Prediction of Insurance Costs

```
# Loading data
library(readr)

insurance_data <- read_csv("insurance_data.csv")
## -- Column specification
## cols(
##   age = col_double(),
##   sex = col_character(),
##   bmi = col_double(),
##   children = col_double(),
##   smoker = col_character(),
##   region = col_character(),
##   charges = col_double()
## )

##Understanding the Data

data <- insurance_data
head(data)

## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
```

```
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest 21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female 25.7        0 no     southeast   3757.
```

```
str(data)
```

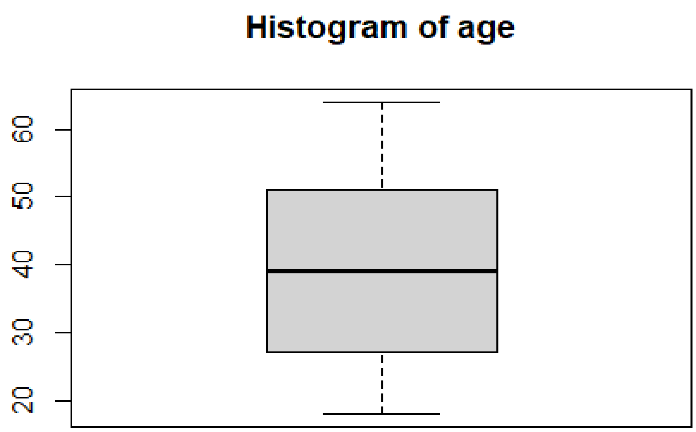
```
## spec_tbl_df [1,338 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
```

```
##   ..   charges = col_double()

##   .. )

# Coverting sex, smoker ad region to data type factor

data$sex <- as.factor(data$sex)

data$smoker <- factor(data$smoker,

                      levels = c("no","yes"),

                      labels = c(0,1))

data$region <- as.factor(data$region)

Summary (data)

##      age            sex            bmi             children          smoker
##  Min.   :18.00   female:662   Min.   :15.96    Min.   :0.000      0:1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30    1st Qu.:0.000      1: 274
##  Median :39.00                Median :30.40    Median :1.000
##  Mean   :39.21                Mean   :30.66    Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69    3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13    Max.   :5.000

##       region              charges
##  northeast:324         Min.   : 1122
##  northwest:325         1st Qu.: 4740
##  southeast:364         Median : 9382
##  southwest:325         Mean   :13270
##                        3rd Qu.:16640
##                        Max.   :63770
```

The dataset contains 1338 observations of 7 variables. The variable charges is the one we have to predict using the following predictors: age, sex, bmi, children, smoker and region. The variable age and bmi are continuous variables, the variables sex, smoker and region are categorical variables. There are no missing variables in the dataset. The five number summary is also displayed above.

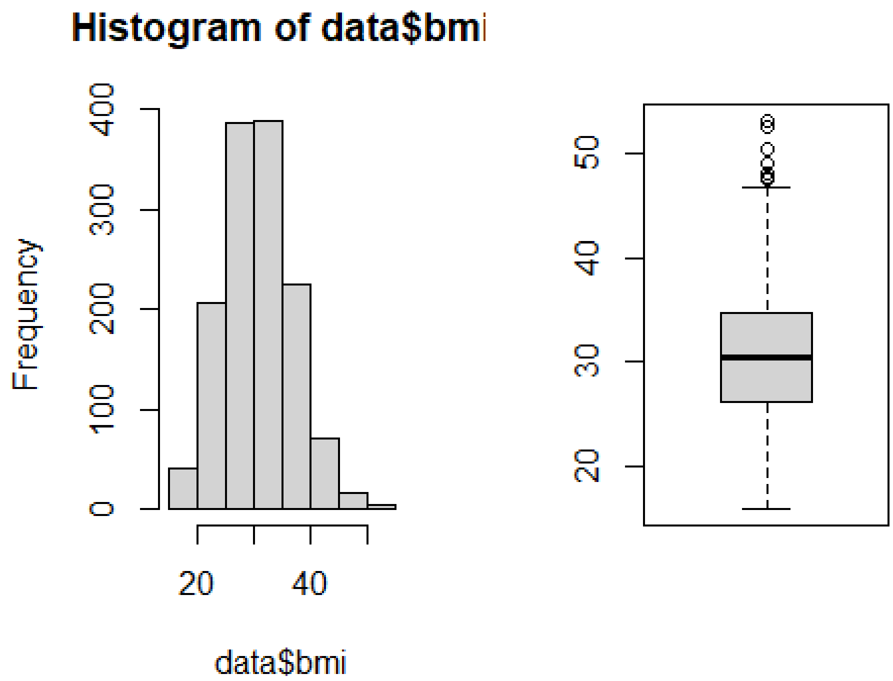
Exploratory data Analysis

```
# Distribution of age in the dataset
boxplot(data$age, main = "Histogram of age")
```



