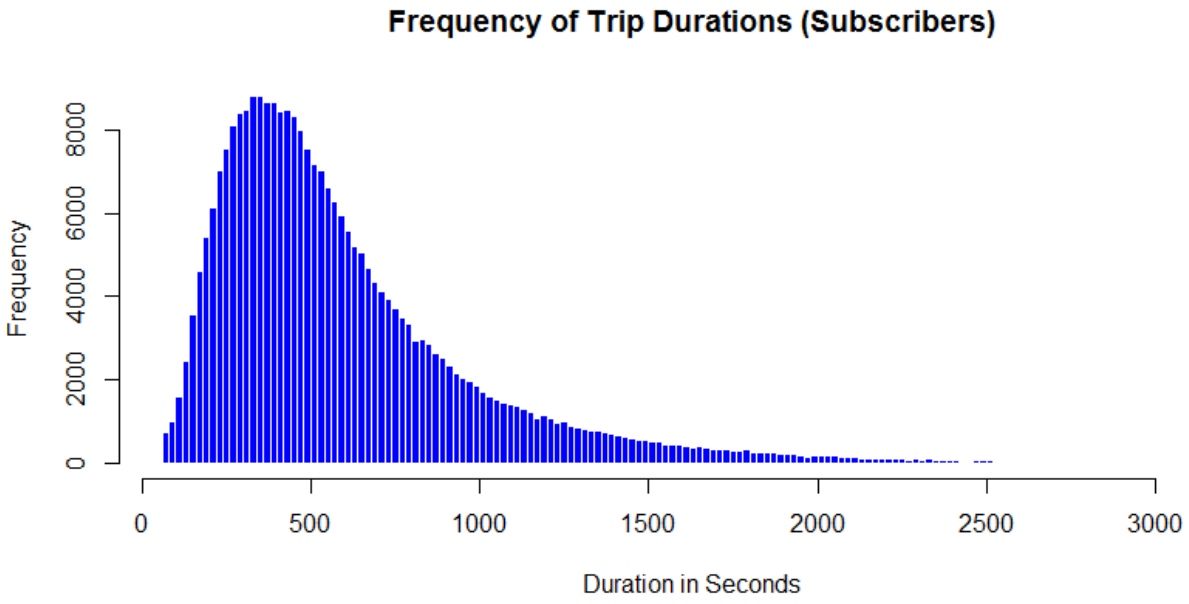
Jeremy Neiman

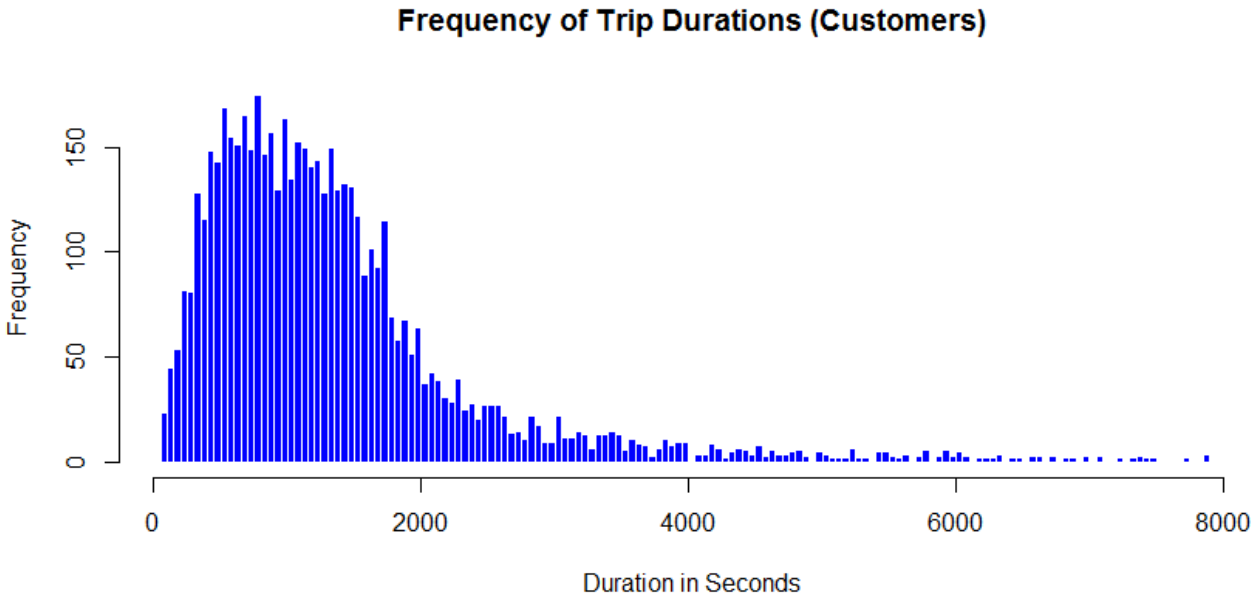
R Challenge #1

