WORLD BANK COUNTRY DATA ANALYSIS

Date 14-05-2018

TERESA RIEDL

A) Introduction

The following report will explore world bank country data. It allows for comparison of countries and answers to the question how different variables such as GDP per capita, energy consumption, mortality rates and various others are linked.

Firstly, it is about understanding the data set, its structure and its gaps. For further processing we identify missing values and either remove the whole row or choose an appropriate approximation as a replacement for the missing value. In order to perform the algorithms, the data needs to be scaled and centred around zero.

We are using two different approaches: Principle Component Analysis (PCA) and Clustering. PCA looks to find a low-dimensional representation of the observations that explain a good fraction of their variance and Clustering looks to find homogenous subgroups among the observations.

We use PCA on both the original unscaled data as well as on the standardized data. The original data doesn't return satisfying results. PCA on the standardized data helps to reduce the number of variables from 19 to two and thereby explains 62.5% of the variance in the data.

Afterwards different clustering methods help us to segment the data. We differentiate between partitioning and hierarchical clustering. In partitioning clustering, we use K-means as well as K-Medoid.

In K-Means each cluster is represented by the centre or the means of the data points belonging to the cluster. We try different methods to choose the appropriate number of cluster and analyse the results for both k=2 and k = 3 clusters.

K-Medoids uses one observation in each cluster for which the average dissimilarities towards the other values in the cluster are minimal. We achieve a clustering with k=2 clusters and an average silhouette width of 0.55 which is a reasonable value.

Hierarchical clustering is a very visual method. Values are clustered pairwise in form of a tree model the so-called dendrogram which is built bottom up. Again, we choose to use three clusters after trying different methods of linkage and get to a valid result with a value of 0.81 for the cophenetic distance.

B) Data Exploration

Originally the country data has 192 observations and 19 variables. The first processing to be done is to name the variables with key-words to make sure that we can easily interpret the data at a later stage. Then we identify the fields with missing values which are 63 in total. We remove rows with more than three missing values and use the Country Classification Data from World Bank to estimate the remaining 21 missing values according to their country groups.

We now have the data "original" in a clean format that we can use for the first PCA analysis. But for later use we are looking at the boxplot of the data and observe that the scales of the variables vary heavily. We have to scale the data and centre it around zero for further processing. Scaling means that we divide the column entries by their column standard deviations. We call the data "variables" and boxplot (Figure 1) the data to make sure the pre-processing was successful.

C) Analysis

1. Principal Component Analysis

Principle component methods are used to summarize and visualize the information contained in multi-variable data sets. Depending on whether the data is quantitative, qualitative of mixed structured there are different Principle Component methods. In this case where are quantitative data set is provided the Principle Component Analysis is the appropriate tool. It allows to extract important information from the multi-variable data and to express this information as a set of few new variables called principle components. PCs correspond to a linear combination of the original variables. The goal of PCA is to identify directions (or principle components) along which the variation in the data is maximal. Thereby PCA reduces the dimensionality to a smaller number of PCs that can be visualized with only minimal loss of information.

PCA assumes that the directions with the largest variances are the most "Important" (i.e. the most principal). The amount of variance retained by each principle component is measured by the so-called eigenvalue. Therefore, PCA is particularly useful when variables within the data set are highly correlated which indicates that there is redundancy in the data. Then the number of variables can be reduced by using a smaller number of new variables – the principle components.

PCA can help to identify hidden pattern in the data, to reduce its dimensionality by removing noise and redundancy and to identify correlated variables.

PCA on the original data

When we perform PCA on the original data the variable with the highest variance - which is without doubt "GDP per capita" – describes very close to 100% of the variance in the data. Looking at the graph (Figure 2) we would not learn anything about other correlations in the data because the scale is so much larger compared to the rest of the variables. The result leaves us with using only one Principal Component. Using un-standardized data only helps us to identify the variables with the largest scale. The Scree Plot (Figure 3) shows that the percentage of explained variance does not increase when adding more PCs.

PCA on the standardized data

We now use the standardized data hoping to be able to identify correlations, redundancies and to reduce the dimensionality of the data. This time the graph for the first two dimensions Figure 4 is more insightful. Biplots represent the PC loading vectors which represent their direction in the feature space along with the scores which are the projections along the directions. The Biplot shows that several variables point into similar directions and are therefore correlated. Strong correlations we can assume for example for "Imports" and "Foreign Investments" as well as for the different Mortality Rates. Those correlations make sense but we can also observe correlations that we might not have expected before such as "Life Expectancy" and "Electricity Access".

PCs provide a low-dimensional linear surface that are closest to the observations. More precisely, the first PC loading vector is the single dimension that lies closest to all observations since such line provides a good summary of the data. Looking at the PCA variables we can observe that most features are highly represented in Dimension / PC number 1. In higher dimensions features are only barely represented.

As a next step we need to define how many PCs to use. The obvious goal is to reduce the dimensionality of the data without loosing too much information contained in the data. The Screeplot in Figure 5 helps to identify a point at which the proportion of variance explained by each subsequent PC drops off. This drop is called an elbow in the curve. Based on the scree plot we can tell that with only 2 PCs we can already explain a decent amount of the information. As we have very tangible data we are having a closer look at the PCs in Figure 6 to see if they are workable for further processing or if we rather keep working with the 19 variables from the original data. As the first PC contains most of the information it is hard to find a clear headline for the variable whereas the second PC would work well as something along the lines of "International Trade Relations".

Another method that helps to decide whether we want to keep working with the PCs or the original variables is the proportion of variance explained. Figure 7 shows the plot of the cumulative proportion of variance explained. If we were to choose only two PCs we would end up explaining only 62.5% of the variance in the data. Therefore, we continue using all 19 variables for the cluster analysis.

2. Cluster Analysis

Clustering methods allow us to partition data into distinct groups so that observations within each group are similar and observations in different groups are different from each other.

Distance measures

To decide whether two observations are similar or dissimilar to each other, we use distance-based approaches. The choice of distance measures is a critical step in clustering. Usually the default distance measure is the Euclidean distance which returns pairwise distances and clusters observations with high/low values of features together. Correlation-based distance considers two observations to be similar if their features are highly correlated, the distance would be zero if two observations are perfectly correlated. Popular correlation-based methods are Pearson correlation distance which measures the degree of a linear relationship between two variables and Spearman correlation distance which calculated the inter-rank distance of two variables.

Literature suggests using correlation-based methods in order to cluster observations with the same overall profiles regardless of their magnitudes. In this report we are looking to identify clusters of countries with similar profiles but don't put them into relation of their size and therefore we if helps to ignore the magnitude and use a correlation-based method. As Pearson is the most commonly used method we use it for the first iteration of the cluster analysis and consider other distance measures if the results are not satisfying.

Partitioning Cluster Analysis

K-Means

During lectures we looked at the centroid partitional clustering method K-Means where each cluster is represented by the centre or means of the data points belonging to the cluster. K-Means is a method sensitive to outliers. The quality of the clusters is defined by their intra class similarity and inter-class dissimilarity.

Probably the most critical step for k-means clustering is to choose the number of clusters you want to compute. This needs to be defined before starting the analysis. We used several different methods in order to select an appropriate number of clusters for the K-Means Clustering:

1. Gap statistic: Compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.
2. Within cluster sums of squares: The direct method where we choose a number of clusters so that adding another cluster doesn't improve the total within cluster sum of squares.

3. Silhouette: The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k.
4. Majority rule: The NbClust function allows counts the results of all methods and chooses the k with the majority of votes.

Figure 8 shows that between all methods both two and three clusters get the same amounts of votes which is eight. We have a look at both K-Means results. For two clusters Figure 9 shows the result and the kmeans function returns that the allocation in two clusters explained 38.6% of the variance. To get a more specific outcome we are looking at the result for three clusters and plot the result in Figure 10. The allocation explains 48.7% of the variance in the data.

When we look closer at the cluster means of the result with two clusters in Table 1, we see that cluster 1 has very high values when it comes to Mortality, Birth and Fertility rates. Despite the low value for life Expectancy the population is still vastly growing compared to cluster 2. The population is mainly living in rural areas and the access to Electricity as well as the distribution of Internet and Mobile subscriptions are very low. The only positive effect thereof is that the CO2 Emission is low but so it the GDP per capita. International trade relations (Foreign Investments, Import and Export) are lower than those for cluster 2. Another difference between the clusters is the inflation rate which is very low in cluster 2 and high in cluster 1. In conclusion one could say that cluster 1 represents underdeveloped countries with a lack in infrastructure development and cluster 2 represents developed countries great infrastructure but shrinking populations.

