Data Analysis on Life Expectancy Dataset

Dataset: - https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

- Descriptive analysis using data visualization
- Inferential analysis using hypothesis testing (t-test, z-test, chi-square test)
- Partial correlations analysis
- Linear and Multivariant Regression Models

```
library(dslabs)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse
1.3.0 --
```

```
## v tibble  3.0.0     v purrr   0.3.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_confli
cts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ppcor)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
lifexp <- read.csv("D:\\Drivers\\academics\\sem 5\\stats\\stats lab\\Life Expectancy Data
.csv")
```

```
View(lifexp)
```

	Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths	Polio	Total.expenditur
1	Afghanistan	2015	Developing	65.0	263	62	0.01	71.279624	65	1154	19.1	83	6	8.16
2	Afghanistan	2014	Developing	59.9	271	64	0.01	73.523582	62	492	18.6	86	58	8.18
3	Afghanistan	2013	Developing	59.9	268	66	0.01	73.219243	64	430	18.1	89	62	8.13
4	Afghanistan	2012	Developing	59.5	272	69	0.01	78.184215	67	2787	17.6	93	67	8.52
5	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87
6	Afghanistan	2010	Developing	58.8	279	74	0.01	79.679367	66	1989	16.7	102	66	9.20
7	Afghanistan	2009	Developing	58.6	281	77	0.01	56.762217	63	2861	16.2	106	63	9.42
8	Afghanistan	2008	Developing	58.1	287	80	0.03	25.873925	64	1599	15.7	110	64	8.33
9	Afghanistan	2007	Developing	57.5	295	82	0.02	10.910156	63	1141	15.2	113	63	6.73
10	Afghanistan	2006	Developing	57.3	295	84	0.03	17.171518	64	1990	14.7	116	58	7.43
11	Afghanistan	2005	Developing	57.3	291	85	0.02	1.388648	66	1296	14.2	118	58	8.70

Showing 1 to 14 of 1,649 entries, 22 total columns

```
Fullifexp<-lifexp
lifexp <- lifexp %>% drop_na()

dim(lifexp)

## [1] 1649   22

range(lifexp$Year)

## [1] 2000 2015

#Adult.Mortality - Adult Mortality Rates on both sexes (probability of dying between 15-6
0 years/1000 population).
#infant.deaths - Number of Infant Deaths per 1000 population.
#Alcohol - Alcohol recorded per capita (15+) consumption (in litres of pure alcohol).
#percentage.expenditure - Expenditure on health as a percentage of Gross Domestic Product
per capita(%).
#Hepatitis.B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%) #how many b
elow one year of age received 3 doses (%)
#BMI - Average Body Mass Index of entire population.
#under.five.deaths - Number of under-five deaths per 1000 population.
#Polio - Polio (Pol3) immunization coverage among 1-year-olds (%).
#Total expenditure - General government expenditure on health as a percentage of total go
vernment expenditure (%).
#Diphtheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among
1-year-olds (%).
#HIV_AIDS - Deaths per 1 000 live births HIV/AIDS (0-4 years).
#GDP - Gross Domestic Product per capita (in USD).
#Population - Population of the country.
#thinness.10.19 years - Prevalence of thinness among children and adolescents for Age 10
to 19 (%).
#thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%).
#Income.composition.of.resources -  in terms of income composition of resources (index ra
nging from 0 to 1).
#Schooling - Number of years of Schooling(years) .

str(lifexp)

## 'data.frame':    1649 obs. of  22 variables:
##  $ Country                        : Factor w/ 193 levels "Afghanistan",..: 1 1 1 1 1 1
1 1 1 1 ...
```

```
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007
2006 ...
##  $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2
2 2 2 2 2 2 2 ...
##  $ 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 19
90 ...
##  $ 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 0.43
3 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 ...
```

```
summary(lifexp)
```

```
##         Country          Year             Status     Life.expectancy
##  Afghanistan:  16   Min.   :2000   Developed : 242   Min.   :44.0
##  Albania    :  16   1st Qu.:2005   Developing:1407   1st Qu.:64.4
##  Armenia    :  15   Median :2008                     Median :71.7
##  Austria    :  15   Mean   :2008                     Mean   :69.3
##  Belarus    :  15   3rd Qu.:2011                     3rd Qu.:75.0
##  Belgium    :  15   Max.   :2015                     Max.   :89.0
##  (Other)    :1557
##  Adult.Mortality infant.deaths        Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.00
##  1st Qu.: 77.0   1st Qu.:   1.00   1st Qu.: 0.810   1st Qu.:   37.44
##  Median :148.0   Median :   3.00   Median : 3.790   Median :  145.10
##  Mean   :168.2   Mean   :  32.55   Mean   : 4.533   Mean   :  698.97
##  3rd Qu.:227.0   3rd Qu.:  22.00   3rd Qu.: 7.340   3rd Qu.:  509.39
##  Max.   :723.0   Max.   :1600.00   Max.   :17.870   Max.   :18961.35
##
##   Hepatitis.B       Measles            BMI        under.five.deaths
##  Min.   : 2.00   Min.   :     0   Min.   : 2.00   Min.   :   0.00
##  1st Qu.:74.00   1st Qu.:     0   1st Qu.:19.50   1st Qu.:   1.00
##  Median :89.00   Median :    15   Median :43.70   Median :   4.00
##  Mean   :79.22   Mean   :  2224   Mean   :38.13   Mean   :  44.22
##  3rd Qu.:96.00   3rd Qu.:   373   3rd Qu.:55.80   3rd Qu.:  29.00
##  Max.   :99.00   Max.   :131441   Max.   :77.10   Max.   :2100.00
```

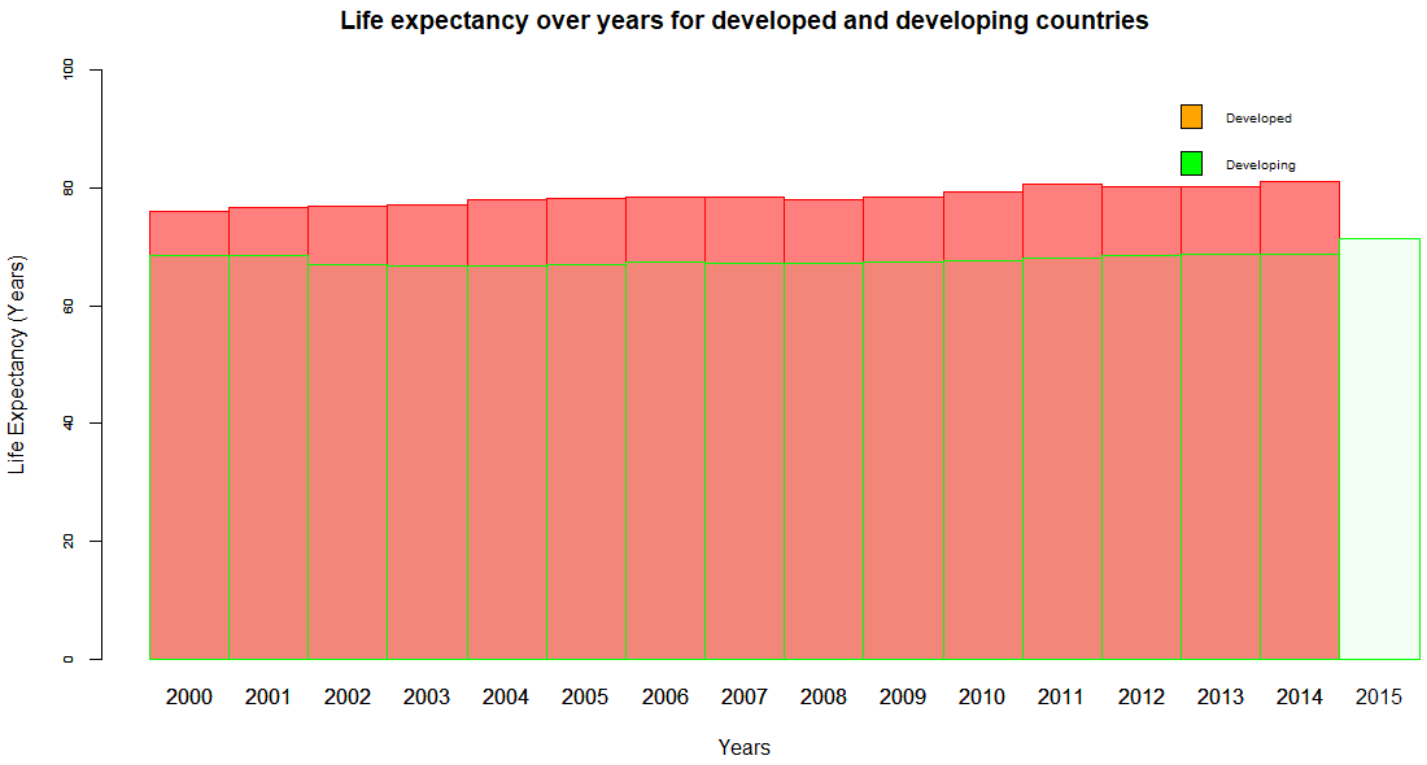
```
##
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.740    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:81.00   1st Qu.: 4.410    1st Qu.:82.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.840    Median :92.00   Median : 0.100
##  Mean   :83.56   Mean   : 5.956    Mean   :84.16   Mean   : 1.984
##  3rd Qu.:97.00   3rd Qu.: 7.470    3rd Qu.:97.00   3rd Qu.: 0.700
##  Max.   :99.00   Max.   :14.390    Max.   :99.00   Max.   :50.600
##
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.100
##  1st Qu.:   462.15   1st Qu.:1.919e+05   1st Qu.: 1.600
##  Median :  1592.57   Median :1.420e+06   Median : 3.000
##  Mean   :  5566.03   Mean   :1.465e+07   Mean   : 4.851
##  3rd Qu.:  4718.51   3rd Qu.:7.659e+06   3rd Qu.: 7.100
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.200
##
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.100     Min.   :0.0000                  Min.   : 4.20
##  1st Qu.: 1.700     1st Qu.:0.5090                  1st Qu.:10.30
##  Median : 3.200     Median :0.6730                  Median :12.30
##  Mean   : 4.908     Mean   :0.6316                  Mean   :12.12
##  3rd Qu.: 7.100     3rd Qu.:0.7510                  3rd Qu.:14.00
##  Max.   :28.200     Max.   :0.9360                  Max.   :20.70
##
```

```
range(lifexp$Life.expectancy)
```

```
## [1] 44 89
```

```
data_corr <- lifexp %>% select_if(is.numeric)  #remove

coef<-cor(data_corr,method = "pearson")         #remove
View(coef)



#Life expectancy distribution in our dataset

hist(lifexp$Life.expectancy, breaks = seq(0,100,10),ylim=c(0,800), xlab = "Life Expectanc
y",
     main="Life expectancy over years for developed and developing countries", col = "gra
y50")
axis(1, at = seq(0, 100, by = 10))
```


```
summary(lifexp$Life.expectancy)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    44.0    64.4    71.7    69.3    75.0    89.0

#Does Status of country have an effect on life expectancy?

Dev<- lifexp%>% filter(Status=="Developed") %>% group_by(Year) %>% summarise(total=mean(L
ife.expectancy))

Devloping<-lifexp%>% filter(Status=="Developing") %>% group_by(Year) %>% summarise(total=
mean(Life.expectancy))

barplot(Dev$total,names.arg = Dev$Year,
        cex.axis = 0.65,ylim=c(0,100),space=c(0),xlab = "Years",
        col=rgb(1, 0, 0, .5),border = "Red",ylab="Life Expectancy (Years)",
        main="Life expectancy over years for developed and developing countries",)

barplot(Devloping$total,names.arg = Devloping$Year,
        cex.axis = 0.65,ylim=c(0,100),space=c(0),xlab = "Years",
        col=rgb(0, 1, 0, .05),border = "Green",ylab="Adult mortality rates",
        main="Life expectancy over years for developed and developing countries",
        add = TRUE)
legend("topright",
       c("Developed","Developing"),
       fill = c("orange","green"),cex=0.65,bty="n"
)
```




```
print("Over the years there isn't much difference in life expectancy though it differs si
gnificantly for developed and developing countries");
```

```
## [1] "Over the years there isn't much difference in life expectancy though it differs s
ignificantly for developed and developing countries"
```

```
# Life expectancy for developed countries in 2014

data_developed<- lifexp%>% filter(Status=="Developed" & Year==2014)
dim(data_developed)
```

```
## [1] 19 22
```

