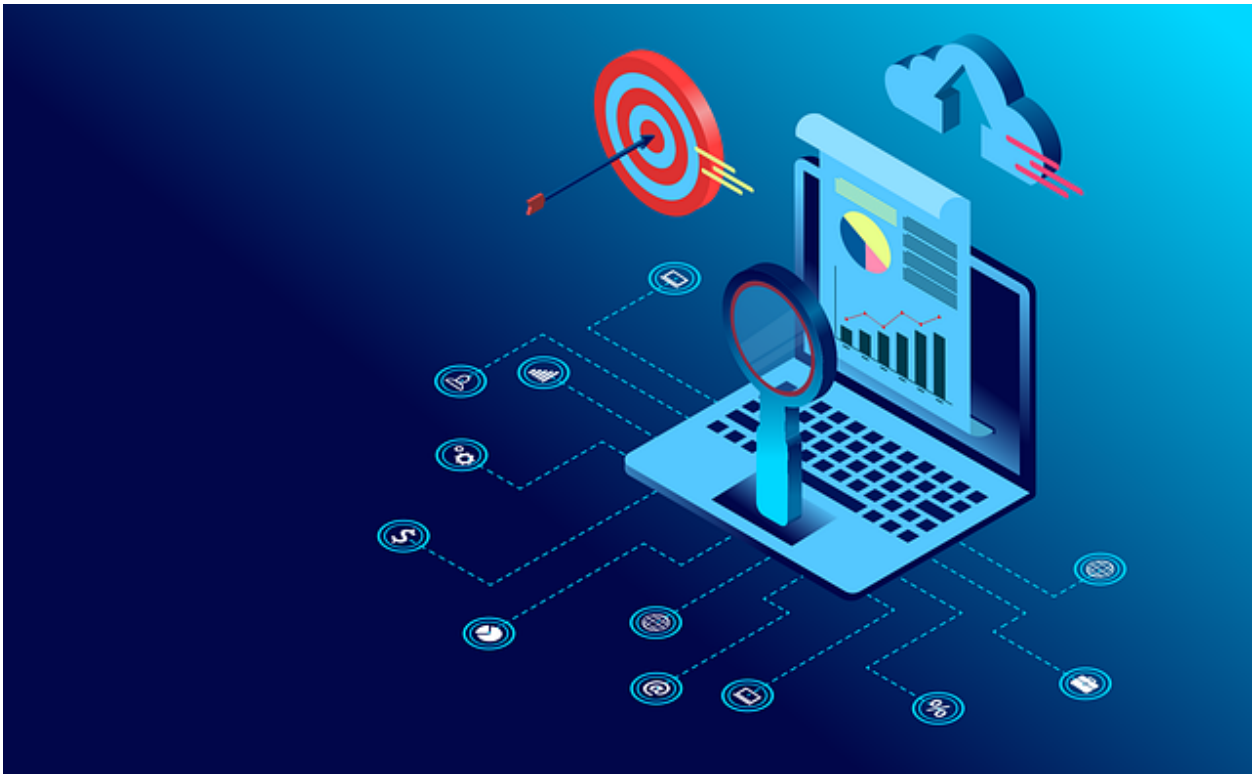
Limpeza e tratamento de dados

Wellington Moreira

2023-12-12



1 Conhecendo a base de dados

```
# Carregando o banco de dados
dados <- read.csv("../../data-cleaning/dados/Churn.csv", sep=";",stringsAsFactors = T ,na.strings = "")

# visualizando primeiras linhas
head(dados)
```

```
##   X0  X1 X2        X3 X4 X4.1       X6 X7 X8 X9      X10 X11
## 1  1 619 RS  Feminino 42    2        0  1  1  1 10134888   1
## 2  2 608 SC  Feminino 41    1  8380786  1  0  1 11254258   0
## 3  3 502 RS  Feminino 42    8  1596608  3  1  0 11393157   1
## 4  4 699 RS  Feminino 39    1        0  2  0  0  9382663   0
## 5  5 850 SC  Feminino 43    2 12551082  1  1  1   790841   0
## 6  6 645 SC Masculino 44    8 11375578  2  1  0 14975671   1
```

Para um melhor entendimento do dataset, renomeamos as variáveis conforme as regras do negócio.

```
# trocando nomes
colnames(dados) <- c(
  "Id","Score","Estado","Genero","Idade","Patrimonio","Saldo","Produtos","TemCartCredito","Ativo","Sala
)

# visualizando primeiras linhas
head(dados)
```

```
##   Id Score Estado    Genero Idade Patrimonio    Saldo Produtos TemCartCredito
## 1  1   619     RS  Feminino    42          2        0        1              1
## 2  2   608     SC  Feminino    41          1  8380786        1              0
## 3  3   502     RS  Feminino    42          8  1596608        3              1
## 4  4   699     RS  Feminino    39          1        0        2              0
## 5  5   850     SC  Feminino    43          2 12551082        1              1
## 6  6   645     SC Masculino    44          8 11375578        2              1
##   Ativo  Salario Saiu
## 1     1 10134888    1
## 2     1 11254258    0
## 3     0 11393157    1
## 4     0  9382663    0
## 5     1   790841    0
## 6     0 14975671    1
```

Agora vamos identificar os tipos de dados que temos neste dataset.

