EDA PÓS CLEANING

> 💡 **Notion Tip:**
> Uso do dataset [churn_clean] para as demonstrações das análises exploratórias.

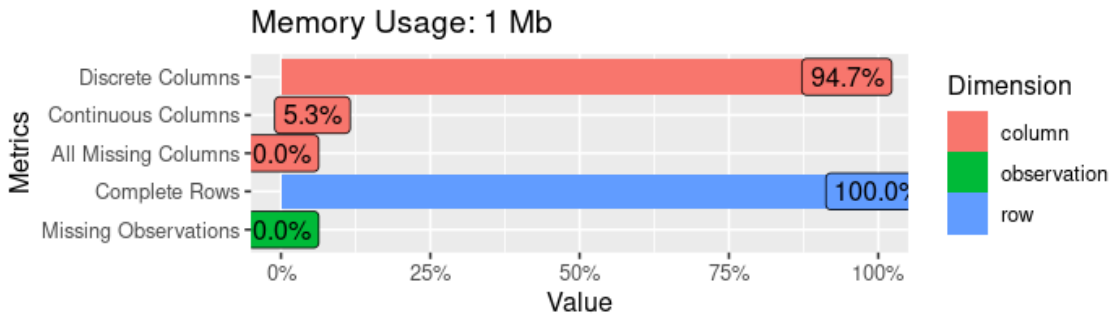
1. Sumário do dataset [churn_clean]

Antes do processo de modelagem, um exame analítico acerca dos dados, após recoding e cleaning, com as verificações das relações entre a variável alvo e as demais variáveis de interesse.

1.1 Verificação do dataset [churn_clean]

O dataset foi limpo e conta atualmente com 21 variáveis, sendo 4 numéricas e 17 categóricas. Sendo TenureYear uma variável numérica réplica da variável Tenure.

```
# Verificando a dimensão do dataset.
dim(churn_clean)
[1] 7032   21

# Simples conferência do dataset.
plot_intro(churn_clean)
```




```
# Aplicando describe aos dados do dataset.
Hmisc::describe(churn_clean)
>
```

```
churn_clean

 21  Variables      7032  Observations
--------------------------------------------------------------------------
Gender
       n  missing distinct
    7032        0        2

Value      Female   Male
Frequency    3483   3549
Proportion  0.495  0.505
--------------------------------------------------------------------------
SeniorCitizen
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   5890  1142
Proportion 0.838 0.162
--------------------------------------------------------------------------
Partner
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   3639  3393
Proportion 0.517 0.483
--------------------------------------------------------------------------
Dependents
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4933  2099
Proportion 0.702 0.298
---------------------------------------------------------------------------------------------------------
PhoneService
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency    680  6352
Proportion 0.097 0.903
---------------------------------------------------------------------------------------------------------
MultipleLines
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4065  2967
Proportion 0.578 0.422
---------------------------------------------------------------------------------------------------------
InternetService
       n  missing distinct
    7032        0        3

Value              DSL Fiber optic          No
Frequency         2416        3096        1520
Proportion       0.344       0.440       0.216
---------------------------------------------------------------------------------------------------------
OnlineSecurity
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   5017  2015
```

```
Proportion 0.713 0.287
---------------------------------------------------------------------------------------------------------
OnlineBackup
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4607  2425
Proportion 0.655 0.345
---------------------------------------------------------------------------------------------------------
DeviceProtection
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4614  2418
Proportion 0.656 0.344
---------------------------------------------------------------------------------------------------------
TechSupport
       n  missing distinct
    7032        0        2

Value        No  Yes
Frequency  4992 2040
Proportion 0.71 0.29
---------------------------------------------------------------------------------------------------------
StreamingTV
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4329  2703
Proportion 0.616 0.384
---------------------------------------------------------------------------------------------------------
StreamingMovies
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   4301  2731
Proportion 0.612 0.388
---------------------------------------------------------------------------------------------------------
Contract
       n  missing distinct
    7032        0        3

Value      Month-to-month       One year       Two year
Frequency            3875           1472           1685
Proportion          0.551          0.209          0.240
---------------------------------------------------------------------------------------------------------
PaperlessBilling
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   2864  4168
Proportion 0.407 0.593
---------------------------------------------------------------------------------------------------------
PaymentMethod
       n  missing distinct
    7032        0        4

Value      Bank transfer (automatic)   Credit card (automatic)     Electronic check     Mailed check
Frequency                       1542                      1521                 2365             1604
Proportion                     0.219                     0.216                0.336            0.228
---------------------------------------------------------------------------------------------------------
```

```
TenureYear
       n  missing distinct
    7032        0        6

lowest : 0-1 ano  highest:  5-6 anos

Value       0-1 ano 1-2 anos 2-3 anos 3-4 anos 4-5 anos 5-6 anos
Frequency      2175     1024      832      762      832     1407
Proportion    0.309    0.146    0.118    0.108    0.118    0.200
-------------------------------------------------------------------------------------------------------
MonthlyCharges
       n  missing distinct     Info     Mean      Gmd      .05      .25      .50      .75      .95
    7032        0     1584        1     64.8    34.38    19.65    35.59    70.35    89.86   107.42

lowest :  18.25  18.40  18.55  18.70  18.75, highest: 118.20 118.35 118.60 118.65 118.75
-------------------------------------------------------------------------------------------------------
TotalCharges
       n  missing distinct     Info     Mean      Gmd      .05      .25      .50      .75      .95
    7032        0     6530        1     2283     2449     49.6    401.4   1397.5   3794.7   6923.6

lowest :   18.80   18.85   18.90   19.00   19.05, highest: 8564.75 8594.40 8670.10 8672.45 8684.80
-------------------------------------------------------------------------------------------------------
Churn
       n  missing distinct
    7032        0        2

Value         No   Yes
Frequency   5163  1869
Proportion 0.734 0.266
-------------------------------------------------------------------------------------------------------
Tenure
       n  missing distinct     Info     Mean      Gmd      .05      .25      .50      .75      .95
    7032        0       72    0.999    32.42    28.07        1        9       29       55       72

lowest :  1  2  3  4  5, highest: 68 69 70 71 72
-------------------------------------------------------------------------------------------------------
```

A função [describe], para o dataset, demonstra:

- que todas as variáveis contam uniformemente com a mesma quantidade de dados;
- que não consta nenhum registro faltante (missing);
- quantos valores distintos cada variável possui;
- a distribuição da frequência quantitativa dos valores de cada variável;
- para as variáveis numéricas, um resumo estatístico sobre dos valores.

```
# Verificação apenas das variáveis numéricas.
> profiling_num(churn_clean)
        variable       mean    std_dev variation_coef
1 MonthlyCharges   64.79821   30.08597      0.4643026
2   TotalCharges 2283.30044 2266.77136      0.9927609
3         Tenure   32.42179   24.54526      0.7570607

  p_01   p_05     p_25     p_50      p_75      p_95      p_99
1 19.2 19.650  35.5875   70.350   89.8625  107.4225  114.7345
2 19.9 49.605 401.4500 1397.475 3794.7375 6923.5900 8039.8830
3  1.0  1.000   9.0000   29.000   55.0000   72.0000   72.0000

     skewness kurtosis      iqr           range_98           range_80
1  -0.2220555 1.743883   54.275   [19.2, 114.7345]   [20.05, 102.645]
```

```
2   0.9614374 2.767513 3393.288   [19.9, 8039.883]    [84.6, 5976.64]
3   0.2376801 1.612311   46.000            [1, 72]            [2, 69]
```

As análises gráficas serão auxiliares para o levantamento sobre o comportamento dos clientes, churners e não-churners, de cada situação, e validação de quais variáveis do conjunto trarão maior relevância para modelagem.

```
# Verificação da estrutura.
> churn_clean %>% str()

churn_clean
tibble [7,032 × 21] (S3: tbl_df/tbl/data.frame)
 $ Gender          : chr [1:7032] "Male" "Female" "Female" "Female" ...
 $ SeniorCitizen   : chr [1:7032] "No" "No" "No" "No" ...
 $ Partner         : chr [1:7032] "No" "No" "No" "Yes" ...
 $ Dependents      : chr [1:7032] "No" "No" "No" "No" ...
 $ PhoneService    : chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...
 $ MultipleLines   : chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ InternetService : chr [1:7032] "DSL" "Fiber optic" "Fiber optic" "Fiber optic" ...
 $ OnlineSecurity  : chr [1:7032] "Yes" "No" "No" "No" ...
 $ OnlineBackup    : chr [1:7032] "Yes" "No" "No" "No" ...
 $ DeviceProtection: chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ TechSupport     : chr [1:7032] "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ StreamingMovies : chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ Contract        : chr [1:7032] "Month-to-month" "Month-to-month" "Month-to-month" "Month-to-month" ...
 $ PaperlessBilling: chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...
 $ PaymentMethod   : chr [1:7032] "Mailed check" "Electronic check" "Electronic check" "Electronic check" ...
 $ Tenure          : num [1:7032] 2 2 8 28 49 10 1 1 47 1 ...
 $ TenureYear      : chr [1:7032] "0-1 ano" "0-1 ano" "0-1 ano" "2-3 anos" ...
 $ MonthlyCharges  : num [1:7032] 53.9 70.7 99.7 104.8 103.7 ...
 $ TotalCharges    : num [1:7032] 108 152 820 3046 5036 ...
 $ Churn           : chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...
```

