University of Essex
Department of Mathematical Sciences

MA317 : GROUP COURSEWORK

Data Analysis of Life Expectancy

Anmol Mahajan (2200735)

Rojitha Repalle (2201010)

Sharon Machado (2202055)

Supervisor: **Stella Hadjiantoni**

December 14, 2022
Colchester

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Life expectancy represents the average lifespan of humans. It is crucial to consider this aspect for each nation as it allows for comparison of factors that one nation lacks in comparison to another to maintain a high life expectancy rate. In this coursework, the influence of socioeconomic growth on life expectancy is examined based on a variety of criteria such as mortality rate, total population, health expenditure, income, etc. The data is obtained from a primary world database that includes the World Development Indicators(WDI). The correlation is calculated between the response and predictor variables. Missing values are imputed using mice and median imputation. Using life expectancy as response variable and other features as predictor variables, a full reduced and best multiple regression model is built on the life expectancy dataset allowing us to predict the life expectancy of a child at birth.

CONTRIBUTIONS OF TEAM MEMBERS :

Question 1 - Sharon Machado
Question 2 - Anmol Mahajan
Question 3 - Rojitha Repalle
Question 4 - Anmol Mahajan

Report - Anmol Mahajan, Rojitha Repalle, Sharon Machado

Presentation:
Descriptive Statistics - Sharon Machado
Investigation of Collinearity Between Predictor Variables - Anmol Mahajan
Linear Regression Model for Life Expectancy in 2020 - Rojitha Repalle
Experimental Design for Analysing Average Life Expectancy Across Continents - Anmol Mahajan

CHAPTER 2

Descriptive Statistics

Descriptive statistics[1] is a field of statistics that deals with collecting, organising, summarising, analysing, and interpreting data thus helping analysts to understand the data better. Descriptive statistics play a vital role in any statistical investigation. It gives a clear overview of the data making it possible to assess the data quality and can be a great place to start conducting more in-depth research when presented effectively.

2.1 Data Summary and Plots

The life expectancy dataset used in this investigation contains information on 217 countries(observations) across different continents around the world having 26 World Development Indicators. The quantitative results obtained by performing descriptive analysis on the dataset are discussed in this section.

The original dataset consists of 29 columns. The description of each column is mentioned below.

Country Name

Country Code

Continent

SP.DYN.LE00.IN - Life expectancy at birth, total (years)

EG.ELC.ACCS.ZS - Access to electricity (% of population)

NY.ADJ.NNTY.KD.ZG -Adjusted net national income (annual % growth)

NY.ADJ.NNTY.PC.KD.ZG - Adjusted net national income per capita (annual % growth)

SH.HIV.INCD.14 - Children (ages 0-14) newly infected with HIV

SE.PRM.UNER - Children out of school, primary

SE.PRM.CUAT.ZS - Educational attainment, at least completed primary, population more than 25 years, total (%) (cumulative)

SE.TER.CUAT.BA.ZS - Educational attainment, at least Bachelors or equivalent, population more than 25 years, total (%) (cumulative)

SP.DYN.IMRT.IN - Mortality rate, infant (per 1,000 live births)

SE.PRM.CMPT.ZS - Primary completion rate, total (% of relevant age group)

SE.ADT.LITR.ZS - Literacy rate, adult total (% of people ages 15 and above)

FR.INR.RINR - Real interest rate (%)

SP.POP.GROW - Population growth (annual %)

EN.POP.DNST - Population density (people per sq. km of land area)

SP.POP.TOTL - Population, total

SH.XPD.CHEX.PC.CD - Current health expenditure per capita, PPP (current international $)

SH.XPD.CHEX.GD.ZS - Current health expenditure (% of GDP)

SL.UEM.TOTL.NE.ZS - Unemployment, total (% of total labor force) (national estimate)

NY.GDP.MKTP.KD.ZG - GDP growth (annual %)

NY.GDP.PCAP.CD - GDP per capita, PPP (current international $)

SP.DYN.CBRT.IN - Birth rate, crude (per 1,000 people)

EG.FEC.RNEW.ZS - Renewable energy consumption (% of total final energy consumption

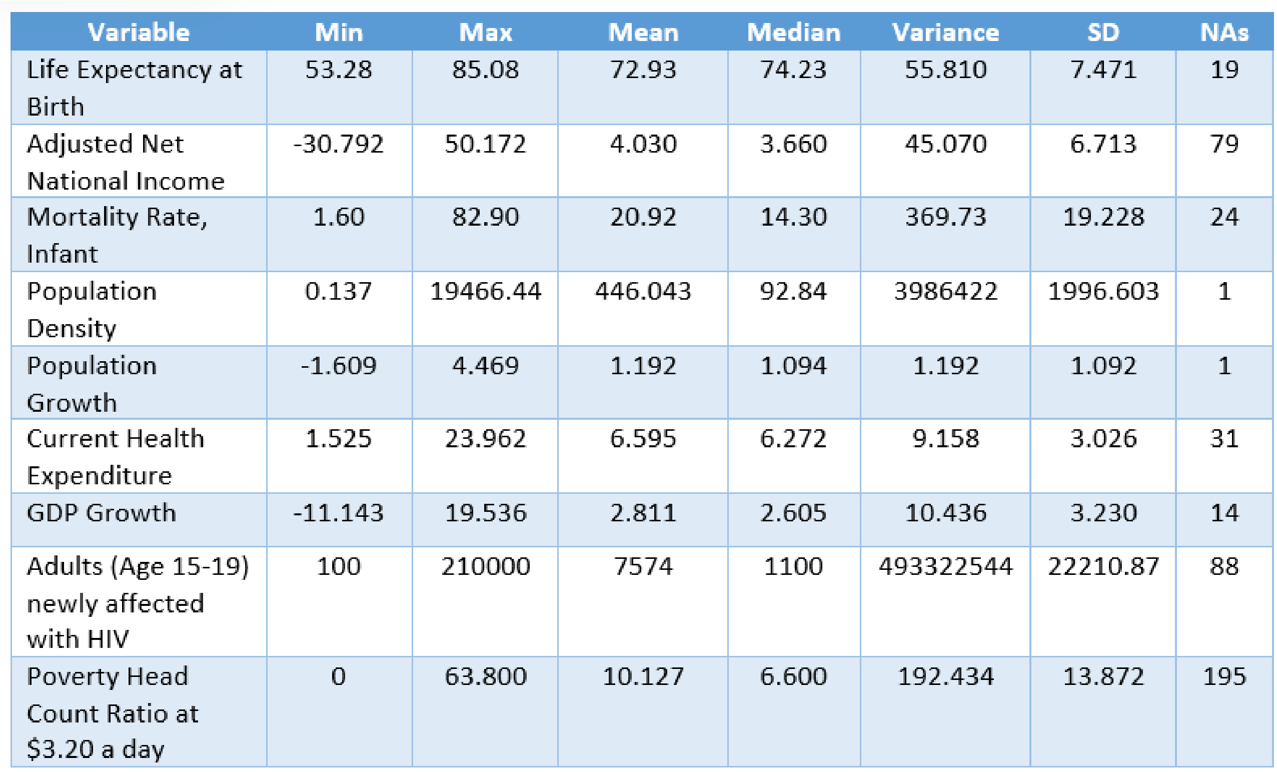
SH.HIV.INCD Adults (ages 15-49) newly infected with HIV

SH.H2O.SMDW.ZS - People using safely managed drinking water services(% of population)

SI.POV.LMIC - Poverty headcount ratio at $3.20 a day (2011 PPP) (% of population)

SE.COM.DURS - Compulsory education, duration (years)

The Life Expectancy at birth is the response variable that is to be predicted using other features present in the dataset. The features - Country name, Country code and Continent are the categorical variables and the other features are continuous in nature. As shown in Figure 2.1, a descriptive analysis of the response variable and a few of the predictor variables are displayed.



Variable	Min	Max	Mean	Median	Variance	SD	NAs
Life Expectancy at Birth	53.28	85.08	72.93	74.23	55.810	7.471	19
Adjusted Net National Income	-30.792	50.172	4.030	3.660	45.070	6.713	79
Mortality Rate, Infant	1.60	82.90	20.92	14.30	369.73	19.228	24
Population Density	0.137	19466.44	446.043	92.84	3986422	1996.603	1
Population Growth	-1.609	4.469	1.192	1.094	1.192	1.092	1
Current Health Expenditure	1.525	23.962	6.595	6.272	9.158	3.026	31
GDP Growth	-11.143	19.536	2.811	2.605	10.436	3.230	14
Adults (Age 15-19) newly affected with HIV	100	210000	7574	1100	493322544	22210.87	88
Poverty Head Count Ratio at $3.20 a day	0	63.800	10.127	6.600	192.434	13.872	195

Figure 2.1: Descriptive Analysis of Predictor and Response Variables

Univariate Analysis of Response Variable

As shown below in Figure 2.2, it represents the plot of density function and histogram of response variable to understand the nature of the distribution. The red dotted line and the green solid line corresponds to the mean and median of the target variable respectively. The mean of the data is found to be 72.9269 and the median is 74.23134. Skewness and Kurtosis[2] are two ways to measure the shape of the distribution. The skewness and kurtosis for the dataset is found to be -0.5874157 and 2.608043 respectively. It is evident that the data is negatively skewed and the distribution is platykurtic, as the coefficient of kurtosis for the response variable is 2.608, less than three.

