Data Bootcamp Problem Set 4 Spring 2020

NYU STERN

Instructions:

- **After completing the assignment, please submit your .ipnyb file to NYU Classes with the following naming convention: Lastname_Firstname_NetID_ProblemSet# (ex. Smith_John_js123_ProblemSet4)**
- Submit your answers in a **Jupyter notebook** with **proper markdowns to indicate problem numbers.**
- **Write the questions in markdown before you provide your answers.**
- See Grading Guidelines under Announcements on NYU Classes.
- In-line comments are helpful but not mandatory.
- Read the given CSV(master_data_ready.csv) to a data frame 'jobs'. Run command
    - jobs = jobs.replace('\\N',np.nan)
- **Do not change this data frame 'jobs' throughout the assignment unless asked to.**
- No explanations are expected at the end of answers, unless requested.
- When displaying dataframes and series, **do not use the 'print()' function**. Display it the way you were taught in class(simply type in the name of the dataframe variable or the .loc/.iloc command line and run the cell)
- You might find the reset_index() and the set_index() functions useful.

Problems:

1. Using the data frame jobs, do the following:

    a) Make job_id the index and display the first 10 rows

    b) From jobs create a new data frame 'jobs_1' by filtering for states Iowa and Oregon

    c) From jobs_1 create a new data frame 'jobs_2' by filtering for cities Cedar Rapids, Des Moines, Portland and Salem

    d) From jobs_2 create a new data frame 'jobs_2i' where state and city are set as indices of jobs_2. Using an appropriate property of data frames, display the MultiIndex object of jobs_2i. What do the parameters 'levels' and 'names' tell you?

    e) Using an appropriate method, push the index 'city' (but not 'state') back to being a column of jobs_2i and display first 5 rows

2. Using the data frame jobs, create a new data frame 'jobs_3' by filtering for the states mentioned in 1.b., for the cities mentioned in 1.c., and additionally for the companies 'Burlington Stores' and 'Info Group NW'. Now do the following:

    a) Update jobs_3 to have state, city and company as its indices. From the updated jobs_3 create a new data frame that has columns and rows switched. Display the new data frame.

    b) Explain (very briefly) what is happening in 2.a. and why doing this might be useful

3. Add a new column 'year' to jobs data frame. This column should contain the corresponding year for each post_date. For example: for post_date value 2017-08-16 'year' value should be 2017.
   Now create a new data frame annual_stats by running the code:
   annual_stats = jobs.groupby(['year','state']).mean()
   Now do the following:

   a) Update annual_stats using an appropriate method which would bring year, state and salary to the same index level. Display annual_stats
   b) Update annual_stats by filtering it for states NY, CA and TX and round the salary to 2 decimal places.
   c) Use an appropriate shaping technique on annual_stats to display salaries of each state for each year, like below representation. Display the data frame.

	2016	2017	2018
NY			
CA			
TX			

   d) Plot a horizontal bar chart from the data frame created in 3.c. for 2016 showing mean salaries of each state. Set appropriate title and x label.