A

PROJECT REPORT

ON

Prediction of life Expectancy

Submitted by

Ashishkumar Dobariya(18IT445)
Mayur Jiyani(18IT446)
Ritul Bathani(18IT447)

Subject: Advanced Programming Practices(3IT04)

November, 2020

Information Technology Department

Birla Vishvakarma Mahavidyalaya Engineering College

(An Autonomous Institution)

Vallabh Vidyanagar – 388120

Gujarat, INDIA

Introduction

As we know that life expectancy is a one powerful factor of any country of the word. The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single data-set. On initial visual inspection of the data showed some missing values. As the data-sets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model data-set. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

method and methodology

The design approach involves pre-processing of data, creative feature engineering and the regression model such as Linear regression and Multiple regression.

2.1) Linear Regression:

For finding a relationship between two continuous variables, Linear regression is useful. One variable is predictor or independent, and the other variable is variable response or dependent. It looks for a relationship that is statistical but not deterministic. It is said that the relationship between two variables is deterministic if the other can express one variable accurately.

$Y=MX+A$

where Y is the dependent variable, X is the independent variable. M is the coefficient factor(Slope).

2.2) Multiple Regression:

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression

is: $Y= a + b1x1 + b2x2 + \quad ....+bnxn$

Following is the description of the parameters used:

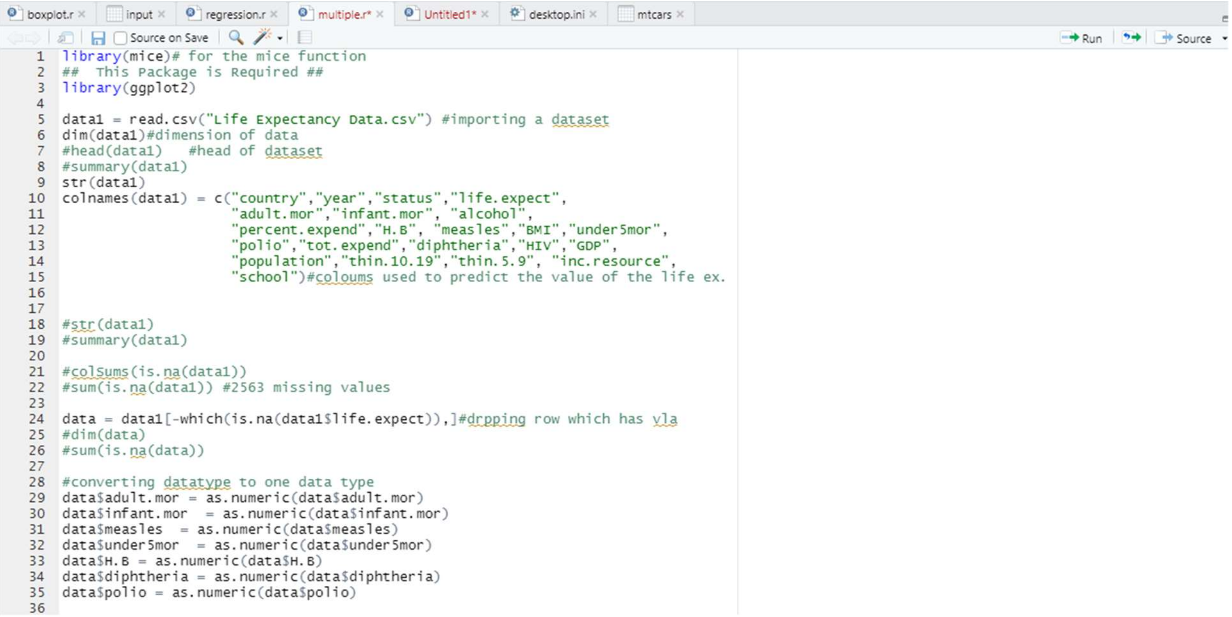
- **Y** is the response variable.
- **a, b1, b2...bn** are the coefficients.
- **x1, x2, ...xn** are the predictor variables.

