CLUSTERING ASSIGNEMNT

Done By

ANUPAMA RAJEEV

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.*
- *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.*

AIM:

1. *Identify top countries that are in severe need of aid*
2. *Categorize them based on health and socio-economic factors*
3. *Suggest top ten countries that needs to be focussed*

Solution:

Use clustering approach to categorize countries based on socio-economic and health factors.

Analysis Approach

The dataset is called "Country-data.csv" (*here after referred to as "cntry"****)***

Below are the steps followed for analysis:

- ✓ **Understanding the Dataset -** *Reading the dataset, doing basic checks like df.info() - datatypes of each column, shape, df.describe() - statistical description of numerical columns.*
- ✓ **Data Cleaning and preparation -** *Involves checking null/missing values, changed values of three columns from % of GDP to actual GDP*
- ✓ **Univariate Analysis -** *Distribution of continuous variables was analyzed using distribution plot, boxplot and statistical description.*
- ✓ **Bivariate Analysis -** *Discovering patterns between two variables which involved Continuous to Continuous – e.g. Scatter plot*
- ✓ **Multivariate Analysis – heatmap was plotted to analyze correlation**
- ✓ **Scaling data and Hopkins score**
- ✓ **K means clustering approach** – *elbow curve and silhouette score methods were performed, cluster profiling and visualization was done.*
- ✓ **Hierarchical approach –** *single and complete linkage methods were done, cluster profiling and visualization was done*
- ✓ **Concluded with the list of top ten countries that need aid immediately.**

Understanding data file name: Country-data.csv

- *dataframe has 167 rows and 10 columns.*
- *Information on missing values and datatypes was checked*
- *Statistical description was looked at.*

```
2]: #reading the country-data.csv file to the dataframe name 'cntry'

    cntry = pd.read_csv("Country-data.csv", encoding = 'utf-8')
```

```
3]: #setting seed to get the same cluster

    np.random.seed(0)
```

```
4]: #viewing first few rows

    cntry.head()
```

4]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2000	10.0000	7.5800	44.9000	1610	9.4400	56.2000	5.8200	553
1	Albania	16.6000	28.0000	6.5500	48.6000	9930	4.4900	76.3000	1.6500	4090
2	Algeria	27.3000	38.4000	4.1700	31.4000	12900	16.1000	76.5000	2.8900	4460
3	Angola	119.0000	62.3000	2.8500	42.9000	5900	22.4000	60.1000	6.1600	3530
4	Antigua and Barbuda	10.3000	45.5000	6.0300	58.9000	19100	1.4400	76.8000	2.1300	12200

```
5]: #checking the dataframe info

    cntry.info()

    #no null values found in any columns
    #country is of type 'object' and all other columns are numerical / continous
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
```

Data cleaning and preparation

1. *Checked for null values. None were found in this dataset.*

2. *For data preparation, three columns named "exports", "imports" and health were taken to change values from percentage of GDP to actual GDP values.*

```
: #checking is there are any miss

  cntry.isnull().sum()

  #this confirms that there are r
```

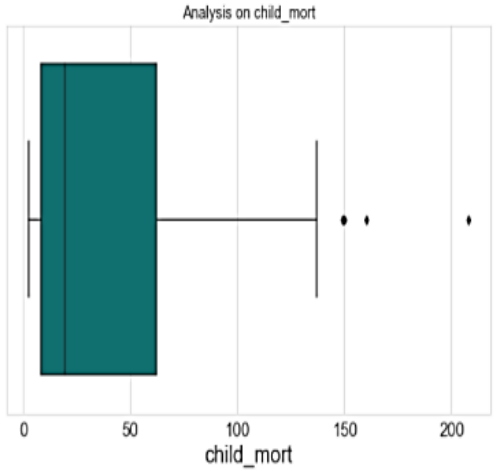
```
: country        0
  child_mort     0
  exports        0
  health         0
  imports        0
  income         0
  inflation      0
  life_expec     0
  total_fer      0
  gdpp           0
  dtype: int64
```

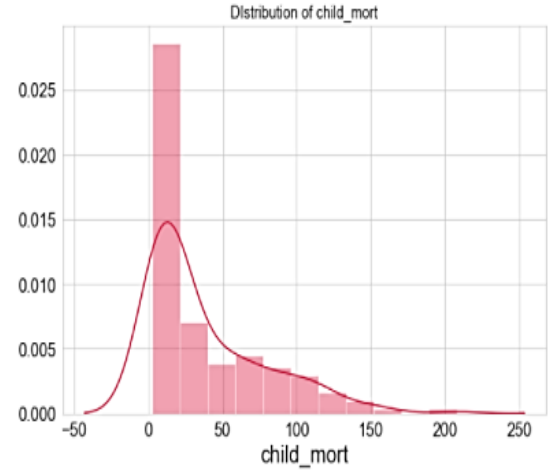
```
#changing 'export' column
cntry.exports = (cntry.exports*cntry.gdpp)/100

#changing 'health' column
cntry.health = (cntry.health*cntry.gdpp)/100

#changing 'imports' column
cntry.imports = (cntry.imports*cntry.gdpp)/100
```