Table 1

	Foreign Investment	Electricity Access	Renewable Energy	CO2 Emission	Inflation	Mobile Subscr.	Internet Use	Exports	Imports	GDP
1	-0.09	-1.23	1.02	-0.64	0.36	-0.99	-0.97	-0.42	-0.13	-0.61
2	0.05	0.63	-0.52	0.33	-0.18	0.51	0.50	0.21	0.07	0.31
	Mortality Male	**Mortality Female**	**Birth Rate**	**Death Rate**	**Mortality Infant**	**Life Expect.**	**Fertility Rate**	**Populat. Growth**	**Urban Populat.**	**#**
1	1.13	0.94	1.13	0.47	1.20	-1.16	1.11	0.55	-0.84	**62**
2	-0.58	-0.48	-0.58	-0.24	-0.62	0.60	-0.57	-0.28	0.43	**121**

Table 2 shows the variable means for three clusters. Cluster 3 has a similar but extremer profile as cluster 1 in the analysis for 2 clusters. The population is growing significantly, people are mainly living in the city and the GDP is very low whilst the inflation is high. Zimbabwe is a well-known representative of cluster 3 in this classification even South Africa would fall in this group of under-developed countries with poor infrastructure. Cluster 1 represents a group of medium-developed countries like Russia, Turkey, Venezuela along with 81 others from all over the world. It sticks out that they made significant improvements in infrastructure development and have therefore much lower mortality and birth rates. At the same time their population is decreasing which could be due to the poor economy and low foreign investments. It sticks out that most regimes with communist tendencies are placed in cluster 1 (China, Cuba, Russia, Venezuela, etc.) along with other countries. Lastly, the 42 countries in cluster 2 represent developed countries like the Unites

States, Japan or France. Their infrastructure has the highest standards in comparison with the other two clusters and their GDP is outstanding. Due to the low birth rates the population is not shrinking which indicated that they are countries of immigration (as opposed to cluster 1).

Table 2

	Foreign Investment	Electricity Access	Renewable Energy	CO2 Emission	Inflation	Mobile Subscr.	Internet Use	Exports	Imports	GDP
1	-0.12	0.53	-0.39	-0.15	0.12	0.28	-0.06	-0.13	-0.11	-0.34
2	0.34	0.71	-0.68	1.16	-0.69	0.82	1.43	0.81	0.35	1.50
3	-0.07	-1.31	1.08	-0.64	0.33	-1.02	-0.97	-0.40	-0.10	-0.61
	Mortality Male	**Mortality Female**	**Birth Rate**	**Death Rate**	**Mortality Infant**	**Life Expect.**	**Fertility Rate**	**Populat. Growth**	**Urban Populat.**	**#**
1	-0.41	-0.20	-0.38	-0.20	-0.40	0.32	-0.44	-0.41	0.03	**84**
2	-0.87	-1.01	-0.88	-0.37	-0.90	1.06	-0.75	0.03	1.05	**42**
3	1.24	1.04	1.21	0.56	1.26	-1.25	1.20	0.59	-0.81	**57**

K-Medoids

Another partitioning clustering method is the K-Medoids algorithm which is related to the K-Means method. The difference is that in K-medoids clustering each cluster is represented by one of the observations in the cluster, the so-called medoid. Medoid describes the member of a cluster that is most centrally located and therefore its average dissimilarity towards the other points in the cluster is minimal.

Compared to K-Means the algorithm is less sensitive to noise and outliers, but we still need to choose the number of k in advance. Average Silhouette Width is the most commonly used method to choose the number of clusters for K-Medoids. The idea is to use different values for k and then the average cluster silhouette is drawn according to the number of clusters. Figure 11 shows that the optimal number of clusters is two for the country data.

Figure 12 shows the clustering. It looks similar to the K-Means clustering with k=2 clusters. Cluster 1 has Korea as its medoid and has 103 observations. Cluster 2 is centred around Malawi and is shared amongst 80 countries. Table 3 shows the medoid variables. Korea represents a cluster with well-established infrastructure and therefore a better life expectancy and coverage of electricity and internet. Malawi represents the opposite, a growing population living in rural areas and a poor internet and electricity coverage. Again, we could classify the two countries into under-developed and developed countries.

Table 3

	Foreign Investment	Electricity Access	Renewable Energy	CO2 Emission	Inflation	Mobile Subscr.	Internet Use	Exports	Imports	GDP
KOR	-0.45	0.72	-1.06	1.01	-0.41	0.32	1.88	0.26	-0.04	0.51
MWI	-0.40	-2.24	1.56	-0.73	0.71	-1.63	-1.11	-0.65	-0.47	-0.68
	Mortality Male	**Mortality Female**	**Birth Rate**	**Death Rate**	**Mortality Infant**	**Life Expect.**	**Fertility Rate**	**Populat. Growth**	**Urban Populat.**	**#**
KOR	-1.00	-1.00	-1.21	-1.00	-0.98	1.16	-1.15	-0.68	1.12	**103**
MWI	1.86	2.14	1.64	0.57	1.21	-1.50	1.59	0.95	-1.71	**80**

To evaluate the result, we have a look at the average silhouette width. The value of 0.55 means that a reasonable structure has been found in the data.

For this K-Medoids clustering we used the pam function from the “cluster” package. For larger applications there is the clara function in the “CLARA” package which considers a small sample of the data with fixed size and then applies the pam algorithm.

Hierarchical Cluster Analysis

As opposed to partitioning clustering this method does not require a pre-defined number of clusters k. Hierarchical clustering can be approached either bottom-up (Agglomerative) or top-down (Divisive). Agglomerative Clustering is by far the most common method and therefore will be used on the countries data set.

In Agglomerative Clustering each object is initially considered as a single-element cluster or leaf. At each step of the algorithm, the two clusters that are most similar are combined into a new bigger cluster or so-called node. This process will be repeated until all points are member of just one single big cluster (root). The visualization is a tree like graph called dendrogram.

In order to group the observations, the linkage function takes the distance information (in our case Pearson correlation distance) and groups pairs of observations into clusters based on their similarity. The newly formed clusters are then linked to each other to create bigger clusters until all objects are linked together in the hierarchical tree. There are single and centroid linkage, but the following three linkage methods are the most commonly used ones:

1. Maximum or complete linkage: distance between two clusters is defined as the maximum of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.
2. Average linkage: compromise between single and complete linkage. Less susceptible to noise and outliers.
3. Ward’s minimum variance linkage: minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are combined.

Once we chose the linkage method and created the dendrogram we need to decide about the height of the fusion of the tree. It is provided on the vertical axis and indicates the distance between two clusters. The higher the height also known as cophenetic distance, the less similar are the clusters.

As a starting point we use the ward minimum variance linkage which divides the observations into three clusters when we cut the tree just below a height of four (Figure 13). Each cluster has a significant amount of observations. In comparison, Figure 14 shows the dendrogram for average linkage which would have one

very big, one medium-sized and one very small cluster. Complete linkage shows a same distribution as the Ward linkage method. Therefore, we continue the analysis with the ward results.

To verify the cluster tree, we assess if the cophenetic distances in the tree reflect the original distances. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix. The closer the value of the correlation coefficient is to 1, the more accurately the clustering solution reflects your data. In the Clustering with Ward linkage we achieve a value of 0.81 for the cophenetic value and values above 0.75 are generally felt to be good.

Alternative ways to verify the cluster tree are the average silhouette width and the gap statistic which we mentioned before when it came to choose an appropriate number of k's for K-Means method. For the sake of completeness in the following we briefly mention the theory behind them.

The average silhouette width measures how close one observation is to another in the same cluster compared to observations in different clusters. Values near 1 imply that the observation is well placed whereas values near 0 mean that it should rather be in another cluster.

The gap statistic compares the observed within-cluster variation to the expected value under the null reference distribution of the data with no obvious clustering (NULL hypothesis).

