(To run my script, you need the following 4 files, which I have put on Google Drive: hubway_stations.csv, boston_weather.csv, hubway_trips.csv, and IMG_ELEV2005.tif; hubway_trips is ~150MB

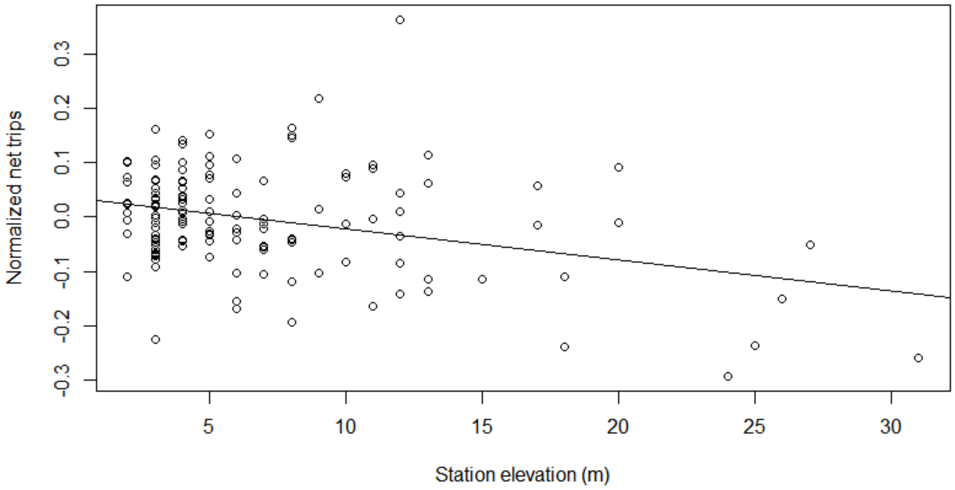
https://drive.google.com/folderview?id=0By0Yeqq6Gk4rMVI5a0tVRXhUWGs&usp=sharing)

For my final project, I analyzed data from the Hubway bikeshare program in Boston (http://hubwaydatachallenge.org/, specifically the 'Hubway trip history data,' and the 'Hack Day Treat'). Hubway allows registered members or "casual" users to take bikes from specific stations, and charges increasingly more for trips as they increase in duration. I looked at 4 questions: 1.) What makes some stations gain or lose bikes over time? 2.) How does the weather effect how people use Hubway? 3.) How similar are trips by casual riders vs registered members? 4.) Does age effect drip duration?

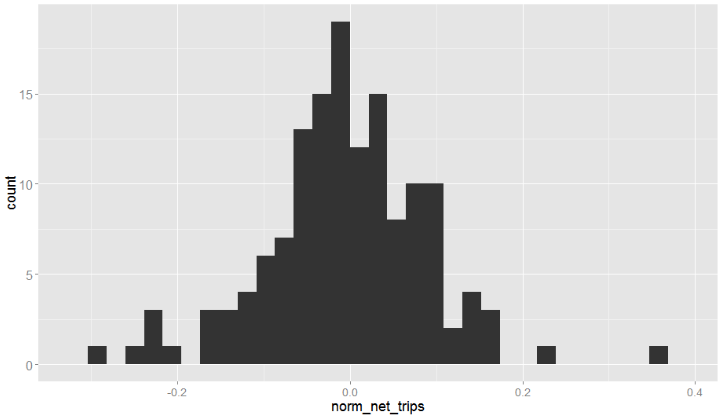
To see which stations gain or lose bikes, I started by looking at how elevation is correlated with bike usage. To get elevation data, I used a GIS package called raster to load in a .tif with an elevation map of Boston, and then looked up the elevation based on longitude and latitude. To determine whether a station gains or loses bikes, I calculated the normalized net trips:

$$\frac{\#\ trips\ ending\ at\ a\ station - \#\ trips\ starting\ at\ a\ station}{total\ number\ of\ trips\ at\ a\ station}$$

If more trips end at a station, meaning a station gains bikes over time, this metric will be positive. I explored the data, and simply plotted the normalized net trips for each station versus the station's elevation:

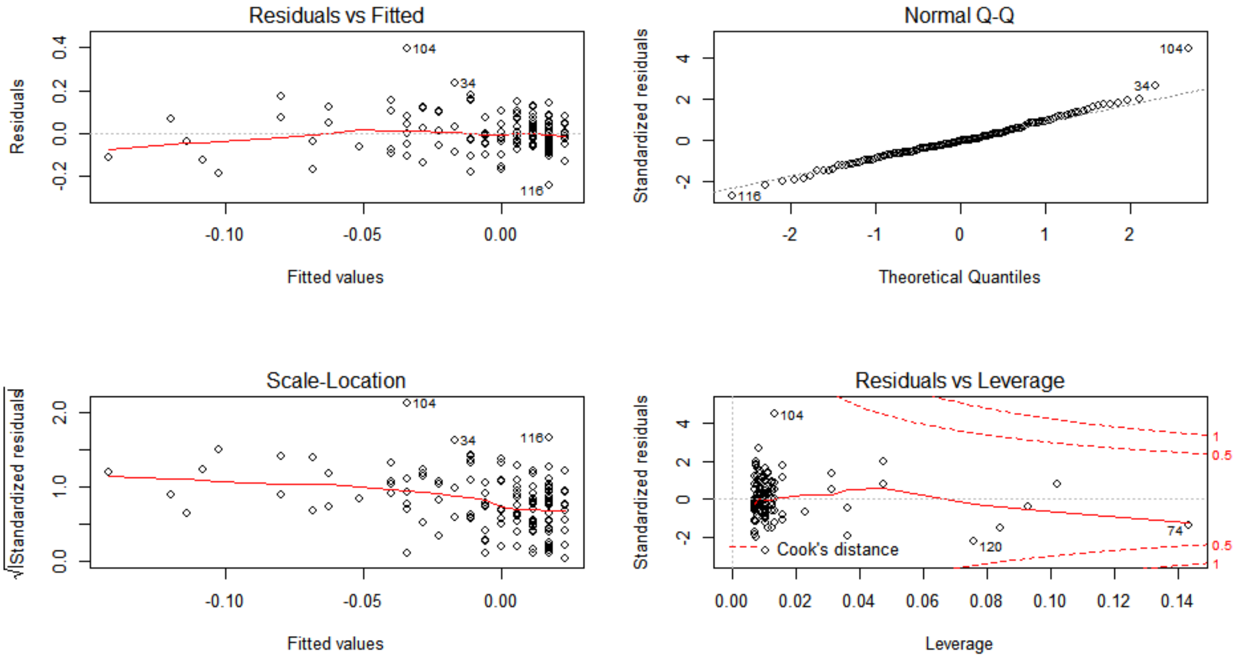


There appeared to be a trend where stations at high elevations lost 10-20% of their bikes on a per-trip basis. I wanted to fit this with a linear model, so I wanted to see if the normalized net trips metric was normal, and plotted a histogram of values:

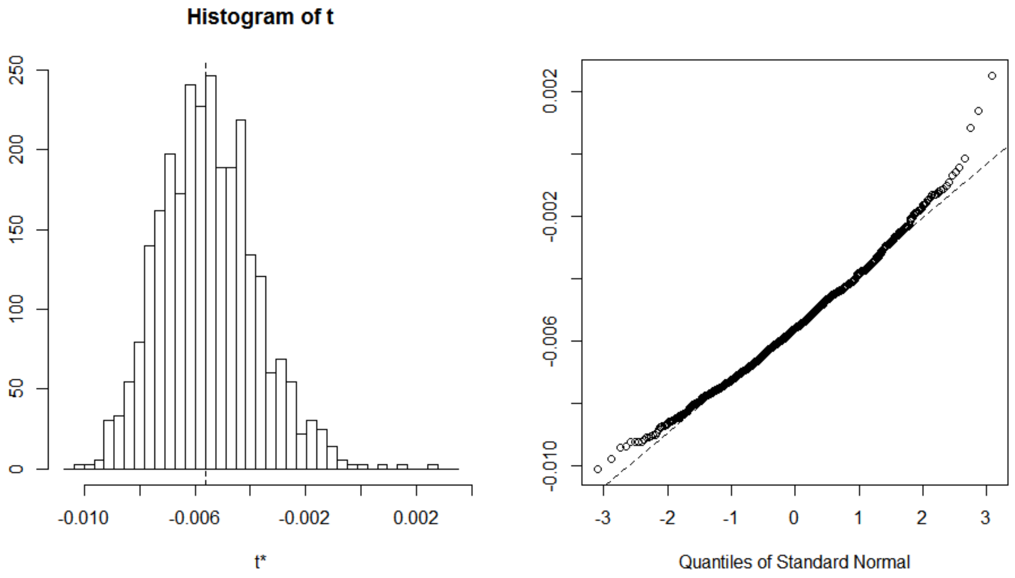


While it looks broadly normal to my eyes, it unfortunately failed a Shapiro-Wilkes test with p~0.001. I could not find a good normalization method to make the distribution normal, so I decided to just run a linear regression anyway, as the slope would be easily interpretable. The linear regression had an $R^2 = 0.11$, which means elevation only explained 11% of the variance in the net trips (fit plotted as the line in the first figure. The slope of the line was -0.005; this means that for every 10m higher a station is located yields a 5% decrease in bikes / trip.

To evaluate the fit, I plotted the residuals (see below). While the residuals fail the Shapiro-Wilkes test (p = 0.005), in general they seem well behaved, with only a slight increase in residuals at extreme elevation values, and a few points with moderate Cook's distance.

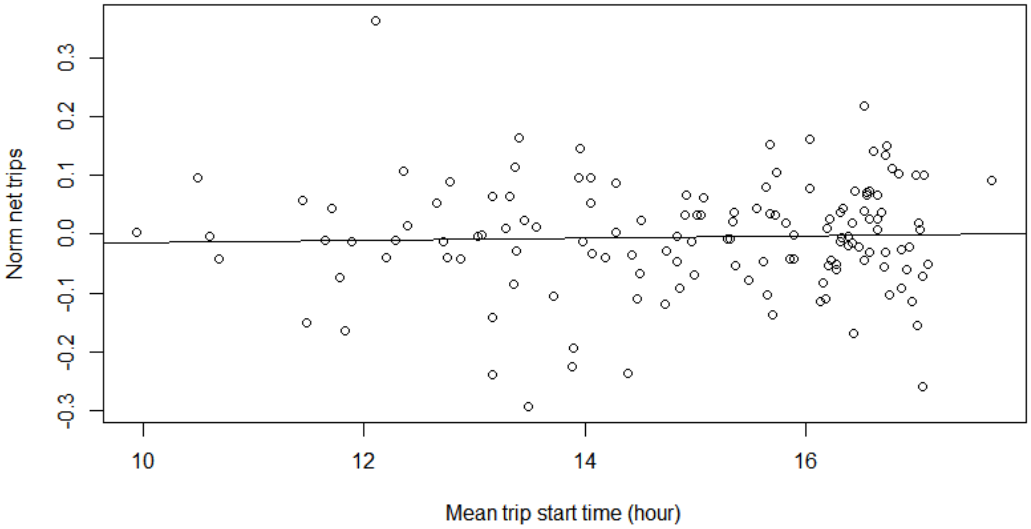


Since the residuals showed a high leverage at high elevation, I wanted to confirm the relationship was real. I bootstrapped the fit 1000 times, and plotted a histogram of the slope coefficient:



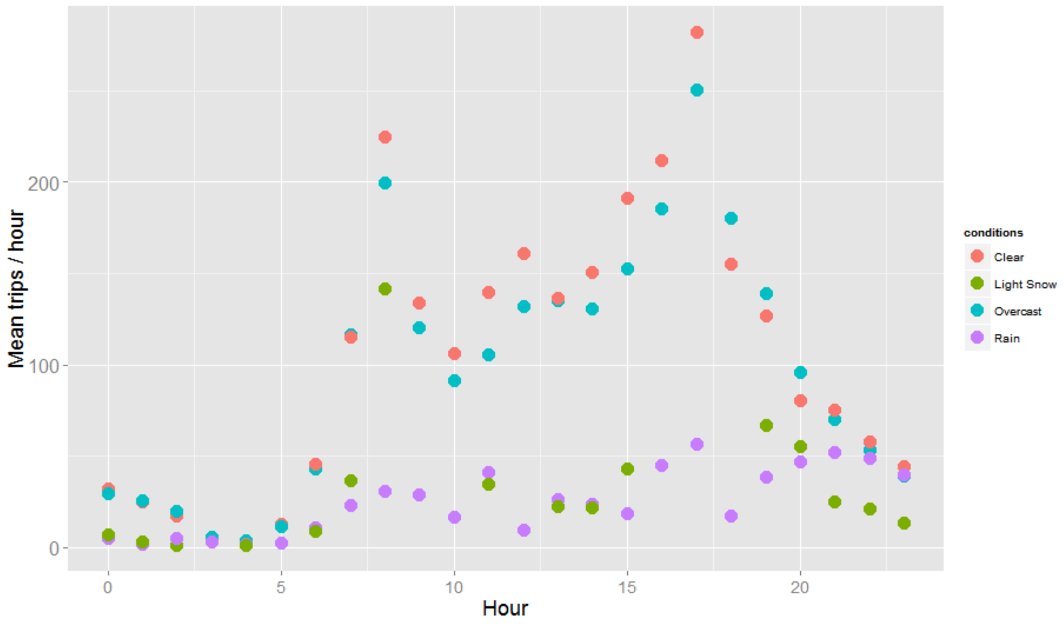
As the graph on the left shows, the bootstrapped slope is rarely near zero; the 95% confidence interval does not cross zero. So it looks like there is a real relationship between elevation and bike loss.

I tried to think of other variables that could influence which stations lose bikes. Since bike loss involves one way trips, I thought it could depend on time of day (e.g. commuting to work and busing home, or biking to parties and taxiing back). I plotted the net trip metric versus the circular mean of trip start times, and then fit it with a linear model (see below). The linear model yielded an $R^2$ of 0.001, so there was no relationship. Given that elevation only accounts for 10% of the variance, factors like local businesses or tourist attractions likely influence which stations gain and lose bikes.

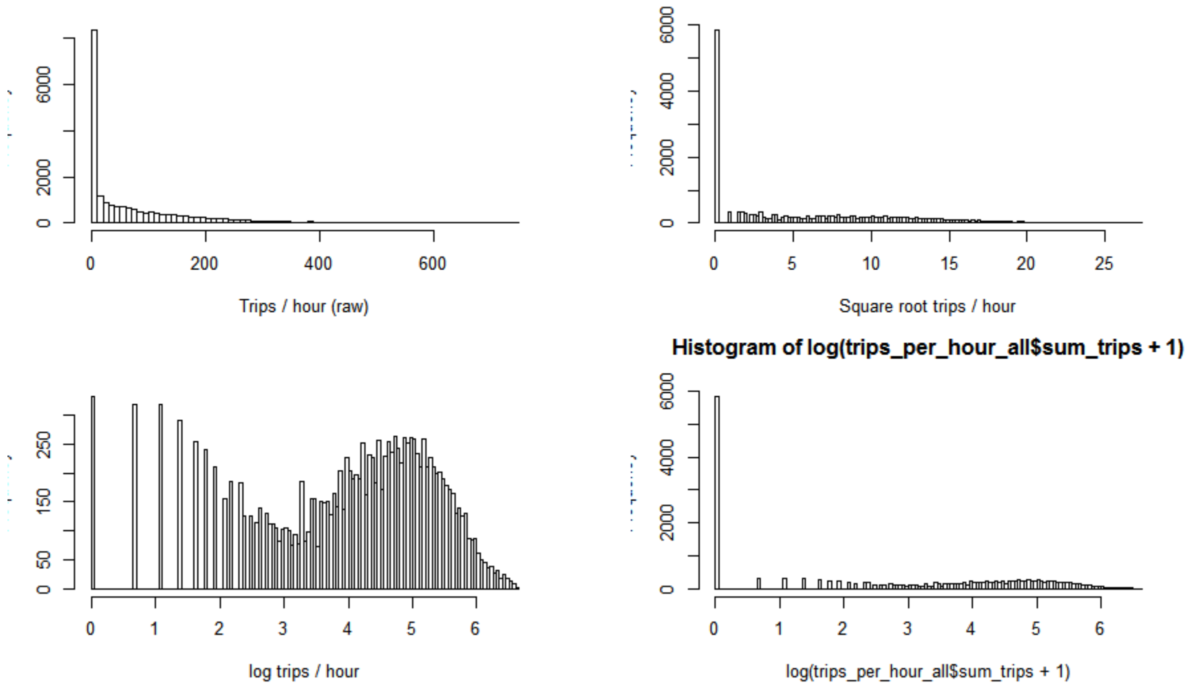


For my next question, I wanted to see how the weather influenced the number of trips. I scraped weather data from Weather Underground using the script provided earlier in the course. When loading the weather, I merged some labels like “Snow” and “Light Snow.” I also dropped hours that had no readings, or multiple readings. If I had more time / knowledge, I could have filled in missing timepoints, interpolating the temperature and conditions.