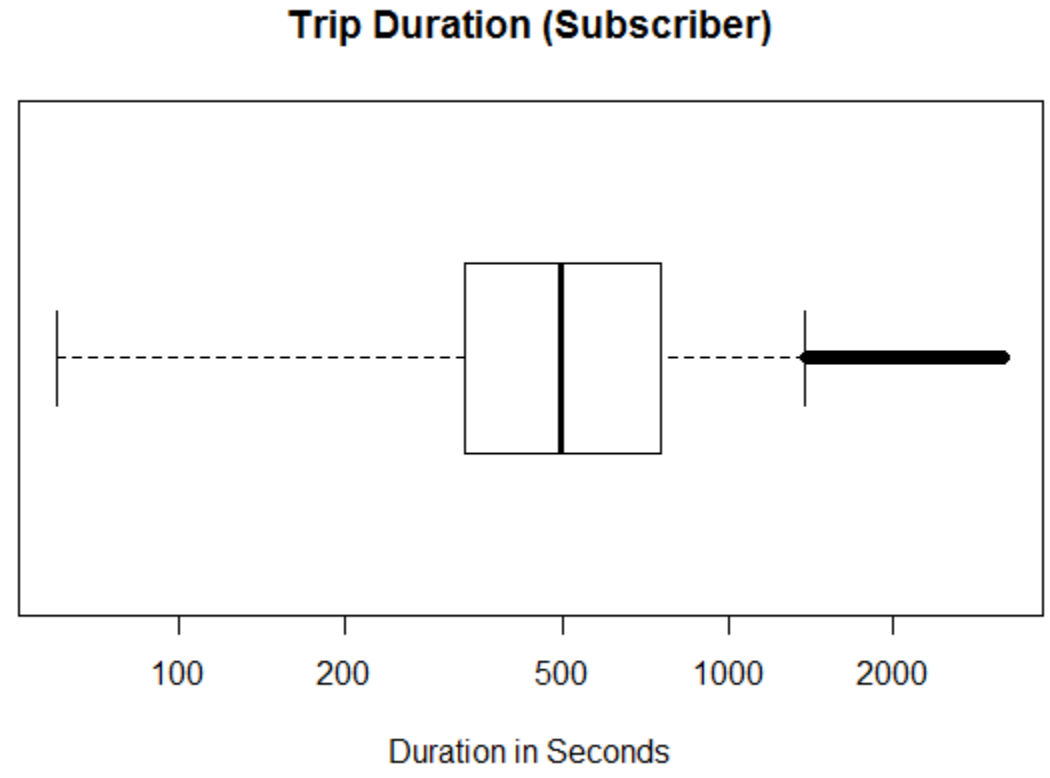
Summary of Subscribers and Customers with and without Outliers

