Predicting Amount of Bike Rental Users

Vishal Saikrishnan
vsaikris

Due Wed, Oct 21, at 8:00PM (Pittburgh time)

Contents

Introduction

As bikes have become more common, new bike sharing systems have been created. These systems are the new generation of bike rentals, where membership, rental, and return are all automatic. Anyone is allowed to easily rent a bike from one location and return that same bike at a different location. Currently, the world has over 500 bike sharing programs. But, along with this system comes effects on traffic, environment, and the user's health. In this paper, we will be exploring the characteristics that are associated with an increased/decreased amount of users. In other words, when does the use of bike sharing increase, and when does it decrease.

Exploratory Data Analysis

Data

For this study, we will analyze a random sample of casual bikers in the Washington D.C. area across 656 days and 4 variables. Because we want to know what characteristics can predict an increase/decrease of bike rental use, we are examining 3 explanatory variables: *Weather, Temperature, and Windspeed*. The response variable is *Casual*. Summaries of the variables are as follows:

Casual: Number of Casual Bike Users

Weather: Type of weather

- 1 = clear, few clouds, partly cloudy

- 2 = mist & cloudy, mist & broken clouds, mist
- 3 = light snow, light rain/thunderstorm/scattered clouds

Temperature: The temperature (scaled as a percentage of overall maximum)- So if The maximum temperature was 108 F, then a temperature of 81 F would be written as 0.75

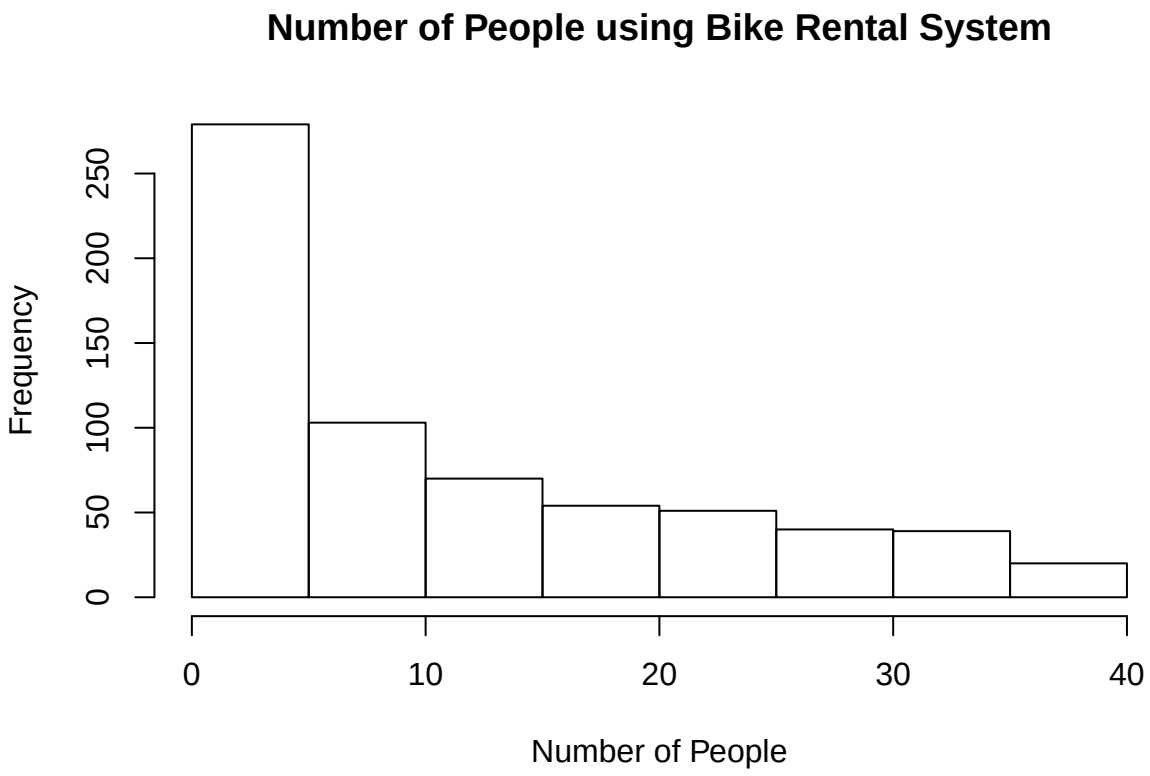
Windspeed: The Windspeed (scaled as percentage of overall maximum) - So if the maximum Windspeed was 30 mph, then a windspeed of 15 mph would be written as 0.5

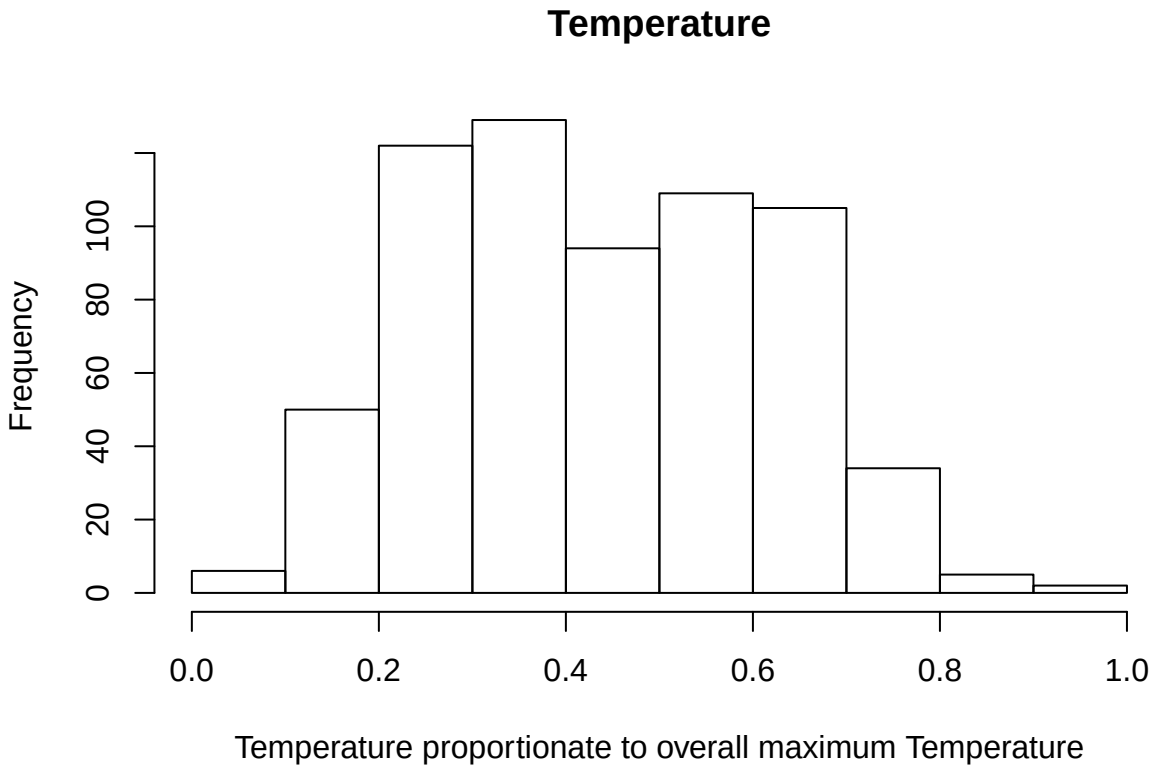
Below are the first few lines of data:

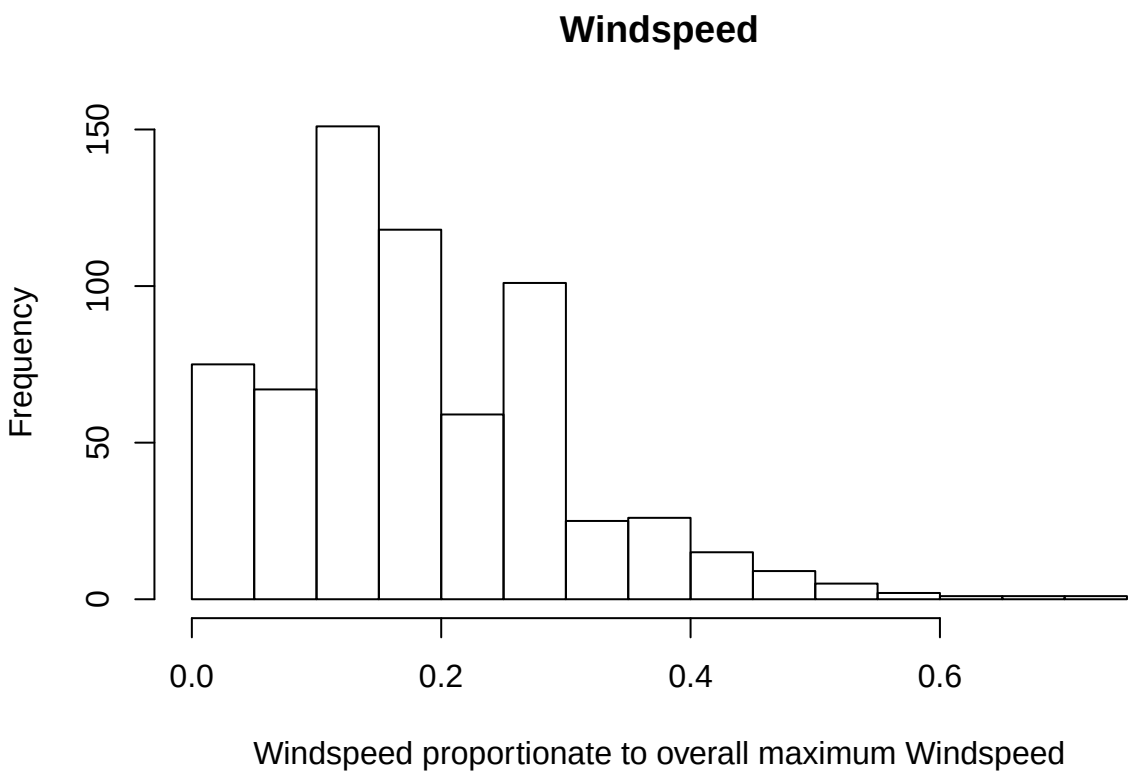
```
## # A tibble: 6 x 4
##   Casual Weather    Temp Windspeed
##    <dbl> <chr>     <dbl>     <dbl>
## 1      5 rain/snow  0.34     0.388
## 2      9 clear      0.34     0.104
## 3      6 misty      0.46     0.224
## 4     25 clear      0.34     0.298
## 5     31 clear      0.54     0.134
## 6     15 clear      0.32     0.254
```

