People Analytics - Regressão Logística

GitHub profile - @marcosmorais94

2023-02-01

1 - Introdução

Resolver o problema para analisar quais fatores influenciam na questão de conflitos Objetivo é identificar as relações com regressão logística e não prever se terá ou não.

Fonte dos dados - IBM Developer: https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/

```r
# Definindo diretório de trabalho
setwd("C:/FCD/R/people_analytics")
getwd()
```

```
## [1] "C:/FCD/R/people_analytics"
```

2 - Carga de Pacotes e Dados

You can also embed plots, for example:

```r
# Carga de pacotes
library(caret)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
## Carregando pacotes exigidos: lattice
```

```r
library(ggplot2)
library(gridExtra)
library(data.table)
library(car)
```

```
## Carregando pacotes exigidos: carData
```

```r
library(caTools)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(rpart)
library(rpart.plot)

# Carga dos dados
bd_rh <- read.csv('dados/people_data.csv')
```

3 - Informações sobre o dataset

```r
# Dimensões do dataset
dim(bd_rh) #23.058 linhas e 30 colunas
```

```
## [1] 23058    30
```

```r
# Tipos de dados
str(bd_rh)
```

```
## 'data.frame':    23058 obs. of  30 variables:
##  $ Age                     : int  41 37 41 37 37 37 41 41 41 41 ...
##  $ Attrition               : chr  "Voluntary Resignation" "Voluntary Resignation" "Voluntary Resignat
##  $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Travel_Rarely" "Travel_Rarely" ..
##  $ Department              : chr  "Sales" "Human Resources" "Sales" "Human Resources" ...
##  $ DistanceFromHome        : int  1 6 1 6 6 6 1 1 1 1 ...
##  $ Education               : int  2 4 2 4 4 4 2 2 2 2 ...
##  $ EducationField          : chr  "Life Sciences" "Human Resources" "Life Sciences" "Marketing" ...
##  $ EnvironmentSatisfaction : int  2 1 2 1 1 1 2 2 2 4 ...
##  $ Gender                  : chr  "Female" "Female" "Female" "Female" ...
##  $ JobInvolvement          : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ JobLevel                : int  2 2 2 2 2 2 2 2 2 4 ...
##  $ JobRole                 : chr  "Sales Executive" "Sales Executive" "Sales Executive" "Sales Execut
##  $ JobSatisfaction         : int  4 4 4 4 4 4 4 4 4 3 ...
##  $ MaritalStatus           : chr  "Single" "Single" "Single" "Single" ...
##  $ MonthlyIncome           : int  5993 5993 5993 5993 5993 5993 5993 5993 5993 14756 ...
##  $ NumCompaniesWorked      : int  8 8 4 5 8 5 8 4 8 2 ...
##  $ OverTime                : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PercentSalaryHike       : int  11 11 11 11 11 11 11 11 11 14 ...
##  $ PerformanceRating       : int  3 4 3 3 3 3 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  1 1 1 1 1 1 1 1 1 3 ...
##  $ StockOptionLevel        : int  0 0 0 0 0 0 0 0 0 3 ...
##  $ TotalWorkingYears       : int  8 8 8 8 8 8 8 8 8 21 ...
##  $ TrainingTimesLastYear   : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ WorkLifeBalance         : int  1 1 1 1 1 1 1 1 1 3 ...
##  $ YearsAtCompany          : int  6 6 6 6 6 6 6 6 6 5 ...
##  $ YearsInCurrentRole      : int  4 4 4 4 4 4 4 4 4 0 ...
##  $ YearsSinceLastPromotion : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ YearsWithCurrManager    : int  5 5 5 5 5 5 5 5 5 2 ...
##  $ Employee.Source         : chr  "Referral" "Referral" "Referral" "Referral" ...
##  $ AgeStartedWorking       : int  33 29 33 29 29 29 33 33 33 20 ...
```

```r
# Resumo estatístico
summary(bd_rh)
```

```
##       Age         Attrition          BusinessTravel      Department
##  Min.   :18.00   Length:23058       Length:23058       Length:23058
##  1st Qu.:30.00   Class :character   Class :character   Class :character
##  Median :36.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :37.04
##  3rd Qu.:43.00
##  Max.   :60.00
##  DistanceFromHome   Education     EducationField     EnvironmentSatisfaction
##  Min.   : 1.000   Min.   :1.000   Length:23058       Min.   :1.00
##  1st Qu.: 2.000   1st Qu.:2.000   Class :character   1st Qu.:2.00
##  Median : 7.000   Median :3.000   Mode  :character   Median :3.00
```

```
##  Mean   : 9.215   Mean   :2.915                      Mean   :2.72
##  3rd Qu.:14.000   3rd Qu.:4.000                      3rd Qu.:4.00
##  Max.   :29.000   Max.   :5.000                      Max.   :4.00
##     Gender          JobInvolvement    JobLevel       JobRole
##  Length:23058       Min.   :1.00   Min.   :1.000   Length:23058
##  Class :character   1st Qu.:2.00   1st Qu.:1.000   Class :character
##  Mode  :character   Median :3.00   Median :2.000   Mode  :character
##                     Mean   :2.73   Mean   :2.044
##                     3rd Qu.:3.00   3rd Qu.:3.000
##                     Max.   :4.00   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome   NumCompaniesWorked
##  Min.   :1.000   Length:23058       Min.   : 1009   Min.   :0.000
##  1st Qu.:2.000   Class :character   1st Qu.: 2900   1st Qu.:1.000
##  Median :3.000   Mode  :character   Median : 4898   Median :2.000
##  Mean   :2.725                      Mean   : 6416   Mean   :2.691
##  3rd Qu.:4.000                      3rd Qu.: 8120   3rd Qu.:4.000
##  Max.   :4.000                      Max.   :19999   Max.   :9.000
##    OverTime         PercentSalaryHike PerformanceRating
##  Length:23058       Min.   :11.00     Min.   :3.000
##  Class :character   1st Qu.:12.00     1st Qu.:3.000
##  Mode  :character   Median :14.00     Median :3.000
##                     Mean   :15.22     Mean   :3.155
##                     3rd Qu.:18.00     3rd Qu.:3.000
##                     Max.   :25.00     Max.   :4.000
##  RelationshipSatisfaction StockOptionLevel TotalWorkingYears
##  Min.   :1.000            Min.   :0.0000   Min.   : 0.00
##  1st Qu.:2.000            1st Qu.:0.0000   1st Qu.: 6.00
##  Median :3.000            Median :1.0000   Median :10.00
##  Mean   :2.713            Mean   :0.7944   Mean   :11.07
##  3rd Qu.:4.000            3rd Qu.:1.0000   3rd Qu.:15.00
##  Max.   :4.000            Max.   :3.0000   Max.   :40.00
##  TrainingTimesLastYear WorkLifeBalance YearsAtCompany  YearsInCurrentRole
##  Min.   :0.000         Min.   :1.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.00   1st Qu.: 2.000
##  Median :3.000         Median :3.000   Median : 5.00   Median : 3.000
##  Mean   :2.804         Mean   :2.762   Mean   : 6.91   Mean   : 4.201
##  3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.00   3rd Qu.: 7.000
##  Max.   :6.000         Max.   :4.000   Max.   :40.00   Max.   :18.000
##  YearsSinceLastPromotion YearsWithCurrManager Employee.Source
##  Min.   : 0.000          Min.   : 0.000       Length:23058
##  1st Qu.: 0.000          1st Qu.: 2.000       Class :character
##  Median : 1.000          Median : 3.000       Mode  :character
##  Mean   : 2.164          Mean   : 4.091
##  3rd Qu.: 3.000          3rd Qu.: 7.000
##  Max.   :15.000          Max.   :17.000
##  AgeStartedWorking
##  Min.   : 0.00
##  1st Qu.:20.00
##  Median :25.00
##  Mean   :25.96
##  3rd Qu.:31.00
##  Max.   :60.00
```

```r
# Visualização do dataset
View(bd_rh)
```

