PRODEGREE DATA SCIENCE
MAJOR PROJECT

Created by: **ROHAN KOKKULA**
rohankokkula01@gmail.com

TABLE OF CONTENT

Objective:

To categorise the countries using socio-economic and health factors that determine the overall development of the country by using K-means and Hierarchical Clustering.

Problem Statement:

The Dataset contains list of Countries with their socio-economic and health factors. With the help of Unsupervised Learning, you need to categorize these countries that are in the direst need of aid. Submit the produced list of countries(minimum 5) to the CEO by selecting either K-means or Hierarchical Clustering Method.

Viewing and Understanding Data

First 5 Rows of the dataset.

	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

Description of the features.

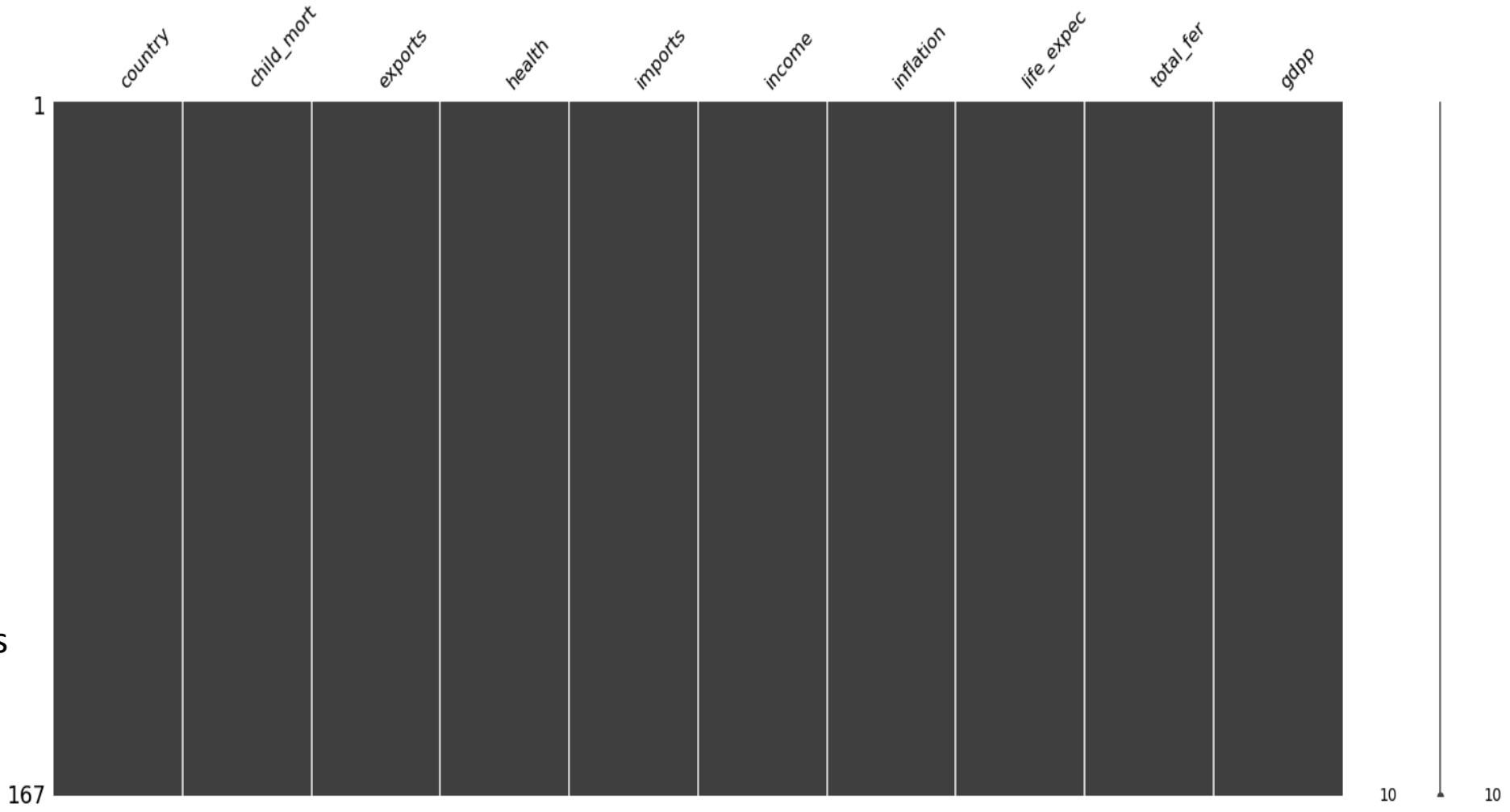
Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as %age of the GDP per capita
health	Total health spending per capita. Given as %age of GDP per capita
imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Visualizing Missing Data

Null values:		NaN values:	
country	0	country	0
child_mort	0	child_mort	0
exports	0	exports	0
health	0	health	0
imports	0	imports	0
income	0	income	0
inflation	0	inflation	0
life_expec	0	life_expec	0
total_fer	0	total_fer	0
gdpp	0	gdpp	0

There are no **Null** or **NaN** values

Matrix Plot



Datatypes, Duplicates & Describing Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   country     167 non-null    object
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64
dtypes: float64(7), int64(2), object(1)
```

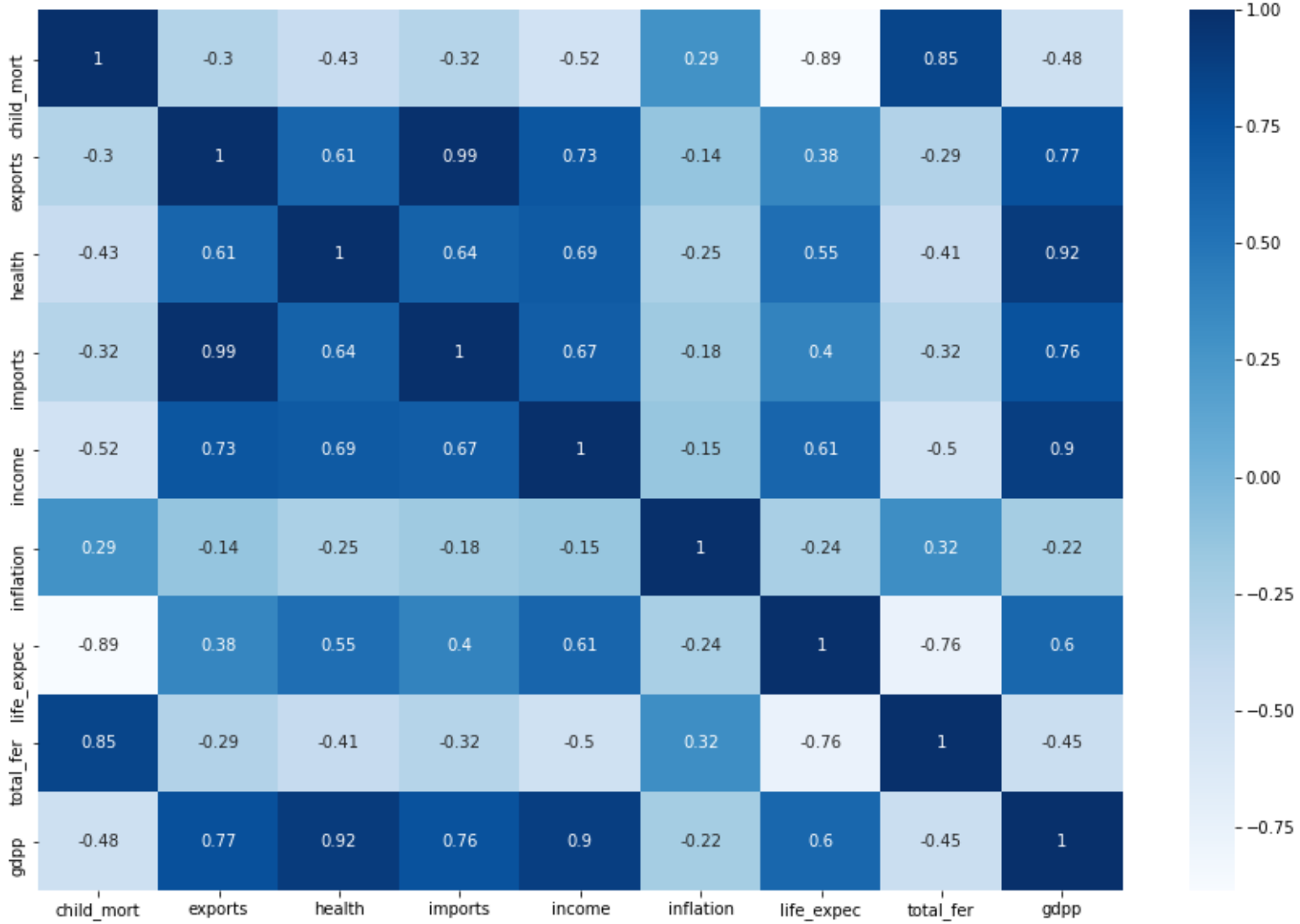
All the Data types are in correct format

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	7420.618847	1056.733204	6588.352108	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	17973.885795	1801.408906	14710.810418	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	1.076920	12.821200	0.651092	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	447.140000	78.535500	640.215000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	1777.440000	321.886000	2045.580000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	7278.000000	976.940000	7719.600000	22800.000000	10.750000	76.800000	3.880000	14050.000000
90%	100.220000	17760.600000	3825.416000	15034.280000	41220.000000	16.640000	80.400000	5.322000	41840.000000
95%	116.000000	31385.100000	4966.701000	24241.560000	48290.000000	20.870000	81.400000	5.861000	48610.000000
99%	153.400000	64794.260000	8410.330400	55371.390000	84374.000000	41.478000	82.370000	6.563600	79088.000000
max	208.000000	183750.000000	8663.600000	149100.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

