R Notebook

**Importing The Libraries**

library(car)

## Warning: package 'car' was built under R version 4.2.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.3

library(e1071)

## Warning: package 'e1071' was built under R version 4.2.3

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.2.3

## Loading required package: Matrix

## Loaded glmnet 4.1-7

library(stats)  
library(utils)  
library(datasets)  
library(ggplot2)  
library(polynom)

## Warning: package 'polynom' was built under R version 4.2.3

**UNDERSTANDING THE DATA**

# Reading The Data set and creating copies for the Data sets   
df = read.csv('Life Expectancy Data.csv')  
  
df\_copy = data.frame(df)   
df\_copy\_1 = subset(df, select = -c(Country,Year,Status))

**To view the Data set**

head(df)

## Country Year Status Life.expectancy Adult.Mortality infant.deaths  
## 1 Afghanistan 2015 Developing 65.0 263 62  
## 2 Afghanistan 2014 Developing 59.9 271 64  
## 3 Afghanistan 2013 Developing 59.9 268 66  
## 4 Afghanistan 2012 Developing 59.5 272 69  
## 5 Afghanistan 2011 Developing 59.2 275 71  
## 6 Afghanistan 2010 Developing 58.8 279 74  
## Alcohol percentage.expenditure Hepatitis.B Measles BMI under.five.deaths  
## 1 0.01 71.279624 65 1154 19.1 83  
## 2 0.01 73.523582 62 492 18.6 86  
## 3 0.01 73.219243 64 430 18.1 89  
## 4 0.01 78.184215 67 2787 17.6 93  
## 5 0.01 7.097109 68 3013 17.2 97  
## 6 0.01 79.679367 66 1989 16.7 102  
## Polio Total.expenditure Diphtheria HIV.AIDS GDP Population  
## 1 6 8.16 65 0.1 584.25921 33736494  
## 2 58 8.18 62 0.1 612.69651 327582  
## 3 62 8.13 64 0.1 631.74498 31731688  
## 4 67 8.52 67 0.1 669.95900 3696958  
## 5 68 7.87 68 0.1 63.53723 2978599  
## 6 66 9.20 66 0.1 553.32894 2883167  
## thinness..1.19.years thinness.5.9.years Income.composition.of.resources  
## 1 17.2 17.3 0.479  
## 2 17.5 17.5 0.476  
## 3 17.7 17.7 0.470  
## 4 17.9 18.0 0.463  
## 5 18.2 18.2 0.454  
## 6 18.4 18.4 0.448  
## Schooling  
## 1 10.1  
## 2 10.0  
## 3 9.9  
## 4 9.8  
## 5 9.5  
## 6 9.2

**Dimensions of the Data set**

dim(df)

## [1] 2938 22

**The Type of Data set**

str(df)

## 'data.frame': 2938 obs. of 22 variables:  
## $ Country : chr "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
## $ Year : int 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...  
## $ Status : chr "Developing" "Developing" "Developing" "Developing" ...  
## $ Life.expectancy : num 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...  
## $ Adult.Mortality : int 263 271 268 272 275 279 281 287 295 295 ...  
## $ infant.deaths : int 62 64 66 69 71 74 77 80 82 84 ...  
## $ Alcohol : num 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...  
## $ percentage.expenditure : num 71.3 73.5 73.2 78.2 7.1 ...  
## $ Hepatitis.B : int 65 62 64 67 68 66 63 64 63 64 ...  
## $ Measles : int 1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...  
## $ BMI : num 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...  
## $ under.five.deaths : int 83 86 89 93 97 102 106 110 113 116 ...  
## $ Polio : int 6 58 62 67 68 66 63 64 63 58 ...  
## $ Total.expenditure : num 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...  
## $ Diphtheria : int 65 62 64 67 68 66 63 64 63 58 ...  
## $ HIV.AIDS : num 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...  
## $ GDP : num 584.3 612.7 631.7 670 63.5 ...  
## $ Population : num 33736494 327582 31731688 3696958 2978599 ...  
## $ thinness..1.19.years : num 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...  
## $ thinness.5.9.years : num 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...  
## $ Income.composition.of.resources: num 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...  
## $ Schooling : num 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...

**Details of the Data set**

summary(df)

## Country Year Status Life.expectancy  
## Length:2938 Min. :2000 Length:2938 Min. :36.30   
## Class :character 1st Qu.:2004 Class :character 1st Qu.:63.10   
## Mode :character Median :2008 Mode :character Median :72.10   
## Mean :2008 Mean :69.22   
## 3rd Qu.:2012 3rd Qu.:75.70   
## Max. :2015 Max. :89.00   
## NA's :10   
## Adult.Mortality infant.deaths Alcohol percentage.expenditure  
## Min. : 1.0 Min. : 0.0 Min. : 0.0100 Min. : 0.000   
## 1st Qu.: 74.0 1st Qu.: 0.0 1st Qu.: 0.8775 1st Qu.: 4.685   
## Median :144.0 Median : 3.0 Median : 3.7550 Median : 64.913   
## Mean :164.8 Mean : 30.3 Mean : 4.6029 Mean : 738.251   
## 3rd Qu.:228.0 3rd Qu.: 22.0 3rd Qu.: 7.7025 3rd Qu.: 441.534   
## Max. :723.0 Max. :1800.0 Max. :17.8700 Max. :19479.912   
## NA's :10 NA's :194   
## Hepatitis.B Measles BMI under.five.deaths  
## Min. : 1.00 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## 1st Qu.:77.00 1st Qu.: 0.0 1st Qu.:19.30 1st Qu.: 0.00   
## Median :92.00 Median : 17.0 Median :43.50 Median : 4.00   
## Mean :80.94 Mean : 2419.6 Mean :38.32 Mean : 42.04   
## 3rd Qu.:97.00 3rd Qu.: 360.2 3rd Qu.:56.20 3rd Qu.: 28.00   
## Max. :99.00 Max. :212183.0 Max. :87.30 Max. :2500.00   
## NA's :553 NA's :34   
## Polio Total.expenditure Diphtheria HIV.AIDS   
## Min. : 3.00 Min. : 0.370 Min. : 2.00 Min. : 0.100   
## 1st Qu.:78.00 1st Qu.: 4.260 1st Qu.:78.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.755 Median :93.00 Median : 0.100   
## Mean :82.55 Mean : 5.938 Mean :82.32 Mean : 1.742   
## 3rd Qu.:97.00 3rd Qu.: 7.492 3rd Qu.:97.00 3rd Qu.: 0.800   
## Max. :99.00 Max. :17.600 Max. :99.00 Max. :50.600   
## NA's :19 NA's :226 NA's :19   
## GDP Population thinness..1.19.years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.10   
## 1st Qu.: 463.94 1st Qu.:1.958e+05 1st Qu.: 1.60   
## Median : 1766.95 Median :1.387e+06 Median : 3.30   
## Mean : 7483.16 Mean :1.275e+07 Mean : 4.84   
## 3rd Qu.: 5910.81 3rd Qu.:7.420e+06 3rd Qu.: 7.20   
## Max. :119172.74 Max. :1.294e+09 Max. :27.70   
## NA's :448 NA's :652 NA's :34   
## thinness.5.9.years Income.composition.of.resources Schooling   
## Min. : 0.10 Min. :0.0000 Min. : 0.00   
## 1st Qu.: 1.50 1st Qu.:0.4930 1st Qu.:10.10   
## Median : 3.30 Median :0.6770 Median :12.30   
## Mean : 4.87 Mean :0.6276 Mean :11.99   
## 3rd Qu.: 7.20 3rd Qu.:0.7790 3rd Qu.:14.30   
## Max. :28.60 Max. :0.9480 Max. :20.70   
## NA's :34 NA's :167 NA's :163