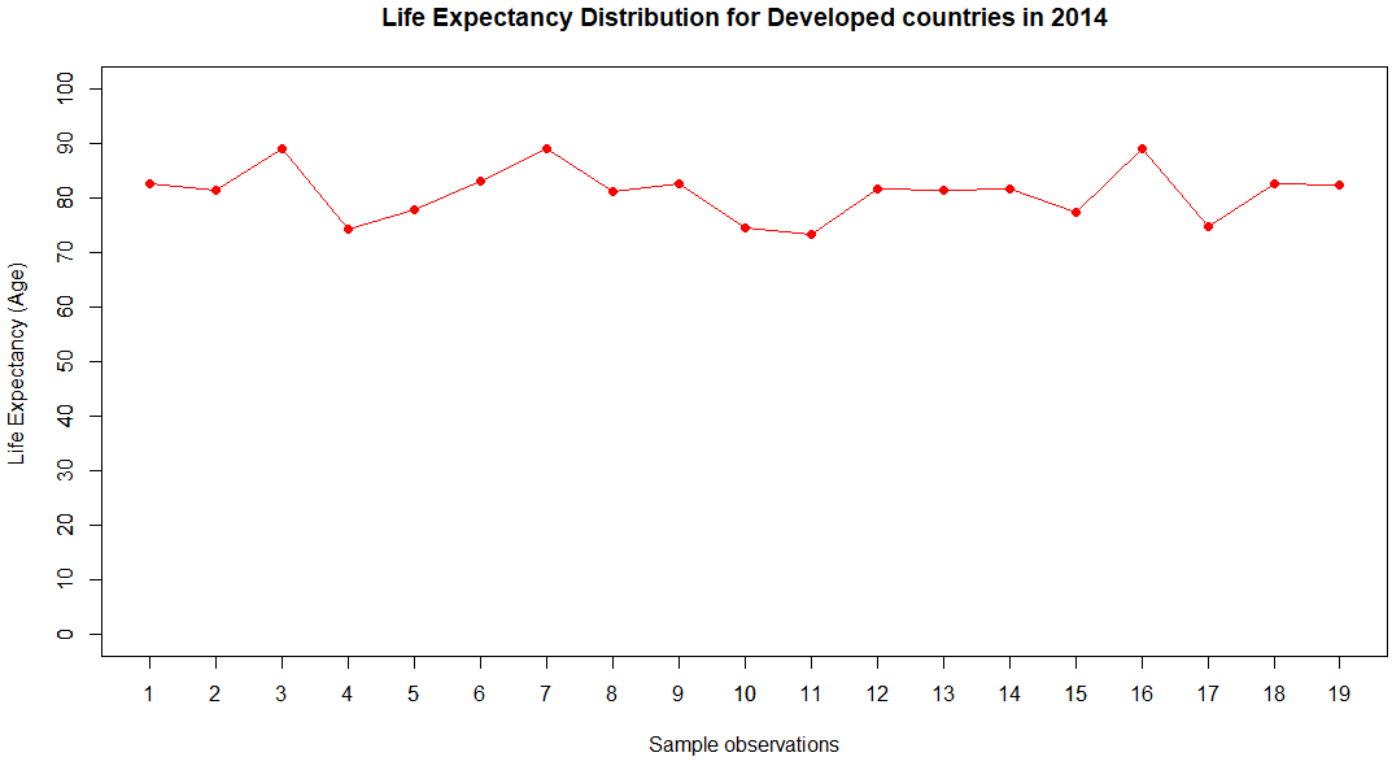
```
Full_data_developed<- Fullifexp%>% filter(Status=="Developed" & Year==2014)
dim(Full_data_developed)
```

```
## [1] 32 22
```

```
mean(data_developed$Life.expectancy)
```

```
## [1] 81.02632
```

```
plot((seq(1:19)), data_developed$Life.expectancy, type = "o", pch = 19, ylim=c(0,100),
     col = "red", ylab = "Life Expectancy (Age) ",
     xlab = "Sample observations",main="Life Expectancy Distribution for Developed countr
ies in 2014")
axis(2, at = seq(0, 100, by = 10))
axis(1, at = seq(1, 19, by = 1))
```




```
# Can average life expectancy for developed countries in 2014 be said as 80 ?

#Null hypothesis life expectancy for developed countries in 2014 = 80
isAccept<- function(tcal,dF,alpha,isTwotailed){
  ttable<-0;
  if(isTwotailed){
    ttable<-abs(qt(alpha/2,df=dF))
  }else{
    ttable<-abs(qt(alpha,df=dF))
  }
  tcal<-abs(tcal)
  if(tcal<ttable){
    cat("The calculated t value is ", tcal," is less than the table t value ",ttable,"\n"
)
    cat("Hence the null hypothesis is accepted")
  }else{
    cat("The calculated t value is ", tcal," is greater than the table t value ",ttable,"
\n")
    cat("Hence the null hypothesis is rejected")
  }
}

res<-t.test(data_developed$Life.expectancy,alt="two.sided",mu=80,conf.level=0.95)
res

##
##  One Sample t-test
##
## data:  data_developed$Life.expectancy
## t = 0.93135, df = 18, p-value = 0.364
## alternative hypothesis: true mean is not equal to 80
## 95 percent confidence interval:
```

```
##  78.71118 83.34145
## sample estimates:
## mean of x
##  81.02632
```

```
isAccept(res$statistic,res$parameter,0.05,TRUE)
```

```
## The calculated t value is  0.9313539  is less than the table t value  2.100922
## Hence the null hypothesis is accepted
```

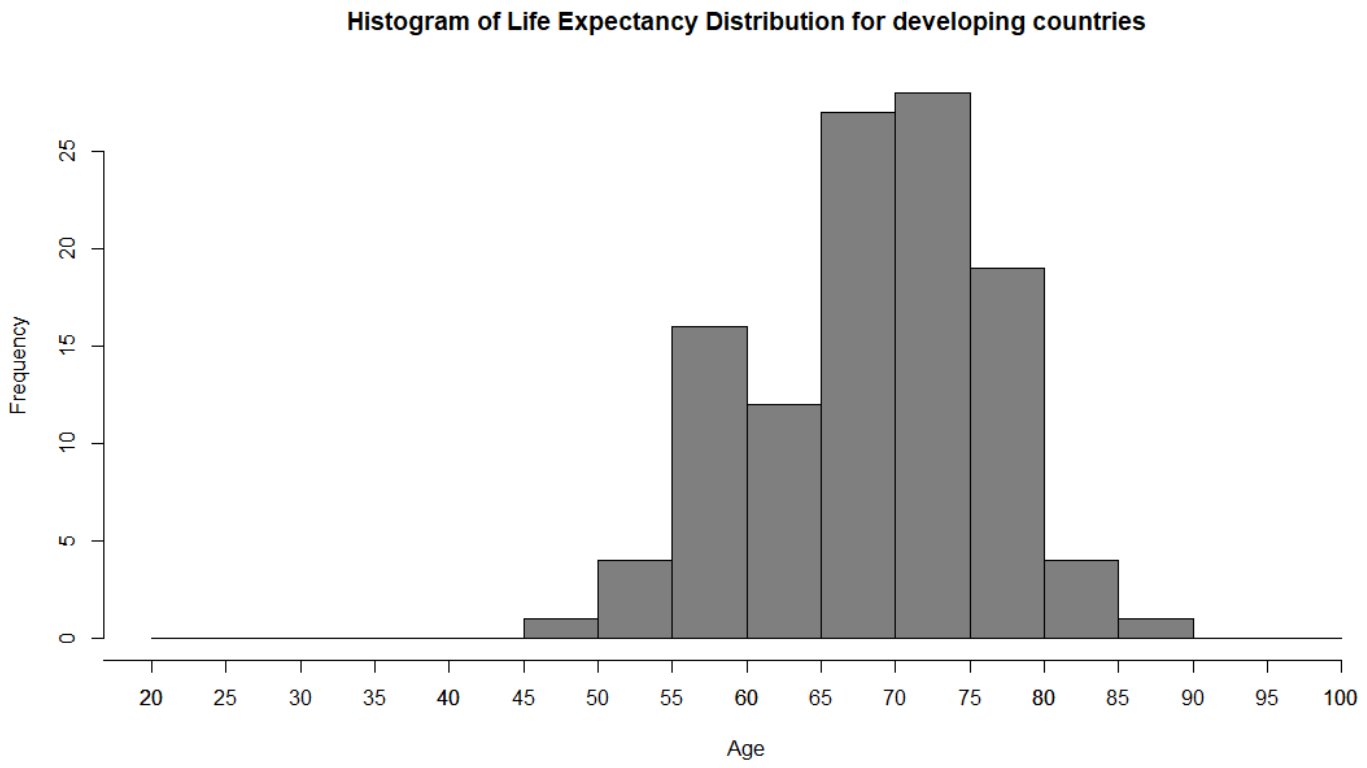
```
#Developing countries in 2014

data_developing<- lifexp%>% filter( Status=="Developing" & Year==2014)
dim(data_developing)
```

```
## [1] 112  22
```

