

Titanic - Kaggle

Marcos de Morais

19/09/2021

Prevendo os sobreviventes no desastre do Titanic

O objetivo desse estudo é prever se um determinado passageiro poderia ou não sobreviver com base em informações como idade, sexo, classe na embarcação e etc.

O desafio foi proposto pelo Kaggle, o dicionário dos dados é:

Variable Survival

0 = No, 1 = Yes

pclass (Ticket class) 1 = 1st, 2 = 2nd, 3 = 3rd

sex Male, Female

Age in years

sibsp # of siblings / spouses aboard the Titanic

parch # of parents / children aboard the Titanic

ticket

Ticket number

fare

Passenger fare

cabin Cabin number

embarked (Port of Embarkation)

C = Cherbourg, Q = Queenstown, S = Southampton

A fonte do dataset é: <https://www.kaggle.com/c/titanic/overview>

O que foi o desastre do Titanic?

O destino do Titanic foi selado em sua viagem inaugural de Southampton, na Inglaterra, à cidade de Nova York. Às 23h40 de 14 de abril de 1912, a lateral do Titanic colidiu com um iceberg no norte do Atlântico, afundando partes do casco do estibordo por uma extensão de quase 100 metros e expondo à água do mar os seis compartimentos dianteiros à prova d'água. A partir daquele instante, o naufrágio era inevitável.

Fonte: <https://www.nationalgeographicbrasil.com/historia/2019/08/como-foi-o-naufragio-e-redescoberta-do-titanic>

Etapa 1 - Coleta dos dados

A coleta dos dados foi com base nos arquivos disponibilizados pelo Kaggle. Foram 3 datasets, um para treino do modelo e outro para teste. Por último, temos o dataset para submissão das respostas no site.

```

# 1 - Carregar dados para Data Frame

#O dataset de treino com 891 observações
df <- read.csv('train.csv')

#O dataset de teste com 417 observações
df2 <- read.csv('test.csv')

#O dataset para registrar as previsões do modelo
submission <- read.csv('gender_submission.csv')

library(dplyr) #pacote para função bind_rows

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

#Dataset único para tratamento/analise dos dados inicial
full_df <- bind_rows(df,df2)
View(full_df)

# Dataset com os passageiros do Titanic, sendo que a coluna
# Survived é a variável target, onde 1 é sobrevivente e 0
# é morto no acidente

```

Etapa 2 - Análise dos Dados

Nesta etapa é possível fazer análises rápidas sobre o conjunto de dados bruto. Assim já fica claro que algumas variáveis precisam ser transformadas.

```

str(full_df)

## 'data.frame':    1309 obs. of  12 variables:
## $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
## $ Survived    : int  0 1 1 1 0 0 0 0 1 1 ...
## $ Pclass      : int  3 1 3 1 3 3 1 3 3 2 ...
## $ Name        : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" ...
## $ Sex         : chr  "male" "female" "female" "female" ...
## $ Age         : num  22 38 26 35 35 NA 54 2 27 14 ...
## $ SibSp       : int  1 1 0 1 0 0 0 3 0 1 ...
## $ Parch       : int  0 0 0 0 0 0 1 2 0 ...
## $ Ticket      : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
## $ Fare        : num  7.25 71.28 7.92 53.1 8.05 ...
## $ Cabin       : chr  "" "C85" "" "C123" ...
## $ Embarked    : chr  "S" "C" "S" "S" ...

#Pclass está classificado como inteiro, na verdade é categórica.
#O mesmo acontece para Sex

