CLUSTERING CASE-STUDY

Presented By –
Khushi Sapra

BUSINESS PROBLEM

- 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.
- Categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most.

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- measurement of the annual growth rate of the GDP deflator
- 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.

INSPECTING THE DATA SET

- Shape of the Data: (167, 10), i.e. the data has 167 countries.
- Head of the Data:

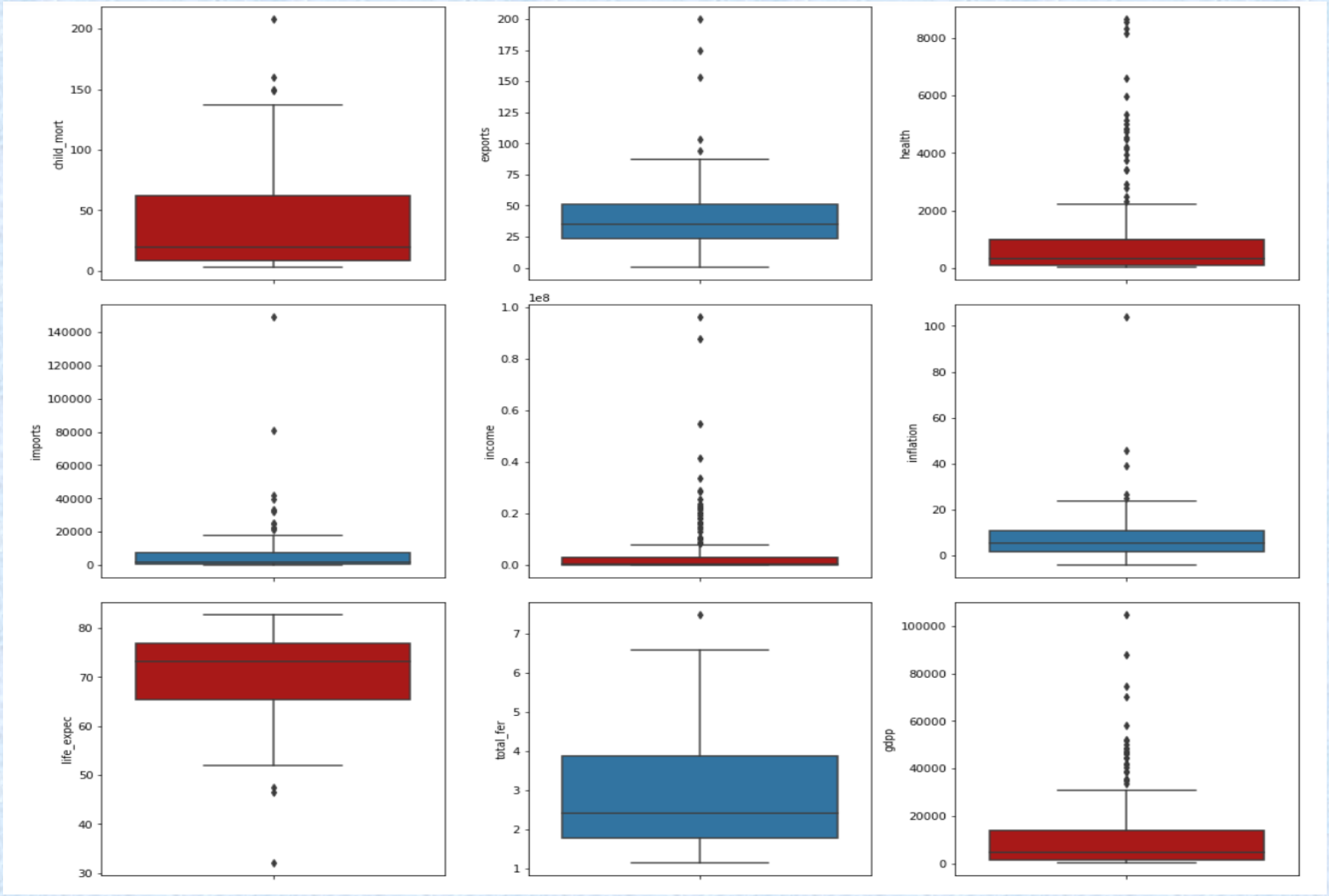
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

- Inspecting the missing values:

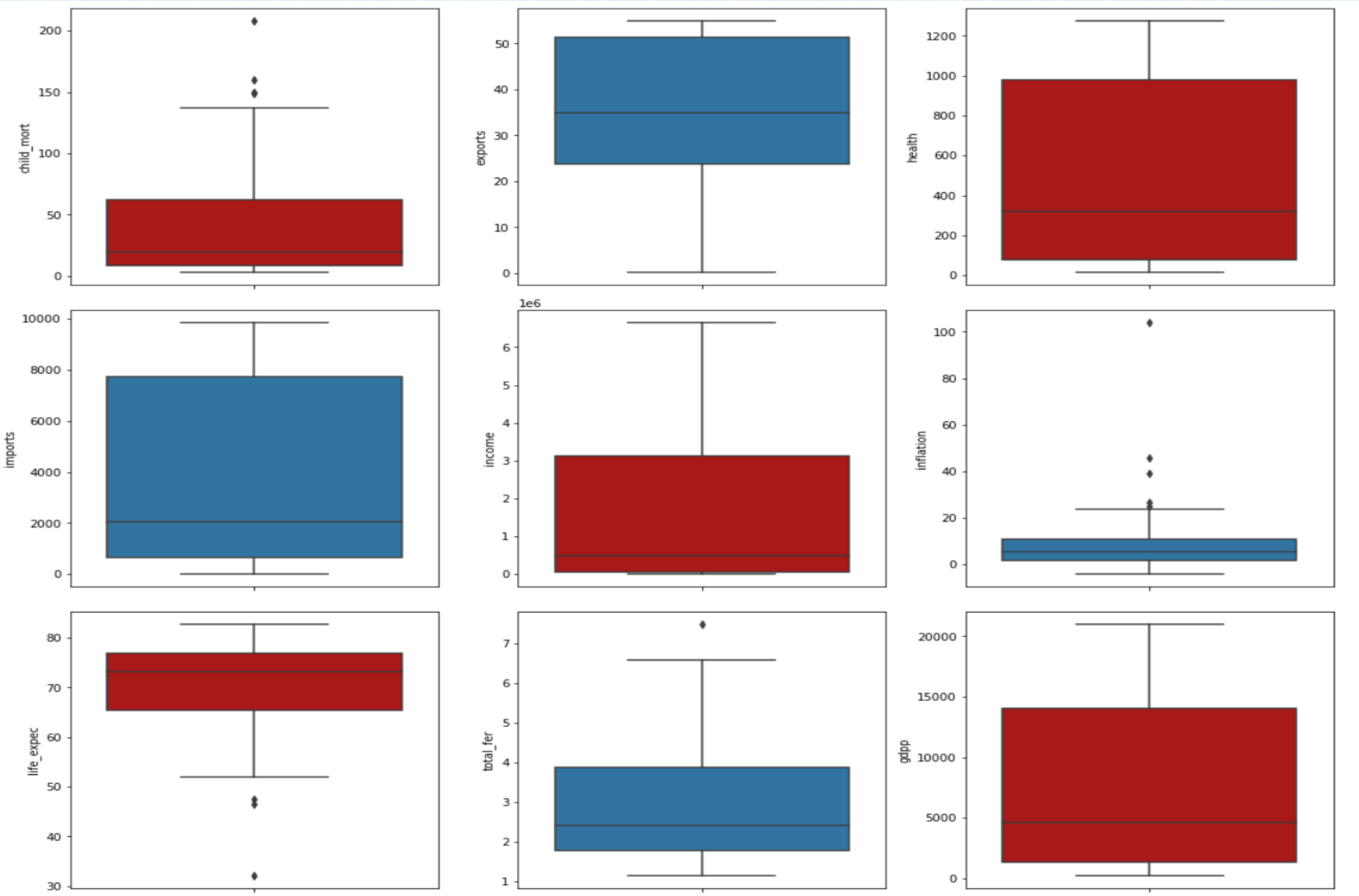
```
country       0
child_mort    0
exports       0
health        0
imports       0
income        0
inflation     0
life_expec    0
total_fer     0
gdpp          0
```

There were no missing values in the data.