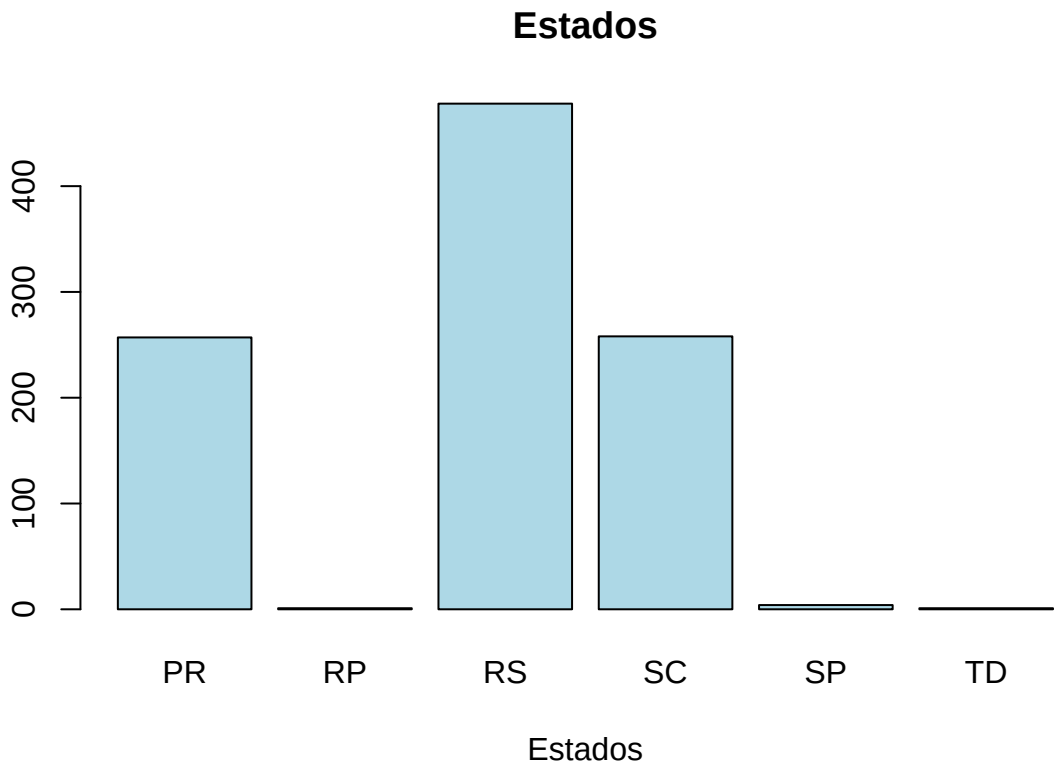
```
# Sumário Estatístico dos Dados
summary(dados)
```

```
##        Id              Score       Estado          Genero        Idade
##  Min.   :   1.0   Min.   :376.0   PR:257   F        :  2   Min.   :-20.0
##  1st Qu.: 251.5   1st Qu.:580.0   RP:  1   Fem      :  1   1st Qu.: 32.0
##  Median : 501.0   Median :653.0   RS:478   Feminino :461   Median : 37.0
##  Mean   : 500.9   Mean   :648.6   SC:258   M        :  6   Mean   : 38.9
##  3rd Qu.: 750.5   3rd Qu.:721.0   SP:  4   Masculino:521   3rd Qu.: 44.0
##  Max.   :1000.0   Max.   :850.0   TD:  1   NA's     :  8   Max.   :140.0
##
##    Patrimonio         Saldo             Produtos     TemCartCredito
##  Min.   : 0.000   Min.   :       0   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 2.000   1st Qu.:       0   1st Qu.:1.000   1st Qu.:0.0000
##  Median : 5.000   Median : 8958835   Median :1.000   Median :1.0000
##  Mean   : 5.069   Mean   : 7164928   Mean   :1.527   Mean   :0.7027
##  3rd Qu.: 8.000   3rd Qu.:12586844   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :10.000   Max.   :21177431   Max.   :4.000   Max.   :1.0000
##
##      Ativo           Salario              Saiu
##  Min.   :0.0000   Min.   :9.677e+03   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:3.029e+06   1st Qu.:0.0000
##  Median :1.0000   Median :8.703e+06   Median :0.0000
##  Mean   :0.5095   Mean   :3.529e+07   Mean   :0.2032
##  3rd Qu.:1.0000   3rd Qu.:1.405e+07   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.193e+10   Max.   :1.0000
##                   NA's   :7
```

2 Explorando os dados

2.1 *Estados*

```
options(width = 300)
counts <- table(dados$Estado)
barplot(counts, main="Estados", xlab="Estados", col="lightblue", border = "black")
```



Podemos ver visualmente que existem dados com valores fora do domínio pré-estabelecido pelas regras de negócio.

