
Estatística para Ciência de Dados



Profa. Rebeca Valgueiro

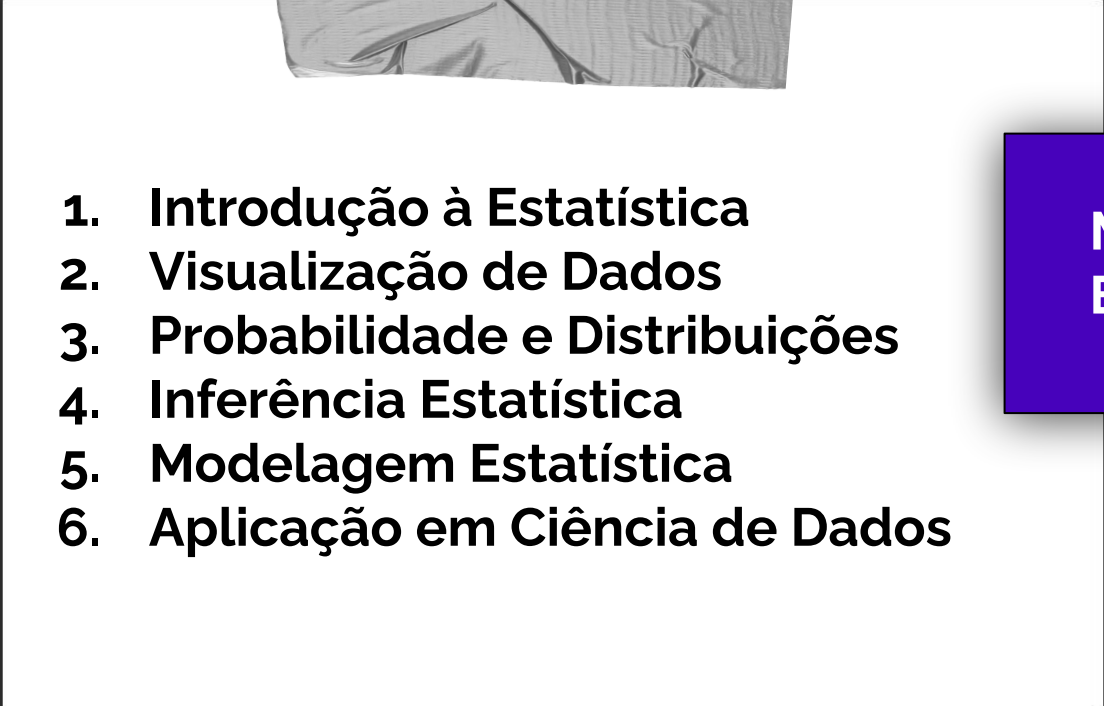
Quem sou eu?

- Graduada em Engenharia Civil
- MBA em Gestão Empresarial
- Trabalho a mais de 4 anos no mercado de tecnologia atuando em projetos de:
 - Desenvolvimento web
 - Desenvolvimento desktop- windows
 - Data Science
 - IA e visão computacional



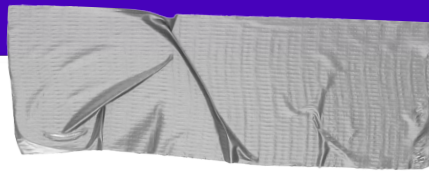
<https://www.linkedin.com/in/rebecavalgueiro/>

— O que vamos aprender?

- 
1. Introdução à Estatística
 2. Visualização de Dados
 3. Probabilidade e Distribuições
 4. Inferência Estatística
 5. Modelagem Estatística
 6. Aplicação em Ciência de Dados



**MUITOS
EXERCÍCIOS**



MATERIAIS DE APOIO

<https://www.w3schools.com/statistics/index.php>

<https://www.kaggle.com/datasets>

E o ChatGPT, Gemini....?



O que é
estatística?