---

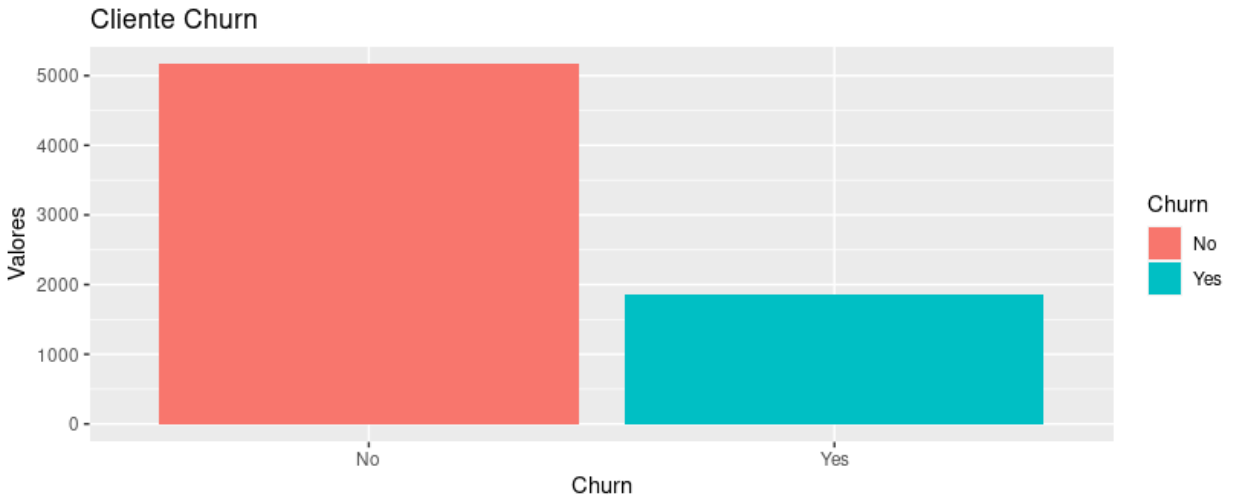
2. Análises Gráficas

As análises gráfica serão auxiliares para o levantamento sobre o comportamento dos clientes, churners e não-churners, de cada situação, e validação de quais variáveis do conjunto trarão maior relevância para modelagem.

2.1 Variável alvo: Churn

Sendo Churn a variável de interesse, a primeira visualização é da distribuição dos clientes que saíram (churners) e dos que permaneceram (não-churners).

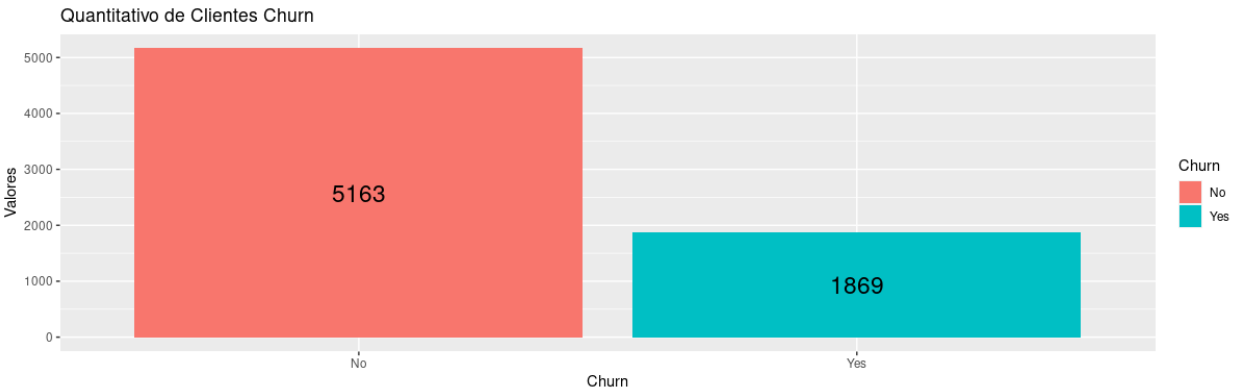
```
# Plotagem com uma visualização simples da composição dos valores (YES | NO).
ggplot(churn_clean, aes(x = Churn)) +
  geom_bar(aes(fill = Churn)) +
  labs(title = "Cliente Churn",
       x = "Churn",
       y = "Valores")
```



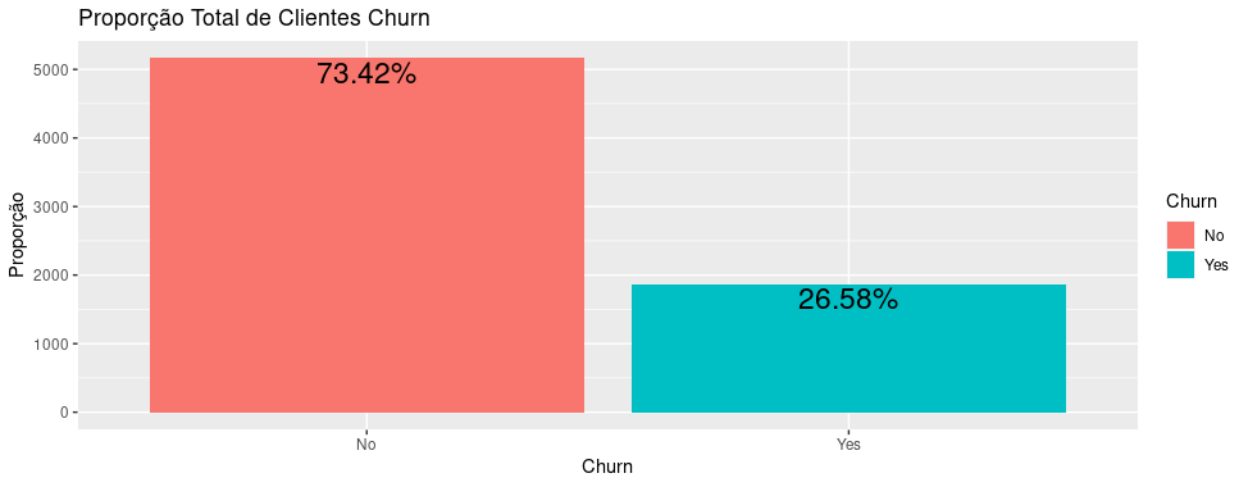
Detalhamento da frequência de Churn X ChurnBin mostrando que ambas contém os mesmos dados, apenas com informações diferentes.

```
# Simples contagem de conferência entre as variáveis.
count(churn_clean,c('Churn', 'ChurnBin'))
>
  Churn ChurnBin freq
1    No        0 5163
2   Yes        1 1869
```

```
# Plotagem dos dados quantitativos dos clientes churners e não-churners.
ggplot(churn_clean, aes(x = Churn)) +
  geom_bar(aes(fill = Churn)) +
  geom_text(aes(label = ..count..),
            stat = "count",
            position=position_stack(vjust=0.5),
            size = 6) +
  labs(title = "Quantitativo de Clientes Churn",
       x = "Churn",
       y = "Valores")
```



```
# Plotagem dos dados percentuais dos clientes churners e não-churners.
ggplot(churn_clean, aes(x = Churn)) +
  geom_bar(aes(fill = Churn)) +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3) +
  labs(title = "Proporção Total de Clientes Churn",
       x = "Churn",
       y = "Proporção")

#-- Usado valores contínuos de frequência intencionalemnte.
```



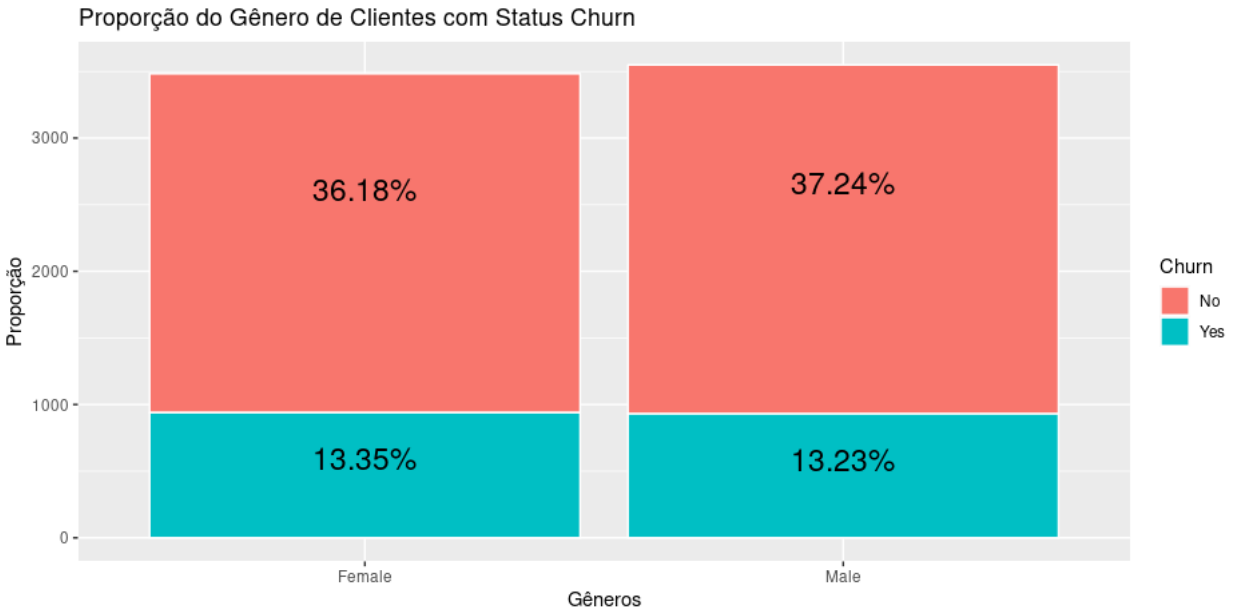
Com base no dados do gráfico pode-se observar que há uma proporção muito menor de clientes que saíram (churners) do que os clientes que permaneceram (não-churners), com aproximadamente 27% de churners do total de registros.

Essa fato não será considerado um viés na modelagem, já que essa proporção é bem realista, em se tratando de empresas de telecomunicações. Portanto, não há preocupações com o balanceamento do conjunto de dados.

