CSCI 4146 – Summer 2017 **Assignment 1**

- 10% of your final grade
- Due **Wednesday, 24 May** (**23:59pm;** before midnight)
- Work independently
- Cite any and all resources used
  - books, websites (other than documentation) like stackoverflow
  - I will google your code …
- Two sides of Data Science (and your mark):
  - Technical // does it work?
  - Conceptual // what did you find?
- Submit a Jupyter Notebook for the whole assignment: **`a1-[your_name].ipynb`**
  - By email (gerow@dal.ca) or upload on Brightspace

Marking: I will download your notebook file and run each cell in sequence. Make sure the cells run in order and that your answers are clearly marked in the code and output. You may use markup cells for explanations, but Python comments are fine. If you do any work outside the notebook to answer the questions, include any derived files with your submission (zip or tar.gz please) and be sure your answers to each question are **in the notebook**.

Q1: Data cleaning and handling

Data in the **us_shipping/** directory contains information about US, inter-state shipping. The **us_shipping.csv** is a contains data about *exports* among states. Make sure you understand the data before you start!

1.a - Import the data into a Pandas dataframe. Convert all numeric values to real (float) values and make sure you properly handle missing data.

1.b - Create a second dataframe with each state as a row, and each mode as a column, where each cell is the value of goods that originated from each state via by the given mode in 2012.

1.d - Write a function that, given this second dataframe and a specified state, returns a Python dictionary of each mode and the proportional value (.5 for 50%) of exports shipped by that mode in 2012. The values of this dictionary should sum to 1 and it should only have keys for modes with >= 0 value (ie. don’t represent modes with no exports). Show that this function works correctly in your notebook.

1.e - Using either dataframe, answer the following two questions: 1) Which modes had the biggest positive and negative differences in **tons-**shipped from 2007 to 2012? 2) Which state decreased the most in the export **value** across all modes from 2007 to 2012? Keep any code you write to answer these questions in your notebook, but also include the answers in the notebook itself.

Q2: Web scraping and Regular Expressions

In the file **trip_advisor/hotel.dat** there are a list of hotel reviews from TripAdvisor. Each entry contains a review, its date, the TripAdvisor username and several ratings.

Managers of the hotel where you work want to understand the demographics of their clients. To do this, we need to know the age range, sex and location of the users, but these features are not in the dataset. However, TripAdvisor makes user profiles public. These profiles can be accessed, as an example for the user "mal51", at:

http://www.tripadvisor.ca/members/mal51

Not all users make their information available, and in some cases profiles in the data file may have since been deleted. Using your knowledge of regular expressions and web-scraping, retrieve and extract the required information from the html source files. With the retrieved data ...

2.a - Create a Pandas dataframe for the scraped data with the following format:

Username	**Age range**	**Sex**	**Location**
Mal51	50-64	Female	United Kingdon
Lucatony	_Null_	_Null_	Ontario

...

and export it to a CSV file. Include this CSV with your final submission.

2.b - Prepare a second dataframe that combines information from the given dataset and the one you just created. This should group individuals by age range and sex with the average overall rating, its standard deviation, and the number of users in each group. This should be something like:

Age range	**Sex**	**Mean rating**	**s.d. rating**	**Group size**
18-24	Female	3.5	2.4	30
18-24	Male	4.0	1.5	20

...

Export this dataframe as a second CSV file and include it in your submission.

2.c - List some conclusions from this aggregation. Describe how the groups differ in terms of ratings.

2.d – Using matplotlib, make plots summarizing this dataframe. Use a different point-style (boxes, Xs, circles, etc.) for sex, put age-ranges on the x-axis and the mean rating on the y-axis.

> **Extra credit**: Add error-bars to each point of +/- 1 s.d. of the mean and / or make the size of the markers proportional to group size. These features may be hard to incorporate into a good plot – make sure the basic information is not obscured!

Keep in mind:

- If age range or location is not available, you should use a suitable null-value (NaN, None, etc.). Be careful with 0s, as they change the mean.

- For simplicity, you can use the **read_csv()** and **write_csv()** Pandas functions.

- To avoid using using regular expressions on the whole html file, use tags to find the correct region first.

- You can use the Python packages **urllib2** or **requests** to retrieve URLs files.