University of Cape Town

Department of Statistical Sciences

STA5077Z Unsupervised Learning

Assignment 1

EVRCHA001 - Charl Everts

19 October 2020

Department of Statistical Sciences

Plagiarism Declaration Form

A copy of this form completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will not be marked.

COURSE CODE: STA5077Z

COURSE NAME: Unsupervised Learning

STUDENT NAME: Charl Everts

STUDENT NUM: EVRCHA001

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed and has been cited and referenced.
3. This tutorial/report/project is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University's plagiarism rules (http://www.uct.ac.za/uct/policies/plagiarism_students.pdf).

Signature: Date: 19 October 2020

Table of Contents

List of Figures

List of Tables

Abbreviations/Symbols

COVID-19	novel coronavirus 2019
PAM	partition around medoids
CLARA	clustering large applications
GDP	gross domestic product
DBSCAN	density-based spatial clustering of applications with noise
PCA	principal component analysis

Abstract

The Coronavirus (COVID-19) Pandemic has swept across the world and is present on every continent on earth. Almost a year has passed since the first case was reported in Wuhan, China, and still very little is known about the novel virus. Scientists and epidemiologists are constantly trying to better understand the virus, especially factors relating to its transmission, fatality rates and regional patterns.

Thus far, the virus has infected at least 40 million people and caused the death of 1.1 million. There is currently no vaccine, with many countries currently experiencing their second wave of infections and reinstating strict lockdown measures.

The objective of this report is to evaluate the dataset provided by conducting partitional and hierarchical clustering analysis to conclude whether or not regional COVID-19 patterns exist. The explicit goals of the report are as follows:

i. Perform an exploratory analysis and report on the general distribution of the variables and report on missing values.

ii. Address issues regarding missing values and justify the chosen approach to handling these values.

iii. Perform cluster analysis on the data and group the countries into 6 clusters, paying special attention to what these clusters indicate about regional patterns of COVID-19.

iv. Produce a summary of the analysis and conclusions.

Chapter 1

Data Pre-Processing

1.1 The Dataset

The research conducted in this project is based on the *Coronavirus Pandemic (COVID-19)* dataset from the popular *Our World in Data* website. The dataset is compiled by a team from the University of Oxford. The website has several interactive map and graphing features which can be used to visualise and better understand all aspects of the COVID-19 pandemic.

The data contained is updated until 2 September 2020 and encapsulates 29 variables for 208 countries. The variables range from COVID-19 specific attributes such as number of cases and deaths to socio-economic factors like extreme poverty, GDP per capita and median age. A full list of variables can be found in the Appendix.

1.2 Data Quality Assessment

The most pressing issue when dealing with the dataset is finding appropriate measures that deal with the vast number of missing values.

1.2.1 Dropped Columns

All the variables relating to the short-term progression of the virus was dropped from the dataset. The reasoning was that all these variables relating to “new cases/new deaths” would not fully encapsulate the overall trend of the virus in the long run. Instead, it focuses on recent developments, which is not entirely useful when investigating worldwide trends since the onset of the pandemic. Additionally, the iso_code and date columns were dropped as it offered no further insight.

1.2.2 Dropped Rows

The dataset was thoroughly inspected to justify the further dropping of entries to clean up the data. It was noted that countries with 5 or more missing values either have extremely small populations or a recent history of political turmoil or civil wars. These are countries like Syria, Western Sahara and South Sudan.

The justification for this is that smaller countries would offer less insight into COVID-19 trends due to their extremely small geographic and population size. Additionally, it is difficult to trust the statistics from war-torn countries. It is nearly impossible to accurately state the number of tests, cases and deaths for such countries. It would furthermore prove to be foolish to attempt value imputation for these states. It would be tantamount to guessing and would guarantee the addition of unwanted noise. These rows were therefore removed before conducting analysis.

1.2.3 Missing Value Imputation

The main strategy in dealing missing values after the initial dropping of rows and columns was to conduct either mean/median value imputation or implement deterministic and stochastic regression imputation.

Deterministic regression imputation replaces missing values with the prediction of the regression model chosen. New values are often too precise and lead to an overestimation of the correlation between the response and predictor variables chosen.

Stochastic regression imputation solves this issue. Stochastic regression adds a random error term to the predicted value and is, therefore, able to reproduce the correlation of the response and predictor variables more appropriately. Both of these methods are generally accepted as being a better alternative to median or mean value imputation.

Both types of regression imputation were achieved using the *mice* package.

1.2.4 Extreme Poverty

To conduct regression imputation on the missing values in the extreme_poverty column, a suitable linear regression has to be identified. Intuitively, GDP_per_capita should be a strong indicator of a country's extreme poverty. More precisely, a very strong inverse relationship should exist between a country's GDP_per_capita and extreme_poverty, as the former is often a strong indicator of a country's wealth and hence the general standard of living. To confirm or deny this hypothesis, a simple linear regression is conducted:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
GDP_per_capita	-0.0006406	0.0001004	-6.378	3.73e-09

Table 1.1: *Summary of simple linear regression on extreme_poverty.*

According to the regression summary, GDP_per_capita is indeed a highly significant predictor of extreme_poverty. However, two missing values are present in the GDP_per_capita column, namely the entries for Cuba and Somalia. The missing value for Cuba was replaced with its latest GDP_per_capita, as per the WorldBank statistics. Somalia proved slightly trickier, as the latest value was from 2010. Instead of using this value, Somalia's missing value was replaced by the average GDP_per_capita of Sierra Leone and Mozambique, since Somalia is consistently ranked between these countries in economic performance.

It is now possible to conduct regression imputation of the missing values. The results from the deterministic regression model were deemed the most appropriate in this case, and the output of 54 missing values (Figure 1.1) was subsequently replaced with these results.

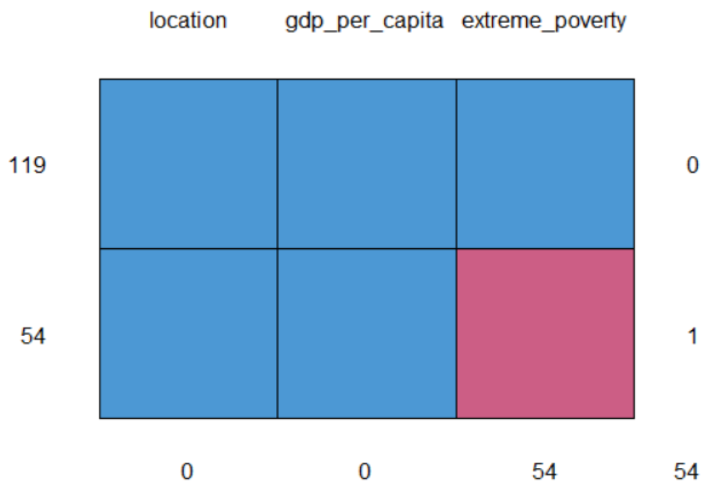


Figure 1.1: *Missing value plot for extreme_poverty.*

1.2.5 Handwashing Facilities

Handwashing facilities is a metric which measures the accessibility of ablution and sanitising facilities to the general public. The value given is a percentage, with 100 meaning handwashing facilities are available to all citizens. The importance of this metric stems from the requirement of proper sanitization of individuals to help curb the spread of COVID-19. One can therefore make the argument that countries with lower levels of handwashing facilities are at risk of an increased chance of virus transmission.

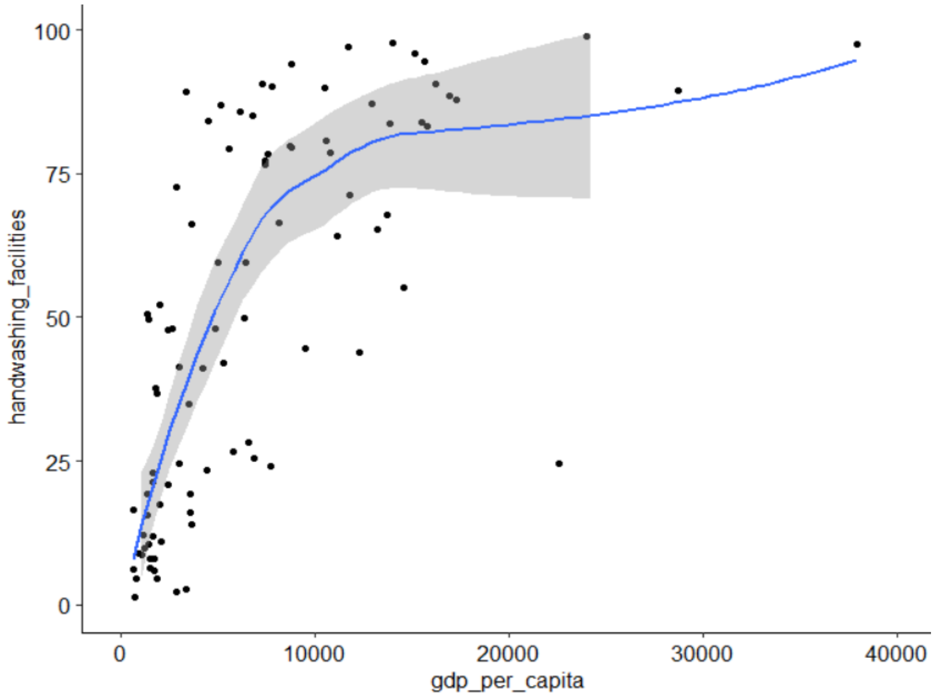


Figure 1.2: *Handwashing facilities as a function of GDP_per_capita.*

Again, GDP_per_capita was used to predict handwashing_facilities, since intuitively they should be closely linked. This relationship confirmed to some extent by Figure 1.2.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
GDP_per_capita	0.003143	0.000388	8.102	2.87e-12

Table 1.2: *Summary of simple linear regression on handwashing_facilities.*

Table 1.2 confirms that GDP_per_capita is highly significant in predicting a country's handwashing facilities metric. Furthermore, there exists a reasonably strong correlation coefficient of 0.654 between the two variables.

