Forecasting Hospital Expenses

Caio di Felice Cunha

Definition of the Business Problem

Understand and predict Hospital expenses

For this analysis, we will use a dataset simulating hypothetical medical expenses for a set of patients spread across 4 US regions. This dataset has 1338 observations and 7 variables.

Stage 1 - Collecting the Data

```
# Step 1 - Collecting the data
expensesdf <- read.csv("expenses.csv")
```

Stage 2 - Exploring and Preparing the Data

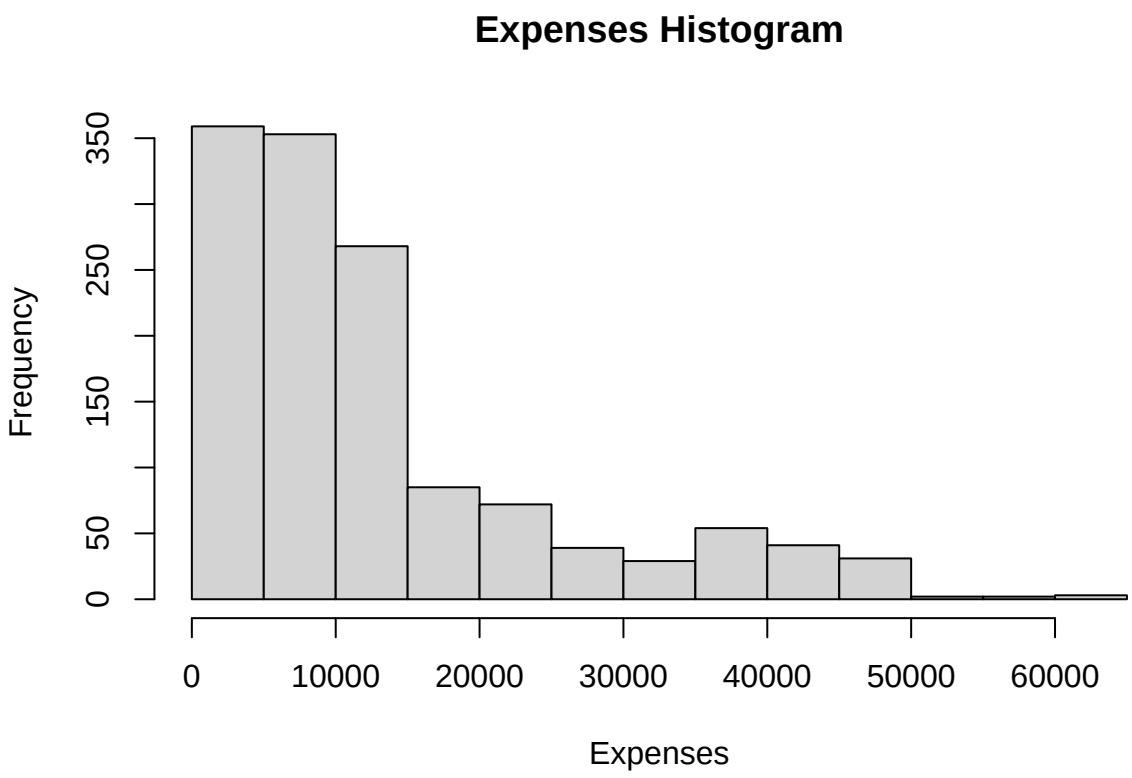
```
# Step 2: Exploring and Preparing the Data
# heading the variables
str(expensesdf)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "woman" "man" "man" "man" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southeast" "south" "south" "north east" ...
##  $ expenses: num  16885 1726 4449 21984 3867 ...
```

```
# Measures of Central Tendency of the variable expenses
summary(expensesdf$expenses)
```

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

```
# Building a histogram
hist(expensesdf$expenses, main = 'Expenses Histogram', xlab = 'Expenses')
```

