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

Aluno: Daniel Barros Candeira

1. **Introdução**

O relatório tem por objetivo demonstrar as etapas percorridas para visualizar 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**

Nessa etapa será realizada a limpeza dos dados, tendo como atividades, remover registros que tenham valores vazios (None, Unknown, ...) para City, State e Shape, manter somente os registros referentes aos 51 estados dos Estado Unidos, remover as variáveis irrelevantes para a análise (Duration, Summary e Posted), Manter somente os registros de Shapes mais populares (com mais de 1000 ocorrências), e por fim, salvar o dataframe final em um arquivo CSV com o nome "df\_OVNI\_limpo".

1. **Resolução**

A fim de realizar os passos identificados na seção 2, foi desenvolvido um script utilizando a linguagem de programação **Python**, e as bibliotecas **Pandasql** e **Pandas**,que servem respectivamente para lidar com dados de forma estatística e com base/arquivos de dados volumosos, e realizar consultas SQL em data frames.

**1º Passo:**

Antes de iniciar a escrita do código, é realizada a importação das bibliotecas a serem utilizadas, como é demonstrado na **Fig 1**.

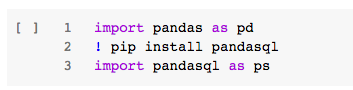


Fig 1. Importando as bibliotecas.

Após isso, importamos o dataframe a ser limpo, **Fig 2**. Após a importação do dataframe, realizamos uma consulta SQL para remover os campos que contenham valores em vazios, ou desconhecidos **Fig 3**.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig 2. Importando o dataframe.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig 3. Realizando a consulta SQL para remover campos em branco, desconhecidos etc.

Uma imagem contendo texto

Descrição gerada automaticamente

Fig 4. Informações do daframe resultante da consulta SQL.

**2º Passo:**

Após realizar a limpeza dos campos, importamos o dataframe com os estados dos Estados Unidos, disponibilizado pelo professor da disciplina na **Fig 5**, para realizar uma consulta SQL que retorna apenas os dados relativos ao EUA, **Fig 6.**

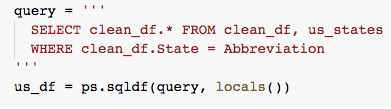


Fig 6. Consulta SQL que retorna os dados dos estados Estados Unidos.

Uma imagem contendo texto, interior

Descrição gerada automaticamente

Fig 8. Informações do dataframe após a consulta SQL da Fig 7.

**3º Passo:**

O terceiro passo é remover as colunas não relevantes do dataframe, **Fig 9.**



Fig 9. Removendo colunas não relevantes.

**4º Passo:**

Após removermos as colunas não relevantes do dataframe, realizamos uma nova consulta SQL, a fim de coletar dados apenas dos formatos que possuem mais de 1000 relatos, **Fig 10**, o retorno da consulta SQL é então utilizado em uma segunda consulta que vai selecionar os dados apenas dos formatos que presentes no dataframe retornado na **Fig 10.**

Uma imagem contendo captura de tela

Descrição gerada automaticamente

Fig 10. Consulta SQL dos formatos mais populares.

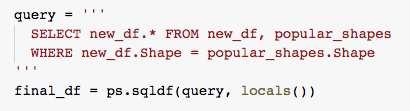


Fig 11. Segunda consulta SQL, selecionando apenas os dados que bateram com os retornados na Fig 10.

**5º Passo:**

Por fim criamos um arquivo CSV que contém os dados retornados da consulta SQL realizada na **Fig 11**.



Fig 12. Criando arquivo CSV.

**Github**:<https://github.com/dan-candeira/Nuforc-Project/tree/master/5.7-Preparacao>