



# **Introdução à Ciência de Dados** **2025.1**

## *Aula 1 - Apresentação*

Prof. Dr. Thiago Pereira da Silva





# Ementa

Introdução à Inteligência de Dados. Processos de preparação, coleta e tratamento de dados. Modelagem de dados. Técnicas de visualização. Mineração de dados. Ferramentas de mineração de dados.

## REFERÊNCIA BÁSICA

- PROVOST, F.; FAWCETT, T. Data Science for Business: What you need to know about data mining and data analytic thinking. O'Reilly Media, 2013
- ZIKOPOULOS, P., & EATON, C. Understanding big data: Analytics for enterprise class hadoop and streaming data. Ed. McGraw. 2011.
- PRAJAPATI, V. Big Data Analytics with R and Hadoop. Packt Publishing Ltd. 2013.

## REFERÊNCIA COMPLEMENTAR

- HAN, J.; KAMBER, M. Data Mining – Concepts and Techniques. Morgan Kaufmann Publishers, 2001. ISBN 1558604898.
- TAN, P.-N.; Steinbach, M.; Kumar, T. Introduction to Data Mining. Addison Wesley, 2005.
- SILVA, L. A.; PERES, S. M.; BOSCAROLI, C. Introdução à Mineração de dados com aplicações em R, 1ª. Edição, Elsevier. 2013.
- BERRY, M. W. & KOGAN, J. Text mining: applications and theory. John Wiley. 2010.
- FACELI, K.; Lorena, A. C.; GAMA, J.; de CARVALHO, A. C. P. L. F. Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina. 1. ed. Rio de Janeiro: LTC. 2011



# Plano de Ensino

## Objetivo Geral

Capacitar os estudantes a compreender e aplicar os fundamentos da Ciência de Dados, abrangendo as etapas de coleta, preparação, tratamento, modelagem e visualização de dados, bem como a utilização de técnicas e ferramentas de mineração de dados, visando a extração de conhecimento útil para a tomada de decisões em diferentes contextos.

## Objetivos Específicos

- Compreender os **conceitos fundamentais** relacionados à Ciência de Dados, incluindo inteligência de dados, mineração de dados e visualização.
- Desenvolver habilidades para **coletar, preparar e tratar dados**, utilizando metodologias e ferramentas adequadas para garantir a qualidade e a integridade das informações.
- Aplicar técnicas de **modelagem de dados** para organizar e estruturar grandes volumes de dados, facilitando sua análise e interpretação.
- Explorar e utilizar **ferramentas de mineração de dados** para identificar padrões e *insights* valiosos a partir de conjuntos de dados complexos.
- Desenvolver capacidades de **visualização de dados**, utilizando diferentes técnicas e ferramentas para representar informações de forma clara e acessível.
- Estimular o **pensamento crítico e analítico**, incentivando a aplicação dos conhecimentos adquiridos em problemas reais e em contextos interdisciplinares.



# Avaliação

Em conformidade com a Resolução CONSEPE 63/18, a avaliação da disciplina será composta da seguinte forma:

- **Três trabalhos em equipe** (dupla ou trio), com **peso 3** (1 ponto cada), focados em Ciência de Dados, nos quais os estudantes aplicarão os conceitos e técnicas aprendidos a problemas do mundo real. Cada trabalho será uma oportunidade de resolver desafios práticos, como coleta, tratamento, análise e visualização de dados, promovendo uma aprendizagem baseada em situações reais e colaborativas.
- **Lista de exercícios,** com **peso 2.**
- **Duas provas individuais,** com **peso 5**, que avaliarão a compreensão teórica dos conteúdos abordados ao longo do curso, garantindo que os alunos tenham uma base sólida para a aplicação prática.



# Conteúdo Programático

- Introdução à Ciência de Dados
- Coleta e Preparação de Dados
- Análise Exploratória de Dados (EDA)
- Modelagem de Dados e Algoritmos
- Mineração de Dados
- Ferramentas e Plataformas de Ciência de Dados
- Projetos Práticos de Ciência de Dados

*A ordem dos tópicos pode mudar de acordo com o andamento das aulas.*



# Agenda

**01**

**Introdução à Ciência  
de Dados**

**02**

**Conceitos  
Fundamentais**

**03**

**Ferramentas de  
Ciência de Dados**

**04**

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

**05**

**Áreas de Aplicação  
da Ciência de Dados**

**06**

**Carreiras em Ciência  
de Dados**



---

**01**

# **Introdução à Ciência de dados**



# O que é Ciência 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 Ciência 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;**
- **Previsão de eventos futuros;**
- **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.



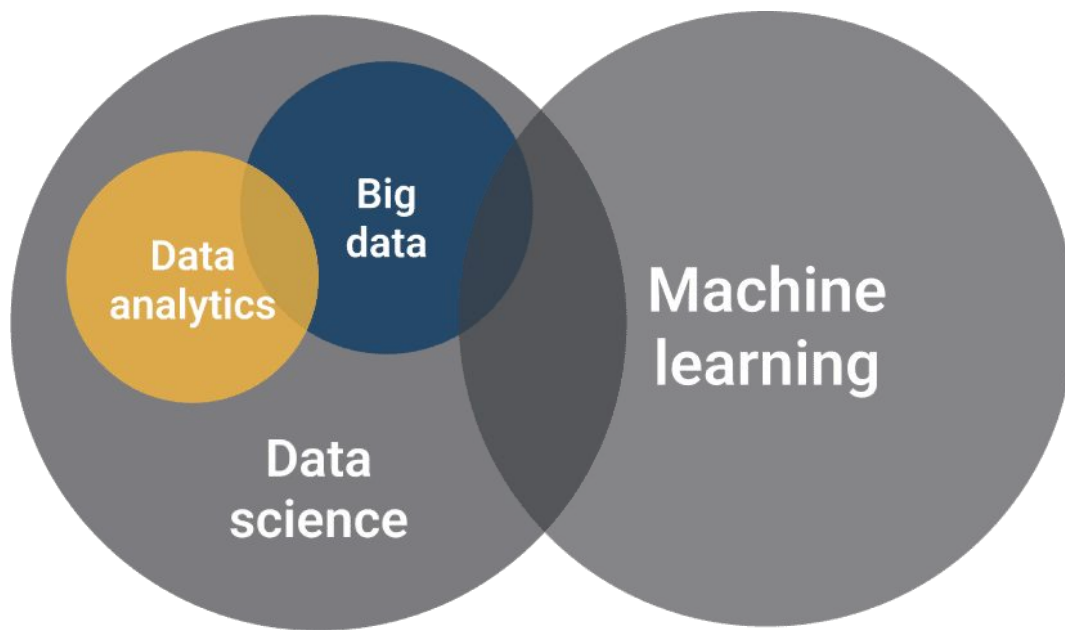
**Conhecimento**

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

# Etapas Gerais do Processo de Ciência de Dados



# Situando à Ciência de Dados



**Fonte:**

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

# Mineração vs Ciência de Dados

- **Mineração de Dados (Data Mining)**

- É um subconjunto da Ciência de Dados.
- Foca na **exploração e análise** de grandes volumes de dados para identificar padrões, relações e insights específicos.

