

CLUSTERING OF COUNTRIES


BY:

DUVVURI SURYATHEJA REDDY

Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

PROBLEM STATEMENT

HELP International is an international humanitarian NGO that is committed to provide financial aid to backward countries. NGO raised funds and now it has to decide on which country to aid. Given data about different countries and their social, economic, health factors, as an analyst, we need to provide countries which are categorised as backward/under developed countries so that NGO can provide financial aid to those countries.

Several white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

SOLUTION METHODOLOGY

DATA CLEANING, TRANSFORMATION

- Imported the data
- Checked for null values, duplicates
- Transformed columns with values in percentages format to absolute values to capture more information



OUTLIER ANALYSIS

- Identified Outliers
- Decided to keep them as removing them might affect ranking of countries for financial aid



EDA, SCALING, HOPKINS STATISTIC

- Univariate and Bivariate analysis is done to check for patterns and correlations
- Scaling is done on data
- Hopkins statistic is computed to check if data has tendency to form clusters



FINAL LIST MAKING

- Further filtration of under developed countries is done to see which countries among the under developed countries require aid the most.
- This filtration is done on GDP first, then on income and then finally on child mortality
- After filtration is done we arrive at the final list of countries



K-MEANS CLUSTERING

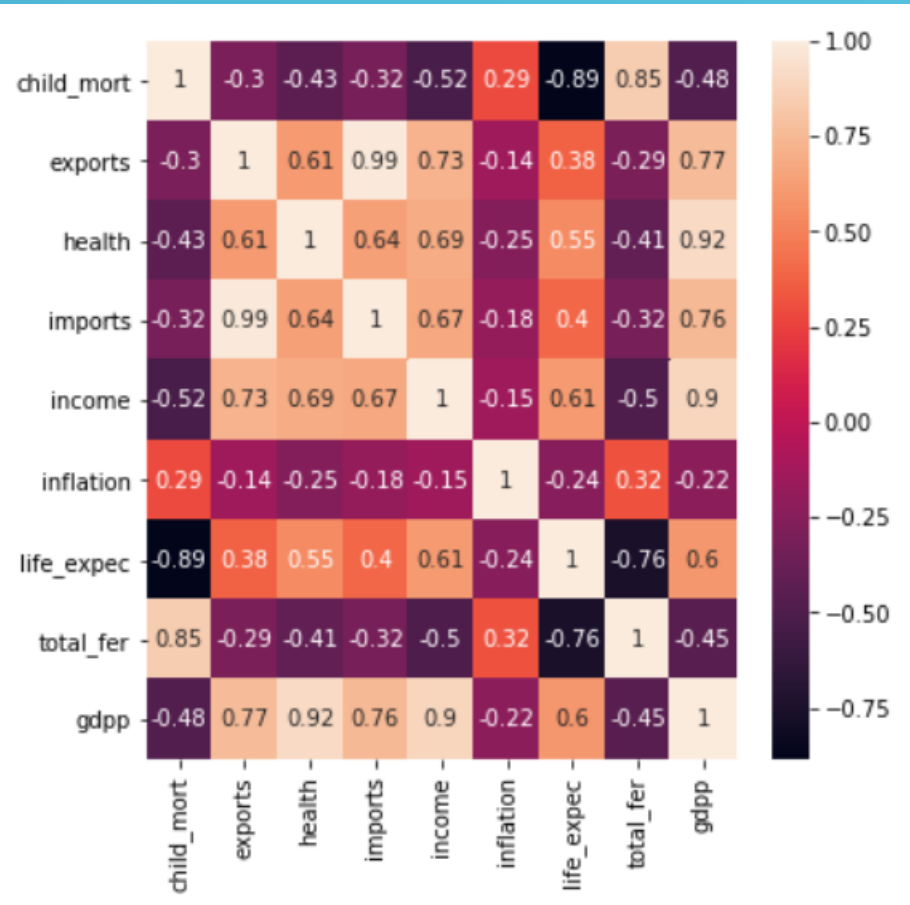
- Based on silhouette analysis, elbow curve method, balanced size of cluster, business objective obtained optimal value of k.
- Visualizing the clusters with various variables and analysing clusters
- Identifying the countries which require aid



