Course Project

Moon, Hyosik

August 24, 2021

1 Required

- **Main objective of the analysis**
  HELP International 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. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

  I will categorise the countries into some groups and suggest the countries in the poorest group. Let's find which countries belong to the poorest group.

- **Brief description of the data set**
  There are 167 rows which are countries and 10 columns explaining the countries' information such as Columns0: country — Description: Name of the country Columns1: child_mort — Description: Death of children under 5 years of age per 1000 live births Columns2: exports — Description: Exports of goods and services per capita. Given as %age of the GDP per capita Columns3: health — Description: Total health spending per capita. Given as %age of GDP per capita Columns4: imports — Description: Imports of goods and services per capita. Given as %age of the GDP per capita Columns5: Income — Description: Net income per person Columns6: Inflation — Description: The measurement of the annual growth rate of the Total GDP Columns7: life_expec — Description: The average number of years a new born child would live if the current mortality patterns are to remain the same Columns8: total_fer — Description: The number of children that would be born to each woman if the current age-fertility rates remain the same. Columns9: gdpp

— Description: The GDP per capita. Calculated as the Total GDP divided by the total population. (Figure 1)

	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
...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	46.6	5.25	52.7	2950	2.62	63.0	3.50	2970
163	Venezuela	17.1	28.5	4.91	17.6	16500	45.90	75.4	2.47	13500
164	Vietnam	23.3	72.0	6.84	80.2	4490	12.10	73.1	1.95	1310
165	Yemen	56.3	30.0	5.18	34.4	4480	23.60	67.5	4.67	1310
166	Zambia	83.1	37.0	5.89	30.9	3280	14.00	52.0	5.40	1460

167 rows × 10 columns

Figure 1: Original data

- **Brief summary of data exploration** In order to cluster the data accurately we need to explore and preprocess. For the accurate distance calculation between data, I preprocessed them according to the following steps: Data types - Remove - Skew - Scaling - Pairplot.

    1. 1. Data types I changed income and gdpp data types from int to float in order to calculate closeness between cluster later.
    2. 2. Remove We will cluster countries, so we don't need the country column and I also deleted NaN values in data.
    3. 3. Skew Check the skewness of the columns and take log1p if the value of skewness is higher than 0.75.
    4. 4. Scaling I used MinMaxScaler to normalize them as an equal size. Picture 2 is the processed data.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.795814	0.441251	0.358608	0.737634	0.182457	0.743428	0.475345	0.840523	0.142343
1	0.390743	0.627680	0.294593	0.752832	0.524170	0.671990	0.871795	0.152242	0.469307
2	0.507689	0.686618	0.146675	0.669351	0.573327	0.798274	0.875740	0.431726	0.483466
3	0.863386	0.777797	0.064636	0.728900	0.426373	0.833137	0.552268	0.875946	0.445234
4	0.281644	0.718482	0.262275	0.789822	0.647057	0.581854	0.881657	0.273452	0.648004
...	...	...	...	...	...	...	...	...	...
162	0.523688	0.722978	0.213797	0.768402	0.296178	0.625700	0.609467	0.537788	0.416995
163	0.397641	0.630967	0.192666	0.560549	0.619567	0.910419	0.854043	0.348536	0.664561
164	0.470168	0.805216	0.312617	0.849466	0.375073	0.768656	0.808679	0.230328	0.283207
165	0.681382	0.640505	0.209447	0.686711	0.374654	0.838696	0.698225	0.706063	0.283207
166	0.775859	0.679660	0.253574	0.666302	0.316093	0.783710	0.392505	0.794243	0.300923

167 rows × 9 columns

Figure 2: Processed data

5. 5. Pairplot Picture 3 is the processed data' pairplot.

