Predicting Casual Users of Bike Sharing System

Ryan Wahler rwahler

Due Wed, March 13, at 11:59PM

Contents


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

Introduction

The most important piece of information for a bike share system is the amount of users it has in a given time period. This information allows the proper allocation of bikes to each of the system's locations on a given day. Additionally the system can also anticipate periods of high and low revenue and high costs from things like maintenance costs and beyond. There are two types of users for a bike sharing system, the intensive user, who's utilization on any given day and can be treated as a constant baseline by the system. The difficulty in allocation of bikes and other functions of the system are given by the casual user who's use is irregular and sporadic. If there are relationships that can be drawn between the habits of the casual user and the conditions on the day the system could be more efficient in its functions, being able to predict total demands of both the fixed intensive user and sporadic casual user.

Exploratory Data Analysis

Data

The sample for this data set was gathered on the hourly casual bike share users from the Washington D.C., Arlington, and DMV area. The sample is made up of 656 observations of the number of casual riders with

the following variables:

Casual: the number of casual bike users in a given hour.

Weather: type of weather in the given hour, (coded as clear, misty, rain/snow).

Temp: temperature(scaled as percentage of overall maximum)

Windspeed: wind speed(scaled as percentage of overall maximum)

For reference this is how the first few rows of the data appear:

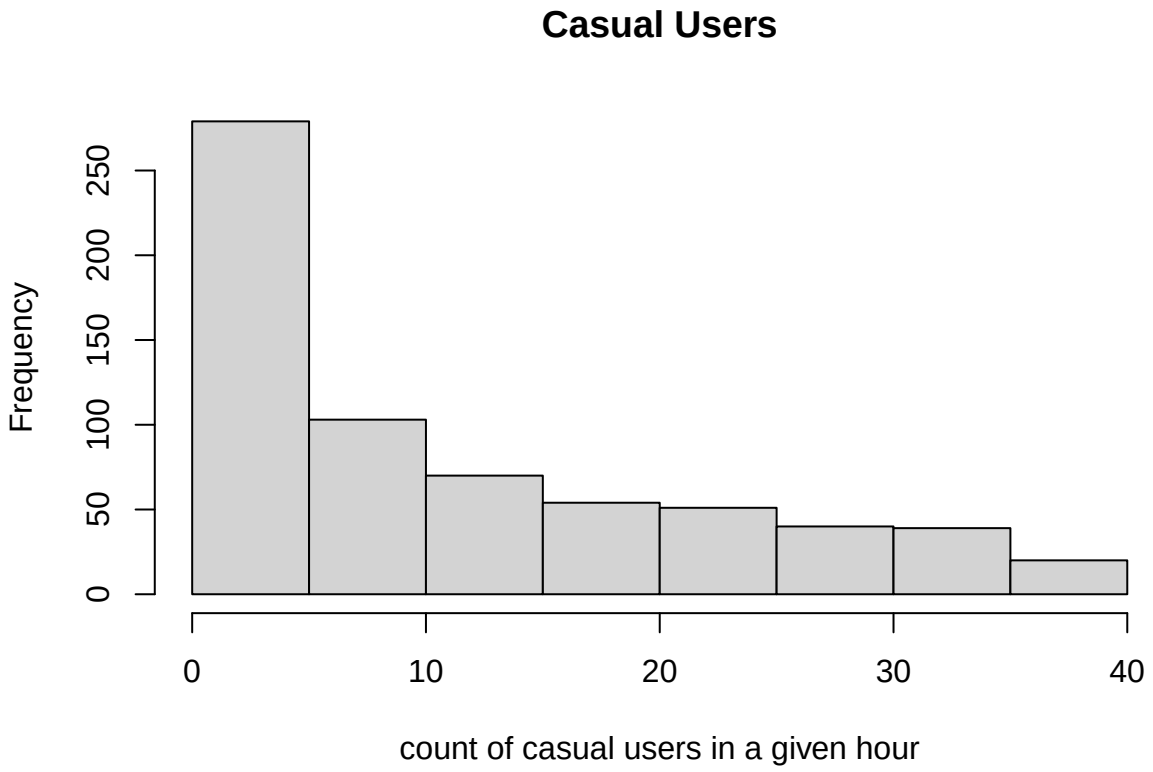
```
bikes
```

```
## # A tibble: 656 x 4
##    Casual Weather    Temp Windspeed
##     <dbl> <chr>     <dbl>     <dbl>
##  1      5 rain/snow  0.34    0.388
##  2      9 clear      0.34    0.104
##  3      6 misty      0.46    0.224
##  4     25 clear      0.34    0.298
##  5     31 clear      0.54    0.134
##  6     15 clear      0.32    0.254
##  7      2 misty      0.56    0.0896
##  8     35 clear      0.64    0.0896
##  9      5 rain/snow  0.54    0.0896
## 10      5 misty      0.28    0
## # i 646 more rows
```

Univariate exploration

Each variable will be explored individually, histograms will be a great aid in identifying the distribution of our continuous quantitative variables, Casual, Temp and Windspeed. While a barplot will give insight to the distribution of the categorical variable Weather. The five number summary also serves to give a numerical look at the each of the variables:

```
hist(bikes$Casual,
     main = "Casual Users",
     xlab = "count of casual users in a given hour")
```

