Python: análise de dados com Pandas I

Poliana N Ferreira

KORU escola

Projetos de Dados

Projetos de Dados

O que envolve?

- Um projeto de dados pode incluir coleta, armazenamento, processamento, análise e visualização de dados.

- Envolve ferramentas para manipulação de dados, algoritmos e modelos de machine learning e infraestrutura de dados.

Kaggle

Plataforma de Projetos e Dados

- Kaggle é uma plataforma de competição de ciência de dados e uma comunidade online de cientistas de dados e profissionais de aprendizado de máquina da Google LLC.

- Possui centenas de bases de dados gratuitas para testar e utilizar em projetos.

Introdução ao Pandas

O que é Pandas?

Pandas é uma biblioteca de Python para trabalhar com dados

- Amplamente utilizada para manipulação e análise de dados.
- Indispensável para cientistas de dados.
- **Eficiente e flexível.**

Como instalar o pandas

Usamos o Pandas em qualquer script Python ou Jupyter Notebook

- Pode ser instalada localmente com `pip install pandas`
- Ou podemos também usar recursos em nuvem como o Google Colab: Pandas já vem instalado lá.

Importamos pandas com o comando

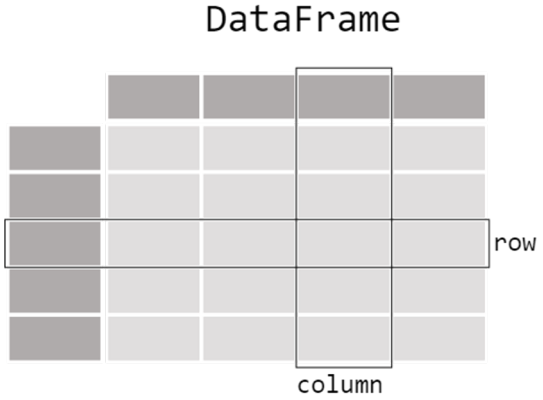
```
import pandas
```

Estrutura dos dados em Pandas

Data Frame

Pandas é uma biblioteca de Python para trabalhar com dados.

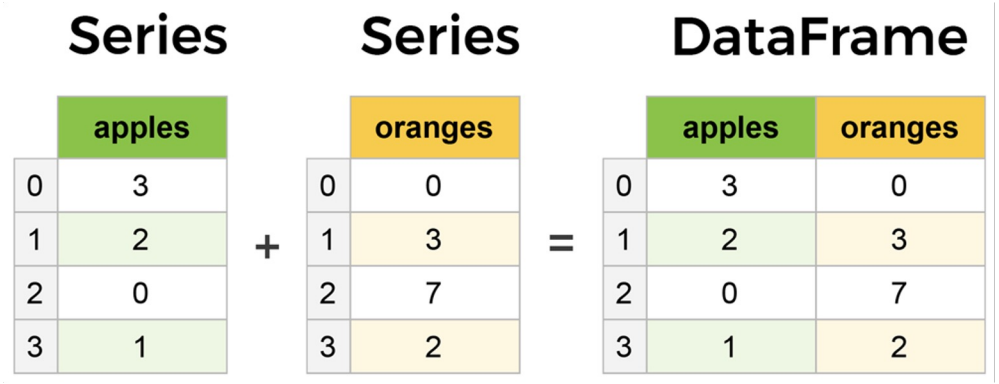
- A principal estrutura de dados do Pandas é o DataFrame.
- Estrutura bidimensional, de tamanho mutável e que pode ser modificada.
- Ele é muito semelhante a uma planilha ou a uma tabela SQL e é ideal para manipular dados tabulares com colunas de diferentes tipos.



Data Frame

Pandas é uma biblioteca de Python para trabalhar com dados

- Uma Série é uma estrutura de dados unidimensional do Pandas.
- Semelhante a um vetor ou a uma coluna em uma tabela.
- Cada Série tem um índice que dá a cada elemento um rótulo.



Importação dos Dados

Criando um Data Frame

Quais as possibilidades?

- Podemos criar a base de dados no Python.

```
import pandas as pd

data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}

df = pd.DataFrame(data, index = ["day1", "day2", "day3"])

print(df)
```

Criando um Data Frame

Quais as possibilidades?

