FACULDADE ESTÁCIO DE SÁ

CURSO: DESENVOLVIMENTO FULL STACK

4º SEMESTRE – MATRÍCULA 202302595341

Repositório GitHub - alaimalmeida/tratantoImensidaoDosDados

ALAIM ALMEIDA DE OLIVEIRA

Tratando a imensidão dos dados

Salvador – BA

2024

1. Introdução

A manipulação de dados é uma tarefa essencial em diversas áreas, como análise de dados, ciência de dados e gestão de projetos. O formato CSV (Comma-Separated Values) é amplamente utilizado para armazenar e transferir dados devido à sua simplicidade e compatibilidade com diversas ferramentas. Neste relatório, abordaremos como ler arquivos CSV utilizando a linguagem Python e a biblioteca Pandas.

2. Objetivo

Demonstrar o processo de leitura de arquivos CSV utilizando a biblioteca Pandas, configurando corretamente parâmetros como separador de colunas, engine e encoding.

Procedimento básicos para criação do projeto

a. Instalar a biblioteca Pandas, caso ainda não esteja instalada:
“pip install pandas”

b. Importar a biblioteca Pandas no código Python:
“import pandas as pd”

c. Definir o nome do arquivo CSV que será lido:
“arquivo_csv = "dados.csv”

d. Ler o arquivo CSV usando a função read_csv, especificando o separador de colunas (sep), a engine (engine='python') e o encoding (encoding='utf-8'):
“tabela = pd.read_csv(arquivo_csv, sep=';', engine='python', encoding='utf-8')”

e. Exibir as primeiras linhas do DataFrame para verificar a leitura dos dados:
“print(tabela)”

Ficando dessa forma do código completo:

projeto.py

```python
import pandas as pd

# Definindo o nome do arquivo CSV
arquivo_csv = "dados.csv"

# Lendo o arquivo CSV
# Usando o separador ';', a engine 'python' e especificando o encoding (se necessário)
tabela = pd.read_csv(arquivo_csv, sep=';', engine='python', encoding='utf-8')

# Exibindo os dados lidos
print(tabela)

```

A parte do arquivo CSV ficou dessa forma:

projeto.py | dados.csv

dados.csv > data

```
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 2
1;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;20201226;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
```

Ao executar o arquivo, fica da seguinte forma:

SAÍDA CONSOLE DE DEPURAÇÃO PORTAS COMENTÁRIOS TERMINAL

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

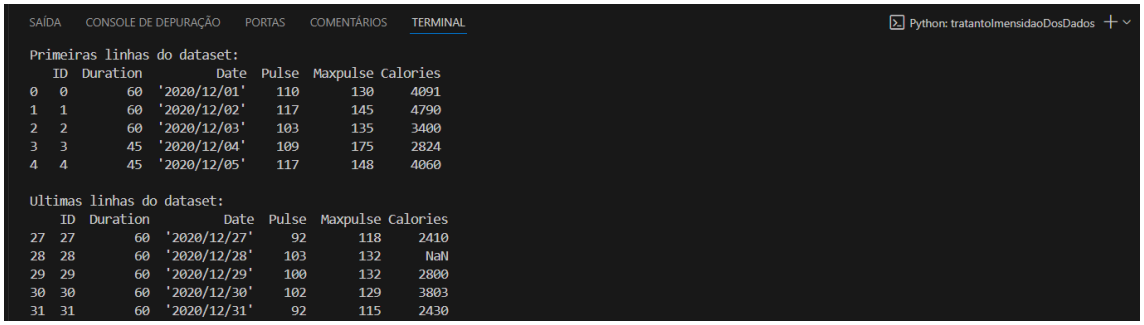
4. Atribua os dados lidos a uma variável;
5. Verifique se os dados foram importados adequadamente:
   a) Imprima as informações gerais sobre o conjunto de dados;
   b) Imprima as primeiras e últimas N linhas do arquivo.
6. Crie uma nova variável e atribua a ela uma cópia do conjunto de dados original (variável criada no passo 4);

Informações gerais sobre o conjunto de dados

SAÍDA CONSOLE DE DEPURAÇÃO PORTAS COMENTÁRIOS TERMINAL Python: tratantoImensidaoDosDados