In Figure 15 the dendrogram is cut into three clusters and three different colours indicate the different groups. For interpretation purposes we look at the cluster means in Table 4. We can compare the results to the K-Means clustering with k=3.

Table 4

	Foreign Investment	**Electricity Access**	**Renewable Energy**	**CO2 Emission**	**Inflation**	**Mobile Subscr.**	**Internet Use**	**Exports**	**Imports**	**GDP**
1	0.17	0.69	-0.50	0.37	-0.44	0.61	0.91	0.28	0.19	0.64
2	-0.07	-0.92	0.86	-0.62	0.30	-0.81	-0.89	-0.34	-0.04	-0.60
3	-0.19	0.63	-0.87	0.60	0.23	0.55	0.11	0.17	-0.29	0.01
	Mortality Male	**Mortality Female**	**Birth Rate**	**Death Rate**	**Mortality Infant**	**Life Expect.**	**Fertility Rate**	**Populat. Growth**	**Urban Populat.**	**#**
1	-0.71	-0.60	-0.92	0.22	-0.82	0.81	-0.82	-0.73	0.58	**70**
2	0.87	0.77	0.92	0.24	0.92	-0.92	0.88	0.40	-0.79	**77**
3	-0.48	-0.48	-0.16	-0.98	-0.35	0.38	-0.28	0.60	0.56	**35**

As a first observation we can state that the variance of the cluster means is closer than in the methods we interpreted before. Cluster one groups countries with good infrastructure, medium pollution values, high life expectancy, high GDPs and low inflation. The group is relatively large and can be classified as high-income cluster. But looking at the countries in this group, a few observations such as Cuba seem suspicious. Cluster 2 represents low income countries, with poor infrastructure and high mortality rates. The population is

growing and living in rural areas. The cluster has the lowest GDP. Cluster 3's GDP value is close to the middle of the two other clusters and can therefore be classified as medium income cluster. Countries in this group have similar values to cluster 1 when it comes to infrastructure and population values but their international trade relations as well as their inflation values are worse. This group could also be referred to as emerging markets including countries like Turkey, China and Mexico.

One downside of hierarchical clustering is that there is no automatic discovering of the optimal number of clusters and it can be difficult to handle the differently sized clusters like in the average linkage example above where one cluster had only one observation. But it has several advantages. It does not need any input parameters besides of the choice of the dissimilarity and it computes a complete hierarchy of clusters which can be nicely visualized.

D) Conclusion

The different clustering methods explored in this report are all producing interpretable results. We could argue for each method in one way or another. Especially the K-Means method with three different clusters produced interesting results. For further exploration I would like to look at different variables. In this dataset we were looking at various variables that are highly correlated such as the different mortality rates. We suspect that value could be added when including other features such as employment statistics or education parameters.