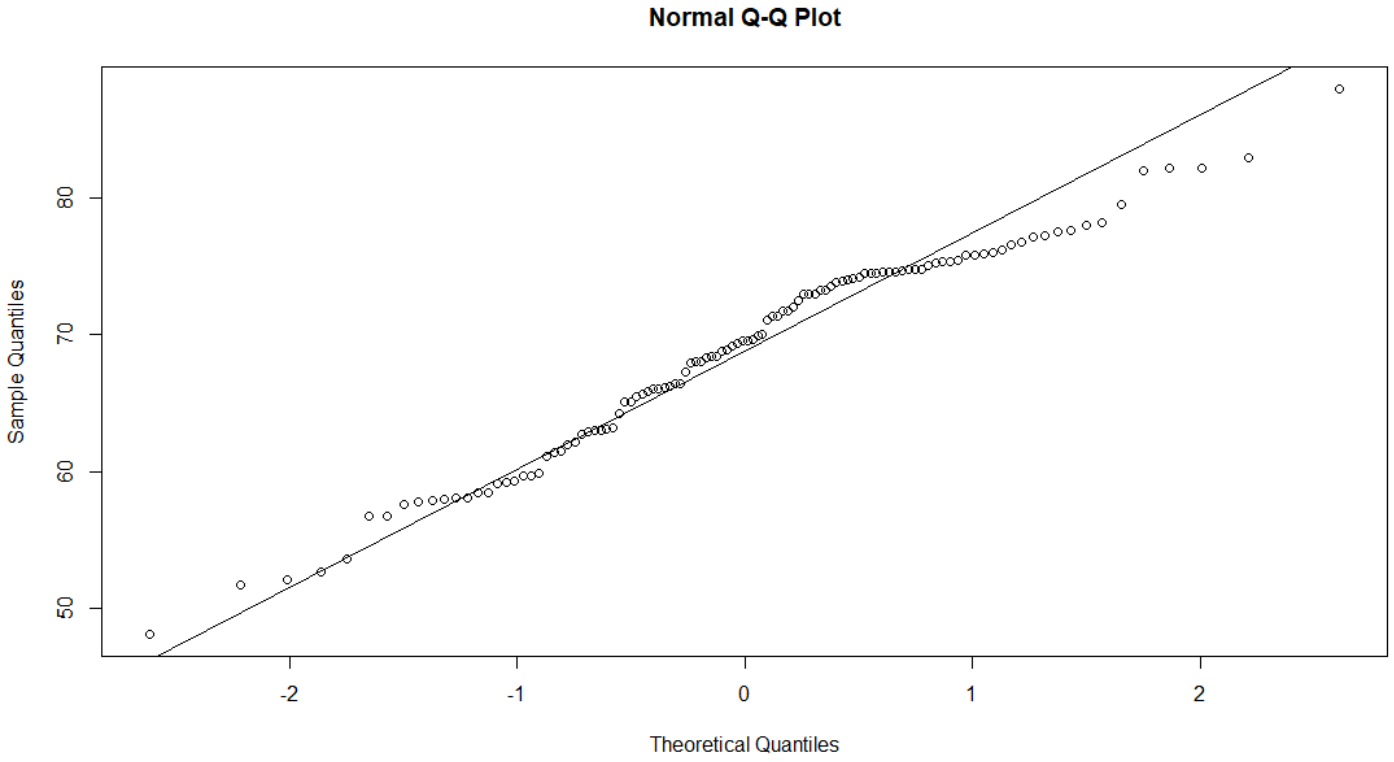
```
Full_data_developing<- Fullifexp%>% filter(Status=="Developing" & Year==2014)
dim(Full_data_developing)
```

```
## [1] 151  22
```

```
hist(data_developing$Life.expectancy, breaks = seq(20,100,5), xlab = "Age",
     main="Histogram of Life Expectancy Distribution for developing countries", col = "gr
ay50")
axis(1, at = seq(0, 100, by = 5))
```




```
#Null hypothesis : the life expectancy of population is 70


#is distribution normal ?
qqnorm(data_developing$Life.expectancy)
qqline(data_developing$Life.expectancy)
```




```
#Z test
isAccept<- function(zcal,alpha,isTwotailed){
  ztable<-0;
  if(isTwotailed){
    ztable<-abs(qnorm(alpha/2))
  }else{
    ztable<-abs(qnorm(alpha))
  }
  zcal<-abs(zcal)
  if(zcal<ztable){
    cat("The calculated z value is ", zcal," is less than the table z value ",ztable,"\n"
)
    cat("Hence the null hypothesis is accepted")
  }else{
    cat("The calculated z value is ", zcal," is greater than the table z value ",ztable,"
\n")
    cat("Hence the null hypothesis is rejected")
  }
}

sd<-sd(data_developing$Life.expectancy)
z<- (mean(data_developing$Life.expectancy)-70)/sd

isAccept(z,0.05,TRUE)  # for 95% confidence interval

## The calculated z value is  0.1618012  is less than the table z value  1.959964
## Hence the null hypothesis is accepted

#Top 10 countries which have highest improvement of life expectancy in 2014 as compared t
o 2000?

high_lifexp<-lifexp%>% filter((Year==2014 | Year == 2000) &Life.expectancy!=0 ) %>%group_
```

```
by(Country,Year) %>% summarise(Life.expectancy) %>% arrange(desc(Life.expectancy))
View(high_lifexp)
```

Filter

	Country	Year	Life.expectancy
1	Belgium	2014	89.0
2	Germany	2014	89.0
3	Portugal	2014	89.0
4	Greece	2014	88.0
5	Chile	2014	83.0
6	Cyprus	2014	83.0
7	Australia	2014	82.7
8	Spain	2014	82.6
9	Italy	2014	82.5
10	Sweden	2014	82.3
11	France	2014	82.2
12	Israel	2014	82.2
13	Canada	2014	82.0
14	Luxembourg	2014	81.7
15	Netherlands	2014	81.7
16	Austria	2014	81.4
17	Malta	2014	81.4
18	Ireland	2014	81.2

Showing 1 to 19 of 192 entries, 3 total columns

```
high_lifexp$Year<- as.factor(high_lifexp$Year)

high_lifexp_s<-high_lifexp %>% spread(Year,Life.expectancy)
View(high_lifexp_s)

high_lifexp_s<- high_lifexp_s%>% mutate(diff= (`2014`- `2000`))
high_lifexp_s <- high_lifexp_s %>% drop_na()
View(high_lifexp_s)
```

	Country	2000	2014	diff
1	Afghanistan	54.8	59.9	5.1
2	Albania	72.6	77.5	4.9
3	Armenia	72.0	74.6	2.6
4	Austria	78.1	81.4	3.3
5	Belarus	68.0	72.0	4.0
6	Belgium	77.6	89.0	11.4
7	Belize	68.3	70.0	1.7
8	Bhutan	62.0	69.4	7.4
9	Botswana	47.8	65.1	17.3
10	Brazil	75.0	74.8	-0.2
11	Bulgaria	71.1	74.3	3.2
12	China	71.7	75.8	4.1
13	Colombia	71.4	74.6	3.2
14	Costa Rica	77.6	79.5	1.9
15	Cyprus	78.1	83.0	4.9
16	Dominican Republic	72.0	73.6	1.6
17	Ecuador	72.8	76.0	3.2
18	El Salvador	69.0	73.3	4.3

Showing 1 to 19 of 60 entries, 4 total columns

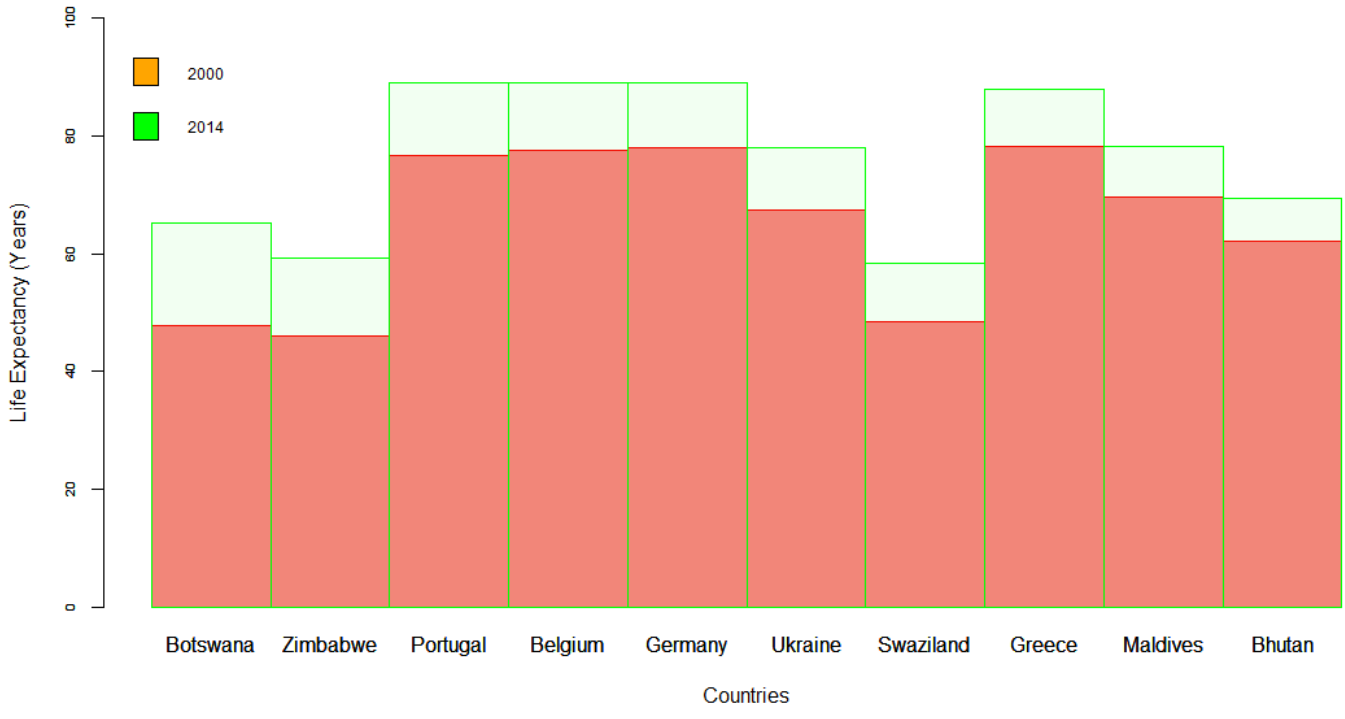
```
#did any country had a fall in life expectancy in 2014 compared to 2000
high_lifexp_s%>% filter(diff<=0)
```

```
## # A tibble: 2 x 4
## # Groups:   Country [193]
##   Country `2000` `2014`  diff
##   <fct>    <dbl>  <dbl> <dbl>
## 1 Brazil      75   74.8  -0.2
## 2 Romania     77   74.8  -2.2
```

```
#Brazil and Romania had a fall in thier life expectancy

top_improved<-high_lifexp_s%>% arrange(desc(diff))
top_improved<-top_improved[1:10,]
View(top_improved) #top improved
```

	Country	2000	2014	diff
1	Botswana	47.8	65.1	17.3
2	Zimbabwe	46.0	59.2	13.2
3	Portugal	76.6	89.0	12.4
4	Belgium	77.6	89.0	11.4
5	Germany	78.0	89.0	11.0
6	Ukraine	67.5	78.0	10.5
7	Swaziland	48.4	58.4	10.0
8	Greece	78.2	88.0	9.8
9	Maldives	69.6	78.2	8.6
10	Bhutan	62.0	69.4	7.4

