Projeto de prática Integrada de ciência de dados, inteligência artificial e machine learning

Relatório de avistamento de Objetos Voadores Não Identificados.

Sprint 1 → Exploração de dados com SQL

Curso: Tecnologia em sistemas para internet

Estudantes:

Brenda Lopes Miranda Teixeira
Mateus Gomes da Silva Fonteles
Rickson Queiroz Marques de Souza
Samuel Araújo Lopes

Professores

Fábio Henrique
Diego Queiroz
Ana Régia

Brasília, julho de 2021

INSTITUTO FEDERAL
Brasília

Sumário

1. Objetivos

Nesta fase do projeto vamos explorar os dados coletados usando linguagem SQL fara realizar as seguintes tarefas:

- Saber a quantidade de linhas, observações ou variáveis que foram coletadas.
- Quantos relatos ocorreram por estado em ordem decrescente?
- Remover possíveis campos vazios (sem estado).
- Limitar a análise aos estados dos Estados Unidos.
- Consulta por cidades, com o objetivo de saber quais contêm o maior número de relatos (cidades que apresentem ao menos 10 relatos).
- Com o dado anterior, responder a seguinte pergunta: por que será que essa é a cidade que possui mais relatos?
- Fazer uma query exclusiva para o estado com maior número de relatos, buscando cidades que possuam um número superior a 10 relatórios. Enfatizar a cidade, a quantidade de relatos e formato do objeto não identificado.

2. Descrição do problema

Com alguns registros possuindo valores vazios, nessa etapa será necessário remover esses registros de valores vazios e manter registros somente dos cinquenta e um estados dos EUA. Feito isso passamos para a exploração dos dados usando SQL para filtrar as aparições por cidades e números de relatos.

3. Desenvolvimento

Este trabalho está sendo desenvolvido usando um Script Python por ser uma linguagem orientada a objetos é bastante maleável, o grupo está utilizando a plataforma Google Colaboratory, assim todos podem modificar e acrescentar o código quando necessário.

3.1. Código implementado

- Primeiramente foram importadas as bibliotecas:

```
!pip install pandasql
import pandasql
import pandas as pd
import numpy as np
```

- Logo foi carregado o arquivo OVNIS.csv em um dataframe:

```
ovnis = pd.read_csv('OVNIS.csv')
ovnis
```

- A primeira demanda é descobrir o número de linhas ou relatos foram coletados:

```
#Saber a quantidade de linhas coletadas
print ("Quantidade de linhas: ", len(ovnis))
```

- Logo, vamos apresentar o dataset por número de relatos por Estado, em ordem decrescente:

```
#Quantos relatos ocorreram por estado em ordem decrescente?
estados = ovnis.State.value_counts()
estados.sort_values(ascending=False)
```

- Logo removemos os registros com valores vazios.

```
#Removendo os NaN
ovnis['State'].dropna()
ovnis['Shape'].dropna()
ovnis['City'].dropna()

ovnis
```

- Logo selecionamos por Estado somente as instâncias de avistamento que ocorreram em estados dos EUA.

```
q = """
   SELECT * from ovnis WHERE State LIKE '%AK%' OR State LIKE '%AL%'
OR State LIKE '%AR%'
OR State LIKE '%AZ%'
OR State LIKE '%CA%'
OR State LIKE '%CO%'
OR State LIKE '%CT%'
OR State LIKE '%DE%'
OR State LIKE '%FL%'
```

```
OR State LIKE '%GA%'
OR State LIKE '%HI%'
OR State LIKE '%IA%'
OR State LIKE '%ID%'
OR State LIKE '%IL%'
OR State LIKE '%IN%'
OR State LIKE '%KS%'
OR State LIKE '%KY%'
OR State LIKE '%LA%'
OR State LIKE '%MA%'
OR State LIKE '%MD%'
OR State LIKE '%ME%'
OR State LIKE '%MI%'
OR State LIKE '%MN%'
OR State LIKE '%MO%'
OR State LIKE '%MS%'
OR State LIKE '%MT%'
OR State LIKE '%NC%'
OR State LIKE '%ND%'
OR State LIKE '%NE%'
OR State LIKE '%NH%'
OR State LIKE '%NJ%'
OR State LIKE '%NM%'
OR State LIKE '%NV%'
OR State LIKE '%NY%'
OR State LIKE '%OH%'
OR State LIKE '%OK%'
OR State LIKE '%OR%'
OR State LIKE '%PA%'
OR State LIKE '%RI%'
OR State LIKE '%SC%'
OR State LIKE '%SD%'
OR State LIKE '%TN%'
OR State LIKE '%TX%'
OR State LIKE '%UT%'
OR State LIKE '%VA%'
OR State LIKE '%VT%'
OR State LIKE '%WA%'
OR State LIKE '%WI%'
OR State LIKE '%WV%'
OR State LIKE '%WY%'
  """
just_us = pandasql.sqldf(q.lower(), locals())
```

```
just_us
```

- Em seguida vamos fazer uma query exclusiva para o estado com maior número de relatos, buscando cidades que possuam um número superior a 10 relatórios. Enfatizar a cidade, a quantidade de relatos e formato do objeto não identificado.

```
#Buscar cidade com maior número de relatos, mostrando cidade, número de
relatos e forma do objeto
query = '''
  SELECT CITY,COUNT(city) as Numero de Ocorrencias, Shape
  FROM just_us GROUP BY city, shape
  HAVING COUNT(city)>=10
  ORDER BY COUNT(city) DESC
  '''

  pandasql.sqldf(query.lower(), locals())
```

- Por fim, fazemos uma consideração sobre as informações coletadas.

O que podemos perceber é que as cidades com mais relatos tem em comum o fato de estarem na costa oeste e na região da Califórnia, onde existe uma menor densidade populacional (por ser uma região desértica) e portanto menos iluminação elétrica, o que leva a uma melhor visibilidade do céu.

O fato de estarem na costa oeste pode também apontar que os objetos avistados sejam manifestações de entidades habitantes do oceano pacífico norte, que podem eventualmente lançar-se aos céus, espalhando assim sua esfera de influência psicológica sobre os habitantes das regiões que estes consideram como seus “campos de colheita”, conforme descrito por HP Lovecraft em sua clássica obra sobre o tema: “O chamado de Cthulhu".

4. Considerações Finais

A parte de exploração dos dados é muito importante para que possamos chegar a conclusões a respeito dos mesmos. Um conjunto de dados muito grande pode fornecer respostas muito interessantes quando sabemos fazer as perguntas corretas, mas pode também não trazer informação alguma, caso não saibamos a forma de abordá-lo.

Referências

Dividir data (dia, mês, ano) em novas colunas - DataFrame Pandas. 2021. Disponível em:

- <https://pt.stackoverflow.com/questions/428236/dividir-data-dia-m%C3%AAs-ano-em-novas-colunas-dataframe-pandas>

H. P. Lovecraft

- O chamado de Cthulu,1926