Using Data Clustering to Understand Life Expectancies in Various Countries

Hank Magan

10/19/2021

Abstract

In this report, similarities and differences between life expectancy at different ages among various nations around the world are analyzed to determine similarities between said countries. The data used was gathered from 26 countries during the 1960's, and includes data from both males and females. This report focuses primarily on the technique of clustering(and various techniques used to perform it), and then applies the outcome to applications outside of the lab. As briefly mentioned, a variety of techniques are used to gain different perspectives on the results of data clustering, with discussion on how each of them manipulate the outcome of the analysis.

Introduction

The data analyzed in this report details life expectancy from 26 different countries from the years 1960-1967. This data is further broken down by projected life expectancy at different ages, as well as by gender. M0, M25, M50 and M75 are the remaining life expectancies for a male of ages 0, 25, 50 and 75, and F0, F25, F50 and F75 are the remaining life expectancies for a female. The dataset is quite small, however meaningful analysis can still be made.

While the data is technically correct and contains no missing values, not all of it is entirely relevant to this report. It is not necessary to know the year for each observation, as it does not offer much in terms of analysis—year is irrelevant (in this context) when determining similar data points. Moreover, it is obvious that data measured in the same country but separated by a couple years will be statistically similar, so it is not necessary to compute this. As such, rows from the same country will be merged. Additionally, some of the data is separated by race within countries. In these instances, the composite data will be removed, and the distinct groups within that country will be analyzed separately.

```r
# average together rows with same name; round to match the dataset of integers
cleanExp <- lifeExp %>% group_by(Country) %>% summarize_if(is.numeric, mean) %>%
            mutate_if(is.numeric, round)

# remove the year column
cleanExp$Year <- NULL

# remove composite US column to focus on white vs. nonwhite data
cleanExp <- subset(cleanExp, Country != "US")

# make the country column names of the rows
cleanExp <- tibble::column_to_rownames(cleanExp, "Country")
```

This analysis focuses on the technique of clustering, which can be simply defined as dividing data into distinct groups as to gain an understanding of similarities between data points.

To accomplish this, a variety of algorithms may be used; these algorithms rely on being able to garner a numerical representation of the distance between two data points. The different ways to do this are called distance measures; examples of distance measures includes Euclidean distance, Manhattan distance, correlation-based distances (such as Pearson correlation distance), among others. Euclidean distance is the method of calculating distance as it is traditionally understood in mathematics. It simply describes the shortest distance between two points on a graph, which includes paths that slant diagonally. On the other hand, Manhattan distance is the sum of the difference in `x` and the difference in `y` between two points. In other words, if Euclidean distance represents the hypotenuse of a triangle (whose vertices represent two points on a graph), Manhattan distance is the sum of its two legs. Euclidean distance is more likely to be influenced by outliers than Manhattan distance. Correlation-based distances disregard geometric closeness, and instead focus on the statistical correlation between the features. Pearson correlation distance is probably the most popular correlation distance method, which measures the degree of linear relationship between variables. Each of these are viable methods, and their use is situational. It is wise to pick the one that improves the performance of the clustering as much as possible.

One method of clustering (using one of the previously discussed distance measures) is partition clustering, in which data is subdivided into a set of `k` groups, where `k` is predefined. Within partition clustering, there are even more techniques, the most popular being k-means clustering. K-means clustering represents clusters using the means of the data points belonging to each cluster. As such, k-means clustering is susceptible to outliers, however is still a reliable method of clustering. With that said, there are other ways to perform clustering.

Hierarchical clustering is an another approach to clustering, in which a tree-like structure is constructed out of identified clusters. Unlike partition clustering, hierarchical clustering does not take a `k` value. The graphic it produces is known as a dendrogram, and is particularly interesting because various levels of similarity can be observed all at once by looking at different levels of the hierarchy. See Figure 1 for a representation of a dendogram. As one can see, it is most definitely a useful way to understand the clustering of a dataset.

Methods

Of course, the analysis in this report largely focuses on clustering. In terms of distance measures, Euclidean and Manhattan distances were used and compared. Moreover, both techniques of partition clustering and hierarchical clustering were used for analysis. So-called "cluster heat maps" were used to visualize distances between observations, as well as to evaluate the differences between distance measures. K-means clustering was used for partition clustering and evaluated graphically.

Results

Firstly, a seed needs to be set in order for some of the algorithms to work properly:

```
# arbitrary value
set.seed(181818888)
```

With that out of the way, a k-means analysis will be conducted on the entire dataset. As mentioned in the introduction, partition clustering methods like k-means analysis require the number of clusters to be specified. As such, variouss values of `k` have been chosen and are modeled below.

