Clustering the Countries by using K-Means for HELP International

Amira Amandanisa

Python Data Science-Sanbercode Batch 29

Introduction

- 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.

Objectives

- Data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis etc.
- **Outlier Analysis:** You must perform the Outlier Analysis on the dataset. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.
- Create model using both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters.
- Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.
- Perform visualisations on the clusters that have been formed using the features selected for building the clustering model

Data Dictionary

- **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

```
country.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Negara            167 non-null    object 
 1   Kematian_anak     167 non-null    float64
 2   Ekspor            167 non-null    float64
 3   Kesehatan         167 non-null    float64
 4   Impor             167 non-null    float64
 5   Pendapatan        167 non-null    int64  
 6   Inflasi           167 non-null    float64
 7   Harapan_hidup     167 non-null    float64
 8   Jumlah_fertiliti  167 non-null    float64
 9   GDPperkapita      167 non-null    int64  
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```

- The dataset contains 167 rows and 10 columns. Out of these 10 columns, 9 are integer type and only 1 object/categorical column is present which is the name of the country.
- There are no **Null** or **Nan** values.

Datatypes, Duplicates and Describing Data

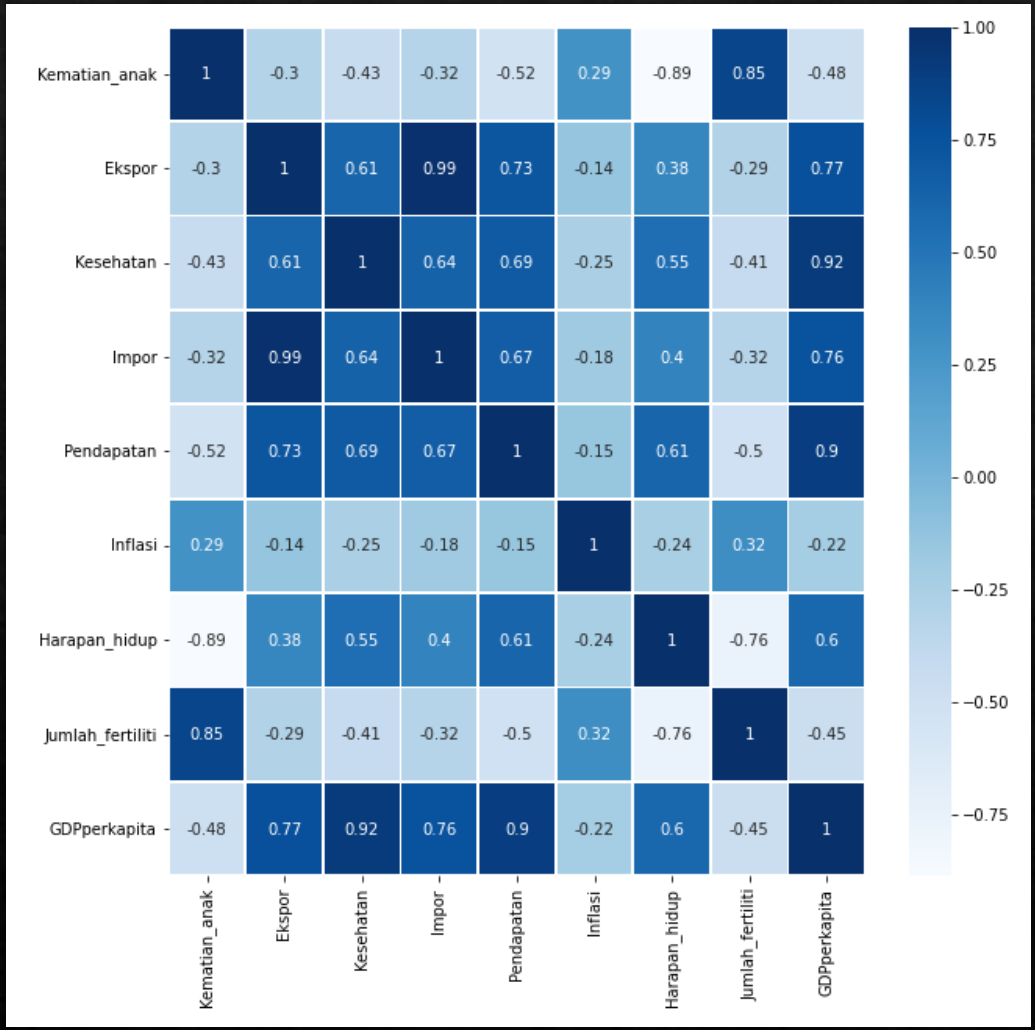
- Checking for outliers by describing percentiles and min max values.

