Análise Exploratória - Machine Learning

Aline SIlva Medeiros, Christian Jacobsen Teixeira, Marcos de Morais Silva

02/04/2022

Introdução

Objetivo do Modelo

O tempo de permanência é um indicador do tempo que o paciente fica internado. Por conta de alguma complicação, uma internação que seria breve se torna algo mais complexo. Dessa forma, o custo médico do paciente aumenta consideralvemente.

O objetivo desse projeto é identificar a probabilidade que um determinado pode fica internado de acordo com CID de alta.

Acurácia do modelo

Devido a questão da generalização, evitando problemas de underfitting (muito erros do modelo) e overfitting (modelo sensível a outliers), uma acurácia em torno de 90% seria o ideal para esse modelo. [bater o martelo com a equipe]

Dicionário de Dados

Para um melhor entendimento do modelo, segue uma definição dos atributos (variáveis) utilizados:

- CARATER - Eletivo ou Urgência (Categórica)
- AHRQ_DIAG_DTL_CATGY_CD - Código do Grupo CID (Categórica)
- tempo_permanencia = Data Internação - Data Alta (Numérica)
- IDADE - Idade do paciente (Numérica)
- GENERO - Sexo do paciente (Categórica)

Fonte dos Dados

- Fonte: AORTA (OPTUM)
- Período: Jan/2018 - Jan/2022

Carga e Prepação do RStudio

Diretório de trabalho

```
setwd('C:/FCD/Projeto_Machine_Learning')
getwd()
```

```
## [1] "C:/FCD/Projeto_Machine_Learning"
```

Pacotes utilizados

```
library(ggplot2)
library(dplyr)
library(readxl)
library(forcats)
library(rmarkdown)
library(rcompanion)
```

Carga do dataset

```
#Carga dos dados
BD <- read_excel(file.choose())

#Filtrando as colunas que serão usadas no modelo
col_names <- c('CARATER','AHRQ_DIAG_DTL_CATGY_CD','tempo_permanencia','IDADE', 'GENERO')
BD_MODELO <- BD[,col_names]
rm(col_names)

#Visualização prévia dos dados
head(BD_MODELO)
```

```
## # A tibble: 6 x 5
##   CARATER  AHRQ_DIAG_DTL_CATGY_CD tempo_permanencia IDADE GENERO
##   <chr>    <chr>                              <dbl> <dbl> <chr>
## 1 Elective 149                                    1    46 Feminino
## 2 Elective 149                                    1    29 Feminino
## 3 Elective 149                                    1    29 Feminino
## 4 Elective 149                                    0    41 Masculino
## 5 Elective 149                                    1    45 Feminino
## 6 Elective 149                                    1    30 Feminino
```

```
str(BD_MODELO)
```

```
## tibble [52,766 x 5] (S3: tbl_df/tbl/data.frame)
##  $ CARATER               : chr [1:52766] "Elective" "Elective" "Elective" "Elective" ...
##  $ AHRQ_DIAG_DTL_CATGY_CD: chr [1:52766] "149" "149" "149" "149" ...
##  $ tempo_permanencia     : num [1:52766] 1 1 1 0 1 1 1 1 2 0 ...
##  $ IDADE                 : num [1:52766] 46 29 29 41 45 30 25 47 47 39 ...
##  $ GENERO                : chr [1:52766] "Feminino" "Feminino" "Feminino" "Masculino" ...
```

```
summary(BD_MODELO)
```

```
##    CARATER          AHRQ_DIAG_DTL_CATGY_CD tempo_permanencia     IDADE
##  Length:52766       Length:52766           Min.   :  0.000   Min.   :  1.0
##  Class :character   Class :character       1st Qu.:  1.000   1st Qu.: 36.0
##  Mode  :character   Mode  :character       Median :  1.000   Median : 44.0
##                                            Mean   :  1.663   Mean   : 45.9
##                                            3rd Qu.:  1.000   3rd Qu.: 55.0
##                                            Max.   :181.000   Max.   :105.0
##     GENERO
##  Length:52766
##  Class :character
##  Mode  :character
##
```

##
##

Em relação as variáveis numéricas, tanto o tempo de permanência quando a idade indicam um distrubuição normal porque o valor da média e da mediana estão bem próximos. Contudo, será feito o teste de shapiro e um gráfico para comprovar essa hipótese.

