Health Analytics

Diego Wenceslau

25/01/2021

Projeto: Podemos Prever o Tempo de Sobrevivência dos Pacientes 1 Ano Após Receberem um Transplante?

Este projeto tem como objetivo, criar um modelo que seja capaz de prever o tempo de sobrevivência dos Pacientes 1 Ano Após Receberem um Transplante.

Usaremos dados reais disponibilizados publicamente.

Os dados foram extraídos do SRTR Database e modificados para que possa ser executado o script na máquina.

Site oficial dos dados:https://www.srtr.org/about-the-data/the-srtr-database/

Definindo o diretório de trabalho

```
setwd("C:/FCD/Business_Analytics/Cap09")
getwd()
```

```
## [1] "C:/FCD/Business_Analytics/Cap09"
```

Liberando os pacotes

```
library(dplyr)
```

```
##
## 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
```

```
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
library(forecast)
library(nnet)
library(neuralnet)
```

```
##
## Attaching package: 'neuralnet'
```

```
## The following object is masked from 'package:dplyr':
##
##     compute
```

Carregando os dados

```
dados <- read.csv("dados/dataset.csv", header = TRUE, na.strings = c(""))
dim(dados)
```

```
## [1] 79100    46
```

Análise Exploratória, Limpeza, Transformação e Manipulação de Dados (Data Wrangling) Visualizando os dados

Visualizar as primeiras linhas

```
head(dados, 5)
```

```
##   ï..DAYSWAIT_CHRON PSTATUS FINAL_MELD_SCORE PTIME    TX_DATE PX_STAT
## 1                 7       0               39    51 12/24/2018       A
## 2                 5       0               19     6 12/23/2018       A
## 3                10       0               22     6 12/28/2018       A
## 4                 9       0               35    27 12/27/2018       A
## 5                 2       0               35    54 12/20/2018       A
##   PX_STAT_DATE AGE ABO GENDER WGT_KG_TCR HGT_CM_TCR BMI_TCR DIAB INIT_AGE
## 1    2/13/2019  30   O      1      56.24     162.60   21.27    1       30
## 2   12/29/2018  63   A      0      81.92     177.80   25.91    1       63
## 3     1/3/2019  48   B      0      78.93     181.10   24.06    1       48
## 4    1/23/2019  54   O      1      63.50     154.94   26.45    1       54
## 5    2/12/2019  71   O      1      75.75     162.56   28.67    1       71
##   ETHCAT REGION PERM_STATE TX_Year TX_PROCEDUR_TY MED_COND_TRR PREV_TX AGE_DON
## 1      1      2         MD    2018            701            1       N      24
## 2      1      3         GA    2018            701            3       N      34
## 3      1     10         OH    2018            701            3       N      42
## 4      1      4         TX    2018            701            2       N      48
## 5      1      3         LA    2018            701            3       N      37
##   GENDER_DON HGT_CM_DON_CALC WGT_KG_DON_CALC BMI_DON_CALC COD_CAD_DON
## 1          1             173            75.0        25.06           3
## 2          0             183            90.0        26.87           3
## 3          1             173           107.0        35.75           1
## 4          1             157            93.0        37.73           2
## 5          1             173            81.6        27.26           3
##   ETHCAT_DON HOME_STATE_DON DIABETES_DON HIST_HYPERTENS_DON
## 1          2             PA            N                  N
## 2          1             GA            N                  N
```

```
## 3          1             NY            Y                  Y
## 4          1             TX            N                  Y
## 5          1             AR            N                  N
##   HIST_IV_DRUG_OLD_DON ABO_DON HIST_CANCER_DON ALCOHOL_HEAVY_DON ABO_MAT
## 1                <NA>       O              N                 N       1
## 2                <NA>       A              N                 N       1
## 3                <NA>       B              N                 N       1
## 4                <NA>       O              N                 N       1
## 5                <NA>       O              N                 N       1
##   COLD_ISCH MALIG HGT_CM_CALC WGT_KG_CALC BMI_CALC TX_MELD LISTYR LiverSize
## 1      4.30     U       162.6        45.0     17.0      No   2018  1721.500
## 2      3.48     U       177.8        85.0     26.9      No   2018  1934.720
## 3      4.95     U       182.9        76.2     22.8      No   2018  1987.348
## 4      3.62     U       154.9        61.1     25.5      No   2018  1669.494
## 5      7.50     U       162.6        70.8     26.8      No   2018  1605.492
##   LiverSizeDon
## 1     2276.860
## 2     2387.360
## 3     2555.460
## 4     2255.140
## 5     2214.884
```

