STA218 Course Project

Susmitha Turaga - 1006184889

Part A: Data Set Selection (5 marks)

(a) Find a data set, either in R or from other published source. Write code to load your data set (if it is in-built) or import it (if it is from another source). If your data set is rather large, only include the first 50 observations. Finally, comment on the source of your data set.

```
library(readxl)
Before50 <- read_excel("/Users/susmi/Desktop/STA218/Before50.xlsx")
head(Before50)
```

```
## # A tibble: 6 x 23
##   Country  Year Status `Life expectanc~ `Adult Mortalit~ `infant deaths` Alcohol
##   <chr>   <dbl> <chr>             <dbl>            <dbl>           <dbl>   <dbl>
## 1 Democr~  2015 Devel~             76                139               6      NA
## 2 Syrian~  2015 Devel~             64.5              293               6      NA
## 3 Chad     2015 Devel~             53.1              356              46      NA
## 4 Belarus  2015 Devel~             72.3              196               0      NA
## 5 Mozamb~  2015 Devel~             57.6              355              60      NA
## 6 Benin    2015 Devel~             60                249              25      NA
## # ... with 16 more variables: percentage expenditure <dbl>, Hepatitis B <dbl>,
## #   Measles  <dbl>,  BMI  <dbl>, under-five deaths  <dbl>, Polio <dbl>,
## #   Total expenditure <dbl>, Diphtheria  <dbl>,  HIV/AIDS <dbl>, GDP <dbl>,
## #   Population <dbl>, thinness  1-19 years <dbl>, thinness 5-9 years <dbl>,
## #   Income composition of resources <dbl>, Schooling <dbl>, ...23 <dbl>
```

```
STA218_data <- read_excel("/Users/susmi/Desktop/STA218/STA218_data.xlsx")
head(STA218_data)
```

```
## # A tibble: 6 x 23
##   Country  Year Status `Life expectanc~ `Adult Mortalit~ `infant deaths` Alcohol
##   <chr>   <dbl> <chr>             <dbl>            <dbl>           <dbl> <lgl>
## 1 Centra~  2015 Devel~             52.5              397              15 NA
## 2 Chad     2015 Devel~             53.1              356              46 NA
## 3 Côte d~  2015 Devel~             53.3              397              57 NA
## 4 Camero~  2015 Devel~             57.3              357              45 NA
## 5 South ~  2015 Devel~             57.3              332              26 NA
## 6 Mozamb~  2015 Devel~             57.6              355              60 NA
## # ... with 16 more variables: percentage expenditure <dbl>, Hepatitis B <dbl>,
## #   Measles  <dbl>,  BMI  <dbl>, under-five deaths  <dbl>, Polio <dbl>,
## #   Total expenditure <lgl>, Diphtheria  <dbl>,  HIV/AIDS <dbl>, GDP <dbl>,
## #   Population <dbl>, thinness  1-19 years <dbl>, thinness 5-9 years <dbl>,
## #   Income composition of resources <dbl>, Schooling <dbl>, ...23 <dbl>
```

Answer: For this project, I have decided to use the Life Expectancy Data provided from the World Health Orgination and United Nation. Over the years, it is clear that there has been a huge development in the health sector and as a result life expectancies in many countries have gone to increase. With this data set, we are able to bring to light the improvement of health and social issues within developing countries as well to truly reflect whether improvements have been made. These statistics span from the years of 2000-2015 concentrating on the data of over 193 countries. The data provides information on 22 different variables, varying from social, economical, and morality factors. In order to determine a smaller sample of the enourmous data set, I have sorted the variables out to get a more precise view into what I want to study. To begin with I applied two filters, I wanted to look at developing countries and their statistics in the 2015 year for the most recent information; I found that out of the 193 countries provided, only 153 were developing countries. In order to choose a randomized sample, I put the 153 countries(Before 50) into an excel sheet and randomized their order using the RAND() function. Once this was done, the information was sorted and the first 50 elements(STA218_data) were taken into consideration. You can find the following CSV file at the link provided below, along with more information regarding the data source:

1) https://www.kaggle.com/kumarajarshi/life-expectancy-who

(b) Within the data set, ensure you have three quantitative variables. Choose one of these to be a response variable (Y) and the other two to be explanatory variables X_1 and X_2. Create objects (vectors) which include the observations on these three variables and name them y, x1 and x2, respectively.

```
y <- c(t((STA218_data[4])))
x1 <- c(t((STA218_data[5])))
x2 <- c(t((STA218_data[13])))
```

(c) Since we will be doing regression in Part C, we need to do some checks on the suitability of the data set. We need to check that for the model regressing Y on X_1 (model 1) and the model regressing Y on X_2 (model 2) that r^2 is at least 0.30. Fit the two linear models using the vectors y, x1 and x2, and then find the two respective values of r^2. Comment on their values and the data set's suitability according to the criterion just described.

```
r1.tflc <- cor(x1, y)^2
r2.tflc <- cor(x2, y)^2
```

Answer: The coefficient of determination or R^2 in regards to X_1 (Adult Mortality) and Y (Life expentancy) is 0.634 or %64. This value fits the requriements of R^2 being greater than 0.30 and simply states that approximately 64% of variation in life expectancy is explained by the changes in Adult Mortality. On a model we might see that the values closely surround the linear regression model, thus the relationship can be viewed as strong. At the same time, R^2 in regards to X_2 (Polio) and Y (Life expentancy) is 0.356 or approximately 36%, which lies right above the requirement listed for the R^2. By understanding this value, it is simple to see that 36% percent of the variation in life expectancy can also be explained by the changes in Polio Rates. We can also interpret that the values here are a bit strayed from the linear regression model for X_2 and Y, which might go to say that the linear relationship is bit weak when compared to the relationship between X_1 and Y.