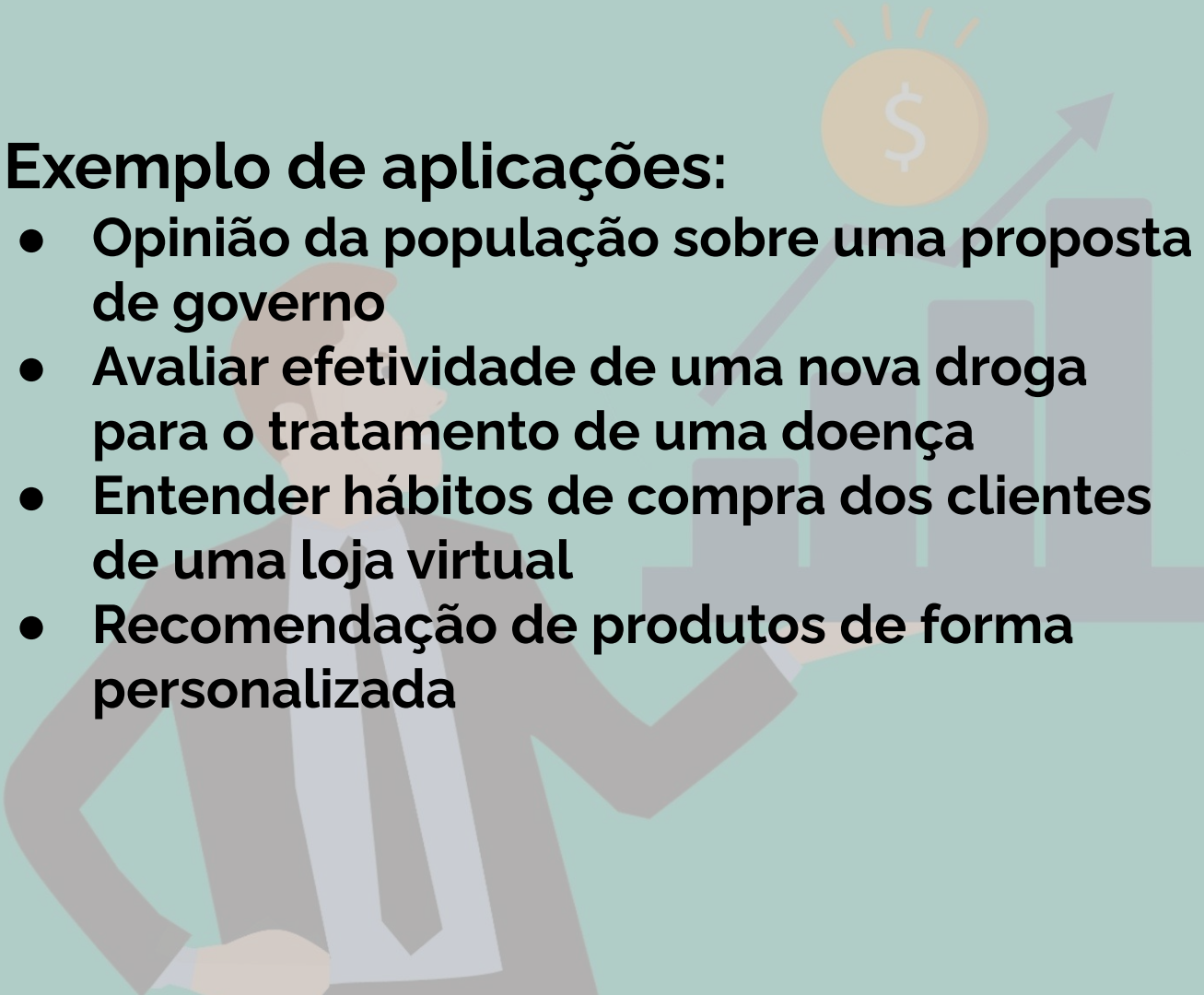
O que é estatística?



Ciência que lida com coleta, análise, interpretação e apresentação dos dados. Auxilia no processo de tomada de decisões.

Exemplo de aplicações:

- Opinião da população sobre uma proposta de governo
- Avaliar efetividade de uma nova droga para o tratamento de uma doença
- Entender hábitos de compra dos clientes de uma loja virtual
- Recomendação de produtos de forma personalizada



E como conseguimos ter acesso aos dados?

