**In this project a dataset about costs for health insurance was used to practice a few machine learning models in R.**

```
library(fastDummies)
library(ggplot2)
library(plotly)
library(hrbrthemes)
library(extrafont)
library(corrgram)
library(caret)
library(caTools)
library(rpart)
library(forecast)
library(ISLR)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
library(dplyr)
library(ConfusionTableR)
library(tidyr)
library(mlbench)
```

```
#Reading the data

df = read.csv("E:/Usuarios/Documentos/R/Medical Insurance/insurance.csv", na.strings="", stringsAsFacto
head(df)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

```
summary(df)
```

```
##       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
```

```
#checking missing data
df[!complete.cases(df),]
```

```
## [1] age      sex      bmi      children smoker   region   charges
## <0 linhas> (ou row.names de comprimento 0)
```

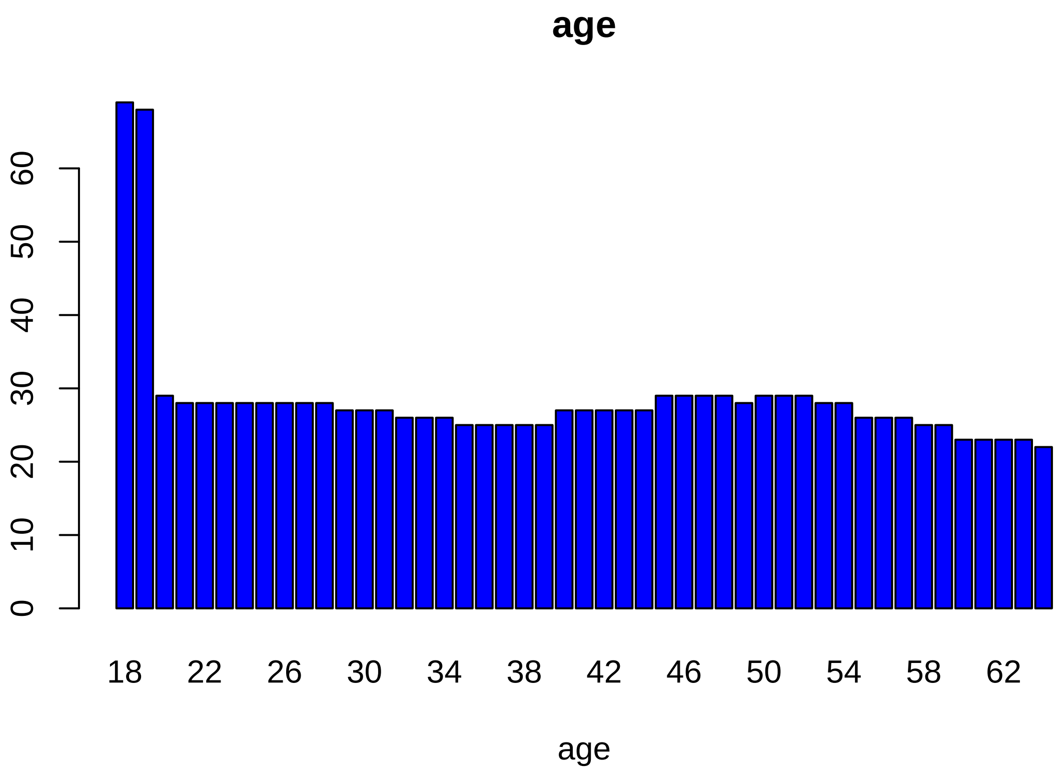
```
#Checking duplicated rows
#df[!duplicated(df),]
sum(duplicated(df))
```

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

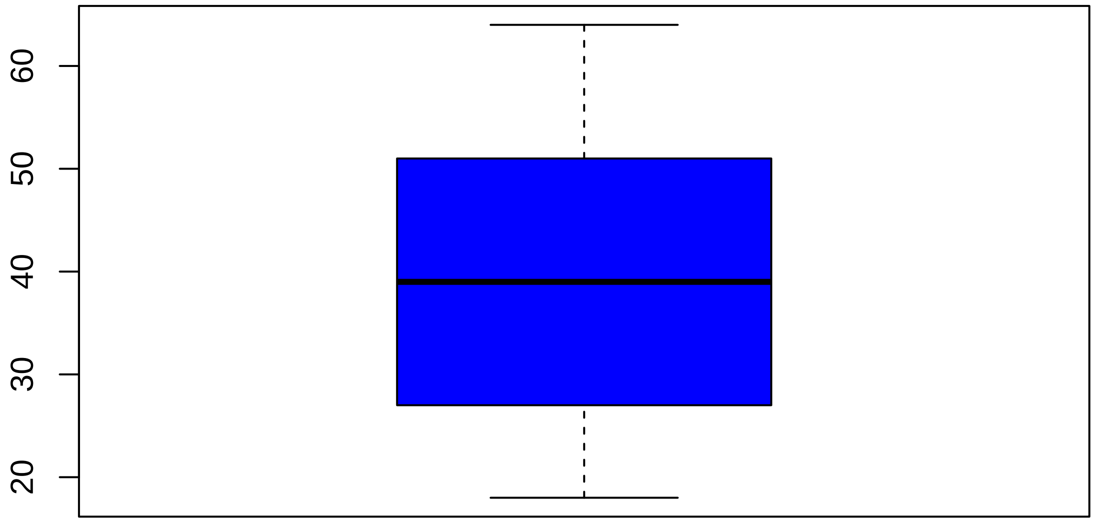
Exploring the variables

1.Age

```
Counts = table(df$age)
barplot(Counts, main="age", xlab="age",  col = c("blue"))
```



```
boxplot(df$age, col = c("blue"))
```

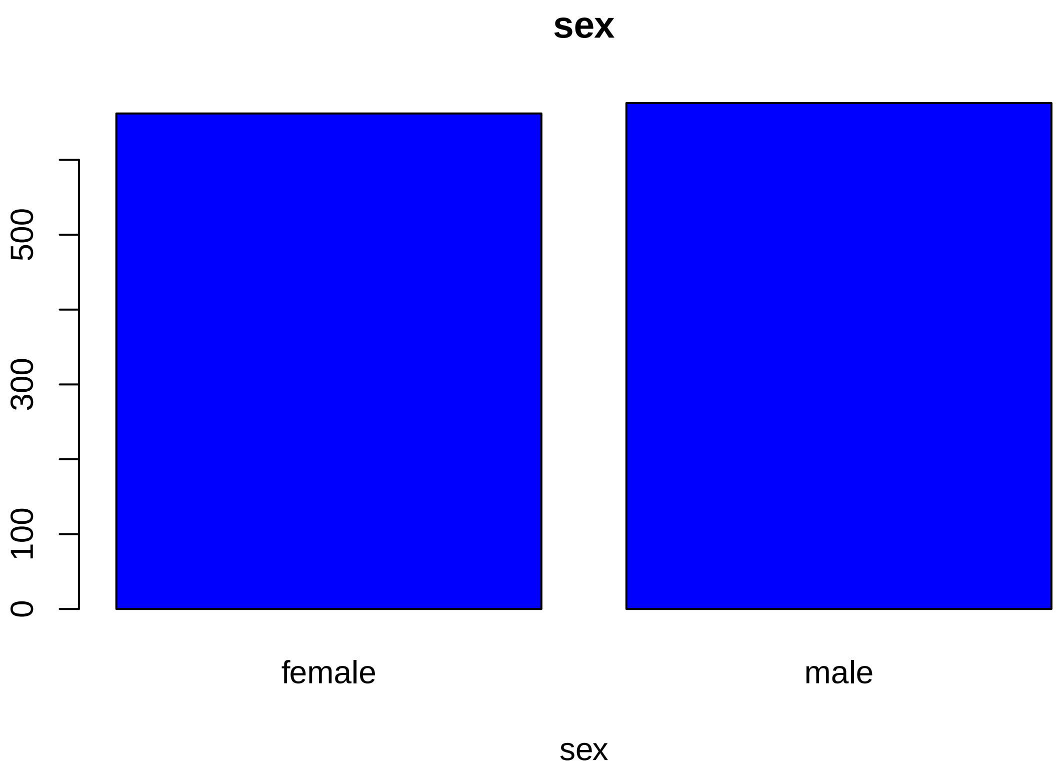


