Country Clustering

For HELP International

By:
Tanu Shri Pant

Aim

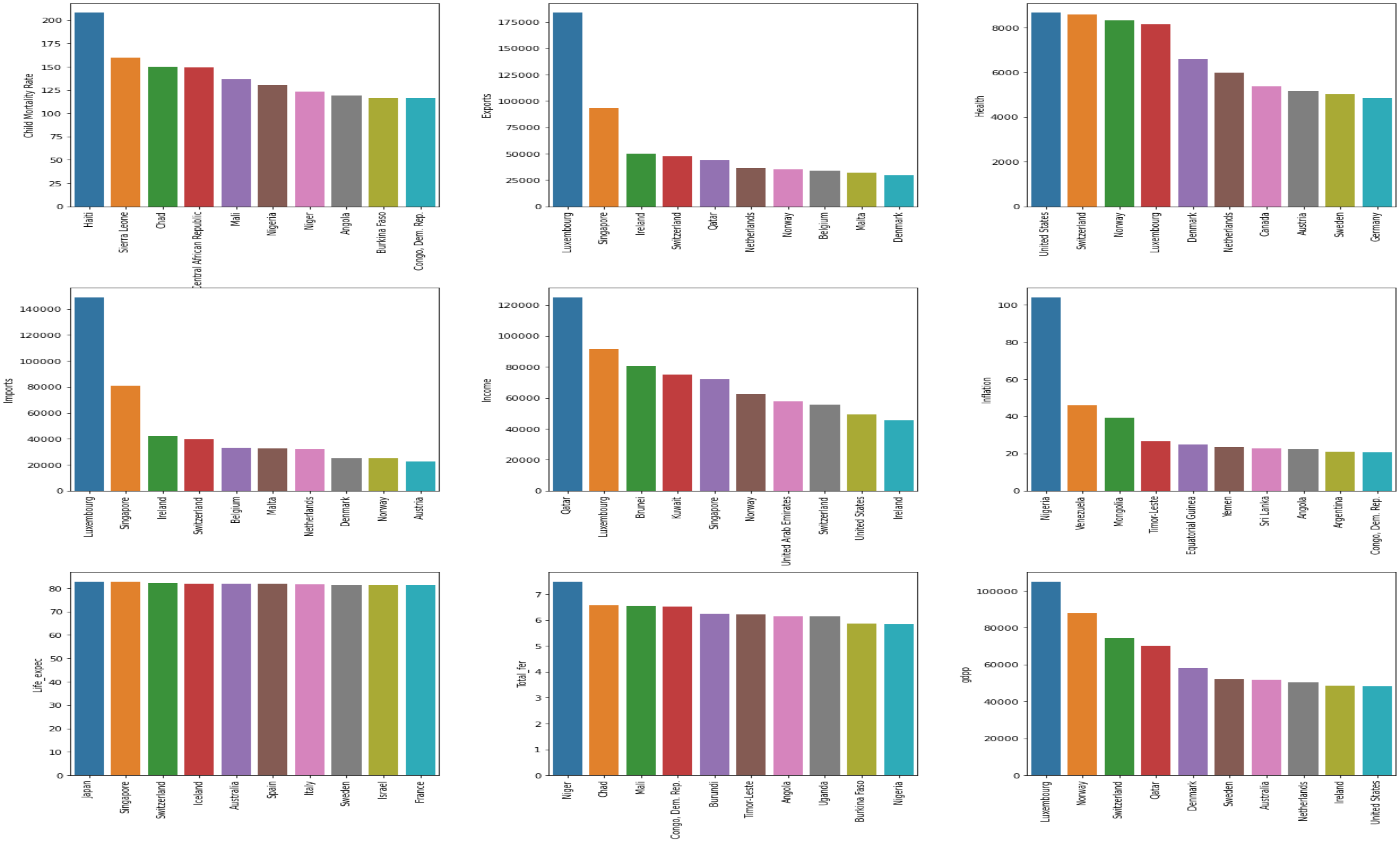
- Strategically and effectively utilise the fund that is raised ($ 10 million)
- Selecting countries that are in direst need of aid.
- Categorising the countries on socio economic and health factors that determine the overall development of the country.

Dataset

- The data set consists of 167 countries.
- Data about the following attributes of the countries has been collected:
    - Child_mort
    - Exports
    - Health
    - Imports
    - Income
    - Inflation
    - Life_Expectancy
    - Total_fer
    - gdpp

Data Cleaning Steps

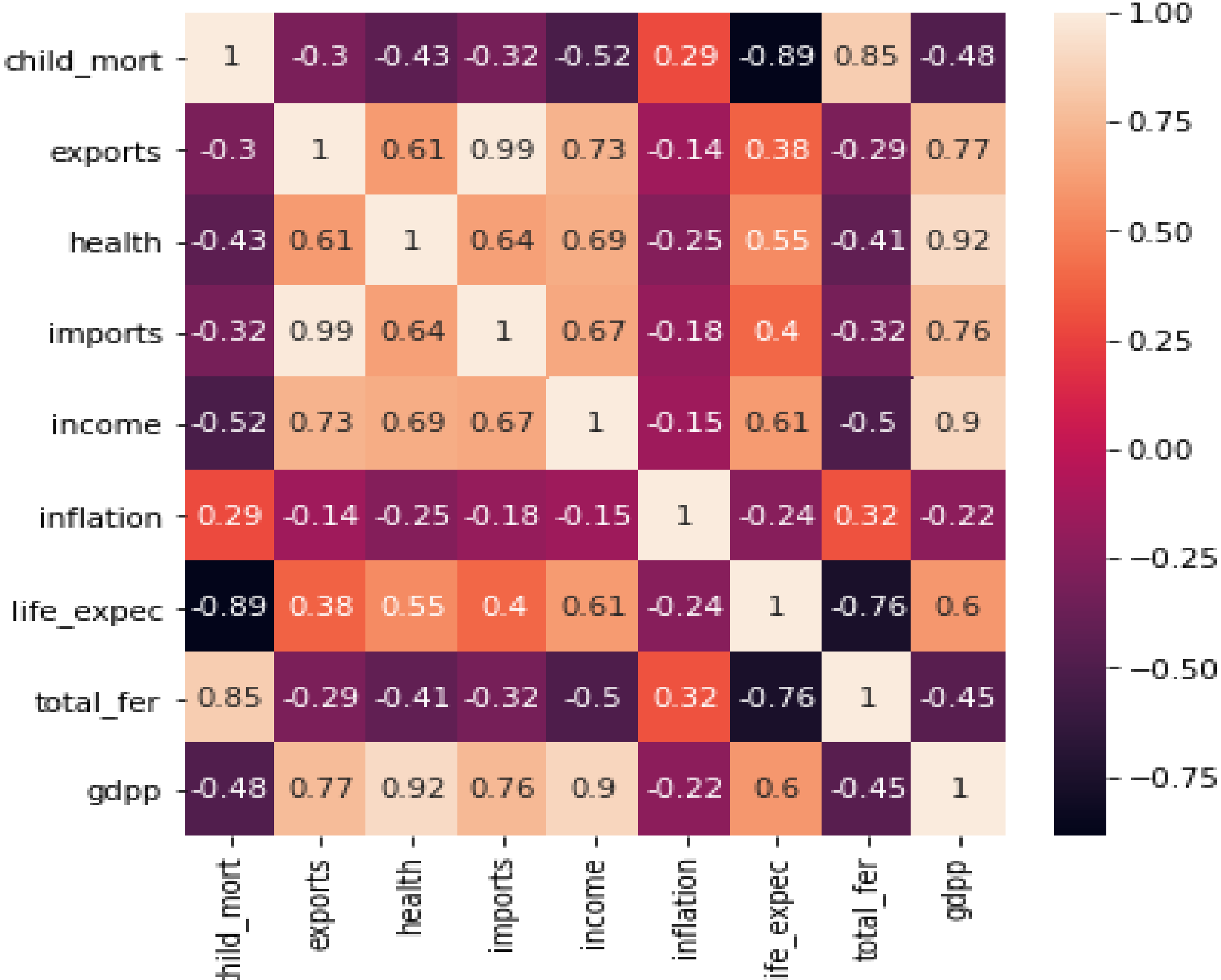
- The dataset did not have any null or missing values present.
- Columns like health, exports, imports were given as percentage of gdp, so converted them to absolute numbers.
- Next page shows plots of top 10 values of attributes with respect to the countries.


