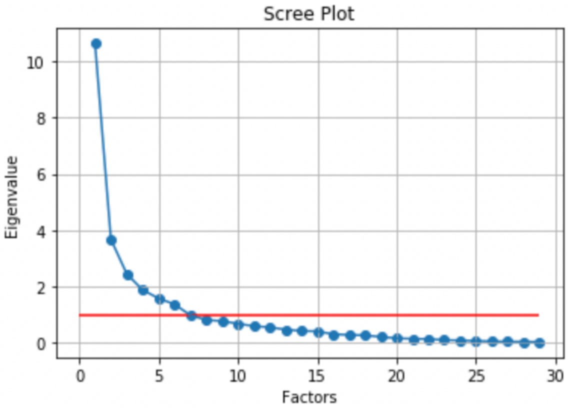
CLUSTERING ANALYSIS ON DIFFERENT COUNTRIES USING VARIOUS INDICATORS.

In this report, we are doing clustering analysis on different countries based on various indicators from the world bank data. First, we do some factor analysis to see what dimension these indicators vary and their representation. And then we do some clustering analysis based on different countries.

To avoid some potential bias caused by many indicators with significant difference, we have chosen growth indicators, instead of non-traditional ones.

By applying correlation matrix on all the indicators, we come up with the indicators which have positive relationship and negative correlation. In order to better understand the relationships between all the indicators, we have to apply factor analysis and reduce the number of indicators.

But the question still remains, how many factors to select? We can plot scree plot to determine the number of factors to select.



The scree plot has number of factors on x-axis and the eigen value in the y-axis. The scree pot above shows that there could be 7 factors.

When we take a look at the heat map of all factors, we basically understand that correlation for variable and the factors. As we know that the 7 factors are representing 7 different aspects of the economy. However, some indicators represent less correlation and some high.

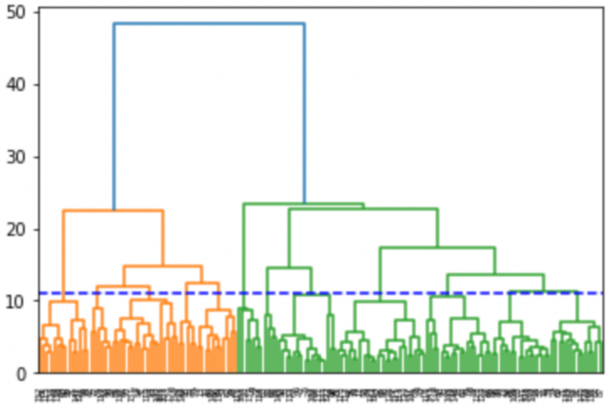
Now, we apply cluster analysis to classify the economies. We would apply most commonly used clustering method – hierarchical clustering with a bottom to top approach.

Before we apply hierarchical clustering, we have to standardize or normalize the indicators to the same scale. To prevent the scale difference between various indicators, we have to standardize the data.

After standardizing the data, we can apply clustering algorithms, by using the library, Agglomerative clustering.

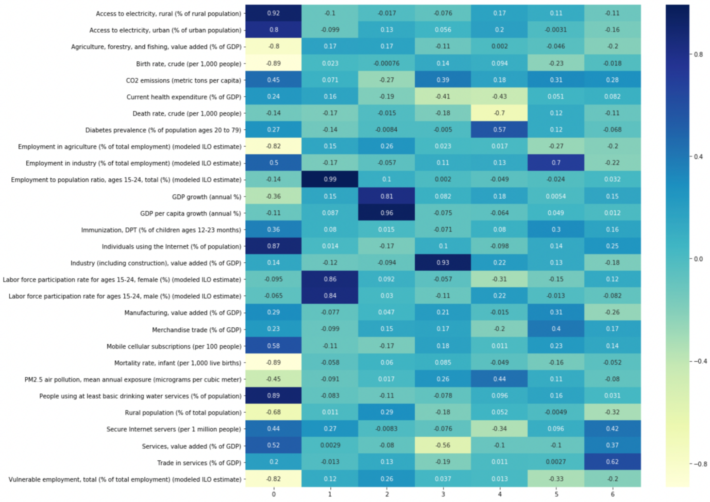
And to visualize the results of the clustering algorithm, we have to visualize it through the dendrogram, a tree like structure or diagram that represents a sequence of splits or merges.

