Analysis Of Bike Rentals
for the city of Washington D.C.

Ramchandra Kommera, Kyle Stiebeling

Introduction:

In this project we have access to the bike rental data for the capital city of D.C. We begin by looking at the parameters that we think might affect the rates of bikes ridden in the city. We want to explore this idea with respect to multiple types of variables; holidays, seasons, temperature, and humidity just to name a few.

Part 1 The Variables in Discussion:

First we look at our data's headers/ titles to see which data would be of interest. Since we are trying to find an overall relationship between bike usage rate and the conditions affecting this bike usage we consider all variables that would affect all the riders, regardless of casual or registered bike users. Below are the factors we considered:

1. **Cnt -** total count of bike users, registered and casual combined
2. **season-** fall,spring,summer,winter
3. **month-** month in the year
4. **Year-**2011 0r 2012 year

5. **workingday-** if it is working day or not
6. **holiday -** whether it a holiday or not
7. **weekday -** what day of the week
8. **windspeed -** speed of wind
9. **Weather -** clear, rain or snow , mist
10. **Atemp -** feeling temperature

Out of all the values here we did not try regression with a few variables because we felt like these variables were not applicable to our analysis. Some of them being: **year***(since we are not comparing a change)*, **month***(felt like 'season' gave us more information about the trends rather than months),* **weekday***(since we were doing a big picture analysis rather than specific analysis, and workday is a more in depth variable relating to the time of a week).*

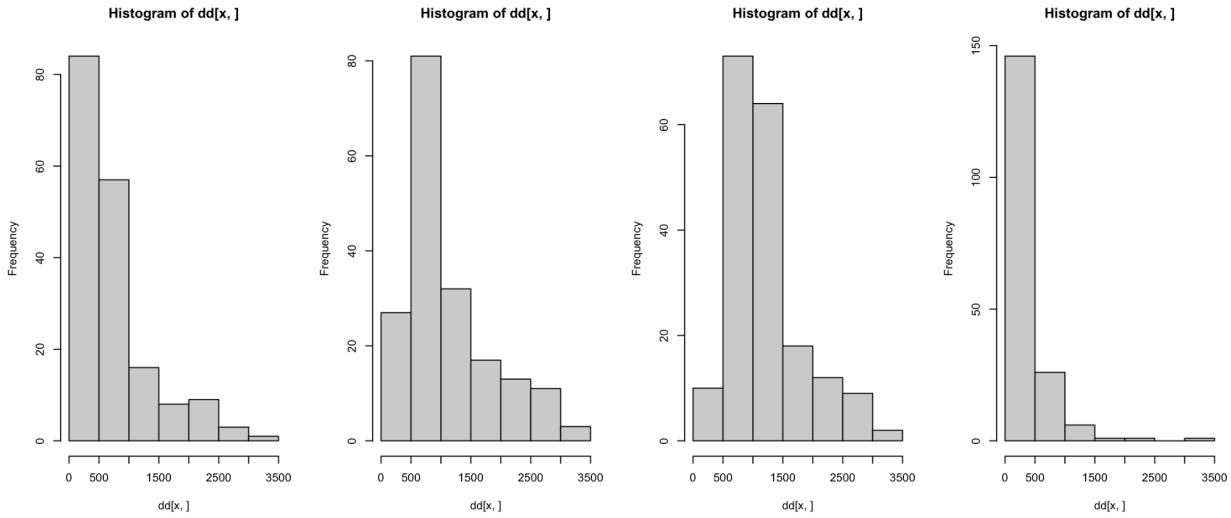
Season:

There are 4 primary seasons in this dataset; Fall, Spring, Summer, Winter. After conducting some normality testing we see that seasons are not normally distributed. But we move on with non parametric tests, and try to find a relationship between seasons and total count of bike rented.

$H_0 =$ *Seasons model is not a good predictor for total bike rental rate*

$H_1 =$ *Seasons model is a good predictor for the total bike rental rate*

Here is the histogram plots for total count of bikes taken in each season. In the order of : Fall, Spring, Summer, Winter respectively.