Graphs and Figures

Figure 1 Boxplot of scaled and centered data

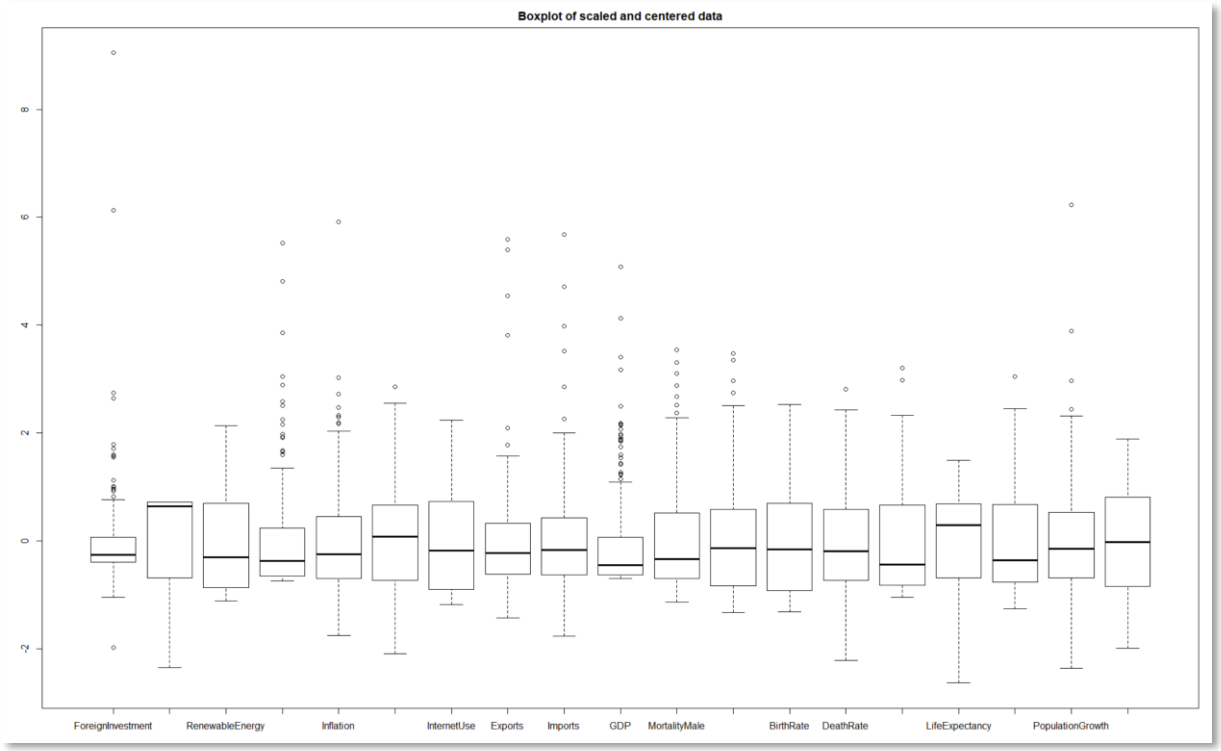


Figure 2 PCA Original Data

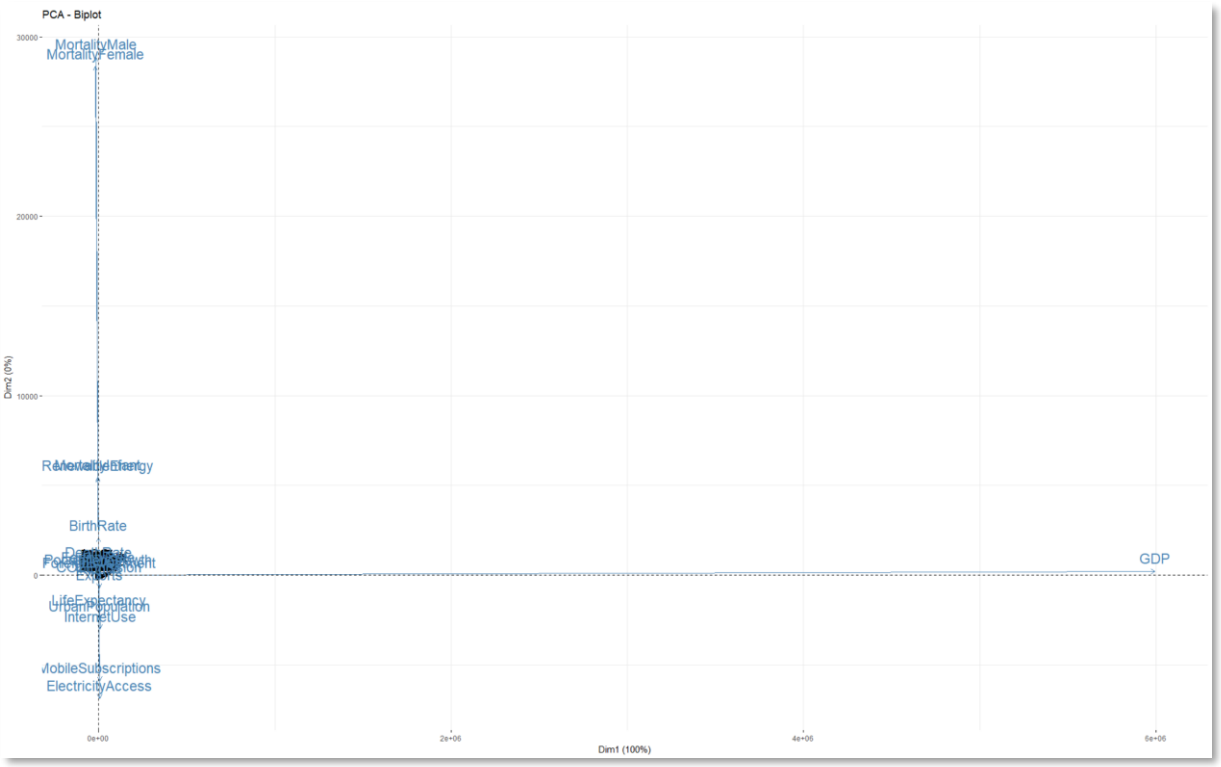


Figure 3 Scree Plot Original Data

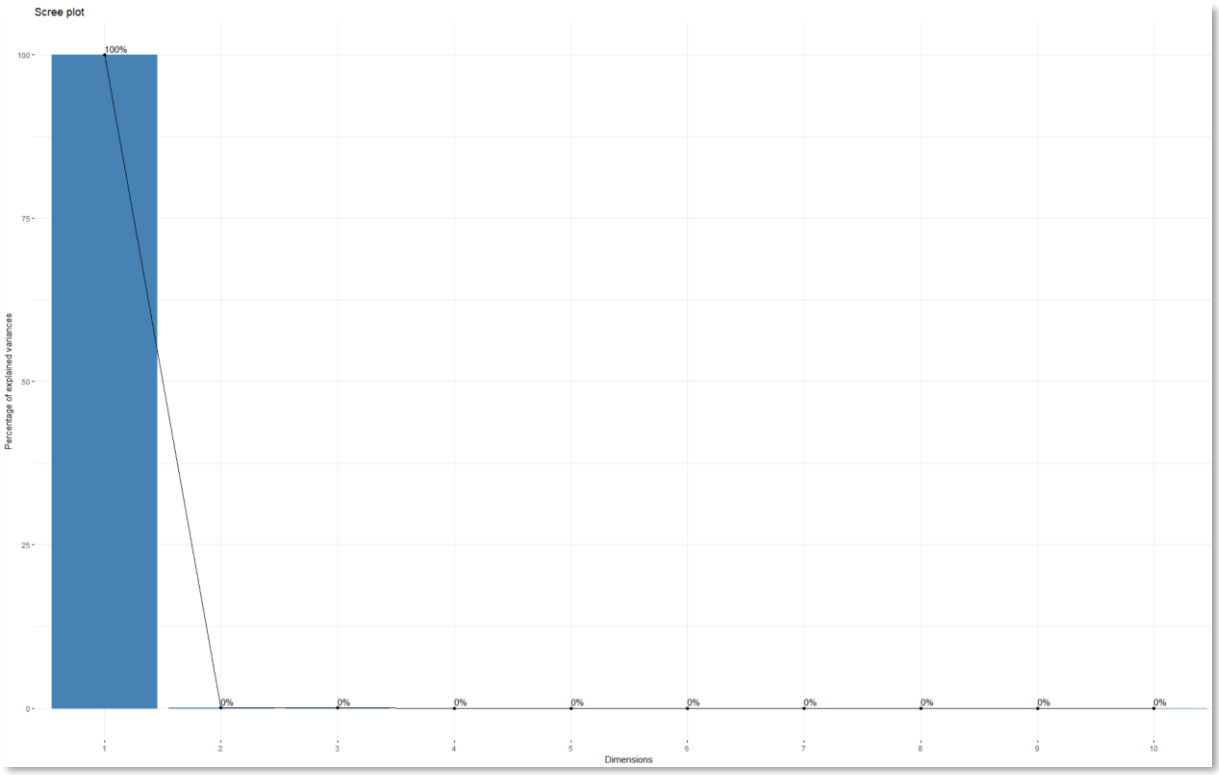


Figure 4 Biplot PCA Standardized Values

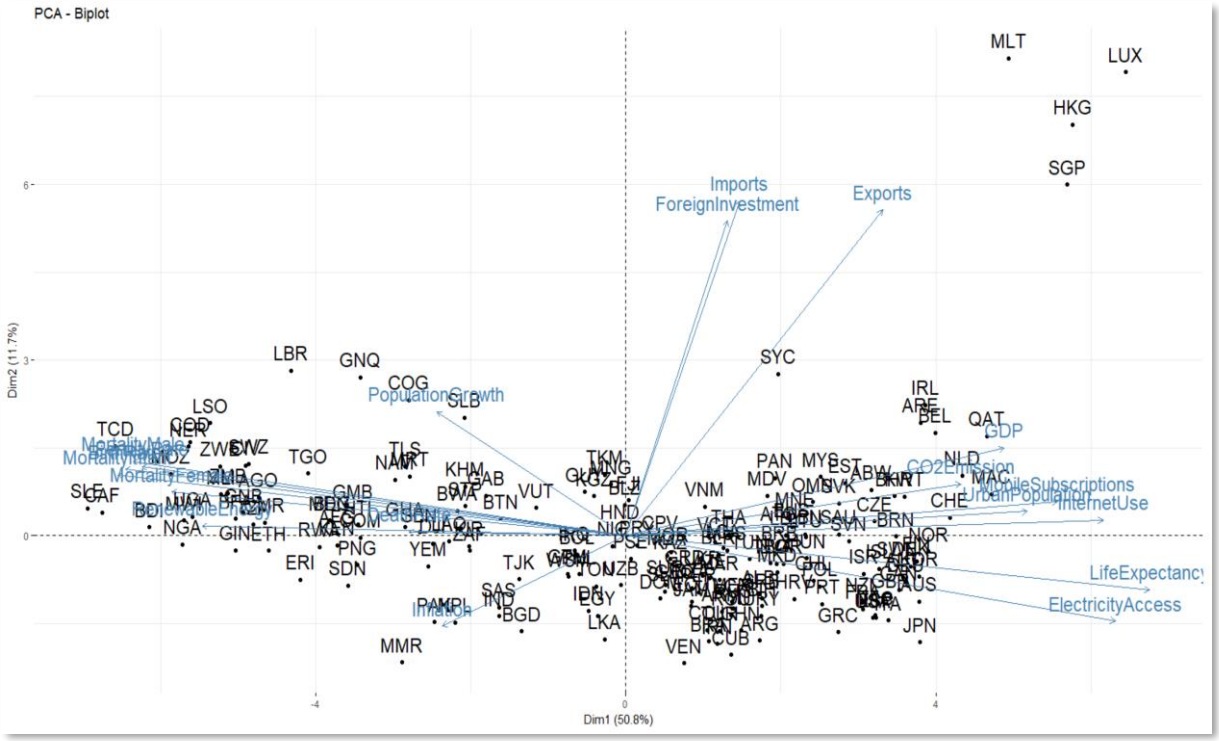


Figure 5 Scree Plot PCA Standardized Values

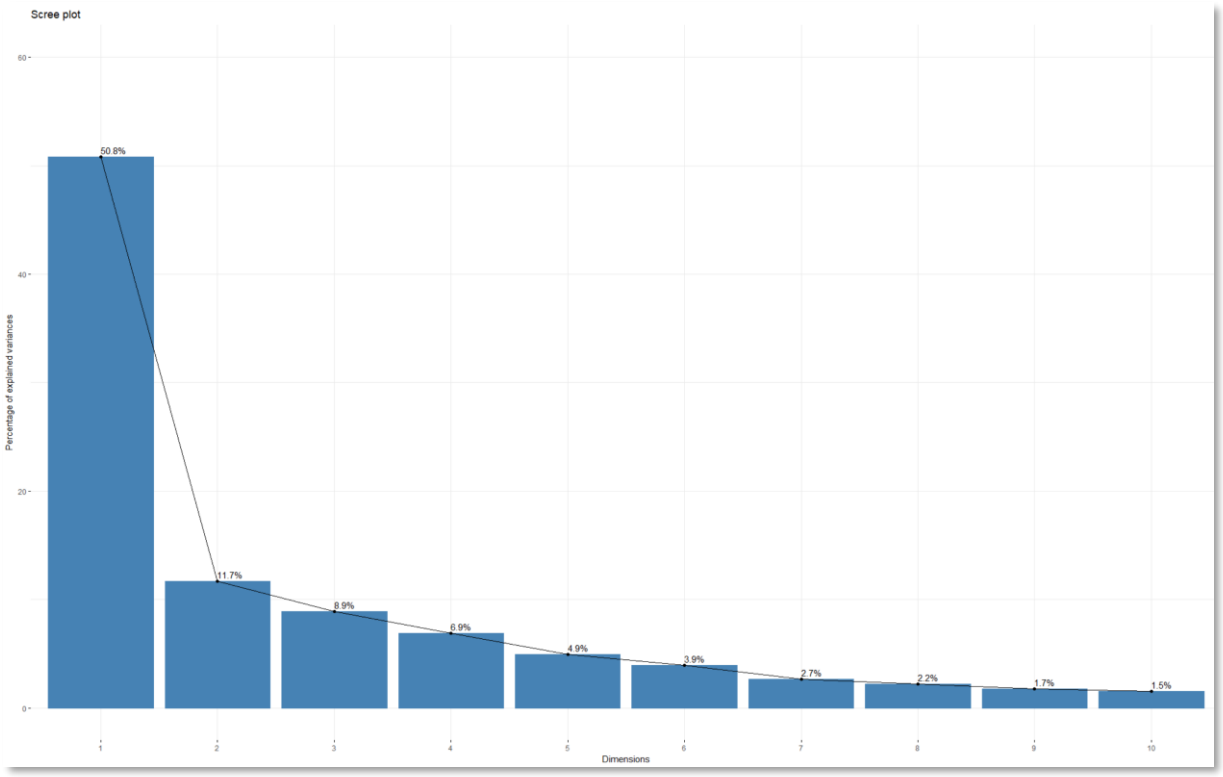


Figure 6 First 7 Principle Components of PCA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
ForeignInvestment	0.184	0.755	-0.115	-0.218	0.032	0.352	0.003
ElectricityAccess	0.888	-0.205	-0.070	-0.123	0.132	-0.139	0.072
RenewableEnergy	-0.768	0.024	-0.043	0.130	-0.259	0.330	0.017
CO2Emission	0.607	0.124	0.371	0.424	0.275	-0.131	-0.361
Inflation	-0.333	-0.218	-0.018	-0.296	0.753	0.380	-0.071
