CSE 351 - Introduction to Data Science (Summer 2020)
Prof. Praveen Tripathi
Homework 1: Exploratory Data Analysis

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.

Data downloading

There are two datasets required for this homework. Please find them attached in blackboard under "Assignment 1". The first dataset is about criminal activity in the Washington, D.C. metro area. The second dataset is regarding the unemployment rates in Washington D.C. based on the Ward (**Note**: Washington, D.C. is geographically divided into 8 wards).

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.

Tasks(50 points)

1. Filter out your data to examine the violent crimes. For 4 fields of your choice(2 numerical & 2 categorical), examine the distribution of the values, and explain any interesting insight you get from this. How do the distributions on violent crimes compare to the non-violent ones(show with visual analysis)? (**10 points**)

2. Often in Data Science, you won’t be able to find one dataset with all the information required for your analyses. Instead, you will have to find datasets from multiple sources and fuse them together yourself to proceed with your analytics. For this task, you are required to combine both datasets given to you (**HINT**: Dataframes can be combined using the pandas merge function).
    a. Which “ward” reported the most criminal activity based on the data presented? Justify with a plot. (**3 points**)
    b. Which “ward” reported the highest average unemployment rates? (**2 points**)
    c. For the ward from part b, plot the trends between the number of crimes occurring in this ward along with the unemployment rate of this ward based on the years chronologically present in the datasets. Note that you may have to clean the data to exclude some years which may not have data for the entire year, as this will skew your analysis. Explain your procedure and what you observed. (**5 points**)

3. **XBLOCK** and **YBLOCK** refer to the coordinates of where a certain crime has taken place.
    a. For the year 2016, plot a scatter plot based on these coordinates, where the points represent the crimes in that year, and the crimes are color coded based on **DISTRICT.** You may have to handle missing values, explain how you handled these rows. (**5 points**)
    b. Plot a scatter plot for the same year as above, where the crimes are color coded based on **OFFENSE**. Explain what you observed(**HINT:** Playing around with the opacity of the points may help you make interesting observations. For seaborn, this can be done using the “alpha” parameter). (**5 points**)

4. Make your best educated guess as to which time of the day was the most dangerous in the D.C. area. Note that we are defining “danger” to be any violent crime for the purpose of this question. Explain your inference with suitable plots. (**5 points**)

5. 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**)

6. 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. For question 2, you may **NOT** merge the datasets manually. In a real world scenario, datasets will be much larger and you'll have to be familiar with the pandas functionality(which is fairly straightforward).
2. If you are not familiar with Python and the mentioned libraries, this may take some time. Please get started early!
3. This assignment must be done **individually** by every student. Your code will be checked thoroughly to detect copying/plagiarism. Do your own work!
4. Please use Piazza to ask any questions.
5. 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)