



Tutorial: Introdução à Análise de Dados com Pandas, Matplotlib e Seaborn

Vitor Moreira Casagrande
Thiago Pereira da Silva



Agenda



01

**Introdução à Análise
de Dados**

02

**Conceitos
Fundamentais**

03

**Ferramentas de
Análise de Dados**

04

**Processo de Análise
de Dados (Pipeline)**

05

**Áreas de Aplicação
da Análise de Dados**

06

**Carreiras em Análise
de Dados**



Quem Somos?



Vitor Moreira Casagrande

- Estudante do curso de Ciência da Computação UFMT Araguaia.
- Entusiasta em Análise de Dados.
- vormoreiracasagrande@hotmail.com.br



Thiago Pereira da Silva

- Doutor em Ciência da Computação.
- Professor do curso de Ciência da Computação UFMT Araguaia.
- thiago.silva@ufmt.br
- <http://lattes.cnpq.br/0241704052892662>

01

Introdução à Análise de dados



O que é Análise de Dados


- É o processo de examinar, limpar, transformar e modelar dados para extrair **informações** úteis, *insights* e apoiar decisões (Foster Provost e Tom Fawcett, 2023).
 - O que os dados estão indicando e como eles podem ser utilizados para resolver problemas.
 - Aquisição de conhecimento.
- Usada em diversas áreas, como negócios, saúde e ciência, e geralmente envolve o uso de ferramentas e técnicas estatísticas e computacionais.





Qual o objetivo da Análise de Dados

Identificar padrões, tendências, correlações e anomalias nos dados que podem ser utilizados para:

- **Tomada de decisões;**
 - **Identificar padrões e tendências;**
 - **Aprimorar processos e operações;**
 - **Identificar novas oportunidades de negócios.**
- 

Dado x Informação x Conhecimento

- **Dado** é informação bruta e sem contexto.
- **Informação** é dado processado e contextualizado.
- **Conhecimento** é a interpretação e aplicação da informação com base em experiência e análise.



Dado

"Aluno 123, 8"



Informação

O aluno de matrícula 123 obteve a nota 8 no exame de matemática.



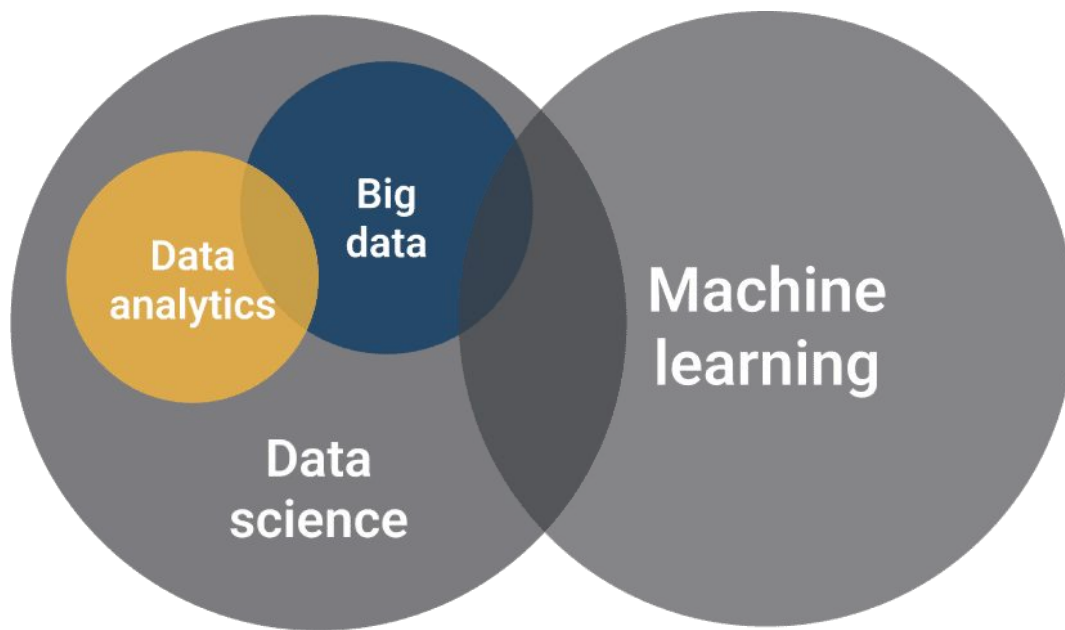
Conhecimento

Alunos com média acima de 7 no exame de matemática geralmente têm um bom desempenho em outras disciplinas.

