FinalAnalysis

2023-12-08

```r
#import global libraries:
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
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
```

```r
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(MASS)
```

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

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

```r
library(faraway)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma
```

```r
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```r
library(tidyr)
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-8
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

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

```r
set.seed(123)#set random seed
```

```r
#load data from given file path
file_path <- "Life Expectancy Data.csv"
Data <- read.csv(file_path)

#display the structure of data, and check if there's empty cells
str(Data)
```

```
## '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 ...
```

```r
colSums(is.na(Data))
```

```
##                         Country                            Year
##                               0                               0
##                          Status                 Life.expectancy
##                               0                              10
##                 Adult.Mortality                   infant.deaths
##                              10                               0
##                         Alcohol          percentage.expenditure
##                             194                               0
##                     Hepatitis.B                         Measles
##                             553                               0
##                             BMI               under.five.deaths
##                              34                               0
##                           Polio               Total.expenditure
##                              19                             226
##                      Diphtheria                        HIV.AIDS
##                              19                               0
##                             GDP                      Population
##                             448                             652
##            thinness..1.19.years              thinness.5.9.years
##                              34                              34
## Income.composition.of.resources                       Schooling
##                             167                             163
```

```r
#data cleaning: for columns with more than 200 null values, fill empty cells with median of this column
Data$'GDP' <- ifelse(is.na(Data$'GDP'),
                     median(Data$'GDP', na.rm = TRUE), Data$'GDP')
Data$'Population' <- ifelse(is.na(Data$'Population'),
                            median(Data$'Population', na.rm = TRUE), Data$'Population')
Data$'Total.expenditure' <- ifelse(is.na(Data$'Total.expenditure'),
                                   median(Data$"Total.expenditure", na.rm = TRUE), Data$'Total.expenditu
Data$'Hepatitis B' <- ifelse(is.na(Data$'Hepatitis.B'),
                             median(Data$'Hepatitis.B', na.rm = TRUE), Data$'Hepatitis.B')

#for status - developed/undeveloped, replace them with 0 and 1
Data$'Status' <- ifelse(Data$'Status' == "Developed", 1, 0)
```

```r
#remove other rows with empty cells
Data <- na.omit(Data)
Data <- unique(Data)

colSums(is.na(Data))
```

```
##                         Country                            Year 
##                               0                               0 
##                          Status                 Life.expectancy 
##                               0                               0 
##                 Adult.Mortality                   infant.deaths 
##                               0                               0 
##                         Alcohol          percentage.expenditure 
##                               0                               0 
##                     Hepatitis.B                         Measles 
##                               0                               0 
##                             BMI               under.five.deaths 
##                               0                               0 
##                           Polio               Total.expenditure 
##                               0                               0 
##                      Diphtheria                        HIV.AIDS 
##                               0                               0 
##                             GDP                      Population 
##                               0                               0 
##             thinness..1.19.years              thinness.5.9.years 
##                               0                               0 
## Income.composition.of.resources                       Schooling 
##                               0                               0 
##                     Hepatitis B 
##                               0
```

```r
nrow(Data)
```

```
## [1] 2088
```

```r
#choose random 1000 observations for this analysis
Data <- sample_n(Data, 1000)
nrow(Data)
```

```
## [1] 1000
```

```r
#summarize data and display first few rows
summary(Data)
```

```
##    Country               Year          Status      Life.expectancy
##  Length:1000        Min.   :2000   Min.   :0.000   Min.   :44.00  
##  Class :character   1st Qu.:2004   1st Qu.:0.000   1st Qu.:65.47  
##  Mode  :character   Median :2008   Median :0.000   Median :72.30  
##                     Mean   :2008   Mean   :0.134   Mean   :69.98  
##                     3rd Qu.:2011   3rd Qu.:0.000   3rd Qu.:75.00  
##                     Max.   :2015   Max.   :1.000   Max.   :89.00  
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.00      
##  1st Qu.: 79.0   1st Qu.:   0.00   1st Qu.: 0.640   1st Qu.:   18.27      
##  Median :143.0   Median :   3.00   Median : 3.075   Median :  100.85      
##  Mean   :160.4   Mean   :  26.70   Mean   : 4.326   Mean   :  688.66      
```

```
##  3rd Qu.:217.0   3rd Qu.:  16.25   3rd Qu.: 7.213   3rd Qu.:  512.78
##  Max.   :723.0   Max.   :1500.00   Max.   :17.870   Max.   :18961.35
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 2.00   Min.   :      0.0   Min.   : 1.40   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:      0.0   1st Qu.:21.27   1st Qu.:   1.00
##  Median :92.00   Median :     12.0   Median :45.30   Median :   3.00
##  Mean   :80.95   Mean   :  2220.6   Mean   :39.28   Mean   :  35.98
##  3rd Qu.:96.25   3rd Qu.:   279.2   3rd Qu.:56.33   3rd Qu.:  20.00
##  Max.   :99.00   Max.   :124219.0   Max.   :76.70   Max.   :2000.00
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.740    Min.   : 4.00   Min.   : 0.10
##  1st Qu.:83.00   1st Qu.: 4.200    1st Qu.:83.00   1st Qu.: 0.10
##  Median :94.00   Median : 5.660    Median :94.00   Median : 0.10
##  Mean   :84.92   Mean   : 5.759    Mean   :85.38   Mean   : 1.76
##  3rd Qu.:97.00   3rd Qu.: 7.190    3rd Qu.:97.00   3rd Qu.: 0.40
##  Max.   :99.00   Max.   :14.390    Max.   :99.00   Max.   :50.60
##       GDP              Population        thinness..1.19.years
##  Min.   :     5.67   Min.   :3.600e+01   Min.   : 0.100
##  1st Qu.:   596.78   1st Qu.:3.684e+05   1st Qu.: 1.800
##  Median :  1766.95   Median :1.387e+06   Median : 3.450
##  Mean   :  6227.53   Mean   :1.260e+07   Mean   : 4.909
##  3rd Qu.:  4772.94   3rd Qu.:4.410e+06   3rd Qu.: 7.000
##  Max.   :115761.58   Max.   :1.294e+09   Max.   :27.200
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.0
##  1st Qu.: 1.80      1st Qu.:0.5450                  1st Qu.:10.6
##  Median : 3.40      Median :0.6770                  Median :12.3
##  Mean   : 4.92      Mean   :0.6413                  Mean   :12.2
##  3rd Qu.: 6.90      3rd Qu.:0.7622                  3rd Qu.:14.0
##  Max.   :28.10      Max.   :0.9360                  Max.   :20.6
##   Hepatitis B
##  Min.   : 2.00
##  1st Qu.:77.00
##  Median :92.00
##  Mean   :80.95
##  3rd Qu.:96.25
##  Max.   :99.00
```

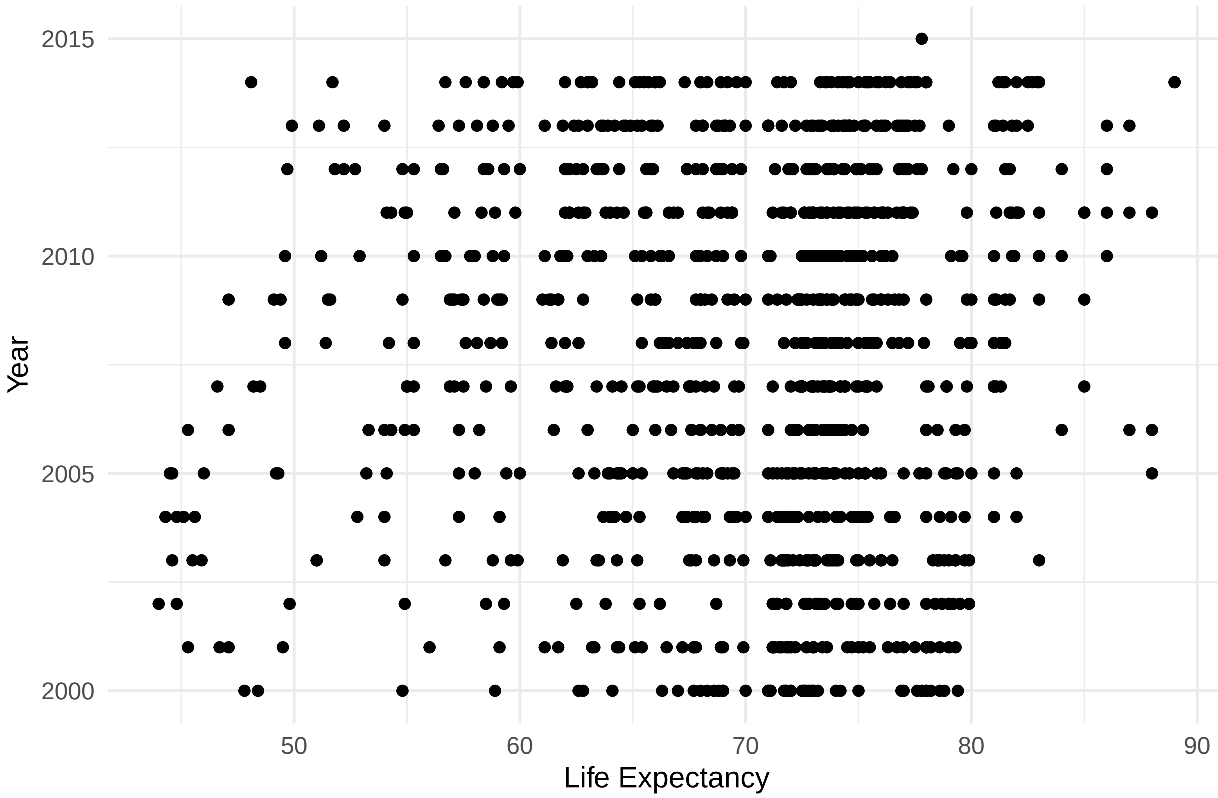
```
head(Data)
```

```
##                                    Country Year Status Life.expectancy
## 1                                   Cyprus 2012      1            80.0
## 2                                  Belarus 2010      0            73.0
## 3 The former Yugoslav republic of Macedonia 2013      0            75.3
## 4                                   Malawi 2004      0            45.1
## 5          Micronesia (Federated States of) 2000      0            67.0
## 6                                 Mongolia 2000      0            62.8
##   Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B
## 1              56             0   10.55            2159.756205          96
## 2             222             0   14.44               8.494095          96
## 3              14             0    1.03               0.000000          97
## 4             615            40    1.11              58.135833          89
## 5             185             0    2.23               0.000000          87
## 6             274             2    2.79              56.431387          93
##   Measles  BMI under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS
```

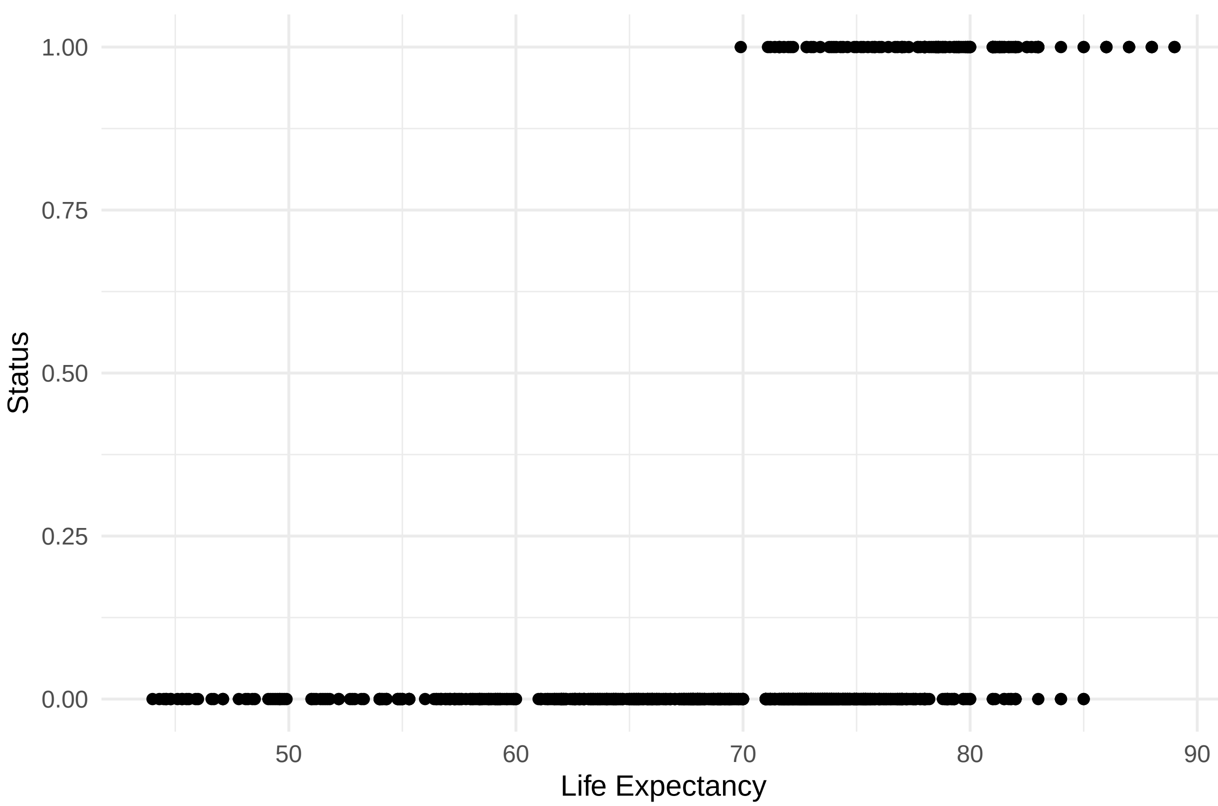
```
## 1       1 58.7                  0    99             7.44         99      0.1
## 2       1 59.3                  1    99             5.55         98      0.1
## 3       4 59.1                  0    98             6.70         98      0.1
## 4    1116 15.5                 65    94             7.82         89     23.4
## 5       0 61.5                  0    85             7.88         85      0.1
## 6     925 38.5                  3    94             4.92         94      0.1
##           GDP Population thinness..1.19.years thinness.5.9.years
## 1 28951.15556     113562                 0.9                1.0
## 2    63.38877     949583                 2.0                2.2
## 3  1766.94760    1386542                 2.2                2.2
## 4   274.22563    1267638                 7.5                7.4
## 5  1766.94760    1386542                 0.3                0.3
## 6   474.21334    2397436                 2.6                2.6
##   Income.composition.of.resources Schooling Hepatitis B
## 1                           0.850      13.8          96
## 2                           0.780      15.5          96
## 3                           0.741      12.9          97
## 4                           0.366      10.0          89
## 5                           0.000       0.0          87
## 6                           0.582       8.9          93
```

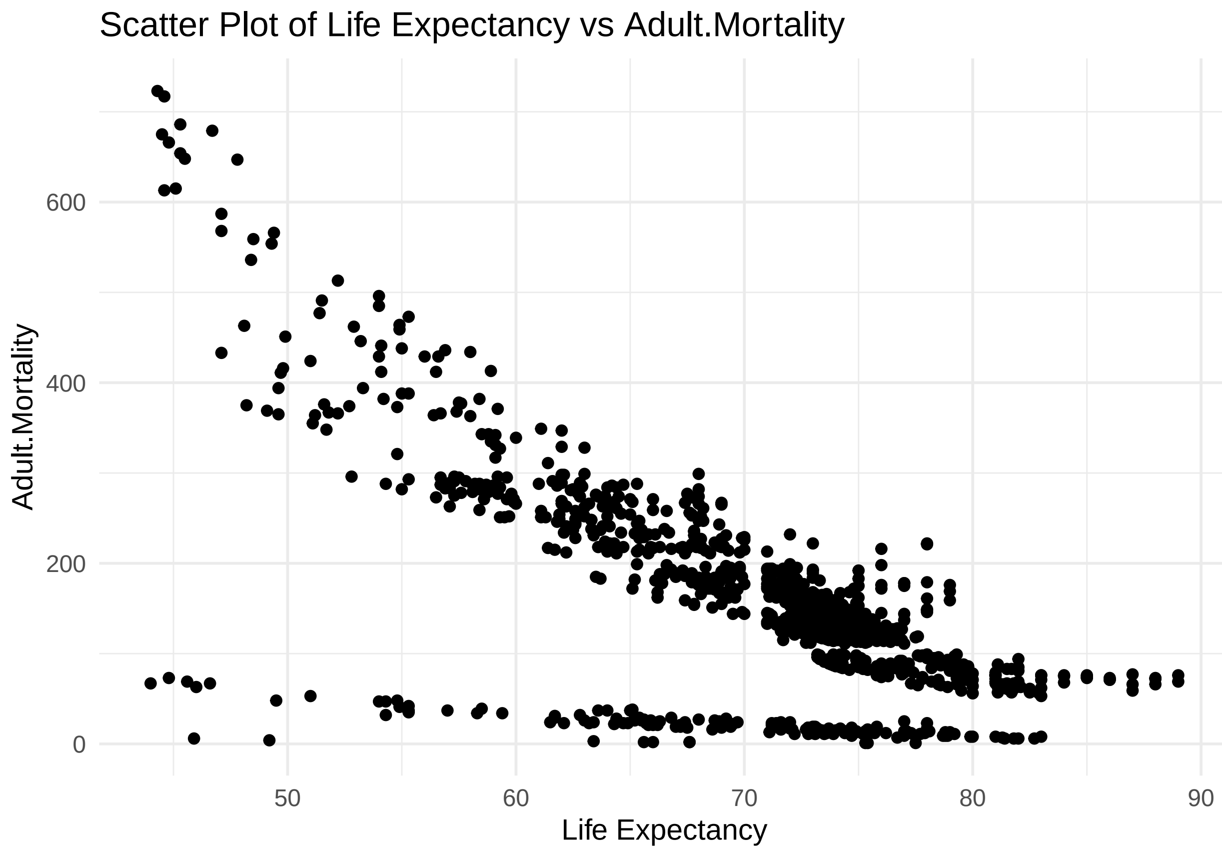
```
#visualize data:
#numeric predictors
numeric_columns <- names(Data[, sapply(Data, is.numeric) & names(Data) != "Life.expectancy"])