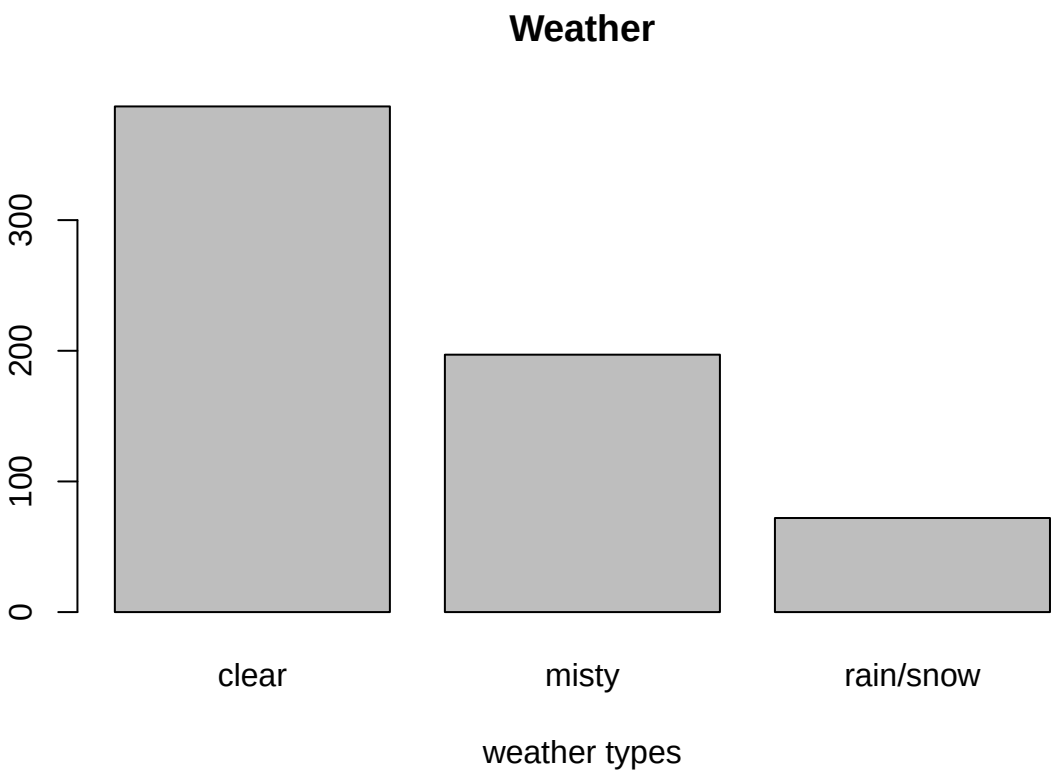



```
summary(bikes$Casual)
```

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

Upon analysis of the graphical and numerical summaries Casual is heavily right skewed with a large difference between mean and median. This strong right skew might give breath to a transformation of the data, that will be entertained later.

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

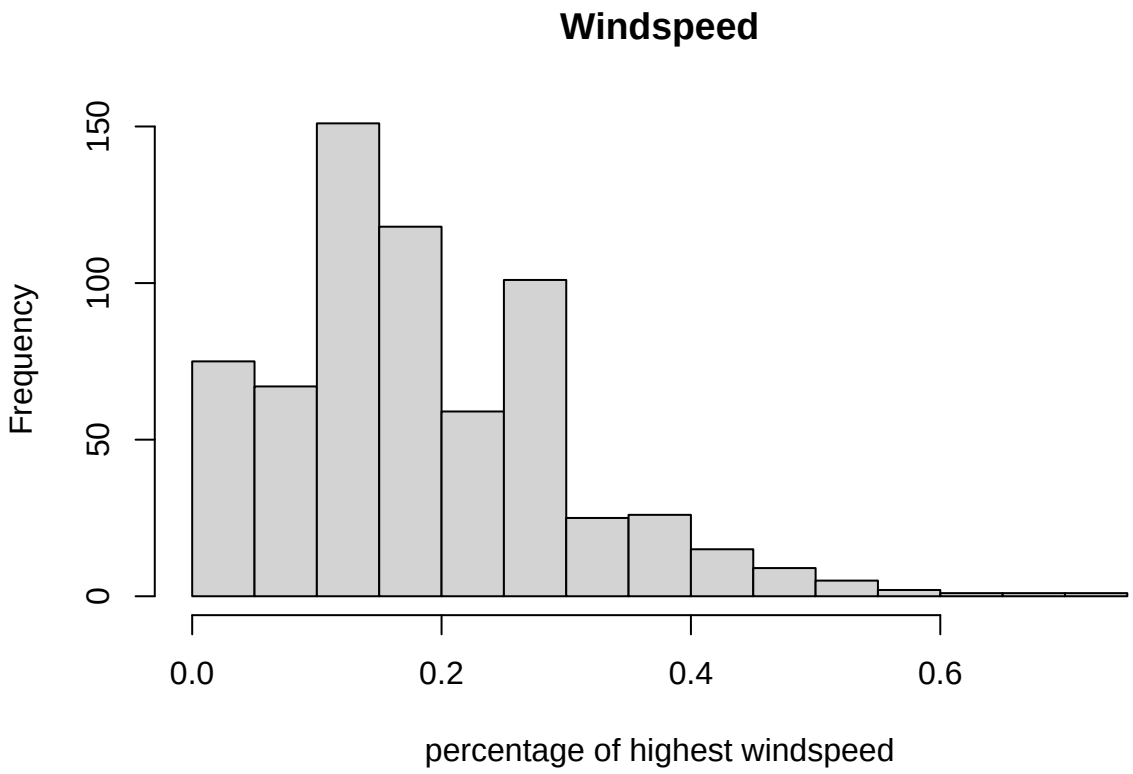



```
table(bikes$Weather)
```

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

Viewing the box plot and summary of Weather, we observe that there are many more cases of clear weather with 387 of the sample having clear weather, 197 of the sample having misty weather and 72 of the sample having some kind of precipitation .

```
hist(bikes$Windspeed,
     main = "Windspeed",
     xlab = "percentage of highest windspeed")
```