Etapas Gerais do Processo de Análise de Dados



Situando à Análise de Dados

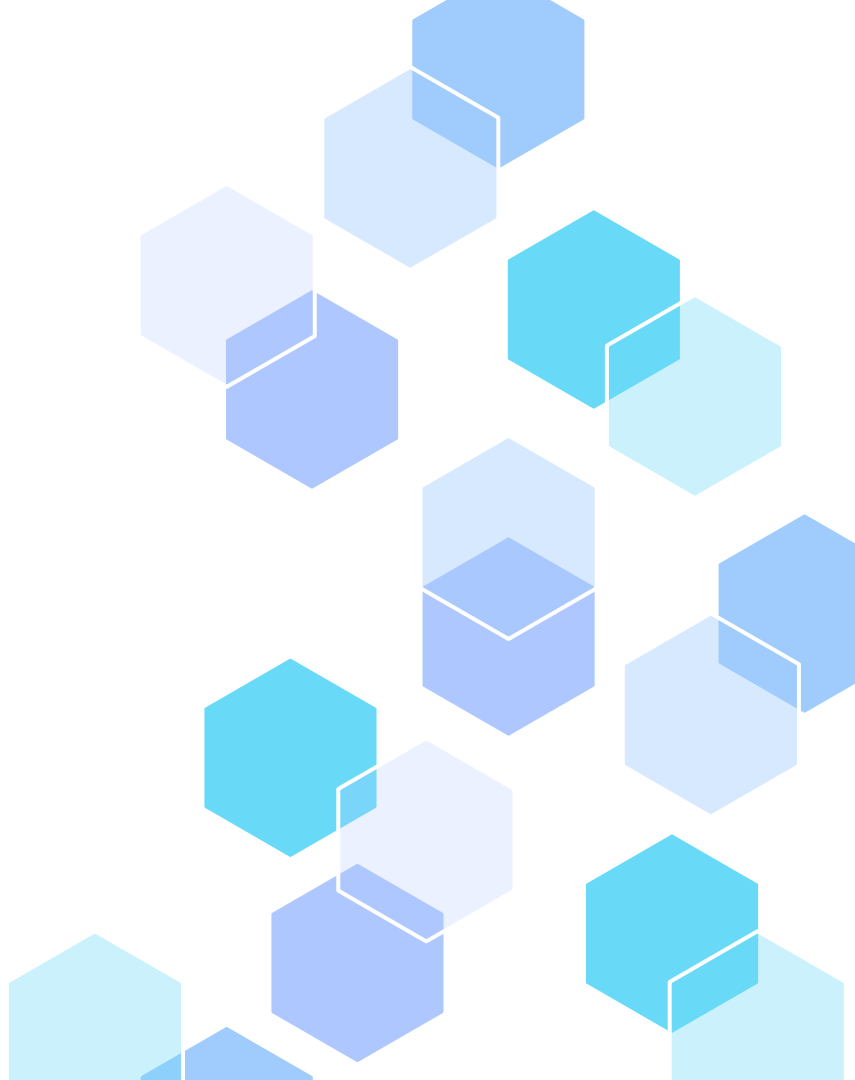


Fonte:

<https://blog.infnet.com.br/data-science/big-data-e-machine-learning-como-sao-usados-em-data-science/>

02

Conceitos Fundamentais



Tipos de Variáveis

Quantitativas (escala qualitativa)

- Discreta – inteiros (Ex. número de filhos, quantidade de reprovados)
- Contínuas – reais (Ex. peso corporal, temperatura)

Qualitativas (ou categóricas)

- Nominais – sem ordenação (Ex. sexo, cor dos olhos, doente/sadio)
- Ordinais – ordenação (Ex. escolaridade (1º, 2º, 3º graus), mês de observação (janeiro, fevereiro,..., dezembro))



Estatística Descritiva

- *Objetivo é sintetizar uma série de valores de mesma natureza, permitindo dessa forma que se tenha uma visão global da variação desses valores (MONTGOMERY; RUNGER 2014).*
- Nas variáveis quantitativas (discretas ou contínuas) às **medidas descritivas** mais comuns buscam responder às questões:
 - Locação (Centralidade)
 - Dispersão (Variabilidade)
 - Associação



Medidas de Localização

Moda

Valor mais frequente na distribuição dos dados. Distribuições podem ser unimodais ou multimodais.

Média

Média Aritmética

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Média Ponderada

$$\bar{x}_p = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_rx_r}{\sum f_r}$$

Mediana

Valor que separa 50% das observações à sua esquerda e 50% à sua direita quando os dados estão em ordenados. Em amostras pares: mediana é a média dos valores centrais.

Medidas de Localização

Moda

X

Média

X

Mediana

- A **moda** é útil em casos onde o valor mais frequente é de interesse.
- A **média** é influenciada por valores extremos (*outliers*); isso não ocorre com a **mediana**.
 - Ex. 2 4 6 8 10
Média = 6 e Mediana = 6
 - Ex. 2 4 6 8 100
Média = 24 e Mediana = 6

Medidas de Dispersão

Desvio Médio

Nível de dispersão, em média, da média aritmética.

$$\overline{DM} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Média

Variância (Desvio Padrão)

Nível de dispersão dos dados estão espalhados em relação à média.