Análise Exploratória dos Dados

Nesta seção é apresentado os resultados de uma análise exploratória para identificar importantes insights para os resultados do modelo.

Tempo Médio de Permanência por Gênero

Essa análise revela que o genêro é uma variável importante para o modelo porque é possível observar uma diferença clara entre a média do tempo de permanência entre mulheres e homens, este tem a maior média.

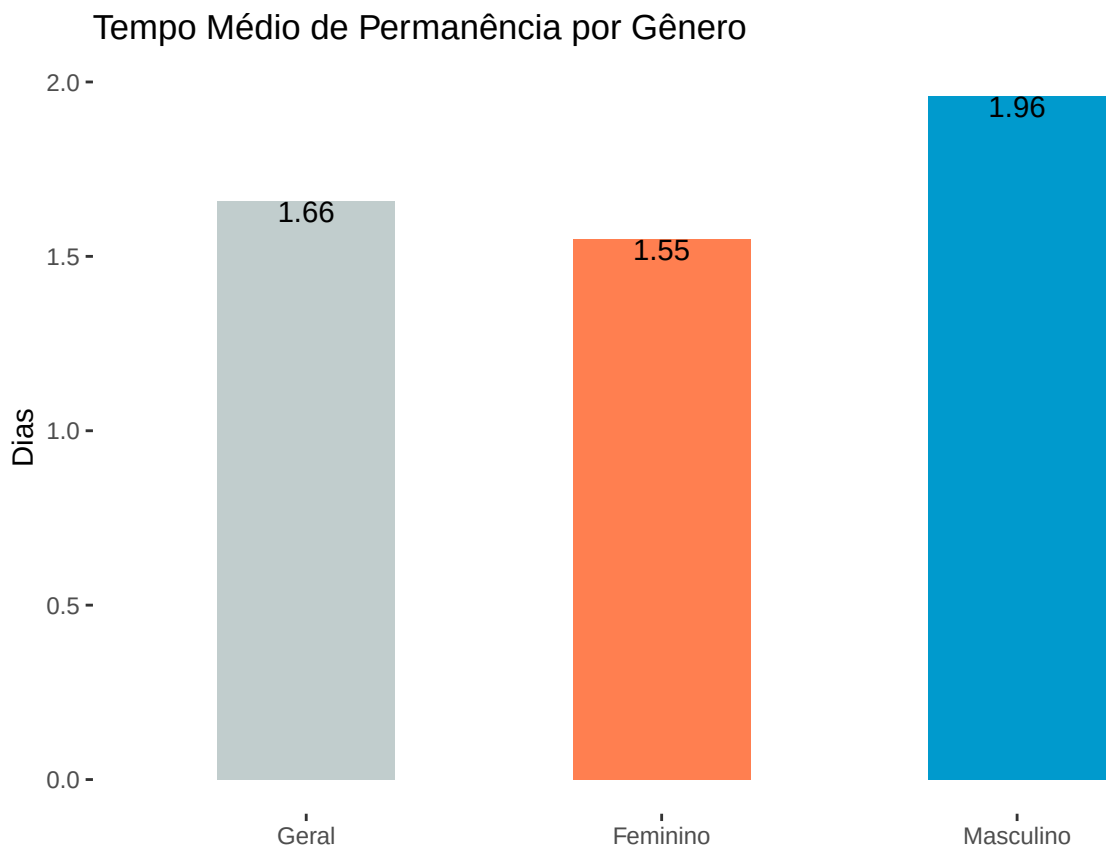
```
#Cálculo da média geral
media_geral <- mean(BD_MODELO$tempo_permanencia)
media_geral
```

```
## [1] 1.662851
```

```
#Média geral como dataframe
linha_geral <- data.frame('Geral', media_geral)

#função agregação do tempo média de permanência por genêro
media_genero <- aggregate(BD_MODELO$tempo_permanencia, list(BD_MODELO$GENERO), FUN = 'mean')
media_genero[3,] <- linha_geral
colnames(media_genero) <- c('Grupo', 'Media')
media_genero[,2] <- round(media_genero$Media, digits = 2)

#gráfico do tempo médio de permanência por genêro
ggplot(media_genero, aes(x = fct_relevel(Grupo, 'Geral'), y = Media)) +
  geom_col(position = 'dodge', width = 0.5, show.legend = F,
           fill = c('coral', 'deepskyblue3', 'azure3')) +
  theme(panel.background = element_blank()) +
  ggtitle('Tempo Médio de Permanência por Gênero') +
  ylab('Dias') +
  xlab(NULL)+
  geom_text(aes(label = Media, vjust = 'top'))
```