```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   27.00   39.00   39.21   51.00   64.00
```

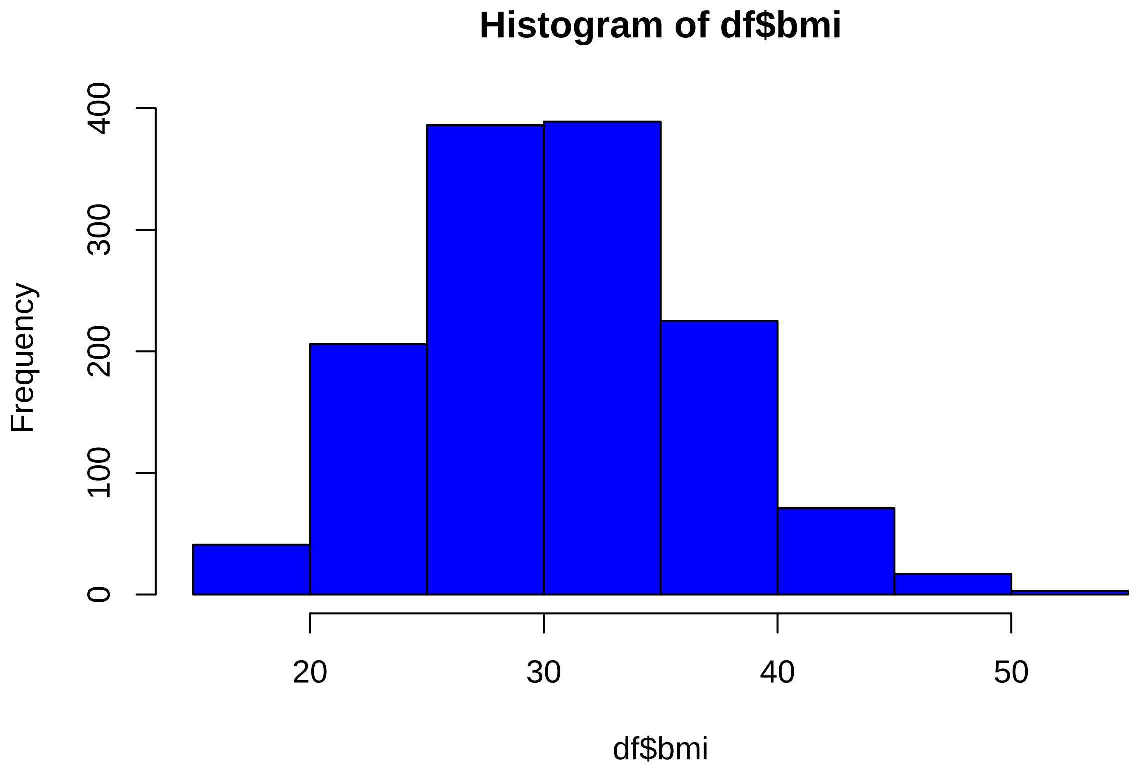
2.Sex

```
Counts = table(df$sex)
barplot(Counts, main="sex", xlab="sex", col = c("blue"))
```

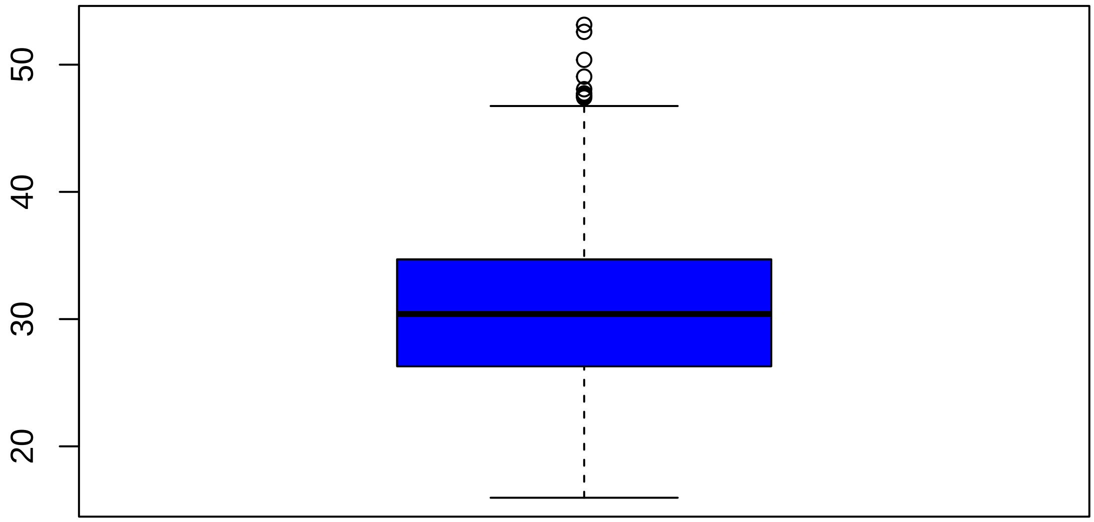


3.BMI

```
hist(df$bmi, col = c("blue"))
```



```
boxplot(df$bmi,  col = c("blue"))
```

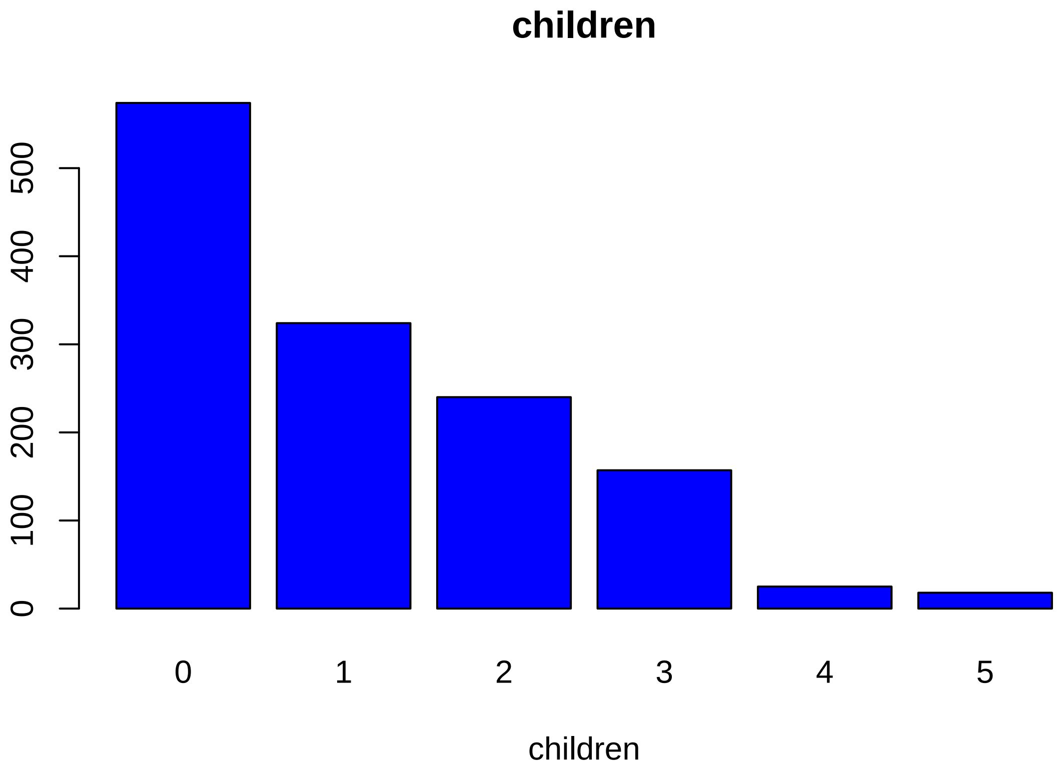


```
summary(df$bmi)
```

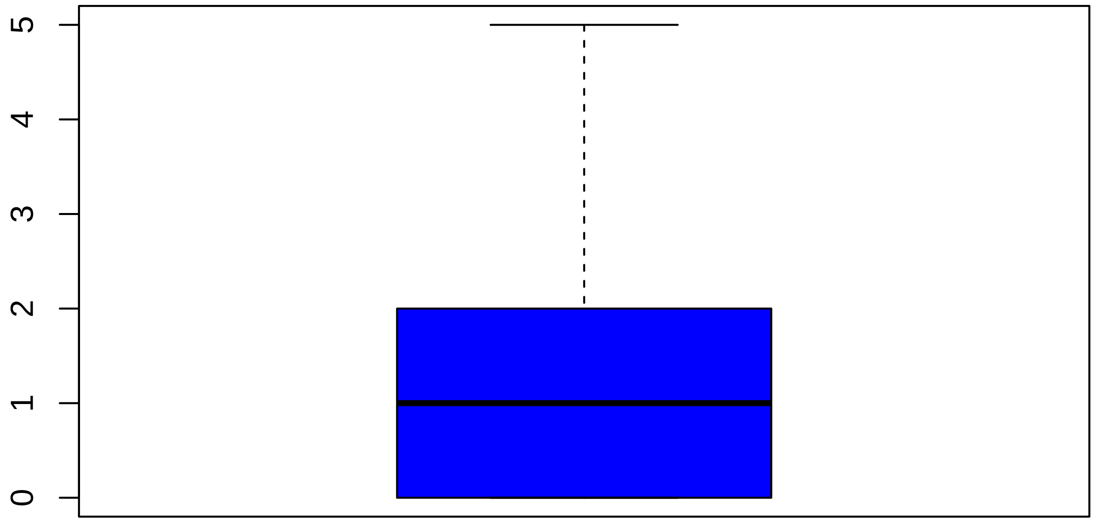
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.96   26.30   30.40   30.66   34.69   53.13
```

4.Children

```
Counts = table(df$children)
barplot(Counts, main="children", xlab="children",col = c("blue"))
```



```
boxplot(df$children,  col = c("blue"))
```

