

MA317 MODELLING EXPERIMENTAL DATA

GROUP COURSEWORK

TITLE: Analysis of World Development Indicators (WDI) Dataset

Group Number: GROUP 20

Group Members: Bhavya Barbhaya

Rudra Mehta

Fatma Demir

Raktim Ghosh

~~Syed Shah~~

Module Supervisor: Dr Stella Hadjiantoni

*March 27, 2022*

Contents

[**Abstract** 1](#_Toc99350340)

[**Word Count** 1](#_Toc99350341)

[**1.** **Introduction** 2](#_Toc99350342)

[**Initial Investigation** 2](#_Toc99350343)

[**2.** **Preliminary Analysis** 7](#_Toc99350351)

[**3.** **Analysis** 8](#_Toc99350352)

[**4.** **Life Expectancy Prediction** 10](#_Toc99350353)

[**5.** **Conclusion** 15](#_Toc99350354)

[**6.** **References** 16](#_Toc99350355)

[**Appendices** 17](#_Toc99350356)

[**Appendix I: R-code with Comments** 17](#_Toc99350357)

[**Appendix II: New Variable Name and Indicator Name** 21](#_Toc99350358)

[**Appendix III: Acronyms/Abbreviations** 21](#_Toc99350359)

[**Contribution** 22](#_Toc99350360)

**Abstract**

Data which is one of the most important assets in the world rightnow and is being captured in an increasing amount every millisecond. With statistical measures, we can pull out different aspects of data. In this project, we evaluated different variables that have significant levels in life expectancy using R Studio. We have first analyse the dataset using both graphical and numerical representation. The next step covered in this report is dealing with missing values in the dataset by either deleting or applying other imputation techniques depending on how much missing values are detected on each variable. Then we have used the Algorithm which can give us the best fit model. Also we have done few assumptions checks so we can conclude that we have the Best fit Model.

**Word Count**

2505

1. **Introduction**

Using the life expectancies dataset for year 2019, we run through numerical and graphical descriptive statistics, to get a high level view of the data along with identification of the response & predictor variables which we use to build the predictive model later on. It is worthwhile mentioning that the dataset is having a lot data quality issues which may have been due to failure of proper observation recording or other challenges as the case may be. We ist tried with complete case analysis to address this but was left with no data due to high missing values almost in all observations. Accordingly we used data imputation technique as suitable for the continuous or categorical data to address the missing values and make it usable by the respective R functions. Subsequently we test the multiple predictor variables for multicollinearity. This is due to the fact as multiple predictors exists in the dataset and there is a linear relationship among two or more of these predictors. We have used Linear model to extrapolate the life expectancies of the given countries outside the given observation.

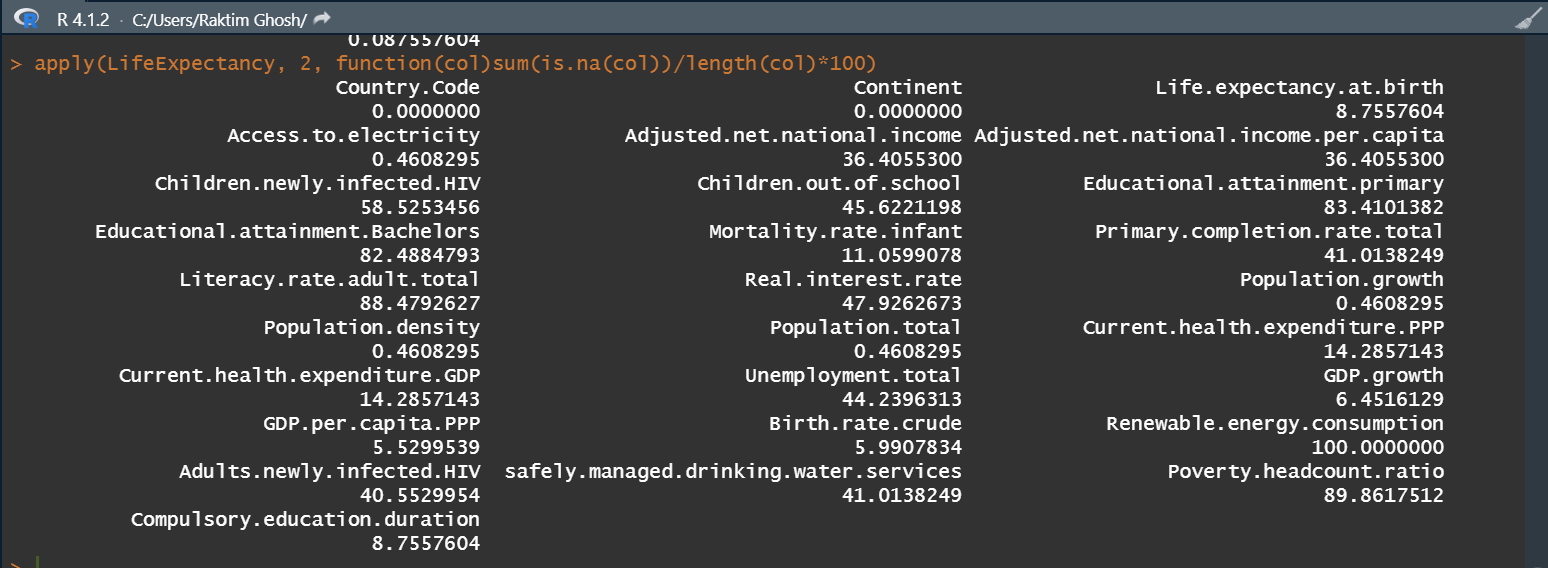
**Initial Investigation**

We browse through the file provided to get an idea of the data. Mainly the Dimensions and categorical attributes which help to give an insight into the data story and aids in data discovery and further steps towards achieving the model.From the file we come to know the following statistics : Number of observation : 217. Number of Predictors : 25. Categorical Data : 3 Responses : 1