However, after conducting mean, stochastic, and deterministic regression on the missing values, none were deemed appropriate. All three methods yielded unrealistic values for the vast majority of countries. For example, Sweden and Libya were granted the same value of 50.5%, when this is objectively unrealistic. Meanwhile, both regression methods yielded results far exceeding 100%, like Luxembourg that yielded a value of 324%. Numerous such examples indicate the inaccuracy of these imputation methods.

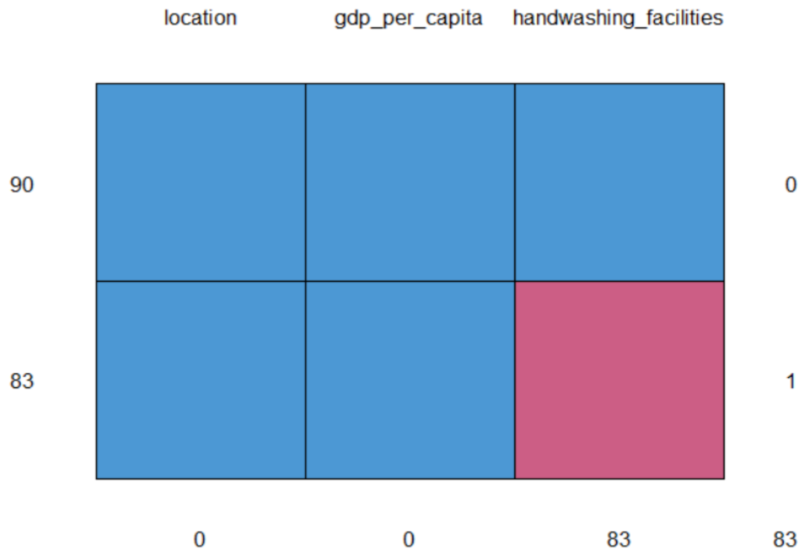


Figure 1.3: *Missing value plot for handwashing_facilities.*

In addition to the vast number of missing values in the handwashing_facilities column, a strong argument exists in support of excluding the column. The presence of handwashing facilities does not guarantee that members of the society will use them. Also, simply having these facilities is not entirely sufficient in curbing the spread of transmission. These facilities also have to be stocked with alcohol-based hand sanitizing materials that may not be readily available in all of the countries represented.

This fact, in conjunction with the almost 50% missing values (Figure 1.3) for the handwashing_facilities column, is deemed sufficient justification to drop the column from the dataset.

1.2.6 Hospital Beds

The importance of the hospital beds per thousand variable stems from the fact that COVID-19 has sent a shockwave through both the public and private sector healthcare systems. This in turn has lead to unprecedented levels of stress on hospital staff and infrastructure. One of the second-order effects of COVID-19 is the potential of overloading the healthcare system and has necessitated the temporary expanding of healthcare systems across the world.

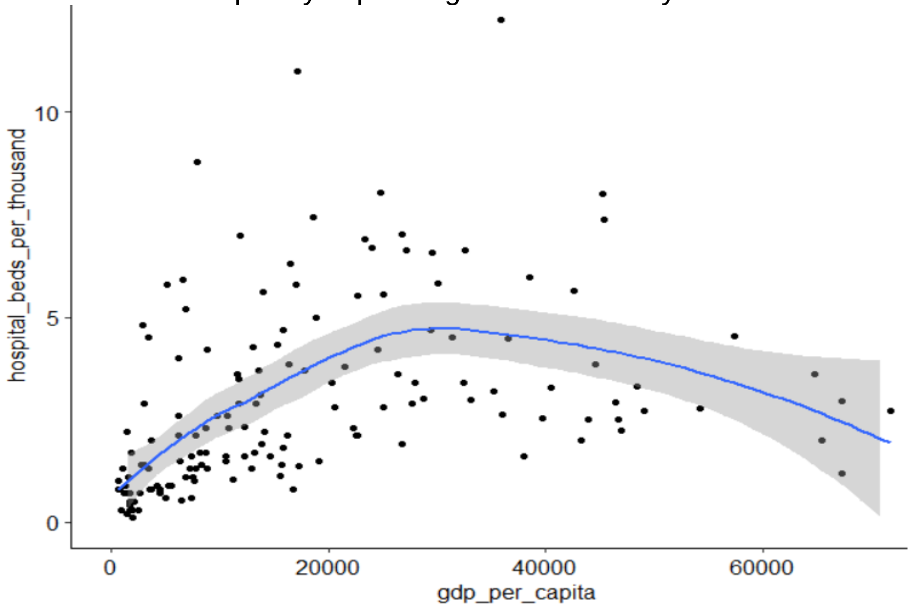


Figure 1.4: *Hospital beds per thousand as a function of GDP_per_capita.*

It is therefore believed that having a high value for hospital beds per thousand lowers the potential risk of overloading the healthcare system of a country. Figure 1.4 reveals no obvious relationship between these to variables, however Table 1.3 indicates that GDP_per_capita is a reasonably strong predictor of hospital beds per thousand.

	Estimate	Std. Error	t value	Pr(>|t|)
GDP_per_capita	3.519e-05	9.087e-06	3.873	0.000158

Table 1.3: *Summary of simple linear regression on hospital beds per thousand.*

The deterministic regression imputation yielded the most realistic results for the hospital beds per thousand column.

1.2.7 Male and Female Smokers

The smoking data for each country represents the percentage of people that smoke per gender. The reasoning behind this variable is that initially, doctors had postulated that smokers might be at a greater risk of death due to a weakened respiratory system. Figure 1.5 indicates 72 missing values across the two columns.

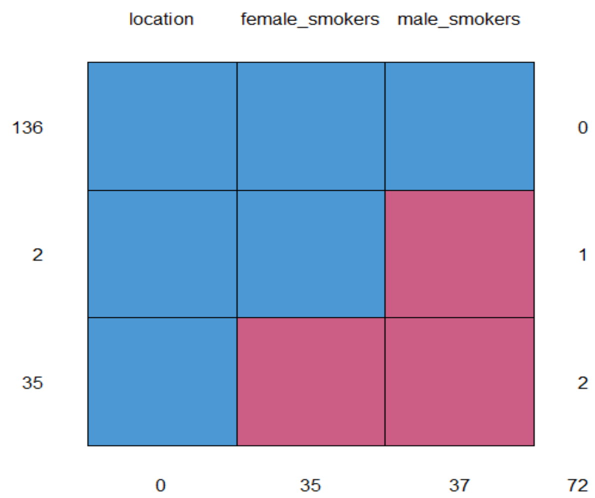


Figure 1.5: *Missing value plot for male and female smokers.*

Mean value imputation was deemed the appropriate strategy to replace missing values. But instead of calculating the worldwide smoking mean value by gender, the smoking values of each continent were calculated by gender. For example, instead of replacing South Africa's smoking statistics with the worlds mean values, it received the mean value from Africa's data. This proved to be more accurate as smoking data by continent varies greatly (Table 1.4). Across the board, mean and median are located close to each other, indicative of symmetric distribution of values. In each instance, the mean is larger than the median. This indicates that the data are skewed to the right.

Continent	Female		Male	
	Mean	Median	Mean	Median
Africa	2.59	1.60	27.43	24.60
Asia	4.81	2.75	39.70	38.30
Europe	23.15	22.95	35.64	35.20
North America	6.93	5.30	22.31	19.75
Oceania	15.38	13.90	29.33	26.00
South America	10.93	7.40	24.66	20.75

Table 1.4: *Summary of central tendency of smoking data by gender and continent.*

However, it was noted that the distribution of smoking by gender and continent often contained substantial outliers. In this case, the missing value was replaced by the median value instead of the mean as it is less susceptible to outliers. However, if the mean and median were similar, the mean was used.

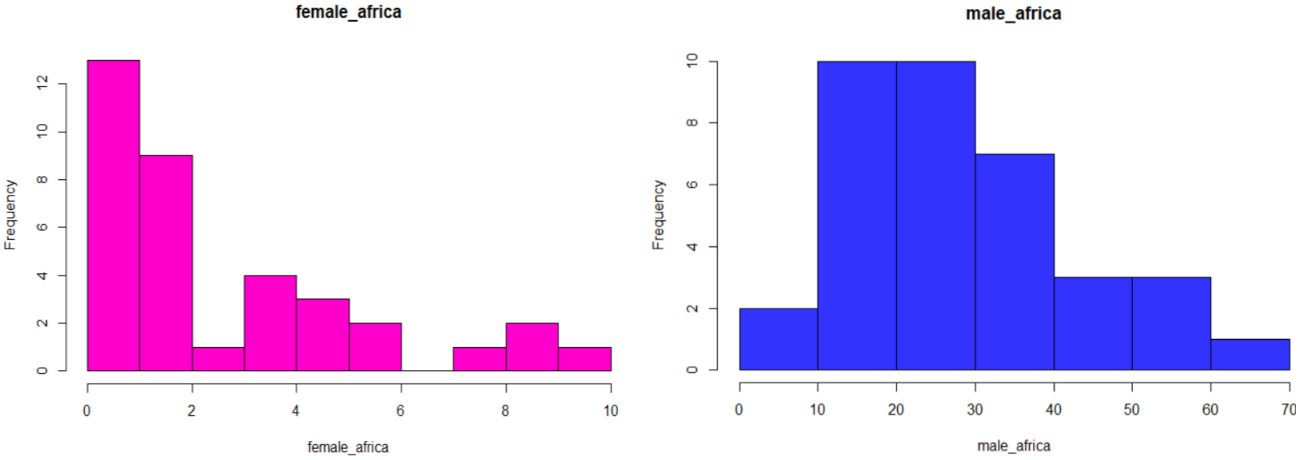


Figure 1.6: *Frequency distribution of female and male smokers in Africa.*

In the case of African smoking data, the female smoker data had similar mean and median values, hence mean-value imputation was conducted. Whilst for male smokers, the mean was larger than the median due to a few outlier countries, hence median-value imputation was conducted. This process was repeated for all continents.

