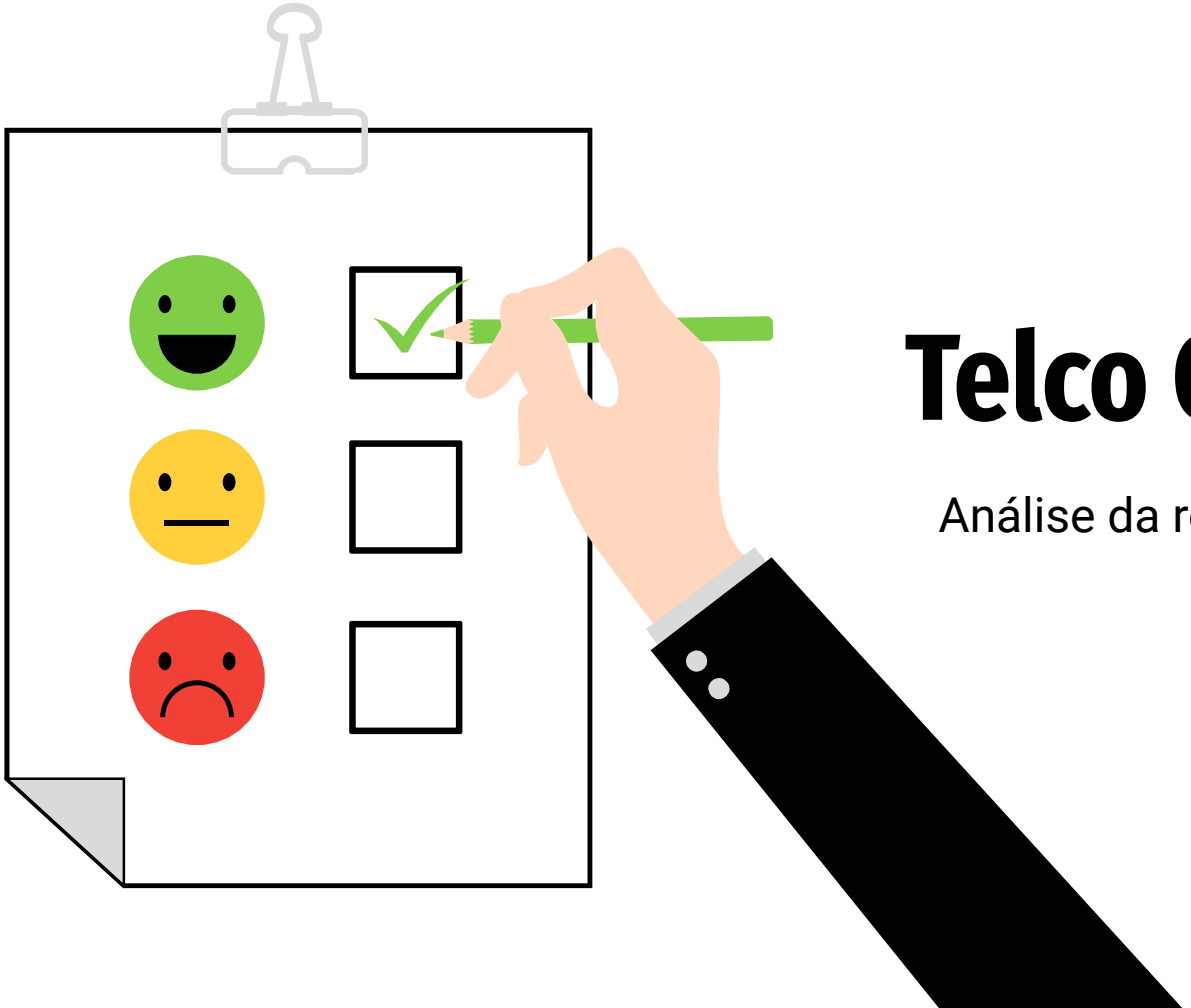


Telco Customers

Análise da rotatividade de clientes

Aluno: Vinicius Mattoso

Customer Churn ?

Customer churn é uma métrica que informa quanto uma empresa perdeu de clientes.

Entender os fatores que levam um ou mais clientes é de extrema importância para a **saúde e planejamento das empresas**

Colunas presentes no dataset

Feature name	Description
customerID	Client ID
gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer churned or not (Yes or No)

Ao lado estão exibidas as colunas disponíveis no dataset.

Como pode-se observar na descrição, muitas das variáveis disponíveis são do tipo categórica.