```

```

summary(full_df)

##   PassengerId      Survived       Pclass        Name
## Min.    :  1  Min.    :0.0000  Min.    :1.000  Length:1309
## 1st Qu.: 328  1st Qu.:0.0000  1st Qu.:2.000  Class  :character
## Median  : 655  Median  :0.0000  Median  :3.000  Mode   :character
## Mean    : 655  Mean    :0.3838  Mean    :2.295
## 3rd Qu.: 982  3rd Qu.:1.0000  3rd Qu.:3.000
## Max.   :1309  Max.   :1.0000  Max.   :3.000
## NA's    :418
##   Sex            Age         SibSp        Parch
## Length:1309      Min.    : 0.17  Min.    :0.0000  Min.    :0.000
## Class  :character  1st Qu.:21.00  1st Qu.:0.0000  1st Qu.:0.000
## Mode   :character  Median :28.00  Median :0.0000  Median :0.000
##                   Mean   :29.88  Mean   :0.4989  Mean   :0.385
##                   3rd Qu.:39.00  3rd Qu.:1.0000  3rd Qu.:0.000
##                   Max.   :80.00  Max.   :8.0000  Max.   :9.000
## NA's    :263
##   Ticket          Fare        Cabin        Embarked
## Length:1309      Min.    : 0.000  Length:1309  Length:1309
## Class  :character  1st Qu.: 7.896  Class  :character  Class  :character
## Mode   :character  Median : 14.454  Mode   :character  Mode   :character
##                   Mean   : 33.295
##                   3rd Qu.: 31.275
##                   Max.   :512.329
## NA's    :1

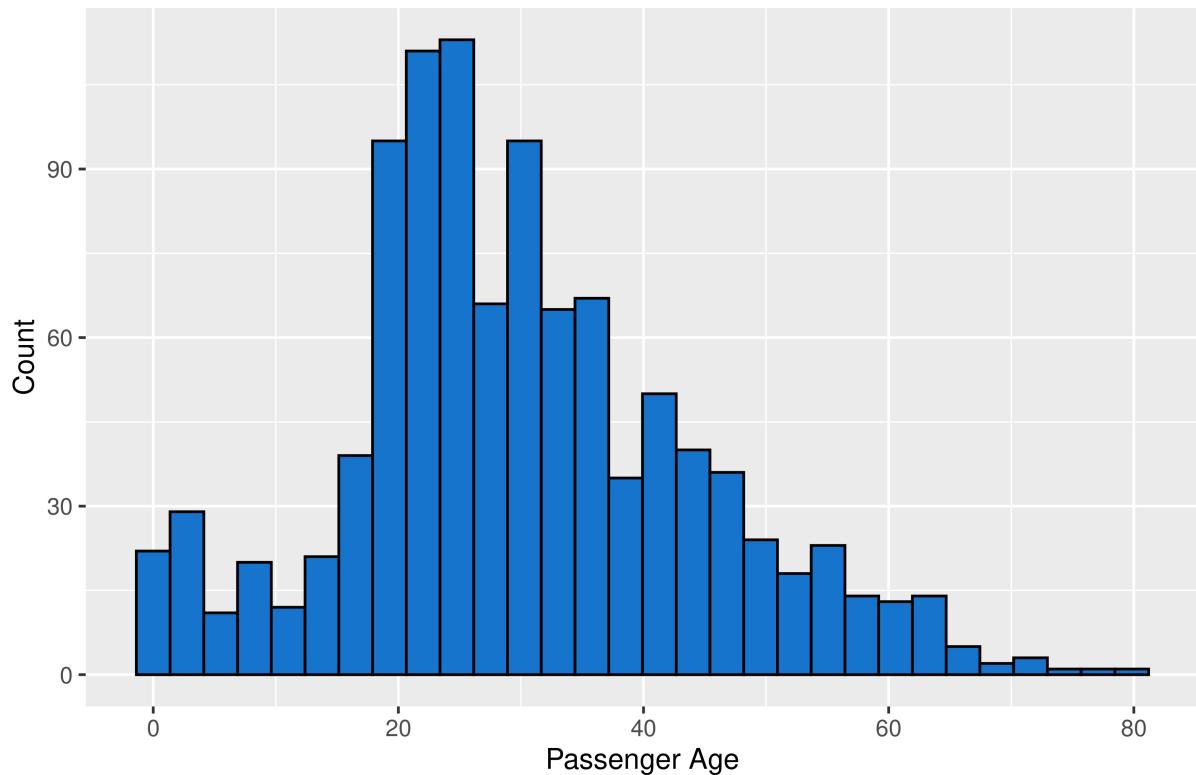
# A variável Age possui alguns valores NAs. Os valores precisam ser tratados mais a frente.

library(ggplot2)
histograma_idade <- ggplot(full_df,aes(x = Age)) +
  geom_histogram(colour = 'Black', fill = 'dodgerblue3') +
  labs(y = 'Count', x = 'Passenger Age', title = 'Age x Survived - Titanic')
histograma_idade

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 263 rows containing non-finite values (stat_bin).

