PRE PROCESSING DATASET

1. Pre-preprocessing Data

O conjunto dados nesse momento tem alguns dados, com características distintas tendo a mesma função, por isso será feito um recondicionamento em algumas variáveis.

As variáveis do dataset [**churn_raw**] estão dispostas conforme as características dos seus dados, classificados da seguinte forma:

Demográficos	Serviços Telefonia	Serviços Internet	Contratuais	Comportamentais
Gender	PhoneService	InternetService	Contract	Churn
SeniorCitizen	MultipleLines	OnlineSecurity	PaperlessBilling	
Partner		OnlineBackup	PaymentMethod	
Dependents		DeviceProtection	Tenure	
		TechSupport	MonthlyCharges	
		StreamingTV	TotalCharges	
		StreamingMovies		

Obs: Sem considerar a variável identificadora customerId.

Tratamento dos dados brutos

Após análises iniciais, alterações são necessárias para padronização dos dados, e dos nomes no intuito de acomodações estéticas.

1. Verificando os nomes das colunas

```
# Conferindo os nomes atuais.
colnames(churn_raw)
>
 [1] "customerID"       "gender"           "SeniorCitizen"    "Partner"          "Dependents"       "tenure"           "PhoneService"
 [8] "MultipleLines"    "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection" "TechSupport"      "StreamingTV"
[15] "StreamingMovies"  "Contract"         "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"     "Churn"
```

Verificada falta de padronização nos nomes dos cabeçalhos das colunas.

---

1. Alterando o header de algumas variáveis para padronização dos nomes

```
# Alteração das clunas customerID, gender e tenure.
names(churn_raw)[names(churn_raw) == 'customerID'] <- 'CustomerID'
names(churn_raw)[names(churn_raw) == 'gender'] <- 'Gender'
names(churn_raw)[names(churn_raw) == 'tenure'] <- 'Tenure'
```

Alteração da posição da coluna Tenure para perto das informações contratuais.

```
# Reordenação da coluna Tenure da 6ª posição para 18ª posição.
churn_raw <- churn_raw %>% relocate(Tenure, .before = MonthlyCharges)
```

Para apoio didático, durante o trabalho foram verificadas alternativas para algumas das tratativas utilizadas ao longo do processo de aprendizagem da Linguagem R, e serão listadas sempre que oportuno.

```
# Variação para alteração do nome de colunas.
# 1
setnames(churn_raw, "customerID", "CustomerID")
# 2
require(dplyr)
df = rename(df, new_col01 = old_col01, new_col02 = old_col02, ...)

# Variação de reordenação simples de colunas.
# 1
churn_raw <- churn_raw %>% relocate(Tenure, .after = PaymentMethod)
# 2
churn_raw <- churn_raw %>% select (CustomerID, Gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines,
InternetService, OnlineSecurity, OnlineBackup, DeviceProtection,,TechSupport, StreamingTV, StreamingMovies,
Contract, PaperlessBilling, PaymentMethod, Tenure, MonthlyCharges, TotalCharges, Churn)
```

Confirmando as alterações realizadas.

```
# Verificação se os nomes e a ordenação ficaram em conformidade.
colnames(churn_raw)
>
 [1] "CustomerID"       "Gender"           "SeniorCitizen"    "Partner"          "Dependents"       "PhoneService"     "MultipleLines"
 [8] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection" "TechSupport"      "StreamingTV"      "StreamingMovies"
[15] "Contract"         "PaperlessBilling" "PaymentMethod"    "Tenure"           "MonthlyCharges"   "TotalCharges"     "Churn"
```

Os nomes das variáveis foram mantidos para efeito de forense posterior com os dados originais. Ainda mais que a simples tradução dos nomes não traria benefícios adicionais ao processo. Os valores também foram mantidos para dar uniformidade aos seus cabeçalhos.

Porém, foi feita uma tabela explicativa dos dados, para fins de entendimento dos tipos de dados que são semelhantes aos serviços utilizados nacionalmente.

---

1. Verificação após alteração

```
# Estrutura após alterações.
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" ...
 $ 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)" ...
 $ Tenure          : num [1:7043] 1 34 2 45 2 8 22 10 28 62 ...
 $ 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" ...
```

```
# Sumarização após alterações.
str(churn_raw)
>
  CustomerID           Gender          SeniorCitizen      Partner           Dependents        PhoneService
 Length:7043        Length:7043        Min.   :0.0000   Length:7043        Length:7043        Length:7043
 Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character   Mode  :character
                                       Mean   :0.1621
                                       3rd Qu.:0.0000
                                       Max.   :1.0000

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



 StreamingTV        StreamingMovies      Contract         PaperlessBilling   PaymentMethod          Tenure
 Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Min.   : 0.00
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 9.00
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :29.00
                                                                                                Mean   :32.37
                                                                                                3rd Qu.:55.00
                                                                                                Max.   :72.00

 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
```

Considerações após análises iniciais:

- O conjunto de dados tem 4 variáveis numéricas e 16 variáveis categóricas
    - A coluna inicial CustomerID é apenas um identificador dos clientes, de forma descaracterizada, e portanto não tem valia para a modelagem.
- Identificados 11 valores nulos na variável TotalCharges, que precisaram ser tratados;
- A coluna SeniorCitizen é a única variável demográfica com valores numéricos (binários 0|1 ao invés de Yes|No);
- A coluna alvo Churn está categorizada com valores Yes|No.

---

2. Recoding e Cleaning Data

Aplicação de tratamentos voltados para a recodificação e a limpeza dos dados, com os últimos ajustes do dataset.

2.1 Tratamento dos valores nulos (missing values)

1. Buscando missing values em em todo o dataset, verifica a existência de valores 'null' na base, retornando (TRUE|FALSE)

```
# Existe algum valor NA na base ?
any(is.na(churn_raw))
> [1] TRUE
```

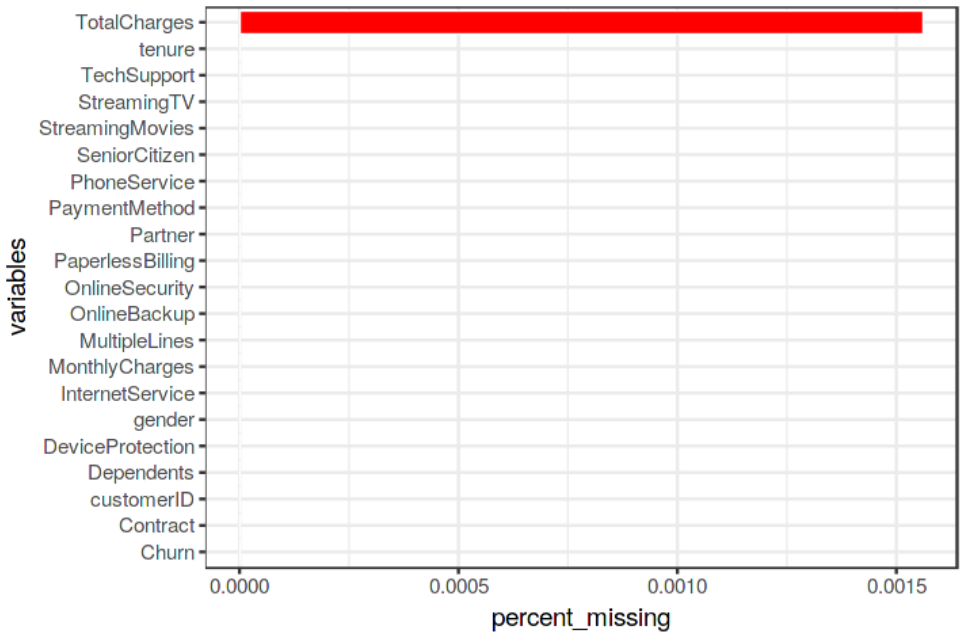
1. Buscando missing values em todas as variáveis

```
# Listagem horizontal com somatório da quantidade.
sapply(churn_raw, function(x) sum(is.na(x)))

> CustomerID           Gender    SeniorCitizen          Partner       Dependents          Tenure     PhoneService
```

```
             0              0              0              0              0              0              0 
  MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV 
             0              0              0              0              0              0              0 
 StreamingMovies         Contract PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn 
             0              0              0              0              0             11              0 
```