#plot relationship btw predictors and Life expectancy
plots <- lapply(numeric_columns, function(col) {
  print(ggplot(Data, aes(x = Life.expectancy, y = !!sym(col))) +
    geom_point() + xlab("Life Expectancy") + ylab(col) +
    ggtitle(paste("Scatter Plot of Life Expectancy vs", col)) +
    theme_minimal())
})
```

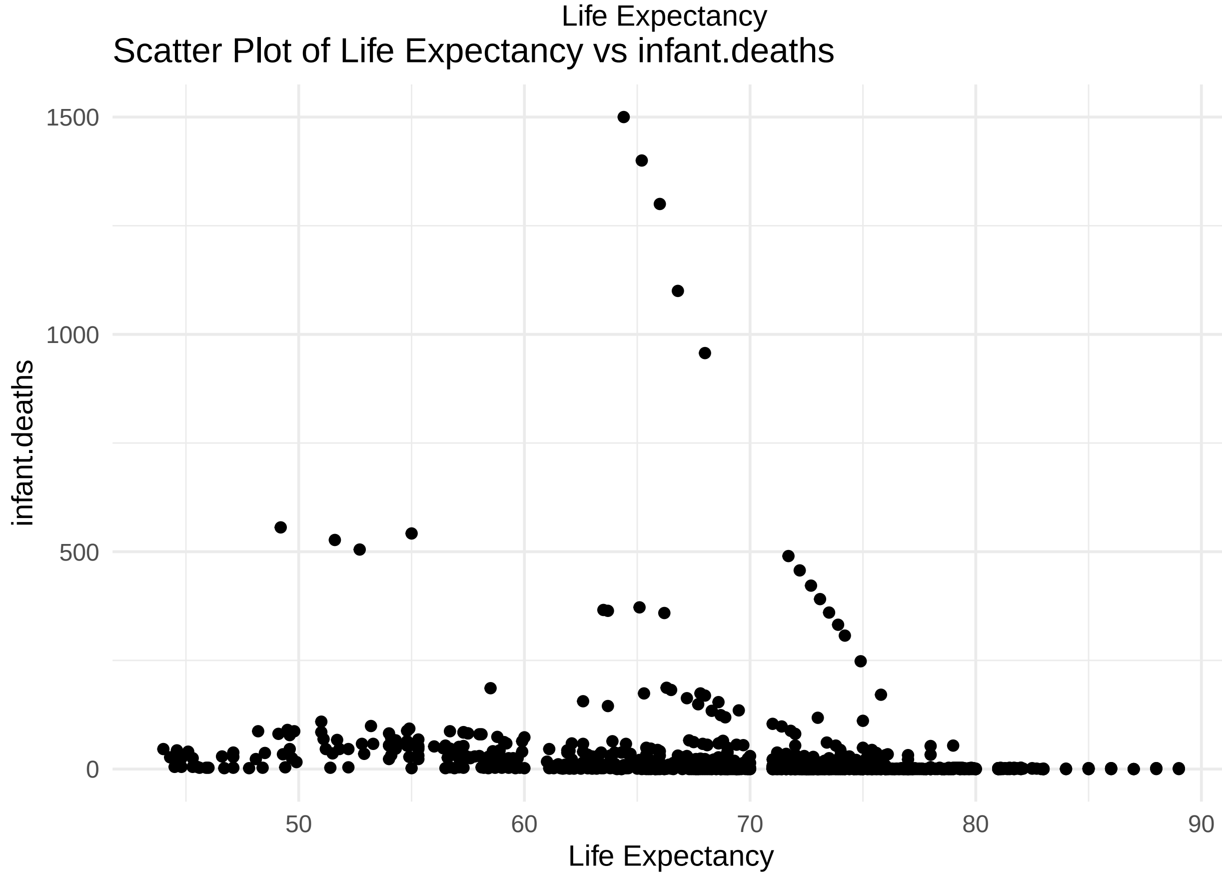
Scatter Plot of Life Expectancy vs Year

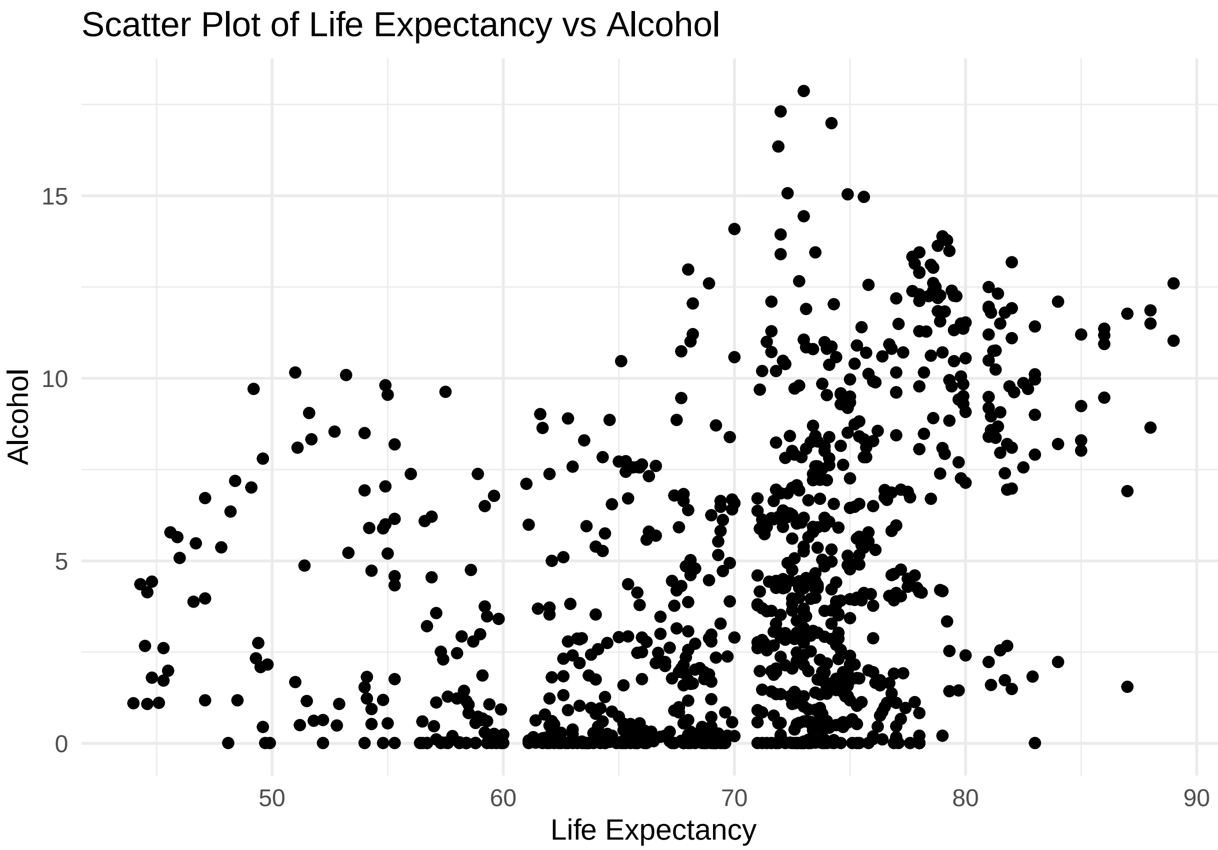


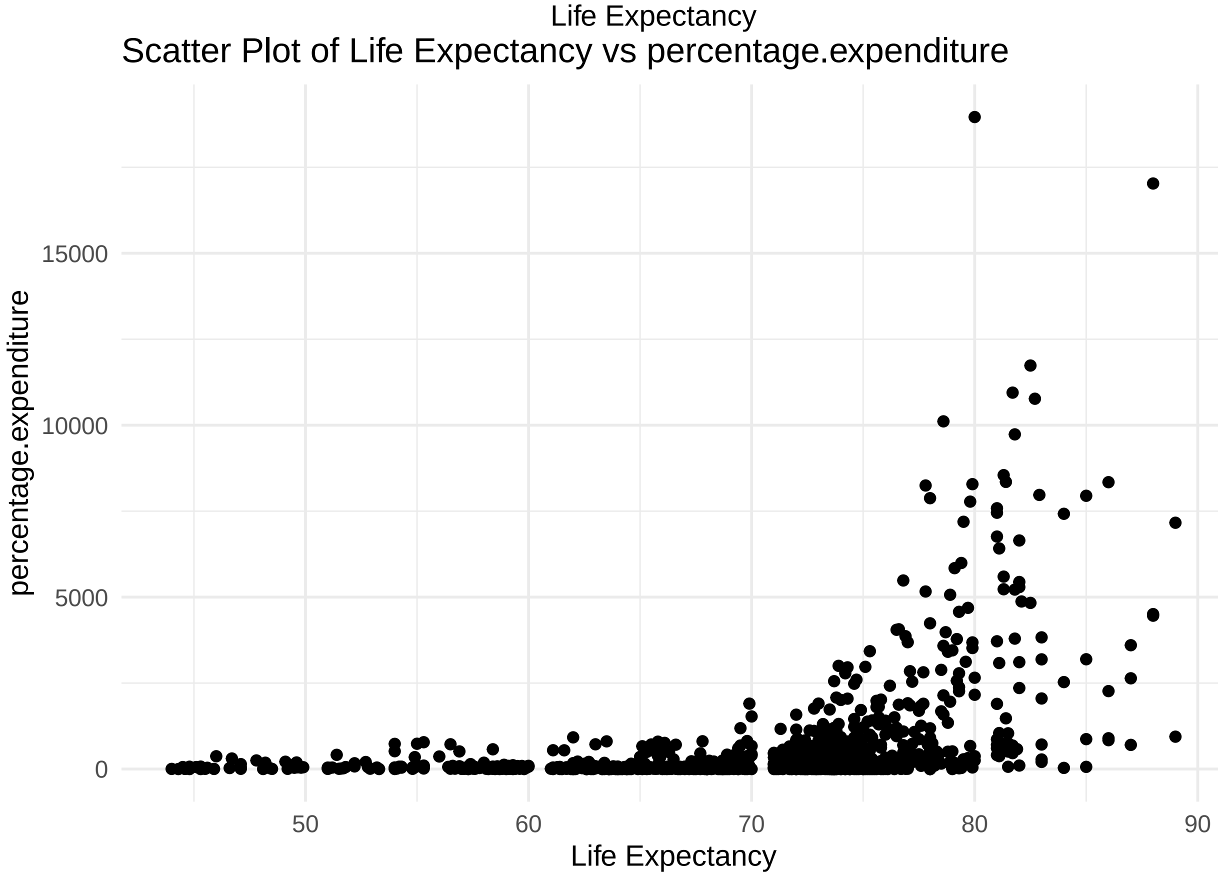
Scatter Plot of Life Expectancy vs Status

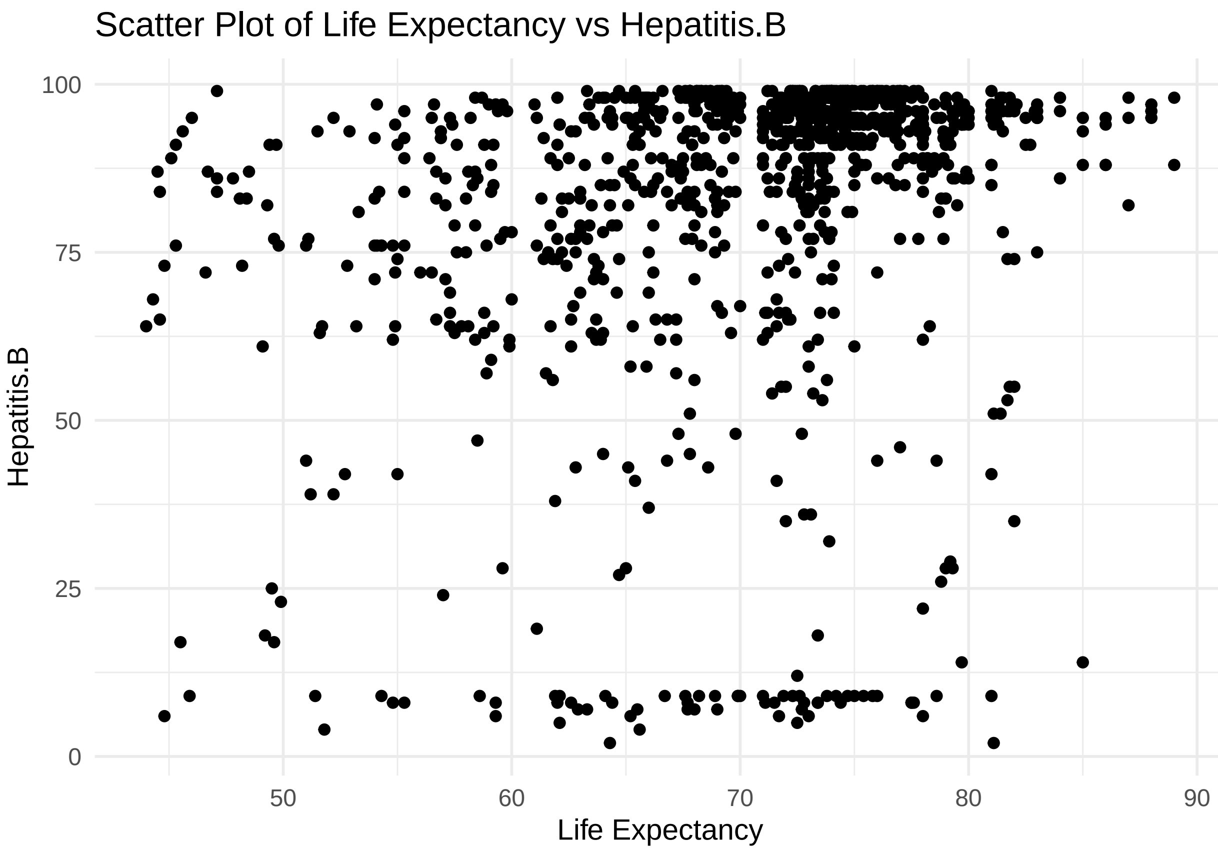


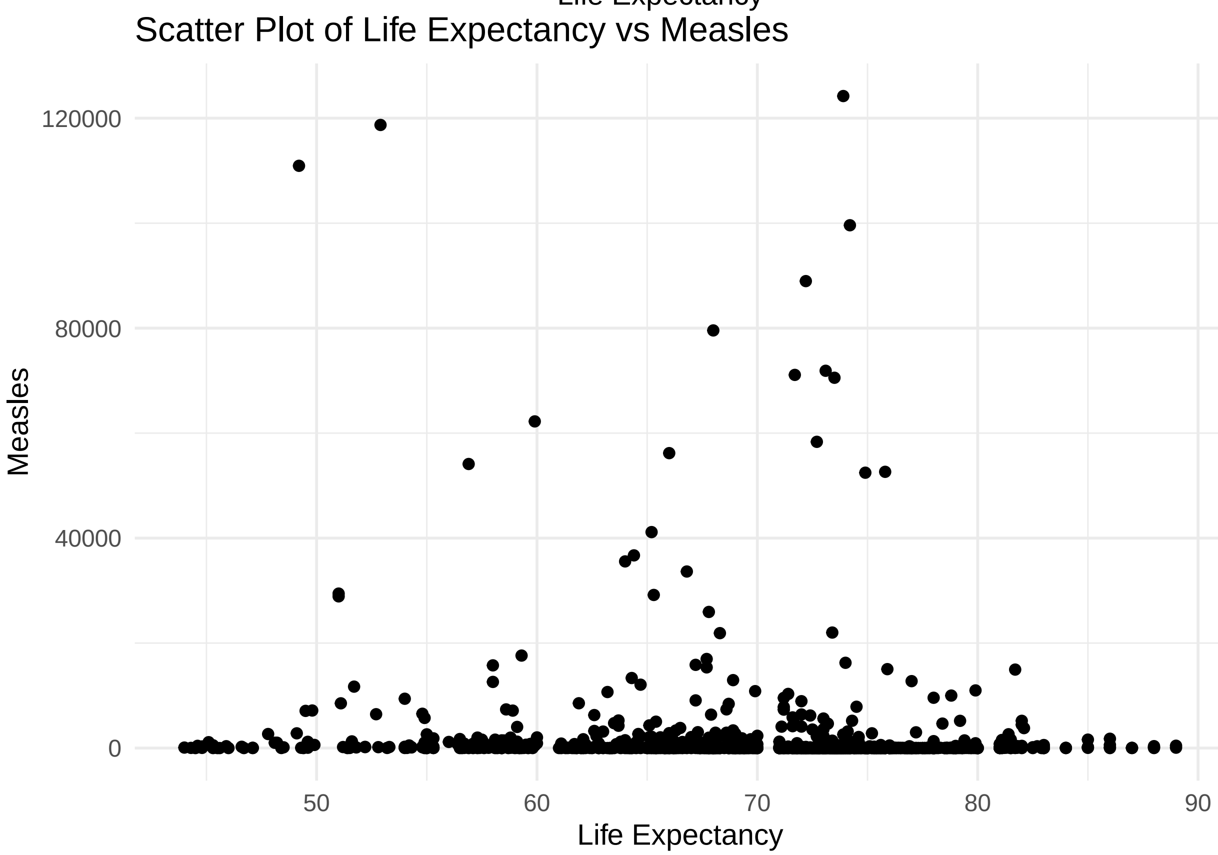

Scatter Plot of Life Expectancy vs Adult.Mortality
Adult.Mortality
600
400
200
0
50
60
70
80
90
Life Expectancy


Scatter Plot of Life Expectancy vs infant.deaths
infant.deaths
1500
1000
500
0
50
60
70
80
90
Life Expectancy

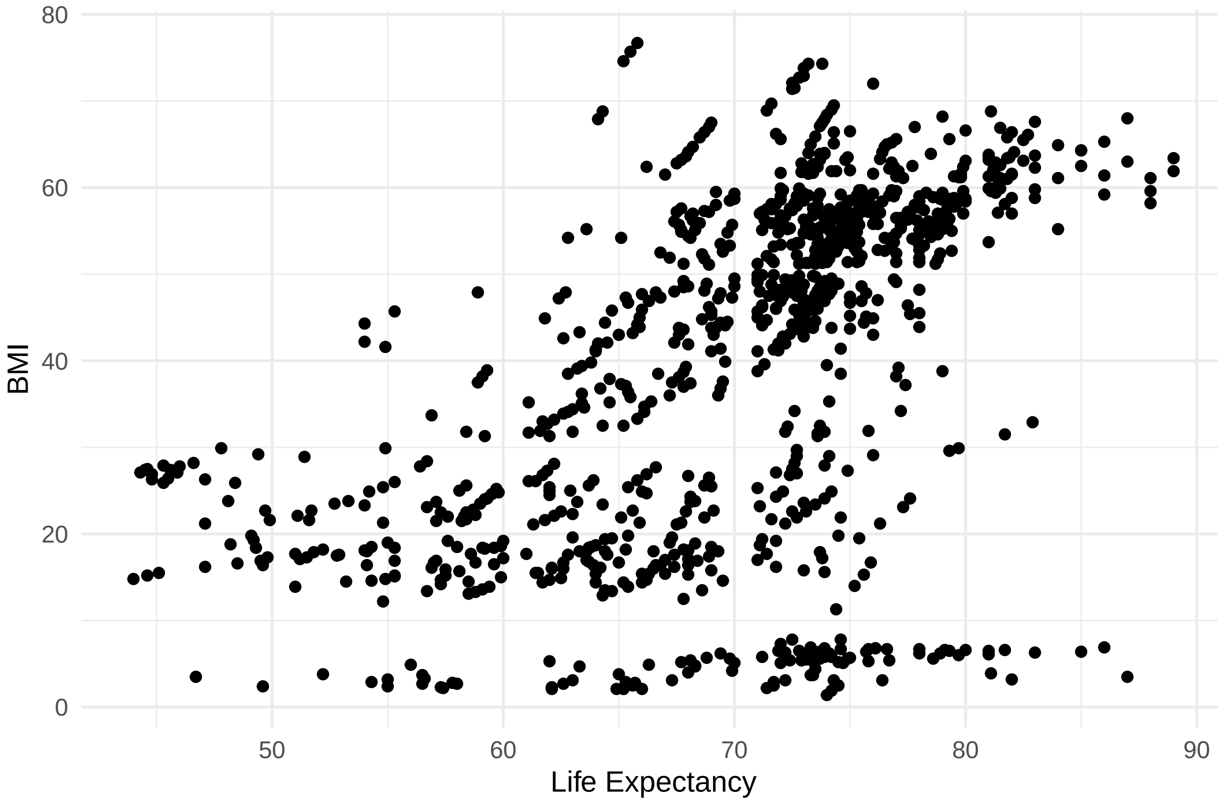

Scatter Plot of Life Expectancy vs Alcohol
Alcohol
0
5
10
15
50
60
70
80
90
Life Expectancy


Scatter Plot of Life Expectancy vs percentage.expenditure
percentage.expenditure
0
5000
10000
15000
50
60
70
80
90
Life Expectancy

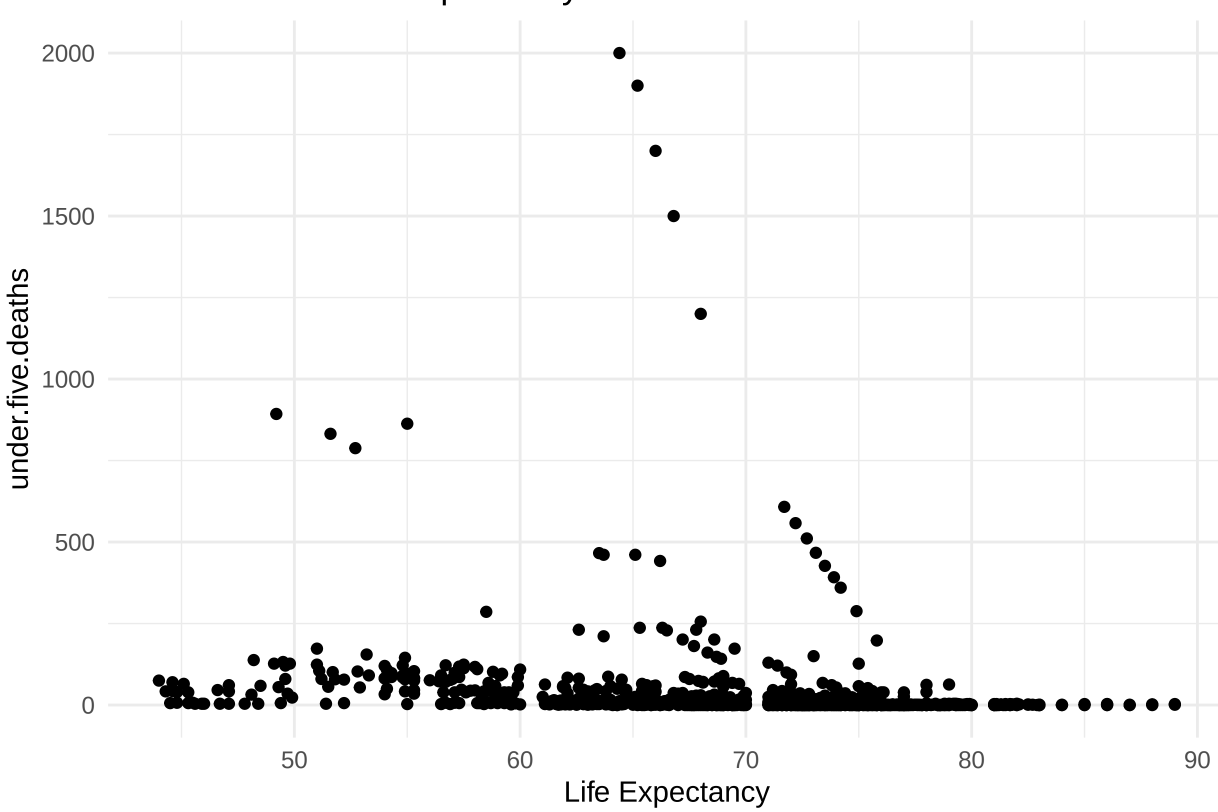

Scatter Plot of Life Expectancy vs Hepatitis.B
100
75
50
25
0
Hepatitis.B
50
60
70
80
90
Life Expectancy


Scatter Plot of Life Expectancy vs Measles
120000
80000
40000
0
Measles
50
60
70
80
90
Life Expectancy

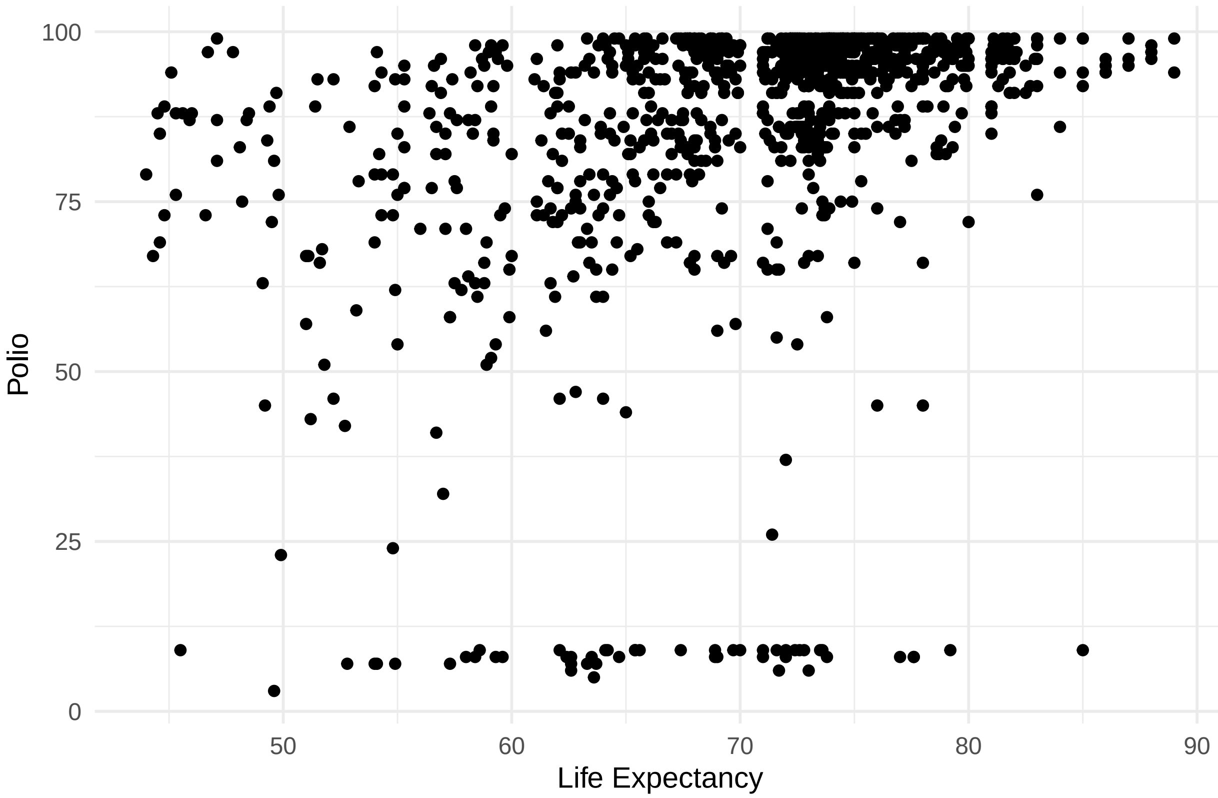
Scatter Plot of Life Expectancy vs BMI