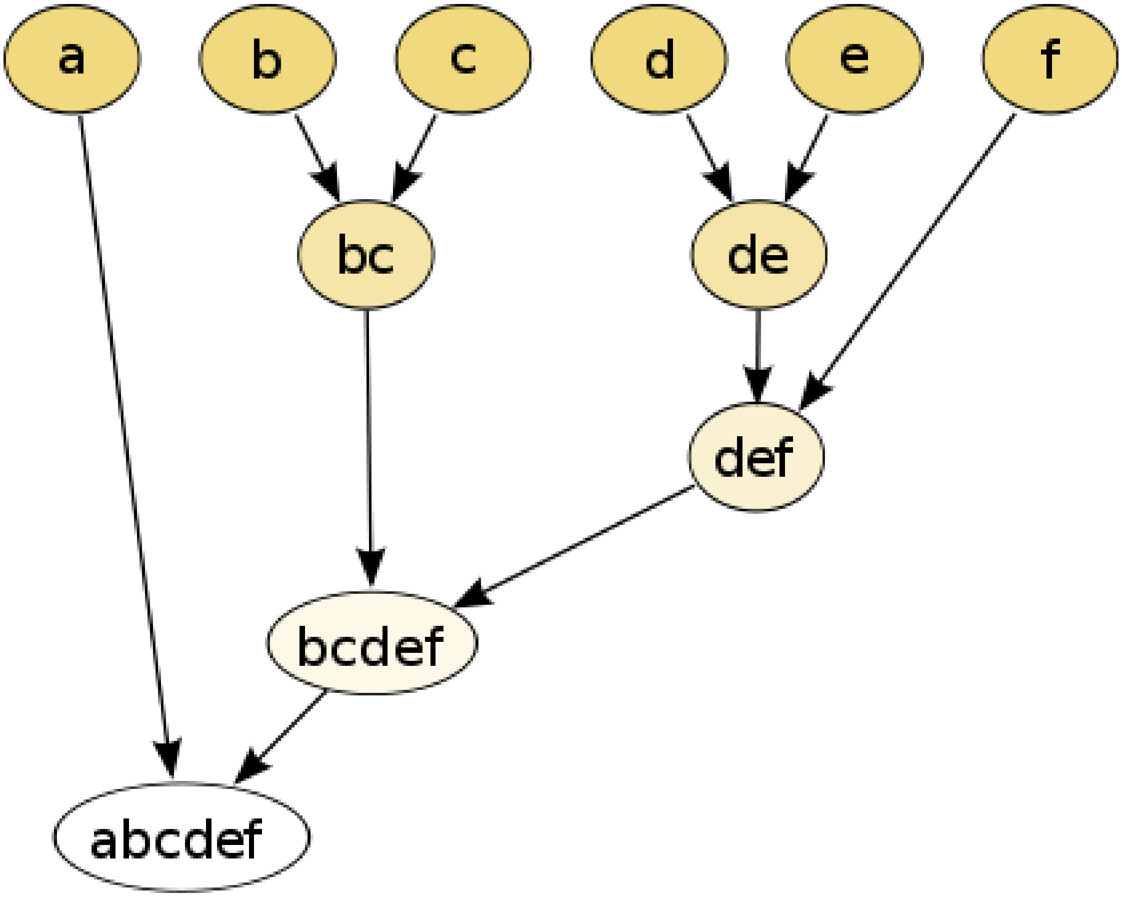
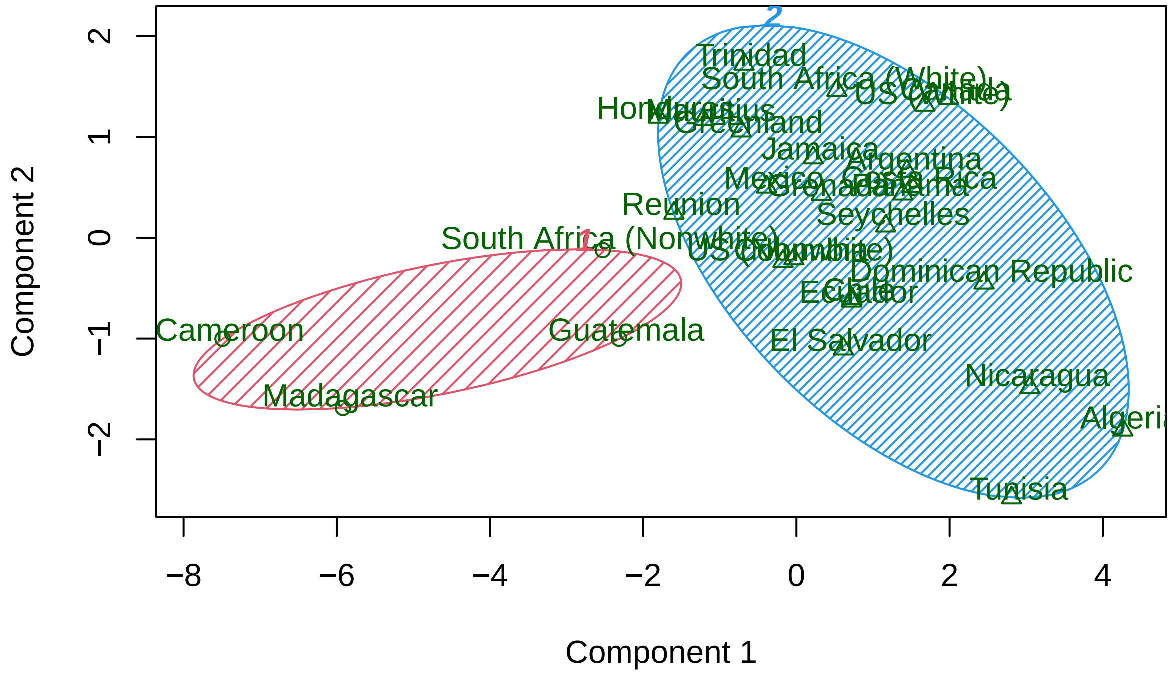
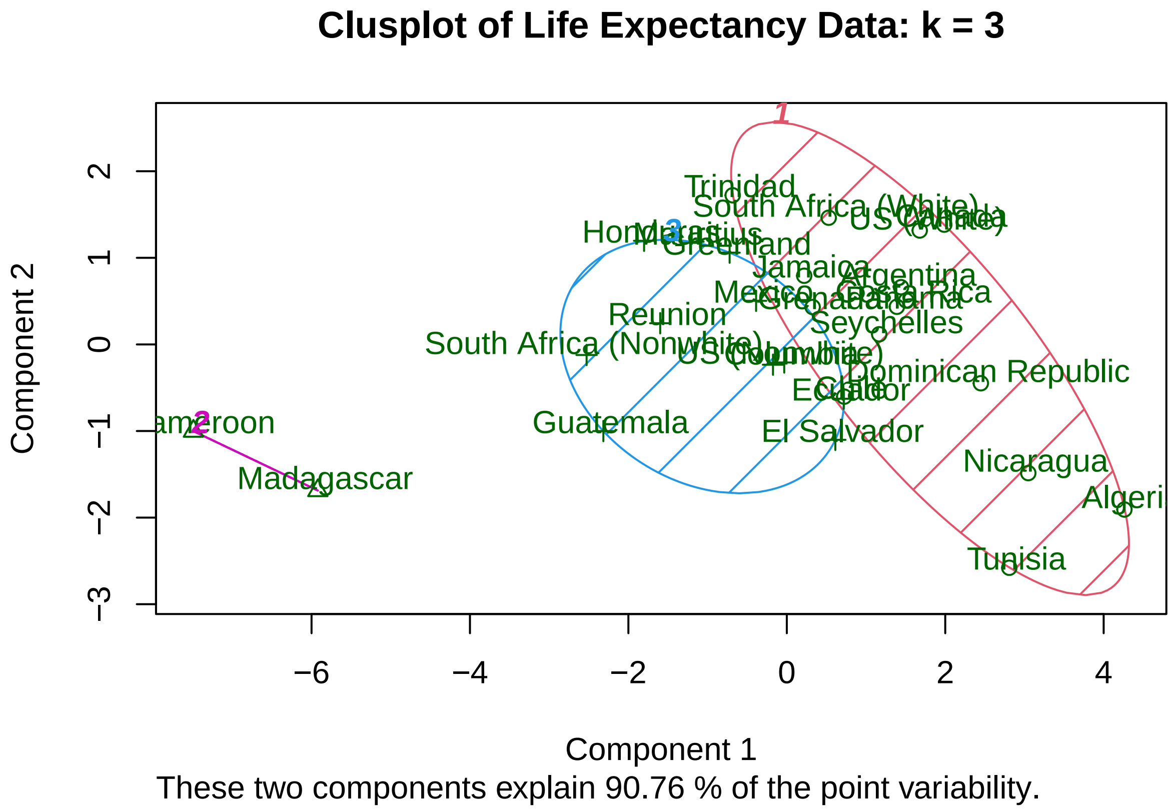


