The Prediction of Casual Bikes Users with Different Factors

XiangCheng Xu xiangchx

Due Mon, July 24, at 11:59PM

Contents

Introduction

In nowadays society, the existence of sharing bikes has greatly changed the way of people's attitude towards ways of traveling, Especially for those who prefer experiencing the nature. Sharing bikes have already become one of the most widely chose ways of public transportation when people are having short-distance tours. Due to people's increasing interest in this new, fashionable way of traveling, this paper will focus on the number of user of sharing bike in the region at Washington D.C./Arlington and VA/MD area; and determine whether there are any possible factors that affect the number of users of sharing bikes.

Exploratory Data Analysis

Data

In Washington D.C./Arlington and VA/MD area, a random sample survey with 656 participants was created, and was used as the data used for analyzing the possible association between the number of users of sharing bike and other variables. To be more specific, in this paper, it focus on the relationship between the number of users, the temperature of the day,the wind speed of the day, and the weather of the day, which can be summarized as following variables:

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

And here are some data in the following:

```
head(bikes.initial)
```

```
## # 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

Before creating a model for predicting the number of users by using data shown above, it is necessary to explore attributes fo each individual. Therefore, this paper will firstly show the summary for each variable

Response Variable

For Number of Casual User

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

```
## [1]  5  9  6 25 31 15
```

Explanatory Variables

For Weather

```
##    Length     Class      Mode
##       656 character character
```

```
## [1] "rain/snow" "clear"     "misty"     "clear"     "clear"     "clear"
```

For Temperature

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

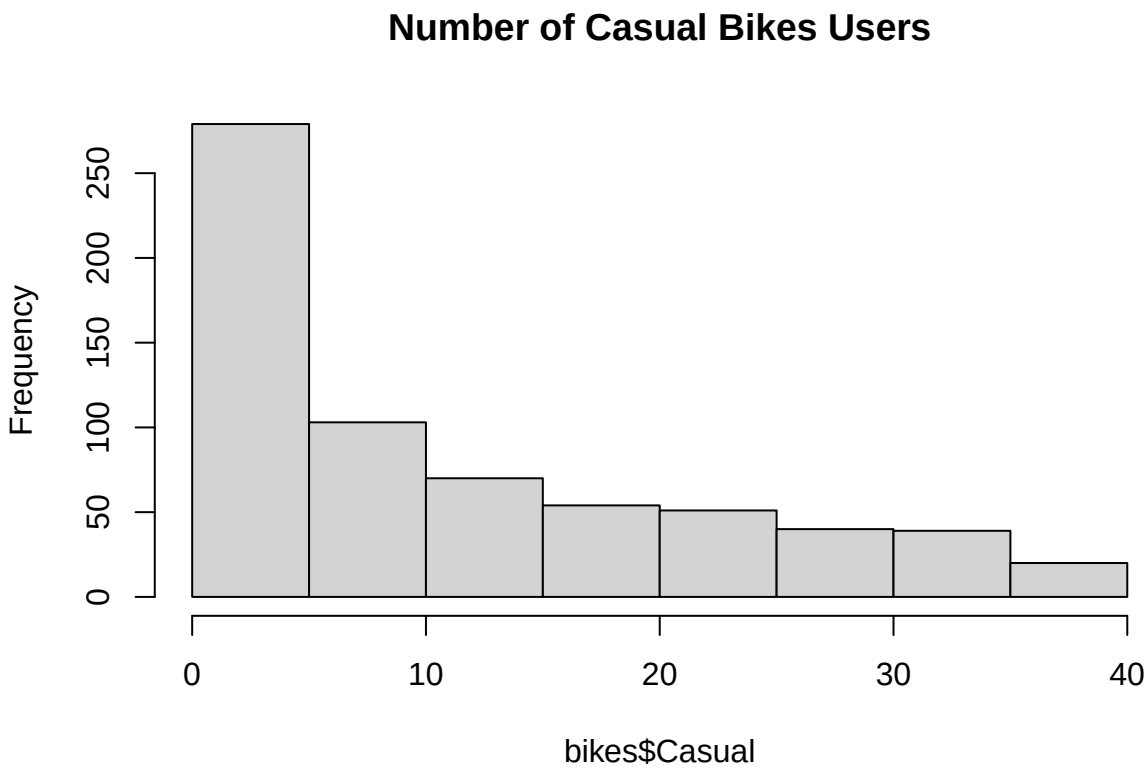
```
## [1] 0.34 0.34 0.46 0.34 0.54 0.32
```

For Wind Speed

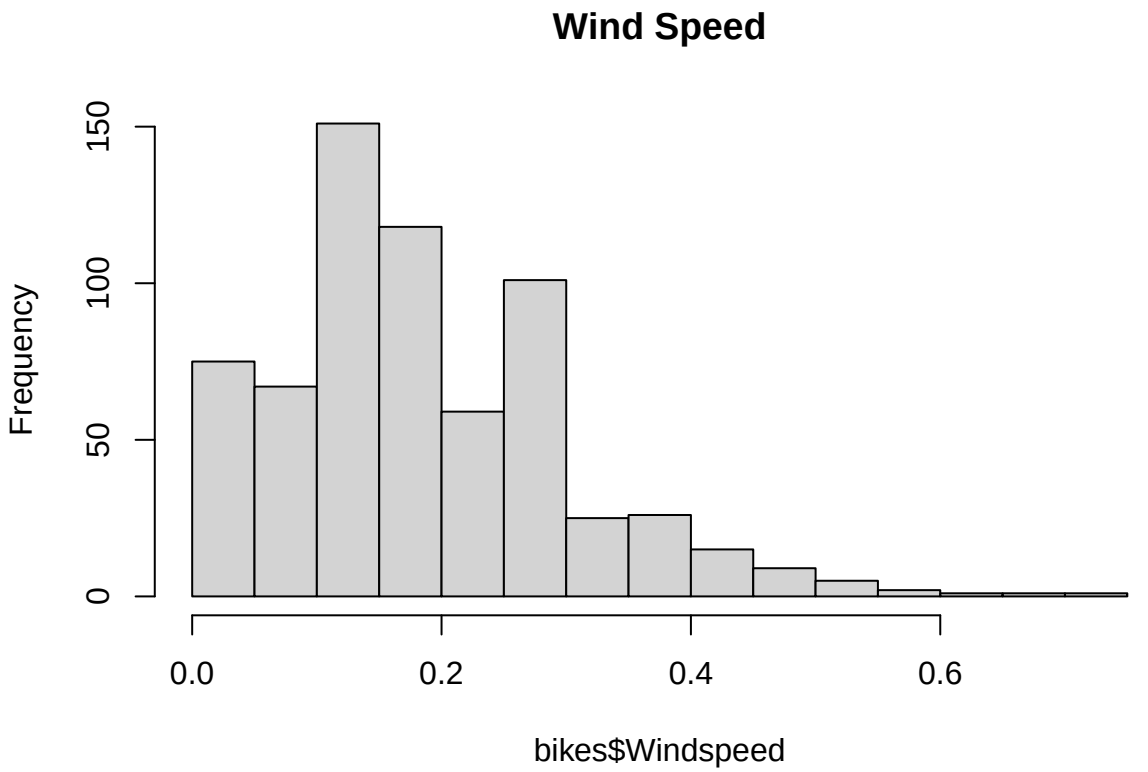
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1045  0.1642  0.1840  0.2537  0.7164
```

```
## [1] 0.3881 0.1045 0.2239 0.2985 0.1343 0.2537
```

