**Life Expectancy Data Analysis**

**Abstract:**

Developed countries have significantly higher life expectancies than underdeveloped countries. There are countless factors that contribute this difference in longevity. For the purpose of simplicity this project will only consider physicians per capita, literacy rate, and GDP per capita. The country’s being compared in this analysis are broken into two groups of three countries, underdeveloped (Madagascar, Afghanistan, and Somalia) and developed (United States, Canada, and France).

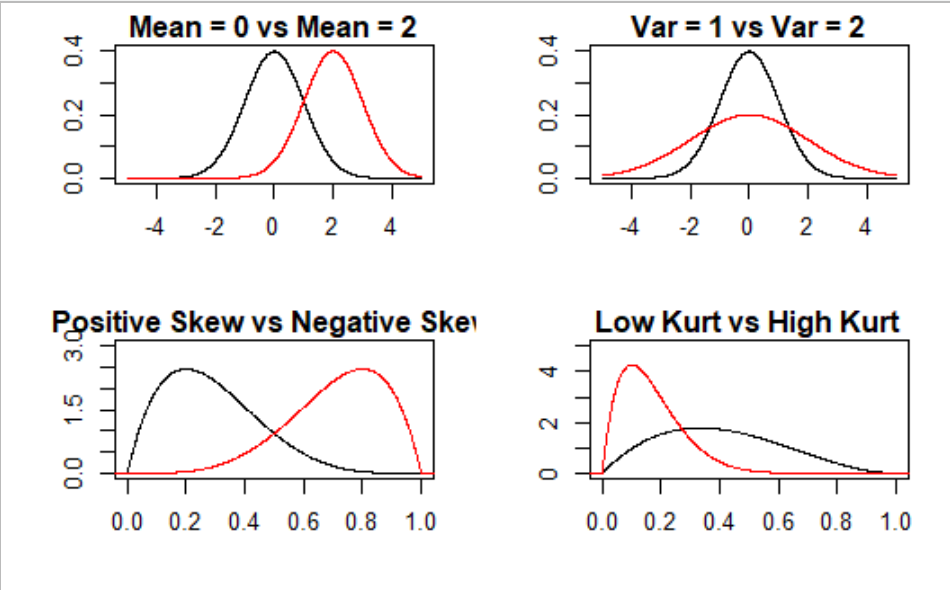
**Introduction:**

Life expectancy is statistical measure of the average time a person has left to live based on what year of birth and the current year. This project will be considering the life expectancies of persons at specific ages in the year 2010 by country.

The distinction between underdeveloped and developed countries is based on economic activity, education level, healthcare availability, industrialization, and other standard of living metrics. Physicians per capita is a good measure of how readily available health care is to the population. Literacy rate is used as an estimator of the country’s education level. Lastly, GDP per capita is representative of a country’s level of infrastructure (i.e. the ability to purchase the resources required for advanced healthcare), as well as the ability of the average citizen to afford healthcare.

**Mathematical Background:**

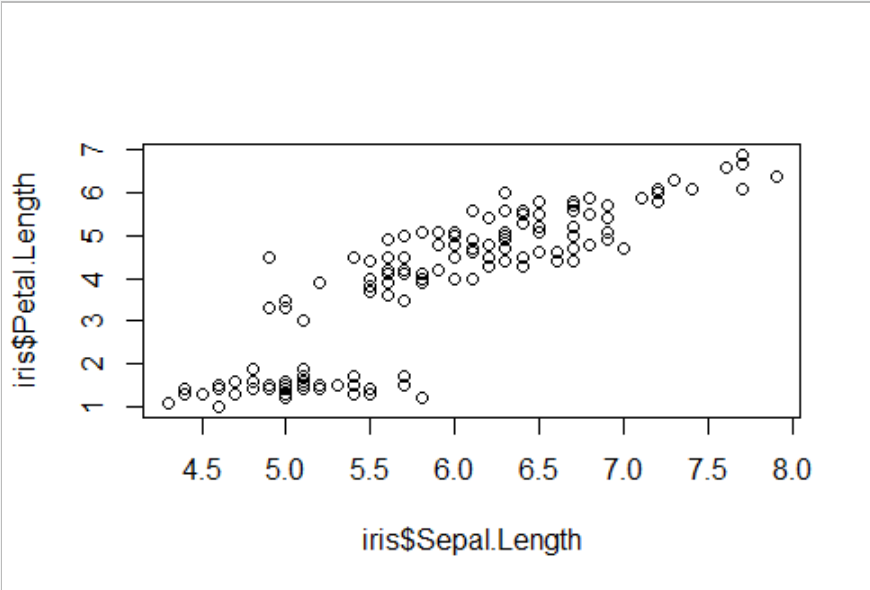
This paper will make use of the statistics of mean, standard deviation, skewness and kurtosis. Mean is the first moment and is a measure of the central location of the data. The second moment, variance measures how dispersed the data is from the mean. The third moment, Skewness is a measure of symmetry, and the fourth moment kurtosis is a measure of tailedness and outliers.



