

TECNOLOGIA EM SISTEMAS PARA INTERNET

**Emanuelle Ferreira da Silva
Lucas Matheus Silva
Neander Wendel Nobre Teixeira**

**RELATÓRIO DE PRÁTICA INTEGRADA
DE
CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL**

Brasília - DF

03/10/2020

Sumário

1. Objetivos	3
2. Descrição do problema	4
3. Desenvolvimento	5
3.1 Código implementado	5
4. Considerações Finais	6
Referências	7

1. Objetivos

O presente projeto se encontra na fase de limpeza dos dados, o principal objetivo é fazer a limpeza dos dados e descobrir os seus respectivos valores.

2. Descrição do problema

O problema a ser desenvolvido nessa etapa do projeto consiste na limpeza dos dados da fase anterior, removendo registros nas colunas da tela que apresenta algum valor vazio (none, Unknown), remoção de variáveis irrelevantes e além de manter registros de dados específicos requisitados no projeto.

3. Desenvolvimento

Para o desenvolvimento dessa etapa foi necessária a utilização da biblioteca pandas e suas funções que auxiliam no decorrer das etapas como a função read para ler o arquivo OVNIS.csv que foi gerado da etapa anterior drop (como alguns argumentos), dropna são utilizadas para realizar a remoção, foi utilizado o pandas sql por meio do query para limitar o data frame para conter apenas os registros dos estados unidos, o for in para contar os shapes da tabela e realizar a limpeza.

3.1 Código implementado

#5.7 Limpeza dos dados

1- Importando a biblioteca pandas, na linha de baixo o arquivo ovnis.csv está sendo carregado em um data frame chamado ovnis_two.

```
6 import pandas as pd
7 ovnis_two = pd.read_csv("OVNIS.csv")
```

2- Por meio da função dropna é possível remover os valores vazios das colunas state, city, shape da tabela ovnis_two.

```
3 ovnis_two["State"].dropna()
4 ovnis_two["City"].dropna()
5 ovnis_two["Shape"].dropna()
```

Removendo os valores em branco por meio da função drop e seus argumentos das colunas state, city e shape.

```
ovnis_two.drop(ovnis_two.index[ovnis_two['State']==None],inplace=True)
ovnis_two.drop(ovnis_two.index[ovnis_two['City']==None],inplace=True)
ovnis_two.drop(ovnis_two.index[ovnis_two['Shape']==None],inplace=True)

ovnis_two
```

A tabela é gerada abaixo.

	Date / Time	City	State	Shape	Duration	Summary	Posted
0	1/29/97 23:15	East Greenwich	RI	Disk	5 minutes	I witnessed a UFO which may be of the Lazar"Sp...	3/7/98
1	1/26/97 22:00	Flagstaff	AZ	Light	six minutes	It traveled at about the same speed we were go...	3/19/09
2	1/25/97 21:00	Marion	WI	Triangle	2 minutes	On a camping trip 3 triangle shaped objects we...	3/7/98
3	1/25/97 06:00	Mount Hope/Binbrook (Canada)	ON	Disk	1/2 hour	A large disk or saucer type object, approximat...	4/28/01
4	1/24/97 19:00	Alta	UT	Other	3 hours plus	A "moving star" similar to a satellite stopped...	3/7/98
...
102386	12/1/17 17:00	Foyil	OK	Formation	All night	At dusk my wife and I noticed a star that look...	12/10/17
102387	12/1/17 04:00	Chesapeake	VA	Light	5 minutes	Light was moving at a constant speed, vanished...	12/4/17
102388	12/1/17 04:00	Boise	ID	Cigar	10 minutes	Large cigar-shaped UFO with visible cabin lights.	12/4/17
102389	12/1/17 02:06	Ras Al khaimah (Oman/UAE)	NaN	Light	10 seconds	UFO PASSING.	12/4/17
102390	12/1/17 01:00	Wasilla	AK	Flash	25 minutes	Flashing, fast moving light over Alaska.	12/4/17

