



EDA NO PYTHON

Análise Exploratória de Dados

Bruno Melo da Silva

Quem sou eu

30 anos,
formada em
Ciências
Atuariais

Trabalhei como Assistente
de Processamento de
dados, Analista de dados e
coordenador de Business
Intelligence

Professor de
Analise de dados

Sergipano

MBA em Data
Science e
Analytics

9 anos na área
de análise de
dados



MENSAGEM

Caro leitor,

Bem-vindo a uma jornada emocionante pelo mundo de EDA no Python, onde a arte da manipulação de dados encontra o poder da programação. Em "EDA no Python", embarcaremos juntos em uma exploração profunda e prática dessas ferramentas essenciais para análise de dados.

O Python é a linguagem que dá vida a grandes análises de dados realizadas pelas grandes empresas do mercado, permitindo-nos extrair informações valiosas de vastos conjuntos de dados de maneira eficiente e precisa.

Que este livro seja seu guia confiável enquanto você mergulha no fascinante mundo do Python. Que seja uma oportunidade de aprendizado e crescimento, capacitando-o a lidar com desafios complexos de forma confiante e eficaz.

Que você possa dominar a EDA e o Python, transformando dados em insights valiosos que impulsionarão o sucesso em seus projetos e carreira.

Boa leitura e que sua jornada no mundo da análise de dados seja repleta de descobertas gratificantes!

Atenciosamente,

Bruno Melo

IDENTIFICAÇÃO VISUAL



Dica

Quando você encontrar este elemento visual, saiba que ele representa uma dica ou uma sugestão relacionada ao conteúdo que está sendo apresentado no texto.



Livro

Esse indicador visual representa um livro que a equipe XY recomenda, e você pode ter certeza de que passou pela nossa avaliação rigorosa.



Leitura

Esse elemento visual sugere uma leitura complementar ao assunto em estudo, uma vez que, se incluirmos todos os detalhes, a apostila se tornaria extensa demais. Assim, disponibilizamos alguns textos que enriquecerão o seu aprendizado.



Curso

Neste elemento visual, você encontrará sugestões de cursos, às vezes gratuitos e outras vezes com um custo mais acessível.



MÓDULO EDA



Em alguma fase de seu trabalho, o pesquisador depara-se com o problema de analisar e entender um conjunto de dados relevante ao seu particular objeto de estudos. Ele necessitará trabalhar os dados para transformá-los em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria.

BUSSAB & MORETTIN

MÓDULO EDA

O que é EDA?

EDA (Exploratory Data Analysis) é a Análise Exploratória de Dados, uma etapa crucial em projetos de dados. É o processo de investigar e resumir um conjunto de dados para entender suas características principais, identificar padrões, detectar anomalias, testar hipóteses e verificar pressupostos com a ajuda de estatísticas descritivas e visualizações.

Objetivos do EDA:

- Compreender os Dados
- Identificar Padrões
- Detectar de Anomalias
- Preparar para Modelagem

Ferramentas e Técnicas:

- Estatística Descritiva
- Visualização de Dados (gráficos)
- Manipulação de Dados

A EDA é uma etapa fundamental na estatística e na ciência de dados, dedicada à investigação inicial dos dados para resumir suas principais características, frequentemente utilizando métodos visuais. Seu objetivo é compreender a estrutura dos dados, identificar padrões, detectar anomalias e testar hipóteses preliminares, sem pressupor modelos estatísticos específicos.

A EDA serve como base para análises estatísticas mais aprofundadas, permitindo que pesquisadores e profissionais obtenham insights valiosos sobre os dados antes de aplicar modelos inferenciais ou preditivos. Ela facilita a compreensão das relações entre variáveis e auxilia na preparação dos dados para análises subsequentes.

MÓDULO EDA

Exemplo

Vamos realizar um exemplo básico manualmente para que você compreenda melhor como fazer a chamada EDA (Exploratory Data Analysis). Suponha que temos as notas de um grupo de 10 alunos em um teste (em uma escala de 0 a 10):

6, 8, 5, 7, 9, 5, 6, 10, 8, 7

Primeiramente, vamos organizar esse conjunto de dados do menor para o maior:

5, 5, 6, 6, 7, 7, 8, 8, 9, 10

Até aqui, apenas organizamos os dados em ordem crescente. Essa simples organização já nos permite identificar o Limite Inferior (LI), o Limite Superior (LS), traçar a mediana e começar a entender o comportamento dos dados.

Agora, avançaremos para a estatística descritiva, começando pelas medidas de posição. Essas medidas são úteis porque muitas vezes queremos traduzir perguntas complexas em respostas simples, e é exatamente isso que elas fazem. As principais medidas de posição são: Média, Mediana e Moda.

- A média representa um “número central” dos dados. Por que as aspas? Porque a média é um valor que, em teoria, deveria representar o conjunto de dados, mas, dependendo do contexto, pode distorcer a realidade.

Por exemplo, imagine que o Confiança, time da cidade de Aracaju (conhecida como a Suíça Brasileira), jogou dois jogos e a média de gols marcados foi 3.

Pausa para reflexão...

Embora esse número seja expressivo, ao analisarmos os dados com mais cuidado, percebemos que no primeiro jogo, contra o time pequeno chamado Sergipe, o Confiança aplicou uma goleada de 6 a 0. Já no segundo jogo, um clássico da Série C contra o Itabaiana, o resultado foi um empate de 0 a 0.

MÓDULO EDA

Média

Percebeu a ironia? A média sozinha não conta toda a história. Vamos pensar em outro exemplo: a média salarial de uma empresa. Suponha que ela seja R\$ 5.000. Você, que está participando de um processo seletivo, pensa: "Quero trabalhar nessa empresa, pois começarei ganhando R\$ 5.000."

Mas, se você fez o curso com o Bruno, já sabe que essa conclusão pode não ser verdadeira. A média pode ser influenciada por salários muito altos ou muito baixos, e a maioria dos funcionários pode ganhar bem menos do que os R\$ 5.000.

Você deve estar pensando: "Eu sei tudo sobre a média." Mas calma aí, vamos dar um sustinho nesse coração curioso. A fórmula da média é escrita assim:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Agora complicou, né? Mas calma! Se você aprendeu que a média é calculada somando todas as observações e dividindo pela quantidade de elementos, você está no caminho certo. Essa fórmula nada mais é do que a forma matemática de representar exatamente isso.

Então, voltando ao nosso conjunto de dados das notas (espero que você se lembre dele!), vamos calcular a média:

$$\text{Somatório} = 5 + 5 + 6 + 6 + 7 + 7 + 8 + 8 + 9 + 10 = 71$$

$$\text{Quantidade} = 10$$

Logo, 71 dividido por 10 = 7,1 ou

$$\text{Média} = \frac{6 + 8 + 5 + 7 + 9 + 5 + 6 + 10 + 8 + 7}{10} = \frac{71}{10} = 7,1$$

MÓDULO EDA

Mediana

A mediana é uma medida de tendência central utilizada para representar o valor **central** de um conjunto de dados organizados em ordem crescente ou decrescente. Diferentemente da média, a mediana não é influenciada por valores extremos (muito altos ou muito baixos).

Você lembra que a primeira coisa que fizemos foi organizar os dados em ordem crescente? Pois bem, o cálculo da mediana depende da quantidade de observações no conjunto de dados. Como assim?

Supondo que a quantidade de observações é ímpar, a mediana será o número central.

Vamos a um exemplo com este conjunto de dados:

1, 2, 3, 4, 5

Aqui, o número central é 3, que está exatamente no meio do conjunto. Esse é o valor da mediana.

Percebeu que o central está destacado, ele é literalmente o número que está no meio do seu dataset, mas quando é par temos um problema porque teremos dois números no meio e como ajustar isso? Vamos utilizar o nosso exemplo de notas:

5, 5, 6, 6, 7, 7, 8, 8, 9, 10

Contando da esquerda para direita o meu quinto número é o 7, da direita para esquerda também, mas como tratar isso? A estatística diz que devemos somar os dois números e dividir por 2, nesse caso será o próprio 7.

$$\text{Mediana} = \frac{7+7}{2} = 7$$

MÓDULO EDA

Moda

A moda é a medida de tendência central que representa o valor que ocorre com maior frequência em um conjunto de dados. Em outras palavras, é o valor mais comum ou repetido no conjunto.



A moda tem um conceito parecido com a ideia de uma roupa que está “na moda”. Olhe para a arara de roupas e perceba que há um modelo que aparece mais vezes. Esse modelo é a moda! O mesmo acontece com o seu conjunto de dados: a moda é o valor que mais se repete.

Agora, pense em um levantamento demográfico onde descobrimos que a média de filhos por mãe é 2. Isso significa que, em média, cada mulher tem dois filhos. Mas, ao analisarmos a moda, percebemos que o número 1 é o que mais se repete. Interessante, não?

Essa análise nos mostra que, apesar da média indicar 2 filhos, é mais provável que uma pessoa tenha apenas 1 filho. É aí que está a sagacidade da interpretação dos dados: a média nem sempre conta toda a história, enquanto a moda pode revelar o comportamento mais comum.

Agora, vamos voltar ao nosso dataset e observar algumas curiosidades sobre a moda:

MÓDULO EDA

5, 5, 6, 6, 7, 7, 8, 8, 9, 10

Ao analisarmos os dados, percebemos que não há uma única moda. Os números 5, 6, 7 e 8 se repetem na mesma frequência (duas vezes cada). Isso significa que o conjunto é multimodal, ou seja, possui várias modas e não uma única.

Para sua melhor interpretação a moda pode ser:

- Unimodal: O conjunto tem apenas uma moda.
Exemplo: {1, 2, 2, 3, 4} → Moda: 2
- Bimodal: O conjunto tem duas modas.
Exemplo: {1, 2, 2, 3, 3, 4} → Modas: 2 e 3
- Multimodal: O conjunto tem mais de duas modas.
Exemplo: {1, 1, 2, 2, 3, 3, 4} → Modas: 1, 2 e 3
- Sem moda: Quando nenhum valor se repete.
Exemplo: {1, 2, 3, 4, 5} → Sem moda

OBS: Das três medidas de posição essa é a única que você pode usar para dados quantitativos e qualitativos.

MÓDULO EDA

Visualização de Dados

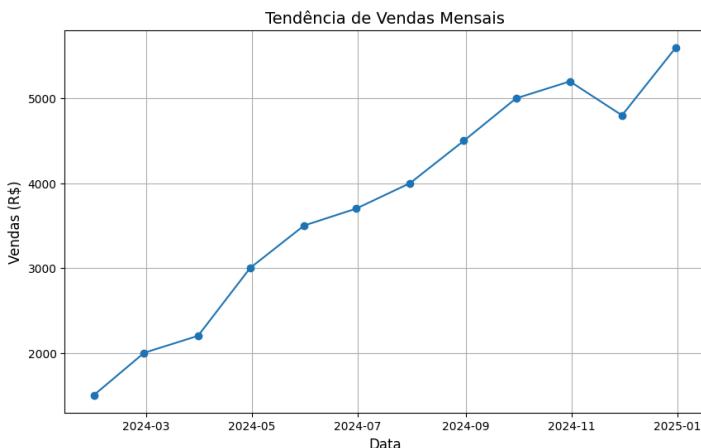
A visualização de dados desempenha um papel essencial na Análise Exploratória de Dados (EDA), ajudando a entender, interpretar e comunicar as características fundamentais de um conjunto de dados. Ela facilita a identificação de padrões, tendências, anomalias e insights que seriam difíceis de detectar apenas com tabelas ou estatísticas descritivas. Vou deixar no Anexo I, dicas para você escolher o melhor gráfico em suas análises.

- Identificação de padrões e tendências

A visualização ajuda a detectar comportamentos recorrentes ou tendências em um conjunto de dados. Gráficos como séries temporais, linhas ou scatter plots são úteis para analisar o comportamento de variáveis ao longo do tempo.

Exemplo prático:

Imagine que você tem os dados de vendas diárias de uma loja ao longo de um ano. Um gráfico de linha pode revelar picos de vendas em datas específicas, como feriados ou períodos de promoção. Isso ajuda a identificar padrões sazonais e tendências de crescimento ou declínio.



MÓDULO EDA

- Detecção de outliers

Outliers são valores que se destacam drasticamente dos demais e podem distorcer análises estatísticas. Gráficos como boxplots e scatter plots tornam esses valores fáceis de visualizar.

Como identificar outliers manualmente?

Para encontrar possíveis valores fora do padrão, podemos usar o método do intervalo interquartil (IQR). Vamos pegar nossa população de dados (dataset), que, completa, representa 100% dos dados. Ao dividir esses dados em quatro partes iguais, obtemos os quartis, que são valores que dividem um conjunto ordenado de dados em quatro partes de tamanhos iguais. Assim, temos:

- $25\% + 25\% + 25\% + 25\% = 100\%$.
-

Definição dos Quartis:

- 1º Quartil (Q1): Representa os 25% inferiores dos dados.
- 2º Quartil (Q2): Abrange os 25% que vão de 25% a 50%, e é onde se encontra a mediana.
- 3º Quartil (Q3): Representa os 25% dos dados entre 50% e 75%.
- 4º Quartil (Q4): Inclui os 25% superiores dos dados, indo de 75% a 100%.

Fórmula para identificar outliers:

Para calcular os outliers, usamos os quartis e o intervalo interquartil (IQR), que é definido como a diferença entre o terceiro e o primeiro quartil:

$$IQR = Q3 - Q1$$

Um dado é considerado outlier se:

- Maior que: $Q3 + (1,5 \times IQR)$
- Menor que: $Q1 - (1,5 \times IQR)$

MÓDULO EDA

- Detecção de outliers

Exemplo Prático:

Vamos analisar a renda mensal de um grupo de pessoas. Suponha que um gráfico boxplot mostre que a maioria ganha entre R\$2.000 e R\$5.000, mas alguns indivíduos têm renda superior a R\$50.000. Esses valores atípicos podem ser investigados para decidir se serão mantidos ou tratados na análise.