2.2 Dealing with Missing Values

The major problem analysts confront is handling the missing values in the dataset. Choosing the appropriate method to handle these missing data is the key to reliable data models.

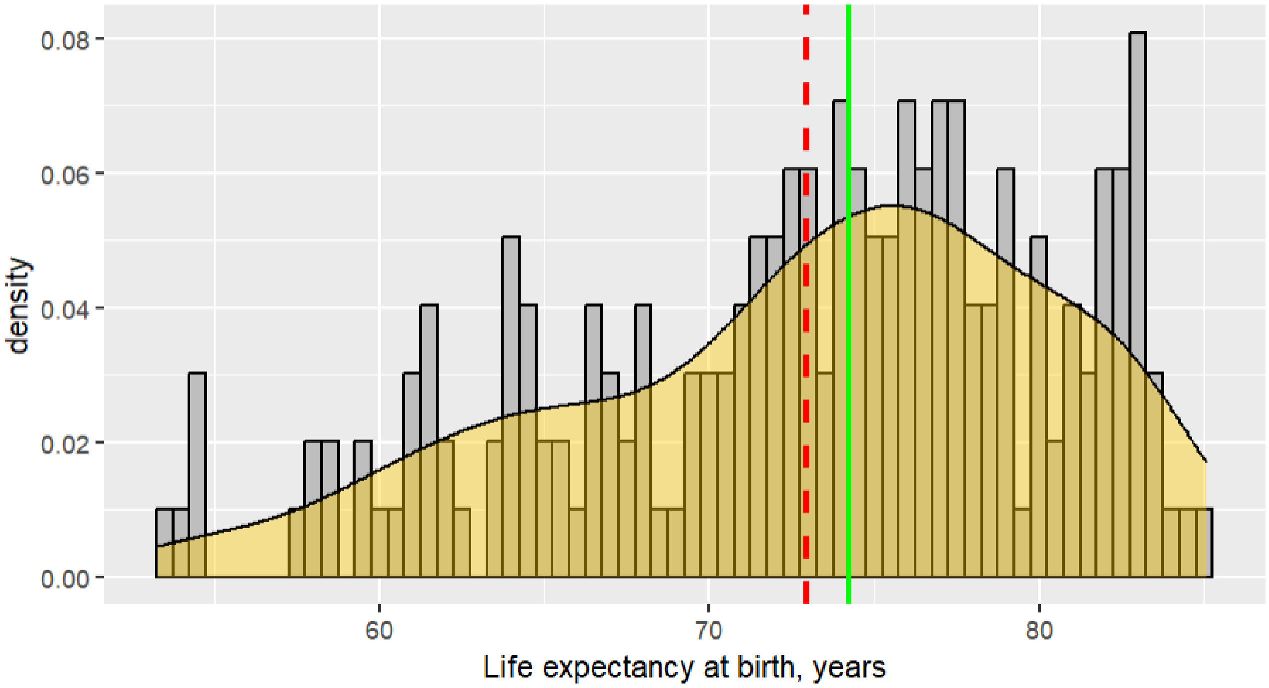


Figure 2.2: Density Function and Histogram of Response Variable

The performance of the predictive model can be severely impacted if the missing values are not appropriately handled. The author[3] categorises the missing data as Missing Data Completely At Random (MCAR), Missing Data At Random (MAR) and Not Missing At Random (NMAR). It is observed that the data follows MCAR. The count of missing values in life expectancy dataset is shown in Figure 2.3.

On analysing the data, it is seen that the response variable have null values in Europe, Australia/Oceania and North America. The predictor variables having null values greater than 80%(EG.FEC.RNEW.ZS, SI.POV.LMIC, SE.ADT.LITR.ZS, SE.TER.CUAT.BA.ZS, SE.PRM.CUAT.ZS) have been eliminated as they cannot contribute to the model. The variable SH.HIV.INCD.14 is dropped as both median and mice imputation cannot give good results. The reason for not using median imputation is that the continents which have more than 80% of missing values for the variable SH.HIV.INCD.14. The mice imputation cannot be applied to SH.HIV.INCD.14 as more than 10% of the data is missing. The next task is to deal with the null values in the rest of the columns and impute them with valid data. The Shapiro's Test is performed on the data to check for normality in predictor variables. From the test and Figure 2.2, it is concluded that the sample of the data does not come from a normal distribution as the p-value is less than 0.05 for many columns. Therefore, the missing values in the predictor variables are dealt by replacing the null values using median imputation, grouping them by continents. As represented in the Table 2.1, the median for the predictor variables for each continent is displayed. The imputation of null values in the response variable is executed using a package called 'mice' in R which provides various methods to deal with missing data. CART is a predictive algorithm which imputes univariate missing data using classification and regression trees. The response variable is imputed using linear regression.

Features	**Africa**	**Asia**	**Australia/Oceania**	**Europe**	**N. America**	**S.America**
EG.ELC.ACCS.ZS	49.35773	100.00000	100.00000	100.00000	100.00000	99.85000
NY.ADJ.NNTY.KD.ZG	5.667937	4.824848	3.373051	3.105359	2.369632	1.364622
NY.ADJ.NNTY.PC.KD.ZG	2.996688	2.906812	1.916541	3.081677	1.096428	1.000331
SH.HIV.INCD.14	920	100	NA	100	100	100
SE.PRM.UNER	121159.0	7923.0	1727.0	3255.5	6652.0	23099.0
SP.DYN.IMRT.IN	39.0	13.2	20.4	3.3	12.9	12.5
SE.PRM.CMPT.ZS	78.94251	99.68049	89.11872	99.01780	93.30409	98.54724
FR.INR.RINR	8.507844	6.230911	5.503926	2.256986	6.427500	10.095083
SP.POP.GROW	2.4840432	1.3557531	0.9157436	0.1989330	0.6310197	1.0881650
EN.POP.DNST	60.11211	125.49794	79.34233	104.16755	221.50427	22.51492
SP.POP.TOTL	12771246.0	13294119.5	167295.0	5401021.5	338253.5	18162846.0
SH.XPD.CHEX.PC.CD	58.49121	293.64357	343.55702	2221.43921	534.27441	490.91093
SH.XPD.CHEX.GD.ZS	5.053260	4.485628	9.761446	7.837886	6.097483	7.763841
SL.UEM.TOTL.NE.ZS	11.905	4.830	6.430	5.360	5.940	8.085
NY.GDP.MKTP.KD.ZG	3.697934	3.976579	1.873541	2.544848	1.720786	1.098468
NY.GDP.PCAP.CD	1359.141	4751.869	5539.451	27749.135	14234.041	6853.693
SP.DYN.CBRT.IN	32.697	17.387	21.833	9.800	14.087	17.661
SH.HIV.INCD	3500	1000	600	500	1400	2000
SH.H2O.SMDW.ZS	23.99069	81.45548	90.90836	97.71257	96.72086	66.60986
SE.COM.DURS	9.0	9.0	9.5	10.0	12.0	14.0