HIERARCHICAL CLUSTERING

- Identifying optimal no of clusters using dendrogram
- Visualizing the clusters with various variables and analyzing the clusters.
- After clustering is done, Cluster which is categorized as under developed has 90 percent of data in it. So decided not to go with hierarchical clustering





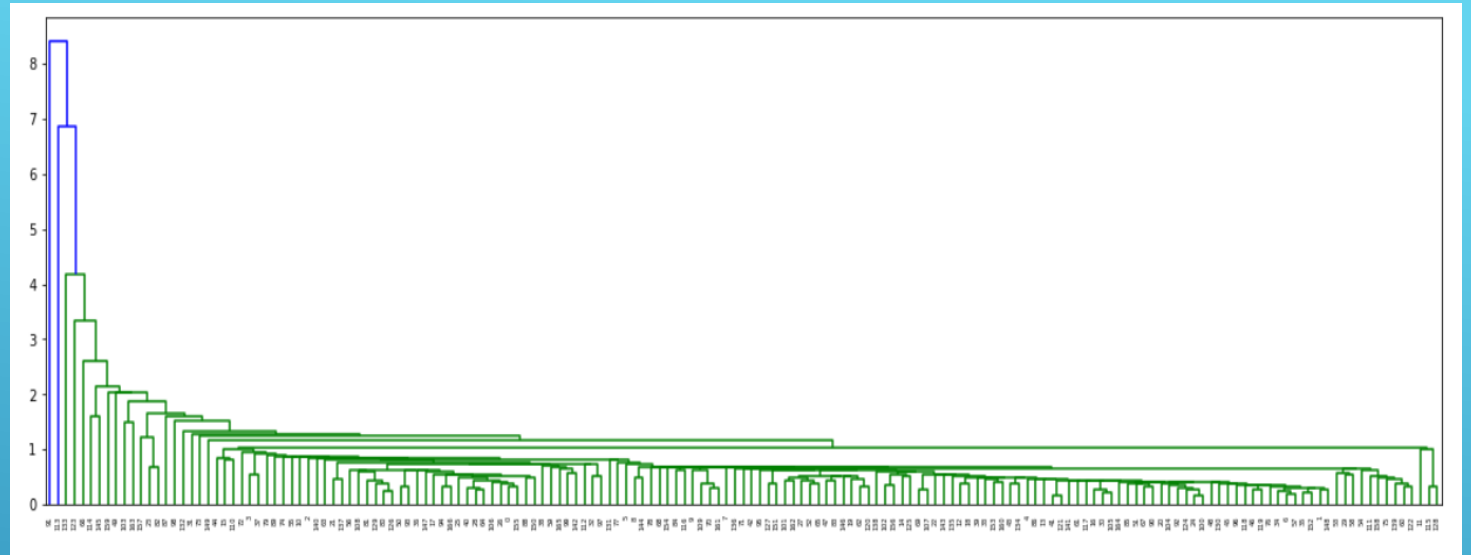
INTRODUCTION AND HEAT MAP

- Data cleaning is done and analyzed null values, duplicates, outliers.
- Scaled data for easy analysis.
- Looking at the correlation matrix we can see that (child mortality, life expectancy), (exports, imports), (health,gdpp), (income,gdpp) have high correlation .