1.2.8 Serbia: aged_70_older

The final missing value was the Serbian entry in the aged_70_older category. The reasoning for the aged_65_older and aged_70_older variables was that increased death rates were associated with the elderly. So, an increased elderly population could result in a greater deaths_per_million value for a country.

The missing value was replaced based on the premise that the proportion of 65-year-olds to 70-year-olds across the world would stay relatively consistent. Hence the worldwide average 65-year-old to 70-year-old ratio was found and applied to the Serbian 65-year-old data.

1.3 Outlier Detection

Due to the large amount of feature in the dataset, it is difficult to identify which observation is indeed an outlier since there are many variables to interpret. Instead, outliers were identified during the clustering process. During partitional clustering, the following four observations repeatedly formed either their clusters, or clustered with each other: US, Brazil, India and Singapore. See section 3.1.

During the hierarchical clustering, six observations repeatedly behaved like outliers, as described in section 3.4: US, Brazil, India, Singapore, Qatar and China.

1.4 Descriptive Statistics

Since the goal of the project is to conclude if regional patterns exist in the COVID-19 pandemic, it makes sense to analyse the data by continent. Boxplots were constructed for the major variables like cases (Appendix), deaths (Appendix) and deaths per million (Figure 1.7). After plotting this data, it was evident that the data required scaling. Feature scaling is a method used to normalize the range of independent variables.

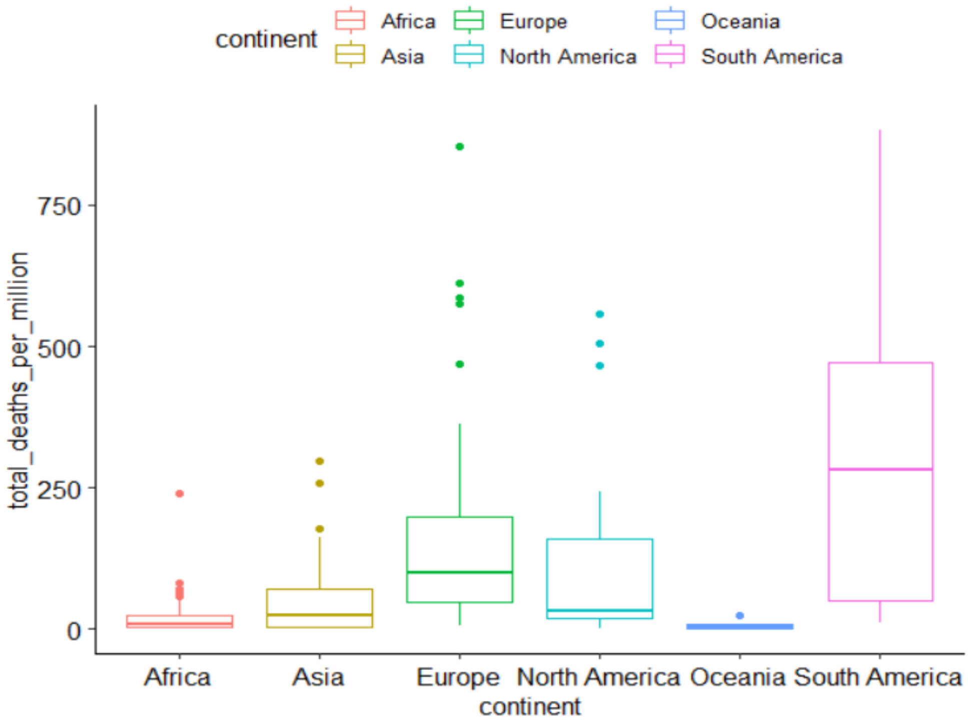


Figure 1.7: *Boxplot of deaths per million by continent.*

A five-number summary is depicted in Figure 1.8. It is interesting to note the range of the four main variables across the dataset. It suggests that although the coronavirus pandemic is widespread, the actual severity thereof is vastly different in different countries. For example, some African countries have been barely affected, whilst countries like the US and Brazil have been shocked by the outbreak.

The main reason for this may be explained by the fact that countries were able to develop policies and prepare lockdown procedures before the virus was present in their country. Take South Africa for example; our government was able to learn from the mistakes made in Italy and France. As soon as our case numbers started to peak, we were forced into quarantine, whereas those countries did not.

	Min	Med	Mean	SD	Max
total_cases	22.0	10524.0	145900.1	625119.0	6075652.0
total_deaths	0.0	203.0	4779.9	18826.7	184689.0
total_cases_per_million	3.0	1469.5	3874.8	5806.2	41302.2
total_deaths_per_million	0.0	30.6	97.0	161.5	881.6

Figure 1.8: *Five-number summary of the four main COVID-19 features.*

Figure 1.9 is a correlation heatmap for each variable in the dataset. Dark blue blocks correspond to 2 sets of variables that are strongly positively correlated, and dark red corresponds to strongly negatively correlated. An interesting observation is that extreme poverty and cardiovascular death rate have a strong negative correlation with total deaths per million. This seems counter-intuitive, as it was initially believed that increased levels of poverty and heart disease in a country would lead to a greater number of COVID-related deaths.

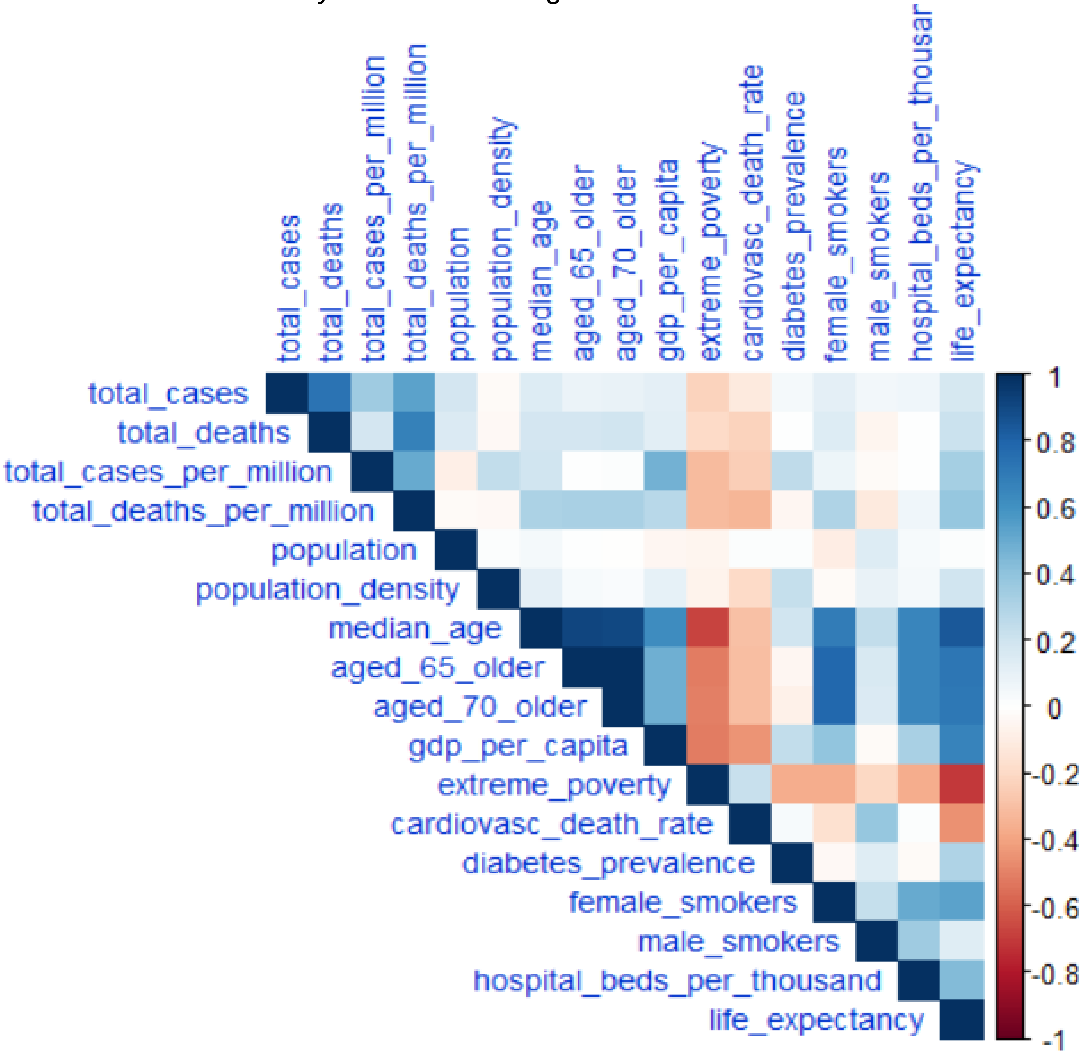


Figure 1.9: *Correlation heatmap.*

Chapter 2

Methodology

2.1 Partitional Clustering

In partitional cluster analysis, the objective is to group objects that share similar features, such that a larger set of objects is divided into smaller subsets. The objects in the resulting subsets will be more similar to other objects in that set, than to objects in other subsets.

Since a response variable is not defined, this is an unsupervised method, which implies that it seeks to find relationships between observations without being trained by a response variable. Clustering allows us to identify which observations are alike and categorize them accordingly.

2.1.1 K-means

K-means clustering is the simplest and the most commonly used clustering method for splitting an unlabelled dataset into a pre-determined set of k-groups.

The algorithm aims to group the intra-cluster data points such that they are as similar as possible, whilst ensuring clusters remain dissimilar to each other. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is minimized. The less variation that exists within clusters, the more homogeneous the observations within each cluster. The method was introduced by MacQueen in 1967.

2.1.2 K-medoids | PAM

The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other the members of the cluster are minimised. It corresponds to the most centrally located point in the cluster. As opposed to k-means clustering, where the centre of a given cluster is the mean value of all observations in a cluster.