Renomeando a primeira coluna

```
colnames(dados)[1]<-'DAYSWAIT_CHRON'
```

Tipos dos dados

```
str(dados)
```

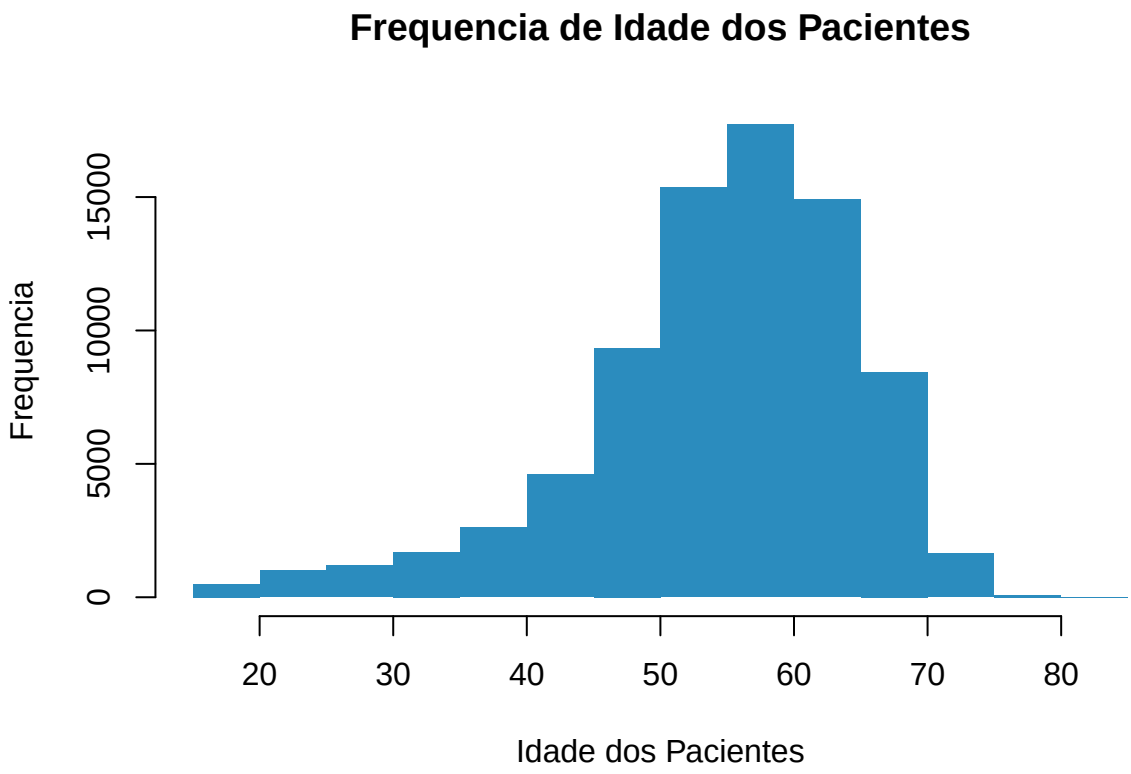
```
## 'data.frame':    79100 obs. of  46 variables:
##  $ DAYSWAIT_CHRON      : int  7 5 10 9 2 6 4 9 1 11 ...
##  $ PSTATUS             : int  0 0 0 0 0 0 0 1 1 0 ...
##  $ FINAL_MELD_SCORE    : int  39 19 22 35 35 19 35 14 36 23 ...
##  $ PTIME               : int  51 6 6 27 54 10 51 0 3 6 ...
##  $ TX_DATE             : Factor w/ 6139 levels "1/1/2003","1/1/2004",..: 1822 1805 1890 1873 1754 181
##  $ PX_STAT             : Factor w/ 2 levels "A","D": 1 1 1 1 1 1 1 2 2 1 ...
##  $ PX_STAT_DATE        : Factor w/ 5774 levels ".","1/1/2003",..: 2046 1818 377 264 2030 377 2013 180
##  $ AGE                 : int  30 63 48 54 71 62 62 56 28 66 ...
##  $ ABO                 : Factor w/ 8 levels "A","A1","A1B",..: 8 1 7 8 8 8 8 1 7 1 ...
##  $ GENDER              : int  1 0 0 1 1 1 1 0 1 0 ...
##  $ WGT_KG_TCR          : num  56.2 81.9 78.9 63.5 75.8 ...
##  $ HGT_CM_TCR          : num  163 178 181 155 163 ...
##  $ BMI_TCR             : num  21.3 25.9 24.1 26.4 28.7 ...
##  $ DIAB                : int  1 1 1 1 1 3 1 1 1 1 ...
##  $ INIT_AGE            : int  30 63 48 54 71 62 62 56 28 66 ...
##  $ ETHCAT              : int  1 1 1 1 1 1 1 4 2 1 ...
##  $ REGION              : int  2 3 10 4 3 3 5 3 3 11 ...
##  $ PERM_STATE          : Factor w/ 56 levels "AK","AL","AR",..: 22 12 38 47 20 20 6 42 2 44 ...
##  $ TX_Year             : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ TX_PROCEDUR_TY      : int  701 701 701 701 701 701 701 701 701 701 ...
##  $ MED_COND_TRR        : int  1 3 3 2 3 1 1 3 1 3 ...
##  $ PREV_TX             : Factor w/ 2 levels "N","Y": 1 1 1 1 1 2 1 1 1 1 ...
```

```
##  $ AGE_DON             : int  24 34 42 48 37 38 23 47 18 35 ...
##  $ GENDER_DON          : int  1 0 1 1 1 0 0 0 0 0 ...
##  $ HGT_CM_DON_CALC     : num  173 183 173 157 173 165 175 163 173 178 ...
##  $ WGT_KG_DON_CALC     : num  75 90 107 93 81.6 55.7 93 49.9 68 67.7 ...
##  $ BMI_DON_CALC        : num  25.1 26.9 35.8 37.7 27.3 ...
##  $ COD_CAD_DON         : int  3 3 1 2 3 3 1 2 3 1 ...
##  $ ETHCAT_DON          : int  2 1 1 1 1 1 4 4 4 1 ...
##  $ HOME_STATE_DON      : Factor w/ 57 levels "AK","AL","AR",..: 42 12 38 48 3 11 5 43 43 45 ...
##  $ DIABETES_DON        : Factor w/ 3 levels "N","U","Y": 1 1 3 1 1 1 1 1 1 1 ...
##  $ HIST_HYPERTENS_DON  : Factor w/ 3 levels "N","U","Y": 1 1 3 3 1 1 1 1 1 1 ...
##  $ HIST_IV_DRUG_OLD_DON: Factor w/ 3 levels "N","U","Y": NA NA NA NA NA NA NA NA NA NA ...
##  $ ABO_DON             : Factor w/ 8 levels "A","A1","A1B",..: 8 1 7 8 8 8 8 4 7 2 ...
##  $ HIST_CANCER_DON     : Factor w/ 3 levels "N","U","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ALCOHOL_HEAVY_DON   : Factor w/ 3 levels "N","U","Y": 1 1 1 1 1 3 1 1 1 1 ...
##  $ ABO_MAT             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ COLD_ISCH           : num  4.3 3.48 4.95 3.62 7.5 5.33 9.14 5.25 4.6 5.6 ...
##  $ MALIG               : Factor w/ 3 levels "N","U","Y": 2 2 2 2 2 2 2 2 2 3 ...
##  $ HGT_CM_CALC         : num  163 178 183 155 163 ...
##  $ WGT_KG_CALC         : num  45 85 76.2 61.1 70.8 ...
##  $ BMI_CALC            : num  17 26.9 22.8 25.5 26.8 47.9 19.5 32.1 33.7 24.9 ...
##  $ TX_MELD             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LISTYR              : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ LiverSize           : num  1722 1935 1987 1669 1605 ...
##  $ LiverSizeDon        : num  2277 2387 2555 2255 2215 ...
```