Child Mortality Rate
0
25
50
75
100
125
150
175
200
Haiti
Sierra Leone
Chad
Central African Republic
Mali
Nigeria
Niger
Angola
Burkina Faso
Congo, Dem. Rep.
Exports
0
25000
50000
75000
100000
125000
150000
175000
Luxembourg
Singapore
Ireland
Switzerland
Qatar
Netherlands
Norway
Belgium
Malta
Denmark
Health
0
2000
4000
6000
8000
United States
Switzerland
Norway
Luxembourg
Denmark
Netherlands
Canada
Austria
Sweden
Germany
Imports
0
20000
40000
60000
80000
100000
120000
140000
Luxembourg
Singapore
Ireland
Switzerland
Belgium
Malta
Netherlands
Denmark
Norway
Austria
Income
0
20000
40000
60000
80000
100000
120000
Qatar
Luxembourg
Brunei
Kuwait
Singapore
Norway
United Arab Emirates
Switzerland
United States
Ireland
Inflation
0
20
40
60
80
100
Nigeria
Venezuela
Mongolia
Timor-Leste
Equatorial Guinea
Yemen
Sri Lanka
Angola
Argentina
Congo, Dem. Rep.
Life_expec
0
10
20
30
40
50
60
70
80
Japan
Singapore
Switzerland
Iceland
Australia
Spain
Italy
Sweden
Israel
France
Total_fer
0
1
2
3
4
5
6
7
Niger
Chad
Mali
Congo, Dem. Rep.
Burundi
Timor-Leste
Angola
Uganda
Burkina Faso
Nigeria
gdpp
0
20000
40000
60000
80000
100000
Luxembourg
Norway
Switzerland
Qatar
Denmark
Sweden
Australia
Netherlands
Ireland
United States

Observations

- Countries like Haiti, Sierra Leone, Chad, Central African Republic have high Child Mortality Rate
- Countries like Luxembourg, Norway, Switzerland, Qatar have high gdpp and are highly developed.
- Countries like Niger, Chad, Mali, Congo Dem Republic have high fertility rates.

Correlation Matrix



	hild_mort	exports	health	imports	income	inflation	ife_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

Observations

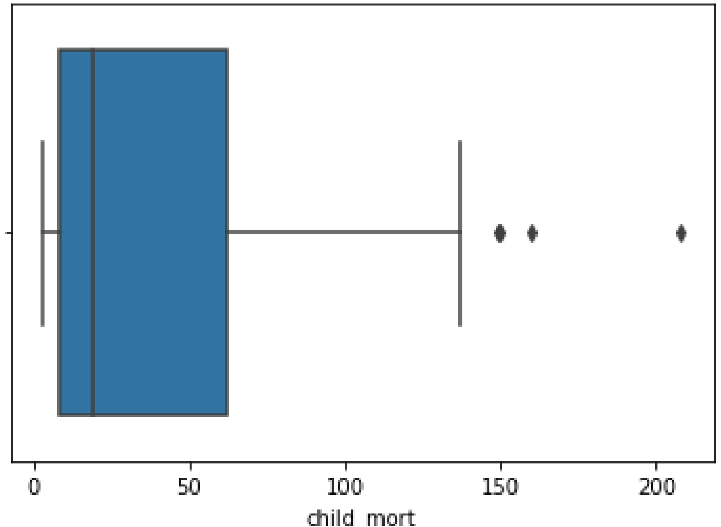
- Imports and exports are highly correlated
- life_expec and child_mort are highly negatively correlated
- gdpp is positively correlated with the health of citizens
- total_fer is highly positively correlated child_mor

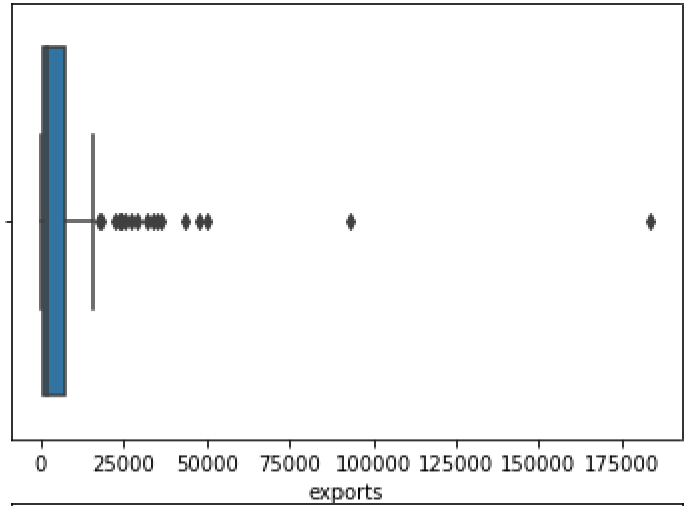
Outlier Treatment

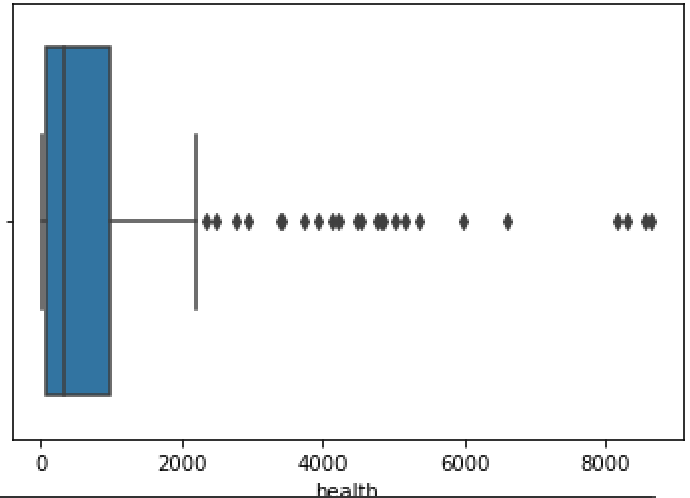
- Performed lower capping for variables like child_mortality, inflation, total_fert.

