Seoul Rental Bike Prediction from Weather Data

April 10, 2022

Authors and Job Assignments

- Eddie Shin (997743615) - model building, EDA and presentation speaker
- Jintong () - data cleansing and responsible for the final report
- Du Han () - model building, model diagnostic and presentation Speaker
- Chenqi () - made powerpoint slides

1. Background and Significance

In this project we are trying to research on how the weather condition such as humidity, windspeed, visibility, temperature, dewpoint temperature influence the bike rental amount in different hours through out the day. In this research we didn't consider the specific days because weather changes could be varied in different hours during the day, however, weather condition could have more power on representing the season's feature. Therefore we omit the dates and hours. Our main research question is whether the weather condition through out the year will influence the bike rental amount. Therefore, we consider to compare the main effect model and main effect interaction to see whether the weather condition can significantly influence the rental bike amount. The Null hypothesis is the weather condition through out the year can influence the bike rental amount

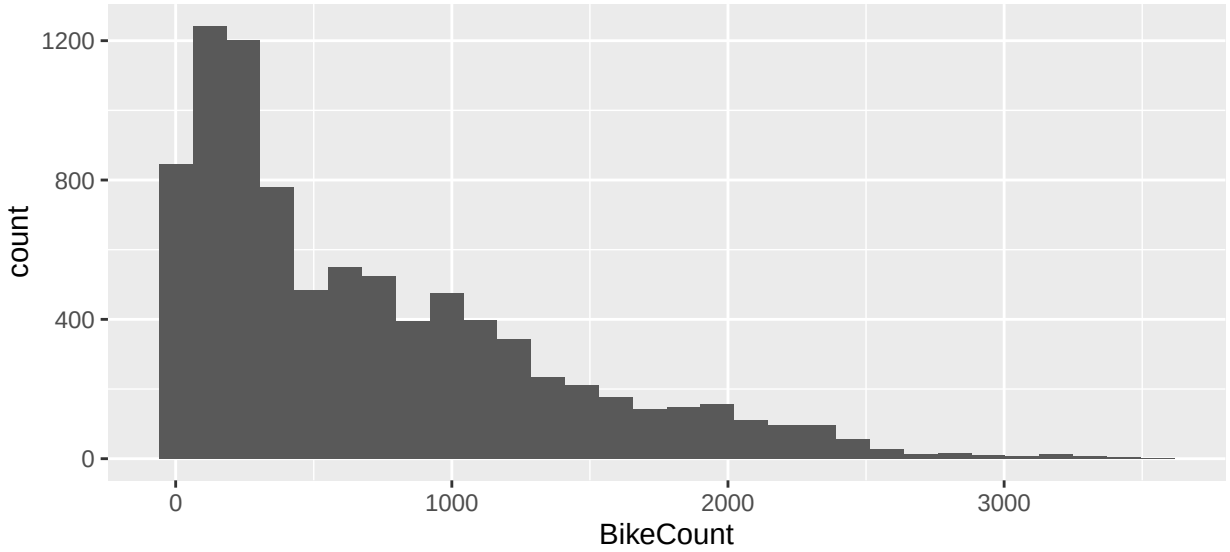
2. Exploratory Data Analysis

2.1 Data Description

The dataset collect 8760 data from between 2017 to 2018, with different bike rental amount in different days during the year and weather condition. In this case study The response variable we choose is Rented.Bike.Count and hour. The explanation variable we choose is seasons and its corresponding temperature,dew point temperature,wind speed,humidity, rainfall in Spring Summer,Autumn seasons. In Winter we consider adding snowfall into the model and interaction between rainfall and snowfall.

