HOMEWORK 1 ASSIGNMENT

Assigned: Sat, 5/31
Deadline: Sat, 5/31 by 11:59PM
(If you need to ask questions before/after class to complete the assignment)

The purpose of this first homework is to get us all exploring data using iPython notebook and pandas/numpy. Because this is the first assignment, "grading" will be simply Attempted/Not Attempted. The only way to "fail" this one is to not try.

In this assignment, we will gain practice using:

- iPython notebook
- RelevantPythonpackages

DATA & CONTEXT

In this assignment, we will explore the passenger list of the Titanic, as provided in a well-known Kaggle competition. For this assignment, we are concerned only with initial exploration. We may build a predictive model later, but not now. The focus of the assignment is to answer the specific questions listed below in the section "Homework Questions."

The dataset is a list of passengers. The second column of the dataset is a "label" for each person indicating whether that person survived (1) or did not survive (0). Here is the Kaggle page with more information on the dataset:

http://www.kaggle.com/c/titanic-gettingStarted/data

Don't worry about downloading the data from Kaggle; we have provided the HW1 dataset for you. To download the dataset:

1. Navigate to this page (yes, it should look like raw text):
   http://goo.gl/huccEr

2. In your browser, select File | Save Page As...
3. In the save dialog that appears, click the Save button. You should now have a file named titanic-train.csv on your computer wherever you save downloads.
4. Move this data file to the directory where you like to do your iPython notebook work and dig in!

Alternatively, you can download ALL of our class materials to date as well as the titanic dataset by downloading a zip file of our class repo so far. To do that, navigate to the following page and click the "Download ZIP" button in the lower right-hand corner:

https://github.com/mike-tamir/GA_DAT7

*Note the GA_DAT7 Git site will be updated as the course progresses.

SUBMITTING YOUR WORK

Please do all your analysis to answer the questions below in an iPython notebook. Include your thinking in the notebook, and show your work as much as possible.

Because we did not get all the way through the Git & Github lab yet, please submit your assignments through Schoology:

https://www.schoology.com/course/113605631

Please zip/compress your iPython notebooks before submitting in order to avoid issues with Schoology.

HOMEWORK QUESTIONS

Please answer the following questions about your data exploration in the iPython notebook. Feel free to explore further.

1. How many passengers are in our passenger list? From here forward, we'll assume our dataset represents the full passenger list for the Titanic.
2. What is the overall survival rate?
3. How many male passengers were onboard?
4. How many female passengers were onboard?
5. What is the overall survival rate of male passengers?
6. What is the overall survival rate of female passengers?
7. What is the average age of all passengers onboard?
    a. How did you calculate this average age?
    b. Did you encounter any problems with this calculation?
    c. If so, how did you address any problems?
8. What is the average age of passengers who survived?
9. What is the average age of passengers who did not survive?
10. At this (early) point in our analysis, what might you infer about any patterns you are seeing?
11. How many passengers are in each of the three classes of service (e.g. First, Second, Third?)
12. What is the survival rate for passengers in each of the three classes of service?
13. Continuing with 10, what else might you conclude?
14. Last, if we were to build a predictive model, which features in the data do you think we should include in the model and which can we leave out? Why?