```

Age x Survived – Titanic

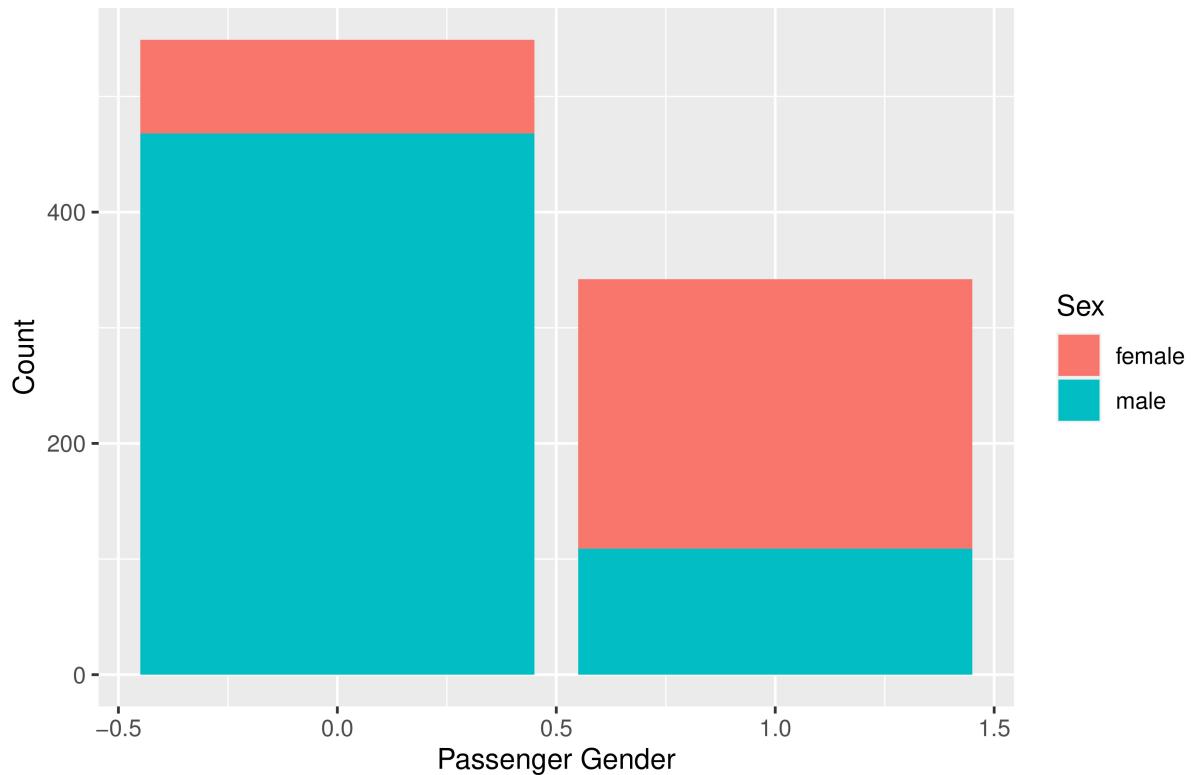


```
# O histograma mostra uma população predominante na casa dos 20 anos.
```

```
Barras_sexo <- ggplot(full_df) +  
  geom_bar(aes(x = Survived, fill = Sex)) +  
  labs(y = 'Count', x = 'Passenger Gender', title = 'Gender x Survived - Titanic')  
Barras_sexo
```

```
## Warning: Removed 418 rows containing non-finite values (stat_count).
```

