Preventing Hospital Expenses

Oracy Martos

November 22, 2018

Preventing Hospital Expenses

For this analysis, we will use a data set simulating hypothetical medical expenses for a set of patients spread across 4 regions of Brazil. This dataset has 1,338 observations and 7 variables.

Step 1 - Data gathering

```
# Data gathering
df <- read.csv('C:\\Users\\Oracy\\Desktop\\DSA_Projetos\\DSA_Projetos\\Big Data Analytics com R e Micros
head(df)
```

```
##   idade   sexo  bmi filhos fumante   regiao   gastos
## 1    19 mulher 27.9      0     sim  sudeste 16884.92
## 2    18  homem 33.8      1     nao      sul  1725.55
## 3    28  homem 33.0      3     nao      sul  4449.46
## 4    33  homem 22.7      0     nao nordeste 21984.47
## 5    32  homem 28.9      0     nao nordeste  3866.86
## 6    31 mulher 25.7      0     nao      sul  3756.62
```

Etapa 2 - Explorando os Dados

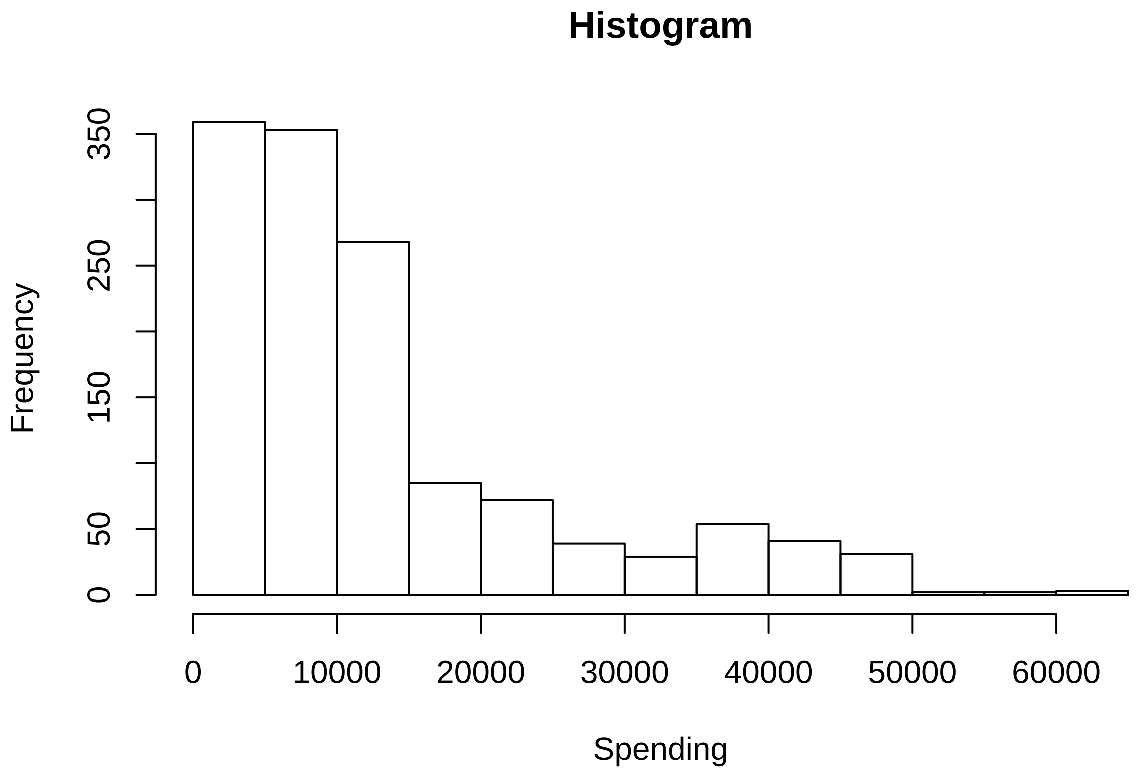
```
# Viewing variables
str(df)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ idade  : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sexo   : Factor w/ 2 levels "homem","mulher": 2 1 1 1 1 2 2 2 1 2 ...
##  $ bmi    : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
##  $ filhos : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ fumante: Factor w/ 2 levels "nao","sim": 2 1 1 1 1 1 1 1 1 1 ...
##  $ regiao : Factor w/ 4 levels "nordeste","norte",..: 3 4 4 1 1 4 4 1 2 1 ...
##  $ gastos : num  16885 1726 4449 21984 3867 ...
```

```
# Central Trend Averages of the variable spending
summary(df[c("gastos")])
```

