Biking in the Rain

Summer Bicycle Commuting – Weather Dependent?

As I enjoyed my cool, rainy bike ride to Statistics class I wondered to myself if other people feel the same way. Is your average commuter cyclist a fair-weather rider, or do they throw on a rain coat and go ride through the puddles? I realized I may be able to answer my question when I passed the Capitol City Eco-Totem, which counts bike traffic in both directions on the path.

Data Qualifications

If I am to answer whether or not bicycle commuters were resilient to the precipitation, I wanted to subset for a few environmental variables.

- Subset for months when school wasn't in session (July, August).
- Subset for months that had reasonably good weather throughout (July, August).
- Subset for days where folks typically work (M-F).

This led me to selecting the months of July and August when the majority of college students had left town, and the weather was typically good.

Data Sources

Dane County has 2 bike path counters which are operated by an organization called eco-public. From scraping the JSON that feeds their website, I was able to get two dataframes which include a date, and count for total bikes that crossed the counter. I combined these two dataframes into one.

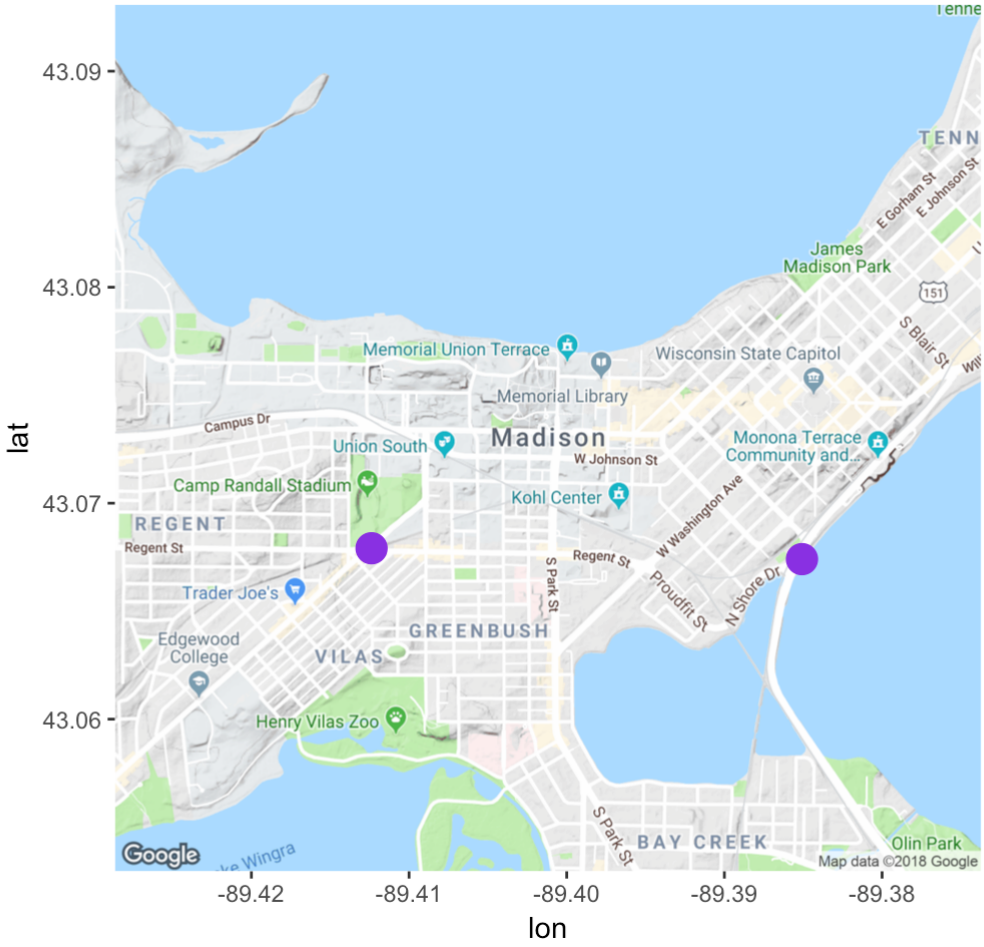
I sourced my weather data from the NOAA station located at the Dane County Regional Airport. I selected years 2015 - 2018.

