Regression\_Project\_Kirti\_Tiwari

Kirti Tiwari

10/20/2019

# Preprocessing

# Question 1 : Data Loading

rm(list=ls())  
library(rio)  
library(car)

## Loading required package: carData

setwd("~/USF/Kirti/USF CourseWork/2nd Sem/QMB-6304")  
taxitrips=import("6304 Regression Project Data.csv")  
colnames(taxitrips)=tolower(make.names(colnames(taxitrips)))  
attach(taxitrips)  
str(taxitrips)

## 'data.frame': 1705421 obs. of 9 variables:  
## $ taxi\_id : int 85 2776 3168 4237 5710 1987 4986 6400 7418 6450 ...  
## $ trip\_seconds: int 180 240 0 480 480 1080 1500 60 180 0 ...  
## $ trip\_miles : num 0.4 0.7 0 1.1 2.71 6.2 18.4 0.2 0 0 ...  
## $ fare : num 4.5 4.45 42.75 7 10.25 ...  
## $ tips : num 0 4.45 5 0 0 0 12 0 2 0 ...  
## $ tolls : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ extras : num 0 0 0 0 0 0 0 0 1.5 1.5 ...  
## $ trip\_total : num 4.5 8.9 47.8 7 10.2 ...  
## $ payment\_type: chr "Cash" "Credit Card" "Credit Card" "Cash" ...

names(taxitrips)

## [1] "taxi\_id" "trip\_seconds" "trip\_miles" "fare"   
## [5] "tips" "tolls" "extras" "trip\_total"   
## [9] "payment\_type"

# Question 2: Creating subset of 100 random records

set.seed(09448090)  
subset.taxitrips=taxitrips[sample(1:nrow(taxitrips),100,replace=FALSE),]

# Question 3 : Data Cleaning

# Checking for missing values if any.  
summary(subset.taxitrips)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 92 Min. : 0 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2267 1st Qu.: 180 1st Qu.: 0.000 1st Qu.: 5.750   
## Median :4946 Median : 480 Median : 0.700 Median : 8.125   
## Mean :4639 Mean : 1058 Mean : 2.312 Mean :12.159   
## 3rd Qu.:6990 3rd Qu.: 780 3rd Qu.: 1.925 3rd Qu.:13.162   
## Max. :8728 Max. :48900 Max. :18.500 Max. :49.500   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. :0.0 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.:0.0 1st Qu.: 6.75   
## Median : 0.000 Median :0 Median :0.0 Median : 9.25   
## Mean : 1.434 Mean :0 Mean :0.6 Mean :14.19   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.:1.0 3rd Qu.:15.49   
## Max. :10.550 Max. :0 Max. :5.0 Max. :60.00   
## payment\_type   
## Length:100   
## Class :character   
## Mode :character   
##   
##   
##

# If the mean of the column returns NA then the data is having NULL/ NA values. But since the means which my data is return are proper and not NA, so it means that the subset is not having any missing values.  
  
# Checking if there are any 0 values in trip\_seconds, because in ideal scenario trip\_seconds should not be zero.If a trip was taken or even started there should be some elapsed time associated with it.  
zero\_tripseconds = subset.taxitrips[which(subset.taxitrips$trip\_seconds == 0.0) ,]  
zero\_tripseconds

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 1338844 1569 0 0.0 30.50 6.10 0 0.0 36.60  
## 1325646 6960 0 0.0 8.45 2.00 0 0.0 10.45  
## 1146944 5462 0 0.0 46.50 9.30 0 0.0 55.80  
## 1314469 2594 0 0.0 5.75 1.44 0 0.0 7.19  
## 811358 4862 0 0.1 3.50 0.00 0 0.0 3.50  
## 197144 7518 0 0.0 11.00 2.20 0 0.0 13.20  
## 1688643 7779 0 0.0 10.00 3.00 0 0.0 13.00  
## 441 4201 0 0.0 8.50 1.70 0 0.0 10.20  
## 1506473 2663 0 0.0 3.25 0.00 0 0.0 3.25  
## 499407 332 0 0.0 10.50 2.10 0 0.0 12.60  
## 367534 2286 0 0.0 9.00 2.25 0 0.0 11.25  
## 1283651 947 0 0.0 49.50 0.00 0 0.0 49.50  
## 1443954 8728 0 0.0 3.25 0.00 0 0.5 3.75  
## 501132 6169 0 0.0 7.25 1.81 0 0.0 9.06  
## payment\_type  
## 1338844 Credit Card  
## 1325646 Credit Card  
## 1146944 Credit Card  
## 1314469 Credit Card  
## 811358 Cash  
## 197144 Credit Card  
## 1688643 Credit Card  
## 441 Credit Card  
## 1506473 Cash  
## 499407 Credit Card  
## 367534 Credit Card  
## 1283651 Credit Card  
## 1443954 Cash  
## 501132 Credit Card

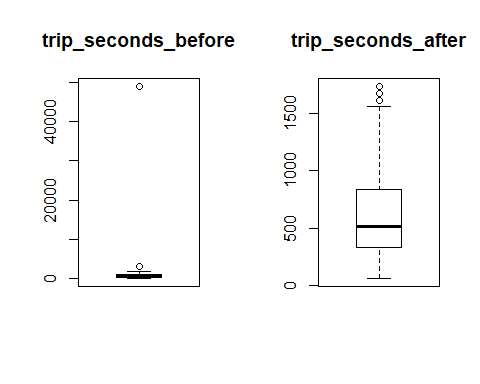
# 14 such records are found, which seems to be bad data, so I will remove them from my subset.  
nonzero.taxitrips = subset.taxitrips[-which(subset.taxitrips$trip\_seconds == 0.0) ,]  
  
# After removing 0 values from trip\_seconds, checking if there are any 0 values in trip\_miles.  
zero\_tripmiles = nonzero.taxitrips[which(nonzero.taxitrips$trip\_miles == 0.0 ),]  
zero\_tripmiles

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 1118698 3542 120 0 4.25 0.0 0 2.5 6.75  
## 724246 727 480 0 11.75 0.0 0 0.0 11.75  
## 1290555 176 420 0 5.50 0.0 0 0.0 5.50  
## 887603 5080 1020 0 15.00 3.0 0 1.0 19.00  
## 1130234 6004 60 0 3.25 0.0 0 0.0 3.25  
## 153138 2907 540 0 8.50 0.0 0 1.0 9.50  
## 58681 8554 300 0 6.00 2.0 0 1.0 9.00  
## 1669630 916 300 0 4.05 0.0 0 0.0 4.05  
## 737786 1386 480 0 8.25 3.0 0 0.0 11.25  
## 564209 7063 600 0 7.75 0.0 0 1.0 8.75  
## 327995 907 1620 0 43.50 9.5 0 4.0 57.00  
## 758439 2766 180 0 4.75 3.0 0 1.0 8.75  
## 209477 5796 360 0 5.75 0.0 0 0.0 5.75  
## 1677877 7695 240 0 5.75 2.0 0 0.0 7.75  
## 1116907 4847 48900 0 0.00 0.0 0 0.0 0.00  
## payment\_type  
## 1118698 Cash  
## 724246 Cash  
## 1290555 Cash  
## 887603 Credit Card  
## 1130234 Cash  
## 153138 Cash  
## 58681 Credit Card  
## 1669630 Cash  
## 737786 Credit Card  
## 564209 Cash  
## 327995 Credit Card  
## 758439 Credit Card  
## 209477 Cash  
## 1677877 Credit Card  
## 1116907 Cash

# Based on the data it looks like that the trip was actually started and there is some fare associated with it. It might be possible that the trip started and the taxi was kept on waiting and probably the trip was cancelled later. So, these cases seems to be the valid scenario, so I keep them as is and wont remove from my data.  
  
# For the remaining columns 0 value seems to be fine and valid.  
# So after removing 14 records for 0 trip\_seconds, we have 86 observations in the dataset now.  
  
# Checking outliers in the data.  
# Outliers for trip\_seconds  
  
par(mfrow=c(1,2))  
outliers\_ts =boxplot(nonzero.taxitrips$trip\_seconds,main="trip\_seconds\_before")$out  
outliers\_ts

## [1] 3120 48900

# We have 2 huge outlier values (48900,3120) which are way to high and because of those the model results will be impacted.  
  