Tipo das colunas presentes no dataset

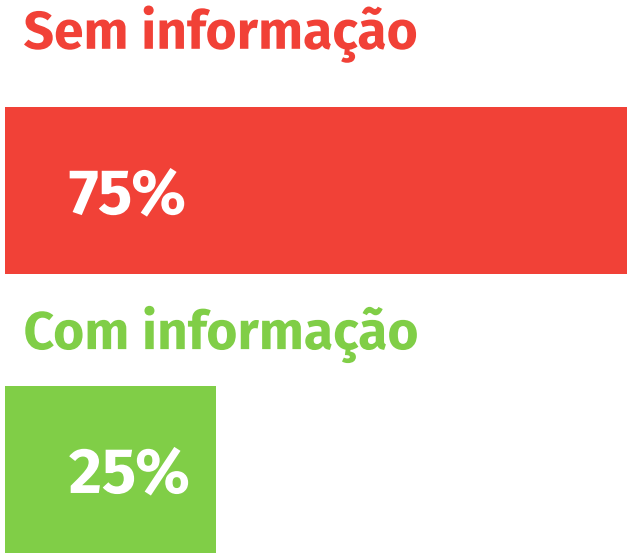
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Como já era de se imaginar, a maioria dos dados presentes nesse dataset são do tipo "**object**", esse tipo de dado requer um cuidado especial para conseguir extrair informações quantitativas do mesmo.

Precisamos converter esse dado para float, já que é um número.

Primeira análise estatística da base de dados

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7032.000000
mean	0.162147	32.371149	64.761692	2283.300441
std	0.368612	24.559481	30.090047	2266.771362
min	0.000000	0.000000	18.250000	18.800000
25%	0.000000	9.000000	35.500000	401.450000
50%	0.000000	29.000000	70.350000	1397.475000
75%	0.000000	55.000000	89.850000	3794.737500
max	1.000000	72.000000	118.750000	8684.800000

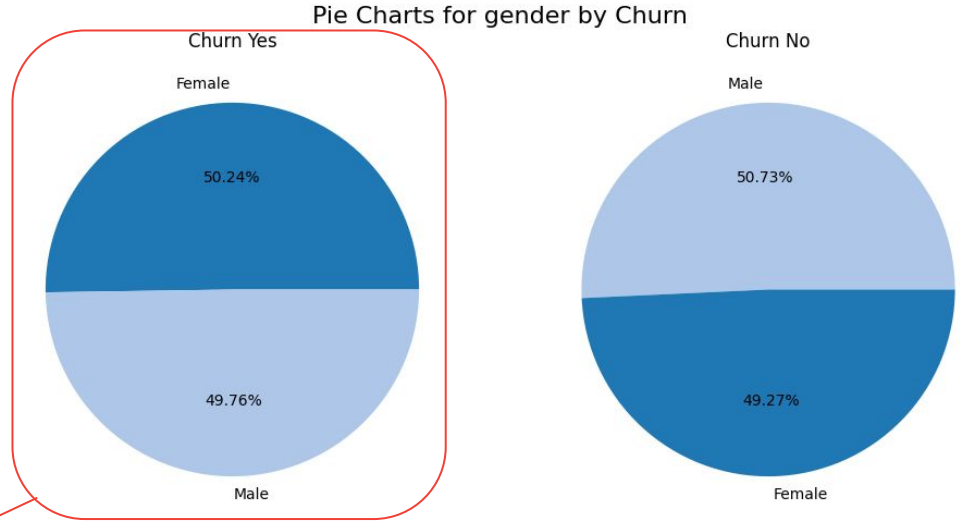


Sem o devido tratamento nos dados, temos uma perda de 75% das informações contidas no dataset. Uma vez que das 20 colunas iniciais só obtivemos informação numérica de 4.

Análise quantitativa da base de dados

	count	unique	top	freq
customerID	7043	7043	7590-VHVEG	1
gender	7043	2	Male	3555
Partner	7043	2	No	3641
Dependents	7043	2	No	4933
PhoneService	7043	2	Yes	6361
MultipleLines	7043	3	No	3390
InternetService	7043	3	Fiber optic	3096
OnlineSecurity	7043	3	No	3498
OnlineBackup	7043	3	No	3088
DeviceProtection	7043	3	No	3095
TechSupport	7043	3	No	3473
StreamingTV	7043	3	No	2810
StreamingMovies	7043	3	No	2785
Contract	7043	3	Month-to-month	3875
PaperlessBilling	7043	2	Yes	4171
PaymentMethod	7043	4	Electronic check	2365
Churn	7043	2	No	5174

Como a base de dados apresenta uma predominância inicial de dados categóricos, se faz necessário uma análise quantitativa da mesma.



