



Trabalho Prático | DGT2823 | Tecnologias

Para desenvolvimento de soluções de big data

DGT2823 - Tecnologias para desenv. De soluções de big data 🌐

Rian Joseph Ramos Felizardo - 202202G23G31

POLO BARREIRO - Belo Horizonte, MG

2025.2 - 6° Semestre Letivo

Objetivo 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);

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. O passo-a-passo de todo o processo de tratamento dos dados é apresentado a seguir, no roteiro de prática.

Estrutura da Base de Dados

ID	Duration	Date	Pulse	Maxpulse	Calories
0	60	2020/12/01	110	130	4091
1	60	2020/12/02	117	145	4790
2	60	2020/12/03	103	135	3400
3	45	2020/12/04	109	175	2824
4	45	2020/12/05	117	148	4060
5	60	2020/12/06	102	127	3000
6	60	2020/12/07	110	136	3740
7	450	2020/12/08	104	134	2533
8	30	2020/12/09	109	133	1951
9	60	2020/12/10	98	124	2690
10	60	2020/12/11	103	147	3293
11	60	2020/12/12	100	120	2507
12	60	2020/12/12	100	120	2507
13	60	2020/12/13	106	128	3453
14	60	2020/12/14	104	132	3793
15	60	2020/12/15	98	123	2750
16	60	2020/12/16	98	120	2152
17	60	2020/12/17	100	120	3000
18	45	2020/12/18	90	112	NaN
19	60	2020/12/19	103	123	3230
20	45	2020/12/20	97	125	2430
21	60	2020/12/21	108	131	3642
22	45	NaN	100	119	2820
23	60	2020/12/23	130	101	3000
24	45	2020/12/24	105	132	2460
25	60	2020/12/25	102	126	3345
26	60	2020/12/26	100	120	2500
27	60	2020/12/27	92	118	2410
28	60	2020/12/28	103	132	NaN
29	60	2020/12/29	100	132	2800
30	60	2020/12/30	102	129	3803
31	60	2020/12/31	92	115	2430

Procedimentos

Utilizando o conjunto de dados fornecido pela seção de contextualização nas microatividades:

Foi dado o nome de picoweb.csv para o arquivo criado seguindo o conjunto de dados fornecidos.

online+retail >  picoweb.csv	
1	ID;Duration;Date;Pulse;Maxpulse;Calories
2	0;60;'2020/12/01';110;130;4091
3	1;60;'2020/12/02';117;145;4790
4	2;60;'2020/12/03';103;135;3400
5	3;45;'2020/12/04';109;175;2824
6	4;45;'2020/12/05';117;148;4060
7	5;60;'2020/12/06';102;127;3000
8	6;60;'2020/12/07';110;136;3740
9	7;450;'2020/12/08';104;134;2533
10	8;30;'2020/12/09';109;133;1951
11	9;60;'2020/12/10';98;124;2690
12	10;60;'2020/12/11';103;147;3293
13	11;60;'2020/12/12';100;120;2507
14	12;60;'2020/12/12';100;120;2507
15	13;60;'2020/12/13';106;128;3453
16	14;60;'2020/12/14';104;132;3793
17	15;60;'2020/12/15';98;123;2750
18	16;60;'2020/12/16';98;120;2152
19	17;60;'2020/12/17';100;120;3000
20	18;45;'2020/12/18';90;112;NaN
21	19;60;'2020/12/19';103;123;3230
22	20;45;'2020/12/20';97;125;2430
23	21;60;'2020/12/21';108;131;3642
24	22;45;NaN;100;119;2820
25	23;60;'2020/12/23';130;101;3000
26	24;45;'2020/12/24';105;132;2460
27	25;60;'2020/12/25';102;126;3345
28	26;60;20201226;100;120;2500
29	27;60;'2020/12/27';92;118;2410
30	28;60;'2020/12/28';103;132;NaN
31	29;60;'2020/12/29';100;132;2800
32	30;60;'2020/12/30';102;129;3803
33	31;60;'2020/12/31';92;115;2430
34	

Passo seguinte

Foi criado um novo script com nome de **picoweb.ipynb** onde em sua execução ele:

- Lê o conteúdo fornecido pelo conjunto de dados.
- Atribui os dados lidos a variável criada (*dados*) que posteriormente será transformada em uma cópia para realizar edição dos dados.
- Nova variável criada com nome (*dados_copy*) onde tratamos os dados conforme o enunciado dos procedimentos:

Ilustração

```
print('-----')
print('----- Verificação de importação correta -----')

print('\nInformações gerais')
print(dados.info())

print('\nPrimeiras 10 linhas:')
print(dados.head(10))

print('\nÚltimas 10 linhas:')
print(dados.tail(10))

print('-----')
print('----- Manipulando o conjunto cópia -----')

dados_copy = dados.copy()

# substitui calories
dados_copy['Calories'] = dados_copy['Calories'].fillna(0)

# substitui os Date nulos
dados_copy['Date'] = dados_copy['Date'].fillna('1900/01/01')

# converte string específica
dados_copy['Date'] = dados_copy['Date'].replace(
    '20201226',
    pd.to_datetime('20201226', format='%Y%m%d')
)

# converte 1900/01/01 em NaT
dados_copy['Date'] = dados_copy['Date'].replace('1900/01/01', pd.NaT)

# agora converte a coluna inteira para datetime
dados_copy['Date'] = pd.to_datetime(dados_copy['Date'], errors='coerce')

print('-----')

✓ 0.0s
```

Finalização

Por fim apos termos os dados modelados e limpos conforme o enunciado temos o resultado da impressão final dos resultados esperados:

```
dados_copy = dados_copy.dropna()
print('-----')
print('----- DataFrame sem Nulos e Finalizado -----')
print('-----')
print(dados_copy)
✓ 0.0s

-----
----- DataFrame sem Nulos e finalizado -----
   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  3748.0
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  2698.0
10   10      60 2020-12-11    103     147  3293.0
11   11      60 2020-12-12    100     120  2507.0
12   12      60 2020-12-13    100     120  2507.0
13   13      60 2020-12-14    106     128  3453.0
14   14      60 2020-12-15    104     132  3793.0
15   15      60 2020-12-16     98     123  2750.0
16   16      60 2020-12-17    100     120  2152.0
17   17      60 2020-12-18     90     112     0.0
18   18      45 2020-12-19    103     123  3230.0
19   19      60 2020-12-20     97     125  2430.0
20   20      45 2020-12-21    108     131  3642.0
21   21      60 2020-12-23    130     101  3000.0
22   22      45 2020-12-24    105     132  2460.0
23   23      60 2020-12-25    102     126  3345.0
24   24      60 2020-12-26    100     120  2500.0
```

Finalização da Atividade Prática

Resultados esperados ⭐

O resultado esperado desta micro atividade é verificar se o aluno é capaz de extrair informações gerais sobre um conjunto de dados utilizando a biblioteca Pandas.

Repositório

https://github.com/rianjsp/microatividades_trab_2010/tree/main