Age is distributed normally as depicted by the boxplot. The lowest age is 18, with the highest being 64.

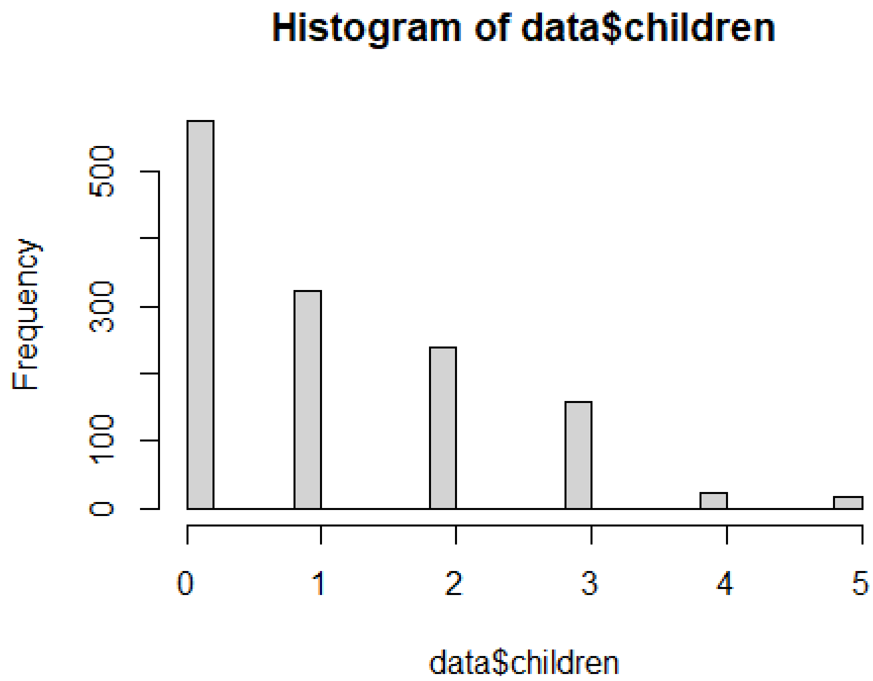
```
# Distribution of bmi in the dataset
par(mfrow = c(1,2))
hist(data$bmi)
boxplot(data$bmi)
```



Bmi is approximately distributed normally. Majority of the bmi is between 20 to 40.

 # Distribution of children in the dataset

hist(data**$**children,breaks=20)



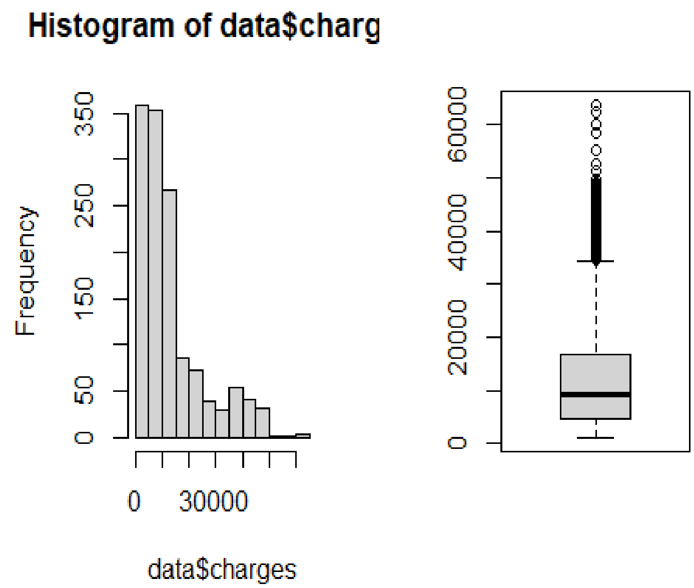
The histogram for children against frequency is right skewed, showing that majority of the people have no children.