EDA analysis - Univariate

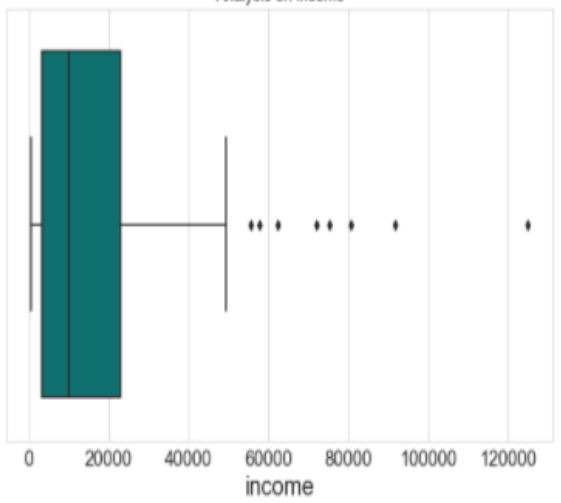
- *Performed univariate analysis on all numeric columns. Outliers and distribution was checked here. Outliers were capped instead of removal*

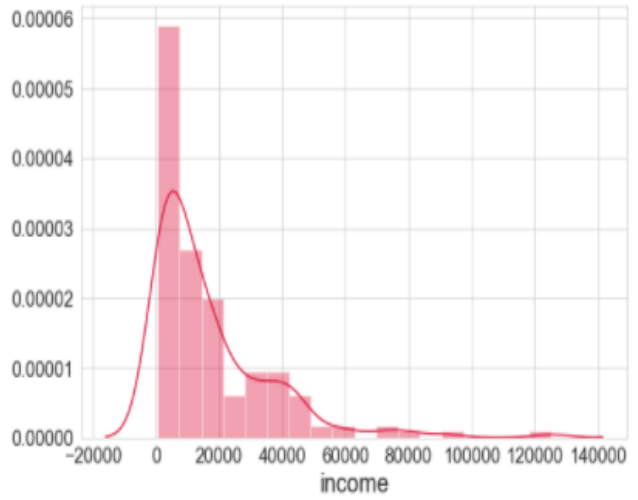
```
Statistical Information of child_mort column
count   167.0000
mean     38.2701
std      40.3289
min       2.6000
25%       8.2500
50%      19.3000
75%      62.1000
max     208.0000
Name: child_mort, dtype: float64
```






```
Statistical Information of income column
count      167.0000
mean     17144.6886
std      19278.0677
min        609.0000
25%       3355.0000
50%       9960.0000
75%      22800.0000
max     125000.0000
Name: income, dtype: float64
```

