Descriptive & Summary Statistics of NYC CitiBike

Linfeng Zhou lz1335@nyu.edu

Question 1: Summary statistics table for January 2015 data

Firstly, we load the data from local dataset and checking the size and dimensions of this file.

```
setwd("E:/Code/NYU-UCSL-2015/R")
citibike_data<-read.csv("201501-citibike-tripdata.csv",header=TRUE)
dim(citibike_data)
```

```
## [1] 285552     15
```

After that, we exploring headers in the .csv file

```
names(citibike_data)
```

```
##  [1] "tripduration"            "starttime"
##  [3] "stoptime"                "start.station.id"
##  [5] "start.station.name"      "start.station.latitude"
##  [7] "start.station.longitude" "end.station.id"
##  [9] "end.station.name"        "end.station.latitude"
## [11] "end.station.longitude"   "bikeid"
## [13] "usertype"                "birth.year"
## [15] "gender"
```

Now, it's time to analyze our mission. Qustion 1 ask us to provide summary statistics tables for

- (1) all January 2015 data

- (2) January 2015 data without outliers

It this situation, we should use z-scores and subset funtion to generate two dataset: first one is the original one, and the second one is clean data (which don't have outliers). Using z-scores(Z>3) ,we can detect outliers from our original data and delete them.

```
#COMPUTING Z-SCORE
tripduration.sd <- sd(citibike_data$tripduration)
tripduration.mean <- mean(citibike_data$tripduration)
z <- (citibike_data$tripduration-tripduration.mean)/tripduration.sd
```

We have tow dataset:one is named bikedata_clean, another is named bikedata_clean.

Using pre-bulit function we calculate the summary statistics for January 2015 data.
(The code is omitted, but you can see it at Rmd file)

After that, we have organize all statistics in one table, which is:

```
bikedata_summary
```

```
##                     bikedata_origi.summary bikedata_clean.summary
## Count                           285552.0000            284255.0000
## Mean                               654.3256               616.4736
## Standard Deviation                 900.7759               421.1332
## Min                                 60.0000                60.0000
## Max                              43023.0000              3355.0000
## Median                             504.0000               502.0000
## 25% Quantiles                      334.0000               333.0000
## 75 Quantiles                       772.0000               766.0000
```

Comment on the folloing question:

- What can you say about the central tendency of this month?

  The average value and median in original data are (654.3255834,504), and in data without outliers are (616.4735713,502)

- What can you say about the spread and dispersion of the data?

  The range and standard deviation in original data are(42963,900.7759282), and the same statistics in data without outliers are(3295,421.1331857)

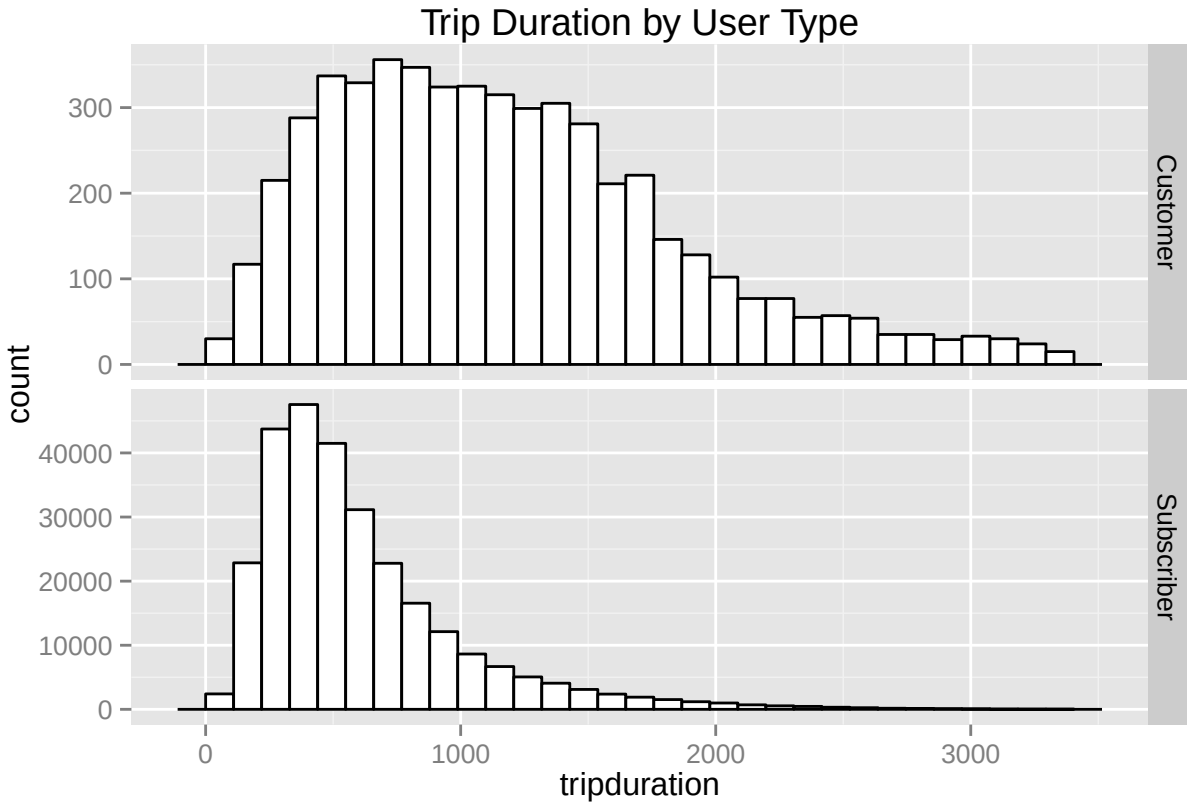
- What is the effect of outliers as per our summary statistics?