CHECKING THE OUTLIERS



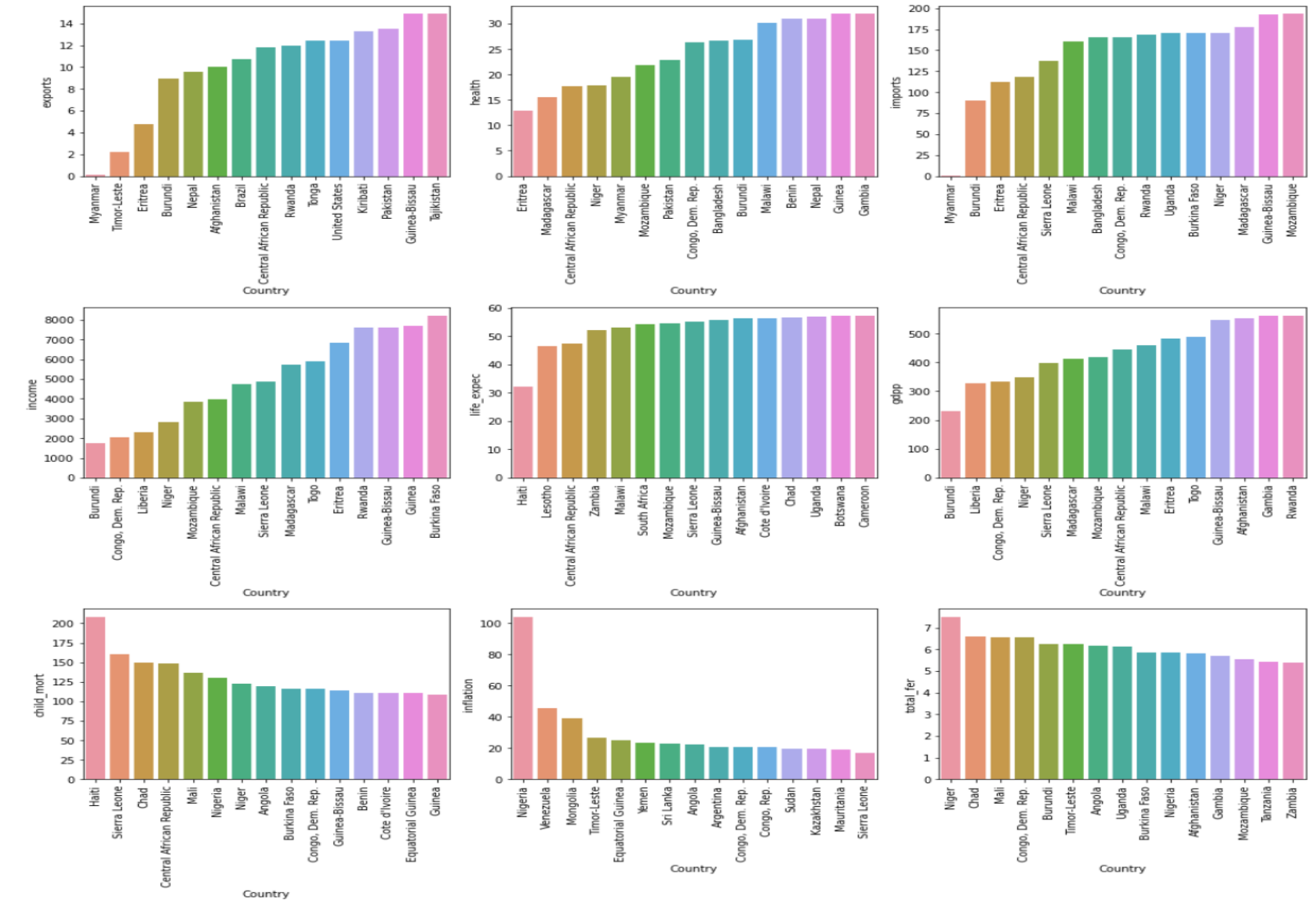
There are outliers in the data. Now, I can either remove the outliers or decide to keep them and go on with the analysis. Now, the purpose of the analysis is to find under developed and poor countries, and for this purpose, I have capped the variables gdpp, income, exports, health and didn't floor them because flooring causes lost in information about the countries that concern us.



The data looks much cleaner of the analysis now as there don't seem to be outliers in the data.

UNIVARIATE ANALYSIS

- I wanted to check what are the bottom countries for each of the variables and I used bas charts to visualize that.



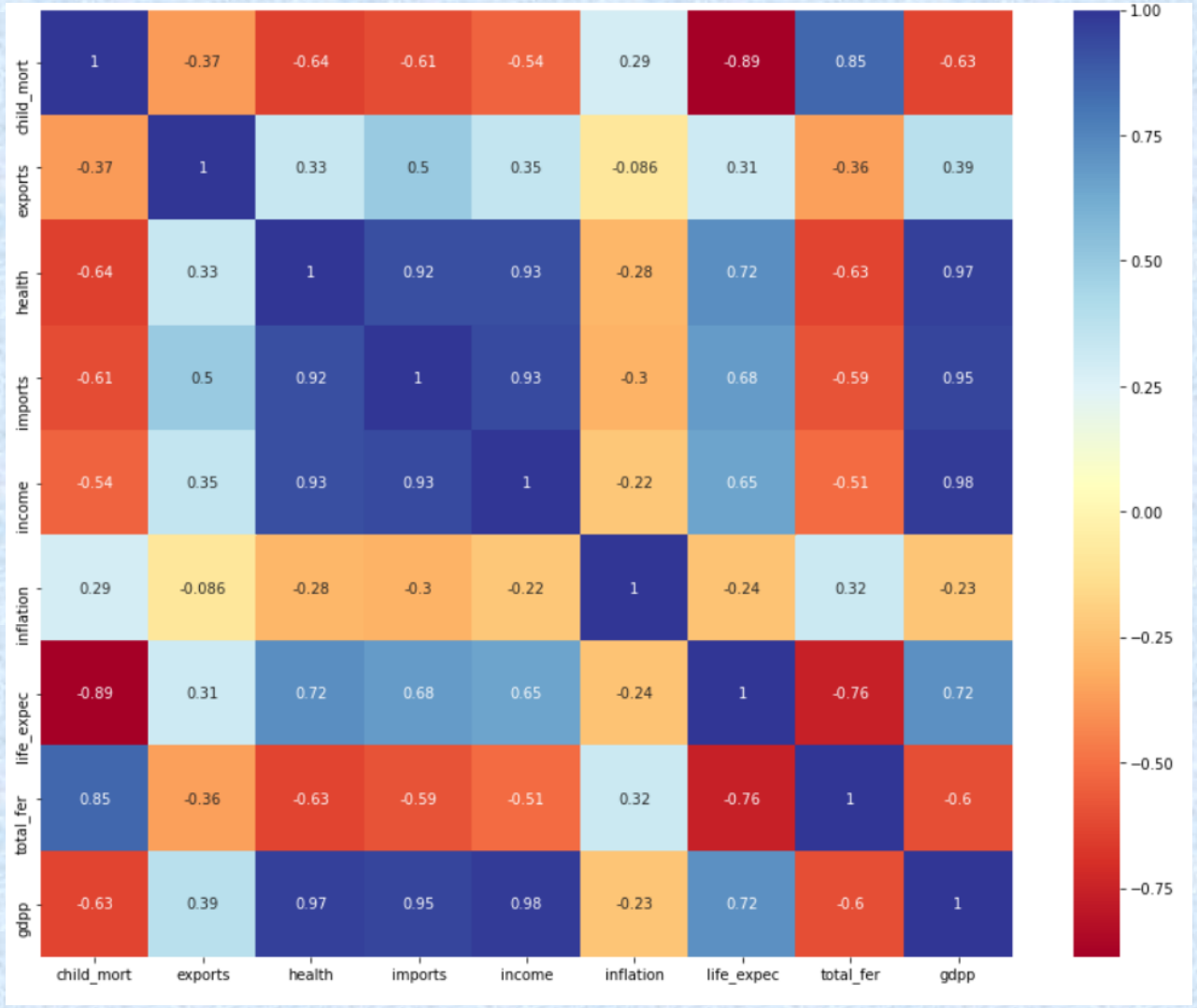
I made a list of countries that appeared in the above graphs and notes the countries that appear in 2 or more graphs. These are the countries I label as backwards using EDA. Later we will check how many of these countries will appear using clustering.

