E14 Final Project

Kelvin A. Mike S. Scott E.

May 6, 2019

Background

Life expectancy is an important statistic that estimates how long a person will live. Life Expectancy Index is defined by the Human Development Reports (HDR) as ”Life expectancy at birth expressed as an index using a minimum value of 20 years and a maximum value of 85 years.” The Life Expectancy Index of a country is affected by many factors such as economic status, demographics, and the level of technological development of a country. In this project, we focused on how GDP per capita, mobile phone subscriptions per 100 people, and percent of land area that is forest correlate to life expectancy in various countries. The HDR define GDP per capita as ”GDP in a particular period divided by the total population in the same period.” Mobile phone subscriptions per 100 people is defined as ”Number of subscriptions for the mobile phone service, expressed per 100 people.” The percent forest area of a country is defined by the HDR as ”Land spanning more than 0.5 hectare with trees taller than 5 metres and a canopy cover of more than 10 percent or trees able to reach these thresholds in situ. It excludes land predominantly under agricultural or urban land use, tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens. Areas under reforestation that have not yet reached but are expected to reach a canopy cover of 10 percent and a tree height of 5 metres are included, as are temporarily unstocked areas resulting from human intervention or natural causes that are expected to regenerate.” All of the data in this project were published in the 2015 Human Development Reports.

We concentrated on how percent forest area, mobile phone subscriptions, and GDP per capita are related to Life Expectancy. We chose these data because we hoped that they would represent different aspects of a country, and we wanted to discover some kind of relationship between the data. We hypothesize that GDP per capita is indicative of the wealth of a country, percent forest area is indicative of the climate of a country, and cell phone subscriptions indicate the level of technological development of a country. The Human Development Reports contain economic, health, environmental, and many other types of data from numerous countries. Most of the data is obtained either from government agencies within a country, or larger multinational groups such as organizations within the United Nations.

Research Questions and Hypotheses

For this project, we sought to analyze how the mobile phone subscriptions per 100 people and the total percent forest area of a country affected the life expectancy index of people living the sampled countries. We wanted to determine if there was some kind of correlation between the data. We also wanted to find out if we could come up with a model with which we can run tests on and potentially predict the life expectancy index of a hypothetical country with respect to a number of mobile phone subscriptions or percent forests area. We expected that a higher number of mobile phone subscriptions per 100 people would be an indicator of wealth, which would therefore correlate with a higher life expectancy index. Our group did not have a strong prediction about how percent forest area and climate would affect life expectancy index of a country. After examining some of our analysis, we also decided to test the relationship between per capita GDP and Life Expectancy Index and per capita GDP and mobile phone subscriptions. We predicted that as the GDP of a country increased, the Life Expectancy Index would also increase because a wealthier country is more likely to have better technologies and more advanced medicines.

Theory

We began our project by computing different statistics about our data such as the sample mean, mode, and variance for the set of data. There is no significant underlying theory used to compute these values.

After computing different test statistics, we created some graphs of the data. While examining the histogram for the number of mobile phone subscriptions per 100 people, we noticed that the data looked like they were normally distributed. We wanted to test if the data were normally distributed, so we normalized our sample data. We then used the pdf for a normal distribution. The pdf, $f(x)$, is given as $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x-\mu^2}{2\sigma^2}}$ where σ is the standard deviation of the data and μ is the mean of the data. Since our data was from 178 countries (n=178), we invoked the Central Limit Theorem and assumed our sample mean to be equal to the actual mean of the data and the sample variance to be equal to the actual variance of the data. The Central Limit Theorem states that the any random variable, regardless of the distribution, will behave as if it is normally distributed when n is large. Using this assumption that n = 178 is large, we were able to create a probability density function for our normalized data, and compare the two to determine if the mobile phone subscriptions data was distributed normally.