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

```r
# Building a Histogram
hist(df$gastos, main = 'Histogram', xlab = 'Spending')
```



```r
# Regions contingency table
table(df$regiao)
```

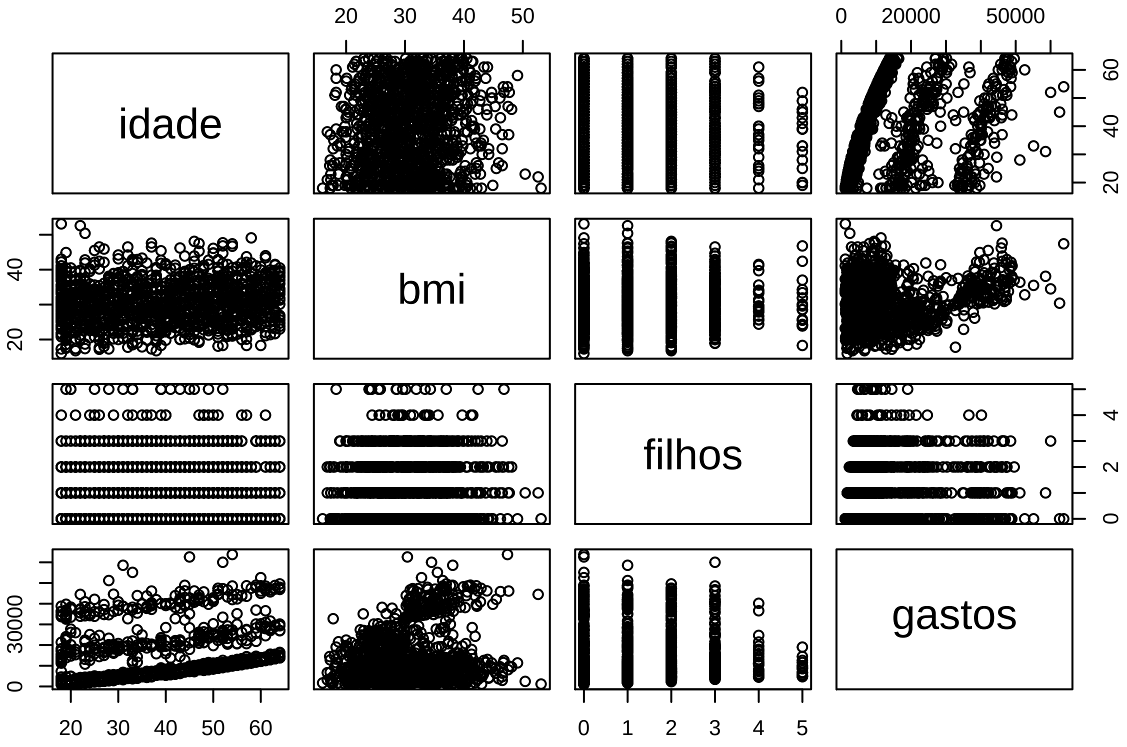
```
##
## nordeste    norte  sudeste      sul
##      325      324      325      364
```

```r
# Exploring relationships among variables: Correlation Matrix
cor(df[c("idade","bmi", "filhos", "gastos")])
```

