Clustering of World Indicator Data

Project Report

DANA 4840-001

Lien Pham (100361334)

Mohammed Ghayaas (100364934)

Ayushi Singh (100359100)

Mary Ann Villamor (100365411)

Table of Contents

1. Introduction 3

1.1. Purpose of the Project.. 3

1.2. Dataset.. 3

1.3. General Description of Data 4

2. Methods 4

In this section we explain the rationale of the clustering method/algorithm along with....... 4

other analysis techniques for clustering analysis. 4

2.1. Characteristics of a good clustering 4

2.2. Basis for Choosing Optimal number of Clusters.................................. 4

2.3. Cluster Algorithm 5

3. Results... 5

3.1. Pre-processing 5

3.2. Clustering Algorithm 6

3.3. Optimal K and Linkage Function... 7

4. Discussions.. 7

4.1. Overall Clustering Results 7

4.2. Characteristics of Clusters 8

4.3. Limitation of the methodology 9

4.4. Expanding Analysis into 3 Clusters 9

4.5. Conclusion... 10

5. Future Research.. 10

6. Acknowledgements 10

7. APPENDIX.. i

7.1. List of Variables i

7.2. Plots and tables ii

1. Introduction

1.1. Purpose of the Project

The aim of this project is to apply clustering algorithms with other techniques that we learned in the Data Analytics Program (DANA4840-001).

1.2. Dataset

1.2.1. Background of Data

We used the 2010 World Bank Economic data for this analysis. The dataset comprises of the economic, education and health factors across 214 countries which is also publicly available at the World Bank website.

1.2.2. Reason for Selecting Data

World indicator has always been an interesting topic. Many of us have general idea on which country may be better that others in terms of economic, health and population factors. We often hear terminologies as "developed countries", "developing countries", "least developed countries", "first world countries", "third world countries", etc.

Through clustering methodology, it is interesting to know:

- if the countries will be grouped in line with general perception that there are 3 general categories of countries;
- if advanced countries (e.g., USA, UK, etc.) will be grouped into same cluster which are separated into the known least developed countries such as African region; and
- what are the characteristics that separate or combine these countries.

1.2.3. Aim of the analysis

Our target audience for this clustering analysis is World Bank who works in every major area of development. They provide a wide array of financial products and technical assistance to help countries to eradicate poverty and increase life's quality. Their specific goals are:

1. *Eradicate poverty and hunger*
2. Achieve universal primary education
3. Promote gender equality and empower women
4. *Reduce child mortality*
5. *Improve maternal health*
6. *Combat HIV/AIDS, malaria, and other diseases*
7. *Ensure environmental sustainability*

The aim of our analysis is to cluster the countries and detect characteristics that would help World Bank in achieving their above goals. To achieve this, we will assess:

a) Cluster that needs financial assistance by reviewing the following indicators:
   - Economy of countries (Goal # 1)
   - Quality of Life (Goal # 4, 5)
   - Health of people (Goal # 6)

b) Cluster that may need to be regulated in terms of environmental sustainability. For this area, we will review the CO2 emissions of the countries within the clusters.

1.3. General Description of Data

There were total of 34 variables used for this analysis, out of which 29 are numerical data. Please refer to the full details of the variables in Appendix 1. CO2 emissions per metric ton (derived from worldbank.org) for 2010 was added in the dataset to assess the environment sustainability factor.

1.3.1. Missing data

There were 1,383 or approximately 20% missing from the total data, from the following variables (refer to Appendix 7.2.1.1. for the bar plot):

1. ARI treatment (% of children under 5 taken to a health provider) – *88.79% missing*;
2. GINI Index – *83.64% missing*;
3. Income share held by lowest 20% - *83.64% missing*;
4. Central government debt, total (% of GDP) – *71.03% missing*; and
5. CPIA gender equality rating (1=low to 6=high) – *64.02% missing*.

Further, there were total of 34 (16%) countries with more than 30% missing data as depicted in Appendix 7.2.1.2.

1.3.2. Variable Redundancies

We noted the following categorical variables which may not be helpful for the analysis:

- Year and Year Code – The data was analyzed for year 2010 only. These variables have constant unit (i.e., 2010 or YR2010) across the observation countries.
- Country_me and Country Code – These data are unique for each observation and serve as identify rather than categories.

There were variables with high correlation (refer to Appendix 7.2.1.3).

2. Methods

In this section we explain the rationale of the clustering method/algorithm along with other analysis techniques for clustering analysis.

2.1. Characteristics of a good clustering

A good clustering will produce high quality clusters in which:

- the intra-class (that is, intra-cluster) similarity is high
- the inter-class similarity is low
- The measured quality of a clustering depends on both the document representation and the similarity measure used

2.2. Basis for Choosing Optimal number of Clusters

Internal criterion is used when we don’t have a ground of truth or expert knowledge, therefore, we use these scores to select the most relevant methods

- Average Silhouette Width

  The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It can be used to study the separation distance between the resulting clusters. The range of silhouette width is [-1, 1]. When value is close to 1, sample is well-clustered and already assigned to a very appropriate cluster. However, when value close to 0, sample could be assigned to another cluster closest to it and the sample lies equally far away from both the clusters. That means it indicates overlapping clusters. Silhouette width is a ratio to show how good a single observation is clustered.