Explorando os dados das variáveis numéricas

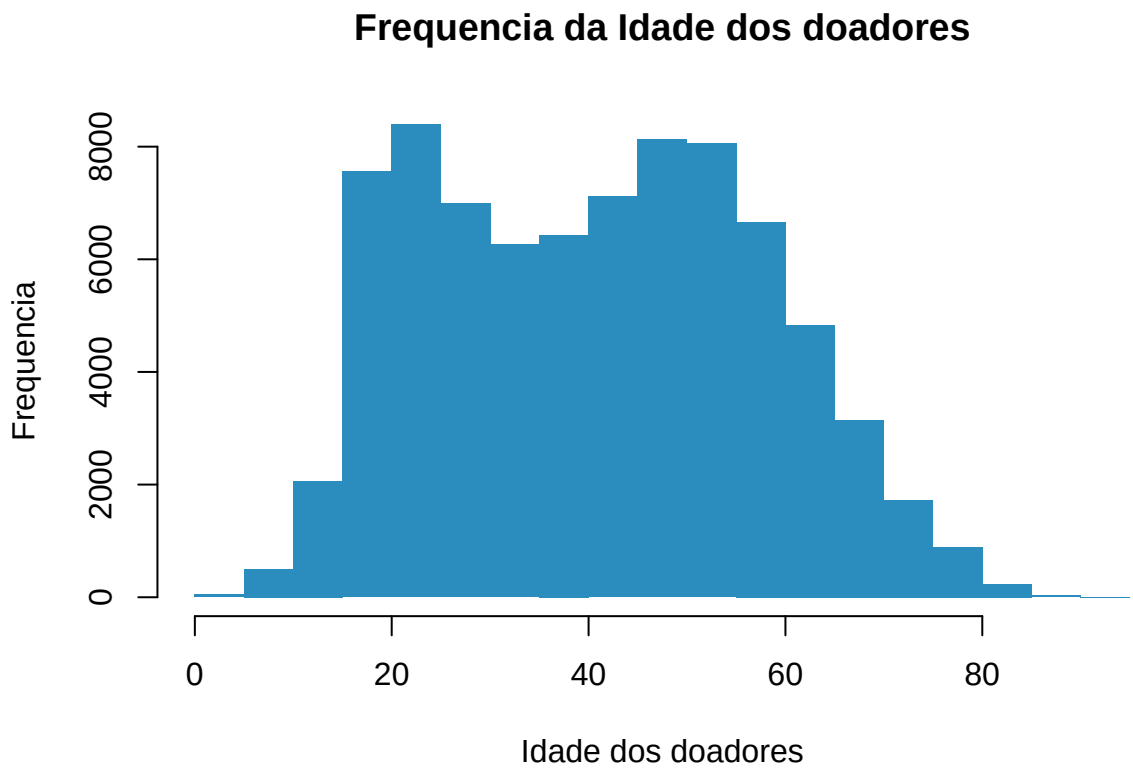
A maioria dos pacientes, tem em torno de 60 anos de idade

```
hist(dados$AGE,
     main = "Frequencia de Idade dos Pacientes",
     xlab = "Idade dos Pacientes", ylab = "Frequencia",
     col = c("#2b8cbe"),
     border = FALSE)
```



