




**Curso:**

Desenvolvimento Full Stack

**Campus:**

POLO JARDIM BRASÍLIA - ÁGUAS LINDAS DE GOIÁS - GO

**Disciplina:**

Missão Prática | Tratando a imensidão dos dados 

Missão Prática | Nível 3 | Mundo 5

**Turma:**

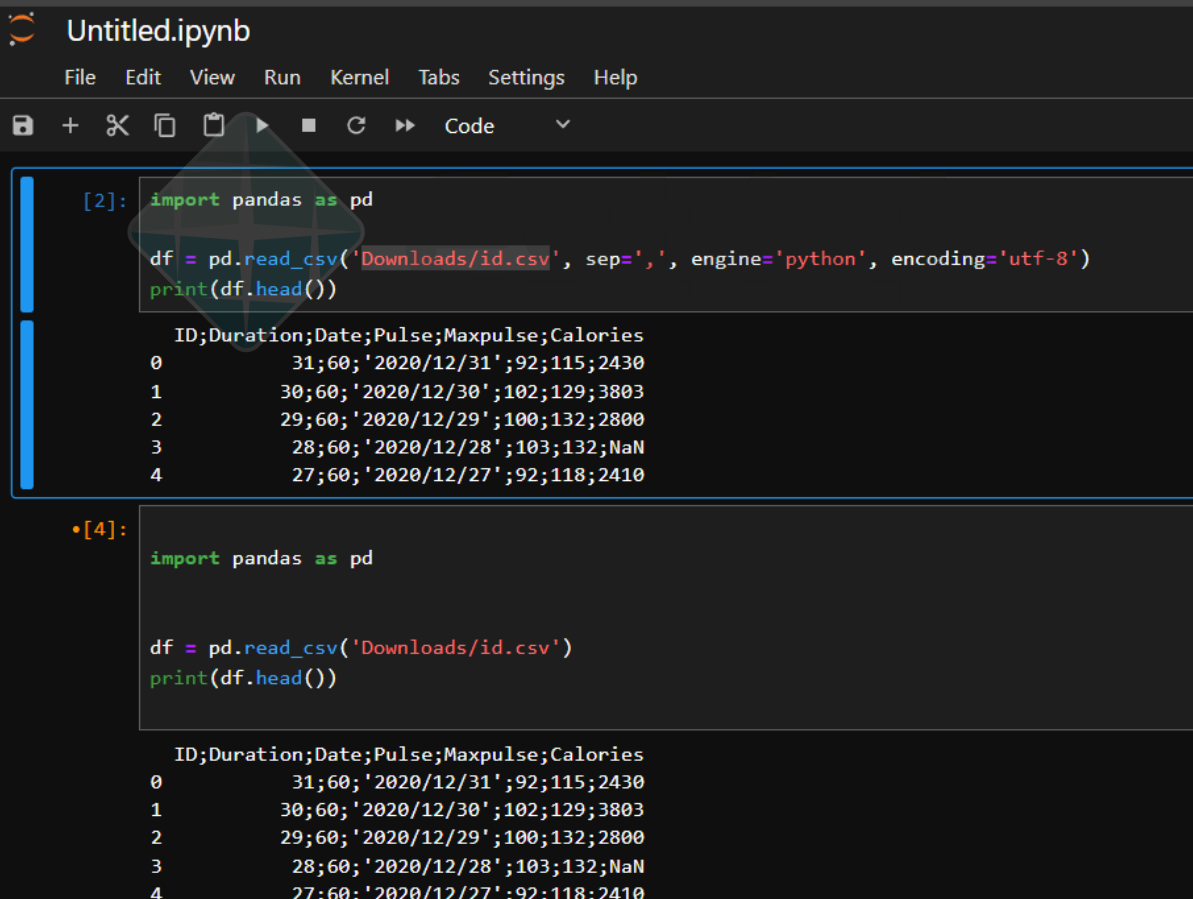
23.4

**Aluno:**

BRUNO SANTIAGO DE OLIVEIRA

# Contextualização

Como Analista de Dados, você recebeu, em um novo projeto, um conjunto de dados. Sua principal tarefa é tratar os dados desse conjunto a fim de que possam ser utilizados para a descoberta de conhecimento através de sua posterior análise e interpretação. Para tal tarefa, você deverá utilizar a linguagem Python e a biblioteca Pandas.



```
Untitled.ipynb
File Edit View Run Kernel Tabs Settings Help

[2]: import pandas as pd

df = pd.read_csv('Downloads/id.csv', sep=',', engine='python', encoding='utf-8')
print(df.head())

ID;Duration;Date;Pulse;Maxpulse;Calories
0      31;60;'2020/12/31';92;115;2430
1      30;60;'2020/12/30';102;129;3803
2      29;60;'2020/12/29';100;132;2800
3      28;60;'2020/12/28';103;132;NaN
4      27;60;'2020/12/27';92;118;2410

•[4]: import pandas as pd

df = pd.read_csv('Downloads/id.csv')
print(df.head())

ID;Duration;Date;Pulse;Maxpulse;Calories
0      31;60;'2020/12/31';92;115;2430
1      30;60;'2020/12/30';102;129;3803
2      29;60;'2020/12/29';100;132;2800
3      28;60;'2020/12/28';103;132;NaN
4      27;60;'2020/12/27';92;118;2410
```

```

•[5]: import pandas as pd

df = pd.read_csv('Downloads/id.csv', sep=';')

print(df.head())

dados = df

print(dados.describe())

```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	'2020/12/31'	92	115	2430
1	30	60	'2020/12/30'	102	129	3803
2	29	60	'2020/12/29'	100	132	2800
3	28	60	'2020/12/28'	103	132	NaN
4	27	60	'2020/12/27'	92	118	2410

	ID	Duration	Pulse	Maxpulse
count	32.000000	32.000000	32.000000	32.000000
mean	14.875000	68.437500	103.500000	128.500000
std	9.664534	70.039591	7.832933	12.998759