We create the regression model using the **lm()** function in R. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

Implementation, result and discussion

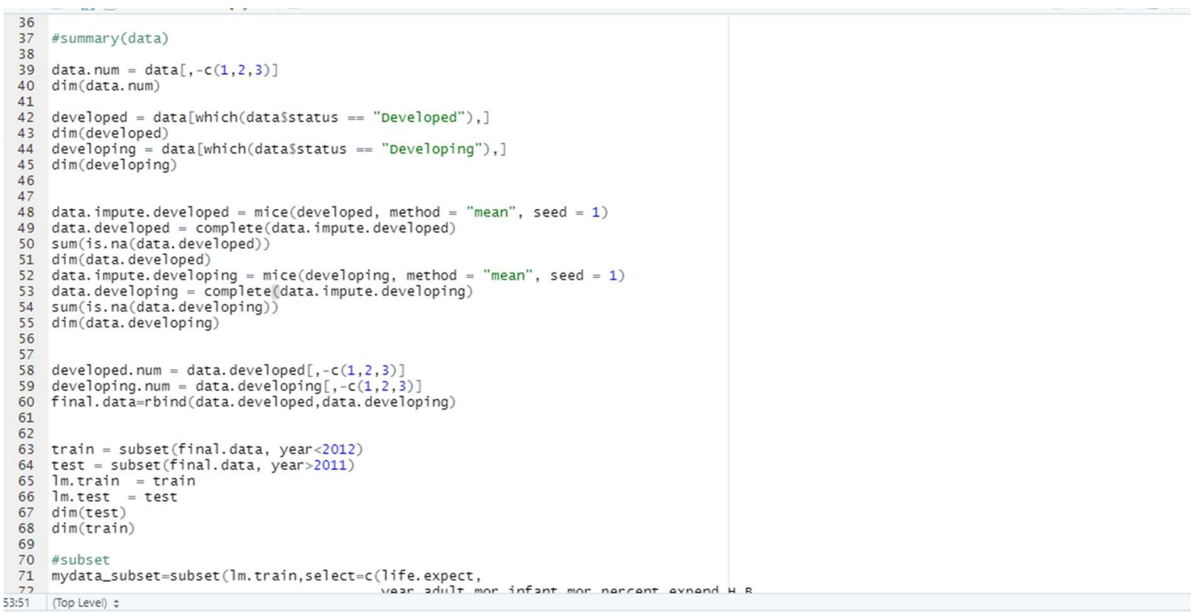
1) dataset importing, data exploration and filling the null value of the data.

we have imported a dataset which is given by the GHO under the WHO, it is in a csv file so we have imported it using a read.csv(path) method;



```
library(mice)# for the mice function
##  This Package is Required ##
library(ggplot2)

data1 = read.csv("Life Expectancy Data.csv") #importing a dataset
dim(data1)#dimension of data
#head(data1)   #head of dataset
#summary(data1)
str(data1)
colnames(data1) = c("country","year","status","life.expect",
                    "adult.mor","infant.mor", "alcohol",
                    "percent.expend","H.B", "measles","BMI","under5mor",
                    "polio","tot.expend","diphtheria","HIV","GDP",
                    "population","thin.10.19","thin.5.9", "inc.resource",
                    "school")#coloums used to predict the value of the life ex.


#str(data1)
#summary(data1)

#colSums(is.na(data1))
#sum(is.na(data1)) #2563 missing values

data = data1[-which(is.na(data1$life.expect)),]#drpping row which has vla
#dim(data)
#sum(is.na(data))

#converting datatype to one data type
data$adult.mor = as.numeric(data$adult.mor)
data$infant.mor  = as.numeric(data$infant.mor)
data$measles  = as.numeric(data$measles)
data$under5mor  = as.numeric(data$under5mor)
data$H.B = as.numeric(data$H.B)
data$diphtheria = as.numeric(data$diphtheria)
data$polio = as.numeric(data$polio)