```
Negara              object
Kematian_anak      float64
Ekspor             float64
Kesehatan          float64
Impor              float64
Pendapatan           int64
Inflasi            float64
Harapan_hidup      float64
Jumlah_fertiliti   float64
GDPperkapita         int64
dtype: object
```

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

	Kematian_anak	Ekspor	Kesehatan	Impor	Pendapatan	Inflasi	Harapan_hidup	Jumlah_fertiliti	GDPperkapita
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

Correlation Heatmap



INSIGHTS FROM CORRELATION MAP

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

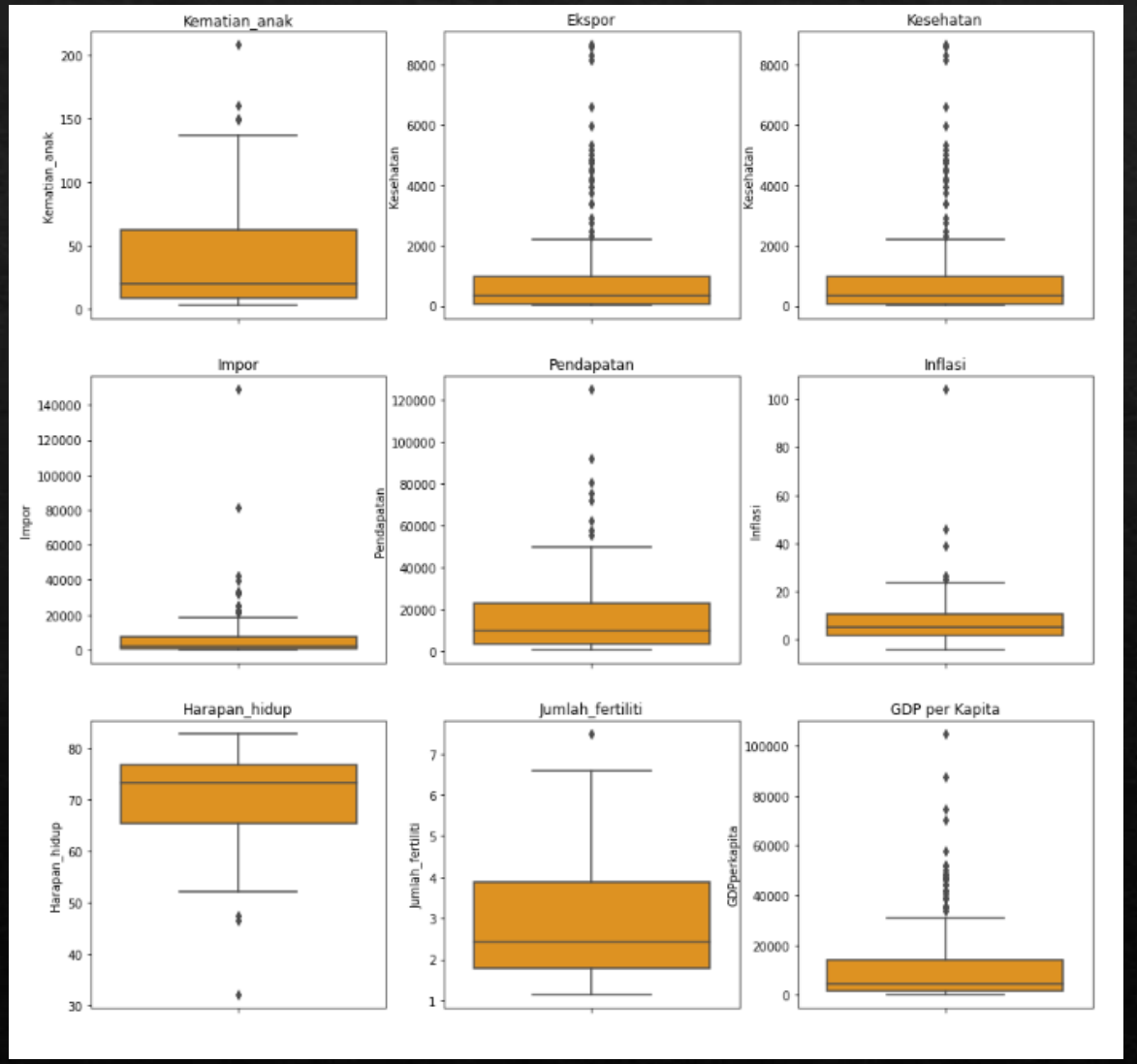
2. **Health, Exports, Income, Imports** are highly correlated with **GDP per capita.**