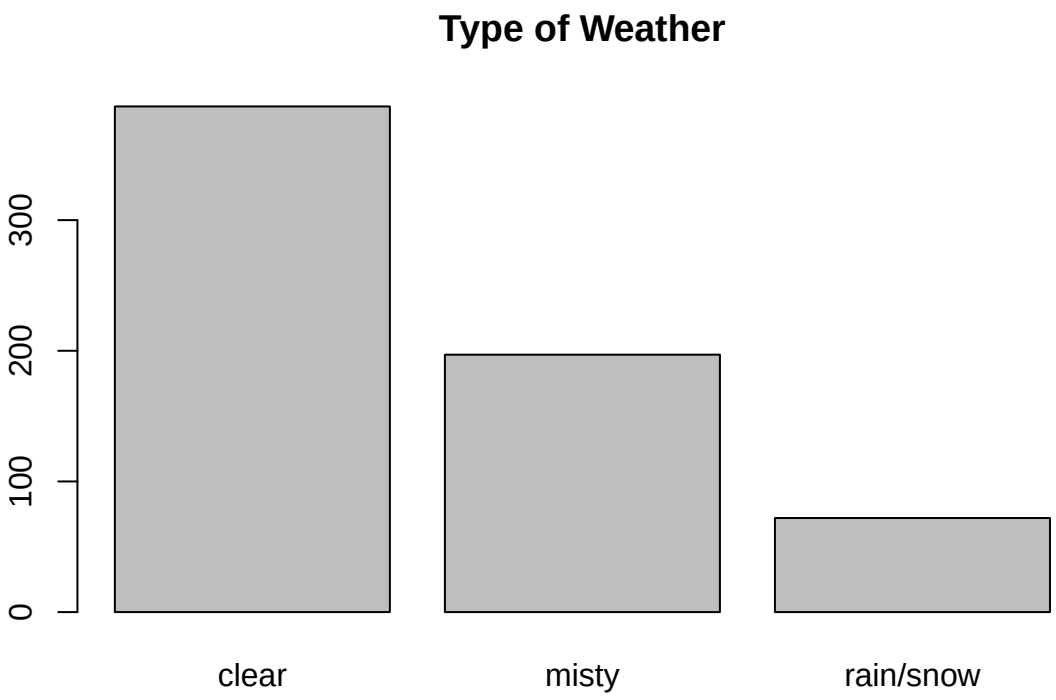
Then it is important to examine about the distribution of each explanatory variable by presenting histograms for quantitative variables and a boxplot for categorical variable.




```
hist(bikes$Windspeed, main = "Wind Speed")
```