Bike Path Counters and Location

http://www.eco-public.com/public2/?id=100020865 (http://www.eco-public.com/public2/?id=100020865)
http://www.eco-public.com/public2/?id=100016754 (http://www.eco-public.com/public2/?id=100016754)

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=madison+wisconsin
&zoom=14&size=640x640&scale=2&maptype=terrain&sensor=false
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=madis
on%20wisconsin&sensor=false
```



Data Used for Analysis:

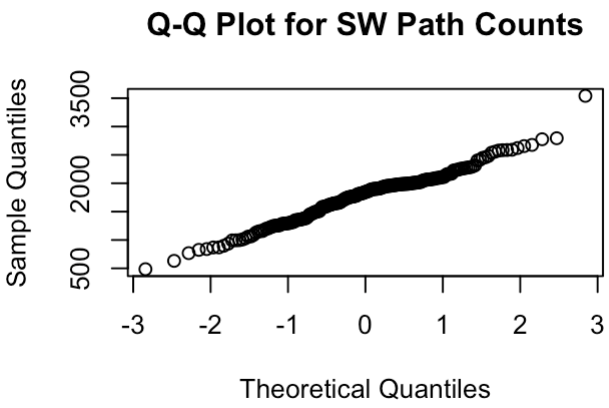
```
##         date count temp temp_min temp_max prcp
## 1 2015-07-01  6491 63.4     55.9     77.0 0.00
## 2 2015-07-02  5464 60.3     46.0     72.0 0.00
## 3 2015-07-03  5586 62.9     46.0     77.0 0.00
## 4 2015-07-06  2802 74.3     61.0     82.9 0.00
## 5 2015-07-07  5239 66.2     59.0     80.1 0.77
## 6 2015-07-08  5804 61.7     51.1     71.1 0.00
```