K-medoids is a robust alternative to k-means as it is less sensitive to noise and outliers. The most common k-medoids clustering methods is the Partitioning Around Medoids (PAM) algorithm popularised by Kaufman and Rousseeuw in 1990.

2.1.3 CLARA

CLARA stands for *clustering large applications* and is an extension of the work done by Kaufman and Rouseeuw. In CLARA, each subset is partitioned into k clusters, using the same algorithm as PAM. Once k representative objects have been selected from the subset, each observation assigned to its nearest medoid.

The mean of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. Each subset contains the medoids obtained from the best sub-dataset.

2.2 Hierarchical Clustering

The divisive clustering algorithm is a top-down clustering approach, initially, all the points in the dataset belong to one cluster, with splitting performed recursively as one moves down the hierarchy. The result is a tree-based representation of the objects called a dendrogram.

Agglomerative Clustering is a bottom-up approach, initially, each data point is a cluster of its own, further pairs of clusters are merged as one moves up the hierarchy. Pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

2.3 Density-Based Clustering

The density-based spatial clustering of applications with noise (DBSCAN) is a popular alternative to the aforementioned partitional and hierarchical clustering algorithms. It does not require a pre-determined number of clusters and works especially well with arbitrarily shaped clusters. The clusters are formed by linking neighbour points together. DBSCAN does however tend to struggle when classifying high-dimensional data, as mentioned in Chapter 3.

Chapter 3

Results

3.1 K-means

To use the K-means algorithm, one needs to specify the number of clusters required in the output. The objective of this report is to find potential continental clusters, hence six clusters (Africa, Asia, North America, South America, Oceania, Europe). However, optimal cluster tests were conducted to gauge the *actual* number of clusters suggested by the algorithm (Figure 3.1).

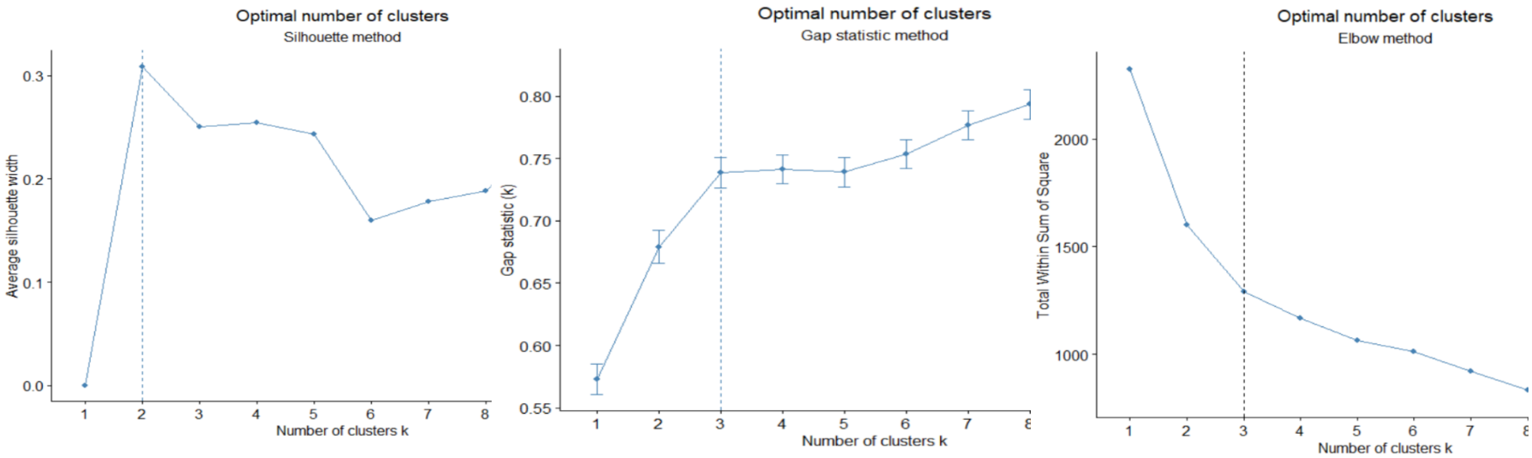


Figure 3.1: *Silhouette, Gap Statistic and Elbow methods to find optimal number of clusters.*

According to Figure 3.1, the recommended number of clusters are two, three and three respectively. It was therefore decided that each of the partitioning algorithms would create models that cluster two, three, four, five and six clusters each. The silhouette widths would then be compared and the optimal model chosen.

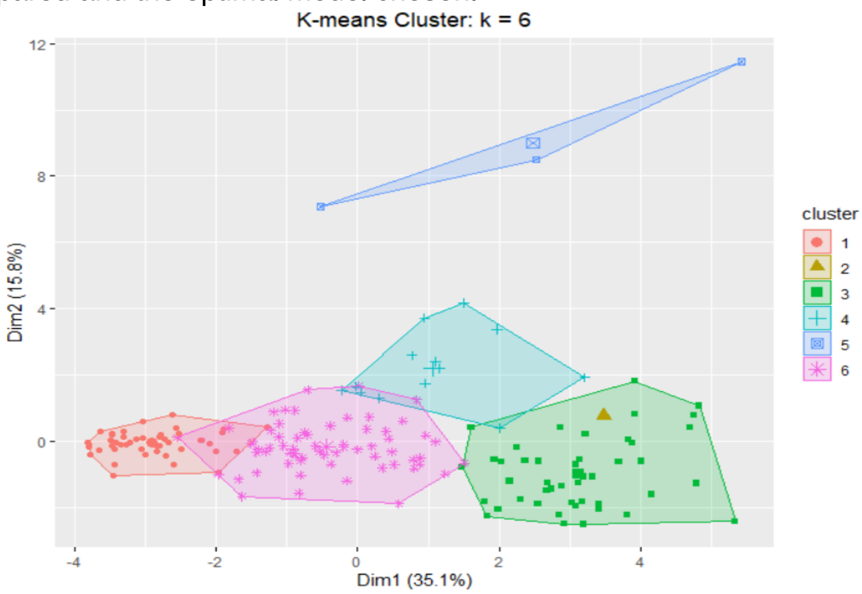


Figure 3.2: *K-means cluster with k = 6.*

Figure 3.2 depicts the final output from the first K-means model with six clusters. The clusters seem well-formed except for clusters two and five that appear to be outliers.

Upon further inspection, it was found that these four outliers were the USA, Brazil, India and Singapore. USA, Brazil and India have been arguably the worst affected COVID-19. Their cases and total deaths are astronomical compared to other countries. While Singapore has proven one of the countries most capable of dealing with the outbreak. It has one of the lowest mortality rates despite being near where the outbreak originated in China. These four observations were removed from the dataset. The new K-means cluster plots are shown in Figure 3.3.

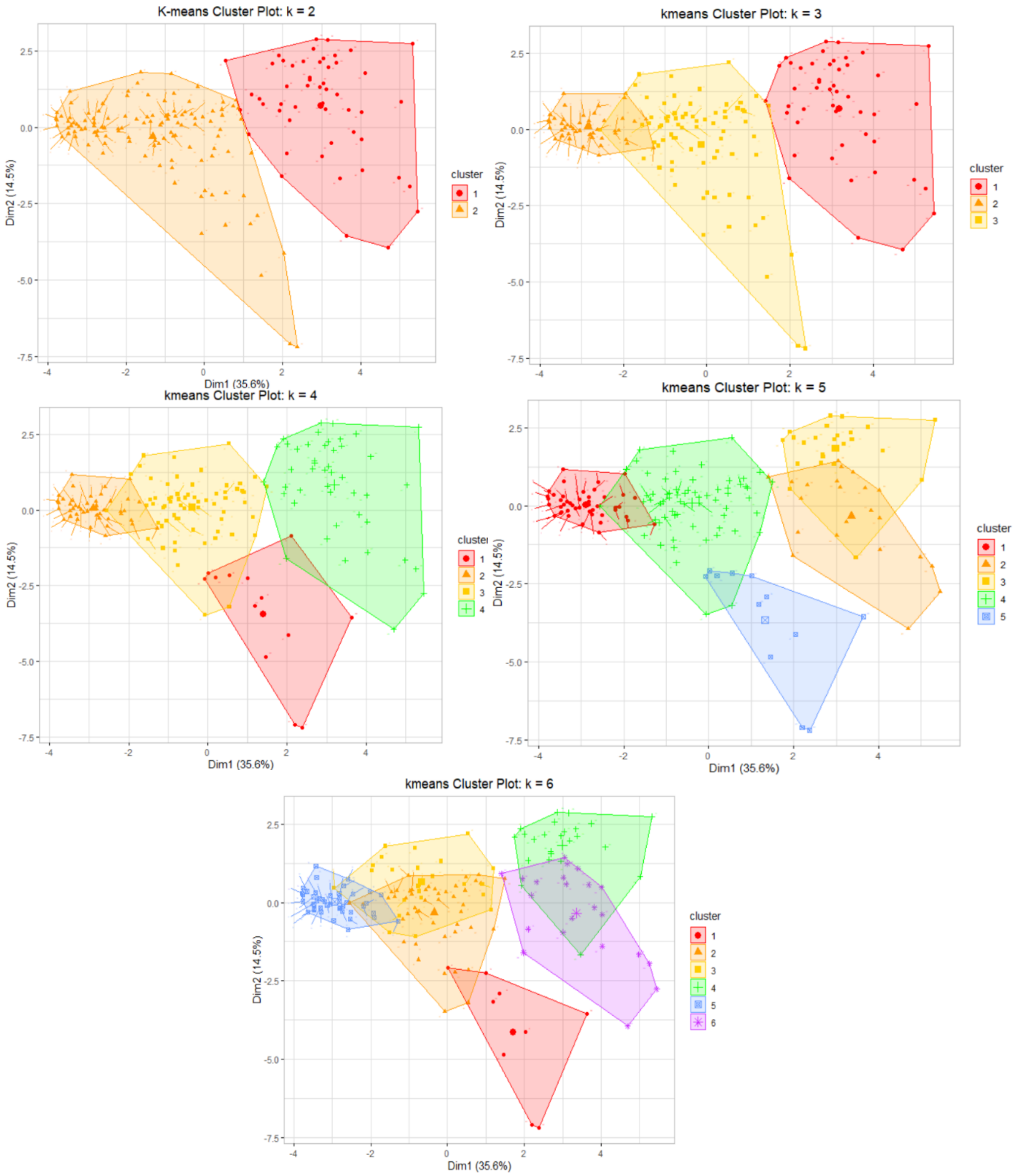


Figure 3.3: *K-means cluster plots for k = 2,3,4,5,6.*

Figure 3.3 visualises how the clusters are transformed between the two and six cluster plots.