Gender x Survived – Titanic

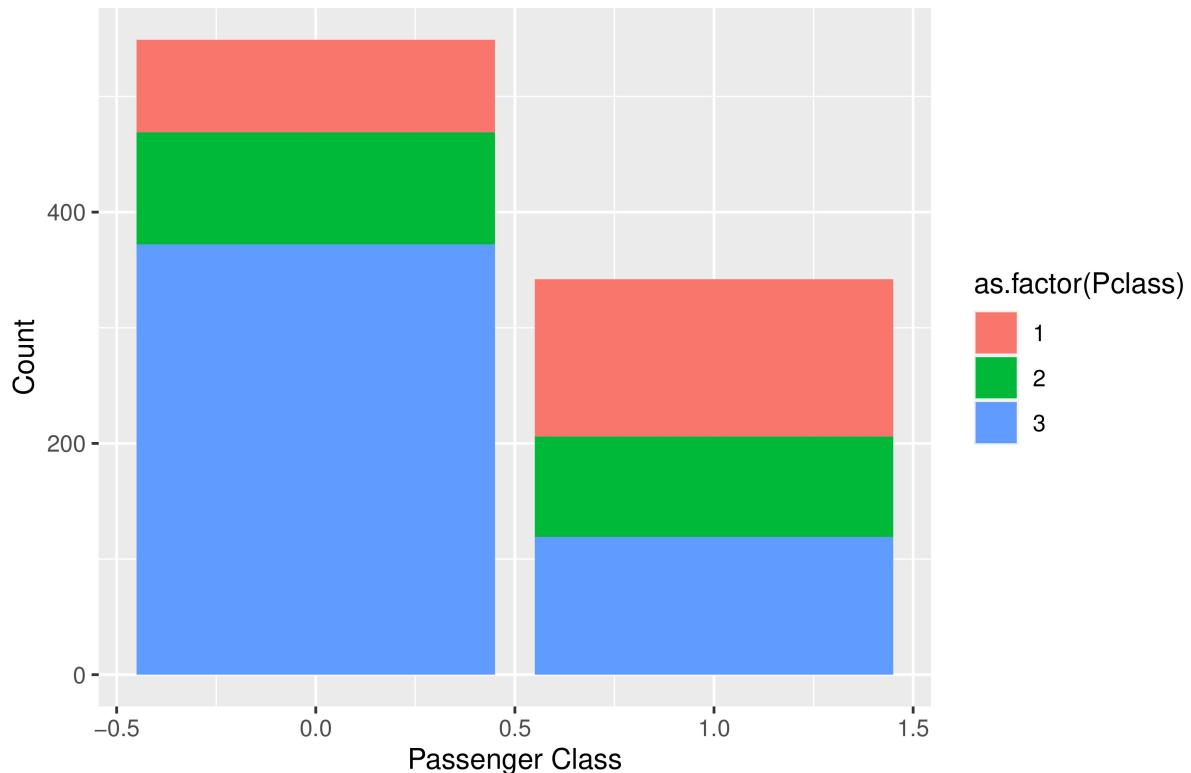


#O gráfico de barras mostra que o maior número de sobreviventes foram mulheres.
#no Titanic, a prioridade de evacuação foram mulheres e crianças primeiro
Um sinal de que é uma variável chave

```
Barras_Classe <- ggplot(full_df) +  
  geom_bar(aes(x = Survived, fill = as.factor(Pclass))) +  
  labs(y = 'Count', x = 'Passenger Class', title = 'Class x Survived - Titanic')  
Barras_Classe
```

```
## Warning: Removed 418 rows containing non-finite values (stat_count).
```

Class x Survived – Titanic



```
# Pelo gráfico de barras, o maior número de sobreviventes eram da 1ª classe  
# o que faz sentido pelo fato deles terem tido prioridade na evacuação  
# Ao contrário da 3ª classe que teve o maior número de vítimas na tragédia  
# Podemos inferir que essa é uma variável chave
```