As we can see clearly that particular seasons have more riders than other seasons. For example the winter bike plot [No.4 plot] shows us high right skeweness and conversely, summer shows the least right skeweness. The less it is right skewed, the more people are taking bikes.
These assumptions can be backed up. Only summer and winter are statistically significant (p-values < 0.05), meaning that it is unlikely that this is due to chance. The value of the F-statistic is 128.8 with a p-value of less than 2.2e-16, which means that the model significantly predicts bike usage *Alternate hypothesis is true,(F-value= 128, p-value<0.05).*

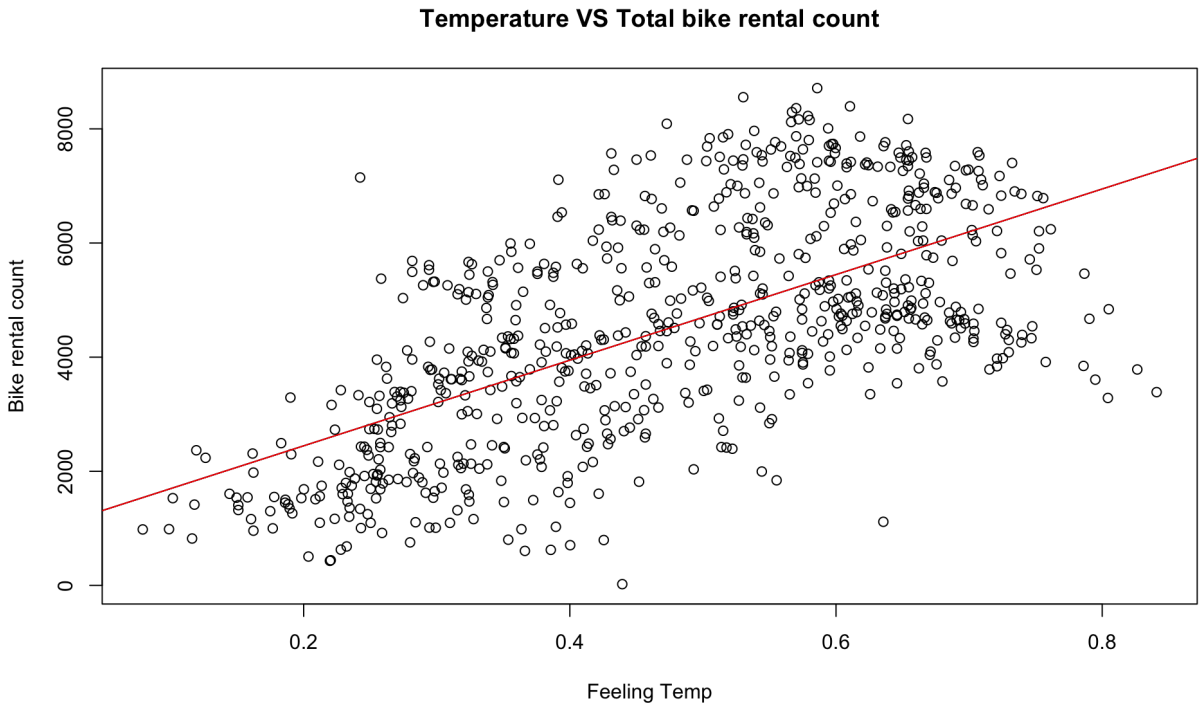
Atemp:

This is the temperature that a person feels opposed to what it actually is. This would possibly make a huge difference since if it is cold outside, people are probably less willing to use a rental bike. We chose ‘atemp’ instead of ‘temp’ because both are effectively almost always the same, but also a person wouldn't generally check the temperature and ride the bike,

but instead go off of how they feel. And hence we decided to use ‘atemp’ instead.