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

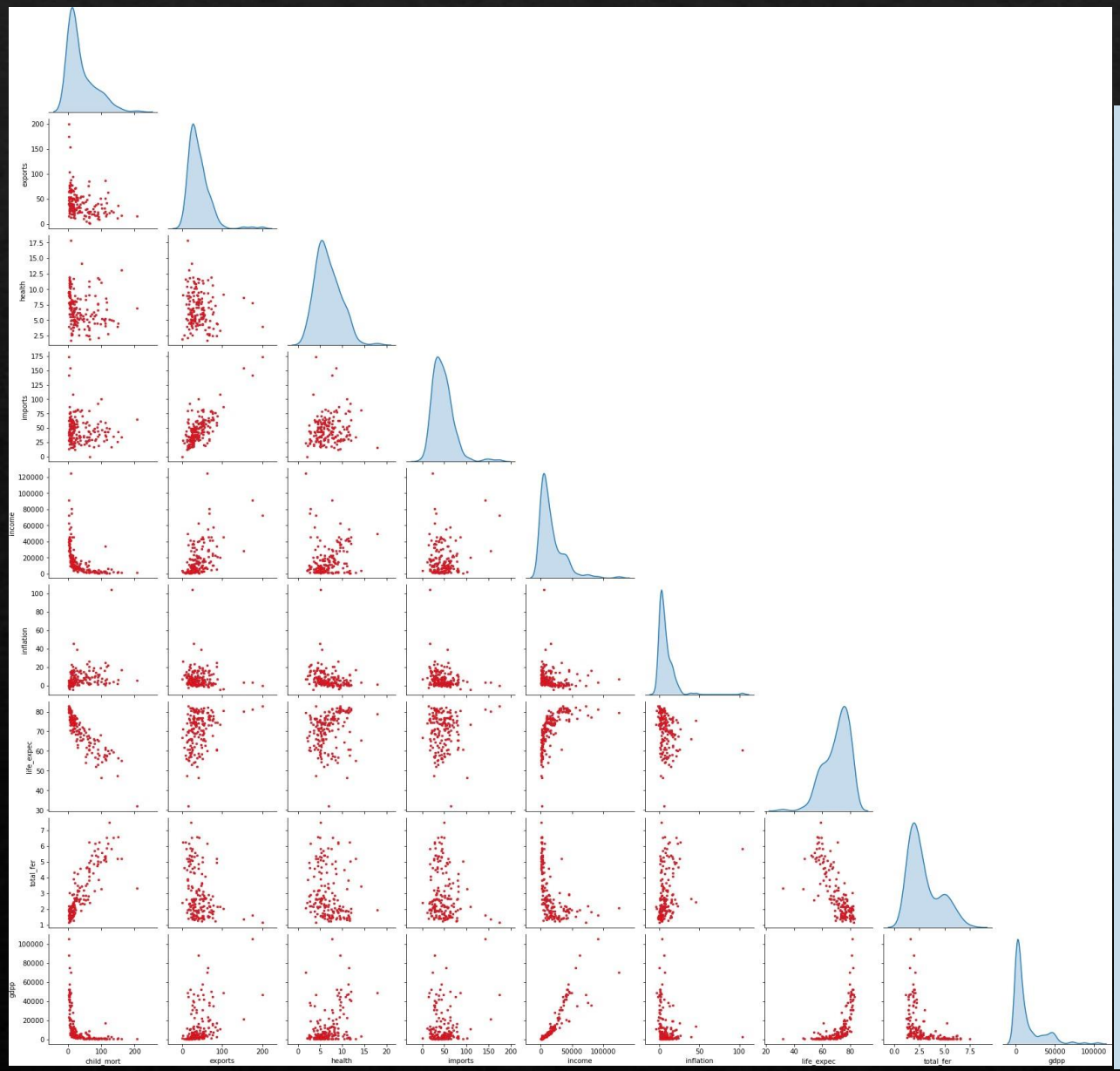
4. **Total fertility** is highly positively correlated with **Child Mortality Rate** and negatively correlated with **Life Expectancy.**

Univariate Analysis

- There is **minimum one outlier** in each of the numerical features.
- Most of outliers can be seen in **GDPP**.
- As our data contains **167 countries**, removing these outliers could increase chances of **removing dire needy** countries
- Example: In case of **child mortality rate**, country with **208** values is being specified as outlier but that country itself **could be in dire** need of aid.
- Removing outlier is **NOT** good option as per the above conditions. So, 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.