In this project, we also used hypothesis testing of regression estimators and correlation coefficients to determine correlations between different sets of data. First, we computed several correlation coefficients. The correlation coefficient σ_{xy} is a way to quickly measure the relationship between two data sets (x and y). A positive value for σ_{xy} suggests a positive correlation, a negative value suggests a negative correlation, and a value of zero suggests no correlation. σ_{xy} is defined as $\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Our group also did hypothesis testing on regression lines fit to the data. We first did hypothesis testing on linear regression lines of the form $y = \theta_0 + \theta_1 x$. The estimated theta values for the regression lines were calculated using a python library. In each case, after the estimate was calculated, a t-test was performed to determine if the value of the the estimate was statistically different from zero and therefore statistically significant. For this t-test, the null hypothesis was $H_0 : \theta_k = 0$ and the alternative hypothesis was $H_0 : \theta_k \neq 0$. The test statistic used, T_0, is defined as $T_0 = \frac{\hat{\theta_k}}{\sqrt{S_e^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}$ where $S_{xx} = \sum(x_i^2) - n(\bar{x}^2)$ and $S_e^2 = \frac{SS_e}{n-2} = \frac{S_{yy} - SS_R}{n-2} =$ $= \frac{\sum(y_i^2) - n(\bar{y}^2) - \beta\hat{S}_{xy}}{n-2} = \frac{\sum(y_i^2) - n(\bar{y}^2) - \beta\sum\widehat{(x_i^2 y_i^2)} - n(\bar{x}^2\bar{y}^2)}{n-2}$. T_0 was then compared to the rejection region determined by a two-tailed t-test using $\alpha = 0.1$ and $n =$ the number of countries. If $-t_{\frac{\alpha}{2}} < T_0 < t_{\frac{\alpha}{2}}$, then we fail to reject our null hypothesis, and the value of the estimator is not significantly different from zero. If T_0 is outside of this region, we reject the null hypothesis and the estimator θ_0 is statistically significant. For θ_1, the computed test statistic was $T_1 = \frac{\hat{\theta_k}}{\sqrt{\frac{S_e^2}{S_{xx}}}}$, and the rejection region is calculated in the same manner as T_0. If the value of T_1 is outside of the fail to reject region, then the value of θ_1 is determined to be statistically significant.

We also performed hypothesis tests on the multi variable regression model based on the equation $y = \theta_0 + \theta_1 x + \theta_2 x^2$. The first step in hypothesis testing is finding the estimators for the multi variable regression. The first step is defining the X matrix: $X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_a & x_a^2 \end{bmatrix}$

From X we can determine the A matrix $A = X^T X$ and the C matrix is define as $C = X^T y$. Using the C and A matrices, we can determine the values of the estimators, the theta values, for the multi variable regression equation. $\hat{\theta} = \begin{bmatrix} \hat{\theta_0} \\ \hat{\theta_1} \\ \hat{\theta_2} \end{bmatrix} = A^{-1}C$

Once we had our theta values, we performed a t-test using the null hypothesis $H_0 : \theta_k = 0$ and the alternative hypothesis $H_A : \theta_k \neq 0$. The test statistic used in the t-test was T_0 which, in this case, was defined as $T_0 = \frac{\hat{\theta_k}}{S_{\theta k}}$. $S_{\theta k}$ is defined as the variance of the theta estimator. In order to calculate $S_{\theta k}$, we must calculate the co-variance matrix, which is the product of $A^{-1}S_e^2$. S_e^2 is defined as $S_e^2 = \frac{SS_e}{n-p}$. To calculate SS_e first find a vector of values on the regression line $\hat{y} = X\hat{\theta}$. Then calculate a vector of residuals $e = y - \hat{y}$. From this SS_e can be calculated as the product of e and e-transpose $SS_e = e^T e$. Next, the co-variance matrix $A^{-1}S_e^2$ is calculated, and the values along the diagonal are the variance estimates for the theta parameters. Finally, the test statistics can be calculated. Then using a t-table, we were able to determine the fail to reject region for our null hypothesis. The region was defined as $-t_{\frac{\alpha}{2}} < T_0 < t_{\frac{\alpha}{2}}$ where alpha $= 0.1$. If the test statistic T_0 falls in this region, we fail to reject our null hypothesis and the theta estimator is not statistically different from zero. If the value of T_0 does not fall within this region, we reject our null hypothesis and the estimator is determined to be statistically significant.

In this project we also computed a regression function with 10 theta estimators. The theta estimates for the data are found using the same procedure used to find the quadratic regression function above. The X matrix, however, will have 11 columns instead of 3. For the 10th degree polynomial case

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 & x_1^6 & x_1^7 & x_1^8 & x_1^9 & x_1^{10} \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 & x_2^6 & x_2^7 & x_2^8 & x_2^9 & x_2^{10} \\ . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . \\ 1 & x_a & x_a^2 & x_a^3 & x_a^4 & x_a^5 & x_a^6 & x_a^7 & x_a^8 & x_a^9 & x_a^{10} \end{bmatrix}$$

A regression model with 10 theta values does provide a highly accurate approximation of the data, but it is not an ideal fit for the data. It is very unlikely that the life expectancy of a country could be affected by 10 different variables. There also is a point where the R^2_{adj} value reveals that it is no longer beneficial to increase the number of parameters to find a fit for the data. For these reasons, we decided that the 10th degree polynomial was not the best fit for the relationship between GDP per capita and Life Expectancy Index. A logarithmic regression fit is also used later in the project. The theory behind this was not explored, and all computations were made strictly by python functions.

