

▼ Desafio:

Você trabalha em uma grande empresa de Cartão de Crédito e o diretor da empresa percebeu que o número de clientes que cancelam seus cartões tem aumentado significativamente, causando prejuízos enormes para a empresa

O que fazer para evitar isso? Como saber as pessoas que têm maior tendência a cancelar o cartão?

O que temos:

Temos 1 base de dados com informações dos clientes, tanto clientes atuais quanto clientes que cancelaram o cartão

```
from google.colab import drive
drive.mount('/content/drive')
```

- Passo 1: Importar a base de dados
- Passo 2: Visualizar e tratar essa base de dados
- Passo 3: "Dar uma olhada" na sua base de dados
- Passo 4: Construir uma análise para identificar o motivo de cancelamento
  - Identificar qual o motivo ou os principais motivos dos clientes estarem cancelando o cartão de crédito

```
import pandas as pd

tabela = pd.read_csv("/content/drive/MyDrive/Minicurso Analise de Dados/ClientesBanco.csv", encoding="latin1")
tabela = tabela.drop("CLIENTNUM", axis=1)
display(tabela)
```

	Categoria	Idade	Sexo	Dependentes	Educação	Estado Civil	Faixa Salarial Anual	Categoria Cartão
0	Cliente	45	M	3	Ensino Médio	Casado	\$60K - \$80K	Blue
1	Cliente	49	F	5	Ensino Superior	Solteiro	Less than \$40K	Blue
2	Cliente	51	M	3	Ensino Superior	Casado	\$80K - \$120K	Blue
3	Cliente	40	F	4	Ensino Médio	Não informado	Less than \$40K	Blue
4	Cliente	40	M	3	Sem ensino formal	Casado	\$60K - \$80K	Blue
...	...	...	...	...	...	...	...	...
10122	Cliente	50	M	2	Ensino Superior	Solteiro	\$40K - \$60K	Blue
10123	Cancelado	41	M	2	Não informado	Divorciado	\$40K - \$60K	Blue
10124	Cancelado	44	F	1	Ensino Médio	Casado	Less than \$40K	Blue
10125	Cancelado	30	M	2	Ensino Superior	Não informado	\$40K - \$60K	Blue
10126	Cancelado	43	F	2	Ensino Superior	Casado	Less than \$40K	Silver

10127 rows × 20 columns

▼ Agora vamos tratar valores vazios e exibir um resumo das colunas da base de dados

```
tabela = tabela.dropna()
display(tabela.info())

display(tabela.describe().round(1))
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10126 entries, 0 to 10126
Data columns (total 20 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Categoria                                 10126 non-null  object
1   Idade                                     10126 non-null  int64
2   Sexo                                     10126 non-null  object
3   Dependentes                             10126 non-null  int64
4   Educação                                 10126 non-null  object
5   Estado Civil                             10126 non-null  object
6   Faixa Salarial Anual                     10126 non-null  object
7   Categoria Cartão                         10126 non-null  object
8   Meses como Cliente                      10126 non-null  int64
9   Produtos Contratados                    10126 non-null  int64
10  Inatividade 12m                          10126 non-null  int64
11  Contatos 12m                             10126 non-null  int64
12  Limite                                    10126 non-null  float64
13  Limite Consumido                         10126 non-null  int64
14  Limite Disponível                       10126 non-null  float64
15  Mudanças Transacoes_Q4_Q1               10126 non-null  float64
16  Valor Transacoes 12m                    10126 non-null  int64
17  Qtde Transacoes 12m                     10126 non-null  int64
18  Mudança Qtde Transações_Q4_Q1           10126 non-null  float64
19  Taxa de Utilização Cartão               10126 non-null  float64
dtypes: float64(5), int64(9), object(6)
memory usage: 1.6+ MB
None
```

	Idade	Dependentes	Meses como Cliente	Produtos Contratados	Inatividade 12m	Contatos 12m	Limite	Limite Consumido	Limite Disponível	Mudanças Transacoes_Q4_Q1	Valor Transacoes 12m
count	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0	10126.0
mean	46.3	2.3	35.9	3.8	2.3	2.5	8632.0	1162.8	7469.2	108.7	4404.7
std	8.0	1.3	8.0	1.6	1.0	1.1	9089.2	815.0	9091.1	353.9	3397.2
min	26.0	0.0	13.0	1.0	0.0	0.0	1438.3	0.0	3.0	0.0	510.0
25%	41.0	1.0	31.0	3.0	2.0	2.0	2555.0	358.0	1324.2	0.6	2156.0
50%	46.0	2.0	36.0	4.0	2.0	2.0	4549.0	1276.0	3473.0	0.7	3899.0
75%	52.0	3.0	40.0	5.0	3.0	3.0	11067.8	1784.0	9860.0	0.9	4741.0