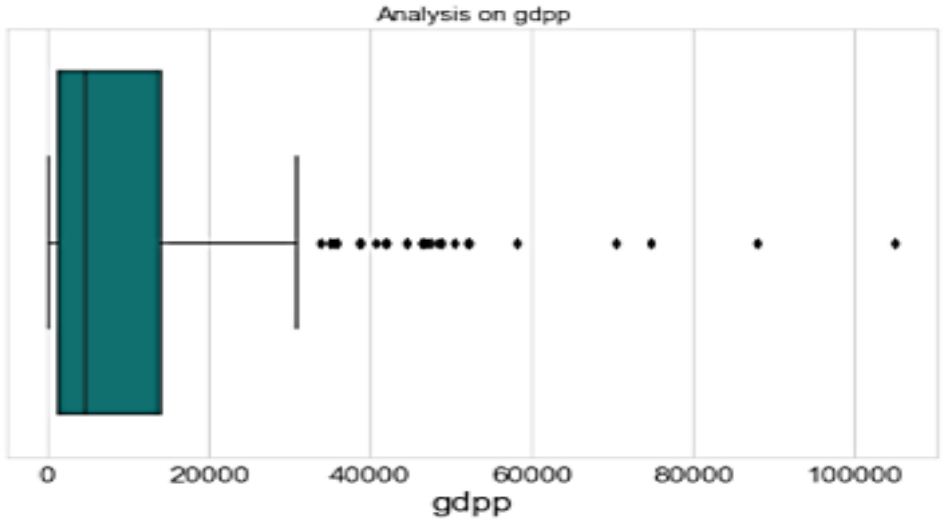


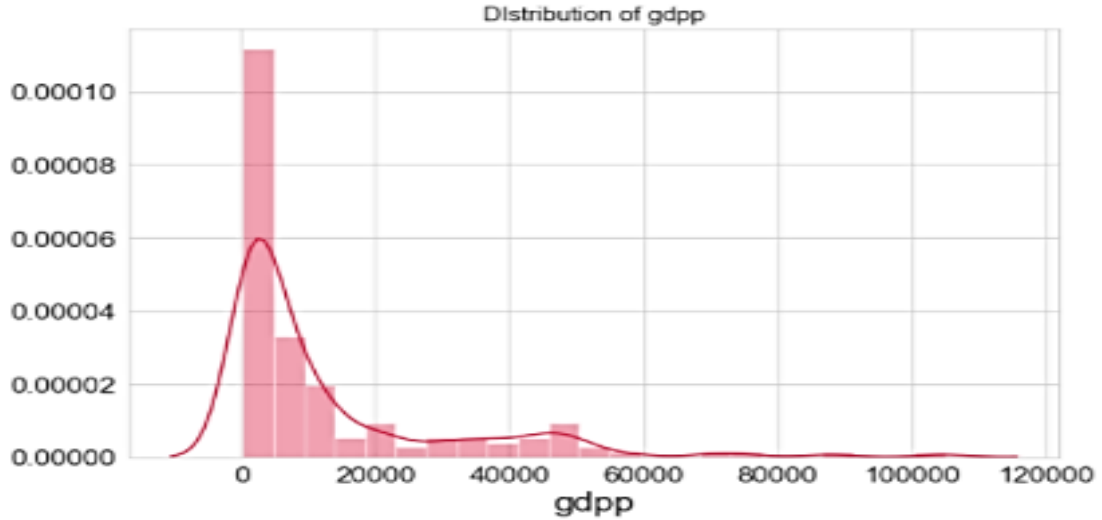


EDA analysis

- *this column indicates **'net GDP'.** As it can be seen from the below analysis of this column, there are outliers on the left and right. Left outliers cannot be removed as they represent countries with lower GDP; and it has to be checked further; hence they are not treated. On the other hand, right outliers can be capped and removed.*

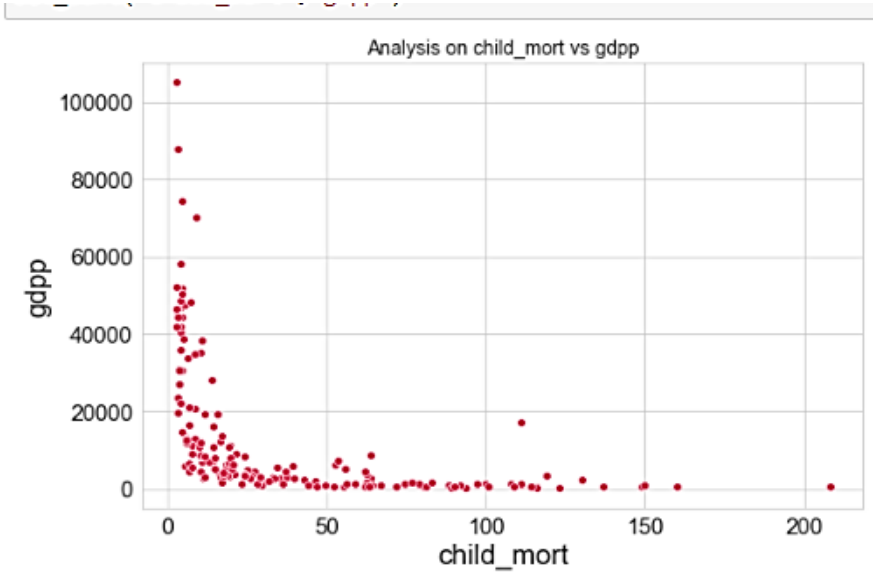
```
Statistical Information of gdpp column
count      167.0000
mean     12964.1557
std      18328.7048
min        231.0000
25%       1330.0000
50%       4660.0000
75%      14050.0000
max     105000.0000
Name: gdpp, dtype: float64
```

