**Data Science Challenge: Dataset of "Green" Taxis in the New York City**

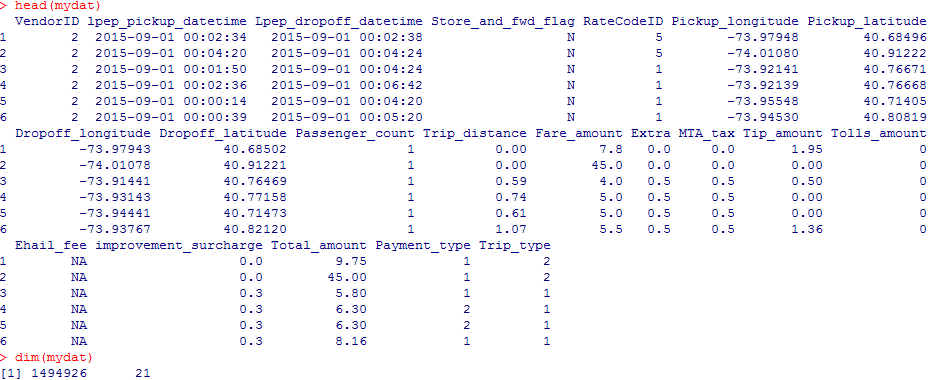
**Weiwei Ouyang**

**12/15/2016**

**Note: All work is finished in r code and the codes for 5 questions are attached with the question sheet.**

**Question 1:**

Firstly I load the Rcurl packages to extract the dataset from website directly. In R, mydat is the datasets containing the “Green” Taxi information.



**Figure 1.1. Head rows of the green taxi dataset**

In summary, there are 1494926 observations (rows) and 21 variables (Columns).

**Question 2:**

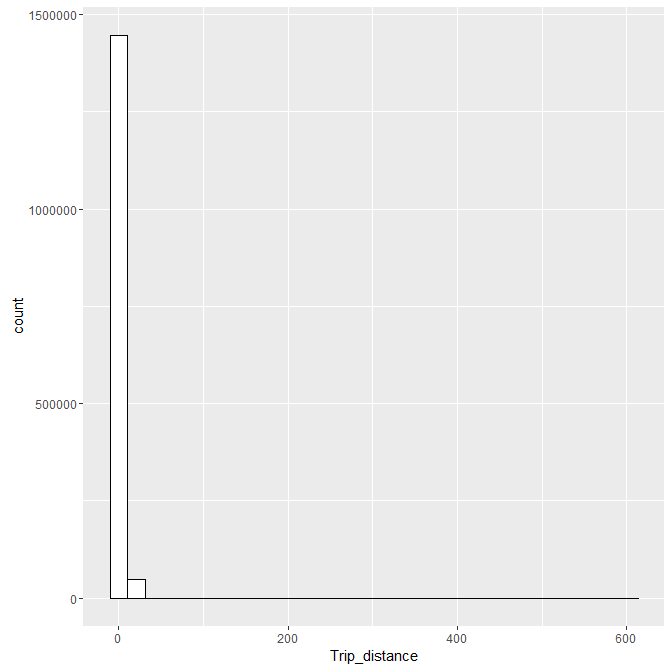
Before plotting the histogram of trip distance, I used the following R codes to check whether there is missing value, zeros and the summary of trip distances, respectively.