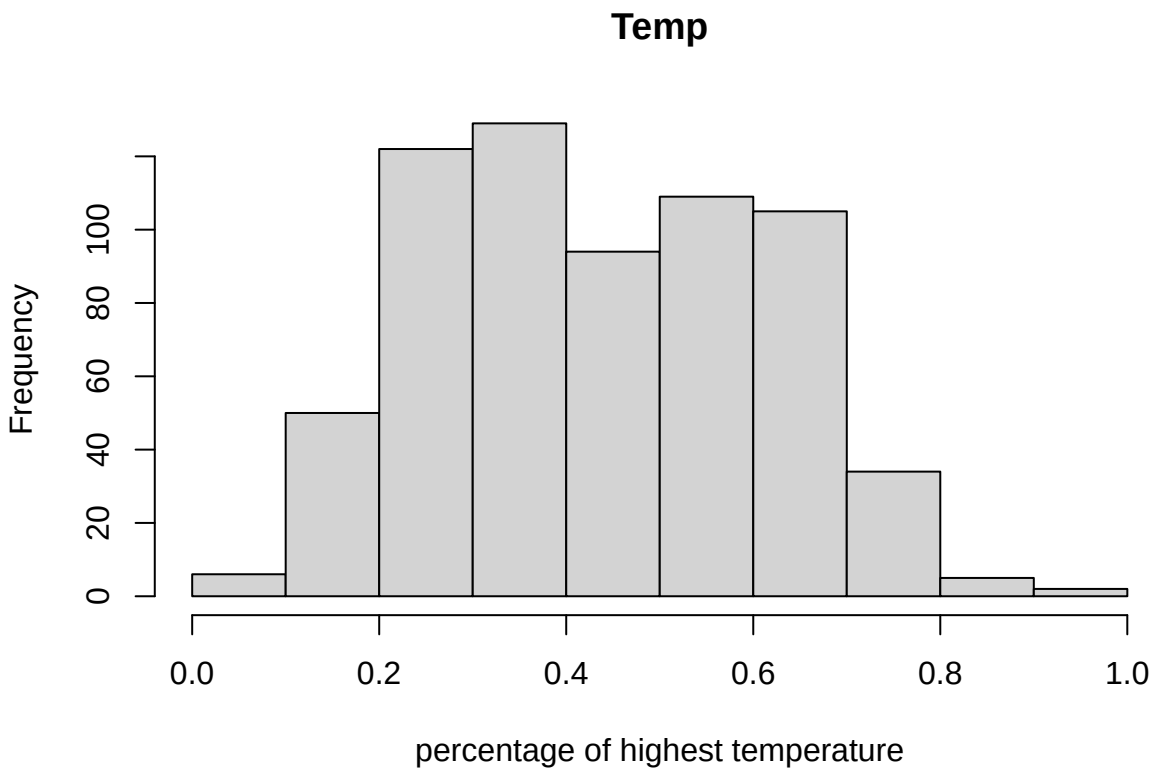



```
summary(bikes$Windspeed)
```

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

The Variable Windspeed produces a histogram that is largely unimodal with spikes around 0.175 and 0.225 with outliers around 0.6. Additionally the median is 0.1642 and the mean 0.1840 a difference given by a slight right skew.

```
hist(bikes$Temp,
     main = "Temp",
     xlab = "percentage of highest temperature")
```



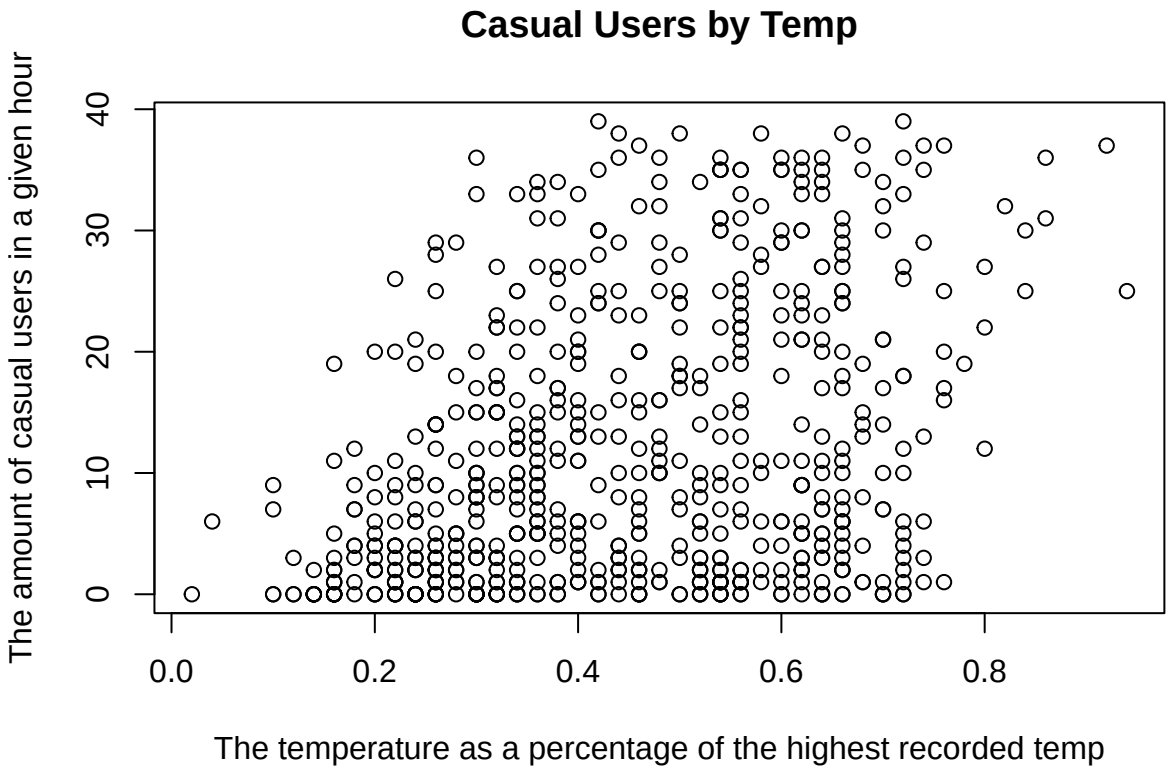

```
summary(bikes$Temp)
```

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

The histogram of Temp is the best looking so far, with a possibly bimodal distribution but we would need more information to distinguish this. Otherwise, the median and mean are nearly equivalent at 0.440 and 0.443 indicating a normal distribution.