**Data Discovery**

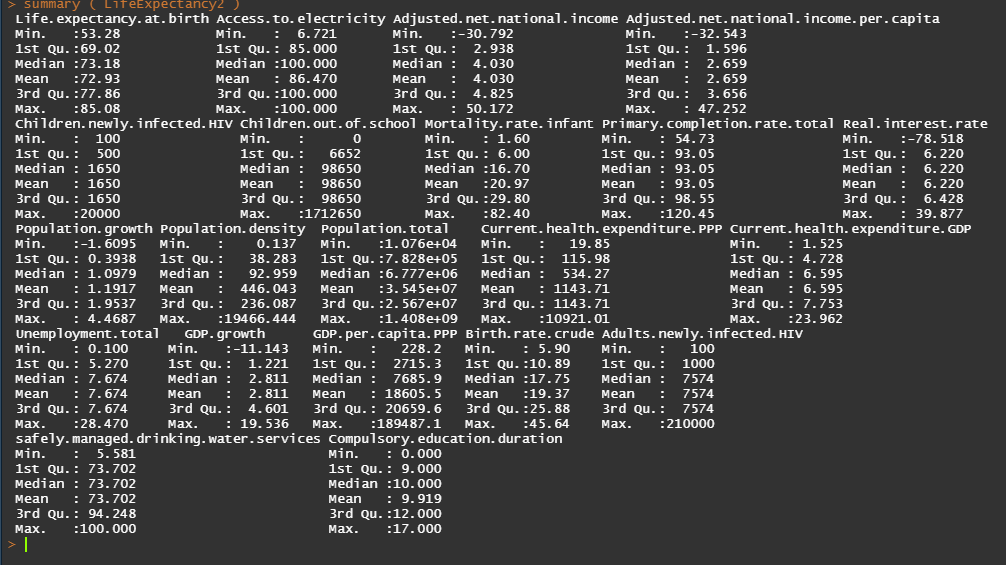
We do a detailed scrutiny of the dataset by importing into the IDE in csv format and run it through various functions of R to check the missing values. As a first step, we try to eliminate the imperfect observations from the dataset but we find we are left with not a sizable sample. We also look for any additional whitespaces / newline characters accidentaly come in the observations, we readjust the file to read the required observations only for the same.

We then use the str() and dim(), to confirm the structure of the data that was loaded in the environment as well as the type of the data we are to deal with. Next we standardise the variable names to remove any ambiguity and better readability and accessibility.

We checked for the percentage of missing values for each variable using the apply() function and generated a graphical representation of the missing values using the aggr () function in the VIM library which was easier to understand and drop attributes with missing values above 70%. 

From the table above, It is seen that Educational.attainment.primary, Educational.attainment.Bachelors,Literacy.rate.adult.total,Renewable.energy.consumption and Poverty.headcount.ratio are having missing values of above 80%, which makes insignificant contribution to the analysis. Imputation on these predictors tend to create noise and make the model biased. Hence we decide to take these out of the dataset-LifeExpectancy2. For the variables with lesser missing values below 70%, we apply mean imputation technique to replace the NAs with the average value of the responses from valid entries in each variable.

Then we perform some basic checks like get a histogram view of the Life.expectancy.at.birth and observe that majority of the frequencies occur between 70-75 years. Seeing the signifance values of the predictors in the premliminary model this likely due to the basic infrastructure and education parameters in most of the coutries. As we see the summary statistics for the overall code.

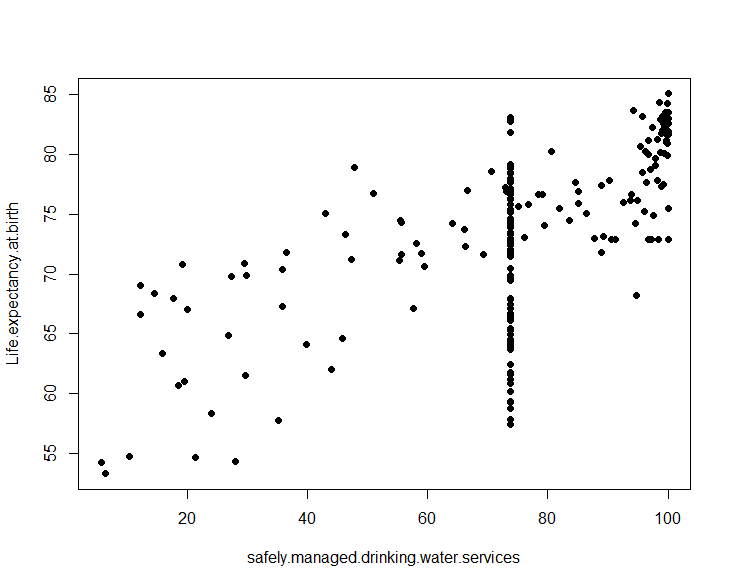
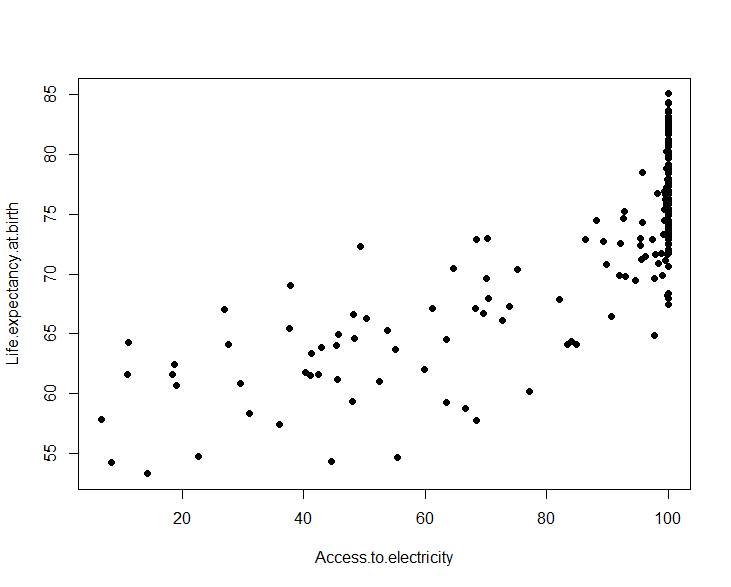


We find that few of basic indicators looks good like "Access.to.electricity", "Primary.completion.rate.total","Unemployment.total".