 # Distribution of charges in the dataset

par(mfrow = **c**(1,2))

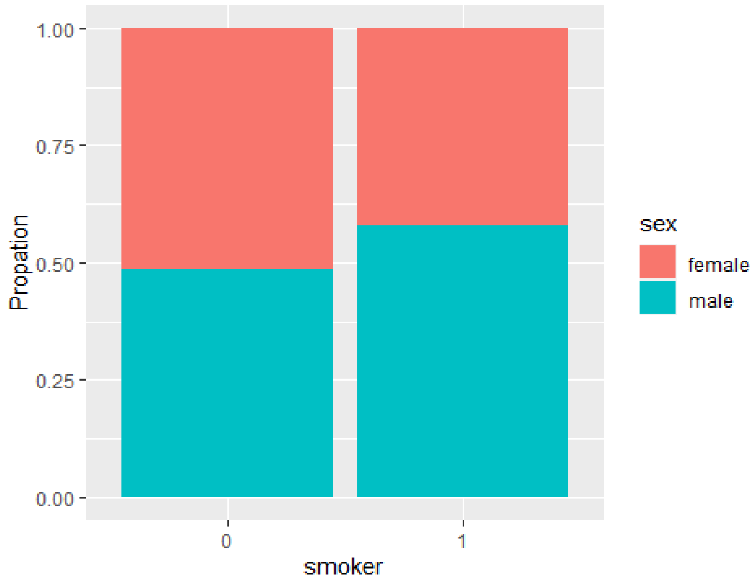
hist(data**$**charges)

boxplot(data**$**charges)



Charges are right skewed in the data with many outliers. The outliers are values above a charge of 30,000 depicted by the boxplot.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

ggplot(data, aes(x=smoker, fill=sex)) +
  geom_bar(position = "fill" )+
  labs(y="Propation")
```

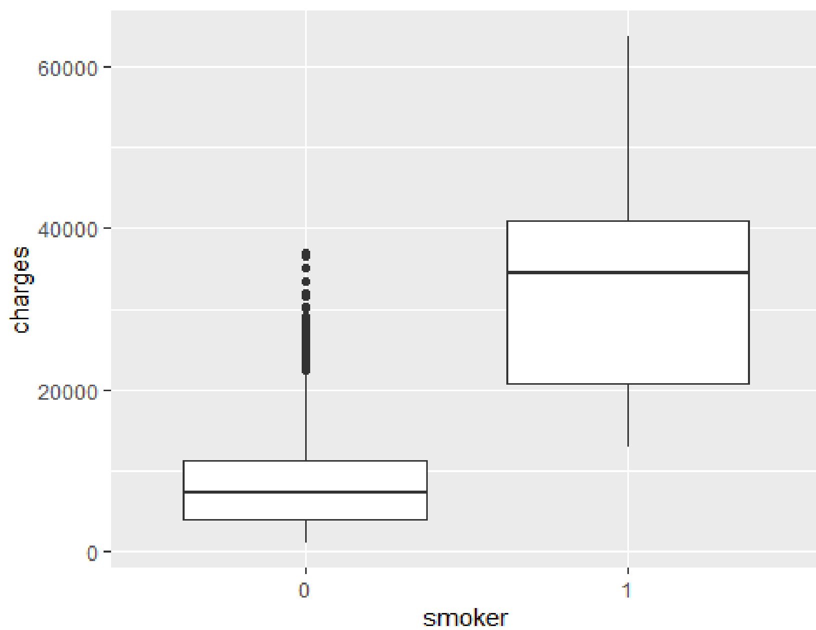


There is a slightly higher proportion of females who do not smoke than males who do not smoke. There is a higher proportion of male smokers than female smokers. The rates between males and females are approximately the same.

