UCSL: R Challenge 1

Christopher Prince [cmp670@nyu.edu]

07/09/2015

1 Summary statistics

1.1 Complete January 2015 data set

For the first part of the challenge, we are asked to provide the basic descriptive statistics for the Citibike data set from January 2015. We begin by importing the .csv data into a data frame named `cbdata` then calculate the statistics in the table below[1]:

symbol	description	value
`cbdata.count`	number of observations in `cbdata`	285552
`tdur.mean`	mean trip duration	654.33
`tdur.var`	variance of trip duration	811397
`tdur.sd`	standard deviation of trip duration	900.78
`tdur.median`	median trip duration	504
`tdur.min`	minimum trip duration	60
`tdur.max`	maximum trip duration	43023
`tdur.range`	total range of trip duration values	42963
`tdur.iqr`	interquartile range of trip duration values	438

The quartiles for this set are:

0%	60
25%	334
50%	504
75%	772
100%	43023

1.2 Removing outliers

Outliers are then removed by z-score. Observations with a z-score greater than 3 (more than three standard deviations from the mean) are removed from the set by creating a subset `cbdata.z3`. The same statistics are calculated for the subset:

symbol	description	value
`cbdata.z3.count`	number of observations in `cbdata.z3`	284255
`tdur.z3.mean`	mean trip duration	616.47
`tdur.z3.var`	variance of trip duration	177353
`tdur.z3.sd`	standard deviation of trip duration	421.13
`tdur.z3.median`	median trip duration	502
`tdur.z3.min`	minimum trip duration	60

[1]See https://github.com/cmprince/UCSL/blob/master/R/ch1/ch1.Rmd for this document's R code.

symbol	description	value
`tdur.z3.max`	maximum trip duration	3355
`tdur.z3.range`	total range of trip duration values	3295
`tdur.z3.iqr`	interquartile range of trip duration values	433

The subset's quartiles are:

0%	60
25%	333
50%	502
75%	766
100%	3355

1.3 Discussion

The **central tendency** for the data in January, after removing the z>3 outliers, is that the average (mean) trip duration is slightly more than 10 minutes (616.47 sec). Half of the trips took less than (and the other half took more than) the median time of 502 sec, about 8 1/2 minutes.

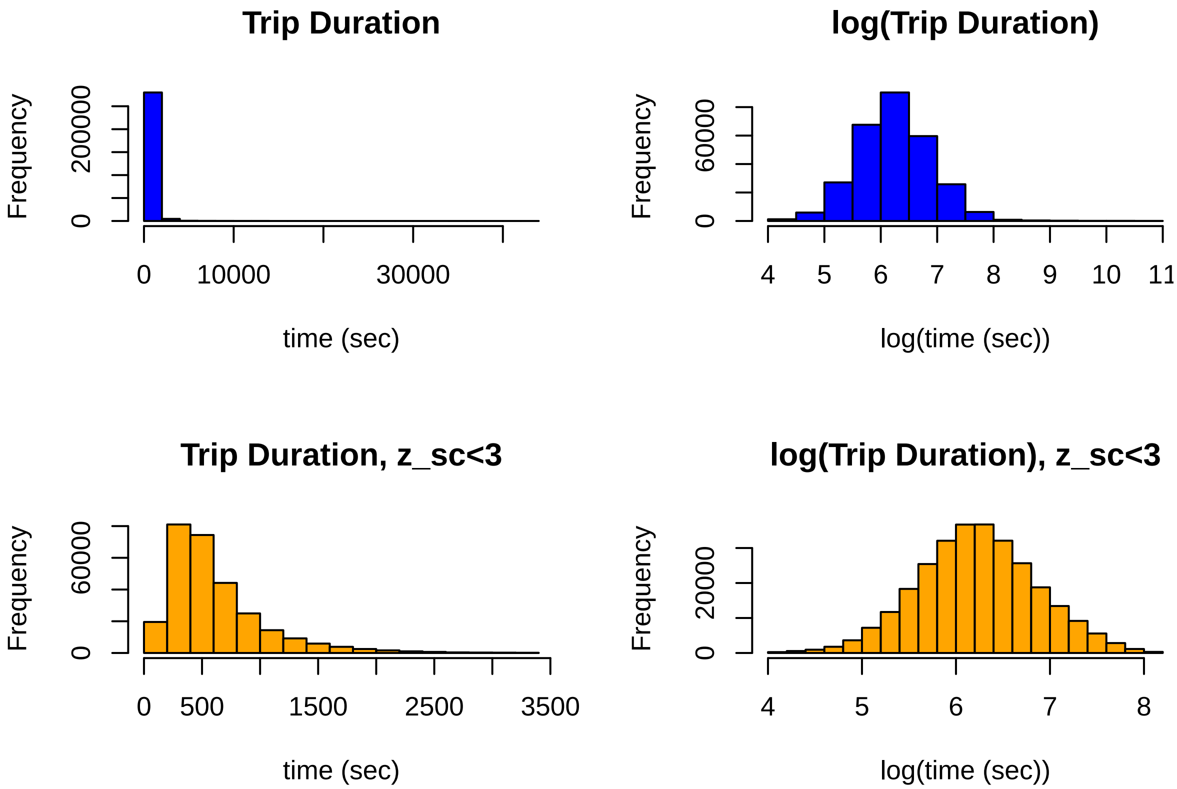
The **dispersion** of the data in January indicate that the middle half of all rides (again, after removing the z>3 outliers), given by the IQR spanned a range of 433 sec. The total range for all of the data is a little under an hour, 3295 sec.