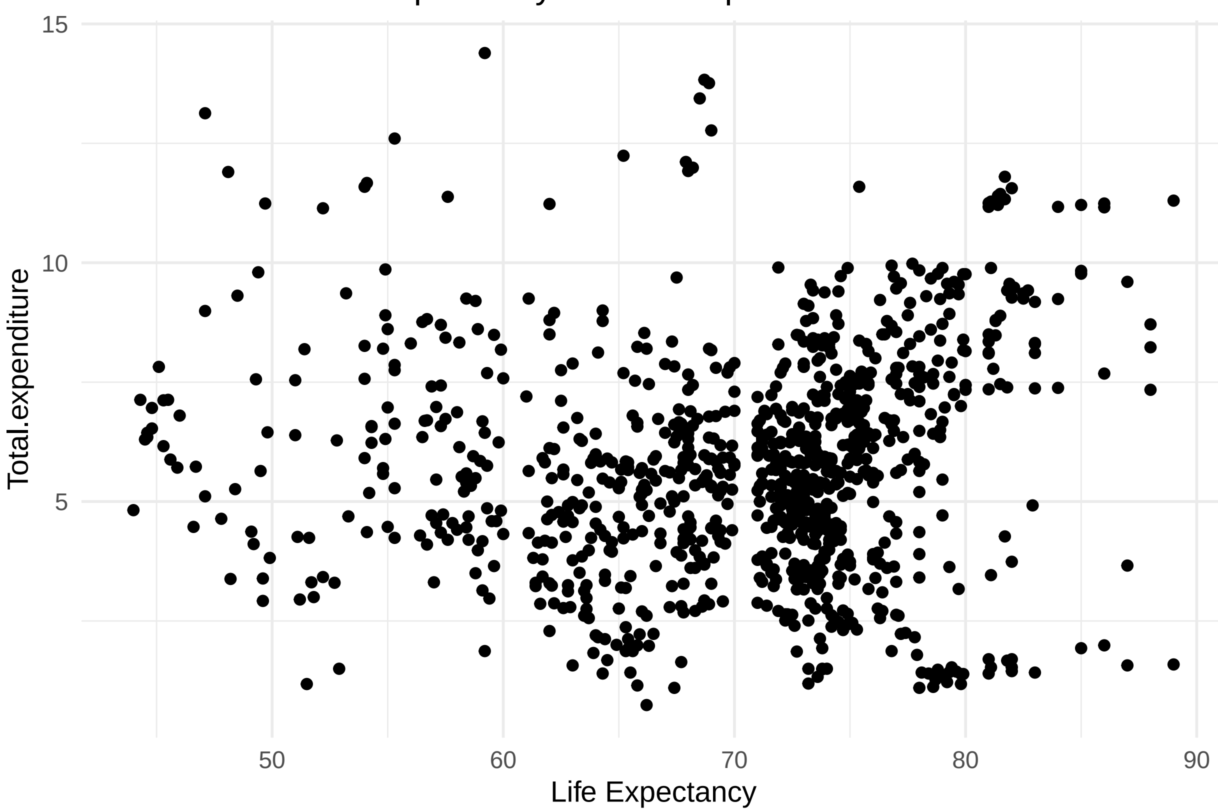
Scatter Plot of Life Expectancy vs under.five.deaths

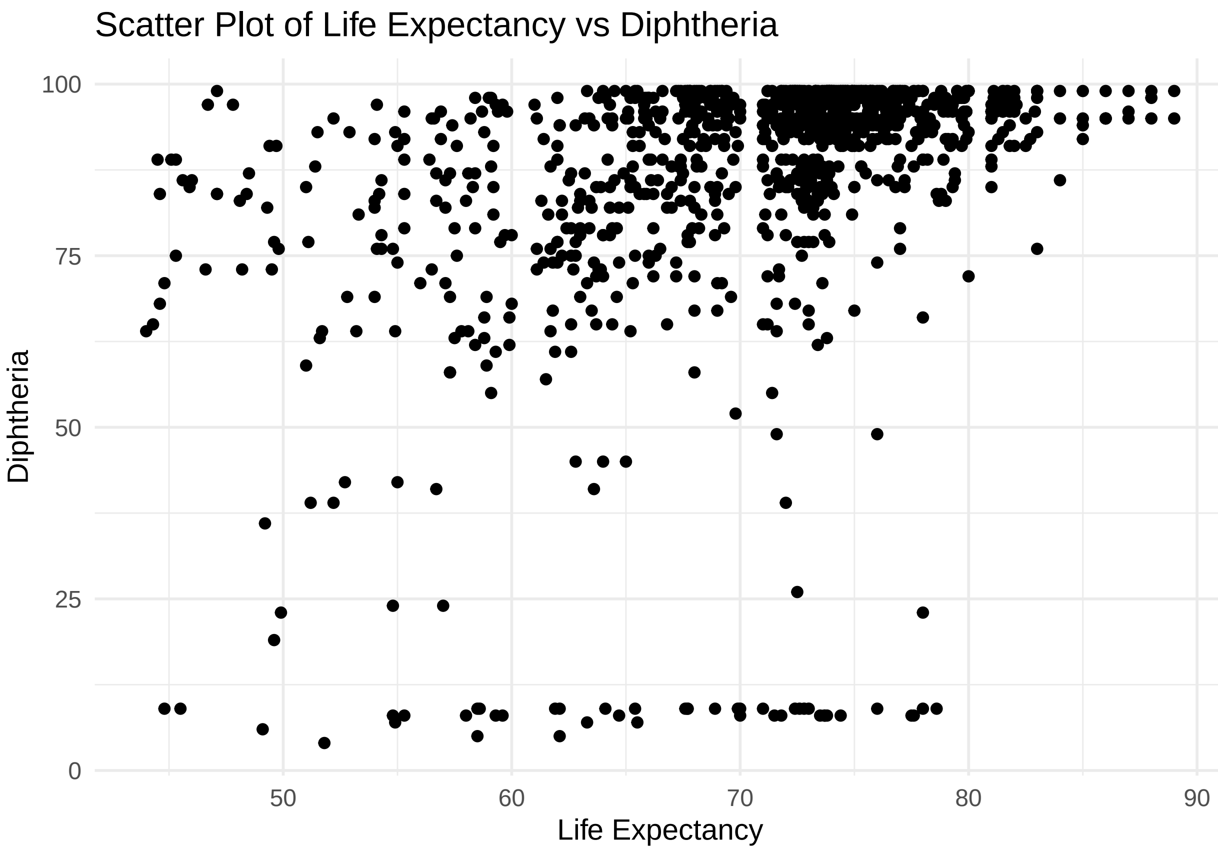


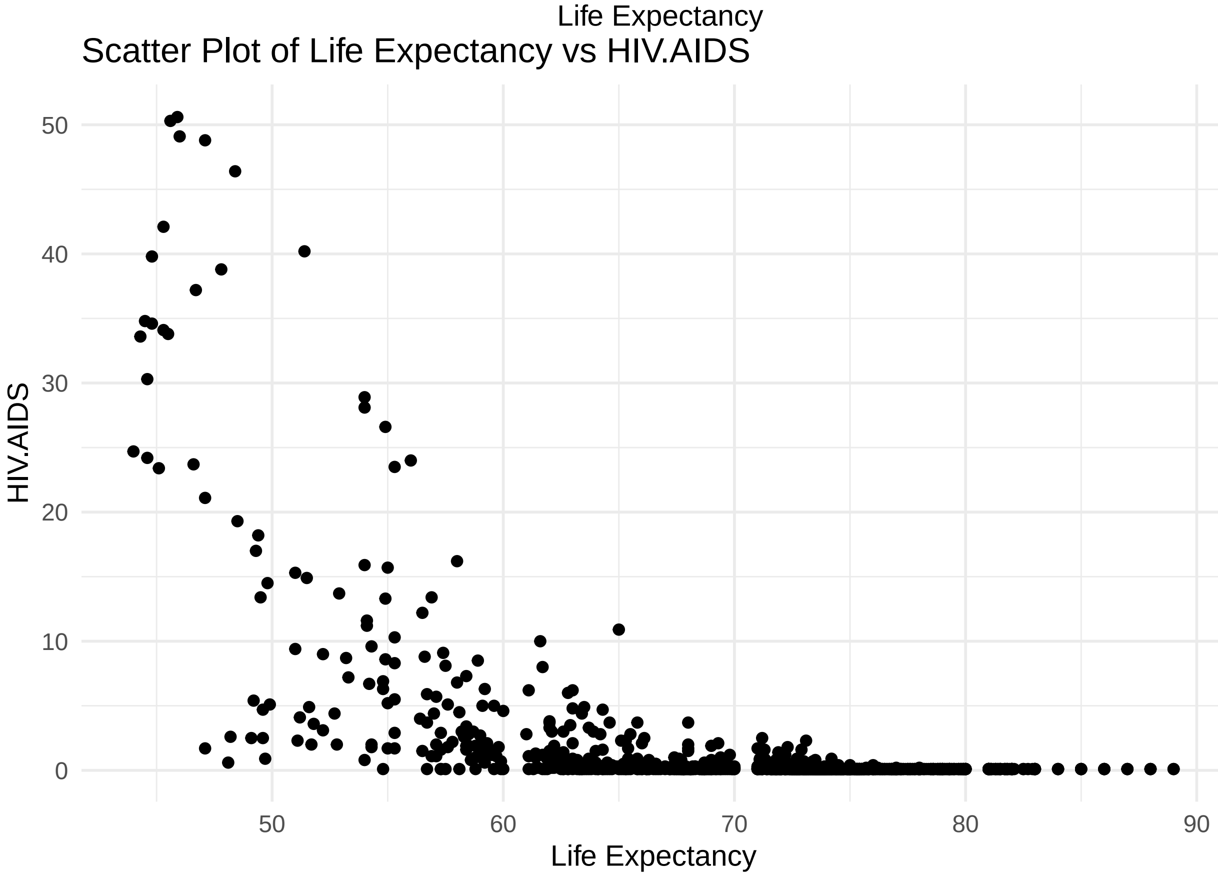
Scatter Plot of Life Expectancy vs Polio

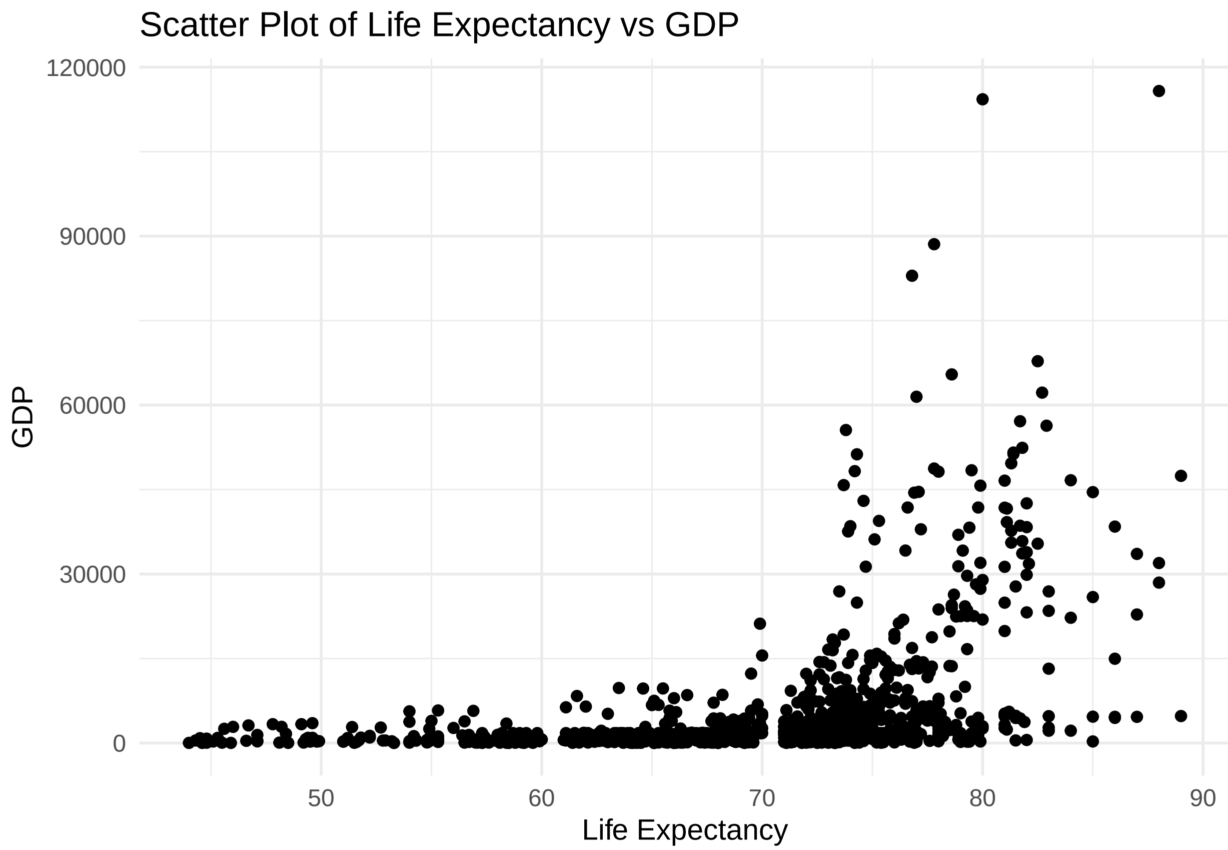


Scatter Plot of Life Expectancy vs Total.expenditure



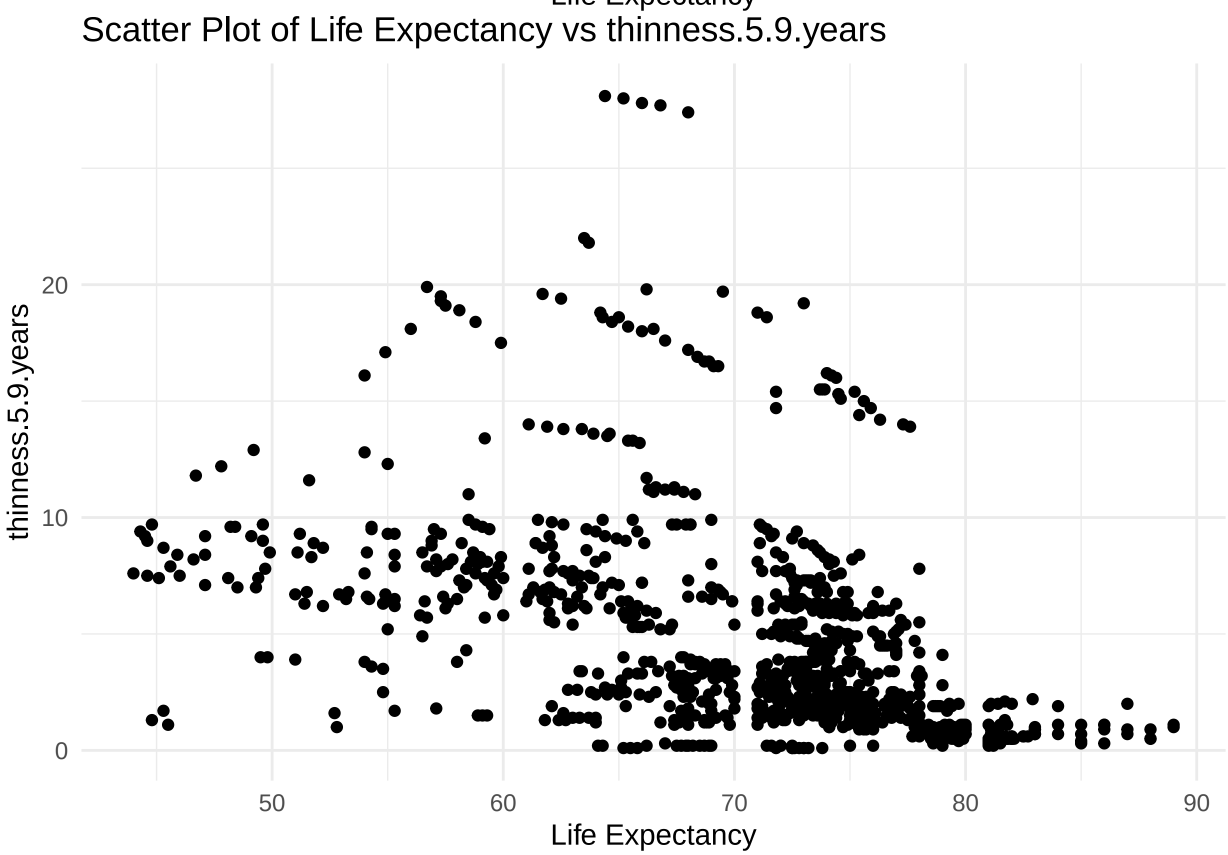

Scatter Plot of Life Expectancy vs Diphtheria
Diphtheria
Life Expectancy
0
25
50
75
100
50
60
70
80
90

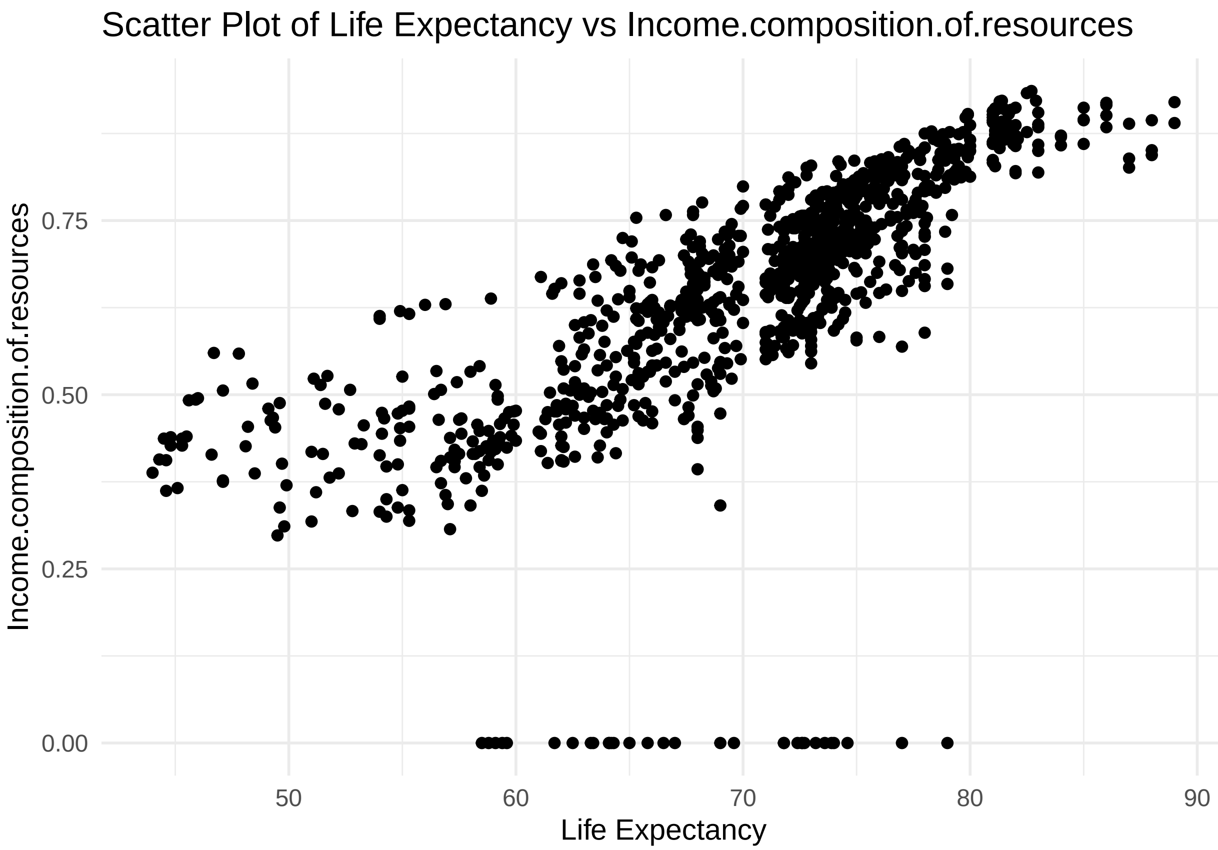

Scatter Plot of Life Expectancy vs HIV.AIDS
HIV.AIDS
Life Expectancy
0
10
20
30
40
50
50
60
70
80
90

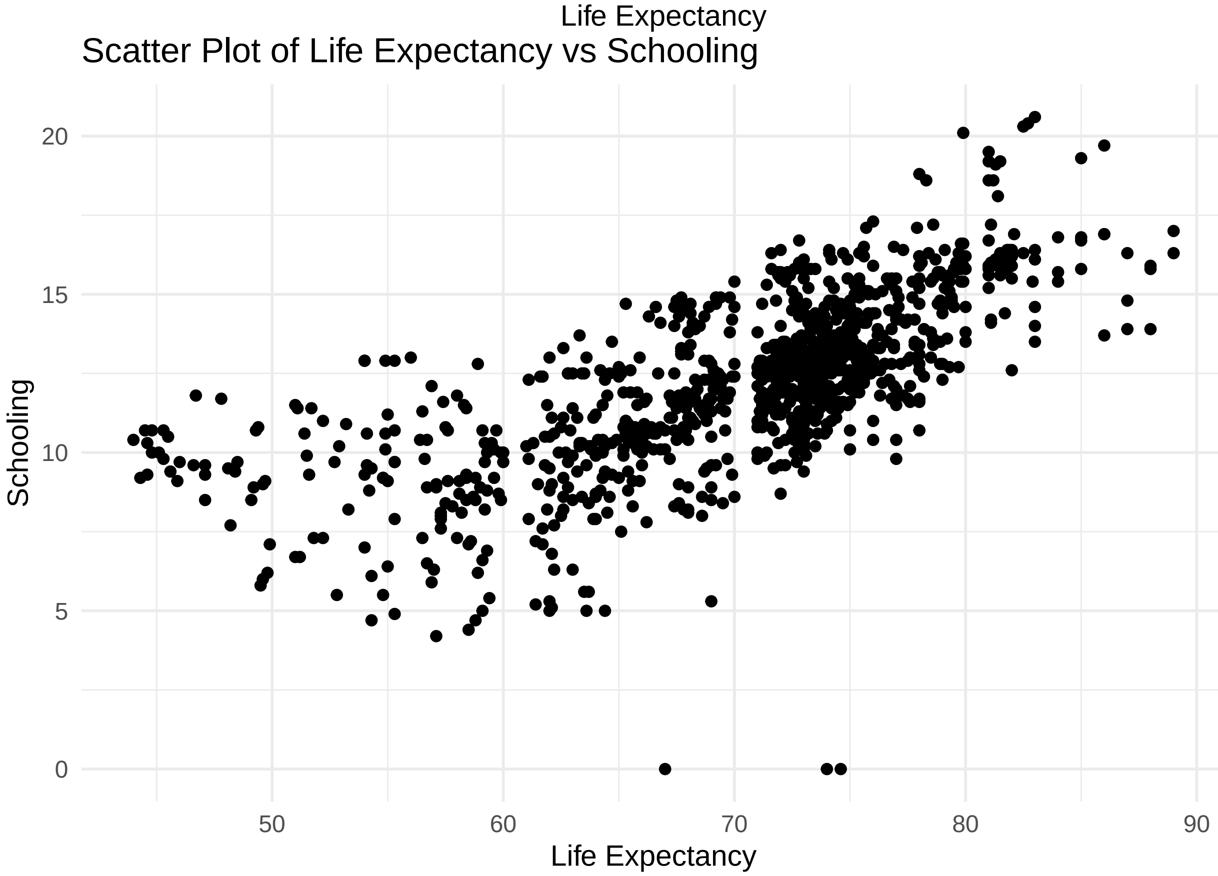

Scatter Plot of Life Expectancy vs GDP
GDP
120000
90000
60000
30000
0
50
60
70
80
90
Life Expectancy

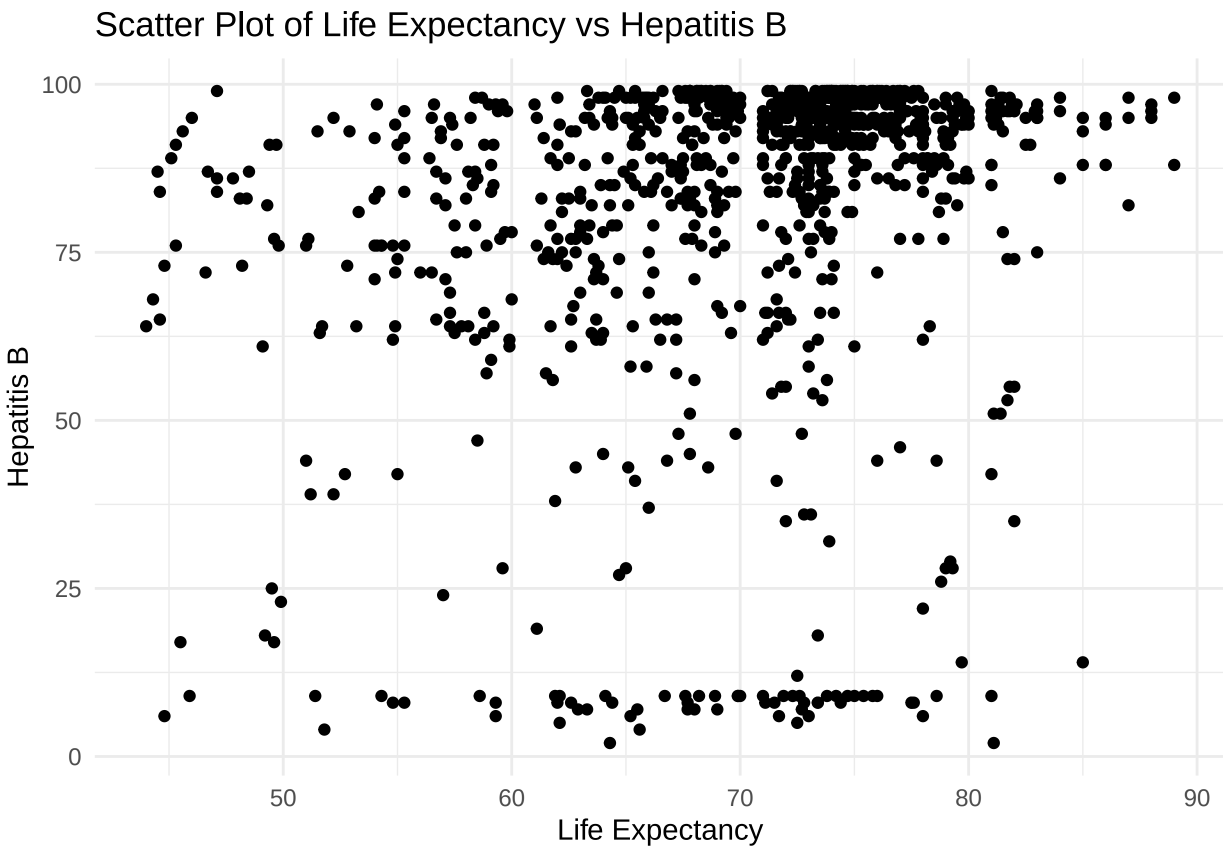

Scatter Plot of Life Expectancy vs Population
Population
1e+09
5e+08
0e+00
50
60
70
80
90
Life Expectancy


Scatter Plot of Life Expectancy vs thinness..1.19.years
thinness..1.19.years
0
10
20
50
60
70
80
90
Life Expectancy

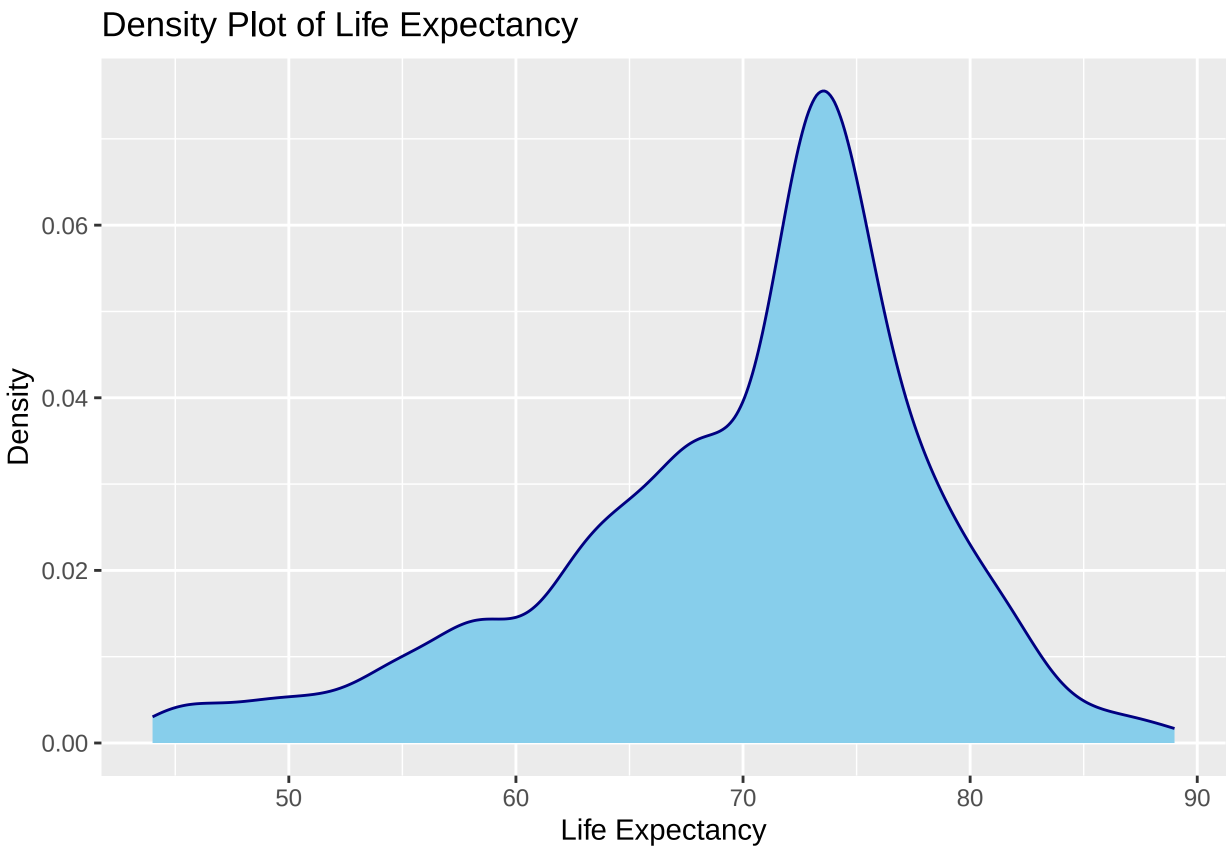

Scatter Plot of Life Expectancy vs thinness.5.9.years
thinness.5.9.years
0
10
20
50
60
70
80
90
Life Expectancy


Scatter Plot of Life Expectancy vs Income.composition.of.resources
Income.composition.of.resources
0.00
0.25
0.50
0.75
50
60
70
80
90
Life Expectancy

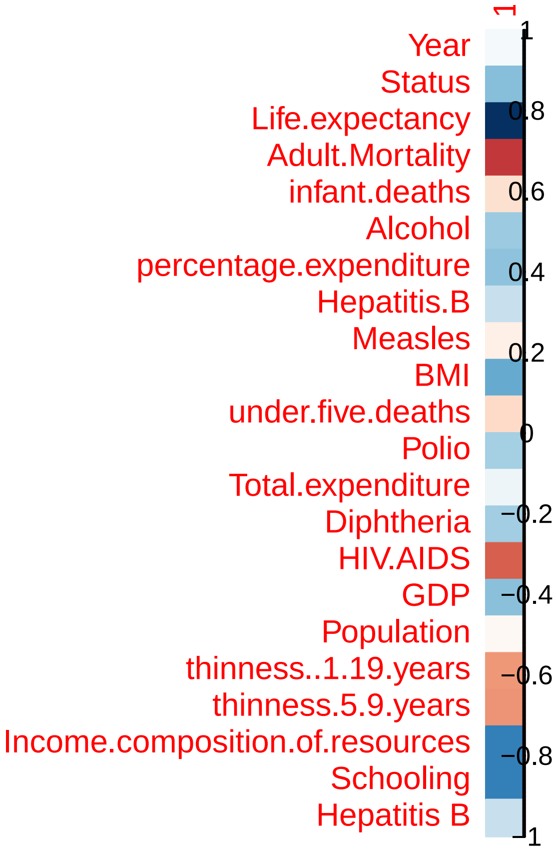

Scatter Plot of Life Expectancy vs Schooling
Schooling
0
5
10
15
20
50
60
70
80
90
Life Expectancy




