Preparação e Análise Exploratória de Dados

ALMEIDA
DATA SCIENCE

Objetivos

- Compreender e aplicar técnicas de preparação de dados
- Difundir a cultura estatística com técnicas descritivas para capacitar o aluno na implementação de técnicas estatísticas

O que você irá aprender ?

1. Conceitos e definições da estatística para preparação de dados
2. Análise Exploratória de Dados: resumo de dados, medidas de dados, análise bivariada, visualização de dados, distribuições de probabilidades, introdução à inferência estatística
3. Pré-processamento de dados: limpeza e transformação de dados, engenharia e seleção de variáveis

Referências Bibliográficas

- BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017. 554 p., il. Inclui índice remissivo e tabelas. ISBN 978-85-472-2022-8.
- BUSSAB, Wilton de O. (coord.). **Elementos de Amostragem**. São Paulo: E. Blucher, 2005. 274 p., : il. ISBN 85-212-0367-5.
- **Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais**, 1ª edição. Andrew Bruce, Peter Bruce. 2019.
- Hastie, T.; Tibshirani, R; e Friedman, J. (2011). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Segunda Edição, Springer Verlag, Nova Iorque.
- James, G.; Witten, D.; Hastie, T.; e Tibshirani, R (2013). **An Introduction to Statistical Learning: with Applications in R** Springer Verlag, Nova Iorque.

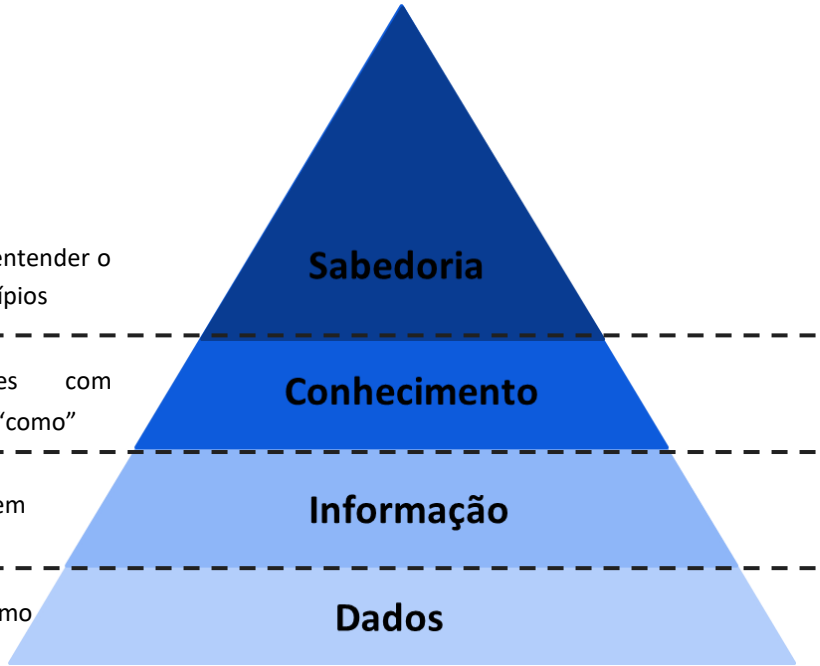
Preparação e Análise Exploratória de Dados

"Dados são o novo petróleo. Mas petróleo bruto não tem valor sem refinamento!”

"Se você não entende seus dados, seu modelo também não vai entender!"

“O melhor modelo do mundo não compensa dados ruins”

Pirâmide dos Dados



Aplicação do conhecimento para entender o "por que", gerando insights e princípios

Combinação de informações com experiência e regras, explicando o "como"

Dados organizados que respondem ao "o que" em um contexto

Elementos brutos coletados, como números, palavras e códigos

Tipos de
Análise de Dados

1 Análise Descritiva

Resume e organiza dados históricos para identificar padrões e tendências

O que aconteceu ?

2 Análise Diagnóstica

Investiga as causas por trás dos padrões identificados

Como aconteceu ?

3 Análise Preditiva

Usa modelos estatísticos e aprendizado de máquina para prever eventos futuros

O que irá acontecer ?

4 Análise Prescritiva

Sugere ações para otimizar resultados com base em simulações e otimização

O que fazer ?

Conceitos estatísticos essenciais para uma Excelente Análise Exploratória de Dados (EDA)

1

Tendência Central

Média, Mediana e Moda ajudam a encontrar o "centro" dos dados

2

Dispersão

Desvio Padrão e Variância indicam o grau de dispersão dos valores

3

Impacto de Valores Extremos

Identificar e tratar outliers é essencial para evitar distorções nas análises

4

Correlação vs. Causalidade

Correlação não implica causalidade; duas variáveis podem estar associadas sem que uma seja a causa da outra

5

Visualização de Dados

Gráficos como histogramas, box plots e dispersões ajudam a interpretar informações de forma clara e objetiva

6

Distribuições de Probabilidade

Normal, Binomial, Poisson são fundamentais para modelar diferentes comportamentos de dados

7

Descritiva vs. Inferencial

Saiba quando resumir os dados e quando fazer previsões

8

Técnicas de Amostragem

Amostragem Aleatória, Estratificada e por Cluster impactam a precisão dos dados

9

Estatística Inferencial

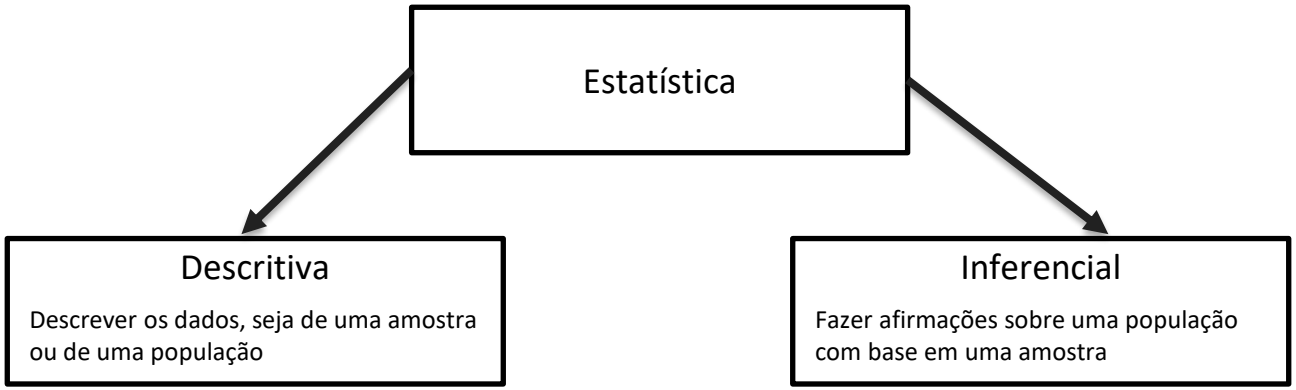
Testes de hipótese e intervalos de confiança orientam decisões

ESTATÍSTICA
DEFINIÇÃO

• **Ciência** que tem por objetivo a coleta, redução, análise e modelagem dos dados [BUSSAB, Wilton de O.; MORETTIN, Pedro A]

• **Conjunto de métodos** adequados para a coleta, organização, resumo e descrição, análise e interpretação de dados observacionais, visando a compreensão de uma realidade específica para embasar a tomada de decisões

