MULTIVARIATE ANALYSIS PROJECT

AIM: To perform Principal Component Analysis and Factor analysis to analyze the data and using various hierarchical and non-hierarchical clustering methods to cluster the countries based on socio-economic and health factors to help and provide funds to poverty driven areas.

DATA DESCRIPTION: 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.
HELP International has 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, the CEO has to make a decision to choose the countries that are in the direst need of aid. Hence, the job as a Data scientist is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then suggest the countries which the CEO needs to focus on the most.

The Data has the following columns as the variables:-

1. Countries: This has the list of countries which are subjected to testing if they are liable to receive the aid through the HELP International. There are 167 countries in the dataset.
2. Child Mortality Rate: This denotes death of children under 5 years of age per 1000 live births. This adds up to the health facility factor of any country and is one of the parameter.
3. Exports: This explains exports of goods and services per capita. In the dataset, this is given in terms of percentage.
4. Health: This variable explains the amount of total health spending per capita. It is given in terms of percentage.
5. Imports: This explains the imports of goods and services per capita. In the dataset, this is explained in terms of age percentage of GDP per capita.
6. Income: This explains the average income of a person in the country. Income is one of the main factors to interpret the economy of a country.
7. Inflation: This variable accounts for total percentage increase in the GDP over the year of a country.
8. Life Expectancy: This variable quantifies the average number of years a new born baby would live given the current mortality rate.
9. Total Fertility Rate: This variable explains the amount of children born to a female given the age fertility rates.
10. GDPP: This explains the GDP per capita. Simple calculation leads us to obtaining GDPP as the total GDP divided by the total population.

Description of the variables in terms of central tendencies and dispersion parameters is given in the table down below:

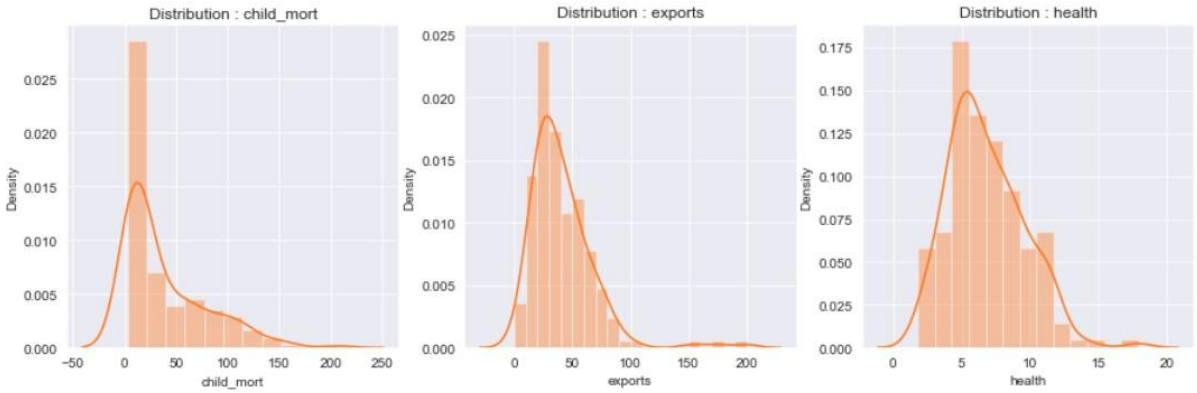
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00	167.00
mean	38.27	41.11	6.82	46.89	17144.69	7.78	70.56	2.95	12964.16
std	40.33	27.41	2.75	24.21	19278.07	10.57	8.89	1.51	18328.70
min	2.60	0.11	1.81	0.07	609.00	-4.21	32.10	1.15	231.00
25%	8.25	23.80	4.92	30.20	3355.00	1.81	65.30	1.79	1330.00
50%	19.30	35.00	6.32	43.30	9960.00	5.39	73.10	2.41	4660.00
75%	62.10	51.35	8.60	58.75	22800.00	10.75	76.80	3.88	14050.00
max	208.00	200.00	17.90	174.00	125000.00	104.00	82.80	7.49	105000.00

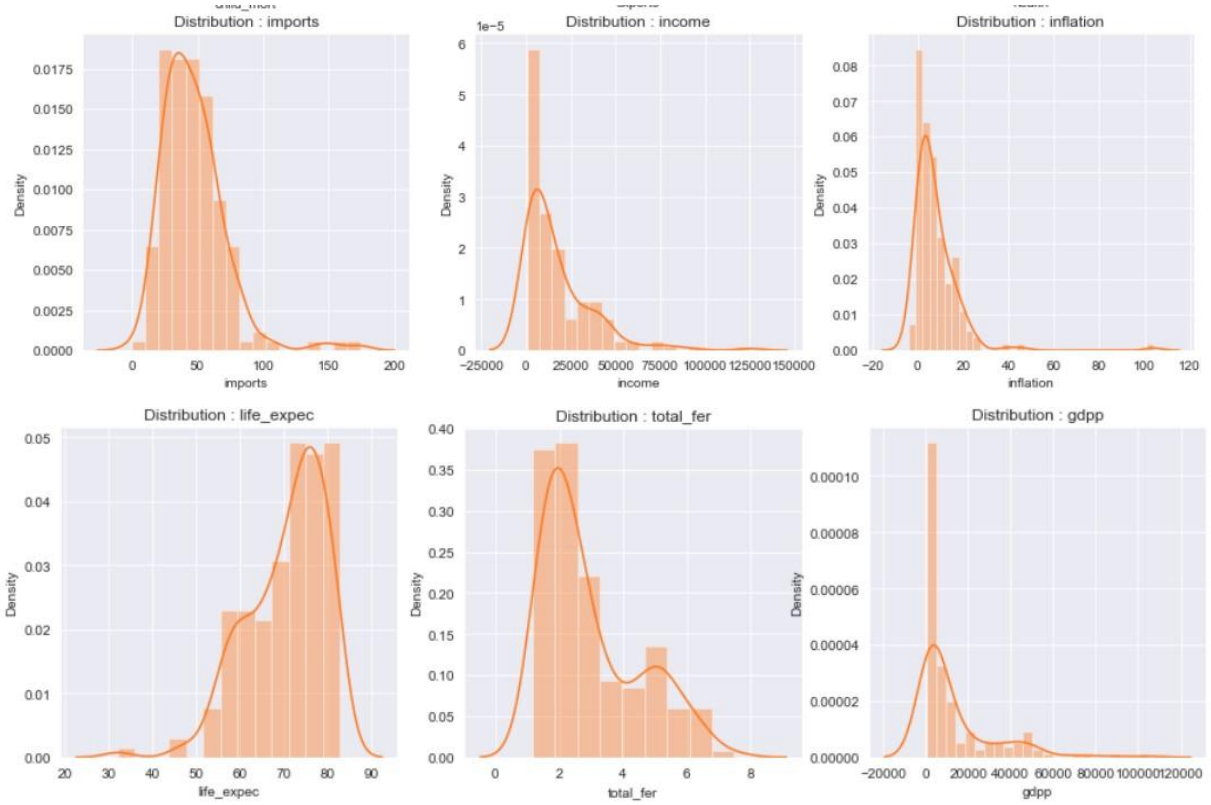
Exploratory Data Analysis:

It is very important to perform EDA on a dataset. EDA gives a clear picture on the contribution of each variable, characteristics of each variable and the interaction between the variables. This will also give us statistical and logical interpretation to perform the latter statistical techniques.

I. Histogram and Densities of each variable.

This will help us understand the skewness of the distribution and graphical interpretation on what the data looks like in terms of a particular variable.





=> Points to Infer:

- From the above histogram and density plot, it is clear that the “health” variable appears to be normally distributed with centering 6.82%.
- “Life-Expectancy” variable appears to be negatively skewed which explains the life expectancy to be more than 55 years in almost all countries. More the Life-Expectancy, the better the health care facilities.
- Rest of the variables tend to be right skewed which is not a great visual especially for certain variables like “Income” where we expect a good income but most countries have income less than 20000 dollars. “Inflation” variable being less is an indication that not many countries have experienced an increase in the net GDP over the year. Similar interpretation can be obtained for “GDPP” variable as well.
- “Country” Variable is not a numeric variable and cannot be interpreted using histograms.

=> Features of Economically backward countries:

- The country's **per capita income is very low.**
- **High Population** that leads to non - availability of resources.
- **Unemployment** due to less resources.
- Low country wealth that leads to **low capital**.
- Inequitable distribution of **wealth** and **income**.
- Lack of proper educational amenities and thus **illiteracy prevails**.
- **Low level of living.**

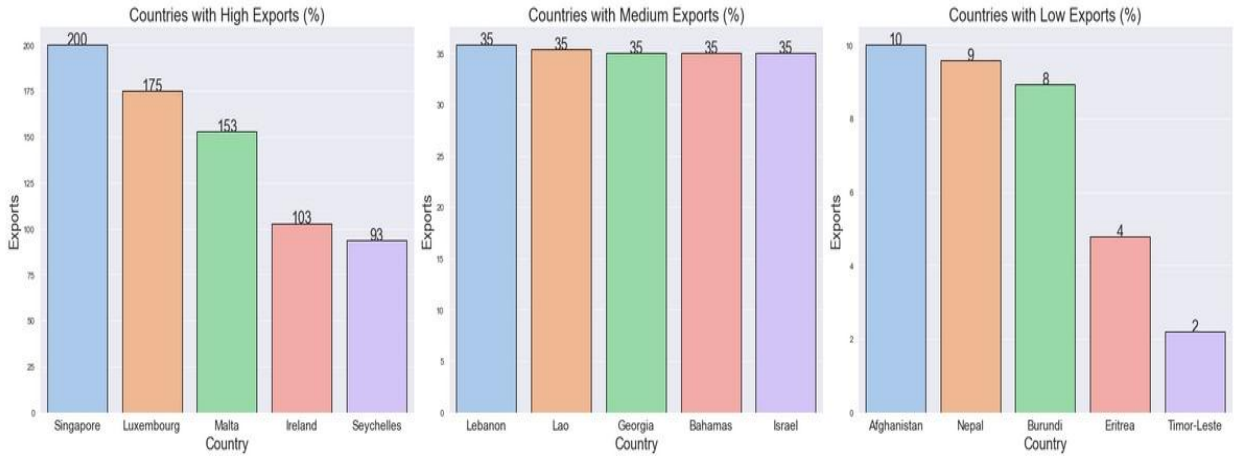
- **No technical advancement.**
- **Poor health services** coupled with high birth & death rates.

HELP International needs to focus on the countries with above characteristics. They are in need of aid.