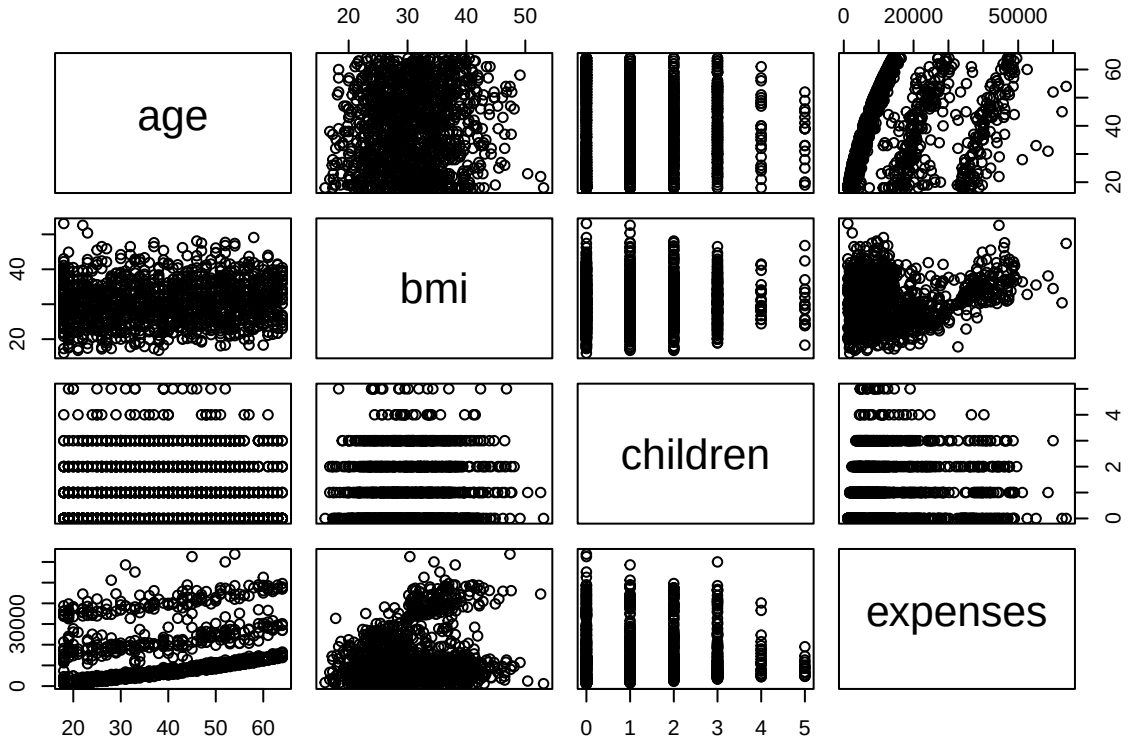



```
# Exploring relationships between variables: Correlation Matrix
cor(expensesdf[c("age", "bmi", "children", "expenses")])
```

```
##                age        bmi   children   expenses
## age      1.0000000 0.10934101 0.04246900 0.29900819
## bmi      0.1093410 1.00000000 0.01264471 0.19857626
## children 0.0424690 0.01264471 1.00000000 0.06799823
## expenses 0.2990082 0.19857626 0.06799823 1.00000000
```