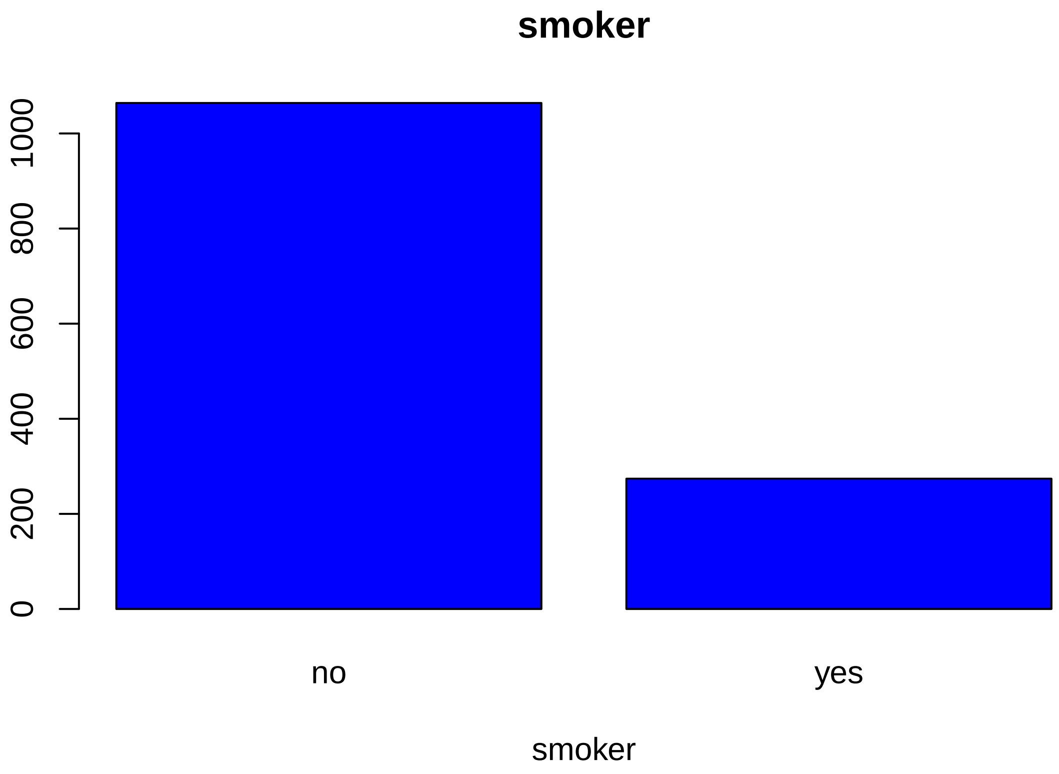


```
summary(df$children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   1.000   1.095   2.000   5.000
```

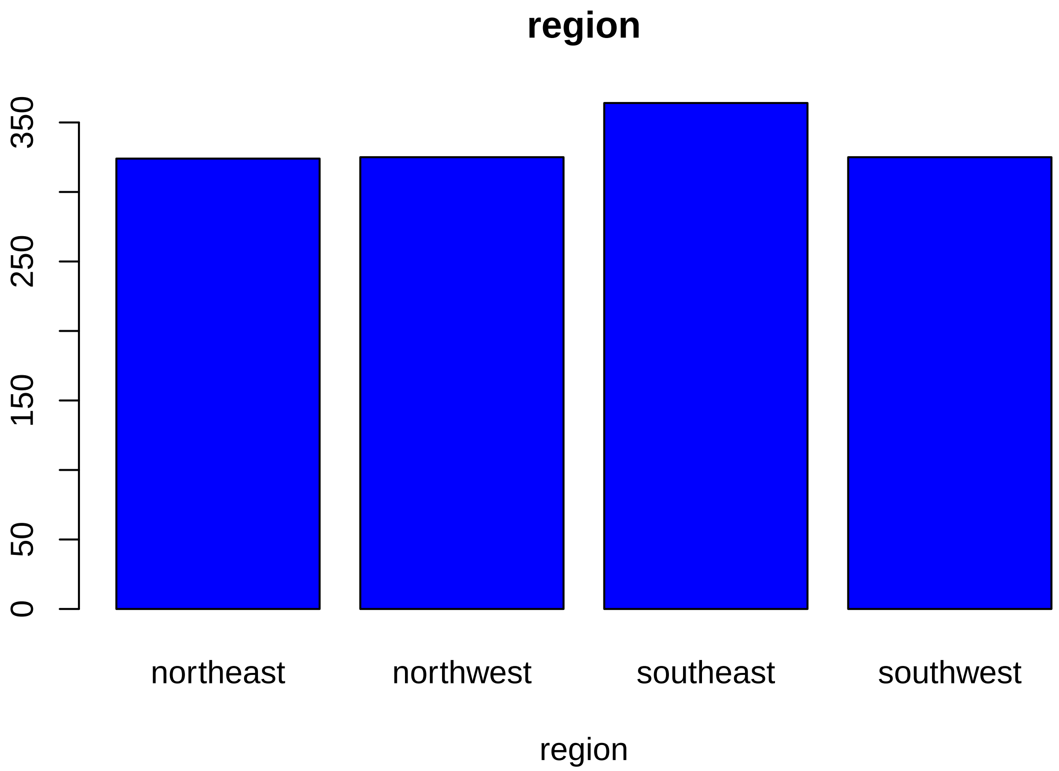
5.Smoker

```
Counts = table(df$smoker)
barplot(Counts, main="smoker", xlab="smoker",col = c("blue"))
```



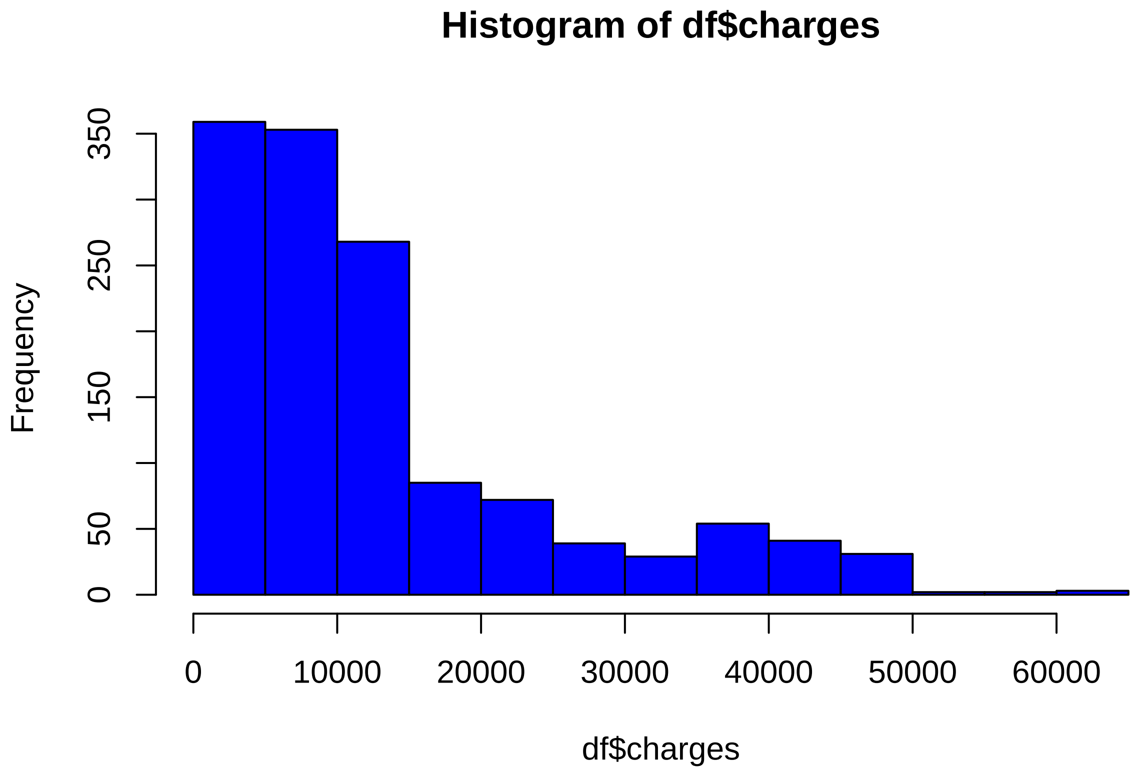
6.Region

```
Counts = table(df$region)
barplot(Counts, main="region", xlab="region",col = c("blue"))
```

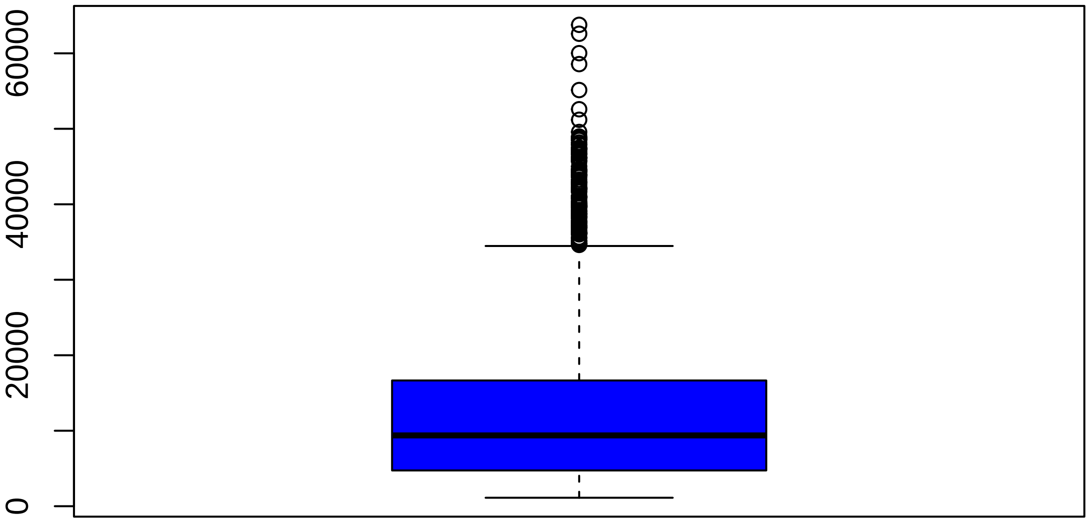


Variable Response - Charges

```
hist(df$charges,col = c("blue"))
```



```
boxplot(df$charges,  col = c("blue"))
```