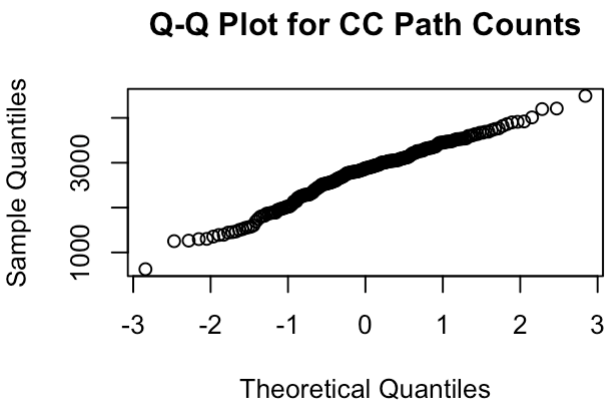
Question

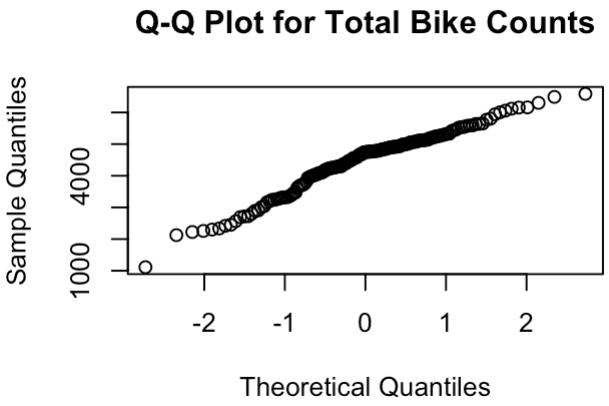
Is there a statistically significant relationship between the number of Madison commuters counted on our bike paths relative to the amount of precipitation on a given weekday in July & August?

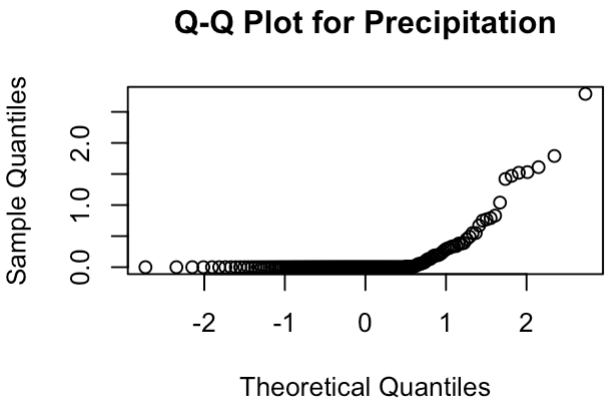
Hø: There is no relationship between precipitation and commuter counts

H_a: There is a relationship between precipitation and commuter counts
Assumptions First step is to visualize our populations and determine whether we have a large enough sample, and whether our sample population is normally distributed.





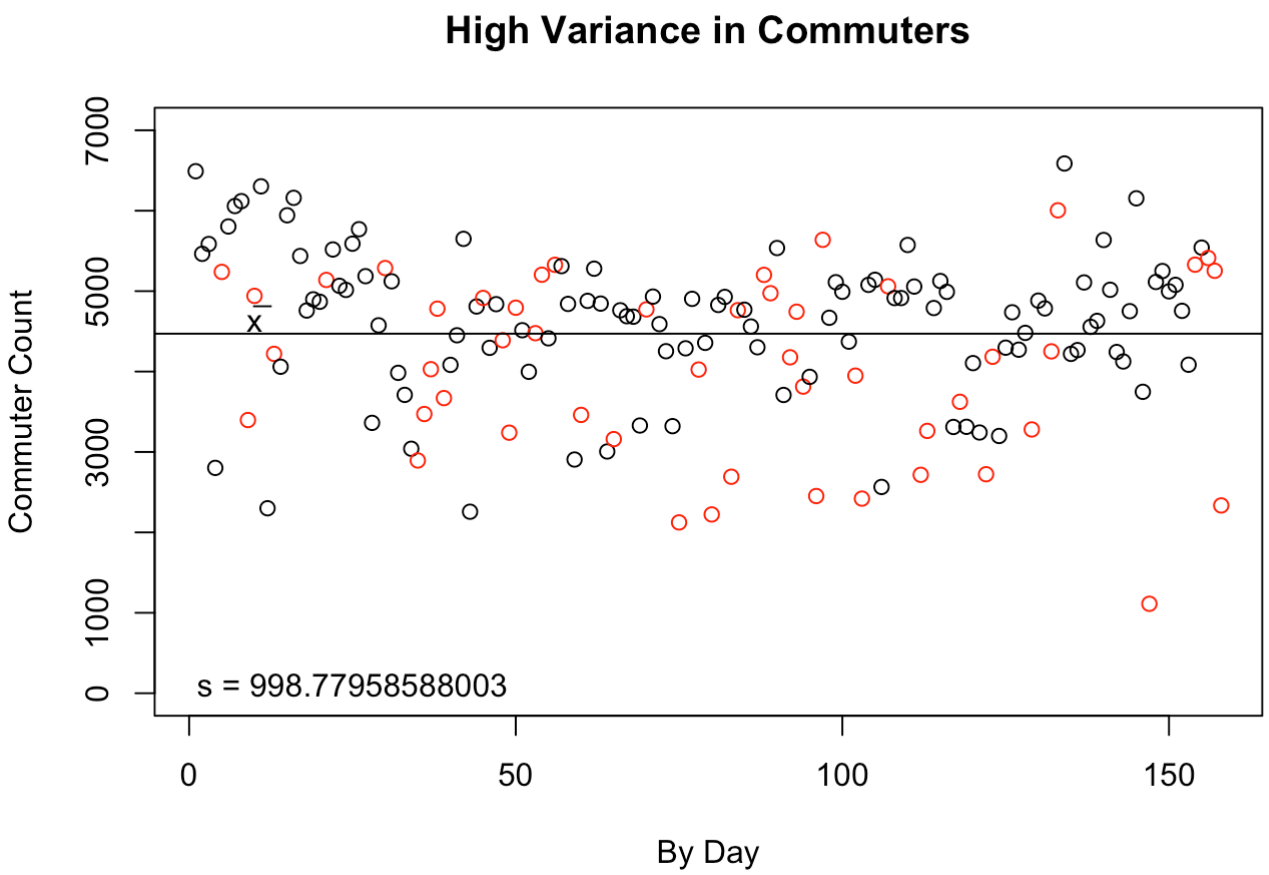