Há praticamente um equilíbrio entre a quantidade de homens e mulheres que deixaram a empresa

Análise da correlação entre os dados numéricos pré-existentes

O total dos custos (**TotalCharges**) do cliente possui forte correlação com o tempo que ele ficou com a empresa (**tenure**)

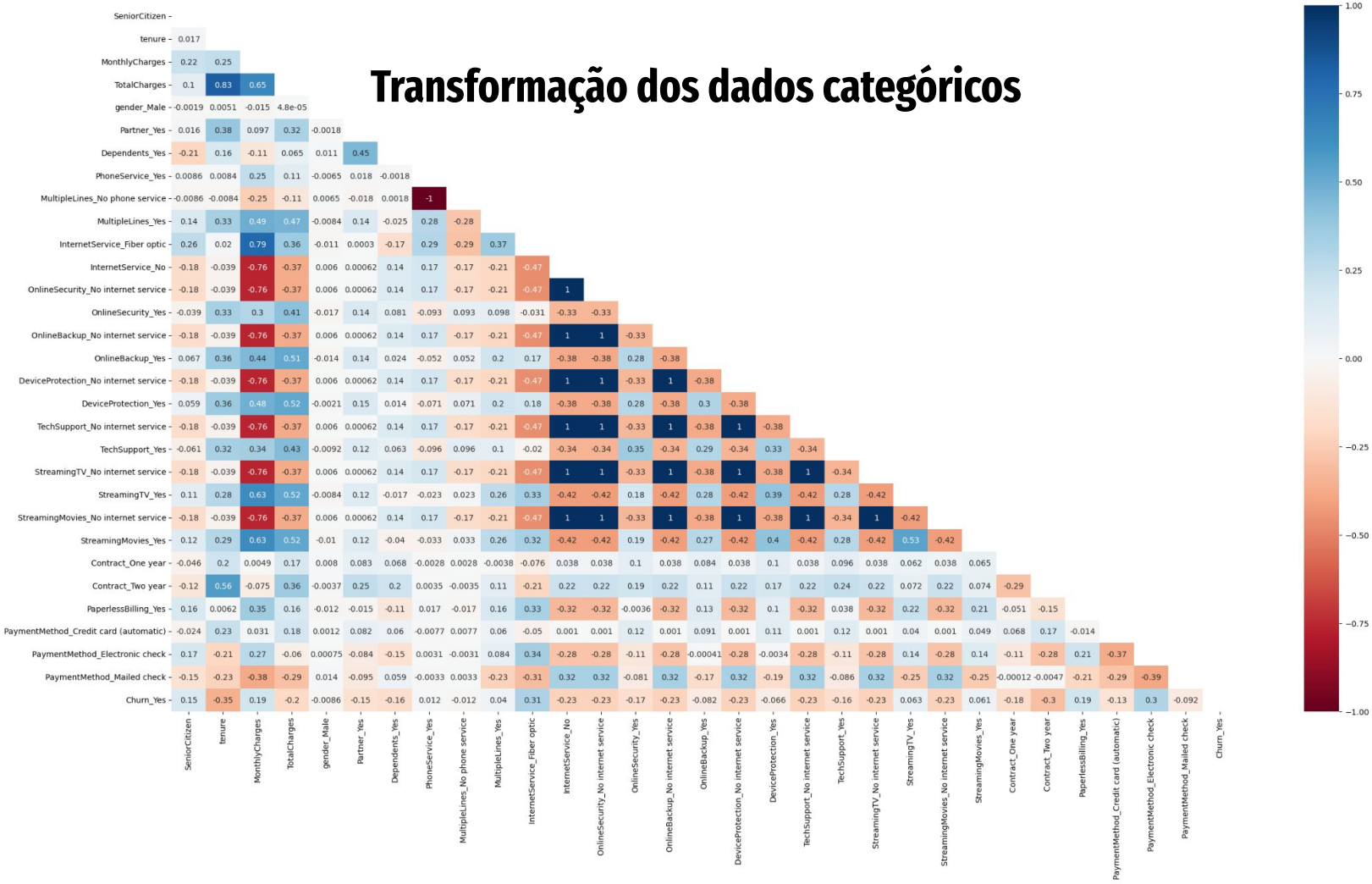
O total dos custos (**TotalCharges**) do cliente também possui forte correlação com a cobrança mensal (**MonthlyCharges**)