- Calinski and Harabasz Score (CH Score)
  CH Score is the ratio of between-group error and within-group error with the penalty ratio when the number of k increases.

In addition, we also used agglomerative coefficient which measures the amount of clustering structure of the dataset. If observations quickly agglomerate into distinct clusters that later agglomerate into a single cluster at much greater dissimilarities, the coefficient will approach 1. In contrast, no clustering for the dataset will have coefficient approaching zero

When above scores are maximized, the clustering has the best performance under the given number of k.

2.3. Cluster Algorithm

We performed clustering methods:

a) Partitional such as K medoid, K means; and
b) Hierarchical (with various linkage function such as Single, Complete, Average and Ward.D2)

To determine the most relevant clustering technique, we assessed the quality and reliability (i.e., goodness of fit) of clustering results for our dataset, using CH score, silhouette score, and agglomerative coefficient.

3. Results

This section displays various analyses and descriptions, including pre-processing, clustering algorithm, and other analysis techniques

3.1. Pre-processing

3.1.1. Handling the Missing Data

Missing data reduces the statistical power of the analysis, which can distort the validity of the results. When dealing with missing data we can use two primary methods such as (a) imputation and (ii) removal of data.

The imputation method develops reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. However, when the portion of missing data is too high, the imputation may affect the effectiveness of the model due to the lack of natural variation. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. Decision on method to implement is based on the analysis why the data is missing.

Missingness pattern

Missing data is categorized as “Missing at Random”. Data missing is relative to the observed data and not specific to the missing values (refer to Appendix 7.2.2. for the plot of missingness pattern).

Removal of missing data

Variables with > 50% missing; and observations/countries with more than 30% missing data (refer to the list in section 1.3.1. and Appendix 7.2.1.) were removed. Imputation of these data would distort the analysis due to the high number of missing data (i.e., biasness). Upon removal of these data, we have 180 observations and 29 variables.

Imputation of missing data

Upon data cleaning, the remaining 71 missing data (approximately 2%). The data is not missing across all observations but only within sub-samples of the data. These missing data can be predicted based on the complete observed data.

Our dataset has multiple variables with high correlation and multicollinearity. Missing data was imputed using MissForest. MissForest is robust to noisy data and multicollinearity, since random-forests have built-in feature selection (evaluating entropy and information gain). KNN-Impute yields poor predictions when datasets have weak predictors or heavy correlation between features.

To ensure that the imputation does not create biasness in our dataset, we performed comparison of the statistical values (e.g., mean, median) before and after imputation. Please refer to Appendix 7.2.3.

3.1.2. Handling the Redundant Variables

We noted that the following variables with high correlation are subsets of other variables:

Redundant Variable	Other variable with overlapping information
1. Life expectancy at birth, total (years)	• Life expectancy at birth, female (years) and • Life expectancy at birth, male (years)
2. Mortality rate, infant (per 1,000 live births) 3. Mortality rate, under-5 (per 1,000 live births)	• Mortality rate, under-5, female (per 1,000) • Mortality rate, under-5, male (per 1,000)

Above variables were removed together with the 4 redundant variables mentioned in section 1.3.2

3.1.3. Outliers

Each observation/country characteristics may differ from each other. In this case, we are expecting to see a country which could be significantly higher or lower than others in terms of global development factors. To prove this scenario, we perform outlier check using Mahalanobis Distance which detected high-income countries (e.g., USA, Japan Singapore, China, etc.) and lower-income countries (i.e., Angola, Niger, etc.) as outliers. *To conclude, outlier test is not applicable for our dataset.*

3.2. Clustering Algorithm

3.2.1. PAM (K Medoids)

The PAM algorithm was implemented using the Euclidian distance dissimilarity matrix performed on the scaled dataset. The algorithm was iterated for 5 trails and for K values between 2 to 10. A maximum silhouette score of 0.3675 and CH score of 105.22 was observed with optimal K = 2 for our dataset.

3.2.2. K means:

Next, we implemented the K means algorithm on the scaled dataset and evaluated the clusters formed using the SW and CH score metrics for k varying between 2 and 10. With this method, a maximum SW score of 0.3652 and CH score of 106.86 for 2 optimal number of clusters was observed.

3.2.3. Agglomerative Hierarchical Clustering:

The hierarchical clustering was performed using the hclust() function from the “cluster” package on the Euclidean distance dissimilarity matrix obtained from the scaled dataset. The algorithm was iterated for all 4 linkage methods. We obtained a maximum Agglomerative coefficient for 'hclust' object of 0.9526 for the “Ward.D2” linkage, followed by complete, average, and single linkages with their coefficients being 0.8496, 0.6895, and 0.4988 respectively. The algorithm was also iterated for K values between 2 to 10 to determine the optimal number of groups in the dataset. “2” was found to be the optimal number with a maximum silhouette score of 0.3694 for the Ward.D2 linkage.