**Names of the Columns of the Dataset**

colnames(df)

## [1] "Country" "Year"   
## [3] "Status" "Life.expectancy"   
## [5] "Adult.Mortality" "infant.deaths"   
## [7] "Alcohol" "percentage.expenditure"   
## [9] "Hepatitis.B" "Measles"   
## [11] "BMI" "under.five.deaths"   
## [13] "Polio" "Total.expenditure"   
## [15] "Diphtheria" "HIV.AIDS"   
## [17] "GDP" "Population"   
## [19] "thinness..1.19.years" "thinness.5.9.years"   
## [21] "Income.composition.of.resources" "Schooling"

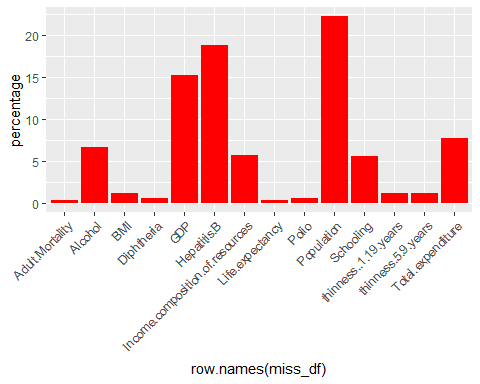
**Missing Values and Handling**

# Finding Out the No of Missing Values in the Datset   
  
miss\_val = sum(is.na(df))  
print(paste0(" The Percentage of missing values in each Data set : ",miss\_val))

## [1] " The Percentage of missing values in each Data set : 2563"

**Graphs Of the Missing Values**

feature <- c()  
percentage <- c()  
for (i in colnames(df)) {  
 miss\_per <- (sum(is.na(df[[i]]))/length(df[[i]]))\*100  
 if (miss\_per > 0) {  
 feature <- c(feature, i)  
 percentage <- c(percentage, miss\_per)  
 }  
}  
  
miss\_df <- data.frame(percentage, row.names=feature)  
library(ggplot2)  
ggplot(miss\_df, aes(x=row.names(miss\_df), y=percentage)) +   
 geom\_bar(stat="identity", fill="red") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



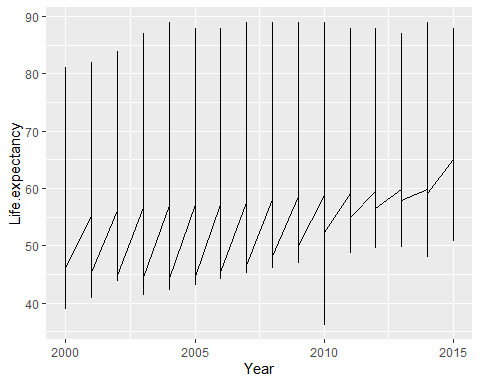
**Find Out the Duplicate values**

df[duplicated(df),]

## [1] Country Year   
## [3] Status Life.expectancy   
## [5] Adult.Mortality infant.deaths   
## [7] Alcohol percentage.expenditure   
## [9] Hepatitis.B Measles   
## [11] BMI under.five.deaths   
## [13] Polio Total.expenditure   
## [15] Diphtheria HIV.AIDS   
## [17] GDP Population   
## [19] thinness..1.19.years thinness.5.9.years   
## [21] Income.composition.of.resources Schooling   
## <0 rows> (or 0-length row.names)

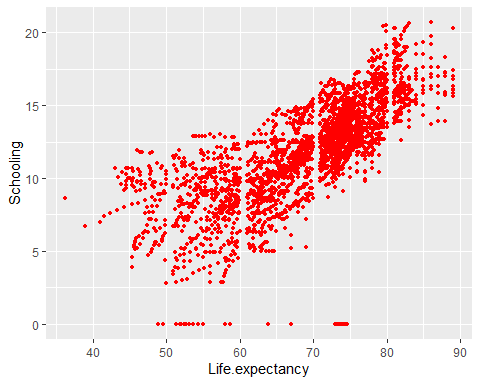
**Line Plot Between year and Life Expectancy**

library(ggplot2)  
  
ggplot(df, aes(x = `Year`, y = `Life.expectancy`)) +  
 geom\_line()



library(ggplot2)  
  
  
ggplot(df, aes(x = `Life.expectancy`, y = Schooling)) +  
 geom\_point(color='red',size =0.9)

## Warning: Removed 170 rows containing missing values (`geom\_point()`).



**Filling The Missing values using Mean**

df[, sapply(df, is.numeric)] <- apply(df[, sapply(df, is.numeric)], 2, function(x) replace(x, is.na(x), mean(x, na.rm = TRUE)))  
  
sum(is.null(df))

## [1] 0

**Pair Plots**

#for (i in 1:ncol(df)) {  
 #for (j in 1:ncol(df)) {  
 #if (i != j) {  
 #plot(df[,i], df[,j], main = paste(colnames(df)[i], " vs ", colnames(df)[j]))  
 #}  
 #}  
#}

# Correalation Matrix and Heat Map :

# Create correlation matrix  
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.2.3

df <- df[, sapply(df, is.numeric)]  
corr <- cor(df)  
  
# Melt correlation matrix  
corr\_melted <- melt(corr)  
  
# Create heatmap  
ggplot(corr\_melted, aes(x=Var1, y=Var2)) +  
 geom\_tile(aes(fill=value)) +  
 scale\_fill\_gradient(low="white", high="green") +  
 geom\_text(aes(label=round(value, 2)), color="black") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

Table, Excel

Description automatically generated

sum(is.na(df))

## [1] 0

sum(is.na(df\_copy))

## [1] 2563

**Filling The Missing Values With Median Values :**

df\_copy[, sapply(df\_copy, is.numeric)] <- apply(df\_copy[, sapply(df\_copy, is.numeric)], 2, function(x) replace(x, is.na(x), median(x, na.rm = TRUE)))

sum(is.na(df\_copy))