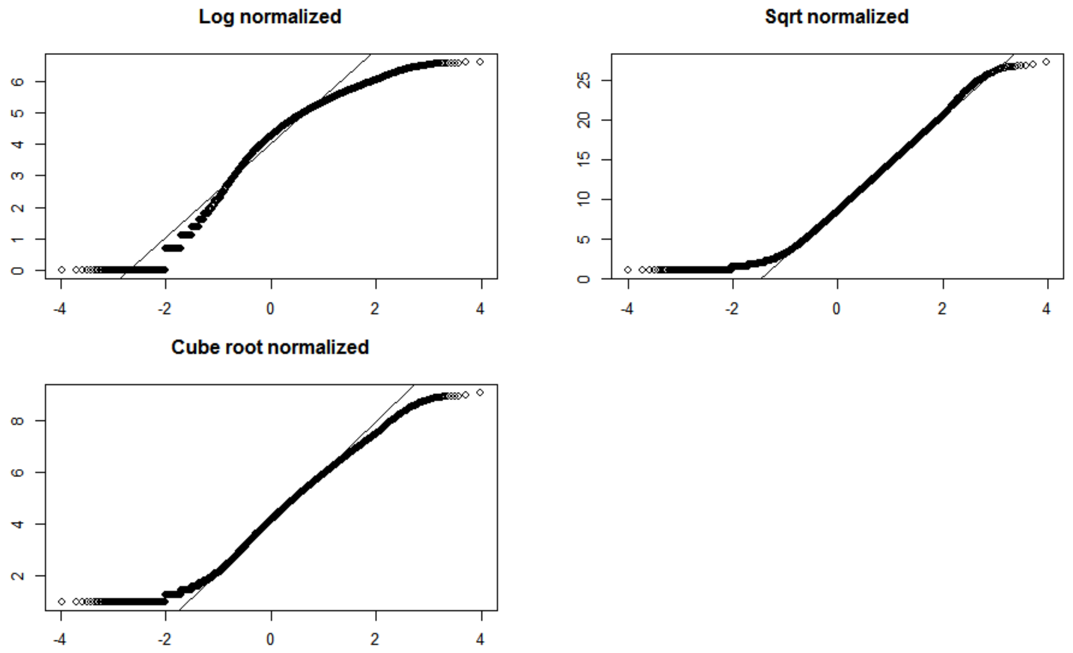
I started with a simple data exploration, and plotted the mean number of trips citywide every hour, for each time of day, and under four weather conditions.



The  # of trips is highly time and weather dependent! People bike the most in dry weather, and during the day. To model the number of trips per hour, I wanted to run a linear regression. I first started by plotting the distribution of trips using various normalizations.