3.3. Optimal K and Linkage Function

According to our analysis, ward D2 has the highest CH score (164.3627) and Silhouette score (0.3694) with optimal k =2, compared to the kmeans, kmedoid, hierarchal clustering with single, average, complete linkages**.** Agglomerative coefficient (0.9527) of the ward D2 clustering seems to give a better structure, in comparison to the other clustering techniques

Refer to Appendix 7.2.4. for the calculation of above scores using k=2 to k=10 for multiple linkage functions.

4. Discussions

4.1. Overall Clustering Results

4.1.1. Optimal K and Appropriate Clustering Techniques

After iterating through K medoid, K means and hierarchical clustering methods, it is determined that k=2 is the best appropriate number for clusters for this dataset and ‘Ward.D2’ is the best linkage methods of hierarchical clustering. Using these best parameters thus determined, we perform the agglomerative clustering and cut the tree into 2 group of countries.

As the result of clustering, we have 2 clusters, cluster 1 consists of 45 countries (Afghanistan, Ethiopia, South Africa, South Sudan, Central African Republic, Nigeria, Rwanda, Yemen, Uganda) and group 2 consists of 106 countries (Australia, Oman, New Zealand, Denmark, Greece, Israel, and others)

According to the World Banks’s criteria, it is quite obvious that 45 countries in group 1 were ranked as the low-income (least developed) countries and 106 countries in group 2 were middle-income (developing) countries and high-income (developed) countries. Therefore, we conclude that the clustering result is reasonable.

4.1.2. Cluster Labels/Description

There is a clear distinction that:

- Cluster 1 countries are low-income countries; and
- Cluster 2 are middle- and high-income countries.

The most important variable for this grouping is “brate” or Birth rate, crude (per 1,000 people) with the importance value being 26.93 (using variable importance in R).

4.2. Characteristics of Clusters

We interpret the characteristics of these clusters by observing the averages of variables in the two groups using the following indicators:

- Economy of countries
- Quality of Life; and
- Health of people

4.2.1. Economy of Countries:

The economy of countries is represented by the variables such as GDP, GDP growth, GDP per capita, Health expenditure per capita, Health expenditure - public and Inflation. The table below compares the means of these variables:

	Variable	**Cluster 1 - Mean**	**Cluster 2 - Mean**
1	GDP	21,481 million	235,213 million
2	GDP growth	5.87	3.41
3	GDP per capita	2,204.46	14,098.43
4	Health expenditure per capita	83.54	1,145.92
5	Health expenditure - public	2.90	4.39
6	Inflation	9.25	4.95

From the above table, we can see that the developed countries represented by cluster 2 has more developed economy and health expenditures as compared to the underdeveloped countries represented by cluster 1. We can also observe that the inflation in underdeveloped countries is more than the developed countries. It is the norm that GDP growth in developed countries in slower than the GDP growth of the under developing countries and higher GDP per capita.

4.2.2. Quality of Life

The quality of life is observed by looking at Life expectancy, mortality rate and population growth. The average life expectancy of male and female in the underdeveloped countries is 57.1 and 59.7 respectively which is lower as opposed to that in cluster 2 with their averages being 71.63 and 77.38 respectively.

The mortality or the death rate in developed countries for both males and females is very less than the underdeveloped countries with the values being 19.47 and 15.98 in cluster 2 against 92.15 and 81.03 in cluster 1.
The average population growth of the developed group is 0.94 whereas that for undeveloped group is 2.58. Therefore, the population is also under control in the developed countries.

4.2.3. Health of People:

We observe better indicators of health of people in developed group than the underdeveloped group. This can be seen by the lesser average birth rate of 16.90 in group 2 than 36.40 in group 1, higher immunization rate against diseases like DPT and measles with average values of 93.75 and 93.37 in the developed group as opposed to lesser immunization rates of 77.88 and 74.89 in the underdeveloped group, and lesser women's share of population living with HIV with an average of 29.88 as opposed to 55.04 in group 1.

4.3. Limitation of the methodology

There are some limitations of hierarchal clustering

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and irregular shapes
  - Breaking large clusters
- In clustering, clusters are inferred from the data without human input (unsupervised learning)

However, in practice, it's a bit less clear: there are many ways of influencing the outcome of clustering: number of clusters, similarity measure, representation of documents and others

4.4. Expanding Analysis into 3 Clusters

To examine the larger group of 106 countries, we further divide this group into two by performing the agglomerative clustering with ward.D2 linkage and k = 3. We thus obtain 3 clusters of sizes 45, 61 and 45 respectively.

4.4.1. Country Grouping using 3 Clusters

Some typical countries in the clusters when K = 3 are:

- Cluster 1: Afghanistan, Ethiopia, Burki Faso, South Africa, South Sudan, Central African Republic, Nigeria, Rwanda, Yemen, Uganda etc.
- Cluster 2: Thailand, Turkey, Vietnam, Nepal, Iran, Mexico, Malaysia, Bhutan, Bangladesh etc.
- Cluster 3: Russia, Australia, Oman, New Zealand, Denmark, Greece, Israel, Germany, Sweden, United Kingdom etc.