Conceitos de Estatística



Tipo de Estrutura de Dados

1

Dados organizados em um formato fixo, geralmente em tabelas com linhas e colunas, armazenados em bancos de dados relacionais

Exemplos:

Tabelas de Excel com colunas fixas

Bancos de dados relacionais (MySQL, PostgreSQL, SQL Server

Dados Estruturados

2

Dados que possuem alguma organização, mas não seguem um modelo rígido como os estruturados. Geralmente, são armazenados em formatos flexíveis e hierárquicos

Exemplos:

páginas da web, documentos XML e arquivos JSON

Dados Semi - Estruturados

3

Dados que não possuem uma organização fixa e não seguem um esquema definido

Exemplos:

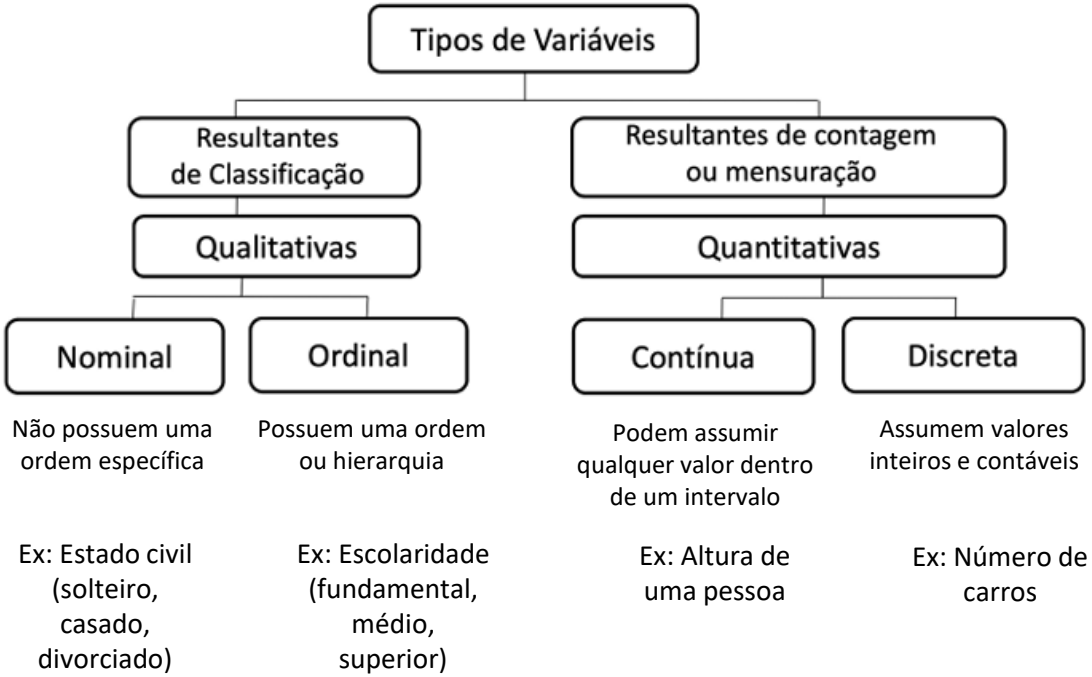
Imagens e vídeos (ex: fotos de veículos na sua base de modelagem)

Áudios e músicas (ex: gravações de call centers para análise de sentimento)

Dados Não - Estruturados

O que é e quais são os Tipos de Variáveis

Variáveis podem ser entendidas como uma **característica** de dada população ou amostra, que podem ser **medidas ou contadas**, e que **variam** entre indivíduos de uma população