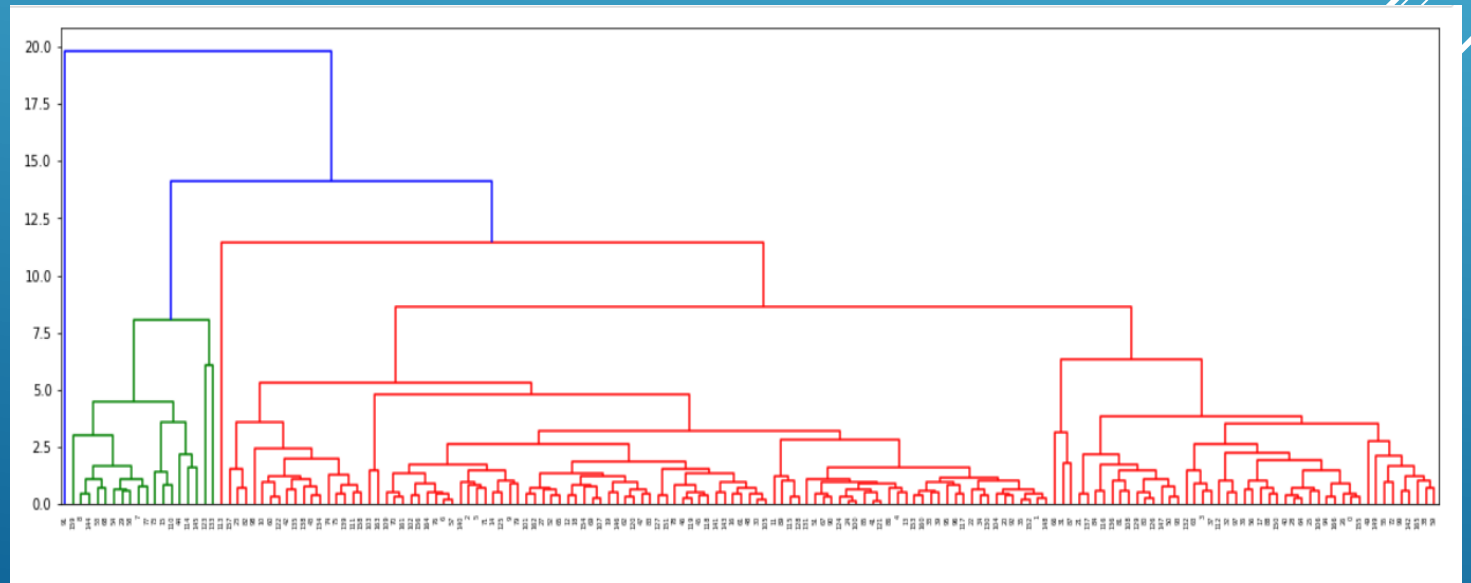
SINGLE LINKAGE

HIERARCHICAL CLUSTERING:

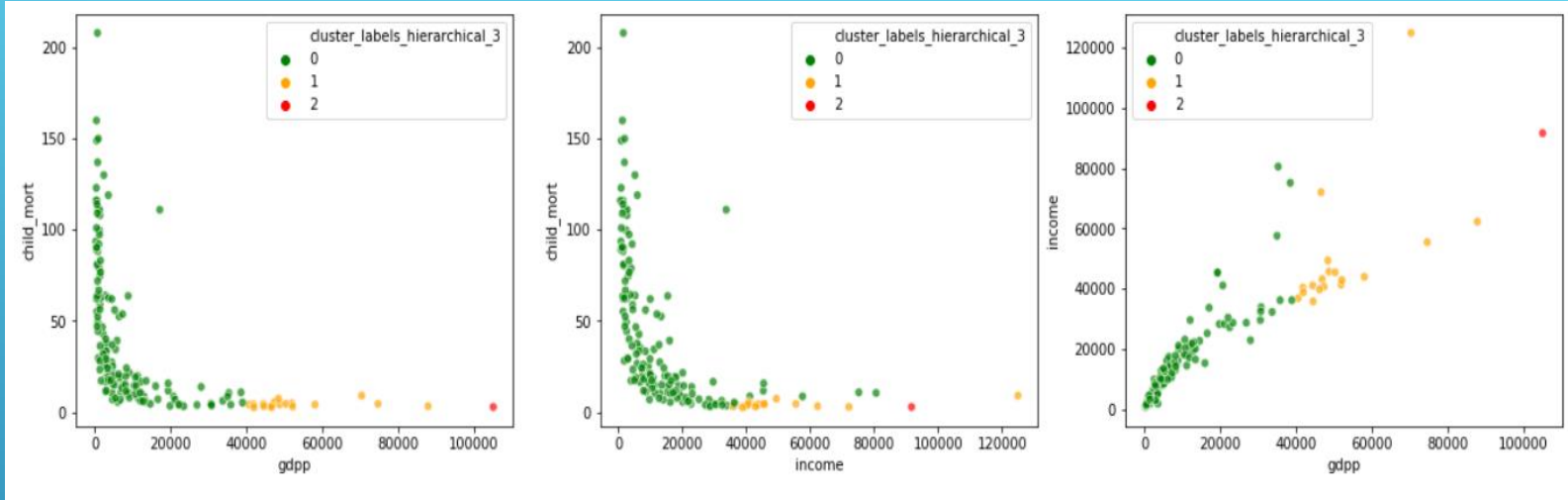
- Going with complete linkage as single linkage is not clear
- Based on dendrogram of complete linkage going with 3 and 4 clusters