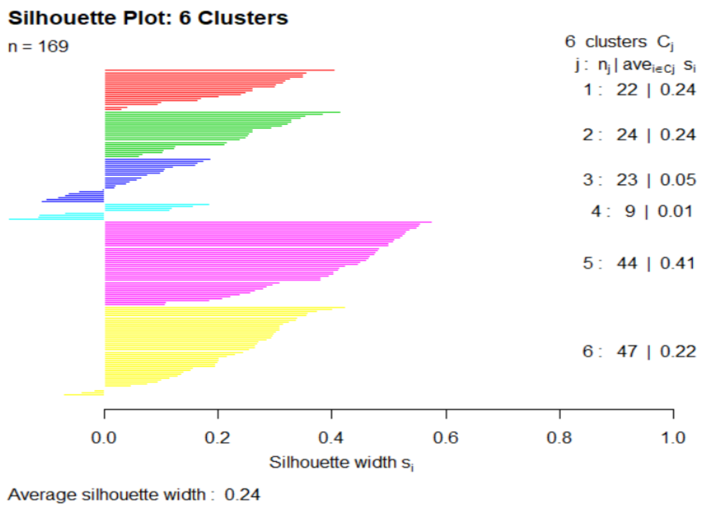


Figure 3.4: *Silhouette plot for 6 cluster K-means.*

A silhouette width in this range is generally ranked as having a weak structure and that the intended cluster structure – regional patterns of COVID-19 – are artificial. The silhouette widths for the k = 2,3,4,5 plots were marginally better, but only the k= 6 cluster model was considered for the sake of the report's objective.

3.2 PAM

The PAM algorithm was conducted and the following six-cluster plot was created (Figure 3.5). The resulting clusters are extremely similar to the shapes found in the K-means algorithm. As with the k-means algorithm, the k = 2,3,4,5 clusters had marginally better silhouette widths but were not considered as they offered no further insight to the report's objective.

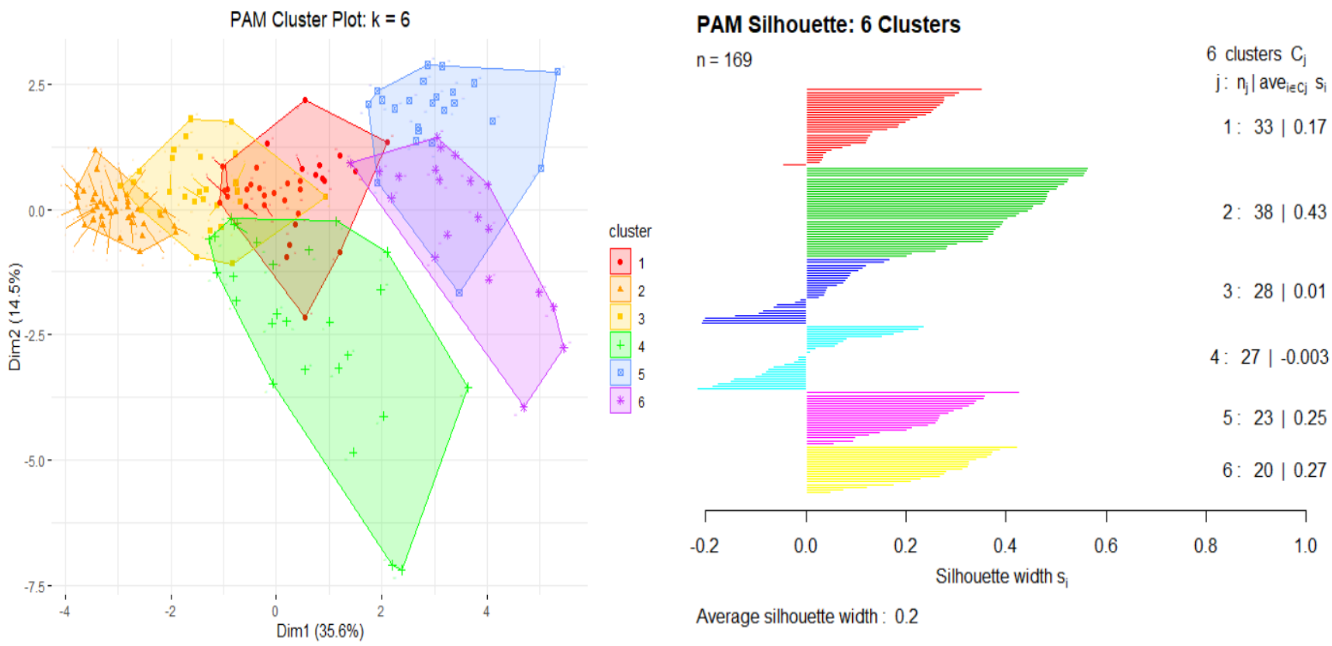


Figure 3.5: *PAM cluster plot and silhouette widths.*

3.3 CLARA

The CLARA algorithm similarly clustered the 169 countries to that of K-means and PAM. The shapes and position of the clusters are virtually identical, although CLARA had a marginally better average silhouette width than PAM, but worse than k-means for their respective six-cluster plots.

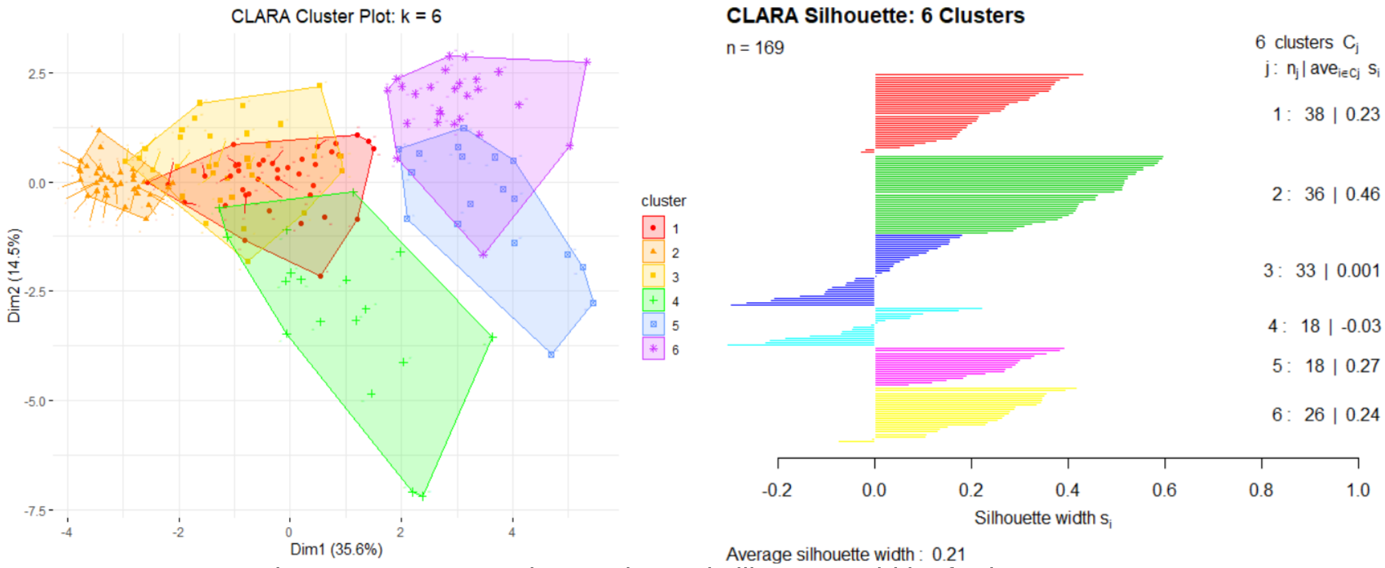


Figure 3.6: *CLARA cluster plot and silhouette widths for k = 6.*

3.4 Hierarchical

The best performing hierarchical clustering algorithm was the complete linkage version, whose dendrogram is depicted in Figure 3.7. Although it was the best performing hierarchical algorithm, it yielded rather poor average silhouette widths of 0.19, which ranks as having virtually no structure whatsoever. The labels in each cluster were removed for the sake of neatness, but can be found in the Appendix.