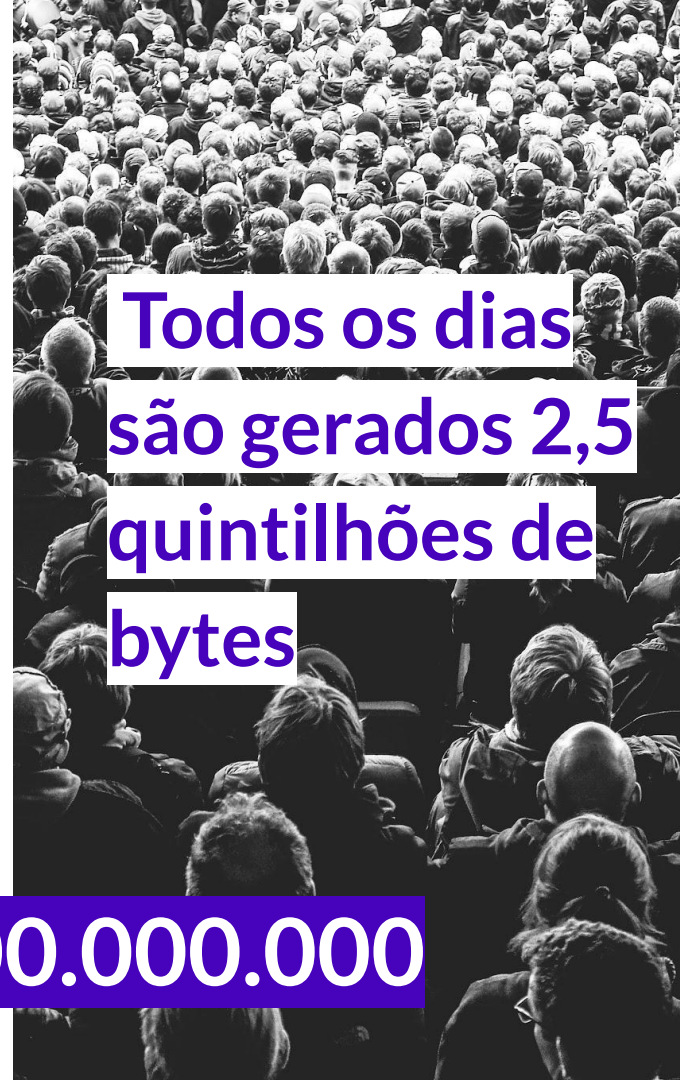
- Censo demográfico brasileiro (1872)
- Índices financeiros (todo dia)
- Pesquisas eleitorais

Desenvolvimento de tecnologias

- Crescente número de dispositivos inteligentes (smartphone, smartwatch, sensores, cctv...)
- Maior variedade de dados (textos, fotos, imagens, áudios, vídeos, redes de relacionamento, localização, UX...)
- Dados coletados a todo momento (ao vivo)

2.500.000.000.000.000.000.000

Todos os dias
são gerados 2,5
quintilhões de
bytes



BIG DATA

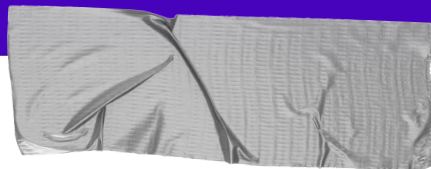
É o processo de coleta, armazenagem, organização, análise e interpretação de **GRANDES** volumes de dados de uma empresa ou mercado de atuação.

Em geral, ele serve para direcionar as companhias em processos de tomada de decisão, resultando em ações mais estratégicas e assertivas.

Associado a capacidade científica e computacional de analisar os dados.



Como podemos tirar proveito do big data?



TOMADA DE DECISÕES!

Apropriar de métodos científicos (estatística) e computacionais que nos permitam extrair o máximo de informação de grandes massas de dados

Maior parte das decisões hoje em dia são tomadas com base na análise de dados

Data is the new gold!

Dados só têm valor se forem analisados.

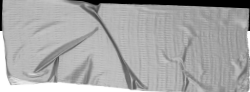
Aplicações de Data Analytics:

- Estratégias de Marketing
- Sistemas Educacionais
- Detecção de Fraudes
- Experiência de usuário
- Sistemas de recomendação
- People analytics

Ciência de Dados

É o estudo científico da criação, validação e transformação de dados para criação de significado e valor.

Cientista de Dados



O que é criar
significado e
valor?

É o profissional que usa métodos científicos para entender e **criar significado e valor** a partir de dados brutos.

Vamos apostar!

Regras do jogo:

- Temos duas máquinas M1 e M2.
- Cada máquina sorteia um número entre 1 e 6.
- Para ganhar você deve acertar qual o número a máquina irá sortear.

Qual número você escolhe para cada máquina?