Procedures

The data used in this project was obtained from the Human Development Index. In order to analyze this data, we first had to trim the raw data to ensure that the data was consistent across data sets. We also ensured that all data was from 2015. All of the data were stored in a CSV file and read into a Pandas DataFrames in a Jupyter Notebook. We then used the built-in functions of a DataFrame to obtain the mean, mode, variance, and median for each of our data sets. We also used a built-in function for the DataFrame to plot the histograms shown in Figures 1 through 4. Our group utilized many functions from the matplotlib.pyplot library for labeling the axes of graphs and inserting lines and test into the figures. Further statistics such as the variance and standard deviation were calculated using built in functions from the pandas library.

When examining the histogram for the data for mobile phone subscriptions, we determined that the data may be distributed normally. To test this, we normalized the phone subscriptions data. We then rounded this data to two decimal places, so that the data could be portrayed well on a histogram. We also used a built-in function to from the scipy.stats library to create a probability density function for a normal distribution with a mean equal to our sample mean and a standard deviation equal to our sample standard deviation. Using matplotlib.pyplot we plotted the histogram and the pdf on the same set of axes. The results are shown in Figure 3.

We also calculated confidence intervals for an estimation of the actual mean of the data for the mobile phone subscriptions data. We calculated intervals of 99% confidence for the actual mean by using a Z-test statistic. The 99% confidence intervals along with the sample mean of the data are shown on a histogram of the data in Figure 4.

Next we began to create regression lines for the data, and do hypothesis testing on the parameters. We first created scatter plots of percent forest area vs. life expectancy index and mobile phone subscriptions per 100 people vs life expectancy index using the matplotlib.pyplot library. Correlation coefficients were then computed using a function from the numpy library. Then the the LinearRegression object from the sklearn.linear_model library was used to create linear regression lines for each scatter plot. The sklearn.linear_model functions returned the theta values for the regression line, and then hypothesis tests were performed using these theta values. Significance tests were done using the equations $T_0 = \frac{\hat{\theta_k}}{\sqrt{S_e^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}})}}$ and $T_1 = \frac{\hat{\theta_k}}{\sqrt{\frac{S_e^2}{S_{xx}}}}$. All of the values for the variables needed for the hypothesis testing were computed using the numpy and stats libraries. The necessary t-values were found in a t-table.

All polynomial regression lines were computed using the polyfit function, which allows the user to pick the degree of the polynomial that will be fit to the data. We determined that the plot of GDP per capita vs. Life Expectancy Index did not appear to have a linear relationship, so we fit polynomial functions to the data. Hypothesis tests were performed on the second degree polynomial to determine if the theta values were statistically significant. The procedure for these hypothesis tests can be found in the Theory section.

After testing multiple polynomial functions and linear regression functions, we decided to try a logarithmic fit for the scatter plots of GDP vs. Life Expectancy and GDP vs. Mobile Phone subscriptions. For all of the logarithmic regression lines, the polyfit function was used with a degree of one. The polyfit function was, however, given to the logarithm of the data instead of the data itself in order to generate a logarithmic fit.

Results

We first created some graphical representations for each set of data to display the distribution of the data, and how the test statistics compare to the data. Below in Figure 1 is a histogram of the data for Percent Forest Area. This histogram gives us valuable information about the percent forest, and how the different test statistics compare to the data as a whole. We created similar histograms for all four data sets.

When examining the histogram for mobile phone subscriptions per 100 people, we thought that the data looked like a normal distribution. The data is shown in Figure 2 below.