*length(which(is.na(mydat$Trip\_distance)))*

*length(which(mydat$Trip\_distance==0))*

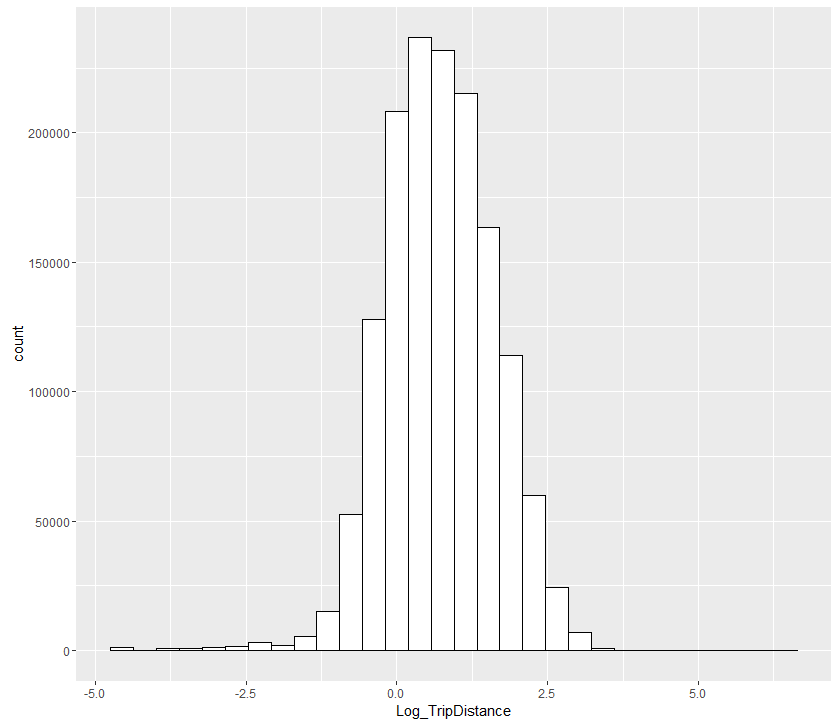
*summary(mydat$Trip\_distance)*

There is no missing value. However, there are 20592 observations with 0 trip distance. The smallest value of trip distance is zero and the largest one is 603.100. Then I loaded the ggplots package for plotting figures.



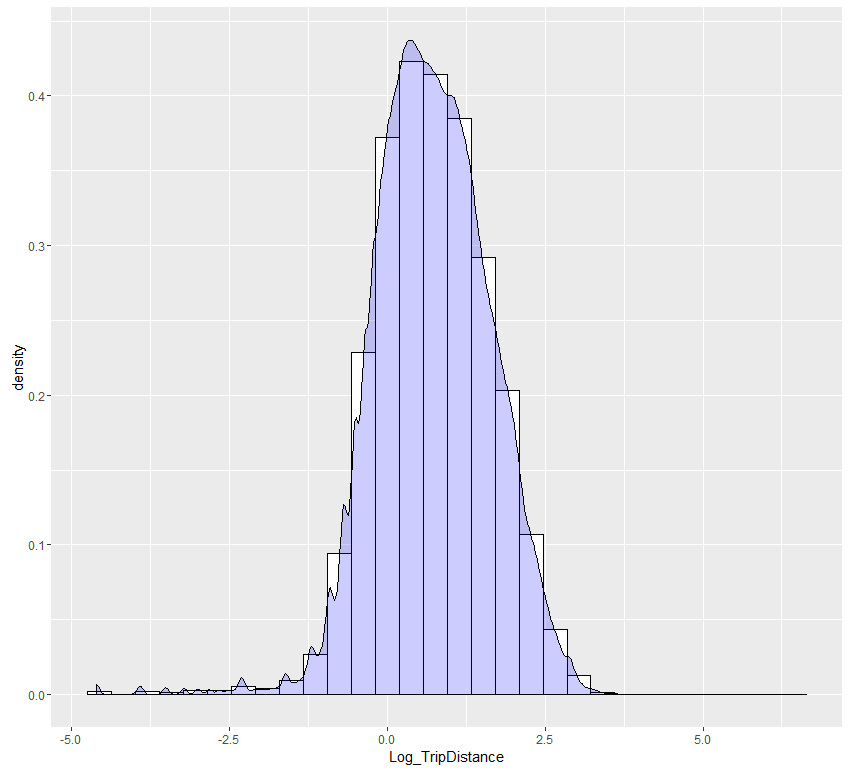
**Figure 2.1. Histogram of Trip Distance**

Seen from Figure 2.1, we observed skewed distribution and it is hard for us to fit the data with appropriate distributions. Herein I do log-transformation to the trip distance value and the histogram of Log(Trip Distance) is also plotted here in Figure 2.2.



