Statistical Analysis on Bike Sharing

Michael Pilson mpilson

Due Wed, October 26, at 11:59PM

Contents


```
library("knitr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("interactions")
library("leaps")
```

Introduction

In recent years, bike sharing has exploded in popularity. Whether you are just a tourist looking to glide around town or a daily commuter who is late to work, bike sharing creates a seamless and speedy way to get from place to another. As someone who prides himself of exercising consistently, bike sharing allows some one to fit exercise into their scheduled without having to make time for it or cram it into their routine throwing off their day. In addition, bike sharing is also much better for the environment than driving promoting a Eco-friendly and athletic lifestyle. To better understand and predict the pattern of use of these bikes, we will look at possible contributing factors of casual use including weather, temperature, and wind speed.

Source: Fanaee-T, Hadi and Gamma, J. "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013); http://capitalbikeshare.com/system-data

Exploratory Data Analysis

To take a closer look at bike sharing, we will be analyzing a sample of data on hourly casual bike share users the from the Washington D.C, Arlington and VA/MD area. The sample contains 656 users and 4 variables: Temperature, Weather, and Windspeed. We will examine the relationship between the hourly casual use of bike sharing and the three explanatory variables to hopefully find some sort of relationship between these

factors and the number of casual hourly bikers. The number of Casual hourly bikers, Weather, Temperature, and Windspeed are characterized as such:

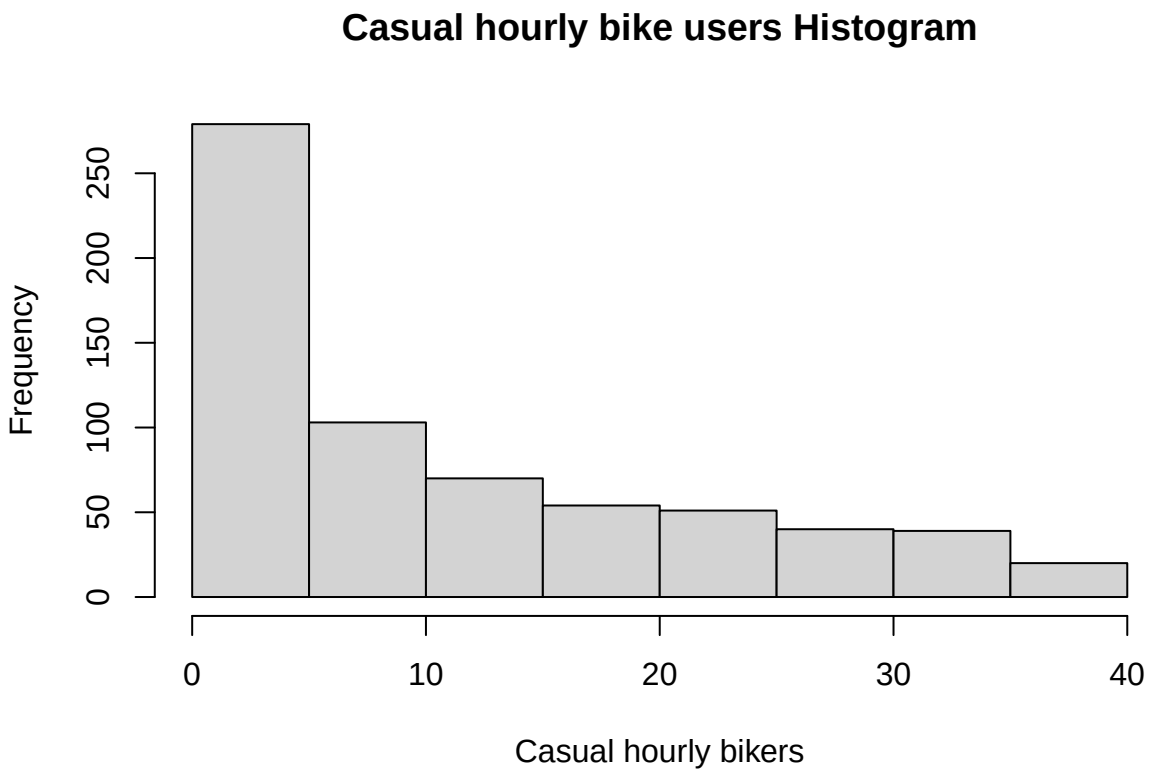
Casual: number of casual hourly bike users (the response variable) Weather: type of weather (in three categories: clear, misty, rain/snow) Temp: temperature (scaled as percentage of overall maximum temperature) Windspeed: windspeed (scaled as percentage of overall maximum windpseed)

The data set columns are organized as Casual, Weather, Temp, then Winspeed across with the row number as the left most column, and each row being an individual data entry (A studied hour)

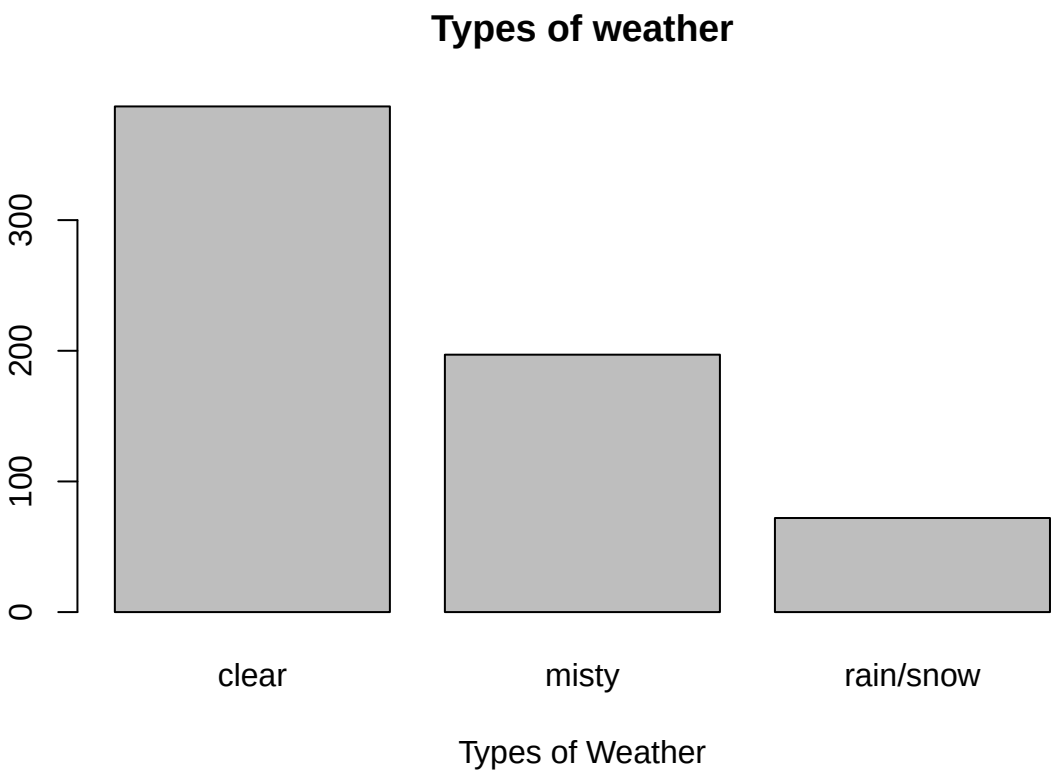
Data

When analyzing a data set, it is very helpful to first look at each variable by itself just to familiarize us with the raw data and the possible patterns or trends. This is known as Univariate exploration. For numerical and continuous variables histograms are incredibly useful for looking at these data sets. For categorical variables, data which falls into a certain group, I will use barplots.

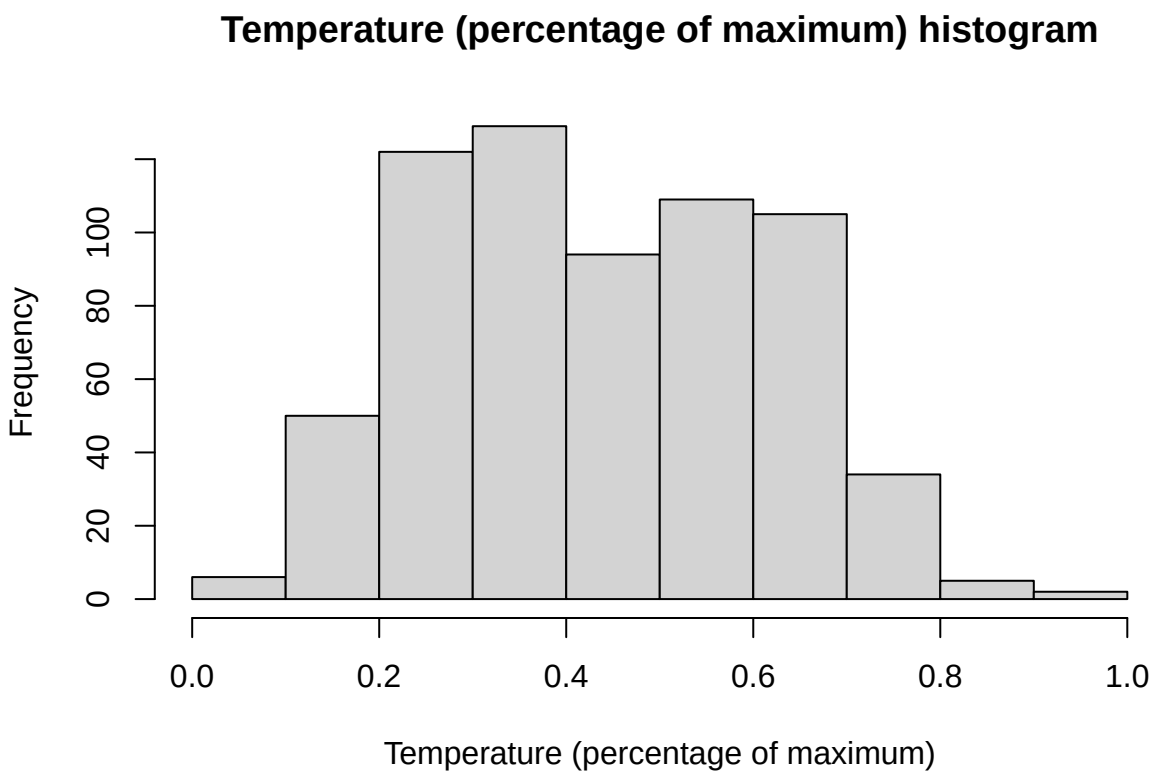
```
Casual <- bikes$Casual
Weather <- bikes$Weather
Temp <- bikes$Temp
Windspeed <- bikes$Windspeed
hist(Casual, xlab = "Casual hourly bikers", main = "Casual hourly bike users Histogram")
```