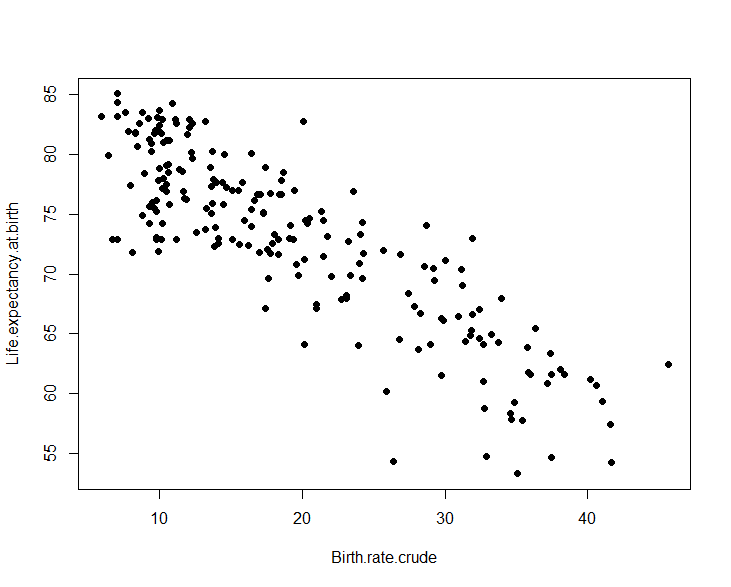
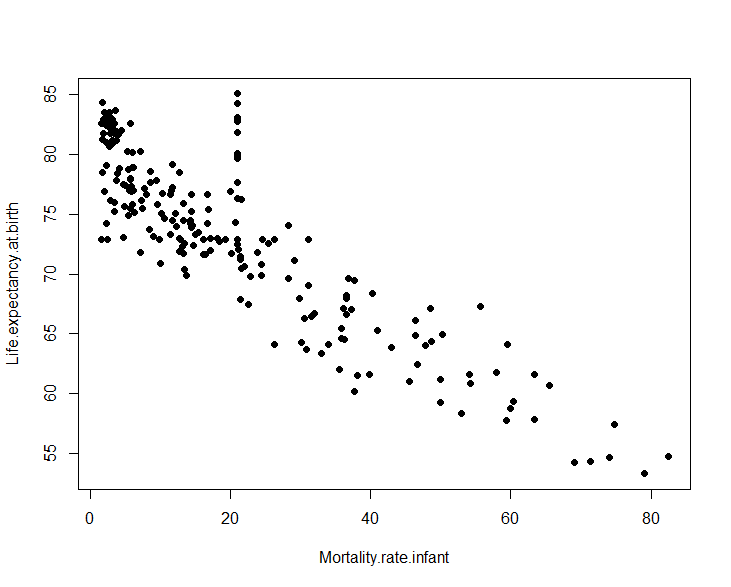
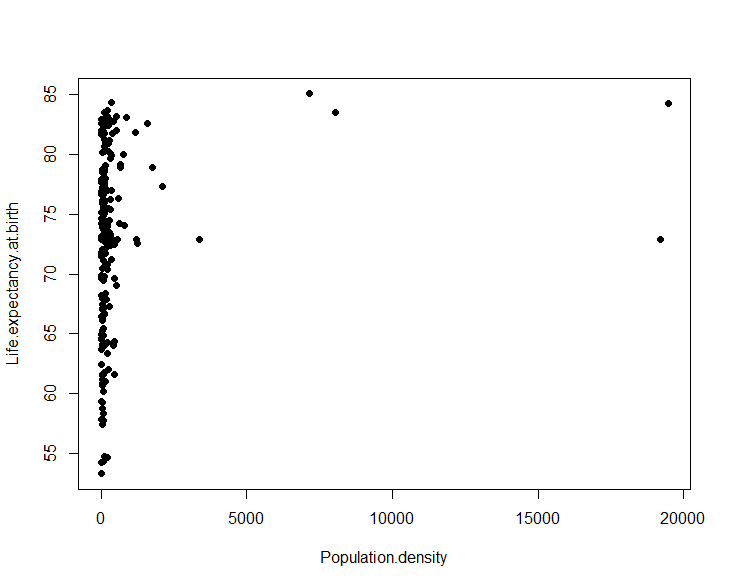
we find the growth parameters like to be moderate across the sample example: "GDP.growth", "GDP.per.capita.PPP", "Unemployment.total" etc.

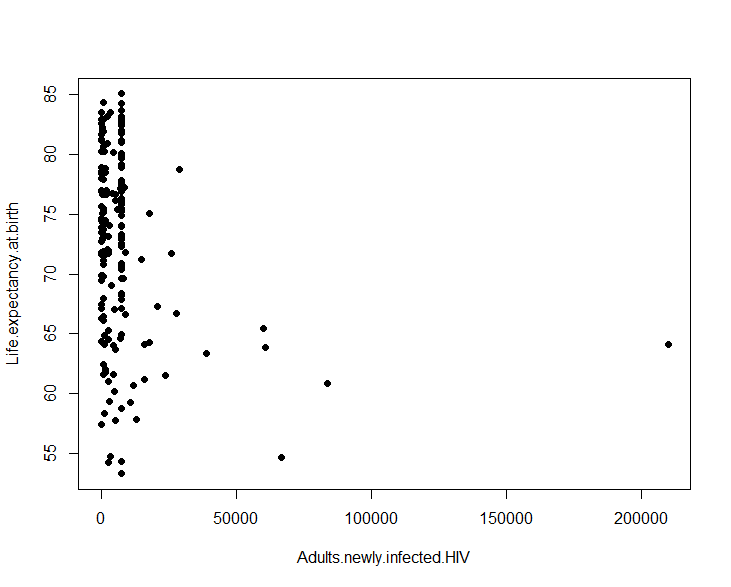
some of the indicators which were worth looking "safely.managed.drinking.water.services" , "Adults.newly.infected.HIV", "Children.newly.infected.HIV" etc these may be in the vorder line of getting to extreme.

**Graphical Analysis**

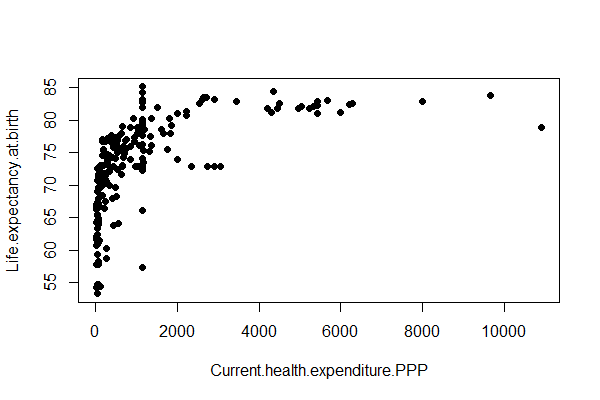
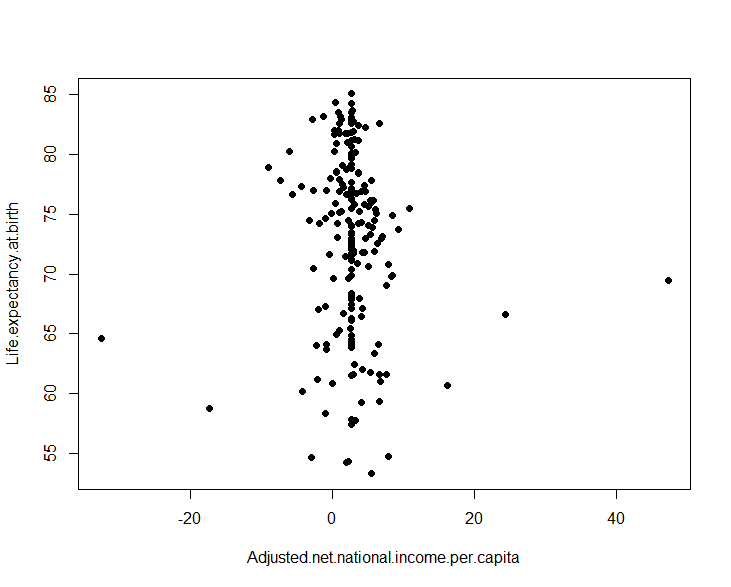
We witness some interesting results from the plot of a few major predictors. A moderate positive correlation can be seen among [Access.to.electricity] & [safely.managed.drinking.water.services] to the life expectancy.