4 - Limpeza e Pré-Processmento

```r
# Classifica os atributos como tipo categórico
bd_rh$Attrition <- as.factor(bd_rh$Attrition)
bd_rh$BusinessTravel <- as.factor(bd_rh$BusinessTravel)
bd_rh$Department <- as.factor(bd_rh$Department)
bd_rh$Education <- as.factor(bd_rh$Education)
bd_rh$EducationField <- as.factor(bd_rh$EducationField)
bd_rh$Employee.Source <- as.factor(bd_rh$Employee.Source)
bd_rh$EnvironmentSatisfaction <- as.factor(bd_rh$EnvironmentSatisfaction)
bd_rh$Gender <- as.factor(bd_rh$Gender)
bd_rh$JobInvolvement <- as.factor(bd_rh$JobInvolvement)
bd_rh$JobLevel <- as.factor(bd_rh$JobLevel)
bd_rh$JobRole <- as.factor(bd_rh$JobRole)
bd_rh$JobSatisfaction <- as.factor(bd_rh$JobSatisfaction)
bd_rh$MaritalStatus <- as.factor(bd_rh$MaritalStatus)
bd_rh$OverTime <- as.factor(bd_rh$OverTime)
bd_rh$PerformanceRating <- as.factor(bd_rh$PerformanceRating)
bd_rh$RelationshipSatisfaction <- as.factor(bd_rh$RelationshipSatisfaction)
bd_rh$StockOptionLevel <- as.factor(bd_rh$StockOptionLevel)
bd_rh$WorkLifeBalance <- as.factor(bd_rh$WorkLifeBalance)

# Confirma se os dados estão como categóricos
str(bd_rh)
```

```
## 'data.frame':    23058 obs. of  30 variables:
##  $ Age                     : int  41 37 41 37 37 37 41 41 41 41 ...
##  $ Attrition               : Factor w/ 3 levels "Current employee",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 3 3 3 3 3 3
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 1 3 1 1 1 3 3 3 3 ...
##  $ DistanceFromHome        : int  1 6 1 6 6 6 1 1 1 1 ...
##  $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 4 2 4 4 4 2 2 2 2 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 1 2 3 1 3 2 2 2 2 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 1 2 1 1 1 2 2 2 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 3 3 ...
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 2 2 2 2 2 2 2 4 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 8 8 8 8 8 8 8 8 4
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 4 4 4 4 4 4 4 4 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ MonthlyIncome           : int  5993 5993 5993 5993 5993 5993 5993 5993 5993 14756 ...
##  $ NumCompaniesWorked      : int  8 8 4 5 8 5 8 4 8 2 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PercentSalaryHike       : int  11 11 11 11 11 11 11 11 11 14 ...
##  $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 1 1 1 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
##  $ StockOptionLevel        : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 4 ...
##  $ TotalWorkingYears       : int  8 8 8 8 8 8 8 8 8 21 ...
##  $ TrainingTimesLastYear   : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
```

```
##  $ YearsAtCompany          : int  6 6 6 6 6 6 6 6 6 5 ...
##  $ YearsInCurrentRole      : int  4 4 4 4 4 4 4 4 4 0 ...
##  $ YearsSinceLastPromotion : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ YearsWithCurrManager    : int  5 5 5 5 5 5 5 5 5 2 ...
##  $ Employee.Source         : Factor w/ 9 levels "Adzuna","Company Website",..: 8 8 8 8 8 8 8 8 8 2 .
##  $ AgeStartedWorking       : int  33 29 33 29 29 29 33 33 33 20 ...
```

```
# Drop dos níveis de fatores com 0 count
dados <- droplevels(bd_rh)
str(bd_rh)
```

```
## 'data.frame':    23058 obs. of  30 variables:
##  $ Age                     : int  41 37 41 37 37 37 41 41 41 41 ...
##  $ Attrition               : Factor w/ 3 levels "Current employee",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 3 3 3 3 3 3
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 3 1 3 1 1 1 3 3 3 3 ...
##  $ DistanceFromHome        : int  1 6 1 6 6 6 1 1 1 1 ...
##  $ Education               : Factor w/ 5 levels "1","2","3","4",..: 2 4 2 4 4 4 2 2 2 2 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 1 2 3 1 3 2 2 2 2 ...
##  $ EnvironmentSatisfaction : Factor w/ 4 levels "1","2","3","4": 2 1 2 1 1 1 2 2 2 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ JobInvolvement          : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 3 3 ...
##  $ JobLevel                : Factor w/ 5 levels "1","2","3","4",..: 2 2 2 2 2 2 2 2 2 4 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 8 8 8 8 8 8 8 8 8 4
##  $ JobSatisfaction         : Factor w/ 4 levels "1","2","3","4": 4 4 4 4 4 4 4 4 4 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ MonthlyIncome           : int  5993 5993 5993 5993 5993 5993 5993 5993 5993 14756 ...
##  $ NumCompaniesWorked      : int  8 8 4 5 8 5 8 4 8 2 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ PercentSalaryHike       : int  11 11 11 11 11 11 11 11 11 14 ...
##  $ PerformanceRating       : Factor w/ 2 levels "3","4": 1 2 1 1 1 1 1 1 1 1 ...
##  $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
##  $ StockOptionLevel        : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 4 ...
##  $ TotalWorkingYears       : int  8 8 8 8 8 8 8 8 8 21 ...
##  $ TrainingTimesLastYear   : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ WorkLifeBalance         : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
##  $ YearsAtCompany          : int  6 6 6 6 6 6 6 6 6 5 ...
##  $ YearsInCurrentRole      : int  4 4 4 4 4 4 4 4 4 0 ...
##  $ YearsSinceLastPromotion : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ YearsWithCurrManager    : int  5 5 5 5 5 5 5 5 5 2 ...
##  $ Employee.Source         : Factor w/ 9 levels "Adzuna","Company Website",..: 8 8 8 8 8 8 8 8 8 2 .
##  $ AgeStartedWorking       : int  33 29 33 29 29 29 33 33 33 20 ...
```

```
summary(bd_rh)
```

```
##       Age                         Attrition               BusinessTravel
##  Min.   :18.00   Current employee     :19370   Non-Travel       : 2344
##  1st Qu.:30.00   Termination          :   87   Travel_Frequently: 4378
##  Median :36.00   Voluntary Resignation: 3601   Travel_Rarely    :16336
##  Mean   :37.04
##  3rd Qu.:43.00
##  Max.   :60.00
##
##                  Department    DistanceFromHome Education
##  Human Resources       : 1010   Min.   : 1.000   1:2659
```

```
##  Research & Development:15040   1st Qu.: 2.000   2:4436
##  Sales                 : 7008   Median : 7.000   3:8930
##                                 Mean   : 9.215   4:6279
##                                 3rd Qu.:14.000   5: 754
##                                 Max.   :29.000
##
##           EducationField EnvironmentSatisfaction    Gender      JobInvolvement
##  Human Resources : 442   1:4490                  Female: 9205   1: 1287
##  Life Sciences   :9513   2:4476                  Male  :13853   2: 5888
##  Marketing       :2484   3:7091                                 3:13644
##  Medical         :7267   4:7001                                 4: 2239
##  Other           :1291
##  Technical Degree:2061
##
##  JobLevel                      JobRole     JobSatisfaction  MaritalStatus
##  1:8594   Sales Executive         :5067   1:4575          Divorced: 5163
##  2:8448   Research Scientist      :4591   2:4371          Married :10543
##  3:3440   Laboratory Technician   :4112   3:6938          Single  : 7352
##  4:1563   Manufacturing Director  :2346   4:7174
##  5:1013   Healthcare Representative:2069
##           Manager                 :1521
##           (Other)                 :3352
##  MonthlyIncome   NumCompaniesWorked OverTime    PercentSalaryHike
##  Min.   : 1009   Min.   :0.000      No :16524   Min.   :11.00
##  1st Qu.: 2900   1st Qu.:1.000      Yes: 6534   1st Qu.:12.00
##  Median : 4898   Median :2.000                  Median :14.00
##  Mean   : 6416   Mean   :2.691                  Mean   :15.22
##  3rd Qu.: 8120   3rd Qu.:4.000                  3rd Qu.:18.00
##  Max.   :19999   Max.   :9.000                  Max.   :25.00
##
##  PerformanceRating RelationshipSatisfaction StockOptionLevel TotalWorkingYears
##  3:19478           1:4331                   0:9873           Min.   : 0.00
##  4: 3580           2:4762                   1:9370           1st Qu.: 6.00
##                    3:7164                   2:2497           Median :10.00
##                    4:6801                   3:1318           Mean   :11.07
##                                                              3rd Qu.:15.00
##                                                              Max.   :40.00
##
##  TrainingTimesLastYear WorkLifeBalance YearsAtCompany  YearsInCurrentRole
##  Min.   :0.000         1: 1263         Min.   : 0.00   Min.   : 0.000
##  1st Qu.:2.000         2: 5374         1st Qu.: 3.00   1st Qu.: 2.000
##  Median :3.000         3:14016         Median : 5.00   Median : 3.000
##  Mean   :2.804         4: 2405         Mean   : 6.91   Mean   : 4.201
##  3rd Qu.:3.000                         3rd Qu.: 9.00   3rd Qu.: 7.000
##  Max.   :6.000                         Max.   :40.00   Max.   :18.000
##
##  YearsSinceLastPromotion YearsWithCurrManager        Employee.Source
##  Min.   : 0.000          Min.   : 0.000       Company Website:5327
##  1st Qu.: 0.000          1st Qu.: 2.000       Seek           :3655
##  Median : 1.000          Median : 3.000       Indeed         :2471
##  Mean   : 2.164          Mean   : 4.091       Jora           :2408
##  3rd Qu.: 3.000          3rd Qu.: 7.000       LinkedIn       :2294
##  Max.   :15.000          Max.   :17.000       Recruit.net    :2283
##                                               (Other)        :4620
```

```
##  AgeStartedWorking
##  Min.   : 0.00
##  1st Qu.:20.00
##  Median :25.00
##  Mean   :25.96
##  3rd Qu.:31.00
##  Max.   :60.00
##
```

```
View(bd_rh)
```