Overall both R^2 values in regards to X_1 and Y as well as X_2 and Y are above the required value, and determine that variation in the dependent variable can be attributed to the independent variable. Therefore given the criteria, the data set is highly suitable.

(d) Secondly, we need to check something called adjusted r^2, which is what we need to calculate for the model regressing y on both x_1 and x_2; call this model 3. Its value should be at least 0.50, and be at least 0.1 larger than the two adjusted r^2 found for models 1 and 2. Fit the model (see sample code below commented out) and find its value. Based on its value, comment on the data set's suitability according to these criteria.

```
model3 <- lm(y~x1+x2)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9094  -2.1774   0.1465   2.3840  10.1403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.904024   3.023579  22.789  < 2e-16 ***
## x1          -0.051500   0.006445  -7.991 2.63e-10 ***
## x2           0.109023   0.027341   3.988 0.000232 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.1 on 47 degrees of freedom
## Multiple R-squared:  0.7268, Adjusted R-squared:  0.7151
## F-statistic: 62.51 on 2 and 47 DF,  p-value: 5.733e-14
```

Answer: The adjusted r^2 for model3 is 0.7151 which fits the requirements that were listed; value should be at least 0.50, and be at least 0.1 larger than the two adjusted r^2. The adjusted r^2 looks at the number of predictors additional and calculates whether they are contributing to the model or not. The value of r^2 only increases when the new predictor improves the model fit more than expected. In this specific case we can see that the value has increased when compared to y~x1 and y~x2, and thus it is clear that the new predictor is adding value to the model. Thus, given the criteria, it is quite clear that the data set is highly suitable.

Part B: Exploratory Data Analysis (15 marks)

(a) Write 4-6 sentences describing the data set, including, in particular, and a description of three quantitative variables, X_1, X_2 and Y. Other exxamples include what the variables measure, their units of measurement (if they have any), other variables that were a part of the data set and how the data was collected.

Answer: My data was taken from a study conducted by World Health Organization, that looked into understanding health factors alongside economic factors that go to affect life expectancy, from a period of 2000-2015. The data was taken from the Global Health Observatory who goes to track the health status of all countries, and this study in particular features information from 193 countries. The variables assessed in the study include immunization, mortality, economic, and social factors dependent on the year, status and country. For the purpose of this project we are taking a look into the data of the top 50 developing countries and their information in the year of 2015. The variables being evaluated are Life expectancy(y) in relation to adult mortality(x1) and polio(x2).

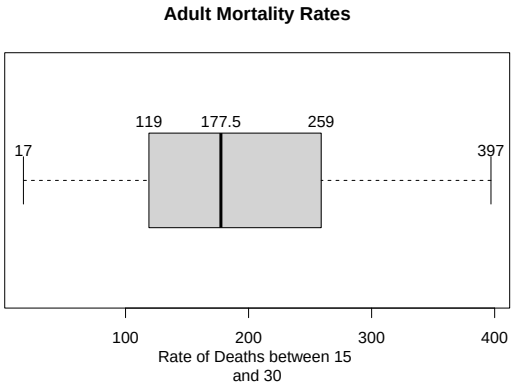
Life expectancy is a hypothetical measure that calculated by understanding the age related death rates of the population of a country. In our study, we can see that the data is represented as the average number of years that a person can expect to live. We also go to take a look at Adult mortality, which is measured as a percentage of the probability of one dying between 15 – 60 years per 1000 people. The last variable that we consider is the percentage of immunization coverage of polio amongst infants within the developing country.

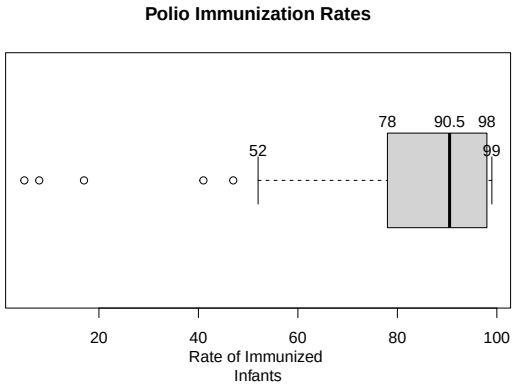
Unarguably, life expectancy is a huge issue amongst third world countries who often don't have the proper amenities to take care of their health. By evaluating the correlation of these variables, I hope to uncover more details about what goes to affect the life expectancy of a country, and how can we a play a role into changing that for future generations.