H_0 = *Feeling Temp model is not a good predictor for total bike rental rate*

H_1 =*Feeling Temp model is a good predictor for the total bike rental rate*



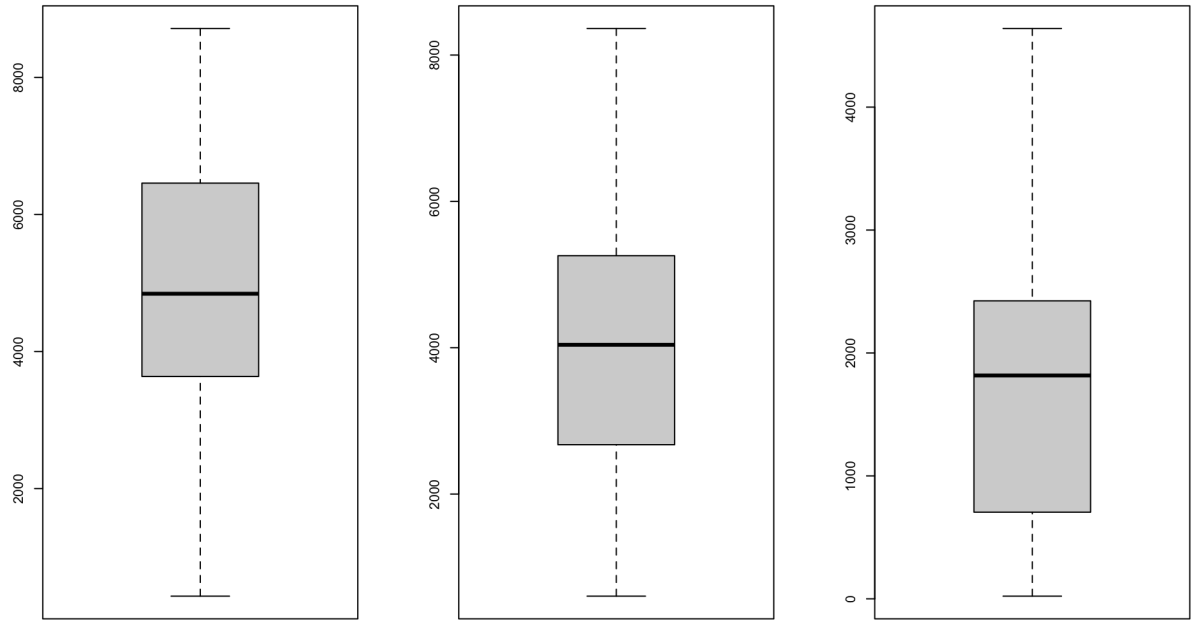
From the graph, as the feeling temperature increases the total bike rental rate also increases. The red line shows us a linear increase. The p-value for ‘atemp’ predicting ‘cnt’ was significant, indicating that we must reject the null hypothesis and accept the alternative hypothesis (p-value<0.05). This means that atemp is a good predictor of bike rental usage.

Weather

The data classified weather into three categories: Clear, mist, and rain and snow.

H_0 = *The weather model is not a good predictor for total bike rental rate*

H_1 = *The weather model is a good predictor for the total bike rental rate*



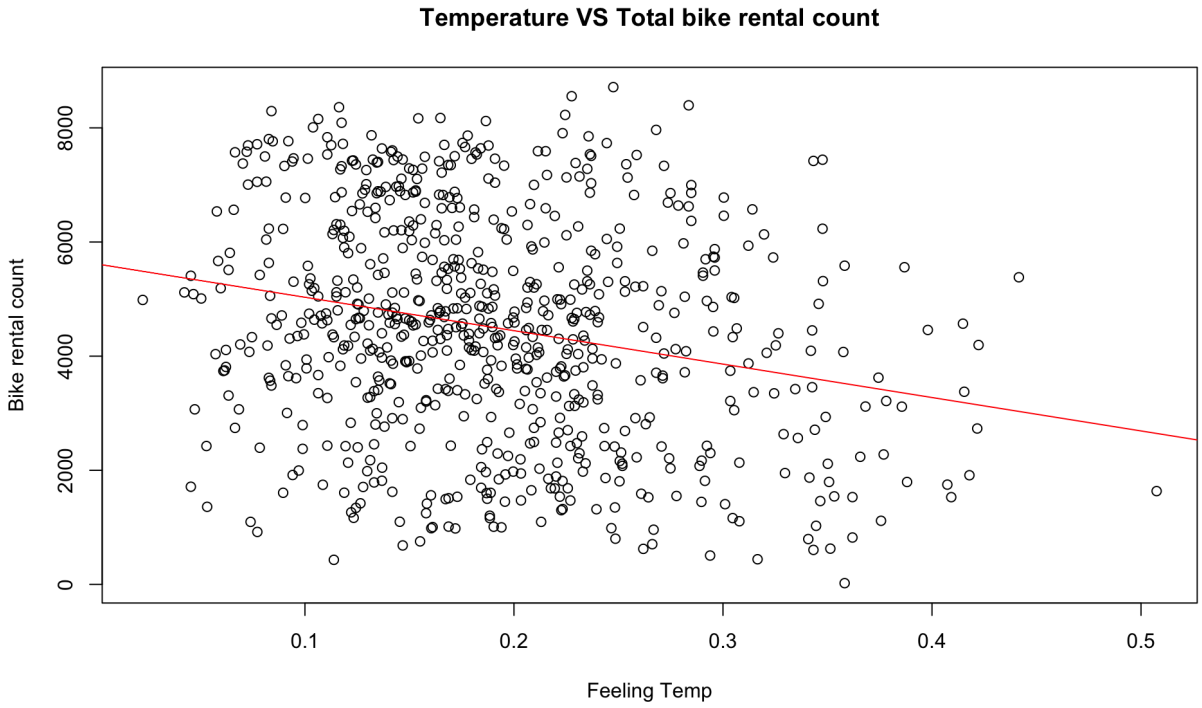
After running a simple regression model, using the ‘weather’ and ‘total bike rental rate’ variables against each other, we obtained a significant p value for all of the predictors (clear, mist, rain and snow). Considering this, we must reject the null hypothesis of no difference and accept the alternative hypothesis, indicating that the weather variable is a good predictor of rental bike usage.