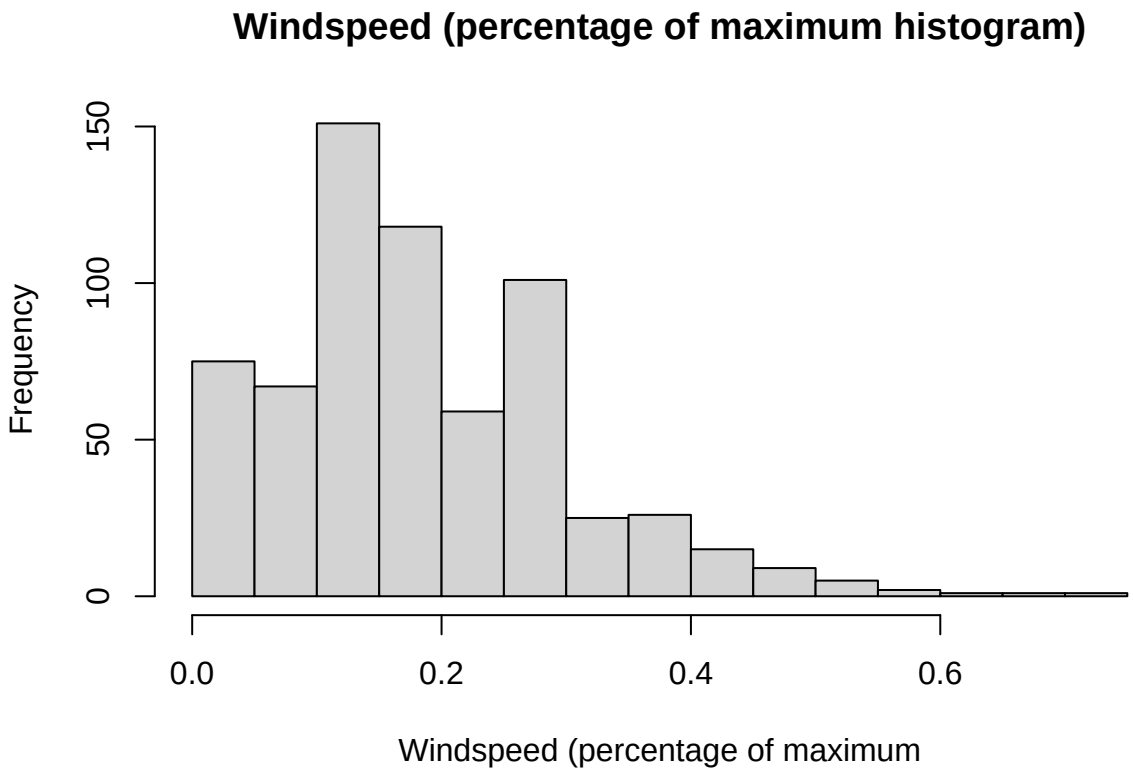

```
barplot(table(bikes$Weather), xlab = "Types of Weather", main = "Types of weather")
```




```
hist(Temp, xlab = "Temperature (percentage of maximum)", main = "Temperature (percentage of maximum) his
```




```
hist(Windspeed, xlab = "Windspeed (percentage of maximum", main = "Windspeed (percentage of maximum his
```



We supplement the univariate graphical summary with numerical summaries, as follows: For casual bike users:

```
summary(Casual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.00    8.00   11.51   20.00   39.00
```

```
sd(Casual)
```

```
## [1] 11.17931
```

For the weather:

```
table(Weather)
```

```
## Weather
##     clear     misty rain/snow
##       387       197        72
```

For the temperature:

```
summary(Temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0200  0.3000  0.4400  0.4429  0.5850  0.9400
```

```
sd(Temp)
```

```
## [1] 0.1755449
```

For windspeed:

```
summary(Windspeed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1045  0.1642  0.1840  0.2537  0.7164
```

```
sd(Windspeed)
```

```
## [1] 0.1206051
```

From the distributions above we can conclude the following:

The majority of the data (Q1-Q3) being between 2 and 20 Casual bikers with a standard deviation of 11.12 casual hourly bikers. The distribution of Casual hourly bikers is unimodal and strongly right skewed with the median being 8 users and the mean being 11.5.

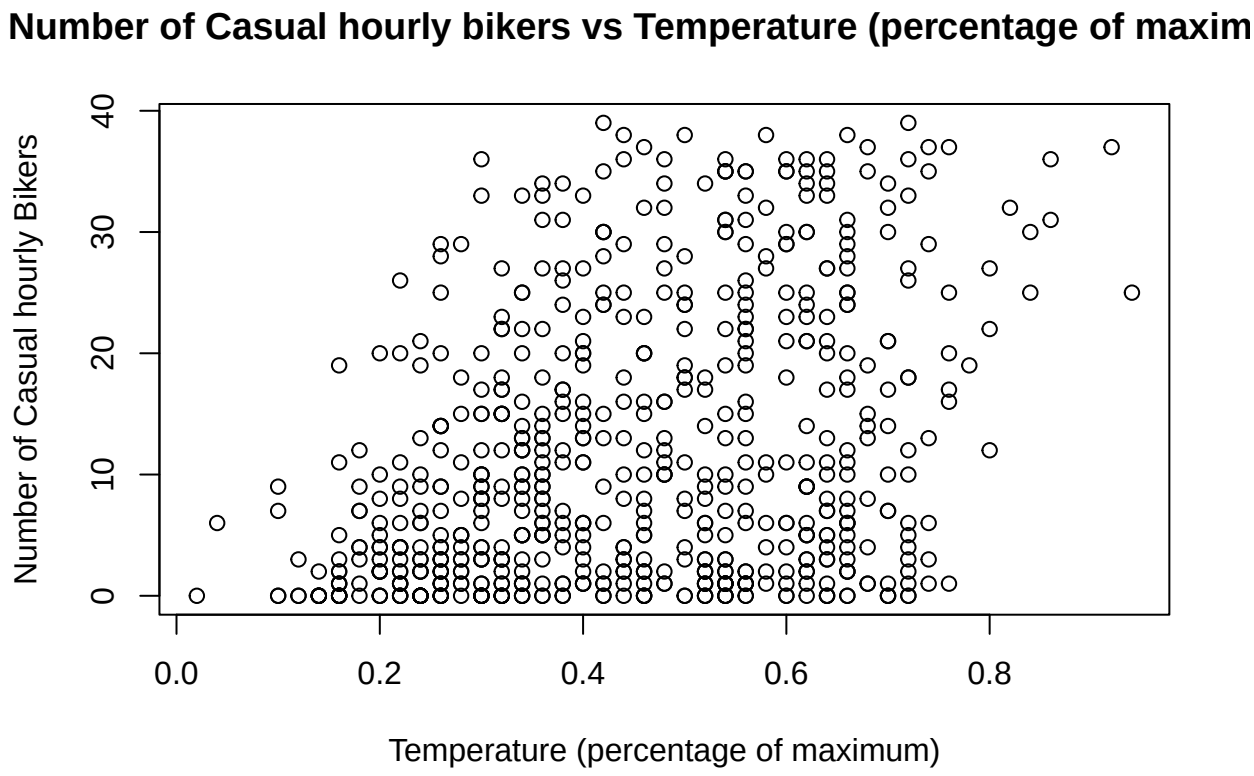
There are 387 clear, 197 misty, and 72 rainy/snowy hours.

The majority of the data (Q1-Q3) being between .3 and .6 of maximum temperature with a standard deviation of .1755 of maximum temperature. The distribution of Temperature is approximately bimodal and symmetrical with the median and mean being .44 of maximum temperature.

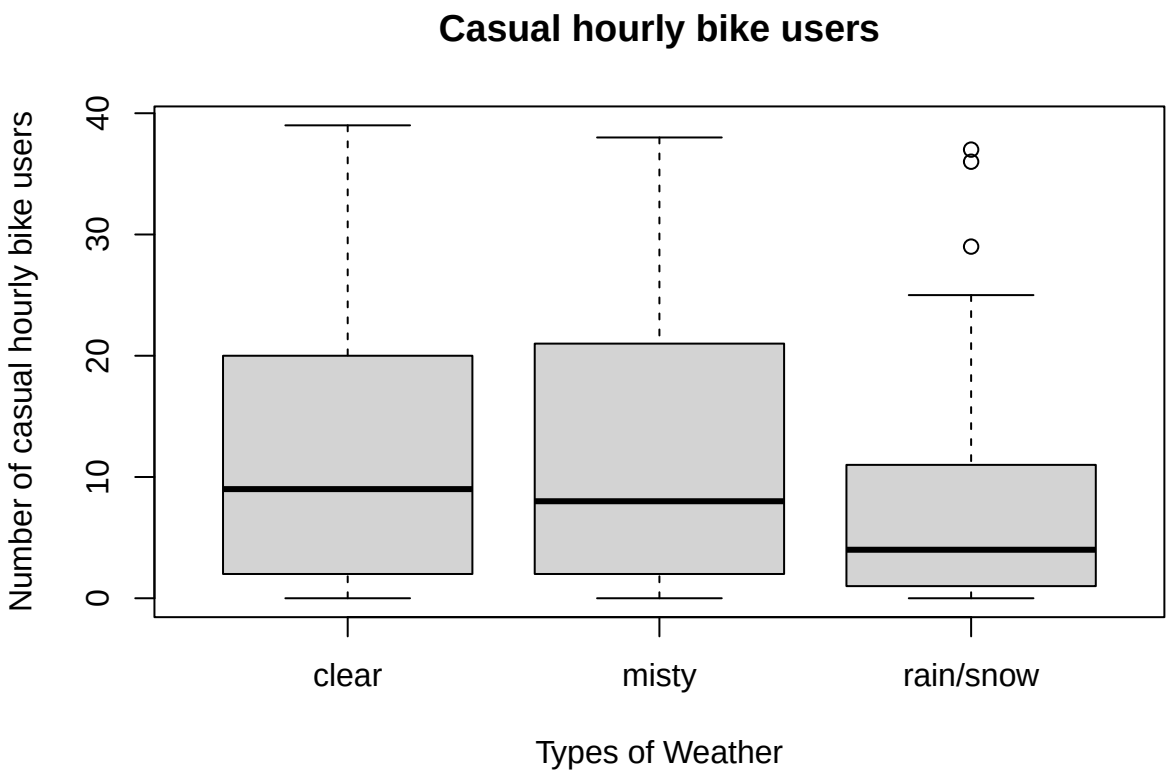
The majority of the data (Q1-Q3) being between .1 and .25 of maximum windspeed with a standard deviation of .12 of maximum windspeed. The distribution of Windspeed is unimodal and skewed rigth with the median of .164 and mean being .184 of maximum windspeed.