MobileSubscr.	0.782	0.082	-0.106	0.088	0.206	-0.235	0.201
InternetUse	0.866	0.037	-0.040	0.351	-0.068	0.167	-0.030
Exports	0.466	0.781	-0.088	-0.047	0.134	-0.110	0.021
Imports	0.206	0.803	-0.296	-0.263	-0.069	-0.118	-0.081
GDP	0.686	0.210	0.210	0.486	-0.066	0.297	-0.095
MortalityMale	-0.893	0.179	-0.089	0.258	0.108	-0.118	0.049
MortalityFemale	-0.823	0.104	-0.273	0.226	0.220	-0.232	0.037
BirthRate	-0.907	0.159	0.266	-0.026	-0.025	0.016	0.105
DeathRate	-0.393	0.009	-0.655	0.557	0.091	0.071	0.008
MortalityInfant	-0.922	0.146	0.063	0.090	0.069	0.031	0.028
LifeExpectancy	0.950	-0.130	0.039	-0.146	-0.138	0.131	-0.014
FertilityRate	-0.877	0.166	0.272	0.054	-0.078	0.070	0.084
PopulationGrowh	-0.342	0.297	0.826	0.066	0.089	-0.124	0.049
UrbanPopulation	0.728	0.058	0.158	0.231	0.144	0.132	0.53