In light of the Q-Q plots that were generated from our data, we can assume normal distributions for all of the bike count data – however it looks like our precipitation sample is not normal as it is left skewed.

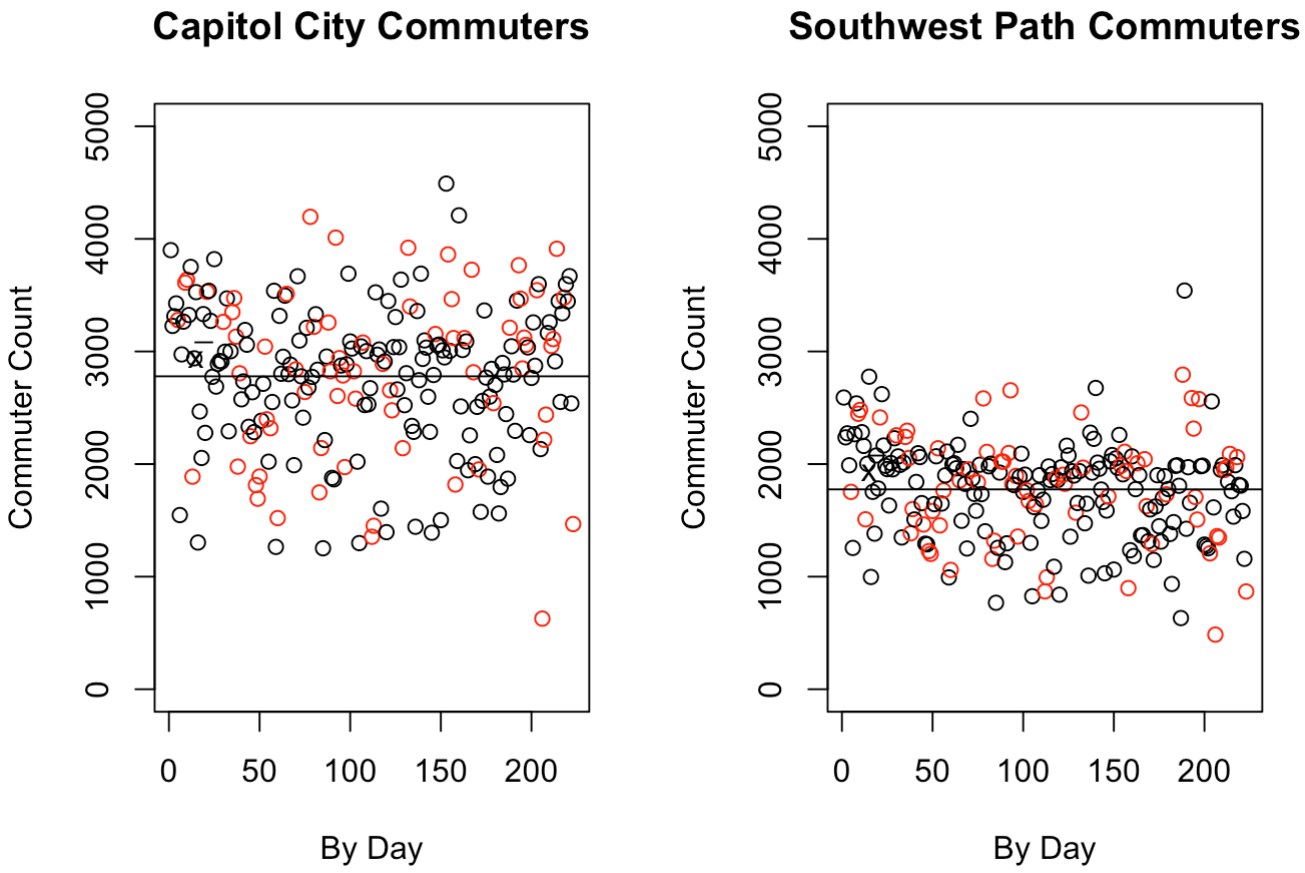
During my exploration, it was also apparent that there is a very high variance in daily commuter numbers for our 158 observations.

s = 987.17



This high variance is also represented in each individual bike path counter, so I will assume this variance is not due to a faulty operation of one of the path counters. Or perhaps long term construction that had one of the counters offline.

s(capitol_city) = 665.70
s(southwest_path) = 441.26

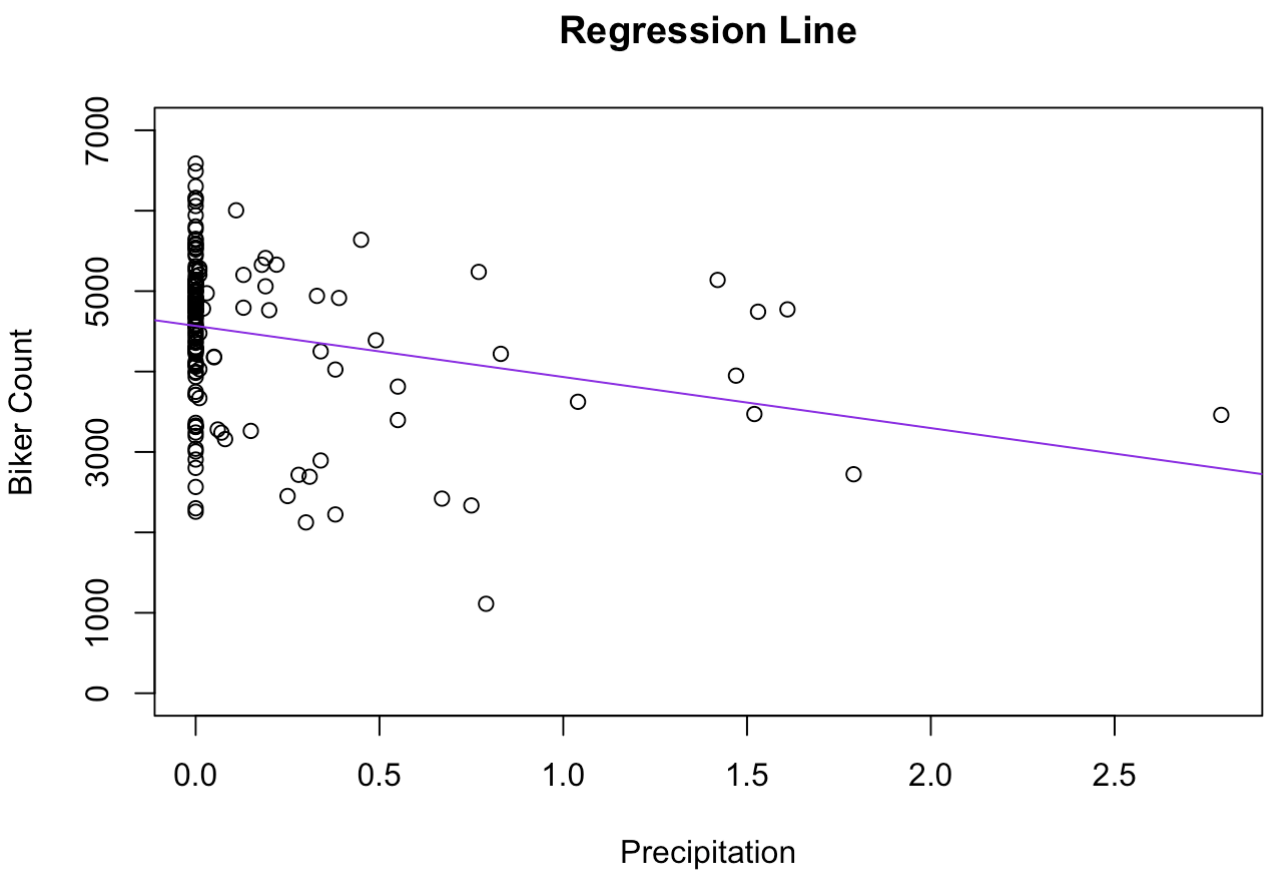


Regression

With the awareness that my precipitation distribution is not normal on account of the majority of data points being 0, I'm going to do a quick regression anyways – mostly for fun.