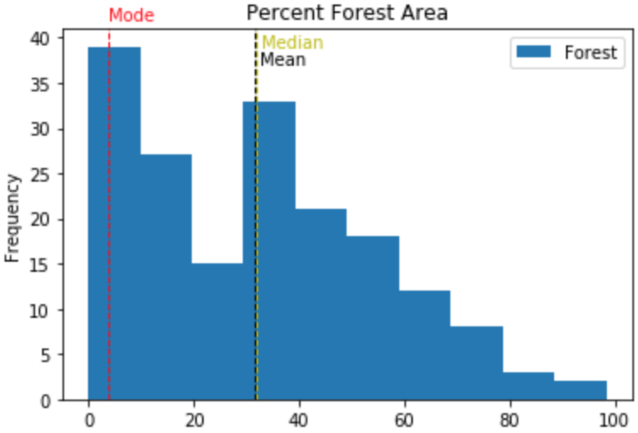

Figure 1: Percent Forest Area Histogram

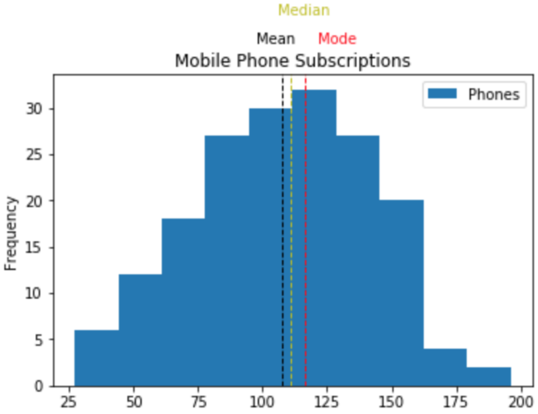

Figure 2: Mobile Phone Subscriptions Histogram

We then normalized our data, and created a histogram of the normalized data. On this histogram we also plotted the probability density function for a normal distribution with our sample mean and sample variance. Figure 3 shows the histogram of the normalized data with the density function.

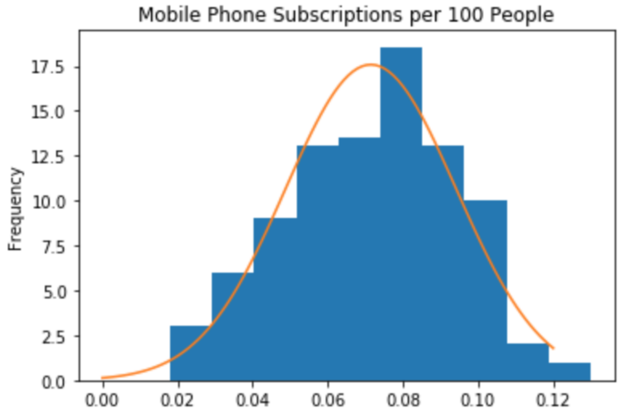

Figure 3: Normalized Phone Subscriptions Histogram

We then decided to determine a confidence interval of the true mean of the mobile phone data assuming that it was normally distributed. We determined these confidence intervals with a 99% confidence (alpha = .01). The intervals and sample mean are shown in Figure 4 below.

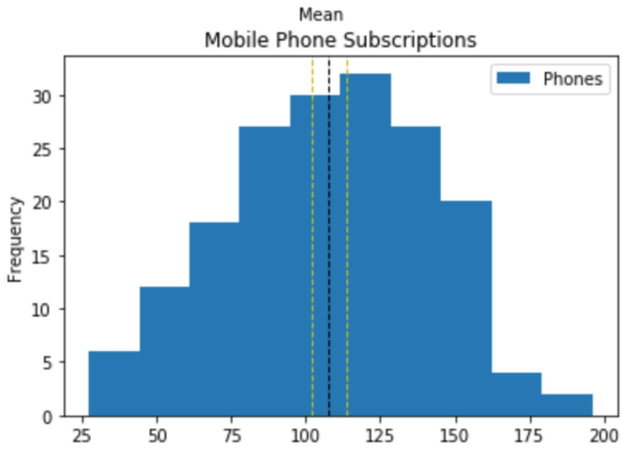

Figure 4: 99% Confidence Interval for Actual Mean estimation

Next, we wanted to test for any correlation between Percent Forest Area and Life

Expectancy Index, and between Mobile Phone Subscriptions and Life Expectancy Index. To do this we first created scatter plots of the data to see if there appeared to be any correlation. The scatter plots are shown below in Figures 5 and 6.

