OUTLIERS ARE PRESENT IN ALL VARIABLES IN THE DATA**
- **We will be treating the outliers BEFORE PCA**

EDA

CORRELATION BETWEEN VARIABLES



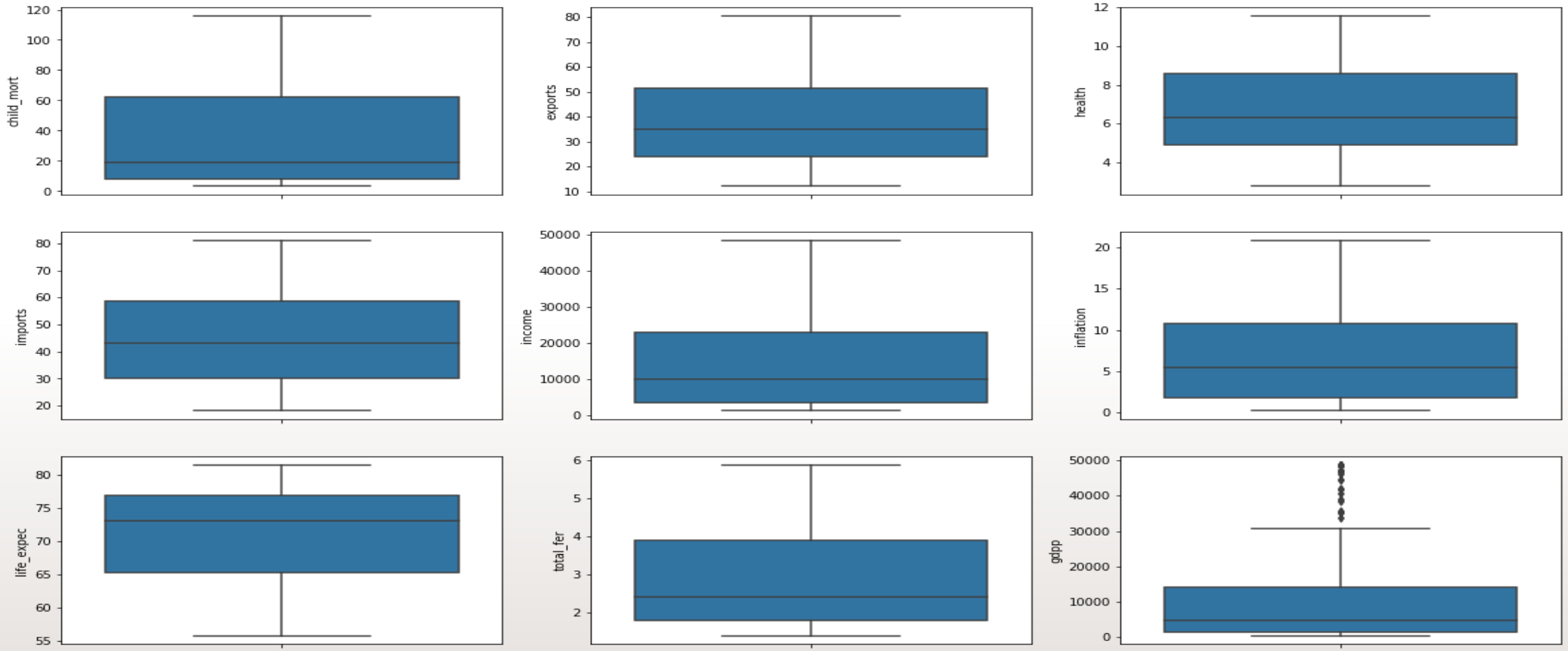
EDA
CORRELATION BETWEEN VARIABLES

- From the heatmap, its evident that correlation exists between the variables
- Income has high positive correlation with gdpp and life_expec
- Exports has high positive correlation with imports
- child_mort has high positive correlation with total_fer
- High negative correlation exists between chid_mort & life_expec, life_expec & total_fer
- Correlation between variables will be treated using PCA

OUTLIER TREATMENT

- Removing outliers will lead to data loss and countries in need of help may be removed in the process
- Instead of removing,
  - we can do floor and capping to treat outliers.
  - Taking 95% and 5% percentile as cap and floor for each column
- After capping & flooring, most of the outliers are removed.
- **Box plot after outlier treatment in next slide**
- GDPP alone has datapoints above whiskers which indicates the presence of countries with high gdp

OUTLIER TREATMENT



SCALING

- Scaling is done to all variables except country to bring it to common scale
- Prevents skewed model
- Opted for **standardisation** method
- Standardisation - *rescales the data to have a mean of zero and a standard deviation of one using (x-mu)/sigma*
  - mu – mean & sigma – std deviation

PCA

- PCA is Principle Components Analysis
- Reduces the dimensionality and multicollinearity (if any)
- Captures maximum variance without information loss
- Fitted on scaled country dataset to check for PC & variance

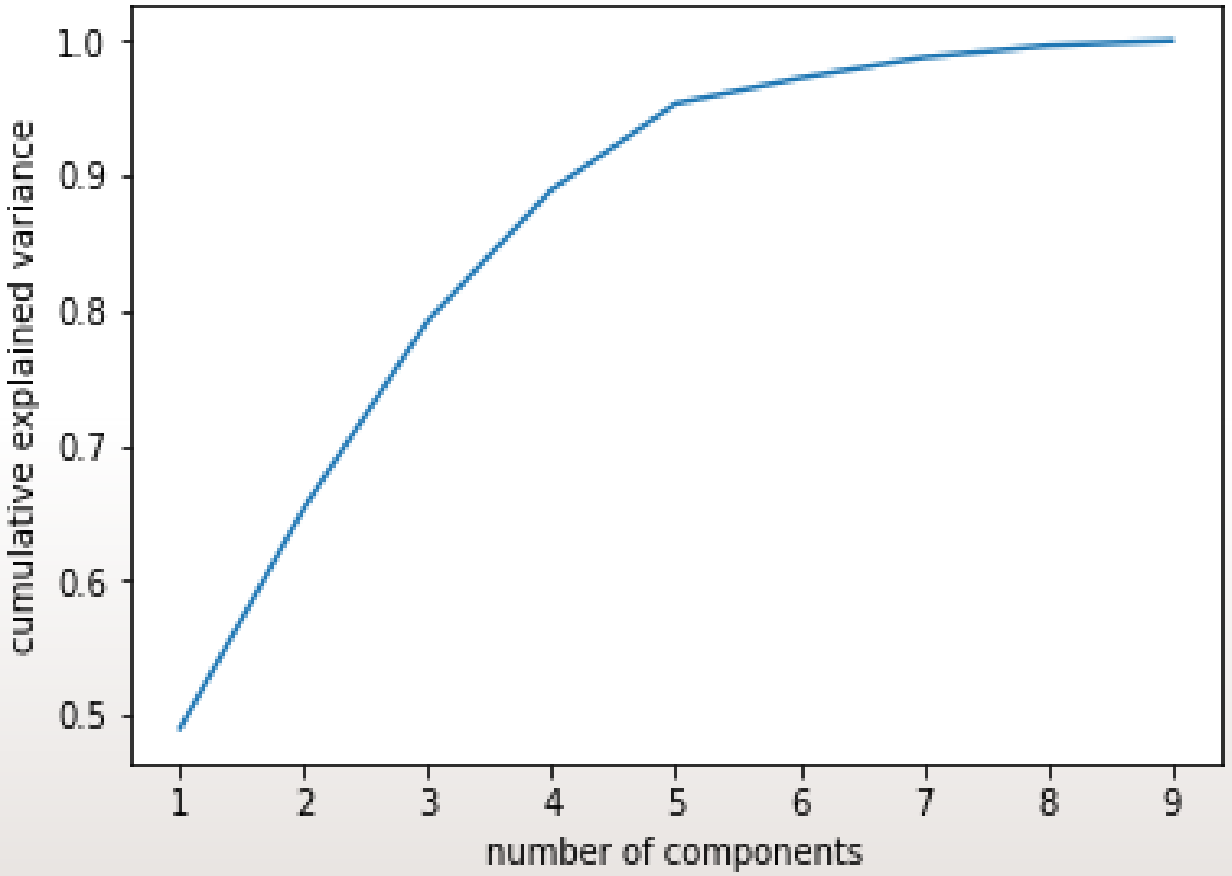
Explained variance ratio for each component

```
In [130]: pca.explained_variance_ratio_

Out[130]: array([0.49018372, 0.16342209, 0.13895338, 0.09695666, 0.06375024,
                 0.01874168, 0.01568208, 0.00866837, 0.00364178])
```

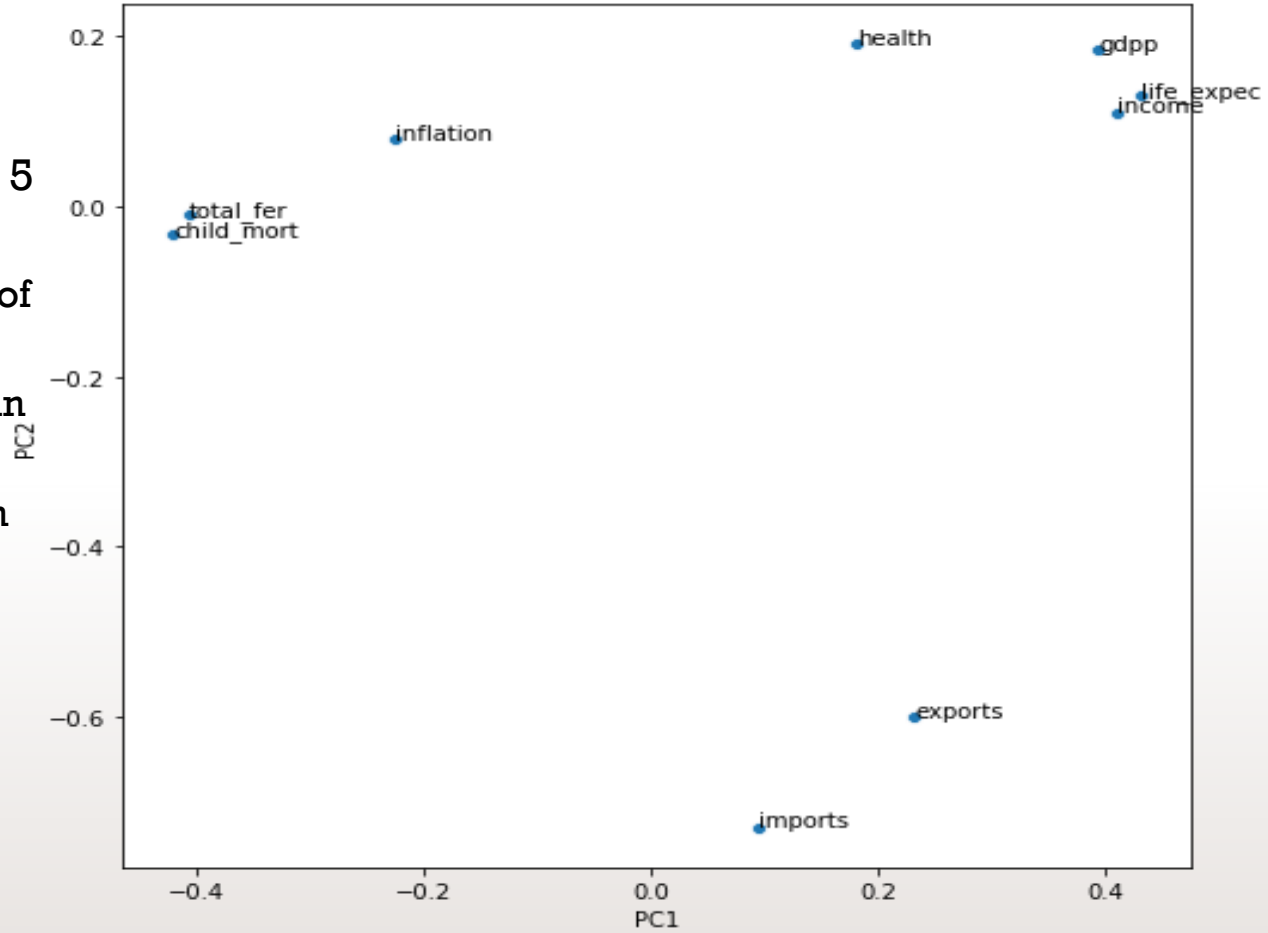
PCA- SCREE PLOT

- Plots cumulative explained variance
- Around 95% of variance is explained by 5 principal components`.
- We will choose 5 components for our modelling



PCA- PC1 VS PC2

- PC1 and PC2 are plotted on X & Y axis
- To visualise the load of variables on the selected 5 principal components
  - High variation of imports and exports in direction of pc1
  - Variation of income, gdpp and life_expec is more in direction of pc1 and less in direction of pc2
  - - High variation of imports and exports in direction of pc1
  - - Variation of income, gdpp and life_expec is more in direction of pc1 and less in direction of pc2
  - - PC2 is in the direction of child_mort inflation & total_ferPC2 is in the direction of child_mort inflation & total_fer



INCREMENTAL PCA

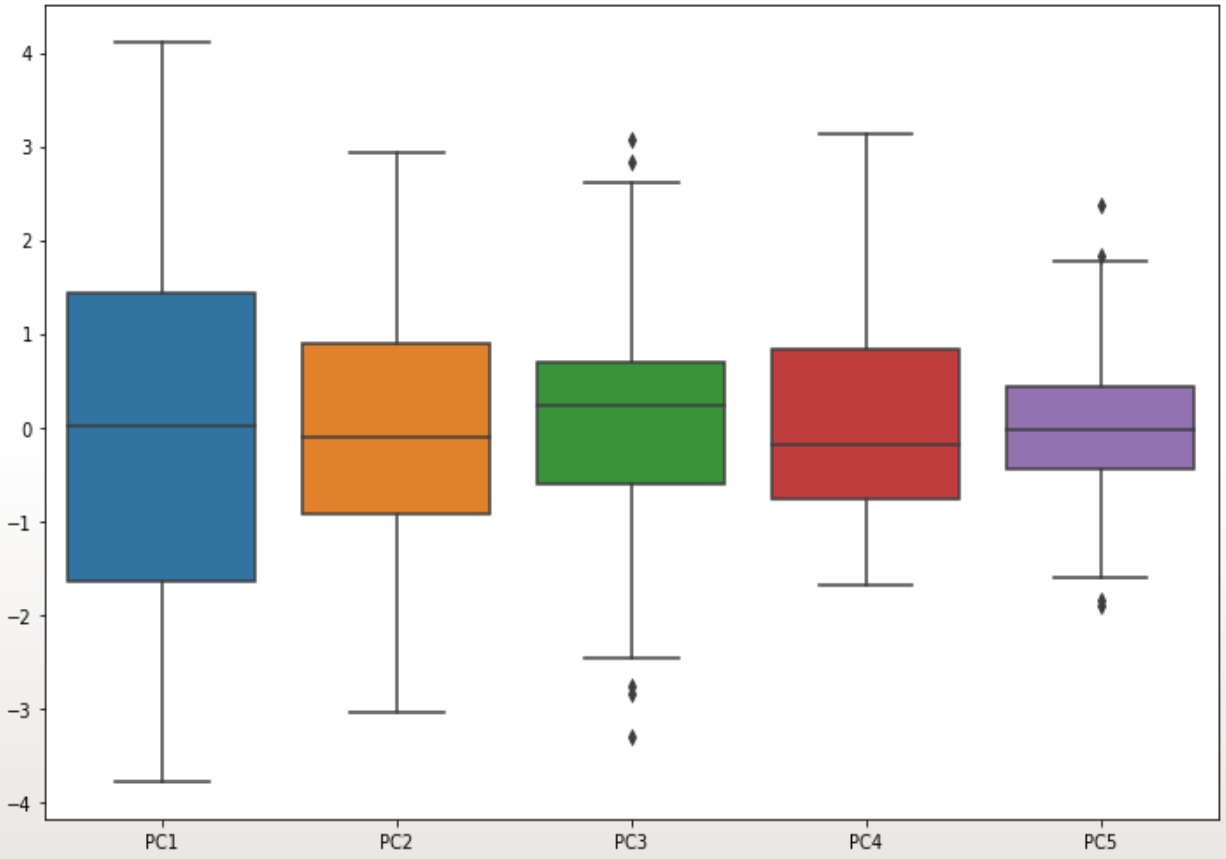
- Performed incremental pca with 5 principle components
- Fit transformed incremental PCA on scaled dataset - to transform the data with respect to 5 PC

Out[137]:

	PC1	PC2	PC3	PC4	PC5
0	-3.276496	-0.384763	1.062327	0.915380	0.050308
1	0.482584	-0.138287	0.316266	-1.430107	0.097874
2	-0.447400	-0.491523	-1.776731	-0.670289	0.515918
3	-3.357084	1.158895	-2.059029	1.732847	0.082381
4	1.245015	0.702900	0.224338	-0.744108	-0.447938

