CURSO - Desenvolvimento full Stack
DISCIPLINA – Tratando a imensidão dos dados
ALUNO (A) – Sfênia Mesquita da Silva Inacio
MATRÍCULA – 202208042341
CAMPUS - Polo Cohatrac/São Luis – MA

Contextualização

Como Analista de Dados, você recebeu, em um novo projeto, um conjunto de dados. Sua principal tarefa é tratar os dados desse 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.

Procedimentos

Após ler um conjunto de dados, fornecidos no enunciado da entrega, compostos pelas conlunas ID;Duration;Date;Pulse;Maxpulse;Calories devemos desenvolver a seguinte programação:

```
#Importing pandas library
import pandas as pand

# Fonte de dados ==> formato CSV
csv_dados = 'picoweb_dados.csv'

# pegando os dados no arquivo CSV
df = pand.read_csv(csv_dados, sep=';', engine='python', encoding='utf-8')

# Imprimindo informações gerais do dataframe
print("Imprimindo informações gerais do DataFrame:")
print(df.info())
print("========="")
```

```
picoweb.ipynb
                 picoweb.ipynb (output) ×
      Imprimindo informações gerais do DataFrame:
      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 32 entries, 0 to 31
      Data columns (total 6 columns):
          Column Non-Null Count Dtype
                                   int64
       0 ID
                   32 non-null
       1 Duration 32 non-null
                                  int64
       2 Date
                   31 non-null
                                 object
       3 Pulse 32 non-null
                                   int64
           Maxpulse 32 non-null
       4
  11
                                   int64
          Calories 30 non-null
                                   float64
```

```
# Imprimindo as 5 primeiras linhas
print("\nPrint das 5 primeiras Linhas:")
print(df.head())
print("==========="")

# Imprimindo as 5 últimas linhas
print("\nPrint das 5 últimas Linhas:")
print(df.tail())
print("========"")
```

```
Print das 5 primeiras Linhas:
       ID Duration
                           Date Pulse Maxpulse Calories
        0
                60
                    '2020/12/01'
                                 110
                                          130
                                                 4091.0
    0
21
    1
       1
                60
                   '2020/12/02'
                                 117
                                          145
                                                 4790.0
                   '2020/12/03'
       2
22
    2
                                 103
                                          135
                60
                                                 3400.0
       3
                   '2020/12/04'
    3
                45
                                 109
                                          175
                                                 2824.0
                    '2020/12/05'
    4
        4
                45
                                 117
                                          148
                                                 4060.0
25
    ______
    Print das 5 Últimas Linhas:
        ID Duration
                           Date
                                 Pulse
                                       Maxpulse Calories
    27 27
                     '2020/12/27'
                                   92
                                           118
                                                  2410.0
                 60
    28 28
                    '2020/12/28'
                                  103
                 60
                                           132
                                                    NaN
    29 29
                 60
                     '2020/12/29'
                                  100
                                           132
                                                  2800.0
    30 30
                    '2020/12/30'
                 60
                                  102
                                           129
                                                  3803.0
    31 31
                     '2020/12/31'
                                   92
                                           115
                 60
                                                  2430.0
```

```
# Criando uma cópia de segurança dos dados
df_copy = df.copy()
```

```
# Substituindo os valores nulos na coluna "Calories" por 0

df_copy['Calories'].fillna(0, inplace=True)
print("\nPrint do DataFrame após substituição dos valores nulos na coluna 'Calories':")
print(df_copy)
print("========""")
```

Pri	nt d	o DataFram	e após substit	uição d	os valores	nulos na	coluna '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	
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	
===	====	=======	========	======	========	=======	

```
# Substituindo os valores nulos na coluna "Date" por "1900/01/01"

df_copy['Date'].fillna('1900/01/01', inplace=True)

print("\nPrint do DataFrame após substituição dos valores nulos na coluna 'Date':")

print(df_copy)

print("========="")
```