O que é e quais são os Tipos de Variáveis

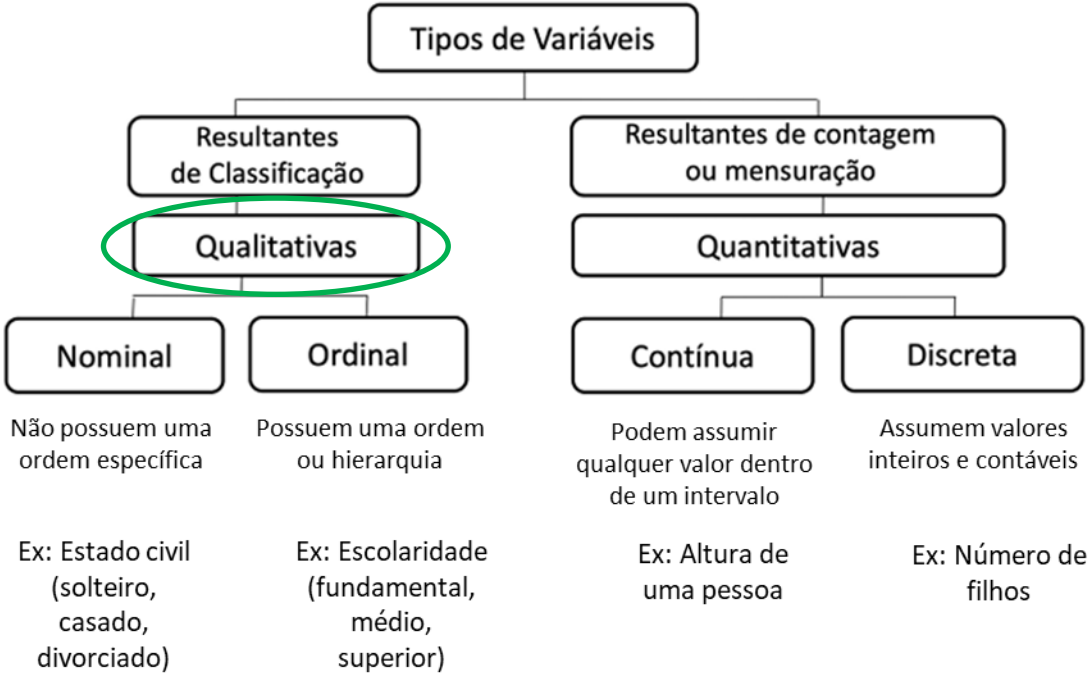
Qualitativas

Também são conhecidas como variáveis categóricas

São variáveis que **não podem ser medidas**; apenas, categorizadas ou contadas

Não permitem cálculos de estatísticas descritivas de posição ou dispersão

Dividem-se em nominais e ordinais



O que é e quais são os Tipos de Variáveis

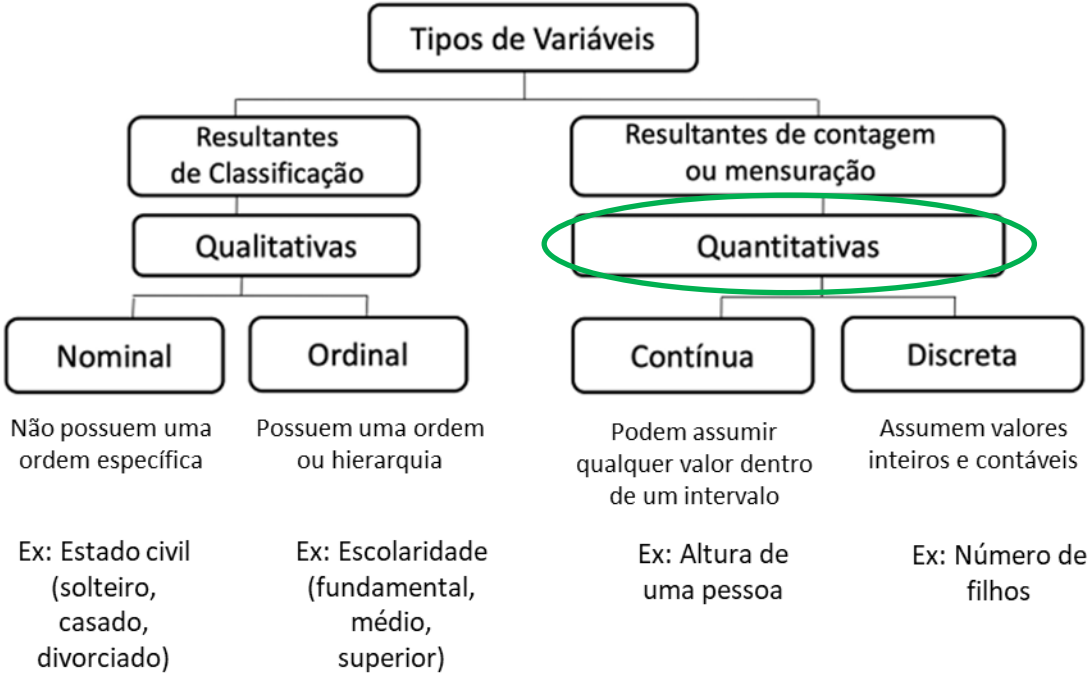
Quantitativas

Também são conhecidas como variáveis métricas

São variáveis que **podem ser medidas**

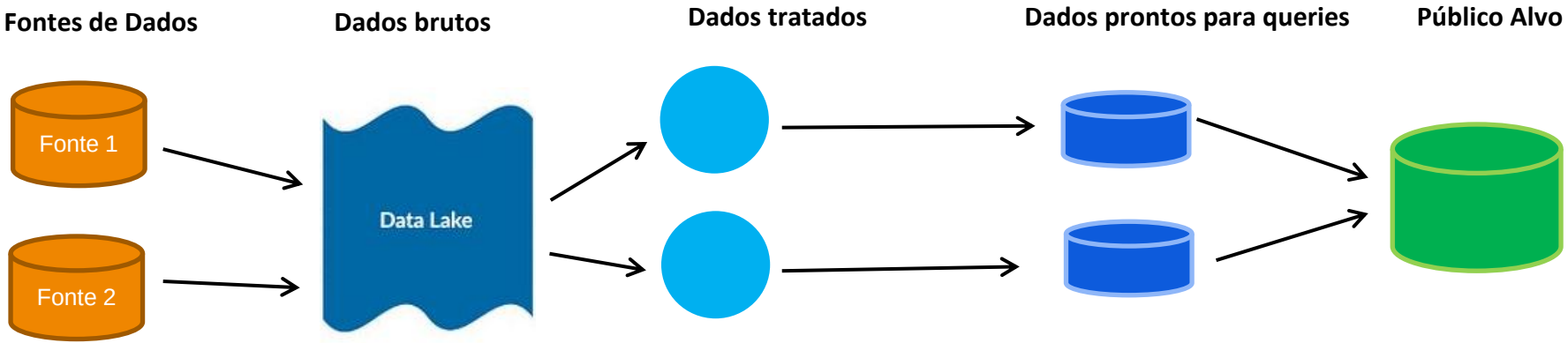
Permitem cálculos de estatísticas descritivas de posição ou dispersão

Dividem-se em contínua e discretas

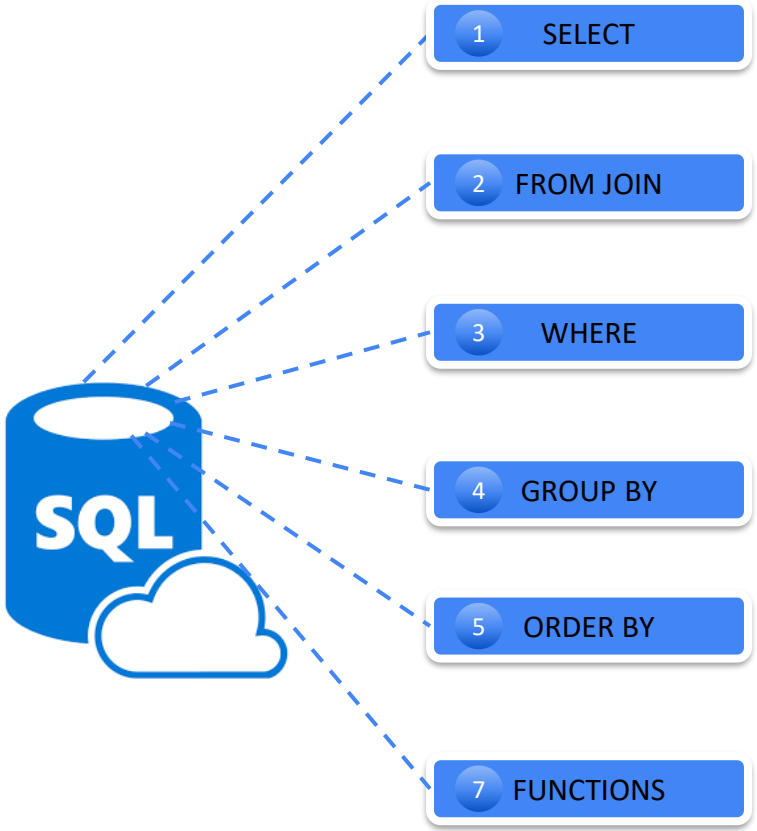


Exemplo Didático

Arquitetura de Dados

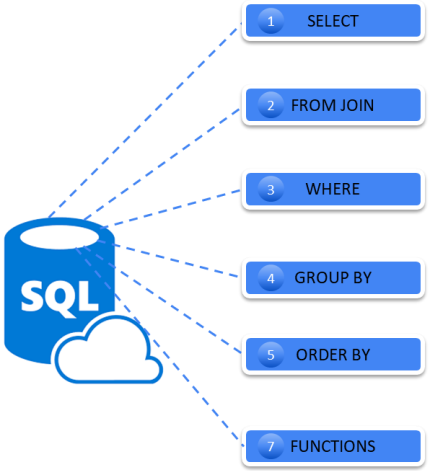


Comandos Básicos da Linguagem SQL




```
SELECT
    c.nome AS Cliente,
    c.cidade AS Cidade,
    p.data_pedido AS Data_Pedido,
    p.valor_total AS Total_Pedido,
    COUNT(i.produto) AS Itens_Distintos,
    AVG(i.preco_unitario) AS Media_Preco_Itens
FROM clientes c
INNER JOIN pedidos p ON c.id = p.cliente_id  -- Relaciona clientes com pedidos
LEFT JOIN itens_pedido i ON p.id = i.pedido_id  -- Relaciona pedidos com itens
WHERE p.valor_total > 100  -- Apenas pedidos acima de R$ 100,00
GROUP BY c.nome, c.cidade, p.data_pedido, p.valor_total  -- Agrupando por cliente e pedido
ORDER BY p.valor_total DESC;  -- Ordenando pelo maior total de pedido
```

