



Curso :Desenvolvimento Full Stack

Aluno:Leonardo Cipriano
Disciplina : Tratando a imensidão dos dados
Código :DGT2823

Missão Prática | Tratando a imensidão dos dados

Este projeto é focado na manipulação e tratamento de dados com a biblioteca Pandas. O objetivo principal é o processamento de dados de um arquivo CSV, corrigindo datas, tratando valores nulos e garantindo a integridade do conjunto de dados.

Baseando-se na especificação "Missão Prática _ Nível 3 _ Mundo 5.pdf" ou no link <https://sway.cloud.microsoft/BbCATE0NxSrJBF5p?ref=Link>

A Missão

O objetivo da missão prática "Tratando a imensidão dos dados" é capacitar o aluno a realizar a limpeza de um conjunto de dados para prepará-lo para análises posteriores. O aluno deve usar Python e a biblioteca Pandas para ler dados de um arquivo CSV, identificar e corrigir valores nulos, formatar corretamente as colunas de data e remover registros inadequados. O desafio também inclui a manipulação de tipos de dados e o uso de métodos para transformar strings em formatos apropriados para análise e mineração de dados.

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. 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.

Etapas

1 - Leitura do arquivo CSV.

Arquivo mp1.py

- Ler o conteúdo do CSV fornecido, atentando-se para a necessidade ou não de incluir parâmetros adicionais como os relativos ao separador dos dados, a engine e o encoding;
- Atribua os dados lidos a uma variável;
- Verifique se os dados foram importados adequadamente;
- Imprima as informações gerais sobre o conjunto de dados;

Missão Prática | Tratando a imensidão dos dados

The screenshot shows the Visual Studio Code interface. The Explorer sidebar on the left lists several projects and files, including 'missao' (containing 'dados.csv'), 'micro4', 'micro1', 'micro2', 'micro3', 'micro5', and 'Missão Prática N3 M5.pdf'. The code editor on the right displays the content of 'mpi1.py':

```
from datetime import datetime
import pandas as pd
import numpy as np
import sys

# 3. Ler o conteúdo do CSV fornecido, especificando separador, engine, e encoding
sys.stdout.reconfigure(encoding='utf-8')

# 4. Atribua os dados lidos a uma variável;
arquivo_csv = 'dados.csv'
dados = pd.read_csv(arquivo_csv, sep=';', engine='python', encoding='utf-8')

# 5 Verifique se os dados foram importados adequadamente:
# a) Imprimir as informações gerais sobre o conjunto de dados
print(" 5.a) Informações gerais sobre o conjunto de dados:")
print(dados.info())
```

Resultado : mp1.py

```
Windows PowerShell x + - PS D:\estudo\estacio\periodo 5-2024\N3\missao_pratica\missao> python mpl.py
5.a) 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     object  
dtypes: int64(4), object(2)
memory usage: 1.6+ KB
None
PS D:\estudo\estacio\periodo 5-2024\N3\missao_pratica\missao> |
```

- Foram impressas as primeiras e últimas N linhas do arquivo. Arquivo mp2.py



Missão Prática | Tratando a imensidão dos dados

Resultado : mp2.y

```
Windows PowerShell
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp2.py
Primeiras 10 linhas do conjunto de dados:
   ID Duration      Date    Pulse Maxpulse Calories
0   0       60 '2020/12/01'    110     130    4091
1   1       60 '2020/12/01'    117     145    4790
2   2       60 '2020/12/01'    103     135    3500
3   3       45 '2020/12/04'    109     175    2820
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

Últimas 10 linhas do conjunto de dados:
   ID Duration      Date    Pulse Maxpulse Calories
22  22       45 '2020/12/23'    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 '2020/12/05'    106     120    2560
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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao>
```

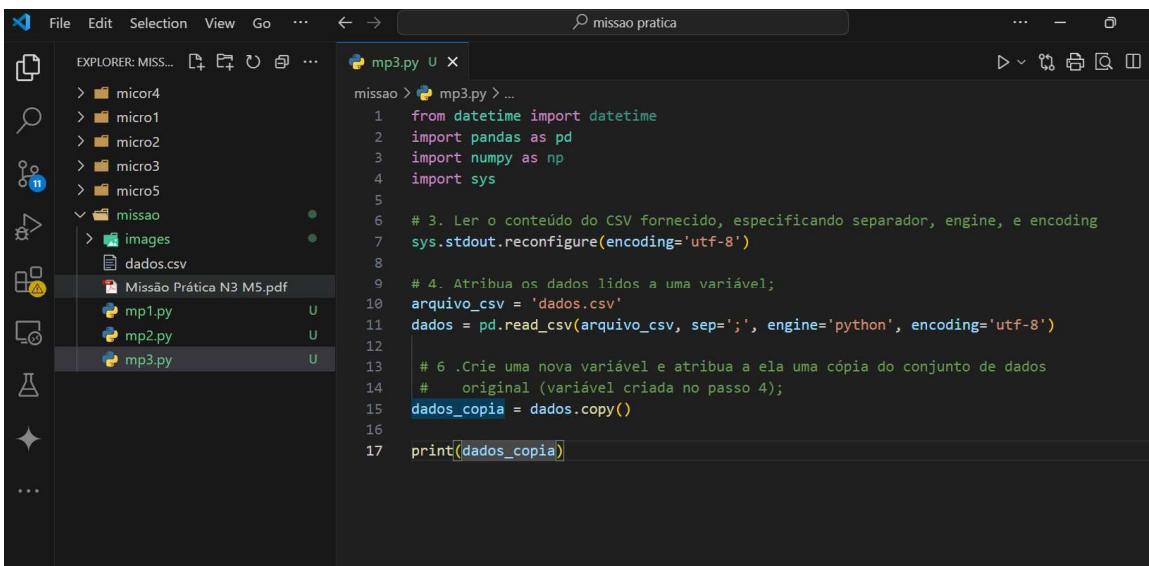
Missão Prática | Tratando a imensidão dos dados

