Factors Affecting Rental Bike Usage

Siddhant Singh

Student Number - 1005242266
Model Validation, Model Building

Mervyn Sanjayan

Student Number - 1005465970
Reasearch, Model Diagnostics

Bhavya Jain

Student Number - 1005604749
Background & Significance, Model Validation

Mahek Prasad

Student Number - 1005877696
Exploratory Data Analysis, Presentation, Model Diagnostics

Case Study Group 4 STAC67

STAC67 Case Study

```
## Loading required package: pacman
```

Background and Significance

Abstract

Currently, Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern.

The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes. The major factors affecting bike demand are weather and holidays.

This case study aims to provide a better insight into the key factors affecting bike rental demand so that an efficient supply of bikes can be formulated.

Introduction

The data was cleaned in three stages. Firstly, the predictor variable names were changed to X_i and the response variable name to Y. Then, the dataset was divided into two equal parts. The first set of data points were used for training and the second set for testing our model. Lastly, NA or missing values were also checked for.

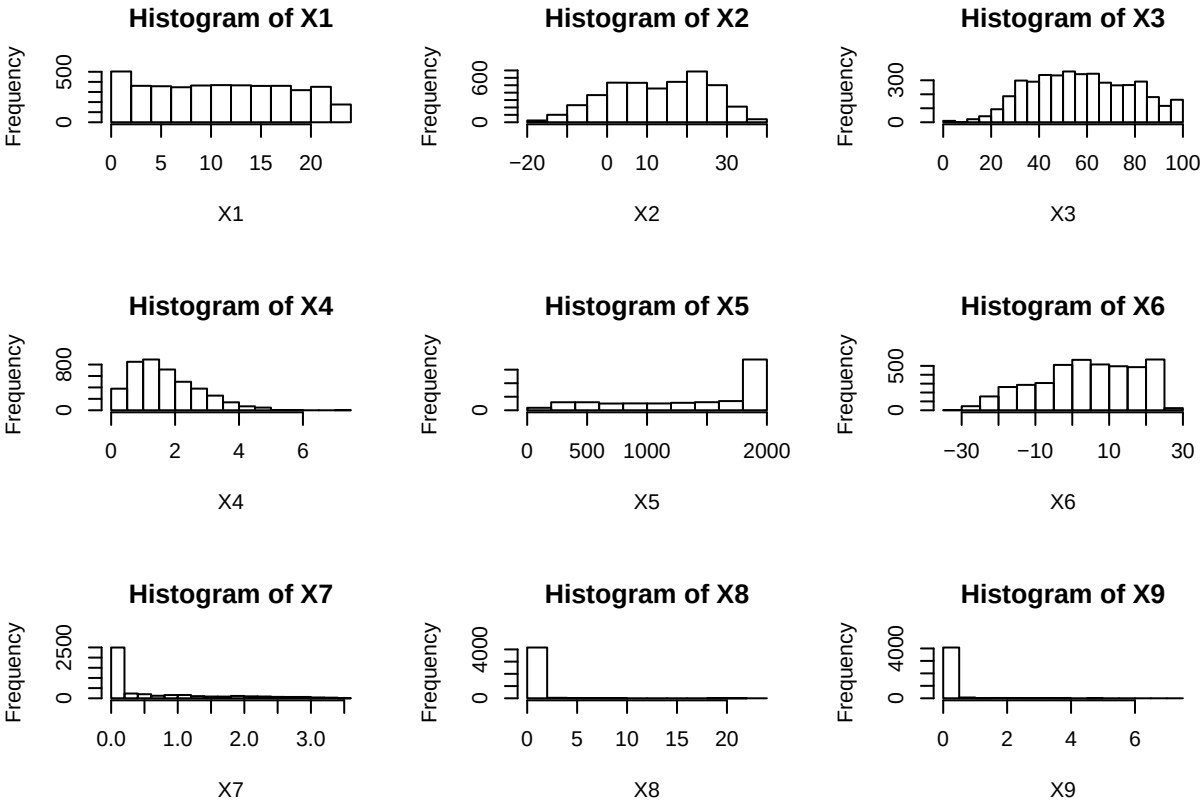
We data was filtered out to only contain rows where the Functional Day (X12) value was true. This is because, the day being functional or not is independent of the other variables as it is caused usually by system failure which is unpredictable. Functional Day is also removed from consideration to be used as a predictor because of this reason. A functional day value being true has no affect on the magnitude of the number of bikes rented and when its false, the number of bikes rented are obviously 0.