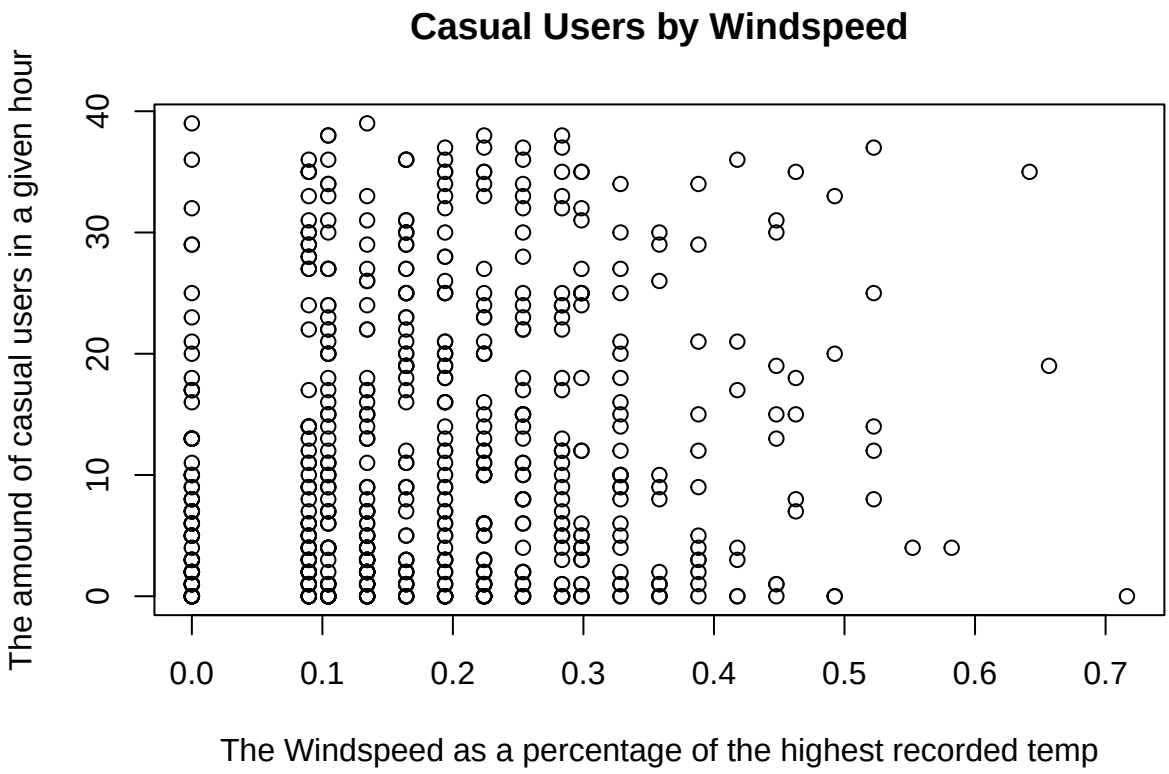
Bivariate exploration

```
plot(Casual ~ Temp,
    data = bikes,
    main = 'Casual Users by Temp',
    xlab = 'The temperature as a percentage of the highest recorded temp',
    ylab = 'The amount of casual users in a given hour')
```



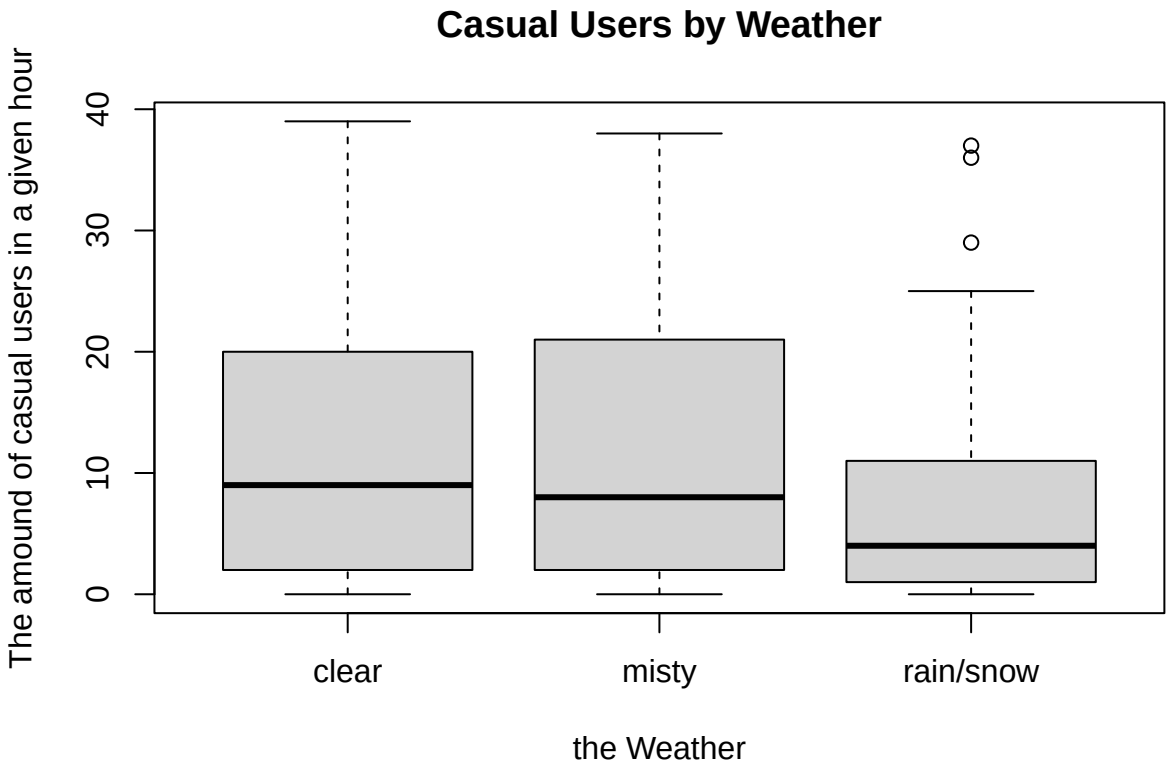
From the scatter plot above, Casual users seem to be positively and linearly associated with the temperature in the given hour, as the temperature increases the amount of casual users seems to increase as well.

```
plot(Casual ~ Windspeed,
    data = bikes,
    main = 'Casual Users by Windspeed',
    xlab = 'The Windspeed as a percentage of the highest recorded temp',
    ylab = 'The amound of casual users in a given hour')
```



The graph of Casual users plotted by wind speed seems to reveal a positive association, that holds a relatively weak association without a clear linear trend.

```
boxplot(Casual ~ Weather,
    data = bikes,
    main = 'Casual Users by Weather',
    xlab = 'the Weather',
    ylab = 'The amound of casual users in a given hour')
```

