Assignment 1

H1B Dataset

## Instruction

This homework is to be done individually. No collaboration permitted with current or past students.

## Objective

* In this Data Gathering Assignment you will download the H1B dataset in the datasets/h1b/<your\_NYU\_username>/ folder shared with the rest of the class, and you will create a documented iPython notebook with the data in a single database
* You will also create a README\_db.txt file that describes the features of the database
* You will produce summary statistics for the dataset and will present your results

## Data Description

The raw H1B datasets contain the record of every H1B application since 2002. The data can be obtained from the following locations:

* Data between 2002-2007 in .txt format:
  + Location: <http://www.flcdatacenter.com/CaseH1B.aspx>
  + Note that there are two datasets for each year: data obtained via the Fax system and data obtained via the Efile system. You will download both.
* Data between 2008-2018 in .xlsx format:
  + Location: <https://www.foreignlaborcert.doleta.gov/performancedata.cfm>
  + Click on the Disclosure Data tab to download the datasets
  + Download the H1B\_Record\_Layout\*.xlsx files (years 2014-18) along with the LCA\_\*.xlsx files (years 2008-14). Both types of files contain the H-1B visa program records.
  + In addition download the PERM\_\*.xlsx data.

## Methodology and Deliverables

### H1B dataset:

Download all the raw datasets into the datasets/h1b/<your\_NYU\_username>/ folder. Create a data\_parse.ipynb iPython notebook where you parse and standardize the 2008-2018 dataset using the Pandas library as follows:

For every year, upload all the H1B data (merge Efile and Fax data whenever necessary) and retain the following fields: date (of application), case\_number, employer, tot\_foreign\_positions (total position requested by employer), naic\_code (industry code of employer when available, or nan when missing), job\_title, job\_code (standard occupational code for job title), soc\_name, approval\_status, annual\_wage, annual\_prevailing\_wage, city (of occupation), state (of occupation), county (of occupation).

### NAIC and SOC codes dataset

Create a dataframe of all the NAIC industry codes and their names.

* The NAIC data can be scraped from here: <https://classcodes.com/naics-code-list/>
* The data can also be downloaded in in excel format from the census bureau: <https://www.census.gov/eos/www/naics/downloadables/downloadables.html>   
  However for the purpose of this exercise, please write a script to scrape the NAIC code definitions form the first link above. Do they all correspond to the same definitions as those in the Census file? Describe your findings.

Create a dataframe of all the SOC industry codes and their names.

* SOC data can be found in the Appendix of form 9035: <https://www.foreignlaborcert.doleta.gov/pdf/ETA_Form_9035CP_Appendix_I_121908.pdf>
* or can be scraped from the Bureau of Labor Statistics site: <https://www.bls.gov/oes/current/oes_stru.htm>

Save the two datasets into a naic.pkl and soc.pkl files.

### Analyze the data and present your results

Using the 2008-2018, answer and address the following questions:

* What are the trends in H1B employment since 2008? Has the average salary changed?
* What about the most popular positions, industries, locations? Which companies have increased hiring H1B employees over the years and which ones have decreased? Pay particular attention to 2008-2009 and compare this period to recent data.
* What job recommendations would you give to your classmates given your analysis of the data?
* Have all your interpretations and analysis in a notebook named h1b.ipynb.
* Be prepared to talk about your results if asked in a class.

### README file

Create a README\_db.txt file containing:

* Description of the h1b\_all.pkl, dataset and all of its fields, as well the naic.pkl and soc.pkl datasets
* Any Notes that you believe a future user the datasets should be aware of

### Extra credit

For extra credit, you will merge the 2002-2007 and 2008-2018 dataset, creating a single Pandas and pickle the dataframes into a h1b\_all.pkl using the pd.DataFrame.to\_pickle() method.

* What additional job trends were prevalent in the 2002-2003 and 2006-2007 periods? How do these compare to 2008-2018?
* Document your results in h1b.ipynb as well as in data\_parse.ipynb so that the grader can understand how you have performed the data cleaning and analysis and follow your work.