```
ggplot(data, aes(x=region, fill=smoker)) +
  geom_bar(position = "fill") +
  labs(y="Proportion")
```

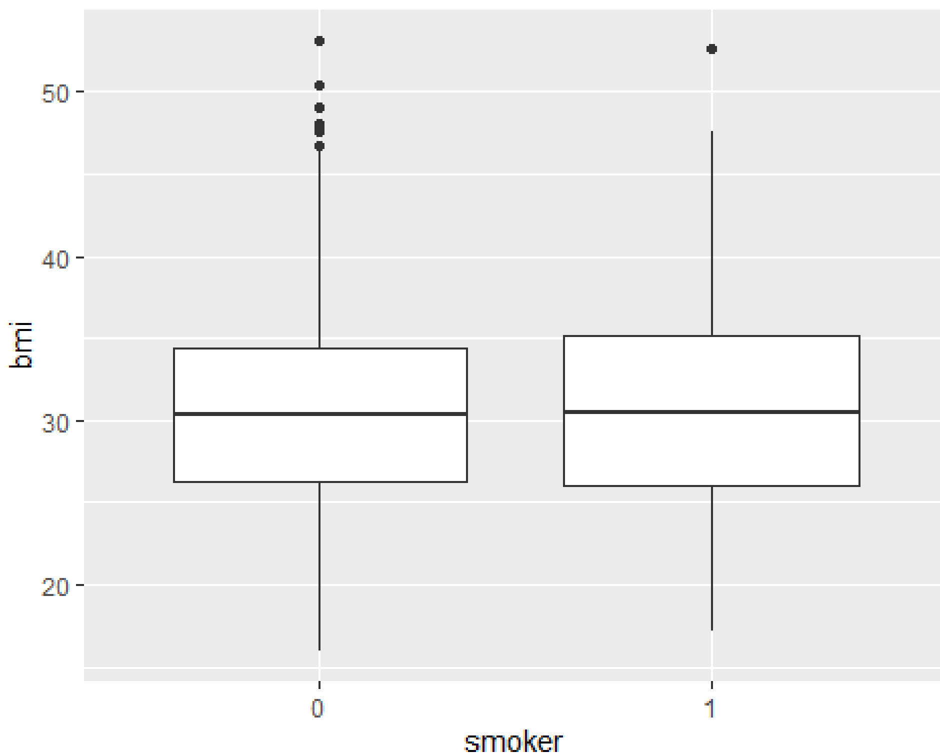
There is a larger proportion of non-smokers than smokers in all the four regions.

```
ggplot(data, aes(smoker, charges)) +
  geom_boxplot()
```



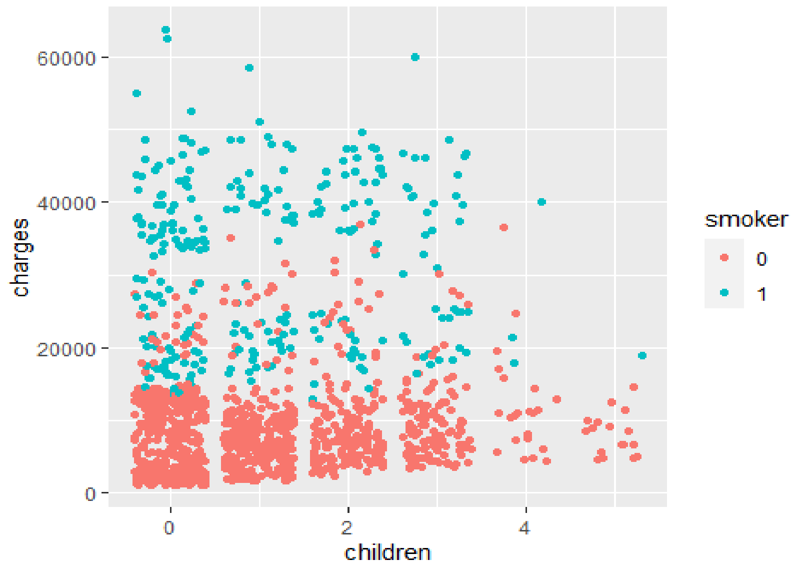
The median for non-smokers is lower than that of smokers. The non-smokers also have lower insurance charges as compared to smokers. The smokers have above 30,000 while the non-smokers have a median of less than 10,000.

