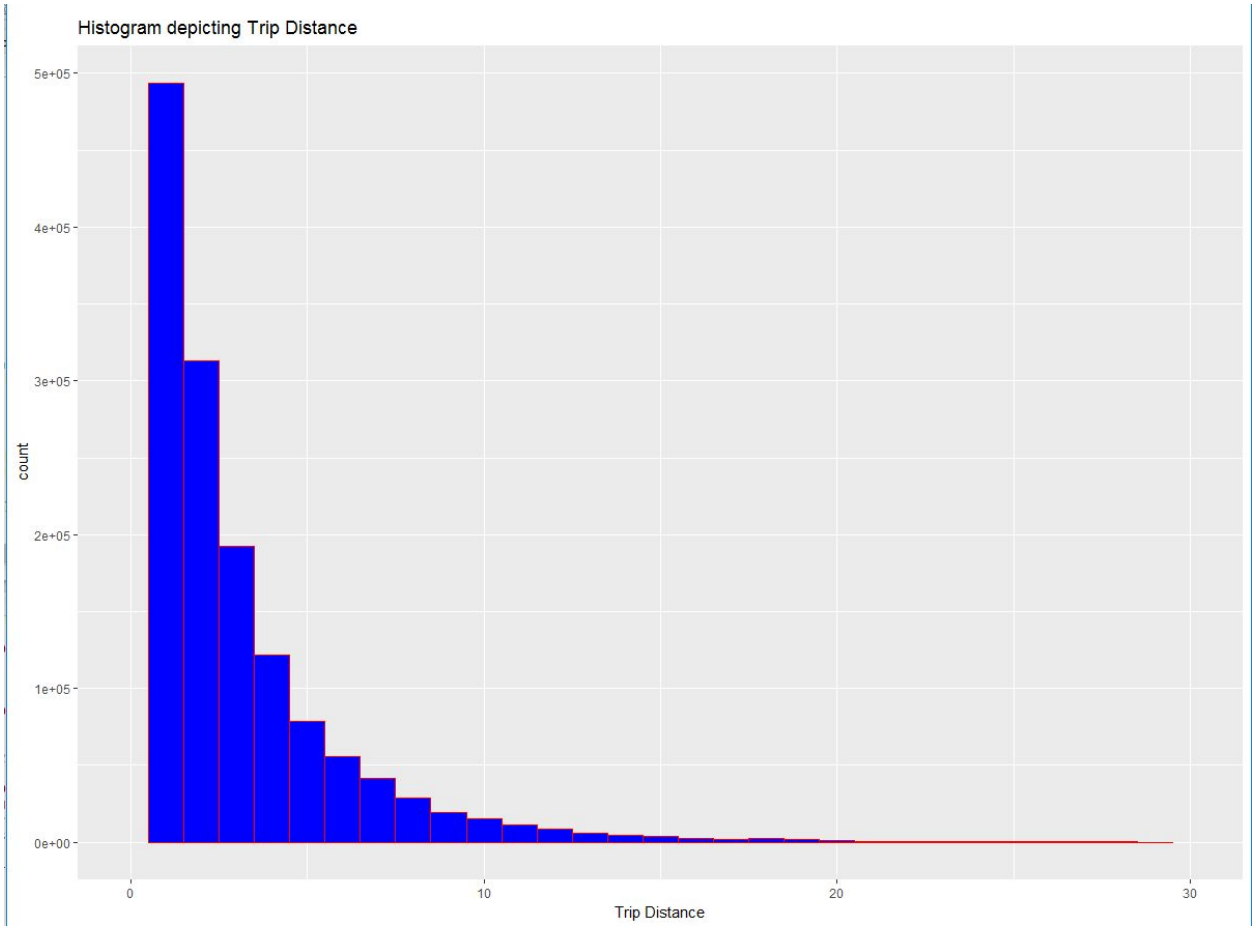
Capital One Data Science Challenge

Q1. Programmatically download and load into your favorite analytical tool the trip data for September 2015. Report how many rows and columns of data you have loaded.
A. I used the fread() function in R available in the data.table package, to import data from a source URL.
There were 1494926 rows and 21 columns loaded into R.

Q2. Plot a histogram of the number of the trip distance ("Trip Distance"). Report any structure you find and any hypotheses you have about that structure.
A. On an initial plot displaying the trip distance across the entire data, I found that most trips concluded at a distance below 30 miles.
I then filtered out the trips which lasted between 0 and 30 miles and only plot a histogram of those trips.