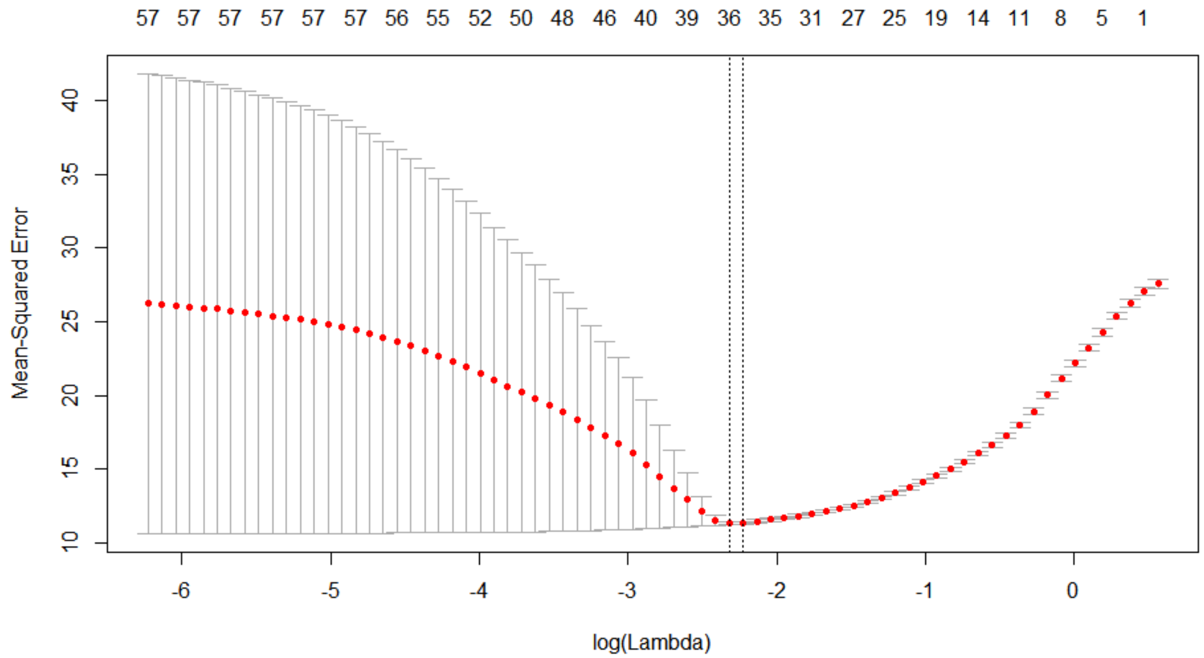


There are a lot of hours with no trips! If I knew how to, I would use a regression that could handle the large number of zero points, like Zero-Inflated Poisson, or Zero-inflated negative binomial. Unfortunately that’s a bit beyond me, so I tried normalizing the data, excluding hours with 0 trips. The Shapiro-Wilkes test fails on large data sets, so I decided to plot the quantile-quantile distribution for the # of trips under various normalizations



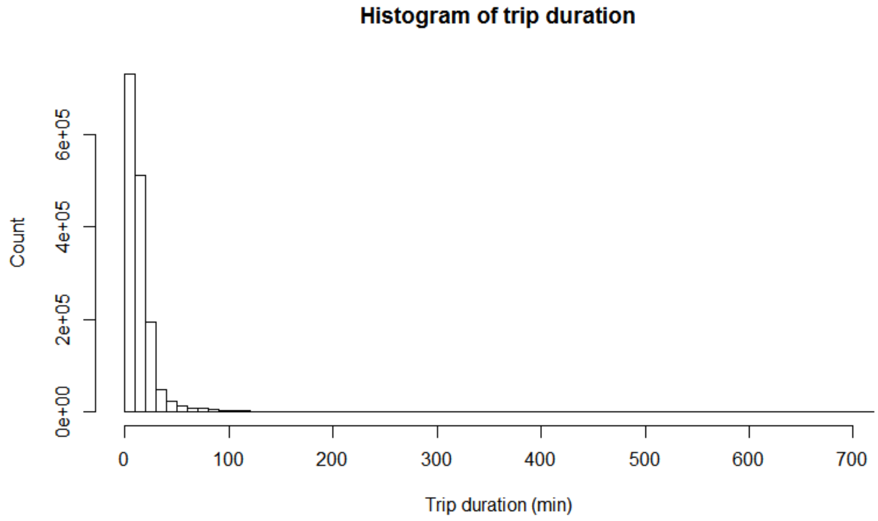
The best normalization seems to be the square root normalization. Now that we have somewhat normal data, we can do regression on it, using inputs of weather and time! I started by trying to do a poisson regression, since this is count data, but that did not work well, as the mean and standard deviation are not poisson. Instead I used a gaussian linear fit. Using all the variables, the linear model yielded an adjusted $R^2$ of 0.61, pretty good. Most coefficients were statistically significant, except for ones like wind speed, or visibility. The AIC score was 76,226.

Since there were a lot of variables, I tried using a cross-validated lasso regression to reduce them. Increasing the number of variables decreased lambda until it got to ~35 variables (see below).