Etapa 3 - Tratamento dos Dados

O próximo passo é remover os valores NAs da Idade e também a criação de uma nova variável. O nome dos passageiros possui o título na frente. A prioridade foi salvar as pessoas com título mais alto, como Condessa

```
#A variável Age possui 177 valores missing, conforme abaixo.  
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.1.1  
## Carregando pacotes exigidos: Rcpp  
## Warning: package 'Rcpp' was built under R version 4.1.1  
## ##  
## ## Amelia II: Multiple Imputation  
## ## (Version 1.8.0, built: 2021-05-26)  
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##  
table(is.na(full_df$Age))
```

```

##  

## FALSE TRUE  

## 1046 263  

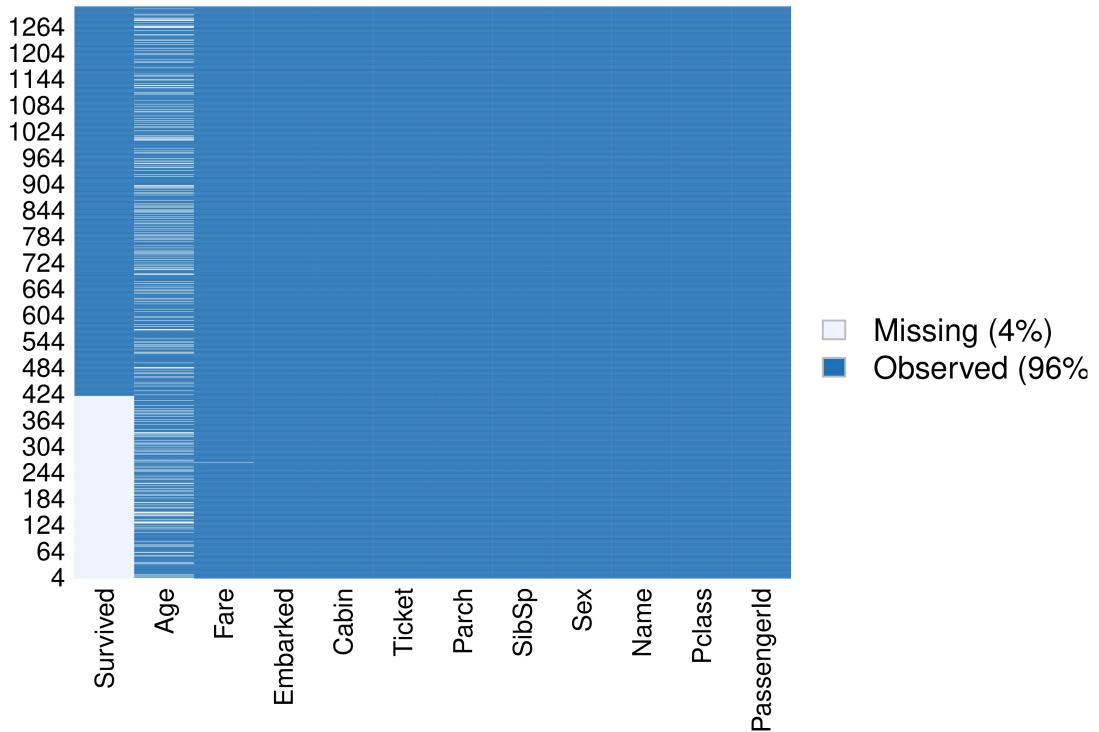
missmap(full_df)  

# Nesse caso, será usado a média dos demais valores para ocupar esses NAs  

missmap(full_df)

```

Missingness Map



```

full_df$Age[is.na(full_df$Age)] <- mean(full_df$Age, na.rm = T)  

full_df$Age <- round(full_df$Age, 0)  

table(is.na(full_df$Age))  

##  

## FALSE  

## 1309  

#Criação de coluna com títulos dos passageiros  

full_df>Title <- gsub("(.* , )|(\\..*)", "", full_df>Name) #separa string na , e no .  

#Classificação dos Rare Titles  

rare_titles <- c("Dona", "Lady", "the Countess", "Capt", "Col", "Don", "Dr", "Major", "Rev", "Sir", "Jon"  

#Ajustes nos títulos para Miss e Mrs  

full_df>Title[full_df>Title == "Mlle"] <- "Miss"  

full_df>Title[full_df>Title == "Ms"] <- "Miss"  

full_df>Title[full_df>Title == "Mme"] <- "Mrs"