```r
#density plot of life expectancy
ggplot(Data, aes(x = Life.expectancy)) +
  geom_density(fill = "skyblue", color = "navy") +
  labs(title = "Density Plot of Life Expectancy",
       x = "Life Expectancy",
       y = "Density")
```




```
#from prev plot we can see some predictors has a significant relatonship with life.expectancy, pick them
#Corelation
correlation_with_life_expec <- cor(Data[, sapply(Data, is.numeric)],
                              Data$`Life.expectancy`,
                              use = "complete.obs")
corrplot(correlation_with_life_expec, method = "color")
```




```r
print(correlation_with_life_expec)
```

```
##                                        [,1]
## Year                             0.04650044
## Status                           0.42964462
## Life.expectancy                  1.00000000
## Adult.Mortality                 -0.70801802
## infant.deaths                   -0.16822601
## Alcohol                          0.36031347
## percentage.expenditure           0.40211144
## Hepatitis.B                      0.22934594
## Measles                         -0.08725120
## BMI                              0.50551468
## under.five.deaths               -0.19146114
## Polio                            0.33442369
## Total.expenditure                0.07107883
## Diphtheria                       0.34921733
## HIV.AIDS                        -0.59698455
## GDP                              0.41993874
## Population                      -0.03040113
## thinness..1.19.years            -0.43938934
## thinness.5.9.years              -0.44125295
## Income.composition.of.resources  0.68606730
## Schooling                        0.68846525
## Hepatitis B                      0.22934594
```

```r
#select predictors with correlation abs value greater than 0.4
selected_columns <- names(correlation_with_life_expec[abs(correlation_with_life_expec[, 1]) > 0.4, 1])
print(selected_columns)
```

```
##  [1] "Status"                          "Life.expectancy"
##  [3] "Adult.Mortality"                 "percentage.expenditure"
##  [5] "BMI"                             "HIV.AIDS"
##  [7] "GDP"                             "thinness..1.19.years"
##  [9] "thinness.5.9.years"              "Income.composition.of.resources"
## [11] "Schooling"
```

```r
#split train/test set, only remove unused columns
Data <- Data[, c(selected_columns)]
head(Data)
```

```
##   Status Life.expectancy Adult.Mortality percentage.expenditure  BMI HIV.AIDS
## 1      1            80.0              56            2159.756205 58.7      0.1
## 2      0            73.0             222               8.494095 59.3      0.1
## 3      0            75.3              14               0.000000 59.1      0.1
## 4      0            45.1             615              58.135833 15.5     23.4
## 5      0            67.0             185               0.000000 61.5      0.1
## 6      0            62.8             274              56.431387 38.5      0.1
##           GDP thinness..1.19.years thinness.5.9.years
## 1 28951.15556                  0.9                1.0
## 2    63.38877                  2.0                2.2
## 3  1766.94760                  2.2                2.2
## 4   274.22563                  7.5                7.4
## 5  1766.94760                  0.3                0.3
## 6   474.21334                  2.6                2.6
##   Income.composition.of.resources Schooling
## 1                           0.850      13.8
## 2                           0.780      15.5
## 3                           0.741      12.9
## 4                           0.366      10.0
## 5                           0.000       0.0
## 6                           0.582       8.9
```

```r
train_indices <- sample(1:nrow(Data), 0.8 * nrow(Data))
train_data <- Data[train_indices, ]
test_data <- Data[-train_indices, ]
```

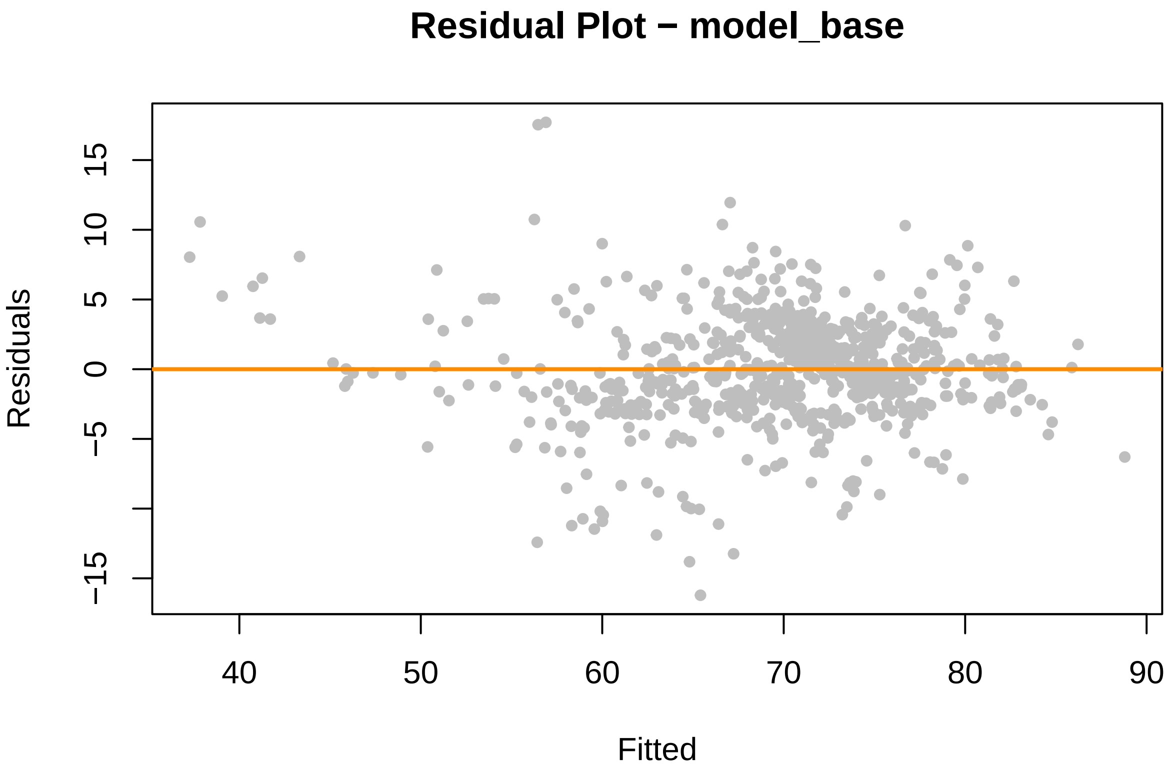
```r
#write two helper function that helps evaluating our model:
#use prediction vs observation, MSE and R2 as main evaluator.
evaluate_model <- function(model) {
  #evaluation based on train set
  print(summary(model))
  print(anova(model))

  #residual plot
  plot(fitted(model), resid(model), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals",
     main = paste("Residual Plot -", deparse(substitute(model))))
  abline(h = 0, col = "darkorange", lwd = 2)

  #BP test
```

```r
  print(bptest(model))

  #Normal Q-Q Plot
  qqnorm(resid(model), main = paste("Normal QQ Plot -", deparse(substitute(model))), col = "darkgrey")
  qqline(resid(model), col = "dodgerblue", lwd = 2)

  #SW normality test
  print(shapiro.test(resid(model)))

  #MSE and R2
  mse <- mean(resid(model)^2)
  cat("Train Set MSE:", mse, "\n")
  r_squared <- summary(model)$r.squared
  cat("Train R-squared:", r_squared, "\n")
}

#evaluation for test set
evaluate_test <- function(model, test_data) {
  #predictions
  predictions <- predict(model, newdata = test_data)
  y = test_data$Life.expectancy
  y_hat = predictions
  SST = sum((y - mean(y)) ^ 2)
  SSR = sum((y_hat - mean(y)) ^ 2)
  SSE = sum((y - y_hat) ^ 2)
  R2 <- 1 - (SSE / SST)

  #calculate residual
  residuals <- test_data$Life.expectancy - predictions

  #MSE calculation
  mse <- mean(residuals^2, na.rm = TRUE)

  plot(test_data$Life.expectancy, y_hat, col = "blue", pch = 20,
       xlab = "Observed Life Expectancy", ylab = "Predicted Life Expectancy",
       main = "Test Set - Observed vs. Predicted Life Expectancy")
  abline(0, 1, col = "red", lwd = 2)

  #Print results
  cat("Test Set MSE:", mse, "\n")
  cat("Test R-squared:", R2, "\n")
}
```

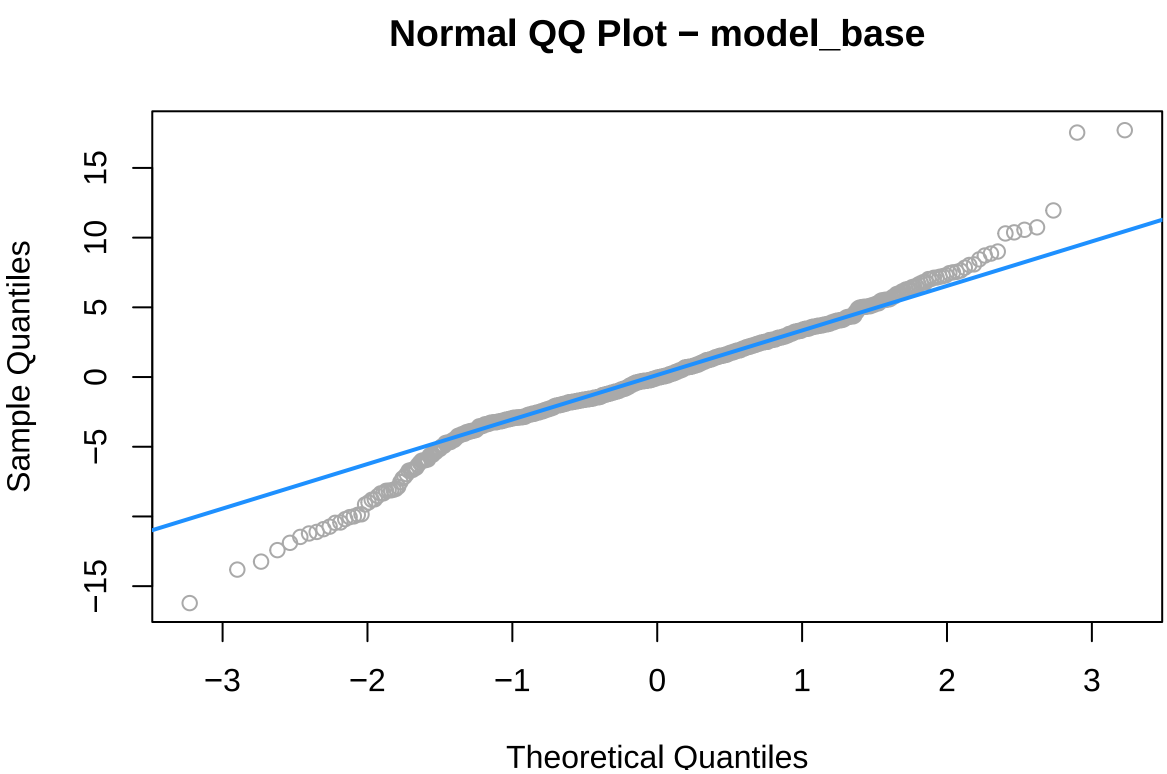
```r
#fit my first model - a multiple linear regression model with all predictors
model_base <- lm(Life.expectancy ~ ., data = train_data)
evaluate_model(model_base)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2138  -2.0071  -0.0484   2.3017  17.7066
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.809e+01  9.378e-01  61.943  < 2e-16 ***
## Status                        8.235e-01  4.952e-01   1.663 0.096700 .
## Adult.Mortality              -1.937e-02  1.451e-03 -13.344  < 2e-16 ***
## percentage.expenditure        3.792e-04  2.050e-04   1.850 0.064704 .
## BMI                           2.905e-02  8.751e-03   3.320 0.000942 ***
## HIV.AIDS                     -4.546e-01  2.599e-02 -17.493  < 2e-16 ***
## GDP                           2.458e-05  2.749e-05   0.894 0.371517
## thinness..1.19.years         -2.276e-02  1.018e-01  -0.224 0.823060
## thinness.5.9.years           -7.571e-02  9.992e-02  -0.758 0.448843
## Income.composition.of.resources  7.641e+00  1.190e+00   6.422 2.32e-10 ***
## Schooling                     7.893e-01  8.108e-02   9.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.792 on 789 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7914
## F-statistic: 304.1 on 10 and 789 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                 Df  Sum Sq Mean Sq   F value    Pr(>F)
## Status                           1  9792.6  9792.6  681.0483 < 2.2e-16 ***
## Adult.Mortality                  1 19647.9 19647.9 1366.4573 < 2.2e-16 ***
## percentage.expenditure           1  1543.0  1543.0  107.3094 < 2.2e-16 ***
## BMI                              1  2828.7  2828.7  196.7254 < 2.2e-16 ***
## HIV.AIDS                         1  4504.1  4504.1  313.2473 < 2.2e-16 ***
## GDP                              1    81.8    81.8    5.6856   0.01734 *
## thinness..1.19.years             1   515.3   515.3   35.8379 3.254e-09 ***
## thinness.5.9.years               1     3.4     3.4    0.2338   0.62882
## Income.composition.of.resources  1  3440.4  3440.4  239.2687 < 2.2e-16 ***
## Schooling                        1  1362.9  1362.9   94.7845 < 2.2e-16 ***
## Residuals                      789 11344.8    14.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

