Homework 1

Assigned: 08/23/2018 | Due: 08/30/2018 at 12am

Your deliverables need to be organized and zipped as below, before submitting to Canvas. Failure to properly name/organize files will result in a 10% penalty.

LastName_FirstName_HW1.zip
- |--- complete_clean.csv
- |--- complete_clean_history.json
- |--- HW1.ipynb

1. OpenRefine Practice
   a. Download the UFO sighting database (ufo-sightings.zip), and extract the two csv files
   b. Launch OpenRefine and create a new Project by importing "complete.csv"
      Create Project → This Computer → (import file) → next → Create Project
   c. Split "datetime" column
      Click the dropdown menu of datetime → Edit column → Split into several columns → change separator to " "(space) → click OK
      The column should be split to two columns: "datetime 1" and "datetime 2"
   d. Rename the column "datetime 1" to "Date"
      Click the dropdown menu of datetime 1 → Edit Column → rename this column → Change to Date → Click OK
   e. Rename the column "datetime 2" to "Time"
   f. Filter out records that have blank values in the "shape" column
      Click the dropdown menu of shape → Facet → customized Facet → Facet by blank
   g. remove the records that have blank values in the "shape" column
      One the facet created in f, click true to show the records with blank shape value → click the dropdown menu of "All" (left most column) → Edit rows → remove all matching rows
   h. Keep records with value "us" in the "country" column
      Hint: Use "text facet"
   i. Change "us" to uppercase "US"
      Dropdown menu → Edit cells → common Transform → to uppercase
   j. Change values in "state" column to uppercase
   k. Remove "duration (hours/min)" column
   l. Change the column "duration (seconds)" to "duration"
   m. make text facet of the shape column
   n. change the cells with value "changed" to "changing"
      click on the edit botton of the entry in the facet panel → change name
   o. change cells with value "round" to "sphere"
   p. change cells with value "flare" to "flash"
   q. export the resulting spreadsheet as comma-separated valued (csv) file
      export → comma-separated value
      name it as "complete_clean.csv"
   r. extract and save your operation history to a json file "complete_clean_history.json"
      Go to the "Undo/Redo" tab → extract → select all steps → save the text to a json file "complete_clean_history.json"

   Deliverable: "complete_clean.csv" and "complete_clean_history.json"

2. Python (pandas, matplotlib) Practice

   Write your answers directly into the “HW1.ipynb” Jupyter notebook. We will use Jupyter notebooks to complete all Python assignments. The correct outputs are already given in “HW1.ipynb”, you just need to fill in the code that generates a similar output. This will allow you to compare and check your answers.

   a. Import pandas package and read the “complete_clean.csv” file into variable “data”
   b. Using Pandas, do the following:
      i. Calculate the average duration of sighting for all records
      ii. Calculate the average duration of sighting for each of the states
      iii. Calculate the occurrences of each of the shapes
      iv. Count all UFO sightings occurred in Georgia (state value equals to “GA”)
      v. Count all sighting in Texas (“TX”) that lasted shorter than 30 seconds or longer than 500 seconds
   c. Import matplotlib
   d. Plot the follow figures:
      i. Barplot: Number of sightings of different shapes
      ii. Barplot: Number of sightings in each state
      iii. Barplot: Average sighting duration of each state
      iv. Lineplot: number of sighting during each month
          Hint: index the dataframe by “Date”, and resample by month