II. Display of countries with High, Medium and Low intensity characteristics.

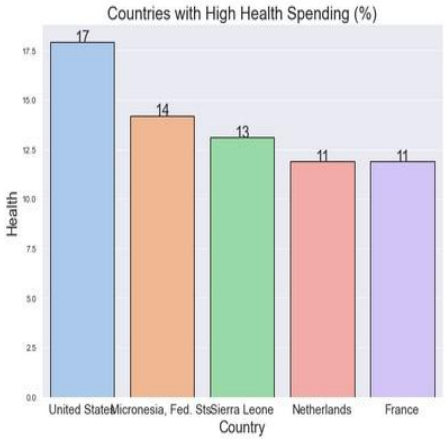
The data has been sorted with respect to each variable and divided into 3 parts. Grouping of the 3 parts have been done based on the features pertaining to backward countries to obtain the performance of each country in each variable and labeling them as High, Medium and Low value of the variable. The top 5 countries in each of these characteristics has been shown in the figures down below:

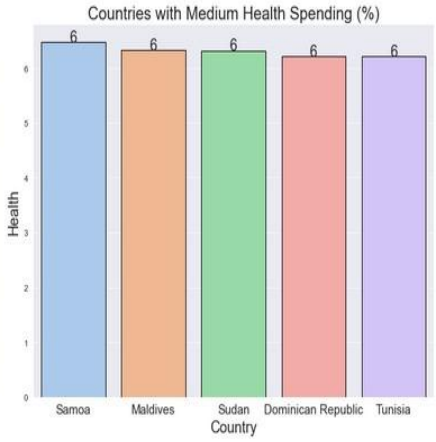
1. Countries - Exports:

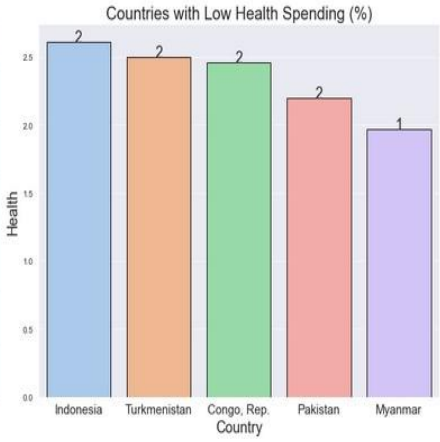


- Exports of a nation are usually goods and services created domestically but sold to other nations. Goods and services exported depend on factors like the geographical location, natural resources, population size & their preference towards specific skills, etc.
- Despite Singapore's population size not being in the top 100, they have the highest number of exports. Luxembourg & Malta have probably followed the same route.
- Afghanistan & Nepal are present in the lower end of exports. Geographical locations of these nations have a heavy influence. Countries with lower exports also have small geographical areas.
- Most of the Asian countries like Lebanon, Laos and Israel have a medium export value around 35.

2. Countries - Health:



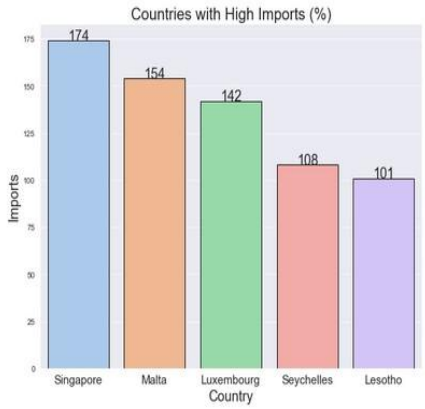


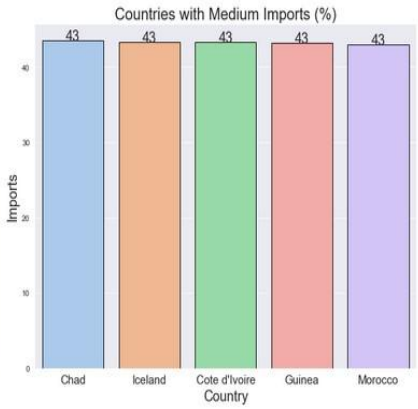


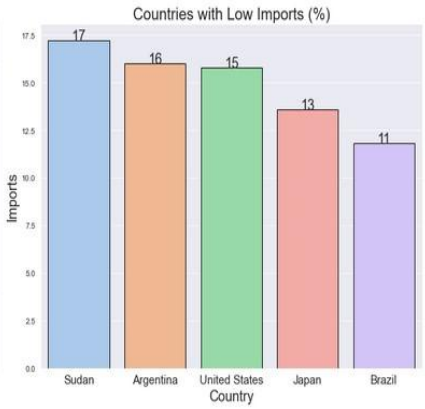
=> Points to Infer:

- The United States is one among the top countries to have a larger percentage of total health spending per capita with 17% individual GDP contribution. The United States is also said to have better health care facilities. Many European countries also showed higher health spending per capita values.
- Some of the African countries showed a medium value for health spending per capita with 6% being the average health spending in such countries.
- Most of the Asian countries have very less value of health spending per capita. It is also a fact that most Asian countries have poor health care facilities and are often ignorant when it comes to health. Not to forget the origin of COVID-19!

3. Countries - Imports:





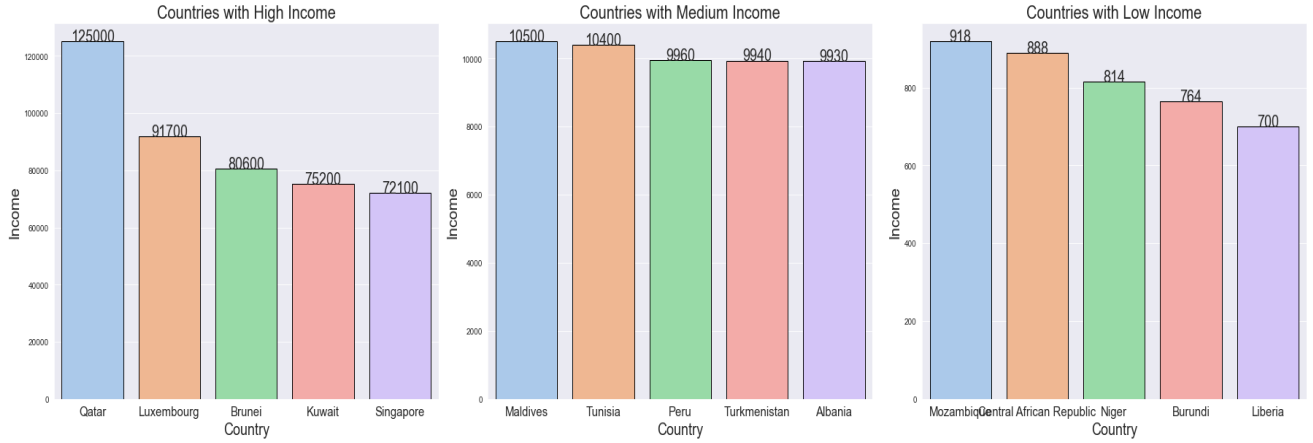


=> Points to Infer:

- When the Export value exceeds the Import value, it is a good sign for the developing countries. The United States has a lesser value of import per capita value followed by Japan. These two countries are said to be developing countries.

- Singapore is said to be Highly developed country and still possesses a higher import value per capita. Imports help the countries achieve a good relationship with other countries.
- Most countries lying in the medium import value are least developed like Chad and Guinea with mean import value percentage as 43%.

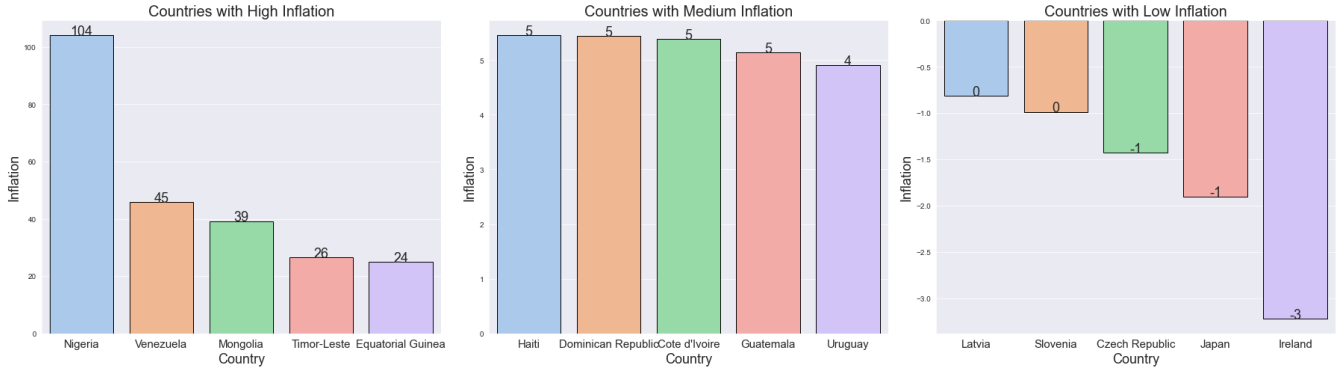
4. Countries - Income:



=> **Points to Infer:**

- Income is one of the important attributes to measure the development of a country. More income implies more GDP.
- Citizens of Qatar have the highest income out of all the countries with a difference of 30k more than the 2nd placed countries. Singapore & Luxembourg are again present in the top 5 of another feature.
- Lower end of the income is dominated by the African nations. This is influenced by the damage done by colonization out of which the nations have not yet recovered.
- The difference in the income of the nations in the top, middle and lower end is quite significant and will have an effect on every other feature.