Dados:

```
salarios = [2500, 2800, 3000, 3200, 3400, 3500, 4000, 15000, 18000]
```

Passo 1: Ordenação e identificação dos quartis:

- Temos 9 observações no total. Como o número é ímpar, a mediana (Q2) é o valor central, que corresponde a 3400.
- Os dados à esquerda (menores) e à direita (maiores) são usados para calcular Q1 e Q3, respectivamente.
 - Q1: Média dos 25% inferiores → 3000
 - Q3: Média dos 25% superiores → 4000

Passo 2: Calcular o IQR:

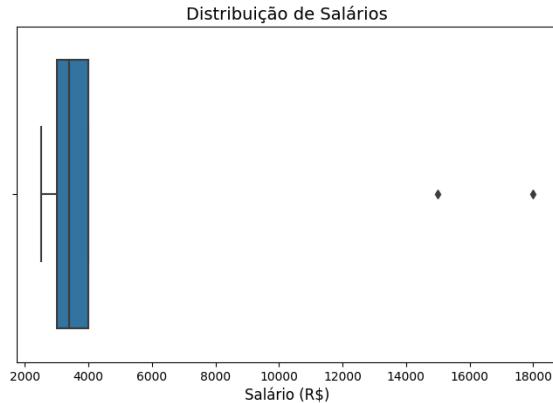
$$\text{IQR} = Q3 - Q1 = 4000 - 3000 = 1000$$

Passo 3: Limites para outliers:

- Limite inferior: $Q1 - (1,5 \times \text{IQR}) = 3000 - (1,5 \times 1000) = 1500$
- Limite superior: $Q3 + (1,5 \times \text{IQR}) = 4000 + (1,5 \times 1000) = 5500$

Os salários 15000 e 18000 são considerados outliers, pois estão acima do limite superior de R\$5500. Podemos usar um boxplot para achar a mesma conclusão que fizemos na mão.

MÓDULO EDA

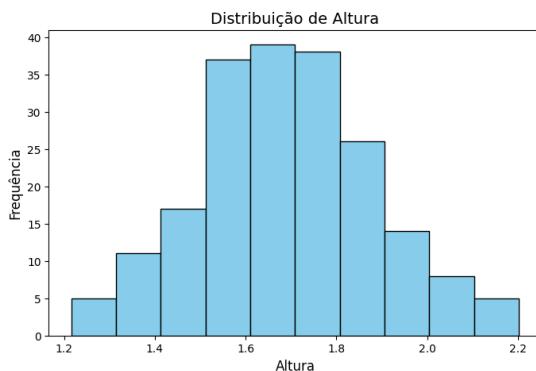


- Compreensão de distribuições

Gráficos como histogramas, gráficos de densidade e gráficos de violino mostram como os dados estão distribuídos, revelando se a distribuição é simétrica, enviesada ou segue uma curva normal.

Exemplo prático:

Ao analisar a altura de uma população, um histograma pode mostrar que a maioria das pessoas tem entre 1,60 m e 1,80 m, enquanto poucas têm alturas abaixo de 1,50 m ou acima de 2,00 m. Esse entendimento é crucial para avaliar a variabilidade dos dados.



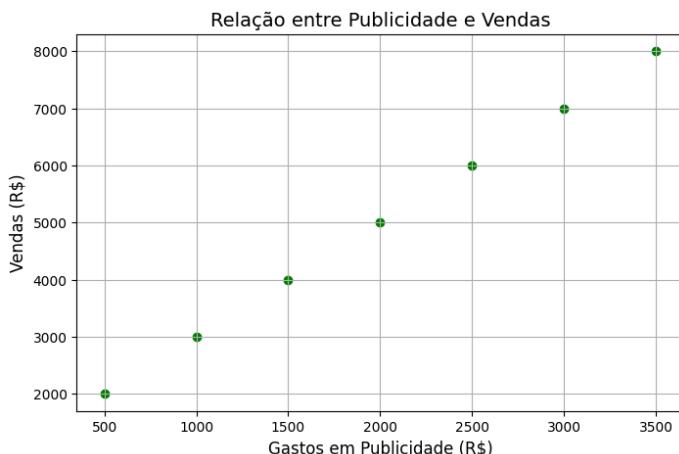
MÓDULO EDA

- Análise de relações entre variáveis

Gráficos bivariados, como scatter plots e heatmaps, ajudam a visualizar como duas variáveis se relacionam. Isso é especialmente útil para identificar correlações ou interdependências.

Exemplo prático:

Considere um conjunto de dados que relaciona a publicidade em mídias sociais e o volume de vendas. Um scatter plot (dispersão/bolhas) pode mostrar que, à medida que os gastos com publicidade aumentam, as vendas também crescem, sugerindo uma correlação positiva.



MÓDULO EDA

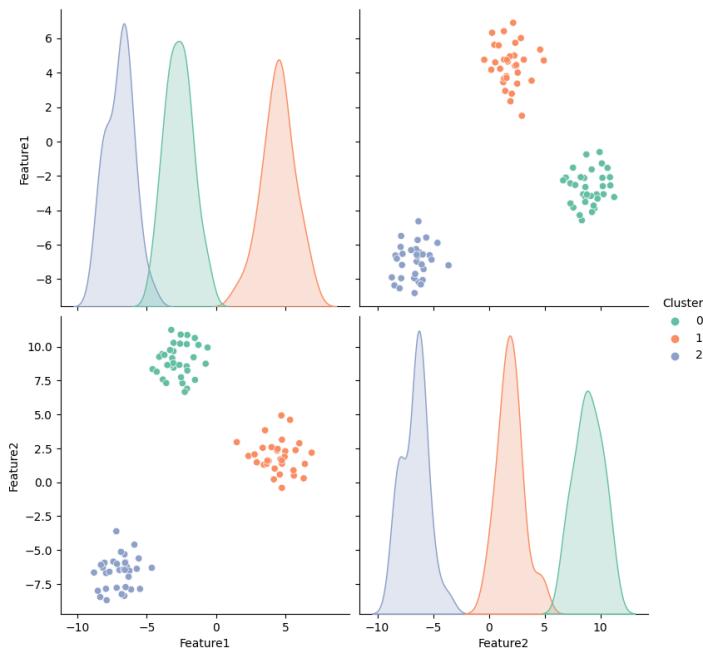
- Segmentação e agrupamento

A visualização permite identificar padrões de agrupamento ou segmentação natural nos dados. Essa abordagem é especialmente relevante quando lidamos com dados multidimensionais.

Exemplo prático:

Ao analisar o perfil de clientes de um e-commerce, gráficos 2D após redução dimensional (como PCA) podem mostrar grupos distintos de clientes com base em comportamentos de compra. Isso pode indicar segmentos como “clientes fiéis”, “compradores sazonais” e “clientes inativos”.

Essa análise não é trivial, tem um nível maior de complexidade



MÓDULO EDA

- Comunicação dos resultados

Além de explorar os dados, a visualização é crucial para apresentar insights de maneira clara e compreensível, mesmo para públicos não técnicos. Gráficos ajudam a contar histórias com os dados e facilitar a tomada de decisão.

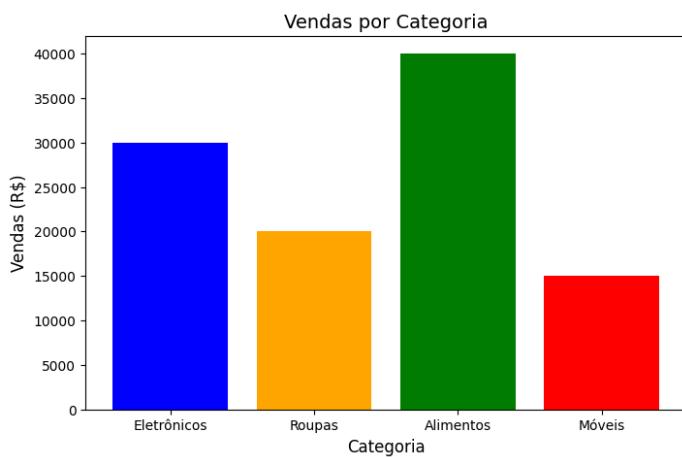
Exemplo prático:

Se você está apresentando o desempenho financeiro de uma empresa para gestores, um gráfico de barras comparando a receita mensal pode ser mais eficiente do que uma tabela cheia de números, destacando facilmente meses de maior e menor faturamento.

Livro



Storytelling com Dados: um Guia Sobre Visualização de Dados Para Profissionais de Negócios. [Link](#)

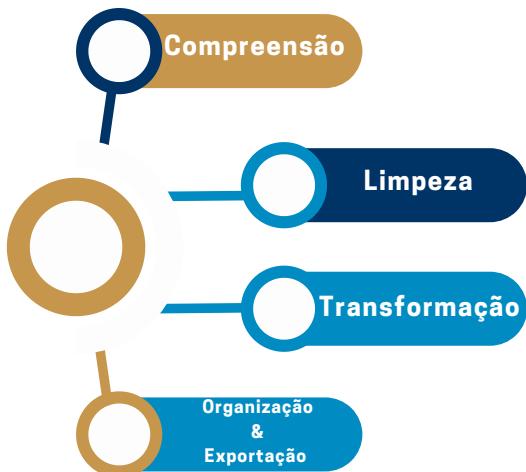


MÓDULO EDA

Manipulação de Dados

A Manipulação de Dados é uma das etapas mais importantes da Análise Exploratória de Dados (EDA). Antes de mergulharmos em ferramentas automatizadas como o Python, vamos entender como realizar essa tarefa manualmente, explorando o processo com um olhar mais prático e didático. A manipulação manual permite compreender melhor os dados, identificar erros e preparar o conjunto de dados para análises futuras.

Etapas da Manipulação de Dados



MÓDULO EDA

Compreensão Inicial dos Dados

Antes de realizar qualquer manipulação, é essencial entender o contexto do conjunto de dados. Isso envolve:

- Identificar o objetivo da análise.
- Conhecer as variáveis disponíveis (números, textos, datas, etc.).
- Observar o tamanho do conjunto de dados (número de linhas e colunas).
- Determinar a origem e a qualidade dos dados.

Exemplo Prático:

Imagine que temos o seguinte conjunto de dados referente às notas de alunos em três disciplinas:

Aluno	Matemática	Português	Ciências
Ana	7	8	9
Bruno	6	5	8
Carla	9	10	-
Diego	5	7	6

Observação Inicial: Existem dados ausentes na coluna “Ciências” para Carla (indicado por um hífen). Isso requer tratamento.

Limpeza dos Dados

Os dados precisam ser organizados para que possam ser analisados corretamente. A limpeza manual envolve as seguintes atividades:

a) Tratamento de Dados Ausentes

- Opção 1: Substituir valores ausentes: Escolher um valor que represente razoavelmente o dado faltante. Por exemplo, usar a média da coluna ou zero

MÓDULO EDA

Opção 2: Remover linhas ou colunas: Caso o valor ausente não seja relevante para a análise.

Exemplo:

- Substituir o valor ausente da coluna "Ciências" pela média das notas dessa disciplina, Atualizando o valor para Carla:

Aluno	Matemática	Português	Ciências
Ana	7	8	9
Bruno	6	5	8
Carla	9	10	7,67
Diego	5	7	6

b) Correção de Inconsistências

Corrigir valores que estejam fora do esperado, como erros de digitação ou unidades diferentes. Por exemplo:

- Transformar "1.000" para "1000".
- Padronizar datas no formato DD/MM/AAAA.

c) Remoção de Duplicatas

Certificar-se de que não existem linhas repetidas. Se houver, elas devem ser removidas para evitar redundância nos cálculos.

Transformação dos Dados

A transformação de dados consiste em reorganizar ou criar novas variáveis para facilitar a análise. As principais transformações incluem:

a) Criação de Colunas Derivadas

Criar novas colunas com base nas existentes. Por exemplo:

- Adicionar uma coluna "Média Geral" para cada aluno:

MÓDULO EDA

Aluno	Matemática	Português	Ciências	Média Geral
Ana	7	8	9	8
Bruno	6	5	8	6,33
Carla	9	10	7,67	8,89
Diego	5	7	6	6

b) Agrupamento e Soma

Se tivermos vários registros para um mesmo aluno ou categoria, podemos somar ou calcular a média para consolidar os dados.

Exemplo: Se houvesse mais de uma nota para cada aluno em diferentes avaliações, poderíamos calcular a média final de cada disciplina.

c) Normalização dos Dados

Redimensionar os dados para que fiquem em uma mesma escala. Por exemplo:

- Converter notas de 0 a 10 para uma escala de 0 a 100 multiplicando por 10.

Organização e Exportação

Após realizar as transformações, é importante reorganizar os dados para facilitar a leitura.

- Ordenar linhas por ordem alfabética ou por valores (ex.: maior para menor).
- Reorganizar colunas de acordo com a relevância.
- Exportar os dados para um formato adequado, como uma tabela impressa ou arquivo CSV.

Aluno	Matemática	Português	Ciências	Média Geral
Ana	7	8	9	8
Bruno	6	5	8	6,33
Carla	9	10	7,67	8,89
Diego	5	7	6	6

MÓDULO EDA

A manipulação de dados é uma forma eficaz de entender os desafios e peculiaridades de um conjunto de dados. Este processo envolve limpeza, transformação e organização, garantindo que os dados estejam prontos para a EDA.

Nas próximas etapas, migraremos esse processo para o Python, onde será possível automatizar e reproduzir facilmente essas tarefas em grandes conjuntos de dados.

Correlação

A correlação é uma medida estatística que indica o grau de relacionamento entre duas variáveis. Em outras palavras, ela nos ajuda a entender se duas coisas estão relacionadas e, se sim, como.

Exemplo:

Imagine que você está analisando dois dados:

1. A quantidade de sorvetes vendidos em um dia.
2. O número de pessoas que sofreram queimaduras de sol no mesmo dia.

Ao organizar esses dados, você percebe que nos dias em que mais sorvetes são vendidos, também há mais casos de queimaduras de sol.

O que isso significa?

- Correlação positiva: Quando a quantidade de sorvetes vendidos aumenta, os casos de queimaduras de sol também aumentam.
- Isso sugere que existe uma relação entre essas duas variáveis, mas atenção: correlação não implica causalidade.

