LOAD ORIGINAL DATASET

1. Carregar os dados do dataset

Os dados originais em CSV são carregados como dataset [**churn**], e gerado um novo com os dados brutos, nomeado como [**churn_raw**].

`spec_tbl_df [7,043 × 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)`.

```
# Importando dataset agrupado original.
churn <- read_csv("Churn.csv")

# Criando um dataset tipo raw para os tratamentos, e mantendo o dataset original como backup.
churn_raw <- churn
```

Arquivo gravado no ambiente.

Arquivo Final	Formato	Total de Linhas	Total de Colunas	Tamanho
Churn.csv	CSV	7.043	21	955 kB

2. Verificações iniciais dos dados

Análise inicial dos dados contidos no dataset, as caraterísticas e tipologia de como os dados estão gravados.

1. Checagem inicial com visualização simples dos dados brutos.

```
# Visualização de todo o dataset.
View(churn_raw)
```

1. Dimensões do dataset.

```
# Dimensão do conjunto de dados.
dim(churn_raw)
> [1] 7043   21
```

Verificado que o conjunto de dados tem 7.043 registros (linhas) com 21 variáveis (colunas).

1. Examinando um amostra inicial.

```
# Listando as variáveis e a estrutura dos dados.
# Demonstratção das 5 primeiras linhas de algumas variáveis.
head(churn_raw)

> A tibble: 5 × 21
  customerID gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines    InternetService OnlineSecurity OnlineBackup
1 7590-VHVEG Female             0 Yes     No              1 No           No phone service DSL             No             Yes
2 5575-GNVDE Male               0 No      No             34 Yes          No               DSL             Yes            No
3 3668-QPYBK Male               0 No      No              2 Yes          No               DSL             Yes            Yes
4 7795-CFOCW Male               0 No      No             45 No           No phone service DSL             Yes            No
5 9237-HQITU Female             0 No      No              2 Yes          No               Fiber optic     No             No
```

1. Verificando a estrutura do dataset para o entendimento dos tipos de variáveis.

```
# Listando as variáveis e a estrutura dos dados.
str(churn_raw)

> spec_tbl_df [7,043 × 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ customerID      : chr [1:7043] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr [1:7043] "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : num [1:7043] 0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr [1:7043] "Yes" "No" "No" "No" ...
 $ Dependents      : chr [1:7043] "No" "No" "No" "No" ...
 $ tenure          : num [1:7043] 1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr [1:7043] "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr [1:7043] "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr [1:7043] "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr [1:7043] "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr [1:7043] "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr [1:7043] "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr [1:7043] "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr [1:7043] "No" "No" "No" "No" ...
 $ StreamingMovies : chr [1:7043] "No" "No" "No" "No" ...
 $ Contract        : chr [1:7043] "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr [1:7043] "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr [1:7043] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ MonthlyCharges  : num [1:7043] 29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num [1:7043] 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr [1:7043] "No" "No" "Yes" "No" ...
```

1. Verificando uma amostra dos dados de cada variável.

```
# Usando outra função para demonstrar os tipos de dados das 21 variáveis.
churn_raw %>% glimpse()

> Columns: 21 Rows: 7,043
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", ...
$ gender           <chr> "Female", "Male", "Male", ...
$ SeniorCitizen    <dbl> 0, 0, 0,...
$ Partner          <chr> "Yes", "No", "No",...
$ Dependents       <chr> "No", "No", "No", ...
$ tenure           <dbl> 1, 34, 2, ...
$ PhoneService     <chr> "No", "Yes", "Yes", ...
$ MultipleLines    <chr> "No phone service", "No", "No", ...
$ InternetService  <chr> "DSL", "DSL", "DSL", ...
$ OnlineSecurity   <chr> "No", "Yes", "Yes", ...
$ OnlineBackup     <chr> "Yes", "No", "Yes", ...
$ DeviceProtection <chr> "No", "Yes", "No", ...
$ TechSupport      <chr> "No", "No", "No", ...
$ StreamingTV      <chr> "No", "No", "No", ...
$ StreamingMovies  <chr> "No", "No", "No", ...
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", ...
$ PaperlessBilling <chr> "Yes", "No", "Yes", ...
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check",  ...
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, ...
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, ...
$ Churn            <chr> "No", "No", "Yes", ...
```