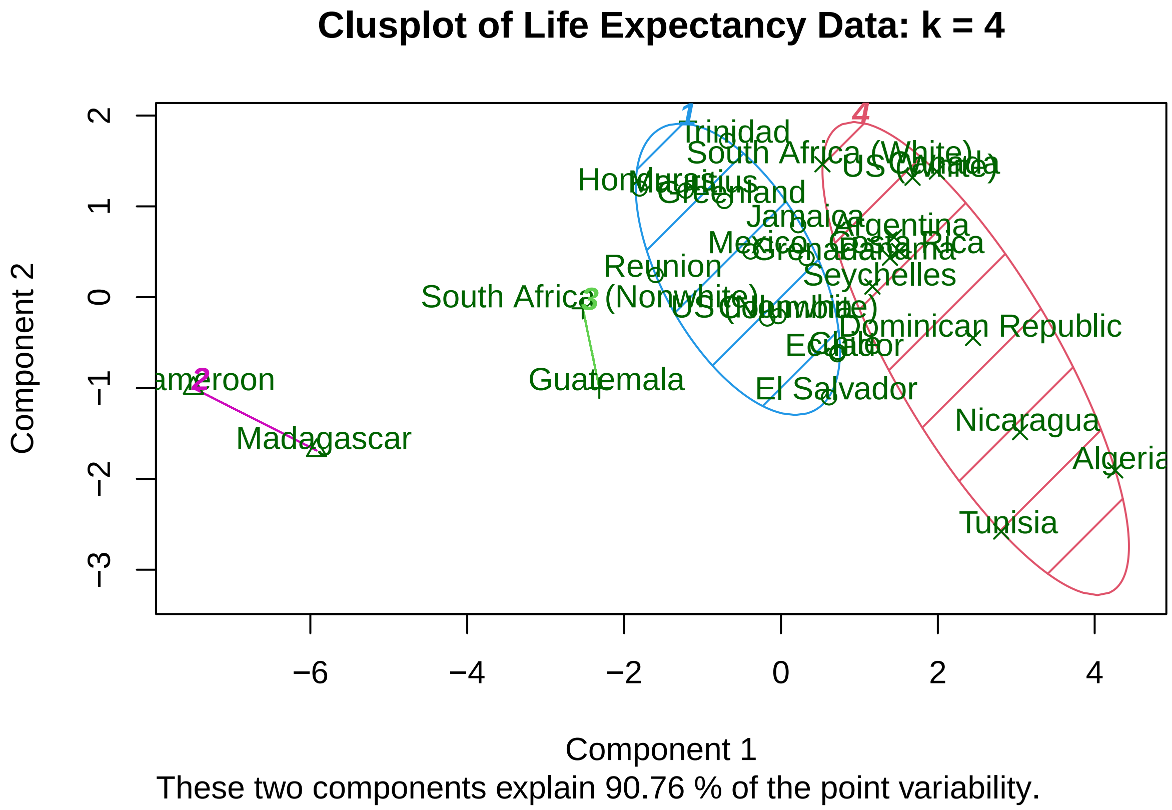
Figure 1: Representation of hierarchical clustering, based on arbitrary data. (2)