```

```

#Classificação dos títulos como rare, apenas os que estão na lista
full_df$title[full_df$title %in% rare_titles] <- "Rare"
unique(full_df$title) #conferência dos valores únicos na coluna Title

## [1] "Mr"      "Mrs"     "Miss"    "Master"   "Rare"

full_df$title <- as.factor(full_df$title) #Coluna como fator
str(full_df)

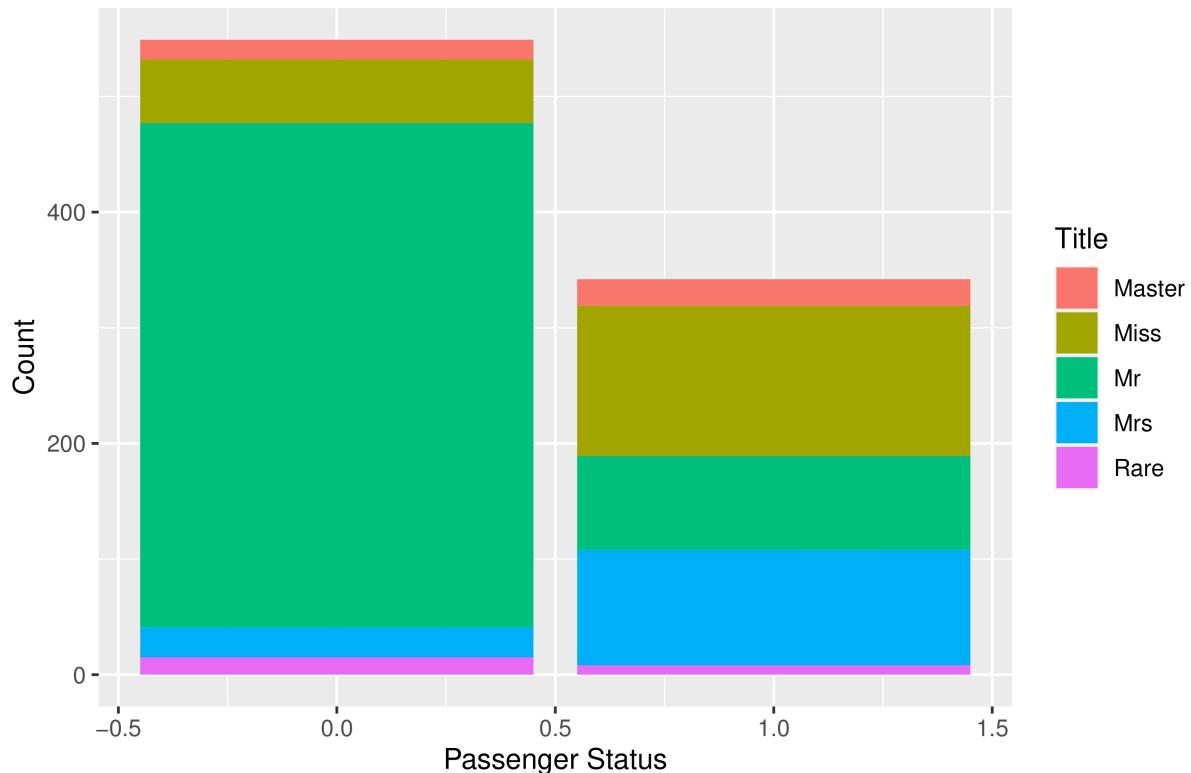
## 'data.frame': 1309 obs. of 13 variables:
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...
## $ Survived    : int 0 1 1 1 0 0 0 0 1 1 ...
## $ Pclass      : int 3 1 3 1 3 3 1 3 3 2 ...
## $ Name        : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" ...
## $ Sex         : chr "male" "female" "female" "female" ...
## $ Age         : num 22 38 26 35 35 30 54 2 27 14 ...
## $ SibSp       : int 1 1 0 1 0 0 0 3 0 1 ...
## $ Parch       : int 0 0 0 0 0 0 0 1 2 0 ...
## $ Ticket      : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
## $ Fare        : num 7.25 71.28 7.92 53.1 8.05 ...
## $ Cabin       : chr "" "C85" "" "C123" ...
## $ Embarked    : chr "S" "C" "S" "S" ...
## $ Title       : Factor w/ 5 levels "Master","Miss",...: 3 4 2 4 3 3 3 1 4 4 ...
rm(rare_titles) #remoção do vetor

Barras_titles <- ggplot(full_df) +
  geom_bar(aes(x = Survived, fill = Title)) +
  labs(y = 'Count', x = 'Passenger Status', title = 'Title x Survived - Titanic')
Barras_titles

## Warning: Removed 418 rows containing non-finite values (stat_count).

```

Title x Survived – Titanic



```
# A maior parte dos sobreviventes foram do título Miss (mulheres jovens e solteiras)
```

```
# Algumas variáveis precisam ser do tipo fator
```

```
fator <- c('Pclass', 'Sex', 'Title', 'Embarked')
full_df[fator] <- lapply(full_df[fator], function(x) as.factor(x))
str(full_df)
```

```
## 'data.frame': 1309 obs. of 13 variables:
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...
## $ Survived    : int 0 1 1 1 0 0 0 0 1 1 ...
## $ Pclass      : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
## $ Name        : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" ...
## $ Sex         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
## $ Age         : num 22 38 26 35 35 30 54 2 27 14 ...
## $ SibSp       : int 1 1 0 1 0 0 0 3 0 1 ...
## $ Parch       : int 0 0 0 0 0 0 1 2 0 ...
## $ Ticket      : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
## $ Fare        : num 7.25 71.28 7.92 53.1 8.05 ...
## $ Cabin       : chr "" "C85" "" "C123" ...
## $ Embarked    : Factor w/ 4 levels "", "C", "Q", "S": 4 2 4 4 4 3 4 4 4 2 ...
## $ Title       : Factor w/ 5 levels "Master", "Miss", ...: 3 4 2 4 3 3 3 1 4 4 ...
rm(fator)
```