5 - Engenharia de Atributos

Nesta etapa vamos incluir alguns atributos que não foram identificadas na base original. Contudo, são informações que podemos incluir a partir do dataset original.

```
# Pior Year of Experience siginifica quantos anos o profissional tem de experiência profissional
bd_rh$PriorYearsOfExperience <- bd_rh$TotalWorkingYears - bd_rh$YearsAtCompany
View(bd_rh)

#Average Tenure é a estabilidade média do profissional no mesmo emprego
bd_rh$AverageTenure <- bd_rh$PriorYearsOfExperience / bd_rh$NumCompaniesWorked
View(bd_rh)

# A Average Tenure produz valores como Inf devido à natureza de sua derivação
# É possível identificar esse valores pelo summary, neste caso na média
summary(bd_rh$AverageTenure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       0       1     Inf       4     Inf     372
```

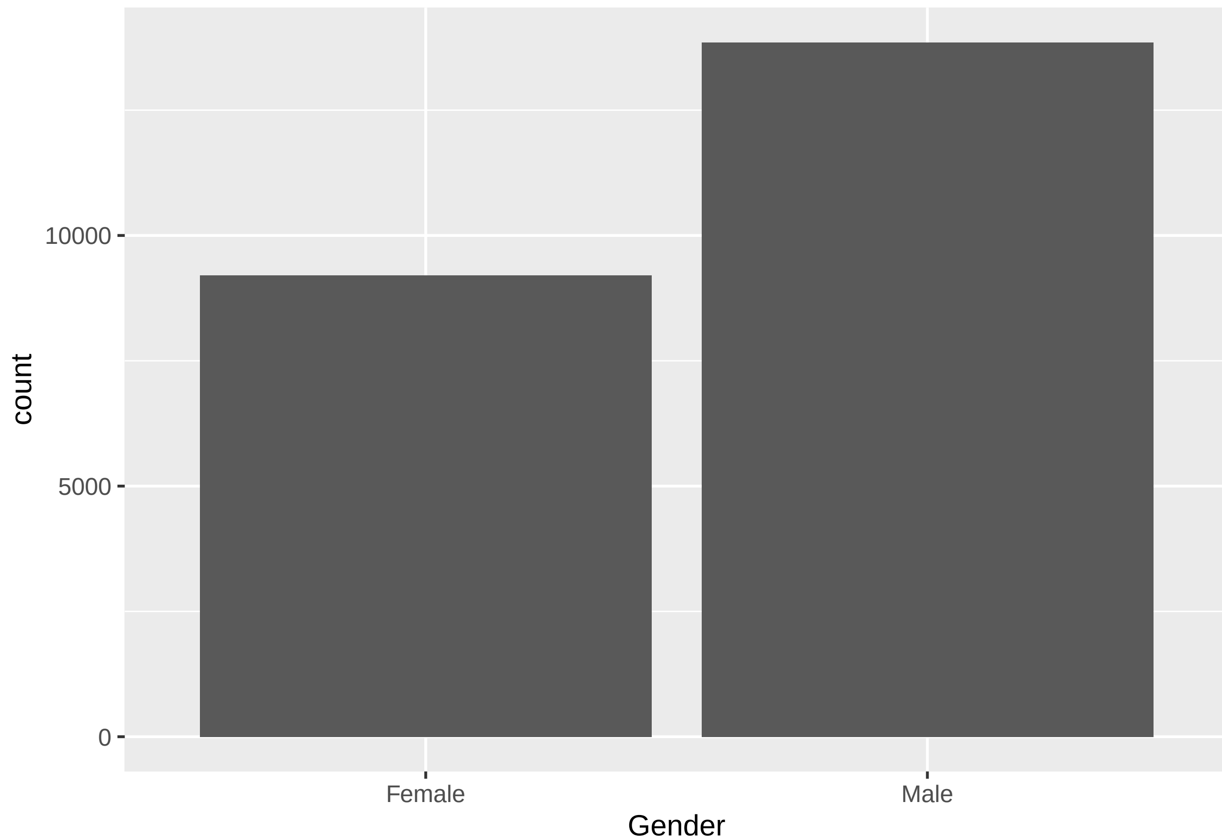
```
# Substituímos para zero, onde tudo que for contrário de finito será igualado a 0.
bd_rh$AverageTenure[!is.finite(bd_rh$AverageTenure)] <- 0

# Confere se ainda há valores Inf
summary(bd_rh$AverageTenure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.3333  1.7725  1.5000 40.0000
```