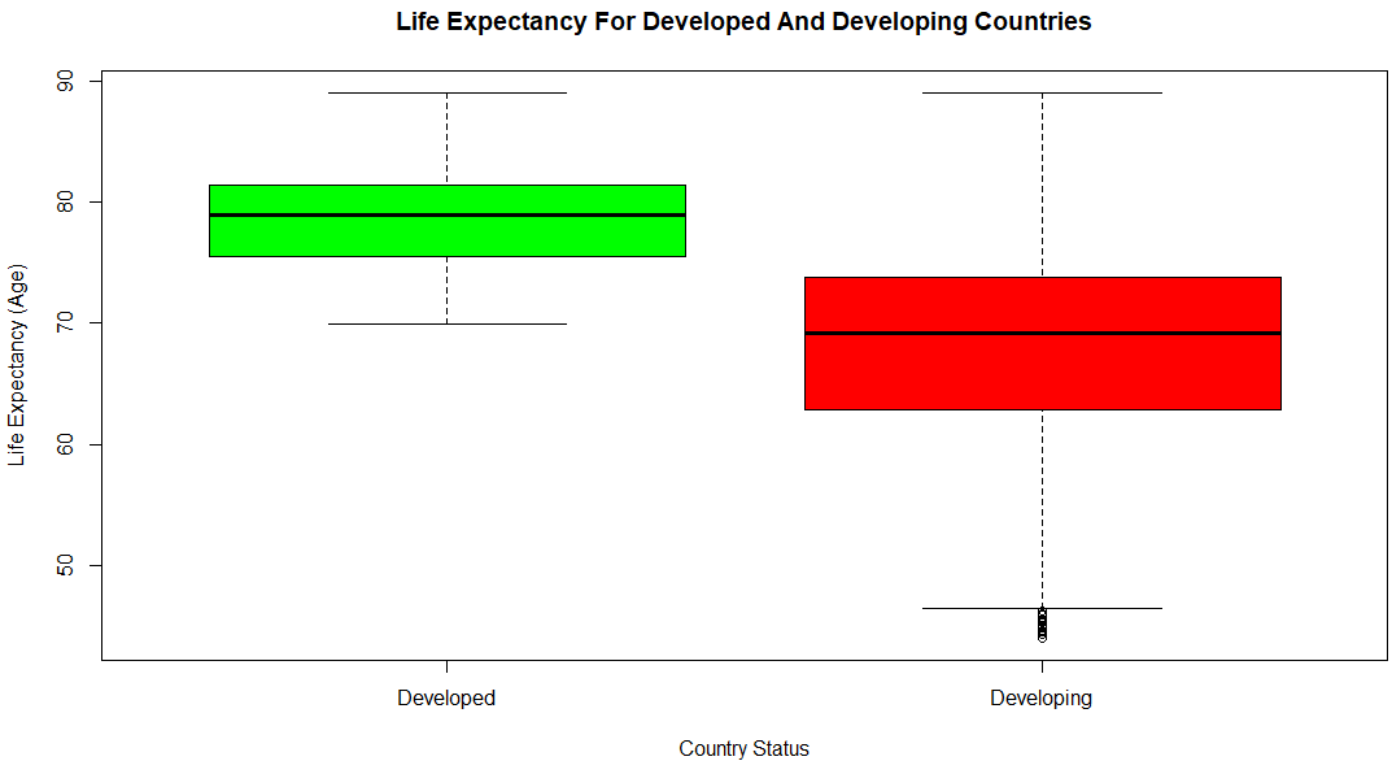
```
barplot(top_improved$`2000`,names.arg = top_improved$Country,
        cex.axis = 0.65,ylim=c(0,100),space=c(0),xlab = "Countries",
        col=rgb(1, 0, 0, .5),border = "Red",ylab="Life Expectancy (Years)")

barplot(top_improved$`2014`,names.arg = top_improved$Country,
        cex.axis = 0.65,ylim=c(0,100),space=c(0),xlab = "Countries",
        col=rgb(0, 1, 0, .05),border = "Green",ylab="Adult mortality rates",
        main="Life expectancy for different countries for year 2000 and 2014",
        add = TRUE)

legend("topleft",
       c("2000","2014"),
       fill = c("orange","green"),cex=0.75,bty="n"
)
```



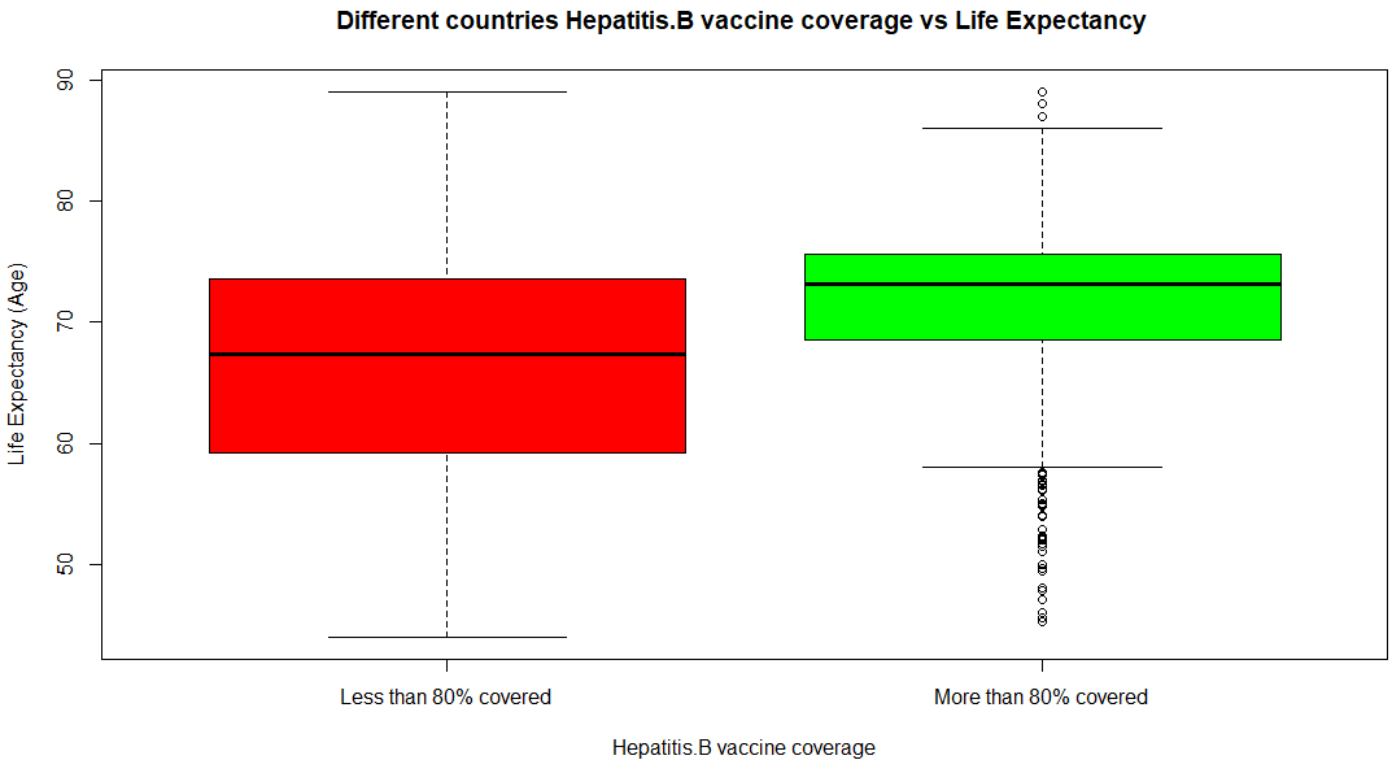
```r
#adult mortality over years in developed and developing countries

morDev<- lifexp%>% filter(Status=="Developed") %>% group_by(Year) %>% summarise(total=mea
n(Adult.Mortality))

morDevloping<-lifexp%>% filter(Status=="Developing") %>% group_by(Year) %>% summarise(tot
al=mean(Adult.Mortality))

barplot(morDev$total,names.arg = morDev$Year,
        cex.axis = 0.65,ylim=c(0,250),space=c(0),xlab = "Years",
        col=rgb(1, 0, 0, .5),border = "Red",ylab="Adult mortality rates",main="Adult mort
ality rates over years")

barplot(morDevloping$total,names.arg = morDevloping$Year,
        cex.axis = 0.65,ylim=c(0,250),space=c(0),xlab = "Years",
        col=rgb(0, 1, 0, .05),border = "Green",ylab="Adult mortality rates",
        main="Avg adult mortality rates over years for developed and developing countries
",
        add = TRUE)
legend("topright",
       c("Developed","Developing"),
       fill = c("orange","green"),cex=0.75,bty="n"
)
```


```r
#Average adult mortality rates over the years have not changed significantly but
#its quite low in developed countries as compared to developing countries


#life exp in dev and devloping countries

boxplot(lifexp$Life.expectancy~lifexp$Status,xlab="Country Status",
        ylab = "Life Expectancy (Age)",main="Life Expectancy For Developed And Developing
```

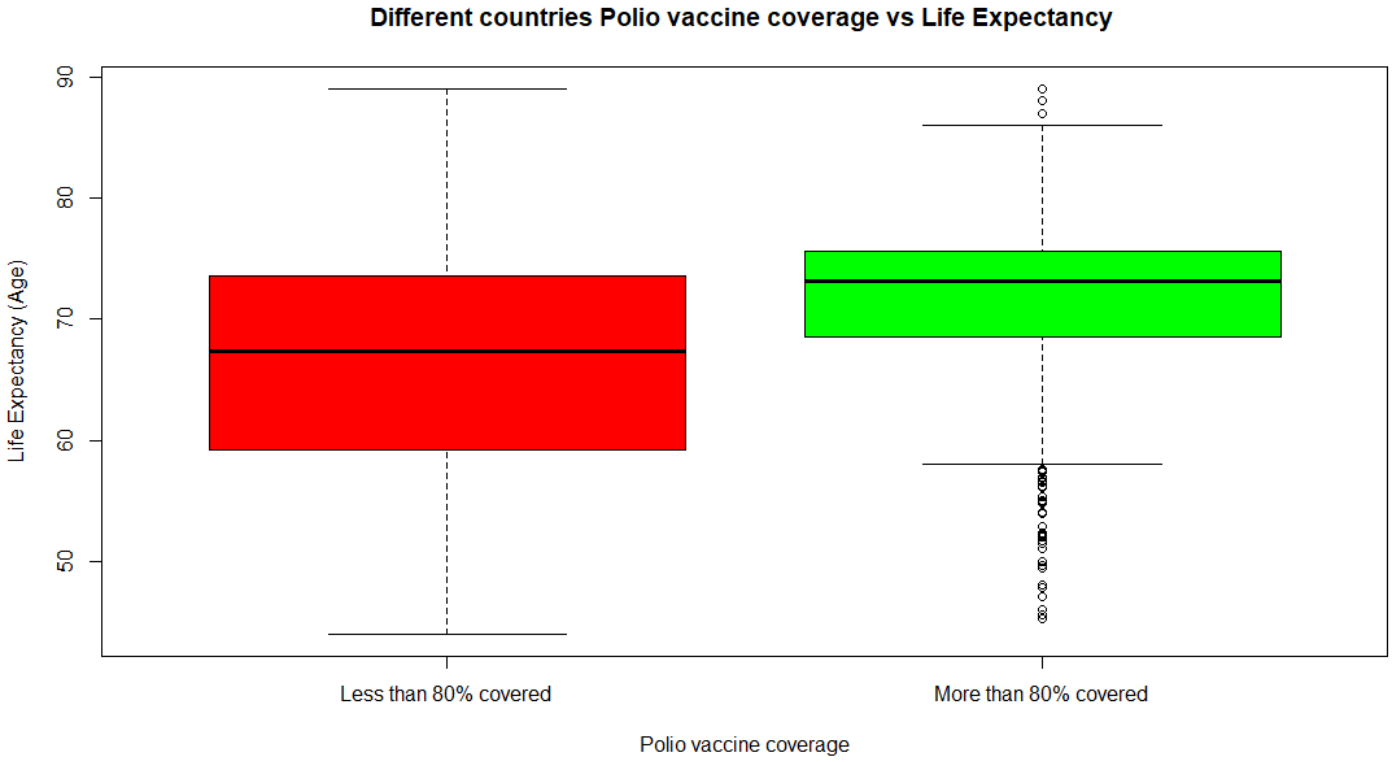
```
Countries",
        col = c("green","red"),
        names = c("Developed","Developing"))
```