- **Ciência de Dados (Data Science)**

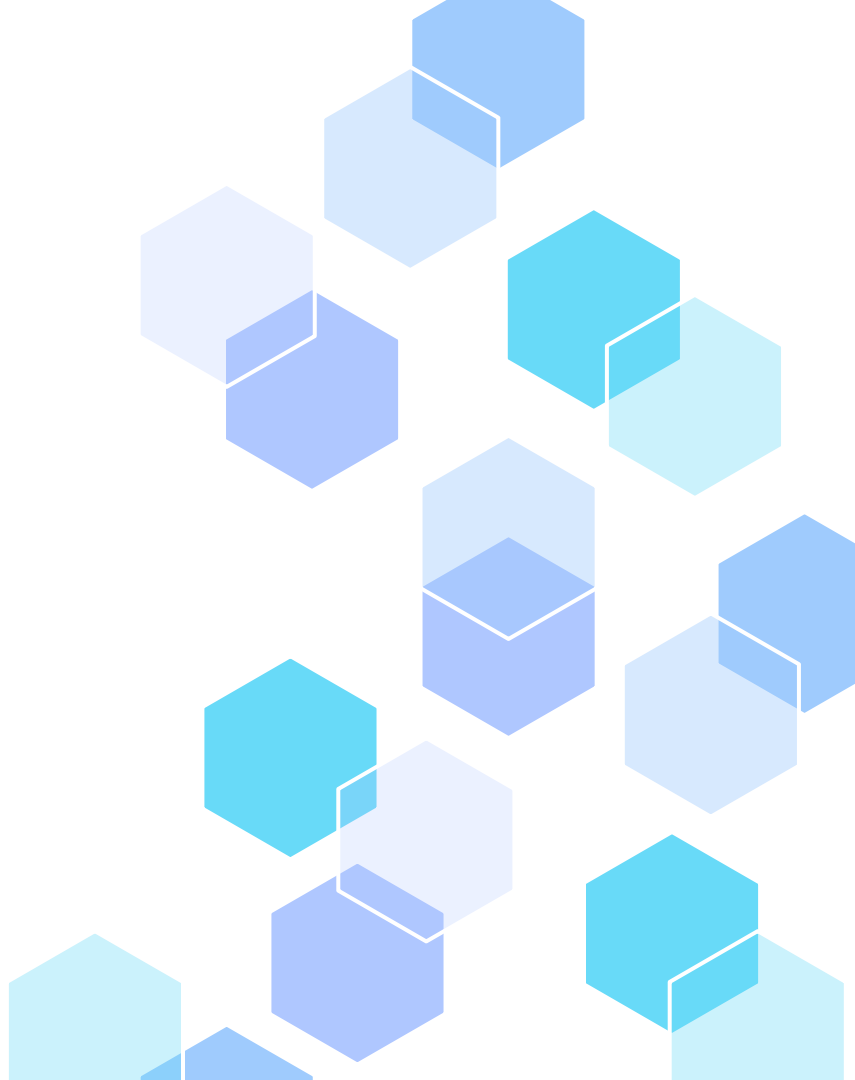
- Mais amplo e engloba todo o **ciclo de manipulação e análise de dados**, indo desde a coleta até a implementação de soluções baseadas em dados.

- Por vezes, a mineração de dados é uma etapa ou técnica usada na Ciência de Dados.

---

**02**

# **Conceitos Fundamentais**



# Tipos de Variáveis

## **Quantitativas** (escala quantitativa)

- 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))



# Exemplo: Tipos de Variáveis

Tabela: Registro de Eventos Escolares

ID do Evento	Data	Hora	Aluno	ID do Aluno	Tipo de Evento	Disciplina	Resultado (Nota ou Status)
1	05/12/2024	08:00	João Silva	A001	Presença	Matemática	Presente
2	05/12/2024	10:00	Maria Oliveira	A002	Teste	Português	8,5
3	05/12/2024	14:00	Carlos Pereira	A003	Entrega de Atividade	Ciências	Completa
4	05/12/2024	16:00	Ana Costa	A004	Chamada de Verificação	Geografia	Ausente

**Tamanho do Conjunto:** 4 registros (eventos) com 8 atributos cada.

**Formato:** Tabela (estruturada).



# Exemplo: Tipos de Variáveis

Atributo	Tipo Geral	Subtipo	Descrição
ID do Evento	Quantitativo	Discreto	Identificador único para cada evento, expresso como um número inteiro.
Data	Qualitativo	Nominal	Representa o dia em que o evento ocorreu, sem hierarquia, mas categórico.
Hora	Qualitativo	Ordinal	Horário do evento, com sequência temporal implícita, indicando uma ordem.
Aluno	Qualitativo	Nominal	Nome do aluno participante, sem ordem lógica.
ID do Aluno	Qualitativo	Nominal	Código único atribuído ao aluno, usado como identificador.
Tipo de Evento	Qualitativo	Nominal	Categoria que descreve o tipo de evento (ex.: "Presença", "Teste"), sem ordem hierárquica.
Disciplina	Qualitativo	Nominal	A matéria associada ao evento, como "Matemática", "Português", etc., sem hierarquia.
Resultado (Nota/Status)	Quantitativo/Qualitativo	Contínuo/Nominal ↓	Pode ser numérico (nota de um teste, como 8,5) ou categórico (presença, completado), dependendo do evento.



# Exemplo: Tipos de Variáveis

## Aplicação dos Dados

- **Análise de Frequência:** Identificar padrões de presença dos alunos.
- **Desempenho Acadêmico:** Avaliar notas e participação em atividades.
- **Relatórios Personalizados:** Gerar resumos para pais, professores ou gestores escolares.
- **Deteção de Problemas:** Monitorar alunos com baixa participação ou desempenho.

## Explicações dos Atributos:

- **Quantitativos:** Representam informações numéricas que podem ser usadas para cálculos (como soma, média, etc.). São discretos (valores inteiros) ou contínuos (valores que podem ter casas decimais, como notas).
- **Qualitativos:** Representam categorias ou atributos que não têm valor numérico, sendo úteis para classificações e agrupamentos. Podem ser:
  - **Nominais:** Sem uma ordem ou hierarquia (como "Aluno" ou "Tipo de Evento").
  - **Ordinais:** Possuem uma ordem ou classificação (como "Hora", onde um horário pode ser sequencialmente maior ou menor que outro).

