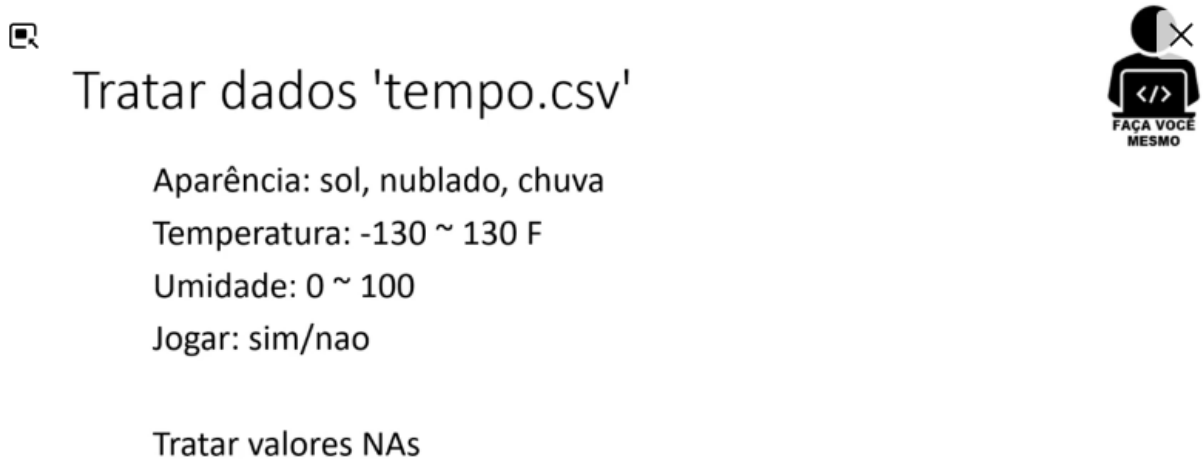
Limpeza de Dados em R - Atividade Prática

Igo da Costa Andrade

2024-07-12

Problema



Carregamento e Visualização dos dados

```
dados = read.csv("../dados/tempo.csv", sep=";", na.strings = "", stringsAsFactors = TRUE)
head(dados)
```

```
##   Aparencia Temperatura Umidade      Vento Jogar
## 1       sol          85      85      FALSO   nao
## 2       sol          80      90 VERDADEIRO   nao
## 3   nublado          83      86      FALSO   sim
## 4     chuva          70      NA      FALSO   sim
## 5     chuva          68      80      FALSO   sim
## 6     chuva          65      70 VERDADEIRO   nao
```

```
# Resumo Estatístico
summary(dados)
```

```
##    Aparencia  Temperatura         Umidade              Vento   Jogar
##  chuva  :5   Min.   :  64.00   Min.   : 65.00   FALSO     :7   nao:5
##  menos  :1   1st Qu.:  69.25   1st Qu.: 70.00   VERDADEIRO:6   sim:9
```

```
##  nublado:3   Median :  73.50   Median : 85.00   NA's      :1
##  sol    :5   Mean   : 155.57   Mean   : 89.77
##              3rd Qu.:  80.75   3rd Qu.: 90.00
##              Max.   :1220.00   Max.   :200.00
##                                NA's   :1
```

