Data set: We will use data collected by the New York City Taxi and Limousine commission about “Green” Taxis. Green Taxis (as opposed to yellow ones) are taxis that are not allowed to pick up passengers inside of the densely populated areas of Manhattan. We will use the data from September 2015. We are using NYC Taxi and Limousine trip record data: (<http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>)

Questions:

**Question 1**

1. Programmatically download and load into your favorite analytical tool the trip data for September 2015.
2. Report how many rows and columns of data you have loaded.

**Question 2**

1. Plot a histogram of the number of the trip distance (“Trip Distance”).
2. Report any structure you find and any hypotheses you have about that structure.

**Question 3**

1. Report mean and median trip distance grouped by hour of day.
2. We’d like to get a rough sense of identifying trips that originate or terminate at one of the NYC area airports. Can you provide a count of how many transactions fit this criterion, the average fare, and any other interesting characteristics of these trips?

**Question 4**

1. Build a derived variable for tip as a percentage of the total fare.
2. Build a predictive model for tip as a percentage of the total fare. Use as much of the data as you like (or all of it). Provide an estimate of performance using an appropriate sample and show your work.

**Question 5**

Distributions

1. Build a derived variable representing the average speed over the course of a trip.
2. Can you perform a test to determine if the average trip speeds are materially the same in all weeks of September? If you decide they are not the same, can you form a hypothesis regarding why they differ?
3. Can you build up a hypothesis of average trip speed as a function of time of day?