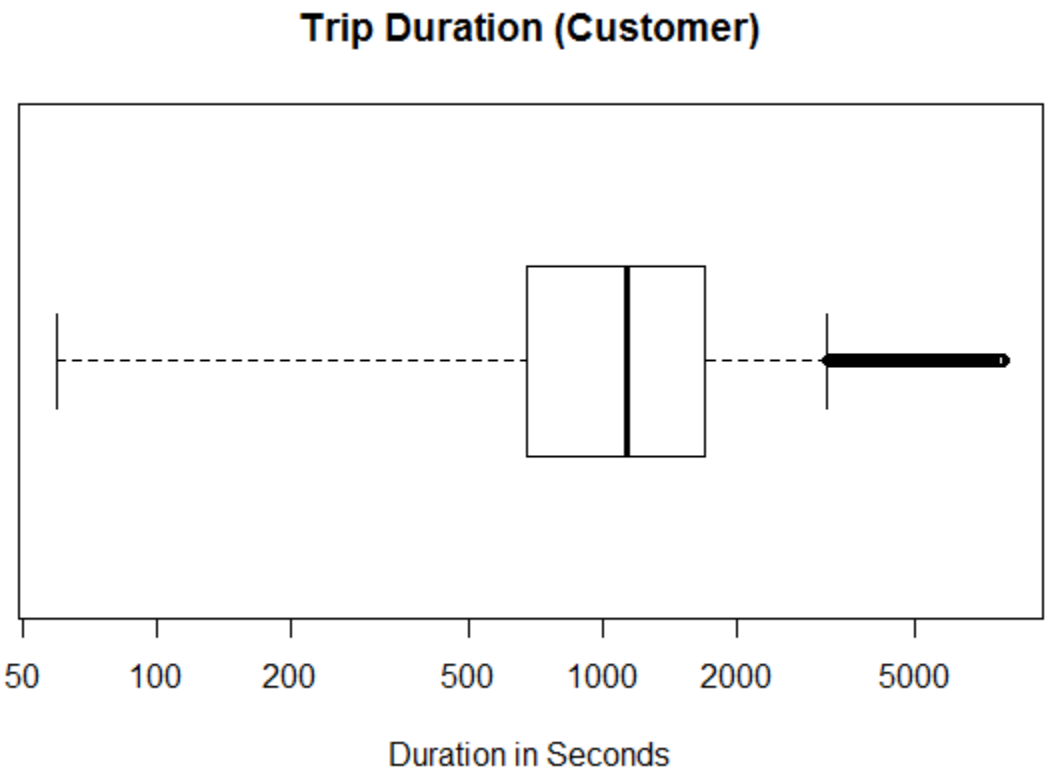
	Count	Min	1st Qu.	Median	Mean	3rd. Qu	Max	σ
Subscribers	279924	60	332	498	635.5	757	43020	848.48
Subscribers (w/o Outliers)	278972	60	331	497	605.3	753	3179	405.22
Customers	5628	60	684	1148	1589	1717	40660	2114.28
Customers (w/o Outliers)	5544	60	679	1129	1401	1689	7931	1142.76

Most rides for subscribers fall within 5-10 minutes and 10-20 minutes for customers. The data is skewed right, with no outliers on the low end and a small number of outliers far to the right. The outliers are so few in number that they don't have a large effect on the mean/median, but they about double the standard deviation and move the maximum values out to 5-14 times greater.


Frequency of Trip Durations (Subscribers)
Frequency
0
2000
4000
6000
8000
0
500
1000
1500
2000
2500
3000
Duration in Seconds


Frequency of Trip Durations (Customers)
Frequency
0
50
100
150
0
2000
4000
6000
8000
Duration in Seconds


Trip Duration (Subscriber)
100
200
500
1000
2000
Duration in Seconds


Trip Duration (Customer)
50
100
200
500
1000
2000
5000
Duration in Seconds

Outliers make it much more difficult to visualize detail in histograms and box plots. If the plots need to accommodate the whole data range, then the bulk of the data will be squashed into one small part of the plot. A histogram with the outliers would have most of the data squashed together in the same bins on the left, while most of the rest of the plot would be largely empty.