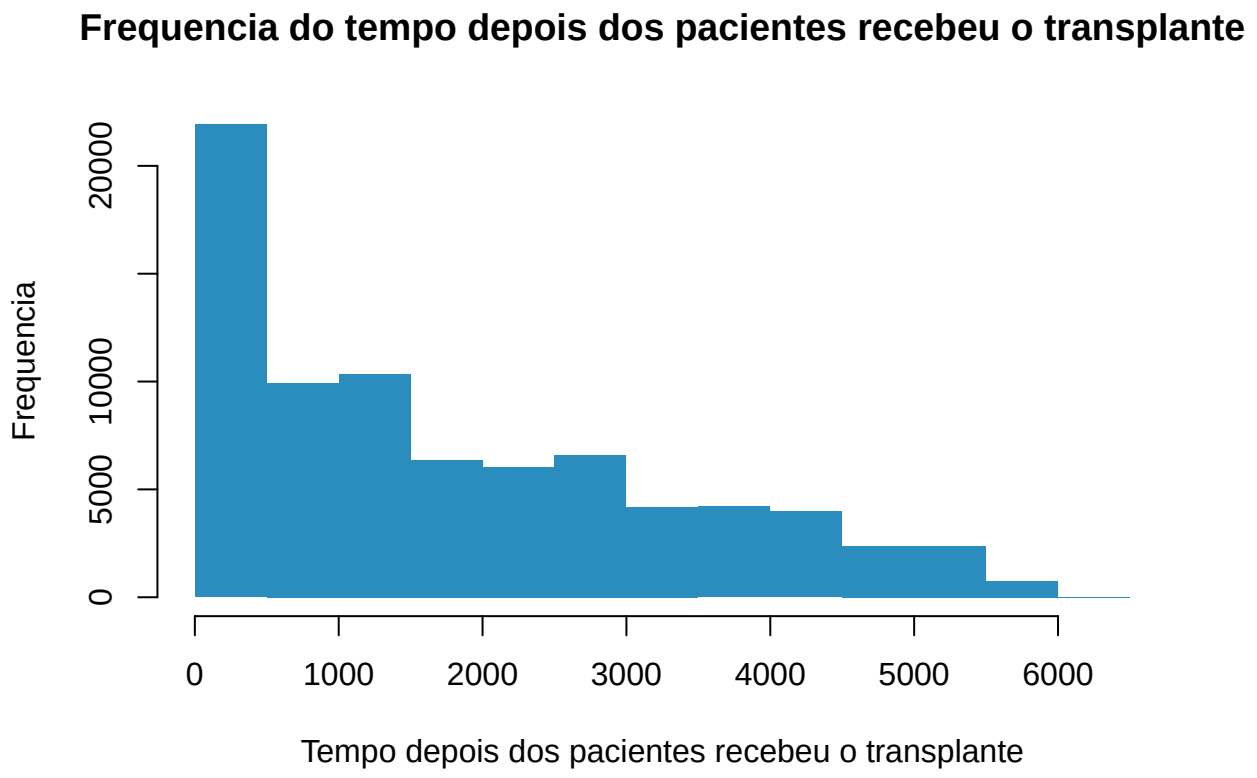
A idade dos doadores de fígado, é bem variada, com quantidade maior de doadores, começando um pouco antes dos 20 anos de idade.

```
hist(dados$AGE_DON,
     main = "Frequencia da Idade dos doadores",
     xlab = "Idade dos doadores", ylab = "Frequencia",
     col = c("#2b8cbe"),
     border = FALSE)
```