**Black line shows (Mean=0, Var=1, Positive Skew, Low Kurt)**

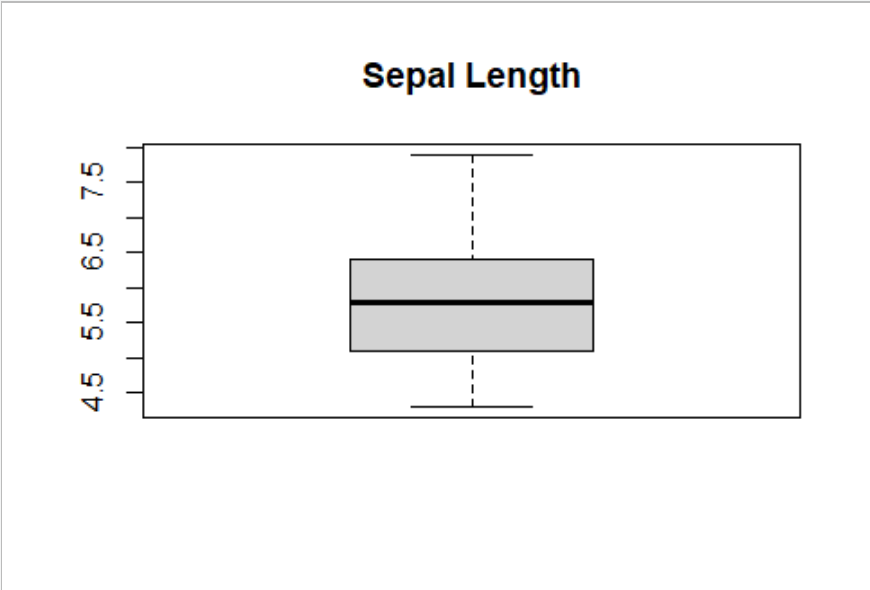
**Red line shows (Mean=2, Var=2, Negative Skew, High Kurt)**

Scatter Plots are used to plot data points with two variables on a graph in order to show a relationship between the two variables



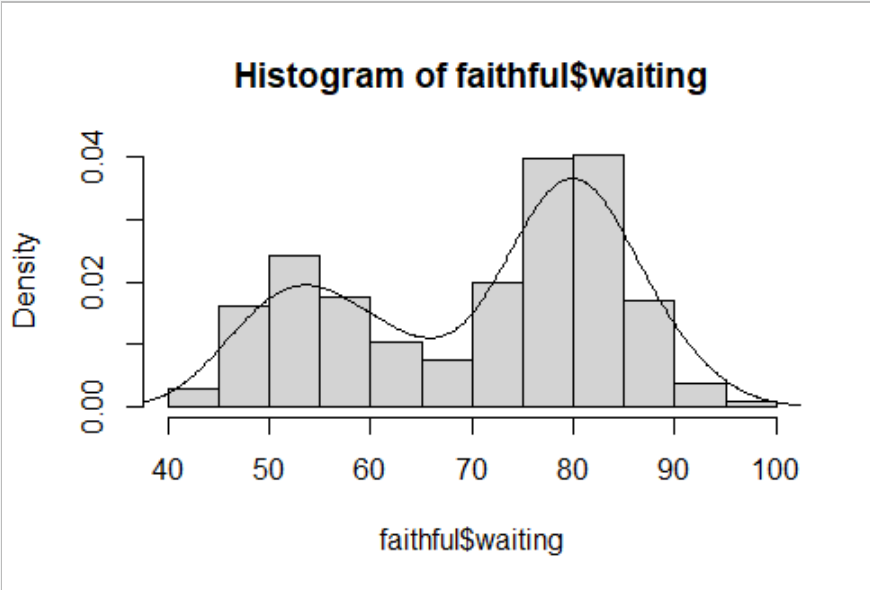
This scatter plot is plotting the petal length vs sepal length from the iris data set. This graph helps to visualize the moderate positive correlation between the two variables.