Conforme visto anteriormente, a variável TotalCharges apresenta 11 valores nulos (NA)

```
# Variação para busca de missing values (listagem vertical de todas as colunas).
map(churn_raw$TotalCharges, ~sum (is.na(.)))

# Sabendo em qual coluna se encontram os valores NA (contagem simples).
sum(is.na(churn_raw$TotalCharges))
> [1] 11

# Forma gráfica para identificar missing values e seu percentual.
options(repr.plot.width = 6, repr.plot.height = 4)

# Gerando um subset [missing_data] para plotagem.
missing_data <- churn_raw %>%
  summarise_all(funs(sum(is.na(.))/n()))

# Configurando o subset.
missing_data <- gather(missing_data, key = "variables",
                       value = "percent_missing")

# Plotagem do subset com a identificação dos missing values.
ggplot(missing_data, aes(x = reorder(variables, percent_missing),
                         y = percent_missing)) +
  geom_bar(stat = "identity", fill = "red",
           aes(color = I('white')), size = 0.3) +
  xlab('variables') +
  coord_flip() +
  theme_bw()
```



1. Consultando valores iguais a zero, e suas proporções em relação ao total de registros

```
# Também reconhece os valores NA e retornam sua proporção.
# É forma variante para identificar se há colunas com valores zerados/NA.
df_status(churn_raw)
>
           variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
```

```
1        CustomerID       0    0.00    0 0.00     0     0 character   7043
2            Gender       0    0.00    0 0.00     0     0 character      2
3     SeniorCitizen    5901   83.79    0 0.00     0     0   numeric      2
4           Partner       0    0.00    0 0.00     0     0 character      2
5        Dependents       0    0.00    0 0.00     0     0 character      2
6      PhoneService       0    0.00    0 0.00     0     0 character      2
7     MultipleLines       0    0.00    0 0.00     0     0 character      3
8   InternetService       0    0.00    0 0.00     0     0 character      3
9    OnlineSecurity       0    0.00    0 0.00     0     0 character      3
10     OnlineBackup       0    0.00    0 0.00     0     0 character      3
11 DeviceProtection       0    0.00    0 0.00     0     0 character      3
12      TechSupport       0    0.00    0 0.00     0     0 character      3
13      StreamingTV       0    0.00    0 0.00     0     0 character      3
14  StreamingMovies       0    0.00    0 0.00     0     0 character      3
15         Contract       0    0.00    0 0.00     0     0 character      3
16 PaperlessBilling       0    0.00    0 0.00     0     0 character      2
17    PaymentMethod       0    0.00    0 0.00     0     0 character      4
18           Tenure      11    0.16    0 0.00     0     0   numeric     73
19   MonthlyCharges       0    0.00    0 0.00     0     0   numeric   1585
20     TotalCharges       0    0.00   11 0.16     0     0   numeric   6530
21            Churn       0    0.00    0 0.00     0     0 character      2
```

Constatado na variável TotalCharges que os 11 valores NA representam aproximadamente 0,16% do total de registros. Exatamente como na variável Tenure, em que também constam 11 valores zerados (diferente de NA).

```
# Confirmando a proporção de 0,16% para os 11 registros, com missing values.
sum(is.na(churn_raw$TotalCharges))/nrow(churn_raw)
>[1] 0.001561834
```

Por conta disso, será feita análise para visualizar os 11 registros de TotalCharges e confrontar com os registros de Tenure.

---

1. Demonstrando os 11 registros que têm valor NULL para TotalCharges

```
# Listando os 11 registros de TotalCharges com valor = NULL.
churn_raw[is.na(churn_raw$TotalCharges),]
```

Essa visualização consegue trazer dados, porém não fica clara se há relação entre as colunas TotalCharges e Tenure.

A disposição das colunas é tal que, as cinco primeiras colunas são de dados demográficos, a 18ª coluna é a variável Tenure, enquanto que TotalCharges é a 20ª e Churn a 21ª.

1. Fazendo um ajuste para que retorne as colunas exatamente conforme é preciso.

```
# Buscar os 11 registros com dados das primeiras 5 colunas do conjunto e colocando ao final as colunas Tenure TotalCharges e Churn
# (verificando se são churnners).
churn_raw[is.na(churn_raw$TotalCharges),c(1:5, 18, 20, 21)]
>
A tibble: 11 × 8
   CustomerID Gender SeniorCitizen Partner Dependents Tenure TotalCharges Churn
 1 4472-LVYGI Female             0 Yes     Yes             0           NA No
 2 3115-CZMZD Male               0 No      Yes             0           NA No
 3 5709-LVOEQ Female             0 Yes     Yes             0           NA No
 4 4367-NUYAO Male               0 Yes     Yes             0           NA No
 5 1371-DWPAZ Female             0 Yes     Yes             0           NA No
 6 7644-OMVMY Male               0 Yes     Yes             0           NA No
 7 3213-VVOLG Male               0 Yes     Yes             0           NA No
 8 2520-SGTTA Female             0 Yes     Yes             0           NA No
 9 2923-ARZLG Male               0 Yes     Yes             0           NA No
10 4075-WKNIU Female             0 Yes     Yes             0           NA No
11 2775-SEFEE Male               0 No      Yes             0           NA No
```

Com essa disposição é possível diagnosticar as informações para análise:

- demonstrado que os 11 valores nulos de TotalCharges (NA) se tratam dos mesmos clientes com valor zerado para Tenure;
- a inspeção da variável Tenure mostra que todos os 11 clientes zerados são assinantes iniciais, com menos de 1 mês de assinatura, que tiveram contabilizados seus valores totais baseado na regra: TotalCharges = MonthlyCharges x Tenure ,

zerando o valor total;

- a fim de não afetar a modelagem, os registros com TotalCharges = NULL, serão deletados.

1. Deleção dos registros NULL com a geração de um novo dataset com nome [**churn_raw_new**]

```
# Gerado dataset [churn_raw_new] sem nenhum missing value.
churn_raw_new <- na.omit(churn_raw)

# Variação para o cleaning.
churn_raw_new <- churn_raw[complete.cases(churn_raw),]
```

1. Validação de que não há valores nulos em nenhuma das variáveis do dataset [**churn_raw_new**]

```
# Existe algum valor NA na base ?
any(is.na(churn_raw_new ))
>
[1] FALSE
```

```
# Identificando se ainda constam colunas com valores NA para TotalCharges.
df_status(churn_raw_new)
>
variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1        CustomerID       0    0.00    0    0     0     0 character   7032
2            Gender       0    0.00    0    0     0     0 character      2
3     SeniorCitizen    5890   83.76    0    0     0     0   numeric      2
4           Partner       0    0.00    0    0     0     0 character      2
5        Dependents       0    0.00    0    0     0     0 character      2
6      PhoneService       0    0.00    0    0     0     0 character      2
7     MultipleLines       0    0.00    0    0     0     0 character      3
8   InternetService       0    0.00    0    0     0     0 character      3
9    OnlineSecurity       0    0.00    0    0     0     0 character      3
10     OnlineBackup       0    0.00    0    0     0     0 character      3
11 DeviceProtection       0    0.00    0    0     0     0 character      3
12      TechSupport       0    0.00    0    0     0     0 character      3
13      StreamingTV       0    0.00    0    0     0     0 character      3
14  StreamingMovies       0    0.00    0    0     0     0 character      3
15         Contract       0    0.00    0    0     0     0 character      3
16 PaperlessBilling       0    0.00    0    0     0     0 character      2
17    PaymentMethod       0    0.00    0    0     0     0 character      4
18           Tenure       0    0.00    0    0     0     0   numeric     72
19   MonthlyCharges       0    0.00    0    0     0     0   numeric   1584
20     TotalCharges       0    0.00    0    0     0     0   numeric   6530
21            Churn       0    0.00    0    0     0     0 character      2
```

2.2 Recodificação de variáveis categóricas.

Após eliminação de valores nulos e zerados para TotalCharges, outros ajustes são necessários para melhorar a qualidade do conjunto de dados.