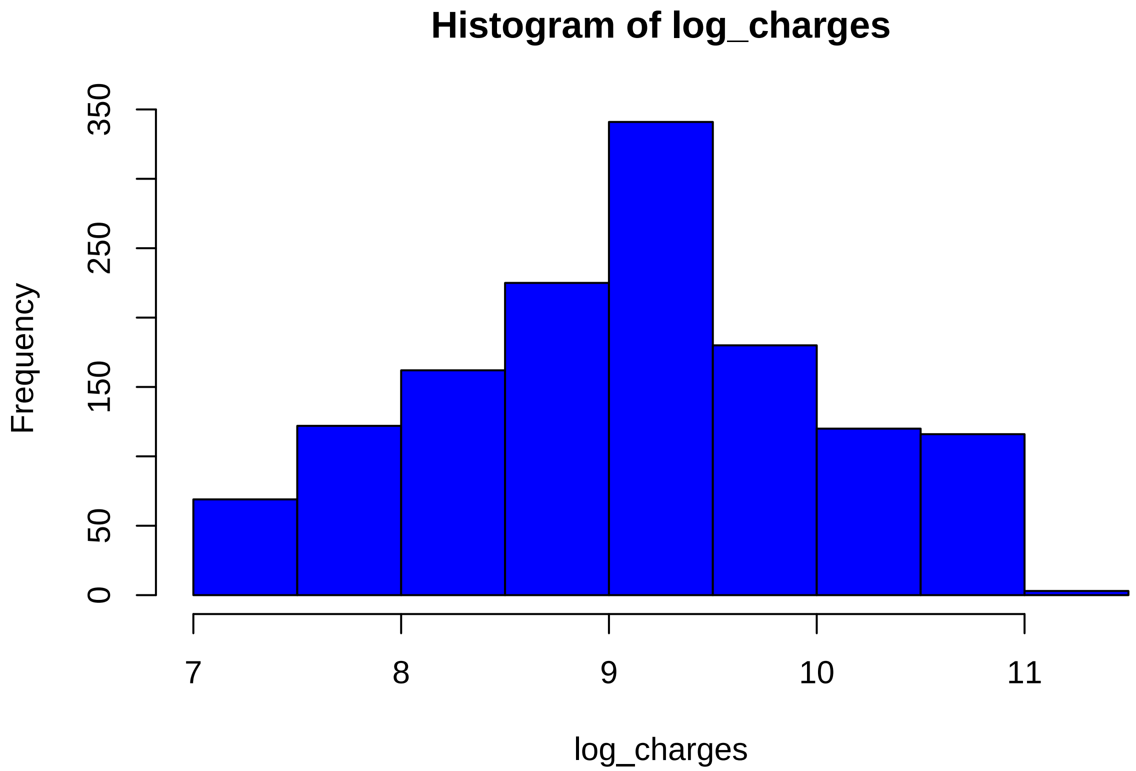


```
summary(df$charges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1122    4740    9382   13270   16640   63770
```

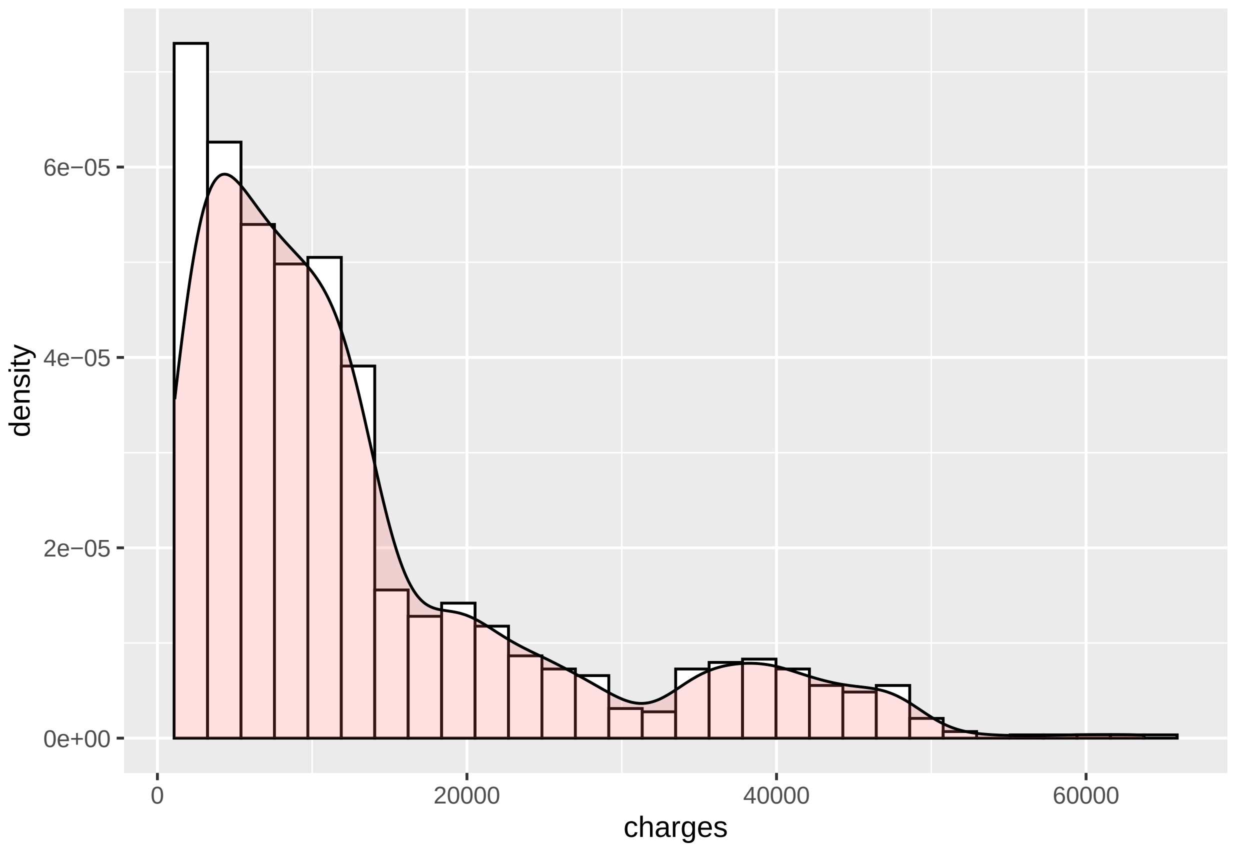
We can see that charges don't have a normal distribution, making a log transformation give us a distribution that tends to normal. For this project that transformation will suffice.

```
log_charges = log(df$charges)
hist(log_charges,col = c("blue"))
```



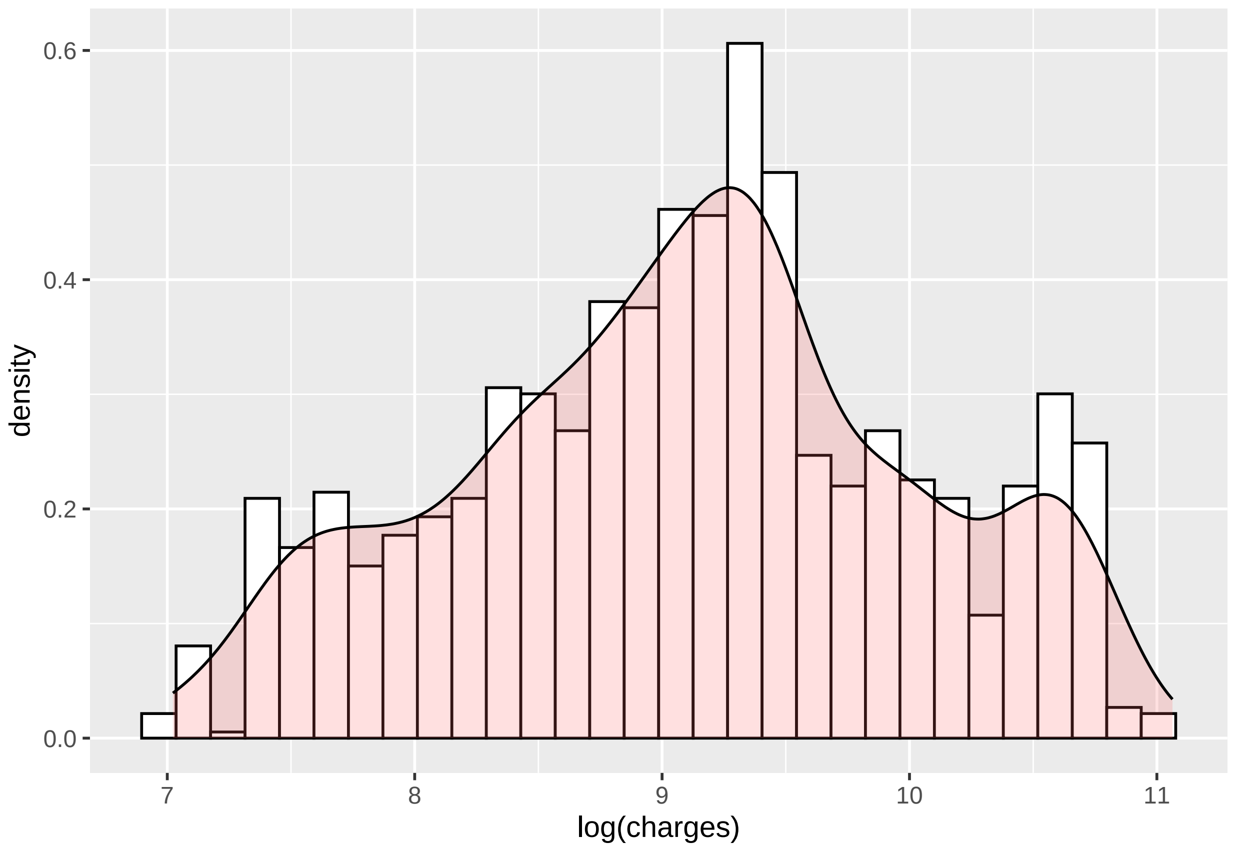
```
# Histogram overlaid with kernel density curve
ggplot(df, aes(x=charges)) +
    geom_histogram(aes(y=after_stat(density)),
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



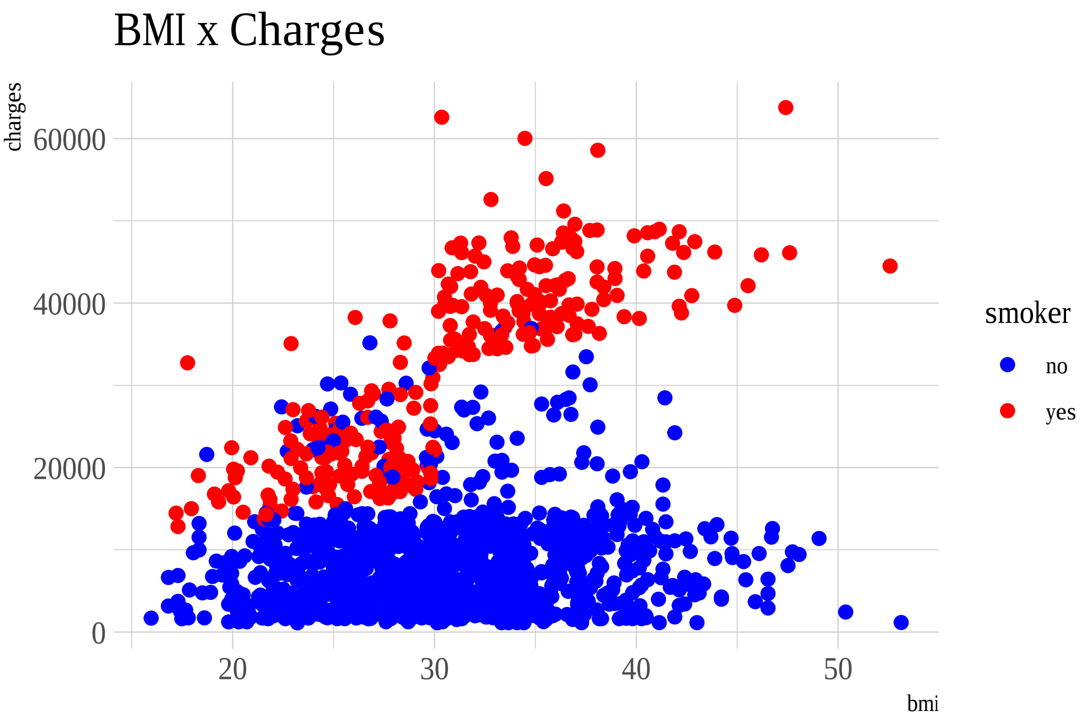
```
ggplot(df, aes(x=log(charges))) +
    geom_histogram(aes(y=after_stat(density)),
                   colour="black", fill="white") +
    geom_density(alpha=.2, fill="#FF6666")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

