Campus: POLO SAGUAÇU - JOINVILLE - SC

Curso: DESENVOLVIMENTO FULL STACK

Disciplina: Tecnologias Para Desenv. De Solucoes de Big Data

Turma: 9001

Período: 2025.2

Aluno: Jederson Borges de Oliveira

Link: https://github.com/JedersonBorges/missao-pratica-DGT2823.git

# DGT2823 - Tecnologias para desenv. de soluções de big data

**Objetivos 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);

Trabalho realizado no Google Colab:

https://colab.research.google.com/drive/19S3uOejOaHcKTQGcBuCPez44xx-4pSUN?usp=sharing

Código:

```python
# 1) Ler o CSV fornecido
tabela_exercicios = pd.read_csv("pico_web.csv", sep=";", engine="python", encoding="utf-8")

# 2) Verificar se os dados foram importados
print("Informações gerais do conjunto de dados:")
print(tabela_exercicios.info())
print("\nPrimeiras 5 linhas:")
print(tabela_exercicios.head())
print("\nÚltimas 5 linhas:")
print(tabela_exercicios.tail())

# 3) Criar uma cópia do dataset original
tabela_exercicios_limpa = tabela_exercicios.copy()

# 4) Substituir valores nulos em 'Calories' por 0
tabela_exercicios_limpa["Calories"] = tabela_exercicios_limpa["Calories"].fillna(0)
print("\nApós substituir nulos em 'Calories':")
print(tabela_exercicios_limpa.head(15))

# 5) Substituir valores nulos em 'Date' por '1900/01/01'
tabela_exercicios_limpa["Date"] = tabela_exercicios_limpa["Date"].fillna("1900/01/01")
print("\nApós substituir nulos em 'Date':")
print(tabela_exercicios_limpa.head(25))

# 6) Tentar converter 'Date' para datetime (vai dar erro por que '1900/01/01' não corresponde ao formato '%Y/%m/%d')
try:
    tabela_exercicios_limpa["Date"] = pd.to_datetime(tabela_exercicios_limpa["Date"], format="%Y/%m/%d")
except Exception as e:
    print("\nErro esperado na conversão de '1900/01/01':", e)

# 7) Substituir '1900/01/01' por NaN e tentar novamente
tabela_exercicios_limpa["Date"] = tabela_exercicios_limpa["Date"].replace("1900/01/01", np.nan)
tabela_exercicios_limpa["Date"] = pd.to_datetime(tabela_exercicios_limpa["Date"], format="%Y/%m/%d", errors="coerce")
print("\nApós corrigir '1900/01/01' para NaN e converter para datetime:")
print(tabela_exercicios_limpa.head(25))

# 8) Corrigir valor '20201226' para '2020/12/26'
tabela_exercicios_limpa["Date"] = tabela_exercicios_limpa["Date"].astype(str).str.replace("20201226", "2020/12/26")
tabela_exercicios_limpa["Date"] = pd.to_datetime(tabela_exercicios_limpa["Date"], format="%Y/%m/%d", errors="coerce")

# 9) Remover aspas simples
tabela_exercicios_limpa["Date"] = tabela_exercicios_limpa["Date"].astype(str).str.replace("'", "")
tabela_exercicios_limpa["Date"] = pd.to_datetime(tabela_exercicios_limpa["Date"], format="%Y/%m/%d", errors="coerce")

print("\nApós corrigir datas inválidas e converter novamente:")
print(tabela_exercicios_limpa.head(30))

# 10) Remover registros nulos restantes (linha 22 com 'Date' nulo)
tabela_exercicios_limpa = tabela_exercicios_limpa.dropna(subset=["Date"])
print("\nApós remover registros nulos (linha 22 eliminada):")
print(tabela_exercicios_limpa.info())
print(tabela_exercicios_limpa.tail())
```

Output:

```
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     float64
dtypes: float64(1), int64(4), object(1)
memory usage: 1.6+ KB
None

Primeiras 5 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

Últimas 5 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

Após substituir nulos em 'Calories':
    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
```

Após substituir nulos em '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 | 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 |

Erro esperado na conversão de '1900/01/01': time data "'2020/12/01'" doesn't match format "%Y/%m/%d", at position 0. You might want to try:
- passing `format` if your strings have a consistent format;
- passing `format='ISO8601'` if your strings are all ISO8601 but not necessarily in exactly the same format;
- passing `format='mixed'`, and the format will be inferred for each element individually. You might want to use `dayfirst` alongside this.