M1



M2



M1



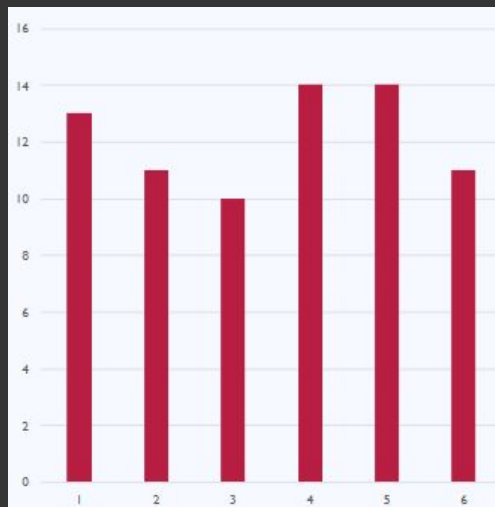
4 2 3 4 5 4 5 4 5 4 1 1 1 3
1 2 3 1 6 2 1 1 3 6 4 4 5 3
1 5 6 5 5 2 2 4 6 4 5 5 6 5
3 5 5 2 4 3 6 1 1 6 5 6 1 1
2 3 1 4 4 6 2 2 6 6 4 3 4 5
2 2 3

M2

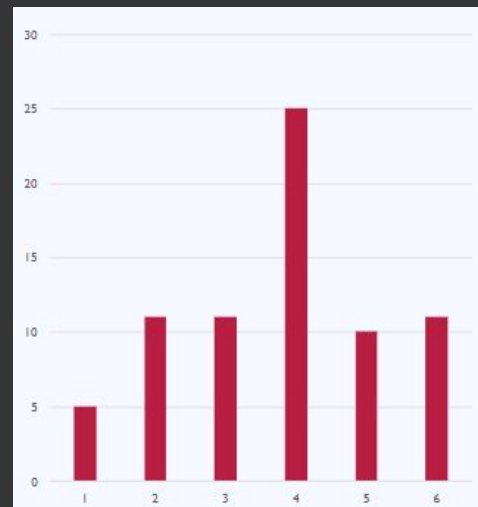


5 3 4 4 3 6 5 6 3 2 4 6 2 3 4
2 5 4 5 2 4 6 2 5 2 4 5 3 6 4
5 4 3 6 4 4 3 4 3 1 4 4 4 5 6
1 2 4 4 4 3 1 4 6 2 5 4 3 4 4
6 5 1 2 2 3 1 6 4 6 2 4 4 2 6
2 6 3 4 5 4 4 3 1 3 3 4 3 6 1
3 4 3 2 2 4 6 4 2 6

M1



M2

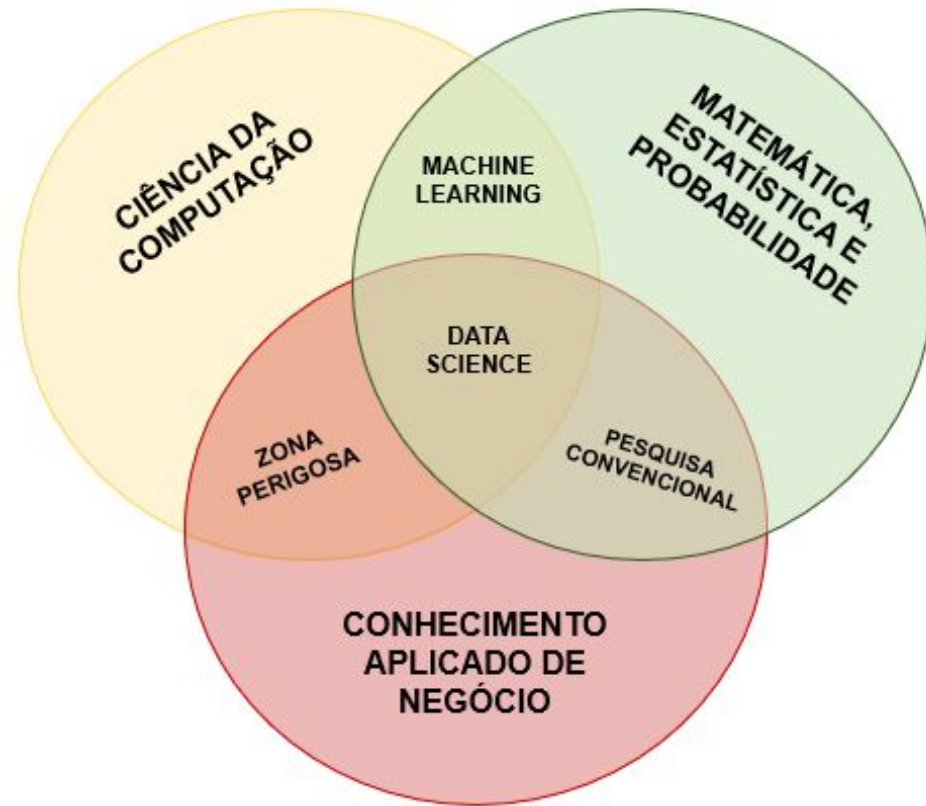


Ciência de dados é
mais que analisar
dados, é
criar/aprimorar
produtos baseados
em dados.