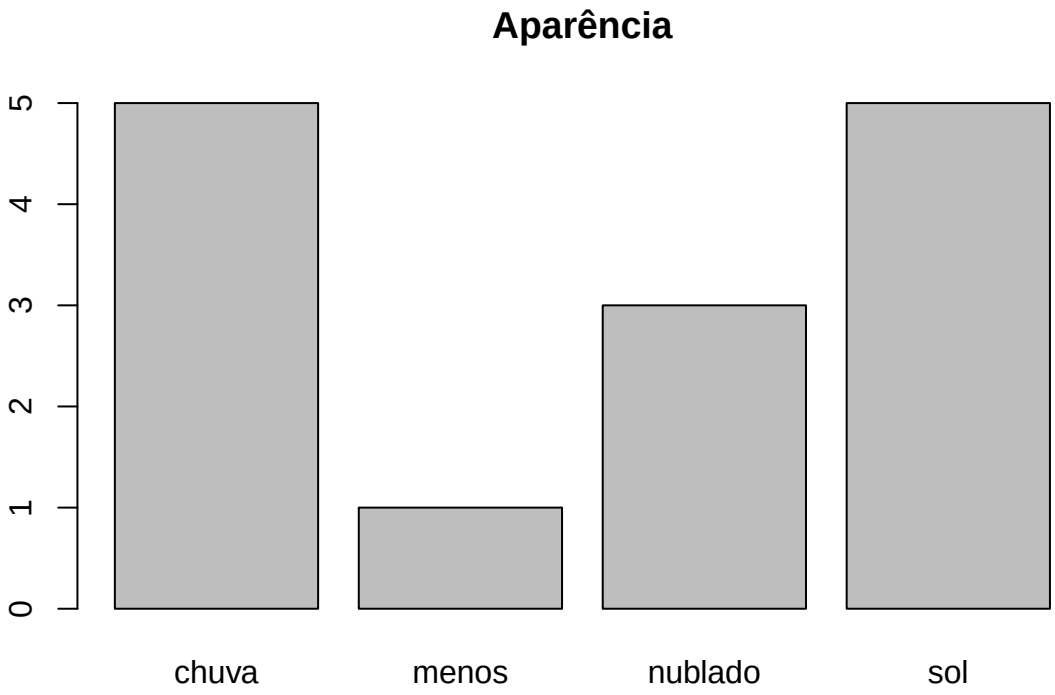
Análise Exploratória de Dados (EDA)

- Aparência

```
table(dados$Aparencia)
```

```
##
##   chuva   menos nublado     sol
##       5       1       3       5
```

```
barplot(table(dados$Aparencia), main="Aparência")
```



- Temperatura

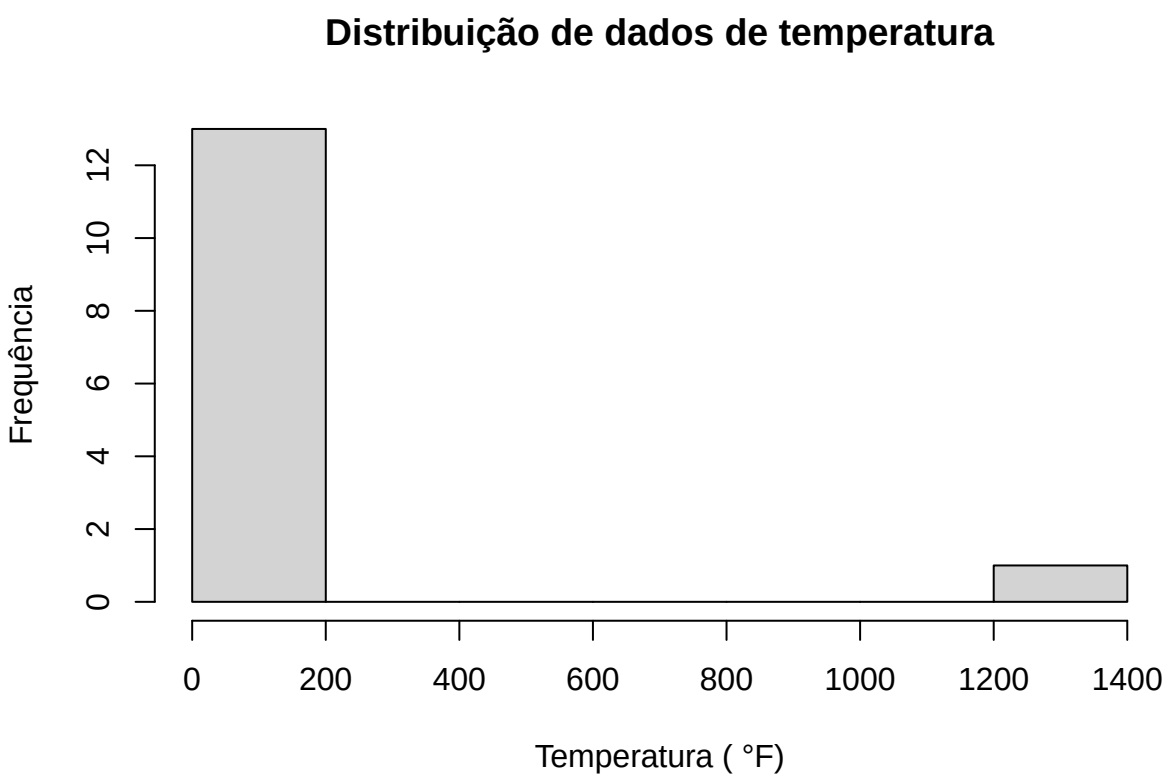
Resumo Estatístico

```
summary(dados$Temperatura)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   64.00   69.25   73.50  155.57   80.75 1220.00
```

```r
# Domínio
Tmin = -130
Tmax = 130
```

```r
hist(dados$Temperatura, main="Distribuição de dados de temperatura", xlab="Temperatura ( °F)", ylab="Fr
```

