JORNADA PYTHON

AULA 02

ANALISANDO DADOS COM PYTHON

Parte 1

Introdução

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

O que Vamos Aprender?

Na aula **Analisando Dados com Python** eu vou abordar um **Projeto de Análise de Cancelamento de Clientes**. Isso quer dizer que vamos fazer uma análise de uma base de dados para verificar essas informações dos clientes e verificar o percentual de cancelamento desses clientes.

Como é um projeto de Análise, o objetivo vai ser fazer análises para ver onde temos os maiores cancelamentos e os motivos e com isso vamos poder propor uma solução para diminuir essa quantidade de cancelamentos.

Isso é basicamente o que uma empresa vai fazer em um projeto, por isso é muito importante a **Análise de Dados**, pois com isso você vai conseguir informações importantes e vai verificar possíveis soluções para os problemas que tem.

Você vai notar que esse tipo de análise de dá um entendimento muito maior dos dados que tem, então “achar” que algo acontece por um motivo não é a melhor solução.

É importante fazer de fato uma análise de dados para que você possa tirar suas próprias conclusões utilizando os números. Principalmente dentro de uma empresa é muito importante essa análise de dados.

Nesse exemplo você vai ver que começamos com um percentual x de cancelamentos, mas ao longo da análise e tratamento dos dados, vai notar que vamos conseguir reduzir drasticamente esse número só com essas análises.

Então é uma solução que a empresa pode seguir para diminuir a quantidade de cancelamentos que está tendo. Lembre-se de que você já tem a base de dados, basta fazer uma análise nela para entender como tudo está funcionando!

Vamos passar por várias etapas utilizando o **Python** para chegar ao nosso resultado, mas vamos te ensinar o que foi feito em cada uma das etapas.

O que Vamos Precisar?

Antes de começar com as explicações, é importante que você saiba o que vai precisar para poder seguir com a aula sem que você fique perdido. Então vou te mostrar o que precisa e já vou deixar os links caso você não tenha ou tenha alguma dúvida sobre como baixar, configurar ou instalar o que for necessário para a aula.

- Editor de Python (**VSCode**) e do **Python** (link do vídeo de instalação do editor e do Python);

- **Jupyter** dentro do VSCode - Aqui vamos instalar o Jupyter dentro do editor de Python para que fique mais fácil visualizar cada etapa do nosso projeto. Dessa forma você vai conseguir ver o passo a passo de forma mais didática e vai entender como funciona cada uma das etapas (link do vídeo instalação);

- Arquivos da Aula – Essa parte é importante, pois vamos disponibilizar a base de dados que será analisada durante a aula, assim como o gabarito com o código completo e comentado (link para download dos arquivos).

Tendo o editor de Python com o Jupyter, o Python e a base de dados você já está apto e iniciar a aula e começar o nosso projeto!

OBS: Além do que foi informado acima, seguindo os passos dos vídeos para instalação, você vai precisar fazer a instalação da biblioteca pandas, então basta abrir o terminal e escrever **pip install pandas**
É importante que faça a instalação da biblioteca plotly (**pip install plotly**) para que possamos criar alguns gráficos para complementar as análises que vamos fazer.

IMPORTANTE: Caso você encontre um erro sobre nbformat quando for criar seus gráficos, pode fazer a instalação desse recurso no terminal do programa (**pip install nbformat**). Dessa forma vai conseguir criar seus gráficos normalmente.

Entendendo a Base de Dados

Antes de iniciar de fato o projeto, é interessante que você entenda a **base de dados** que vai utilizar e o que cada informação significa para ter um melhor entendimento.

Aqui eu vou te mostrar essa base de dados já dentro do nosso editor de Python, pois ela está em extensão **.csv**, o que dificulta um pouco a visualização dos dados.

Essa é uma base de dados de cancelamento de clientes em uma empresa fictícia.

Mas lembre-se de que isso é bem semelhante ao que vai encontrar no seu dia a dia de trabalho, então o que vamos fazer aqui é algo que você vai poder replicar no seu trabalho!

Então é possível que se depare com bases de dados bem similares a essa.

	idade	sexo	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	assinatura	duracao_contrato	total_gasto	meses_ultima_interacao	cancelou
0	30.0	Female	39.0	14.0	5.0	18.0	Standard	Annual	932.00	17.0	1.0
1	65.0	Female	49.0	1.0	10.0	8.0	Basic	Monthly	557.00	6.0	1.0
2	55.0	Female	14.0	4.0	6.0	18.0	Basic	Quarterly	185.00	3.0	1.0
3	58.0	Male	38.0	21.0	7.0	7.0	Standard	Monthly	396.00	29.0	1.0
4	23.0	Male	32.0	20.0	5.0	8.0	Basic	Monthly	617.00	20.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
881661	42.0	Male	54.0	15.0	1.0	3.0	Premium	Annual	716.38	8.0	0.0
881662	25.0	Female	8.0	13.0	1.0	20.0	Premium	Annual	745.38	2.0	0.0
881663	26.0	Male	35.0	27.0	1.0	5.0	Standard	Quarterly	977.31	9.0	0.0
881664	28.0	Male	55.0	14.0	2.0	0.0	Standard	Quarterly	602.55	2.0	0.0
881665	31.0	Male	48.0	20.0	1.0	14.0	Premium	Quarterly	567.77	21.0	0.0

