Urban Mobility: exploring taxi trip fares in the San Francisco Bay Area

In this project, you will explore taxi trip fares in the San Francisco Bay Area. You will study the effects of traffic-related delays on the extra charges passengers have to pay.

Data

The data is named *SF_taxi_data.csv*, and contains San Francisco Taxi data from 9/1/2012 to 9/17/2012. The dataset consists of 50,000 taxi trips taken in the Bay Area during that time period. For each trip you are given the departure time, arrival time, passenger fare, departure lat/lon coordinates, arrival lat/lon coordinates, departure taz (Traffic Analysis Zone), arrival taz, and the distance between origin and destination in miles.

Data might have outliers and errors. To make sure that you have removed all the outliers, do the following:

- Get trips where the number of passengers is only 1, and remove fares less than \$3.5, since base fare = \$3.5, these are likely to be errors. Do the rest of analysis on this filtered data.

Data exploration

Analysis 1. SFMTA fare calculation table states that the fare for a trip of x miles will be at least $3.5 + 0.55 \times (5x - 1)$. Explore how this compares to the actual relationship between fares and distance that you can find using regression methods that you have learned in class.

Analysis 2. Turns out the TAZ that contains SFO (San Francisco International Airport) is the TAZ that generates the most trips. find the most popular "deptaz" and save the taz id as sfo_taz. Then split the dataset into two groups of trips that originated or ended at SFO, and the rest of the trips. What can you say about these two groups of data? How different they are? (your answers need to be quantitative, based on some measures or metrics)

Machine learning

Analysis 3. In most trips, the actual fare was way higher than the base one (the equation given in analysis 1). The reason for the extra fare paid by pax is traffic delays (and also drivers taking longer trips than necessary, sometimes due to congestion). Let's see if this extra surcharge is related to the length and duration of the trip. This will help us predict it before a trip started. Fit a linear regression of travel distance vs. extra cost (the cost above the base fare) on the data from Day 1 to Day 10 of September, and see how the fitted line predicts the fares for data from Day 11 to Day 17? What is the error in prediction?

Analysis 4. Both length and duration of a trip seem to have an association with extra fares. Longer trips are more expensive. Do the same analysis in (3), but for travel duration vs. extra cost, and identify whether trip length or trip duration seem to be a better predictor of the extra travel fare?

Linear regression vs. k-nearest neighbor

Analysis 5. Take the data of fares and distances from day 1 to day 10 of September as training data, and data from day 11 to day 17 as testing data. Apply both linear regression and k-nearest neighbor (with $k = \{1, 5, 10\}$) to identify the fares in the testing data given the distances and compare the results. Which method is better and why? What is the difference between performance of k-NN with different values of k?