- Importar de um CSV.

```
import pandas as pd

df = pd.read_csv('data.csv')

print(df)
```

Criando um Data Frame

Quais as possibilidades?

- Importar de um Excel.

```
import pandas as pd
# Read Excel file
df = pd.read_excel('c:/apps/courses_schedule.xlsx')
print(df)
```

Limpeza e Preparação

Comandos básicos

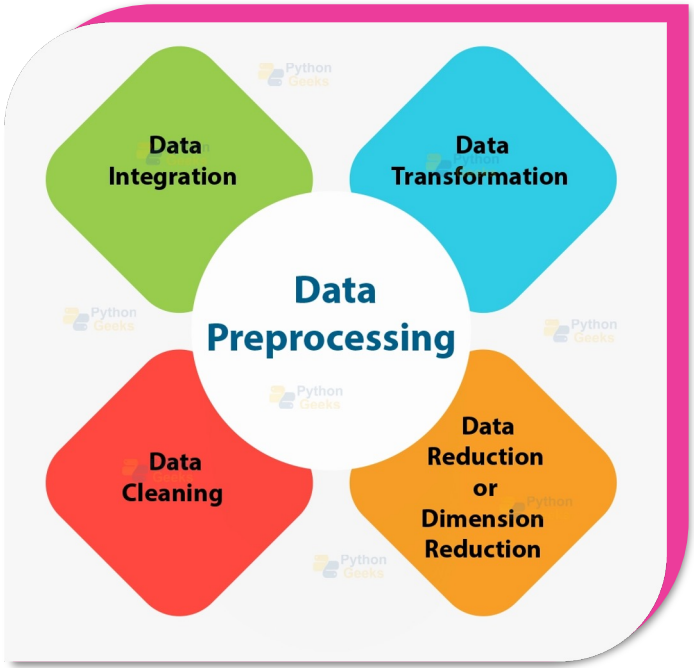
Observando a base de dados de forma inicial

- Info: dá informações gerais sobre dados nulos, colunas e tipos de dados.
- Describe: dá informações estatísticas sobre cada coluna.

```
df.info()

df.describe()
```

Pré-Processamento dos Dados



Limpeza dos Dados

- Limpeza de dados significa consertar dados incorretos em seu conjunto de dados.

- Dados incorretos podem ser:

    - Células vazias;
    - Dados em formato errado;
    - Dados errados;
    - Duplicatas.

Limpeza dos Dados

Exemplos

```
df.dropna(inplace = True)
```

```
df.fillna(130, inplace = True)
```

```
x = df["Calories"].mean()

df["Calories"].fillna(x, inplace = True)
```

```
df['Date'] = pd.to_datetime(df['Date'])
```

```
df.loc[7, 'Duration'] = 45
```

```
for x in df.index:
  if df.loc[x, "Duration"] > 120:
    df.loc[x, "Duration"] = 120
```

```
df.drop_duplicates(inplace = True)
```

Exemplo - Limpeza e Preparação dos Dados

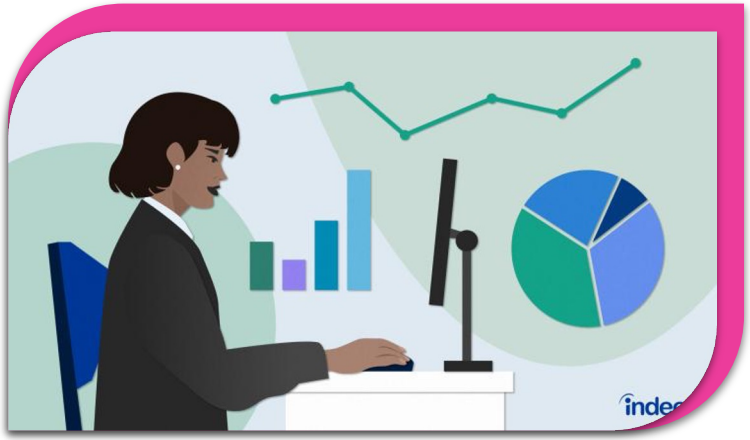
python™

Manipulação e Análise

Análise dos Dados

Análises: Estatística Descritiva e Gráficos

- O Pandas possui funções específicas para isso também.
- Baseado em Numpy e Matplotlib.



Exemplo – Subindo nosso projeto no Github

python™

Agora vamos avaliar a aula?

Obrigada!

escola