Explicando melhor

Essa relação ocorre porque ambos os eventos (sorvetes e queimaduras de sol) estão ligados a um terceiro fator: o calor.

MÓDULO EDA

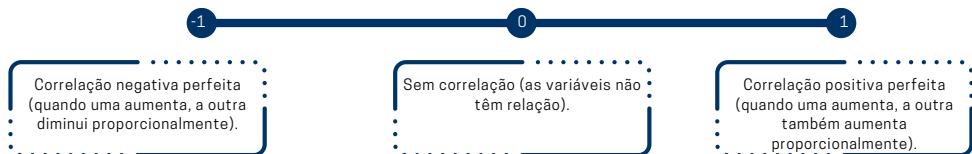
Nos dias mais quentes:

- As pessoas compram mais sorvetes para se refrescar.
- Também passam mais tempo ao ar livre, aumentando o risco de queimaduras de sol.

Portanto, o calor é a causa em comum que afeta ambas as variáveis.

Como calcular a correlação?

A correlação é representada por um valor que vai de -1 a 1:



Vamos utilizar a correlação de Pearson (r), vou demonstrar a fórmula aqui, mas não fique assustado.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

Onde:

- x_i e y_i são os valores das variáveis X e Y .
- \bar{x} e \bar{y} são as médias de X e Y .

- Quantidade de sorvetes vendidos (X): [10,20,30,40,50]
- Queimaduras de sol (Y): [5,7,15,20,25]

Passo 1: Calcular as médias (\bar{x} e \bar{y})

$$\bar{x} = \frac{\sum X}{n} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30$$

$$\bar{y} = \frac{\sum Y}{n} = \frac{5 + 7 + 15 + 20 + 25}{5} = 14.4$$

MÓDULO EDA

Passo 2: Calcular as diferenças em relação às médias

$$(x_i - \bar{x}) = [10 - 30, 20 - 30, 30 - 30, 40 - 30, 50 - 30] = [-20, -10, 0, 10, 20]$$

$$(y_i - \bar{y}) = [5 - 14.4, 7 - 14.4, 15 - 14.4, 20 - 14.4, 25 - 14.4] = [-9.4, -7.4, 0.6, 5.6, 10.6]$$

Passo 3: Calcular os produtos das diferenças $((x_i - \bar{x})(y_i - \bar{y}))$

$$(x_i - \bar{x})(y_i - \bar{y}) = [-20 \cdot -9.4, -10 \cdot -7.4, 0 \cdot 0.6, 10 \cdot 5.6, 20 \cdot 10.6]$$

$$= [188, 74, 0, 56, 212]$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 188 + 74 + 0 + 56 + 212 = 530$$

Passo 4: Calcular os quadrados das diferenças

$$(x_i - \bar{x})^2 = [-20^2, -10^2, 0^2, 10^2, 20^2] = [400, 100, 0, 100, 400]$$

$$\sum(x_i - \bar{x})^2 = 400 + 100 + 0 + 100 + 400 = 1000$$

$$(y_i - \bar{y})^2 = [-9.4^2, -7.4^2, 0.6^2, 5.6^2, 10.6^2] = [88.36, 54.76, 0.36, 31.36, 112.36]$$

$$\sum(y_i - \bar{y})^2 = 88.36 + 54.76 + 0.36 + 31.36 + 112.36 = 287.2$$

Passo 5: Substituir na fórmula

Agora que temos todos os valores, substituímos na fórmula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

$$r = \frac{530}{\sqrt{1000} \cdot \sqrt{287.2}}$$

$$r = \frac{530}{31.62 \cdot 16.95}$$

$$r = \frac{530}{535.99} \approx 0.99$$

O coeficiente de correlação $r=0.99$ indica uma forte correlação positiva entre a venda de sorvetes e os casos de queimadura de sol.

MÓDULO EDA

Isso significa que, conforme as vendas de sorvetes aumentam, o número de queimaduras de sol também tende a aumentar (mas lembre-se de que correlação não implica causalidade!).

Correlação não implica causalidade!

Outra observação correlação é para dados numéricos, caso o objetivo seja comparação de dados categóricos você vai utilizar análise de correspondência.

Teste de Hipótese

O teste de hipótese é uma ferramenta estatística fundamental para a tomada de decisão baseada em dados. Ele permite que você avalie uma suposição sobre uma população com base em dados amostrais, utilizando um processo lógico e matemático.

Imagine que você é um cientista tentando entender se uma nova medicação é mais eficaz do que o tratamento atual. Você não pode testar toda a população, então analisa uma amostra e utiliza um teste de hipótese para decidir se há evidências suficientes para concluir que a nova medicação é realmente melhor.

Passos de um Teste de Hipótese

Definir as Hipóteses

H_0, H_1

- Hipótese nula (H_0): Representa o estado inicial ou a ausência de efeito. Exemplo: "A média da nova medicação é igual à média da antiga."
- Hipótese alternativa (H_1): Contradiz a hipótese nula. Exemplo: "A média da nova medicação é maior que a da antiga."

Escolher um nível de significância (α)



O valor de α define a probabilidade máxima de rejeitar H_0 quando ela é verdadeira. Comumente, usa-se 5% ($\alpha=0.05$).

MÓDULO EDA

Escolher o teste estatístico adequado

Dependendo do tipo de dado e do problema, utiliza-se testes como o teste t, z, χ^2 , entre outros.

Calcular a estatística do teste e o valor p

- A estatística do teste mede a diferença entre os dados observados e o esperado sob H_0 .
- O valor p é a probabilidade de observar os dados (ou algo mais extremo) assumindo H_0 verdadeira.

Tomar a decisão

- Se $p < \alpha$: Rejeita-se H_0 (não há evidências para aceitar H_0 / não há evidências para rejeitar H_1 / aceitamos H_1 / A média da nova medicação é maior que a da antiga).
- Se $p \geq \alpha$: Não rejeitamos H_0 (não há evidências suficientes para aceitar H_1 / não há evidências suficientes para rejeitar H_0 / aceitamos H_0 / A média da nova medicação é igual à média da antiga).

Exemplo: Vamos realizar um teste t de Student para duas amostras independentes.

Você deseja verificar se duas turmas (A e B) têm médias significativamente diferentes em uma prova.

Dados amostrais:

- Turma A: [75, 80, 85, 90, 95]
- Turma B: [70, 75, 80, 85, 90]

μ = média, estatisticamente μ é usado para população e x^- é para amostra.

Hipóteses:

- H_0 (Hipótese nula): Não há diferença significativa entre as médias das turmas ($x^-A=x^-B$).
- H_1 (Hipótese alternativa): Há uma diferença significativa entre as médias ($x^-A \neq x^-B$).

Nível de significância (α): 0,05 (5%).

MÓDULO EDA

Passo 1: Calcular as médias (\bar{x}_A e \bar{x}_B)

$$\bar{x}_A = \frac{\sum X_A}{n_A} = \frac{75 + 80 + 85 + 90 + 95}{5} = 85$$
$$\bar{x}_B = \frac{\sum X_B}{n_B} = \frac{70 + 75 + 80 + 85 + 90}{5} = 80$$

Passo 2: Calcular os desvios padrão (s_A e s_B)

A fórmula do desvio padrão amostral é:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

Para Turma A:

$$s_A = \sqrt{\frac{(75 - 85)^2 + (80 - 85)^2 + (85 - 85)^2 + (90 - 85)^2 + (95 - 85)^2}{5 - 1}}$$
$$s_A = \sqrt{\frac{100 + 25 + 0 + 25 + 100}{4}} = \sqrt{\frac{250}{4}} = \sqrt{62.5} \approx 7.91$$

Para Turma B:

$$s_B = \sqrt{\frac{(70 - 80)^2 + (75 - 80)^2 + (80 - 80)^2 + (85 - 80)^2 + (90 - 80)^2}{5 - 1}}$$
$$s_B = \sqrt{\frac{100 + 25 + 0 + 25 + 100}{4}} = \sqrt{\frac{250}{4}} = \sqrt{62.5} \approx 7.91$$

Passo 3: Calcular o valor de t

A fórmula para o t-teste para duas amostras independentes é:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Substituindo os valores:

$$t = \frac{85 - 80}{\sqrt{\frac{7.91^2}{5} + \frac{7.91^2}{5}}}$$

$$t = \frac{5}{\sqrt{\frac{62.5}{5} + \frac{62.5}{5}}}$$

$$t = \frac{5}{\sqrt{12.5 + 12.5}} = \frac{5}{\sqrt{25}} = \frac{5}{5} = 1.0$$

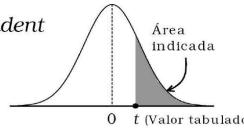
MÓDULO EDA

Passo 4: Determinar o valor crítico

Com $\alpha=0.05$ e graus de liberdade ($df = nA + nB - 2 = 5 + 5 - 2 = 8$), consultamos uma tabela t:

- Valor crítico t para $\alpha=0.05$ (bicaudal) e $df = 8 \pm 2.306$

Tabela 5 Distribuição t de Student

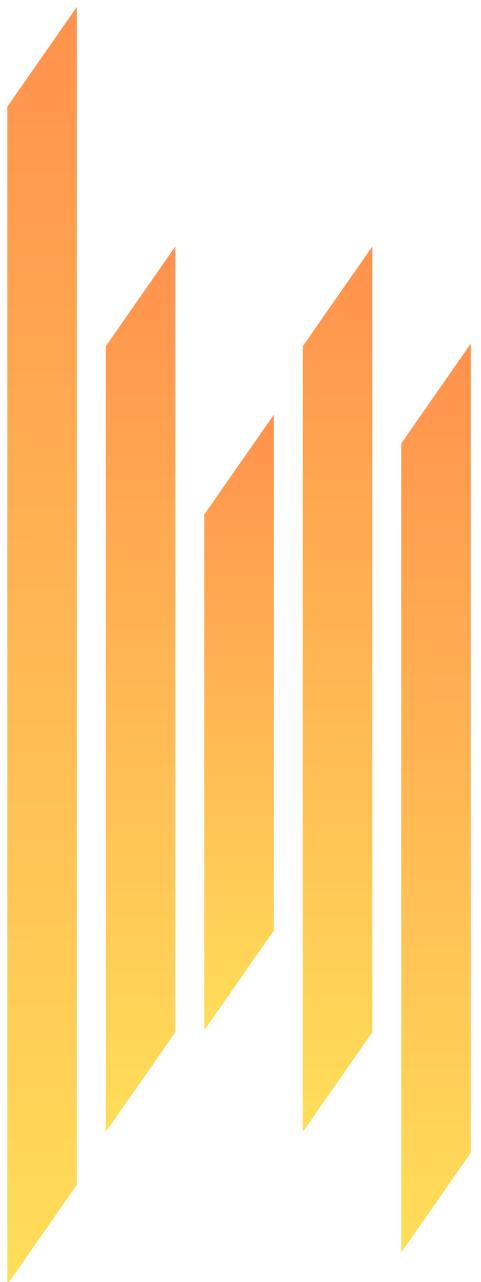


gl	Área na cauda superior								
	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496

Passo 5: Concluir

- O valor calculado ($t=1,0$) está dentro do intervalo aceito ($-2,306 \leq t \leq 2,306$).
- Portanto, não rejeitamos H_0 .

Desta forma, não há evidências suficientes para afirmar que as médias das turmas são significativamente diferentes.



MÓDULO EDA NO PYTHON



A matemática é o alfabeto que Deus usou para escrever o Universo.

Galileu Galilei

MÓDULO EDA

Exemplo 1 - Dataset titanic

Primeiro passo imports de bibliotecas:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Segundo passo importar banco de dados:

```
# Carregar um conjunto de dados (exemplo: Titanic)
df = sns.load_dataset('titanic')
df.head()
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

Terceiro passo entender o nosso banco de dados:

```
# 1. Compreensão dos Dados
print("Informações gerais sobre o dataset:")
print(df.info())

print("\nEstatísticas descritivas:")
print(df.describe(include='all')) # Inclui tanto numéricos

print("\nTipos de dados e valores ausentes:")
print(df.isnull().sum())

print("\nNúmero de duplicatas:")
print(df.duplicated().sum())
```

MÓDULO EDA

```
Informações gerais sobre o dataset:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 891 entries, 0 to 890  
Data columns (total 15 columns):  
 #   Column      Non-Null Count  Dtype     
---    
 0   survived    891 non-null   int64    
 1   pclass      891 non-null   int64    
 2   sex         891 non-null   object    
 3   age         714 non-null   float64   
 4   sibsp       891 non-null   int64    
 5   parch       891 non-null   int64    
 6   fare        891 non-null   float64   
 7   embarked    889 non-null   object    
 8   class        891 non-null   category    
 9   who          891 non-null   object    
 10  adult_male  891 non-null   bool     
 11  deck         203 non-null   category    
 12  embark_town  889 non-null   object    
 13  alive        891 non-null   object    
 14  alone        891 non-null   bool     
dtypes: bool(2), category(2), float64(2), int64(4), object(5)  
memory usage: 80.7+ KB  
None  
  
Estatísticas descritivas:  
   survived   pclass   sex      age     sibsp     parch \\\br/>count  891.000000 891.000000 891 714.000000 891.000000 891.000000  
unique   NaN        NaN        2   NaN        NaN        NaN  
top      NaN        NaN        male  NaN        NaN        NaN  
freq     NaN        NaN        577  NaN        NaN        NaN  
mean    0.383838 2.308642  NaN 29.699118 0.523008 0.381594  
std     0.486592 0.836071  NaN 14.526497 1.102743 0.806057  
min     0.000000 1.000000  NaN 0.420000 0.000000 0.000000  
25%    0.000000 2.000000  NaN 20.125000 0.000000 0.000000  
50%    0.000000 3.000000  NaN 28.000000 0.000000 0.000000  
75%    1.000000 3.000000  NaN 38.000000 1.000000 0.000000  
max     1.000000 3.000000  NaN 80.000000 8.000000 6.000000  
  
   fare embarked class  who  adult_male  deck  embark_town  alive \\\br/>count  891.000000 889 891 891 891 203 889 891  
unique   NaN        3 3 3 2 7 3 2  
top      NaN        S Third man True C Southampton no  
freq     NaN        644 491 537 537 59 644 549  
mean    32.204208  NaN  NaN  NaN  NaN  NaN  NaN  NaN  
...  
dtype: int64  
  