102391 rows × 7 columns

3- Utilizando a biblioteca pandasql por meio do query para limitar o data frame ovnis_two para conter os apenas as ocorrências dos estados unidos.

```

5  q = ""
6  | SELECT * from ovnis_two where State LIKE '%AK%' OR State LIKE '%AL%'
7  OR State LIKE '%AR%'
8  OR State LIKE '%AZ%'
9  OR State LIKE '%CA%'
10 OR State LIKE '%CO%'
11 OR State LIKE '%CT%'
12 OR State LIKE '%DE%'
13 OR State LIKE '%FL%'
14 OR State LIKE '%GA%'
15 OR State LIKE '%HI%'
16 OR State LIKE '%IA%'
17 OR State LIKE '%ID%'
18 OR State LIKE '%IL%'
19 OR State LIKE '%IN%'
20 OR State LIKE '%KS%'
21 OR State LIKE '%KY%'
22 OR State LIKE '%LA%'
23 OR State LIKE '%MA%'
24 OR State LIKE '%MD%'
25 OR State LIKE '%ME%'
26 OR State LIKE '%MI%'
27 OR State LIKE '%MN%' OR State LIKE '%MO%' OR State LIKE '%MS%' OR State LIKE '%MT%' OR State LIKE '%NC%' OR State LIKE '%ND%'
28 OR State LIKE '%NE%'
29 OR State LIKE '%NH%'
30 OR State LIKE '%NJ%'

```

```

29 OR State LIKE '%NH%'
30 OR State LIKE '%NJ%'
31 OR State LIKE '%NM%'
32 OR State LIKE '%NV%'
33 OR State LIKE '%NY%'
34 OR State LIKE '%OH%'
35 OR State LIKE '%OK%'
36 OR State LIKE '%OR%'
37 OR State LIKE '%PA%'
38 OR State LIKE '%RI%'
39 OR State LIKE '%SC%'
40 OR State LIKE '%SD%'
41 OR State LIKE '%TN%'
42 OR State LIKE '%TX%'
43 OR State LIKE '%UT%'
44 OR State LIKE '%VA%'
45 OR State LIKE '%VT%'
46 OR State LIKE '%WA%'
47 OR State LIKE '%WI%'
48 OR State LIKE '%WV%'
49 OR State LIKE '%WY%'
50
51 #atribuindo o resultado da consulta sql a variavel estados unidos
52 estados_unidos = pandasql.sqldf(q.lower(), locals())
53 estados_unidos

```

A tabela gerada contém somente os registros dos estados unidos.

	Date / Time	City	State	Shape	Duration	Summary	Posted
0	1/29/97 23:15	East Greenwich	RI	Disk	5 minutes	I witnessed a UFO which may be of the Lazar"Sp...	3/7/98
1	1/26/97 22:00	Flagstaff	AZ	Light	six minutes	It traveled at about the same speed we were go...	3/19/09
2	1/25/97 21:00	Marion	WI	Triangle	2 minutes	On a camping trip 3 triangle shaped objects we...	3/7/98
3	1/24/97 19:00	Alta	UT	Other	3 hours plus	A "moving star" similar to a satellite stopped...	3/7/98
4	1/23/97 18:30	North Kingstown	RI	Triangle	3 minutes	1/23/97: Huge black triangle w/ rearguard heli...	1/27/05
...
90587	12/1/17 17:00	New Rochelle	NY	Sphere	5 minutes	Red light sphere moving in vertical and horizo...	12/4/17
90588	12/1/17 17:00	Foyil	OK	Formation	All night	At dusk my wife and I noticed a star that look...	12/10/17
90589	12/1/17 04:00	Chesapeake	VA	Light	5 minutes	Light was moving at a constant speed, vanished...	12/4/17
90590	12/1/17 04:00	Boise	ID	Cigar	10 minutes	Large cigar-shaped UFO with visible cabin lights.	12/4/17
90591	12/1/17 01:00	Wasilla	AK	Flash	25 minutes	Flashing, fast moving light over Alaska.	12/4/17

90592 rows x 7 columns

4-Criando um novo data frame para armazenar a nova tabela, por meio da função drop as colunas duration. summary e posted da tabela foram removidas.

```

4 df_OVNI_limpo = estados_unidos.drop(columns=['Duration','Summary','Posted'], axis=1)
5 df_OVNI_limpo

```

Resultado da nova tabela se as colunas duration. summary e posted.

