CSE 351 - Introduction to Data Science (Spring 2022)
Assignment 1 (Homework1): Exploratory Data Analysis

Due Date: Assignment1 is due by 11:59 PM EST on Sunday, March 06, 2022

This is an individual assignment which should be done independently by each student. This homework will investigate doing exploratory data analysis in iPython. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis, by working with datasets where you have some basic sense of familiarity.

Python Installation

Instead of installing python and other tools manually, we suggest installing Anaconda, which is a Python distribution with package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found here. Installation instructions can be found here.

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Packages that I believe you may use for this homework include:

- pandas
- NumPy
- matplotlib
- seaborn

The packages above are very well documented and can be found online.

Another modern alternative is Google Colaboratory. This is another option for those who want to run their Jupyter notebook remotely instead of installing the required packages locally.

Data

The assignment is based on New York City Airbnb Open Data [2]. The main task is to mine the data and uncover interesting observation about the different hosts and areas. You will need to submit your code files in three different formats (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you needed to produce the resulting tables and figures. The submission steps have been discussed below. [Data also available along with the assignment]

This assignment is worth 70 points.

Tasks

1. Examine the data, there may be some anomalies in the data, and you will have to clean the data before you move forward to other tasks. Explain what you did to clean the data. (**10 Points**)

2. Examine how the prices of the Airbnb changes with the change in the neighborhood.

a. Find Top 5 and Bottom 5 neighborhood based on the price of the Airbnb in that neighborhood (select only neighborhoods with more than 5 listings). (**10 Points**)
b. Analyze, the price variation between different neighborhood group, and plot these trends. (**5 Points**)

3. Select a set of the most interesting features. Do a pairwise Pearson correlation analysis on all pairs of these variables. Show the result with a heat map and find out most positive and negative correlations. (**5 points**)

4. The Latitude and Longitude of all the Airbnb listings are provided in the dataset.
   a. Plot a scatter plot based on these coordinates, where the points represent the location of an Airbnb, and the points are color coded based on the neighborhood group feature. (**5 Points**)
   b. Now again, plot a scatter plot based on these coordinates, where the points represent the location of an Airbnb, and the points are color coded based on the price of the particular Airbnb, where price of the listing is less than 1000. Looking at the graph can you tell which neighborhood group is the most expensive. (**5 Points**)

5. Word clouds are useful tool to explore the text data. Extract the words from the name of the Airbnb and generate a word cloud. (**5 Points**)

6. Find out which areas has the busiest (hosts with high number of listings) host? Are there any reasons, why these hosts are the busiest, considers factors such as availability, price, review, etc.? Bolster you reasoning with different plots and correlations. (**10 Points**)

7. Create two plots (at least one unique plot not used above) of your own using the dataset that you think reveals something very interesting. Explain what it is, and anything else you learned. (**10 Points**)

8. Visual Appeal and Layout - For all the tasks above, please include an explanation wherever asked and make sure that your procedure is documented (suitable comments) as well as you can. Don't forget to label all plots and include legends wherever necessary as this is key to making good visualizations! Ensure that the plots are visible enough by playing with size parameters. Be sure to use appropriate color schemes wherever possible to maximize the ease of understandability. Everything must be laid out in a python notebook(.ipynb). (**5 Points**)

Submission

1. This assignment must be done individually by every student. Your code will be checked thoroughly to detect copying/plagiarism. Do your own work!
2. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!
3. Please use Piazza to ask any questions.
4. Submit everything through Blackboard. You will need to upload:
   1. The Jupyter notebook all your work is in (.ipynb file)
   2. Python file (export the notebook as .py)
   3. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse351_hw1_lastname_firstname_sbuid.ipynb
2. cse351_hw1_lastname_firstname_sbuid.py
3. cse351_hw1_lastname_firstname_sbuid.pdf

References

[1]. Installation instructions, courtesy of Professor Steven Skiena (CSE 519)
[2]. Data set https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Please keep in mind that:
(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.
(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.