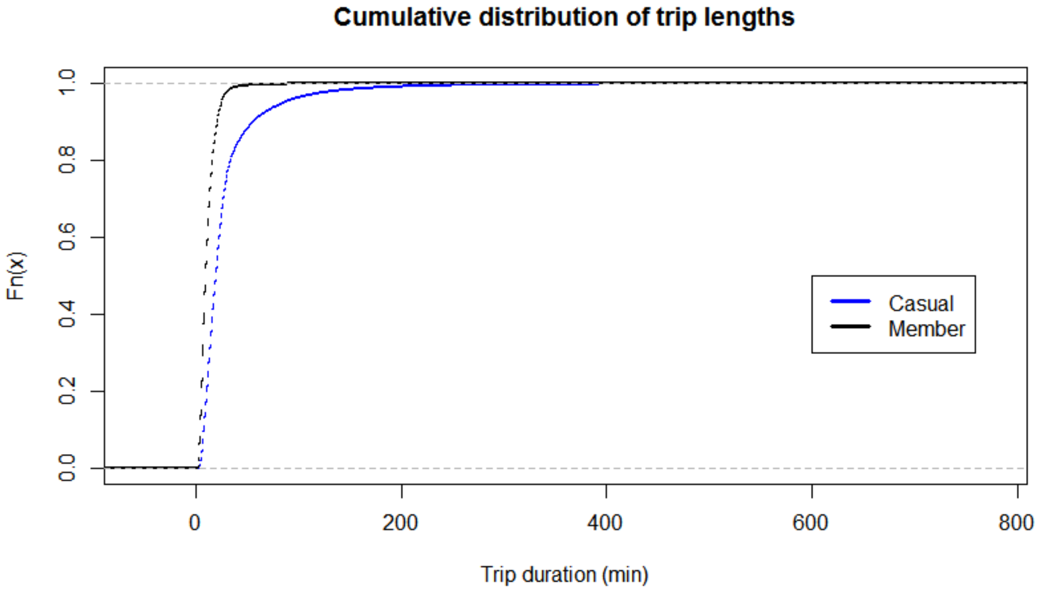


Of interest, some of the wind directions were chosen by the lasso, while others were not. Since rain and wind are correlated, I calculated the mean precipitation for each wind direction, and found easterly winds generally had higher precipitation. Since only some of the wind directions were important, I tried running the regression using the lassoed variables, and excluding wind direction. This yielded an AIC of 76,270, worse than before! When I ran the regression with the lassoed variables, including the wind, it yielded an AIC of 76, 227, nearly the same as before! It seems all the weather and time variables are important to determine how many trips there are (I could only get a lower AIC by using a step-wise model selection for AIC).

For a simple question, I wanted to know how similar casual versus member rides were. Like the # of trips / hour, the trip duration was also highly skewed:

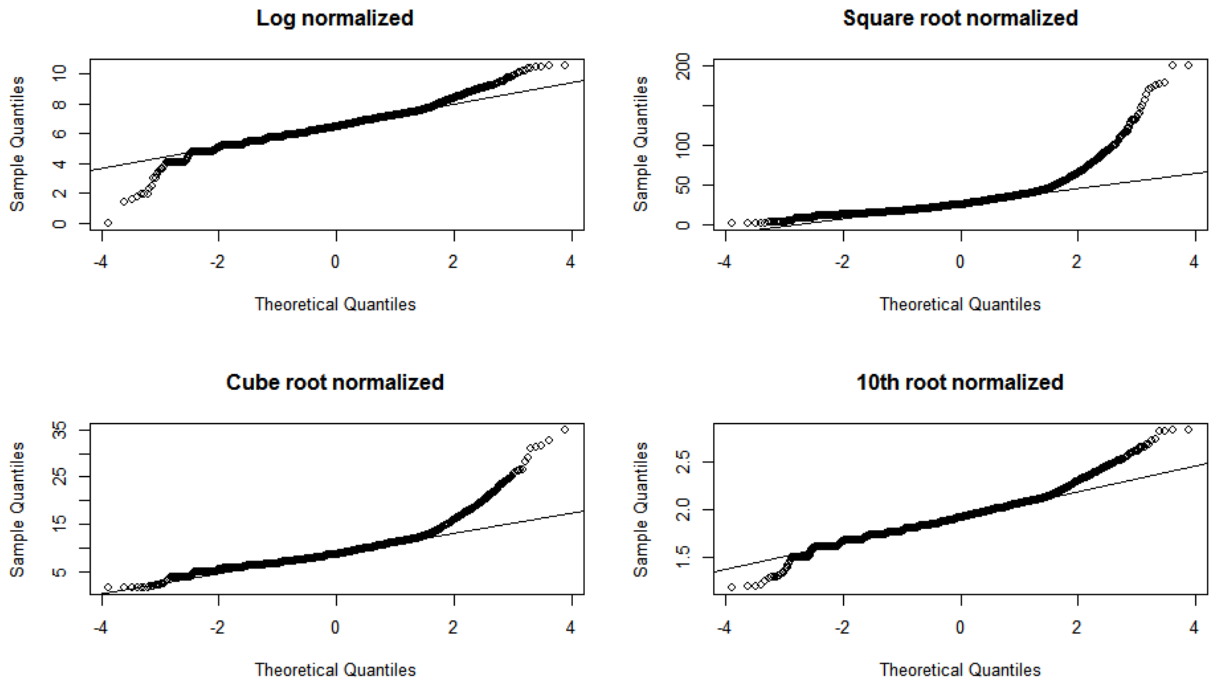


To investigate the distribution, I plotted the cumulative distribution of ride durations for casual versus registered riders:

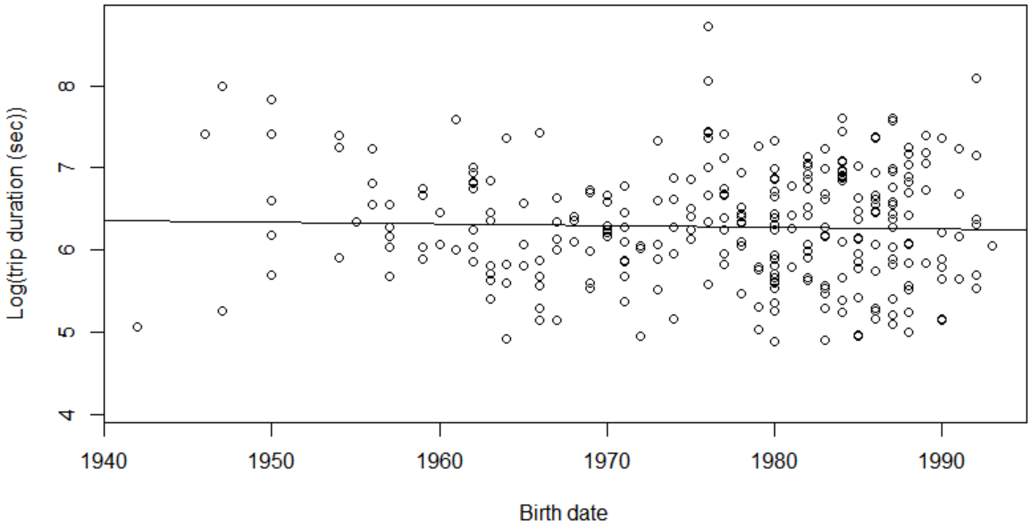


Since these distributions are not normal, I used a KS-test, and found they were different with p < 0.0001. Surprisingly (to me at least), casual riders take longer trips! Perhaps since the don't have to be anywhere, they are just touring the city.

Finally, I wanted to correlate rider age with trip duration, thinking that older riders might take shorter trips. I started by trying to find a good normalization for ride duration by plotting quantiles:



This time I chose the log normalization. Then I plotted the log trip duration against rider birth date:



There doesn’t seem to be much of a trend. I fit the model with a linear fit, which yielded an R$^2$ of approximately zero. Biking is ageless.

For future directions, there are a few possibilities. One, I could try modeling the stations as a graph, with number of trips between them defining edge weights. You could then try to determine what makes certain stations well connected. I also think there is more to investigate with figuring out why certain stations lose bikes over time. Here you might try to find out distance between stations, or investigate specific stations that are losing bikes, and find out what is unique about them (although that is less systematic). Looking at some other people’s analyses, I could have expanded my weather regression to include weeday vs weekend, or try to factor in the casual / member riders ratio.