Figure 7 Cumulative Proportion of Variance Explained Plot Standardized Data

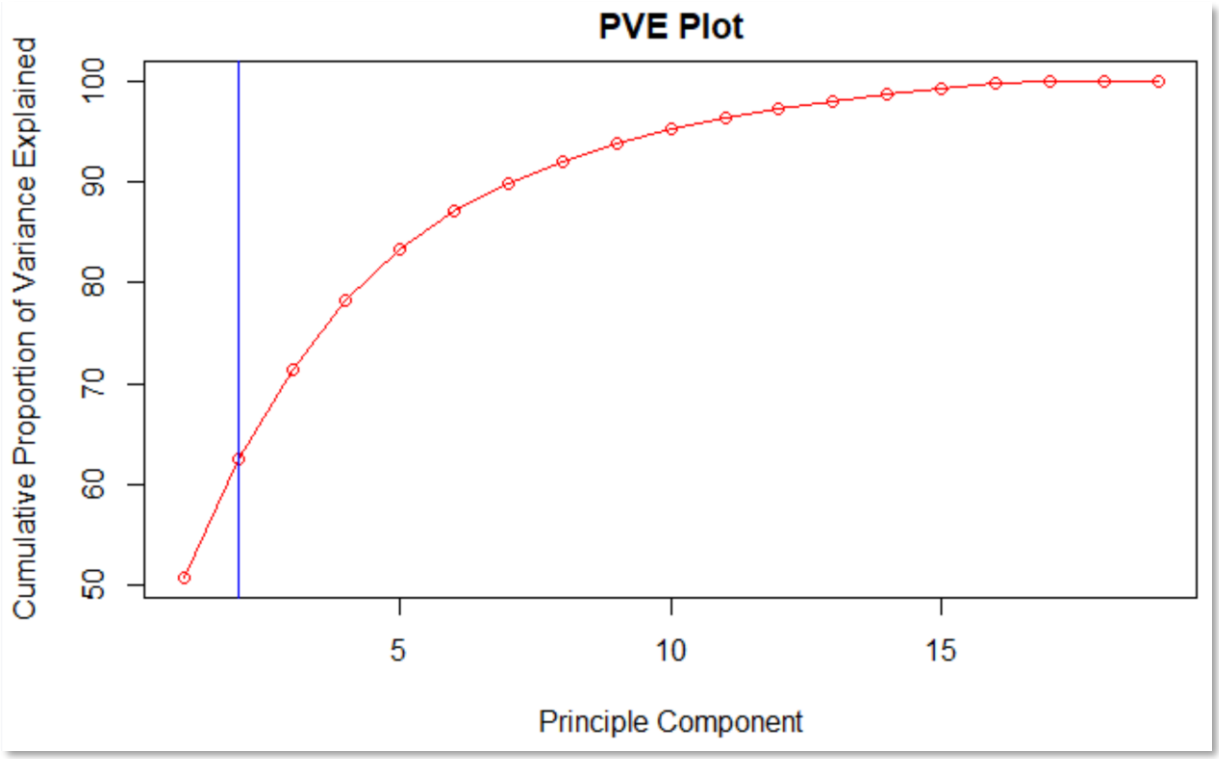


Figure 8 Optimal Number of Clusters based on Majority Rule

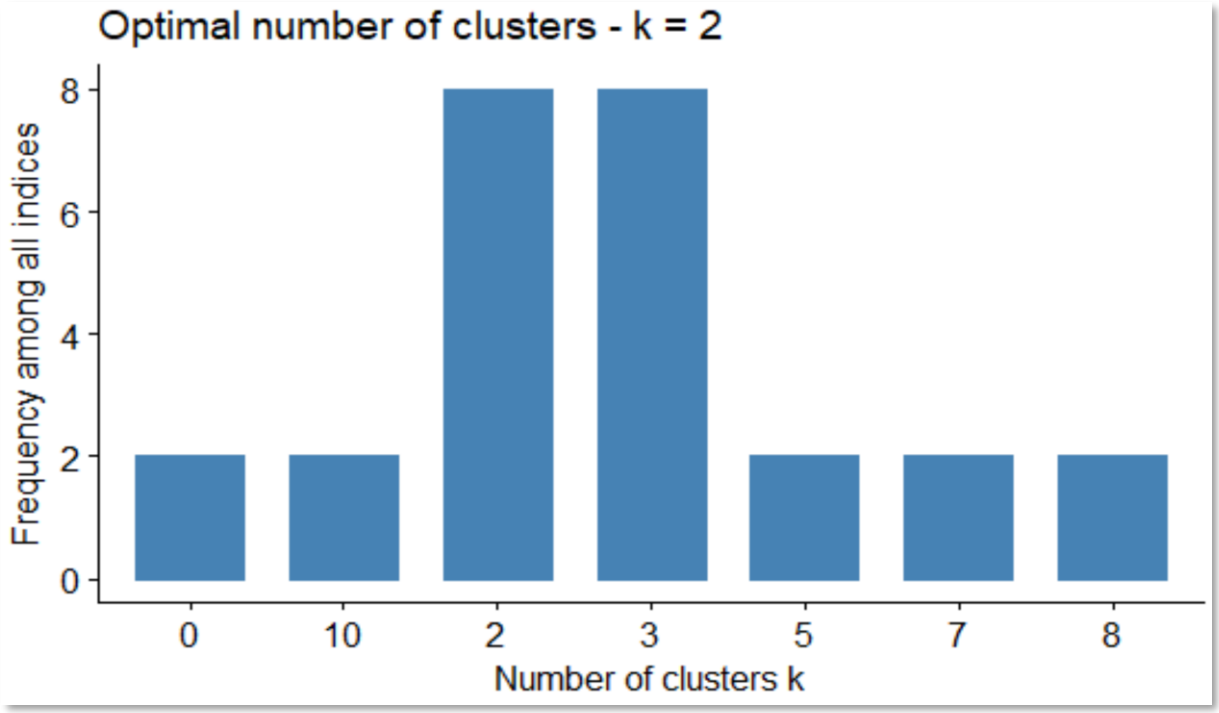


Figure 9 Visualization of K-Means Clustering with k=2

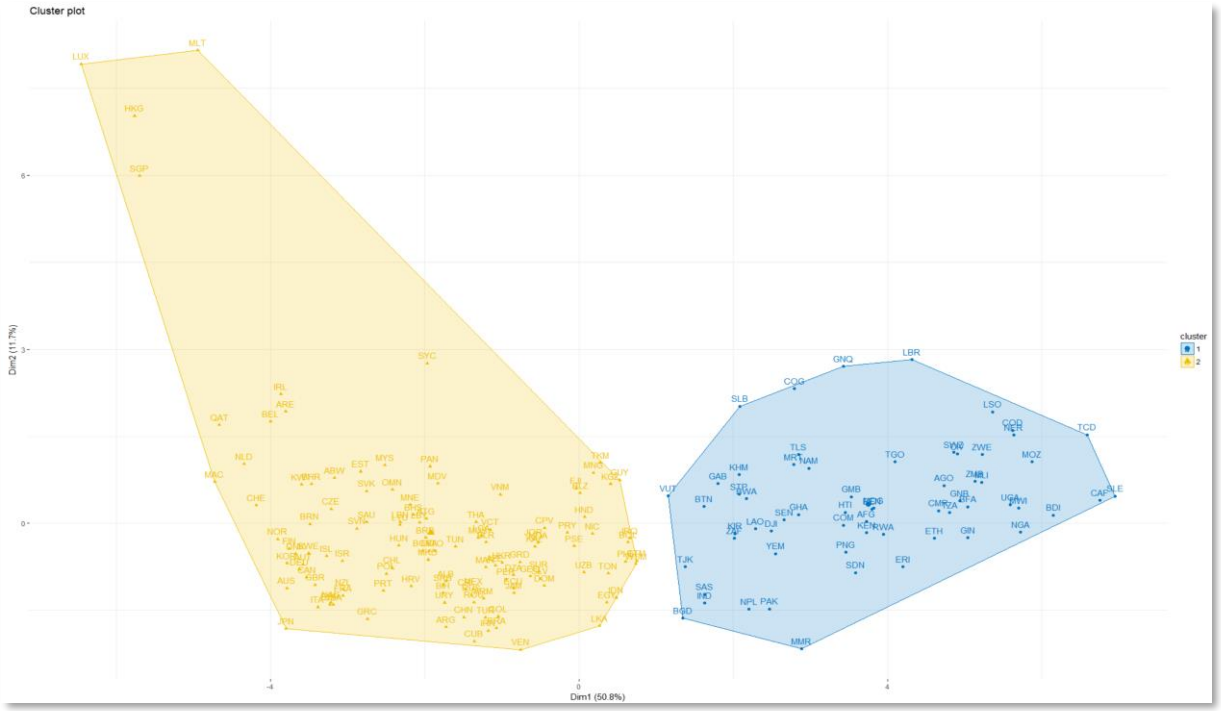


Figure 10 Visualization of K-Means Clustering with k=3

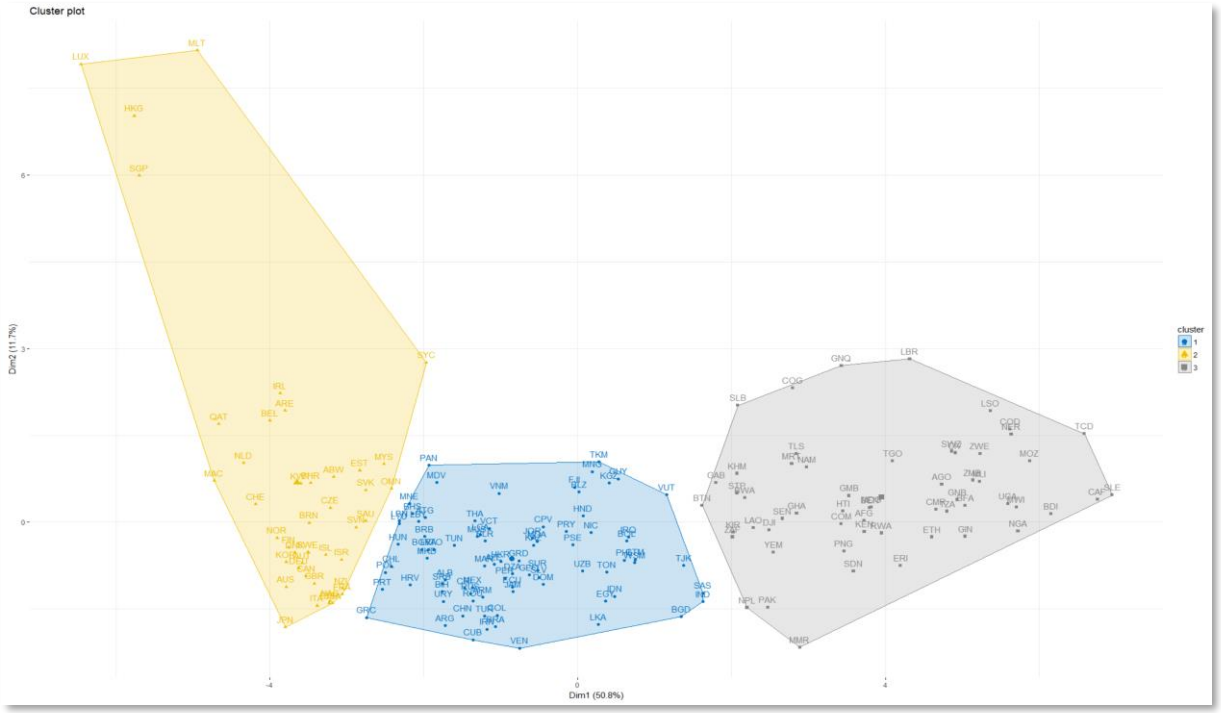


