Tutorial 6 - Handle Missing Data replace function

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv('sample_data_tutorial_06.csv')
df
```

Out[2]:

	day	temperature	windspeed	event
0	1/1/2017	32	6	Rain
1	1/2/2017	-99999	7	Sunny
2	1/3/2017	28	-99999	Snow
3	1/4/2017	-99999	7	No event
4	1/5/2017	32	-88888	Rain
5	1/6/2017	31	2	Sunny
6	1/6/2017	34	5	No event

In [3]:
```
newdf = df.replace(-99999,np.NaN)
newdf
```

Out[3]:

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/2/2017	NaN	7.0	Sunny
2	1/3/2017	28.0	NaN	Snow
3	1/4/2017	NaN	7.0	No event
4	1/5/2017	32.0	-88888.0	Rain
5	1/6/2017	31.0	2.0	Sunny
6	1/6/2017	34.0	5.0	No event

In [4]:
```
newdf = df.replace([-99999, -88888],np.NaN)
newdf
```

Out[4]:

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/2/2017	NaN	7.0	Sunny
2	1/3/2017	28.0	NaN	Snow
3	1/4/2017	NaN	7.0	No event
4	1/5/2017	32.0	NaN	Rain
5	1/6/2017	31.0	2.0	Sunny
6	1/6/2017	34.0	5.0	No event

In [5]:
```
newdf = df.replace({
        'temperature': -99999,
        'windspeed': [-99999, -88888],
        'event': 'No event'
    }, np.NaN)
newdf
```

Out[5]:

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/2/2017	NaN	7.0	Sunny
2	1/3/2017	28.0	NaN	Snow
3	1/4/2017	NaN	7.0	NaN
4	1/5/2017	32.0	NaN	Rain
5	1/6/2017	31.0	2.0	Sunny
6	1/6/2017	34.0	5.0	NaN

In [6]:
```
# Podemos gerar um mapa das alterações que queremos fazer:
newdf = df.replace({
    -99999: np.NaN,
    -88888: np.NaN,
    'No event': 'Sunny'
})
newdf
```

Out[6]:

	day	temperature	windspeed	event
0	1/1/2017	32.0	6.0	Rain
1	1/2/2017	NaN	7.0	Sunny
2	1/3/2017	28.0	NaN	Snow
3	1/4/2017	NaN	7.0	Sunny
4	1/5/2017	32.0	NaN	Rain
5	1/6/2017	31.0	2.0	Sunny
6	1/6/2017	34.0	5.0	Sunny

In [7]:
```
# Importando outro csv com algumas unidades que precisam ser limpas!
df = pd.read_csv('sample_data_tutorial_06a.csv')
df
```

Out[7]:

	day	temperature	windspeed	event
0	1/1/2017	32 F	6 mph	Rain
1	1/2/2017	-99999	7mph	Sunny
2	1/3/2017	28	-99999	Snow
3	1/4/2017	-99999	7	No event
4	1/5/2017	32C	-88888	Rain
5	1/6/2017	31	2	Sunny
6	1/6/2017	34	5	No event

In [8]:
```
# É necessário usar o 'regex' (regular expression)
# No caso abaixo estamos substituindo todas as letras (de A a Z - maiúscula e minús
cula) por vazio ('')
newdf = df.replace('[A-Za-z]', '', regex=True)
newdf
```

Out[8]:

	day	temperature	windspeed	event
0	1/1/2017	32	6	
1	1/2/2017	-99999	7	
2	1/3/2017	28	-99999	
3	1/4/2017	-99999	7	
4	1/5/2017	32	-88888	
5	1/6/2017	31	2	
6	1/6/2017	34	5	

In [9]:
```
# Observe, no caso anterior, que ele removeu o que pedimos mas também removeu toda
 coluna 'event'
# Para fazer as substituições somente em determinadas colunas é preciso utilizar o
 dicionário:
newdf = df.replace({
    'temperature': '[A-Za-z]',
    'windspeed': '[A-Za-z]'
}, '', regex=True)
newdf
```

Out[9]:

	day	temperature	windspeed	event
0	1/1/2017	32	6	Rain
1	1/2/2017	-99999	7	Sunny
2	1/3/2017	28	-99999	Snow
3	1/4/2017	-99999	7	No event
4	1/5/2017	32	-88888	Rain
5	1/6/2017	31	2	Sunny
6	1/6/2017	34	5	No event