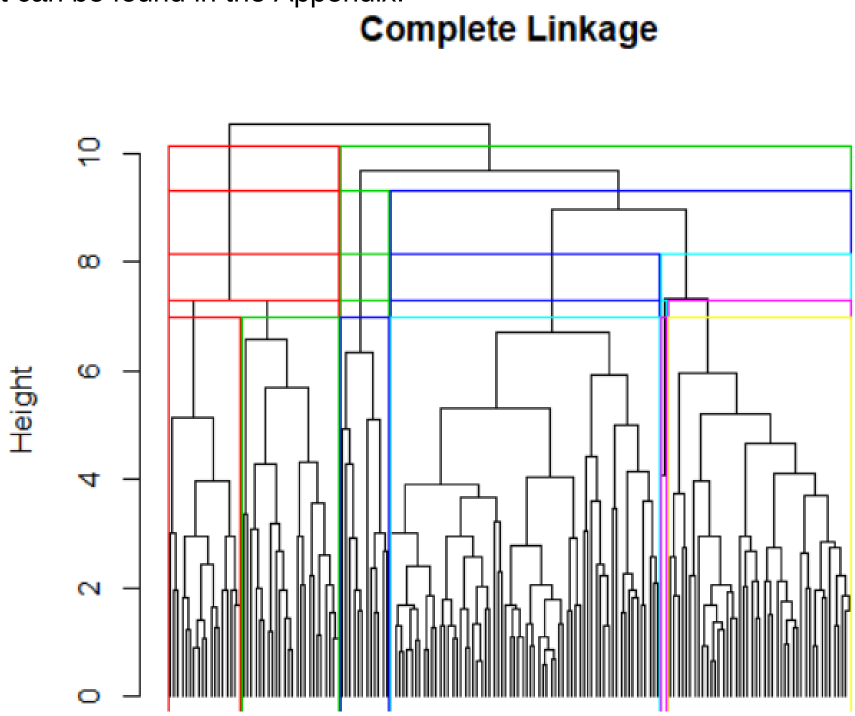


Figure 3.7: *Complete linkage hierarchical clustering dendrogram.*

3.5 DBSCAN

Unfortunately, the DBSCAN algorithm failed to yield intelligible results. After many attempts, it continued to group each country in the same cluster, regardless of changing the `eps` and `minPts` parameters.

3.6 Principal Component Analysis

Principal component analysis was conducted to see if improvements could be made through further dimensionality reduction. Figure 3.8 is a plot indicating the contribution of variables to the two most important dimensions – essentially a variable importance plot.

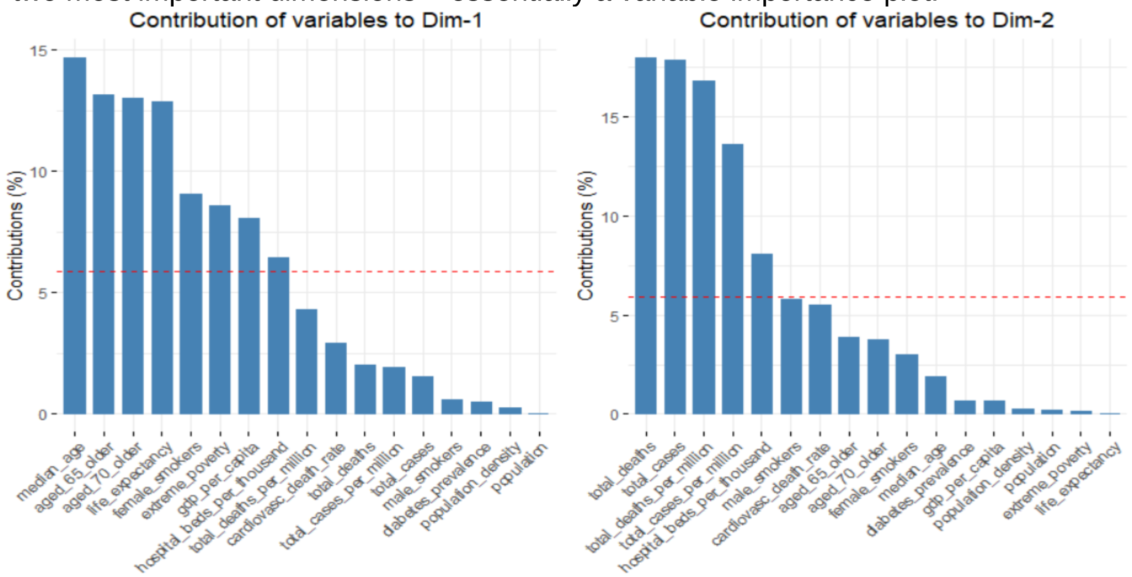


Figure 3.8: *Individual dimension variable importance plot.*

Figure 3.9 is a dual variable importance plot that depicts the contribution of each variable to the first two dimensions considered. Unfortunately, after multiple attempts at dimensionality reduction, the new dataset failed to improve the average silhouette widths in the respective algorithms, hence the old data set was used instead.

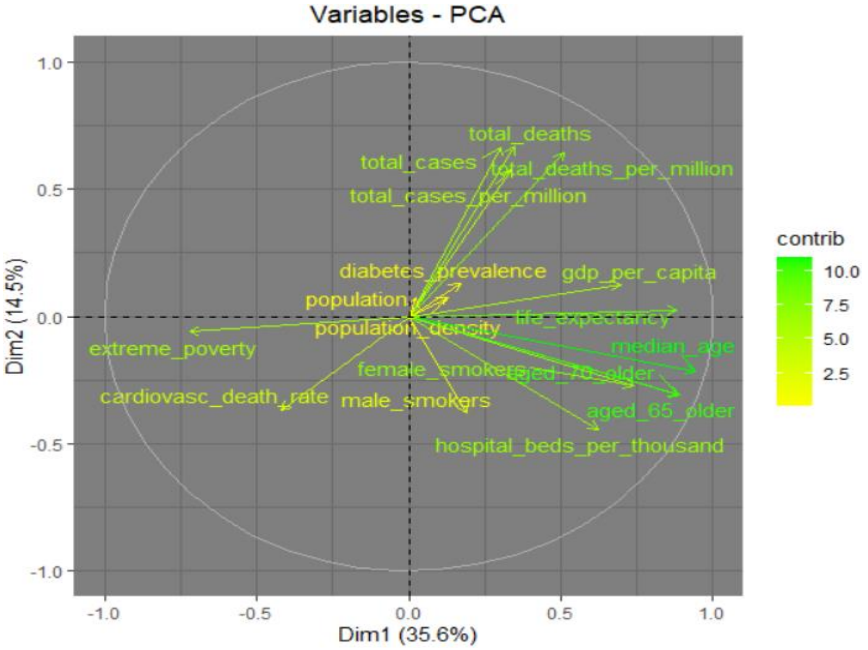


Figure 3.9: *Dual variable importance plot.*

Chapter 4

Conclusion

In this report, partitional and hierarchical clustering algorithms were implemented using the COVID-19 dataset available from the “Our World in Data” online repository. The main objective was to establish whether or not regional patterns exist in the COVID-19 outbreak. The data contained information on 208 countries and 29 variables up to the 2nd of September 2020.

Two main strategies were implemented in dealing with missing values. The first was mean and median value imputation. The second was a slightly more sophisticated, stochastic and deterministic regression imputation. After conducting data pre-processing, the dataset contained 169 entries and 17 variables. Table 4.1 ranks the algorithms based on average silhouette width.

Clustering Algorithm	Average Silhouette Width
K-means	0.24
CLARA	0.21
PAM	0.20
Complete Linkage	0.19
DBSCAN	N/A

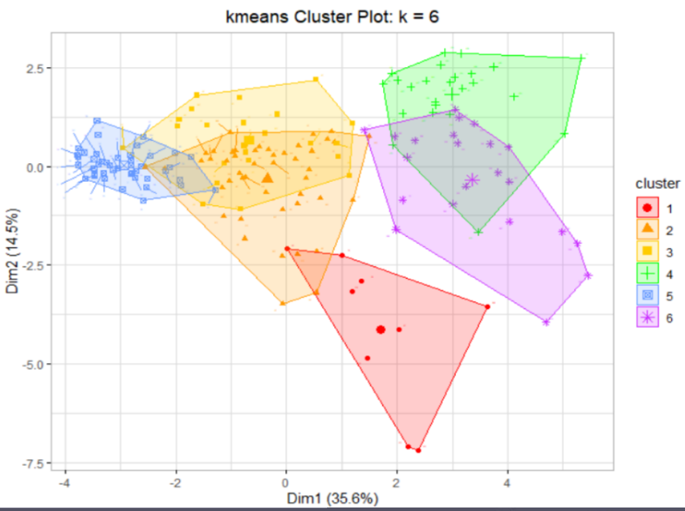
Table 4.1: *Clustering algorithms ranked by average silhouette width.*

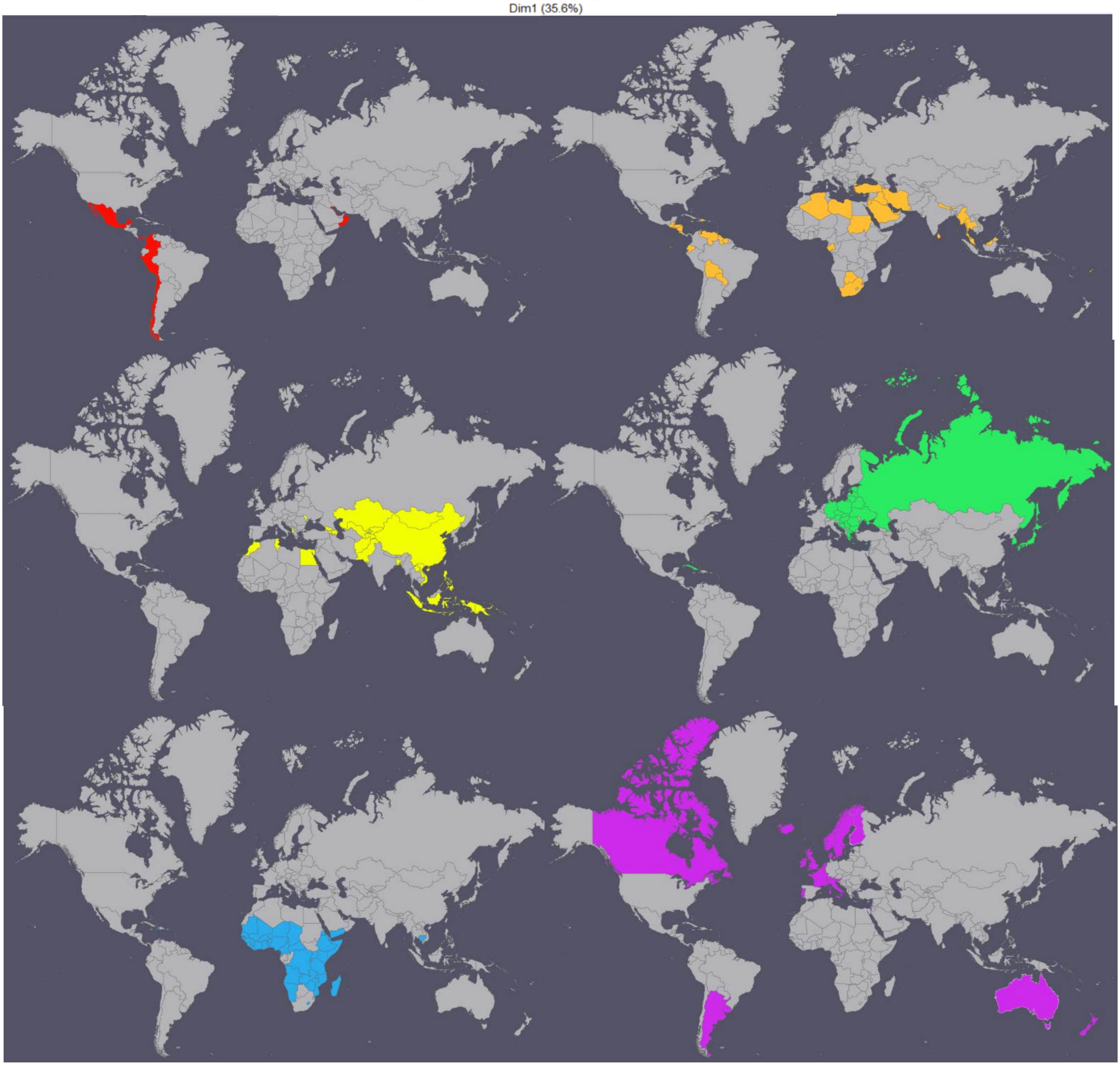
The visuals on the following page offer an intriguing insight into the clustering of the k = 6 K-means algorithm. Each colour in the cluster plot relates to the same colour on the world map, allowing for a much easier and visually stimulating view of the clustered data.

