**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>

**Contextualização**

Para resolução das micro atividades será necessário ter em mãos um conjunto de dados no formato CSV. Tais dados podem ser obtidos a partir de fontes gratuitas, disponíveis na Web (como, por exemplo, o dataset disponível em <https://archive.ics.uci.edu/dataset/352/online+retail>), assim como um ambiente contendo o interpretador da linguagem python.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto. Ainda em relação aos conjuntos de dados, a atividade “pico web” terá como base o seguinte conjunto de dados (que deverá ser copiado e salvo num arquivo “csv”, usando como separados de colunas o“;”):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 | 450 | '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 | '2020/12/12' | 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 | NaN |
| 19 | 60 | '2020/12/19' | 103 | 123 | 3230 |
| 20 | 45 | '2020/12/20'; | 97 | 125 | 2430 |
| 21 | 60 | '2020/12/21' | 108 | 131 | 3642 |
| 22 | 45 | NaN | 100 | 119 | 2820 |
| 23 | 60 | '2020/12/23' | 130 | 101 | 3000 |
| 24 | 45 | '2020/12/24' | 105 | 132 | 2460 |
| 25 | 60 | '2020/12/25' | 102 | 126 | 3345 |
| 26 | 60 | ‘2020/12/26’ | 100 | 120 | 2500 |
| 27 | 60 | '2020/12/27' | 92 | 118 | 2410 |
| 28 | 60 | '2020/12/28' | 103 | 132 | NaN |
| 29 | 60 | '2020/12/29' | 100 | 132 | 2800 |
| 30 | 60 | '2020/12/30' | 102 | 129 | 3803 |
| 31 | 60 | '2020/12/31' | 92 | 115 | 2430 |

O uso do dataframe acima é imprescindível, uma vez que ele contém dados não válidos que deverão ser tratados posteriormente. Vide as linhas 18 e 28 (coluna Calories); 22 e 26 (coluna Date).

**Micro atividade 1: Descrever como ler um arquivo CSV usando a biblioteca Pandas (Python)**

**Criação do Script de leitura local**

**Tela preta com letras brancas

O conteúdo gerado por IA pode estar incorreto.**

* **Importar a biblioteca pandas, necessária para a manipulação de dados;**
* **Definir o arquivo CSV que conterá os dados informados e a sua localização;**
* **Ler o arquivo CSV utilizando os seguintes parâmetros:**
  + **sep: define o separador de colunas, que é o ponto e vírgula;**
  + **engine: define a linguagem utilizada, que é o Python;**
  + **encoding: define a codificação do arquivo, como ‘uft-8’, que é a mais comum.**
* **Exibe o conteúdo lido do arquivo CSV, no terminal.**

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

**Criação do Script de leitura remoto**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.**

* Necessário efetuar a instalação do pacote: ucimlrepo;
* Criar um novo arquivo: csv\_externo.py;
* O código do exemplo foi fornecido pelo link, fornecido no sway da missao;
  + <https://archive.ics.uci.edu/dataset/352/online+retail>

No arquivo criado:

* Importar o pacote instalado para localizar os dados desejados;
* Selecionar o id dos dados a serem exibidos;
* Separar os dados de forma ampla e de forma específica;
* Exibir ambas as formas no terminal.

Em cima como foi exibido no VsCode, abaixo como consta no site:

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

**Micro atividade 2: Descrever como criar um subconjunto de dados a partir de um conjunto existente usando a biblioteca Pandas (Python)**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.**

* **Utilizando-se o mesmo arquivo da micro atividade anterior;**
* **Foi criado um subconjunto contendo apenas as colunas 'ID', 'Date' e 'Calories';**
* **Foi exibido esse subconjunto.**

**Micro atividade 3: Descrever como configurar o número máximo de linhas a serem exibidas na visualização de um conjunto de dados usando a biblioteca Pandas (Python)**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Tabela

O conteúdo gerado por IA pode estar incorreto.**

* **Utilizando-se o mesmo arquivo da micro atividade 01;**
* **Foi solicitado o uso de uma propriedade de exibição do Pandas:**
  + **display.max\_rows definindo seu valor para 9999;**
  + **Porém, para exibir os valores de ‘dados’, foi necessário utilizar o método to\_string() para converter o DataFrame em uma string completa.**

**OBS: Foram exibidos apenas 31 linhas devido a quantidade de linhas que o arquivo ‘dados’ possui, porém o código foi ajustado para exibir até 9999 linhas.**

**Micro atividade 4: Descrever como exibir as primeiras e últimas “N” linhas de um conjunto de dados usando a biblioteca Pandas (Python)**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.**

* **Utilizando-se o mesmo arquivo da micro atividade 01;**
  + **Foi solicitado exibir as 10 primeiras e as 10 últimas linhas do arquivo DataFrame: ‘dados’;**

**Micro atividade 5: Descrever como exibir informações gerais sobre as colunas, linhas e dados de um conjunto de dados usando a biblioteca Pandas (Python)**

**Utilizando-se o mesmo arquivo da micro atividade 01, foi solicitado:**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Exibir informações gerais sobre o conjunto de dados: suas colunas, linhas e dados**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Total de Linhas e Colunas**

**Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.**

**Verificar a quantidade de dados nulos, caso existam**

**Uma imagem contendo Texto

O conteúdo gerado por IA pode estar incorreto.**

**O tipo de dado de cada coluna**

**Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.**

**A quantidade de memória utilizada pelo conjunto de dados**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**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.

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Texto

O conteúdo gerado por IA pode estar incorreto.**

1. Atribua os dados lidos a uma variável;
2. 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.
3. Crie uma nova variável e atribua a ela uma cópia do conjunto de dados original (variável criada no passo 4);

**Texto

O conteúdo gerado por IA pode estar incorreto.**

**Tela preta com letras brancas

O conteúdo gerado por IA pode estar incorreto. Tela preta com letras brancas

O conteúdo gerado por IA pode estar incorreto.**

1. 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;

**Calendário

O conteúdo gerado por IA pode estar incorreto.**

1. 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’;

**Calendário

O conteúdo gerado por IA pode estar incorreto.**

1. 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;

Uma imagem contendo Calendário

O conteúdo gerado por IA pode estar incorreto.

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

1. 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’;
2. 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;

**Uma imagem contendo Tabela

O conteúdo gerado por IA pode estar incorreto.**

1. 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;
2. Imprima o dataframe e verifique se todas as transformações foram executadas conforme solicitado nos passos anteriores.

**Uma imagem contendo Calendário

O conteúdo gerado por IA pode estar incorreto.**