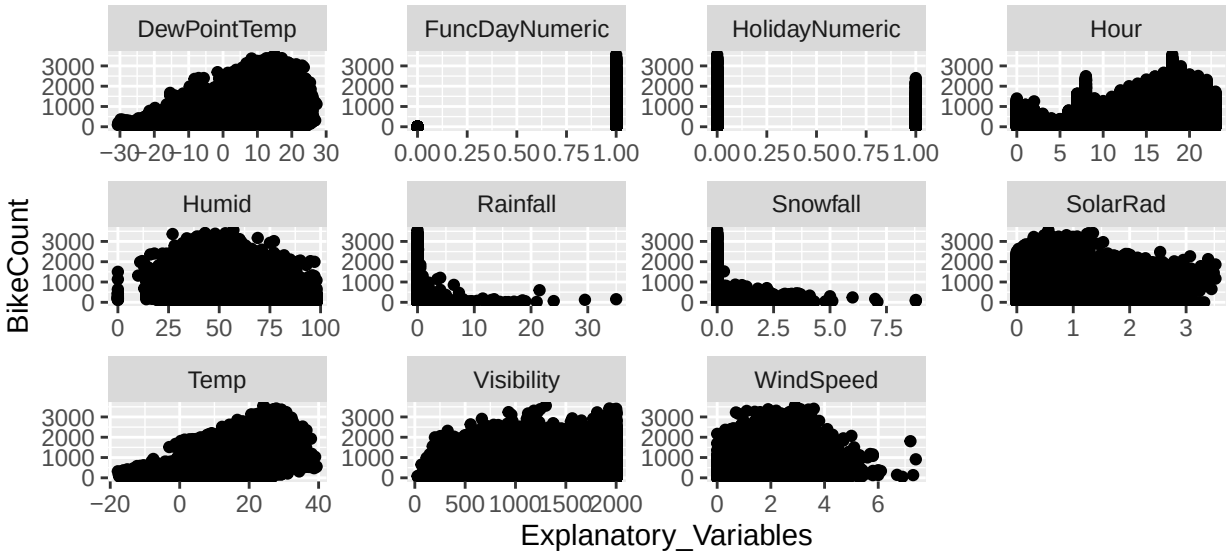
In this model the main effect is temperature, humidity, windspeed, visibility, dewpoint temperature, solar radiation, snowfall, seasons, holiday, functional day

2.2 Distribution of Response Variable

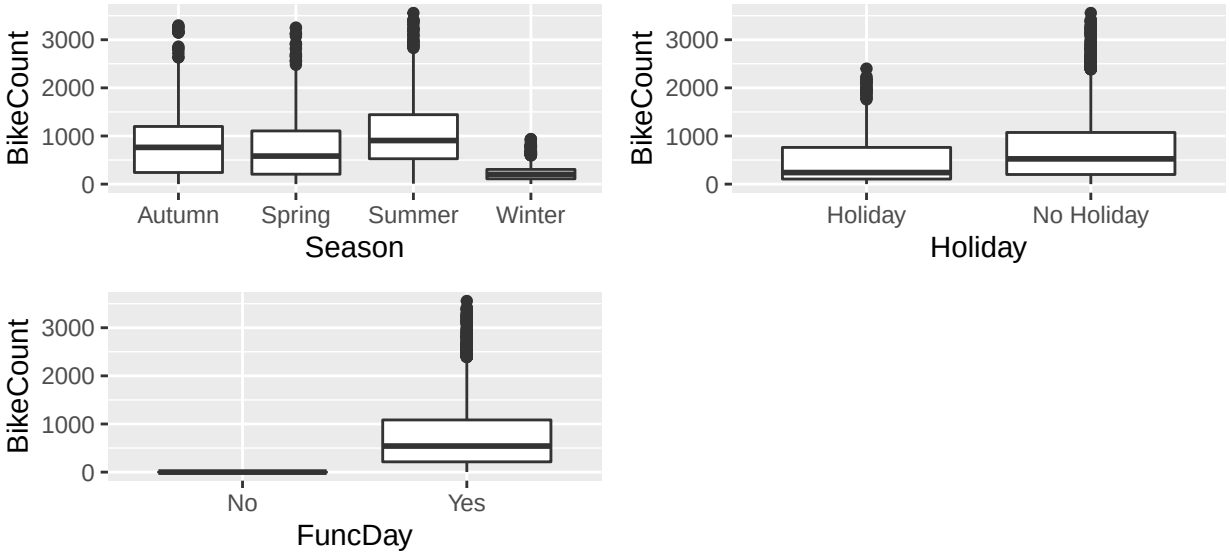


The distribution of our response variable reveals that it's skewed to the right. In order to maximize the predictability of the model, we might need some sort of transformation on our response variable.