Número de duplicatas:  
107
```

MÓDULO EDA

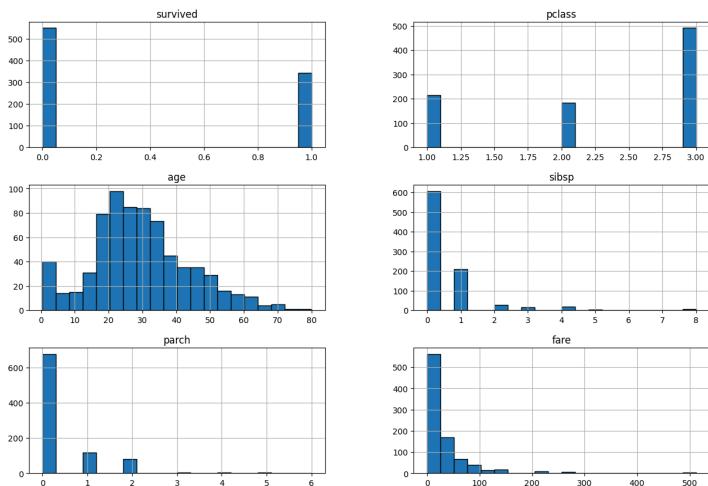
Quarto passo identificar padrões:

```
# Analisar a distribuição das variáveis numéricas
df.hist(bins=20, figsize=(15, 10), edgecolor='black')
plt.suptitle("Distribuição das Variáveis Numéricas")
plt.show()

# Analisar a relação entre variáveis
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Matriz de Correlação")
plt.show()

# Boxplot para verificar a relação entre 'survived' e 'age'
plt.figure(figsize=(8, 6))
sns.boxplot(x='survived', y='age', data=df)
plt.title("Distribuição de Idade por Status de Sobrevida")
plt.show()
```

Distribuição das Variáveis Numéricas

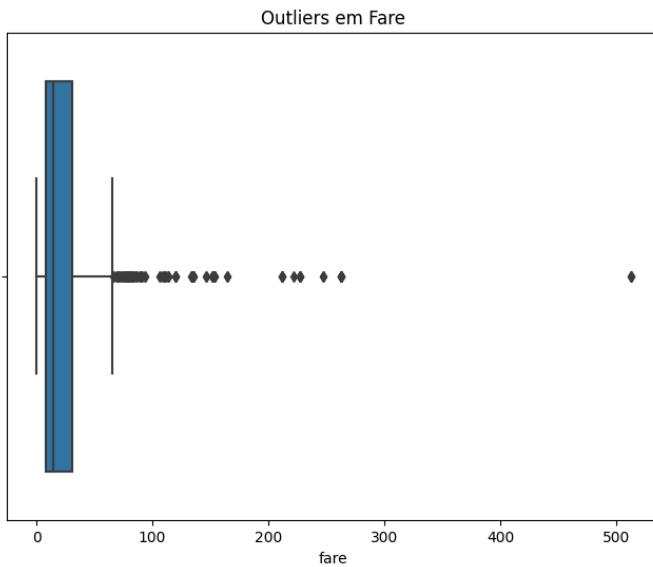


MÓDULO EDA

Quinto passo detecção de anomalias:

```
# Identificar outliers usando Boxplot para 'fare' (tarifa paga)
plt.figure(figsize=(8, 6))
sns.boxplot(x=df['fare'])
plt.title("Outliers em Fare")
plt.show()

# Identificar valores extremos no resumo estatístico
Q1 = df['fare'].quantile(0.25)
Q3 = df['fare'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['fare'] < Q1 - 1.5 * IQR) | (df['fare'] > Q3 + 1.5 * IQR)]
print("\nNúmero de outliers em 'fare':", len(outliers))
```



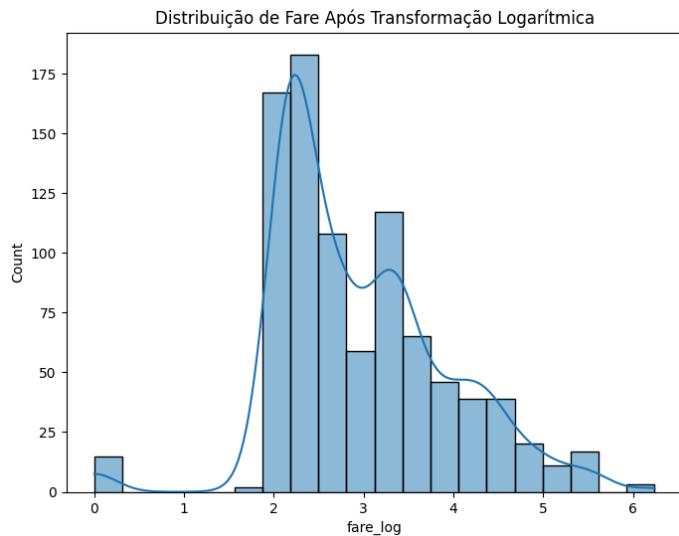
MÓDULO EDA

Sexto passo preparação para modelagem:

```
# 4. Preparação para Modelagem
# Lidando com valores ausentes
df_cleaned = df.copy()
df_cleaned['age'].fillna(df['age'].median(), inplace=True) # Preenche com a mediana
df_cleaned['embark_town'].fillna('Unknown', inplace=True) # Substitui com 'Unknown'
df_cleaned.dropna(subset=['fare'], inplace=True) # Remove linhas onde 'fare' está ausente

# Verificar a normalização de variáveis numéricas (exemplo: escala logarítmica para 'fare')
df_cleaned['fare_log'] = np.log1p(df_cleaned['fare']) # log(1 + fare)

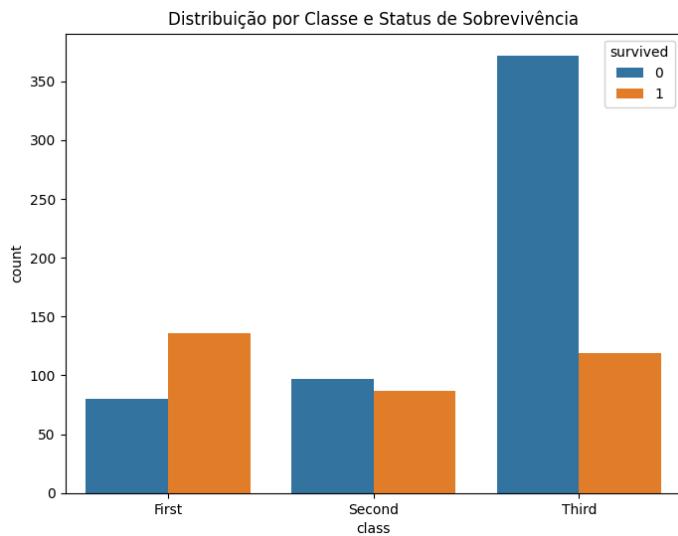
# Visualizar a nova distribuição
plt.figure(figsize=(8, 6))
sns.histplot(df_cleaned['fare_log'], kde=True, bins=20)
plt.title("Distribuição de Fare Após Transformação Logarítmica")
plt.show()
```



MÓDULO EDA

Sétimo passo análise adicional: Visualizar variáveis categóricas:

```
# 5. Análise adicional: Visualizar variáveis categóricas
plt.figure(figsize=(8, 6))
sns.countplot(x='class', hue='survived', data=df_cleaned)
plt.title("Distribuição por Classe e Status de Sobrevivência")
plt.show()
```



MÓDULO EDA

Exemplo 2 - Quantidade de sorvetes vendidos e queimaduras de sol

```
import pandas as pd

# Dados: Quantidade de sorvetes vendidos e queimaduras de sol
dados = {
    'Sorvetes': [10, 20, 30, 40, 50],
    'Queimaduras': [5, 7, 15, 20, 25]
}

# Criar um DataFrame
df = pd.DataFrame(dados)

# Calcular a correlação de Pearson
correlacao = df['Sorvetes'].corr(df['Queimaduras'])

print("Coeficiente de Correlação de Pearson:", correlacao)
```

Exemplo 3 - Comparação das notas entre as turmas

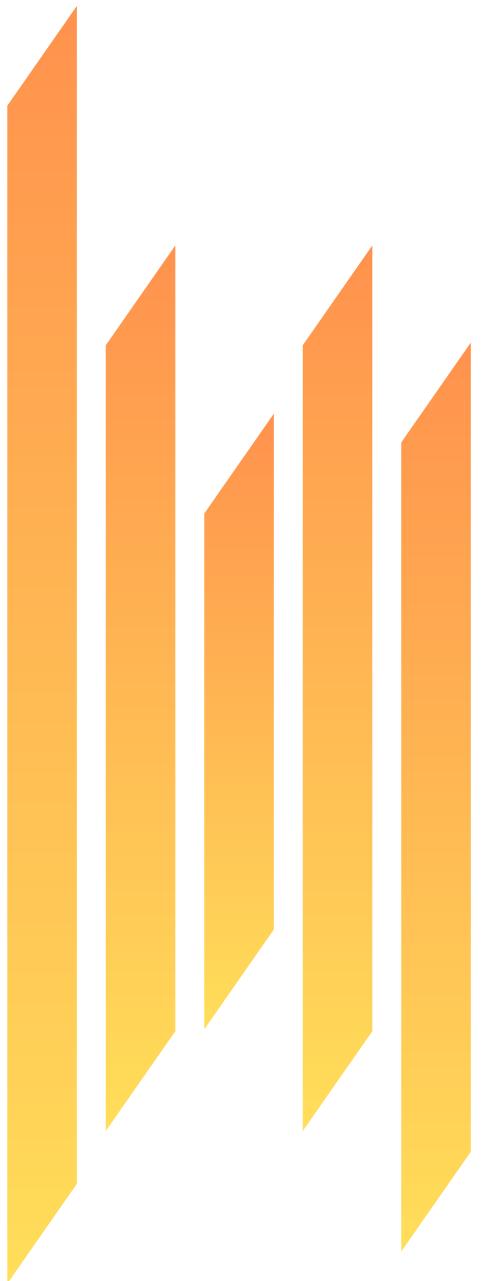
```
import scipy.stats as stats