72	Pri	nt d	o DataFram	e após substit	uição d	os valores	nulos na	coluna	'Date':
73		ID	Duration	Date	Pulse	Maxpulse	Calories		
74	0	0	60	'2020/12/01'	110	130	4091.0		
75	1	1	60	'2020/12/02'	117	145	4790.0		
76	2	2	60	'2020/12/03'	103	135	3400.0		
77	3	3	45	'2020/12/04'	109	175	2824.0		
78	4	4	45	'2020/12/05'	117	148	4060.0		
79	5	5	60	'2020/12/06'	102	127	3000.0		
80	6	6	60	'2020/12/07'	110	136	3740.0		
81	7	7	450	'2020/12/08'	104	134	2533.0		
82	8	8	30	'2020/12/09'	109	133	1951.0		
83	9	9	60	'2020/12/10'	98	124	2690.0		
84	10	10	60	'2020/12/11'	103	147	3293.0		
85	11	11	60	'2020/12/12'	100	120	2507.0		
86	12	12	60	'2020/12/12'	100	120	2507.0		
87	13	13	60	'2020/12/13'	106	128	3453.0		
88	14	14	60	'2020/12/14'	104	132	3793.0		
89	15	15	60	'2020/12/15'	98	123	2750.0		
90	16	16	60	'2020/12/16'	98	120	2152.0		
91	17	17	60	'2020/12/17'	100	120	3000.0		
92	18	18	45	'2020/12/18'	90	112	0.0		
93	19	19	60	'2020/12/19'	103	123	3230.0		
94	20	20	45	'2020/12/20'	97	125	2430.0		
95	21	21	60	'2020/12/21'	108	131	3642.0		
96	22	22	45	1900/01/01	100	119	2820.0		
97	23	23	60	'2020/12/23'	130	101	3000.0		
98	24	24	45	'2020/12/24'	105	132	2460.0		
99	25	25	60	'2020/12/25'	102	126	3345.0		
100	26	26	60	20201226	100	120	2500.0		
101	27	27	60	'2020/12/27'	92	118	2410.0		
102	28	28	60	'2020/12/28'	103	132	0.0		
103	29	29	60	'2020/12/29'	100	132	2800.0		
104	30	30	60	'2020/12/30'	102	129	3803.0		
105	31	31	60	'2020/12/31'	92	115	2430.0		
106	===	====	=======	===========	======	=======	=======		

```
# Corrigindo o formato das datas
df_copy['Date'] = df_copy['Date'].str.strip("'")
df_copy['Date'] = df_copy['Date'].astype(str).replace({'20201226': '2020/12/26'})
df_copy['Date'] = pand.to_datetime(df_copy['Date'], format='%Y/%m/%d', errors='coerce')
print("\nPrint doDataFrame após correção para o formato data id 26 '20201226':")
print(df_copy)
print("=========="")
```

Pri	nt d	oDataFrame	e após corre	ção pa	ra o forma	to data id	26 '20201226':
	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	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	
===	====	=======	=========	======	=======	========	=

```
# Transformando a coluna 'Date' para datetime
df_copy['Date'] = pand.to_datetime(df_copy['Date'], format='%Y/%m/%d', errors='coerce')
print("\nPrint do DataFrame após transformação da coluna 'Date' em datetime:")
print(df_copy)
print("========"")
```

Pri	nt d	o DataFran	ne após tran	ısformaç	ão da colu	na 'Date'	em datetime:
	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	2 3	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	
===	====	========	========	======	=======	========	===

```
# Mudando, na coluna Date, o valor '1900/01/01' para 'NaN' - Not a Number

df_copy['Date'].replace(pand.Timestamp('1900-01-01'), pand.NaT, inplace=True)
print("\nPrint do DataFrame após alteração, na coluna Date, do valor '1900/01/01' para 'NaN' - Not a Number")
print(df_copy)
print("========"")
```

Pri	int d	o DataFran	ne após alte	ração,	na coluna	Date, do valor	'1900/01/01'	para	'NaN' -	Not a	Number
	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					
===					=========						

```
# Excluindo os registros com valores nulos na coluna 'Date'
df_clean = df_copy.dropna(subset=['Date'])
print("\nPrint do DataFrame após remoção dos registros com valores nulos na coluna 'Date':")
print(df_clean)
print("========="")
```

ъ.			•	~ .		1	1		,	10 1 1	
Pri						com valores	nulos	na	coluna	Date:	
	ID	Duration			Maxpulse						
0	0		2020-12-01	110	130	4091.0					
1	1		2020-12-02	117	145	4790.0					
2 3	2		2020-12-03	103	135	3400.0					
4	3 4		2020-12-04	109	175	2824.0					
5			2020-12-05 2020-12-06	117	148	4060.0					
6	5 6		2020-12-00	102 110	127 136	3000.0 3740.0					
7	7		2020-12-07	104	134	2533.0					
8	8		2020-12-08	104	133	1951.0					
9	9		2020-12-03	98	124	2690.0					
10	10		2020-12-10	103	147	3293.0					
11	11		2020-12-12	100	120	2507.0					
12	12		2020-12-12	100	120	2507.0					
13	13		2020-12-13	106	128	3453.0					
14	14		2020-12-14	104	132	3793.0					
15	15		2020-12-15	98	123	2750.0					
16	16		2020-12-16	98	120	2152.0					
17	17		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		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					
===	====	=======	=======	======	=======	========					