2.3 Distribution of Explanatory Variables against Response Variable



2.4 Boxplots of Categorical Variables



We have three categorical variables in the dataset, `Season`, `Holiday` and `Functional Day`. Based on the distributions of the box plots above, we can make some inferences about each categorical variable. For `Season`, we can clearly see that the rental activity during winter is significantly less than other seasons. For `Holiday`, we can observe the pattern that more people are active with sharing bikes during non-holiday periods. Lastly, for `Functional Day`, we have close to 0 rental activity when `FuncDay` is equal to 0 and active sharing when `FuncDay` is equal to 1.

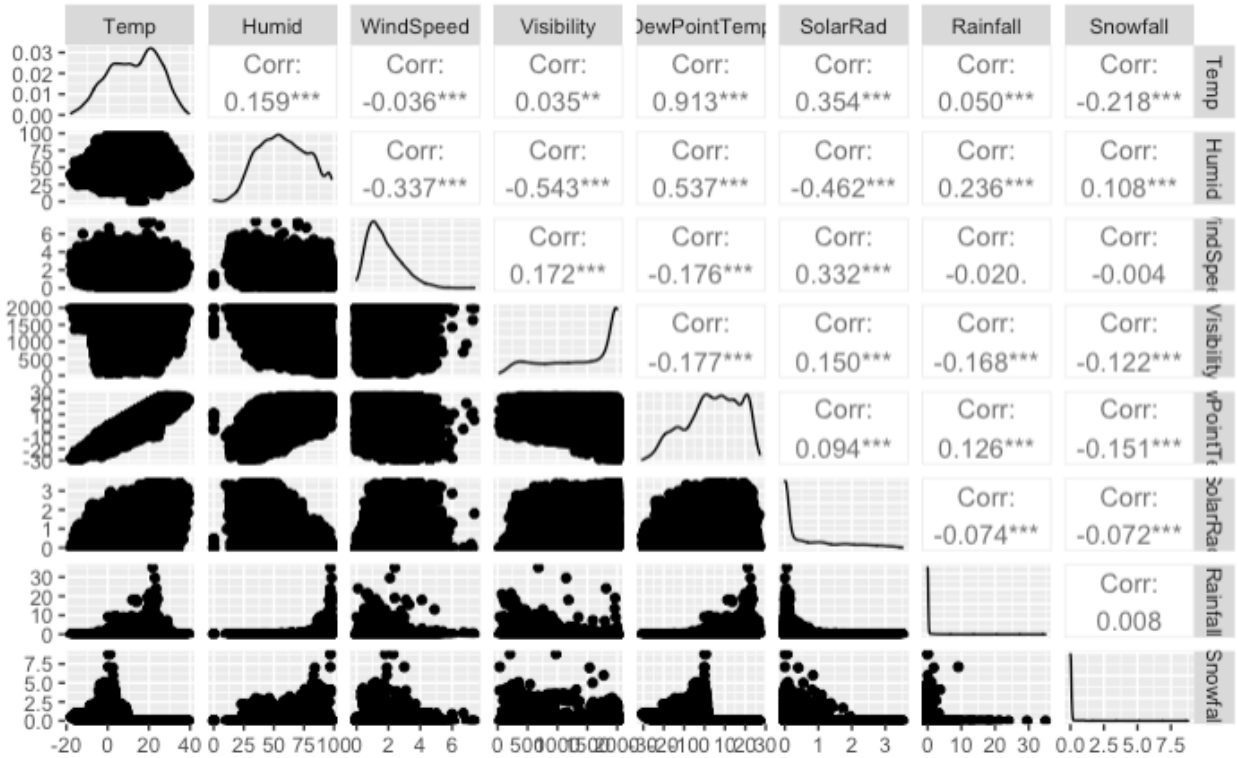
2.5 Summary Statistics of Quantitative Variables

```
##       Temp             Humid          WindSpeed       Visibility
##  Min.   :-17.80   Min.   : 0.00   Min.   :0.000   Min.   :  27
##  1st Qu.:  3.50   1st Qu.:42.00   1st Qu.:0.900   1st Qu.: 940
##  Median : 13.70   Median :57.00   Median :1.500   Median :1698
##  Mean   : 12.88   Mean   :58.23   Mean   :1.725   Mean   :1437
##  3rd Qu.: 22.50   3rd Qu.:74.00   3rd Qu.:2.300   3rd Qu.:2000
##  Max.   : 39.40   Max.   :98.00   Max.   :7.400   Max.   :2000
##   DewPointTemp        SolarRad         Rainfall          Snowfall
##  Min.   :-30.600   Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000
##  1st Qu.: -4.700   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000
##  Median :  5.100   Median :0.0100   Median : 0.0000   Median :0.00000
##  Mean   :  4.074   Mean   :0.5691   Mean   : 0.1487   Mean   :0.07507
##  3rd Qu.: 14.800   3rd Qu.:0.9300   3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   : 27.200   Max.   :3.5200   Max.   :35.0000   Max.   :8.80000
```