```
There are 0 duplicates in dataset
```

Checking for outliers by describing percentiles and min max values

Correlation Heatmap



INSIGHTS FROM CORRELATION MAP

1. **Exports** is highly correlated with **Imports.**

2. **Health, Exports, Income, Imports** are highly correlated with **GDP per capita.**

3. **Child Mortality Rate** is having high negative correlation with **Life Expectancy.**

4. **Total fertility** is highly positively correlated with **Child Mortality Rate** and negatively correlated with **Life Expectancy.**

Univariate Analysis

INSIGHTS FROM BOXPLOTS

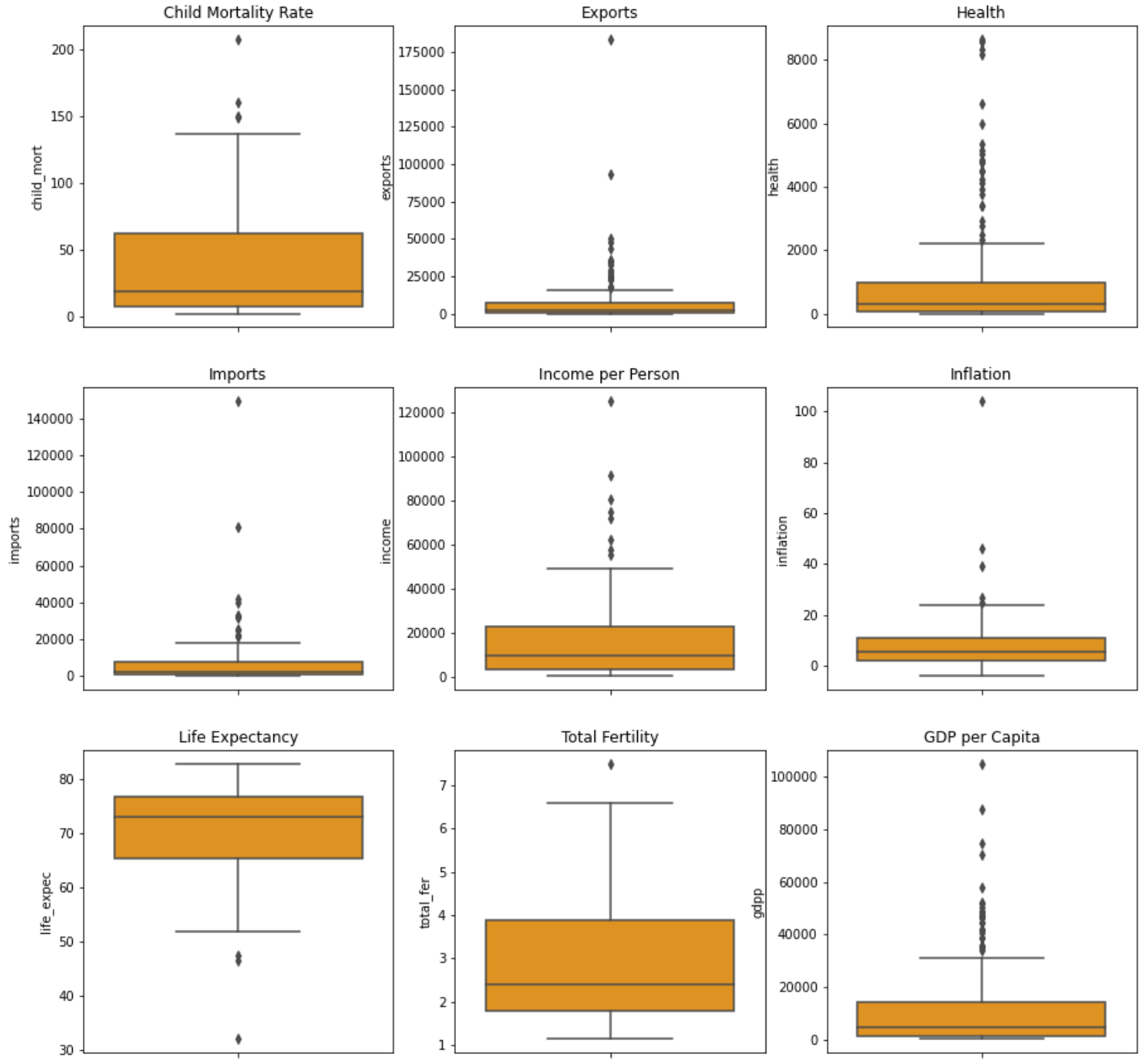
1. There is **minimum one outlier** in each of the numerical features.

2. Most Outliers can be seen in **GDPP**.

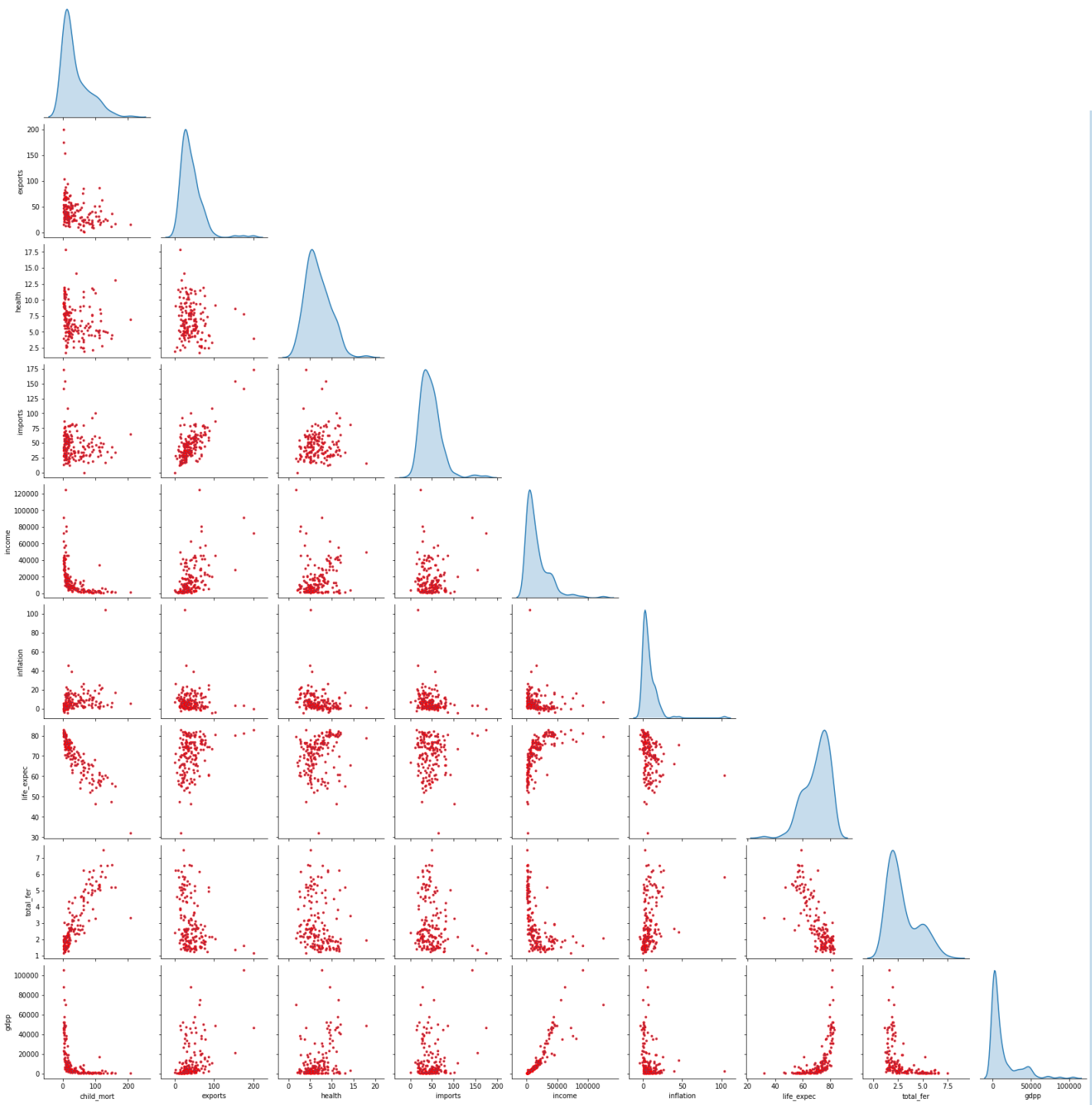
3. As our data contains only **167 countries**, Removing these outliers could increase chances of **removing dire needy** countries.

4. Example, In case of **Child Mortality Rate**, Country with **208** value is being specified as outlier but that country itself could be in **dire need of aid.!**

5. Removing outliers is **NOT** a good option as per the above conditions.
Hence, I choose to **KEEP** outliers.



Bivariate Analysis



INSIGHTS FROM PAIRPLOT

1. Univariate Analysis(KDE)

1. Only **life expectancy** is **right-skewed** whereas all the rest features are left-skewed.

2. Total Fertility and **GDPP** are bimodal whereas all the rest features are unimodal.