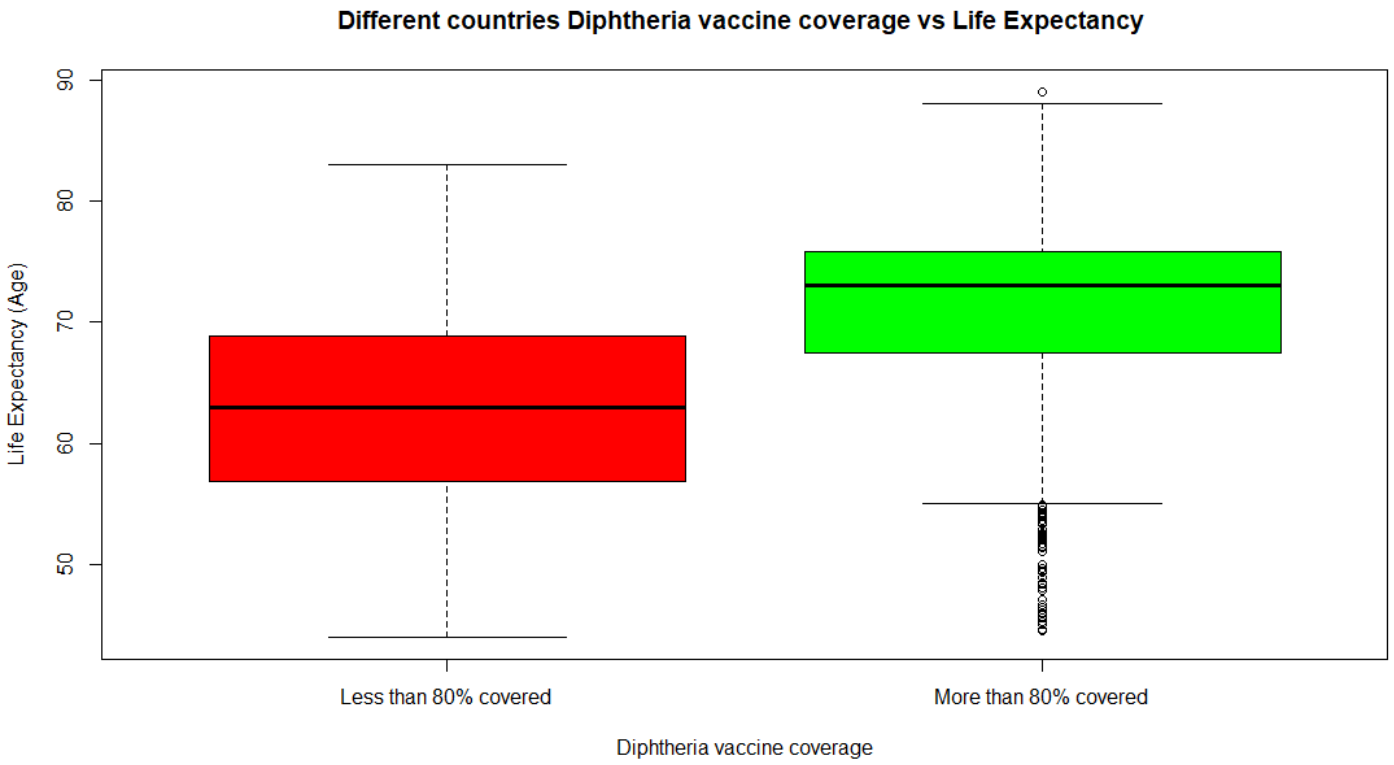

```
#vaccination status

lifexp1 <- lifexp %>%
  mutate(Hepatitis.B = ifelse(Hepatitis.B < 90, "lt80", "mt80"),
         Polio = ifelse(Polio < 80, "lt80", "mt80"),
         Diphtheria = ifelse(Diphtheria < 80, "lt80", "mt80"),
         Hepatitis.B = as.factor(Hepatitis.B),
         Polio = as.factor(Polio),
         Diphtheria = as.factor(Diphtheria))


# corr between Hepatitis B Coverage and life expectancy

coeff<-cor(lifexp$Hepatitis.B,lifexp$Life.expectancy, method = "pearson")
coeff  #v low

## [1] 0.1999353

boxplot(lifexp1$Life.expectancy~lifexp1$Hepatitis.B,xlab="Hepatitis.B vaccine coverage",
        ylab = "Life Expectancy (Age)",main="Different countries Hepatitis.B vaccine cove
rage vs Life Expectancy ",
        col = c("red","green"),
        names = c("Less than 80% covered","More than 80% covered"))
```



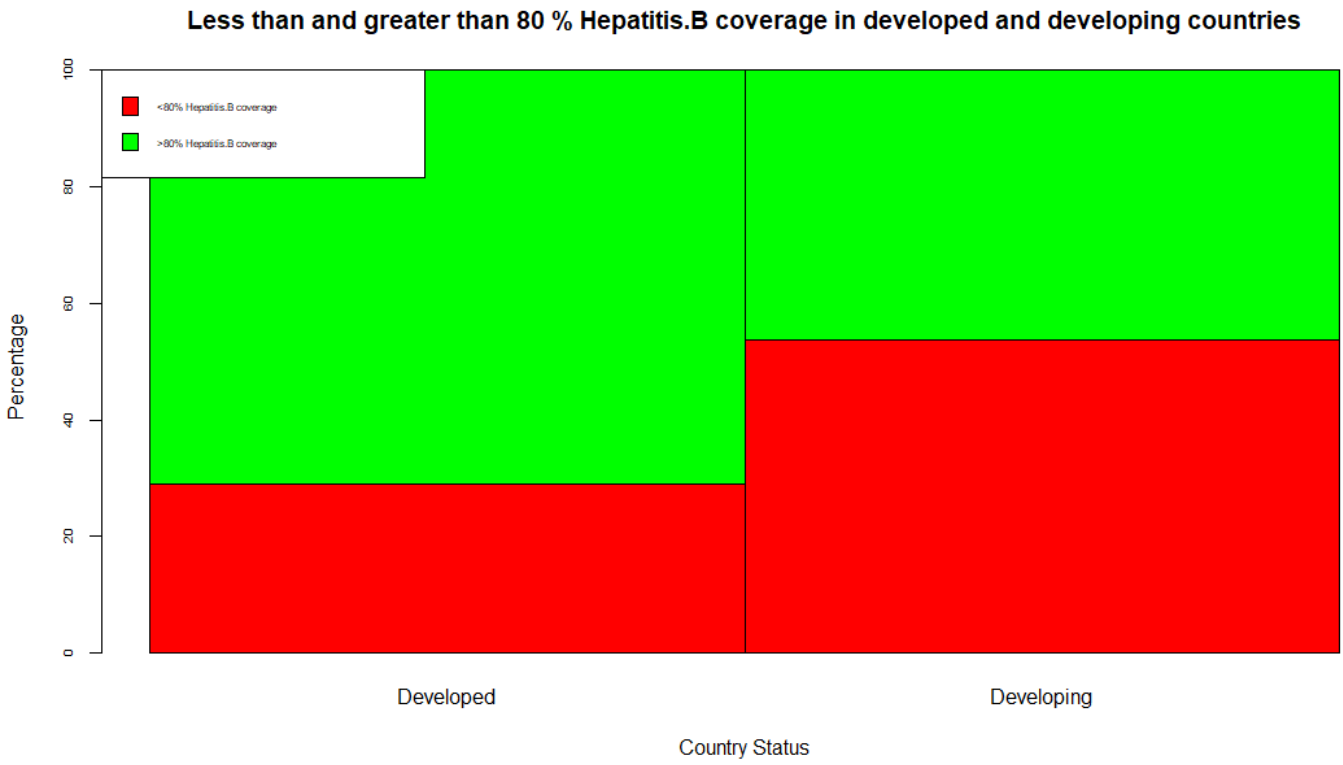

```
# Polio Coverage and life expectancy

coeff<-cor(lifexp$Polio,lifexp$Life.expectancy, method = "pearson")
coeff  #

## [1] 0.3272944

boxplot(lifexp1$Life.expectancy~lifexp1$Hepatitis.B,xlab="Polio vaccine coverage",
        ylab = "Life Expectancy (Age)",main="Different countries Polio vaccine coverage v
s Life Expectancy ",
        col = c("red","green"),
        names = c("Less than 80% covered","More than 80% covered"))
```




```
coeff<-cor(lifexp$Diphtheria,lifexp$Life.expectancy, method = "pearson")
coeff

## [1] 0.3413312

boxplot(lifexp1$Life.expectancy~lifexp1$Diphtheria,xlab="Diphtheria vaccine coverage",
        ylab = "Life Expectancy (Age)",main="Different countries Diphtheria vaccine cover
age vs Life Expectancy ",
        col = c("red","green"),
        names = c("Less than 80% covered","More than 80% covered"))
```




```
y <- lifexp1 %>% group_by(Status, Hepatitis.B) %>%  count()%>% spread(Hepatitis.B, n, fil
l = 0L)
y<- y%>%  mutate( lt80p=(lt80/(lt80+mt80))*100,mt80p=(mt80/(lt80+mt80))*100  )

View(y)
barplot(height = t(y[c("lt80p","mt80p" )]),
        names.arg = y$Status,col = c("red","green"),space=c(0),cex.axis = 0.65,xlab = "Co
untry Status",ylab="Percentage ",main="Less than and greater than 80 % Hepatitis.B covera
ge in developed and developing countries")

legend("topleft",
       c("<80% Hepatitis.B coverage",">80% Hepatitis.B coverage"),
       fill = c("red","green"),cex=0.5
)
```



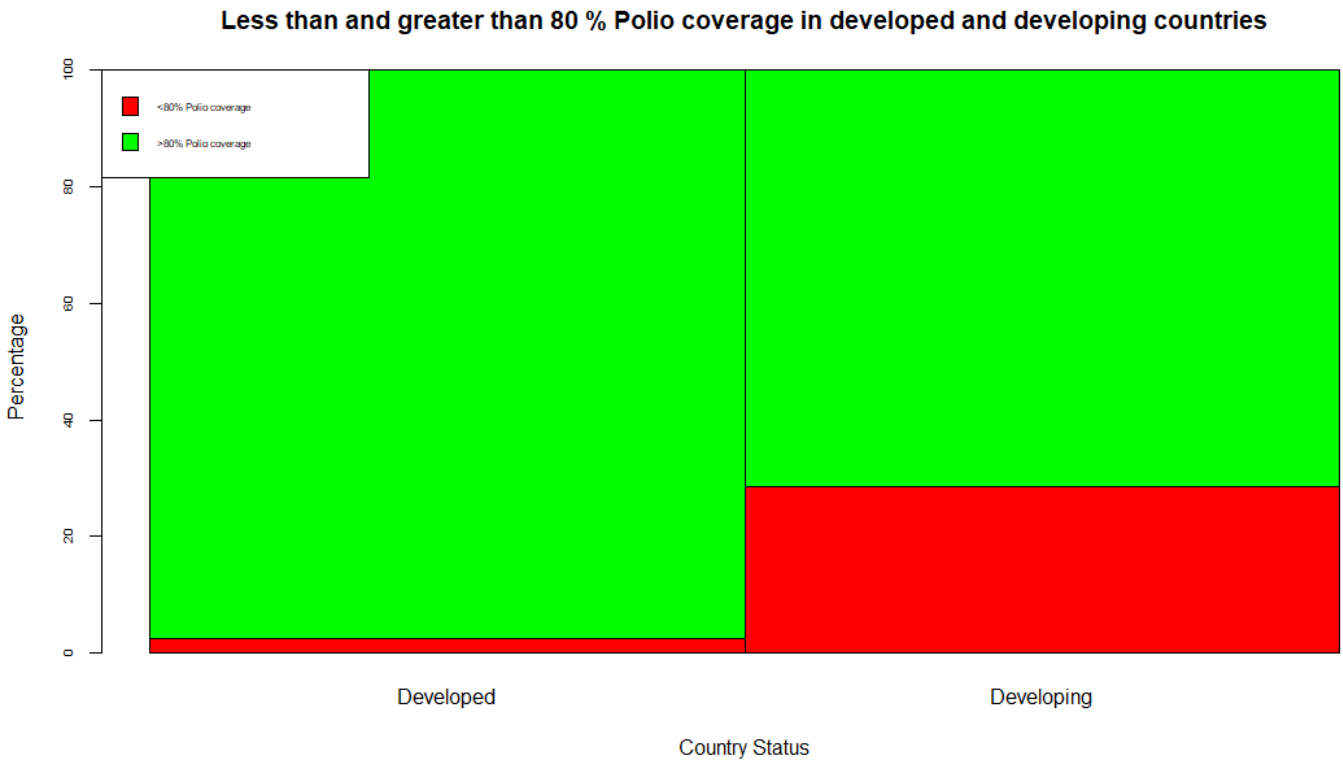

```
# Does the staus of country affect Hepatitis.B coverage significantly ?

#chi square test
#function to compare x value calculated and table x value
acceptChi<-function(Xcal,alpha,df){
  XTable<-qchisq(p=alpha,df,lower.tail=FALSE)
  if(abs(Xcal)<=abs(XTable)){
    cat("The calculated X value is ", abs(Xcal)," is less than the table X value ",abs(XT
able),"\n")
    print("There isn't much difference in the groups of data ")
  }else{
    cat("The calculated X value is ", abs(Xcal)," is greater than the table X value ",abs
(XTable),"\n")
    print("There is a significant difference in the groups of data")
  }
}


df<-y[c(4,5)]
df<-data.frame(df)
rownames(df) <- c("Developed","Developing")
#class(df)
#dim(df)
#str(df)
df

##                lt80p    mt80p
## Developed  28.92562 71.07438
## Developing 53.73134 46.26866

res<-chisq.test(df,correct=FALSE)
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  df
## X-squared = 12.688, df = 1, p-value = 0.000368
```