A grande maioria das pessoas que recebeu o transplante de fígado, vivem cerca de 500 dias.

```
hist(dados$PTIME,
     main = "Frequencia do tempo depois dos pacientes recebeu o transplante",
     xlab = "Tempo depois dos pacientes recebeu o transplante", ylab = "Frequencia",
     col = c("#2b8cbe"),
     border = FALSE)
```

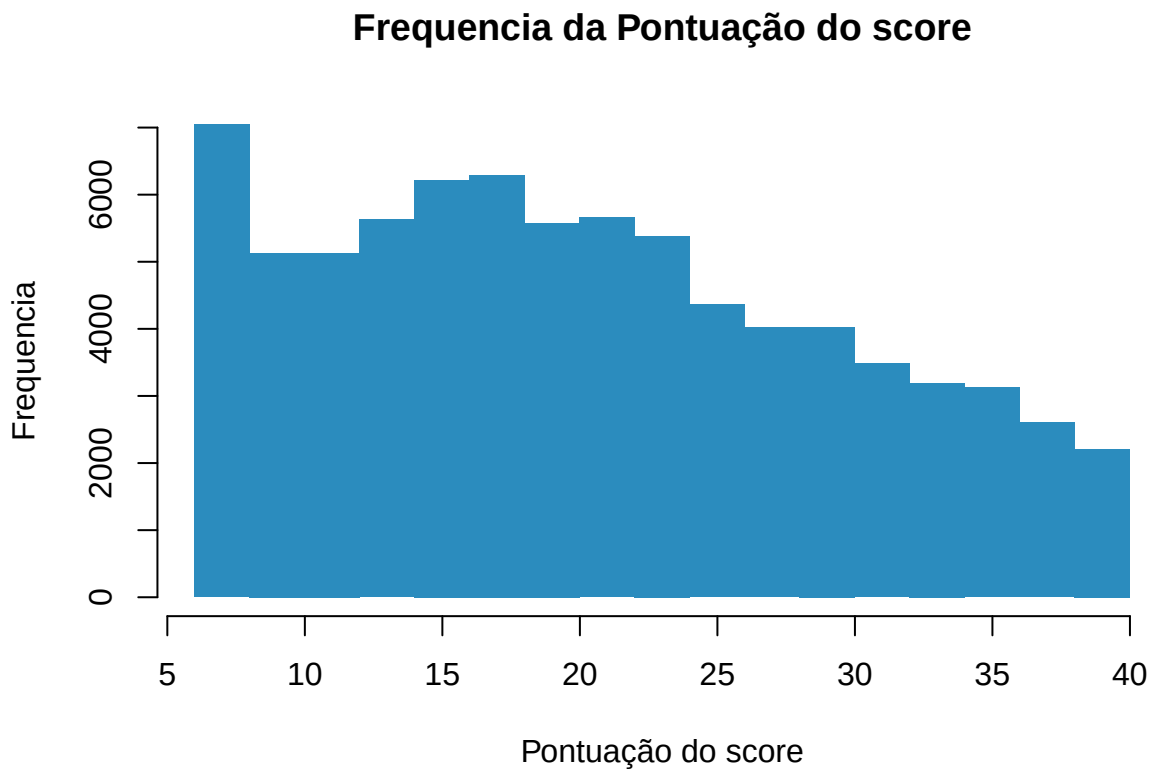


Tempo de espera é cerca de 1000 dias para uma pessoa receber o transplante.

```
hist(dados$DAYSWAIT_CHRON,
     main = "Frequencia do Tempo de espera para uma pessoa receber o transplante",
     xlab = "Tempo de espera", ylab = "Frequencia",
     col = c("#2b8cbe"),
     border = FALSE)
```

Pontuação do score da doença hepática que é qualquer condição que danifica o fígado e impede seu bom funcionamento.

```
hist(dados$FINAL_MELD_SCORE,
     main = "Frequencia da Pontuação do score",
     xlab = "Pontuação do score", ylab = "Frequencia",
     col = c("#2b8cbe"),
     border = FALSE)
```



Explorando os dados das variáveis categóricas

```
dados$DIAB <- as.factor(dados$DIAB)
table(dados$DIAB)
```

```
##
##     1     2     3     4     5   998
## 57017  1520 16476   309  2939   828
```

A quantidade mostra mais pacientes sobreviveram ao transplante. 0 -> não veio a óbito 1 -> veio a óbito

```
dados$PSTATUS <- as.factor(dados$PSTATUS)
table(dados$PSTATUS)
```

```
##
##     0     1
## 55634 23466
```

Quantidade de paciente que recebe transplante. 0 -> Masculino 1 -> Feminino

```
dados$GENDER <- as.factor(dados$GENDER)
table(dados$GENDER)
```

```
##
##     0     1
## 53312 25788
```

Quantidade de doadores. 0 -> Masculino 1 -> Feminino

```
dados$GENDER_DON <- as.factor(dados$GENDER_DON)
table(dados$GENDER_DON)
```

```
##
##     0     1
## 47310 31790
```