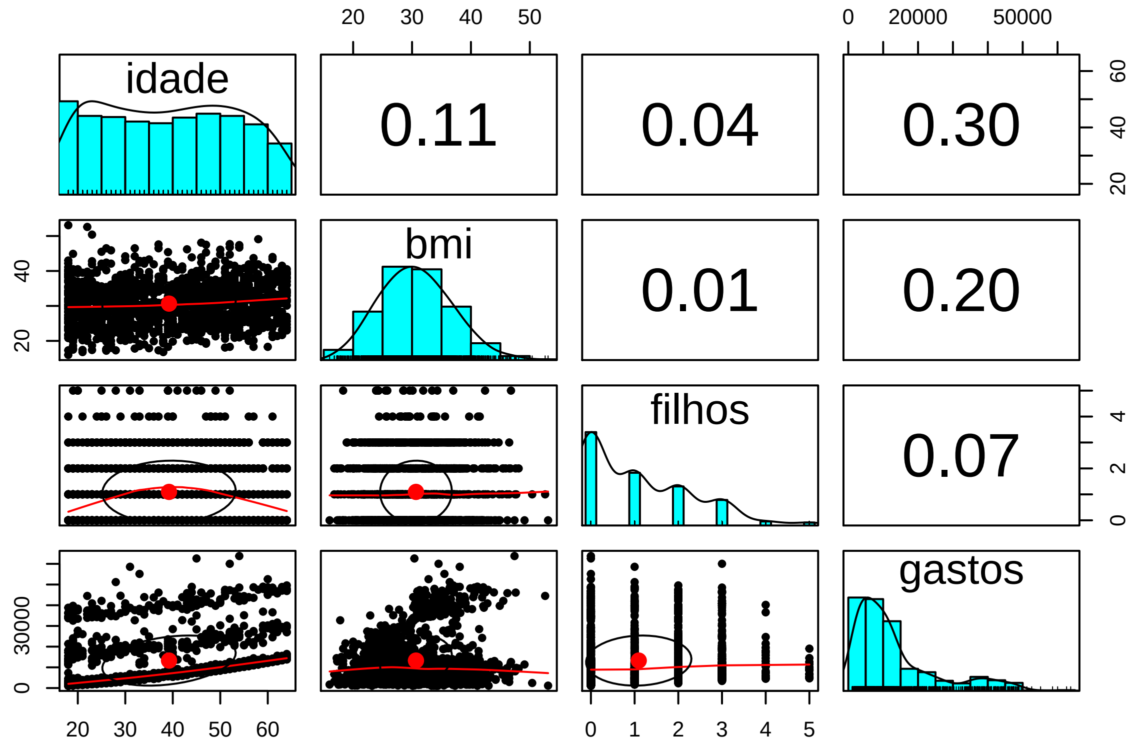
```
##            idade        bmi     filhos     gastos
## idade  1.0000000 0.10934101 0.04246900 0.29900819
## bmi    0.1093410 1.00000000 0.01264471 0.19857626
## filhos 0.0424690 0.01264471 1.00000000 0.06799823
## gastos 0.2990082 0.19857626 0.06799823 1.00000000
```

```r
# None of the correlations in the matrix are considered strong, but there are some interesting associat
# For example, age and bmi (BMI) appear to have a weak positive correlation, which means that
# As age increases, body mass tends to increase. There is also a positive correlation
# Moderate between age and expenditure, in addition to the number of children and expenses. These assoc
# that as the average age, body mass and number of children increases, the expected cost of health insu
# Viewing relationship between variables: Scatterplot
# Note that there is no clear relationship between the variables
pairs(df[c("idade", "bmi", "filhos", "gastos")])

# Scatterplot Matrix
# Font: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs#use-the-r-package-psych
#install.packages ("psych")
library(psych)
```



```
pairs.panels(df[c("idade", "bmi", "filhos", "gastos")], method = "pearson", # correlation method
             density = TRUE, # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```



```
# This graphic provides more information about the relationship between variables
```

Step 3: Training the Model

```
# Font: https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/lm
str(df)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ idade  : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sexo   : Factor w/ 2 levels "homem","mulher": 2 1 1 1 1 2 2 2 1 2 ...
##  $ bmi    : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
##  $ filhos : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ fumante: Factor w/ 2 levels "nao","sim": 2 1 1 1 1 1 1 1 1 1 ...
##  $ regiao : Factor w/ 4 levels "nordeste","norte",..: 3 4 4 1 1 4 4 1 2 1 ...
##  $ gastos : num  16885 1726 4449 21984 3867 ...
```

```
model <- lm(gastos ~ idade + filhos + bmi + sexo + fumante + regiao, df)

# Similar to the previous item
model_2 <- lm(gastos ~ ., df) # "." is the same as type all variables

# Viewing the coefficients
# Font: https://stackoverflow.com/questions/6577058/extract-regression-coefficient-values
```

```
model_summary <- summary(model)
model_summary
```

```
##
## Call:
## lm(formula = gastos ~ idade + filhos + bmi + sexo + fumante +
##     regiao, data = df)
##
## 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)   -12425.7     1000.7 -12.418  < 2e-16 ***
## idade            256.8       11.9  21.586  < 2e-16 ***
## filhos           475.7      137.8   3.452 0.000574 ***
## bmi              339.3       28.6  11.864  < 2e-16 ***
## sexomulher       131.3      332.9   0.395 0.693255
## fumantesim     23847.5      413.1  57.723  < 2e-16 ***
## regiaonorte      352.8      476.3   0.741 0.458976
## regiaosudeste   -606.5      477.2  -1.271 0.203940
## regiaosul       -682.8      478.9  -1.426 0.154211
## ---
## 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
```

