Lab #4
Data Science I

Before you begin this lab you will need to read pages 146-158 of Python Data Science Handbook https://jakevdp.github.io/PythonDataScienceHandbook/03.07-merge-and-join.html. In this excerpt you will learn about 'pd.merge' which we will use to complete this Lab. For this lab we will use 'listings.csv' and 'reviews.csv' to investigate data for AirBnB hosts in Boston. The data was accessed from http://insideairbnb.com/get-the-data.html on July 16, 2019. Please answer the questions below using the dataset. For all plots, make sure your plot has (1) labeled x and y axis, (2) a title, and (3) a legend when appropriate.

1. Read in the file 'listings' that contains the listings for the Boston AirBnBs. How many entries are in the file? How many unique identifiers are there for AirBnB listings? Hint: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.nunique.html
2. Read in the 'reviews' file that contains the date of reviews for the Boston AirBnB listings. How many entries are in the file? How many unique identifiers are there for AirBnB listings? Do all of the listings have a corresponding review?
3. Merge the review and listing files. Do a merge using the function 'pd.merge' so that all of the entries from listing are in the merged data frame and have an NaN if there is no corresponding entry in reviews. How many rows do you now have? How many unique AirBnB ids do you have? How many rows were filled in with NaN?
4. Merge the review and listing files again using 'pd.merge'. This time do a merge so that all of the entries from reviews are in the dataframe (and therefore any entries from listings that do not have a review are dropped). How many rows do you now have? How many unique AirBnB ids do you have?
5. Accoring to AirBnB superhosts 'are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests'. How many unique AirBnBs are hosted by a 'superhost' in Boston? How many superhosts are there in Boston? How many super hosts do not have any reviews in the 'reviews' dataset? What is the median number of reviews that a superhost has in the 'reviews' dataset?
6. Make boxplots to compare the number of reviews of superhosts to non-superhosts. Hint: the boxplots may look better if you log the counts! Add 1 to the number of counts so that you can log values of 0.

Rubric:

Code Style (5 points) Is code organized well and commented?

Submission (5 points) Was the lab submitted as an html document on Canvas? Does the html document contain a link for a GitHub repository that contains your code?

Participation (5 points) When the instructor or TA came into the breakout room were you working together and screen sharing.

Correctness (35 points, ~ 6 points per questions)