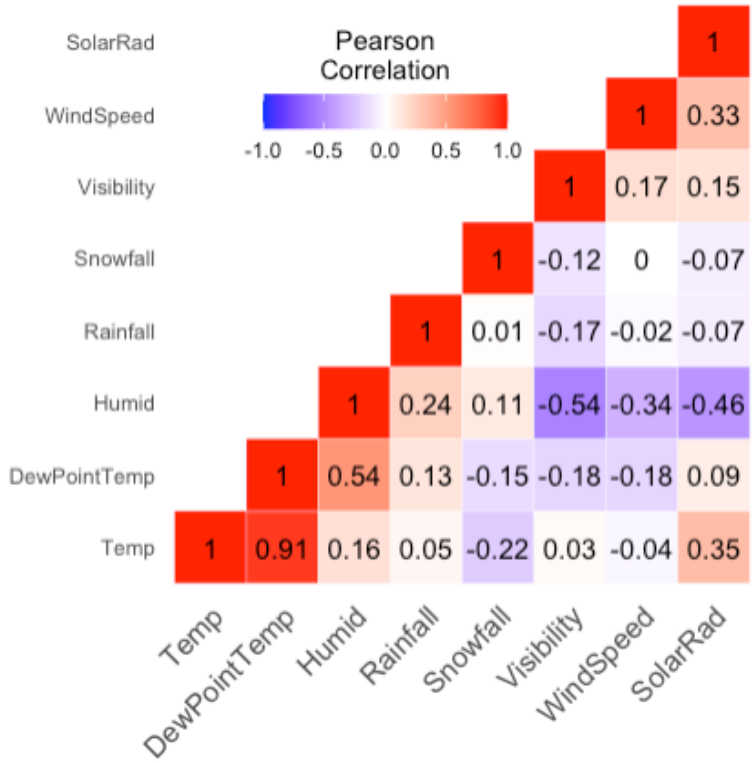
The table above summarizes the key statistics of each quantitative variable. Each variable represents weather condition in Seoul, Korea.