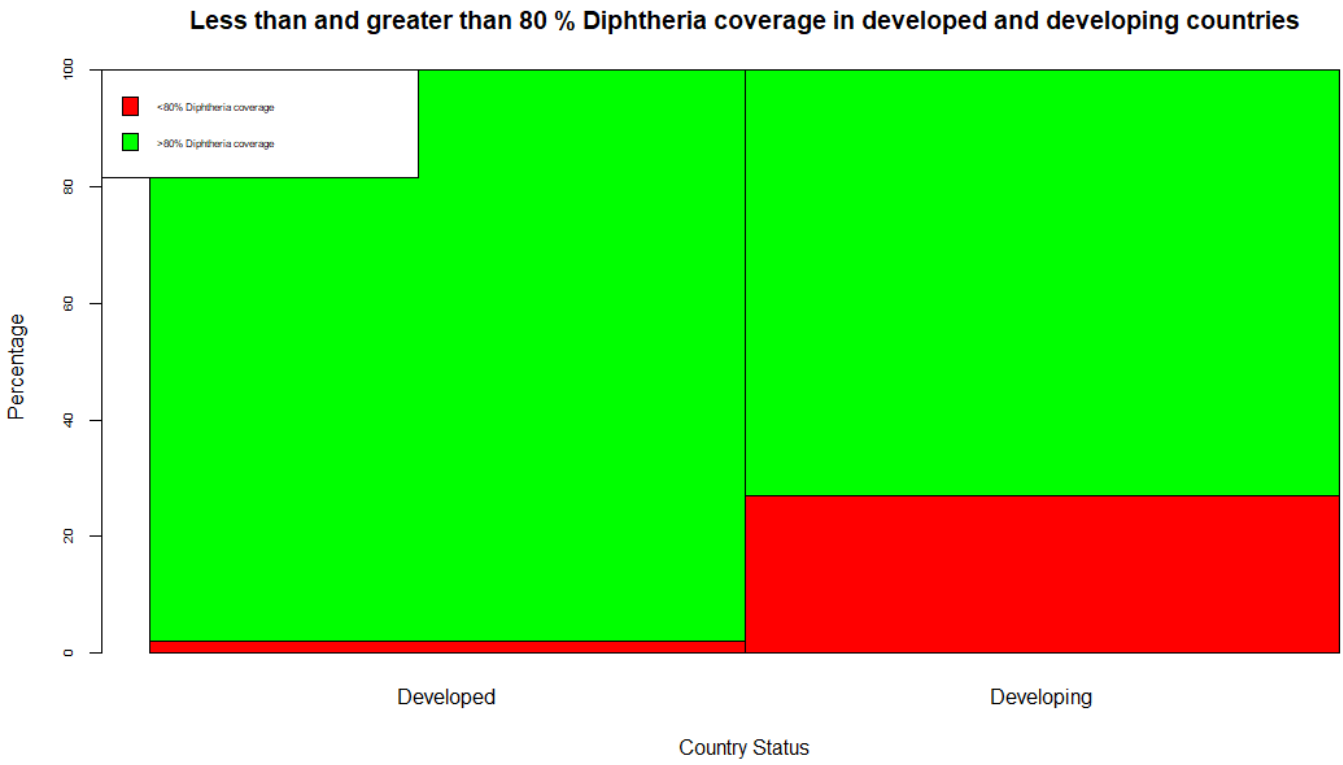
```
acceptChi(res$statistic,0.05,res$parameter)
```

```
## The calculated X value is  12.68811  is greater than the table X value  3.841459
## [1] "There is a significant difference in the groups of data"
```

```
# Hence development status affect the Hepatitis.B coverage


#Does developed or devloping status of country affect polio immunization?

y <- lifexp1 %>% group_by(Status, Polio) %>%  count()%>% spread(Polio, n, fill = 0L)
y<- y%>%  mutate( lt80p=(lt80/(lt80+mt80))*100,mt80p=(mt80/(lt80+mt80))*100  )

#View(y)
barplot(height = t(y[c("lt80p","mt80p" )]),
        names.arg = y$Status,col = c("red","green"),space=c(0),cex.axis = 0.65,xlab = "Co
untry Status",
        ylab="Percentage ",
        main="Less than and greater than 80 % Polio coverage in developed and developing
countries")

legend("topleft",
       c("<80% Polio coverage",">80% Polio coverage"),
       fill =  c("red","green"),cex=0.5
)
```



```r
#chi square test Does the staus of country affect Hepatitis.B coverage significantly ?

#function to compare x value calculated and table x value
acceptChi<-function(Xcal,alpha,df){
  XTable<-qchisq(p=alpha,df,lower.tail=FALSE)
  if(abs(Xcal)<=abs(XTable)){
    cat("The calculated X value is ", abs(Xcal)," is less than the table X value ",abs(XT
able),"\n")
    print("There isn't much difference in the groups of data ")
  }else{
    cat("The calculated X value is ", abs(Xcal)," is greater than the table X value ",abs
(XTable),"\n")
    print("There is a significant difference in the groups of data")
  }
}


df<-y[c(4,5)]
df<-data.frame(df)
rownames(df) <- c("Developed","Developing")
#class(df)
#dim(df)
#str(df)
res<-chisq.test(df,correct=FALSE)
res

##
##  Pearson's Chi-squared test
##
## data:  df
## X-squared = 26.048, df = 1, p-value = 3.331e-07

acceptChi(res$statistic,0.05,res$parameter)

## The calculated X value is  26.04784  is greater than the table X value  3.841459
## [1] "There is a significant difference in the groups of data"

# Hence development status affect the Polio coverage


#Does developed or devloping status of country affect Diphteria immunization?

y <- lifexp1 %>% group_by(Status, Diphtheria) %>%  count()%>% spread(Diphtheria, n, fill
= 0L)
y<- y%>%  mutate( lt80p=(lt80/(lt80+mt80))*100,mt80p=(mt80/(lt80+mt80))*100  )

barplot(height = t(y[c("lt80p","mt80p" )]),
        names.arg = y$Status,col = c("red","green"),space=c(0),cex.axis = 0.65,xlab = "Co
untry Status",
        ylab="Percentage ",
        main="Less than and greater than 80 % Diphtheria coverage in developed and develo
ping countries")

legend("topleft",
       c("<80% Diphtheria coverage",">80% Diphtheria coverage"),
       fill =  c("red","green"),cex=0.5
)
```




```
#chi square test Does the staus of country affect Diphtheria coverage significantly ?

#function to compare x value calculated and table x value
acceptChi<-function(Xcal,alpha,df){
  XTable<-qchisq(p=alpha,df,lower.tail=FALSE)
  if(abs(Xcal)<=abs(XTable)){
    cat("The calculated X value is ", abs(Xcal)," is less than the table X value ",abs(XT
able),"\n")
    print("There isn't much difference in the groups of data ")
  }else{
    cat("The calculated X value is ", abs(Xcal)," is greater than the table X value ",abs
(XTable),"\n")
    print("There is a significant difference in the groups of data")
  }
}


df<-y[c(4,5)]
df<-data.frame(df)
rownames(df) <- c("Developed","Developing")
#class(df)
#dim(df)
#str(df)
res<-chisq.test(df,correct=FALSE)
res

##
##  Pearson's Chi-squared test
##
## data:  df
## X-squared = 25.128, df = 1, p-value = 5.364e-07
```

```r
acceptChi(res$statistic,0.05,res$parameter)
```

```
## The calculated X value is  25.12823  is greater than the table X value  3.841459
## [1] "There is a significant difference in the groups of data"
```

```r
# Hence development status affect the Diphtheria coverage
```

```r
#correlation
data_corr <- lifexp %>% select_if(is.numeric)
coef<-cor(data_corr,method = "pearson")
View(coef)
```