Box and whiskers plots are used for visualizing locality in the minimum, first quartile, median, third quartile, and maximum. The plot also shows the spread and skewness of the data.



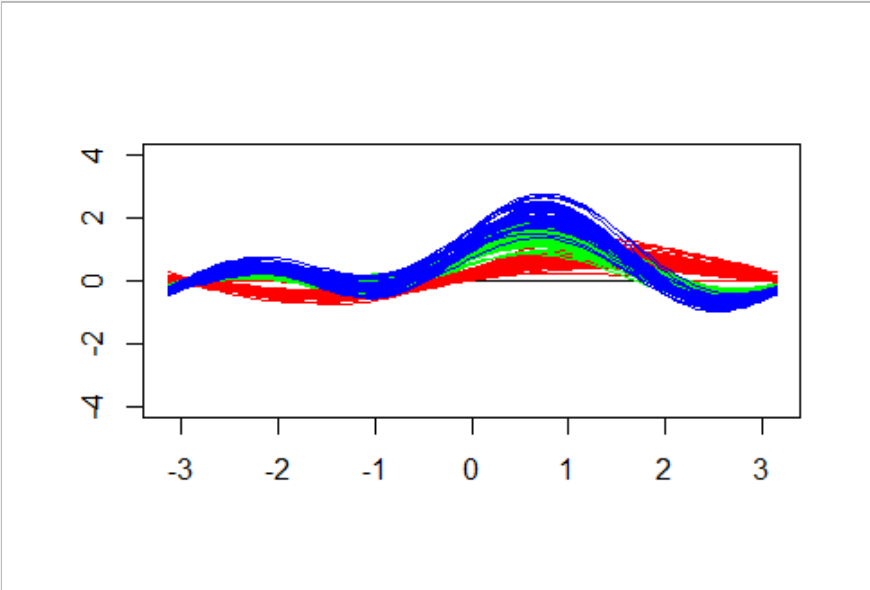
This box and whisker plot of the Sepal length shows most of the sepal lengths in the data set fall between 5 and 6.5, as well a slightly positive skew in the data.

Kernel density estimation is a non-parametric function used to estimate an unknown density function for which we have data for. Often useful for bimodal data.



This data for waiting times between eruptions of old faithful geyser in Yellowstone would be impossible to construct a density function for using standard probability density functions, because of the two peaks.

Andrews curves are used to visualize high dimensional data though the use of data transformation. Clear distinction and structure in the curve can help to class independent variables that have similar characteristics.



This Andrews curve of the iris data shows a clear distinction between two of the species

Monte Carlo simulations are used when probability weights are assigned to possible outcomes for an experiment. Then, numerous simulations are ran by generating random numbers and assigning them values based on the probability weights.

