Medical Cost Personal Dataset Analysis

Contents

1 Introduction

The goal of this project is to explore the Medical Insurance Dataset to predict accurately the medical costs based on some explanatory variables. We are interested to find out which variables affect the costs and We will apply linear regression model to select the best fit for our analysis.

2 Knowing the Data

The *Medical Cost Personal Dataset* has been obtained from Kaggle provided from "Machine Learning with R" by Brett Lantz.

This dataset contains 1338 observations and 7 variables including:

- age: age of individuals
- sex: gender; female or male
- bmi: Body mass index, a measure of body fat based on height and weight
- children: Number of children of individuals
- smoker: smoker or non-smoker
- region: the residential area of individuals in the US, northeast, southeast, southwest, northwest
- charges: Individual medical costs billed by health insurance

The response variable (dependent) that we are going to predict is "charges".

3 Loading libraries

```
library(readr)
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## 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(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.1.3
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.1.3
```

```
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.1.3
```

```
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.1.3
```

```
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 4.1.3
```

```
##
## Attaching package: 'modelr'
```

```
## The following object is masked from 'package:broom':
##
##     bootstrap
```

4 Importing Dataset

```
insurance <- read_csv("D:/_UniPD/Semster 2/Statistical Learning B/PROJECT/Dataset/insurance.csv")
```

```
## Rows: 1338 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
attach(insurance)
```

5 Knowing the Dataset

```
str(insurance)
```

```
## 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()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

According to`str()` we find out that we are dealing with 3 categorical variables including sex, smoker, region and 4 numerical variables. We convert them to factors and take a look at `summary(insurance)` to explore data further.

```
# converting categorical variables to factors
insurance$sex <- as.factor(insurance$sex)
insurance$smoker <- as.factor(insurance$smoker)
insurance$region <- as.factor(insurance$region)

summary(insurance)
```

```
##       age            sex           bmi           children        smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 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
```

Briefly, from the `summary()` it is evident that sex and region variables are evenly distributed. The age of individuals is between 18 and 64 years old. Furthermore, the number of non-smokers is approximately 4 times more than the number of smokers. The number of children is from 0 to 5. Also, the average charge is 13279 USD.

6 Data Cleaning

```r
# Checking for duplicated rows
cat("number of duplicated rows = ",sum(duplicated(insurance)))
```

```
## number of duplicated rows =  1
```

```r
# Removing the duplicated row
insurance <- distinct(insurance)
dim(insurance)
```

```
## [1] 1337    7
```

```r
# Checking for null values
cat("number of null values = ", sum(is.na(insurance)))
```

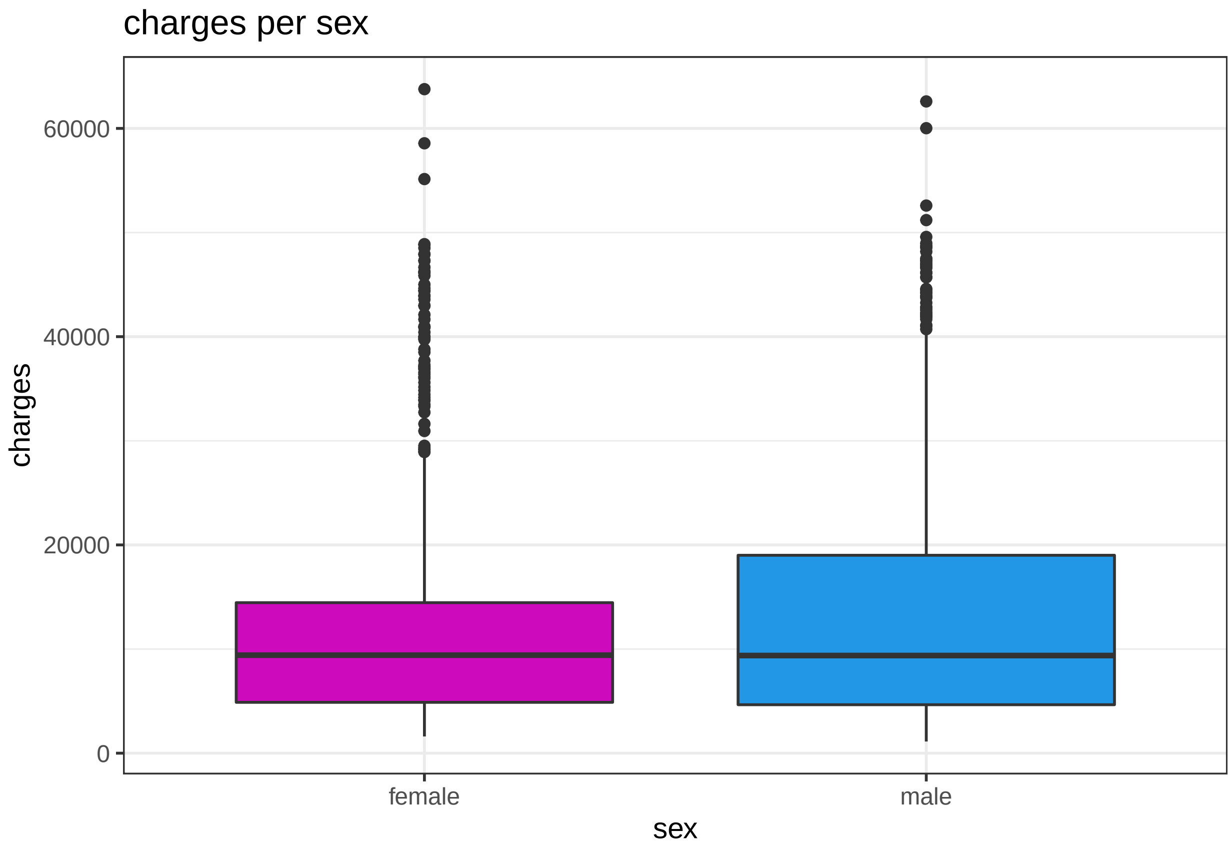
```
## number of null values =  0
```

We do not have any NA's and there was only one duplicated row which is removed.

7 Exploratory Data Analysis (EDA)

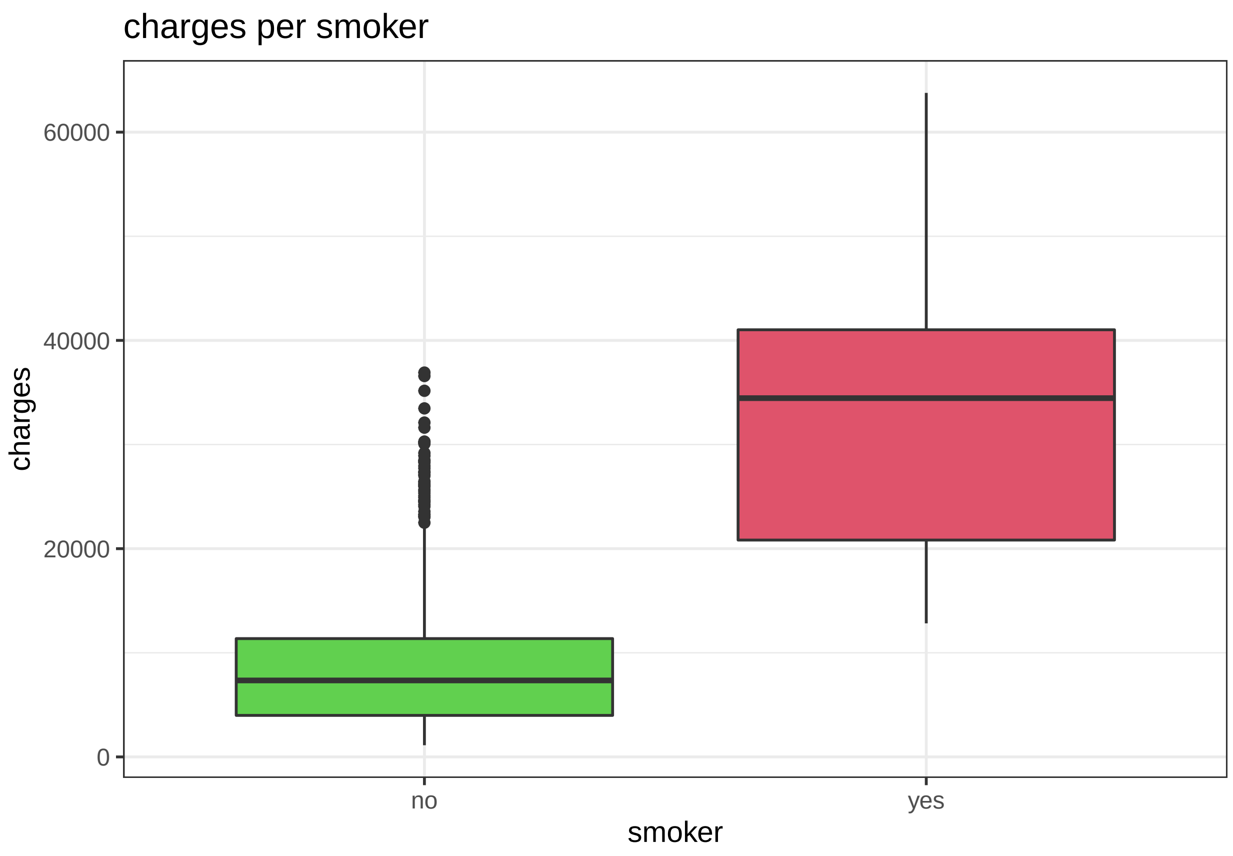
7.1 Boxplots

```
ggplot(data = insurance,aes(sex,charges)) + geom_boxplot(fill = c(6,4)) +
  theme_bw() + ggtitle("charges per sex")
```



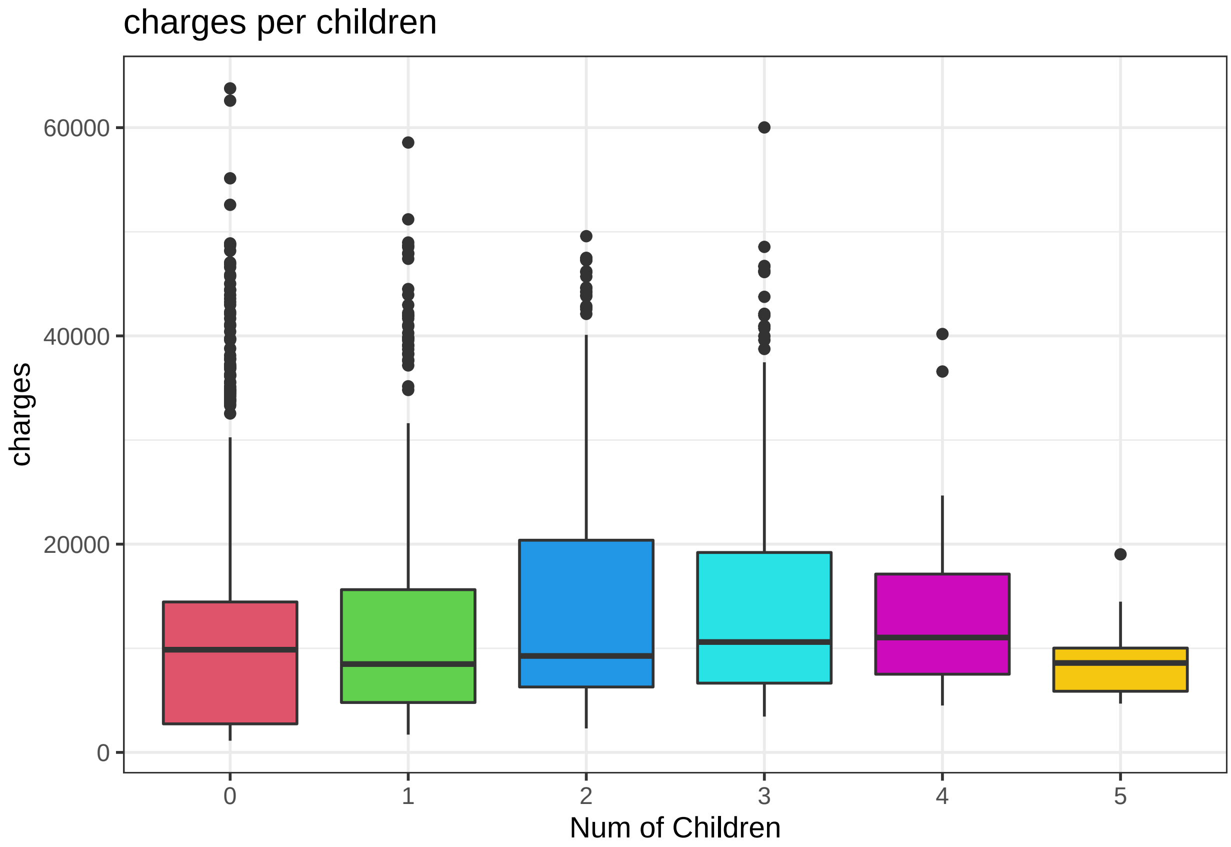
Insight: the plot shows that there is not a notable difference in charges for male and females. So, we can conclude that the charges is not affected by gender.

