MA615 Midterm Project

Gahyun (Grace) Kim

10/21/2019

Introduction

This report attempts to examine the possible relationships between variables in the World Bank Data from R. The World Bank Data contains records collected from 1960 to 2018 for various characteristics of a plethora of countries. The variables examined in this report are the following: GDP, life expectancy, infant mortality, urban population growth by percentage, urban population, nitrous oxide emission, carbon dioxide emission, number of nurses and midwives per 1000 people, number of physicians per 1000 people, and mortality for ages under 5 measured per 1000 people. The countries that were selected for analysis were the United States, Mexico, Sweden, Ghana, Republic of Korea, Philippines, Australia, and Russia.

Data Set

To display the data set in a more intuitive manner, the iso2c country codes were removed. The regions for each country were also removed because this paper focuses on individual countries, not regional differences. The country names and variables examined are displayed in the following lines of code. An example of one of the observations is also shown below.

```
unique(data$country)
```

```
## [1] "Australia"          "Ghana"              "Korea, Rep."
## [4] "Mexico"             "Philippines"        "Russian Federation"
## [7] "Sweden"             "United States"
```

```
names(data)
```

```
##  [1] "country"                    "date"
##  [3] "co2_emission"               "GDP"
##  [5] "infant_mortality"           "life_expectancy"
##  [7] "n2o_emission"               "nurses_midwives_per_1000"
##  [9] "physicians_per_1000"        "under5_mortality_per_1000"
## [11] "urban_pop"                  "urban_pop_growth_pct"
```

```
head(data, 1)
```

```
##      country date co2_emission      GDP infant_mortality life_expectancy
## 37 Australia 1996      16.5018 21861.33              5.6        78.07805
##    n2o_emission nurses_midwives_per_1000 physicians_per_1000
## 37      60474.8                  10.3597              2.5199
##    under5_mortality_per_1000 urban_pop urban_pop_growth_pct
## 37                       6.8  15521685             1.159392
```

Correlation Plot

Correlation plots graphically display a correlation matrix with the variables used. The following correlation plot represents the correlation between different pairs of variables by color. Variables that are positively correlated are displayed in blue, and variables that are negatively correlated are displayed in red. The intensity of the color is proportional to the absolute value of the correlation coefficient.