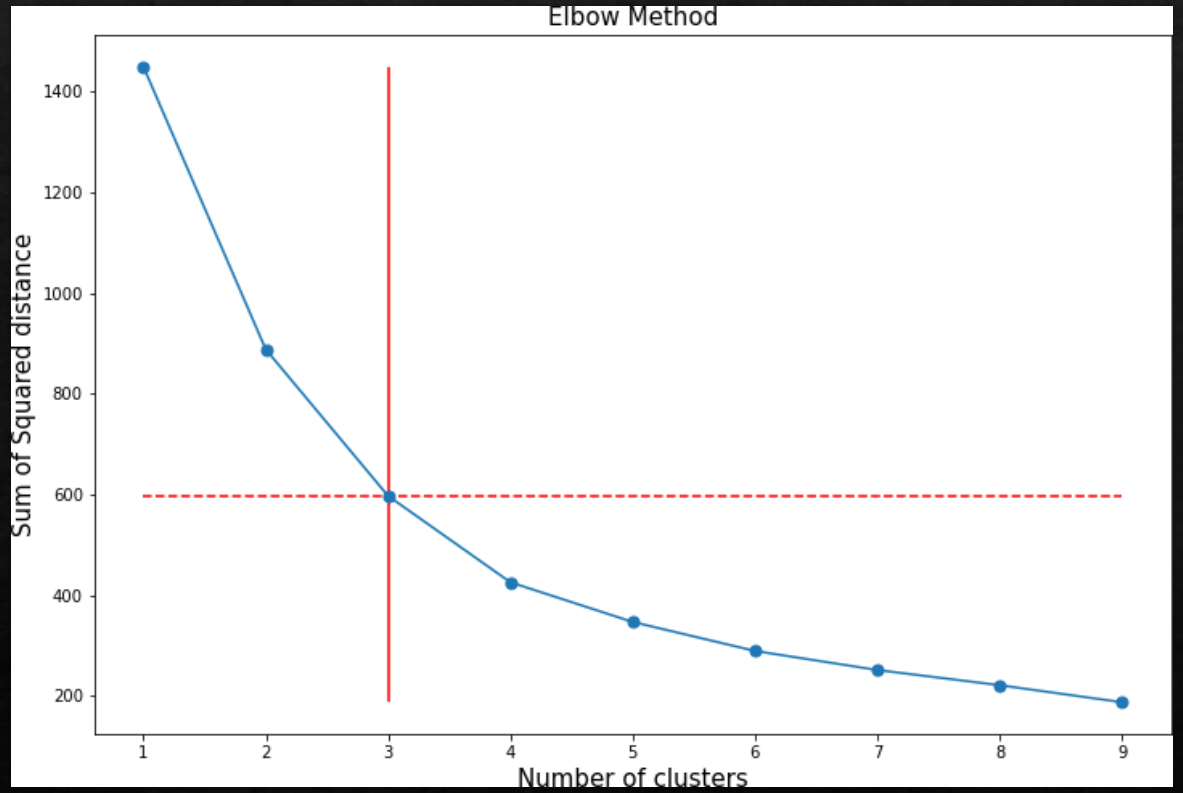
2. Bivariate Analysis

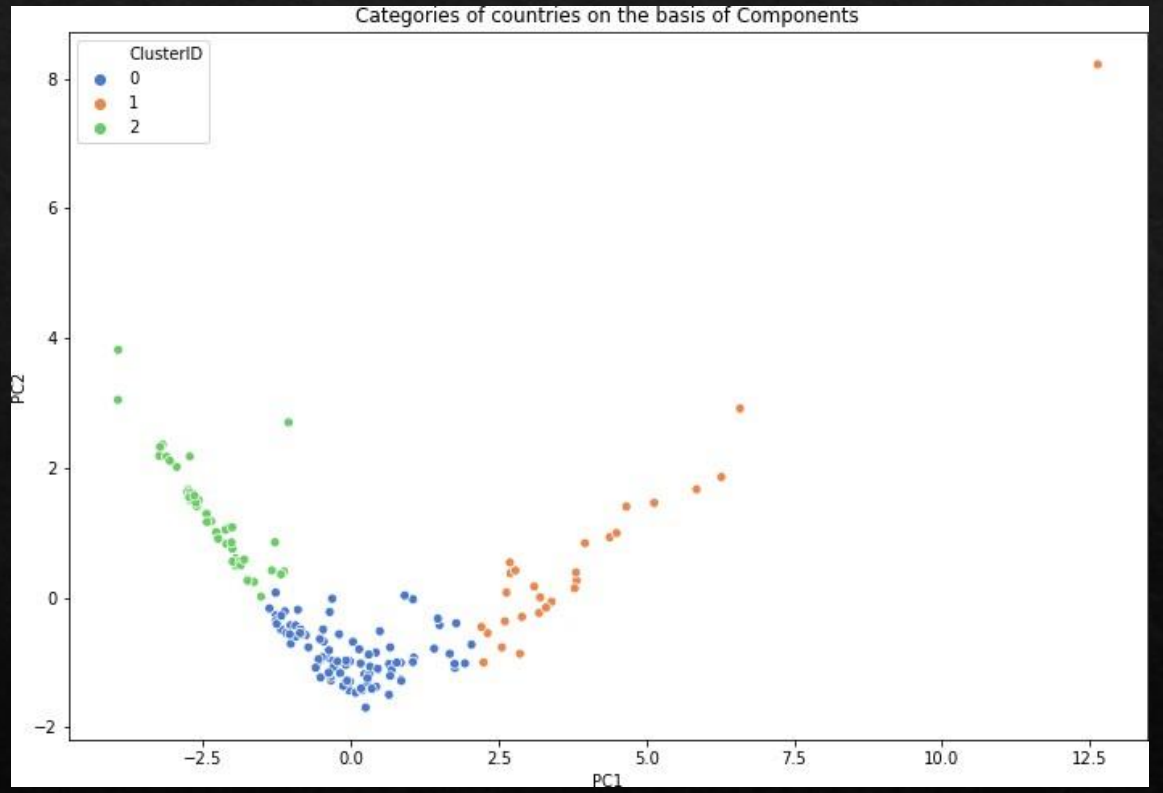
1. Linear relationship is found between **[gdpp – income], [imports – exports], [total_fer - child_mort]**

