This project is designed to test your skill and intuition about real world data. For the project, we will use data collected by the New York City Taxi and Limousine commission about "Green" Taxis. Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. We will use the data from September 2017. We are using NYC Taxi and Limousine trip record data: (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml).

What to submit – **All the source codes** that are used to solve the following problems, **the screenshots of results of each problem**. You can submit multiple documents (include .ipynb file) if needed. Your documents should show the clear work/solution you have conducted to solve the problems.

Required Questions: Please answer completely all four required questions.
Question 1

- Programmatically download and load into your favorite analytical tool the trip data for September 2017.
- Report how many rows and columns of data you have loaded.

Question 2

- Plot a histogram of the number of the trip distance ("Trip Distance").
- Report any structure you find and any hypotheses you have about that structure.

Question 3

- Report mean and median trip distance grouped by hour of day.
- We'd like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criteria, the average fair, and any other interesting characteristics of these trips.

Question 4

- Build a derived variable for tip as a percentage of the total fare.
- Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). We will validate a sample.

Question 5
Undergraduate student:
Choose only one of these options to answer for Question 5.
Graduate student:
Choose two of these options to answer for Question 5.

- *Option A*: Distributions

    - Build a derived variable representing the average speed over the course of a trip.
    - Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?
    - Can you build up a hypothesis of average trip speed as a function of time of day?

- *Option B*: Visualization
    - Can you build a visualization (interactive or static) of the trip data that helps us understand intra- vs. inter-borough traffic? What story does it tell about how New Yorkers use their green taxis?

- *Option C*: Search
    - We're thinking about promoting ride sharing. Build a function that given point a point P, find the k trip origination points nearest P.
        - For this question, point P would be a taxi ride starting location picked by us at a given LAT-LONG.
    - As an extra layer of complexity, consider the time for pickups, so this could eventually be used for real time ride sharing matching.
    - Please explain not only how this can be computed, but how efficient your approach is (time and space complexity)

- *Option D*: Anomaly Detection
    - What anomalies can you find in the data? Did taxi traffic or behavior deviate from the norm on a particular day/time or in a particular location?
    - Using time-series analysis, clustering, or some other method, please develop a process/methodology to identify out of the norm behavior and attempt to explain why those anomalies occurred.

- *Option E*: Your own curiosity!
    - If the data leaps out and screams some question of you that we haven't asked, ask it and answer it! Use this as an opportunity to highlight your special skills and philosophies.