Clustering Assignment

Identify countries that are in dire need of financial aid

Problem Statement

- Based on the data about socio-economic indicators for countries, we have to identify countries that are most vulnerable, and are in need of financial help in time of disaster and calamities .
- Objectively, we have to cluster similar countries and group, based on features – child mortality, exports, health, imports, income, inflation, life expectancy, total fertility and gdpp

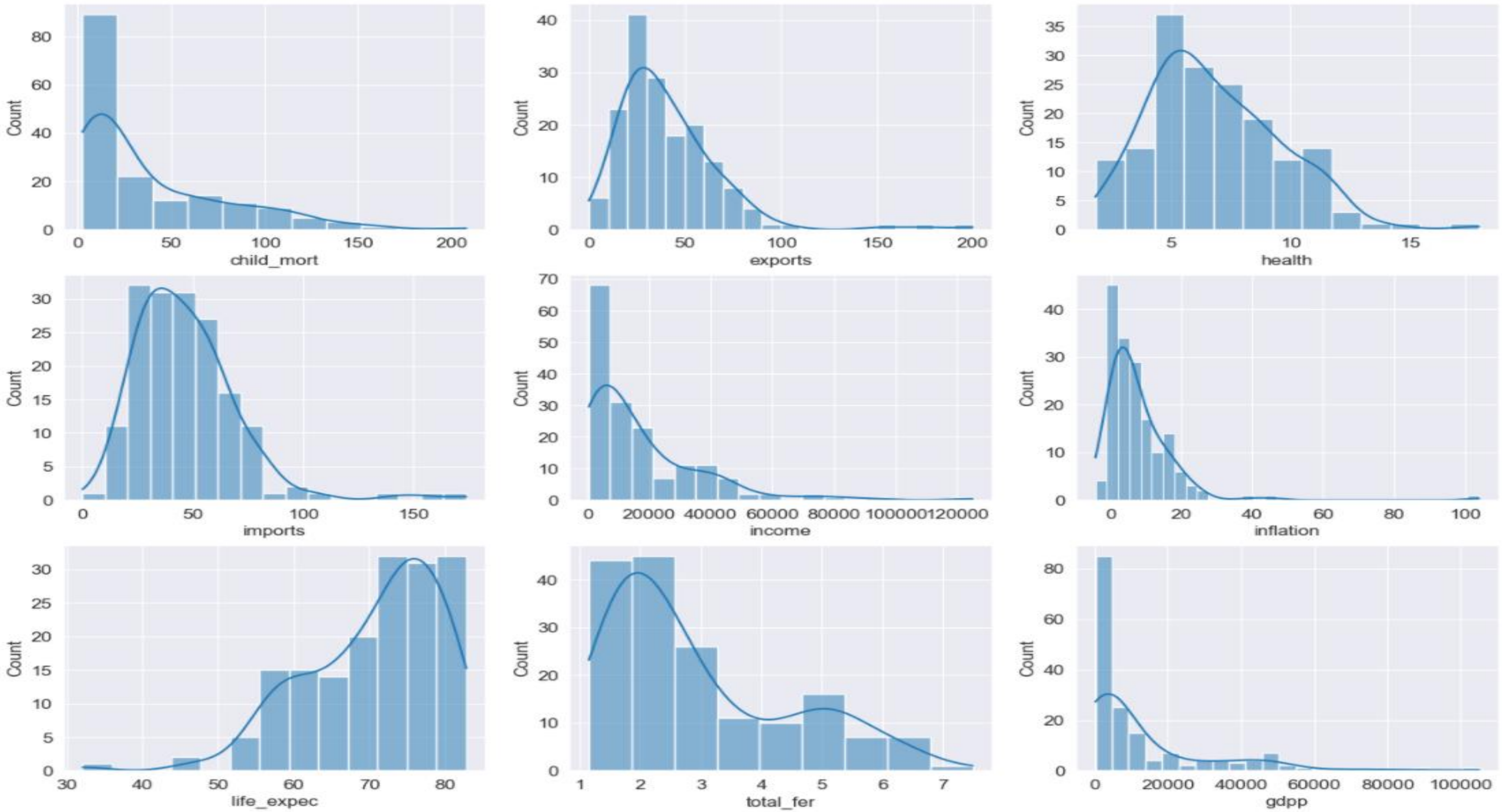
Steps

- Understanding data and EDA
- Data processing
- Model – kmeans
- Model – kmeans (outlier removed)
- Model – hierarchical clustering (single and complete linkage)
- Model results

