Pedaling Patterns and Factors Influencing Casual Bike Ridership

Joseph Ouyang josephou

Due Wed, October 22, at 11:59PM

Contents

Introduction

Transportation networks represent a significant economic sector, shaping how cities function and affecting both ecological outcomes and community well-being. The rise of bike sharing programs has provided cities with a practical tool for addressing congestion and pollution, now operating in hundreds of locations globally with substantial ridership. As metropolitan areas continue to expand and face mounting challenges related to mobility and environmental quality, analytical approaches have become essential for managing these transportation alternatives effectively. In the present paper, we focus on casual bike sharing ridership patterns in the Washington D.C./Arlington, VA/MD area, and determine whether there are any specific factors that contribute to hourly usage among casual riders.

Exploratory Data Analysis

Data

Referencing Capital Bikeshare's casual bike usage data, we analyze a random sample of 656 hours and 4 variables. After that, we analyze the relationship between causal bike usage (our response variable) with the three predictor variables (weather, temperature, and windspeed). Summaries of the variables are listed below:

`Casual`: number of bike users (measured by the number of bikers in an hour block)

`Weather`: type of weather (falling into three categories of clear, misty, and rain/snow)

`Temperature`: temperature, (scaled as percentage of the dataset's overall maximum, with numbers closer to 1 being warmer)

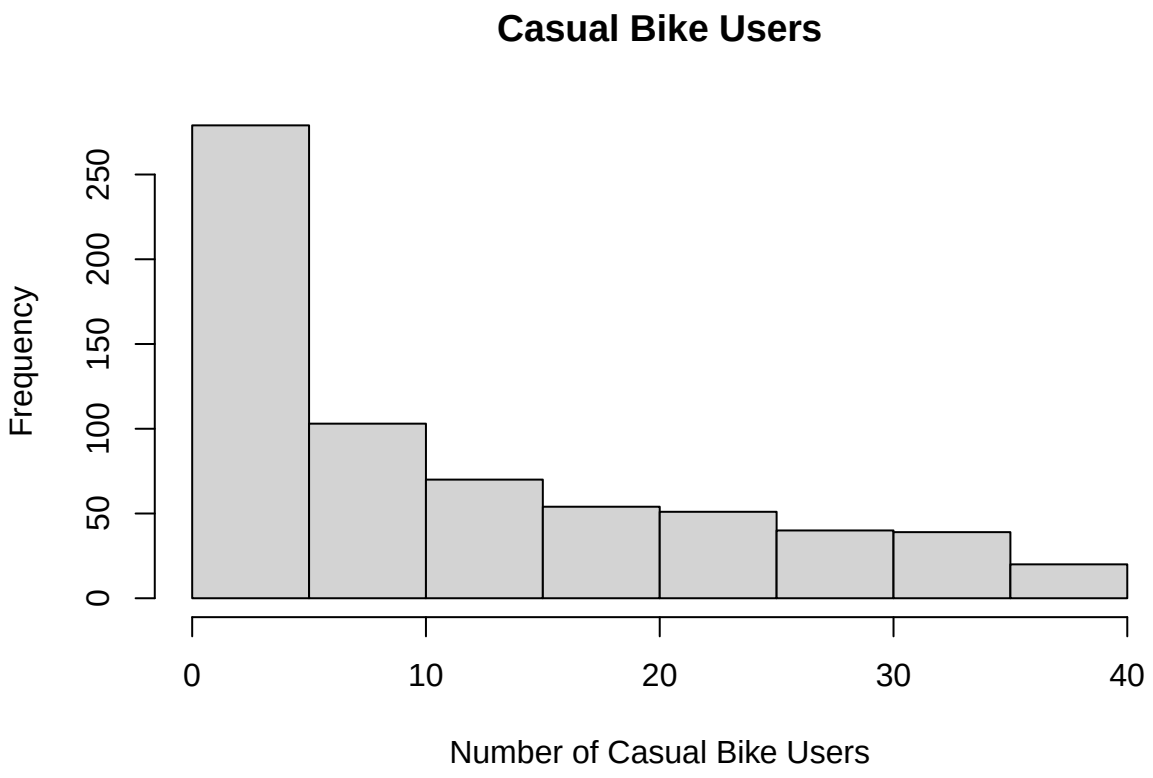
`Windspeed`: windspeed, (scaled as percentage of the dataset's overall maximum, with numbers closer to 1 being windier)