```

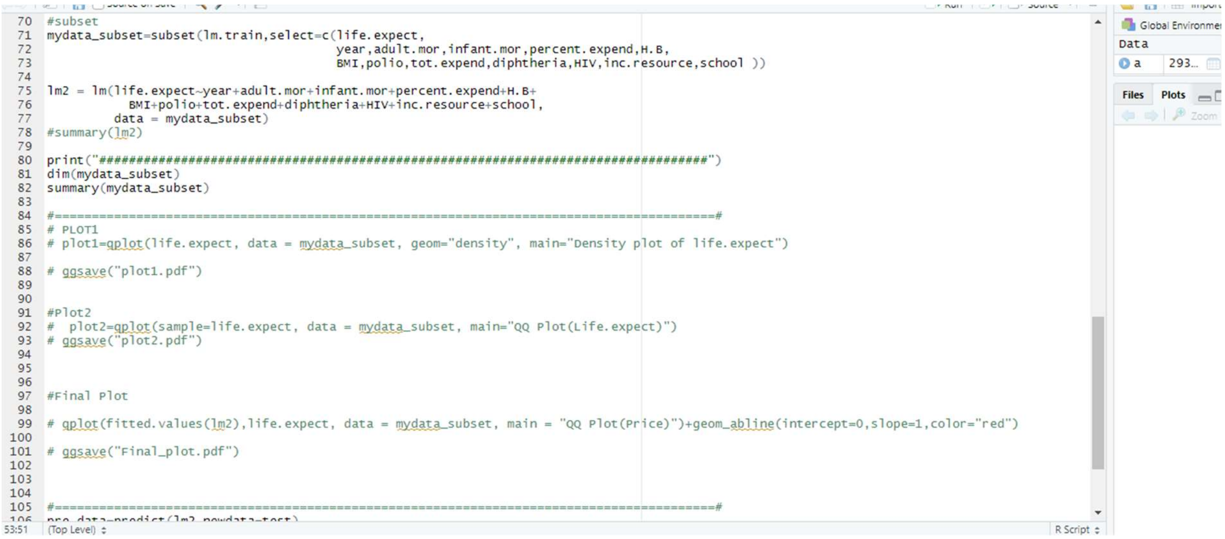
We are plotting a graph using a ggplot2 library. We are using a some of all columns to predict the life expectancy. We are dropping the NULL value of a life expectancy because we are predicting a life expectancy so we are not filling the null value of that column. So we are dropping a row which has a null value of life expectancy. We will have to convert all the data type to a one data type. So we can use than to predict the life expectancy.



```
#summary(data)

data.num = data[,-c(1,2,3)]
dim(data.num)

developed = data[which(data$status == "Developed"),]
dim(developed)
developing = data[which(data$status == "Developing"),]
dim(developing)


data.impute.developed = mice(developed, method = "mean", seed = 1)
data.developed = complete(data.impute.developed)
sum(is.na(data.developed))
dim(data.developed)
data.impute.developing = mice(developing, method = "mean", seed = 1)
data.developing = complete(data.impute.developing)
sum(is.na(data.developing))
dim(data.developing)


developed.num = data.developed[,-c(1,2,3)]
developing.num = data.developing[,-c(1,2,3)]
final.data=rbind(data.developed,data.developing)


train = subset(final.data, year<2012)
test = subset(final.data, year>2011)
lm.train  = train
lm.test  = test
dim(test)
dim(train)

#subset
mydata_subset=subset(lm.train,select=c(life.expect,
                                        year,adult.mor,infant.mor,percent.expend,H.B
```

To handle the null value of the other columns we have divide the dataset in to the developing and developed countries because the all possible parameter of the both type of country cannot be matched. So differentiate them and then fill the values of the null values. We have called a mice function to fill the null value of a whole dataset using mean method. Here we are neglecting first 3 columns of the dataset because first Colum is country and remaining columns are year and status (developed, developing) which we are not using for the prediction.

We have divided the value in train and test dataset in a train based on the year.