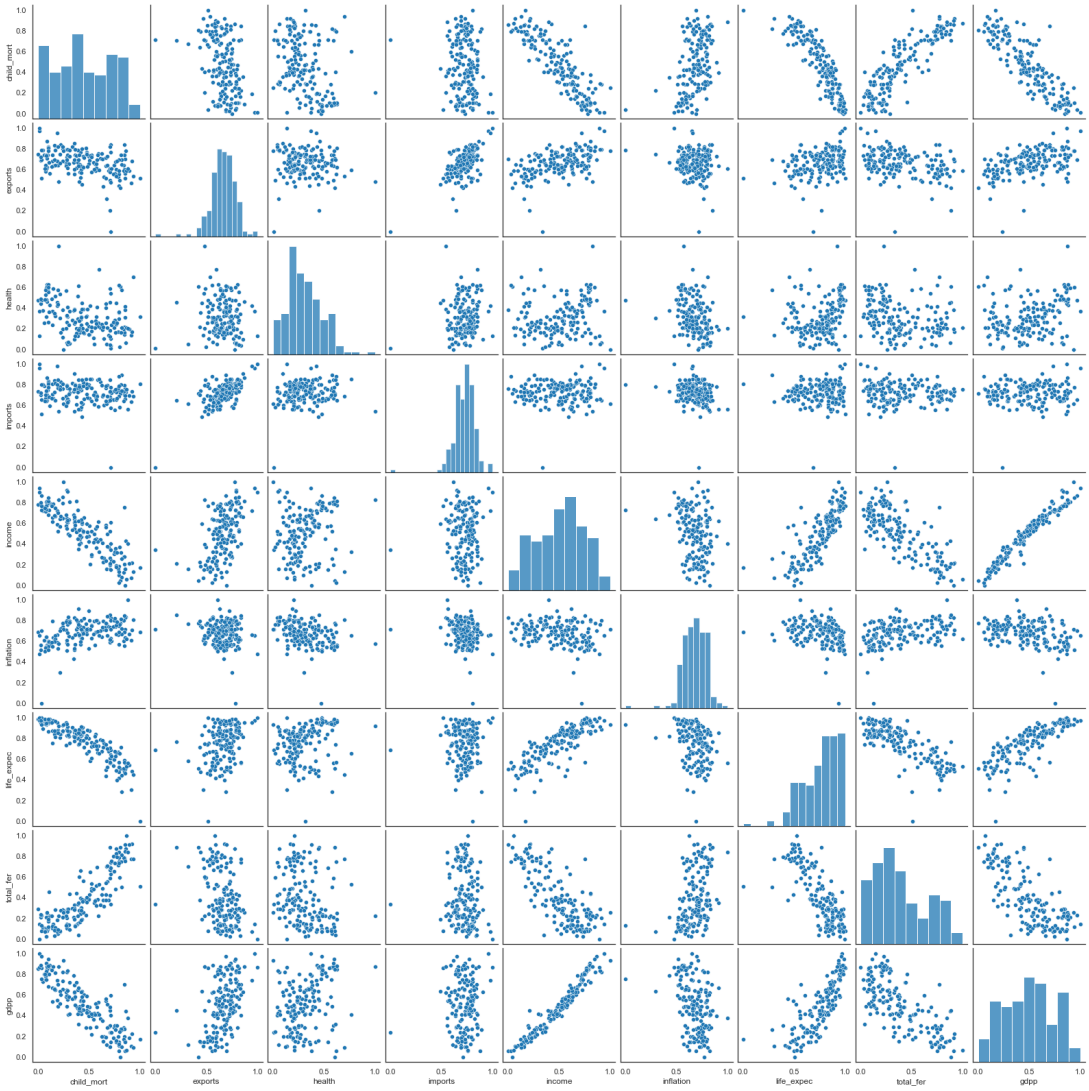

Figure 3: Pairplot of the data

- **Summary of training at least three models** I implemented three different clustering algorithms which are 'K-Means, Hierarchical Agglomerative Clustering, and Non-negative '.

    1. K-Means. I tested the kmeans algorithm with 10 different K values to find the optimal one. I found the optimal K=4, by using the elbow method (Figure 4).