Clusters 1 (red), 2 (orange) and 6 (purple) have no real structure relating to regional patters. However, clusters 3 (yellow), 4 (green) and 5 (blue) have more distinct regional patterns. Cluster 3 is almost entirely situated in Asia, with cluster 4 a combination of Asia and Europe and cluster 5 predominantly in Africa.

Although these results seem far more convincing than the average silhouette width metric of 0.24, one should be hesitant of drawing COVID-19 related conclusions based on the graphic. Since the dataset contains an array of features, not all relating to COVID-19, some are health-related and some are socio-economic indicators, it is difficult to make claims about COVID-19 alone. All these variables should be taken into context before conclusions are made.

I think it would be unwise to conclude that regional COVID-19 patterns exist. Instead, the conclusion should be made that an overlap (albeit small) exists between the spread of COVID-19 and socio-economic variables contained within the dataset, leading to the perceived regional patterns in the world map graphic.

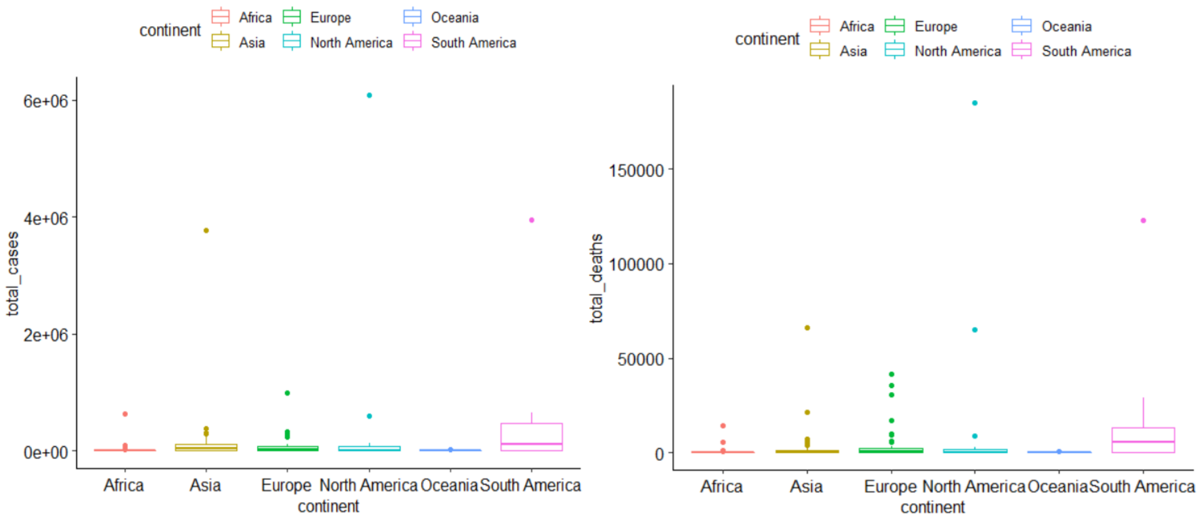

kmeans Cluster Plot: k = 6
Dim2 (14.5%)
Dim1 (35.6%)
2.5
0.0
-2.5
-5.0
-7.5
-4
-2
0
2
4
cluster
1
2
3
4
5
6



Appendix

iso_code
continent
location
date
total_cases
new_cases
new_cases_smoothed
total_deaths
new_deaths
new_deaths_smoothed
total_cases_per_million
new_cases_per_million
new_cases_smoothed_per_million
total_deaths_per_million
new_deaths_per_million
new_deaths_smoothed_per_million
population
population_density
median_age
aged_65_older
aged_70_older
gdp_per_capita
extreme_poverty
cardiovasc_death_rate
diabetes_prevalence
female_smokers
male_smokers
handwashing_facilities
hospital_beds_per_thousand
life_expectancy

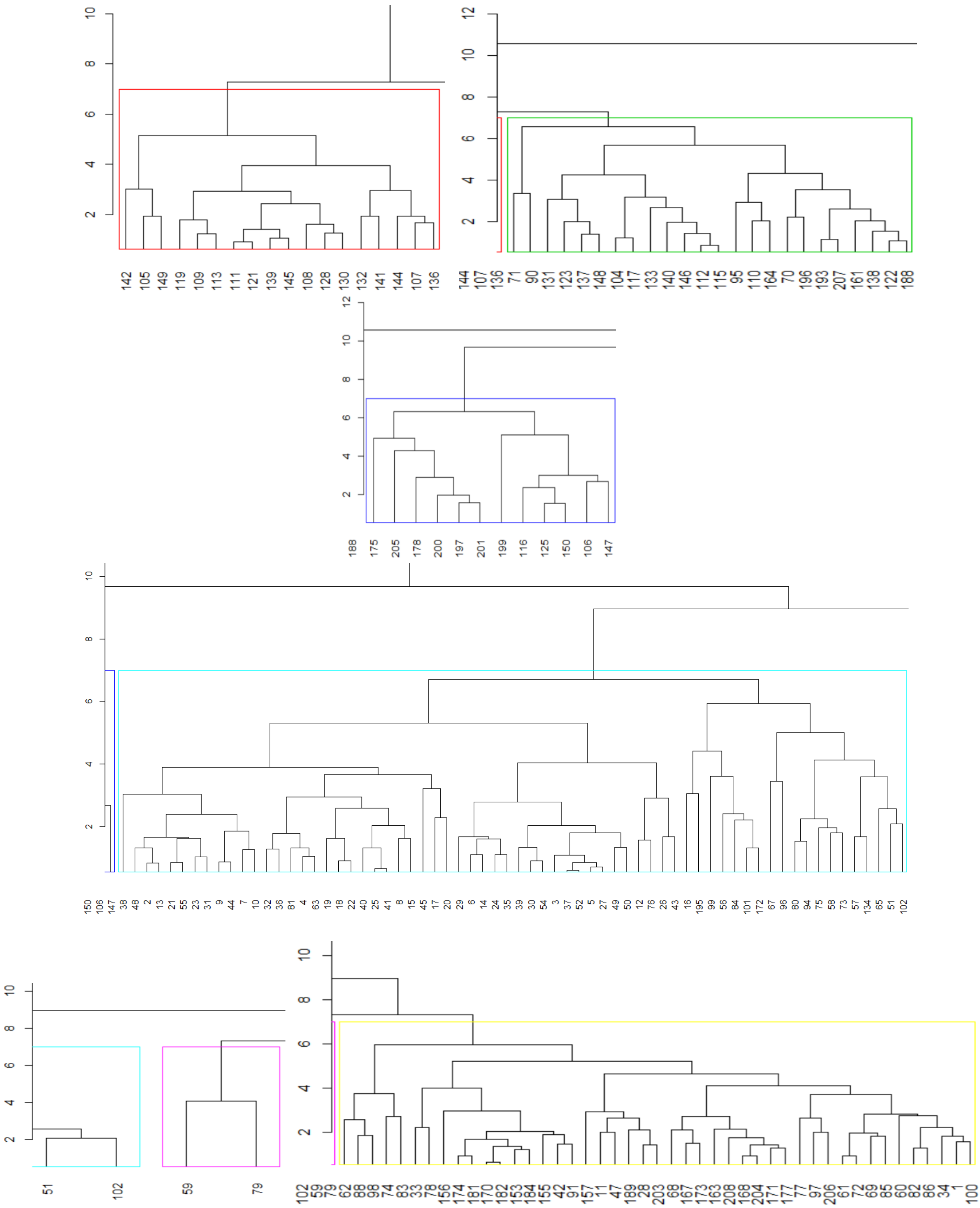
A-1: All 29 variables in the original dataset



A-2: Boxplots for total cases and deaths by continent. It is evident that the data required scaling before being implemented in the models.