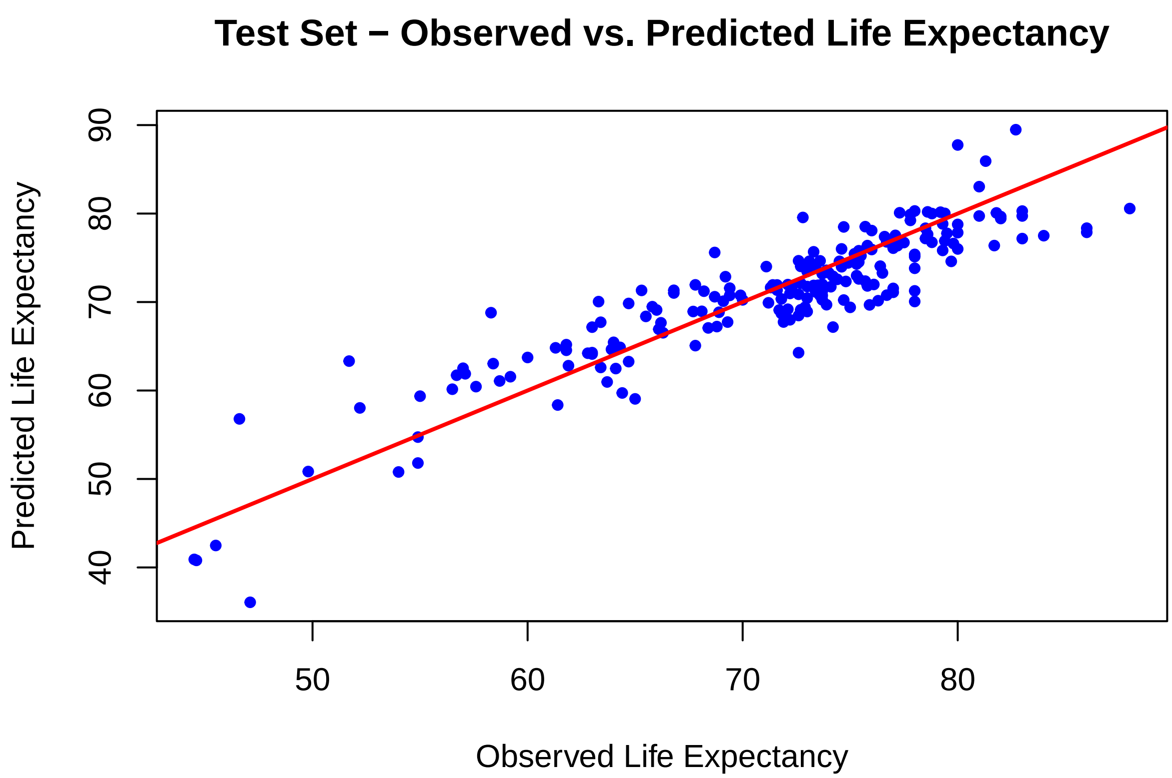



```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 121.11, df = 10, p-value < 2.2e-16
```



```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.97074, p-value = 1.418e-11
##
## Train Set MSE: 14.18098
## Train R-squared: 0.7939732
```

```r
#calculate SST etc and R2 for test set
evaluate_test(model_base, test_data)
```




```
## Test Set MSE: 13.20732
## Test R-squared: 0.8091199
```

```r
#the anova test shows there exist issue with predictors, so we do a VIF test
vif_value <- car::vif(model_base)
print(vif_value)
```

```
##                            Status                  Adult.Mortality
##                          1.492273                         1.627661
##            percentage.expenditure                              BMI
##                          5.812907                         1.700131
##                          HIV.AIDS                              GDP
##                          1.348801                         5.727922
##              thinness..1.19.years               thinness.5.9.years
##                         11.565617                        11.517557
## Income.composition.of.resources                        Schooling
##                          2.636653                         2.785809
```

```r
#hypothesis test for linearity between thiness 1.19 and thinness 5.9
#H0: non exist, H1: exist - p = 2.2e-16 < 0.05
```

```r
colinear_model <- lm(thinness..1.19.years ~ thinness.5.9.years, data = train_data)
summary(colinear_model)
```

```
##
## Call:
## lm(formula = thinness..1.19.years ~ thinness.5.9.years, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6900  -0.1722  -0.0291   0.2536  17.2649
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.26171    0.06997    3.74 0.000197 ***
## thinness.5.9.years  0.93971    0.01033   90.92  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.331 on 798 degrees of freedom
## Multiple R-squared:  0.912,  Adjusted R-squared:  0.9119
## F-statistic:  8267 on 1 and 798 DF,  p-value: < 2.2e-16
```

```r
#based on interpretation above, build next model, removing thinness 5.9
model_VIF_reduced <- lm(Life.expectancy ~ Adult.Mortality + BMI + HIV.AIDS + GDP + thinness..1.19.years
               Income.composition.of.resources + Schooling, data = train_data)

vif_value2 <- car::vif(model_VIF_reduced)
print(vif_value2)
```

```
##                 Adult.Mortality                             BMI
##                        1.617241                        1.676392
##                        HIV.AIDS                             GDP
##                        1.346659                        1.256854
##            thinness..1.19.years Income.composition.of.resources
##                        1.511341                        2.622933
##                       Schooling
##                        2.662509
```

```r
#this passes VIF test, so we
train_data <- train_data[, !colnames(train_data) %in% c('thinness.5.9.years')]
test_data <- test_data[, !colnames(test_data) %in% c('thinness.5.9.years')]
head(train_data)
```

```
##     Status Life.expectancy Adult.Mortality percentage.expenditure  BMI HIV.AIDS
## 962      0            58.9             413              123.75334 47.9      8.5
## 918      0            65.2              28              162.29037 74.6      0.1
## 145      0            74.1             126               43.08717 51.8      0.1
## 645      0            68.3             183              119.45712  4.7      0.2
## 627      1            81.7              57            10947.02327 58.1      0.1
## 335      0            74.0              86             2009.57560 68.4      0.1
##           GDP thinness..1.19.years Income.composition.of.resources Schooling
## 962  849.9542                  6.6                           0.638      12.8
## 918 1297.2851                  0.1                           0.576      11.9
## 145  495.2549                  6.0                           0.697      12.6
## 645 1377.8214                  3.0                           0.657      11.9
```

```
## 627 57134.7770                  1.4                             0.903      15.8
## 335 38497.6170                  3.3                             0.790      13.5
```

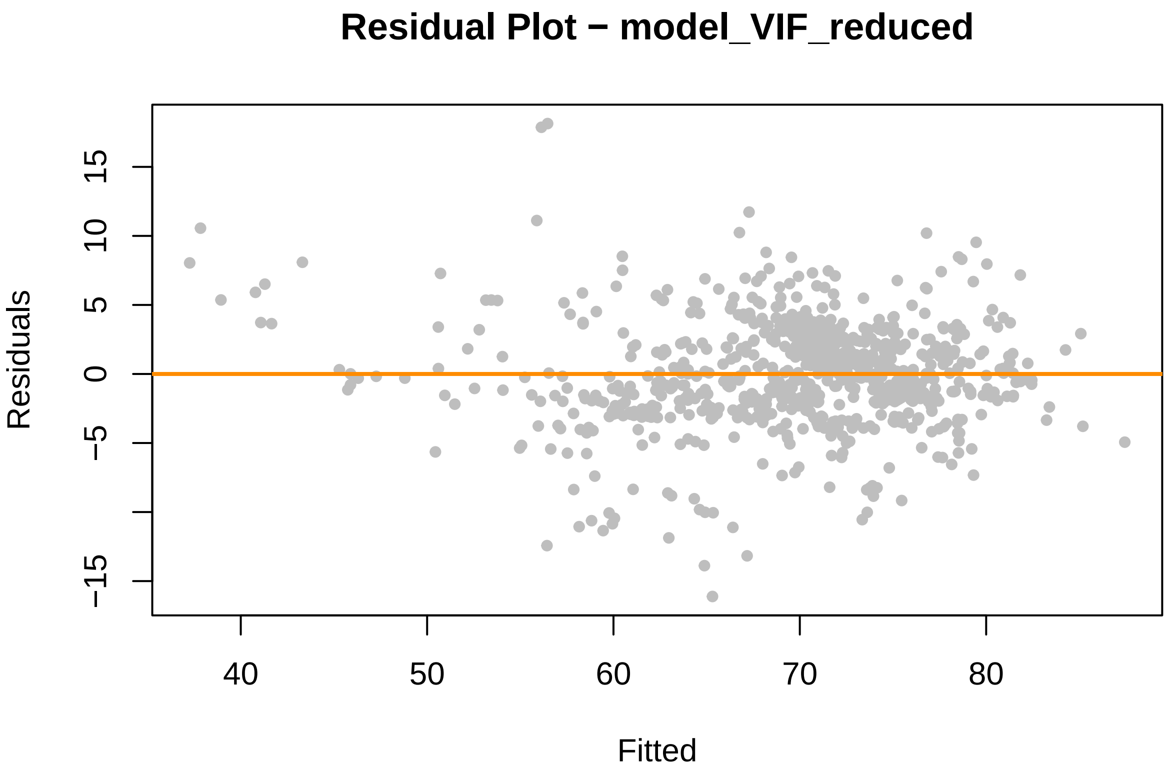
```r
head(test_data)
```

```
##    Status Life.expectancy Adult.Mortality percentage.expenditure  BMI HIV.AIDS
## 2       0            73.0             222               8.494095 59.3      0.1
## 3       0            75.3              14               0.000000 59.1      0.1
## 8       0            82.0              94            5291.234786 57.0      0.1
## 15      0            75.0             153             345.339056 47.4      0.1
## 21      0            65.8              21              11.136087 26.2      0.9
## 24      0            72.6             122               4.409153 28.2      0.3
##            GDP thinness..1.19.years Income.composition.of.resources Schooling
## 2     63.38877                  2.0                           0.780      15.5
## 3   1766.94760                  2.2                           0.741      12.9
## 8  33874.74255                  0.6                           0.857      15.5
## 15  2697.96137                  1.5                           0.582      10.7
## 21   199.57146                  6.3                           0.533      10.1
## 24   367.42945                  7.3                           0.632      13.1
```

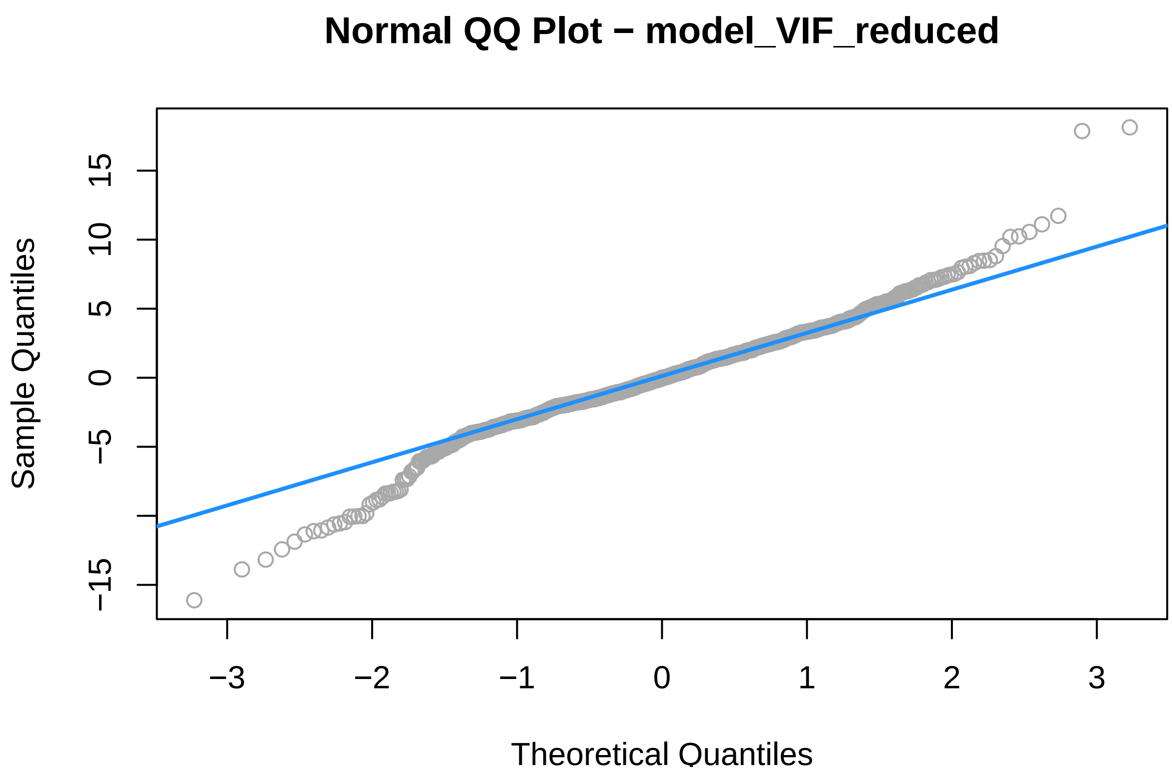
```r
#evaluate this model:
evaluate_model(model_VIF_reduced)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + BMI + HIV.AIDS +
##     GDP + thinness..1.19.years + Income.composition.of.resources +
##     Schooling, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1132  -1.9794  -0.0096   2.2351  18.1348
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.773e+01  9.294e-01  62.113  < 2e-16 ***
## Adult.Mortality                 -1.951e-02  1.452e-03 -13.443  < 2e-16 ***
## BMI                              2.778e-02  8.719e-03   3.186  0.00150 **
## HIV.AIDS                        -4.533e-01  2.606e-02 -17.396  < 2e-16 ***
## GDP                              7.654e-05  1.292e-05   5.924 4.70e-09 ***
## thinness..1.19.years            -1.059e-01  3.691e-02  -2.870  0.00421 **
## Income.composition.of.resources  7.627e+00  1.191e+00   6.405 2.57e-10 ***
## Schooling                        8.317e-01  7.953e-02  10.458  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.805 on 792 degrees of freedom
## Multiple R-squared:  0.7918, Adjusted R-squared:  0.7899
## F-statistic: 430.2 on 7 and 792 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Adult.Mortality                   1 26244.3 26244.3 1812.834 < 2.2e-16 ***
## BMI                               1  4478.3  4478.3  309.343 < 2.2e-16 ***
```

```
## HIV.AIDS                          1  4056.4  4056.4  280.197 < 2.2e-16 ***
## GDP                               1  2422.9  2422.9  167.363 < 2.2e-16 ***
## thinness..1.19.years              1   812.4   812.4   56.119 1.824e-13 ***
## Income.composition.of.resources   1  4001.3  4001.3  276.393 < 2.2e-16 ***
## Schooling                         1  1583.3  1583.3  109.365 < 2.2e-16 ***
## Residuals                       792 11465.7    14.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

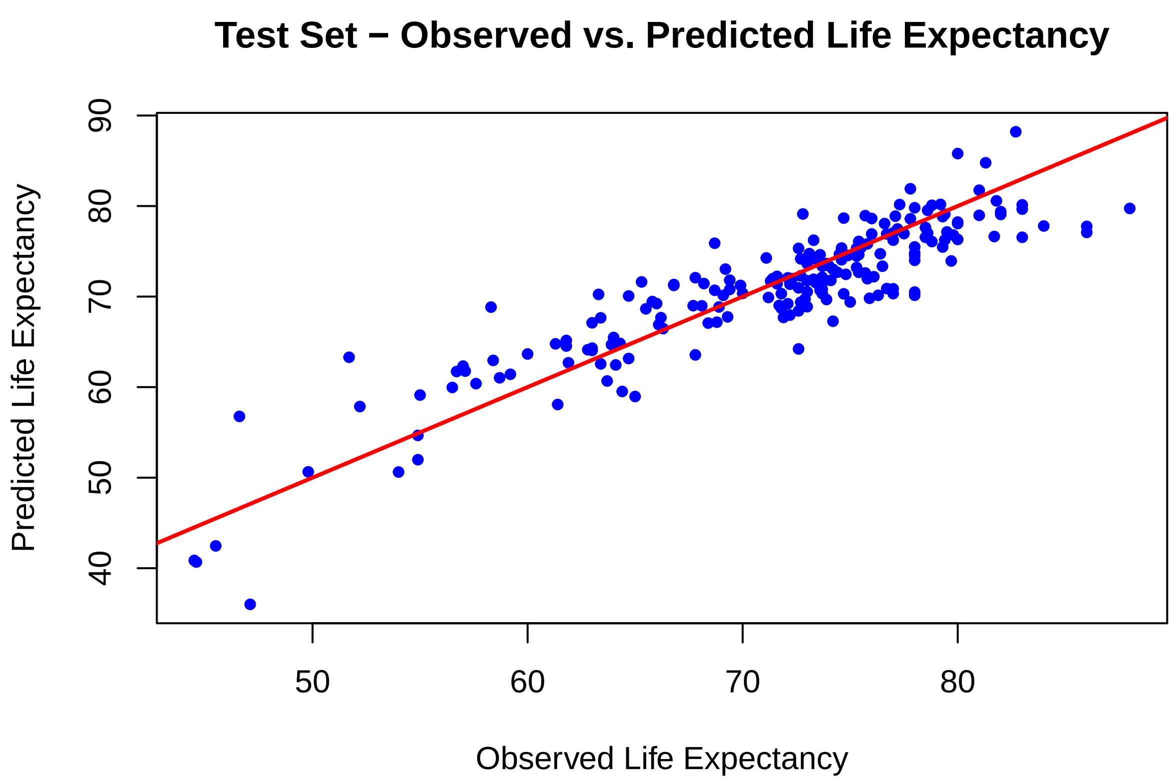



```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 111.86, df = 7, p-value < 2.2e-16
```



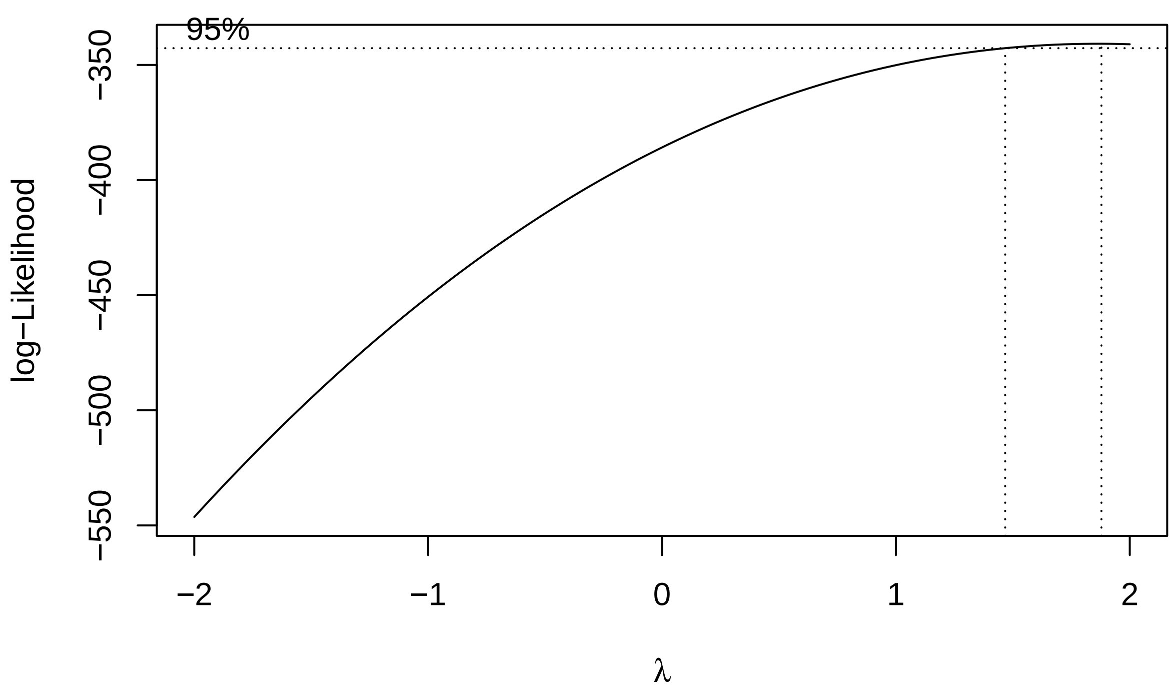

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.9716, p-value = 2.334e-11
##
## Train Set MSE: 14.33215
## Train R-squared: 0.791777
```

```
evaluate_test(model_VIF_reduced, test_data)
```

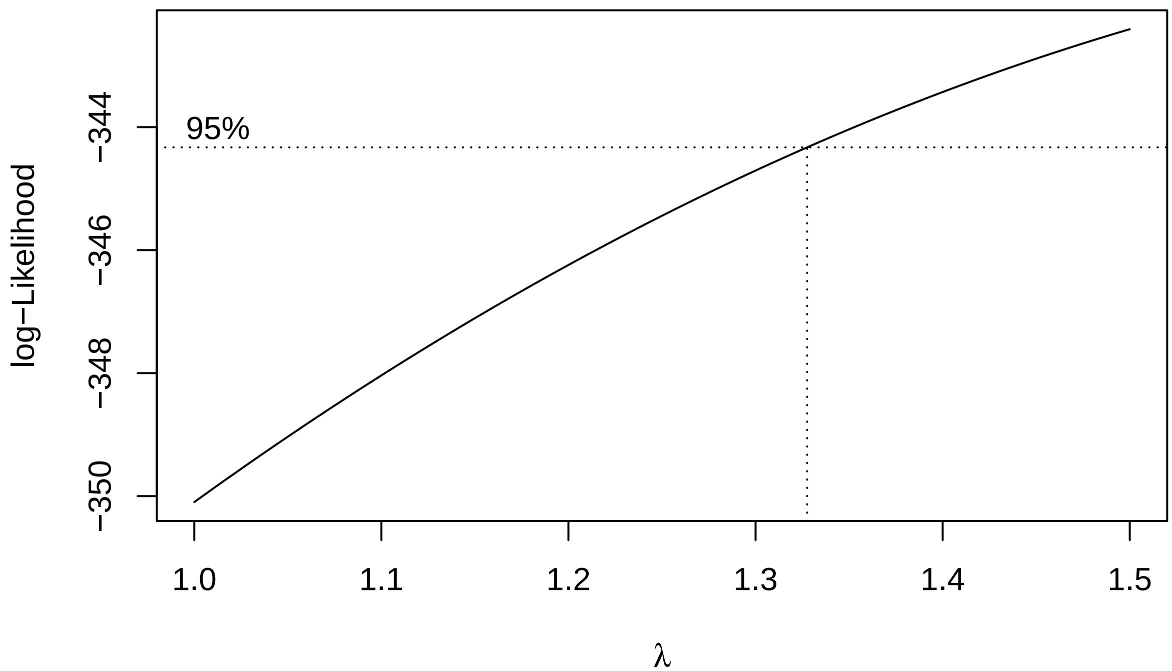



```
## Test Set MSE: 13.54154
## Test R-squared: 0.8042896
```

```
#do a box plot to see if we should shift y values:
boxcox_result <- boxcox(model_VIF_reduced, plotit = TRUE)
```