```
ggplot(data, aes(smoker, bmi)) +
  geom_boxplot()
```



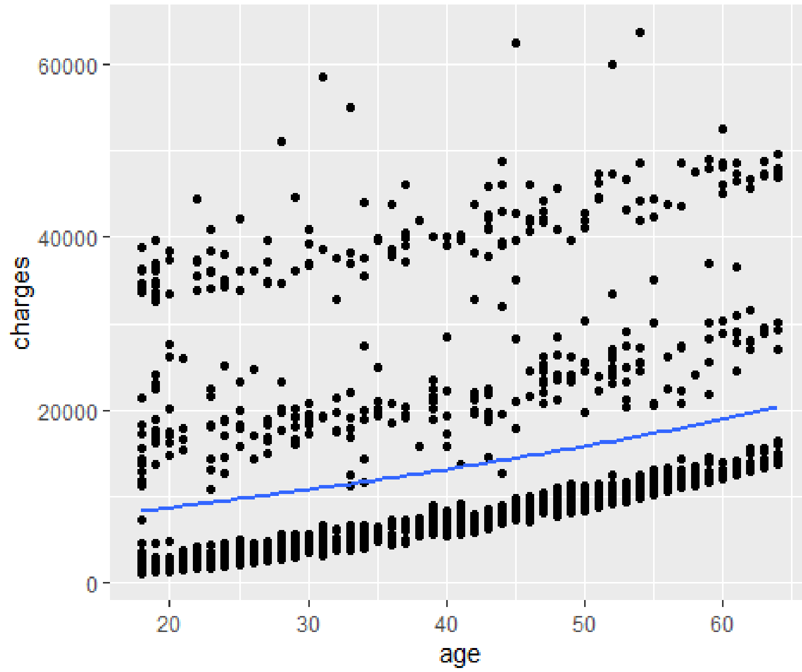
The median for both smokers and non-smokers are approximately equal.

```
ggplot(data, aes(x=children, y=charges, color=smoker)) +
  geom_jitter()
```

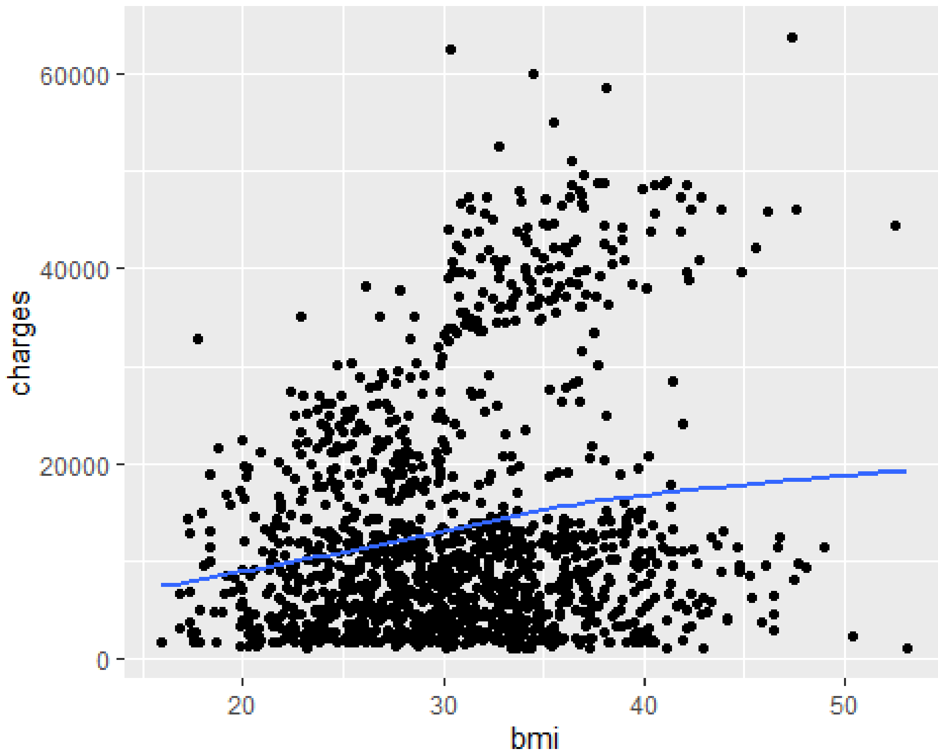


Majority of the people have children 2 or fewer children. Smokers have higher charges than non-smokers. On average, people with more children pay higher charges as opposed to people with no children.