It’s also obserbed that Life Expectancy exhibits a strong negative correlation with [Mortality.rate.infant] & [Birth.rate.crude].

[Population.density] & [Adults.newly.infected.HIV] features a distinctive pattern with Life Expectancy.



[Current.health.expenditure] & [Adjusted.net.national.income.per.capita] displays non linear but distinctive pattern which may be opf interest to further analysis.

1. **Preliminary Analysis**

Before diving into the preliminary analysis, we started off with sorting and cleaning the dataset as we analysed that there were some missing values in the Dataset. Below stated are the steps that we carried out before performing the preliminary analysis.

**2.1 Loading Dataset and Removing the Missing Values**

As the heading suggests, we have loaded the dataset onto the R-studio and we have successfully loaded the dataset with read.csv() command and assigned it to a variable called “LifeExpectancy”, after which we changed the column names of our dataset for better understanding using the colnames() function(we changed it to the same column names provided in the group coursework file). As we further observed the Life Expectancy Data, we saw that many variables have missing values in almost all variables so we went ahead and identified the percentage of missing values in each of the variable. We used the apply() function included in the tidyverse package and successfully got the percentage of missing values in each variable as you can see in figure 1.1 below:-

Graphical user interface, text, application

Description automatically generated

From the figure 1.1, We can identify that Some variables have missing values which are more than 70% like Educational.attainment.primary, Educational.attainment.Bachelors, Literacy.rate.adult.total, Renewable.energy.consumption, Poverty.headcount.ratio . As the variables had more than 70% of missing values and that cannot be considered while going ahead with our linear model because it will hamper our analysis of the model and won’t fetch us a good accuracy while doing the analysis. So, we went ahead and Dropped those variables using their index numbers and created a new data frame “LifeExpectancy2” . Also, some variables like Country.Code and Continent had character values which we checked with the class() function and those cannot be considered in a linear model, therefore we dropped them as well. After dropping these variables, we had a much cleaner version of the dataset but we still had some good amount of missing values inside other variables as well. So to deal with those missing values we tried and tested 2 methods.

**2.1.1 Imputation with Median Method**

As the name suggests, we used the median() function for all the remaining variables below 70% to replace the missing values with the median of all the values in each variable. Once the values were replaced. We ran the linear model LifeExpectancy2 in order to identify that all the values have been successfully replaced and to Compare the Multiple R-Squared Value with the Imputation with Mean Method(2.1.2).

**2.1.2 Imputation with Mean Method**

Using the mean() function we replaced the rest of the missing values with the mean of all the values in each variables in LifeExpectancy2. Here, also We ran the linear model in order to compare it with the Linear model we had in Imputation with Median Method and we analysed that Imputation of Mean has better R-Squared values and Also it is very efficient and easy to work with compared to the Median Method.

So after doing the following steps we have successfully completed a preliminary analysis and to conclude this analysis we have successfully removed the variables with high(More than 70%) missing values, removed character variables and Imputed the Mean values in place of the remaining Missing Values to help with the further analysis of Life Expectancy and to make the model more accurate and efficient to work with.

1. **Analysis**

Carrying out the analysis of any dataset is a step-by-step procedure and the first step in analysis is to get a Correlation matrix (Heatmap) and using that Matrix we can keep Life.expectancy.at.birth as our intercept and identify the significant values useful to us for creating a model.

**3.1 Correlation Matrix**

We have installed and used the Library Corrplot and applied the cor() function on the Lifeexpectancy2 dataframe which provided us with a correlation matrix and using that correlation matrix we generated a heatmap using corrplot() inorder to identify the most significant and least significant variables. So According to the Heatmap provided in figure 1.2, we identified that the following variables "Adjusted.net.national.income", "Adjusted.net.national.income.per.capita", "Children.newly.infected.HIV", "Children.out.of.school", "Mortality.rate.infant","Primary.completion.rate.total", "Real.interest.rate", "Population.growth", "Population.density", "Population.total", "Current.health.expenditure.GDP", "Unemployment.total", "GDP.growth", "Birth.rate.crude", "Adults.newly.infected.HIV", "Compulsory.education.duration" are least significant as their correlation coefficients were less than 0.5 when compared to the intercept. For getting a better fit and unbiased Linear Model we need to consider significant variables in our dataframe keeping Life.expectancy.at.birth as our intercept.

Chart, scatter chart

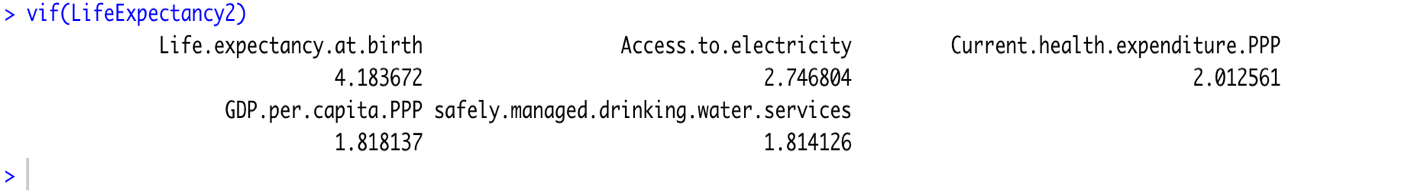
Description automatically generated

**3.2 Dropping the Least Significant Variables**

We created a droppingvar variable which contains all the variable names that are needed to be dropped out of LifeExpectancy2 and dropped those variables so that we are only left with the Most significant and significant variables with correlation coefficients more than 0.5.

**3.3 Variance Inflation Factor(VIF)**

Variance Inflation Factor(VIF) is used to identify Multicollinearity in a set of regression variables. So using the vif() function we identified the correlation between variables as shown in the figure 1.3. As we can clearly see that all the vif values lie between 1 to 5 so that indicates moderate correlation between the variables left in the dataframe LifeExpectancy2. So its safe to assume that now we can go ahead and create Linear Regression Model based on Significant variables that affect our life expectancy at birth and gives us the best fit model of it.



1. **Life Expectancy Prediction**

**Life Expectancy Prediction**

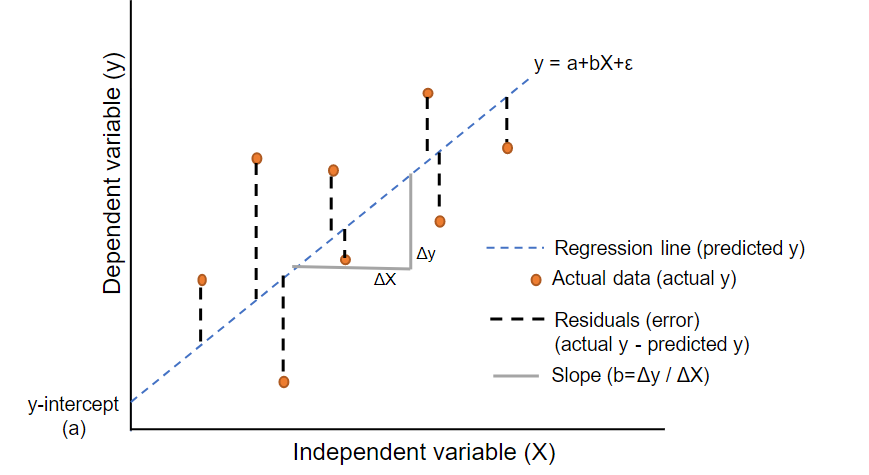
Now the main question comes that is there any way to predict life expectancy based on discussed features? Yes, but we need to decide first that which supervised algorithm that fits our task.