# Removing outliers for trip\_seconds from the dataset  
clean.taxitrips = nonzero.taxitrips[-which(nonzero.taxitrips$trip\_seconds %in% outliers\_ts),]  
boxplot(clean.taxitrips$trip\_seconds,main="trip\_seconds\_after")$out

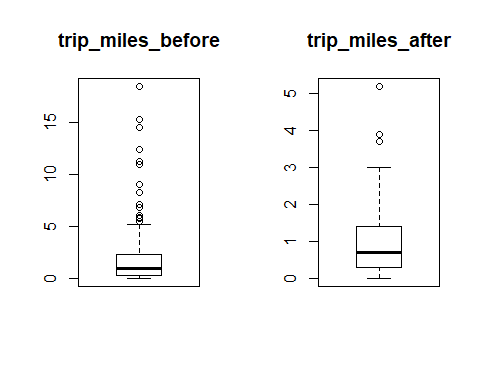


## [1] 1680 1680 1740 1620

par(mfrow=c(1,1))  
  
# Checking outliers for trip\_miles  
par(mfrow=c(1,2))  
outliers\_tm = boxplot(clean.taxitrips$trip\_miles,main="trip\_miles\_before")$out  
outliers\_tm

## [1] 18.5 5.5 6.1 14.5 5.8 12.4 7.1 11.3 5.9 15.3 8.3 6.8 6.8 9.0  
## [15] 11.0

# Removing outliers for trip\_miles from the dataset  
clean.taxitrips = clean.taxitrips[-which(clean.taxitrips$trip\_miles %in% outliers\_tm),]  
boxplot(clean.taxitrips$trip\_miles,main="trip\_miles\_after")$out

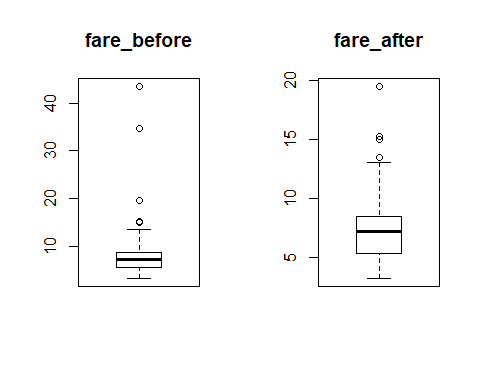


## [1] 5.2 3.7 3.9

par(mfrow=c(1, 1))  
  
# Checking outliers for fare  
par(mfrow=c(1,2))  
boxplot(clean.taxitrips$fare,main="fare\_before")$out

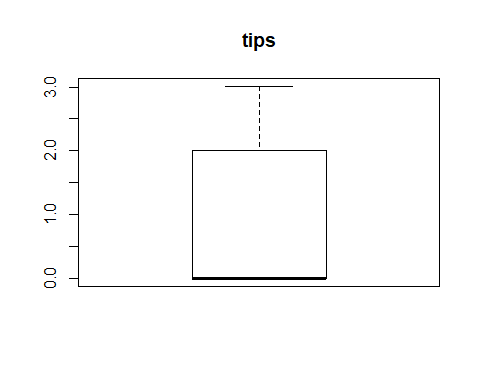
## [1] 34.75 15.00 19.50 15.00 15.25 43.50

# Removing 2 extreme outliers for fare from the dataset  
clean.taxitrips = clean.taxitrips[-which(clean.taxitrips$fare == 43.50),]  
clean.taxitrips = clean.taxitrips[-which(clean.taxitrips$fare == 34.75),]  
boxplot(clean.taxitrips$fare,main="fare\_after")$out



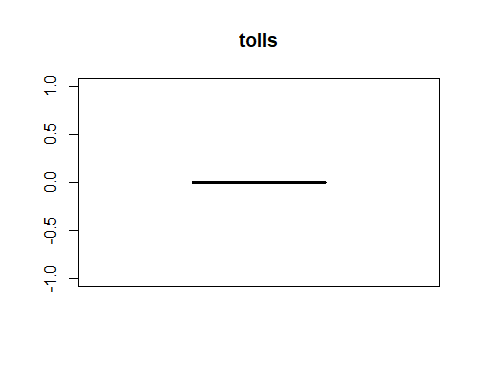
## [1] 15.00 19.50 15.00 15.25 13.50

par(mfrow=c(1, 1))  
  
# After cleaning Outliers from trip\_seconds, trip\_miles and fare, checking outliers for remaining variables.  
boxplot(clean.taxitrips$tips,main="tips")$out



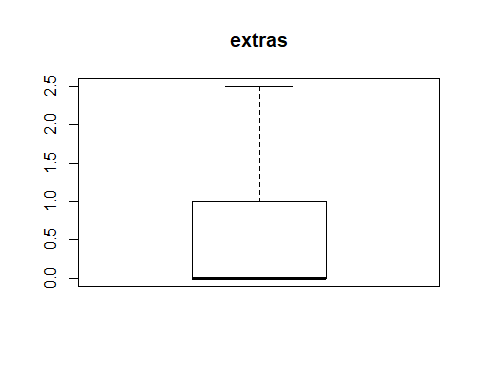
## numeric(0)

# No outliers for tips  
boxplot(clean.taxitrips$tolls,main="tolls")$out



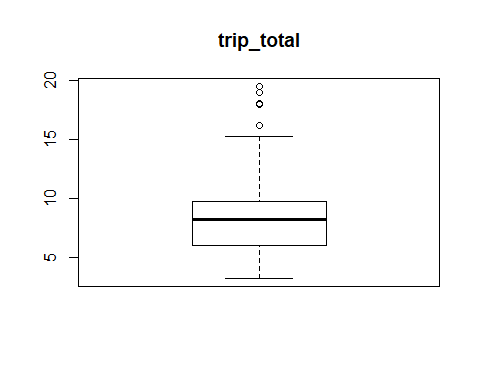
## numeric(0)

# No outliers for tolls  
boxplot(clean.taxitrips$extras,main="extras")$out



## numeric(0)

# No outliers for extras  
boxplot(clean.taxitrips$trip\_total,main="trip\_total")$out



## [1] 19.00 19.50 18.00 18.06 16.20

# 4 outliers for trip\_total but there are not very extreme, so I will keep them as is.

# So finally we are left with 67 observations which is the primary dataset for analysis

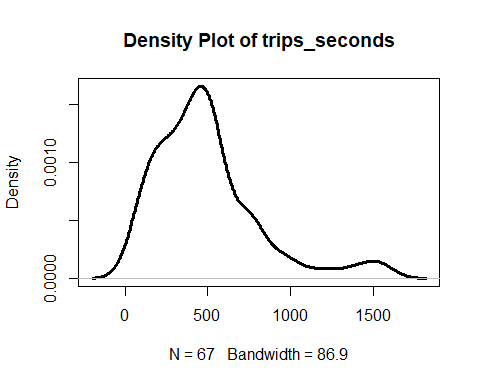
# Analysis Based on clean sample data

# Question 1: Summary & Density Plots for all continuous variables

# trip\_seconds  
summary(clean.taxitrips$trip\_seconds)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60.0 270.0 480.0 492.5 570.0 1560.0

plot(density(clean.taxitrips$trip\_seconds),lwd=3, main="Density Plot of trips\_seconds")

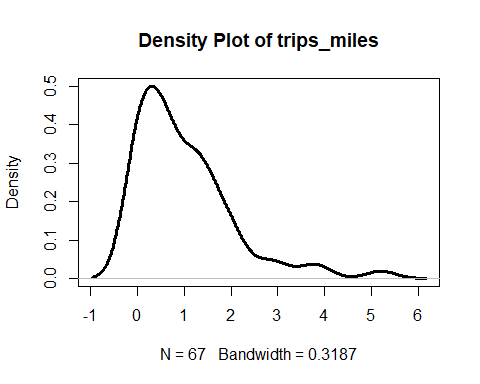


# Based on the above density plot for trip\_seconds ,the plot is approximately normal with most of the values concentrated around the mean (around 500) but it also has some extreme values (>1000) at the tail in right and hence it is right skewed rather than normally distributed

#trip\_miles  
summary(clean.taxitrips$trip\_miles)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.300 0.700 1.012 1.400 5.200

plot(density(clean.taxitrips$trip\_miles),lwd=3, main="Density Plot of trips\_miles")

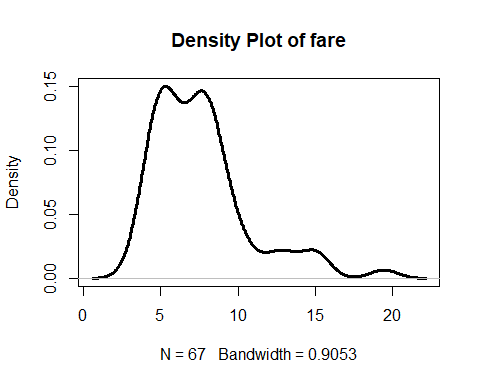


# The trip\_miles density plot shows majority of values are accumulated below mean (~1) and has many extreme values present at the tail on the right side, so it is also skewed right

# fare  
summary(clean.taxitrips$fare)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.250 5.375 7.250 7.598 8.500 19.500

plot(density(clean.taxitrips$fare),lwd=3, main="Density Plot of fare")

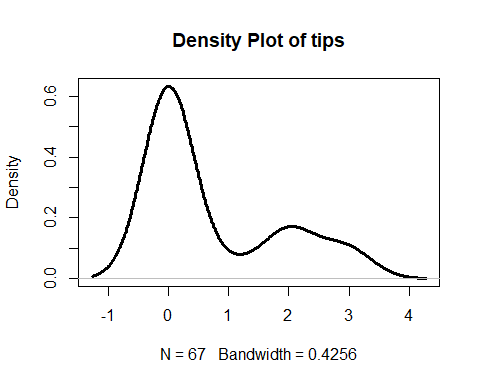


# The density plot for fare is bimodal ( i.e. 2 peaks) ,with the tail extending towards right and hence skewed towards right

# tips  
summary(clean.taxitrips$tips)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.7158 2.0000 3.0100

plot(density(clean.taxitrips$tips),lwd=3, main="Density Plot of tips")



# The density plot for tips is also bimodal but the one of the peaks is much lower than the other. Most of the values are concentrated around the 1st peak which are less than 1 making model look like more normally distributed, but we have another peak at around 2 and some extreme values after that which could possibly can be considered as outliers and because of those the model is skewed towards right.