(b) Make appropriate displays of the distributions of x1 and x2. Make appropriate comments on their distributions based on these plots; for instance, you may comment on shape and spread.

```
boxplot(x1, main = "Adult Mortality Rates", xlab = "Rate of Deaths between 15
and 30", horizontal = TRUE)
text(x = boxplot.stats(x1)$stats[2:4], labels = boxplot.stats(x1)$stats[2:4], y = 1.25)
text(x = boxplot.stats(x1)$stats[5], labels = boxplot.stats(x1)$stats[5], y = 1.125)
text(x = boxplot.stats(x1)$stats[1], labels = boxplot.stats(x1)$stats[1], y = 1.125)

boxplot(x2, main = "Polio Immunization Rates", xlab = "Rate of Immunized
Infants", horizontal = TRUE)
text(x = boxplot.stats(x2)$stats[2:4], labels = boxplot.stats(x2)$stats[2:4], y = 1.25)
text(x = boxplot.stats(x2)$stats[5], labels = boxplot.stats(x2)$stats[5], y = 1.125)
text(x = boxplot.stats(x2)$stats[1], labels = boxplot.stats(x2)$stats[1], y = 1.125)
```





Answer: Boxplot displays are particularly useful for understanding the distribution of quantitative data. They provide us with a visual summary that allows us to compare and identity the data in a very simple and concise manner. Below are the trends that need to be noted after interpreting our boxplots for each of our explanatory variables:

- **X1:** Taking a look at the distribution of the adult morality rates, we can come to a couple of conclusions. Starting at a rate of 17%, ending at 397% and centered at 177.5%, it is clear to see that the data is widely spread about the mean, and promises larger variability. Half of the values fall within a short interval to the left, and majority of the values fall within a larger interval to the right, the data is evidently skewed to the right. There exists a higher frequency of mortality rates that lie around the mean, within the dataset. Even though the data is fairly well spread compared to the distribution of polio immunization rates, it seems that there exist no extreme outliers. Note that most of the data falls under the given range of the dataset, thus, we can go to assume that there were minimal errors when collecting the data.
- **X2:** The distribution of polio immunization rates is far much different when compared to adult morality rates. The range of the data starts at 52%, ends at 99% and is centered at 90.5%. Given this information, is it clear to see that the boxplot is not widely spread about the mean, instead it is sitting closer to higher end of the range, and promises lower variability. The data set is also skewed to the left, which means that there exists a higher frequency of immunization rates that are greater than 52% and less than 90.5% within our data set; the mean of our data is less than the median. There also seem to exist 5 outliers that fall outside the lower range of our values. Note that there might have been some error when collecting the data, and this might go to affect future interpretations and calculations of this dataset.

(c) Numerically summarize y, x1 and x2 by finding their means, medians, IQRs and variances. Make sure to print all their values and label (or comment) appropriately.

```
nsummary_y <- data.frame(MEASURE = c("Mean", "Median", "IQR", "Variance",
"Standard deviation"), VALUE = c(mean(y), median(y), IQR(y), var(y),
sd(y)))

nsummary_x1 <- data.frame(MEASURE = c("Mean", "Median", "IQR", "Variance",
"Standard deviation"), VALUE = c(mean(x1), median(x1), IQR(x1), var(x1), sd(x1)))

nsummary_x2 <- data.frame(MEASURE = c("mean", "median", "IQR", "variance",
"Standard deviation"), VALUE = c(mean(x2), median(x2), IQR(x2), var(x2), sd(x2)))

library(knitr)
kable(nsummary_y, caption = "Summary for Y")
kable(nsummary_x1, caption = "Summary for X1")
kable(nsummary_x2, caption = "Summary for X2")
```

Table 1: Summary for Y

MEASURE	VALUE
Mean	67.910000
Median	69.300000
IQR	14.125000
Variance	59.004592
Standard deviation	7.681445

Table 2: Summary for X1

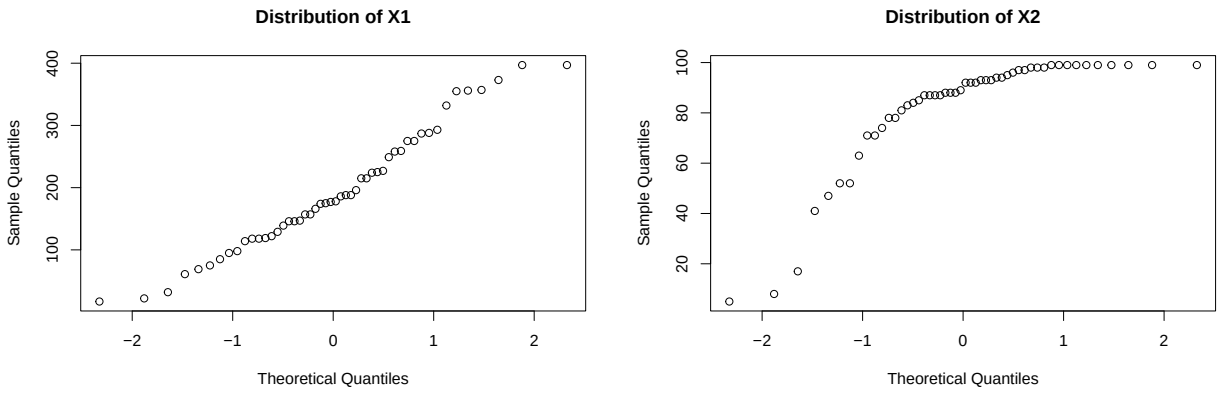
MEASURE	VALUE
Mean	193.02000
Median	177.50000
IQR	139.00000
Variance	9818.38735
Standard deviation	99.08778