Resumindo os dados brutos do dataset [churn_raw]: 7043 linhas (registros distintos de clientes) e 21 colunas (atributo ou variáveis). Sendo “Churn” é a variável alvo, para a modelagem.

```
# Total de registros e clolunas( cada coluna um atributo/variável).
Rows: 7043 Columns: 21

# Descrição da tipologia das variáveis.
── Column specification ────────────────────────────────────────────────────────
chr (17): customerID, gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtec
          TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, Churn.
dbl  (4): SeniorCitizen, tenure, MonthlyCharges, TotalCharges.
```

1. Informações descritivas do dataset.

```
# Sumário com as informaçõs descritivas de cada variável.
summary(churn_raw)

>
 customerID         gender             SeniorCitizen    Partner            Dependents         tenure
 Length:7043        Length:7043        Min.   :0.0000   Length:7043        Length:7043        Min.   : 0.00
 Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 9.00
 Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character   Median :29.00
                                       Mean   :0.1621                                        Mean   :32.37
                                       3rd Qu.:0.0000                                        3rd Qu.:55.00
                                       Max.   :1.0000                                        Max.   :72.00

#---
 PhoneService       MultipleLines
 Length:7043        Length:7043
 Class :character   Class :character
 Mode  :character   Mode  :character

#---

 InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
 Length:7043        Length:7043        Length:7043        Length:7043
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character

 TechSupport        StreamingTV        StreamingMovies
 Length:7043        Length:7043        Length:7043
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character

#---

 Contract          PaperlessBilling   PaymentMethod
 Length:7043       Length:7043        Length:7043
 Class :character  Class :character   Class :character
 Mode  :character  Mode  :character   Mode  :character

#---

 MonthlyCharges   TotalCharges     Churn
 Min.   : 18.25   Min.   :  18.8   Length:7043
 1st Qu.: 35.50   1st Qu.: 401.4   Class :character
 Median : 70.35   Median :1397.5   Mode  :character
 Mean   : 64.76   Mean   :2283.3
 3rd Qu.: 89.85   3rd Qu.:3794.7
 Max.   :118.75   Max.   :8684.8
                  NA's   :11
```

1. Dados apenas das variáveis numéricas.

```
# Obtendo tabela com indicadores para as variáveis numéricas (ajustado manualmente).
profiling_num(churn_raw)
>
> profiling_num(churn_raw)
        variable         mean      std_dev variation_coef   skewness kurtosis      iqr
1  SeniorCitizen    0.1621468    0.3686116      2.2733201  1.8332422 4.360777    0.000
2         tenure   32.3711487   24.5594810      0.7586843  0.2394887 1.612761   46.000
3 MonthlyCharges   64.7616925   30.0900471      0.4646273 -0.2204775 1.742781   54.350
4   TotalCharges 2283.3004408 2266.7713619      0.9927609  0.9614374 2.767513 3393.288


        variable p_01   p_05   p_25     p_50      p_75    p_95     p_99         range_98
1  SeniorCitizen  0.0  0.000   0.00    0.000     0.000    1.00    1.000          [0, 1]
2         tenure  1.0  1.000   9.00   29.000    55.000   72.00   72.000         [1, 72]
3 MonthlyCharges 19.2 19.650  35.50   70.350    89.850  107.40  114.729  [19.2, 114.729]
4   TotalCharges 19.9 49.605 401.45 1397.475  3794.738 6923.59 8039.883 [19.9, 8039.883]
```

Destaques para:

- os dados da variável CustomerID são meramente identificadores descaracterizados de clientes, e não farão parte de análises;

- identificados 11 valores NA (null/em branco) na coluna TotalCharges;
- verificado que na coluna SeniorCitizen os valores estão com formato numérico, ao invés de categorizados com nas demais variáveis demográficas.