Windspeed

The wind speed variable informs us on how fast the wind was moving on a specific day

H_0 = *The wind speed model is not a good predictor for total bike rental rate*

H_1 = *The wind speed model is a good predictor for the total bike rental rate*



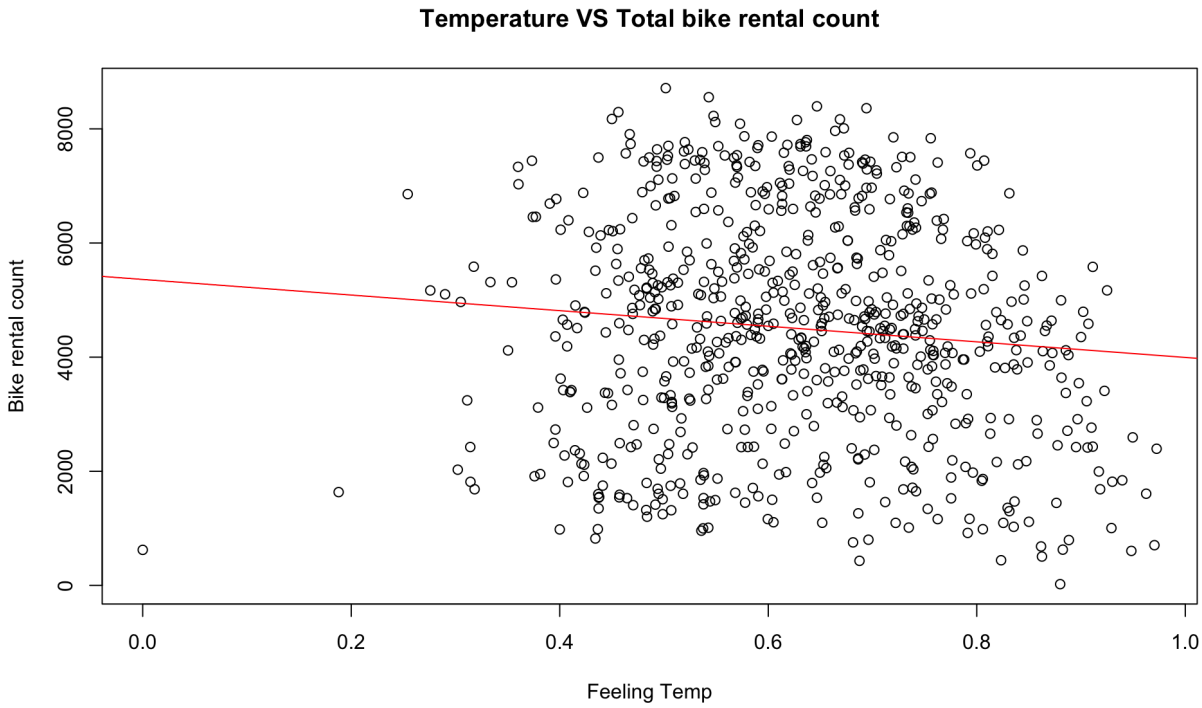
After running a simple regression model using the ‘wind speed’ and ‘total bike rental rate’ variables against each other, we obtained a significant p value under 0.001. Considering this, we must reject the null hypothesis of no difference and accept the alternative hypothesis, indicating that the wind speed variable is a good predictor of rental bike usage.

Humidity

The humidity variable informs us on the normalized humidity on a certain day.