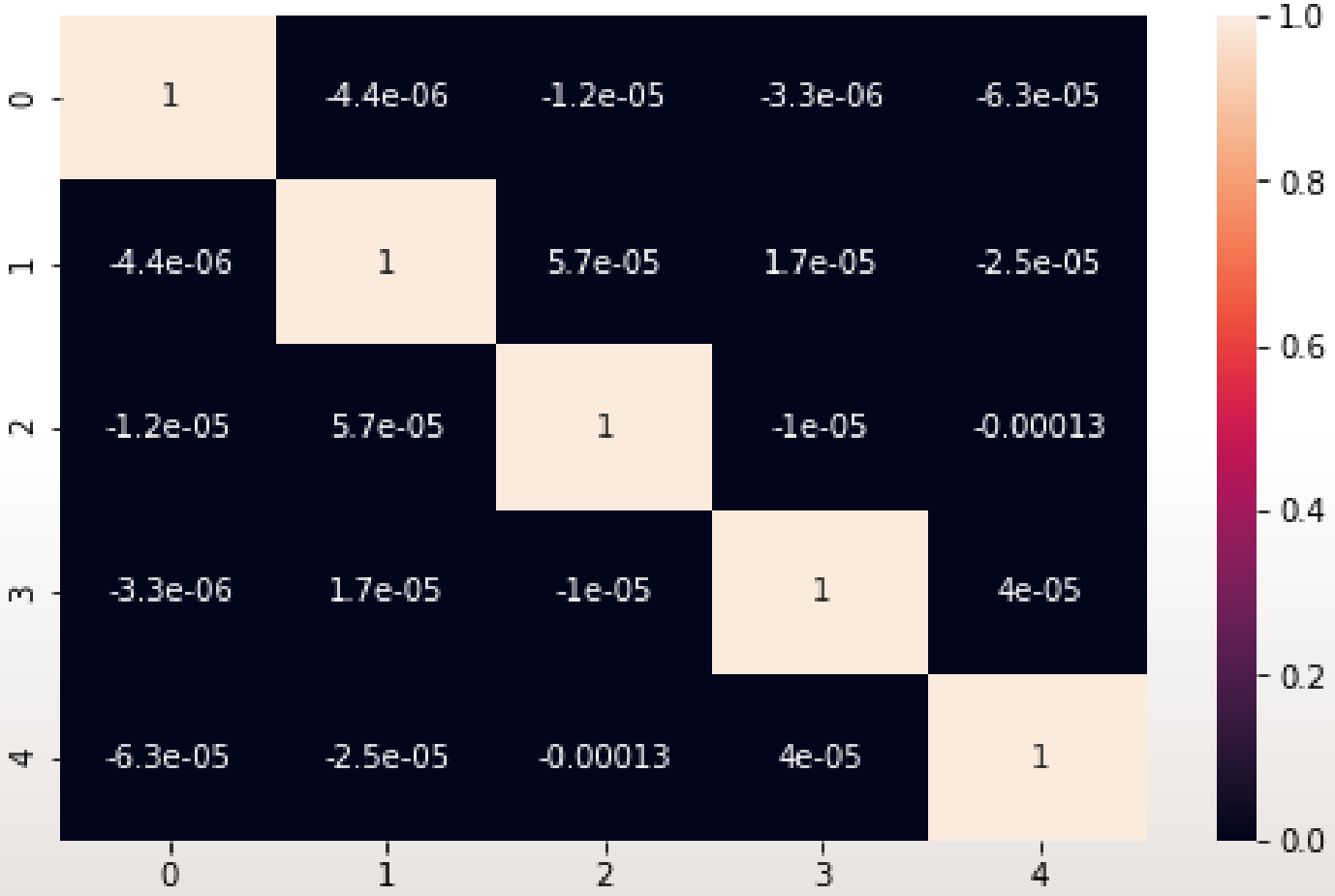
OUTLIERS POST PCA

- PC3 & PC5 shows minimal outliers
- As the extreme datapoints are already capped and floored during outlier treatment before PCA, **I do not want to treat them again.**
- **Removing them will lead to missing of countries which needs aid**



CORRELATION OF PC

- Correlation between the principal components is almost equal to zero
- **It means the PC's are NOT CORRELATED or NO RELATIONSHIP bn PC's**



CHECK FOR CLUSTER TENDENCY

- **HOPKINS STATISTIC** - is a statistic which gives a value which indicates the cluster tendency
- Value between {0.7, …, 0.99}, means higher tendency to cluster
- We have a value of 0.74 – we can go ahead with clustering

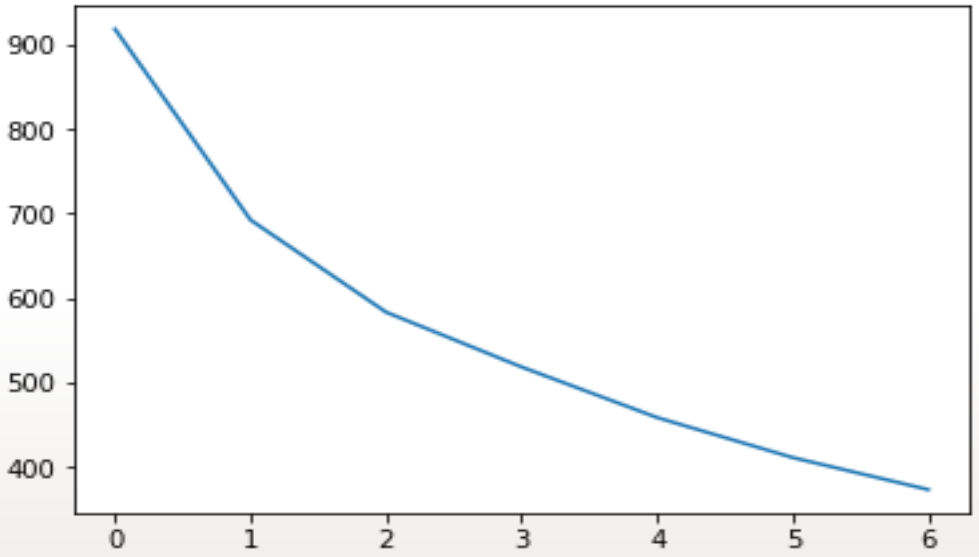
CLUSTERING

- CLUSTERING will be done on **PCA modified dataset**
- Both K Means & Hierarchical clustering will be performed
- One of the methods will be chosen to report the final list of countries

K-MEANS CLUSTERING

OPTIMAL NO OF CLUSTERS

- **Elbow method**



- **Silhouette analysis**

```
For n_clusters=2, the silhouette score is 0.3
For n_clusters=3, the silhouette score is 0.29
For n_clusters=4, the silhouette score is 0.27
For n_clusters=5, the silhouette score is 0.27
For n_clusters=6, the silhouette score is 0.27
For n_clusters=7, the silhouette score is 0.29
For n_clusters=8, the silhouette score is 0.28
```

From both the methods, 3 no of clusters seems fine.

K MEANS WITH K = 3

- K-means instantiated with n_clusters=3, max_iter=50 & random_state=100
- And fitted on pca modified dataset which gave us cluster labels
- Cluster id was assigned to both PCA data frame & country data frame

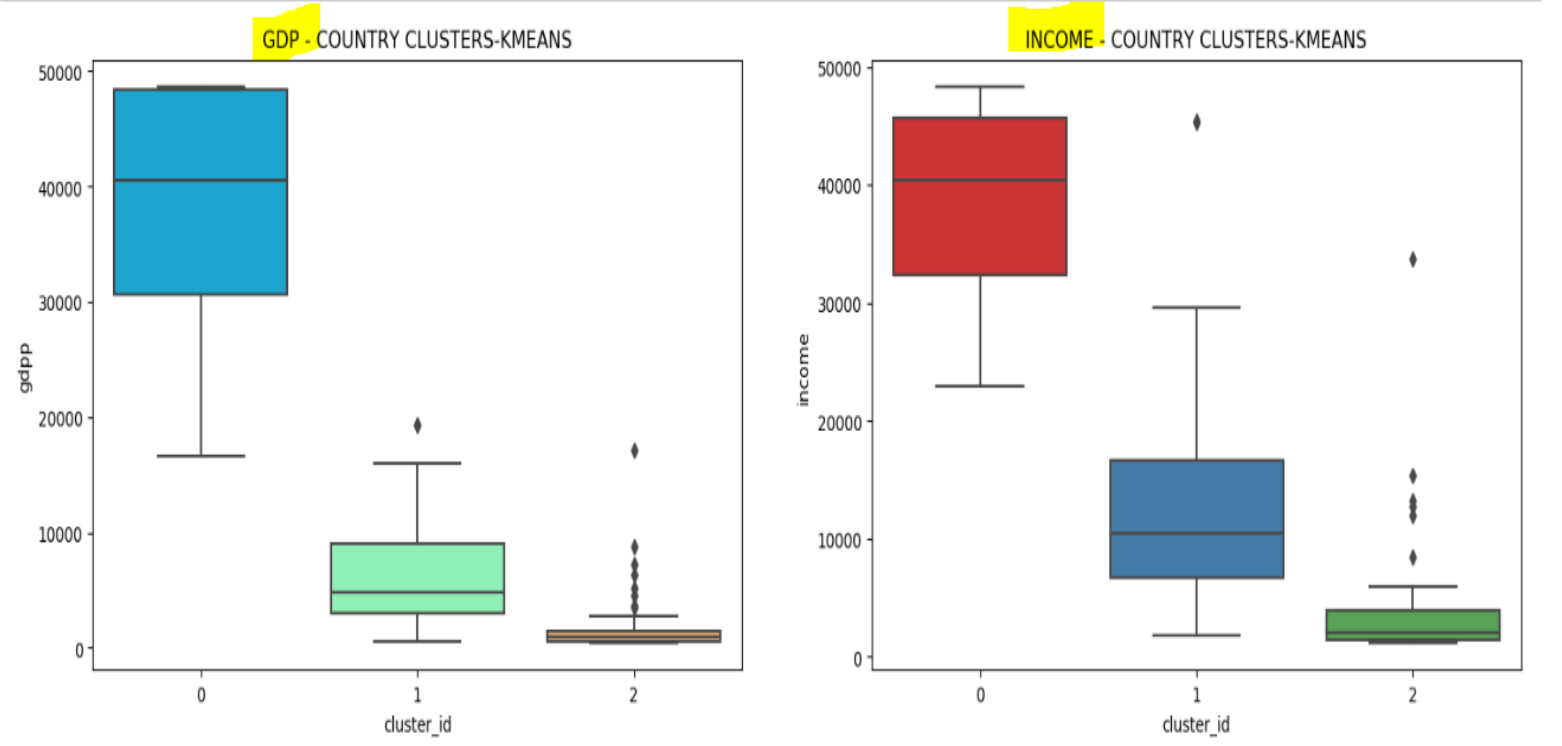
|:

	PC1	PC2	PC3	PC4	PC5	cluster_id	country
0	-3.276496	-0.384763	1.062327	0.915380	0.050308	2	Afghanistan
1	0.482584	-0.138287	0.316266	-1.430107	0.097874	1	Albania
2	-0.447400	-0.491523	-1.776731	-0.670289	0.515918	1	Algeria
3	-3.357084	1.158895	-2.059029	1.732847	0.082381	2	Angola
4	1.245015	0.702900	0.224338	-0.744108	-0.447938	1	Antigua and Barbuda

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	12.12	7.58	44.9	1610.0	9.44	56.2	5.820	553.0	2
1	Albania	16.6	28.00	6.55	48.6	9930.0	4.49	76.3	1.650	4090.0	1
2	Algeria	27.3	38.40	4.17	31.4	12900.0	16.10	76.5	2.890	4460.0	1
3	Angola	116.0	62.30	2.85	42.9	5900.0	20.87	60.1	5.861	3530.0	2
4	Antigua and Barbuda	10.3	45.50	6.03	58.9	19100.0	1.44	76.8	2.130	12200.0	1

CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

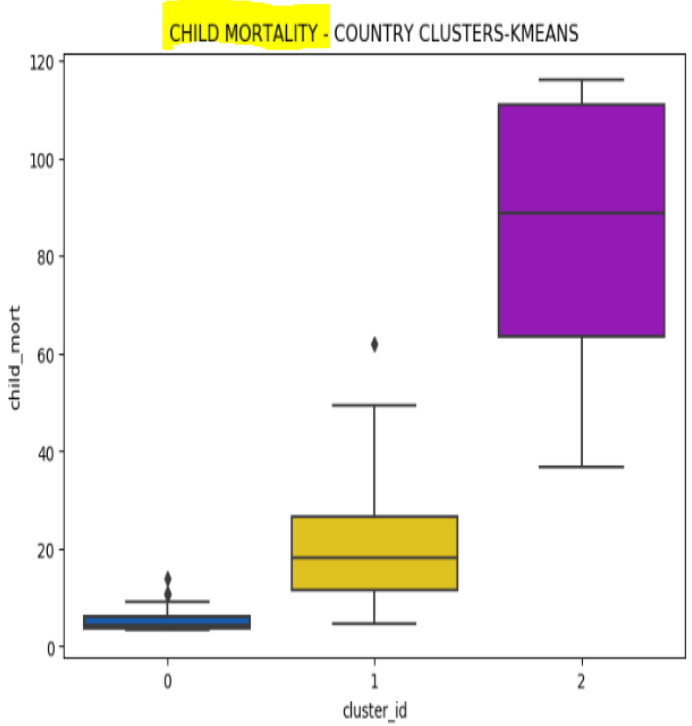
Gdpp and income variation for each cluster of countries

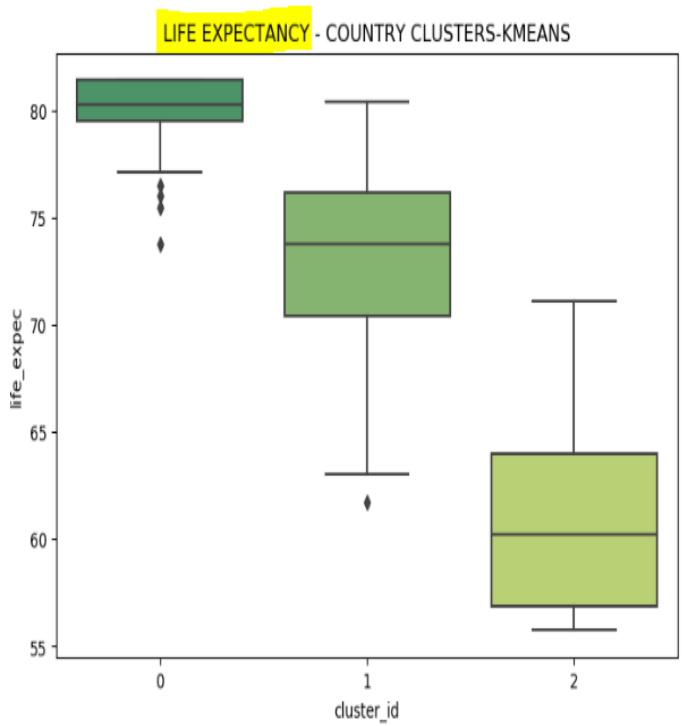


- Cluster 0 - HIGH gdpp, income - DEVELOPED COUNTRIES
- Cluster 2 - VERY LOW gdpp, income - UNDER - DEVELOPED COUNTRIES
- Cluster - LOW gdpp, AVERAGE income 1 - DEVELOPING COUNTRIES
- Its evident that cluster 2 needs AID

CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

Child mortality & life expectancy variation for each cluster of countries

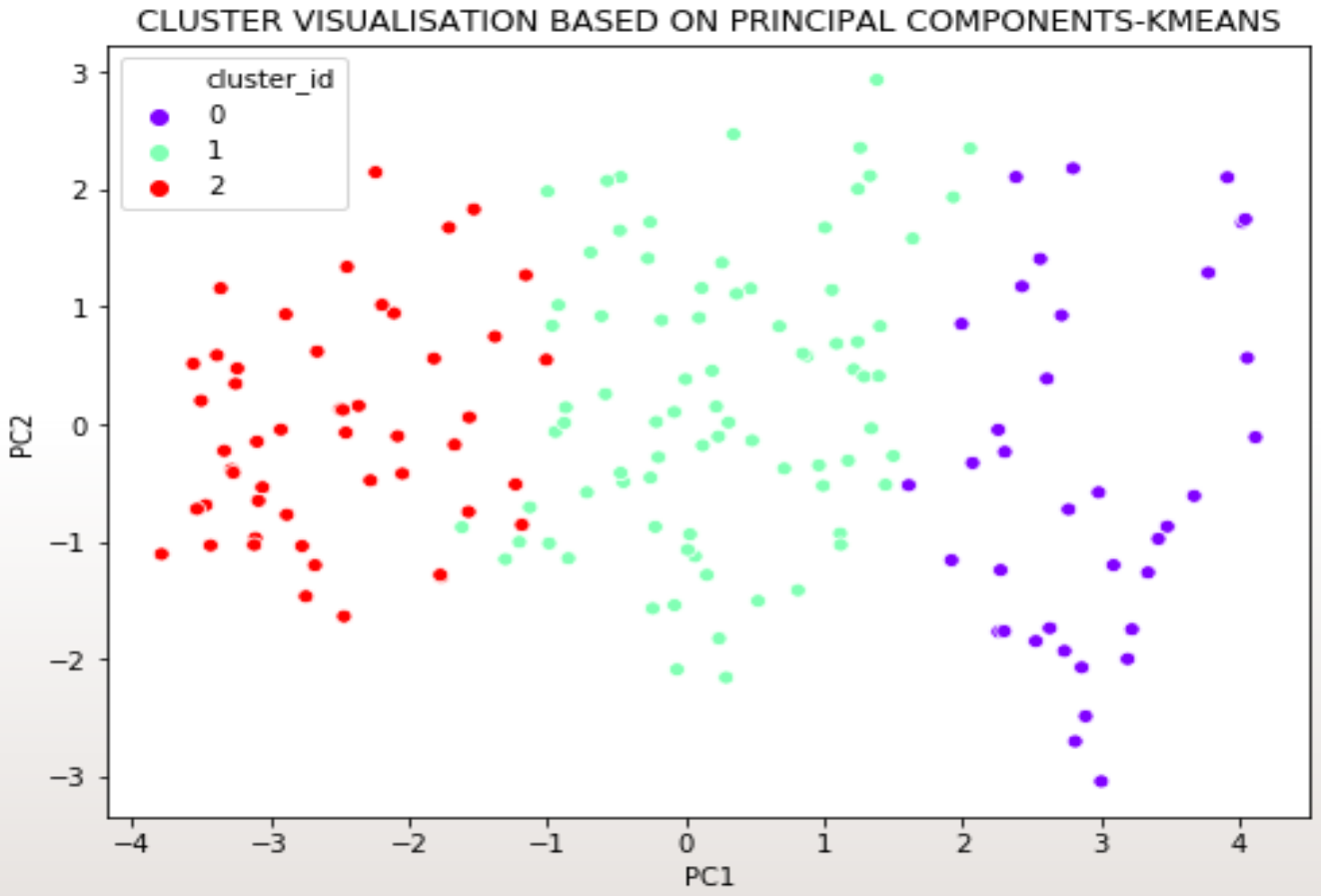




- Cluster 0 - HIGH gdpp, income - DEVELOPED COUNTRIES
- Cluster 2 - VERY LOW gdpp, income - UNDER - DEVELOPED COUNTRIES
- Cluster - LOW gdpp, AVERAGE income 1 - DEVELOPING COUNTRIES
- Its evident that cluster 2 needs AID