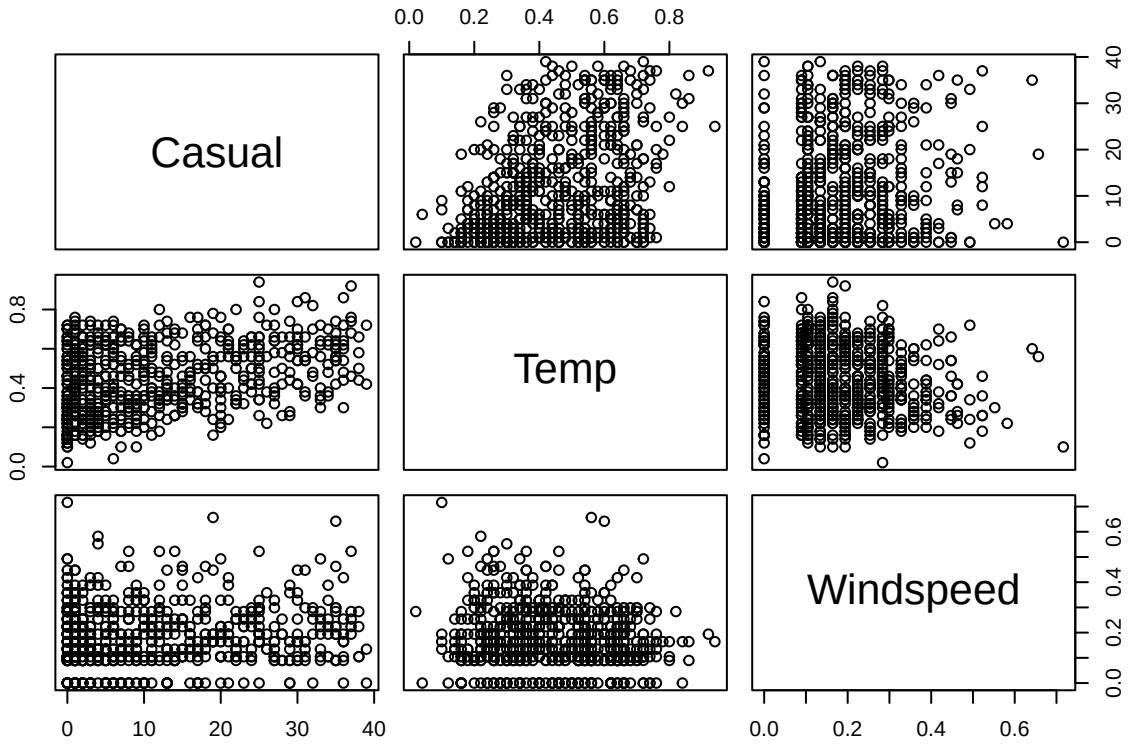


The difference in heights of the boxplots and their quartile measures does indicate an association between weather and the casual users. This is especially clear with the rain/snow weathering gathering the least casual users on average. But this relationship is not observedly so strong as each box does have a large portion of overlap in comparison to the others.

Modeling

As discovered in the bivariate exploratory data analysis, each of the possible explanatory variables did have some association with the response variable. Below is the matrix of correlations, to not these are relatively small coefficients of correlation between the respective explanatory variables and the response. In appeasing the linearity assumption, upon observation of the pairs plot there is no other clear relationship such as a quadratic one between any of the variables. This indicates that the model may yield a low R^2 score and not be a good predictor of the number of Casual users given the explanatory variables.

```
bikes.only.quant <- subset(bikes,
                           select = c(Casual, Temp, Windspeed))
pairs(bikes.only.quant)
```




```
cor(bikes.only.quant)
```

```
##                Casual       Temp  Windspeed
## Casual     1.0000000  0.3584811  0.1078030
## Temp       0.3584811  1.0000000 -0.1277144
## Windspeed 0.1078030 -0.1277144  1.0000000
```

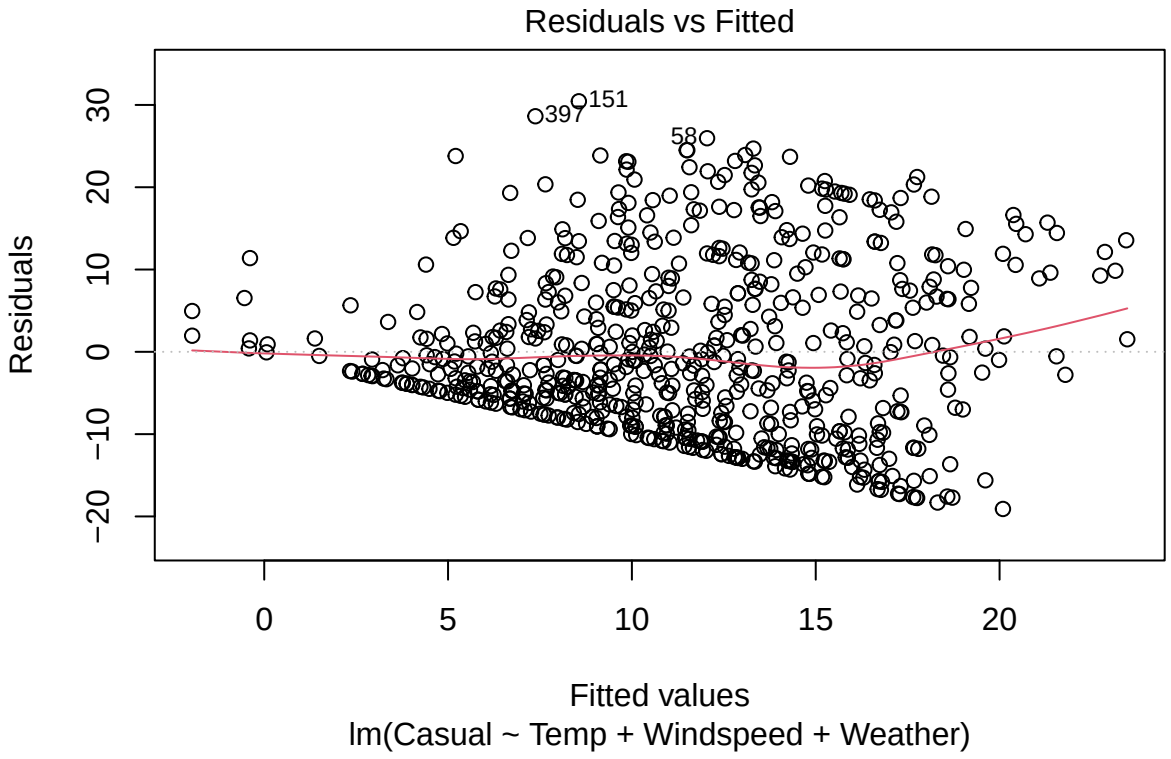
To rule out any form of multicollinearity we will be looking for any erroneous correlation between the explanatory variables. Formally checking the variation inflation factors (vif) of each of the explanatory variables in a multilinear regression model, will indicate which explanatory variables are unusable if any.

