Predicting_medical_expenses

Felipe Henrique da Silva

20/04/2021

Project 3 - Predicting Hospital Expenses

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

Step 1 - Collecting the Data

Here is the data collection, in this case a csv file.

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

Step 2: Exploring and Preparing the Data

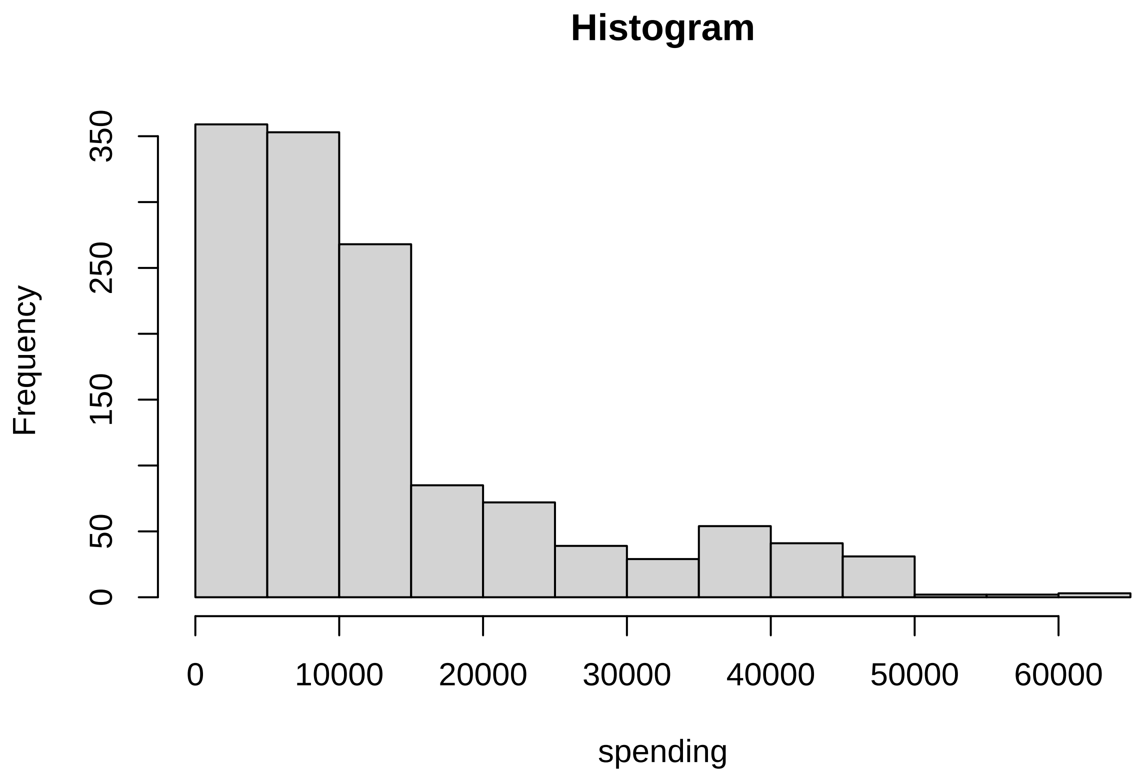
```
# Viewing the variables (Age,Sex,bmi,children,smoking, region, spending)
str(despesas)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ idade  : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sexo   : chr  "mulher" "homem" "homem" "homem" ...
##  $ 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: chr  "sim" "nao" "nao" "nao" ...
##  $ regiao : chr  "sudeste" "sul" "sul" "nordeste" ...
##  $ gastos : num  16885 1726 4449 21984 3867 ...
```

```
# Central Trend Means of the spending variable
summary(despesas$gastos)
```

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

```
# Building a histogram
hist(despesas$gastos, main = 'Histogram', xlab = 'spending')
```




```
# Contingency table for regions
table(despesas$regiao)
```

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

```
# Exploring the relationship between variables: Correlation Matrix
cor(despesas[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
```

```
# 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
# 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 associat
# that as age, body mass and number of children increase, the expected cost of health insurance goes up

# Viewing the relationship between variables: Scatterplot
# Realize that there is no clear relationship between the variables
pairs(despesas[c("idade", "bmi", "filhos", "gastos")])
colunas_numericas <- sapply(despesas, is.numeric)
colunas_numericas
```

```
##    idade     sexo      bmi   filhos fumante   regiao   gastos
##     TRUE    FALSE     TRUE     TRUE   FALSE    FALSE     TRUE
```