# Dados das duas turmas
turma_a = [75, 80, 85, 90, 95]
turma_b = [70, 75, 80, 85, 90]

# Teste t para amostras independentes
t_stat, p_value = stats.ttest_ind(turma_a, turma_b)

# Exibir resultados
print("Estatística t:", t_stat)
print("Valor p:", p_value)

# Conclusão
alpha = 0.05
if p_value < alpha:
    print("Rejeitamos H0: Há diferença significativa entre as médias.")
else:
    print("Não rejeitamos H0: Não há diferença significativa entre as médias.")
```



ANEXO 1

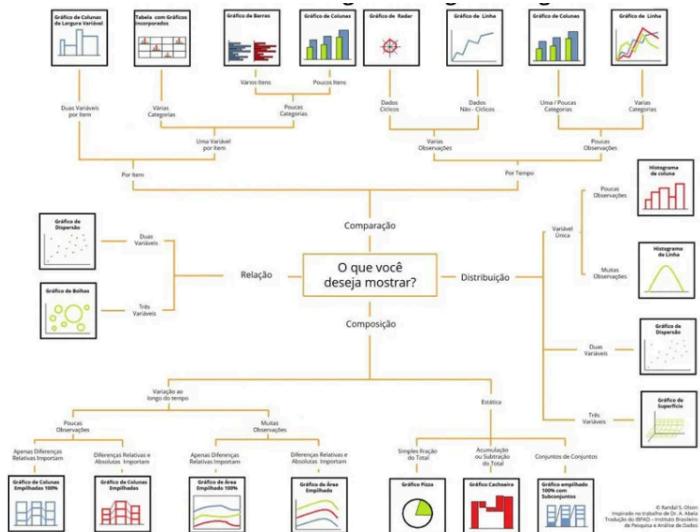


O gráfico não é para você.

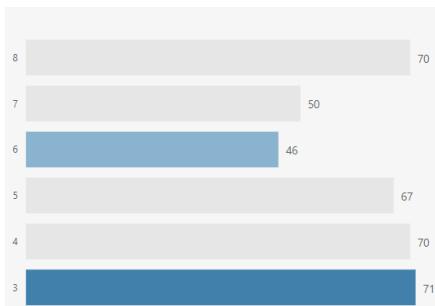
Gustavo Esteves

MÓDULO EDA

Como escolher o seu gráfico?



BARRAS

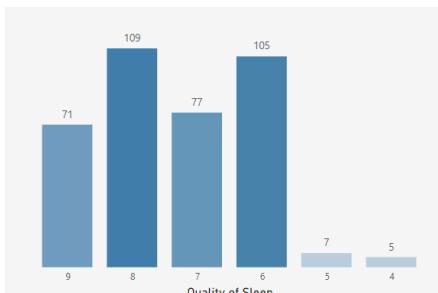


O gráfico de barras é uma representação visual de dados estatísticos ou quantitativos por meio de barras retangulares. Cada barra representa uma categoria ou grupo específico e o comprimento da barra está proporcional à quantidade ou valor associado a essa categoria.

- Comparação de Categorias
- Tendências ao Longo do Tempo
- Partes de um Todo (Gráfico de Barras Empilhadas)
- Classificação de Dados Nominais
- Facilidade de Comparação Entre Diversas Categorias
- Visualização de Dados Discretos
- Comparação de Dados Categóricos

MÓDULO EDA

COLUNAS



O gráfico de barras é uma representação visual de dados estatísticos ou quantitativos por meio de barras retangulares. Cada barra representa uma categoria ou grupo específico e o comprimento da barra está proporcional à quantidade ou valor associado a essa categoria.

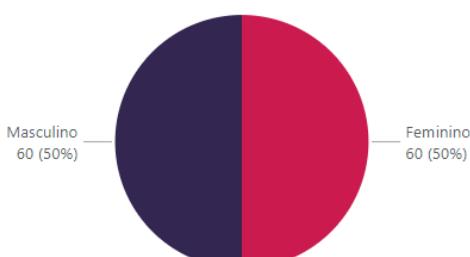
- Comparação de Categorias
- Tendências ao Longo do Tempo
- Partes de um Todo (Gráfico de Barras Empilhadas)
- Classificação de Dados Nominais
- Facilidade de Comparação Entre Diversas Categorias
- Visualização de Dados Discretos
- Comparação de Dados Categóricos

LINHAS

O gráfico de linhas é uma representação visual de dados em que informações são exibidas através de uma série de pontos conectados por linhas. Esse tipo de gráfico é particularmente útil para mostrar a relação entre duas variáveis contínuas, geralmente ao longo de um eixo de tempo. Cada ponto no gráfico de linhas representa um valor específico da variável, e as linhas conectam esses pontos na sequência em que aparecem.



PIZZA



O gráfico de pizza é uma representação visual que utiliza um círculo dividido em fatias para mostrar a proporção ou a distribuição percentual de um todo. Cada fatia representa uma parte específica ou categoria, e a área da fatia é proporcional à porcentagem que ela contribui para o total. Esse tipo de gráfico é eficaz para destacar a participação relativa de categorias quando o número de categorias é limitado.

O gráfico de pizza é comumente utilizado para representar a distribuição percentual de categorias, especialmente quando se quer enfatizar a participação relativa de cada categoria em relação ao todo.

MÓDULO EDA

ROSCA

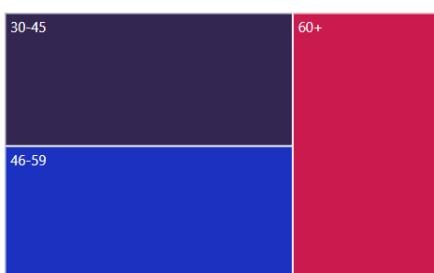
O gráfico de rosca é uma variação do gráfico de pizza que possui um espaço vazio no centro, criando uma estrutura anelar. Cada fatia no anel representa uma categoria, e a área do anel é proporcional à porcentagem que cada categoria contribui para o total. O espaço central vazio pode ser usado para destacar informações adicionais ou para melhorar a legibilidade do gráfico.

O gráfico de rosca é utilizado quando se deseja manter a noção de participação percentual de categorias, mas também oferece uma oportunidade para adicionar elementos visuais ou informações ao espaço central vazio, tornando-o mais versátil em alguns contextos.

Ambos os gráficos de pizza e rosca são eficazes para representar distribuições percentuais de categorias, mas é importante usá-los com moderação, especialmente quando o número de categorias é grande, pois a interpretação pode se tornar mais desafiadora.

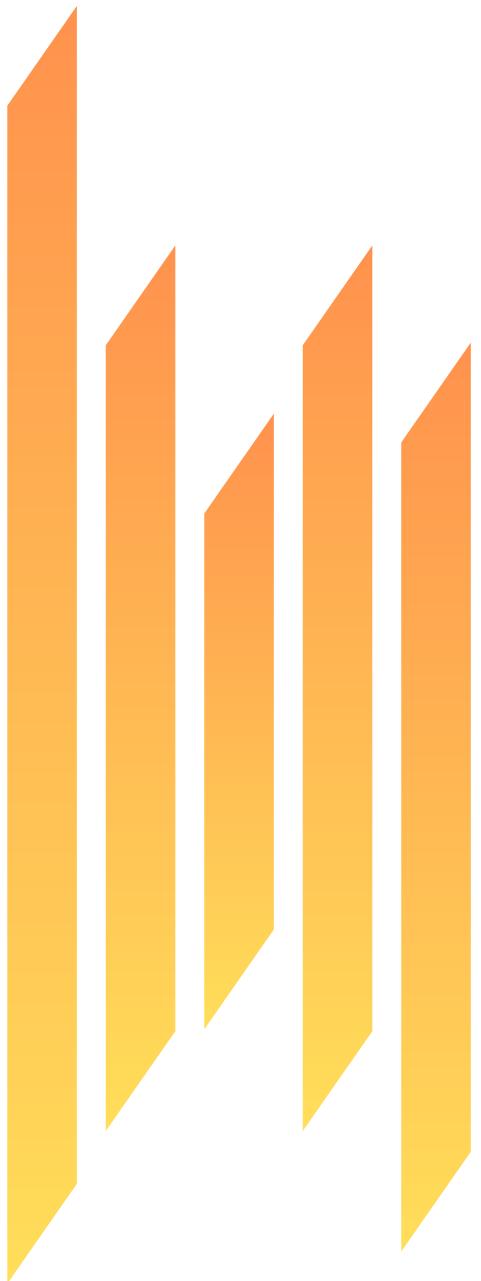


TREEMAP



O gráfico treemap é uma representação visual hierárquica que exibe dados em uma estrutura retangular, dividida em sub-retângulos coloridos ou áreas proporcionais à quantidade que representam. Esse tipo de gráfico é particularmente útil para visualizar a hierarquia e a distribuição de dados em categorias e subcategorias de maneira intuitiva.

O gráfico treemap é frequentemente usado quando se deseja representar a distribuição proporcional de dados em categorias e subcategorias de uma maneira visualmente eficaz. Ele é útil para mostrar a contribuição relativa de diferentes partes para o todo.



ANEXO 2



Estudo sem treino é igual a entretenimento.

Heitor Sasaki

MÓDULO EDA

EXERCÍCIOS

Iniciante

Distribuição de Idades

- Dataset: [Titanic](#) (kaggle)
- Tarefa:
 - Calcule a média, mediana, moda e desvio padrão da coluna de idade.
 - Pinte um histograma para visualizar a distribuição das idades.
 - Há outliers na distribuição de idades?

Contagem de Gêneros

- Dataset: [Titanic](#) (seaborn)
- Tarefa:
 - Calcule a quantidade de passageiros por gênero.
 - Faça um gráfico de barras para visualizar as contagens.
 - Existe alguma diferença no número de sobreviventes entre homens e mulheres?

Análise de Preço de Casas

- Dataset: [House Prices](#) (Kaggle)
- Tarefa:
 - Encontre os valores mínimo, máximo, médio e mediano da coluna de preços (SalePrice).
 - Pinte um boxplot da coluna de preços.
 - O preço das casas é simétrico ou assimétrico?

MÓDULO EDA

EXERCÍCIOS

Intermediário

Correlação entre Variáveis

- Dataset: [Iris](#) (seaborn/UCI)
- Tarefa:
 - Calcule a matriz de correlação entre as variáveis numéricas.
 - Visualize a correlação com um heatmap.
 - Quais variáveis apresentam a maior correlação?

Detecção de Outliers

- Dataset: [Tips](#) (seaborn/Kaggle)
- Tarefa:
 - Use boxplots para identificar outliers na coluna de gorjetas (tip).
 - Plote um scatterplot entre total_bill e tip para investigar a relação entre as variáveis.
 - Existe alguma tendência entre o valor total da conta e a gorjeta?

Distribuição de Notas Escolares

- Dataset: [Students Performance](#) (Kaggle)
- Tarefa:
 - Faça um histograma das notas de matemática (math score).
 - Separe a análise por gênero e plote gráficos comparativos.
 - Os homens ou as mulheres tiveram um desempenho superior?

MÓDULO EDA

EXERCÍCIOS

Avançado

Análises multivariadas e exploração avançada com insights

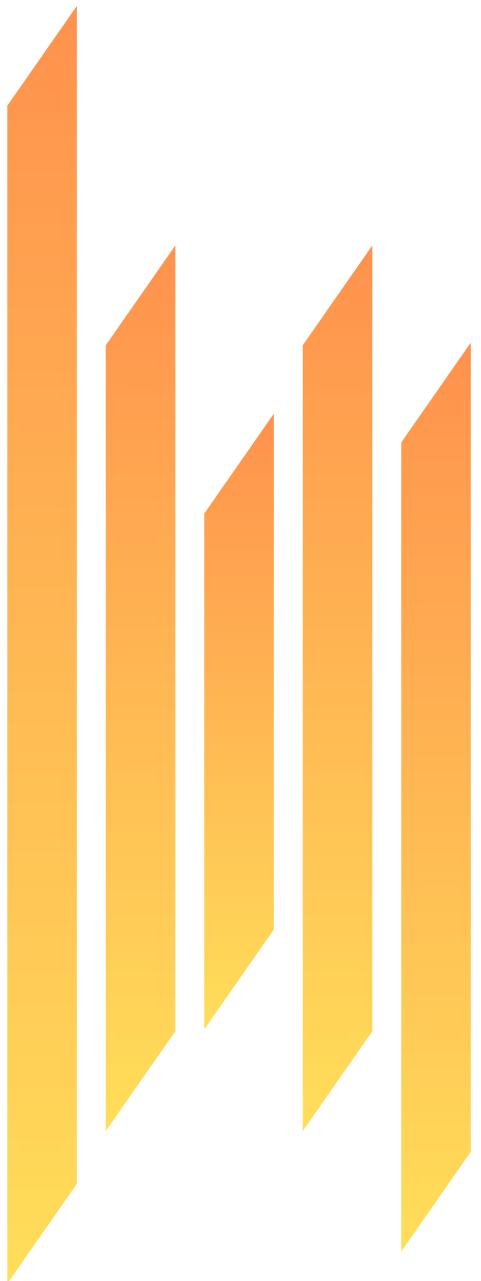
- Dataset: [Titanic](#) (seaborn/Kaggle)
- Tarefa:
 - Relacione a taxa de sobrevivência com classe do ticket, idade e gênero.
 - Utilize gráficos como boxplots, countplots e gráficos de barras empilhados.
 - Quais fatores tiveram maior impacto na sobrevivência?

Análise de Vendas e Desempenho

- Dataset: [Superstore Sales](#) (Kaggle)
- Tarefa:
 - Descubra os top 5 produtos que mais geraram lucro e os que mais deram prejuízo.
 - Faça um gráfico de tendência de vendas por mês.
 - Existe alguma sazonalidade nas vendas?

Análise de Dados de Airbnb

- Dataset: [Airbnb NYC Listings](#) (Kaggle)
- Tarefa:
 - Compare os preços de hospedagem por bairro.
 - Verifique a relação entre número de reviews e preço.
 - Quais bairros têm os preços médios mais altos? E os mais acessíveis?



ANEXO 3



É notável uma ciência que começou com jogos de azar tenha se tornado o mais importante objeto do conhecimento humano.

Pierre Simon Laplace

MÓDULO EDA

RESPOSTAS

Iniciante

Distribuição de Idades

- Dataset: [Titanic](#) (kaggle)
- Tarefa:
 - Calcule a média, mediana, moda e desvio padrão da coluna de idade.

```
## Calcule a média, mediana, moda e desvio padrão da coluna de idade.

import numpy as np
from statistics import mode

media_titanic_age = np.nanmean(titanic['Age'])
mediana_titanic_age = np.nanmedian(titanic['Age'])
moda_titanic_age = mode(titanic['Age'])

print(f'Média da Idade é {media_titanic_age:.2f}')
print(f'Mediana da Idade é {mediana_titanic_age}')
print(f'Moda da Idade é {moda_titanic_age}')
```

Ignorando os null

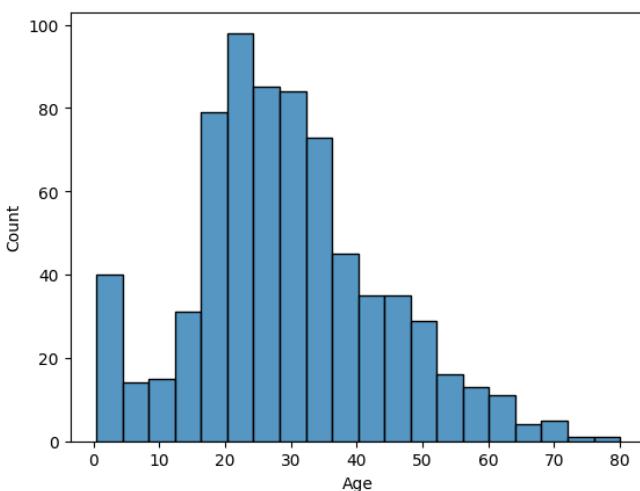
- Plote um histograma para visualizar a distribuição das idades.

```
## Pinte um histograma para visualizar a distribuição das idades.

import seaborn as sns

sns.histplot(titanic['Age'])
```

Aqui já respondeu o exercício.



MÓDULO EDA

RESPOSTAS

Iniciante

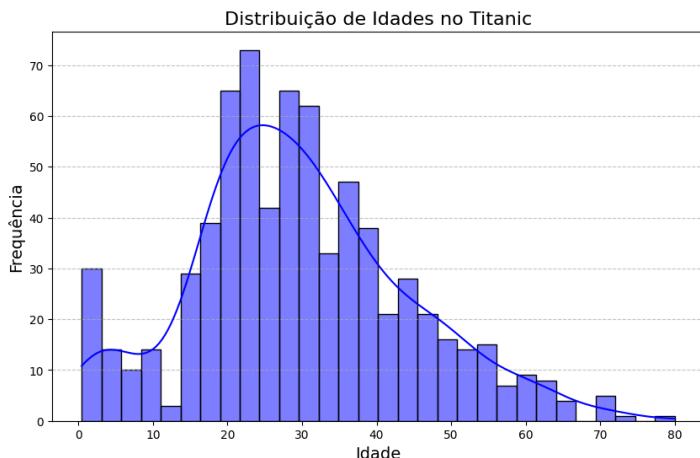
```
import matplotlib.pyplot as plt
import seaborn as sns

# Plotando o histograma com personalização
plt.figure(figsize=(10, 6)) # Define o tamanho do gráfico
sns.histplot(titanic['Age'], bins=30, kde=True, color='blue')

# Adicionando título e rótulos
plt.title('Distribuição de Idades no Titanic', fontsize=16)
plt.xlabel('Idade', fontsize=14)
plt.ylabel('Frequência', fontsize=14)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Mostrando o gráfico
plt.show()
```

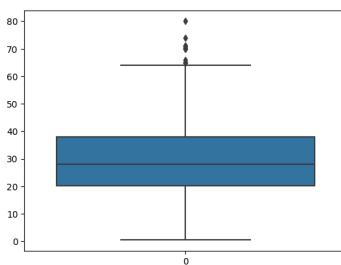
Mas podemos deixar mais elegante.



- Há outliers na distribuição de idades?