2. Bivariate Analysis

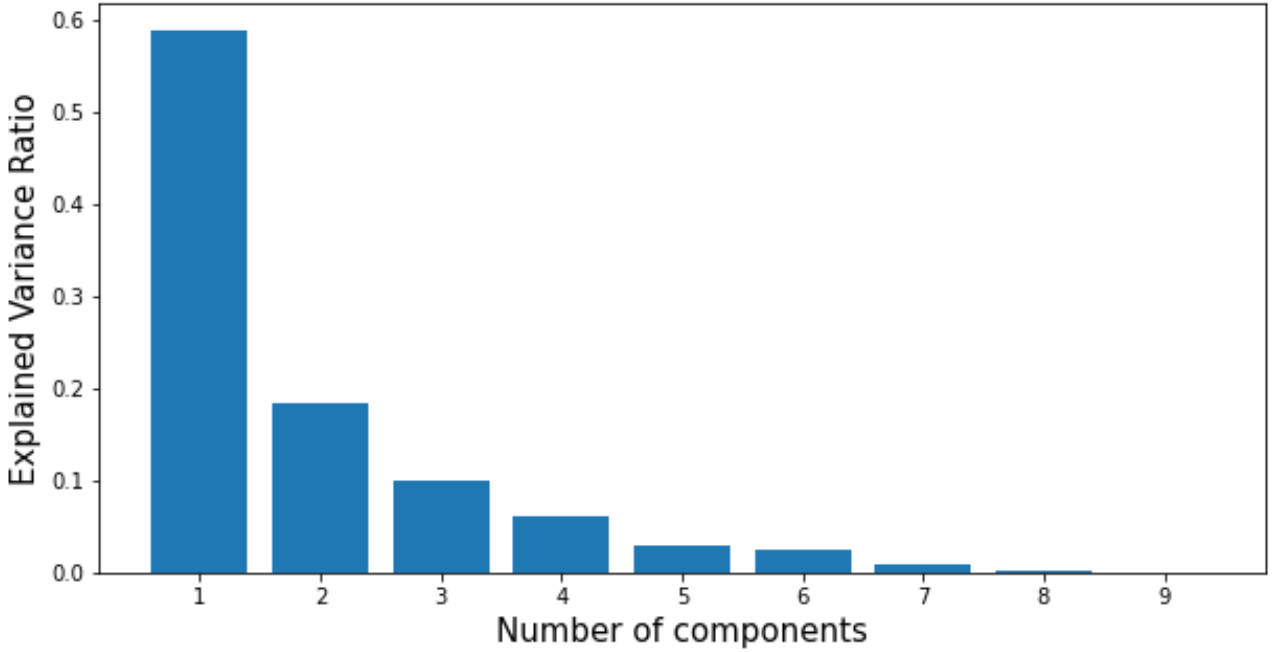
1. Linear relationship is found between **[gdpp – income], [imports – exports], [total_fer - child_mort]**

2. If **GDPP** is **HIGH**:

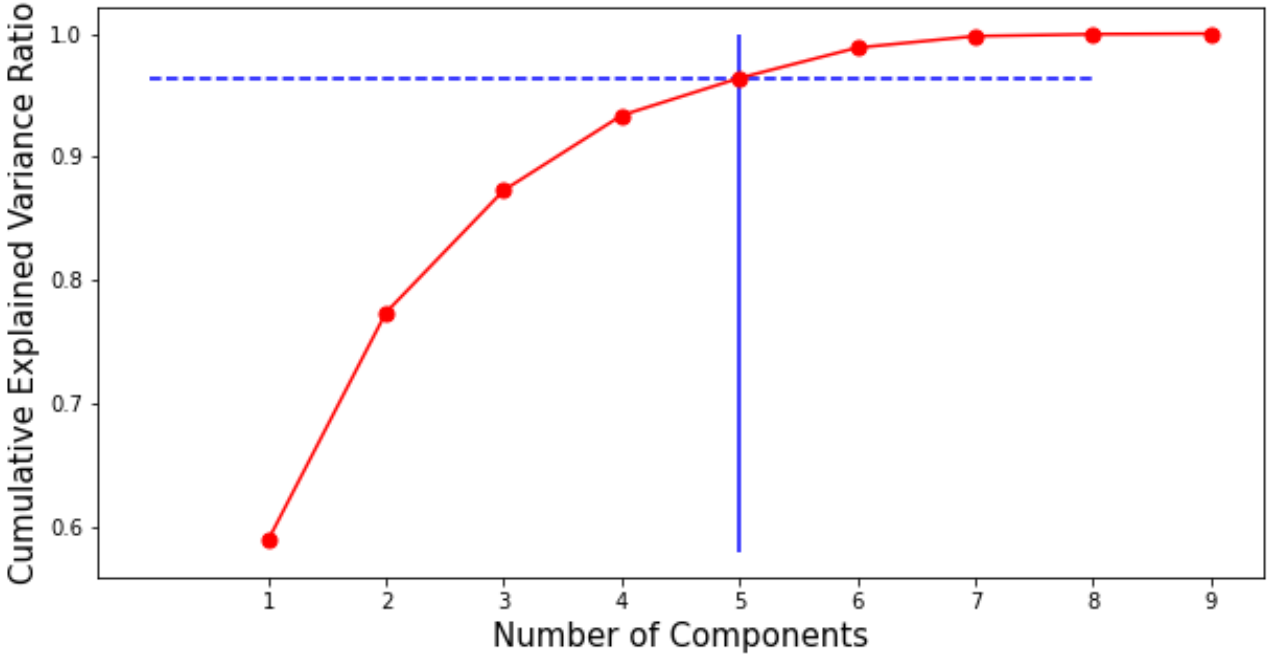
child mortality is **LOW**
income is **HIGH**
inflation is **LOW**
life expectancy is **HIGH**
total fertility is **LOW**
health, imports and **exports** are **AVG**

Principal Component Analysis

BARPLOT

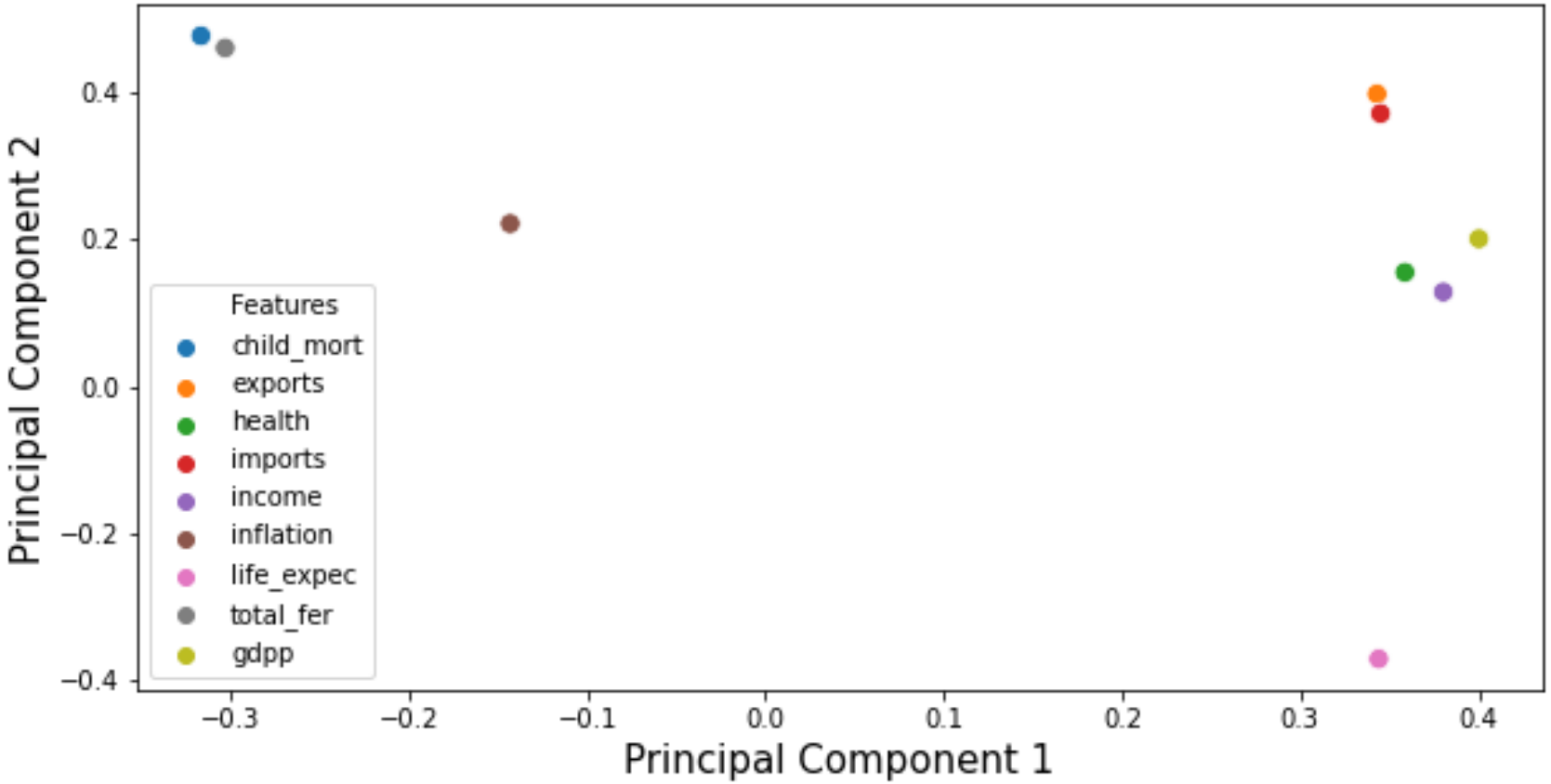


SCREE PLOT



From above plots, Around **96%** of the information is being explained by **5** principal components.

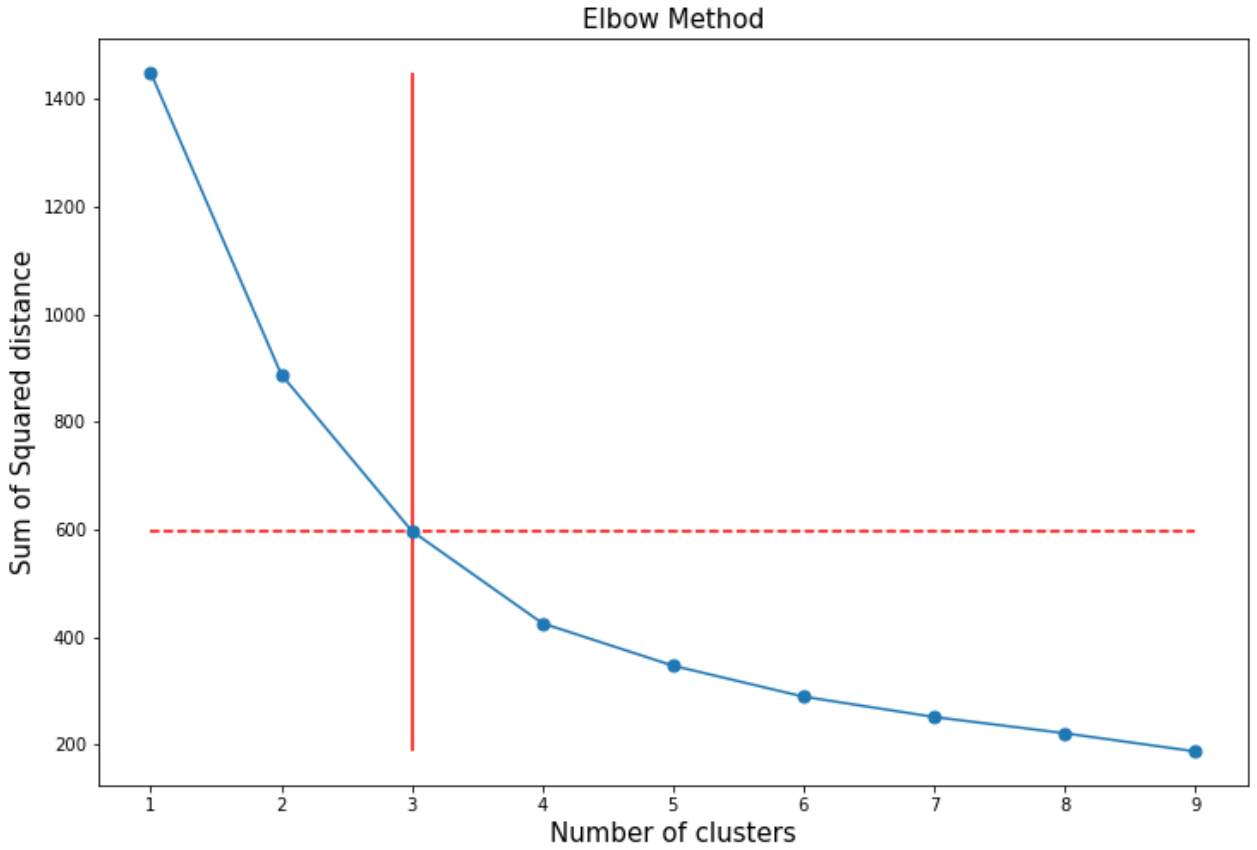
Visualizing 2 Principal Components

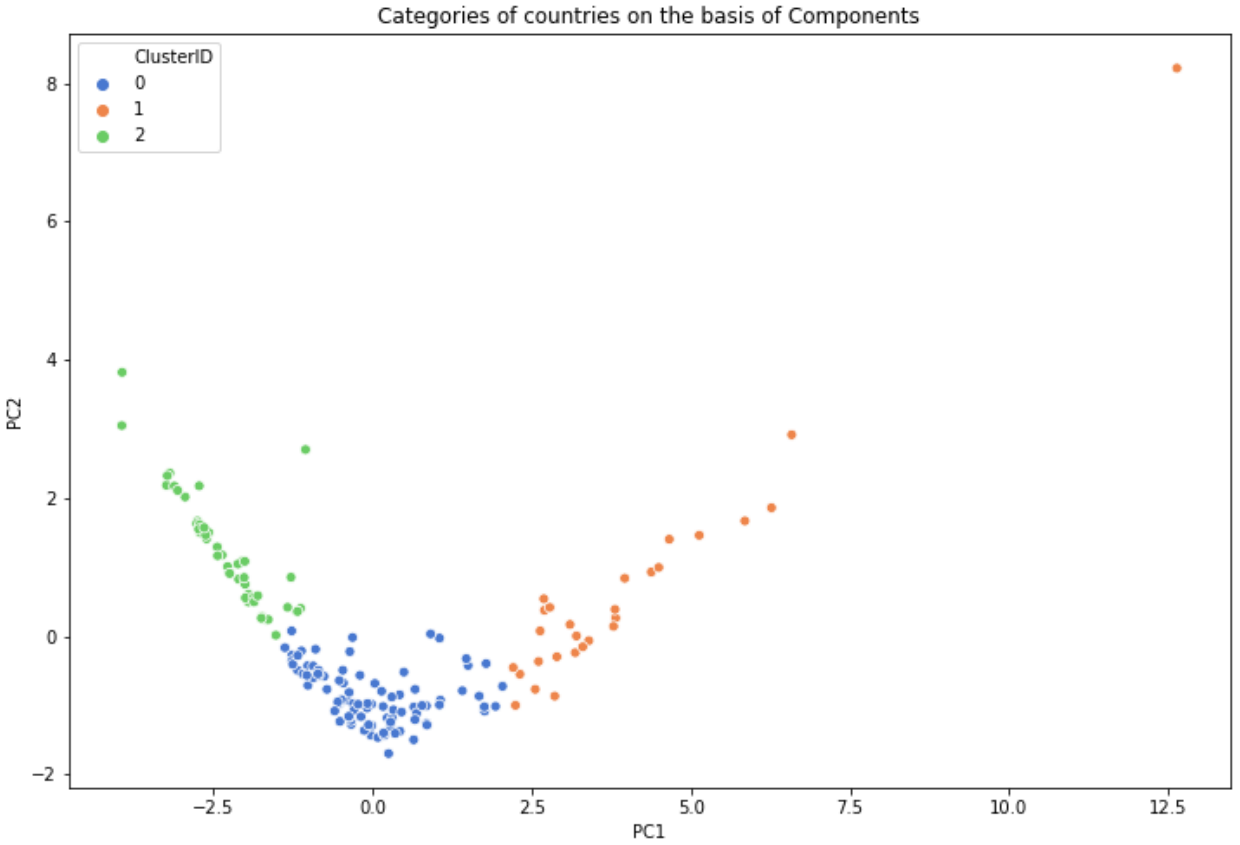


1. From the above plot, we can see the **first component** is in the direction where the **imports, exports, gdpp, income, health, life_expec** are heavy and **second component** is in the direction where **child_mort , total_fer** is more.

2. If we recall, correlation between **imports** and **exports** was **0.99**. Now we can surely confirm it by looking the above plot.

K Means Clustering Algorithm

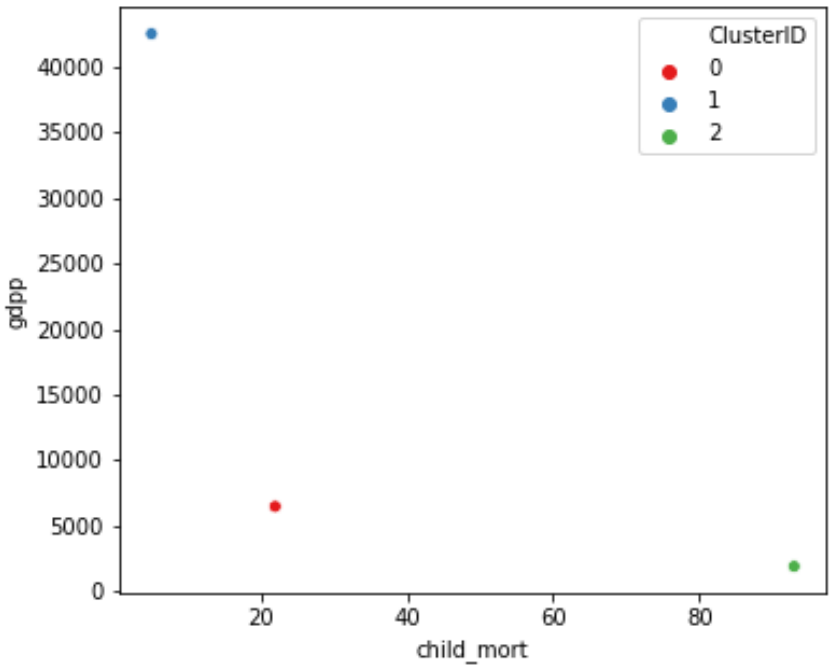


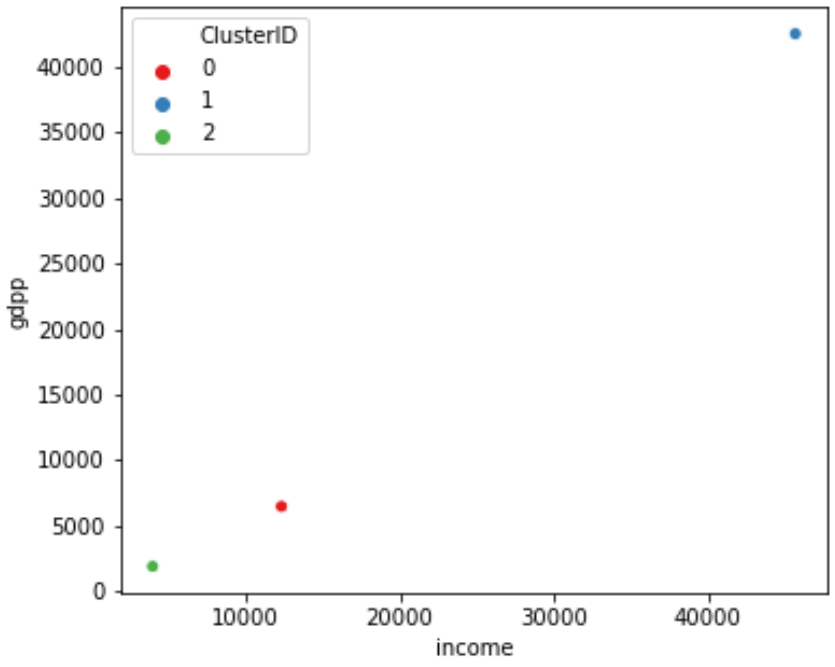


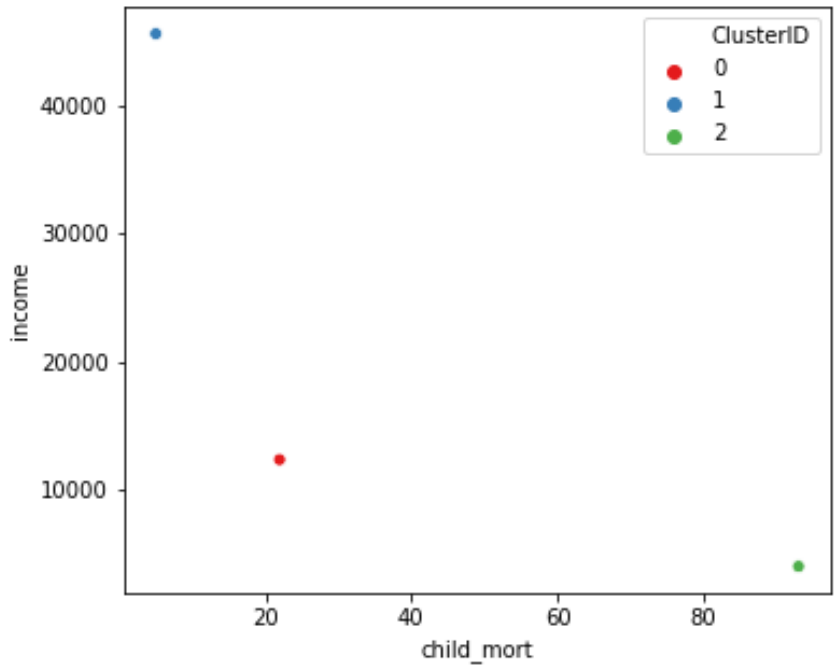
As per **Elbow** method, we'll select no. of clusters as **3**

Plotting PC1 & PC2 scatterplot w.r.t ClusterID