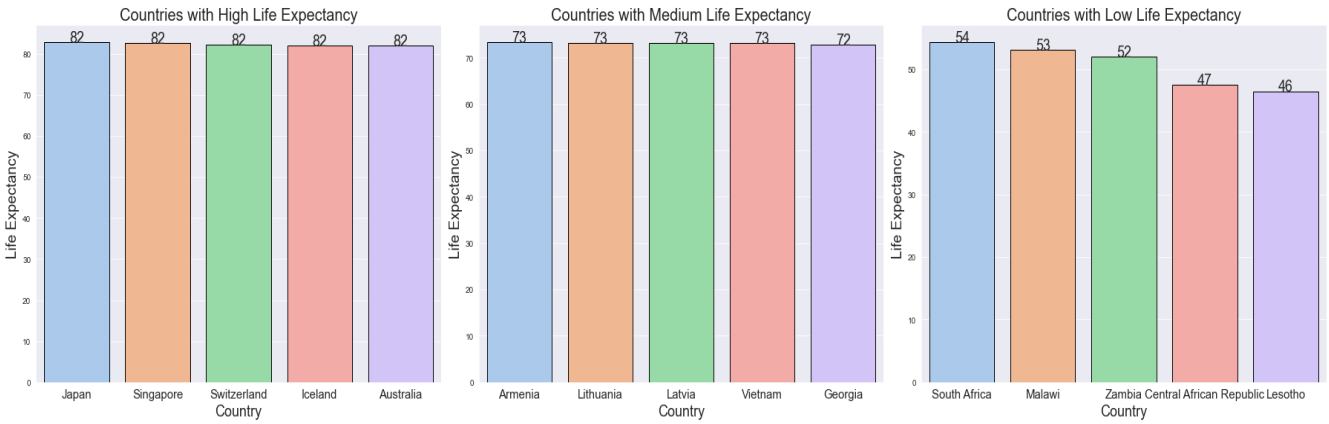
5. Country - Inflation:



=> Points to Infer:

- Higher inflation reduces the purchasing power of the citizens. Countries present at the top end of inflation have a devastating economic situation. Having such high inflation is a risk to the existence of the nation.
- Similarly, the lower end of inflation has negative values i.e known as deflation. It signals an economy in a downward spiral leading to a recession or even a depression.

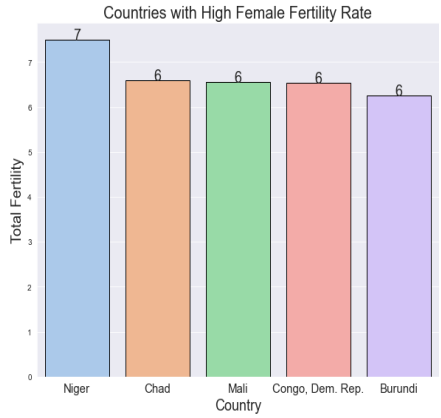
6. Country - Life Expectancy:

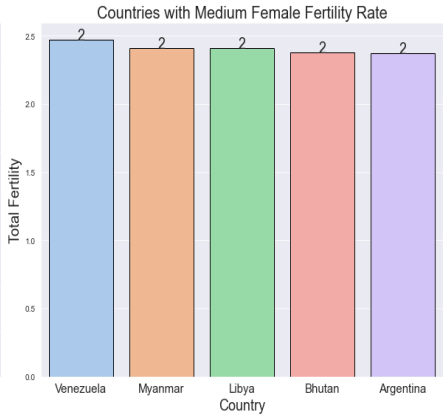


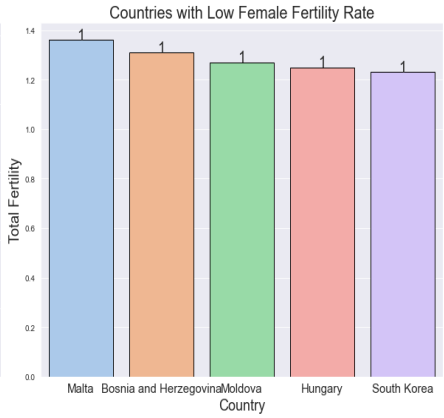
=> Points to Infer:

- Life Expectancy depends a lot on mental state as well as the lifestyle adopted by the citizens. Singapore is again present in the top 5 of this feature. Life Expectancy also depends on the health care facilities in the country.
- None of the countries with a high life expectancy are present in the top 5 of countries with higher health spending per capita value. (Refer graph above “health spending per capita)
- African countries are again present in the lower end for another feature which signifies the poor health care facilities and lifestyle deprivation in these countries.

7. Country - Total Fertility rate:



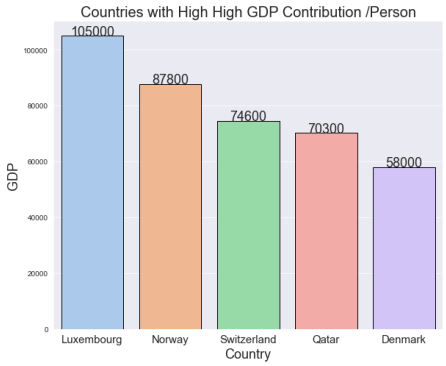


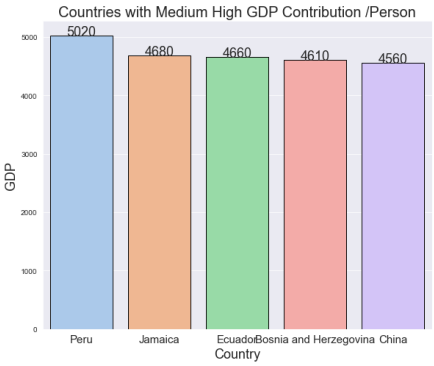


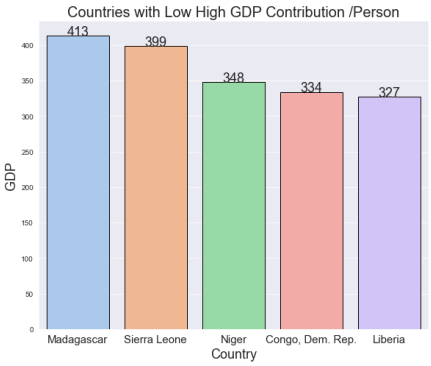
=> Points to Infer:

- Most African countries have a Total Fertility Rate greater than 6 which is not a good sign considering its income levels and health care facilities.
- Many Asian countries have a medium total fertility rate with mean TFR 2.
- Low total fertility rates produces a rapidly aging population and declining population rate.

8. Country - GDPP:





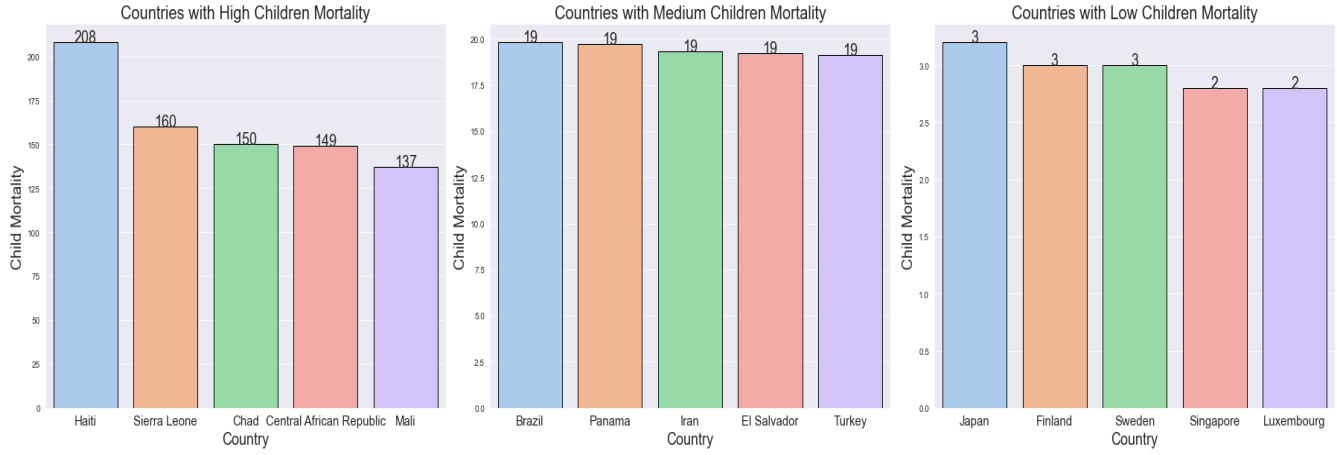


=> Points to Infer:

- This feature is a bit complicated since for example, China is one of the leading super developed countries but also possesses a greater population. As a result it belongs to the category of medium GDPP contribution. It may also happen that a country with the smallest population might yield a larger GDPP value.
- Luxembourg is again present in the top ranks. Switzerland & Qatar are present in the top 5 similar to income.

- Lower end is again dominated by African nations that label them as economically backward.

9. Country - Mortality Rate:



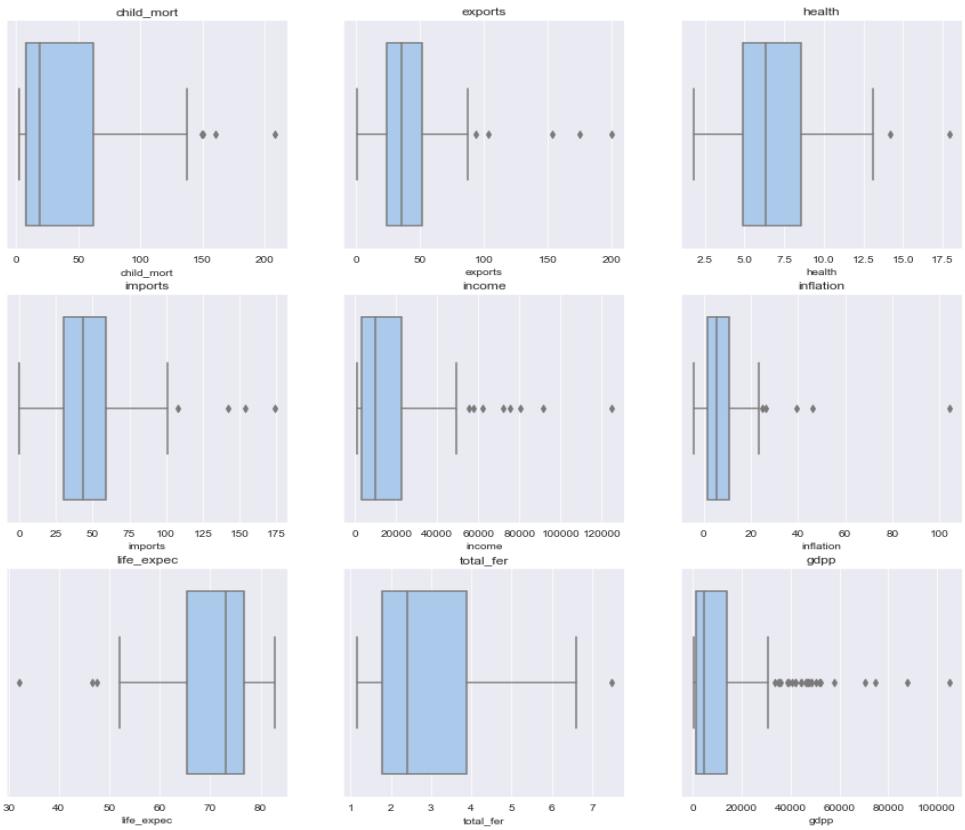
=> Points to Infer:

- Child Mortality rate may indicate a lot of situations. It may be because of lack of proper health care facilities or religious-political matters. Child Mortality rate is a very bad sign for developing countries since it may decline the population size.
- Haiti has the highest children's deaths. African countries have significant positions in this statistic.
- At the other extreme of child mortality rate, countries from Asia and Europe have some solid presence.