From above, we can observe that 45 underdeveloped countries earlier in cluster 1 (when k =2) remained the same in cluster 1 when k =3. The 106 developed countries earlier represented by cluster 2 were divided into two clusters of 61 and 45 respectively where cluster 2 represented slightly lesser developed and cluster 3 represents the highly developed / rich countries / superpowers.

4.4.2. Characteristics of each Clusters (Refer to Box Plot in Appendix 7.2.5)

Economy, Quality of Life and Health

We can observe similar trends in the economy, quality of life and people's health in cluster 2 and cluster 3 (That is lesser developed and fully developed countries) when K = 3 as we observed earlier with K = 2. The highly developed countries of cluster 3 have higher average GDP, Per capita GDP, health expenditure per capita and lower inflation than the lesser developed countries of cluster 2.

Industrialization / CO2 Emission

The development of a country is majorly governed by the industries thriving in that country. On one hand, industrialization brings about employment to the citizens of a country, contribute to the income, economy, and GDP of a country, but on the other hand, industrialization releases adverse biproducts that negatively affect the environment overall. The quantity of Co2 emission and checking on the same becomes of interest while we are looking into the characteristics of the country

In our analysis with two clusters, the average CO2 emission by the developed countries was far greater than that of the underdeveloped group with their averages being 4.96 as opposed to 0.71 MT per capita respectively. When these countries are divided into three clusters, we can see a similar trend of relatively higher CO2 emission as the countries are developing that is the most developed countries contribute to the high CO2 emission to the environment.

Therefore, this factor becomes of interest. As the world bank on one side is helping countries towards development, it becomes necessary to curb the adverse effects of development, like the CO2 emission to the environment.

4.5. Conclusion

According to the World Bank's indicator report, the results of cluttered countries align with the WB's criteria to access the lower income, middle income, and high-income countries. However, we will need to look deeper into the cluster 2 in which some countries such as Bulgaria, Polland, Hungary, Polland, and Romania are clustered into group 3, which were high – income countries

We observed the adverse effect of industrialization and a country's development on the environmental sustainability. The World Bank shall undertake efforts to check and curb the high CO2 emission from developed countries alongside the other goals.

5. Future Research

- Look for more variables such as criminal rate, clean water quality access rate, literacy rate to support World Bank's goals/decision making if there is a specific goal
- Perform analysis in cluster 2 to detect lower middle-income countries
- Try the clustering on the most recent dataset (2021) to detect the changes in clustering, trends, and patterns
- Cophenetic correlation coefficient is another way to evaluate how well the cluster hierarchy represents the original object-by-object dissimilarity space

6. Acknowledgements

References used in this report are as follows:

a) The World Bank - https://data.worldbank.org/ - Source for the following information:
   - Global CO2 emission data https://data.worldbank.org/CO2indicator
   - 2010 World Development Indicator Report

b) DANA 4840 Notes
   - Chapter 2: Partitioning Clustering
   - Chapter 3: Hierarchical Clustering

c) How to Deal with Missing Data, Masters in Data Science (https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data)

d) Practical Guide To Cluster Analysis in R sthda.com Edition 1 Unsupervised Machine Learning - Alboukadel Kassambara.