$H_0 =$ *The wind speed model is not a good predictor for total bike rental rate*

$H_1 =$ *The wind speed model is a good predictor for the total bike rental rate*



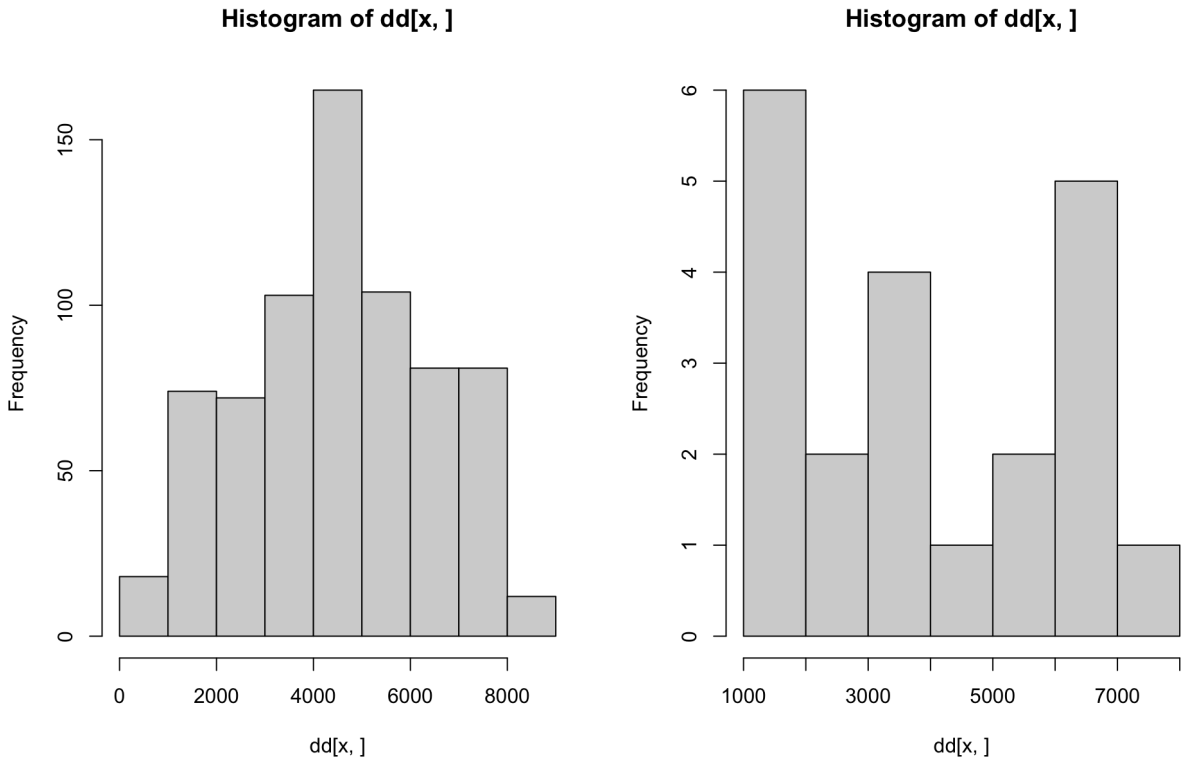
After running a simple regression model using the 'humidity' and 'total bike rental rate' variables against each other, we obtained a significant p value under 0.001. Considering this, we must reject the null hypothesis of no difference and accept the alternative hypothesis, indicating that the humidity variable is a good predictor of rental bike usage.

Holiday

The holiday variable informs us if a specific day is considered to be a holiday.

$H_0 =$ *The wind speed model is not a good predictor for total bike rental rate*