Now that we have looked at the data by themselves, we can now start looking at how the variables impact or are related to one another. This is known as Bivariate exploration. Since we care about the number of Casual hourly bikers, we use the other variables to predict that.

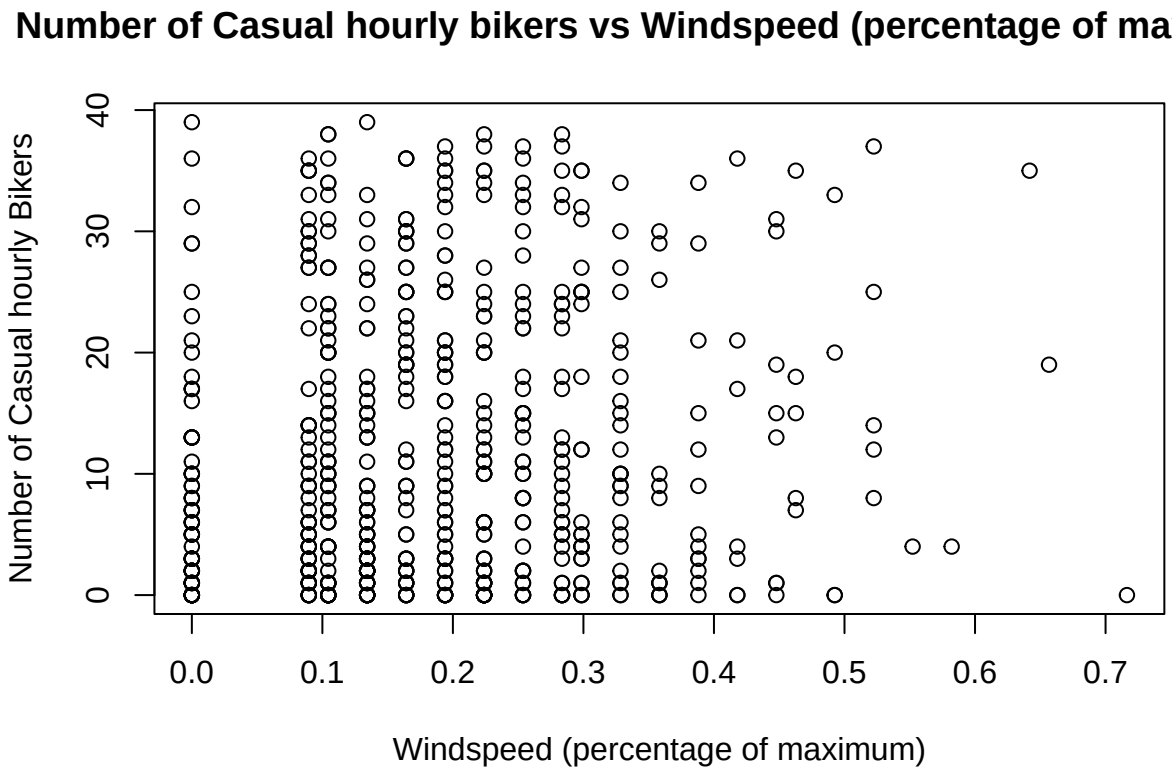
```
plot(Casual ~ Temp, xlab = "Temperature (percentage of maximum)", ylab = "Number of Casual hourly Bikers
```




```
boxplot(Casual ~ Weather, data = bikes, xlab = "Types of Weather",
   ylab = "Number of casual hourly bike users", main = "Casual hourly bike users")
```




```
plot(Casual ~ Windspeed, xlab = "Windspeed (percentage of maximum)", ylab = "Number of Casual hourly Bi
```



The number of casual hourly bikers seems to be generally increasing with increased temperature; however, there seems to be a large concentration of points near 0 as the data is bounded due to it being a percentage. The scatter plot of number of casual hourly bikers vs temperature (percentage of maximum) shows a weak positive linear relationship.

Clear and misty weather have an almost identical spread and relationship to number of casual hourly bikers. Both having a center around ten and the majority (Q1-Q3) of the data being between 2 and 20. Rainy/Snowy, however, produces on average a lower number of casual hourly bikers and has a much tighter spread with a center of 4. The majority of the data is between 1 and 10 (Q1-Q3), but it has a few outliers around 30 casual hourly bikers. The box plot of weather shows all three weather types: clear, misty,and rainy/snowy are all skewed right (Due to Casual Hourly Bikers being skewed).

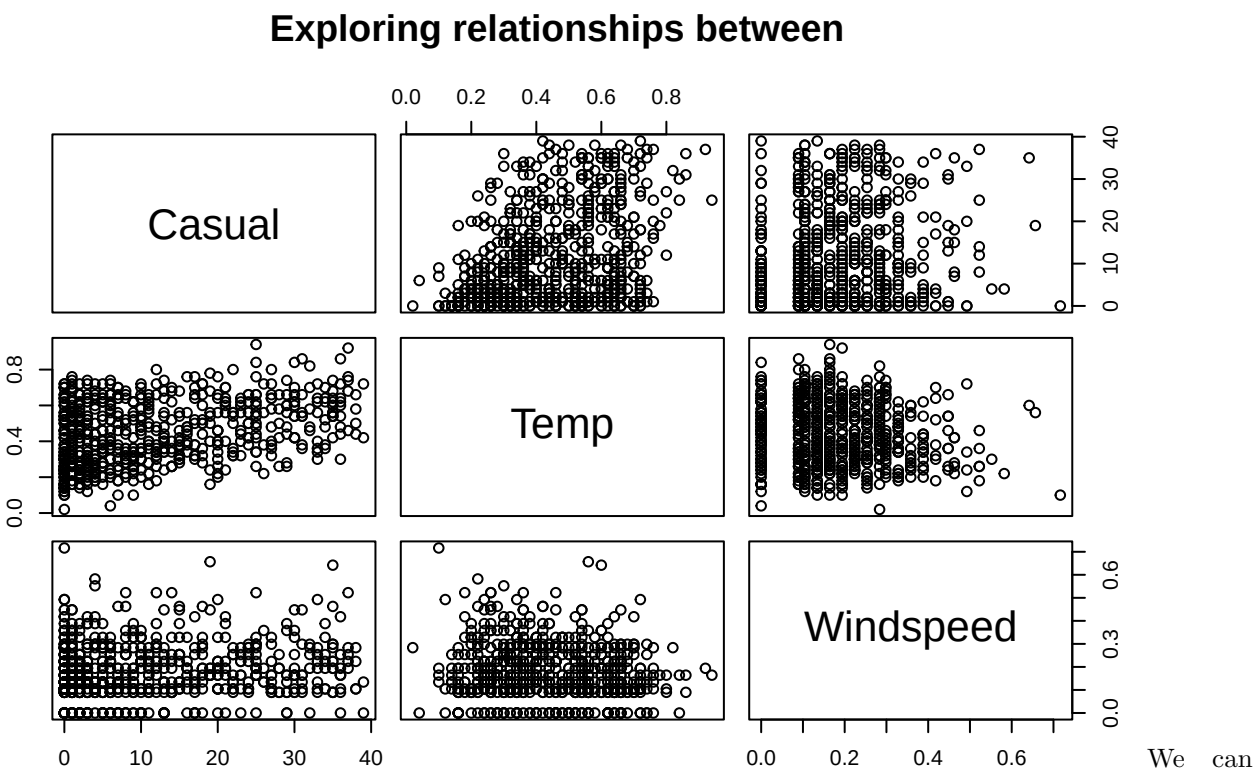
The scatter plot of number of casual hourly bikers vs windspeed (percentage of maximum) shows a nonlinear relationship as the points appear to be along multiple vertical lines throughout the plot. This is concerning and will be addressed later.

Modeling

Now that we have looked at how each variable is related to the response variable, the number of Casual hourly bikers, we can now start to build our model. We start by looking at the histogram of our response variable. It looks skewed right, indicating that a transformation might be needed. For now, we will leave the number of Casual hourly bikers unchanged but we will address that later.