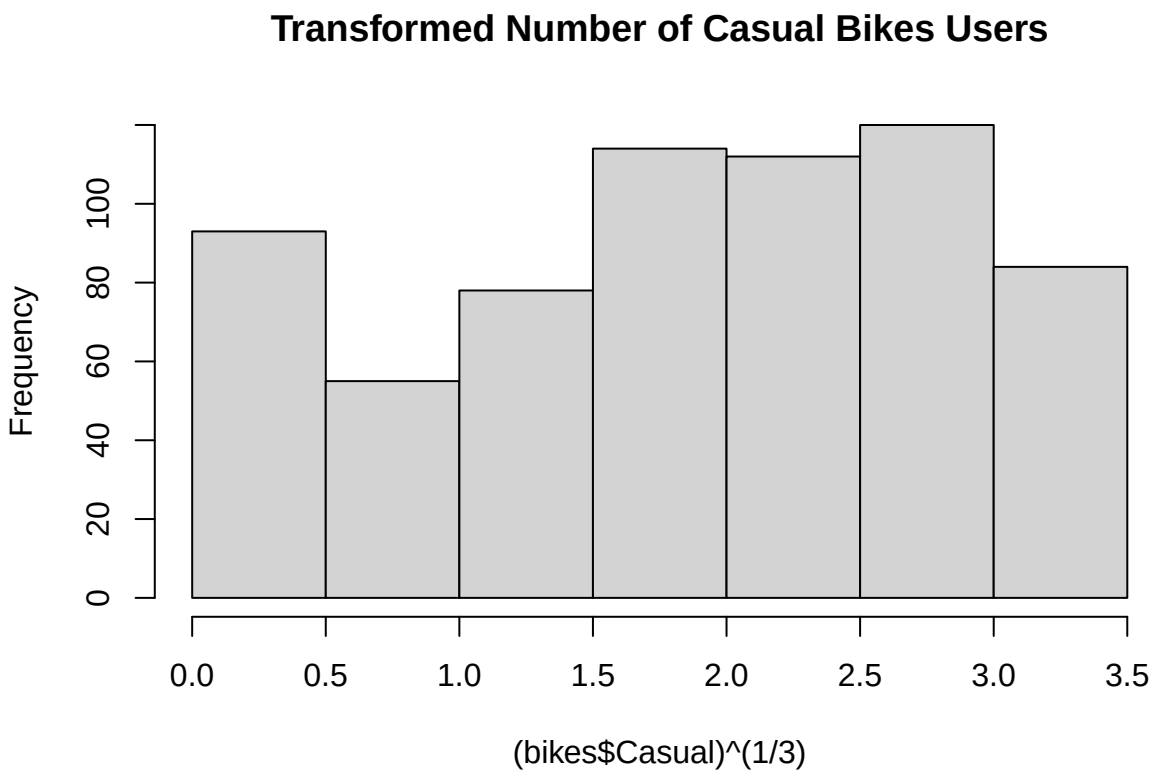



```
barplot(table(bikes$Weather),
        main = "Type of Weather")
```

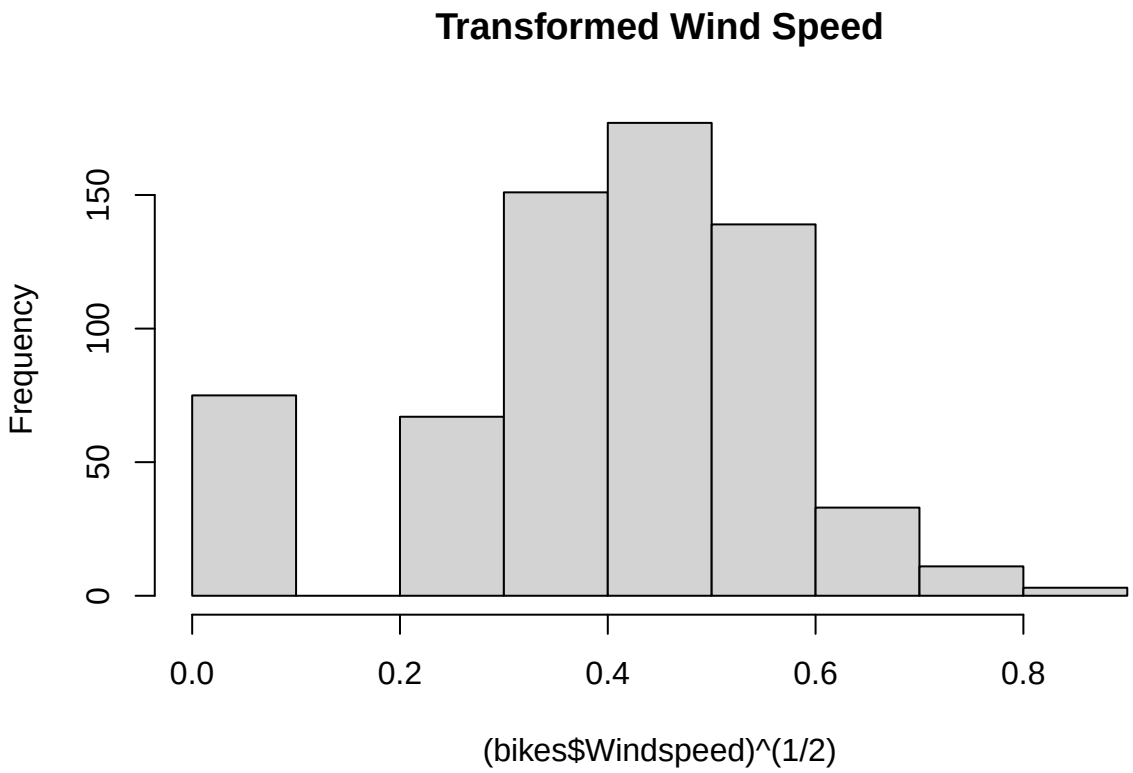


According to graph, it is not hard to see that histograms for 'Casual' and 'Windspeed' contain quite high skewness which can influence the result when computing the final model. Therefore it is necessary to do some transformation to 'Casual' and 'Windspeed' to increase the normality of these two variables. After trying transformed these two variables with different coefficient, the following code will show how do these two variables change:

```
hist((bikes$Casual)^(1/3), main = "Transformed Number of Casual Bikes Users")
```



```
hist((bikes$Windspeed)^(1/2), main = "Transformed Wind Speed")
```



