Loading A CSV Into Pandas

01 May 2016 / Python / Data Wrangling

Want to learn more? I recommend these Python books: Python for Data Analysis (http://amzn.to/2ljV9wY), Python Data Science Handbook (http://amzn.to/2m0mgMB), and Introduction to Machine Learning with Python (http://amzn.to/2mjYiwK).

import modules

```
import pandas as pd
import numpy as np
```

Create dataframe (that we will be importing)

```
raw_data = {'first_name': ['Jason', 'Molly', 'Tina', 'Jake', 'Amy'],
        'last_name': ['Miller', 'Jacobson', ".", 'Milner', 'Cooze'],
        'age': [42, 52, 36, 24, 73],
        'preTestScore': [4, 24, 31, ".", "."],
        'postTestScore': ["25,000", "94,000", 57, 62, 70]}
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name', 'age', 'preTestScore', 'postTestScore'])
df
```

	first_name	last_name	age	preTestScore	postTestScore
0	Jason	Miller	42	4	25,000
1	Molly	Jacobson	52	24	94,000
2	Tina	.	36	31	57
3	Jake	Milner	24	.	62
4	Amy	Cooze	73	.	70

Save dataframe as csv in the working director

```
df.to_csv('../data/example.csv')
```

Load a csv

```
df = pd.read_csv('../data/example.csv')
df
```

	Unnamed: 0	first_name	last_name	age	preTestScore	postTestScore
0	0	Jason	Miller	42	4	25,000
1	1	Molly	Jacobson	52	24	94,000
2	2	Tina	.	36	31	57
3	3	Jake	Milner	24	.	62
4	4	Amy	Cooze	73	.	70

Load a csv with no headers

```
df = pd.read_csv('../data/example.csv', header=None)
df
```

	0	1	2	3	4	5
0	NaN	first_name	last_name	age	preTestScore	postTestScore
1	0.0	Jason	Miller	42	4	25,000
2	1.0	Molly	Jacobson	52	24	94,000
3	2.0	Tina	.	36	31	57
4	3.0	Jake	Milner	24	.	62
5	4.0	Amy	Cooze	73	.	70

Load a csv while specifying column names


```
df = pd.read_csv('../data/example.csv', names=['UID', 'First Name', 'Last Name', 'Age', 'Pre-Test Score', 'Post-Test S
df
```

	UID	First Name	Last Name	Age	Pre-Test Score	Post-Test Score
0	NaN	first_name	last_name	age	preTestScore	postTestScore
1	0.0	Jason	Miller	42	4	25,000
2	1.0	Molly	Jacobson	52	24	94,000
3	2.0	Tina	.	36	31	57
4	3.0	Jake	Milner	24	.	62
5	4.0	Amy	Cooze	73	.	70

Load a csv with setting the index column to UID

```
df = pd.read_csv('../data/example.csv', index_col='UID', names=['UID', 'First Name', 'Last Name', 'Age', 'Pre-Test Scc
df
```

	First Name	Last Name	Age	Pre-Test Score	Post-Test Score
UID					
NaN	first_name	last_name	age	preTestScore	postTestScore
0.0	Jason	Miller	42	4	25,000
1.0	Molly	Jacobson	52	24	94,000
2.0	Tina	.	36	31	57
3.0	Jake	Milner	24	.	62
4.0	Amy	Cooze	73	.	70

Load a csv while setting the index columns to First Name and Last Name

```
df = pd.read_csv('../data/example.csv', index_col=['First Name', 'Last Name'], names=['UID', 'First Name', 'Last Name'
df
```

		UID	Age	Pre-Test Score	Post-Test Score
First Name	Last Name				
first_name	last_name	NaN	age	preTestScore	postTestScore
Jason	Miller	0.0	42	4	25,000
Molly	Jacobson	1.0	52	24	94,000
Tina	.	2.0	36	31	57
Jake	Milner	3.0	24	.	62
Amy	Cooze	4.0	73	.	70

Load a csv while specifying "." as missing values

```
df = pd.read_csv('../data/example.csv', na_values=['.'])
pd.isnull(df)
```

	Unnamed: 0	first_name	last_name	age	preTestScore	postTestScore
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	False	False	False
3	False	False	False	False	True	False
4	False	False	False	False	True	False

Load a csv while specifying "." and "NA" as missing values in the Last Name column and "." as missing values in Pre-Test Score column

```
sentinels = {'Last Name': ['.', 'NA'], 'Pre-Test Score': ['.']}
```

```
df = pd.read_csv('../data/example.csv', na_values=sentinels)
df
```

	Unnamed: 0	first_name	last_name	age	preTestScore	postTestScore
0	0	Jason	Miller	42	4	25,000
1	1	Molly	Jacobson	52	24	94,000
2	2	Tina	.	36	31	57
3	3	Jake	Milner	24	.	62
4	4	Amy	Cooze	73	.	70

Load a csv while skipping the top 3 rows

```
df = pd.read_csv('../data/example.csv', na_values=sentinels, skiprows=3)
df
```

	2	Tina	.	36	31	57
0	3	Jake	Milner	24	.	62
1	4	Amy	Cooze	73	.	70

Load a csv while interpreting "," in strings around numbers as thousands seperators

```
df = pd.read_csv('../data/example.csv', thousands=',')
df
```

	Unnamed: 0	first_name	last_name	age	preTestScore	postTestScore
0	0	Jason	Miller	42	4	25000
1	1	Molly	Jacobson	52	24	94000
2	2	Tina	.	36	31	57
3	3	Jake	Milner	24	.	62
4	4	Amy	Cooze	73	.	70

Find an error or bug? Have a suggestion?

Everything on this site is avaliable on GitHub. Head on over and submit an issue. (https://github.com/chrisalbon/notes_on_data_science_machine_learning_and_artificial_intelligence/issues/new) You can also message me directly on Twitter (https://twitter.com/chrisalbon).