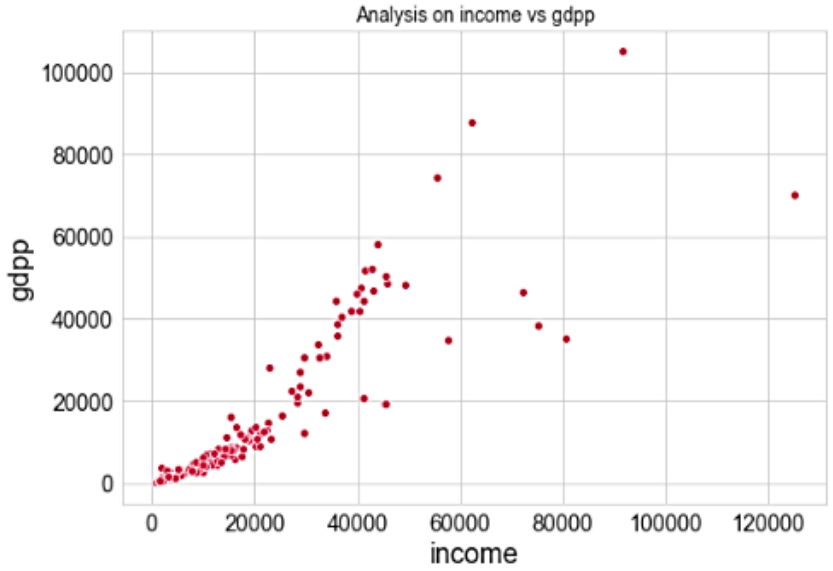




EDA analysis

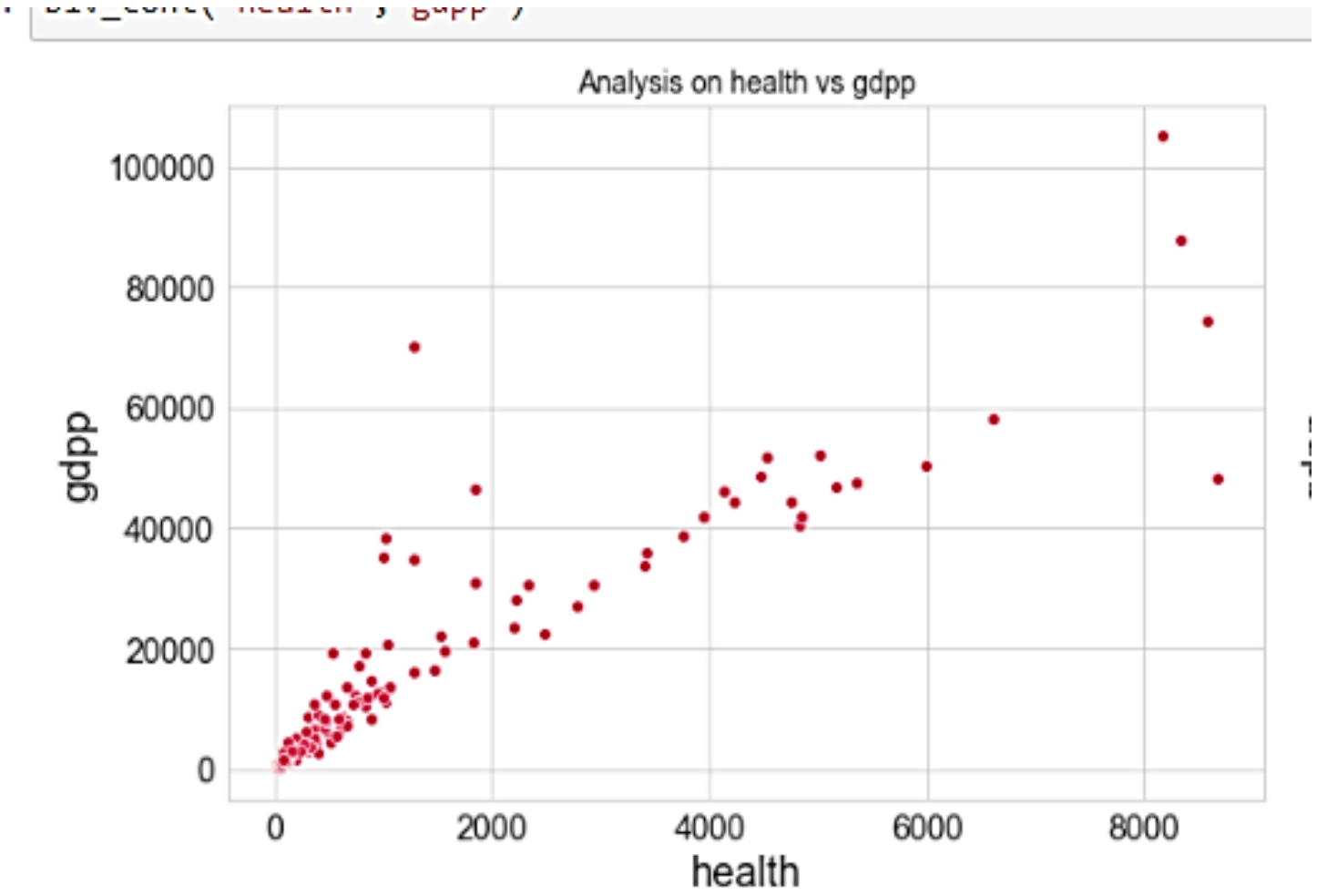
- *First scatter plot is between GDP and child mortality rate; and it indicates that as child mortality rate increases, it affects the total GDP negatively.*
- *Second scatter plot is between GDP and income; and it indicates that as income increases, it affects the total GDP positively.*