2.2.1 Alteração dos valores da coluna SeniorCitizen

Os valores da variável estão definidos como 0|1, os valores serão recodificados para Yes|No a fim de compatibilizar como variável categórica, e permitir melhores interpretações posteriores.

```
# Visualização simples da frequência dos dados antes do recondicionamento.
# Disposição tabular horizontal, e identificação da variável.
apply(churn_raw[c("SeniorCitizen")], 2, table)
>
SeniorCitizen
0          5890
1          1142
```

```
# Variações para visualização da frequência.

# 1 table.
table(churn_raw$SeniorCitizen)
> 0    1
 5890 1142

# 2 base - distribuição idêntica ao [table].
base::table(churn_raw_new$SeniorCitizen)

# 3 prop.table - verificação com basse percentual dos dados.
prop.table(table(churn_raw_new$SeniorCitizen))
> 0         1
0.8375995 0.1624005
#---
prop.table(table(churn_raw_new$SeniorCitizen))*100
> 0        1
83.75995 16.24005
```

Aplicação do recondicionamento.

```
# Troca dos valores.
churn_raw_new$SeniorCitizen <- as.character(mapvalues(churn_raw_new$SeniorCitizen,
                                                      from=c("0","1"),
                                                      to=c("No", "Yes")))

# Variação do recondicionamento.
churn_raw_new$SeniorCitizen <- as.character(ifelse(churn_raw_new$SeniorCitizen == 1, "Yes", "No"))
```

```
# Visualização da frequência dos dados após recondicionamento.
apply(churn_raw_new[c("SeniorCitizen")], 2, table)
>
 SeniorCitizen
  No      5890
  Yes     1142

# Detalhamento das frequências dos valores da coluna SeniorCitizen após recodificação, com apenas os valores Yes|No.
tab1(churn_raw_new$SeniorCitizen, sort.group = "decreasing", cum.percent = TRUE)
>
churn_raw_new$SeniorCitizen :
        Frequency Percent Cum. percent
No           5890    83.8         83.8
Yes          1142    16.2        100.0
Total        7032   100.0        100.0
```

Recondicionamento dos dados demográficos finalizado.

2.2.2 Alteração dos valores dos dados de serviço de telefonia.

Análise dos dados de serviços de telefonia para identificar se há necessidade de recodificação.

Verificando os dados das variáveis PhoneService e MultipleLines:

```
# Fazendo uma análise da frequência dos dados de cada variável.
apply(churn_raw_new[c("PhoneService", "MultipleLines")], 2, table)
$PhoneService       $MultipleLines
  No   Yes           No     No phone service    Yes
 680  6352         3385                  680   2967

# Uma análise cruzada entre os dados das duas variáveis.
CrossTable(churn_raw_new$PhoneService,
           churn_raw_new$MultipleLines,
           prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

Total Observations in Table:  7032

                           |                   churn_raw_new$MultipleLines          |                  |
---------------------------|------------------|------------------|------------------|------------------|
churn_raw_new$PhoneService |               No | No phone service |              Yes |        Row Total |
---------------------------|------------------|------------------|------------------|------------------|
```

```
                        No |               0 |             680 |               0 |             680 |
                           |         327.332 |        5737.757 |         286.911 |                 |
                           |           0.000 |           1.000 |           0.000 |           0.097 |
                           |           0.000 |           1.000 |           0.000 |                 |
                           |           0.000 |           0.097 |           0.000 |                 |
---------------------------|-----------------|-----------------|-----------------|-----------------|
                       Yes |            3385 |               0 |            2967 |            6352 |
                           |          35.042 |         614.243 |          30.715 |                 |
                           |           0.533 |           0.000 |           0.467 |           0.903 |
                           |           1.000 |           0.000 |           1.000 |                 |
                           |           0.481 |           0.000 |           0.422 |                 |
---------------------------|-----------------|-----------------|-----------------|-----------------|
             Column Total |            3385 |             680 |            2967 |            7032 |
                           |           0.481 |           0.097 |           0.422 |                 |
---------------------------|-----------------|-----------------|-----------------|-----------------|
```

Constatado que a variável MultipleLines é **dependente** da variável PhoneService. Como isso os 680 valores 'No phone service' representam os mesmos 680 registros com valores 'No' para PhoneService.

Assim, fácil notar que os 680 clientes que não possuem serviço de telefonia, também não possuem nenhuma linha adicional. Os resultados podem ser recodificados para 'No phone service', para fins de análise futura.

Aplicação do recondicionamento.

```
# Recodificando a resposta 'No phone service' para 'No'.
churn_raw_new$MultipleLines <- as.character(mapvalues(churn_raw_new$MultipleLines,
                                                      from=c("No phone service"),
                                                      to=c("No")))


# Variação para recodificação.
churn_raw_new <- data.frame(lapply(churn_raw_new, function(x) {
                                    gsub("No", "No phone service", x)}
                                  )
                            )
```

```
# Fazendo nova análise cruzada da frequência após recondicionamento.
apply(churn_raw_new[c("PhoneService", "MultipleLines")], 2, table)
>
    PhoneService MultipleLines
No           680          4065
Yes         6352          2967

# Detalhamento das frequências dos valores da coluna MultipleLines após recodificação, com apenas os valores Yes|No.
tab1(churn_raw_new$MultipleLines, sort.group = "decreasing", cum.percent = TRUE)
>
churn_raw_new$MultipleLines :
        Frequency Percent Cum. percent
No           4065    57.8         57.8
Yes          2967    42.2        100.0
  Total      7032   100.0        100.0
```

Recondicionamento dos dados se serviços de telefonia finalizado.

2.2.3 Alteração dos valores das variáveis adicionais ao InternetService.

Análise dos dados de serviços de internet para identificar se há necessidade de recodificação.

Analogamente as variáveis OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies são dependentes da variável InternetService, pois quando o valor de Serviço de Internet é 'No', os valores das 6 colunas a seguir mostram 'No Internet Service'.

Analisando os dados da variável InternetService.

```
# Verificando a frequência dos dados de serviço de internet.
CrossTable(churn_raw_new1$InternetService, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)
>  |         DSL | Fiber optic |          No |
   |-------------|-------------|-------------|
```

```
   |        2416 |        3096 |        1520 |
   |       0.344 |       0.440 |       0.216 |
   |-------------|-------------|-------------|
#--- Existem 1520 clientes que não possuem serviço de internet.
```

Analisando cada variável de serviço adicional de internet com os dados da variáveis InternetService.

```
# Comparação das variáveis de serviço de internet para verificação de dependência.
table(churn_raw_new[, c("InternetService", "OnlineSecurity")])

>              OnlineSecurity
InternetService   No No internet service  Yes
    DSL         1240                   0 1176
    Fiber optic 2257                   0  839
    No             0                1520    0
#---

table(churn_raw_new[, c("InternetService", "OnlineBackup")])

>               OnlineBackup
InternetService   No No internet service  Yes
    DSL         1334                   0 1082
    Fiber optic 1753                   0 1343
    No             0                1520    0
#---

table(churn_raw_new[, c("InternetService", "DeviceProtection")])

>               DeviceProtection
InternetService   No No internet service  Yes
    DSL         1355                   0 1061
    Fiber optic 1739                   0 1357
    No             0                1520    0
#---

table(churn_raw_new[, c("InternetService", "TechSupport")])
              TechSupport
InternetService   No No internet service  Yes
    DSL         1242                   0 1174
    Fiber optic 2230                   0  866
    No             0                1520    0
#---

table(churn_raw_new[, c("InternetService", "StreamingTV")])

>               StreamingTV
InternetService   No No internet service  Yes
    DSL         1463                   0  953
    Fiber optic 1346                   0 1750
    No             0                1520    0
#---

table(churn_raw_new[, c("InternetService", "StreamingMovies")])

>               StreamingMovies
InternetService   No No internet service  Yes
    DSL         1436                   0  980
    Fiber optic 1345                   0 1751
    No             0                1520    0
```

Fácil verificar que todos os serviços adicionais contém 1520 registros com o valor 'No internet service'.