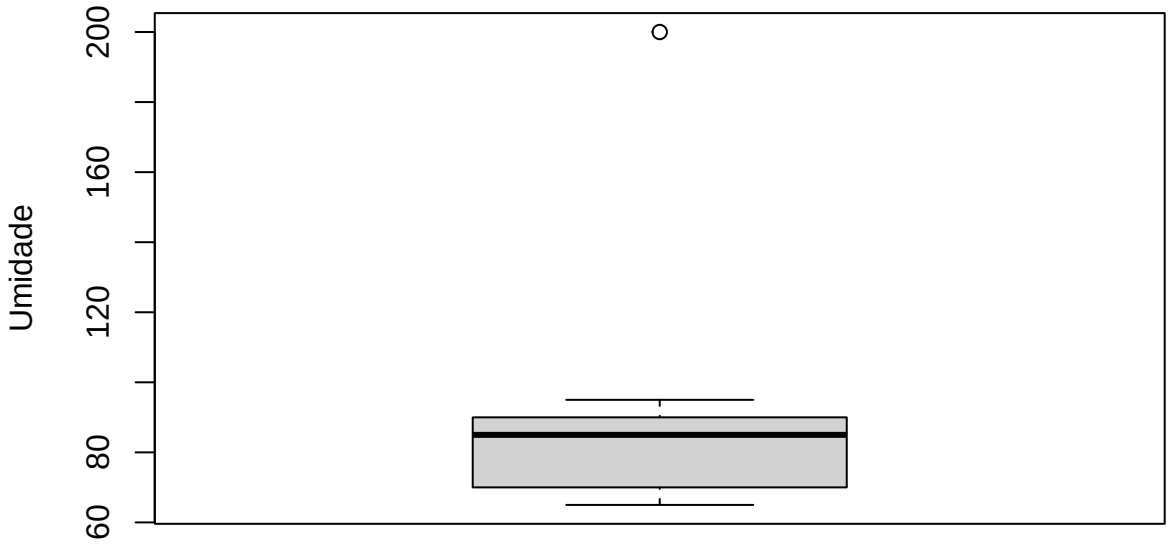


- Umidade

```r
# Resumo Estatístico
summary(dados$Umidade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   65.00   70.00   85.00   89.77   90.00  200.00       1
```

```r
# Boxplot
boxplot(dados$Umidade, ylab="Umidade")
```