Table 2.1: Median Values Per Continent

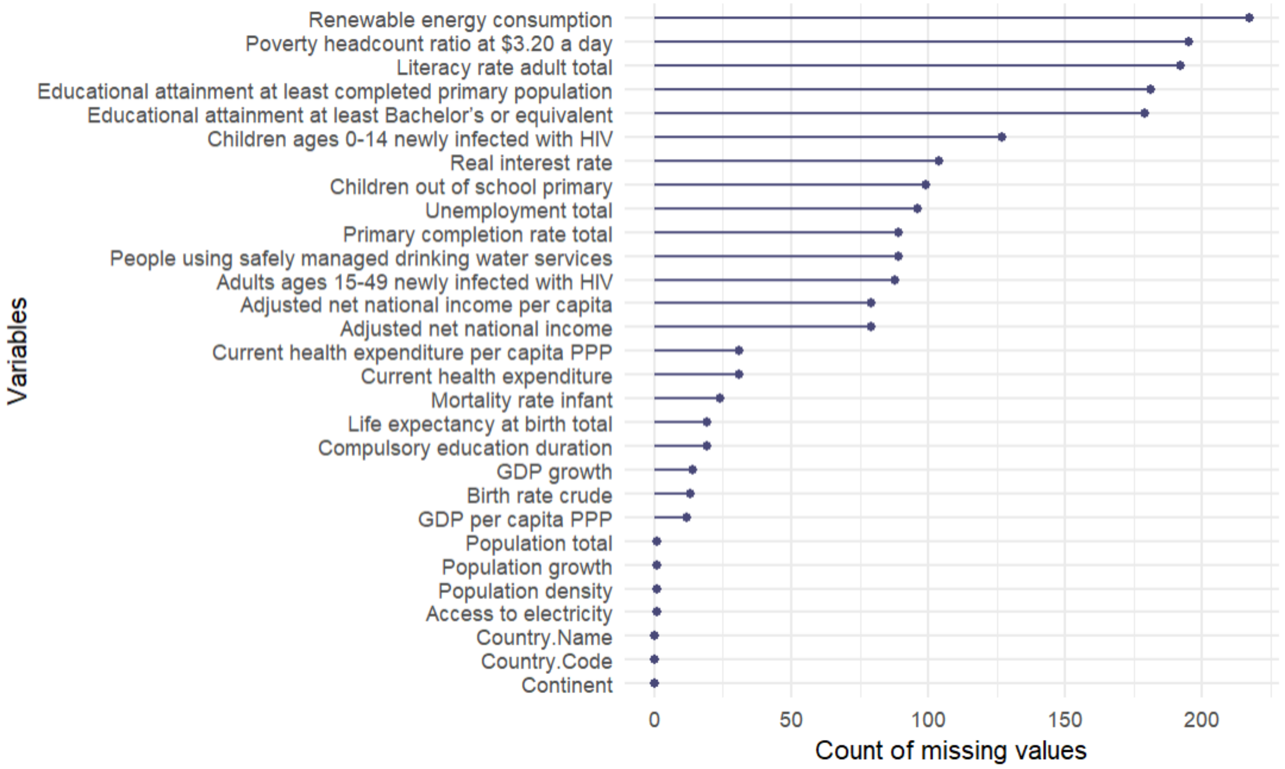


Figure 2.3: Count of Missing Values

CHAPTER 3

Investigation of Collinearity Between Predictor Variables

Multicollinearity in regression analysis occurs when one or more independent predictor variables are highly correlated with each other. It is problematic as it increases the variance of the regression coefficients, making the model unstable. One approach to identify multicollinearity is examining the correlation between each pair of predictor variables. There are two ways for detecting multicollinearity[4].

1. **Correlation Matrix / Correlation Plot :** It is used to find the bivariate relationship between two independent variables.

2. **Variance Inflation Factor (VIF) :** It is used to identify the correlation between an independent variable and the rest of the independent variables. VIF = 1 indicates no correlation between the variables, VIF = 1 to 5 signifies moderate correlation. VIF > 5 indicates that the variables are highly correlated and have potentially severe connection and can reduce the adequacy of the model.

3.1 Analysis of Data using Correlation Matrix

The correlation plot in the Figure 3.1 describes 11 features that have mild or no correlation (less than 0.4) with the response variable. Hence, it is excluded from the dataset. The predictor variables are further analysed for VIF to look for multicollinearity as VIF is helpful in

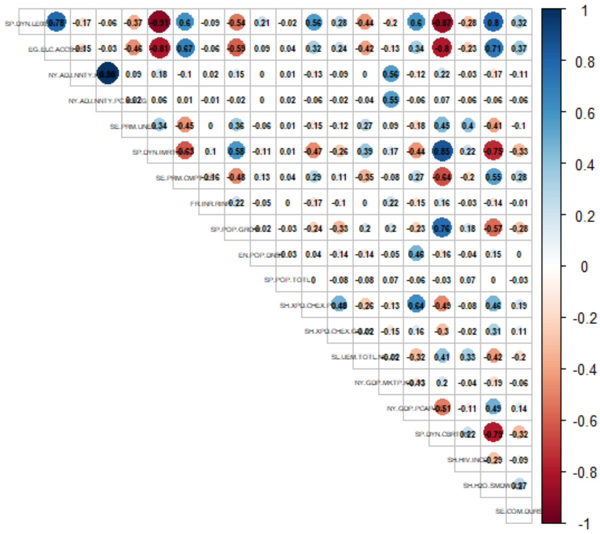


Figure 3.1: Correlation Matrix Plot

identifying multicollinearity between a variable and linear combination of several variables rather than looking at pairwise correlations.

3.2 Analysis of Data using VIF

To investigate collinearity for each of the predictor variables in the model, VIF function is performed. Based on the VIF values obtained, the variable Birth rate(SP.DYN.CBRT.IN) possess VIF greater than five. Therefore, it is dropped and VIF is recomputed. The VIF values for remaining predictor variables is less than five. Hence the model is created based on the final eight variables which are EG.ELC.ACCS.ZS, SE.PRM.CMPT.ZS, SH.XPD.CHEX.PC.CD, NY.GDP.PCAP.CD, SH.H20.SMDW.ZS, SP.DYN.IMRT.IN, SP.POP.GROW and SL.UEM.TOTL.NE.ZS.

CHAPTER 4

Linear Regression Model for Life Expectancy in 2020

Simple linear regression enables to analyse and study of the connection between two continuous variables. The dataset in consideration has several multiple independent variables and one dependent variable. The Multiple Linear Regression Model is used to predict life expectancy in various countries across different continents using the predictor variables. Based on missing value imputation and VIF calculation, there are nine variables for creating the multiple regression model which includes eight independent variables and one dependent variable. The Figure 4.1 represents the simple linear regression models for each of the variables against the response variable.

The summary statistics is shown in Figure 4.2 after fitting the multiple linear regression model with selected eight variables and Life expectancy at birth as the response variable. The larger value of the t-statistic(31.242) indicates that the standard error is small and the coefficient is not equal to zero. The predictor variables are said to be statistically significant if the p-value is less than 0.05. On interpreting the results of the summary statistics, it is observed that the p-value is greater than 0.05 for four predictor variables which also can be seen in Figure 4.2. The Backward Elimination Technique is employed to select the best predictors for our model. Initially the model is implemented using all the selected variables. Subsequently in next iterations, the variable with higher p-value is discarded and models are built until the predictors are statistically significant. Here, three iterations are carried out until the