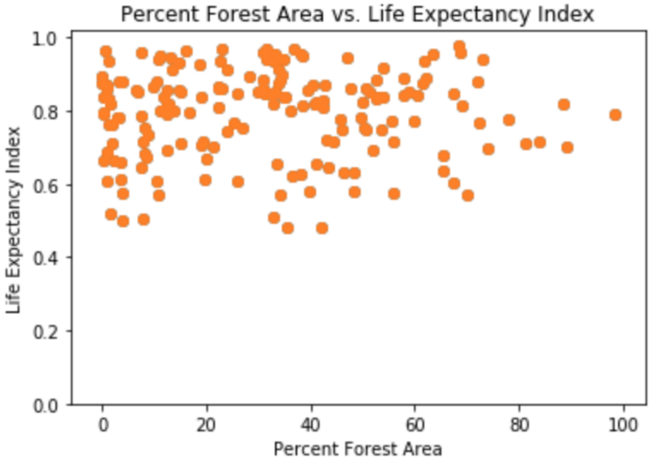

Figure 5: Percent Forest Area vs Life Expectancy Scatter Plot

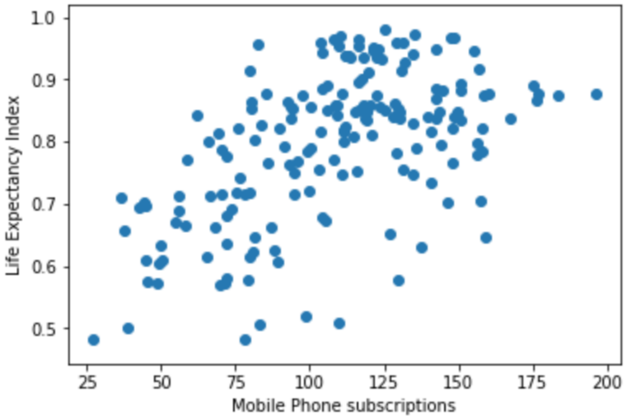

Figure 6: Phone Subscriptions vs Life Expectancy Scatter Plot

We then calculated correlation coefficients. The correlation coefficients were calculated using a python library. The correlation coefficient for Percent Forest Area and Life expectancy index is $\sigma = 0.02640931$ The correlation coefficient for Mobile Phone Subscriptions and Life expectancy index is $\sigma = 0.54686818$.

Next, we fit linear regression lines to both data sets. Figure 7 and Figure 8 show the computer generated linear regression lines for Percent Forest Area and Mobile Phone subscriptions respectively.

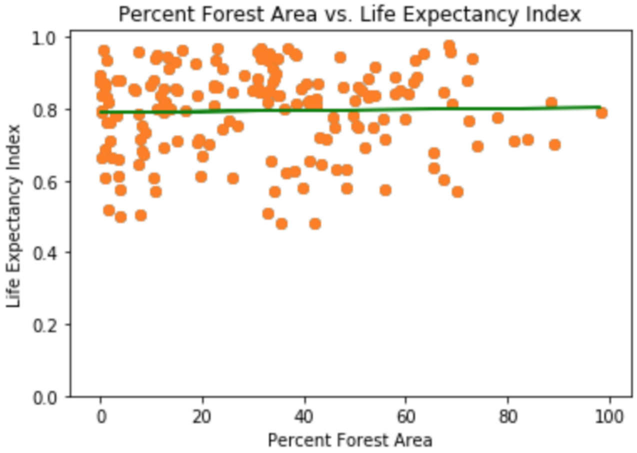

Figure 7: Linear Regression of Forest Area vs Life Expectancy Index

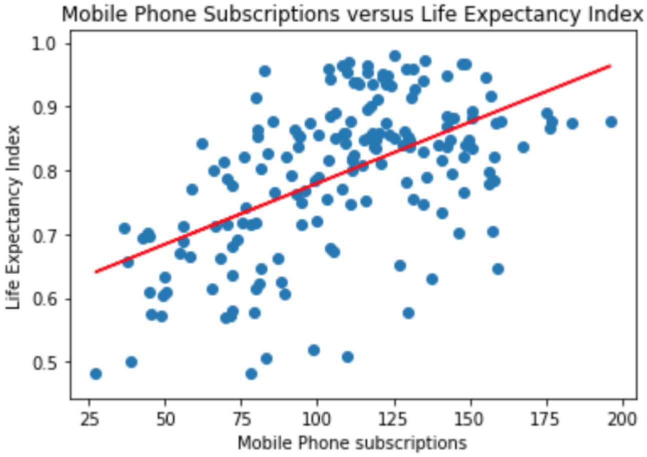

Figure 8: Linear Regression of Phone Subscriptions vs Life Expectancy Index

From the Linear regression line for Percent Forest Area versus Life Expectancy

Index, the slope of line, $\theta_1 = 0.00014$, and the intercept of the line, $\theta_0 = 0.79030$ From this, $T_{\theta 0}$ is calculated to be $T_{\theta 0} = 3.57920$. The confidence interval for a t-test where alpha = .05 and n = 178 is t = +/- 1.6535. Since $T_{\theta 0}$ is outside of this region, we reject our null hypothesis, and the value is statistically significant. $T_{\theta 1}$ is also calculated calculated to be $T_{\theta 1} = 0.02426$. The confidence interval for a t-test where alpha = .05 and n = 178 is t = +/- 1.6535. Since $T_{\theta 0}$ is within of this region, we fail to reject our null hypothesis and the value is not statistically significant.