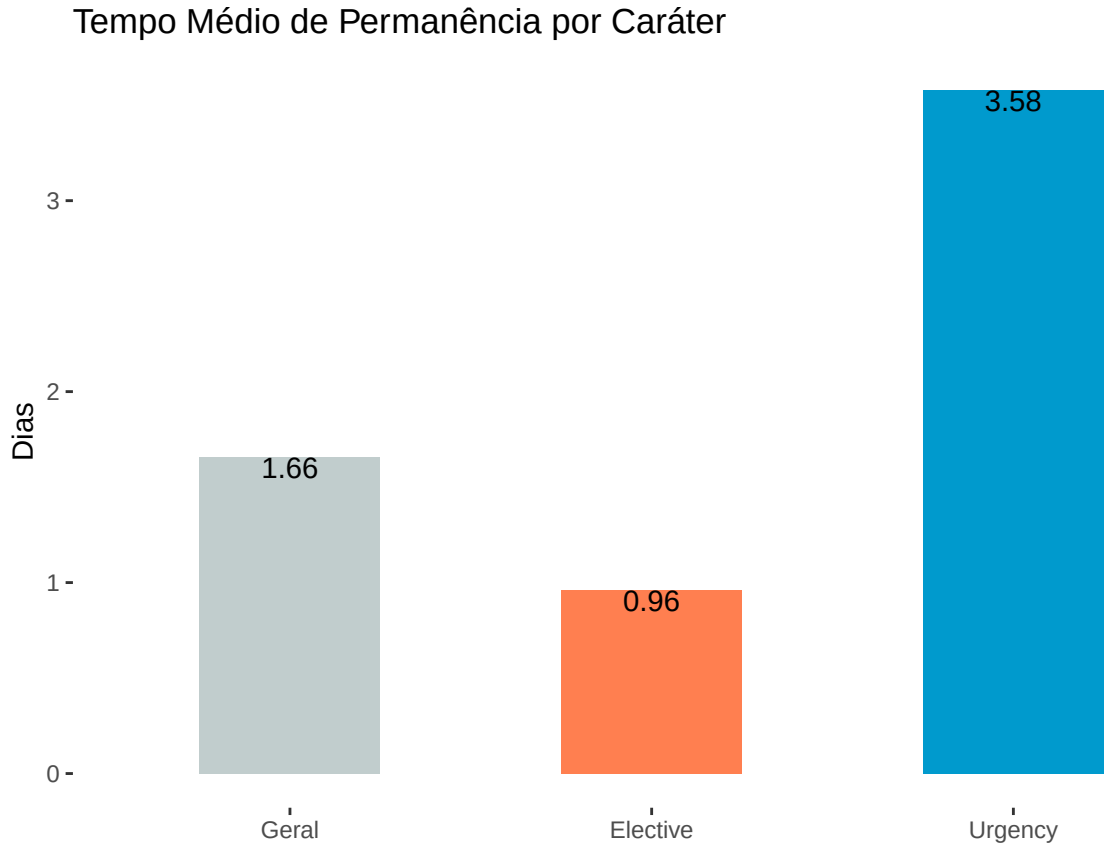


Tempo médio de permanência por caráter

Nesse cenário podemos observar que o tempo médio de permanência é muito afetado pelo caráter, fica claro a diferença entre a média geral e a média da urgência. Isso é um comportamento esperado visto que o paciente que entre no regime de urgência tende a ficar mais dias porque está mais sucetível a alguma complicação ao longo da internação.

```
#Função de agregação do tempo médio de permanência por caráter
media_carater <- aggregate(BD_MODELO$tempo_permanencia, list(BD_MODELO$CARATER), FUN = 'mean')
media_carater[3,] <- linha_geral
colnames(media_carater) <- c('Grupo', 'Media')
media_carater[,2] <- round(media_carater$Media, digits = 2)

#Gráfico do tempo de permanência por caráter