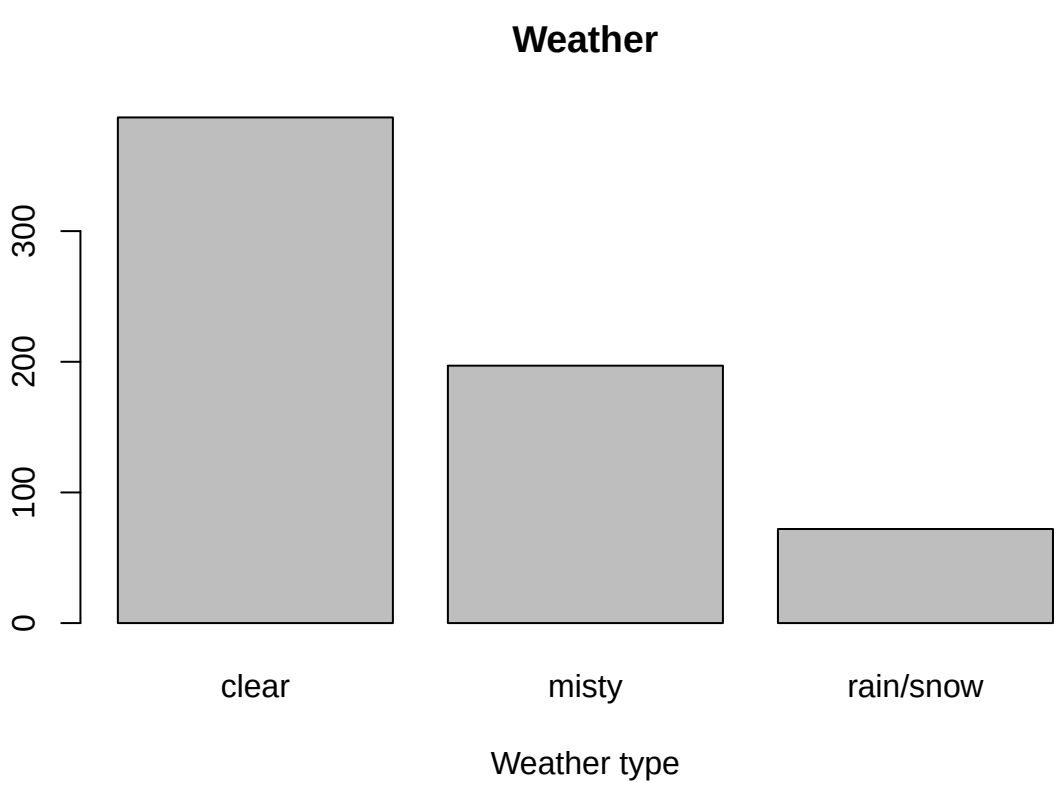
Univariate Exploration

To begin our analysis, we are exploring each variable by itself. Here are the histograms for the quantitative variables and a barplot for the categorical variable of Weather.


Number of People using Bike Rental System
Frequency
0
50
100
150
200
250
0
10
20
30
40
Number of People

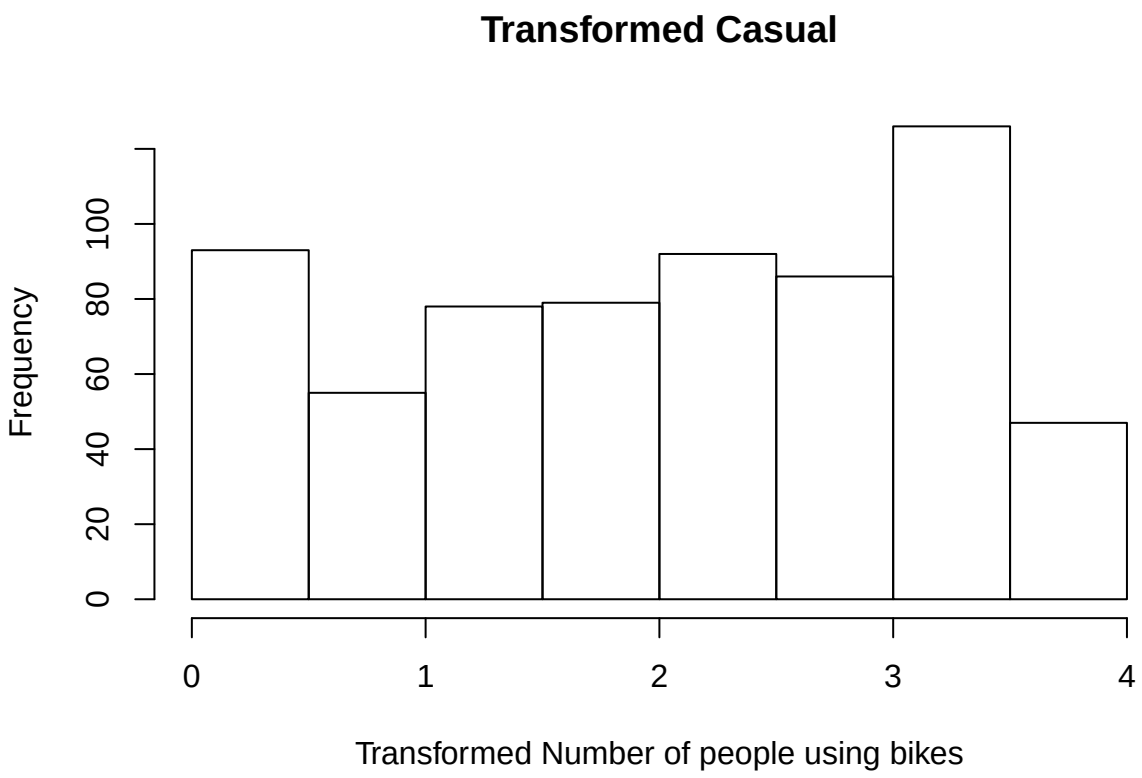

Temperature
Frequency
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Temperature proportionate to overall maximum Temperature

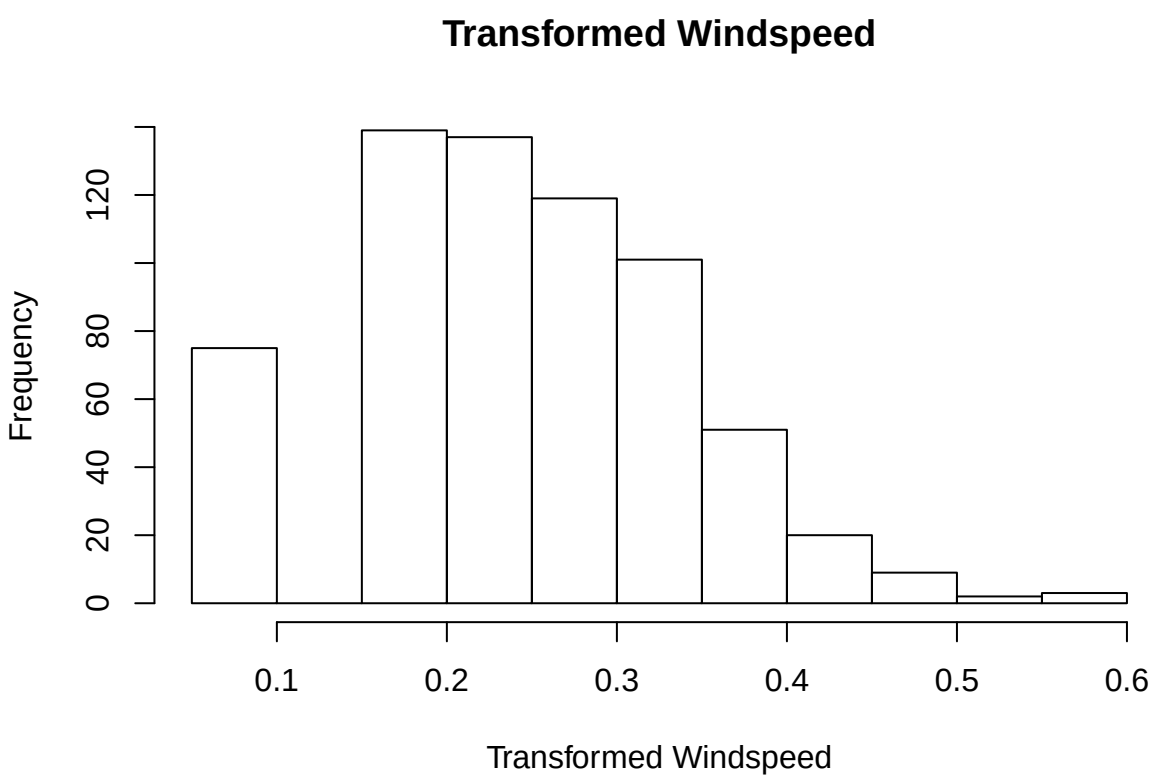

Windspeed
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
Windspeed proportionate to overall maximum Windspeed



We can see from the histograms for variables *Casual* and *Windspeed* that they are both skewed right. There are two solutions to this problem. We can either transform the variables using a log shift, or just simply remove the rows with 0s in both *Casual* and *Windspeed*. However, there are above 70 rows in each variable with 0, so we cannot remove them. Plus, 0 bikers or 0 Windspeed is valuable data and key to understanding the relationships between variables. Therefore, we will have to transform these variables using a log shift so that we get a normal spread. Shifting the data, though, causes complications in the explanations and analysis that follow. We will have to slightly adjust our conclusions (if we choose a model with transformations), which is better than simply removing the data. So I am making the decision to conduct a log shift on both *Casual* and *Windspeed*. The shift will be +1.1 for each variable, and then we will log that value.

Here are the shifted histograms for *Casual* and *Windspeed*:


Transformed Casual
Frequency
0
20
40
60
80
100
0
1
2
3
4
Transformed Number of people using bikes



As you can see by these histograms, they look more symmetric and normal than the other non-transformed histograms. We supplement the histograms with the following numerical summaries:

For Casual:

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

For Transformed Casual:

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

For Temperature:

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

For Windspeed:

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

For Transformed Windspeed:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09531 0.18606 0.23444 0.24576 0.30284 0.59686
```

For Weather:

```
##
##     clear     misty rain/snow
```

```
##         387         197          72
```

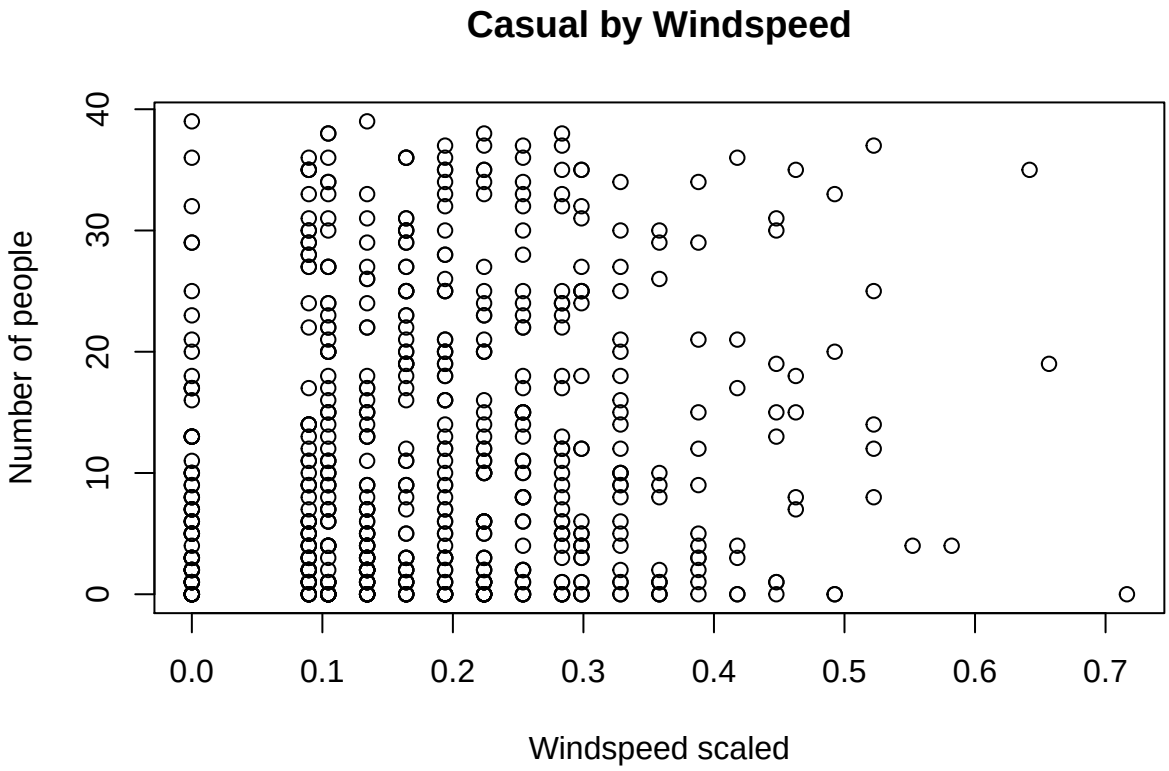
From the histograms and the summaries of the variables, here are my observations:

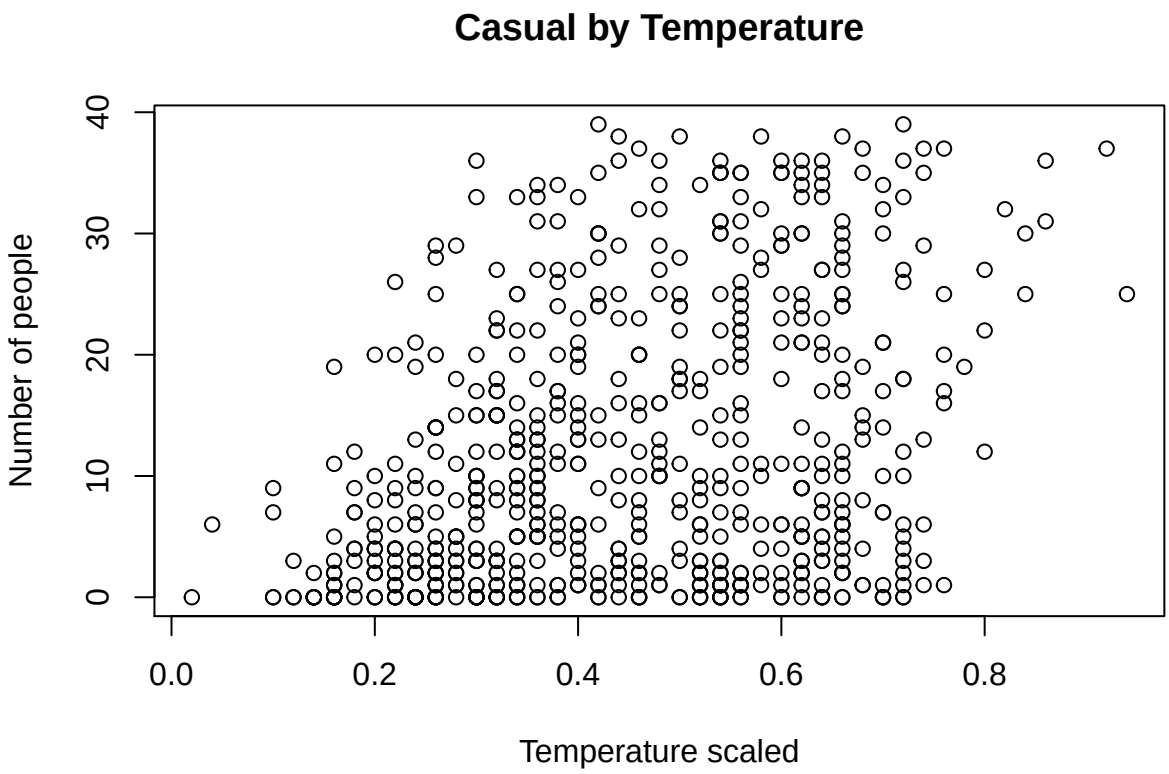
- Distrbution of **Casual** is skewed right. The mean is higher than the median by 3, which proves the skewness. The max is 39, while the min is 0 and the median is 8.
- Distribution of **Transformed Casual** is very symmetric, and approximately even across all bars. The median and mean are very close to eachother, indicating a normal distribution. The IQR is around 2.8 with the range being around 3.6. There seems to be a spike around 3-3.5, which makes sense because this clumps all the original **Casual** data from around 20-40 people.
- Distribution of **Temperature** is normally distributed and symmetric. The mean and median are almost exactly the same, but the histogram shows that this variable almost looks bimodal. IQR is around 0.28 and the range is 0.92. This suggests that in our study, we had a wide range of Temperatures, but the majority of temperatures stayed between the IQR and close to the mean.
- Distribution of **Windspeed** is skewed right. The mean is higher than the median by 0.02. The max windspeed is 0.71, while the third quartile is at 0.25. The median is 0.16. These numbers prove the right skewness of this variable.
- Distribution of **Transformed Windspeed** is still slightly skewed right, but the histogram looks much better than the untransformed variable. With the numbers from the summary, we see that the mean and median are relatively close, the IQR is around 0.12 and range is about 0.5. One thing odd about the histogram is that there are no transformed data points between 0.1 and 0.15.
- The barplot shows us that **Weather** was mostly clear. There were 387 days of clear weather, 197 days of misty weather, and 72 days of rain/snow.

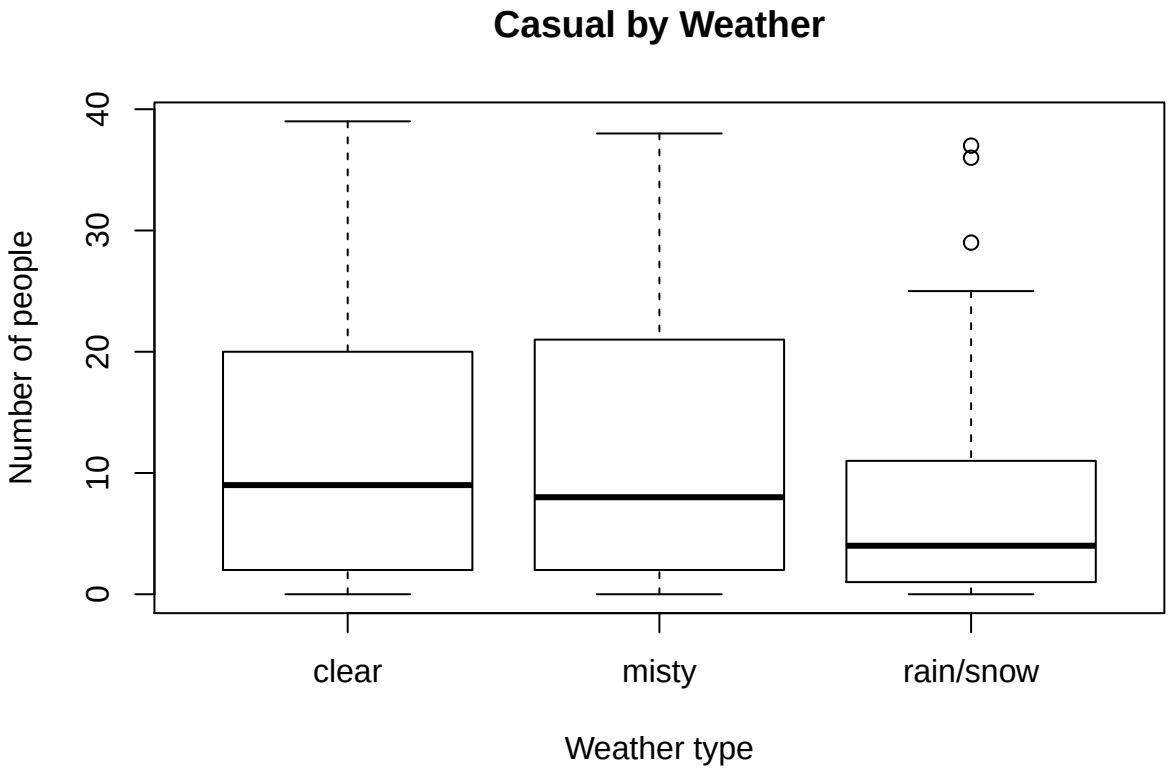
Bivariate Exploration

Now, after univariate exploration, we can graphically show the relationship between the variables. We will be using **Casual, Temperature, Windspeed, and Weather** as the four variables.

Here are the following scatterplots between each explanatory variable and the response variable:


Casual by Windspeed
Number of people
0
10
20
30
40
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Windspeed scaled


Casual by Temperature
Number of people
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
Temperature scaled

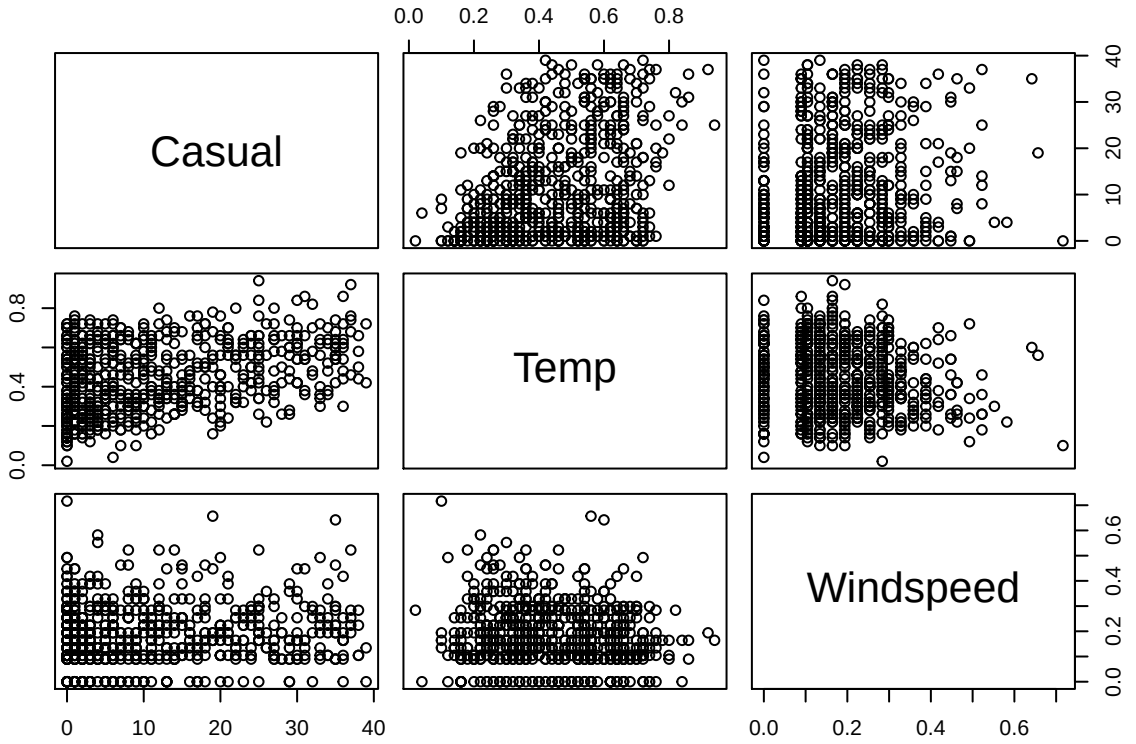


Analysis of Bivariate Exploration

- The scatterplot between **Casual** and **Windspeed** look mostly random. It looks very weakly correlated, but just slightly positive. This would mean as windspeed increases, the number of bike users increases.
- The scatterplot between **Casual** and **Temperature** looks positive and moderately correlated. The association definitely looks stronger than the association between Casual and Windspeed. Theoretically, as Temperature increases, the number of bike users will too.
