Data Bootcamp Problem Set 3 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_ProblemSet2)**
- Submit your answers in a **Jupyter notebook** with **proper markdowns to indicate problem numbers**. **Please include the actual question from this document in a markdown before you display your answers.**
- In-line comments are helpful but not mandatory.
- Evaluation will be done mostly by verifying number of rows in your result.
- In-line comments are preferred for this assignment but not mandatory.
- No explanations are expected at the end of answers, but please **include the code** that you wrote to obtain the answers and **display the output.**

Problems:

1. Use read_csv() to read the responses of a previous semesters entry poll from this source : https://raw.githubusercontent.com/NYUDataBootcamp/Materials/master/Data/entry_poll_fall16.csv

a) Read the attached file "master_data_ready.csv" and assign it to the variable 'ep'.

b) How many variables and responses are there? What are these variables and responses in terms of the dataset?

c) Display the datatypes of the variables.

d) Change the variable names to something shorter.

e) Describe what the following code does: **ep[list(ep)[1]].value_counts()**. Suggestion: Break it into two or more statements and explain them one at a time.

2. Read the data set to your notebook. Modify the dataset to replace \N values everywhere in the dataset with NaN. Capitalize the first letter of each column name and assign the new names to the columns of the dataset.
3. Using an appropriate pandas method present a column wise summary of nulls (or NaN).
4. Drop columns "latitude" and "longitude" from the modified data set and save to a new data frame. Let's call it Master_Clean.
5. Using Master_Clean: Display jobs for the state "NY" that were posted in 2018.
6. Using Master_Clean: Display jobs from all states other than the states "NY" and "CA" that offer a salary higher than 80,000. Do **NOT** include jobs that have NaN values for state column.
7. Using Master_Clean: How many jobs have NaN in the Time_to_fill column? Replace all these nulls (or NaN) with the median of this column. Display first 25 and last 25 entries.
8. Using Master_Clean: Display only company, posting date and salary of jobs that are listed for the state "NJ". After filtering, exclude NaN from all columns.
9. Using Master_Clean: Use the .isin() method to display company, posting date, time to fill and salaries of jobs in the states "NC", "PA" or "TX". After filtering, exclude NaN from all columns.

10. Using Master_Clean: Use the contains method to display city, state, county and regions of jobs that have “New York” or “Los Angeles-Long” in their region state value.
11. Using Master_Clean: Display company, vertical, posting date and salaries for all jobs posted during the years 2017 and 2018 in “New York” or “Seattle” region states that had salaries higher than 80,000. Exclude NaN from company column.
12. Using Master_Clean: Calculate the percentage of jobs offering a six-figure salary or higher in the state “NY” in 2018.
13. Using Master_Clean: Calculate the percentage of job postings for the company “Wells Fargo” where the salary offered is above the mean of all salaries in the states of “NY” or “CA”.