Figure 11 K-Medoid Optimal Number of Clusters based on Average Silhouette Width

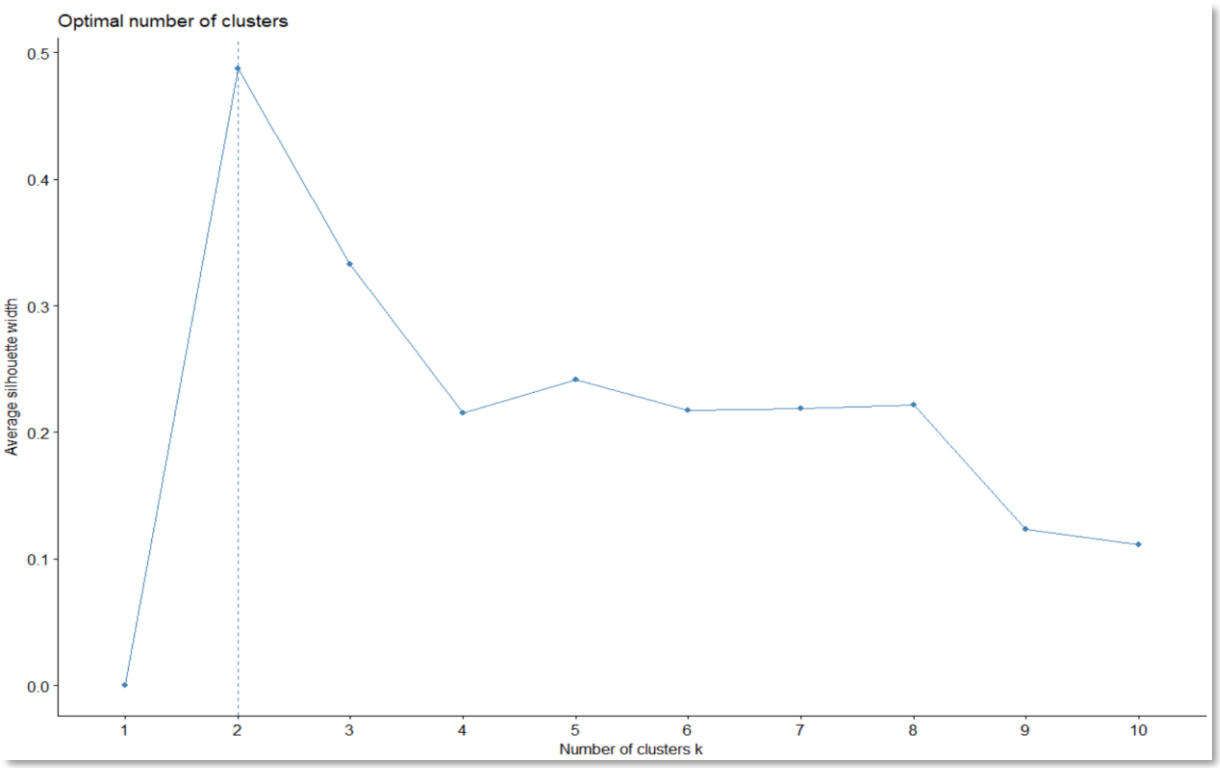


Figure 12 K-Medoid Clustering k=2

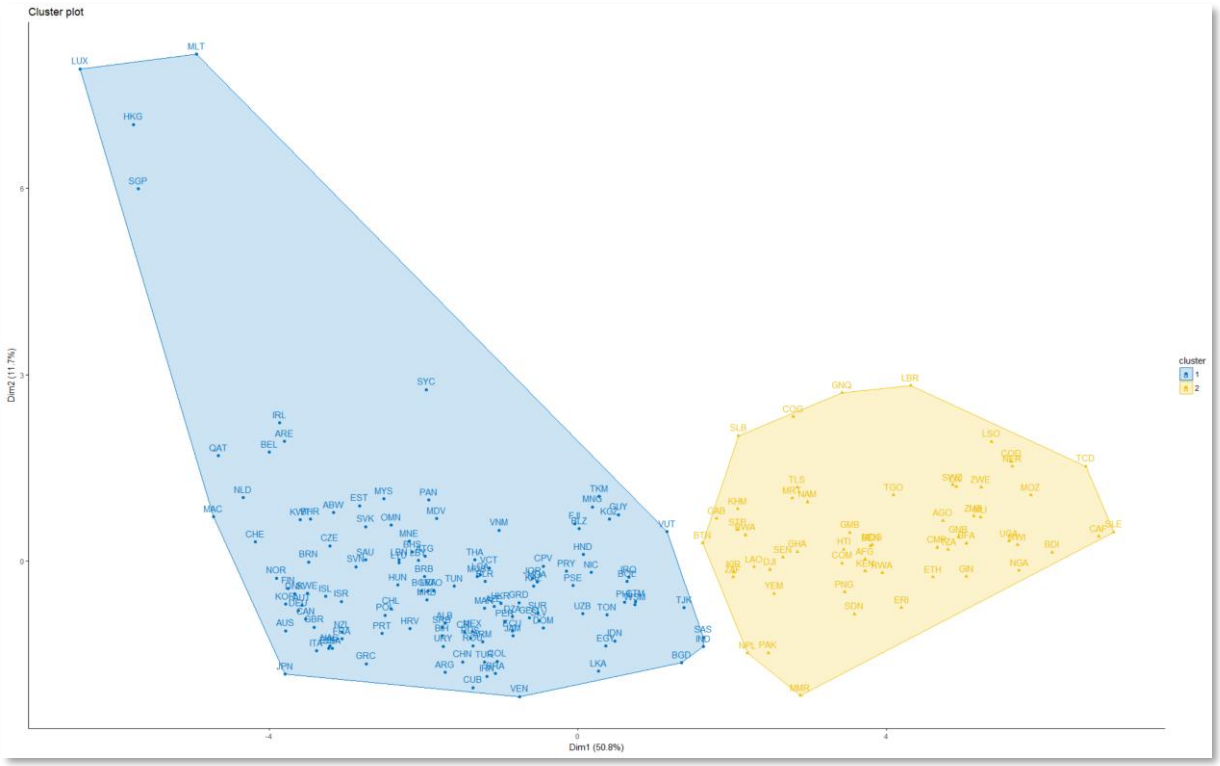


Figure 13 Dendrogram Ward Linkage

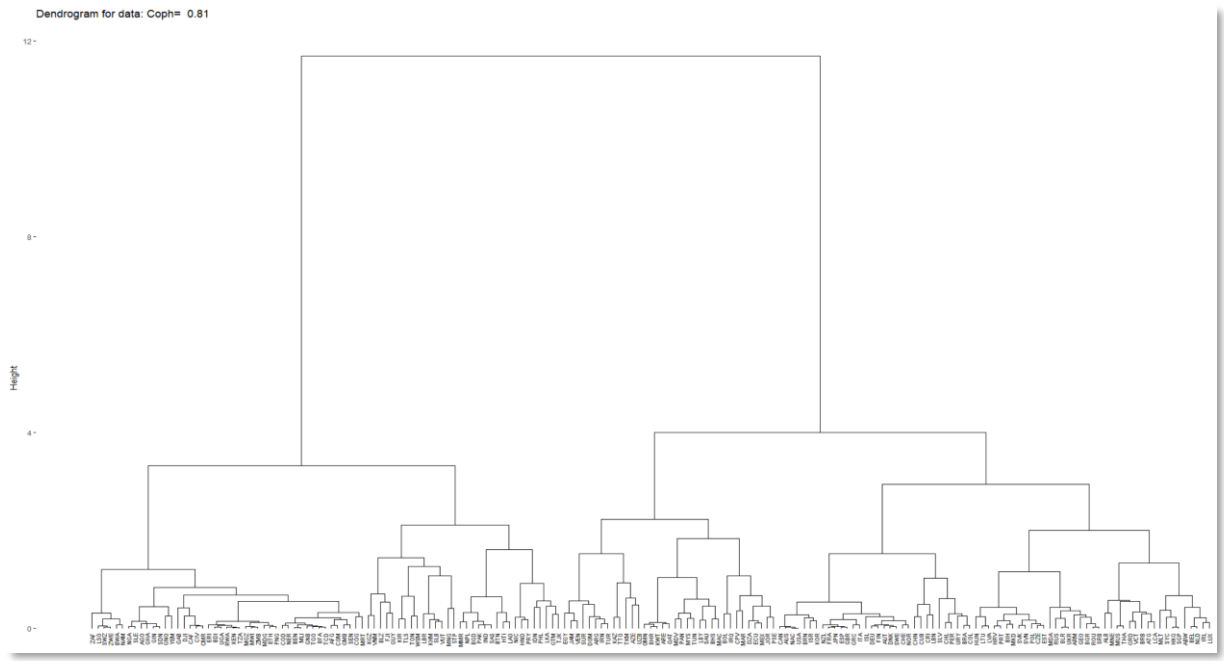


Figure 14 Dendrogram Average Linkage

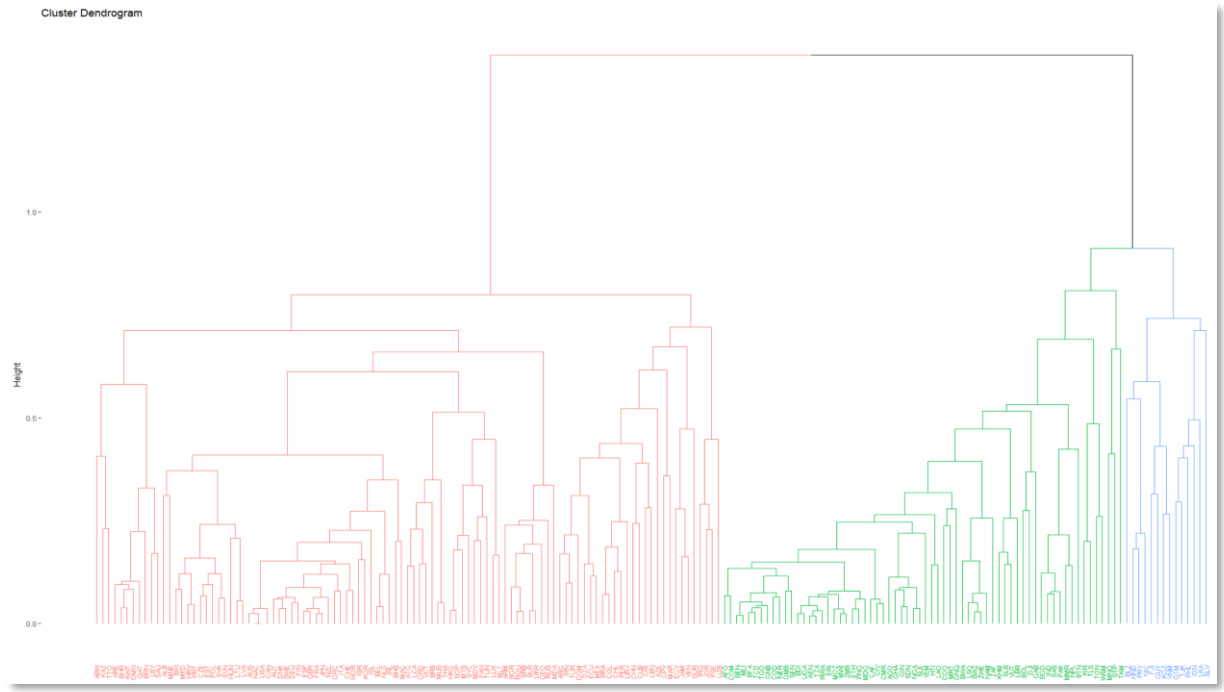


Figure 15 Dendrogram Ward Linkage (colour coded)