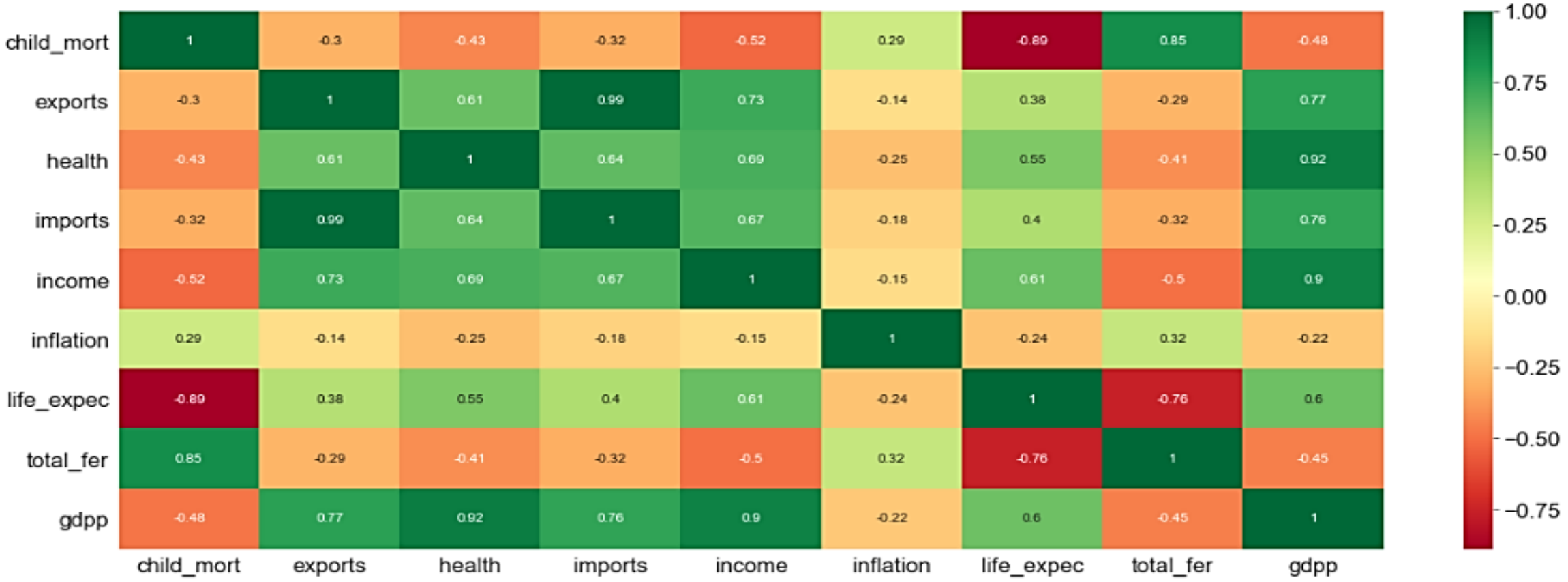
EDA analysis

- *this scatter plot is between GDP and health; and it indicates that as health spending rate is higher, it affects the total GDP positively.*