7. APPENDIX

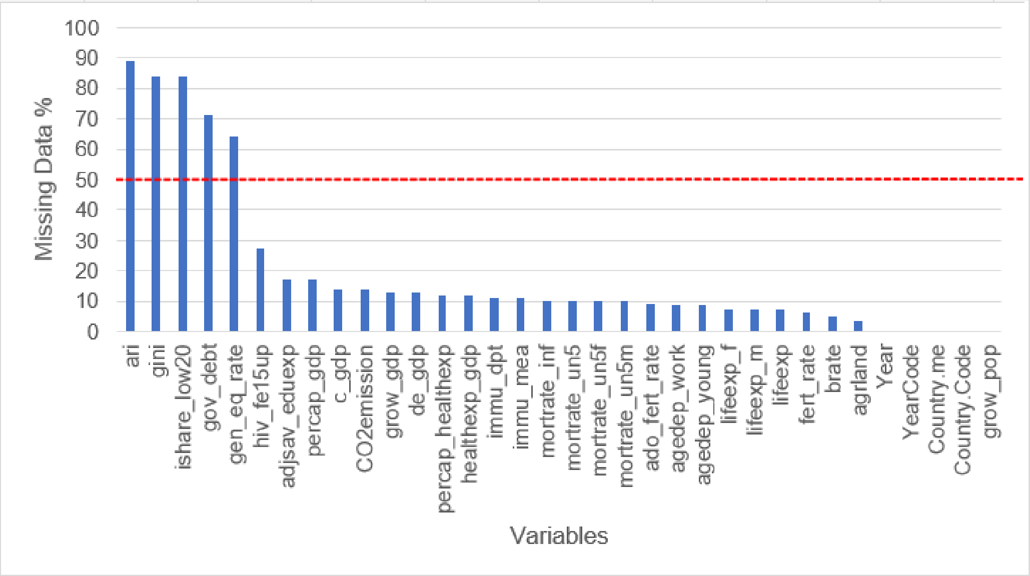
7.1. List of Variables

	Variables	Type	Description
1	Year	Char	Year
2	YearCode	Char	YearCode
3	Country me	Char	Country Name
4	Country Code	Char	Country Code
5	ari	Num	ARI treatment (% of children under 5 taken to a health provider)
6	adjsav_eduexp	Num	Adjusted savings: education expenditure (% of GNI)
7	ado_fert_rate	Num	Adolescent fertility rate (births per 1,000 women ages 15-19)
8	agedep_work	Num	Age dependency ratio (% of working-age population)
9	agedep_young	Num	Age dependency ratio, young (% of working-age population)
10	agrland	Num	Agricultural land (% of land area)
11	brate	Num	Birth rate, crude (per 1,000 people)
12	gen_eq_rate	Num	CPIA gender equality rating (1=low to 6=high)
13	gov_debt	Num	Central government debt, total (% of GDP)
14	fert_rate	Num	Fertility rate, total (births per woman)
15	c_gdp	Num	GDP (constant 2005 US$)
16	grow_gdp	Num	GDP growth (annual %)
17	percap_gdp	Num	GDP per capita, PPP (constant 2005 international $)
18	gini	Num	GINI index
19	percap_healthexp	Num	Health expenditure per capita (current US$)
20	healthexp_gdp	Num	Health expenditure, public (% of GDP)
21	ishare_low20	Num	Income share held by lowest 20%
22	de_gdp	Num	Inflation, GDP deflator (annual %)
23	lifeexp_f	Num	Life expectancy at birth, female (years)
24	lifeexp_m	Num	Life expectancy at birth, male (years)
25	lifeexp	Num	Life expectancy at birth, total (years)
26	mortrate_inf	Num	Mortality rate, infant (per 1,000 live births)
27	mortrate_un5	Num	Mortality rate, under-5 (per 1,000 live births)
28	mortrate_un5f	Num	Mortality rate, under-5, female (per 1,000)
29	mortrate_un5m	Num	Mortality rate, under-5, male (per 1,000)
30	grow_pop	Num	Population growth (annual %)
31	immu_dpt	Num	Immunization, DPT (% of children ages 12-23 months)
32	immu_mea	Num	Immunization, measles (% of children ages 12-23 months)
33	hiv_fe15up	Num	Women's share of population ages 15+ living with HIV (%)
34	CO2	Num	CO2 emission per metric ton

7.2. Plots and tables

7.2.1. Missing Data

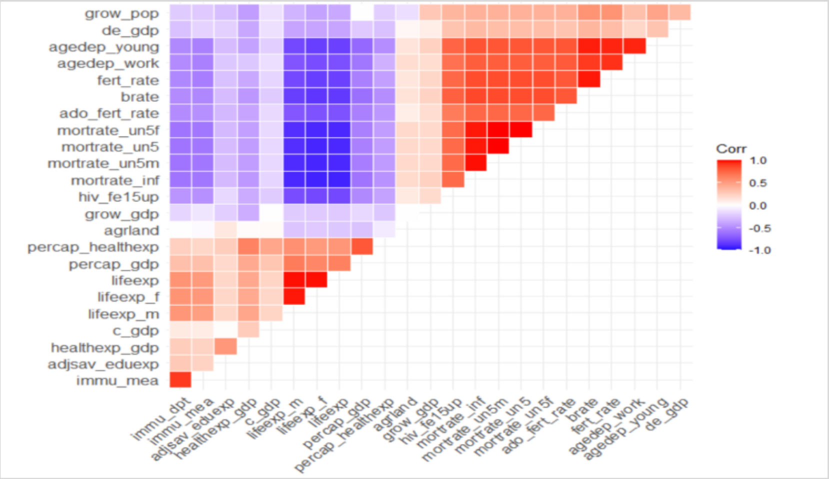
7.2.1.1. Rate of missing data per variable



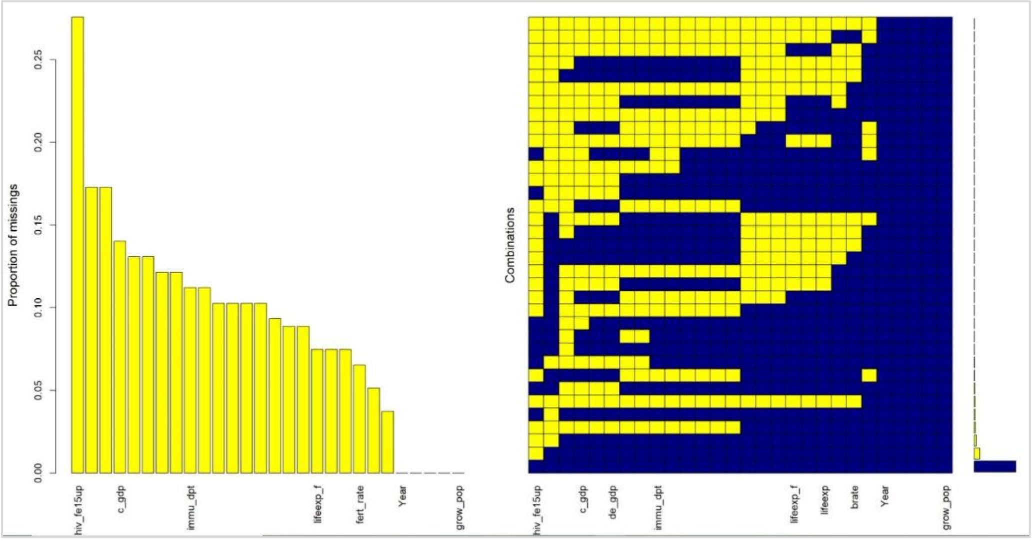
7.2.1.2. Rate of missing data per country