III. Boxplot of explanatory variables:

Box plots are crucial for summarizing dataset distribution in a visually clear way. They highlight key measures, making it easy to compare groups and identify patterns. These plots are simple yet powerful tools for researchers, offering insights and aiding informed decision-making based on data distribution.

Following is the depiction of Box Plots of the variables:



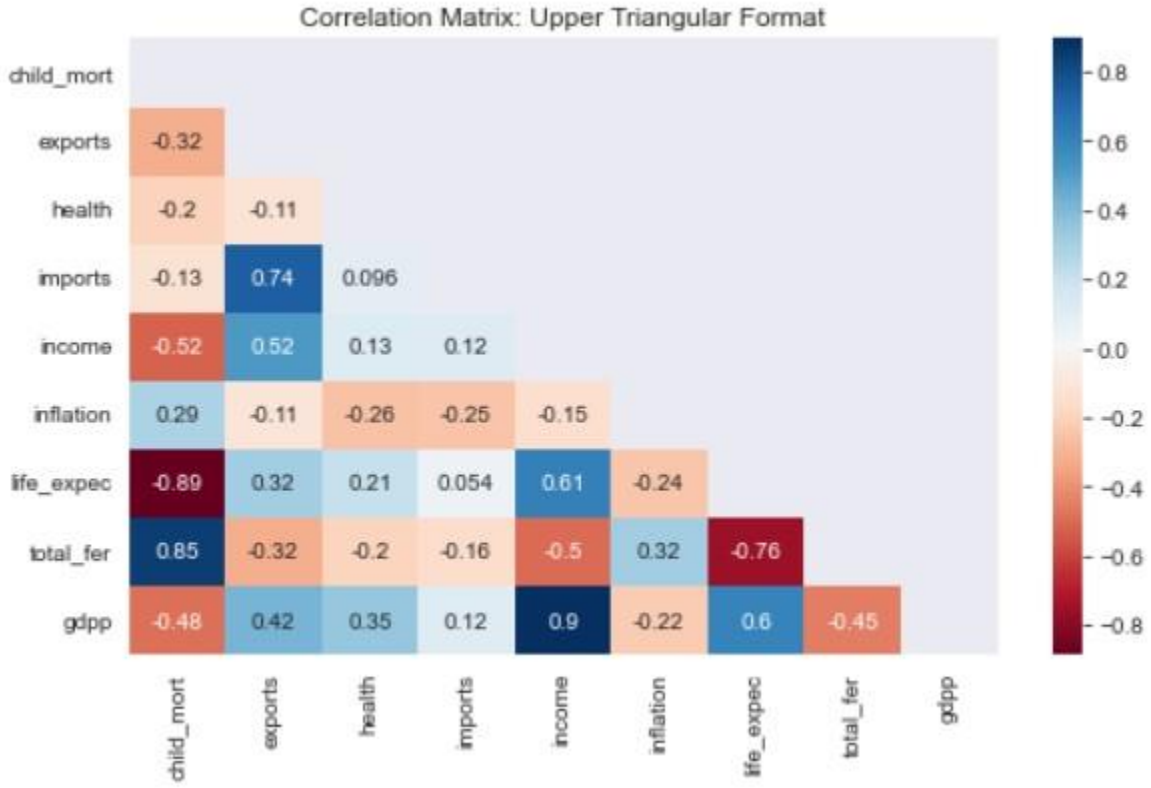
=> Summary of Exploratory Data Analysis:

- From the visualizations and the list of features of economically backward nations, a host of insights can be gained!
- When it comes to health conditions, African countries hold higher ranks in all the wrong situations. They hold a significant presence in high child mortality rate, low life expectancy and high total fertility rate.
- All these problems are already pretty serious and hence it is very important to assist them during the periods of unforeseen turmoil. Despite such numbers, Haiti grabs the top spot with high values of child mortality rate. Asian & European countries are present at the other end of it.
- US citizens are the highest spenders on their health however they are not present in the top 5 ranks of life expectation & total fertility rate. None of the countries with a high life expectancy are present in the top 5 of health. Asian countries crowd the lower end of health.
- Singapore, Malta, Luxembourg & Seychelles are present in the top 5 of exports as well as imports. Population size and geographical locations play a pivotal role when it comes to imports and exports.
- Sudan is the only African nation with low imports and Brazil has the lowest imports out of all.
- African countries display very high values of inflation whereas countries from all the continents can be found with low inflation values.

- Citizens of Qatar are the highest paid with Singapore & Luxembourg again grabbing spots in top 5 of income.
- For gdpp, Luxembourg is in the top ranks. Switzerland & Qatar are present in the top 5 similar to income.
- African nations are present in the lower end of income as well as gdpp. Colonization has had a huge toll on the African nations.

Correlation Matrix (Heat Map)

Correlation is an important statistical measure to check association between two variables. Logically, the variables seem to have some association. For example, Life Expectancy and Health, GDPP and Income. Following is a heat map which depicts the correlation between pair of variables:



=> Points to Infer:

- Many features have relationships with each other.
- Child mortality clearly increases when income, gdpp & exports decreases. Rise in inflation also leads to high child mortality cases. Economic conditions unfortunately act as an important factor!
- Rise in exports clearly increases gdpp, income & imports.
- Spending on health has a small rise in life expectancy and also decreases the child mortality.

- Income & gdpp display a very high 0.9 correlation value. From the health perspective, high income has led to higher life expectancy but decreases the total fertility by some significant margin.
- As expected, high inflation has a negative effect on the financial features. High inflation displays a high total fertility rate and child mortality. This describes the typical features of a backward nation.
- According to the data, higher life expectancy displays a low total fertility. Higher gdpp has lead more spending on health.

Note: An important thing to understand is there are some groups of features which have a similar association with other group variables. The groups can be identified by observing the correlation heat map.

The 3 categories of the features are :

- Health : child mortality rate, health, life expectancy, total fertility rate.
- Trade : imports, exports
- Finance : income, inflation, gdpp

Hence, these features are dissolved into these categories and normalized.

Before proceeding to Principal component analysis and factor analysis, the data needs to be standardized to yield better results.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.29	-1.14	0.28	-0.08	-0.81	0.16	-1.62	1.90	-0.68
1	-0.54	-0.48	-0.10	0.07	-0.38	-0.31	0.65	-0.86	-0.49
2	-0.27	-0.10	-0.97	-0.64	-0.22	0.79	0.67	-0.04	-0.47
3	2.01	0.78	-1.45	-0.17	-0.59	1.39	-1.18	2.13	-0.52
4	-0.70	0.16	-0.29	0.50	0.10	-0.60	0.70	-0.54	-0.04

PRINCIPAL COMPONENT ANALYSIS:

Principal Component Analysis (PCA) is a statistical technique widely used in data analysis and dimensionality reduction. It involves transforming a set of correlated variables into a new set of uncorrelated variables, called principal components. These components are linear combinations of the original variables and are arranged in order of the variance they capture, with the first component explaining the most variance.

The primary goal of PCA is to simplify complex datasets while retaining as much variability as possible. By doing so, it becomes easier to visualize patterns, reduce noise, and facilitate subsequent analyses. PCA finds applications in various fields, such as image processing, finance, biology, and engineering. It is employed for tasks like feature extraction, noise reduction, and identifying underlying patterns in data. In essence, PCA provides a powerful tool for understanding and summarizing the inherent structure within high-dimensional datasets, contributing to more efficient and meaningful data analysis.

Principal Component $Y_i = e^T_i X$ for i=1,2,..9 here.

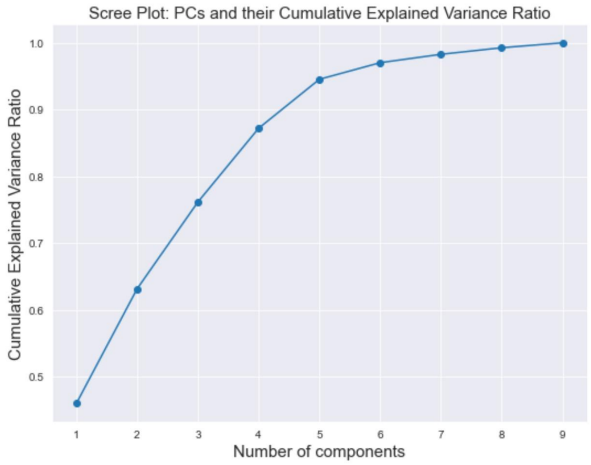
Where e_i's are the eigenvectors of the correlation matrix.

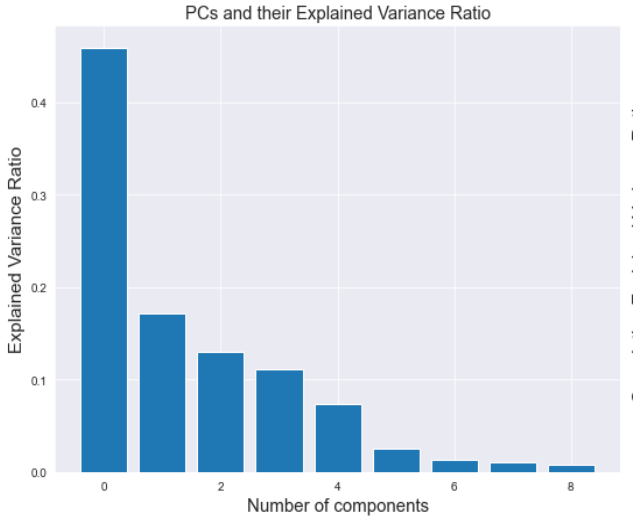
Obtained Principal components are given below in the table:

	Features	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0	child_mort	-0.42	0.19	-0.03	0.37	-0.17	0.20	-0.08	-0.68	0.33
1	exports	0.28	0.61	0.14	0.00	0.06	-0.06	-0.71	-0.01	-0.12
2	health	0.15	-0.24	-0.60	0.46	0.52	0.01	-0.25	0.07	0.11
3	imports	0.16	0.67	-0.30	-0.07	0.26	-0.03	0.59	-0.03	0.10
4	income	0.40	0.02	0.30	0.39	-0.25	0.16	0.10	0.35	0.61
5	inflation	-0.19	-0.01	0.64	0.15	0.71	0.07	0.10	-0.01	-0.03
6	life_expec	0.43	-0.22	0.11	-0.20	0.11	-0.60	0.02	-0.50	0.29
7	total_fer	-0.40	0.16	0.02	0.38	-0.14	-0.75	0.03	0.29	-0.03
8	gdpp	0.39	-0.05	0.12	0.53	-0.18	0.02	0.24	-0.25	-0.63

Considering all the principal components is a sheer waste of statistical methods since the main goal to perform PCA is to reduce the dimension. Scree plot helps one to understand the total variance explained by each of the principal components. If we stop considering the principal components at 90%, the job is done.