Regiões.

```
dados$REGION <- as.factor(dados$REGION)
table(dados$REGION)
```

```
##
##     1     2     3     4     5     6     7     8     9    10    11
##  2802  9435 14070  7436  9962  2505  6503  5458  4745  7656  8528
```

Anos que foram utilizado na pesquisa da coleta de dados

```
dados$TX_Year <- as.factor(dados$TX_Year)
table(dados$TX_Year)
```

```
##
## 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
##    1 1456 2948 3717 4062 4475 4501 4459 4641 4583 4745 4751 4898 5128 5430 6239
## 2017 2018
## 6519 6547
```

Se tem tumor maligimo

```
dados$MALIG <- as.factor(dados$MALIG)
table(dados$MALIG)
```

```
##
##     N     U     Y
## 42828 21290 14982
```

Histórico de câncer do paciente

```
dados$HIST_CANCER_DON <- as.factor(dados$HIST_CANCER_DON)
table(dados$HIST_CANCER_DON)
```

```
##
##     N     U     Y
## 76040   398  2660
```

Quantidade de 61600 considerando apenas os pacientes que sobreviveram ao primeiro ano de cirurgia

```r
dados1 <- dados %>%
  filter(PTIME > 365) %>%
  mutate(PTIME = (PTIME - 365))

dim(dados1)
```

```
## [1] 61600    46
```

Quantidade de 23348 dos pacientes que sobreviveram ao primeiro ano da cirurgia. Filtramos os que permaneceram vivos por até três anos depois da cirurgia.

```r
dados2 <- dados1 %>%
  filter(PTIME <= 1095)

dim(dados2)
```

```
## [1] 23348    46
```

Vamos separar variáveis numéricas e categóricas

```r
dados_num <- dados2[,!unlist(lapply(dados2, is.factor))]
dim(dados_num)
```

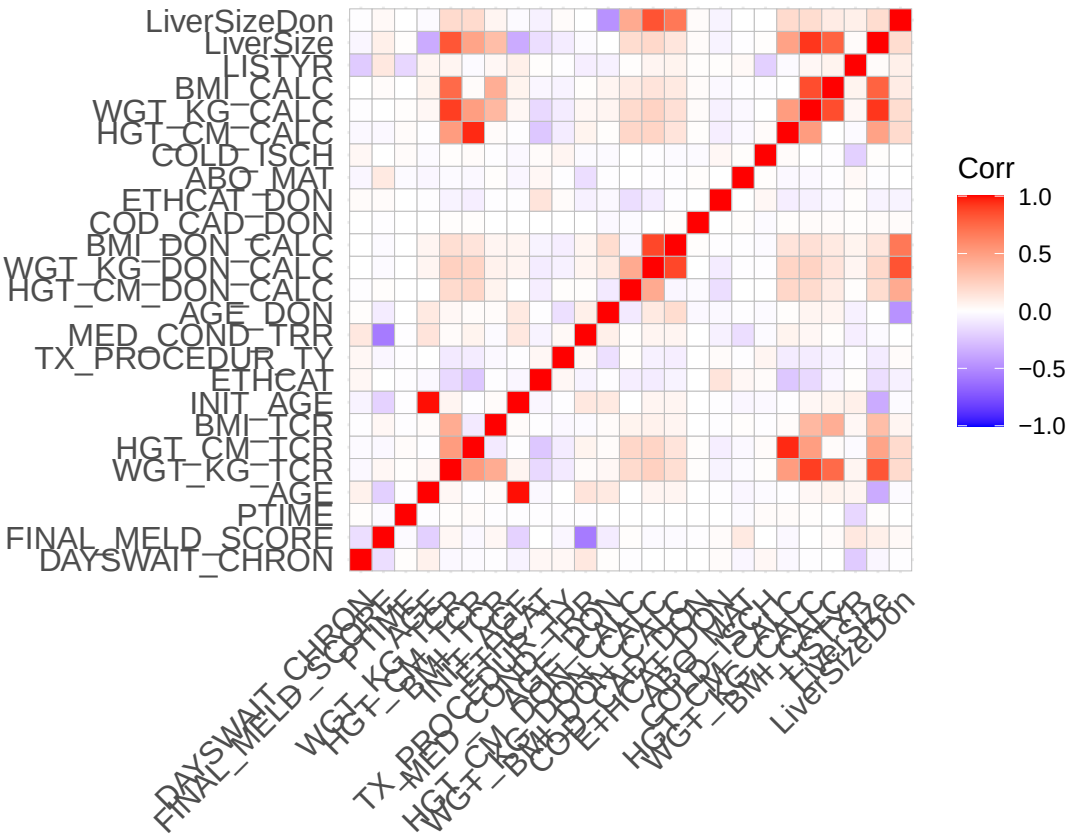
```
## [1] 23348    25
```

```r
dados_fator <- dados2[,unlist(lapply(dados2, is.factor))]
dim(dados_fator)
```

```
## [1] 23348    21
```

