Instituto Federal de Brasília

Professor: Fábio Henrique Monteiro Oliveira

Nome: Vanessa de Andrade Formiga

Brasília 31 de Outubro de 2019

**Relatório da Exploração de Dados**

**Problema:**

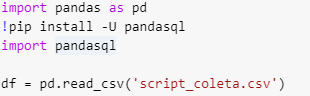
Nesta etapa, a questão que deveria ter ser resolvida era que precisa ser feita uma seleção dos dados encontrados em cada um dos campos contidos na base de dados, e partir da integração entre as várias tabelas, foram feitas consultas específicas para terem como resultado a exploração dos dados, e também por meio das consultas específicas podem ser retiradas informações mais detalhados dos dados encontrados no arquivo geram isso, pode ter sido solucionado umas das questões a serem esclarecidas a partir, das consultas realizadas, que era qual o motiva de que determinada cidade tinha mais registros de relatos publicados, caso não tivesse essa consulta não teria tido informações necessárias para responderem esse questionamento.

**Por que será que essa é a cidade que possui mais relatos?**

Porque a cidade de Phoenix possui maior quantidade de relatos registrados no portal, o motivo e porque possui registros na história de que foram visualizados luzes encontradas na cidade durante o período de vinte anos, com isso selecionaram para serem feitas a coleta dos dados. E esses acontecimentos ocorrem durante os anos de 1997, e esse fato repetiu-se durante os anos de 2007, 2008 e 2009 em seguidas repetições. Isso pode ser uma das motivações encontrados para que a cidade de Phoenix no estado do Arizona ter uma maior quantidade de relatos.

**Solução encontrada para Resolver o problema:**

Utilizou comandos sql, que por meio dele, realizou-se consultas de sql. Para isso o trabalho dividiu-se em várias fases. A primeira atividade, ocorreu com a importação e a instalação das bibliotecas necessárias. Depois disso teve a sua realização a leitura do arquivo em formato .csv.



A saída da primeira tela teve como resultado a instalação da biblioteca necessária.

Uma imagem contendo captura de tela

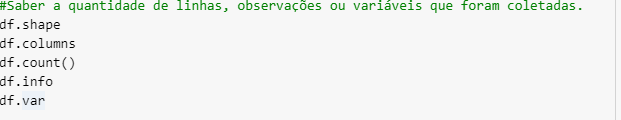
Descrição gerada automaticamente

A saída da segunda tela após a leitura do Data Frame coletado na etapa anterior, teve como resultado todas as informações contidas no arquivo.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

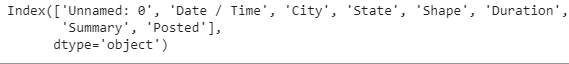
A segunda fase teve com a resolução da primeira questão proposta, que era coletar a quantidade de linhas, observações ou variáveis que foram coletadas.



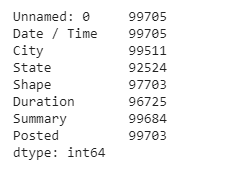
Para fazer as consultas foram utilizadas as funções shape, columns, count, info e var. Na Shape que o valor total de linhas e teve como saída após esse comando.



Para a função column que mostra todas as colunas contidas no arquivo e a saída gerada e mostrada na tela a seguir



Depois de fazer a pesquisa com o count, e gerado a quantidade de dados em cada coluna e a saída gerada possui como demostrado na imagem a seguir.



A penúltima função usada nessa consulta e o info que serve para informações. A saída teve como resposta a imagem exibida a seguir.

Uma imagem contendo texto

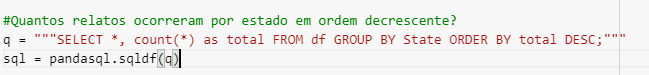
Descrição gerada automaticamente

E por último utilizou-se a função e a saída teve como resolução a tela seguir:

Uma imagem contendo texto

Descrição gerada automaticamente

A terceira fase, teve como a questão a ser respondida, a quantidade de relatos ocorridos por estados em ordem decrescente. Para solucionar essa questão foram selecionados todos os campos da base de dados coletada, e em seguida utilizou o count para coletar os estados, depois agrupou por estado e ordenou pelo apelido criado e no final utilizou o DESC, para organizar em ordem Decrescente.