COMPLETE LINKAGE



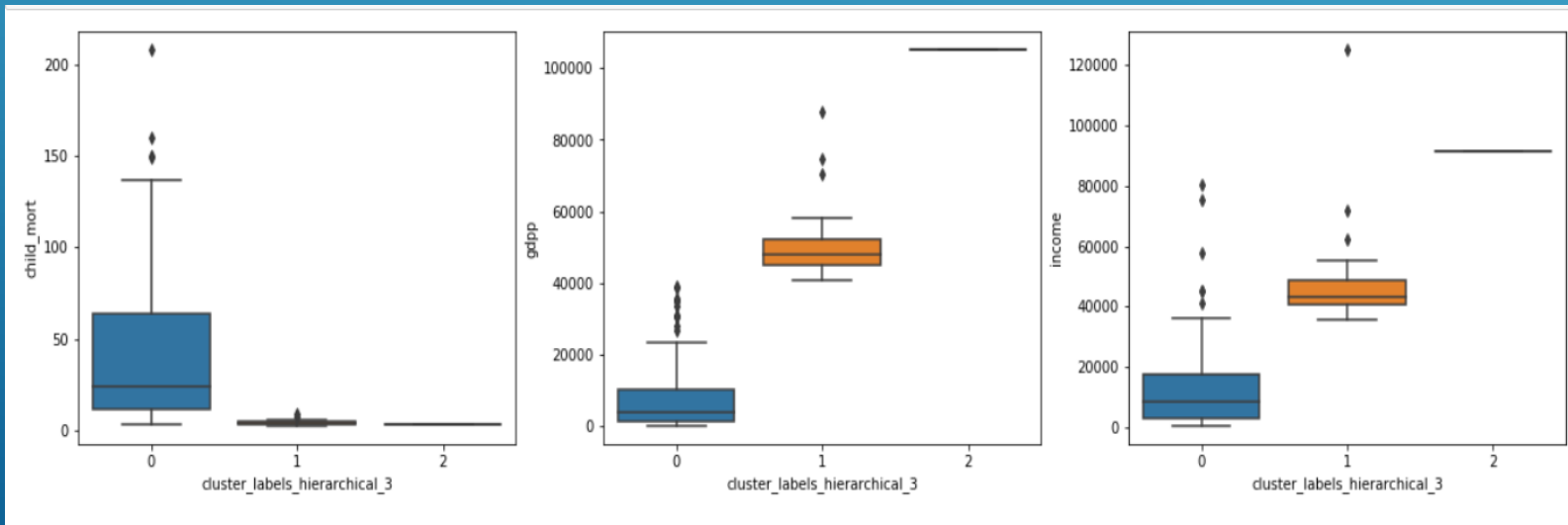
Analyzing clusters using 2 variables



3 CLUSTERS MODEL

Based on both plots we can see cluster 0 has high mortality rate, less income, less GDP. Therefore cluster 0 can be categorized as under developed