- The boxplot between **Casual** and **Weather** shows us that clear and misty weather could almost be interchangeable. Both of them have approximately the same median, and IQR. In fact, the boxes look almost identical. Rain/snow weather, however, has a considerably lower median than clear/misty weather. This suggests that clear/misty weather doesn't affect biker showout, but rain and snow does. Rain and snow data also has a few outliers, which is expected in a random sample of 656 days.

Modeling

We now look to build a linear model between the variables that could predict the amount of bike users on any day, given weather, windspeed and temperature. Before we make the model, let's look at the correlations between the explanatory variables. We will only be using the quantitative variables for this comparison.



There doesn't to be any strong correlation between Temperature and Windspeed. To make sure, lets check the numerical correlations.

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

Looking at this data, I believe there are 3 appropriate models that could be created:

- The first model would use use Casual as the response variable, and Weather, Temperature, and Windspeed as the explanatory variables. No transformations in this model. I would like to see the R^2 value and if the beta values are statistically significant for the raw data.
- The second model would use shifted Casual as the response variable and Weather, Temperature, and Shifted Windspeed as the explanatory variables. If the diagnostics of the first model don't look good, we will see if this model with the transformed variables improve them.
- The third model wouldn't have Windspeed. Casual would be the response variable, while Weather and Temperature are the only two explanatory variables. If we look at the scatterplot between Casual and Windspeed, there seems to be absolutely no correlation at all (0.10). However, I can't remove a variable just because it is weakly correlated with the response variable. This is why I want to create this model and see if the R^2 and beta values are significantly higher than the other two models. Plus, if the diagnostics are better, I may have reasonable justification to drop Windspeed and use this model.
- I will also check if an interaction model is appropriate for each model stated above.

Remember: I have to check the diagnostics for all three models. Combining analysis of the diagnostics, R^2, and statistically significant beta values, I will choose the best model and use it for predicting the response variable.

Model 1

Model 1:

y = Casual

x1 = Temp

x2 = Weather

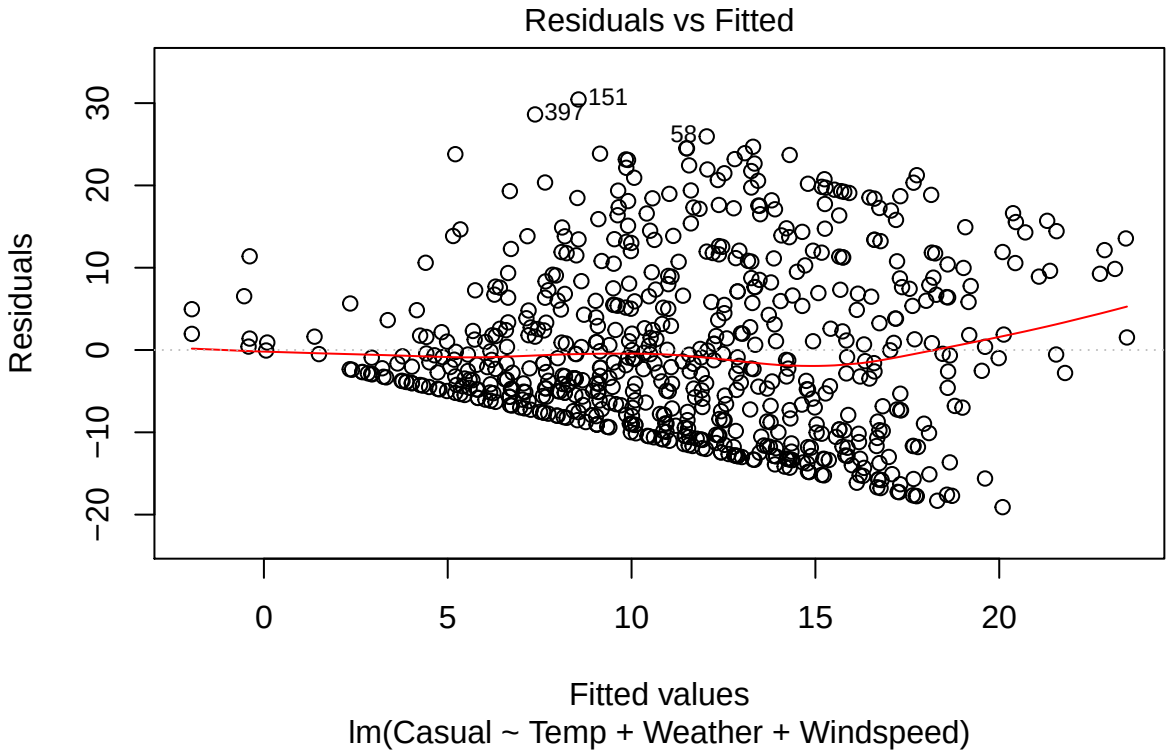
x3 = Windspeed

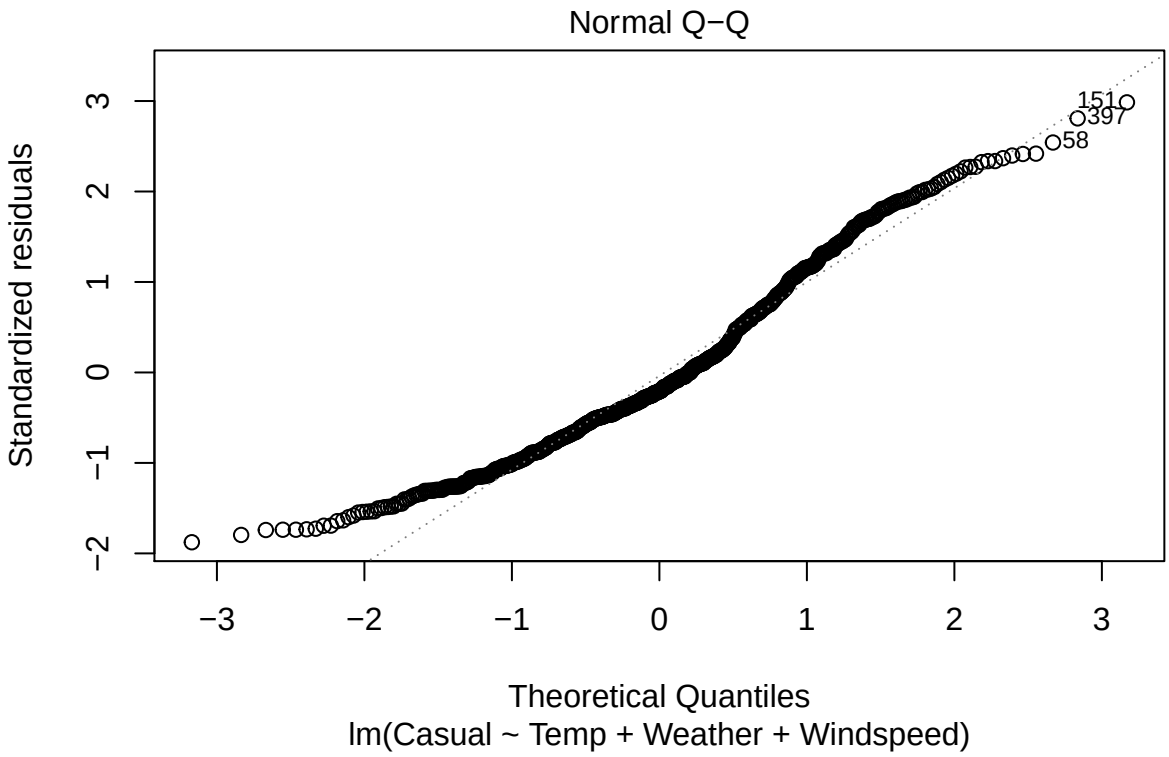
Here are the vifs, the model, and the diagnostics:

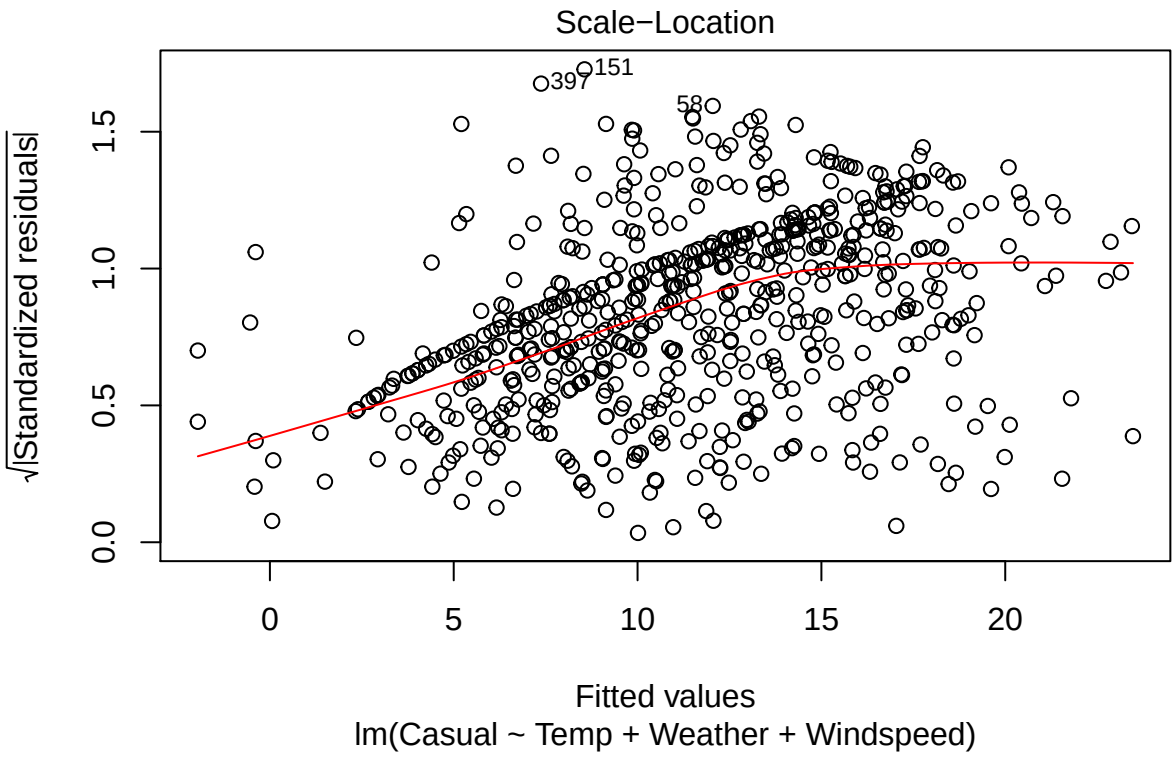
```
##                GVIF Df GVIF^(1/(2*Df))
## Temp       1.020677  1        1.010285
## Weather    1.006894  2        1.001719
## Windspeed 1.019349  1        1.009628
```

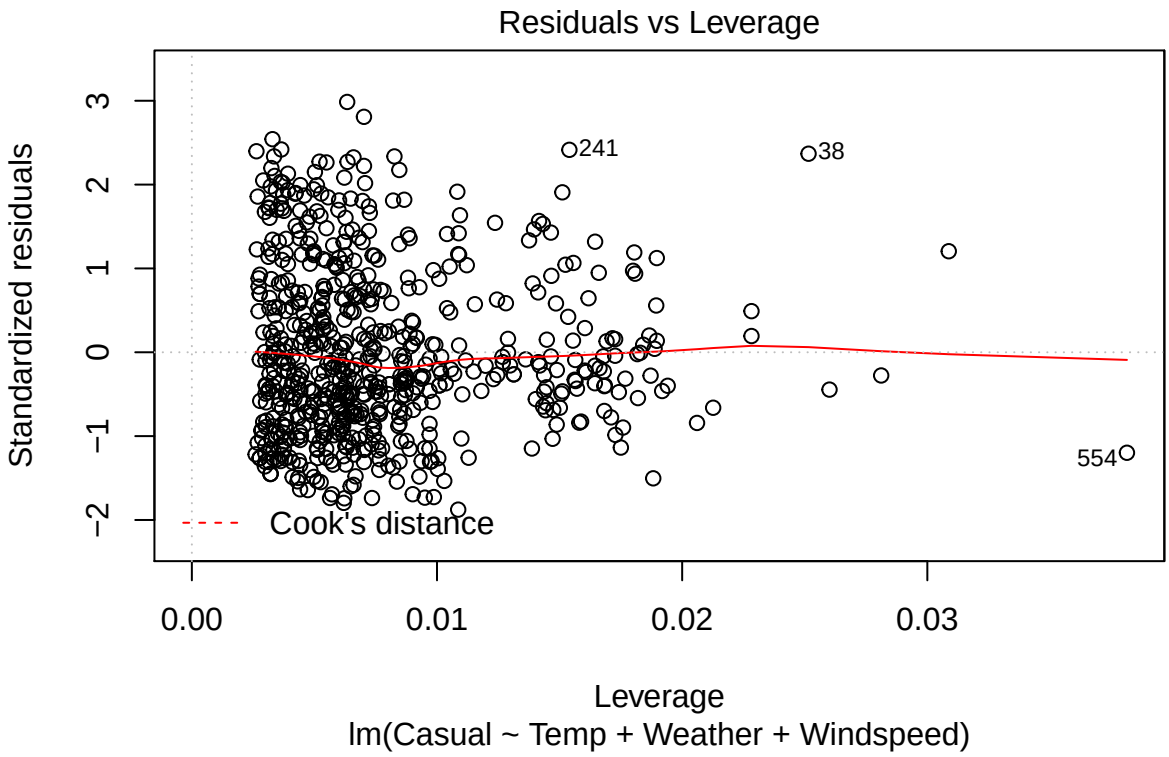
All the VIFs are below 2.5, suggesting NO dangerous multicollinearity.

Model #1 and Diagnostics:




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Casual ~ Temp + Weather + Windspeed)
151
397
58
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
5
10
15
20
397
151
58
Fitted values
lm(Casual ~ Temp + Weather + Windspeed)