881666 rows × 11 columns

INFORMAÇÕES DA BASE DE DADOS - CLIENTES

Idade
Sexo – Masculino e Feminino
Tempo_como_cliente
Frequencia_uso – A frequência de uso desse cliente
Ligacoes_callcenter – Quantas vezes o cliente ligou ao Call Center da empresa
Dias_atraso – Dias em que o cliente esteve em atraso
Assinatura – O plano que o cliente possui
Duracao_contrato – Tempo de duração do contrato (Mensal, Trimestral e Anual)
Total_gasto
Meses_ultima_interação – Meses desde a última interação do cliente
Cancelou – Se o cliente cancelou (1) ou não (0) sua assinatura

Entendendo a solução final

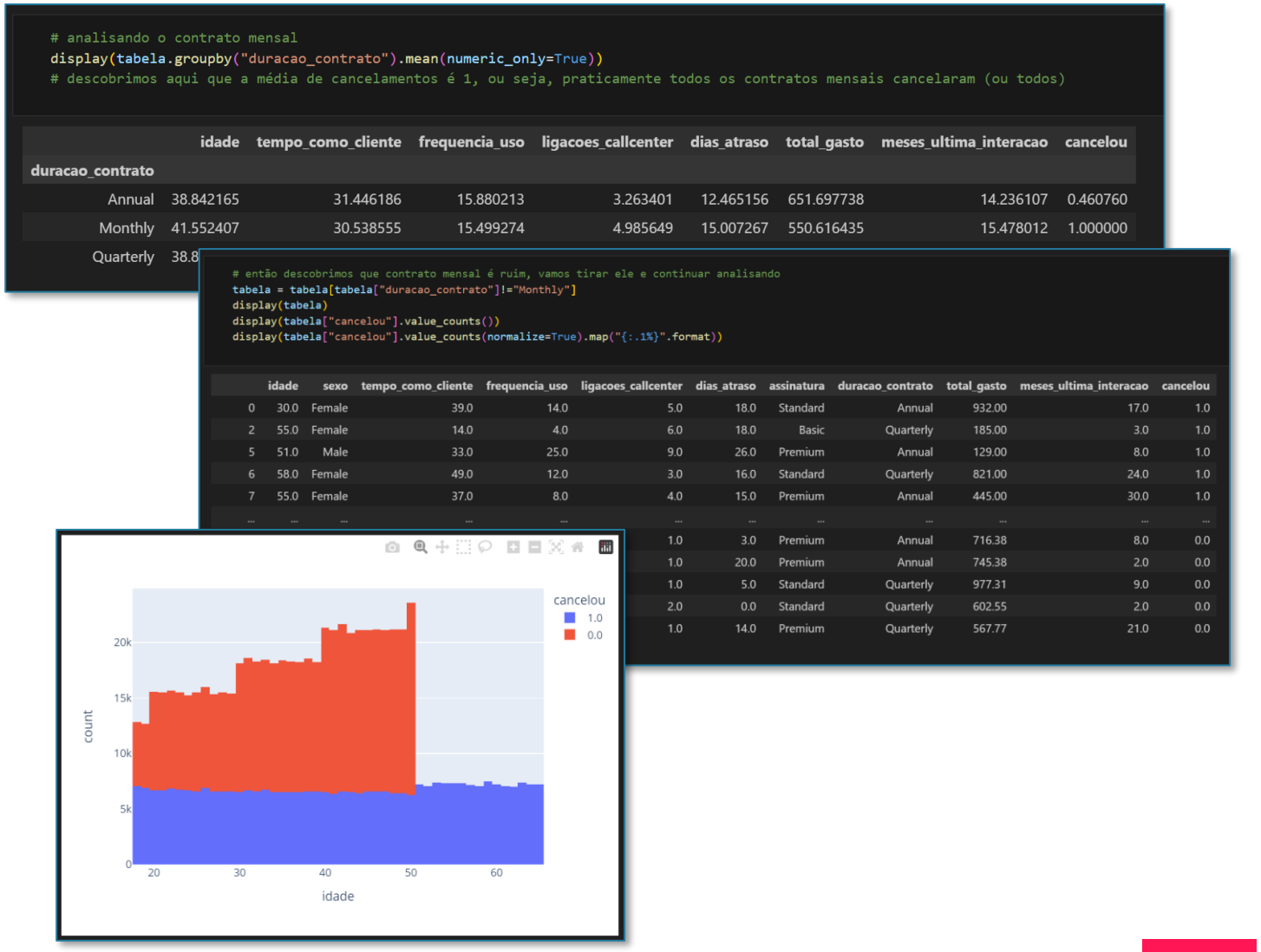
O nosso exemplo é uma base de dados de uma empresa onde vamos analisar o cancelamento dos clientes.

