Assignment 2: Data Exploration (10 points)

In this assignment we will work with a (fake) dataset encompassing people's age, number of steps taken and income. The main purpose of this assignment relates to data wrangling / data cleaning - but we will also follow through on some basic analysis and predictions with this data.

Tasks / Learning Goals

Hands-on experience with:

- Data Wrangling: Connecting together data from
- Data Visualization: Looking at data to check for problems
- Data Cleaning: Removing 'bad' data
- Basic Analysis & Prediction: Looking at how different variables relate to each other

Due Date

11:59 pm Tuesday, April 25th, submitted on TritonED.

Submitting Assignments

You will submit a Jupyter notebook file (.ipynb) to TritonED. Make sure that the file you submit has the following filename (filled in with your student ID number):

'A2_A#########.ipynb'

Grading Rubric

There are 4 parts to this assignment, with the following point values:

Part 1: Loading & Combining	2 points
Part 2: Data Cleaning	5 points
Part 3: Data Analysis	1 points
Part 4: Predictions	2 points

Detailed Instructions

These instructions outline what you will be doing in the assignment, and suggest things to explore with the data. Step by step, explicit instructions are written in the notebook - and you are graded based on what is specified in the notebook - these here are merely guidelines.

Part 1 - Data Wrangling

You have been given two files: one JSON and the other CSV. The JSON has four variables, though you only need two: ID and income. The CSV has three variables; you'll need all three.

Main tasks:

- Import both into pandas dataframes.
- Join these two datasets by ID.

Part 2 - Data Cleaning

Before analysis, data must be cleaned of missing or unusable data. Once the datasets are joined into a single dataframe, your next step is to remove garbage data and then sub-select based on analysis needs.

- Income has missing data (NaNs). You need to remove these.
- Plot histograms (using matplotlib - see the linked example) to check distributions of the 3 variables of interest (age, steps, income)
    - http://matplotlib.org/1.2.1/examples/pylab_examples/histogram_demo.html
    - What do you notice, especially in the steps data?
        - Age and steps are relatively normally (Gaussian) distributed, while income is not.
- Since income is not normally distributed, we're going to log transform it, so that it becomes more normal. Transform income using a log10 transform and plot its histogram again. Now what do you notice? What's happening at $0.00?
- Note that the histogram for steps show a lot of values equal to -1, while income shows a lot equal to $0.00. In this dataset, -1 steps per day is an impossible value, and here means the data are missing so much be removed, while $0.00 is a real income, meaning unemployed.
- Remove all rows where steps equal is equal to -1 (an impossible value),
- Remove all rows and where age is less than 18 (pretend regulatory requirements).
- Once the data are cleaned, you should also perform outlier removal. This is part of the art of data analysis!
    - Remove all rows where the steps data is more than +/- 3 standard deviations.
    - Note that **when** you perform outlier removal--before or after removing the -1 values from steps, can have a **big** effect on the results!
    - For the steps data, what are the -3 and +3 std cutoff points for each of the three variables? What do these numbers look like if you apply this criterion to steps **before** getting rid of the -1 values?

Part 3 - Basic Data Analysis
Now we're going to explore some basic relationships between these three variables.

Main Tasks:

- Look at how the three variables correlate with one another (Pearson correlations)
- Look at their plots using pandas.scatter_matrix()

Part 4 - Predictions
Let's try some basic predictions using linear models.

Main Tasks:

- Fit some simple linear models, predicting income from age and/or steps