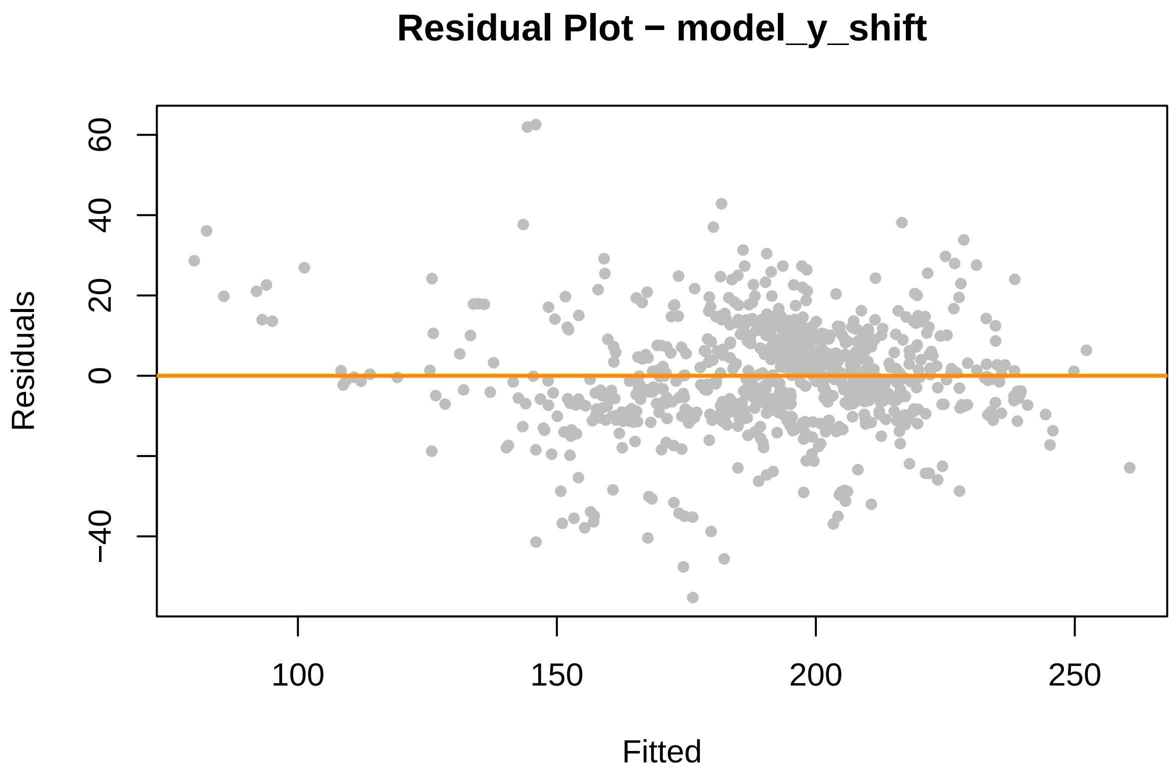
```r
boxcox(model_VIF_reduced, lambda <- seq(1, 1.5, by = 0.05), plotit = TRUE)
```



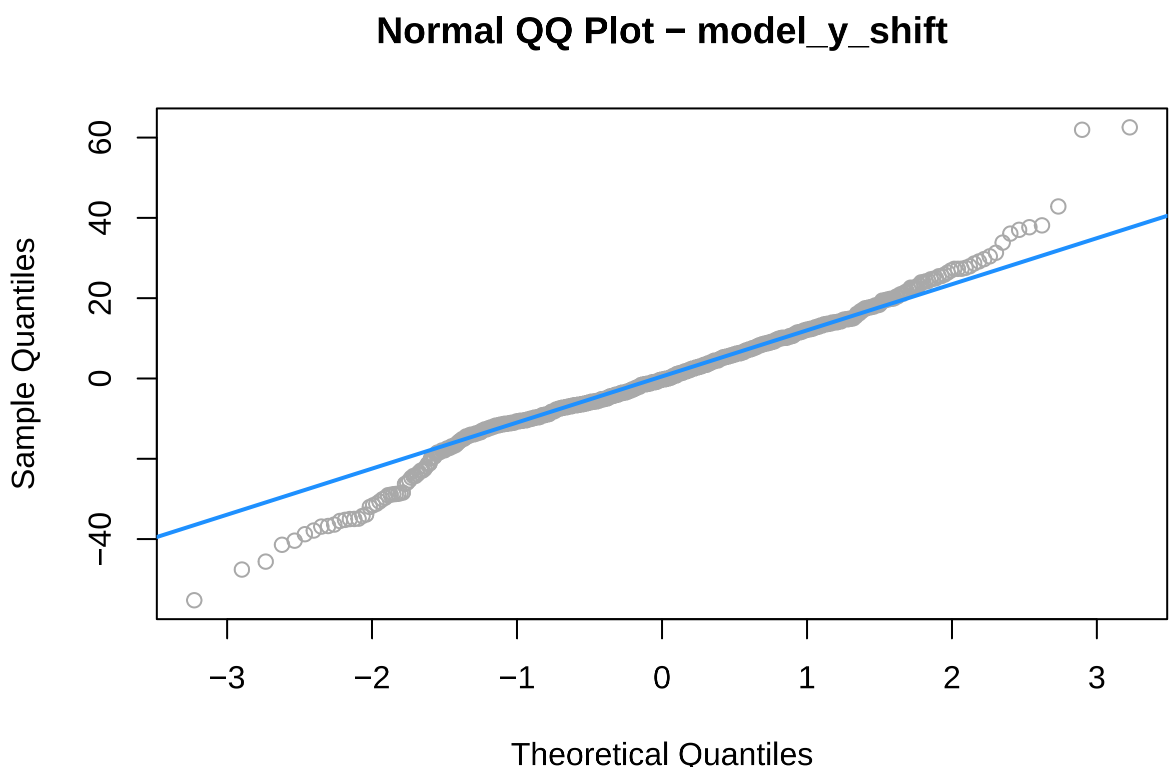

```r
#from the plot, use l = 0.5
model_y_shift <- lm((((Life.expectancy ^ 1.3) - 1) /1.3) ~ ., data = train_data)
evaluate_model(model_y_shift)
```

```
##
## Call:
## lm(formula = (((Life.expectancy^1.3) - 1)/1.3) ~ ., data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.232  -7.225  -0.285   8.256  62.546
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.498e+02  3.309e+00  45.277  < 2e-16 ***
## Status                          3.524e+00  1.749e+00   2.014 0.044327 *
## Adult.Mortality                -6.795e-02  5.128e-03 -13.251  < 2e-16 ***
## percentage.expenditure          1.457e-03  7.243e-04   2.011 0.044619 *
## BMI                             1.033e-01  3.080e-02   3.354 0.000834 ***
## HIV.AIDS                       -1.527e+00  9.181e-02 -16.629  < 2e-16 ***
## GDP                             8.983e-05  9.712e-05   0.925 0.355308
## thinness..1.19.years           -3.401e-01  1.309e-01  -2.599 0.009514 **
## Income.composition.of.resources  2.688e+01  4.204e+00   6.393 2.78e-10 ***
## Schooling                       2.795e+00  2.865e-01   9.757  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 790 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7891
## F-statistic: 333.2 on 9 and 790 DF,  p-value: < 2.2e-16
##
```

```
## Analysis of Variance Table
##
## Response: (((Life.expectancy^1.3) - 1)/1.3)
##                                Df Sum Sq Mean Sq   F value    Pr(>F)    
## Status                          1 126740  126740  705.9649 < 2.2e-16 ***
## Adult.Mortality                 1 237693  237693 1323.9894 < 2.2e-16 ***
## percentage.expenditure          1  20822   20822  115.9819 < 2.2e-16 ***
## BMI                             1  34909   34909  194.4498 < 2.2e-16 ***
## HIV.AIDS                        1  50714   50714  282.4858 < 2.2e-16 ***
## GDP                             1   1033    1033    5.7543   0.01668 *  
## thinness..1.19.years            1   6538    6538   36.4185 2.446e-09 ***
## Income.composition.of.resources   1  42852   42852  238.6912 < 2.2e-16 ***
## Schooling                       1  17091   17091   95.1981 < 2.2e-16 ***
## Residuals                     790 141827     180                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

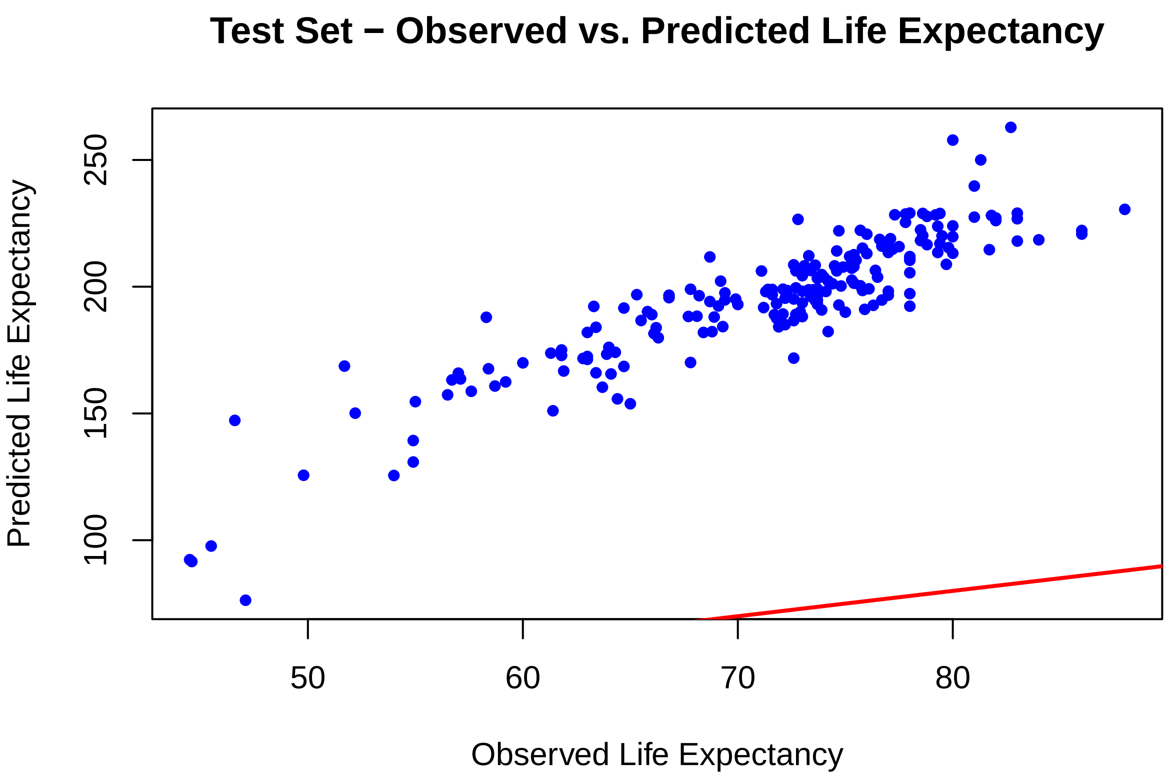



```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 117.28, df = 9, p-value < 2.2e-16
```




```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.97474, p-value = 1.549e-10
##
## Train Set MSE: 177.2838
## Train R-squared: 0.7914981
```

```
evaluate_test(model_y_shift, test_data)
```



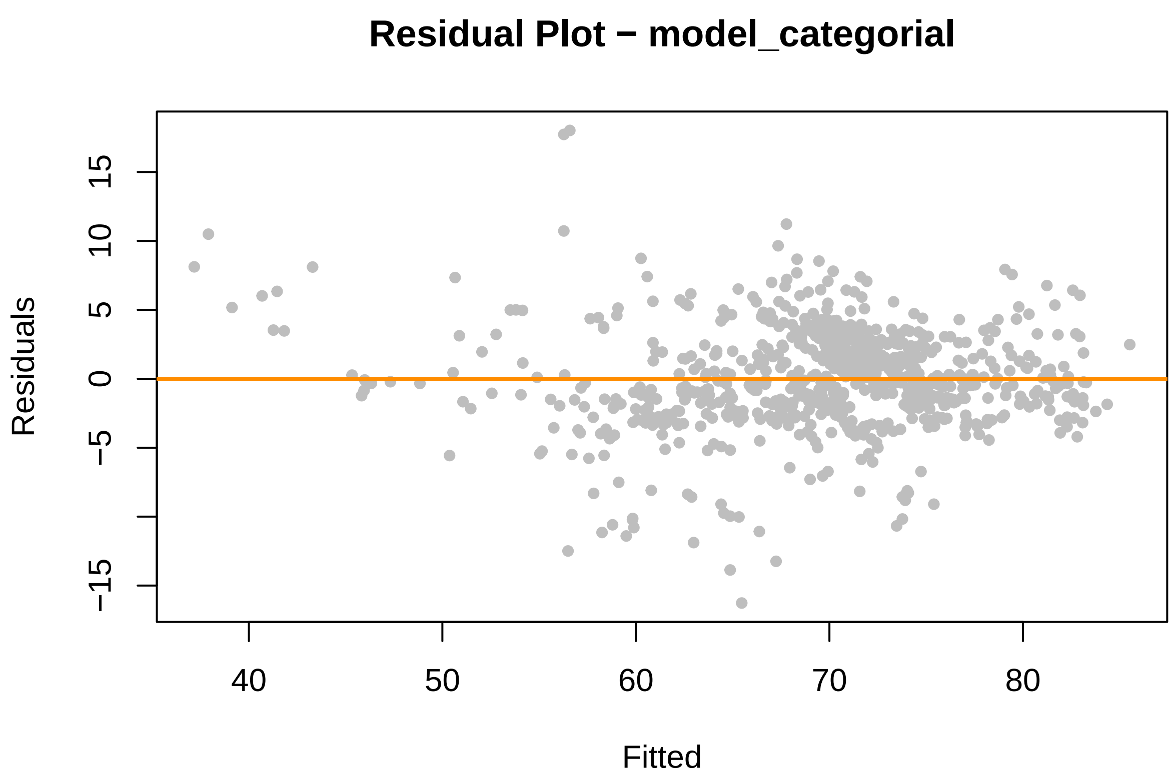

```
## Test Set MSE: 15761.39
## Test R-squared: -226.7931
```

```
#from the catastratpic result we can see that including 1.3 shift is not a good idea.
#here comes our next model with multiplying categorial variable
model_categorial <- lm(Life.expectancy ~ Status + Adult.Mortality + BMI + HIV.AIDS + GDP + thinness..1.
                       + Income.composition.of.resources + Schooling + I(Adult.Mortality*Status) + I(BM
                       + I(HIV.AIDS*Status) + I(GDP*Status) + I(thinness..1.19.years*Status)
                       + I(Income.composition.of.resources*Status) + I(Schooling*Status),
                       data = train_data)

evaluate_model(model_categorial)
```

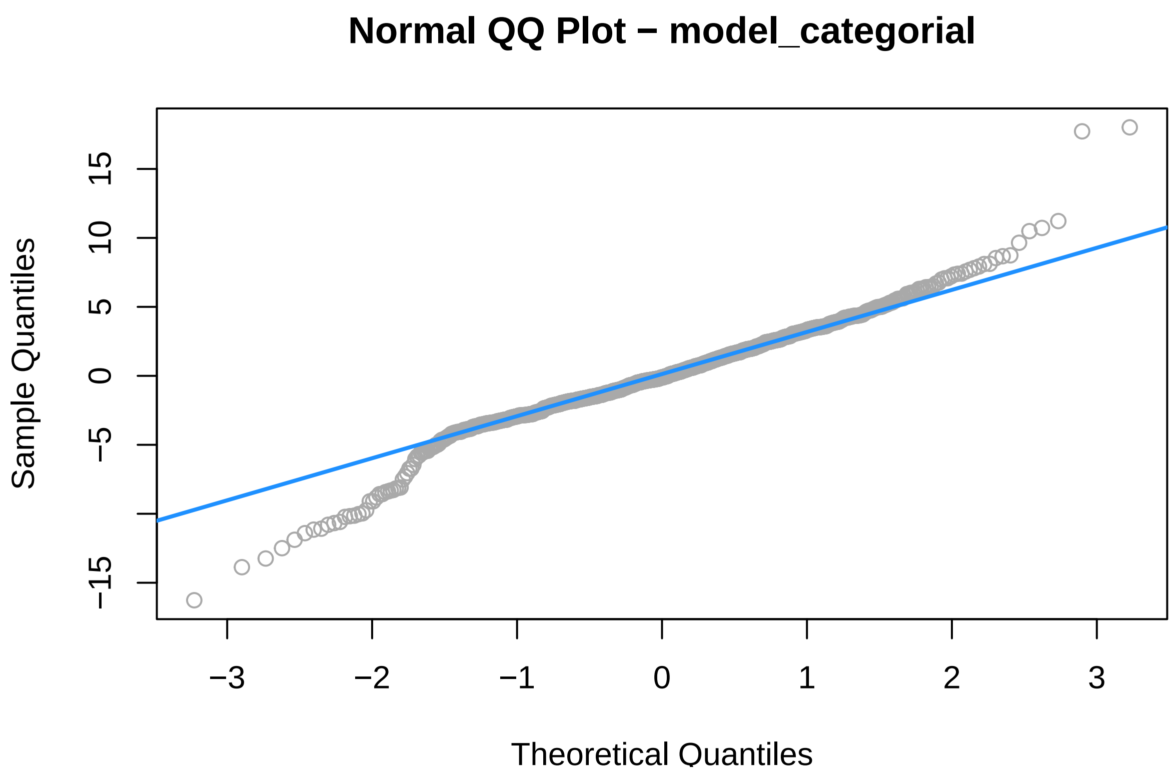
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + BMI +
##     HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources +
##     Schooling + I(Adult.Mortality * Status) + I(BMI * Status) +
##     I(HIV.AIDS * Status) + I(GDP * Status) + I(thinness..1.19.years *
##     Status) + I(Income.composition.of.resources * Status) + I(Schooling *
##     Status), data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2685  -1.9277  -0.1131   2.1881  18.0149
##
## Coefficients: (1 not defined because of singularities)
##                                              Estimate Std. Error t value
## (Intercept)                                 5.755e+01  9.483e-01  60.685
## Status                                     -1.683e+01  9.852e+00  -1.708
```

```
## Adult.Mortality                              -1.952e-02  1.438e-03 -13.578
## BMI                                           3.748e-02  9.304e-03   4.029
## HIV.AIDS                                     -4.523e-01  2.544e-02 -17.777
## GDP                                           5.758e-05  1.843e-05   3.123
## thinness..1.19.years                         -7.252e-02  3.668e-02  -1.977
## Income.composition.of.resources               6.645e+00  1.181e+00   5.627
## Schooling                                     8.508e-01  8.360e-02  10.177
## I(Adult.Mortality * Status)                   1.846e-02  8.592e-03   2.148
## I(BMI * Status)                              -4.749e-02  2.417e-02  -1.965
## I(HIV.AIDS * Status)                                 NA         NA      NA
## I(GDP * Status)                              -3.722e-05  2.892e-05  -1.287
## I(thinness..1.19.years * Status)             -1.758e+00  6.179e-01  -2.845
## I(Income.composition.of.resources * Status)  5.058e+01  1.335e+01   3.788
## I(Schooling * Status)                        -1.308e+00  3.049e-01  -4.289
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## Status                                       0.088005 .
## Adult.Mortality                               < 2e-16 ***
## BMI                                          6.15e-05 ***
## HIV.AIDS                                      < 2e-16 ***
## GDP                                          0.001855 **
## thinness..1.19.years                         0.048399 *
## Income.composition.of.resources              2.55e-08 ***
## Schooling                                     < 2e-16 ***
## I(Adult.Mortality * Status)                  0.031995 *
## I(BMI * Status)                              0.049738 *
## I(HIV.AIDS * Status)                               NA
## I(GDP * Status)                              0.198520
## I(thinness..1.19.years * Status)             0.004562 **
## I(Income.composition.of.resources * Status) 0.000163 ***
## I(Schooling * Status)                        2.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.7 on 785 degrees of freedom
## Multiple R-squared:  0.8049, Adjusted R-squared:  0.8014
## F-statistic: 231.3 on 14 and 785 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                              Df  Sum Sq Mean Sq   F value
## Status                                        1  9792.6  9792.6  715.4108
## Adult.Mortality                               1 19647.9 19647.9 1435.4024
## BMI                                           1  3331.0  3331.0  243.3494
## HIV.AIDS                                      1  4339.4  4339.4  317.0206
## GDP                                           1  1212.3  1212.3   88.5671
## thinness..1.19.years                          1   533.2   533.2   38.9557
## Income.composition.of.resources               1  3411.0  3411.0  249.1967
## Schooling                                     1  1395.2  1395.2  101.9294
## I(Adult.Mortality * Status)                   1    17.1    17.1    1.2465
## I(BMI * Status)                               1    42.2    42.2    3.0796
## I(GDP * Status)                               1    14.6    14.6    1.0641
## I(thinness..1.19.years * Status)              1   302.1   302.1   22.0693
```

```
## I(Income.composition.of.resources * Status)   1    29.2    29.2    2.1344
## I(Schooling * Status)                          1   251.8   251.8   18.3956
## Residuals                                    785 10745.1    13.7
##                                                 Pr(>F)
## Status                                       < 2.2e-16 ***
## Adult.Mortality                              < 2.2e-16 ***
## BMI                                          < 2.2e-16 ***
## HIV.AIDS                                     < 2.2e-16 ***
## GDP                                          < 2.2e-16 ***
## thinness..1.19.years                         7.090e-10 ***
## Income.composition.of.resources              < 2.2e-16 ***
## Schooling                                    < 2.2e-16 ***
## I(Adult.Mortality * Status)                    0.26455
## I(BMI * Status)                                0.07967 .
## I(GDP * Status)                                0.30260
## I(thinness..1.19.years * Status)             3.104e-06 ***
## I(Income.composition.of.resources * Status)    0.14443
## I(Schooling * Status)                        2.018e-05 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



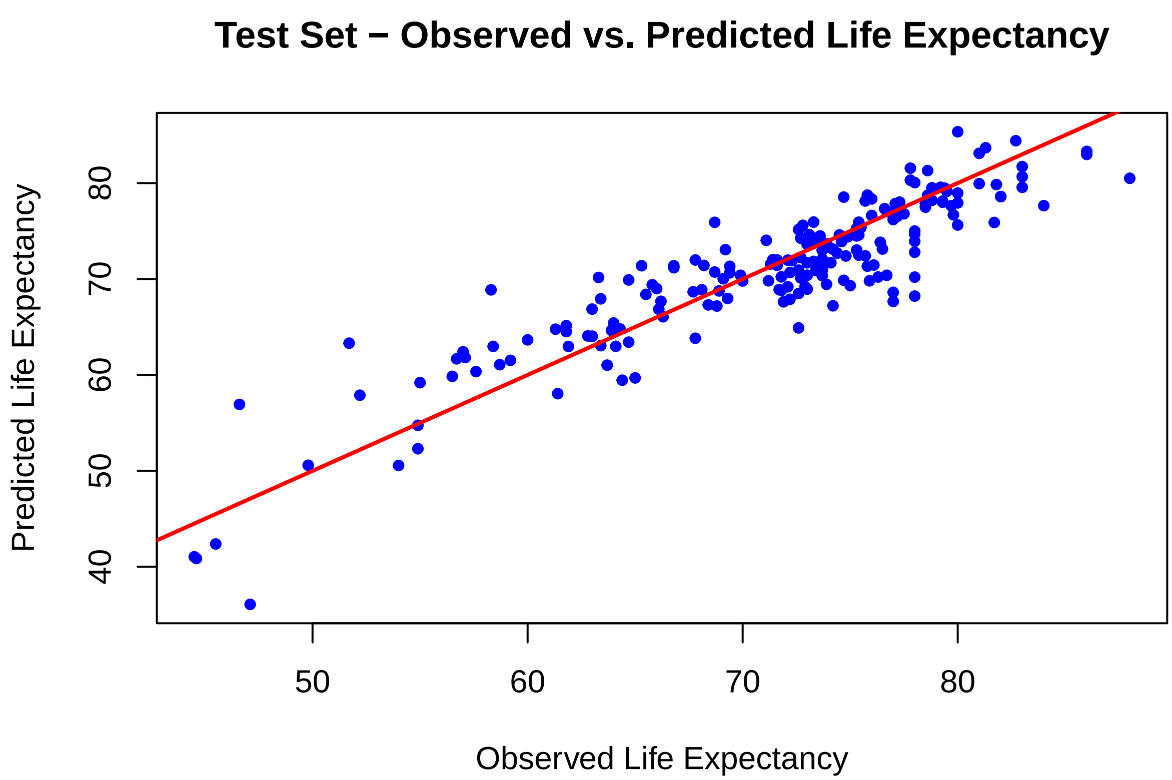

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 127.42, df = 14, p-value < 2.2e-16
```




```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.96504, p-value = 6.644e-13
##
## Train Set MSE: 13.43141
## Train R-squared: 0.8048633
```

```
evaluate_test(model_categorial, test_data)
```

```
## Warning in predict.lm(model, newdata = test_data): prediction from
## rank-deficient fit; attr(*, "non-estim") has doubtful cases
```




```
## Test Set MSE: 12.68117
## Test R-squared: 0.8167241
```