▼ Vamos avaliar como está a divisão entre Clientes x Cancelados

```
qtde_categoria = tabela["Categoria"].value_counts()
display(qtde_categoria)

qtde_categoria_perc = tabela["Categoria"].value_counts(normalize=True)
display(qtde_categoria_perc)

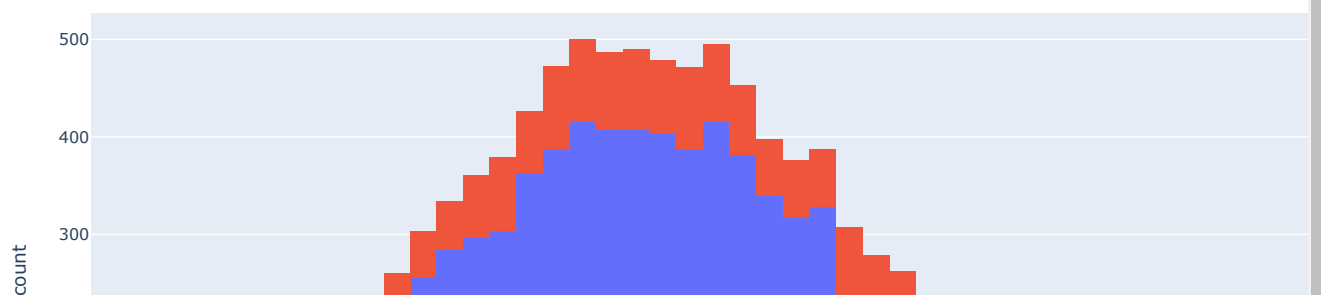
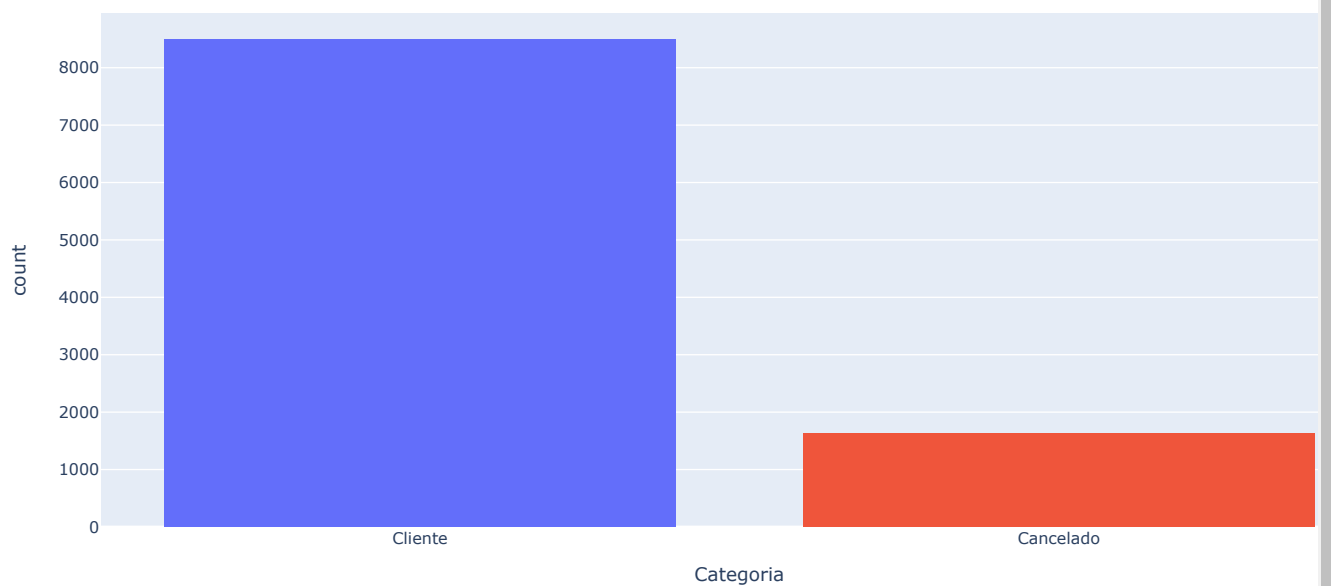
Cliente      8499
Cancelado    1627
Name: Categoria, dtype: int64
Cliente      0.839325
Cancelado    0.160675
Name: Categoria, dtype: float64
```

▼ Temos vários formas de descobrir o motivo de cancelamento

- Podemos olhar a comparação entre Clientes e Cancelados em cada uma das colunas da nossa base de dados, para ver se essa informação traz algum insight novo para a gente

```
import plotly.express as px

for coluna in tabela:
    grafico = px.histogram(tabela, x=coluna, color="Categoria")
    grafico.show()
```



## Informações retiradas da análise

- Me parece que quanto mais produtos contratados um cliente tem, menor a chance dele cancelar
- E quanto mais transações e quanto maior o valor de transação, menor a chance dele cancelar
- Quanto maior a quantidade de contatos que a pessoa teve que fazer, maior a chance dela cancelar