min	0.000000	30.000000	90.000000	101.000000
25%	6.750000	60.000000	100.000000	120.000000
50%	14.500000	60.000000	102.500000	127.500000
75%	23.250000	60.000000	106.500000	132.250000
max	31.000000	450.000000	130.000000	175.000000

•[6]:

```
print(dados.info())
```

```
print(dados.head())
```

```
print(dados.tail())
```

```
3    Pulse    32 non-null    int64
4    Maxpulse 32 non-null    int64
5    Calories 30 non-null    object
dtypes: int64(4), object(2)
memory usage: 1.6+ KB
None
```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	'2020/12/31'	92	115	2430
1	30	60	'2020/12/30'	102	129	3803
2	29	60	'2020/12/29'	100	132	2800
3	28	60	'2020/12/28'	103	132	NaN
4	27	60	'2020/12/27'	92	118	2410

	ID	Duration	Date	Pulse	Maxpulse	Calories
27	3	45	'2020/12/04'	109	175	2824
28	2	60	'2020/12/03'	103	135	3400
29	1	60	'2020/12/02'	117	145	4790
30	1	60	'2020/12/21'	108	131	3642
31	0	60	'2020/12/01'	110	130	4091

•[7]:

```
# Cria uma cópia do DataFrame original
dados_copia = dados.copy()

# Verifica se a cópia foi criada com sucesso
print(dados_copia.head())
print(dados.columns)
```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	'2020/12/31'	92	115	2430
1	30	60	'2020/12/30'	102	129	3803
2	29	60	'2020/12/29'	100	132	2800
3	28	60	'2020/12/28'	103	132	NaN
4	27	60	'2020/12/27'	92	118	2410

Index(['ID', 'Duration', 'Date', 'Pulse', 'Maxpulse', 'Calories'], dtype='object')

•[12]:

```
print(dados.columns)

# Substitui valores nulos na coluna 'Calories' por 0
dados['Calories'].fillna(0, inplace=True)
```

Index(['ID', 'Duration', 'Date', 'Pulse', 'Maxpulse', 'Calories'], dtype='object')

```

•[14]: import pandas as pd
        # Substitui valores nulos na coluna 'Date' por '1900/01/01'
        dados['Date'].fillna('1900/01/01', inplace=True)

        # Transforma os dados da coluna 'Date' em datetime
        dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], errors='coerce')

        print(dados_copia)