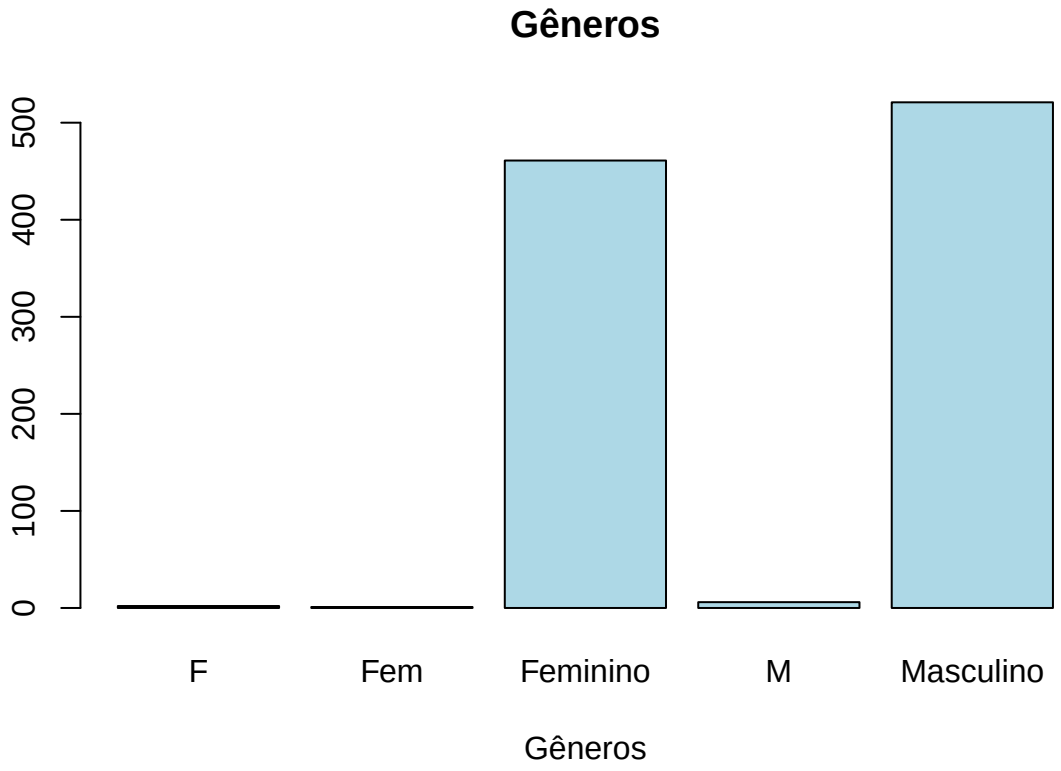
Como os dados são para a Região Sul do Brasil, vemos aqui a distribuição destes dados.

```
summary(dados$Estado)
```

```
##  PR  RP  RS  SC  SP  TD
## 257   1 478 258   4   1
```

2.2 *Gênero*

```
counts <- table(dados$Genero)
barplot(counts, main="Gêneros", xlab="Gêneros", col="lightblue")
```



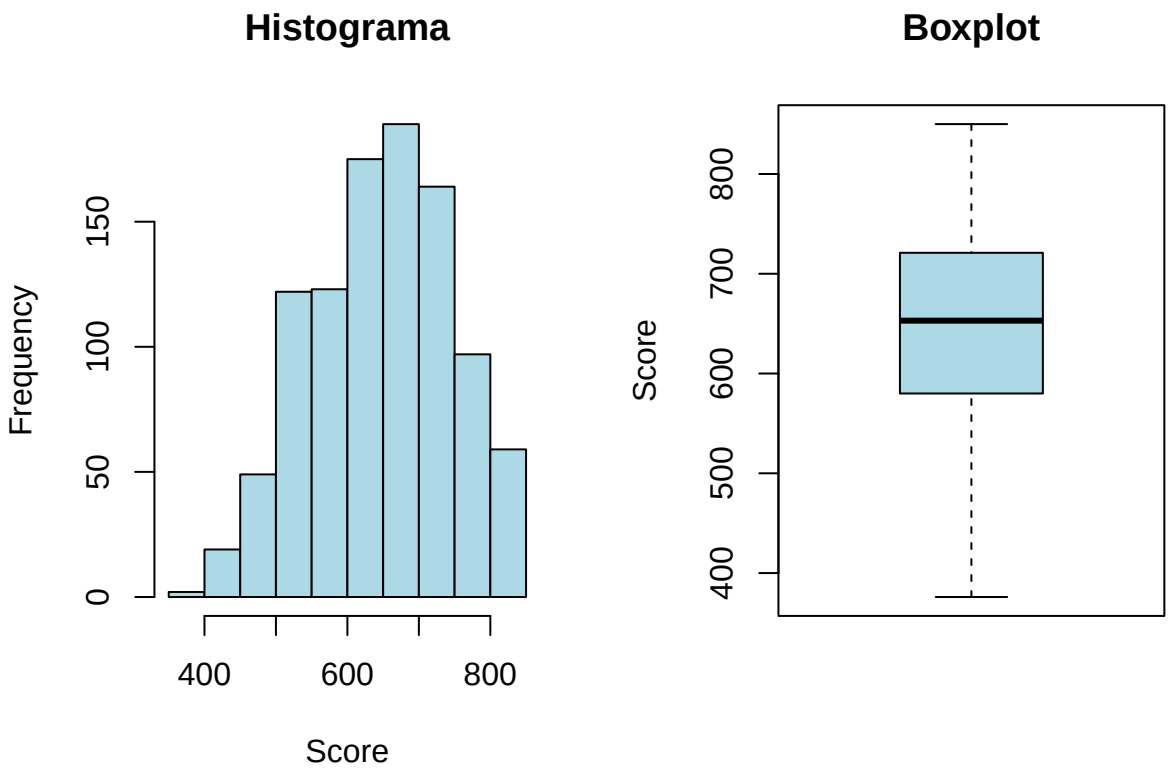
Aqui percebemos a falta de padronização dos dados e a existência de valores faltantes.

```
summary(dados$Genero)
```

```
##         F       Fem  Feminino         M Masculino      NA's
##         2         1       461         6       521         8
```

2.3 *Score*

```
par(mfrow = c(1, 2))
hist(dados$Score, main="Histograma",xlab = "Score", col = "lightblue", border = "black")
boxplot(dados$Score, main = "Boxplot", ylab = "Score", col = "lightblue", border = "black")
```