	Min	Med	Mean	SD	Max
total_cases	22.0	10524.0	145900.1	625119.0	6075652.0
total_deaths	0.0	203.0	4779.9	18826.7	184689.0
total_cases_per_million	3.0	1469.5	3874.8	5806.2	41302.2
total_deaths_per_million	0.0	30.6	97.0	161.5	881.6
population	97928.0	9904608.0	44183015.0	156081787.0	1439323774.0
population_density	2.0	81.3	204.0	643.1	7915.7
median_age	15.1	29.6	30.3	9.1	48.2
aged_65_older	1.1	6.2	8.6	6.2	27.0
aged_70_older	0.5	3.8	5.5	4.2	18.5
gdp_per_capita	661.2	11840.8	18342.2	19526.6	116935.6
extreme_poverty	0.0	4.2	13.2	17.8	77.6
cardiovasc_death_rate	79.4	244.0	259.9	117.2	724.4
diabetes_prevalence	1.0	7.1	7.6	3.9	22.0
female_smokers	0.1	5.2	9.1	9.6	44.0
male_smokers	7.7	27.8	31.1	12.9	78.1
hospital_beds_per_thousand	0.1	2.4	2.9	2.3	13.1
life_expectancy	53.3	74.1	72.5	7.5	84.6

A-3: Comprehensive five-number summary table for the 17 variables used in the final dataset.



A-4: Fully labelled complete linkage dendrogram from section 3.4 – unnamed since only the best model was used (k-means).

	location		location		location		
1	Algeria	34	Morocco	69	Iraq	105	Belarus
2	Angola	35	Mozambique	70	Israel	106	Belgium
3	Benin	36	Namibia	71	Japan	107	Bosnia and Herzegovina
4	Botswana	37	Niger	72	Jordan	108	Bulgaria
5	Burkina Faso	38	Nigeria	73	Kazakhstan	109	Croatia
6	Burundi	39	Rwanda	74	Kuwait	110	Cyprus
7	Cameroon	40	Sao Tome and Principe	75	Kyrgyzstan	111	Czech Republic
8	Cape Verde	41	Senegal	76	Laos	112	Denmark
9	Central African Republic	42	Seychelles	77	Lebanon	113	Estonia
10	Chad	43	Sierra Leone	78	Malaysia	115	Finland
11	Comoros	44	Somalia	79	Maldives	116	France
12	Congo	45	South Africa	80	Mongolia	117	Germany
13	Cote d'Ivoire	47	Sudan	81	Myanmar	119	Greece
14	Democratic Republic of Congo	48	Swaziland	82	Nepal	121	Hungary
15	Djibouti	49	Tanzania	83	Oman	122	Iceland
16	Egypt	50	Togo	84	Pakistan	123	Ireland
17	Equatorial Guinea	51	Tunisia	85	Palestine	125	Italy
18	Eritrea	52	Uganda	86	Philippines	128	Latvia
19	Ethiopia	54	Zambia	87	Qatar	130	Lithuania
20	Gabon	55	Zimbabwe	88	Saudi Arabia	131	Luxembourg
21	Gambia	56	Afghanistan	89	Singapore	132	Macedonia
22	Ghana	57	Armenia	90	South Korea	133	Malta
23	Guinea	58	Azerbaijan	91	Sri Lanka	134	Moldova
24	Guinea-Bissau	59	Bahrain	94	Tajikistan	136	Montenegro
25	Kenya	60	Bangladesh	95	Thailand	137	Netherlands
26	Lesotho	61	Bhutan	96	Timor	138	Norway
27	Liberia	62	Brunei	97	Turkey	139	Poland
28	Libya	63	Cambodia	98	United Arab Emirates	140	Portugal
29	Madagascar	64	China	99	Uzbekistan	141	Romania
30	Malawi	65	Georgia	100	Vietnam	142	Russia
31	Mali	66	India	101	Yemen	144	Serbia
32	Mauritania	67	Indonesia	102	Albania	145	Slovakia
33	Mauritius	68	Iran	104	Austria	146	Slovenia

A-5: Full list of locations and attached numbers.

	location
147	Sweden
148	Switzerland
149	Ukraine
150	United Kingdom
153	Antigua and Barbuda
155	Bahamas
156	Barbados
157	Belize
161	Canada
163	Costa Rica
164	Cuba
167	Dominican Republic
168	El Salvador
170	Grenada
171	Guatemala
172	Haiti
173	Honduras
174	Jamaica
175	Mexico
177	Nicaragua
178	Panama
181	Saint Lucia
182	Saint Vincent and the Grenadines
184	Trinidad and Tobago
186	United States
188	Australia
189	Fiji
193	New Zealand
195	Papua New Guinea
196	Argentina
197	Bolivia
198	Brazil
199	Chile
200	Colombia
201	Ecuador
203	Guyana
204	Paraguay
205	Peru
206	Suriname
207	Uruguay
208	Venezuela

location	continent	cluste	location	continen	cluste	location	continen	cluste	location	continen	cluste	location	continen	cluste	location	continen	cluster
Bahrain	Asia	1	Algeria	Africa	2	Egypt	Africa	3	Japan	Asia	4	Angola	Africa	5	Israel	Asia	6
Kuwait	Asia	1	Botswan	Africa	2	Morocco	Africa	3	South Kor	Asia	4	Benin	Africa	5	Belgium	Europe	6
Oman	Asia	1	Comoros	Africa	2	Tunisia	Africa	3	Austria	Europe	4	Burkina	Africa	5	Denmark	Europe	6
Qatar	Asia	1	Gabon	Africa	2	Afghanis	Asia	3	Belarus	Europe	4	Burundi	Africa	5	Finland	Europe	6
Mexico	North Ameri	1	Libya	Africa	2	Armenia	Asia	3	Bosnia & I	Europe	4	Cameroo	Africa	5	France	Europe	6
Panama	North Ameri	1	Mauritiu	Africa	2	Azerbaij	Asia	3	Bulgaria	Europe	4	Cape Ve	Africa	5	Iceland	Europe	6
Chile	South Amer	1	Seychell	Africa	2	Banglad	Asia	3	Croatia	Europe	4	Central A	Africa	5	Ireland	Europe	6
Colombi	South Amer	1	South Af	Africa	2	China	Asia	3	Cyprus	Europe	4	Chad	Africa	5	Italy	Europe	6
Peru	South Amer	1	Sudan	Africa	2	Georgia	Asia	3	Czech Rep	Europe	4	Congo	Africa	5	Luxemb	Europe	6
			Bhutan	Asia	2	Indonesi	Asia	3	Estonia	Europe	4	Cote d'Iv	Africa	5	Malta	Europe	6
			Brunei	Asia	2	Kazakhs	Asia	3	Germany	Europe	4	Democra	Africa	5	Netherla	Europe	6
			Iran	Asia	2	Kyrgyzst	Asia	3	Greece	Europe	4	Djibouti	Africa	5	Norway	Europe	6
			Iraq	Asia	2	Laos	Asia	3	Hungary	Europe	4	Equatori	Africa	5	Portugal	Europe	6
			Jordan	Asia	2	Mongoli	Asia	3	Latvia	Europe	4	Eritrea	Africa	5	Slovenia	Europe	6
			Lebanon	Asia	2	Pakistan	Asia	3	Lithuania	Europe	4	Ethiopia	Africa	5	Sweden	Europe	6
			Malaysia	Asia	2	Philippin	Asia	3	Macedoni	Europe	4	Gambia	Africa	5	Switzerla	Europe	6
			Maldives	Asia	2	Tajikista	Asia	3	Monteneg	Europe	4	Ghana	Africa	5	United K	Europe	6
			Myanma	Asia	2	Timor	Asia	3	Poland	Europe	4	Guinea	Africa	5	Canada	North An	6
			Nepal	Asia	2	Uzbekist	Asia	3	Romania	Europe	4	Guinea-B	Africa	5	Australia	Oceania	6
			Palestin	Asia	2	Vietnam	Asia	3	Russia	Europe	4	Kenya	Africa	5	New Zea	Oceania	6
			Saudi Ar	Asia	2	Albania	Europe	3	Serbia	Europe	4	Lesotho	Africa	5	Argentin	South Ar	6
			Sri Lank	Asia	2	Moldova	Europe	3	Slovakia	Europe	4	Liberia	Africa	5	Uruguay	South Ar	6
			Thailand	Asia	2	Papua N	Oceania	3	Ukraine	Europe	4	Madagas	Africa	5			
			Turkey	Asia	2				Cuba	North Am	4	Malawi	Africa	5			
			United A	Asia	2							Mali	Africa	5			
			Antigua	North An	2							Mauritan	Africa	5			
			Bahama	North An	2							Mozamb	Africa	5			
			Barbado	North An	2							Namibia	Africa	5			
			Belize	North An	2							Niger	Africa	5			
			Costa Ri	North An	2							Nigeria	Africa	5			
			Dominic	North An	2							Rwanda	Africa	5			
			El Salva	North An	2							Sao Tom	Africa	5			
			Grenada	North An	2							Senegal	Africa	5			
			Guatema	North An	2							Sierra Le	Africa	5			
			Hondura	North An	2							Somalia	Africa	5			
			Jamaica	North An	2							Swazilan	Africa	5			
			Nicaragu	North An	2							Tanzani	Africa	5			
			Saint Lu	North An	2							Togo	Africa	5			
			Saint Vir	North An	2							Uganda	Africa	5			
			Trinidad	North An	2							Zambia	Africa	5			
			Fiji	Oceania	2							Zimbabw	Africa	5			
			Bolivia	South Ar	2							Cambod	Asia	5			
			Ecuador	South Ar	2							Yemen	Asia	5			
			Guyana	South Ar	2							Haiti	North An	5			
			Paragua	South Ar	2												
			Surinam	South Ar	2												
			Venezue	South Ar	2												

A-6: Full list of countries in the final (best) k-means model with k = 6 as depicted on the maps in section 4.

References

Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2017.

https://statisticsglobe.com/regression-imputation-stochastic-vs-deterministic/

https://www.datanovia.com/en/courses/partitional-clustering-in-r-the-essentials/

https://techvidvan.com/tutorials/cluster-analysis-in-r/

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/clara