Figure down below gives the scree plot and the variance explained by each principal component:

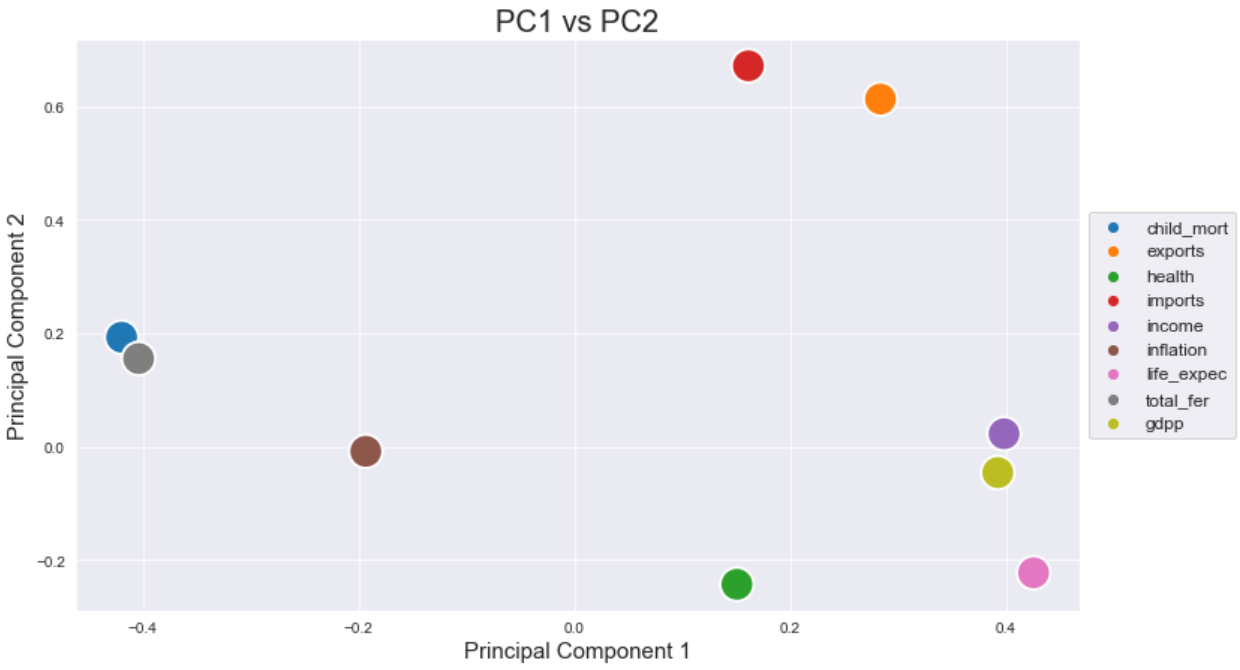




Considering the first 5 components will explain 90% of the variance but from the scree plot, after 5 components, there is not much variance left to be explained and a steady decrease can be observed.

The first two principal components are very crucial as they alone explain more variance. A plot of these two principal components will help interpret the variables dominant in these two principal components.

Figure down below is a plot of PC1 v/s PC2:

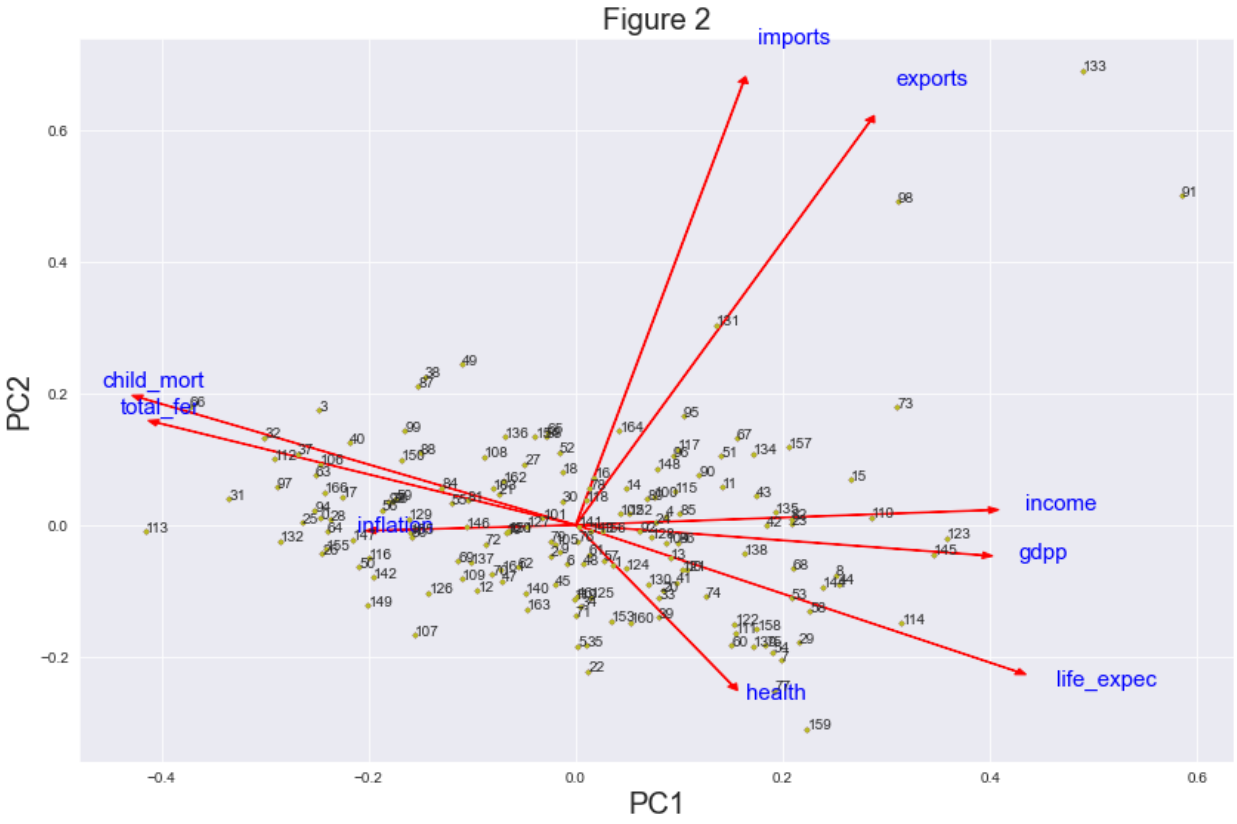


- The first principal component is gravitated more towards “total fertility rates”, “life expectancy”, “health”, “child mortality” and “gdpp”. This component explains the contribution of a healthy population to the overall development. There are a lot of health factors that govern this component
- The second principal component is gravitated more towards “exports” and “imports” which can be interpreted as association of the country with its fellow countries for its dependence and contribution through imports and exports.

BI-PLOT ANALYSIS:

In a biplot, the length of the lines approximates the variances of the variables. The longer the line, the higher the variance. The angle between the lines, or, to be more precise, the cosine of the angle between the lines, approximates the correlation between the variables they represent. The closer the angle is to 90, or 270 degrees, the smaller the correlation. An angle of 0 or 180 degrees reflects a correlation of 1 or –1, respectively.

Following is the bi-plot analysis of PC1 and PC2:



Points to Infer:

- Life Expectancy and imports seem to be completely uncorrelated. Life expectancy is dominated as a variable in PC1 and imports is dominated as a variable in PC2.

- Child Mortality, imports and exports has the maximum variance whereas inflation has the minimum variance.
- Child Mortality and Life expectancy are in contrast.
- (Income, GDPP) and (Child Mortality, Total fertility rate) seem to have a high positive correlation and they dominate in PC1.

FACTOR ANALYSIS:

Factor Analysis is a statistical method used for identifying underlying patterns, or latent factors, within a set of observed variables. It assumes that observed variables are influenced by common factors and unique factors specific to each variable. The technique aims to capture the shared variance among variables and reduce them to a smaller number of latent factors.

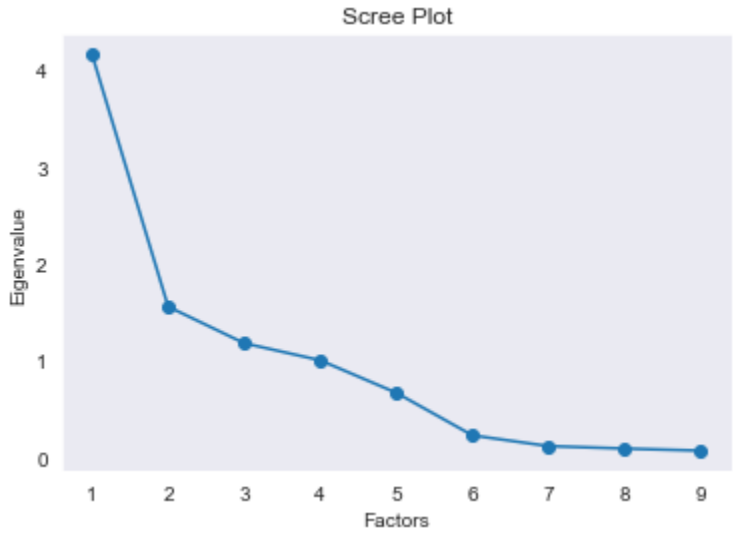
The primary purpose of Factor Analysis is dimensionality reduction and simplifying complex datasets by revealing the essential structure. It helps uncover the relationships between observed variables, making it easier to interpret and understand the underlying factors influencing the data. Factor Analysis finds applications in various fields, including psychology, economics, and marketing, where it is used for constructing indices, simplifying survey data, or identifying latent constructs.

Firstly, one has to observe the plot of Eigenvalues of correlation matrix. This will give a knowledge about the number of factors to include in the factor model.