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

Amplitude

Δ = maior valor – menor valor

Medidas de Dispersão

Desvio Médio

Variância
(Desvio Padrão)

Amplitude

- Desvio médio é menos sensível a valores extremos (*outliers*).
- Um **desvio padrão** pequeno indica que os valores estão mais próximos da média, enquanto um desvio padrão grande indica uma dispersão maior em relação à média.

Ex. Conjunto de dados [2,4,6,8,10]

Média = 6

Amplitude = 8

Variância = 8

Desvio Padrão = 2,83

Desvio Médio = 2,4

Ex. Conjunto de dados [2,4,6,8,**100**]

Média = 24

Amplitude = 98

Variância = 1448

Desvio Padrão = 38,05

Desvio Médio = 30,4

Medidas de Associação

Covariância

Indica a direção do relacionamento entre duas variáveis.

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Positiva: Variáveis aumentam ou diminuem juntas.

Negativa: Uma variável aumenta enquanto a outra diminui.

Coeficiente de Correlação de Pearson

Força e a direção do relacionamento linear entre duas variáveis.

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Desvio padrão

+1: Correlação perfeita positiva (as variáveis aumentam ou diminuem juntas).

-1: Correlação perfeita negativa (uma variável aumenta enquanto a outra diminui).

0: Nenhuma correlação linear

Medidas de Associação

Covariância

- O Coeficiente de Correlação de Pearson é normalizado entre -1 e 1. Quanto mais próximos de -1 e 1, mais relacionadas estão as variáveis.

Ex.

$X=[1,2,3,4,5,6,7,8,9,10]$

$Y=[2,4,5,4,6,8,7,10,9,12]$

$\text{Cov}(X,Y)=8.05$

$r \approx 0.96$ (correlação positiva forte)

Coeficiente de Correlação de Pearson

Ex.

$X=[1,2,3,4,3,6,7,82,9,10]$

$Y=[2,10,5,4,26,8,7,10,9,2]$

$\text{Cov}(X,Y)=9.49$

$r \approx 0.06$ (correlação fraca)

Quartis e Percentis

Quartis

Dividem um conjunto de dados ordenado em quatro partes iguais.

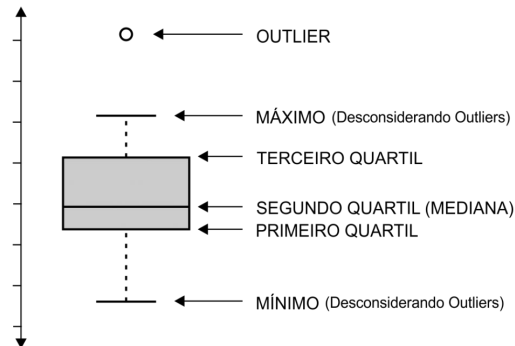
- **Q1 (Primeiro Quartil):** O valor que separa os 25% menores dados.
- **Q2 (Segundo Quartil ou Mediana):** O valor que separa os 50% dos dados (mediana).
- **Q3 (Terceiro Quartil):** O valor que separa os 75% menores dados.

Percentis

Dividem o conjunto de dados em 100 partes iguais.

- Percentil 50 é a mediana.
- Percentis 25 e 75 são Q1 e Q3, respectivamente

Boxplot





03

Ferramentas de Análises de Dados

Linguagem R



- Linguagem de programação.
- Análise de dados.
- Estatística.
- Visualização de dados.

<https://www.r-project.org/>

Python



- Versátil e Simples.
- Alta aplicabilidade (desenvolvimento web, análise de dados, inteligência artificial, etc).
- Alta gama de bibliotecas.

<https://www.python.org/>

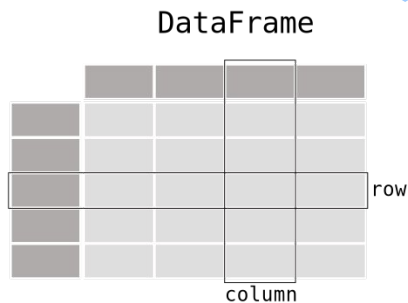
Pandas



- Biblioteca de código aberto para manipulação e análise de dados em Python.
- Focada em operações de dados tabulares, como em planilhas ou bancos de dados.
- Estrutura de dados do Pandas:
 - Séries e *Dataframes*.
- Ampla comunidade e documentação.
- Suporte para grandes volumes de dados.
- Integração com outras ferramentas de análise e aprendizado de máquina.