Range	No of Countries	Rate %
>=80%	5	2.34%
>=60% < 80%	6	2.80%
>=50% <60%	11	5.14%
>=30% <50%	12	5.61%
>=20% <30%	11	5.14%
> 20%	169	78.97%
Total	**214**	**100.00%**

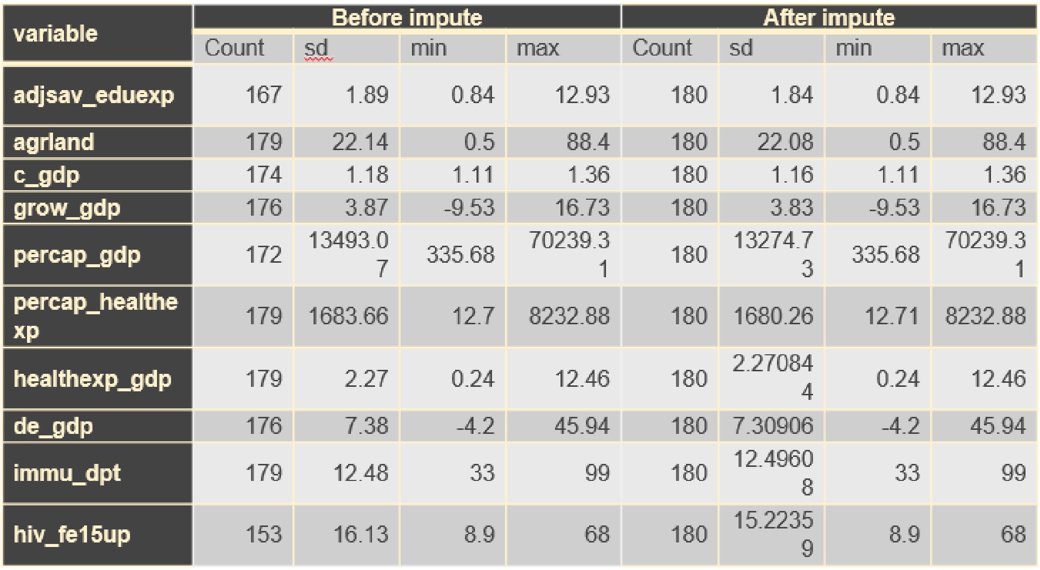
7.2.1.3. Correlation Matrix



7.2.2. Missingness Pattern



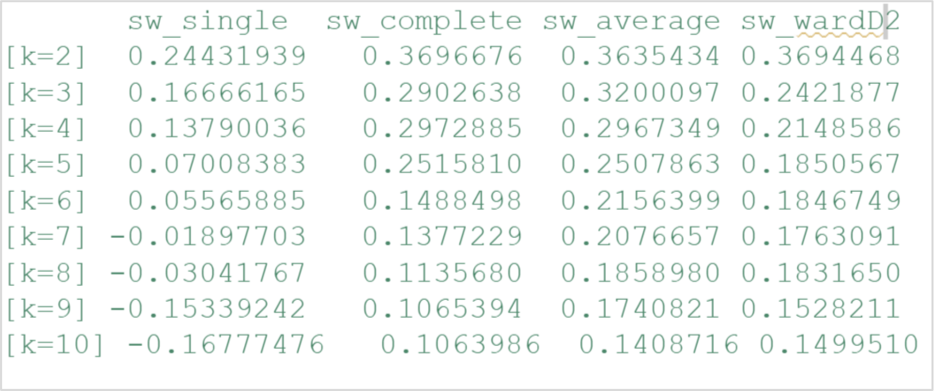
7.2.3. Statistics before and after imputation.



variable	Before impute				After impute			
	Count	sd	min	max	Count	sd	min	max
adjsav_eduexp	167	1.89	0.84	12.93	180	1.84	0.84	12.93
agrland	179	22.14	0.5	88.4	180	22.08	0.5	88.4
c_gdp	174	1.18	1.11	1.36	180	1.16	1.11	1.36
grow_gdp	176	3.87	-9.53	16.73	180	3.83	-9.53	16.73
percap_gdp	172	13493.07	335.68	70239.31	180	13274.73	335.68	70239.31
percap_healthexp	179	1683.66	12.7	8232.88	180	1680.26	12.71	8232.88
healthexp_gdp	179	2.27	0.24	12.46	180	2.270844	0.24	12.46
de_gdp	176	7.38	-4.2	45.94	180	7.30906	-4.2	45.94
immu_dpt	179	12.48	33	99	180	12.49608	33	99
hiv_fe15up	153	16.13	8.9	68	180	15.22359	8.9	68