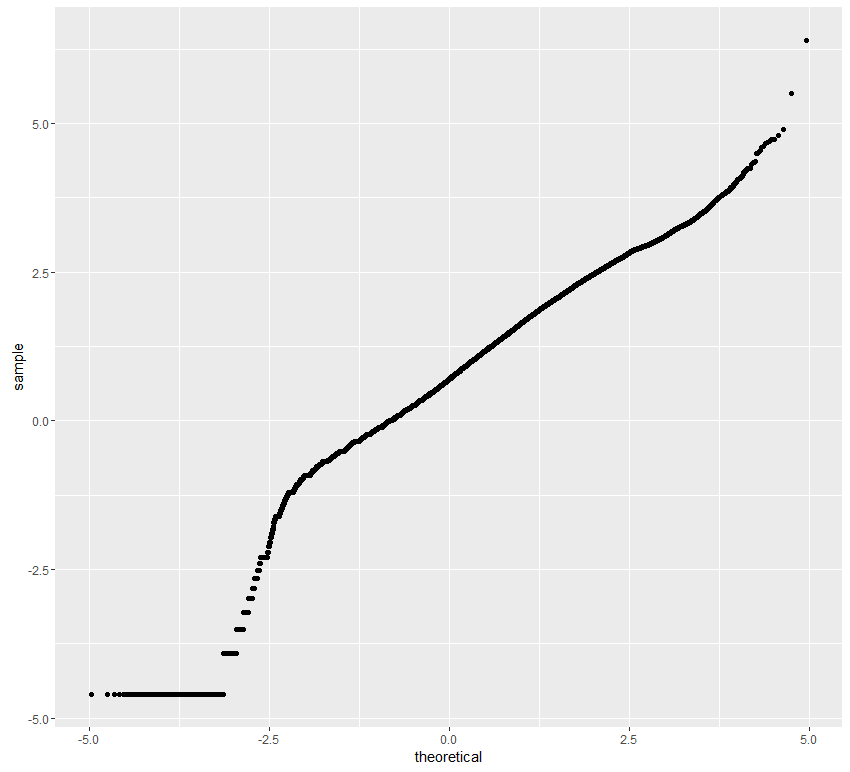
**Figure 2.2. Histogram of Log(Trip Distance)**

After the log-transformation, the distribution of Log(Trip Distance) is approximately like symmetric distribution. Figure 2.3 showed the density estimation for Log(Trip\_Distance).



**Figure 2.3. Density Estimation of Log(Trip Distance)**

The existence of 20592 zero values lead to the serious skewed problem. The QQ plot is also displayed here. The serious violation of normal distribution is mostly induced by the zero values. For the majority of observations, we can assume that they approximately follow normal distribution.

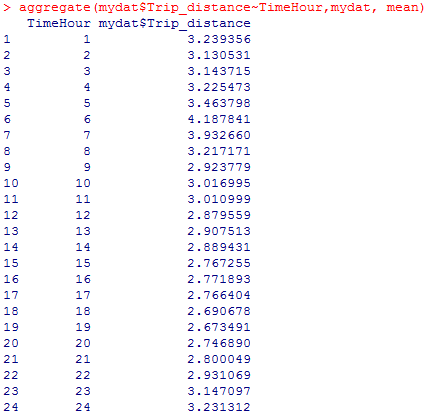


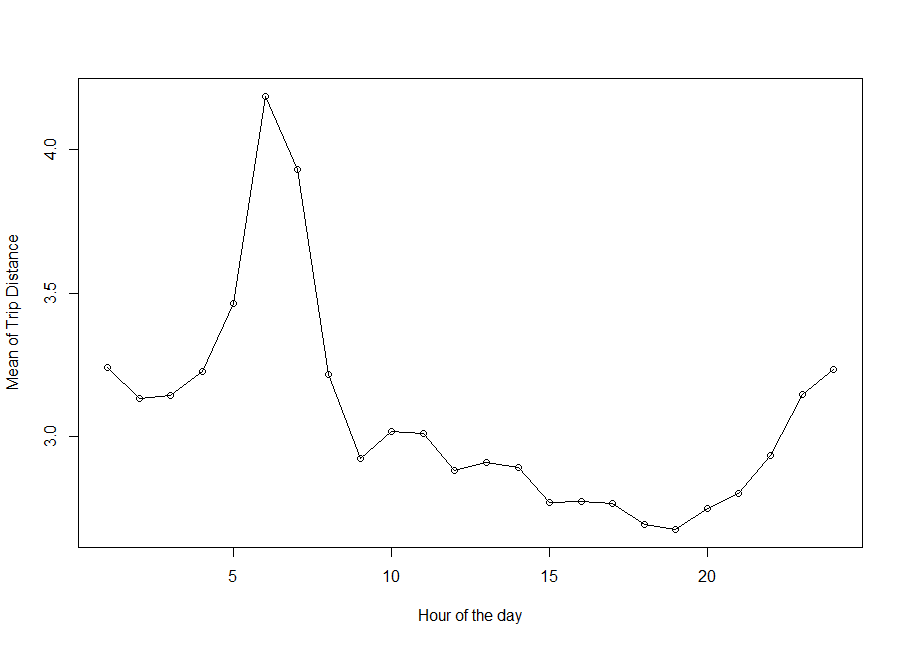
**Figure 2.4. Q-Q plot of Log(Trip Distance)**