2 - *Tratamento de valores nulos: Valores nulos na coluna 'Calories' substituídos por 0.*

Arquivo mp3.py

Aqui foi criado uma nova variável e atribua a ela uma cópia do conjunto de dados original. Nessa nova variável, contendo uma cópia dos dados:

- Foram impressos o conjunto de dados afim de verificar se a mudança acima foi aplicada com sucesso;



```
File Edit Selection View Go ... ← → ⌂ missao pratica
EXPLORER: MISSAO PRATICA
  > micor4
  > micro1
  > micro2
  > micro3
  > micro5
  > missao
    > images
    > dados.csv
    Missao Pratica N3 M5.pdf
  mp1.py
  mp2.py
  mp3.py

mp3.py U X
missao > mp3.py > ...
1  from datetime import datetime
2  import pandas as pd
3  import numpy as np
4  import sys
5
6  # 3. Ler o conteúdo do CSV fornecido, especificando separador, engine, e encoding
7  sys.stdout.reconfigure(encoding='utf-8')
8
9  # 4. Atribua os dados lidos a uma variável;
10 arquivo_csv = 'dados.csv'
11 dados = pd.read_csv(arquivo_csv, sep=';', engine='python', encoding='utf-8')
12
13 # 6 .Crie uma nova variável e atribua a ela uma cópia do conjunto de dados
14 #   original (variável criada no passo 4);
15 dados_copia = dados.copy()
16
17 print(dados_copia)
```

Missão Prática | Tratando a imensidão dos dados

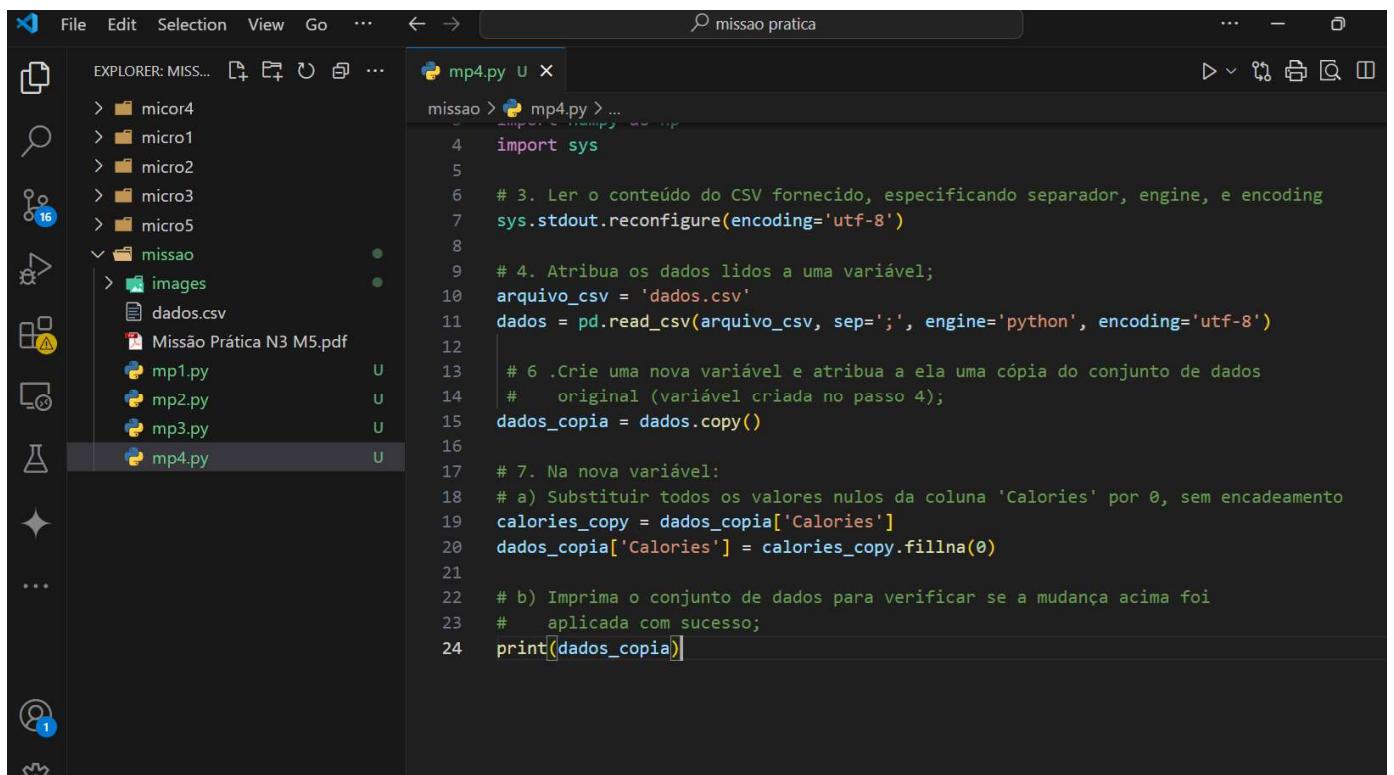
Resultados : mp3.y

```
Windows PowerShell x + v
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp3.py
   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  21      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 '2020/12/26'  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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao>
```