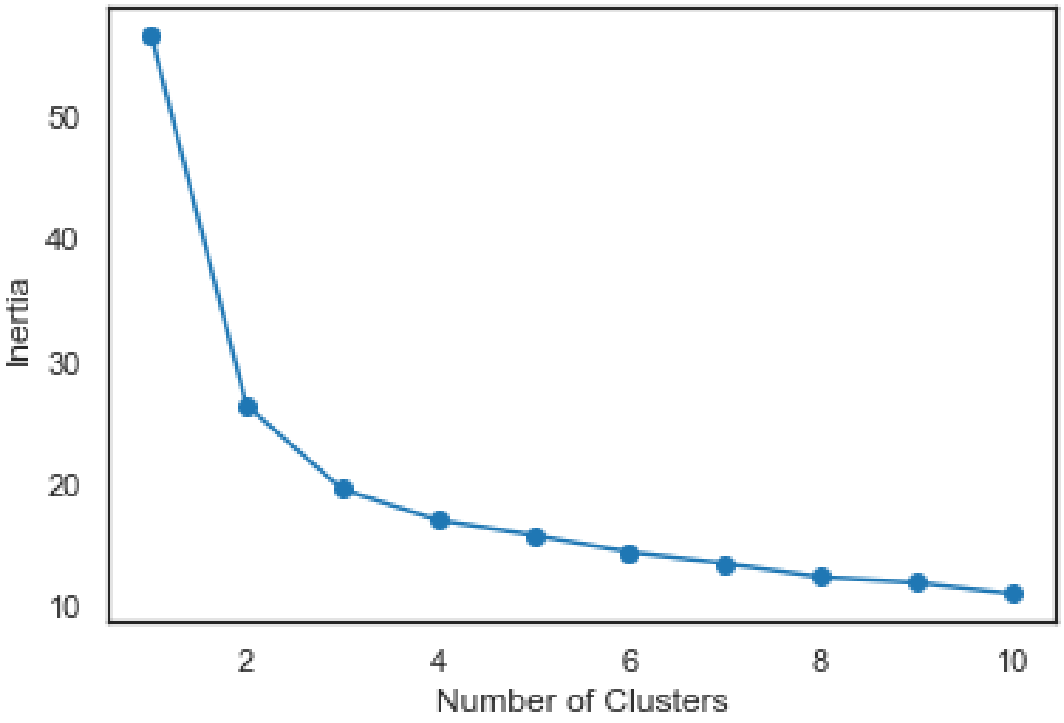

Figure 4: K-Means' inertia plot

Figure 5 is the clustering results, so we can know that 1-cluster is the poorest group. So, we need to support the countries in 1-cluster first.

cluster	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.129433	0.732979	0.416889	0.737676	0.776359	0.578930	0.928550	0.165223	0.810373
1	0.800601	0.558411	0.282951	0.694050	0.204551	0.721227	0.537307	0.752040	0.197456
2	0.354626	0.672646	0.298709	0.712367	0.582728	0.697757	0.840806	0.231976	0.551980
3	0.579522	0.683200	0.238490	0.742472	0.436810	0.711540	0.702649	0.468389	0.404059

Figure 5: K-means Clustering result

The countries in 1-cluster are that:
cluster country
1 Burundi
1 Liberia

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

2. AgglomerativeClustering. I implemented AgglomerativeClustering algorithm with 4 clusters (Figure 6). We can know that 3-cluster is the poorest group.

agglom	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0.359571	0.680830	0.267383	0.703380	0.608648	0.716925	0.841341	0.244788	0.569483
1	0.627178	0.667506	0.219853	0.719066	0.407978	0.731520	0.673639	0.526629	0.373842
2	0.125854	0.721675	0.453126	0.746053	0.748856	0.552394	0.930811	0.157002	0.794844
3	0.802275	0.549799	0.328154	0.720383	0.168553	0.692531	0.518799	0.761291	0.173326