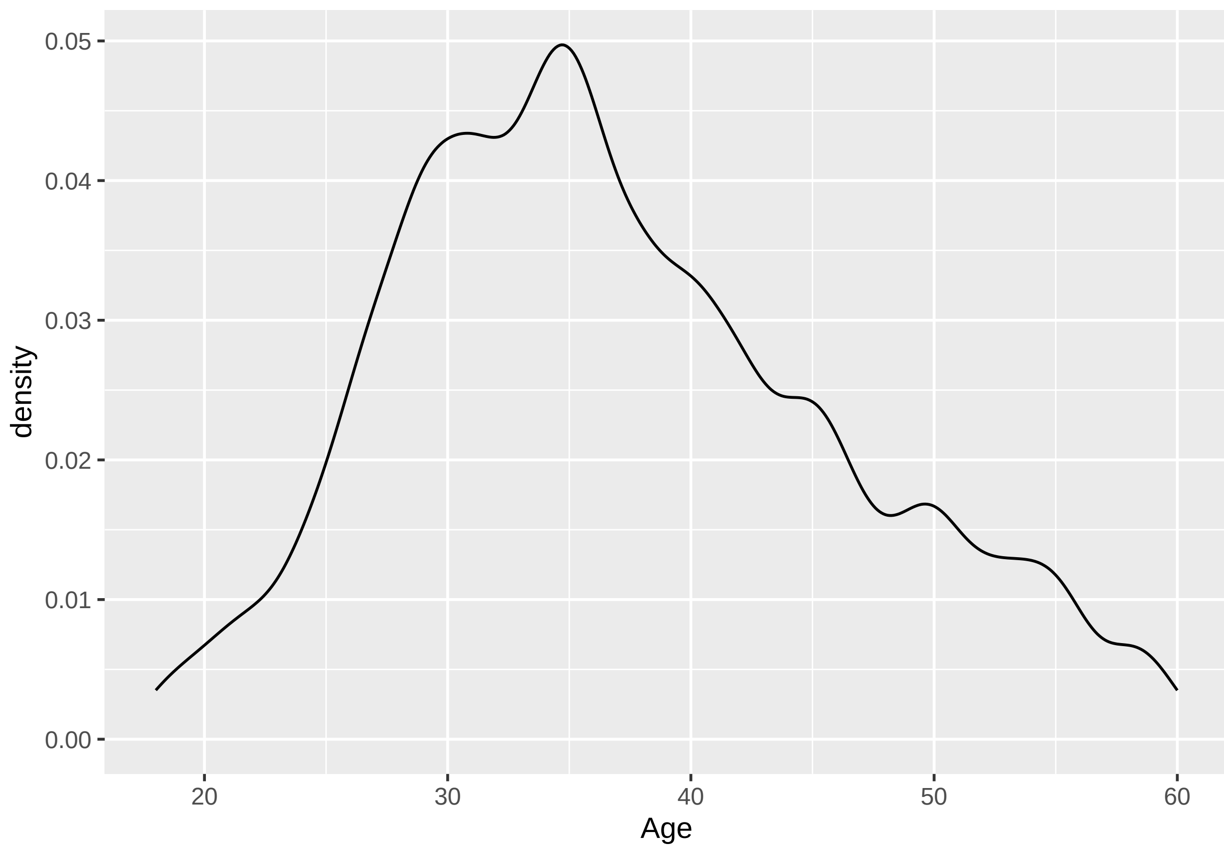
```
View(bd_rh)
```

6 - Análise Exploratória

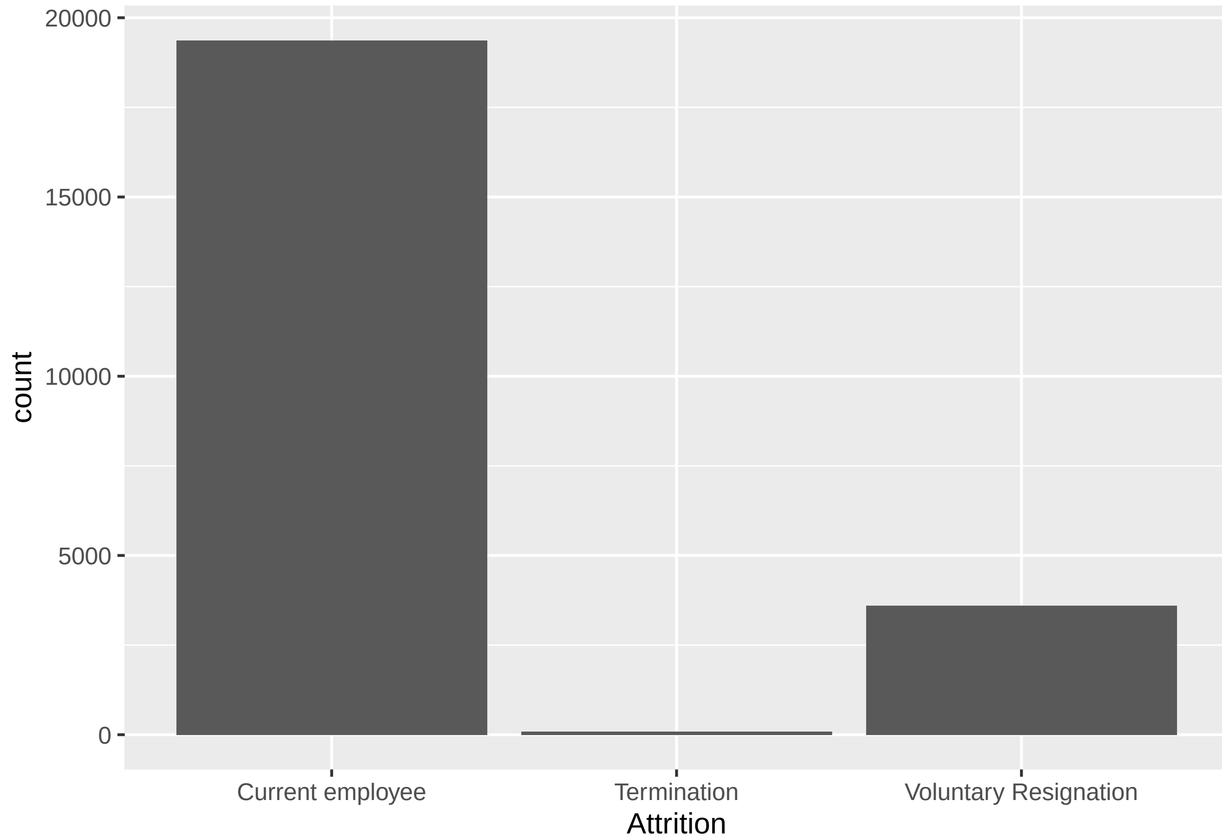
```
# Plots de análise univariada

# Contagem por genêro
# Aqui vemos que a base de dados temos mais homens que mulheres na base
ggplot(bd_rh) + geom_bar(aes(x = Gender))
```



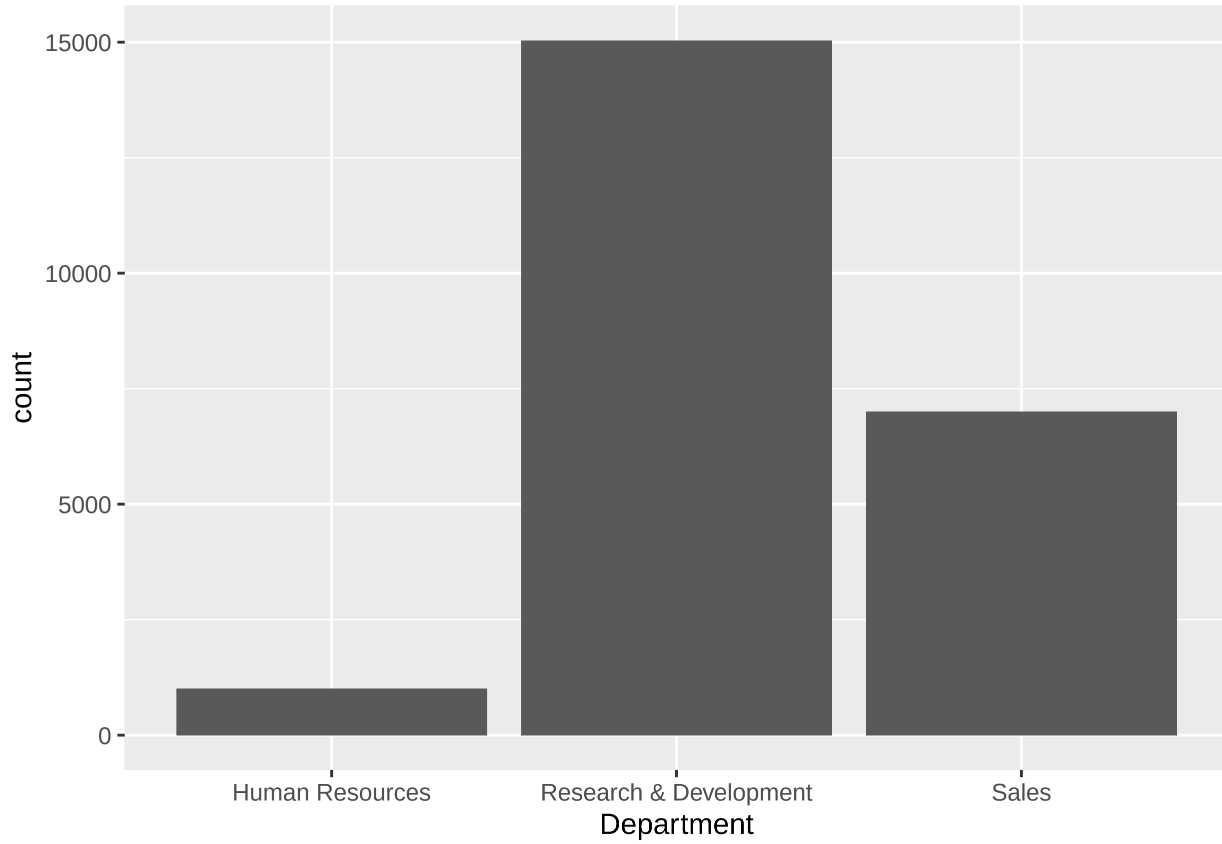

```r
# Idade dos profissionais da IBM
# A idade é em torno de 30 a 35 anos em suas grande maioria,
# indica que temos um boa parte dos profissionais com uma idade média
ggplot(bd_rh) + geom_density(aes(x = Age))
```