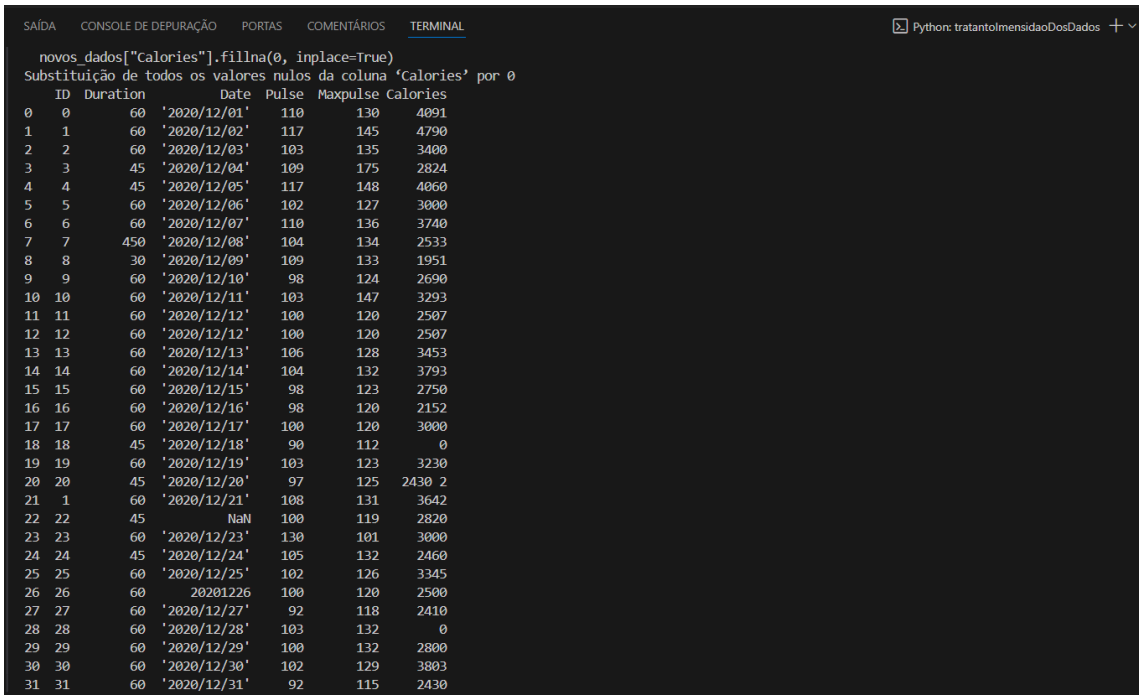
```
Informações gerais do dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        32 non-null     int64
 1   Duration  32 non-null     int64
 2   Date      31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

Primeiras e últimas linhas



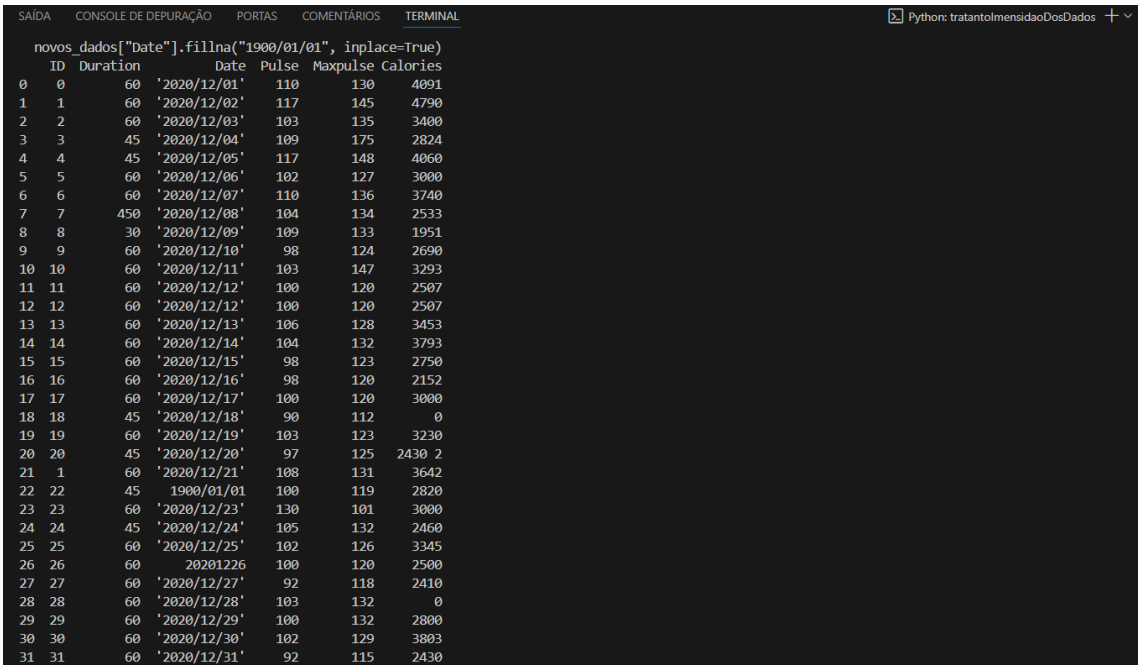