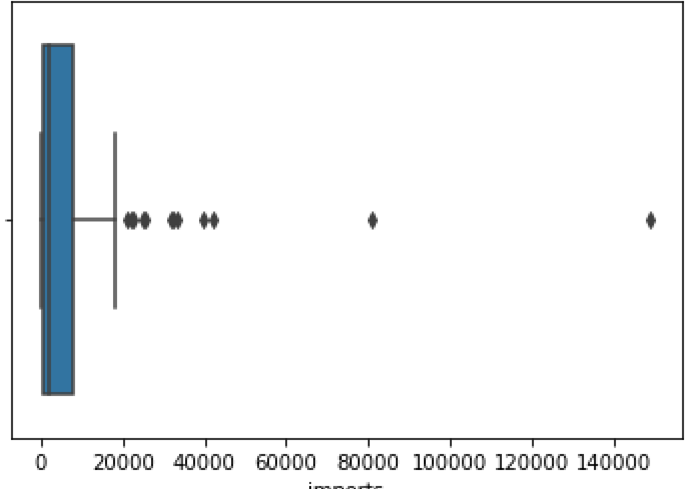
- For other attributes like gdpp, health performed upper capping as they could affect our quality of clustering and such values are not on scope of our analysis.

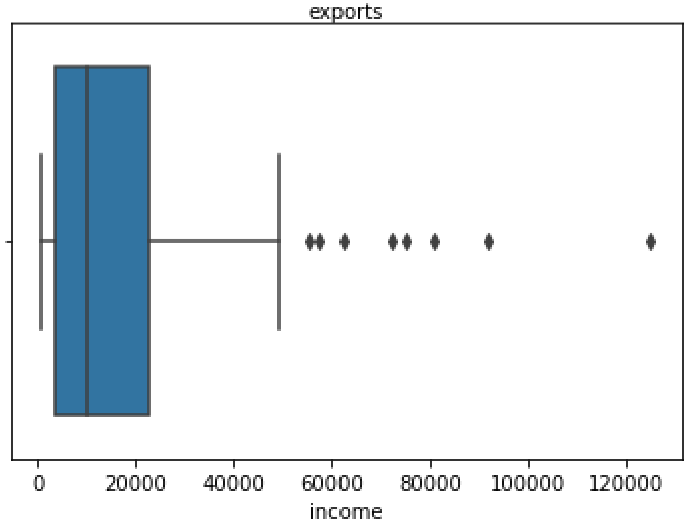
Boxplots before capping

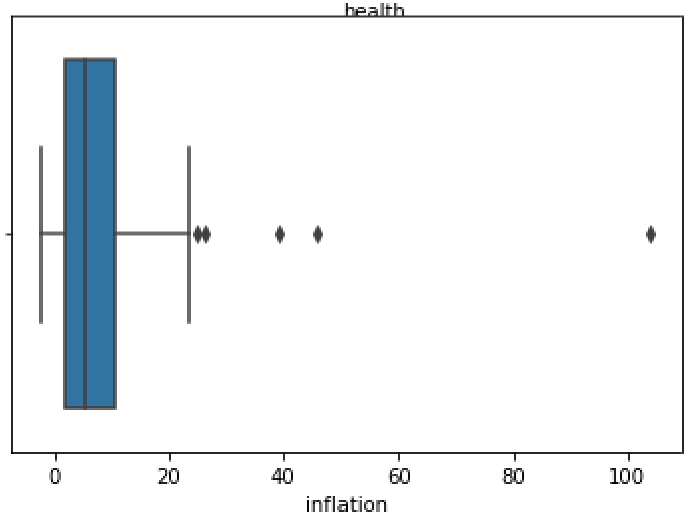


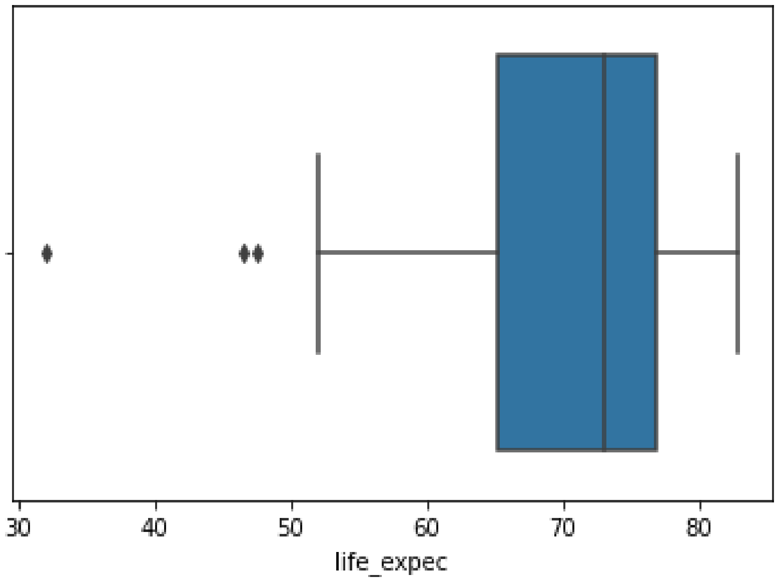


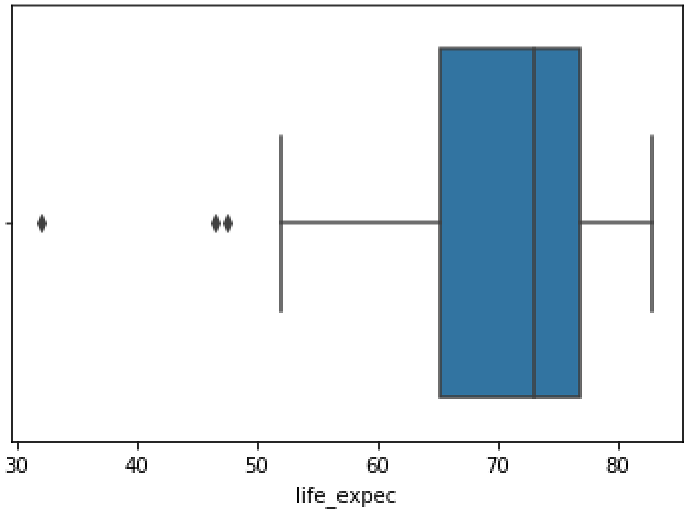


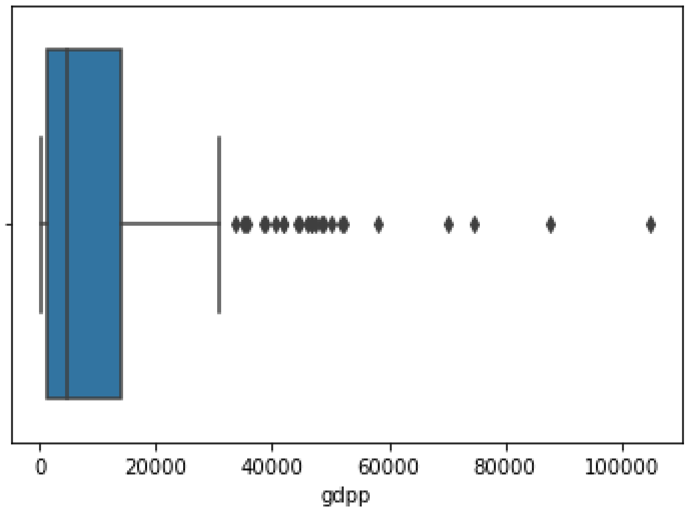







30
40
50
60
70
80
life_expec


30
40
50
60
70
80
life_expec


0
20000
40000
60000
80000
100000
gdpp

Measuring the cluster tendency of data

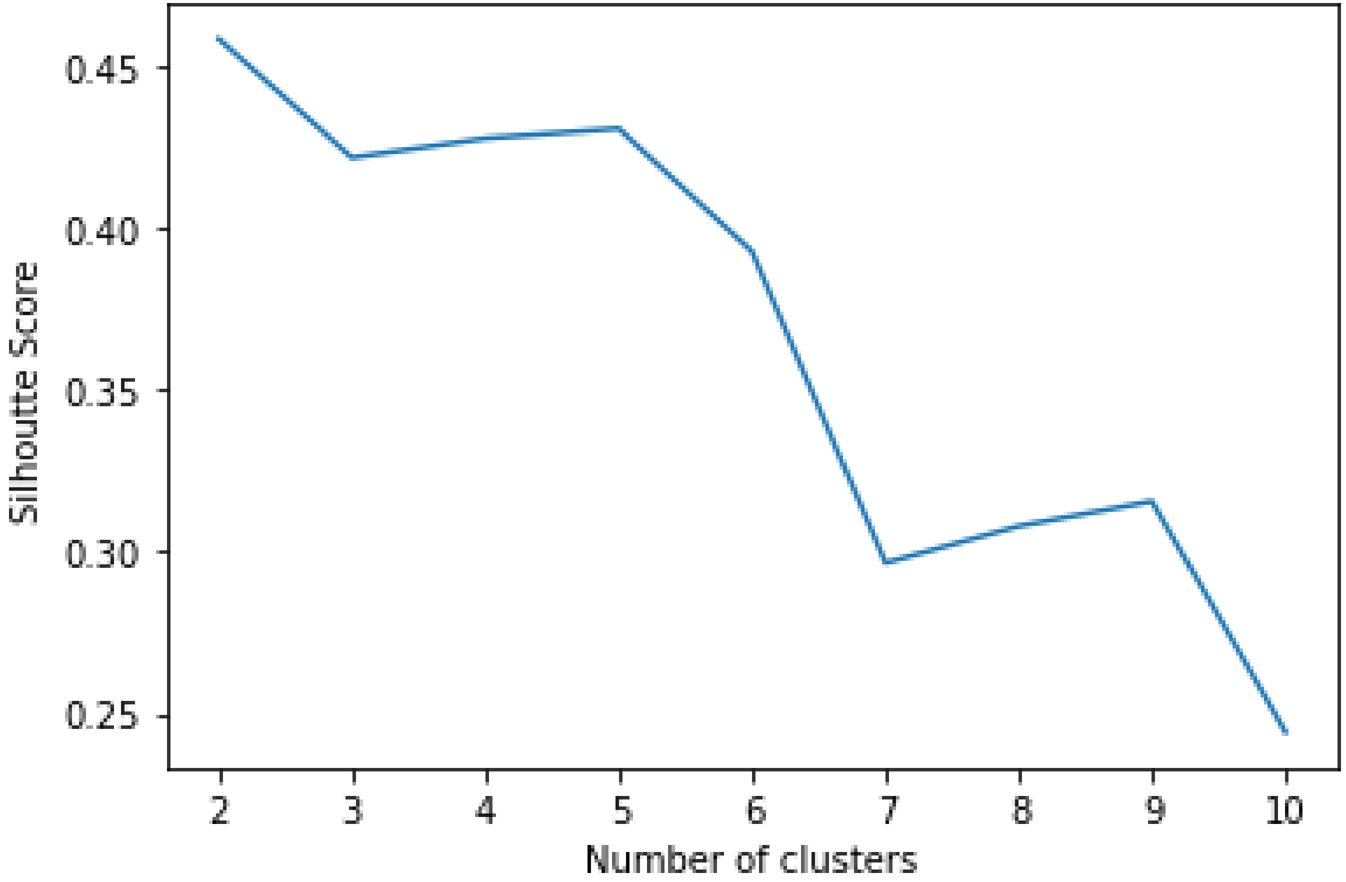
- Conducted Hopkins Test to determine how much the data is suitable for clustering
- Hopkins statistic coming out to be > 85% which is a good measure.

Finding the best value of k

- Used two measures to decide the value of k:
  - SSD (Sum of Squared Distances)
  - Silhouette Score

Silhouette score graph

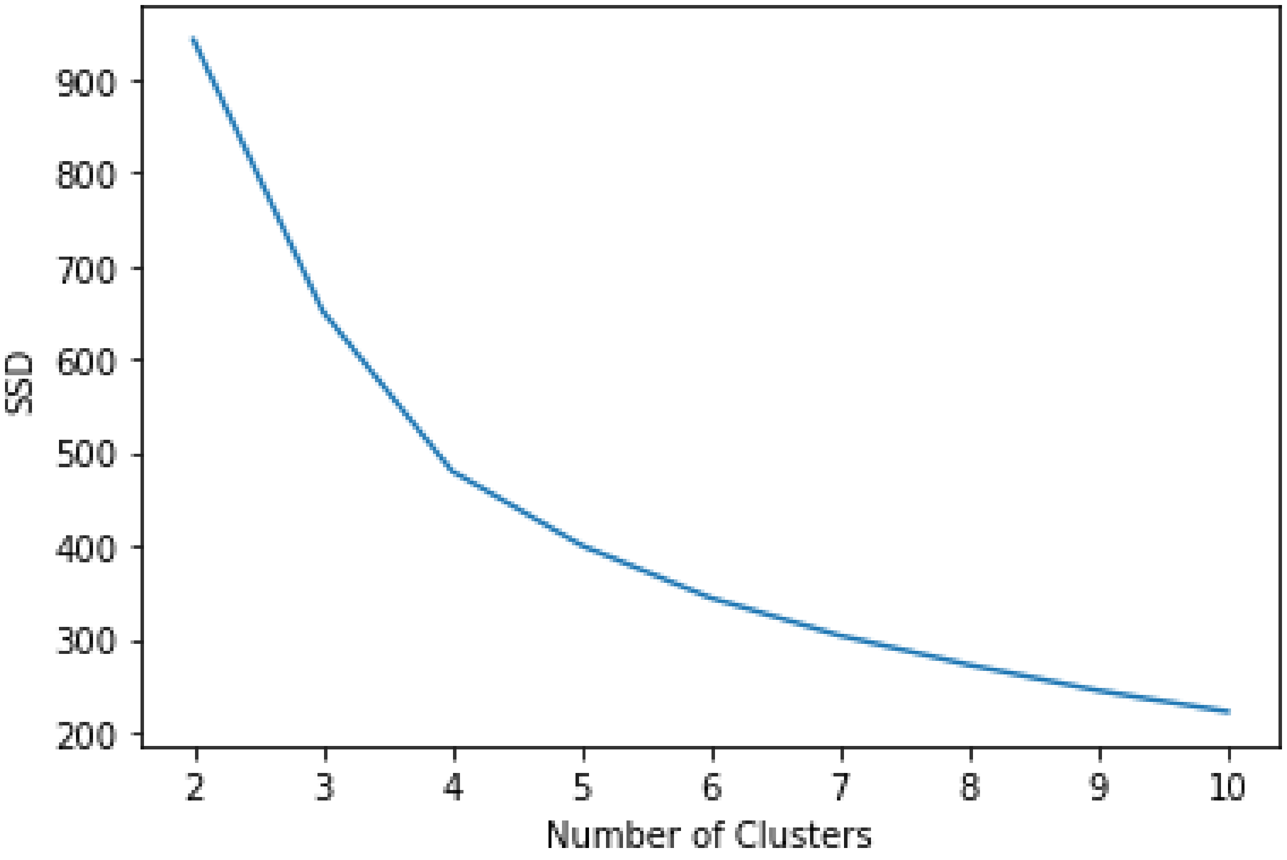
At k = 5 we see a high silhouette score



Sum of Squared Distances (Elbow Curve)

- We see a bend (elbow) in the plot at k = 3 and a slight bend at k = 5.