# tolls  
summary(clean.taxitrips$tolls)

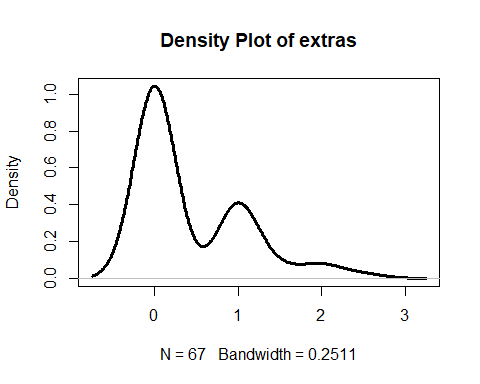
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 0 0 0

# There is no data in tolls so there is no use of generating the denisty plot for it.

# extras  
summary(clean.taxitrips$extras)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.4254 1.0000 2.5000

plot(density(clean.taxitrips$extras),lwd=3, main="Density Plot of extras")

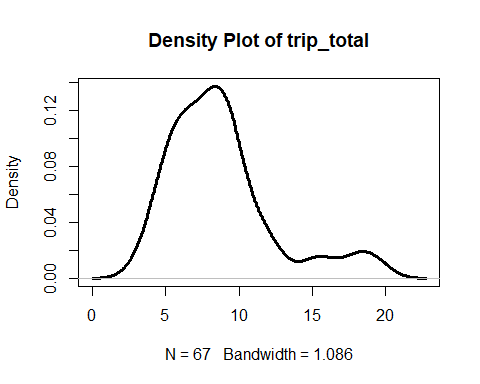


# Similary for extras, based on the density plot it is bimodel as it has 2 peaks and because of more values concentrated towards right, it is skewed right

# trip\_total  
summary(clean.taxitrips$trip\_total)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.250 6.000 8.250 8.739 9.750 19.500

plot(density(clean.taxitrips$trip\_total),lwd=3, main="Density Plot of trip\_total")



# The density plot for trip\_total is also bimodal but the one of the peaks is much lower than the other. Most of the values are concentrated around the 1st peak which equal to the mean (~8.7) making model look like more normally distributed and we also some extreme values after that because of which model is skewed more towards right.

# Question 2: number of cases in each level of payment\_type

# can use library plyr to get the count of the each cases   
library(plyr)  
count(clean.taxitrips, vars=c("payment\_type"))

## payment\_type freq  
## 1 Cash 45  
## 2 Credit Card 22

# can use table function to get the count  
table(clean.taxitrips$payment\_type)

##   
## Cash Credit Card   
## 45 22

# Question 3: correlation matrix using all continuous variables except taxi\_id. Also removing payment\_type as it is categorical variable and tolls because there is no data

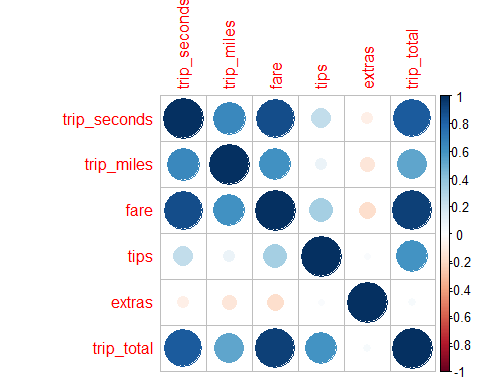
library(corrplot)

## corrplot 0.84 loaded

cor(clean.taxitrips[-c(1,6,9)])

## trip\_seconds trip\_miles fare tips extras  
## trip\_seconds 1.00000000 0.64497155 0.8890960 0.24591046 -0.08721228  
## trip\_miles 0.64497155 1.00000000 0.6057883 0.08199138 -0.13928517  
## fare 0.88909602 0.60578828 1.0000000 0.33194314 -0.17025190  
## tips 0.24591046 0.08199138 0.3319431 1.00000000 0.02924469  
## extras -0.08721228 -0.13928517 -0.1702519 0.02924469 1.00000000  
## trip\_total 0.83020395 0.52564722 0.9376142 0.59483524 0.03903476  
## trip\_total  
## trip\_seconds 0.83020395  
## trip\_miles 0.52564722  
## fare 0.93761424  
## tips 0.59483524  
## extras 0.03903476  
## trip\_total 1.00000000

corrplot(cor(clean.taxitrips[-c(1,6,9)]))



#A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. Bigger and darker coloured circles indicate higher correlation whereas smaller and lighter coloured circles indicate lower correlation. Blue is for positive and red colour represents negative correlation.The diagonal can be ignored as it shows the correlation of a variable with itself.So,the diagonal has dark blue with large size indicating a correlation of 1(The variable with itself). Further we can see the next bigger and darker circle is between fare and trip\_total and it shows they are highly correlated followed by circle between trip\_seconds and fare. As the circle becomes smaller lighter blue in color it means they don't have strong correlation between them. For ex: the circle between tips and trip\_miles is very small and too faded, which means they both are very weakly correlated. There are some circles which have red color, for ex : between trip\_seconds and extras, which means correlation between them is negative. And if the value of one increase then others will decrease. The blue ones are positively correlated and if value of one increases others will also increase and the intensity of increase will depend on the strength of the correlation.

# Question 4: Using fare as the dependent variable,build a regression model using trip\_seconds, trip\_miles, and payment\_type as potential independent variables

taxitrips.out=lm(fare~trip\_seconds+trip\_miles+payment\_type, data=clean.taxitrips)  
summary(taxitrips.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + trip\_miles + payment\_type,   
## data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2127 -0.7785 -0.2551 0.3364 4.7517   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.1738797 0.3336216 9.513 8.45e-14 \*\*\*  
## trip\_seconds 0.0082096 0.0007294 11.256 < 2e-16 \*\*\*  
## trip\_miles 0.1751904 0.2230191 0.786 0.435   
## payment\_typeCredit Card 0.6183777 0.3802898 1.626 0.109   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.442 on 63 degrees of freedom  
## Multiple R-squared: 0.8007, Adjusted R-squared: 0.7912   
## F-statistic: 84.34 on 3 and 63 DF, p-value: < 2.2e-16

Interpretation:

The regression equation will be: fare= 3.174+0. 0.0082*trip\_seconds +0.175*trip\_miles +0.618\*payment\_type

Based on the above results, the intercept and trip\_seconds are significant (as p-values < 0.05) whereas trip\_miles and payment\_type are insignificant (p-value=0.435 and 0.109 respectively which are greater than 0.05). We may say that based on this model might not be a strong relationship between fare and trip\_miles and payment\_type.

For Intercept :

Based on the model the beta coefficient is 3.1738797 and p-value is 8.45e-14 which is way too lower than .05 (approx. to 0) so we can reject the null hypothesis (intercept = 0) as based on the p-value we are sure that intercept is not equal to 0

For Slopes :

trip\_seconds : As the slope is positive, there is positive correlation between fare and trip\_seconds. Everytime with the increase of 1000 seconds there will be increase in fare by 8.2 dollars.As the p-value is 2e-16 way to less than 0.05 , so we can reject the null hypothesis (slope = 0) and trip\_seconds is having very significant impact on fare.

trip\_miles:

Here also, the slope is positive, so we have positive correlation between trip\_miles and fare. But it is not having much significant contribution to the model as p-value is >0.05

payment\_type:

Here also, the slope is positive, so we have positive correlation between payment\_type and fare. But it is not having much significant contribution to the model as p-value is >0.05

confint(taxitrips.out)

## 2.5 % 97.5 %  
## (Intercept) 2.507190059 3.8405693  
## trip\_seconds 0.006752057 0.0096671  
## trip\_miles -0.270477847 0.6208586  
## payment\_typeCredit Card -0.141570885 1.3783262

As shown above p-value for the coefficients for intercept and trip\_seconds are < 0.05 so they are rejecting the null hypothesis and we can say based on their confidence intervals none of them is crossing 0 and there is significant impact of these coefficient on the charges.

• For a trip with zero trip\_seconds, zero trip\_miles and payment\_type, fare could be anywhere from 2.5 to 3.8 dollars.

• For every 1000 times increase in trip\_seconds, there could be a increment in fare which will vary anywhere from 6.7 to 9.6 dollars.

• For every 10 miles increase in trip\_miles, fare can vary between -2.7 to 6.2 dollars. But it’s insignificant according to the model.

• Depending on the customer is using credit card as payment\_type or not the fare will vary between -0.14 to 1.3 dollars.But it’s insignificant according to the model.