The countries are:-

'Sierra Leone', 'Congo, Dem. Rep.', 'Zambia', 'Central African Republic', 'Gambia', 'Equatorial Guinea', 'Nigeria', 'Timor-Leste', 'Afghanistan', 'Bangladesh', 'Eritrea', 'Uganda', 'Madagascar', 'Haiti', 'Mali', 'Myanmar', 'Chad', 'Burkina Faso', "Cote d'Ivoire", 'Benin', 'Togo', 'Angola', 'Malawi', 'Nepal', 'Mozambique', 'Niger', 'Rwanda', 'Guinea', 'Burundi', 'Guinea-Bissau', 'Pakistan', 'Liberia'

BIVARIATE ANALYSIS

- I plotted a heatmap and pairplot to find out how the different factors are related to each other.



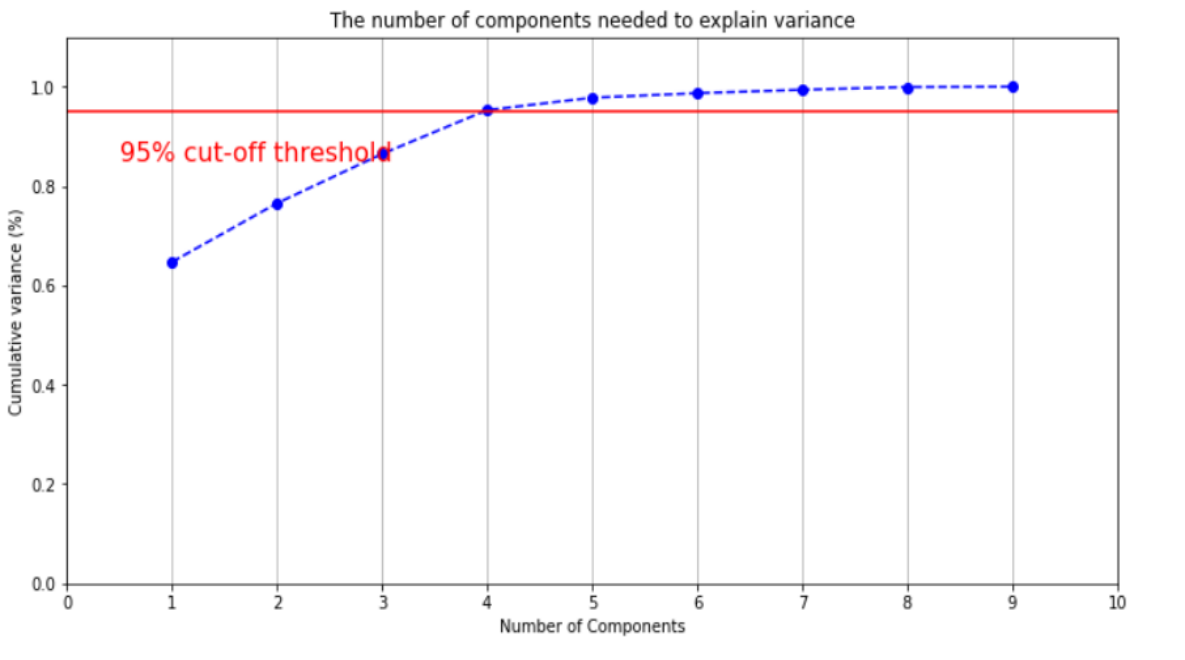
• There is a very high negative correlation between life expectancy and child_mort, total_fer and life expectancy.

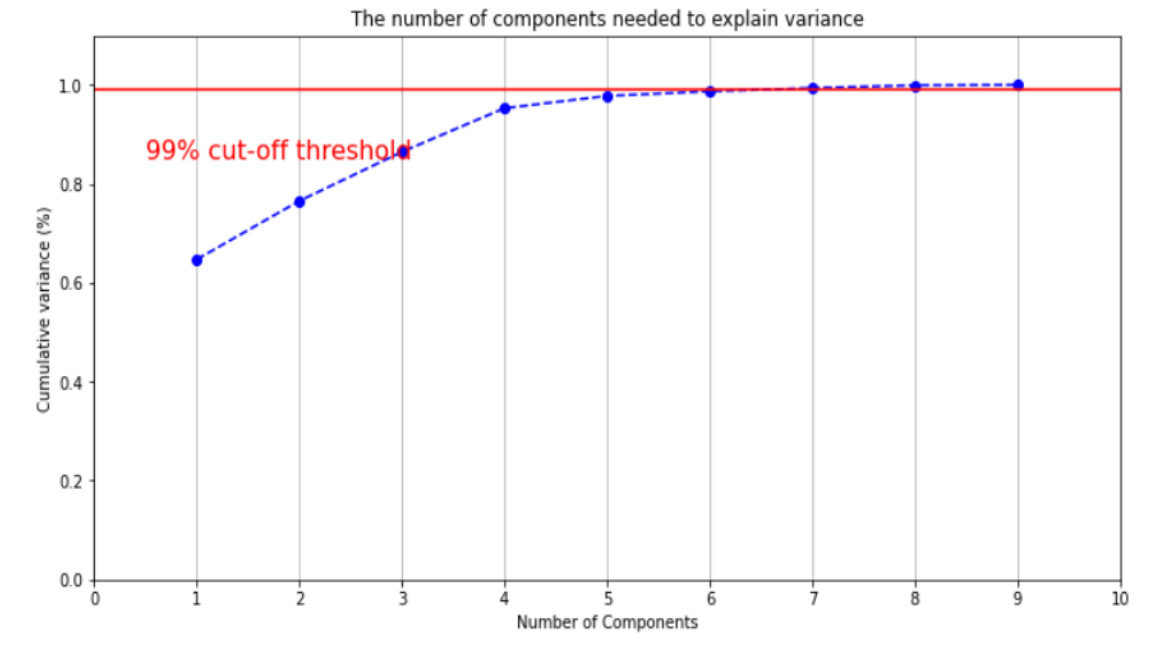
• There is a very high positive correlation between income and gdpp, life_expec and income, imports and exports
I researched a bit about whether or not collinearity a problem in clustering and how the problem differs from multi-collinearity issue in regression.

• When variables used in clustering are collinear, some variables get a higher weight than others. If two variables are perfectly correlated, they effectively represent the same concept. But that concept is now represented twice in the data and hence gets twice the weight of all the other variables. The final solution is likely to be skewed in the direction of that concept, which could be a problem if it's not anticipated. In the case of multiple variables and multi-collinearity, the analysis is in effect being conducted on some unknown number of concepts that are a subset of the actual number of variables being used in the analysis. So our approach to the analysis could either be –

• Dropping some of the variables
- **Using feature reduction techniques like- PCA.**
- **Dropping information results in loss of information and hence we'll try to explore PCA.**