```
##
## Call:
## lm(formula = Casual ~ Temp + Weather + Windspeed, 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 ***
## Weathermisty       0.3417     0.8968   0.381  0.70333
## Weatherrain/snow  -4.2943     1.3154  -3.265  0.00115 **
## Windspeed         15.0479     3.3460   4.497 8.15e-06 ***
## ---
## 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
```

Looking at the residual plot, the errors look mostly independent. We see a "line" of errors at the bottom, but the red line also takes a dip at the same spot. The errors also have mean 0 and a constant sigma. There are a few outliers at the top, but as a whole, this residual plot checks off 3/4 of the assumptions. With the QQ plot, almost all of the dots are on, or very close to the line. The bottom dots trail off, but we can conclude that these errors are normal. There are two insignicant beta values–the Intercept and one of the dummy

variables for the categorical variable. But since the other dummy variable is statistically significant (0.00115 < 0.05), we conclude that the categorical variable is stasitically significant. Finally, the R^2 value 0.1679 is, and the F test shows the model is statistically significant

Model 2

Model 2:

y = Casual (transformed)

x1 = Temperature

x2 = Weather

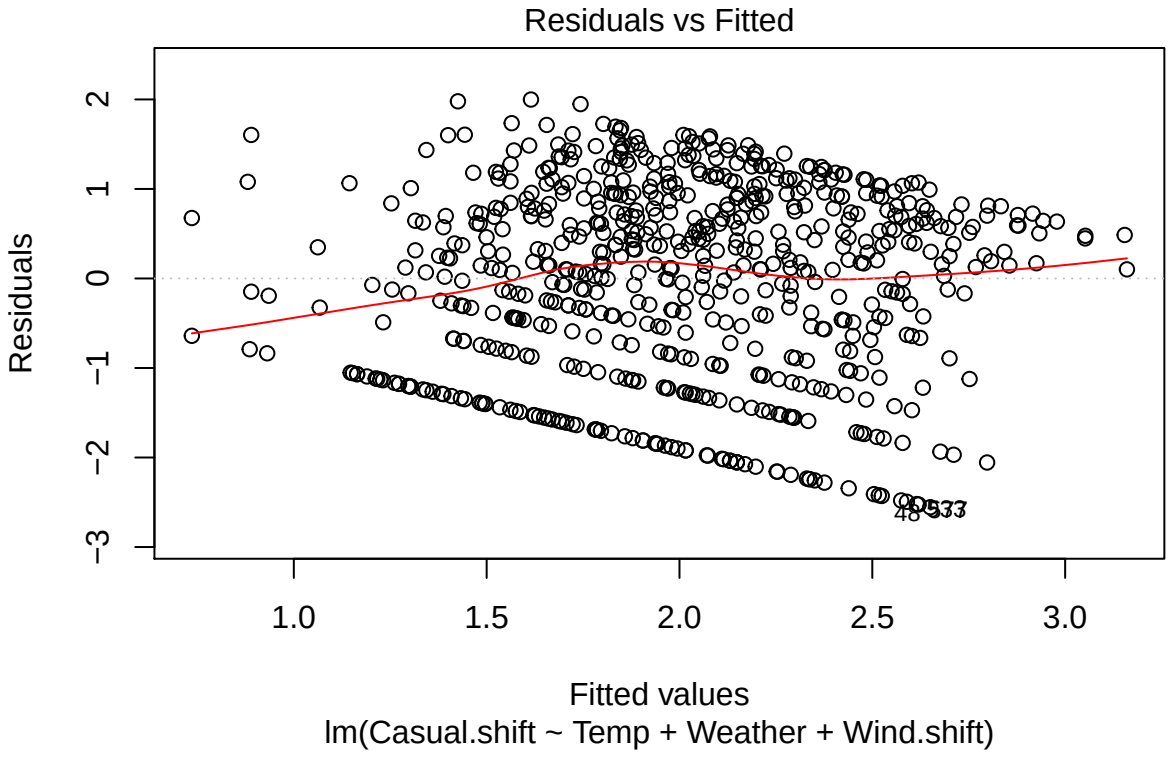
x3 = Windspeed (transformed)

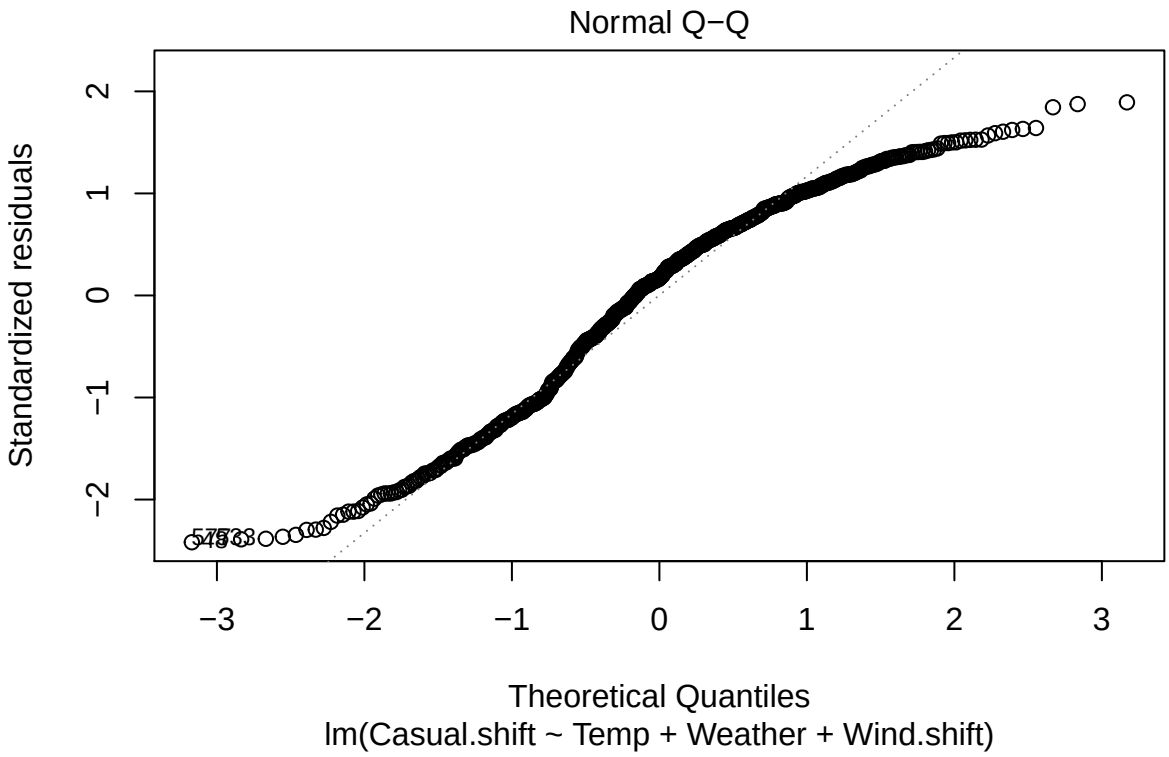
Here are the vifs, the model, and the diagnostics:

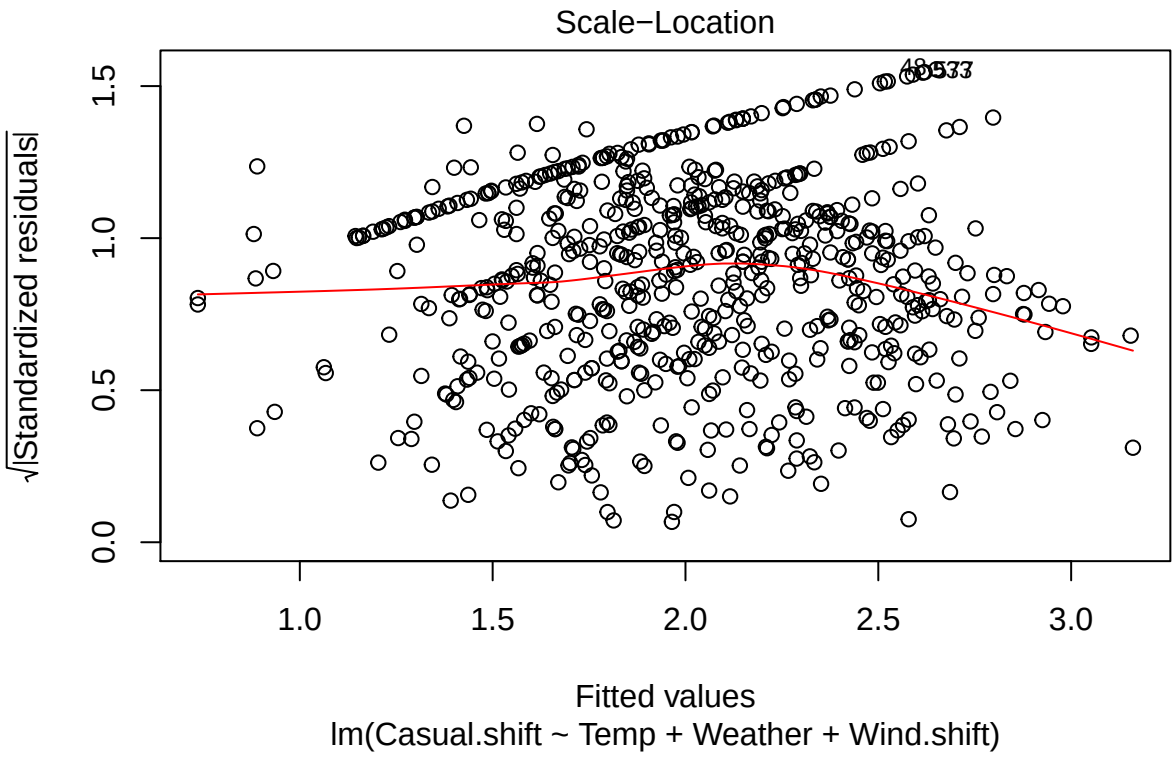
```
##                GVIF Df GVIF^(1/(2*Df))
## Temp       1.019726  1        1.009815
## Weather    1.007123  2        1.001776
## Wind.shift 1.018617  1        1.009266
```

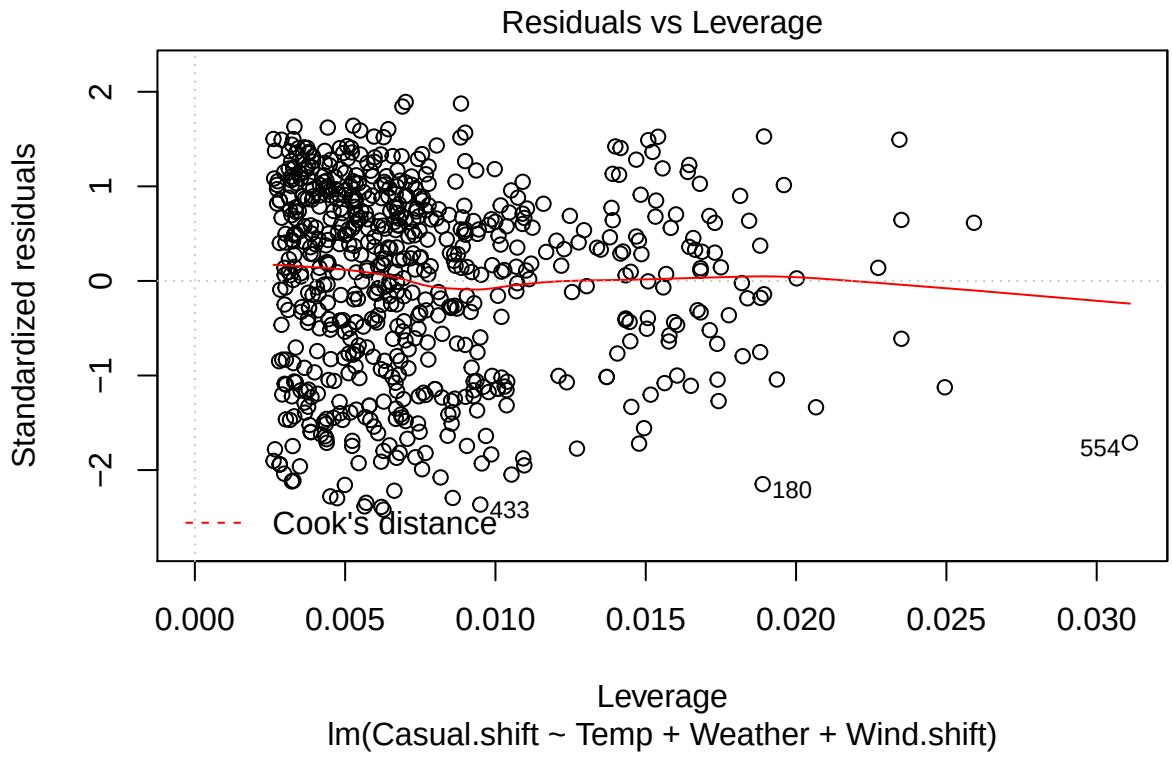
All the Vifs are under 2.5, suggesting NO dangerous multicollinearity.

Model #2 and Diagnostics:




Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Casual.shift ~ Temp + Weather + Wind.shift)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.0
1.5
2.0
2.5
3.0
Fitted values
lm(Casual.shift ~ Temp + Weather + Wind.shift)