```
ggplot(data = insurance,aes(smoker,charges)) + geom_boxplot(fill = c(3,2)) +
  theme_bw() + ggtitle("charges per smoker")
```



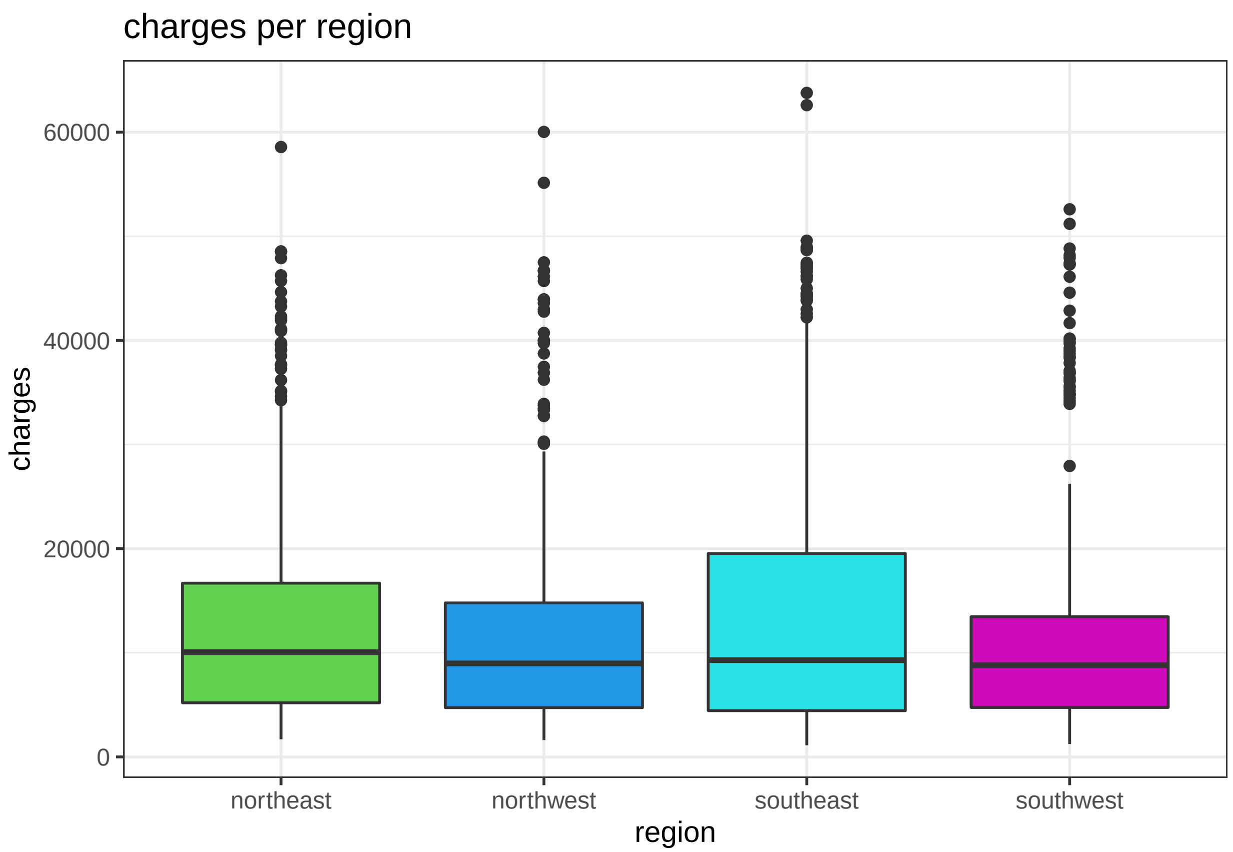
Insight: smokers have higher medical charges compared with non-smokers; therefore, it seems to be a significant predictor for our analysis.

```
ggplot(data = insurance,aes(as.factor(children),charges)) + geom_boxplot(fill = c(2:7)) +
  theme_bw() + ggtitle("charges per children") +
  xlab("Num of Children")
```



Insight: We normally expect to witness more medical charges for people with more children but, as we can see from this plot, people with 5 children, on average, have lower medical costs in comparison with other people with 0 to 4 children. It is surprising to know that groups with no child have higher insurance charges than other groups.

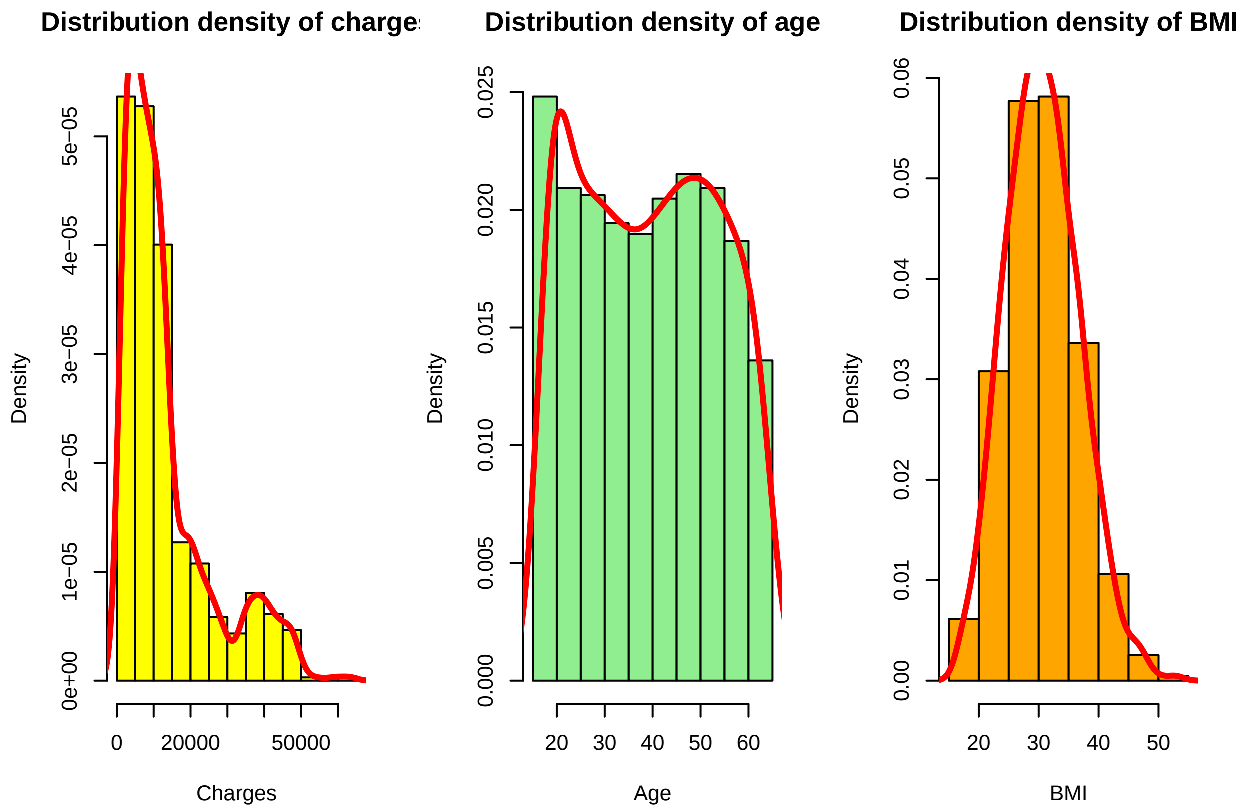
```r
ggplot(data = insurance,aes(region,charges)) + geom_boxplot(fill = c(3:6)) +
  theme_bw() + ggtitle("charges per region")
```



Insight: There is no impact on charges based on different regions. Thus, region seems to be an insignificant variable to our analysis.

7.2 Histograms and Density

```
par(mfrow = c(1,3))

hist(charges, prob = TRUE,  col = "yellow", main = "Distribution density of charges",
     xlab="Charges")
lines(density(charges), col="red", lwd = 3)

hist(age, prob = TRUE, col = "lightgreen", main = "Distribution density of age",
     xlab="Age")
lines(density(age), col="red", lwd = 3)

hist(bmi,prob = TRUE, col = "orange", main = "Distribution density of BMI",
     xlab="BMI")
lines(density(bmi), col="red", lwd = 3)
```

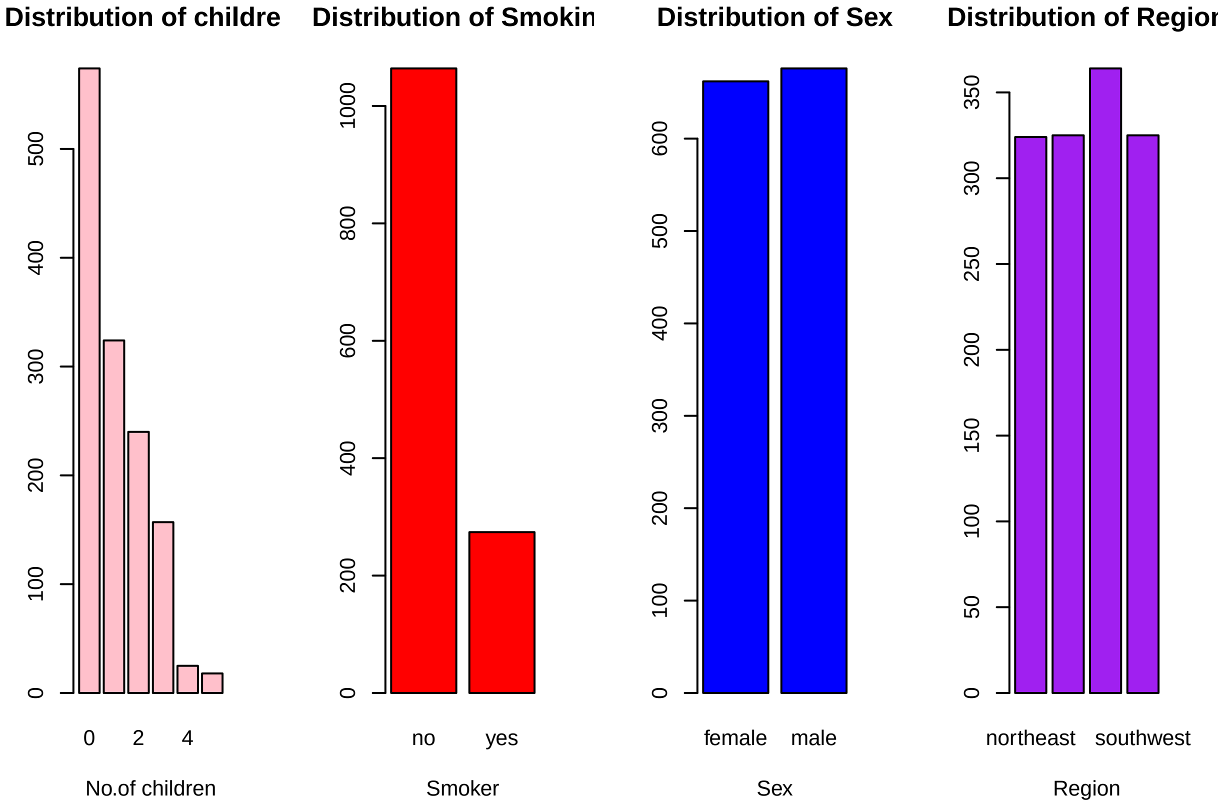