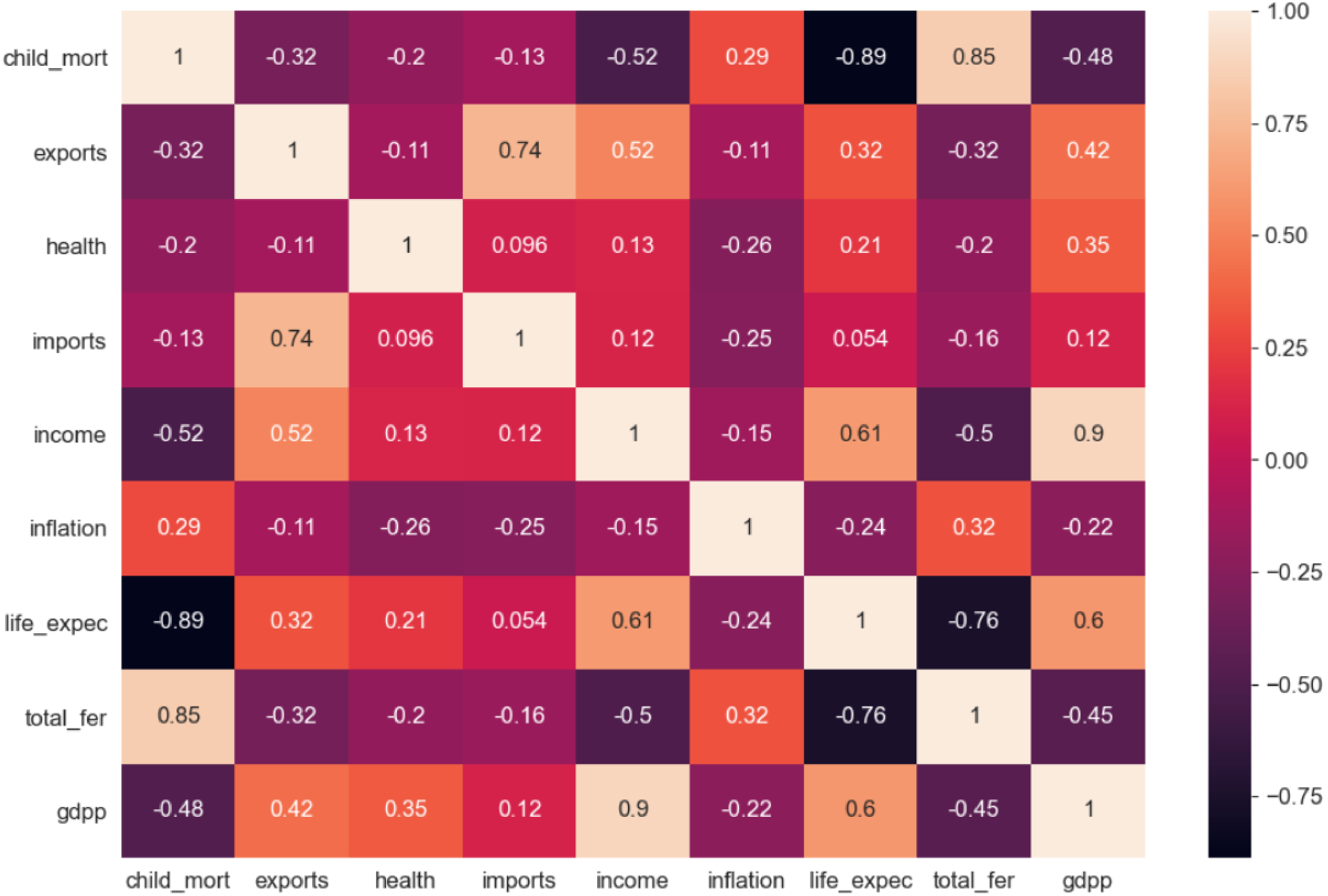
EDA

- We have data about 167 countries.
- Dataset is clean and no has no null values.
- Countries can be grouped in under-developed, developing and developed.
- Features like child mortality, inflation and total fertility are negative indicators (increase in them, is bad)
- All other are positive features, indicating progress of the country.

Data Histograms



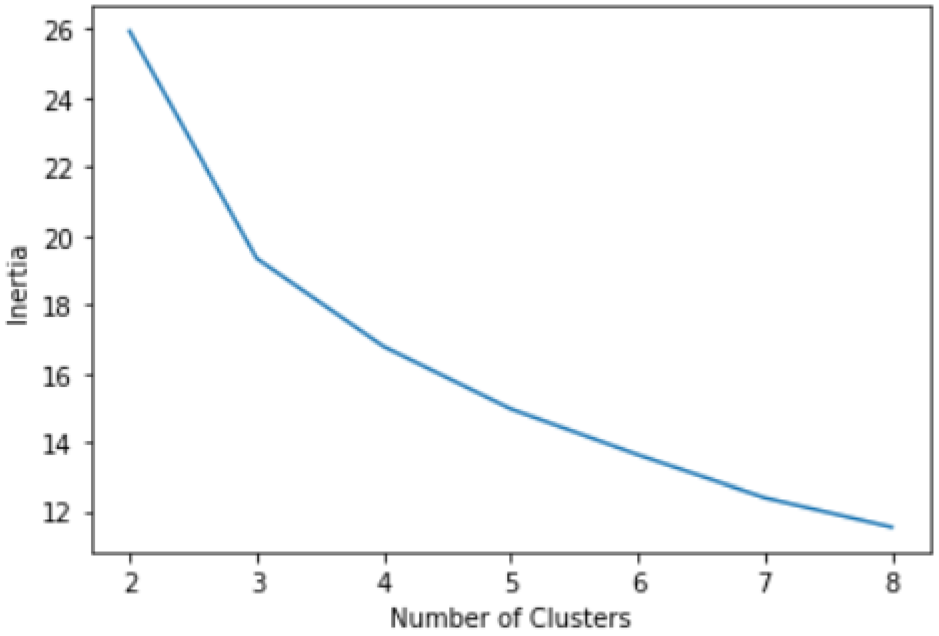
Co-relation plots



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.32	-0.2	-0.13	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.32	1	-0.11	0.74	0.52	-0.11	0.32	-0.32	0.42
health	-0.2	-0.11	1	0.096	0.13	-0.26	0.21	-0.2	0.35
imports	-0.13	0.74	0.096	1	0.12	-0.25	0.054	-0.16	0.12
income	-0.52	0.52	0.13	0.12	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.11	-0.26	-0.25	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.32	0.21	0.054	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.32	-0.2	-0.16	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.42	0.35	0.12	0.9	-0.22	0.6	-0.45	1

Kmeans (k – number of clusters)

- Scaled the data.
- Elbow, method used to identify optimal k
- Optimal k found = 3
- Inertia : Sum of square distances from cluster centre (SSD)