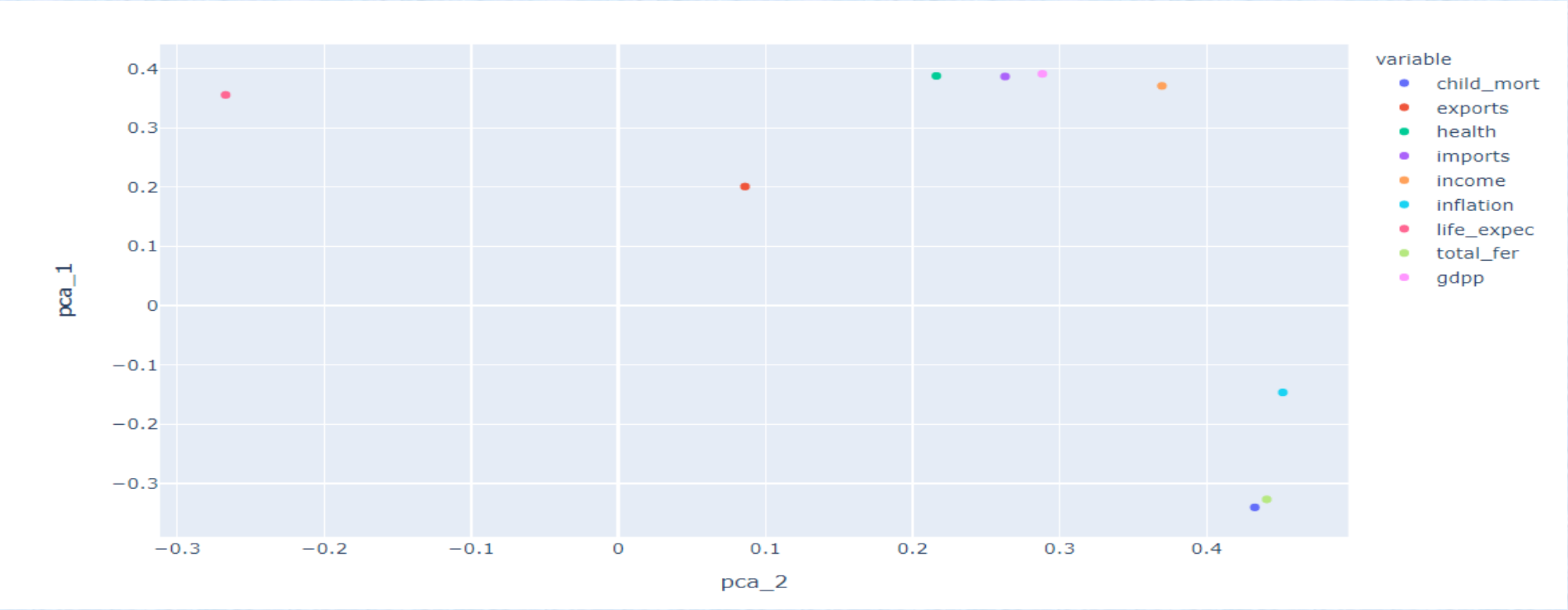
- **Then I standardized the data using Standard Scaler.**
- **After that, I went ahead with PCA. For PCA, I first needed to decide how many number of the components I ant to include in my analysis. I decided it using a screeplot.**





- **Choosing 4 components can explain 95% of the variance in the data and 5 components can explain 99% of the variance in the data and hence we'll go ahead with 5 components.**

- For the sake of some inspection, I tried visualizing PCA 1 with PCA 2.



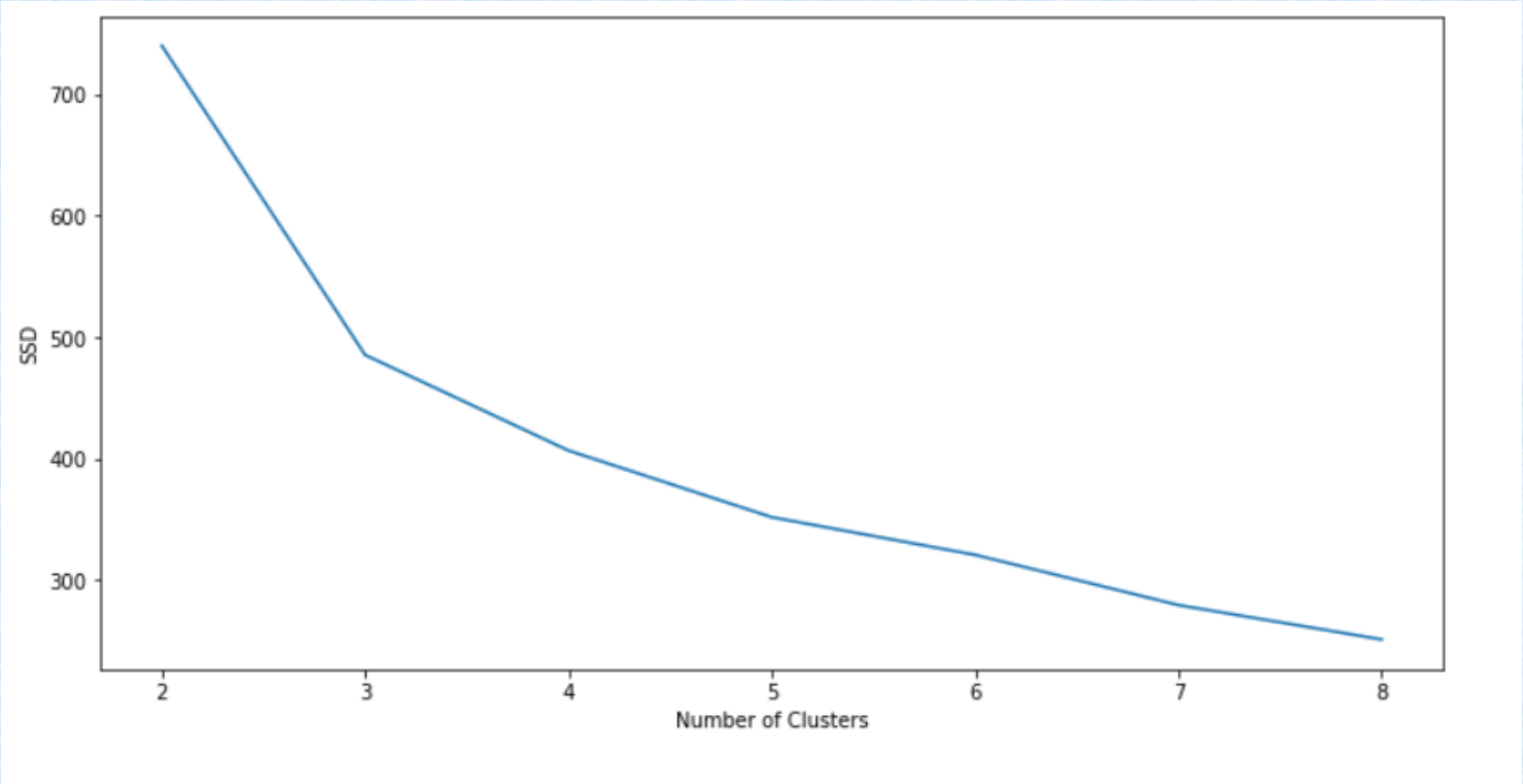
- Pca_1 apparently has the a high value of life expectancy and Pca_2 has a high value of child mortality.
- Now that our data is ready, we'll go ahead with the clustering. Before starting with clustering, we'll first check whether or not the data is suitable for clustering using the Hopkins statistics. I calculated John Hopkins 10 times to get the following result:

[0.851400191092 37, 0.8693026247673185, 0.787180505500869, 0.8165709195215118, 0.7623759304716791, 0.8257843437226186, 0.8415115783913175, 0.8821297039290928, 0.8397255074101452, 0.8180877489196433]

- As all the values lie between {0.7, ..., 0.99}, the data has a high tendency to cluster.