Etapa 4 - Preparação do dataset para a modelagem preditiva

Nesta etapa é preciso dividir os dados em treino e teste, assim como ele estavam originalmente Os dados de treino estavam até o índice 892 e o restante fica para teste.

```
df_train <- full_df[1:891,]
df_test <- full_df[892:1309,]
View(df_train)
View(df_test)
```

Etapa 5 - Modelo de Regressão Logística

```
model_logistic <- glm(Survived ~ Pclass + Parch + Sex + Age + SibSp + Embarked + Title,
                       data = df_train, family = 'binomial')
summary(model_logistic)

##
## Call:
## glm(formula = Survived ~ Pclass + Parch + Sex + Age + SibSp +
##       Embarked + Title, family = "binomial", data = df_train)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -2.3710 -0.5555 -0.3796  0.5423  2.5537
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.326e+01 1.327e+03  0.025  0.98000
## Pclass2     -1.212e+00 2.940e-01 -4.124 3.72e-05 ***
## Pclass3     -2.337e+00 2.707e-01 -8.633 < 2e-16 ***
## Parch      -3.124e-01 1.304e-01 -2.396  0.01659 *
## Sexmale    -1.625e+01 8.375e+02 -0.019  0.98452
## Age        -2.596e-02 9.107e-03 -2.851  0.00436 **
## SibSp      -5.213e-01 1.234e-01 -4.223 2.41e-05 ***
## EmbarkedC  -1.246e+01 1.029e+03 -0.012  0.99034
## EmbarkedQ  -1.258e+01 1.029e+03 -0.012  0.99024
## EmbarkedS  -1.295e+01 1.029e+03 -0.013  0.98996
## TitleMiss -1.676e+01 8.375e+02 -0.020  0.98404
## TitleMr   -3.489e+00 5.334e-01 -6.541 6.13e-11 ***
## TitleMrs  -1.604e+01 8.375e+02 -0.019  0.98472
## TitleRare -3.597e+00 7.821e-01 -4.600 4.23e-06 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1186.66 on 890 degrees of freedom
## Residual deviance: 727.57 on 877 degrees of freedom
## AIC: 755.57
##
## Number of Fisher Scoring iterations: 14
# Lista com as variáveis mais relevantes para o modelo de regressão.
library(caret)
```

```

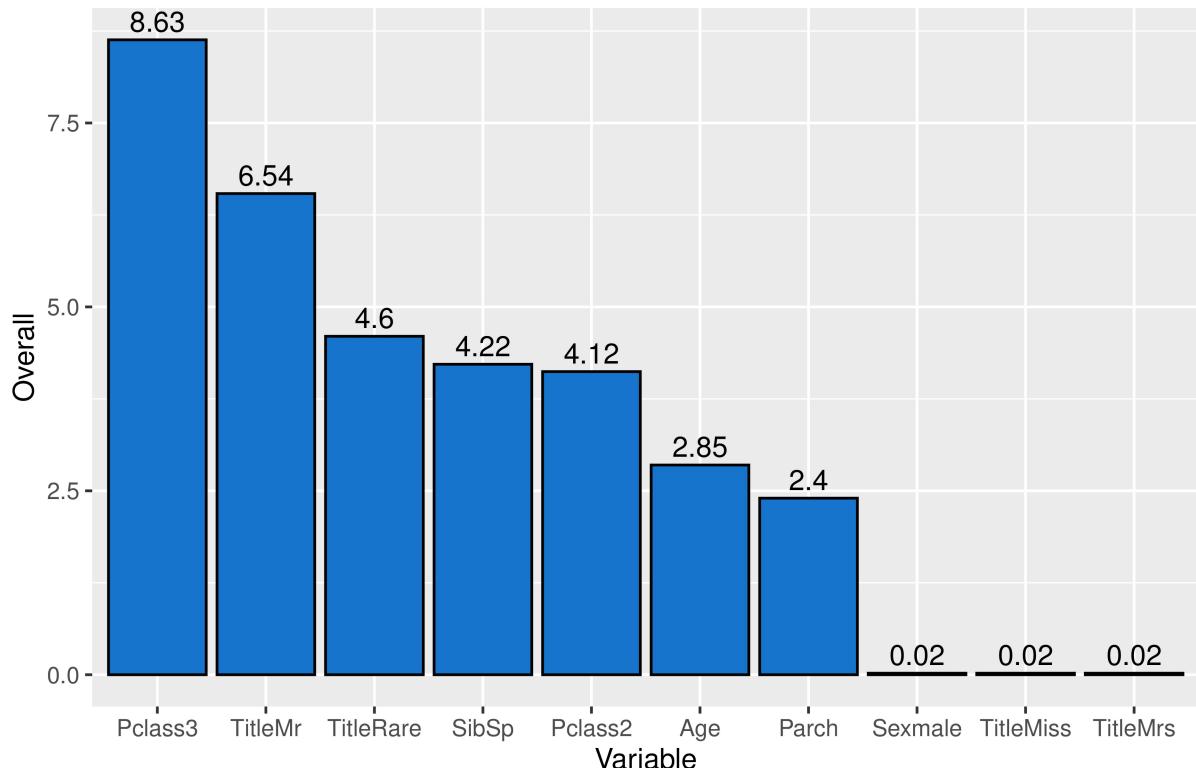
## Warning: package 'caret' was built under R version 4.1.1
## Carregando pacotes exigidos: lattice
variavel_modelo <- varImp(model_logistic)
plot_var <- variavel_modelo %>% arrange(desc(Overall)) %>% top_n(10)

## Selecting by Overall
plot_var <- round(plot_var,2)
plot_var$class <- row.names(plot_var)

# Gráfico que mostra as variáveis mais relevantes no modelo de regressão logística
ggplot(plot_var,
       aes(x = reorder(class,-Overall), y = Overall)) +
  geom_col(colour = 'Black', fill = 'dodgerblue3') +
  labs(x = 'Variable', title = 'Top 10 Feature - Logistic Regression', y = 'Overall') +
  geom_text(aes(label = Overall, vjust = -0.4))