Since these histograms looks quite normally distributed, therefore, it is reasonable to save the transformed data by using the following code and use it for computing the final required model.

```
#Saving the transformed data into the data frame --- "bikes"
bikes$transCasual <- (bikes$Casual)^(1/3)
bikes$transWindspeed <- (bikes$Windspeed)^(1/2)
```

The summary of adjusted/transformed variable is shown, as follows:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.260   2.000   1.856   2.714   3.391
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3233  0.4052  0.3916  0.5037  0.8464
```

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

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

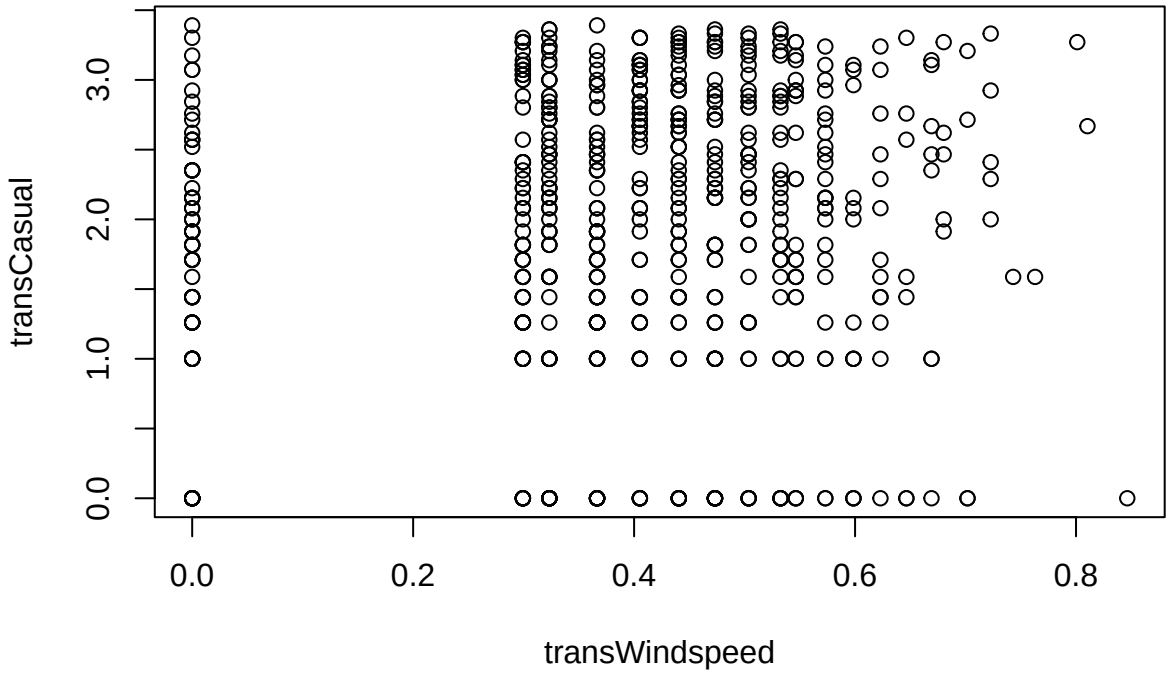
By considering the summarized data presented by R, the mean and the median of each variable are very similar, which proves that all variables have been transformed to have high normality. And according to the summarized data and graph, the distribution of **transformed Casaul** is unimodel, but not very symmetric. If we want a more accurate or a better distributed data, it is necessary to obtain more data. For the mean and median of transformed Causal, its mean is 1.856, and median is 2.000, which are close to each other. Distributions of both the **temperature** and the **transformed wind speed** are both unimodel and symmetric ,which can also seen as with great normality. For the temperature, it has the mean of 0.44 and the median of 0.442, which is very close; meanwhile, for the transformed wind speed, it also have quite similar mean and median, which are 0.4052 and 0.3916. Last but not the least, for the categorical variable, **Weather**,

we can see from the summary of data that 387 of them are clear, 197 of them are misty, and 72 of them are rain or snow.

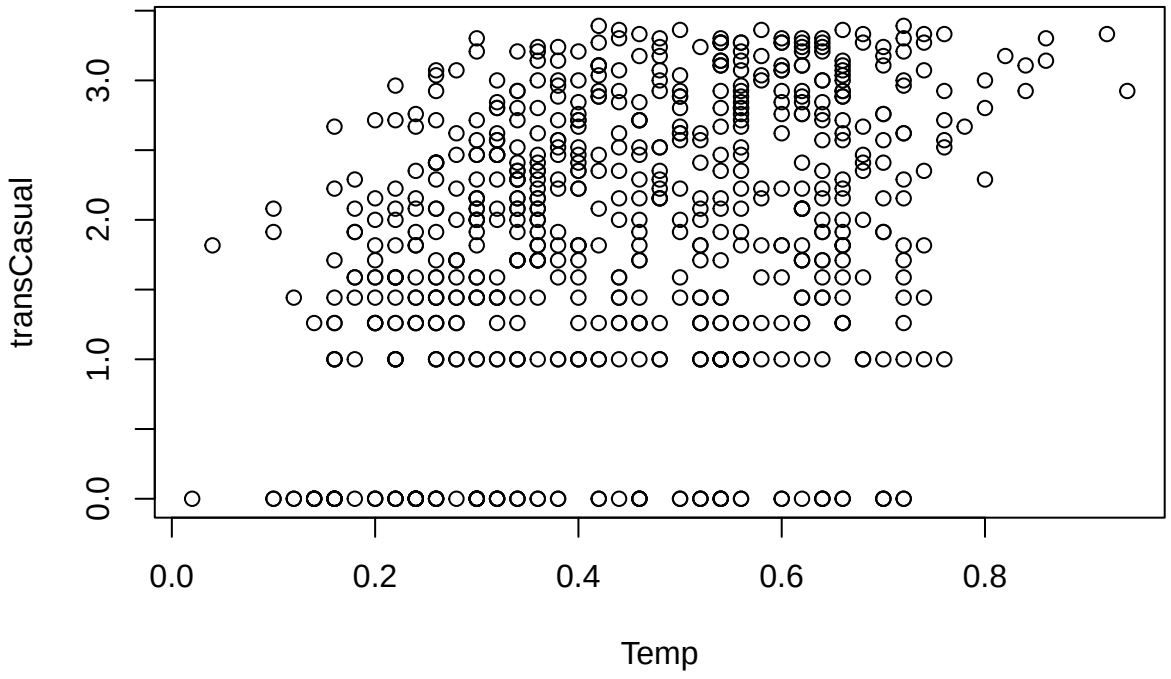
Bivariate Exploration

After exploring and analyzing each of the variable individually, it is eventually possible to graphically determine the association between response variable and each explanatory variable by creating plots such as scatter plots or boxplot, as follows

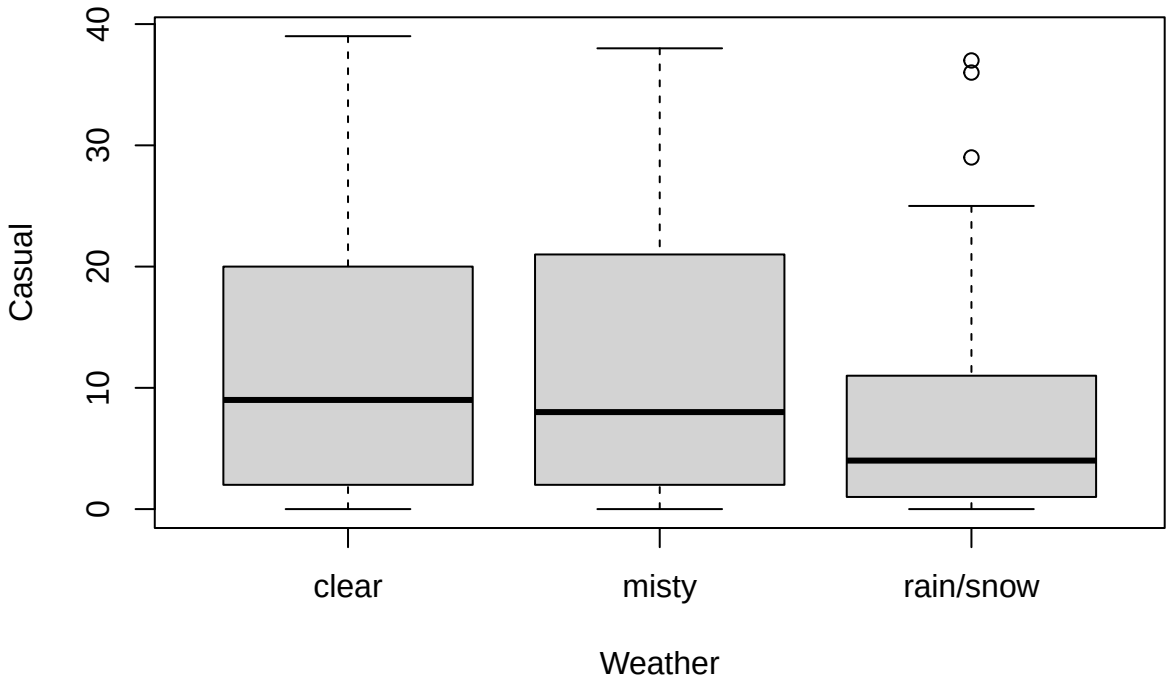
```
plot(transCasual ~ transWindspeed, data = bikes)
```