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

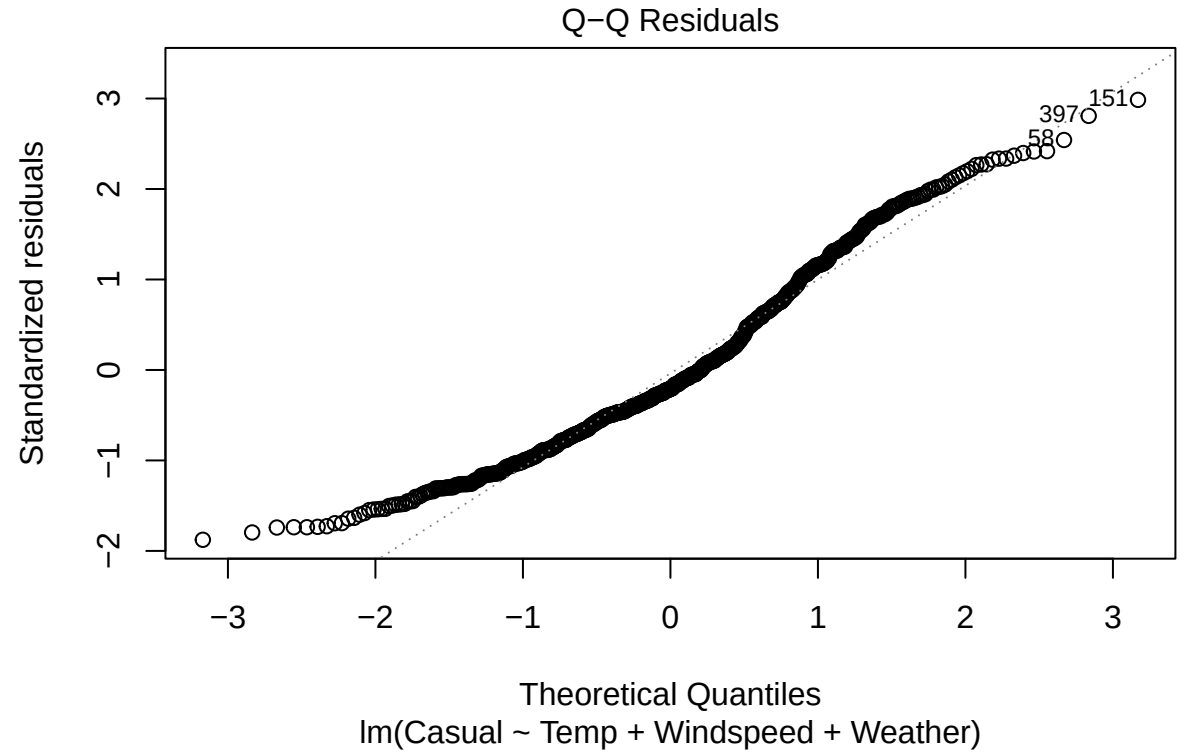
```
##                 GVIF Df GVIF^(1/(2*Df))
## Temp       1.020677  1        1.010285
## Windspeed 1.019349  1        1.009628
## Weather    1.006894  2        1.001719
```

Each of the GVIF values (given above) are below 2.5 and should be considered a non-issue, and can now proceed to fulfilling the four error assumptions.

```
plot(casualOriginal.full.mod, which = 1)
```




```
plot(casualOriginal.full.mod, which = 2)
```



By the residual plot we realize little deviation from the 0 mean across the domain, apart from the larger residuals around the 2.5 mark on the X axis. And with the absence of any obvious patters across the domain, the constant standard deviation, mean 0 and independence assumptions are justified. As for the QQ plot, throughout the domain, the points largely stay quite fitted to the line with the exception of small deviations