```r
par(mfrow = c(1,1))
```

Insight:

- Distribution of charges is right-skewed indicating the mean larger than median. The highest density of charges is between 1000 to 13000
- Ages range from 18 to 64 years old. observations are more frequent in ages between 18 to 20
- bmi has a normal distribution so we do not have many individuals who are extremely obese or underweight

```r
par(mfrow = c(1,4))

barplot(height = table(children),col = "pink",  main = "Distribution of children",
        xlab="No.of children")

barplot(height = table(smoker), col="red", main = "Distribution of Smoking",
        xlab = "Smoker")

barplot(height = table(sex), col="blue", main = "Distribution of Sex",
        xlab = "Sex")

barplot(height = table(region), col="purple", main = "Distribution of Region",
        xlab='Region')
```



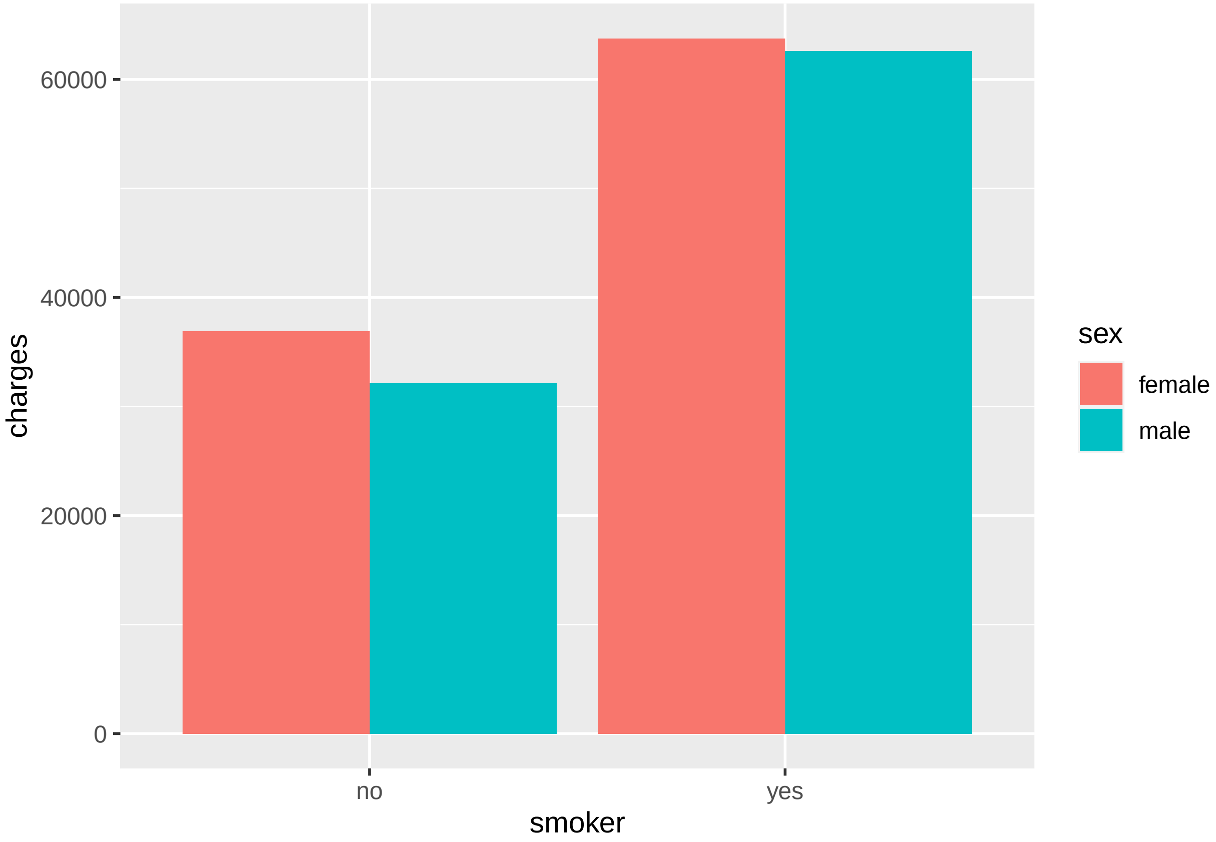
```r
par(mfrow = c(1,1))
```

Insight:

- the number of individuals with no child is much more than the ones with 5 children.
- the number of non-smokers is almost 4 times the number of smokers
- population of females and males are almost equal
- 4 different regions have almost the same number of observations with a little more in SouthEast region

Moreover, it is useful to take a look at the charges for smoker and non-smoker based on sex:

```r
ggplot(insurance, aes(x = smoker, y = charges, fill = sex)) +
  geom_col(position = "dodge")
```

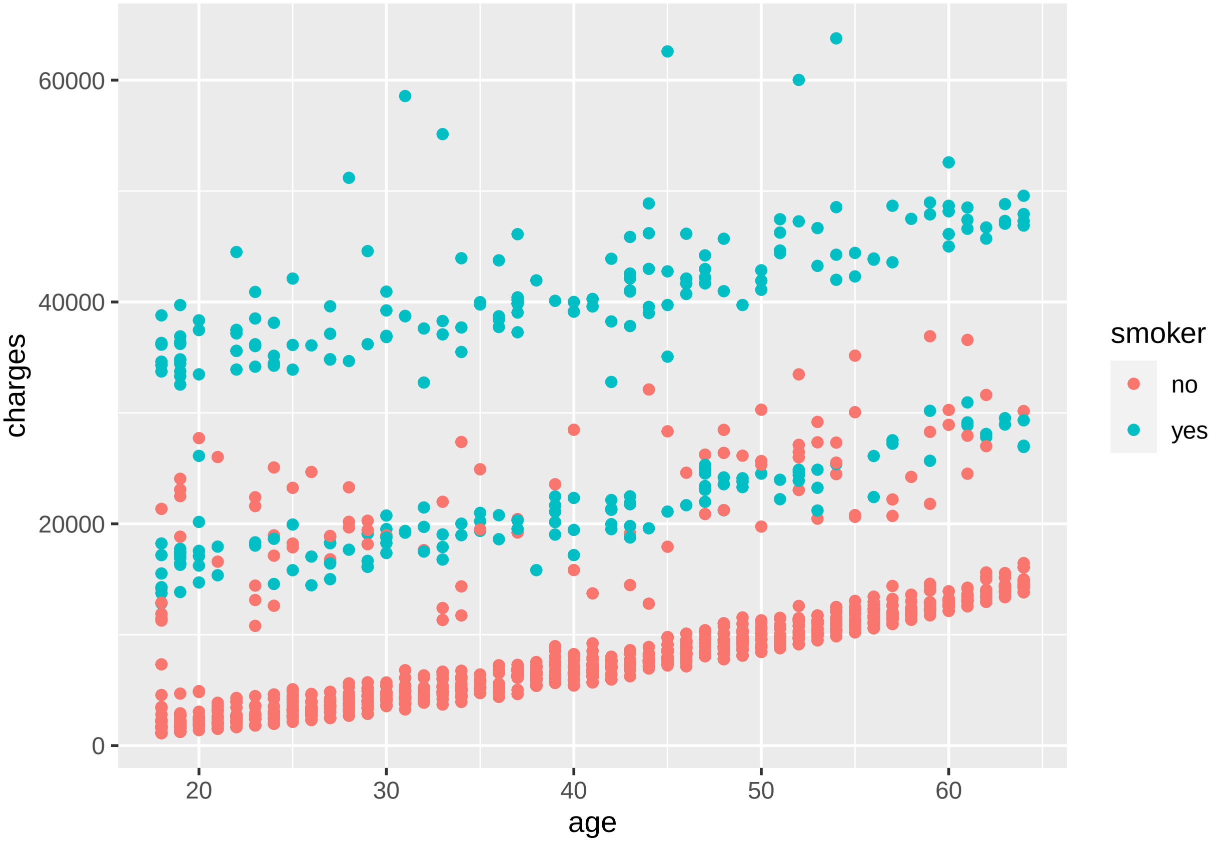


Insight: generally, gender does not affect the charges no matter they are smoker or non-smoker. However, females display a slightly higher charges.

7.3 scatter plots

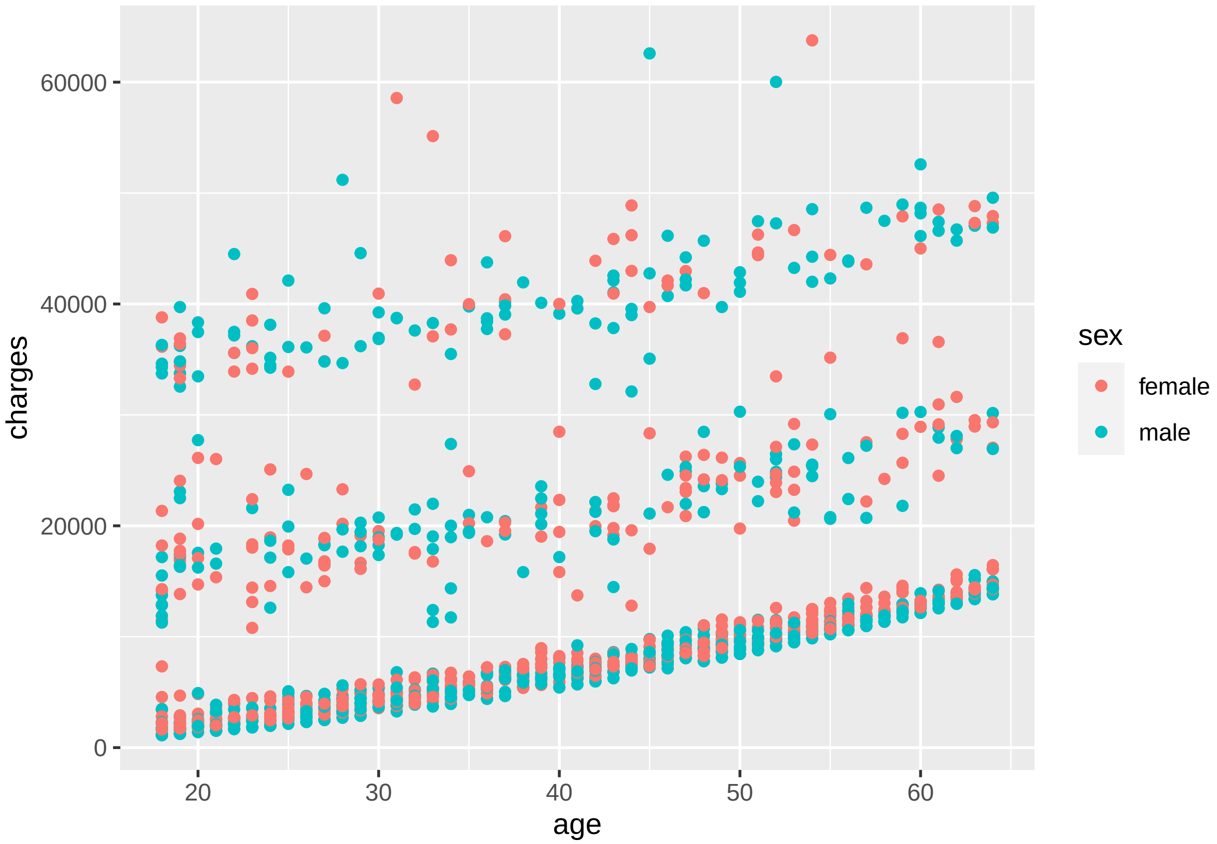
we are exploring the interaction between bmi, age and smoker and their effect on medical costs