In our bivariate analysis, we saw that two numerical variables have a relationship with number of casual hourly bikers (temperature and weather). Before we consider building the model though we need to consider the possibility of multicollinearity, a relationship between the explanatory variables whcih could impact the relationship with the response variable. A good and convenient model to use to explore these relationships is the pairs plot. We will include windspeed and temperature and exclude weather as it is categorical.

```
bikes.no.weather <- subset(bikes, select = c(Casual, Temp, Windspeed))
pairs(bikes.no.weather, main = "Exploring relationships between ")
```



We can conclude that Temperature and Windspeed are not correlated as seen in the pairs plot above, but just to make sure we will check with the vif's.

```
Casual.all <- lm(Casual ~ Temp + Windspeed + Weather, data = bikes)
car::vif(Casual.all)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## Temp       1.020677  1        1.010285
## Windspeed 1.019349  1        1.009628
## Weather    1.006894  2        1.001719
```

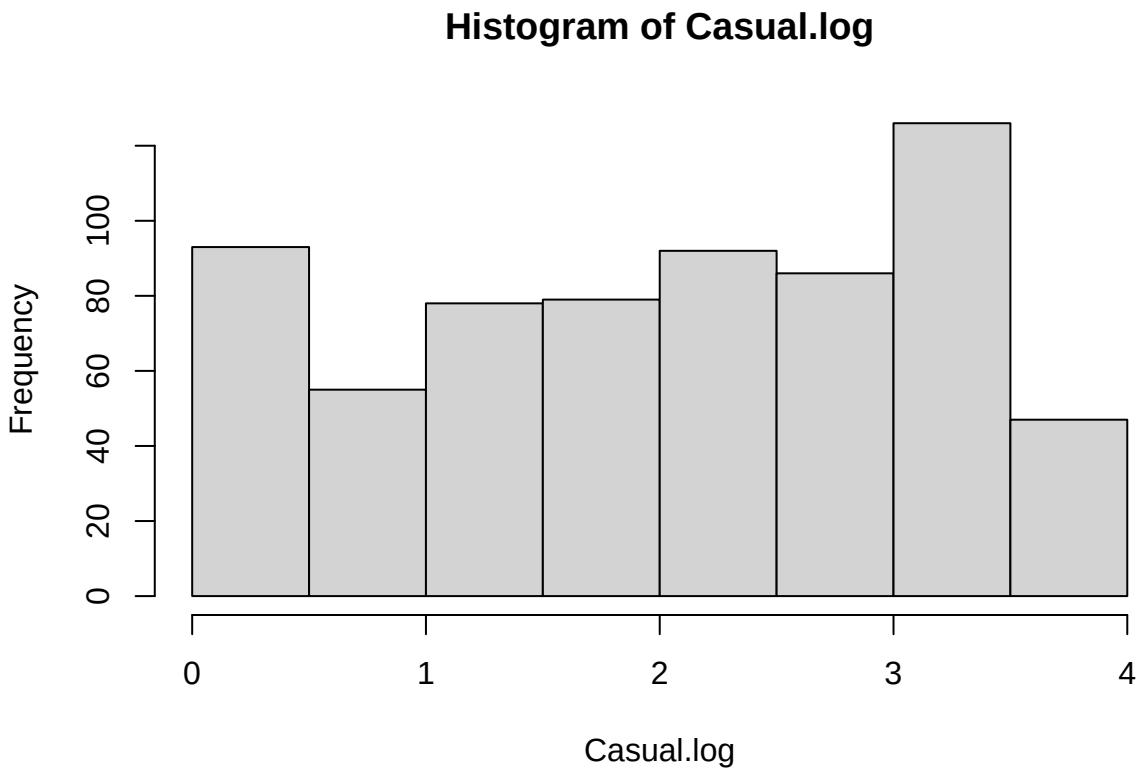
The gvifs (which are treated the same as vifs) all fall below 2.5 so no danger of muticollinearity. We can construct the model as we checked and there appears to be no multi collinearity.

Now to address the right skewness of Casual hourly Bikers. we will transform the distribution using two methods: logging it or putting to a higher power. To log it properly, We will first have to adjust for any 0 or negative values. Since the minimum is zero, We can add 1.1 and then log. To put it to a higher power, We will use multiple higher powers to see the best fit.

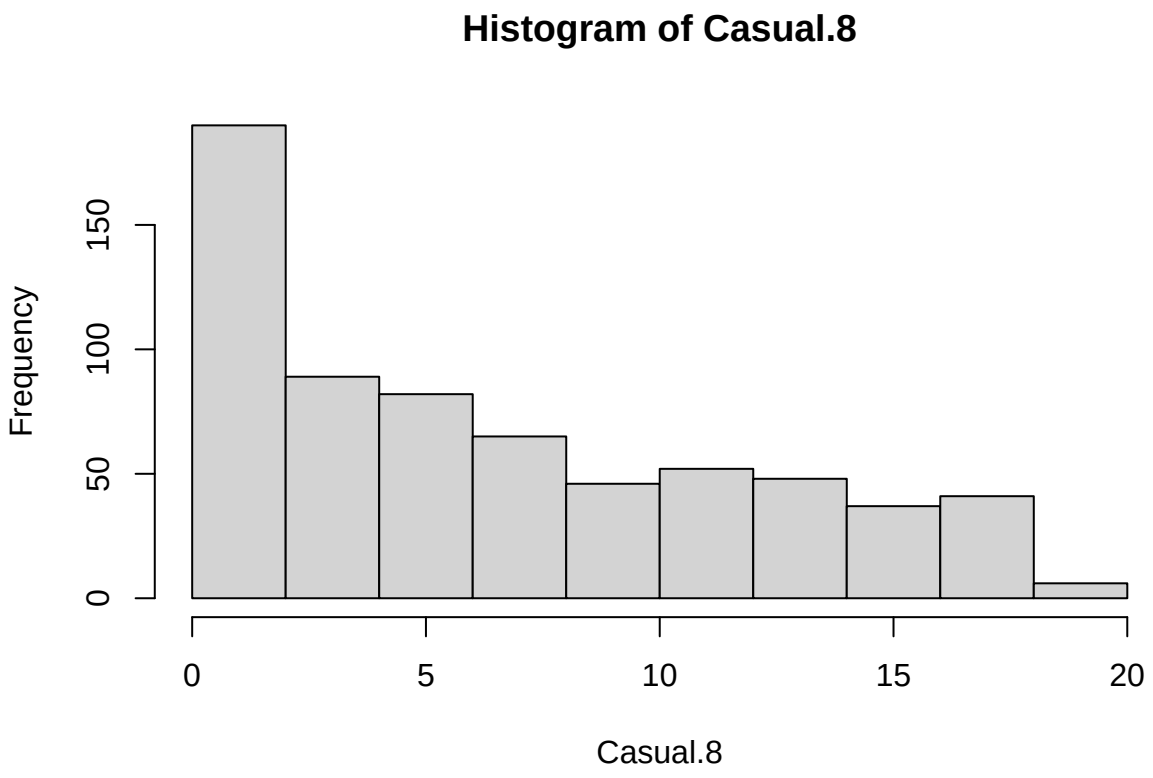
```
Casual.8 <- (Casual)^.8
Casual.6 <- (Casual)^.6
Casual.4 <- (Casual)^.4
Casual.2 <- (Casual)^.2
min(bikes$Casual)
```

```
## [1] 0
```

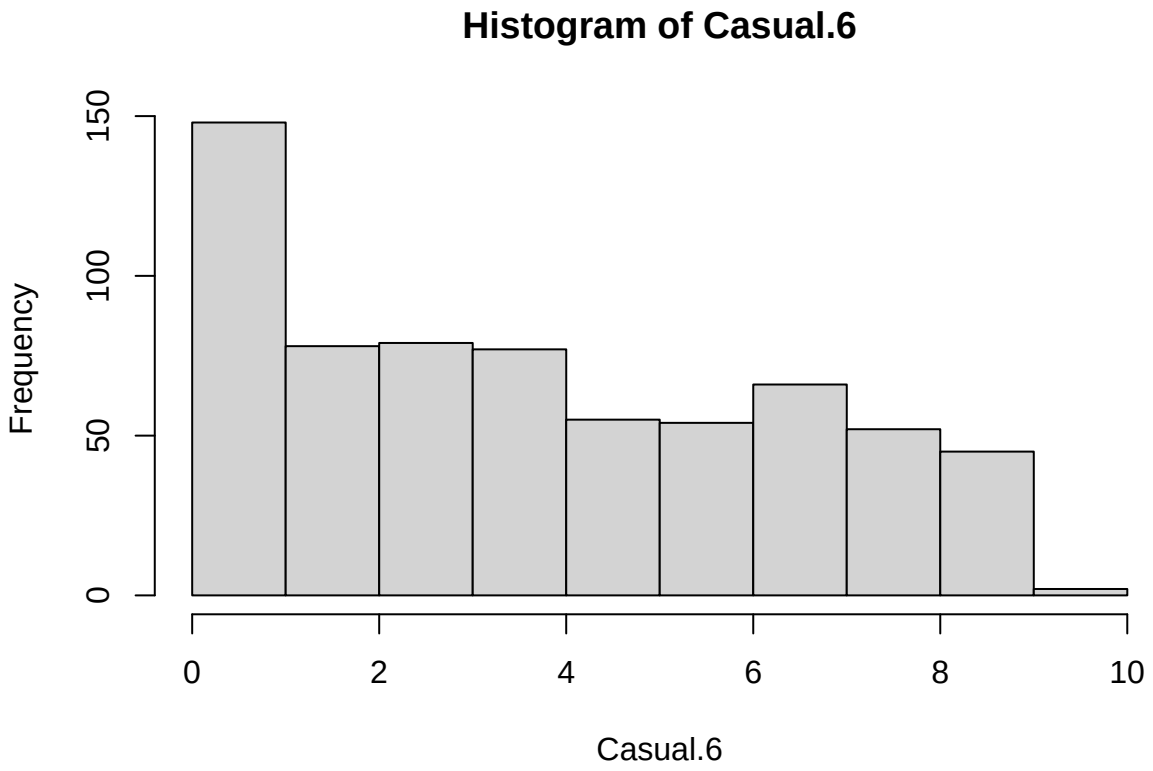
```
Casual.shifted <- bikes$Casual + 1.1
Casual.log <- log(Casual.shifted)
hist(Casual.log)
```



