

## Faculdade Estácio - Polo Curitiba - Centro

**Curso:** Desenvolvimento Full Stack

**Disciplina:** Tratando a imensidão dos dados

**Número da Turma:** RPG0033

**Semestre Letivo:** 5

**Integrante:** Mariana Lucas Fernandes Onório

**Repositório:** <https://github.com/MariLFO/estacio-mundo5-missao-nivel-3>

### Sumário:

<b>Faculdade Estácio - Polo Curitiba - Centro</b> .....	<b>1</b>
Sumário:.....	1
1. Título da Prática:.....	2
2. Objetivos da Prática:.....	2
3. Códigos do roteiro:.....	2
4. Resultados da execução dos códigos:.....	3
Microatividade 1: Descrever como ler um arquivo CSV usando a biblioteca Pandas (Python).....	3
Microatividade 2: Descrever como criar um subconjunto de dados a partir de um conjunto existente usando a biblioteca Pandas (Python).....	4
Microatividade 3: Descrever como configurar o número máximo de linhas a serem exibidas na visualização de um conjunto de dados usando a biblioteca Pandas (Python).....	5
Microatividade 4: Descrever como exibir as primeiras e últimas “N” linhas de um conjunto de dados usando a biblioteca Pandas (Python).....	6
Microatividade 5: Descrever como exibir informações gerais sobre as colunas, linhas e dados de um conjunto de dados usando a biblioteca Pandas (Python).....	7
Missão Prática   Vamos interligar as coisas com a nuvem!.....	8

# 1. Título da Prática:

**RPG0033**

Tratando a imensidão dos dados

## 2. Objetivos da Prática:

- Descrever como ler um arquivo CSV usando a biblioteca Pandas (Python);
- Descrever como criar um subconjunto de dados a partir de um conjunto existente usando a biblioteca Pandas (Python);
- Descrever como configurar o número máximo de linhas a serem exibidas na visualização de um conjunto de dados usando a biblioteca Pandas (Python);
- Descrever como exibir as primeiras e últimas “N” linhas de um conjunto de dados usando a biblioteca Pandas (Python); Descrever como exibir informações gerais sobre as colunas, linhas e dados de um conjunto de dados usando a biblioteca Pandas (Python);

## 3. Códigos do roteiro:

<https://github.com/MariLFO/estacio-mundo5-missao-nivel-3>

## 4. Resultados da execução dos códigos:

Microatividade 1: Descrever como ler um arquivo CSV usando a biblioteca Pandas (Python)

```
microatividade1.py M X
microatividade1.py > ...
1 import pandas as pd
2
3 # Variável para armazenar os dados
4 dados = pd.read_csv('dados.csv', sep=';', engine='python', encoding='utf-8')
5
6 # Salva as alterações no arquivo CSV
7 dados.to_csv('dados_gerados_microatividade1.csv', index=False)
8
9 # Exibe os dados da variável
10 print(dados)
11
```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** PORTS AZURE Python + - - - ^ X

/Users/D/Development/Estacio/estacio-mundo5-missao-nivel-3/env/bin/python /Users/D/Development/Estacio/estacio-mundo5-missao-nivel-3/microatividade1.py

The default interactive shell is now zsh.  
To update your account to use zsh, please run `chsh -s /bin/zsh`.  
For more details, please visit <https://support.apple.com/kb/HT208050>.