This choosing k = 4 as number of clusters.



Performed K means on various k values

- For k = 5 following was the cluster distribution:

Cluster ID	Number of countries
0	30
1	86
2	3
3	47
4	1

- K = 4

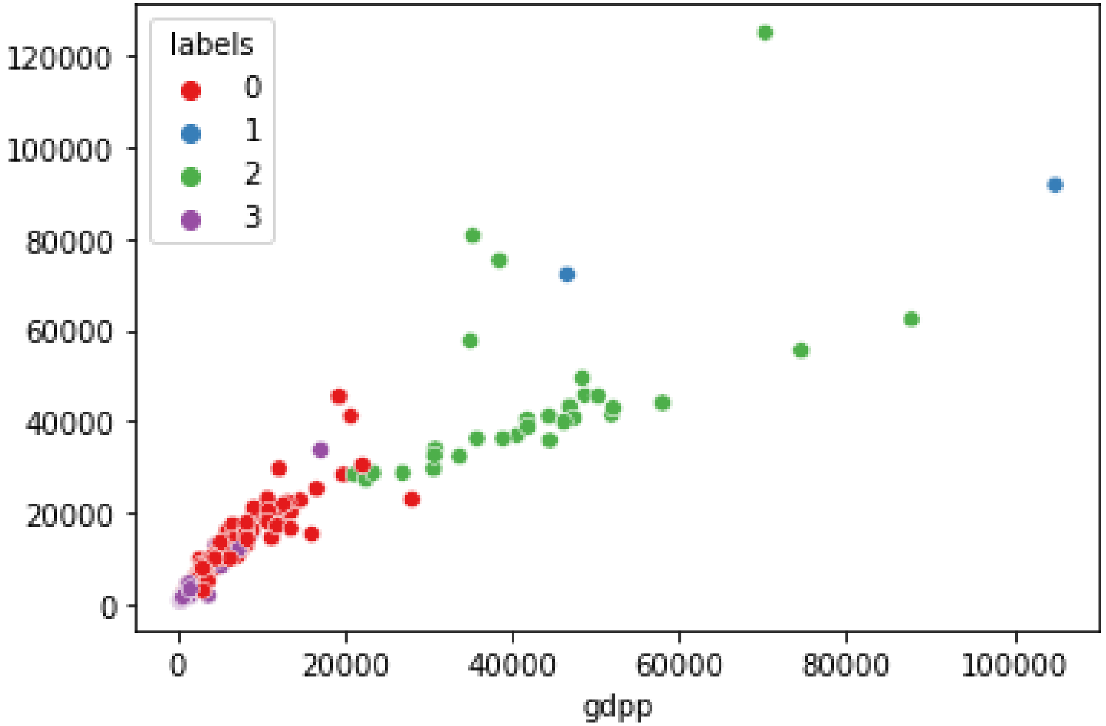
Cluster ID	Number of Countries
0	87
1	2
2	30
3	48