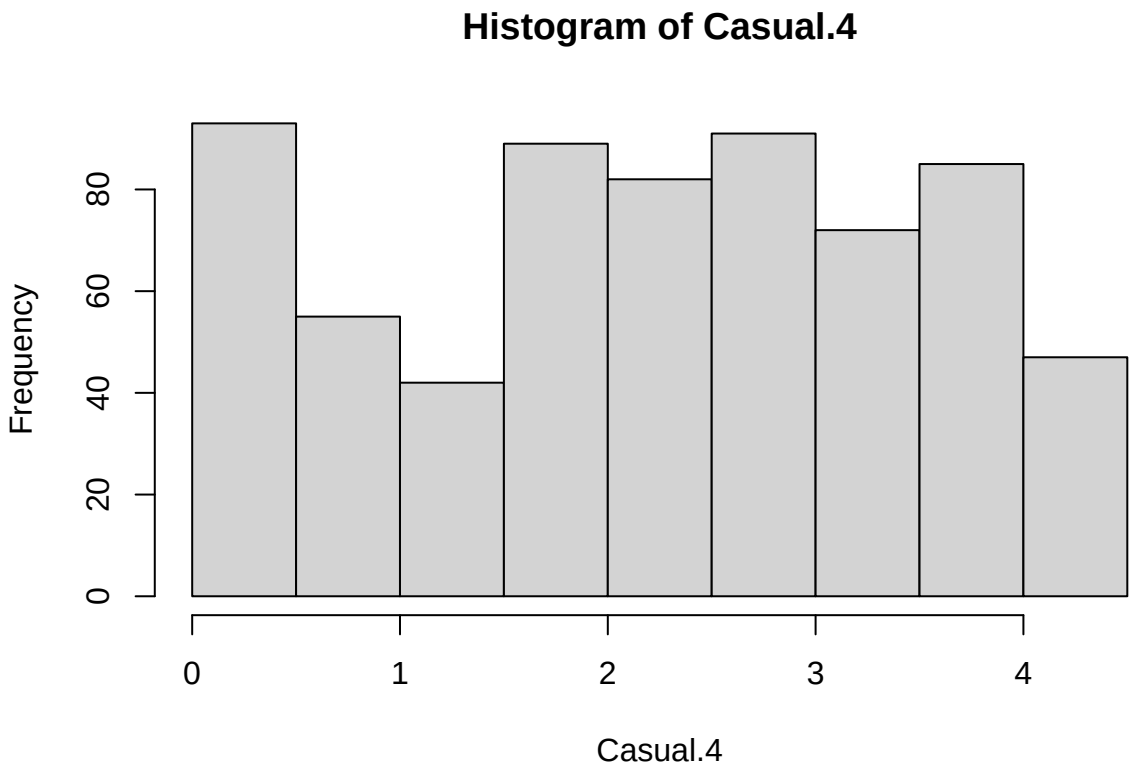

```
hist(Casual.8)
```



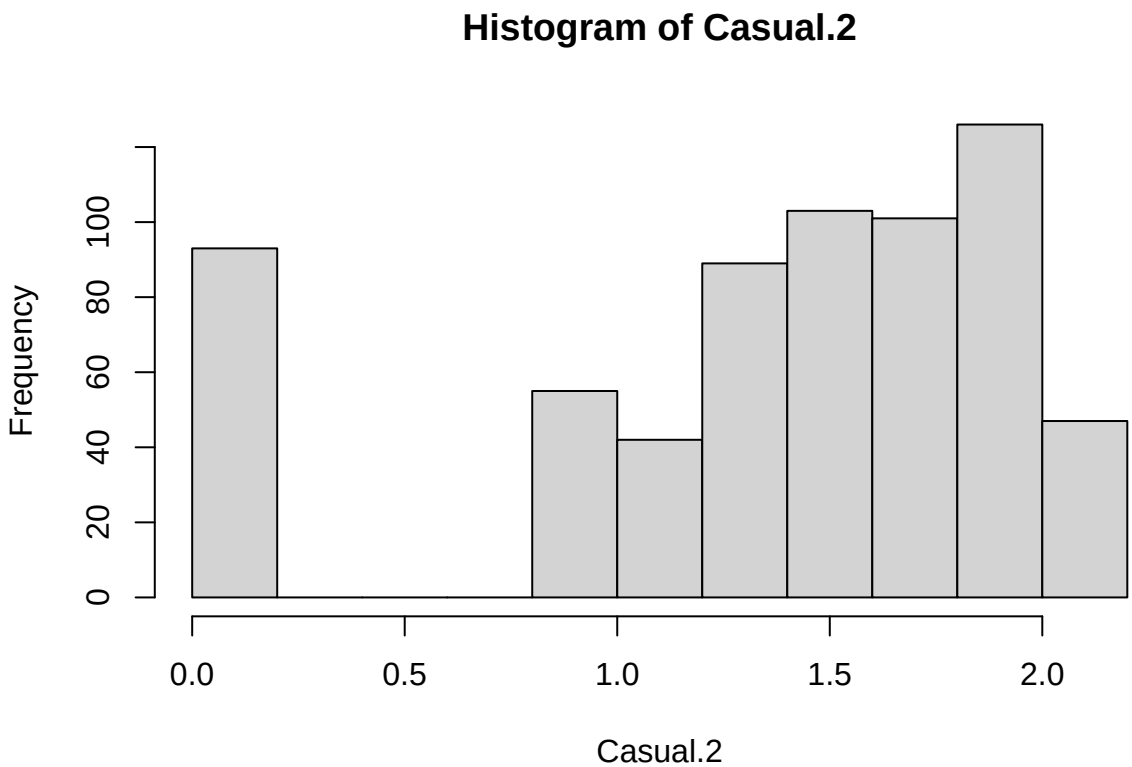

```
hist(Casual.6)
```




```
hist(Casual.4)
```




```
hist(Casual.2)
```



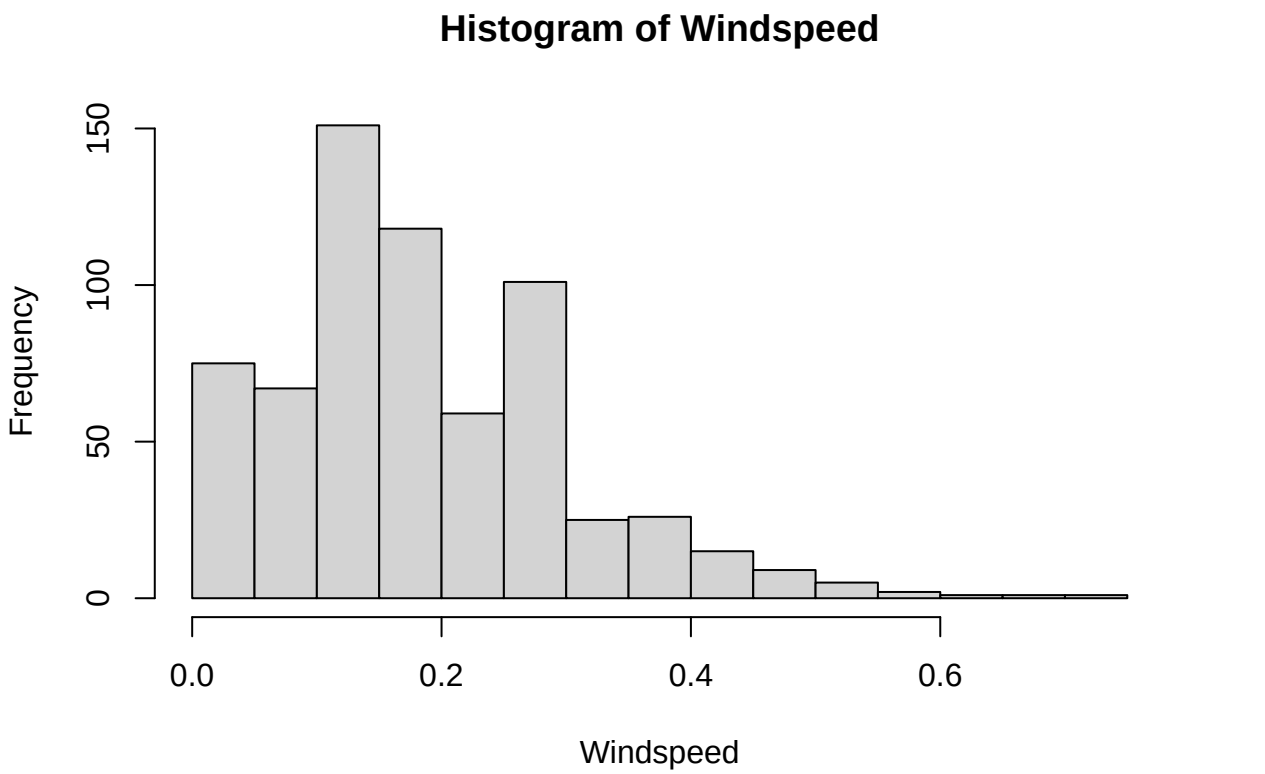
Now with the Casual hourly bikers distribution appearing a little more normal, we can now start to set up our linear regression model. Since the log of the distribution and the distribution^.4 seem to be similar fits, We will stick with the log transformation.

Cautiously, We will include both the transformmed (log) and the unadjusted distribution to compare and

make sure the best or most accurate model is not missed.

Unfortunately, as seen in the scatter plot of Casual hourly bikers and windspeed there seems to be no relationship so We will leave windspeed out of the model. We attempted to transform it with a log but the lack of a relationship persisted as seen below.