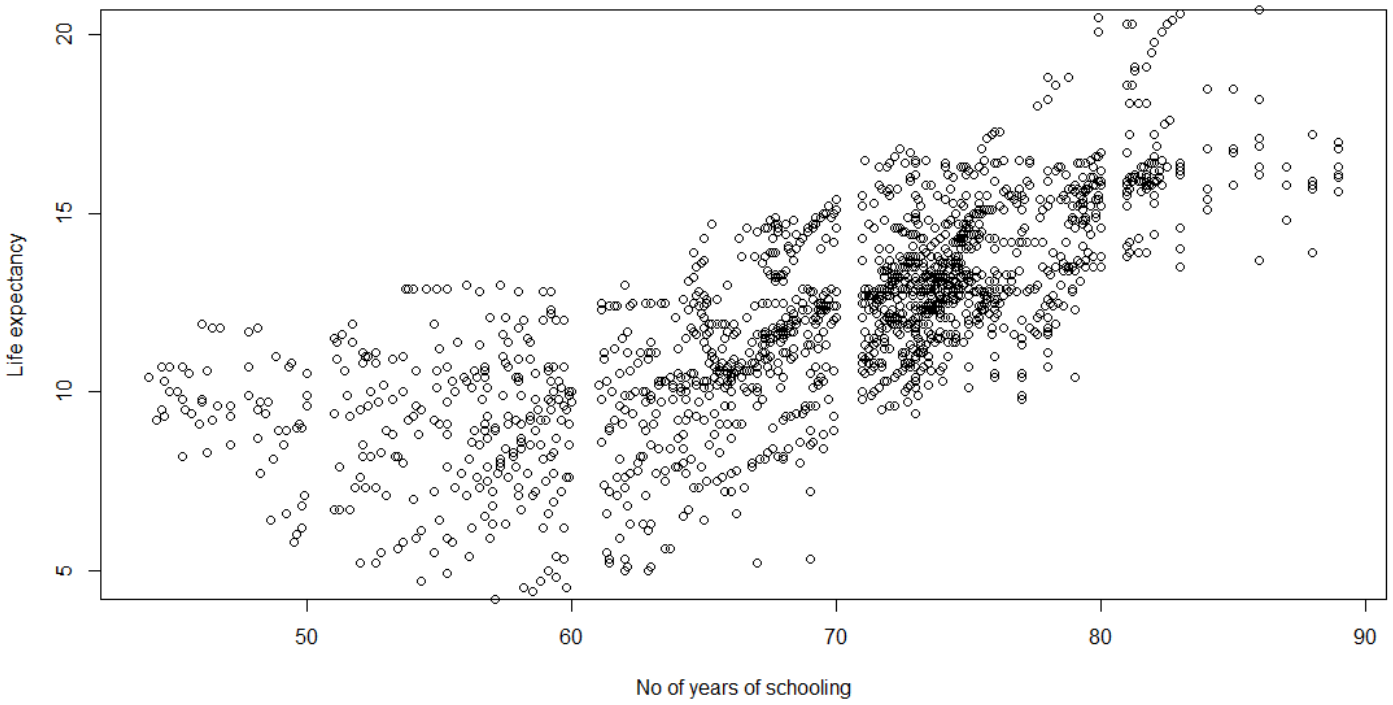
Filter

	Year	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure	Hepatitis.B	Measles	BMI	under.five.deaths	Polio
Year	1.000000000	0.05077103	-0.037091782	0.008029128	-0.11336476	0.06955347	0.11489709	-0.053822046	0.005739061	0.01047859	-0.01669880
Life.expectancy	0.050771035	1.00000000	-0.702523062	-0.169073804	0.40271832	0.40963082	0.19993528	-0.068881222	0.542041588	-0.19226530	0.32729440
Adult.Mortality	-0.037091782	-0.70252306	1.000000000	0.042450237	-0.17553509	-0.23760989	-0.10522544	-0.003966685	-0.351542478	0.06036503	-0.19985300
infant.deaths	0.008029128	-0.16907380	0.042450237	1.000000000	-0.10621692	-0.09076463	-0.23176894	0.532679832	-0.234425154	0.99690562	-0.15692881
Alcohol	-0.113364764	0.40271832	-0.175535086	-0.106216917	1.00000000	0.41704736	0.10988939	-0.050110235	0.353396205	-0.10108216	0.24031453
percentage.expenditure	0.069553468	0.40963082	-0.237609890	-0.090764632	0.41704736	1.00000000	0.01676017	-0.063070789	0.242738243	-0.09215806	0.12862605
Hepatitis.B	0.114897092	0.19993528	-0.105225443	-0.231768937	0.10988939	0.01676017	1.00000000	-0.124799993	0.143301786	-0.24076603	0.46333080
Measles	-0.053822046	-0.06888122	-0.003966685	0.532679832	-0.05011023	-0.06307079	-0.12479999	1.000000000	-0.153245464	0.51750556	-0.05785013
BMI	0.005739061	0.54204159	-0.351542478	-0.234425154	0.35339621	0.24273824	0.14330179	-0.153245464	1.000000000	-0.24213740	0.18626797
under.five.deaths	0.010478594	-0.19226530	0.060365026	0.996905622	-0.10108216	-0.09215806	-0.24076603	0.517505563	-0.242137398	1.00000000	-0.17116419
Polio	-0.016698803	0.32729440	-0.199853000	-0.156928805	0.24031453	0.12862605	0.46333080	-0.057850133	0.186267965	-0.17116419	1.00000000
Total.expenditure	0.059492777	0.17471764	-0.085226535	-0.146951117	0.21488509	0.18387236	0.11332668	-0.113582738	0.189468964	-0.14580310	0.11976798
Diphtheria	0.029640586	0.34133123	-0.191428759	-0.161871004	0.24295143	0.13481324	0.58898993	-0.058605907	0.176294503	-0.17844819	0.60924547
HIV.AIDS	-0.123404990	-0.59223629	0.550690745	0.007711547	-0.02711264	-0.09508499	-0.09480197	-0.003521854	-0.210896746	0.01947593	-0.10788547
GDP	0.096421485	0.44132181	-0.255034733	-0.098092020	0.44343279	0.95929886	0.04184950	-0.064767590	0.266113973	-0.10033126	0.15680869
Population	0.012566893	-0.02230498	-0.015011838	0.671758310	-0.02888023	-0.01679214	-0.12972265	0.321946377	-0.081415982	0.65867969	-0.04538657
thinness..1.19.years	0.019756611	-0.45783819	0.272230044	0.463415256	-0.40375499	-0.25503460	-0.12940595	0.180641506	-0.547017514	0.46478470	-0.16406959
thinness.5.9.years	0.014122422	-0.45750829	0.286722882	0.461907925	-0.38620819	-0.25563544	-0.13325099	0.174946217	-0.554093981	0.46228938	-0.17448925

```r
#correlation btw life expectancy and schooling is v high
coef<-cor(data_corr$Life.expectancy,data_corr$Schooling, method = "pearson")
cat("Pearson correlation between Life Expectancy and Schooling: ",coef)
```

```
## Pearson correlation between Life Expectancy and Schooling:  0.72763
```

```r
plot(data_corr$Life.expectancy,data_corr$Schooling,xlab="No of years of schooling",ylab="
Life expectancy");
```




```
pcor.test(data_corr$Life.expectancy,data_corr$Schooling,data_corr$Income.composition.of.resources,method = c("pearson"))$estimate
```

```
## [1] 0.3766893
```

```
print("large decrease in correlation and hence life expectancy and schooling have spurious relationship")
```

```
## [1] "large decrease in correlation and hence life expectancy and schooling have spurious relationship"
```

```
coef<-cor(data_corr$Life.expectancy,data_corr$Income.composition.of.resources, method = "pearson")
coef
```

```
## [1] 0.7210826
```

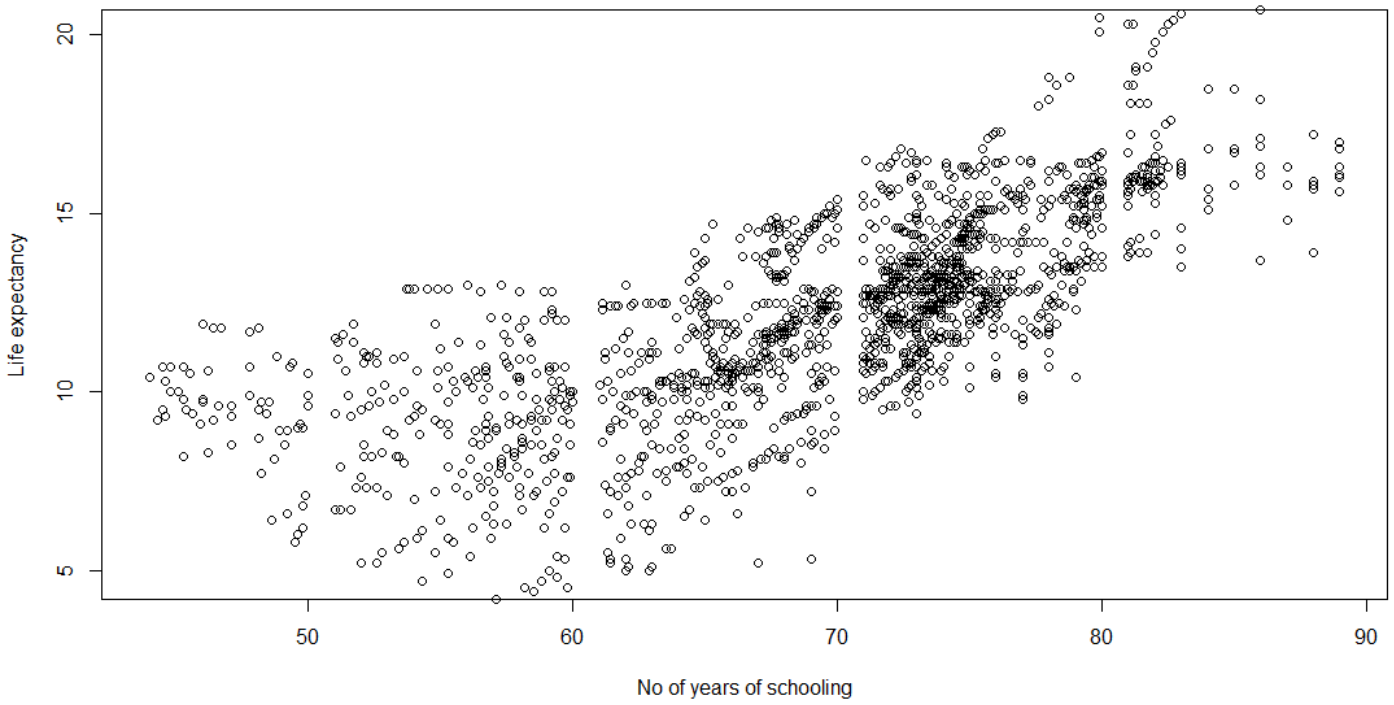
```
pcor.test(data_corr$Life.expectancy,data_corr$Income.composition.of.resources,data_corr$GDP,method = c("pearson"))$estimate
```

```
## [1] 0.6525817
```

```
print("not a major difference hance life expectancy and income composition share a direct realtionship")
```

```
## [1] "not a major difference hance life expectancy and income composition share a direct realtionship"
```

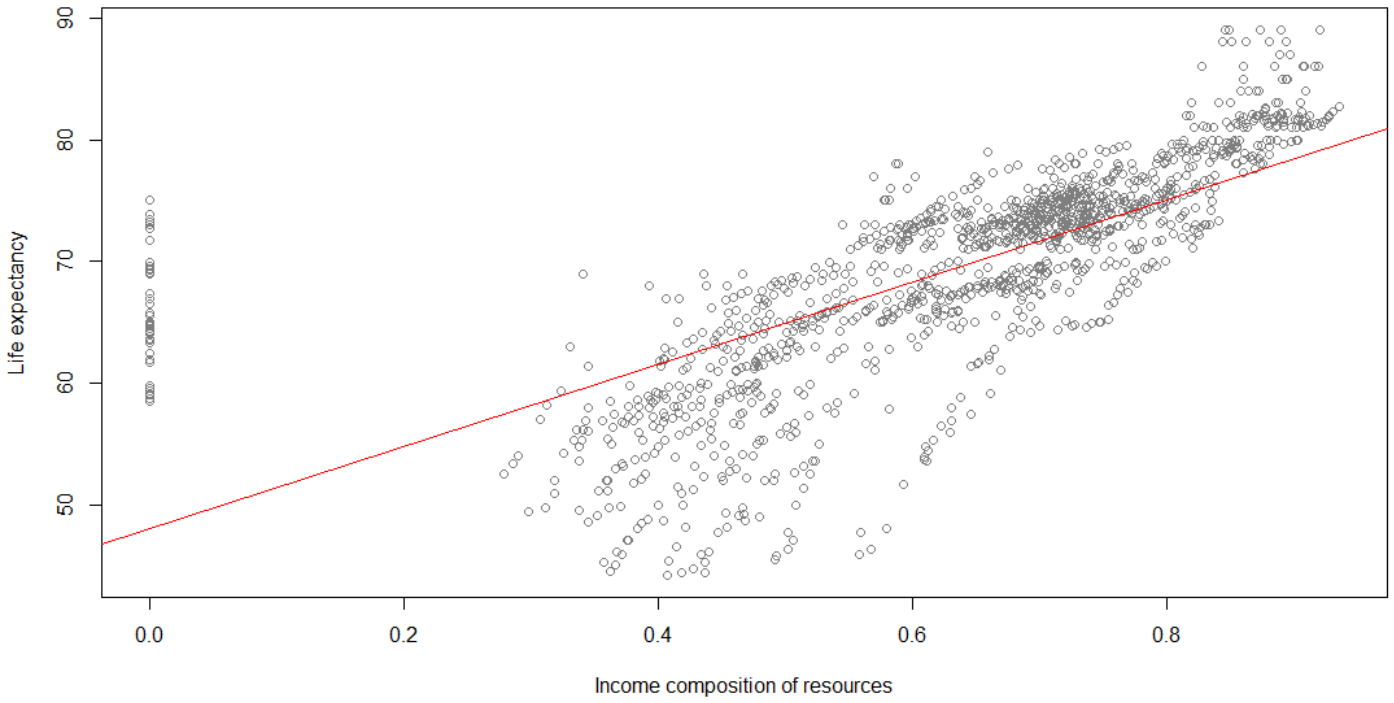
```
model<-lm(Life.expectancy ~ Income.composition.of.resources,data=lifexp)
plot(Life.expectancy ~ Income.composition.of.resources,data=lifexp,
     xlab="Income composition of resources",ylab="Life expectancy",col="gray50");
abline(reg=lm(Life.expectancy~Income.composition.of.resources,data=lifexp),col="red")
```




```
coeffs<-coefficients(model)
coeffs
```

```
##                    (Intercept) Income.composition.of.resources
##                       47.42175                          34.64574
```

```
#divide the datset into 2 for prediction

set.seed(42)
rows <- sample(nrow(lifexp))

lifexp_100 <- lifexp[rows, ]
split <- round(nrow(lifexp_100) * 0.80)
lifexp_80<-lifexp_100[1:split, ]
lifexp_20<-lifexp_100[(split + 1):nrow(lifexp_100), ]
View(lifexp_80)
View(lifexp_20)


#linear regression

model<-lm(Life.expectancy ~ Income.composition.of.resources,data=lifexp_80)
plot(Life.expectancy ~ Income.composition.of.resources,data=lifexp_80,
     xlab="Income composition of resources",ylab="Life expectancy",col="gray50");
abline(reg=model,col="red")
```