```
summary(lm(df$count~df$prcp))
```

```
##
## Call:
## lm(formula = df$count ~ df$prcp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2954.1  -478.2   192.3   591.6  2018.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4568.19      82.59  55.309  < 2e-16 ***
## df$prcp      -635.51     193.30  -3.288  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 969 on 156 degrees of freedom
## Multiple R-squared:  0.0648, Adjusted R-squared:  0.0588
## F-statistic: 10.81 on 1 and 156 DF,  p-value: 0.001248
```



Results

In light of our regression analysis, it appears there is a statistically significant relationship between commuter counts and precipitation on any given day in July or August.

p = .00125

Since our p-value is less than .05, we could cautiously consider rejecting **Hø**, however since our precipitation data was not from a normal population I can't fully support this significance. It is interesting nonetheless, and gret exploration.

T-Test

Since our precipiation sample is not distributed normally and our ridership counts are, this is the perfect example to subset the commuter counts data into two sets (precipitation/non-precipitation) and perform a Welch's T-Test. This should enable us to determine a relationship between the mean commuter counts on rainy days versus non rainy days. I have subsetted the data into two groups, with one being 49 counts on days of precipitation, and the other being 49 counts on days of no precipitation.

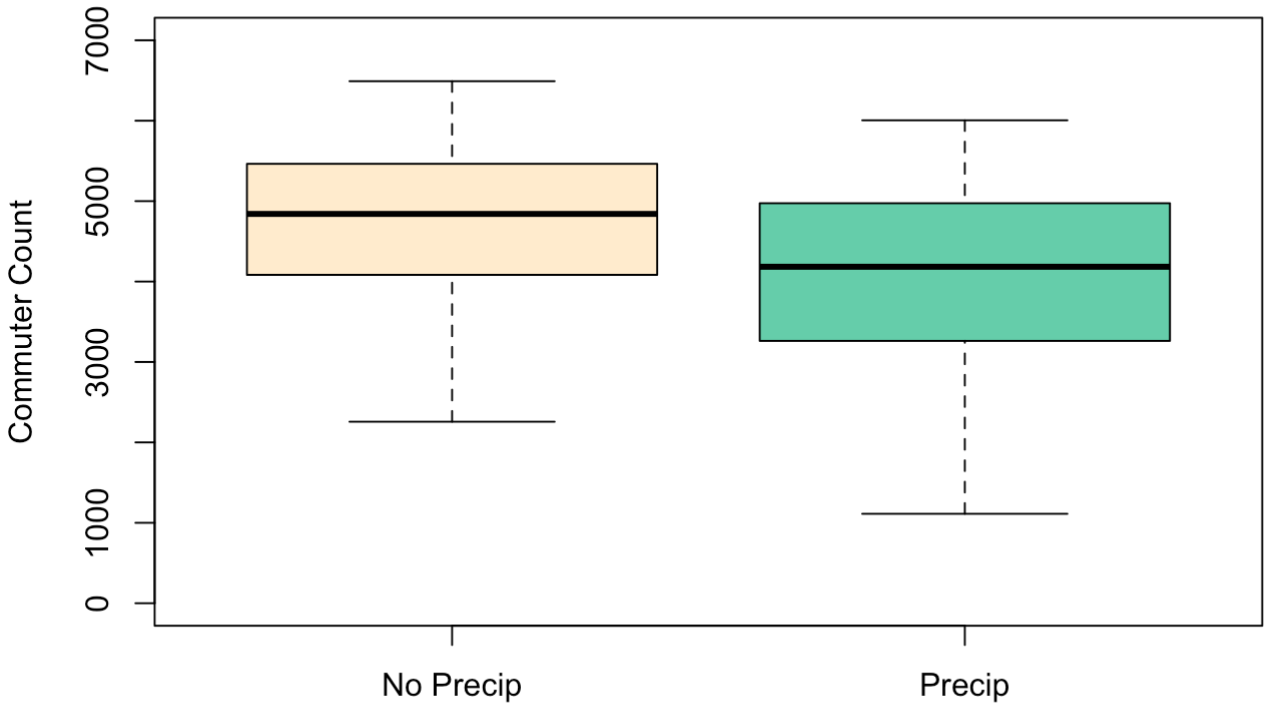
Question

Is there a statistically significant relationship between the number of Madison commuters counted on our bike paths relative to the amount of precipitation on a given weekday in July & August?

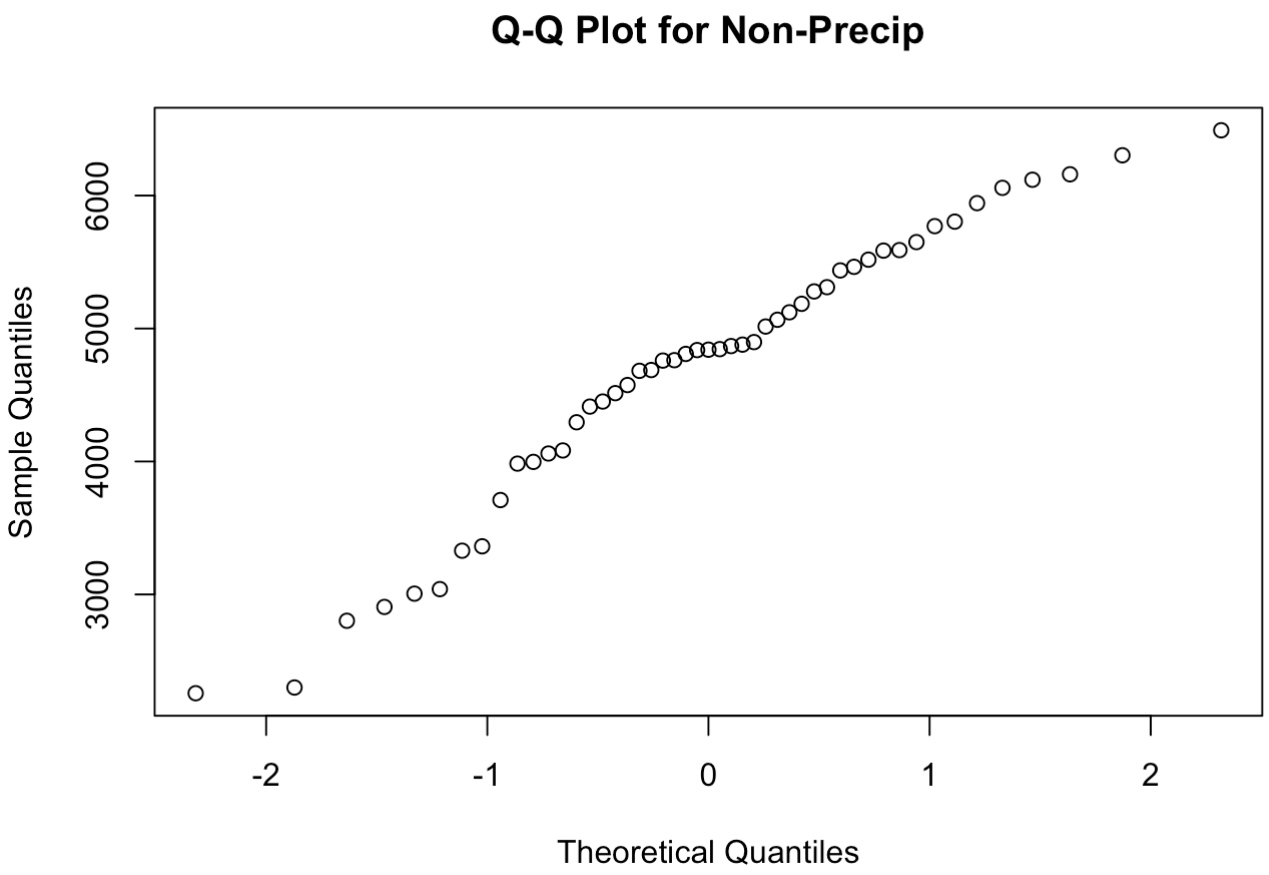
Hø: μ(commuter_counts_with_precip) - μ(commuter_counts_no_precip) = 0

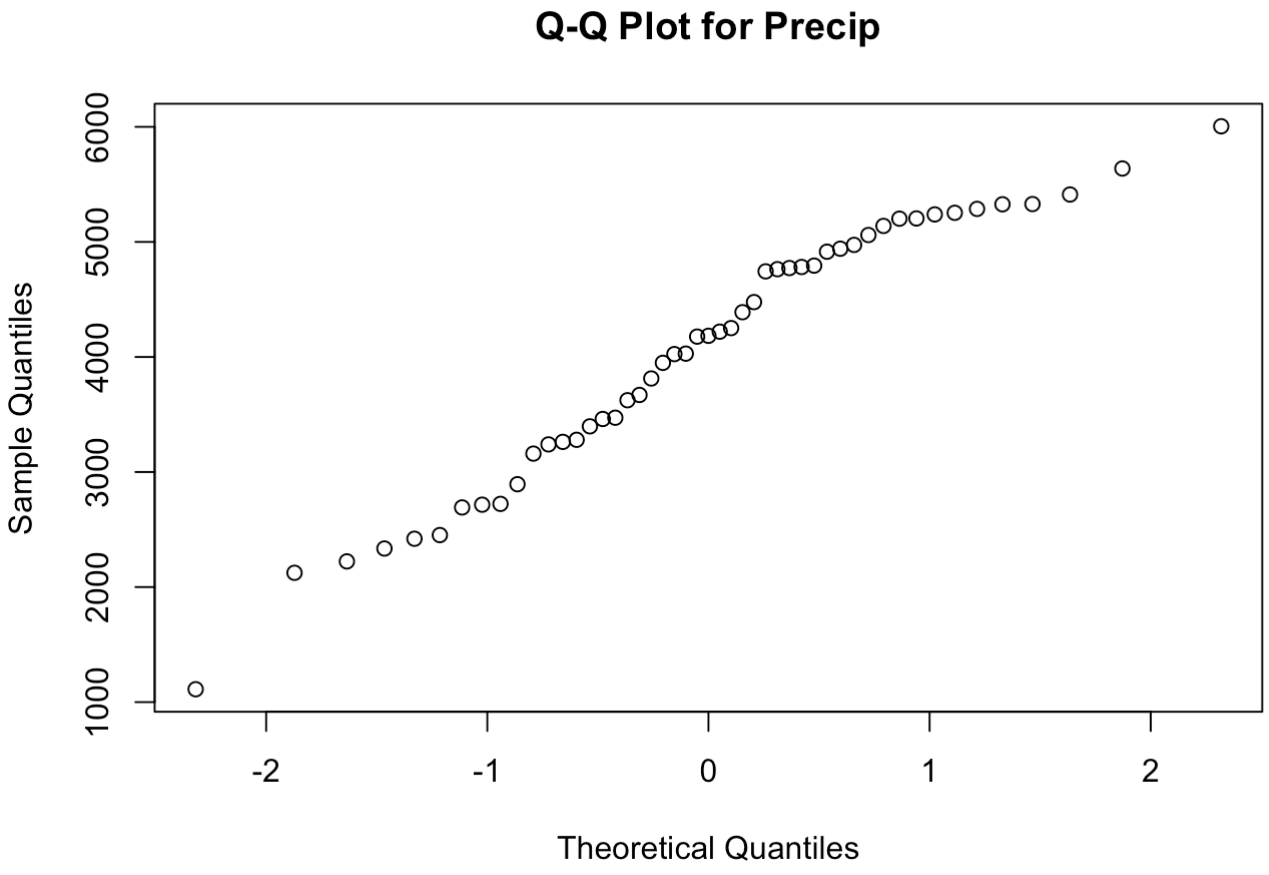
H_a: μ(commuter_counts_with_precip) - μ(commuter_counts_no_precip) ≠ 0