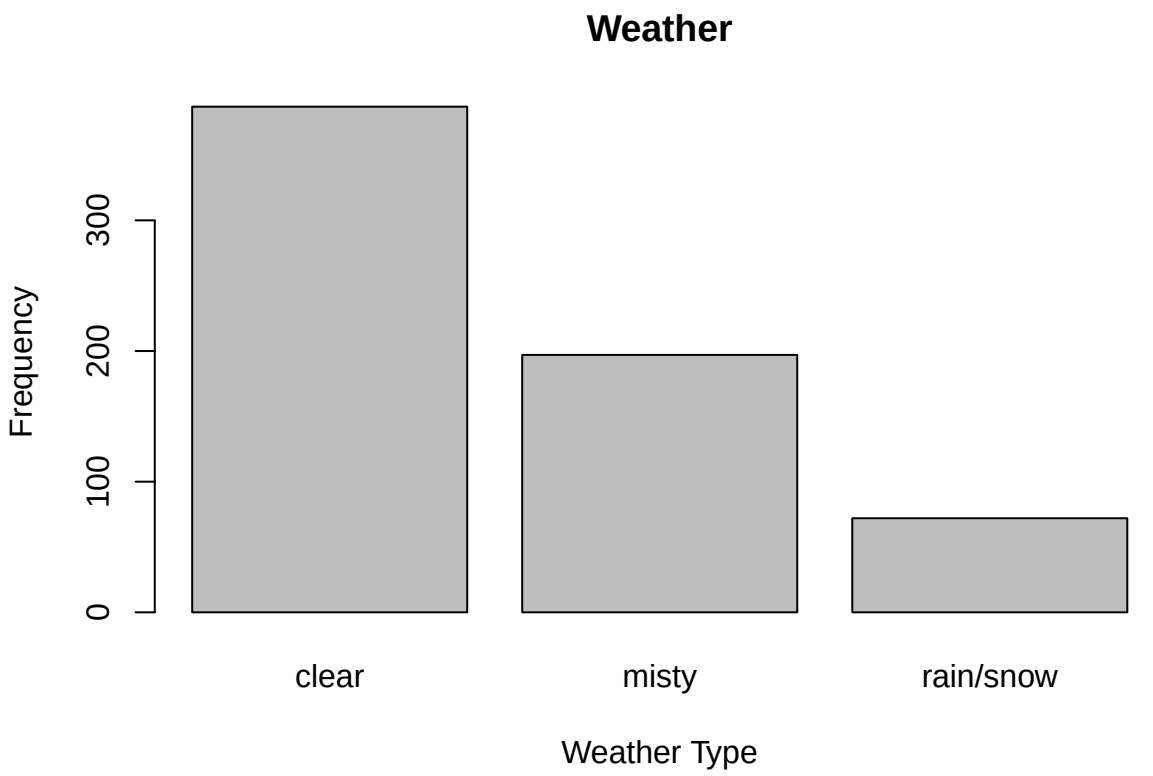
Below are the first few lines from the data set:

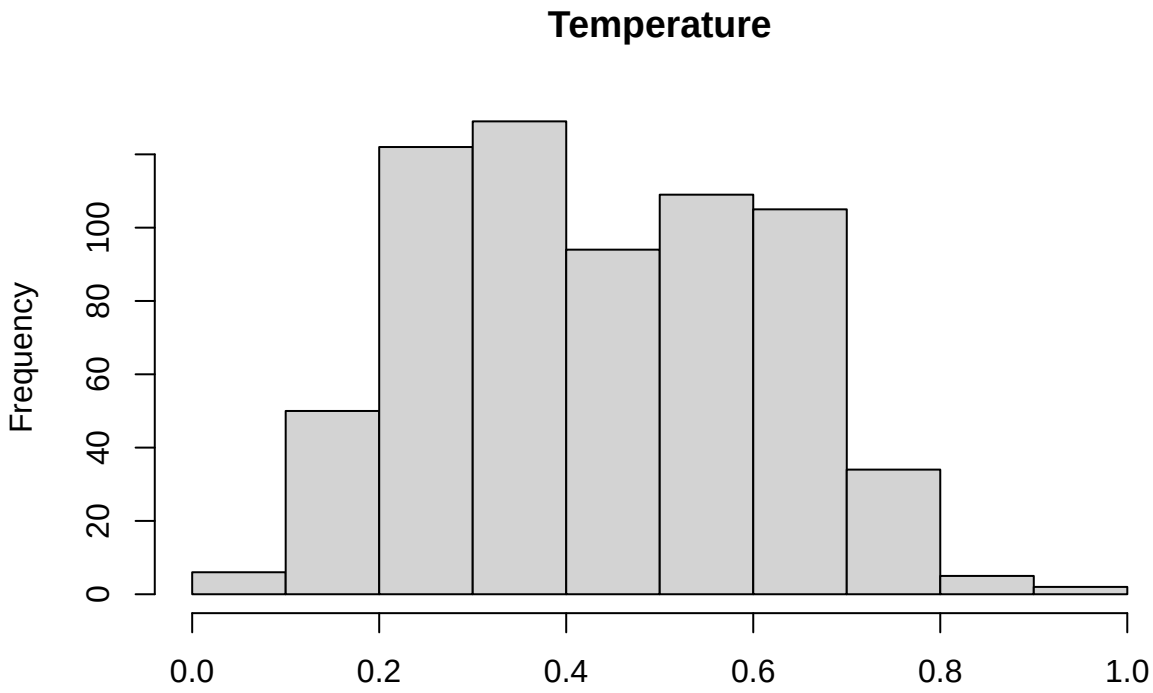
```
## # A tibble: 6 x 4
##   Casual Weather    Temp Windspeed
##    <dbl> <chr>     <dbl>     <dbl>
## 1      5 rain/snow  0.34     0.388
## 2      9 clear      0.34     0.104
## 3      6 misty      0.46     0.224
## 4     25 clear      0.34     0.298
## 5     31 clear      0.54     0.134
## 6     15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

Univariate EDA

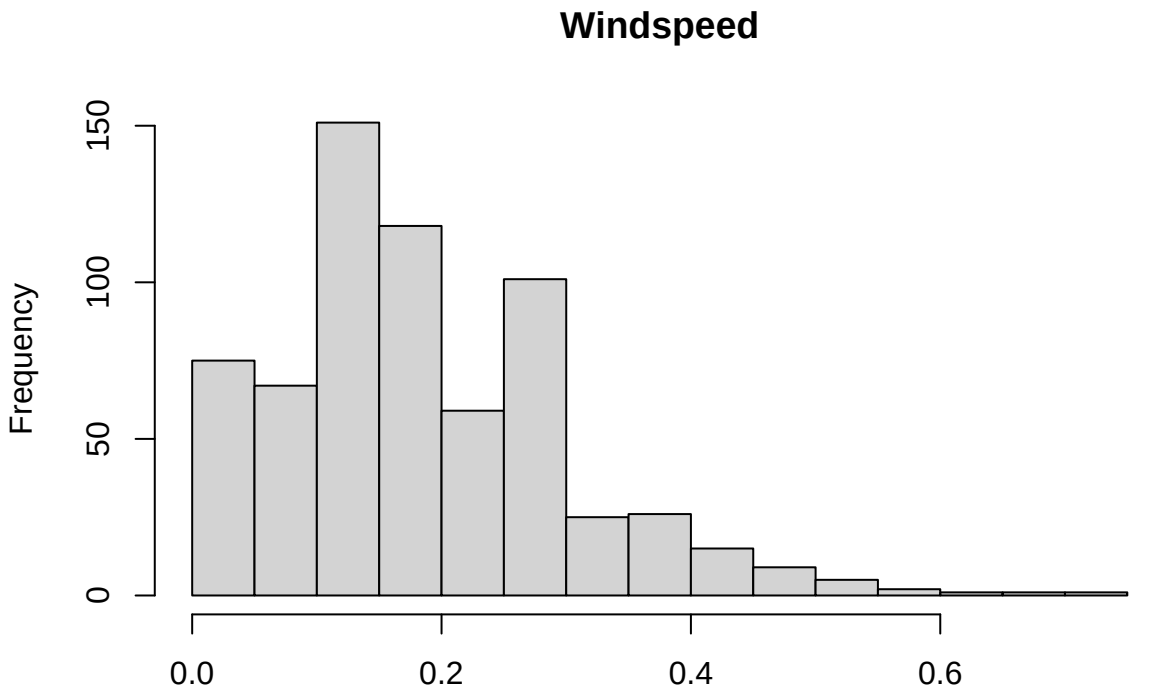
First, we will analyze the response variable and all the predictor variables individually. Our methodology will be to use a histogram for the continuous variables (Casual, Temperature, and Windspeed) and bar graph for our categorical variable (Weather).


Casual Bike Users
Frequency
0
50
100
150
200
250
0
10
20
30
40
Number of Casual Bike Users


Weather
Frequency
0
100
200
300
clear
misty
rain/snow
Weather Type



Temperature (scaled as percentage of dataset's overall maximum, from 0 to 1)



Windspeed (scaled as percentage of dataset's overall maximum, from 0 to 1)

Additionally, we include numerical summaries of each variable listed below:

For `Casual`:

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

For `Weather`:

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

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
```