# Question :5 Investigating relevant interactions.

# Using Kitchen Sink Model: we will use all variable except taxi\_id  
taxitrips\_kit.out=lm(fare~. -taxi\_id , data=clean.taxitrips)  
summary(taxitrips\_kit.out)

## Warning in summary.lm(taxitrips\_kit.out): essentially perfect fit: summary  
## may be unreliable

##   
## Call:  
## lm(formula = fare ~ . - taxi\_id, data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.017e-14 1.415e-16 5.502e-16 7.781e-16 1.648e-15   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.196e-16 1.284e-15 -1.710e-01 0.86482   
## trip\_seconds -7.732e-18 2.694e-18 -2.870e+00 0.00566 \*\*   
## trip\_miles 1.202e-16 4.901e-16 2.450e-01 0.80712   
## tips -1.000e+00 1.242e-15 -8.050e+14 < 2e-16 \*\*\*  
## tolls NA NA NA NA   
## extras -1.000e+00 6.027e-16 -1.659e+15 < 2e-16 \*\*\*  
## trip\_total 1.000e+00 2.811e-16 3.558e+15 < 2e-16 \*\*\*  
## payment\_typeCredit Card -3.625e-16 2.515e-15 -1.440e-01 0.88587   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.027e-15 on 60 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.195e+31 on 6 and 60 DF, p-value: < 2.2e-16

AIC(taxitrips\_kit.out)

## [1] -4281.054

As trip\_total has very high correlation with fare, so using it in the model is giving a warning messgae (“Warning message: In summary.lm(taxitrips\_kit.out) : essentially perfect fit: summary may be unreliable”) and it is making the model way to perfect to analyze. So will remove trip\_total column as well

# removing taxi\_id and trip\_total  
taxitrips\_kit.out=lm(fare~. -taxi\_id -trip\_total, data=clean.taxitrips)  
summary(taxitrips\_kit.out)

##   
## Call:  
## lm(formula = fare ~ . - taxi\_id - trip\_total, data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5654 -0.7643 -0.1707 0.3309 4.5001   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6088036 0.3589471 10.054 1.45e-14 \*\*\*  
## trip\_seconds 0.0075857 0.0007502 10.112 1.16e-14 \*\*\*  
## trip\_miles 0.2558779 0.2208494 1.159 0.2511   
## tips 1.1292647 0.4959141 2.277 0.0263 \*   
## tolls NA NA NA NA   
## extras -0.5234240 0.2677010 -1.955 0.0551 .   
## payment\_typeCredit Card -1.8027047 1.1222574 -1.606 0.1134   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.379 on 61 degrees of freedom  
## Multiple R-squared: 0.8235, Adjusted R-squared: 0.809   
## F-statistic: 56.91 on 5 and 61 DF, p-value: < 2.2e-16

AIC(taxitrips\_kit.out)

## [1] 240.893

#Adjusted R-square came down to 80.9% from 100% previously. We will also remove the tolls as it doesn't have any value and there is no impact of that column.  
  
# removing taxi\_id, trip\_total and tolls  
taxitrips\_kit.out=lm(fare~. -taxi\_id -trip\_total-tolls, data=clean.taxitrips)  
summary(taxitrips\_kit.out)

##   
## Call:  
## lm(formula = fare ~ . - taxi\_id - trip\_total - tolls, data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5654 -0.7643 -0.1707 0.3309 4.5001   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6088036 0.3589471 10.054 1.45e-14 \*\*\*  
## trip\_seconds 0.0075857 0.0007502 10.112 1.16e-14 \*\*\*  
## trip\_miles 0.2558779 0.2208494 1.159 0.2511   
## tips 1.1292647 0.4959141 2.277 0.0263 \*   
## extras -0.5234240 0.2677010 -1.955 0.0551 .   
## payment\_typeCredit Card -1.8027047 1.1222574 -1.606 0.1134   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.379 on 61 degrees of freedom  
## Multiple R-squared: 0.8235, Adjusted R-squared: 0.809   
## F-statistic: 56.91 on 5 and 61 DF, p-value: < 2.2e-16

AIC(taxitrips\_kit.out)

## [1] 240.893

After applying Kitchen Sink Model and excluding 3 columns taxi\_id, trip\_total and tolls, we are getting Adjusted R-square of 80.9% and the terms which have significant impact are Intercept, trip\_seconds and tips

#Stepwise Regression (both direction):  
#We will apply stepwise regression first to check the model performance and see which variable have significant impact.  
taxitrips\_step.out=step(lm(fare~. -taxi\_id -trip\_total-tolls, data=clean.taxitrips),direction="both")

## Start: AIC=48.76  
## fare ~ (taxi\_id + trip\_seconds + trip\_miles + tips + tolls +   
## extras + trip\_total + payment\_type) - taxi\_id - trip\_total -   
## tolls  
##   
## Df Sum of Sq RSS AIC  
## - trip\_miles 1 2.552 118.52 48.214  
## <none> 115.96 48.755  
## - payment\_type 1 4.905 120.87 49.531  
## - extras 1 7.268 123.23 50.828  
## - tips 1 9.858 125.82 52.221  
## - trip\_seconds 1 194.384 310.35 112.711  
##   
## Step: AIC=48.21  
## fare ~ trip\_seconds + tips + extras + payment\_type  
##   
## Df Sum of Sq RSS AIC  
## <none> 118.52 48.214  
## - payment\_type 1 3.66 122.18 48.251  
## + trip\_miles 1 2.55 115.96 48.755  
## - extras 1 7.93 126.45 50.553  
## - tips 1 8.08 126.59 50.631  
## - trip\_seconds 1 402.17 520.69 145.381

summary(taxitrips\_step.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type,   
## data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5233 -0.7285 -0.2730 0.4588 4.2344   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5975990 0.3598062 9.999 1.5e-14 \*\*\*  
## trip\_seconds 0.0081624 0.0005627 14.505 < 2e-16 \*\*\*  
## tips 0.9930768 0.4831106 2.056 0.0440 \*   
## extras -0.5453826 0.2677655 -2.037 0.0459 \*   
## payment\_typeCredit Card -1.5197176 1.0983763 -1.384 0.1714   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.383 on 62 degrees of freedom  
## Multiple R-squared: 0.8196, Adjusted R-squared: 0.808   
## F-statistic: 70.42 on 4 and 62 DF, p-value: < 2.2e-16

vif(taxitrips\_step.out)

## trip\_seconds tips extras payment\_type   
## 1.130198 9.686265 1.036231 9.325634

#Stepwise Regression (backward direction):  
taxitrips\_step\_b.out=step(lm(fare~. -taxi\_id -trip\_total-tolls, data=clean.taxitrips),direction="backward")

## Start: AIC=48.76  
## fare ~ (taxi\_id + trip\_seconds + trip\_miles + tips + tolls +   
## extras + trip\_total + payment\_type) - taxi\_id - trip\_total -   
## tolls  
##   
## Df Sum of Sq RSS AIC  
## - trip\_miles 1 2.552 118.52 48.214  
## <none> 115.96 48.755  
## - payment\_type 1 4.905 120.87 49.531  
## - extras 1 7.268 123.23 50.828  
## - tips 1 9.858 125.82 52.221  
## - trip\_seconds 1 194.384 310.35 112.711  
##   
## Step: AIC=48.21  
## fare ~ trip\_seconds + tips + extras + payment\_type  
##   
## Df Sum of Sq RSS AIC  
## <none> 118.52 48.214  
## - payment\_type 1 3.66 122.18 48.251  
## - extras 1 7.93 126.45 50.553  
## - tips 1 8.08 126.59 50.631  
## - trip\_seconds 1 402.17 520.69 145.381

summary(taxitrips\_step\_b.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type,   
## data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5233 -0.7285 -0.2730 0.4588 4.2344   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5975990 0.3598062 9.999 1.5e-14 \*\*\*  
## trip\_seconds 0.0081624 0.0005627 14.505 < 2e-16 \*\*\*  
## tips 0.9930768 0.4831106 2.056 0.0440 \*   
## extras -0.5453826 0.2677655 -2.037 0.0459 \*   
## payment\_typeCredit Card -1.5197176 1.0983763 -1.384 0.1714   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.383 on 62 degrees of freedom  
## Multiple R-squared: 0.8196, Adjusted R-squared: 0.808   
## F-statistic: 70.42 on 4 and 62 DF, p-value: < 2.2e-16

vif(taxitrips\_step\_b.out)

## trip\_seconds tips extras payment\_type   
## 1.130198 9.686265 1.036231 9.325634

Based on the output of the stepwise regression for both and backward direction we got same Adjusted R-square which is 80.8% and the terms which are significant are trips\_seconds,tips and extras.Last model (for both and backward) is giving least AIC, so we can say it is best fit model as of now.