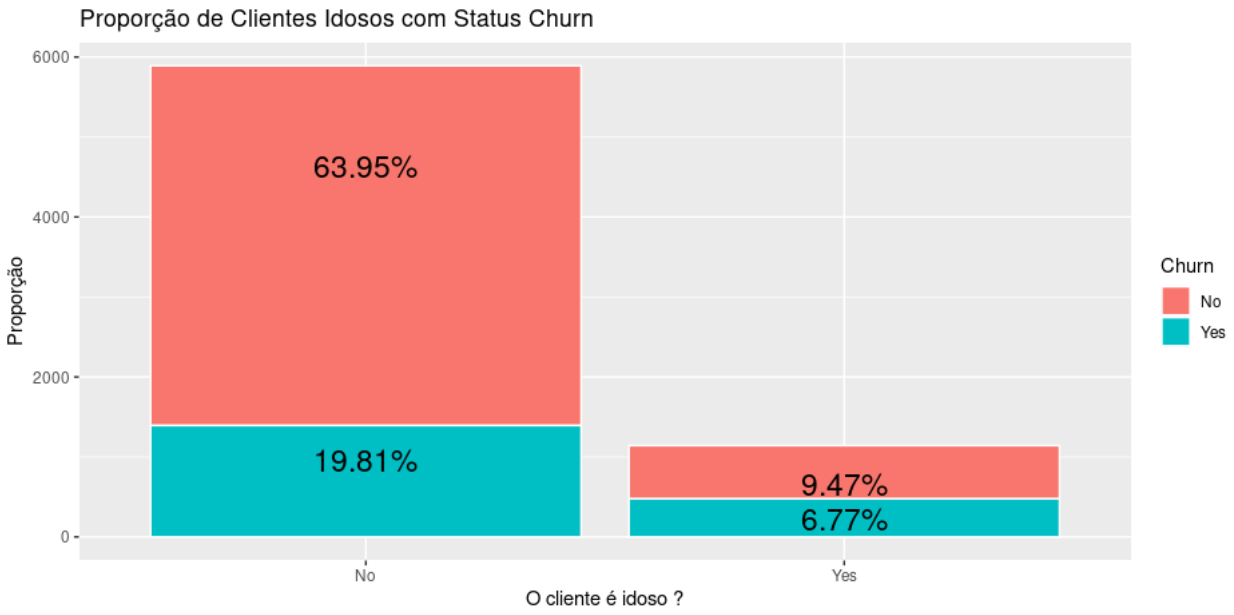
2.2 Dados Demográficos

Verificação dos percentuais de churners dentro das variáveis demográficas.

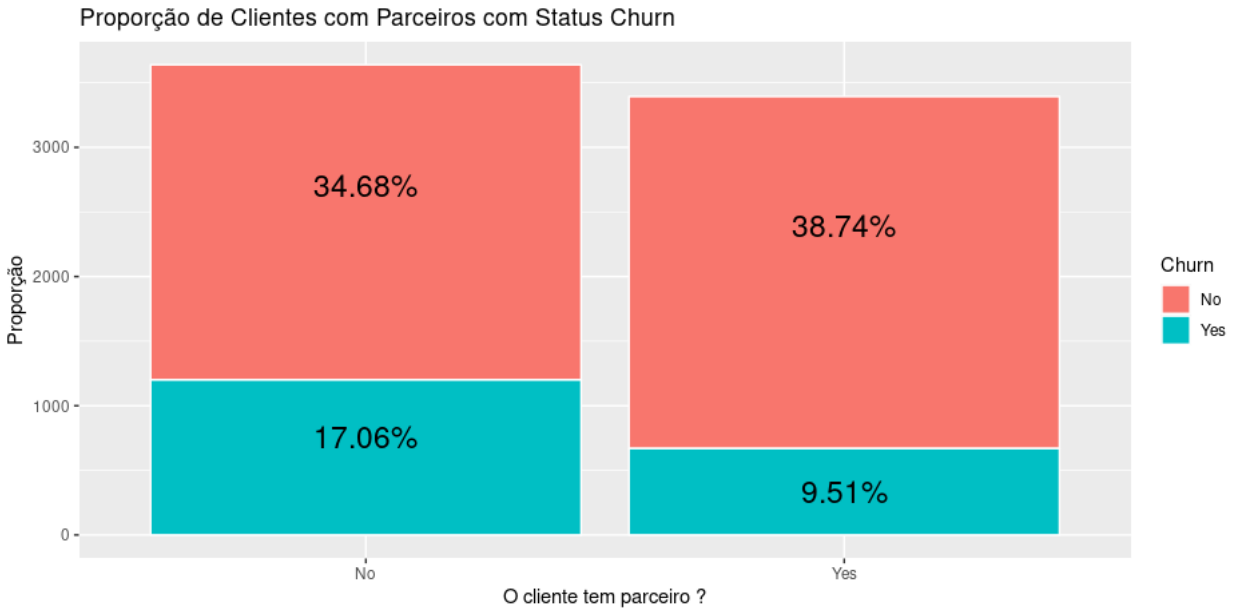
```
# Plotagem stacked com Percentuais sobre a frequencia.

# Gênero X Churn.
ggplot(churn_clean,aes(x = Gender, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.8),
            size = 6) +
  labs(title = "Proporção do Gênero de Clientes com Status Churn",
       x = "Gêneros",
       y = "Proporção")
```



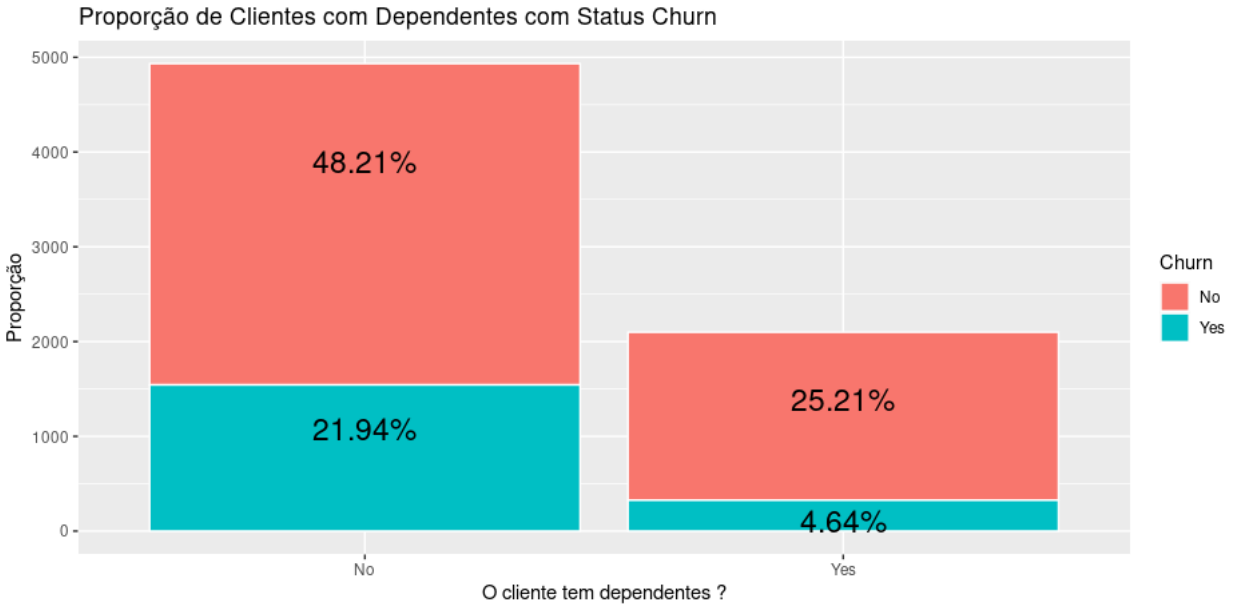

```
# Idoso X Churn.
ggplot(churn_clean,aes(x = SeniorCitizen, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.8),
            size = 6) +
  labs(title = "Proporção de Clientes Idosos com Status Churn",
       x = "O cliente é idoso ?",
       y = "Proporção")
```




```
# Parceiro X Churn.
ggplot(churn_clean,aes(x = Partner, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -250,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.8),
            size = 6) +
  labs(title = "Proporção de Clientes com Parceiros com Status Churn",
       x = "O cliente tem parceiro ?",
       y = "Proporção")
```



```
# Dependentes X Churn.
ggplot(churn_clean, aes(Dependents, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.8),
            size = 6) +
  labs(title = "Proporção de Clientes com Dependentes com Status Churn",
       x = "O cliente tem dependentes ?",
       y = "Proporção")
```