Comandos Básicos da Linguagem SQL




```sql
SELECT
    c.nome AS Cliente,
    c.cidade AS Cidade,
    p.data_pedido AS Data_Pedido,
    p.valor_total AS Total_Pedido,
    COUNT(i.produto) AS Itens_Distintos,
    AVG(i.preco_unitario) AS Media_Preco_Itens
FROM clientes c
INNER JOIN pedidos p ON c.id = p.cliente_id  -- Relaciona clientes com pedidos
LEFT JOIN itens_pedido i ON p.id = i.pedido_id  -- Relaciona pedidos com itens
WHERE p.valor_total > 100  -- Apenas pedidos acima de R$ 100,00
GROUP BY c.nome, c.cidade, p.data_pedido, p.valor_total  -- Agrupando por cliente e pedido
ORDER BY p.valor_total DESC;  -- Ordenando pelo maior total de pedido
```

✓ **SELECT** → Escolhemos as colunas desejadas.
✓ **FROM** clientes → A tabela principal é clientes.
✓ **INNER JOIN** pedidos → Pegamos os pedidos do cliente.
✓ **LEFT JOIN** itens_pedido → Pegamos os itens do pedido (se houver).
✓ **WHERE** p.valor_total > 100 → Filtramos pedidos com valor acima de R$ 100,00.
✓ **GROUP BY** → Agrupamos por cliente e pedido para aplicar funções agregadoras.
✓ **COUNT**(i.produto) → Contamos os produtos diferentes no pedido.
✓ **AVG**(i.preco_unitario) → Calculamos o preço médio dos itens.
✓ **ORDER BY** p.valor_total DESC → Ordenamos pelo maior total de pedido.

Exemplo Didático

Público Alvo

Colunas: variáveis

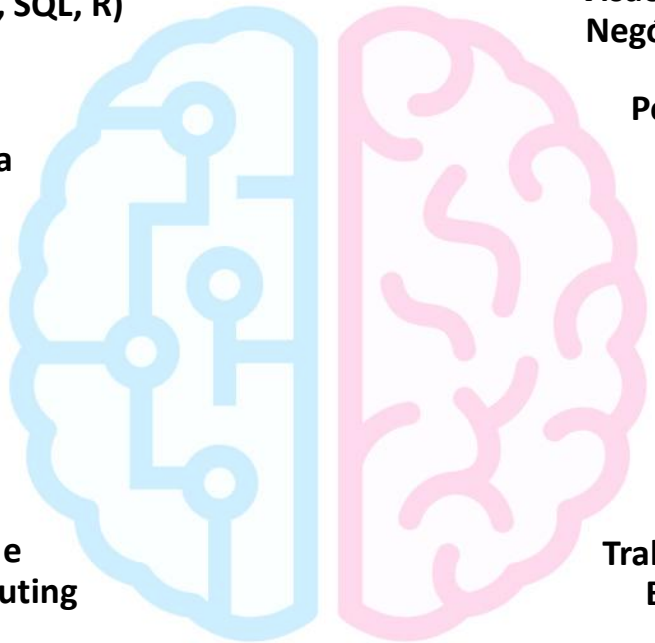
linhas: observações

Cliente	Cidade	Data_Pedido	Total_Pedido	Itens_Distintos	Media_Preco_Itens
João	São Paulo	20/02/2024	350	3	80
Maria	Curitiba	18/02/2024	210	2	105
Pedro	Salvador	15/02/2024	150	1	150

HARD Skills

são suas habilidades **técnicas**

- **Programação (Python, SQL, R)**
- **Machine Learning & Modelagem Estatística**
- **Manipulação e Engenharia de Dados**
- **Visualização de Dados**
- **Big Data e Cloud Computing**



SOFT Skills

são suas habilidades **comportamentais**

- **Visão de Negócio**
- **Pensamento Crítico**
- **Comunicação e Storytelling**
- **Resiliência**
- **Trabalho em Equipe**

Macro Etapas

Análise Exploratória de Dados

1

Problema
(Identifica o problema de negócio)

2

Dados
(Identifica as fontes e Coleta)

3

Análise
(Realiza análise exploratória)

4

Conclusão EDA
(resposta ao “o que” no contexto do problema de negócio)

Resumo de Dados

Medidas de Tendência Central

As medidas de tendência central são estatísticas que indicam o valor típico ou central de um conjunto de dados

Medidas de Variação

São medidas estatísticas que caracterizam o quanto um conjunto de dados está disperso em torno de sua tendência central

Resumo de Dados

Medidas de Tendência Central - Média

- Representam o valor típico ou central do conjunto de dados
- A média é útil quando os dados são simétricos e não possuem valores extremos (*outliers*)
- **Média (Média Aritmética)**: A média é calculada somando todos os valores e dividindo pelo número total de observações

Fórmula

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Onde:

- $\bar{X}$ = Média
- X_i = Cada valor da amostra
- n = Número total de observações

Exemplo:

Se temos os valores **5, 8, 10, 12, 15**, a média será:

$$\bar{X} = \frac{5 + 8 + 10 + 12 + 15}{5} = \frac{50}{5} = 10$$

Resumo de Dados

Medidas de Tendência Central - Mediana

- Valor central quando os dados estão ordenados
- A mediana é ideal quando os dados possuem **distribuição assimétrica** ou **outliers**, pois não é afetada por valores extremos

Fórmula

- **Se n for ímpar:** A mediana é o valor que está exatamente no meio.

$$\text{Mediana} = X_{\left(\frac{n+1}{2}\right)}$$

- **Se n for par:** A mediana é a média dos dois valores centrais.

$$\text{Mediana} = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

Exemplo:

Valores ordenados: **5, 8, 10, 12, 15**

- Como há 5 valores ($n = 5$, ímpar), a mediana é o terceiro valor: **10**.

Valores ordenados: **3, 7, 9, 12, 14, 18**

- Como há 6 valores ($n = 6$, par), a mediana é:

$$\frac{9 + 12}{2} = \frac{21}{2} = 10.5$$

Resumo de Dados

Medidas de Tendência Central - Moda

- A moda é o valor que aparece com maior frequência no conjunto de dados
- A moda é útil quando se deseja saber **o valor mais comum** em um conjunto de dados categóricos ou discretos
  - ✓ **Sem moda:** Nenhum valor se repete
  - ✓ **Moda única (unimodal):** Apenas um valor tem a maior frequência
  - ✓ **Multimodal:** Dois ou mais valores têm a mesma maior frequência

Exemplo 01:

Valores: **3, 5, 7, 8, 8, 9, 10**

A moda é **8**, pois aparece duas vezes

Exemplo 02:

Valores: **2, 4, 4, 6, 6, 8, 10**

O conjunto é **bimodal** (duas modas: **4 e 6**)

Resumo de Dados

Medidas de Tendência Central - Resumo: Qual usar?