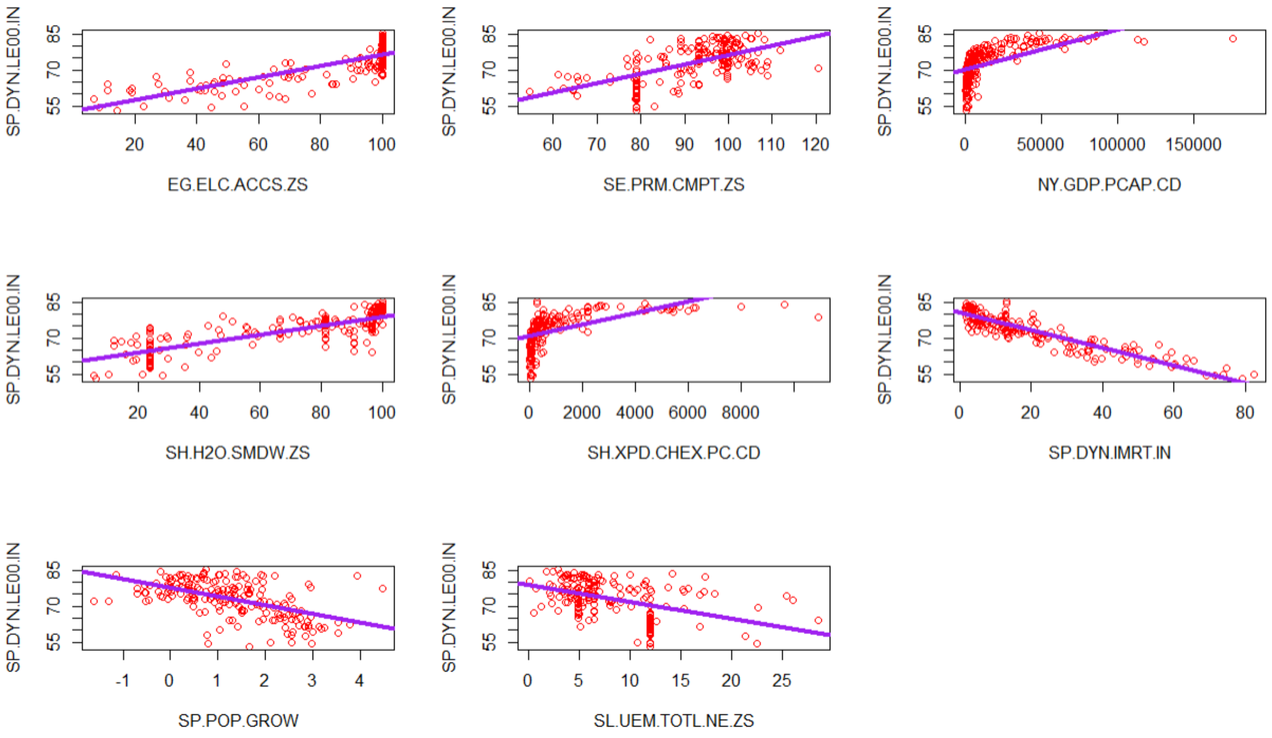
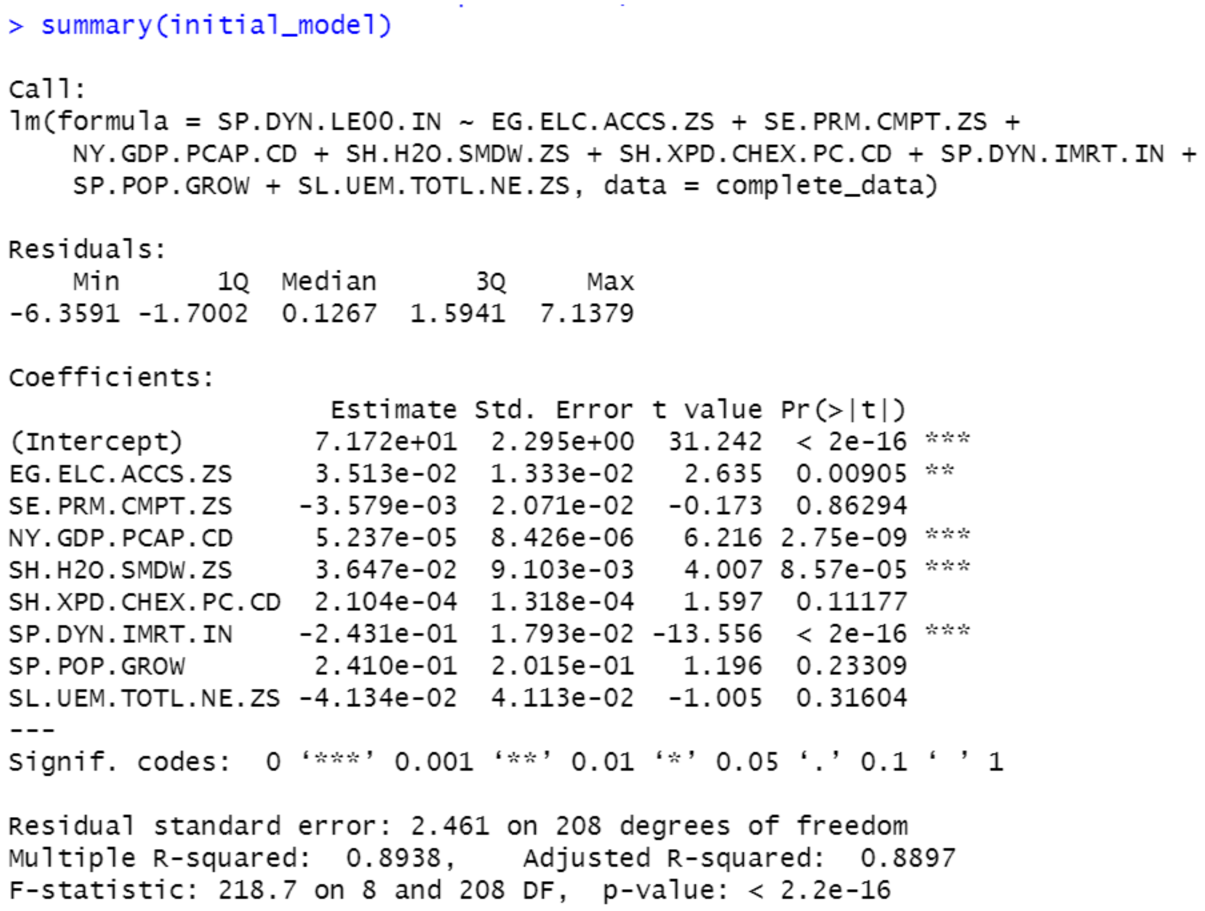


Figure 4.1: Linear Regression Models

most significant predictors are obtained. For each iteration, the variables SE.PRM.CMPT.ZS, SE.PRM.UNER and SL.UEM.TOTL.NE.ZS respectively are discarded from the scope and three models are derived.

The criterion to determine the best regression model is determined by implementing two types of tests, namely Alkaline Information Criterion (AIC) and Mallows' Cp[5] Test. After executing both the tests, it is observed that there is less difference between final_model2 and final_model3 and it can be seen clearly in their p-values as well. This occurrence is the result of imputing missing values in the response variable using linear regression method. Therefore, there are cases when either of the models become more significant. The Multiple Regression model is given by:

Life Expectancy At Birth = β_0 + β_1 (Access to electricity) + β_2 (GDP per capita) + β_3 (People using safely managed drinking water services) + β_4 (Current health expenditure per capita) - β_5 (Mortality rate, infant)



```
> summary(initial_model)

Call:
lm(formula = SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SE.PRM.CMPT.ZS +
    NY.GDP.PCAP.CD + SH.H2O.SMDW.ZS + SH.XPD.CHEX.PC.CD + SP.DYN.IMRT.IN +
    SP.POP.GROW + SL.UEM.TOTL.NE.ZS, data = complete_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3591 -1.7002  0.1267  1.5941  7.1379

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.172e+01  2.295e+00  31.242  < 2e-16 ***
EG.ELC.ACCS.ZS     3.513e-02  1.333e-02   2.635  0.00905 **
SE.PRM.CMPT.ZS    -3.579e-03  2.071e-02  -0.173  0.86294
NY.GDP.PCAP.CD     5.237e-05  8.426e-06   6.216 2.75e-09 ***
SH.H2O.SMDW.ZS     3.647e-02  9.103e-03   4.007 8.57e-05 ***
SH.XPD.CHEX.PC.CD  2.104e-04  1.318e-04   1.597  0.11177
SP.DYN.IMRT.IN    -2.431e-01  1.793e-02 -13.556  < 2e-16 ***
SP.POP.GROW        2.410e-01  2.015e-01   1.196  0.23309
SL.UEM.TOTL.NE.ZS -4.134e-02  4.113e-02  -1.005  0.31604
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.461 on 208 degrees of freedom
Multiple R-squared:  0.8938,    Adjusted R-squared:  0.8897
F-statistic: 218.7 on 8 and 208 DF,  p-value: < 2.2e-16
```

Figure 4.2: Summary Statistics of Multiple Regression Model

CHAPTER 5

Experimental Design for Analysing Average Life Expectancy Across Continents

The imputed data is used to investigate the life Expectancy across the continents. The mean of the Life Expectancy at birth is calculated, grouping them by continent. It can be inferred numerically from Table 5.1 and visually in Figure 5.1 that Europe has the highest mean value for life expectancy, which implies Europe has a higher life expectancy and Africa has the least Life Expectancy compared to the other continents and the rest of the continent pairs have a difference of 1-2 years.

Africa	Asia	Australia/Oceania	Europe	North America	South America
64.11014	74.61739	73.33691	79.57199	76.16346	75.09100

Table 5.1: Mean of Life Expectancy Across Continents

To understand the differences in average life expectancy across various continents and to assess the statistical significance between their means, the one-way ANOVA is chosen for analysis as there is one categorical independent variable(Continent) and one continuous dependent variable(Life Expectancy). The following two hypothesis is tested:

H0 (Null Hypothesis): Average life expectancy across all the continents is same

H1 (Alternate Hypothesis): Average life expectancy across all the continents is not same

The aov() fuction is used to execute a one-way ANOVA. The summary() function is used