ggplot(media_carater, aes(x = fct_relevel(Grupo, 'Geral'), y = Media)) +
  geom_col(fill = c('coral', 'deepskyblue3', 'azure3'),
           position = 'dodge', width = 0.5, show.legend = F) +
  theme(axis.title.x = element_blank(), panel.background = element_blank()) +
  ggtitle('Tempo Médio de Permanência por Caráter') +
  ylab('Dias') +
  xlab(NULL) +
  geom_text(aes(label = Media, vjust = 'top'))
```



Idade dos pacientes

A variável da idade apresenta uma distribuição normal. Conforme visto anteriormente, a primeira evidência foi observada no sumário estatístico do dataset.

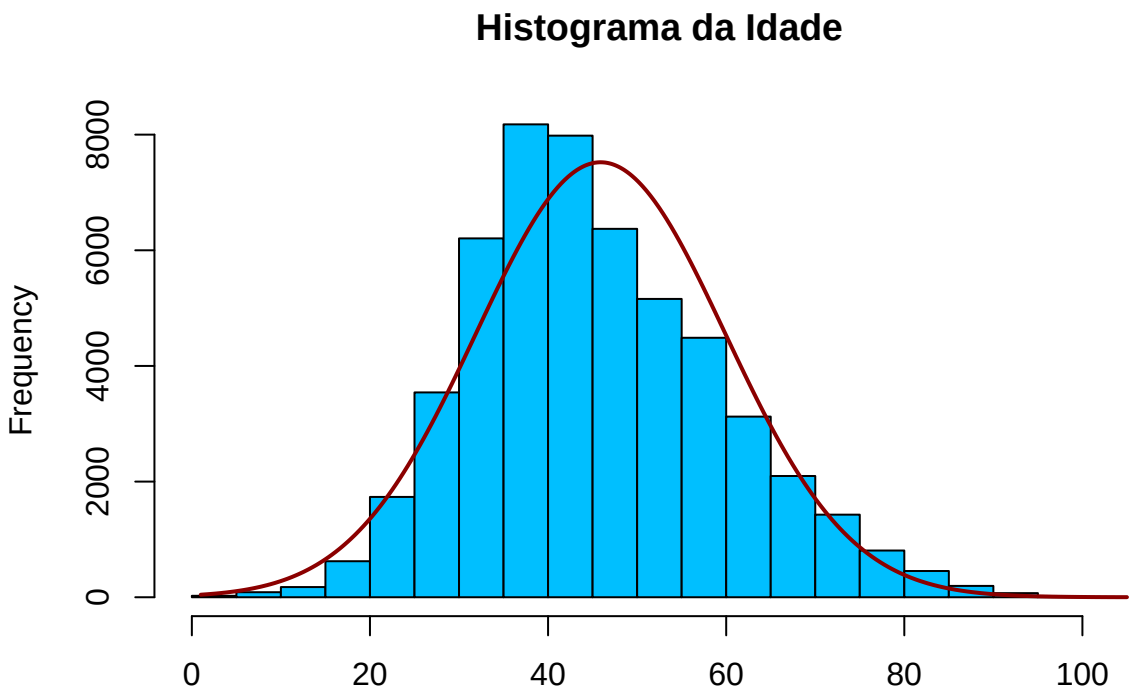
Nesta análise da idade, o histograma apresenta um comportamento esperado de uma distribuição normal.