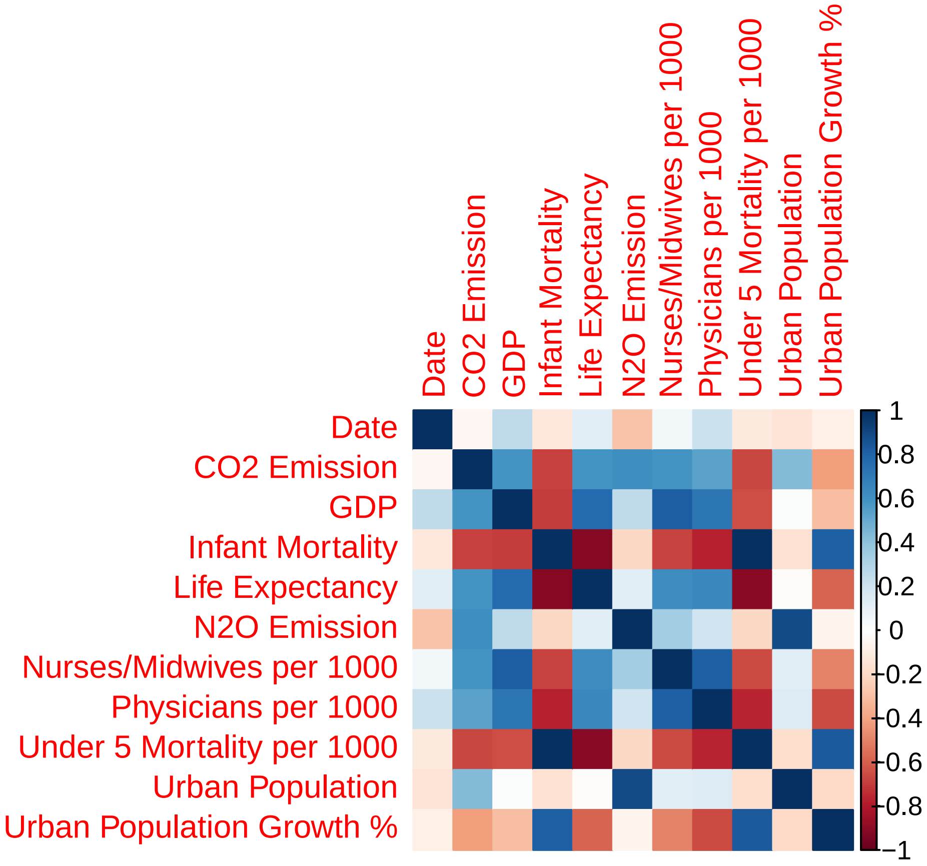


Figure 1: Correlation plot for selected variables in World Bank Data

According to this correlation plot, there are a few pairs of variables that exhibit a strong correlation. Pairs with visibly strong positive correlations are infant mortality vs. under 5 mortality per 1000 and nitrous oxide emission vs. urban population. Variable pairs with visibly strong negative correlations are infant mortality vs. life expectancy, life expectancy vs. under 5 mortality per 1000, and infant mortality vs. physicians per 1000.

Scree Plot

The scree plot in this report displays the number of data clusters against the sum of squares within each group. The "elbow" in the line, the point where the graph seems to level off, represents that factors to the left of that point should be retained as significant. Because this scree plot lacks a clear "elbow" where the graph begins to level off, a cluster analysis is performed to further investigate how to divide the data.

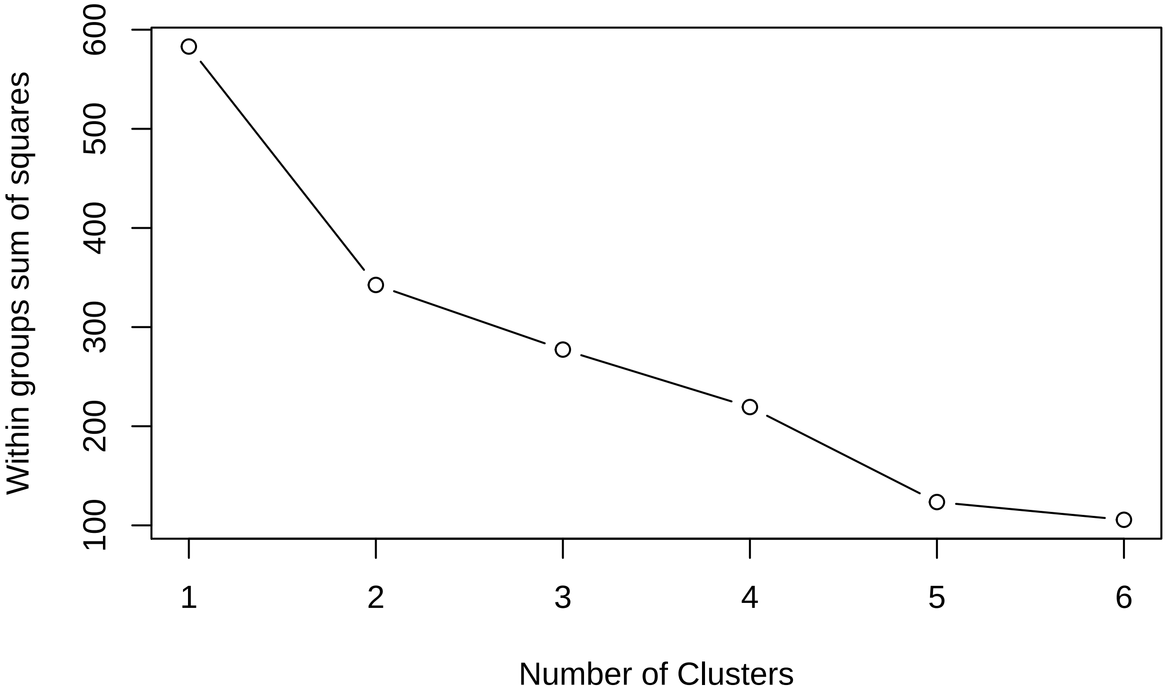


Figure 2: Scree plot for World Bank Data

Cluster Analysis

Cluster analysis is the process of determining which data points can be grouped into how many clusters. Clusters are small collections of data points that are similar to each other within the same cluster. Data in one cluster is not similar to data in another. Through cluster analysis, one can understand how to divide large amounts of data for exploratory analysis and discover latent variables.