We have lot of algorithms for regression tasks and each algorithm has its pros and cons. One algorithm might fetch superior results compared to other algorithms but it might lack in terms of explainability. Let’s say if explainability is not compromised the development of such complex algorithms is a very tiring task. So an optimal algorithm must be accurate, easy to develop and explainable.

For Instance, Linear Regression is comparatively Explainable and Simple Algorithm. Deployment of Linear Regression requires nominal efforts but on the other side when the data is non – linear it lacks accuracy. Complex algorithms perform better on Non-Linear datasets, but then the model lacks explainability.

**Linear Regression**

Linear Regression is a regression algorithm with a linear approach. It is a supervised regression where we try to predict a continuous value of a given data point by generalizing the data we have. The Linear Part indicated the linear approach for generalization of the data.



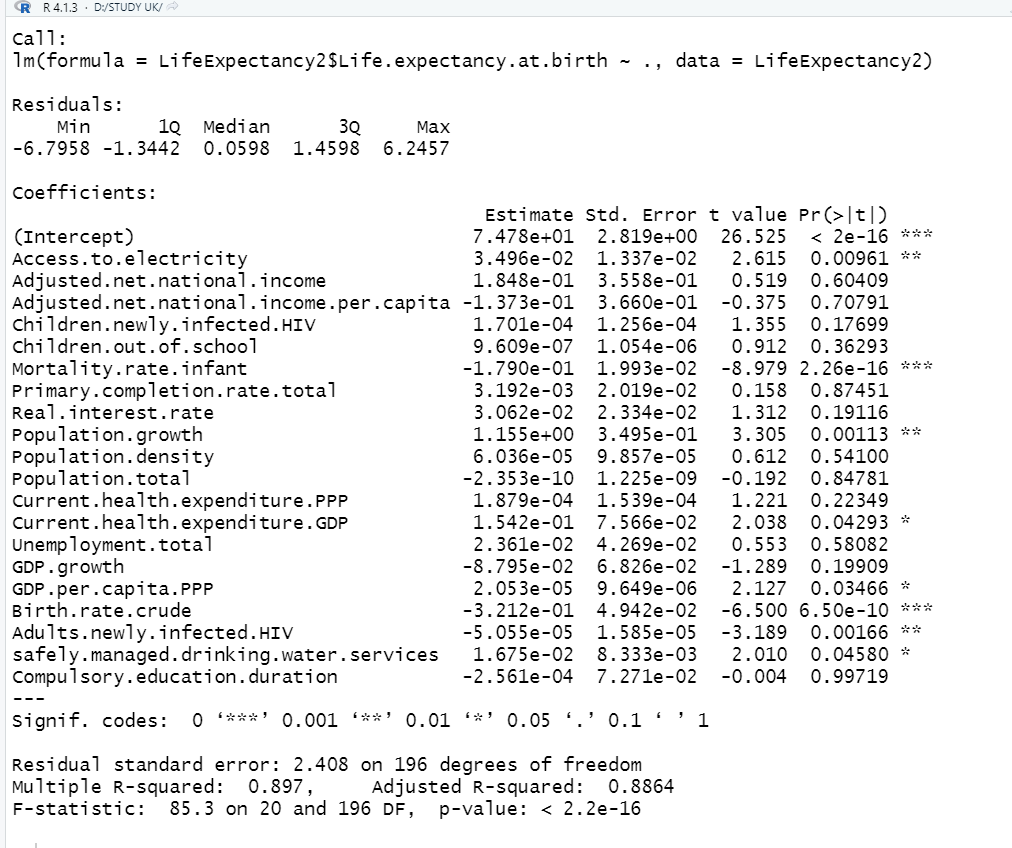
Source :- Google Images

Here the idea is to predict the dependent variable (Y) using a given independent variable (X). All this can be accomplished by fitting a best fit line in the data. A line that is providing the least sum of residual error is the best fit line or regression line

Let’s Predict the Life expectancy by using linear Regression

* Creating a Model

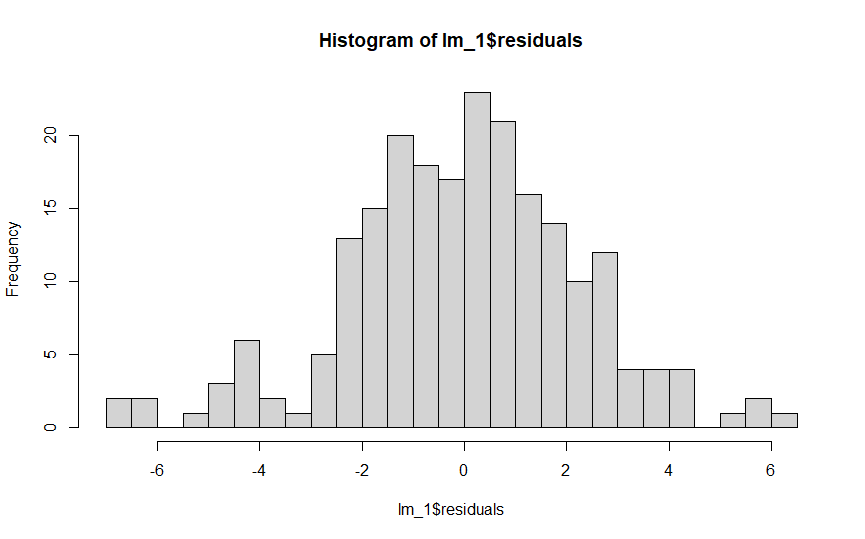
Running the Linear regression on the data which has all the variables including least significant variables.