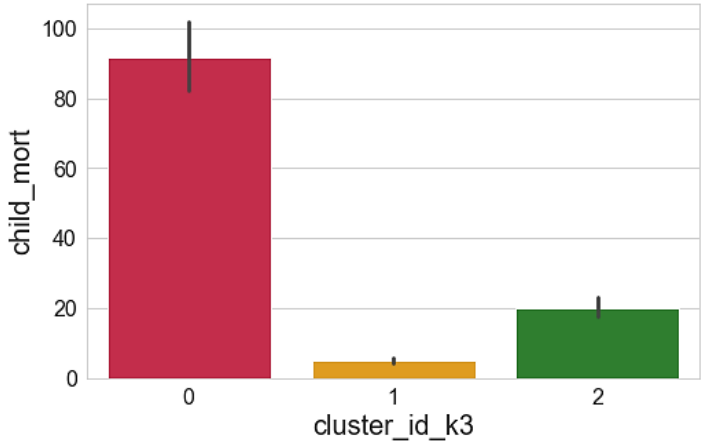
Correlation

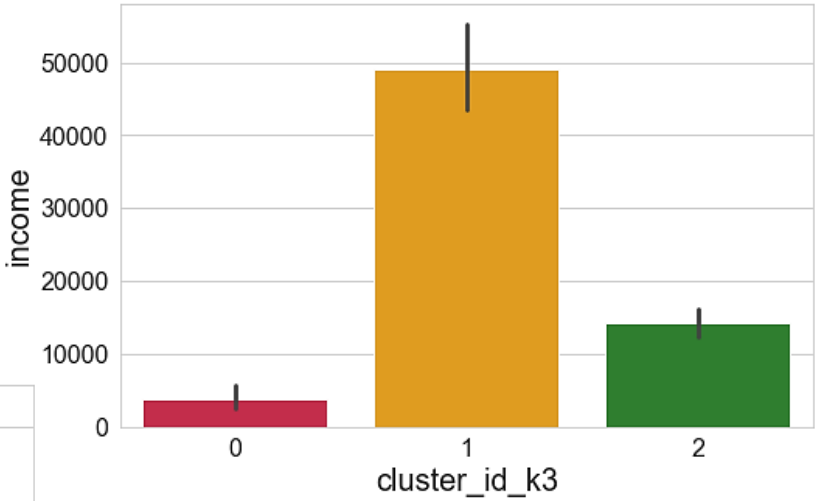
- *Even though there are features that are highly correlated, they won't be dropped at the moment.*

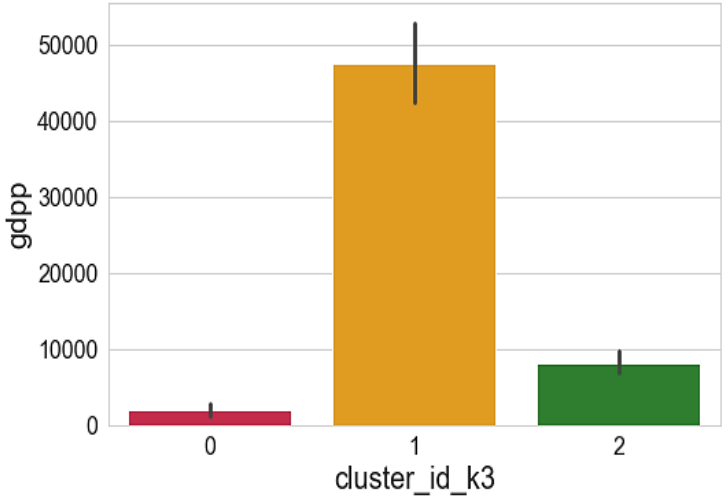


Approach after scaling KMEANS

1. *analysis with k means, k = 3 cluster was taken since that gave balanced data points between the clusters*
2. *cluster profiling and visualisation was done using plots*
3. ***concluded that cluster 0 had datapoints with lowest GDP, lowest income and highest child mortality rate***
4. ***48 countries** were present in this cluster; out of which top ten were listed based on high child_mort and lowest GDP*