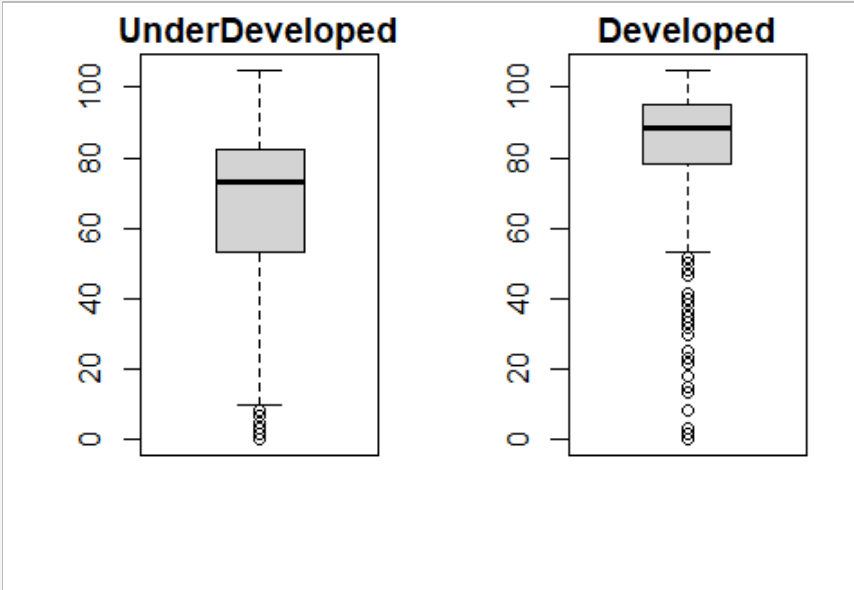
**Numerical Experiments:**

**Example of Data Set (2010)**

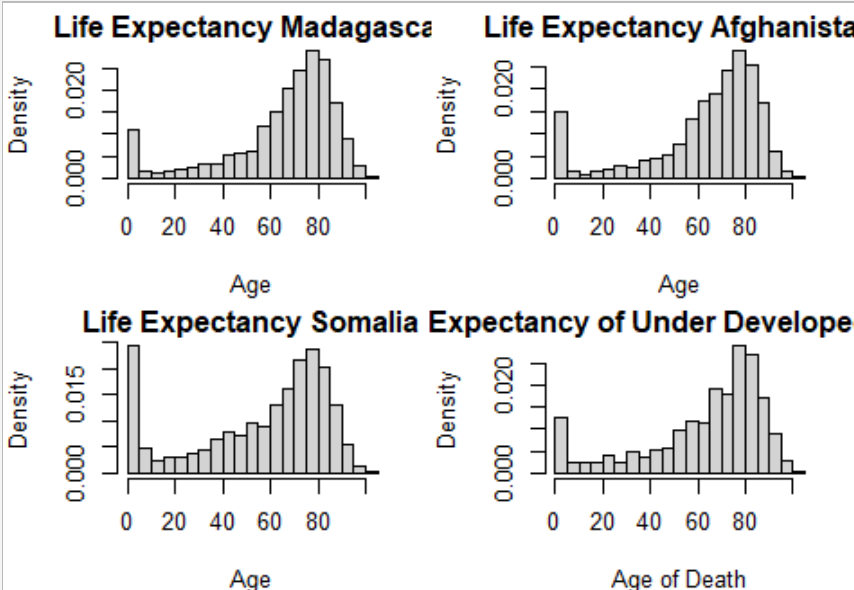
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | **0** | **1** | **5** | **10** | **…** | **95** | **100+** |
| Madagascar | **100000** | **96232** | **94550** | **93552** | **…** | **1239** | **181** |
| United States | **100000** | **99404** | **99303** | **99245** | **…** | **9980** | **2239** |

This data shows 100,000 people being born in 2010 and every five years shows the expected remaining living people from the original 100,000. Using this data we can run a Monte Carlo simulation by dividing these numbers by 100,000 and subtracting them from 1. This gives us the probability that someone dies by this age. Then we generate numbers from runif() between 0 and 1, and whatever 2 age probabilities the random number falls between signifies the age that number “dies” at. Example: runif(x)=.65, P(dead by 70)=.62, P(dead by 75)=.64, P(dead by 80)=.68. In this case we will count this random number as dying at 80.