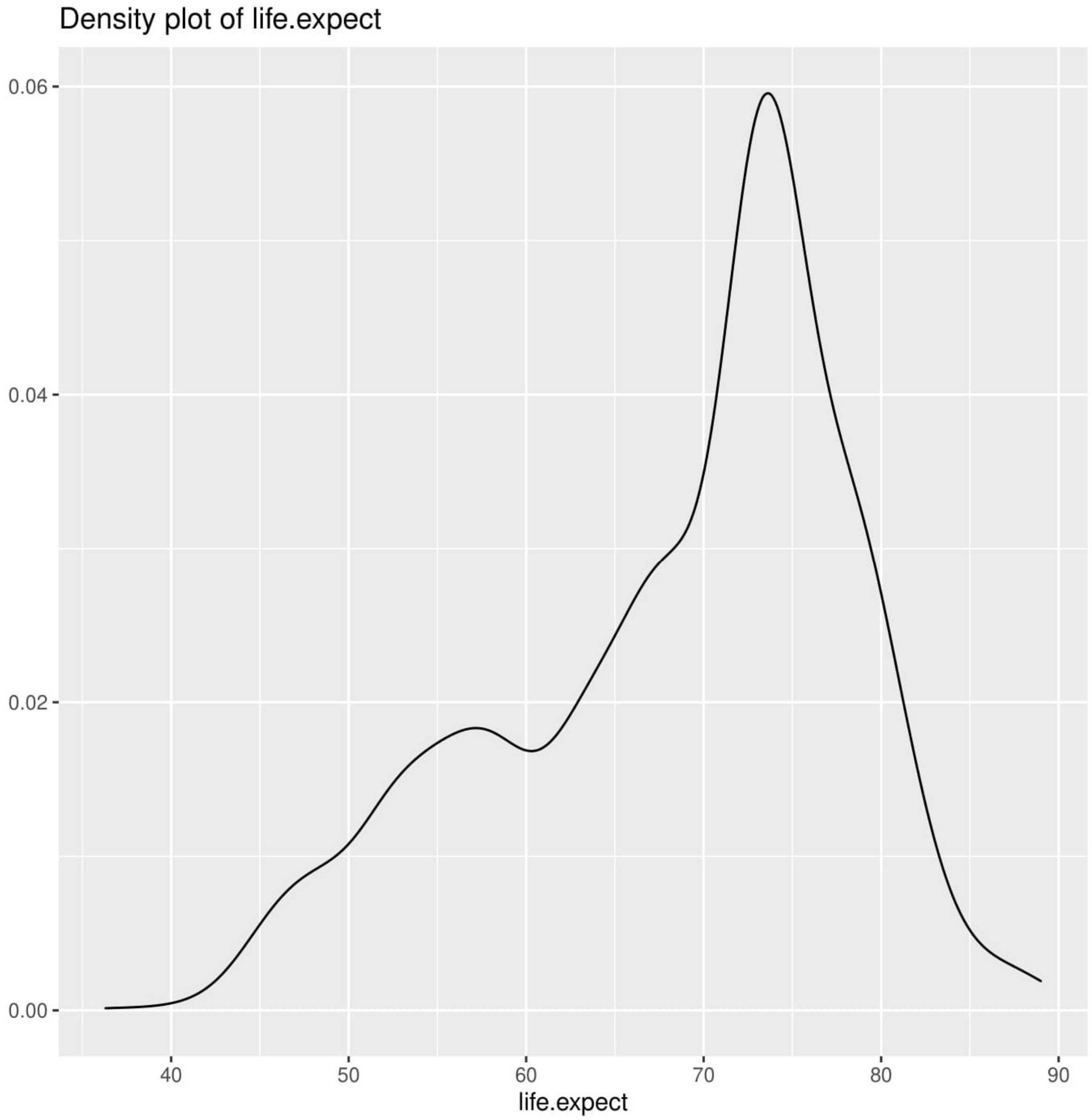


```
#subset
mydata_subset=subset(lm.train,select=c(life.expect,
                                        year,adult.mor,infant.mor,percent.expend,H.B,
                                        BMI,polio,tot.expend,diphtheria,HIV,inc.resource,school ))

lm2 = lm(life.expect~year+adult.mor+infant.mor+percent.expend+H.B+
           BMI+polio+tot.expend+diphtheria+HIV+inc.resource+school,
         data = mydata_subset)
#summary(lm2)

print("##################################################################################")
dim(mydata_subset)
summary(mydata_subset)

#==================================================================================#
# PLOT1
# plot1=qplot(life.expect, data = mydata_subset, geom="density", main="Density plot of life.expect")

# ggsave("plot1.pdf")


#Plot2
#  plot2=qplot(sample=life.expect, data = mydata_subset, main="QQ Plot(Life.expect)")
# ggsave("plot2.pdf")



#Final Plot

# qplot(fitted.values(lm2),life.expect, data = mydata_subset, main = "QQ Plot(Price)")+geom_abline(intercept=0,slope=1,color="red")

# ggsave("Final_plot.pdf")



#==================================================================================#
pre.data=predict(lm2,newdata=test)
```