```
sns.boxplot(titanic['Age'])
```

Aqui já respondeu o exercício.



MÓDULO EDA

RESPOSTAS

Iniciante

Mas podemos apresentar uma resposta mais completa.

```
# Calcular Q1, Q3 e IQR
Q1 = titanic['Age'].quantile(0.25) # Primeiro quartil
Q3 = titanic['Age'].quantile(0.75) # Terceiro quartil
IQR = Q3 - Q1 # Intervalo interquartil

# Limites inferior e superior
limite_inferior = Q1 - 1.5 * IQR
limite_superior = Q3 + 1.5 * IQR

# Filtrar os outliers
outliers = titanic[(titanic['Age'] < limite_inferior) | (titanic['Age'] > limite_superior)]

print(f"Número de outliers: {len(outliers)}")
outliers.head()
```

Quais os limites, IQR, Quantos Outliers tem e quem são.

Número de outliers: 11												
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
33	34	0	2 Whedon, Mr. Edward H	male	66.0	0	0	CA 24579	10.5000	NaN	S	
54	55	0	1 Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.9792	B30	C	
96	97	0	1 Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C	
116	117	0	3 Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	NaN	Q	
280	281	0	3 Duane, Mr. Frank	male	65.0	0	0	336439	7.7500	NaN	Q	

MÓDULO EDA

RESPOSTAS

Iniciante

Contagem de Gêneros

- Dataset: [Titanic](#) (seaborn)
- Tarefa:
 - Calcule a quantidade de passageiros por gênero.

```
✓ titanic['Sex'].value_counts()
   Sex
   male      577
   female    314
   Name: count, dtype: int64
```

- Faça um gráfico de barras para visualizar as contagens.

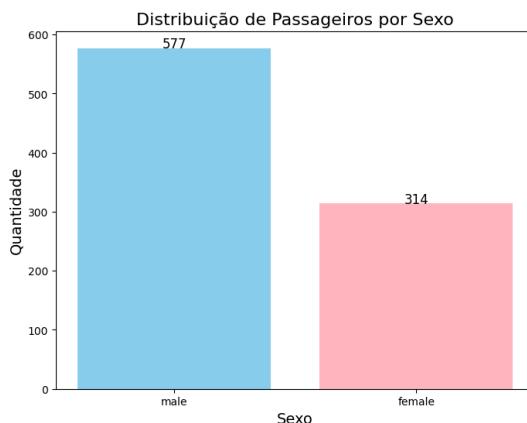
```
# Contar o número de passageiros por sexo
sex_counts = titanic['Sex'].value_counts()

# Criar o gráfico de barras
plt.figure(figsize=(8, 6))
plt.bar(sex_counts.index, sex_counts.values, color=['skyblue', 'lightpink'])

# Adicionar rótulos e título
plt.title('Distribuição de Passageiros por Sexo', fontsize=16)
plt.xlabel('Sexo', fontsize=14)
plt.ylabel('Quantidade', fontsize=14)

# Adicionar os valores acima das barras
for i, count in enumerate(sex_counts.values):
    plt.text(i, count + 10, str(count), ha='center', fontsize=12)

plt.show()
```



MÓDULO EDA

RESPOSTAS

Iniciante

Contagem de Gêneros

- Existe alguma diferença no número de sobreviventes entre homens e mulheres?

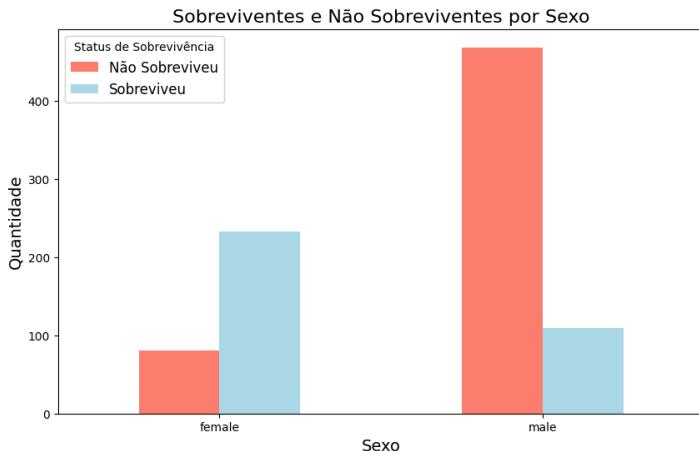
```
survival_counts = titanic.groupby(['Sex', 'Survived']).size().unstack()

# Criar o gráfico de colunas agrupadas
survival_counts.plot(kind='bar', figsize=(10, 6), color=['salmon', 'lightblue'])

# Adicionar título e rótulos
plt.title('Sobreviventes e Não Sobreviventes por Sexo', fontsize=16)
plt.xlabel('Sexo', fontsize=14)
plt.ylabel('Quantidade', fontsize=14)
plt.legend(['Não Sobreviveu', 'Sobreviveu'], title='Status de Sobrevivência', fontsize=12)

# Mostrar o gráfico
plt.xticks(rotation=0)
plt.show()
```

Para ratificar essa diferença podemos evoluir para um teste de hipótese



MÓDULO EDA

RESPOSTAS

Iniciante

Análise de Preço de Casas

- Dataset: [House Prices](#) (Kaggle)
- Tarefa:
 - Encontre os valores mínimo, máximo, médio e mediano da coluna de preços (SalePrice).

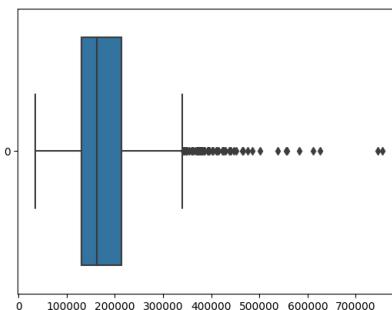
```
max = np.max(house_prices['SalePrice'])
min = np.min(house_prices['SalePrice'])
media = np.average(house_prices['SalePrice'])

print(f'A casa com o maior valor é R$ {max} e com menor valor R$ {min}, enquanto que o valor médio é R$ {media:.2f}')
```

A casa com o maior valor é R\$ 755000 e com menor valor R\$ 34900, enquanto que o valor médio é R\$ 180921.20

- Plote um boxplot da coluna de preços.

```
sns.boxplot(house_prices['SalePrice'], orient='h')
```



- O preço das casas é simétrico ou assimétrico?

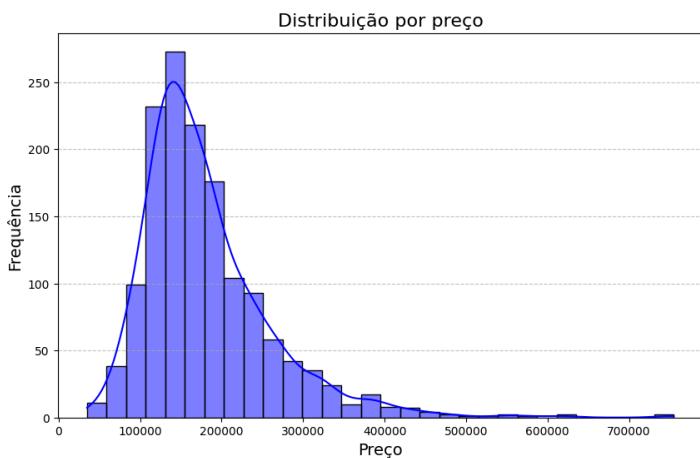
```
# Plotando o histograma com personalização
plt.figure(figsize=(10, 6)) # Define o tamanho do gráfico
sns.histplot(house_prices['SalePrice'], bins=30, kde=True, color='blue')

# Adicionando título e rótulos
plt.title('Distribuição por preço', fontsize=16)
plt.xlabel('Preço', fontsize=14)
plt.ylabel('Frequência', fontsize=14)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Mostrando o gráfico
plt.show()
```

MÓDULO EDA

RESPOSTAS



Percebemos que é um gráfico assimétrico a direita, mas você pode fazer o cálculo e comprovar isso.

MÓDULO EDA

RESPOSTAS

Intermediário

Correlação entre Variáveis

- Dataset: [Iris](#) (seaborn/UCI)
- Tarefa:
 - Calcule a matriz de correlação entre as variáveis numéricas.

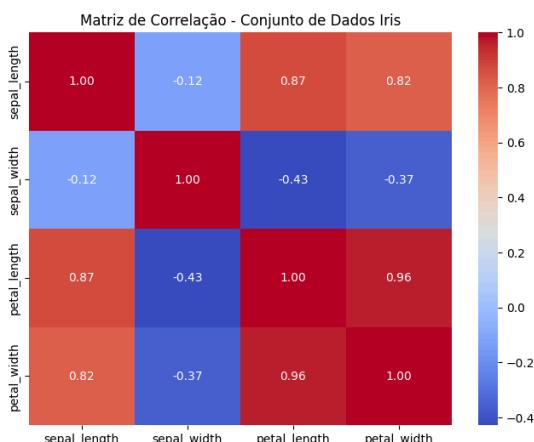
```
iris = sns.load_dataset('iris')
iris.head()
iris.drop(columns= ['species'], inplace=True)
correlation_matrix = iris.corr()
print(correlation_matrix)
```

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.117570	0.871754	0.817941
sepal_width	-0.117570	1.000000	-0.428440	-0.366126
petal_length	0.871754	-0.428440	1.000000	0.962865
petal_width	0.817941	-0.366126	0.962865	1.000000

Questão resolvida, mas podemos deixar mais palatável.

- Visualize a correlação com um heatmap.

```
correlation_matrix = iris.corr()
# Plotar a matriz de correlação como um mapa de calor
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', cbar=True)
plt.title('Matriz de Correlação - Conjunto de Dados Iris')
plt.show()
```



MÓDULO EDA

RESPOSTAS

Intermediário

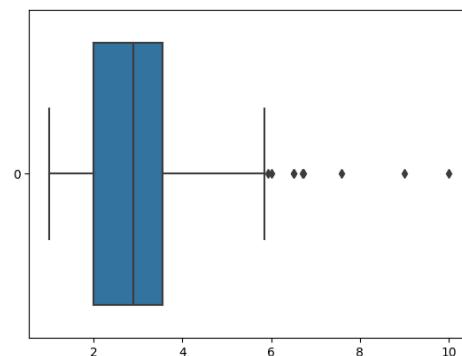
- Quais variáveis apresentam a maior correlação?

Não precisa de código aqui e você pode inferir graficamente.

Detecção de Outliers

- Dataset: [Tips](#)(seaborn/Kaggle)
- Tarefa:
 - Use boxplots para identificar outliers na coluna de gorjetas (tip).

```
#%%  
  
import pandas as pd  
  
tips = pd.read_csv('tips.csv')  
tips.head()
```



```
import seaborn as sns  
  
sns.boxplot(tips['tip'], orient='h')  
  
Run Cell | Run Above | Debug Cell  
vNS  
# Calcular Q1, Q3 e IQR  
Q1 = tips['tip'].quantile(0.25) # Primeiro quartil  
Q3 = tips['tip'].quantile(0.75) # Terceiro quartil  
IQR = Q3 - Q1 # Intervalo interquartil  
  
# Limite inferior e superior  
limite_inferior = Q1 - 1.5 * IQR  
limite_superior = Q3 + 1.5 * IQR  
  
# Filtrar os outliers  
outliers = tips[(tips['tip'] < limite_inferior) | (tips['tip'] > limite_superior)]  
  
print(f"Número de outliers: {len(outliers)}")  
outliers.head()
```

Mesmo código de um exercício que já fizemos.

	total_bill	tip	sex	smoker	day	time	size
23	39.42	7.58	Male	No	Sat	Dinner	4
47	32.40	6.00	Male	No	Sun	Dinner	4
59	48.27	6.73	Male	No	Sat	Dinner	4
141	34.30	6.70	Male	No	Thur	Lunch	6
170	50.81	10.00	Male	Yes	Sat	Dinner	3

MÓDULO EDA

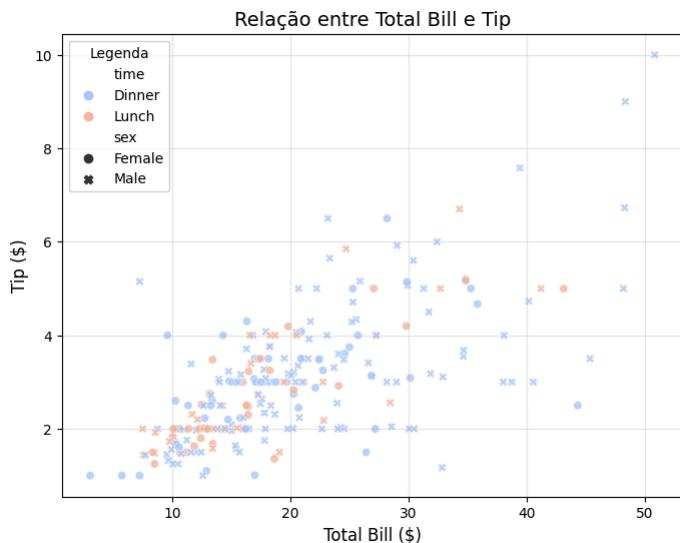
RESPOSTAS

Intermediário

- Pinte um scatterplot entre total_bill e tip para investigar a relação entre as variáveis.