O objetivo é tratar essa base de dados, para que possamos analisar de forma eficiente e fazer algumas análises úteis para a empresa.

Sendo uma delas, verificar o motivo dos cancelamentos e tentar diminuir esse parâmetro.

Você vai notar que um dos principais motivos de cancelamento são as pessoas que possuem contratos mensais.

Só que essa não é a única análise que pode ser feita na base de dados. Como temos muitas informações, podemos fazer diversas outras análises também!



Parte 2

O que é o Python

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

O que é o Python?

O Python, é uma linguagem de programação.

Ok.... Mas o que é uma linguagem de programação??

Assim como temos diferentes línguas para falarmos, existem diversas línguas que nos permitem "falar" com os computadores.

Entre as línguas de produção, o Python é uma das mais fáceis de aprender e uma das que mais cresce no mundo em termos de utilização.

Pode ser utilizado em diversas áreas:

- Data Science;
- Automação de processos;
- Desenvolvimento de sites;
- Inteligência artificial;
- Vários outros

Curiosidade: Seu nome apesar de geralmente ser vinculado a cobra, não tem essa origem.... Na verdade, ele é uma homenagem a um grupo de comédia inglês chamado Monty Python.

Jupyter Notebook no Visual Studio Code

Como nós vamos utilizar o **Jupyter Notebook** dentro do **Visual Studio Code** você vai notar que ele tem uma estrutura um pouco diferente.

Vamos ter uma estrutura em blocos que facilita a visualização e entendimento do que estamos fazendo, pois dessa forma conseguimos fazer por partes e já te mostrar os resultados.

Só que você precisa tomar um certo cuidado quanto a isso, pois quando roda um bloco de código apenas o editor de código não vai rodar todos (você tem essa opção também).

Isso é **MUITO IMPORTANTE**, pois se você faz uma edição no bloco 1 e não roda ele novamente, o restante do seu código ainda vai estar utilizando a informação antiga.

Então para evitar problemas, principalmente nos gráficos que fica mais visual, sempre rode todos os blocos para ter todas as informações atualizadas!

Parte 3

Importando a Base de Dados

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

Importando a base de dados

Tendo tudo instalado e com a base de dados na pasta onde criou o seu arquivo (**main.ipynb**). É importante que o seu arquivo tenha essa extensão para que você consiga utilizar o Jupyter dentro do Visual Studio Code (VSCode).

Nós já podemos começar a importação da nossa base de dados.

```
import pandas as pd

tabela = pd.read_csv("cancelamentos.csv")
tabela = tabela.drop("CustomerID", axis=1)
display(tabela)
```

✓ 2.0s

Vamos iniciar com a importação da biblioteca pandas (1ª linha de código). Essa é uma biblioteca muito utilizada em análise de dados, então é uma biblioteca muito importante para essa área.

Após importar a biblioteca pandas nós vamos utilizar o **pd.read_csv** para ler o nosso arquivo que está no formato csv.

A ideia é atribuir essa tabela a uma variável, que nesse caso vai se chamar **tabela**. É interessante você colocar nomes intuitivos para saber do que se tratam seus dados, então tabela é um bom nome para iniciar.

Logo abaixo nós vamos utilizar o comando **tabela.drop** para remover a coluna **CustomerID** da nossa base de dados, pois essa informação não é útil e não adiciona nada na nossa análise.

É apenas o número do cliente, por esse motivo podemos remover essas informações logo no início.

Para finalizar vamos utilizar o comando display para visualizar a nossa base de dados.

	idade	sexo	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	assinatura	duracao_contrato	total_gasto	meses_ultima_interacao	cancelou
0	30.0	Female	39.0	14.0	5.0	18.0	Standard	Annual	932.00	17.0	1.0
1	65.0	Female	49.0	1.0	10.0	8.0	Basic	Monthly	557.00	6.0	1.0
2	55.0	Female	14.0	4.0	6.0	18.0	Basic	Quarterly	185.00	3.0	1.0
3	58.0	Male	38.0	21.0	7.0	7.0	Standard	Monthly	396.00	29.0	1.0
4	23.0	Male	32.0	20.0	5.0	8.0	Basic	Monthly	617.00	20.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
881661	42.0	Male	54.0	15.0	1.0	3.0	Premium	Annual	716.38	8.0	0.0
881662	25.0	Female	8.0	13.0	1.0	20.0	Premium	Annual	745.38	2.0	0.0
881663	26.0	Male	35.0	27.0	1.0	5.0	Standard	Quarterly	977.31	9.0	0.0
881664	28.0	Male	55.0	14.0	2.0	0.0	Standard	Quarterly	602.55	2.0	0.0
881665	31.0	Male	48.0	20.0	1.0	14.0	Premium	Quarterly	567.77	21.0	0.0