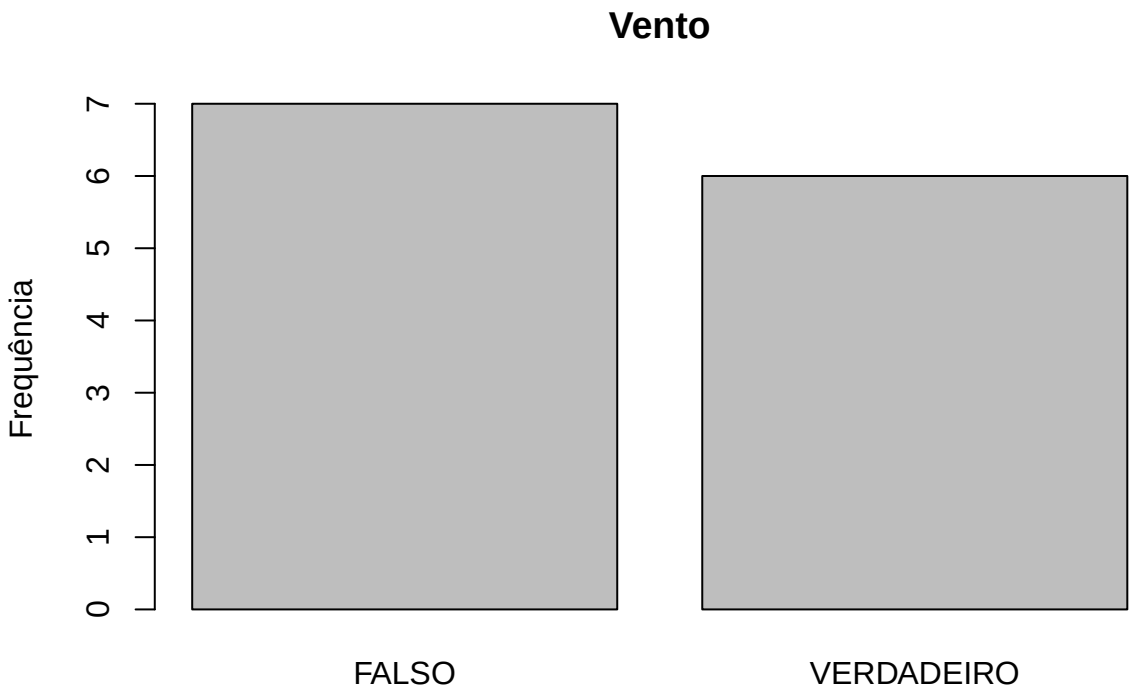

```
# Domínio
Umin = 0
Umax = 100
```

- Vento

```
summary(dados$Vento)
```

```
##      FALSO VERDADEIRO       NA's
##          7          6          1
```

```
barplot(table(dados$Vento), main="Vento", ylab="Frequência")
```

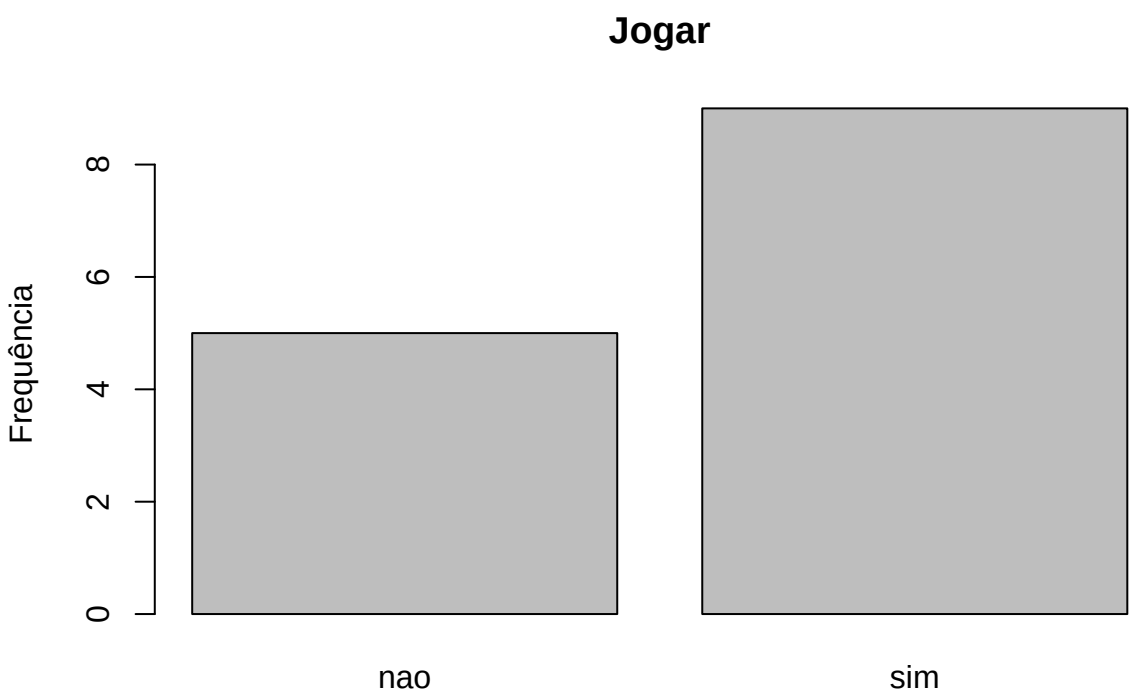


- Jogar

```
summary(dados$Jogar)
```

```
## nao sim
##   5   9
```

```
barplot(table(dados$Jogar), main="Jogar", ylab="Frequência")
```



Não existem valores faltantes nem ajustes necessários para a coluna Jogar.