Stepwise model selection is used to get the initial model and thereafter, several diagnostic measures like Boxcox tranformation, weighted least squares are applied to get a better fit and satisfy all the Gauss-Markov assumptions for linear regression.

Variable Information

Variable	Description
Rented Bike Count (Y)	Count of bikes rented at each hour
Hour (X1)	Hour of the day
Temperature (X2)	Temperature in Celcius
Humidity (X3)	Humidity in %
Windspeed (X4)	Wind speed in m/s
Visibility (X5)	Distance at which an object or light can be clearly discerned
Dew Point Temperature (X6)	The temperature at which the air is saturated with moisture. (in Celcius)
Solar Radiation (X7)	Solar Radiation in MJ/m2
Rainfall (X8)	Rainfall in mm
Snowfall (X9)	Snowfall in cm
Seasons (X10)	Seasons one of Winter, Summer, Spring, Autumn
Holiday (X11)	Holiday/No Holiday
Functional Day (X12)	Functional or non-funcitonal hours

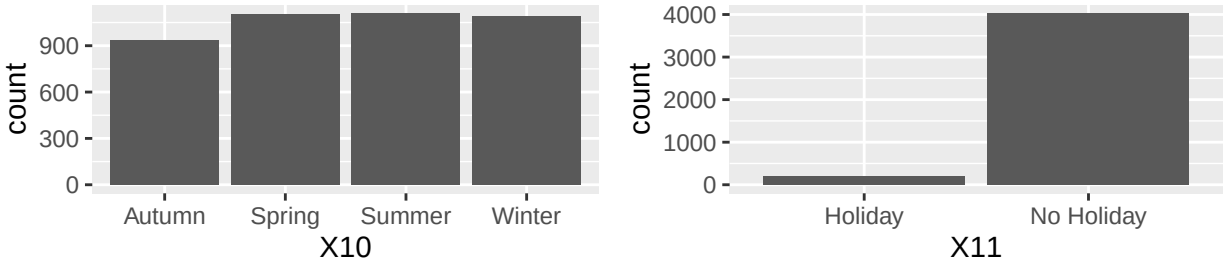
Histograms of Continuous Variables



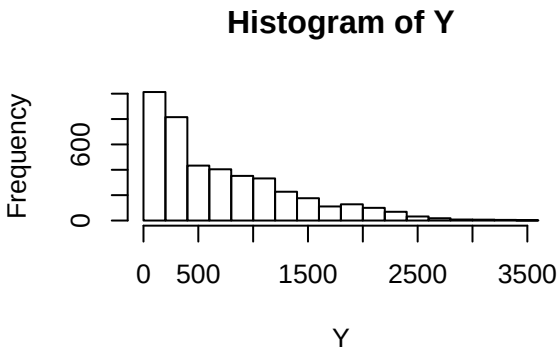

```
if (!any(is.na(data))) {
  print("No NA values were found in the dataset")
}
```

```
## [1] "No NA values were found in the dataset"
```

