

```
panda@ufscar:~$ intro
```

Introdução ao pandas em Python

Por Joãozinho



O que é pandas?

- Pacote do Python, criado para manipular dados tabulares
- Muitas similaridades com SQL
 - `SELECT` vira indexação com `[]`
 - `WHERE` vira indexação booleana
 - `GROUP BY`, `JOIN`, `UNION`...
- Análise exploratória, limpeza dos dados, interface com outras bibliotecas de data science
- Não confundir com o nome do grupo (foi mal)



Mexer com dados não é uma tarefa fácil

- Dados faltantes
- Formatos e encodings inutilizáveis
 - Exemplo: dados advindos de scraping
- Dados duplicados
- Outliers
- Falta de documentação
- ...E outras tarefas de pré-processamento

Ninguém: a

O mundo se todos os dados fossem limpos automaticamente:



You better work, data engineer!



[Work B**ch - Britney Spears](#)

Antes, um pouco sobre arrays em Python

```
import numpy as np
```

```
minha_matriz = np.arange(0, 20).reshape(4,5)
```

```
print(minha_matriz)
```

```
[[ 0  1  2  3  4]
```

```
 [ 5  6  7  8  9]
```

```
[10 11 12 13 14]
```

```
[15 16 17 18 19]
```

```
[20 21 22 23 24]]
```

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

`minha_matriz[linha, coluna]`

`print(minha_matriz[1,2])`

`7`

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

```
minha_matriz[linha, coluna]
```

```
print(minha_matriz[-1,-1])
```

24



0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

`minha_matriz[linha, coluna]`

`print(minha_matriz[1,:])`

`[5 6 7 8 9]`

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

`minha_matriz[linha, coluna]`

`print(minha_matriz[:4,2:4])`

`[[2 3]`

`[7 8]`

`[12 13]`

`[17 18]]`

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19
20	21	22	23	24

Outra maneira!

```
minha_matriz[linha, coluna]
```

```
print(minha_matriz[:-1,2:4])
```

```
[[ 2  3]
 [ 7  8]
 [12 13]
 [17 18]]
```



Exemplo: Livros do John Green

- Vamos criar um dataframe simples contendo os livros do John Green
- Título, data de publicação, se eu já li, minha avaliação e a média de avaliação no Goodreads



Series [pandas.Series](#)

- Estrutura mais básica do pandas
- Array nomeado ou não, de 1 dimensão e que possui um índice
- Pode ser uma observação ou uma característica
- Usada para formar dataframes ou é derivada de dataframes



```
livros_jg = ['Quem é Você, Alasca?', 'O Teorema Katherine', 'Deixe a Neve Cair',  
            'Cidades de Papel', 'Will e Will', 'A Culpa É das Estrelas',  
            'Tartarugas Até Lá Embaixo']
```

```
livros_jg_series = pd.Series(livros_jg)
```

```
livros_jg_series
```

```
livros_jg_series
```

```
0      Quem é Você, Alasca?  
1      O Teorema Katherine  
2      Deixe a Neve Cair  
3      Cidades de Papel  
4      Will e Will  
5      A Culpa É das Estrelas  
dtype: object
```

```
datas = pd.to_datetime(['03/03/2005', '21/09/2006', '02/10/2008',  
                        '16/10/2008', '06/04/2010', '10/01/2012'], dayfirst=True)  
datas_series = pd.Series(datas)
```

datas_series

```
0    2005-03-03  
1    2006-09-21  
2    2008-10-02  
3    2008-10-16  
4    2010-04-06  
5    2012-01-10  
dtype: datetime64[ns]
```

```
lidos = pd.Series([True, False, False, True, True, False])  
lidos
```

```
0    True  
1   False  
2   False  
3    True  
4    True  
5   False  
dtype: bool
```

```
minha_nota = pd.Series([5, np.nan, np.nan, 4.5, 4, np.nan])  
minha_nota
```

```
0    5.0  
1    NaN  
2    NaN  
3    4.5  
4    4.0  
5    NaN  
dtype: float64
```

```
nota_gr = pd.Series([4.02, 3.57, 3.76, 3.80, 3.77, 4.21])  
nota_gr
```

```
0    4.02  
1    3.57  
2    3.76  
3    3.80  
4    3.77  
5    4.21  
dtype: float64
```



```
df = pd.DataFrame({'livros': livros_jg_series, 'ano_publicacao': datas,
                    'lido': lidos, 'minha_nota': minha_nota,
                    'nota_goodreads': nota_gr})
df
```

	livros	ano_publicacao	lido	minha_nota	nota_goodreads
0	Quem é Você, Alasca?	2005-03-03	True	5.0	4.02
1	O Teorema Katherine	2006-09-21	False	NaN	3.57
2	Deixe a Neve Cair	2008-10-02	False	NaN	3.76
3	Cidades de Papel	2008-10-16	True	4.5	3.80
4	Will e Will	2010-04-06	True	4.0	3.77
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95



