Python para Ciência de Dados e Inteligência Artificial

Aula 05: Introdução ao Pandas.

IMT – Instituto Mauá de Tecnologia

Março/2023

Prof. Jones Egydio
jones.egydio@maua.br

Introdução ao Pandas

- O **Pandas** é a principal biblioteca para manipulação, limpeza, transformação e análise de dados no Python. Foi criada por *Wes McKinney* em 2008.
  - O pandas fornece estruturas de dados de alta performance, fáceis de usar e flexíveis, que permitem trabalhar com dados em formato de planilha (dataframe), séries temporais e outras formas de dados tabulares;
  - É uma ferramenta poderosa para lidar com dados em Python, com funcionalidades que permitem carregar e salvar dados em diferentes formatos, manipular dados ausentes, fazer junção de dados, transformações de dados, agregações, filtragem, ordenação e demais tarefas;
  - Algumas das principais funcionalidades do pandas incluem a leitura e escrita de diversos formatos de arquivos de dados, como CSV, Excel, SQL, JSON e HTM;
  - O pandas é amplamente utilizado em diversas áreas, como finanças, ciência de dados, pesquisa acadêmica e engenharia. É uma ferramenta essencial para aqueles que trabalham com dados em Python.

```
import pandas as pd
```

Introdução ao Pandas

- **Series e DataFrame:** Existem dois objetos básicos no Pandas, o Series e o DataFrame **.**

  DataFrame: É a estrutura de dados mais utilizada do Pandas, pode ser entendida como uma Tabela, onde cada entrada pode ser entendida como uma linha com um ou vários atributos (colunas):

```
df = pd.DataFrame({"Aprovado": [65, 45], "Reprovado": [35, 55]})
print(df)

df2 = pd.DataFrame({"Azul": ["Gosto", "Indiferente"], "Vermelho": ["Odeio", "Amo"]})
print(df2)
```

Introdução ao Pandas

- **Series e DataFrame:** Existem dois objetos básicos no Pandas, o Series e o DataFrame .

  Series: É a estrutura de dados mais simples do Pandas, pode ser entendida como uma lista, mas na verdade é uma coluna de um DataFrame:

```
a = pd.Series([1, 2, 3, 4, 5])
print(a)
```

Introdução ao Pandas

- **Lendo arquivos de dados:** Por mais que seja interessante criar DataFrames e Series, normalmente iremos importar bases de dados disponíveis em arquivos externos:

```
df = pd.read_csv("/content/drive/My Drive/Colab Notebooks/Python
 para CD e IA/netflix_titles.csv")

print(df.shape)

print(df.columns)

df.head()
```

Introdução ao Pandas

- Seleção:

```
df["type"]

df.type

df.type[0] #ou df["type"][0]

df.iloc[0]

df.iloc[:,0]

df.loc[1:3]

df.loc[0,'release_year']
```

Introdução ao Pandas

- Seleção:

```
df.loc[df.release_year > 2019]

df.loc[(df.release_year > 2019) & (df.type == "Movie")]

df.loc[(df.release_year > 2019) & (df.type == "Movie"),["title"]]

df.loc[(df.release_year > 2019) & ((df.type == "Movie") | (df.typ
e == "TV Show")),["title"]]

df.loc[df.release_year.isin([2019, 2020])]

df.loc[df.director.notna()]
```

Introdução ao Pandas

- Indexação:

```
df.loc[df.release_year > 2019]

df.loc[(df.release_year > 2019) & (df.type == "Movie")]

df.loc[(df.release_year > 2019) & (df.type == "Movie"),["title"]]

df.loc[(df.release_year > 2019) & ((df.type == "Movie") | (df.typ
e == "TV Show")),["title"]]

df.loc[df.release_year.isin([2019, 2020])]

df.loc[df.director.notna()]
```

Introdução ao Pandas

- Indexação:

```
df.set_index("show_id")
```

Introdução ao Pandas

- Exercício – enviar via Open LMS até 25/3/2023:

1. Selecione a coluna *country* do DataFrame e atribua a um novo DataFrame
2. Selecione apenas o primeiro valor da coluna *country* do DataFrame e atribua a uma variável
3. Selecione a primeira linha do DataFrame e atribua a um novo DataFrame
4. Selecione os 10 primeiros valores da coluna *country* do DataFrame e atribua a um novo DataFrame
5. Selecione as linhas 1, 2, 3, 10 e 50 da coluna *release_year* do DataFrame e atribua a um novo DataFrame
6. Crie um novo DataFrame que contenha as colunas *type, title, country* e *release_year* apenas das linhas 1, 2 e 10
7. Crie um novo DataFrame que contenha as colunas *title* e *country* das 100 primeiras linhas
8. Crie um novo DataFrame que contenha apenas as produções Brasileiras
9. Crie um novo DataFrame que contenha apenas os filmes (*Movie*) Brasileiros dos anos de 2019 e 2020

Referências bibliográficas

MENEZES, N. N. C., Introdução à programação com Python, 3ª Edição. São Paulo: Editora Novatec, 2019.

Notas de aula: Prof. Anderson Harayashini Moreira, pós-graduação em CD e IA, IMT, 2022.