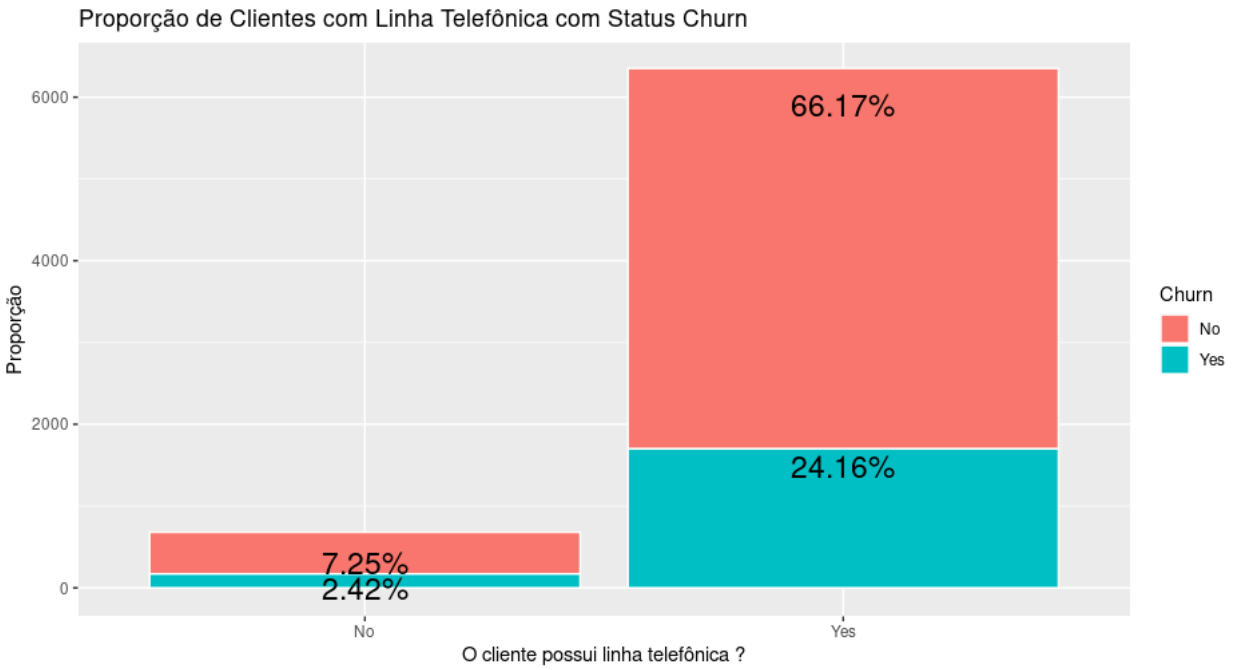


Com base nos gráficos, pode-se observar que:

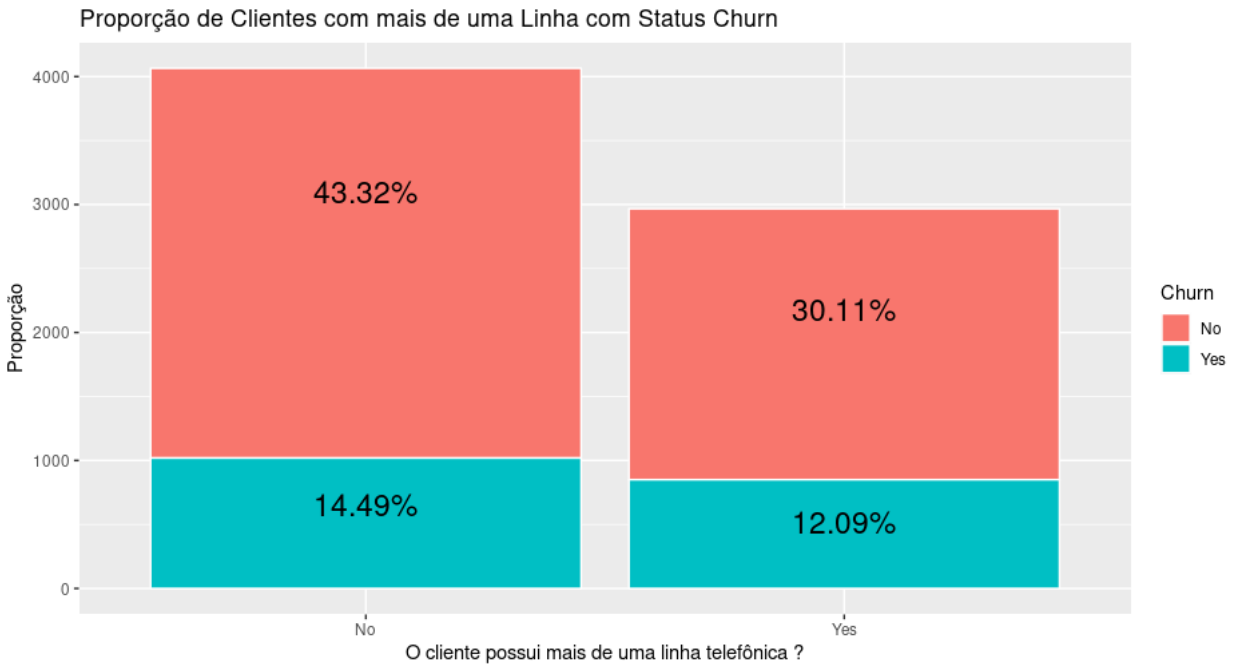
- a amostra de churners está equilibrada entre os gêneros;
- a amostra de chuners tem maior proporção para clientes com menos de 65 anos;
- para clientes sem parceiros e sem dependentes há maior proporção de churners;

2.3 Dados Serviços Telefônicos

```
# Compartivo entre dados serviços telefônicos e Churn.

# Telefonia X Churn.
ggplot(churn_clean, aes(PhoneService, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com Linha Telefônica com Status Churn",
       x = "O cliente possui linha telefônica ?",
       y = "Proporção")
```




```
# Mais de 1 Linha  X Churn.
ggplot(churn_clean, aes(MultipleLines, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.8),
            size = 6) +
  labs(title = "Proporção de Clientes com mais de uma Linha com Status Churn",
       x = "O cliente possui mais de uma linha telefônica ?",
       y = "Proporção")
```

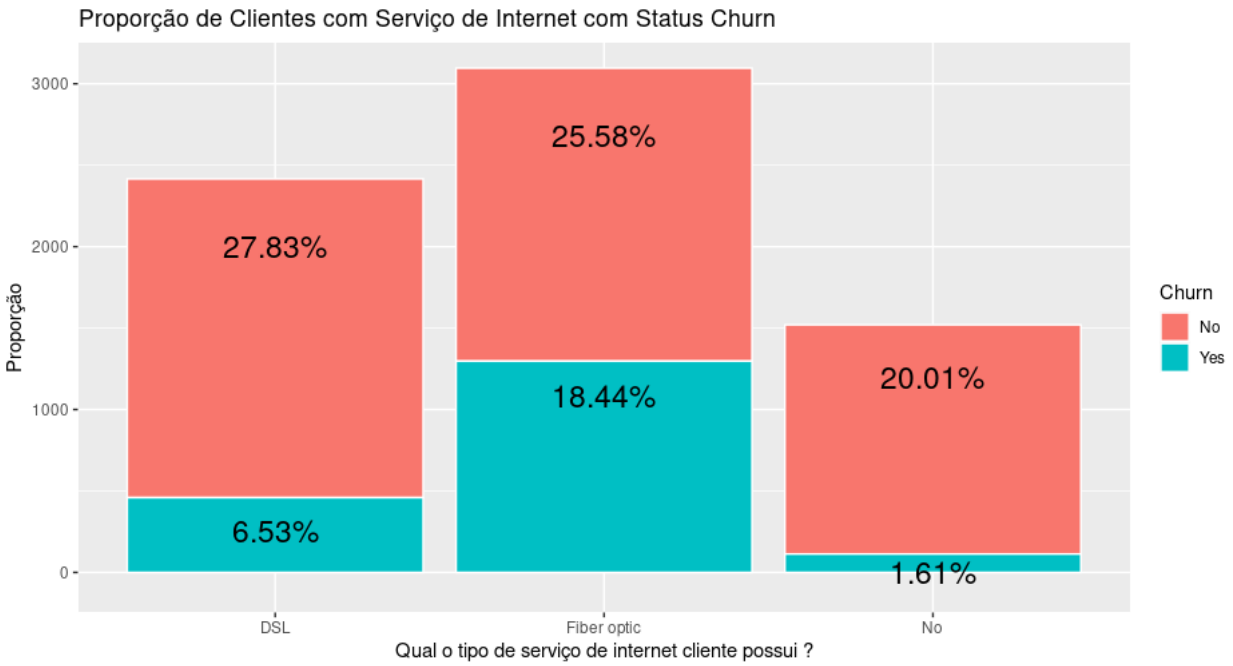


Com base nos gráficos, pode-se observar que:

- a amostra de chuners é maior para clientes que possuem serviço telefônico;
- a amostra de churners está equilibrada para clientes que possuem uma ou mais linhas.

2.4 Dados Serviços Internet

```
# Compartivo entre dados serviços de internet e Churn.

# Serviço de Internet X Churn.
ggplot(churn_clean, aes(InternetService, fill = Churn)) +
geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com Serviço de Internet com Status Churn",
       x = "Qual o tipo de serviço de internet cliente possui ?",
       y = "Proporção")
```

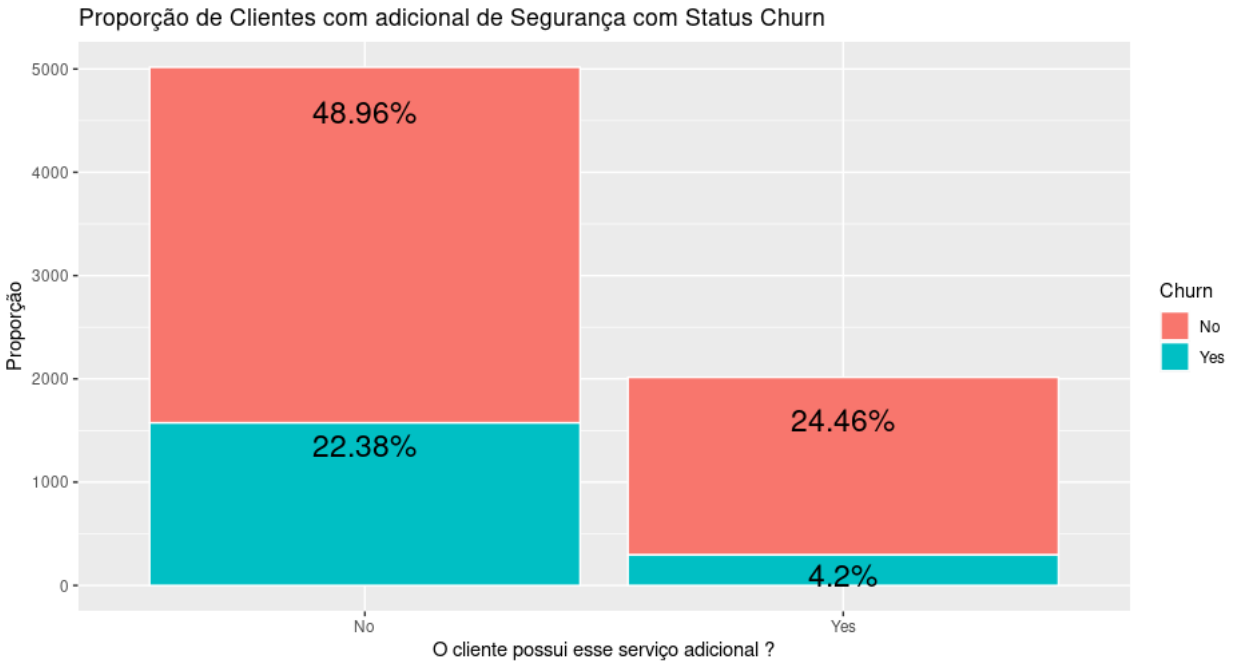


Com base nos gráficos, pode-se observar que:

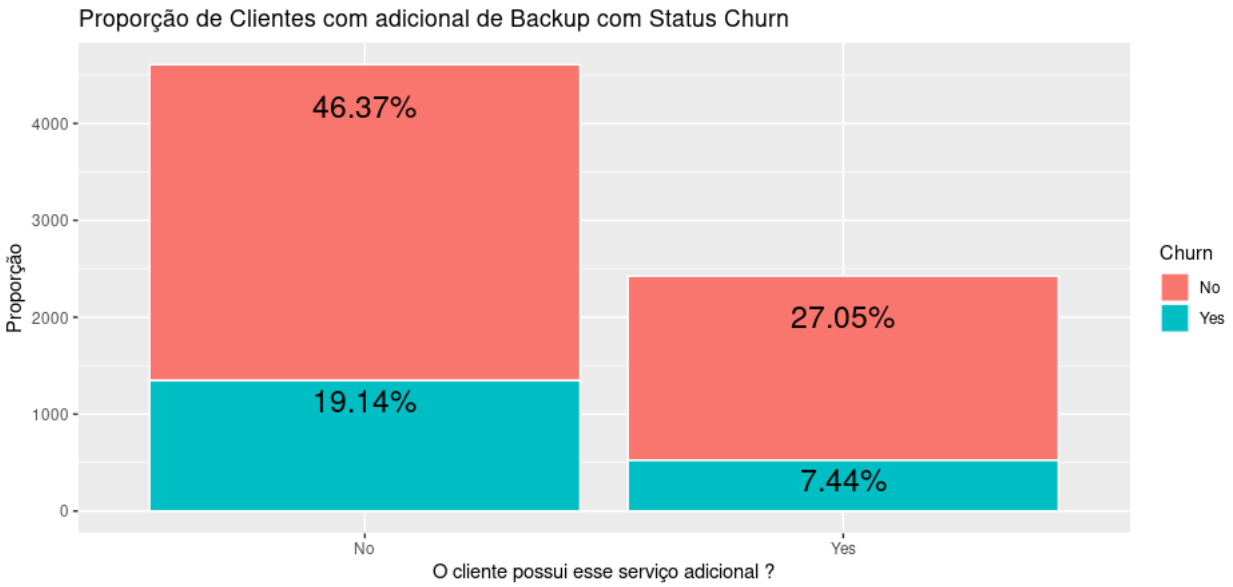
- a amostra de churners tem maior proporção para clientes com conexão por fibra óptica.

Separando os gráficos dos serviço adicionais de internet, que estão todos atrelados ao tipo de conexão que cada cliente possui.

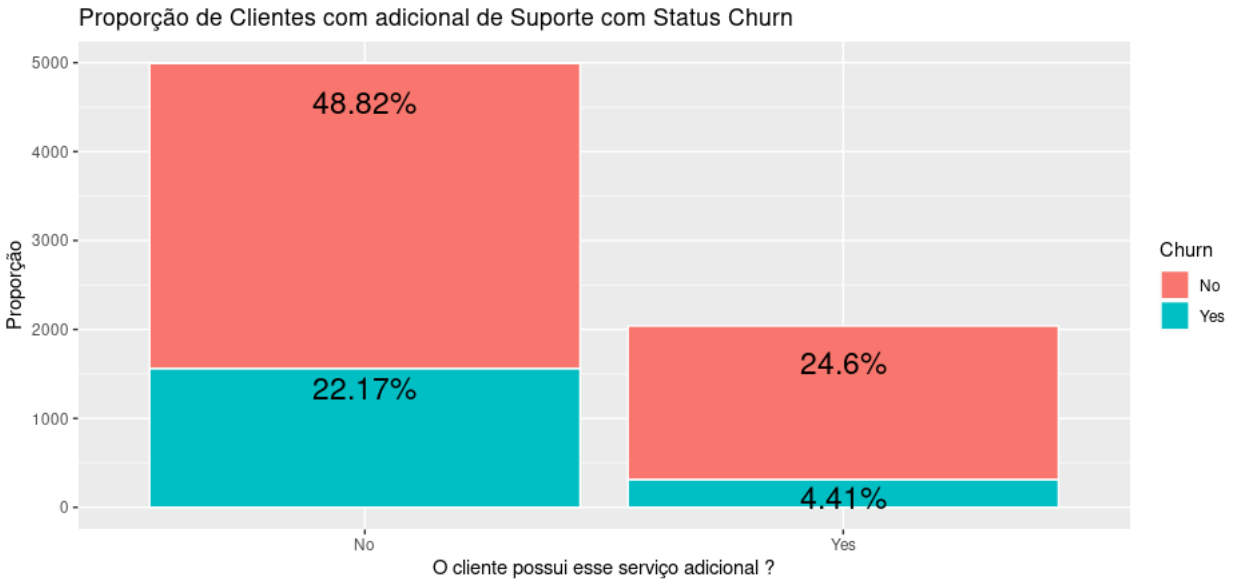
```
# Compartivo entre dados serviços adicionais de internet e Churn.

# Plotagem para Online Security.
ggplot(churn_clean, aes(OnlineSecurity, fill = Churn)) +
 geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com adicional de Segurança com Status Churn",
       x = "O cliente possui esse serviço adicional ?",
       y = "Proporção")
```



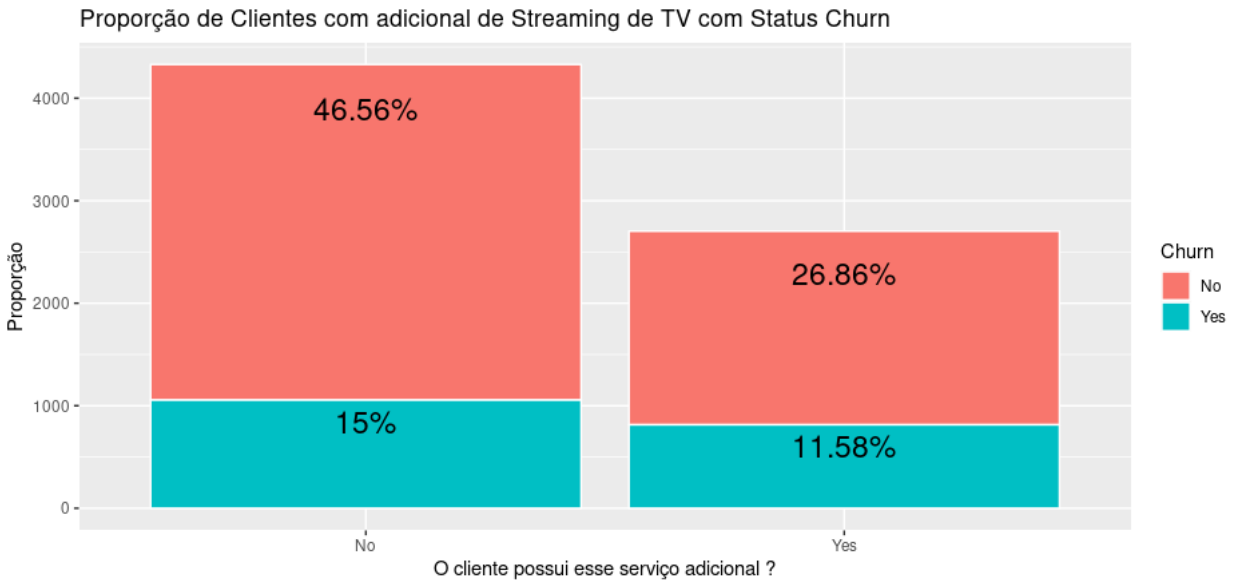

```
# Compartivo entre dados serviços adicionais de internet e Churn.

# Plotagem para Online Backup.
ggplot(churn_clean, aes(OnlineBackup, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com adicional de Backup com Status Churn",
       x = "O cliente possui esse serviço adicional ?",
       y = "Proporção")
```



