Exploratory Data Analysis

Data Summary and Visualization

Team Members:

- Darryl Pinto (dp6417@rit.edu)
- Naman Kothari (nsk2400@rit.edu)
- Renzil Dourado (rd9012@rit.edu)

Topic: New York City Taxi Trip Duration Prediction [1]

Outlier Analysis:

We observed that there are no missing values in the training data

```
> sum(is.na(train))
[1] 0
```

To detect outliers, one of the major attribute is going to be distance. As a result, we introduce a new attribute in our data trip-distance which is computed using the pickup and drop-off coordinates.

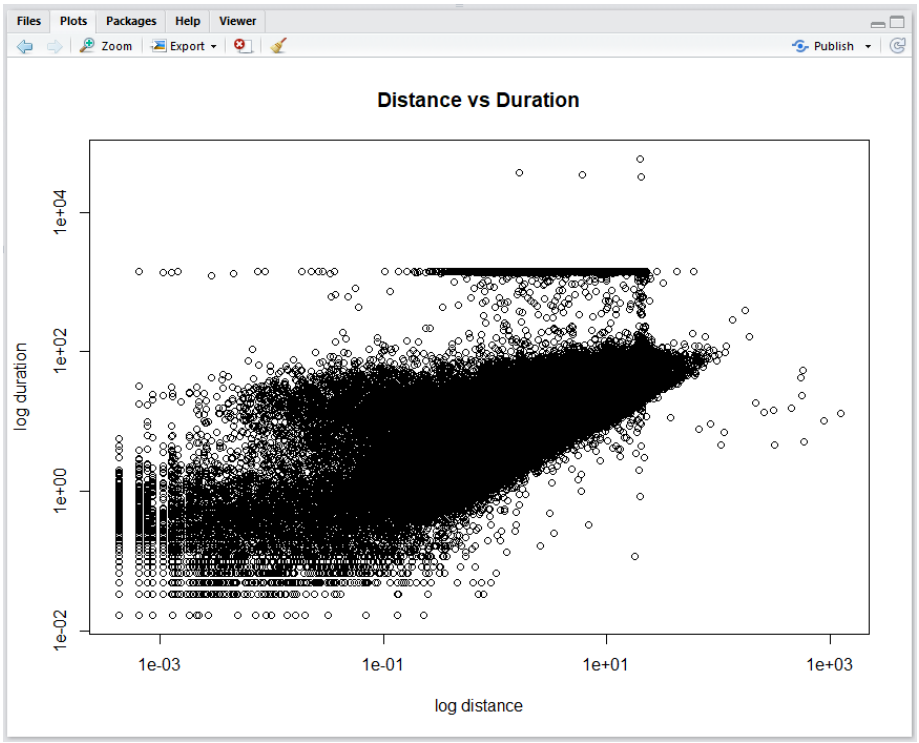

Figure 1

We see a positive correlation between trip-distance and trip- duration which is expected in Figure 1.

We assume that no taxi-trips within the city will be longer than a day. Hence, we considered the trips whose duration is greater than 24 hours as outliers and removed them.