- `Temp` measures temperature during specific hour in Celsius. The mean `Temp` in the dataset is 12.88292, and the range of `Temp` is from -17.8 to 39.4 degree Celsius.
- `Humid` measures the humidity of the day. The mean of `Humid` is 58.23 and its range is from 0% to 98%.
- `WindSpeed` is used to describe how strong the wind is during the day in m/s. The mean of `WindSpeed` is 1.724909, and its range is from 0 m/s to 7.4 m/s.
- `Visibility` measures how far people can see during the day with the unit of 10m. The mean `Visibility` is 1436.826m with the range between 27m and 2000m.
- `DewPointTemp` is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. The mean of `DewPointTemp` is 4.073813 Celsius, and its range forms from -30.6 C to 27.2 C.
- `SolarRad` is an abbreviated form of Solar Radiation (MJ/m2) which is a general term for the electromagnetic radiation emitted by the sun. The mean `SolarRad` is 0.5691107 with the range from 0 to 3.52(MJ/m2).
- `Rainfall` is the amount of rain during specific hour of the day in mm. The mean `Rainfall` is 0.1486872(mm) and the range is from 0mm to 35mm.
- `Snowfall` is the amount of snowfall during the specific hour of the day in cm. The mean `Snowfall` in the dataset is 0.07506849 cm, and it ranges from 0cm to 8.8cm.

2.6 Correlations between Explanatory Variables in Matrix and Heatmap Forms



Based on the pairs of explanatory variables, we can detect a linear relationship between `Temp` and `DewPointTemp` which means this linear relationship may cause some problems related to multicollinearity.

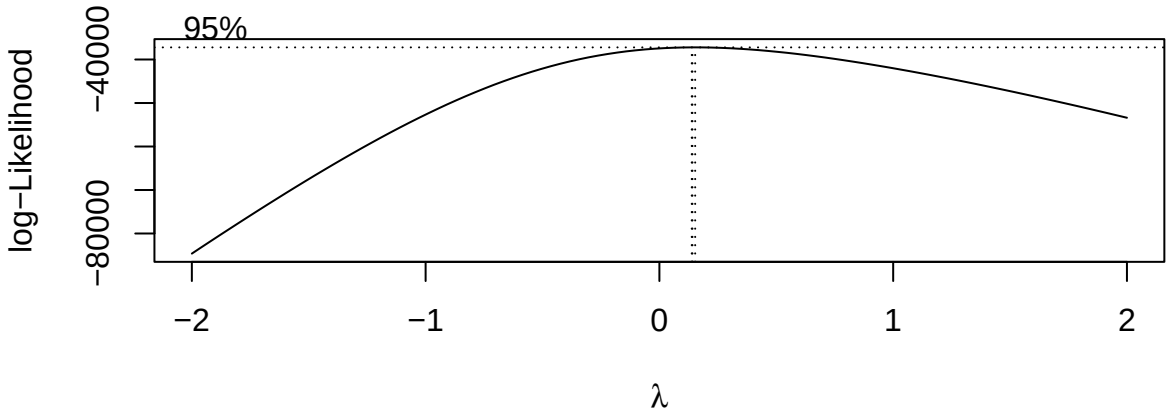


Based on the correlation heat map above, we can quantitatively verify that the correlation between `Temp` and `DewPointTemp` has the highest value of 0.91 which means there is a strong positive relationship between the two variables.

3. Model Selection and Validation

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df  logLik     AIC     BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1     0.478         0.477  466.      617.       0    13 -66250. 132531. 132637.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

The initial main-effect only model without any modification returns a 47% R-squared. Also both AIC and BIC are very high, 132531 and 132637 respectively. We decided to perform Box-Cox transformation on our response variable.



The optimal lambda of Box-Cox transformation is 0.141, and we used this lambda to transform the response variable. After the transformation is performed, we have the following results.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.813         0.813 0.245     2931.       0    13  -85.2  200.  307.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