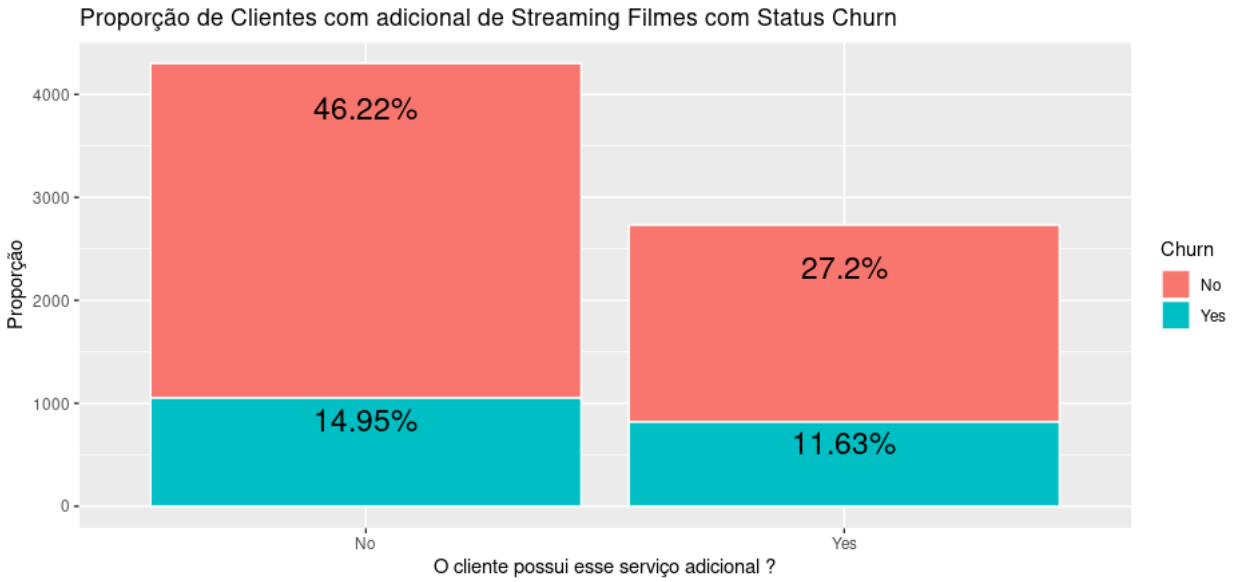

```
# Compartivo entre dados serviços adicionais de internet e Churn.

# Plotagem para Tech Support.
ggplot(churn_clean, aes(OnlineBackup, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com adicional de Suporte com Status Churn",
       x = "O cliente possui esse serviço adicional ?",
       y = "Proporção")
```



```
# Compartivo entre dados serviços adicionais de internet e Churn.

# Plotagem para Streaming TV.
ggplot(churn_clean, aes(StreamingTV, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com adicional de Streaming de TV com Status Churn",
       x = "O cliente possui esse serviço adicional ?",
       y = "Proporção")
```




```
# Compartivo entre os serviços adicionais de internet e Churn.

# Plotagem para Streaming Movies.
ggplot(churn_clean, aes(StreamingMovies, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com adicional de Streaming Filmes com Status Churn",
       x = "O cliente possui esse serviço adicional ?",
       y = "Proporção")
```

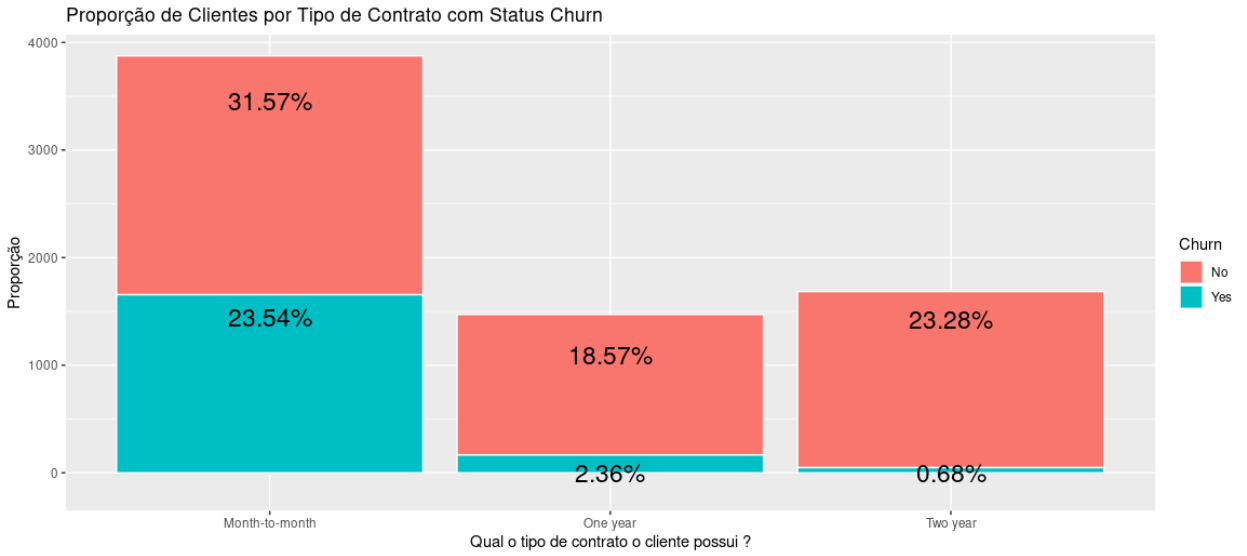


Com base nos gráficos, pode-se observar que:

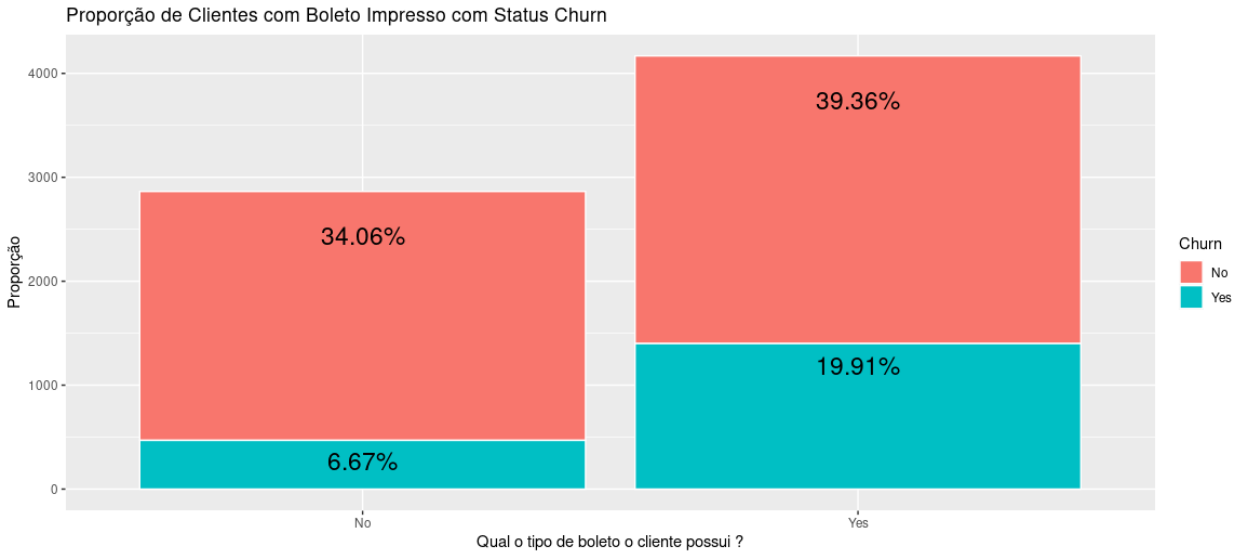
- há maior proporção de churners para os clientes que não têm serviços adicionais habilitados em seus planos de internet.

2.5 Dados Contratuais

```
# Compartivo entre dados contratuais e Churn.

# Plotagem para Contract.
ggplot(churn_clean, aes(Contract, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes por Tipo de Contrato com Status Churn",
       x = "Qual o tipo de contrato o cliente possui ?",
       y = "Proporção")
```

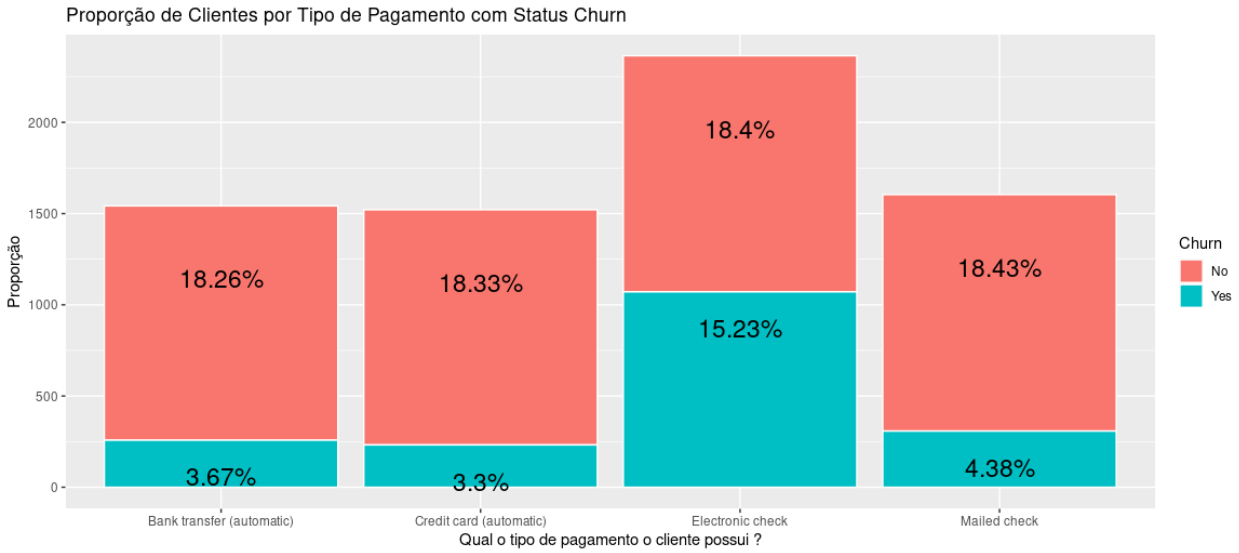



```
# Plotagem para Paperless Billing.
ggplot(churn_clean, aes(PaperlessBilling, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes com Boleto Impresso com Status Churn",
       x = "Qual o tipo de boleto o cliente possui ?",
       y = "Proporção")
```




```
# Plotagem para Payment Method.
ggplot(churn_clean, aes(PaymentMethod, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
```

```
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_stack(.99),
            size = 6) +
  labs(title = "Proporção de Clientes por Tipo de Pagamento com Status Churn",
       x = "Qual o tipo de pagamento o cliente possui ?",
       y = "Proporção")
```