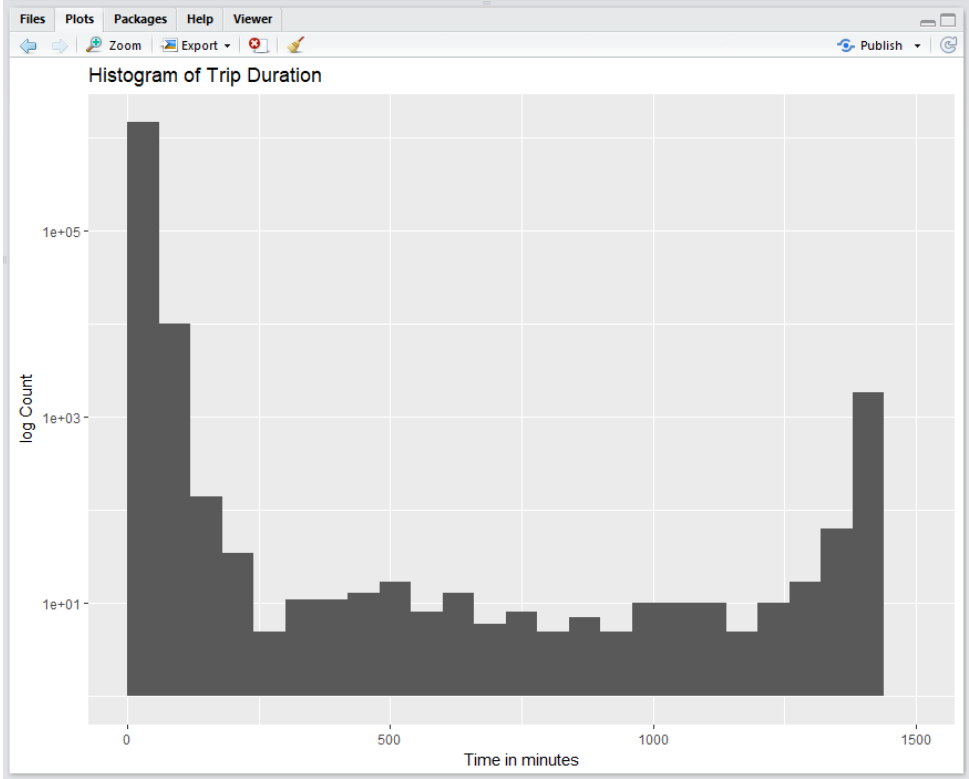

Figure 2

As we can see in figure 2, the count of trips in the first bin is very high.

We investigate the first bin. In the histogram shown in figure 3, we notice that the count of trips whose duration is less than a minute is very high. This may be because of incorrect data collection or immediate cancellation after booking the taxi. We do not consider this as a taxi trip and so these instances are dropped from our data.

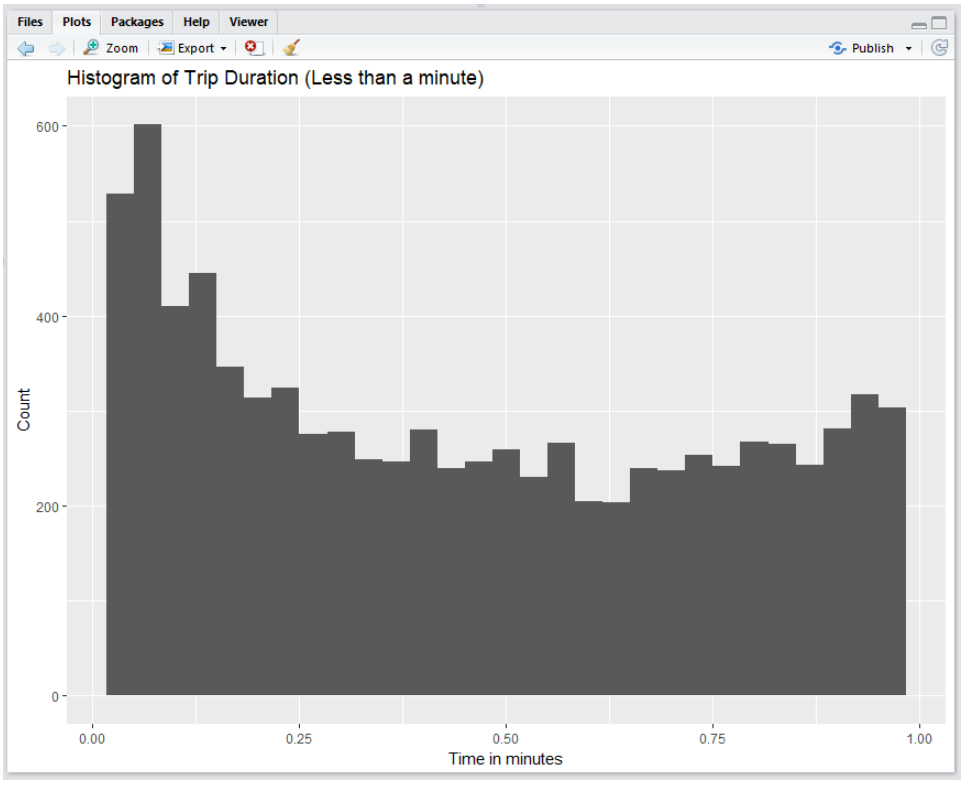

Figure 3

Now we consider the trips with a duration greater than 15 hours.