Medida	Quando usar?	Exemplo
Média	Quando os dados são simétricos e não há outliers	Determinar a **renda média** dos clientes de uma loja
Mediana	Quando há outliers ou distribuição assimétrica	Identificar o **valor típico de financiamento** sem ser influenciado por clientes com crédito muito alto
Moda	Para identificar o valor mais comum em dados categóricos ou discretos	Descobrir **o canal de pagamento mais usado** pelos clientes

Resumo de Dados

Medidas de Tendência Central

As medidas de tendência central são estatísticas que indicam o valor típico ou central de um conjunto de dados

Medidas de Variação

São medidas estatísticas que caracterizam o quanto um conjunto de dados está disperso em torno de sua tendência central

Resumo de Dados

Medidas de Variação - Amplitude

- A amplitude é a diferença entre o maior e o menor valor do conjunto de dados

Fórmula

$$A = X_{\max} - X_{\min}$$

Exemplo:

Valores: **4, 8, 6**

$$A = 8 - 4 = 4$$

Resumo de Dados

Medidas de Variação – Variância

- A variância mede a dispersão dos dados em relação à média
- Quanto maior a variância, mais espalhados estão os valores

Fórmula

- **Para uma população** (σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

- **Para uma amostra** (s^2):

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- X_i = Cada valor do conjunto de dados
- μ = Média populacional
- $\bar{X}$ = Média amostral
- N = Tamanho da população
- n = Tamanho da amostra

Resumo de Dados

Medidas de Variação – Desvio-padrão

- O desvio padrão é a raiz quadrada da variância e tem a mesma unidade dos dados originais.

Fórmula

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}$$

Se $s^2 = 4$, então:

$$s = \sqrt{4} = 2$$

Interpretação: O desvio padrão indica que os dados tendem a variar **± 2** em torno da média

Resumo de Dados

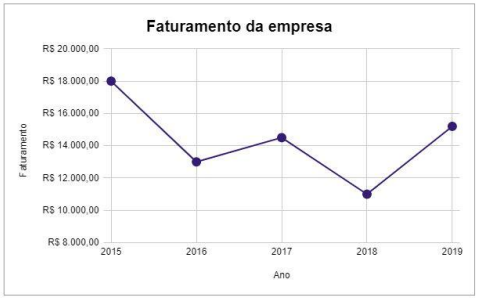
Medidas de Dispersão - Resumo: Qual usar?

Medida	Definição	Exemplo
Desvio Padrão	Mede a dispersão dos dados em relação à média	Medir a inconsistência no tempo de entrega de pedidos de um restaurante
Amplitude	Mede a diferença entre maior e menor valor	Comparar a altura mínima e máxima de jogadores em um time de basquete
Coeficiente de Variação	Mede a dispersão relativa (%)	Comparar a variação no peso dos pacotes em uma empresa de logística

Tipos de Gráficos

Linhas

Provavelmente o elemento visual mais usado para dados em série temporal. Eles mostram a relação entre duas variáveis, e no mais das vezes usados para rastrear mudanças ou tendências ao longo do tempo

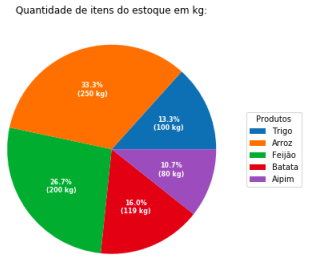


Barras

Eles são indicados quando existem dados nominais ou dados numéricos bem segmentados entre diferentes categorias. Podem ser orientados na vertical ou na horizontal

Setores

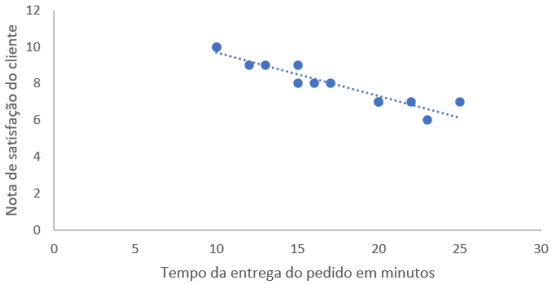
Devem ser utilizados para ilustrar proporções relativas de uma medida específica. Não utilizar quando existir muitas categorias



Tipos de Gráficos

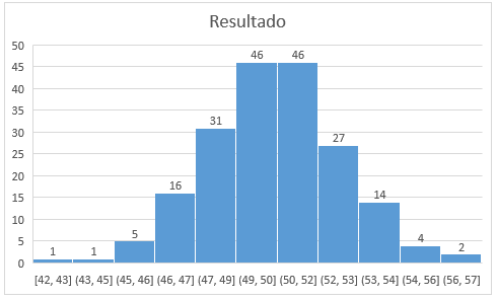
Dispersão

Costumam ser utilizados para explorar a relação entre duas ou três variáveis. É uma maneira eficiente de explorar a existência de tendências, concentrações e valores discrepantes



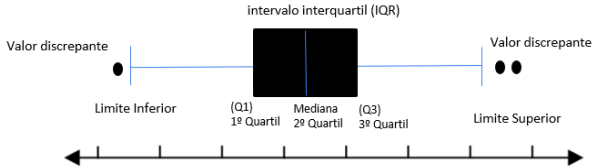
Histograma

São utilizados para mostrar a distribuição de frequência de uma variável ou mais variáveis



Boxplot

O Boxplot, também conhecido como diagrama de caixa, é uma ferramenta estatística gráfica que permite visualizar a distribuição, a variabilidade e os valores discrepantes (outliers) de um conjunto de dados



Análise Bivariada

A **análise bivariada** examina a relação entre **duas variáveis** para identificar padrões, correlações e associações. Dependendo do tipo de variável (quantitativa ou qualitativa), diferentes técnicas são aplicadas.

Em todas as situações, o objetivo é encontrar as **possíveis relações** ou associações entre as duas variáveis

Análise Bivariada

As duas variáveis são quantitativas

Quando as **duas variáveis** são **quantitativas**, as observações são provenientes de mensurações, e técnicas como **gráficos de dispersão e correlação** são apropriadas

As duas variáveis são qualitativas

Quando as variáveis são **qualitativas**, os dados são resumidos em **tabelas de contingências** (dupla entrada)

Uma variável é qualitativa e outra é quantitativa

Quando temos **uma variável qualitativa** e **outra quantitativa**, em geral analisamos o que acontece com a variável quantitativa quando os dados são **categorizados** de acordo com os diversos atributos da variável qualitativa

Gráfico de dispersão entre **duas variáveis quantitativas**

- O **diagrama de dispersão** (ou gráfico de dispersão) é um gráfico utilizado para visualizar a relação entre duas variáveis quantitativas (numéricas).
- Cada ponto no gráfico representa uma observação e suas coordenadas são definidas pelos valores das duas variáveis analisadas

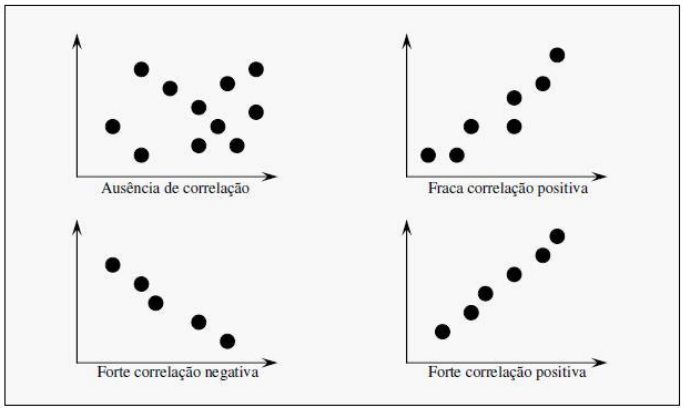
Principais usos do diagrama de dispersão

- Identificar padrões e tendências entre duas variáveis numéricas
- Detectar correlações positivas, negativas ou inexistentes
- Visualizar outliers (valores discrepantes)

Coeficiente de Correlação de Pearson

duas variáveis são quantitativas

O **diagrama de dispersão** e a **correlação** estão diretamente relacionado



Diagramas de Dispersão.