```
boxplot(non_rain$count, rain_counts$count, ylim=c(0,7000),names=c("No Precip","Precip"),
 col=c('blanchedalmond','aquamarine3'), ylab='Commuter Count')
```



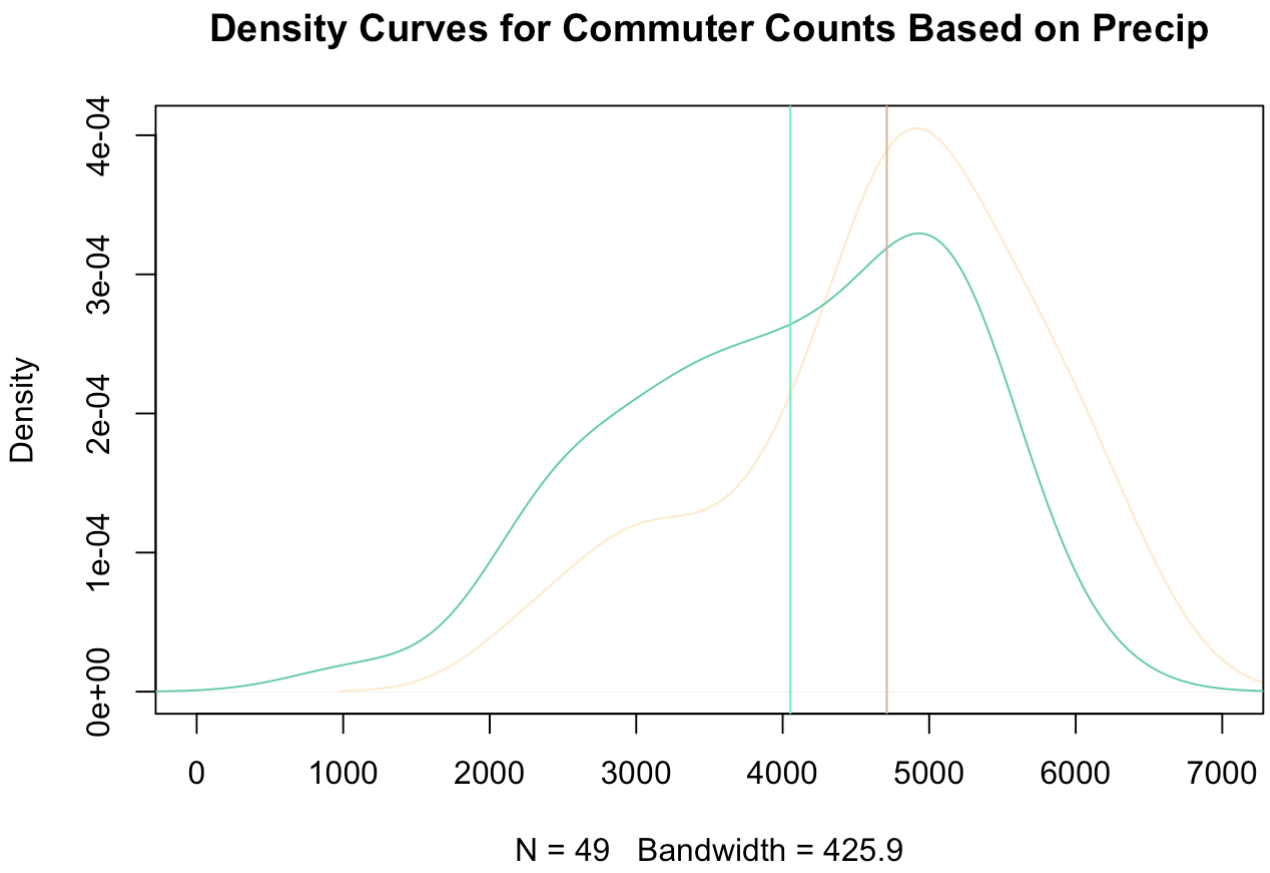
Assumptions Let's examine the Q-Q plots of these data and see if we have normal distributions.


Q-Q Plot for Non-Precip
Sample Quantiles
3000
4000
5000
6000
-2
-1
0
1
2
Theoretical Quantiles


Q-Q Plot for Precip
Sample Quantiles
1000
2000
3000
4000
5000
6000
-2
-1
0
1
2
Theoretical Quantiles

In light of the Q-Q plots, it appears that both our Precip/Non-Precip samples come from normal populations to which we can perform a T-test on.

```
plot(density(non_rain$count), col='blanchedalmond', xlim=c(0,7000), main='Density Curves
 for Commuter Counts Based on Precip')
lines(density(rain_counts$count), col='aquamarine3')
abline(v=mean(non_rain$count), col='tan')
abline(v=mean(rain_counts$count), col='aquamarine2')
```



By visualizing this graph it would be hard to tell if there's a significiant difference between the two curves as both populations have roughly the same count as their highest density. By visualizing their means, we can see how each population is swayed in either direction.

Welch T-Test

Now we can perform our Welch T-Test:

```
##
##  Welch Two Sample t-test
##
## data:  rain_counts$count and non_rain$count
## t = -3.0009, df = 95.478, p-value = 0.003434
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1094.3191  -222.9462
## sample estimates:
## mean of x mean of y
##  4052.082  4710.714
```

Results

p-value = 0.003434

With a p-value < .05, we can reject **Hø**. The data are strong evidence that the population mean commuter counts are different on days of precipitation v. days of no precipitation.

Additional Questions

- Perform the same test for commuter countage during the academic year – are college students more likely to ride in the rain? Do they have a choice?
- Look at other variables such as temperature and wind speed.