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

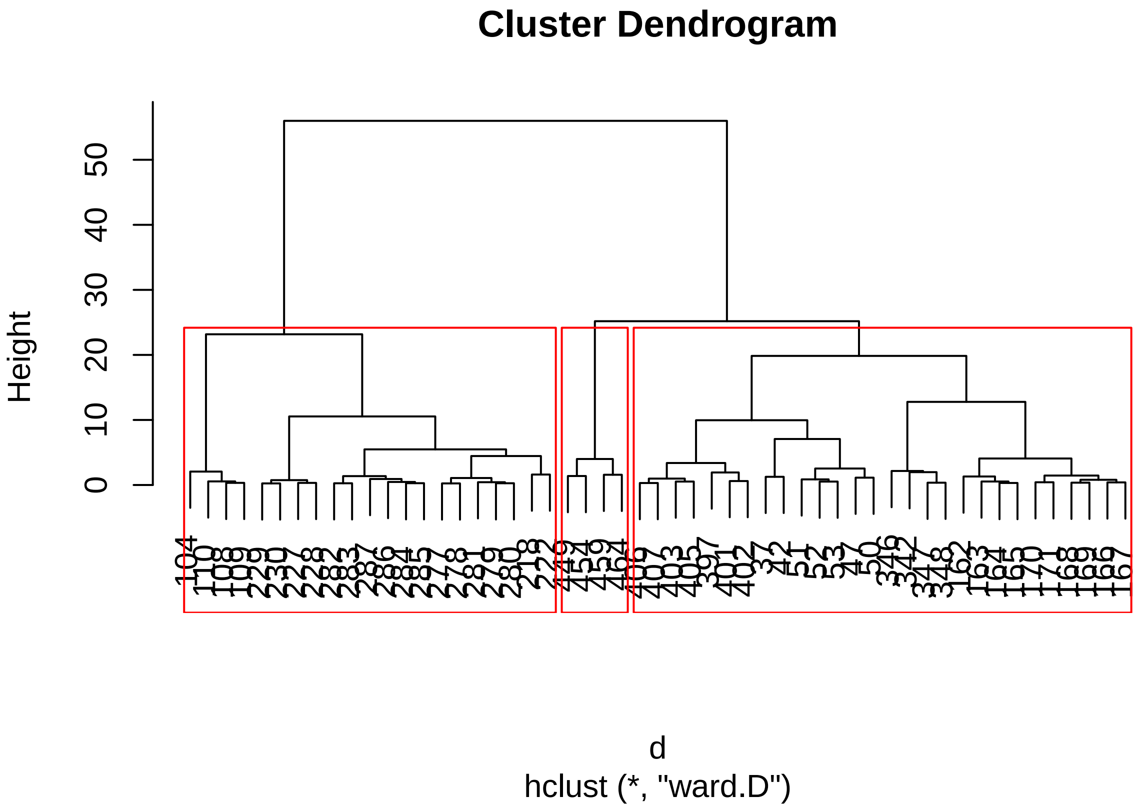


Figure 3: Cluster dendogram for World Bank Data

Cluster dendograms are the main tools used for visually representing a cluster solution for hierarchical clustering methods. Each line on the very bottom of the dendogram represents an observation from the data.