Coeficiente de Correlação de Pearson

A **correlação** mede a intensidade e o sentido da relação entre duas variáveis numéricas (X e Y) e é geralmente expressa pelo **coeficiente de correlação de Pearson (*r*)** , que varia entre **-1 e 1**

- **r>0** → Correlação **positiva** (quando uma variável aumenta, a outra tende a aumentar)
- **r<0 → Correlação negativa** (quando uma variável aumenta, a outra tende a diminuir)
- **r=0 → Nenhuma correlação** (as variáveis não possuem relação linear)

Inferência Estatística

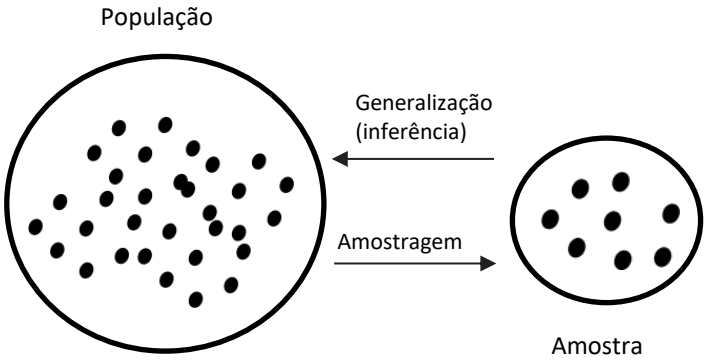
A **Inferência Estatística** é a área da estatística que desenvolve métodos para tirar conclusões sobre uma **população** com base em uma **amostra** extraída dessa população

Conceito

População e Amostra

População

População é o conjunto de **todos** os elementos sob investigações



Amostra

Amostra é qualquer **subconjunto** da população

Inferência Estatística – Tipos de Amostragem

Existem dois grupos principais de métodos de amostragem

Métodos de Amostragem Probabilística:

- Aleatória simples
- Sistemática
- Estratificada
- Conglomerados

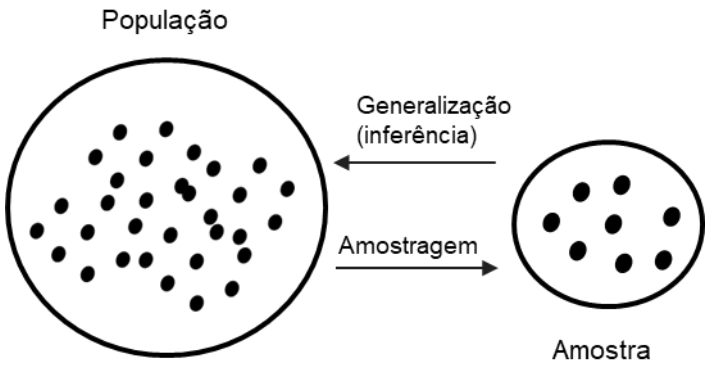
Métodos de Amostragem Não Probabilística:

- Conveniência
- Julgamento
- Cotas

Inferência Estatística – Usos de Amostragem

- Pesquisa Eleitoral
- Pesquisa com clientes
- Desenvolvimentos de modelos estatísticos, aprendizado de máquina
  - ✓ Amostra de Treinamento
  - ✓ Amostra de Validação
  - ✓ Amostra de Teste

Etapas da Amostragem



1 **Definição do Problema e Objetivo da Pesquisa**

- Determinar o que se deseja estudar e qual é o objetivo da amostragem
- Identificar a população-alvo para garantir que a amostra seja representativa

2 **Definição da População-Alvo**

- Delimitar claramente o grupo de interesse, especificando critérios de inclusão e exclusão
- Exemplo: "Clientes que contrataram empréstimo nos últimos 12 meses"

3 **Escolha do Método de Amostragem**

- Amostragem probabilística (quando todos os elementos têm chance conhecida de serem selecionados)
- Amostragem não probabilística (quando a seleção não é totalmente aleatória)

4 **Determinação do Tamanho da Amostra**

- Definir a quantidade necessária para garantir representatividade estatística
- Depende de fatores como erro amostral, nível de confiança e variabilidade da população

5 **Coleta da Amostra**

- Aplicar a metodologia escolhida para selecionar os indivíduos da população
- Garantir que os dados sejam coletados corretamente, sem viés ou erros

6 **Verificação e Tratamento da Amostra**

- Avaliar a qualidade da amostra e verificar se houve viés de seleção
- Comparar características da amostra com a população para garantir representatividade

7 **Análise dos Dados e Generalização**

- Realizar as análises estatísticas considerando as características da amostra
- Interpretar os resultados e avaliar sua aplicabilidade à população-alvo

Por que Usar Amostras em Data Science?

1 Redução de Custo e Tempo

- Processar grandes volumes de dados é caro e demorado
- Exemplo: Testar um modelo de Machine Learning com x% dos dados antes de rodar na base completa

2 Facilidade para Testes e Experimentação

- Permite testar diferentes abordagens rapidamente
- Exemplo: Avaliar impacto de features usando uma amostra antes de treinar o modelo final

3 Dados Difíceis de Coletar

- Nem sempre é possível coletar todos os dados da população
- Exemplo: Pesquisas eleitorais usam amostras para estimar intenção de voto

4 Limitação de Recursos Computacionais

- Processamento de big data exige otimização
- Exemplo: Executar algoritmos de clustering em uma amostra antes da base completa