CLUSTER ANALYSIS
WRT PRINCIPAL COMPONENTS

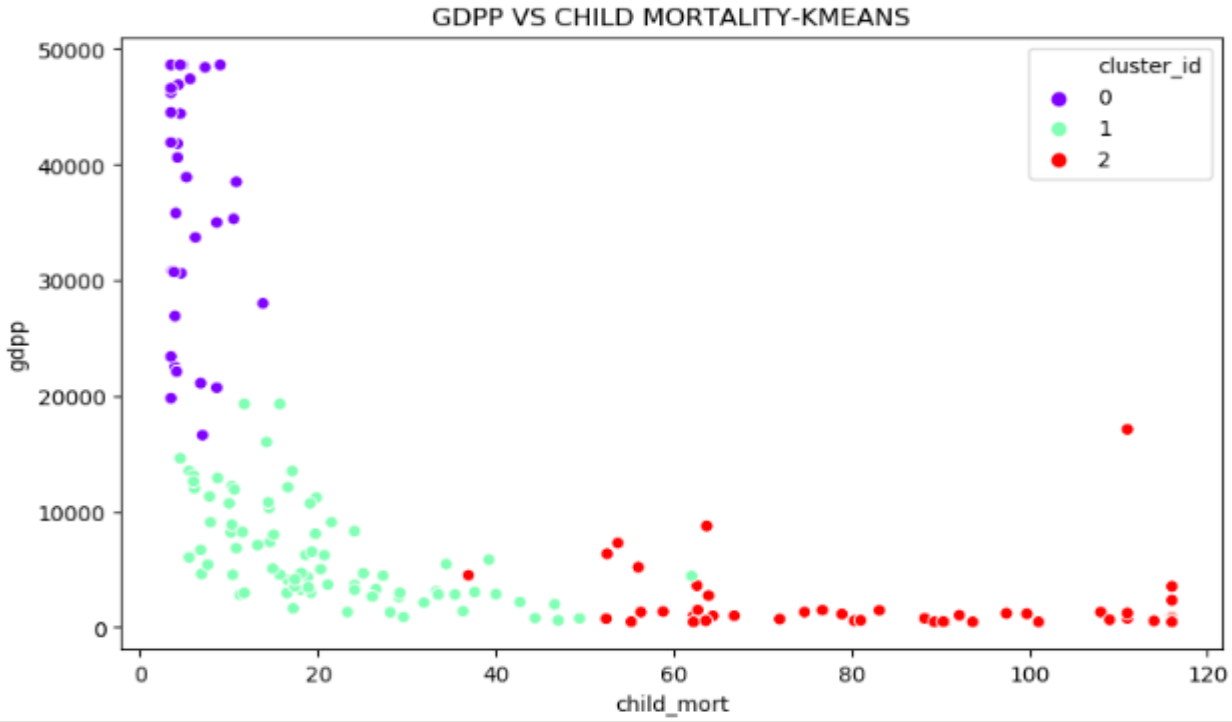
- Visualising first 2 principal components PC1 & PC2
- Done using scatter plot of all the countries, differentiating the clusters
- 3 unique clusters can be seen.



CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

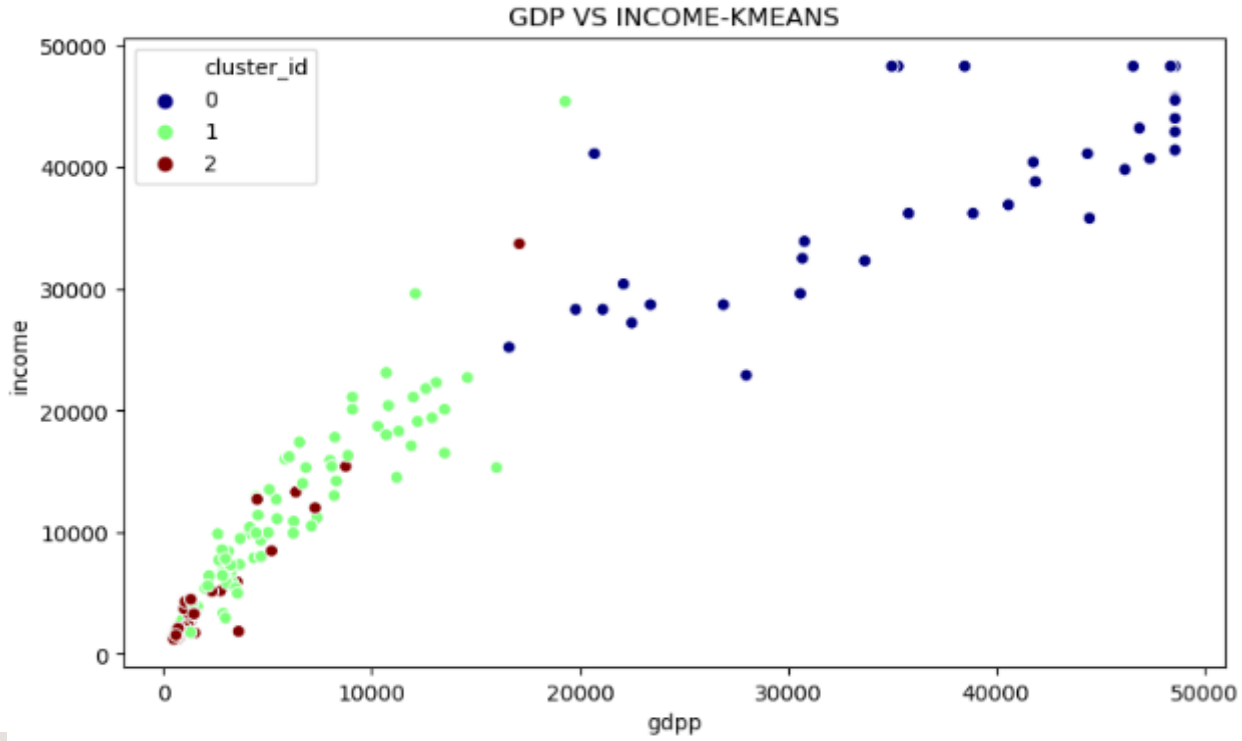
GDPP VS CHILD MORTALITY

- Lower the GDP of clusters, higher is the child mortality rate



GDP VS INCOME

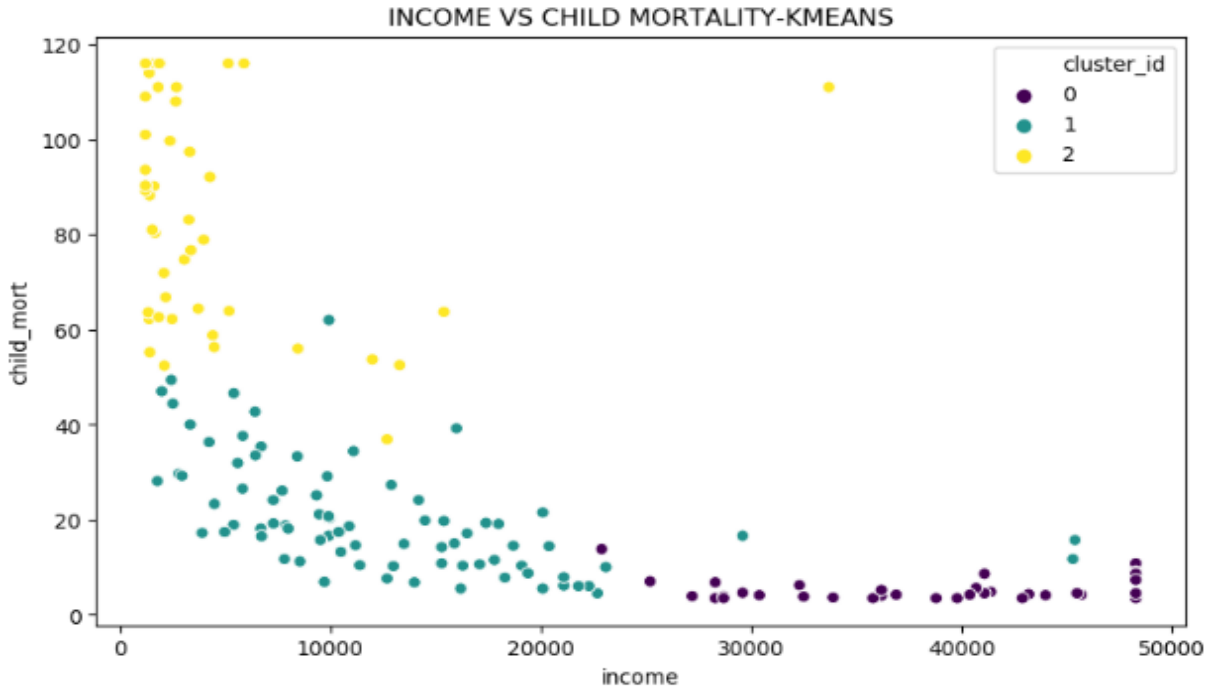
- When gdp is more, income per person is also more



CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

INCOME VS CHILD MORTALITY

- Countries having high income have low child mortality rate and vice versa



LIFE_EXPEC VS HEALTH

- No definite pattern observed

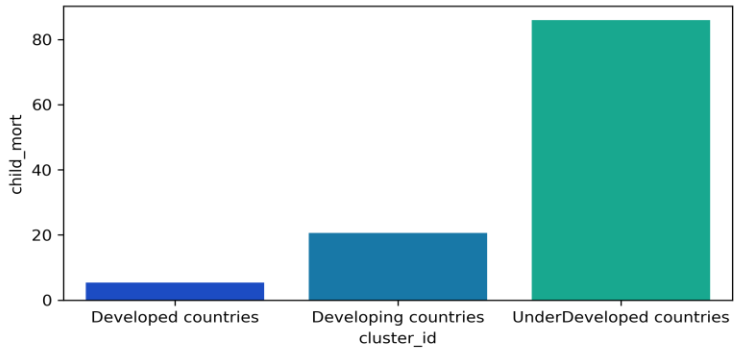
MEAN ANALYSIS OF CLUSTERS

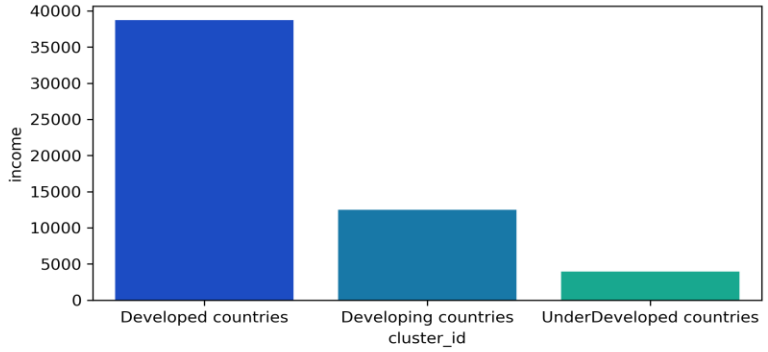
- Mean data is taken for all countries for each clusters
- Analysing each variable by its mean and clusters will give clear bifurcation of under-developed countries form the rest
- Cluster_id 's have been renamed for visualisation needs
  - cluster_id 0 -> DEVELOPED COUNTRIES
  - cluster_id 1 -> DEVELOPING COUNTRIES
  - cluster_id 2 -> UNDER - DEVELOPED COUNTRIES

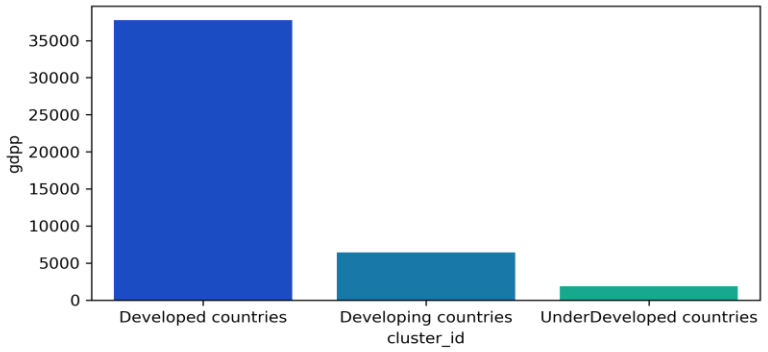
	cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	5.344324	49.753514	8.624108	45.197838	38711.081081	2.834373	79.805405	1.767838	37745.675676
1	1	20.656250	41.138125	6.234038	48.001750	12506.625000	7.130223	73.000000	2.297750	6422.325000
2	2	85.936000	28.659600	6.246880	41.267400	3966.540000	9.927900	60.716400	4.776180	1884.902000

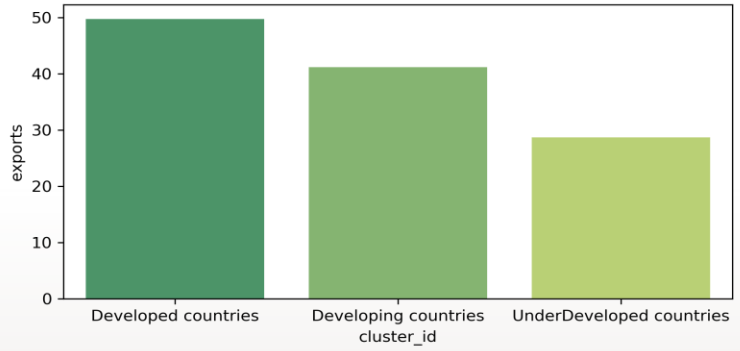
	cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Developed countries	5.344324	49.753514	8.624108	45.197838	38711.081081	2.834373	79.805405	1.767838	37745.675676
1	Developing countries	20.656250	41.138125	6.234038	48.001750	12506.625000	7.130223	73.000000	2.297750	6422.325000
2	UnderDeveloped countries	85.936000	28.659600	6.246880	41.267400	3966.540000	9.927900	60.716400	4.776180	1884.902000

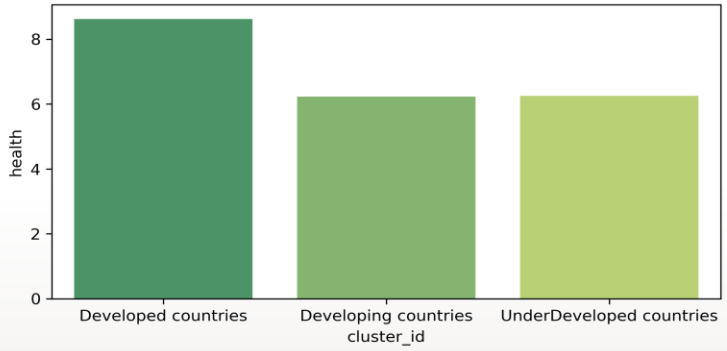
MEAN ANALYSIS OF CLUSTERS

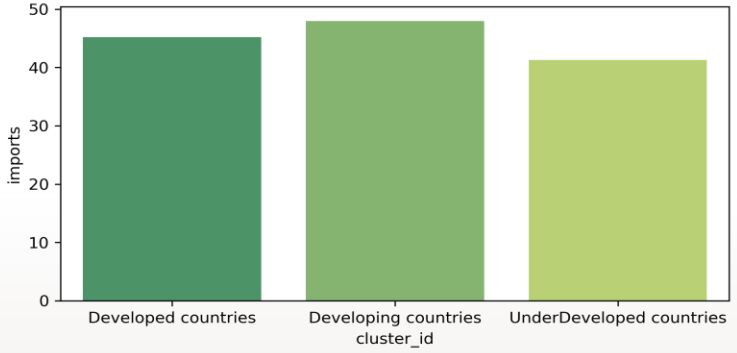


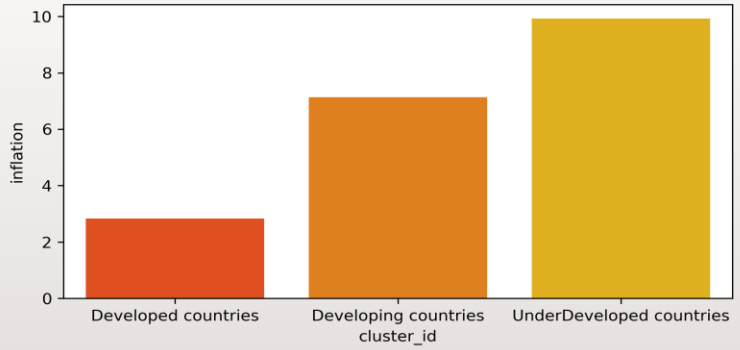


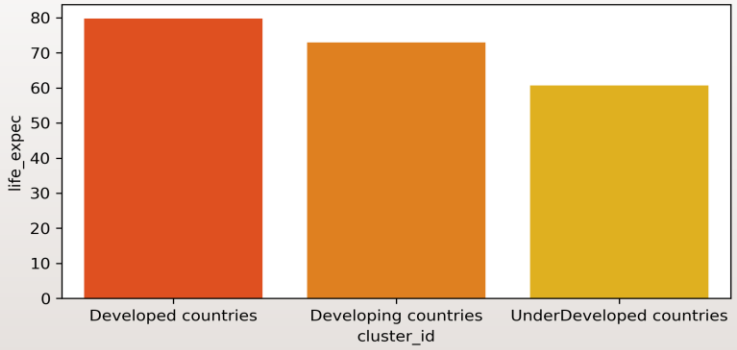


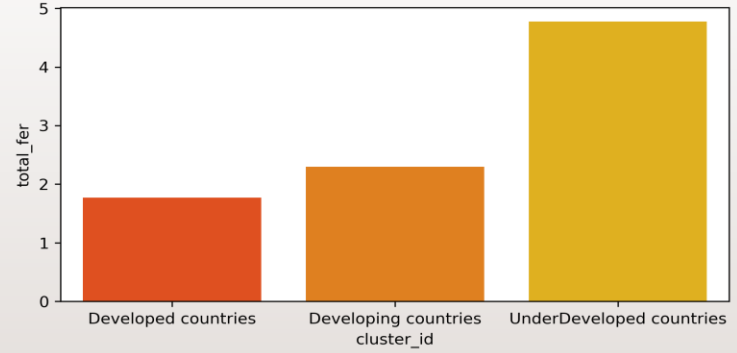












MEAN ANALYSIS OF CLUSTERS

- From the above plots and table which shows mean analysis of clusters of all variables, we can observe the following:
- Only gdpp, income & child_mort show considerable difference in clusters.
- Developed countries:
  - HIGH - gdp, income, health spending, exports, life expectancy
  - Avg - imports
  - VERY LOW - child mortality, inflation, total fertility
- Developing countries:
  - AVG - health spending,exports, life expectancy, total fertility,child mortality, inflation,income
  - HIGH - imports
  - LOW – gdp
- **Under-Developed countries:**
  - VERY LOW - gdp, income,exports, life expectancy,Imports
  - Avg - health spending
  - HIGH - child mortality, inflation, total fertility

