# Pandas

Professora: Lorena Barberia

DCP-USP

*lorenabarberia@usp.br*

Universidade de São Paulo

# Tópicos da Aula



Universidade de São Paulo

# O que vimos até agora?

- Até agora, aprendemos em Python:
  - Uso básico do Python.
  - Variáveis e tipos de dados.
  - Estruturas de dados (listas, tuplas, dicionários).
  - Estruturas condicionais (if, else, elif).
  - Loops (for, while).
  - Definição e uso de funções.
  - Numpy para computação numérica.

Universidade de São Paulo

# Recapitulação

Nas três primeiras semanas do curso, já exploramos muita coisa de Python. Fomos do nível completamente iniciante e estamos adentrando no intermediário. Na aula de hoje, descobriremos o Pandas e sua utilidade para a análise de dados.

Universidade de São Paulo

# O que é o Pandas?

- Pandas é uma biblioteca de análise de dados em Python. Seu nome vem de *Panel Data*, dados em painel.
- Desenvolvida por Wes McKinney durante os anos 2000.
- Projetada para lidar com dados de maneira eficiente. E tinha em mente replicações de métodos econométricos da época.
- É uma das bibliotecas mais populares para análise de dados em Python.

Universidade de São Paulo

# História do Pandas

- Criado por Wes McKinney enquanto trabalhava em finanças quantitativas.
- Inspirado nas funcionalidades de manipulação de dados do R e do Excel.
- Lançado como código aberto em 2008.
- Desde então, tornou-se uma ferramenta essencial para cientistas de dados.

Universidade de São Paulo

# Principais Conceitos

- **DataFrame**: Tabela bidimensional, semelhante a uma planilha.
- **Series**: Estrutura unidimensional, semelhante a uma coluna de uma planilha.
- **Index**: Etiquetas para identificar linhas e colunas.

Universidade de São Paulo

# Funcionalidades Principais

- Leitura e escrita de dados em diversos formatos.
- Manipulação e limpeza de dados ausentes.
- Fatiamento, indexação e seleção de dados.
- Agregação e transformação de dados.
- Visualização de dados.

Universidade de São Paulo

# Relação do Pandas e Numpy

- Pandas é construído sobre o NumPy.
- Pandas fornece estruturas de dados de alto nível para análise de dados tabulares.
- NumPy fornece arrays multidimensionais e funções para operações numéricas eficientes.

Universidade de São Paulo

# Diferenças Chave

- NumPy é mais adequado para operações numéricas e matriciais.
- Pandas é otimizado para análise de dados tabulares e manipulação de séries temporais.
- NumPy usa arrays homogêneos, enquanto Pandas permite tipos de dados heterogêneos.
- Pandas oferece recursos de indexação mais flexíveis.

Universidade de São Paulo

# O que são Series em Pandas?

- Uma das estruturas de dados fundamentais em Pandas.
- Similar a uma matriz unidimensional ou coluna em uma planilha.
- Pode armazenar dados de diferentes tipos.
- Possui um índice associado que identifica cada elemento.

Universidade de São Paulo

# Criando Series em Pandas

- Series podem ser criadas a partir de listas, arrays, dicionários, etc.
- Exemplo:

```
import pandas as pd
data = [1, 2, 3, 4, 5]
minha_serie = pd.Series(data)
```

Universidade de São Paulo

# Indexação em Series

Os elementos de uma Series podem ser acessados por índice:

```
# Criando a serie

data = [1, 2, 3, 4, 5]
minha_serie = pd.Series(data)

# Indexando
minha_serie[0]

#Output:

1
```

Universidade de São Paulo

# Seleção em Series

Pode-se usar operadores lógicos para seleção condicional. Por exemplo, se eu quiser os valores maiores que 3 da serie anterior:

```
# Criando a serie

data = [1, 2, 3, 4, 5]
minha_serie = pd.Series(data)

# Indexando
minha_serie[minha_serie > 3]

#Output:

4, 5
```

Universidade de São Paulo

# Operações Matemáticas

Series suportam operações matemáticas elementares:

```
# Criando a serie

data = [1, 2, 3, 4, 5]
minha_serie = pd.Series(data)

# Indexando
minha_serie + 10

#Output:

[11,12,13,14,15]
```

Universidade de São Paulo

# Funções Estatísticas

Pandas fornece funções para cálculos estatísticos em Series:

```
# Criando a serie

data = [1, 2, 3, 4, 5]
minha_serie = pd.Series(data)

# Média
minha_serie.mean()

#Output:

3
```

Universidade de São Paulo

# Conclusão

- Series em Pandas são estruturas de dados poderosas.
- Permitem armazenar, manipular e analisar dados de forma eficiente.
- São a base para operações mais avançadas em Pandas.