# We try to see the model output after sqauring the terms.  
# Squaring trip\_miles first  
  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2586 -0.7029 -0.2435 0.5483 4.1687   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.1874213 0.4162762 10.059 1.71e-14 \*\*\*  
## trip\_seconds 0.0070706 0.0007497 9.431 1.87e-13 \*\*\*  
## tips 1.0074737 0.4787574 2.104 0.0395 \*   
## extras -0.5918026 0.2585508 -2.289 0.0256 \*   
## payment\_typeCredit Card -1.4396761 1.0876151 -1.324 0.1906   
## trip\_miles -0.5843757 0.3998020 -1.462 0.1491   
## I(trip\_miles^2) 0.2496647 0.1007014 2.479 0.0160 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.324 on 60 degrees of freedom  
## Multiple R-squared: 0.8399, Adjusted R-squared: 0.8239   
## F-statistic: 52.45 on 6 and 60 DF, p-value: < 2.2e-16

AIC(taxitrips\_sq.out)

## [1] 236.3584

vif(taxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 2.187469 10.372044 1.053438 9.970034   
## trip\_miles I(trip\_miles^2)   
## 6.527611 6.931769

# Adjusted R-square increased a bit to 82.3% and sqaured trip\_miles term is also having some signifance impact. Also, based on the vif we can see that it is low and we dont have multicollinearity between the terms which is good because ideally vif should be less than 10 for a good model  
# Introducing cube to the equation  
  
taxitrips\_cu.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(trip\_miles^3), data = clean.taxitrips)  
summary(taxitrips\_cu.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(trip\_miles^3), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2848 -0.7049 -0.2514 0.5527 4.1438   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.2157290 0.4604967 9.155 6.31e-13 \*\*\*  
## trip\_seconds 0.0070635 0.0007574 9.325 3.29e-13 \*\*\*  
## tips 0.9916405 0.4942032 2.007 0.0494 \*   
## extras -0.5918000 0.2606834 -2.270 0.0269 \*   
## payment\_typeCredit Card -1.4138767 1.1100998 -1.274 0.2078   
## trip\_miles -0.6899309 0.8134304 -0.848 0.3998   
## I(trip\_miles^2) 0.3213633 0.4905320 0.655 0.5149   
## I(trip\_miles^3) -0.0106645 0.0713820 -0.149 0.8817   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.335 on 59 degrees of freedom  
## Multiple R-squared: 0.8399, Adjusted R-squared: 0.821   
## F-statistic: 44.23 on 7 and 59 DF, p-value: < 2.2e-16

AIC(taxitrips\_cu.out)

## [1] 238.3331

vif(taxitrips\_cu.out)

## trip\_seconds tips extras payment\_type   
## 2.196226 10.872001 1.053438 10.217280   
## trip\_miles I(trip\_miles^2) I(trip\_miles^3)   
## 26.580927 161.797908 72.607434

# Introducing cubic term has not changed Adjusted R-square much but it has removed the impact of sqaured trip\_miles term also. Also, vif for some terms has increased way to high which tells us that there is high multicorrelation between those terms.So will remove cubic term and will try squaring the other terms like tips and extras  
  
# squaring tips terms along with miles  
  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(tips^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(tips^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2478 -0.6937 -0.0864 0.4636 4.1834   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.1832328 0.4176049 10.017 2.41e-14 \*\*\*  
## trip\_seconds 0.0070487 0.0007526 9.366 2.82e-13 \*\*\*  
## tips -1.0650220 2.6582696 -0.401 0.6901   
## extras -0.6194600 0.2616917 -2.367 0.0212 \*   
## payment\_typeCredit Card 0.4913386 2.6691938 0.184 0.8546   
## trip\_miles -0.5362138 0.4056223 -1.322 0.1913   
## I(trip\_miles^2) 0.2395887 0.1018113 2.353 0.0220 \*   
## I(tips^2) 0.5002003 0.6310208 0.793 0.4311   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.328 on 59 degrees of freedom  
## Multiple R-squared: 0.8416, Adjusted R-squared: 0.8228   
## F-statistic: 44.77 on 7 and 59 DF, p-value: < 2.2e-16

AIC(taxitrips\_sq.out)

## [1] 237.6487

vif(taxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 2.190428 317.784607 1.072504 59.677130   
## trip\_miles I(trip\_miles^2) I(tips^2)   
## 6.677432 7.041531 126.144360

# squared tips term is removing the signifiance of tips also. vif for some terms has increased way to high which tells us that there is high multicorrelation between those terms.Will remove it and introduce squared term for extras  
  
# squaring extras terms along with miles  
  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(extras^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(extras^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2668 -0.7147 -0.2310 0.5161 4.1543   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.1853643 0.4196841 9.973 2.85e-14 \*\*\*  
## trip\_seconds 0.0071049 0.0007697 9.231 4.72e-13 \*\*\*  
## tips 1.0006886 0.4834349 2.070 0.0428 \*   
## extras -0.7464651 0.7078676 -1.055 0.2959   
## payment\_typeCredit Card -1.4181416 1.1001038 -1.289 0.2024   
## trip\_miles -0.5831275 0.4030224 -1.447 0.1532   
## I(trip\_miles^2) 0.2461884 0.1025760 2.400 0.0196 \*   
## I(extras^2) 0.0922821 0.3926957 0.235 0.8150   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.335 on 59 degrees of freedom  
## Multiple R-squared: 0.84, Adjusted R-squared: 0.8211   
## F-statistic: 44.26 on 7 and 59 DF, p-value: < 2.2e-16

AIC(taxitrips\_sq.out)

## [1] 238.2958

vif(taxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 2.269078 10.409174 7.771924 10.039697   
## trip\_miles I(trip\_miles^2) I(extras^2)   
## 6.528745 7.079001 7.844312

# Not much change in Adjusted R-square but the vif for the terms seems to be with 10 which is good. Before deciding this model as best model we will check for few more scenarios  
  
# Introducing squared terms for trip\_miles,extras and tips  
  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(extras^2)+I(tips^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(extras^2) + I(tips^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2571 -0.6850 -0.1049 0.4866 4.1671   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.1808148 0.4210253 9.930 4.02e-14 \*\*\*  
## trip\_seconds 0.0070877 0.0007724 9.176 6.79e-13 \*\*\*  
## tips -1.1030648 2.6831801 -0.411 0.683   
## extras -0.7970676 0.7128960 -1.118 0.268   
## payment\_typeCredit Card 0.5442140 2.6976538 0.202 0.841   
## trip\_miles -0.5340803 0.4089281 -1.306 0.197   
## I(trip\_miles^2) 0.2354586 0.1037710 2.269 0.027 \*   
## I(extras^2) 0.1057318 0.3942756 0.268 0.790   
## I(tips^2) 0.5075057 0.6366264 0.797 0.429   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.339 on 58 degrees of freedom  
## Multiple R-squared: 0.8418, Adjusted R-squared: 0.8199   
## F-statistic: 38.57 on 8 and 58 DF, p-value: < 2.2e-16

AIC(taxitrips\_sq.out)

## [1] 239.5656

vif(taxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 2.270860 318.675411 7.834034 59.997653   
## trip\_miles I(trip\_miles^2) I(extras^2) I(tips^2)   
## 6.679960 7.200134 7.858702 126.375765

# Adjusted R-sqaure dropped a bit and vif for some terms has increased a lot, so this model is not better that previous only.  
  
# Squared term for trip\_miles and trip\_seconds  
  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(trip\_seconds^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(trip\_seconds^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3900 -0.6979 -0.1563 0.5518 3.7927   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.428e+00 5.207e-01 6.585 1.36e-08 \*\*\*  
## trip\_seconds 1.104e-02 1.876e-03 5.886 2.00e-07 \*\*\*  
## tips 9.775e-01 4.628e-01 2.112 0.03890 \*   
## extras -5.468e-01 2.506e-01 -2.182 0.03309 \*   
## payment\_typeCredit Card -1.433e+00 1.051e+00 -1.364 0.17777   
## trip\_miles -1.037e+00 4.336e-01 -2.391 0.02002 \*   
## I(trip\_miles^2) 4.478e-01 1.300e-01 3.443 0.00106 \*\*   
## I(trip\_seconds^2) -3.353e-06 1.460e-06 -2.296 0.02526 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.279 on 59 degrees of freedom  
## Multiple R-squared: 0.853, Adjusted R-squared: 0.8356   
## F-statistic: 48.91 on 7 and 59 DF, p-value: < 2.2e-16

AIC(taxitrips\_sq.out)

## [1] 232.6254

vif(taxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 14.676622 10.380307 1.059920 9.970104   
## trip\_miles I(trip\_miles^2) I(trip\_seconds^2)   
## 8.226126 12.382622 20.278011

# Adjusted R-square has increased to 83.5 which is the best which we got till now.trip\_miles and square of trip\_miles are also having signifiance along with trip\_seconds, tips,extras,square of trip\_seconds. Also, vif looks good, although some of values are slightly greater than 10 but they not too high, everything below 20, which means those variables will have slighly multicollinearity amoung them. But overall this model looks the best

# Question : 6 Best Model

# Out of all the models ran above the the best model output was generated when we introduced Squared term for trip\_miles and trip\_seconds

