CSE 91, Fall 2017
Project 1, Part 1
Due: Thursday, October 12, 11:59pm

Part 1. Data Wrangling: Loading, Parsing, and Combining

In this assignment you will be working with a (fake) dataset containing a person's age, number of steps taken by that person and their income. The main purpose of part one is to get experience with data wrangling. We will use the *datascience* Python package to load and manipulate the data.

We have provided a Jupyter notebook for you to work with. It has detailed, step by step instructions on what to do and hints about what tools you should use. Be prepared to look up examples online and in the Python documentation. A large part of this project is about figuring out how to do things that you've never done before.

Part 1.1. Data Loading

Some formats can be loaded into a table nicely (like csv), others require some additional work (like JSON). For example, if you load JSON directly into a table, you will get this:


```
[{"id":22453,"last_name":"Adams","first_name":"Jason","income":13024.21},
 {"id":49241,"last_name":"Adams","first_name":"Paula","income":61877.51}]
```


It is hard to understand and simply incorrect. There is no header in a JSON file, so the first person becomes the header. What we want is a clean, formatted table that makes sense to us:

ID	LAST	FIRST	INCOME
22453	Adams	Jason	13024.21
49241	Adams	Paula	61877.51

Part 1.2. Data Parsing

You need will need to *parse* the data before it can be analyzed. You will need to:

- Drop unnecessary formatting information like braces and colons, keeping only relevant data.
- Create a new table with meaningful labels and add data there.
    - Make sure every row gets added.
- Drop unneeded columns.

Part 1.3. Combining Tables

Data often comes from different sources. In our project we have two files: csv and json. They have something common: ID. Therefore instead of having two separate tables, you will join them by the shared field, then reorder and relabel the columns for consistency and readability.

Submission Details

For this project, you may work alone or with one other person in either section of CSE 91, using the pair programming approach.

For help with your projects, instructors are available during the scheduled discussion section times on Thursdays.

To start working, log in to Jupyterhub and then click the assignment link available on the course website and Piazza.

Work through the notebook, and if you need to stop, submit the file, as only the most recent submission will be graded.

To submit your file, run the cell

```
_ = ok.submit()
```

This project is due at 11:59pm on Thursday, October 12.

Part 2. Data Cleaning

Data cleaning involves ensuring that all the data you will use in your analysis is of the proper format, and that all the data is meaningful. The main purpose of part two is to get experience with data cleaning.

Submission: TBA