From the Linear regression line for Mobile Phone Subscriptions versus Life Expectancy Index, the slope of line, $\theta_1 = 0.00191$, and the intercept of the line, $\theta_0 = 0.58843$ From this, $T_{\theta 0}$ is calculated to be $T_{\theta 0} = 264.05956$. The confidence interval for a t-test where alpha = .05 and n = 178 is t = +/- 1.6535. Since $T_{\theta 0}$ is outside of this region, we reject our null hypothesis, and the value is statistically significant. $T_{\theta 1}$ is also calculated calculated to be $T_{\theta 1} = 97.06872$. The confidence interval for a t-test where alpha = .05 and n = 178 is t = +/- 1.6535. Since $T_{\theta 0}$ is outside of this region, we reject our null hypothesis, and the value is statistically significant.

We proceeded to our analysis of a correlation between GDP and Life Expectancy Index. We first created a histogram of the GDP data.

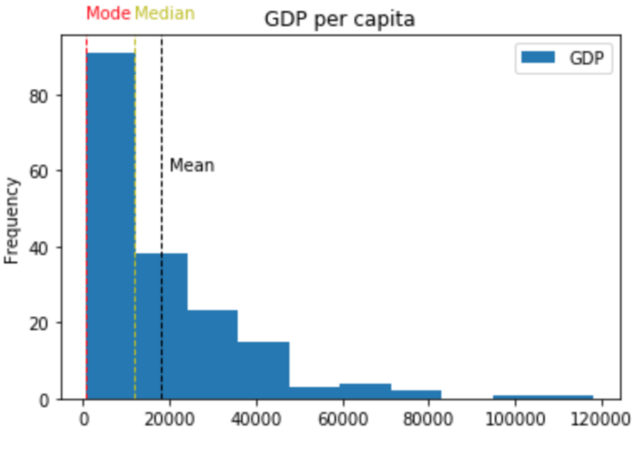

Figure 9: Histogram of GDP Data

We then began to test for a correlation between GDP and Life Expectancy Index. The calculated correlation coefficient $\sigma = 0.64157$. Then a regression line was fit to the data. It is shown in Figure 10 below.

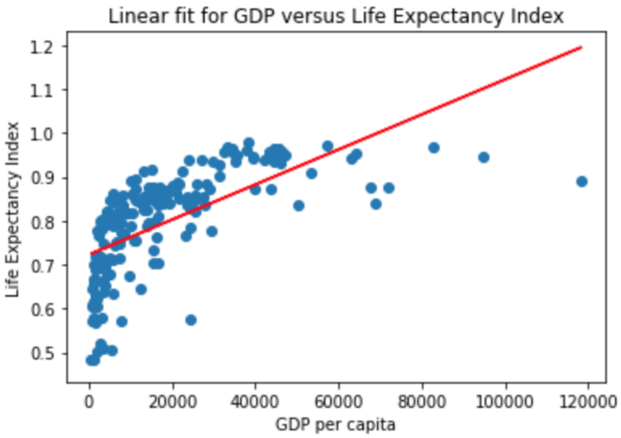

Figure 10: Linear Regression of GDP vs. Life Expectancy Index

The slope $\theta_1 = 3.99997e - 06$ and the y-intercept of the line $\theta_0 = 0.72243$. From these values the values of the test statistics T_0 and T_1 were calculated to be 0.00487 and 0.00071 respectively. The confidence interval for a t-test where alpha = .05 and n = 178 is t = +/- 1.6535. Since $T_{\theta 0}$ and $T_{\theta 1}$ are both inside of this region, we fail to reject both null hypotheses, so the values are not statistically significant.

We then fit a quadratic regression function to the GDP versus Life Expectancy Index Data. The result is shown in Figure 11 below.

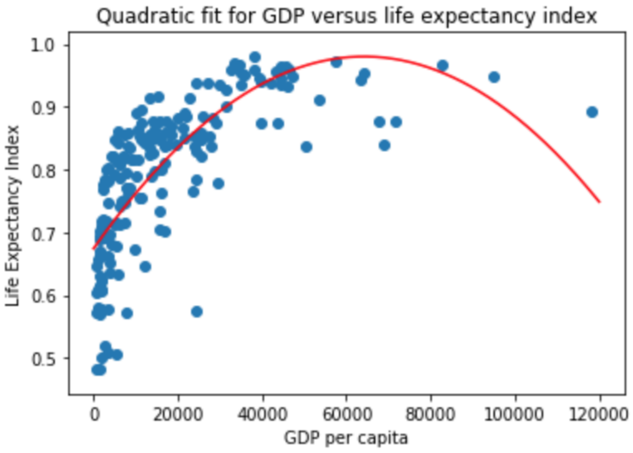

Figure 11: Quadratic Regression of GDP vs. Life Expectancy Index

For the quadratic regression function, $\theta_2 = -7.42320e^{-11}, \theta_1 = 9.52948e^{-6}$, and $\theta_0 = 6.74079e^{-1}$