	Date / Time	City	State	Shape
0	1/29/97 23:15	East Greenwich	RI	Disk
1	1/26/97 22:00	Flagstaff	AZ	Light
2	1/25/97 21:00	Marion	WI	Triangle
3	1/24/97 19:00	Alta	UT	Other
4	1/23/97 18:30	North Kingstown	RI	Triangle
...
90587	12/1/17 17:00	New Rochelle	NY	Sphere
90588	12/1/17 17:00	Foyil	OK	Formation
90589	12/1/17 04:00	Chesapeake	VA	Light
90590	12/1/17 04:00	Boise	ID	Cigar
90591	12/1/17 01:00	Wasilla	AK	Flash

90592 rows × 4 columns

5- Contando os objetos da coluna shape por meio da função value_counts.

```
2 contando_shape = df_OVNI_limpo['Shape'].value_counts()
3 contando_shape
```


Resultado

```
Light      19383
Circle     9546
Triangle   8601
Fireball   7638
Unknown    6538
Other       6125
Sphere     6032
Disk       4408
Oval       3963
Formation  2977
Changing   2318
Cigar      1962
Flash      1738
Rectangle  1475
Cylinder   1396
Diamond    1349
Chevron    1095
Teardrop   850
Egg        766
Cone       374
Cross      290
light       49
other       16
triangle    15
sphere      7
circle      6
```

6- Utilizando o for in para contar os shapes e excluir os que possuem valores menores que 1000.

```
1 for i in contando_shape[contando_shape < 1000].index: df_OVNI_limpo.drop(df_OVNI_limpo.index[df_OVNI_limpo['Shape'] == i], inplace =
2
3 df_OVNI_limpo
```

Resultado da tabela.



	Date / Time	City	State	Shape
0	1/29/97 23:15	East Greenwich	RI	Disk
1	1/26/97 22:00	Flagstaff	AZ	Light
2	1/25/97 21:00	Marion	WI	Triangle
3	1/24/97 19:00	Alta	UT	Other
4	1/23/97 18:30	North Kingstown	RI	Triangle
...
90587	12/1/17 17:00	New Rochelle	NY	Sphere
90588	12/1/17 17:00	Foyil	OK	Formation
90589	12/1/17 04:00	Chesapeake	VA	Light
90590	12/1/17 04:00	Boise	ID	Cigar
90591	12/1/17 01:00	Wasilla	AK	Flash

88197 rows × 4 columns

7-Salvando o data frame final em um arquivo csv.

```
3 df_OVNI_limpo.to_csv("df_OVNI_limpo.csv", index=False)
```

4. Considerações Finais

Por meio dessa etapa do projeto podemos concluir que é muito importante realizamos a limpeza dos dados para a melhor análise e compreensão dos dados.

Referências

FIGUEIREDO, Vinicius. Seus Primeiros Passos com Data Scientist: Introdução ao Pandas. **Data Hackers**. São Paulo, 30 de maio de 2018. Disponível em:

<<https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-simples-ao-pandas-1e15ea37fa1>>. Acesso em: 08 de Set. de 2020.

FAMETHEMES. Biblioteca seaborn com o matplotlib. **Vooo - Insights**. c2020

Disponível em: <<https://www.vooo.pro/insights/biblioteca-seaborn-com-o-matplotlib/>>. Acesso em: 26 de set. de 2020

MATHEUS, Yuri. Matplotlib uma biblioteca Python para gerar gráficos interessantes. **alura. São Paulo**, 27 de Novembro de 2018. Disponível em:

<<https://www.alura.com.br/artigos/criando-graficos-no-python-com-a-matplotlib>>. Acesso em: 25 de set. de 2020.