## [1] 0

**Filling the Missing value with Interpolation :**

for (col in colnames(df\_copy\_1)) {  
 df\_copy\_1[[col]] <- approx(seq\_along(df\_copy\_1[[col]])[!is.na(df\_copy\_1[[col]])], df\_copy\_1[[col]][!is.na(df\_copy\_1[[col]])], seq\_along(df\_copy\_1[[col]]))$y  
}

**Skewness Of the Dataset**

skewness\_1 <- function(x) {  
 n <- length(x)  
 m3 <- sum((x - mean(x))^3)/(n-1)  
 s3 <- sd(x)^3  
 skewness <- m3/s3  
 return(skewness)  
}  
  
for (col in names(df\_copy\_1)) {  
 if(is.numeric(df\_copy\_1[[col]])){  
 print(paste(col, "-:", skewness\_1(df\_copy\_1[[col]])))  
 }  
}

## [1] "Life.expectancy -: -0.635856669158128"  
## [1] "Adult.Mortality -: 1.17172589519657"  
## [1] "infant.deaths -: 9.78030062030323"  
## [1] "Alcohol -: 0.609030000411964"  
## [1] "percentage.expenditure -: 4.64888453275825"  
## [1] "Hepatitis.B -: -1.58984428227641"  
## [1] "Measles -: 9.43490490012689"  
## [1] "BMI -: -0.217067070312807"  
## [1] "under.five.deaths -: 9.48860103228038"  
## [1] "Polio -: -2.08628016732132"  
## [1] "Total.expenditure -: 0.619390409871752"  
## [1] "Diphtheria -: -2.06115278591113"  
## [1] "HIV.AIDS -: 5.39243871874848"  
## [1] "GDP -: 3.30244457730235"  
## [1] "Population -: 17.22294470789"  
## [1] "thinness..1.19.years -: 1.67634269718188"  
## [1] "thinness.5.9.years -: 1.7384013998031"  
## [1] "Income.composition.of.resources -: -1.16698627027287"  
## [1] "Schooling -: -0.616449433074087"

**Normalizing the Skewed Datasets**

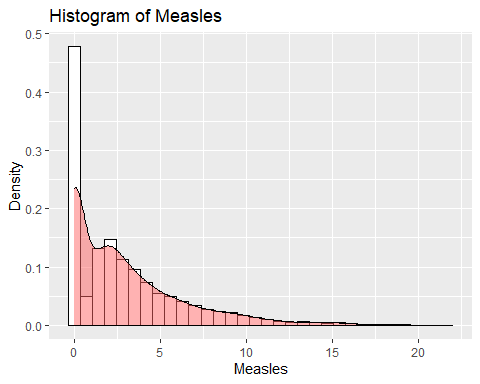
Measles\_sqrt <- sqrt(sqrt(df\_copy\_1$Measles))  
skewness(Measles\_sqrt)

## [1] 1.658537

df\_copy\_1$Measles <- Measles\_sqrt

ggplot(df\_copy\_1, aes(x=Measles)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.7, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of Measles", x="Measles", y="Density")

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.

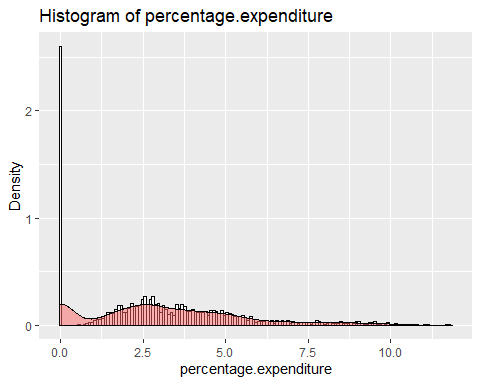


Percentage\_expe\_sqrt <- sqrt(sqrt(df\_copy\_1$percentage.expenditure))  
skewness(Percentage\_expe\_sqrt)

## [1] 0.7289899

df\_copy\_1$percentage.expenditure <- Percentage\_expe\_sqrt

ggplot(df\_copy\_1, aes(x=percentage.expenditure)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.08, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of percentage.expenditure", x="percentage.expenditure", y="Density")

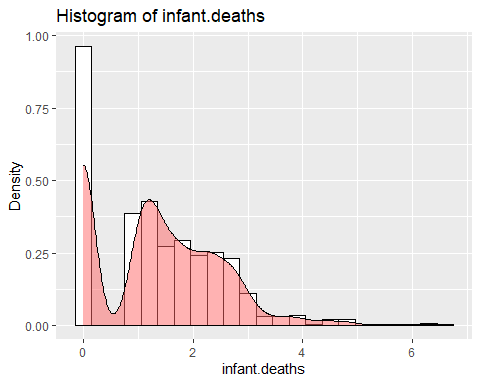


infant.deaths\_sqrt <- sqrt(sqrt(df\_copy\_1$infant.deaths))  
skewness(infant.deaths\_sqrt)

## [1] 0.6797275

df\_copy\_1$infant.deaths <- infant.deaths\_sqrt

ggplot(df\_copy\_1, aes(x=infant.deaths)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.3, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of infant.deaths", x="infant.deaths", y="Density")

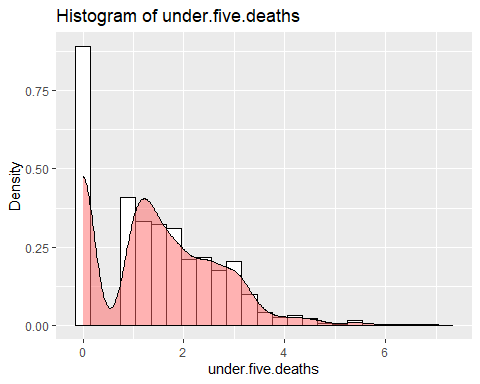


under.five.deaths\_sqrt <- sqrt(sqrt(df\_copy\_1$under.five.deaths))  
skewness(under.five.deaths\_sqrt)

## [1] 0.7276725

df\_copy\_1$under.five.deaths <- under.five.deaths\_sqrt

ggplot(df\_copy\_1, aes(x=under.five.deaths)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.3, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of under.five.deaths", x="under.five.deaths", y="Density")



Polio\_sqrt <- sqrt(sqrt(df\_copy\_1$Polio ))  
skewness(Polio\_sqrt)

## [1] -2.957654

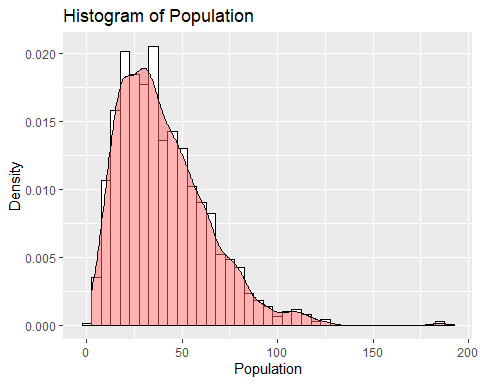
df\_copy\_1$Polio <- Polio\_sqrt

Population\_sqrt <- sqrt(sqrt(df\_copy\_1$Population))  
skewness(Population\_sqrt)