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**

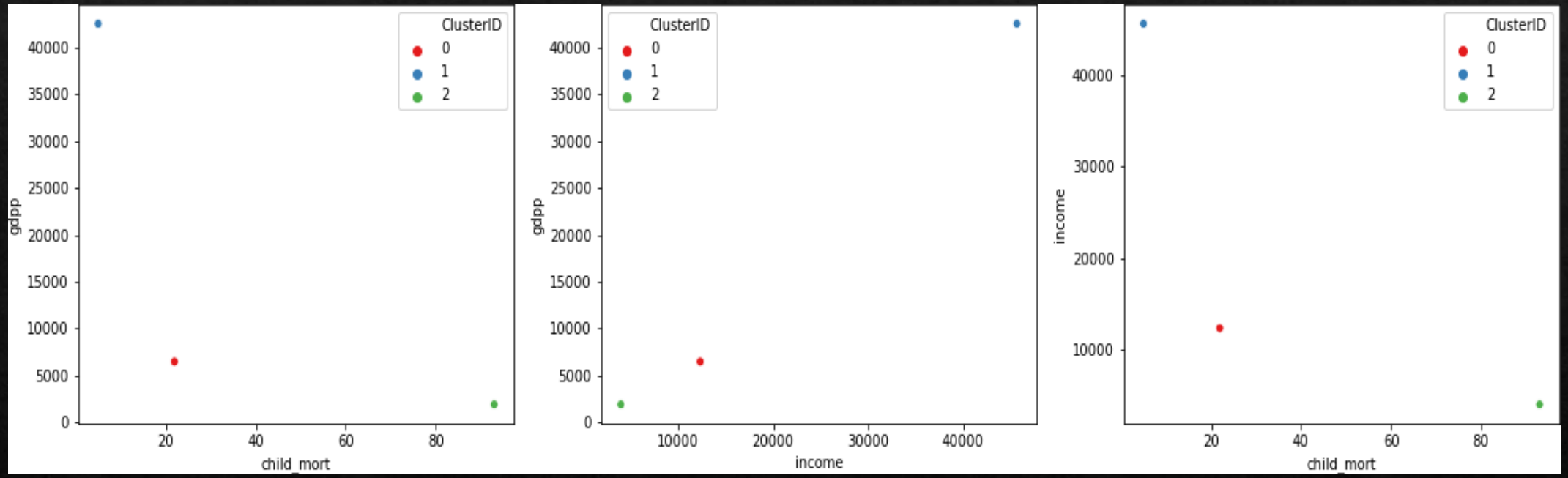
K Means Clustering Algorithm





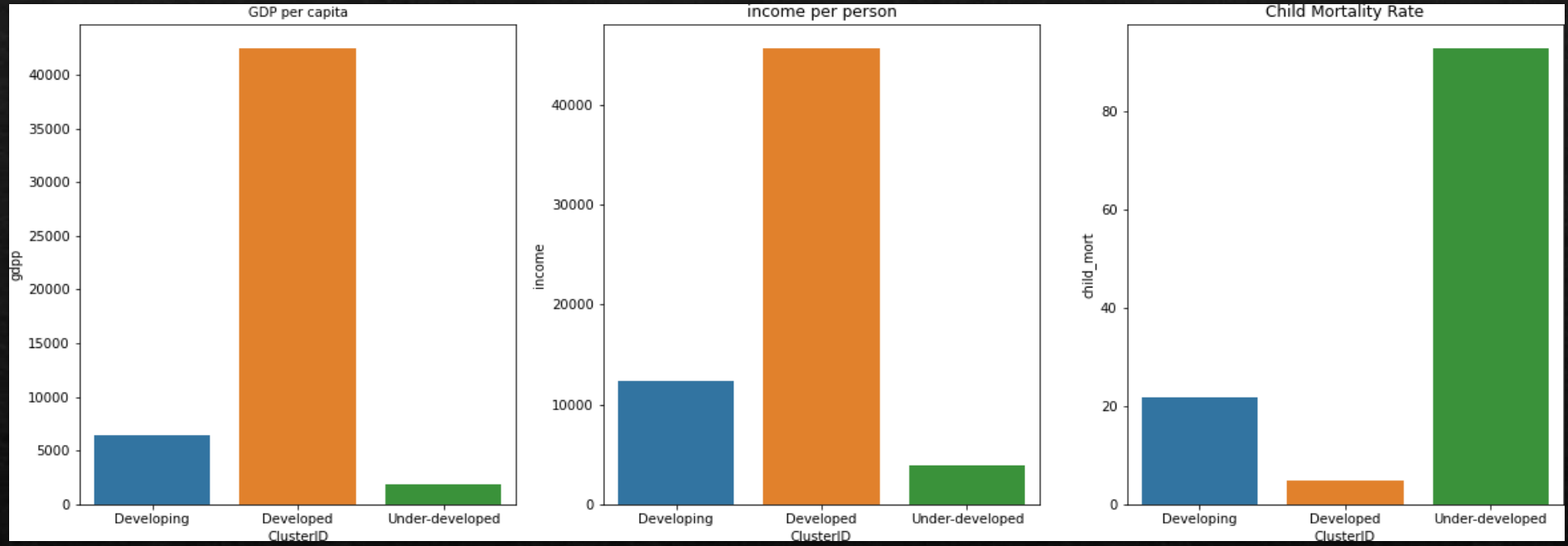
As per Elbow method, we'll selece no of clusters as 3.

Renaming Clusters base on Means