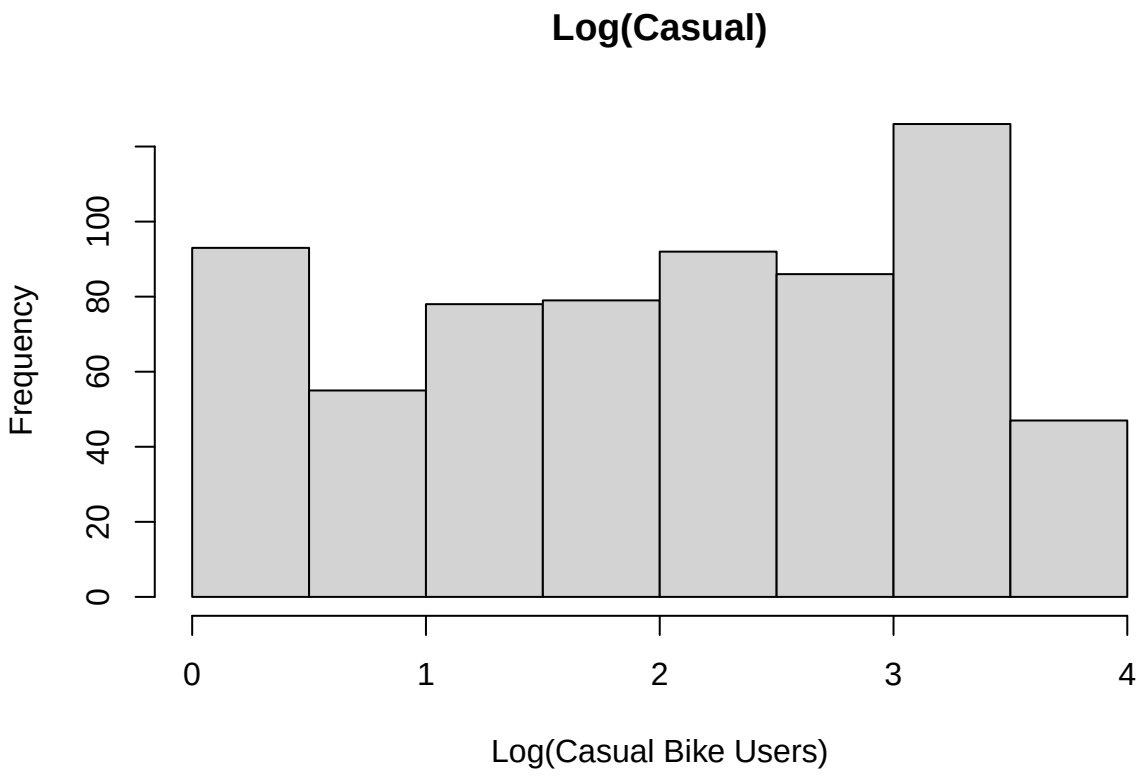
Looking at the plotted graphs and numerical summaries above, we can make several comments about our variables. The distribution of **Casual** Bike Users is right skewed and unimodal, with the mean being greater than the median. The distribution has a peak of about 275 hour blocks at 0 to 5 casual bike users, is centered around 8 casual bike users, and ranges from 0 to 39 casual bike users. For **Weather**, we observe 387 clear hours, 197 misty hours, and 72 rainy/snowy hours. The **Temperature** distribution appears to be unimodal and roughly symmetric, with the mean and median being very similar. The distribution is centered around 0.44, or 44% of the dataset's overall maximum temperature and has an interquartile range of 0.285. The **Windspeed** distribution appears to be right skewed and unimodal, but there also exists a smaller peak at around 0.3 or 30% of the dataset's overall maximum windspeed. Overall, the distribution is centered around 0.1642 with windspeed percentages ranging from 0 to 0.7164.