Before modelling, we will decide the optimal number of clusters. It can be done using 2 approaches:-

- Using SSD
- Using Silhouette Analysis

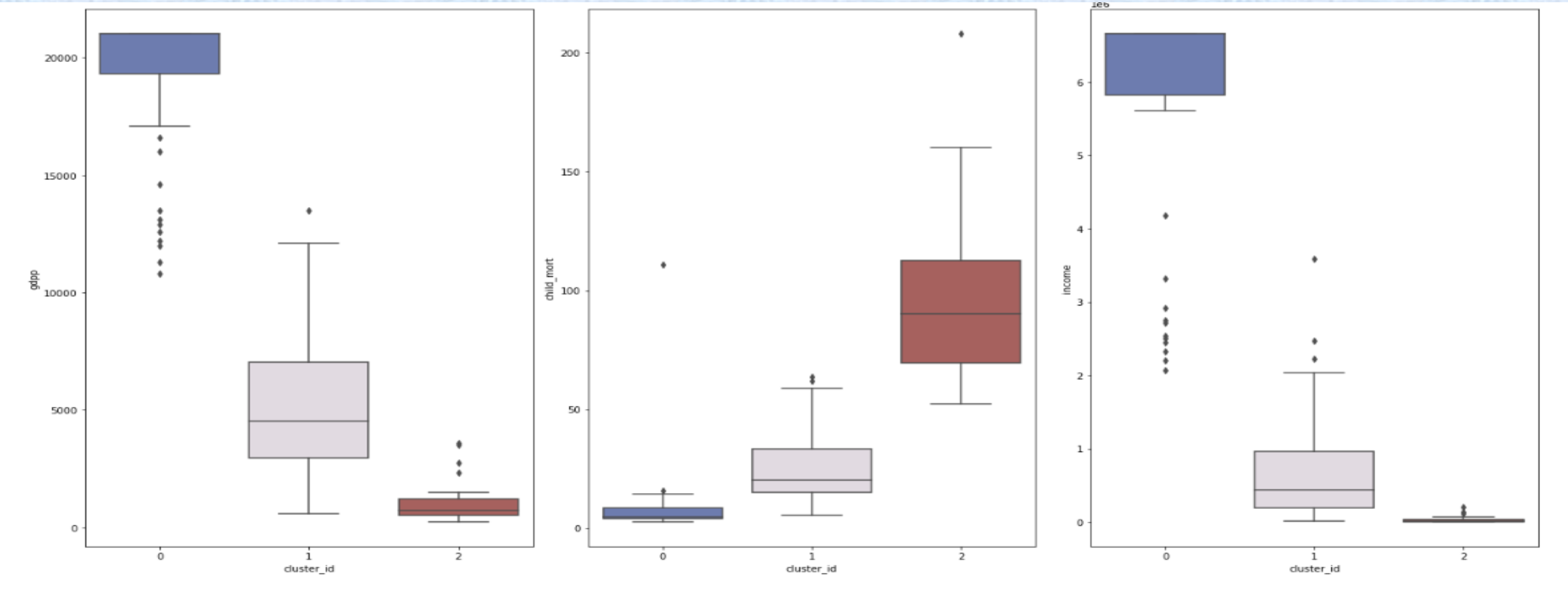


- For n_clusters=2, the silhouette score is 0.4551979910680447
- For n_clusters=3, the silhouette score is 0.4214605163153466
- For n_clusters=4, the silhouette - score is 0.4251436949848127 5
- For n_clusters=5, the silhouette score is 0.3597928789586653 5
- For n_clusters=6, the silhouette score is 0.3455373384573384 7
- For n_clusters=7, the silhouette score is 0.3050765503741525 4
- For n_clusters=8, the silhouette score is 0.2845075343171167

From both SSD and Silhouette Analysis, I conclude that 3 is the optimal number of clusters for our analysis.

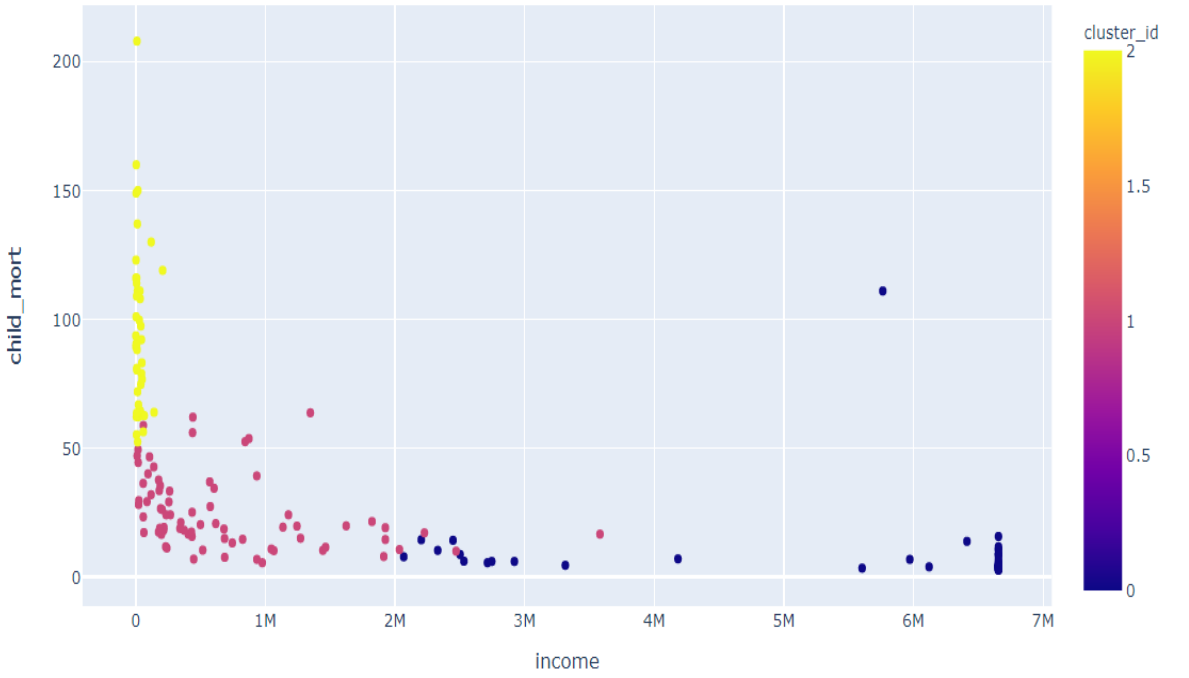
K-MEANS CLUSTERING

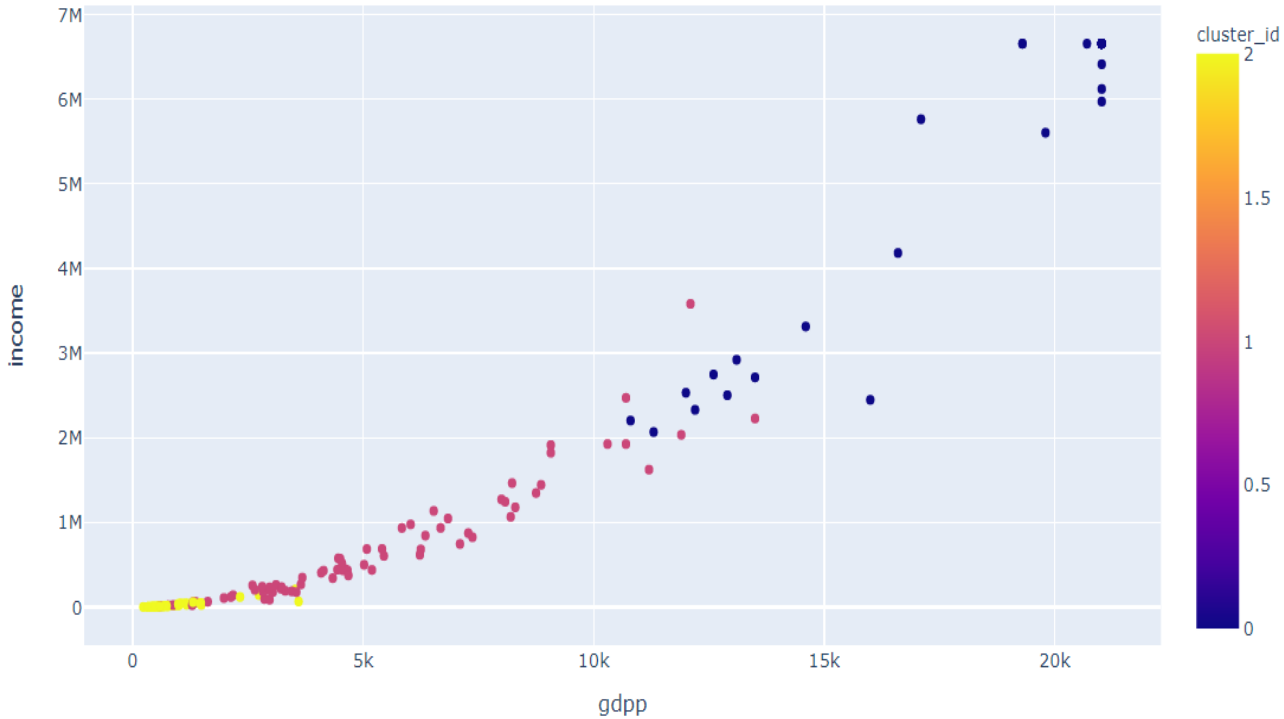
- I made three clusters using K-Clusters and plotted the gdpp, child mortality, and income for different clusters using boxplot.