Tratamento de dados

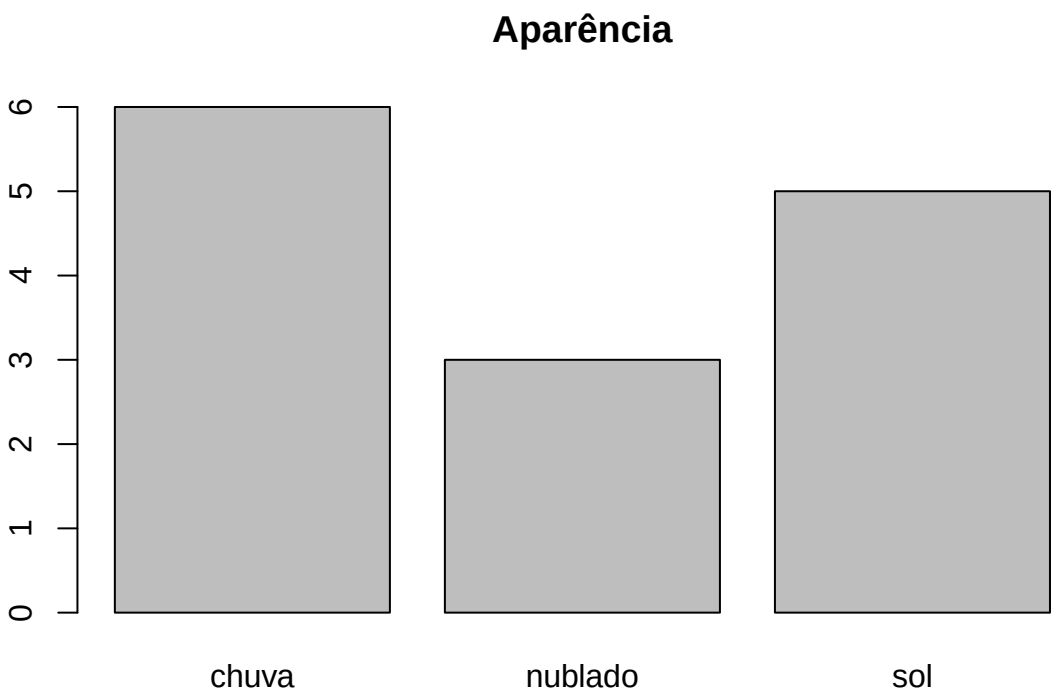
- Aparência

```
# Tabela de frequências
freq_table = table(dados$Aparencia)
# Índice da tabela com maior frequência
max_idx = which.max(freq_table)
# Nome correspondente ao índice de maior frequência
max_value_name = names(freq_table)[max_idx]
max_value_name
```

```
## [1] "chuva"
```

```
# Valores de domínio para a coluna Aparencia
dominio = c("sol", "nublado", "chuva")
# Substituição dos valores fora de domínio
dados[!dados$Aparencia %in% dominio, ]$Aparencia = max_value_name
```

```r
# Ajuste
dados$Aparencia = factor(dados$Aparencia)
# Resultado
barplot(table(dados$Aparencia), main="Aparência")
```