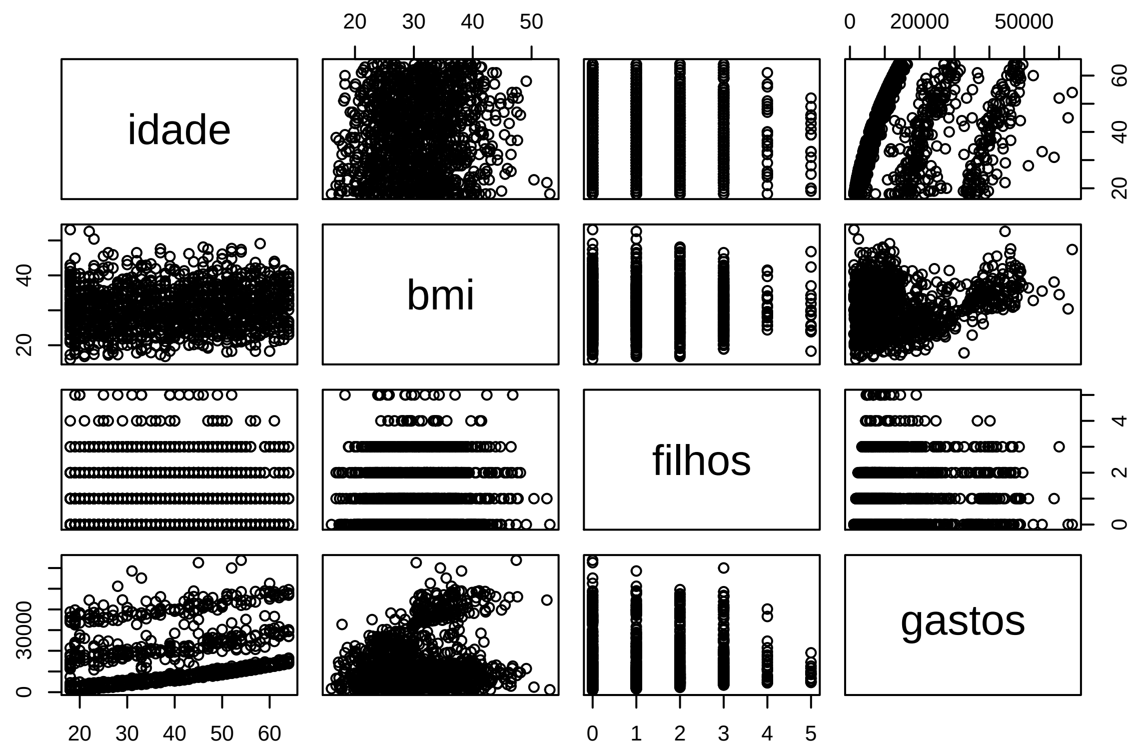
```
data_cor <- cor(despesas[,colunas_numericas])
data_cor
```

```
##            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
```

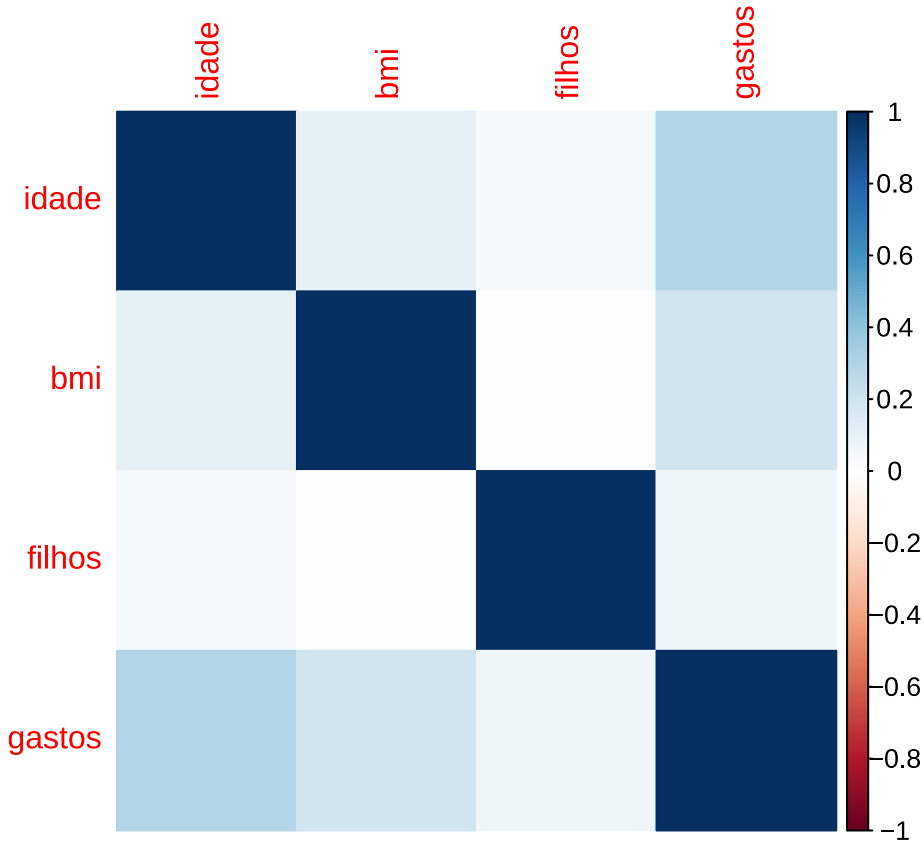
```
# install.packages('corrgram')
# install.packages('corrplot')
library(corrplot)
```

```
## corrplot 0.84 loaded
```