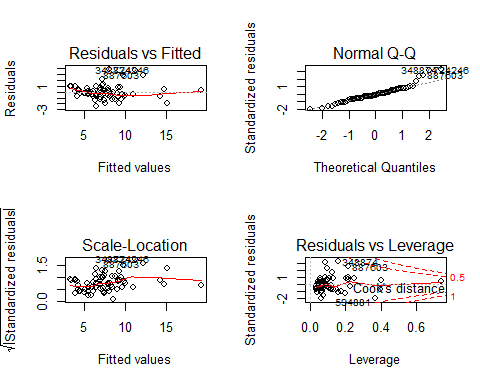
Consolidated results in the below table from all the models which are applied above. And based on those we can see that the last model which has squared terms for trip\_miles and trip\_seconds is giving us the best fit. It has greater adjusted R-square, less AIC as compared to what we get with other squared term model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Terms** | **Squared Terms** | **Adjusted R square** | **AIC/ minimum AIC** | **DF** | **Significant terms** |
| Kitchen Sink | All except taxi id | None | 100 | -4281.1 | 60 | trip\_seconds, tips,extras,trip\_total |
| Kitchen Sink | Removing taxi id and trip\_total | None | 80.9 | 240.893 | 61 | trip\_seconds,tips |
| Kitchen Sink | excluding taxi\_id,trip\_total and tolls | None | 80.8 | 240.893 | 61 | trip\_seconds,tips |
| Stepwise regression - both direction | excluding taxi\_id,trip\_total and tolls | None | 80.8 | 48.21 (model removed miles) | 62 | trip\_seconds,tips,extras |
| Stepwise regression - backward | excluding taxi\_id,trip\_total and tolls | None | 80.8 | 48.21 (model removed miles) | 62 | trip\_seconds,tips,extras |
| Squaring terms | excluding taxi\_id,trip\_total and tolls | trip\_miles | 82.3 | 236.358 | 60 | trip\_seconds,tips,extras,square(trip\_miles) |
| Cubic terms | excluding taxi\_id,trip\_total and tolls | trip\_miles, cubic for trip\_miles | 82.1 | 238.333 | 59 | trip\_seconds,tips,extras |
| Sqauring terms | excluding taxi\_id,trip\_total and tolls | trip\_miles,tips | 82.28 | 237.649 | 59 | trip\_Seconds,extras,square(trip\_miles) |
| Squaring terms | excluding taxi\_id,trip\_total and tolls | trip\_miles,extras | 82.11 | 238.296 | 59 | trip\_Seconds,tips,square(trip\_miles) |
| Squaring terms | excluding taxi\_id,trip\_total and tolls | trip\_miles,extras,tips | 81.99 | 239.566 | 59 | trip\_Seconds,square(trip\_miles) |
| Squaring terms | excluding taxi\_id,trip\_total and tolls | trip\_miles,trip\_seconds | 83.56 | 232.625 | 59 | trip\_seconds,miles,tips,extras,square(trip\_miles),square(trip\_Seconds) |

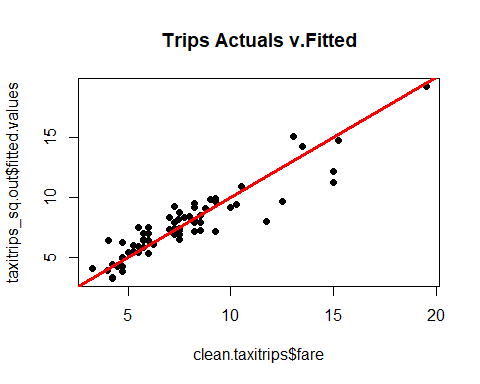
# Generating Regression output again for the best model  
taxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(trip\_seconds^2), data = clean.taxitrips)  
summary(taxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(trip\_seconds^2), data = clean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3900 -0.6979 -0.1563 0.5518 3.7927   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.428e+00 5.207e-01 6.585 1.36e-08 \*\*\*  
## trip\_seconds 1.104e-02 1.876e-03 5.886 2.00e-07 \*\*\*  
## tips 9.775e-01 4.628e-01 2.112 0.03890 \*   
## extras -5.468e-01 2.506e-01 -2.182 0.03309 \*   
## payment\_typeCredit Card -1.433e+00 1.051e+00 -1.364 0.17777   
## trip\_miles -1.037e+00 4.336e-01 -2.391 0.02002 \*   
## I(trip\_miles^2) 4.478e-01 1.300e-01 3.443 0.00106 \*\*   
## I(trip\_seconds^2) -3.353e-06 1.460e-06 -2.296 0.02526 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.279 on 59 degrees of freedom  
## Multiple R-squared: 0.853, Adjusted R-squared: 0.8356   
## F-statistic: 48.91 on 7 and 59 DF, p-value: < 2.2e-16

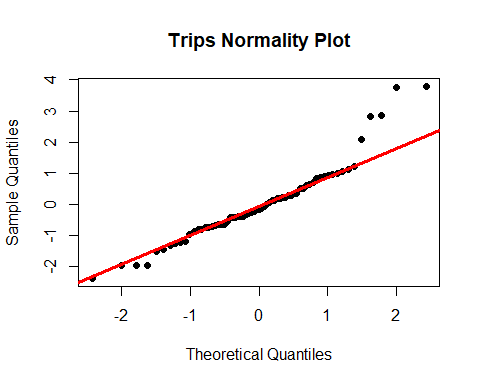
# LINE Conformity  
  
par(mfrow=c(2,2))  
plot(taxitrips\_sq.out)



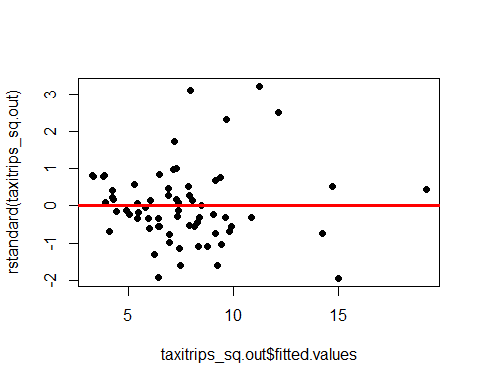
par(mfrow=c(1,1))  
  
# Linearity:  
plot(clean.taxitrips$fare,taxitrips\_sq.out$fitted.values,pch=19,main="Trips Actuals v.Fitted")  
abline(0,1,lwd=3,col="red")



# Based on the above scatter plot we see that there is some linear relationship when the fare are less than 10 dollars but after that the values are scattered and it has some outliers. So we can say it sort of conform's Linearity but not very strongly.  
  
  
# Normality  
qqnorm(taxitrips\_sq.out$residuals,pch=19,main="Trips Normality Plot")  
qqline(taxitrips\_sq.out$residuals,lwd=3,col="red")



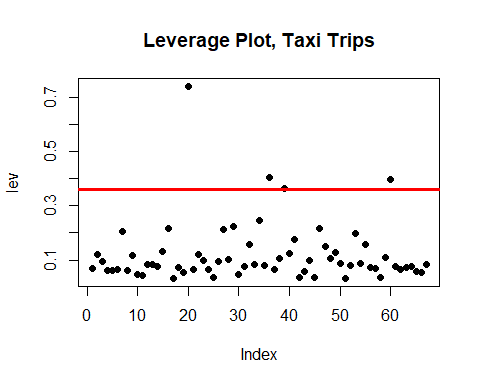
# Ideally for a plot to be considered as normal it should have all the points falling on the line. But from the qq plot generated above we can see that its mostly normally distributed in the center but it has few outliers above and below the qq line which deviates it away. So, we can say it sort of conform's normality.  
  
  
# Equality of Variances  
plot(taxitrips\_sq.out$fitted.values,rstandard(taxitrips\_sq.out),pch=19)  
abline(0,0,col="red",lwd=3)



# From the above plot, we can see that majority of the observations are concentrated around 0 because of which it looks like equally distributed but we have many outliers and extreme values also, so the model is not conforming Equality of variances.  
  
  
# Independence  
  
# We look for independence when it is time-series data and here we don’t have any time-series

# Question 7: #Identifying high leverage points.

lev=hat(model.matrix(taxitrips\_sq.out))  
plot(lev,pch=19,main="Leverage Plot, Taxi Trips")  
abline(3\*mean(lev),0,col="red",lwd=3)



clean.taxitrips[lev>(3\*mean(lev)),]

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 879479 8206 1560 5.2 19.50 0.00 0 0 19.50  
## 61239 4640 1500 3.7 15.25 0.00 0 0 15.25  
## 594881 3086 1200 3.9 13.05 3.01 0 2 18.06  
## 312033 8387 1380 2.9 13.50 2.70 0 0 16.20  
## payment\_type  
## 879479 Cash  
## 61239 Cash  
## 594881 Credit Card  
## 312033 Credit Card

lev\_points = clean.taxitrips[lev>(3\*mean(lev)),1]  
  
  
# Removing high leverage points  
lev.taxitrips = clean.taxitrips[-which(clean.taxitrips$taxi\_id %in% lev\_points),]  
  
# Rerunning final model and Checking output after removing high leverage points  
trip\_lev.out=lm(formula = fare ~ trip\_seconds + tips +   
extras + payment\_type + trip\_miles+I(trip\_miles^2)+I(trip\_seconds^2), data = lev.taxitrips)  
summary(trip\_lev.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(trip\_seconds^2), data = lev.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2308 -0.6787 -0.1001 0.5080 4.0235   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.823e+00 6.324e-01 6.045 1.36e-07 \*\*\*  
## trip\_seconds 8.293e-03 2.759e-03 3.006 0.00399 \*\*   
## tips 1.188e+00 4.739e-01 2.506 0.01520 \*   
## extras -3.777e-01 2.624e-01 -1.439 0.15570   
## payment\_typeCredit Card -1.802e+00 1.043e+00 -1.727 0.08981 .   
## trip\_miles -1.534e+00 6.324e-01 -2.425 0.01861 \*   
## I(trip\_miles^2) 8.018e-01 2.610e-01 3.072 0.00330 \*\*   
## I(trip\_seconds^2) -3.357e-07 2.661e-06 -0.126 0.90010   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.234 on 55 degrees of freedom  
## Multiple R-squared: 0.7777, Adjusted R-squared: 0.7494   
## F-statistic: 27.49 on 7 and 55 DF, p-value: 8.344e-16

# After removing the leverage points the model adjusted R-sqaure has decreased the signifiance of extra and sqaured trip\_seconds has removed.