Em relação ao teste de Shapiro, o valor de foi $p > 0.05$ indicando uma forte evidência de que os dados apresentam uma distruibuição normal.

Por fim, plot QQ do teste de normalidade mostra que os dados apresentam uma distribuição normal. A linha do gráfico representa, em teoria, como os dados deveriam estar dispostos caso a distruibuição fosse normal. Os círculos representam o valor observado no dataset e desta forma é observado que os pontos de dados acompanham a curvua normal.

Estes testes são fundamentais porque muitos modelos de Machine Learning aceitam apenas dados de uma distribuição normal. Caso contrário, é necesário usar modelos com métodos não-paramétricos.

```
# Histograma Idade
plotNormalHistogram(BD_MODELO$IDADE,
                    prob = F,
                    col = 'deepskyblue',
                    border = 'black',
                    main = 'Histograma da Idade',
                    linecol = 'darkred',
                    xlab = NULL,
                    lwd = 2)
```

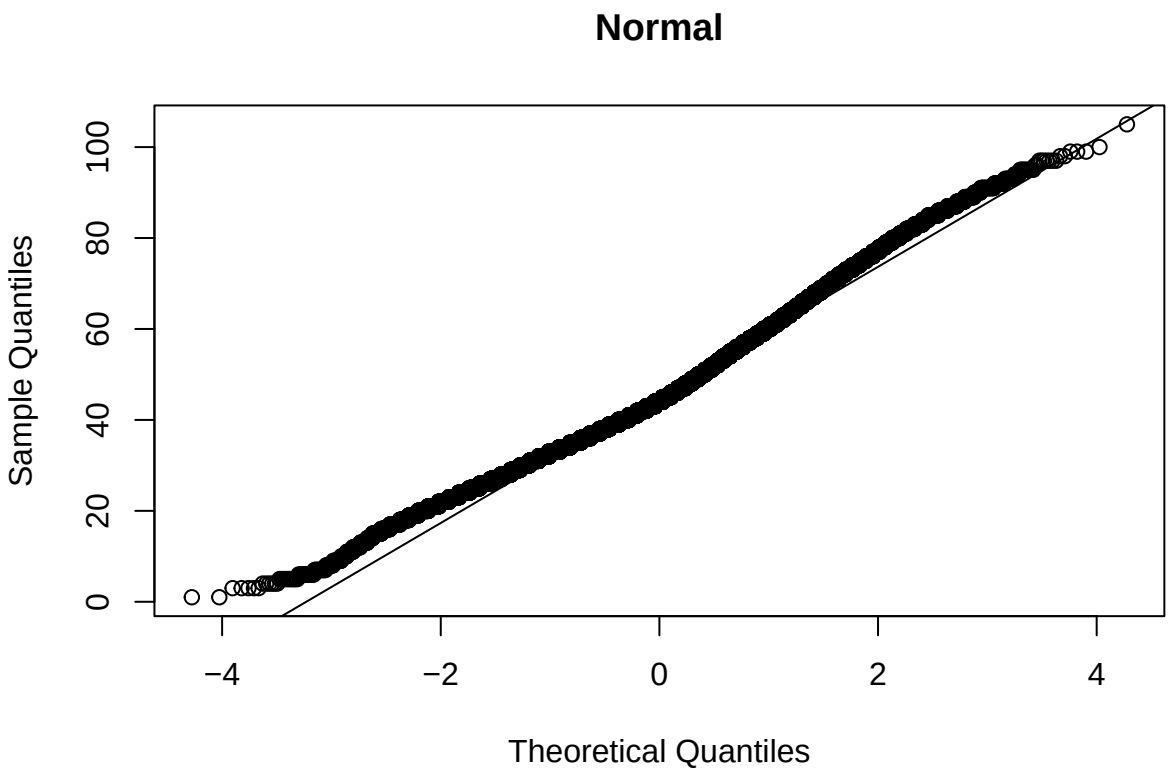



```
# Teste de Normalidade - Idade

# Teste Shapiro
# Foi usado 100 amostras aleatórios para o teste
teste_shapiro <- (sample(BD_MODELO$IDADE, 100))
shapiro.test(teste_shapiro)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  teste_shapiro
## W = 0.98582, p-value = 0.3631
```

```
# Plot teste normalidade
qqnorm(BD_MODELO$IDADE, main = 'Normal')
qqline(BD_MODELO$IDADE)
```