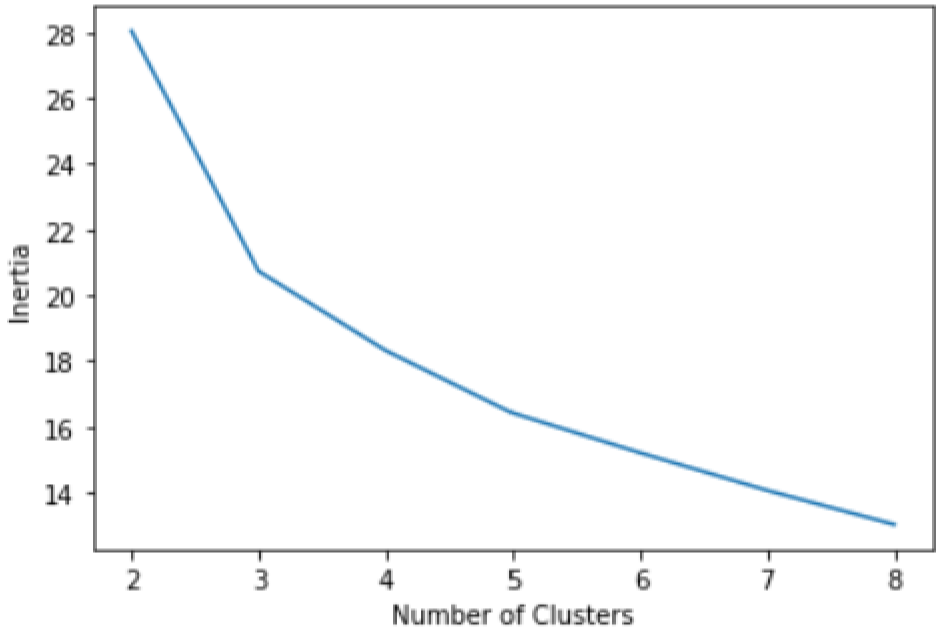
Finding optimal k by elbow method

Outliers detection

- Outliers based on positive features.
- We have only 167 data points.
- Given our objective, we cannot discard outliers based on negative features.
- Outliers cleaned based on GDPp feature.
- Only countries with GDPp less than - **Q3 + 1.5 x (Q3 – Q1)**

kmeans – no_outliers

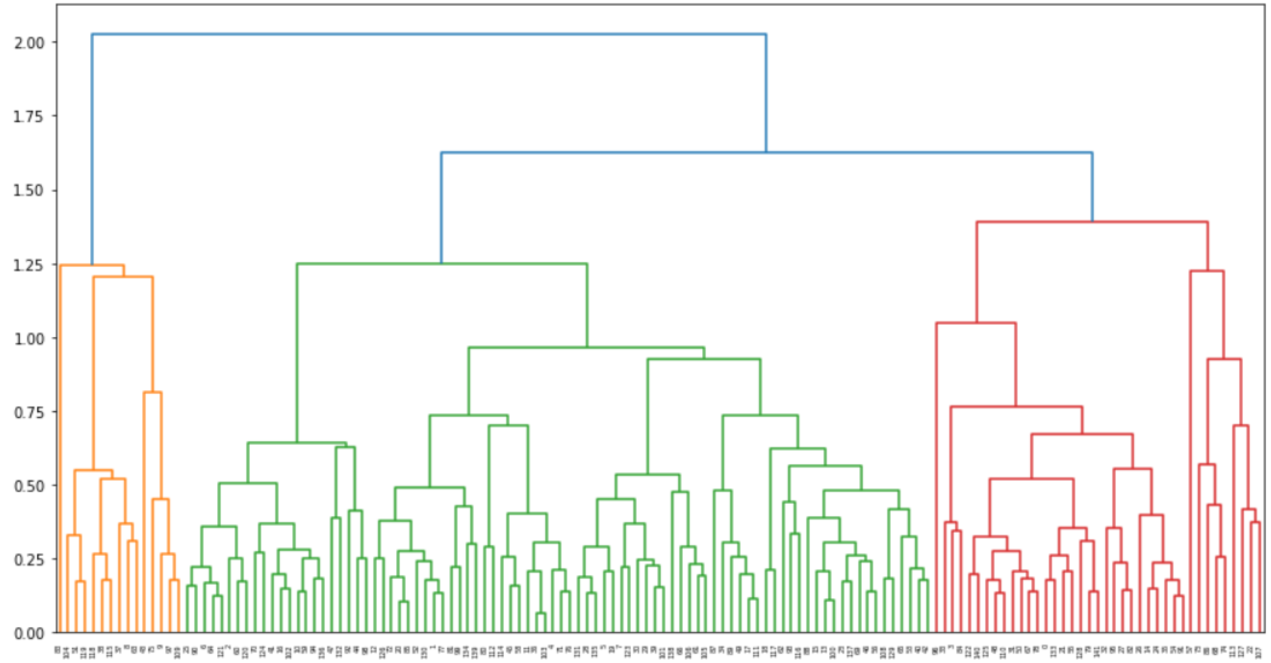
- Scale the data with MinMaxScaler()
- Elbow, method used to identify optimal k
- Optimal k found = 3
- Inertia : Sum of square distances from cluster centre (SSD)