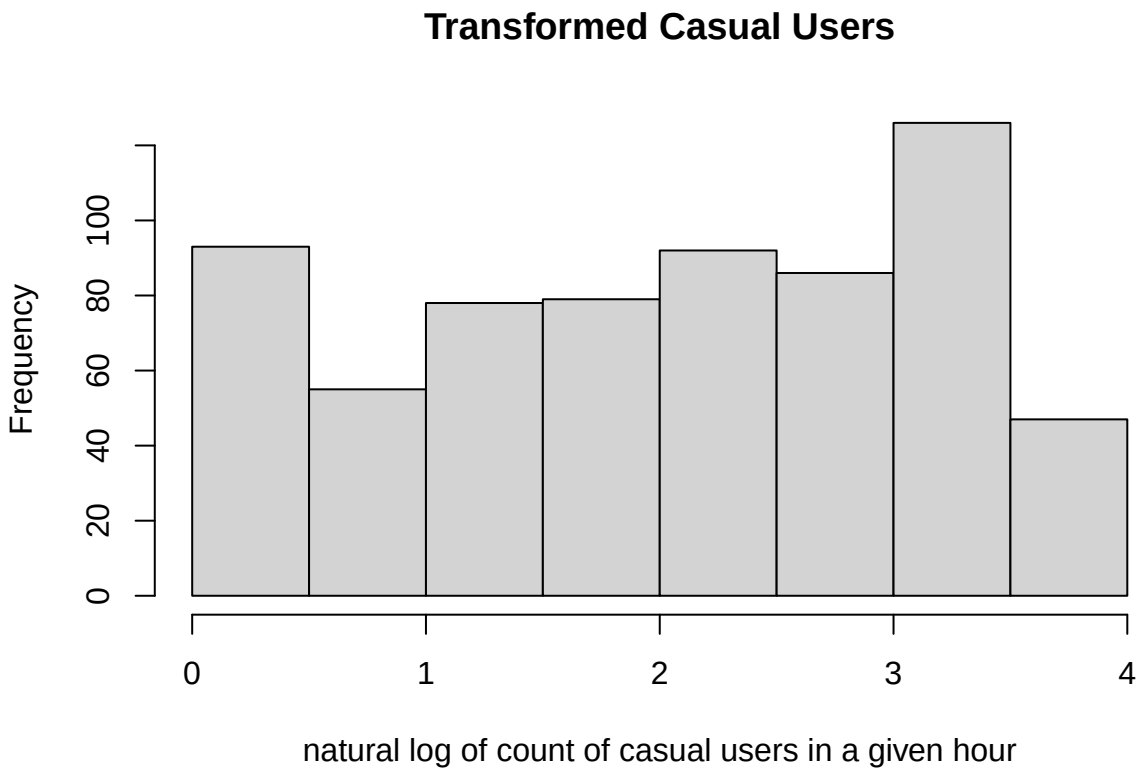
on either end.

In our EDA of our response variable, Casual, we identified a strong right skew that would complicate the interpretations and findings surrounding the data. To preform a transformation to achieve a more Gaussian distribution across the casual variable we will the natural log of each measure of casual users. Additionally because the minimum of the Casual variable is 0 we actually take the log of each measure of casual users plus 1.1.

```
min(bikes$Casual)
```

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

```
log.Casual <- log(bikes$Casual + 1.1)
casual.full.mod <- lm(log.Casual ~ Temp + Windspeed + Weather,
                      data = bikes)
hist(log.Casual,
     main = "Transformed Casual Users",
     xlab = "natural log of count of casual users in a given hour")
```

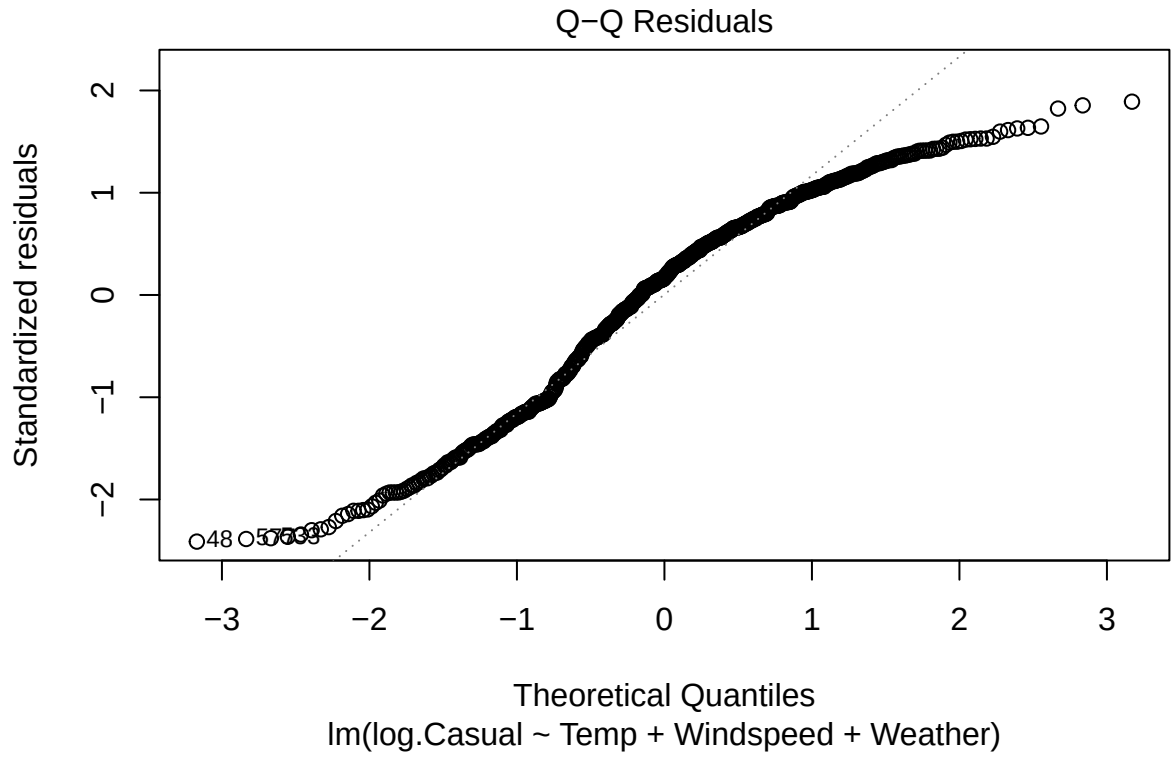



```
summary(log.Casual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09531 1.13140 2.20827 2.00820 3.04927 3.69138
```