Todas as variáveis de serviço adicionais de internet devem ser recondicionado para identificarem 'No internet service' como 'No'.

Aplicação do recondicionamento.

```
# Alterando todas as variáveis dependendtes trocando o valor 'No internet service' por 'No'.
churn_raw_new$InternetService[churn_raw_new$InternetService=="No"] <- "No internet service"
churn_raw_new$OnlineSecurity[churn_raw_new$OnlineSecurity=="No internet service"] <- "No"
churn_raw_new$OnlineBackup[churn_raw_new$OnlineBackup=="No internet service"] <- "No"
churn_raw_new$DeviceProtection[churn_raw_new$DeviceProtection=="No internet service"] <- "No"
churn_raw_new$TechSupport[churn_raw_new$TechSupport=="No internet service"] <- "No"
```

```
churn_raw_new$StreamingTV[churn_raw_new$StreamingTV=="No internet service"] <- "No"
churn_raw_new$StreamingMovies[churn_raw_new$StreamingMovies=="No internet service"] <- "No"
```

Verificação das frequências e seus percentuais com o ANTES, e o DEPOIS após o ajuste das variáveis.

```
# As frequências ANTES, com 3 valores, as frequências DEPOIS com apenas 2 valores.

CrossTable(churn_raw_new$OnlineSecurity, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)
>
|                  No | No internet service |                 Yes |
|---------------------|---------------------|---------------------|
|                3497 |                1520 |                2015 |
|               0.497 |               0.216 |               0.287 |
|---------------------|---------------------|---------------------|

|                  No  |                 Yes |
|----------------------|---------------------|
|                5017 |                2015 |
|               0.713 |               0.287 |
|---------------------|---------------------|

CrossTable(churn_raw_new$OnlineBackup, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

|                  No | No internet service |                 Yes |
|---------------------|---------------------|---------------------|
|                3087 |                1520 |                2425 |
|               0.439 |               0.216 |               0.345 |
|---------------------|---------------------|---------------------|

|                  No  |                 Yes |
|----------------------|---------------------|
|                4607 |                2425 |
|               0.655 |               0.345 |
|---------------------|---------------------|

CrossTable(churn_raw_new$DeviceProtection, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

|                  No | No internet service |                 Yes |
|---------------------|---------------------|---------------------|
|                3094 |                1520 |                2418 |
|               0.440 |               0.216 |               0.344 |
|---------------------|---------------------|---------------------|

|                  No  |                 Yes |
|----------------------|---------------------|
|                4614 |                2418 |
|               0.656 |               0.344 |
|---------------------|---------------------|

CrossTable(churn_raw_new$TechSupport, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

|                  No | No internet service |                 Yes |
|---------------------|---------------------|---------------------|
|                3472 |                1520 |                2040 |
|               0.494 |               0.216 |               0.290 |
|---------------------|---------------------|---------------------|

|                  No  |                 Yes |
|----------------------|---------------------|
|                4992 |                2040 |
|               0.710 |               0.290 |
|---------------------|---------------------|

CrossTable(churn_raw_new$StreamingTV, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

|                  No | No internet service |                 Yes |
|---------------------|---------------------|---------------------|
|                2809 |                1520 |                2703 |
|               0.399 |               0.216 |               0.384 |
|---------------------|---------------------|---------------------|

|                  No  |                 Yes |
|----------------------|---------------------|
|                4329 |                2703 |
|               0.616 |               0.384 |
|---------------------|---------------------|
```

```
CrossTable(churn_raw_new$StreamingMovies, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)

|                   No | No internet service |                  Yes |
|----------------------|---------------------|----------------------|
|                 2781 |                1520 |                 2731 |
|                0.395 |               0.216 |                0.388 |
|----------------------|---------------------|----------------------|

|                  No  |                  Yes |
|----------------------|----------------------|
|                 4301 |                 2731 |
|                0.612 |                0.388 |
|----------------------|----------------------|

Total Observations in Table:  7032
#--- Para todas as variáveis os percentuais para valores = Yes foram mantidos,
#--- e aumentados, para os valores = No internet service, do percentual referente ao valor = No.
#--- Para a variável InternetService, houve apenas a alteraçaõ do nome, sem modificações no quantitativo.
```

Verificação tabular para o resultado da recodificação dos dados se serviços de internet.

```
# Visualização da frequência cruzada dos valores em tablução simples.

table(churn_raw_new[, c("InternetService", "OnlineSecurity")])
>
                     OnlineSecurity
InternetService                        No  Yes
  DSL                                1240 1176
  Fiber optic                        2257  839
  No internet service                1520    0
#---

table(churn_raw_new[, c("InternetService", "OnlineBackup")])
>
                     OnlineBackup
InternetService                        No  Yes
  DSL                                1334 1082
  Fiber optic                        1753 1343
  No internet service                1520    0
>
#---

table(churn_raw_new[, c("InternetService", "DeviceProtection")])
>
                    DeviceProtection
InternetService                        No  Yes
  DSL                                1355 1061
  Fiber optic                        1739 1357
  No internet service                1520    0
#---

table(churn_raw_new[, c("InternetService", "TechSupport")])
>
                    TechSupport
InternetService                        No  Yes
  DSL                                1242 1174
  Fiber optic                        2230  866
  No internet service                1520    0
#---

table(churn_raw_new[, c("InternetService", "StreamingTV")])
>
                    StreamingTV
InternetService                        No  Yes
  DSL                                1463  953
  Fiber optic                        1346 1750
  No internet service                1520    0
#---

table(churn_raw_new[, c("InternetService", "StreamingMovies")])
                    StreamingMovies
InternetService                        No  Yes
  DSL                                1436  980
  Fiber optic                        1345 1751
  No internet service                1520    0
#---
```

```
# Verificando a estrutura do dataset após as alterações.
str(churn_raw_new)
>
tibble [7,032 × 21] (S3: tbl_df/tbl/data.frame)
 $ CustomerID      : chr [1:7032] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ Gender          : chr [1:7032] "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : chr [1:7032] "No" "No" "No" "No" ...
 $ Partner         : chr [1:7032] "Yes" "No" "No" "No" ...
 $ Dependents      : chr [1:7032] "No" "No" "No" "No" ...
 $ PhoneService    : chr [1:7032] "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr [1:7032] "No" "No" "No" "No" ...
 $ InternetService : chr [1:7032] "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr [1:7032] "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr [1:7032] "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr [1:7032] "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr [1:7032] "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr [1:7032] "No" "No" "No" "No" ...
 $ StreamingMovies : chr [1:7032] "No" "No" "No" "No" ...
 $ Contract        : chr [1:7032] "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr [1:7032] "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr [1:7032] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ Tenure          : num [1:7032] 1 34 2 45 2 8 22 10 28 62 ...
 $ MonthlyCharges  : num [1:7032] 29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num [1:7032] 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr [1:7032] "No" "No" "Yes" "No" ...
```

Recondicionamento dos dados se serviços de internet finalizado, agora com 3 variáveis numéricas e 18 categóricas.