Finding optimal k by elbow method

Hierarchical Clustering

- Solution with complete linkage (single linkage solution too skewed)

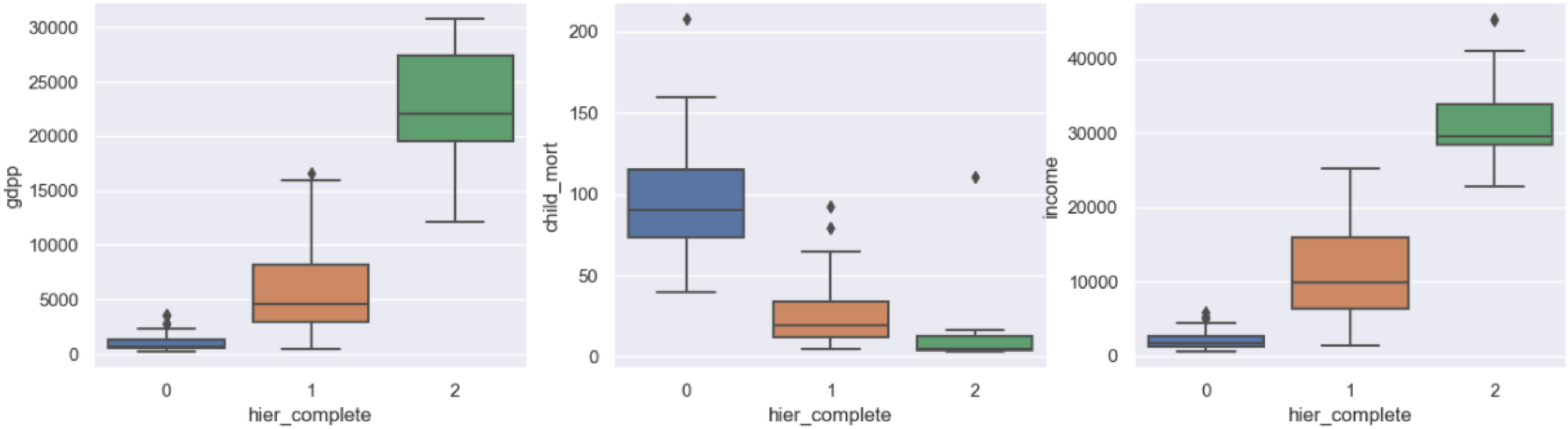