Transformação dos dados categóricos

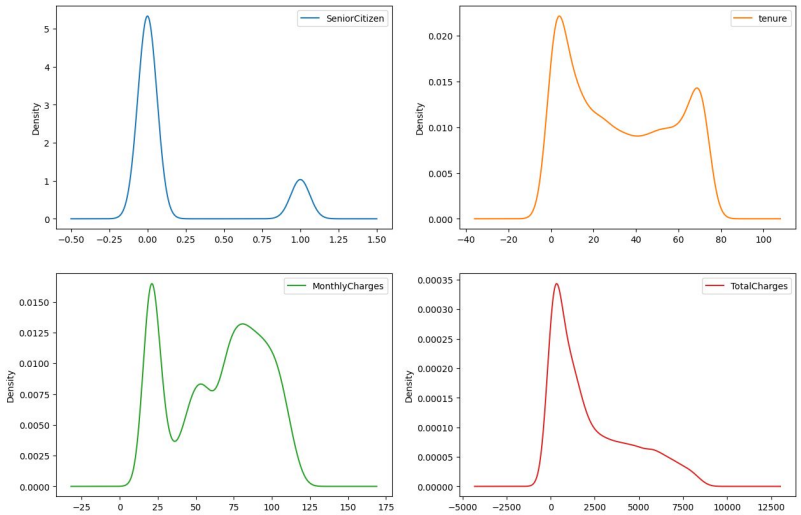
Podemos transformar os dados categóricos em dados numéricos de duas maneiras: **Dummy Encoding** e **OneHotEncoder**