```
##
## Call:
## lm(formula = Casual.shift ~ Temp + Weather + Wind.shift, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5550 -0.8262  0.1780  0.8334  1.9987
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.62812    0.17313   3.628 0.000308 ***
## Temp              2.27067    0.23822   9.532  < 2e-16 ***
## Weathermisty      0.01105    0.09293   0.119 0.905365
## Weatherrain/snow -0.41753    0.13630  -3.063 0.002279 **
## Wind.shift        1.69646    0.45589   3.721 0.000215 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 651 degrees of freedom
## Multiple R-squared:  0.1429, Adjusted R-squared:  0.1377
## F-statistic: 27.14 on 4 and 651 DF,  p-value: < 2.2e-16
```

From the Residual plot, we see that the data points are mostly independent, aside from some residuals that look like they form a line at the bottom of the plot. We can also see that the errors have mean 0 and constant sigma. When we look at the QQ plot, we see that the middle bulk of points fall on, or very close to the line. The top and bottom, specifically the top, doesn't follow the line, suggesting that the errors might not be completely normal. We need to keep this in mind before we choose the right model. All the beta values are

significant except the beta2 value, which corresponds with one of the dummy variables for the categorical variable. However, since the other dummy variable is statistically significant ($0.022 < 0.05$), we conclude that the categorical variable IS statistically significant. Finally, we see that the R^2 value is 0.1429, which is smaller than Model 1. So far, this model looks a bit worse than Model 1.

Model 3

Model 3:

y = Casual

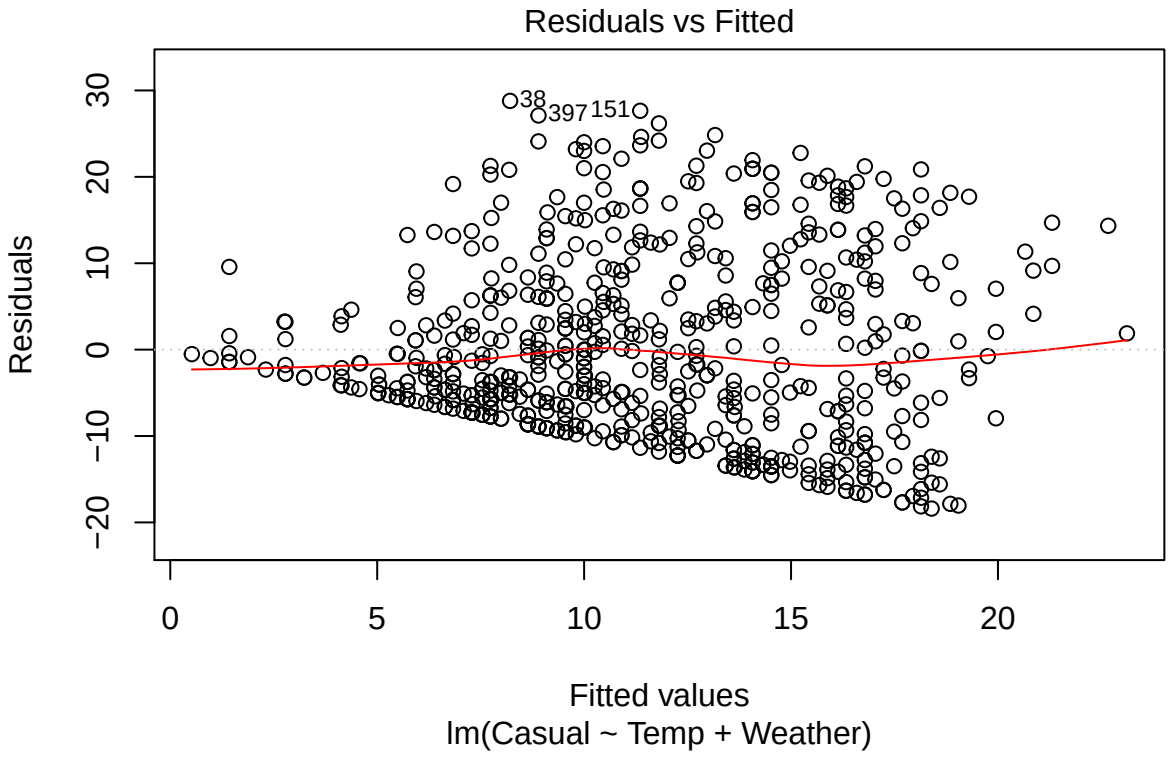
x1 = Temp

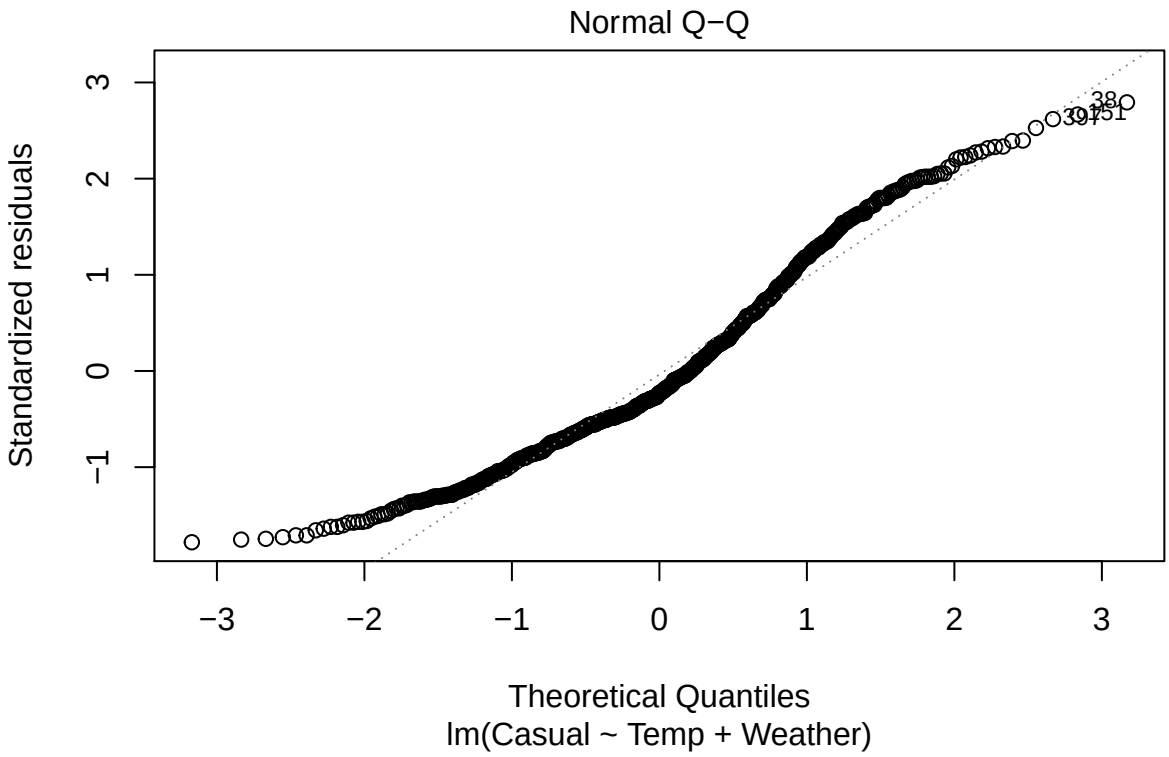
x2 = Weather

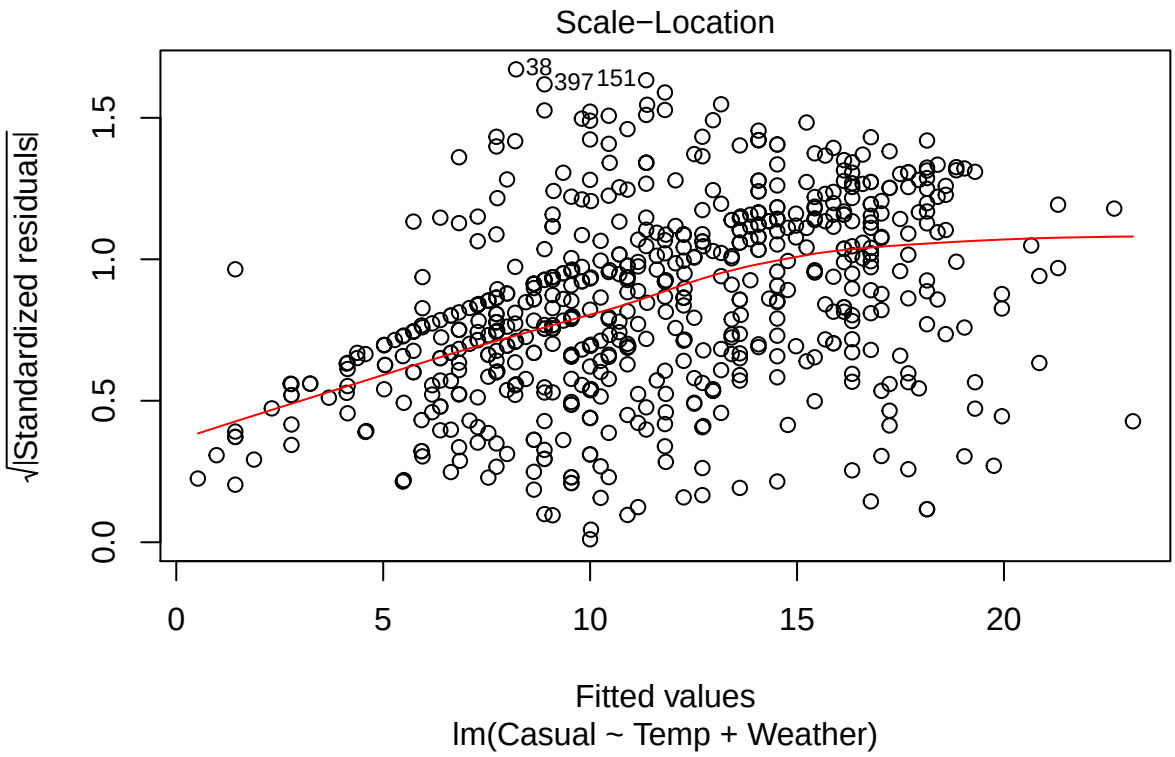
```
##             GVIF Df GVIF^(1/(2*Df))
## Temp    1.00416  1        1.002078
## Weather 1.00416  2        1.001038
```

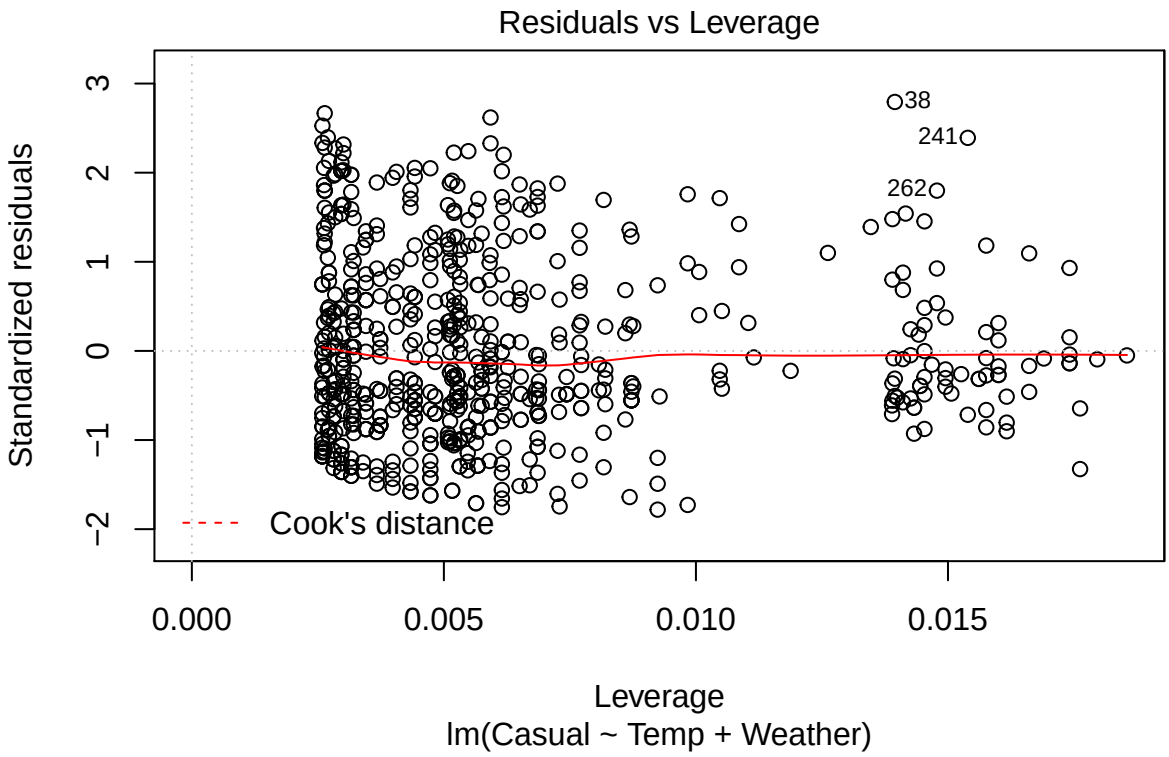
The VIFs are less than 2.5, suggesting there is NO dangerous multicollinearity.

Model #3 and diagnostics:




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Casual ~ Temp + Weather)
−3
−2
−1
0
1
2
3
38
39
751


Scale−Location
38
397
151
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
5
10
15
20
Fitted values
lm(Casual ~ Temp + Weather)