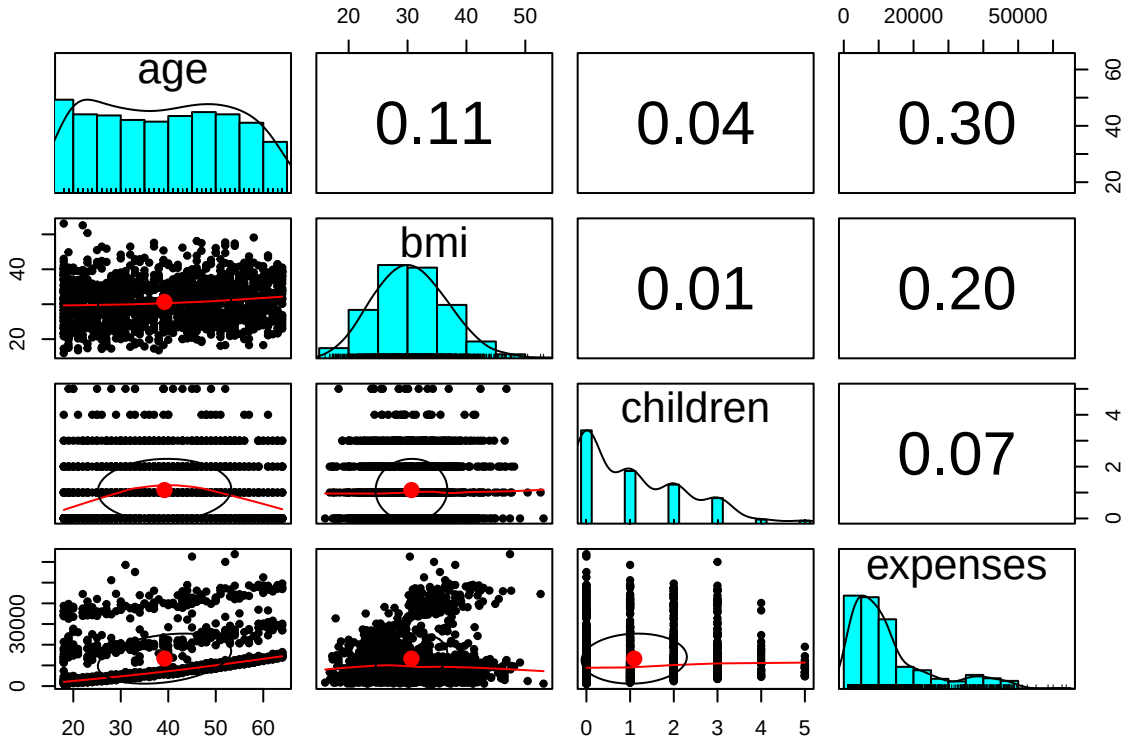
None of the correlations in the matrix are considered strong, but there are some interesting associations. For example, age and bmi (BMI) seem to have a weak positive correlation, meaning that with increasing age, body mass tends to increase. There is also a positive correlation moderate between age and expenses, in addition to the number of children and expenses. These associations imply that as you age, body mass and number of children increase, the expected cost of health insurance goes up.

```
# Visualizing relationships between variables: Scatterplot
# Realize that there is no clear relationship between the variables
pairs(expensesdf[c("age", "bmi", "children", "expenses")])
```




```
# Scatterplot Matrix
#install.packages("psych")
library(psych)

# This plot provides more information about the relationship between variables
pairs.panels(expensesdf[c("age", "bmi", "children", "expenses")])
```



Stage 3 - Training the model (using the training data)

```
model <- lm(expenses ~ ., data = expensesdf)

# Visualizing the coefficients
model
```

```
##
## Call:
## lm(formula = expenses ~ ., data = expensesdf)
##
## Coefficients:
##      (Intercept)               age          sexwoman               bmi
##         -12072.9             256.8             131.4             339.3
##         children         smokeryes  regionnorth east       regionsouth
##            475.7           23847.5            -352.8           -1035.6
##  regionsoutheast
##           -959.3
```

```
# Anticipating medical expensesdf
# Here we check the expenses predicted by the model which must be equal to
#the training data
forecast1 <- predict(model)
head(forecast1)
```

```
##         1         2         3         4         5         6
## 25292.740  3458.281  6706.619  3751.868  5598.626  3704.606
```

```
# Predicting expenses with test data
expensesdftest <- read.csv("expenses-test.csv")
head(expensesdftest)
```

```
##   age   sex  bmi children smoker     region
## 1  52 woman 26.6        0     no north east
## 2  27   man 27.1        0     no      south
## 3  26 woman 29.9        1     no      south
## 4  24 woman 22.2        0     no      south
## 5  34 woman 33.7        1     no  southeast
## 6  53 woman 33.3        0     no      north
```

```
forecast2 <- predict(model, expensesdftest)
head(forecast2)
```

```
##          1          2          3          4          5          6
## 10086.3947  3020.9027  4321.1161   719.2169  7741.4208 12969.2660
```