881666 rows × 11 columns

Parte 4

Tratamento de Dados

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

Removendo Informações Vazias

Antes de começar com as análise é essencial que você faça o tratamento de dados, assim evita trazer erro por conta de dados desnecessários ou até inexistentes.

Você deve ter visto na visualização da base de dados que nós temos **881.666 linhas** de informação.

Só que como você deve saber, nem toda base de dados é completa, então nós vamos ter informações em branco, e essas informações podem atrapalhar nossas análises.

Por isso nós vamos utilizar o comando **tabela.dropna** para remover as informações vazias da nossa tabela.

```
# identificando e removendo valores vazios
display(tabela.info())
tabela = tabela.dropna()
display(tabela.info())
```

✓ 0.3s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 881666 entries, 0 to 881665
Data columns (total 11 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   idade                  881664 non-null  float64
 1   sexo                   881664 non-null  object
 2   tempo_como_cliente     881663 non-null  float64
 3   frequencia_uso         881663 non-null  float64
 4   ligacoes_callcenter    881664 non-null  float64
 5   dias_atraso            881664 non-null  float64
 6   assinatura             881661 non-null  object
 7   duracao_contrato       881663 non-null  object
 8   total_gasto            881664 non-null  float64
 9   meses_ultima_interacao 881664 non-null  float64
 10  cancelou               881664 non-null  float64
dtypes: float64(8), object(3)
memory usage: 74.0+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 881659 entries, 0 to 881665
Data columns (total 11 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   idade                  881659 non-null  float64
 1   sexo                   881659 non-null  object
 2   tempo_como_cliente     881659 non-null  float64
 3   frequencia_uso         881659 non-null  float64
 4   ligacoes_callcenter    881659 non-null  float64
 5   dias_atraso            881659 non-null  float64
 6   assinatura             881659 non-null  object
 7   duracao_contrato       881659 non-null  object
 8   total_gasto            881659 non-null  float64
 9   meses_ultima_interacao 881659 non-null  float64
 10  cancelou               881659 non-null  float64
dtypes: float64(8), object(3)
memory usage: 80.7+ MB
```

Na imagem a esquerda você consegue ver as informações dos dados em cada uma das colunas. Com isso, pode notar que não temos **881.666 informações não vazias** (ou seja, informações preenchidas).

Por esse motivo que vamos utilizar o **dropna**, para remover essas informações vazias e padronizar nossa base de dados para manter a mesma quantidade de informações em todas as colunas.

Parte 5

Análise de Dados

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

Verificando a Taxa de Cancelamento

Agora que a nossa base de dados já está tratada, nós podemos dar início a análise de dados propriamente dita.

Vamos começar analisando a taxa de cancelamento da empresa, pois o objetivo é diminuir essa taxa. Então vamos ter que verificar qual é essa taxa e descobrir de onde vem essa taxa.

```
# quantas pessoas cancelaram e não cancelaram
display(tabela["cancelou"].value_counts())
display(tabela["cancelou"].value_counts(normalize=True).map("{:.1%}".format))
```

✓ 0.0s

```
cancelou
1.0    499993
0.0    381666
Name: count, dtype: int64

cancelou
1.0    56.7%
0.0    43.3%
Name: proportion, dtype: object
```

Podemos utilizar o **tabela[“cancelou”].value_counts** para verificar a quantidade de informações que temos na coluna “cancelou” da tabela.

Dessa forma vamos poder visualizar com o **display** qual é a proporção de clientes que cancelaram e que continuam com a assinatura.

Você deve ter notado que temos duas visualizações, a primeira vai mostrar a contagem dos dados, quem cancelou (1) e quem não cancelou (0).

Já a segunda visualização vamos normalizar e formatar em percentual, assim fica mais fácil para analisar qual a proporção de clientes que estão cancelando o serviço.

Veja que temos um total de **56,7%** de cancelamento das assinaturas, então mais da metade dos clientes estão cancelando o serviço.

Concorda que nenhuma empresa quer algo similar, não é mesmo? Então vamos descobrir de onde vem esse número tão alto!

Verificando o Cancelamento por Contrato

IMPORTANTE: Na parte de análise de dados, não tem uma informação correta para ser analisada logo de cara. Esse é um processo que vai tomar tempo, pois você de fato precisa analisar os dados e entender o que está acontecendo na sua base de dados. Então pode ser que demore mais em alguns casos para encontrar o que procura antes de propor sua solução.