```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410
5	26	60	NaT	100	120	2500
6	25	60	2020-12-25	102	126	3345
7	24	45	2020-12-24	105	132	2460
8	23	60	2020-12-23	130	101	3000
9	22	45	NaT	100	119	2820
10	20	45	2020-12-20	97	125	2430 2
11	19	60	2020-12-19	103	123	3230
12	18	45	2020-12-18	90	112	NaN
13	17	60	2020-12-17	100	120	3000
14	16	60	2020-12-16	98	120	2152
15	15	60	2020-12-15	98	123	2750
16	14	60	2020-12-14	104	132	3793
17	13	60	2020-12-13	106	128	3453
18	12	60	2020-12-12	100	120	2507
19	11	60	2020-12-12	100	120	2507
20	10	60	2020-12-11	103	147	3293
21	9	60	2020-12-10	98	124	2690
22	8	30	2020-12-09	109	133	1951
23	7	450	2020-12-08	104	134	2533
24	6	60	2020-12-07	110	136	3740
25	5	60	2020-12-06	102	127	3000
26	4	45	2020-12-05	117	148	4060
27	3	45	2020-12-04	109	175	2824

```

•[15]: import pandas as pd
# Substitui '1900/01/01' por NaN na coluna 'Date'
dados_copia['Date'] = dados_copia['Date'].replace('1900/01/01', pd.NaT)

# Transforma a coluna 'Date' para datetime
dados_copia['Date'] = pd.to_datetime(dados_copia['Date'])

print(dados_copia.head())

```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410

```

•[16]: import pandas as pd
# Substitui '20201226' por '2020/12/26' na coluna 'Date'
dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')

# Transforma a coluna 'Date' para datetime
dados_copia['Date'] = pd.to_datetime(dados_copia['Date'])

print(dados_copia.head())

```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410

```
•[17]: import pandas as pd
# Transforma a coluna 'Date' para datetime
dados_copia['Date'] = pd.to_datetime(dados_copia['Date'])

print(dados_copia.head())
```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410

```
•[18]: # Remove registros com valores nulos na coluna 'Date'
dados_copia.dropna(subset=['Date'], inplace=True)

print(dados_copia.head())
```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410



```
[20]: # Imprime o DataFrame para verificar as transformações
print(dados_copia)
```

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	31	60	2020-12-31	92	115	2430
1	30	60	2020-12-30	102	129	3803
2	29	60	2020-12-29	100	132	2800
3	28	60	2020-12-28	103	132	NaN
4	27	60	2020-12-27	92	118	2410
6	25	60	2020-12-25	102	126	3345
7	24	45	2020-12-24	105	132	2460
8	23	60	2020-12-23	130	101	3000
10	20	45	2020-12-20	97	125	2430 2
11	19	60	2020-12-19	103	123	3230
12	18	45	2020-12-18	90	112	NaN
13	17	60	2020-12-17	100	120	3000
14	16	60	2020-12-16	98	120	2152
15	15	60	2020-12-15	98	123	2750
16	14	60	2020-12-14	104	132	3793
17	13	60	2020-12-13	106	128	3453
18	12	60	2020-12-12	100	120	2507
19	11	60	2020-12-12	100	120	2507
20	10	60	2020-12-11	103	147	3293
21	9	60	2020-12-10	98	124	2690
22	8	30	2020-12-09	109	133	1951
23	7	450	2020-12-08	104	134	2533
24	6	60	2020-12-07	110	136	3740
25	5	60	2020-12-06	102	127	3000
26	4	45	2020-12-05	117	148	4060
27	3	45	2020-12-04	109	175	2824
28	2	60	2020-12-03	103	135	3400
29	1	60	2020-12-02	117	145	4790
30	1	60	2020-12-21	108	131	3642
31	0	60	2020-12-01	110	130	4091

## CONCLUSÃO

Conclui que, após seguir o roteiro proposto para o tratamento dos dados utilizando Python e a biblioteca Pandas, adquiri uma compreensão mais profunda sobre a manipulação e análise de dados. No processo, consegui importar e verificar o conjunto de dados, lidar com valores nulos e realizar transformações necessárias para garantir a integridade e usabilidade dos dados.

Ao ler o CSV e atribuir os dados a uma variável, pude conferir as informações gerais e as primeiras e últimas linhas do arquivo, o que me permitiu identificar e corrigir problemas como valores nulos e formatos de data inconsistentes. A substituição de valores nulos por 0 e a transformação das datas foram etapas cruciais, especialmente ao enfrentar e resolver erros relacionados a formatos de data.

O desafio de transformar o valor "20201226" em um formato de data apropriado e a remoção de registros com valores nulos foram oportunidades valiosas para aplicar métodos específicos e garantir a qualidade dos dados. No final, o tratamento eficaz dos dados me preparou para utilizá-los de forma mais eficiente na análise e descoberta de conhecimento, confirmando meu domínio sobre as operações básicas da biblioteca Pandas e a importância dessas habilidades no trabalho com conjuntos de dados reais.