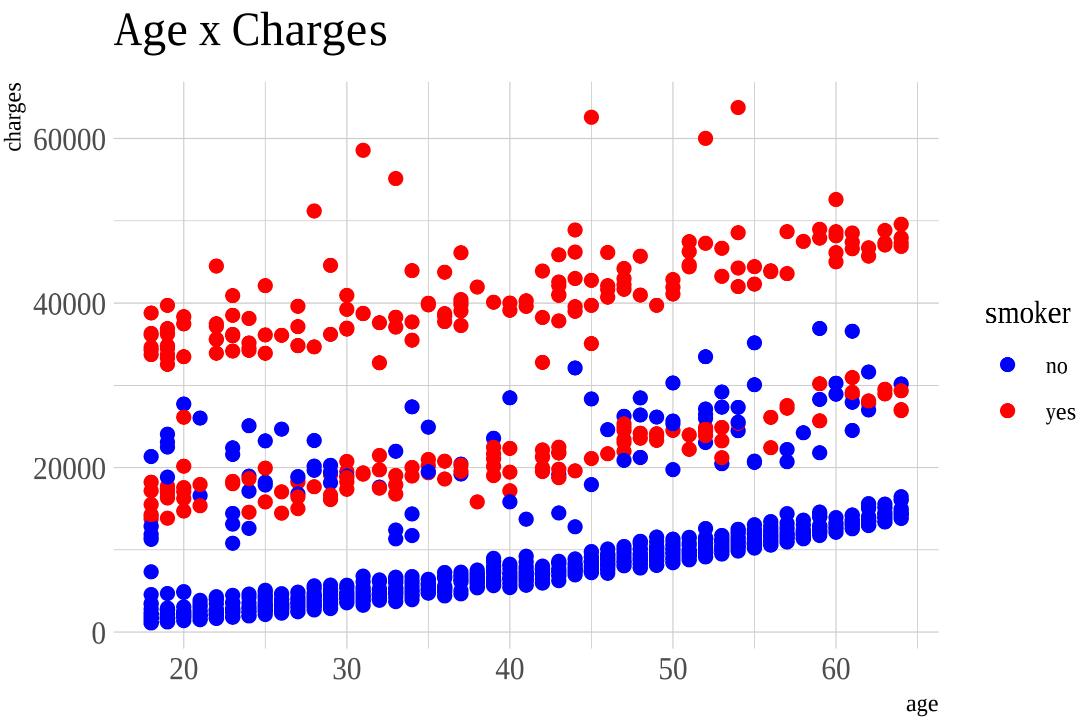


Let's explore how other variables behave in relation to our response variable.

```
plot<-ggplot(df, aes(x=bmi, y=charges, color=smoker)) +
    geom_point(size=2) +
    theme_ipsum() +
    scale_color_manual(values=c("blue", "red"))
print(plot + ggtitle("BMI x Charges"))
```



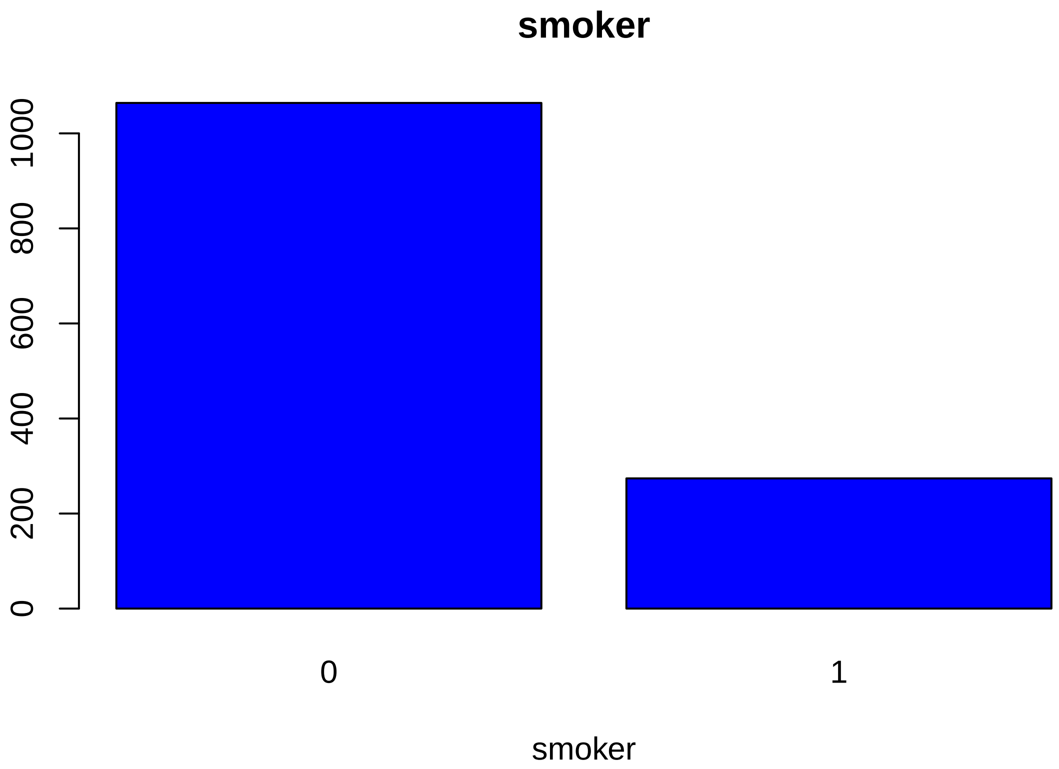
```
plot<-ggplot(df, aes(x=age, y=charges, color=smoker)) +
    geom_point(size=2) +
    theme_ipsum() +
    scale_color_manual(values=c("blue", "red"))
print(plot + ggtitle("Age x Charges"))
```



From those plots we can see that people who smoke have higher costs. It seems that's the biggest correlation with the costs even without checking the correlation. Cost also seems to increase slightly with Age and BMI. We will check further correlations in the pre-processing step.