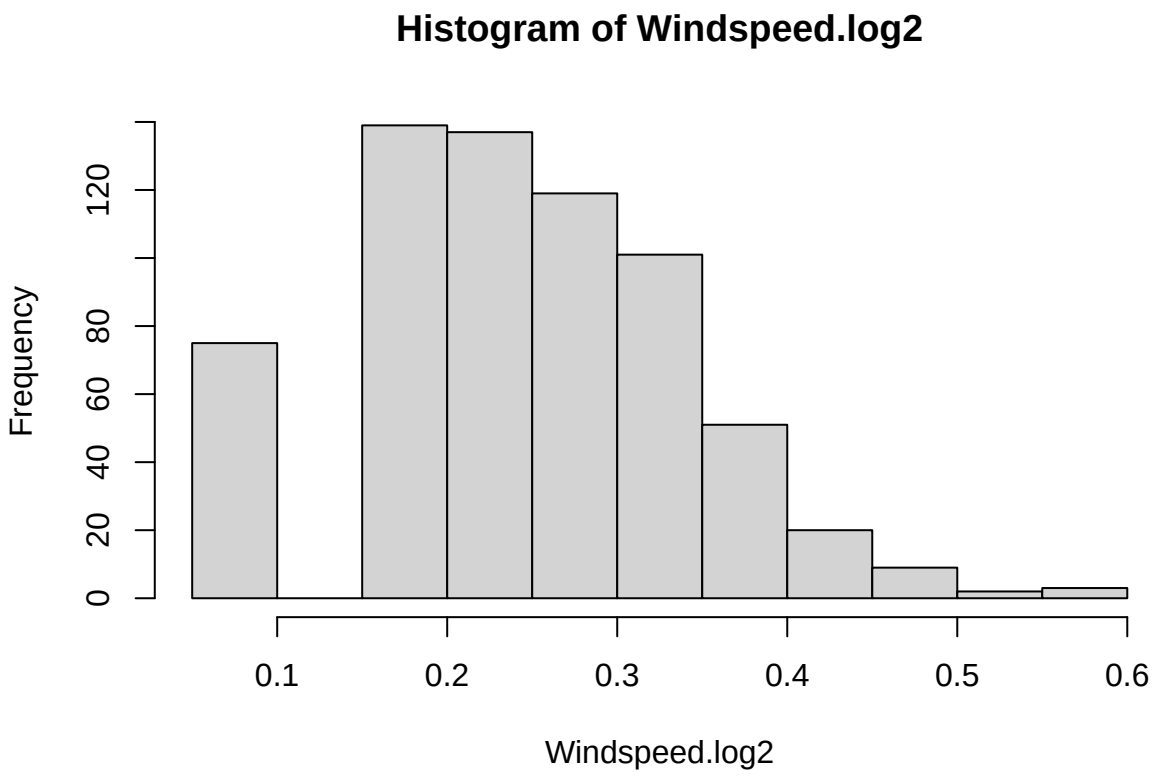
```
hist(Windspeed)
```



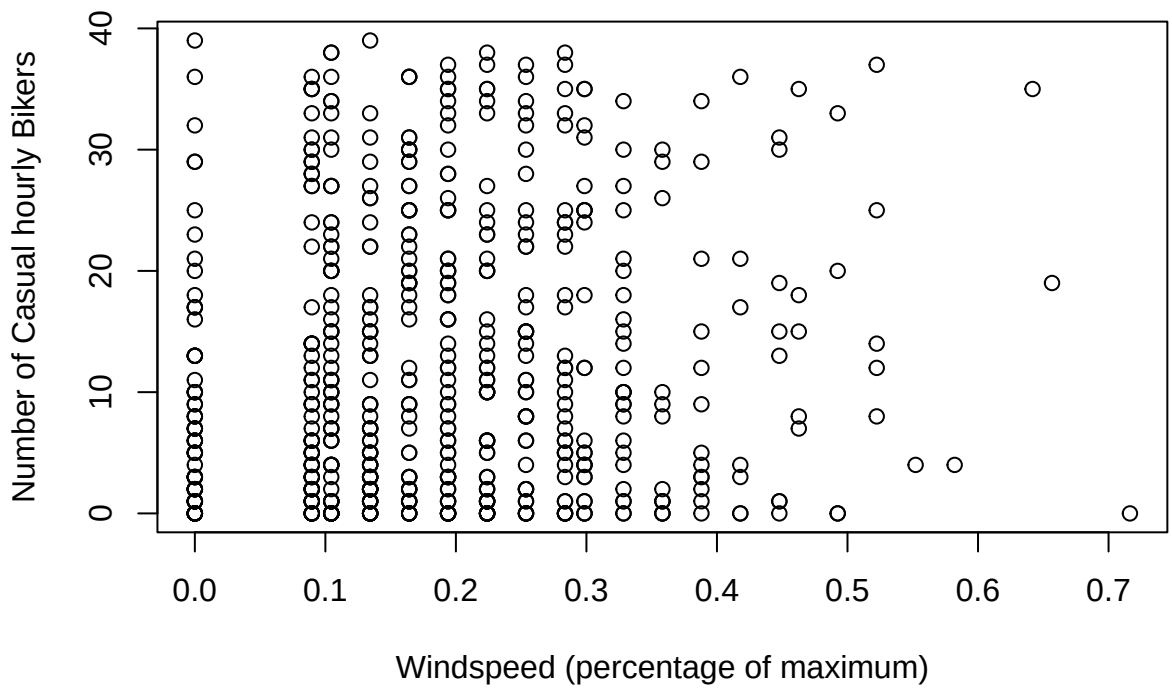

```
min(Windspeed)
```

```
## [1] 0
```

```
Windspeed.shifted <- bikes$Windspeed + 1.1
Windspeed.log2 <- log(Windspeed.shifted)
hist(Windspeed.log2)
```

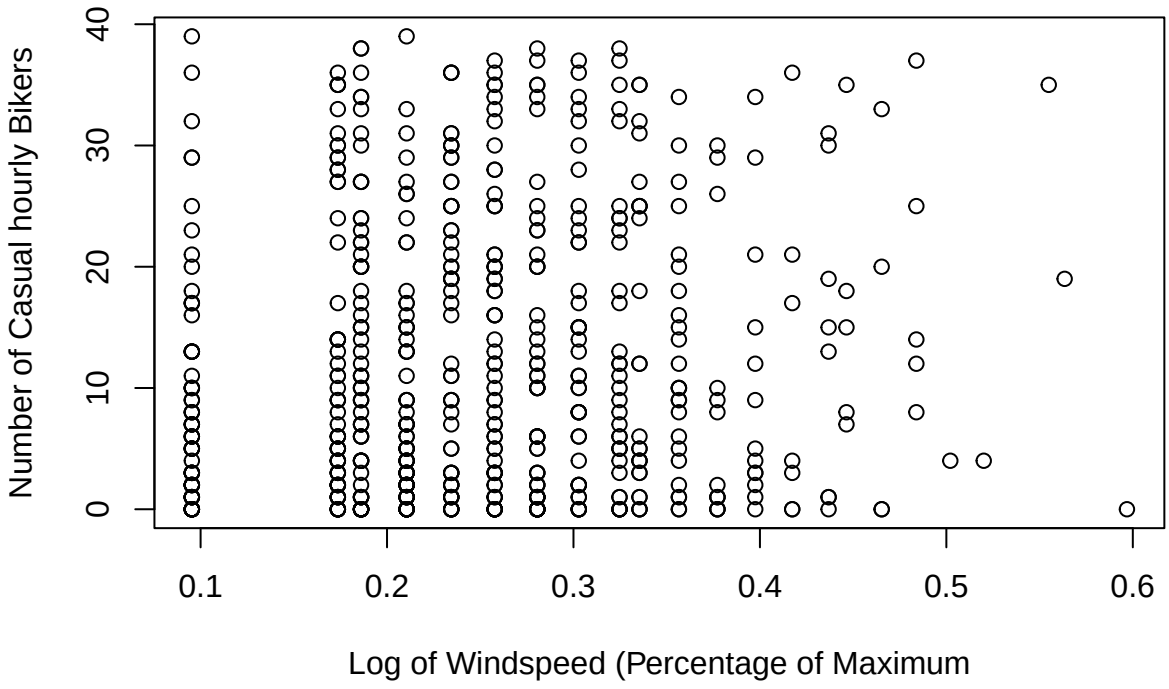



```
plot(Casual ~ Windspeed, xlab = "Windspeed (percentage of maximum)", ylab = "Number of Casual hourly Bi
```



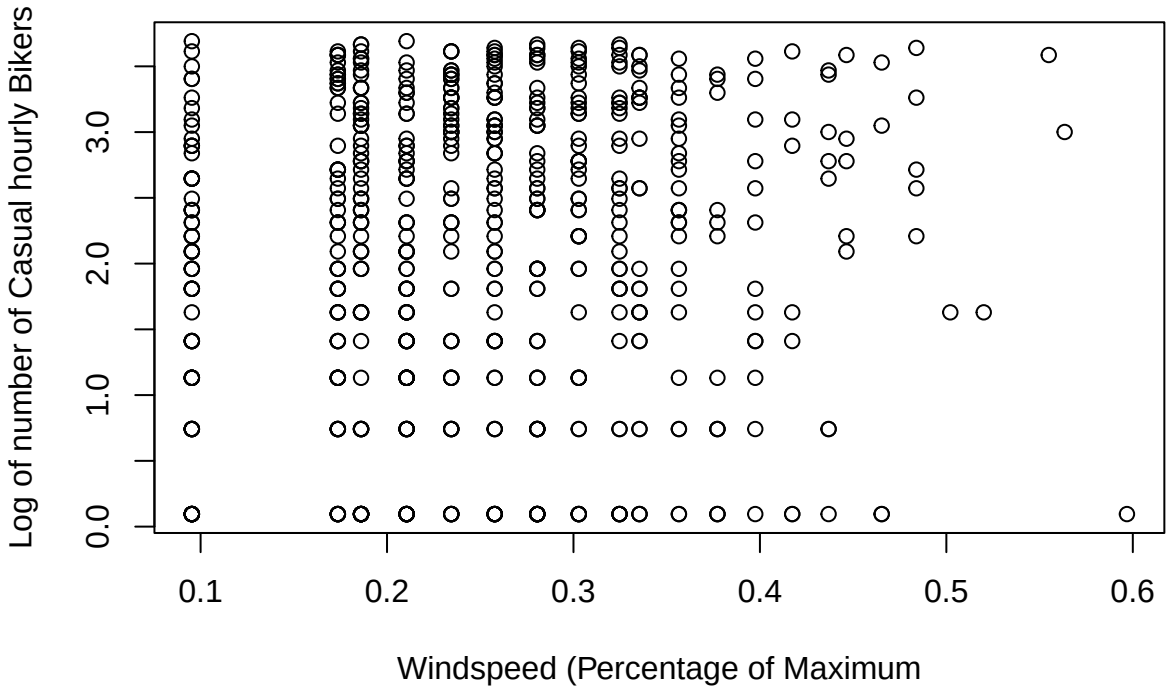

```
Windspeed.log <- log(Windspeed)
plot(Casual ~ Windspeed.log2, xlab = "Log of Windspeed (Percentage of Maximum", ylab = "Number of Casua
```