Bivariate EDA

Before we continue with Bivariate EDA, we will first transform some of the variables, namely **Casual** and **Windspeed**, which are both right skewed. We will apply log transformations to both.

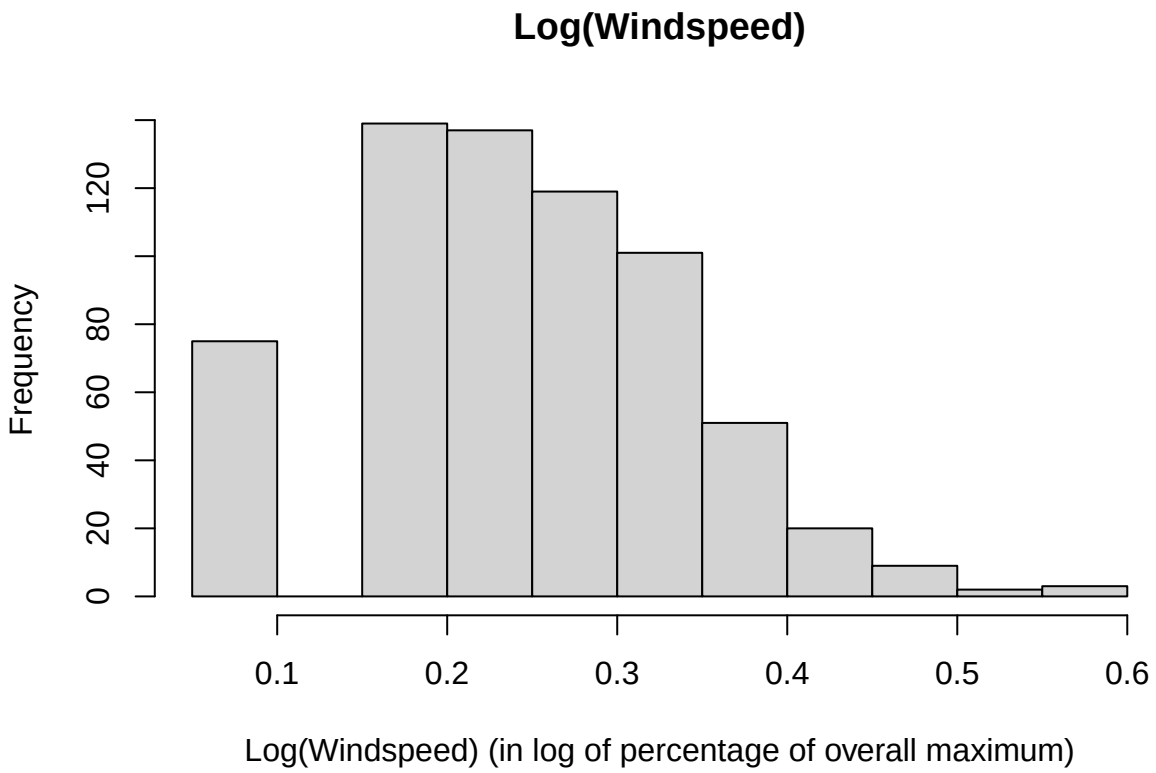
For `Casual`:

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

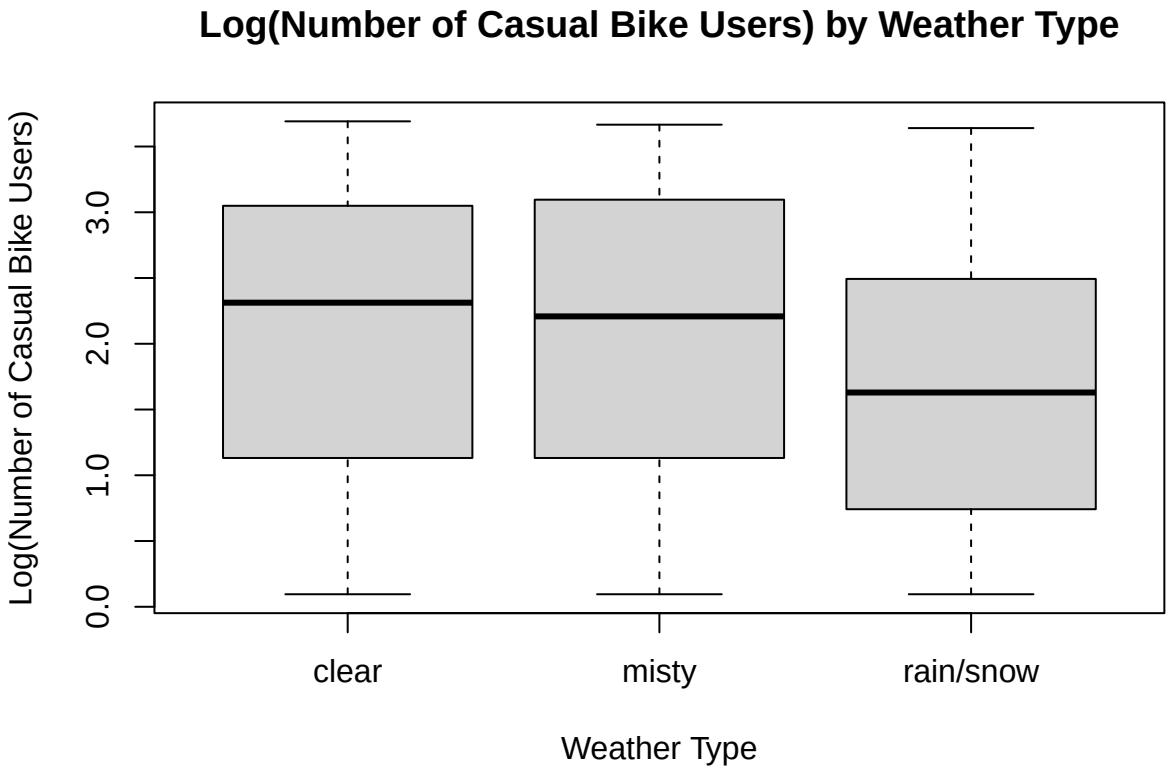


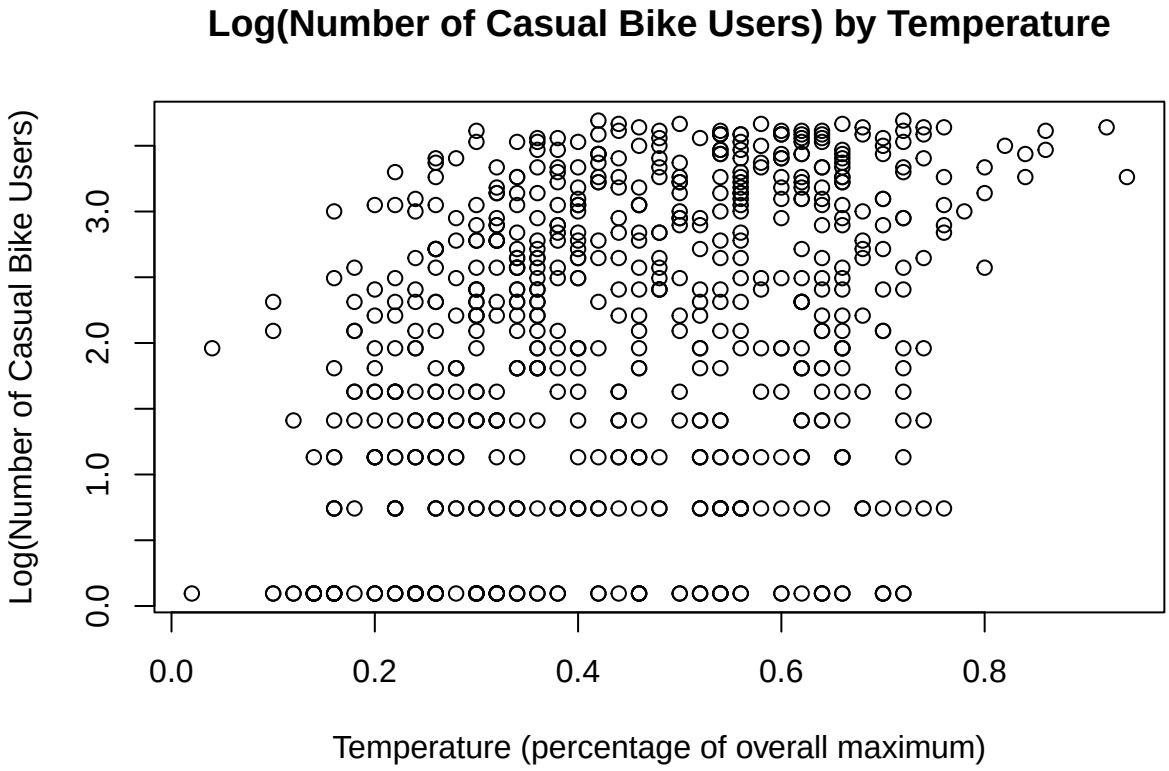
For `Windspeed`:

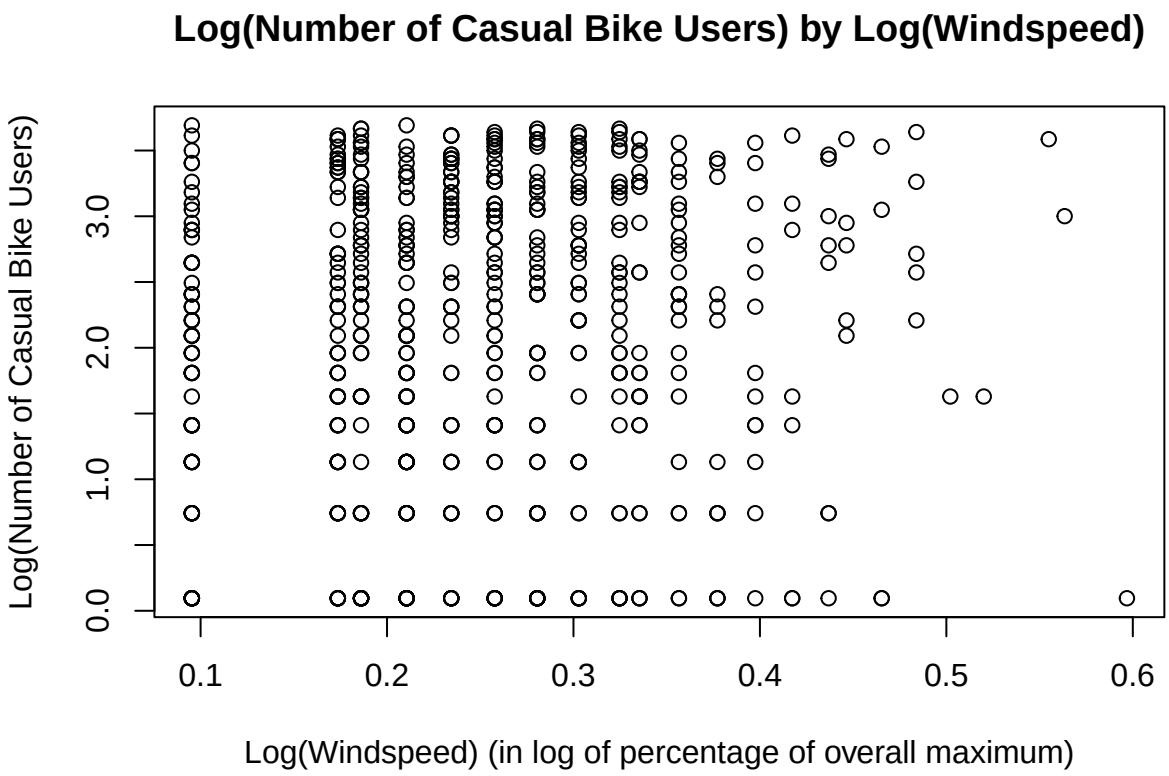
```
bikes$log.windspeed <- log(bikes$Windspeed + 1.1)
```



After applying the log transformations, the histogram of log(Casual) appears approximately normal, indicating that the transformation successfully stabilized variance and reduced skewness. In contrast, log(Windspeed) still shows a slight right skew, though it is noticeably weaker than before. Despite this remaining skew, the transformation improves model fit and better satisfies the assumptions of linear regression. A small gap between 0.1 and 0.2 on the log(Windspeed) histogram is visible, but this is simply an artifact of the dataset, reflecting fewer observations in that range rather than any underlying issue with the transformation. Overall, these refinements allow us to confidently proceed to bivariate EDA, examining how each predictor variable relates to the response variable, Casual, as presented on the following page:


Log(Number of Casual Bike Users) by Weather Type
Log(Number of Casual Bike Users)
0.0
1.0
2.0
3.0
clear
misty
rain/snow
Weather Type


Log(Number of Casual Bike Users) by Temperature
Log(Number of Casual Bike Users)
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
Temperature (percentage of overall maximum)



Referencing the graphs above, we observe that the median of `log(Casual)` is lower for rainy/snowy `Weather`, when compared to the medians of clear and misty weather, which have similar higher numbers of casual biker users. To generalize, when the weather gets worse, the number of casual bike users decreases. The log of the number of casual bike users seems to be positively associated with `Temp`. As temperature increases, the number of casual bike users also increases. The log of the number of casual users does not seem to be related to the log of `Windspeed`, as we observe no clear pattern to the data. The points on the scatterplot seem to be randomly distributed with no clear upward or downward trend, which suggests no linear relationship at worst, or a very weak relationship at best.

Modeling

After conducting univariate and bivariate EDA to visualize the relationships among our variables, we will now construct our linear regression model to predict the number of casual bike users.

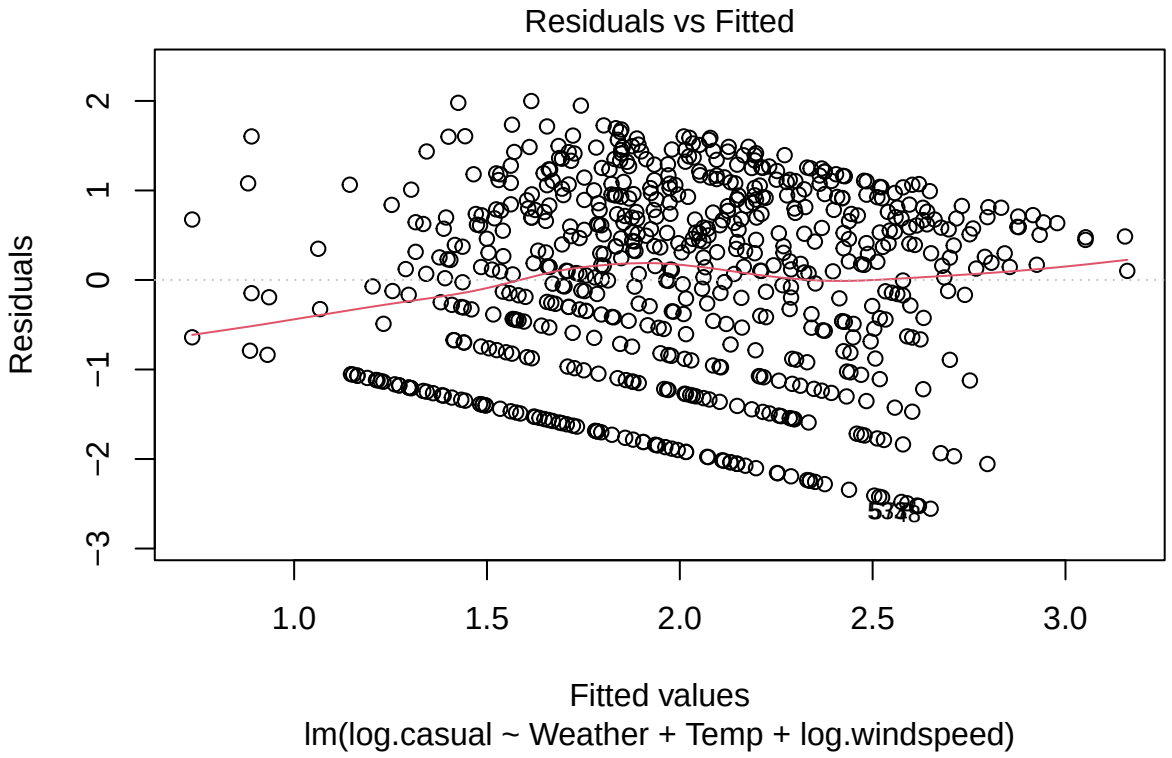
At first glance, there appears to be no clear relationship between the log of the number of casual bike users and the log of windspeed. Upon further analysis, however, there appears to be a very weak positive linear relationship between the two variables. After analyzing the regression summary, `log(Windspeed)` was determined to be statistically significant at the 5% level and meaningfully improved the adjusted R^2 when combined with other predictors. Therefore, `log(Windspeed)` was kept in the new regression model. `Temp` exhibited a significant positive relationship with the log of the number of casual bike users according to both the EDA and regression summary, so it was also included in the new regression model.