```
import matplotlib.pyplot as plt

# Criar um scatterplot entre total_bill e tip
plt.figure(figsize=(6, 6))
sns.set(style='white', palette='coolwarm', alpha=0.8)
sns.lmplot(x='total_bill', y='tip', hue='time', style='sext', palette='coolwarm', alpha=0.8)
plt.title('Relação entre Total Bill e Tip', fontsize=14)
plt.xlabel('Total Bill ($)', fontsize=12)
plt.ylabel('Tip ($)', fontsize=12)
plt.legend('time', fontsize=10)
plt.grid(alpha=0.3)
plt.show()
```



- Existe alguma tendência entre o valor total da conta e a gorjeta?

[Interprete o gráfico](#)

MÓDULO EDA

RESPOSTAS

Intermediário

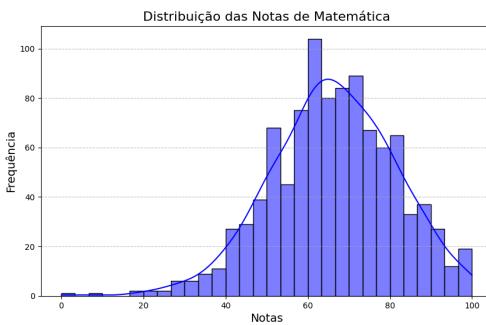
Distribuição de Notas Escolares

- Dataset: [Students Performance](#) (Kaggle)
- Tarefa:
 - Faça um histograma das notas de matemática (math score).

```
# Plotando o histograma com personalização
plt.figure(figsize=(10, 6)) # Define o tamanho do gráfico
sns.histplot(students['math score'], bins=30, kde=True, color='blue')

# Adicionando título e rótulos
plt.title('Distribuição das Notas de Matemática', fontsize=16)
plt.xlabel('Notas', fontsize=14)
plt.ylabel('Frequência', fontsize=14)
plt.grid(axes='y', linestyle='--', alpha=0.7)

# Mostrando o gráfico
plt.show()
```



- Separe a análise por gênero e plote gráficos comparativos.

```
fig, axes = plt.subplots(1, 3, figsize=(18, 6), sharey=True)

# Math Score
sns.boxplot(data=students, x='gender', y='math score', ax=axes[0], palette='Set2')
axes[0].set_title('Distribuição de Math Scores por Gênero')
axes[0].set_xlabel('Gênero')
axes[0].set_ylabel('Math Score')

# Reading Score
sns.boxplot(data=students, x='gender', y='reading score', ax=axes[1], palette='Set2')
axes[1].set_title('Distribuição de Reading Scores por Gênero')
axes[1].set_xlabel('Gênero')
axes[1].set_ylabel('Reading Score')

# Writing Score
sns.boxplot(data=students, x='gender', y='writing score', ax=axes[2], palette='Set2')
axes[2].set_title('Distribuição de Writing Scores por Gênero')
axes[2].set_xlabel('Gênero')
axes[2].set_ylabel('Writing Score')

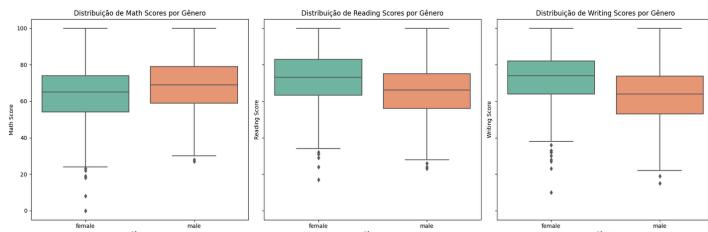
plt.tight_layout()
plt.show()
```

MÓDULO EDA

RESPOSTAS

Intermediário

Distribuição de Notas Escolares



- Os homens ou as mulheres tiveram um desempenho superior?

```
import pandas as pd
from scipy.stats import ttest_ind

# Carregar o dataset
df = pd.read_csv("StudentsPerformance.csv")

# Separar os dados por gênero
homens = df[df['gender'] == 'male']
mulheres = df[df['gender'] == 'female']

# Calcular as médias
media_homens = homens[['math score', 'reading score', 'writing score']].mean()
media_mulheres = mulheres[['math score', 'reading score', 'writing score']].mean()

print("Médias - Homens:")
print(media_homens)
print("\nMédias - Mulheres:")
print(media_mulheres)

# Teste T para cada tipo de nota
resultados = {}
for score in ['math score', 'reading score', 'writing score']:
    stat, p_value = ttest_ind(homens[score], mulheres[score], equal_var=False)
    resultados[score] = {'t-statistic': stat, 'p-value': p_value}

print("\nResultados do Teste T:")
for score, res in resultados.items():
    print(f"({score}): t-statistic = {res['t-statistic']:.2f}, p-value = {res['p-value']:.4f}")
```

```
Médias - Homens:
math score      68.728216
reading score   65.473029
writing score   63.311203
dtype: float64

Médias - Mulheres:
math score      63.633205
reading score   72.608108
writing score   72.467181
dtype: float64

Resultados do Teste T:
math score: t-statistic = 5.40, p-value = 0.0000
reading score: t-statistic = -7.97, p-value = 0.0000
writing score: t-statistic = -10.00, p-value = 0.0000
```

Existe?

MÓDULO EDA

EXERCÍCIOS

Avançado

Análises multivariadas e exploração avançada com insights

- Dataset: [Titanic](#) (seaborn/Kaggle)
- Tarefa:
 - Relacione a taxa de sobrevivência com classe do ticket, idade e gênero.
 - Utilize gráficos como boxplots, countplots e gráficos de barras empilhados.
 - Quais fatores tiveram maior impacto na sobrevivência?

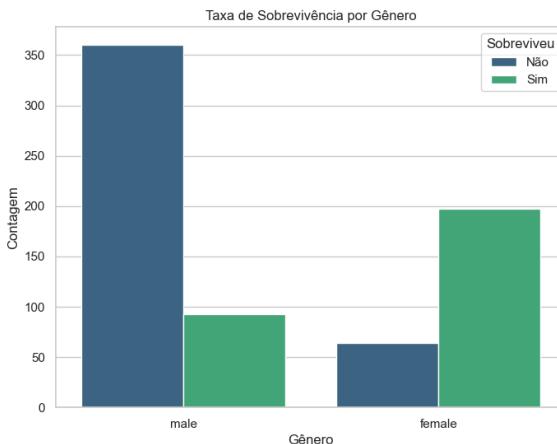
```
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

# Carregar dataset Titanic do Seaborn
df = sns.load_dataset("titanic")

# Remover entradas com valores nulos para simplicidade
df = df.dropna(subset=['age', 'pclass', 'sex', 'survived'])

# Configuração de estilo
sns.set_theme(style="whitegrid")

# 1. Taxa de sobrevivência por gênero
plt.figure(figsize=(8, 6))
sns.countplot(data=df, x="sex", hue="survived", palette="viridis")
plt.title("Taxa de Sobrevivência por Gênero")
plt.xlabel("Gênero")
plt.ylabel("Contagem")
plt.legend(title="Sobreviveu", labels=["Não", "Sim"])
plt.show()
```



MÓDULO EDA

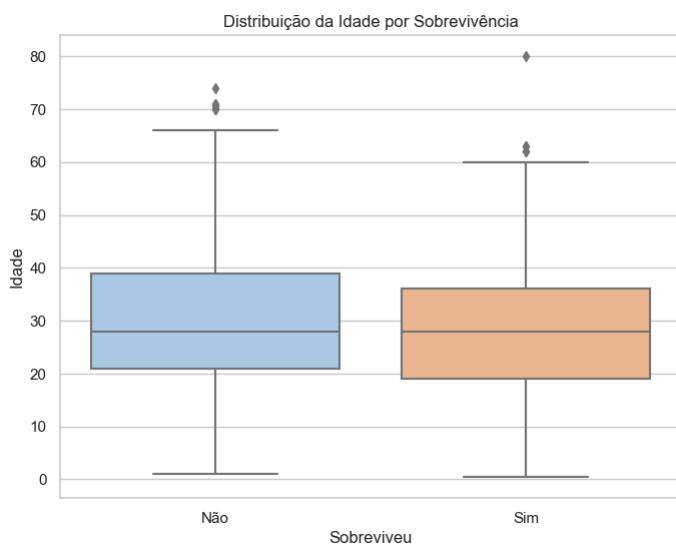
EXERCÍCIOS

Avançado

Análises multivariadas e exploração avançada com insights

- Dataset: [Titanic](#) (seaborn/Kaggle)
- Tarefa:
 - Relacione a taxa de sobrevivência com classe do ticket, idade e gênero.
 - Utilize gráficos como boxplots, countplots e gráficos de barras empilhados.
 - Quais fatores tiveram maior impacto na sobrevivência?

```
# 2. Boxplot: Idade e Sobrevidência
plt.figure(figsize=(8, 6))
sns.boxplot(data=df, x="survived", y="age", palette="pastel")
plt.title("Distribuição da Idade por Sobrevidência")
plt.xlabel("Sobrevideu")
plt.ylabel("Idade")
plt.xticks([0, 1], ["Não", "Sim"])
plt.show()
```



MÓDULO EDA

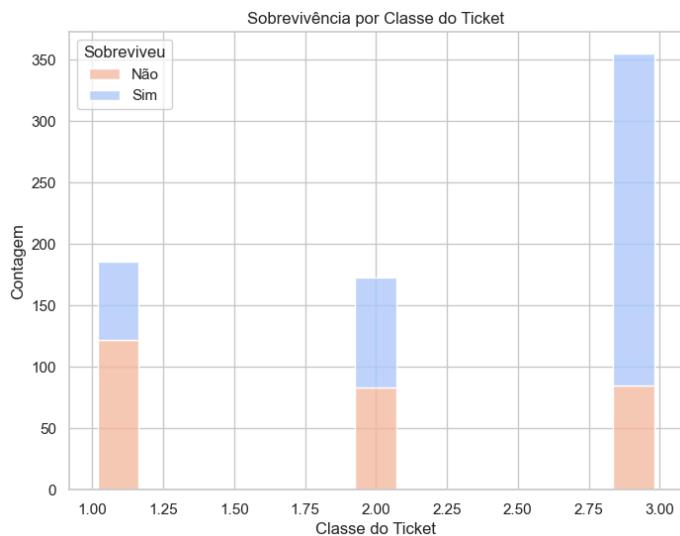
EXERCÍCIOS

Avançado

Análises multivariadas e exploração avançada com insights

- Dataset: [Titanic](#) (seaborn/Kaggle)
- Tarefa:
 - Relacione a taxa de sobrevivência com classe do ticket, idade e gênero.
 - Utilize gráficos como boxplots, countplots e gráficos de barras empilhados.
 - Quais fatores tiveram maior impacto na sobrevivência?

```
# 3. Gráfico de Barras Empilhado: Classe do Ticket e Sobrevidência
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x="pclass", hue="survived", multiple="stack", palette="coolwarm", shrink=0.8)
plt.title("Sobrevidência por Classe do Ticket")
plt.xlabel("Classe do Ticket")
plt.ylabel("Contagem")
plt.legend(title="Sobreviveu", labels=["Não", "Sim"])
plt.show()
```



MÓDULO EDA

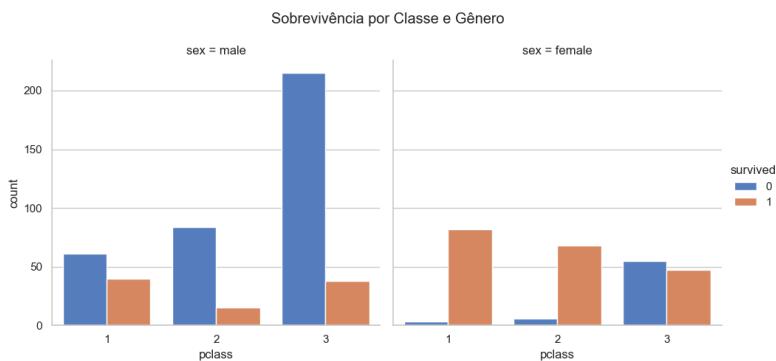
EXERCÍCIOS

Avançado

Análises multivariadas e exploração avançada com insights

- Dataset: [Titanic](#) (seaborn/Kaggle)
- Tarefa:
 - Relacione a taxa de sobrevivência com classe do ticket, idade e gênero.
 - Utilize gráficos como boxplots, countplots e gráficos de barras empilhados.
 - Quais fatores tiveram maior impacto na sobrevivência?

```
# 4. Taxa de sobrevivência por gênero e classe do ticket
plt.figure(figsize=(10, 6))
sns.catplot(data=df, x="pclass", hue="survived", col="sex", kind="count", height=5, aspect=1, palette="muted")
plt.subplots_adjust(top=0.85)
plt.suptitle("Sobrevivência por Classe e Gênero")
plt.show()
```



MÓDULO EDA

EXERCÍCIOS

Avançado

Análise de Vendas e Desempenho

- Dataset: [Superstore Sales](#) (Kaggle)
- Tarefa:
 - Descubra os top 5 produtos que mais geraram lucro e os que mais deram prejuízo.
 - Faça um gráfico de tendência de vendas por mês.
 - Existe alguma sazonalidade nas vendas?

