dados

October 16, 2020

In [10]:
```
# Pandas is used for data manipulation
import pandas as pd # bibliote Pandas é usada para manipulação de dados
# CSV file
data = pd.read_csv('data/iris-with-errors.csv', header=(0))
print("Número de linhas e colunas:",data.shape)
data.head(25)
```

```
Número de linhas e colunas: (25, 5)
```

Out[10]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	duplicada
1	5.1	3.5	1.4	0.2	duplicada
2	?	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	5.1	3.5	1.4	0.2	duplicada
5	NaN	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	duplicada
8	5.4	3.9	1.7	0.4	duplicada
9	4.6	3.4	1.4	NaN	setosa
10	5	3.4	1.5	0.2	setosa
11	4.4	2.9	1.4	0.2	duplicada
12	4.9	3.1	1.5	0.1	setosa
13	5.4	3.7	1.5	0.2	setosa
14	4.4	2.9	1.4	0.2	duplicada
15	4.8	3.4	1.6	0.2	setosa
16	4.8	3	1.4	0.1	setosa
17	4.4	2.9	1.4	0.2	duplicada
18	4.3	3	1.1	0.1	setosa
19	5.8	4	1.2	0.2	setosa
20	5.7	4.4	1.5	0.4	setosa
21	5.4	3.9	1.3	?	setosa
22	5.1	3.5	1.4	0.3	setosa
23	5.7	?	1.7	0.3	setosa
24	NaN	3.8	1.5	0.3	setosa

In [11]:
```
import numpy as np
# Substitui Nan por um caracter desejado
```

```
data = data.replace('?', np.nan)
data.head(25)
```

Out[11]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	duplicada
1	5.1	3.5	1.4	0.2	duplicada
2	NaN	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	5.1	3.5	1.4	0.2	duplicada
5	NaN	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	duplicada
8	5.4	3.9	1.7	0.4	duplicada
9	4.6	3.4	1.4	NaN	setosa
10	5	3.4	1.5	0.2	setosa
11	4.4	2.9	1.4	0.2	duplicada
12	4.9	3.1	1.5	0.1	setosa
13	5.4	3.7	1.5	0.2	setosa
14	4.4	2.9	1.4	0.2	duplicada
15	4.8	3.4	1.6	0.2	setosa
16	4.8	3	1.4	0.1	setosa
17	4.4	2.9	1.4	0.2	duplicada
18	4.3	3	1.1	0.1	setosa
19	5.8	4	1.2	0.2	setosa
20	5.7	4.4	1.5	0.4	setosa
21	5.4	3.9	1.3	NaN	setosa
22	5.1	3.5	1.4	0.3	setosa
23	5.7	NaN	1.7	0.3	setosa
24	NaN	3.8	1.5	0.3	setosa

In [12]:
```
# remove as linhas com NaN
data = data.dropna()
print("Número de linhas e colunas:",data.shape)
data.head(25)
```

```
Número de linhas e colunas: (19, 5)
```

Out[12]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	duplicada
1	5.1	3.5	1.4	0.2	duplicada
3	4.7	3.2	1.3	0.2	setosa
4	5.1	3.5	1.4	0.2	duplicada
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	duplicada
8	5.4	3.9	1.7	0.4	duplicada
10	5	3.4	1.5	0.2	setosa
11	4.4	2.9	1.4	0.2	duplicada
12	4.9	3.1	1.5	0.1	setosa

```
13           5.4          3.7           1.5          0.2     setosa
14           4.4          2.9           1.4          0.2  duplicada
15           4.8          3.4           1.6          0.2     setosa
16           4.8            3           1.4          0.1     setosa
17           4.4          2.9           1.4          0.2  duplicada
18           4.3            3           1.1          0.1     setosa
19           5.8            4           1.2          0.2     setosa
20           5.7          4.4           1.5          0.4     setosa
22           5.1          3.5           1.4          0.3     setosa
```

In [13]:
```
# Remove as linhas duplicadas
data = data.drop_duplicates()
data.head(25)
```

Out[13]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	duplicada
3	4.7	3.2	1.3	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	duplicada
10	5	3.4	1.5	0.2	setosa
11	4.4	2.9	1.4	0.2	duplicada
12	4.9	3.1	1.5	0.1	setosa
13	5.4	3.7	1.5	0.2	setosa
15	4.8	3.4	1.6	0.2	setosa
16	4.8	3	1.4	0.1	setosa
18	4.3	3	1.1	0.1	setosa
19	5.8	4	1.2	0.2	setosa
20	5.7	4.4	1.5	0.4	setosa
22	5.1	3.5	1.4	0.3	setosa