- Temperatura

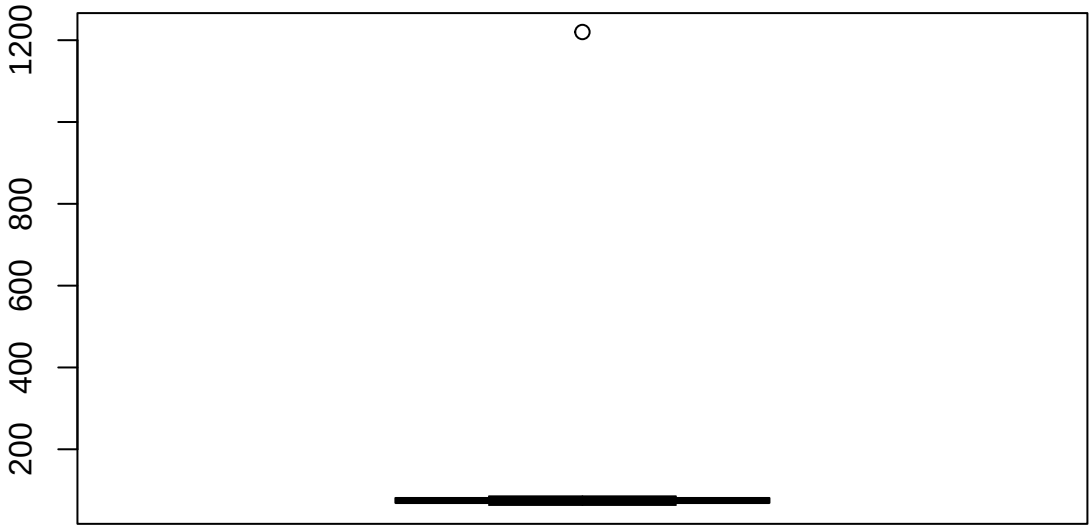
```r
summary(dados$Temperatura)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   64.00   69.25   73.50  155.57   80.75 1220.00
```

```r
# Domínio
Tmin = -130
Tmax = 130
```

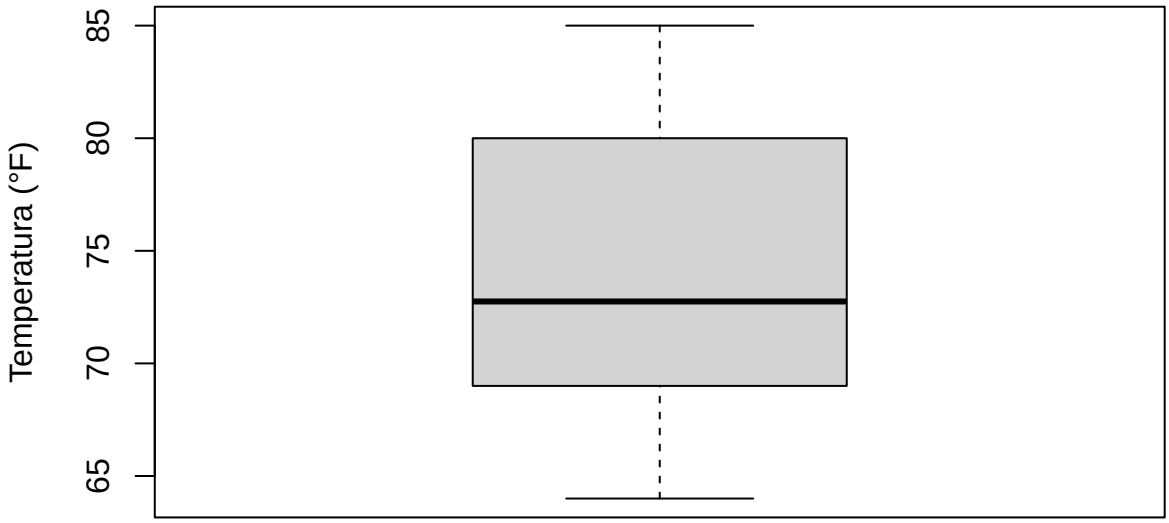
Dado que os valores de domínio da Temperatura estão entre -130°F e 130°F, existe 1 observações fora do domínio, conforme mostrado no boxplot abaixo:

```r
boxplot(dados$Temperatura)
```