```
plot(transCasual ~ Temp, data = bikes)
```



```r
boxplot(Casual ~ Weather, data = bikes)
```



Through plots that are dranw by R studio, it is possible to have a rough understanding about the relationship between the response variable, **Transformed Casual**, and explanatory variables, Temperature, **Transformed Wind Speed**, and Types of Weather. However, since some of relationships are not obvious enough to see by eyes, (*i.e Scatterplot of transCasual and transWindspeed*), it is to create some simple linear regression models to see these relationship. Even though the data presented by those SLR models may not truly indicate the actual correlation coefficient between variables, it is still possble for us the see the direction of relationship in a large picture, as follows:

```r
Casual.Windspeed.mod <- lm(transCasual ~ transWindspeed, data = bikes)
Casual.Temp.mod <- lm(transCasual ~ Temp, data = bikes)
Casual.Weather.mod <- lm(Casual ~ Weather, data = bikes)
summary(Casual.Windspeed.mod)
```

```
##
## Call:
## lm(formula = transCasual ~ transWindspeed, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0886 -0.5879  0.1314  0.8169  1.7360
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.65526    0.09577  17.284   <2e-16 ***
## transWindspeed  0.51199    0.22323   2.294   0.0221 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.001 on 654 degrees of freedom
## Multiple R-squared:  0.007979,   Adjusted R-squared:  0.006462
## F-statistic:  5.26 on 1 and 654 DF,  p-value: 0.02213
```