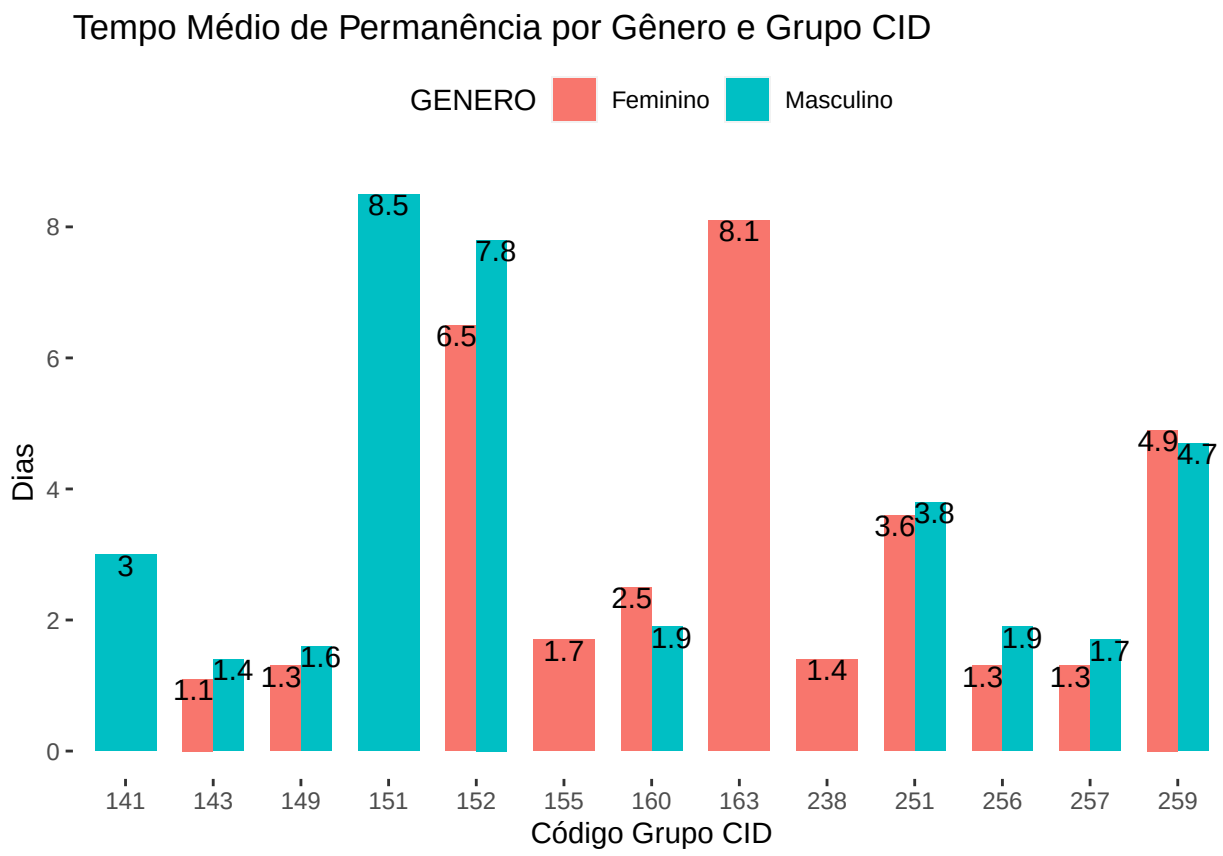


##Tempo médio de permanência por genêro e Grupo CID Neste gráfico é possível observar que os grupos de CID e genêro tem um comportamento onde as médias acompanham a mesma tendência do tempo médio de permanência, lembrando que esse gráfico representa o Top 10 por quantidade do grupo CID.

```
#Dataframe agrupado por genêro e código grupo CID
#Inclusão do tempo média de permanência e quantidade totais de dias (top 10)
#Classificação decrescente por quantidade
BD_Grafico <- BD_MODELO %>%
  group_by(GENERO, AHRQ_DIAG_DTL_CATGY_CD) %>%
  summarize(Media = round(mean(tempo_permanencia), digits = 1), Qtd = n()) %>%
  arrange(desc(Qtd)) %>%
  top_n(10)
```

```
## `summarise()` has grouped output by 'GENERO'. You can override using the
## `.groups` argument.
## Selecting by Qtd
```

```
#Plot do gráfico
View(BD_Grafico)
ggplot(BD_Grafico, aes(fill = GENERO, y = Media, x = AHRQ_DIAG_DTL_CATGY_CD)) +
  geom_bar(position = 'dodge', stat = 'identity', width = 0.7) +
  ggtitle('Tempo Médio de Permanência por Gênero e Grupo CID') +
  theme(legend.position = 'top', panel.background = element_blank()) +
  ylab('Dias') +
  xlab('Código Grupo CID') +
  geom_text(aes(label = Media, vjust = "top"), position = position_dodge(0.9))
```



Tempo médio de permanência por caráter e grupo CID