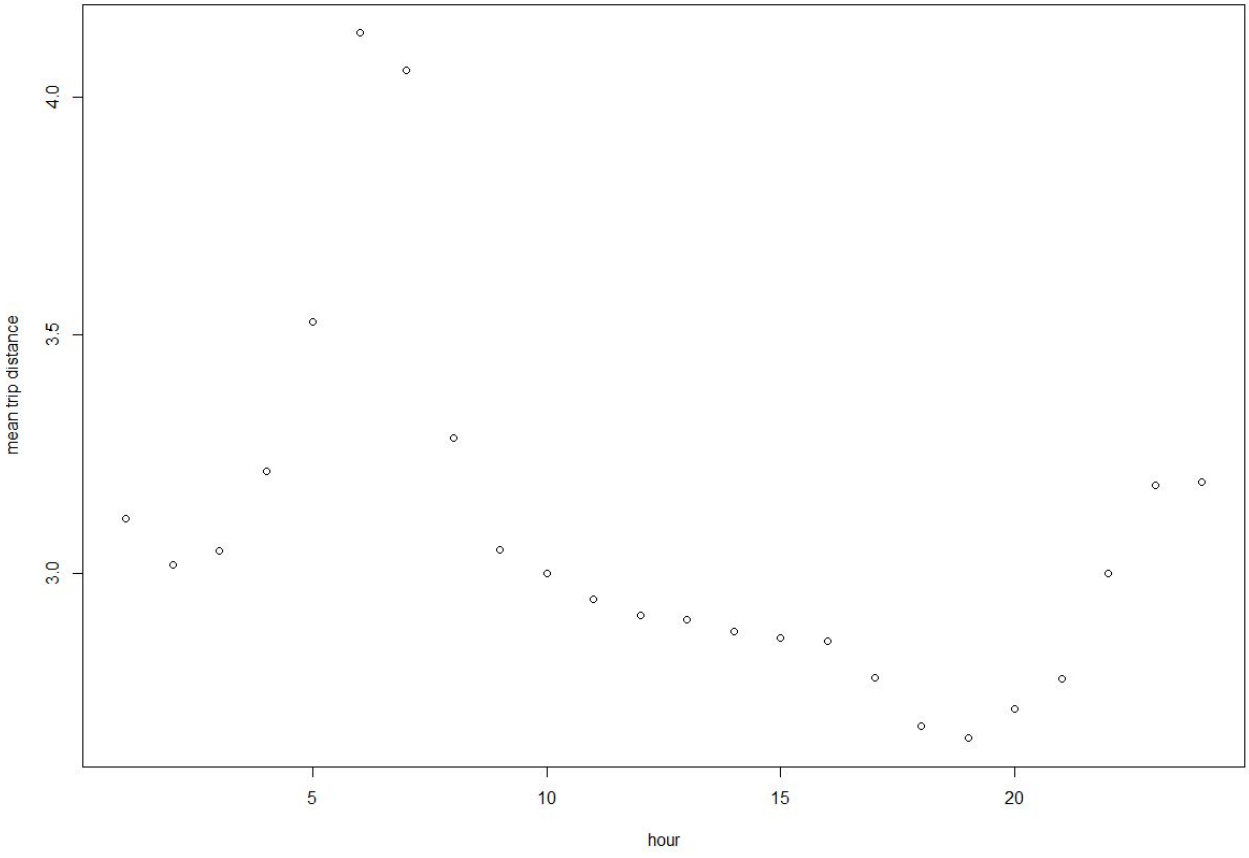


It is observed that most trips range between 0 and 10 miles, with journeys of 0-3 miles having most frequency. As trip distance increases, frequency of the trips decreases. Therefore trip

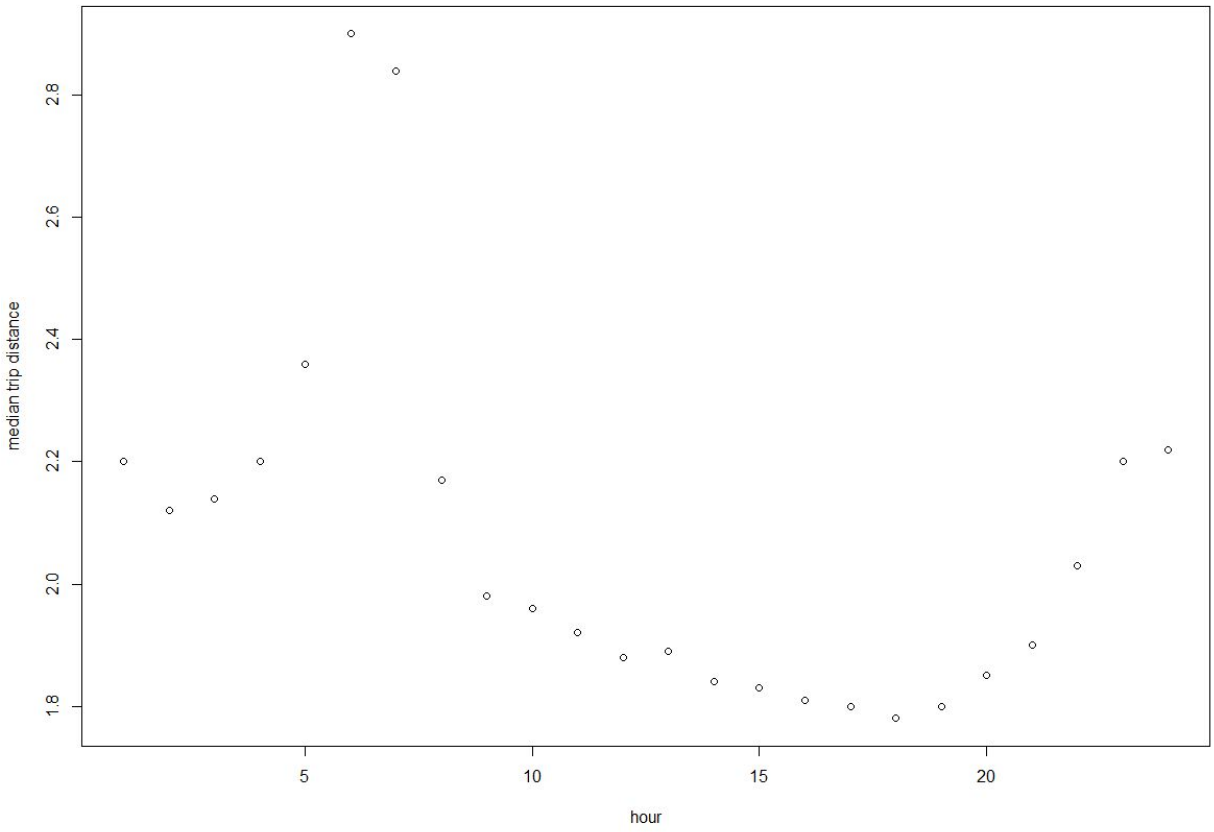
distance and count of the trips is inversely related. This can possibly mean that Green Taxis are typically used for shorter rides.
Only 467 trips have had trips longer than 30 miles. These can safely be assumed as outliers.
Hypothesis : Green Taxis are typically used for short commutes