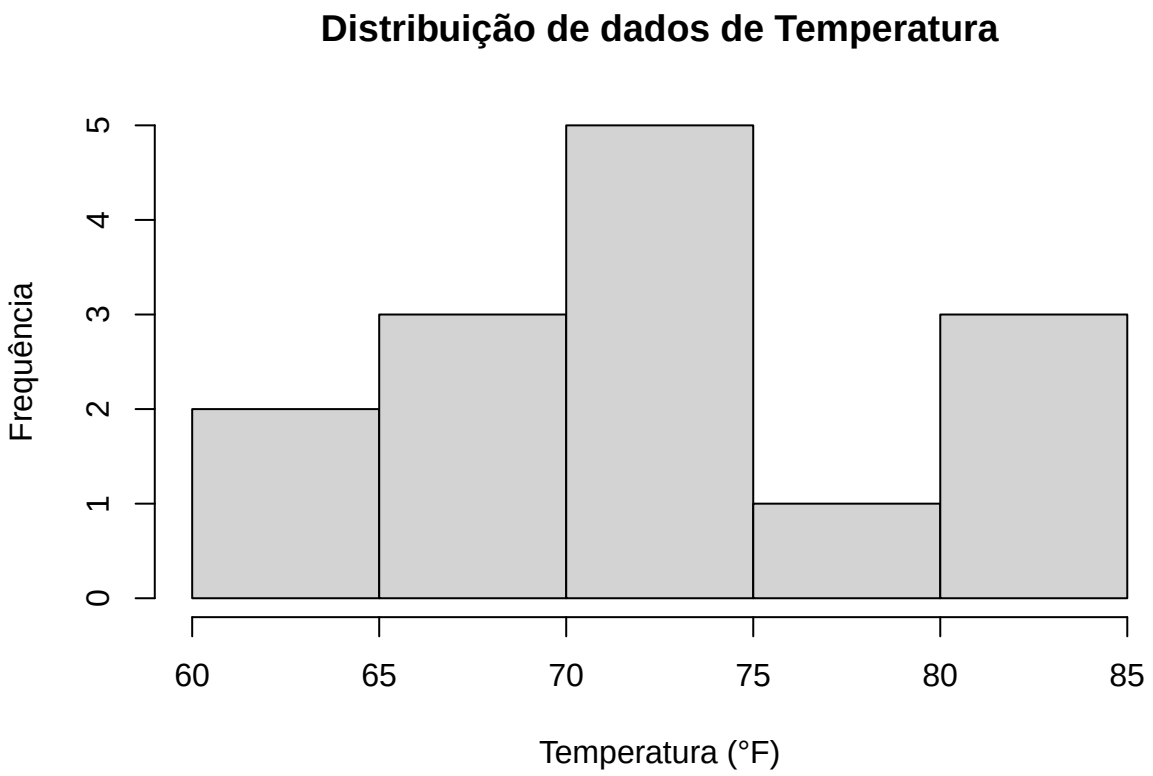


Substituiremos o valor fora de domínio pela mediana dos dados de temperatura:

```
dados[dados$Temperatura < Tmin | dados$Temperatura > Tmax, ]$Temperatura = median(dados$Temperatura, na
# Resultado
boxplot(dados$Temperatura, ylab="Temperatura (°F)")
```




```
hist(dados$Temperatura, main="Distribuição de dados de Temperatura", xlab="Temperatura (°F)", ylab="Fre
```

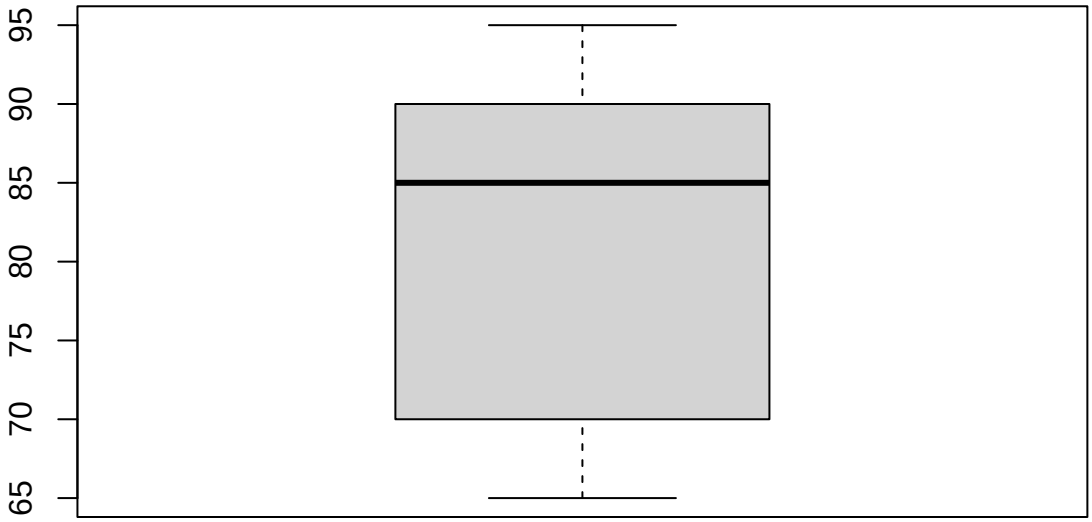


- Umidade

```
dados[dados$Umidade < Umin | dados$Umidade > Umax | is.na(dados$Umidade), ]$Umidade = median(dados$Umida
summary(dados$Umidade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.00   71.25   85.00   81.21   89.00   95.00
```

```
boxplot(dados$Umidade)
```



- Vento

Há 1 valor faltante na coluna de Vento. Vamos substituí-lo pelo valor mais frequente dessa coluna.