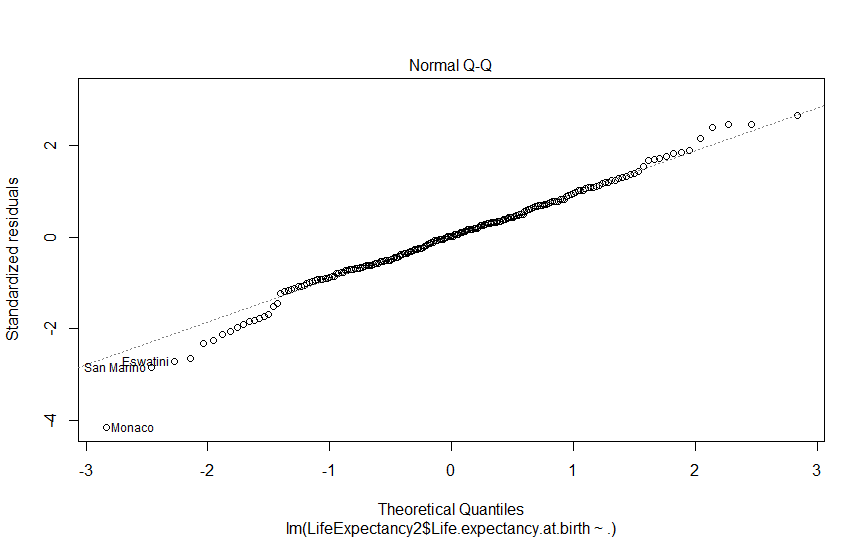
Coefficients interpretations: - Since there will no condition where all independent variables are 0, then the intercept relatively has no meaning in this context. About the coefficients, interesting findings can be seen there, some of the numerical Variables may give negative effects

Adj. R-squared interpretation: Approximately 88.64% of the observed variation can be explained by the model’s inputs, this is a good result, indicating that we are on the right path to create good linear model.

For Validation of this model let’s check distribution of residuals.

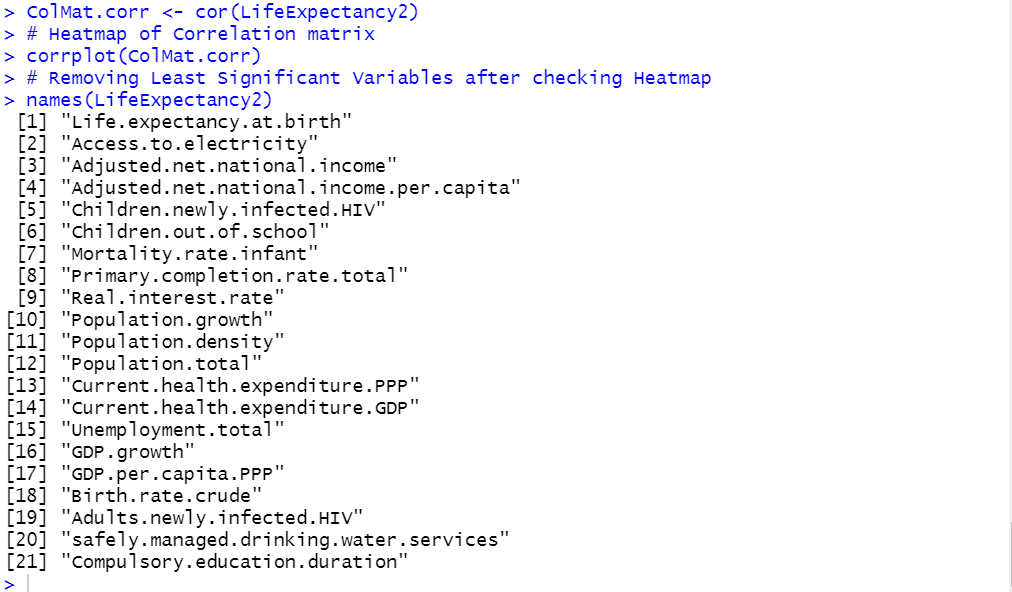


The Residual distribution is approximately normal, having a mean close to zero but there are some variable which are least significant

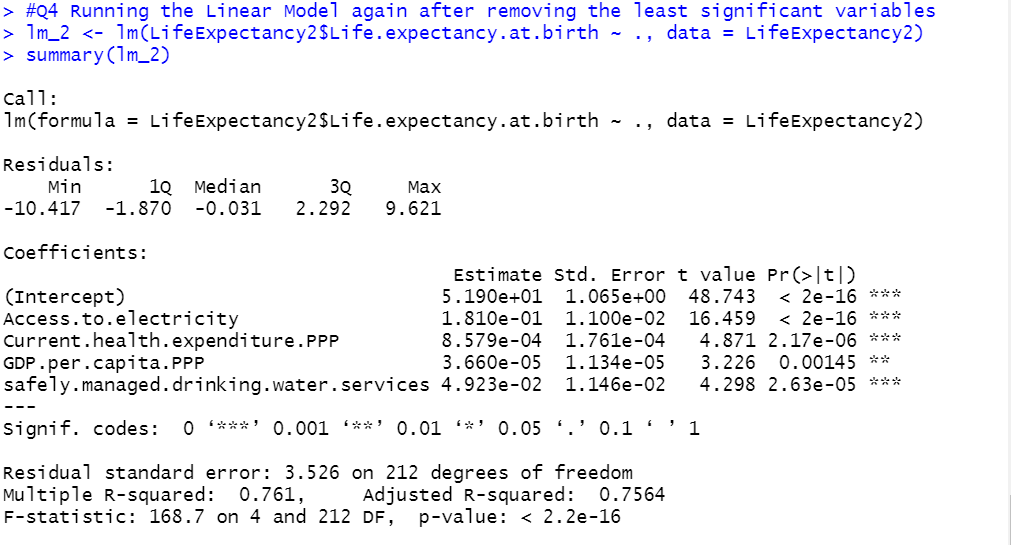


Most of the residuals gathered on the centre line, indicates they are distributed normally

Now let’s remove some of least significant variables and then run & check the linear model again.



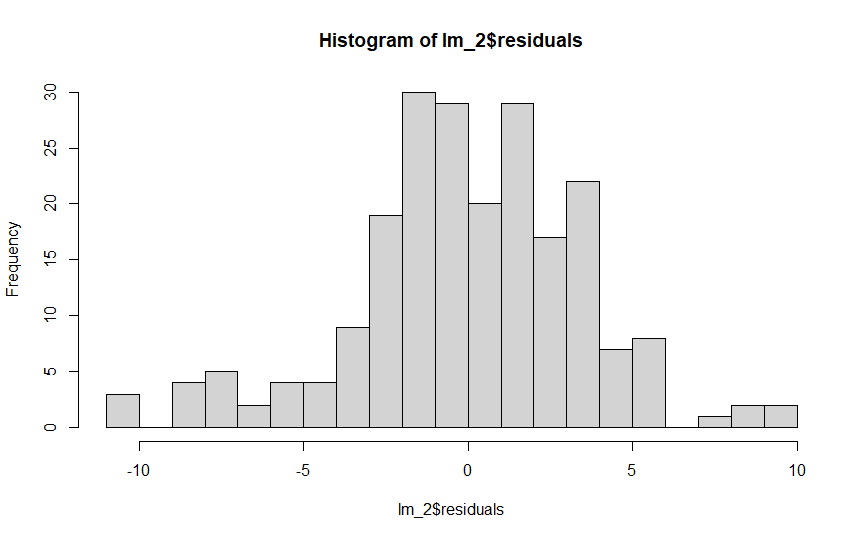
Now running the linear model after removing the least significant variables .