```
ggplot(data, aes(x=age, y=charges)) +

  geom_point() +

  geom_smooth(se = F)


## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

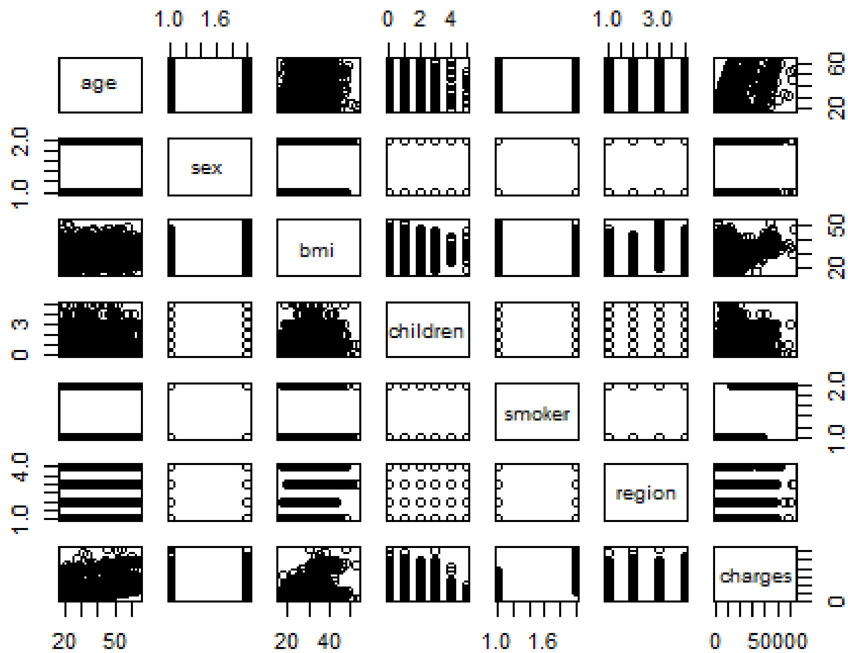



```
ggplot(data, aes(x=bmi, y=charges)) +
  geom_point() +
  geom_smooth(se = F)

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
data_corr <- pairs(data)
```

In the first plot, we see that there is a trend that with older age the charges increase. There are also three groups/lines visible. In the second plot we see some sort of trend that with increasing bmi the charges increase, however this is not very clear.




```
data_corr

## NULL
```

Regression Analysis

Splitting the data into a train set and test set

```
library(caTools)

## Warning: package 'caTools' was built under R version 4.0.5

set.seed(200)

sample <- sample.split(data$charges,
```

```
                     SplitRatio = 0.75)
train_data <- subset(data, sample == T)
test_data <- subset(data, sample == F)
dim(train_data)

## [1] 1003    7

dim(test_data)

## [1] 335   7
```

MODEL BUILDING

```
model_1 <- lm("charges ~ age + sex + bmi + children + region + smoker", data = data)
summary(model_1)

##
## Call:
## lm(formula = "charges ~ age + sex + bmi + children + region + smoker",
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age               256.9       11.9  21.587  < 2e-16 ***
## sex              -131.3      332.9  -0.394 0.693348
## bmi               339.2       28.6  11.860  < 2e-16 ***
## children          475.5      137.8   3.451 0.000577 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## smoker1         23848.5      413.1  57.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16

r_sq_1 <- summary(model_1)$r.squared
r_sq_1 #0.750913

## [1] 0.750913

predict_1 <- predict(model_1, newdata = test_data)
residuals_1 <- test_data$charges - predict_1
rmse_1 <- sqrt(mean(residuals_1^2))
rmse_1 #6805.822
```

```
## [1] 6805.822

model_2 <- lm("charges ~ age + bmi + children + region + smoker", data = data)

summary(model_2)

##

## Call:

## lm(formula = "charges ~ age + bmi + children + region + smoker",

##     data = data)

##

## Residuals:

##      Min       1Q   Median       3Q      Max

## -11367.2  -2835.4   -979.7   1361.9  29935.5

##

## Coefficients:

##                  Estimate Std. Error t value Pr(>|t|)

## (Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***

## age                256.97      11.89  21.610  < 2e-16 ***

## bmi                338.66      28.56  11.858  < 2e-16 ***

## children           474.57     137.74   3.445 0.000588 ***

## regionnorthwest   -352.18     476.12  -0.740 0.459618

## regionsoutheast  -1034.36     478.54  -2.162 0.030834 *

## regionsouthwest   -959.37     477.78  -2.008 0.044846 *

## smoker1          23836.30     411.86  57.875  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