Neste gráfico é possível observar que o caráter ter uma influência bem maior no tempo de permanência dos pacientes. Podemos assumir que ao entrar em regime de urgência, a quantidade de dias internados aumenta bastante.

Esse gráfico demonstra uma visão bem interessante para tratamendo dos valores NAs. O gráfico mostra que pelo tempo médio de permanência podemos assumir se o caráter foi urgência ou eletivo.

```
#Dataframe agrupado por caráter e código grupo CID
#Inclusão do tempo média de permanência e quantidade totais de dias (top 10)
#Classificação decrescente por quantidade
BD_Grafico2 <- BD_MODELO %>%
  group_by(CARATER, AHRQ_DIAG_DTL_CATGY_CD) %>%
  summarize(Media = round(mean(tempo_permanencia), digits = 0), Qtd = n()) %>%
  arrange(desc(Qtd)) %>%
  top_n(10)
```

```
## `summarise()` has grouped output by 'CARATER'. You can override using the
## `.groups` argument.
## Selecting by Qtd
```

```
#Plot do gráfico
ggplot(BD_Grafico2, aes(fill = CARATER, y = Media, x = AHRQ_DIAG_DTL_CATGY_CD)) +
  geom_bar(position = 'dodge', stat = 'identity', width = 0.7) +
  ggtitle('Tempo Médio de Permanência por Caráter e Grupo CID') +
  ylab('Dias') +
  xlab('Código Grupo CID') +
```

```
  theme(legend.position = 'top', panel.background = element_blank()) +
  geom_text(aes(label = Media, vjust = "top", hjust = 'center'), position = position_dodge(1))
```