Figure 6: AgglomerativeClustering

The countries in 3-cluster are that:
cluster country
3 Burundi
3 Liberia
3 Congo, Dem. Rep.
3 Niger
3 Sierra Leone
3 Madagascar
3 Mozambique
3 Central African Republic
3 Malawi
3 Eritrea
3 Togo
3 Guinea-Bissau
3 Afghanistan
3 Gambia
3 Rwanda
3 Burkina Faso
3 Uganda
3 Guinea
3 Haiti
3 Tanzania
3 Mali
3 Benin
3 Comoros

3 Chad
3 Kenya
3 Senegal
3 Pakistan
3 Lesotho
3 Cameroon
3 Kiribati
3 Micronesia, Fed. Sts.
3 Timor-Leste

3. Non-negative Matrix Fatorization. This is dimensionality reduction method. I reduced the dimensionality to 4. Figure 7 is the variance result of the each dimension. I clustered the countries based on the highest variance. Thus, countries were categorized into 3 groups (Figure 8). We can know that 3-cluster is the poorest group.

	group_1	group_2	group_3	group_4
Afghanistan	0.28667	0.49735	0.89040	0.30651
Albania	1.50710	0.37657	0.30011	0.38546
Algeria	1.59316	0.16495	0.51036	0.40605
Angola	0.97912	0.03153	0.89530	0.47337
Antigua and Barbuda	1.77767	0.35644	0.24444	0.40758
...	...	...	...	...
Vanuatu	0.88894	0.27232	0.53916	0.43907
Venezuela	1.94251	0.22343	0.44248	0.31305
Vietnam	1.08311	0.34679	0.41494	0.50442
Yemen	0.94301	0.25223	0.76377	0.39275
Zambia	0.60853	0.27087	0.83463	0.37762

163 rows × 4 columns

Figure 7: Variance of NMF

group	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
group_1	0.371378	0.682329	0.306698	0.717780	0.590539	0.672666	0.820852	0.302391	0.578454
group_2	0.702675	0.529379	0.630205	0.809536	0.172742	0.639535	0.607002	0.639251	0.229242
group_3	0.808240	0.590512	0.282724	0.721842	0.178896	0.703037	0.500286	0.772541	0.172284

Figure 8: Non-negative Matrix Fatorization clustering result

The countries in 3-cluster are that:

cluster	country
group_3	Burundi
group_3	Congo, Dem. Rep.
group_3	Niger
group_3	Sierra Leone
group_3	Madagascar
group_3	Mozambique
group_3	Central African Republic
group_3	Malawi
group_3	Eritrea
group_3	Togo
group_3	Guinea-Bissau
group_3	Afghanistan
group_3	Gambia
group_3	Burkina Faso
group_3	Uganda
group_3	Guinea
group_3	Haiti
group_3	Tanzania
group_3	Mali
group_3	Benin
group_3	Comoros
group_3	Chad
group_3	Kenya
group_3	Senegal
group_3	Lesotho
group_3	Mauritania
group_3	Cote d'Ivoire
group_3	Solomon Islands
group_3	Ghana
group_3	Cameroon
group_3	Zambia

- **Explanation of your final model** Based on the three algorithms, top-10 poorest countries are almost same but in the case of nmf some of poorest countries are not included such as 'Liberia' and it clustered the countries into 3 groups which are more broad. Thus, I think K-means (or HAC) is the final model that I choose.

- **Summary Key Findings and Insights** Although NMF shows a little bit different

results compared to K-means and HAC, the result is almost same. If I have to save resources I think NMF will be the best model. And we need to help top-10 poorest countries such as
1 Burundi
1 Liberia
1 Congo, Dem. Rep.
1 Niger
1 Sierra Leone
1 Madagascar
1 Mozambique
1 Central African Republic
1 Malawi
1 Eritrea
.

- **Suggestions for next steps** It will be interesting to analyse the reason why the NMF clusters our data into 3 groups not 4 groups which means the highest variance belonged only into 3 groups. Does it mean that we can reduce the principle components into 3 not 4?