The eigenvalues, variance explained by each variable and the cumulative variance is given in the table below:

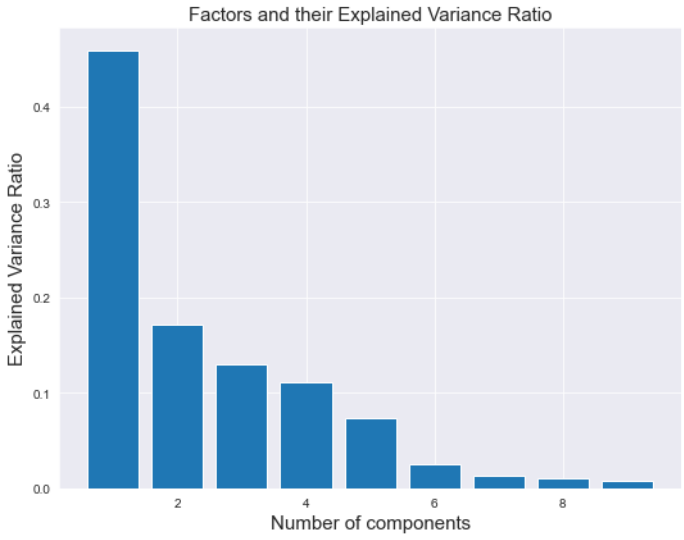
	Eigen_val	Proportion Var	Cumulative Var
1	4.16	0.46	0.46
2	1.56	0.17	0.63
3	1.18	0.13	0.76
4	1.00	0.11	0.87
5	0.66	0.07	0.95
6	0.22	0.02	0.97
7	0.11	0.01	0.98
8	0.09	0.01	0.99
9	0.07	0.01	1.00

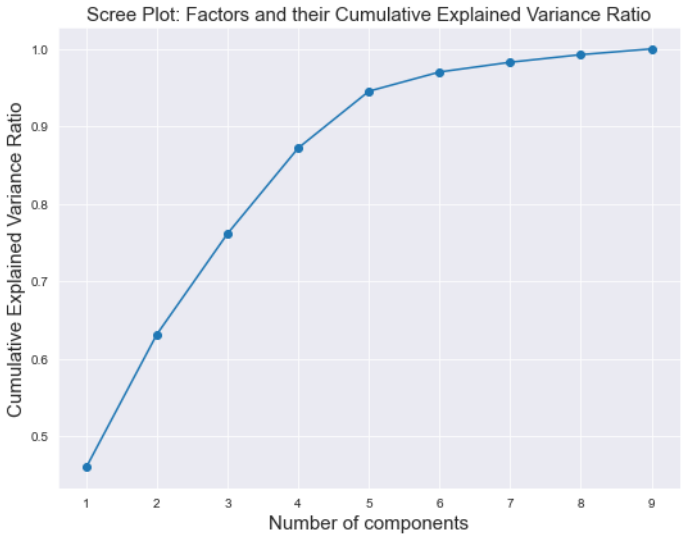
Scree plot of the eigenvalues is given below:



The graph decreases in a lesser rate steadily after the fifth eigenvalue. This is an indication that the factor model will be well estimated for 5 factors.

- The Scree plot and graph of the number of components v/s proportion of variance is shown below:





Similar interpretation can be obtained in this as well. Factor model will be well estimated for 5 factors.

Factor loadings for 5-Factor Model is given below in the table:

	1	2	3	4	5	Communalities	Specific Variance
child_mort	-0.861817	0.256833	0.115015	0.409425	-0.076482	0.995397	0.010627
exports	0.568741	0.683238	0.227697	-0.095659	-0.091966	0.859735	0.146289
health	0.279642	-0.172964	-0.318951	0.384753	0.498945	0.606827	0.399197
imports	0.335173	0.887453	-0.193368	-0.058923	0.233680	0.995384	0.010640
income	0.800834	-0.019713	0.378113	0.274732	-0.165687	0.887624	0.118400
inflation	-0.399224	-0.071246	0.791957	-0.242559	0.380153	0.995003	0.011021
life_expec	0.840554	-0.273433	0.013631	-0.218926	0.033578	0.830538	0.175486
total_fer	-0.777274	0.170440	0.127801	0.287318	-0.033744	0.733228	0.272796
gdpp	0.809975	-0.082308	0.277621	0.506070	-0.001798	0.996018	0.010007

=> Points to Infer:

- The first factor has major loadings on Child Mortality rate, income, life expectancy, total fertility rate and GDPP. Child Mortality and total fertility rate is negatively loaded in the factor F1. This factor can be interpreted as **economic stability** of a country.
- The second factor is majorly loaded with exports and imports. This factor can be interpreted as **external affairs** with other countries which plays a major role in the development of a nation.

Factor loadings by performing varimax rotation is given below:

	1	2	3	4	5	Communalities	Specific Variance
child_mort	0.969352	-0.183766	-0.087113	0.090161	-0.079162	0.995397	0.010627
exports	-0.206445	0.379718	0.788353	0.000063	-0.226780	0.859735	0.146289
health	-0.115885	0.115793	-0.015044	-0.121243	0.751707	0.606827	0.399197
imports	-0.024956	-0.031958	0.980653	-0.125212	0.127987	0.995384	0.010640
income	-0.367372	0.855194	0.142919	-0.024433	-0.016787	0.887624	0.118400
inflation	0.175518	-0.055869	-0.106624	0.962733	-0.151167	0.995003	0.011021
life_expec	-0.829378	0.361349	0.029963	-0.059453	0.087547	0.830538	0.175486
total_fer	0.805192	-0.208096	-0.119232	0.144349	-0.080852	0.733228	0.272796
gdpp	-0.285421	0.910695	0.095198	-0.072393	0.266239	0.996018	0.010007

=> **Points to Infer:**

- By performing varimax rotation, every factor has a set of unique variables describing it.
- **Factor 1:** This is majorly loaded with child mortality, life expectancy and total fertility rate. This can be interpreted as the **population parameter.**
- **Factor 2:** This is majorly loaded with income and GDPP. This can be interpreted as the **economic parameter.**
- **Factor 3:** This is majorly loaded with imports and exports and can be interpreted as the **external affairs parameter.**
- **Factor 4:** This is majorly loaded with Inflation and can be interpreted as **Inflation parameter.**
- **Factor 5:** This can be interpreted as the **Health parameter.**
- It seems varimax rotation has provided a clear interpretation on the factors.

CLUSTERING METHODS:

The last stage of the project after interpretation of variables and the forming principal components is forming clusters. The main aim of the project was to suggest the countries that need AID by HELP International.

Two kinds of Clustering methods have been used:

1. Hierarchical Clustering- Single Linkage, Complete Linkage and Average Linkage.
2. Non-Hierarchical Clustering- K-Means.

The distance matrix is the correlation matrix.

I. Hierarchical Clustering: Single Linkage Method:

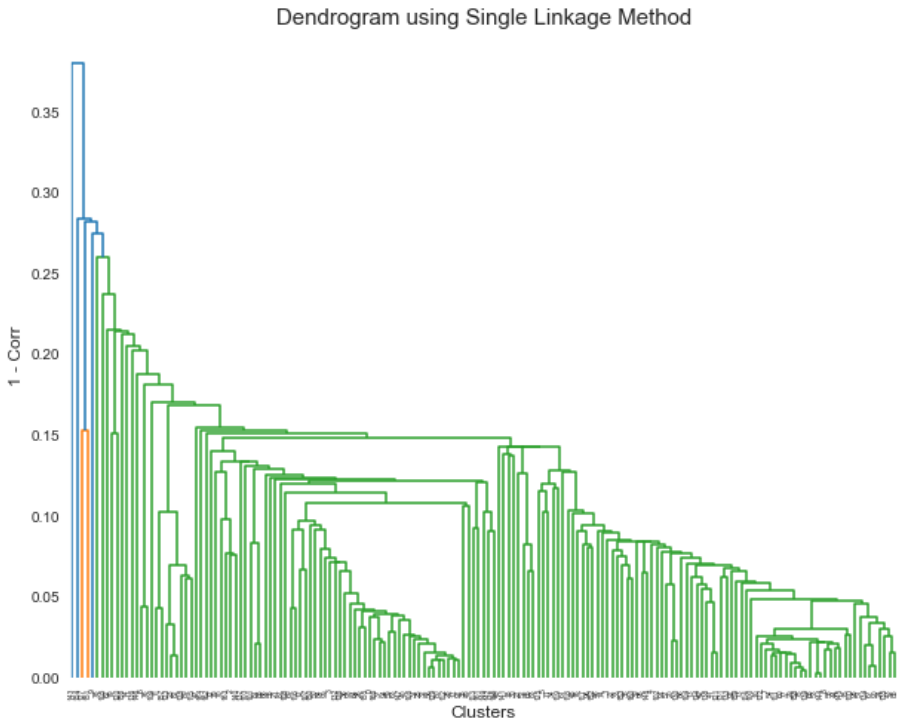
Single Linkage is a hierarchical clustering method used in data analysis. It connects clusters based on the closest data points, creating a dendrogram that reveals relationships within the dataset. It is particularly useful for identifying tight-knit groups in proximity. Single Linkage finds applications in biology, computer science, and pattern recognition. By emphasizing local similarities, it aids in understanding the hierarchical structure of data. Overall, Single Linkage offers a straightforward yet effective approach to hierarchical clustering, contributing to the exploration of relationships and patterns in diverse datasets.

Here the distance between the clusters (UV) and W is given by:

$$d_{(UV),W} = min\{d_{UV}, d_W\}$$

Usually Single Linkage method is not preferred since the distance between the clusters is calculated using a minimum which makes the maximum distance vanish but is a wrong step as larger distances will still be clustered just because one of the distances is very small.

The Dendrogram for Single Linkage method is given below:



Single Linkage method has not worked well for the dataset and as a result this is not considered in the selection of the best clustering method.