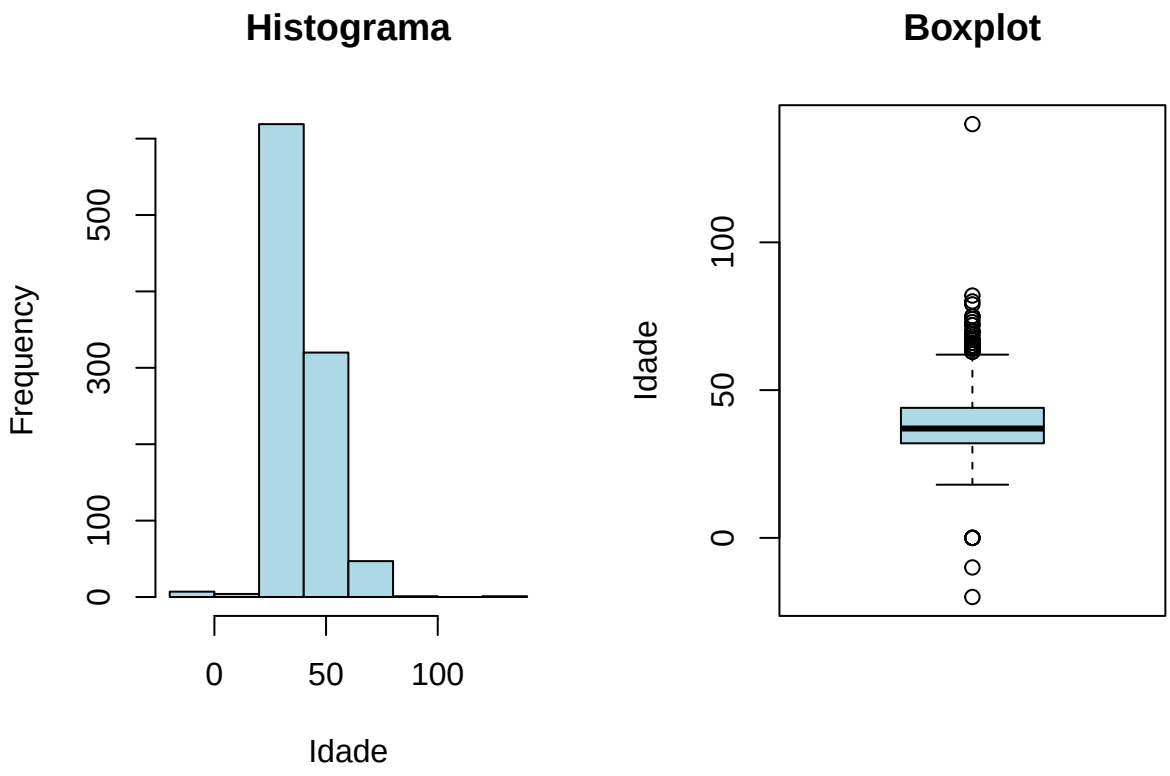
Aqui temos dados sem inconsistências conforme as regras de negócio.

```
summary(dados$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   376.0   580.0   653.0   648.6   721.0   850.0
```

2.4 *Idades*

```
par(mfrow = c(1, 2))
hist(dados$Idade, main="Histograma", xlab="Idade", col="lightblue", border="black")
boxplot(dados$Idade, main="Boxplot", ylab="Idade", col="lightblue", border="black")
```



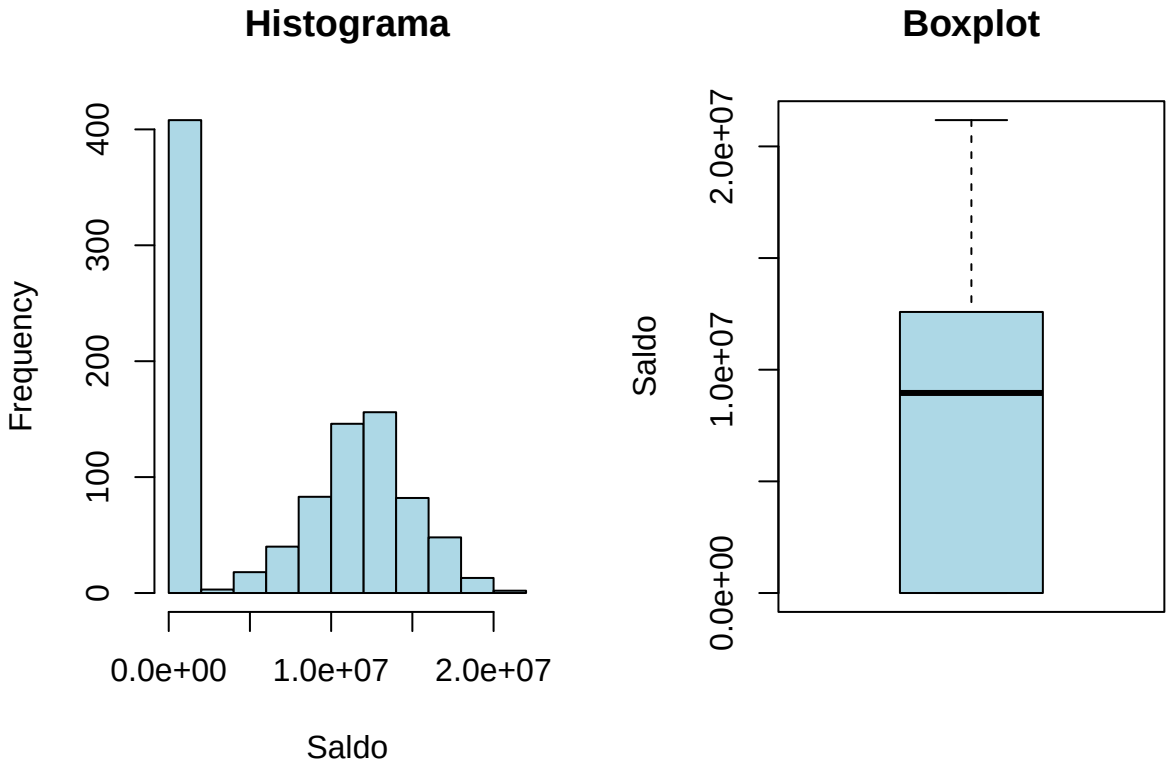
Aqui percebemos que existem valores de idade fora do domínio estabelecidos pelas regras de negócio, podemos perceber idades negativas e muito acima de 100 anos.

```
summary(dados$Idade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -20.0    32.0    37.0    38.9    44.0   140.0
```