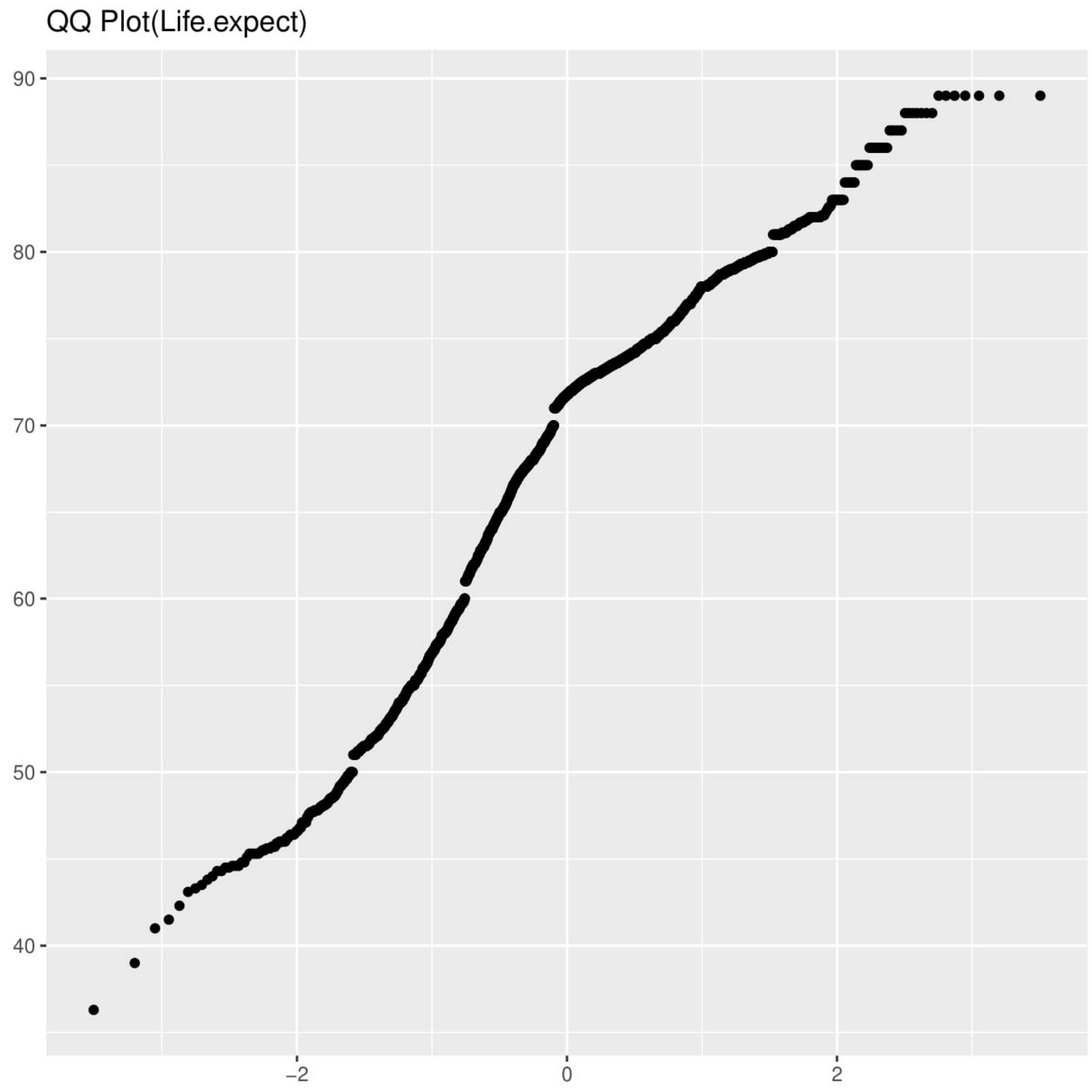
We are building a model using some of the specific columns not all columns.