Coefficients interpretations: - Since there will no condition where all independent variables are 0, then the intercept relatively has no meaning in this context. About the coefficients, interesting findings can be seen there, some of the numerical Variables may give negative effects

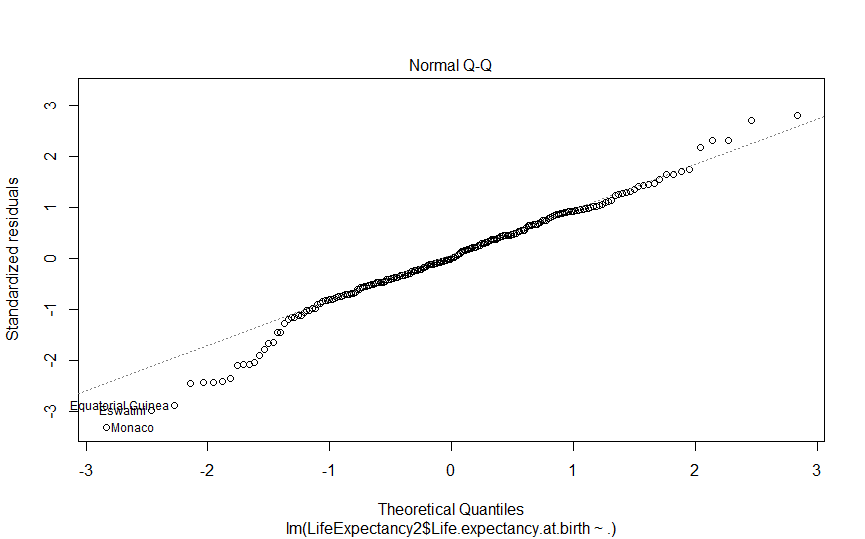
Adj. R-squared interpretation: Approximately 75.64% of the observed variation can be explained by the model’s inputs, this is a good result, indicating that we are on the right path to create good linear model.

For Validation of this model let’s check distribution of residuals.



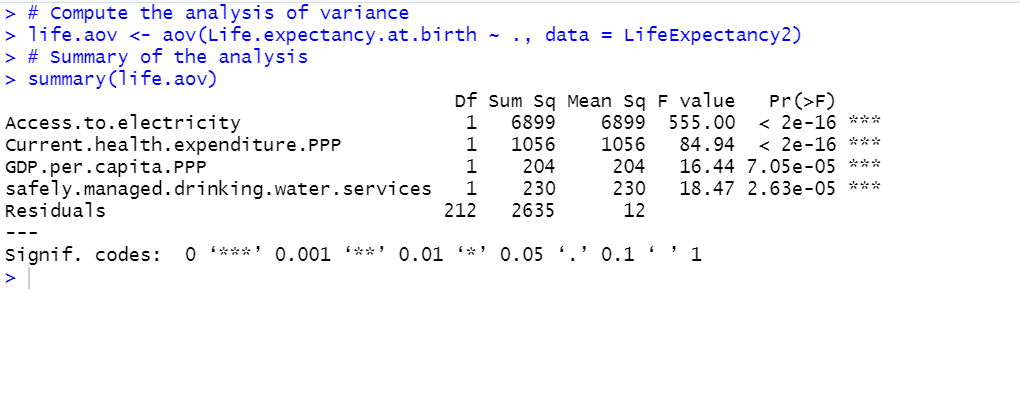
The Residual distribution is approximately normal, having a mean close to zero.

Now lets plot residuals on the QQ plot



Most of the residuals gathered on the centre line, indicates they are distributed normally.

Now let’s compute the analysis of the Variance



As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

Finally we have covered the linear regression and implemented it to predict the Life Expectancy. The Linear model seems to be fit to predict the Life Expectancy based on the Adj R- Squared Value, Error Value and also pass 2 of the Assumptions Check.

Linear Model can be used to explain the linear correlation between Life Expectancy and the selected Independent Variables. One can extend their life span by adopting a healthy lifestyle, proper education and lot other parameters but demographic roles play an important role. In our analysis we found that Hong Kong SAR China has the highest life Expectancy at the birth and Monaco has the highest GDP per capita and United States has the highest health expenditure. At the same time Central African Republic has the least life Expectancy at the Birth. A country’s GDP and Income Composition affects the life expectancy more broadly.

1. **Conclusion**

To Conclude the study included by the examination of the “World Development Indicators (WDI)” derived from such a Global Central database in order to obtain a Linear Model in explaining average life expectancy in 2019. In this report we have covered almost the Analysis of the given data and also we have used Linear model which is the bestfit model. we have covered the linear regression and implemented it to predict the Life Expectancy. The Linear model seems to be fit to predict the Life Expectancy based on the Adj R- Squared Value, Error Value and also pass 2 of the Assumptions Check.

Linear Model can be used to explain the linear correlation between Life Expectancy and the selected Independent Variables. One can extend their life span by adopting a healthy lifestyle, proper education and lot other parameters but demographic roles play an important role. In our analysis we found that Hong Kong SAR China has the highest life Expectancy at the birth and Monaco has the highest GDP per capita and United States has the highest health expenditure. At the same time Central African Republic has the least life Expectancy at the Birth. A country’s GDP and Income Composition affects the life expectancy more broadly.

1. **References**
2. https://cran.r-project.org
3. Google.com
4. Youtube.com
5. Front Page Logo. “University of Essex Logo”. [Online]. Available: http://www.essex.ac.uk/.

**Appendices**

**Appendix I: R-code with Comments**

# Load Dataset

LifeExpectancy<- read.csv("Life\_Expectancy\_Data1.csv",row.names=1)

view(LifeExpectancy)

#Changing Column names

colnames(LifeExpectancy) <- c (

"Country.Code",

"Continent",

"Life.expectancy.at.birth",

"Access.to.electricity",

"Adjusted.net.national.income",

"Adjusted.net.national.income.per.capita",

"Children.newly.infected.HIV",

"Children.out.of.school",

"Educational.attainment.primary",

"Educational.attainment.Bachelors",

"Mortality.rate.infant",

"Primary.completion.rate.total",

"Literacy.rate.adult.total",

"Real.interest.rate",

"Population.growth",

"Population.density",

"Population.total",

"Current.health.expenditure.PPP",

"Current.health.expenditure.GDP",

"Unemployment.total",

"GDP.growth",

"GDP.per.capita.PPP",

"Birth.rate.crude",

"Renewable.energy.consumption",

"Adults.newly.infected.HIV",

"safely.managed.drinking.water.services",

"Poverty.headcount.ratio",

"Compulsory.education.duration"

)