(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D\$ /Users/D/Development/Estacio/estacio-mundo5-missao-nivel-3/microatividade1.py

ID	Duration	Date	Pulse	Maxpulse	Calories
0	0	'2020/12/01'	110	130	4091.0
1	1	'2020/12/02'	117	145	4790.0
2	2	'2020/12/03'	103	135	3400.0
3	3	'2020/12/04'	109	175	2824.0
4	4	'2020/12/05'	117	148	4060.0
5	5	'2020/12/06'	102	127	3000.0
6	6	'2020/12/07'	110	136	3740.0
7	7	'2020/12/08'	104	134	2533.0
8	8	'2020/12/09'	109	133	1951.0
9	9	'2020/12/10'	98	124	2690.0
10	10	'2020/12/11'	103	147	3293.0
11	11	'2020/12/12'	100	120	2507.0
12	12	'2020/12/12'	100	120	2507.0
13	13	'2020/12/13'	106	128	3453.0
14	14	'2020/12/14'	104	132	3793.0
15	15	'2020/12/15'	98	123	2750.0
16	16	'2020/12/16'	98	120	2152.0
17	17	'2020/12/17'	100	120	3000.0
18	18	'2020/12/18'	90	112	NaN
19	19	'2020/12/19'	103	123	3230.0
20	20	'2020/12/20'	97	125	2430.0
21	21	'2020/12/21'	108	131	3642.0
22	22	NaN	100	119	2820.0
23	23	'2020/12/23'	130	101	3000.0
24	24	'2020/12/24'	105	132	2460.0
25	25	'2020/12/25'	102	126	3345.0
26	26	20201226	100	120	2500.0
27	27	'2020/12/27'	92	118	2410.0
28	28	'2020/12/28'	103	132	NaN
29	29	'2020/12/29'	100	132	2800.0
30	30	'2020/12/30'	102	129	3803.0
31	31	'2020/12/31'	92	115	2430.0

(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D\$

Microatividade 2: Descrever como criar um subconjunto de dados a partir de um conjunto existente usando a biblioteca Pandas (Python)

```
microatividade1.py  microatividade2.py U X
microatividade2.py > ...
1  import pandas as pd
2
3  # Variável para armazenar os dados
4  dados = pd.read_csv('dados.csv', sep=';', engine='python', encoding='utf-8')
5
6  # Nova variável contendo subconjunto de dados com apenas três colunas
7  subconjunto_dados = dados[['ID', 'Date', 'Calories']]
8
9  # Salva as alterações em um novo arquivo CSV
10 subconjunto_dados.to_csv('dados_gerados_microatividade2.csv', index=False)
11
12 # Exibe os dados da nova variável
13 print(subconjunto_dados)
14
```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** PORTS AZURE

Python + - [ ] [X] ... ^ X

```
(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D$ /Users/D/Development/Estacio/estacio-mundo5-missao-nivel-3/env/bin/python /Users/D/Development/Estacio/e
stacio-mundo5-missao-nivel-3/microatividade2.py
ID      Date      Calories
0 0    '2020/12/01'    4091.0
1 1    '2020/12/02'    4790.0
2 2    '2020/12/03'    3400.0
3 3    '2020/12/04'    2824.0
4 4    '2020/12/05'    4060.0
5 5    '2020/12/06'    3000.0
6 6    '2020/12/07'    3740.0
7 7    '2020/12/08'    2533.0
8 8    '2020/12/09'    1951.0
9 9    '2020/12/10'    2690.0
10 10   '2020/12/11'    3293.0
11 11   '2020/12/12'    2507.0
12 12   '2020/12/12'    2507.0
13 13   '2020/12/13'    3453.0
14 14   '2020/12/14'    3703.0
15 15   '2020/12/15'    2750.0
16 16   '2020/12/16'    2152.0
17 17   '2020/12/17'    3000.0
18 18   '2020/12/18'      NaN
19 19   '2020/12/19'    3230.0
20 20   '2020/12/20'    2430.0
21 21   '2020/12/21'    3642.0
22 22      NaN    2820.0
23 23   '2020/12/23'    3000.0
24 24   '2020/12/24'    2460.0
25 25   '2020/12/25'    3345.0
26 26    20201226    2500.0
27 27   '2020/12/27'    2410.0
28 28   '2020/12/28'      NaN
29 29   '2020/12/29'    2800.0
30 30   '2020/12/30'    3803.0
31 31   '2020/12/31'    2430.0
(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D$
```