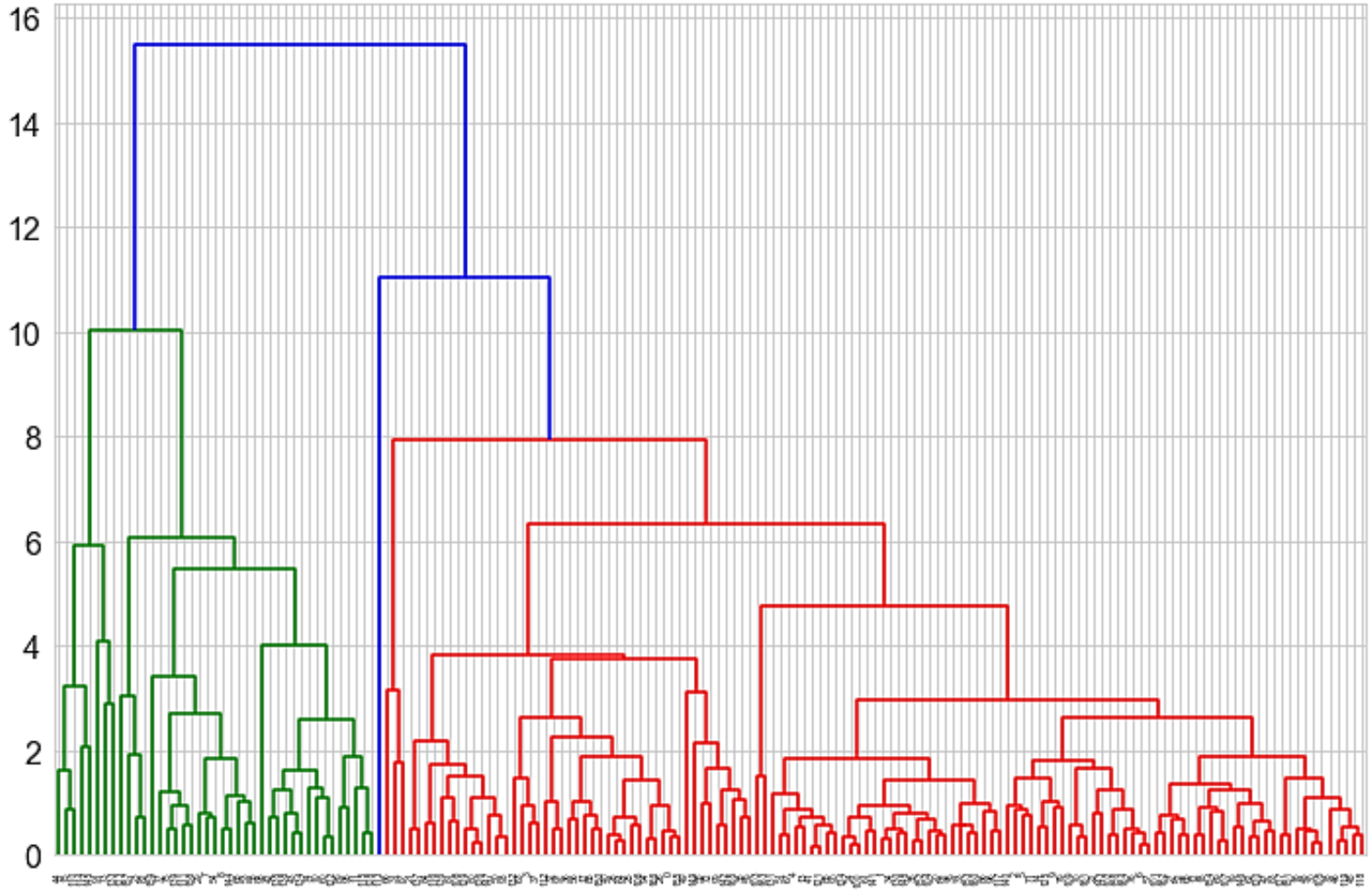






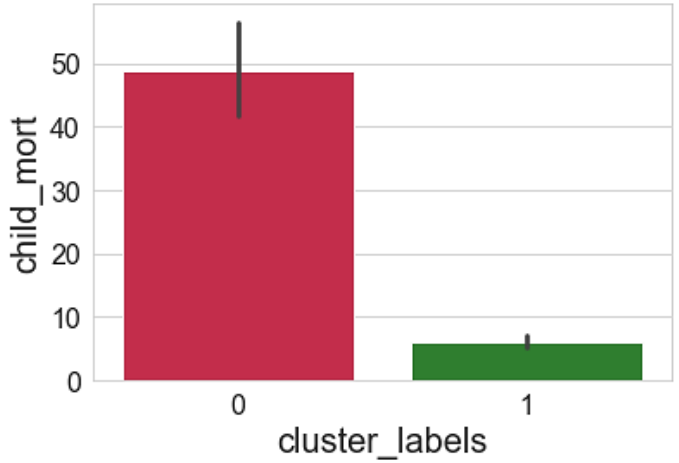
Approach after scaling Hierarchical

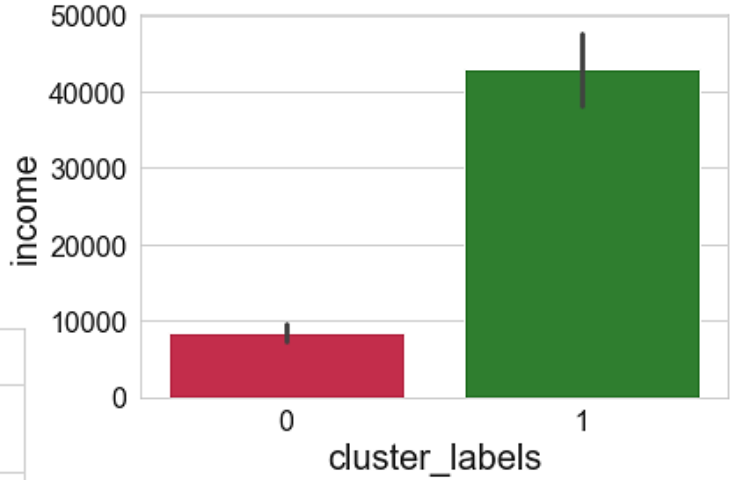
1. *simple and complete linkage was performed*
2. *complete linkage dendrogram was considered for further analysis as it was more readable*
3. *Analysis was done at k = 2; and performed further analysis*
4. *cluster profiling and visualisation was done using plots*