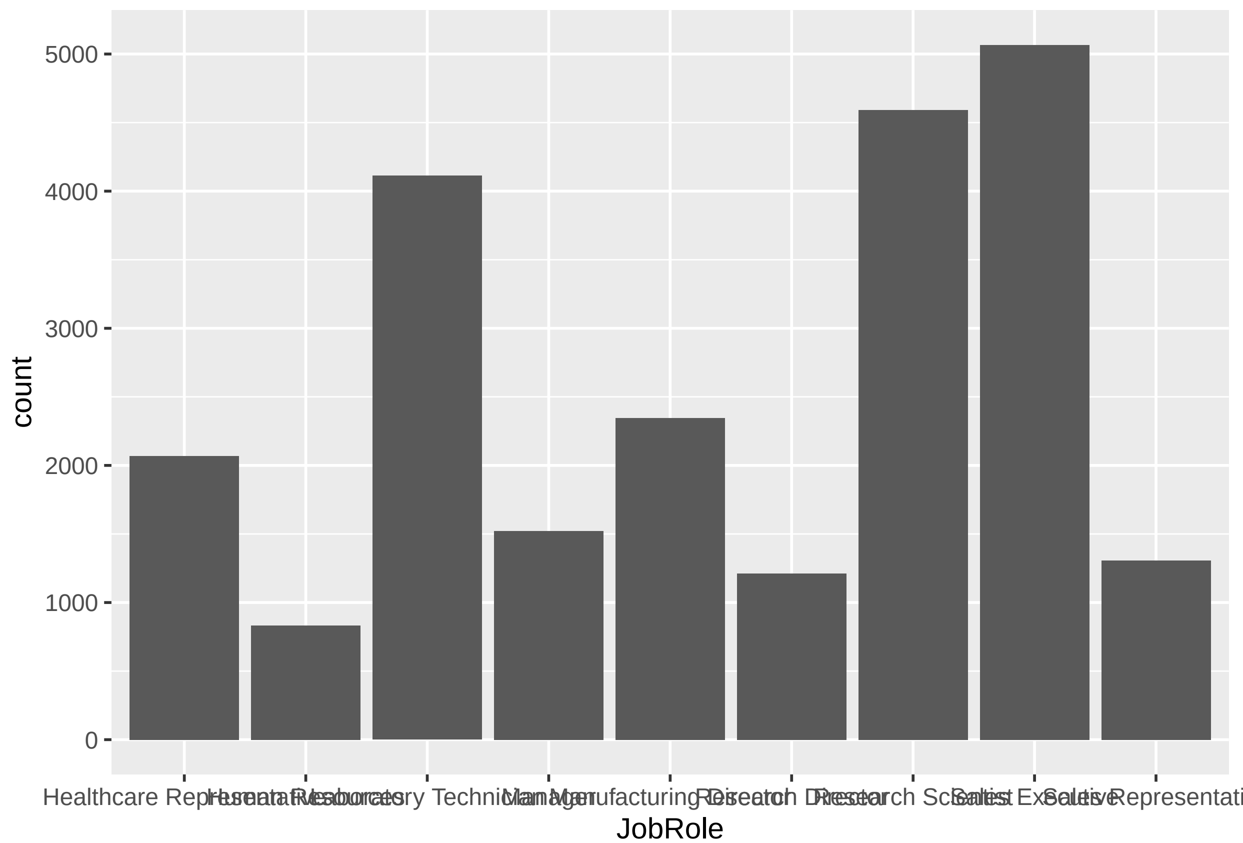

```
# Situação atual dos profissionais da base de dados
# A grande maioria continua empregado, uma boa parcela escolheu a demissão voluntária
# Essa fatia da demissão voluntária pode ter respostas interessantes do motivo da saída dos profissinai
# A menor parcela são os demitidos, que pode ter insights interessantes também relacionados aos motivos
ggplot(bd_rh) + geom_bar(aes(x = Attrition))
```



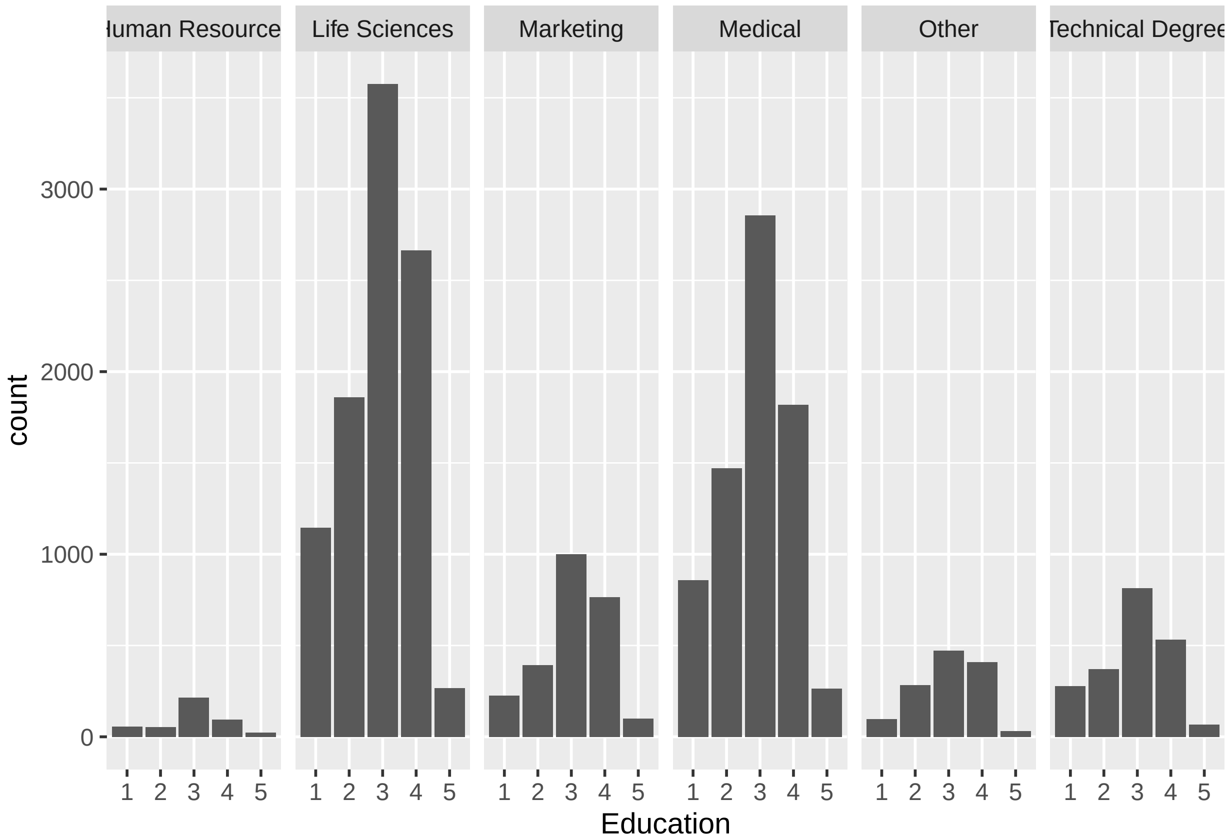

```r
# Contagem por Departamento
# Neste gráfico vemos que a maioria dos profissionais pertecem a área de pesquisa e desenvolvimento
# O que indica que temos na análise uma área que tende a ser muito estressante dentro da empresa
# Historicamente a área de Pesquisa e Desenvolvimento é muito cobrada por resultados
ggplot(bd_rh) + geom_bar(aes(x = Department))
```



