CSE 519 -- Data Science (Fall 2018)
Prof. Steven Skiena
Homework 2: Exploratory Data Analysis in iPython
Due: Tuesday, September 25, 2018 (8:30 AM)

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 a data set where you have some basic sense of familiarity.

This homework is based New York Taxi Fare Prediction on Kaggle, revolving around predicting the fare of a taxi ride given a pickup and a drop off location. More than just data exploration, you must also join the challenge and submit your model before the deadline, to get a score feedbacked from Kaggle. You are to explore the data and uncover interesting observations about the New York Taxi operations. You will need to submit all your results in a single google form and your code files in three different format (.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 downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest to install **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 at here. Installing instruction can be found here. A useful instruction about Anaconda in Youtube 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. Some packages I believe you will definitely use for this homework are as following:

- pandas
- scikit-learn
- numpy
- matplotlib
- seaborn

Tasks (100 pts)

1. Take a look at the training data. There may be anomalies in the data that you may need to factor in before you start on the other tasks. Clean the data first to handle these issues. Explain what you did to clean the data (in bulleted form). (10 pt)
2. Compute the Pearson correlation between the following: (9 pt)
    a. Euclidean distance of the ride and the taxi fare
    b. time of day and distance traveled
    c. time of day and the taxi fare

    Which has the highest correlation?
3. For each subtask of (2), create a plot visualizing the relation between the variables. Comment on whether you see non-linear or any other interesting relations. (9 pt)
4. Create an exciting plot of your own using the dataset that you think reveals something very interesting. Explain what it is, and anything else you learned. (15 pt)
5. Generate additional features like those from (2) from the given data set. What additional features can you create? (10 pt)
6. Set up a simple linear regression model to predict taxi fare. Use your generated features from the previous task if applicable. How well/badly does it work? What are the coefficients for your features? Which variable(s) are the most important one? (12 pt)
7. Consider external datasets that may be helpful to expand your feature set. Give bullet points explaining all the datasets you could identify that would help improve your predictions. If possible, try finding such datasets online to incorporate into your training. List any that you were able to use in your analysis. (10 pt)
8. Now, try to build a better prediction model that works harder to solve the task. Perhaps it will still use linear regression but with new features. Perhaps it will preprocess features better (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values). Perhaps it will use a different machine learning approach (e.g. nearest neighbors, random forests, etc). Briefly explain what you did differently here versus the simple model. Which of your models minimizes the squared error? (10 pt)
9. Predict all the taxi fares for instances at file “sample_submission.csv”. Write the result into a csv file and submit it to the website. You should do this for every model you develop. Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation. (15 pt)

Be honest. This is your first modelling experience, and I am hoping to see you learned something, not just where you are ranked on the leaderboard.

Rules of the Game

This assignment must be done **individually by each student**. It is not a group activity.

1. 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!
2. All of your written responses will be submitted through a form during submission. It may make sense to keep your answers inside your notebook and copy it over into the form when you are ready to submit.
3. We will discuss topics like linear regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
4. To ensure that you are who you are when submitting your models, have your Kaggle profile show your face as well as a Stony Brook affiliation.
5. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
6. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2018/cse519.

Submission

Submit everything through Google classroom. As mentioned above, 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)

For everything else, you will fill out a separate Google Form. These will include the responses to all of the task questions above. You will also need to link your Kaggle profile. It is recommended that you have a look at the Google response form for the questions asked and write all your responses in a local document. Once you feel comfortable with your responses, you may record your final responses in the form and then submit.