Cluster Analysis Assignment

Nyasha Mashanda

2020-10-18

Contents

Contents

List of Figures

List of Tables

Table 1: Acronyms

Variable	Description
iso_code	ISO 3166-1 alpha-3 – three-letter country codes
continent	Continent of the geographical location
location	Geographical location
date	Date of observation
total_cases	Total confirmed cases of COVID-19
new_cases	New confirmed cases of COVID-19
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed)
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people
new_cases_smoothed_per_million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people
total_deaths_per_million	Total deaths attributed to COVID-19 per 1,000,000 people
new_deaths_per_million	New deaths attributed to COVID-19 per 1,000,000 people
new_deaths_smoothed_per_million	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people
population_density	Number of people divided by land area, measured in square kilometers
median_age	Median age of the population, UN projection for 2020
aged_65_older	Share of the population that is 65 years and older, most recent year available
aged_70_older	Share of the population that is 70 years and older in 2015
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars)
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017
female_smokers	Share of women who smoke, most recent year available
male_smokers	Share of men who smoke, most recent year available
handwashing_facilities	Share of the population with basic handwashing facilities on premises
hospital_beds_per_thousand	Hospital beds per 1,000 people, most recent year available since 2010
life_expectancy	Life expectancy at birth in 2019

Abstract

The novel COVID-19 corona virus is still not well understood and there are many open questions related to patterns in its spread. The goal of this assignment is to discover if there are any regional patterns that exist using cluster analysis.

The assignment uses COVD-19 pandemic data collected from the Our World In Data site (European Center for Disease Prevention and Control 2020). The data contains 29 indicators related to the COVID-19 cases for 208 countries. The data set is updated daily from when the pandemic started. For this assignment, a subset of the data will be used; this subset consists of all the information on the pandemic on the 02 of September 2020.

There are three kinds of clustering methods that will be explored in this analysis: hierarchical, partitioning and density based methods. In order to determine any regional patterns, the number of clusters will be limited to 6, resembling the six regions that are: Africa, Asia, Europe, North America, Oceania and South America.

1 What about standardising the data????

2 Exploratory Data Analysis

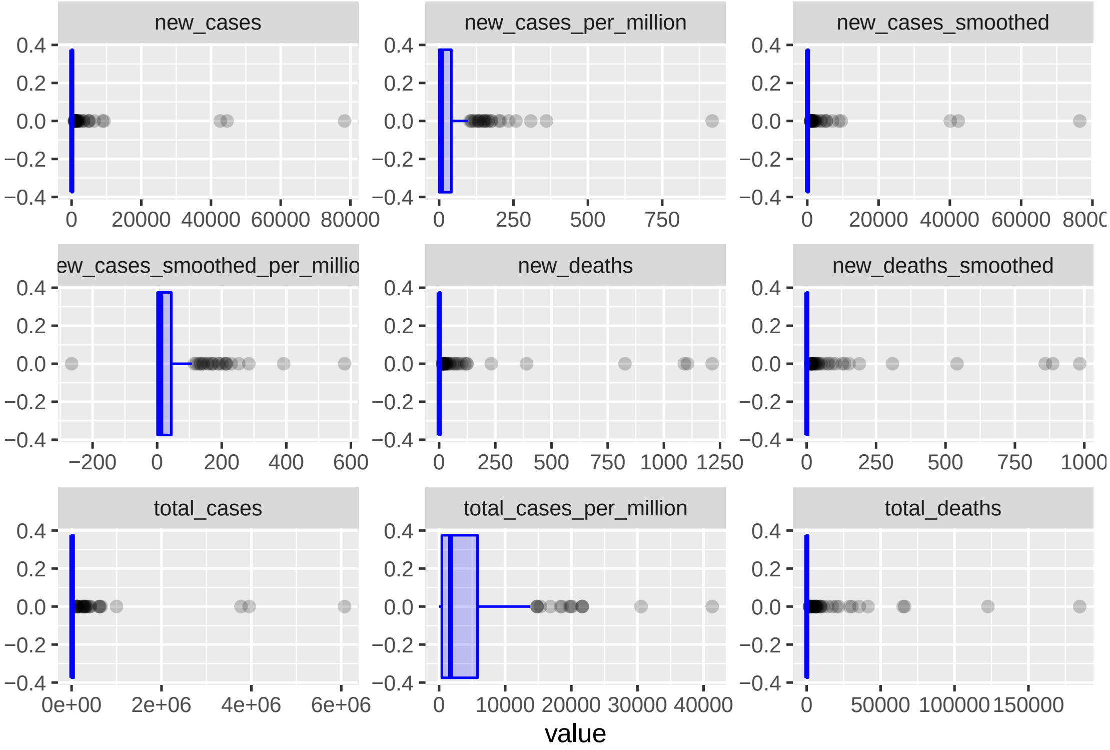
2.1 Importing the Data

The data was exported to a dataframe from an excel file “owid-covid-data.xlsx”. The first step was to check if the data has been imported correctly using the head() and tail() functions. This file was confirmed to have been imported correctly. The next step included checking the structure of the dataframe and also the variables in the data (See variables in Table 1). Location, iso_code, continent and date were found to be in character format while the rest of the variables are numerical. The date variable is expected to be in a date format. However, since this column has only one date, the column will not be used for analysis and will be removed from the dataframe.

The next step was to visualize the distribution of the variables and this was done using box and whisker plots. This is a very important step as it may highlight outliers and incorrectly recorded values that are out of the expected range.

2.2 Distribution of variables

In plotting the distribution of numerical variables, the variables were divided into three groups. This makes it easier to analyse the variables. The variable in the first group are shown in Fig **??**.



Most variabled have many outleiers that make them seem as if thier values are less spread. This indicates that a clustering method that is less susceptible to outliers can be used.

Taking a deeper look into the data, it is immediately obvious that new_cases_smoothed_per_million might has an incorrectly recorded value given that it has a negative outlier. To deal with the negative value in new_cases_smoothed_per_million the data was explored to identify the observation with the negative value. This data point was found to be belonging to Luxembourg. Upon further investigation, Luxembourg seems to have negative values for new_cases_smoothed and new_cases_smoothed_per_million. This is probably because the new_cases_smoothed_per_million is derived from new_cases_smoothed.

For variables in Fig @ref(fig:group-2, there are numerous outliers in each of the variables.