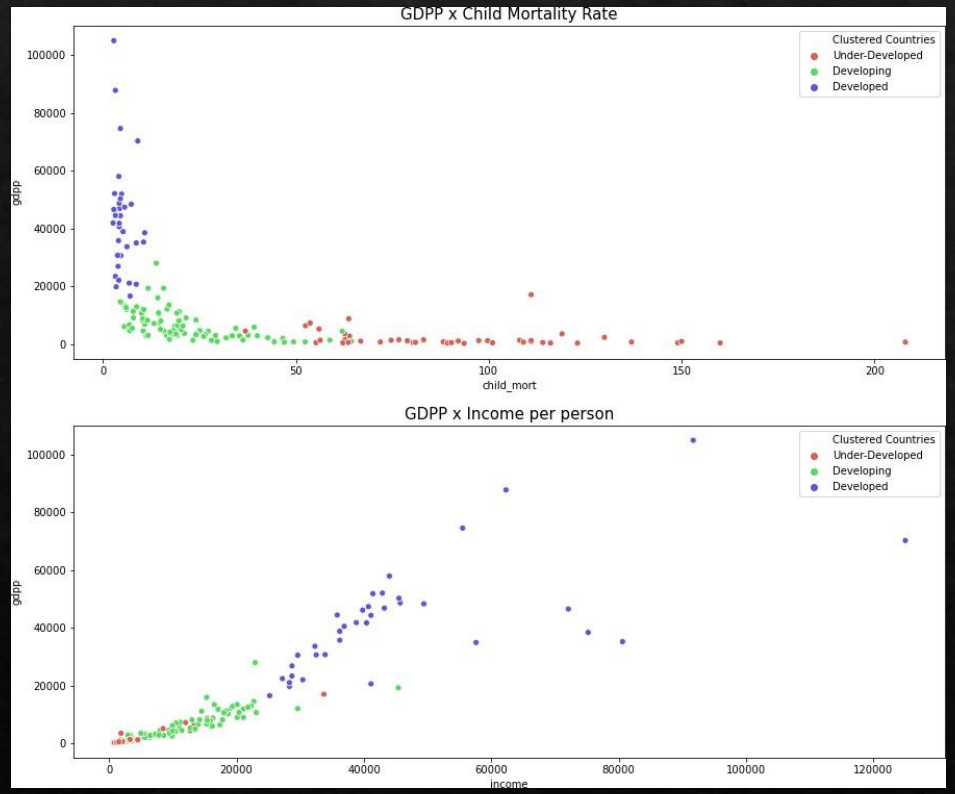
1. Countries with high GDPP, high income and low Child Mortality Rate are Developed countries
2. Countries with average GDPP, average income and average Child Mortality Rate are Developing countries.
3. Countries with low GDPP, low income and high Child Mortality Rate are under developing countries

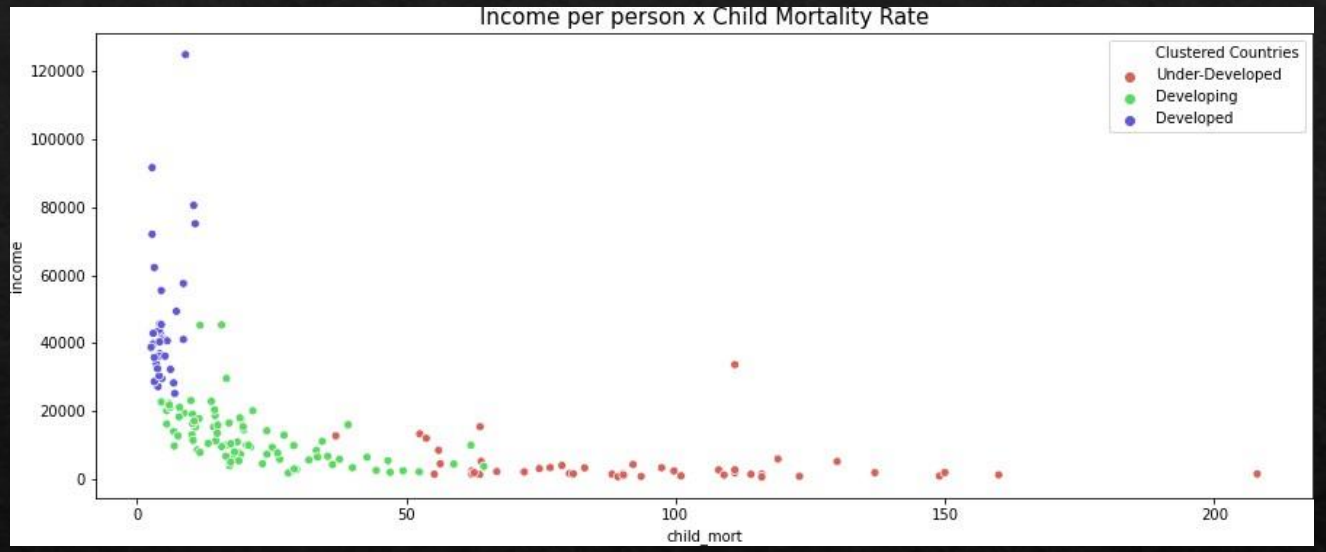
Univariate Analysis on Clustered Countries



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 habing 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 at least child mortality rate.

Bivariate Analysis on Clustered Countries





In gdpp x child_mort, there is some clustering when gdpp is low, there child mort is high, which is true for under-developed countries in reality.

In gdpp x income, there is come clustering where gdpp is average, there income is average, which is true for developing countries in reality

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.