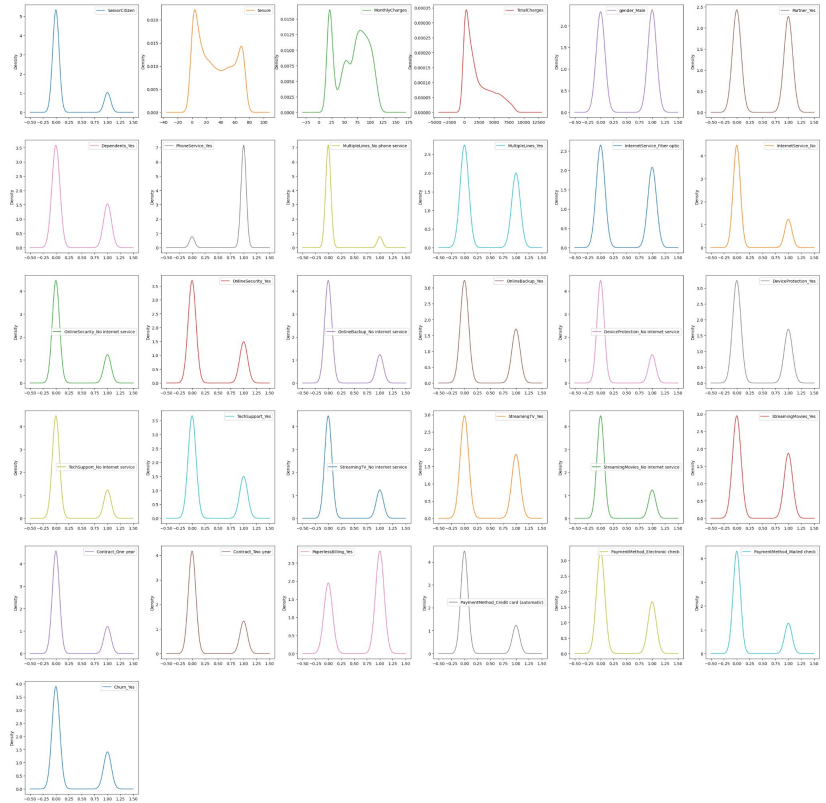
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 31 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   SeniorCitizen                          7043 non-null   int64
 1   tenure                                 7043 non-null   int64
 2   MonthlyCharges                         7043 non-null   float64
 3   TotalCharges                           7032 non-null   float64
 4   gender_Male                            7043 non-null   float64
 5   Partner_Yes                            7043 non-null   float64
 6   Dependents_Yes                         7043 non-null   float64
 7   PhoneService_Yes                       7043 non-null   float64
 8   MultipleLines_No phone service         7043 non-null   float64
 9   MultipleLines_Yes                      7043 non-null   float64
 10  InternetService_Fiber optic            7043 non-null   float64
 11  InternetService_No                     7043 non-null   float64