#Q.1

attach ( LifeExpectancy)

summary ( LifeExpectancy)

stat.desc( LifeExpectancy[,c("Access.to.electricity","Adjusted.net.national.income","Mortality.rate.infant","Population.density","Current.health.expenditure.PPP", "Adjusted.net.national.income.per.capita", "Birth.rate.crude","Adults.newly.infected.HIV","safely.managed.drinking.water.services" )])

# Summary statistics

#viewing missing data for sample dataset.

library (mice)

library (VIM)

badata <- LifeExpectancy[,4:ncol(LifeExpectancy)]

misbadata <- apply( badata, 2, function(col)sum(is.na(col))/length(col)\*100)

aggr\_plot <- aggr ( badata , col = c('navyblue','grey'), numbers= TRUE, sortVars=TRUE, labels=names(badata),cex.axis=.9,gap=3,

ylab=c("Histogram-Missing Data","Pattern"), dev.new (width=100, height=50))

# checking the retationship b/w the predictor and response variables.

plot ( Life.expectancy.at.birth ~ Access.to.electricity ,pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ safely.managed.drinking.water.services, pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Mortality.rate.infant ,pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Birth.rate.crude ,pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Population.density ,pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Adults.newly.infected.HIV , pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Current.health.expenditure.PPP ,pch = 20, cex = 1.4 )

plot ( Life.expectancy.at.birth ~ Adjusted.net.national.income.per.capita ,pch = 20, cex = 1.4 )

# checking the life expectance base on the recorded response in dataset

hist ( Life.expectancy.at.birth ,xlab = "Life.expectancy" )

qplot ( x=Mortality.rate.infant, y=Birth.rate.crude , geom="boxplot" )

#Q.2

#Identifying Percentage of NA Values in Columns

library(tidyverse)

apply(LifeExpectancy, 2, function(col)sum(is.na(col))/length(col))

#Removing Variables with more than 70% NA Values, Country Code and Continent Column

LifeExpectancy2 = LifeExpectancy[,c(-1:-2,-9,-10,-13,-24,-27)]

view(LifeExpectancy2)

apply(LifeExpectancy2, 2, function(col)sum(is.na(col))/length(col))

#Imputed Median for all NA Values

for(i in 1:ncol(LifeExpectancy2)) {

LifeExpectancy2[ , i][is.na(LifeExpectancy2[ , i])] <- median(LifeExpectancy2[ , i], na.rm = TRUE)

}

view(LifeExpectancy2)

# Imputed Mean of all the Variables

for(i in 1:ncol(LifeExpectancy2)) {

LifeExpectancy2[ , i][is.na(LifeExpectancy2[ , i])] <- mean(LifeExpectancy2[ , i], na.rm = TRUE)

}

#Running linear model with the data having least significant variables just for comparison with the Data in which we have removed least significant variables

lm\_1 <- lm(LifeExpectancy2$Life.expectancy.at.birth ~ ., data = LifeExpectancy2)

summary(lm\_1)

hist(lm\_1$residuals, breaks = 5)

plot(lm\_1,which= 2)

#Q.3

# Collinearity Matrix

library(corrplot)

ColMat <- cor(LifeExpectancy2)

print(ColMat)

ColMat.corr <- cor(LifeExpectancy2)

# Heatmap of Correlation matrix

corrplot(ColMat.corr)

# Removing Least Significant Variables after checking Heatmap

names(LifeExpectancy2)

droping<- c("Adjusted.net.national.income", "Adjusted.net.national.income.per.capita", "Children.newly.infected.HIV", "Children.out.of.school", "Mortality.rate.infant","Primary.completion.rate.total", "Real.interest.rate", "Population.growth", "Population.density", "Population.total", "Current.health.expenditure.GDP", "Unemployment.total", "GDP.growth", "Birth.rate.crude", "Adults.newly.infected.HIV", "Compulsory.education.duration")

LifeExpectancy2 <- LifeExpectancy2[, !(names(LifeExpectancy2) %in% drop)]

view(LifeExpectancy2)

#Applying VIF on the data

vif(LifeExpectancy2)

#Q4

#Running the Linear Model again after removing the least significant variables

lm\_2 <- lm(LifeExpectancy2$Life.expectancy.at.birth ~ ., data = LifeExpectancy2)

summary(lm\_2)

plot(lm\_2,which= 2)

hist(lm\_2$residuals, breaks = 20)

# Compute the analysis of variance

life.aov <- aov(Life.expectancy.at.birth ~ ., data = LifeExpectancy2)

# Summary of the analysis

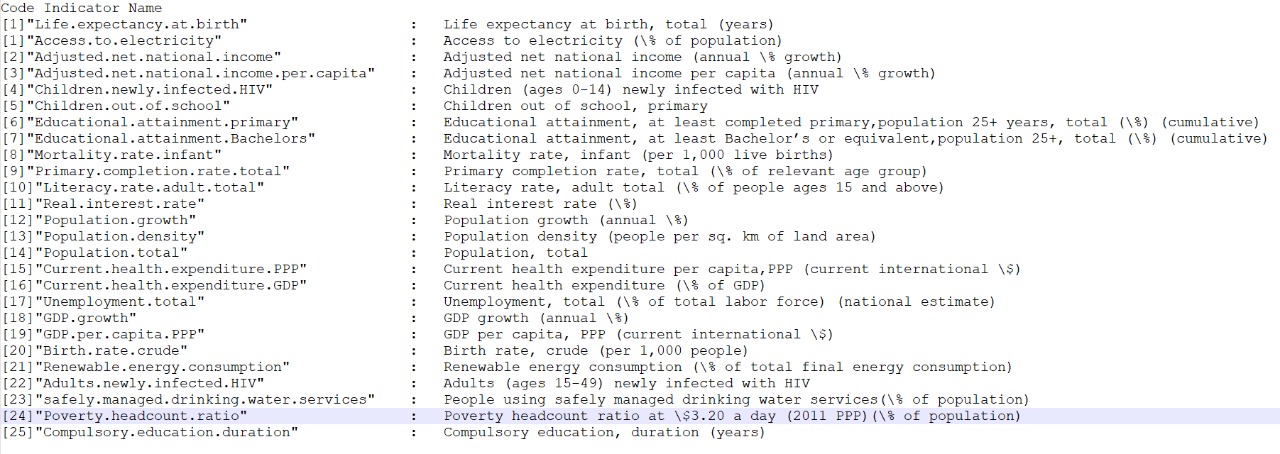
summary(life.aov)

# Summary Function

summary(LifeExpectancy2)

**Appendix II: New Variable Name and Indicator Name**

Below are the new variable name and indicator name of the WDI dataset used in this project:



**Appendix III: Acronyms/Abbreviations**

***Acronym Expanded Term***

IDE Integrated Development Environment

**Contribution**

Bhavya Barbhaya (2111120) - Q4, Report, PPT

Fatma Demir (2111495) - Q3, Report, PPT

Rudra Mehta (2111901) - Q2, Report, PPT

Raktim Ghosh (2110480) – Q1, Report , PPT