Exercise 3a: Practical Pandas

This is a quiz given in Roger Peng Coursera (https://www.coursera.org) class Computing for Data Analysis (https://www.coursera.org/course/compdata).

Sourced from Research Computing MeetUp's (https://github.com/ResearchComputing/Meetup-Fall-2013) Python course.

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

data = pd.read_csv(os.path.join('data', 'ozone.csv'))
```

In [2]:
```
print data.head()
```

```
   Ozone  Solar.R  Wind  Temp  Month  Day
0     41      190   7.4    67      5    1
1     36      118   8.0    72      5    2
2     12      149  12.6    74      5    3
3     18      313  11.5    62      5    4
4    NaN      NaN  14.3    56      5    5
```

Print the column names of the dataset to the screen, one column name per line.

In [ ]:
```

```

Extract the first 2 rows of the data frame and print them to the console. What does the output look like?

In [ ]:
```

```

How many observations (i.e. rows) are in this data frame?

In [ ]:
```

```

Extract the last 2 rows of the data frame and print them to the console. What does the output look like?

In [ ]:
```

```

What is the value of Ozone in the 47th row?

In [ ]:
```

```

How many missing values are in the Ozone column of this data frame?

In [ ]:
```

```

What is the mean of the Ozone column in this dataset? Exclude missing values (coded as NA) from this calculation.

In [ ]:

Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?

In [ ]:

What is the mean of "Temp" when "Month" is equal to 6?

In [ ]:

What was the maximum ozone value in the month of May (i.e. Month = 5)?

In [ ]:

In [ ]:

Exercise 3b: Functions with Pandas

Kaggle has a nice challenge based on titanic passenger data:

"One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class."

We'll use the 892 line training set from Kaggle (http://www.kaggle.com/c/titanic-gettingStarted/data), located in the data directory of this lesson.

In [3]:
```
import os
import pandas as pd
```

In [30]:
```
a=pd.read_csv('./data/titanic.csv')
```

Read the data

- Read with `pandas.read_csv`
- Verify that you get `891` lines

What percent of the people survied?

- Yes, they survived: `d.survived == 1`

In [ ]:

Function

Write a function that returns a dictionary of the number of `survived`, `not survived`, and `unknown`. Here is the example function call:

```
print titanic_function(data)

{'survived': 342, 'not survived': 549}
```

In [2]:

What percent of males survived? Females?

In []:

Next Steps

Recommended Resources

Name	Description
Official Pandas Tutorials (http://pandas.pydata.org/pandas-docs/stable/tutorials.html)	Wes & Company's selection of tutorials and lectures
Julia Evans Pandas Cookbook (https://github.com/jvns/pandas-cookbook)	Great resource with eamples from weather, bikes and 311 calls
Learn Pandas Tutorials (https://bitbucket.org/hrojas/learn-pandas)	A great series of Pandas tutorials from Dave Rojas
Research Computing Python Data PYNBs (https://github.com/ResearchComputing/Meetup-Fall-2013/tree/master/python)	A super awesome set of python notebooks from a meetup-based course exclusively devoted to pandas