2.5 *Saldo*

```
par(mfrow = c(1, 2))
hist(dados$Saldo, main="Histograma", xlab="Saldo", col="lightblue", border="black")
boxplot(dados$Saldo, main="Boxplot", ylab="Saldo", col="lightblue", border="black")
```



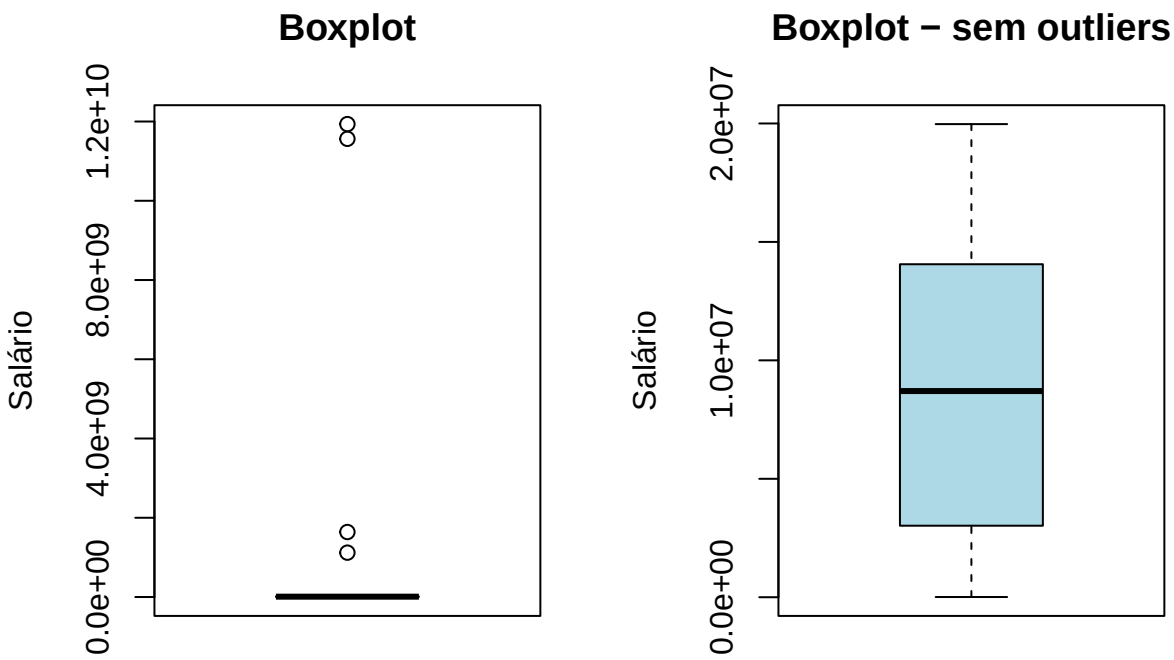
Aqui temos dados sem inconsistências conforme as regras de negócio.

```
summary(dados$Saldo)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0        0  8958835  7164928 12586844 21177431
```

2.6 *Salários*

```
par(mfrow = c(1, 2))
boxplot(dados$Salario, main="Boxplot", ylab="Salário", col="lightblue", border="black")
boxplot(dados$Salario, main="Boxplot - sem outliers", ylab="Salário", col="lightblue", border="black", c
```



Podemos perceber inconsistência de dados com valores faltantes e dispersão de dados com outliers conforme métricas estabelecidas pelas regras de negócio

```
summary(dados$Salario)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 9.677e+03 3.029e+06 8.703e+06 3.529e+07 1.405e+07 1.193e+10         7
```

2.7 *Duplicidade*

```
x <-  dados[duplicated(dados$Id),]
x
```

```
##    Id Score Estado   Genero Idade Patrimonio   Saldo Produtos TemCartCredito Ativo  Salario Saiu
## 82 81   665     RS Feminino    34          1 9664554        2              0     0 17141366    0
```

Podemos pelo identificador perceber que existe um dado duplicado na base de dados.

2.8 *Valores faltantes*

```
dados[!complete.cases(dados),]
```

```
##       Id Score Estado    Genero Idade Patrimonio    Saldo Produtos TemCartCredito Ativo  Salario Saiu
## 18    18   549     SC  Feminino    24          9        0        2              1     1       NA    0
## 53    53   788     RS  Feminino    33          5        0        2              0     0       NA    0
## 65    65   603     PR      <NA>    26          4 10916637        1              1     1  9284067    0
## 85    86   493     RS      <NA>    46          4        0        2              1     0   190766    0
## 180 181   754     PR  Feminino    55          3 16160881        1              1     0       NA    1
## 214 215   676     RS      <NA>    34          1  6309501        1              1     1  4064581    0
## 297 298   714     RS      <NA>    31          4 12516926        1              1     1 10663689    0
## 331 332   656     RS Masculino    50          7        0        2              0     1       NA    0
## 371 372   801     SC      <NA>    42          4 14194767        1              1     1  1059829    0
## 427 428   492     PR Masculino    39         10 12457665        2              1     0       NA    0
## 502 503   692     RS      <NA>    54          5        0        2              1     1  8872184    0
## 551 552   721     PR  Feminino    36          3  6525307        2              1     0       NA    0
## 964 965   529     SC      <NA>    63          4  9613411        3              1     0 10873296    1
## 970 971   649     PR      <NA>    70          9 11685471        2              0     1 10712579    0
## 984 985   614     PR  Feminino    35          6 12810028        1              0     0       NA    1
```

Aqui pegamos todos os registros que contém dados faltantes no dataset.