```
##
## 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
```

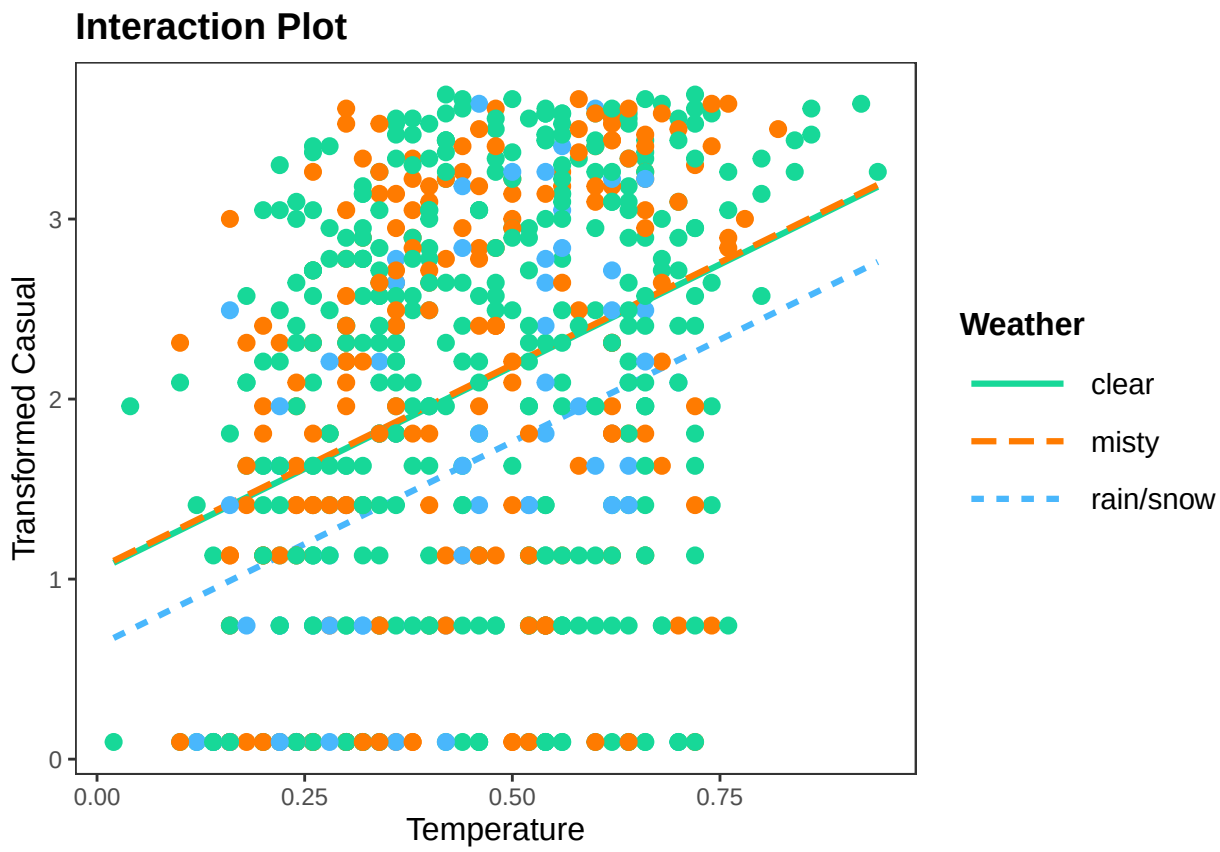
The diagnostics look eerily similar to Model #1. Errors look independent, have mean 0 and a constant sigma. The QQ plot shows that the errors are mostly normal. There are a few outliers, but aside from that, the diagnostics show the error assumptions to be true. Again, there are two beta values that are insignificant–the intercept and dummy variable 1 for the categorical variable. But since the other dummy variable is statistically significant ($0.00249 < 0.05$), the categorical variable is statistically significant. The R^2 is also a bit lower than Model #1 at 0.142.

Interaction Model?

Before I choose the model, I need to check if an interaction model would fit better for this study.

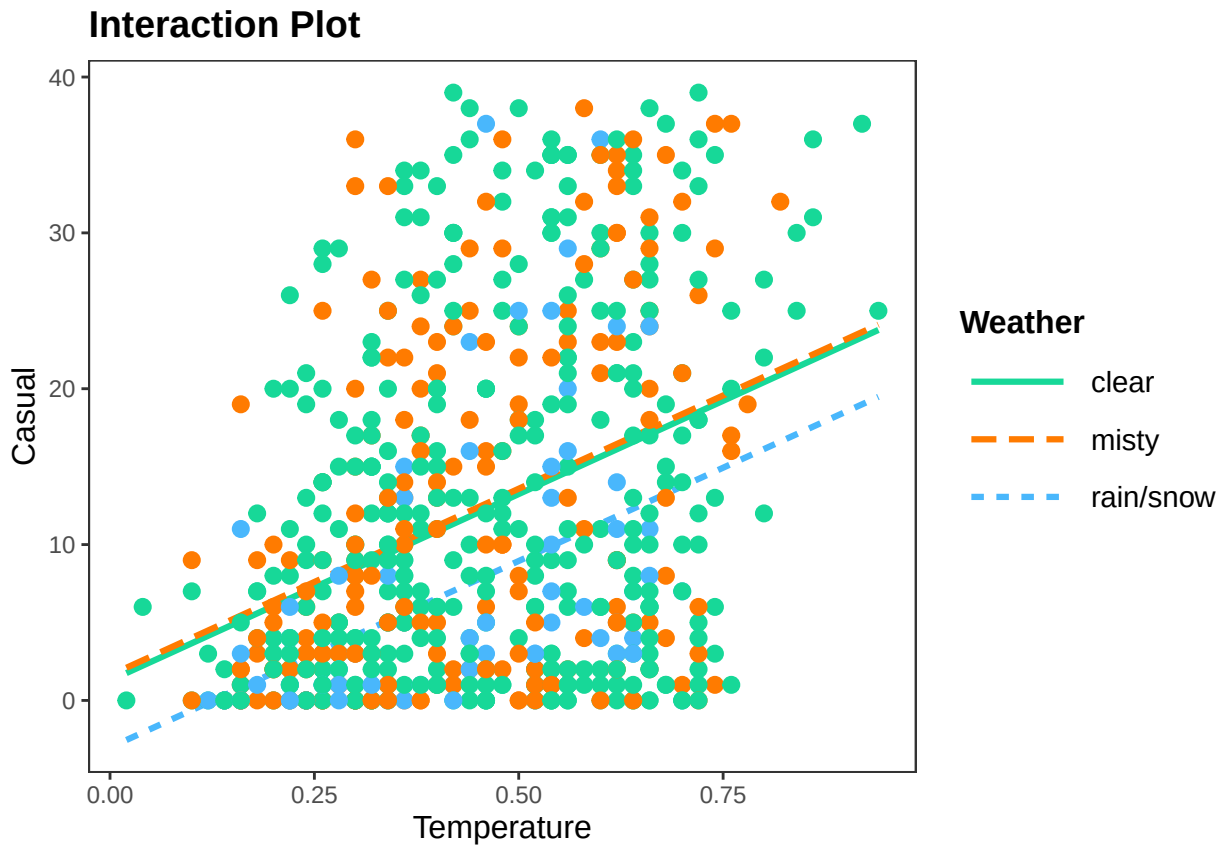
Here are the following interaction plots for all three proposed models

For Proposed Model 1:



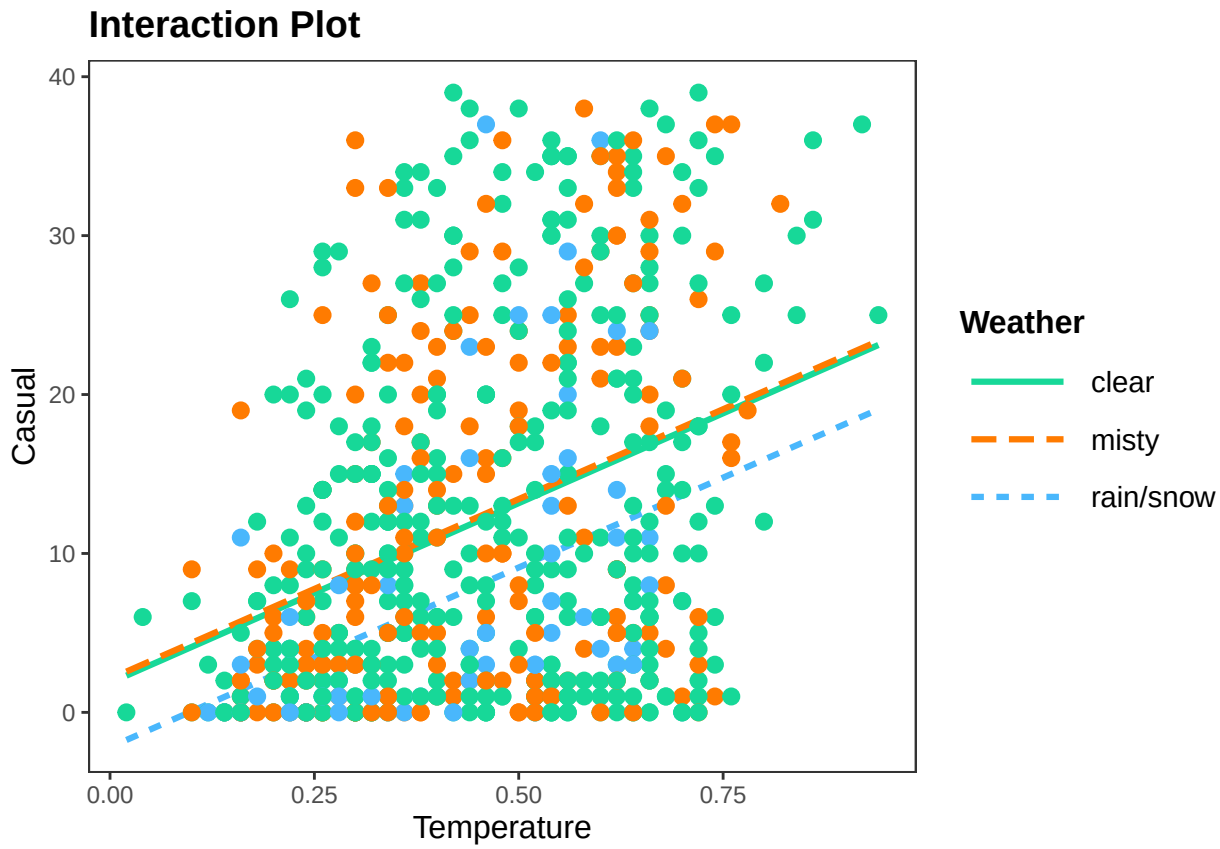
Since the slopes are approximately parallel, an interaction model is NOT needed

For Proposed Model #2:



Since the slopes are approximately parallel, an interaction model is NOT needed

For Proposed Model #3:



Since the slopes are approximately parallel, an interaction model is NOT needed

Decision

Analyzing the diagnostics, R^2 values, and the p-values, Model #1 seems to be the best fit for this data. It has the highest R^2 value and the best diagnostics. The only knock is the Intercept is not statistically significant ($0.27 > 0.05$), but we can overlook this because the significance of beta0 isn't relevant to whether there is a relationship or not. Overall, Model #1 is the best in diagnostics, the R^2 value, and p-values.

As a reminder, here is Model #1:

```
##
## Call:
## lm(formula = Casual ~ Temp + Weather + Windspeed, 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 ***
## Weathermisty       0.3417     0.8968   0.381  0.70333
## Weatherrain/snow  -4.2943     1.3154  -3.265  0.00115 **
## Windspeed         15.0479     3.3460   4.497 8.15e-06 ***
```

```
## ---
## 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
```

Additionally, we notice that all the beta values are not 0, suggesting a linear relationship, just like we predicted in the univariate exploration. F-test p-value is 2.2*10^-16, which is less than 0.05, meaning the model is statisticlaly signficant. The beta value for Windspeed is less than the beta value for Temperature, which corresponds with our EDA because we noticed that Temperature and Casual was a bit more positively correlated than Windspeed and Casual. This model's VIFs were all less than 2.5, so we can confidently say there isn't multicollinearity between the explanatory variables. Taking all of this into account, Model #1 explains that Weather, Temperature and Windspeed are moderately associated with biker turnout. Specifically, both Temperature and Windspeed have a direclty proportional relationship with Casual. If Temp goes up, so do bike users, and if Windspeed goes up, so do bike users. As weather gets worse, from clear/misty/cloudy to rain/snow, bike users decrease.

Prediction

Now that we have picked the best model, we want to predict Casual when Weather is misty/cloudy, Temperature is 0.75, and Windspeed is 0.25. Remember Temperature and Windspeed are both scaled (NOT Transformed), which is why their values are between 0 and 1.

Parameters for Model #1:

beta0 = -1.4947

beta1 = 23.9334

beta2 = 0.3417

beta3 = -4.2943

beta4 = 15.0479

Error = 10.23

Here is Model #1's regression equation:

```
-1.4947 + 23.9334(Temperature) + 0.3417(Weathermisty) -4.2943(Weatherrain/snow) + 15.0479(Windspeed)
```

[Weathermisty is 1 for misty, 0 for otherwise. Weatherrain/snow is 1 for rain/snow, 0 for otherwise clear weather is the baseline]

At a Temp of 0.75, misty weather, and windspeed of 0.25:

```
## [1] 20.55902
```

We predict that at a Scaled Temperature of 0.75, a scaled temperature of 0.25, and misty weather, there are 20.5 casual bike users. Looking back at the distribution of *Casual*, we see that this number is slightly above the 3rd quartile, so 20.5 bikers is a relatively high amount of people.

Discussion

In this study, we learned that the number of bike users is slightly related to Temperature, Weather, and Windspeed. We tried to transform both Casual and Windspeed to get better diagnostics, but it ended up giving worse diagnostics. In **Model #1**, which includes Casual, Temperature, Weather and Windspeed, all

error assumptions look good, but the Intercept is not statistically significant. There was also a noticeable difference in biker turnout when the weather changed. If it was rainy/snowy, on average, there were 4 less bikers.

One limitation of my model is that the R^2 value is low at 0.16. This suggests that maybe the explanatory variables in this study aren't that strongly correlated with the response variable. We also noted a few outliers from the residual plot. We should explore who these outliers were and why they used the bike rental system on those specific days.

Also, it would be interesting to see if maybe there is another characteristic that affects bike users. Perhaps it could be day of the week, since many people have more free time on the Weekends rather than the weekdays.

Overall, studies using statistics will be very useful in understanding mobility in the city. It could help the digital bike sharing company on what their next steps are. Or it could even help the taxi or bus industry because they might need to know what days they should expect more passengers (on days where theres less biker turnout).