2.2.4 Validação do recoding realizado nas variáveis demográficas e de serviços.

```
# Após o recoding, todas as variáveis Demográficas e de Serviços ficaram com valores Yes|No para seus registros.
# Para algumas variáveis, a categorização não foi alterada, casos de:
# Gender (Male|Female) e InternetService (DSL|Fiber optic|No).
glimpse(churn_raw_new)
>
Rows: 7,032 Columns: 21
$ CustomerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW", "9237-HQI…
$ Gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Male", "Fe…
$ SeniorCitizen    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "No",…
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes", "No", "Ye…
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "Ye…
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "Yes"…
$ MultipleLines    <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "No", "N…
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber optic", "Fiber …
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "No", "Yes", …
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "No", "Yes", "…
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Yes", "No", "…
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes", "No", "No…
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Yes", "No", "N…
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "No…
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One year", "Mont…
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "No",…
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "Bank transfe…
$ Tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 25, 69, 52, 7…
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.75, 104.80, 5…
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949.40, 301.90,…
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "…
```

Verificação se após recoding, algum dos procedimentos causou avaria ou nulidade em algum dado.

```
# Consulta simples se há algum valor NA em todo o dataset.
any(is.na(churn_raw_new))
[1] FALSE
```

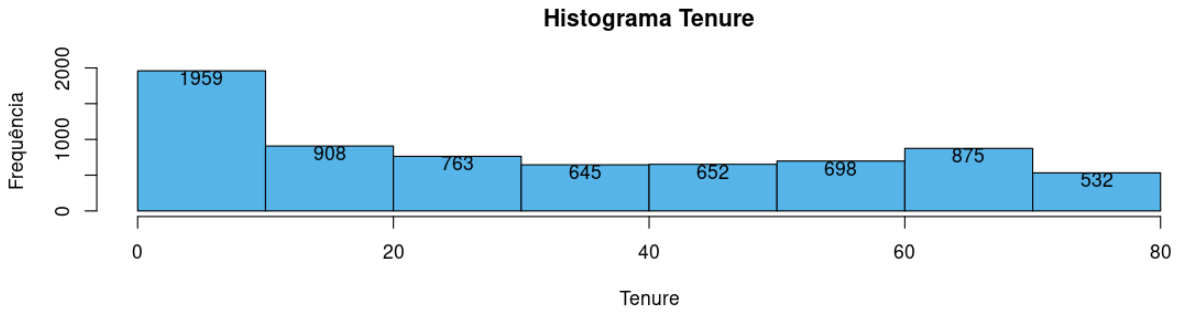
Verificação se após recoding, algum dos procedimentos causou avaria ou nulidade em algum dado.

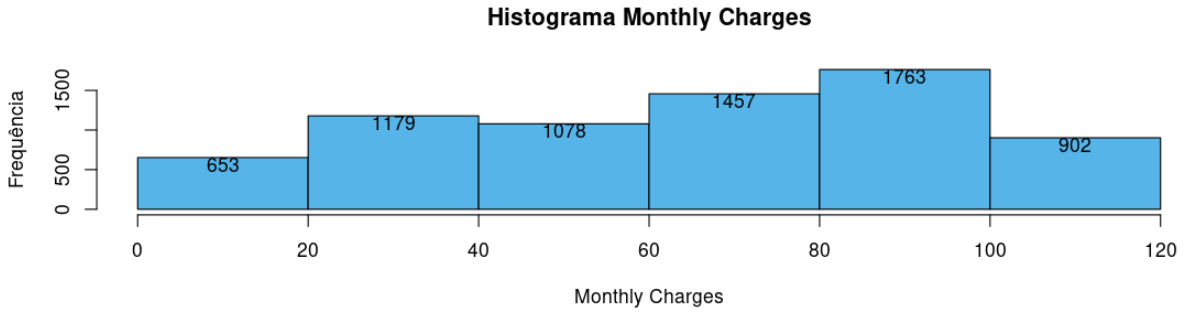
2.3 Análises antes da recodificação de variáveis numéricas.

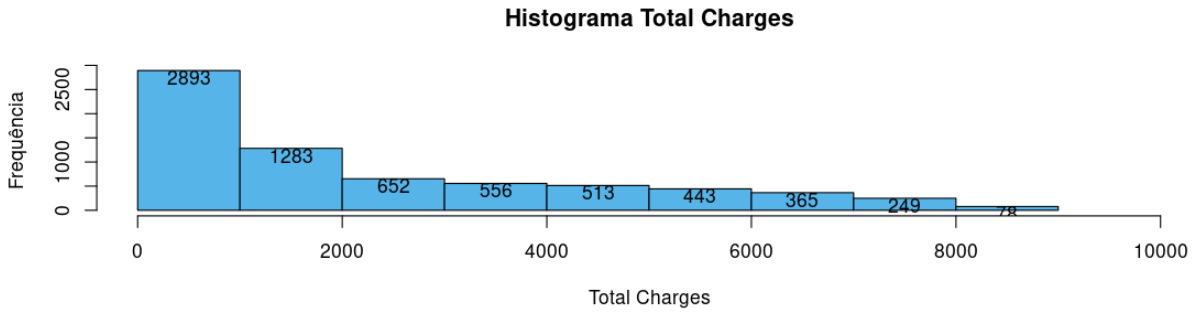
Antes de finalizar o recoding, iniciar algumas análises gráficas para verificações simples de outliers, variação e amplitude.

2.3.1 Distribuição Histográfica

```
# Verificação da distribuição.
t <- hist(churn_raw_new$Tenure,
     xlim=c(0,80),
     breaks=10,
     main='Histograma Tenure',
     xlab='Tenure',
     ylab='Frequência',
     col = "#56B4E9")

text(t$mids, t$counts, labels=t$counts, adj=c(0.5, 0.99))

mc <- hist(churn_raw_new$MonthlyCharges,
           xlim=c(0,120),
           breaks=5,
           main='Histograma Monthly Charges',
           xlab='Monthly Charges',
           ylab='Frequência',
           col = "#56B4E9")

text(mc$mids, mc$counts, labels=mc$counts, adj=c(0.5, 1))

tc <- hist(churn_raw_new$TotalCharges,
           xlim=c(0,10000),
           breaks=8,
           main='Histograma Total Charges',
           xlab='Total Charges',
           ylab='Frequência',
           col = "#56B4E9")

text(tc$mids, tc$counts, labels=tc$counts, adj=c(0.5,1))
```





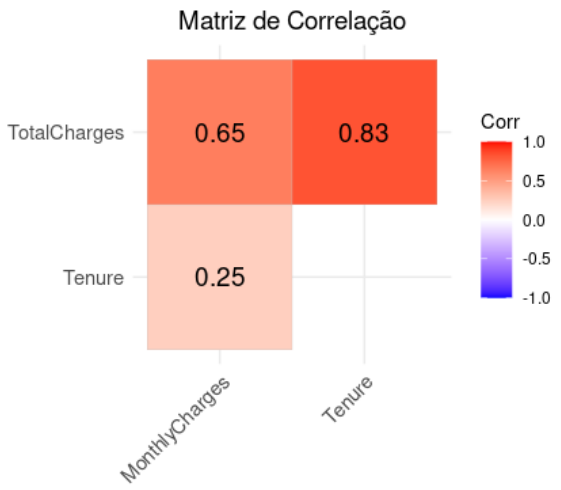


Para as variáveis numéricas, Tenure, MonthlyCharges e TotalCharges, não foram verificados outliers.