### Microatividade 3: Descrever como configurar o número máximo de linhas a serem exibidas na visualização de um conjunto de dados usando a biblioteca Pandas (Python)

```
microatividade1.py  microatividade2.py  microatividade3.py U X
microatividade3.py > ...
1  import pandas as pd
2
3  # Configura a propriedade max_rows para 9999
4  pd.set_option('display.max_rows', 9999)
5
6  # Variável para armazenar os dados
7  dados = pd.read_csv('dados.csv', sep=';', engine='python', encoding='utf-8')
8
9  # Subconjunto de dados com apenas três colunas
10 subconjunto_dados = dados[['ID', 'Date', 'Calories']]
11
12 # Salva as alterações no arquivo CSV
13 dados.to_csv('dados_gerados_microatividade3.csv', index=False)
14
15 # Imprime/exibe os dados da variável usando o método to_string()
16 print(dados.to_string())
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS AZURE Python + - ☐ ☒ ... ^ X

● (env) MacBooks-MBP:estacio-mundo5-missao-nivel-3 D\$ /Users/D/Development/Estacio/estacio-mundo5-missao-nivel-3/env/bin/python /Users/D/Development/Estacio/e

```
stacio-mundo5-missao-nivel-3/microatividade3.py
ID Duration Date Pulse Maxpulse Calories
0 0 60 '2020/12/01' 110 130 4091.0
1 1 60 '2020/12/02' 117 145 4790.0
2 2 60 '2020/12/03' 103 135 3400.0
3 3 45 '2020/12/04' 109 175 2824.0
4 4 45 '2020/12/05' 117 148 4060.0
5 5 60 '2020/12/06' 102 127 3000.0
6 6 60 '2020/12/07' 110 136 3740.0
7 7 450 '2020/12/08' 104 134 2533.0
8 8 30 '2020/12/09' 109 133 1951.0
9 9 60 '2020/12/10' 98 124 2690.0
10 10 60 '2020/12/11' 103 147 3293.0
11 11 60 '2020/12/12' 100 120 2507.0
12 12 60 '2020/12/12' 100 120 2507.0
13 13 60 '2020/12/13' 106 128 3453.0
14 14 60 '2020/12/14' 104 132 3793.0
15 15 60 '2020/12/15' 98 123 2750.0
16 16 60 '2020/12/16' 98 120 2152.0
17 17 60 '2020/12/17' 100 120 3000.0
18 18 45 '2020/12/18' 90 112 NaN
19 19 60 '2020/12/19' 103 123 3230.0
20 20 45 '2020/12/20' 97 125 2430.0
21 21 60 '2020/12/21' 108 131 3642.0
22 22 45 NaN 100 119 2820.0
23 23 60 '2020/12/23' 130 101 3000.0
24 24 45 '2020/12/24' 105 132 2460.0
25 25 60 '2020/12/25' 102 126 3345.0
26 26 60 20201226 100 120 2500.0
27 27 60 '2020/12/27' 92 118 2410.0
28 28 60 '2020/12/28' 103 132 NaN
29 29 60 '2020/12/29' 100 132 2800.0
30 30 60 '2020/12/30' 102 129 3803.0
31 31 60 '2020/12/31' 92 115 2430.0
```

○ (env) MacBooks-MBP:estacio-mundo5-missao-nivel-3 D\$ ☐

Microatividade 4: Descrever como exibir as primeiras e últimas “N” linhas de um conjunto de dados usando a biblioteca Pandas (Python)

```
microatividade1.py microatividade2.py microatividade3.py M microatividade4.py U X
microatividade4.py > ...
7 dados = pd.read_csv('dados.csv', sep=';', engine='python', encoding='utf-8')
8
9 # Subconjunto de dados com apenas três colunas
10 subconjunto_dados = dados[['ID', 'Date', 'Calories']]
11
12 # Salva as alterações no arquivo CSV
13 dados.to_csv('dados_gerados_microatividade4.csv', index=False)
14
15 # Exibe os dados da variável usando o método to_string()
16 print("\nDados original utilizando to_string():")
17 print(dados.to_string())
18
19 # Exibe as primeiras 10 linhas do conjunto de dados original
20 print("\nPrimeiras 10 linhas do conjunto de dados original:")
21 print(dados.head(10))
22
23 # Exibe as últimas 10 linhas do conjunto de dados original
24 print("\nÚltimas 10 linhas do conjunto de dados original:")
25 print(dados.tail(10))
26
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS AZURE