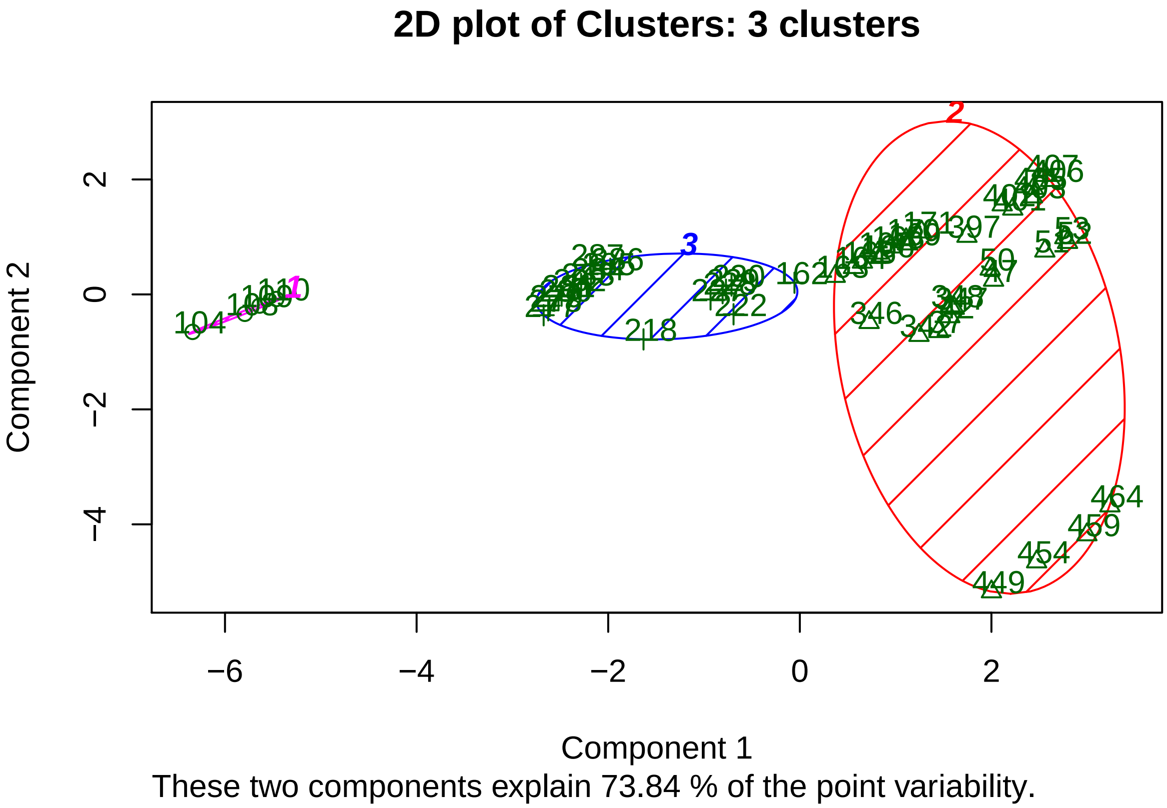


Figure 4: 2D cluster plot for World Bank Data

This 2D plot of clusters displays which observations would be clustered together if we were to group the given data set into 3 clusters.

