Capstone Project 1 Proposal

New York City Taxi Trip Duration

What is the problem you want to solve?

Building a model that predicts the total ride duration of taxi trips in New York City. Our primary dataset is released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

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 done otherwise?

NYC Taxi and Limousine Commission. With my analysis, the client can provide the trip estimation time to the users as well as the price.

What data are you using? How will you acquire the data?

I will be using a data that was provided by the client available in Kaggle:
https://www.kaggle.com/c/nyc-taxi-trip-duration/overview

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution:

I will start by exploring the data, have an overall overview of each column and print the statistics. Next, I will start the cleaning process, look if there are any null values. The data wrangling part will include adding new columns such as the distance and the speed. For this, I will use a third party API, probably google or open street map. I will also make some changes in the units of some columns such as the duration, I would rather have it in minutes instead of seconds. I will create few DataFrames based on pick up location, drop off location, trip duration and eliminate outliers. Once my data is clean and ready to use, I will conduct my analysis based on Geohashing pick up and drop location, pick up time of the day and day of the week.

What are your deliverables? Typically, this includes code, a paper, or a slide deck:

I will deliver a .CSV file that has basically two columns, one of them is the trip ID and the other one has the the trip duration estimate.