**Conceitos em Pandas**

Arquivo contendo as anotações sobre pandas, seus conceitos básicos e outras coisas.

Anotações baseadas em pesquisas, retiradas também de trechos das aulas do curso da Stack Academy – Data Science, e também com ajuda do ChatGPT

# O que é o Pandas?

Pandas é uma **biblioteca do Python** feita para **análise e manipulação de dados**.  
Pensa nele como um “Excel turbinado dentro do Python”: com ele você consegue **ler, transformar, analisar e exportar dados** de maneira muito mais flexível e automatizada do que em planilhas manuais.

**Para que serve na prática?**  
No dia a dia, usamos Pandas para:

* **Ler dados** de diferentes fontes: CSV, Excel, SQL, JSON, entre outros.
* **Explorar dados**: ver as primeiras linhas, descrever estatísticas, entender o formato.
* **Transformar dados**: limpar colunas, tratar valores nulos, converter tipos.
* **Analisar dados**: agrupar, filtrar, ordenar, calcular médias/somas, gerar tabelas resumidas.
* **Exportar dados**: salvar de volta em CSV, Excel ou até mandar para um banco.

**Principais estruturas do Pandas**

* **Series**: uma coluna de dados (como uma coluna do Excel).
* **DataFrame**: uma tabela com linhas e colunas (como uma planilha inteira) que suporta diferentes tipos de dados (inteiro, float, string, etc)

# Tipos de arquivos suportados pelo pandas

O **Pandas** consegue ler e gravar dados em vários formatos. Vou te listar os **principais tipos de arquivos suportados** que realmente usamos no mercado:

**📄 Arquivos de texto/planilhas**

* **CSV** → pd.read\_csv() e df.to\_csv()
* **TSV** (tab separado) → também com read\_csv(sep="\t")
* **TXT** → idem, se for estruturado com separadores
* **Excel (.xls, .xlsx, .xlsm, .xlsb, .odf, .ods, .odt)**
  + pd.read\_excel() e df.to\_excel()
  + Suporta múltiplas abas de uma planilha.

**🗄️ Bancos de dados**

* **SQL (MySQL, PostgreSQL, SQLite, etc.)**
  + pd.read\_sql() e df.to\_sql() (precisa de conexão via SQLAlchemy ou driver).

**🌐 Estruturados em texto**

* **JSON** → pd.read\_json() e df.to\_json()
* **HTML (tabelas)** → pd.read\_html()
* **XML** → pd.read\_xml()

**📦 Outros formatos comuns em ciência de dados**

* **Parquet** → pd.read\_parquet() e df.to\_parquet()
* **Feather** → pd.read\_feather() e df.to\_feather()
* **ORC** → pd.read\_orc() e df.to\_orc()
* **Stata (.dta)** → pd.read\_stata() e df.to\_stata()
* **SPSS (.sav)** → pd.read\_spss()
* **HDF5 (.h5)** → pd.read\_hdf() e df.to\_hdf()
* **Pickle (.pkl)** → pd.read\_pickle() e df.to\_pickle()

👉 Ou seja: **se é um formato tabular ou semi-estruturado, o Pandas provavelmente consegue lidar**.

💡 **No dia a dia**:

* CSV é o mais comum (troca de dados entre sistemas).
* Excel é onipresente em empresas (relatórios, planilhas financeiras, etc).
* SQL é essencial (quase sempre os dados “moram” em um banco).
* Parquet está se tornando padrão em ambientes de Big Data (porque é compacto e rápido).

# Exemplo básico de um DataFrame:

