MissingValuesinPandas

November 17, 2016

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

In [2]: ufo = pd.read_csv('http://bit.ly/uforeports')

In [3]: ufo.tail()
```

Out[3]:

	City	Colors Reported	Shape Reported	State	Time
18236	Grant Park	NaN	TRIANGLE	IL	12/31/2000 23:00
18237	Spirit Lake	NaN	DISK	IA	12/31/2000 23:00
18238	Eagle River	NaN	NaN	WI	12/31/2000 23:45
18239	Eagle River	RED	LIGHT	WI	12/31/2000 23:45
18240	Ybor	NaN	OVAL	FL	12/31/2000 23:59

```
In [4]: ufo.isnull().tail() # true if missing, false if not missing
```

Out[4]:

	City	Colors Reported	Shape Reported	State	Time
18236	False	True	False	False	False
18237	False	True	False	False	False
18238	False	True	True	False	False
18239	False	False	False	False	False
18240	False	True	False	False	False

```
In [5]: ufo.notnull().tail()
```

Out[5]:

	City	Colors Reported	Shape Reported	State	Time
18236	True	False	True	True	True
18237	True	False	True	True	True
18238	True	False	False	True	True
18239	True	True	True	True	True
18240	True	False	True	True	True

```
In [6]: ufo.isnull().sum() # number of missing values in each column
```

Out[6]:

```
City                  25
Colors Reported    15359
Shape Reported      2644
State                  0
Time                   0
dtype: int64
```

```
In [8]: pd.Series([True, False, True]).sum()

Out[8]: 2

In [10]: ufo[ufo.City.isnull()] # look at portion -> NaN
```

Out[10]:

	City	Colors Reported	Shape Reported	State	Time
21	NaN	NaN	NaN	LA	8/15/1943 0:00
22	NaN	NaN	LIGHT	LA	8/15/1943 0:00
204	NaN	NaN	DISK	CA	7/15/1952 12:30
241	NaN	BLUE	DISK	MT	7/4/1953 14:00
613	NaN	NaN	DISK	NV	7/1/1960 12:00
1877	NaN	YELLOW	CIRCLE	AZ	8/15/1969 1:00
2013	NaN	NaN	NaN	NH	8/1/1970 9:30
2546	NaN	NaN	FIREBALL	OH	10/25/1973 23:30
3123	NaN	RED	TRIANGLE	WV	11/25/1975 23:00
4736	NaN	NaN	SPHERE	CA	6/23/1982 23:00
5269	NaN	NaN	NaN	AZ	6/30/1985 21:30
6735	NaN	NaN	FORMATION	TX	4/1/1992 2:00
7208	NaN	NaN	CIRCLE	MI	10/4/1993 17:30
8828	NaN	NaN	TRIANGLE	WA	10/30/1995 21:30
8967	NaN	NaN	VARIOUS	CA	12/8/1995 18:00
9273	NaN	NaN	TRIANGLE	OH	5/1/1996 3:00
9388	NaN	NaN	OVAL	CA	6/12/1996 12:00
9587	NaN	NaN	EGG	FL	8/24/1996 15:00
10399	NaN	NaN	TRIANGLE	IL	6/15/1997 23:00
11625	NaN	NaN	CIRCLE	TX	6/7/1998 7:00
12441	NaN	RED	FIREBALL	WA	10/26/1998 17:58
15767	NaN	NaN	RECTANGLE	NV	1/21/2000 11:30
15812	NaN	NaN	LIGHT	NV	2/2/2000 3:00
16054	NaN	GREEN	NaN	FL	3/11/2000 3:30
16608	NaN	NaN	SPHERE	NY	6/15/2000 15:00

```
In [11]: ufo.shape

Out[11]: (18241, 5)

In [12]: ufo.dropna(how='any').shape # drop any rows with NaN in any of five columns
         # there is an inplace that is set to false, so does not change original data

Out[12]: (2486, 5)

In [13]: # only drop rows with NaN for all columns
         ufo.dropna(how='all').shape

Out[13]: (18241, 5)

In [14]: # drop row if either City or Shape Reported are missing
         ufo.dropna(subset=['City', 'Shape Reported'], how='any').shape
```

```
Out[14]: (15576, 5)

In [15]: # drop row if BOTH City or Shape Reported are missing
         ufo.dropna(subset=['City', 'Shape Reported'], how='all').shape

Out[15]: (18237, 5)

In [17]: # filling missing values
         # how many times each occurrence
         # by default, missing values are excluded
         ufo['Shape Reported'].value_counts().head()

Out[17]: LIGHT       2803
         DISK        2122
         TRIANGLE    1889
         OTHER       1402
         CIRCLE      1365
         Name: Shape Reported, dtype: int64

In [18]: # missing values are included => NaN = 2644
         ufo['Shape Reported'].value_counts(dropna=False).head()

Out[18]: LIGHT       2803
         NaN         2644
         DISK        2122
         TRIANGLE    1889
         OTHER       1402
         Name: Shape Reported, dtype: int64

In [20]: ufo['Shape Reported'].fillna(value='VARIOUS', inplace=True)

In [21]: # look at the head of the Shape Reported => all the NaNs in Shape Reported
         # converted to VARIOUS
         ufo['Shape Reported'].value_counts(dropna=False).head()

Out[21]: VARIOUS     2977
         LIGHT       2803
         DISK        2122
         TRIANGLE    1889
         OTHER       1402
         Name: Shape Reported, dtype: int64

In [ ]:
```