Clusplot of Life Expectancy Data: k = 2



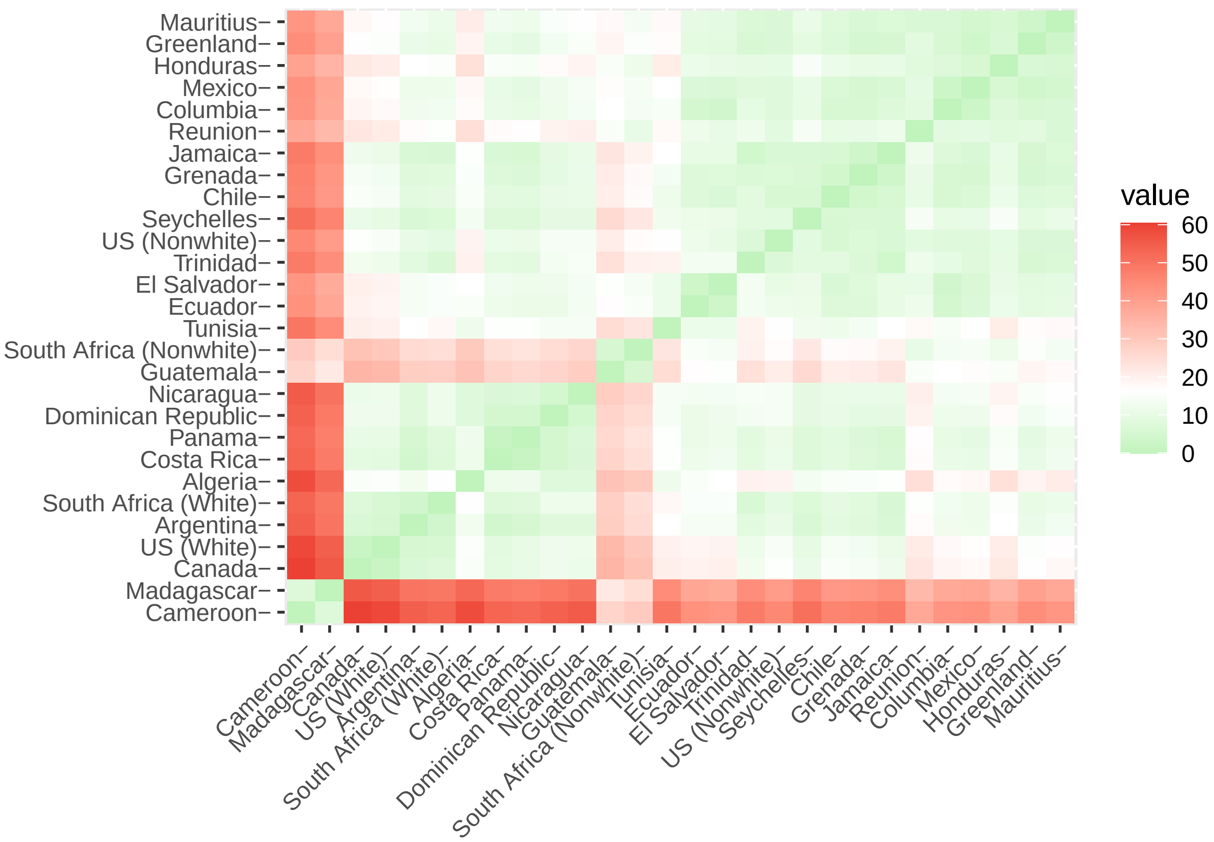
These two components explain 90.76 % of the point variability.