<https://pandas.pydata.org/>

Pandas



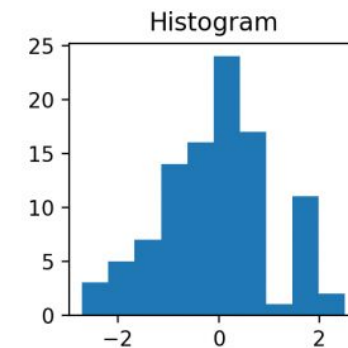
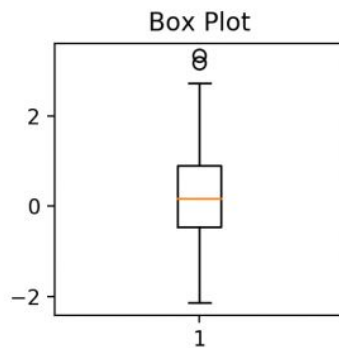
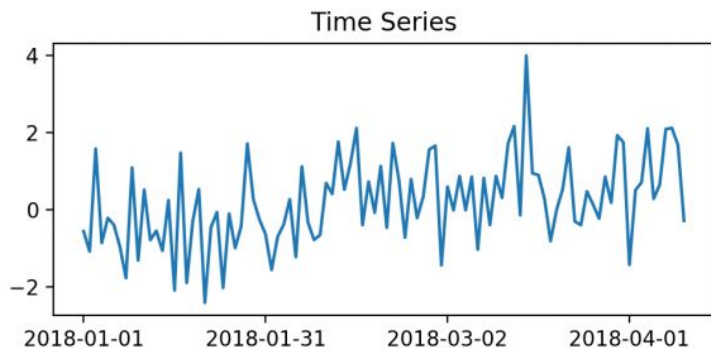
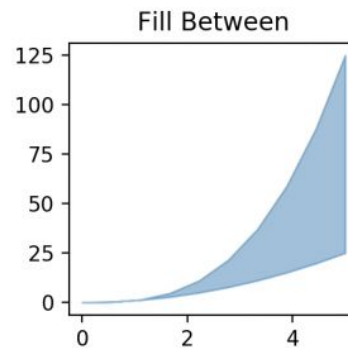
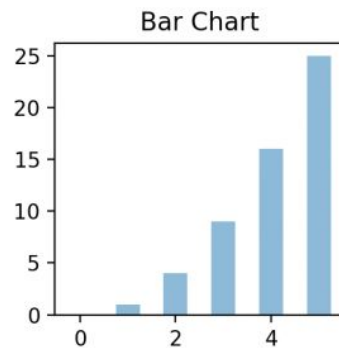
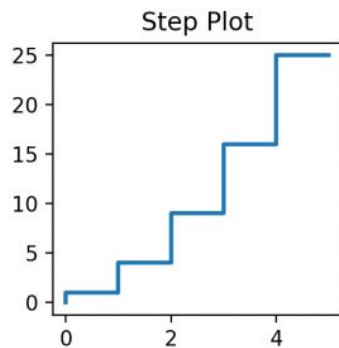
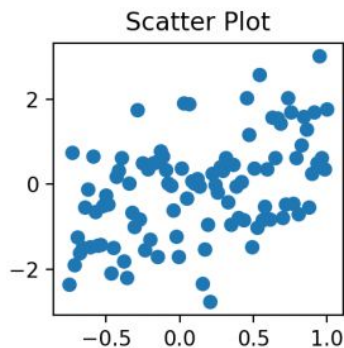
Series 1			Series 2			Series 3			Dataframe			
INDEX	DATA		INDEX	DATA		INDEX	DATA		INDEX	SERIES 1	SERIES 2	SERIES 3
0	A		0	1		0	[1, 2]		0	A	1	[1, 2]
1	B		1	2		1	A		1	B	2	A
2	C	&	2	3	&	2	1	=	2	C	3	1
3	D		3	4		3	(4, 5)		3	D	4	(4, 5)
4	E		4	5		4	{"a": 1}		4	E	5	{"a": 1}
5	F		5	6		5	6		5	F	6	6

matplotlib

- Biblioteca de código aberto para criação de gráficos e visualizações 2D.
 - Gráficos simples até visualizações mais complexas e customizadas
- Gráficos de linha
- Gráficos de barras
- Histogramas
- Boxplot
- Integração com Pandas, Numpy e Seaborn. (**Foco do tutorial!**)