# 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 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

$\Delta$  = 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)

# Exemplo: Estatística Descritiva

ID do Evento	Data	Hora	Aluno	ID do Aluno	Tipo de Evento	Disciplina	Resultado (Nota ou Status)
1	01/12/2024	08:00	João Silva	A001	Presença	Matemática	Presente
2	01/12/2024	10:00	Maria Oliveira	A002	Teste	Português	7,0
3	01/12/2024	14:00	Carlos Pereira	A003	Entrega de Atividade	Ciências	Completa
4	01/12/2024	16:00	Ana Costa	A004	Chamada de Verificação	Geografia	Ausente
5	02/12/2024	08:00	João Silva	A001	Teste	Matemática	8,5
6	02/12/2024	10:00	Maria Oliveira	A002	Entrega de Atividade	Português	Completa
7	02/12/2024	14:00	Carlos Pereira	A003	Presença	Ciências	Presente
8	02/12/2024	16:00	Ana Costa	A004	Teste	Geografia	6,5
9	03/12/2024	08:00	João Silva	A001	Chamada de Verificação	Matemática	Presente
10	03/12/2024	10:00	Maria Oliveira	A002	Teste	Português	9,0
11	03/12/2024	14:00	Carlos Pereira	A003	Entrega de Atividade	Ciências	Incompleta
12	03/12/2024	16:00	Ana Costa	A004	Presença	Geografia	Presente

# Exemplo: Estatística Descritiva

Aluno	Número de Participações	Média de Notas	Número de Atividades/Provas Entregues
João Silva	3	8,25	2
Maria Oliveira	4	8,25	2
Carlos Pereira	3	6,5	1
Ana Costa	3	7,0	1

## Resumo das Medidas Estatísticas:

Métrica	Resultado
Média de Notas	7,75
Mediana de Notas	7,625
Moda de Notas	Não aplicável
Desvio Padrão de Notas	0,82
Correlação (Frequência de Participação e Notas)	0,45 (moderada positiva)

# Quartis e Percentis

## Quartis

Dividem um conjunto de dados ordenado em quatro partes iguais.

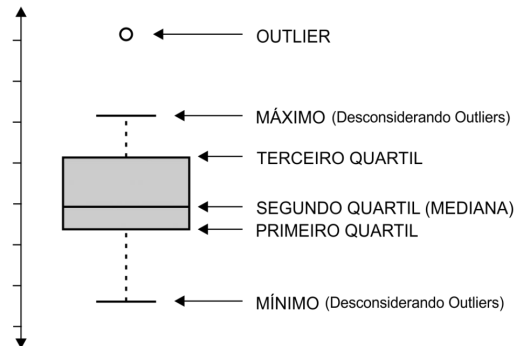
- **Q1 (Primeiro Quartil):** O valor que separa os 25% menores dados.
- **Q2 (Segundo Quartil ou Mediana):** O valor que separa os 50% dos dados (mediana).
- **Q3 (Terceiro Quartil):** O valor que separa os 75% menores dados.

## 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



# Quartis e Percentis

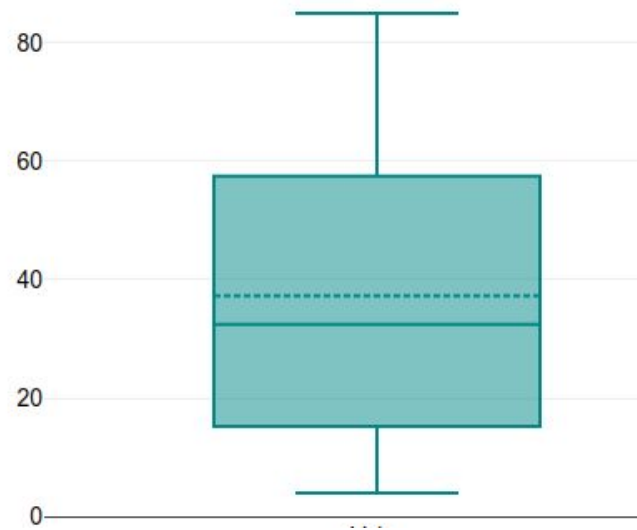
## Quartis

### Exemplo:

Considere o conjunto de dados ordenado:

[4, 8, 15, 16, 23, 42, 50, 60, 70, 85]

- **Q1 (primeiro quartil)** - os primeiros 25% dos dados:
  - Q1 está na posição 2.5 (média entre os valores da posição 2 e 3):  
 $Q1 = 8 + 15/2 = 11.5$
- **Q2 (mediana ou segundo quartil)** - 50% dos dados estão abaixo dele:
  - Q2 está na posição 5.5 (média entre os valores da posição 5 e 6):  
 $Q2 = 23 + 42/2 = 32.5$
- **Q3 (terceiro quartil)** - 75% dos dados:
  - Q3 está na posição 8.5 (média entre os valores da posição 8 e 9):  
 $Q3 = 60 + 70/2 = 65.0$

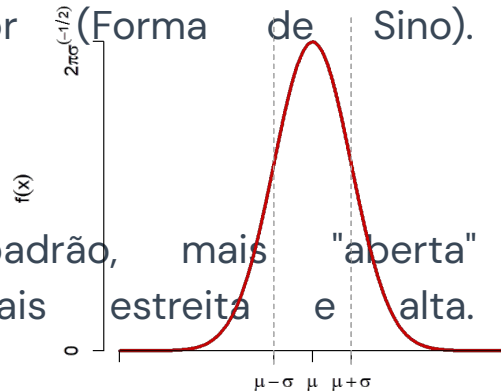


# Distribuições

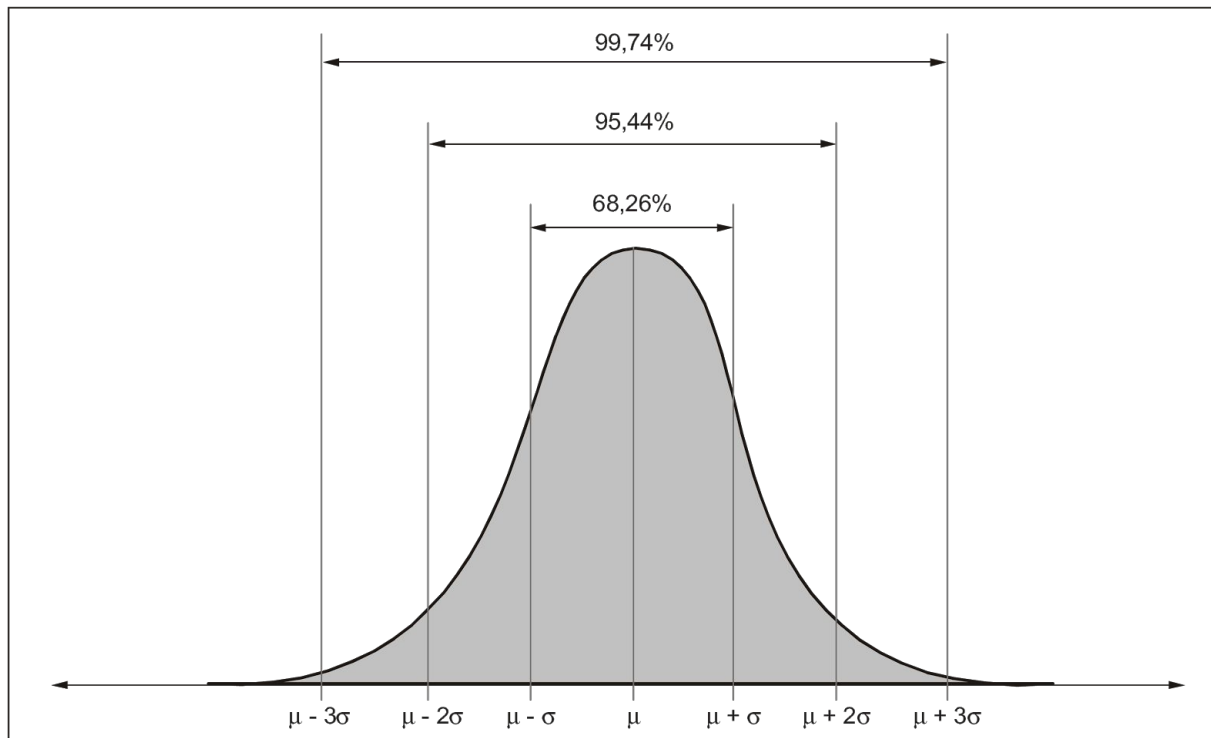
- Descreve como os valores dessa variável estão distribuídos ou distribuídos ao longo de um conjunto de dados ou de uma população.
  - Indica a **frequência ou probabilidade de ocorrência** de diferentes valores ou intervalos de valores para essa variável.
- **Exemplo:** Ao observar a variável *peso* dos alunos do curso é fácil notar que o valor assumido pela variável **varia** de aluno para aluno. Os dados dos pesos dos alunos apresentam variabilidade. O padrão desta variabilidade é a distribuição da variável.