**Question 3**

Firstly, we derive a new variable TimeHour to represent the drop-off hour of the day from 00:00-24:00. (TimeHour=1,2,…,24). Next we calculated the mean and median grouped by hour of day. The results by R is shown here:

*Mean grouped by hour of day*

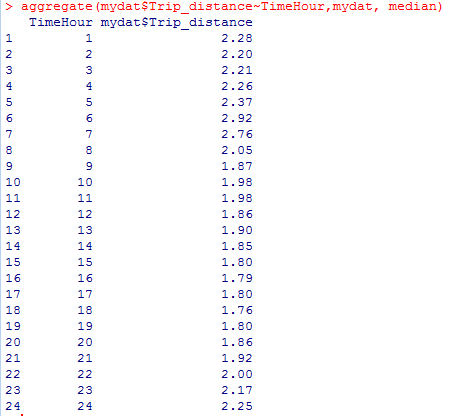


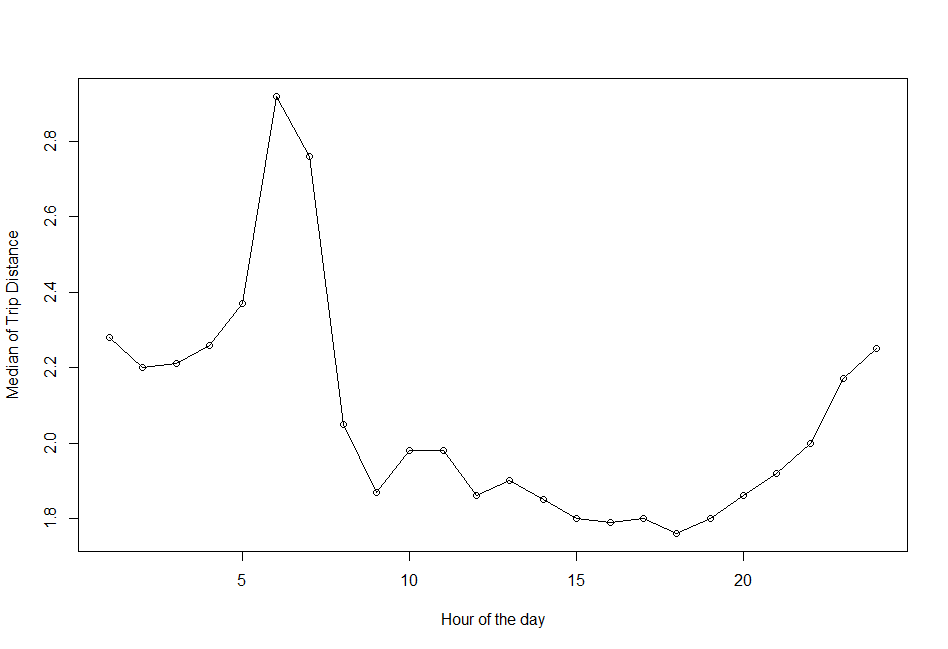


**Figure 3.1 Mean grouped by hour of the day**

It was observed from Figure 3.1 that there was a very clear increase peak in mean of the trip distance approximately from 5:00 AM to 6:00 AM. The minimum mean appeared around 18:00 (6:00 PM) to 19:00 (7:00 PM) at evening.