Missão Prática | Tratando a imensidão dos dados

- Aqui foram substituídos todos os valores nulos da coluna 'Calories' por 0;

Arquivo : mp4.py



The screenshot shows a code editor window with the file 'mp4.py' open. The left sidebar displays a file tree with several Python files ('mp1.py', 'mp2.py', 'mp3.py', 'mp4.py') and a CSV file ('dados.csv'). The main editor area contains the following Python code:

```
EXPLORER: MISSAO PRATICA
File Edit Selection View Go ...
missao > mp4.py > ...
4 import sys
5
6 # 3. Ler o conteúdo do CSV fornecido, especificando separador, engine, e encoding
7 sys.stdout.reconfigure(encoding='utf-8')
8
9 # 4. Atribua os dados lidos a uma variável;
10 arquivo_csv = 'dados.csv'
11 dados = pd.read_csv(arquivo_csv, sep=';', engine='python', encoding='utf-8')
12
13 # 6 .Crie uma nova variável e atribua a ela uma cópia do conjunto de dados
14 # original (variável criada no passo 4);
15 dados_copia = dados.copy()
16
17 # 7. Na nova variável:
18 # a) Substituir todos os valores nulos da coluna 'Calories' por 0, sem encadeamento
19 calories_copy = dados_copia['Calories']
20 dados_copia['Calories'] = calories_copy.fillna(0)
21
22 # b) Imprima o conjunto de dados para verificar se a mudança acima foi
23 # aplicada com sucesso;
24 print(dados_copia)
```



Missão Prática | Tratando a imensidão dos dados

Resultado : mp4.py

ID	Duration	Date	Pulse	Maxpulse	Calories
0	60	'2028/12/01'	116	138	4891
1	60	'2028/12/02'	117	145	4790
2	60	'2028/12/03'	103	135	3400
3	45	'2028/12/04'	109	175	2824
4	45	'2028/12/05'	117	148	4060
5	60	'2028/12/06'	102	127	3800
6	60	'2028/12/07'	118	136	3740
7	450	'2028/12/08'	104	134	2533
8	8	'2028/12/09'	109	133	1951
9	9	'2028/12/10'	98	124	2690
10	60	'2028/12/11'	103	147	3293
11	11	'2028/12/12'	100	128	2507
12	60	'2028/12/12'	100	128	2507
13	13	'2028/12/13'	106	128	3453
14	60	'2028/12/14'	104	132	3793
15	60	'2028/12/15'	98	123	2750
16	60	'2028/12/16'	98	120	2152
17	60	'2028/12/17'	100	128	3000
18	45	'2028/12/18'	90	112	NaN
19	60	'2028/12/19'	103	123	3230
20	45	'2028/12/20'	97	125	2438.2
21	1	'2028/12/21'	108	131	3642
22	22	NaN	100	119	2820
23	23	'2028/12/23'	130	101	3000
24	45	'2028/12/24'	105	132	2460
25	25	'2028/12/25'	102	126	3345
26	60	'2028/12/26'	100	128	2500
27	60	'2028/12/27'	92	118	2410
28	60	'2028/12/28'	103	132	NaN
29	60	'2028/12/29'	100	132	2800
30	60	'2028/12/30'	102	129	3803
31	60	'2028/12/31'	92	115	2430

ID	Duration	Date	Pulse	Maxpulse	Calories
0	60	'2028/12/01'	116	138	4891
1	1	'2028/12/02'	117	145	4790
2	2	'2028/12/03'	103	135	3400
3	3	'2028/12/04'	109	175	2824
4	4	'2028/12/05'	117	148	4060
5	5	'2028/12/06'	102	127	3800
6	6	'2028/12/07'	118	136	3740
7	7	'2028/12/08'	104	134	2533
8	8	'2028/12/09'	109	133	1951
9	9	'2028/12/10'	98	124	2690
10	10	'2028/12/11'	103	147	3293
11	11	'2028/12/12'	100	128	2507
12	12	'2028/12/12'	100	128	2507
13	13	'2028/12/13'	106	128	3453
14	14	'2028/12/14'	104	132	3793
15	15	'2028/12/15'	98	123	2750
16	16	'2028/12/16'	98	128	2152
17	17	'2028/12/17'	100	128	3000
18	18	'2028/12/18'	90	112	0
19	19	'2028/12/19'	103	123	3230
20	20	'2028/12/20'	97	125	2438.2
21	1	'2028/12/21'	108	131	3642
22	22	45	NaN	100	2820
23	23	'2028/12/23'	130	101	3000
24	24	'2028/12/24'	105	132	2460
25	25	'2028/12/25'	102	126	3345
26	26	'2028/12/26'	100	128	2500
27	27	'2028/12/27'	92	118	2410
28	28	'2028/12/28'	103	132	0
29	29	'2028/12/29'	100	132	2800
30	30	'2028/12/30'	102	129	3803
31	31	'2028/12/31'	92	115	2430

Missão Prática | Tratando a imensidão dos dados

3 - Formatação de datas: A coluna 'Date' foi ajustada, substituindo valores nulos por '1900/01/01', corrigindo formatação de strings e transformando a coluna para datetime

- Substitua os valores nulos da coluna 'Date' por '1900/01/01';
 - Imprimir o conjunto de dados e confira se a mudança foi aplicada com sucesso;