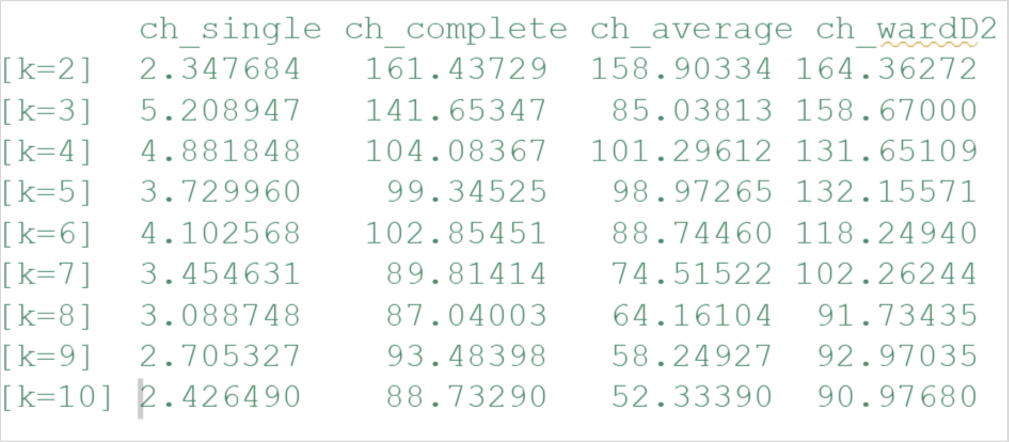
7.2.4. Calculating Optimal K and Linkage Function

7.2.4.1. Silhouette Score



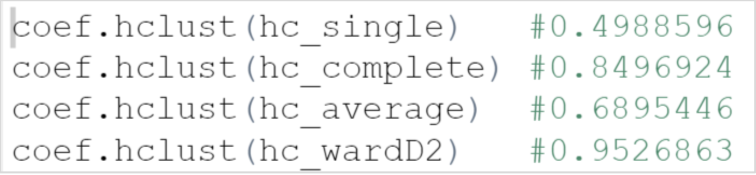
```
        sw_single   sw_complete sw_average sw_wardD2
[k=2]   0.24431939    0.3696676  0.3635434 0.3694468
[k=3]   0.16666165    0.2902638  0.3200097 0.2421877
[k=4]   0.13790036    0.2972885  0.2967349 0.2148586
[k=5]   0.07008383    0.2515810  0.2507863 0.1850567
[k=6]   0.05565885    0.1488498  0.2156399 0.1846749
[k=7]  -0.01897703    0.1377229  0.2076657 0.1763091
[k=8]  -0.03041767    0.1135680  0.1858980 0.1831650
[k=9]  -0.15339242    0.1065394  0.1740821 0.1528211
[k=10] -0.16777476    0.1063986  0.1408716 0.1499510
```

7.2.4.2. Silhouette Score, CH Score and Agglomerative Coefficient



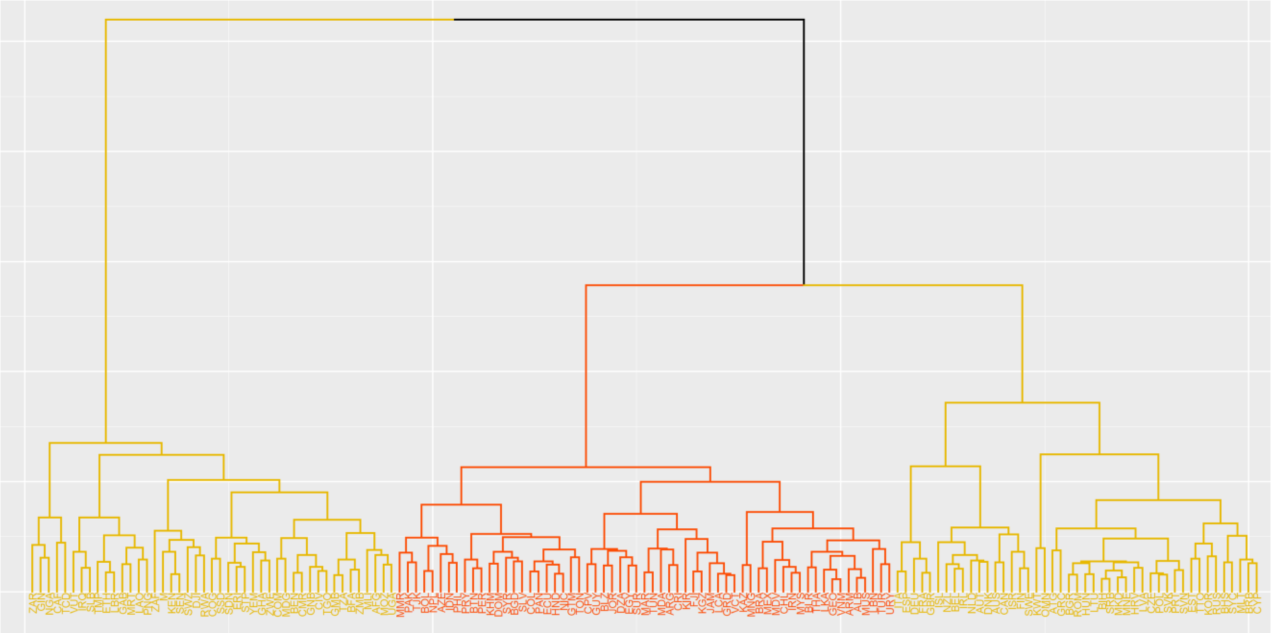
```
        ch_single ch_complete ch_average ch_wardD2
[k=2]   2.347684   161.43729  158.90334 164.36272
[k=3]   5.208947   141.65347   85.03813 158.67000
[k=4]   4.881848   104.08367  101.29612 131.65109
[k=5]   3.729960    99.34525   98.97265 132.15571
[k=6]   4.102568   102.85451   88.74460 118.24940
[k=7]   3.454631    89.81414   74.51522 102.26244
[k=8]   3.088748    87.04003   64.16104  91.73435
[k=9]   2.705327    93.48398   58.24927  92.97035
[k=10]  2.426490    88.73290   52.33390  90.97680
```