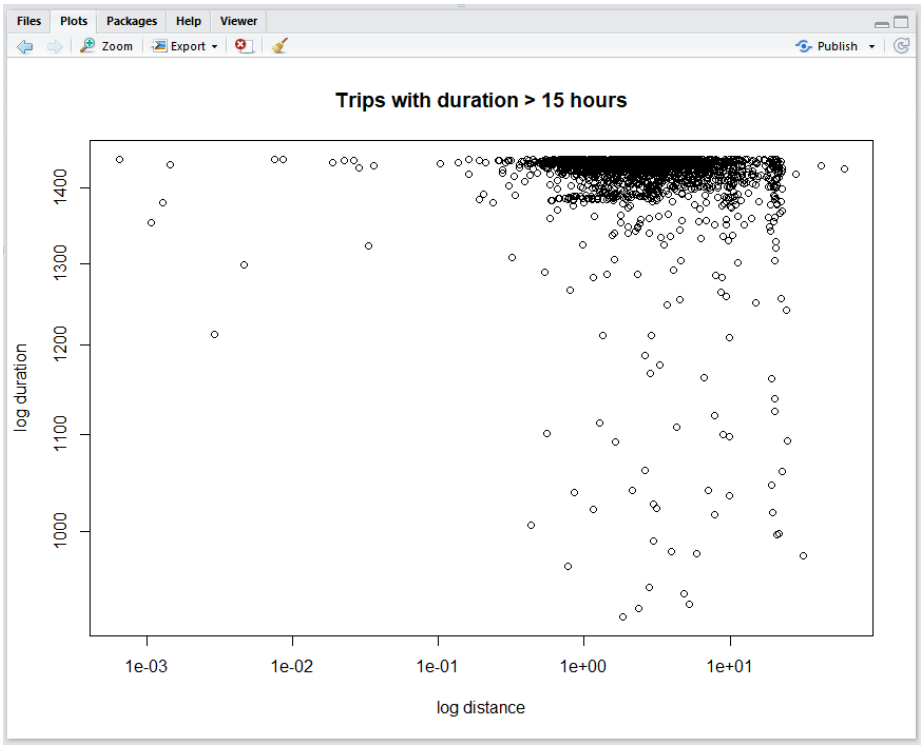

Figure 4

In the scatterplot shown on figure 4, we can see that a lot of trips have a distance in between 10-100 kilometers. The time taken for such trips should be much lesser than 15 hours irrespective of traffic or other conditions. Hence such data is not considered in our analysis. Such instances may be generated when the driver might not have marked the end of a trip once the passenger gets off.

Feature Analysis:

At this point, we believe that we have a clean data set and then proceed to analyze each feature individual. First, we plot the pickup locations of the trip using the Leaflet package.