```
model
```

```
##
## Call:
## lm(formula = gastos ~ idade + filhos + bmi + sexo + fumante +
##     regiao, data = df)
##
## Coefficients:
##   (Intercept)          idade         filhos            bmi     sexomulher
##      -12425.7          256.8          475.7          339.3          131.4
##    fumantesim    regiaonorte  regiaosudeste      regiaosul
##       23847.5          352.8         -606.5         -682.8
```

```
# Preventing medical expenses
# Font: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.lm.html
predicting <- predict(model)
class(predicting)
```

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

```
head(predicting)
```

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

Step 4: Evaluating Model Performance

```
# More details about the model
# Font: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
summary(model)
```

```
## 
## Call:
## lm(formula = gastos ~ idade + filhos + bmi + sexo + fumante + 
##     regiao, data = df)
## 
## 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)    -12425.7     1000.7 -12.418  < 2e-16 ***
## idade             256.8       11.9  21.586  < 2e-16 ***
## filhos            475.7      137.8   3.452 0.000574 ***
## bmi               339.3       28.6  11.864  < 2e-16 ***
## sexomulher        131.3      332.9   0.395 0.693255    
## fumantesim      23847.5      413.1  57.723  < 2e-16 ***
## regiaonorte       352.8      476.3   0.741 0.458976    
## regiaosudeste    -606.5      477.2  -1.271 0.203940    
## regiaosul        -682.8      478.9  -1.426 0.154211    
## ---
## 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
```

Step 5: Optimizing Model Performance

```
# Adding a variable with twice the age value
df$idade2 <- df$idade * 2
#df$idade2
#df$idade

# Adding a Bookmark to BMI> = 30
# Font: https://www.datamentor.io/r-programming/ifelse-function/
df$bmi30 <- ifelse(df$bmi >= 30, 1, 0)
#df$bmi30
```

```r
# Creating the final template
str(df)
```

```
## 'data.frame':    1338 obs. of  9 variables:
##  $ idade  : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sexo   : Factor w/ 2 levels "homem","mulher": 2 1 1 1 1 2 2 2 1 2 ...
##  $ bmi    : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
##  $ filhos : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ fumante: Factor w/ 2 levels "nao","sim": 2 1 1 1 1 1 1 1 1 1 ...
##  $ regiao : Factor w/ 4 levels "nordeste","norte",..: 3 4 4 1 1 4 4 1 2 1 ...
##  $ gastos : num  16885 1726 4449 21984 3867 ...
##  $ idade2 : num  38 36 56 66 64 62 92 74 74 120 ...
##  $ bmi30  : num  0 1 1 0 0 0 1 0 0 0 ...
```

```r
model_3 <- lm(gastos ~ idade + idade2 + sexo + filhos + bmi30 * fumante + regiao, df)

summary(model_3)
```

```
##
## Call:
## lm(formula = gastos ~ idade + idade2 + sexo + filhos + bmi30 *
##     fumante + regiao, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18829.3  -1872.4  -1306.7   -582.2  24710.6
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2713.616    470.681  -5.765 1.01e-08 ***
## idade             265.486      8.812  30.126  < 2e-16 ***
## idade2                 NA         NA      NA       NA
## sexomulher        479.924    247.474   1.939   0.0527 .
## filhos            524.041    102.338   5.121 3.49e-07 ***
## bmi30             201.748    281.354   0.717   0.4735
## fumantesim      13360.059    445.408  29.995  < 2e-16 ***
## regiaonorte       278.942    353.726   0.789   0.4305
## regiaosudeste    -881.758    353.916  -2.491   0.0128 *
## regiaosul        -308.196    348.633  -0.884   0.3768
## bmi30:fumantesim 19856.483    612.121  32.439  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4502 on 1328 degrees of freedom
## Multiple R-squared:  0.8627, Adjusted R-squared:  0.8618
## F-statistic: 927.4 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Fim