As we can see from the result table above, our new R-squared is 81.33% with AIC of 200.3632 and BIC of 306.5324. Thus, the Box-Cox transformation definitely helped improve the model in terms of R-squared quite significantly while reducing both AIC and BIC. We can further improve the model with `stepAIC` function to remove insignificant variables.

```
##
## Call:
## lm(formula = y_transformed ~ Humid + WindSpeed + Visibility +
##     DewPointTemp + SolarRad + Rainfall + factor(Season) + factor(Holiday) +
##     factor(FuncDay), data = bike)
##
## Coefficients:
##               (Intercept)                      Humid
##                 4.748e-01                 -1.158e-02
##                 WindSpeed                 Visibility
##                 2.276e-02                 -1.658e-05
##              DewPointTemp                   SolarRad
##                 1.909e-02                 -2.628e-02
##                  Rainfall       factor(Season)Spring
##                -5.749e-02                 -1.112e-01
##      factor(Season)Summer       factor(Season)Winter
##                -1.363e-01                 -2.308e-01
## factor(Holiday)No Holiday         factor(FuncDay)Yes
##                 1.150e-01                  2.540e+00
```

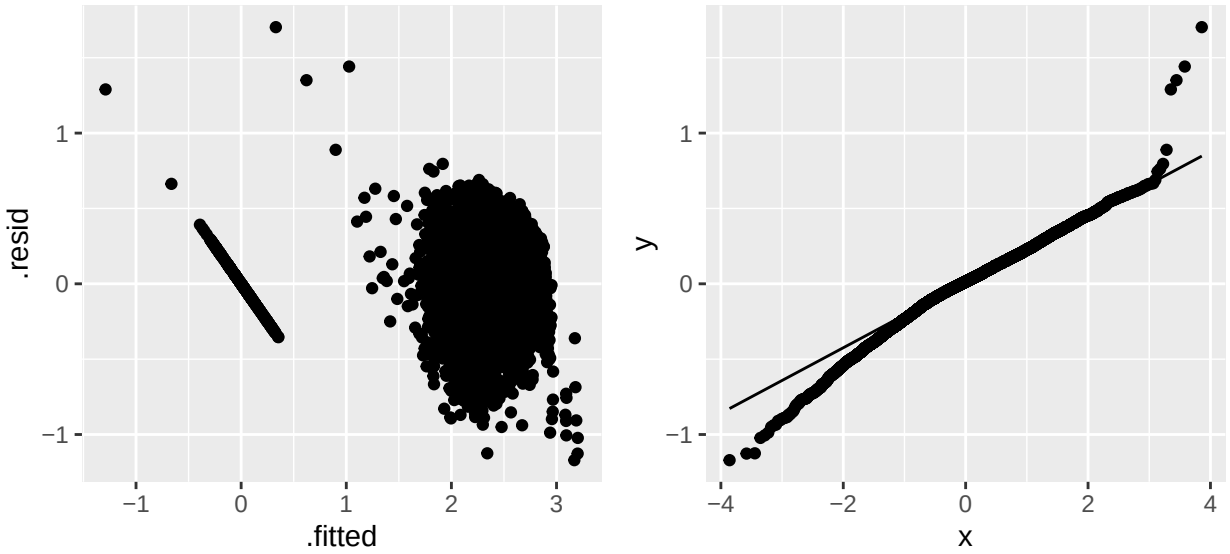
```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.813         0.813 0.245     3464.       0    11  -86.0  198.  290.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

According to the result of `stepAIC` function, we could remove `Temp` and `Snowfall` from the model while maintaining the same R-squared. Before we finalize our model, we also tested the significance of interaction terms. We added a few interaction terms in the model to see if there is any improvement in terms of R-squared.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.817         0.816 0.242     2994.       0    13  -10.1  50.2  156.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

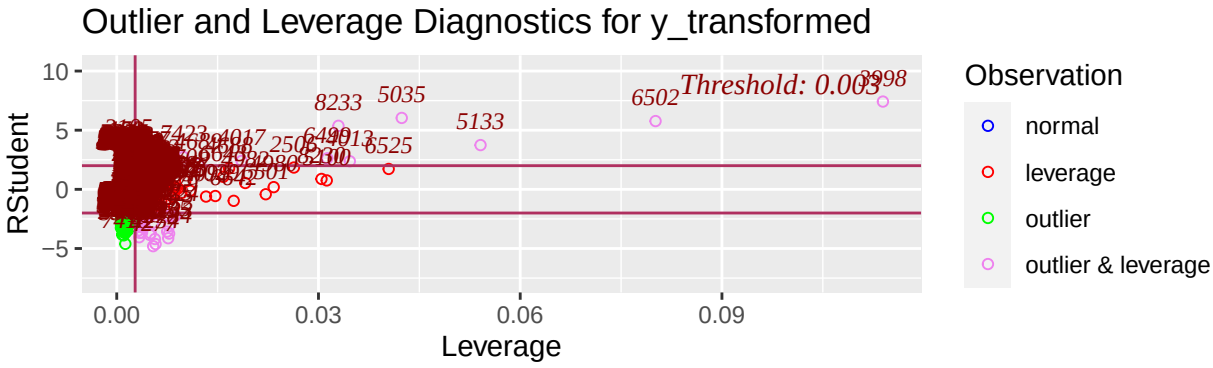
Based on the table above, the R-squared remained unchanged with multiple interaction terms added. Therefore, we concluded that it's an efficient decision to choose the main-effect only model over the main-effect and interactions model.