# Distribuição Normal (Gaussiana)

- Descreve fenômenos em que a maior parte dos dados se concentra em torno de um **valor central** e diminui conforme se afastam desse valor (Forma de Sino).
- É descrita por dois **Parâmetros**:
  - Média ( $\mu$ ): onde a maior parte dos dados está concentrada.
  - Desvio padrão ( $\sigma$ ): quanto maior o desvio padrão, mais "aberta" e "achatada" será a curva; quanto menor, mais estreita e alta.
- **Propriedades**:
  - **Simetria**: A distribuição normal é simétrica em torno de sua média. Ou seja, os valores à esquerda da média têm a mesma distribuição que os valores à direita.
  - **Regra empírica (68-95-99.7)**: Aproximadamente:
    - 68% dos dados estão dentro de 1 desvio padrão da média ( $\mu \pm \sigma$ ),
    - 95% dos dados estão dentro de 2 desvios padrões ( $\mu \pm 2\sigma$ ),
    - 99.7% dos dados estão dentro de 3 desvios padrões ( $\mu \pm 3\sigma$ ).



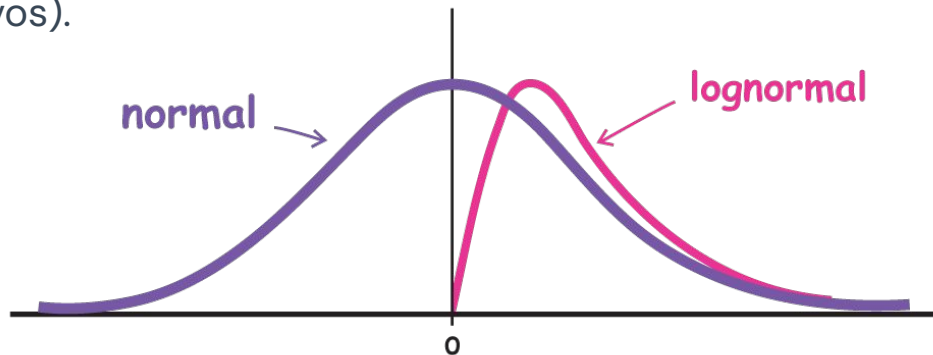
# Distribuição Normal (Gaussiana)





# Distribuição Lognormal

- Distribuição de probabilidade contínua que descreve variáveis cujos **logaritmos seguem uma distribuição normal**.
  - Se uma variável aleatória  $X$  segue uma distribuição lognormal, então  $\ln(X)$  (o logaritmo natural de  $X$ ) segue uma distribuição normal.
- A distribuição lognormal tem uma forma assimétrica (não simétrica), com um longo "rabo" à direita.
  - Isso a torna útil para modelar variáveis que têm uma tendência a crescer exponencialmente ou que são naturalmente limitadas em uma direção (tipicamente, valores positivos).





