Our team is composed of three members, Anna Jurgensen, Greg DeVore, and Ryan Blosser. We have chosen to use the classroom-supplied New York City cab data for our capstone project.

We believe that one of the most important components of a modeling project is defensibility. It is one thing to plug a bunch of features into a model, but it’s another thing to know what they mean. To that end, one of the first things we needed to do is determine what data we are looking at. Many of the columns are self-explanatory, but we needed to research some of the other ones. We found a few other transportation data sources on [WWW.CITYOFNEWYORK.US](http://WWW.CITYOFNEWYORK.US) to help us explain a few of these columns, such as “Store\_and\_fwd\_flag”.

We also needed to come up with a question we wanted to answer and felt it was important to approach it from the business’ perspective. What would the business want to know? We decided it would be good to be able to predict trip distance, so we will be using regression modelling techniques for this project.

After deciding on this, we felt that data cleaning and feature generation would be necessary. For example, there are negative values for various payment categories, so we will be removing these. As far as generating features, any time dates or times are introduced, it is good practice to parse these out and categorize the components (day of week, weekend/weekday, hour of day, is/is not peak time etc). We will also be leveraging an external shapefile to translate the latitude and longitude values into zip codes so as to bin these into fewer categories.