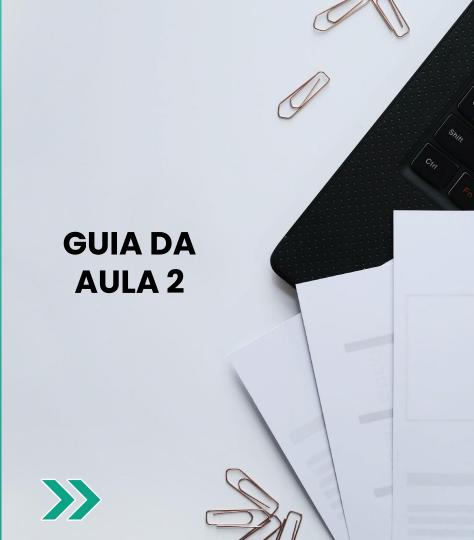


Profissão: Analista de dados





DATA WRANGLING I







Selecione e filtre dados



DataFrame



Acompanhe aqui os temas que serão tratados na videoaula







1. Série

Uma **série** é uma **coluna** de um **dataframe**. Para selecionar uma coluna utilizamos:

```
notação (similiar a indexação de listas Python):
serie = dataframe['<nome-da-coluna>']

Nota: Repare no uso das chaves simples [ ] .
```

• Exemplo: Coluna linguagem de programação do datafra github_df:

```
In []: linguagem_serie = github_df['language']
In []: linguagem_serie
In []: type(linguagem_serie)
```





Seleção

• **Exemplo**: Indexação simples com método loc (similar a lista Python):

```
In []: top_1_linguagem = linguagem_serie.loc[0]
In []: top_1_linguagem
In []: type(top_1_linguagem)
```

• **Exemplo**: Fatiamento ou *slicing* com método loc (similar a lista Python):

```
In []: top_5_linguagem = linguagem_serie.loc[0:5]
In []: top_5_linguagem
In []: type(top_5_linguagem)
```







• **Exemplo**: Filtro funcional:

```
In [ ]:
    linguagem_serie [lambda linguagem: linguagem == 'python']
```

• Exemplo: Filtro funcional com novos índices:

```
In []:
    linguagem_serie[
        lambda linguagem: linguagem == 'python'
].reset_index (drop=True)
```





2. DataFrame

Um conjunto de colunas ou séries é um novo dataframe.

Para selecionar um conjunto de colunas utilizamos a seguinte notação:

```
novo_dataframe = dataframe[['<nome-da-coluna-a>', '<nome-da-coluna- b>', ...]]
```

Nota: Repare no uso das chaves duplas [[]].

• Exemplo: Colunas ranking e linguagem de programação do dataframe github df:

```
In []: ranking_linguagem_df = github_df[['ranking', 'language']]
In []: ranking_linguagem_df
In []: type(ranking_linguagem_df)
```





Seleção

• **Exemplo**: Indexação simples (linha) com método loc (similar a lista Python):

```
In []: top_1_linguagem = ranking_linguagem_df.loc[0]
In []: top_1_linguagem
In []: type(top_1_linguagem)
```

• Exemplo: Indexação simples (linha e coluna) com método loc (similar a lista Python):

```
In []: top_1_linguagem = github_df.loc[0, ['ranking', 'language']]
In []: top_1_linguagem
In []: type(top_1_linguagem)
```





• **Exemplo**: Fatiamento ou *slicing* (linhas) com método loc (similar a lista Python):

```
In []: top_5_ranking_linguagem = ranking_linguagem_df.loc[0:5]
In []: top_5_ranking_linguagem
In []: type(top_5_ranking_linguagem)
```

• **Exemplo**: Fatiamento ou *slicing* (linhas e colunas) com método _{loc} (similar a lista Python):

```
In []: top_5_ranking_linguagem = github_df.loc[0:5, ['ranking', 'language']]
In []: top_5_ranking_linguagem
In []: type(top_5_ranking_linguagem)
```





Filtros

• **Exemplo**: Filtro com o método query :

```
In []:    ranking_linguagem_df .query('language == "python"')

In []:    ranking_linguagem_df .query('language == "python" & ranking > 5')

In []:    ranking_linguagem_df .query('language == "python" | language == "go"')
```