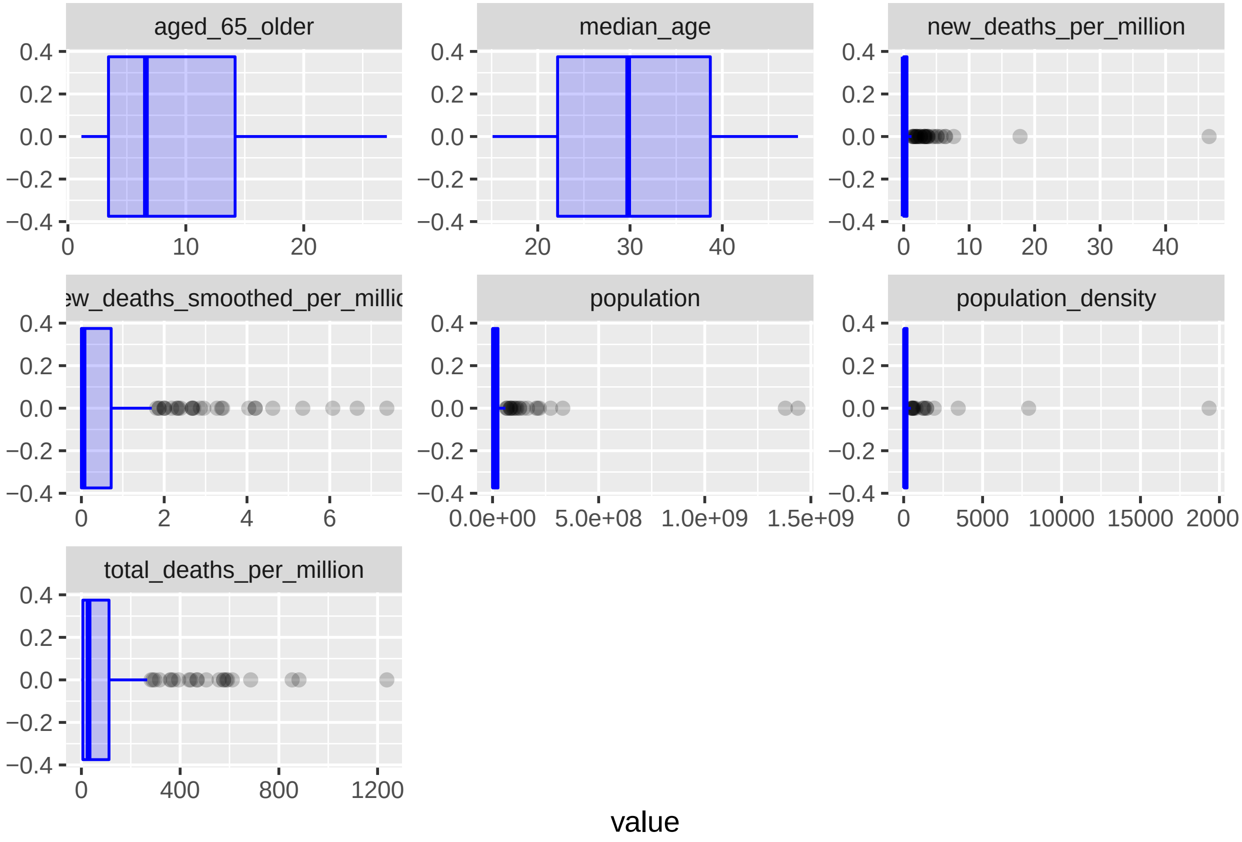


Figure 1: Group 2 of Variables

Median_age and aged_65_older values are much more spread. The numbers seem resasonable given that a smaller percentage of the population is age above 65 and usually the median age of most countries is expected to be below 50 given that there usually more young people than older people in country. Countries like Japan and Italy seem to have the highest percentage of older people with a median age of 48.2 and 47.9. The Niger and Uganda have the youngest population with a median age of 16.1 and 15.4. In general European countries seems to have an older society while Africa countries have a much younger population.

Monaco and Singapore have the highest population densities while China has the highest population. As a result these countries stand out as outliers with regards to population variables.

New Deaths per million are highest in north and South America.

New cases smoothed per million show a negative number(outlier). This is most likely a entry error and the values will be replaced with a regional average.

The histograms also suggest that there is a negative value in new_cases_smoothed_per_million.