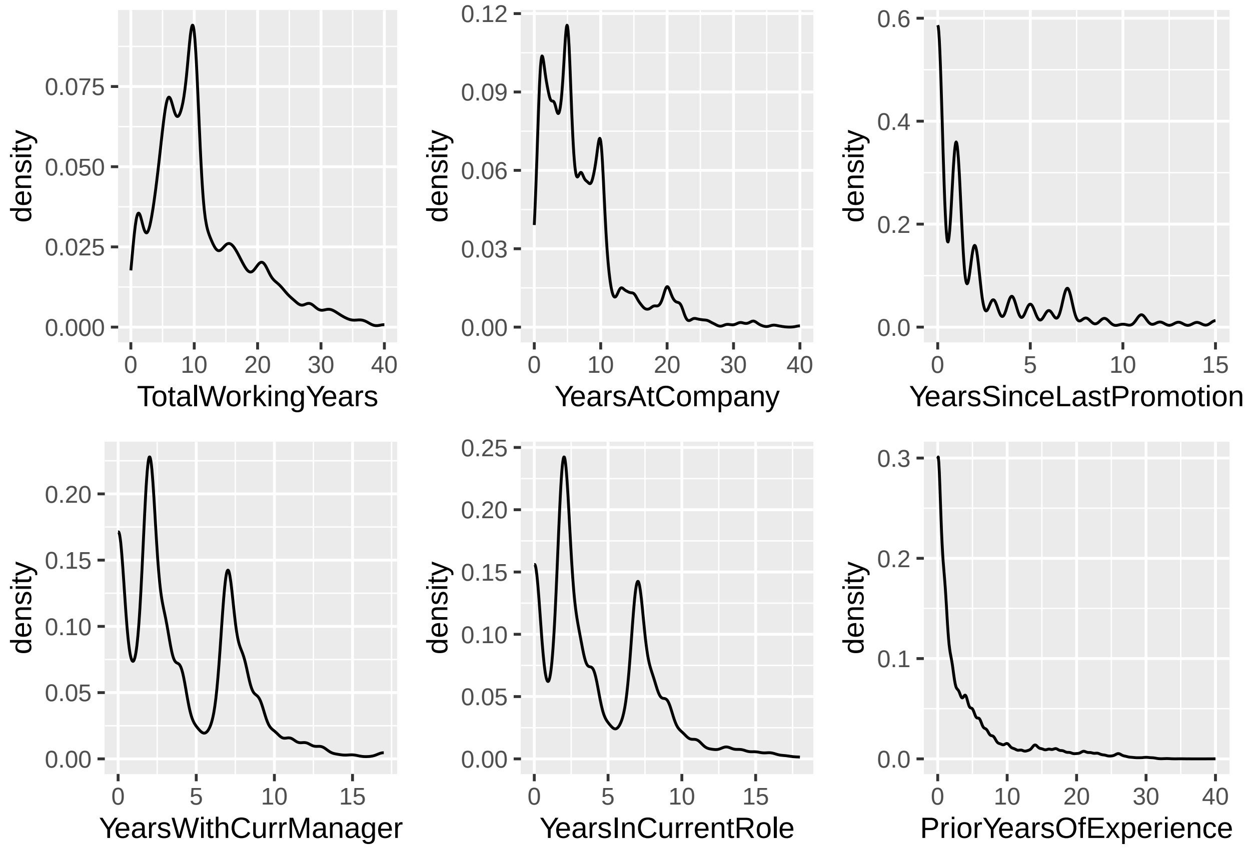

```
# Contagem por Cargo
# A maioria dos profissionais é executivo de vendas seguido de pesquisadores
# Geralmente são cargos que possuem muita cobranças e prazos curtos de entrega, o que implica em stress
ggplot(bd_rh) + geom_bar(aes(x = JobRole))
```




```
# Análise por Formação Acadêmica
# Neste gráfico vemos que a formação de Life Sciences, que engloba biotecnologia por exemplo, é a maior
# Outro ponto importante é que a área médica tem muitos representantes na base e formação técnica possu
ggplot(bd_rh) + geom_bar(aes(x = Education)) + facet_grid(~EducationField)
```




```r
# Multiplot Grid
# Aqui vamos plotar uma série de gráfico sobre atributos relacionados ao tempo,
# como anos de experiência e tempo com o mesmo gestor
p.TotalWorkingYears        <- ggplot(bd_rh) + geom_density(aes(TotalWorkingYears))
p.YearsAtCompany           <- ggplot(bd_rh) + geom_density(aes(YearsAtCompany))
p.YearsSinceLastPromotion <- ggplot(bd_rh) + geom_density(aes(YearsSinceLastPromotion))
p.YearsWithCurrManager     <- ggplot(bd_rh) + geom_density(aes(YearsWithCurrManager))
p.YearsInCurrentRole       <- ggplot(bd_rh) + geom_density(aes(YearsInCurrentRole))
p.PriorYearsOfExperience   <- ggplot(bd_rh) + geom_density(aes(PriorYearsOfExperience))


# Organiza no grid
grid.arrange(p.TotalWorkingYears,
             p.YearsAtCompany,
             p.YearsSinceLastPromotion,
             p.YearsWithCurrManager,
             p.YearsInCurrentRole,
             p.PriorYearsOfExperience,
             nrow = 2,
             ncol = 3)
```




```r
# Alguns dados interessantes são que a medida que os gestores mudam, os cargo dos profissionais também m
# Outro detalhe que chama a atenção são que temos um pico na casa dos 10 anos de trabalho na empresa, ap
# esse tempo vemos uma queda que se mantém ao longo do período.

# Tempo de experiência anterior
# Vamos descobrir a proporção de funcionários com menos de alguns anos de experiência
# (valores escolhidos: 1, 3, 5, 7, 10 anos)
length(which(bd_rh$PriorYearsOfExperience < 1)) / length(bd_rh$PriorYearsOfExperience)
```

```
## [1] 0.3246596
```

```r
length(which(bd_rh$PriorYearsOfExperience < 3)) / length(bd_rh$PriorYearsOfExperience)
```

```
## [1] 0.5828346
```

```r
length(which(bd_rh$PriorYearsOfExperience < 5)) / length(bd_rh$PriorYearsOfExperience)
```

```
## [1] 0.7085177
```

```r
length(which(bd_rh$PriorYearsOfExperience < 7)) / length(bd_rh$PriorYearsOfExperience)
```

```
## [1] 0.7952121
```

```r
length(which(bd_rh$PriorYearsOfExperience < 10)) / length(bd_rh$PriorYearsOfExperience)
```

```
## [1] 0.8589644
```

```r
# 58% dos funcionários têm menos de 3 anos de experiência de trabalho antes de entrar na IBM
# Possíveis problemas: conjuntos de habilidades subdesenvolvidos, base de jovens funcionários,
# mentalidade de "trabalho" imatura.
```

```r
# Idade
# Apenas 22% dos funcionários têm menos de 30 anos, a base de funcionários não é exatamente
# tão jovem como o esperado.
length(which(bd_rh$Age < 30)) / length(bd_rh$Age)
```

```
## [1] 0.2165409
```

```r
# # Educação
summary(bd_rh$Education)
```

```
##    1    2    3    4    5
## 2659 4436 8930 6279  754
```

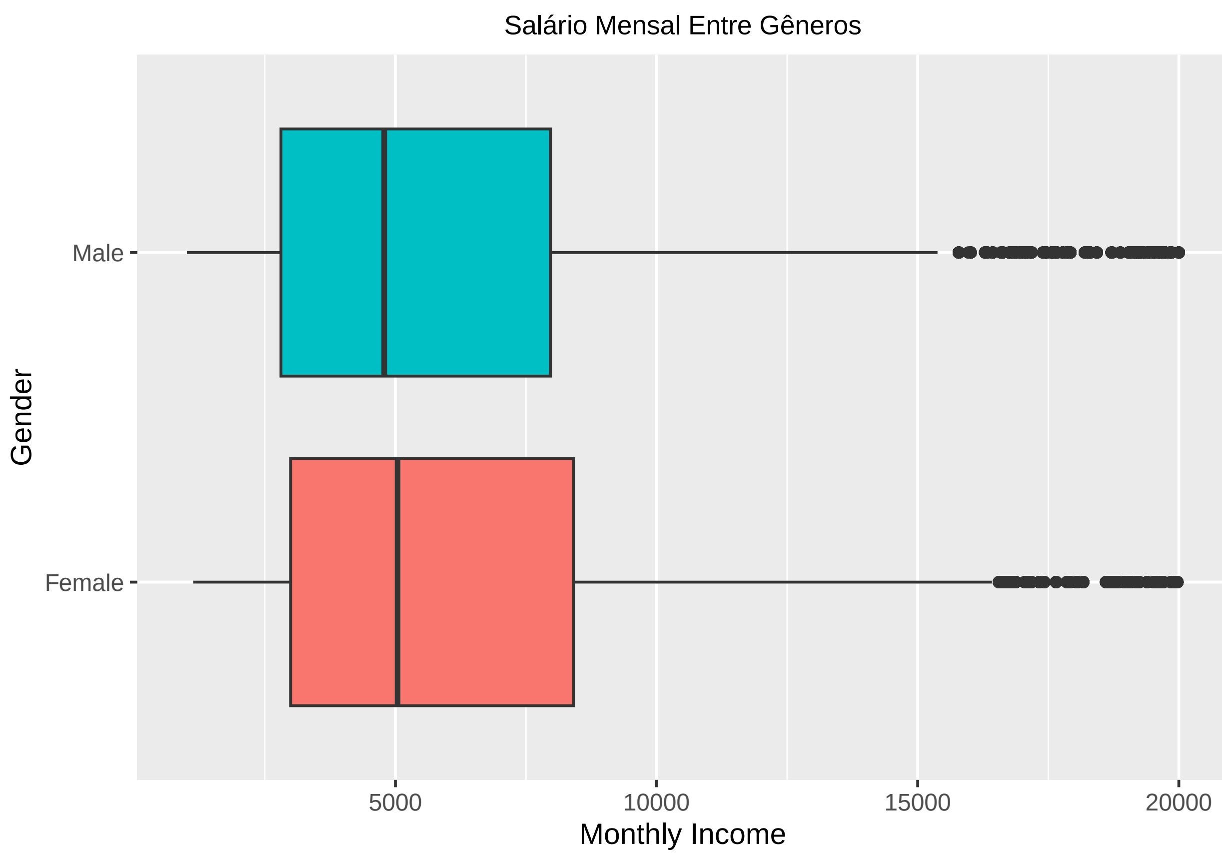
```r
length(which(bd_rh$Education == 3)) / length(bd_rh$Education)
```

```
## [1] 0.3872842
```

```r
length(which(bd_rh$Education == 4)) / length(bd_rh$Education)
```

```
## [1] 0.2723133
```

```r
# Cerca de 39% dos funcionários são graduados e 27% realizaram o mestrado.
# A busca pelo ensino superior pode ter levado a uma diminuição da experiência de trabalho.

# Verificando a diferença salarial entre homens e mulheres.
ggplot(data = subset(bd_rh, !is.na(Gender)), aes(Gender, MonthlyIncome, fill = Gender)) +
  geom_boxplot() +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5, size = 10)) +
  labs(x = "Gender", y = "Monthly Income", title = "Salário Mensal Entre Gêneros") +
  coord_flip()
```




```
# As mulheres ganham um pouco mais, em média, desconsiderando todos os outros fatores.
```

7 - Modelagem Preditiva

Objetivo inicial é criar 4 versões do modelo preditivo com o algoritmo de Regressão Logística.