```
Primeiras linhas do dataset:
   ID  Duration          Date  Pulse  Maxpulse Calories
0   0        60  '2020/12/01'    110       130     4091
1   1        60  '2020/12/02'    117       145     4790
2   2        60  '2020/12/03'    103       135     3400
3   3        45  '2020/12/04'    109       175     2824
4   4        45  '2020/12/05'    117       148     4060

Ultimas linhas do dataset:
    ID  Duration          Date  Pulse  Maxpulse Calories
27  27        60  '2020/12/27'     92       118     2410
28  28        60  '2020/12/28'    103       132      NaN
29  29        60  '2020/12/29'    100       132     2800
30  30        60  '2020/12/30'    102       129     3803
31  31        60  '2020/12/31'     92       115     2430
```

7. Nessa nova variável, contendo uma cópia dos dados:
   a) Substitua todos os valores nulos da coluna ‘Calories’ por 0;
   b) Imprima o conjunto de dados para verificar se a mudança acima foi aplicada com
   c) sucesso;




```
  novos_dados["Calories"].fillna(0, inplace=True)
Substituição de todos os valores nulos da coluna ‘Calories’ por 0
    ID  Duration          Date  Pulse  Maxpulse Calories
0    0        60  '2020/12/01'    110       130     4091
1    1        60  '2020/12/02'    117       145     4790
2    2        60  '2020/12/03'    103       135     3400
3    3        45  '2020/12/04'    109       175     2824
4    4        45  '2020/12/05'    117       148     4060
5    5        60  '2020/12/06'    102       127     3000
6    6        60  '2020/12/07'    110       136     3740
7    7       450  '2020/12/08'    104       134     2533
8    8        30  '2020/12/09'    109       133     1951
9    9        60  '2020/12/10'     98       124     2690
10  10        60  '2020/12/11'    103       147     3293
11  11        60  '2020/12/12'    100       120     2507
12  12        60  '2020/12/12'    100       120     2507
13  13        60  '2020/12/13'    106       128     3453
14  14        60  '2020/12/14'    104       132     3793
15  15        60  '2020/12/15'     98       123     2750
16  16        60  '2020/12/16'     98       120     2152
17  17        60  '2020/12/17'    100       120     3000
18  18        45  '2020/12/18'     90       112        0
19  19        60  '2020/12/19'    103       123     3230