*Median grouped by hour of day*



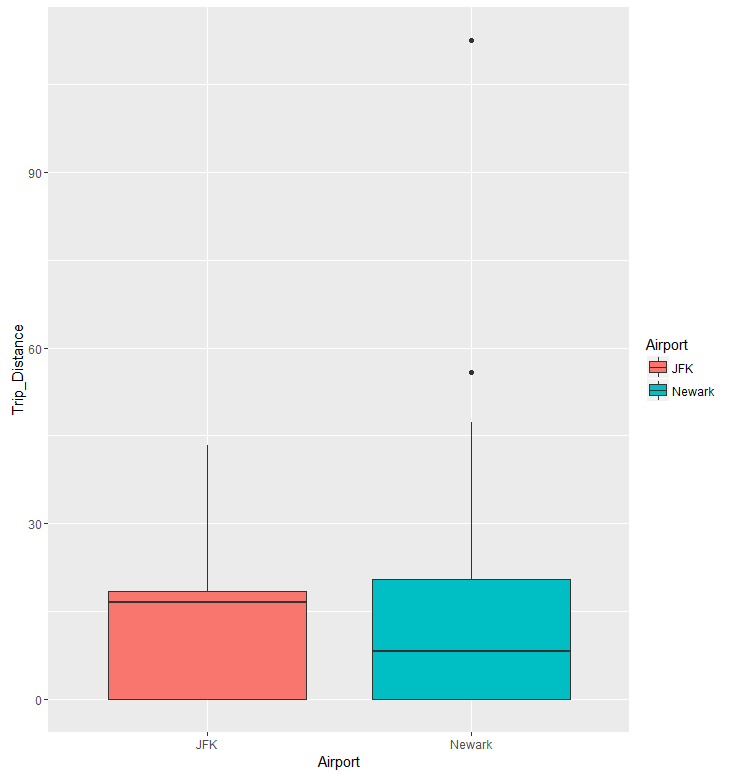


**Figure 3.2 Median grouped by hour of the day**

The results from median showed a similar result as that of mean.

There are two local airports in New York Region in this dataset: JFK and Newark. For variable RateCode ID, 2=JFK and 3=Newark. Calculated by R, there are 4435 transactions for JFK and the average fare is 49.02187. There are 1117 transactions for Newark and the average fare is 48.79857. The p value of welch t test between average fare of the two airport is 0.8198 that indicate there is no significant difference between average fare of JFK and that of Newwark.

Another interesting characteristic is the trip distance between the two airports. Figure 3.3 indicated that the transactions of JFK had a higher trip distance than Newark on average.



**Figure 3.3 Boxplot of Trip Distance between JFK and Newark**

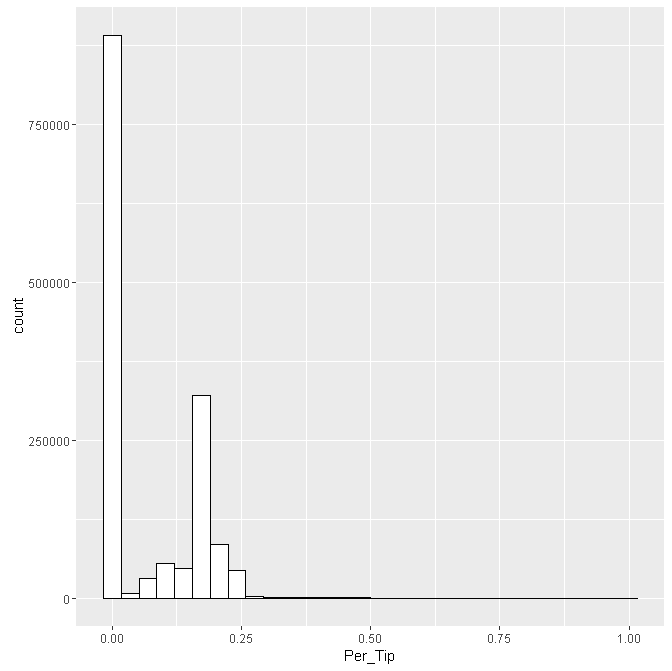
**Question 4**

Firstly we derived a new variable for tip as a percentage of the total fare and is defined as

This is the response variable we need to predict.

