Relatório UFO Data – Parte 4

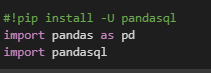
Levi Alves de Freitas Junior

**Problema**

Realizar uma limpeza no DataFrame OVNI, retirar campos em Branco, Unknown e None, realizar uma filtragem com pandasql para explorar e eliminar os dados. Vamos utilizar a base ovni\_data gerada no primeiro relatório.

Etapa inicial criar um novo notebook no colab research. Siga as instruções a seguir e veja as imagens ilustrativas:

**1** - Após criar um novo notebook, podemos começar a criar nosso código e inicialmente precisamos importar as bibliotecas pandas e para usar o banco de dados vamos importar o pandasql.



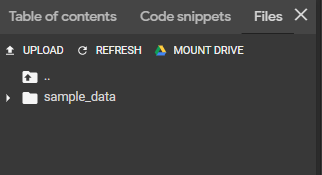
Primeiro selecione o comando - !pip install -U pandasql e aperte CTRL+ SHIFT + ENTER para instalar o pandasql no colab, logo após inicialize a célula no colab para importar o pandas e o pandasql,

Obs. O import do pandasql pode não funcionar antes de utilizar o comando pip para utilizá-lo.

**2** – Agora vamos importar a base de dados

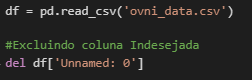
Lembre-se de importar a base de dados no colab para podermos chama-la no código, siga as instruções abaixo para colocar a base.

No menu esquerdo selecione Files, e clique em upload, selecione a base, onde você salvou



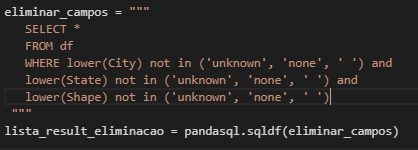
Aguarde um pouco e sua base estará inserida no colab.

Podemos prosseguir agora e chamar nossa base de dados.



O comando ‘del’ retiramos a coluna não nomeada pois ela não terá nenhuma utilidade para o projeto.

**3** - Remover registros que tenham valores vazios (None, Unknown, ...) para City, State e Shape.



Utilizamos o pandasql para realizar a filtragem e remoção dos dados, selecionamos todos os dados do dataframe DF e na cláusula WHERE colocamos a restrição dos itens a não serem retornados após a pesquisa.

Como resultado recebemos o dataframe a seguir:

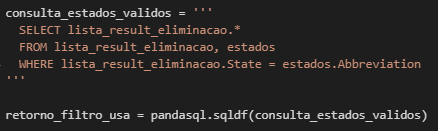


**4 –** Manter os registros referentes aos 51 estados dos Estados Unidos.

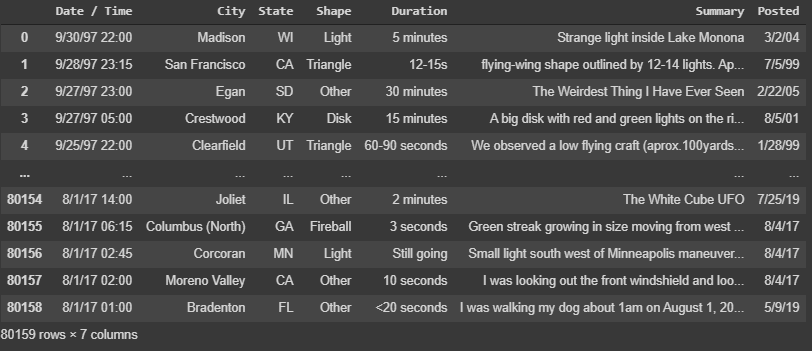
Para esta etapa baixe o arquivo no GitHub chamado states.csv e logo após importe-o no seu Files do colab.



Agora comparamos os estados do arquivo states.csv com os estados de lista\_result\_eliminacao



Este bloco de pandasql retornará um dataframe mais filtrado, vamos observar



Observe se seu resultado seja igual à quantidade de linhas que ele retorna que no exemplo acima, no caso, 80159 linhas.

**5 –** Remover variáveis irrelevantes para análise (Duration, Summary e Posted).

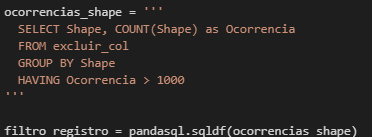


Utilizamos o drop para retirar as colunas Duration, Summary e Posted. O drop é uma biblioteca utilizada no pandas para eliminar as colunas do dataframe, e utilizamos o axis para referenciar o eixo do dataframe, por exemplo (horizontal = 1 ou vertical = 0).

Selecione excluir\_col e aperte CTRL + SHIFT + ENTER e seu resultado será o seguinte:

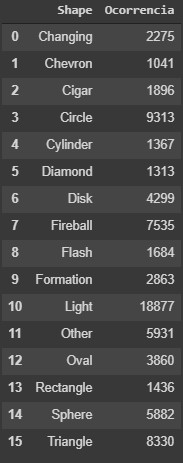


**6** – Manter os registros de Shapes mais populares (com mais de 1000 ocorrências).

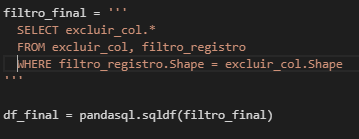


Aqui selecionamos o Shape e o COUNT (Shape) do dataframe excluir\_col, agrupamos o Shape que tenham o número de ocorrências maior que 1000.

Recebemos como resultado uma pesquisa mais filtrada:



**7 –** Retornar o dataframe criado com os campos Date/ Time, City, State e Shape comparando com a coluna Shape criada anteriormente.



Agora no WHERE comparamos as colunas Shape de filtro\_registro que foi criada no tópico anterior e a coluna Shape de excluir\_col, o exemplo que retiramos 3 campos com o DROP.

E nosso dataframe final terá o seguinte resultado:



8 – Crie um novo DataFrame



**Referências**

Pandas sql - <https://pypi.org/project/pandasql/>

Pandas - <https://pandas.pydata.org/>