## Problem Definition

Our group (Greg, Anna and Ryan) decided to use the 2013 New York taxi dataset given to us in class. The first thing we needed to come up with was a business question to answer. We decided to devote our efforts to attempt to predict the amount of time a trip would take. This has significant business impact because it has a profound effect on driver scheduling. Another benefit to solving for this is that if we can predict the time, we can also predict the money to charge (similar to Uber, where the cost is known before a person accepts a ride). After cleanup, our data set has just over 1.5 million rides.

## Exploratory Data Analysis

Once we decided on a question, we sought out to understand the data that we were given. It is hard to make a defensible model when the features we are given are unknown to us. The model can have a very good fit, but if we cannot effectively communicate our findings, our model would very likely be dismissed by our business partner, particularly if the results did not align with their preconceived notions. As it turns out, many of the columns in the datasets are self-explanatory, but others are not. We did some research on the internet to try to find the meaning behind some these other columns. For example, our original intuition was that VENDOR\_ID might have been the cab company, but as it turns out, according to this site ([http://iquantny.tumblr.com/post/107245431809/how-software-in-half-of-nyc-cabs-generates-52?lipi=urn%3Ali%3Apage%3Ad\_flagship3\_messaging%3BOO1WYcrmTt6AzLIIyl%2BirA%3D%3D](http://iquantny.tumblr.com/post/107245431809/how-software-in-half-of-nyc-cabs-generates-52?lipi=urn%253Ali%253Apage%253Ad_flagship3_messaging%253BOO1WYcrmTt6AzLIIyl%252BirA%253D%253D) ) it is a payment processing company. We also found a data dictionary (<http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf?lipi=urn%3Ali%3Apage%3Ad_flagship3_messaging%3BOO1WYcrmTt6AzLIIyl%2BirA%3D%3D>) for a similar taxi dataset that explained columns like STORE\_AND\_FWD\_FLAG. There is a danger to using these, as it is possible that the column in one dataset is not the same as something named very similarly in another dataset, but we were confident in our belief that some of these previously unknown columns were in fact the same.

Once we felt like we understood the data, the next step was to go through the process of data cleansing, one of the most important parts of the data science process. The files we were given were mostly clean, but they did come with a few issues. First, we made sure to remove any trips the recorded a distance of zero miles or a duration of zero seconds. We also removed any trips that seemed too short or too long (under 0.1 miles/over 10 miles, or under 60 seconds/over 50 minutes). To further filter the data, we calculated the average speed for each trip and removed any extreme outliers (under 2 mph/over 60 mph). In addition, there were a few duplicates in the data. We noticed that a few of the latitude and longitude coordinates were reversed, causing a problem when trying to map them to ZIP codes.

Once we felt like we understood the data, we performed exploratory analysis and gained a few insights. First, most trips are short, the median trip distance and time are just over a mile, and around 10 minutes, respectively. Also, early morning trips are the fastest, while afternoon trips are slowed down due to traffic. Finally, the longest trips by ZIP code are to and from the various regional airports, not surprisingly.

## Feature Engineering

After getting a handle on the data we were given, we thought of ways that we should augment our data. Most questions often need complementary data sources to come up with satisfactory answers. While we were given latitude and longitude values for pick up and drop off locations, we felt that using ZIP codes would give us more of a discrete variable to use in our model. We also thought that weather might play a big role in differing trip times of the same distance from the same location, so we found a dataset that had hourly weather data collected at JFK airport in NYC. There are other datasets that we felt like we could not use. We thought that any social event data might be problematic because it would be very incomplete and affect our model.

With all of these datasets, we started to brainstorm new features we could generate. One of the natural things to do was to split up the date and time values into chunks. For example, we thought the hour of the day and the day of the week would be important, since they can be indicators of traffic. In addition, we created binary features indicating whether a trip occurred on a holiday or a weekend.

One of the most important features that was not modeled in the data set is traffic. The distance for a given trip is fixed, but the time taken to transverse that distance can vary greatly depending on the en route traffic. To try and simulate traffic, we created three new features. The first is the direction of a trip, which is one of 16 cardinal directions (N, NNE, NE, ENE, etc.). The second and third is the region in which the trip originated and the number of rides occurring in the same region at the hour that the taxi ride started. The regions were determined by using the kmeans algorithm to cluster pickup locations (in terms of latitude and longitude) into 16 distinct clusters. The cluster ID and number of rides in the cluster during the pickup hour were added as two additional features.

Finally, there were a few features we felt we could not put into our model. We felt we could not use most of the money related columns. For one, we don’t know how much money a cab ride will cost until it is over. The “total amount” features acts as leakage.

## Current Technical Blockers

Currently, a Random Forest model is being used to predict the travel time, as it performed the best when compared to other models (linear regression, single decision tree, or boosted decision trees). The model is capturing around 75% of the variance in the response variable. We would like this value to be higher, and we plan to try and get there hyper-parameter tuning.

## Current Progress/What’s Left

So far, we’ve selected our data set, performed data cleaning and exploratory analysis, and begun the model creation and evaluation process. The remaining steps are to tune the parameters of our model to reduce the error in the predicted trip time (the model is being evaluated using the RMSE metric). The current best RMSE on the trip time is ~210 seconds, or 3.5 minutes.