# Question 8: Creating new subset of 100 random records with new seed which is 09448095

set.seed(09448095)  
newsubset.taxitrips=taxitrips[sample(1:nrow(taxitrips),100,replace=FALSE),]  
  
# Data Cleaning  
# Checking for missing values if any.  
summary(newsubset.taxitrips)

## taxi\_id trip\_seconds trip\_miles fare   
## Min. : 413 Min. : 0 Min. : 0.000 Min. : 3.750   
## 1st Qu.:2466 1st Qu.: 300 1st Qu.: 0.000 1st Qu.: 6.250   
## Median :4127 Median : 540 Median : 1.035 Median : 7.625   
## Mean :4416 Mean : 651 Mean : 2.565 Mean :12.455   
## 3rd Qu.:6564 3rd Qu.: 780 3rd Qu.: 2.525 3rd Qu.:13.375   
## Max. :8696 Max. :3060 Max. :21.600 Max. :52.750   
## tips tolls extras trip\_total   
## Min. : 0.000 Min. :0 Min. : 0.0000 Min. : 3.750   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.: 0.0000 1st Qu.: 7.237   
## Median : 1.000 Median :0 Median : 0.0000 Median : 9.500   
## Mean : 1.764 Mean :0 Mean : 0.9425 Mean : 15.162   
## 3rd Qu.: 2.000 3rd Qu.:0 3rd Qu.: 1.0000 3rd Qu.: 15.312   
## Max. :16.950 Max. :0 Max. :32.0000 Max. :101.700   
## payment\_type   
## Length:100   
## Class :character   
## Mode :character   
##   
##   
##

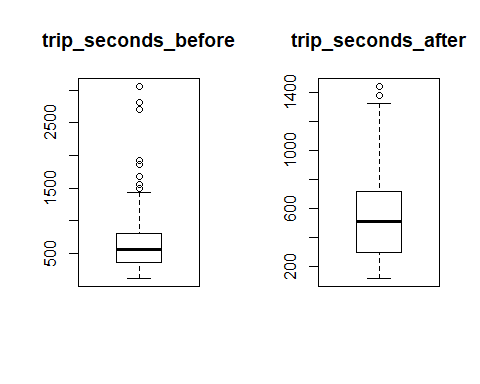
# No NA/ NULLs in the means so we dont have any missing data  
# Checking if there are any 0 values in trip\_seconds, because in ideal scenario trip\_seconds should not be zero.If a trip was taken or even started there should be some elapsed time associated with it.  
nzero\_tripseconds = newsubset.taxitrips[which(newsubset.taxitrips$trip\_seconds == 0.0) ,]  
nzero\_tripseconds

## taxi\_id trip\_seconds trip\_miles fare tips tolls extras trip\_total  
## 913164 1796 0 0 5.50 1.65 0 0 7.15  
## 1282974 5897 0 0 6.75 1.35 0 0 8.10  
## 184723 4464 0 0 6.00 1.20 0 0 7.20  
## 964509 7147 0 0 4.75 0.00 0 4 8.75  
## 1140579 8552 0 0 9.75 1.95 0 0 11.70  
## 957972 2746 0 0 18.75 3.75 0 0 22.50  
## 688758 773 0 0 5.85 2.15 0 0 8.00  
## 918123 7056 0 0 27.25 5.45 0 0 32.70  
## payment\_type  
## 913164 Credit Card  
## 1282974 Credit Card  
## 184723 Credit Card  
## 964509 Cash  
## 1140579 Credit Card  
## 957972 Credit Card  
## 688758 Credit Card  
## 918123 Credit Card

# 8 such records are found, which seems to be bad data, so I will remove them from my subset.  
nnonzero.taxitrips = newsubset.taxitrips[-which(newsubset.taxitrips$trip\_seconds == 0.0) ,]  
  
# Checking outliers in the data.  
# Outliers for trip\_seconds  
  
par(mfrow=c(1,2))  
outliers\_nts =boxplot(nnonzero.taxitrips$trip\_seconds,main="trip\_seconds\_before")$out  
outliers\_nts

## [1] 2820 1680 1560 1500 1920 1500 1860 2700 1560 3060

# Removing outliers for trip\_seconds from the dataset  
newclean.taxitrips = nnonzero.taxitrips[-which(nnonzero.taxitrips$trip\_seconds %in% outliers\_nts),]  
boxplot(newclean.taxitrips$trip\_seconds,main="trip\_seconds\_after")$out

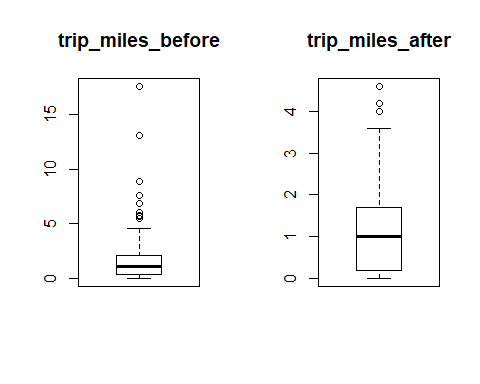


## [1] 1380 1440

par(mfrow=c(1,1))  
  
# Checking outliers for trip\_miles  
par(mfrow=c(1,2))  
outliers\_ntm = boxplot(newclean.taxitrips$trip\_miles,main="trip\_miles\_before")$out  
outliers\_ntm

## [1] 5.8 5.5 17.6 6.0 13.1 5.7 7.6 6.9 8.9

# Removing outliers for trip\_miles from the dataset  
newclean.taxitrips = newclean.taxitrips[-which(newclean.taxitrips$trip\_miles %in% outliers\_ntm),]  
boxplot(newclean.taxitrips$trip\_miles,main="trip\_miles\_after")$out

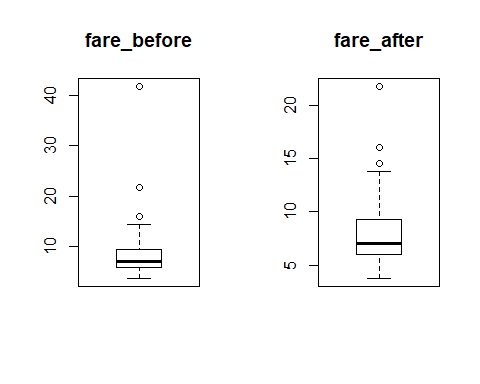


## [1] 4.0 4.2 4.6

par(mfrow=c(1, 1))  
  
# Checking outliers for fare  
par(mfrow=c(1,2))  
boxplot(newclean.taxitrips$fare,main="fare\_before")$out

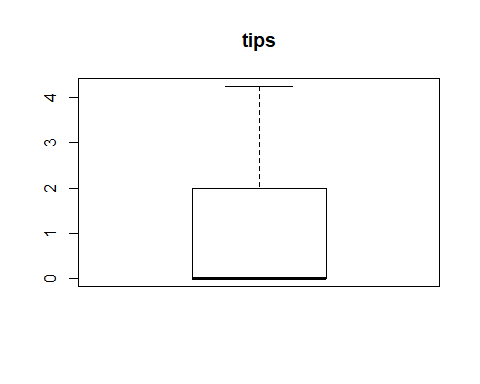
## [1] 16.00 21.75 41.75

# Removing 1 extreme outlier for fare from the dataset  
newclean.taxitrips = newclean.taxitrips[-which(newclean.taxitrips$fare == 41.75),]  
boxplot(newclean.taxitrips$fare,main="fare\_after")$out



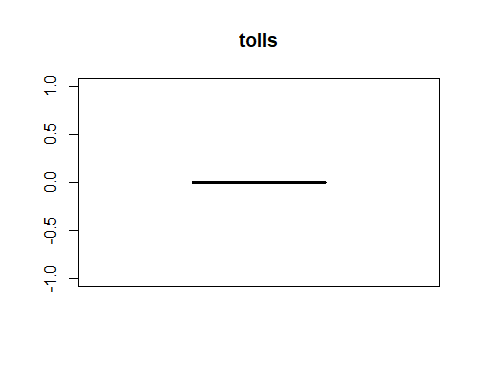
## [1] 16.00 21.75 14.50

par(mfrow=c(1, 1))  
  
# After cleaning Outliers from trip\_seconds, trip\_miles and fare, checking outliers for remaining variables.  
boxplot(newclean.taxitrips$tips,main="tips")$out