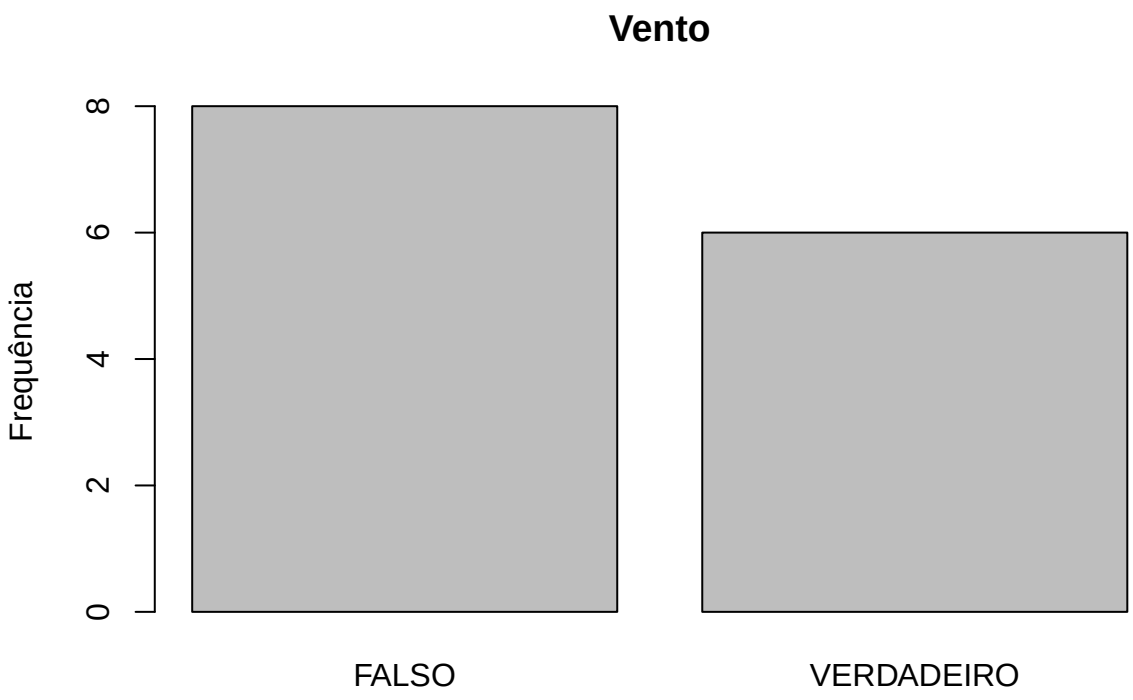
```
# Tabela de frequências
freq_table <- table(dados$Vento)
# Índice mais frequente
idx_max <- which.max(freq_table)
# Valor do índice mais frequente
idx_max_value <- names(freq_table)[idx_max]
idx_max_value
```

```
## [1] "FALSO"
```

```
# Substituição do valor faltante pelo valor mais frequente
dados[is.na(dados$Vento), ]$Vento = idx_max_value
# Resultado
summary(dados$Vento)
```

```
##      FALSO VERDADEIRO
##          8          6
```

```
# Barplot
barplot(table(dados$Vento), main="Vento", ylab="Frequência")
```



Resultado final

```
summary(dados)
```

```
##    Aparencia  Temperatura       Umidade            Vento   Jogar
##  chuva  :6   Min.   :64.00   Min.   :65.00   FALSO     :8   nao:5
##  nublado:3   1st Qu.:69.25   1st Qu.:71.25   VERDADEIRO:6   sim:9
##  sol    :5   Median :72.75   Median :85.00
##              Mean   :73.68   Mean   :81.21
##              3rd Qu.:78.75   3rd Qu.:89.00
##              Max.   :85.00   Max.   :95.00
```