<https://matplotlib.org/>

matplotlib





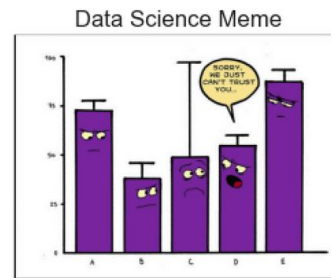
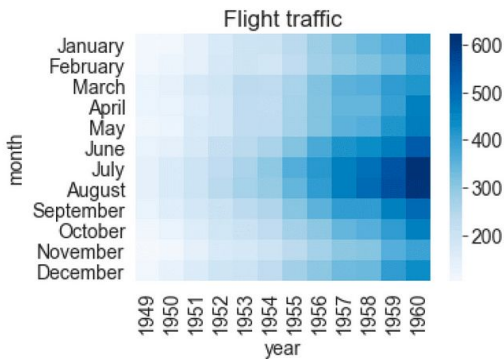
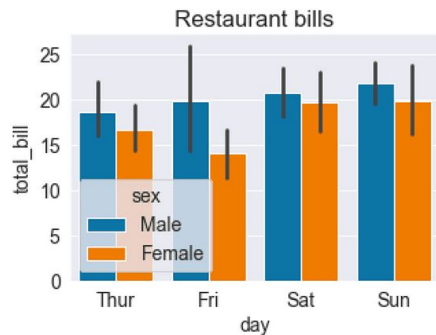
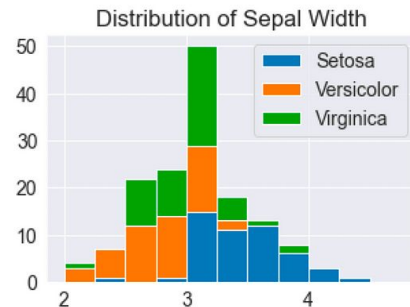
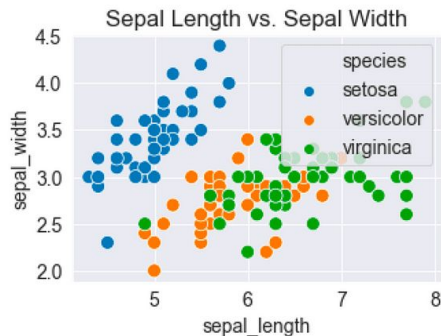
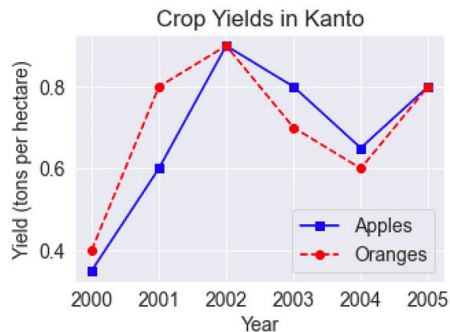
seaborn

- Biblioteca de visualização de dados baseada no Matplotlib.
- Fornece uma interface de alto nível para gráficos estatísticos, com estilo e paletas de cores aprimoradas.
- Visualizações Estatísticas.
- Estilo e Paleta de Cores.
- Integração com Pandas e Matplotlib.