Universidade de São Paulo

# O que são DataFrames?

- Estruturas de dados bidimensionais.
- Tabelas ou planilhas de dados.
- Compostas por linhas e colunas.

Universidade de São Paulo

# Criando Dataframes

```python
import pandas as pd

data = {
    'Nome': ['Alice', 'Bob', 'Carol'],
    'Idade': [25, 30, 35],
    'Cidade': ['São Paulo', 'Rio de Janeiro', 'Belo Horizonte']
}

df = pd.DataFrame(data)
```

Universidade de São Paulo

# O que vamos aprender de dataframes

Hoje, iremos aprender diversas coisas que podemos fazer com dataframes em Pandas:

- Indexação, Seleção e Filtragem;
- Operação em Colunas (Ou Series);
- Aplicando Funções em series e dataframes (e *Mapping*);
- Como calcular estatísticas descritivas e estatísticas resumo;
- Correlação e Covariância;
- Remover valores duplicados e contagem de ocorrências (Frequência) de valores.

Universidade de São Paulo

# Importação de Dados

Finalmente trabalharemos com dados reais em Python. Há várias maneiras diferentes de importar um arquivo de dados em Python (seja excel, .csv, Stata, etc.), e há maior variação quando se trata do Colab.

Universidade de São Paulo

# Importação de Dados

Por exemplo, para ler .csv, o pandas oferece o método read.csv().

```
caminho_arquivo = 'sample_data/mnist_train_small.csv'

df = pd.read_csv(caminho_arquivo)
```

Universidade de São Paulo

# Importação de Dados do Excel

Para ler planilhas de excel, o pandas oferece o método read.excel().

```
caminho_arquivo = 'sample_data/mnist_train_small.xlsx'

df = pd.read_excel(caminho_arquivo)
```

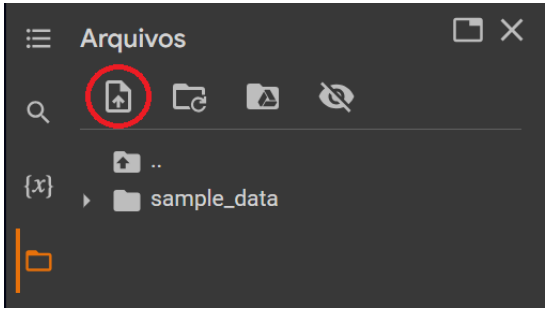
Universidade de São Paulo

# Importando dados no Colab

No Colab, por ser em nuvem, você precisa fazer algumas coisas antes de importar o dado, já que ele não possui acesso direto aos arquivos do seu computador. Uma primeira forma é inserir manualmente os arquivos na sua seção do Colab.

Universidade de São Paulo

# Inserção manual no Colab



# Montando o Drive

Uma outra forma é montar o seu drive no colab e utilizar arquivos do drive diretamente. Por exemplo:

```
from google.colab import drive

drive.mount('/content/drive')

caminho = 'content/drive/MyDrive/projeto/dados.csv'

df = pd.read_csv(caminho)
```

Universidade de São Paulo

# Limpeza e Manipulação de Dados

Por fim, também trabalharemos como limpar e manipular dados em Python, e as principais ferramentas que são disponibilizadas pelo Pandas.

Universidade de São Paulo

# Importância da Limpeza de Dados

- Dados brutos frequentemente contêm ruído, erros e valores ausentes.
- Limpeza de dados é essencial para análises precisas.
- Essa pode ser uma das partes mais trabalhosas e demoradas da análise de dados.

# Técnicas de Limpeza e Manipulação

- Remoção de dados duplicados.
- Tratamento de valores ausentes.
- Filtragem de outliers.
- Conversão de tipos de dados.
- Renomeação de colunas.

Universidade de São Paulo

# Conclusão

- Limpeza e manipulação de dados são etapas cruciais na preparação de dados para análise.
- Pandas fornece ferramentas poderosas para realizar essas tarefas de forma eficiente.
- Dados limpos e bem preparados são a base para análises precisas e inferências significativas.

Universidade de São Paulo

# Considerações Finais

A aula de hoje buscou dar uma visão geral do Pandas, o que são series e dataframes, como importar dados no Python, e Limpeza e Manipulação dos dados. Na próxima aula, iremos para a visualização de dados.

Universidade de São Paulo

# Laboratório

Agora, vocês irão se reunir em grupos e trabalharão no laboratório da aula de hoje (Lab_Aula4.ipynb). Os laboratórios estão na nossa pasta do drive. Pedimos que façam uma cópia do arquivo em uma subpasta de acordo com o seu turno. Nomeiem essa subpasta de acordo com o grupo (sobrenome1_sobrenome2_lab1) e coloquem as informações dos alunos em uma célula de *Markdown* no início do laboratório (Nome, NUSP, curso e se é da graduação ou pós).

Universidade de São Paulo

# Dúvidas?

Universidade de São Paulo