2nd Capstone Project Proposal

Which New York Taxi Trip is the fastest?

1. What is the problem you want to solve?
    - Which features determine the duration of the taxi trips in New York City, and what is their relative importance?
    - Where and when do the fastest and slowest trips originate?
    - How correlated is the trip duration with the distance between pick-up and drop-off? How are the average speeds of the trips distributed?
    - Create a regression model that best predicts the trip durations as well as the average speeds of the trips.

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?
    - Taxi companies are interested in optimizing the trips within the city limits as well as between cities. In other words, they are interested in knowing which trips, originated at different locations and times, are likely to be faster given the same distances between origins and at destinations.
    - Citizens who use cabs frequently are themselves interested in knowing where and when it may be convenient to catch a cab in order to save time.

3. What data are you going to use for this? How will you acquire this data?

    - The data will be downloaded from the following webpage:

      https://www.kaggle.com/c/nyc-taxi-trip-duration

    - The dataset includes pickup time, drop-off time, geo-coordinates, of pick-up and drop-off, number of passengers, and several other variables (total 11 variables). It was prepared by the NYC Taxi and Limousine Commission.

4. In brief, outline your approach to solving this problem (knowing this might change later).
    - I will visualize the features I consider important in the raw dataset.
    - I will perform the feature engineering generating new features such as month, day and hour of the pickup, as well as the map distance between pickup and drop-off locations.
    - I will determine the degree of correlation between the input variables.
    - I will determine the degree of correlation between the duration of the trip (the target variable) and the input variables.
    - Towards the end, I will first try different models to determine the one that best predicts the trip durations from the test set. Second, I will try to do the same not for the duration of the trip itself, but for the new variable called “average speed of the trip”. Third, I will look at the relative importance of features.

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.
    - Code.
    - Analyses, visualizations, model in the form of a report.
    - Slide deck.