03

# Ferramentas de Ciência 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/>

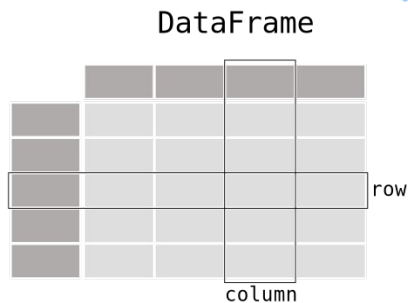
# Pandas



- 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



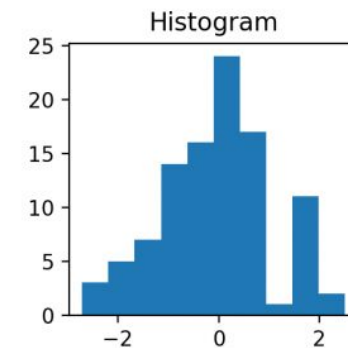
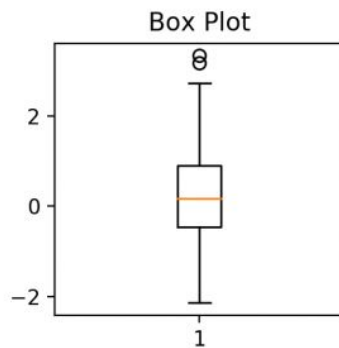
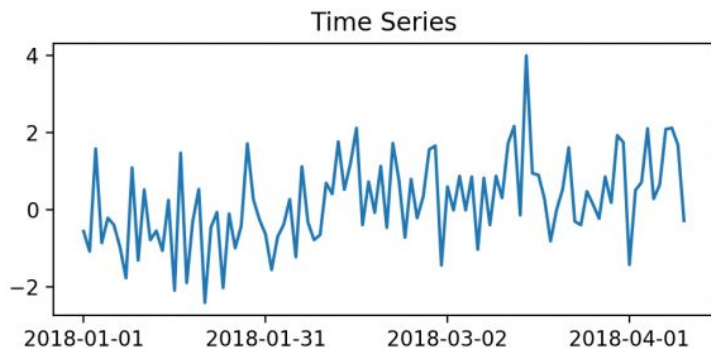
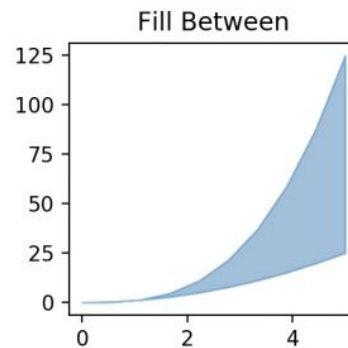
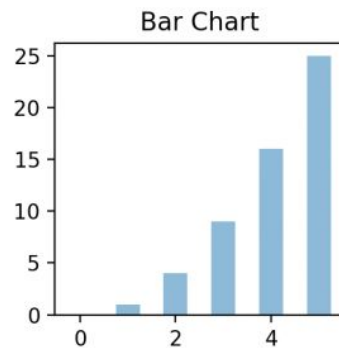
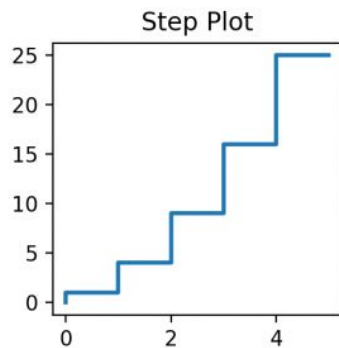
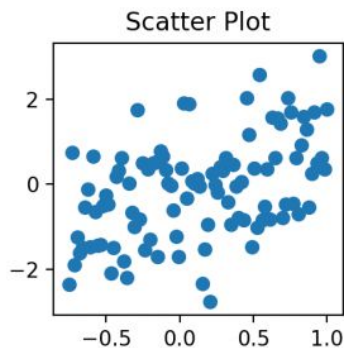
Series 1		Series 2		Series 3		Dataframe			
INDEX	DATA	INDEX	DATA	INDEX	DATA	INDEX	SERIES 1	SERIES 2	SERIES 3
0	A	0	1	0	[1, 2]	0	A	1	[1, 2]
1	B	1	2	1	A	1	B	2	A
2	C	2	3	2	1	2	C	3	1
3	D	3	4	3	(4, 5)	3	D	4	(4, 5)
4	E	4	5	4	{"a": 1}	4	E	5	{"a": 1}
5	F	5	6	5	6	5	F	6	6

# 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







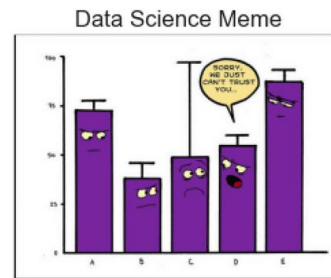
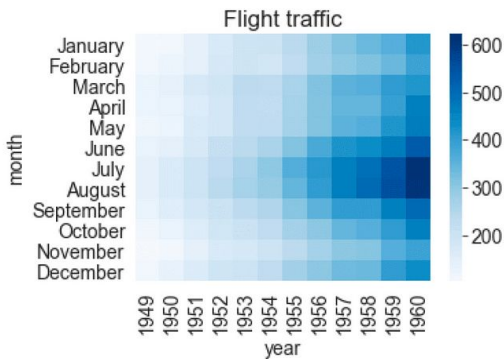
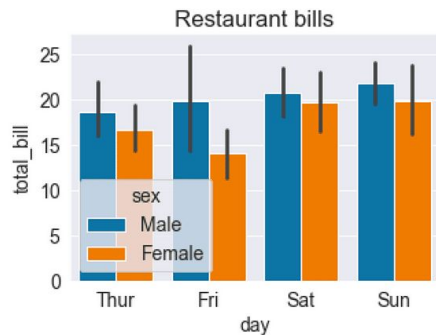
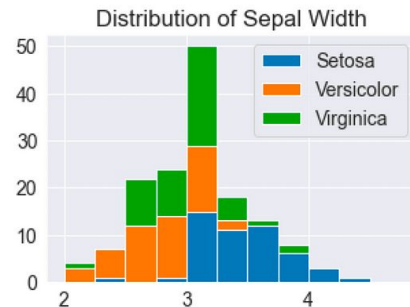
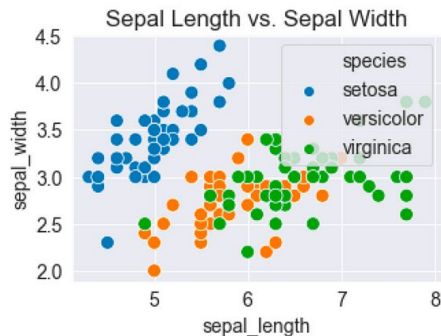
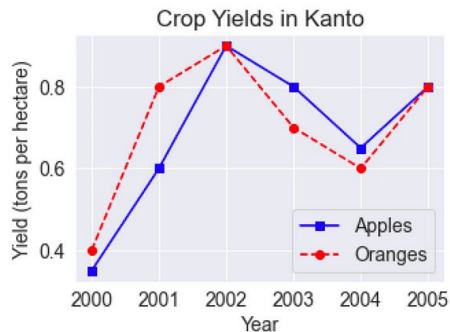
# seaborn

- 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

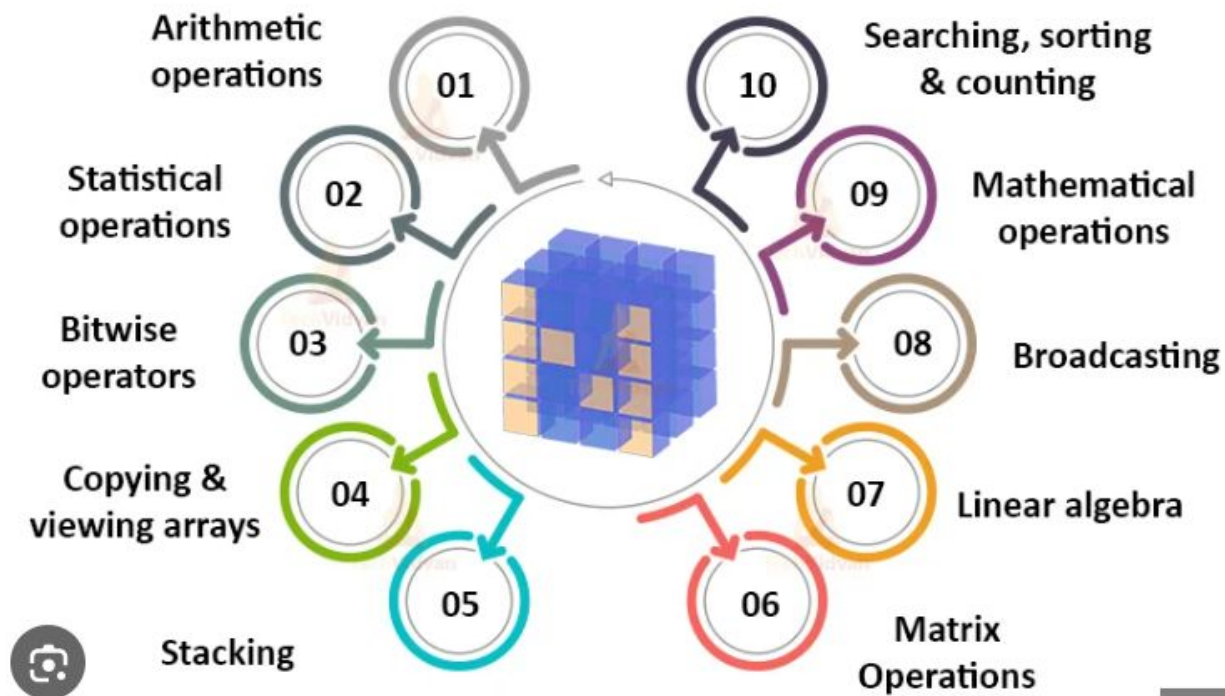




- Biblioteca de código aberto para computação numérica.
- Fornece suporte para arrays e matrizes multidimensionais, além de funções matemáticas avançadas.
- Escrito em C para performance, o NumPy é extremamente rápido em comparação com listas Python
  - Especialmente em operações com grandes conjuntos 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/>