2.3.2 Correlação

A correlação visa mensurar a relação entre as variáveis numéricas contínuas do conjunto, a fim de evitar uma multicolinearidade, que podem ampliar efeitos de algum vício, na modelagem de regressão.

```
# Valores para Correlação Tenure, MonthlyCharges, TotalCharges.
ggcorrplot(round(cor(churn_raw_new[,c(18, 19, 20)]),2),
           title = "Matriz de Correlação",
           hc.order=TRUE,
           lab=TRUE, lab_size = 5,
           type = "upper") +
  theme(plot.title=element_text(hjust = 0.5, size = 14),
        axis.text.y = element_text(size = 11),
        axis.text.x = element_text(size = 11),
        legend.text = element_text(size = 9))

#---

# Variação simples da Correlação com uso de círculos.

# Valores para correlação Tenure, MonthlyCharges, TotalCharges.
Mat_Corr <- cor(churn_raw_new[,c(18,19,20)])

# Resumo dos valores de Correlação entre as variáveis numéricas.
Mat_Corr
>
                  Tenure MonthlyCharges TotalCharges
Tenure         1.0000000      0.2468618    0.8258805
MonthlyCharges 0.2468618      1.0000000    0.6510648
TotalCharges   0.8258805      0.6510648    1.0000000

# Plotagem da Matriz de Correlação, com os valores associados.
ggcorrplot(round(cor(churn_raw_new[,c(18, 19, 20)]),2),
           title = "Matriz de Correlação",
           hc.order=TRUE,
           lab=TRUE, lab_size = 5,
           type = "upper") +
  theme(plot.title=element_text(hjust = 0.5, size = 14),
        axis.text.y = element_text(size = 11),
        axis.text.x = element_text(size = 11),
        legend.text = element_text(size = 9))
```



Verificando a sumarização das variáveis com valores que representam os gastos em dinheiro, temos o range do quanto o valor total varia entre seu mínimo e seu máximo, apresentando uma variabilidade de mais de 8.600 pontos. Em contrapartida, a variável de gastos mensais tem uma variação de apenas 100 pontos, sendo assim a mais indicada para o trabalho com relação aos gastos de clientes.

```
# Sumário das variáveis TotalCharges e MonthlyCharges.
> summary(churn_raw_new$TotalCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   18.8   401.4  1397.5  2283.3  3794.7  8684.8

> summary(churn_raw_new$MonthlyCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.25   35.59   70.35   64.80   89.86  118.75
```

Com a análise de Correlação entre os valores numéricos fica evidenciado que:

- Forte correlação positiva (0,83) entre Tenure e TotalCharges;
- Fraca correlação positiva (0,25) entre Tenure e MonthlyCharges.
- Moderada correlação positiva (0,65) entre MonthlyCharges e TotalCharges.

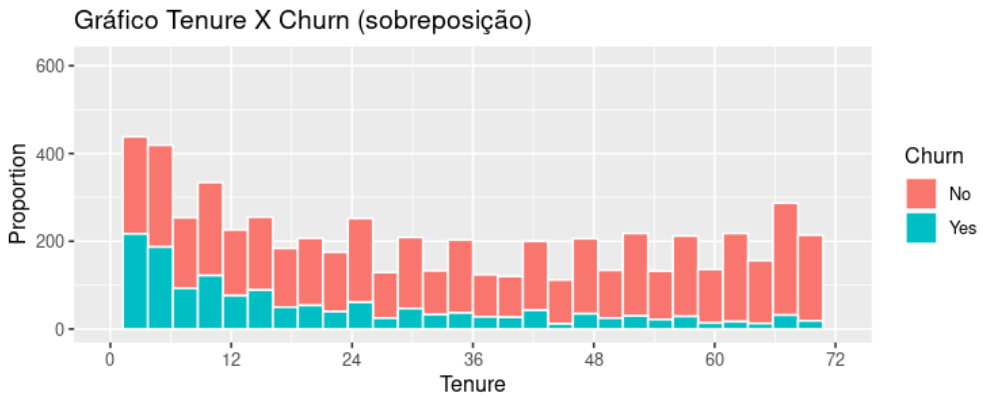
Assim, caso seja necessária a eliminação de um atributo relacionado com valores em dinheiro, a variável TotalCharges será a escolhida, por ter a maior variabilidade e por ter correlação direta com a variável MonthlyCharges.

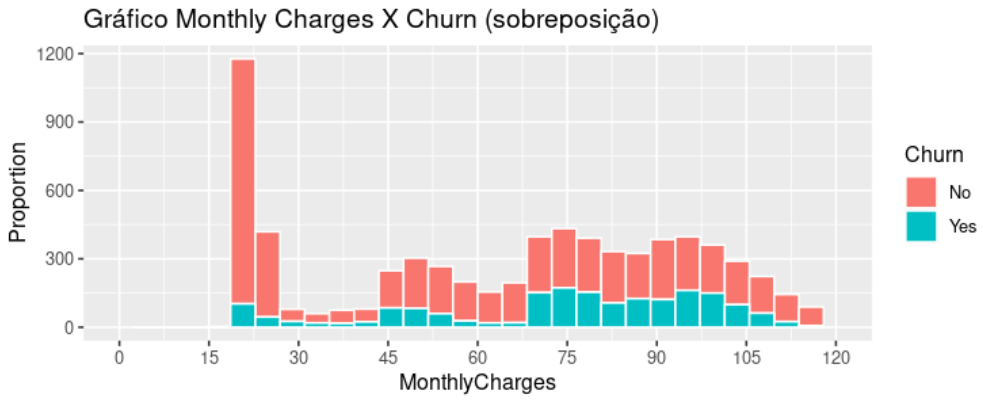
2.3.3 Relação entre as variáveis numéricas e Churn

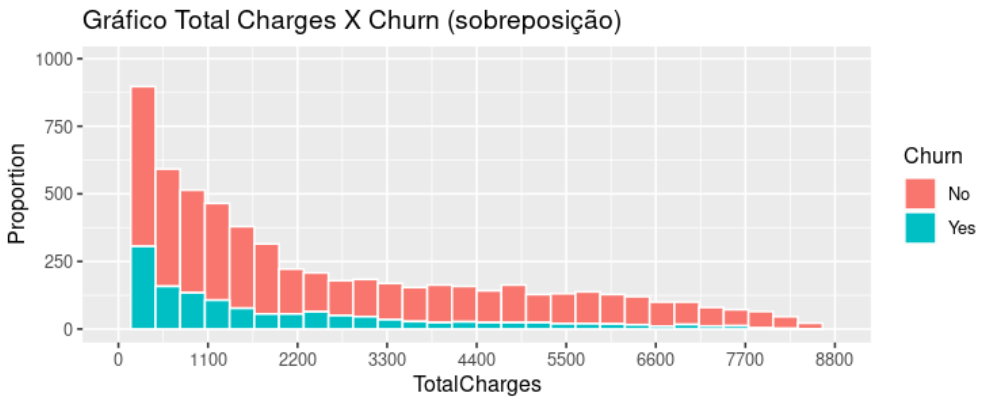
A seguir uma demonstração da relação de churners entre as variáveis numéricas e a variável de alvo Churn, antes dos recondicioamentos, para verificar como estão associadas, e se há maneira de verificar qual variável apresenta informações relevantes para a modelagem.

```
# Gráficos de Barras com sobreposição dos valores YES | NO para Churn.

# Tenure X Churn.
ggplot(data = churn_raw_new,
       aes(x = Tenure,
           fill = Churn)) +
  geom_histogram(colour = 'white',
                 position = 'stack') +
  scale_x_continuous(
    breaks = seq(0, 72, 12),
    limits=c(0, 72)) +
  ggtitle("Gráfico Tenure X Churn (sobreposição)") +
```

```
  labs(y = "Proportion")


# MonthlyCharges X Churn.
ggplot(data = churn_raw_new,
       aes(x = MonthlyCharges,
           fill = Churn)) +
  geom_histogram(colour = 'white',
                 position = 'stack') +
  scale_x_continuous(
    breaks = seq(0, 120, 15),
    limits=c(0, 120)) +
  ggtitle("Gráfico Monthly Charges X Churn (sobreposição)") +
  labs(y = "Proportion")

# TotalCharges X Churn.
ggplot(data = churn_raw_new,
       aes(x = TotalCharges,
           fill = Churn)) +
  geom_histogram(colour = 'white',
                 position = 'stack') +
                 scale_x_continuous(
                   breaks = seq(0, 8800, 1100),
                   limits=c(0, 8800)) +
  ggtitle("Gráfico Total Charges X Churn (sobreposição)") +
  labs(y = "Proportion")
```







Com esses gráficos fica evidenciado:

- Forte relação entre churners e valores de Tenure menor do que 1 ano;
- Forte relação entre churners e valores de TotalCharges mais baixos;
- Dispersa relação entre churners e valores MonthlyCharges.