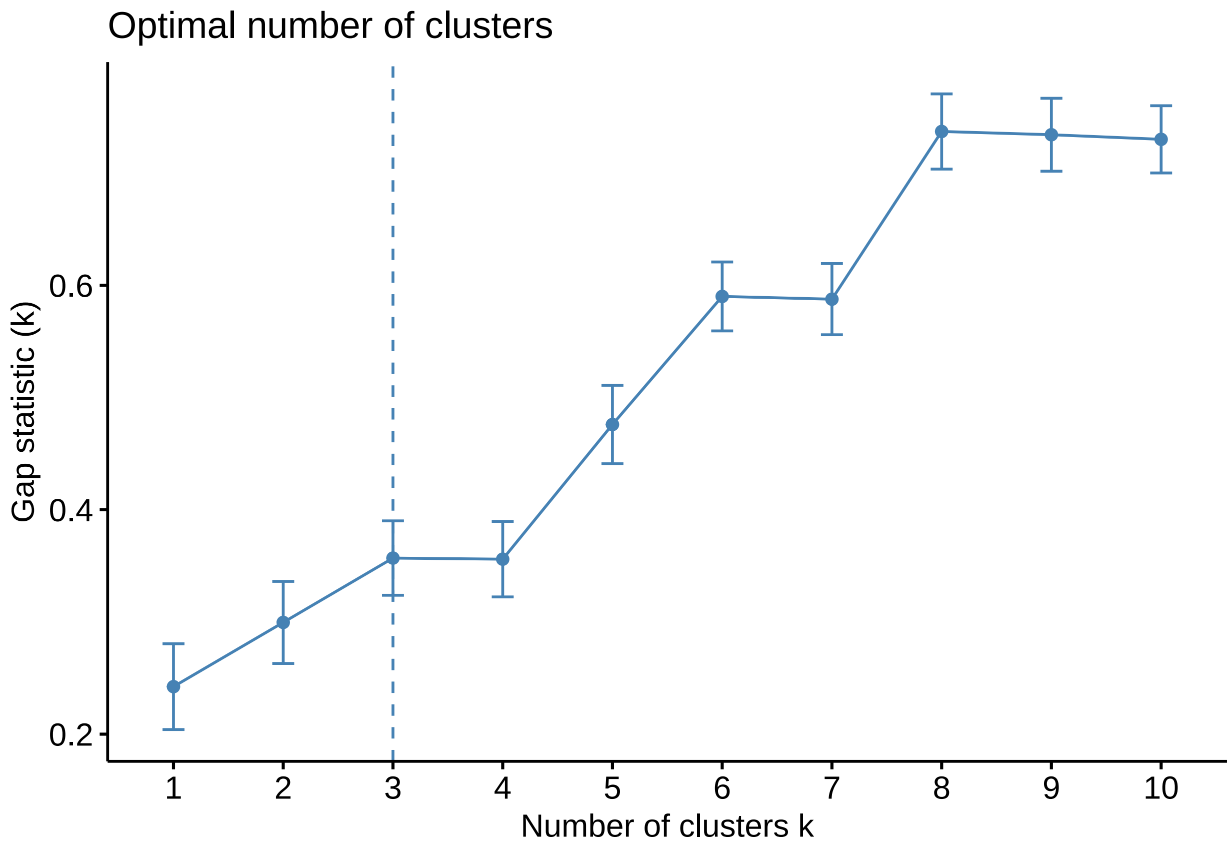


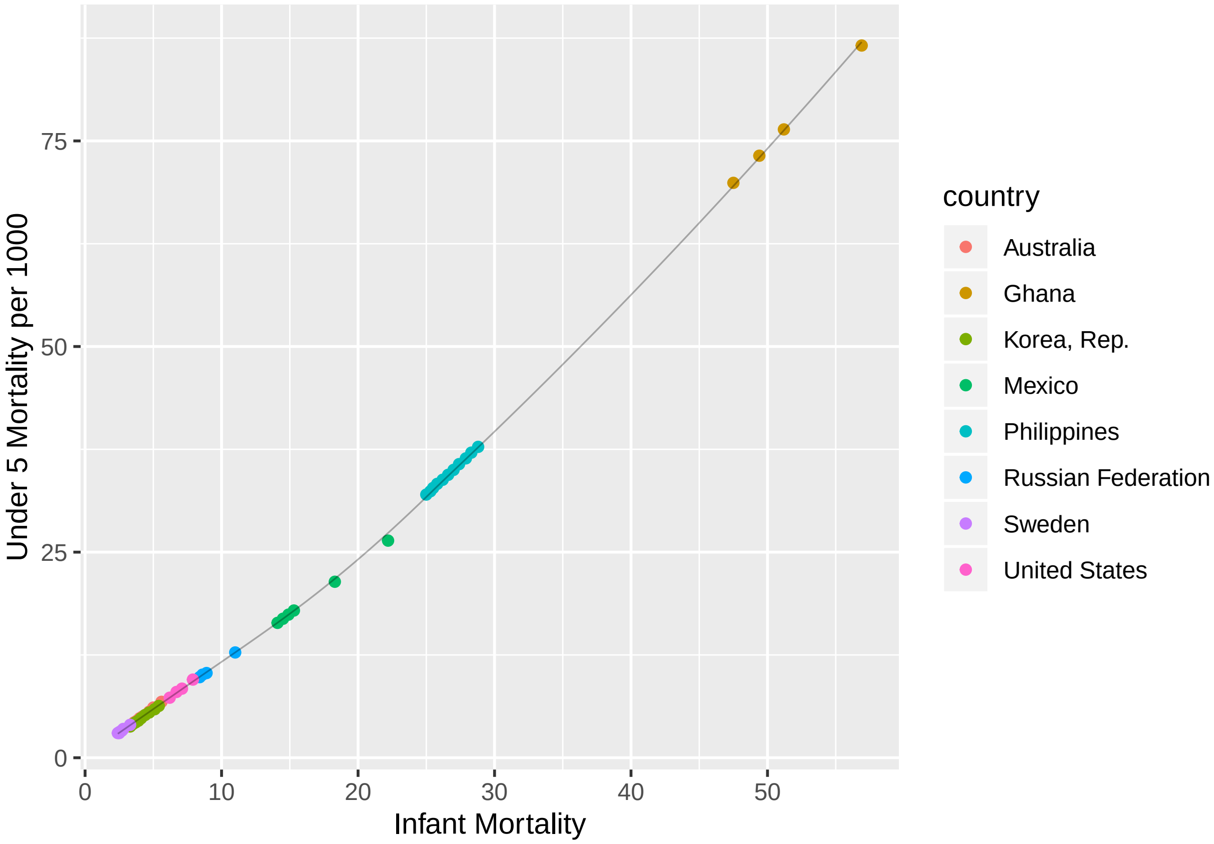
Figure 5: Optimal number of clusters for World Bank Data

The following plot represents how many clusters are optimal for the given data set by plotting the number of clusters against the k-means gap statistic. As the dendogram demonstrated above, the optimal number of clusters for this data set is three.