Renaming Clusters base on Means





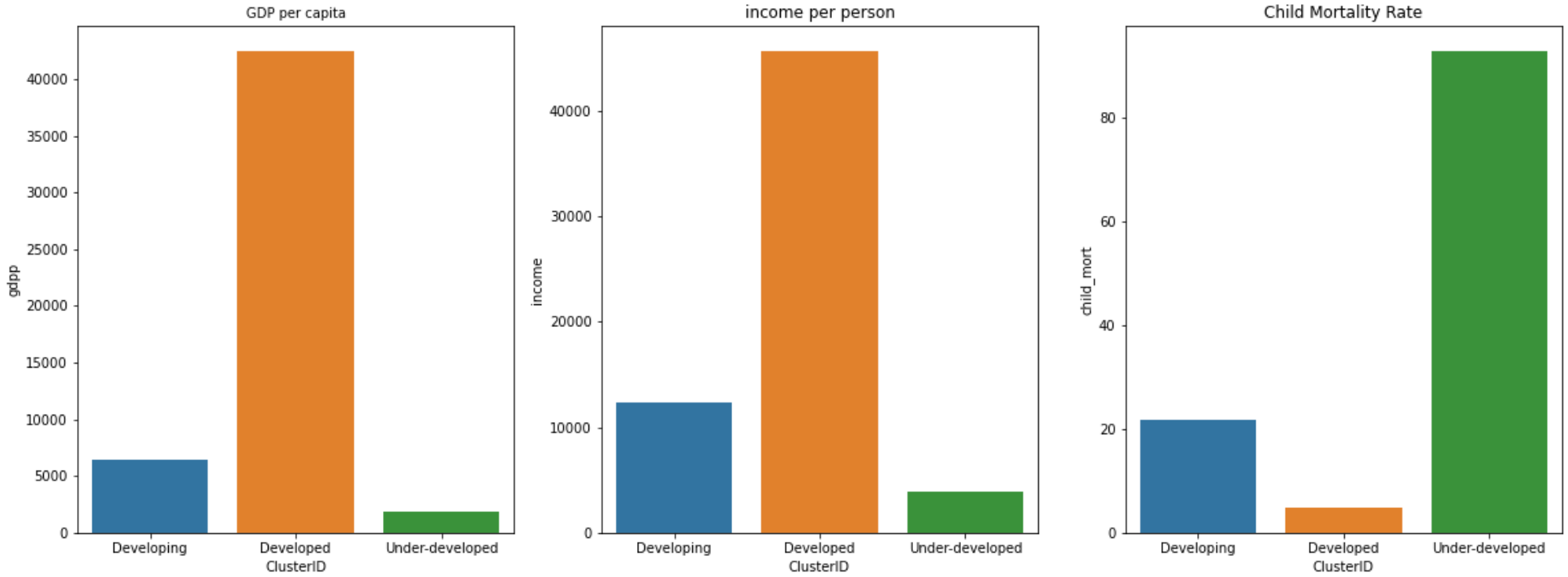


Renaming ClusterIDs as per gdpp, income & child mortality rate mean values.

1. Countries with **high GDPP** , **high Income** and **low Child Mortality Rate** are **Developed** countries
2. Countries with **average GDPP**, **average Income** and **average Child Mortality Rate** are **Developing** countries
3. Countries with **low GDPP**, **low Income** and **high Child Mortality Rate** are **Under-developing** countries

0 = Developing countries , 1 = Developed countries , 2 = Under-developed countries

Univariate Analysis on Clustered Countries

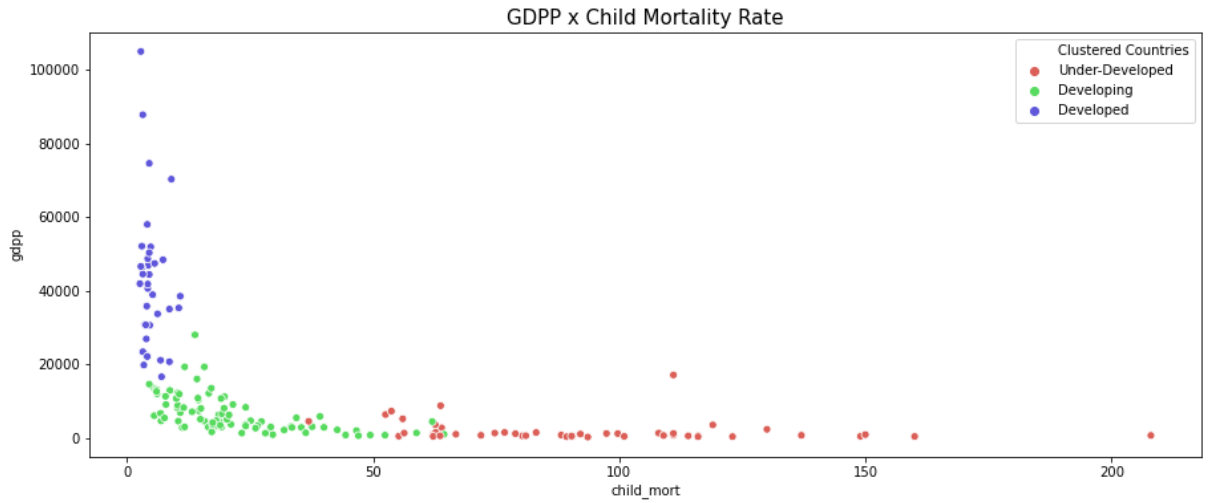


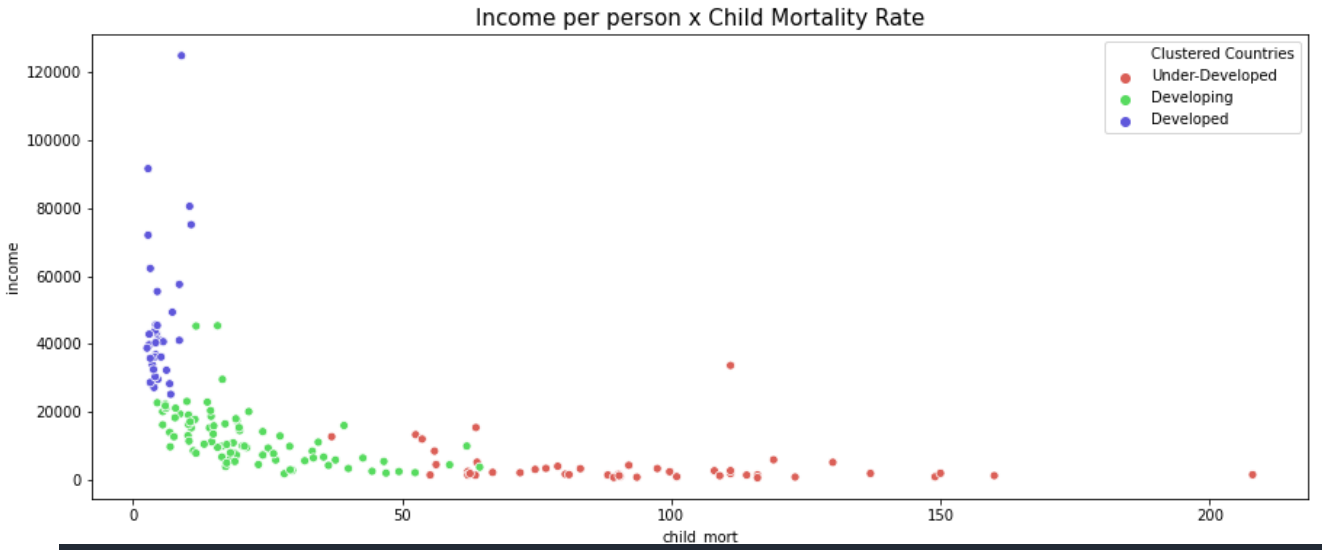
1. All the **developed** countries are having **high GDPP**, **developing** countries are having **average GDPP** and **Under-developed** countries are having the **least GDPP** values.

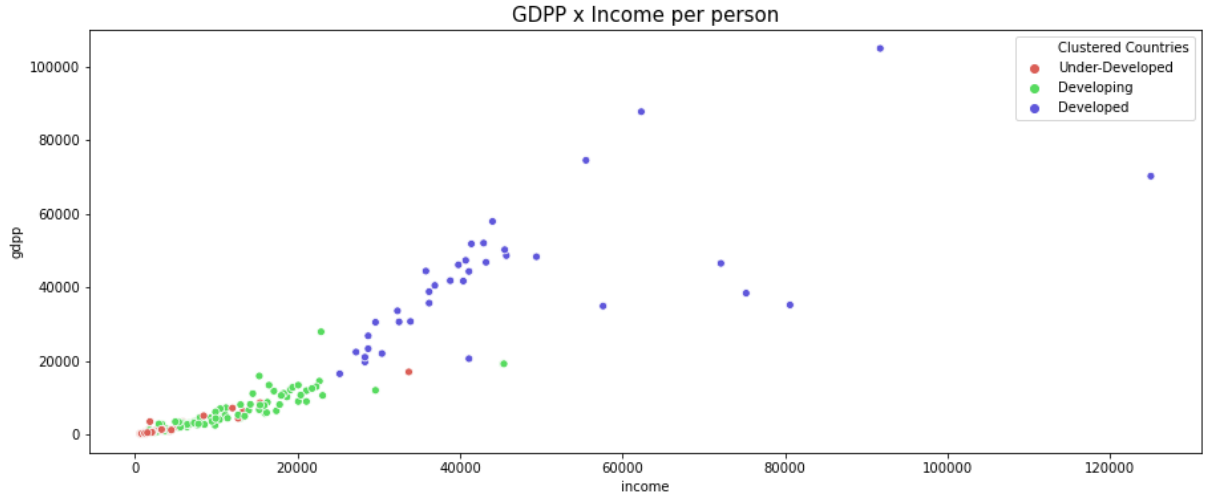
2. All the **developed** countries are having **high income** per person, **developing** countries are having **average income** per person and **Under-developed** countries are having the **least income** per person.

3. All the **developed** countries are having **low Child mortality rate**, **developing** countries are having **average child mortality rate** and **Under-developed** countries are having the **least child mortality rate**.