## [1] 1.192727

df\_copy\_1$Population <- Population\_sqrt

ggplot(df\_copy\_1, aes(x=Population)) +   
 geom\_histogram(aes(y=..density..), binwidth=5, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of Population", x="Population", y="Density")

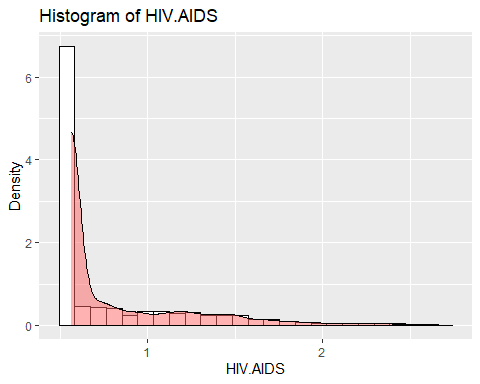


HIV.AIDS\_sqrt <- sqrt(sqrt(df\_copy\_1$HIV.AIDS))  
skewness(HIV.AIDS\_sqrt)

## [1] 1.884533

df\_copy\_1$HIV.AIDS <- HIV.AIDS\_sqrt

ggplot(df\_copy\_1, aes(x=HIV.AIDS)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.09, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of HIV.AIDS", x="HIV.AIDS", y="Density")

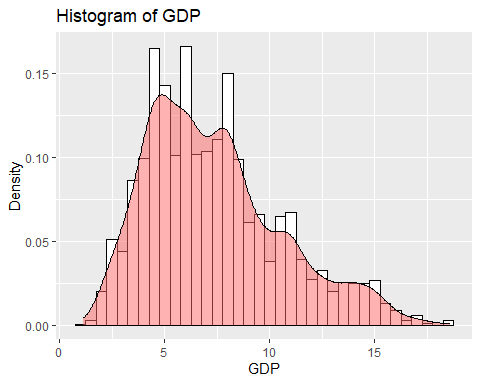


GDP\_sqrt <- sqrt(sqrt(df\_copy\_1$GDP))  
skewness(GDP\_sqrt)

## [1] 0.7475398

df\_copy\_1$GDP <- GDP\_sqrt

ggplot(df\_copy\_1, aes(x=GDP)) +   
 geom\_histogram(aes(y=..density..), binwidth=0.5, color="black", fill="white") +   
 geom\_density(alpha=0.5, fill="#FF6666") +  
 labs(title="Histogram of GDP", x="GDP", y="Density")



**Skewness of the Datasets after Normalizing**

skewness\_1 <- function(x) {  
 n <- length(x)  
 m3 <- sum((x - mean(x))^3)/(n-1)  
 s3 <- sd(x)^3  
 skewness <- m3/s3  
 return(skewness)  
}  
  
for (col in names(df\_copy\_1)) {  
 if(is.numeric(df\_copy\_1[[col]])){  
 print(paste(col, "is:", skewness\_1(df\_copy\_1[[col]])))  
 }  
}

## [1] "Life.expectancy is: -0.635856669158128"  
## [1] "Adult.Mortality is: 1.17172589519657"  
## [1] "infant.deaths is: 0.679958962571999"  
## [1] "Alcohol is: 0.609030000411964"  
## [1] "percentage.expenditure is: 0.729238115659287"  
## [1] "Hepatitis.B is: -1.58984428227641"  
## [1] "Measles is: 1.65910196085473"  
## [1] "BMI is: -0.217067070312807"  
## [1] "under.five.deaths is: 0.727920277782387"  
## [1] "Polio is: -2.95866101548017"  
## [1] "Total.expenditure is: 0.619390409871752"  
## [1] "Diphtheria is: -2.06115278591113"  
## [1] "HIV.AIDS is: 1.88517466390931"  
## [1] "GDP is: 0.747794352536623"  
## [1] "Population is: 1.19313265642152"  
## [1] "thinness..1.19.years is: 1.67634269718188"  
## [1] "thinness.5.9.years is: 1.7384013998031"  
## [1] "Income.composition.of.resources is: -1.16698627027287"  
## [1] "Schooling is: -0.616449433074087"

**Feature Selection**

**Correlation Method**

correlations <- cor(df\_copy\_1)  
sorted\_correlations <- sort(correlations['Life.expectancy',], decreasing=TRUE)  
  
for (i in 1:length(sorted\_correlations)) {  
 print(paste(names(sorted\_correlations)[i], ':', sorted\_correlations[i]))  
}

## [1] "Life.expectancy : 1"  
## [1] "Schooling : 0.704294298045188"  
## [1] "Income.composition.of.resources : 0.690883034766278"  
## [1] "BMI : 0.556433178433153"  
## [1] "GDP : 0.54457339936761"  
## [1] "Diphtheria : 0.474084429604973"  
## [1] "percentage.expenditure : 0.423768265757413"  
## [1] "Alcohol : 0.400318941908566"  
## [1] "Polio : 0.358585307880295"  
## [1] "Hepatitis.B : 0.248073596079946"  
## [1] "Total.expenditure : 0.210744089883623"  
## [1] "Population : -0.014133660075424"  
## [1] "Measles : -0.315516631175639"  
## [1] "thinness.5.9.years : -0.472484735902466"  
## [1] "thinness..1.19.years : -0.476808562984325"  
## [1] "infant.deaths : -0.55954857333208"  
## [1] "under.five.deaths : -0.58507078777411"  
## [1] "Adult.Mortality : -0.695730283490939"  
## [1] "HIV.AIDS : -0.786991201010715"

library(ggplot2)  
  
ggplot(data = reshape2::melt(cor(df\_copy\_1)), aes(x = Var1, y = Var2, fill = value)) +  
 geom\_tile(aes(fill=value)) +  
 scale\_fill\_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +  
 geom\_text(aes(label=round(value, 2)), color="black") +  
   
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.01, hjust=0.01)) +  
 labs(title="Correlation Heatmap") +   
 theme(plot.title = element\_text(size=20),  
 axis.text.y = element\_text(size= 10),  
 axis.text.x = element\_text(size= 10),  
 legend.text = element\_text(size= 10),  
 legend.title = element\_text(size=10))

Chart, treemap chart

Description automatically generated

**Lasso Regression Method**

data\_n = df\_copy\_1  
trainIndex <- sample(nrow(data\_n), floor(0.75 \* nrow(data\_n)), replace = FALSE)  
train <- data\_n[trainIndex, ]  
  
test <- data\_n[-trainIndex, ]  
# Feature selection using Lasso regression  
x\_1 <- as.matrix(train[,2:ncol(train)])  
y\_1 <- train[,1]  
  
lassoModel <- cv.glmnet(x\_1, y\_1, alpha = 0.5)  
coef(lassoModel)