Analyzing clusters using 1 variable

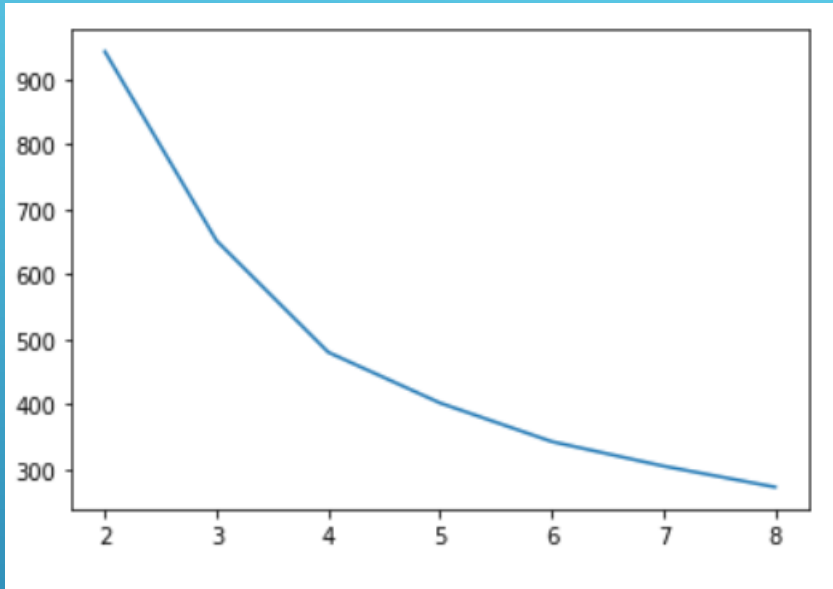


10 countries in under developed countries categorized by hierarchical clustering

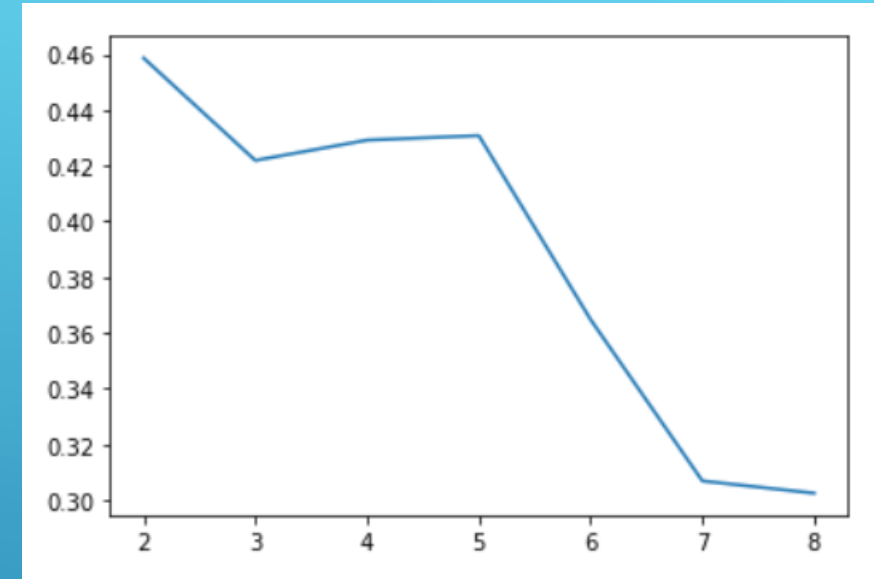
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels_hierarchical_3
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.440	56.2	5.82	553	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.490	76.3	1.65	4090	0
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.100	76.5	2.89	4460	0
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.400	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.440	76.8	2.13	12200	0
5	Argentina	14.5	1946.70	834.3000	1648.000	18700	20.900	75.8	2.37	10300	0
6	Armenia	18.1	669.76	141.6800	1458.660	6700	7.770	73.3	1.69	3220	0
9	Azerbaijan	39.2	3171.12	343.3920	1208.880	16000	13.800	69.1	1.92	5840	0
10	Bahamas	13.8	9800.00	2209.2000	12236.000	22900	-0.393	73.8	1.86	28000	0
11	Bahrain	8.6	14386.50	1028.7900	10536.300	41100	7.440	76.0	2.16	20700	0