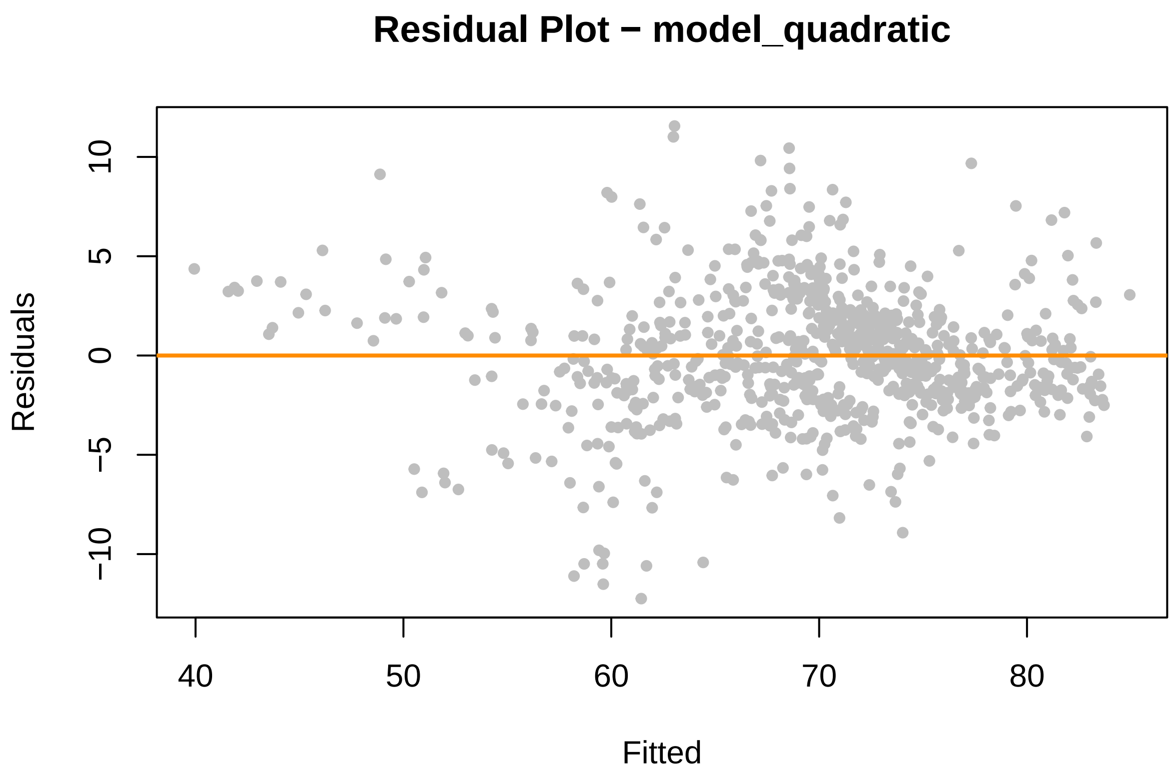
```r
#anova(model_categorial)
```

```r
#model quadratic: removed predictors with high anova pr, and added quadratic form
model_quadratic <- lm(Life.expectancy ~ Status + Adult.Mortality + BMI + HIV.AIDS + GDP
                      + thinness..1.19.years + Income.composition.of.resources + Schooling + I(Adult.Mo
                      + I(BMI^2) + I(HIV.AIDS^2) + I(GDP^2) + I(thinness..1.19.years^2)
                      + I(Income.composition.of.resources^2) + I(Schooling^2) + I(thinness..1.19.years*S
                      + I(Schooling*Status), data = train_data)
evaluate_model(model_quadratic)
```

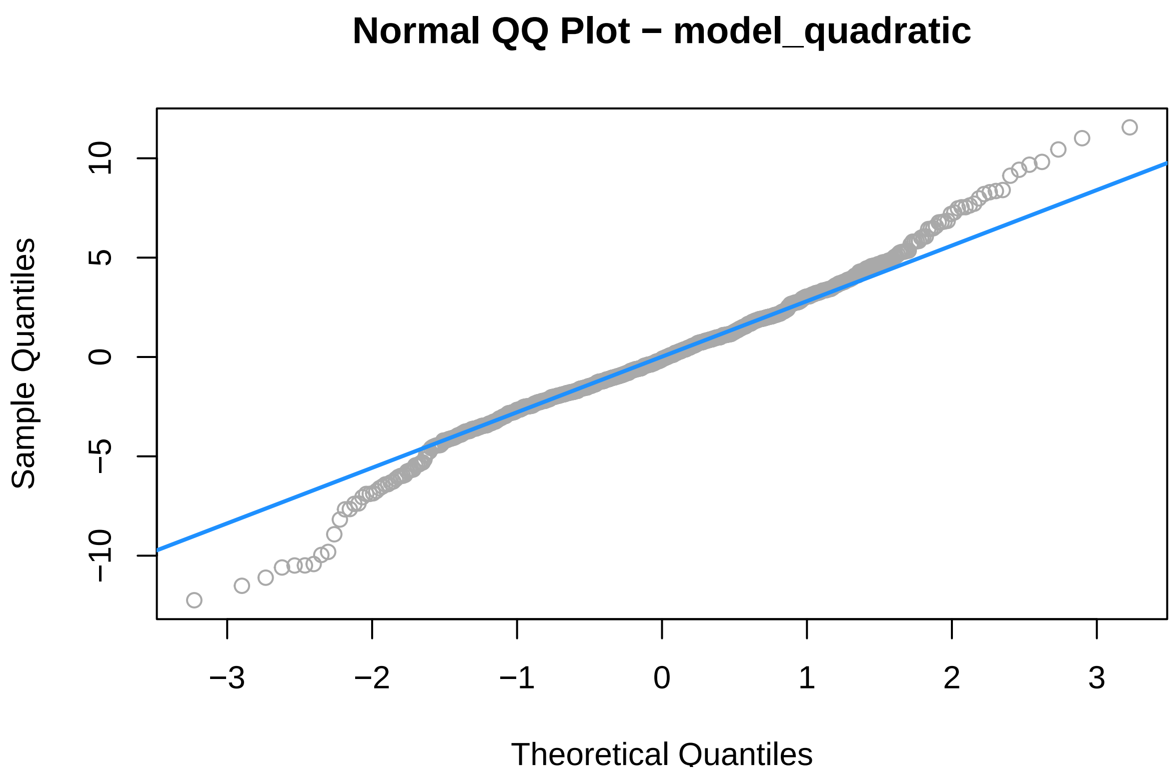
```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + BMI +
##     HIV.AIDS + GDP + thinness..1.19.years + Income.composition.of.resources +
##     Schooling + I(Adult.Mortality^2) + I(BMI^2) + I(HIV.AIDS^2) +
##     I(GDP^2) + I(thinness..1.19.years^2) + I(Income.composition.of.resources^2) +
##     I(Schooling^2) + I(thinness..1.19.years * Status) + I(Schooling *
##     Status), data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2427  -1.8701  -0.1095   1.9015  11.5569
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           6.471e+01  1.473e+00  43.925  < 2e-16 ***
## Status                                7.676e+00  4.260e+00   1.802 0.071957 .
```

```
## Adult.Mortality                     -9.801e-03  2.886e-03  -3.396 0.000719 ***
## BMI                                  4.457e-02  2.672e-02   1.668 0.095728 .
## HIV.AIDS                            -8.328e-01  8.242e-02 -10.104  < 2e-16 ***
## GDP                                 -1.944e-05  2.617e-05  -0.743 0.457787
## thinness..1.19.years                -2.240e-01  8.077e-02  -2.773 0.005689 **
## Income.composition.of.resources     -2.798e+01  3.100e+00  -9.026  < 2e-16 ***
## Schooling                            7.659e-01  2.739e-01   2.797 0.005289 **
## I(Adult.Mortality^2)                -5.313e-06  6.372e-06  -0.834 0.404634
## I(BMI^2)                            -5.798e-04  3.851e-04  -1.505 0.132638
## I(HIV.AIDS^2)                        1.101e-02  1.769e-03   6.223 7.94e-10 ***
## I(GDP^2)                             2.746e-10  3.675e-10   0.747 0.455136
## I(thinness..1.19.years^2)            7.679e-03  3.710e-03   2.070 0.038781 *
## I(Income.composition.of.resources^2)  4.683e+01  3.881e+00  12.065  < 2e-16 ***
## I(Schooling^2)                      -2.497e-02  1.349e-02  -1.851 0.064527 .
## I(thinness..1.19.years * Status)    -1.915e+00  4.678e-01  -4.095 4.67e-05 ***
## I(Schooling * Status)               -3.568e-01  2.604e-01  -1.370 0.171000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.265 on 782 degrees of freedom
## Multiple R-squared:  0.8486, Adjusted R-squared:  0.8453
## F-statistic: 257.9 on 17 and 782 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                       Df  Sum Sq Mean Sq   F value    Pr(>F)
## Status                                 1  9792.6  9792.6  918.7770 < 2.2e-16
## Adult.Mortality                        1 19647.9 19647.9 1843.4368 < 2.2e-16
## BMI                                    1  3331.0  3331.0  312.5251 < 2.2e-16
## HIV.AIDS                               1  4339.4  4339.4  407.1384 < 2.2e-16
## GDP                                    1  1212.3  1212.3  113.7437 < 2.2e-16
## thinness..1.19.years                   1   533.2   533.2   50.0295 3.365e-12
## Income.composition.of.resources        1  3411.0  3411.0  320.0346 < 2.2e-16
## Schooling                              1  1395.2  1395.2  130.9043 < 2.2e-16
## I(Adult.Mortality^2)                   1    98.4    98.4    9.2322 0.0024571
## I(BMI^2)                               1     0.1     0.1    0.0070 0.9332869
## I(HIV.AIDS^2)                          1   815.4   815.4   76.5014 < 2.2e-16
## I(GDP^2)                               1    17.5    17.5    1.6403 0.2006589
## I(thinness..1.19.years^2)              1    63.1    63.1    5.9208 0.0151864
## I(Income.composition.of.resources^2)   1  1824.3  1824.3  171.1583 < 2.2e-16
## I(Schooling^2)                         1    69.5    69.5    6.5172 0.0108726
## I(thinness..1.19.years * Status)       1   159.1   159.1   14.9246 0.0001212
## I(Schooling * Status)                  1    20.0    20.0    1.8776 0.1709996
## Residuals                            782  8334.8    10.7
##
## Status                               ***
## Adult.Mortality                      ***
## BMI                                  ***
## HIV.AIDS                             ***
## GDP                                  ***
## thinness..1.19.years                 ***
## Income.composition.of.resources      ***
## Schooling                            ***
```

```
## I(Adult.Mortality^2)                **
## I(BMI^2)
## I(HIV.AIDS^2)                       ***
## I(GDP^2)
## I(thinness..1.19.years^2)           *
## I(Income.composition.of.resources^2) ***
## I(Schooling^2)                      *
## I(thinness..1.19.years * Status)    ***
## I(Schooling * Status)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

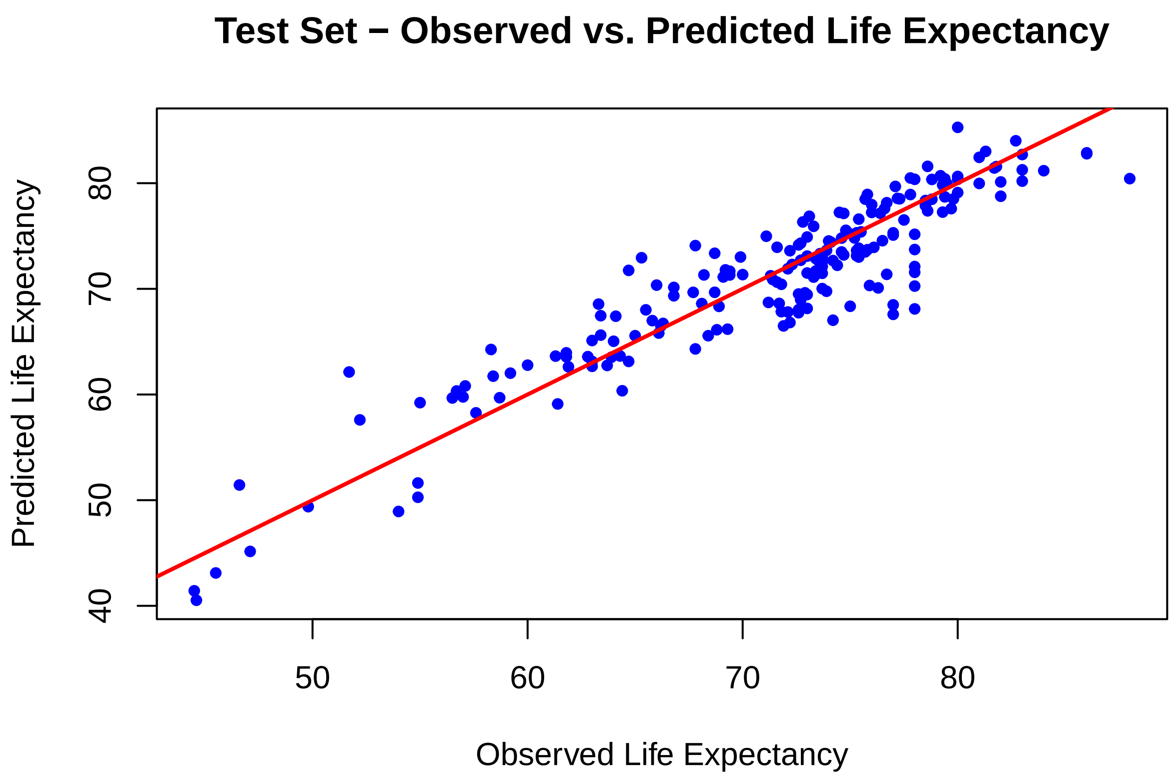



```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 100.11, df = 17, p-value = 8.502e-14
```



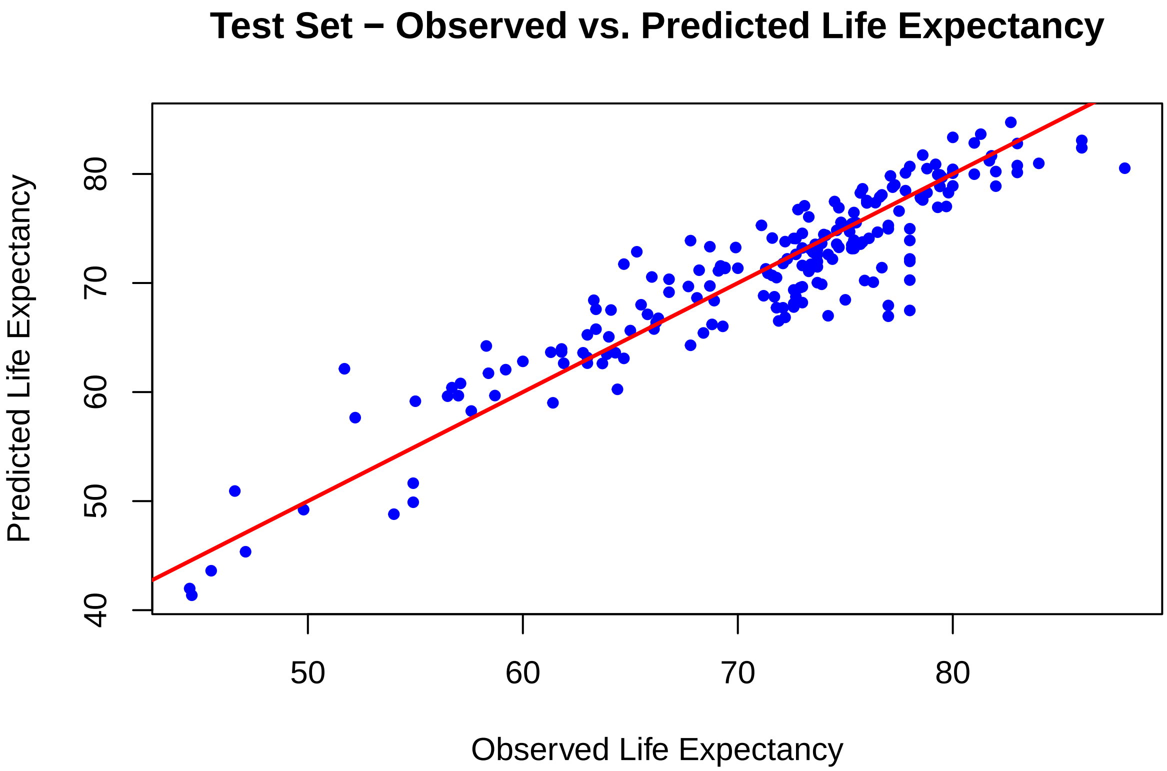

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98262, p-value = 3.804e-08
##
## Train Set MSE: 10.41847
## Train R-squared: 0.8486364
```

```r
evaluate_test(model_quadratic, test_data)
```



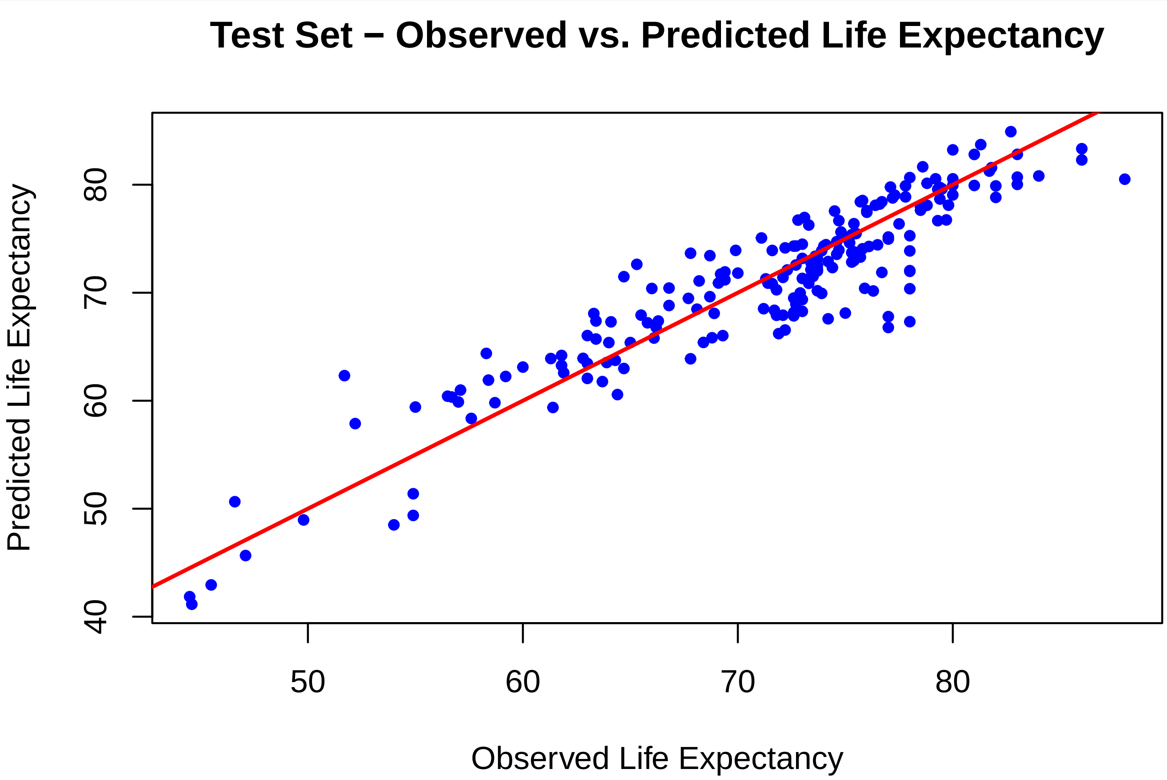

```
## Test Set MSE: 10.21839
## Test R-squared: 0.8523176
```

```
#use this as full model, do a AIC and BIC backward selection:
model_AIC <- step(model_quadratic, direction = "backward", trace = 0)
model_BIC <- step(model_quadratic, direction = "backward", k = log(nrow(train_data)), trace = 0)

evaluate_test(model_AIC, test_data)
```




```
## Test Set MSE: 10.30219
## Test R-squared: 0.8511065
```

```r
evaluate_test(model_BIC, test_data)
```




```
## Test Set MSE: 10.51515
```

```
## Test R-squared: 0.8480288
```

```r
#compare the number of predictor used
num_predictors_AIC <- sum(!is.na(coef(model_AIC)))
cat("Number of predictors in AIC-selected model:", num_predictors_AIC, "\n")
```

```
## Number of predictors in AIC-selected model: 14
```