Esqueci de algo... 🤔

```
tartarugas = pd.Series({'livro': 'Tartarugas Até Lá Embaixo', 'ano_publicacao': pd.to_datetime('10/10/2017'),  
                        'lido': True, 'minha_nota': 4.5, 'nota_goodreads': 3.95})
```

tartarugas

livro	Tartarugas Até Lá Embaixo
ano_publicacao	2017-10-10 00:00:00
lido	True
minha_nota	4.5
nota_goodreads	3.95
dtype:	object

Agora sim! 😞

```
df = df.append(tartarugas, ignore_index=True)
df
```

	livro	ano_publicacao	lido	minha_nota	nota_goodreads
0	Quem é Você, Alasca?	2005-03-03	True	5.0	4.02
1	O Teorema Katherine	2006-09-21	False	NaN	3.57
2	Deixe a Neve Cair	2008-10-02	False	NaN	3.76
3	Cidades de Papel	2008-10-16	True	4.5	3.80
4	Will e Will	2010-04-06	True	4.0	3.77
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95

Análise exploratória

- Descobrir padrões nos dados
- Encontrar problemas
- Testar hipóteses
- Confirmar intuições
- Se exibir pros amigos



	livro	data_publicacao	lido	minha_notas	nota_goodreads
0	Quem é Você, Alasca?	2005-03-03	True	5.0	4.02
1	O Teorema Katherine	2006-09-21	False	NaN	3.57
2	Deixe a Neve Cair	2008-10-02	False	NaN	3.76
3	Cidades de Papel	2008-10-16	True	4.5	3.80
4	Will e Will	2010-04-06	True	4.0	3.77
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95

```
df['livro']
```

```
0      Quem é Você, Alasca?
1      O Teorema Katherine
2      Deixe a Neve Cair
3      Cidades de Papel
4      Will e Will
5      A Culpa É das Estrelas
6      Tartarugas Até Lá Embaixo
Name: livro, dtype: object
```

```
df[['data_publicacao', 'livro']]
```

	data_publicacao	livro
0	2005-03-03	Quem é Você, Alasca?
1	2006-09-21	O Teorema Katherine
2	2008-10-02	Deixe a Neve Cair
3	2008-10-16	Cidades de Papel
4	2010-04-06	Will e Will
5	2012-01-10	A Culpa É das Estrelas
6	2017-10-10	Tartarugas Até Lá Embaixo

	livro	data_publicacao	lido	minha_notas	nota_goodreads
0	Quem é Você, Alasca?	2005-03-03	True	5.0	4.02
1	O Teorema Katherine	2006-09-21	False	NaN	3.57
2	Deixe a Neve Cair	2008-10-02	False	NaN	3.76
3	Cidades de Papel	2008-10-16	True	4.5	3.80
4	Will e Will	2010-04-06	True	4.0	3.77
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95

```
df['livro']
```

```
0    Quem é Você, Alasca?
1    O Teorema Katherine
2    Deixe a Neve Cair
3    Cidades de Papel
4    Will e Will
5    A Culpa É das Estrelas
6    Tartarugas Até Lá Embaixo
Name: livro, dtype: object
```

```
df.loc[:, 'livro']
```

```
0    Quem é Você, Alasca?
1    O Teorema Katherine
2    Deixe a Neve Cair
3    Cidades de Papel
4    Will e Will
5    A Culpa É das Estrelas
6    Tartarugas Até Lá Embaixo
Name: livro, dtype: object
```

```
df.iloc[:, 0]
```

```
0    Quem é Você, Alasca?
1    O Teorema Katherine
2    Deixe a Neve Cair
3    Cidades de Papel
4    Will e Will
5    A Culpa É das Estrelas
6    Tartarugas Até Lá Embaixo
Name: livro, dtype: object
```

```
df.loc[5:6]
```

	livro	data_publicacao	lido	minha_notas	nota_goodreads
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95

	livro	ano_publicacao	lido	minha_notas	nota_goodreads
0	Quem é Você, Alasca?	2005-03-03	True	5.0	4.02
1	O Teorema Katherine	2006-09-21	False	NaN	3.57
2	Deixe a Neve Cair	2008-10-02	False	NaN	3.76
3	Cidades de Papel	2008-10-16	True	4.5	3.80
4	Will e Will	2010-04-06	True	4.0	3.77
5	A Culpa É das Estrelas	2012-01-10	False	NaN	4.21
6	Tartarugas Até Lá Embaixo	2017-10-10	True	4.5	3.95

```
df.describe()
```

	minha_notas	nota_goodreads
count	4.000000	7.000000
mean	4.500000	3.868571
std	0.408248	0.208761
min	4.000000	3.570000
25%	4.375000	3.765000
50%	4.500000	3.800000
75%	4.625000	3.985000
max	5.000000	4.210000

```
df['nota_goodreads'].describe()
```

```
count    6.000000
mean     3.855000
std      0.225278
min      3.570000
25%     3.762500
50%     3.785000
75%     3.965000
max      4.210000
Name: nota_goodreads, dtype: float64
```



```
df.groupby(by=df['lido'])['nota_goodreads'].describe()
```

	count	mean	std	min	25%	50%	75%	max
lido								
False	3.0	3.846667	0.328684	3.57	3.665	3.76	3.985	4.21
True	3.0	3.863333	0.136504	3.77	3.785	3.80	3.910	4.02

Brigado 🙄 Agora vamos ao Jupyter!!!!