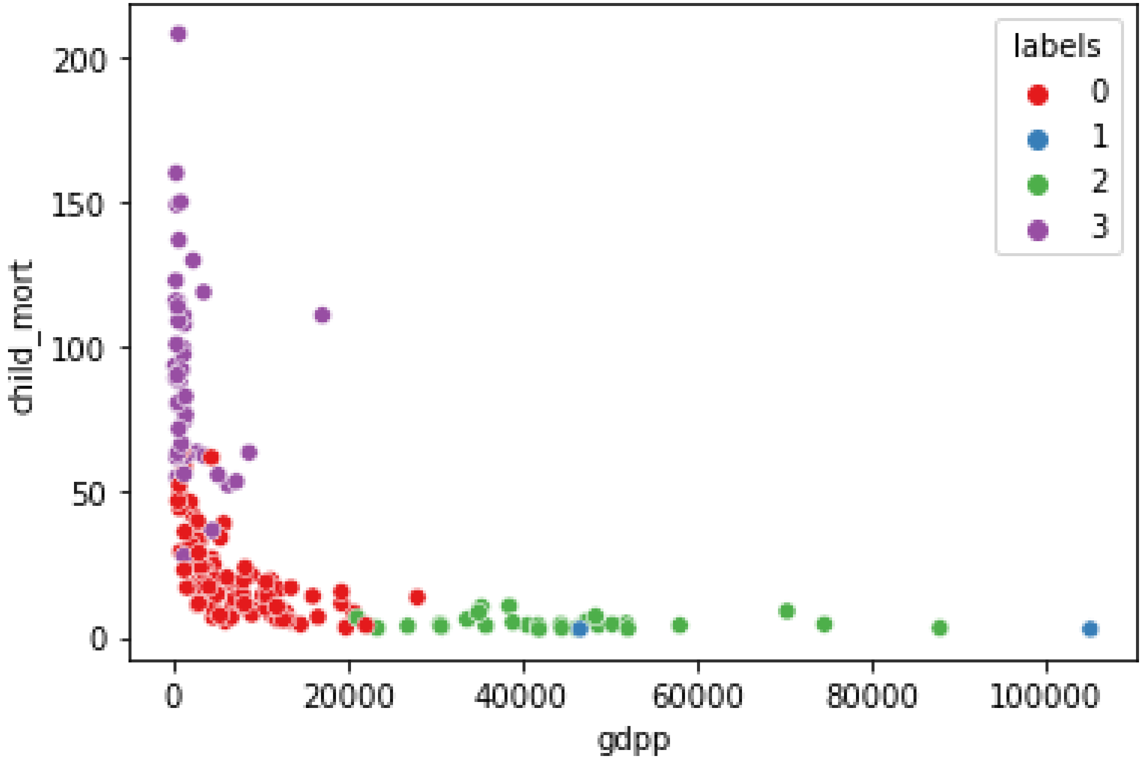
- K = 3

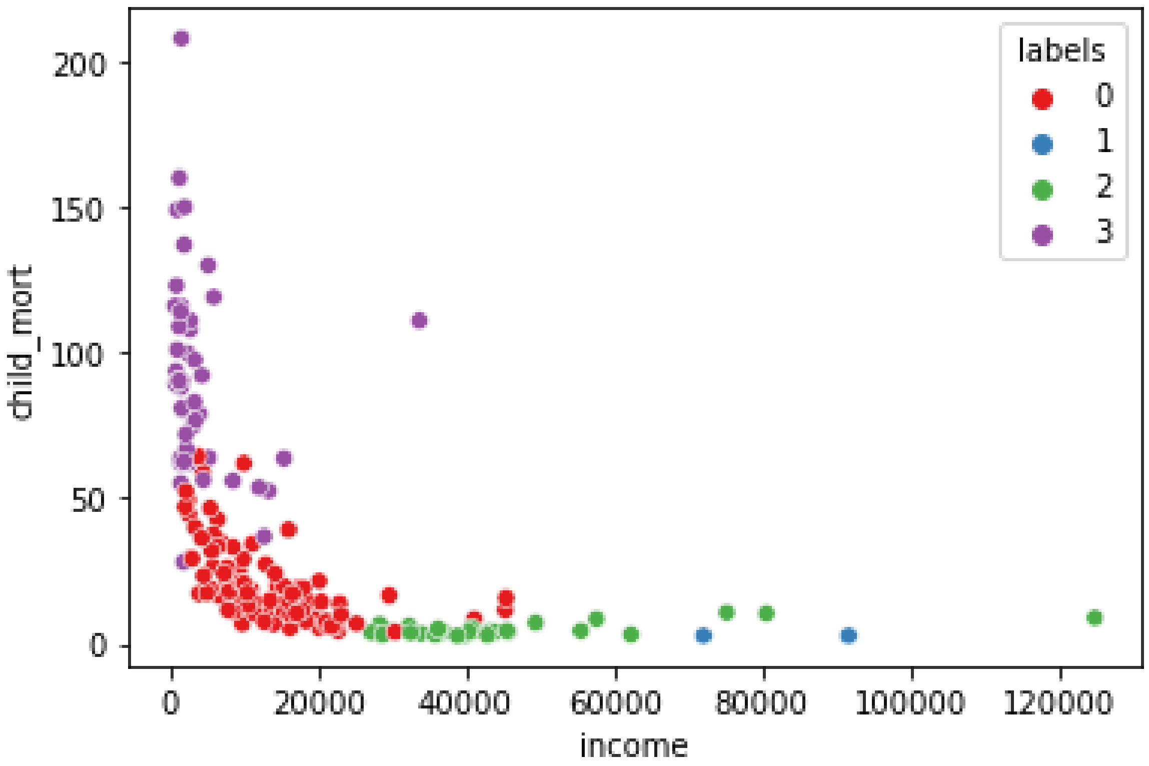
Cluster ID	Number of Countries
0	48
1	28
2	91

Taking k = 4 for further analysis

Visualising gdpp, income, labels in a scatter plot






labels
0
1
2
3
child_mort
200
150
100
50
0
income
0
20000
40000
60000
80000
100000
120000

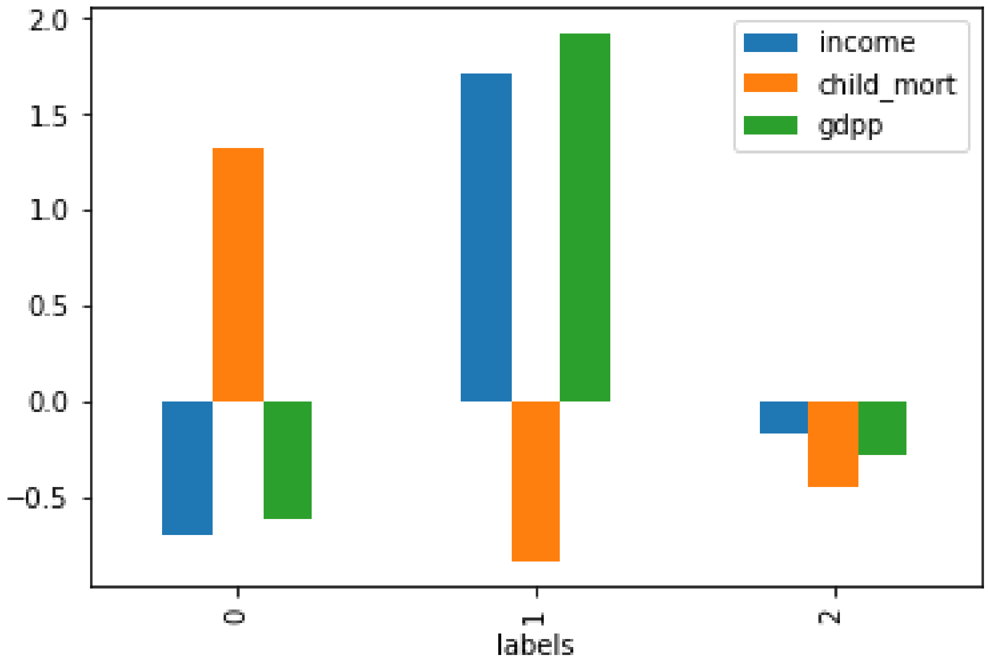
Observations

- Observations:
- Lower income countries have high child mortality
- As income of people in a country is increasing gdpn also increases
- Countries with high gdp have low child mortality.

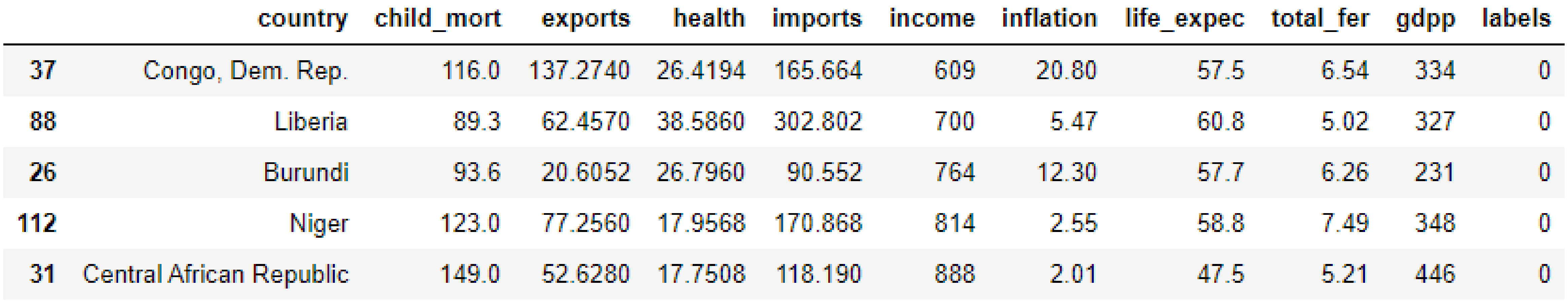
Target Cluster

- Top Attributes for the target cluster selection:
  - Gdpp, income, child_mort
  - The target cluster would have low gdpp, low income and high child_mort
  - In order to choose that cluster performed cluster profiling

We see cluster id 0 has high child_mort , low income and low gdpp.