Missão Prática | Tratando a imensidão dos dados

Resultado de mp5.py

```

MPS          x  MP4          x  +  v
D:\estudo\estacio\periodo 5-2024\W3\missao pratica\missao> python mp4.py
ID Duration Date Pulse Maxpulse Calories
0 60 '2028/12/01' 110 130 4091
1 60 '2028/12/02' 117 145 4798
2 60 '2028/12/03' 103 135 3480
3 45 '2028/12/04' 189 175 2874
4 45 '2028/12/05' 117 148 4060
5 60 '2028/12/06' 102 127 3000
6 60 '2028/12/07' 110 136 3708
7 450 '2028/12/08' 104 134 2533
8 38 '2028/12/09' 109 133 1951
9 60 '2028/12/10' 98 124 2699
10 60 '2028/12/11' 103 147 3293
11 60 '2028/12/12' 100 120 2587
12 60 '2028/12/13' 106 120 2567
13 60 '2028/12/14' 106 128 3453
14 60 '2028/12/15' 104 132 3793
15 60 '2028/12/16' 98 123 2758
16 60 '2028/12/17' 98 120 2152
17 60 '2028/12/18' 100 120 3060
18 45 '2028/12/19' 90 112 0
19 60 '2028/12/20' 103 123 3230
20 45 '2028/12/21' 97 125 2430 2
1 60 '2028/12/22' 108 131 3642
22 45 NaN 109 119 2820
23 60 '2028/12/23' 130 101 3000
24 45 '2028/12/24' 105 132 2460
25 60 '2028/12/25' 102 126 3345
26 60 20281226 100 120 2500
27 60 '2028/12/27' 92 118 2410
28 60 '2028/12/28' 103 132 0
29 60 '2028/12/29' 100 132 2800
30 60 '2028/12/30' 102 129 3883
31 60 '2028/12/31' 92 115 2430
D:\estudo\estacio\periodo 5-2024\W3\missao pratica\missao>

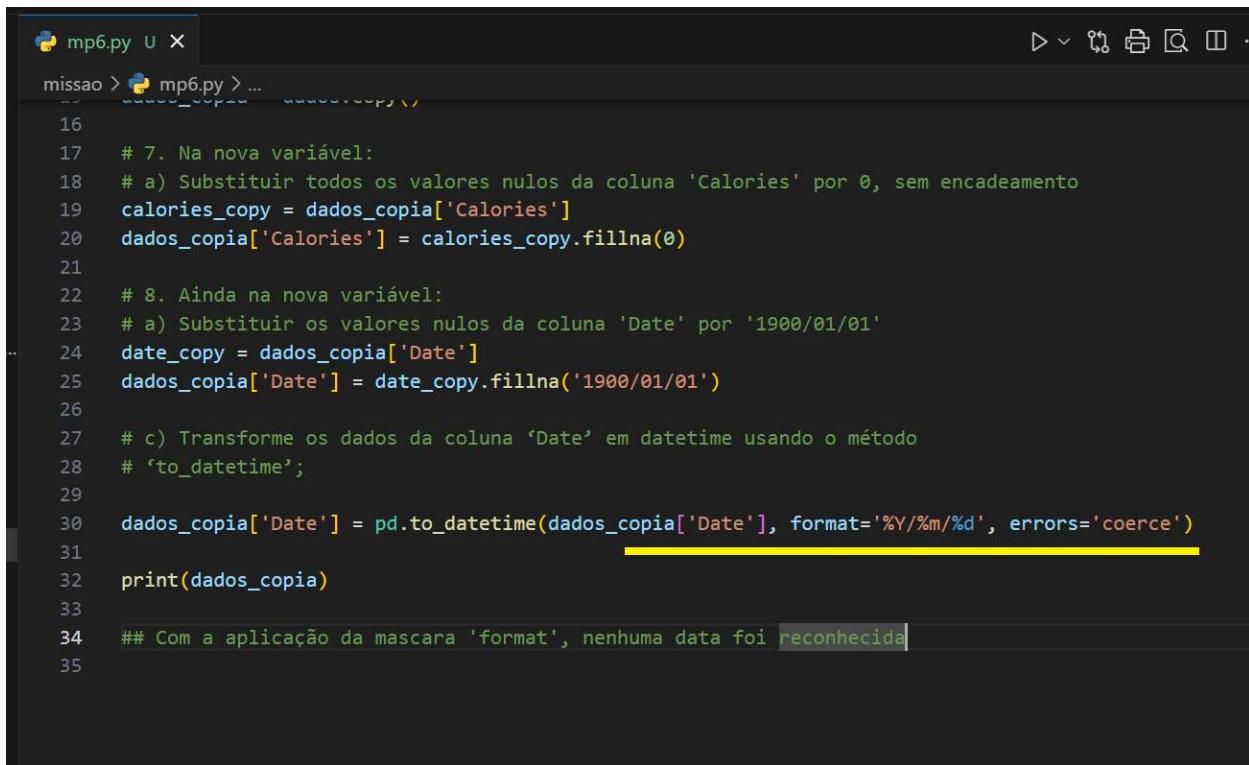
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp5.py
ID Duration Date Pulse Maxpulse Calories
0 0 60 '2028/12/01' 110 130 4091
1 1 60 '2028/12/02' 117 145 4798
2 2 60 '2028/12/03' 103 135 3400
3 3 45 '2028/12/04' 109 175 2824
4 4 45 '2028/12/05' 117 148 4060
5 5 60 '2028/12/06' 102 127 3000
6 6 60 '2028/12/07' 110 136 3748
7 7 450 '2028/12/08' 104 134 2533
8 8 38 '2028/12/09' 109 133 1951
9 9 60 '2028/12/10' 98 124 2699
10 10 60 '2028/12/11' 103 147 3293
11 11 60 '2028/12/12' 100 120 2587
12 12 60 '2028/12/13' 100 120 2507
13 13 60 '2028/12/14' 106 128 3453
14 14 60 '2028/12/15' 104 132 3793
15 15 60 '2028/12/16' 98 123 2758
16 16 60 '2028/12/17' 98 120 2152
17 17 60 '2028/12/18' 100 120 3060
18 18 45 '2028/12/19' 90 112 0
19 19 60 '2028/12/20' 103 123 3230
20 20 45 '2028/12/21' 97 125 2430 2
21 21 60 '2028/12/22' 108 131 3642
22 22 45 1900/01/01 100 119 2820
23 23 60 '2028/12/23' 106 101 3000
24 24 45 '2028/12/24' 105 120 2460
25 25 60 '2028/12/25' 102 126 3345
26 26 60 20281226 100 120 2500
27 27 60 '2028/12/27' 92 118 2410
28 28 60 '2028/12/28' 103 132 0
29 29 60 '2028/12/29' 100 132 2800
30 30 60 '2028/12/30' 102 129 3883
31 31 60 '2028/12/31' 92 115 2430
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao>