## 19 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 65.8659927721  
## Adult.Mortality -0.0124957503  
## infant.deaths .   
## Alcohol 0.0872774984  
## percentage.expenditure 0.2718374842  
## Hepatitis.B .   
## Measles .   
## BMI 0.0130905775  
## under.five.deaths -0.8913906577  
## Polio 0.1150745966  
## Total.expenditure 0.0460016441  
## Diphtheria 0.0299532821  
## HIV.AIDS -9.2602190100  
## GDP 0.0481875650  
## Population 0.0035676225  
## thinness..1.19.years -0.0275805853  
## thinness.5.9.years -0.0004712236  
## Income.composition.of.resources 6.7240467657  
## Schooling 0.3977962724

**Anova Feature Slection score method**

X <- df\_copy\_1[, -1]  
y <- df\_copy\_1[, 1]  
  
  
fs <- apply(X, 2, function(x) summary(aov(y ~ x))[[1]][1, 4])  
  
features\_map <- data.frame(feature = names(fs), score = fs)  
features\_map <- features\_map[order(features\_map$score, decreasing = TRUE), ]  
  
print("The feature scores generated using ANOVA method are: ")

## [1] "The feature scores generated using ANOVA method are: "

for (i in 1:nrow(features\_map)) {  
 print(paste0("Feature ", i, ": ", features\_map$feature[i], " - ", features\_map$score[i]))  
}

## [1] "Feature 1: HIV.AIDS - 4777.22665632284"  
## [1] "Feature 2: Schooling - 2889.74889319136"  
## [1] "Feature 3: Adult.Mortality - 2754.37051311568"  
## [1] "Feature 4: Income.composition.of.resources - 2681.19684779012"  
## [1] "Feature 5: under.five.deaths - 1528.09447948579"  
## [1] "Feature 6: infant.deaths - 1338.24216681776"  
## [1] "Feature 7: BMI - 1316.71733393934"  
## [1] "Feature 8: GDP - 1237.7757047412"  
## [1] "Feature 9: thinness..1.19.years - 863.891726115104"  
## [1] "Feature 10: Diphtheria - 851.194967910277"  
## [1] "Feature 11: thinness.5.9.years - 843.812169338012"  
## [1] "Feature 12: percentage.expenditure - 642.652842179943"  
## [1] "Feature 13: Alcohol - 560.300534606921"  
## [1] "Feature 14: Polio - 433.226702343313"  
## [1] "Feature 15: Measles - 324.594622322818"  
## [1] "Feature 16: Hepatitis.B - 192.531415987745"  
## [1] "Feature 17: Total.expenditure - 136.457263899143"  
## [1] "Feature 18: Population - 0.586613561295334"

**Finalizing the Data set and Reduce the Data sets with the Potential Predictors**

final\_df =df\_copy\_1[,c( "Life.expectancy","Schooling","HIV.AIDS","Adult.Mortality","infant.deaths","Alcohol",  
 "Measles","Total.expenditure","BMI","GDP","Income.composition.of.resources")]  
poly\_df = final\_df  
head(final\_df)

## Life.expectancy Schooling HIV.AIDS Adult.Mortality infant.deaths Alcohol  
## 1 65.0 10.1 0.5623413 263 2.806066 0.01  
## 2 59.9 10.0 0.5623413 271 2.828427 0.01  
## 3 59.9 9.9 0.5623413 268 2.850270 0.01  
## 4 59.5 9.8 0.5623413 272 2.882121 0.01  
## 5 59.2 9.5 0.5623413 275 2.902783 0.01  
## 6 58.8 9.2 0.5623413 279 2.932972 0.01  
## Measles Total.expenditure BMI GDP Income.composition.of.resources  
## 1 5.828428 8.16 19.1 4.916447 0.479  
## 2 4.709679 8.18 18.6 4.975209 0.476  
## 3 4.553728 8.13 18.1 5.013436 0.470  
## 4 7.265813 8.52 17.6 5.087589 0.463  
## 5 7.408833 7.87 17.2 2.823300 0.454  
## 6 6.678189 9.20 16.7 4.850046 0.448

sum(is.na(final\_df))

## [1] 0

str(final\_df)

## 'data.frame': 2938 obs. of 11 variables:  
## $ Life.expectancy : num 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...  
## $ Schooling : num 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...  
## $ HIV.AIDS : num 0.562 0.562 0.562 0.562 0.562 ...  
## $ Adult.Mortality : num 263 271 268 272 275 279 281 287 295 295 ...  
## $ infant.deaths : num 2.81 2.83 2.85 2.88 2.9 ...  
## $ Alcohol : num 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...  
## $ Measles : num 5.83 4.71 4.55 7.27 7.41 ...  
## $ Total.expenditure : num 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...  
## $ BMI : num 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...  
## $ GDP : num 4.92 4.98 5.01 5.09 2.82 ...  
## $ Income.composition.of.resources: num 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...

**Spliiting The Dataset into Traing and Testing set**

trainIndex <- sample(seq\_len(nrow(final\_df)), floor(0.78 \* nrow(final\_df)))  
train <- final\_df[trainIndex, ]  
test <- final\_df[-trainIndex, ]  
  
# fit linear regression model  
lm.fit <- lm(Life.expectancy ~ ., data = train)  
  
  
str(train[,-1])

## 'data.frame': 2291 obs. of 10 variables:  
## $ Schooling : num 11.7 6.6 11.2 15.3 12.6 14.6 9.6 12.5 11.7 7.2 ...  
## $ HIV.AIDS : num 0.562 1.707 1.174 0.562 1.218 ...  
## $ Adult.Mortality : num 145 426 237 162 256 153 198 78 197 217 ...  
## $ infant.deaths : num 1.19 2.41 1 0 1.19 ...  
## $ Alcohol : num 6.23 7.13 7.84 11.46 1.82 ...  
## $ Measles : num 0 6.77 0 1 0 ...  
## $ Total.expenditure : num 8.44 4.22 5.73 7.55 6.66 ...  
## $ BMI : num 54 13.2 38.4 61.1 37.9 53.1 31.3 28.5 33.6 26.1 ...  
## $ GDP : num 4.15 3.83 3.47 10.67 8.99 ...  
## $ Income.composition.of.resources: num 0.642 0.318 0.613 0.816 0.698 0.777 0.529 0.81 0.709 0.475 ...

sum(is.na(train))