Exploratory Data Analysis for Variables

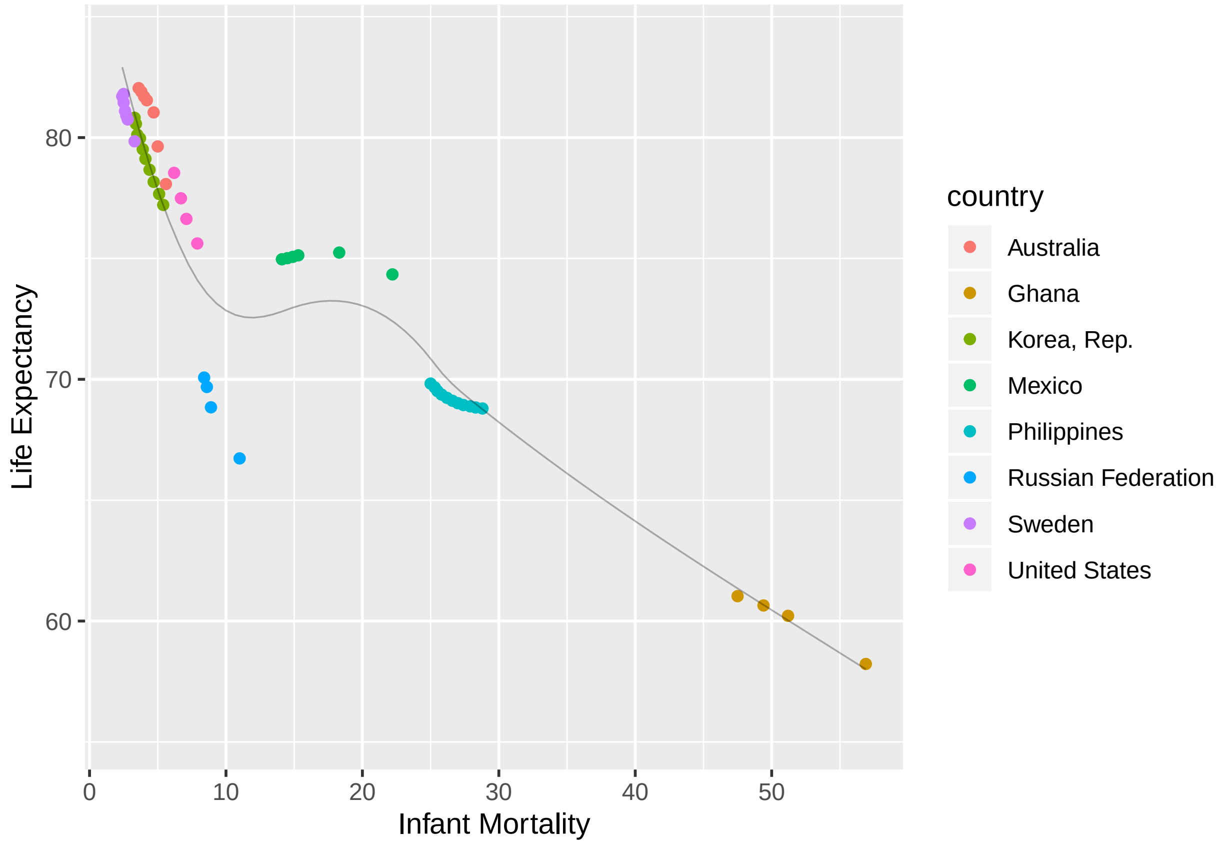
To visualize the variable pairs that seemed to be correlated according to the correlation plot, scatterplots with each pair were made.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