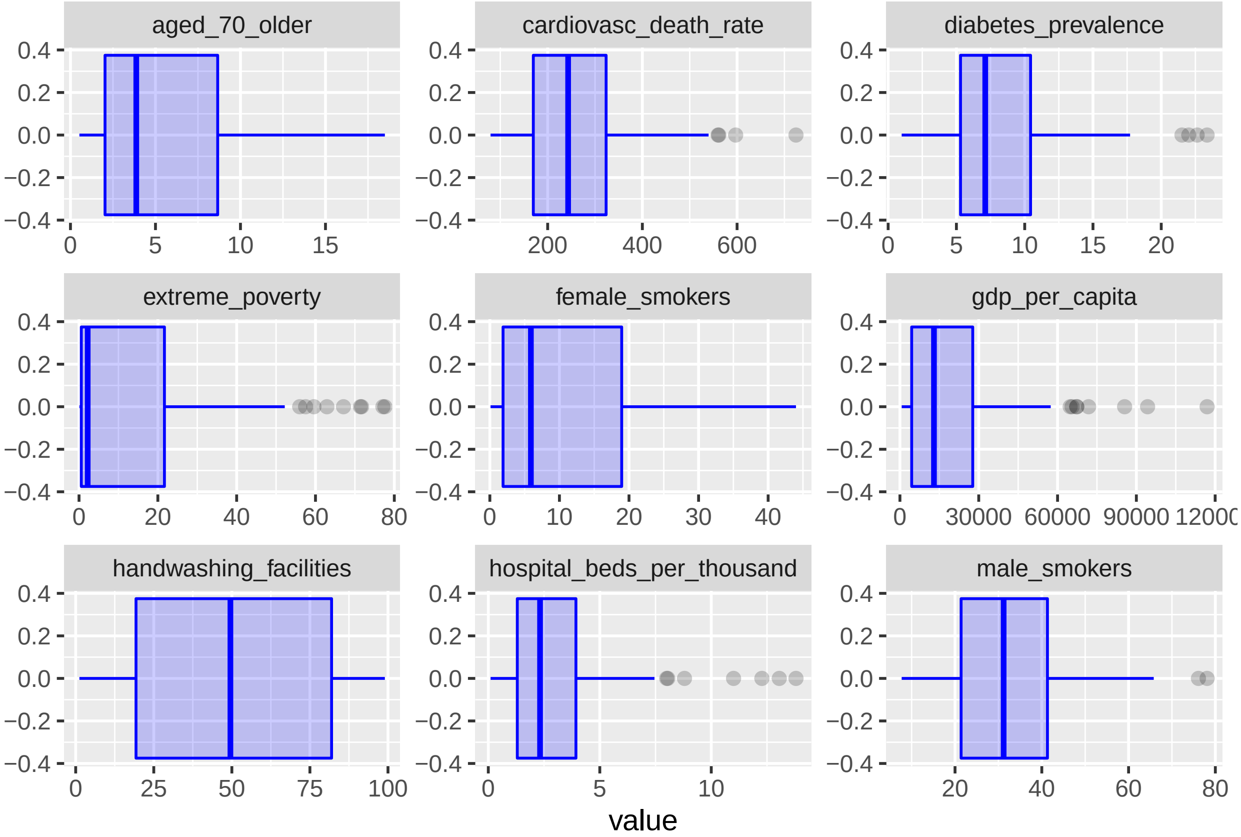


Figure 2: Group 3 of Variables

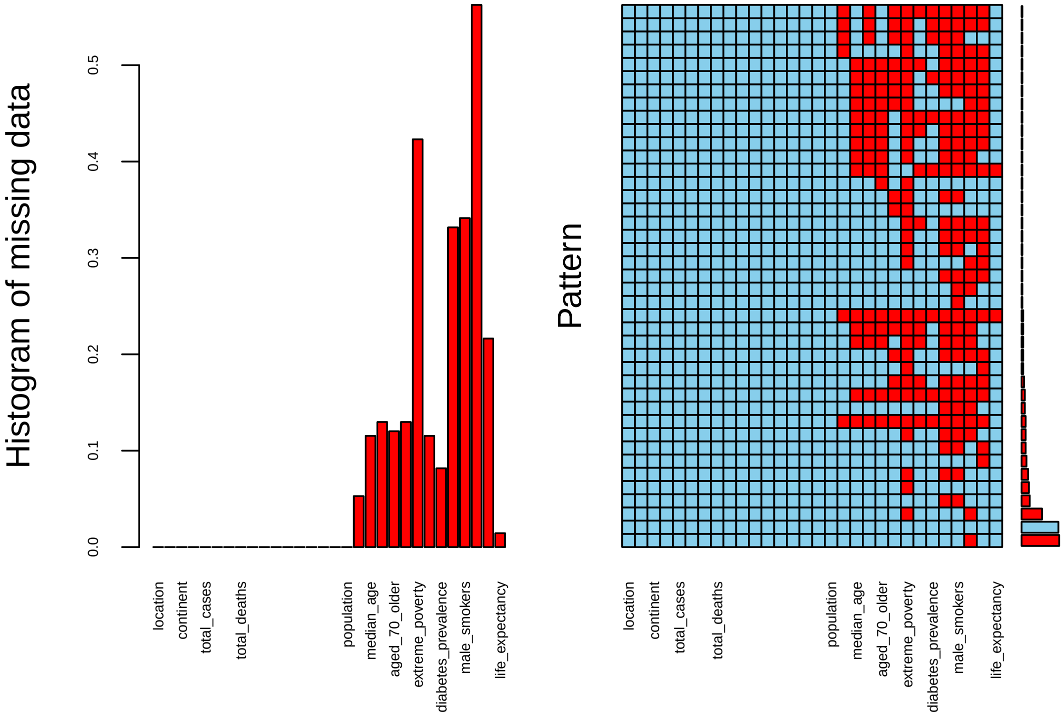
The observations for these variables look more spread with less outliers. The variables also suggest that there are generally more male smokers than female smokers. Furthermore, the data show that it is mostly european countries that are wealthy or less poor. The numbers seem to be in the expected ranges.

2.3 Missing values

Running a summary on the dataframe shows that there are various columns with missing data.

Table 2: Missing data visualisation

Variable	Count	Percentage
handwashing_facilities	117	0.56250
extreme_poverty	88	0.42308
male_smokers	71	0.34135
female_smokers	69	0.33173
hospital_beds_per_thousand	45	0.21635
aged_65_older	27	0.12981
gdp_per_capita	27	0.12981
aged_70_older	25	0.12019
median_age	24	0.11538
cardiovasc_death_rate	24	0.11538
diabetes_prevalence	17	0.08173
population_density	11	0.05288
life_expectancy	3	0.01442



As show in (Table 2), a number of variables have a very high number of missing values including handwashing_facilities, extreme_poverty, male smokers and female smokers. It is generally not ggod working with data or columns that have a high percentage of missing data. Therefore various methods of imputing the missing data will be observed in the next section.

2.4 Imputation by data scrapping

There are various sources that were used to collect the data including the European Center for Disease Prevention and Control. However some of the data can be found on wikipedia and various internet sources. Consequently, looking for the data before applying an imputation algorithms is much better. Although the data was not exactly be the same, in most cases the values were found to be very similar thereby justify the imputation.

Population density missing values were imputed using the data from wikipedia (Wikipedia 2020a). The data for all the countries with missing values was found except for the Falkland Islands.

Median age missing values were imputed using the data found from the CIA (CIA 2020). Using this site only three countries had values for median age that could not be found.

aged_65_older missing values were imputed using 2019 data from wikipedia (Wikipedia 2020b). Only data for Syria, Taiwan and British Virgin Islands. The rest of the countries values could not be imputed using this method.

Median age above 75 years old was ignored since there was not much data on the internet regarding this variable.

Extreme poverty is a population with an income of less than $1.90 a day. From the data the Share of the population living in extreme poverty, most recent year available since 2010. The data from wikipedia (Wikipedia 2018). The data is very similar to that available without missing values, therefore it will be better to use the data for imputing.

Most countries with 0 percent were having missing values

Cardiovascular death rate was skipped due to insufficient data to fill in the missing values.

female and male smokers info not enough handwashing facilities skipped because data is not enough. hospital bed per thousand informatiion was found to be too old for example Chad which had data for 2005. Given that there are many years between the time the data was collected and this year the data was ignored.

Only three observations were found to have missing life expectancy values. Imputation was done using the world bank data on life expectancy (The World Bank Data 2020). The life expectancy of Guernsey, Jersey and Kosovo were found to be 82.6, 80.6, 71.95 years respectively.

No data was found on the internet woith regards to the rest of the variables on their missing data. There alternative methods had to be applied.

With regards to population density, extreme poverty, gdp per capita, hospital beds per thousand, aged 65 and older and aged 70 and older, a better imputation method would be replacing missing values with column regional means given that countries in the same regions are more likely to have the same of these variables.