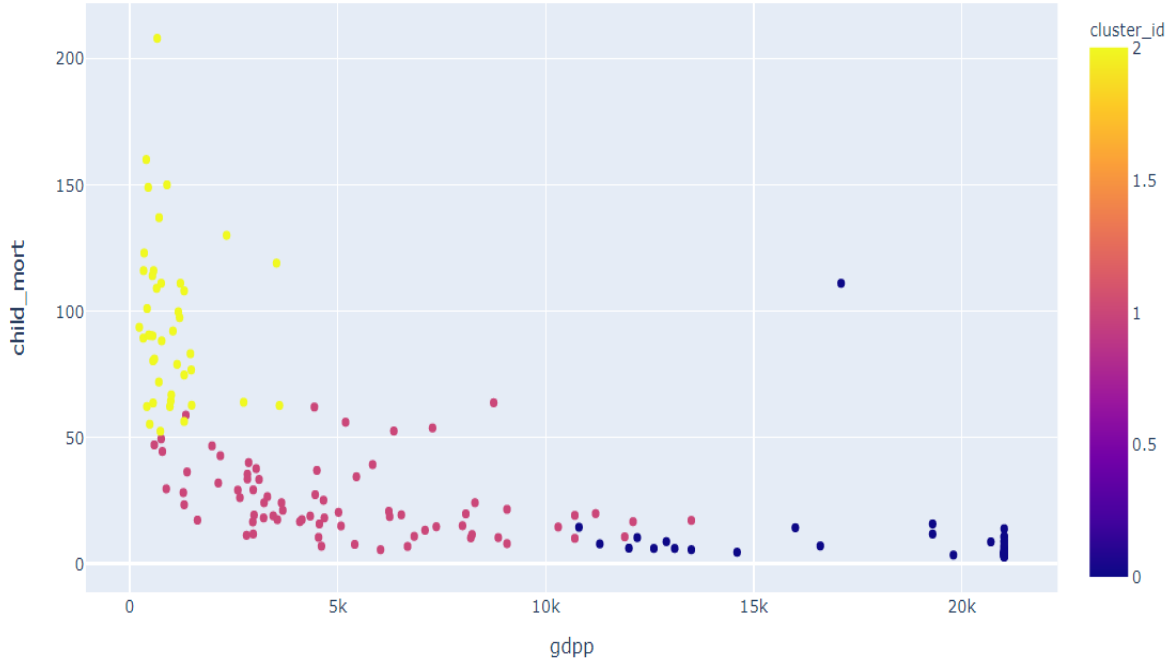


- It's clearly visible from the boxplots that cluster 2 consists of countries that have low income and gdpp but a very high child mortality rate and hence this cluster is the cluster of interest for us.

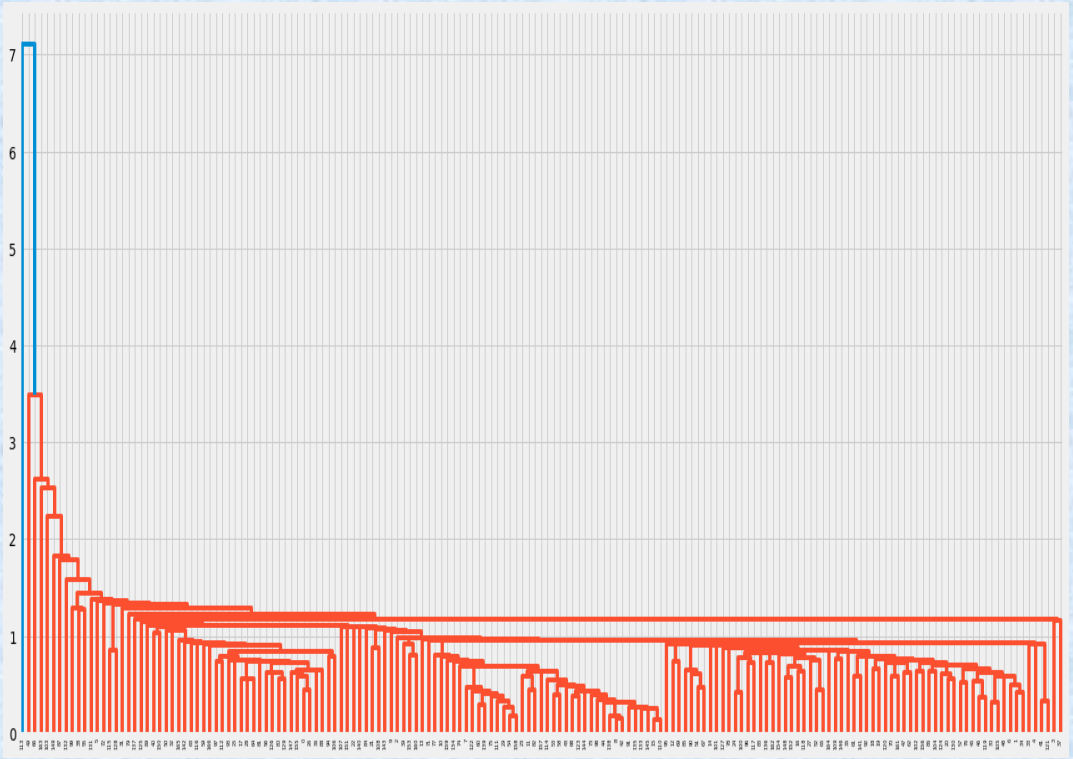
Next we can visualize the clusters formed for all the three variables pairwise.