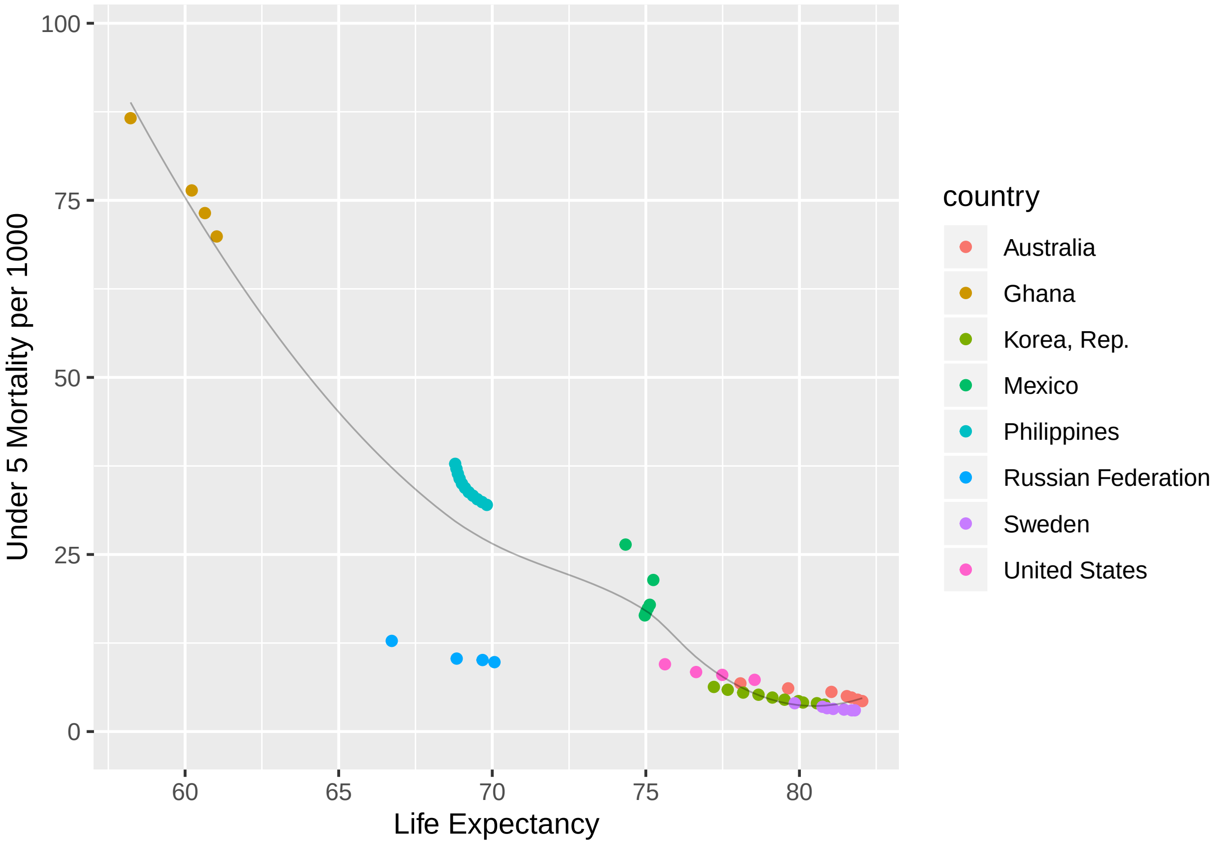
The first scatterplot displays infant mortality plotted against under 5 mortality per 1000. Because the two variables measure nearly the same quality, a strong positive correlation was observed in this scatterplot.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