Correlação entre as variáveis numéricas Para variáveis categóricas usamos associação

```r
df_corr <- round(cor(dados_num, use = "complete.obs"), 2)
ggcorrplot(df_corr)
```



Agora vamos padronizar os dados de treino e teste de forma separada. # Executamos o procedimento anterior, mas de forma separada em cada subset.

```
set.seed(1)
index <- sample(1:nrow(dados2), dim(dados2)[1]*.7)
dados_treino <- dados2[index,]
dados_teste <- dados2[-index,]
```

Vamos separar variáveis numéricas e categóricas (treino)

```
dados_treino_num <- dados_treino[,!unlist(lapply(dados_treino, is.factor))]
dim(dados_treino_num)
```

```
## [1] 16343    25
```

```
dados_treino_fator <- dados_treino[,unlist(lapply(dados_treino, is.factor))]
dim(dados_treino_fator)
```

```
## [1] 16343    21
```

Vamos separar variáveis numéricas e categóricas (teste)

```
dados_teste_num <- dados_teste[,!unlist(lapply(dados_teste, is.factor))]
dim(dados_teste_num)
```

```
## [1] 7005   25
```

```
dados_teste_fator <- dados_teste[,unlist(lapply(dados_teste, is.factor))]
dim(dados_teste_fator)
```

```
## [1] 7005   21
```

Padronização dados de treino

```
dados_treino_num_norm <- scale(dados_treino_num)
dados_treino_final <- cbind(dados_treino_num_norm, dados_treino_fator)
dim(dados_treino_final)
```

```
## [1] 16343    46
```

Padronização dados de teste

```
dados_teste_num_norm <- scale(dados_teste_num)
dados_teste_final <- cbind(dados_teste_num_norm, dados_teste_fator)
dim(dados_teste_final)
```

```
## [1] 7005   46
```

Filtra os anos de 2001 e 2002

```
dados_treino_final <- dados_treino_final %>%
  filter(TX_Year != 2001) %>%
  filter(TX_Year != 2002)
```

```
dados_teste_final <- dados_teste_final %>%
  filter(TX_Year != 2001) %>%
  filter(TX_Year != 2002)
```

Cria novamente o modelo agora com o outro dataset de treino

```
modelo_v1 <- lm(PTIME ~ FINAL_MELD_SCORE +
                  REGION +
                  LiverSize +
                  LiverSizeDon +
                  ALCOHOL_HEAVY_DON +
                  MALIG +
                  TX_Year,
                data = dados_treino_final)

summary(modelo_v1)
```

```
##
## Call:
## lm(formula = PTIME ~ FINAL_MELD_SCORE + REGION + LiverSize +
##     LiverSizeDon + ALCOHOL_HEAVY_DON + MALIG + TX_Year, data = dados_treino_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.78477 -0.26397 -0.00602  0.32273  1.98940
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.313873   0.300577  -1.044 0.296393
## FINAL_MELD_SCORE   -0.002193   0.005565  -0.394 0.693577
## REGION2            -0.006503   0.031024  -0.210 0.833964
## REGION3            -0.033365   0.029779  -1.120 0.262549
## REGION4             0.012703   0.031875   0.399 0.690237
## REGION5             0.027234   0.030825   0.883 0.376983
## REGION6             0.057101   0.041459   1.377 0.168439
## REGION7            -0.013949   0.033254  -0.419 0.674870
## REGION8             0.080145   0.034082   2.352 0.018706 *
## REGION9            -0.007366   0.035613  -0.207 0.836138
## REGION10            0.014251   0.032071   0.444 0.656791
## REGION11            0.011422   0.031608   0.361 0.717821
## LiverSize           0.012879   0.005458   2.360 0.018298 *
## LiverSizeDon        0.005114   0.005448   0.939 0.347891
## ALCOHOL_HEAVY_DONU -0.114221   0.040030  -2.853 0.004331 **
## ALCOHOL_HEAVY_DONY  0.018036   0.014601   1.235 0.216756
## MALIGU             -0.146375   0.021319  -6.866 6.84e-12 ***
## MALIGY             -0.110179   0.018874  -5.838 5.40e-09 ***
## TX_Year2004         0.346311   0.302151   1.146 0.251750
## TX_Year2005         0.313592   0.301255   1.041 0.297914
## TX_Year2006         0.304034   0.301130   1.010 0.312681
## TX_Year2007         0.267007   0.301163   0.887 0.375315
## TX_Year2008         0.284077   0.301141   0.943 0.345524
## TX_Year2009         0.293203   0.301130   0.974 0.330231
## TX_Year2010         0.315918   0.301159   1.049 0.294191
## TX_Year2011         0.336151   0.301216   1.116 0.264447
## TX_Year2012         0.414068   0.301289   1.374 0.169361
## TX_Year2013         0.572770   0.300947   1.903 0.057030 .
## TX_Year2014         1.606153   0.299694   5.359 8.47e-08 ***
## TX_Year2015         1.002783   0.300052   3.342 0.000834 ***
## TX_Year2016         0.087718   0.300131   0.292 0.770086
## TX_Year2017        -0.763483   0.300268  -2.543 0.011010 *
## TX_Year2018        -1.003027   0.303191  -3.308 0.000941 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6687 on 15808 degrees of freedom
##   (376 observations deleted due to missingness)
## Multiple R-squared:  0.5542, Adjusted R-squared:  0.5533
## F-statistic: 614.2 on 32 and 15808 DF,  p-value: < 2.2e-16
```