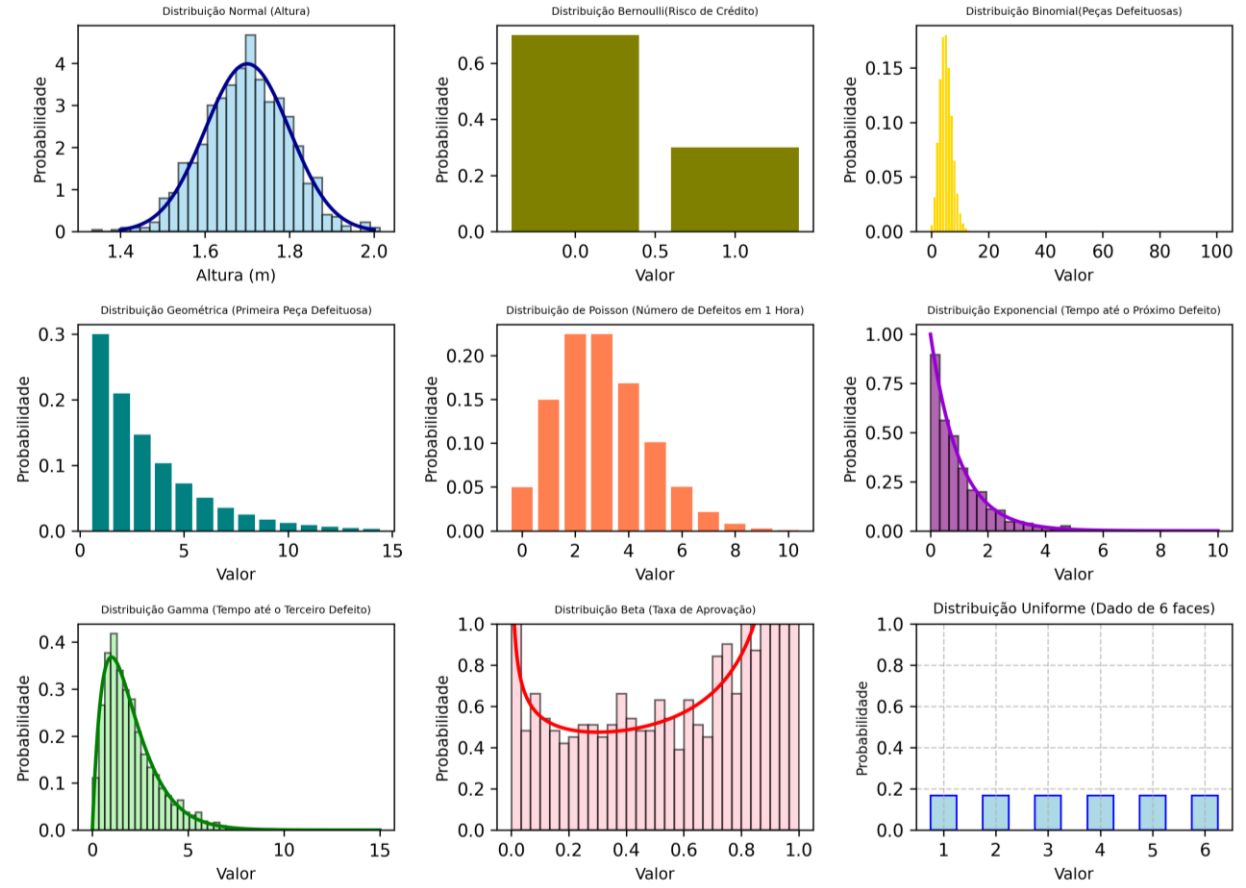
Distribuições de Probabilidade

1. **Compreensão dos Dados**: conhecer a distribuição ajuda a entender o comportamento e as características dos dados (média, variabilidade, etc.)

2. **Inferência Estatística**: Permite realizar estimativas precisas, testes de hipóteses e intervalos de confiança

3. **Escolha do Modelo**: A distribuição influencia a escolha do modelo adequado para análise ou predição

4. **Simulações e Modelagem**: Permite realizar simulações para prever cenários futuros de forma mais precisa



Exemplos

Preparação de Dados

Data Prep

1

Problema
(Identifica o problema de negócio)

2

Dados
(Identifica as fontes e Coleta)

3

Análise
(Realiza análise exploratória)

4

Conclusão EDA
(resposta ao “o que” no contexto do problema de negócio)

5

Pré-processamento
(Realiza tratamentos necessários aos dados no contexto do problema de negócio)

6

Conclusão Data Prep
(Apresenta dados tratados no contexto do problema de negócio)

Preparação de Dados - Data Prep

Só devemos **TRATAR** nossos dados **DEPOIS** de fazer o Split no Dataset para evitar: ***Data Leakage*** e ***Overfitting***

SEMPRE devemos realizar o split **ANTES** do pré-processamento dos dados, garantindo que os conjuntos de treino e teste sejam tratados como entidades completamente separadas

Preparação de Dados
Data Prep

1

Limpeza de Dados

- A limpeza de dados envolve a identificação e correção de inconsistências.
- Tratamento de valores ausentes, Tratamento de valores duplicados, Correção de erros tipográficos e padronização e Detecção e tratamento de outliers

2

Transformação de Dados

- A transformação de dados visa estruturar os dados para melhorar a análise e modelagem.
- Normalização vs. Padronização, Transformação de variáveis categóricas e Transformação de variáveis temporais

Preparação de Dados
Data Prep

3 Engenharia de Recursos

- Engenharia de recursos envolve criar novas features que melhorem o desempenho do modelo.
- Combinação de variáveis, Transformações matemáticas, Agrupamentos e estatísticas agregadas e Detecção de variáveis altamente correlacionadas

4 Seleção de Variáveis

- A seleção de recursos visa reduzir a dimensionalidade e aumentar a interpretabilidade do modelo
- Métodos de seleção: *Filter Methods*: Baseados em estatísticas (ex.: mutual information, correlação)
- Wrapper Methods: Testa combinações de features com um modelo específico (ex.: RFE - Recursive Feature Elimination)
- Embedded Methods: Métodos internos aos modelos, como regularização L1/L2

Linguagem Python: o que é e para que serve ?

Linguagem Python

Python é uma linguagem de programação de alto nível desenvolvida entre 1980 e 1990

É amplamente utilizada em desenvolvimento de *software*, automação, análise de dados e inteligência artificial

O Python é a linguagem mais utilizada em *Data Science*

Versatilidade

Python é uma linguagem de programação de uso geral amplamente utilizada para coleta e engenharia de dados, análise, Web Scraping, desenvolvimento de aplicativos web e diversas outras aplicações

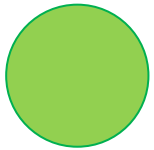
Bibliotecas Poderosas para *Data Science*

Ampla variedade de bibliotecas disponíveis como Pandas, NumPy e Seaborn facilitam a manipulação, análise e visualização de dados, tornando os processos mais eficientes e acessíveis.

Facilidade no Uso para Inteligência Artificial

Diversas bibliotecas para aplicação em projetos de Inteligência Artificial, como Scikit-learn, CatBoost, XGBoost, LightGBM e PyCaret, que oferecem funções prontas para a criação e otimização de modelos.

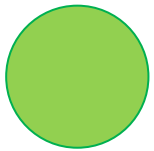
Linguagem Python - Bibliotecas Poderosas para *Data Science*



Pandas (Manipulação de Dados)

- ✓ A biblioteca é essencial para manipulação e análise de dados em Python
- ✓ Permite ler e escrever dados em diversos formatos (CSV, Excel, SQL) e realiza limpeza e pré-processamento de dados, com funções para tratar valores ausentes e substituí-los
- ✓ A biblioteca permite seleção e filtragem de dados usando índices e condições, além de realizar operações agregadas e estatísticas com funções como **groupby()**, **mean()** e **sum()**
- ✓ Também facilita a combinação de conjuntos de dados com **merge()** e **concat()**, permitindo unir ou concatenar DataFrames de forma eficiente

Linguagem Python - Bibliotecas Poderosas para *Data Science*



Numpy (operações numéricas)

- ✓ O NumPy é uma biblioteca essencial para computação científica em Python, especializada no suporte a matrizes e arrays multidimensionais
- ✓ Sua principal estrutura de dados é o array, que contém elementos homogêneos e de tamanho fixo, permitindo operações rápidas e eficientes
- ✓ O NumPy oferece funções matemáticas avançadas para operações vetoriais e matriciais, como soma, multiplicação e álgebra linear, além de métodos para criar arrays (zeros(), ones(), arange()) e realizar indexação e fatiamento avançados
- ✓ Ele facilita a manipulação de grandes volumes de dados numéricos, promovendo eficiência por meio de operações vetorizadas

Linguagem Python - Bibliotecas Poderosas para *Data Science*

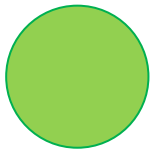
Matplotlib (Vizualização de dados)

- ✓ O Matplotlib é uma das bibliotecas mais populares para visualizações em Python, oferecendo diversas opções para criar gráficos 2D, como linhas, barras, dispersão e histogramas.
- ✓ Sua principal estrutura de dados é a Figura e os Eixos, que permitem um controle detalhado sobre o estilo, tamanho e formato dos gráficos.
- ✓ Com a função pyplot, é possível gerar gráficos de forma simples e personalizá-los com títulos, legendas e rótulos.