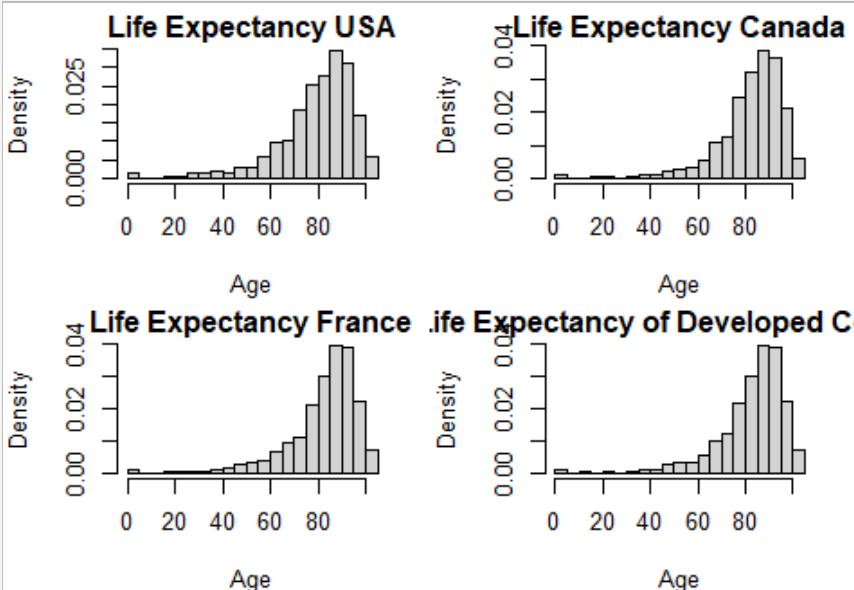
Running this simulation many times and keeping count of the “death” age of each random number we can create the following life expectancy graphs by country (Underdeveloped and Developed data are produced by averaging the other 3 countries data.)



**Boxplots of Underdeveloped and Developed Countries Life Expectancy**



**Histograms of Life Expectancy of Underdeveloped Countries**



**Histograms of Developed Countries Life Expectancies**

Boxplot Data Under Developed Countries:

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 53.33 73.33 63.84 82.08 105.00

Boxplot Data Developed Countries:

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 78.33 88.33 83.80 95.00 105.00

Histogram Data:

Mean SD Skewness Kurtosis

Madagascar 67.81750 23.73723 -1.4144029 4.507603

Afghanistan 65.57250 24.98393 -1.3745509 4.155765

Somalia 58.11750 28.96400 -0.7986955 2.408252

USA 81.85750 16.67262 -1.7423113 7.520760

Canada 84.66750 14.77705 -2.0785264 10.109937

France 84.87500 14.96567 -1.8711515 8.492799

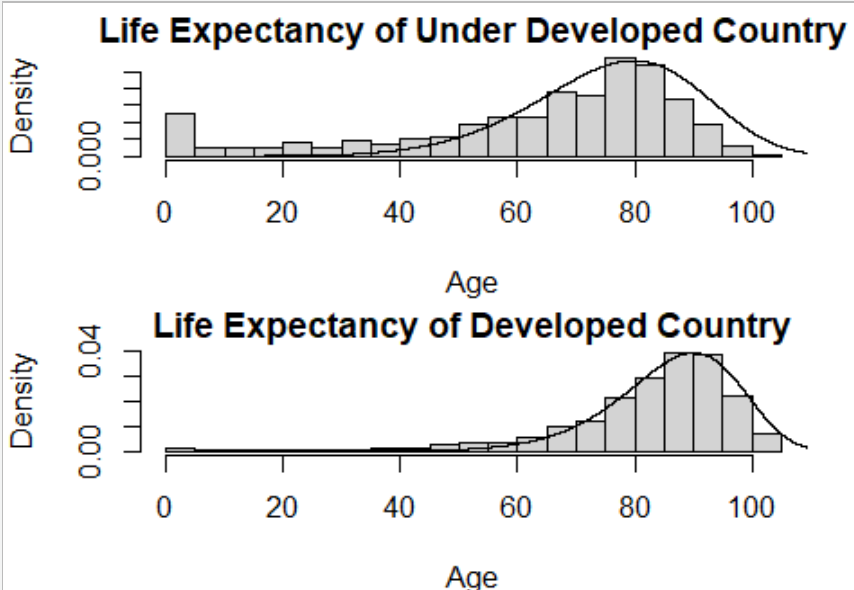
**Under Developed 63.83583 25.89505 -1.1958831 3.690540**