Avaliação do modelo

Com dados de treino

```
modelo_v1_pred_1 = predict(modelo_v1, newdata = dados_treino_final)
accuracy(modelo_v1_pred_1, dados_treino_final$PTIME)
```

```
##                       ME      RMSE       MAE      MPE     MAPE
## Test set -4.341561e-15 0.6680439 0.4746271 40.09682 107.2935
```

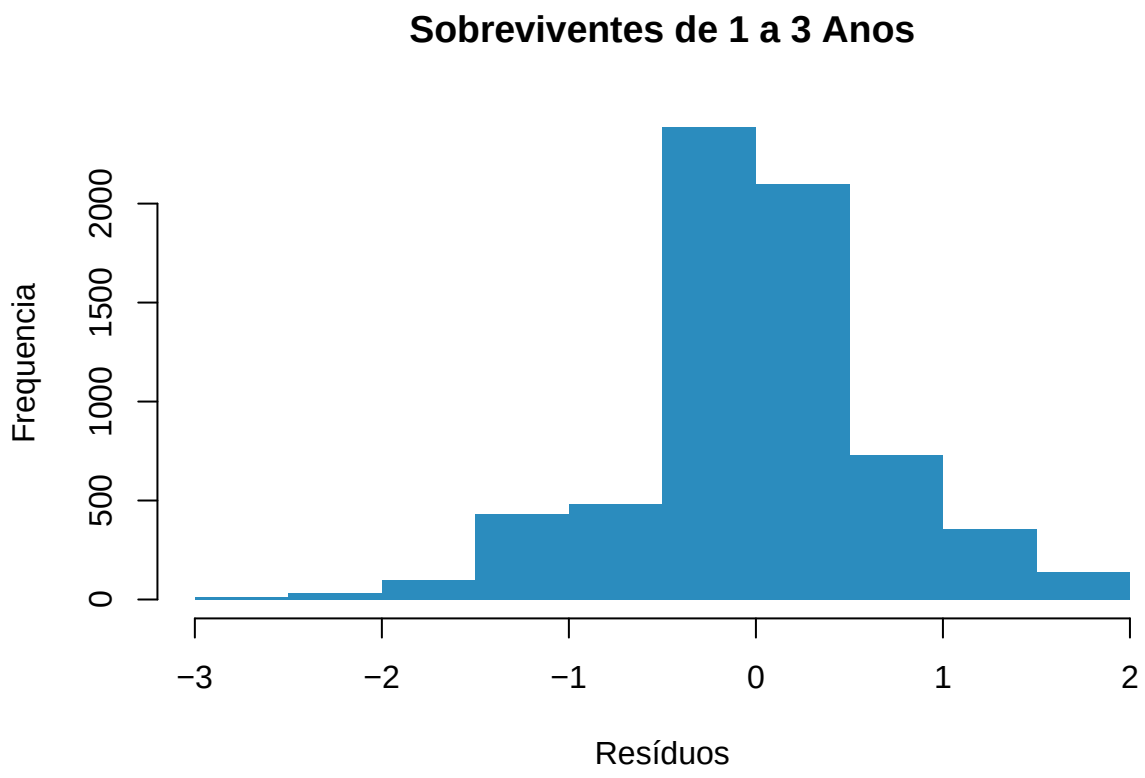
Com dados de teste

```
modelo_v1_pred_2 = predict(modelo_v1, newdata = dados_teste_final)
accuracy(modelo_v1_pred_2, dados_teste_final$PTIME)
```

```
##                     ME      RMSE       MAE      MPE     MAPE
## Test set 0.001993727 0.6657701 0.4676246 17.04116 111.9695
```

Distribuição do erro de validação

```
par(mfrow = c(1,1))
residuos <- dados_teste_final$PTIME - modelo_v1_pred_2
hist(residuos,
      xlab = "Resíduos", ylab = "Frequencia",
      main = "Sobreviventes de 1 a 3 Anos",
      col = c("#2b8cbe"),
      border = FALSE)
```



Conclusão: O modelo conseguil prever o tempo de sobrevivência dos pacientes 1 ano após receberem um transplante.