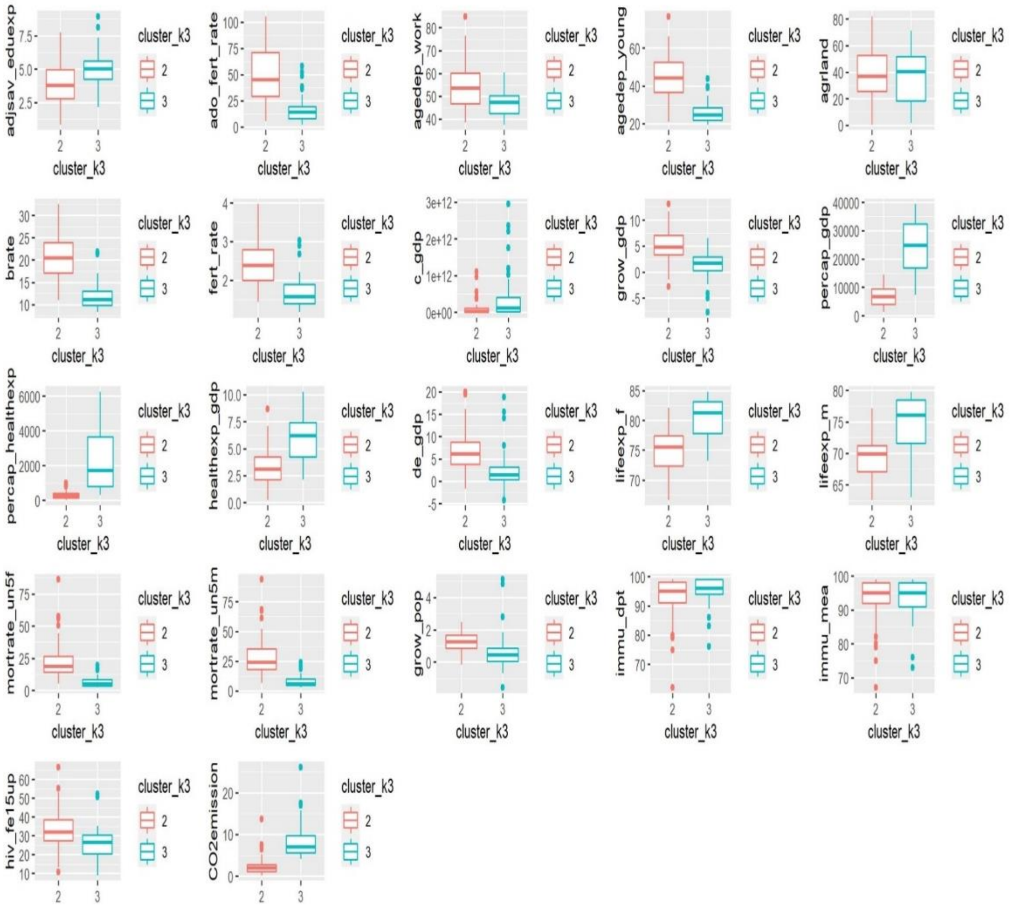
7.2.4.3. Silhouette Score, CH Score and Agglomerative Coefficient



```
coef.hclust(hc_single)    #0.4988596
coef.hclust(hc_complete)  #0.8496924
coef.hclust(hc_average)   #0.6895446
coef.hclust(hc_wardD2)    #0.9526863
```

7.2.5. Box Plot Comparison of Cluster 2 and Cluster 3




adjsav_eduexp
ado_fert_rate
agedep_work
agedep_young
agrland
brate
fert_rate
c_gdp
grow_gdp
percap_gdp
percap_healthexp
healthexp_gdp
de_gdp
lifeexp_f
lifeexp_m
mortrate_un5f
mortrate_un5m
grow_pop
immu_dpt
immu_mea
hiv_fe15up
CO2emission
cluster_k3
2
3