 12  OnlineSecurity_No internet service     7043 non-null   float64
 13  OnlineSecurity_Yes                     7043 non-null   float64
 14  OnlineBackup_No internet service       7043 non-null   float64
 15  OnlineBackup_Yes                       7043 non-null   float64
 16  DeviceProtection_No internet service   7043 non-null   float64
 17  DeviceProtection_Yes                   7043 non-null   float64
 18  TechSupport_No internet service        7043 non-null   float64
 19  TechSupport_Yes                        7043 non-null   float64
 20  StreamingTV_No internet service        7043 non-null   float64
 21  StreamingTV_Yes                        7043 non-null   float64
 22  StreamingMovies_No internet service    7043 non-null   float64
 23  StreamingMovies_Yes                    7043 non-null   float64
 24  Contract_One year                      7043 non-null   float64
 25  Contract_Two year                      7043 non-null   float64
 26  PaperlessBilling_Yes                   7043 non-null   float64
 27  PaymentMethod_Credit card (automatic)  7043 non-null   float64
 28  PaymentMethod_Electronic check         7043 non-null   float64
 29  PaymentMethod_Mailed check             7043 non-null   float64
 30  Churn_Yes                              7043 non-null   float64
dtypes: float64(29), int64(2)
memory usage: 1.7 MB
```

Transformação dos dados categóricos



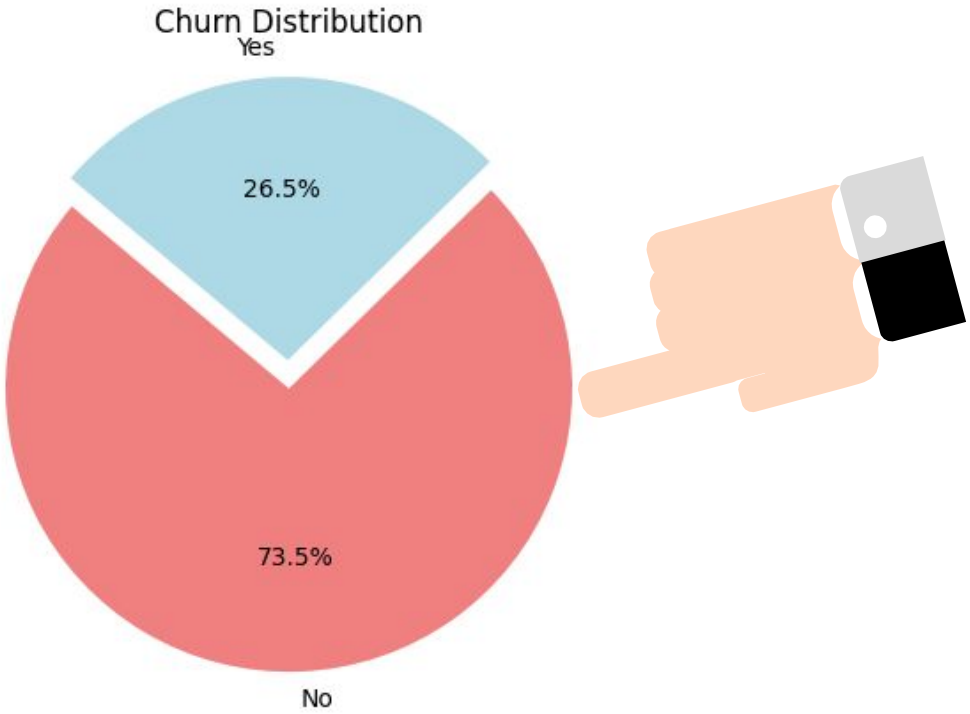
Resultado da transformação





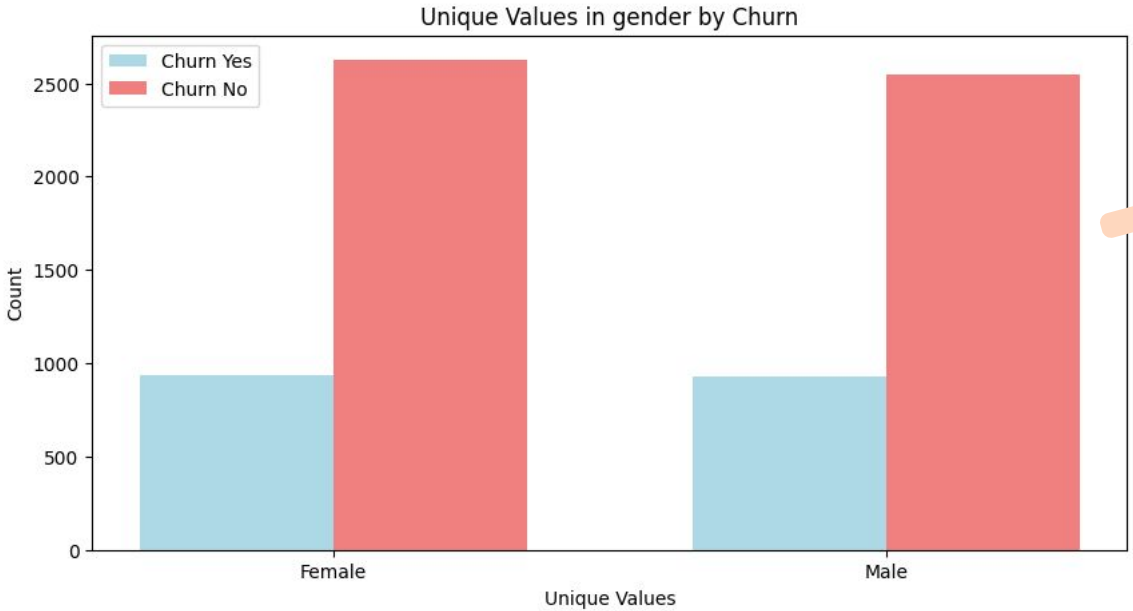
Ampliamos nosso poder de análise gráfica na base de dados.

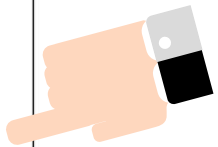
Análise da base de dados



Uma base de dados extremamente desbalanceada, o que pode elevar a dificuldade da criação de modelos para fazer essa classificação.

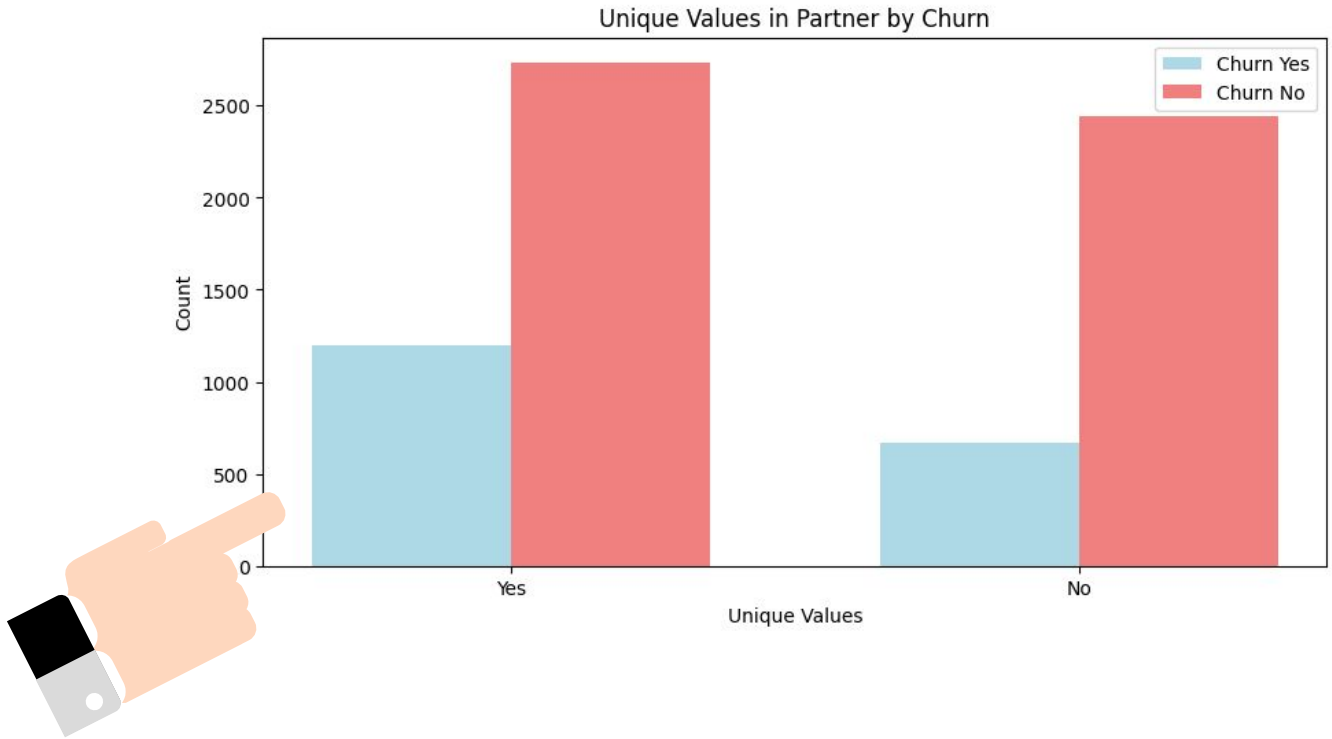
Análise da base de dados





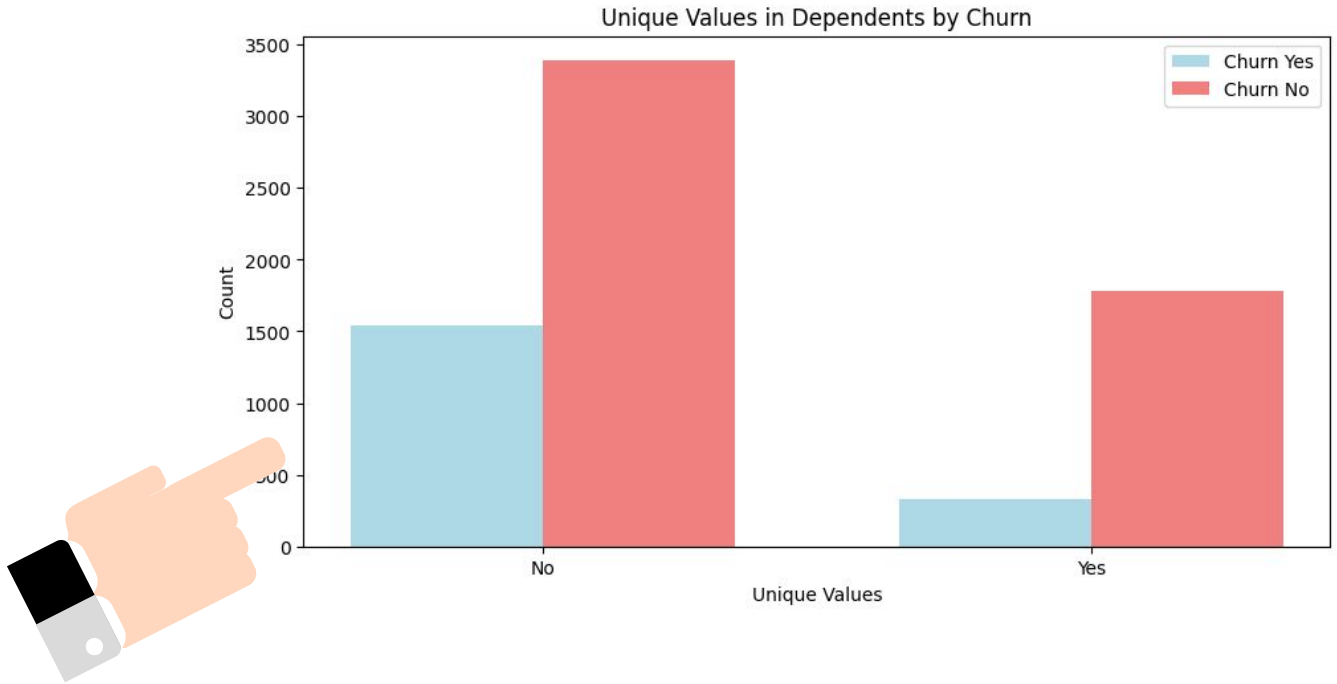