2.4 Recodificação de variáveis numéricas.

Os dados numéricos precisam ser analisados para verificar se alguma recodificação é necessária, a fim de favorecer a modelagem.

2.4.1 Tenure

Tenure representa o período de tempo, em meses, que o cliente ficou fidelizado na companhia. Para melhorar os padrões relativos ao tempo de cada cliente, será estratificado em 6 níveis, em que cada nível representa um ano de contrato.

1. Verificação da variação de tempo

```
# Sumário da variável antes da transformação para factor.
summary(churn_raw_new$Tenure)
>
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    9.00   29.00   32.42   55.00   72.00
```

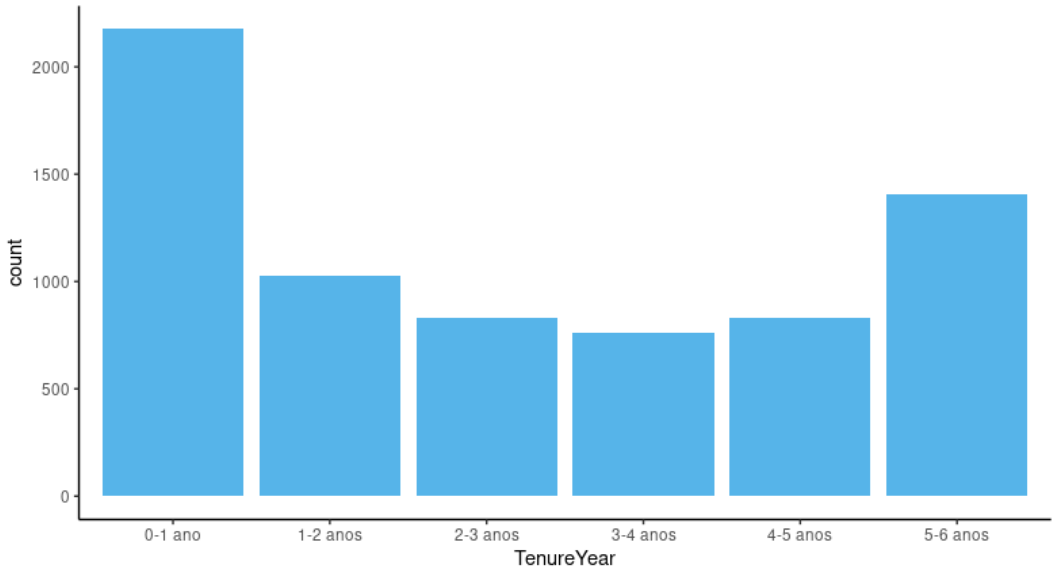
Como a permanência mínima é de 1 mês e a permanência máxima é de 72 meses, o agrupamento do tempo de contrato, em anos, dará maior condição de análises histográficas, por isso estratificando os tempos de contrato ficarão da seguinte maneira:

- “0-12 Meses = 0-1 ano”;
- “12–24 Meses = 1-2 anos”;
- “25–36 Meses = 2-3 anos”;
- “37-48 Meses = 3-4 anos”;
- “49–60 Meses = 4-5 anos”;
- “61 a 72 Meses = 5-6 anos”.

1. Promovendo a estratificação da variação de tempo

```
# Efetuando a estratificação.
churn_raw_new %>%
  mutate(TenureYear = case_when(Tenure <= 12 ~ "0-1 ano",
                                Tenure > 12 & Tenure <= 24 ~ "1-2 anos",
                                Tenure > 24 & Tenure <= 36 ~ "2-3 anos",
                                Tenure > 36 & Tenure <= 48 ~ "3-4 anos",
```

```
                                Tenure > 48 & Tenure <= 60 ~ "4-5 anos",
                                Tenure > 60 & Tenure <= 72 ~ "5-6 anos")) -> churn_raw_new

# Verificando a estratificação.
# 1 - table.
table(churn_raw_new$TenureYear )
>
   0-1 ano 1-2 anos 2-3 anos 3-4 anos 4-5 anos 5-6 anos
      2175     1024      832      762      832     1407

# 2 - CrossTable.
CrossTable(churn_raw_new$TenureYear , prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)
>
   |   0-1 ano |  1-2 anos |  2-3 anos |  3-4 anos |  4-5 anos |  5-6 anos |
   |-----------|-----------|-----------|-----------|-----------|-----------|
   |      2175 |      1024 |       832 |       762 |       832 |      1407 |
   |     0.309 |     0.146 |     0.118 |     0.108 |     0.118 |     0.200 |
   |-----------|-----------|-----------|-----------|-----------|-----------|

# 3 - Histograma.
base::table(churn_raw_new$TenureYear)

ggplot(churn_raw_new, aes(churn_raw_new$TenureYear)) +
  geom_bar(fill = "#56B4E9") +
  theme_classic() + xlab("TenureYear")
```




```
# Alterando a posição da coluna TenureYear que ficou no final do dataset para a 18ª posição.
# E alterando Tenure para o final.
churn_raw_new <- chuchurn_raw_new %>% relocate(TenureYear , .before = MonthlyCharges)
churn_raw_new <- churn_raw_new %>% relocate(Tenure, .after = Churn)

# Conferindo a estrutura do dataset após estratificação.
str(churn_raw_new)
>
tibble [7,032 × 22] (S3: tbl_df/tbl/data.frame)
 $ CustomerID      : chr [1:7032] "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ Gender          : chr [1:7032] "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : chr [1:7032] "No" "No" "No" "No" ...
 $ Partner         : chr [1:7032] "Yes" "No" "No" "No" ...
 $ Dependents      : chr [1:7032] "No" "No" "No" "No" ...
 $ PhoneService    : chr [1:7032] "No phone service" "Yes" "Yes" "No phone service" ...
 $ MultipleLines   : chr [1:7032] "No phone service" "No phone service" "No phone service" "No phone service" ...
 $ InternetService : chr [1:7032] "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr [1:7032] "No internet service" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr [1:7032] "Yes" "No internet service" "Yes" "No internet service" ...
 $ DeviceProtection: chr [1:7032] "No internet service" "Yes" "No internet service" "Yes" ...
 $ TechSupport     : chr [1:7032] "No internet service" "No internet service" "No internet service" "Yes" ...
 $ StreamingTV     : chr [1:7032] "No internet service" "No internet service" "No internet service" ...
```

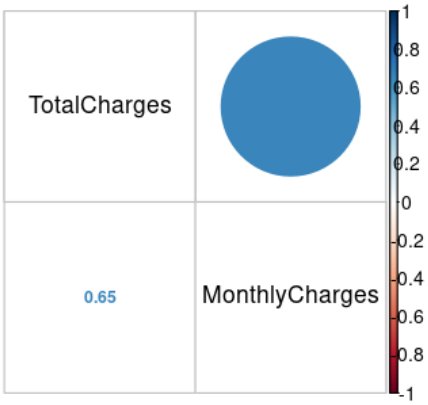
```
 $ StreamingMovies : chr [1:7032] "No internet service" "No internet service" "No internet service" ...
 $ Contract        : chr [1:7032] "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr [1:7032] "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr [1:7032] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ TenureYear      : chr [1:7032] "0-1 ano" "2-3 anos" "0-1 ano" "3-4 anos" ...
 $ MonthlyCharges  : num [1:7032] 29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num [1:7032] 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr [1:7032] "No" "No" "Yes" "No" ...
 $ Tenure          : num [1:7032] 1 34 2 45 2 8 22 10 28 62 ...
```

2.4.2 Justificativa para a retirada da variável Totalcharges

Os valores MonthlyCharges e TotalCharges representam o quanto é gasto com os serviços de telefonia e de internet, ou seja existem duas variáveis para a mesma finalidade. Porém, com a constatação da Correlação entre os valores, não há necessidade da coexistência de ambos os valores para a modelagem, assim sendo, a variável TotalCharges será removida do dataset.

