**Atividade de Ciências de Dados**

**Aluno:** Heberth Vinícius Amarante Severo

**Descrição do problema:**

Essa é a quarta parte da nossa atividade de Web Scraping, nela iremos modificar o arquivo csv original e gerar um novo arquivo csv com dados mais concisos. Iremos excluir algumas colunas, tirar os campos vazios, dentre outros. Todas as etapas dessa atividade serão descritas e explicadas a seguir.

**1ª Etapa**

Para o início do arquivo no Colab é necessário fazer a importação de algumas funções do python, a Figura 1 mostra todas as funções utilizadas para a nossa atividade.

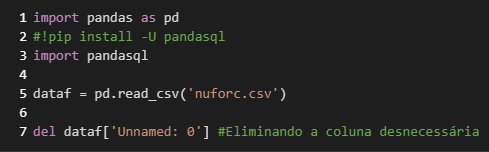


Figura 1 Importação das funções

A primeira etapa é bem simples, nela há apenas as importações que estão presentes até a terceira linha, a leitura do arquivo csv na linha de número cinco, esse arquivo csv foi obtido na primeira parte do projeto. Na linha sete foi necessário deletar uma coluna do nosso csv, pois ela foi gerada automaticamente e não faz parte dos dados coletados.

**2ª Etapa**

Nessa etapa da atividade é necessário deletar algumas linhas do nosso arquivo csv. As linhas que iremos deletar são aquelas que estão vazias, com o valor NONE ou UNKNOWN, a Figura 2 mostra como foi realizado a exclusão dessas linhas.

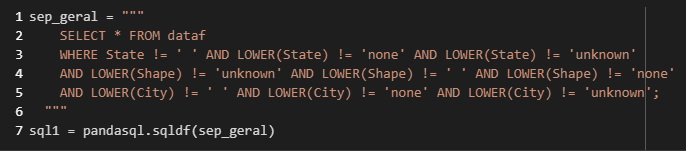


Figura 2 Removendo os campos vazios de Shape, State e City

**3ª Etapa**

Nessa etapa iremos filtrar nosso data frame, para realizar esse processo é necessário utilizar um arquivo csv, chamado de *estados\_usa*, que foi disponibilizado pelo professor, que contém o nome dos estados e a abreviação deles, e comparar esse arquivo com o data frame gerado na Figura 2.

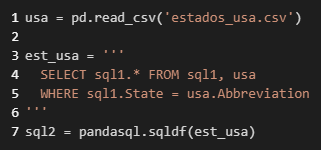


Figura 3 Deixando apenas os dados dos Estados Unidos

A Figura 4 mostra o resultado desse bloco de código, nele contém apenas os dados que são dos Estados Unidos.

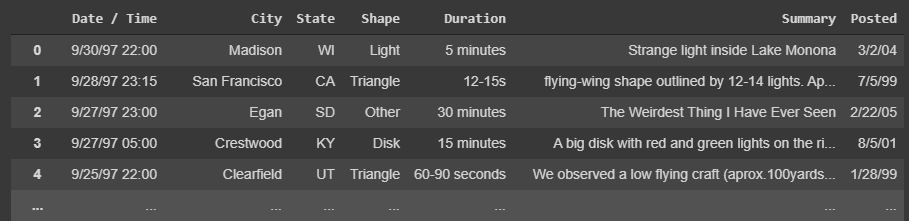


Figura 4 Resultado gerado pelo código da Figura 3

**4ª Etapa**

Essa etapa é a exclusão de algumas colunas do data frame, essas colunas não são essenciais para o entendimento do arquivo. As colunas que iremos deletar são Duration, Summary e Posted, como mostra a Figura 5.



Figura 5 Deletando as colunas Duration, Summary e Posted

Após a execução do bloco de código anterior, o nosso data frame terá uma quantidade de colunas inferior ao que tinha antes. A Figura 6 mostra o arquivo atualizado.



Figura 6 Data frame após a exclusão das colunas

**5ª Etapa**

Nessa etapa iremos fazer um filtro e calcular os Shapes do data frame, iremos selecionar apenas os Shapes que possuem mais de mil aparições, a Figura 7 mostra o trecho de código para fazer essa etapa da atividade.

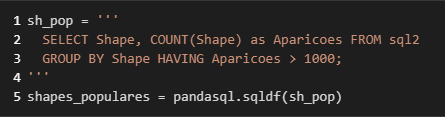
****

Figura 7 Filtrando os Shapes

**6ª Etapa**

Essa é a última etapa dessa atividade, nela vamos juntar o nossa data frame com o data frame dos Shapes gerado após a execução da Figura 7. Após juntar os data frames iremos gerar um arquivo csv após a limpeza dos dados, a Figura 8 mostra o último trecho de código dessa atividade.

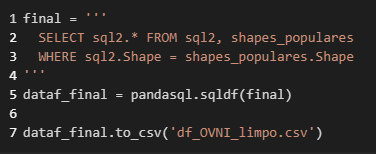


Figura 8 Criação do data frame final e criação do csv

O trecho que vai da primeira linha até a quinta linha, é para juntar os data frames. A linha sete é onde colocamos o nome desejado para o csv, para enfim criarmos o nosso arquivo.

**Observação**

Este relatório estará no GitHub, cujo link está a seguir: <https://github.com/heberth-hvas/icd_5>

**Referências**

**Google Colab:** <https://colab.research.google.com/notebooks/welcome.ipynb>

**Documentação Pandas:** <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.append.html>

**Drop do Pandas:** <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html>

**Delete column from Pandas DataFrame:** <https://stackoverflow.com/questions/13411544/delete-column-from-pandas-dataframe>