```
#%%

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Carregar o dataset
df = pd.read_csv("train.csv")

df.head()

# Converter a coluna Order Date para datetime
df['Order Date'] = pd.to_datetime(df['Order Date'], dayfirst=True)

# Criar colunas para ano e mês
df['YearMonth'] = df['Order Date'].dt.to_period('M')

# 1. Top 5 Produtos por Lucro
produto_lucro = df.groupby('Product Name')['Sales'].sum().sort_values(ascending=False)
top5_lucro = produto_lucro.head(5)
top5_prejuizo = produto_lucro.tail(5)

print("Top 5 Produtos com Maior Lucro:")
print(top5_lucro)

print("\nTop 5 Produtos com Maior Prejuízo:")
print(top5_prejuizo)
```

```
Top 5 Produtos com Maior Lucro:
Product Name
Canon imageCLASS 2200 Advanced Copier           61599.824
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind 27453.384
Cisco TelePresence System EX90 Videoconferencing Unit 22638.480
HON 5400 Series Task Chairs for Big and Tall    21870.576
GBC DocuBind TL300 Electric Binding System      19823.479
Name: Sales, dtype: float64

Top 5 Produtos com Maior Prejuízo:
Product Name
Acme Serrated Blade Letter Opener             7.632
Grip Seal Envelopes                          7.072
Xerox 20                                     6.480
Avery 5                                      5.760
Eureka Disposable Bags for Sanitaire Vibra Groomer I Upright Vac 1.624
Name: Sales, dtype: float64
```

MÓDULO EDA

EXERCÍCIOS

Avançado

Análise de Vendas e Desempenho

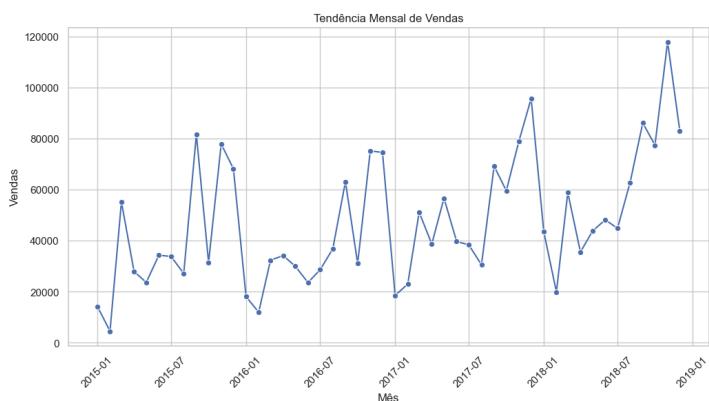
- Dataset: [Superstore Sales](#) (Kaggle)
- Tarefa:
 - Descubra os top 5 produtos que mais geraram lucro e os que mais deram prejuízo.
 - Faça um gráfico de tendência de vendas por mês.
 - Existe alguma sazonalidade nas vendas?

```
# 2. Gráfico de Tendência Mensal de Vendas

# Agrupe por 'YearMonth' e calcule as vendas mensais
vendas_mensais = df.groupby('YearMonth')[['Sales']].sum().reset_index()

# Certifique-se de que o eixo X é categórico (ordenado por tempo)
vendas_mensais['YearMonth'] = pd.to_datetime(vendas_mensais['YearMonth'])

# Pinte o gráfico de linha
plt.figure(figsize=(12, 6))
sns.lineplot(data=vendas_mensais, x='YearMonth', y='Sales', marker='o')
plt.title("Tendência Mensal de Vendas")
plt.xlabel("Mês")
plt.ylabel("Vendas")
plt.xticks(rotation=45)
plt.show()
```



MÓDULO EDA

EXERCÍCIOS

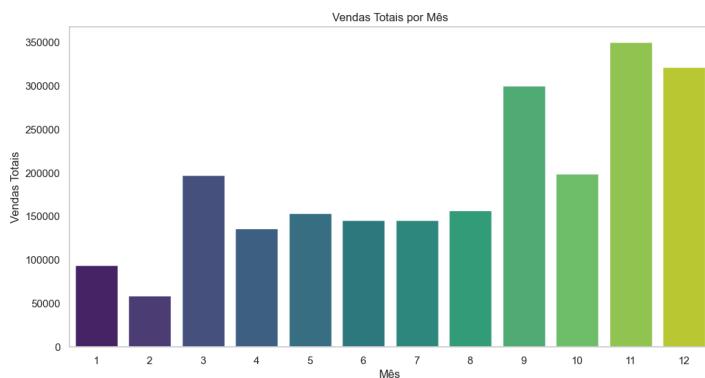
Avançado

Análise de Vendas e Desempenho

- Dataset: [Superstore Sales](#) (Kaggle)
- Tarefa:
 - Descubra os top 5 produtos que mais geraram lucro e os que mais deram prejuízo.
 - Faça um gráfico de tendência de vendas por mês.
 - Existe alguma sazonalidade nas vendas?

```
# 3. Investigação de Sazonalidade
df['Month'] = df['Order Date'].dt.month
sazonalidade = df.groupby('Month')['Sales'].sum()

plt.figure(figsize=(12, 6))
sns.barplot(x=sazonalidade.index, y=sazonalidade.values, palette="viridis")
plt.title("Vendas Totais por Mês")
plt.xlabel("Mês")
plt.ylabel("Vendas Totais")
plt.grid(axis='y')
plt.show()
```



MÓDULO EDA

EXERCÍCIOS

Avançado

Análise de Dados de Airbnb

- Dataset: [Airbnb NYC Listings](#) (Kaggle)
- Tarefa:
 - Compare os preços de hospedagem por bairro.
 - Verifique a relação entre número de reviews e preço.
 - Quais bairros têm os preços médios mais altos? E os mais acessíveis?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Verifique se as colunas relevantes estão disponíveis
df = pd.read_csv('AB_NYC_2019.csv')

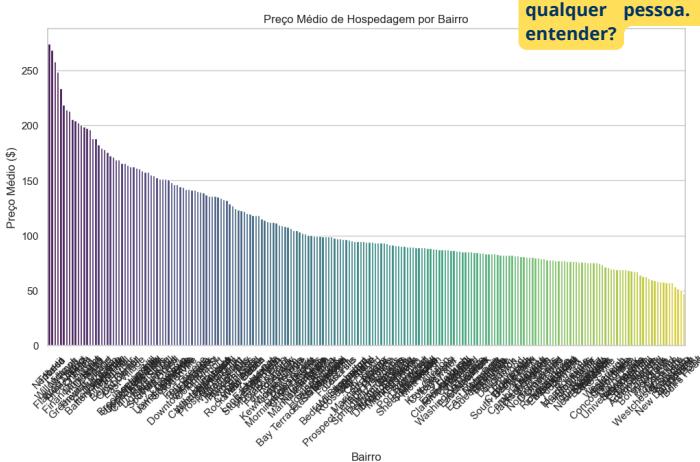
# Garantir que as colunas 'price' e 'number_of_reviews' são numéricas
df['price'] = pd.to_numeric(df['price'], errors='coerce')
df['number_of_reviews'] = pd.to_numeric(df['number_of_reviews'], errors='coerce')

# Remova outliers (opcional, para melhorar a visualização)
df = df[df['price'] <= 500]

# 1. Comparar os preços médios por bairro
preco_bairros = df.groupby('neighbourhood')['price'].mean().sort_values(ascending=False)

# Plotar os preços médios por bairro
plt.figure(figsize=(12, 6))
sns.barplot(x=preco_bairros.index, y=preco_bairros.values, palette="viridis")
plt.title("Preço Médio de Hospedagem por Bairro")
plt.ylabel("Preço Médio ($)")
plt.xlabel("Bairro")
plt.xticks(rotation=45)
plt.show()
```

Respondemos a questão, mas imagine entregar isso para qualquer pessoa. Você consegue entender?



MÓDULO EDA

EXERCÍCIOS

Avançado

Análise de Dados de Airbnb

- Dataset: [Airbnb NYC Listings](#) (Kaggle)
- Tarefa:
 - Compare os preços de hospedagem por bairro.
 - Verifique a relação entre número de reviews e preço.
 - Quais bairros têm os preços médios mais altos? E os mais acessíveis?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Verifique se as colunas relevantes estão disponíveis
print(df.columns)

# Garantir que as colunas 'price' são numéricas
df['price'] = pd.to_numeric(df['price'], errors='coerce')

# Remova outliers (opcional, para melhorar a visualização)
df = df[df['price'] <= 500]

# Agrupamento por bairro
preco_bairros = df.groupby('neighbourhood')['price'].mean().sort_values(ascending=False)

# 5 bairros mais caros
mais_caros = preco_bairros.head(5)

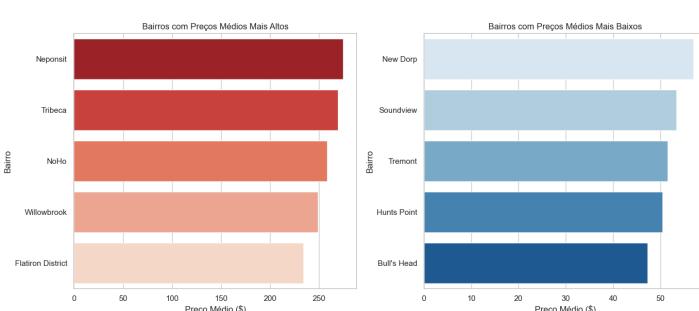
# 5 bairros mais baratos
mais_baratos = preco_bairros.tail(5)

# Criar gráficos lado a lado
plt.figure(figsize=(14, 6))

# Gráfico dos 5 bairros mais caros
plt.subplot(1, 2, 1)
sns.barplot(x=mais_caros.values, y=mais_caros.index, palette="Reds_r")
plt.title("Bairros com Preços Médios Mais Altos")
plt.xlabel("Preço Médio ($)")
plt.ylabel("Bairro")

# Gráfico dos 5 bairros mais baratos
plt.subplot(1, 2, 2)
sns.barplot(x=mais_baratos.values, y=mais_baratos.index, palette="Blues")
plt.title("Bairros com Preços Médios Mais Baixos")
plt.xlabel("Preço Médio ($)")
plt.ylabel("Bairro")

plt.tight_layout()
plt.show()
```



MÓDULO EDA

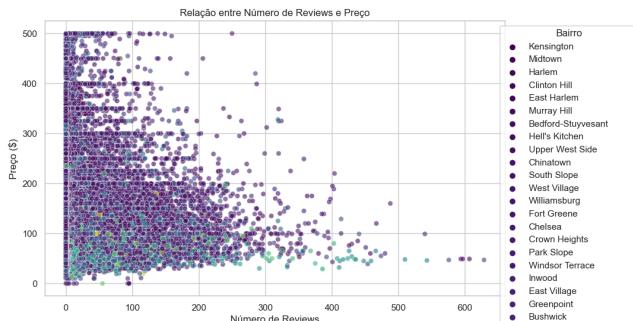
EXERCÍCIOS

Avançado

Análise de Dados de Airbnb

- Dataset: [Airbnb NYC Listings](#) (Kaggle)
- Tarefa:
 - Compare os preços de hospedagem por bairro.
 - Verifique a relação entre número de reviews e preço.
 - Quais bairros têm os preços médios mais altos? E os mais acessíveis?

```
# 2. Relação entre número de reviews e preço
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x="number_of_reviews", y="price", alpha=0.6, hue='neighbourhood', palette="viridis")
plt.title("Relação entre Número de Reviews e Preço")
plt.xlabel("Número de Reviews")
plt.ylabel("Preço ($)")
plt.legend(loc='upper right', title="Bairro", bbox_to_anchor=(1.3, 1))
plt.show()
```



Essa legenda infinita, vamos precisar de uma folha A0

MÓDULO EDA

EXERCÍCIOS

Avançado

Análise de Dados de Airbnb

- Dataset: [Airbnb NYC Listings](#) (Kaggle)
- Tarefa:
 - Compare os preços de hospedagem por bairro.
 - Verifique a relação entre número de reviews e preço.
 - Quais bairros têm os preços médios mais altos? E os mais acessíveis?

```
from scipy.stats import pearsonr
import seaborn as sns
import matplotlib.pyplot as plt

# Garantir que as colunas relevantes sejam numéricas
df['price'] = pd.to_numeric(df['price'], errors='coerce')
df['number_of_reviews'] = pd.to_numeric(df['number_of_reviews'], errors='coerce')

# Removendo valores nulos
df_corr = df[['price', 'number_of_reviews']].dropna()

# Calculando a correlação de Pearson
correlation, p_value = pearsonr(df_corr['price'], df_corr['number_of_reviews'])

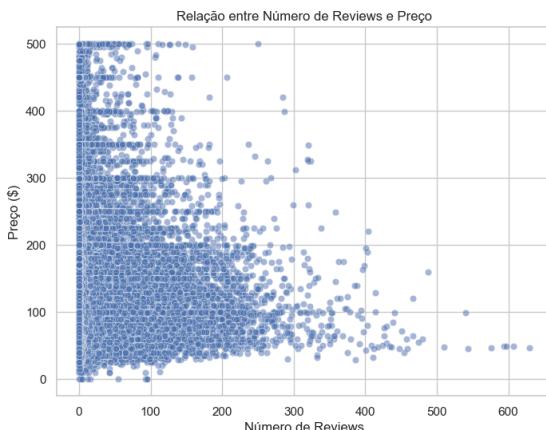
print(f"Coefficiente de Correlação (Pearson): {correlation:.2f}")
print(f"P-valor: {p_value:.2e}")

# Interpretando os resultados
if p_value < 0.05:
    print("Há uma relação estatisticamente significativa entre preço e número de reviews.")
else:
    print("Não há uma relação estatisticamente significativa entre preço e número de reviews.")

# Visualizando a relação com um scatter plot
plt.figure(figsize=(8, 6))
sns.scatterplot(x='number_of_reviews', y='price', data=df_corr, alpha=0.5)
plt.title("Relação entre Número de Reviews e Preço")
plt.xlabel("Número de Reviews")
plt.ylabel("Preço ($)")
plt.show()
```

Correlação significativa e correlação alta são diferentes, interprete a saída.

Coefficiente de Correlação (Pearson): -0.05
P-valor: 4.02e-26
Há uma relação estatisticamente significativa entre preço e número de reviews.





E assim, chegamos ao final desta incrível jornada pelo universo fascinante do Power BI. Espero que esta leitura tenha sido não apenas educativa, mas também uma explosão de criatividade e insights para você!

Espero que este ebook tenha sido mais do que uma fonte de aprendizado. Que ele tenha despertado em você a curiosidade, a criatividade e a vontade de descobrir os segredos que os dados têm a revelar.

Antes de nos despedirmos, quero compartilhar algumas “boas práticas extras” para você dominar a EDA com um olhar analítico, mas também cheio de estilo e personalidade:

- Explore os dados como um detetive
- Transforme números em narrativas visuais
- Cuide da limpeza dos dados
- Deixe os números falarem, mas faça as perguntas certas
- Compartilhe insights com clareza e propósito

Que cada análise seja uma tela em branco para sua criatividade fluir, e que cada relatório seja uma narrativa envolvente. Continue explorando, continue aprendendo e, acima de tudo, continue se divertindo no emocionante mundo do Python!

Agradeço por ter me acompanhado nesta jornada, e que seus futuros relatórios sejam brilhantes.

Até a próxima aventura analítica!



@Data_Bruno

@Data_Bruno