```
ggplot(insurance, aes(age, charges, color = smoker)) +
  geom_point()
```



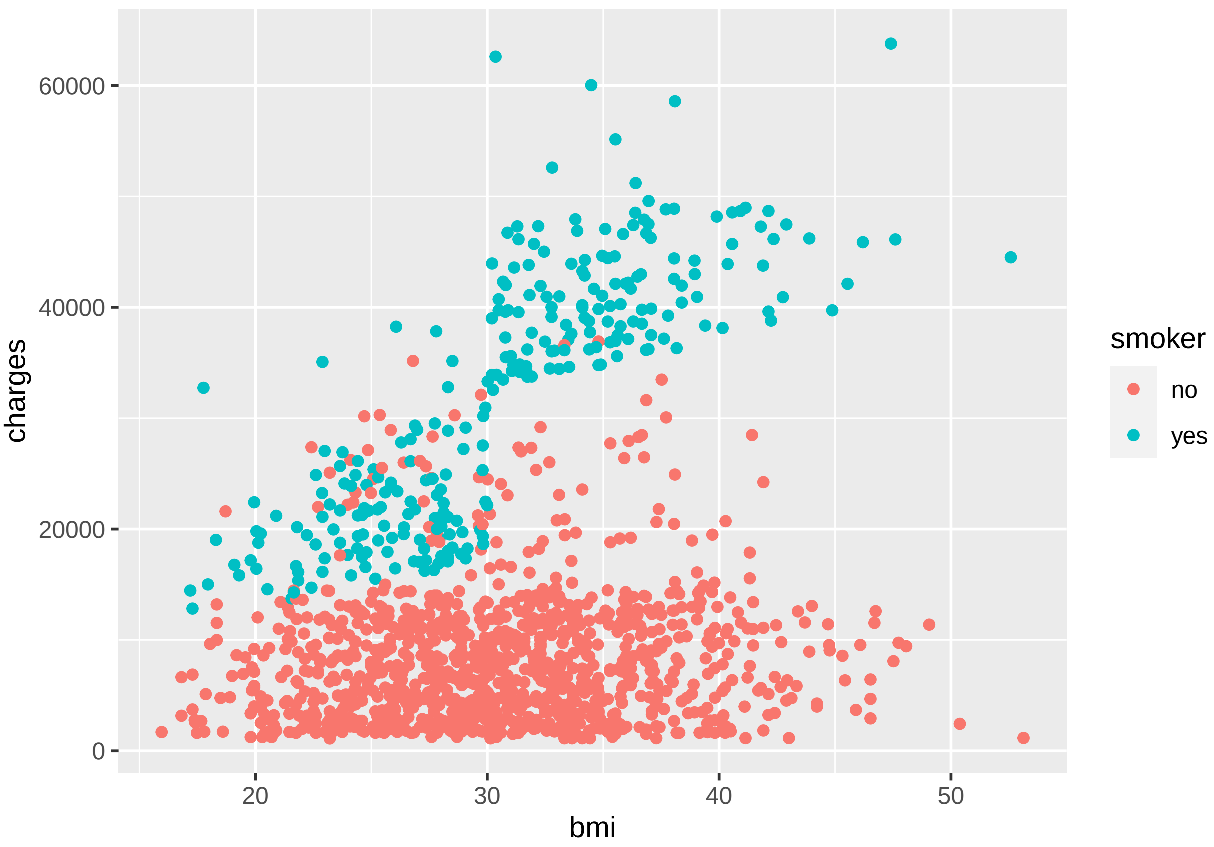
Insight: we can infer that older people spend more on medical expenses; more specifically, smokers' medical costs is massively higher than non-smokers as age increases.

```
ggplot(insurance, aes(age, charges, color = sex)) +
  geom_point()
```



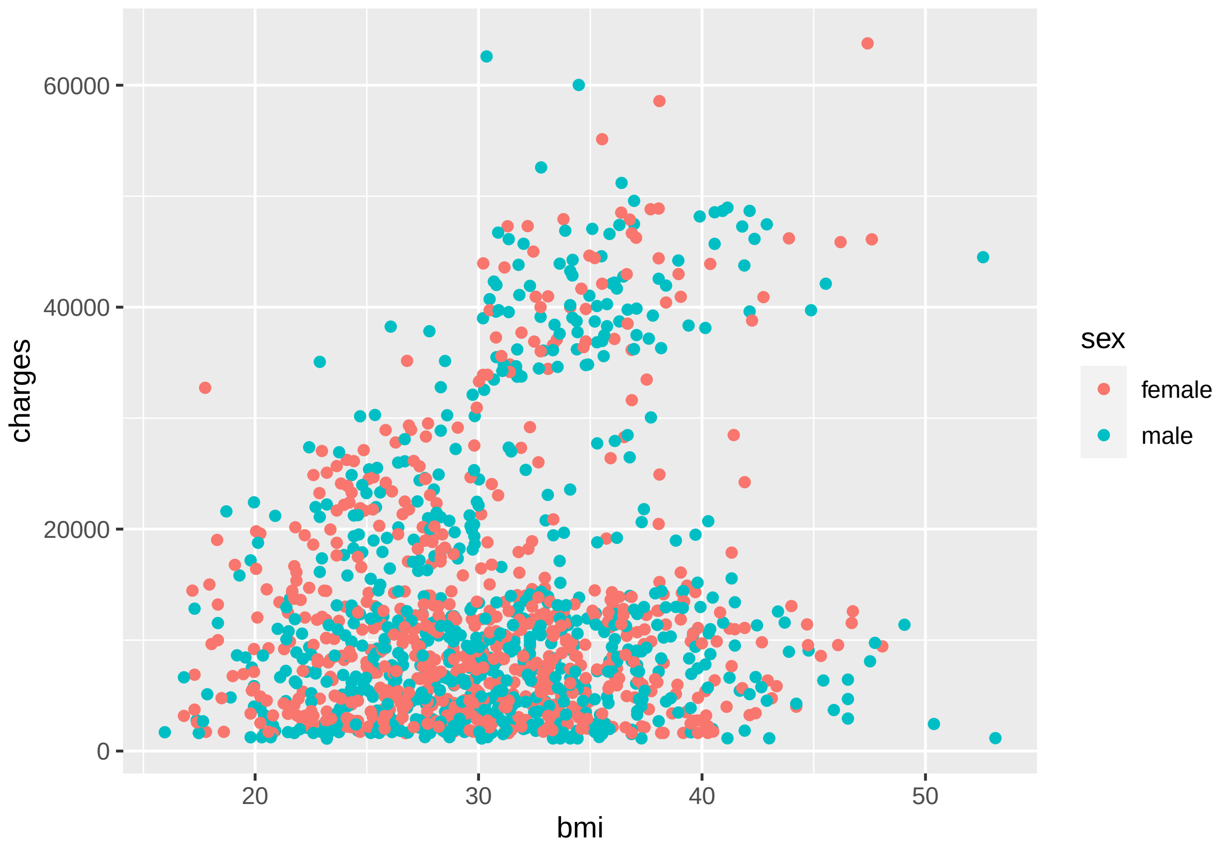
Insight: from this plot we can deduce that charges does not depend on sex because there is no specific pattern.

```
ggplot(insurance, aes(bmi, charges, color = smoker)) +
  geom_point()
```



Insight: medical cost for smokers steeply augments for people with higher bmi specially for those who are obese (bmi>30). In contrast, non-smokers' bmi does not significantly affect medical charges.

```
ggplot(insurance, aes(bmi, charges, color = sex)) +
  geom_point()
```



Insight: there is no evident increase on charges considering bmi in association with sex.

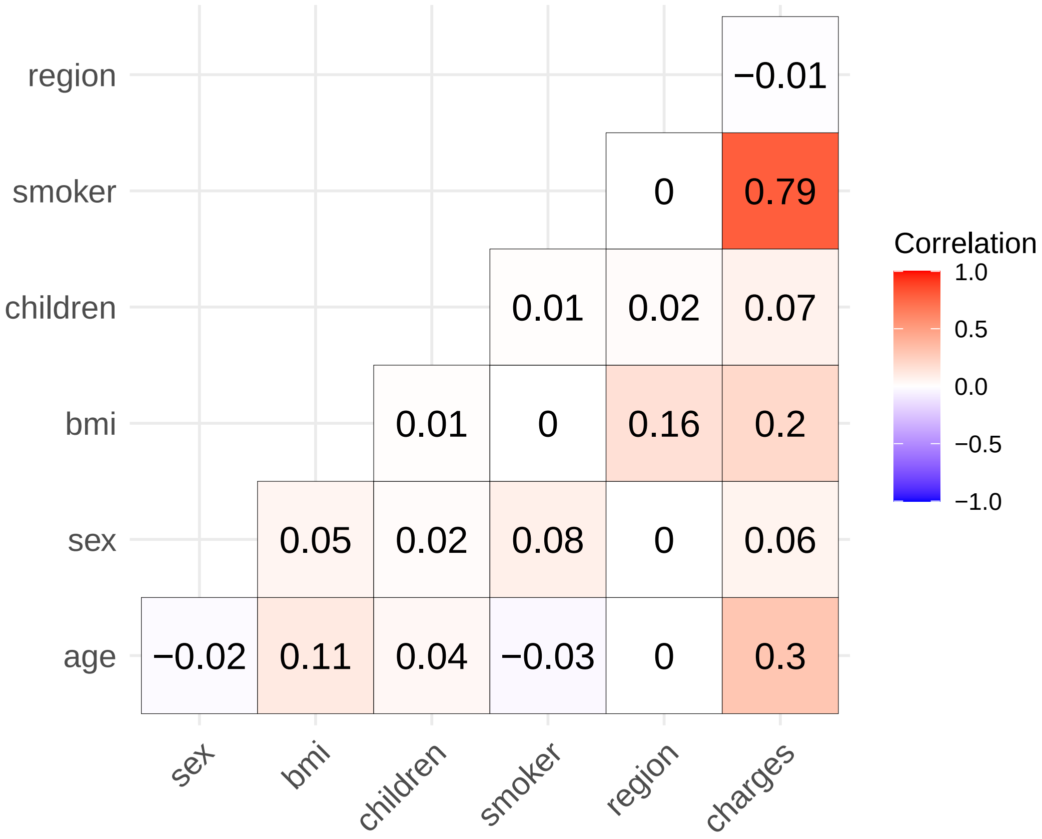
7.4 Correlation between variables

In order to check the correlation between variables we need to convert variables to numeric first.

