**Relatório de exploração dos dados coletados do site *nuforc.org***

Aluno: Daniel Barros Candeira

1. **Introdução**

O relatório tem por objetivo demonstrar as etapas percorridas para explorar os dados coletados do site *nuforc.org* entre 07 de setembro de 1997 e 07 de agosto de 2017, site que contém relatos de “visualização” de OVNIs – Objetos Voadores Não Identificados, a fim de identificar informações relevantes.

1. **Problema**

Após a realização da coleta, nessa etapa do projeto são realizados alguns passos a fim de melhor compreender os dados coletados. Sendo o 1º passo identificar a quantidade de linhas, observações ou variáveis que foram coletadas, 2º identificar quantos relatos ocorreram por estado em ordem decrescente, 3º remover possíveis campos vazios (sem estado), 4º limitar a análise aos estados dos Estados Unidos, 5º consultar por cidades, com o objetivo de saber quais contêm o maior número de relatos (cidades que apresentem ao menos 10 relatos), 6º com os dados obtidos no passo anterior, responder a seguinte pergunta : por que será que essa é a cidade que possui mais relatos? e 7º realizar uma *query* exclusiva para o estado com maior número de relatos, buscando cidades que possuam um número superior a 10 relatórios. Enfatizando a cidade, a quantidade de relatos e formato do objeto não identificado.

1. **Resolução**

A fim de realizar os passos identificados na seção 2, foi desenvolvido um script utilizando a linguagem de programação **Python**, e a bibliotecas **Requests, BeautifulSoup** e **Pandas**, que servem respectivamente para realizar requisições HTTP[[1]](#footnote-1) que retornam o código HTML que compõe a página, possibilitar a navegação dentro das tags HTML, e para lidar com dados de forma estatística e com base/arquivos de dados volumosos.

**1º Passo:**

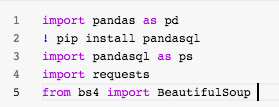


Fig1. Importando bibliotecas.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig2. Importando a base de dados.

Antes da escrita dos códigos, é realizada a importação das bibliotecas na **Fig1** e a importação a base de dados (arquivo CSV) que está hospedado no Github com a utilização da biblioteca **pandas** na **Fig2**, como resultado da importação da base é retornado um dataframe que contém todas a linhas e colunas salvas no arquivo CSV, no final da linha de importação é utilizada a função ***drop*** a fim de remover uma coluna irrelevante do dataframe.

Uma imagem contendo texto

Descrição gerada automaticamente

Fig3. Verificando as variáveis presentes no dataframe.

Após a importação dos dados, utilizamos a função info() na **Fig3,** que retorna todas as variáveis presentes no dataframe, assim como a quantidade de dados não nulos entre outras informações.

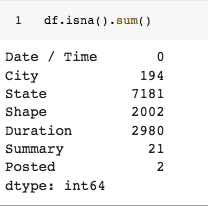


Fig4. Verificando quantidade de campos vazios.

A função isna().sum() **Fig4**, retorna a soma de todos os campos vazios de cada uma das colunas presentes no dataframe.

**2º e 3º Passos:**

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig5. Selecionando estados por número de relatos, em ordem decrescente.

O próximo passo realizado foi a identificação de quantos relatos ocorreram por estado em ordem decrescente, para isso, o estudante optou por utilizar uma consulta SQL[[2]](#footnote-2), que conta a quantidade de relatos de cada estado, agrupa os dados por estado e ordena-os em ordem decrescente. Como retorno se tem os estados e o número de relatos de cada estado.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig6. Selecionando estados por número de relatos, em ordem decrescente, sem campos de estado em branco.

O código mostrado na **Fig6** realiza a praticamente a mesma função do código mostrado na **Fig5** com a diferença que no segundo código são excluídos da consulta as linhas da tabela que não possuem valor na coluna *State*, que contém a sigla do estado.

**4º Passo:**

Contudo, os resultados retornados no passo anterior contêm dados de relatos de diferentes países. A fim diminuir a quantidade de dados a ser analisada, optou-se por explorar apenas relatos de estados estadunidenses.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig7. Coletando estados dos EUA.