Table 3: Summary for X2

MEASURE	VALUE
mean	82.06000
median	90.50000
IQR	19.00000
variance	545.52694
Standard deviation	23.35652

(d) Make quantile-quantile plots of x1 and x2, and make appropriate observations and comparisons about their distributions.

```
qqnorm(x1, main = "Distribution of X1")
qqnorm(x2, main = "Distribution of X2")
```



Answer:

1) Distribution of X1: We can begin by understanding that the points do seem to fall about a linear form. At the same time, if we look a tad bit closer we can see that most of our points lie within the upper 50% of our data set. The data set is skewed to the right, which goes to say that there is a higher frequency of values that lie close to our mean. There are no extreme outliers that go beyond the trend of the plotted data points, however, if we look closely we can see that after the second standard deviation above the mean, there exists a slight jump between the points and from there onwards the ends of our quantile plot taper off. From this trend, it can be interpreted that there exist some extreme values that fall close to the higher boundaries of the range and might go to affect calculations later on. Overall, given the trends of values in X1, we can make the assumption that this data, for the most part, follows a normal distribution.

2) Distribution of X2: The distribution of polio immunization rates is much different than the distribution of adult mortality rates. The data is extremely skewed to the left, with majority of the observations lying near the mean of the data set. We can also go to make the assumption that there is a higher frequency of data that falls under the upper bound of our dataset. There is a heavy peak in the middle of the data which promises that this data set might have a lot of repeated values that lie near the mean, which might also fall quite close to the higher range; values may not be entirely unique. Towards the third deviation below the mean, we can also notice that there exist some extreme outliers, which might go to suggest that there might have been error when collecting the data.

Overall, when comparing the two plots a couple of conclusions can be made:

- Adult mortality rates have a higher level of variability when compared to Polio Immunization Rates
- Polio Immunization Rates have extreme outliers that might go to affect any future assumptions we make about the values and the spread of the data, whereas we can believe that adult mortality rates will be closer in terms of accuracy.
- Both graphs are definitely skewed to either side, so there is minimal spread amongst all the data.
- Both graphs seem to have majority of their values lie around the mean, thus it is important to note that the data set might not be unique, and there might be a higher frequency of a certain set of values. **Important to note: this might go to affect data analysis later on.**

(e) Choose one of your variables X_1 or X_2. Write code to break (divide) the observations on your selected variable (x1 or x2) into 5 categories (intervals); try using the function `which` to do this. Make sure each value (observation) falls in one of your categories. Then, display the categories and counts in a table.

```
library(data.table)
intervals = c(levels(cut(x1, 5)))
count <- c(table(cut(x1, 5)))
freq_data <- data.table(Intervals = intervals, Count = count)

library(knitr)
kable(freq_data, caption = "Frequency Table")
```

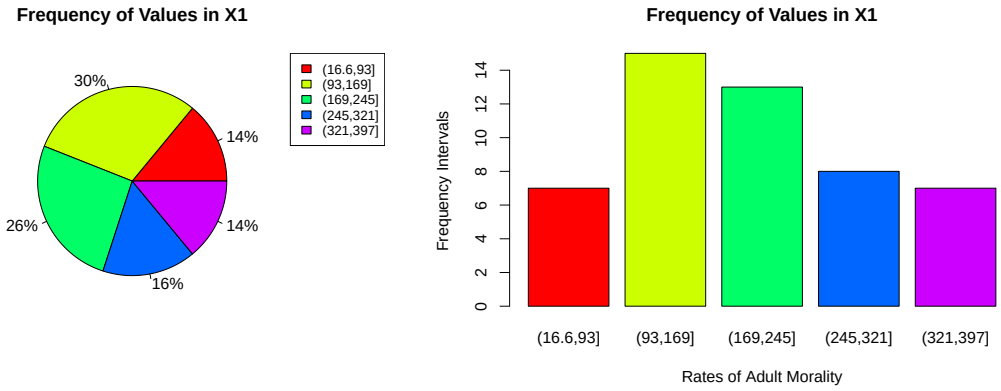
Table 4: Frequency Table

Intervals	Count
(16.6,93]	7
(93,169]	15
(169,245]	13
(245,321]	8
(321,397]	7

(f) Create both a pie chart and a bar plot of the categorical data found in Part B (e), with appropriate labels.

```
par(mfrow=c(1,1))
percent <- round(count/sum(count)*100)
percent <- paste(percent,"%",sep="")
pie(count, labels = percent, main = "Frequency of Values in X1",
col=rainbow(length(count)))
legend("topright", c(intervals), cex = 0.8, fill = rainbow(length(count)))
```

```r
barplot(count, main = "Frequency of Values in X1", xlab = "Rates of Adult Morality",
ylab = "Frequency Intervals", col = rainbow(length(count)), names.arg = intervals)
```