```
ins <- insurance

ins$sex <- as.factor(ins$sex)
ins$smoker <- as.factor(ins$smoker)
ins$region <- as.factor(ins$region)
ins$children <- as.factor(ins$children)

ins$sex <- as.numeric(ins$sex)
ins$smoker <- as.numeric(ins$smoker)
ins$region <- as.numeric(ins$region)
ins$children <- as.numeric(ins$children)

cor(ins)
```

```
##                   age          sex         bmi    children       smoker
## age       1.000000000 -0.019813556 0.109343610 0.041536210 -0.025587118
## sex      -0.019813556  1.000000000 0.046397060 0.017847542  0.076595850
## bmi       0.109343610  0.046397060 1.000000000 0.012754658  0.003746217
## children  0.041536210  0.017847542 0.012754658 1.000000000  0.007331222
## smoker   -0.025587118  0.076595850 0.003746217 0.007331222  1.000000000
## region    0.001626251  0.004936187 0.157574500 0.016258480 -0.002358321
## charges   0.298308213  0.058044496 0.198400831 0.067389351  0.787234367
##                region      charges
## age       0.001626251  0.298308213
## sex       0.004936187  0.058044496
## bmi       0.157574500  0.198400831
## children  0.016258480  0.067389351
## smoker   -0.002358321  0.787234367
## region    1.000000000 -0.006546563
## charges  -0.006546563  1.000000000
```

```r
corr <- round(cor(ins), 3)

ggcorrplot(corr, type = "lower", lab = TRUE, outline.color = "black",
           lab_size = 5, legend.title = "Correlation")
```



Based on the correlation plot above, the strongest correlation is between smoker and charges with the value of 0.79; age and bmi variables also tend to have a a correlation with charges but the relationship is not so strong. There is no collinearity between independent variables.

7.5 introdicuing the new variable “obese”

Indeed, based on the scatter plot charges~bmi, it is noteworthy to introduce a new column named “obese”(bmi>30) to use in our analysis.

```
# Introducing a new column "obese"
insurance$obese <- ifelse(insurance$bmi >= 30 , "yes" , "no")
```

8 Regression Models

To begin our analysis, we apply a Full model where all covariates are included in the linear model. Then, we try to select the most significant variables through comparing the models.

First of all we split our data into Train and Test: 75% data is considered as the train-set and 25% as the test-set

```
# Spllitting data into Train and Test
set.seed(134)
partition <- floor(0.75*nrow(insurance))
train.numbers <- sample(seq_len(nrow(insurance)), partition, replace = FALSE)
train <- insurance[train.numbers, ]
test <- insurance[-train.numbers, ]
```

```
# MODEL 1 (Full Model)
model.1 <- lm(charges ~. , data = train)
summary(model.1)
```

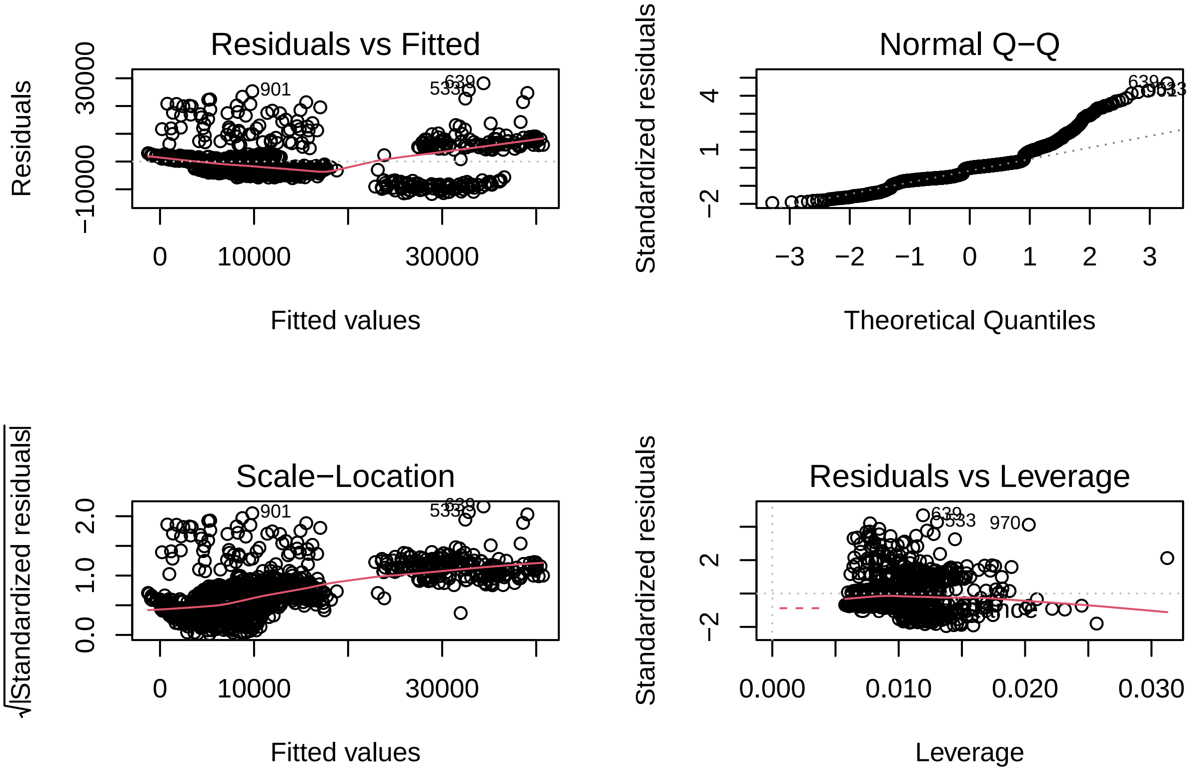
```
##
## Call:
## lm(formula = charges ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11719.8  -3615.0    -38.1   1619.5  28194.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7728.68    1472.67  -5.248 1.88e-07 ***
## age                261.80      13.75  19.042  < 2e-16 ***
## sexmale             31.98     384.66   0.083  0.93376
## bmi                137.37      53.52   2.567  0.01041 *
## children           425.47     157.81   2.696  0.00713 **
## smokeryes        23626.39     474.60  49.781  < 2e-16 ***
## regionnorthwest   -119.92     549.69  -0.218  0.82735
## regionsoutheast   -647.77     556.63  -1.164  0.24481
## regionsouthwest   -966.34     553.09  -1.747  0.08092 .
## obeseyes          3164.91     634.53   4.988 7.21e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 6053 on 992 degrees of freedom
## Multiple R-squared:  0.7536, Adjusted R-squared:  0.7513
## F-statistic:   337 on 9 and 992 DF,  p-value: < 2.2e-16
```

```
BIC(model.1)
```

```
## [1] 20360.81
```

```
par(mfrow = c(2,2))
plot(model.1)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 1
glance(model.1) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma    AIC    BIC   p.value
##           <dbl> <dbl>  <dbl>  <dbl>     <dbl>
## 1         0.751 6053. 20307. 20361. 1.74e-294
```

```
# RMSE and MAE
data.frame(
  R2 = rsquare(model.1, data = insurance),
  RMSE = rmse(model.1, data = insurance),
  MAE = mae(model.1, data = insurance)
)
```

```
##          R2     RMSE      MAE
## 1 0.7553929 5987.339 4283.494
```

Interpretation:

The result of high value of F-statistic : 337 and a very low P-value: < 2.2e-16 shows that the null hypothesis is rejected therefore we can conclude that there is a potential relationship between covariates and the response variable. significant covariates, based on their small P-values, are smoker,age, and obese respectively. Furthermore, two independent variables of sex and region are not significant based on their P-value which is higher than %5. Adjusted R-squared is about 0.75 which shows a pretty decent fit for our model. It means that 75% of variability of data can be explained by this model.

According to smoker's coefficient, we expect the charges of smokers to be,on average, 23626.39 US Dollars more than non-smokers.

Diagnostic plots depict that our model is not predicting data very well because of existence of outliers. the residual vs. fitted values exhibits a U-shaped pattern which provides an evidence of non-linearity in the data; therefor, we try to transform the response variable to tackle this problem and compare the behavior of the model.

8.1 *Transformation of the response variable*

8.1.1 log transformation

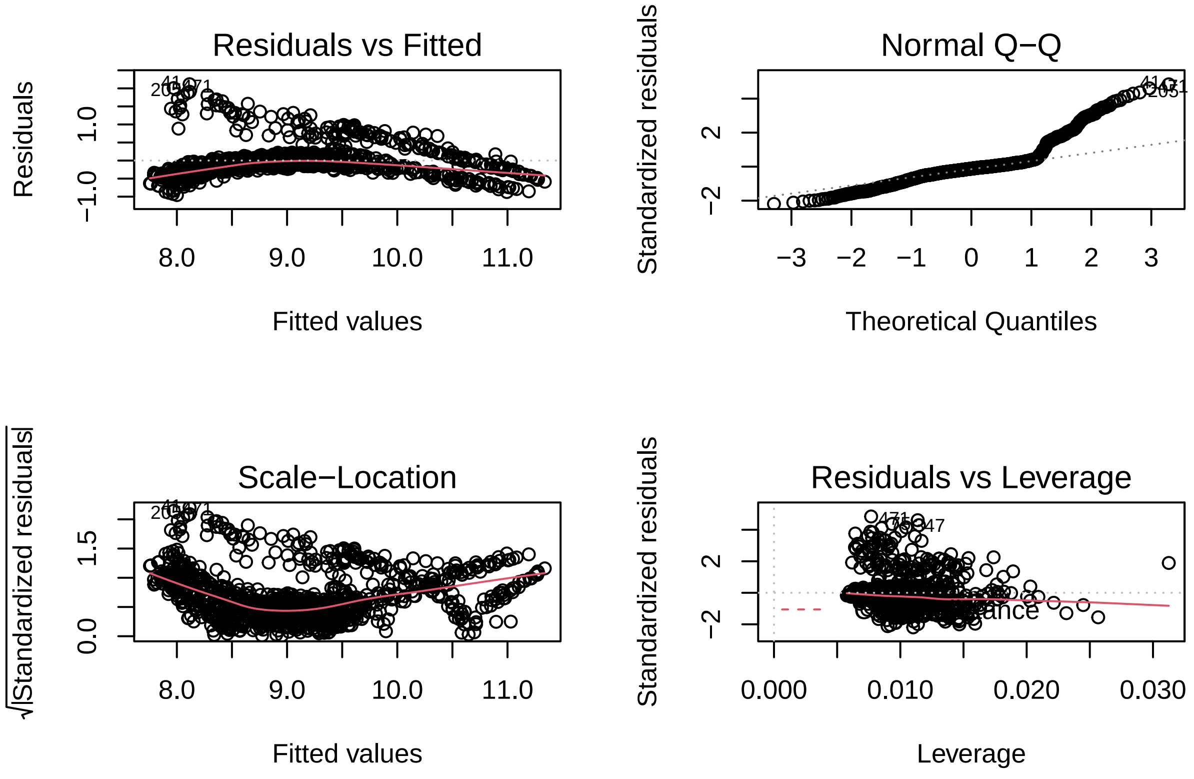
```
# MODEL 2 (Full Model_Log Transformation)
model.2 <- lm(log(charges) ~. , data = train)
summary(model.2)
```

