Missão Prática – Mundo 05 – Nível 03

Gilvan Pereira de Oliveira – 2023.01.53256-61197

Polo Centro – São Lourenço Da Mata – PE

RPG0033 – TRATANDO A IMENSIDÃO DOS DADOS – 9001 – 2025.1

https://github.com/GilvanPOliveira/FullStack/tree/main/Mundo05/tratandoDados

Missão Prática | Tratando a imensidão dos dados

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.

tratando_dados.py

```python
import pandas as pd
import numpy as np

# 04
dados = pd.read_csv('dados.csv', sep=';', engine='python', encoding='utf-8')

# 05
print("Informacoes Gerais do Dataset:")
dados.info()
print("\nPrimeiras Linhas:")
print(dados.head())
print("\nUltimas Linhas:")
print(dados.tail())

# 06
dados_copia = dados.copy()

# 07
dados_copia['Calories'] = dados_copia['Calories'].fillna(0)
print("\nApos substituir nulos em 'Calories' por 0:")
print(dados_copia)

# 08
dados_copia['Date'] = dados_copia['Date'].fillna('1900/01/01')
print("\nApos substituir nulos em 'Date' por '1900/01/01':")
print(dados_copia)
dados_copia['Date'] = (
    dados_copia['Date']
    .astype(str)
    .str.strip("'")
    .str.strip('"')
)

```

```python

dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')
try:
    dados_copia['Date'] = pd.to_datetime(
        dados_copia['Date'],
        format='%Y/%m/%d'
    )
except Exception as e:
    print("\nErro na conversao de 'Date':", e)

# 09
dados_copia['Date'] = dados_copia['Date'].replace('1900/01/01', np.nan)
dados_copia['Date'] = pd.to_datetime(
    dados_copia['Date'],
    format='%Y/%m/%d',
    errors='coerce'
)
print("\nApos substituir '1900/01/01' por NaN e converter 'Date':")
print(dados_copia)

# 10
dados_copia = dados_copia.dropna(subset=['Date'])

# 11
dados_copia['Date'] = pd.to_datetime(
    dados_copia['Date'], format='%Y/%m/%d', errors='coerce')
print("\nApos correcao do valor '20201226' e conversao final de 'Date':")
print(dados_copia)

# 12
dados_copia = dados_copia.dropna(subset=['Date'])
print("\nDataset final apos remocao de registros com 'Date' nulo:")
print(dados_copia)

```

4. Atribua os dados lidos a uma variável;
5. Verifique se os dados foram importados adequadamente:
    1. Imprima as informações gerais sobre o conjunto de dados;
    2. 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);

```
[Running] python -u "                              \missaoPratica\tratando_dados.py"
Informacoes 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     float64
dtypes: float64(1), int64(4), object(1)
memory usage: 1.6+ KB
```

Primeiras Linhas:

	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

Ultimas Linhas:

	ID	Duration	Date	Pulse	Maxpulse	Calories
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

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

Apos substituir nulos em 'Calories' por 0:

	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	0.0
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	0.0
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

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

Apos substituir nulos em 'Date' por '1900/01/01':

	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	0.0
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	1900/01/01	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	0.0
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

9. Tendo seguido todas as instruções anteriores, ao executar o passo anterior você deverá ter encontrado um erro informando que o valor ‘1900/01/01’ não corresponde ao formato ‘%Y/%m/%d’. Para resolver esse problema:
    1. Substitua, na coluna ‘Date’, o valor ‘1900/01/01’ por ‘NaN’;
    2. Utilizando o método ‘to_datetime’, repita o passo de transformação dos dados da coluna ‘Date’ para datetime;
    3. Imprima o conjunto de dados para verificar se as mudanças acima foram aplicadas com sucesso;

```
Apos substituir '1900/01/01' por NaN e converter 'Date':
    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       0.0
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        NaT    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 2020-12-26    100       120    2500.0
27  27        60 2020-12-27     92       118    2410.0
28  28        60 2020-12-28    103       132       0.0
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
```

OBS: o valor está como NaT, pois no Pandas, é atribuído para date: NaT (not a time), e não, NaN (not a number).

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;

```
Apos correcao do valor '20201226' e conversao final de 'Date':
    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       0.0
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
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 2020-12-26    100       120    2500.0
27  27        60 2020-12-27     92       118    2410.0
28  28        60 2020-12-28    103       132       0.0
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
```

12. Por fim, remova os registros contendo valores nulos. Nesse ponto, apenas a coluna ‘Date’ possui um registro que atende a essa premissa (linha 22). Logo, utilize-a como base para realizar a transformação solicitada;
13. Imprima o dataframe e verifique se todas as transformações foram executadas conforme solicitado nos passos anteriores.

```
Dataset final apos remocao de registros com 'Date' nulo:
    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       0.0
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
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 2020-12-26    100       120    2500.0
27  27        60 2020-12-27     92       118    2410.0
28  28        60 2020-12-28    103       132       0.0
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

[Done] exited with code=0 in 0.733 seconds
```