```

Missão Prática | Tratando a imensidão dos dados

- Transformar os dados da coluna ‘Date’ em datetime usando o método ‘to_datetime’;

Arquivo mp6.py



```
mp6.py  U  X
missao > mp6.py > ...
16
17  # 7. Na nova variável:
18  # a) Substituir todos os valores nulos da coluna 'Calories' por 0, sem encadeamento
19  calories_copy = dados_copia['Calories']
20  dados_copia['Calories'] = calories_copy.fillna(0)
21
22  # 8. Ainda na nova variável:
23  # a) Substituir os valores nulos da coluna 'Date' por '1900/01/01'
24  date_copy = dados_copia['Date']
25  dados_copia['Date'] = date_copy.fillna('1900/01/01')
26
27  # c) Transforme os dados da coluna 'Date' em datetime usando o método
28  # 'to_datetime';
29
30  dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], format='%Y/%m/%d', errors='coerce')
31
32  print(dados_copia)
33
34  ## Com a aplicação da máscara 'format', nenhuma data foi reconhecida
35
```

Aqui devido a formatação foi gerado um erro onde nenhuma data foi reconhecida

Missão Prática | Tratando a imensidão dos dados

Resultado mp6.y

```
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp6.py
          ID Duration      Date  Pulse Maxpulse Calories
 0         0       60    NaT     110      130    4091
 1         1       60    NaT     117      145    4790
 2         2       60    NaT     103      135    3400
 3         3       45    NaT     109      175    2824
 4         4       45    NaT     117      148    4060
 5         5       60    NaT     102      127    3000
 6         6       60    NaT     110      136    3740
 7         7      450    NaT     104      134    2533
 8         8       30    NaT     109      133    1951
 9         9       60    NaT      98      124    2690
10        10      60    NaT     103      147    3293
11        11      60    NaT     100      120    2507
12        12      60    NaT     100      120    2507
13        13      60    NaT     106      128    3453
14        14      60    NaT     104      132    3793
15        15      60    NaT      98      123    2750
16        16      60    NaT      98      120    2152
17        17      60    NaT     100      120    3000
18        18      45    NaT      90      112      0
19        19      60    NaT     103      123    3230
20        20      45    NaT      97      125    2430  2
21        1       60    NaT     108      131    3642
22        22      45 1900-01-01     100      119    2820
23        23      60    NaT     130      101    3000
24        24      45    NaT     105      132    2460
25        25      60    NaT     102      126    3345
26        26      60    NaT     100      120    2500
27        27      60    NaT      92      118    2410
28        28      60    NaT     103      132      0
29        29      60    NaT     100      132    2800
30        30      60    NaT     102      129    3803
31        31      60    NaT      92      115    2430
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> |
```

Missão Prática | Tratando a imensidão dos dados

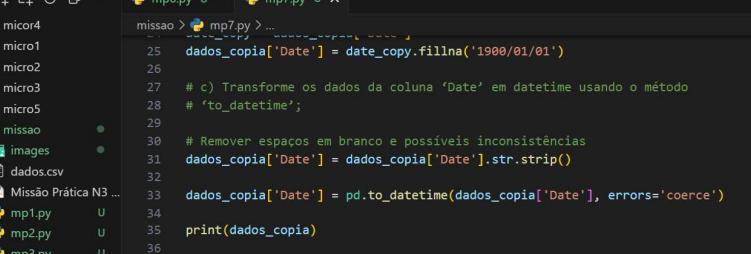
4 - Remoção de registros inválidos: Registros com valores nulos restantes na coluna 'Date' foram removidos.

- Após seguir todas as instruções anteriores, ao executar o passo anterior foi encontrado um erro informando que o valor '1900/01/01' não corresponde ao formato '%Y/%m/%d'.

