Stats R Lab Core Challenge: Linear Model & Prediction

Linfeng Zhou

Sunday, August 02, 2015

1 Summary statistics

Firstly, we load the data and using summary() function to obtain summary statistics.

The summary statistics of all data are:

```
summary(bikedata_origi$tripduration)
```

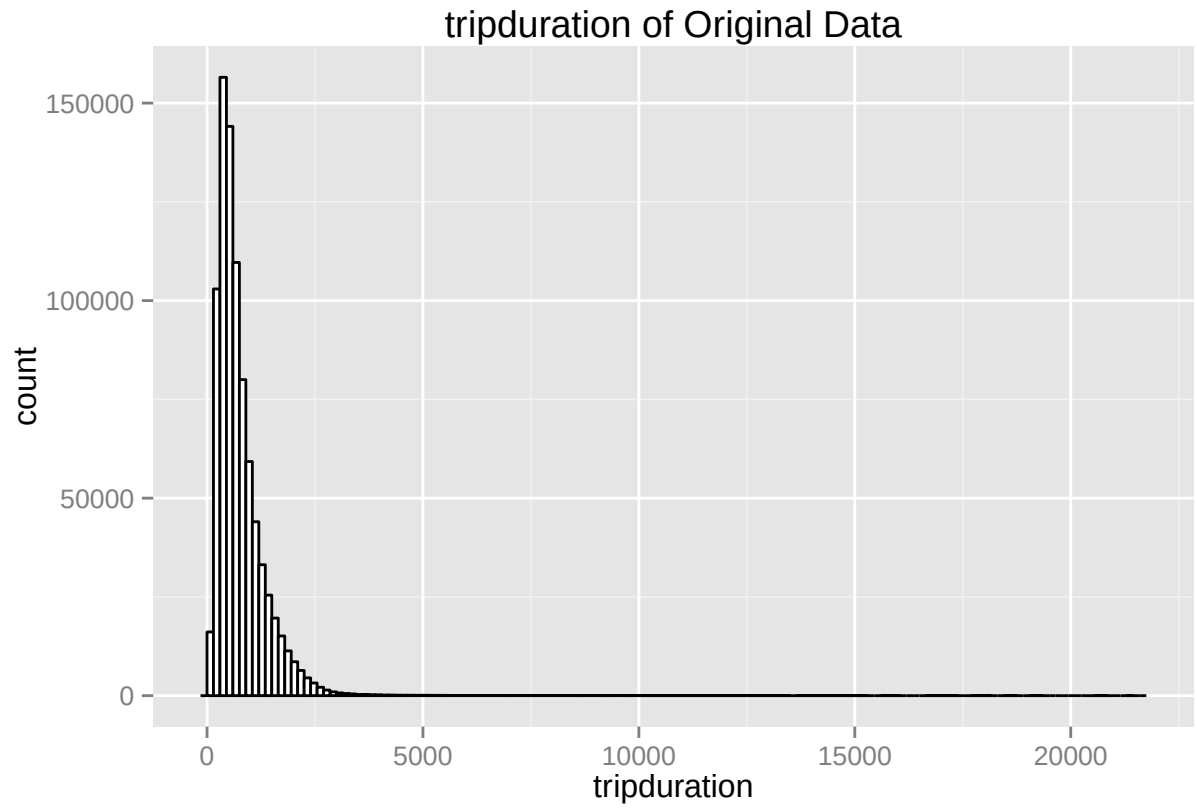