```
summary(Casual.Temp.mod)
```

```
##
## Call:
## lm(formula = transCasual ~ Temp, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3991 -0.6148  0.1399  0.7476  1.7263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9874     0.1001   9.862   <2e-16 ***
## Temp          1.9606     0.2102   9.328   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9443 on 654 degrees of freedom
## Multiple R-squared:  0.1174, Adjusted R-squared:  0.1161
## F-statistic:    87 on 1 and 654 DF,  p-value: < 2.2e-16
```

```
summary(Casual.Weather.mod)
```

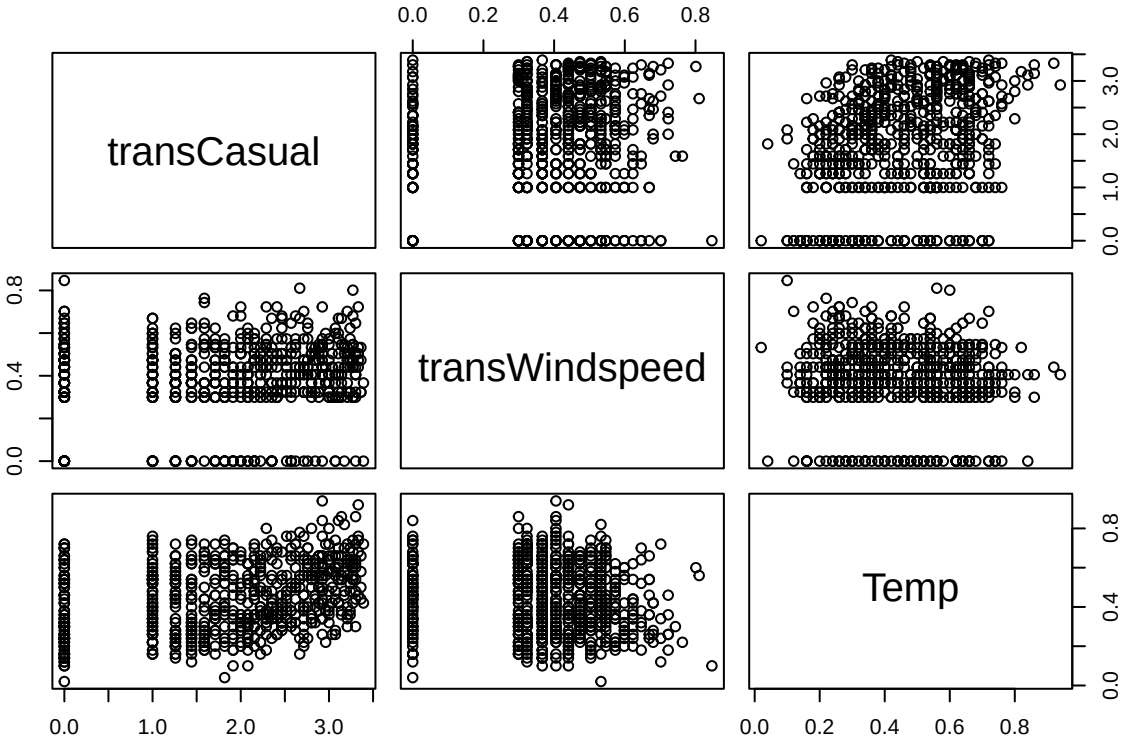
```
##
## Call:
## lm(formula = Casual ~ Weather, data = bikes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.085  -9.277  -3.085   7.915  29.556
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.0853     0.5645  21.410  < 2e-16 ***
## Weathermisty      -0.2325     0.9719  -0.239  0.81102
## Weatherrain/snow  -4.6408     1.4252  -3.256  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.1 on 653 degrees of freedom
## Multiple R-squared:  0.01639,    Adjusted R-squared:  0.01337
## F-statistic: 5.439 on 2 and 653 DF,  p-value: 0.004544
```

After doing a test about models of each explanatory variable with the response variable, it is not hard to see that according to scatterplots, the boxplot, and summaries of rough SLR models, the relationship between the response variable and explanatory variables can be understood by the following context.

- TransCasual has a positive relationship with transWindspeed. As the value of transWindspeed increases, the value of transCasual increase
- TransCasual has a positive linear relationship with temp(temperature). As the value of temp increases, the value of transCasual increases
- TransCasual has a negative relationship with types of weather besides 'clear'. In general, when the other variables maintain same, if the weather is either misty or rain/snow, there will have decrease in the value of transCasual.