Due to hand_washing_facilities having missing data greater than 70%, the variable was removed since it is most likely that the imputed values will be far from the actual values.

After imputation using the data scrapping and regional averages, the looks like below:

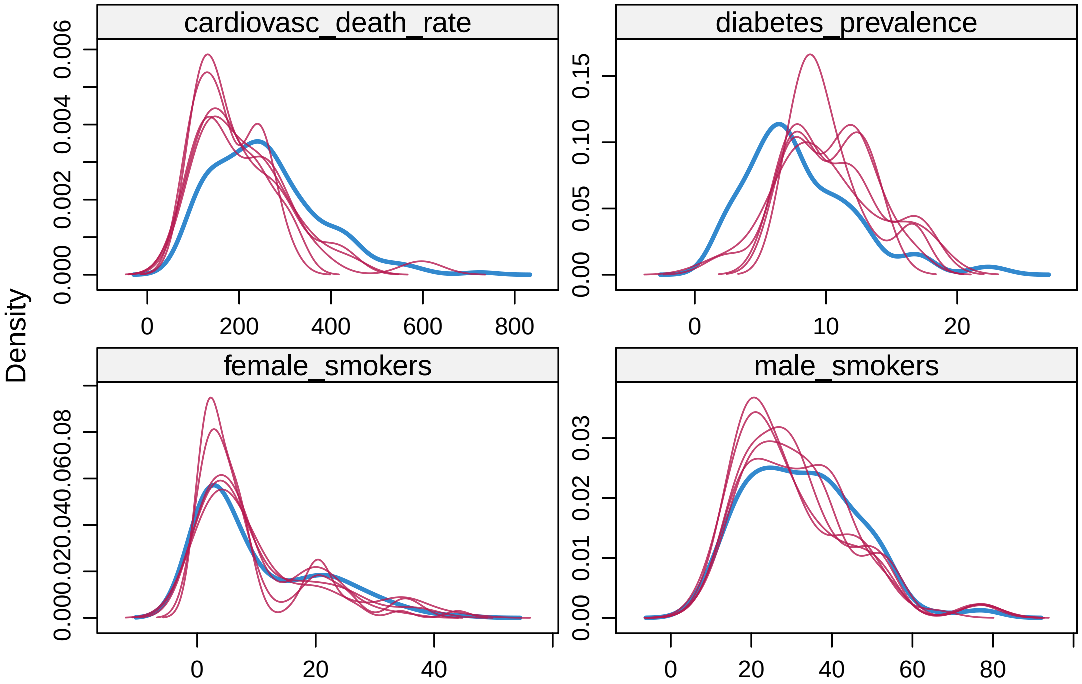
Now the highest percentage of missing values is found in male smokers and female smokers having 34% and 33 % missing values. For the these variables and other that have missing values different imputation algorithms will be explored and the one that gives the best results will be chosen.

The results show that handwashing_facilites is missing 56.25% of the values followed by extreme_poverty missing 42.3% of the values. In total, 78.4% of the samples have missing values.

2.5 Mice

The mice package allows multivariate imputation by chained equations. MICE assumes that the missing data are Missing at Random (MAR), which means that the propensity for a data point to be missing is not related to the missing data but is related to some of the observed data (Karen Grace-Martin 2011). In this section, the MICE package will be used on the remaining variables with missing data that are cardiovasc_death_rate, diabetes prevalence, female and male smokers.

The number of multiple imputations was set to 5, method was set to classification and regression trees and the maximum number of iterations set to 500. 5 datasets were created and the goodness of fit is shown in the plots below. The graphs of imputed values(red) closely resembles that of the available values(blue). This shows that the imputation process was good.



Going forward the second dataset was randomly selected from the five datasets. This will not affect the outcome significantly given that the datasets have very similar properties. The continent, date and iso_code variables were removed from the dataset as they will not be useful in the final analysis. The dataset was named mice_data_for_clustering.

2.6 Hmisc

Hmisc is a multiple purpose package useful for data analysis, high – level graphics, imputing missing values, advanced table making, model fitting & diagnostics (linear regression, logistic regression & cox regression) etc. Hmisc assumes linearity in the variables being predicted. The package was also used to impute values on the missing values.

mi (Multiple imputation with diagnostics) package provides several features for dealing with missing values. Like other packages, it also builds multiple imputation models to approximate missing values. And, uses predictive mean matching method like the mice package.

There is also the Amelia package and mi package that are used for imputation but these will not be considered for this analysis.

The percentage of population aged_65_older looks to have regional patterns therefore it is justified to use regional medians/averages for imputation. Since aged_70_older can is closely related to aged_65_older it is decided to use the same method.

The imputations were made on the dataset and the resulting dataset was named df_Hmisc_for_clustering. The two datasets were used for clustering and the results were compared.

3 Cluster Analysis

In cluster analysis, the number of clusters of interest is 6 given that we have six distinct regions in the data that are Africa, Asia, Europe, North America, Oceania and South America. In order to quantify the goodness of fit for each clustering method, the average silhouette method which determines the quality of the clustering. The optimum number of clusters in a dataset will have the maximum average silhouette width.

Table 3: HCClust

Method	Result
Complete	0.75
Single	0.83
Average	0.79
Median	0.78
Centroid	0.78

Table 4: K-means

continent	clusters_6	count	Percentage
Africa	2	1	1.82
Africa	3	3	5.45
Africa	4	29	52.73
Africa	5	16	29.09
Africa	6	6	10.91
Asia	1	2	4.35
Asia	2	2	4.35
Asia	3	6	13.04
Asia	4	21	45.65
Asia	5	11	23.91
Asia	6	4	8.70
Europe	3	2	4.00
Europe	4	41	82.00
Europe	5	3	6.00
Europe	6	4	8.00
North America	2	1	2.78
North America	3	1	2.78
North America	4	32	88.89
North America	5	2	5.56
Oceania	4	7	87.50
Oceania	5	1	12.50
South America	2	1	7.69
South America	4	6	46.15
South America	5	4	30.77
South America	6	2	15.38