Part C: Univariate Regression (15 marks)

(a) Calculate the correlation between x_1 and x_2 in two ways: (i) Using the formula in the notes, which includes finding means, sums of squares, etc and (ii) using the appropriate in-built function. Interpret the value of the correlation in the context of your data set.

```r
#(i) Use the formula to calculate the correlation
x1bar <- mean(x1)
x2bar <- mean(x2)
SSx1 <- sum((x1-x1bar)^2)
SSx2 <- sum((x2-x2bar)^2)
SSxx <- sum((x1-x1bar)*(x2-x2bar))
r <- SSxx/sqrt(SSx1*SSx2)
sprintf("Correlation: %s", r)
```

```
## [1] "Correlation: -0.398561672667916"
```

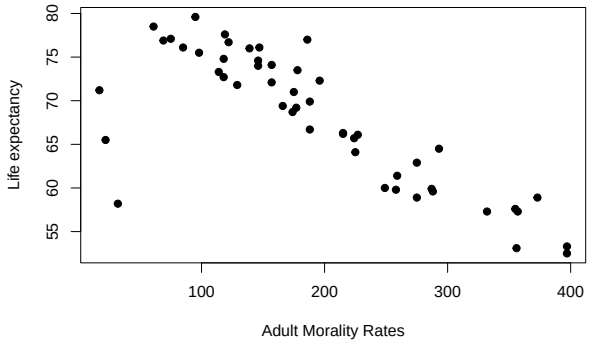
```r
#(ii) The built in function to calculate the correlation
sprintf("Correlation: %s", cor(x1, x2))
```

```
## [1] "Correlation: -0.398561672667916"
```

Answer: The value of correlation tells us about the type of association between the two variables, and gives us no idea about the casual relationship between the data. After calculations, it seems the correlation between x_1 and x_2 is -0.3985617. Since it is a negative value, we can go to make the assumption that x_1 and x_2 negatively associated, which means that they are moving in opposite directions of each other. As x_1 increases in value, x_2 will decrease; similarly, if x_2 decreases in value,x_1 will increase. The value is also noted to be below 0, which means that there is a weak and unreliable relationship between the two variables. Overall, there exists a weak negative linear association between x_1 and x_2.

(b) Choose one of your quantitative variables X_1 or X_2 and make a scatterplot of y and your chosen variable. Based on the scatterplot, comment on their relationship.

```
plot(x1, y, pch=19, xlab = "Adult Morality Rates", ylab = "Life expectancy")
```



Answer: For this question, I chose the quantitative variable x_1 in relation to y. After creating the scatterplot it is clear to see that there is a negative linear association between adult mortality rates and life expectancy. For the most part, the lines do seem to be following a pattern and the points do lay close together. It can be said that the linear association is strong, with the exception of a few outliers that lay past the pattern of current plot, along the y direction. We can go to assume that these outliers have very little effect on the regression line. Overall, there is a strong, negative, linear association between adult mortality rates, and life expectancy.

(c) Using your data, find the least-squares regression line, regressing your chosen response variable Y on your chosen explanatory variable from Part C (b), in two ways: (i) from the formulas in the notes (finding all the necessary sums of squares, etc.) and (ii) using the appropriate in-built function. Interpret the values of the intercept and slope in the context of your data set.

```
#(i) Use the formula to calculate the least-squares regression line
x1bar <- mean(x1)
ybar <- mean(y)
SSxx <- sum((x1-x1bar)^2)
SSyy <- sum((y-ybar)^2)
SSxy <- sum((x1-x1bar)*(y-ybar))
slope <- SSxy/SSxx
intercept <- ybar - (slope*x1bar)
sprintf("Intercept: %s, Slope: %s", intercept, slope)
```

```
## [1] "Intercept: 79.8274702454358, Slope: -0.061742152343984"
```

```
#(ii) The built in function to calculate the least-squares regression line
model1 <- lm(y~x1)
sprintf("Intercept: %s, Slope: %s", model1[[1]][1], model1[[1]][2])
```

```
## [1] "Intercept: 79.8274702454358, Slope: -0.0617421523439841"
```

Answer: The slope goes to explain the amount that y changes for every unit increase in x_1. In the context of this data set, we can interpret this as for every unit increase in adult mortality rates, the value of life expectancy decreases by approximately 0.062 units. As for the intercept, if in the case the adult morality rate within a country was 0%, it means that the predicted life expectancy for the average developing country in 2015 would be around 80 years.

(d) Find the value of the coeffecient of determination for the model in Part C (c). Interpret its value in the context of your data set.

```
r.tflc <- cor(x1, y)^2
sprintf("Co-effecient of determination: %s", r.tflc)
```

```
## [1] "Co-effecient of determination: 0.63433384087288"
```

Answer: After calculation of the values, the co-efficient of determination in regards to x_1 and y is 0.6343338. The interpretation of this value, in regards to the data set, means that 63% of the variation in life expectancy is explained by the changes in Adult Mortality. Since the co-efficient is quite high, we can also go to believe that the 64% of the data fits the regression model of life expectancy and adult mortality.