Estatística descritiva (ou simples): visa tornar os dados coletados mais fáceis de entender. (geração de tabelas, gráficos e medidas)

Estatística Inferencial: parte da estatística que nos permite tirar conclusões sobre uma população com base em uma amostra



Como se entrega resultados?

Insights acionáveis

Tipos de análises:

- **Descritiva:** o que aconteceu no passado?
- **Diagnóstica:** o que está acontecendo agora e por quê?
- **Preditiva:** O que vai acontecer e por quê?
- **Prescritiva:** O que devemos fazer sabendo o que pode acontecer?


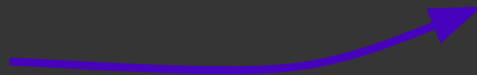
Conceitos Fundamentais da Estatística

1. POPULAÇÃO

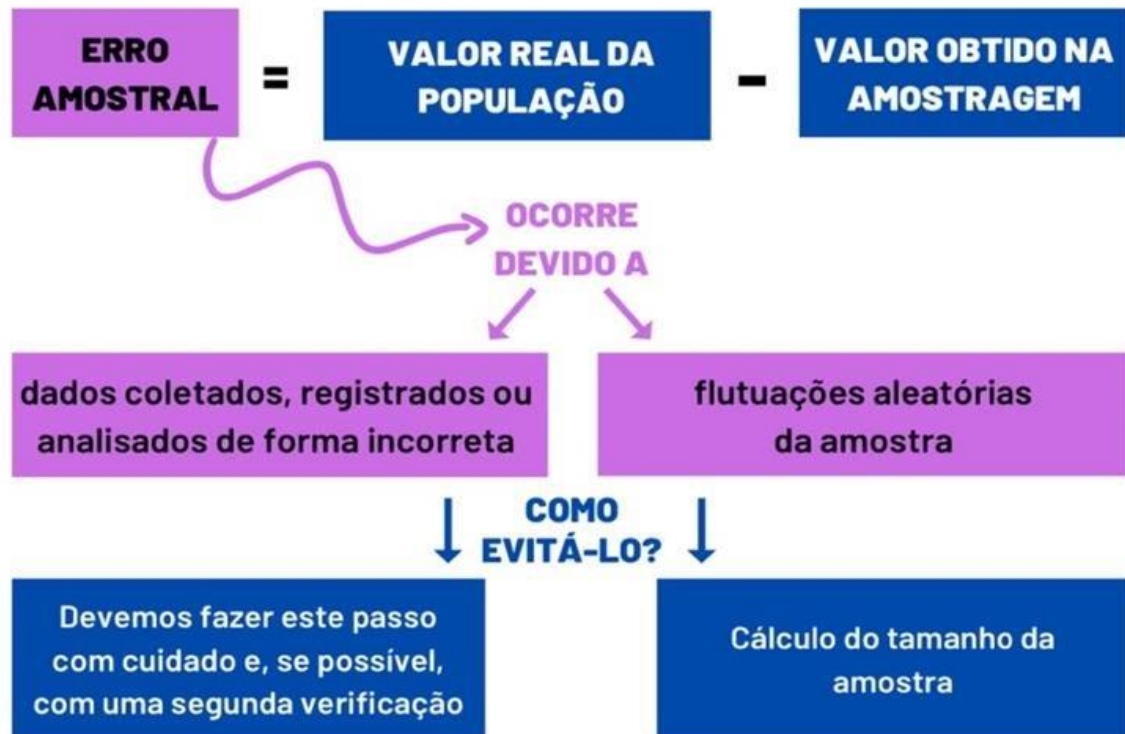
Conjunto de todos os elementos sob investigação