- We can have a look at the clusters which fall in cluster id 0
- Sorting the countries in increasing order of income, gdpp and decreasing order child_mort we get:



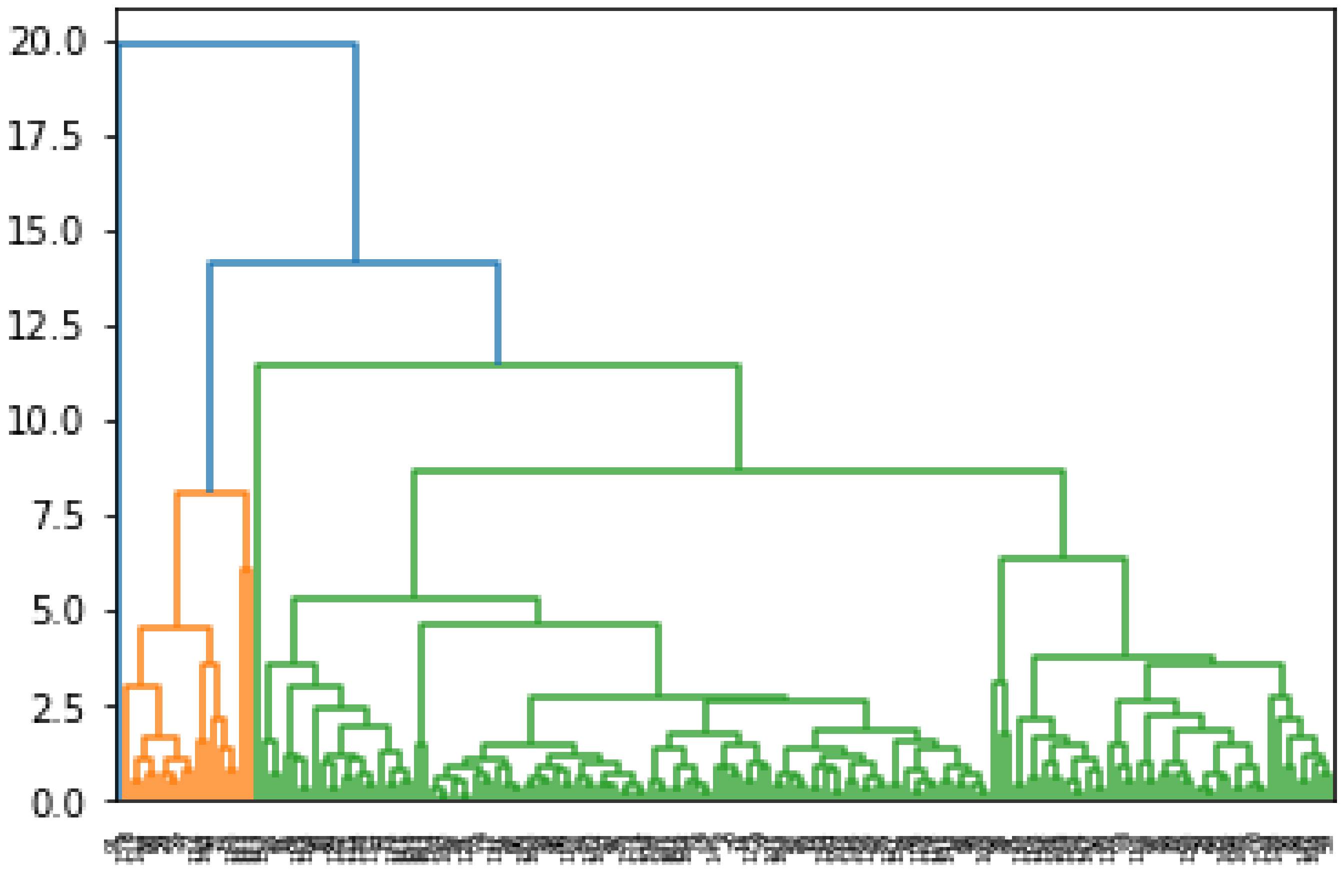
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	0
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	0
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	0

Target Countries

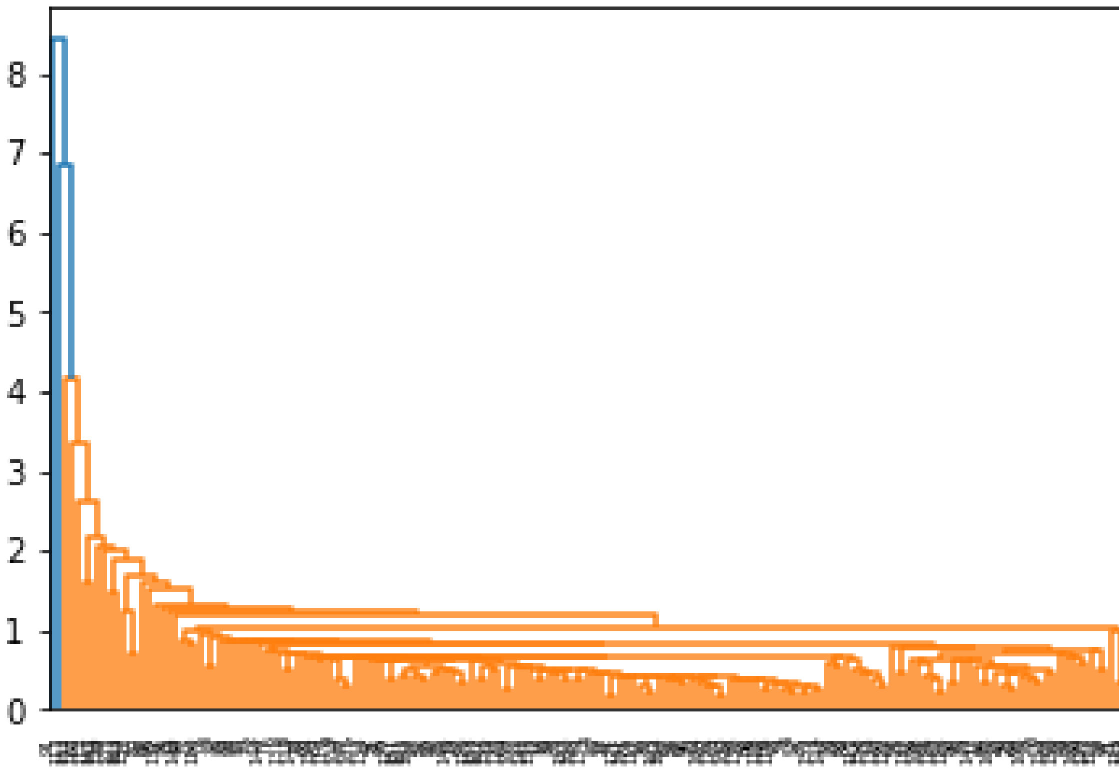
- Congo, Dem. Rep.
- Liberia
- Burundi
- Niger
- Central African Republic

Hierarchical Clustering

- Performed clustering using another method for the problem statement – Hierarchical clustering.
- Following are the dendrogram using complete linkage amd single linkage:



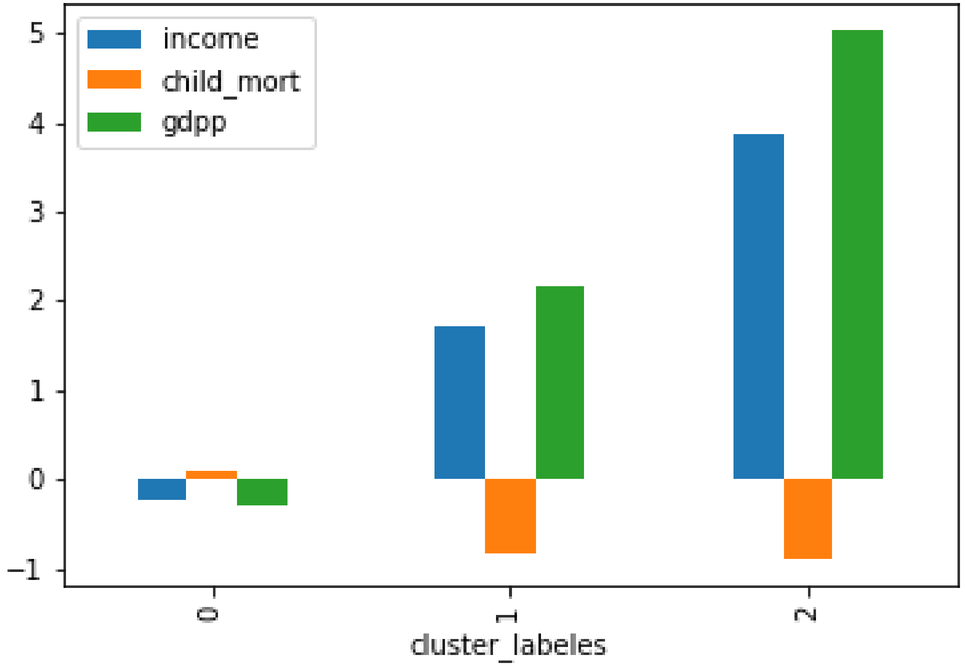
Complete Linkage



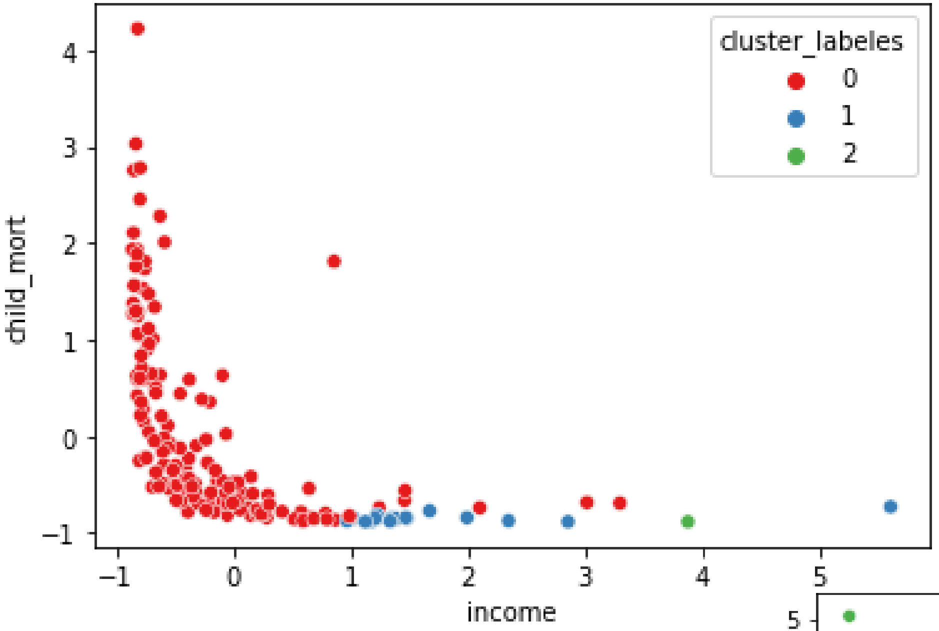
Single Linkage

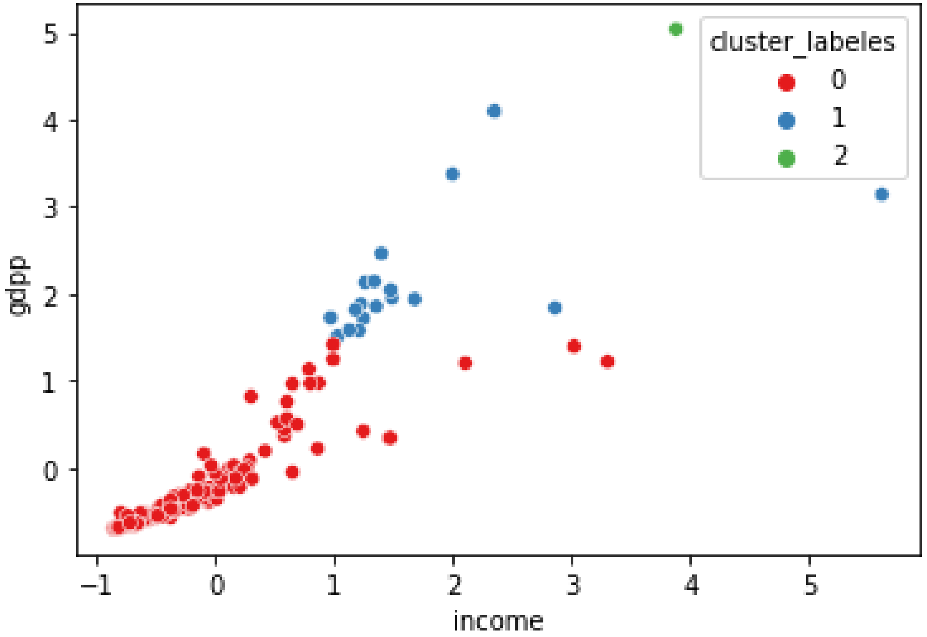
We are going with complete linkage as single linkage is not so clear.

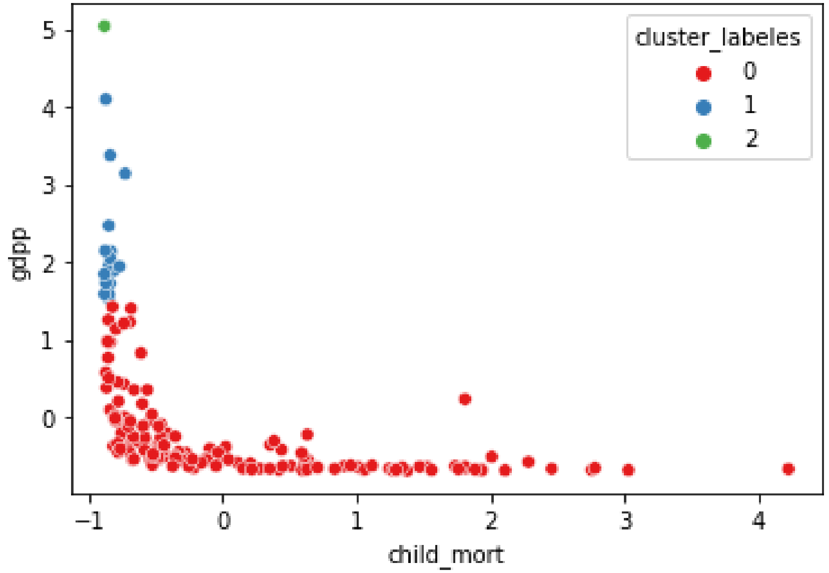
Choosing to cut the dendrogram at k = 3 we get the following distribution of income, child_mort and gdpp.



Visualising complete linkage clusters







Observations

- Lower income countries have high child mortality
- As income of people in a country is increasing gdpn also increases
- Countries with high gdp have low child mortality.

Target Cluster

- Selecting cluster with low income, low gdpp and high child_mort as the target cluster.
- Target Countries:
  - Congo, Dem. Rep.
  - Liberia
  - Burundi
  - Niger
  - Central African Republic

Conclusions

- The top 5 countries that are in dire need of help with factors like gdpp, income, child_mort taken into account are:
    - Congo, Dem. Rep.
    - Liberia
    - Burundi
    - Niger
    - Central African Republic

  HELP International can go forward and invest their funds in the above 5 countries