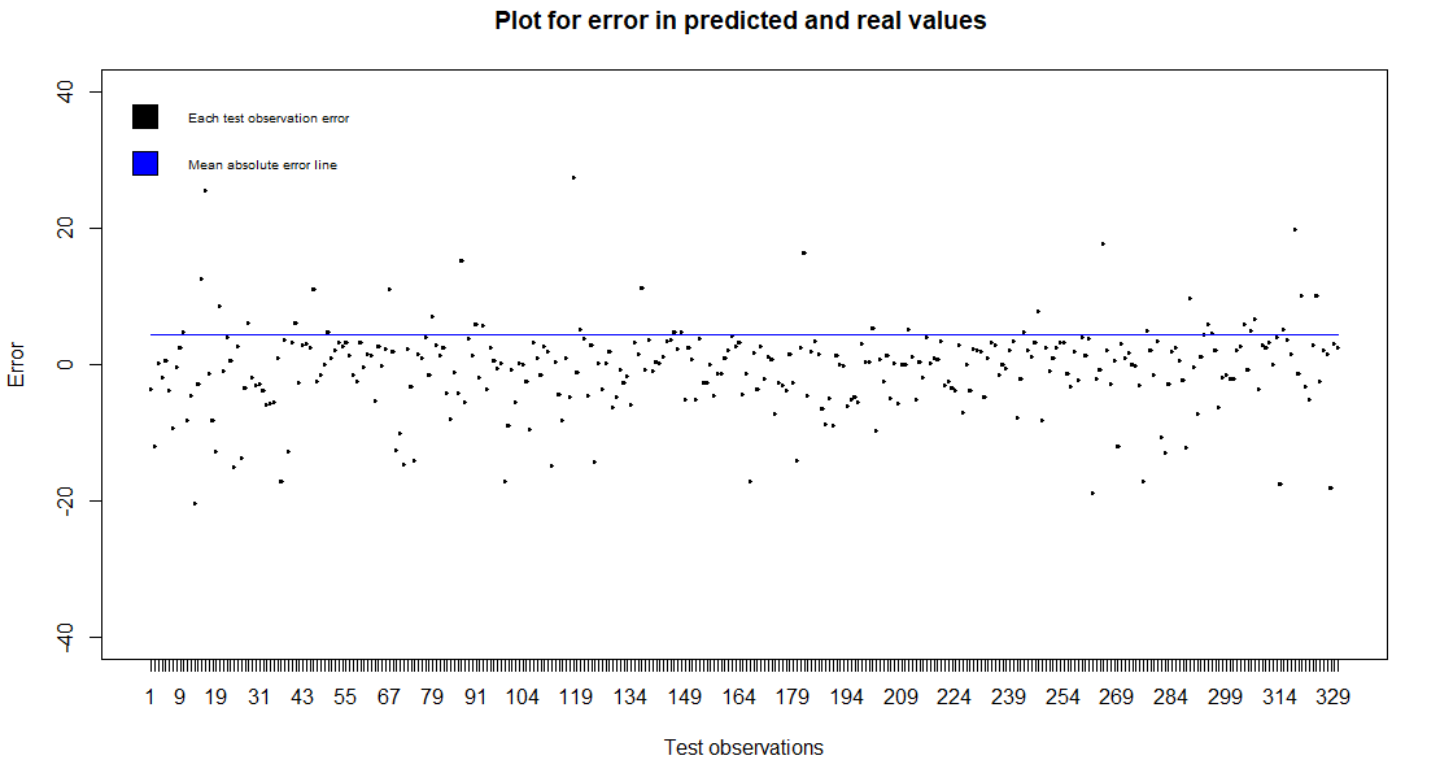

```
coeffs<-coefficients(model)
coeffs
```

```
##                     (Intercept) Income.composition.of.resources
##                        48.14638                         33.69854
```

```
predicted<-predict(model,lifexp_20)

View(predicted)

error_info = data.frame(real =lifexp_20$Life.expectancy,prediction = predicted)

error_info<- error_info %>% mutate(error=error_info[,1] - error_info[,2] );
View(error_info)
mean_abs_err = mean(abs(error_info$error))
mean_abs_err
```

```
## [1] 4.290697
```

```
x<- as.factor(seq(1:330))
plot(x,error_info$error,type = "l",pch=1,ylim=c(-40,40),xlab="Test observations",ylab = "
Error",main="Plot for error in predicted and real values ")
lines(x,rep(mean_abs_err,330),col="Blue")
legend("topleft",
       c("Each test observation error","Mean absolute error line"),
       fill = c("Black","blue"),cex=0.65,bty="n"
)
```

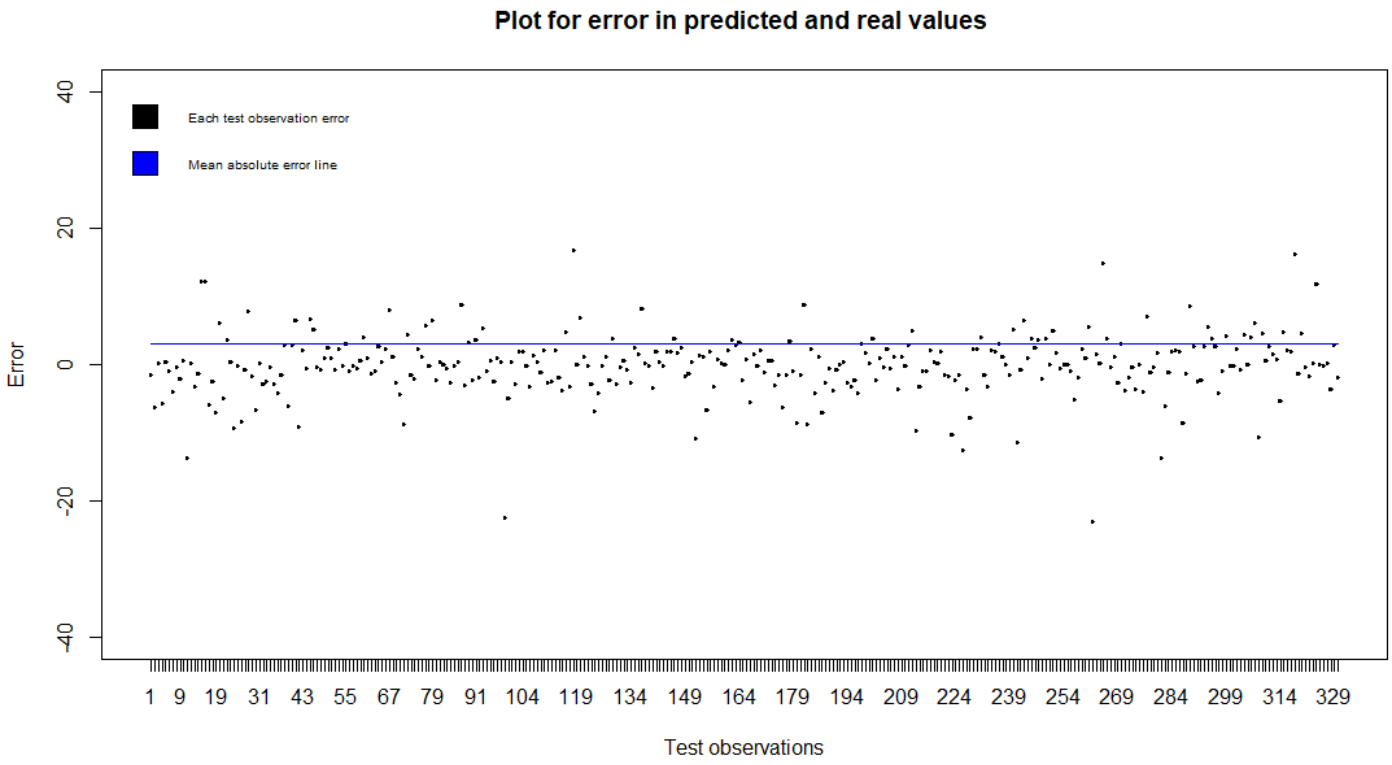



```
rmse = sqrt(mean(error_info$error ^ 2))
rmse

## [1] 6.193577

#multiple regression

my_model<-lm(Life.expectancy ~ Income.composition.of.resources+Adult.Mortality+Alcohol+GD
P+thinness..1.19.years+thinness.5.9.years,data=lifexp_80)
#plot(Life.expectancy ~ Income.composition.of.resources,data=lifexp_80,
 #    xlab="Income composition of resources",ylab="Life expectancy",col="gray50");
#abline(reg=my_model,col="red")
coeffs<-coefficients(my_model)
coeffs

##                       (Intercept) Income.composition.of.resources
##                      6.274058e+01                      2.016409e+01
##                   Adult.Mortality                           Alcohol
##                     -3.139423e-02                     -3.081762e-02
##                               GDP              thinness..1.19.years
##                      8.158393e-05                     -1.363737e-01
##               thinness.5.9.years
##                     -1.037149e-01

predicted<-predict(my_model,lifexp_20)

#View(predicted)

error_info = data.frame(real =lifexp_20$Life.expectancy, prediction = predicted)


error_info<- error_info %>% mutate(error=error_info[,1] - error_info[,2] );
View(error_info)
```

```r
mean_abs_err = mean(abs(error_info$error))
mean_abs_err
```

```
## [1] 3.067707
```

```r
x<- as.factor(seq(1:330))
plot(x,error_info$error,type = "l",pch=1,ylim=c(-40,40),xlab="Test observations",ylab = "
Error",main="Plot for error in predicted and real values ")
lines(x,rep(mean_abs_err,330),col="Blue")
legend("topleft",
       c("Each test observation error","Mean absolute error line"),
       fill = c("Black","blue"),cex=0.65,bty="n"
)
```




```r
rmse = sqrt(mean(error_info$error ^ 2))
rmse
```

```
## [1] 4.542652
```