```
r_sq_2 <- summary(model_2)$r.squared
r_sq_2 #0.7508839
```

```
## [1] 0.7508839
```

```
predict_2 <- predict(model_2, newdata = test_data)
residuals_2 <- test_data$charges - predict_2
rmse_2 <- sqrt(mean(residuals_2^2))
rmse_2 #6805.863
```

```
## [1] 6805.863
```

```
library(car)
```

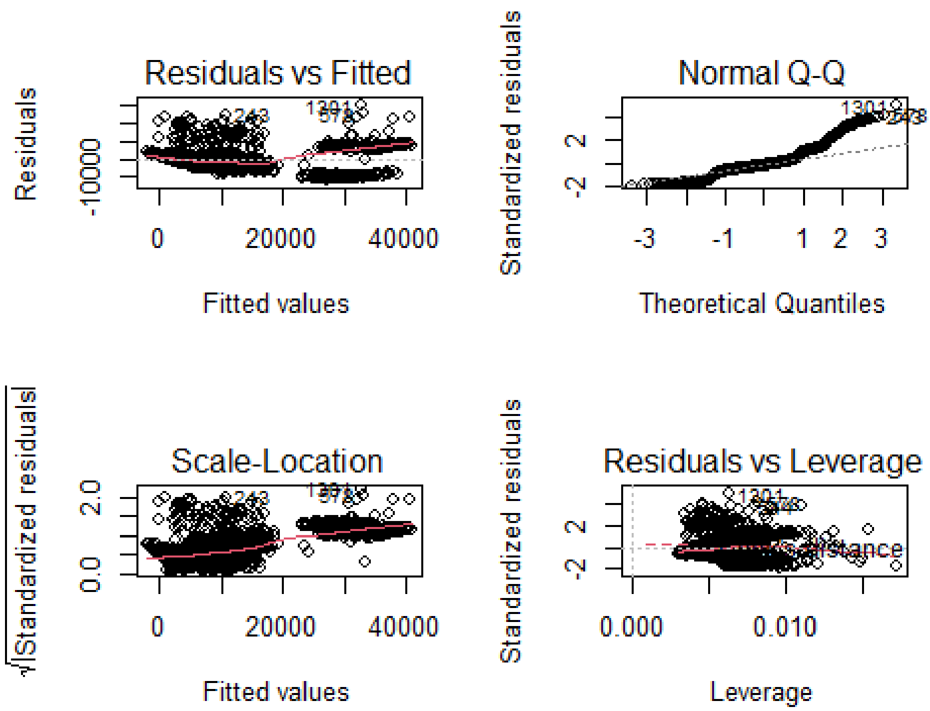
```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```
# vif
vif(model_2)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.016188  1        1.008061
## bmi      1.104197  1        1.050808
## children 1.003714  1        1.001855
## region   1.098870  3        1.015838
## smoker   1.006369  1        1.003179

# tolerance
1/vif(model_2)

##               GVIF        Df GVIF^(1/(2*Df))
## age      0.9840702 1.0000000       0.9920031
## bmi      0.9056351 1.0000000       0.9516486
## children 0.9962997 1.0000000       0.9981481
## region   0.9100262 0.3333333       0.9844092
## smoker   0.9936716 1.0000000       0.9968308

mean(vif(model_2))

## [1] 1.153939

par(mfrow = c(2,2))
plot(model_2)
```



Some variables from the model are not significant (sex), while others are significant (age, bmi, smoker, children and region). Training the model will happen without non-significant variables to find out if the model gets improved. After the training, the performance of the two models is similar. Model_1 has an R square of 0.750913 with a root mean square error of 6805.822. Model_2 has an R square of 0.7508839 with a root mean square error of 6805.863.

Model_2 will get used since it is simpler than model_1.

To check the assumptions for regression, muticollinearity gets checked using the *vif* function, for which the mean *vif* for model_2 is 1.153939. The smallest possible value of *vif* is 1; hence, there is minimal correlation amongst the independent variables leading to a small amount of inflation.