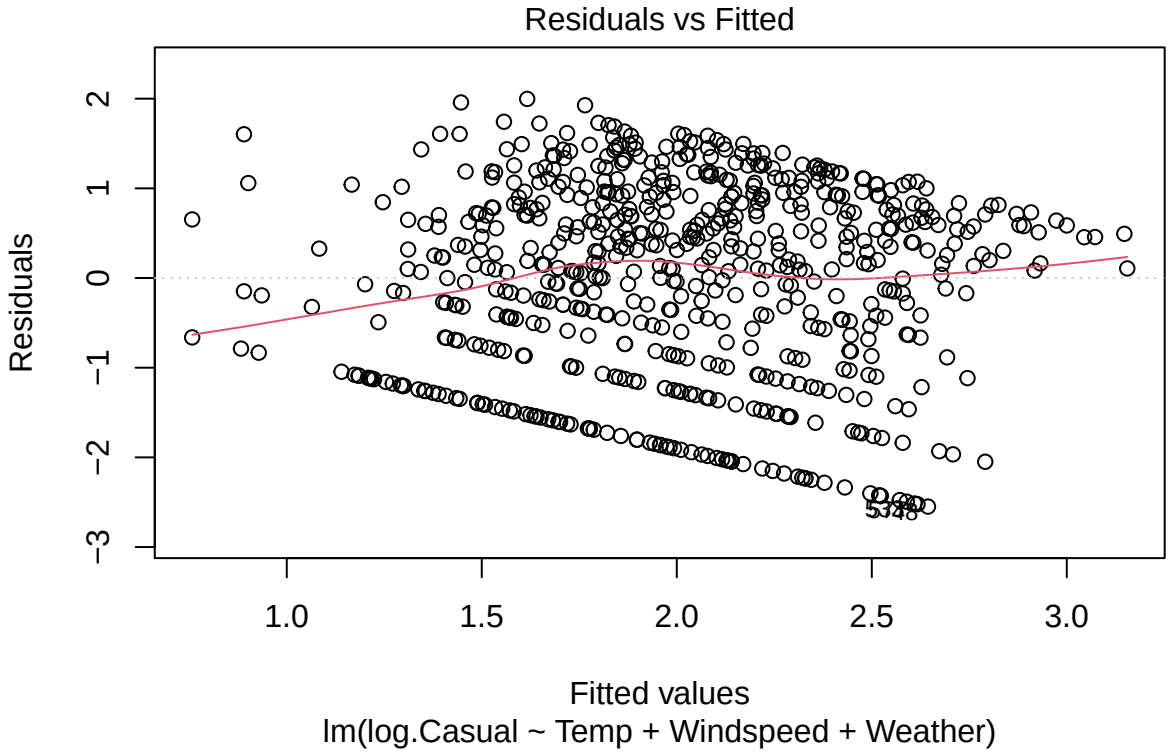
The Transformed Casual variable gives a largely symmetric and unimodal histogram with a mean of 2.01 and median of 2.21 a difference that does not indicate a large skew. From here we constructed a new linear model, log.Casual by Temp, WindSpeed and Weather. However on renewed attempts to fulfill the error assumptions with the transformed Casual Users the resulting QQplot and residual plot actually had worse outlooks

```
plot(casual.full.mod, which = 2)
```



Here in comparison to our first QQ plot, there are much higher deviations from the line at towards either end of the domain. While this new QQ plot may fulfill the normality assumption and the deviations could be permissible, it is a cause for concern.

```
plot(casual.full.mod, which = 1)
```



In this new residual plot, the mean zero assumption is less clearly justified, where especially at the end of the range there is deviation from mean zero. Again while this may be passing for our assumptions, it is still cause for concern and makes accepting this linear model as a reasonable predictor as questionable. Calculating and comparing the R^2 score of our two linear models, the linear model of transformed Casual actually had a smaller R^2 score of .142 while the linear model of untransformed Casual had a R^2 of .168. Based on the more difficult QQ and residual plots and the lower R^2 score, we will revert to the linear model of untransformed Casual.

```
summary(casualOriginal.full.mod)
```

```
##
## Call:
## lm(formula = Casual ~ Temp + Windspeed + Weather, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.092  -7.514  -2.109   6.751  30.443
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.4947     1.3749  -1.087  0.27738
## Temp              23.9334     2.3003  10.405  < 2e-16 ***
## Windspeed         15.0479     3.3460   4.497 8.15e-06 ***
## Weathermisty       0.3417     0.8968   0.381  0.70333
## Weatherrain/snow  -4.2943     1.3154  -3.265  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 651 degrees of freedom
## Multiple R-squared:  0.1679, Adjusted R-squared:  0.1627
## F-statistic: 32.83 on 4 and 651 DF,  p-value: < 2.2e-16
```

The model is significant by its F-test which gives a p-value of 2.2 X 10^(-16). All of our explanatory variables are also significant. To note, the p-value of Weathermisty is quite large and would suggest that it is insignificant and should be thrown out however the other category of the categorical variable is significant, Weatherrain/snow, compelling us to leave in the categorical variable Weather.

By running the 'regsubsets' function it was determined that including all the explanatory variables given gave the largest R^2 score from all possible combinations. Noting the R^2 score was still quite low, with only 16.8% of variation of casual users being determined by Windspeed, Temp and Weather. Also while attempting higher order models for Windspeed, and Temp, this linear model still gave the largest R^2 score. All the while not having violating VIF values, having proper residual diagnostics that preformed better than with transformations of the response variable, and not 'overfitting' to the sample. We can be reasonably confident that higher Temp, higher Windspeed and misty weather results in larger amounts of casual users, which follows our EDA finding the, relatively weak, positive associations of Windspeed and Temperature. Additionally the categorical variable Weather is seen to have a positive coefficient value for misty weather and a negative coefficient value for rain/snow weather, which were reflected in the box plot during EDA.

Prediction

To now have a prediction of the casual users in a given hour with a scaled windspeed of .25 and a scaled temperature of .75 and is calculated below:

```
-1.4947 + 23.9334 + 15.0479 + 0.3417
```

```
## [1] 37.8283
```

To note this prediction is made without transformations to the data and an incorporation of all given explanatory variables, because as noted above this combination created the best R^2 score while fulfilling the error assumptions.

Discussion

Throughout this discussion we discovered that casual users is predicted by quantitative variables Temp, Wind and the categorical variable Weather.

For further discussion, though our EDA we learned that there is not a large difference between misty weather and clear weather, while rain/snow weather is significant. It would be interesting if there were better categories for the weather that differentiated the data more distinctly. Additionally, what is still of concern is the skewness of the casual response variable, that possibly complicates interpretations of the model with higher a higher response value. In other words there are not many samples observes in the upper range of the domain of casual users, and any interpretation of the model in this area would likely be not as reliable. If possible it would be beneficial to gather more data concerning higher casual rider counts in a given hour, and the explanatory variables that come along with those observations.

The analytics built here will service the bike share industry in their allocation of resources and stations for the bikes in the DMV area. This data will also help out city planners that are attempting to make transportation more equitable for people that choose cycling over other forms of transportation and so on.