Bar Plots for Categorical Variables

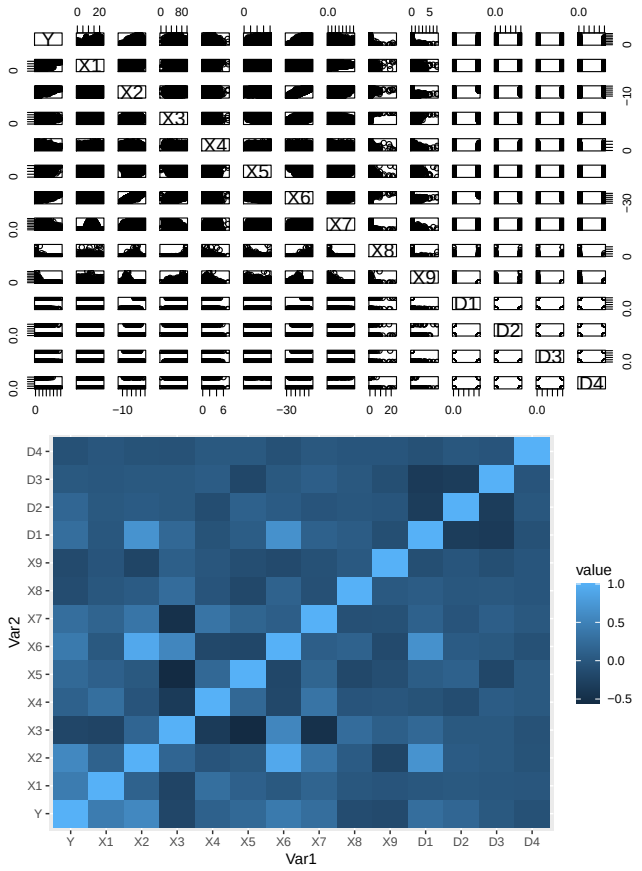


Histogram of Response Variable Y



Exploratory Data Analysis

Correlation Between Continuous Variables



From the correlation matrix, there looks to be a very strong correlation (=0.91468312) between $X2$ (Temperature) and $X6$ (Dew Point Temperature). So, to avoid multi-collinearity, the $X6$ variable is dropped. Also there is a strong correlation (=0.6759473864) between our dummy variable $D1$ and $X2$ so we can drop $D1$ as well. It is also hypothesized that temperature (X2) and hour of the day (X1) will have the strongest positive effect on rental bike usage due to their relatively high correlation with Y. These high correlation values can be observed by the lighter shade of blue in the second plot.

Model

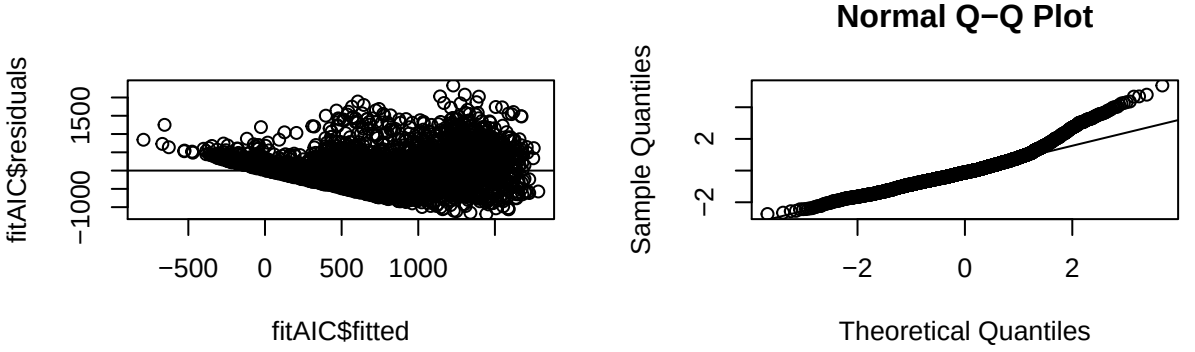
Model Selection

For model selection, both direction (forward and backward) stepwise model selection procedure is used taking into consideration the full model (with all predictors apart from interaction terms) and the simplest model (with no predictors) to find the predictors that are most important.

```
##
## Call:
## lm(formula = Y ~ X2 + X1 + X3 + D2 + X8 + X7 + D3 + D4 + X9 +
##     X4, data = trainingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1187.16  -272.54   -51.31   207.08  2327.86
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  444.1122    33.8202  13.132  < 2e-16 ***
## X2            32.4902     0.6818  47.652  < 2e-16 ***
## X1            28.2319     1.0497  26.896  < 2e-16 ***
## X3            -8.5573     0.4380 -19.537  < 2e-16 ***
## D2           271.0711    17.1120  15.841  < 2e-16 ***
## X8           -67.0507     6.5500 -10.237  < 2e-16 ***
## X7           -93.7071    10.3704  -9.036  < 2e-16 ***
## D3           111.6234    16.2538   6.868 7.49e-12 ***
## D4          -151.1035    30.8085  -4.905 9.71e-07 ***
## X9            39.8339    15.8041   2.520   0.0118 *
## X4            11.4061     7.2590   1.571   0.1162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 433.6 on 4221 degrees of freedom
## Multiple R-squared:  0.5462, Adjusted R-squared:  0.5451
## F-statistic:   508 on 10 and 4221 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value of this model is satisfactory but not adequate (=0.5451). Several diagnostic procedures will be followed to better this measure.

Model Validation



As can be seen from the residual plot, the equal variance assumption is clearly violated. Along with that, the residuals also don't seem to be evenly distributed. Also, from the QQ plot, it can be observed that the points

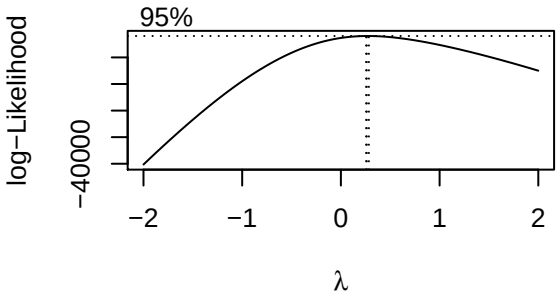
stray away from the line for a huge chunk indicating that the residuals don't come from a normal distribution.

Model Diagnosis

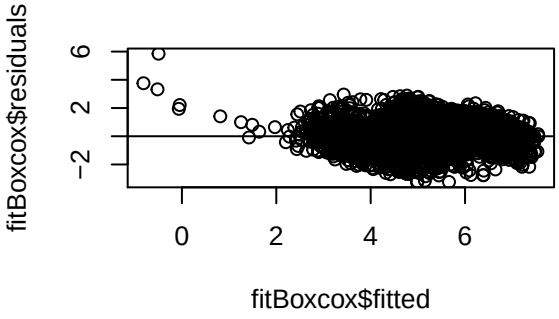
Boxcox Tranformation

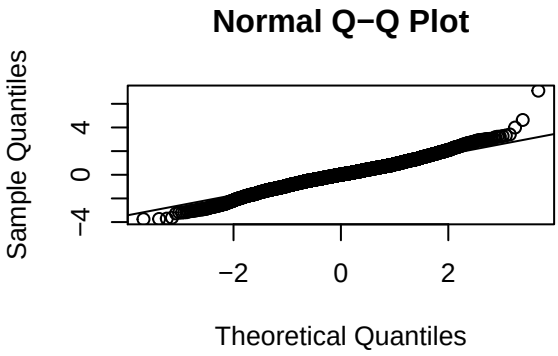
```
##      Lambda
## 1 0.2626263
```