Q3. Report mean and median trip distance grouped by hour of day. 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 criteria, the average fair, and any other interesting characteristics of these trips.
A. The mean trip distance grouped by hour sees a gradual decrease in value starting from 08:00 hour upto 20:00 hour, after which it increases again. There is a spike in hour 5 and 6 of the day.

Mean trip distance by hour



Median trip distance by hour



For trips near airports, I classified trips as near/not near airports by checking the difference in latitudes/longitudes of the pickup/dropoff location and the latitudes/longitudes of established airports. I then performed some analysis on the data classified as near airports.
Statistics:

1. Percentage of trips originating or ending near airports = 26.23%
2. Average fare of trips originating or ending near airports = 13.77$
3. Average fare of trips NOT originating or ending near airports = 12.10$
4. Average tip of trips originating or ending near airports = 1.1$
5. Average tip of trips NOT originating or ending near airports = 1.29$
6. Tipping ratio = (Tip % of airport trips)/(Tip % of other trips) = 0.75% i.e Trips involving airports generate only 75% the amount of tips as other trips.
7. Mean passenger count of trips originating or ending near airports = 1.44
8. Mean passenger count of trips NOT originating or ending near airports = 1.34

Conclusion : Mean passenger count of trips near airports is higher than the trips which are not near airports. This may mean that trips to or from airports may involve groups of people.

Also the fact that even though there are more passengers in these trips but tipping is worse is surprising. It may mean that people tend to stick to a budget while travelling(flying).

Q4. Build a derived variable for tip as a percentage of the total fare. 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). We will validate a sample.

A. The derived variable for tip as a percentage of fare was based on the ratio of tip amount and Fare amount for each trip.

Predictive Modeling:

My training data was limited to the trips which had a fare amount between $10 and 20$, and tip amount between $1 and $2. I used two approaches for modelling:

1. Linear regression
2. Random Forests

My test was was the trips which had a fare amount between $20 and $25 and tip amount between $2 and $2.5

I compared the two models based on the root mean squared errors they generated on comparing the predicted values of the test data to their actual values.

Random Forests did a better job of minimizing RMSE, hence I picked the random forest model. Linear regression generated a RMSE of 3% tip percentage, whereas Random forest model generated a RMSE of only 1.3%

Q5. *Option A*: Distributions. Build a derived variable representing the average speed over the course of a trip. 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? Can you build up a hypothesis of average trip speed as a function of time of day?

A. Since speed is distance/time, I cleaned the data to only include those trips who's time duration was not 0 seconds.

I computed the duration for each trip as Dropoff time - Pickup Time, and we are already given the trip distance.

Using these two values, I built a derived variable for average trip as Trip Distance/Trip Duration.

To check if average trip speeds are same in all weeks of September, I grouped the data according to their week number. Then I performed a T-test(null hypothesis: speeds must be same in each week) on each pair of weeks based on their average speed values. None of the p-values returned for these weeks were <0.05. Which led me to retain the null hypothesis that average trip speeds are materially the same in all weeks of September.

I used linear regression to train average trip speeds over time of day and I observed that the model returned a negative coefficient for timeofday. This led me to believe that as time of day increases, average trip speeds go down. This may be due to the fact that traffic increases later on in the day leading to decrease of speeds.