---

3 Tratamento de dados

Aqui iniciamos o tratamento com base nas inconsistências encontradas:

1. **Estados** apresenta dados fora de domínio.
2. **Gênero** contém dados faltantes e falta de padronização.
3. **Idades** apresenta valores fora de domínio.
4. **Salário** contém dados faltantes e outliers.
5. **Duplicidades** apresenta uma duplicata na base de dados.
6. **Valores faltantes** a base de dados apresenta registros com valores faltantes.

3.1 Estados

```
summary(dados$Estado)
```

```
##  PR  RP  RS  SC  SP  TD
## 257   1 478 258   4   1
```

Sendo um dado categórico, uma estratégia é utilizar a moda (valor que mais se repete) para atribuir ou substituir valores. E assim, vamos pegar todos os valores que não estão no vetor e os substituir pela moda:

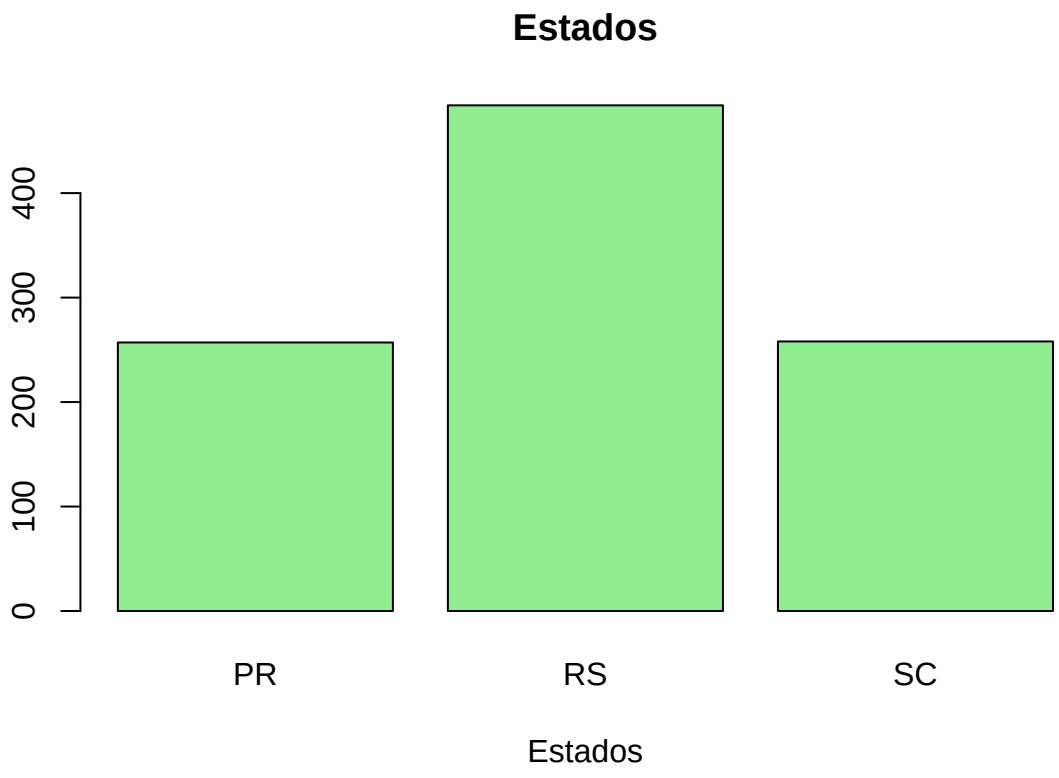
```
dados[!dados$Estado %in% c("RS","SC","PR"),]$Estado <- "RS"
```

Feito isso, vamos excluir os levels substituidos e confirmar as modificações feitas.

```
dados$Estado <- factor(dados$Estado)
summary(dados$Estado)
```

```
##  PR  RS  SC
## 257 484 258
```

```
counts <- table(dados$Estado)
barplot(counts, main="Estados", xlab="Estados", col="lightgreen", border = "black")
```



3.2 Gênero

Assim como *Idades*, esta variável também é categórica e seu tratamento será o mesmo, onde também utilizaremos da moda como valor para padronização.

```
summary(dados$Genero)
```

```
##         F       Fem  Feminino         M Masculino      NA's
##         2         1       461         6       521         8
```

Podemos perceber que a moda destes dados é *Masculino e assim,* padronizaremos *F* e *Fem* para *Feminino, M* e *NAs* para *Masculino.*