Removing outliers had little effect on the quartile calculations: the median shifted just `tdur.median - tdur.z3.median =` 2 sec, and the IQR was reduced by only `tdur.iqr - tdur.z3.iqr =` 5 sec. However, the mean shifted by a significant amount, `tdur.mean - tdur.z3.mean =` 37.85 sec. This is due to removing `cbdata.count - cbdata.z3.count =` 1297 data points skewing the mean. The overall range reduced from 42963 sec to 3295 sec.

2 Visualization

2.1 Histograms for `cbdata` and `cbdata.z3`

Here we plot histograms for both the full set and z-score reduced set. The log-transformed data is also plotted, which is particularly useful for the full data set.



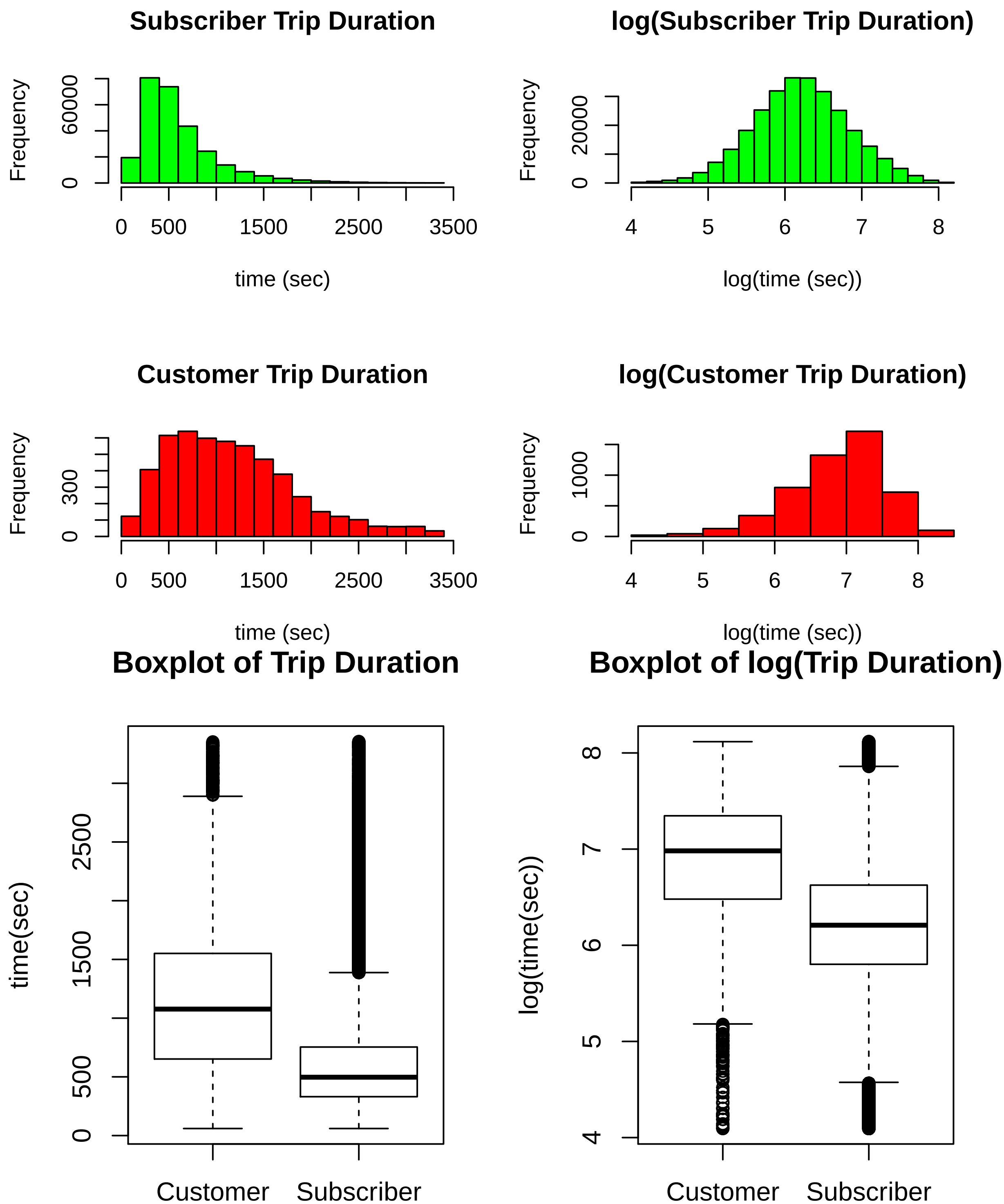
Visually, the log-transformed data appears to fit a normal distribution, though other tests (which we'll no doubt learn and are out of this assignment's scope) can assess this.

2.2 Visualization by `usertype`

First we use the `subset` function to parse out new data frames by `usertype`. There are two `usertype`s, `Subscriber` and `Customer`. After creating the subsets for both the full data set and the z-score reduced set, we count the observations and verify that we haven't missed any blanks or mislabeled points.

	`cbdata`	`cbdata.z3`
subset sums	285552	284255
total counts	285552	284255

Now we visualize the subsets by producing histograms and boxplots.



2.3 Discussion

Outliers in data sets will create longer tails in histograms and more data points beyond the boxplot whiskers. To visualize the complete data set, there will be a loss of resolution in both types of graphs.

In a histogram, there will be many sparsely populated bins in the tails, with most of the observations piled into just a few bins around the median.

In a boxplot, the outliers will dominate the axis along which the values are plotted due to their range. Effectively this compresses the IQR into a smaller space on the graph, making the visualization less effective.

From the discussion of the descriptive statistics above, we can hypothesize that the histograms and boxplots for the data sets including outliers will suffer the effects identified above.

Note that the shapes of the histograms for the two different usertypes are markedly different. Indeed, the log-transform of the `Subscriber` subset has the same normal-looking shape as the outlier-removed data at large. This is not surprising since `Subscriber`s account for 98.03% of the data. The histogram of the `Customer` subset, however, has a much different shape, and the log-transform does *not* appear normally-shaped. This suggests that the trip patterns of `Customer`s and `Subscriber`s are significantly different.