```
display(tabela["duracao_contrato"].value_counts(normalize=True))
display(tabela["duracao_contrato"].value_counts())
```

✓ 0.1s

```
duracao_contrato
Annual       0.401964
Quarterly    0.400448
Monthly      0.197588
Name: proportion, dtype: float64

duracao_contrato
Annual       354395
Quarterly    353059
Monthly      174205
Name: count, dtype: int64
```

Já vimos como está a relação do cancelamento dos clientes, agora podemos dar uma olhada como está a duração do contrato desses clientes.

Lembrando que temos **3 tipos de contrato**: mensal, trimestral e anual. Então é interessante ver como está essa proporção para verificar se isso pode ser um fator que afeta diretamente o cancelamento do serviço.

Com essa simples análise você já nota que temos a seguinte proporção na duração dos contratos:

- **Anual** 40,19%
- **Trimestral** 40,00%
- **Mensal** 19,75%

Veja que a temos uma divisão quase igual entre os planos anual e trimestral, mas o plano mensal já fica atrás com quase 20%.

O que podemos fazer é analisar as informações dos contratos para verificar como estão distribuídas e verificar se algum deles tem um percentual maior de cancelamento.

Analisando as Informações dos Conratos

```
# analisando o contrato mensal
display(tabela.groupby("duracao_contrato").mean(numeric_only=True))
# descobrimos aqui que a média de cancelamentos é 1, ou seja, praticamente todos os contratos mensais cancelaram (ou todos)
```

✓ 0.0s

duracao_contrato	idade	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	total_gasto	meses_ultima_interacao	cancelou
Annual	38.842165	31.446186	15.880213	3.263401	12.465156	651.697738	14.236107	0.460760
Monthly	41.552407	30.538555	15.499274	4.985649	15.007267	550.616435	15.478012	1.000000
Quarterly	38.830938	31.419916	15.886662	3.265245	12.460863	651.427783	14.234544	0.460255

Aqui nós vamos utilizar o **groupby** para agrupar as informações da coluna **duração_contrato** e depois fazer a média das informações que temos na tabela.

Isso vai nos dar uma informação mais geral de cada um desses planos, e podemos verificar se tem alguma informação importante.

Com as informações agrupadas, é possível notar que os clientes do plano **Mensal**, possuem uma **média de cancelamento igual a 1**, ou seja, praticamente todos os clientes que utilizam esse plano fizeram o cancelamento do serviço.

Esse já é um ponto importante dentro da nossa análise, pois existe um plano dessa empresa, onde praticamente todos os clientes fazem o cancelamento do serviço.

Removendo o Contrato Mensal

Sabendo que o contrato mensal é ruim para a empresa, nós podemos remover as informações desse contrato específico e continuar analisando.

Aqui vale lembrar que nem sempre que encontrar algo que seja ruim na sua análise de dados você retira e para por ali. A ideia é ir analisando até que chegue em um valor aceitável dentro do seu projeto.

Então é importante definir esse "valor aceitável" ou seu objetivo para não ficar trabalhando sem ter um ponto de parada.

```
# então descobrimos que contrato mensal é ruim, vamos tirar ele e continuar analisando
tabela = tabela[tabela["duracao_contrato"]!="Monthly"]
display(tabela)
display(tabela["cancelou"].value_counts())
display(tabela["cancelou"].value_counts(normalize=True).map("{:.1%}".format))
```

✓ 0.1s

	idade	sexo	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	assinatura	duracao_contrato	total_gasto	meses_ultima_interacao	cancelou
0	30.0	Female	39.0	14.0	5.0	18.0	Standard	Annual	932.00	17.0	1.0
2	55.0	Female	14.0	4.0	6.0	18.0	Basic	Quarterly	185.00	3.0	1.0
5	51.0	Male	33.0	25.0	9.0	26.0	Premium	Annual	129.00	8.0	1.0
6	58.0	Female	49.0	12.0	3.0	16.0	Standard	Quarterly	821.00	24.0	1.0
7	55.0	Female	37.0	8.0	4.0	15.0	Premium	Annual	445.00	30.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
881661	42.0	Male	54.0	15.0	1.0	3.0	Premium	Annual	716.38	8.0	0.0
881662	25.0	Female	8.0	13.0	1.0	20.0	Premium	Annual	745.38	2.0	0.0
881663	26.0	Male	35.0	27.0	1.0	5.0	Standard	Quarterly	977.31	9.0	0.0
881664	28.0	Male	55.0	14.0	2.0	0.0	Standard	Quarterly	602.55	2.0	0.0
881665	31.0	Male	48.0	20.0	1.0	14.0	Premium	Quarterly	567.77	21.0	0.0

707454 rows × 11 columns

Veja que a proporção de cancelamentos já caiu para **46,1%** nessa análise, mas esse número ainda é muito alto.

Então vamos continuar analisando para chegar em um valor aceitável de cancelamentos que não esteja perto dos 50%.