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   390.0   606.0   767.6   965.0 21560.0
```

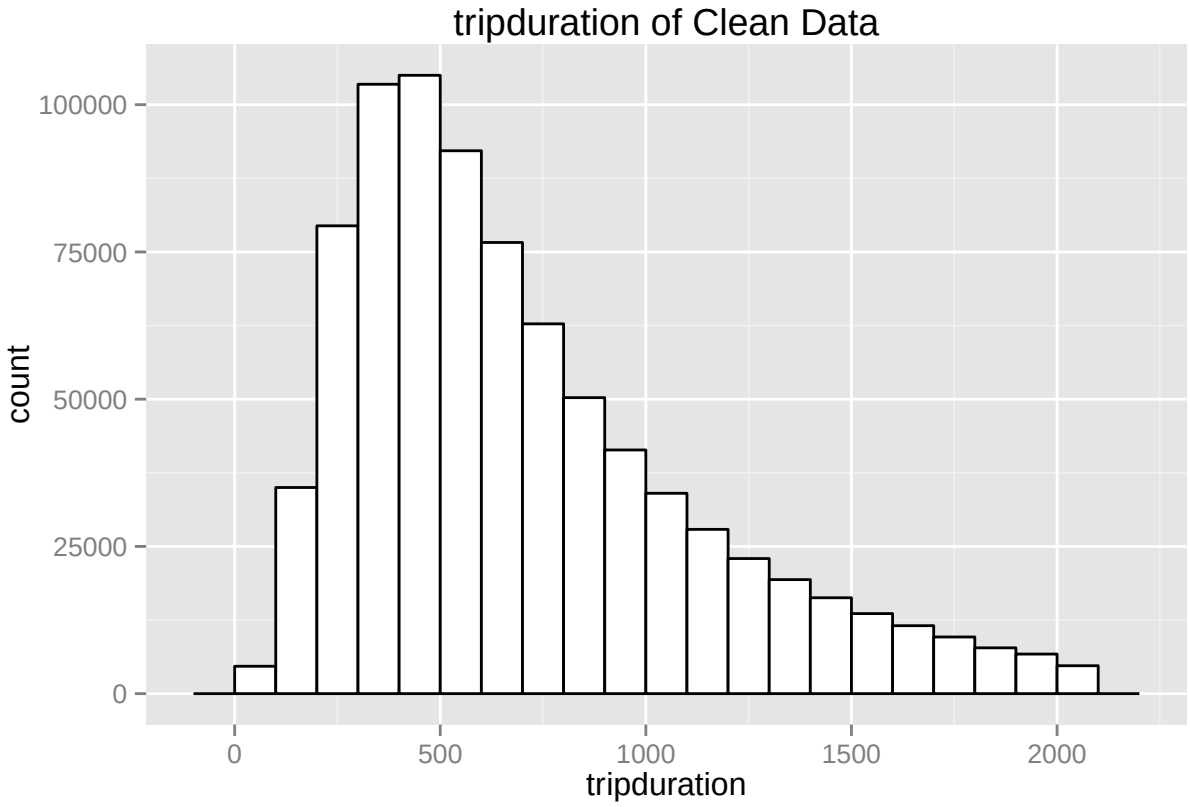
The summary statistics of clean data are:

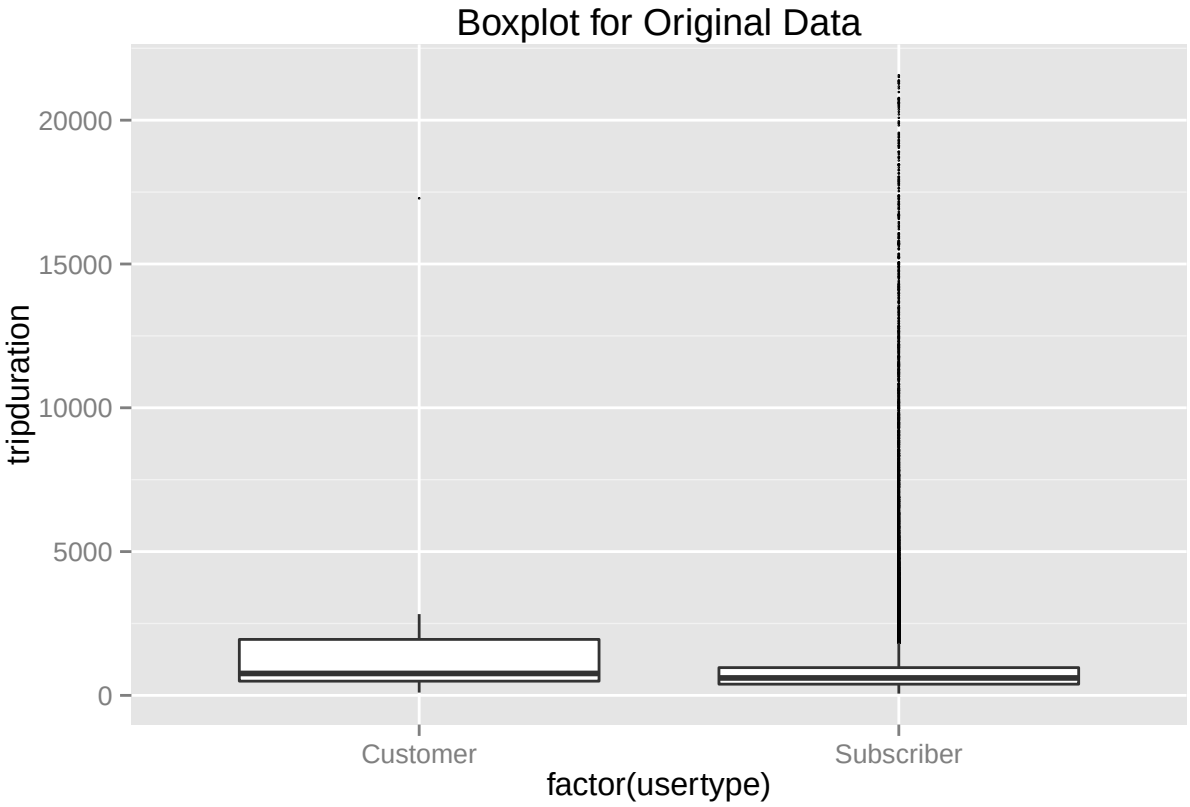
```
summary(bikedata_clean$tripduration)
```

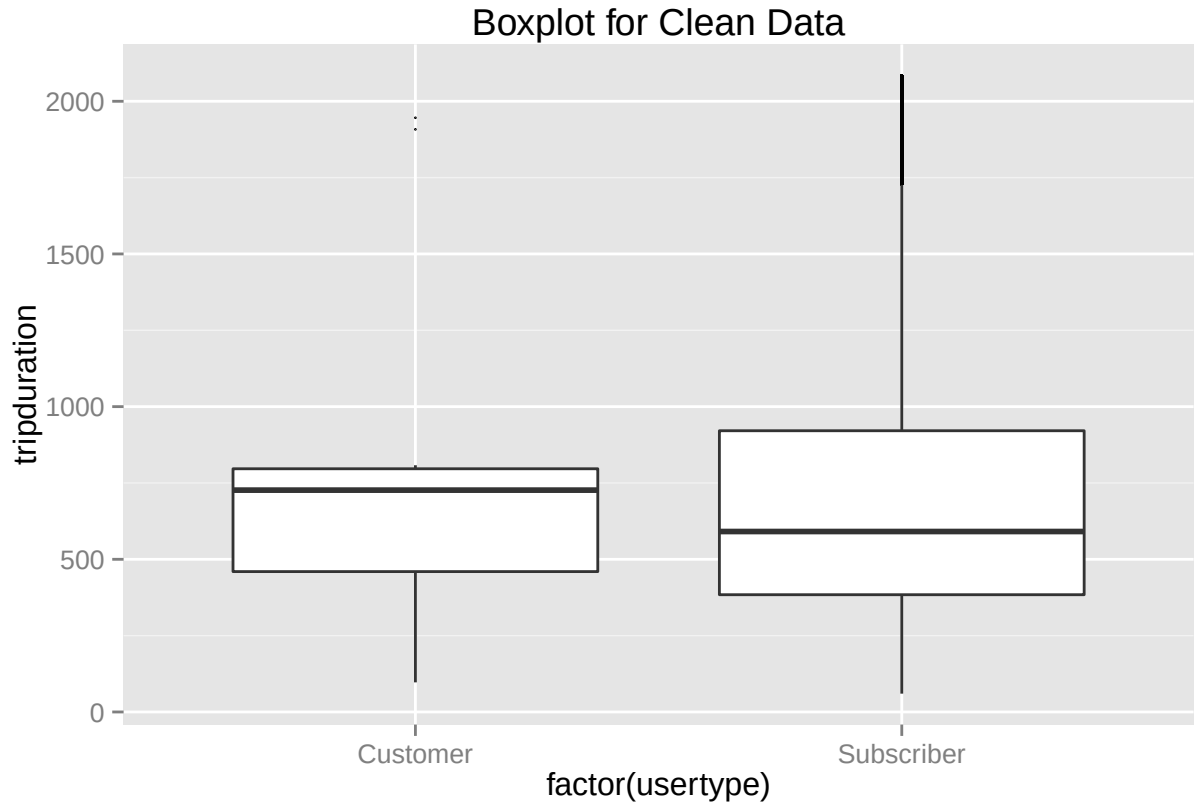
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   384.0   591.0   700.9   921.0  2085.0
```