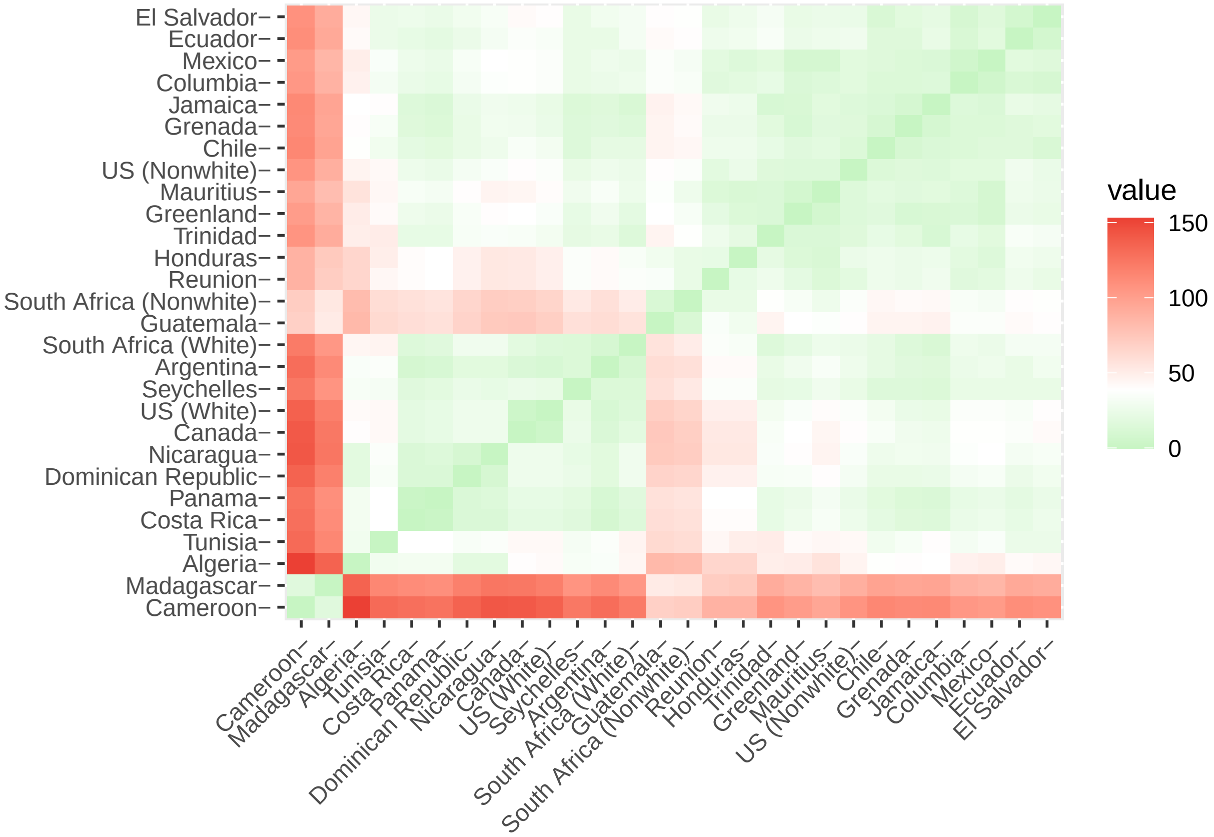
Clusplot of Life Expectancy Data: k = 3
1
2
3
Trinidad
South Africa (White)
US (White)
Canada
Honduras
Mauritius
Greenland
Jamaica
Argentina
Mexico
Grenada
Costa Rica
Panama
Reunion
Seychelles
South Africa (Nonwhite)
US (Nonwhite)
Columbia
Dominican Republic
Ecuador
Chile
ameroon
Madagascar
Guatemala
El Salvador
Nicaragua
Algeri
Tunisia
Component 2
2
1
0
−1
−2
−3
−6
−4
−2
0
2
4
Component 1
These two components explain 90.76 % of the point variability.