<https://seaborn.pydata.org/>



seaborn

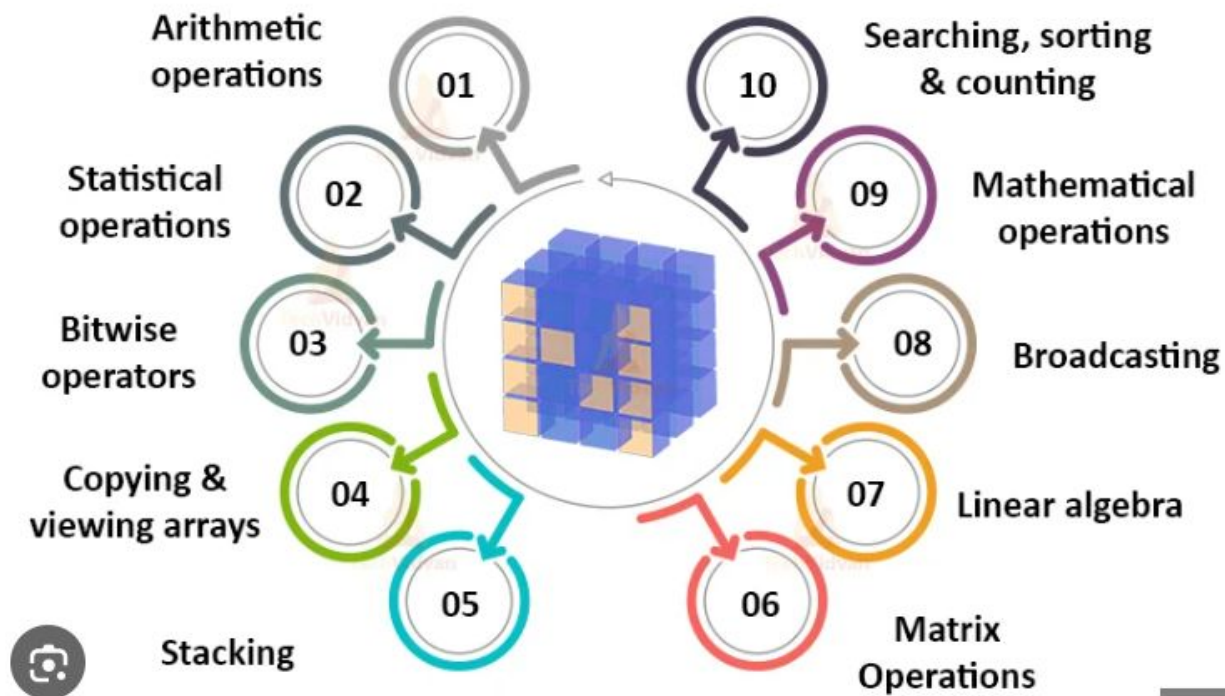




- Biblioteca de código aberto para computação numérica.
- Fornece suporte para arrays e matrizes multidimensionais, além de funções matemáticas avançadas.
- Escrito em C para performance, o NumPy é extremamente rápido em comparação com listas Python
 - Especialmente em operações com grandes conjuntos de dados.
- Integração com Pandas, SciPy e Scikit-Learn, e amplamente usado em ciência de dados e aprendizado de máquina.

<https://numpy.org/>

Uses of NumPy





- Ferramenta de código aberto para criação e compartilhamento de documentos que integram código, texto, gráficos e visualizações.
- Utilizado em análise de dados, aprendizado de máquina, pesquisa e ensino.
- Ambiente Interativo.
- Células de código e de *Markdown*.
- Execução Interativa.

<https://jupyter.org/>

<https://jupyter.org/try-jupyter/lab/>

jupyterlab

Google colab

- Ambiente de notebooks baseado em Jupyter que permite executar código Python diretamente no navegador.
- Armazena e processa dados na nuvem.
- Notebooks Compartilháveis e com colaboração em Tempo real.
- Processadores de alto desempenho (GPUs e TPUs).
- Integração com Google Drive e GitHub.
- Tempo limite de Sessão na modalidade gratuita.

<https://colab.google/>

04

Processo de Análise de Dados



Etapas de Análise de Dados

1. Coleta de Dados
2. Limpeza de Dados
3. Exploração e Visualização Inicial
4. Análise Exploratória e Modelagem
5. Interpretação e Apresentação de Resultados

Coleta de Dados:

- Processo de obtenção de informações relevantes para análise, investigação e tomada de decisões:
 - Pesquisas
 - Experimentos
 - Observações
 - Sensores e IoT
 - Web Scraping
 - etc.



Limpeza de Dados

- Processo de preparação dos dados para análise, eliminando ou corrigindo inconsistências, valores ausentes e erros.
- Etapa essencial para garantir que os dados sejam precisos, consistentes e relevantes para a análise.
- Valores ausentes.
- ***Outliers.***
- Duplicatas.



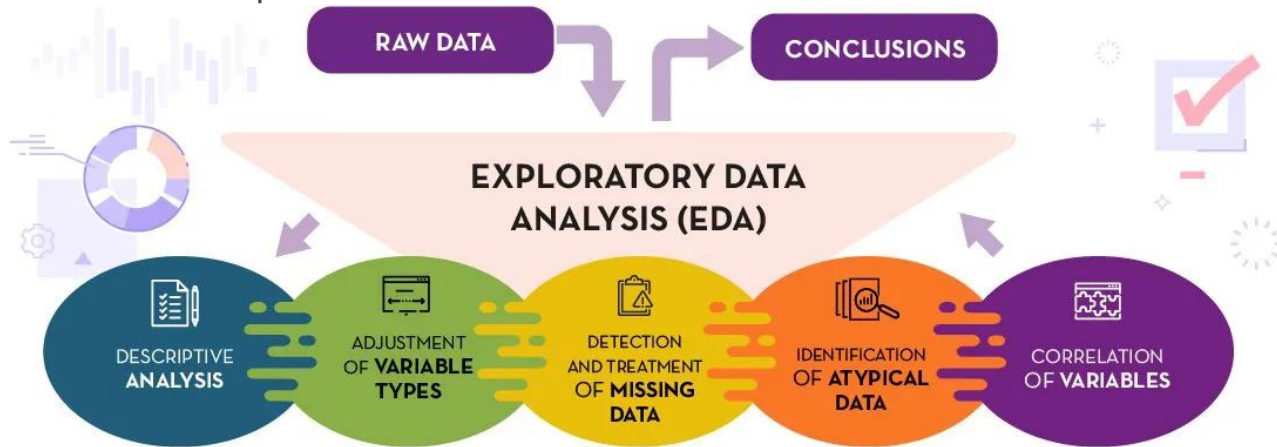
Exploração e Visualização Inicial

- Primeira análise exploratória dos dados para identificar padrões, tendências e possíveis problemas.
- Ajuda a formular hipóteses e direcionar as próximas etapas da análise.
- **Entendimento geral dos dados**
- Identificar padrões e tendências.
- Detecção de anomalias.
- **Estatísticas descritivas.**
- Análise de distribuição e correlação de dados.