Stage 4 - Evaluating the Model's Performance

```
# More details about the model
summary(model)
```

```
##
## Call:
## lm(formula = expenses ~ ., data = expensesdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -12072.9      999.6 -12.077  < 2e-16 ***
## age                 256.8       11.9  21.586  < 2e-16 ***
## sexwoman            131.3      332.9   0.395 0.693255
## bmi                 339.3       28.6  11.864  < 2e-16 ***
## children            475.7      137.8   3.452 0.000574 ***
## smokeryes         23847.5      413.1  57.723  < 2e-16 ***
## regionnorth east   -352.8      476.3  -0.741 0.458976
## regionsouth       -1035.6      478.7  -2.163 0.030685 *
## regionsoutheast    -959.3      477.9  -2.007 0.044921 *
## ---
## 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: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Stage 5: Optimizing Model Performance

```
# Adding a variable with twice the value of ages
expensesdf$age2 <- expensesdf$age ^ 2
```

```
# Adding an indicator for BMI >= 30
expensesdf$bmi30 <- ifelse(expensesdf$bmi >= 30, 1, 0)

head(expensesdf)
```

```
##   age   sex  bmi children smoker     region expenses age2 bmi30
## 1  19 woman 27.9        0    yes  southeast 16884.92  361     0
## 2  18   man 33.8        1     no      south  1725.55  324     1
## 3  28   man 33.0        3     no      south  4449.46  784     1
## 4  33   man 22.7        0     no north east 21984.47 1089     0
## 5  32   man 28.9        0     no north east  3866.86 1024     0
## 6  31 woman 25.7        0     no      south  3756.62  961     0
```

```
# Creating the final model
model_v2 <- lm(expenses ~ age + age2 + children + bmi + sex +
                  bmi30 * smoker + region, data = expensesdf)

summary(model_v2)
```

```
##
## Call:
## lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
##     smoker + region, data = expensesdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17297.1  -1656.0  -1262.7   -727.8  24161.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -357.7636  1364.4505  -0.262 0.793205
## age                -32.6181    59.8250  -0.545 0.585690
## age2                 3.7307     0.7463   4.999 6.54e-07 ***
## children           678.6017   105.8855   6.409 2.03e-10 ***
## bmi                119.7715    34.2796   3.494 0.000492 ***
## sexwoman           496.7690   244.3713   2.033 0.042267 *
## bmi30             -997.9355   422.9607  -2.359 0.018449 *
## smokeryes        13404.5952   439.9591  30.468  < 2e-16 ***
## regionnorth east  -279.1661   349.2826  -0.799 0.424285
## regionsouth       -828.0345   351.6484  -2.355 0.018682 *
## regionsoutheast  -1222.1619   350.5314  -3.487 0.000505 ***
## bmi30:smokeryes  19810.1534   604.6769  32.762  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4445 on 1326 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8653
## F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

```
# test data

expensesdftest <- read.csv("expenses-test.csv")
head(expensesdftest)
```

```
##   age   sex  bmi children smoker     region
## 1  52 woman 26.6        0     no north east
## 2  27   man 27.1        0     no      south
## 3  26 woman 29.9        1     no      south
## 4  24 woman 22.2        0     no      south
## 5  34 woman 33.7        1     no  southeast
## 6  53 woman 33.3        0     no      north
```

```
forecast3 <- predict(model, expensesdftest)
class(forecast3)
```

```
## [1] "numeric"
```

```
head(forecast3)
```

```
##          1          2          3          4          5          6
## 10086.3947  3020.9027  4321.1161   719.2169  7741.4208 12969.2660
```

Disclaimer:

Disclaimer: a good part of this project was largely done in the Data Science Academy, Big Data Analytics with R and Microsoft Azure Machine Learning course (part of the Data Scientist training)

End