

TECNOLOGIA EM SISTEMAS PARA INTERNET

Daniel Evangelista Pereira Ribson Coelho Cardoch Valdés Douglas Seidi Shibata

RELATÓRIO DE PRÁTICA INTEGRADA DE CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

29/09/2020

Brasília - DF

2020

Sumário

1. Objetivos	3
2. Descrição do problema	4
3.1 Código implementado	5
4. Considerações Finais	11
Referências	12

1. Objetivos

Realizar a limpeza dos dados que se apresentam nulos, em branco ou possuem os valores Unknown. Para que se possa realizar uma exploração mais precisa de dados. Carregar os dados do csv Ovnis para um dataframe, remover as variáveis irrelevantes para análise, manter os formatos mais populares que contêm mais de 1000 ocorrências e salvar em um novo dataframe e converte-lo para csv.

2. Descrição do problema

Realizar a limpeza dos dados que são irrelevantes para a análise. E delimitar os dados da coluna dos states para apenas os estados dos norte americanos.

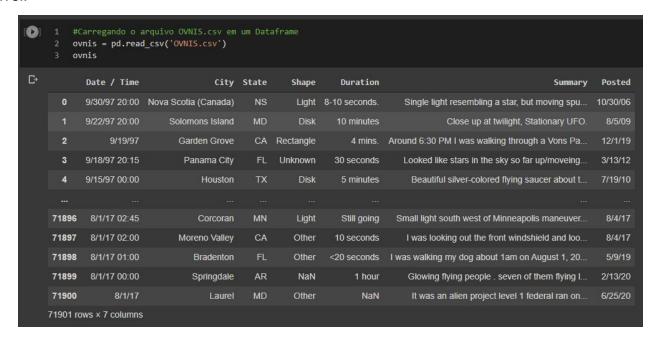
3. Desenvolvimento

As tecnologias utilizadas para a elaboração desta segunda fase do projeto, será a linguagem python e algumas bibliotecas para ajudar no desenvolvimento, com a biblioteca pandas e pandasql para a análise dos dados e remoção dos dados irrelevantes, por meio do notebook e o ambiente de desenvolvimento Google Colab.

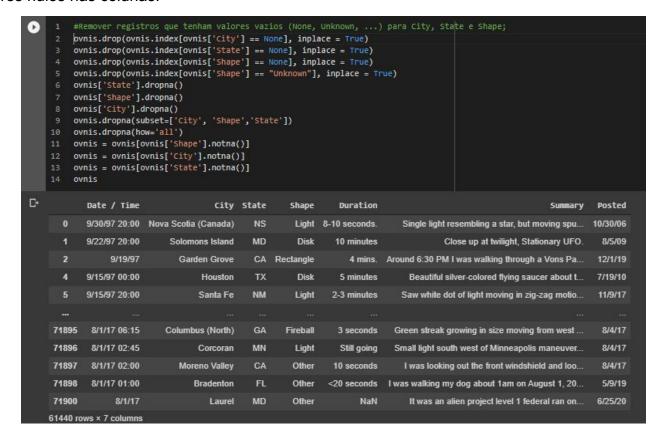
3.1 Código implementado

5.7 Limpeza dos dados

Carregando o arquivo Ovnis com a biblioteca pandas e atribuindo a uma variável.



Remoção dos valores que se apresentam nulos, vazios ou desconhecidos nas colunas City, State e Shape. Utilizando as funções drop e especificando o index e a coluna para ser removida. Depois utilizando a função dropna para remover os valores nulos nas colunas.



Executando a query para limitar apenas aos estados dos estados Unidos.

```
#Manter somente os registros referentes aos (51 estados dos Estados Unidos)
 q = """

✓ SELECT * from ovnis where

 STATE LIKE 'AL'
 OR STATE LIKE 'AK'
 OR STATE LIKE 'AZ'
 OR STATE LIKE 'AR'
 OR STATE LIKE 'CA'
 OR STATE LIKE 'CO'
 OR STATE LIKE 'CT'
 OR STATE LIKE 'DE'
 OR STATE LIKE 'DC'
 OR STATE LIKE 'FL'
 OR STATE LIKE 'GA'
 OR STATE LIKE 'HI'
 OR STATE LIKE 'ID'
 OR STATE LIKE 'IL'
 OR STATE LIKE 'IN'
 OR STATE LIKE 'IA'
 OR STATE LIKE 'KS'
 OR STATE LIKE 'KY'
 OR STATE LIKE 'LA'
 OR STATE LIKE 'ME'
 OR STATE LIKE 'MT'
 OR STATE LIKE 'NE'
```