**Step 1. Data Cleaning and Manipulating**

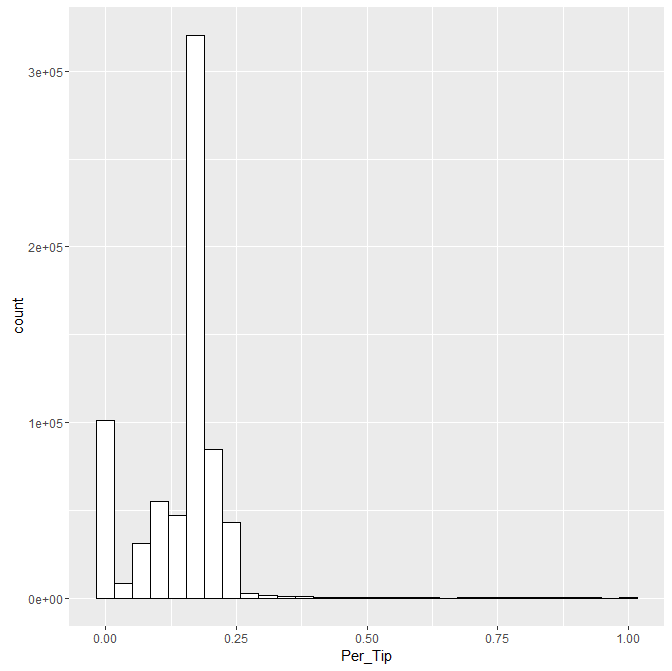
Before model building, we will have a look at the distribution of *Per\_Tip* by plotting a simple histogram as followings:



**Figure 4.1 Histogram of Percentage Tip for all trips**

Seen from Figure 4.1 most of the trips had a reported tip percentage of 0%. Among the trips with 0% tip, we found about 87.85% of the passengers paid with cash than credit card. Further analysis showed that the average tip percentage paid in cash was close to 0 () while the average tip percentage for trips paid with credit card was almost 14.15%. One of the possible reason for this phenomenon is that taxi drivers were more likely to report less tips then what they received when they were paid in cash. Cash pay is a way to evade tax and taxi companies will earn some of the tips if the trip was paid with credit card.

Therefore before building the predictive model analyzing taxi data, we would not choose the trips paid in cash or similar types that lead to serious sample bias and only focused on the trips paid by card. Figure 4.2 showed the histogram of percentage tip for trips paid with credit card.



**Figure 4.2 Histogram of Percentage Tip for trips with credit card**

Now the distribution of percentage tip was now much better. In total there are 701287 trips paid with credit card, among which 320 trips had 0 total amounts and their percentage tips went to infinity. In addition, we check the missing data for each variable and found no missing value except for variable Ehail\_fee. Thus we remove the 320 trips and 700967 trips underwent model building procedure.

**Step 2. Model Selection**

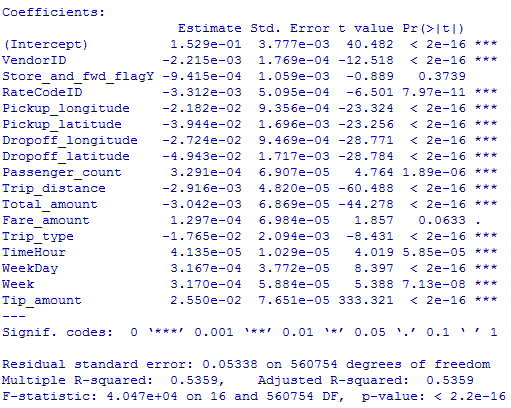
We need to select the appropriate variable in developing the predicting models.

|  |  |
| --- | --- |
| Dependent Variable | Description |
| Per\_Tip | Tip as a percentage |
| Possible Predictors | **Description** |
| VendorID | 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc. |
| Dropoff\_longitude | Longitude where the meter was timed off. |
| Dropoff\_ latitude | Latitude where the meter was timed off. |
| Pickup\_longitude | Longitude where the meter was engaged. |
| Pickup\_latitude | Latitude where the meter was engaged. |
| Passenger\_count | The number of passengers in the vehicle. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| RateCodeID | 1= Standard rate,2=JFK,3=Newark,4=Nassau or Westchester,5=Negotiated fare,6=Group ride |
| Store\_and\_fwd\_flag | Y= store and forward trip, N=not a store and forward trip |
| Fare\_amount | The time-and-distance fare calculated by the meter |
| Total\_amount | The total paid money |
| Trip\_type | 1= Street-hail, 2= Dispatch |
| Weekday | 1=Mon, 2=Tue,3=Wed,4=Thus,5=Friday,6=Sat,7=Sun |
| Week | 1=Week1, 2=Week2,3=Week3,4=Week4,5=week5 |
| TimeHour | 1~24 represent the hour interval by one hour per day |
| Tip\_amount | The paid tip amount |