7.1 - Modelo v1

```
# Primeira versão do modelo com algumas variáveis
# Esse modelo é como um balizador sem a divisão de treino e teste
modelo_v1 <- glm(Attrition ~ Age + Department + DistanceFromHome + Employee.Source +
                 JobRole + MaritalStatus + AverageTenure + PriorYearsOfExperience +
                 family = binomial,
                 data = bd_rh)

summary(modelo_v1)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Department + DistanceFromHome +
##     Employee.Source + JobRole + MaritalStatus + AverageTenure +
##     PriorYearsOfExperience + Gender + Education + EducationField,
##     family = binomial, data = bd_rh)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.4738  -0.6239  -0.4962  -0.3553   2.7405
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.515415   0.198808  -2.593 0.009527 **
## Age                             -0.046402   0.002434 -19.062  < 2e-16 ***
## DepartmentResearch & Development -0.402413   0.102837  -3.913 9.11e-05 ***
## DepartmentSales                  0.041108   0.106275   0.387 0.698901
## DistanceFromHome                 0.022014   0.002497   8.816  < 2e-16 ***
## Employee.SourceCompany Website   0.200175   0.074567   2.684 0.007264 **
## Employee.SourceGlassDoor        -0.002062   0.089568  -0.023 0.981630
## Employee.SourceIndeed           -0.048126   0.088966  -0.541 0.588545
## Employee.SourceJora              0.202494   0.084534   2.395 0.016602 *
## Employee.SourceLinkedIn         -0.086527   0.090292  -0.958 0.337911
## Employee.SourceRecruit.net      -0.024145   0.088800  -0.272 0.785699
## Employee.SourceReferral          0.222132   0.147177   1.509 0.131226
## Employee.SourceSeek              0.039192   0.079096   0.495 0.620253
## JobRoleHuman Resources           0.092163   0.125250   0.736 0.461832
## JobRoleLaboratory Technician     0.313456   0.079749   3.931 8.48e-05 ***
## JobRoleManager                  -0.370055   0.121400  -3.048 0.002302 **
## JobRoleManufacturing Director   -0.091942   0.094178  -0.976 0.328937
## JobRoleResearch Director        -0.326907   0.125855  -2.597 0.009391 **
## JobRoleResearch Scientist        0.102218   0.078537   1.302 0.193080
## JobRoleSales Executive          -0.030434   0.079097  -0.385 0.700414
## JobRoleSales Representative      0.484732   0.095181   5.093 3.53e-07 ***
## MaritalStatusMarried             0.179376   0.053279   3.367 0.000761 ***
## MaritalStatusSingle              0.740422   0.053393  13.867  < 2e-16 ***
## AverageTenure                   -0.016927   0.009230  -1.834 0.066663 .
## PriorYearsOfExperience           0.018901   0.005353   3.531 0.000414 ***
## GenderMale                       0.033768   0.038421   0.879 0.379467
## Education2                       0.096221   0.068965   1.395 0.162951
## Education3                       0.129656   0.061109   2.122 0.033862 *
## Education4                       0.120603   0.066456   1.815 0.069558 .
## Education5                      -0.221560   0.134302  -1.650 0.099001 .
## EducationFieldLife Sciences     -0.149802   0.143779  -1.042 0.297462
## EducationFieldMarketing         -0.122315   0.152984  -0.800 0.423984
## EducationFieldMedical           -0.176829   0.145066  -1.219 0.222859
## EducationFieldOther             -0.170949   0.161651  -1.058 0.290274
## EducationFieldTechnical Degree   0.183255   0.154276   1.188 0.234898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20272  on 23057  degrees of freedom
## Residual deviance: 18904  on 23023  degrees of freedom
## AIC: 18974
##
## Number of Fisher Scoring iterations: 5
```

```r
# Análise por VIF
# VIF é uma função determina a correlação entre variáveis
# Neste caso, analisamos qual variável impacta mais na variável preditora, neste caso o Atrittion
```

```r
# Aqui vemos que o JobRole, AverageTenure e PiorYearsofExperience são as variáveis mais influntes no mo
# Basicamente é o cargo, estabilidade média no mesmo emprego e anos de experiência anteriores
# Ou seja, o perfil é de um profissional mais senior e estavél no emprego
vif_modelo1 <- vif(modelo_v1)
View(vif_modelo1)
```

7.2 - Divisão de Dados em treino e Teste

```r
# Vamos dividir os dados em treino e teste.

# Vamos trabalhar com os dados sem registros de demitidos.
dados_rh_1 <- bd_rh[bd_rh$Attrition != 'Termination',]
dados_rh_1 <- droplevels(dados_rh_1)

# Divisão de treino e teste
index_treino <- sample.split(Y = dados_rh_1$Attrition, SplitRatio = 0.7)
dados_rh_1_treino <- subset(dados_rh_1, train = T)
dados_rh_1_teste <- subset(dados_rh_1, train = F)
```

7.3 - Modelo v2

```r
# Segunda versão do modelo com dados de treino
modelo_v2 <- glm(Attrition ~ Age + Department + DistanceFromHome + Employee.Source +
                   JobRole + MaritalStatus + AverageTenure + PriorYearsOfExperience + Gender +
                   Education + EducationField,
                 family = binomial,
                 data = dados_rh_1_treino)

summary(modelo_v2)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Department + DistanceFromHome +
##     Employee.Source + JobRole + MaritalStatus + AverageTenure +
##     PriorYearsOfExperience + Gender + Education + EducationField,
##     family = binomial, data = dados_rh_1_treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4484  -0.6177  -0.4918  -0.3558   2.7300
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.499751   0.199492  -2.505 0.012241 *
## Age                              -0.044889   0.002446 -18.348  < 2e-16 ***
## DepartmentResearch & Development -0.427955   0.103053  -4.153 3.28e-05 ***
## DepartmentSales                   0.025684   0.106499   0.241 0.809423
## DistanceFromHome                  0.020372   0.002522   8.076 6.69e-16 ***
## Employee.SourceCompany Website    0.183335   0.074868   2.449 0.014334 *
## Employee.SourceGlassDoor          0.006274   0.089680   0.070 0.944229
## Employee.SourceIndeed            -0.080908   0.089734  -0.902 0.367244
## Employee.SourceJora               0.183678   0.084958   2.162 0.030618 *
## Employee.SourceLinkedIn          -0.079145   0.090405  -0.875 0.381325
```

```
## Employee.SourceRecruit.net       -0.050665   0.089444  -0.566 0.571095
## Employee.SourceReferral           0.230121   0.147168   1.564 0.117897
## Employee.SourceSeek              -0.005837   0.079828  -0.073 0.941708
## JobRoleHuman Resources            0.107348   0.125753   0.854 0.393302
## JobRoleLaboratory Technician      0.314968   0.080707   3.903 9.52e-05 ***
## JobRoleManager                   -0.402633   0.123788  -3.253 0.001144 **
## JobRoleManufacturing Director    -0.083426   0.095273  -0.876 0.381221
## JobRoleResearch Director         -0.292195   0.126243  -2.315 0.020637 *
## JobRoleResearch Scientist         0.111877   0.079359   1.410 0.158608
## JobRoleSales Executive           -0.028140   0.079873  -0.352 0.724611
## JobRoleSales Representative       0.478077   0.096067   4.977 6.47e-07 ***
## MaritalStatusMarried              0.176289   0.053865   3.273 0.001065 **
## MaritalStatusSingle               0.747383   0.053896  13.867  < 2e-16 ***
## AverageTenure                    -0.021245   0.009467  -2.244 0.024825 *
## PriorYearsOfExperience            0.019787   0.005399   3.665 0.000248 ***
## GenderMale                        0.030982   0.038752   0.800 0.424000
## Education2                        0.067584   0.069195   0.977 0.328712
## Education3                        0.092553   0.061236   1.511 0.130684
## Education4                        0.071013   0.066760   1.064 0.287461
## Education5                       -0.233758   0.134267  -1.741 0.081685 .
## EducationFieldLife Sciences      -0.148858   0.143810  -1.035 0.300620
## EducationFieldMarketing          -0.106268   0.152995  -0.695 0.487317
## EducationFieldMedical            -0.202212   0.145203  -1.393 0.163736
## EducationFieldOther              -0.137807   0.161652  -0.852 0.393940
## EducationFieldTechnical Degree    0.180977   0.154552   1.171 0.241608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19951  on 22970  degrees of freedom
## Residual deviance: 18626  on 22936  degrees of freedom
## AIC: 18696
##
## Number of Fisher Scoring iterations: 5
```

```
# Análise VIF Modelo 2
# Os dados permanceram os mesmos que o modelo 1.
# A explicação é que removemos poucas linhas (removido pessoas demitidas)
# e mantivemos basicamente os mesmos atributos que o modelo anterior
vif_modelo2 <- vif(modelo_v2)
View(vif_modelo2)

# Previsões modelo 2
threshold <- 0.5
previsoes_v2 <- predict(modelo_v2, type = 'response', newdata = dados_rh_1_teste)
previsoes_finais_v2 <- ifelse(previsoes_v2 > threshold, 'Voluntary Resignation', 'Current employee')
table(dados_rh_1_teste$Attrition, previsoes_finais_v2)
```

```
##                        previsoes_finais_v2
##                         Current employee Voluntary Resignation
##   Current employee                 19328                    42
##   Voluntary Resignation             3523                    78
```

