MIDTERM ASSIGNMENT – DAT8

Deadline to Submit via Schoology: Tuesday, August 20 by 6:30PM

The purpose of this assignment is to get us exploring data using iPython notebook and pandas/numpy.

In this assignment, we will gain practice using:

- iPython notebook
- Relevant Python packages

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 as part of this assignment. 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; the dataset is already in this repo in the midterm/data folder directory.

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.

Please submit your assignments through Schoology. Please zip/compress your iPython notebooks before submitting in order to avoid issues with Schoology not accepting files with “unusual” filename extensions.

Notebooks should be named according to the following standard:

DAT8_midterm_FIRSTNAME_LASTNAME

QUESTIONS

Please answer the following questions about your data exploration in the iPython notebook. Feel free to explore further. These questions are a guide and a minimum, not a limit ;-)

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. 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?