Since the quantitative variables `Temp` and `log(Windspeed)` are in the new regression model, we will check for multicollinearity. First we examine the variance inflation factors (vifs) of each variable in the new regression model that predicts the response variable `Casual`.

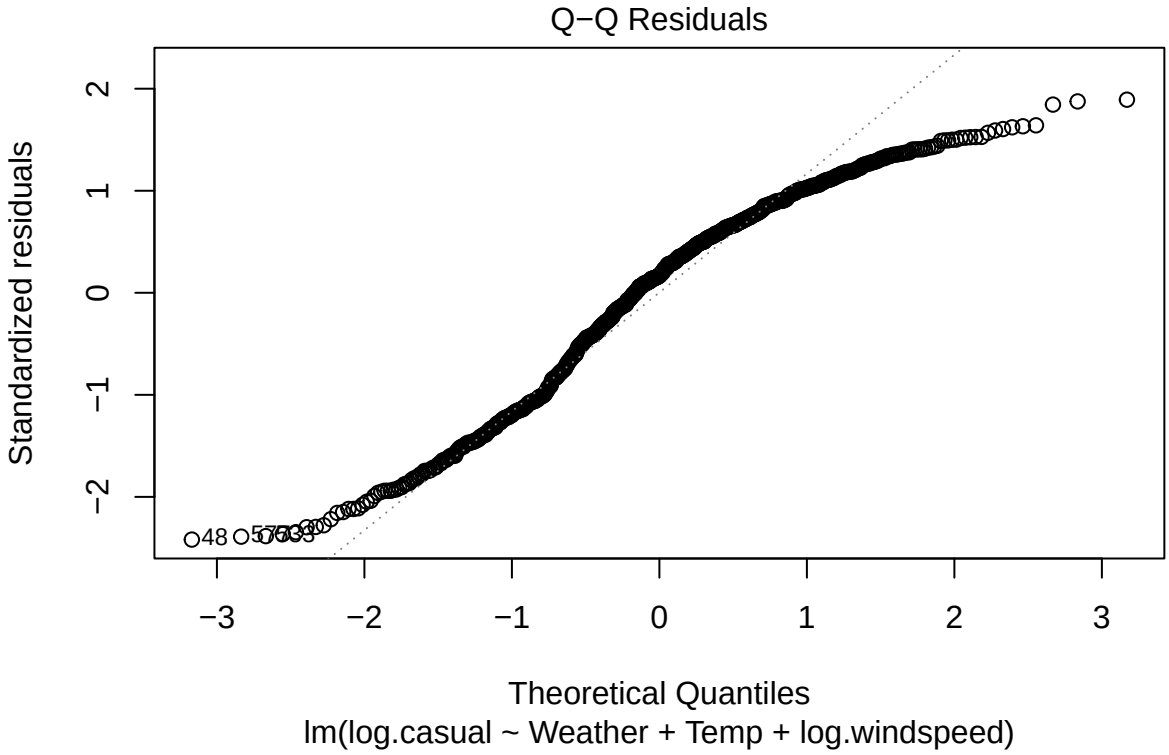
```
##                   GVIF Df GVIF^(1/(2*Df))
## Weather       1.007123  2        1.001776
## Temp          1.019726  1        1.009815
## log.windspeed 1.018617  1        1.009266
```

The vifs of `Weather`, `Temp`, and `log(Windspeed)` are 1.007123, 1.019726, and 1.018617, respectively. These vif values are acceptable because they are all below 2.5. Hence, we can move forward with the model because there are no significant multicollinearity issues.

From our new multiple linear regression model predicting Number of Casual Users from `Weather`, `Temp`, and `log(Windspeed)`, we now examine the residual plot and qq plot below:



On the residual plot, we examine that there is no clear pattern, aside from a slight downward slide of the lower residuals. Among all the models tested, this residual plot appeared the most stable, as the others displayed similar patterns without showing any clearer improvement. Overall, the independence assumption is reasonably satisfied. Since there is generally constant spread above and below the zero line as we scan left to right, the constant spread assumption is reasonably satisfied. Looking from left to right, the mean of the residuals seems to change from slightly positive to slightly negative. However, when we take into consideration the whole residual plot, the mean of the residuals is roughly centered around zero. Therefore, the mean zero assumption is reasonably satisfied.