## [1] 0

# fit polynomial regression model  
  
summary(lm.fit)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0885 -1.9341 0.0628 2.2848 11.6083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.3318701 0.5612783 119.962 < 2e-16 \*\*\*  
## Schooling 0.4304373 0.0449522 9.575 < 2e-16 \*\*\*  
## HIV.AIDS -9.7259459 0.2766186 -35.160 < 2e-16 \*\*\*  
## Adult.Mortality -0.0146680 0.0008863 -16.549 < 2e-16 \*\*\*  
## infant.deaths -0.9396443 0.1078167 -8.715 < 2e-16 \*\*\*  
## Alcohol 0.0985627 0.0250358 3.937 8.50e-05 \*\*\*  
## Measles -0.0283901 0.0302986 -0.937 0.348851   
## Total.expenditure 0.1440241 0.0345717 4.166 3.22e-05 \*\*\*  
## BMI 0.0170054 0.0050447 3.371 0.000762 \*\*\*  
## GDP 0.1856304 0.0320650 5.789 8.05e-09 \*\*\*  
## Income.composition.of.resources 7.8239069 0.6967680 11.229 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.815 on 2280 degrees of freedom  
## Multiple R-squared: 0.8383, Adjusted R-squared: 0.8375   
## F-statistic: 1182 on 10 and 2280 DF, p-value: < 2.2e-16

\*\*Fit the Model

lm.fit

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) Schooling   
## 67.33187 0.43044   
## HIV.AIDS Adult.Mortality   
## -9.72595 -0.01467   
## infant.deaths Alcohol   
## -0.93964 0.09856   
## Measles Total.expenditure   
## -0.02839 0.14402   
## BMI GDP   
## 0.01701 0.18563   
## Income.composition.of.resources   
## 7.82391

\*\*Vif score

vif(lm.fit)

## Schooling HIV.AIDS   
## 3.515640 2.033889   
## Adult.Mortality infant.deaths   
## 1.894465 2.371736   
## Alcohol Measles   
## 1.562835 1.772043   
## Total.expenditure BMI   
## 1.138222 1.600751   
## GDP Income.composition.of.resources   
## 1.688509 3.229370

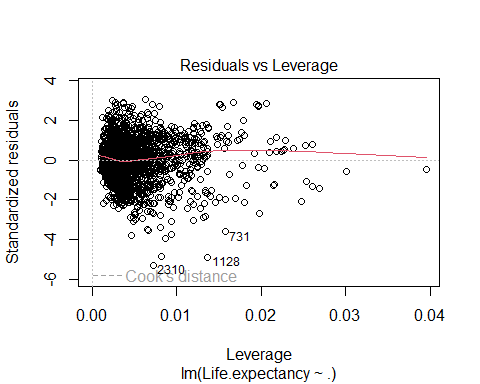
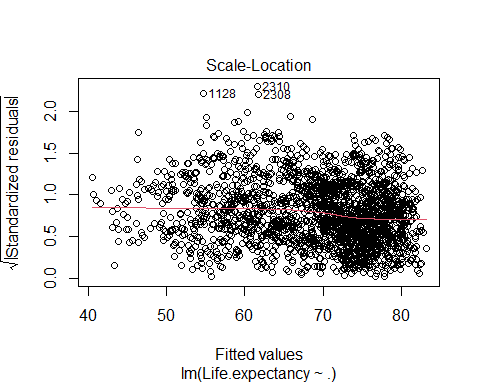
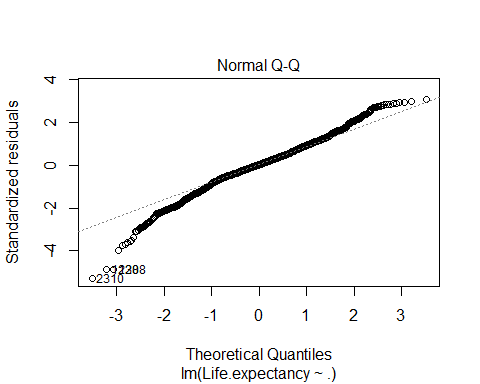
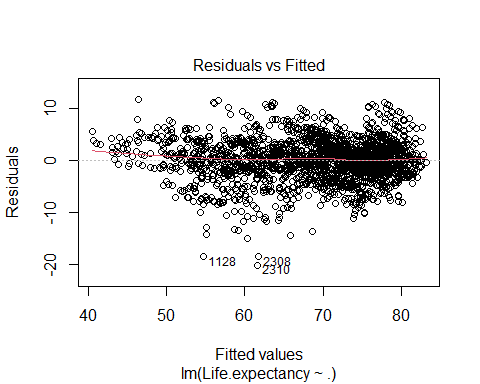
\*\*Anova Test

fit.aov <- anova(lm.fit)  
  
tab <- as.table(cbind(  
 'SS' = c("SSR(x1, x2, x1:X2)" = sum(fit.aov[1:3, 2]),  
 "SSR(X1)" = fit.aov[1, 2],  
 "SSR(X2|X1)" = fit.aov[2, 2],  
 "SSR(X1:X2|X1,X2)" = fit.aov[3, 2],  
 "SSE(X1,X2,X1:X2)" = fit.aov[4, 2],  
 "Total" = sum(fit.aov[, 2])),  
  
 'Df' = c( sum(fit.aov[1:3, 1]),  
 fit.aov[1, 1],  
 fit.aov[2, 1],  
 fit.aov[3, 1],  
 fit.aov[4, 1],  
 sum(fit.aov$Df)),  
  
 'MS' = c( sum(fit.aov[1:3, 2]) / sum(fit.aov[1:3, 1]),  
 fit.aov[1, 3],  
 fit.aov[2, 3],  
 fit.aov[3, 3],  
 fit.aov[4, 3],  
 NA)  
))  
  
round(tab, 2)

## SS Df MS  
## SSR(x1, x2, x1:X2) 164571.33 3.00 54857.11  
## SSR(X1) 102968.48 1.00 102968.48  
## SSR(X2|X1) 55583.57 1.00 55583.57  
## SSR(X1:X2|X1,X2) 6019.27 1.00 6019.27  
## SSE(X1,X2,X1:X2) 3181.43 1.00 3181.43  
## Total 205114.49 2290.00

**Multiple Regression Plots**

plot(lm.fit)



**R-Squared**

# calculate r-squared values for both models  
lm.r\_squared <- summary(lm.fit)$r.squared  
print(paste0("Linear regression r-squared: ", lm.r\_squared))

## [1] "Linear regression r-squared: 0.838259166402846"

*Post processing* **Further Reducing the Model**

mod\_df <- subset(final\_df, select = -c(Schooling, infant.deaths, Income.composition.of.resources))  
head(mod\_df)

## Life.expectancy HIV.AIDS Adult.Mortality Alcohol Measles Total.expenditure  
## 1 65.0 0.5623413 263 0.01 5.828428 8.16  
## 2 59.9 0.5623413 271 0.01 4.709679 8.18  
## 3 59.9 0.5623413 268 0.01 4.553728 8.13  
## 4 59.5 0.5623413 272 0.01 7.265813 8.52  
## 5 59.2 0.5623413 275 0.01 7.408833 7.87  
## 6 58.8 0.5623413 279 0.01 6.678189 9.20  
## BMI GDP  
## 1 19.1 4.916447  
## 2 18.6 4.975209  
## 3 18.1 5.013436  
## 4 17.6 5.087589  
## 5 17.2 2.823300  
## 6 16.7 4.850046