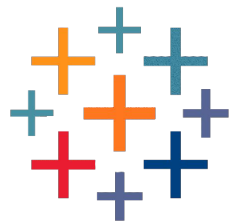
<https://jupyter.org/try-jupyter/lab/>

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/>



# + a b l e a u

- Ferramenta para visualização de dados;
- Criação de dashboards interativos.

<https://www.tableau.com/>



# Power BI

- Visualização de dados corporativos e relatórios.
  - <https://app.powerbi.com/>





# Machine learning

- Scikit-learn: Biblioteca para aprendizado de máquina em Python.
- TensorFlow/Keras: Frameworks para deep learning.
- PyTorch: Popular para redes neurais.
- XGBoost/LightGBM: Modelos de aprendizado de máquina baseados em árvores.
- River Framework (*online machine learning*)



# Big Data e Armazenamento



- Apache Hadoop: Processamento de grandes volumes de dados.
  - Apache Spark: Processamento em tempo real.
  - Google BigQuery: Análise de dados em grande escala.
- 
- 

# Deploy e Automação

- Docker: Para criar contêineres de aplicações.
- Kubernetes: Orquestração de contêineres.
- Streamlit/Flask/Django: Para criar dashboards e APIs.

---

**04**

# **Processo de Ciência de Dados**

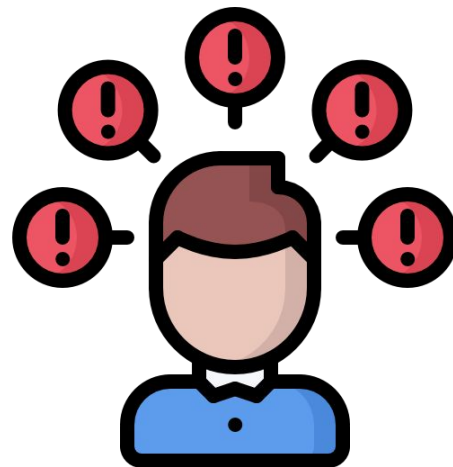


# Etapas de Ciência de Dados

1. Definição do Problema
2. Coleta de Dados
3. Limpeza de Dados
4. Exploração e Visualização Inicial
5. Análise Exploratória e Modelagem
6. Interpretação e Apresentação de Resultados
7. Monitoramento e Manutenção

# Definição do Problema:

- **O que é:** Identificar claramente o objetivo do projeto e o problema que se busca resolver.
- **Perguntas-chave:** Qual é o impacto esperado? Quais são os resultados desejados?
- **Exemplo:** Prever a rotatividade de clientes em uma empresa.



# 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.





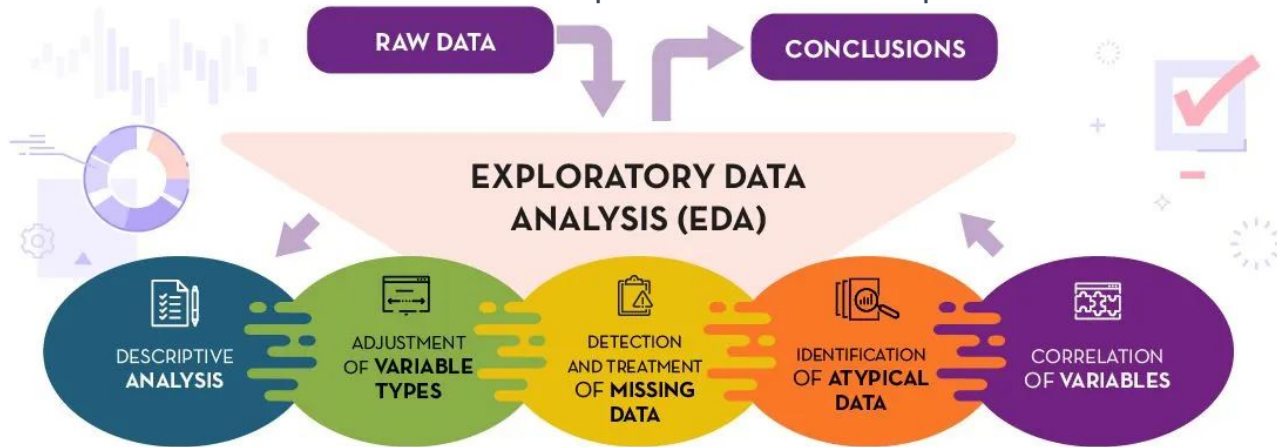
# Exploração e Visualização Inicial

- 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 estatísticos ou de aprendizado de máquina.



# 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).

# Monitoramento e Manutenção



- **O que é:** Acompanhar o desempenho do modelo ao longo do tempo e ajustá-lo conforme necessário.
- **Considerações:** Dados novos podem mudar o desempenho do modelo.
- **Exemplo:** Re-treinar o modelo mensalmente com novos dados de clientes.

# Vamos ler um material interessante!



**Afinal, como se desenvolve um projeto de Data Science?**

<https://medium.com/techbloghotmart/afinal-como-se-desenvolve-um-projeto-de-data-science-233472996c34>

---

**05**

# **Áreas de Aplicação da Ciência de Dados**



# Negócios e Marketing



- **Segmentação de Clientes (exemplo)**
- Previsão de Vendas
- Análise de Sentimento

# Exemplo: Segmentação de Clientes

## Problema:

Uma empresa de e-commerce quer segmentar seus clientes para criar campanhas de marketing mais eficazes. Atualmente, ela envia as mesmas promoções para todos, mas percebe que isso não gera resultados satisfatórios.

## Etapas do Processo de Ciência de Dados na Segmentação:

### 1. Definição do Problema:

- **Objetivo:** Identificar grupos de clientes com base em comportamentos de compra e características demográficas para campanhas personalizadas.
- **Perguntas:** Quais fatores diferenciam os clientes? Quais produtos ou serviços são mais atrativos para cada grupo?