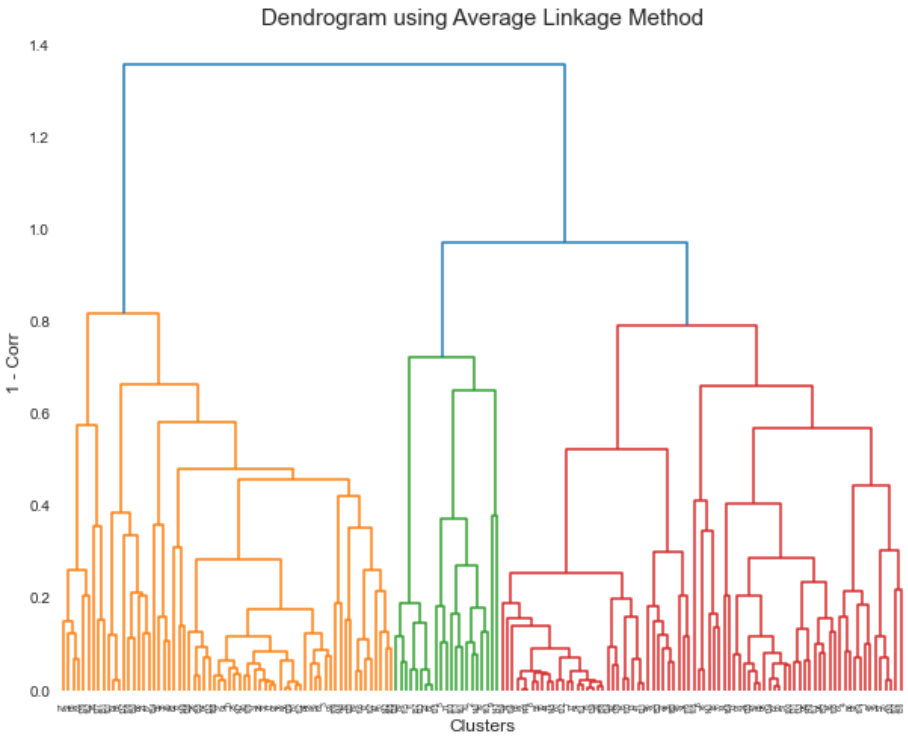
II. Hierarchical Clustering: Average Linking Method -

Average Linkage is a hierarchical clustering method that connects clusters based on the average distance between all pairs of data points from different clusters. It forms a dendrogram, revealing relationships within the dataset. This method is valuable for identifying clusters with a balanced internal structure.

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

where $|C_i|$ and $|C_j|$ are the sizes of clusters C_i and C_j,

The dendrogram of average linkage for the dataset is given below:



The orange, green and red coloured groups in the dendrogram form 3 clusters and the clusters have been listed below:

Cluster 0: Afghanistan, Angola, Bangladesh, Benin, Bolivia, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Egypt, Equatorial Guinea, Eritrea, Fiji, Gabon, Gambia, Ghana, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, India, Iraq, Jordan, Kenya, Kiribati, Kyrgyz Republic, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Micronesia, Fed. Sts., Mozambique, Myanmar, Namibia, Nepal, Niger, Pakistan, Philippines, Rwanda, Samoa, Senegal, Sierra Leone, Solomon Islands, South Africa, Sudan, Tajikistan, Tanzania, Timor-Leste, Togo, Tonga, Turkmenistan, Uganda, Vanuatu, Yemen, Zambia

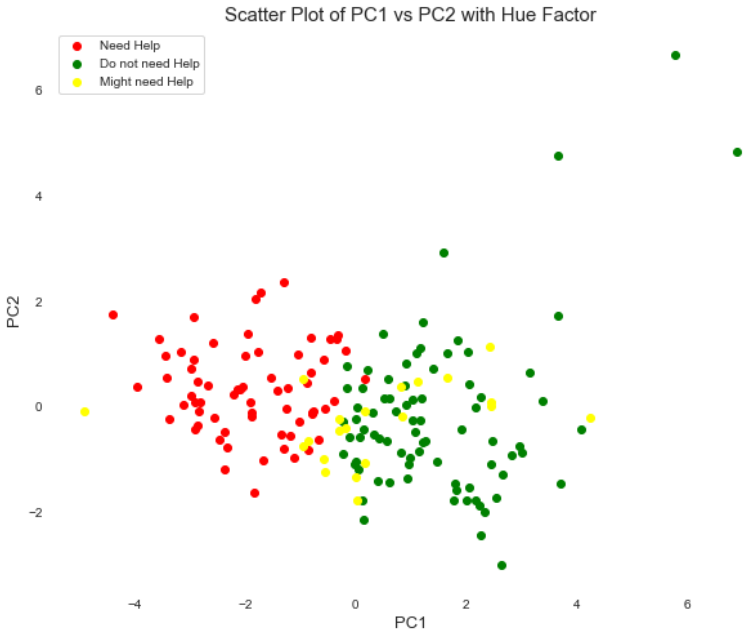
Cluster 1: Albania, Antigua and Barbuda, Armenia, Australia, Austria, Bahamas, Barbados, Belarus, Belgium, Belize, Bhutan, Bosnia and Herzegovina, Brazil, Bulgaria, Canada, Cape Verde, Chile, China, Colombia, Costa Rica, Croatia, Cyprus, Czech Republic, Denmark, Dominican Republic, Ecuador, El Salvador, Estonia, Finland, France, Georgia, Germany, Greece, Grenada, Hungary, Iceland, Ireland, Israel, Italy, Jamaica, Japan, Latvia, Lebanon, Lithuania, Luxembourg, Macedonia, FYR, Malaysia, Maldives, Malta, Mauritius, Moldova, Montenegro, Morocco, Netherlands, New Zealand, Norway, Panama, Paraguay, Peru, Poland, Portugal, Romania, Serbia, Seychelles, Singapore, Slovak Republic, Slovenia, South Korea, Spain, St. Vincent and the Grenadines, Sweden, Switzerland, Thailand, Tunisia, Turkey, Ukraine, United Kingdom, United States, Uruguay, Vietnam

Cluster 2: Algeria, Argentina, Azerbaijan, Bahrain, Brunei, Indonesia, Iran, Kazakhstan, Kuwait, Libya, Mongolia, Nigeria, Oman, Qatar, Russia, Saudi Arabia, Sri Lanka, Suriname, United Arab Emirates, Uzbekistan, Venezuela

- Cluster 0 has majorly African and Asian countries which have poor economic states, bad health care facilities, income levels, high child mortality rates, etc. **These countries need AID.**

- Cluster 1 has American and European countries which are super developed and have strong economic conditions. **These countries do not need help.**
- Cluster 2 has mixed countries which may need help.

Scatter Plot of PC1 v/s PC2 with hue factor:



- From the above plot and the biplot of the PC, we can see clear clusters of countries that need and don't need help.
- Countries with higher value of PC1 have been listed under safe zone which is clear since PC1 describes the economic conditions of the country including GDPP

Silhouette Coefficient: Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

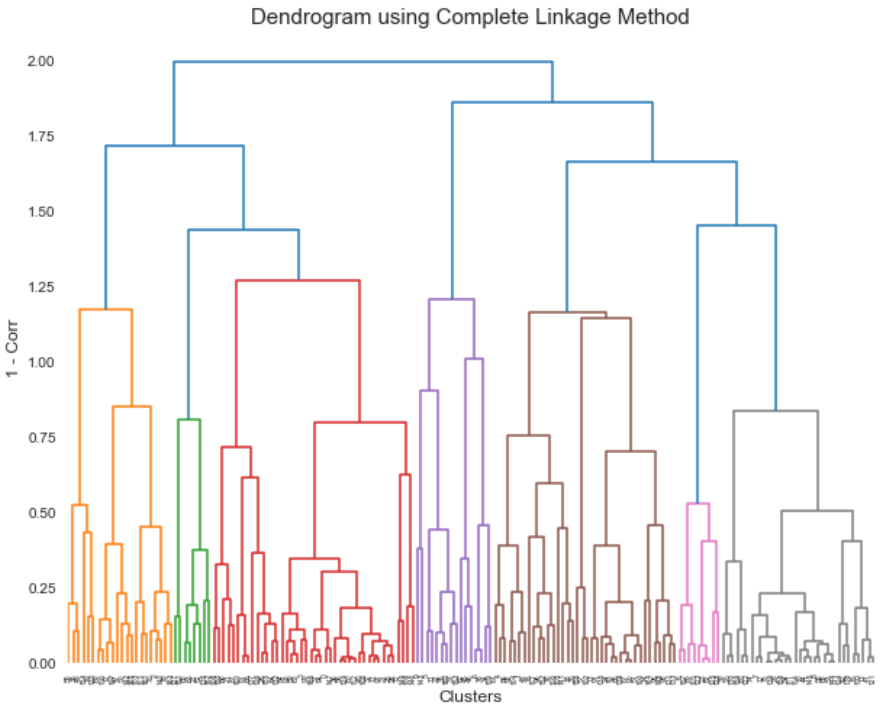
Silhouette score for Average Linkage comes up to be around 0.413.

III. Hierarchical Clustering: Complete Linkage Method-

Complete Linkage is a hierarchical clustering method that connects clusters based on the maximum distance between any pair of data points from different clusters. It constructs a dendrogram, revealing relationships within the dataset. This method is effective for identifying clusters with compact and well-separated structures.

$$d_{(UV),W} = max\{d_{UV}, d_W\}$$

Dendrogram of Complete Linkage for the dataset is given below:



The 3 clusters formed are listed below:

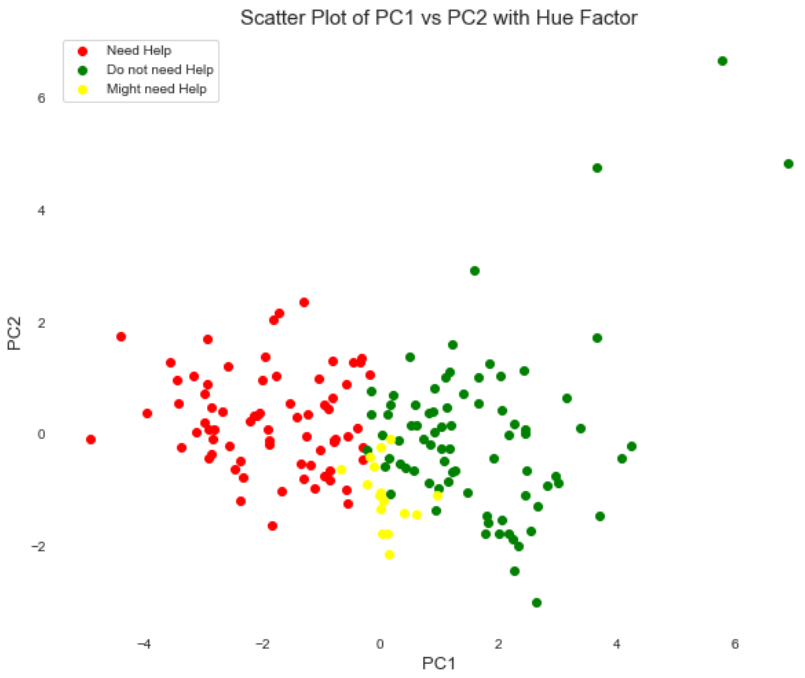
Cluster 0: Afghanistan, Algeria, Angola, Bangladesh, Benin, Bolivia, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Egypt, Equatorial Guinea, Eritrea, Fiji, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Guyana, Haiti, India, Indonesia, Iraq, Kazakhstan, Kenya, Kiribati, Kyrgyz Republic, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Micronesia, Fed. Sts., Mongolia, Mozambique, Myanmar, Namibia, Nepal, Niger, Nigeria, Pakistan, Philippines, Rwanda, Samoa, Senegal, Sierra Leone, Solomon Islands, South Africa, Sri Lanka, Sudan, Tajikistan, Tanzania, Timor-Leste, Togo, Tonga, Turkmenistan, Uganda, Uzbekistan, Vanuatu, Venezuela, Yemen, Zambia