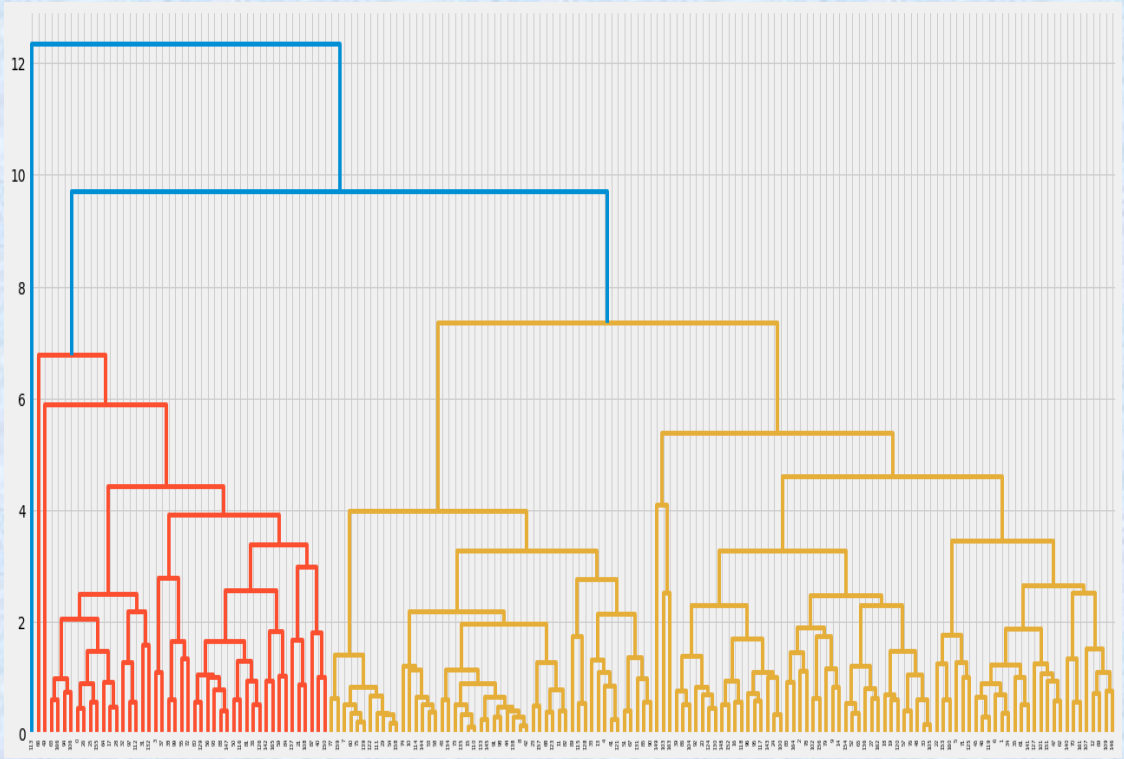




HIERARCHICAL CLUSTERING



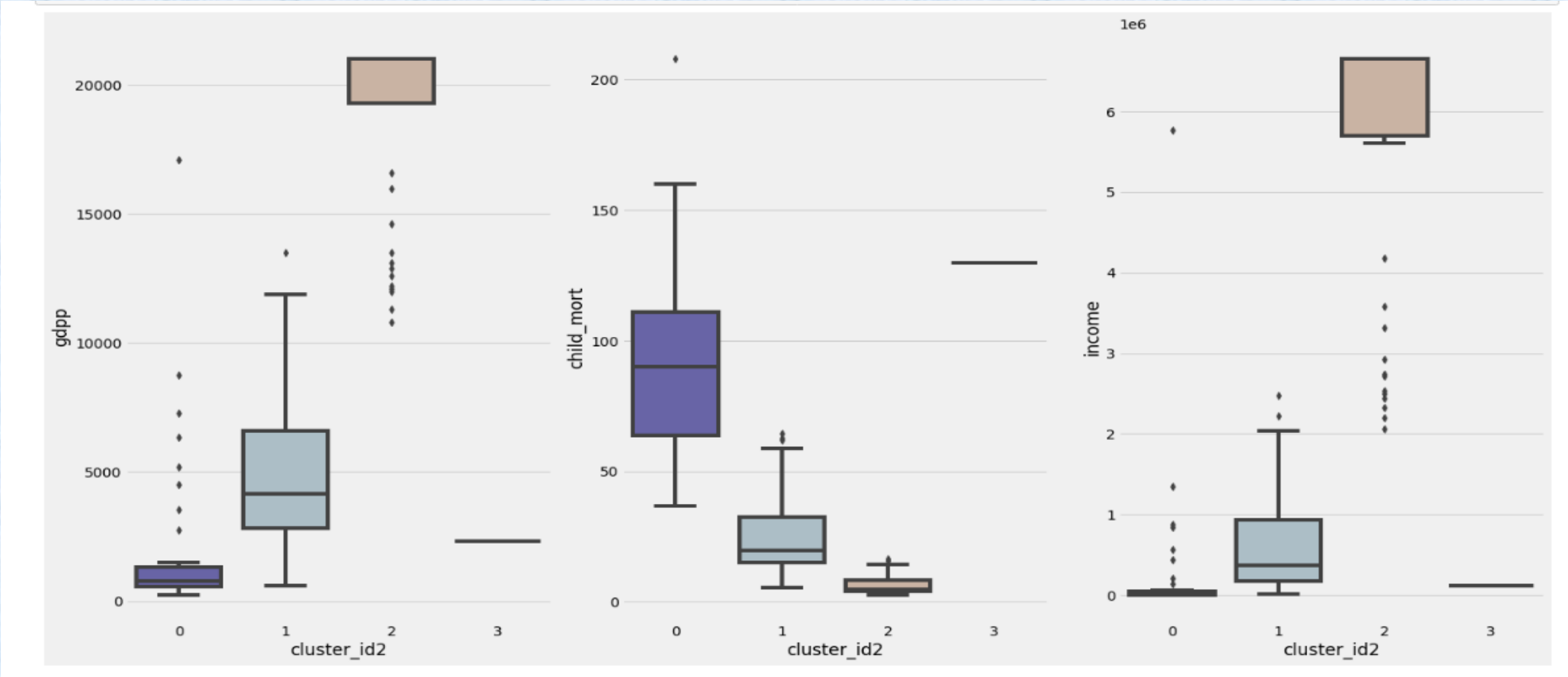
Single Linkage



Complete Linkage

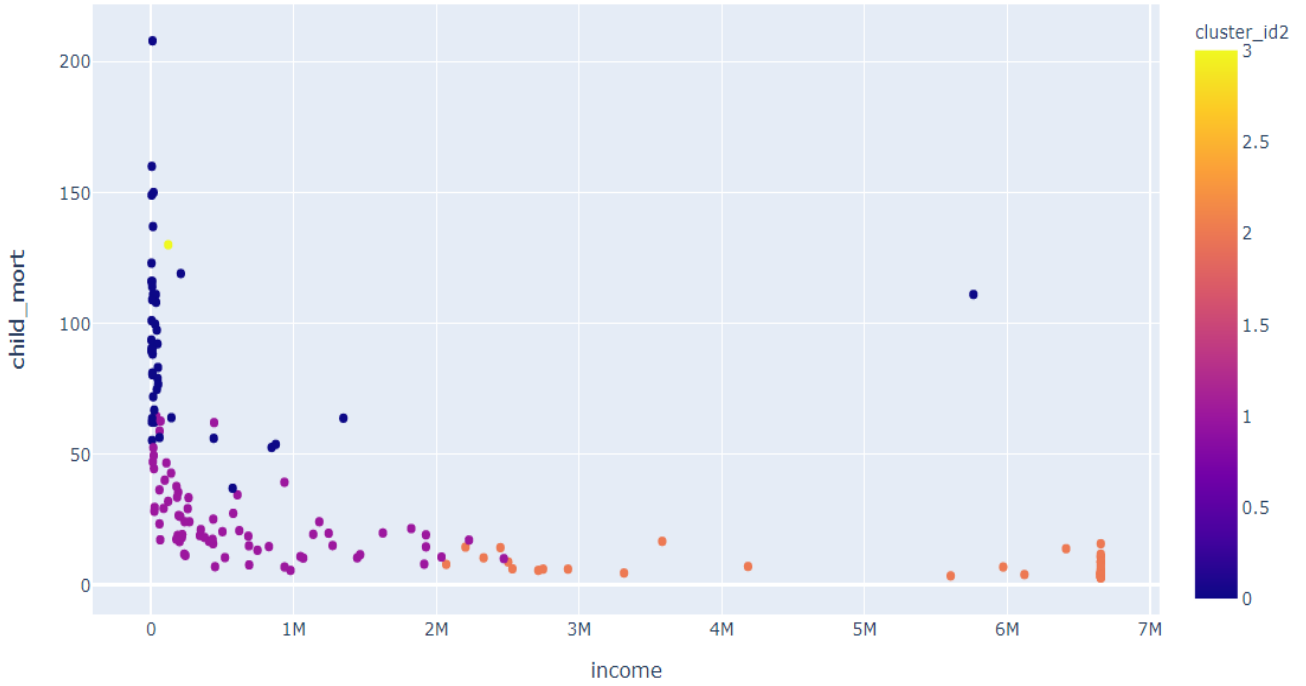
Complete linkage has a much better dendrogram that the simple linkage. Also, in complete linkage, we can either choose number of clusters to be 2 or 4. I'd go with 4 as 2 will result a large number of countries.