Then we are printing a graph of the model

1. histogram

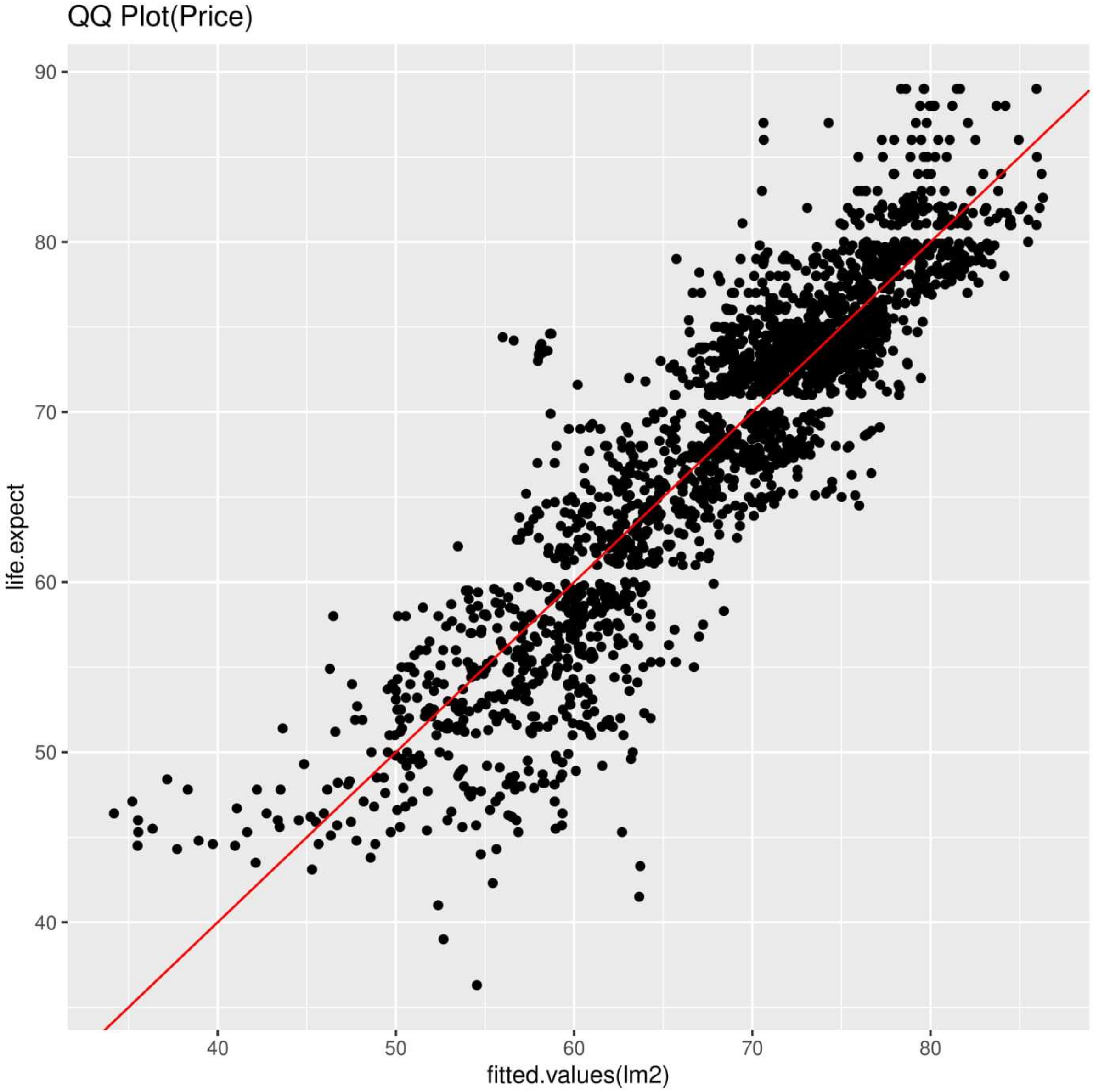


This is the life expectancy histo graph of train dataset values. We can clearly say that graph(data) is not normally distributed. we can also see it using QQ plot or (normal distribution).



Here the graph is in continuous because we have removed life expectancy of the year>2011. Here graph is not equally distributed.

Final graph is scatter plot of the data vs life.expe.



This all graphs are for the train data but likewise we can plot the graph for the test data type.

Conclusion

We can predict the life expectancy using this model of any country by knowing some req. parameter. This project is very imported for the new country to predict whose data is not in this dataset file.

REFERENCES

1) We got the data set from here: https://www.kaggle.com/kumarajarshi/life-expectancy-who
2) https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data
3) https://www.statmethods.net/stats/regression.html
4) https://www.kaggle.com/sivararamakrishnanv/life-expectancy-capstone
5) https://www.rdocumentation.org/packages/mice/versions/3.11.0/topics/mice