**Developed 83.80000 15.47178 -1.8973297 8.707832**

Looking at the boxplot data it is evident that both countries have negatively skewed data. Also, the boxplot shows that all quartiles of life expectancy are lower for underdeveloped countries than for developed countries. More interestingly, the spread of the data for underdeveloped countries is much large, meaning there is a larger age range to have a high probability of death. Even speaking to the kurtosis, there is significantly more outliers for the developed countries because dying young in a developed country is considered an outlier.

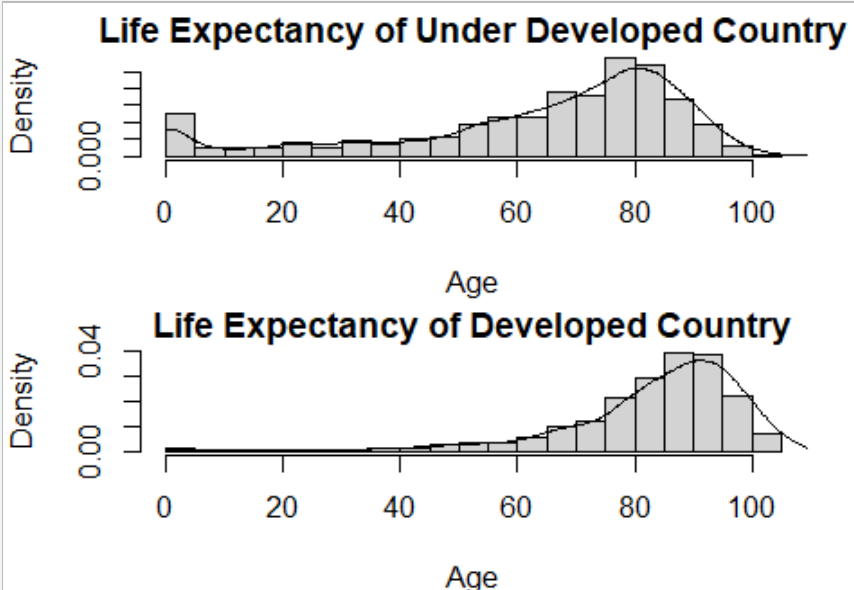
Further, looking at the histograms confirms that the life expectancies are higher for the developed countries for all cases. Most notably, the infant mortality rates for underdeveloped countries is much higher. The negative skew of the data is very evident and more pronounced in the developed countries. Also, comparing the sharpness of peaks and weight of the tails, the kurtosis for developed countries is higher than for the underdeveloped countries.

In order to make more precise predictions about life expectancy, a continuous probability distribution is need. The standard probability distribution to represent mortality is the Weibull distribution. Plotting the Weibull distribution with shape parameter=6.2 and scale parameter=82 over the underdeveloped countries histogram and plotting the Weibull distribution with shape parameter=9.7 and scale parameter =91 over the developed countries histogram, the following is produced.



**Life Expectancy histograms with Weibull curve**

While the Weibull curve does a decent job at capturing the developed countries data, it does a very poor job at showing the infant mortality rate of the underdeveloped countries. This is because high infant mortality rates create bimodal data, which cannot be created by any parametric probability density function. For cases like this a kernel density estimator can be used to create a non-parametric function that captures both peaks of the data.



**Kernel Density Estimator using gaussian kernel with bw=3 starting at x=0**

The kernel is able to capture the structure of the data significantly better than the Weibull curve, especially the high death rates of the young people in underdeveloped countries.

Using global rankings for life expectancy, physicians per capita, literacy rate, and GDP per capita, the following data and graphs can be produced.