20  20        45  '2020/12/20'     97       125  2430 2
21   1        60  '2020/12/21'    108       131     3642
22  22        45           NaN    100       119     2820
23  23        60  '2020/12/23'    130       101     3000
24  24        45  '2020/12/24'    105       132     2460
25  25        60  '2020/12/25'    102       126     3345
26  26        60      20201226    100       120     2500
27  27        60  '2020/12/27'     92       118     2410
28  28        60  '2020/12/28'    103       132        0
29  29        60  '2020/12/29'    100       132     2800
30  30        60  '2020/12/30'    102       129     3803
31  31        60  '2020/12/31'     92       115     2430
```

8. Ainda na nova variável:
   a) Substitua os valores nulos da coluna ‘Date’ por ‘1900/01/01’;
   b) Imprima o conjunto de dados e confira se a mudança foi aplicada com sucesso;
   c) Transforme os dados da coluna ‘Date’ em datetime usando o método
   d) ‘to_datetime’;



9. Tendo seguido todas as instruções anteriores, ao executar o passo anterior você
10. deverá ter encontrado um erro informando que o valor ‘1900/01/01’ não
11. corresponde ao formato ‘%Y/%m/%d’. Para resolver esse problema:

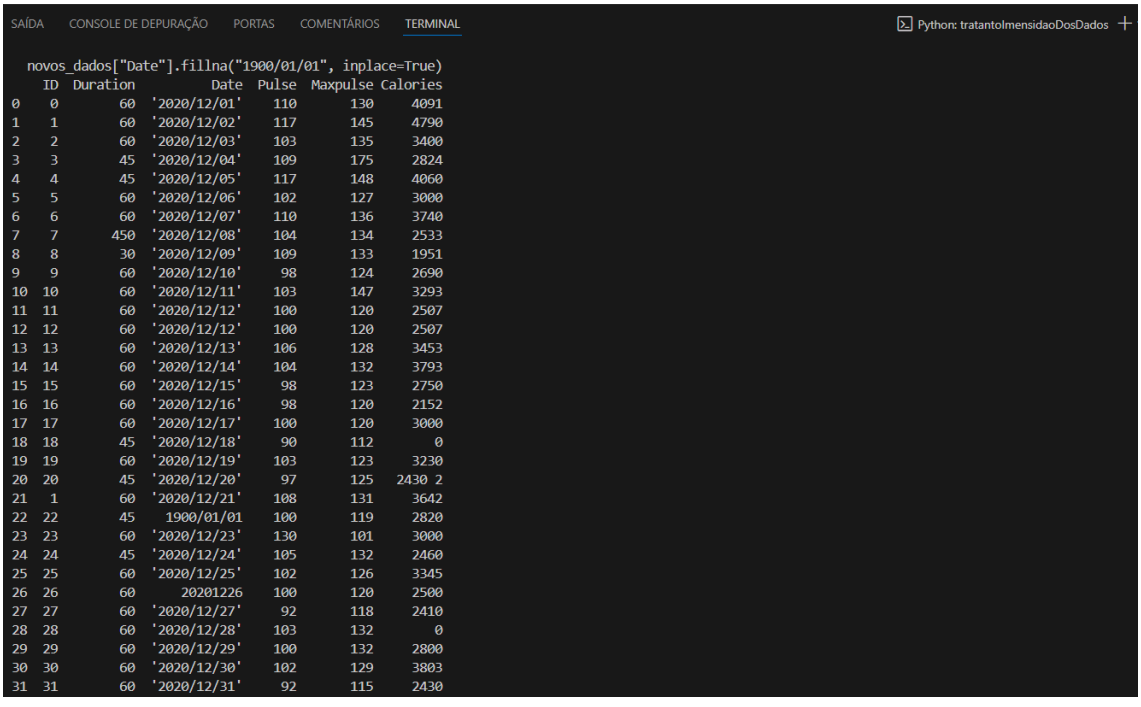
    a) Substitua, na coluna ‘Date’, o valor ‘1900/01/01’ por ‘NaN’;

    b) Utilizando o método ‘to_datetime’, repita o passo de transformação dos dados da

    c) coluna ‘Date’ para datetime;

    d) Imprima o conjunto de dados para verificar se as mudanças acima foram

    e) aplicadas com sucesso;



OBS: O valor aparece como NaT porque, no Pandas, essa é a representação padrão para dados ausentes em colunas do tipo datetime, enquanto NaN é usado para valores ausentes em colunas numéricas ou de texto.

10. Nesse ponto, você deverá ter esbarrado em outro erro, informando agora que o valor "20201226" não corresponde ao formato "'%Y/%m/%d'" . Você precisará, agora, na coluna ‘Date”, transformar especificamente esse valor, atualmente uma string, para o formato datetime. Para isso você deverá combinar os métodos ‘replace’ e ‘to_datetime’;
11. Após o passo anterior, execute novamente a transformação de todos os dados da coluna ‘Date’ para o formato datetime (usando o to_datetime). Imprima o conjunto de dados atual para verificar se todas as transformações foram executadas com sucesso;