Seaborn (Vizualização de dados)

- ✓ O Seaborn, construído sobre o Matplotlib, facilita a criação de gráficos estatísticos de alta qualidade, proporcionando uma interface mais simples e eficiente para gráficos complexos, como distribuições, correlações e categóricos, com estilos e paletas de cores atraentes por padrão.
- ✓ Ele se integra bem com o Pandas, permitindo criar visualizações diretamente de DataFrames com funções como **sns.barplot()**, **sns.heatmap()** e **sns.pairplot()**

Linguagem Python - Bibliotecas Poderosas para *Data Science*



scikit-learn (Aprendizado de Máquina)

- ✓ O Scikit-learn é uma das bibliotecas mais populares para aprendizado de máquina em Python, oferecendo uma ampla gama de ferramentas e algoritmos para tarefas supervisionadas e não supervisionadas, como classificação, regressão, clustering, redução de dimensionalidade, seleção de modelos e validação cruzada.
- ✓ Ele é altamente integrado com outras bibliotecas como NumPy e Pandas, facilitando a manipulação de dados e a implementação de modelos.
- ✓ A biblioteca inclui diversos algoritmos de aprendizado de máquina, como Regressão Linear, SVM, Árvores de Decisão, KNN, Random Forests e K-Means.
- ✓ Além disso, oferece ferramentas para pré-processamento de dados (normalização, padronização e imputação de valores ausentes) e funcionalidades para avaliar e ajustar o desempenho dos modelos, com métricas de avaliação (precisão, recall, F1-score) e técnicas de ajuste de hiperparâmetros, como grid search.

Onde encontrar Datasets ?

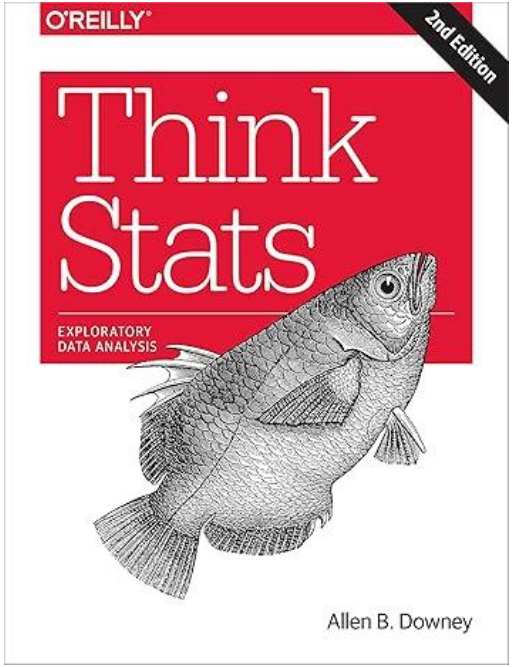
Os datasets são bases de dados específicas que servem de amostras para treinamentos de algoritmos de inteligência artificial ou para outros tipos de projetos de Data Science

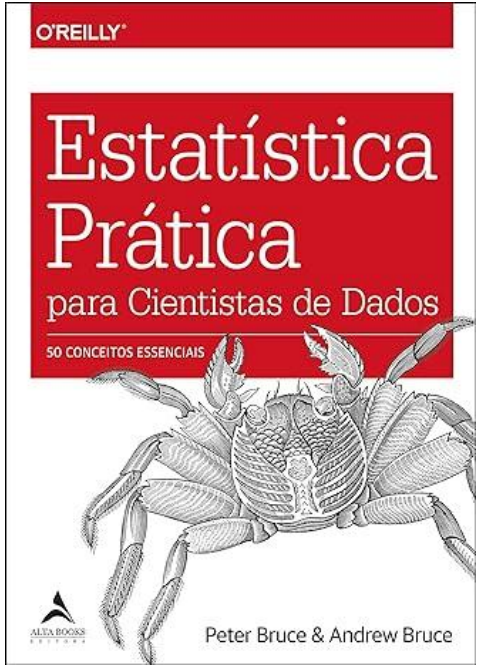
Kaggle: (https://www.kaggle.com/datasets): Uma das plataformas mais populares para datasets. O Kaggle oferece uma ampla gama de datasets em diversas áreas, como saúde, finanças, esportes, e muito mais. Além de datasets, a plataforma também oferece competições de aprendizado de máquina e notebooks interativos.

UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
O UCI Repository é uma das fontes mais tradicionais de datasets para aprendizado de máquina. Ele contém diversos datasets em várias áreas, como biologia, medicina, ciência social e engenharia.

Data.gov (https://www.data.gov/)
Um repositório do governo dos EUA que oferece acesso a uma enorme quantidade de dados públicos sobre uma ampla gama de temas, incluindo educação, saúde, transporte, e meio ambiente.

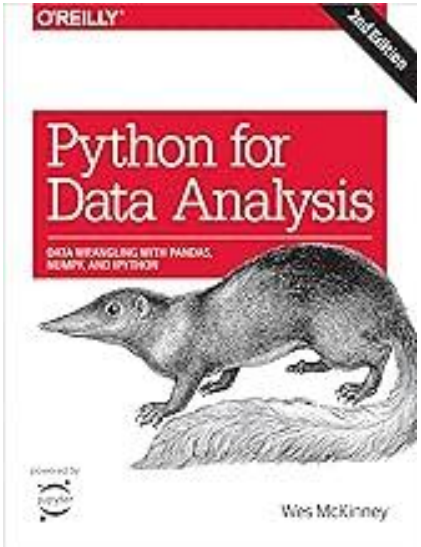
Indo Além

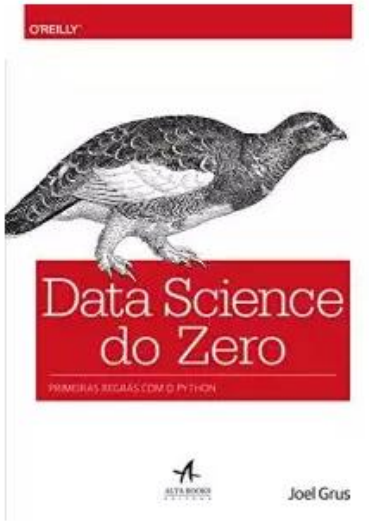

O'REILLY
2nd Edition
Think Stats
EXPLORATORY DATA ANALYSIS
Allen B. Downey


O'REILLY
Estatística Prática
para Cientistas de Dados
50 CONCEITOS ESSENCIAIS
Peter Bruce & Andrew Bruce


cole nussbaumer knaflic
storytelling com dados
um guia sobre visualização de dados para profissionais de negócios

Indo Além


O'REILLY
2nd Edition
Python for Data Analysis
DATA WRANGLING WITH PANDAS, NUMPY, AND IPYTHON
Wes McKinney


O'REILLY
Data Science do Zero
Joel Grus