23	23	60	'2020/12/23'	130	101	3000.0
24	24	45	'2020/12/24'	105	132	2460.0
25	25	60	'2020/12/25'	102	126	3345.0
26	26	60	20201226	100	120	2500.0
27	27	60	'2020/12/27'	92	118	2410.0
28	28	60	'2020/12/28'	103	132	NaN
29	29	60	'2020/12/29'	100	132	2800.0
30	30	60	'2020/12/30'	102	129	3803.0
31	31	60	'2020/12/31'	92	115	2430.0

Primeiras 10 linhas do conjunto de dados original:

ID	Duration	Date	Pulse	Maxpulse	Calories	
0	0	60	'2020/12/01'	110	130	4091.0
1	1	60	'2020/12/02'	117	145	4790.0
2	2	60	'2020/12/03'	103	135	3400.0
3	3	45	'2020/12/04'	109	175	2824.0
4	4	45	'2020/12/05'	117	148	4060.0
5	5	60	'2020/12/06'	102	127	3000.0
6	6	60	'2020/12/07'	110	136	3740.0
7	7	450	'2020/12/08'	104	134	2533.0
8	8	30	'2020/12/09'	109	133	1951.0
9	9	60	'2020/12/10'	98	124	2690.0

Últimas 10 linhas do conjunto de dados original:

ID	Duration	Date	Pulse	Maxpulse	Calories	
22	22	45	NaN	100	119	2820.0
23	23	60	'2020/12/23'	130	101	3000.0
24	24	45	'2020/12/24'	105	132	2460.0
25	25	60	'2020/12/25'	102	126	3345.0
26	26	60	20201226	100	120	2500.0
27	27	60	'2020/12/27'	92	118	2410.0
28	28	60	'2020/12/28'	103	132	NaN
29	29	60	'2020/12/29'	100	132	2800.0
30	30	60	'2020/12/30'	102	129	3803.0
31	31	60	'2020/12/31'	92	115	2430.0

o (env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D\$

## Microatividade 5: Descrever como exibir informações gerais sobre as colunas, linhas e dados de um conjunto de dados usando a biblioteca Pandas (Python)

```
microatividade1.py  microatividade2.py  microatividade3.py  microatividade4.py  microatividade5.py U x
microatividade5.py > ...
25 print(dados.tail(10))
26
27 # Imprime as informações gerais sobre o conjunto de dados
28 print("\nInformações gerais sobre o conjunto de dados:")
29 dados_info = dados.info()
30
31 # Total de linhas
32 total_linhas = dados.shape[0]
33 print(f"\nTotal de linhas: {total_linhas}")
34
35 # Total de colunas
36 total_colunas = dados.shape[1]
37 print(f"Total de colunas: {total_colunas}")
38
39 # Quantidade de dados nulos por coluna
40 dados_nulos = dados.isnull().sum()
41 print(f"\nQuantidade de dados nulos por coluna:\n{dados_nulos}")
42
43 # Quantidade de memória utilizada pelo conjunto de dados
44 memoria_utilizada = dados.memory_usage(deep=True).sum()

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS AZURE
26 26 60 20201226 100 120 2500.0
27 27 60 '2020/12/27' 92 118 2410.0
28 28 60 '2020/12/28' 103 132 NaN
29 29 60 '2020/12/29' 100 132 2800.0
30 30 60 '2020/12/30' 102 129 3803.0
31 31 60 '2020/12/31' 92 115 2430.0

Informações gerais sobre o conjunto de dados:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 6 columns):
# Column Non-Null Count Dtype
---
0 ID 32 non-null int64
1 Duration 32 non-null int64
2 Date 31 non-null object
3 Pulse 32 non-null int64
4 Maxpulse 32 non-null int64
5 Calories 30 non-null float64
dtypes: float64(1), int64(4), object(1)
memory usage: 1.6+ KB

Total de linhas: 32
Total de colunas: 6

Quantidade de dados nulos por coluna:
ID 0
Duration 0
Date 1
Pulse 0
Maxpulse 0
Calories 2
dtype: int64

Quantidade de memória utilizada: 3575 bytes
(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D$
```