7.4 - Modelo v3

```r
# Terceira versão do modelo com dados de treino e sem variáveis de educação
modelo_v3 <- glm(Attrition ~ Age + Department + DistanceFromHome + Employee.Source +
                   JobRole + MaritalStatus + AverageTenure + PriorYearsOfExperience + Gender,
                 family = binomial,
                 data = dados_rh_1_treino)
summary(modelo_v3)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Department + DistanceFromHome +
##     Employee.Source + JobRole + MaritalStatus + AverageTenure +
##     PriorYearsOfExperience + Gender, family = binomial, data = dados_rh_1_treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3428  -0.6201  -0.4941  -0.3619   2.7143
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.594443   0.163302  -3.640 0.000272 ***
## Age                              -0.044338   0.002361 -18.781  < 2e-16 ***
## DepartmentResearch & Development -0.455831   0.097648  -4.668 3.04e-06 ***
## DepartmentSales                   0.006375   0.100798   0.063 0.949567
## DistanceFromHome                  0.023945   0.002219  10.792  < 2e-16 ***
## Employee.SourceCompany Website    0.185836   0.074684   2.488 0.012835 *
## Employee.SourceGlassDoor          0.004131   0.089469   0.046 0.963174
## Employee.SourceIndeed            -0.084488   0.089587  -0.943 0.345638
## Employee.SourceJora               0.182141   0.084629   2.152 0.031378 *
## Employee.SourceLinkedIn          -0.073833   0.090249  -0.818 0.413300
## Employee.SourceRecruit.net       -0.058670   0.089241  -0.657 0.510903
## Employee.SourceReferral           0.237922   0.146800   1.621 0.105078
## Employee.SourceSeek              -0.006818   0.079571  -0.086 0.931717
## JobRoleHuman Resources            0.099083   0.125594   0.789 0.430163
## JobRoleLaboratory Technician      0.312339   0.080556   3.877 0.000106 ***
## JobRoleManager                   -0.418085   0.123665  -3.381 0.000723 ***
## JobRoleManufacturing Director    -0.079696   0.095061  -0.838 0.401826
## JobRoleResearch Director         -0.308958   0.126075  -2.451 0.014263 *
## JobRoleResearch Scientist         0.119993   0.079265   1.514 0.130071
## JobRoleSales Executive           -0.023432   0.079774  -0.294 0.768961
## JobRoleSales Representative       0.483836   0.095952   5.042 4.60e-07 ***
## MaritalStatusMarried              0.176480   0.053793   3.281 0.001035 **
## MaritalStatusSingle               0.747665   0.053772  13.904  < 2e-16 ***
## AverageTenure                    -0.019906   0.009465  -2.103 0.035453 *
## PriorYearsOfExperience            0.019187   0.005400   3.553 0.000381 ***
## GenderMale                        0.033764   0.038690   0.873 0.382838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19951  on 22970  degrees of freedom
## Residual deviance: 18668  on 22945  degrees of freedom
```

```
## AIC: 18720
##
## Number of Fisher Scoring iterations: 5
```

```
# Análise VIF Modelo 3
# Os primeiros registros se manteram (obRole, AverageTenure e PiorYearsofExperience)
# Contudo tivemos uma queda em Departament ao remover Education do treinamento do modelo
vif_modelo3 <- vif(modelo_v3)
View(vif_modelo3)

# Previsões modelo 3
threshold <- 0.5
previsoes_v3 <- predict(modelo_v3, type = 'response', newdata = dados_rh_1_teste)
previsoes_finais_v3 <- ifelse(previsoes_v3 > threshold, 'Voluntary Resignation', 'Current employee')
table(dados_rh_1_teste$Attrition, previsoes_finais_v3)
```

```
##                        previsoes_finais_v3
##                         Current employee Voluntary Resignation
##   Current employee                 19328                    42
##   Voluntary Resignation             3541                    60
```

7.5 - Modelo v4

```
# Quarta versão do modelo com dados de treino e sem variáveis de educação e genero
modelo_v4 <- glm(Attrition ~ Age + Department + DistanceFromHome + Employee.Source +
                   JobRole + MaritalStatus + AverageTenure + PriorYearsOfExperience,
                 family = binomial,
                 data = dados_rh_1_treino)

summary(modelo_v4)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + Department + DistanceFromHome +
##     Employee.Source + JobRole + MaritalStatus + AverageTenure +
##     PriorYearsOfExperience, family = binomial, data = dados_rh_1_treino)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3360  -0.6192  -0.4939  -0.3622   2.7205
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -0.569968   0.160865  -3.543 0.000395 ***
## Age                               -0.044408   0.002359 -18.822  < 2e-16 ***
## DepartmentResearch & Development -0.457114   0.097648  -4.681 2.85e-06 ***
## DepartmentSales                    0.004776   0.100790   0.047 0.962208
## DistanceFromHome                   0.023979   0.002218  10.810  < 2e-16 ***
## Employee.SourceCompany Website     0.185968   0.074691   2.490 0.012780 *
## Employee.SourceGlassDoor           0.004217   0.089473   0.047 0.962404
## Employee.SourceIndeed             -0.082065   0.089543  -0.916 0.359412
## Employee.SourceJora                0.182210   0.084632   2.153 0.031321 *
## Employee.SourceLinkedIn           -0.073105   0.090254  -0.810 0.417948
## Employee.SourceRecruit.net        -0.058149   0.089234  -0.652 0.514631
## Employee.SourceReferral            0.240776   0.146746   1.641 0.100844
```

```
## Employee.SourceSeek              -0.006816   0.079577  -0.086 0.931742
## JobRoleHuman Resources            0.100479   0.125614   0.800 0.423769
## JobRoleLaboratory Technician      0.315123   0.080478   3.916 9.02e-05 ***
## JobRoleManager                   -0.419678   0.123673  -3.393 0.000690 ***
## JobRoleManufacturing Director    -0.082962   0.094978  -0.873 0.382397
## JobRoleResearch Director         -0.310452   0.126056  -2.463 0.013785 *
## JobRoleResearch Scientist         0.120223   0.079252   1.517 0.129277
## JobRoleSales Executive           -0.023015   0.079761  -0.289 0.772925
## JobRoleSales Representative       0.482258   0.095927   5.027 4.97e-07 ***
## MaritalStatusMarried              0.175136   0.053769   3.257 0.001125 **
## MaritalStatusSingle               0.745551   0.053714  13.880  < 2e-16 ***
## AverageTenure                    -0.019985   0.009465  -2.112 0.034727 *
## PriorYearsOfExperience            0.019266   0.005398   3.569 0.000358 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19951  on 22970  degrees of freedom
## Residual deviance: 18668  on 22946  degrees of freedom
## AIC: 18718
##
## Number of Fisher Scoring iterations: 5
```

```
# Análise VIF Modelo 4
# Pouca mudança em relação a V2 do modelo
vif_modelo4 <- vif(modelo_v4)
View(vif_modelo4)

# Previsões modelo 4
threshold <- 0.5
previsoes_v4 <- predict(modelo_v4, type = 'response', newdata = dados_rh_1_teste)
previsoes_finais_v4 <- ifelse(previsoes_v4 > threshold, 'Voluntary Resignation', 'Current employee')
table(dados_rh_1_teste$Attrition, previsoes_finais_v4)
```

```
##                        previsoes_finais_v4
##                         Current employee Voluntary Resignation
##   Current employee                 19326                    44
##   Voluntary Resignation             3545                    56
```

8 - Conclusão final

Com base nas informações, o modelo 2 teve um desempenho mais interessante na visão da análise de variáveis. É um proposta de solução, mas seria interessante balancear as classes e criar varíaveis dummy. Com essas etapas, o modelo teria um resultado mais assertivo, contudo o objetivo final de analisar quais atributos mais influenciam no modelo foi feito e isso demonstra as possibilidades com Machine Learning.