Life Rank Phys Rank Literacy Rank GDP per Capita Rank

Madagascar 47 35 28 9

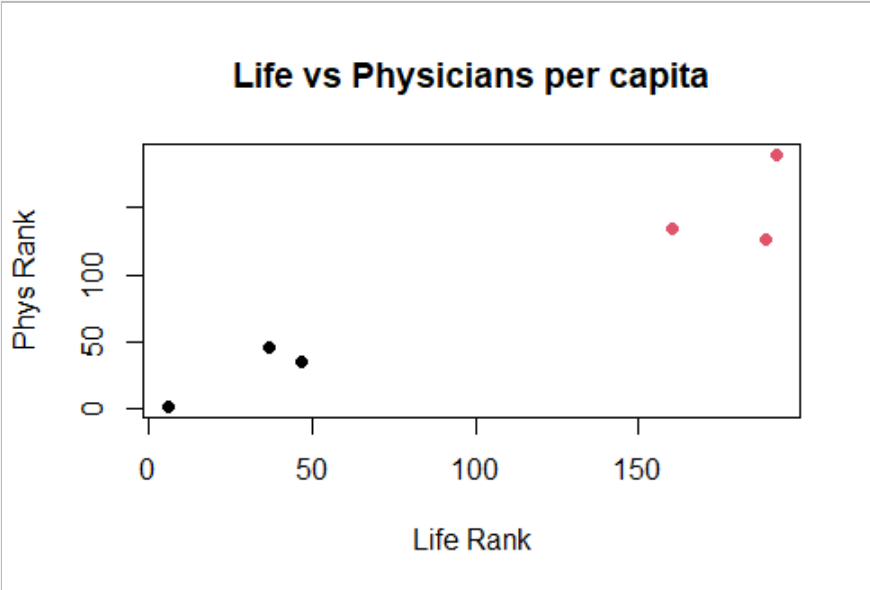
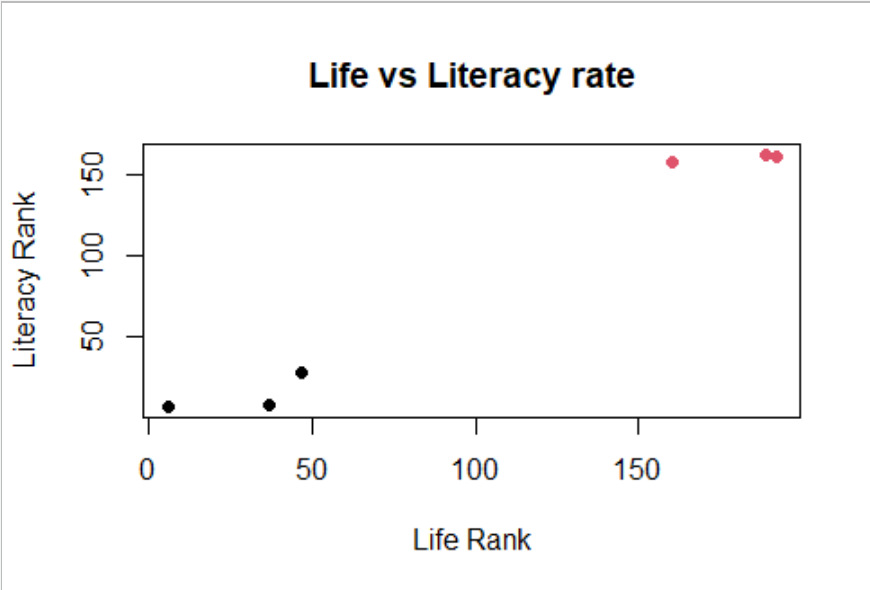
Afghanistan 37 46 8 15