TOP 10 COUNTRIES – CLUSTER WISE

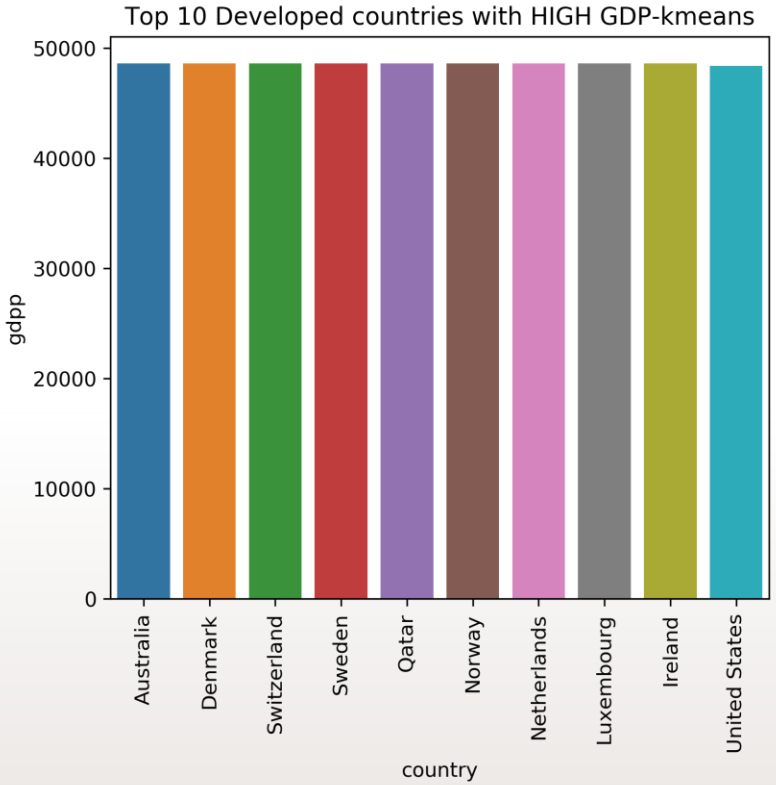
- 3 new data frames subsetted for each cluster
- groupby & sort_values functions are used instead of binning
- Top 10 countries are grouped and sorted with respect to their gdp, income per person & child mortality

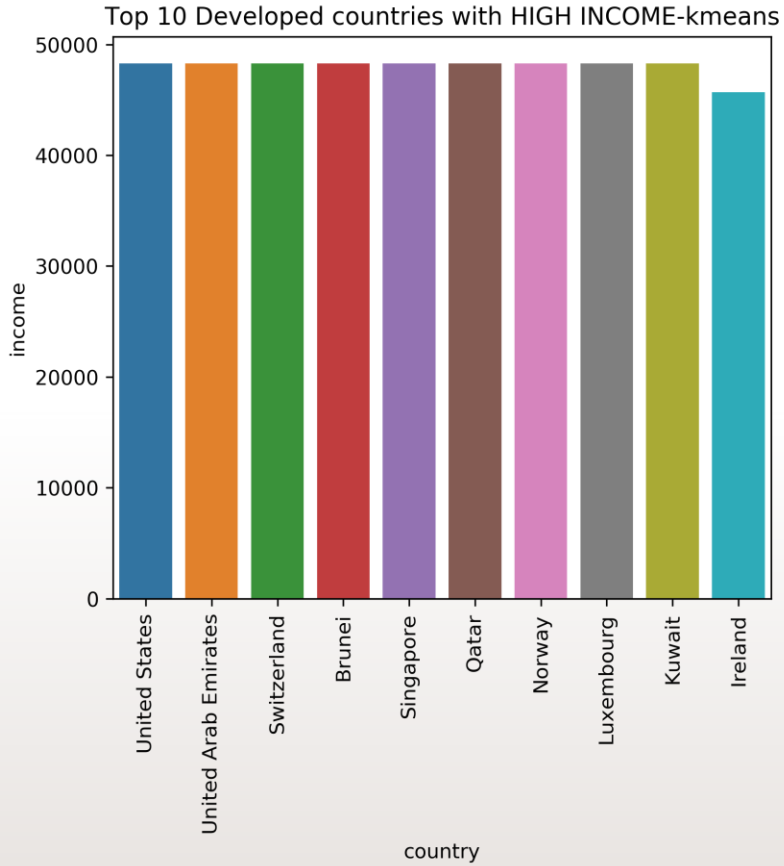
```
print(developed_df.shape)
print(developing_df.shape)
print(UnderDeveloped_df.shape)
```

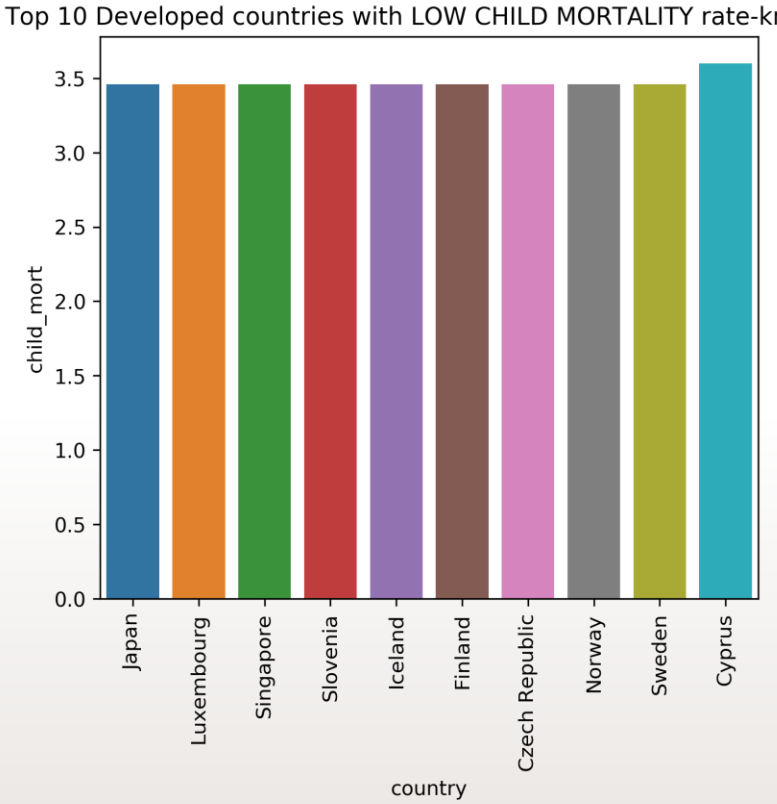
```
(37, 11)
(80, 11)
(50, 11)
```

TOP 10 DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







TOP 10 DEVELOPED COUNTRIES

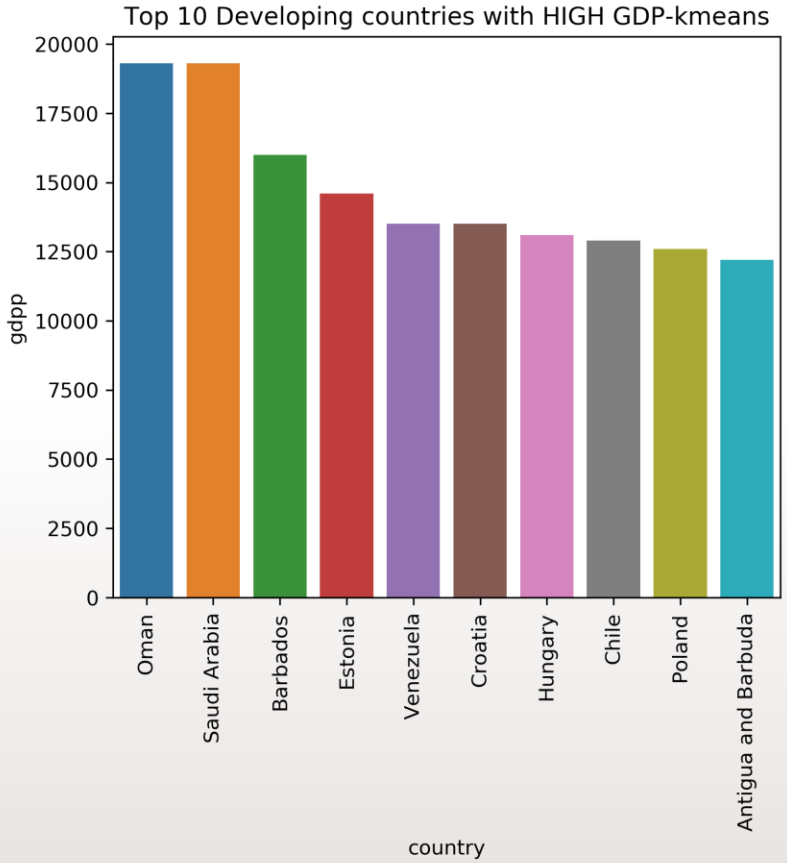
WRT GDPP,INCOME & CHILD_MOR

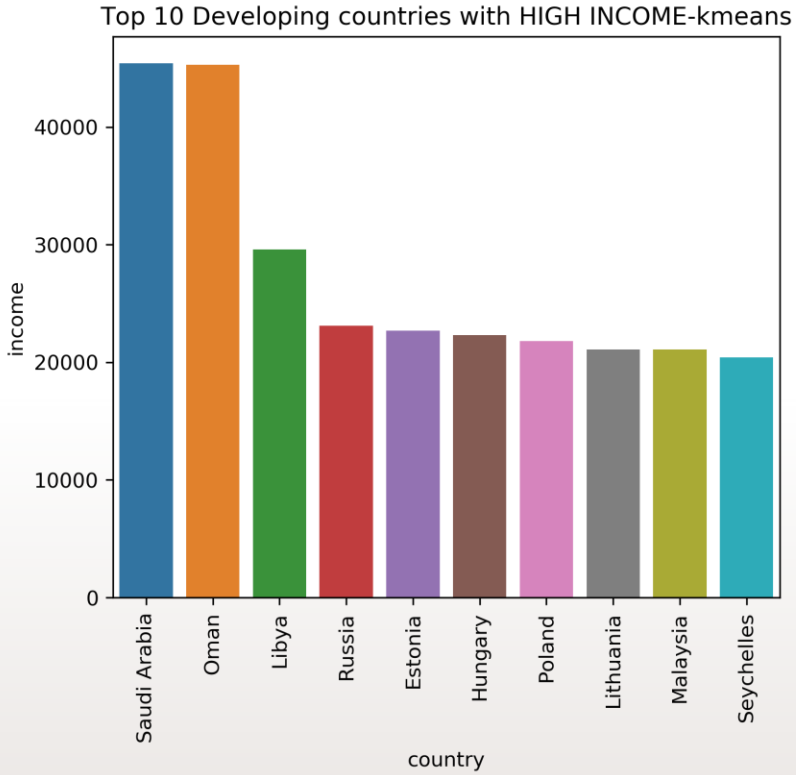
- The developed countries have very high gdp and income per person and very low child mortality rate
- Below are the top developed countries:

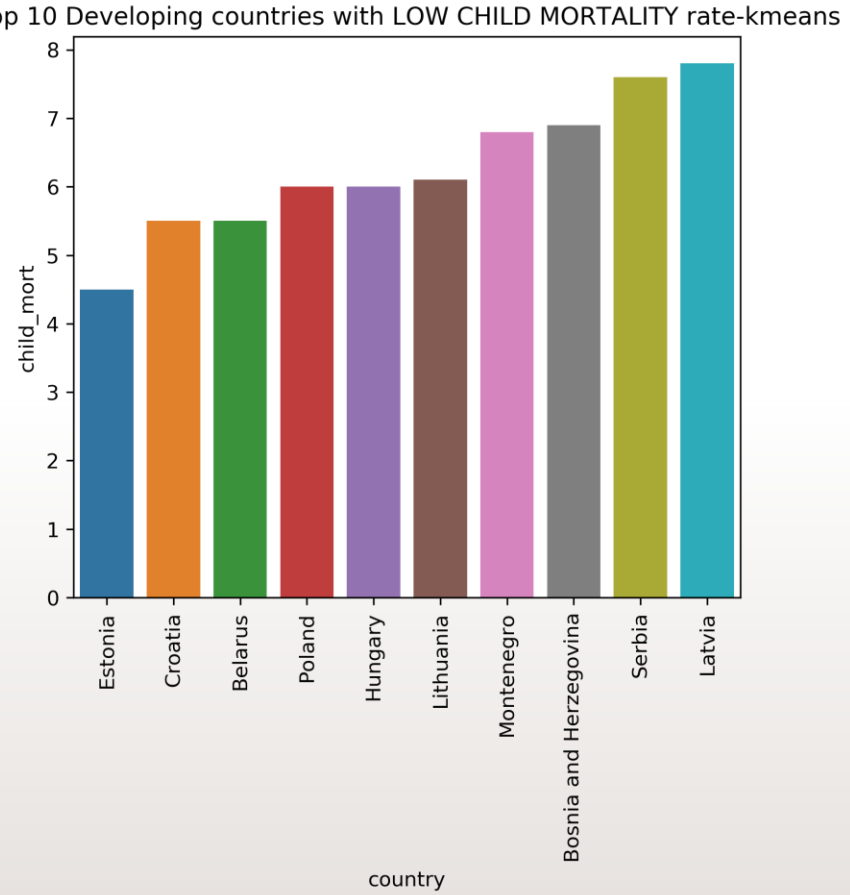
- Australia
- Denmark
- Switzerland
- Sweden
- Qatar
- Norway
- Netherlands
- Luxemburg
- Ireland
- United States
- UAE
- Kuwait
- Singapore
- Iceland
- Japan
- Finland

TOP 10 DEVELOPING COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







TOP 10 DEVELOPING COUNTRIES

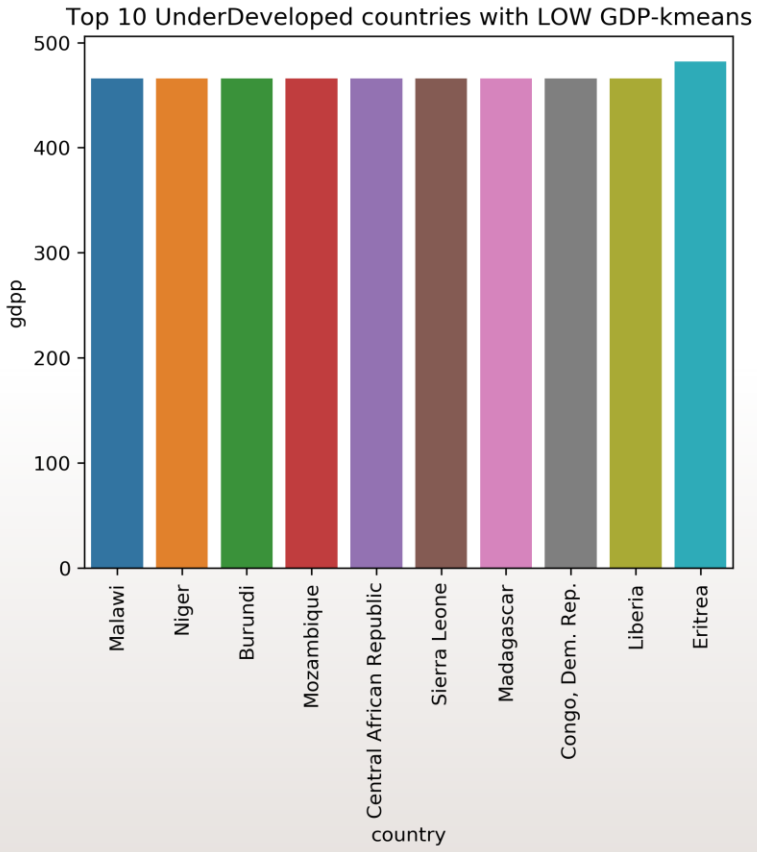
WRT GDPP,INCOME & CHILD_MOR

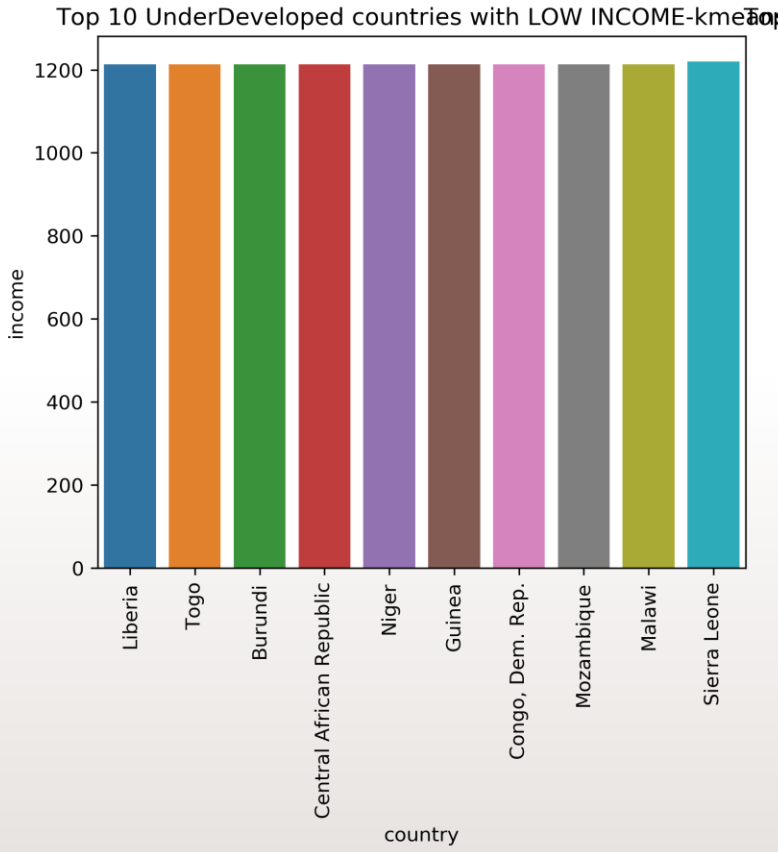
- The developing countries have very low gdp and income per person and average child mortality rate
- Below are the top developing countries:

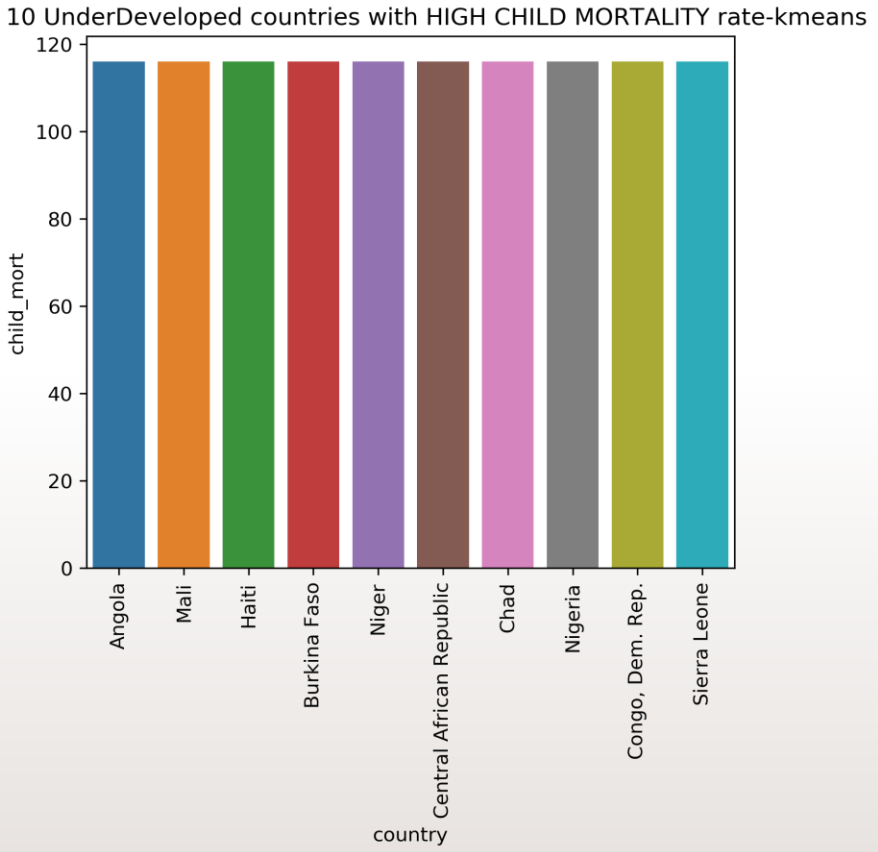
- Oman
- Saudi Arabia
- Libya
- Malaysia
- Hungary
- Lithuania
- Poland
- Croatia
- Estonia
- Serbia
- Barbados
- Russia
- Belarus
- Latvia

TOP 10 UNDER DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







TOP 10 UNDER DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR

- As per K-means clustering, below are the countries which are in dire need of aid. The below countries have very low gdp and income per person and very high child mortality rate

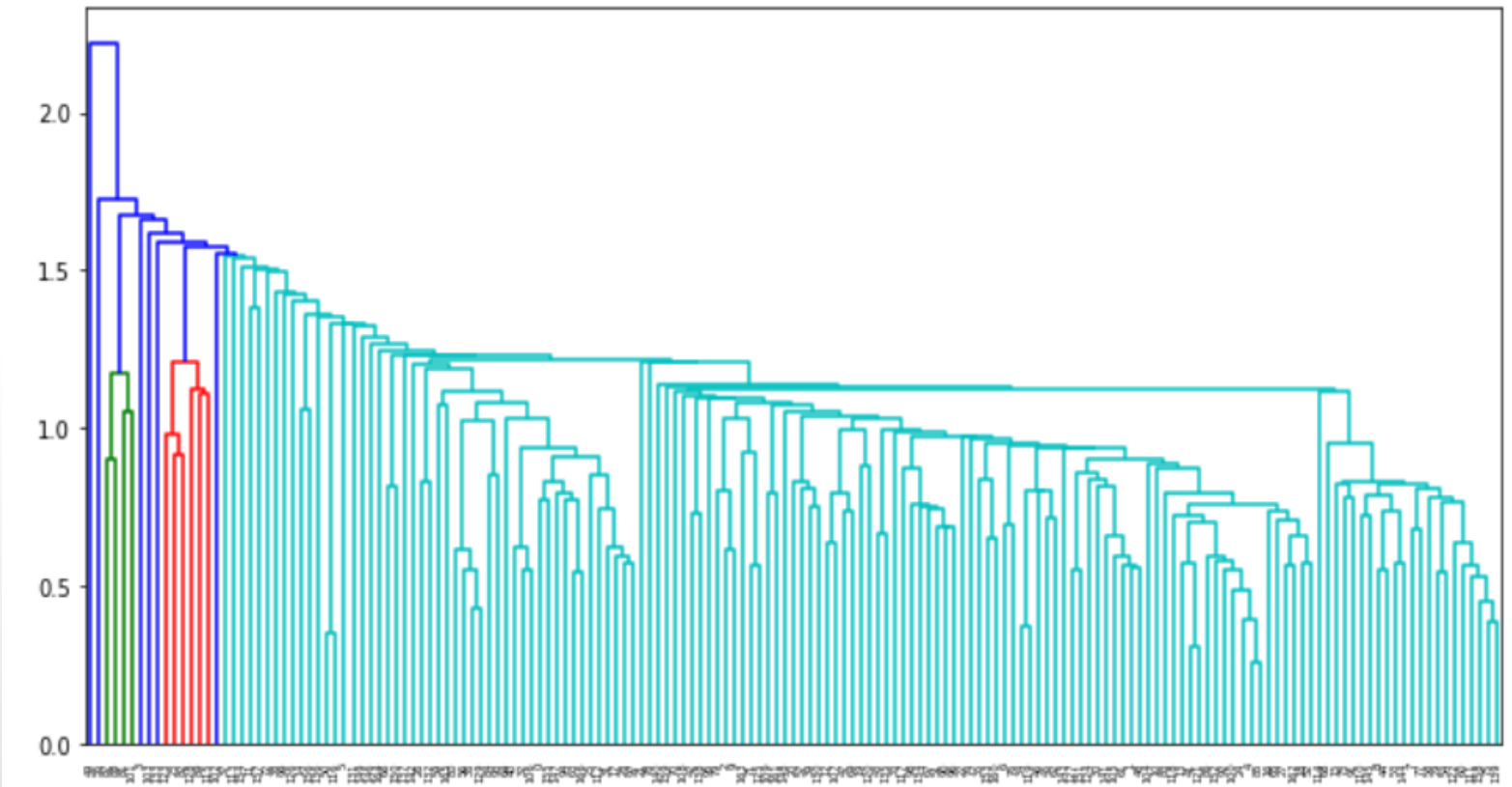
- Burundi
- Malawi
- Niger
- Mozambique
- Central African republic
- Sierre Leone
- Congo, Dem rep
- Liberia
- Eritrea
- Togo
- Guinea
- Madagascar
- Angola
- Mali
- Haiti
- Burkino Faso
- Chad
- Nigeria

HIERARCHICAL CLUSTERING

OPTIMAL NO OF CLUSTERS
DENDROGRAM

Single linkage

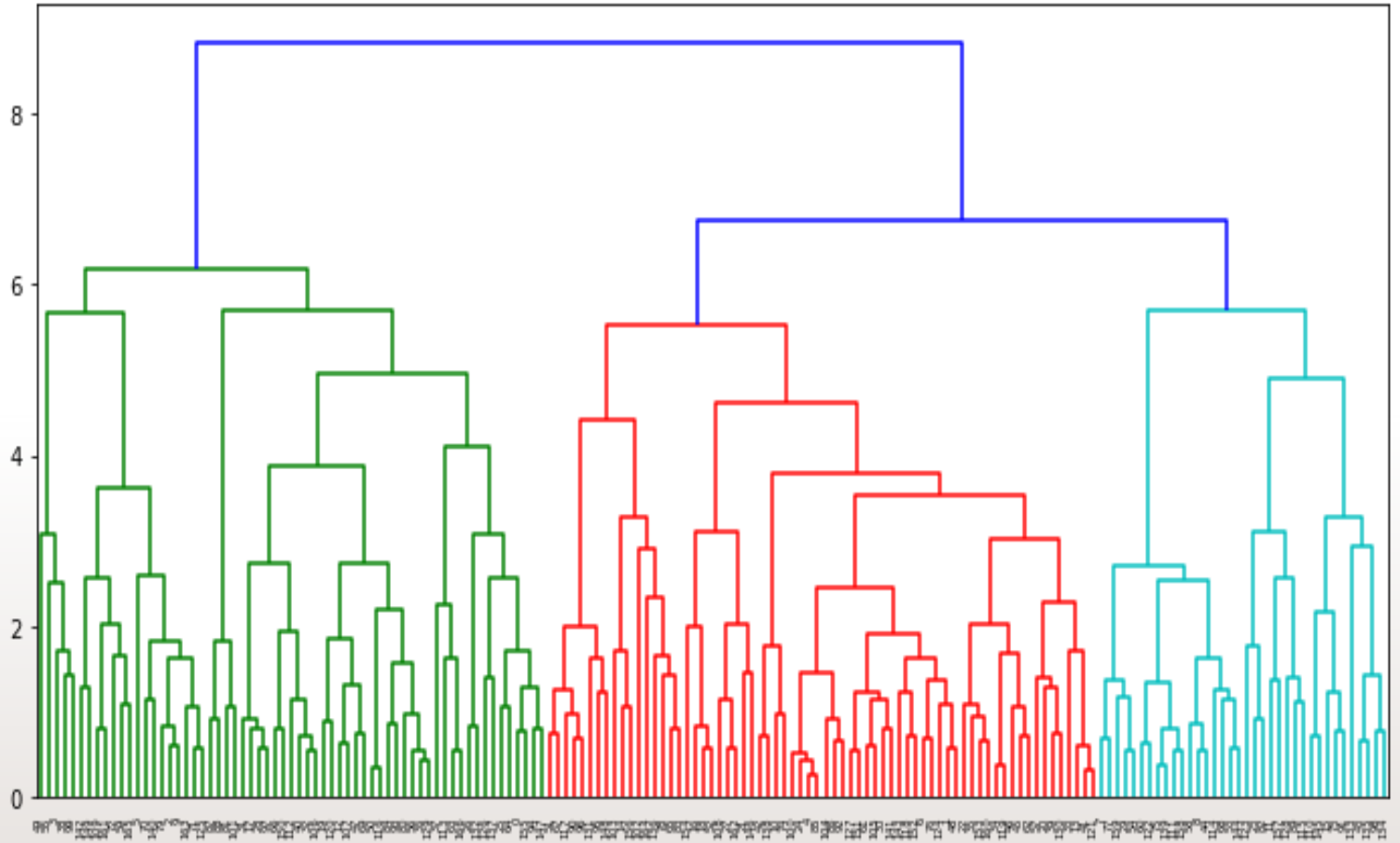
- Hierarchical done using single linkage & Euclidean metric
- As the dendrogram is not very much interpretable using single linkage, we shall use **complete** linkage



OPTIMAL NO OF CLUSTERS
DENDROGRAM

COMPLETE linkage

- Hierarchical done AGAIN using COMPLETE linkage & Euclidean metric
- we can infer from the dendrogram that 3 clusters can be formed when we cut it at a height 6.5



HIERARCHICAL WITH 3 CLUSTERS

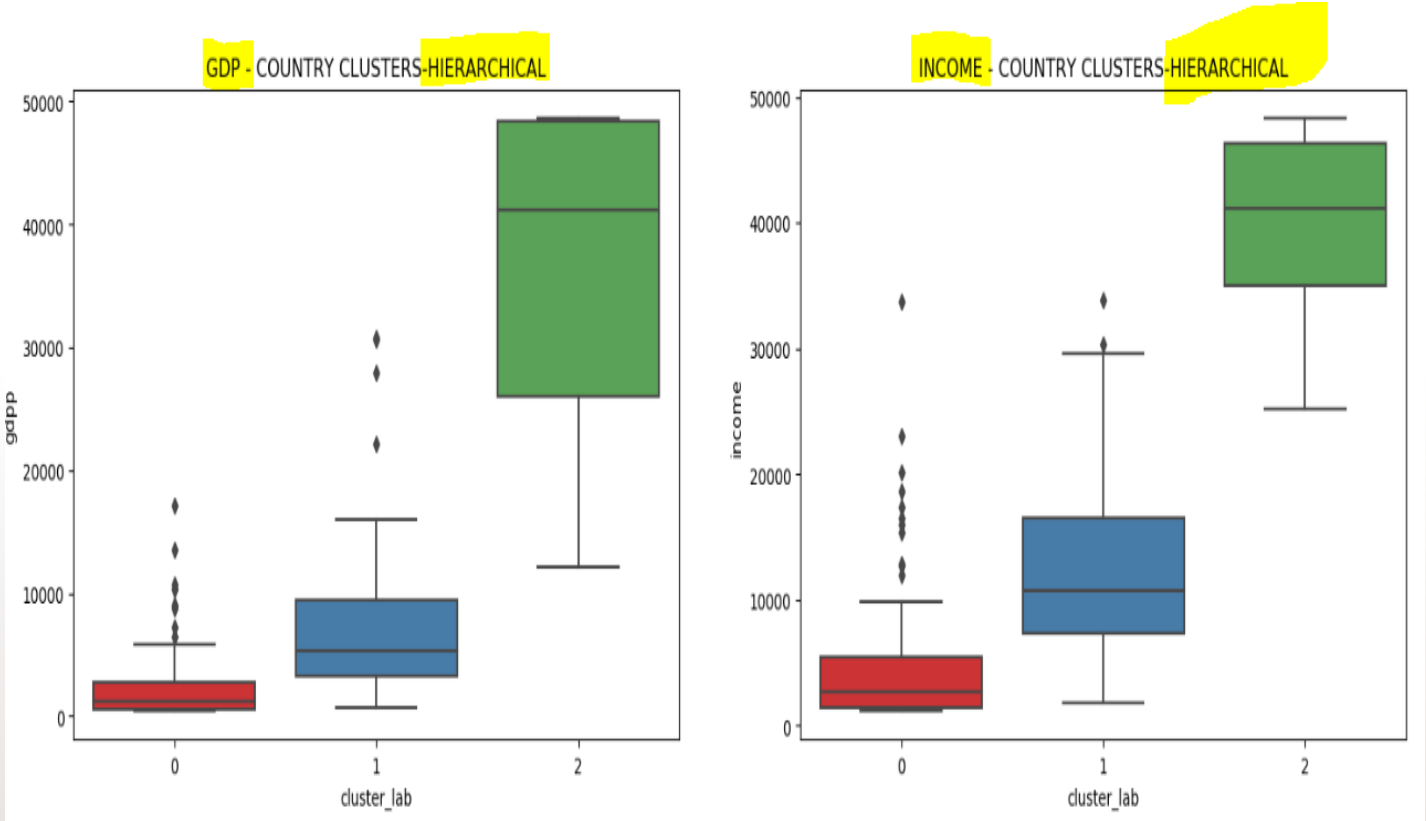
- Hierarchical clustering was done with 3 number of clusters
- Resulted in cluster labels
- Cluster labels was assigned to both PCA data frame & country data frame

	PC1	PC2	PC3	PC4	PC5	cluster_lab	country
0	-3.276496	-0.384763	1.062327	0.915380	0.050308	0	Afghanistan
1	0.482584	-0.138287	0.316266	-1.430107	0.097874	1	Albania
2	-0.447400	-0.491523	-1.776731	-0.670289	0.515918	0	Algeria
3	-3.357084	1.158895	-2.059029	1.732847	0.082381	0	Angola
4	1.245015	0.702900	0.224338	-0.744108	-0.447938	1	Antigua and Barbuda

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_lab
0	Afghanistan	90.2	12.12	7.58	44.9	1610.0	9.44	56.2	5.820	553.0	0
1	Albania	16.6	28.00	6.55	48.6	9930.0	4.49	76.3	1.650	4090.0	1
2	Algeria	27.3	38.40	4.17	31.4	12900.0	16.10	76.5	2.890	4460.0	0
3	Angola	116.0	62.30	2.85	42.9	5900.0	20.87	60.1	5.861	3530.0	0
4	Antigua and Barbuda	10.3	45.50	6.03	58.9	19100.0	1.44	76.8	2.130	12200.0	1

CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

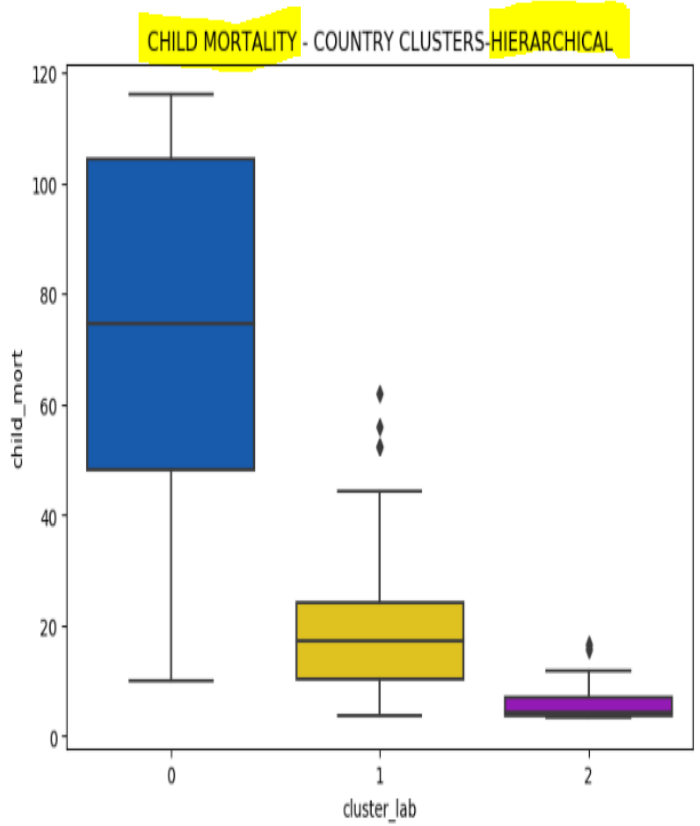
Gdpp and income variation for each cluster of countries

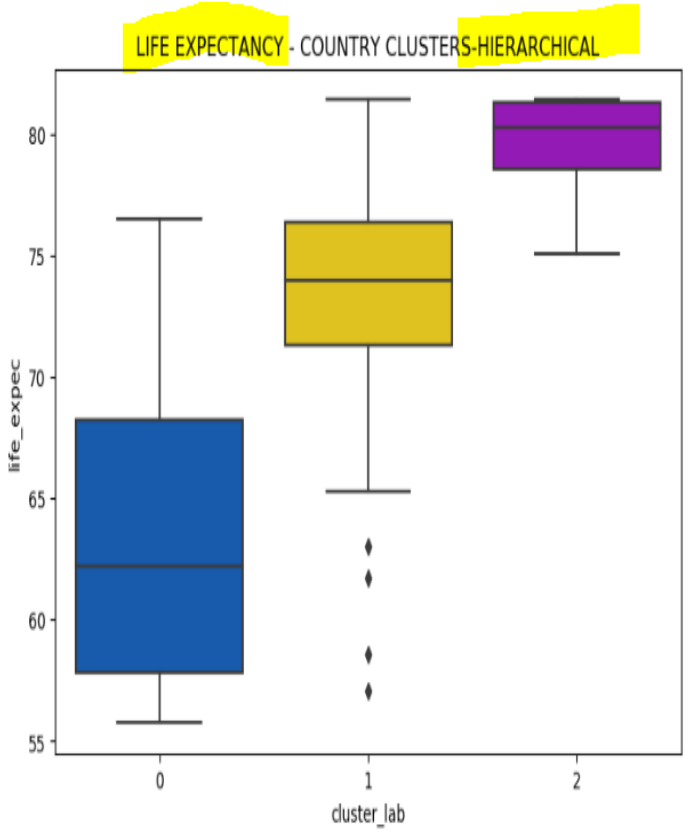


- Cluster 0 - VERY LOW gdpp, income - UNDER - DEVELOPED COUNTRIES
- Cluster 1 - LOW gdpp, AVERAGE income - DEVELOPING COUNTRIES
- Cluster 2 - HIGH gdpp, income - DEVELOPED COUNTRIES
- Its evident that cluster 0 needs AID

CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

Child mortality & life expectancy variation for each cluster of countries

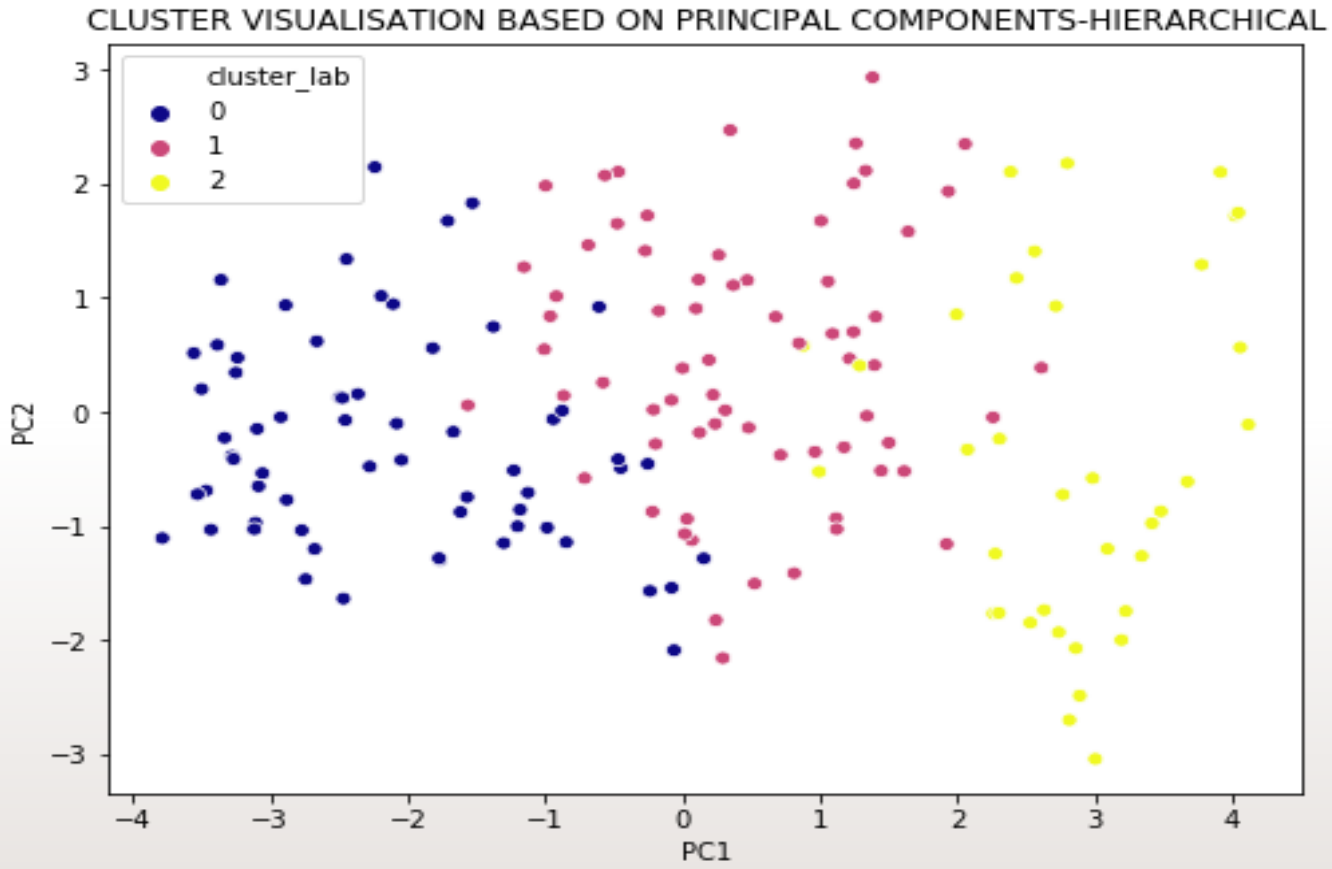




- Cluster 0 - VERY LOW gdpp, income - UNDER - DEVELOPED COUNTRIES
- Cluster 1 - LOW gdpp, AVERAGE income - DEVELOPING COUNTRIES
- Cluster 2 - HIGH gdpp, income - DEVELOPED COUNTRIES
- Its evident that cluster 0 needs AID

CLUSTER ANALYSIS
WRT PRINCIPAL COMPONENTS

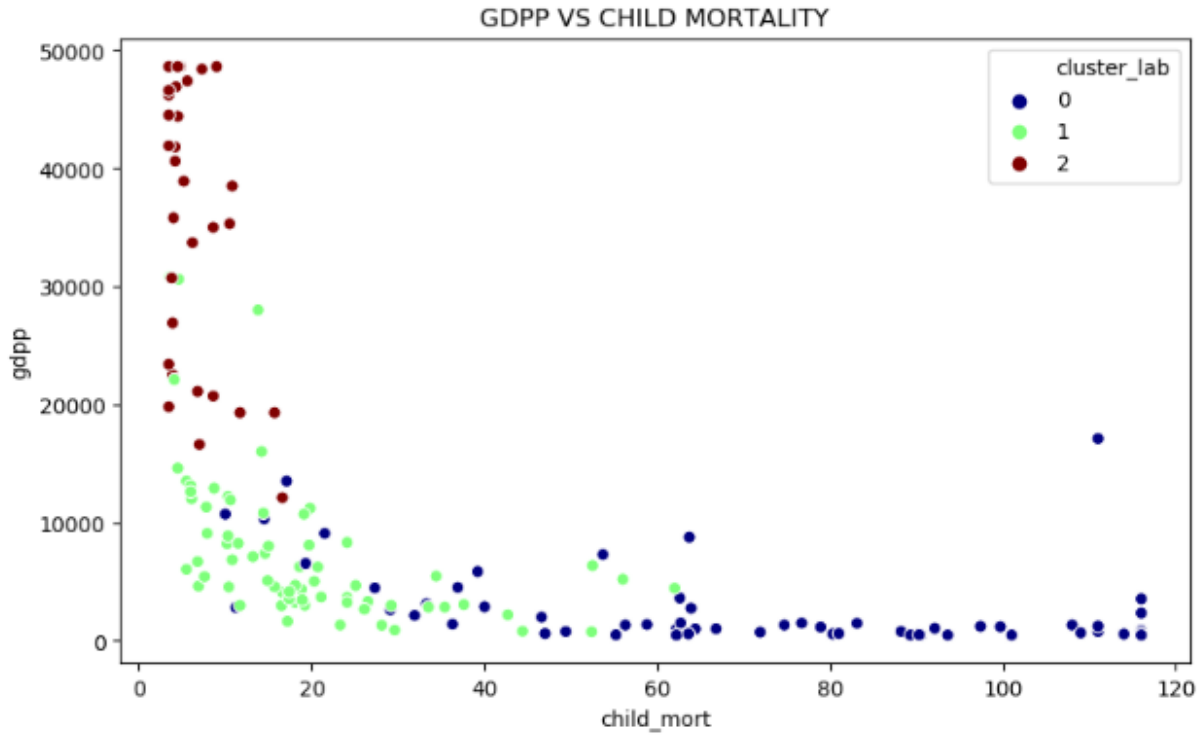
- Visualising first 2 principal components PC1 & PC2
- Done using scatter plot of all the countries, differentiating the clusters
- 3 unique clusters can be seen.



CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

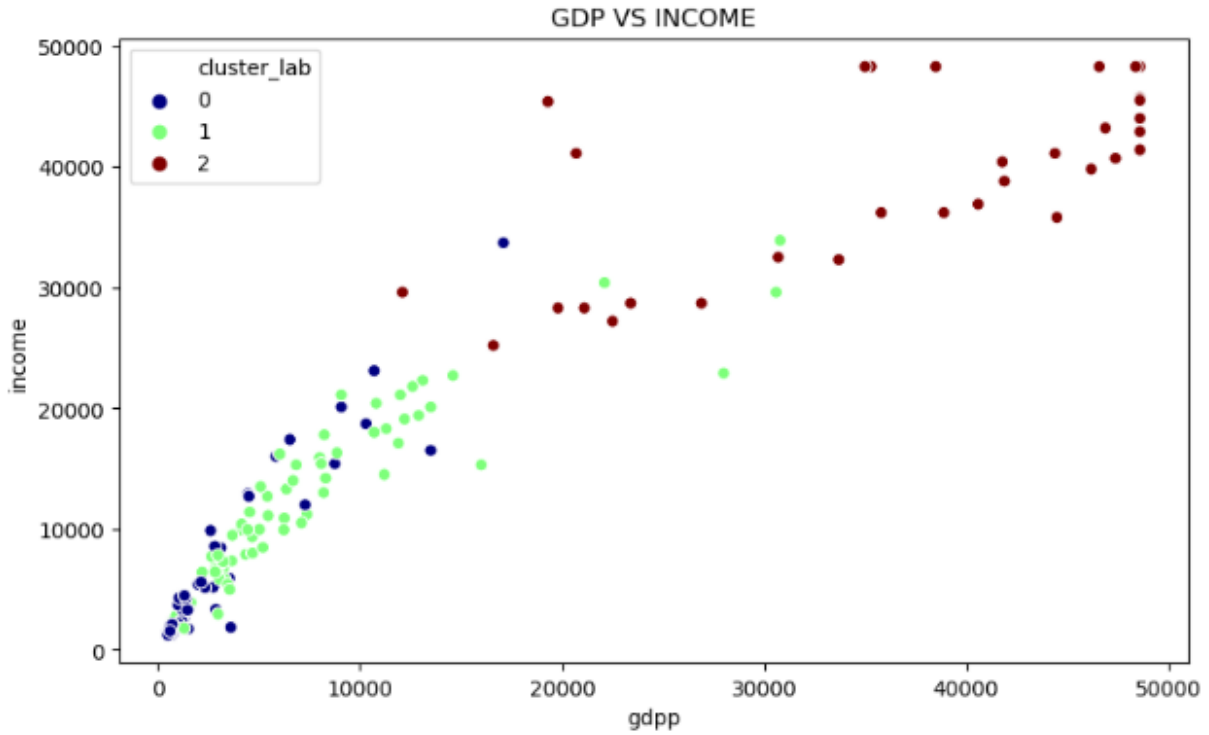
GDPP VS CHILD MORTALITY

- Same as K-means



GDP VS INCOME

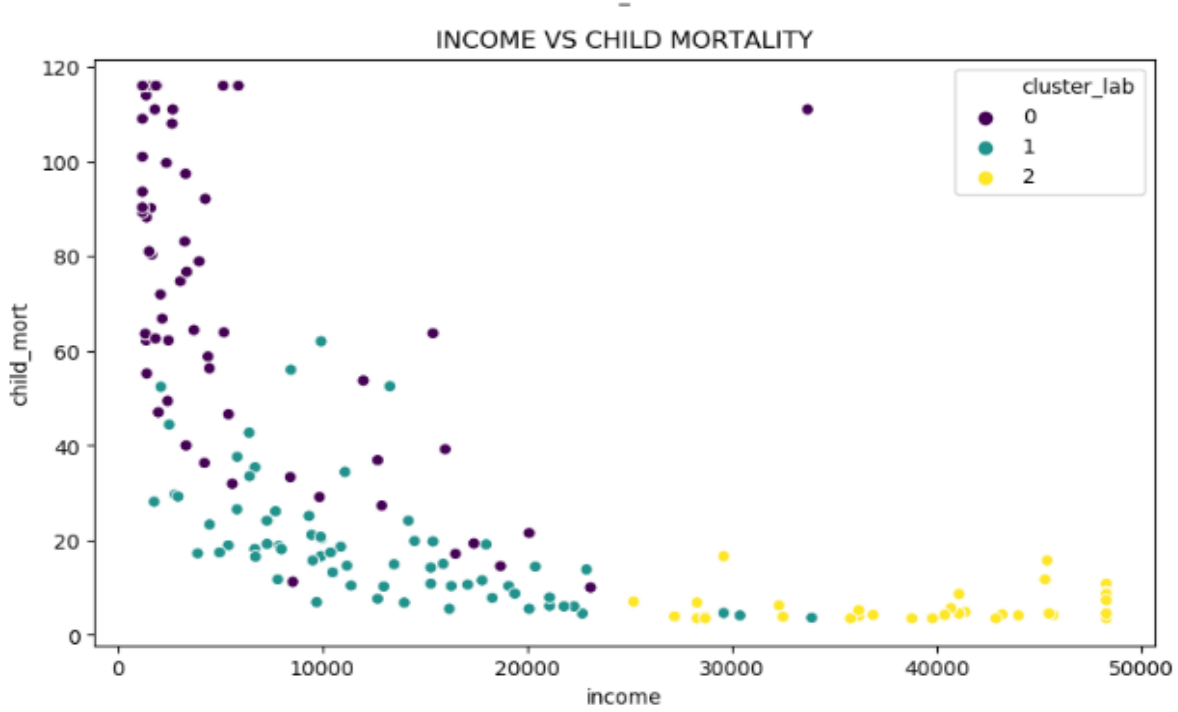
- Same as K-means



CLUSTER ANALYSIS
WRT ORIGINAL VARIABLES

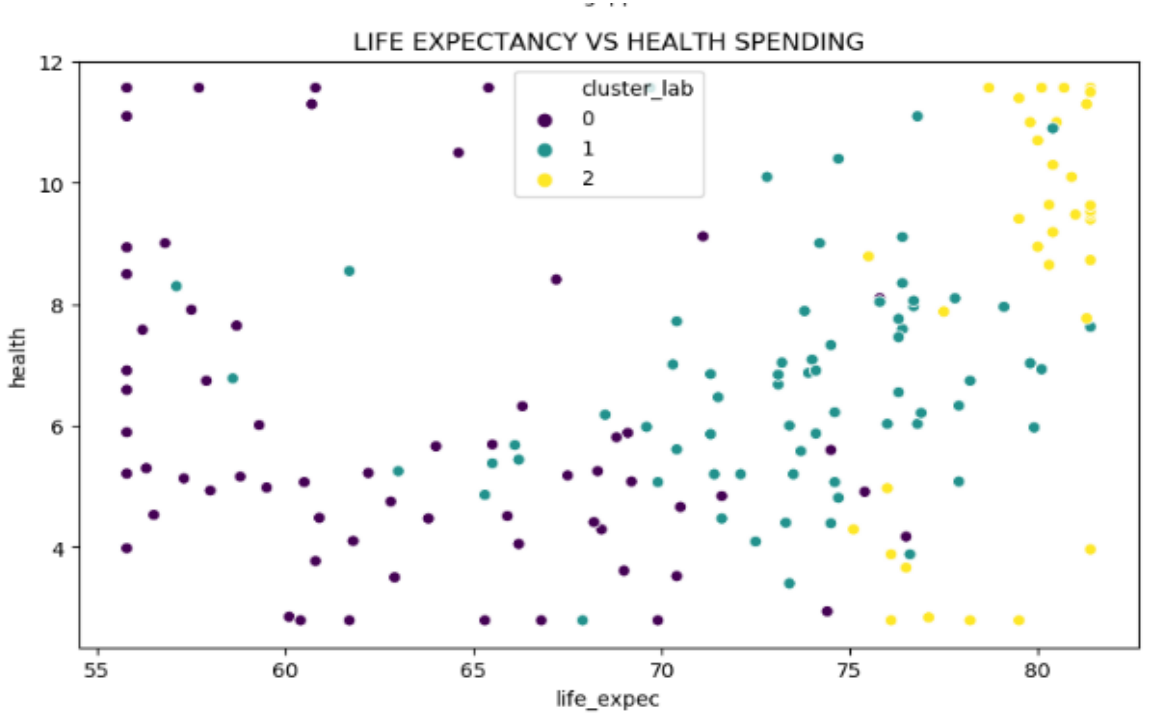
INCOME VS CHILD MORTALITY

- Same as k-means



LIFE_EXPEC VS HEALTH

- No definite pattern observed



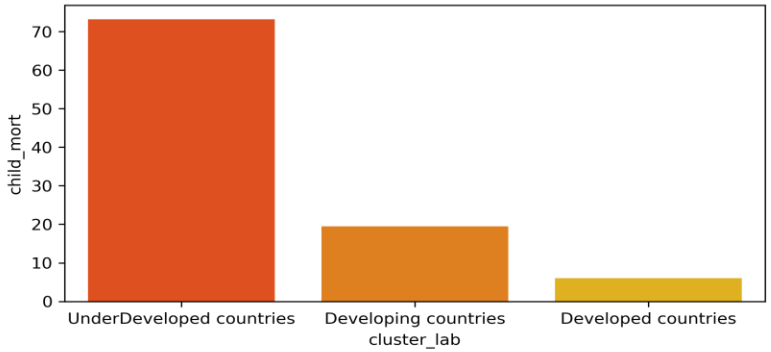
MEAN ANALYSIS OF CLUSTERS

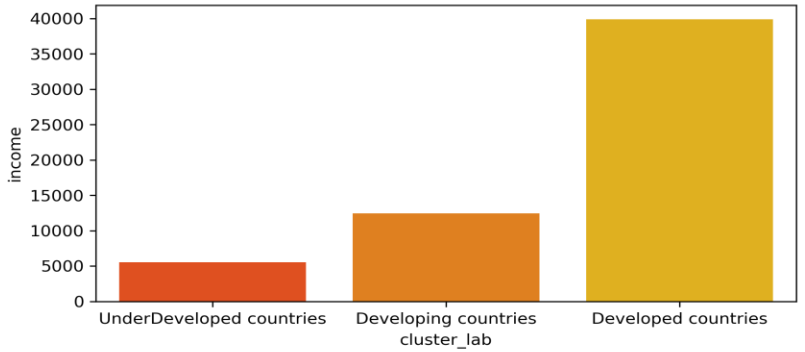
- Mean data is taken for all countries for each clusters
- Analysing each variable by its mean and clusters will give clear bifurcation of under-developed countries form the rest
- Cluster_lab 's have been renamed for visualisation needs
  - cluster_lab 0 -> UNDER - DEVELOPED COUNTRIES
  - cluster_lab 2 -> DEVELOPED COUNTRIES
  - cluster_lab 1 -> DEVELOPING COUNTRIES

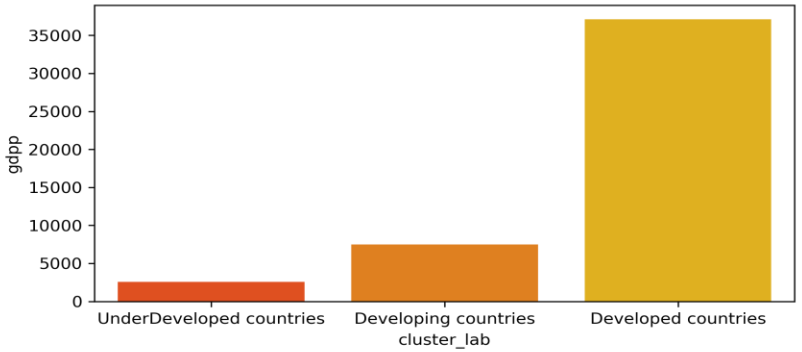
	cluster_lab	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	73.168254	28.398413	5.941349	37.554444	5540.111111	11.023254	63.297143	4.275698	2549.001587
1	1	19.438235	42.994559	6.679721	52.964412	12446.029412	4.889282	73.176471	2.245441	7474.176471
2	2	5.990000	51.449444	8.378694	44.675556	39883.611111	4.020789	79.580556	1.832778	37105.277778

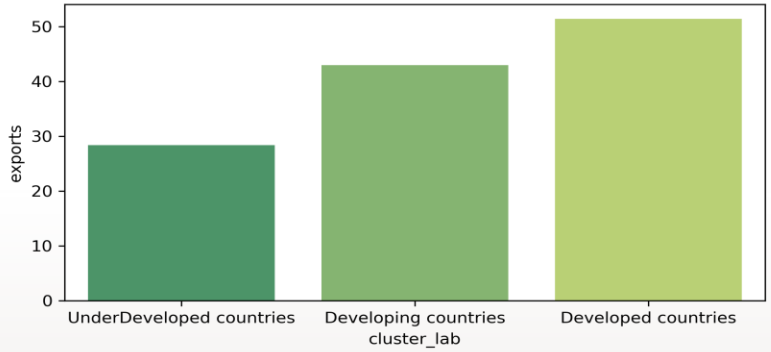
	cluster_lab	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	UnderDeveloped countries	73.168254	28.398413	5.941349	37.554444	5540.111111	11.023254	63.297143	4.275698	2549.001587
1	Developing countries	19.438235	42.994559	6.679721	52.964412	12446.029412	4.889282	73.176471	2.245441	7474.176471
2	Developed countries	5.990000	51.449444	8.378694	44.675556	39883.611111	4.020789	79.580556	1.832778	37105.277778

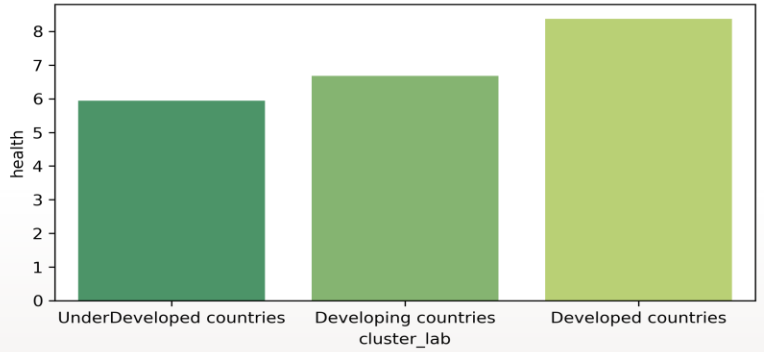
MEAN ANALYSIS OF CLUSTERS

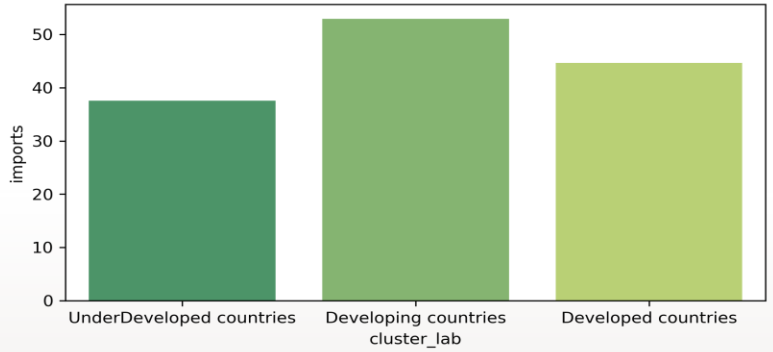


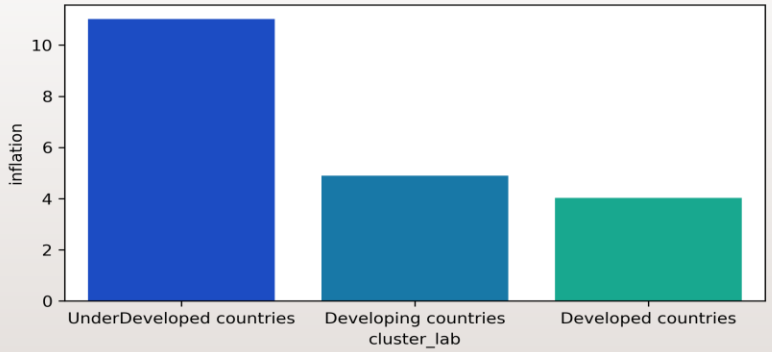


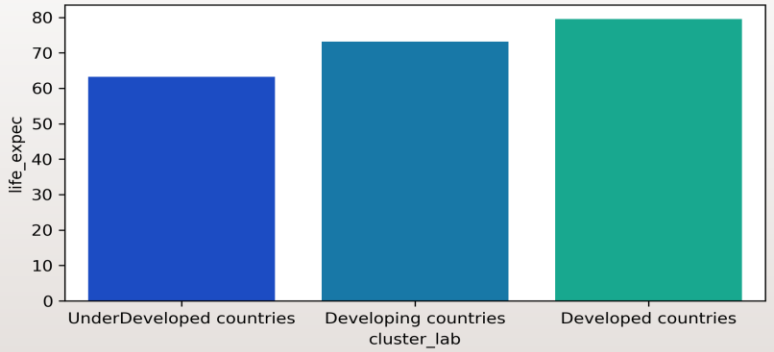


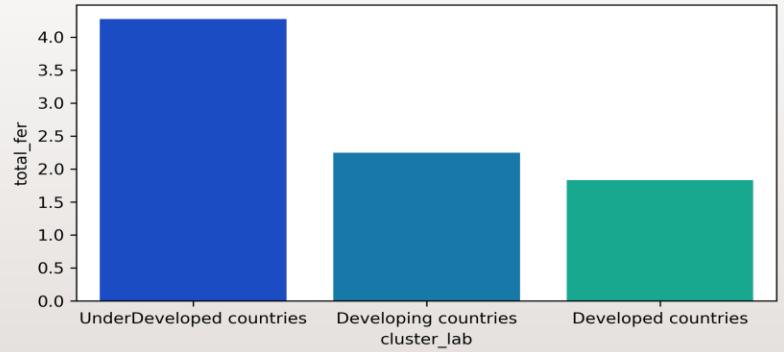












MEAN ANALYSIS OF CLUSTERS

- From the mean analysis of cluster of all variables in hierarchical clustering, we can observe that its same as kmeans:
- Only gdpp, income & child_mort show considerable difference in clusters.
- Developed countries:
  - HIGH - gdp, income, health spending, exports, life expectancy
  - Avg - imports
  - VERY LOW - child mortality, inflation, total fertility
- Developing countries:
  - AVG - health spending,exports, life expectancy, total fertility,child mortality, inflation,income
  - HIGH - imports
  - LOW – gdp
- **Under-Developed countries:**
  - VERY LOW - gdp, income,exports, life expectancy,Imports
  - Avg - health spending
  - HIGH - child mortality, inflation, total fertility

TOP 10 COUNTRIES – CLUSTER WISE

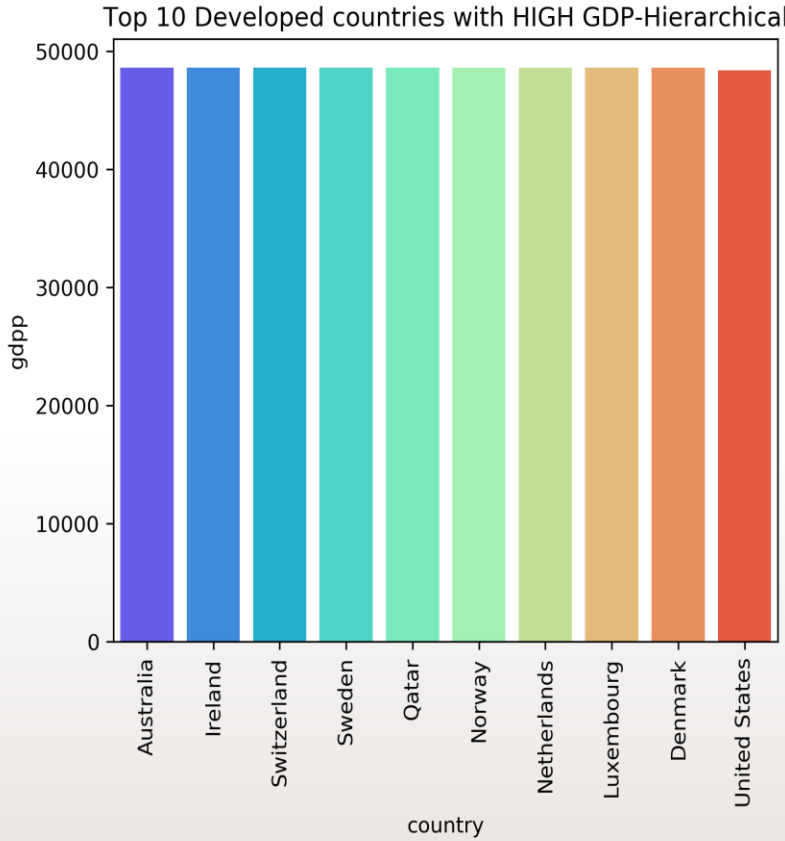
- 3 new data frames subsetted for each cluster
- groupby & sort_values functions are used instead of binning
- Top 10 countries are grouped and sorted with respect to their gdp, income per person & child mortality

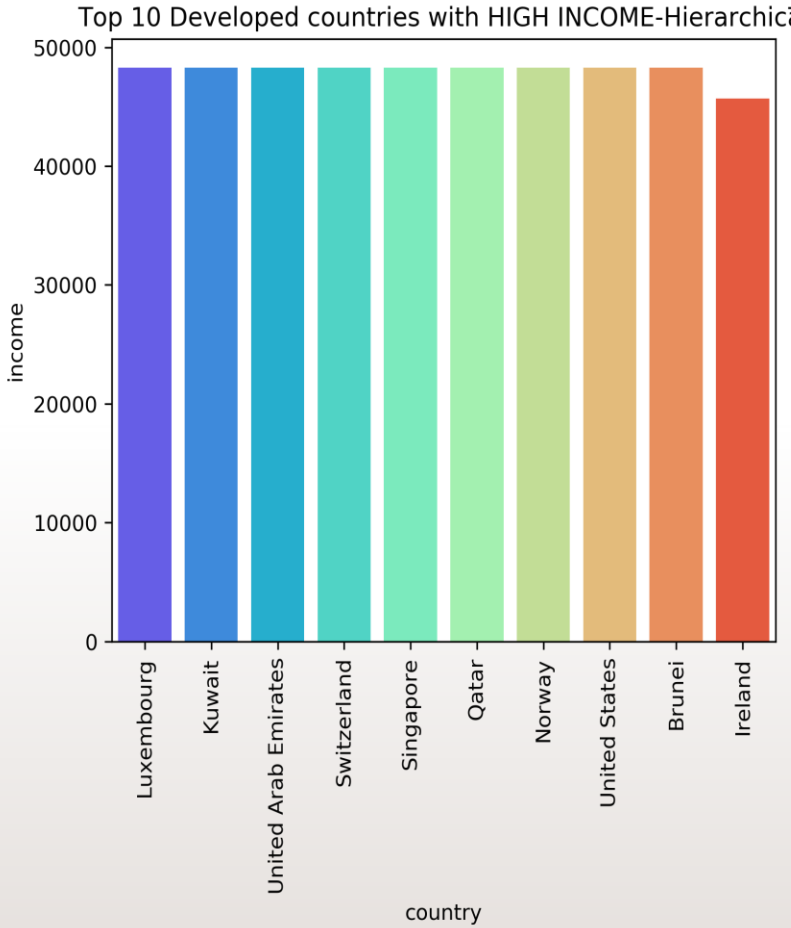
```
print(developed_hie.shape)
print(developing_hie.shape)
print(UnderDeveloped_hie.shape)
```