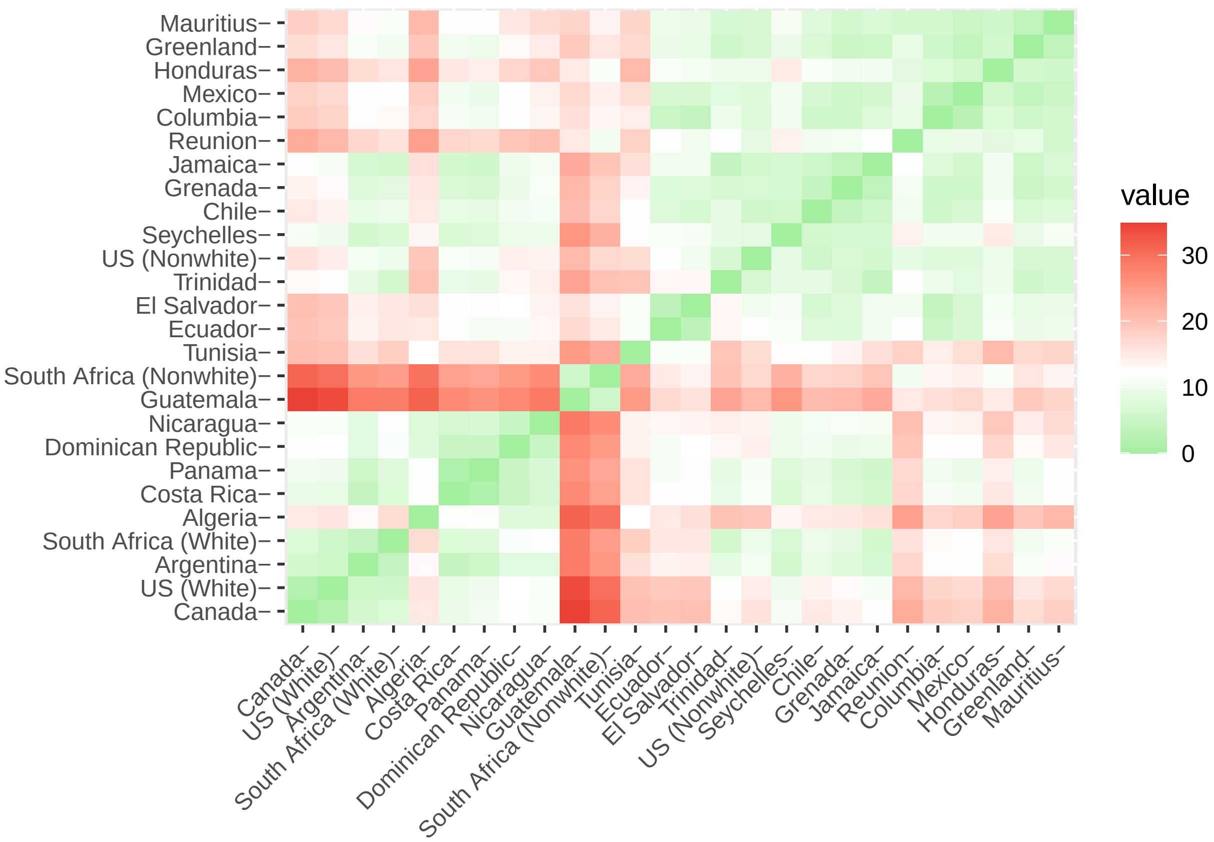
From these graphs, a few things can be observed. It seems that the model is avoiding grouping the outliers, namely Cameroon, Madagascar, South Africa (Nonwhite), and Guatemala. This is unsurprising, as these four countries have the lowest life expectancy out of all the countries in the dataset by a significant amount, causing their distance to be rather large. These countries are also rather poor, and a cluster analysis of the same countries but with GDP (or a similar statistic) would likely provide similar results—the clusters almost seem as if they are based around the countries' wealth. Additionally, it is interesting how the "white" and "nonwhite" life expectancy in the US and South Africa is different enough to be clustered separately, something which speaks to the effects of the socioeconomic divide in each of these countries.