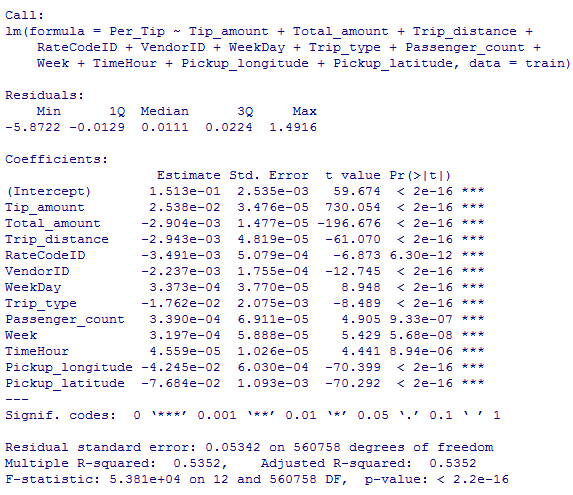
**Table 4.1 Description of response variable and predictors**

Table 4.1 give a general summary of the response variable and 14 predictors that are considered in statistical modeling. From the date information from the dataset, we create three new variables here: Weekday, Week and TimeHour to represent the day of the week, the week in September, 2015 and the time interval during the day, respectively. The aim of feature selection is to reduce the numbers of variables included in the model selection procedure.

We spitted the data and 80% is the train data while 20% is the test data. Here I adopted the AIC as measure to select the optimal model. In the test data, the mean square error (MSE) is compared with the variance of the training dataset. If MSE is smaller than the variance of Per\_tip in test dataset, the predictive model is considered as a good model. We also compared the correlation between estimated Per\_tip and Per\_tip of test datasets. Adjusted is also been compared.



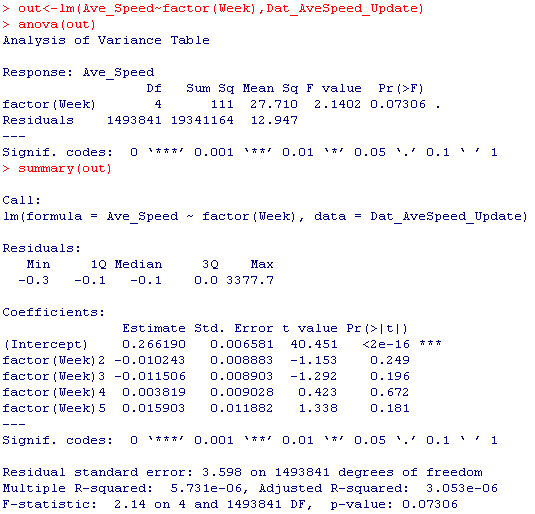
Only Store\_and\_fwd\_flagY and Fare\_amount is not significant by partial *t* test. For Fare\_amount, it has strong correlation with Total\_amount and the multicollinearity exists in this full model. Over half of the variance of Per\_tip can be explained by the full model. Then we conducted multiple regression to select the optimal model based on the smallest AIC. The results are as followings. All the variables are significant.



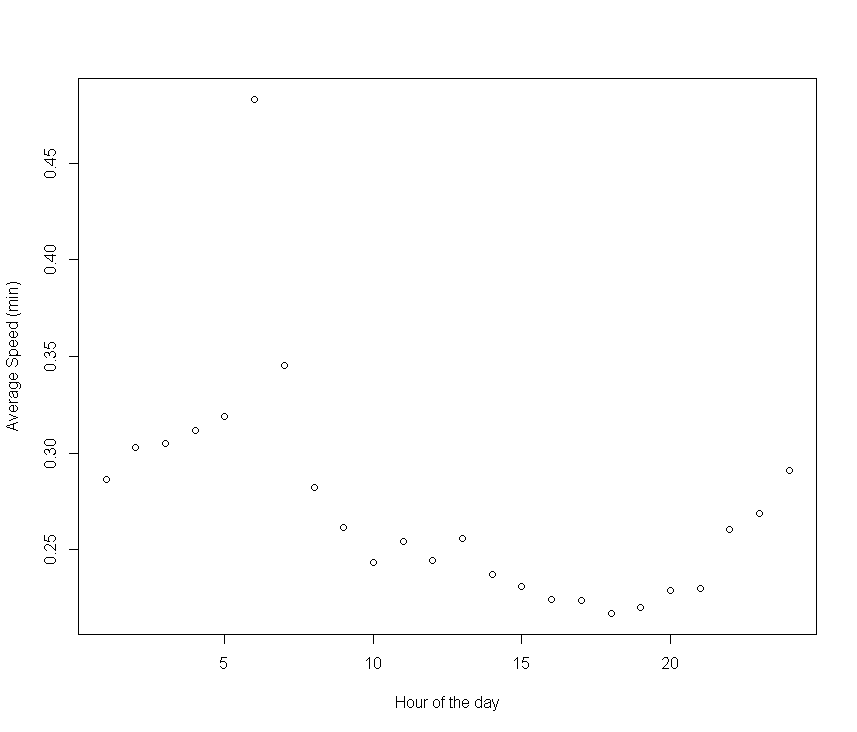
We applied this model to the test dataset and the MSE is 0.0031194, which is smaller than variance of Per\_tip 0.006143956 in the test model. And the correlation of predicted Per\_tip and the actual test Per\_tip is 0.7045. There are many 0 values in Per\_tip and it is hard to model 0 using the general linear models. For such 0 values, some zero-inflated models can be conducted.