umber of Casual hourly bikers vs Log of Windspeed (Percentage of Ma

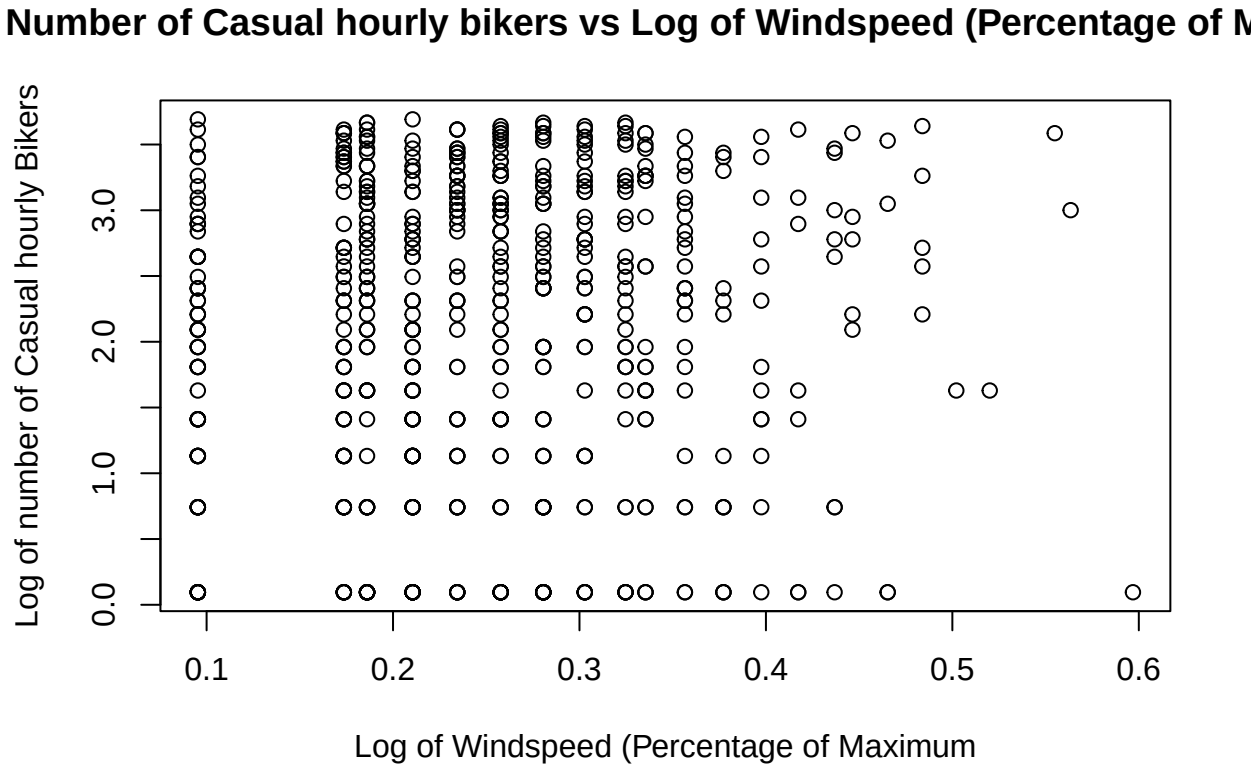



```
plot(Casual.log ~ Windspeed.log2, xlab = "Windspeed (Percentage of Maximum", ylab = "Log of number of Ca
```

.og Number of Casual hourly bikers vs Windspeed (Percentage of Maxi

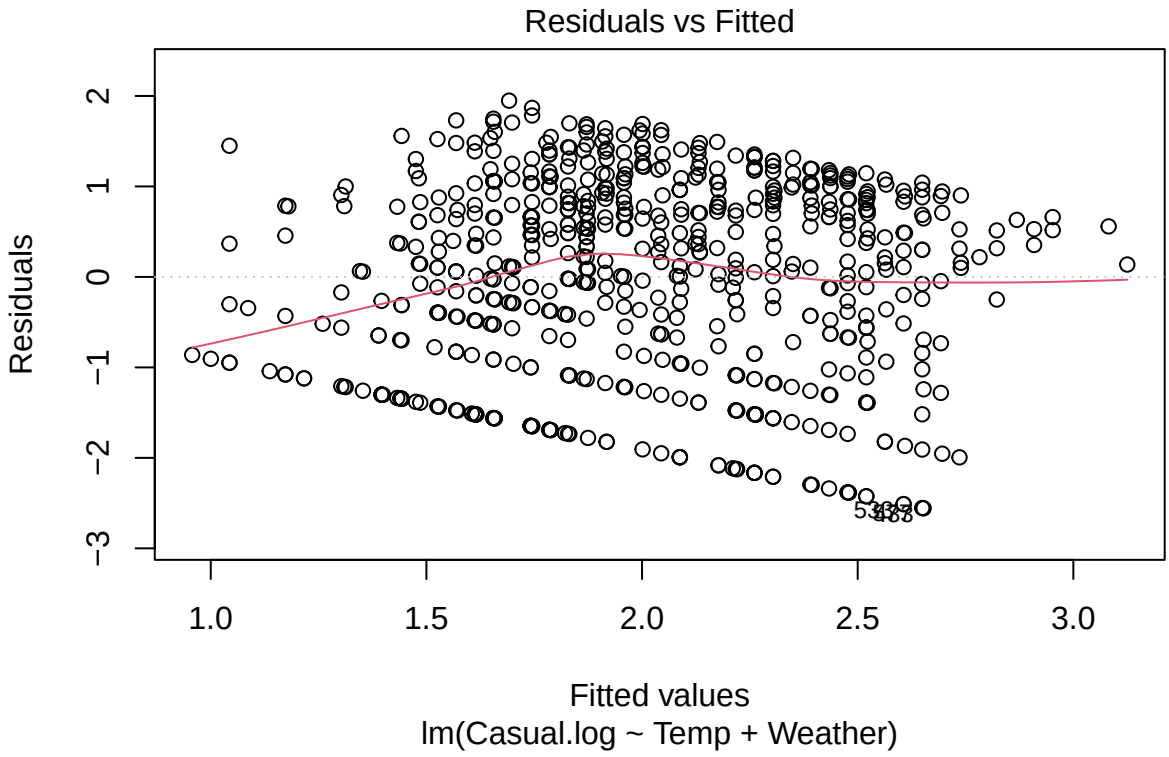



```
plot(Casual.log ~ Windspeed.log2, xlab = "Log of Windspeed (Percentage of Maximum", ylab = "Log of numbe
```

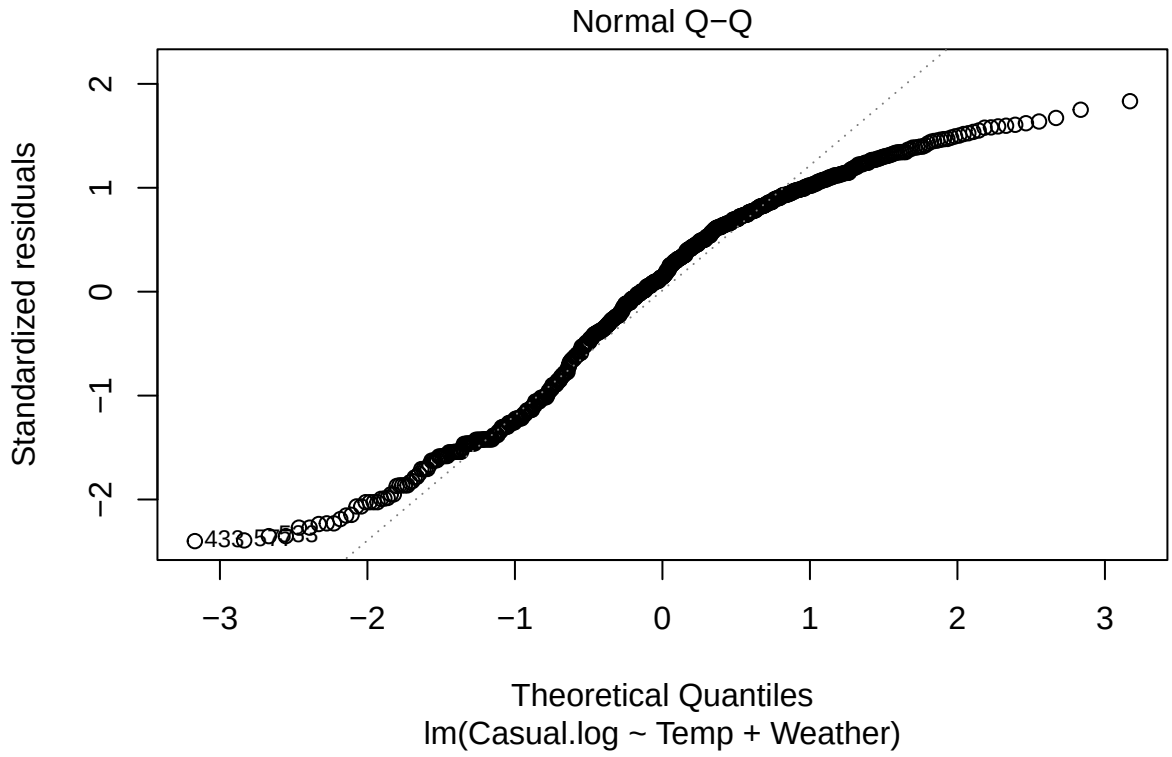


So our model will be using Weather and Temperature as the explanatory variables. Both of these variables seem to be relatively normal so no transformation will be needed. We will be using the transformed Casual hourly bikers and the unadjusted distribution to explore the best fit.

```
bikes.log.all <- lm(Casual.log ~ Temp + Weather, data = bikes)
plot(bikes.log.all, which=1)
```