A justificativa de manter-se o valor mensal gasto, e retirar o valor total, é para evitar a multicolinearidade, além disso a variável TotalCharges tem maior range de valores. Abaixo, reforçando as demonstrações, numérica e gráfica, da Correlação entre os dois valores.

```
# Validação da correlação numérica entre os valores TotalCharges e MonthlyCharges.
Corr_TotalXMonthly <- cor(churn_raw_new[c('MonthlyCharges','TotalCharges')], use= 'complete')
Corr_TotalXMonthly.melted <- melt(Corr_TotalXMonthly)
Corr_TotalXMonthly
>
               MonthlyCharges TotalCharges
MonthlyCharges      1.0000000    0.6510648
TotalCharges        0.6510648    1.0000000

# Gráfico da correlação.
churn_raw_new %>%
  dplyr::select (TotalCharges, MonthlyCharges) %>%
  cor() %>%
  corrplot.mixed(upper = "circle",
                 tl.col = "black",
                 number.cex = 0.7)
```



Sumário dos dois valores.

```
# Sumário da variável TotalCharges e MonthlyCharges.
summary(churn_raw_new$TotalCharges)
>
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   18.8   401.4  1397.5  2283.3  3794.7  8684.8

summary(churn_raw_new$MonthlyCharges)
>
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.25   35.59   70.35   64.80   89.86  118.75
```

```
#--- Verificando que há uma maior amplitude entre os valores totais do que entre os valores mensais.
```

3 Cleaning final

O dataset [**churn_raw_new**] após os recondicionamentos e as tratativas sofrerá por fim o clenaning com a remoção das colunas CustomerId, Tenure e TotalCharges, que não serão mais necessárias, e não farão parte de nenhuma estrutura da modelagem.

Bem como terá a adição de uma coluna de identificação do Churn, em formato binomial numérico ao invés de apenas categórico.

3.1 Dataset final

Após o cleaning do dataset [**churn_raw_new**] será renomeado para [**churn_clean**].

```
# Alteração do dataset.
churn_clean <- churn_raw_new

# Remoção das colunas tratadas e que não serão mais necessárias para a modelagem.

# Remoção da coluna MonthlyCharges.
churn_clean$TotalCharges <- NULL

# Remoção da coluna Tenure.
churn_clean$Tenure <-NULL

# Remoção da coluna CustomerID.
churn_clean$CustomerID <- NULL
```

3.2 Coluna ChurnBin

A coluna Churn é categórica com valores Yes|No e para alguma modelagem será conveniente que os dados estejam em formato binomial de 0|1. A fim de gerar condições para essa situação, será adicionada uma coluna ChurnBin com esses valores.

```
# Criando uma nova coluna atribuindo valores 0|1 para No|Yes.
churn_clean$ChurnBin = 0
churn_clean$ChurnBin[churn_clean$Churn == "Yes"] = 1
churn_clean$ChurnBin[churn_clean$Churn == "No"] = 0

# CrossTable entre Churn e ChurnBin.
CrossTable(churn_clean$Churn, prop.t=FALSE, prop.r=FALSE, prop.c=FALSE)
>
Total Observations in Table:  7032


          |         0 |         1 |
          |-----------|-----------|
          |      5163 |      1869 |
          |     0.734 |     0.266 |
          |-----------|-----------|

CrossTable(churn_clean$ChurnBin, prop.t=FALSE, prop.r=FALSE, prop.c=FALSE)
>
Total Observations in Table:  7032


          |         0 |         1 |
          |-----------|-----------|
          |      5163 |      1869 |
          |     0.734 |     0.266 |
          |-----------|-----------|
```

Verificação final do dataset [**churn_clean**] com as variáveis dispostas conforme abaixo:

Demográficos	Serviços Telefonia	Serviços Internet	Contratuais	Comportamentais
Gender	PhoneService	InternetService	Contract	Churn
SeniorCitizen	MultipleLines	OnlineSecurity	PaperlessBilling	
Partner		OnlineBackup	PaymentMethod	

Demográficos	Serviços Telefonia	Serviços Internet	Contratuais	Comportamentais
Dependents		DeviceProtection	TenureYear	
		TechSupport	TotalCharges	
		StreamingTV		
		StreamingMovies		

```
# Dimensão do conjunto de dados.
dim(churn_clean)
>
[1] 7032   20

# Também reconhece os valores NA e retornam sua proporção.
df_status(churn_clean)
>
           variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
1            Gender       0    0.00    0    0     0     0 character      2
2     SeniorCitizen       0    0.00    0    0     0     0 character      2
3           Partner       0    0.00    0    0     0     0 character      2
4        Dependents       0    0.00    0    0     0     0 character      2
5      PhoneService       0    0.00    0    0     0     0 character      2
6     MultipleLines       0    0.00    0    0     0     0 character      2
7   InternetService       0    0.00    0    0     0     0 character      3
8    OnlineSecurity       0    0.00    0    0     0     0 character      2
9      OnlineBackup       0    0.00    0    0     0     0 character      2
10 DeviceProtection       0    0.00    0    0     0     0 character      2
11      TechSupport       0    0.00    0    0     0     0 character      2
12      StreamingTV       0    0.00    0    0     0     0 character      2
13  StreamingMovies       0    0.00    0    0     0     0 character      2
14         Contract       0    0.00    0    0     0     0 character      3
15 PaperlessBilling       0    0.00    0    0     0     0 character      2
16    PaymentMethod       0    0.00    0    0     0     0 character      4
17      TenureYear        0    0.00    0    0     0     0 character      6
18     TotalCharges       0    0.00    0    0     0     0   numeric   6530
19            Churn       0    0.00    0    0     0     0 character      2
20         ChurnBin    5163   73.42    0    0     0     0   numeric      2


# Verificação final da estrutura do novo dataset, tratado e limpo.
str(churn_clean)
>
tibble [7,032 × 20] (S3: tbl_df/tbl/data.frame)
 $ Gender          : chr [1:7032] "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : chr [1:7032] "No" "No" "No" "No" ...
 $ Partner         : chr [1:7032] "Yes" "No" "No" "No" ...
 $ Dependents      : chr [1:7032] "No" "No" "No" "No" ...
 $ PhoneService    : chr [1:7032] "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr [1:7032] "No" "No" "No" "No" ...
 $ InternetService : chr [1:7032] "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr [1:7032] "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr [1:7032] "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr [1:7032] "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr [1:7032] "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr [1:7032] "No" "No" "No" "No" ...
 $ StreamingMovies : chr [1:7032] "No" "No" "No" "No" ...
 $ Contract        : chr [1:7032] "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr [1:7032] "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr [1:7032] "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
 $ TenureYear      : chr [1:7032] "0-1 ano" "2-3 anos" "0-1 ano" "3-4 anos" ...
 $ TotalCharges    : num [1:7032] 29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr [1:7032] "No" "No" "Yes" "No" ...
 $ ChurnBin        : num [1:7032] 0 0 1 0 1 1 0 0 1 0 ...
```