```

Top 10 Feature – Logistic Regression



```

# No modelo de regressão, caso ele seja maior que 0.5, pode-se considerar como positivo.
# Nesse caso, como 1.
resultado_logistic <- predict(model_logistic, newdata = df_test, type = 'response')
aux_rl <- ifelse(resultado_logistic > 0.5, 1, 0)

# Gerar arquivo de resposta para o Kaggle!
submission_kaggle <- submission
submission_kaggle$Survived <- aux_rl
write.table(submission_kaggle, 'submission_titanic.csv', sep = ",", row.names = F)

```

```

#Score no Kaggle foi 0.77272 (77,2%)

contributors() # Resposáveis pelo R

citation('ggplot2') #Pacote ggplot2

## 
## To cite ggplot2 in publications, please use:
## 
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
## 
## A BibTeX entry for LaTeX users is
## 
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }

citation('caret') #Pacote caret

## 
## To cite package 'caret' in publications use:
## 
##   Max Kuhn (2021). caret: Classification and Regression Training. R
##   package version 6.0-88. https://CRAN.R-project.org/package=caret
## 
## A BibTeX entry for LaTeX users is
## 
##   @Manual{,
##     title = {caret: Classification and Regression Training},
##     author = {Max Kuhn},
##     year = {2021},
##     note = {R package version 6.0-88},
##     url = {https://CRAN.R-project.org/package=caret},
##   }

citation('dplyr') #Pacote dplyr

## 
## To cite package 'dplyr' in publications use:
## 
##   Hadley Wickham, Romain François, Lionel Henry and Kirill Müller
##   (2021). dplyr: A Grammar of Data Manipulation. R package version
##   1.0.7. https://CRAN.R-project.org/package=dplyr
## 
## A BibTeX entry for LaTeX users is
## 
##   @Manual{,
##     title = {dplyr: A Grammar of Data Manipulation},
##     author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller},

```

```

##     year = {2021},
##     note = {R package version 1.0.7},
##     url = {https://CRAN.R-project.org/package=dplyr},
##   }
citation('Amelia') #Pacote Amelia

##
## To cite Amelia in publications use:
##
## James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A
## Program for Missing Data. Journal of Statistical Software, 45(7),
## 1-47. URL https://www.jstatsoft.org/v45/i07/.
##
## A BibTeX entry for LaTeX users is
##
## @Article{,
##   title = {{Amelia II}: A Program for Missing Data},
##   author = {James Honaker and Gary King and Matthew Blackwell},
##   journal = {Journal of Statistical Software},
##   year = {2011},
##   volume = {45},
##   number = {7},
##   pages = {1--47},
##   url = {https://www.jstatsoft.org/v45/i07/},
## }

```