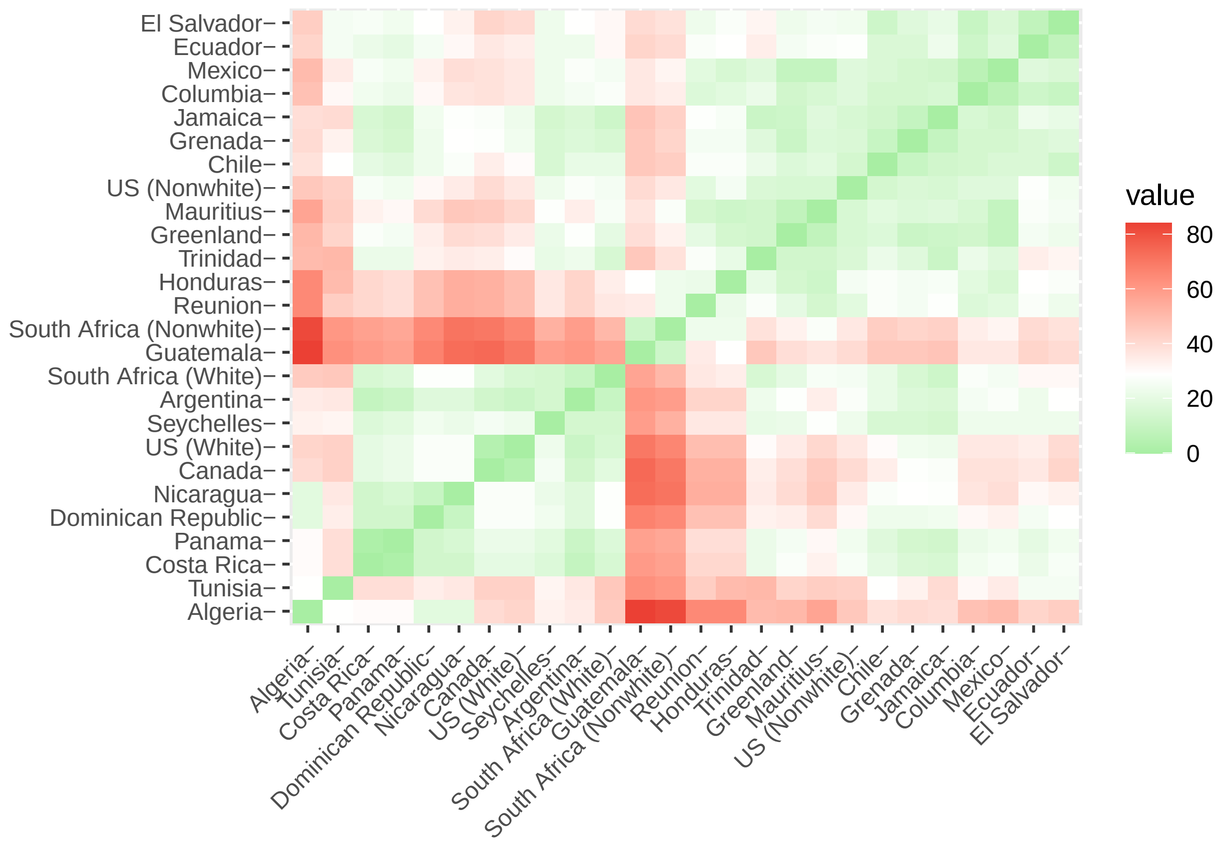
With this, heat maps of the data may be observed. These heat maps visualize a matrix of distances which are generated using the specified methods. In this report, both Euclidean and Manhattan distances are used, as modeled below (Euclidean in the first graphic, Manhattan in the second).


Mauritius
Greenland
Honduras
Mexico
Columbia
Reunion
Jamaica
Grenada
Chile
Seychelles
US (Nonwhite)
Trinidad
El Salvador
Ecuador
Tunisia
South Africa (Nonwhite)
Guatemala
Nicaragua
Dominican Republic
Panama
Costa Rica
Algeria
South Africa (White)
Argentina
US (White)
Canada
Madagascar
Cameroon
value
60
50
40
30
20
10
0