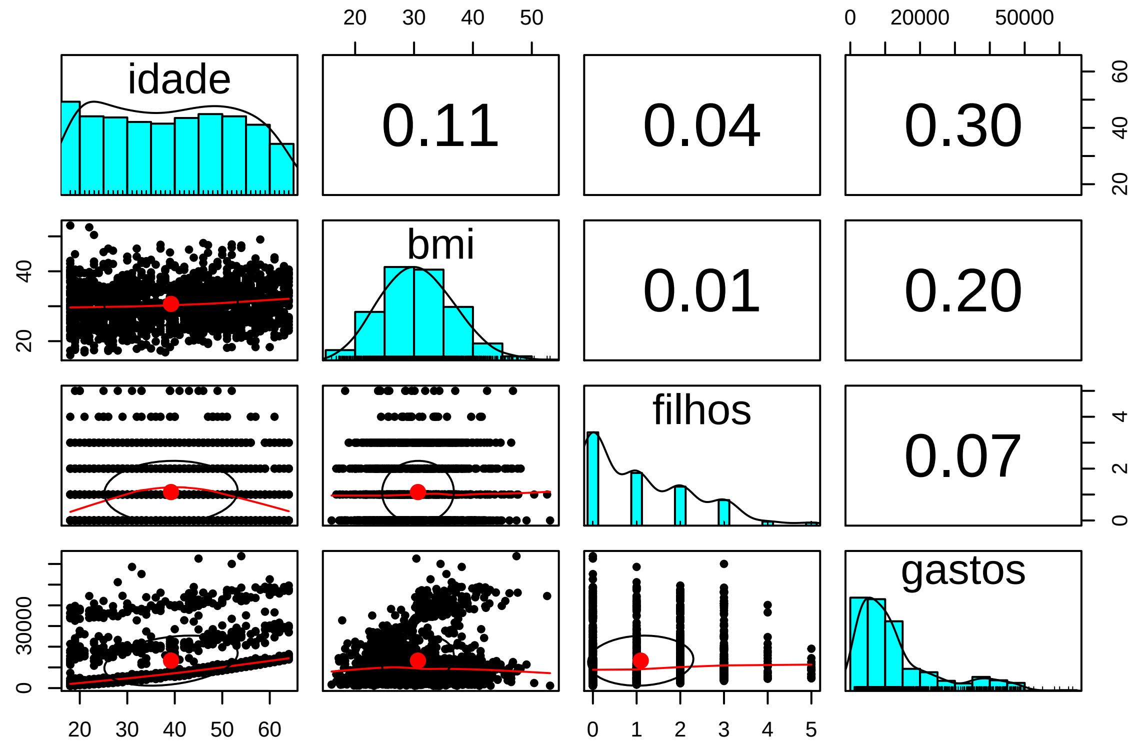
```
library(corrgram)
```




```
corrplot(data_cor, method = 'color')
```




```r
# Scatterplot Matrix
# install.packages("psych")
library(psych)
# This chart provides more information about the relationship between variables
pairs.panels(despesas[c("idade", "bmi", "filhos", "gastos")])
```