3.1 Heirachical clustering

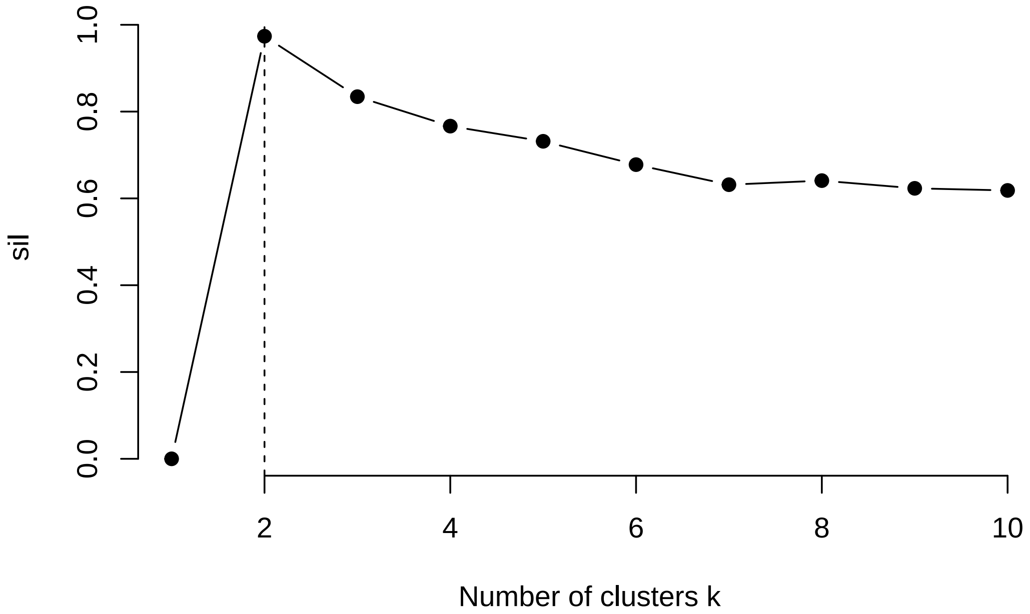
Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. In this excercise complete, single, average, median and centroid linkage were tested.

The results from heirachical clustering are shown in :

The randomizing, re-running, averaging and final run is a very good advice.

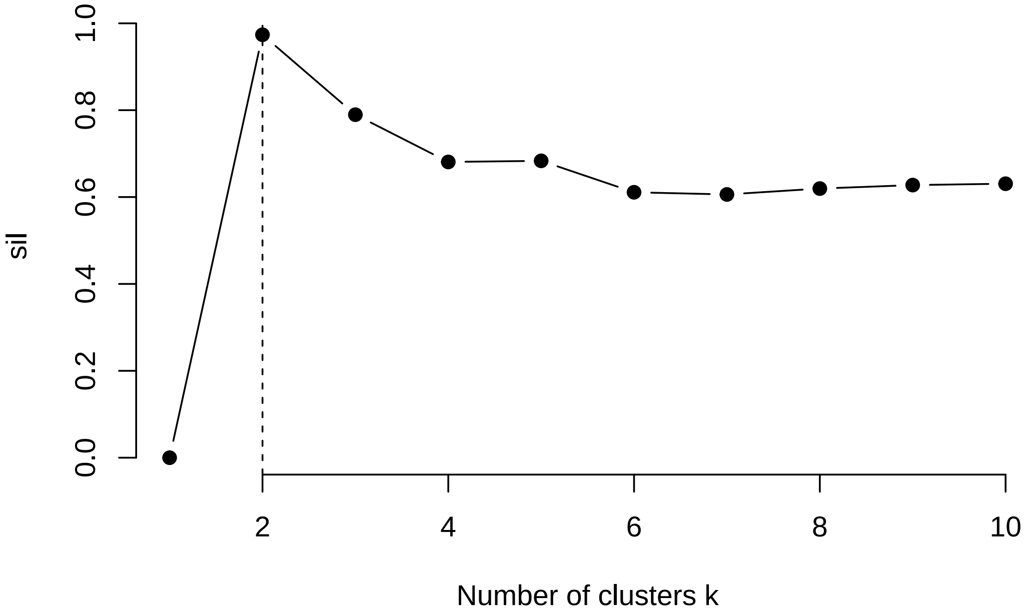
The results show that single linkage had the best results with a average silhouette width of 0.83. China 3, India 4, Indonesia 5, Pakistan 2, Brazil 2, US 6 are outliers in this data. The rest of the countries fall within the same group. Although the average silhouette width is high, the results do not demonstrate any regional patterns for the six continents.

3.2 K-means



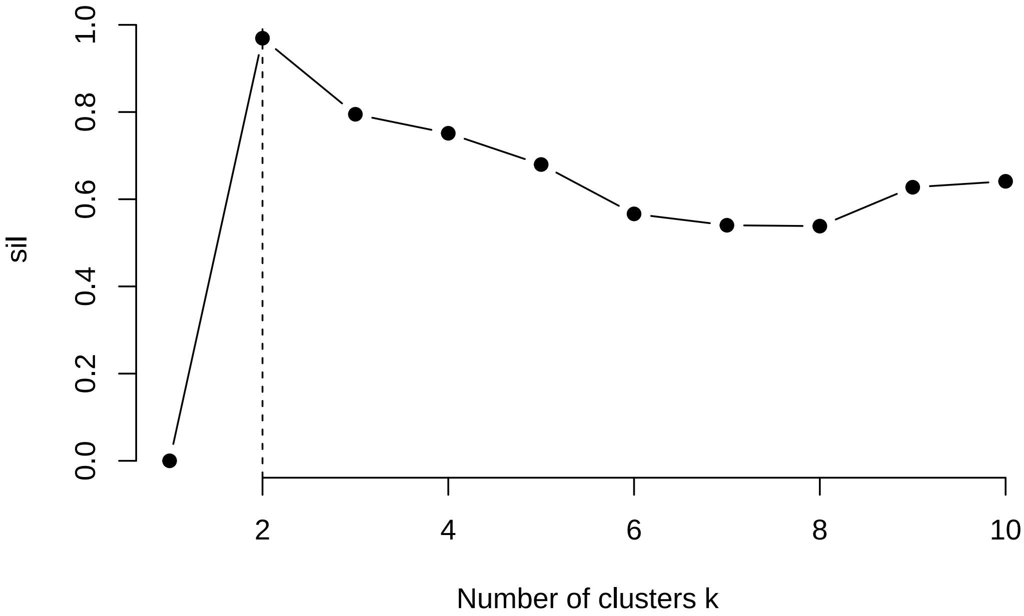
K-means clustering results show that the average silhouette width for six clusters is 0.677787 for 6 clusters. The diagram suggest that the best number of cluster in the data is 2 with an ASW of 0.9736113. India and China form part of their own clusters while other countries are in the second cluster. More than 80% of african countries fall in groups 4 and 5, around 70 % of Asian countries fall in cluster 4 and 5, 88% of North American countries fall in cluster 4, 87% of Oceania countries fall in clusters 4, 76 of south American Countries fall in cluster 4 and 6. This information seem to suggests that there are mainly two clusters that is 4 and 5. This is further supported by the ASW of 0.9736113 for two clusters.