(e) Find all the residuals for the fit in Part C (c) in two ways: (i) coding the calculation and (ii) using an appropriate in-built function.

```
#(i) Coding the calculation
fit <- lm(y ~ x1)
b <- coef(fit)
resi <- y - (b[1] + b[2] * x1)
sprintf("Residuals: %s", resi)
```

```
##  [1] "Residuals: -2.81583576487409"   "Residuals: -4.74726401097743"
##  [3] "Residuals: -2.01583576487409"   "Residuals: -0.485521858633447"
##  [5] "Residuals: -2.02907566723305"   "Residuals: -0.309006163321421"
##  [7] "Residuals: -19.6517213704283"   "Residuals: 2.1023525788703"
##  [9] "Residuals: -3.94837835084014"   "Residuals: -2.44573037036835"
## [11] "Residuals: -4.09799494068788"   "Residuals: -2.20747252271234"
## [13] "Residuals: -4.45367431178373"   "Residuals: -2.43625278834389"
## [15] "Residuals: 0.0516216491598556"  "Residuals: -1.83548596803935"
## [17] "Residuals: 2.76298039135158"    "Residuals: -12.9691428938681"
## [19] "Residuals: -0.297228120383323"  "Residuals: 0.287998336648613"
## [21] "Residuals: -0.352907491479186"  "Residuals: -0.252907491479192"
## [23] "Residuals: -1.51994560476676"   "Residuals: -0.384335737582532"
## [25] "Residuals: 0.300890719449427"   "Residuals: -0.178272956334396"
## [27] "Residuals: 1.68005439523324"    "Residuals: 1.97740641476145"
## [29] "Residuals: -7.57785365558803"   "Residuals: -0.0627325930618241"
## [31] "Residuals: 1.96604767256973"    "Residuals: 4.57399161398511"
## [33] "Residuals: 0.158103731154355"   "Residuals: 0.511135121778423"
## [35] "Residuals: 4.66263287179341"    "Residuals: 3.18688399678591"
## [37] "Residuals: 3.96604767256973"    "Residuals: 3.7868839967859"
## [39] "Residuals: 2.25810373115435"    "Residuals: 1.72326068427468"
## [41] "Residuals: 4.75468893037802"    "Residuals: 1.52061270380288"
## [43] "Residuals: 5.34862614912988"    "Residuals: 4.4050723405303"
## [45] "Residuals: 1.33273826629915"    "Residuals: 8.65657009054527"
## [47] "Residuals: 1.90319118036304"    "Residuals: 5.11984588349834"
## [49] "Residuals: 2.43880104754727"    "Residuals: 5.63803422724271"
```

```r
#(ii) Using an appropriate built in function
fit <- lm(y ~ x1)
resi <- unname(residuals(fit))
sprintf("Residuals: %s", resi)
```

```
##  [1] "Residuals: -2.81583576487428"   "Residuals: -4.74726401097747"
##  [3] "Residuals: -2.01583576487413"   "Residuals: -0.485521858633491"
##  [5] "Residuals: -2.02907566723309"   "Residuals: -0.309006163321454"
##  [7] "Residuals: -19.6517213704283"   "Residuals: 2.10235257887026"
##  [9] "Residuals: -3.94837835084018"   "Residuals: -2.44573037036839"
## [11] "Residuals: -4.09799494068791"   "Residuals: -2.20747252271237"
## [13] "Residuals: -4.45367431178377"   "Residuals: -2.43625278834393"
## [15] "Residuals: 0.0516216491598169"  "Residuals: -1.83548596803939"
## [17] "Residuals: 2.76298039135153"    "Residuals: -12.9691428938682"
## [19] "Residuals: -0.297228120383367"  "Residuals: 0.287998336648577"
## [21] "Residuals: -0.352907491479223"  "Residuals: -0.252907491479229"
## [23] "Residuals: -1.51994560476679"   "Residuals: -0.384335737582571"
## [25] "Residuals: 0.300890719449382"   "Residuals: -0.17827295633444"
## [27] "Residuals: 1.68005439523321"    "Residuals: 1.97740641476141"
## [29] "Residuals: -7.57785365558807"   "Residuals: -0.0627325930618591"
## [31] "Residuals: 1.96604767256969"    "Residuals: 4.57399161398507"
## [33] "Residuals: 0.158103731154321"   "Residuals: 0.511135121778379"
## [35] "Residuals: 4.66263287179336"    "Residuals: 3.18688399678587"
## [37] "Residuals: 3.96604767256969"    "Residuals: 3.78688399678587"
## [39] "Residuals: 2.25810373115432"    "Residuals: 1.72326068427464"
## [41] "Residuals: 4.75468893037798"    "Residuals: 1.52061270380284"
## [43] "Residuals: 5.34862614912985"    "Residuals: 4.40507234053026"
## [45] "Residuals: 1.3327382662991"     "Residuals: 8.65657009054524"
## [47] "Residuals: 1.903191180363"      "Residuals: 5.1198458834983"
## [49] "Residuals: 2.43880104754723"    "Residuals: 5.63803422724268"
```