### 2. Coleta de Dados:

- **Fontes:**
  - Dados transacionais: histórico de compras (frequência, valor, tipo de produto).
  - Dados demográficos: idade, localização, gênero.
  - Dados comportamentais: frequência de visita ao site, uso de cupons de desconto.
- **Exemplo de dados coletados:**
  - Cliente A: 25 anos, comprou 5 vezes no último mês, gasta em média R\$ 200 por compra.
  - Cliente B: 40 anos, comprou 1 vez no último mês, gasta R\$ 500 por compra.



# Exemplo: Segmentação de Clientes

(continuação)

## 3. Preparação de Dados:

- Limpar dados ausentes (ex.: preencher ou excluir registros sem gênero informado).
- Normalizar variáveis, como escalas de gasto mensal e frequência de compras.
- Criar variáveis derivadas, como "tempo médio entre compras".

## 4. Análise Exploratória de Dados (EDA):

- Identificar padrões, como:
  - Clientes jovens compram mais itens baratos e frequentemente.
  - Clientes mais velhos compram menos vezes, mas gastam mais por compra.

## 5. Modelagem:

- **Algoritmo:** Clusterização não supervisionada, como K-means.
- **Entrada:** Variáveis como idade, gasto médio, frequência de compras.
- **Saída:** Agrupamento de clientes em clusters (ex.: 3 grupos).
  - Cluster 1: Jovens com baixo gasto e alta frequência.
  - Cluster 2: Adultos com gasto moderado e frequência média.
  - Cluster 3: Clientes premium com gasto alto e baixa frequência.

# Exemplo: Segmentação de Clientes

(continuação)

## 6. Avaliação:

- Validar a consistência dos clusters usando métricas como **silhouette score**.
- Verificar se os clusters fazem sentido para o negócio.

## 7. Implementação:

- Desenvolver campanhas específicas:
  - **Cluster 1:** Ofertas relâmpago para atrair mais compras.
  - **Cluster 2:** Oferecer frete grátis para aumentar a frequência de compras.
  - **Cluster 3:** Promoções exclusivas ou programas de fidelidade.

## 8. Monitoramento e Manutenção:

- Atualizar os clusters regularmente com novos dados de clientes.
- Medir os resultados das campanhas (ex.: aumento na taxa de conversão).

# Finanças e Bancos

- **Detecção de Fraudes (exemplo)**
- Análise de Crédito
- Gestão de Riscos
- Entendimento do Mercado



# Exemplo: Detecção de Fraudes

## Problema:

Um banco deseja implementar um sistema automatizado para detectar transações fraudulentas em tempo real, minimizando perdas financeiras e protegendo os clientes.

## Etapas do Processo de Ciência de Dados na Detecção de Fraudes:

### 1. Definição do Problema:

- **Objetivo:** Identificar e bloquear transações potencialmente fraudulentas antes que sejam processadas.
- **Perguntas:**
  - Quais características indicam uma transação fraudulenta?
  - Como equilibrar precisão e tempo de resposta?

### 2. Coleta de Dados:

- **Fontes:**
  - Dados históricos de transações.
  - Informações sobre clientes (idade, localização, histórico de crédito).
  - Dados externos, como padrões de fraudes conhecidas.
- **Exemplo de dados coletados:**
  - Valor da transação, localização, dispositivo utilizado, hora do dia.
  - Status de transação (fraudulenta ou legítima).

# Exemplo: Detecção de Fraudes

(continuação)

## 3. Preparação de Dados:

- Limpeza de valores ausentes e inconsistências nos dados.
- Normalização de variáveis (ex.: valores monetários, localização geográfica).
- Criação de novas variáveis:
  - Tempo médio entre transações.
  - Distância geográfica entre transações consecutivas.
  - Desvios em padrões usuais de gasto.

## 4. Análise Exploratória de Dados (EDA):

- Identificar padrões:
  - Transações de valor alto feitas em dispositivos desconhecidos têm maior probabilidade de serem fraudulentas.
  - Mudanças bruscas no local de acesso podem ser indicativos de fraude.

# Exemplo: Detecção de Fraudes

(continuação)

## 5. Modelagem:

- **Algoritmos comuns:**
  - Classificação supervisionada: Random Forest, Gradient Boosting (XGBoost, LightGBM).
  - Detecção não supervisionada: Isolation Forest, Autoencoders (para identificar anomalias).
- **Entrada:** Variáveis como valor, localização, dispositivo, hora.
- **Saída:** Probabilidade de fraude.
- **Exemplo:**
  - Transação A: probabilidade de fraude = 95% (bloqueada).
  - Transação B: probabilidade de fraude = 10% (processada).

## 6. Avaliação:

- **Métricas:**
  - Precisão: porcentagem de fraudes corretamente identificadas.
  - Recall: porcentagem de fraudes identificadas em relação ao total de fraudes.
  - F1-score: equilíbrio entre precisão e recall.
- **Exemplo:** O modelo atinge 90% de precisão e 85% de recall.

# Exemplo: Detecção de Fraudes

(continuação)

## 7. Implementação:

- Integração com sistemas de pagamento do banco.
- Configuração para bloquear automaticamente transações com alta probabilidade de fraude e notificar o cliente.
- Monitoramento em tempo real.

## 8. Monitoramento e Manutenção:

- Atualizar o modelo com novos padrões de fraude.
- Ajustar o modelo para evitar falsos positivos (transações legítimas bloqueadas).

## Impacto:

- Redução de prejuízos financeiros causados por fraudes.
- Aumento da confiança dos clientes no sistema bancário.
- Processos mais eficientes e menos dependência de revisões manuais.

# Saúde

- Diagnóstico Precoce e Prognóstico
- Pesquisa de Genética e Genômica
- **Monitoramento de Pacientes (exemplo)**
- Planejamento de Saúde Pública





# Exemplo: Monitoramento de Pacientes

## Problema:

Um hospital deseja monitorar pacientes com doenças cardíacas para identificar sinais precoces de arritmia ou outros problemas, permitindo intervenções rápidas e reduzindo o risco de internações emergenciais.

## Etapas do Processo de Ciência de Dados no Monitoramento de Pacientes:

### 1. Definição do Problema

- **Objetivo:** Detectar anomalias em sinais vitais que possam indicar problemas cardíacos.
- **Perguntas:**
  - Quais padrões de sinais vitais estão associados a complicações?
  - Como integrar o monitoramento em tempo real com alertas para médicos?

### 2. Coleta de Dados

- **Fontes:**
  - Dados de dispositivos vestíveis (smartwatches, monitores cardíacos).
  - Registros hospitalares: histórico médico, medicamentos prescritos.
  - Sensores IoT: dispositivos que monitoram frequência cardíaca, pressão arterial, saturação de oxigênio, ECG.
- **Exemplo de dados coletados:**
  - Frequência cardíaca: 72 bpm.
  - Pressão arterial: 120/80 mmHg.
  - ECG: padrão normal/sinais de arritmia.