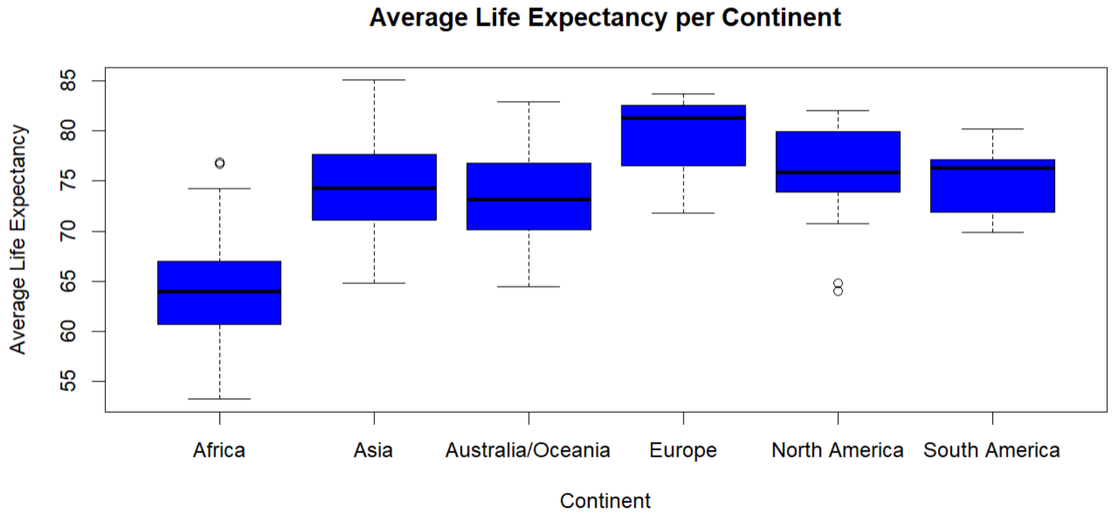
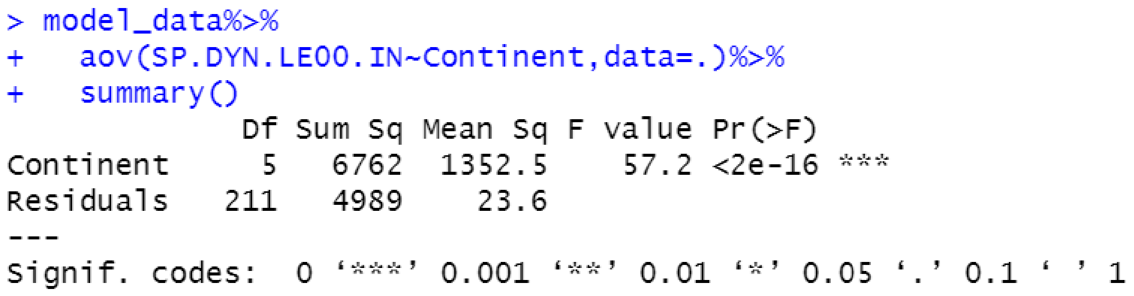


Figure 5.1: Boxplot of Life Expectancy Across Continents



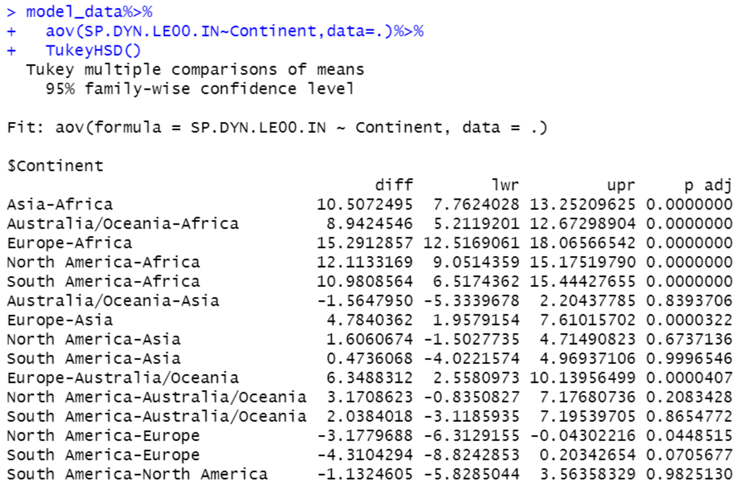
```
> model_data%>%
+   aov(SP.DYN.LE00.IN~Continent,data=.)%>%
+   summary()
             Df Sum Sq Mean Sq F value Pr(>F)    
Continent     5   6762  1352.5    57.2 <2e-16 ***
Residuals   211   4989    23.6                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.2: ANOVA Test

to interpret the results of the ANOVA analysis. On investigating, it can be seen in Figure 5.2 that the variable Continent has a significant F-value and the p-value is so small that it is almost close to zero. This evidence is sufficient to reject the null hypothesis and accept the alternate hypothesis.

Post hoc tests are used to determine which continents differ from each other with respect to the average life expectancy. Tukey’s HSD test is carried out to pairwise identify which sample differs from the other sample. The Tukey’s test is summarised in Figure 5.3 and Figure 5.4. Based on the Tukey’s Post Hoc Test, it is inferred that Africa has a significant average life expectancy difference with all the other continents. The test also shows that Europe has significant average life expectancy difference with Asia, North America and Australia/Oceania.

The standardised residual plot and the Normal Q-Q plot are commonly used for checking



```
> model_data%>%
+   aov(SP.DYN.LE00.IN~Continent,data=.)%>%
+   TukeyHSD()
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SP.DYN.LE00.IN ~ Continent, data = .)

$Continent
                                      diff        lwr         upr     p adj
Asia-Africa                     10.5072495  7.7624028 13.25209625 0.0000000
Australia/Oceania-Africa         8.9424546  5.2119201 12.67298904 0.0000000
Europe-Africa                   15.2912857 12.5169061 18.06566542 0.0000000
North America-Africa            12.1133169  9.0514359 15.17519790 0.0000000
South America-Africa            10.9808564  6.5174362 15.44427655 0.0000000
Australia/Oceania-Asia          -1.5647950 -5.3339678  2.20437785 0.8393706
Europe-Asia                      4.7840362  1.9579154  7.61015702 0.0000322
North America-Asia               1.6060674 -1.5027735  4.71490823 0.6737136
South America-Asia               0.4736068 -4.0221574  4.96937106 0.9996546
Europe-Australia/Oceania         6.3488312  2.5580973 10.13956499 0.0000407
North America-Australia/Oceania  3.1708623 -0.8350827  7.17680736 0.2083428
South America-Australia/Oceania  2.0384018 -3.1185935  7.19539705 0.8654772
North America-Europe            -3.1779688 -6.3129155 -0.04302216 0.0448515
South America-Europe            -4.3104294 -8.8242853  0.20342654 0.0705677
South America-North America     -1.1324605 -5.8285044  3.56358329 0.9825130
```

Figure 5.3: Tukey's HSD Post Hoc Test

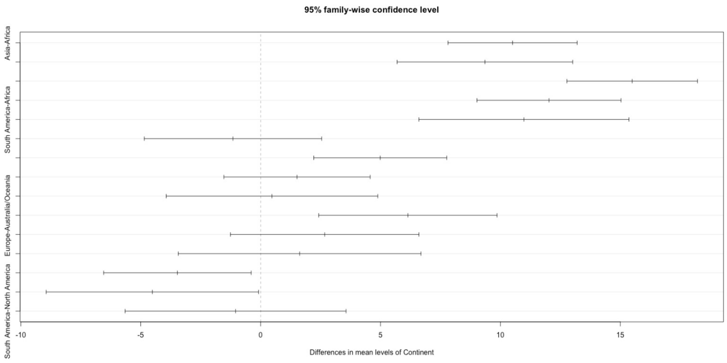


Figure 5.4: Tukey's HSD Confidence Interval Plot

the normality of the data. The standardised residual plot signifies that the residuals are normally distributed. The Q-Q plot of the residuals is approximately linear substantiating the condition that the residuals are normally distributed on the slope.

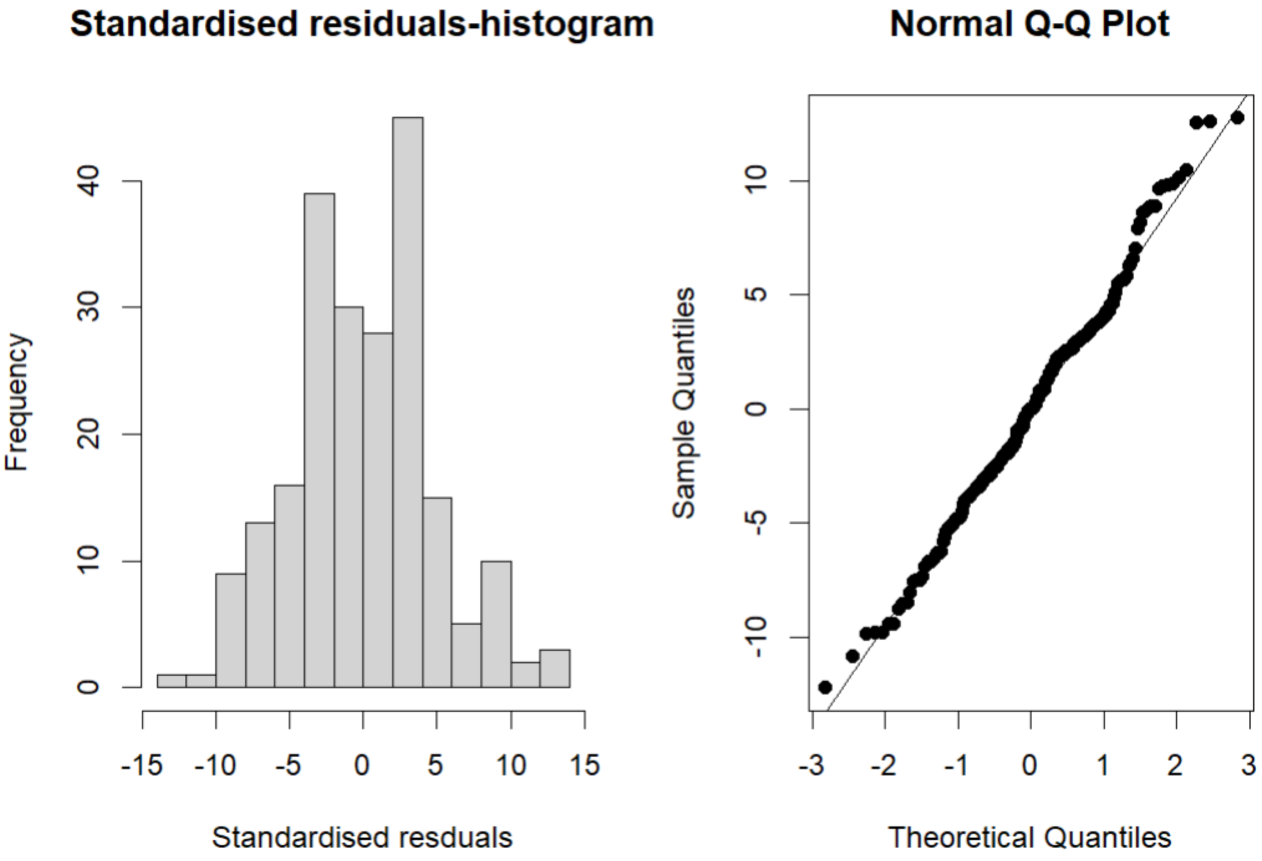


Figure 5.5: Standardised Residuals Plot and Normal Q-Q Plot

CHAPTER 6

Appendix

6.1 R Code for Data Summary and Plots