	57 u	# Executa o seu comando SQL e retorna um dataframe usa_df = pandasql.sqldf(q.lower(), locals()) usa_df							
C>		Date / Time	City	State	Shape	Duration		Summary	Posted
	0	9/22/97 20:00	Solomons Island	MD	Disk	10 minutes	Close up	at twilight, Stationary UFO.	8/5/09
	1	9/19/97	Garden Grove	CA	Rectangle	4 mins.	Around 6:30 PM I was w	alking through a Vons Pa	12/1/19
	2	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-col	ored flying saucer about t	7/19/10
	3	9/15/97 20:00	Santa Fe	NM	Light	2-3 minutes	Saw white dot of ligh	t moving in zig-zag motio	11/9/17
	4	9/15/97 20:00	Kent	WA	Sphere	5 minutes	Was looking thru a tel	escope at the moon in 19	3/17/17
	59081	8/1/17 06:15	Columbus (North)	GA	Fireball	3 seconds	Green streak growing	in size moving from west	8/4/17
	59082	8/1/17 02:45	Corcoran	MN	Light	Still going	Small light south west	of Minneapolis maneuver	8/4/17
	59083	8/1/17 02:00	Moreno Valley	CA	Other	10 seconds	I was looking out the	e front windshield and loo	8/4/17
	59084	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog at	oout 1am on August 1, 20	5/9/19
	59085	8/1/17	Laurel	MD	Other	None	It was an alien pro	ject level 1 federal ran on	6/25/20
	59086 r	ows × 7 columns				<u> </u>			

Removendo as colunas irrelevantes para a análise utilizando a função drop e atribuindo a uma nova variável que irá receber o novo dataframe limpo.

0	2 li		eis irrelevantes drop(["Duration"		
₽		Date / Time	City	State	Shape
	0	9/22/97 20:00	Solomons Island	MD	Disk
	1	9/19/97	Garden Grove	CA	Rectangle
	2	9/15/97 00:00	Houston	TX	Disk
	3	9/15/97 20:00	Santa Fe	NM	Light
	4	9/15/97 20:00	Kent	WA	Sphere
	59081	8/1/17 06:15	Columbus (North)	GA	Fireball
	59082	8/1/17 02:45	Corcoran	MN	Light
	59083	8/1/17 02:00	Moreno Valley	CA	Other
	59084	8/1/17 01:00	Bradenton	FL	Other
	59085	8/1/17	Laurel	MD	Other
	59086 го	ws × 4 columns	1		

Verificando os formatos mais populares para que sejam removidos os que apresentam menos de 1000 ocorrências.

```
#Manter somente os registros de Shapes mais populares (com mais de 1000 ocorrências);
        shapes_counts = limpo['Shape'].value_counts()
        shapes_counts
Light 13936
Circle 7445
Triangle 5939
              5911
   Fireball
   Sphere
               4548
   Other
               4106
   Oval
               2849
   Disk
               2799
               2126
   Formation
   Changing
                1601
   Cigar
                 1303
   Flash
                 1248
   Rectangle 1131
   Cylinder
                 981
   Diamond
                938
   Chevron
                 698
    Teardrop
                 591
                 459
   Egg
   Cone
   Cross
   Name: Shape, dtype: int64
```

Verificando os valores abaixo de 1000 ocorrências e depois excluindo estes valores do dataframe.

```
[34]
         below 1000 = shapes counts[shapes counts< 1000]
         below 1000
Cylinder
                981
    Diamond
                938
    Chevron
                698
    Teardrop
               591
                459
    Egg
    Cone
               262
    Cross
               215
    Name: Shape, dtype: int64
         limpo.drop(limpo.index[limpo['Shape'] == "Cylinder"], inplace = True)
[35] 1
         limpo.drop(limpo.index[limpo['Shape'] == "Diamond"], inplace = True)
         limpo.drop(limpo.index[limpo['Shape'] == "Chevron"], inplace = True)
         limpo.drop(limpo.index[limpo['Shape'] == "Teardrop"], inplace = True)
         limpo.drop(limpo.index[limpo['Shape'] == "Egg"], inplace = True)
         limpo.drop(limpo.index[limpo['Shape'] == "Cone"], inplace = True)
         limpo.drop(limpo.index[limpo['Shape'] == "Cross"], inplace = True)
```

Gerando um csv a partir da limpeza dos dados que foram solicitados. E depois lendo o csv gerado e mostrando na tela

[36]	<pre>#Salvar o dataframe final em um arquivo CSV com o nome "df_OVNI_limpo" limpo.to_csv('df_OVNI_limpo.csv',index=False) df_OVNI_limpo = pd.read_csv('df_OVNI_limpo.csv') df_OVNI_limpo = df_OVNI_limpo[df_OVNI_limpo['Shape'].notna()] df_OVNI_limpo</pre>								
C→		Date / Time	City	State	Shape				
3	0	9/22/97 20:00	Solomons Island	MD	Disk				
	1	9/19/97	Garden Grove	CA	Rectangle				
	2	9/15/97 00:00	Houston	TX	Disk				
	3	9/15/97 20:00	Santa Fe	NM	Light				
	4	9/15/97 20:00	Kent	WA	Sphere				
	54937	8/1/17 06:15	Columbus (North)	GA	Fireball				
	54938	8/1/17 02:45	Corcoran	MN	Light				
	54939	8/1/17 02:00	Moreno Valley	CA	Other				
	54940	8/1/17 01:00	Bradenton	FL	Other				
	54941	8/1/17	Laurel	MD	Other				
54942 rows × 4 columns									

4. Considerações Finais

Nesta parte do script foi realizado a remoção dos dados que não são essenciais para a continuidade do projeto e por não apresentar um padrão de dados nulos, ou irrelevantes, a remoção acabou sendo um pouco difícil, mas foi obtido o resultado esperado.

Referências

Raymond, Pandas DataFrame Plot - Bar Chart Disponível em https://kontext.tech/column/code-snippets/399/pandas-dataframe-plot-bar-chart Acesso em 27 de setembro de 2020