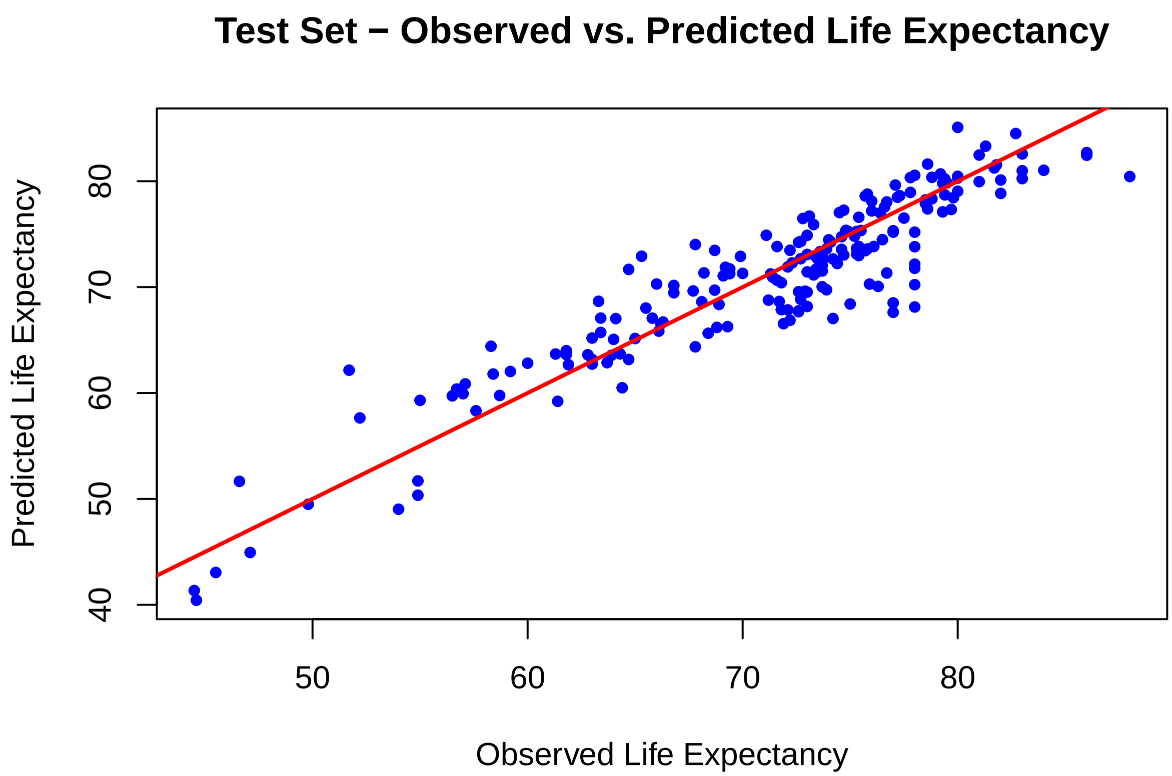
```r
num_predictors_BIC <- sum(!is.na(coef(model_BIC)))
cat("Number of predictors in BIC-selected model:", num_predictors_BIC, "\n")
```

```
## Number of predictors in BIC-selected model: 11
```

```r
#extract predictors and train variable
X <- model.matrix(model_quadratic)[, -1]
y <- train_data$Life.expectancy

#array of lambda to try
lambda_values <- 10^seq(10, -2, length = 100)

#ridge regression with cross validation
ridge_cv_model <- cv.glmnet(X, y, alpha = 0, lambda = lambda_values)

#get best lambda
best_lambda <- ridge_cv_model$lambda.min
cat("Best Lambda:", best_lambda, "\n")
```

```
## Best Lambda: 0.01
```

```r
#construct a ridge model with best lambda from CV,
#train model
model_ridge <- glmnet(X, y, alpha = 0, lambda = best_lambda)

#format corresponding test dataset
X_test <- model.matrix(model_quadratic, data = test_data)[, -1]
y_test <- test_data$Life.expectancy

#prediction
y_hat <- predict(model_ridge, s = best_lambda, newx = X_test)


R2 <- 1 - sum((y_test - y_hat)^2) / sum((y_test - mean(y_test))^2)
MSE <- mean((y_test - y_hat)^2)

plot(y_test, y_hat, col = "blue", pch = 20,
     xlab = "Observed Life Expectancy", ylab = "Predicted Life Expectancy",
     main = "Test Set - Observed vs. Predicted Life Expectancy")
abline(0, 1, col = "red", lwd = 2)
```




```r
cat("Ridge Test Set MSE:", MSE, "\n")
```

```
## Ridge Test Set MSE: 10.23504
```

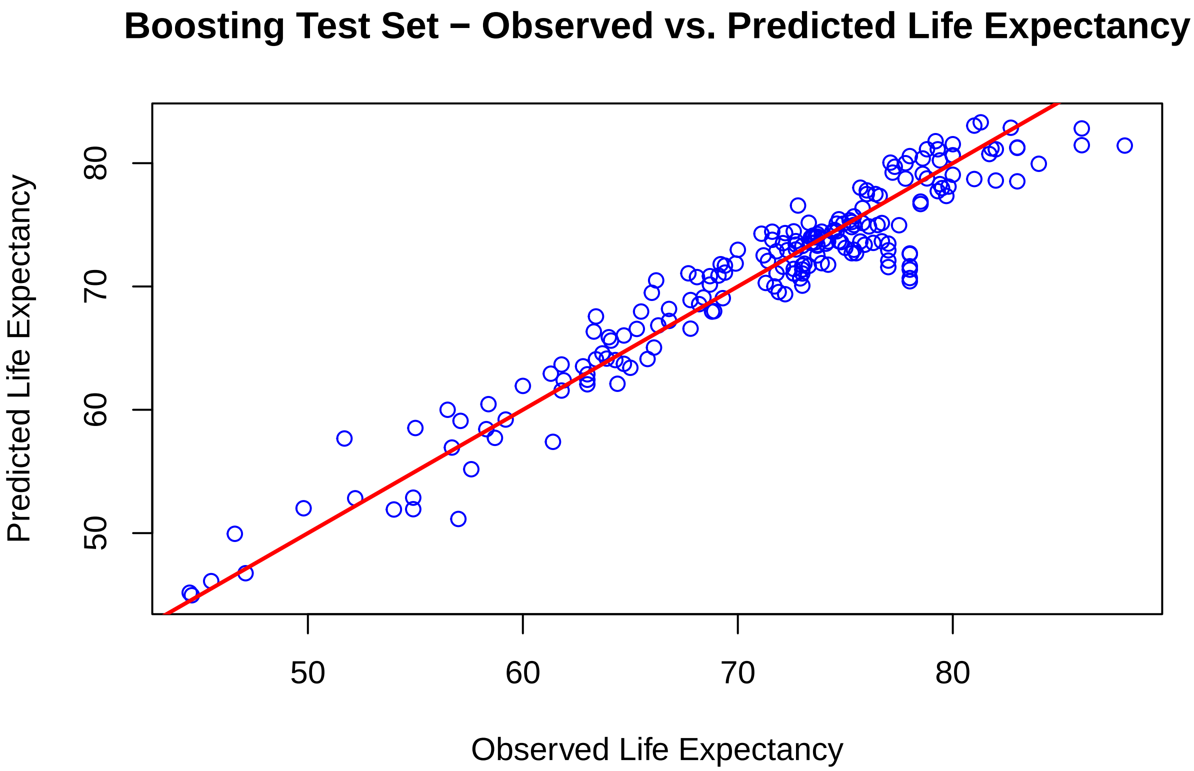
```r
cat("Ridge Test Set R-squared:", R2, "\n")
```

```
## Ridge Test Set R-squared: 0.8520771
```

```r
#create boosting model with 100 trees
model_boost <- gbm(Life.expectancy ~ ., data = train_data, distribution = "gaussian", n.trees = 100, in

#evaluate model on test set
y <- test_data$Life.expectancy
y_hat <- predict(model_boost, newdata = test_data)
```

```
## Using 100 trees...
```

```r
# Plot observed vs. predicted values
plot(y, y_hat, col = "blue",
     xlab = "Observed Life Expectancy", ylab = "Predicted Life Expectancy",
     main = "Boosting Test Set - Observed vs. Predicted Life Expectancy")
abline(0, 1, col = "red", lwd = 2)
```




```r
#MSE
MSE <- mean((y - y_hat)^2)
cat("Boosting Test Set MSE:", MSE, "\n")
```

```
## Boosting Test Set MSE: 5.389733
```

```r
#R2
R2 <- 1 - (sum((y - y_hat)^2) / sum((y - mean(y_hat))^2))
cat("Boosting Test Set R-Squared:", R2, "\n")
```

```
## Boosting Test Set R-Squared: 0.9221696
```

```r
#bagging model
model_bagging <- randomForest(Life.expectancy ~ Status + Adult.Mortality + BMI + HIV.AIDS + GDP
                          + thinness..1.19.years + Income.composition.of.resources + Schooling + I(Adult.Mo
                          + I(BMI^2) + I(HIV.AIDS^2) + I(GDP^2) + I(thinness..1.19.years^2)
                          + I(Income.composition.of.resources^2) + I(Schooling^2) + I(thinness..1.19.years*
                          + I(Schooling*Status), data = train_data,
                          mtry = 20, importance = TRUE, ntree = 150, oob = TRUE)
```

```
## Warning in randomForest.default(m, y, ...): invalid mtry: reset to within valid
## range
```

```r
# Evaluate the bagging model

summary(model_bagging)
```

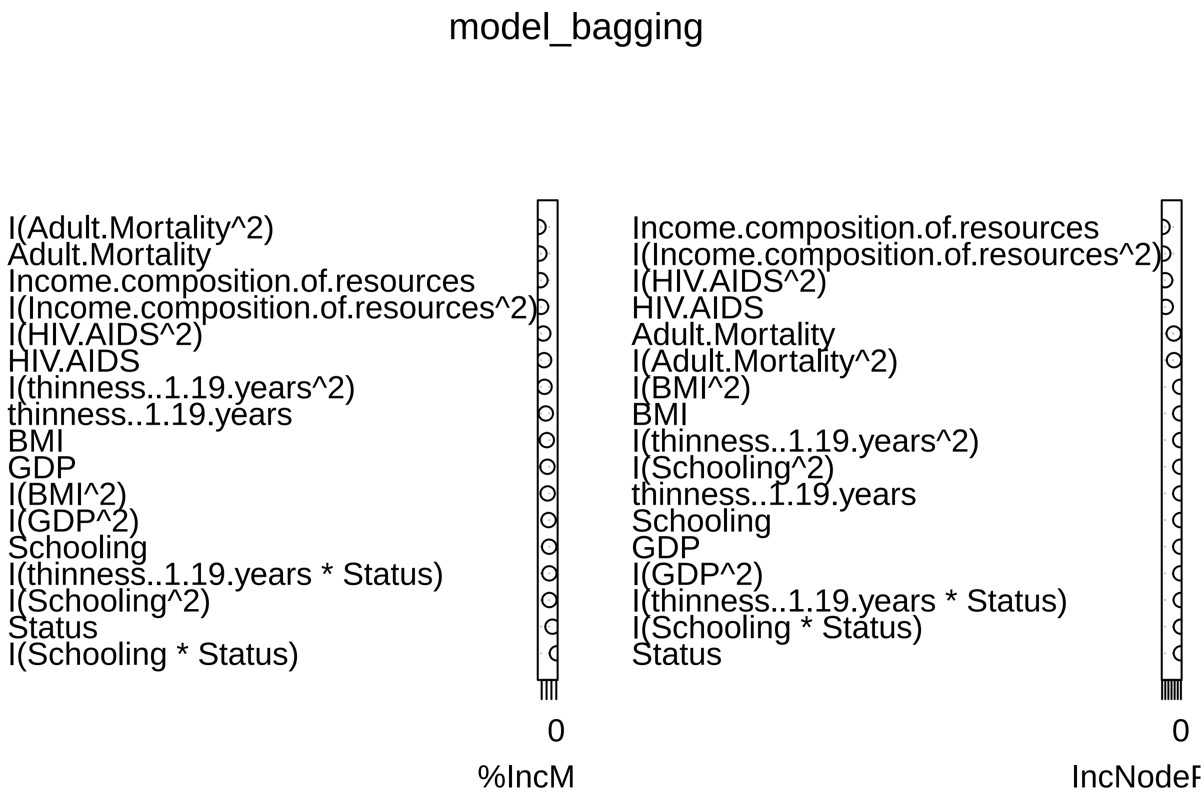
```
##                 Length Class  Mode
## call              7    -none- call
## type              1    -none- character
## predicted       800    -none- numeric
## mse             150    -none- numeric
```

```
## rsq             150    -none- numeric
## oob.times       800    -none- numeric
## importance       34    -none- numeric
## importanceSD     17    -none- numeric
## localImportance   0    -none- NULL
## proximity         0    -none- NULL
## ntree             1    -none- numeric
## mtry              1    -none- numeric
## forest           11    -none- list
## coefs             0    -none- NULL
## y               800    -none- numeric
## test              0    -none- NULL
## inbag             0    -none- NULL
## terms             3    terms  call
```

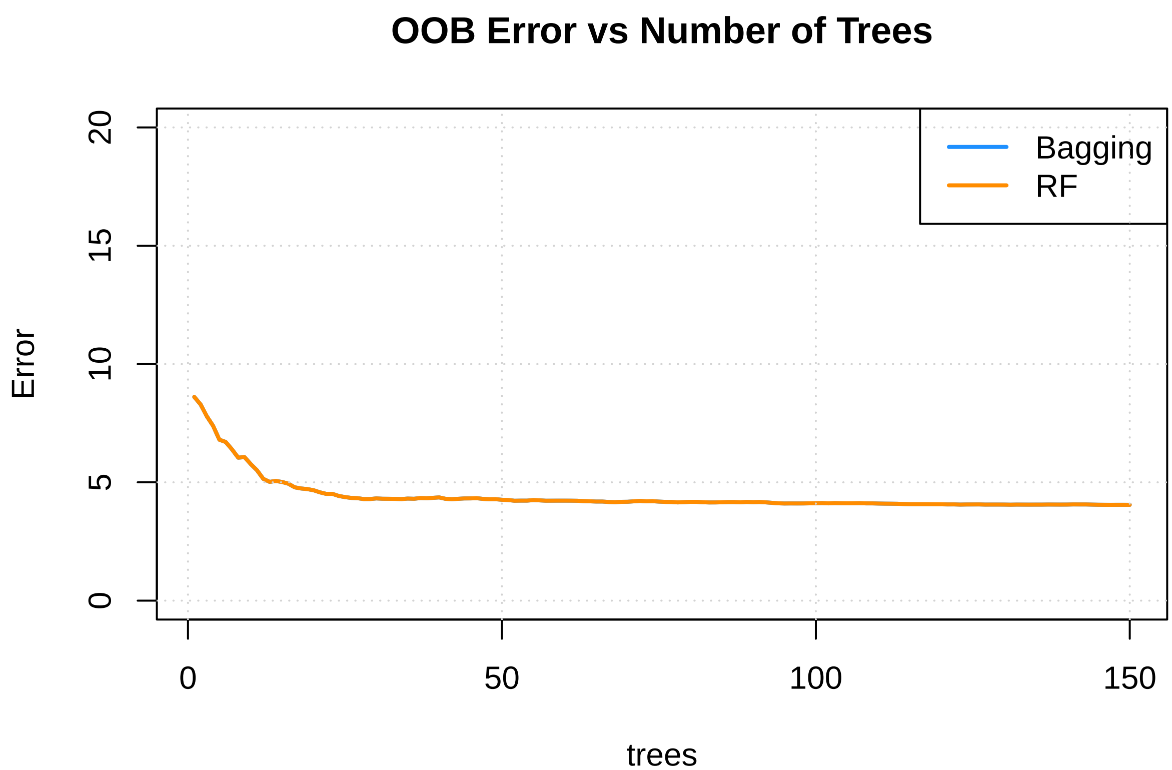
```r
# 1st col: OOB sample error based (i.e. prediction accuracy based)
# 2nd col: SSE based
importance(model_bagging)
```

```
##                                         %IncMSE IncNodePurity
## Status                                3.9009500      8.599175
## Adult.Mortality                      17.3691042   4556.000097
## BMI                                   9.7360829    442.756585
## HIV.AIDS                             12.5008563   9696.647447
## GDP                                   9.0973260    380.259694
## thinness..1.19.years                 10.7687782    406.694868
## Income.composition.of.resources      16.3427463  11759.208603
## Schooling                             7.3307657    392.813607
## I(Adult.Mortality^2)                 18.2891780   4395.600469
## I(BMI^2)                              8.8944010    471.374966
## I(HIV.AIDS^2)                        13.3002806  10066.424374
## I(GDP^2)                              7.9712462    361.732254
## I(thinness..1.19.years^2)            12.0550879    429.972167
## I(Income.composition.of.resources^2) 15.6734661  11262.174880
## I(Schooling^2)                        7.0918014    429.196442
## I(thinness..1.19.years * Status)      7.1127263     93.826994
## I(Schooling * Status)                -0.6490934     30.645217
```

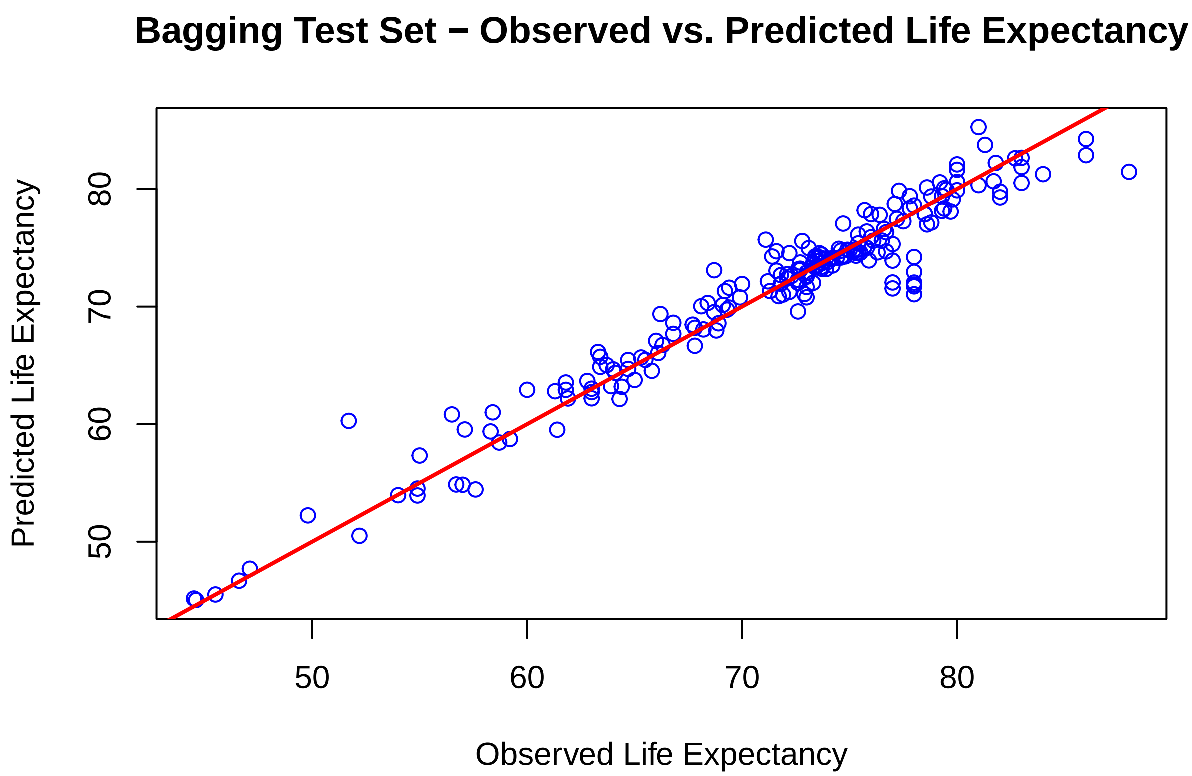
```r
varImpPlot(model_bagging)
```




```r
#see how oob error decreases as we put more trees.
plot(model_bagging, col = "dodgerblue", lwd = 2,
     main = "OOB Error vs Number of Trees", ylim=c(-0,20))
plot(model_bagging, add=TRUE,col = "darkorange", lwd = 2,
     main = "Bagged Trees: Error vs Number of Trees", xlim=c(0,10))
legend("topright",c("Bagging","RF"),col=c("dodgerblue","darkorange"),lwd=2)
grid()
```




```r
y <- test_data$Life.expectancy
y_hat <- predict(model_bagging, newdata = test_data)

# Plot observed vs. predicted values
plot(y, y_hat, col = "blue",
     xlab = "Observed Life Expectancy", ylab = "Predicted Life Expectancy",
     main = "Bagging Test Set - Observed vs. Predicted Life Expectancy")
abline(0, 1, col = "red", lwd = 2)
```




```
#MSE
MSE <- mean((y - y_hat)^2)
cat("Bagging Test Set MSE:", MSE, "\n")
```

```
## Bagging Test Set MSE: 3.945235
```

```
#R2
R2 <- 1 - (sum((y - y_hat)^2) / sum((y - mean(y_hat))^2))
cat("Bagging Test Set R-Squared:", R2, "\n")
```

```
## Bagging Test Set R-Squared: 0.9429815
```