2. AMOSTRA

Subconjunto da população

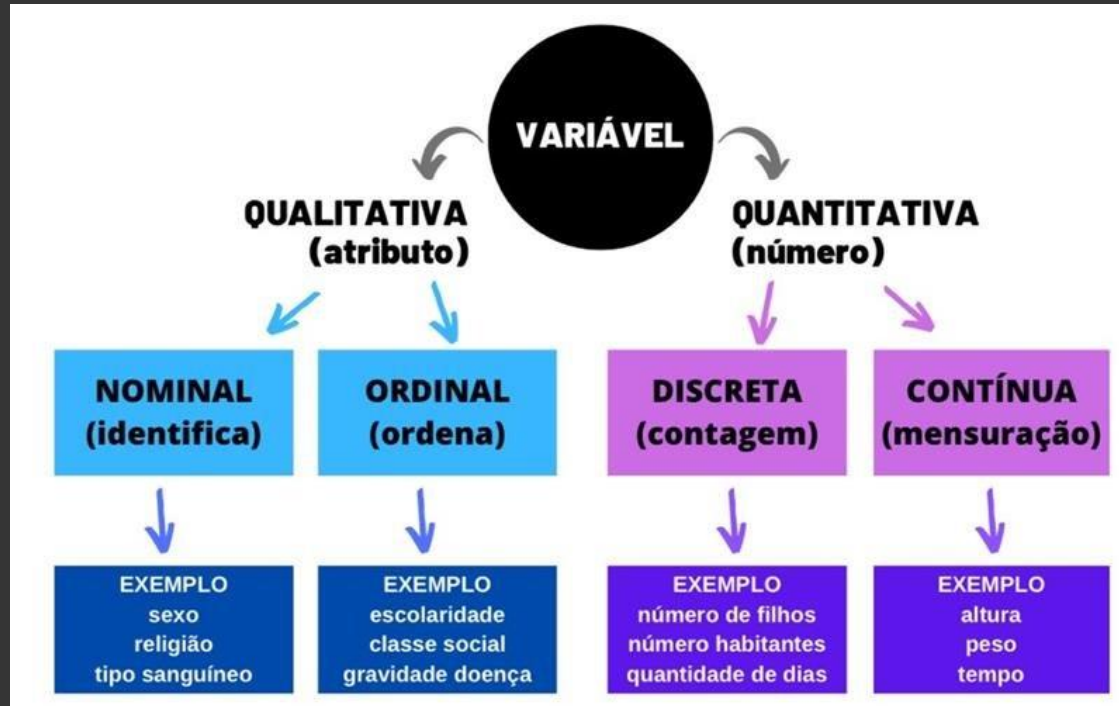


É importante que a amostra seja escolhida de tal maneira que seja REPRESENTATIVA da população



3. VARIÁVEL DE INTERESSE

Característica a ser observada nos indivíduos da amostra



Exemplos

- Opinião da população sobre uma proposta de governo
 - ♦ **População:** Todos os habitantes? Maiores de 16 anos?
 - ♦ **Amostra:** algum subconjunto dessa população. Como selecionar?
 - ♦ **Variável de interesse:** Concorda ou não, nota, ...
- Entender hábitos de compra dos clientes de uma loja virtual
 - ♦ **População:** Todos os clientes da loja
 - ♦ **Amostra:** Todos os clientes? Clientes dos últimos 90 dias?
 - ♦ **Variável de interesse:** Como podemos caracterizar os hábitos de compra? (horário da compra, valor da compra, local de entrega da loja...)

Não tem certo e errado > Temos que analisar caso a caso

PRÁTICA



Google Colab

<https://colab.research.google.com/>



Jupyter Notebook

+ DATASETS



Pandas



O que é?

- Biblioteca Python fundamental para manipulação e análise de dados estruturados.
- Oferece estruturas de dados poderosas e flexíveis:
 - Series: Arrays unidimensionais rotulados (como colunas).
 - DataFrame: Tabelas bidimensionais com colunas de diferentes tipos.

Por que é importante?

- Organização e estruturação eficiente de dados.
- Facilita a limpeza e o pré-processamento de dados: tratamento de dados ausentes, duplicatas, etc.
- Permite a seleção, filtragem e transformação de dados.
- Suporte para análise exploratória de dados (EDA) inicial.
- Integração com outras bibliotecas importantes (NumPy, Matplotlib).