Para resolver esse problema:

- Substitua, na coluna 'Date', o valor '1900/01/01' por 'NaN';
 - Utilizando o método 'to_datetime', repita o passo de transformação dos dados da coluna 'Date' para datetime;
- arquivo mp7.py

Missão Prática | Tratando a imensidão dos dados



```
File Edit Selection View Go ... missao pratica

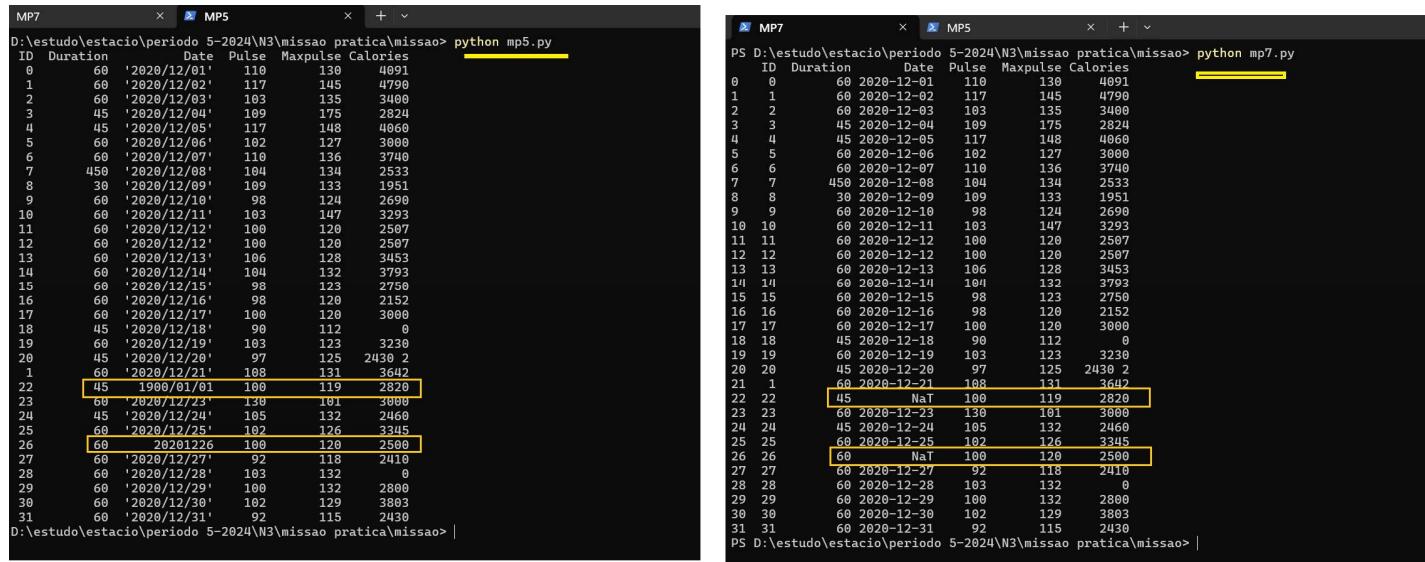
E... F... O... B... mp6.py U mp7.py U

missao > mp7.py > ...
25 dados_copia['Date'] = date_copy.fillna('1900/01/01')
26
27 # c) Transforme os dados da coluna 'Date' em datetime usando o método
28 # 'to_datetime';
29
30 # Remover espaços em branco e possíveis inconsistências
31 dados_copia['Date'] = dados_copia['Date'].str.strip()
32
33 dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], errors='coerce')
34
35 print(dados_copia)
36
37 # Aqui foi removida a formatação e todos os itens abaixo foram satisfeitos
38
39 # 9 - Tendo seguido todas as instruções anteriores, ao executar o passo anterior
40 # você deverá ter encontrado um erro informando que o valor '1900/01/01' não
41 # corresponde ao formato '%Y/%m/%d'. Para resolver esse problema:
42
43 # a) Substitua, na coluna 'Date', o valor '1900/01/01' por 'NaN';
44 # b) Utilizando o método 'to_datetime', repita o passo de transformação dos
45 #     dados da coluna 'Date' para datetime;
46 # c) Imprima o conjunto de dados para verificar se as mudanças acima foram
47 # aplicadas com sucesso;
```

- Imprima o conjunto de dados para verificar se as mudanças acima foram aplicadas com sucesso