Step 3: Training the Model

```
modelo <- lm(gastos ~ idade + filhos + bmi + sexo + fumante + regiao,
             data = despesas)
modelo
```

```
##
## Call:
## lm(formula = gastos ~ idade + filhos + bmi + sexo + fumante +
##     regiao, data = despesas)
##
## 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
```

```
# Similar to the previous item
modelo <- lm(gastos ~ ., data = despesas)

# Viewing the coefficients
modelo
```

```
##
## Call:
## lm(formula = gastos ~ ., data = despesas)
##
```

```
## Coefficients:
##   (Intercept)          idade     sexomulher            bmi         filhos
##      -12425.7          256.8          131.4          339.3          475.7
##    fumantesim    regiaonorte  regiaosudeste      regiaosul
##       23847.5          352.8         -606.5         -682.8
```

```r
# Predicting medical expenses
previsao <- predict(modelo)
class(previsao)
```

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

```r
head(previsao)
```

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

#Step 4: Assessing the Model's Performance

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

```
##
## Call:
## lm(formula = gastos ~ ., data = despesas)
##
## 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 ***
## sexomulher       131.3      332.9   0.395 0.693255
## bmi              339.3       28.6  11.864  < 2e-16 ***
## filhos           475.7      137.8   3.452 0.000574 ***
## 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 the Model's Performance

```r
# Adding a variable with twice the age value

# One of the main differences from regression modeling to other Machine Learning techniques is that reg
# typically leaves the selection of the model specification characteristics to the analyst. Consequentl
# if we have enough information on how the selection of variables is related to the result, we can use
# information to specify the model and thus improve performance.
```

```r
# In linear regression, the relationship between the independent variable and the dependent variable is
# although this may not always be true. For example, the effect of age on medical expenses may not be c
# through all ages. Medical treatment may be disproportionately higher among the older population.

# Linear regression responds by the formula: y = A + Bx

# However, in some situations, we may want to include a non-linear relationship, adding a higher order
# treating the model as polynomial. Therefore, the formula will be: y = A + B1x + B2x^2

# The difference between these two equations is that the additional item B2 (Beta coefficient) will be
#thus capturing the impact of age as a function of age squared.

# By adding age and age2 to the model, this will allow us to separate the linear and non-linear impact
# ** The creation of the age2 variable could lead to questions about multicollinearity. See an explanat

despesas$idade2 <- despesas$idade ^ 2

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

# Creating the final model
modelo_v2 <- lm(gastos ~ idade + idade2 + filhos + bmi + sexo +
                  bmi30 * fumante + regiao, data = despesas)

summary(modelo_v2)
```

```
##
## Call:
## lm(formula = gastos ~ idade + idade2 + filhos + bmi + sexo +
##     bmi30 * fumante + regiao, data = despesas)
##
## 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)        -636.9298  1361.0589  -0.468 0.639886
## idade               -32.6181    59.8250  -0.545 0.585690
## idade2                3.7307     0.7463   4.999 6.54e-07 ***
## filhos              678.6017   105.8855   6.409 2.03e-10 ***
## bmi                 119.7715    34.2796   3.494 0.000492 ***
## sexomulher          496.7690   244.3713   2.033 0.042267 *
## bmi30              -997.9355   422.9607  -2.359 0.018449 *
## fumantesim        13404.5952   439.9591  30.468  < 2e-16 ***
## regiaonorte         279.1661   349.2826   0.799 0.424285
## regiaosudeste      -942.9958   350.1754  -2.693 0.007172 **
## regiaosul          -548.8684   352.1950  -1.558 0.119372
## bmi30:fumantesim 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
```

```
# ** Multicollinearity

# The creation of age2 could lead to questions about multicollinearity. But what is multicollinearity?

# Multicollinearity is a common problem when estimating linear regression models, including logistic re
# occurs when there is a high correlation between the predictive variables, generating unreliable estim
# This phenomenon is certainly something that requires special attention from the Data Scientist, but i

# In our project, multicollinearity is not a problem. Multicollinearity needs to be verified and resolv
# independent effect of two variables that are correlated. In our case, we are not interested in assess
# regardless of age and age2. Whenever a study involving age is carried out, it is a good practice to i
# square to reduce the effect of age on the modeling process, because as we saw the relationship of the
# age (dependent) may not necessarily be linear. Multicollinearity will be present, but it will not aff
```

END