```
##
## Call:
## lm(formula = log(charges) ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95621 -0.20583 -0.05165  0.07547  2.11486
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.123537   0.106688  66.770  < 2e-16 ***
## age              0.034250   0.000996  34.387  < 2e-16 ***
## sexmale         -0.056611   0.027867  -2.031 0.042472 *
## bmi              0.009419   0.003877   2.429 0.015308 *
## children         0.100676   0.011432   8.806  < 2e-16 ***
## smokeryes        1.540279   0.034383  44.798  < 2e-16 ***
## regionnorthwest -0.059152   0.039822  -1.485 0.137759
## regionsoutheast -0.143690   0.040325  -3.563 0.000384 ***
## regionsouthwest -0.148147   0.040069  -3.697 0.000230 ***
## obeseyes          0.073709   0.045969   1.603 0.109152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4385 on 992 degrees of freedom
## Multiple R-squared:  0.7703, Adjusted R-squared:  0.7682
## F-statistic: 369.7 on 9 and 992 DF,  p-value: < 2.2e-16
```

```
BIC(model.2)
```

```
## [1] 1257.321
```

```
par(mfrow = c(2,2))
plot(model.2)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 2
glance(model.2) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.768 0.438 1203. 1257. 1.24e-309
```

```
data.frame(
  R2 = rsquare(model.2, data = insurance),
  RMSE = rmse(model.2, data = insurance),
  MAE = mae(model.2, data = insurance)
)
```

```
##          R2      RMSE       MAE
## 1 0.7680326 0.4423151 0.2825195
```

By log transformation of the response variable the behavior of the model improved. We witnessed increased value of Adjusted R-squared to 0.77 and F-Statistic to 369.7 and BIC from 20360.81 to 1257.321; so this model appears to be a much better fit but there is still some problem in diagnostic plots.

8.1.2 Check for another transformation

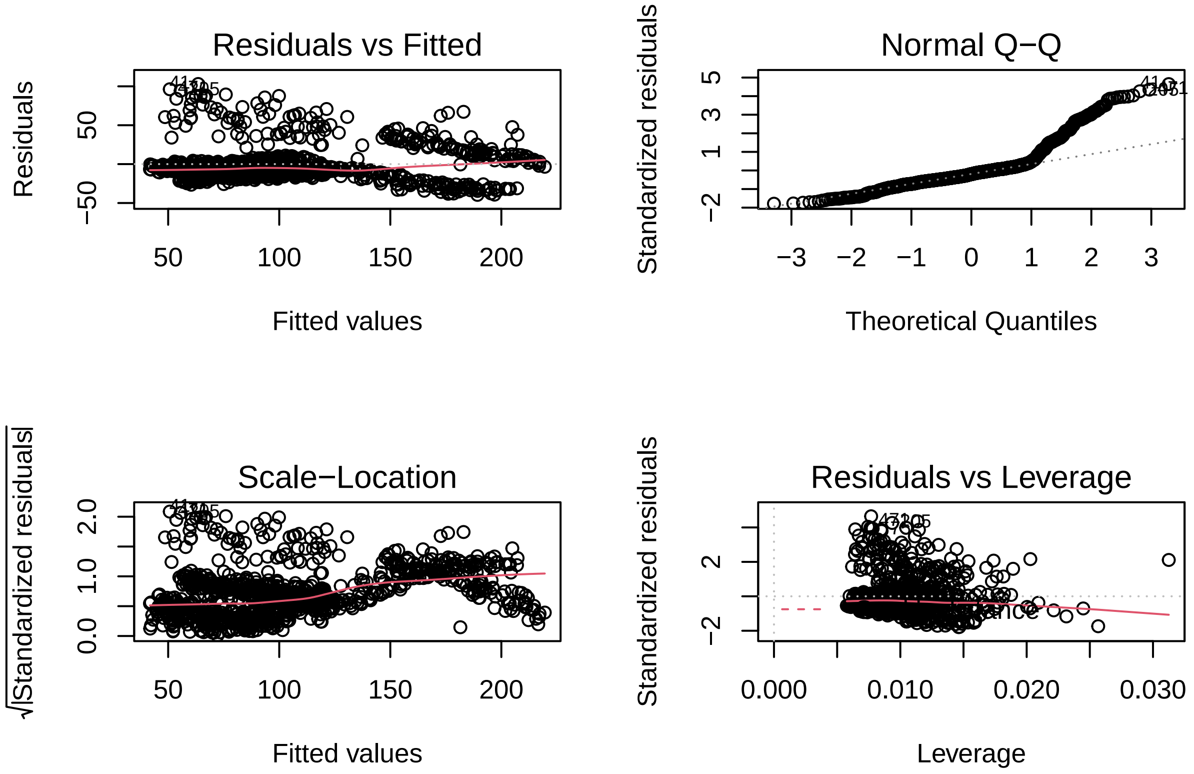
```
# MODEL 3 (Full Model_ Transformation (charges)^(0.5))
model.3 <- lm((charges)^(1/2) ~. , data = train)
summary(model.3)
```

```
##
## Call:
## lm(formula = (charges)^(1/2) ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.530 -12.134  -4.334   3.812 102.987
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.14872    5.41304   2.244 0.025030 *
## age              1.40113    0.05054  27.726  < 2e-16 ***
## sexmale         -1.07087    1.41388  -0.757 0.448990
## bmi              0.48549    0.19672   2.468 0.013760 *
## children         3.17899    0.58005   5.481 5.38e-08 ***
## smokeryes       90.01167    1.74448  51.598  < 2e-16 ***
## regionnorthwest -1.66916    2.02048  -0.826 0.408934
## regionsoutheast -4.63597    2.04599  -2.266 0.023674 *
## regionsouthwest -5.50311    2.03298  -2.707 0.006908 **
## obeseyes         8.73266    2.33233   3.744 0.000191 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.25 on 992 degrees of freedom
## Multiple R-squared:  0.7834, Adjusted R-squared:  0.7814
## F-statistic: 398.7 on 9 and 992 DF,  p-value: < 2.2e-16
```

```
BIC(model.3)
```

```
## [1] 9126.347
```

```
par(mfrow = c(2,2))
plot(model.3)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 3
glance(model.3) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC p.value
##           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1         0.781  22.2 9072. 9126.       0
```

```
data.frame(
  R2 = rsquare(model.3, data = insurance),
  RMSE = rmse(model.3, data = insurance),
  MAE = mae(model.3, data = insurance)
)
```

```
##          R2     RMSE      MAE
## 1 0.7816839 22.30563 14.96255
```

From the plot, it is obvious that transforming the response variable to (charges)^0.5 did not provide a noticeable improvement compared to log transformation, it even resulted in a much more BIC equal to 9126.347. Consequently, we keep the log transformation of the dependent variable.

Besides, we try a polynomial regression model to find out if it is going to be a better fit.

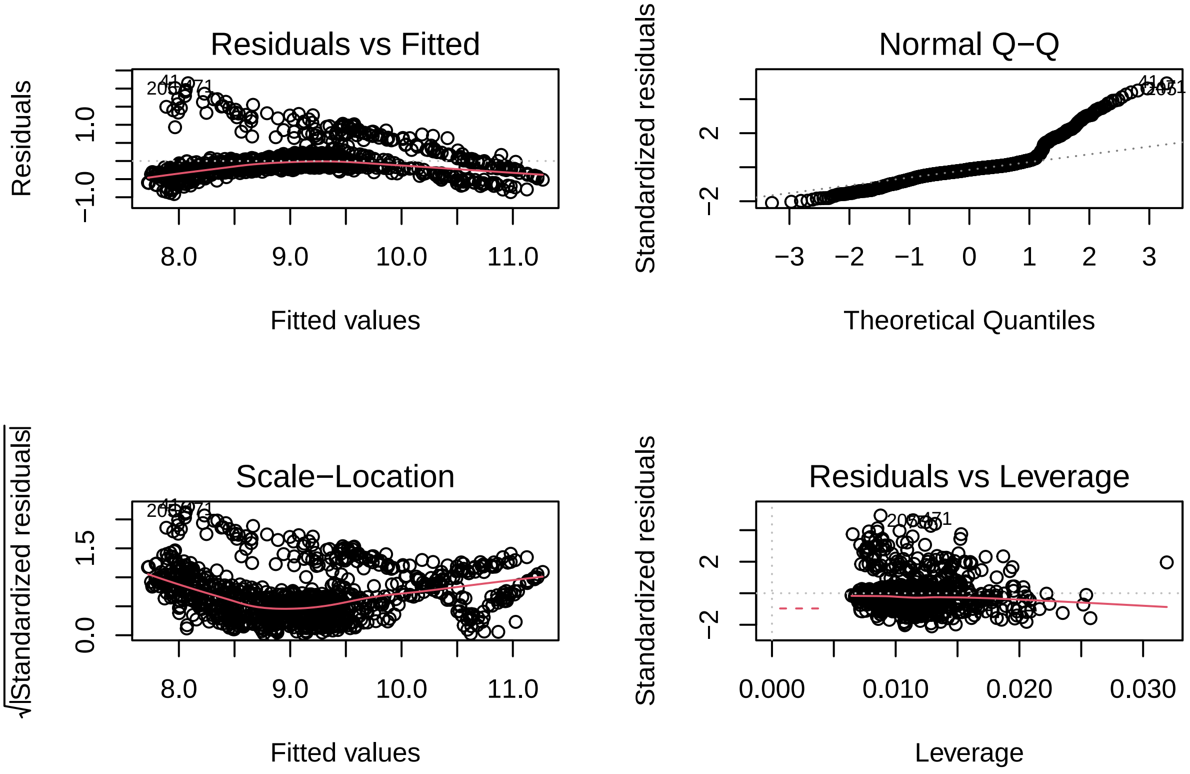
```
# MODEL 4 (Full Model_Polynomial)
model.4 <- lm(log(charges) ~. + poly(age , 2) , data = train)
summary(model.4)
```

```
##
## Call:
## lm(formula = log(charges) ~ . + poly(age, 2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91297 -0.20112 -0.05713  0.06626  2.14759
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.140152   0.106718  66.907  < 2e-16 ***
## age              0.034274   0.000994  34.480  < 2e-16 ***
## sexmale         -0.056449   0.027809  -2.030 0.042635 *
## bmi              0.008939   0.003875   2.307 0.021276 *
## children         0.092383   0.011981   7.711 3.04e-14 ***
## smokeryes        1.541237   0.034314  44.916  < 2e-16 ***
## regionnorthwest -0.057854   0.039744  -1.456 0.145804
## regionsoutheast -0.140772   0.040262  -3.496 0.000492 ***
## regionsouthwest -0.145699   0.040000  -3.642 0.000284 ***
## obeseyes         0.082214   0.046026   1.786 0.074368 .
## poly(age, 2)1          NA         NA      NA       NA
## poly(age, 2)2   -1.045655   0.461383  -2.266 0.023645 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4376 on 991 degrees of freedom
## Multiple R-squared:  0.7715, Adjusted R-squared:  0.7692
## F-statistic: 334.6 on 10 and 991 DF,  p-value: < 2.2e-16
```

```
BIC(model.4)
```

```
## [1] 1259.051
```

```
par(mfrow = c(2,2))
plot(model.4)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 4
glance(model.4) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.769 0.438 1200. 1259. 1.91e-309
```

