**Lab #4 - Data Wrangling**

The notebook should be well organized. Each section should be **clearly labeled with the exercise (and part) that it addresses** (e.g., Exercise #1a, #1b, #2) in a Markdown cell block. Use (clear and concise) comments as needed to help describe each step of your process. All notebook cells that contain essential steps should be executed and the output should be visible, so as to demonstrate your successful completion of the exercise. If you cannot complete an exercise in its entirety, you should make an effort to demonstrate your intermediate progress in order to maximize partial credit, and move forward as best as possible. You may submit any written answers to the exercises in the notebook as text cells.

**Academic Integrity**: Each group is expected to submit its own original work. You may consult with the instructor and the TAs for help, but collaboration with other groups should be kept to a minimum. Submissions that contain significant similarities will be reported directly to the Office of Student Conduct.

**Instructions/Background**

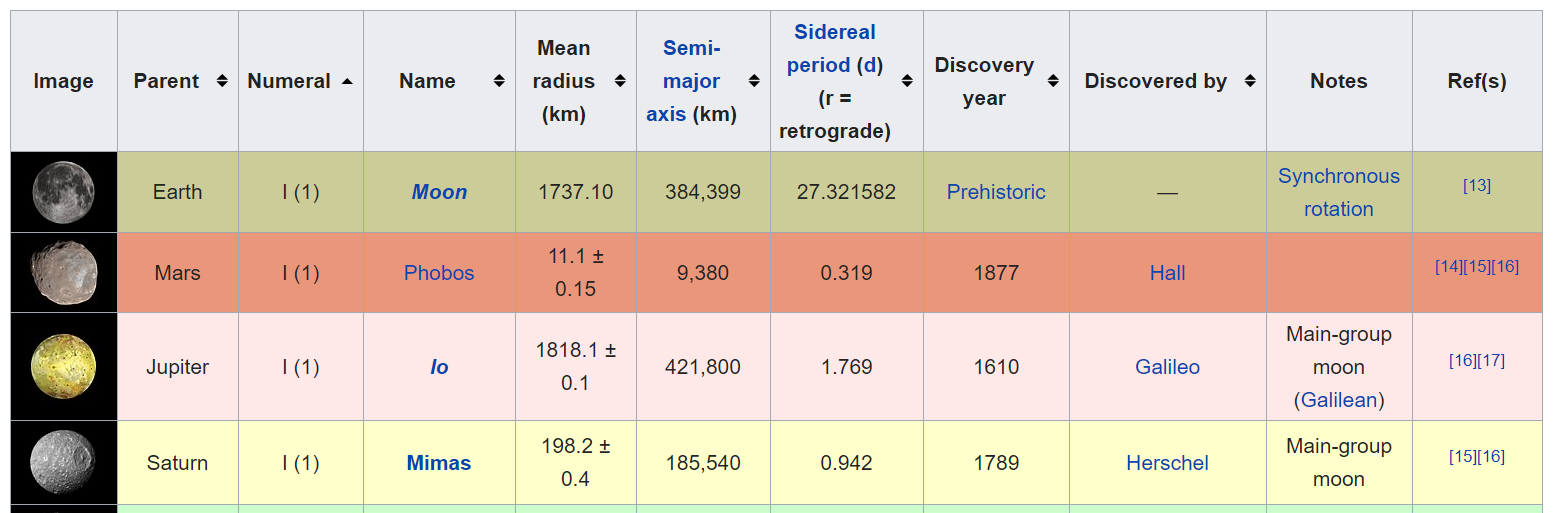
This lab is designed to independently test your understanding of importing, cleaning, summarizing, and analyzing data. There are fewer explicit steps provided than in previous labs.

**Exercise #1 - Data Import**

The following link contains a table of information about the moons in our solar system:

<https://en.wikipedia.org/wiki/List_of_natural_satellites>

We will be working with the following table. Note that there are several tables on the Wikipedia page.



**Step 1.1 Practice with BeautifulSoup**

**Step 1.1.a** Inspect the table shown above in your web browser. What is the class of the table tag that contains this table? Write your answer in a Markdown cell.

**Step 1.1.b** Create an HTTP request for the page and pass it to the BeautifulSoup parser. User .find\_all to return a list of all the tables that have the class you wrote above.  Save that list as a variable. How many tables are there with that class?

**Step 1.1.c** Check each of the list entries for the table we are interested in. What is the index of the table we are interested in within the list created with .find\_all? Show just that entry from the list; that is, the HTML contents that show our table of interest.

**Step 1.2 Import Using Pandas**

As an alternative to BeautifulSoup, we can import the data using Pandas. Use pd.read\_html() to read in the table. Use the "match" parameter to find only the table we are interested in. Show the first 10 rows of the DataFrame.

**Exercise #2 - Data Validation and Cleaning**

**Step 2.1** Perform four data cleaning/validation tasks.

* Rename/name the columns.
* Reorder the columns in a way that makes the table easy to read/understand.
* Check the data types of the columns.
* What columns identify each observation in the dataset? Set the index. (Note: You will need to make the index consist of two columns or create a new column that combines two columns.)

**Step 2.2 Additional data cleaning**

Choose at least one additional column of the data to clean. Here are some cleaning ideas, but you are not limited to these: convert the column's data type, extract bad characters, or create a dummy variable. It may be helpful to choose your research question for the next exercise and then return to this step to choose the column to clean based on your question.

**Exercise #3 - Analysis**

Given the tidy data that you've now processed, formulate a question about this data. Then, perform the appropriate analysis needed to address the question and describe your results. You will be evaluated on both the quality of your question and the quality of your analysis and description.