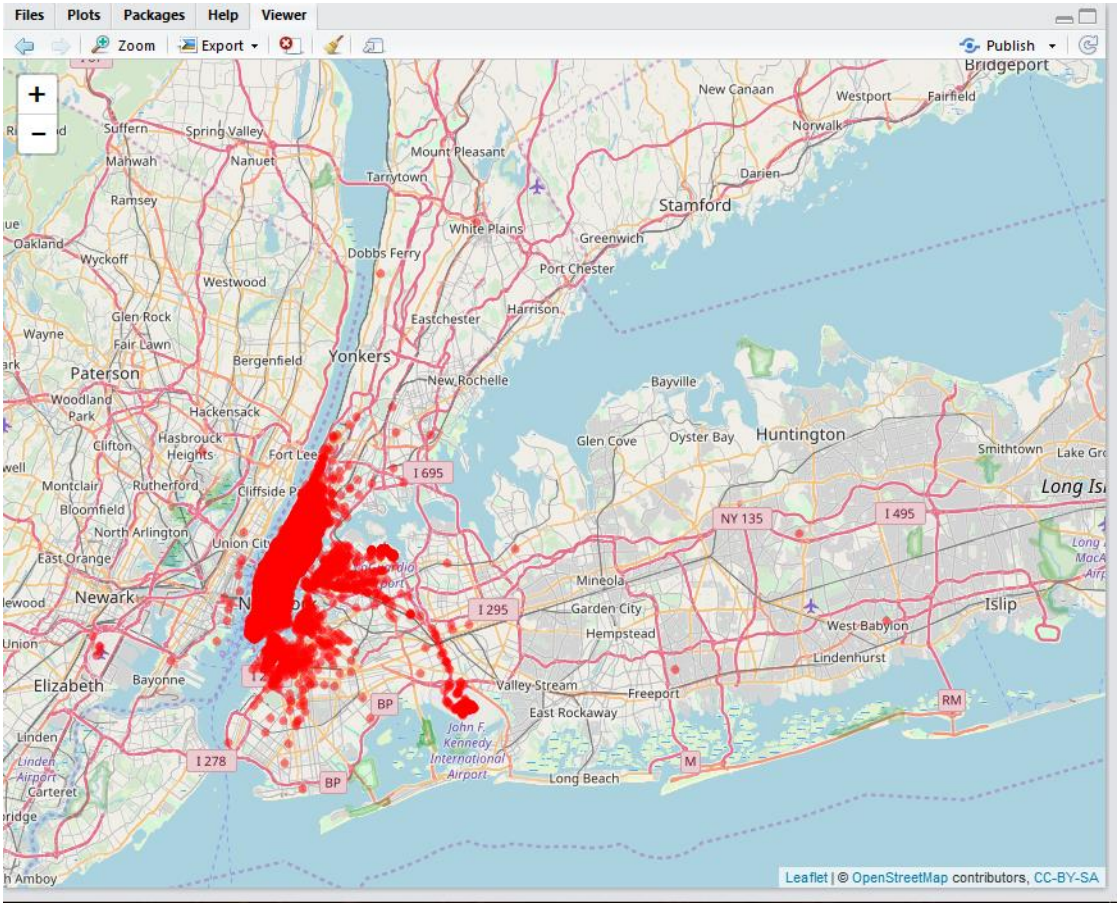

Figure 5

It is observed in figure 5 that most of the pickup locations are within downtown Manhattan, Brooklyn, JFK airport and La Guardia airport. There a few trips originating outside of NYC, but we still consider that they are valid trips and their drop off might be in NYC. [2] [3]

Now we plot the drop-off locations in NYC:

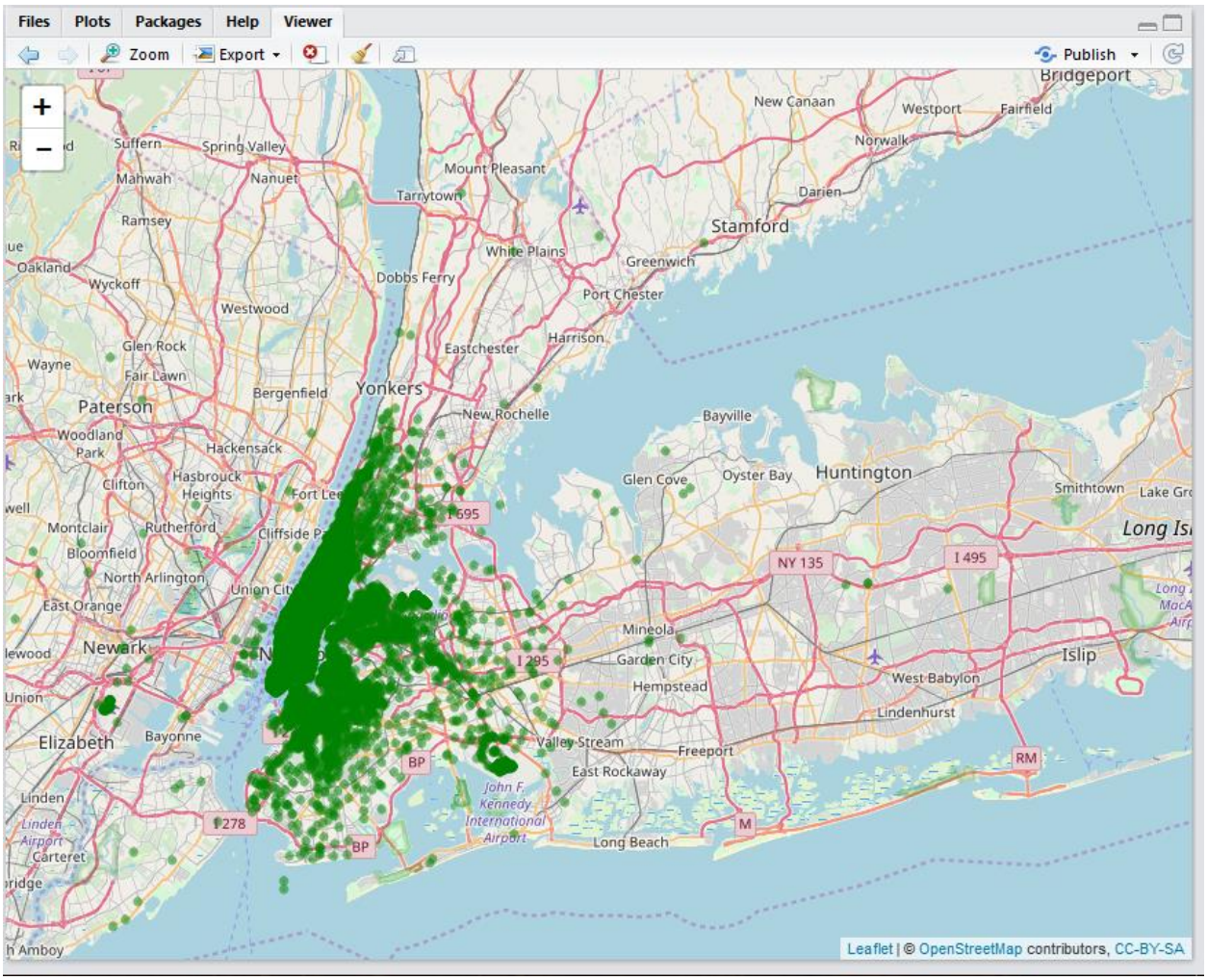

Figure 6

We notice in figure 6 that the most of drop-offs happen in downtown Manhattan, Brooklyn, La Guardia airport and JFK airport. On comparing the drop-off and pickup locations in the Brooklyn area, there are a lot more drop-offs in this area than the pickups. This may be because people use taxi for getting back home in Brooklyn from their start location. [2] [3]

The busiest pickup points in NYC are plotted in the map below in figure 7:

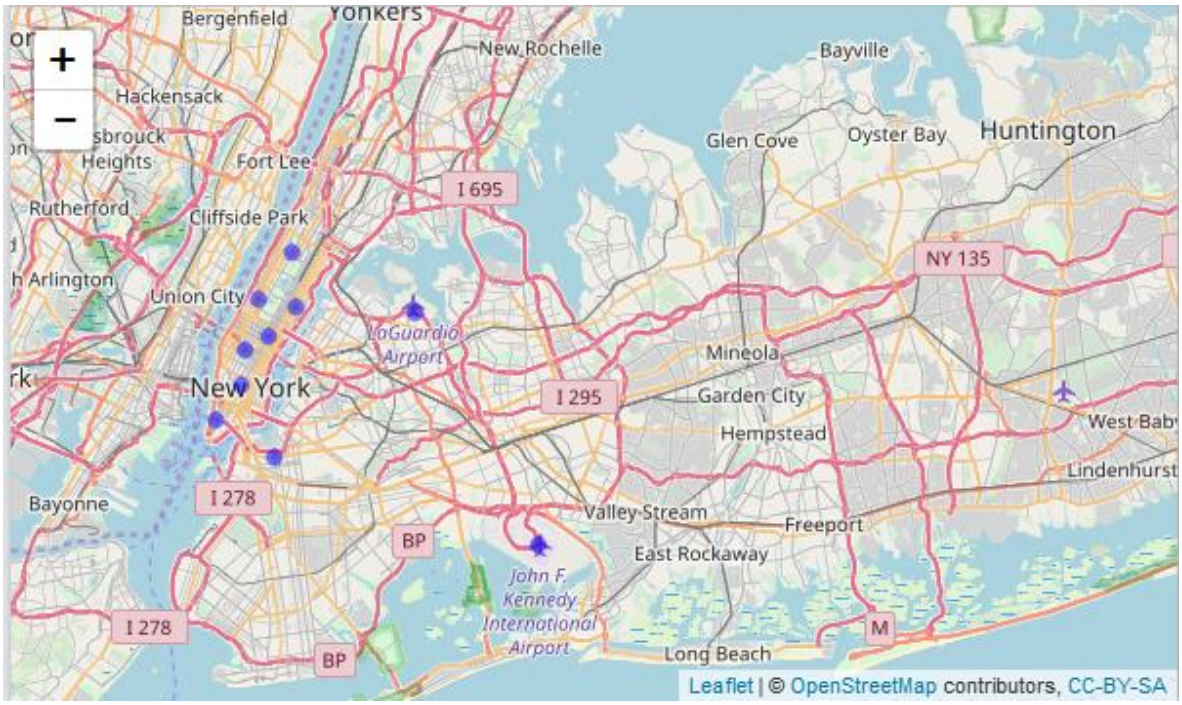

Figure 7

As it can be seen above, downtown Manhattan, La Guardia airport and JFK airport and the places where commuters use the taxi the most. [3]
The following graph shows the daily distribution of taxi-trips.

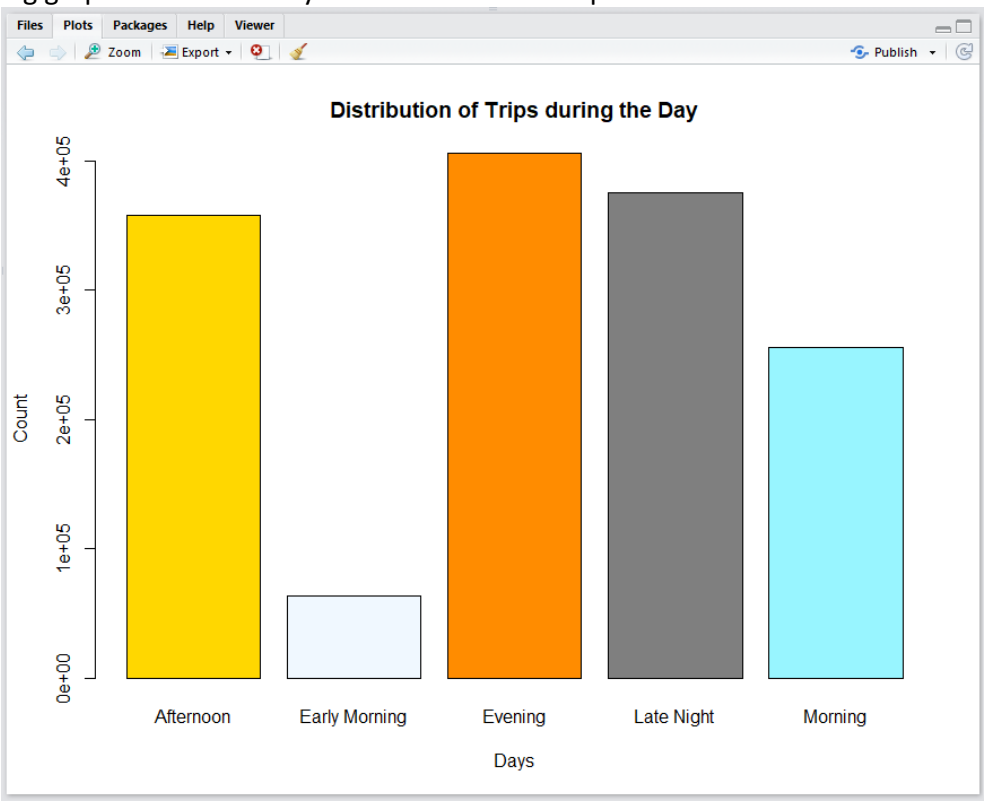

Figure 8

We have divided the time into 5 slots:

- 04 AM - 06 AM: Early Morning
- 07 AM - 10 AM: Morning
- 11 AM - 03 PM: Afternoon
- 04 PM - 08 PM: Evening
- 09 PM - 03 AM: Late Night

The most number of trips happen in afternoon, evening and Late Night. This shows that New-Yorkers use other modes of commuter services to travel to their workplace. Also, such a high number of evenings and nights suggest that New-Yorkers use taxi-trips for late-night parties.

The dataset considers that there are 2 vendors, vendor 1 and vendor 2.

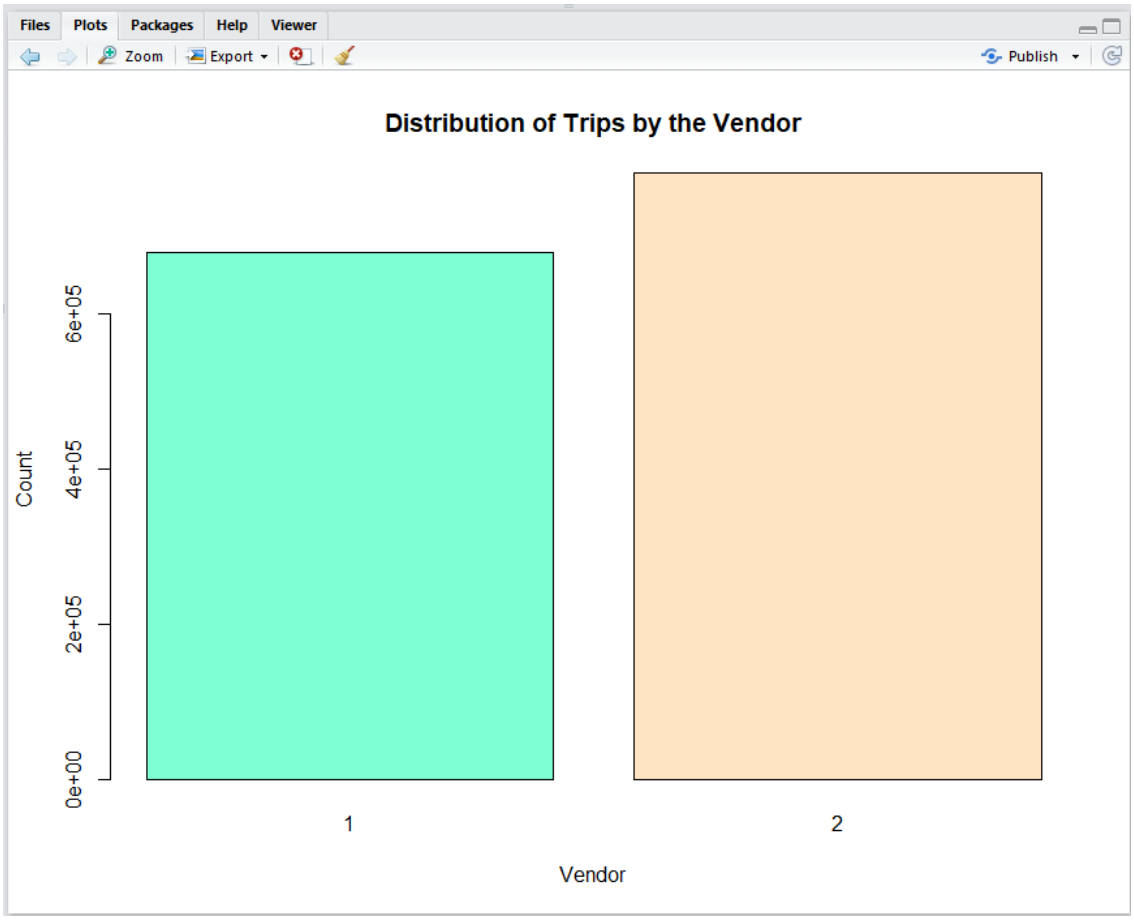

Figure 9

Taxis owned by Vendor 2 has more trips than those owned by vendor 1 which is shown in figure 9.