Bivariate Analysis on Clustered Countries





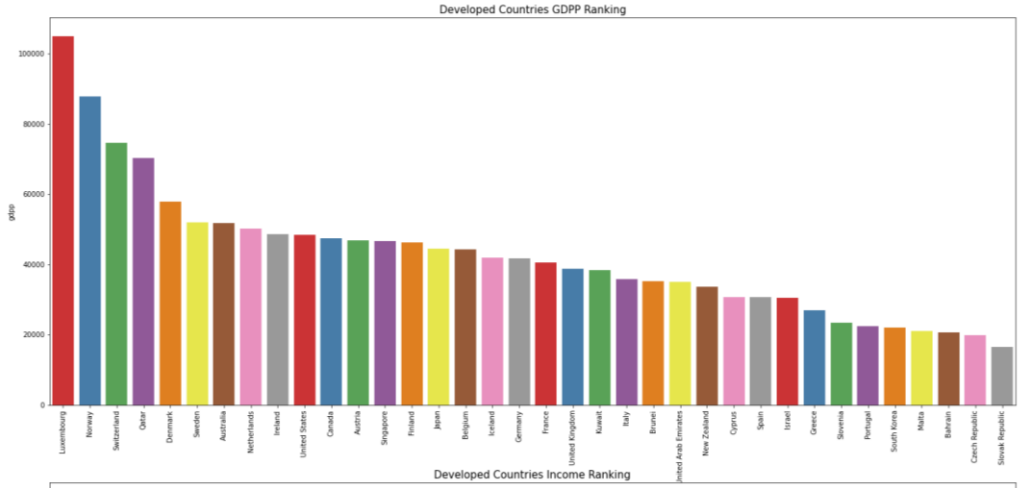


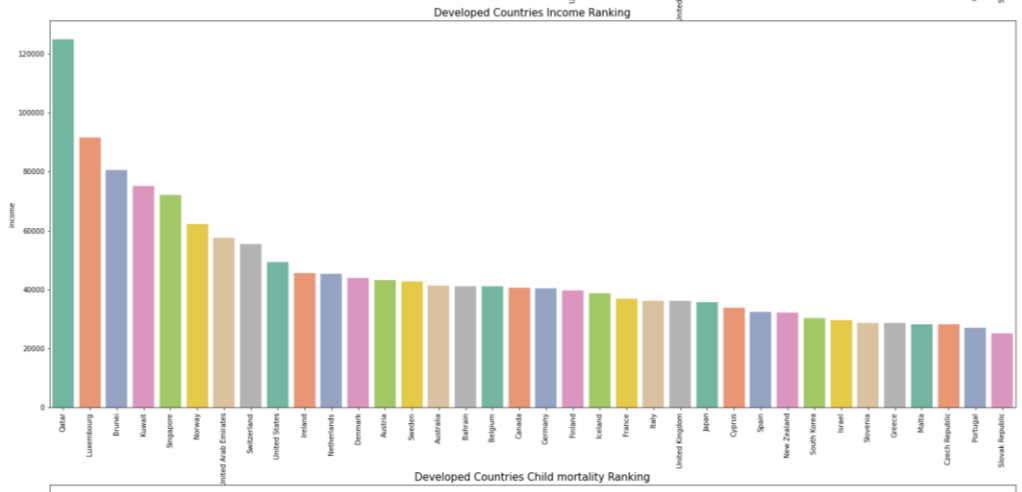
INSIGHTS FROM SCATTER PLOTS

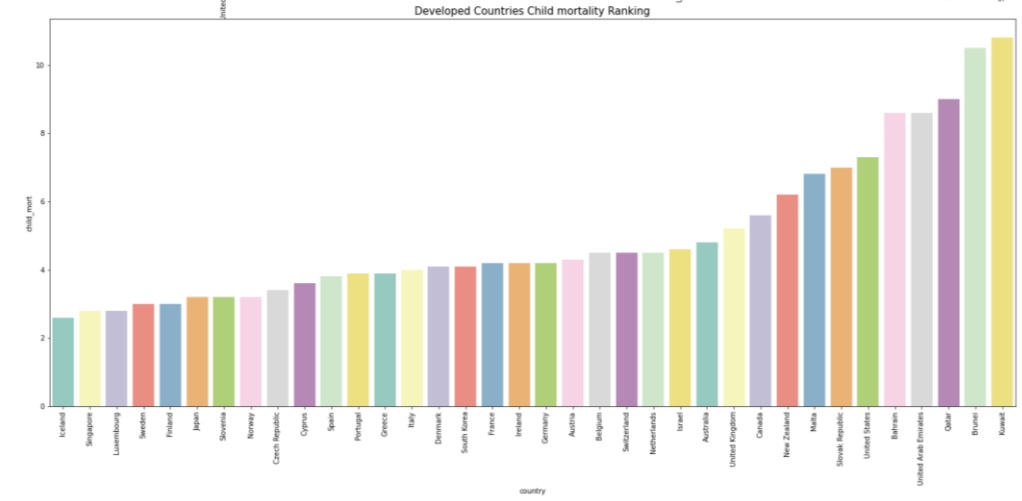
1. In **gdpp x child_mort**, there is some clustering where **gdpp is low**, there **child-mort is high**, which is true for **Under-developed** countries in reality.

2. In **gdpp x income**, there is some clustering where **gdpp is average**, there **income is average**, which is true for **Developing** countries in reality.

3. In **income x child_mort**, there is some clustering where if **income is high**, then **child mortality is low**, which is true for **Developed** countries in reality.







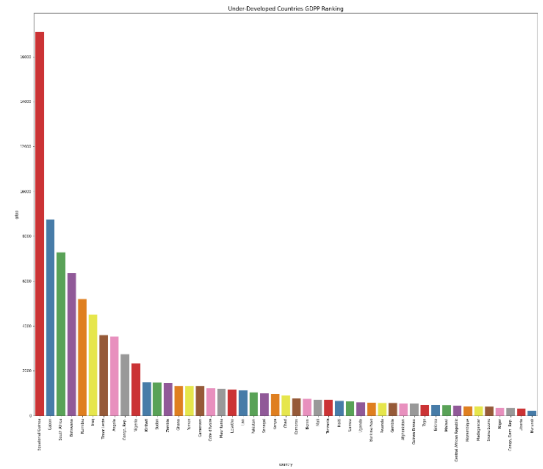
Developed Countries - 36

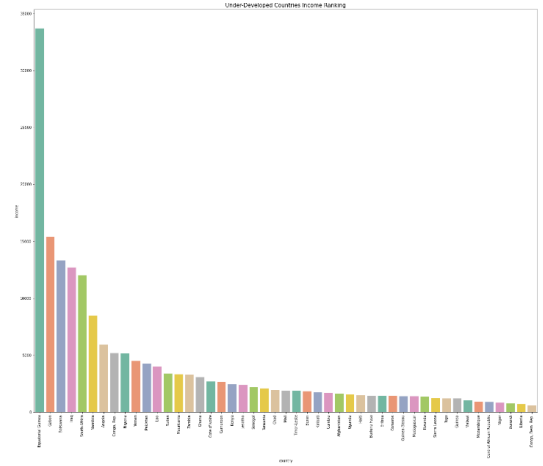
TOP **10** Developed Countries based on:

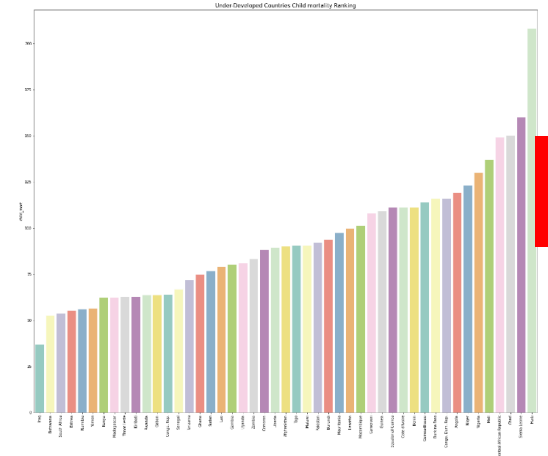
HIGH GDPP	**HIGH INCOME**	**LOW CHILD MORTALITY**
Luxembourg	Qatar	Iceland
Norway	Luxembourg	Singapore
Switzerland	Brunei	Luxembourg
Qatar	Kuwait	Sweden
Denmark	Singapore	Finland
Sweden	Norway	Japan
Australia	United Arab Emirates	Slovenia
Netherlands	Switzerland	Norway
Ireland	United States	Czech Republic
United States	Ireland	Cyprus

Note: The subplots can be clearly seen in the code outputs.

Under-Developed Countries - 47







TOP **10** Under-Developed Countries based on:

	LOW GDPP	LOW INCOME	HIGH CHILD MORTALITY