3.3 K-Mediods



The ASW for six clusters is 0.610983 while that of 2 clusters is the maximum at 0.9736113. The information also suggest that the optimum number of clusters is 2. Over 90% of African countries belong to group 1, 2 and 3, 75% of Asian countries belong to group 1, 2 and 3. 86% of European countries belong to group 2 and 3, 91% of North American countries belong to group 2 and 3, 75% of Oceania countries belong to group 3 and more than 90% of South American countries belong to group 1, 2 and 3. These results show that there are at most three main clusters in the data and there is no specific regional pattern that can be seen. The PAM method also suggests that the optimum number of clusters in 2.

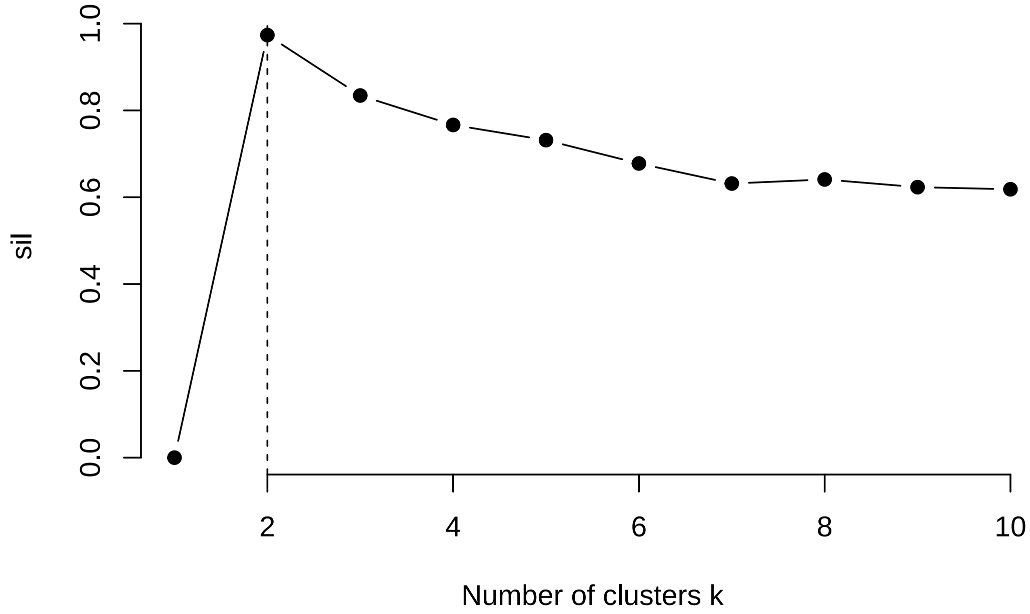
3.4 Clara

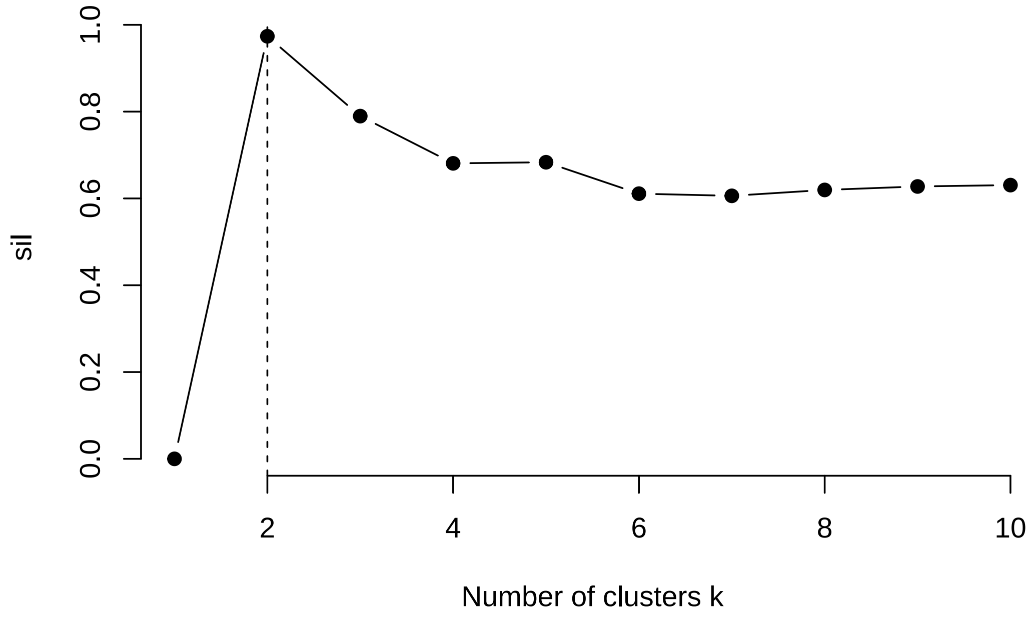


More than 91% of African countries are found in group 1, 2 and 3. More than 77% of Asian countries are found in group 1, 2 and 3. More than 76% of countries in Europe are found in groups 2 and 3. More than 91% of North American countries are found in groups 2 and 3. More than 87% of Oceanian countries are found in groups2 and 3 while more than 90% of South American countries are found in group 1, 2 and 3. Again, the information suggest that there are at most the main clusters in the data.

Density based scanning was tested and in order to get six main clusters on the data the epsilon value was set to 47 000 and the min points to 3. This AWS was -0.21 which indicates that the clusters are of poor quality and objects are less similar in their own cluster. This may well mean that the data either has too many of too few clusters. The DBSCAN method also pick up many outliers in the data and very few countries belonging to groups. This method will not be explored for further analysis given that it does not give any good results.

The randomizing, re-running, averaging and final run is a very good advice.

