How do I Complete this Project?

This project is connected with the Introduction to Data Analysis course, but depending on your background knowledge, you may not need to take the whole class to complete this project.

Introduction

For the final project, you will conduct your own data analysis and create a file to share that documents your findings. You should start by taking a look at your dataset and brainstorming what questions you could answer using it. Then you should use Pandas and NumPy to answer the questions you are most interested in, and create a report sharing the answers. You will not be required to use statistics or machine learning to complete this project, but you should make it clear in your communications that your findings are tentative. This project is open-ended in that we are not looking for one right answer.

Step One - Choose Your Data Set

Choose one of the following datasets to analyze for your project:

- Titanic Data - Contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic. You can view a description of this dataset on the Kaggle website, where the data was obtained.
- Baseball Data - A data set containing complete batting and pitching statistics from 1871 to 2014, plus fielding statistics, standings, team stats, managerial records, post-season data, and more. This dataset contains many files, but you can choose to analyze only the one(s) you are most interested in.

  Choose the comma-delimited version, which contains CSV files.

Step Two - Get Organized

Eventually you'll want to submit your project (and share it with friends, family, and employers). Get organized before you begin. We recommend creating a single folder that will eventually contain:

- The **report** communicating your findings
- Any **Python code** you wrote as part of your analysis
- The **data set** you used (which you will not need to submit)

You may wish to use IPython notebook, in which case you can submit both the code you wrote and the report of your findings in the same document. Otherwise, you will need to submit your report and code separately.

Step Three - Analyze Your Data

Brainstorm some questions you could answer using the data set you chose, then start answering those questions. Here are some ideas to get you started:

- Titanic Data
    - What factors made people more likely to survive?
- Baseball Data
    - What is the relationship between different performance metrics? Do any have a strong negative or positive relationship?
    - What are the characteristics of baseball players with the highest salaries?

Try and suggest questions that promote looking at relationships between multiple variables. You should aim to analyze at least one dependent variable and three independent variables in your investigation. Make sure you use NumPy and Pandas where they are appropriate!

Step Four - Share Your Findings

Once you have finished analyzing the data, create a report that shares the findings you found most interesting. You might wish to use IPython notebook to share your findings alongside the code you used to perform the analysis, but you can also use another tool if you wish.

Step Five - Review

Use the **Project Rubric** to review your project. If you are happy with your submission, then you're ready to submit your project. If you see room for improvement, keep working to improve your project.