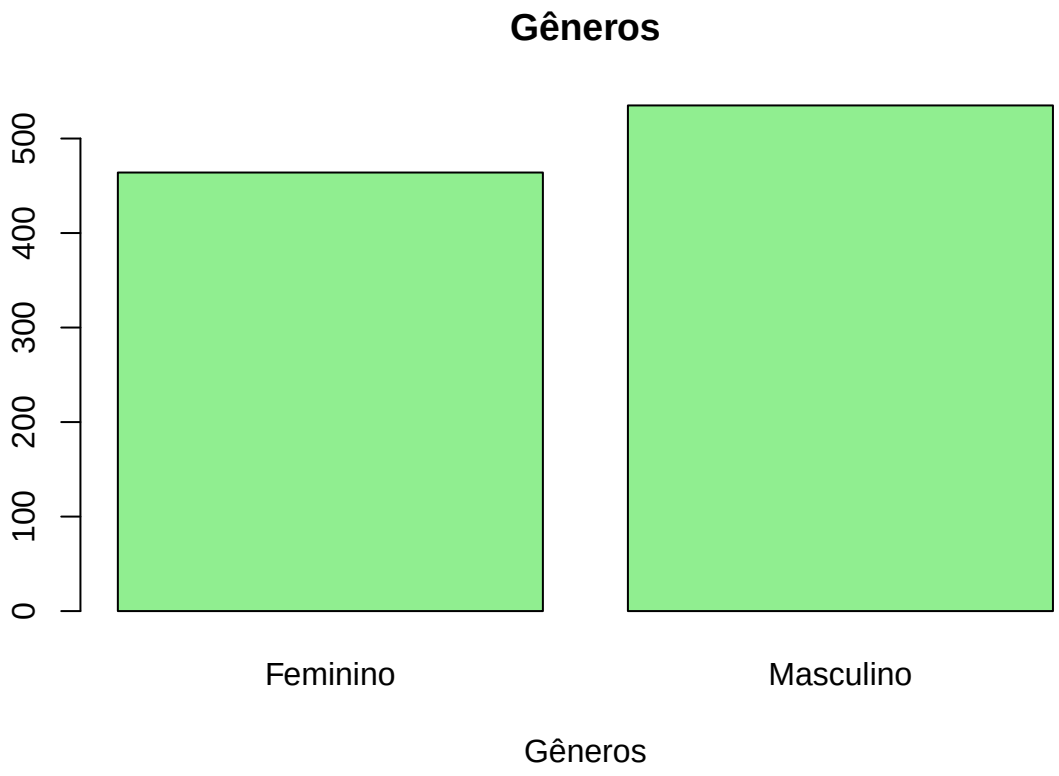
```
dados[is.na(dados$Genero) | dados$Genero == "M" ,]$Genero <- "Masculino"
dados[dados$Genero == "F" | dados$Genero == "Fem", ]$Genero <- "Feminino"
```

Removemos os levels não mais utilizados e os conferimos novamente.

```
dados$Genero =   factor(dados$Genero)
summary(dados$Genero)
```

```
##  Feminino Masculino
##       464       535
```

```
counts <- table(dados$Genero)
barplot(counts, main="Gêneros", xlab="Gêneros", col="lightgreen")
```



3.3 Idades

Aqui iniciamos o tratamento de variáveis numéricas, sendo uma estratégia a utilização da mediana.

```
summary(dados$Idade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -20.0    32.0    37.0    38.9    44.0   140.0
```

Conforme as regras de negócio buscamos os valores fora do domíno.

```
dados[dados$Idade<0 | dados$Idade>110 ,]$Idade
```

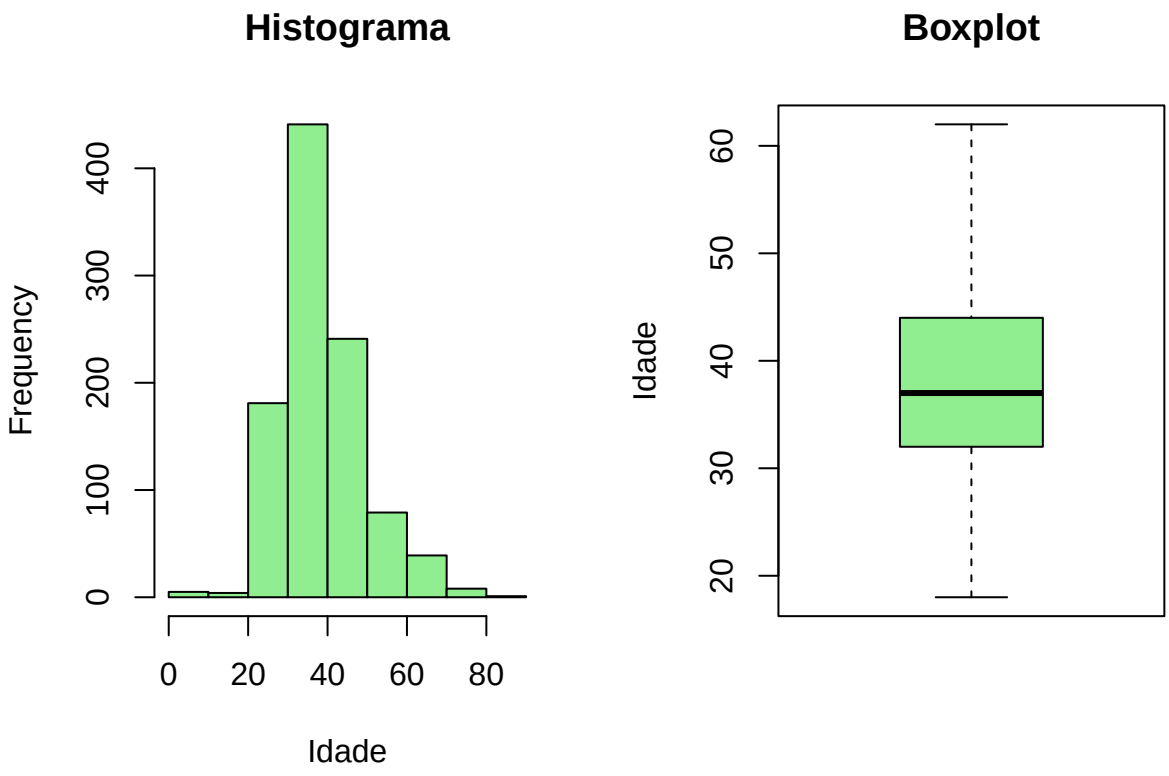
```
## [1] -10 -20 140
```

Após regra estabelecida, iniciamos a substituição e faremos novamente uma verificação.

```
dados[dados$Idade<0 | dados$Idade>110 ,]$Idade <- median(dados$Idade)
summary(dados$Idade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    32.0    37.0    38.9    44.0    82.0
```

```
par(mfrow = c(1, 2))
hist(dados$Idade, main="Histograma", xlab="Idade", col="lightgreen", border="black")
boxplot(dados$Idade, main="Boxplot", ylab="Idade", col="lightgreen", border="black", outline=F)
```