```
cancelou
0.0    381666
1.0    325788
Name: count, dtype: int64

cancelou
0.0    53.9%
1.0    46.1%
Name: proportion, dtype: object
```

Análise de Assinaturas

Como ainda temos um número bem alto de cancelamentos vamos agora fazer uma análise nas assinaturas para verificar se podemos tirar alguma conclusão para melhorar esse índice de cancelamentos.

Primeiro vamos fazer a **contagem dos valores na coluna de assinaturas** para saber quantas assinaturas temos em cada um dos planos.

Em seguida vamos **agrupar as informações por assinatura** e obter a **média** das linhas para cada uma das colunas.

```
# chegamos agora em menos da metade de pessoas cancelando, mas ainda temos muitas pessoas ai, vamos continuar analisando
display(tabela["assinatura"].value_counts(normalize=True))
display(tabela.groupby("assinatura").mean(numeric_only=True))
# vemos que assinatura é quase 1/3, 1/3, 1/3
# e que os cancelamentos são na média bem parecidos, então fica difícil tirar alguma conclusão da média, vamos precisar ir mais a fundo
```

✓ 0.1s

```
assinatura
Standard    0.339648
Premium     0.338138
Basic       0.322215
Name: proportion, dtype: float64
```

assinatura	idade	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	total_gasto	meses_ultima_interacao	cancelou
Basic	38.904813	32.316031	15.876921	3.310021	12.507054	648.642614	14.240814	0.475188
Premium	38.817814	30.977869	15.889673	3.235886	12.433427	653.337633	14.231150	0.452338
Standard	38.790478	31.048621	15.883393	3.249275	12.450690	652.566793	14.234280	0.454714

Na primeira análise podemos verificar que temos praticamente a mesma quantidade em cada uma das assinaturas, ou seja, temos praticamente **1/3 em cada assinatura**.

E na segunda análise temos que os valores de cancelamento também são muito parecidos.

O que fazer agora?

Não podemos excluir nenhuma informação, pois os dados são praticamente iguais. Isso quer dizer que vamos ter que ir mais fundo na nossa análise de dados.

Foi o que eu comentei anteriormente, nem sempre vamos achar logo de cara o que precisamos!

Análises Gráficas

Como a última análise não foi muito boa para poder verificar quais informações poderiam ser removidas, vamos criar alguns gráficos, pois dessa forma fica muito mais fácil visualizar os dados e obter as informações que de fato estão aumentando o número de cancelamentos nessa empresa.

```
# vamos criar gráfico, porque só com números tá difícil de visualizar
import plotly.express as px