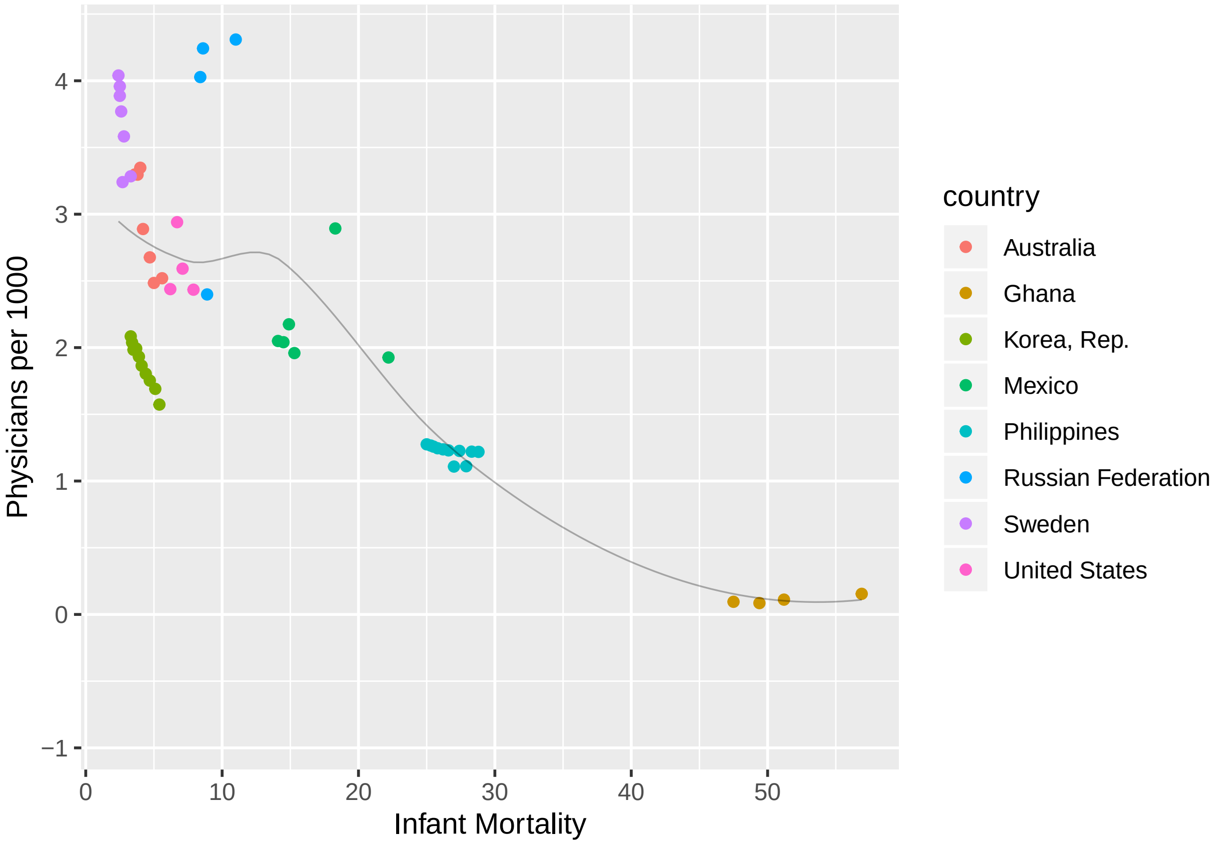
The second scatterplot displays infant mortality plotted against life expectancy. As expected, countries with higher life expectancy have lower infant mortality and vice versa. However, unlike the infant mortality vs. under 5 mortality per 1000 plot above, there are some outliers that make the best fit line deviate from linear behavior.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The third scatterplot displays life expectancy plotted against under 5 mortality per 1000. There is an overall negative correlation according to the plot. This plot seems to exhibit a trend that is generally similar to the second plot because under 5 mortality per 1000 and infant mortality are similar qualities, as shown in the first plot.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The final scatterplot shows infant mortality plotted against the number of physicians per 1000 people. Based on this plot, we can deduce that having more physicians may lower infant mortality across nations.

Conclusion

This project examined 12 variables from the World Bank Data. The countries chosen were the United States, Mexico, Sweden, Ghana, Republic of Korea, Philippines, Australia, and Russia. The variables chosen were GDP, life expectancy, infant mortality, urban population growth by percentage, urban population, nitrous oxide emission, carbon dioxide emission, number of nurses and midwives per 1000 people, number of physicians per 1000 people, and mortality for ages under 5 measured per 1000 people.

When the variables were plotted against each other in a correlation plot, it was observed that there were five pairs of variables that had a visibly strong positive or negative correlation. A cluster analysis was performed to see how the data could be divided into groups for analysis. When hierarchical clustering was performed via euclidean distance, it was found that grouping the data into three clusters would be appropriate. Using three as the optimal number of clusters was confirmed when the number of clusters was plotted against the k-means gap statistic.

When four variable pairs from the correlation plot were visualized as scatterplots, the correlations between some variables looked clearer. Upon comparing infant mortality and under 5 mortality per 1000, it is evident in the plots that the two measurements determine similar qualities. A plot with the two variables displays a strong positive correlation, and plots with life expectancy plotted against the two variables have similar negative correlations. Another insightful scatterplot was the plot with infant mortality and the number of physicians per 1000. This scatterplot showed that countries with more physicians have a higher chance of having lower infant mortality rates.