4. Model Diagnostics

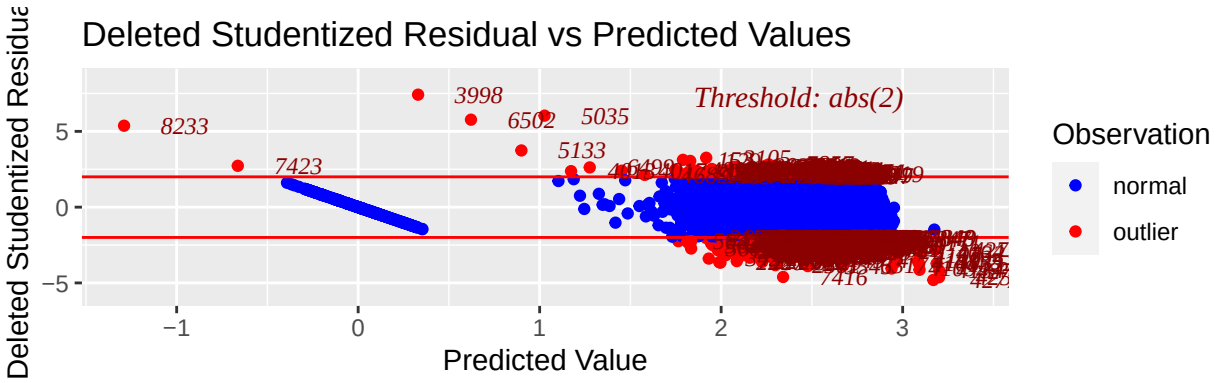


In this section, we tested the main assumptions of a linear model which are equal variance and normality of errors. As we can see from the plots above, the dataset which constructed the current final model has some violations with the assumptions of the model. The residual plot has two significantly different patterns while the normality QQ line has some deviations on both ends. Therefore, we need to clean some data on our dataset to mitigate these issues.

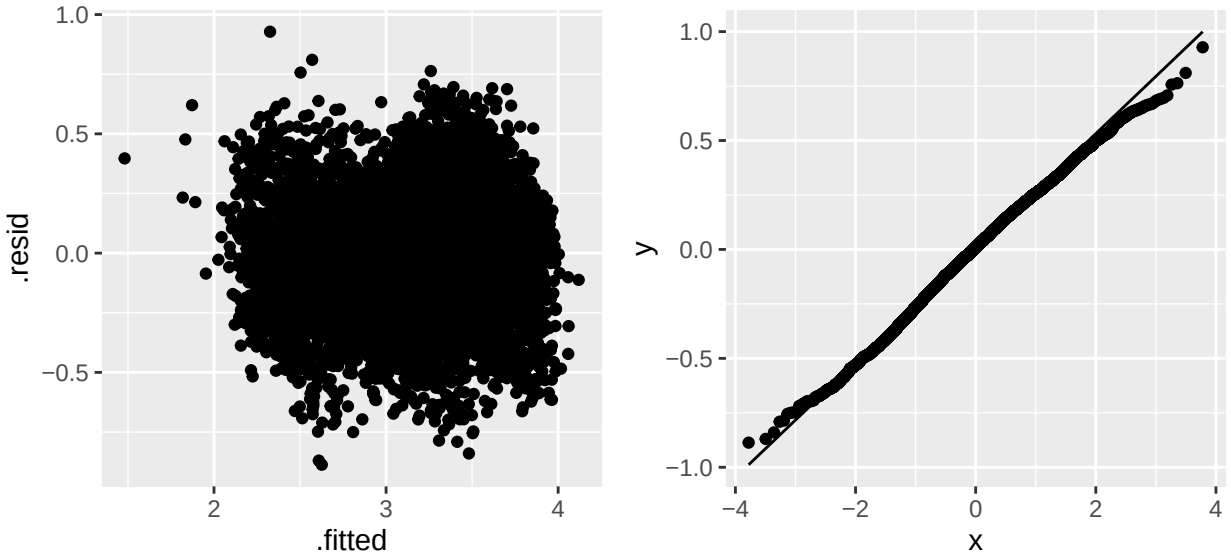
We used Leverage and Studentized Deleted Residuals to identify outliers from the model which are shown below.



Based on the residual leverage plot, we've identified the outliers from the dataset.



Based on Studentized Deleted Residual measure, we've also identified the outliers from the dataset. The following shows what happens when we remove the outliers from the dataset and re-fit the model.



Based on the new residual plot and new normal QQ line plot, it's clear that we've successfully fixed the violations of the linear regression assumptions with the residual plot and the normal QQ line. The new residual plot no longer has the straight line pattern on the left and most of the points are scattered around 0 without any specific pattern. For the normal QQ line, the end-behaviors have been mitigated causing the shape of the line look more straight.

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.748         0.747 0.259     1710.       0    11  -428.  883.  971.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Based on the summary of the new fit model, the R-squared of the new model without outliers reduced to 74% from 81%. However, we still believe that 74% R-squared is still high enough to justify removing the outliers and preserving the linear regression assumptions. Therefore, our remedial actions can be justified.

6. Conclusion