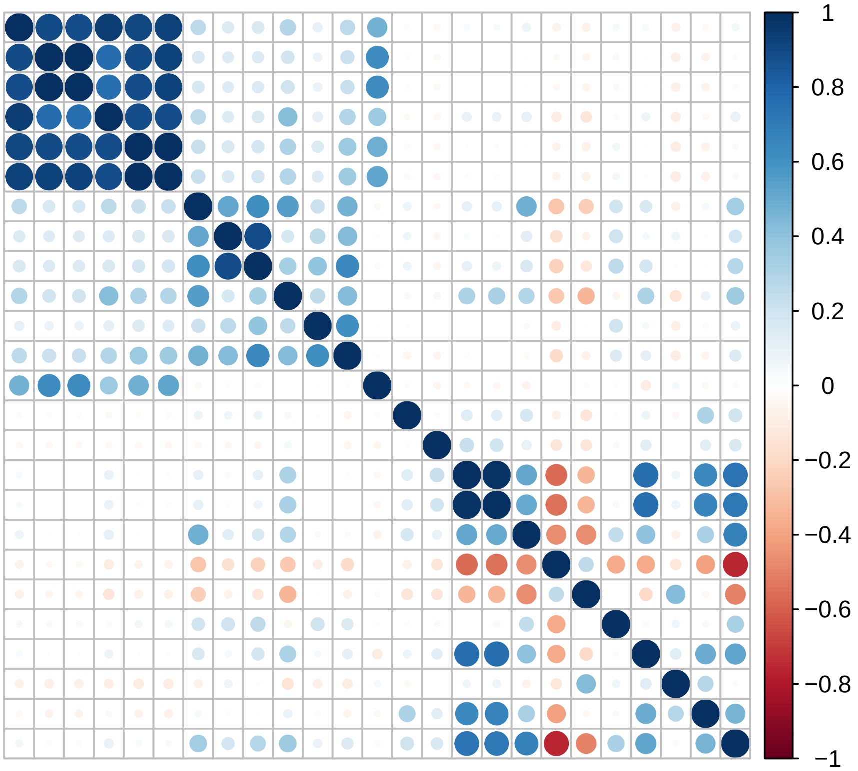




The dataset is very small and this results in Clara performing worse than PAM.

4 Dimension reduction

A correlation plot shows that a number of variables in the dataset are highly correlated. Therefore, the size of the dataset can be reduced but keeping most of the information in the dataset using dimension reduction techniques. In this exercise PCA will be used to reduce the size of the data and clustering will be performed on the reduced data.

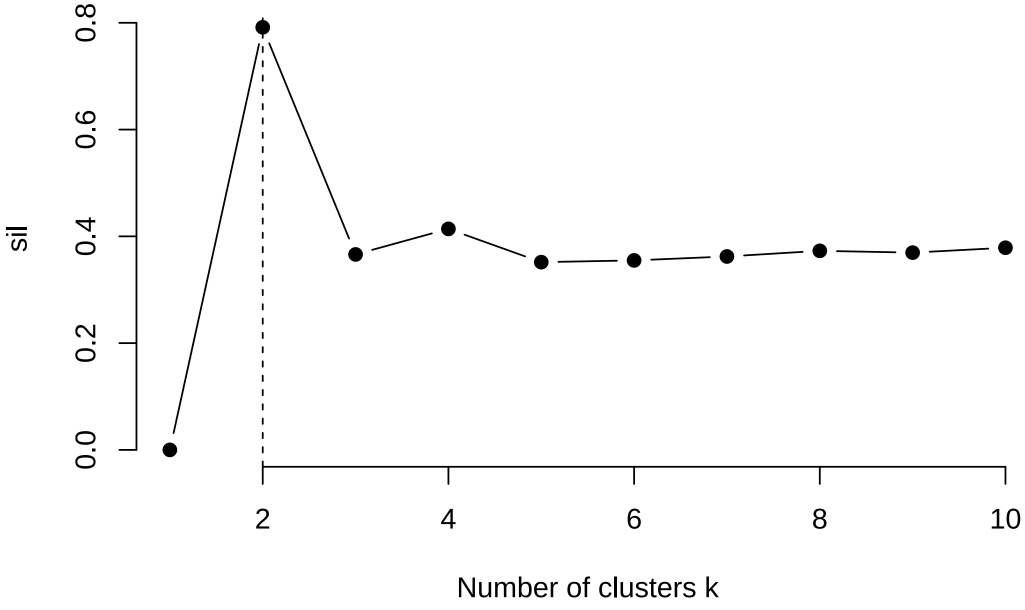


The scree plot shows that most of the variance in the data can be found in the first 4 principal componenents. The first 4 principal compnents were taken to represent the whole dataset.

The ASW for k-means clustering for all cluster numbers was found to be generally lower than that of the unreduced dataset. For 6 clusters, the ASW is 0.3549331 and for 2 clusters 0.7914952. The value for six clusters is very low to suggest any reasonable regional patterns given that the AWS is low.

Table 5: HCClust

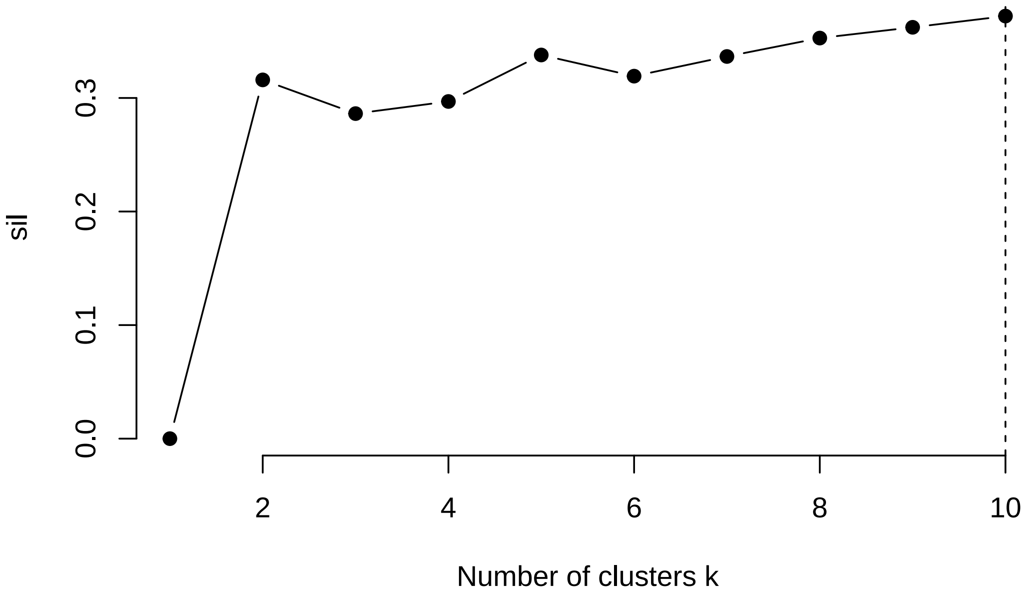
Method	Result
Complete	0.32
Single	0.50
Average	0.30
Median	0.09
Centroid	0.45