# Exemplo: Monitoramento de Pacientes

(continuação)

## 3. Preparação de Dados

- **Etapas:**
  - Tratamento de dados ausentes, como lacunas em medições.
  - Sincronização temporal dos dados de diferentes dispositivos.
  - Normalização para lidar com variações em escalas (ex.: valores de ECG vs. frequência cardíaca).

## 4. Análise Exploratória de Dados (EDA)

- Identificar padrões nos dados de saúde:
  - Aumento repentino na frequência cardíaca pode preceder episódios de fibrilação atrial.
  - Queda na saturação de oxigênio pode ser um sinal precoce de insuficiência respiratória.
- Visualização de séries temporais para identificar tendências e anomalias.

# Exemplo: Monitoramento de Pacientes

(continuação)

## 5. Modelagem

- **Algoritmos comuns:**
  - **Classificação supervisionada:** Redes neurais ou SVM para prever condições específicas.
  - **Análise de séries temporais:** Modelos como LSTM (redes neurais recorrentes) para prever anomalias em sinais vitais.
  - **Deteção de anomalias:** Isolation Forest ou Autoencoders para identificar padrões atípicos.
- **Entrada:** Dados como frequência cardíaca, ECG, pressão arterial.
- **Saída:** Alerta de risco, como "Sinais de arritmia detectados, consulte um médico imediatamente."
- **Exemplo:**
  - Paciente A: "Normal."
  - Paciente B: "Alerta! Frequência cardíaca 150 bpm com ECG irregular."

# Exemplo: Monitoramento de Pacientes

(continuação)

## 6. Avaliação

- **Métricas:**
  - Precisão e recall para prever condições de risco corretamente.
  - Sensibilidade do sistema a pequenos desvios nos sinais vitais.
- **Exemplo:** O modelo identifica 95% dos eventos de arritmia com 90% de precisão.

## 7. Implementação

- Integração com aplicativos móveis para pacientes e sistemas hospitalares.
- Notificações automáticas para médicos em casos críticos.
- Painéis de monitoramento centralizados no hospital para acompanhamento em tempo real.

# Exemplo: Monitoramento de Pacientes

(continuação)

## 8. Monitoramento e Manutenção

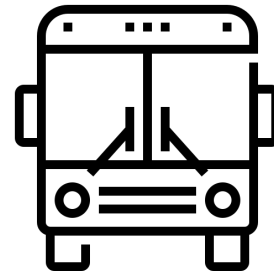
- Atualizar o modelo com novos dados de pacientes.
- Ajustar parâmetros com base no feedback médico para melhorar a precisão.
- Implementar medidas para lidar com alarmes falsos positivos.

## Impacto do Projeto:

- Redução de complicações graves devido à intervenção precoce.
- Melhora na qualidade de vida de pacientes com condições crônicas.
- Maior eficiência no uso de recursos hospitalares, priorizando casos urgentes.

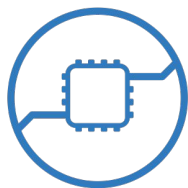
# Setor Público e Governamental

- 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.





# Agenda

**01**

Introdução à Ciência  
de Dados

**02**

Conceitos  
Fundamentais

**03**

Ferramentas de  
Ciência de Dados

**04**

Processo de Ciência  
de Dados (Pipeline)

**05**

Áreas de Aplicação  
da Ciência de Dados

**07**

Práticas  
Recomendadas

**08**

Montando o  
Ambiente





---

**06**

# **Carreiras em Ciência de Dados**

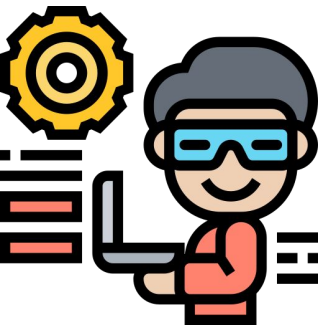


# Data Analytics



- 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



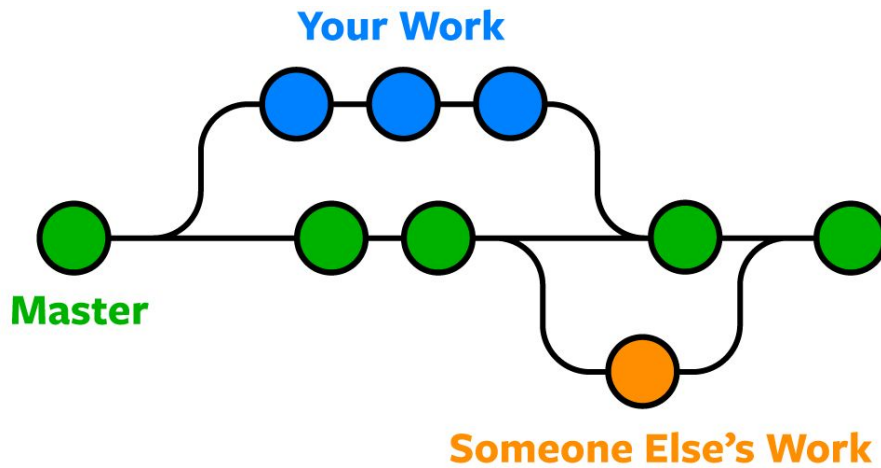
- Documentar o código e as análises é fundamental para que outros possam entender o projeto.
- Usar *Markdown* para comentários e explicações.

# 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
```

# Exercício

## Caracterização e Pré-processamento dos Dados

1. Considere o histórico de cotações do preço da soja em grãos (saca de 60 kg) no estado do Mato Grosso, fornecido pelo site Agrolink. Faça a caracterização inicial dos tipos dos atributos e avalie as Medidas de Localização (Centralidade), Dispersão (Variabilidade) e Associação, se aplicáveis.  
<https://www.agrolink.com.br/cotacoes/historico/mt/soja-em-grao-sc-60kg>
2. Considere o dataset sobre preços de **Produtos Agrícolas**, fornecido no arquivo Anexo, faça o que se pede:
  - a. Faça a caracterização inicial dos tipos dos atributos e avalie as medidas de Localização (Centralidade), Dispersão (Variabilidade) e Associação, se aplicáveis.
  - b. Analise o espalhamento das observações dos atributos de preços de soja e milho. Utilize técnicas gráficas como Boxplots para verificar a presença de outliers.
  - c. Faça a análise de correlação e covariância entre os atributos de preços da soja e Milho.

# 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.
- ESCOVEDO, Tatiana; MARQUES, Thiago; KALINOWSKI, Marcos. Introdução à Estatística para Ciência de Dados. São Paulo: Casa do Código, 2020. Disponível em: <https://www.casadocodigo.com.br>. Acesso em: 5 dez. 2024.
- ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. Introdução a Data Science: Algoritmos de Machine Learning e Métodos de Análise. São Paulo: Casa do Código, 2020. Disponível em: <https://www.casadocodigo.com.br>. Acesso em: 5 dez. 2024.

---

# Obrigado!

Alguma dúvida?

thiago.silva@ufmt.br

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