O estudante optou por utilizar uma tabela disponível no site Wikipédia ao invés de fazer a entrada dos estados manualmente. Na **Fig7** é realizado uma solicitação ao site a fim de coletar o HTML, e utilizando a biblioteca **BeautifulSoup** para coletar a tabela de estados presentes no site, por fim os dados foram passado para a função read\_html da biblioteca **pandas** convertendo os dados em um dataframe, tendo como retorno uma tabela com diversos dados dos estados estadunidenses.



Fig8. Renomeando as colunas selecionadas do dataframe.

Após isso foram selecionadas apenas as colunas que continham o nome dos estados e sua respectiva abreviação, que foram renomeadas.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig9. Coletando dados dos estados estadunidenses.

Com as siglas dos estados armazenados em um dataframe, realizamos uma consulta SQL com o objetivo de retornar apenas os dados dos estados dos EUA.

**5º Passo:**

Uma imagem contendo captura de tela

Descrição gerada automaticamente

­­Fig10. Verificando relatos por cidade.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig11. Selecionando cidades com ao menos 10 relatos.

Para retornar as cidades que possuem ao menos 10 relatos, foram realizados duas consultas SQL, sendo o retorno da figura que se encontra na **Fig10** um dataframe com a quantidade de relatos separados por cidade e a que se encontra na **Fig11** retorna apenas as que possuem ao menos 10 relatos e as ordena de forma decrescente, sem contar as cidades não identificadas ou deixadas em branco.

**6º Passo:**

Com base nos dados coletados anteriormente, por que será que essa é a cidade que possui mais relatos?

**R:** Na cidade de Phoenix em março de 1997, ocorreu uma série de relatos de aparição de OVNIs, milhares de pessoas reportaram diferentes luzes, com formações triangulares, que também foram reportados em 2007 e 2008. Observando-se os dados pode-se identificar que o formato dos objetos relatados foram flashes, que podem estar diretamente ligados ao que ocorreu nos anos de 1997, 2007 e 2008.

**7º Passo:**

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig12. Selecionando estado com maior número de relatos.

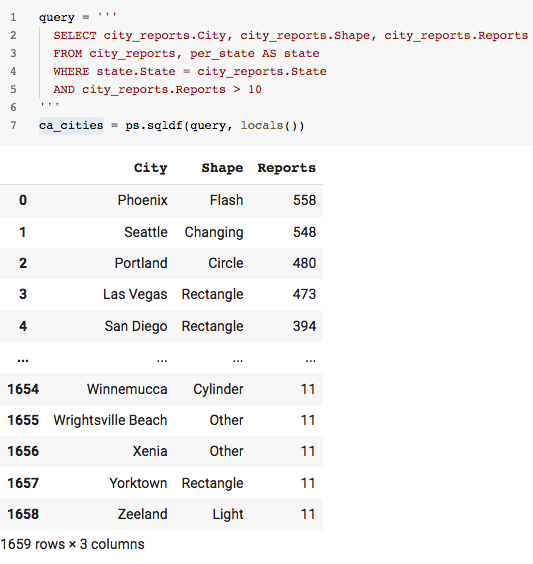


Fig13. Selecionando a cidade com mais de 10 relatos no estado da Califórnia.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig14. Selecionando a cidade com mais de 10 relatos no estado da Califórnia, sem formatos do objeto (OVNI) desconhecido ou em branco.

Por fim nós realizamos duas consultas demonstradas em **Fig12** e **Fig13** que realizam respectivamente uma busca do estado com maior número de relatos e o último seleciona todos as cidades que pertence a esse estado e possuem mais de 10 relatos. O código na **Fig14** apenas se diferencia do apresentado na **Fig13** por remover dos resultados da busca os relatos que possuem objetos (OVNIs) com formato não identificado ou vazios.

**Github**: <https://github.com/dan-candeira/Nuforc-Project>

**Bibliografia**

<https://en.wikipedia.org/wiki/Phoenix_Lights>

1. Protocolo de Transferência de Hipertexto. [↑](#footnote-ref-1)
2. Structured Query Language – Linguagem de Consulta Estruturada. [↑](#footnote-ref-2)