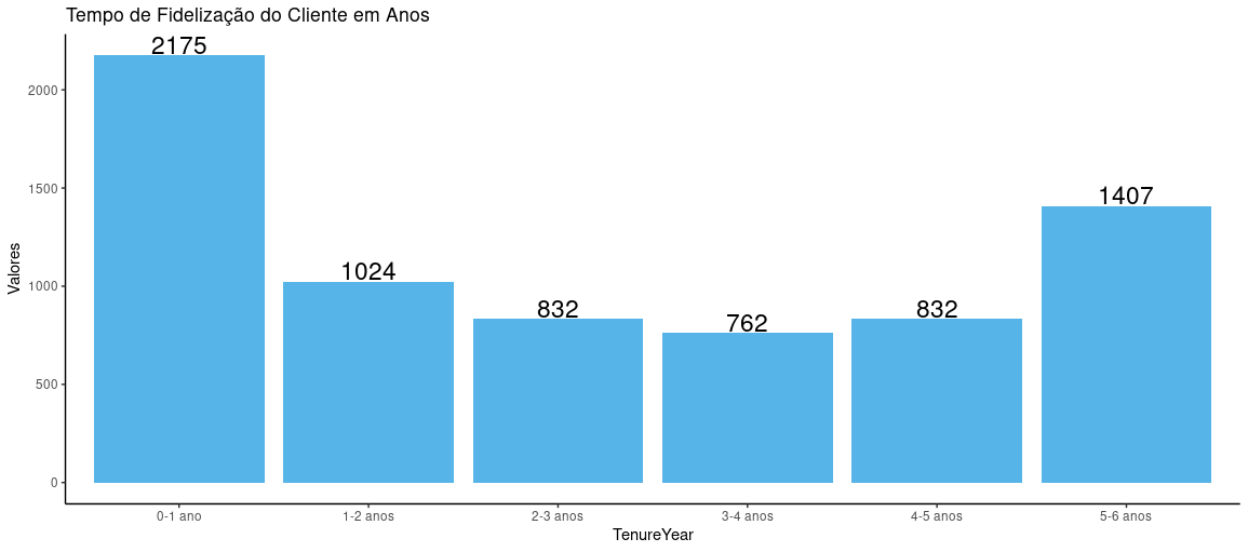
Com base nesses gráficos, pode-se observar que:

- a amostra de churners tem maior proporção para clientes com subscrição mensal (pré-pago);
- a amostra de churners tem maior proporção para clientes com fatura do tipo online;
- a amostra de churners tem maior proporção para clientes que efetuam pagamento online.

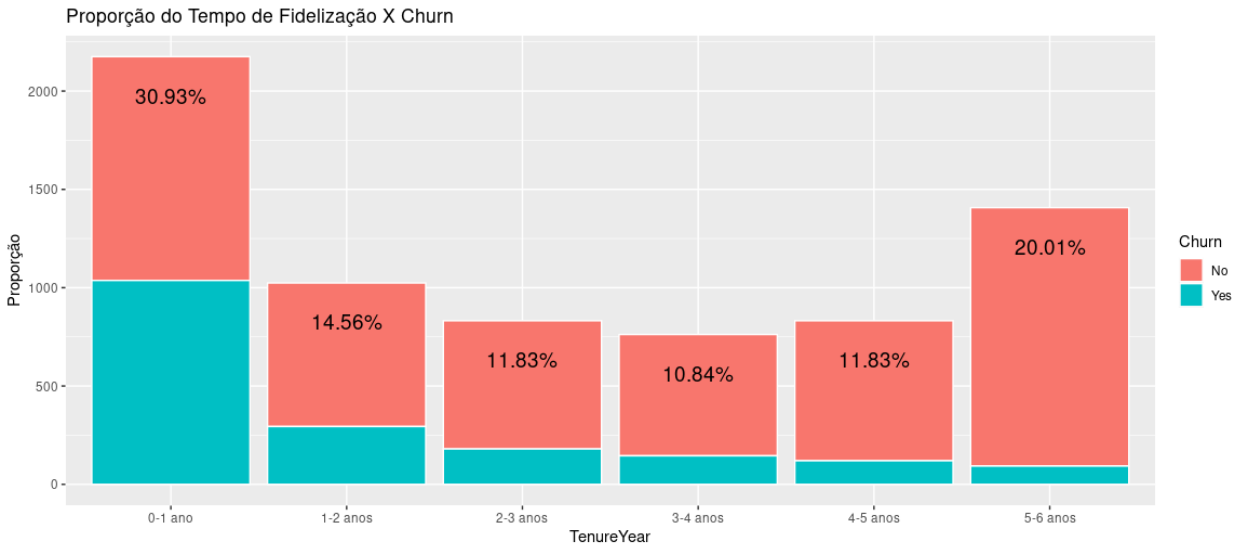
2.6 Dados Tempo de Fidelização

Visualização do quantitativo de clientes em cada ano:

```
# Distribuição da frequência de clientes entre os tempos de contrato estratificados.
base::table(churn_clean$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

# Grafico com as frequencias por ano.
ggplot(churn_clean, aes(churn_clean$TenureYear)) +
  geom_bar(fill = "#56B4E9") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.1) +
  theme_classic() +
  labs(title = "Tempo de Fidelização do Cliente em Anos",
       x = "TenureYear",
       y = "Valores")
```

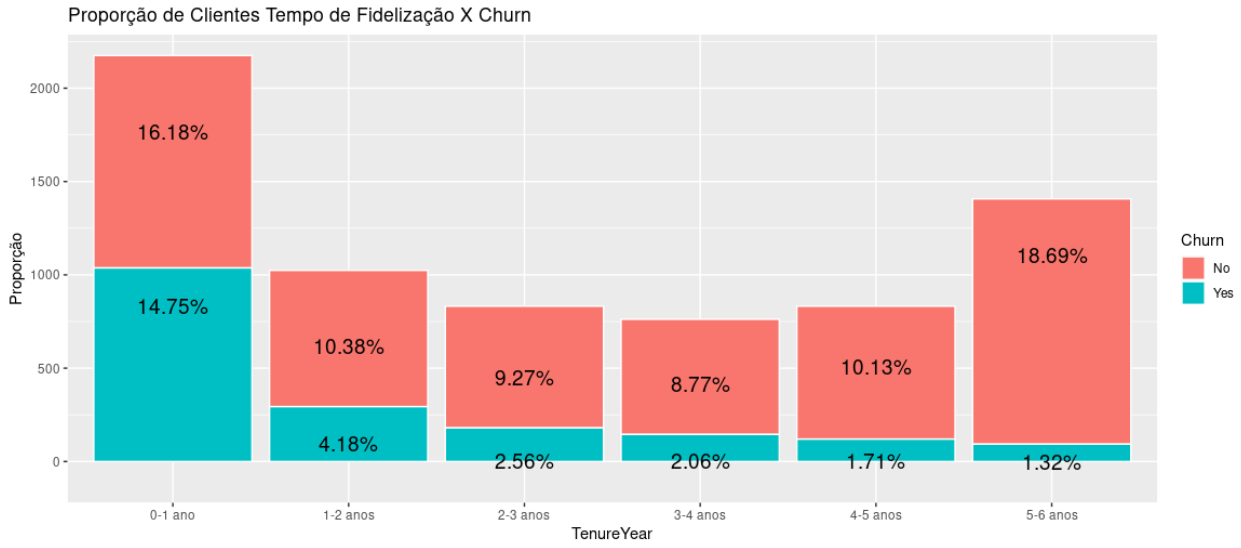



```
# # Plotagem para percentual de clientes por Tenure Year.
ggplot(churn_clean, aes(x = TenureYear)) +
  geom_bar(aes(fill = Churn), colour = 'white') +
  geom_text(aes(y = ..count.. -200,
            label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 5) +
  labs(title = "Proporção do Tempo de Fidelização X Churn",
       x = "TenureYear",
       y = "Proporção")
```

