Gabriel Henrique dos Santos – Matrícula 202208292411

Polo: Limeira Centro

Turma 2023.3

MUNDO 5 - Tratando a Imensidão Dos Dados

Resultados de cada uma das alterações solicitadas no roteiro da prática:

Visualizando os dados selecionados e montando uma tabela:

```
data = pd.read_csv('Database.txt', sep=';', encoding='utf-8', engine='python')

# Atribua os dados lidos a uma variável e verifique se foram importados adequadamente
# Informações gerais sobre o conjunto de dados
print("Informações gerais do conjunto de dados:")
print(data.info())
data = pd.DataFrame (data)

print("Tabela de dados")
print (data)
```

```
Informações gerais do 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
Tabela de dados
    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
```

Visualizando as primeiras e ultimas linhas do conteúdo:

```
# Imprima as primeiras e últimas N linhas do arquivo (suponhamos N=5)
print("\nPrimeiras 5 linhas do conjunto de dados:")
print(data.head(5))
print("\nÚltimas 5 linhas do conjunto de dados:")
print(data.tail(5))
```

```
Primeiras 5 linhas do conjunto de dados:
   ID  Duration          Date  Pulse  Maxpulse Calories
0   0        60  '2020/12/01'    110       130     4091
1   1        60  '2020/12/02'    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

Últimas 5 linhas do conjunto de dados:
    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
```

Substituindo valores nulos por “0” na coluna “calories”

```
# Substitua todos os valores nulos da coluna 'Calories' por 0
data_copy['Calories'].fillna(0, inplace=True)
print("\nConjunto de dados após substituir nulos em 'Calories' por 0:")
print(data_copy)
```

```
data_copy['Calories'].fillna(0, inplace=True)

Conjunto de dados após substituir nulos em '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
```

Substituindo valores nulos por ”1900/01/01” na coluna “Date”

```
# Substitua os valores nulos da coluna 'Date' por '1900/01/01'
data_copy['Date'].fillna('1900/01/01', inplace=True)
print("\nConjunto de dados após substituir nulos em 'Date' por '1900/01/01':")
print(data_copy)
```

```
Conjunto de dados após substituir nulos em 'Date' por '1900/01/01':
    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    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      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
```

Tratando a coluna “date” como “datetime” e corrigindo erros

```python
# Tente transformar a coluna 'Date' para datetime, e trate o erro se houver
try:
    data_copy['Date'] = pd.to_datetime(data_copy['Date'], format='%Y/%m/%d')
except ValueError as e:
    print("\nErro encontrado ao tentar converter 'Date' para datetime:", e)

# Substitua '1900/01/01' por NaN e tente novamente a conversão
data_copy['Date'].replace('1900/01/01', np.nan, inplace=True)
data_copy['Date'] = pd.to_datetime(data_copy['Date'], errors='coerce')

print("\nConjunto de dados após corrigir 'Date' e converter para datetime:")
print(data_copy)

# Corrija o valor específico '20201226' na coluna 'Date'
data_copy['Date'] = data_copy['Date'].replace('20201226', '2020/12/26')
data_copy['Date'] = pd.to_datetime(data_copy['Date'], errors='coerce')
```

```
Conjunto de dados após correções adicionais e conversão para datetime:
    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        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
```

Removendo linhas com valor nulo e apresentando resultado:

```python
# Remova registros com valores nulos (somente na coluna 'Date')
data_cleaned = data_copy.dropna(subset=['Date'])

# Imprima o dataframe final e confirme se todas as transformações foram realizadas
print("\nDataFrame final após todas as transformações:")
print(data_cleaned)
```

```
DataFrame final após todas as transformações:
    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
17  17        60 2020-12-17    100       120     3000
17  17        60 2020-12-17    100       120     3000
18  18        45 2020-12-18     90       112        0
17  17        60 2020-12-17    100       120     3000
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
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
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
```