The variances were calculated to be $\theta_2 = 7.84852e^{-23}, \theta_1 = 5.28762e^{-13}, \theta_0 = 9.79469e^{-05}$. Using these values, the test statistics were calculated to be $T_0 = 6882.08360, T_1 = 18022230.03584, T_2 = -945808341061.9396$ For a 95% confidence interval, alpha = .05, t = +/- 1.6535 In all cases, the test statistic T_k is inside of the rejection region, so we reject our null hypothesis for all three estimators, and all three are statistically significant.

A 10th degree polynomial regression function was also fit to the data. The 10th degree fit of GDP versus Life Expectancy is shown in Figure 12 below.

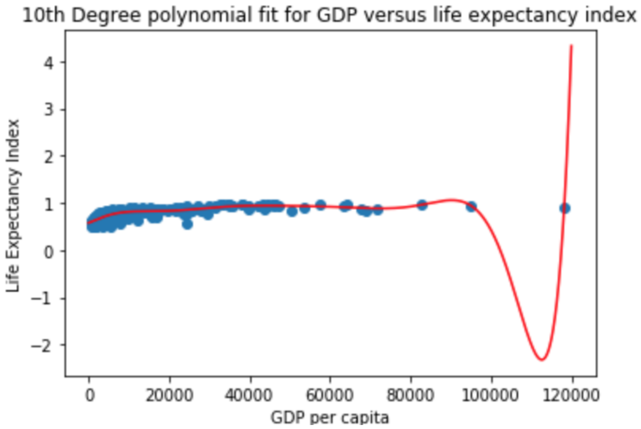

Figure 12: 10th Degree Regression of GDP vs. Life Expectancy Index

After further examination of the data, we fit a logarithmic regression function to the GDP versus Life Expectancy Index data. The result is shown in the figure below.
We also created scatter plots for the GDP per capita versus the number of Mobile Phone Subscriptions per 100 people. Then we fit a linear regression line and a logarithmic regression line to the data. The linear and logarithmic regression lines are shown in the figures below.