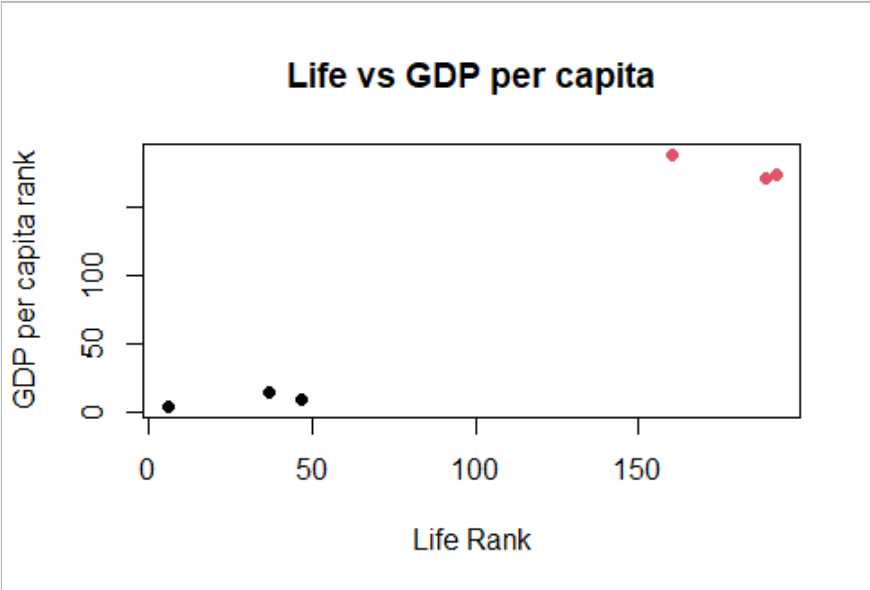
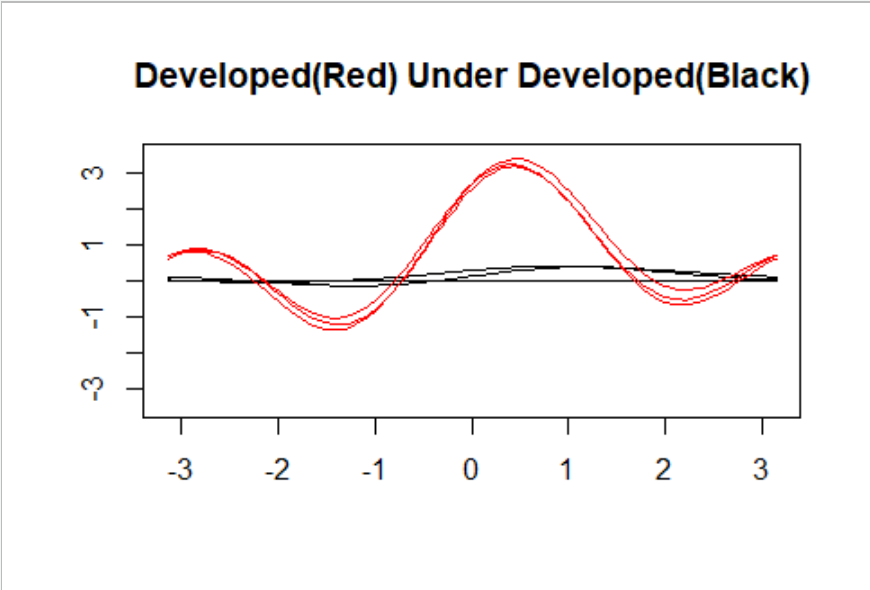
Somalia 6 1 7 4

USA 160 134 157 189

Canada 189 126 162 172

France 192 190 161 174

Although there is a limited amount of data to create plots with, there is still clear positive correlation between the data. With more data, more conclusions could be made about the strength of correlation, but from the limited sample it seems that physicians per capita has the strongest linear correlation with life expectancy. This makes sense intuitively, because the number of trained healthcare workers per person should have a strong correlation and even causation for the health of the general public. Additionally, the Andrew’s curve shows that there is a clear distinction in groups between developed and underdeveloped countries.

**Future work:**

In the context of an insurance company knowing how a countries life expectancy data is correlated with other economic and medical factors can help a company to create a more accurate prediction of future life expectancies, allowing them to create more accurate models for Life insurance premiums.

**References:**

**“World Population Prospects - Population Division.” United Nations, United Nations, https://population.un.org/wpp/.**

**Mahler, Daniel Gerszon, et al. “World Bank Open Data.” Data, 13 Apr. 2022, https://data.worldbank.org/.**