```
##   R.squared
## 1 0.6329822
```



Upon doing a Boxcox transformation of the response variance, a lambda value of ≈ 0.26 is obtained. When using this transformation, we get an improved adjusted R-squared value of 0.63298 and improved residual and QQ plots.



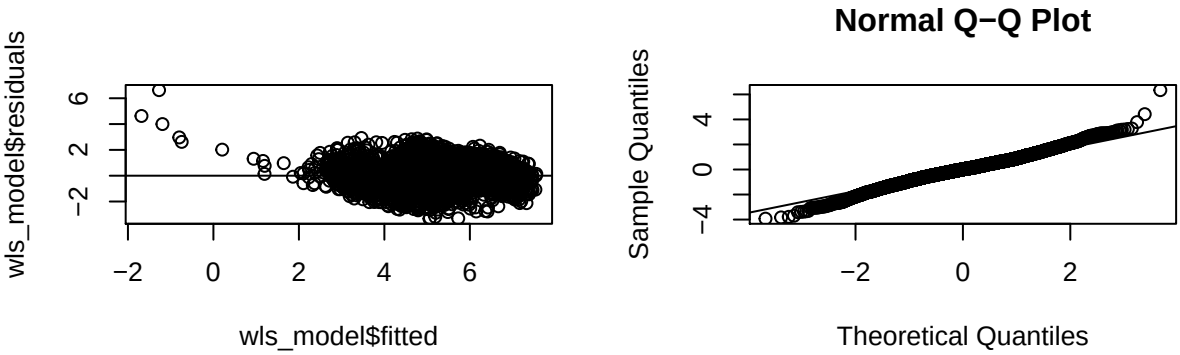


The variances look comparatively more evenly distributed except some outliers and the QQ plot almost lies on a straight line.

Weighted Least Squares

On doing a weighted least squares regression, there is not much affect on the adjusted R-squared value as it can still be rounded to the same 0.64 as seen above.