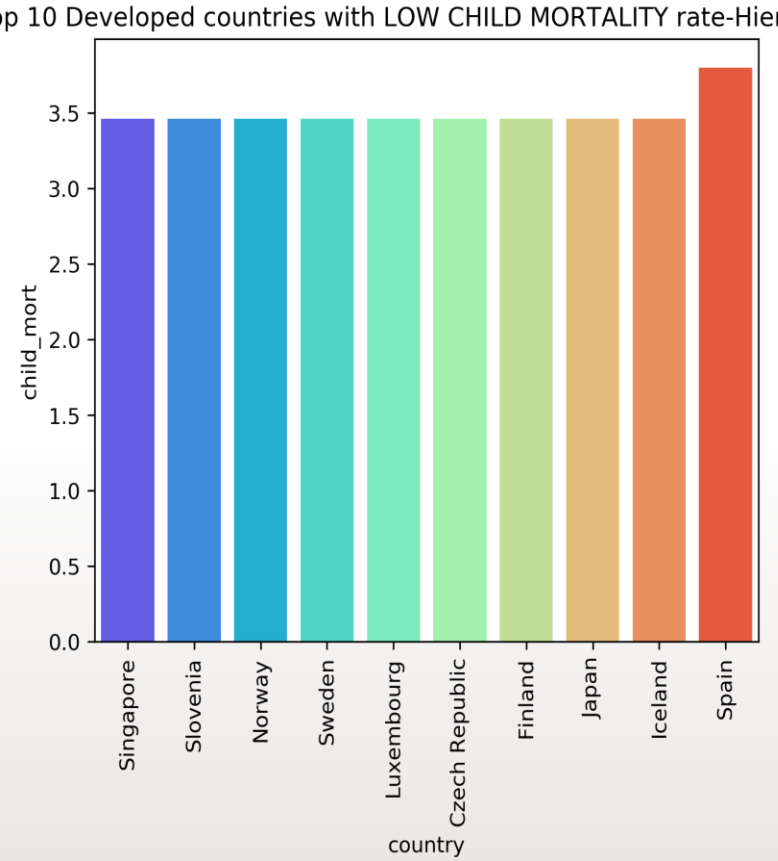
```
(36, 11)
(68, 11)
(63, 11)
```

TOP 10 DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







TOP 10 DEVELOPED COUNTRIES

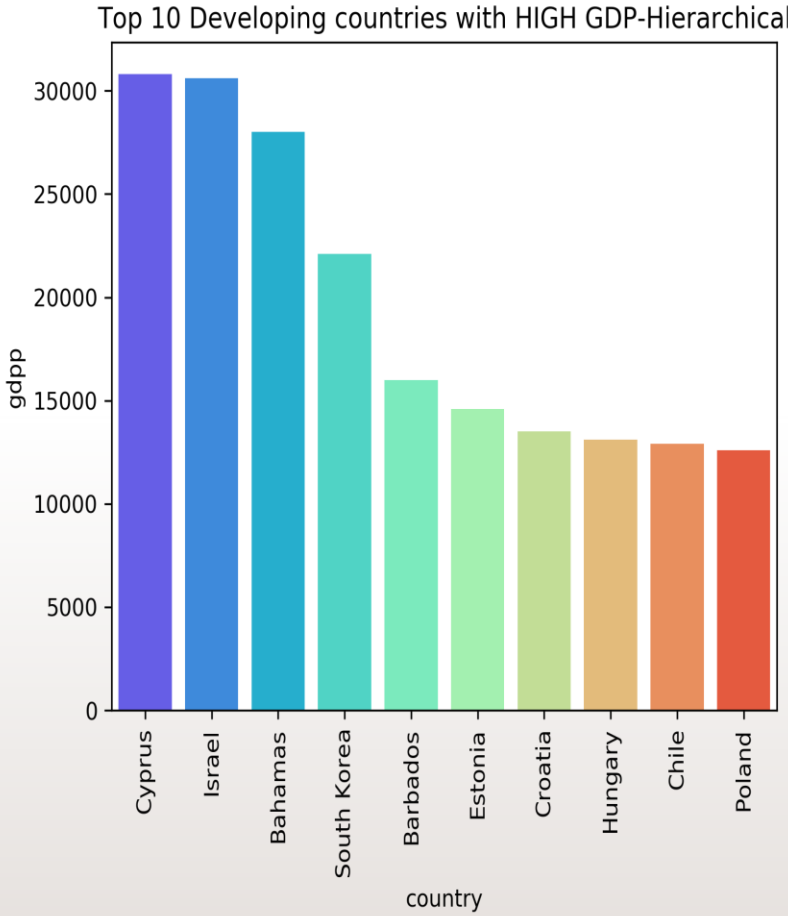
WRT GDPP,INCOME & CHILD_MOR

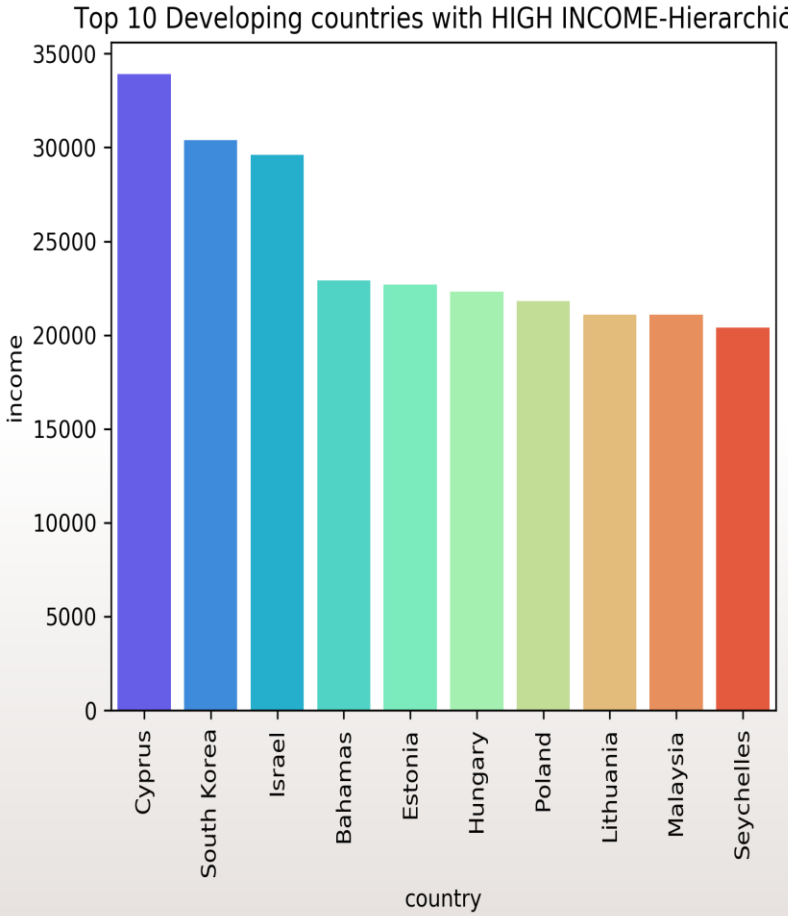
- The developed countries have very high gdp and income per person and very low child mortality rate
- Below are the top developed countries & same as K-means:

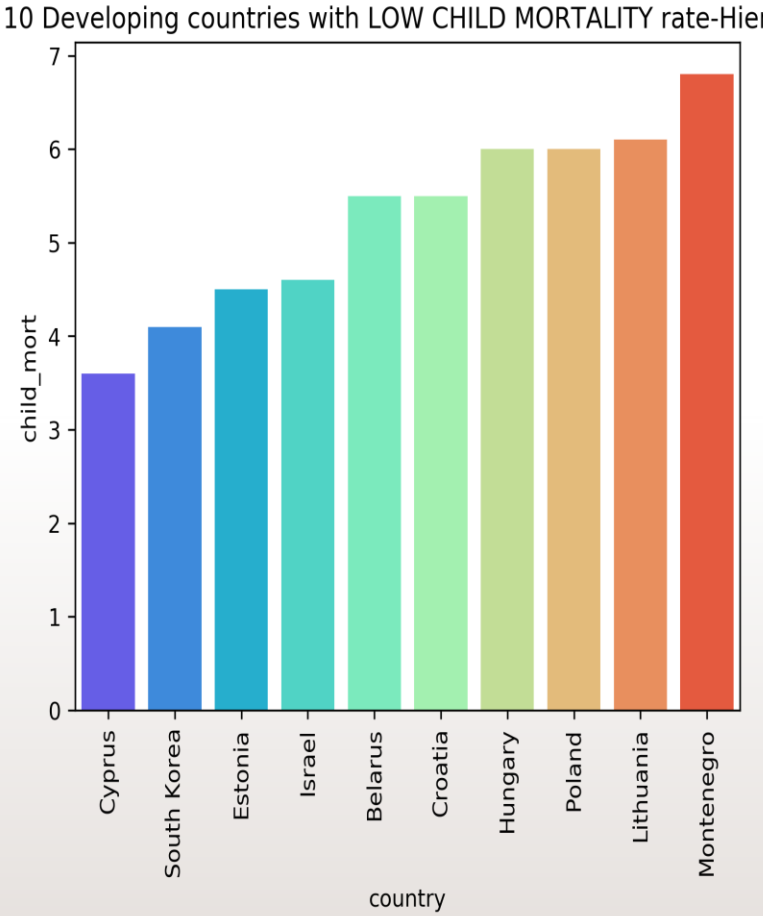
- Australia
- Denmark
- Switzerland
- Sweden
- Qatar
- Norway
- Netherlands
- Luxemburg
- Ireland
- United States
- UAE
- Kuwait
- Singapore
- Iceland
- Japan
- Finland

TOP 10 DEVELOPING COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







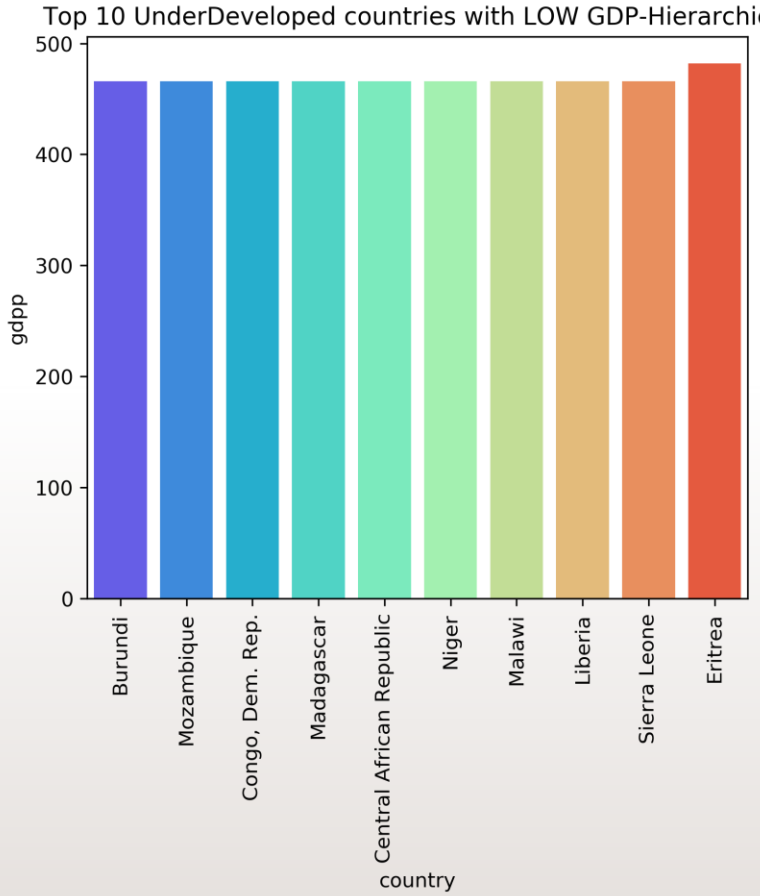
TOP 10 DEVELOPING COUNTRIES

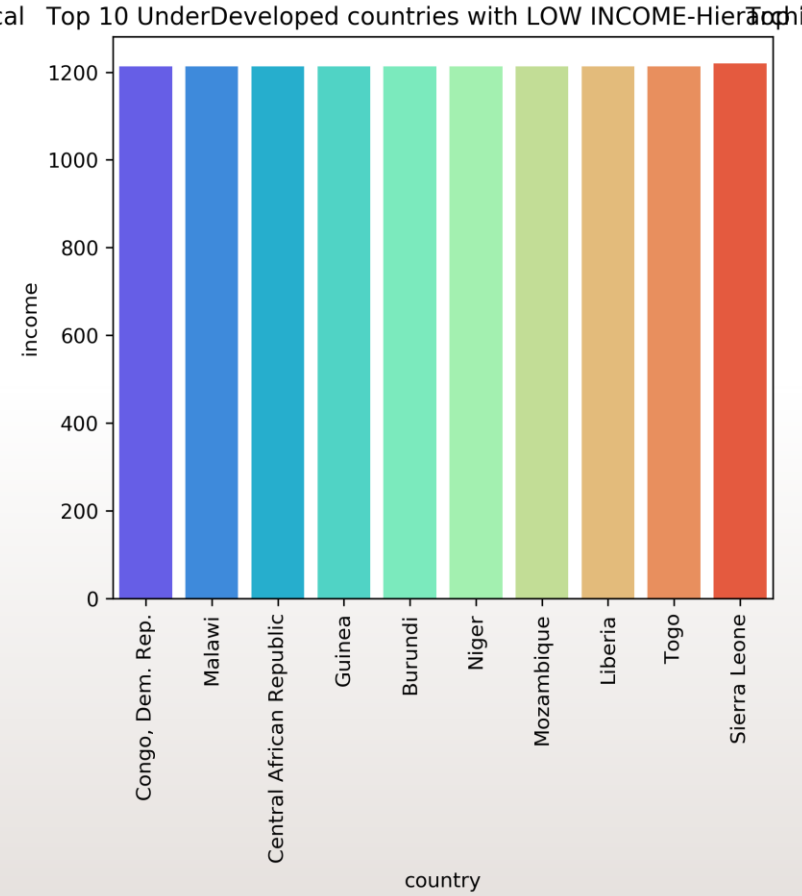
WRT GDPP,INCOME & CHILD_MOR

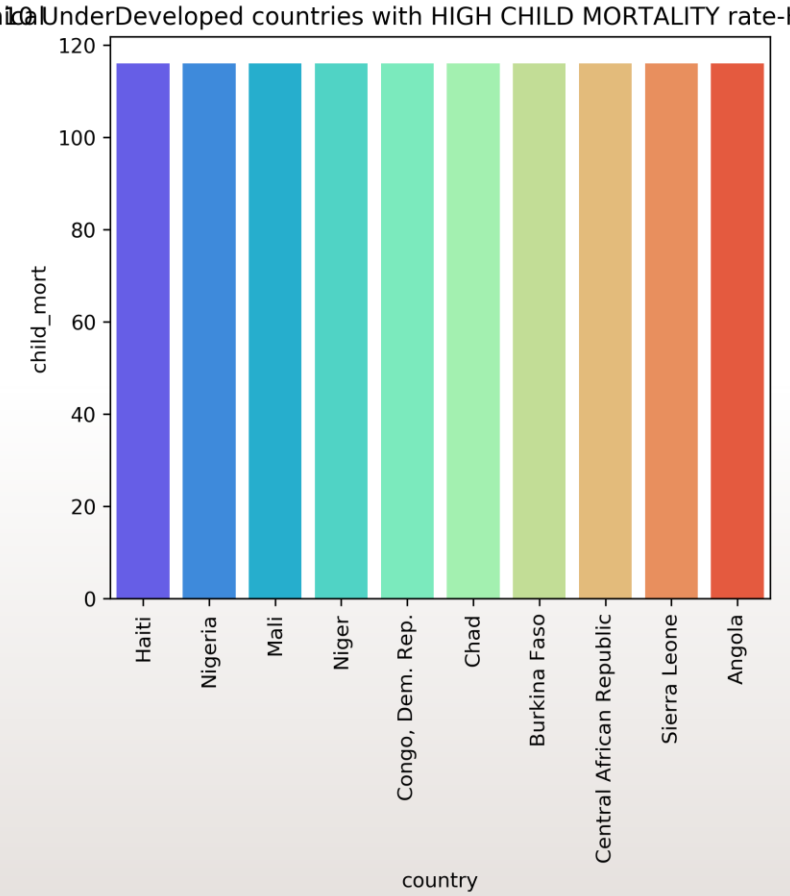
- The developing countries have very low gdp and income per person and average child mortality rate
- Below are the top developing countries & vary slightly from kmeans clustering:
  - CYPRUS
  - Israel
  - Bahamas
  - South Korea
  - Estonia
  - Croatia
  - Hungary
  - Chile
  - Poland
  - Malaysia
  - Lithuania
  - Barbados
  - Belarus

TOP 10 UNDER DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR







TOP 10 UNDER DEVELOPED COUNTRIES

WRT GDPP,INCOME & CHILD_MOR

- Hierarchical gave same results as k-means for cluster of under developed countries
- The below countries have very low gdp and income per person and very high child mortality rate

- Burundi
- Malawi
- Niger
- Mozambique
- Central African republic
- Sierre Leone
- Congo, Dem rep
- Liberia
- Eritrea
- Togo
- Guinea
- Madagascar
- Angola
- Mali
- Haiti
- Burkino Faso
- Chad
- Nigeria

SUMMARY

- The main aim of the exercise was to is to categorize the countries using some socio-economic and health factors that determine the overall development of the country
- After initial data inspection, cleaning, EDA & scaling, Principal component analysis was done and with the obtained principal components, K-means clustering and hierarchical clustering with single and complete linkage were performed
- K-means with K=3 perfectly differentiated 3 clusters Cluster 0 was developed countries, cluster 1 - developing countries & cluster 2 – Under-Developed countries
- Upon clustering using hierarchical method, countries were segmented into 3 clusters as cluster 2 - Developed, cluster 1- developing & cluster 0 - Under-Developed countries
- Among other factors, Income, GDP and child mortality showed huge difference between clusters in both k-means & hierarchical

K-MEANS VS HIERARCHICAL

- Both methods yielded the same results as below, only cluster labels were different:
  - Developed countries visualisations showed HIGH gdp and income, LOW child mortality rate
  - Developing countries showed low gdp, income and average child mortality rate
  - Under developed countries visualisations depicted VERY LOW gdp, income & VERY HIGH child mortality rate
- Top 10 countries were the same in both methods only with difference of few developing countries
- **As hierarchical dendrogram clearly showed 3 different clusters, am going ahead with results of hierarchical clustering**

RECOMMENDATIONS

- Upon above analysis, Below are the UNDER-DEVELOPED countries which are in dire need of aid and i suggest the below 18 UNDER-DEVELOPED countries for HELP INTERNATIONAL CEO to focus the most:

- Burundi
- Malawi
- Niger
- Mozambique
- Central African Republic
- Sierre Leone
- Congo, Dem Rep
- Liberia
- Eritrea
- Togo
- Guinea
- Madagascar
- Angola
- Mali
- Haiti
- Burkino Faso
- Chad
- Nigeria

THANK YOU