```
library(moments)
library('tidyverse')
library('mice')
library(ggcorrplot)
library(naniar)
library(data.table)
library('psych')
library('corrplot')
library('rstatix')
library(car)
library(olsrr)
library(AICcmodavg)
library(patchwork)
library(dplyr)
#loading data to dataframe
df<-read.csv('Life_Expectancy_Data1.csv')
data<-data.frame(df)
attach(data)

#descriptive statistics
data_summary<-summary(data)
View(data_summary)
#creating life-expectancy missing data subset
subset_SP.DYN.LE00.IN <- subset(data,is.na(SP.DYN.LE00.IN))

#For`` Skewness and Kurtosis Value

skewness(data$SP.DYN.LE00.IN,na.rm = TRUE)
kurtosis(data$SP.DYN.LE00.IN,na.rm = TRUE)

#checking continents having missing values in Life-Expectancy Column - SP.DYN.LE00.IN
subset_Europe<-data %>% filter((Continent=="Europe"))
subset_Euro<-data.frame(subset_Europe$Continent,subset_Europe$Country.Name,subset_Europe$SP.DYN.LE00.IN)
subset_Euro
sum(!is.na(subset_Euro$subset_Europe.Continent)) #total Europe values
sum(is.na(subset_Euro$subset_Europe.SP.DYN.LE00.IN)) #missing values
```

```
subset_Australia_Oceania<-data %>% filter((Continent=="Australia/Oceania"))
subset_Aus_Ocean<-data.frame(subset_Australia_Oceania$Continent,subset_Australia_Oceania$Country.Name,subset_Australia_Oceania$SP.DYN.LE00.IN)
subset_Aus_Ocean
sum(!is.na(subset_Aus_Ocean$subset_Australia_Oceania.Continent)) #total Australia/Oceania values
sum(is.na(subset_Aus_Ocean$subset_Australia_Oceania.SP.DYN.LE00.IN)) #missing values

subset_North_America<-data %>% filter((Continent=="North America"))
subset_Nor_Amer<-data.frame(subset_North_America$Continent,subset_North_America$Country.Name,subset_North_America$SP.DYN.LE00.IN)
subset_Nor_Amer
sum(!is.na(subset_Nor_Amer$subset_North_America.Continent)) #total North America values
sum(is.na(subset_Nor_Amer$subset_North_America.SP.DYN.LE00.IN)) #missing values


#scatter-plots for checking data distribution

par( mfrow= c(3,2) )
plot(EG.ELC.ACCS.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,10,maxColorValue=255), pch=16)
plot(NY.ADJ.NNTY.KD.ZG,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(NY.ADJ.NNTY.PC.KD.ZG,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SH.HIV.INCD.14,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SE.PRM.UNER,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SE.PRM.CUAT.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SE.TER.CUAT.BA.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SP.DYN.IMRT.IN,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SE.ADT.LITR.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(FR.INR.RINR,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SP.POP.GROW,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(EN.POP.DNST,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SP.POP.TOTL,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SH.XPD.CHEX.PC.CD,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SH.XPD.CHEX.GD.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SL.UEM.TOTL.NE.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(NY.GDP.MKTP.KD.ZG,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(NY.GDP.PCAP.CD,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SP.DYN.CBRT.IN,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SH.HIV.INCD,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SH.H2O.SMDW.ZS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SI.POV.LMIC,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)
plot(SE.COM.DURS,SP.DYN.LE00.IN, main="Scatterplot", col=rgb(0,100,0,50,maxColorValue=255), pch=16)


IQR <- IQR(SP.DYN.LE00.IN,na.rm = TRUE)
StdDev <- sd(SP.DYN.LE00.IN,na.rm = TRUE)
Median = median(SP.DYN.LE00.IN,na.rm = TRUE)
Mean = mean(SP.DYN.LE00.IN,na.rm = TRUE)

dev.off()

#Density function and histogram of Response Variable
density_plot = ggplot(data, aes(x=SP.DYN.LE00.IN)) +
  geom_histogram(aes(y=..density..), # Histogram with density instead of count on y-axis
                 binwidth=.5,
                 colour="black", fill="grey") +
  geom_density(alpha=.5, fill="#FFCC0011") + geom_vline(aes(xintercept=Median),
                                                        color="green", size=1)
density_plot+geom_vline(aes(xintercept=Mean), color="red", linetype="dashed", size=1)

#data-frame to visualise missing values
missing_data<-data.frame(data)

#Giving meaningful names to the variables to plot count of missing values graph
colnames(missing_data)[4] <- c(’Life expectancy at birth total’)
colnames(missing_data)[5:29] <- c(’Access to electricity’,
                                  ’Adjusted net national income’,
                                  ’Adjusted net national income per capita’,
                                  ’Children ages 0-14 newly infected with HIV’,
                                  ’Children out of school primary’,
                                  ’Educational attainment at least completed primary population’,
                                  ’Educational attainment at least Bachelorâs or equivalent’,
                                  ’Mortality rate infant’,’Primary completion rate total’,
                                  ’Literacy rate adult total’,
                                  ’Real interest rate’,’Population growth’,
                                  ’Population density’,’Population total’,
                                  ’Current health expenditure per capita PPP’,
                                  ’Current health expenditure’,’Unemployment total’,’GDP growth’,
                                  ’GDP per capita PPP’,’Birth rate crude’,
                                  ’Renewable energy consumption’,
                                  ’Adults ages 15-49 newly infected with HIV’,
                                  ’People using safely managed drinking water services’,
                                  ’Poverty headcount ratio at $3.20 a day’,
                                  ’Compulsory education duration’)

# Plotting missing values

gg_miss_var(missing_data)+labs(y = "Count of missing values")

# percentage of missing-values
initial_percent_missing<-(colMeans(is.na(data)))*100
View(initial_percent_missing)
```

6.2 R Code for Imputation

```
#dropping  predictor variables with >80% missing values
data <- data %>% select(-c(EG.FEC.RNEW.ZS,SI.POV.LMIC,SE.TER.CUAT.BA.ZS,SE.PRM.CUAT.ZS,SE.ADT.LITR.ZS,SH.HIV.INCD.14))



#replacing missing values with median continent wise

#calculating median values continent wise
median_values_per_continent<-data%>%
  group_by(Continent)%>%summarise_at(vars(EG.ELC.ACCS.ZS,
                                          NY.ADJ.NNTY.KD.ZG,
                                          NY.ADJ.NNTY.PC.KD.ZG,
                                          SE.PRM.UNER,
                                          SP.DYN.IMRT.IN,
                                          SE.PRM.CMPT.ZS,
                                          FR.INR.RINR,
                                          SP.POP.GROW,
                                          EN.POP.DNST,
                                          SP.POP.TOTL,
                                          SH.XPD.CHEX.PC.CD,
                                          SH.XPD.CHEX.GD.ZS,
                                          SL.UEM.TOTL.NE.ZS,
                                          NY.GDP.MKTP.KD.ZG,
                                          NY.GDP.PCAP.CD,
                                          SP.DYN.CBRT.IN,
                                          SH.HIV.INCD,
                                          SH.H2O.SMDW.ZS,
                                          SE.COM.DURS),list(name=median),
                                          na.rm=TRUE)

View(median_values_per_continent)