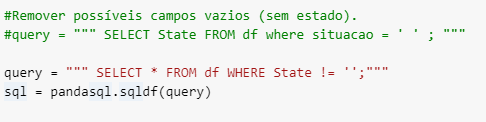


Depois dessa solução teve como resultado.

Uma imagem contendo captura de tela

Descrição gerada automaticamente

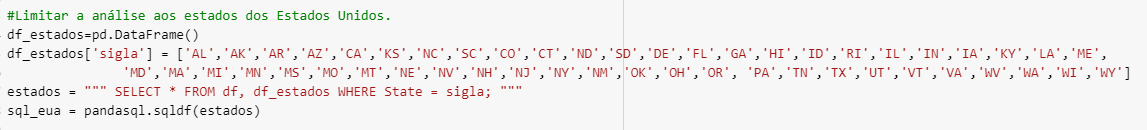
A quarta fase, precisa fazer as remoções de possíveis campos contidos na tabela ou sem estados. O comando utilizado foi feito porque precisava selecionar todos os campos do data frame, onde os estados sejam diferentes dos espaços vazios.



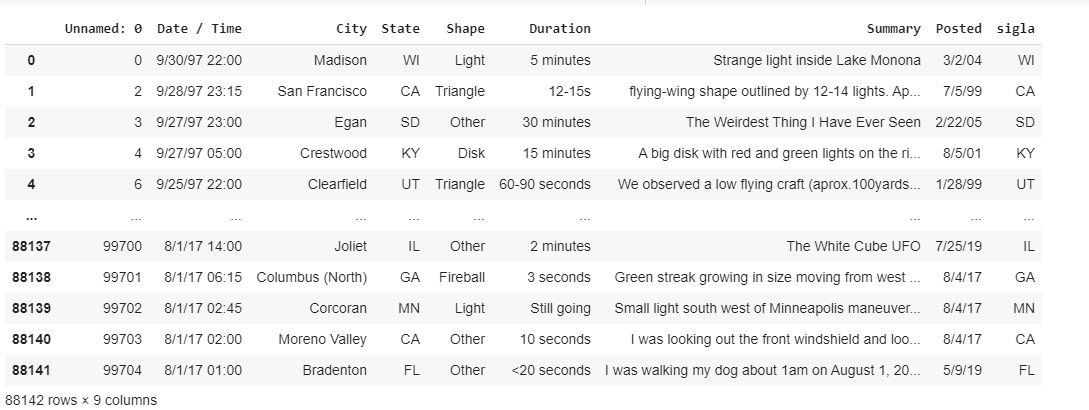
Essa tela teve como resultado o seguinte.



A quinta fase, era preciso limitar a análise aos estados dos Estados Unidos, para isso precisou criar um Data frame com todos os estados que formam o país. Para criar essa tarefa criou uma lista com todos os estados.



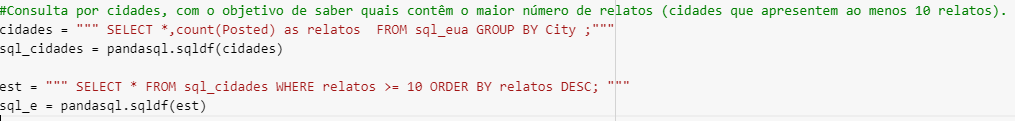
A saída dessa tela teve como resultado após a consulta, a tela a seguir.



A sexta fase, teve como a consulta por cidades com o objetivo de saber quais as cidades que contêm o maior número de relatos (cidades que apresentam ao menos 10 relatos.