**Question 5 (Optional A)**

To calculate the average speed, we firstly derive a new variable to represent the Trip time (mins). Then the average speed can be calculated as . 1080 trips was removed as missing value or Inf value. Then ANOVA analysis was conducted between average speed and the week of September (i.e.1~5).The results was shown as followings:

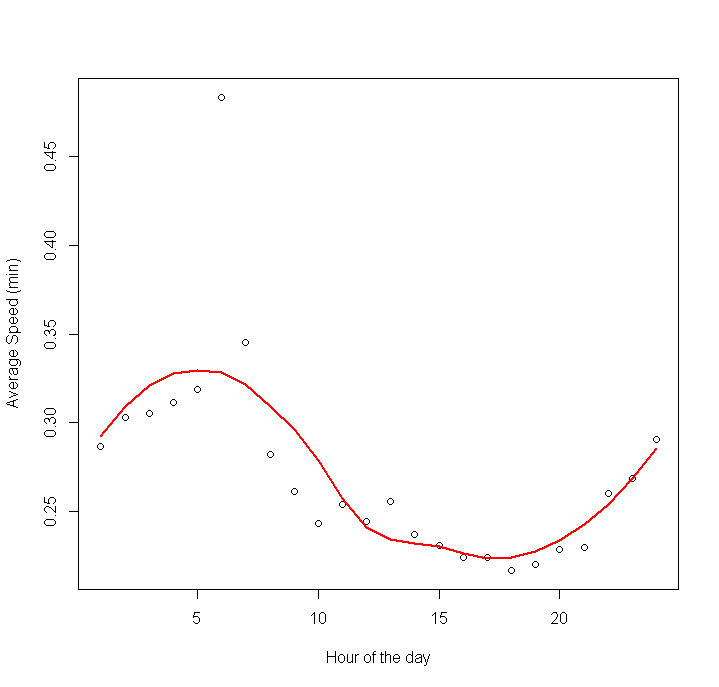


The p value of F test is 0.07306, which is larger than 0.05. Then we failed to reject that the average trip speeds are materially the same in all weeks of September. We can plot the mean of average speed grouped by hour of the day as followings.



**Figure 5.1 Mean of Speed grouped by hour of the day**

Seen from Figure5.1, there was a very clear increase in average speed until approximately 5:00 AM to 6:00 AM and reached the peak at almost 6:00 AM. The minimum average speed appeared round evening that is the busiest work-to-home peak time was. The average speed varies significantly over different hour of the day. Next using the loess() function in R, a smooth line is fitted to the points below:



**Figure 5.2 Fitted line of Mean of Speed grouped by hour of the day**

This fitted function clearly indicated the trend that is consistent with our conclusion that avrega speed is dependent on the hour of day. During the morning before work time, the traffic is not busy and the speed is higher while the speed will decrease significantly during the busy day time. And the speed would increase again at night.

**Discussion**

1. For predictive model, I only tried the stepwise model selection. The methods like random forest and Mallow , BIC can also be used for model selection.
2. Interaction can be included in the model.
3. Location variable is closely related. PCA regression can be applied as another alternative.