1.	Equatorial Guinea	Equatorial Guinea	Iraq
2.	Gabon	Gabon	Botswana
3.	South Africa	Botswana	South Africa
4.	Botswana	Iraq	Eritrea
5.	Namibia	South Africa	Namibia
6.	Iraq	Namibia	Yemen
7.	Timor-Leste	Angola	Kenya
8.	Angola	Congo, Rep.	Madagascar
9.	Congo, Rep.	Nigeria	Timor-Leste
10.	Nigeria	Yemen	Kiribati

OVERALL CONDITIONS*

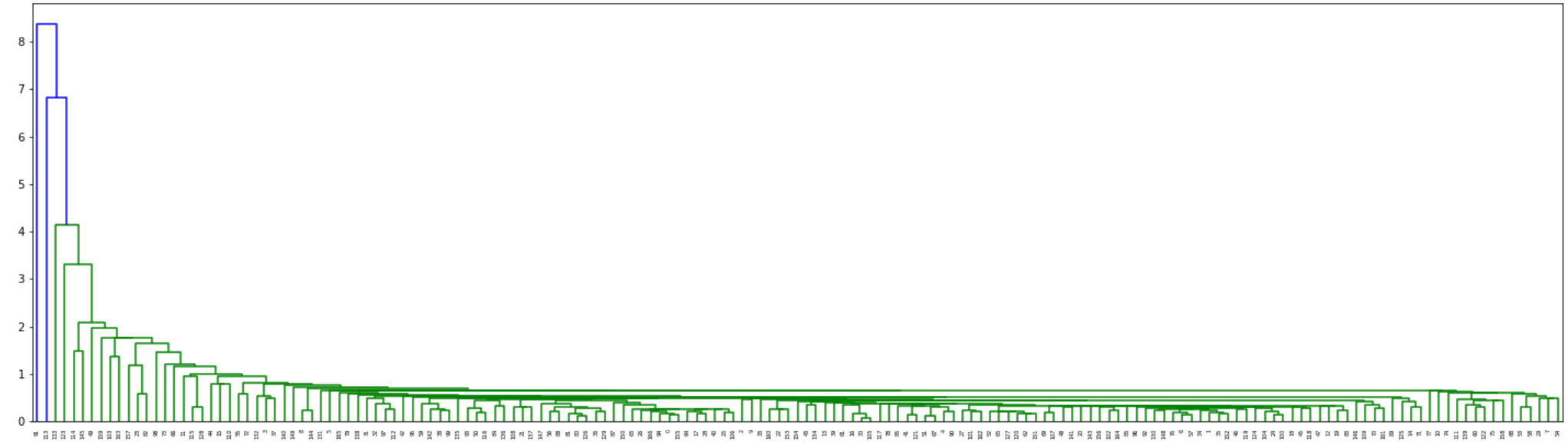
Burundi
Congo, Dem. Rep.
Niger
Sierra Leone
Mozambique
Central African Republic
Guinea-Bissau
Burkina Faso
Guinea
Haiti

OVERALL CONDITIONS* = Low GDPP + Low Income + High Child Mortality Rate
These are the top 10 countries which are in DIRE need of aid among all the under-developed Countries

Note: The subplots can be clearly seen in the code outputs.

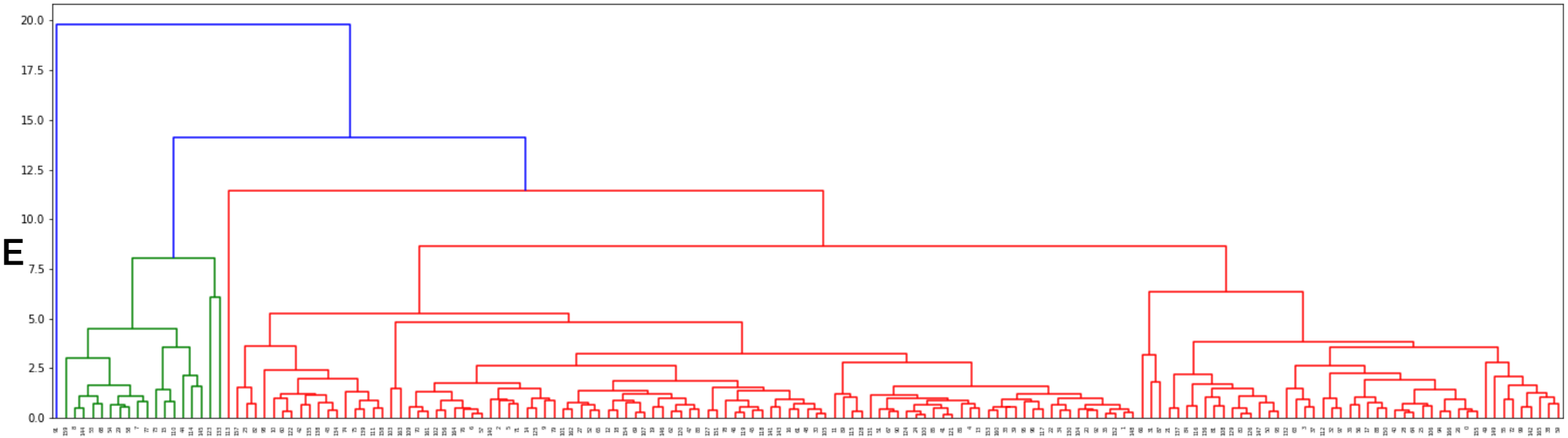
Hierarchical Clustering

SINGLE LINKAGE



COMPLETE LINKAGE

3 Clusters at 12.5



Renaming Clusters

K Means Clustering

ClusterID	gdpp	child_mort	income
Developing	7979.912088	20.357143	13968.021978
Developed	48114.285714	5.046429	50178.571429
Under-developed	1909.208333	91.610417	3897.354167

Hierarchical Clustering with Complete Linkage Method

H_ClusterID	gdpp	child_mort	income
0	12470.812121	37.929091	16765.533333
1	105000.000000	2.800000	91700.000000
2	2330.000000	130.000000	5150.000000

By comparing averages of K-means and Hierarchical Clustering, we can conclude that
Cluster 2 belongs to Under-Developed Countries,