#Applying Shapiro Test to understand predictor variable normality

shapiro.test(EG.ELC.ACCS.ZS)
shapiro.test(NY.ADJ.NNTY.KD.ZG)
shapiro.test(NY.ADJ.NNTY.PC.KD.ZG)
shapiro.test(SE.PRM.UNER)
shapiro.test(SP.DYN.IMRT.IN)
shapiro.test(SE.PRM.CMPT.ZS)
shapiro.test(FR.INR.RINR)
shapiro.test(SP.POP.GROW)
shapiro.test(EN.POP.DNST)
shapiro.test(SP.POP.TOTL)
shapiro.test(SH.XPD.CHEX.PC.CD)
shapiro.test(SH.XPD.CHEX.GD.ZS)
shapiro.test(SL.UEM.TOTL.NE.ZS)
shapiro.test(NY.GDP.MKTP.KD.ZG)
shapiro.test(NY.GDP.PCAP.CD)
shapiro.test(SP.DYN.CBRT.IN)
shapiro.test(SH.HIV.INCD)
shapiro.test(SH.H2O.SMDW.ZS)
shapiro.test(SE.COM.DURS)


#Adding missing values in the columns
setDT(data)

data[, EG.ELC.ACCS.ZS := ifelse(is.na(EG.ELC.ACCS.ZS),
                                median(EG.ELC.ACCS.ZS, na.rm = TRUE),
                                EG.ELC.ACCS.ZS), by = Continent]
data[, NY.ADJ.NNTY.KD.ZG := ifelse(is.na(NY.ADJ.NNTY.KD.ZG),
                                   median(NY.ADJ.NNTY.KD.ZG, na.rm = TRUE),
                                   NY.ADJ.NNTY.KD.ZG), by = Continent]

data[,  NY.ADJ.NNTY.PC.KD.ZG:= ifelse(is.na(NY.ADJ.NNTY.PC.KD.ZG),
                                      median(NY.ADJ.NNTY.PC.KD.ZG, na.rm = TRUE),
                                      NY.ADJ.NNTY.PC.KD.ZG), by = Continent]

data[, SE.PRM.UNER := ifelse(is.na(SE.PRM.UNER),
                             median(SE.PRM.UNER, na.rm = TRUE),
                             SE.PRM.UNER), by = Continent]

data[, SP.DYN.IMRT.IN := ifelse(is.na(SP.DYN.IMRT.IN),
                                median(SP.DYN.IMRT.IN, na.rm = TRUE),
                                SP.DYN.IMRT.IN), by = Continent]

data[, SE.PRM.CMPT.ZS := ifelse(is.na(SE.PRM.CMPT.ZS),
                                median(SE.PRM.CMPT.ZS, na.rm = TRUE),
                                SE.PRM.CMPT.ZS), by = Continent]

data[, FR.INR.RINR := ifelse(is.na(FR.INR.RINR),
                             median(FR.INR.RINR, na.rm = TRUE),
                             FR.INR.RINR), by = Continent]

data[, SP.POP.GROW := ifelse(is.na(SP.POP.GROW),
                             median(SP.POP.GROW, na.rm = TRUE),
                             SP.POP.GROW), by = Continent]
data[,  EN.POP.DNST:= ifelse(is.na(EN.POP.DNST),
                             median(EN.POP.DNST, na.rm = TRUE),
                             EN.POP.DNST), by = Continent]

data[,  SP.POP.TOTL:= ifelse(is.na(SP.POP.TOTL),
                             median(SP.POP.TOTL, na.rm = TRUE),
                             SP.POP.TOTL), by = Continent]


data[,SL.UEM.TOTL.NE.ZS  := ifelse(is.na(SL.UEM.TOTL.NE.ZS),
                                   median(SL.UEM.TOTL.NE.ZS, na.rm = TRUE),
                                   SL.UEM.TOTL.NE.ZS), by = Continent]

data[,NY.GDP.MKTP.KD.ZG  := ifelse(is.na(NY.GDP.MKTP.KD.ZG),
                                   median(NY.GDP.MKTP.KD.ZG, na.rm = TRUE),
                                   NY.GDP.MKTP.KD.ZG), by = Continent]

data[,  SP.DYN.CBRT.IN:= ifelse(is.na(SP.DYN.CBRT.IN),
                                median(SP.DYN.CBRT.IN, na.rm = TRUE),
                                SP.DYN.CBRT.IN), by = Continent]
data[,  SH.HIV.INCD:= ifelse(is.na(SH.HIV.INCD),
                             median(SH.HIV.INCD, na.rm = TRUE),
                             SH.HIV.INCD), by = Continent]

data[, SH.XPD.CHEX.PC.CD := ifelse(is.na(SH.XPD.CHEX.PC.CD),
                                   median(SH.XPD.CHEX.PC.CD, na.rm = TRUE),
                                   SH.XPD.CHEX.PC.CD), by = Continent]
data[, SE.COM.DURS := ifelse(is.na(SE.COM.DURS),
                             median(SE.COM.DURS, na.rm = TRUE),
                             SE.COM.DURS), by = Continent]

data[, NY.GDP.PCAP.CD := ifelse(is.na(NY.GDP.PCAP.CD),
                                median(NY.GDP.PCAP.CD, na.rm = TRUE),
                                NY.GDP.PCAP.CD), by = Continent]


data[, SH.H2O.SMDW.ZS := ifelse(is.na(SH.H2O.SMDW.ZS),
                                median(SH.H2O.SMDW.ZS, na.rm = TRUE),
                                SH.H2O.SMDW.ZS), by = Continent]

data[, SH.XPD.CHEX.GD.ZS := ifelse(is.na(SH.XPD.CHEX.GD.ZS),
                                   median(SH.XPD.CHEX.GD.ZS, na.rm = TRUE),
                                   SH.XPD.CHEX.GD.ZS), by = Continent]



#Imputing the Life-Expectancy Column values based on MICE Imputation
miceData <- mice(data,m=5,method = 'cart')
complete_data<-complete(miceData)
final_percent_missing<-(colMeans(is.na(data)))*100
View(final_percent_missing)
```

6.3 R Code for Investigating Collinearity

```
#can features negatively correlated with the target variable be used?

#correlation matrix of data
numeric_columns<-data.frame(complete_data)
numeric_columns <- numeric_columns %>%
  select(-c(Country.Name,Country.Code,Continent))
cor_matrix <- cor_mat(numeric_columns)
View(cor_matrix)

#Correlation Matrix plot

corrplot.mixed(cor(numeric_columns),
               lower = FALSE,
               upper = "circle",
               tl.col = "black",addCoef.col = 1,
               number.cex = 0.5,tl.cex = 0.35)


#Correlated columns with life expectancy

#High positive correlation
```

```
#EG.ELC.ACCS.ZS
#SE.PRM.CMPT.ZS
#NY.GDP.PCAP.CD
#SH.H2O.SMDW.ZS
#SH.XPD.CHEX.PC.CD
#---------------
#High negative correlation
#SP.DYN.IMRT.IN
#SP.DYN.CBRT.IN
#SP.POP.GROW
#SL.UEM.TOTL.NE.ZS

#dropping  unwanted features from the complete_data
complete_data <- complete_data %>%
  select(-c(NY.ADJ.NNTY.KD.ZG,
            NY.ADJ.NNTY.PC.KD.ZG,
            FR.INR.RINR,
            SP.POP.TOTL,
            NY.GDP.MKTP.KD.ZG,
            SH.HIV.INCD,
            EN.POP.DNST,
            SE.COM.DURS,
            SH.XPD.CHEX.GD.ZS,
            SE.PRM.UNER
  ))

setDT(complete_data)
#Calculate the VIF for feature selection

initial_model_for_vif<-lm(SP.DYN.LE00.IN~
                            EG.ELC.ACCS.ZS+
                            SE.PRM.CMPT.ZS+
                            SH.XPD.CHEX.PC.CD+
                            NY.GDP.PCAP.CD+
                            SH.H2O.SMDW.ZS+
                            SP.DYN.IMRT.IN+
                            SP.DYN.CBRT.IN+
                            SP.POP.GROW+
                            SL.UEM.TOTL.NE.ZS,
                            data = complete_data)

vif(initial_model_for_vif)


#Based on the vif values, dropping independent variable SP.DYN.CBRT.IN

initial_model_for_vif_1<-lm(SP.DYN.LE00.IN~
                              EG.ELC.ACCS.ZS+
                              SE.PRM.CMPT.ZS+
                              SH.XPD.CHEX.PC.CD+
                              NY.GDP.PCAP.CD+
                              SH.H2O.SMDW.ZS+
                              SP.DYN.IMRT.IN+
                              SP.POP.GROW+
                              SL.UEM.TOTL.NE.ZS,
                              data = complete_data)
vif(initial_model_for_vif_1)

#final data for regression model building based on VIF < 5
complete_data<-complete_data%>%
  select(Country.Name,
         Country.Code,
         Continent,
         SP.DYN.LE00.IN,
         EG.ELC.ACCS.ZS,
         SE.PRM.CMPT.ZS,
         SH.XPD.CHEX.PC.CD,
         NY.GDP.PCAP.CD,
         SH.H2O.SMDW.ZS,
         SP.DYN.IMRT.IN,
         SP.POP.GROW,
         SL.UEM.TOTL.NE.ZS
  )
```

6.4 R Code for Multiple Linear Regression Modelling