2 Histogram and boxplot




tripduration of Clean Data
count
100000
75000
50000
25000
0
0
500
1000
1500
2000
tripduration


Boxplot for Original Data
tripduration
20000
15000
10000
5000
0
Customer
Subscriber
factor(usertype)



3 Regression analysis

```
##
## Call:
## lm(formula = tripduration ~ age, data = origidata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -657.00  -25.15   -4.10   23.99  980.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  645.517     55.234  11.687  < 2e-16 ***
## age            3.113      0.885   3.518 0.000739 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 198 on 76 degrees of freedom
## Multiple R-squared:   0.14,  Adjusted R-squared:  0.1287
## F-statistic: 12.37 on 1 and 76 DF,  p-value: 0.0007387

##
## Call:
## lm(formula = tripduration ~ age, data = cleandata)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -568.16  -24.35   -2.18   24.98 1069.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 584.1925    52.5131  11.125  < 2e-16 ***
## age           2.7854     0.8414   3.311  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 188.3 on 76 degrees of freedom
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.1145
## F-statistic: 10.96 on 1 and 76 DF,  p-value: 0.001426
```

4 Compare the two models and comment on the following

- Which model is better?

  From the regression analysis' summary, we can see that the R-square of original dataset is larger than the dataset without outliers, therefore, we conclude that first model is better.

- What is the effect of age in trip duration?

  As we can see the coefficients of regression models, the coefficient of age is positive, which means age has a positive effect in trip duration.

- What is the effect of outliers in these models?

  The outliers increase coefficients of models not only on age parameters but also on intercept, which may cause overfitting .

- From this analysis, what would you recommend Citi Bike in terms of a more equitable and socially responsible program for NYC and other cities around the world?

  According to our model, most users are in short-distance driving. Therefore we can set more citi bike in order to fit people's need.