On the qq plot, we observe that the residuals mostly follow the diagonal line, aside from some minor deviations at the tails. Small departures from normality at the tails, however, do not invalidate the normality assumption because this data comes from a large sample. With the assumptions from the residual plot and qq plot being reasonably satisfied, there is no need to try other transformations. Therefore, we can proceed with this multiple linear regression model. The regression analysis summary from this final model is shown below:

```
##
## Call:
## lm(formula = log.casual ~ Weather + Temp + log.windspeed, 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 ***
## Weathermisty      0.01105    0.09293   0.119 0.905365
## Weatherrain/snow -0.41753    0.13630  -3.063 0.002279 **
## Temp              2.27067    0.23822   9.532  < 2e-16 ***
## log.windspeed     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
```

Justification

We consider this a reasonable model to predict `Casual` since the red line of the residual plot is relatively smooth, meaning the linearity condition is reasonably satisfied. Some other models we constructed included interaction terms between `Weather` and `Temp`, and between `Weather` and `log(Windspeed)`. Although adding interaction terms slightly increased the adjusted R^2 value (by about 1%), the improvement did not justify a model switch because all of the interaction terms were not statistically significant at the 5% level.

Additionally, we constructed a quadratic model with an additional `(Temp)^2` term. While the quadratic model increased the adjusted R^2 and more accurately reflected the observed curvature between temperature and ridership, including this term did not reflect the client's request for a **linear** regression model.

Analysis

The regression F-test p-value is about 2.2x10^(-16) which is less than 0.05, implying statistical significance. With the exception of misty weather showing no statistically significant difference in number of casual bike users when compared to the baseline clear weather, all the other predictor coefficients are significant in the model with p-values less than 0.05.

We observe a negative coefficient value (-0.41753) for the categorical dummy variable `Weatherrain/snow` which is consistent with the side-by-side box plots, showing that rainy/snowy weather, on average, correlates to lower numbers of casual bike users. The positive `Temp` coefficient (2.27067) is consistent with the EDA because the log of the number of casual bike users is shown to increase when temperature increases. The positive `log(Windspeed)` coefficient (1.69646) is also consistent with the EDA and regression summary because the log of the number of casual bike users is shown to increase when the log of windspeed increases.

Conclusion

In summary, we determine this multiple linear regression model to be the best fit because of the following reasons: the signs of the coefficients reflect the EDA, most of the coefficients are themselves statistically significant with p-values less than 0.05, and this model's R^2 value of 0.1429 is as high as it can get when balancing explanatory power with model simplicity. The relatively low R^2 value of 0.1429 is expected for behavioral data like bike-share usage, since ridership is influenced by many unobserved factors such as time of day, day of week, proximity to stations, and personal or social motivations that extend beyond weather and temperature. In conclusion, we are confident that clear and misty weather, higher temperatures, and faster windspeeds are associated with higher numbers of casual bike users in a given hour.

Prediction

With our newly constructed multiple linear regression model, we can now predict the number of casual bike users for an hour with misty/cloudy weather, a scaled temperature of 0.75, and a scaled windspeed of 0.25.

The predicted number of casual bike users is shown below:

```
0.62812 + 0.01105 + 2.27067*(0.75) + 1.69646*log(0.25 + 1.1)
```

```
## [1] 2.851288
```

Since this predicted value is under the log transformation, we will exponentiate to calculate the actual predicted number of casual bike users under the specified conditions above.

```
exp(2.851288) - 1.1
```

```
## [1] 16.21006
```

The predicted number of casual bike users for an hour with misty/cloudy weather, a scaled temperature of 0.75, and a scaled windspeed of 0.25 is **16.21006**. In context, 16.21006 casual bike users sits roughly between the mean (11.51) and Q3 (20.00) of the response variable's (number of casual bike users) distribution.

Discussion

Through this statistical analysis we learned that the number of casual bike users is related to the type of weather, temperature, and windspeed. In a given hour with clear or misty weather, relatively high temperatures, and relatively fast windspeeds, the number of casual bike users is higher, on average.

One possible source of concern is the low outliers displayed on the residual plot. Those outliers may represent unusually low ridership hours that the model fails to capture, such as during extreme weather conditions, late-night periods, or system maintenance times not reflected in the dataset. Another possible area of concern is the subtle downward slide and growing spread of the residuals as we scan left to right. This pattern suggests that the relationship between temperature and ridership could be more complex or even non-linear, potentially leveling off or reversing more sharply during very hot hours.

The similarity in number of casual bike users for clear weather and misty weather is interesting and would warrant further investigation, because one would think that misty weather would decrease ridership in a given time period as bikers would not be able to see clearly. Furthermore, the possible non-linear relationship between temperature and number of casual bike users could be an area for further investigation. Statisticians could determine the approximate temperature when the rate of ridership increase starts to significantly drop off, which could provide practical insights relating to budgetary saving for bike-share companies.

Overall, this study highlights how environmental conditions shape short-term transportation behavior and demonstrates the value of data-driven modeling in urban planning. By understanding how temperature, weather, and windspeed influence ridership, cities can optimize bike availability and maintenance schedules to match demand. Ultimately, incorporating such analytical insights can help promote more efficient, sustainable, and user-responsive transportation systems.