(f) Choose a value of your explanatory variable, call it $x0$, that is outside the range of the original values. Calculate the predicted value of your response variable corresponding to $x0$. Make a comment on its value and whether it is trustworthy.

```r
range(x1)
```

```
## [1]  17 397
```

```r
x0 <- 450
y0 <- intercept + (slope*x0)
sprintf("X0 Value: %s, Y[X0]: %s", x0, y0)
```

```
## [1] "X0 Value: 450, Y[X0]: 52.043501690643"
```

Answer: My chosen value outside of the range of the data was $450(x0)$, after plugging in the value within the least-squares regression line, we get the value of 52.0435. This value is quite accurate given the data that we have on hand. At a value of 397 we have an output of 52.5, now comparing those values with the output of when $x = 450$, we can see that the outputs are quite similar and close in range. As well, given the we determined that there is a strong relationship between the data points, and our co-efficient of determination is 63%, we can definitely go to say that this value is quite trustworthy. It matches the trend of the data, and doesn't fall too far from the current values given to us.

Part D: Inference (10 marks)

(a) Construct two 95% confidence intervals using the t-distribution for the two true population means μ_1 and μ_2, which denote the mean of X_1 and X_2 respectively.

```
#Confidence Interval for X1
X1_CI <- c(t.test(x1, mu = mean(x1))[["conf.int"]][1],
t.test(x1, mu =  mean(x1))[["conf.int"]][2])

#Confidence Interval for X2
X2_CI <- c(t.test(x1, mu = mean(x2))[["conf.int"]][1],
t.test(x2, mu = mean(x2))[["conf.int"]][2])


print(paste0("Confidence Interval for X1:", sprintf("[%4.3f, %4.3f]",
(X1_CI)[1], (X1_CI)[2])))
```

```
## [1] "Confidence Interval for X1:[164.860, 221.180]"
```

```
print(paste0("Confidence Interval for X2:", sprintf("[%4.3f, %4.3f]",
(X2_CI)[1], (X2_CI)[2])))
```

```
## [1] "Confidence Interval for X2:[164.860, 88.698]"
```

(b) Consider the hypotheses: $H_0 : \mu_1 = a$ versus $H_1 : \mu_1 \neq a$, where you can choose an appropriate value of a (say something near the sample mean of x1). Without using `t.test`, at 1% level of significance, carry out a formal hypothesis test using the p-approach showing all your calculations and steps, including a formal conclusion.

```
# Mean of X1
mean(x1)
```

```
## [1] 193.02
```

```
# Standard Deviation of X1
sd(x1)
```

```
## [1] 99.08778
```

```
# Test Statistic
z <- (mean(x1) - 200) / (sd(x1)/sqrt(50))
z
```

```
## [1] -0.4981044
```

STEP 1: STATE THE LEVEL OF SIGNIFICANCE.
We have $a = 0.01$

STEP 2: STATE THE HYPOTHESES.

$$H_0 : \mu_1 = 200 \text{ versus } H_1 : \mu_1 \neq 200$$

STEP 3: CALCULATE THE TEST STATISTIC.

$$\begin{aligned} z &= \frac{\overline{x} - 200}{\frac{\sigma}{\sqrt{50}}} \\ &= \frac{193.02 - 200}{\frac{99.08778}{\sqrt{50}}} \\ &= -0.4981044 \end{aligned}$$

STEP 4: FIND THE P-VALUE.
Since our alternative is two sided, we are going to have to calculate the P value, by taking the P value from the absolute value of z, subtract that from 1 and multiply it by 2.

$$\begin{aligned} P(Z \geq |-0.4981044|) &= 2P(Z \geq 0.4981044) \\ &= 2[1 - P(Z \leq 0.4981044)] \\ &= 2[1 - 0.6907948] \\ &= 2 * 0.3092052 \\ &= 0.6184104 \end{aligned}$$

STEP 5: STATE A CONCLUSION.
At a significance $a = 0.01$, since p-value = $0.6184104 \geq a = 0.01$, we fail to reject H_0. That is, at a 1% level of significance, we have insufficient statistical evidence that the true mean rate of adult mortality within developing countries differs from 200%

(c) Consider the same hypotheses in (b). Without using `t.test`, carry out a formal hypothesis test using the rejection region approach, showing all of your calculations.

STEP 1: STATE THE LEVEL OF SIGNIFICANCE.
We have $a = 0.01$

STEP 2: STATE THE HYPOTHESES.

$$H_0 : \mu_1 = 200 \text{ versus } H_1 : \mu_1 \neq 200$$

STEP 3: DETERMINE THE REJECTION REGION.
For $a = 0.01$, we can determine the rejection region as $z \leq$ -2.575 or $z \geq 2.575$

STEP 4: CALCULATE THE VALUE OF THE TEST STATISTIC.

$$\begin{aligned} z &= \frac{\overline{x} - 200}{\frac{\sigma}{\sqrt{50}}} \\ &= \frac{193.02 - 200}{\frac{99.08778}{\sqrt{50}}} \\ &= -0.4981044 \end{aligned}$$