```
data.frame(
  R2 = rsquare(model.4, data = insurance),
  RMSE = rmse(model.4, data = insurance),
  MAE = mae(model.4, data = insurance)
)
```

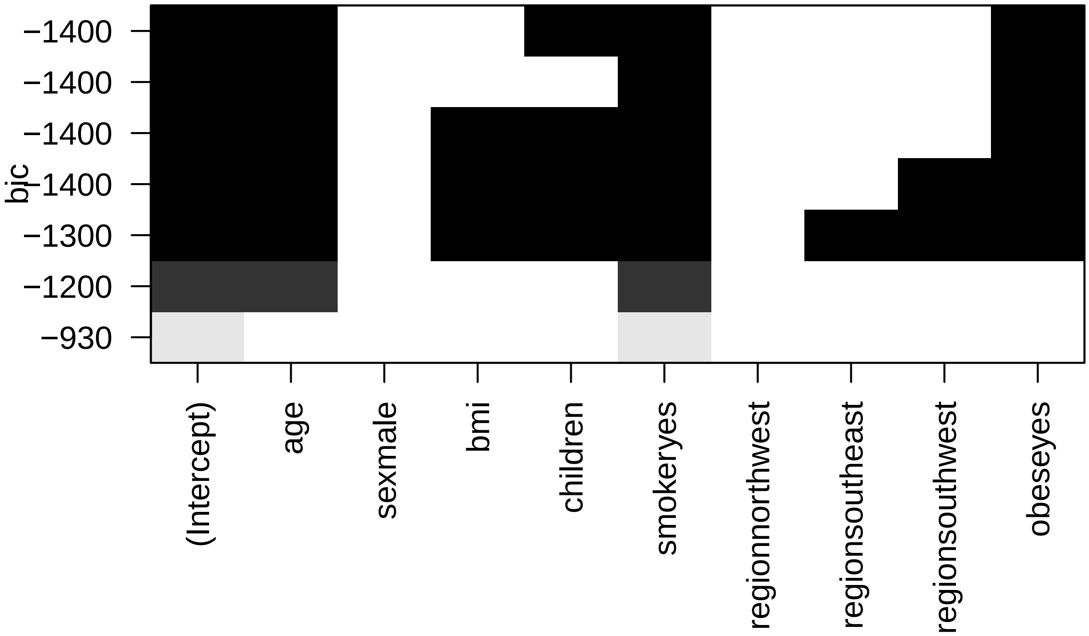
```
## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading

## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading

## Warning in predict.lm(model, data): prediction from a rank-deficient fit may be
## misleading
##         R2     RMSE       MAE
## 1 0.769781 0.440647 0.2821346
```

- **To summarize**: Applying polynomial regression model did not contribute to increasing Adjusted R-squared, F-statistic, or decreasing BIC to a great extent. As a result, we still select MODEL 2 (Full Model_log Transformation) to address the non-linearity problem instead of choosing a polynomial model.

The next step is to select the most significant variables.

8.2 Variable Selection Stepwise Backward Elimination

```
regfit.bwd <- regsubsets((charges) ~.  , data = train, nvmax=7, method="backward")
bwd.summary <- summary(regfit.bwd)
plot(regfit.bwd, scale="bic")
```



```
#Looking at the Outmat Matrix
bwd.summary$outmat
```

```
##          age sexmale bmi children smokeryes regionnorthwest regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 4  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 5  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 7  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
##          regionsouthwest obeseyes
## 1  ( 1 ) " "             " "
## 2  ( 1 ) " "             " "
## 3  ( 1 ) " "             "*"
## 4  ( 1 ) " "             "*"
## 5  ( 1 ) " "             "*"
## 6  ( 1 ) "*"             "*"
## 7  ( 1 ) "*"             "*"
```

```
bwd.summary$bic
```

```
## [1]  -926.7341 -1242.5269 -1358.5605 -1359.2961 -1357.9305 -1353.4822 -1348.0765
```

```
which.min(bwd.summary$bic)
```

```
## [1] 4
```

With this method we ended up with the model in row 4 which selected the following variables as the most significant ones: age, children, smoker, obese

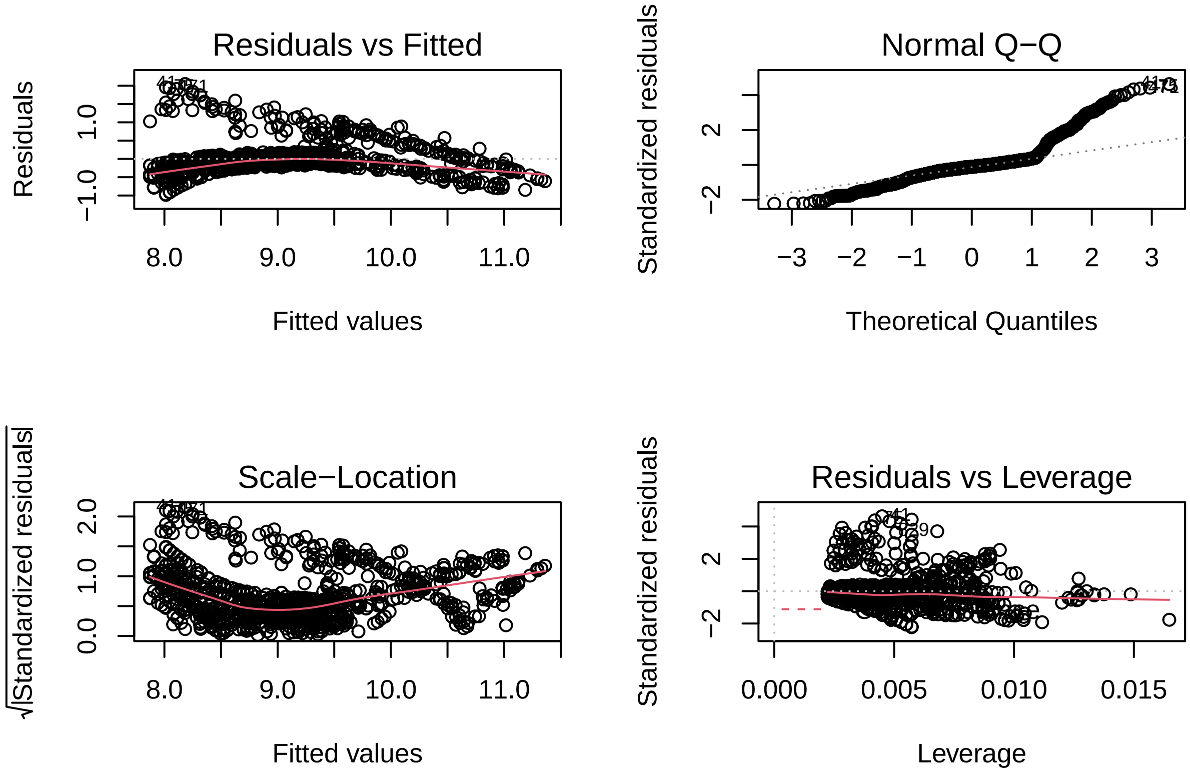
```
# MODEL 5 Reduced Model Based on selected variables by Backward Elimination
model.5 <- lm(log(charges) ~ age + children + smoker + obese , data = train)
summary(model.5)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + children + smoker + obese,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98223 -0.20022 -0.04999  0.08519  2.04687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.253519   0.045004 161.175  < 2e-16 ***
## age         0.034427   0.001002  34.345  < 2e-16 ***
## children    0.100802   0.011524   8.747  < 2e-16 ***
## smokeryes   1.530233   0.034522  44.327  < 2e-16 ***
## obeseyes    0.140335   0.028178   4.980 7.48e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.443 on 997 degrees of freedom
## Multiple R-squared:  0.7644, Adjusted R-squared:  0.7635
## F-statistic: 808.7 on 4 and 997 DF,  p-value: < 2.2e-16
```

```
BIC(model.5)
```

```
## [1] 1248.247
```

```
par(mfrow = c(2,2))
plot(model.5)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 5
glance(model.5) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC   p.value
##           <dbl> <dbl> <dbl> <dbl>     <dbl>
## 1         0.763 0.443 1219. 1248. 4.02e-311
```

```
data.frame(
  R2 = rsquare(model.5, data = insurance),
  RMSE = rmse(model.5, data = insurance),
  MAE = mae(model.5, data = insurance)
)
```

```
##          R2      RMSE       MAE
## 1 0.7623534 0.4476937 0.2845681
```

8.3 Including intraction effect of smoker*obese

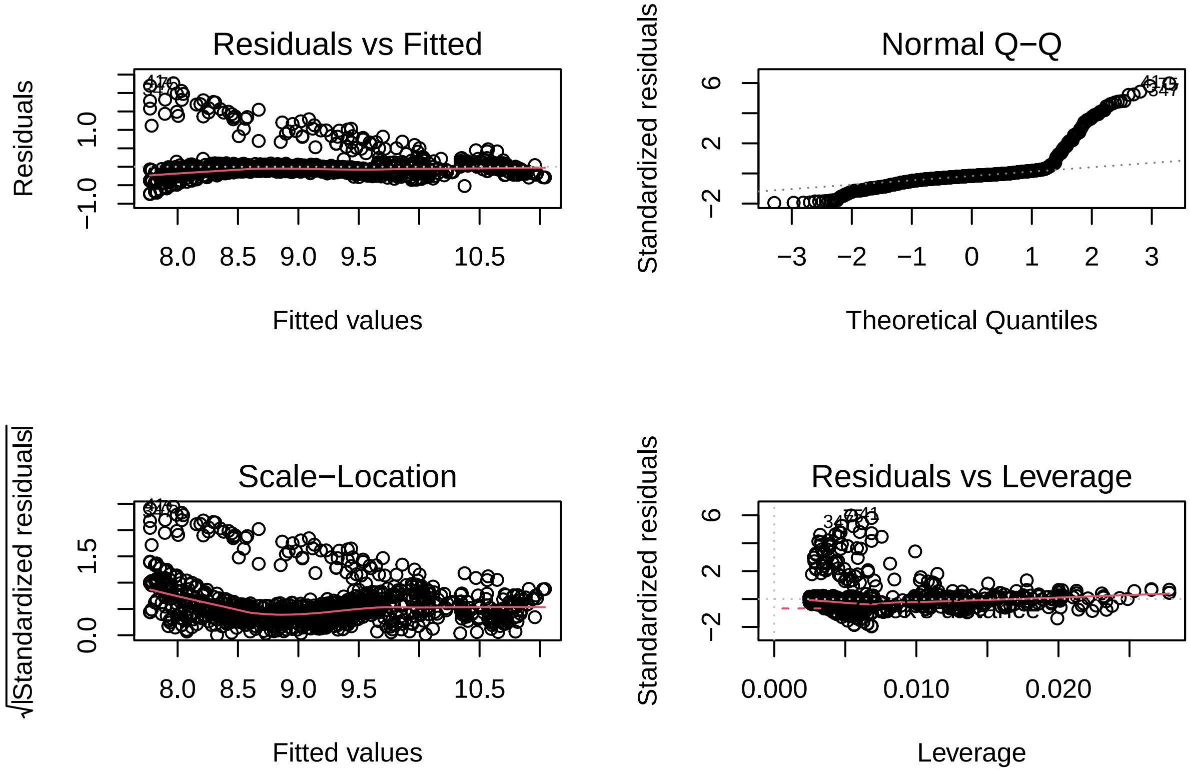
```
# MODEL 6 Reduced Model Based on Backward Elimination including
# interaction of smoker*age and smoker*obese
model.6 <- lm(log(charges) ~ age + children + smoker*age + smoker*obese , data = train)
summary(model.6)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + children + smoker * age + smoker *
##     obese, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74072 -0.13731 -0.06067  0.01030  2.26494
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.0351928  0.0426020 165.138   <2e-16 ***
## age                   0.0416713  0.0009686  43.022   <2e-16 ***
## children              0.1097584  0.0098862  11.102   <2e-16 ***
## smokeryes             2.4658538  0.0892119  27.640   <2e-16 ***
## obeseyes             -0.0132487  0.0271611  -0.488    0.626
## age:smokeryes        -0.0333096  0.0020914 -15.927   <2e-16 ***
## smokeryes:obeseyes    0.7062960  0.0594444  11.882   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3795 on 995 degrees of freedom
## Multiple R-squared:  0.8274, Adjusted R-squared:  0.8264
## F-statistic:   795 on 6 and 995 DF,  p-value: < 2.2e-16
```