Com a relação de distribuição de gênero, essa base possui praticamente uma divisão de 50%.

Análise da base de dados



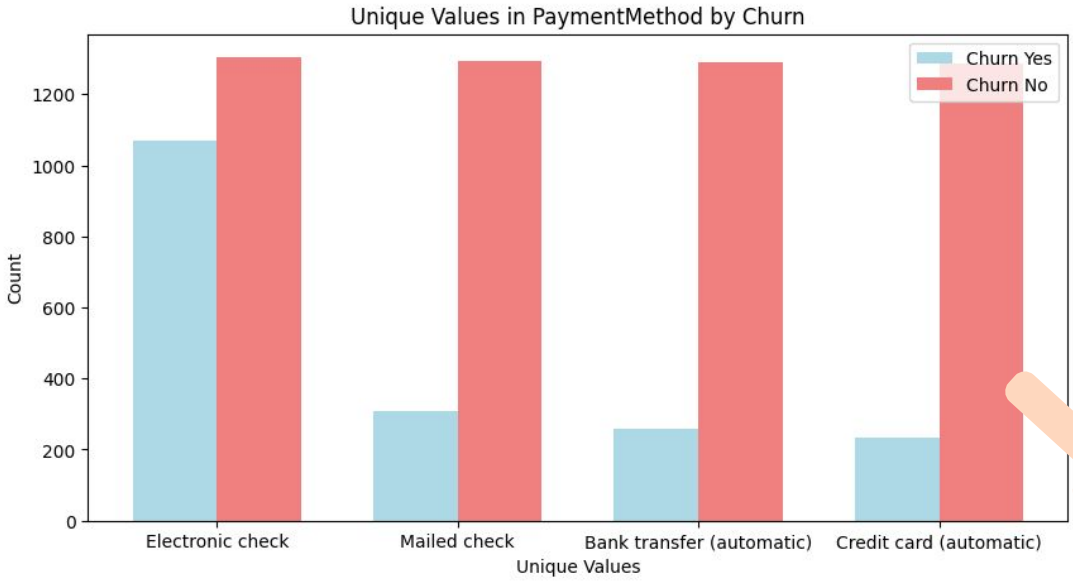
Na base de dados há uma predominância de solteiros.

Análise da base de dados



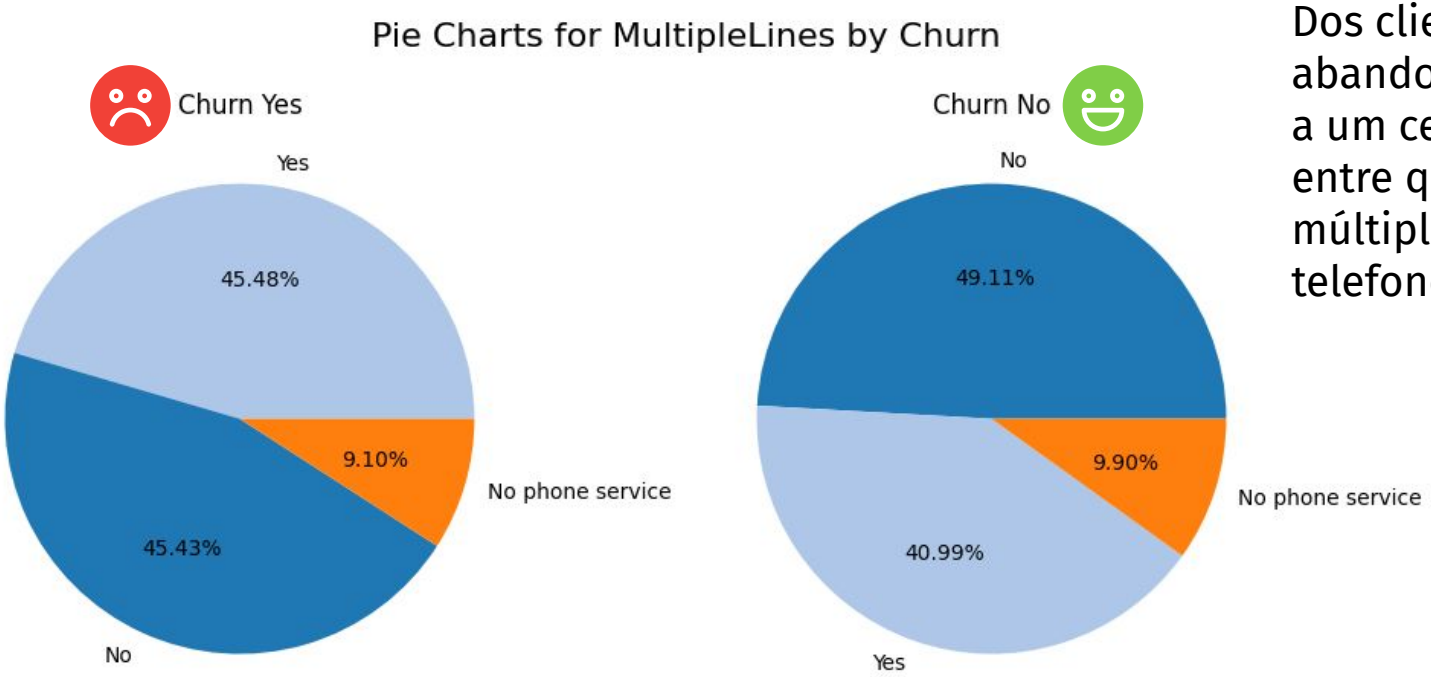
Dos clientes que abandonaram o serviço, a maioria não tinha dependentes.

Análise da base de dados



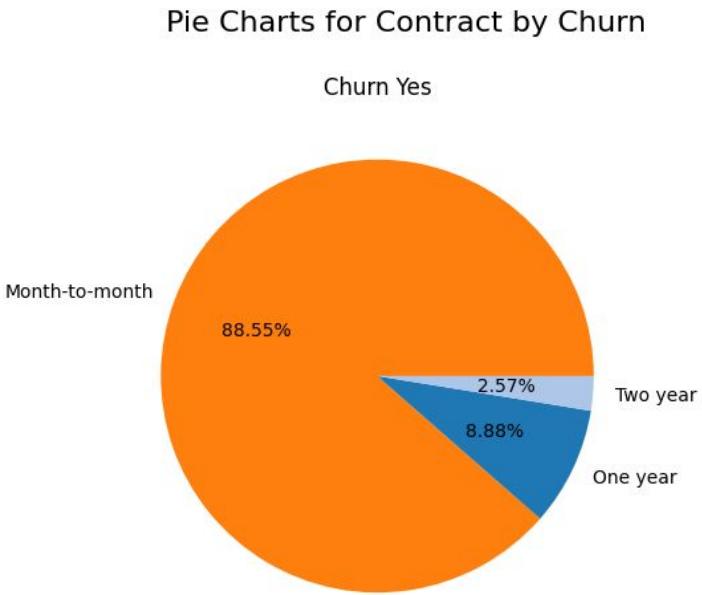
Dos clientes que abandonaram o serviço, a maioria utilizava cheque eletrônico com método de pagamento

Análise da base de dados



Dos clientes que abandonaram o serviço, a um certo equilíbrio entre quem tinha múltiplas linhas de telefone.

Análise da base de dados



Dos clientes que abandonaram o serviço, **88.55%** tinham contratos do tipo **mensal**.

Uma análise mais detalhada pode ser encontrada no notebook

https://github.com/vinicius-mattoso/MVP_Sprint_Analise_de_Dados/blob/main/MVP_Sprint_Analise_de_Dados_Notebook_Churn.ipynb