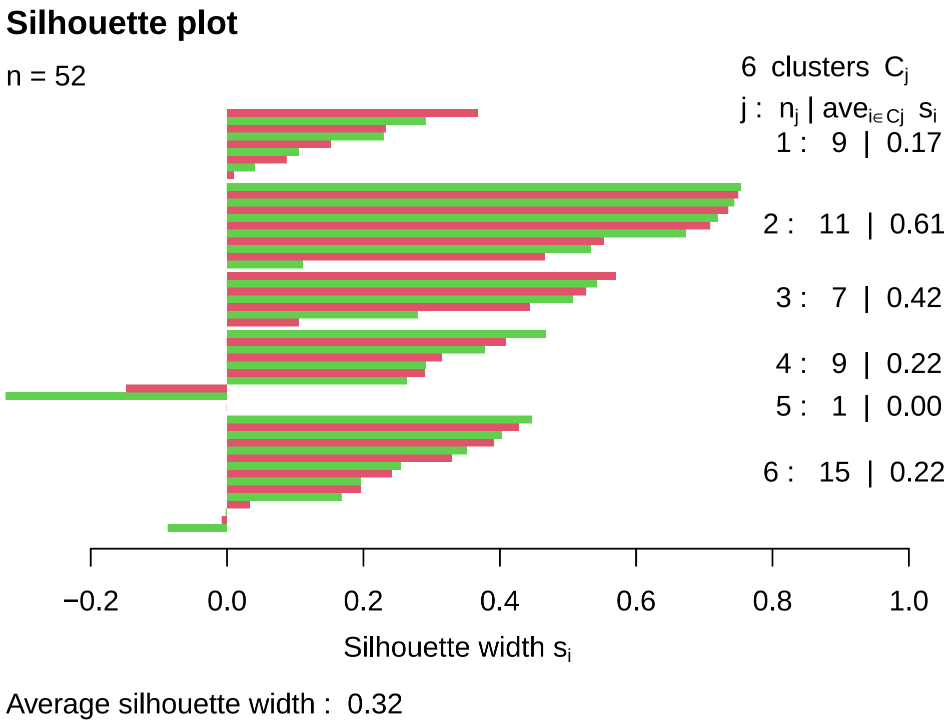
```
hc_results <- read_excel("HeirachicalResultsHmisc.xlsx")
knitr::kable(hc_results, digits = 5, caption = "HCClust") %>%
  kable_styling(full_width = F, font_size = 7) %>%
  column_spec(1, border_left = T) %>%
  column_spec(2, border_right = T)
```

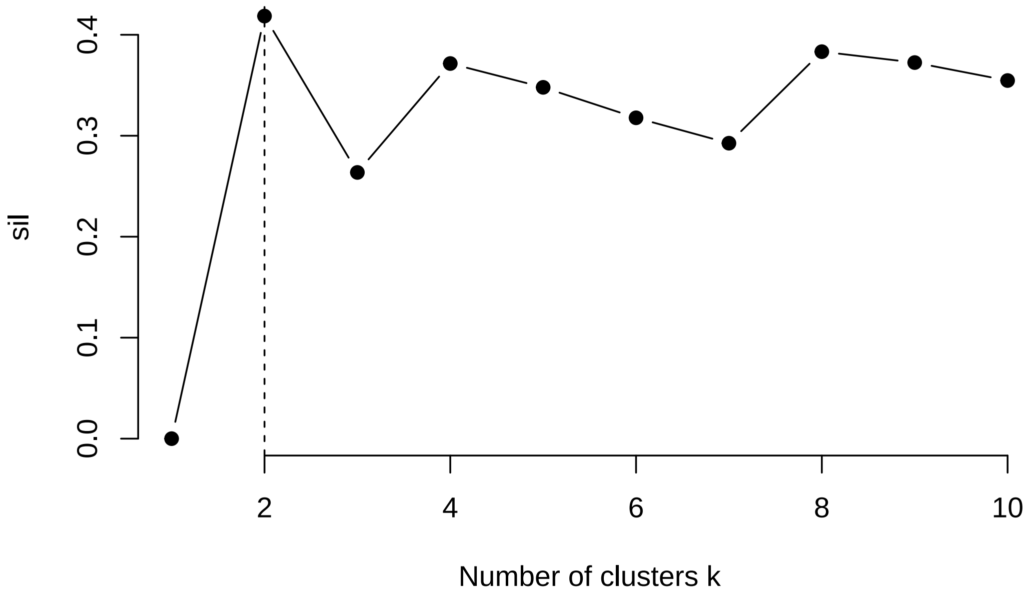
```
tt = tibble(Country = hc[["labels"]], Group = silhouette(cutree(hc_single,6),dist.out)[,1])
```

The AWS for all the heirachical clustering methods are less than or equal to 0.5 for 6 clusters. This also shows that the counties are less likely to be divided ino six continental regions thereby no evidence for regional patterns.



The K-mediods method gives an ASW of 0.32 for six clusters. It also suggests that there are probably more than six clusters for the data.





```
## [1] 0.3177602

## [1] 0.4185095

## [1] 0.1684077 0.6131996 0.4249841 0.2159589 0.0000000 0.2229429
```

The dataset is very small and this results in Clara performing worse than PAM.

In general, reducing the dimension of the data led to a worse performance in the clustering algorithm. Therefore, it is best to use the full dataset for clustering.

5 Conclusion

The results show that the best algorithm is xys The optimum number of clusters is The following countries are outliers Regional patterns on a continental scale do not exist in the data

This code was optimised for reading

References

CIA. 2020. *The World Fact Book*. https://www.cia.gov/library/publications/the-world-factbook/fields/343rank.html.

European Center for Disease Prevention and Control. 2020. *Coronavirus Source Data*. https://ourworldindata.org/coronavirus-source-data.

Karen Grace-Martin. 2011. *What Is the Difference Between Mar and Mcar Missing Data?* https://www.theanalysisfactor.com/mar-and-mcar-missing-data/.

The World Bank Data. 2020. *Life Expectancy at Birth*. https://data.worldbank.org/indicator/SP.DYN.LE00.IN.

Wikipedia. 2018. *List of Countries by Percentage of Population Living in Poverty*. https://en.wikipedia.org/wiki/List_of_countries_by_percentage_of_population_living_in_poverty.

———. 2020a. *List of Countries and Dependencies by Population Density*. https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density.

———. 2020b. *Population Aged 65 or Above*. https://en.wikipedia.org/wiki/List_of_countries_by_age_structure.