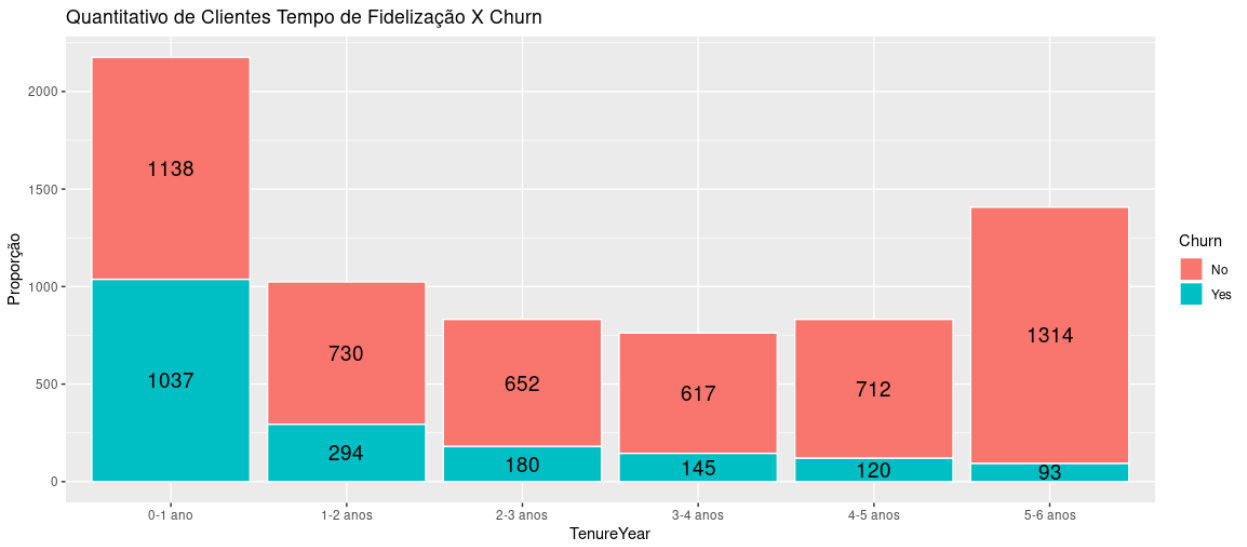



```
# Plotagem do percentual de clientes por ano X Churn.
ggplot(churn_clean,aes(x = TenureYear, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(y = ..count.. -200,
```

```
              label = paste0(round(prop.table(..count..),4) * 100, '%')),
          stat = 'count',
          position = position_stack(.99),
          size = 5) +
  labs(title = "Proporção de Clientes Tempo de Fidelização X Churn",
       x = "TenureYear",
       y = "Proporção")
```




```
# Compartivo quantitativo de churners para cada ano.

# Plotagem de churners por Tenure Year.
ggplot(churn_clean, aes(TenureYear, fill = Churn)) +
  geom_bar(position = "stack", colour = 'white') +
  geom_text(aes(label = ..count..),
            stat = "count",
            position=position_stack(vjust=0.5),
            size = 5) +
  labs(title = "Quantitativo de Clientes Tempo de Fidelização X Churn",
       x = "TenureYear",
       y = "Proporção")
```



Com base nesses gráficos, pode-se observar que:

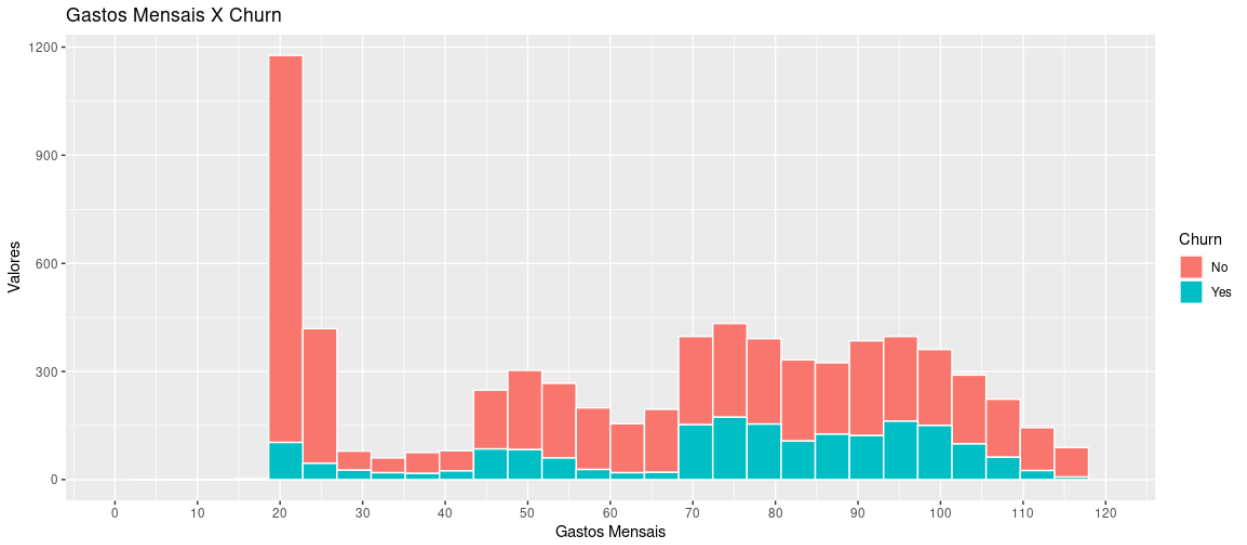
- a amostra de churners tem uma maior proporção para clientes com tempo de até 1 ano de relacionamento;
- há visivelmente um decréscimo de churners conforme aumenta o tempo de fidelização.

2.7 Dados Gastos Mensais

Os dados contínuos para os gastos de cada cliente é um gráfico com variabilidade de valores e, com isso, a plotagem de gráfico com valores fica muito prejudicada.

Visualizações mais simples (sem valores ou percentuais) podem fazer o papel da verificação dos dados, sem perder a objetividade do conteúdo.

```
# Compartivo quantitativo de churners pelos gastos mensais.

# Plotagem de churners por MonthlyCharges.
ggplot(data = churn_clean,
       aes(x = MonthlyCharges,
           fill = Churn)) +
  scale_x_continuous(
    breaks = seq(0, 120, 10),
    limits=c(0, 120)) +
  geom_histogram(colour = 'white') +
labs(title = "Gastos Mensais X Churn",
     x = "Gastos Mensais",
     y = "Valores")
```



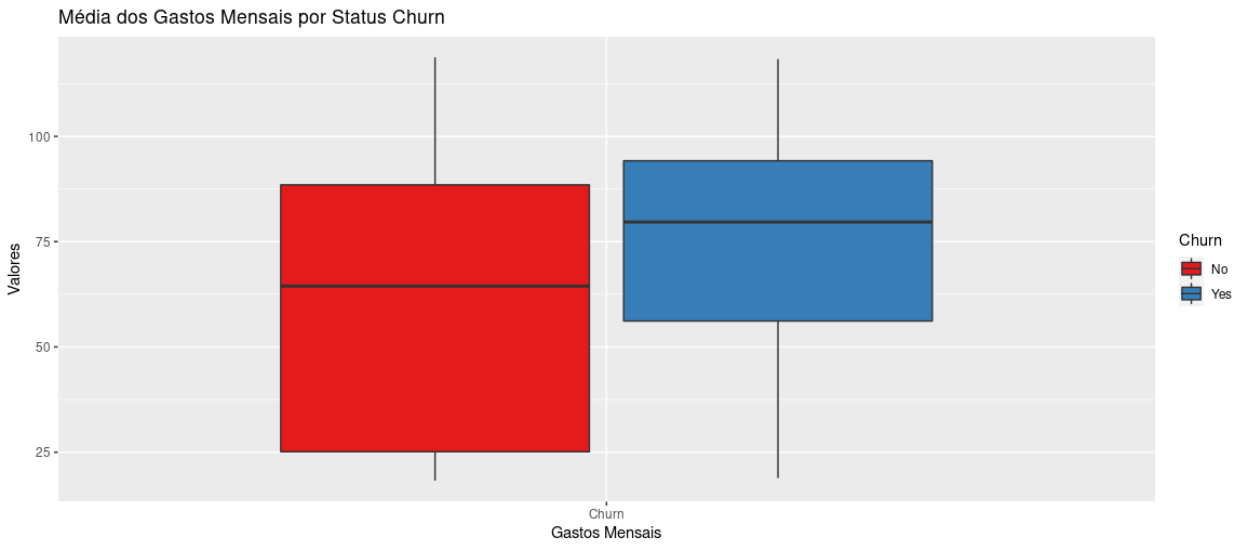

```
# Verificando informçãoes estatísticas sobre a variável.
profiling_num(churn_clean)
>
       variable     mean  std_dev variation_coef         range_98     skewness kurtosis    iqr
1 MonthlyCharges 64.79821 30.08597      0.4643026  [19.2, 114.7345]  -0.2220555 1.743883 54.275

       variable  p_01  p_05    p_25  p_50    p_75     p_95     p_99
1 MonthlyCharges  19.2 19.65 35.5875 70.35 89.8625 107.4225 114.7345
```

Com base nesses gráficos, pode-se observar que:

- a amostra de churners tem uma maior proporção para clientes com valores de gastos mensais acima da média de $64,80.

É possível ainda corroborar essa análise com uso das médias de gastos entre os churners e os não-churners com um gráfico do tipo [boxplot].

```
# Média de MonthlyCharges por tipo de Churn.
data_means <- aggregate(churn_clean$MonthlyCharges,
                        list(churn_clean$Churn),
                        mean)
data_means
>
data_means
  Group.1 mean.churn_clean$MonthlyCharges
1      No                        61.30741
2     Yes                        74.44133

# Boxplot para média de MonthlyCharges.
options(repr.plot.width = 6, repr.plot.height = 2)

ggplot(churn_clean, aes(x="Churn", y=MonthlyCharges)) +
  geom_boxplot(aes(fill=Churn)) +
  scale_fill_brewer(palette="Set1") +
  labs(title = "Média dos Gastos Mensais por Status Churn",
       x = "Gastos Mensais",
       y = "Valores")
```



O valor médio dos gastos mensais é maior entre os churners.

3. Conclusões Preliminares

3.1 Conclusões com base nas Análises Exploratória

Com as análises desses gráficos, é possível avaliar em quais situações há churners (Churn = Yes), para cada variável categórica do conjunto de dados. Bem como demonstra aspectos em que o cliente é mais propenso ao churn:

Com base nas avalições iniciais há mais churners para clientes:

- sem dependentes;
- sem parceiros;
- com menos de 65 anos;
- com serviço de telefonia habilitado - maior em apenas uma linha;
- com serviço de internet habilitado - maior em Fibra Óptica;
- com serviços online adicionais não habilitados;
- com contratos pré-pago (mês a mês);
- com fatura online sem papel;
- com pagamento de fatura por meio eletrônico;
- com média de gastos mensais superior à $65,00 .

Dentre as informações avaliadas, importante destacar que o gênero não foi identificado como um fator para churn.