## Missão Prática | Vamos interligar as coisas com a nuvem!

oatividade1.py

microatividade2.py

microatividade3.py

microatividade4.py

microatividade5.py

missao\_pratica.py U X

missao\_pratica.py > ...

```
41 dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], format='%Y/%m/%d', errors='coerce')
42 print("\nConjunto de dados após transformar 'Date' para datetime:")
43 print(dados_copia)
44
45 # Transformar especificamente o valor "20201226" para o formato datetime
46 dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')
47
48 # Repetindo novamente a transformação dos dados da coluna 'Date' para datetime
49 dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], format='%Y/%m/%d', errors='coerce')
50 print("\nConjunto de dados após corrigir o valor '20201226' e transformar 'Date' para datetime:")
51 print(dados_copia)
52
53 # Remoção de todos os registros contendo valores nulos
54 dados_copia.dropna(subset=['Date'], inplace=True)
55 print("\nConjunto de dados após remover registros com valores nulos em 'Date':")
56 print(dados_copia)
```

PROBLEMS

OUTPUT

DEBUG CONSOLE

TERMINAL

PORTS

AZURE

Python + ▾

🗑️ ⌵ ✕

Conjunto de dados após corrigir o valor '20201226' e transformar 'Date' para datetime:

	ID	Duration	Date	Pulse	Maxpulse	Calories
0	0	60	NaT	110	130	4091.0
1	1	60	NaT	117	145	4790.0
2	2	60	NaT	103	135	3400.0
3	3	45	NaT	109	175	2824.0
4	4	45	NaT	117	148	4060.0
5	5	60	NaT	102	127	3000.0
6	6	60	NaT	110	136	3740.0
7	7	450	NaT	104	134	2533.0
8	8	30	NaT	109	133	1951.0
9	9	60	NaT	98	124	2690.0
10	10	60	NaT	103	147	3293.0
11	11	60	NaT	100	120	2507.0
12	12	60	NaT	100	120	2507.0
13	13	60	NaT	106	128	3453.0
14	14	60	NaT	104	132	3793.0
15	15	60	NaT	98	123	2750.0
16	16	60	NaT	98	120	2152.0
17	17	60	NaT	100	120	3000.0
18	18	45	NaT	90	112	0.0
19	19	60	NaT	103	123	3230.0
20	20	45	NaT	97	125	2430.0
21	21	60	NaT	108	131	3642.0
22	22	45	NaT	100	119	2820.0
23	23	60	NaT	130	101	3000.0
24	24	45	NaT	105	132	2460.0
25	25	60	NaT	102	126	3345.0
26	26	60	NaT	100	120	2500.0
27	27	60	NaT	92	118	2410.0
28	28	60	NaT	103	132	0.0
29	29	60	NaT	100	132	2800.0
30	30	60	NaT	102	129	3803.0
31	31	60	NaT	92	115	2430.0

Conjunto de dados após remover registros com valores nulos em 'Date':  
Empty DataFrame  
Columns: [ID, Duration, Date, Pulse, Maxpulse, Calories]  
Index: []

(env) MacBook-MBP:estacio-mundo5-missao-nivel-3 D\$

Ln 53, Col 55 Spaces: 4 UTF-8 LF Python 3.8.9 ('env': venv)