Result and Visualizations-

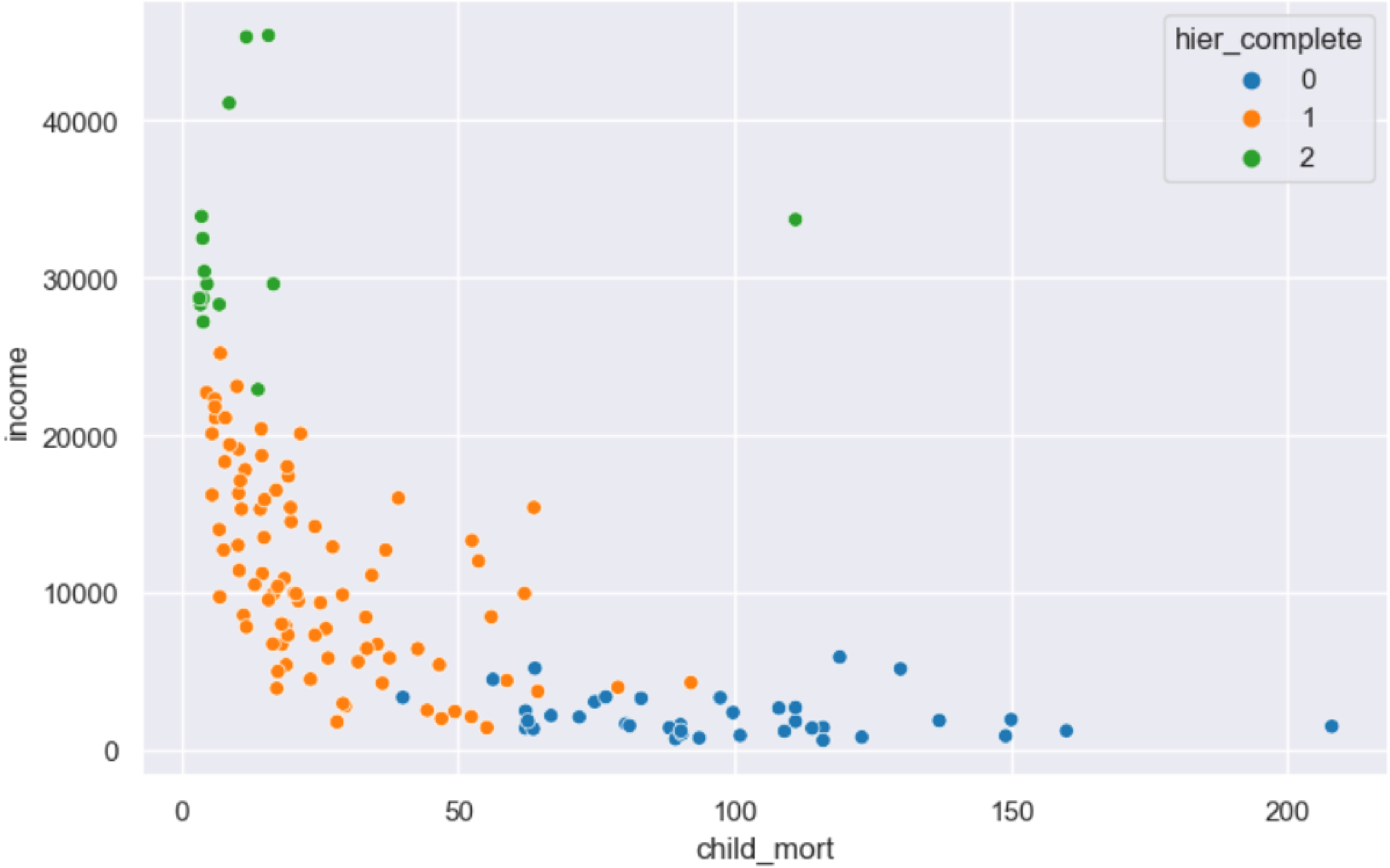
39 countries were common out of all models, that were clustered as under-developed.

Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d’Ivoire, Gambia, Ghana, Guinea, Guinea-Bissau, Haiti, Kenya, Kiribati, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Micronesia, Fed. Sts., Mozambique, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen, Zambia

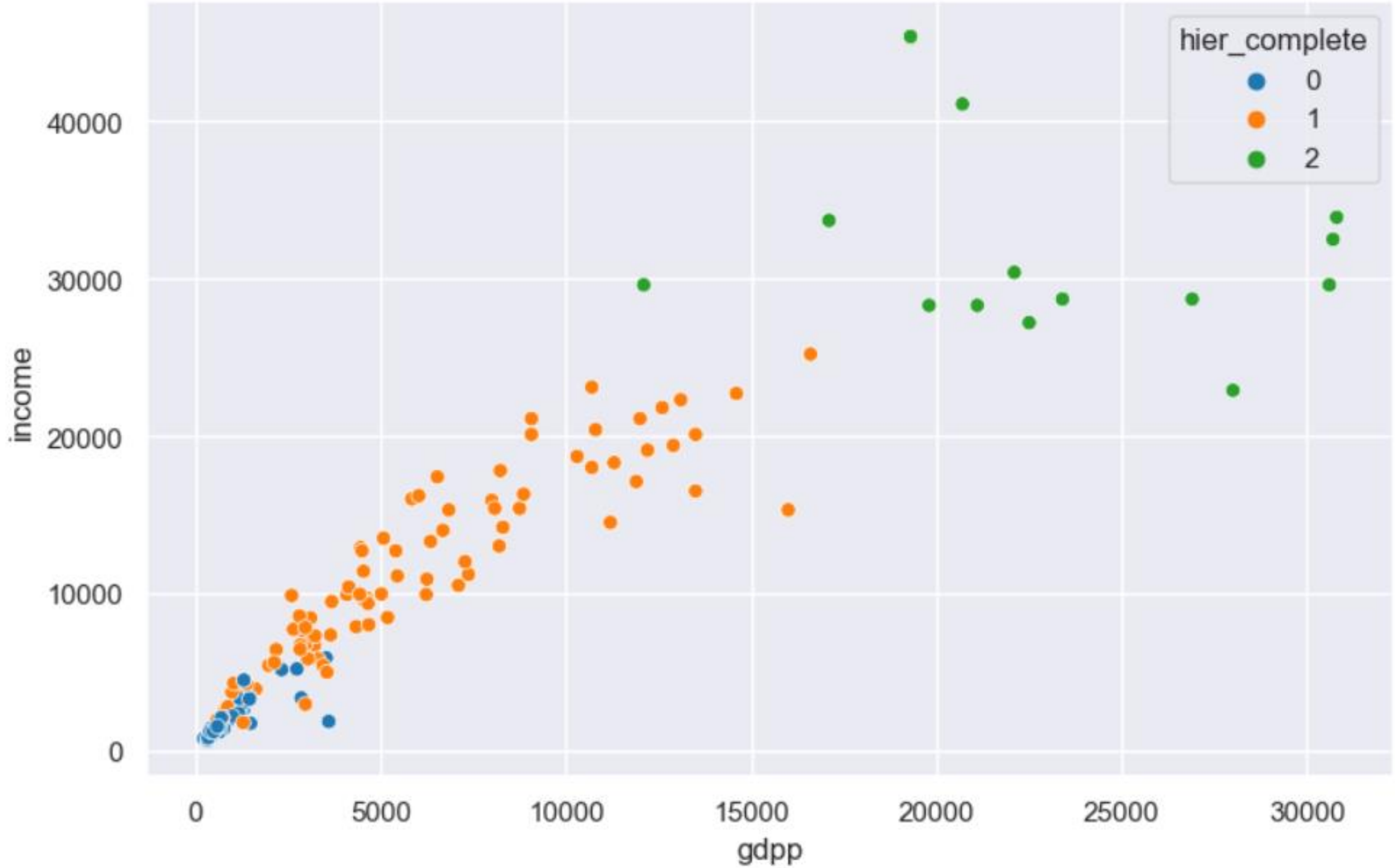
Cluster 0 - under-developed countries



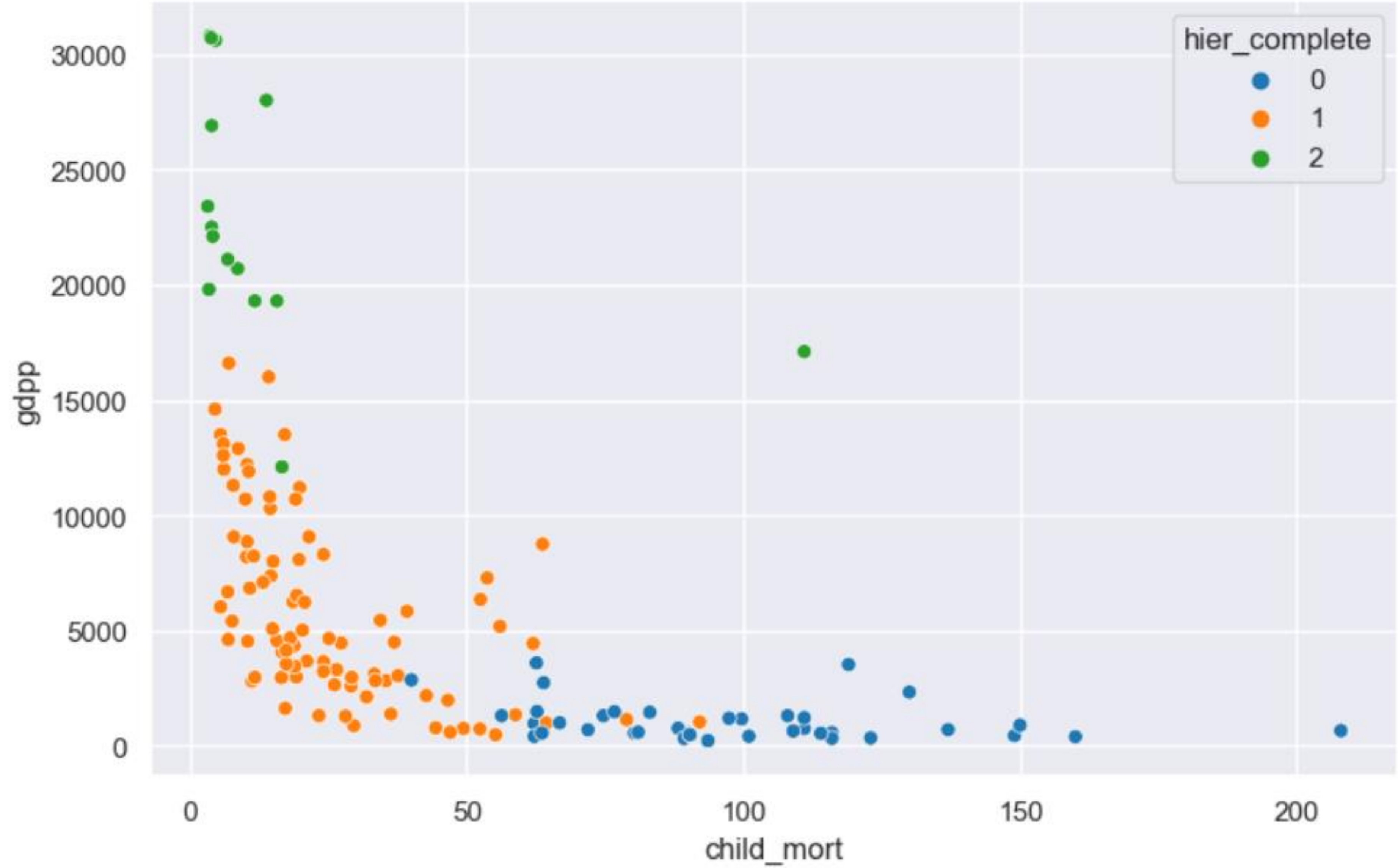
Income vs child_mort (cluster 0 – underdeveloped)



Income vs GDPp (cluster 0 – underdeveloped)



GDPp vs child mort (cluster 0 – underdeveloped)


hier_complete
● 0
● 1
● 2
gdpp
child_mort