INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Brasília
Campus Brasília

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. Desenvolvimento **5**

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.

```
#Carregando o arquivo OVNIS.csv em um Dataframe
ovnis = pd.read_csv('OVNIS.csv')
ovnis
```

	Date / Time	City	State	Shape	Duration	Summary	Posted
0	9/30/97 20:00	Nova Scotia (Canada)	NS	Light	8-10 seconds.	Single light resembling a star, but moving spu...	10/30/06
1	9/22/97 20:00	Solomons Island	MD	Disk	10 minutes	Close up at twilight, Stationary UFO.	8/5/09
2	9/19/97	Garden Grove	CA	Rectangle	4 mins.	Around 6:30 PM I was walking through a Vons Pa...	12/1/19
3	9/18/97 20:15	Panama City	FL	Unknown	30 seconds	Looked like stars in the sky so far up/moveing...	3/13/12
4	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-colored flying saucer about t...	7/19/10
...	...	...	...	...	...	...	...
71896	8/1/17 02:45	Corcoran	MN	Light	Still going	Small light south west of Minneapolis maneuver...	8/4/17
71897	8/1/17 02:00	Moreno Valley	CA	Other	10 seconds	I was looking out the front windshield and loo...	8/4/17
71898	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog about 1am on August 1, 20...	5/9/19
71899	8/1/17 00:00	Springdale	AR	NaN	1 hour	Glowing flying people . seven of them flying l...	2/13/20
71900	8/1/17	Laurel	MD	Other	NaN	It was an alien project level 1 federal ran on...	6/25/20

71901 rows × 7 columns

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.

```python
#Remover registros que tenham valores vazios (None, Unknown, ...) para City, State e Shape;
ovnis.drop(ovnis.index[ovnis['City'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['State'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['Shape'] == None], inplace = True)
ovnis.drop(ovnis.index[ovnis['Shape'] == "Unknown"], inplace = True)
ovnis['State'].dropna()
ovnis['Shape'].dropna()
ovnis['City'].dropna()
ovnis.dropna(subset=['City', 'Shape','State'])
ovnis.dropna(how='all')
ovnis = ovnis[ovnis['Shape'].notna()]
ovnis = ovnis[ovnis['City'].notna()]
ovnis = ovnis[ovnis['State'].notna()]
ovnis
```

	Date / Time	City	State	Shape	Duration	Summary	Posted
0	9/30/97 20:00	Nova Scotia (Canada)	NS	Light	8-10 seconds.	Single light resembling a star, but moving spu...	10/30/06
1	9/22/97 20:00	Solomons Island	MD	Disk	10 minutes	Close up at twilight, Stationary UFO.	8/5/09
2	9/19/97	Garden Grove	CA	Rectangle	4 mins.	Around 6:30 PM I was walking through a Vons Pa...	12/1/19
4	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-colored flying saucer about t...	7/19/10
5	9/15/97 20:00	Santa Fe	NM	Light	2-3 minutes	Saw white dot of light moving in zig-zag motio...	11/9/17
...	...	...	...	...	...	...	...
71895	8/1/17 06:15	Columbus (North)	GA	Fireball	3 seconds	Green streak growing in size moving from west ...	8/4/17
71896	8/1/17 02:45	Corcoran	MN	Light	Still going	Small light south west of Minneapolis maneuver...	8/4/17
71897	8/1/17 02:00	Moreno Valley	CA	Other	10 seconds	I was looking out the front windshield and loo...	8/4/17
71898	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog about 1am on August 1, 20...	5/9/19
71900	8/1/17	Laurel	MD	Other	NaN	It was an alien project level 1 federal ran on...	6/25/20

61440 rows × 7 columns

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'
```

```
55  """
56  # Executa o seu comando SQL e retorna um dataframe
57  usa_df = pandasql.sqldf(q.lower(), locals())
58  usa_df
```

	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 walking through a Vons Pa...	12/1/19
2	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-colored flying saucer about t...	7/19/10
3	9/15/97 20:00	Santa Fe	NM	Light	2-3 minutes	Saw white dot of light moving in zig-zag motio...	11/9/17
4	9/15/97 20:00	Kent	WA	Sphere	5 minutes	Was looking thru a telescope 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 front windshield and loo...	8/4/17
59084	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog about 1am on August 1, 20...	5/9/19
59085	8/1/17	Laurel	MD	Other	None	It was an alien project level 1 federal ran on...	6/25/20

59086 rows × 7 columns

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

```
#Remover variáveis irrelevantes para a análise (Duration, Summary e Posted);
limpo = usa_df.drop(["Duration", 'Summary' ,'Posted'],axis=1)
limpo
```

	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 rows × 4 columns

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
Fireball      5911
Sphere        4548
Other         4106
Oval          2849
Disk          2799
Formation     2126
Changing      1601
Cigar         1303
Flash         1248
Rectangle     1131
Cylinder       981
Diamond        938
Chevron        698
Teardrop       591
Egg            459
Cone           262
Cross          215
Name: Shape, dtype: int64
```

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

```
[34]  1  below_1000  = shapes_counts[shapes_counts< 1000]
      2  below_1000
```

```
Cylinder    981
Diamond     938
Chevron     698
Teardrop    591
Egg         459
Cone        262
Cross       215
Name: Shape, dtype: int64
```

```
[35]  1  limpo.drop(limpo.index[limpo['Shape'] == "Cylinder"], inplace = True)
      2  limpo.drop(limpo.index[limpo['Shape'] == "Diamond"], inplace = True)
      3  limpo.drop(limpo.index[limpo['Shape'] == "Chevron"], inplace = True)
      4  limpo.drop(limpo.index[limpo['Shape'] == "Teardrop"], inplace = True)
      5  limpo.drop(limpo.index[limpo['Shape'] == "Egg"], inplace = True)
      6  limpo.drop(limpo.index[limpo['Shape'] == "Cone"], inplace = True)
      7  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]  1  #Salvar o dataframe final em um arquivo CSV com o nome "df_OVNI_limpo"
      2  limpo.to_csv('df_OVNI_limpo.csv',index=False)
      3  df_OVNI_limpo = pd.read_csv('df_OVNI_limpo.csv')
      4  df_OVNI_limpo = df_OVNI_limpo[df_OVNI_limpo['Shape'].notna()]
      5  df_OVNI_limpo
```

	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
...	...	...	...	...
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