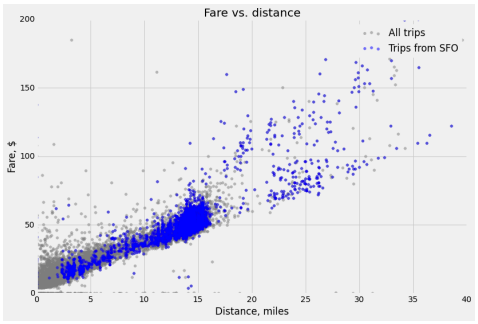
CE 88
Homework 6
Due 3/8/2016 (1 week)

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



Data. We will continue working on the taxi trips dataset explored in Minilab 6. The data are located in '*data/SF_taxi_data.csv*'. Minilab 6 contains useful code for calculating trip durations amongst giving other useful tips.

Problem 1 (2 points). We would like to focus on the data subset where all factors that may hinder our conclusions are eliminated. Filter out all the trips with number of passengers ('num') is greater than 1. Filter out all trips where fare is recorded as zero (these are most likely wrong). Further, filter out all trips originating or terminating at SFO (TAZ id 239). How many trips are there in the resulting subset?

First 1/5 mile	$3.50
Each additional 1/5th mile	$0.55
Each minute of waiting, or traffic time delay	$0.55

Problem 2 (4 points). Produce a scatter plot of the trip fares paid as a function of the travel distance. A taxi company in the Bay Area has to charge single passenger trips as shown in the table (source: SFMTA). Assume there is no congestion or any other traffic-related delays in the area, and waiting time is negligible for the majority of trips. In this case, no surcharge on traffic time delay is applied. Write a formula to compute the trip fare in this 'ideal' case when it is only a function of distance. This is a theoretical lower bound on the trip fare. Add a corresponding line to your scatterplot. Can you see that passengers pay significant extra charges due to traffic delays?

Additionally, there is a 150% surcharge on the 'out-of-town' trips longer than 15 miles (see the detailed definition at SFMTA page). Can you see two 'branches' of the two different types of trips (>15miles) in the scatter plot? Filter out all trips >15 miles to approach the next question.

Problem 3 (4 points). Compute the 'extra' charge passengers face due to traffic delays by subtracting the theoretical lower bound fare from the actual amount paid. Produce a scatterplot of this extra charge as a function of the trip duration. Is it more likely, on average, to experience longer traffic delays when taking longer trips? Justify your answer with regression (or correlation) analysis.

Your submission must contain both a PDF file and the original ipynb.