Após corrigir '1900/01/01' para NaN e converter para datetime:

| | ID | Duration | Date | Pulse | Maxpulse | Calories |
|---|---|---|---|---|---|---|
| 0 | 0 | 60 | NaT | 110 | 130 | 4091.0 |
| 1 | 1 | 60 | NaT | 117 | 145 | 4790.0 |
| 2 | 2 | 60 | NaT | 103 | 135 | 3400.0 |
| 3 | 3 | 45 | NaT | 109 | 175 | 2824.0 |
| 4 | 4 | 45 | NaT | 117 | 148 | 4060.0 |
| 5 | 5 | 60 | NaT | 102 | 127 | 3000.0 |
| 6 | 6 | 60 | NaT | 110 | 136 | 3740.0 |
| 7 | 7 | 450 | NaT | 104 | 134 | 2533.0 |
| 8 | 8 | 30 | NaT | 109 | 133 | 1951.0 |
| 9 | 9 | 60 | NaT | 98 | 124 | 2690.0 |
| 10 | 10 | 60 | NaT | 103 | 147 | 3293.0 |
| 11 | 11 | 60 | NaT | 100 | 120 | 2507.0 |
| 12 | 12 | 60 | NaT | 100 | 120 | 2507.0 |
| 13 | 13 | 60 | NaT | 106 | 128 | 3453.0 |
| 14 | 14 | 60 | NaT | 104 | 132 | 3793.0 |
| 15 | 15 | 60 | NaT | 98 | 123 | 2750.0 |
| 16 | 16 | 60 | NaT | 98 | 120 | 2152.0 |
| 17 | 17 | 60 | NaT | 100 | 120 | 3000.0 |
| 18 | 18 | 45 | NaT | 90 | 112 | 0.0 |
| 19 | 19 | 60 | NaT | 103 | 123 | 3230.0 |
| 20 | 20 | 45 | NaT | 97 | 125 | 2430.0 |
| 21 | 21 | 60 | NaT | 108 | 131 | 3642.0 |
| 22 | 22 | 45 | NaT | 100 | 119 | 2820.0 |
| 23 | 23 | 60 | NaT | 130 | 101 | 3000.0 |
| 24 | 24 | 45 | NaT | 105 | 132 | 2460.0 |

```
Após corrigir datas inválidas e converter novamente:
    ID  Duration Date  Pulse  Maxpulse  Calories
0    0        60  NaT    110       130    4091.0
1    1        60  NaT    117       145    4790.0
2    2        60  NaT    103       135    3400.0
3    3        45  NaT    109       175    2824.0
4    4        45  NaT    117       148    4060.0
5    5        60  NaT    102       127    3000.0
6    6        60  NaT    110       136    3740.0
7    7       450  NaT    104       134    2533.0
8    8        30  NaT    109       133    1951.0
9    9        60  NaT     98       124    2690.0
10  10        60  NaT    103       147    3293.0
11  11        60  NaT    100       120    2507.0
12  12        60  NaT    100       120    2507.0
13  13        60  NaT    106       128    3453.0
14  14        60  NaT    104       132    3793.0
15  15        60  NaT     98       123    2750.0
16  16        60  NaT     98       120    2152.0
17  17        60  NaT    100       120    3000.0
18  18        45  NaT     90       112       0.0
19  19        60  NaT    103       123    3230.0
20  20        45  NaT     97       125    2430.0
21  21        60  NaT    108       131    3642.0
22  22        45  NaT    100       119    2820.0
23  23        60  NaT    130       101    3000.0
24  24        45  NaT    105       132    2460.0
25  25        60  NaT    102       126    3345.0
26  26        60  NaT    100       120    2500.0
27  27        60  NaT     92       118    2410.0
28  28        60  NaT    103       132       0.0
29  29        60  NaT    100       132    2800.0

Após remover registros nulos (linha 22 eliminada):
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        0 non-null      int64
 1   Duration  0 non-null      int64
 2   Date      0 non-null      datetime64[ns]
 3   Pulse     0 non-null      int64
 4   Maxpulse  0 non-null      int64
 5   Calories  0 non-null      float64
dtypes: datetime64[ns](1), float64(1), int64(4)
memory usage: 0.0 bytes
None
Empty DataFrame
Columns: [ID, Duration, Date, Pulse, Maxpulse, Calories]
Index: []
```