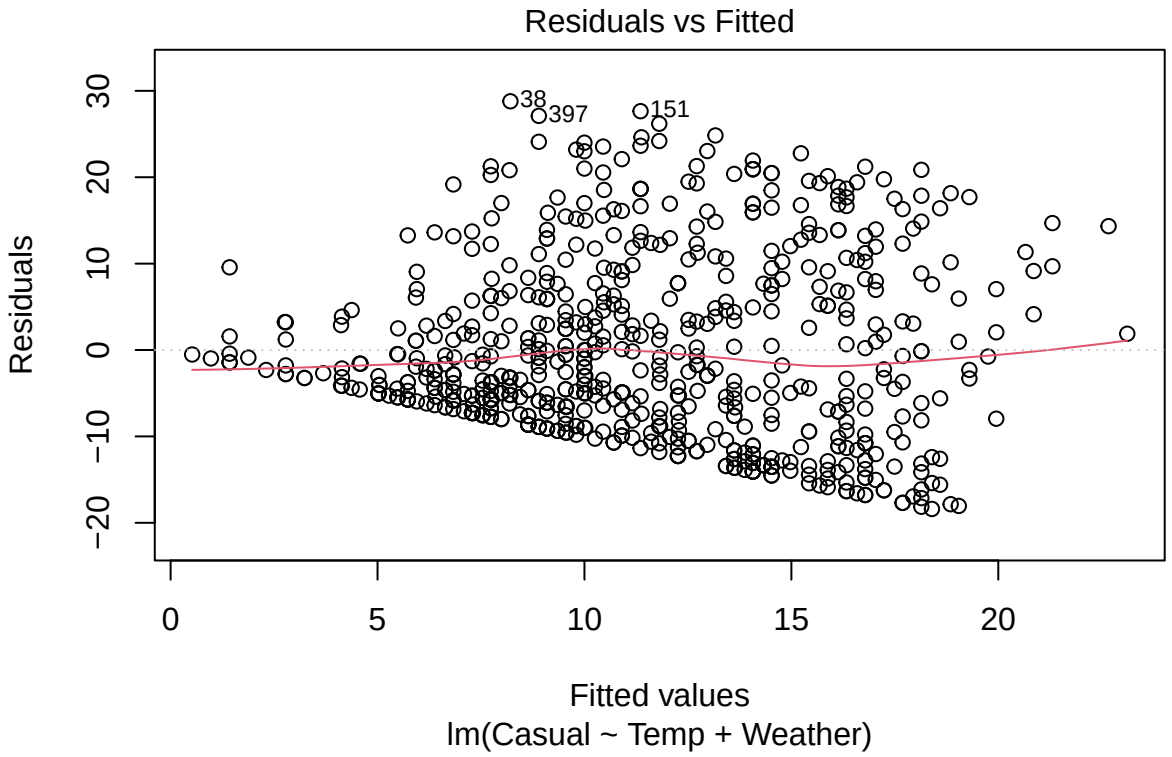
```
plot(bikes.log.all, which=2)
```



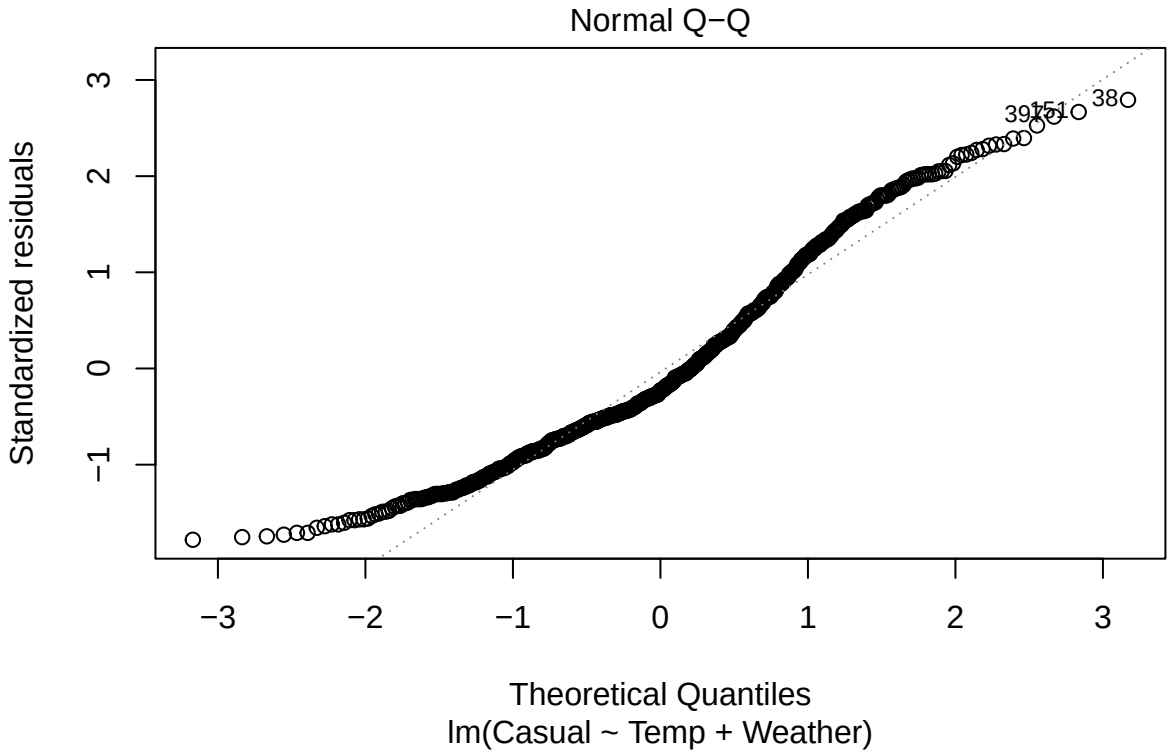

```
summary(bikes.log.all)
```

```
##
## Call:
## lm(formula = Casual.log ~ Temp + Weather, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5574 -0.8526  0.1506  0.8784  1.9485
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.09353    0.12089   9.045  < 2e-16 ***
## Temp              2.16115    0.23872   9.053  < 2e-16 ***
## Weathermisty      0.00318    0.09381   0.034  0.97297
## Weatherrain/snow -0.39590    0.13751  -2.879  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 652 degrees of freedom
## Multiple R-squared:  0.1247, Adjusted R-squared:  0.1207
## F-statistic: 30.97 on 3 and 652 DF,  p-value: < 2.2e-16
```

```
bikes.all <- lm(Casual ~ Temp + Weather, data=bikes)
plot(bikes.all, which=1)
```




```
plot(bikes.all, which=2)
```




```
summary(bikes.all)
```

```
##
## Call:
```

```
## lm(formula = Casual ~ Temp + Weather, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.397  -7.465  -2.380   6.704  28.789
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.8566     1.1724   1.584  0.11376
## Temp              22.6174     2.3150   9.770  < 2e-16 ***
## Weathermisty       0.2558     0.9098   0.281  0.77868
## Weatherrain/snow  -4.0497     1.3335  -3.037  0.00249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.38 on 652 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.138
## F-statistic: 35.97 on 3 and 652 DF,  p-value: < 2.2e-16
```

As seen in the linear regression equations of both models, R has created two dummy variables two dummy variables "Weather Misty" and "Weather Rain/Snow" are understood to be 0,0 for Clear, 1,0 for Misty, and 0,1 and for Rain/Snow.

As seen by the residual plot, the qq plot, and the summary of the model (LogCasual), this does not seem to accurate. The residuals are not normal, do not have a constant spread, do have a approximate mean of 0, and are not independent as there seem to be distinct patterns (downward sloping lines) shown in the graph. The qqplot appears troublesome as the points fall on the line in the middle but veer off substantially at the tails of the distribution. The coefficients b1, b2, b3 all are reasonable and line up with behaviors witnessed in the bivariate exploration so no violation there (b1 = +, b2(cloud/misty) = +, b3(snowy/rainy) = 1, base line approx = b2). With an r squared value of .12, only 12% that means that only 12% of the variation of the data can be attributed to the linear relationship. All in all, a very weak model.

As seen by the residual plot, the qq plot, and the summary of the model(Original Casual), this seems to be slightly more accurate. The residuals are again not normal, do not have a constant spread, do have a approximate mean of 0, and are not independent as there seem to patterns within the graph. The qqplot appears a little more promising as the points fall on the line in the middle but veer off slightly at the tails of the distribution. The coefficients b1, b2, b3 all are reasonable and line up with behaviors witnessed in the bivariate exploration so no violation there (b1 = +, b2(cloud/misty) = +, b3(snowy/rainy) = 1, base line approx = b2). With an r squared value of .14, only 14% that means that only 14% of the variation of the data can be attributed to the linear relationship. All in all, a very weak model but better than the transformed one.

So after comparing the r squared values, we will be using the unadjusted model.

Prediction

Using the model, we can predict number of casual hourly bikers with a .75 scaled temperature, .25 scaled windspeed, and on a misty or cloudy day. Since the model excludes windspeed, we can ignore it.

```
1.857+22.62*.75+.256
```

```
## [1] 19.078
```

I predict that on a hour with a .75 scaled temperature, .25 scaled windspeed, and on a misty or cloudy day, there will be about 19 casual hourly bikers using the bike sharing.

Discussion

In this project, we were tasked with with building a model that hopefully can accurate predict the number of casual hourly bikers in in the Washington D.C/Virginia. We learned that the response variable is related to the temperature (Percentage of Maximum), the windspeed (Percentage of Maximum), and the Weather. There was no multicollinearity issues between the numerical predictors (Temperature and Windspeed) so no issues there.

In the model that includes Temperature, Windspeed, and Weather, only 2 predictors are significant: Weather and Temperature.We noted that based on the plot of Number of Casualy Hourly Bikers ~ Windspeed showed no relationship between the two despite transformations on the two. So windspeed was excluded from the model.

When constructing the model, we chose to explore the possibilities of transformations of Casual Hourly bikers as from the histogram it was clear that it was strong skewed to the right. So we constructed two models one transformed with the logarithmic function and the original Casual hourly bikers distribution. Then after looking at the summaries of linear regression models and their residual and qqplots we concluded that the unadjusted model looked better as the qqplot was a better fit on the line. However, overall the model was a very poor fit on the data set and the relationship between the explanatory and response variables. The residual plots for both distrbutions violated multiple assumptions (constant spread, normal, independent) and the r squared values were very low (12% of transformed, 14% for original).

These residual violations, low r squared values, and other unusual behaviors are most likely due to the treatment of data for Temperature and Windpseed. Both were scaled or percentage of the maximum which introduced hard boundaries for the data 0 and 1. Thes rigid borders led to the data behaving very abnormal especially when treated as predictor variables (as seen in the scatter plots with Casual hourly bikers). Due to this reason, the data and model as a whole was not effective or accurate in prediction.

Overall, data analytics could play beneficial role in bike sharing. Companies should continue to use data and statistical models as with the right variables and proper data collection, one could create a useful model. I would be interested in seeing a prediction model for bike sharing with different variables such as location in the area, time of the year, and others.