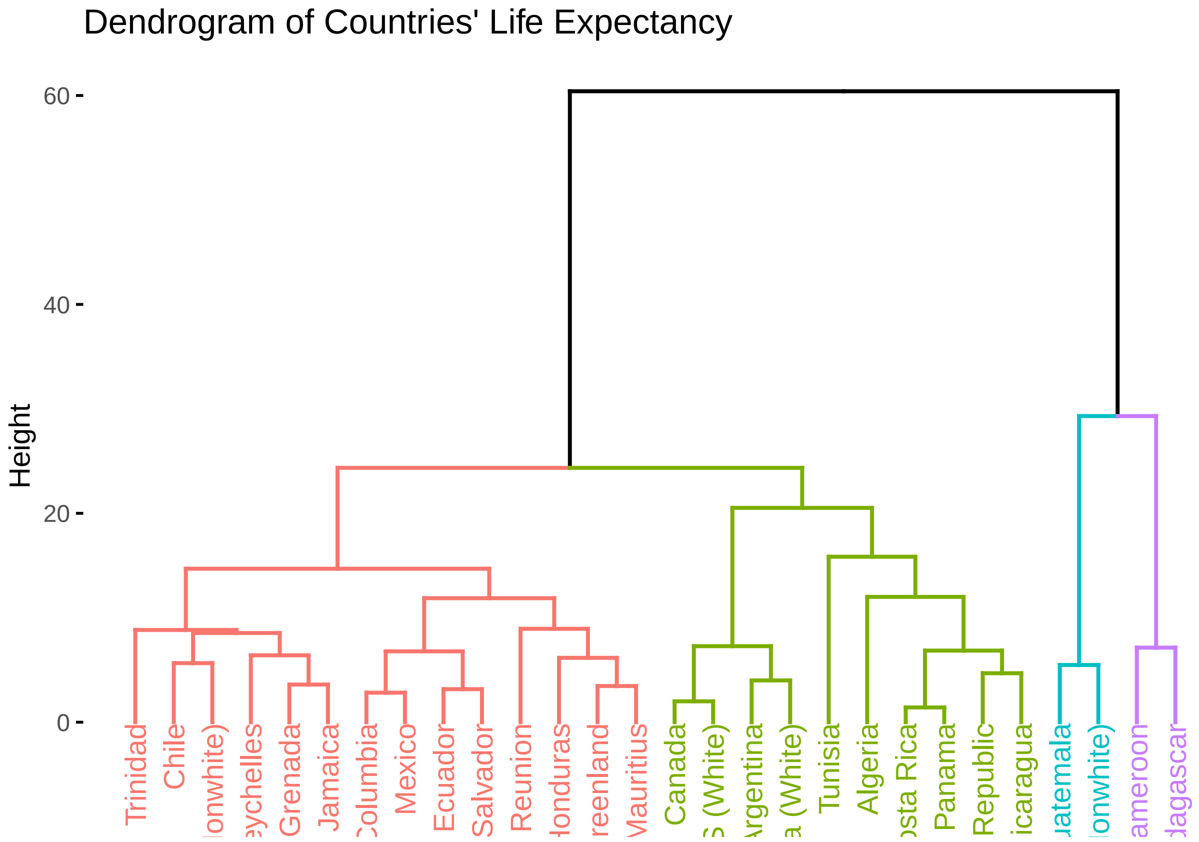
Despite using different distance methods, the two graphics appear more or less the same. With that said, they both follow a distinct pattern in which the outliers (Cameroon and Madagascar) overpower much of the other clusters. Briefly excluding the two from the data yields the following:


Mauritius
Greenland
Honduras
Mexico
Columbia
Reunion
Jamaica
Grenada
Chile
Seychelles
US (Nonwhite)
Trinidad
El Salvador
Ecuador
Tunisia
South Africa (Nonwhite)
Guatemala
Nicaragua
Dominican Republic
Panama
Costa Rica
Algeria
South Africa (White)
Argentina
US (White)
Canada
Canada
US (White)
Argentina
South Africa (White)
Algeria
Costa Rica
Panama
Dominican Republic
Nicaragua
Guatemala
South Africa (Nonwhite)
Tunisia
Ecuador
El Salvador
Trinidad
US (Nonwhite)
Seychelles
Chile
Grenada
Jamaica
Reunion
Columbia
Mexico
Honduras
Greenland
Mauritius
value
30
20
10
0



This provides a graphic which is more telling for the rest of the data, but also goes to show how much of an influence the outliers have on the analysis. Looking closely at both heat maps, there is a noticeable divide between the rich, more first-world countries, and the poorer countries. Additionally, countries that are relatively close geographically and perhaps even culturally seem to have a tendency to cluster together. For example: US (white) and Canada, Costa Rica and Panama, Ecuador and El Salvador, etc.

Another way the data can be analyzes is (as discussed) through hierarchical clustering. To do so, a dendrogram will be constructed, and clusters will be evaluated that way, like so:



This graphic is yet another perspective on what has been demonstrated through partition clustering. The outliers are rather isolated, only connected through the root of the graph. This demonstration furthers the idea that countries which are close geographically and culturally have similar life expectancy (branches connecting two countries are easy to point out). Moreover, the similarities between individual countries, rather than between groups of countries, is more easily observed on this graphic. Again, it seems that, for the most part, the wealthiest countries are clustered together, with those who are poor grouped together and those who are somewhere in between together.

Conclusion

Through clustering, a few things may be concluded. Firstly, although not necessarily the focus of the report, a country's wealth is correlated with its life expectancy. Secondly, Guatemala, South Africa (Nonwhite), Cameroon, Madagascar (particularly Cameroon and Madagascar) have life expectancies which are significantly lower than the average, emphasizes how these countries are relatively poor and unstable. Thirdly, countries which are similar geographically and culturally tend to have similar life expectancy. Lastly, race can have a significant effect on life expectancy, speaking to the effects of the socieconomic divide in the some of these countries.

References

1. https://towardsdatascience.com/log-book-guide-to-distance-measuring-approaches-for-k-means-clustering-f137807e8e21
2. https://en.wikipedia.org/wiki/Hierarchical_clustering

3. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
4. https://en.wikipedia.org/wiki/Euclidean_distance