Store and Forward Flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. We plotted a graph to compare how many taxis use store and forward technique.

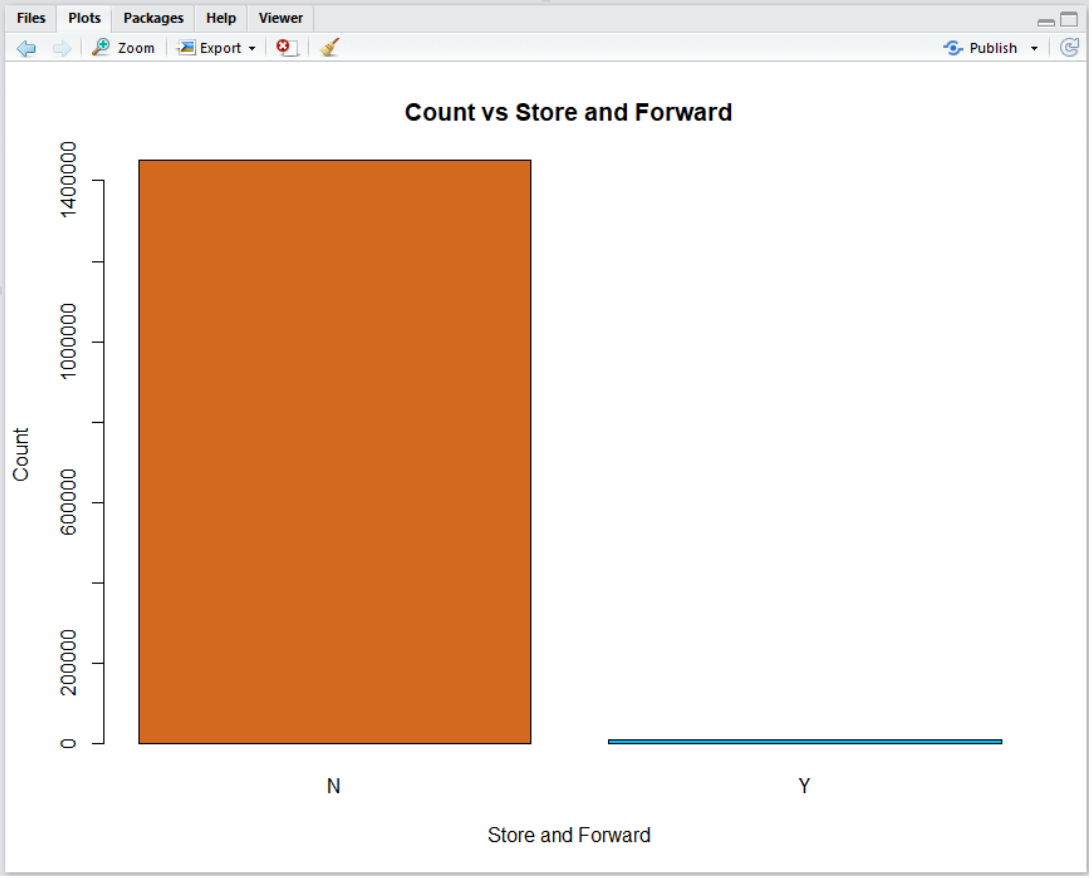

Figure 10

As seen in figure 10, the number of taxis using Store and forward technique is very less. Hence, this shows that most of the vehicles have an internet connection to connect to the server to record the trip.

References:

[1] Data Set: https://www.kaggle.com/c/nyc-taxi-trip-duration/data

[2] ShashankBhushan, 'Trip Visualization and Analysis', https://www.kaggle.com/monkeydunkey/trip-visualization-and-analysis, September 29, 2017

[3] Heads or Tails, 'NYC Taxi EDA - Update: The fast & the curious', https://www.kaggle.com/headsortails/nyc-taxi-eda-update-the-fast-the-curious, October 1, 2017