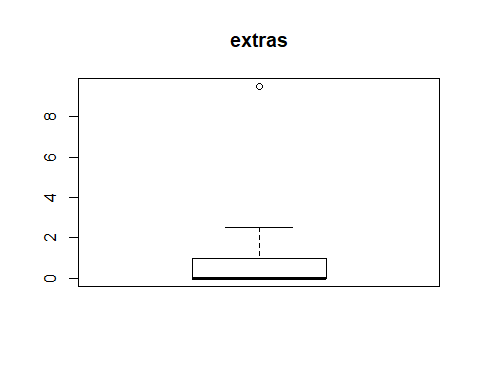
## numeric(0)

# No outliers for tips  
boxplot(newclean.taxitrips$tolls,main="tolls")$out



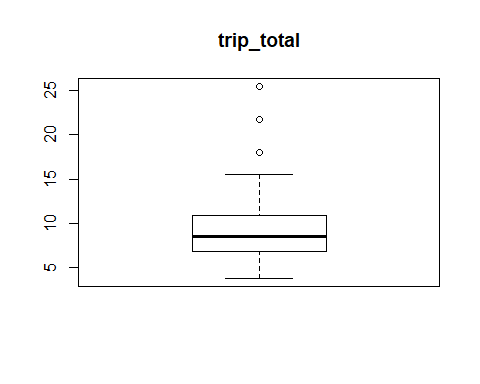
## numeric(0)

# No outliers for tolls  
boxplot(newclean.taxitrips$extras,main="extras")$out



## [1] 9.5

# 1 outliers for extras  
boxplot(newclean.taxitrips$trip\_total,main="trip\_total")$out



## [1] 18.00 25.50 21.75

# 3 outliers for trip\_total but there are not very extreme, so I will keep them as is.  
  
# So finally we are left with 72 observations which is the primary dataset for analysis

# In order to check if the best model generated in Question 6 is best fit for this new set of data, we will take base model first (generate by Question 4)

# Base Model : using fare as dependent and trip\_seconds, trip\_miles and payment\_type as independent variables(Question 4)   
   
newtaxitrips.out =lm(formula = fare ~ trip\_seconds + payment\_type + trip\_miles, data = newclean.taxitrips)  
summary(newtaxitrips.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + payment\_type + trip\_miles,   
## data = newclean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6657 -0.8894 -0.1861 0.4451 9.6789   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.4775785 0.5207715 4.758 1.06e-05 \*\*\*  
## trip\_seconds 0.0099933 0.0009325 10.717 2.99e-16 \*\*\*  
## payment\_typeCredit Card -0.1090273 0.4133235 -0.264 0.79275   
## trip\_miles 0.5266866 0.1876393 2.807 0.00652 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.709 on 68 degrees of freedom  
## Multiple R-squared: 0.6994, Adjusted R-squared: 0.6861   
## F-statistic: 52.74 on 3 and 68 DF, p-value: < 2.2e-16

AIC(newtaxitrips.out)

## [1] 287.3788

vif(newtaxitrips.out)

## trip\_seconds payment\_type trip\_miles   
## 1.111159 1.013102 1.120907

# We got Adjusted R-square of 68.6% and trip\_seconds and trip\_miles are having an significant impact on fare.

# Applying the best model generated out of all the models ran above the model output came when we introduced Squared term for trip\_miles and trip\_seconds and comparing to based model  
  
  
newtaxitrips\_sq.out=lm(formula = fare ~ trip\_seconds + tips + extras   
+ payment\_type + trip\_miles+I(trip\_miles^2)+I(trip\_seconds^2), data = newclean.taxitrips)  
summary(newtaxitrips\_sq.out)

##   
## Call:  
## lm(formula = fare ~ trip\_seconds + tips + extras + payment\_type +   
## trip\_miles + I(trip\_miles^2) + I(trip\_seconds^2), data = newclean.taxitrips)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9606 -0.6422 -0.1529 0.4375 8.6736   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.108e+00 8.845e-01 5.776 2.44e-07 \*\*\*  
## trip\_seconds 1.795e-03 3.464e-03 0.518 0.605983   
## tips 3.076e-01 4.537e-01 0.678 0.500230   
## extras -1.042e-01 1.892e-01 -0.551 0.583470   
## payment\_typeCredit Card -5.808e-01 1.020e+00 -0.570 0.570973   
## trip\_miles -1.021e+00 4.725e-01 -2.161 0.034483 \*   
## I(trip\_miles^2) 4.489e-01 1.248e-01 3.598 0.000625 \*\*\*  
## I(trip\_seconds^2) 6.776e-06 3.157e-06 2.146 0.035658 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.555 on 64 degrees of freedom  
## Multiple R-squared: 0.7659, Adjusted R-squared: 0.7402   
## F-statistic: 29.9 on 7 and 64 DF, p-value: < 2.2e-16

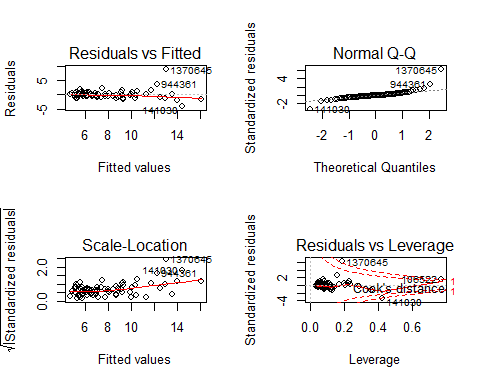
AIC(newtaxitrips\_sq.out)

## [1] 277.3908

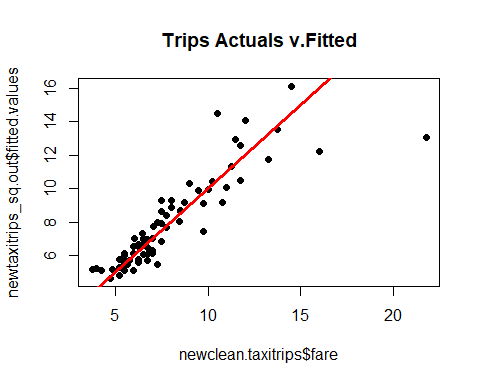
vif(newtaxitrips\_sq.out)

## trip\_seconds tips extras payment\_type   
## 18.521558 8.491391 1.610683 7.452003   
## trip\_miles I(trip\_miles^2) I(trip\_seconds^2)   
## 8.586636 9.190013 18.267170

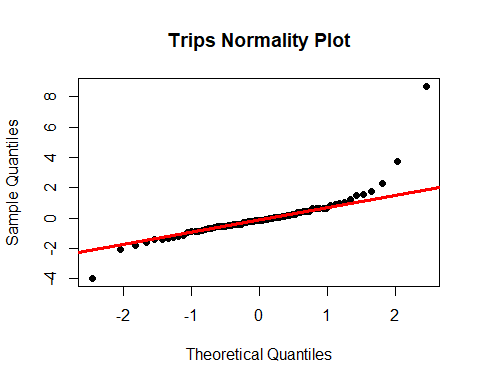
# Adjusted R-square increased to 74% and significant terms are trip\_miles, square of trip\_miles and trip\_seconds. AIC has decreased as compared to base model. vif is low so there is less multicollinearity between the terms.  
  
# LINE Conformity  
  
par(mfrow=c(2,2))  
plot(newtaxitrips\_sq.out)



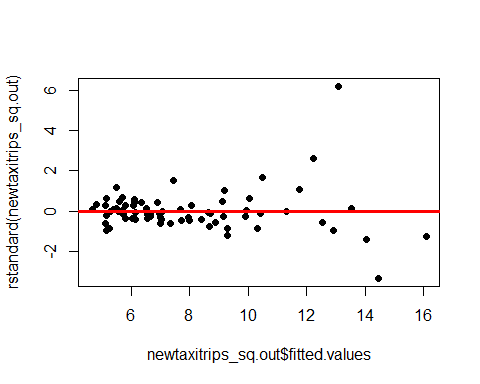
par(mfrow=c(1,1))  
  
#Linearity:  
plot(newclean.taxitrips$fare,newtaxitrips\_sq.out$fitted.values,pch=19,main="Trips Actuals v.Fitted")  
abline(0,1,lwd=3,col="red")



#Based on the above scatter plot we see that there is linear relationship when the fare is less than 15 dollars and there are few outliers also specially at the upper end. In this case we can strongly conform the Linearity.  
  
#Normality  
qqnorm(newtaxitrips\_sq.out$residuals,pch=19,main="Trips Normality Plot")  
qqline(newtaxitrips\_sq.out$residuals,lwd=3,col="red")



# From the qq plot generated above we can see that its mostly normally distributed in the center and has couple of outliers above and below the qq line which deviates it away. It strongly conforms the normality.  
  
  
#Equality of Variances  
plot(newtaxitrips\_sq.out$fitted.values,rstandard(newtaxitrips\_sq.out),pch=19)  
abline(0,0,col="red",lwd=3)



#From the above plot, we can see that majority of the observations are concentrated around 0 because of which it looks like equally distributed but we have few outliers and extreme values also, so for this case we can say that model is sort of conforming equalirt of variances.  
  
#Independence  
  
#We look for independence when it is time-series data and here we don’t have any time-series

# Based on above analysis we can say that the best model generated in step 6 fits fine with the new dataset as well.