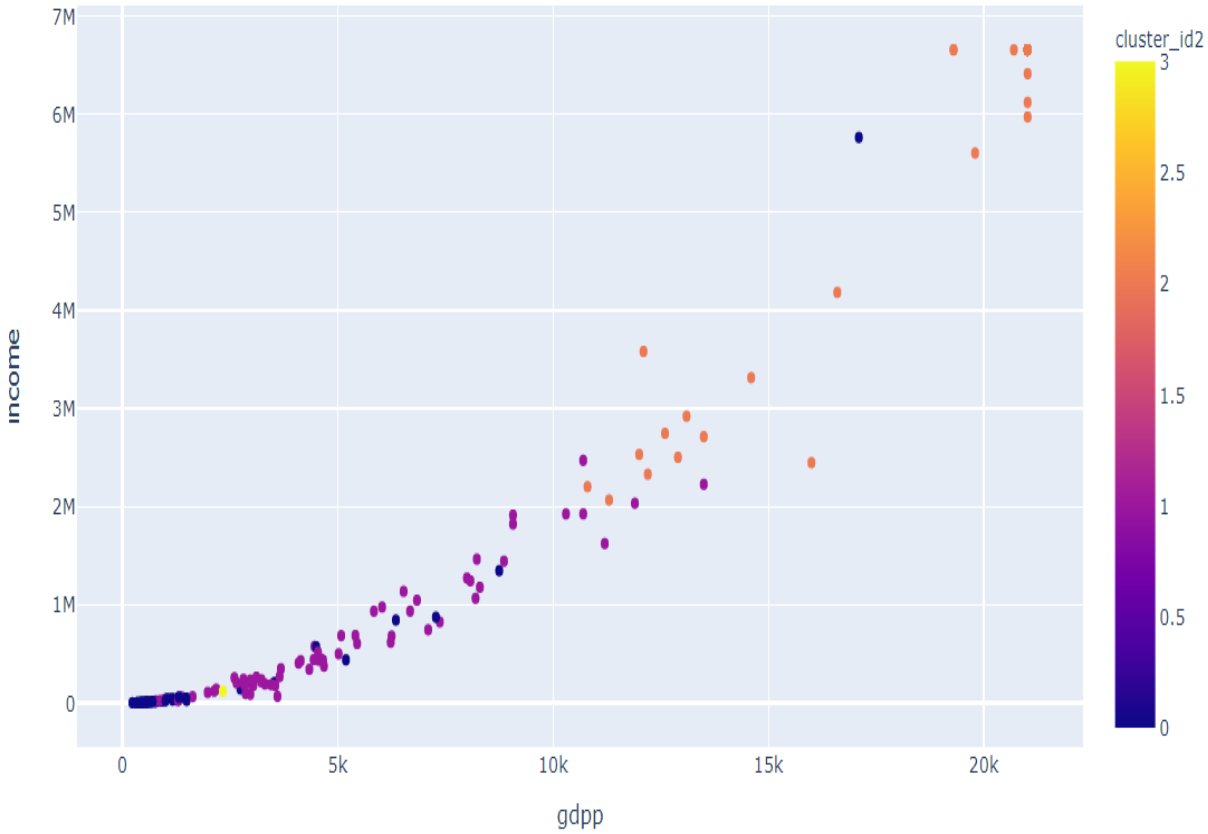
- I made three clusters using Hierarchical Clustering and plotted the gdpp, child mortality, and income for different clusters using boxplot.

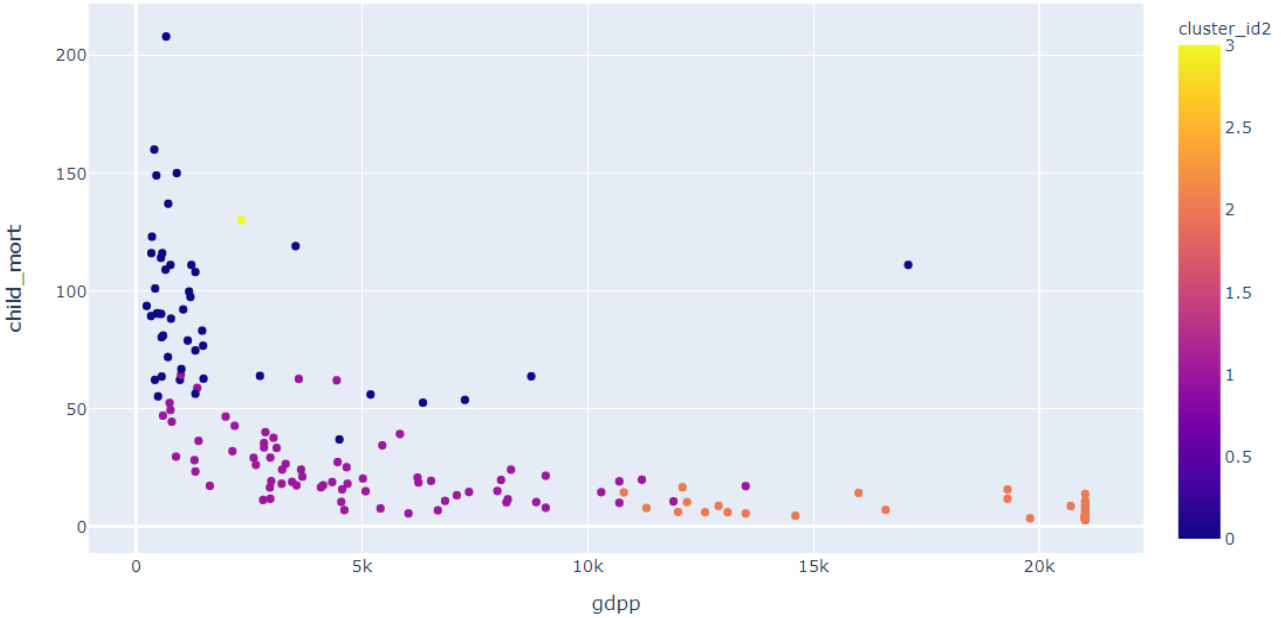


- It's clearly visible from the boxplots that cluster 0 consists of countries that have low income and gdpp but a very high child mortality rate and hence this cluster is the cluster of interest for us.

Next we can visualize the clusters formed for all the three variables pairwise.







CHOOSING THE COUNTRIES

- **Countries we got using K Means Clustering:**

```
['Afghanistan', 'Angola', 'Benin', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo,
Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Eritrea', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Kenya',
'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Myanmar', 'Niger',
'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Sudan', 'Tajikistan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda',
'Yemen', 'Zambia']
```

- **Countries we got using hierarchical clustering**

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

```
Number of countries from k-means: 43
Number of countries from hierarchical: 45
```

- **Countries that we got using hierarchical but not by K Means**

```
{'Botswana', 'Equatorial Guinea', 'Gabon', 'Iraq', 'Namibia', 'South Africa'}
```

- **Countries that we got using K Means but not hierarchical**

```
{'Myanmar', 'Nigeria', 'Tajikistan', 'Timor-Leste'}
```

- **We'll check which of these countries are present in the backward countries we calculated using the EDA.**

```
{'Botswana', 'Equatorial Guinea', 'Myanmar', 'Nigeria', 'South Africa', 'Tajikistan', 'Timor-Leste'}
```

- I'll add these countries to the countries intersection of the countries we got using the hierarchical and k-means clustering.

Countries in need to be funded are:-

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

There are 46 countries that are been identified in the need of the fundings. But as we need to decrease the number of countries suggested so hat we could give more help to give a proper funding to these countries. As according to the business need and the questions, I feel that GDPP, Child Mortality and Income are the somewhat more important variables here, I'll delete those countries that have more than median GDPP and Income and Less than median Child Mortality.

We got our 5 countries by doing so.

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_id2
Central African Republic	149.0	11.8	17.7508	118.190	3960.48	2.01	47.5	5.21	446.0	2	0
Congo, Dem. Rep.	116.0	41.1	26.4194	165.664	2034.06	20.80	57.5	6.54	334.0	2	0
Niger	123.0	22.2	17.9568	170.868	2832.72	2.55	58.8	7.49	348.0	2	0
Mozambique	101.0	31.5	21.8299	193.578	3846.42	7.64	54.5	5.56	419.0	2	0
Sierra Leone	160.0	16.8	52.2690	137.655	4867.80	17.20	55.0	5.20	399.0	2	0

CONCLUSION

- Countries in the need of financial aid are:-

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

- Countries in dire need of financial aid are:-
- 1. Congo, Dem. Rep
- 2. Niger
- 3. Sierra Leone
- 4. Mozambique
- 5. Central African Republic


THANK YOU
THANK YOU
THANK YOU
THANK YOU