Under-Developed Countries-47

	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

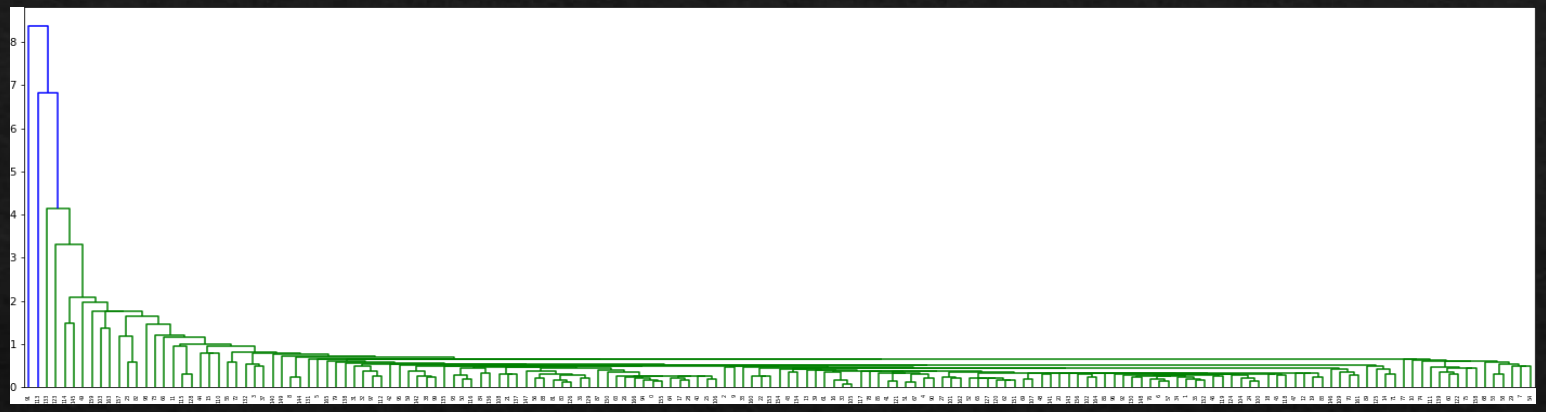
OVERALL CONDITIONS*

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

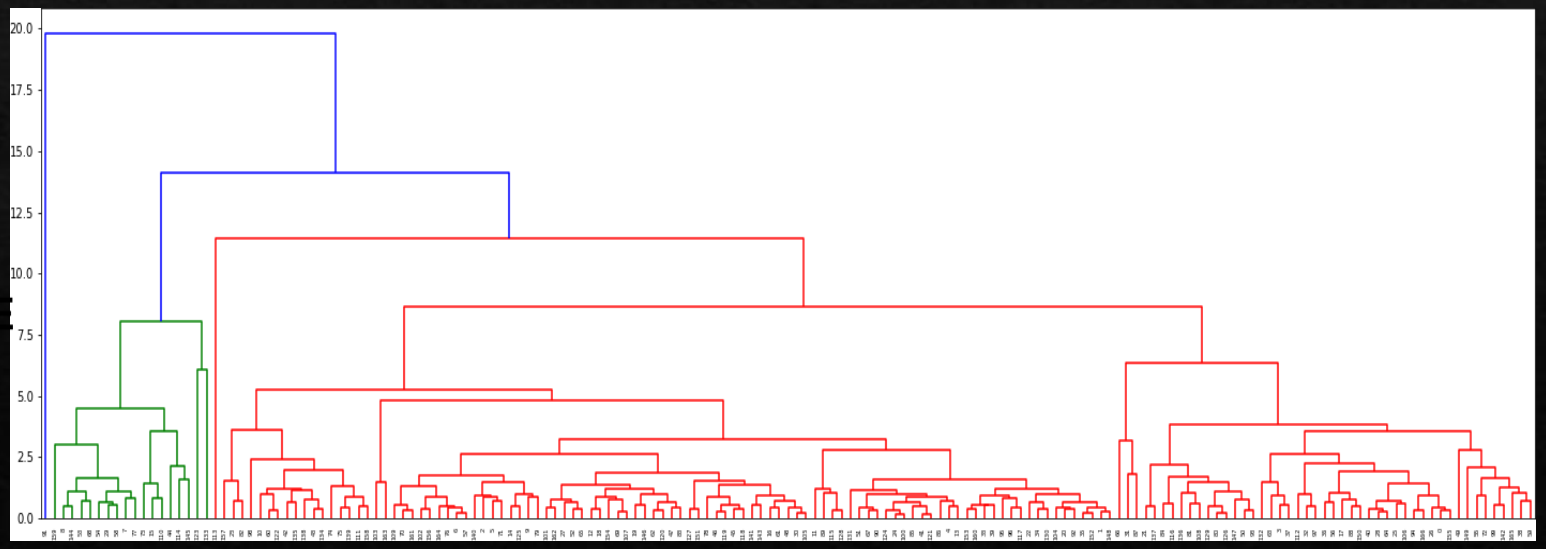
OVERALL CONDITIONS* = Low GDPP + Low Income + High Child Mortality Rate
These are the top 10 countries which are in DIRE need of aid among all the under-developed Countries

Hierarachical Clustering

- Single Linkage



- Complete Linkage(3 clusters at 12.5)



COMPLETE LINKAGE

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 average 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

	Negara	GDPperkapita	Pendapatan	Kematian_anak
26	Burundi	231	764	93.6
37	Congo, Dem. Rep.	334	609	116.0
112	Niger	348	814	123.0
132	Sierra Leone	399	1220	160.0
106	Mozambique	419	918	101.0
31	Central African Republic	446	888	149.0
94	Malawi	459	1030	90.5
150	Togo	488	1210	90.3
64	Guinea-Bissau	547	1390	114.0
0	Afghanistan	553	1610	90.2

- Since main focus was to find out countries which are in dire need of aids as per socio economic factors, I've calculated only the under developed countries based on Mean values on child mortality, income and GDPP.

Conclusion

- After grouping all the countries into 3 groups by using socio-economis factors, we can determine the overall development of the countries.
- The countries are catogorised into list of develop countries, developing countries and under-developed countries.
- In developed countries, we can see GDP per capita and income is high whereas death of children under 5 years age per 1000 live births (children mortality in data) is very low, which is expected.
- In developing countries and under developed countries, the GDP per capita and income are low and children mortality is high. Specially for under-developed countries, the death rate of children is very high.