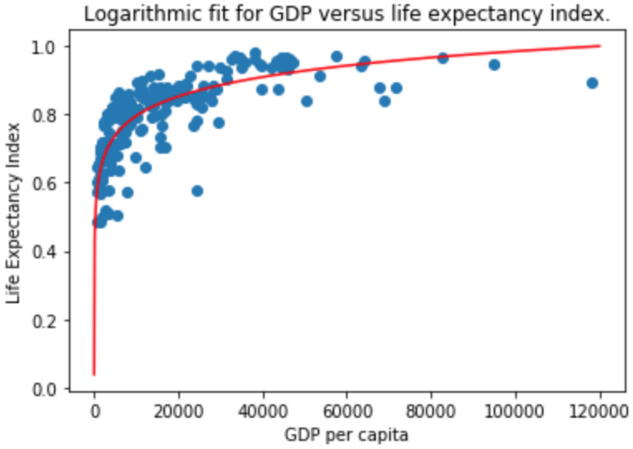

Figure 13: Logarithmic Regression of GDP vs. Life Expectancy Index

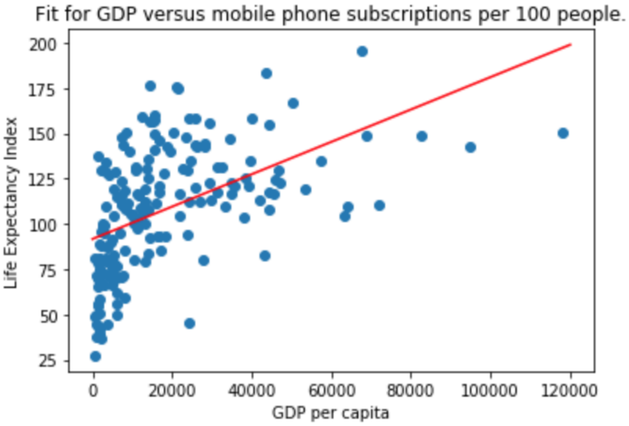

Figure 14: Linear Regression of GDP vs. Mobile Phone Subscriptions

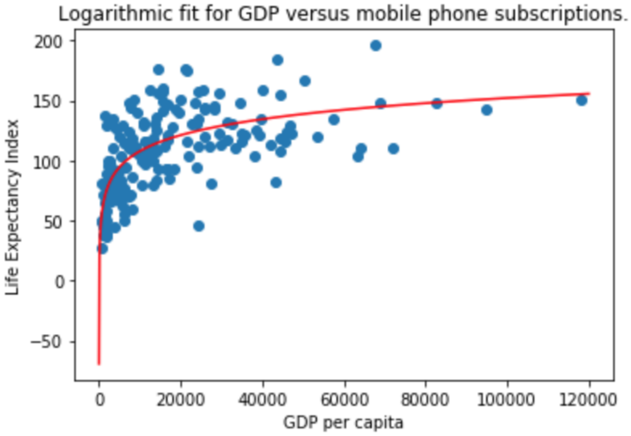

Figure 15: Logarithmic Regression of GDP vs. Mobile Phone Subscriptions

Analysis

The generated histograms provided interesting data about the world as a whole. In Figure 1, for example, the mode of the data on the histogram reveals that most of the countries in the world have less than 10% forest area. The histograms for both percent forest area and GDP per capita suggest that the underlying processes are not normally distributed, because they are both skewed graphs. The histogram in Figure 2 is also interesting because it suggests that the number of mobile phone subscriptions per 100 people throughout the world is distributed normally. This conjecture is further supported by Figure 3 and how well the pdf aligns with the histogram of normalized data.

Originally we were amazed by the alignment of our sample mean and the 99% confidence interval for our prediction of the mean in Figure 4. However, after further investigation, we realized that in order to compute this confidence interval, we had to assume that the data was normally distributed. This assumption determines the confidence interval based on our sample mean, therefore this graphical representation does not provide any large insights into any of our research questions or hypotheses.

The scatter plot in Figure 5, along with the correlation coefficient for these two sets of data, suggest that there is no clear correlation between percent forest area and life expectancy. Our analysis of the regression line showed that the value for the slope of the line was not statistically significant from zero, which supports the suggestion that there is no correlation between the data sets. These results help to answer our research question regarding how percent forest area of a country affects life expectancy.

The scatter plot of mobile phone subscriptions versus life expectancy index suggests some type of positive correlation between mobile phone subscriptions and life expectancy. The correlation coefficient is over 0.5, which also supports a correlation. The hypothesis testing of the regression line revealed that both the y-intercept and the slope of the data were statistically significant. These results suggest a positive, linear correlation between life expectancy and mobile phone subscriptions per 100 people. It appears that these results suggest that countries with more cell phone subscriptions have a higher life expectancy, but it is difficult to argue that an increase cell phone subscriptions will cause an increase in life expectancy.

The analysis of the correlation between GDP per capita and life expectancy was proven to be unable to be described by a linear regression line because the theta values were not statistically significant. Since the linear fit was not satisfactory, we fit a quadratic function to the scatter plot of GDP per capita versus life expectancy index. Hypothesis testing of this quadratic regression function revealed that all of the parameters are statistically significant. The quadratic fit, however, began to decrease significantly as GDP became very large. This made us try to fit a logarithmic function to the data as shown in Figure 13. The logarithmic fit appeared to give the best fit. The same process was used for the analysis of GDP vs Mobile phone subscriptions, and we came to the same conclusion that the logarithmic fit was the best fit for the data.

Recommendations and Conclusions

In conclusion, we observed a statistically significant, positive correlation between life expectancy and mobile phone subscriptions per 100 people. This correlation made logical sense based on the assumption that a country with more mobile phone subscriptions was more technologically advanced than those with less phone subscriptions. While this correlation appears to exist, it is very unlikely that increased mobile phone subscriptions causes an increase in life expectancy. In order to prove a causal relationship, more research should be done to investigate the effects of having a mobile phone and life expectancy.

We also observe that the per capita GDP is positively correlated with life expectancy index, and with the mobile phone subscriptions. We determined that a logarithmic regression function was the function of best fit for the graphs of GDP versus life expectancy and mobile phone subscriptions. It is logical that an increase of GDP will lead to an increase in both technology and life expectancy. It is, however, unclear why a logarithmic fit is the best option for regression functions. We may speculate that once the average wealth of a person reaches a certain point, the medicine available becomes uniform and futher increases in wealth will have a negligible affect on the life expectancy of a country. In order to further our understanding the relationship between GDP and life expectancy, more research must be conducted to determine how GDP affects the life expectancy of a country.

Our analysis provides us with useful information, but does not provide us with the underlying phenomenon that explains the various distributions we discovered for each data set. We suggest that more work should be done in understanding the various factors that lead to the distribution of each data set.