#Modelling

After having both univariate and bivariate exploration to see attitudes of the data itself and the relationship between variable in both graphical and data-based ways, it is the time to fit a reasonable, logical linear regression model. Since all quantitative variables have unimodel and symmetric distribution, which can be understand as being normally distributed. Meanwhile, since we have already proved that every explanatory variable has some relationships with the response variable, all explanatory variables might be useful when constructing a model. Nevertheless, before having the final model, we still need to examine the multicollinearity issue between every explanatory variable. In order to check that issue, pairs graph can be a great tool to measure, as follows:



Through the result of the pairs graph does not show quite obvious relationship transWindspeed and Temp. This shows that testing the vif value of each variable tends to be an effective to see that whether there are any problems with the collinearity for explanatory variable in this situation.

Here are the results of the VIF test.(The GVIF value at here is also understanded as the VIF value)

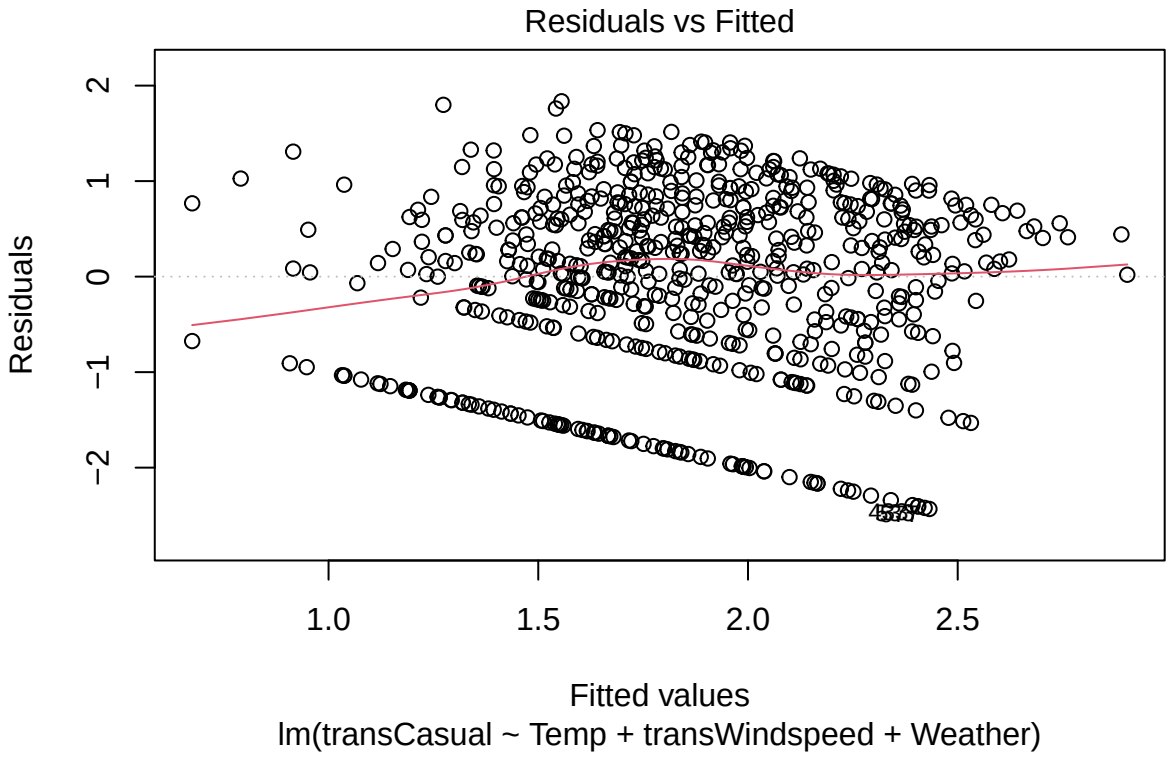
```
##                    GVIF Df GVIF^(1/(2*Df))
## Temp           1.013998  1        1.006975
## transWindspeed 1.013142  1        1.006549
## Weather        1.007457  2        1.001859
```

From the data of the test, it is not hard to see that each of the GVIF value of every explanatory variable is smaller than 2.5, which can a be good sign that there is no multicollinearity issue between quantitative explanatory variables.

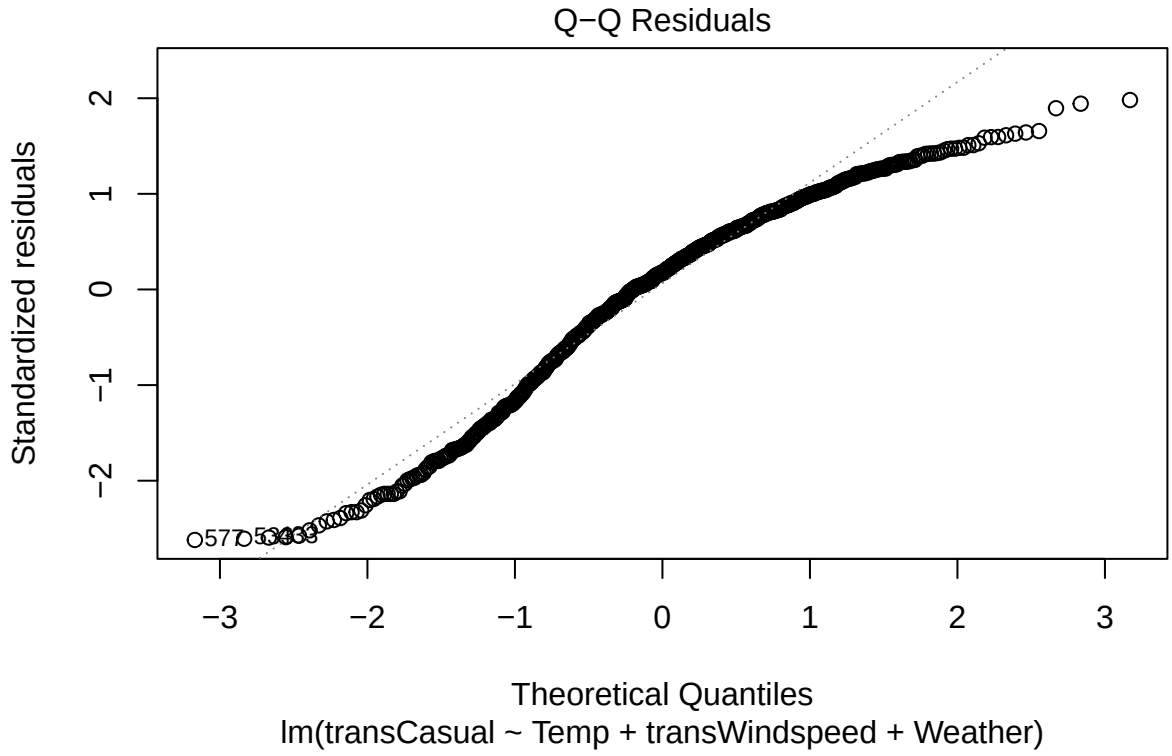
As a matter of fact, it is the time to see about residual plots and qq plots.

Graphs followed is the QQ plot and the residual plot of the predicted model:

```
plot(bikes.mod,
     which = 1)
```




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



From the residual diagnostic plots of models with explanatory variables of 'transWindspeed', 'Temp', and 'Weather', it is clear that we can do some interpretation from them.

- From the residual plot, even though at the bottom part of the residual plot, it seems to have some of the trend in residuals, which can be interpret as for this project it might not be suitable to use the linear regression model, or as the existence of heteroscedasticity in explanatory variables, we can still see that overall there is a constant spread, and independence for residual points. Most importantly, the mean of the residual plot is quite close to the zero. As a matter of fact, it may not be as perfect as what we expected, but it is still in the acceptable range since most of residuals fulfill the assumption that is required in linear regression model.
- From the QQ plot, we notice that at the end part of the plot, there are some deviation exist, and there are also some outliers in that part of the plot as well. In a large scale, they may not have any severe or greatly harmful impact on the model itself. At the same time, most of points on the plot is close to the fitted line, which can be a strong evidence of the existence of normality. So it might be the best QQ plot that we can created by using the linear regression model and given data.

With all those prior test done above, eventually the linear regression model can be created.

```
##
## Call:
## lm(formula = transCasual ~ Temp + transWindspeed + Weather, data = bikes.transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4325 -0.5980  0.1633  0.7199  1.8357
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.710018   0.139199   5.101 4.44e-07 ***
## Temp              2.013178   0.208782   9.642  < 2e-16 ***
## transWindspeed    0.745006   0.209059   3.564 0.000393 ***
## Weathermisty      0.005235   0.081683   0.064 0.948914
## Weatherrain/snow -0.357287   0.119783  -2.983 0.002963 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9315 on 651 degrees of freedom
## Multiple R-squared:  0.1452, Adjusted R-squared:  0.1399
## F-statistic: 27.64 on 4 and 651 DF,  p-value: < 2.2e-16
```

By using the R studio, we can see all related coefficients about the model of predicting the value of the *Casual bikes user*. However, the linearity of the model, and also other model that we did before for testing relationship between response variable and other variables are not as good as what we expected since the R square coefficient is only 0.1452 for this model, which is not a high value. On the other side, try to compare it with every other model that is created before, it shows a better R square value, and it also has an quite high F-statistic with 27.64(at least it is a high value for other model that based on the real-life data).

Meanwhile, the significance test for our model is great owing to the fact that for every coefficient quantitative variable and the dummy variables for categorical variables, they are with high statistic significance level —— p-values a lot more smaller than 0.05. Even for the F-statistic, it has a p-value smaller than 2.2e-16. Therefore, it is save to say that this model can be a great estimation for the population model.

The coefficient for most of the explanatory variable in the final model is consistent with the on in rough models constructed at the EDA section. To be more specific, there are positive relationship between response variable 'transCausal' and explanatory variables 'Temp', 'transWindspeed', and there is negative relationship between response variable 'transCausal' and categorical explanatory varible 'rain/snow'. However, what we didn't expect is the inconsistence in the coefficient of categorical explanatory variable 'misty', which can be the problem of whether the data, the methodology that is used for testing, or the model that is chosen. Still, overall, it is still a reasonable model just like what we predicted in EDA section.

In conclusion, even though model may not predict the reality situation in a extreme accurate way due to reasons mentioned above, it is still the best linear regression model what we can get from the given dataset

Prediction

For the prediction of the number of casual users for an hour with misty/cloudy weather, a scaled temperature of 0.75, and a scaled windspeed of 0.25, the calculation formula and the result are shown, as follows

```
(2.013178*0.75 + 0.745006*(0.25^(1/2)) + 0.005235 + 0.710018) ^ 3
```

```
## [1] 17.52817
```

As a result, the predicted number of casual users for an hour with misty/cloudy weather, a scaled temperature of 0.75, and a scaled windspeed of 0.25 is 17.52817

Discussion

In this project, we used the linear regression model to predict the number of Casual bikes user for an hour with explanatory variables of temperature, wind speed and types of weather.

During the process of doing the project, there are still some problems that arises. In a specific way, major problems can be

- Deviation and outliers exists in QQ plot. So for the next time, it is necessary to select more data to see whether such 'error' can be eliminated
- Some of the trend in residuals. Therefore, for the next time, try some other models instead only use the linear regression model, or try to test the existence of heteroscedasticity in explanatory variables.
- The low in the R-square value of the model. So next time, try to use some more complex models such as quadratics regression model or other models that are more suitable.

I strongly believe that if these question can be solved, it is possible for us to conclude a more accurate model for making prediction to the real life cases.