Resultado mp7.py

Missão Prática | Tratando a imensidão dos dados



```
D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp5.py
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 45 8 '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 '2020/12/13' 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 0
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 '1986/01/01' 100 119 2820
23 60 '2020/12/23' 130 181 3000
24 45 '2020/12/24' 105 132 2160
25 60 '2020/12/25' 102 126 3245
26 60 20201226 100 128 2500
D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao>

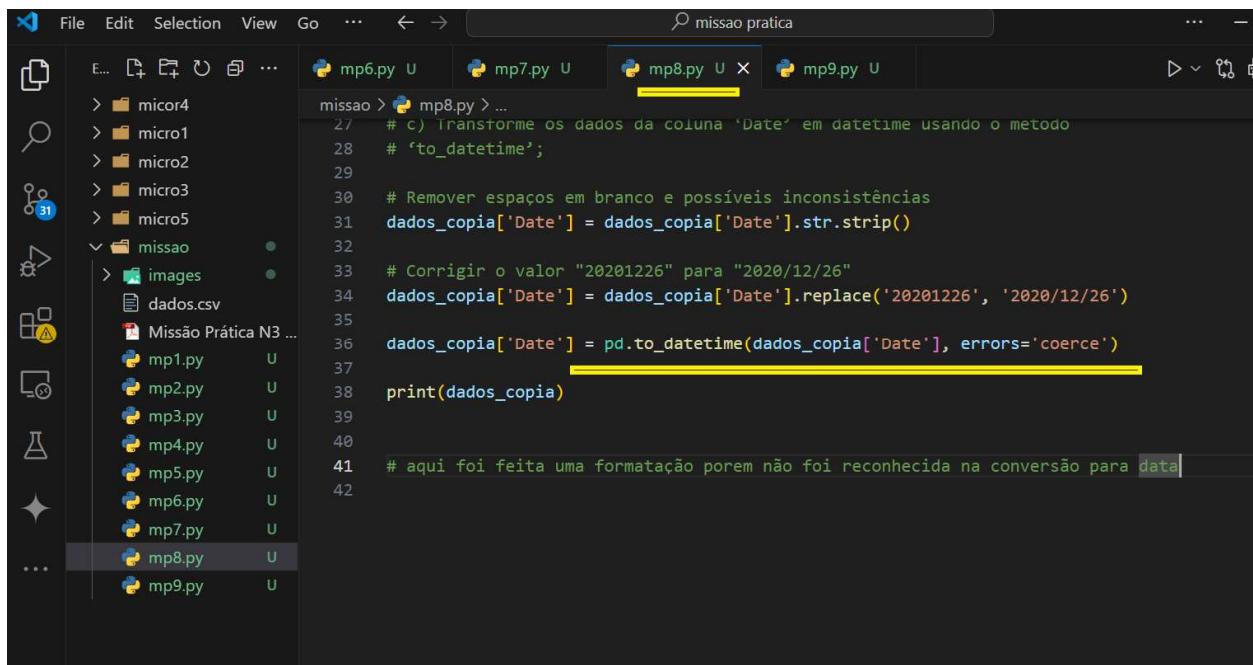
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp7.py
ID Duration Date Pulse Maxpulse Calories
0 0 60 2020-12-01 110 130 4091
1 1 60 2020-12-02 117 145 4790
2 2 60 2020-12-03 103 135 3400
3 3 45 2020-12-04 109 175 2824
4 4 45 2020-12-05 117 148 4060
5 5 60 2020-12-06 102 127 3000
6 6 60 2020-12-07 110 136 3740
7 7 45 8 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 21 60 2020-12-21 108 131 3642
22 22 45 NaT 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 NaT 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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao>
```

Aqui pode ser notado que os valores incompatíveis não foram convertidos

- Nesse ponto, foi notado outro erro, informando agora que o valor "20201226" não corresponde ao formato "%Y/%m/%d" .

- arquivo mp8.py

Missão Prática | Tratando a imensidão dos dados

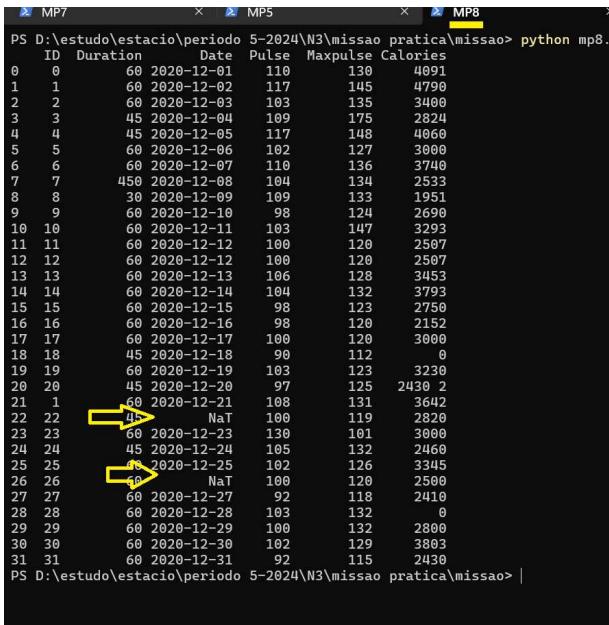


The screenshot shows a code editor interface with a dark theme. The left sidebar contains icons for file operations like new, open, save, and search, along with a tree view of project files. The main area displays the content of the file `mp8.py`. The code is written in Python and performs data manipulation on a CSV file named `dados.csv`.

```
missao > mp8.py > ...
27 # c) Transforme os dados da coluna 'Date' em datetime usando o metodo
28 # 'to_datetime';
29
30 # Remover espaços em branco e possíveis inconsistências
31 dados_copia['Date'] = dados_copia['Date'].str.strip()
32
33 # Corrigir o valor "20201226" para "2020/12/26"
34 dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')
35
36 dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], errors='coerce')
37
38 print(dados_copia)
39
40
41 # aqui foi feita uma formatação porém não foi reconhecida na conversão para data|
```