Análise Exploratória e Modelagem

- Explorar os dados mais aprofundada, buscando relações e padrões que possam guiar a modelagem.
- Análise de correlação.
- Visualizações avançadas (*heatmaps, pairplots, etc*).
- Testes Estatísticos.
- Construção de modelos preditivos ou classificatórios.



Interpretação e Apresentação de Resultados

- Processo de traduzir os resultados de análises e modelagem para *insights* compreensíveis e relevantes para o problema em questão.
- Comunicação clara dos *insights*, das conclusões e das implicações dos resultados para as partes interessadas.
 - **Explicar as conclusões**
 - **Fornecer recomendações**
 - **Apoiar a tomada de decisões**
- Adicionalmente podem ser usados Dashboards interativos (Power BI, Tableau, etc).

05

Áreas de Aplicação da Análise de Dados



Negócios e Marketing



- Segmentação de Clientes
- Previsão de Vendas
- Análise de Sentimento

Finanças e Bancos

- Detecção de Fraudes
- Análise de Crédito
- Gestão de Riscos
- Entendimento do Mercado



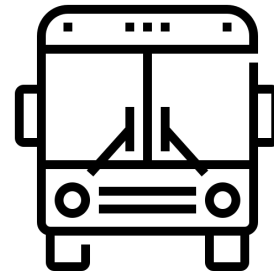
Saúde

- Diagnóstico Precoce e Prognóstico
- Pesquisa de Genética e Genômica
- Monitoramento de Pacientes
- Planejamento de Saúde Pública



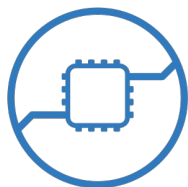
Setor Público e Governamental

- Planejamento Urbano e de Infraestrutura
- Previsão de Desastres Naturais
- Análise de Crimes
- Monitoramento Ambiental

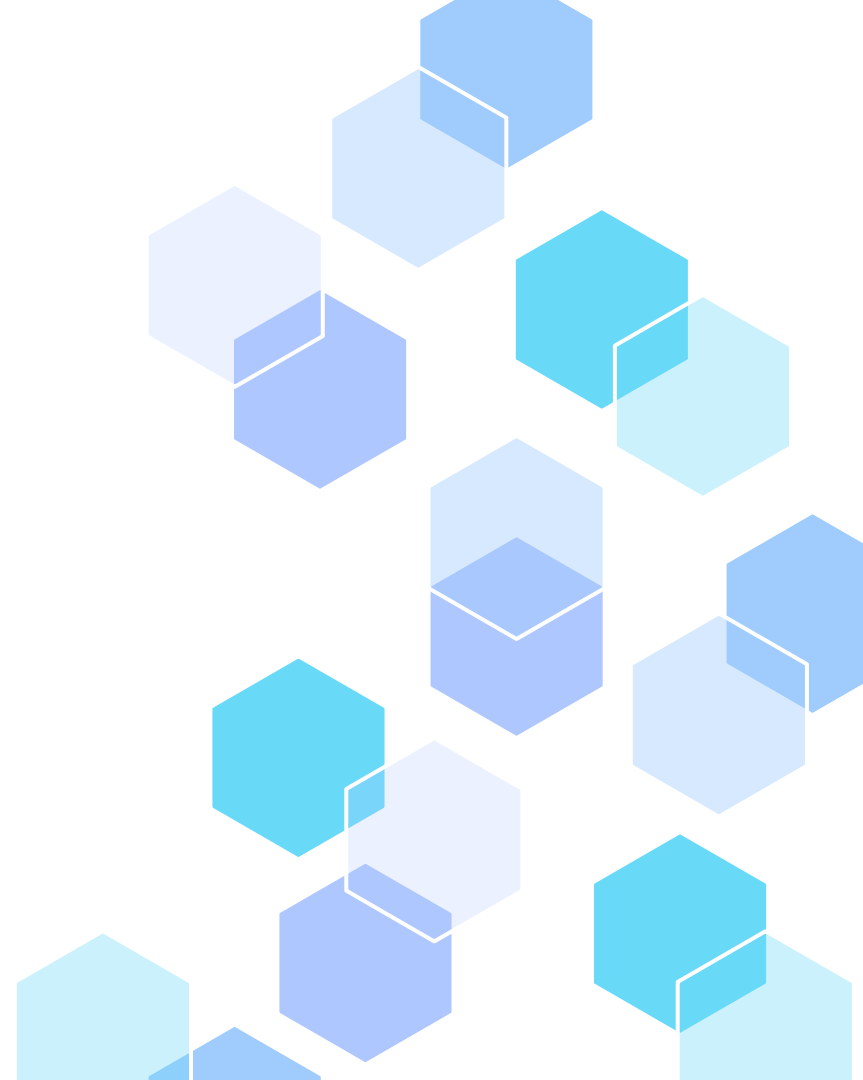


Ciência e Pesquisa

- **Análise** de Grandes Conjuntos de Dados (astrofísica, biologia molecular).
- Modelagem matemática para **prever** fenômenos físicos, biológicos e químicos.
- **Descoberta** de medicamentos através da análise de dados clínicos para identificar compostos promissores e simular testes de medicamentos.