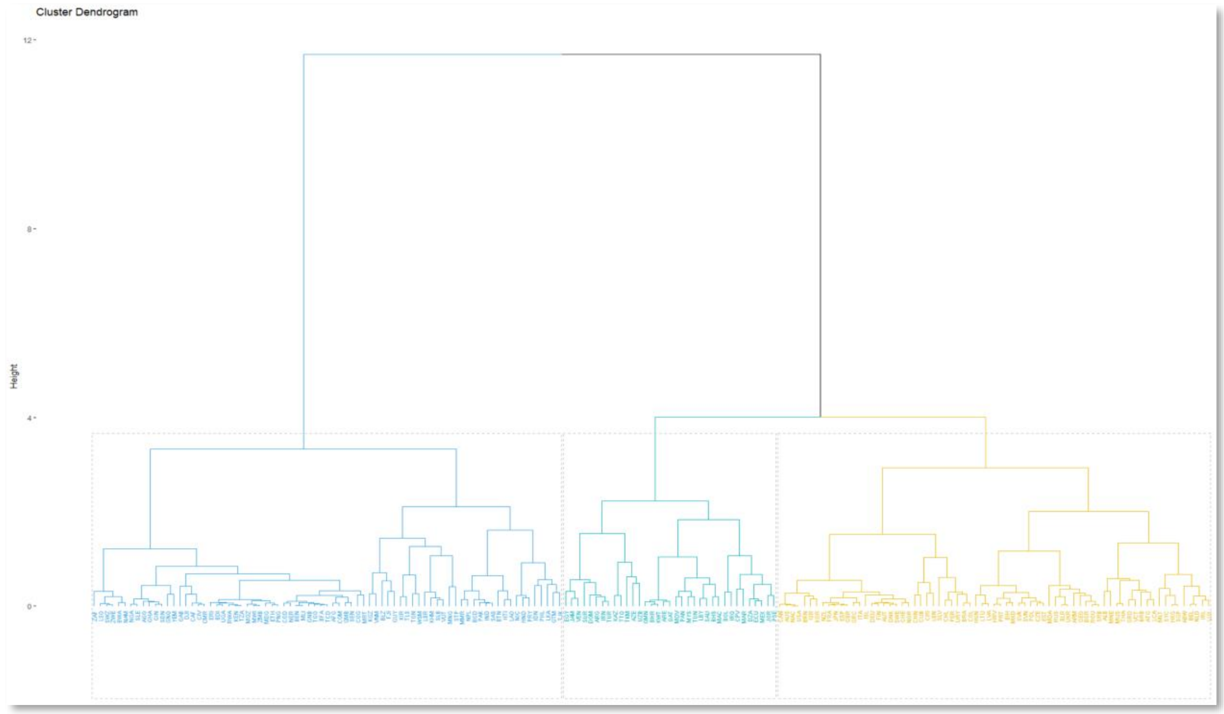
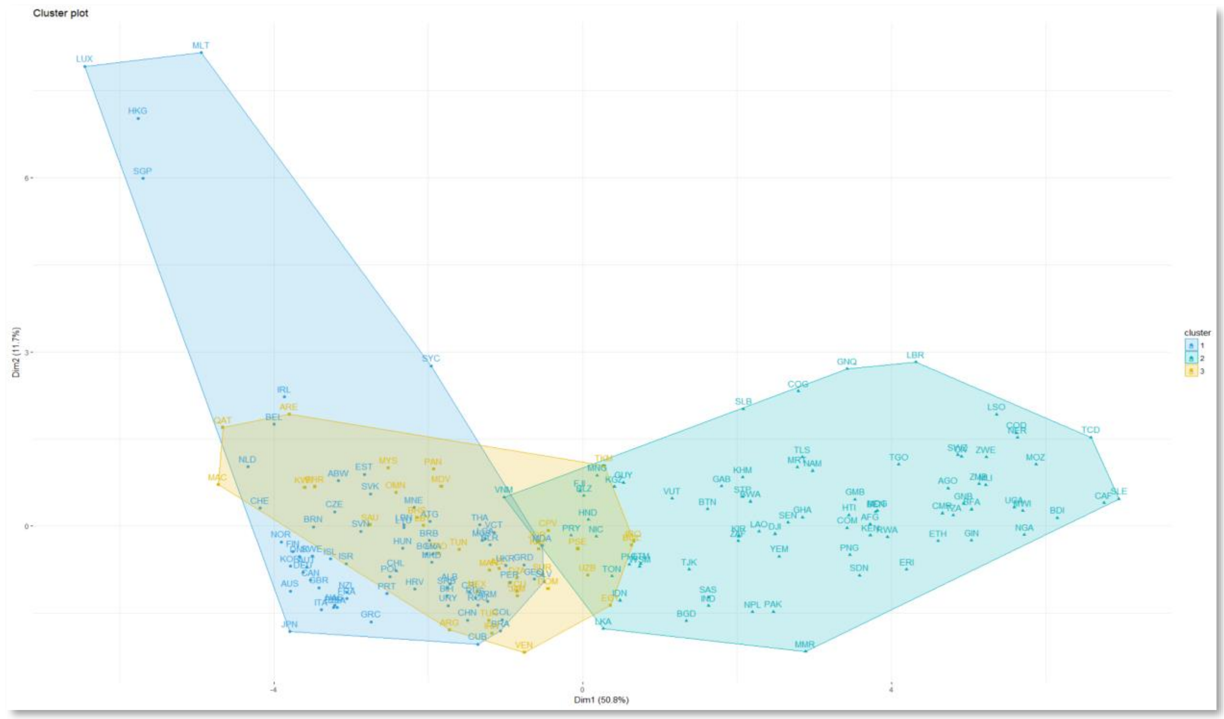


Figure 16 Cluster Plot Ward Linkage Result



References

‘‘Introduction to Statistical Learning’’ by Gareth James, Daniela Witte, Trevor Hastie and Robert Tibshirani

‘‘The Elements of Statistical Learning’’ by Trevor Hastie, Robert Tibshirani and Jerome Friedman

‘‘Practical Guide To Cluster Analysis in R’’ – Alboukadel Kassambara