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

Pandas



Instalação `pip install pandas`

Importação `import pandas as pd`

Pandas - Estruturas de Dados



Séries: Uma única coluna de dados, como uma lista, mas com rótulos (índices) para cada valor.

```
idades = pd.Series([25, 30, 22, 28, 35]) #Criando a série a partir da lista
print("Series de idades:\n", idades) #Visualização da série
print("Índices da Series:", idades.index) #Visualização dos índices
print("Valores da Series:", idades.values) #Visualização dos valores
```

Pandas - Estruturas de Dados



Séries: Uma única coluna de dados, como uma lista, mas com rótulos (índices) para cada valor.

Criando uma Series com índices personalizados

```
notas = pd.Series([8.5, 9.0, 7.5, 6.0], index=['Ana', 'Bruno', 'Carla', 'Daniel'])
```

```
print("Series de notas com índices personalizados:\n", notas)
```

```
print("Nota do Bruno:", notas['Bruno'])
```

Pandas - Estrutura de Dados



DataFrame: Tabela completa, com linhas e colunas nomeadas. Cada coluna é uma Series. É a estrutura mais utilizada para dados tabulares..

Criando um DataFrame a partir de um dicionário de listas

```
dados = {  
    'Nome': ['Alice', 'Bob', 'Charlie', 'David'],  
    'Idade': [25, 30, 22, 28]  
}  
  
df = pd.DataFrame(dados)  
print("DataFrame:\n", df)  
print("Coluna 'Nome':\n", df['Nome'])  
print("Coluna 'Idade' como Series:\n", df['Idade'])
```

Pandas - Leitura de arquivos



```
df = pd.read_csv('funcionario.csv')  
print(df)
```

```
df=pd.read_excel('funcionarios.xlsx', sheet_name='nome')  
print(df)
```

****** precisa de > pip install openpyxl

df.head() #imprime as primeiras 5 linhas

df.tail() #imprime as últimas 5 linhas

Pandas - Funcionalidades



Seleção e Filtragem de Dados: Permite acessar subconjuntos específicos dos seus dados.

Selecionando uma coluna

```
nomes = df['Nome']
```

```
print("Nomes:\n", nomes)
```

Selecionando múltiplas colunas

```
info = df[['Nome', 'Cidade']]
```

```
print("Nome e Cidade:\n", info)
```

Filtrando linhas com base em uma condição

```
mais_de_25 = df[df['Idade'] > 25]
```

```
print("Pessoas com mais de 25 anos:\n", mais_de_25)
```

Pandas - Funcionalidades



Manipulação de Dados: inclui adicionar novas colunas, modificar existentes, **remover colunas** ou linhas.

```
print("DataFrame com a coluna 'País':\n", df)
```

Modificando valores de uma coluna

```
df.loc[df['Nome'] == 'Alice', 'Idade'] = 26
```

```
print("DataFrame com a idade de Alice atualizada:\n", df)
```

Removendo uma coluna

```
df_sem_pais = df.drop('País', axis=1) # axis=1 indica coluna
```

```
print("DataFrame sem a coluna 'País':\n", df_sem_pais)
```


Pandas - Funcionalidades



```
vendas = {  
    'Produto': ['A', 'B', 'A', 'C', 'B', 'C', 'A', 'B'],  
    'Região': ['Norte', 'Sul', 'Norte', 'Leste', 'Sul', 'Leste', 'Sul', 'Norte'],  
    'Valor': [100, 150, 120, 80, 200, 90, 130, 160]  
}  
df_vendas = pd.DataFrame(vendas)
```

df.info() #informativo dos tipos de dados do dataFrame

df.describe() #retorna um resumo estatístico do DataFrame

df['Valor'].sum() #retorna a soma dos valores

df['Valor'].mean() #retorna a média dos valores

Pandas - Funcionalidades



Agrupamento de Dados: Permite agrupar linhas com base em valores de uma ou mais colunas

Agregar dados iguais

```
df.groupby('Região')
```

Métricas de dados agrupados

```
df.groupby('Região')['Valor'].mean()
```

```
df.groupby('Região')['Valor'].sum()
```

```
df.groupby('Região')['Valor'].max()
```

```
df.groupby('Região')['Valor'].min()
```

Pandas - Funcionalidades



Ordenando dados: Permite ordenar os dados das colunas em ordem crescente ou decrescente

Ordenando em ordem crescente

```
df.sort_values(by='Valor')
```

Ordenando em ordem decrescente

```
df.sort_values(by='Valor', ascending = False)
```



Obrigada!
Bons estudos