### Pre-processing data

```
#Changing yes and no to 0 and 1 in the column smoker
df$smoker<-ifelse(df$smoker=="yes",1,0)
Counts = table(df$smoker)
barplot(Counts, main="smoker", xlab="smoker",  col = c("blue"))
```



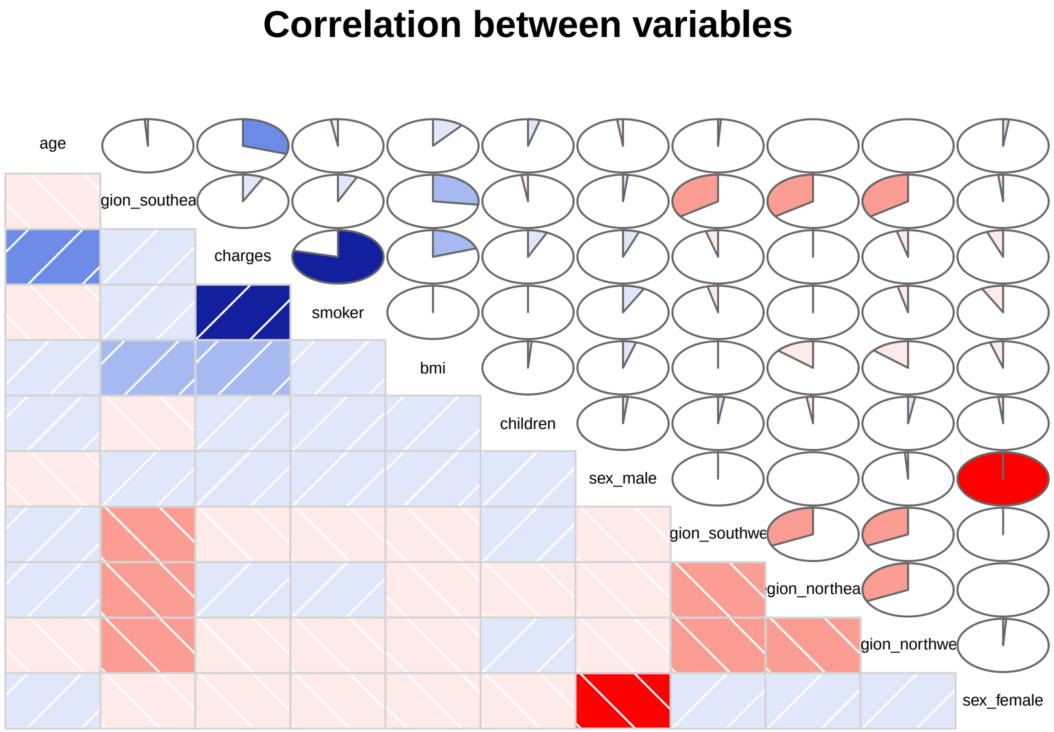
```
#One hot encoding on the categorical variables remaining
df <- dummy_cols(df,select_columns = "sex")
df <- dummy_cols(df,select_columns = "region")
```

```
df = subset(df, select = -c(region) )
df = subset(df, select = -c(sex) )
```

```
#Checking the correlation of our variables
cor(cor(df))
```

```
##                           age         bmi    children      smoker     charges
## age               1.000000000  0.10640381 -0.04474510 -0.04287092  0.27287447
## bmi               0.106403812  1.00000000 -0.10163657 -0.01718504  0.16151499
## children         -0.044745096 -0.10163657  1.00000000 -0.10467372 -0.08116170
## smoker           -0.042870924 -0.01718504 -0.10467372  1.00000000  0.92139533
## charges           0.272874468  0.16151499 -0.08116170  0.92139533  1.00000000
## sex_female        0.047469334 -0.09425464 -0.02755352 -0.15030057 -0.12558313
## sex_male         -0.047469334  0.09425464  0.02755352  0.15030057  0.12558313
## region_northeast  0.008364004 -0.26365942 -0.03651218  0.02318241  0.01254347
## region_northwest -0.002118621 -0.26357731  0.06462370 -0.06801243 -0.08568219
## region_southeast -0.020283841  0.50836250 -0.07097172  0.12123180  0.15317537
## region_southwest  0.015590976 -0.01794667  0.04812693 -0.08549361 -0.09145913
##                    sex_female     sex_male region_northeast region_northwest
## age               0.047469334 -0.047469334      0.008364004     -0.002118621
## bmi              -0.094254643  0.094254643     -0.263659424     -0.263577306
## children         -0.027553519  0.027553519     -0.036512183      0.064623697
## smoker           -0.150300570  0.150300570      0.023182412     -0.068012431
## charges          -0.125583128  0.125583128      0.012543467     -0.085682192
## sex_female        1.000000000 -1.000000000      0.007545555      0.027643224
## sex_male         -1.000000000  1.000000000     -0.007545555     -0.027643224
## region_northeast  0.007545555 -0.007545555      1.000000000     -0.299347330
## region_northwest  0.027643224 -0.027643224     -0.299347330      1.000000000
## region_southeast -0.043550315  0.043550315     -0.361527173     -0.366305472
## region_southwest  0.011545616 -0.011545616     -0.313282321     -0.310669169
##                  region_southeast region_southwest
## age                   -0.02028384       0.01559098
## bmi                    0.50836250      -0.01794667
## children              -0.07097172       0.04812693
## smoker                 0.12123180      -0.08549361
## charges                0.15317537      -0.09145913
## sex_female            -0.04355031       0.01154562
## sex_male               0.04355031      -0.01154562
## region_northeast      -0.36152717      -0.31328232
## region_northwest      -0.36630547      -0.31066917
## region_southeast       1.00000000      -0.34612202
## region_southwest      -0.34612202       1.00000000
```

```
corrgram(df, order = TRUE, lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt,
```



Smoker variable has the highest correlation with our response variable as we suspected. Age and BMI don't have a strong correlation, but are the next ones with the highest correlation with charges variable.

Next we will apply log in the response variable and make some simple re-scaling in age and BMI.

```
df$charges <- log(df$charges)
```

```
df$bmi <- (df$bmi)/100
df$age <- (df$age)/100
```

```
#Splitting data into train-test samples
set.seed(42)
split = sample.split(Y=df$charges, SplitRatio=0.8)
train = df[split,]
test = df[!split,]

dim(train)
```

```
## [1] 1070   11
```

```
dim(test)
```

```
## [1] 268  11
```

### Multiple Linear Regression

Our first ML analysis is a multiple linear regression to determine the medical costs. ‘Smoker’, ‘Age’ and ‘BMI’ are our independent variables in this scenario.

```
lr_model = lm(charges ~ smoker + age + bmi, data = train)
lr_model
```

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi, data = train)
##
## Coefficients:
## (Intercept)       smoker          age          bmi
##       7.087        1.513        3.440        1.142
```

```
summary(lr_model)
```

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25413 -0.22381 -0.03957  0.10452  2.05147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.08731    0.08034  88.214  < 2e-16 ***
## smoker       1.51327    0.03562  42.488  < 2e-16 ***
## age          3.44030    0.10295  33.419  < 2e-16 ***
## bmi          1.14196    0.23520   4.855 1.38e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4715 on 1066 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.7354
## F-statistic: 991.6 on 3 and 1066 DF,  p-value: < 2.2e-16
```

```
Prediction <-predict(lr_model, newdata=test)
results <- data.frame(Actual= test$charges, Prediction)
head(results)
```

```
##      Actual Prediction
## 1  9.734176   9.572843
## 2  7.453302   8.092201
## 4  9.998092   8.481888
## 13 7.510345   8.271411
## 16 7.516018   8.021887
## 17 9.287055   9.227758
```

```
RSQUARE = function(y_actual,y_predict){
  cor(y_actual,y_predict)^2
}
MAPE = function(y_actual,y_predict){
  mean(abs((y_actual-y_predict)/y_actual))*100
}
```

```
model_R_Squared = RSQUARE(test$charges, Prediction)
model_R_Squared
```

```
## [1] 0.7836317
```

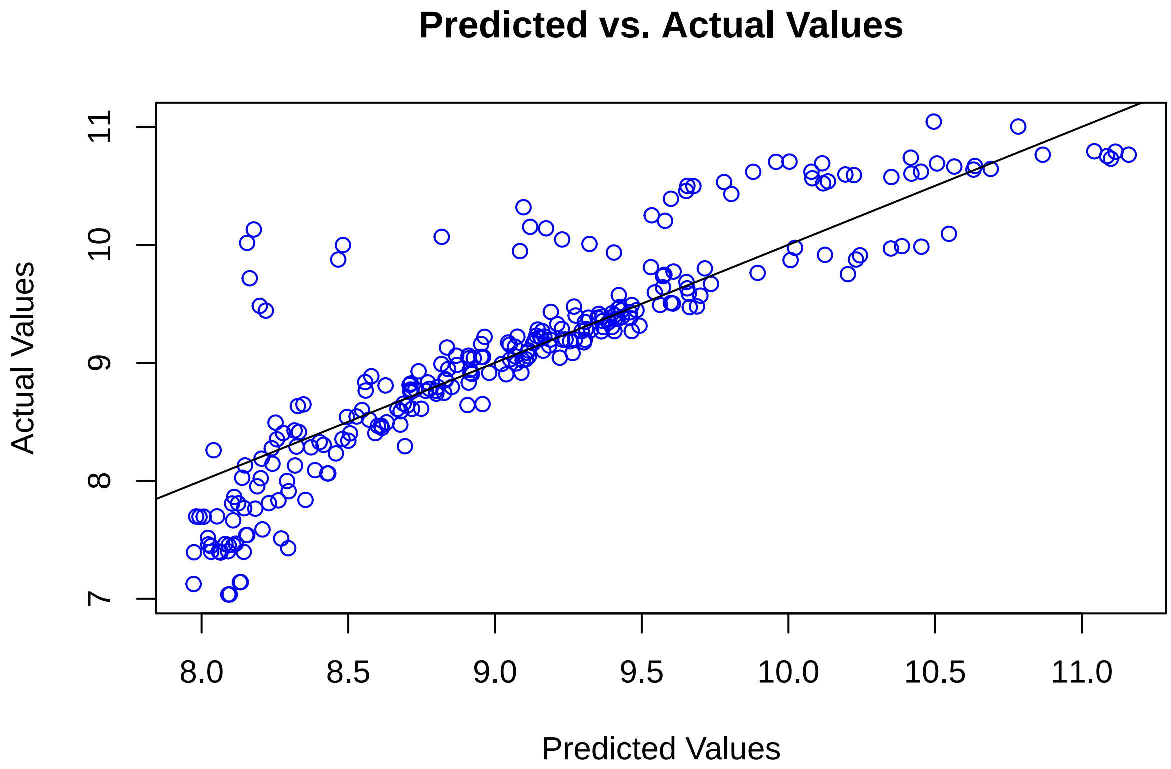
```
model_MAPE = MAPE(test$charges, Prediction)
model_MAPE
```

```
## [1] 3.207903
```

```
Accuracy_Linear = 100 - model_MAPE
Accuracy_Linear
```

```
## [1] 96.7921
```

```
plot(Prediction, y= test$charges,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs. Actual Values',col = c("blue"))
abline(a=0, b=1)
```