```
##   R.squared
## 1 0.6349566
```



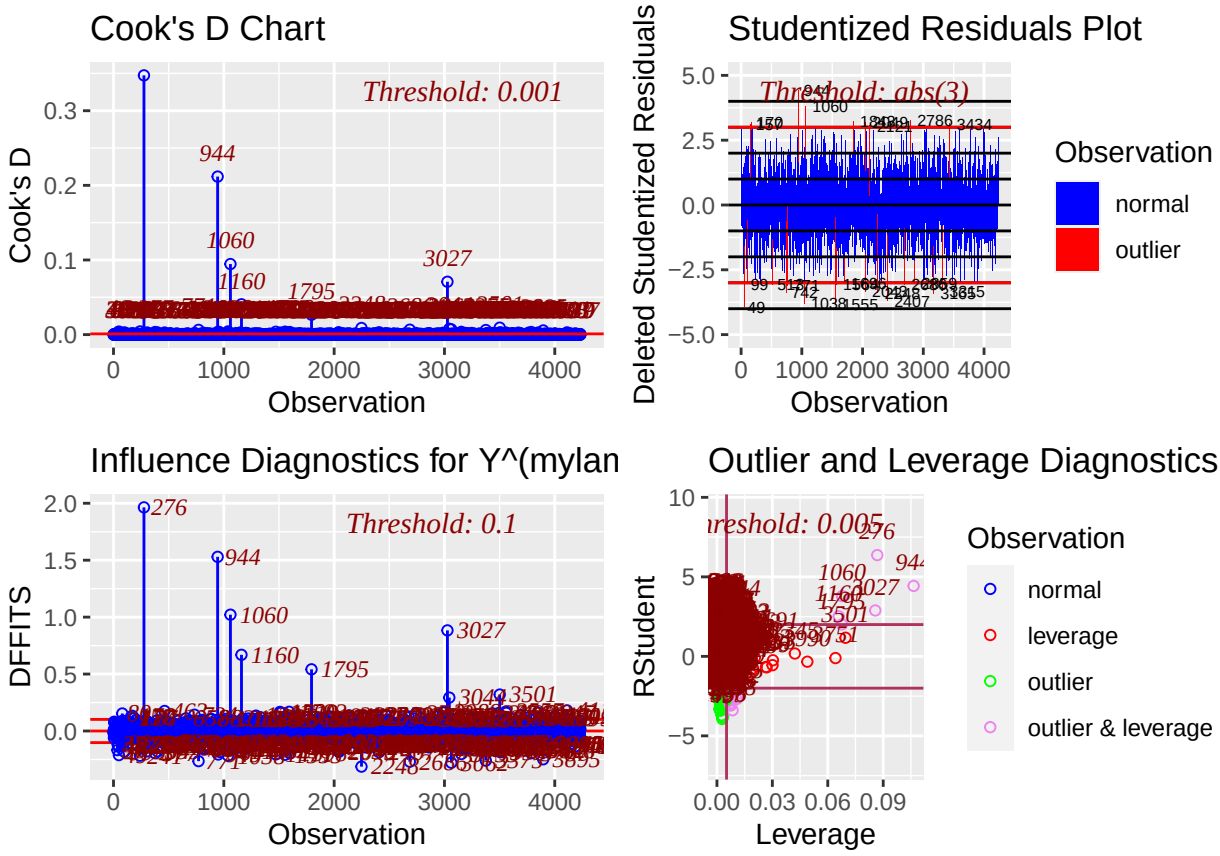
There is also not any noticeable change in the residual and QQ plots.

Outlier Detection

In this section, we would try identify the numerous outliers present in the model to give a better understanding of the model and the dataset to the reader.

```
## Warning: Removed 1 rows containing missing values (position_stack).
```

```
## Warning: Removed 3701 rows containing missing values (geom_text).
```



By taking the union of outliers from all techniques with non-null outputs, a confident list of all outliers can be created. The model outputted by removing these outliers had a considerable improvement in the adjusted R-squared value at 0.71. So, this model is the final model chosen for the problem at hand.

```
##   R.squared
## 1 0.7106186
```

Final Model

The model obtained by outlier removal is:

```
##
## Call:
## lm(formula = boxcox_formula, data = striped_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39392 -0.46524  0.00095  0.46209  2.68768
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4066441  0.0621112  70.948  < 2e-16 ***
## X1           0.0638399  0.0018632  34.263  < 2e-16 ***
## X2           0.0747152  0.0012046  62.026  < 2e-16 ***
## X3          -0.0173887  0.0008175 -21.270  < 2e-16 ***
## X4           0.0069165  0.0129828   0.533    0.594
## X7          -0.1092668  0.0183357  -5.959 2.75e-09 ***
## X8          -0.4514658  0.0225042 -20.061  < 2e-16 ***
## X9          -0.0175294  0.0301276  -0.582    0.561
## D2           0.6943519  0.0302172  22.979  < 2e-16 ***
## D3           0.3037648  0.0287959  10.549  < 2e-16 ***
## D4          -0.5313945  0.0595279  -8.927  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7432 on 4011 degrees of freedom
## Multiple R-squared:  0.7113, Adjusted R-squared:  0.7106
## F-statistic: 988.4 on 10 and 4011 DF,  p-value: < 2.2e-16
```

But, the X4 and X8 variables are not statistically significant and can be dropped. So, a simpler model could be the one with those variables removed. The final model obtained through this study is:

```
outlierFitReduced <- lm(formula = Y^(mylambda) ~ X1 + X2 + X3 + X7 + X8  + D2+ D3 + D4, data = striped_
summary(outlierFitReduced)
```

```
##
## Call:
## lm(formula = Y^(mylambda) ~ X1 + X2 + X3 + X7 + X8 + D2 + D3 +
##     D4, data = striped_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38793 -0.46425  0.00301  0.46378  2.68816
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4213644  0.0572624  77.212  < 2e-16 ***
## X1           0.0640130  0.0018103  35.361  < 2e-16 ***
## X2           0.0748273  0.0011524  64.932  < 2e-16 ***
## X3          -0.0175329  0.0007962 -22.021  < 2e-16 ***
## X7          -0.1082907  0.0176713  -6.128 9.75e-10 ***
## X8          -0.4505078  0.0224441 -20.072  < 2e-16 ***
## D2           0.6931310  0.0299256  23.162  < 2e-16 ***
```

```
## D3           0.3061548  0.0285497  10.724  < 2e-16 ***
## D4          -0.5293624  0.0594604  -8.903  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.743 on 4013 degrees of freedom
## Multiple R-squared:  0.7113, Adjusted R-squared:  0.7107
## F-statistic:  1236 on 8 and 4013 DF,  p-value: < 2.2e-16
```

$$Y = 0.0640130X_1 + 0.0748273X_2 - 0.0175329X_3 - 0.1082907X_7 - 0.4505078X_8 + 0.6931310D_2 + 0.3061548D_3 - 0.5293624D_4 + 4.4$$

Where D_2 , D_3 are dummy variables for season each representing Autumn and Spring season respectively while $D4$ is a dummy variable representing if there was a Holiday at that day or not.

Model Interpretation

Out of the qualitative variables we can see that $X3$,$X7$ and $X8$ have a negative effect on the response variable Y keeping all the other factors constant, while $D4$ is the only qualitative variable showing this behaviour. Our model is also significant by the F-test with a p-value $< 2.2 \times 10^{-16}$.Also all of the features selected are statistically significant proven by the extremely small p-values $< 2.2 \times 10^{-16}$ for their t-tests.

Validate Final Model

MSPR

```
##       MSE      MSPR
## 0.5520884 0.8612766
```

As the MSPR and MSE values are fairly close, we can say that our selected model is a good fit for the data at hand. The testing dataset was used to calculate MSPR here. Below is the MSPR and MSE for the outlier removed model:

VIF

From the VIF table obtained from the "vif" function of the "car" library, there are no GVIF values > 10, so there is no serious multi-collinearity issue.

```
##       X1       X2       X3       X7       X8       D2       D3       D4
## 1.126879 1.448157 1.815479 1.767300 1.069630 1.111473 1.129791 1.013302
```

```
##   Mean.VIF
## 1 1.310251
```

The mean VIF value is also not much larger than 1. So, we do not need to worry about multi-collinearity.

Discussion/Conclusion

As seen from the summary of the final model, nearly all variables are statistically significant as noted by their extremely small p-values. Variables X_1, X_2, X_3, X_7, X_8, X_{10} and X_{11} are all important for predicting Y. Below is a table of adjusted R-squared values acquired when removing each of the variables. The lowest value of corresponsing adjusted R-squared means the variable is more important. So, this list, in sorted order, lists the variables from most important to least important.

```
##   Variables R.2.After.Removal
## 2        X2         0.3686534
## 1        X1         0.5659070
```

```
## 3        X3         0.5847498
## 6        D3         0.5981638
## 7        D4         0.5986375
## 5        X8         0.5986624
## 4        X7         0.6322130
```

From the table, we can see a more holistic picture of which variables are more important. It can be observed that X2 (temperature) and X1 (hour of day) are the most important factors affecting rental bike usage.

Research done by studies such as *Investigation on the effects of weather and calendar events on bike-sharing* by Kyoungok Kim which led to similar findings in how factors such as functional days and temperature had the strongest positive effect on rental bike usage.

Since we used outlier detection to remove outliers, one limitation of our study might be that the data points might be a biased based on the outlier detection techniques we used. Also, the R-squared value is not that high (≈ 0.71). In the future, an approach other than linear regression might result in a better model.

References

1. Kyoungok Kim (2018), Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations, *Journal of Transport Geography*, Volume 66, Pages 309-320, ISSN 0966-6923. Retrieved from https://doi.org/10.1016/j.jtrangeo.2018.01.001