Hierarchical clustering is not giving good solution as 90 percent of data is going into one single cluster . So going for K-means Clustering for better clustering

ELBOW CURVE METHOD



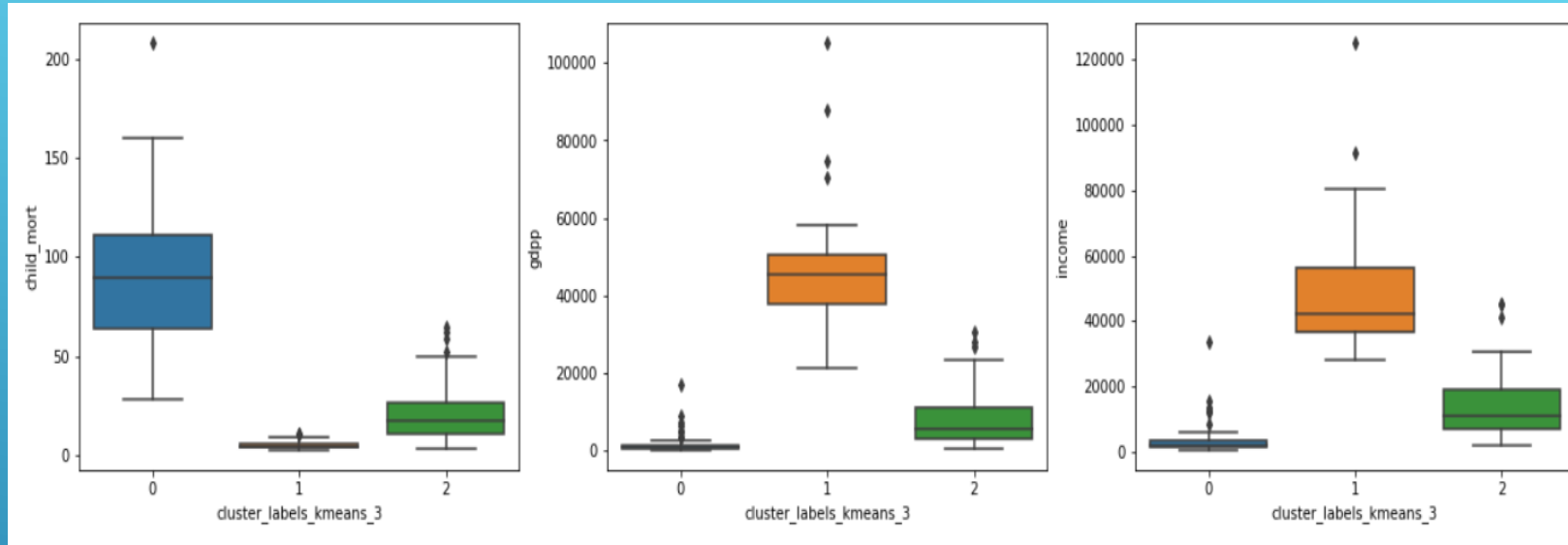
SILHOUETTE ANALYSIS



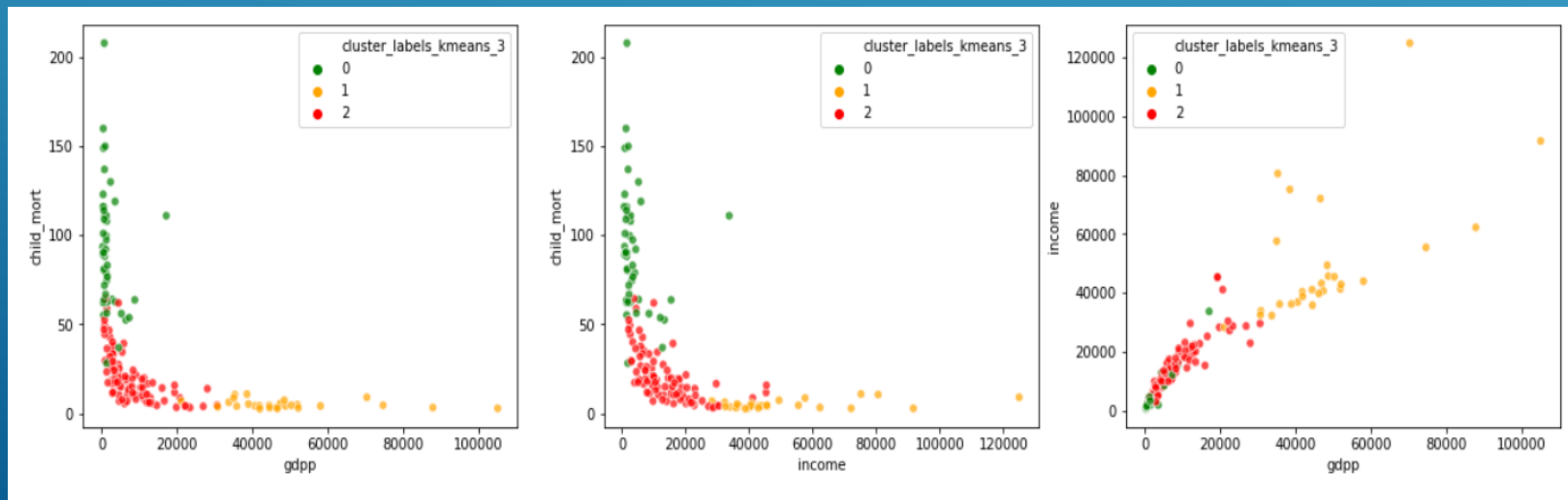
Based on both plots 2,3,4,5 seem to optimal k ,therefore created 4 models respectively.

3 seem to good choice of k as bigger drop is seen at 3 in elbow curve method, also has balanced cluster size. Also in terms of business aspect too 3 is good choice as it is interpretable as under developed, developing and developed countries.

Analyzing clusters using 2 variables



Analyzing clusters using 1 variables



FINAL MODEL

- Cluster 0 has high child mortality ,less income, less GDPP. Therefore it consists of under developed countries
- Cluster 1 has less child mortality ,high income, high GDPP. Therefore it consists of developed countries
- Cluster 2 is in mid of both. Therefore it consists of developing countries

TOP 10 COUNTRIES – K MEANS

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels_kmeans_3	development_status
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	0	Under_developed
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0	Under_developed
112	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	0	Under_developed
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	0	Under_developed
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419	0	Under_developed
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	0	Under_developed
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459	0	Under_developed
150	Togo	90.3	196.1760	37.3320	279.624	1210	1.18	58.7	4.87	488	0	Under_developed
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.97	55.6	5.05	547	0	Under_developed
0	Afghanistan	90.2	55.3000	41.9174	248.297	1610	9.44	56.2	5.82	553	0	Under_developed

RECOMMENDATIONS :

The final list of countries are Burundi, Niger, Sierra Leone, Mozambique, Central African Republic, Malawi, Togo, Guinea-Bissau, Afghanistan, Congo Dem Rep.

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231
Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334
Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348
Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399
Mozambique	101.0	131.9850	21.8299	193.578	918	7.64	54.5	5.56	419
Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446
Malawi	90.5	104.6520	30.2481	160.191	1030	12.10	53.1	5.31	459
Togo	90.3	196.1760	37.3320	279.624	1210	1.18	58.7	4.87	488
Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.97	55.6	5.05	547
Afghanistan	90.2	55.3000	41.9174	248.297	1610	9.44	56.2	5.82	553