STEP 5: STATE A CONCLUSION.
At a significance $a = 0.01$, our value of -0.4981044 > -2.575, and -0.4981044 > -2.575, thus, we fail to reject H_0 since our value does not fall under the rejection region. That is, at a 1% level of significance, we have insufficient statistical evidence that the true mean rate of adult mortality within developing countries differs from 200%

(d) Consider the same hypotheses in (b). Carry out a hypothesis test at 5% level of significance but now using the in-built function `t.test`. Make an appropriate conclusion based on the output.

```
t.test(x1, alternative = "two.sided", mu = 200, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x1
## t = -0.4981, df = 49, p-value = 0.6206
## alternative hypothesis: true mean is not equal to 200
## 95 percent confidence interval:
##  164.8596 221.1804
## sample estimates:
## mean of x
##    193.02
```

Conclusion: Given that the p-value 0.6206 is greater than our significance level of 5% we can conclude that we fail to reject H_0. In other words, when $a = 0.05$, we have insufficient statistical evidence that the true mean of adult mortality rate in the average developing country differs from 200%.

Part E: Multivariate Regression (5 marks)

(a) Continuing with your response variable Y, now fit the least-squares regression line regressing Y on both X_1 and X_2 using your vectors of observations y, x1 and x2.

```
model2 <- lm(y~x1+x2)
coefficients(model2)
```

```
## (Intercept)          x1          x2
## 68.90402448 -0.05149972  0.10902331
```

(b) Find the appropriate value of adjusted r^2 for this new model and compare it to that found in Part C (d). What do these values tell you about a preferred model?

```
ls.line <- lm(y~x1)

#Adjusted R in Part C (d)
summary(ls.line)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6517  -1.9707   0.2231   2.3936   8.6566
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 79.827470   1.464983  54.490  < 2e-16 ***
## x1          -0.061742   0.006766  -9.125 4.65e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.693 on 48 degrees of freedom
## Multiple R-squared:  0.6343, Adjusted R-squared:  0.6267
## F-statistic: 83.27 on 1 and 48 DF,  p-value: 4.646e-12
```

```
#Adjusted R in Part E (b)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9094  -2.1774   0.1465   2.3840  10.1403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.904024   3.023579  22.789  < 2e-16 ***
## x1          -0.051500   0.006445  -7.991 2.63e-10 ***
## x2           0.109023   0.027341   3.988 0.000232 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.1 on 47 degrees of freedom
## Multiple R-squared:  0.7268, Adjusted R-squared:  0.7151
## F-statistic: 62.51 on 2 and 47 DF,  p-value: 5.733e-14
```

Answer: The adjusted r^2 value was 0.6267 for the model in Part C (c), when calculating the adjusted r^2 value for the current model we got 0.7151. A higher adjusted r^2 means that the additional input variables added onto the model, and providing some value. In other words, when including the polio immunization rates into the model, we can see that there is a higher amount of variability being accounted for in our data. That said, since the addition of polio immunization rates only increased our adjusted value of r^2 , the new term improves the model and provides more accuracy. Thus, these values tell us that our preferred model of use would be the current model with y regressing on both x_1 and x_2, as there is a higher and accurate explanation of variability within our data.

(c) Find a 95% confidence interval for each of the regression coefficients for X_1 and X_2 in the model in Part E (a). Make conclusions about the significance of the variables in the model.

```
summary(model1)
t <- qt(0.975, 47)
x1_interval <- c(-0.051500 + t(0.006445), -0.051500 - t(0.006445))

x2_interval <- c(0.109023 + t(0.027341), 0.109023 - t(0.027341))

x1_p <-  2*2.63e-10

x2_p <- 2*0.000232
```

```r
# 95% Confidence Interval for X1
x1_interval

# 95% Confidence Interval for X2
x2_interval

#P-Value for X1
x1_p

#P-Value for X2
x2_p
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.6517  -1.9707   0.2231   2.3936   8.6566
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.827470   1.464983  54.490  < 2e-16 ***
## x1          -0.061742   0.006766  -9.125 4.65e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.693 on 48 degrees of freedom
## Multiple R-squared:  0.6343, Adjusted R-squared:  0.6267
## F-statistic: 83.27 on 1 and 48 DF,  p-value: 4.646e-12
##
## [1] -0.045055 -0.057945
## [1] 0.136364 0.081682
## [1] 5.26e-10
## [1] 0.000464
```

Answer: Given the calculations, and the summary of our model we can see that we have the respective p-values; X1: 5.26e-10 and X2: 0.000464. Based on these, we can come to a conclusion that x_1 nor x_2 are not significant at a 5% level of significance. If the coefficient of x_1 (Adult Mortality Rates) was infact 0, the probability of observing a value of b1, at least as extreme as -0.051500 would be 5.26e-10; extremely low probability. At the same time if the coefficient of x_2 (polio immunization rates) was in fact 0, the probability of observing a value of b2, at least as extreme as 0.109023 would be 0.000464; yet again an extremely low probability. It can be concluded that at a 5% level of significance, both the rate of Adult morality and Polio Immunization are not significant. In other words, changes within the rate of Life Expectancy is not associated with changes in the rates of Adult Morality alongside Polio Immunizations.