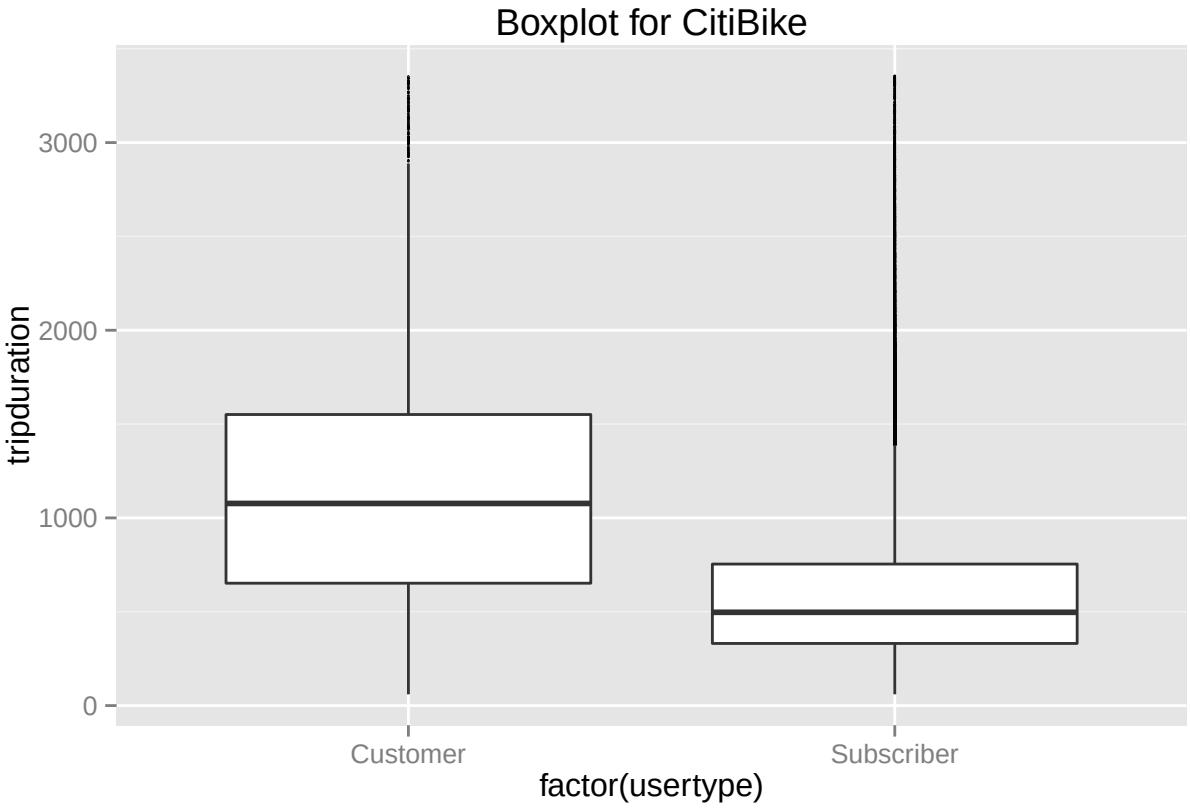
  As we can see the summary statistics table, effects of outliers are: 1. Raised the average of dataset, the average value of original data is bigger than clean data. 2. Raised the standard deviation of the dateset. Other statistics like Max Value, Quantiles Value get a corresponding rise.

Question 2: Visualize the trip duration by user type

Using ggplot2, plots of trip duration data by user type are :

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```


Trip Duration by User Type
count
300
200
100
0
Customer
40000
30000
20000
10000
0
Subscriber
0
1000
2000
3000
tripduration


Boxplot for CitiBike
tripduration
3000
2000
1000
0
Customer
Subscriber
factor(usertype)

Comment on the following:

- How does outliers affect histograms and/or box plots?

  It will affect on widths of histograms and amount of outliers on boxplots.

- What do you expect the histogram of this month would look like if you were to plot it with outliers

  Histograms may have a long tail, and box plots may show outliers if there are too many outliers.