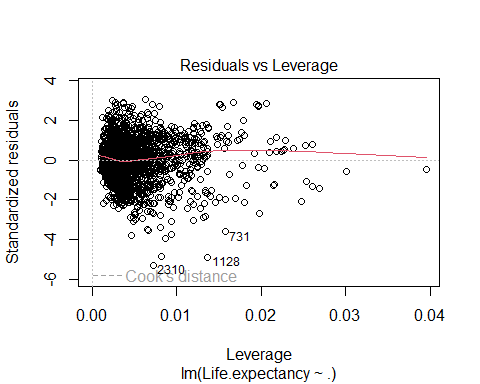
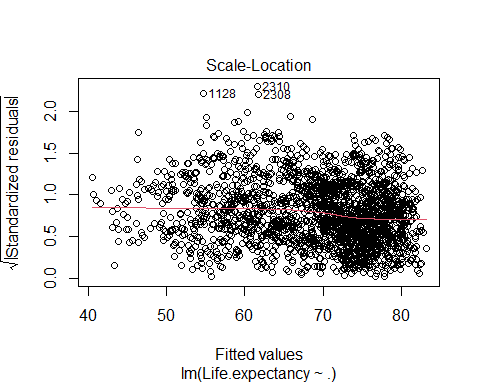
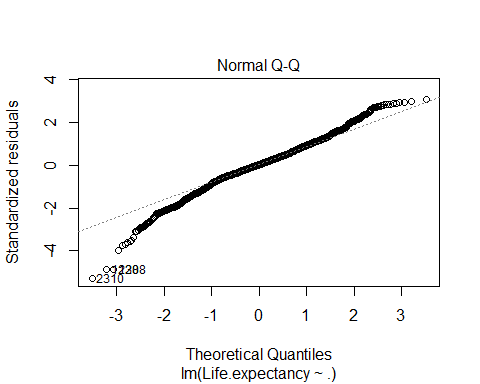
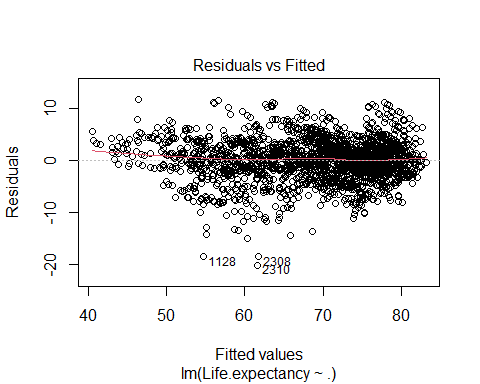
lm.fit\_mod <- lm(Life.expectancy ~ ., data = mod\_df)  
  
summary(lm.fit\_mod)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = mod\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.1915 -2.5818 0.1761 2.6791 14.2115   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.376e+01 4.176e-01 176.621 < 2e-16 \*\*\*  
## HIV.AIDS -1.108e+01 2.652e-01 -41.779 < 2e-16 \*\*\*  
## Adult.Mortality -1.818e-02 8.631e-04 -21.062 < 2e-16 \*\*\*  
## Alcohol 3.427e-01 2.331e-02 14.702 < 2e-16 \*\*\*  
## Measles -2.267e-01 2.469e-02 -9.184 < 2e-16 \*\*\*  
## Total.expenditure 1.470e-01 3.486e-02 4.216 2.56e-05 \*\*\*  
## BMI 4.895e-02 4.930e-03 9.931 < 2e-16 \*\*\*  
## GDP 5.143e-01 2.954e-02 17.413 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.349 on 2930 degrees of freedom  
## Multiple R-squared: 0.7914, Adjusted R-squared: 0.7909   
## F-statistic: 1588 on 7 and 2930 DF, p-value: < 2.2e-16

lm.r\_squared <- summary(lm.fit)$r.squared  
print(paste0("Linear regression r-squared: ", lm.r\_squared))

## [1] "Linear regression r-squared: 0.838259166402846"

plot(lm.fit)



**Polynomial Regression**

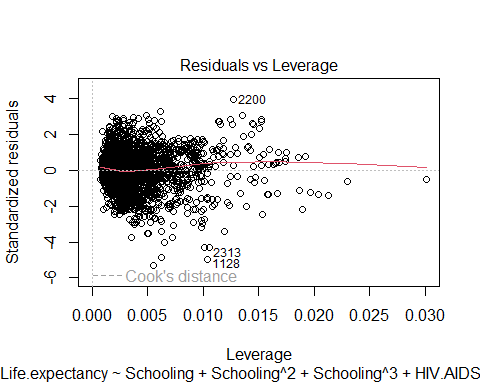
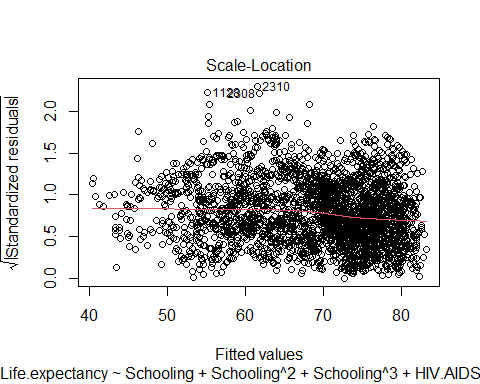
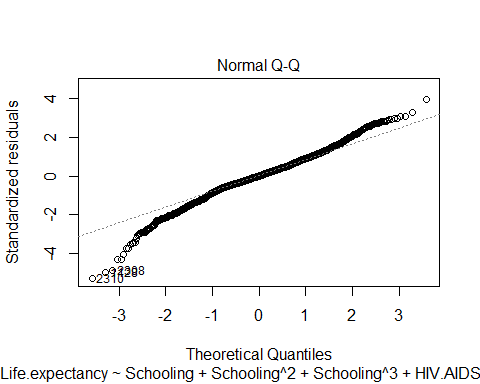
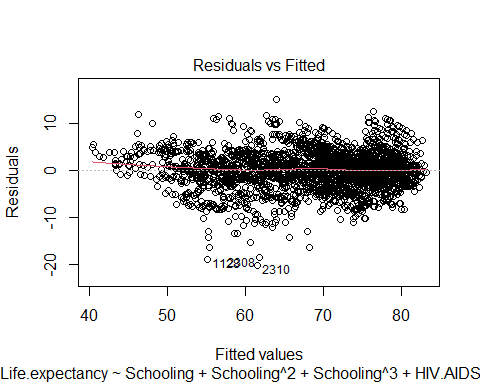
poly\_reg = lm(formula = Life.expectancy ~ Schooling + Schooling^2 + Schooling^3 + HIV.AIDS + HIV.AIDS^2 + HIV.AIDS^3 + Adult.Mortality + Adult.Mortality^2 + Adult.Mortality^3 + infant.deaths + infant.deaths^2 + infant.deaths^3 + Alcohol + Alcohol^2 + Alcohol^3 + Measles + Measles^2 + Measles^3 + Total.expenditure + Total.expenditure^2 + Total.expenditure^3 + BMI + BMI^2 +BMI^3 + GDP + GDP^2 + GDP^3 + Income.composition.of.resources + Income.composition.of.resources^2 + Income.composition.of.resources^3,data = poly\_df)  
  
summary(poly\_reg)