Resultado mp8.py

Missão Prática | Tratando a imensidão dos dados



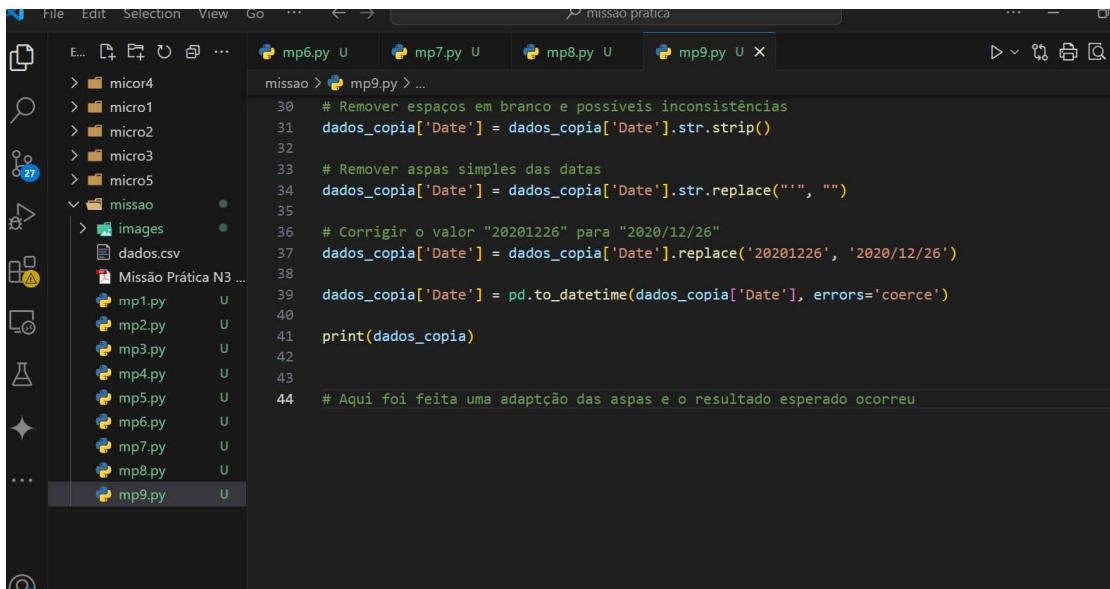
```
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp8
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp8
   ID Duration      Date    Pulse Maxpulse Calories
 0  0       60 2020-12-01    110     130    4091
 1  1       60 2020-12-02    117     145    4790
 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    3800
 6  6       60 2020-12-07    110     136    3740
 7  7       45 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 21      60 2020-12-21    108     131    3642
22 22      60 2020-12-21    NaT     100     119    2820
23 23      60 2020-12-23    130     101    3800
24 24      45 2020-12-24    105     132    2460
25 25      60 2020-12-25    102     126    3345
26 26      60 2020-12-26    NaT     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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> |
```

Missão Prática | Tratando a imensidão dos dados

Aqui foi preciso, na coluna ‘Date”, transformar especificamente esse valor,mutualmente uma string, para o formato datetime.

- Após o passo anterior, foi executada novamente a transformação de todos os dados da coluna ‘Date’ para o formato datetime (usando o to_datetime).

Arquivo mp9.py



```
File Edit Selection View Go ... < > missao práta ... mp6.py U mp7.py U mp8.py U mp9.py X
missao > mp9.py > ...
30     # Remover espaços em branco e possíveis inconsistências
31     dados_copia['Date'] = dados_copia['Date'].str.strip()
32
33     # Remover aspas simples das datas
34     dados_copia['Date'] = dados_copia['Date'].str.replace("'", "")
35
36     # Corrigir o valor "20201226" para "2020/12/26"
37     dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')
38
39     dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], errors='coerce')
40
41     print(dados_copia)
42
43
44     # Aqui foi feita uma adaptação das aspas e o resultado esperado ocorreu
```

- Foi Impresso o conjunto de dados atual para verificar se todas as transformações foram executadas com sucesso;

Missão Prática | Tratando a imensidão dos dados

Resultado mp9.py

```
MP7      MP5      MP8      MP9      + 
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp9.py
   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 21      60 2020-12-21 108    131    3642
22 22      45 1900-01-01 100    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 2020-12-26 100    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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> |
```



Missão Prática | Tratando a imensidão dos dados

- Por fim, foi removido os registros contendo valores nulos. Nesse ponto, apenas a coluna 'Date' possui um registro.

Arquivo mp10.py

The screenshot shows a Jupyter Notebook environment with the following details:

- File Bar:** File, Edit, Selection, View, Go, ..., search bar containing "missao pratica".
- Left Sidebar:** Includes icons for file operations like Open, Save, Run, and Help, along with a tree view of the project structure:
 - micor4
 - micra1
 - micro2
 - micro3
 - micro5
 - missao
 - dados.csv
 - Missão Prática N3 ...
 - mp1.py
 - mp2.py
 - mp3.py
 - mp4.py
 - mp5.py
 - mp6.py
 - mp7.py
 - mp8.py
 - mp9.py
 - mp10.py
- Code Cell:** The current cell contains the following Python code:

```
17
18     date_copy = dados_copia['Date']
19     dados_copia['Date'] = date_copy.fillna('1900/01/01')
20
21
22     dados_copia['Date'] = dados_copia['Date'].str.strip()
23
24 # Remover aspas simples das datas
25     dados_copia['Date'] = dados_copia['Date'].str.replace("'", "")
26
27 # Corrigir o valor "20201226" para "2020/12/26"
28     dados_copia['Date'] = dados_copia['Date'].replace('20201226', '2020/12/26')
29
30     dados_copia['Date'] = pd.to_datetime(dados_copia['Date'], errors='coerce')
31
32 # 12. Por fim, remova os registros contendo valores nulos. Nesse ponto, apenas a
33 #       coluna 'Date' possui um registro que atende a essa premissa (linha 22). Logo,
34 #       utilize-a como base para realizar a transformação solicitada;
35
36     dados_copia = dados_copia.dropna(subset=['Date'])
37
38 # 12. Imprima o dataframe e verifique se todas as transformações foram
39 #       executadas conforme solicitado nos passos anteriores.
40
41 print(dados_copia)
42
43
```

Missão Prática | Tratando a imensidão dos dados

- Impressão do dataframe , verificação de todas as transformações que foram executadas conforme solicitado nos passos anteriores.

Resultado mp10.py

```
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> python mp10.py
   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   21       60 2020-12-21     108      131    3642
22   22       45 1900-01-01     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 2020-12-26     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
PS D:\estudo\estacio\periodo 5-2024\N3\missao pratica\missao> |
```