The number of clusters finally formed is completely based on our requirement. The more the number of clusters, the more the information is detailed. If there are few clusters, then the economies may not be classified as per our judgment.



Based on the dendrogram above, there could be 12 clusters. Based on this, we can apply agglomerative clustering on the data set by 12 clusters and Euclidian distance and the linkage as ward's method.

After clustering many countries into 12 clusters, we perform a heat map, we observe that the 20 different variables have high correlation with only 7 factors as mentioned before.



	0	1	2	3	4	5	6
Access to electricity, rural (% of rural population)	0.92	-0.1	-0.017	-0.076	0.17	0.11	-0.11
Access to electricity, urban (% of urban population)	0.8	-0.099	0.13	0.056	0.2	-0.0031	-0.16
Agriculture, forestry, and fishing, value added (% of GDP)	-0.8	0.17	0.17	-0.11	0.002	-0.046	-0.2
Birth rate, crude (per 1,000 people)	-0.89	0.023	-0.00076	0.14	0.094	-0.23	-0.018
CO2 emissions (metric tons per capita)	0.45	0.071	-0.27	0.39	0.18	0.31	0.28
Current health expenditure (% of GDP)	0.24	0.16	-0.19	-0.41	-0.43	0.051	0.082
Death rate, crude (per 1,000 people)	-0.14	-0.17	-0.015	-0.18	-0.7	0.12	-0.11
Diabetes prevalence (% of population ages 20 to 79)	0.27	-0.14	-0.0084	-0.005	0.57	0.12	-0.068
Employment in agriculture (% of total employment) (modeled ILO estimate)	-0.82	0.15	0.26	0.023	0.017	-0.27	-0.2
Employment in industry (% of total employment) (modeled ILO estimate)	0.5	-0.17	-0.057	0.11	0.13	0.7	-0.22
Employment to population ratio, ages 15-24, total (%) (modeled ILO estimate)	-0.14	0.99	0.1	0.002	-0.049	-0.024	0.032
GDP growth (annual %)	-0.36	0.15	0.81	0.082	0.18	0.0054	0.15
GDP per capita growth (annual %)	-0.11	0.087	0.96	-0.075	-0.064	0.049	0.012
Immunization, DPT (% of children ages 12-23 months)	0.36	0.08	0.015	-0.071	0.08	0.3	0.16
Individuals using the Internet (% of population)	0.87	0.014	-0.17	0.1	-0.098	0.14	0.25
Industry (including construction), value added (% of GDP)	0.14	-0.12	-0.094	0.93	0.22	0.13	-0.18
Labor force participation rate for ages 15-24, female (%) (modeled ILO estimate)	-0.095	0.86	0.092	-0.057	-0.31	-0.15	0.12
Labor force participation rate for ages 15-24, male (%) (modeled ILO estimate)	-0.065	0.84	0.03	-0.11	0.22	-0.013	-0.082
Manufacturing, value added (% of GDP)	0.29	-0.077	0.047	0.21	-0.015	0.31	-0.26
Merchandise trade (% of GDP)	0.23	-0.099	0.15	0.17	-0.2	0.4	0.17
Mobile cellular subscriptions (per 100 people)	0.58	-0.11	-0.17	0.18	0.011	0.23	0.14
Mortality rate, infant (per 1,000 live births)	-0.89	-0.058	0.06	0.085	-0.049	-0.16	-0.052
PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)	-0.45	-0.091	0.017	0.26	0.44	0.11	-0.08
People using at least basic drinking water services (% of population)	0.89	-0.083	-0.11	-0.078	0.096	0.16	0.031
Rural population (% of total population)	-0.68	0.011	0.29	-0.18	0.052	-0.0049	-0.32
Secure Internet servers (per 1 million people)	0.44	0.27	-0.0083	-0.076	-0.34	0.096	0.42
Services, value added (% of GDP)	0.52	0.0029	-0.08	-0.56	-0.1	-0.1	0.37
Trade in services (% of GDP)	0.2	-0.013	0.13	-0.19	0.011	0.0027	0.62
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	-0.82	0.12	0.26	0.037	0.013	-0.33	-0.2

Based on clusters characteristics, we further divided the entire clusters into 4 big categories – Most develop countries, developing, less developed and undeveloped economies.

Under most developed countries we see – US, UK and France. Similarly, we see countries from clusters 3,4,10 are least developed countries in economy.

It would be interesting if we would compare our clustering and classification with the world's bank data of high, upper middle, lower middle and low-income levels.