```
BIC(model.6)
```

```
## [1] 950.2432
```

```
par(mfrow = c(2,2))
plot(model.6)
```



```
par(mfrow = c(1,1))
```

```
# Metrics for model 6
glance(model.6) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC p.value
##           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1         0.826 0.380  911.  950.       0
```

```
data.frame(
  R2 = rsquare(model.6, data = insurance),
  RMSE = rmse(model.6, data = insurance),
  MAE = mae(model.6, data = insurance)
)
```

```
##          R2      RMSE       MAE
## 1 0.8227052 0.3866913 0.2056036
```

8.4 Concave transformation of the response variable with smoker*obese

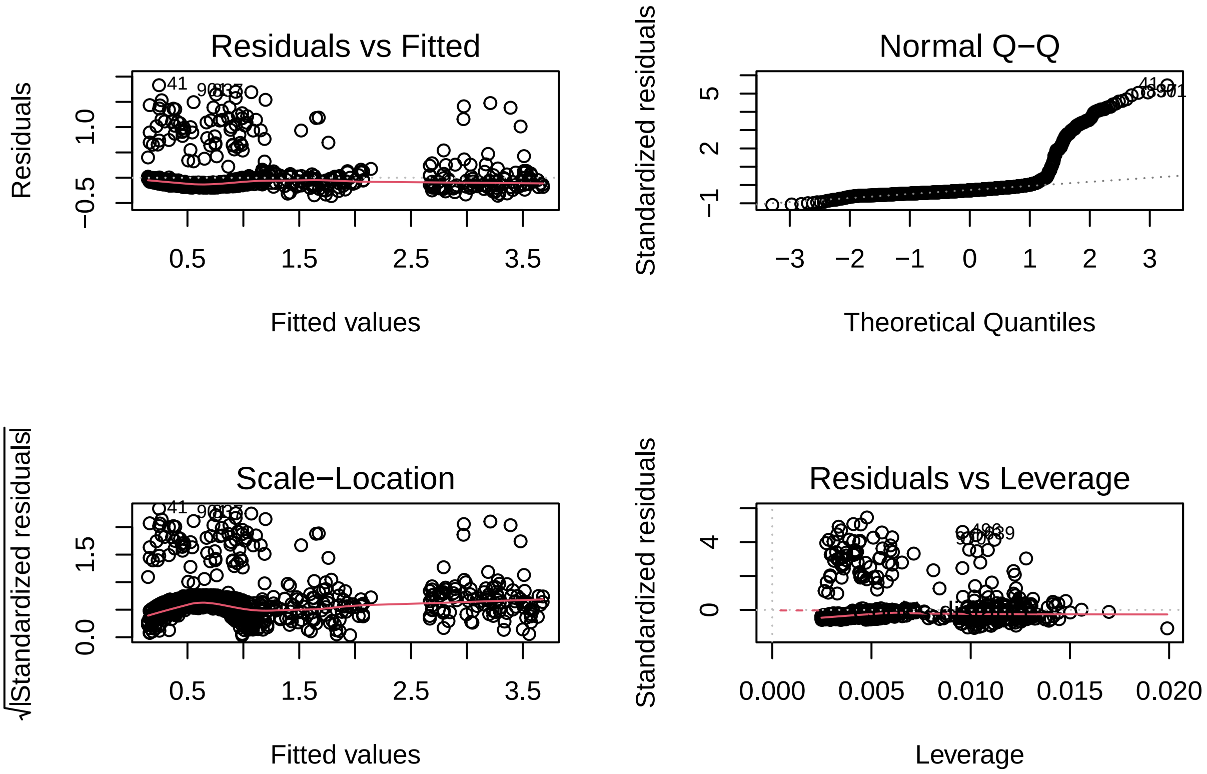
```
# MODEL 7 Concave transformation of the response variable with smoker*obese
model.7 <- lm(charges/mean(charges) ~ age + children + smoker*obese , data = train)
summary(model.7)
```

```
##
## Call:
## lm(formula = charges/mean(charges) ~ age + children + smoker *
##     obese, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36309 -0.14079 -0.09875 -0.04068  1.82897
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.2133475  0.0345143  -6.181 9.25e-10 ***
## age                 0.0200493  0.0007613  26.337  < 2e-16 ***
## children            0.0432727  0.0087591   4.940 9.14e-07 ***
## smokeryes           1.0020317  0.0374485  26.758  < 2e-16 ***
## obeseyes            0.0152741  0.0240367   0.635    0.525
## smokeryes:obeseyes  1.5028468  0.0524683  28.643  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3364 on 996 degrees of freedom
## Multiple R-squared:  0.8636, Adjusted R-squared:  0.8629
## F-statistic:  1261 on 5 and 996 DF,  p-value: < 2.2e-16
```

```
BIC(model.7)
```

```
## [1] 702.6071
```

```
par(mfrow = c(2,2))
plot(model.7)
```



```
# Metrics for model 7
glance(model.7) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```
## # A tibble: 1 x 5
##   adj.r.squared sigma   AIC   BIC p.value
##           <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1         0.863 0.336  668.  703.       0
```

```
data.frame(
  R2 = rsquare(model.7, data = insurance),
  RMSE = rmse(model.7, data = insurance),
  MAE = mae(model.7, data = insurance)
)
```

```
##          R2      RMSE       MAE
## 1 0.8609659 0.3399417 0.1863047
```

9 Remarks

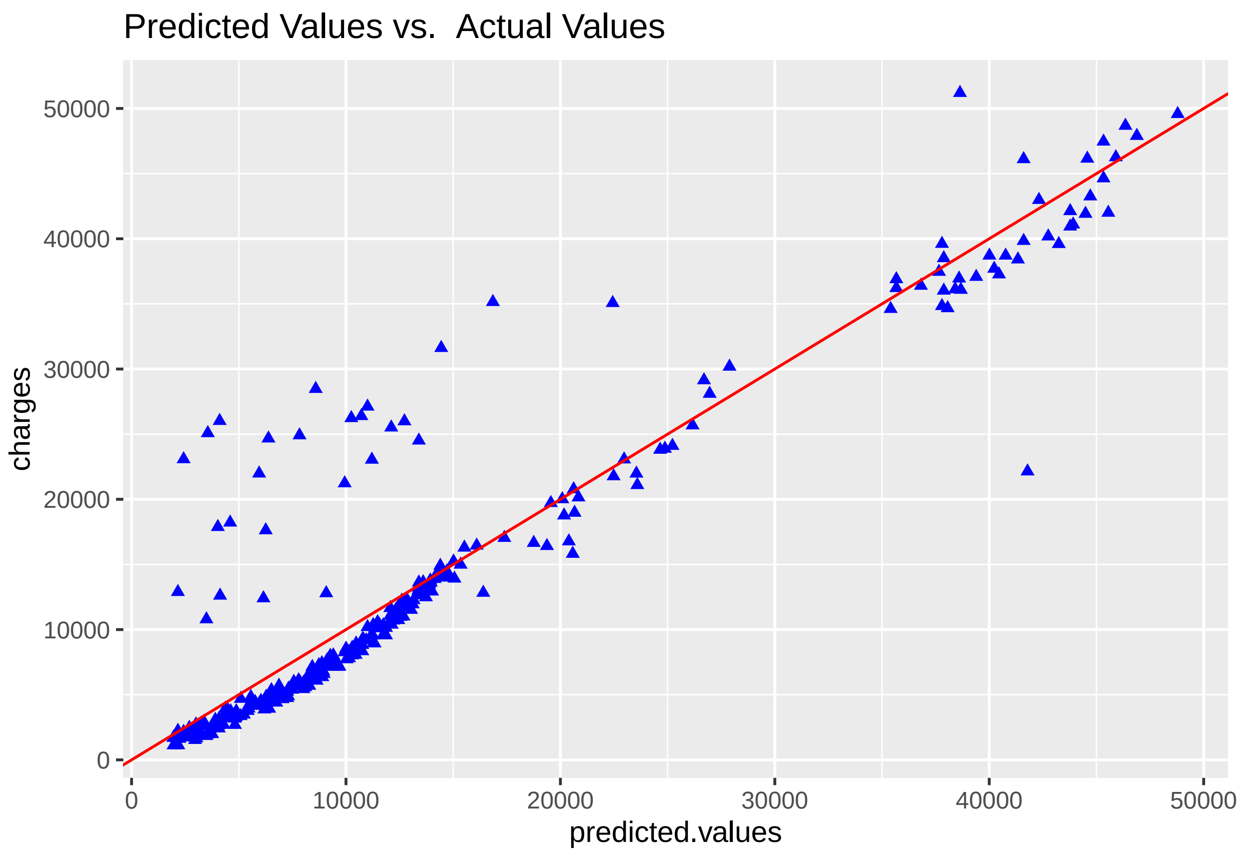
So, the model we selected as the optimal model is MODEL 7 with Adjusted R-squared: 0.8264 which has been significantly improved compared to other 5 models.it implies that 0.8264 variation of charges can be explained by our select independent variables in MODEL 6.

- The Residual Vs. Fitted Values plot shows a straight line indicating that the relationship is linear
- In the normal Q-Q plot residuals are not normally distributed
- The Scale-Location plot shows the assumption of equal variances is satisfied (homoscedasticity)
- The Residuals vs Leverage plot does not display any influential points within the cook's distance

We are going to use this model for our prediction of charges on test set as follows;

10 Predicted Values vs Actual Values

```
test$predicted <- predict(model.7, newdata = test)

predicted.values <- (test$predicted)* mean(charges)

test %>%
  ggplot(aes(x = predicted.values , y = charges)) +
  geom_point(shape = "triangle", color = "blue") +
  geom_abline(color = "red") +
  ggtitle("Predicted Values vs.  Actual Values")
```



This plot shows that our model (MODEL 7) is a good fit for our test set and provides a relatively robust prediction even though it may struggle due to presence of outliers.

The best model for predicting the medical costs is with interaction terms. The best BIC and adjusted R-squared achieved are 702.60 and 0.86 respectively.

In our analysis, we found out that the most significant predictors of charges were the variables related to the age, children, and interaction of obese and smoker.

our model (MODEL 7) is a good fit for our test set and provides a relatively robust prediction even though it may struggle due to presence of outliers.

As we have seen, some observations had relatively high insurance charges, although they were young and non-smokers; they might have had an accident, surgery or something that affected their charges in such a manner.

As a consequence, insurance companies need to collect more accurate data regarding such observations and even include more variables such as their medical history, the hospitals they visited.

11 Bibliography

[1] MIRI CHOI. (2018). Medical Cost Personal Datasets. Retrieved [25/06/22] from https://www.kaggle.com/datasets/mirichoi0218/insurance