A primeira consulta ocorreu porque era precisa selecionar todos os campos, e depois retornou após o uso do count, a quantidade de relatos e agrupados por cidades.

A segunda consultado precisou informar a onde teve mais relatos igual ou maior que 10 ordenados por relatos e exibidos em ordem decrescente.



O primeiro comanda para saber as cidades;

Uma imagem contendo captura de tela

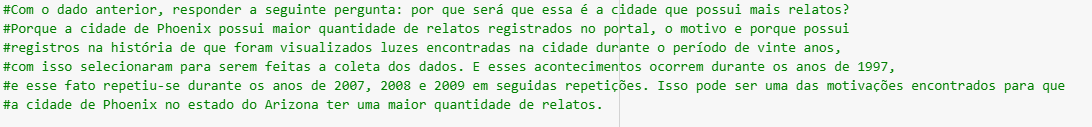
Descrição gerada automaticamente

A resposta do segundo comanda e o seguinte:

Uma imagem contendo captura de tela

Descrição gerada automaticamente

A sétima fase proposta do projeto e a resposta da questão porque que determinada cidade possui maior número de relatos.



A oitava fase proposta teve como atividade a criação de um query exclusiva. Para criar isso, precisou realizar três consultas. A primeira consultada selecionou todos os campos, e contado as postagens do Data Frame dos estados agrupando os estados. Em seguida foram selecionados os campos State, City e Shape do Data frame mostrado na etapa anterior e ordenados por relados. A terceira consulta foram feitas a partir das tabelas juntamente com cidades onde a quantidade de relatos seja maior que 10.

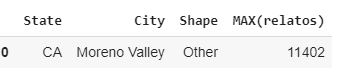
Uma imagem contendo captura de tela

Descrição gerada automaticamente

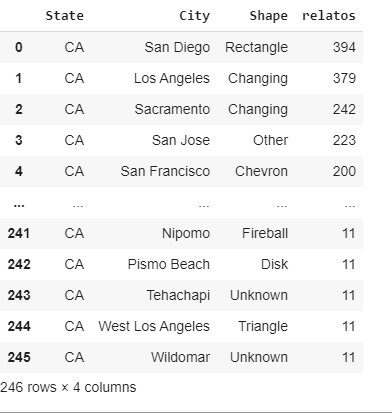
A saída do primeiro comando foi o seguinte

Uma imagem contendo texto

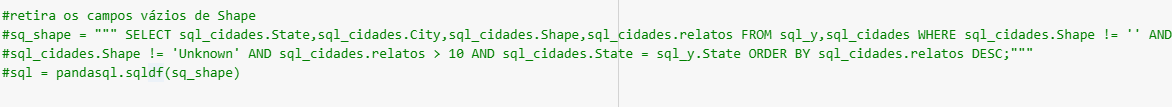
Descrição gerada automaticamente  
A saída do segundo comando foi o seguinte:



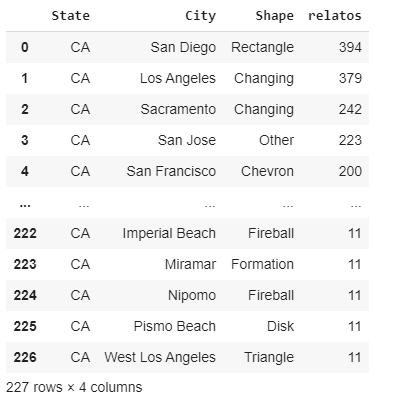
A saída do terceiro comando do terceiro comando foi o seguinte



Para realizar a última precisou retirar os campos vazios e Unknown de Shape.



A saída teve como resultado a tela a seguir.



**Ferramentas Utilizadas:**

O Colab para desenvolvimento e o Microsoft Excel para fazer a leitura e a visualização do arquivo em formato .csv.

**Link Para o Colab:**

github.com/vanessaformiga/trabalho-de-ci-ncia-de-dados

**Bibliografia**

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