3.4 Salários

Conforme relatado, trataremos aqui dados faltantes, e também utilizaremos a mediana para esta variável numérica.

```
summary(dados$Salario)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 9.677e+03 3.029e+06 8.703e+06 3.529e+07 1.405e+07 1.193e+10         7
```

Substituindo NAs pela mediana e após verficando novamente os dados.

```
dados[is.na(dados$Salario),]$Salario <- median(dados$Salario,na.rm = T)
dados[!complete.cases(dados$Salario),]
```

```
##  [1] Id             Score          Estado         Genero         Idade          Patrimonio     Saldo
## <0 linhas> (ou row.names de comprimento 0)
```

Iniciamos agora o tratamento de outliers, uma estratégia é utilizar o desvio padrão como métrica de comparação. Aqui vemos valores que passam em duas vezes o valor do desvio.

```
desv <- sd(dados$Salario, na.rm = T)
desv
```

```
## [1] 528720617
```

```
dados[dados$Salario >= 2 *desv  , ]$Salario
```

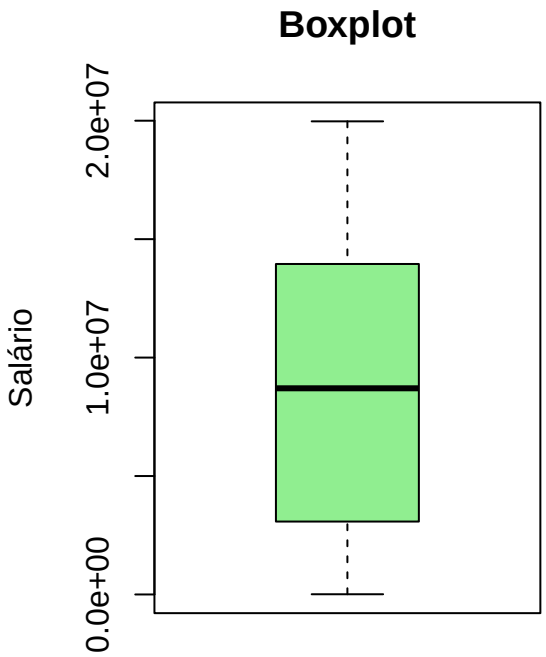
```
## [1] 11934688000 11563829000  1640178900  1119811900
```

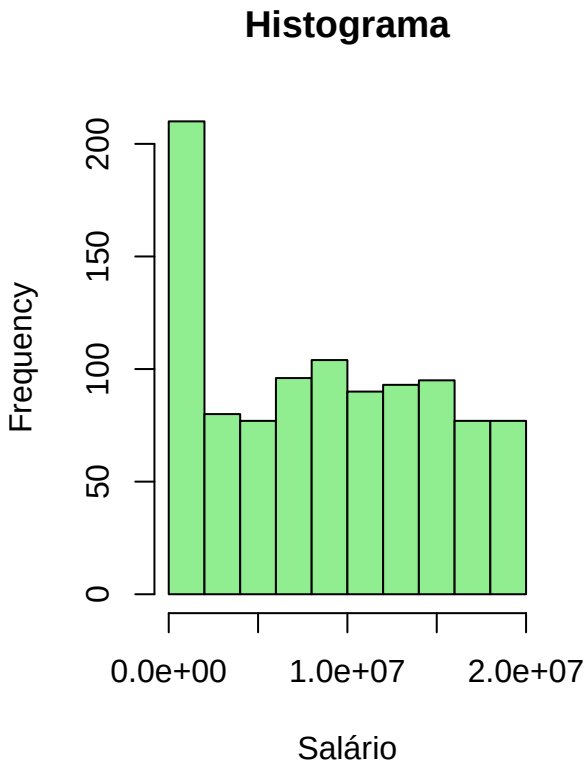
Para estes, utilizamos a mediana para reatribuição e ao final conferimos novamente a base de dados.

```
dados[dados$Salario >= 2 *desv  , ]$Salario <- median(dados$Salario)
dados[dados$Salario >= 2 *desv  , ]$Salario
```

```
## numeric(0)
```

```
par(mfrow = c(1, 2))
boxplot(dados$Salario, main="Boxplot", ylab="Salário", col="lightgreen", border="black")
hist(dados$Salario, main="Histograma", xlab="Salário", col="lightgreen", border="black")
```


Boxplot
Salário
0.0e+00
1.0e+07
2.0e+07


Histograma
Frequency
0
50
100
150
200
0.0e+00
1.0e+07
2.0e+07
Salário

3.5 Duplicidade

Como relatado a base apresenta uma duplicidade a qual podemos identificar pelo *ID*

```
x <-  dados[duplicated(dados$Id),]
x
```

```
##    Id Score Estado   Genero Idade Patrimonio   Saldo Produtos TemCartCredito Ativo  Salario Saiu
## 82 81   665     RS Feminino    34          1 9664554        2              0     0 17141366    0
```

Faremos sua exclusão pelo indice 82 e verificamos novamente.

```
dados <- dados[-c(82),]
dados[dados$Id == x$Id ,]
```

```
##    Id Score Estado   Genero Idade Patrimonio   Saldo Produtos TemCartCredito Ativo  Salario Saiu
## 81 81   665     RS Feminino    34          1 9664554        2              0     0 17141366    0
```

```
x <-  dados[duplicated(dados$Id),]
x
```

```
##  [1] Id             Score          Estado         Genero         Idade          Patrimonio     Saldo
## <0 linhas> (ou row.names de comprimento 0)
```

3.6 Valores faltantes

Por fim confirmamos o tratamento de valores faltantes.

```
dados[!complete.cases(dados),]
```

```
##  [1] Id             Score          Estado         Genero         Idade          Patrimonio     Saldo
## <0 linhas> (ou row.names de comprimento 0)
```