### Decision Tree

We will use the same variables to predict the charges with a decision tree model.

```
#train
tree_model = rpart(charges ~ smoker + bmi + age, data=train)
```

```
tree_model
```

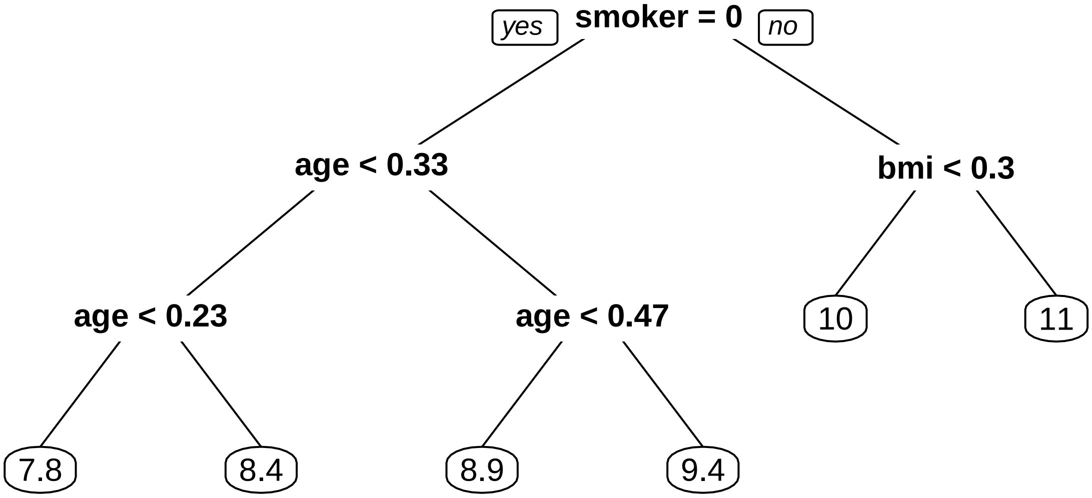
```
## n= 1070
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 1070 898.398200  9.103598
##    2) smoker< 0.5 849 470.919200  8.793176
##      4) age< 0.325 316 150.236600  8.123337
##        8) age< 0.225 139  67.231810  7.805579 *
##        9) age>=0.225 177  57.948210  8.372876 *
##      5) age>=0.325 533  94.839000  9.190303
##       10) age< 0.465 230  29.500560  8.888004 *
##       11) age>=0.465 303  28.365410  9.419771 *
##    3) smoker>=0.5 221  31.379350 10.296120
##      6) bmi< 0.3001 110   5.681614  9.966382 *
##      7) bmi>=0.3001 111   1.884914 10.622900 *
```

```
summary(tree_model)
```

```
## Call:
## rpart(formula = charges ~ smoker + bmi + age, data = train)
##   n= 1070
##
##           CP nsplit rel error    xerror       xstd
## 1 0.44089547      0 1.0000000 1.0016587 0.03531510
## 2 0.25138465      1 0.5591045 0.5604633 0.02278057
## 3 0.04115439      2 0.3077199 0.3210662 0.02069712
## 4 0.02789034      3 0.2665655 0.2733930 0.02003407
## 5 0.02650586      4 0.2386752 0.2642151 0.02208108
## 6 0.01000000      5 0.2121693 0.2282833 0.02162514
##
## Variable importance
## smoker    age    bmi
##     55     40      4
##
## Node number 1: 1070 observations,    complexity param=0.4408955
##   mean=9.103598, MSE=0.8396245
##   left son=2 (849 obs) right son=3 (221 obs)
##   Primary splits:
##       smoker < 0.5      to the left,  improve=0.44089550, (0 missing)
##       age    < 0.355    to the left,  improve=0.22833170, (0 missing)
##       bmi    < 0.23625  to the left,  improve=0.01704224, (0 missing)
##
## Node number 2: 849 observations,    complexity param=0.2513846
##   mean=8.793176, MSE=0.5546751
##   left son=4 (316 obs) right son=5 (533 obs)
##   Primary splits:
##       age < 0.325    to the left,  improve=0.47958020, (0 missing)
##       bmi < 0.238575 to the left,  improve=0.01973163, (0 missing)
##   Surrogate splits:
##       bmi < 0.20955  to the left,  agree=0.637, adj=0.025, (0 split)
##
## Node number 3: 221 observations,    complexity param=0.02650586
##   mean=10.29612, MSE=0.141988
##   left son=6 (110 obs) right son=7 (111 obs)
##   Primary splits:
##       bmi < 0.3001   to the left,  improve=0.7588692, (0 missing)
##       age < 0.435    to the left,  improve=0.1148139, (0 missing)
##   Surrogate splits:
##       age < 0.255    to the right, agree=0.534, adj=0.064, (0 split)
##
## Node number 4: 316 observations,    complexity param=0.02789034
##   mean=8.123337, MSE=0.4754324
##   left son=8 (139 obs) right son=9 (177 obs)
##   Primary splits:
##       age < 0.225    to the left,  improve=0.16678110, (0 missing)
##       bmi < 0.234325 to the left,  improve=0.01632879, (0 missing)
##   Surrogate splits:
##       bmi < 0.221825 to the left,  agree=0.589, adj=0.065, (0 split)
##
```

```
## Node number 5: 533 observations,    complexity param=0.04115439
##   mean=9.190303, MSE=0.1779343
##   left son=10 (230 obs) right son=11 (303 obs)
##   Primary splits:
##       age < 0.465    to the left,  improve=0.3898505, (0 missing)
##       bmi < 0.34785  to the left,  improve=0.0154752, (0 missing)
##   Surrogate splits:
##       bmi < 0.20025  to the left,  agree=0.576, adj=0.017, (0 split)
##
## Node number 6: 110 observations
##   mean=9.966382, MSE=0.05165104
##
## Node number 7: 111 observations
##   mean=10.6229, MSE=0.01698121
##
## Node number 8: 139 observations
##   mean=7.805579, MSE=0.4836821
##
## Node number 9: 177 observations
##   mean=8.372876, MSE=0.327391
##
## Node number 10: 230 observations
##   mean=8.888004, MSE=0.1282633
##
## Node number 11: 303 observations
##   mean=9.419771, MSE=0.09361522
```

```
prp(tree_model)
```



```
predic_tree = predict(tree_model, test)
head(predic_tree)
```

```
##        1        2        4       13       16       17
## 9.966382 7.805579 8.888004 8.372876 7.805579 9.419771
```

```
comp_tree = cbind(predic_tree, test$charges, predic_tree - test$charges)
```

```
head(comp_tree)
```

```
##    predic_tree
## 1     9.966382 9.734176  0.2322058
## 2     7.805579 7.453302  0.3522764
## 4     8.888004 9.998092 -1.1100873
## 13    8.372876 7.510345  0.8625317
## 16    7.805579 7.516018  0.2895608
## 17    9.419771 9.287055  0.1327160
```

```
accuracy(predic_tree, test$charges)
```

```
##                   ME      RMSE       MAE        MPE     MAPE
## Test set -0.0416007 0.3715601 0.2435343 -0.6673216 2.752852
```

### Classification with Decision Tree

Now, let's change our point of view.

People can lie about their smoking habits when filling their register for insurance. This can be configured as fraud since it will generate higher insurance costs.

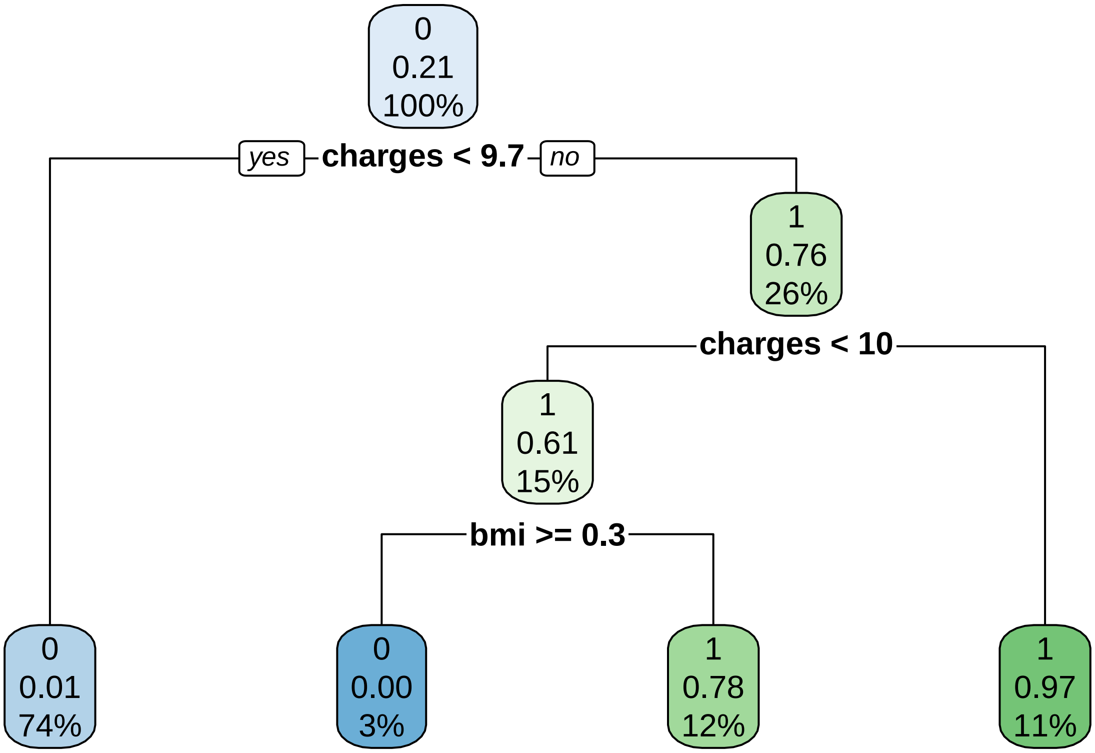
Suppose that we already have the medical costs, we want to determine if people smoke or not.

```
tree_model_class = rpart(smoker ~ ., data=train, method="class")
tree_model_class
```