From this project, we've learned that the main-effect only model with just Box-Cox transformation is still good enough to predict the number of rental bikes from the weather data with the average R-squared of 78%. Also, quantitative variables such as `Humid`, `WindSpeed`, `Visibility`, `Dew Point Temperature`, `Rainfall`, `Seasons`, `Holiday` and `Functional Day` are significant predictors of the number of rental bikes in Seoul. In order to make sure our linear model is a valid model, we needed to check the core assumptions of the model which are equal variance of the errors and the normality of the errors. In order to correct the assumptions, we used different kinds of measures learned from the lectures to identify outlying and influential observations. With proper modification of the dataset, we were able to ensure the model assumptions and make the model more reliable for its purposes.

In terms of limitations, a linear model usually works well when the dataset has reasonable linear relationship between a model's response variable and its explanatory variables. When the dataset doesn't have a clear linear relation, it's difficult to see meaningful results from applying linear regression techniques. Of course, we can utilize advanced data cleansing techniques and feature engineering to make the dataset more suitable for linear models but sometimes it's more effective to apply non-linear prediction techniques or machine learning algorithms which may be more appropriate. In the real world, practitioners in the field are equipped with advanced regression techniques for various types of dataset and nowadays machine learning is one of the hottest fields for generating useful insights from big data.

7. References

GGPLOT2 : Quick correlation matrix heatmap - R software and Data Visualization. STHDA. (n.d.). Retrieved April 7, 2022, from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

Gov, E. (n.d.). Solar radiation basics. Energy.gov. Retrieved April 7, 2022, from https://www.energy.gov/eere/solar/solar-radiation-basics

Colier, N. (n.d.) Replace values in R, "yes" to 1 and "no" to 0. Retrieved April 7, 2022, from https://stackoverflow.com/questions/43986118/replace-values-in-r-yes-to-1-and-no-to-0