for coluna in tabela.columns:
    grafico = px.histogram(tabela, x=coluna, color="cancelou", width=600)
    grafico.show()
```

✓ 5.3s

Para criação dos gráficos nós vamos utilizar a **biblioteca plotly.express**, então se você ainda não instalou a biblioteca basta ir até o terminal e escrever **pip install plotly**.

OBS: Caso tenha problemas na visualização dos gráficos depois de instalar a biblioteca, volte ao terminal e escreva **pip install nbformat**.

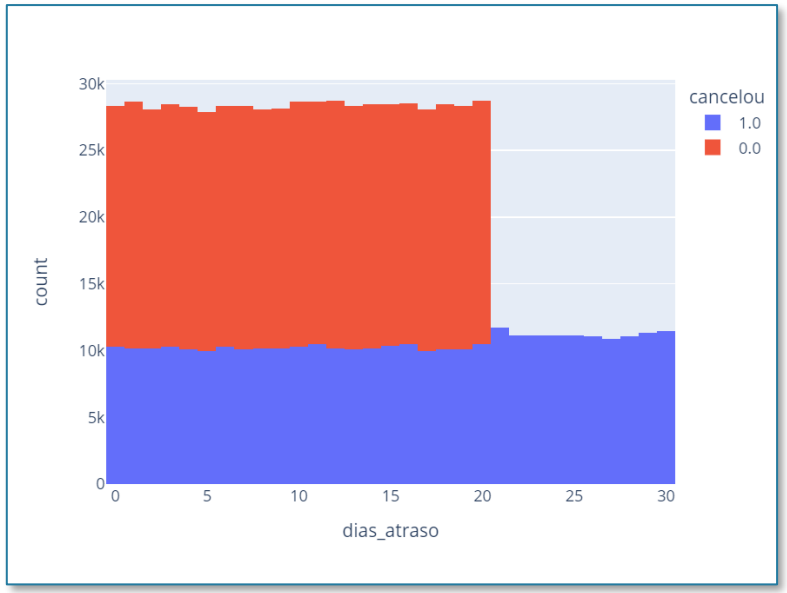
Feito isso, vamos utilizar a estrutura de repetição For para percorrer cada uma das colunas da nossa tabela. Com isso vamos criar um gráfico de Histograma com cada uma das colunas, assim podemos analisar cada uma das informações e verificar como elas se comportam em relação aos cancelamentos da empresa.

Análises Gráficas

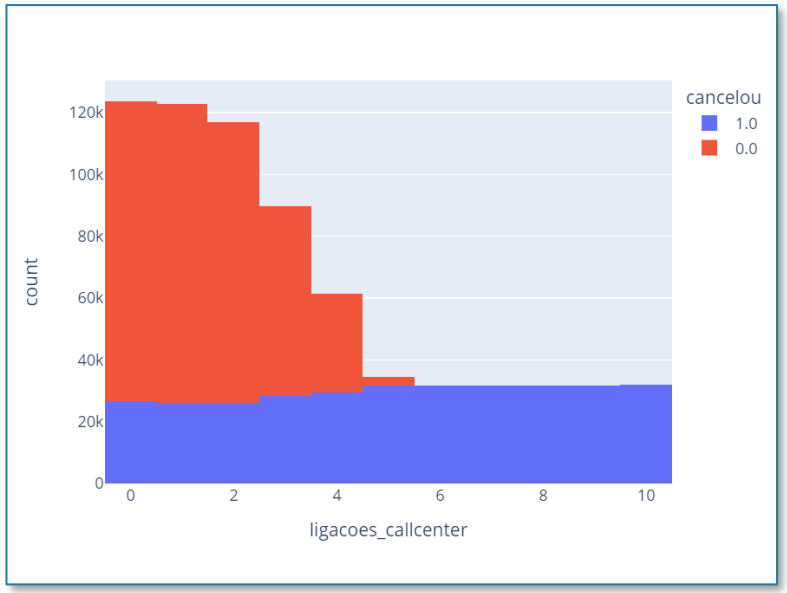
Na base de dados nós temos 11 colunas, isso quer dizer que vamos ter 11 gráficos para analisar.

Aqui vamos focar em 2 desses gráficos, mas você pode dar uma olhada nos outros também para verificar se há possibilidades de melhorar ainda mais esse percentual de cancelamento.

Vamos analisar o gráfico de **Dias de Atraso** e **Ligações ao Call Center**.



Nesse gráfico é possível notar que clientes com **mais de 20 dias de atraso** cancelam suas assinaturas.



Já nesse outro gráfico é possível notar que clientes com **mais de 5 ligações** ao call center cancelam suas assinaturas.

Análises Gráficas

```
# com os graficos a gente consegue descobrir muita coisa:
# dias atraso acima de 20 dias, 100% cancela
# ligações call center acima de 5 todo mundo cancela

tabela = tabela[tabela["ligacoes_callcenter"]<5]
tabela = tabela[tabela["dias_atraso"]<=20]
display(tabela)
display(tabela["cancelou"].value_counts())
display(tabela["cancelou"].value_counts(normalize=True).map("{:.1%}".format))

# se resolvermos isso, já caímos para 18% de cancelamento
# é claro que 100% é utópico, mas com isso já temos as principais causas (ou talvez 3 das principais):
# - forma de contrato mensal
# - necessidade de ligações no call center
# - atraso no pagamento
```

✓ 0.0s

	idade	sexo	tempo_como_cliente	frequencia_uso	ligacoes_callcenter	dias_atraso	assinatura	duracao_contrato	total_gasto	meses_ultima_interacao	cancelou
6	58.0	Female	49.0	12.0	3.0	16.0	Standard	Quarterly	821.00	24.0	1.0
7	55.0	Female	37.0	8.0	4.0	15.0	Premium	Annual	445.00	30.0	1.0
9	64.0	Female	3.0	25.0	2.0	11.0	Standard	Quarterly	415.00	29.0	1.0
13	48.0	Female	35.0	25.0	1.0	13.0	Basic	Annual	518.00	17.0	1.0
19	42.0	Male	15.0	16.0	2.0	14.0	Premium	Quarterly	262.00	16.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
881661	42.0	Male	54.0	15.0	1.0	3.0	Premium	Annual	716.38	8.0	0.0
881662	25.0	Female	8.0	13.0	1.0	20.0	Premium	Annual	745.38	2.0	0.0
881663	26.0	Male	35.0	27.0	1.0	5.0	Standard	Quarterly	977.31	9.0	0.0
881664	28.0	Male	55.0	14.0	2.0	0.0	Standard	Quarterly	602.55	2.0	0.0
881665	31.0	Male	48.0	20.0	1.0	14.0	Premium	Quarterly	567.77	21.0	0.0

464479 rows × 11 columns