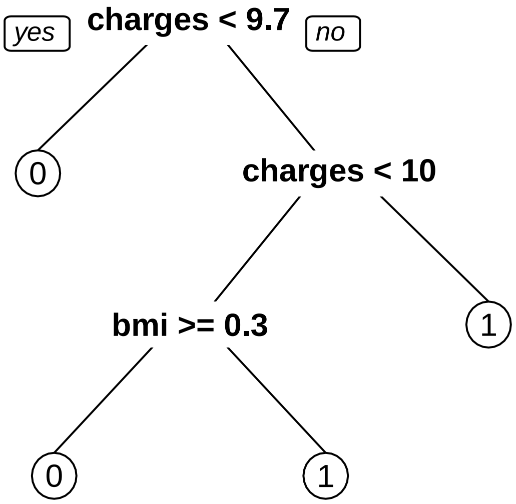
```
## n= 1070
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 1070 221 0 (0.79345794 0.20654206)
##    2) charges< 9.686588 790   8 0 (0.98987342 0.01012658) *
##    3) charges>=9.686588 280  67 1 (0.23928571 0.76071429)
##      6) charges< 10.41852 163  64 1 (0.39263804 0.60736196)
##       12) bmi>=0.300675 36   0 0 (1.00000000 0.00000000) *
##       13) bmi< 0.300675 127  28 1 (0.22047244 0.77952756) *
##      7) charges>=10.41852 117   3 1 (0.02564103 0.97435897) *
```

```
predic_tree_class = predict(tree_model_class, test, type ="class")
#predic_tree_class
```

```
rpart.plot(tree_model_class)
```



```
prp(tree_model_class)
```



```r
rf_class <- predict(tree_model_class, newdata = test, type = "class")
predictions <- cbind(data.frame(train_preds=rf_class,
                                test$smoker))
#predictions
```

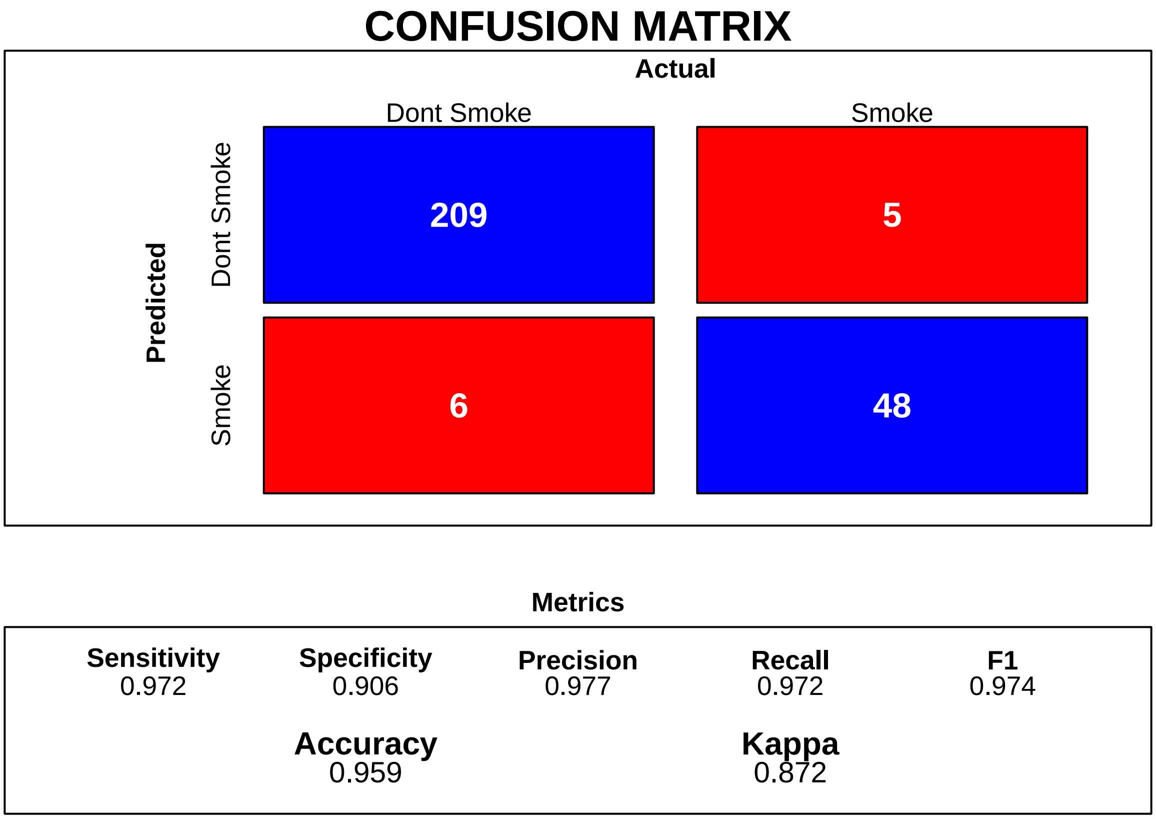
```r
cm <- caret::confusionMatrix(predictions$train_preds, as.factor(predictions$test.smoker))
print(cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 209   5
##          1   6  48
##
##                Accuracy : 0.959
##                  95% CI : (0.9277, 0.9793)
##     No Information Rate : 0.8022
##     P-Value [Acc > NIR] : 5.786e-14
##
##                   Kappa : 0.8716
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9721
##             Specificity : 0.9057
##          Pos Pred Value : 0.9766
##          Neg Pred Value : 0.8889
##              Prevalence : 0.8022
##          Detection Rate : 0.7799
##    Detection Prevalence : 0.7985
##       Balanced Accuracy : 0.9389
##
##        'Positive' Class : 0
##
```

```
draw_confusion_matrix <- function(cm) {

  layout(matrix(c(1,1,2)))
  par(mar=c(2,2,2,2))
  plot(c(100, 345), c(300, 450), type = "n", xlab="", ylab="", xaxt='n', yaxt='n')
  title('CONFUSION MATRIX', cex.main=2)

  # create the matrix
  rect(150, 430, 240, 370, col='blue')
  text(195, 435, 'Dont Smoke', cex=1.2)
  rect(250, 430, 340, 370, col='red')
  text(295, 435, 'Smoke', cex=1.2)
  text(125, 370, 'Predicted', cex=1.3, srt=90, font=2)
  text(245, 450, 'Actual', cex=1.3, font=2)
  rect(150, 305, 240, 365, col='red')
  rect(250, 305, 340, 365, col='blue')
  text(140, 400, 'Dont Smoke', cex=1.2, srt=90)
  text(140, 335, 'Smoke', cex=1.2, srt=90)

  # add in the cm results
  res <- as.numeric(cm$table)
  text(195, 400, res[1], cex=1.6, font=2, col='white')
  text(195, 335, res[2], cex=1.6, font=2, col='white')
  text(295, 400, res[3], cex=1.6, font=2, col='white')
  text(295, 335, res[4], cex=1.6, font=2, col='white')

  # add in the specifics
  plot(c(100, 0), c(100, 0), type = "n", xlab="", ylab="", main = "Metrics", xaxt='n', yaxt='n')
  text(10, 85, names(cm$byClass[1]), cex=1.2, font=2)
  text(10, 70, round(as.numeric(cm$byClass[1]), 3), cex=1.2)
  text(30, 85, names(cm$byClass[2]), cex=1.2, font=2)
  text(30, 70, round(as.numeric(cm$byClass[2]), 3), cex=1.2)
  text(50, 85, names(cm$byClass[5]), cex=1.2, font=2)
  text(50, 70, round(as.numeric(cm$byClass[5]), 3), cex=1.2)
  text(70, 85, names(cm$byClass[6]), cex=1.2, font=2)
  text(70, 70, round(as.numeric(cm$byClass[6]), 3), cex=1.2)
  text(90, 85, names(cm$byClass[7]), cex=1.2, font=2)
  text(90, 70, round(as.numeric(cm$byClass[7]), 3), cex=1.2)

  # add in the accuracy information
  text(30, 35, names(cm$overall[1]), cex=1.5, font=2)
  text(30, 20, round(as.numeric(cm$overall[1]), 3), cex=1.4)
  text(70, 35, names(cm$overall[2]), cex=1.5, font=2)
```

```r
  text(70, 20, round(as.numeric(cm$overall[2]), 3), cex=1.4)
}
```

```r
draw_confusion_matrix(cm)
```



Let’s check our False Positives:

Instead of thinking about it just as a prediction error from the model, we can also look at it as people who stated they don’t smoke but our model predicts they do. Considering the strong correlation of the smoker variable with the charges, this could mean financial loss to the company if those people are lying in their register. It sounds reasonable for the company to investigate those cases and similar profiles in the future.