PARTE 2 DO TUTORIAL - SEXTA-FEIRA





Agenda

01

Introdução à Análise
de Dados

02

Conceitos
Fundamentais

03

Ferramentas de
Análise de Dados

04

Processo de Análise
de Dados (Pipeline)

05

Áreas de Aplicação
da Análise de Dados

07

Práticas
Recomendadas

08

Montando o
Ambiente



06

Carreiras em Análise de Dados

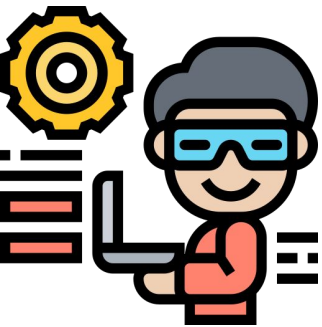


Data Analytics



- Responsável pela coleta, limpeza, análise e interpretação de dados para produzir relatórios e gerar *insights*.
- Excel, SQL, Python, ferramentas de visualização de dados.

Cientista de Dados



- Responsável pelo desenvolvimento de modelos preditivos, machine learning e análise exploratória avançada dos dados.
- Python, machine learning, deep learning, estatística, SQL, Hadoop, Spark.



Engenheiro de Dados

- Responsável pela construção e manutenção de infraestruturas para coleta, armazenamento e processamento de grandes volumes de dados
- Habilidades em programação, frameworks de machine learning, DevOps, cloud computing







07

Práticas Recomendadas



Planejamento e organização

- Dividir o projeto em pequenas tarefas e definir um cronograma
 - Ferramentas: Trello, Notion
- 
- 

Documentação



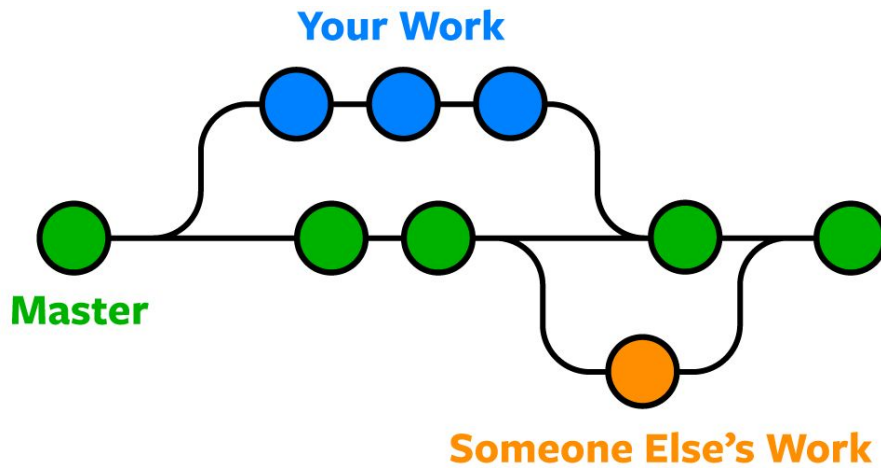
- Documentar o código e as análises é fundamental para que outros possam entender o projeto.
- Usar *Markdown* para comentários e explicações.

Controle de Versão



git

- Controle de versão permite acompanhar o histórico de mudanças no código e facilita o trabalho colaborativo.
- GitHub.



Ambientes Virtuais



- Um ambiente virtual é um espaço isolado no sistema onde é possível instalar dependências específicas para um projeto.
- Isolamento de dependências.
- Reprodutibilidade.
- Facilidade de manutenção, permitindo atualizar pacotes sem afetar outros pacotes.





08

**Montando o
Ambiente**

Montando o Ambiente Virtual

1. Crie o ambiente virtual

```
python3 -m venv nome_do_ambiente
```

2. Ative o ambiente virtual

```
source nome_do_ambiente/bin/activate
```

3. Instale o Jupyter Notebook no ambiente virtual

```
pip install jupyter
```

4. Adicione o ambiente virtual ao Jupyter Notebook

```
pip install ipykernel  
python -m ipykernel install --user --name=nome_do_ambiente
```

5. Inicie o Jupyter Notebook

```
jupyter notebook
```

Referências

- PROVOST, Foster; FAWCETT, Tom. Data science for business: what you need to know about data mining and data-analytic thinking. 1. ed. Sebastopol: O'Reilly Media, 2013.
- MONTGOMERY, Douglas C.; RUNGER, George C. Estatística aplicada e probabilidade para engenheiros. 6. ed. Rio de Janeiro: LTC, 2014.

Obrigado!

Alguma dúvida?

vitormoreiracasagrande@hotmail.com

thiago.silva@ufmt.br

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