##   
## Call:  
## lm(formula = Life.expectancy ~ Schooling + Schooling^2 + Schooling^3 +   
## HIV.AIDS + HIV.AIDS^2 + HIV.AIDS^3 + Adult.Mortality + Adult.Mortality^2 +   
## Adult.Mortality^3 + infant.deaths + infant.deaths^2 + infant.deaths^3 +   
## Alcohol + Alcohol^2 + Alcohol^3 + Measles + Measles^2 + Measles^3 +   
## Total.expenditure + Total.expenditure^2 + Total.expenditure^3 +   
## BMI + BMI^2 + BMI^3 + GDP + GDP^2 + GDP^3 + Income.composition.of.resources +   
## Income.composition.of.resources^2 + Income.composition.of.resources^3,   
## data = poly\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0988 -1.9763 0.0518 2.2318 14.9758   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.5842926 0.4867453 138.849 < 2e-16 \*\*\*  
## Schooling 0.4169463 0.0389141 10.715 < 2e-16 \*\*\*  
## HIV.AIDS -9.6659788 0.2371008 -40.767 < 2e-16 \*\*\*  
## Adult.Mortality -0.0144805 0.0007674 -18.870 < 2e-16 \*\*\*  
## infant.deaths -0.9453013 0.0942088 -10.034 < 2e-16 \*\*\*  
## Alcohol 0.1233136 0.0218041 5.656 1.70e-08 \*\*\*  
## Measles -0.0668474 0.0268716 -2.488 0.0129 \*   
## Total.expenditure 0.1352579 0.0306180 4.418 1.03e-05 \*\*\*  
## BMI 0.0177496 0.0044492 3.989 6.79e-05 \*\*\*  
## GDP 0.1895129 0.0280834 6.748 1.80e-11 \*\*\*  
## Income.composition.of.resources 7.6730577 0.5933148 12.933 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.809 on 2927 degrees of freedom  
## Multiple R-squared: 0.8401, Adjusted R-squared: 0.8396   
## F-statistic: 1538 on 10 and 2927 DF, p-value: < 2.2e-16

Poly.r\_squared <- summary(poly\_reg)$r.squared  
print(paste0("Linear regression r-squared: ", Poly.r\_squared))

## [1] "Linear regression r-squared: 0.840131284586183"

plot(poly\_reg)



summary(poly\_reg)

##   
## Call:  
## lm(formula = Life.expectancy ~ Schooling + Schooling^2 + Schooling^3 +   
## HIV.AIDS + HIV.AIDS^2 + HIV.AIDS^3 + Adult.Mortality + Adult.Mortality^2 +   
## Adult.Mortality^3 + infant.deaths + infant.deaths^2 + infant.deaths^3 +   
## Alcohol + Alcohol^2 + Alcohol^3 + Measles + Measles^2 + Measles^3 +   
## Total.expenditure + Total.expenditure^2 + Total.expenditure^3 +   
## BMI + BMI^2 + BMI^3 + GDP + GDP^2 + GDP^3 + Income.composition.of.resources +   
## Income.composition.of.resources^2 + Income.composition.of.resources^3,   
## data = poly\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0988 -1.9763 0.0518 2.2318 14.9758   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.5842926 0.4867453 138.849 < 2e-16 \*\*\*  
## Schooling 0.4169463 0.0389141 10.715 < 2e-16 \*\*\*  
## HIV.AIDS -9.6659788 0.2371008 -40.767 < 2e-16 \*\*\*  
## Adult.Mortality -0.0144805 0.0007674 -18.870 < 2e-16 \*\*\*  
## infant.deaths -0.9453013 0.0942088 -10.034 < 2e-16 \*\*\*  
## Alcohol 0.1233136 0.0218041 5.656 1.70e-08 \*\*\*  
## Measles -0.0668474 0.0268716 -2.488 0.0129 \*   
## Total.expenditure 0.1352579 0.0306180 4.418 1.03e-05 \*\*\*  
## BMI 0.0177496 0.0044492 3.989 6.79e-05 \*\*\*  
## GDP 0.1895129 0.0280834 6.748 1.80e-11 \*\*\*  
## Income.composition.of.resources 7.6730577 0.5933148 12.933 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.809 on 2927 degrees of freedom  
## Multiple R-squared: 0.8401, Adjusted R-squared: 0.8396   
## F-statistic: 1538 on 10 and 2927 DF, p-value: < 2.2e-16

**Polynomial Equation**

poly\_reg

##   
## Call:  
## lm(formula = Life.expectancy ~ Schooling + Schooling^2 + Schooling^3 +   
## HIV.AIDS + HIV.AIDS^2 + HIV.AIDS^3 + Adult.Mortality + Adult.Mortality^2 +   
## Adult.Mortality^3 + infant.deaths + infant.deaths^2 + infant.deaths^3 +   
## Alcohol + Alcohol^2 + Alcohol^3 + Measles + Measles^2 + Measles^3 +   
## Total.expenditure + Total.expenditure^2 + Total.expenditure^3 +   
## BMI + BMI^2 + BMI^3 + GDP + GDP^2 + GDP^3 + Income.composition.of.resources +   
## Income.composition.of.resources^2 + Income.composition.of.resources^3,   
## data = poly\_df)  
##   
## Coefficients:  
## (Intercept) Schooling   
## 67.58429 0.41695   
## HIV.AIDS Adult.Mortality   
## -9.66598 -0.01448   
## infant.deaths Alcohol   
## -0.94530 0.12331   
## Measles Total.expenditure   
## -0.06685 0.13526   
## BMI GDP   
## 0.01775 0.18951   
## Income.composition.of.resources   
## 7.67306

#hiv is sample bias or insignificant or biased result.

**Aic Test**

ply\_aic = AIC(poly\_reg)  
ply\_aic

## [1] 16209.43

lm\_aic = AIC(lm.fit\_mod)  
lm\_aic

## [1] 16985.04

print(paste0("Polynomial r-squared: ", Poly.r\_squared))

## [1] "Polynomial r-squared: 0.840131284586183"

print(paste0("Linear regression r-squared: ", lm.r\_squared))

## [1] "Linear regression r-squared: 0.838259166402846"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**END**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_