```
cancelou
0.0    379032
1.0     85447
Name: count, dtype: int64

cancelou
0.0    81.6%
1.0    18.4%
Name: proportion, dtype: object
```

Analisando apenas os dois gráficos indicados, nós podemos manter na tabela as ligações ao call center que são menores do que 5.

E podemos manter também as informações onde os dias de atraso são menores do que 0.

Com isso podemos visualizar a base de dados e fazer o cálculo novamente para verificar o percentual de cancelamento.

Veja que só removendo as informações dessas duas colunas passamos de um percentual de quase 50% de cancelamento para **18,4%.**

Conseguiu perceber que tivemos que fazer várias etapas até chegar em um valor “aceitável”?

Não é apenas fazer um único tratamento que vamos ter esse resultado.

Parte 6

Conclusão

{JORNADA}
PYTHON
100% GRATUITO E ONLINE

Análises Gráficas

Nós começamos o problema com um taxa de cancelamento de **56,7%**. Após o primeiro tratamento conseguimos diminuir um pouco e atingimos **46,1%**.

No final com ajuda dos gráficos conseguimos ajustar nossa base de dados e chegamos ao percentual de **18,4%** em cancelamentos.

Consegue ver que cada parte da nossa análise é fundamental para chegar o resultado que queremos?

Nessa empresa já seria um avanço muito grande sair de **56%** cancelamentos para **18%** não é mesmo?

E fazendo todas essas análises e tratamentos nós temos como mostrar para empresa onde estão os problemas e o que pode ser feito para minimizar os cancelamentos.

Claro que você pode se aprofundar mais e diminuir ainda mais esse percentual, mas por motivos óbvios nunca vamos chegar em **0%** de cancelamentos.

Mas podemos verificar o que é necessário para reduzir esse número ao máximo.

Se você reparar no gráfico de Idade dos clientes, vai notar que os clientes com idade superior a 50 cancelam suas assinaturas.

Então esse poderia ser mais um parâmetro a ser ajustado no seu modelo para melhorar ainda mais o seu projeto.

Só com esse outro ajuste já teríamos um percentual de cancelamento igual a **12,1%**.

Caso fizesse o ajuste no total gasto, conseguimos um cancelamento de **4,8%**.

Mas vale lembrar que não se trata apenas de diminuir o percentual de cancelamento, é muito importante verificar se é viável.

Imagine que o plano mais caro seja R$200,00, como podemos fazer um ajuste para remover as pessoas que gastaram menos de 500 para chegar em **4%**, sendo que elas precisam gastar dinheiro até chegar a esse nível?

Por isso é muito importante analisar os dados e verificar a viabilidade e não só pensar em chegar a 0% de cancelamentos, pois isso será impossível. Precisamos fazer isso de forma correta e eficiente, por isso poderíamos ter determinado que até **20%** seria um percentual ideal.

Análises Gráficas

É muito importante falar sobre viabilidade, pois certas análises podem não fazer sentido.

A que comentei do valor gasto por exemplo, não faria sentido. Se fosse apenas para obter um número no final poderíamos eliminar todos os cancelamentos da base de dados.

Com isso teríamos 0% de cancelamentos, mas não faria nenhum sentido.

Por isso é importante analisar com calma e verificar se de fato é viável fazer esses ajustes.

Geralmente quando for fazer um projeto desse tipo você vai conversar com o responsável tanto para verificar o que é possível fazer quanto para definir uma meta.

Nesse caso poderíamos ter definido a meta de 20% de cancelamentos. Que já seria muito melhor do que os mais de 50% que tínhamos inicialmente.

Então fica como desafio fazer essas outras análises e verificar se são as únicas possíveis, se existem outros tratamentos que podem ser feitos para minimizar o percentual de cancelamento.

Lembrando sempre de manter a coerência dentro do projeto, veja se de fato faz sentido fazer tal tratamento, até porque a empresa terá que implementar o que você propuser.

Em relação as ligações ao call center por exemplo, a empresa pode verificar onde melhorar para resolver o problema evitando que as ligações sejam mais altas.

Agora fica o desafio, inclusive, você pode começar desde o início para replicar o que aprendeu e fixar ainda mais os conhecimentos.

{ JORNADA } PYTHON

100% GRATUITO E ONLINE

Ainda não segue a gente no Instagram e nem é inscrito no nosso canal do Youtube? Então corre lá!

@hashtagprogramacao

youtube.com/hashtag-programacao