Cluster 1 belongs to Developed Countries
Cluster 0 belongs to Developing Countries.

RESULTS:

Since Main Focus was to find out countries which are in dire need of aid as per socio-economic factors, I've calculated only the Under-developed countries' based on Mean values of Child Mortality, Income and GDPP.
After data binning, Hierarchical clustering gave only 5 countries which satisfied overall conditions.

KMEANS CLUSTERING

1. Burundi
2. Congo, Dem. Rep.
3. Niger
4. Sierra Leone
5. Mozambique
6. Central African Republic
7. Guinea-Bissau
8. Burkina Faso
9. Guinea
10. Haiti
11. Mali
12. Benin
13. Chad
14. Lesotho
15. Mauritania
16. Cote d'Ivoire
17. Cameroon

HIERARCHICAL CLUSTERING

1. Sierra Leone
2. Central African Republic
3. Haiti
4. Mali
5. Chad

ClusterID value counts of H-Clustering,
0: 165 countries,
1: 1 country,
2: 1 country

Therefore due to imbalance counts, making visualization plots would be inappropriate.

I Choose K-Means Clustering Algorithm over Hierarchical Clustering Algorithm because:

1. The clusterID value counts were properly divided and visualizing each cluster was possible.
2. The countries in dire need of aid by K-Means (17) were more than by Hierarchical Clustering(5)

Conclusion

After comparing both **K-means** and **Hierarchical clustering** method, I am going with the **K-means** outcomes as the plots are clearly visible. As in both the methods, the produced mean for underdeveloped countries was almost same. i.e. deciding no. of clusters as **3** was profitable.

After grouping all the countries into **3 groups** by using some socio-economic and health factors, we can determine the **overall** development of the countries.

Here, the countries are categorised into list of **developed** countries, **developing** countries and **under-developed** countries.

In **Developed** countries, we can see the **GDP per capita** and **income** is **high** where as Death of children under 5 years of age per 1000 live births i.e. **child-mort** is very **low**, which is expected.

In **Developing** countries and **Under-developed** countries, the **GDP per capita** and **income** are **low** and **child-mort** is **high.** Specially, for **under-developed** countries, the **death rate** of children is very **high.**

Recommendations

- From bar chats, we can clearly see the socio-economic and heath situation of the under developed countries.
- In countries like Haiti, Sierra Leone, Chad, etc., the death rate of children under 5 years of age per 1000 (child-mort) is high.

- Countries like Burundi, Congo, Niger, etc., GDP per capita is very low & the income per person is also low. So, these countries are considered as Poor countries.
- If Child Mortality Rate is decreased and GDPP , Income is increased in Under-developed countries, the need will be resolved.