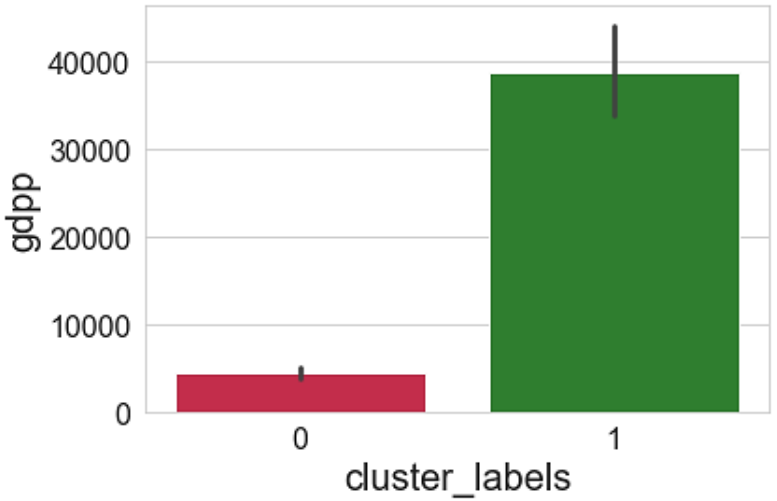


Approach after scaling Hierarchical

1. ***concluded that cluster 0 had datapoints with lowest GDP, lowest income and highest child mortality rate***
2. ***126 countries were present*** *in this cluster; out of which top ten were listed based on high child_mort and lowest GDP*







TOP 10 COUNTRIES IN SEVERE NEED FOR AID BASED ON LOWEST GDPP RATE

1. *Burundi*
2. *Liberia*
3. *Congo, Dem. Rep*
4. *Niger*
5. *Sierra Leone*
6. *Madagascar*
7. *Mozambique*
8. *Central African Republic*
9. *Malawi*
10. *Eritrea*

country	child_mort	exp	gdpp	cluster_id_k3	c
Burundi	93.6000	20.	231.0000	0	
Liberia	89.3000	62.	327.0000	0	
Congo, Dem. Rep.	116.0000	137.	334.0000	0	
Niger	123.0000	77.	348.0000	0	
Sierra Leone	160.0000	67.	399.0000	0	
Madagascar	62.2000	103.	413.0000	0	
Mozambique	101.0000	131.	419.0000	0	
Central African Republic	149.0000	52.	446.0000	0	
Malawi	90.5000	104.	459.0000	0	
Eritrea	55.2000	23.	482.0000	0	

TOP 10 COUNTRIES IN SEVERE NEED FOR AID BASED ON HIGHEST CHILD MORT.

1. *Haiti*
2. *Sierra Leone*
3. *Chad*
4. *Central African Republic*
5. *Mali*
6. *Nigeria*
7. *Niger*
8. *Angola*
9. *Congo, Dem. Rep*
10. *Burkina Faso*

	country	child_mort	
	Haiti	208.0000	1
	Sierra Leone	160.0000	
	Chad	150.0000	3
	Central African Republic	149.0000	
	Mali	137.0000	1
	Nigeria	130.0000	5
	Niger	123.0000	
	Angola	119.0000	21
	Congo, Dem. Rep.	116.0000	1
	Burkina Faso	116.0000	1

CONCLUSION and SOLUTION

- *K-Means approach with number of clusters being 3 gave a better and accurate result.*

- *Not only did it give the same list of countries as the output like other methods, but also **all three clusters were well balanced in terms of datapoints distribution**.*

- *Final list will have 48 countries that needs immediate aid are listed.*

- *Since we need countries that are low in socio-economical and health factors, we can not exclude any countries from the dataset as a part of outlier treatment.*

Afghanistan	Gambia	Namibia
Angola	Ghana	Niger
Benin	Guinea	Nigeria
Botswana	Guinea-Bissau	Pakistan
Burkina Faso	Haiti	Rwanda
Burundi	Iraq	Senegal
Cameroon	Kenya	Sierra Leone
Central African Republic	Kiribati	Solomon Islands
Chad	Lao	South Africa
Comoros	Lesotho	Sudan
Congo, Dem. Rep.	Liberia	Tanzania
Congo, Rep.	Madagascar	Timor-Leste
Cote d'Ivoire	Malawi	Togo
Equatorial Guinea	Mali	Uganda
Eritrea	Mauritania	Yemen
Gabon	Mozambique	Zambia