Cluster 1: Albania, Antigua and Barbuda, Australia, Austria, Bahamas, Bahrain, Barbados, Belarus, Belgium, Belize, Bhutan, Bosnia and Herzegovina, Brunei, Bulgaria, Canada, Cape Verde, Costa Rica, Croatia, Cyprus, Czech Republic, Denmark, El Salvador, Estonia, Finland, France, Georgia, Germany, Greece, Grenada, Hungary, Iceland, Ireland, Israel, Italy, Japan, Jordan, Kuwait, Latvia, Lebanon, Libya, Lithuania, Luxembourg, Macedonia, FYR, Malaysia, Maldives, Malta, Mauritius, Moldova, Montenegro, Morocco, Netherlands, New Zealand, Norway, Oman, Panama, Paraguay, Poland, Portugal, Qatar, Romania, Russia, Saudi Arabia, Serbia, Seychelles, Singapore, Slovak Republic, Slovenia, South Korea, Spain, St. Vincent and the Grenadines, Sweden, Switzerland, Thailand, Tunisia, Ukraine, United Arab Emirates, United Kingdom, United States, Vietnam

Cluster 2: Argentina, Armenia, Azerbaijan, Brazil, Chile, China, Colombia, Dominican Republic, Ecuador, Guatemala, Iran, Jamaica, Peru, Suriname, Turkey, Uruguay

- As in Average linkage, most of the african and asian countries have been clustered together in Cluster 0. These countries need AID as most of them are underdeveloped, lack proper health care facilities and have less GDPP.
- Cluster 1 has super developed countries in the list (most of them) which are super secure and do not need further assistance.
- Cluster 2 is again a mix.

The scatter plot of PC1 v/s PC2 with hue factor is shown below:



=>Points to Infer:

- The plot clearly gives a picture on which countries need help and are depicted by green color(Cluster 1). They lie on the higher side of PC1 which indicate that their economic conditions are secure.
- The countries in Cluster 2 are depicted by red color and lie on the lower level of PC1 which indicate their economic levels are not great.

Silhouette score for complete linkage is 0.3409.

IV. Non Hierarchical Clustering: K-Means:

K-means is a clustering algorithm that partitions a dataset into K clusters, where each data point belongs to the cluster with the nearest mean. It iteratively refines cluster assignments until convergence, aiming to minimize the sum of squared distances between data points and their cluster centroids. K-means is widely used in various fields, including machine learning, image analysis, and customer segmentation. Its simplicity and efficiency make it a popular choice for clustering tasks, providing a practical approach to identifying coherent groups within datasets.

The three clusters formed using K-Means Algorithm are listed below:

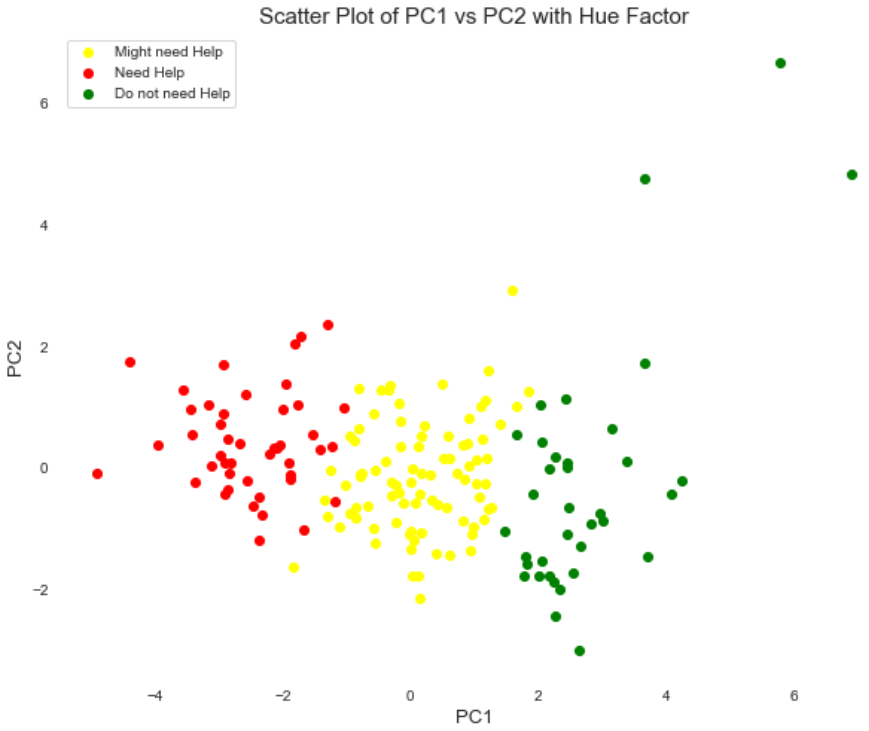
Cluster 0: Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bahamas, Bangladesh, Barbados, Belarus, Belize, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Cambodia, Cape Verde, Chile, China, Colombia, Costa Rica, Croatia, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Fiji, Georgia, Grenada, Guatemala, Guyana, Hungary, India, Indonesia, Iran, Iraq, Jamaica, Jordan, Kazakhstan, Kyrgyz Republic, Latvia, Lebanon, Libya, Lithuania, Macedonia, FYR, Malaysia, Maldives, Mauritius, Micronesia, Fed. Sts., Moldova, Mongolia, Montenegro, Morocco, Myanmar, Nepal, Oman, Panama, Paraguay, Peru, Philippines, Poland, Romania, Russia, Samoa, Saudi Arabia, Serbia, Seychelles, Solomon Islands, Sri Lanka, St. Vincent and the Grenadines, Suriname, Tajikistan, Thailand, Tonga, Tunisia, Turkey, Turkmenistan, Ukraine, Uruguay, Uzbekistan, Vanuatu, Venezuela, Vietnam

Cluster 1: Afghanistan, Angola, Benin, 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, Kenya, Kiribati, Lao, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Tanzania, Timor-Leste, Togo, Uganda, Yemen, Zambia

Cluster 2: Australia, Austria, Bahrain, Belgium, Brunei, Canada, Cyprus, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Qatar, Singapore, Slovak Republic, Slovenia, South Korea, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States

- Cluster 2 has developed countries which have great economic conditions.
- Cluster 1 has many African and Asian countries that have lesser GDPP, High mortality rates and etc.
- Cluster 0 has mixed countries which may need further help.

The Scatter plot of PC1 v/s PC2 with hue factor for K-Means is given below:



Points to Infer:

- The plot clearly gives a picture on which countries need help and are depicted by green color(Cluster 2). They lie on the higher side of PC1 which indicate that their economic conditions are secure.
- The countries in Cluster 1 are depicted by red color and lie on the lower level of PC1 which indicate their economic levels are not great.
- There are many countries that are lying in the yellow cluster that need further examination to judge their ability to receive AID.

Silhouette Score for K-Means is 0.2856.

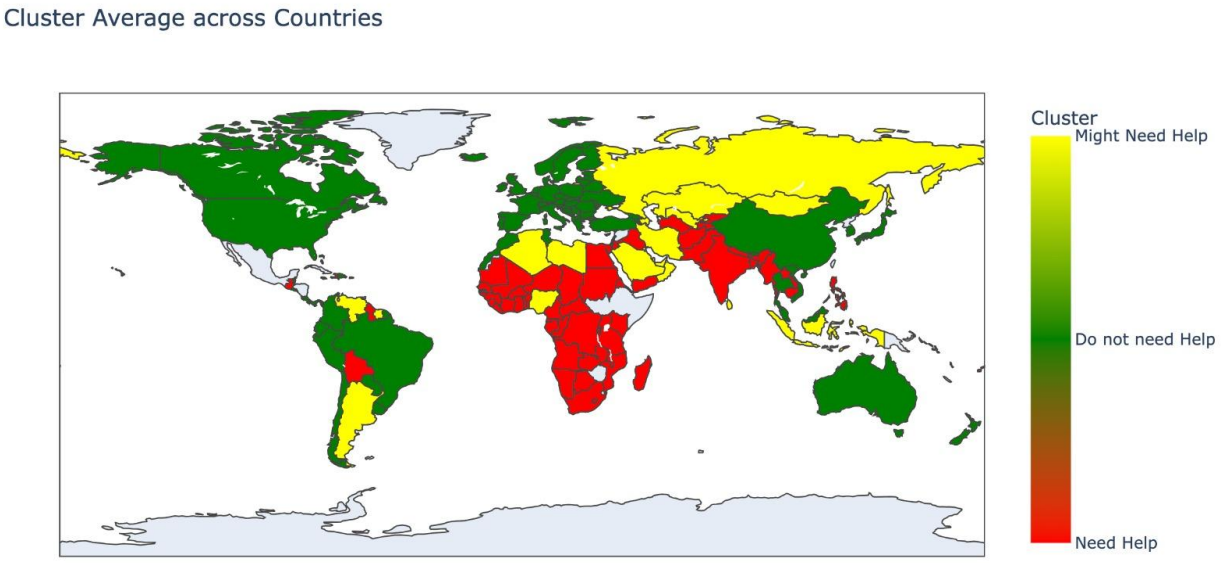
Conclusion: We select the algorithm with the highest silhouette score among the three performed.

Average Linkage Method yielded a maximum silhouette score of **0.413.**

A higher silhouette score is desired for clustering because it indicates better-defined and more distinct clusters within the data. Clusters with higher silhouette scores are easier to interpret because the boundaries

between clusters are more distinct. This facilitates a more straightforward understanding of the underlying patterns or groups present in the data.

Finally, the figure down below gives the world map with the countries coloured according to their need for AID from HELP International.



The countries with red colour highlighted are in need of AID provided by HELP International. The CEO can focus on these countries.