H_1 = *The wind speed model is a good predictor for the total bike rental rate*



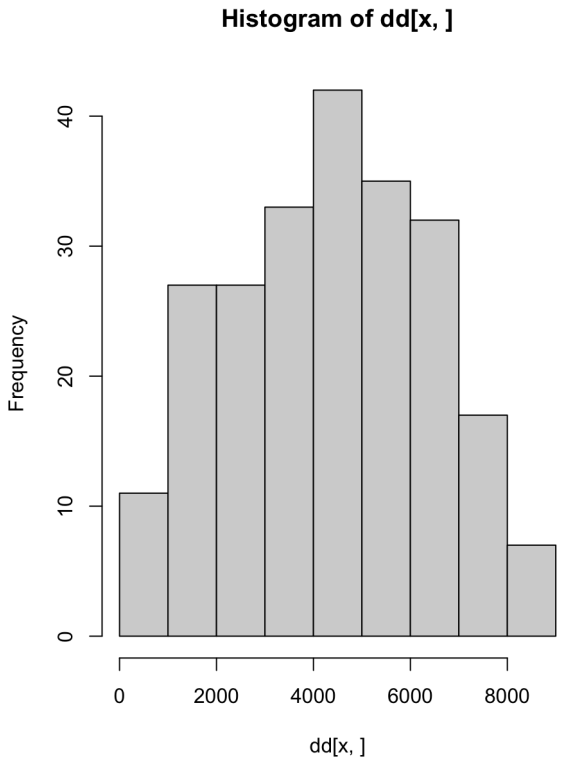
After running a simple regression model using the ‘holiday’ and ‘total bike rental rate’ variables against each other, we obtained a non-significant p value of 0.0648. Considering this, we must accept the null hypothesis of no difference and reject the alternative hypothesis, indicating that the holiday variable is not a good predictor of rental bike usage.

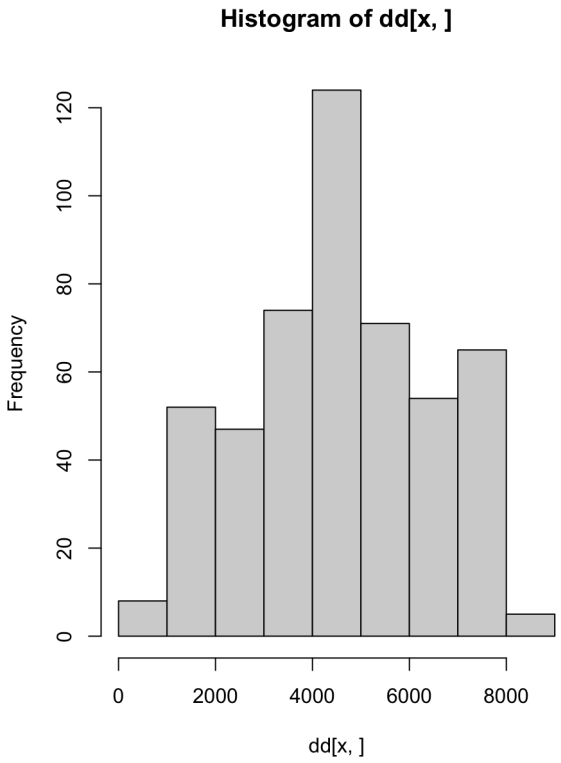
Workday

The workday variable tells us if a specific day is considered a workday or not. A one indicates that it is a work day while a zero indicates a weekend or holiday.