```
#simple linear model

#Access to Electricity - EG.ELC.ACCS.ZS
par(mfrow=c(3,3))
model1<-lm(SP.DYN.LE00.IN~EG.ELC.ACCS.ZS,data = complete_data)
plot(EG.ELC.ACCS.ZS, SP.DYN.LE00.IN , col ="red")
abline(model1, lwd = 3, col="purple")
summary(model1)

#Primary completion rate, total (\% of relevant age group) - SE.PRM.CMPT.ZS
model2<-lm(SP.DYN.LE00.IN~SE.PRM.CMPT.ZS,data = complete_data)
plot(SE.PRM.CMPT.ZS, SP.DYN.LE00.IN , col ="red")
abline(model2, lwd = 3, col="purple")
summary(model2)


#GDP per capita, PPP (current international \$) - NY.GDP.PCAP.CD
model3<-lm(SP.DYN.LE00.IN~NY.GDP.PCAP.CD,data = complete_data)
plot(NY.GDP.PCAP.CD, SP.DYN.LE00.IN , col ="red")
abline(model3, lwd = 3, col="purple")
summary(model3)

#People using safely managed drinking water services - SH.H2O.SMDW.ZS
model4<-lm(SP.DYN.LE00.IN~SH.H2O.SMDW.ZS,data = complete_data)
plot(SH.H2O.SMDW.ZS, SP.DYN.LE00.IN , col ="red")
abline(model4, lwd = 3, col="purple")
summary(model4)

#Current health expenditure per capita,PPP - SH.XPD.CHEX.PC.CD
model5<-lm(SP.DYN.LE00.IN~SH.XPD.CHEX.PC.CD,data = complete_data)
plot(SH.XPD.CHEX.PC.CD, SP.DYN.LE00.IN , col ="red")
abline(model5, lwd = 3, col="purple")
summary(model5)

#Mortality rate, infant (per 1,000 live births) - SP.DYN.IMRT.IN
model6<-lm(SP.DYN.LE00.IN~SP.DYN.IMRT.IN,data = complete_data)
plot(SP.DYN.IMRT.IN, SP.DYN.LE00.IN , col ="red")
abline(model6, lwd = 3, col="purple")
summary(model6)

#Population growth (annual \%) - SP.POP.GROW
model7<-lm(SP.DYN.LE00.IN~SP.POP.GROW,data = complete_data)
plot(SP.POP.GROW, SP.DYN.LE00.IN , col ="red")
abline(model7, lwd = 3, col="purple")
summary(model7)

#Unemployment, total (\% of total labor force) - SL.UEM.TOTL.NE.ZS
model8<-lm(SP.DYN.LE00.IN~SL.UEM.TOTL.NE.ZS,data = complete_data)
plot(SL.UEM.TOTL.NE.ZS, SP.DYN.LE00.IN , col ="red")
abline(model8, lwd = 3, col="purple")
```

```
summary(model8)

#initial multiple regression model
# EG.ELC.ACCS.ZS
# SE.PRM.CMPT.ZS,
# NY.GDP.PCAP.CD,
# SH.H2O.SMDW.ZS,
# SH.XPD.CHEX.PC.CD,
# SP.DYN.IMRT.IN
#SP.POP.GROW
#SL.UEM.TOTL.NE.ZS



initial_model= lm(SP.DYN.LE00.IN ~
                    EG.ELC.ACCS.ZS+
                    SE.PRM.CMPT.ZS+
                    NY.GDP.PCAP.CD+
                    SH.H2O.SMDW.ZS+
                    SH.XPD.CHEX.PC.CD+
                    SP.DYN.IMRT.IN+
                    SP.POP.GROW+
                    SL.UEM.TOTL.NE.ZS,
                  data=complete_data)

summary(initial_model)


vif(initial_model)


#For model building, we will execute Backward Elimination

#Step 1-As SE.PRM.CMPT.ZS has highest p-value in 'initial_model', remove it
final_model1<-lm(SP.DYN.LE00.IN ~
                   EG.ELC.ACCS.ZS+
                   NY.GDP.PCAP.CD+
                   SH.H2O.SMDW.ZS+
                   SH.XPD.CHEX.PC.CD+
                   SP.DYN.IMRT.IN+
                   SP.POP.GROW+
                   SL.UEM.TOTL.NE.ZS,
                 data=complete_data)
summary(final_model1)

#Step 2 - As SL.UEM.TOTL.NE.ZS has the highest p-value, we will remove it

#considering top three features
final_model2<-lm(SP.DYN.LE00.IN ~
                   EG.ELC.ACCS.ZS+
                   NY.GDP.PCAP.CD+
                   SH.H2O.SMDW.ZS+
                   SH.XPD.CHEX.PC.CD+
                   SP.DYN.IMRT.IN+
                   SP.POP.GROW,
                 data=complete_data)

summary(final_model2)

#Step 3 - As SP.POP.GROW  has the highest p-value , we will remove it
final_model3<-lm(SP.DYN.LE00.IN ~
                   EG.ELC.ACCS.ZS+
                   NY.GDP.PCAP.CD+
                   SH.H2O.SMDW.ZS+
                   SH.XPD.CHEX.PC.CD+
                   SP.DYN.IMRT.IN,
                 data=complete_data)


summary(final_model3)

#comparing the models for best regression model
summary(final_model1)
summary(final_model2)
summary(final_model3)

#AIC (Alkaline Information Criterion) TEST for model selection

aic_model_test<-list(final_model1,final_model2, final_model3)

mod.names <- c(’final_model1’, ’final_model2’, ’final_model3’)

aic_model<-aictab(cand.set = aic_model_test, modnames = mod.names)
aic_model

#run mallows cp test
ols_mallows_cp(final_model1, initial_model)
ols_mallows_cp(final_model2, initial_model)
ols_mallows_cp(final_model3, initial_model)



#Constructing data containing Continent and Life Expectancy only
model_data<-complete_data%>%
  filter(Continent %in%
           c("Africa",
             "Asia",
             "Australia/Oceania",
             "Europe",
             "North America",
             "South America"))%>%
  select(Continent,SP.DYN.LE00.IN)

View(model_data)

#Summarise life expectancy mean by continent

model_data%>%
  group_by(Continent)%>%
  summarise(Average_By_Continent=mean(SP.DYN.LE00.IN))%>%
  arrange(Average_By_Continent)

boxplot(model_data$SP.DYN.LE00.IN~
          model_data$Continent,main=’Average Life Expectancy per Continent’,
          xlab=’Continent’, col="blue", ylab = "Average Life Expectancy",)
model_data%>%
  aov(SP.DYN.LE00.IN~Continent,data=.)%>%
  summary()

aov_model<-model_data%>%
  aov(SP.DYN.LE00.IN~Continent,data=.)


model_data%>%
  aov(SP.DYN.LE00.IN~Continent,data=.)%>%
  TukeyHSD()

tukey.SP.DYN.LE00.IN<-TukeyHSD(aov_model)
plot(tukey.SP.DYN.LE00.IN)

model_data$residulas1 <- aov_model$residuals
par(mfrow=c(1,2))
```

```
hist(model_data$residulas1, main="Standardised residuals-histogram",
     xlab="Standardised resduals")

qqnorm(model_data$residulas1,pch=19)
qqline(model_data$residulas1)

shapiro.test(model_data$residulas1)
```

Chapter 7

Conclusion

In the Life Expectancy dataset, the variables having null values more than 80% are dropped. The missing values are imputed with median value for the predictor variables and using mice imputation for the response variable. The multiple linear regression model is built and examined with the final predictor variables after feature selection. Backward Elimination Technique is applied to eliminate variables that do not contribute significantly to the model. The AIC and Mallows' Cp metric is used to obtain the best model among the three models obtained during the backward elimination. The difference in mean life expectancy across continents is analysed using ANOVA Test and Tukey's HSD Test. By pair-wise comparison, it is found that Europe has the highest mean life expectancy and Africa the least.

Bibliography

[1] Laerd Statistics (2018). Understanding Descriptive and Inferential Statistics. [online] Laerd Statistics.

[2] Menon, K. (2022). The Complete Guide To Skewness And Kurtosis | Simplilearn. [online]

[3] AREZKI, Y. (2020). Handling missing data MCAR, MAR and MNAR (Part I). [online] kaggle.com.

[4] Pulagam, S. (2020). How to detect and deal with Multicollinearity. [online] Medium.

[5] online.stat.psu.edu. (2018). 11.3 - Best Subsets Regression, Adjusted R-Sq, Mallows Cp | STAT 462. [online]

[6] Rdocumentation.org. (2019). R Documentation and manuals | R Documentation. [online]

[7] cran.r-project.org. (n.d.). An Introduction to corrplot Package. [online]