H_0 = *The wind speed model is not a good predictor for total bike rental rate*

H_1 = *The wind speed model is a good predictor for the total bike rental rate*





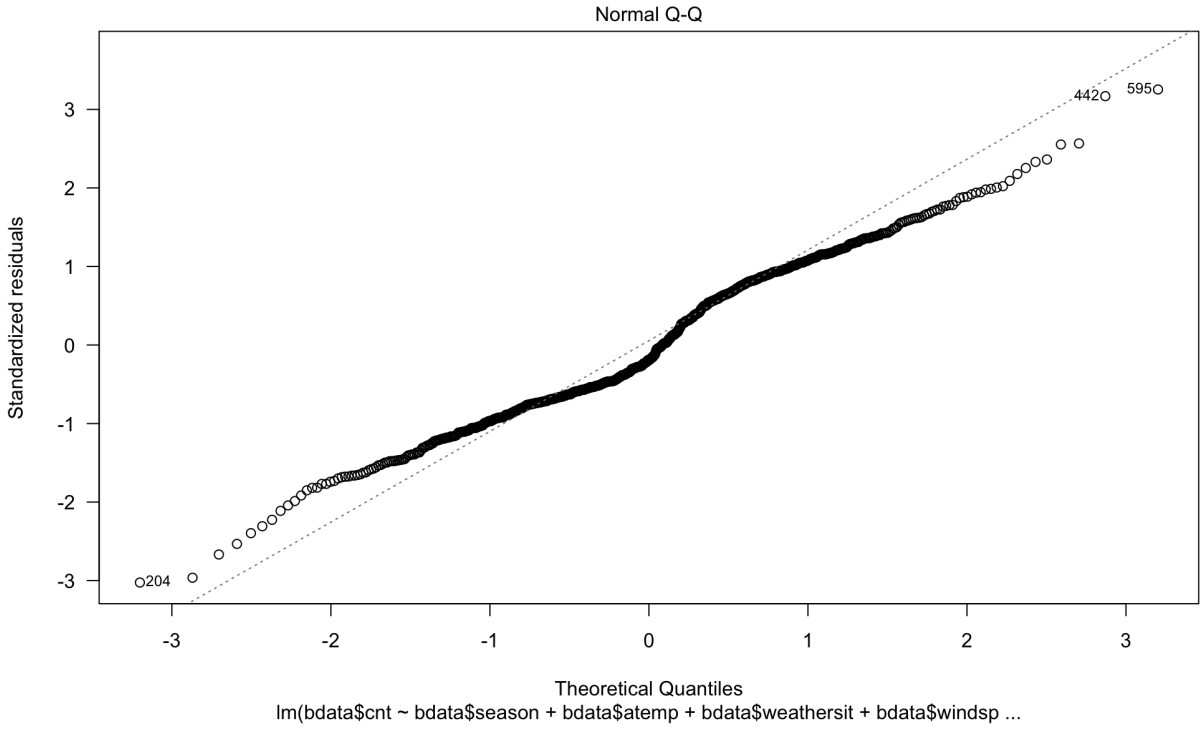
After running a simple regression model using the ‘workday’ and ‘total bike rental rate’ variables against each other, we obtained a non-significant p value of 0.0985. Considering this, we must accept the null hypothesis of no difference and reject the alternative hypothesis, indicating that the workday variable is not a good predictor of rental bike usage.

Part 2: Multiple regression

After conducting multiple simple regression plots in order to determine which ones had the most correlation with the rental bike usage variable, we obtain significant p values for the season, atemp, weather, wind speed, and

humidity variables; therefore, these variables will be used in the multiple regression. We did not obtain significant p values for the holiday and workday variables, therefore we did not include them in the multiple regression. The multiple r-squared values for this plot is 0.5523, meaning that 55% of the variation in the plot can be explained by the model, indicating that the model does significantly explain bike rental usage.

Below are the results of our multiple regression analysis, and the graph of our residuals on a qqnorm plot. The qqnorm plot shows how the residuals pan out against our predictive multiple regression model.



Output for Multiple Regression:

```
Call:
lm(formula = bdata$cnt ~ bdata$season + bdata$atemp + bdata$weathersit +
    bdata$windspeed + bdata$hum)

Residuals:
    Min      1Q  Median      3Q     Max
-3919.2  -937.3  -247.1  1082.8  4167.2

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4451.8      382.7  11.633  < 2e-16 ***
bdata$seasonspring             -522.9      151.2  -3.459 0.000573 ***
bdata$seasonsummer             -864.0      186.4  -4.636 4.21e-06 ***
bdata$seasonwinter            -1501.3      153.5  -9.782  < 2e-16 ***
bdata$atemp                    6645.8      527.6  12.597  < 2e-16 ***
bdata$weathersitmist           -215.1      128.2  -1.678 0.093868 .
bdata$weathersitrain or snow  -1872.0      328.5  -5.698 1.77e-08 ***
bdata$windspeed               -2996.6      678.7  -4.415 1.16e-05 ***
bdata$hum                     -2670.9      464.1  -5.754 1.29e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1303 on 722 degrees of freedom
Multiple R-squared:  0.5523,    Adjusted R-squared:  0.5473
F-statistic: 111.3 on 8 and 722 DF,  p-value: < 2.2e-16
```

When we break the model down by each variable, we can see that the misty weather variable has now obtained a p value of 0.093868, rendering it insignificant ($p > 0.05$). Every other variable we tested for has returned significant in explaining rental bike usage, but to varying degrees. The season data, specifically the spring variable, was still significant in explaining bike usage with a p-value of 0.000573 ($p < 0.05$), while others had more significant p-values. Two of the variables (the season of winter

and atemp) obtained the smallest p-values, both of 2e-16, indicating a very significant correlation to rental bike usage ($p < 0.05$).
We can conclude that all of the variables included in the multiple regression model, with the exception of misty weather, were significant in predicting bike usage.

Collaboration

- The coding work was split fairly evenly between us both, although Ramchandra did slightly more in that realm. The write up was also split fairly evenly between us, although I (Kyle) did slightly more in this realm.