Seoul Biking Sharing Regression Analysis

Caleb Tran, Anthony Chen, Jacob Bianchi, Aryan Sunkersett, Janet Yu

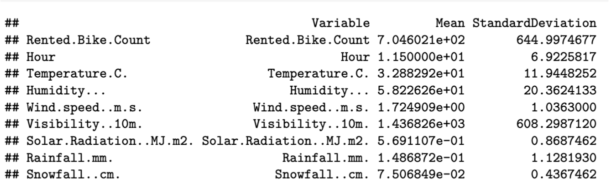
Introduction

With our project, we're looking to explore how various weather conditions and seasons affect the demand for rental bikes in the Seoul Bike Sharing System from 2017-2018. Our group sourced our data from the UC Irvine Machine Learning Repository per the Seoul Bike Sharing System. We chose to model our relationship between Number of Rented Bikes (as the response variable) and {Rented Bike Count, Hour, Temperature (°C), Humidity, Wind Speed (m/s), Visibility (10m), Solar Radiation (MJ/m2), Rainfall (mm), Snowfall (cm)}, and investigated relationships using various transformation methods and variable selection to decide on a final best-fit model. It should be noted that our dataset consists of time series data, which implies that predictors are correlated despite our model assumption of independent predictors. Our paper will dive into more about our data, a detailed run-through of our method selection methodology, our final model, possible limitations, and real world applications of our findings.

Data Description

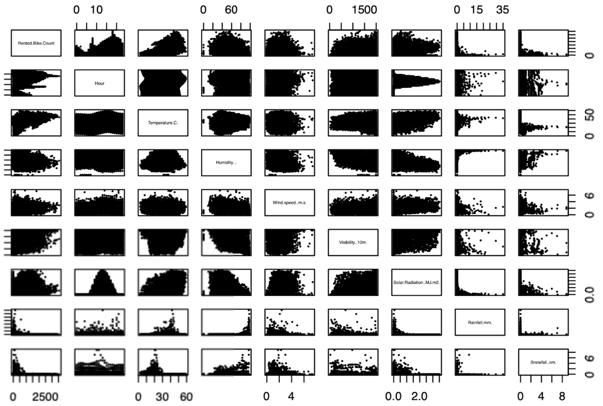
Below are some summary statistics and scatterplot matrix of our variables:

Mean & Standard Deviations:

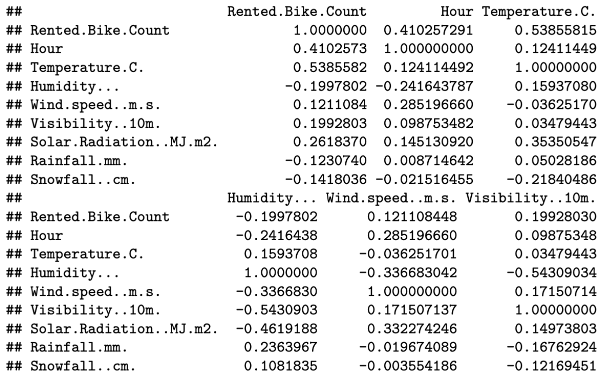


```
##                                           Variable         Mean StandardDeviation
## Rented.Bike.Count                 Rented.Bike.Count 7.046021e+02       644.9974677
## Hour                                           Hour 1.150000e+01         6.9225817
## Temperature.C.                       Temperature.C. 3.288292e+01        11.9448252
## Humidity...                             Humidity... 5.822626e+01        20.3624133
## Wind.speed..m.s.                   Wind.speed..m.s. 1.724909e+00         1.0363000
## Visibility..10m.                   Visibility..10m. 1.436826e+03       608.2987120
## Solar.Radiation..MJ.m2. Solar.Radiation..MJ.m2. 5.691107e-01         0.8687462
## Rainfall.mm.                           Rainfall.mm. 1.486872e-01         1.1281930
## Snowfall..cm.                         Snowfall..cm. 7.506849e-02         0.4367462
```

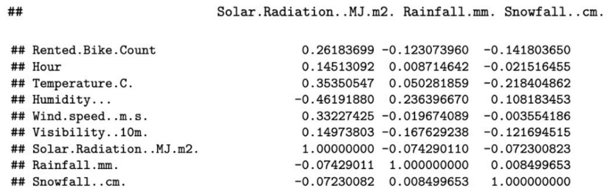
Pairs Plot:



Correlations:



```
##                         Rented.Bike.Count         Hour Temperature.C.
## Rented.Bike.Count               1.0000000  0.410257291     0.53855815
## Hour                            0.4102573  1.000000000     0.12411449
## Temperature.C.                  0.5385582  0.124114492     1.00000000
## Humidity...                    -0.1997802 -0.241643787     0.15937080
## Wind.speed..m.s.                0.1211084  0.285196660    -0.03625170
## Visibility..10m.                0.1992803  0.098753482     0.03479443
## Solar.Radiation..MJ.m2.         0.2618370  0.145130920     0.35350547
## Rainfall.mm.                   -0.1230740  0.008714642     0.05028186
## Snowfall..cm.                  -0.1418036 -0.021516455    -0.21840486
##                         Humidity... Wind.speed..m.s. Visibility..10m.
## Rented.Bike.Count        -0.1997802      0.121108448       0.19928030
## Hour                     -0.2416438      0.285196660       0.09875348
## Temperature.C.            0.1593708     -0.036251701       0.03479443
## Humidity...               1.0000000     -0.336683042      -0.54309034
## Wind.speed..m.s.         -0.3366830      1.000000000       0.17150714
## Visibility..10m.         -0.5430903      0.171507137       1.00000000
## Solar.Radiation..MJ.m2.  -0.4619188      0.332274246       0.14973803
## Rainfall.mm.              0.2363967     -0.019674089      -0.16762924
## Snowfall..cm.             0.1081835     -0.003554186      -0.12169451
```



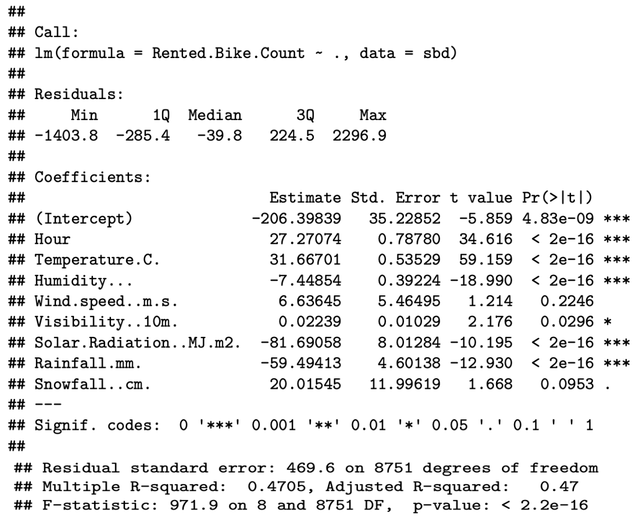
```
##                         Solar.Radiation..MJ.m2. Rainfall.mm. Snowfall..cm.

## Rented.Bike.Count                    0.26183699 -0.123073960  -0.141803650
## Hour                                 0.14513092  0.008714642  -0.021516455
## Temperature.C.                       0.35350547  0.050281859  -0.218404862
## Humidity...                         -0.46191880  0.236396670   0.108183453
## Wind.speed..m.s.                     0.33227425 -0.019674089  -0.003554186
## Visibility..10m.                     0.14973803 -0.167629238  -0.121694515
## Solar.Radiation..MJ.m2.              1.00000000 -0.074290110  -0.072300823
## Rainfall.mm.                        -0.07429011  1.000000000   0.008499653
## Snowfall..cm.                       -0.07230082  0.008499653   1.000000000
```

Results and Interpretation

First, we fit the full model, untransformed, with multiple linear regression.

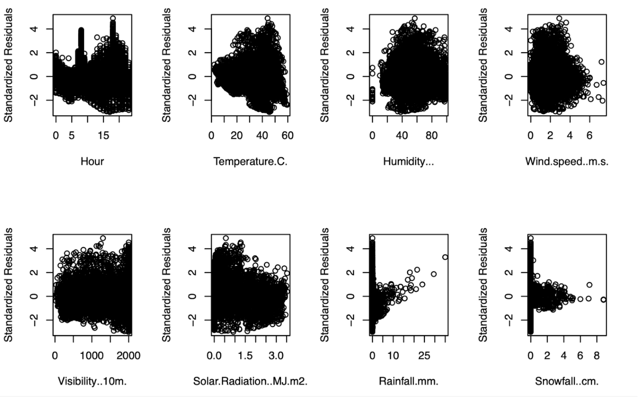
Summary Output for Full Model:

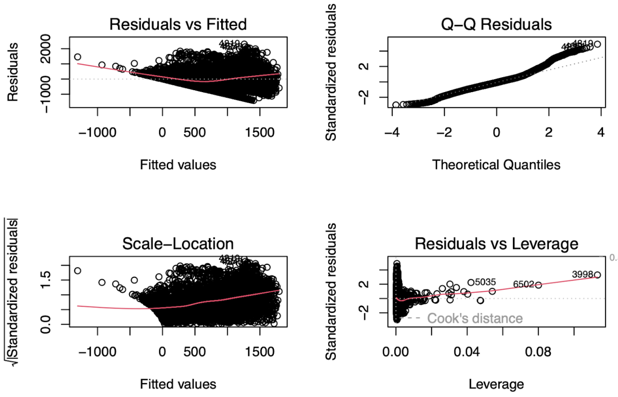


```
##
## Call:
## lm(formula = Rented.Bike.Count ~ ., data = sbd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1403.8  -285.4   -39.8   224.5  2296.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -206.39839   35.22852  -5.859 4.83e-09 ***
## Hour                     27.27074    0.78780  34.616  < 2e-16 ***
## Temperature.C.           31.66701    0.53529  59.159  < 2e-16 ***
## Humidity...              -7.44854    0.39224 -18.990  < 2e-16 ***
## Wind.speed..m.s.          6.63645    5.46495   1.214   0.2246
## Visibility..10m.          0.02239    0.01029   2.176   0.0296 *
## Solar.Radiation..MJ.m2. -81.69058    8.01284 -10.195  < 2e-16 ***
## Rainfall.mm.            -59.49413    4.60138 -12.930  < 2e-16 ***
## Snowfall..cm.            20.01545   11.99619   1.668   0.0953 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 469.6 on 8751 degrees of freedom
## Multiple R-squared:  0.4705, Adjusted R-squared:   0.47
## F-statistic: 971.9 on 8 and 8751 DF,  p-value: < 2.2e-16
```

From the scatterplot matrix, we can observe some relationships between Rented Bike Count and some predictors, such as Temperature and Hour. We can also see some correlation between some pairs of predictors, indicating that multicollinearity may be present and variable selection may be required later on. From the model summary, we noticed that the Adjusted R^2 is low (0.47), which is an indicator of a bad fit of the model. However, we see that the majority of the predictors are significant (with the exception of Wind Speed and Snowfall), and the overall F-test indicates that at least one of the slopes are significant.

To check the model's validity and see if a transformation is necessary, we also look at the residual plots for each predictor and the diagnostic plots:





Most of the residual plots seem to show a random scatter around the horizontal axis, which implies that the linearity assumption is roughly satisfied with the exception of the Rainfall predictor. From the diagnostic plots, 1. the variance seems to have an increasing trend, 2. the normality of the errors is slightly tailed, and 3. there are leverage points (e.g. case 3998 and 6502) that need to be investigated. These two hour counts correspond to dates that are not special holidays in Korea, there is nothing really unique about these leverage points. We argue that the model assumptions are somewhat violated overall, so we'll explore a transformed version of the model to see if the model will improve, but we will keep this full, non-transformed model in mind.

Here is the regression equation before we apply any transformations/model selection:

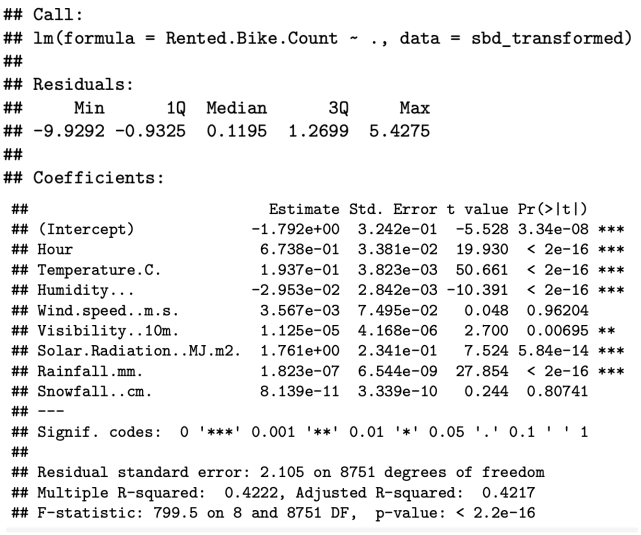
$$\begin{gathered}
Rented.Bike.Count = -206.29839 + (27.27074 \times \text{Hour}) + (31.66701 \times \text{Temperature}) + (-7.44854 \times \text{Humidity}) \\
+ (6.63645 \times \text{Wind Speed}) + (0.02239 \times \text{Visibility}) + (-81.69058 \times \text{Solar Radiation}) \\
+ (-59.59413 \times \text{Rainfall}) + (20.01545 \times \text{Snowfall})
\end{gathered}$$

Interpretation of Slopes:

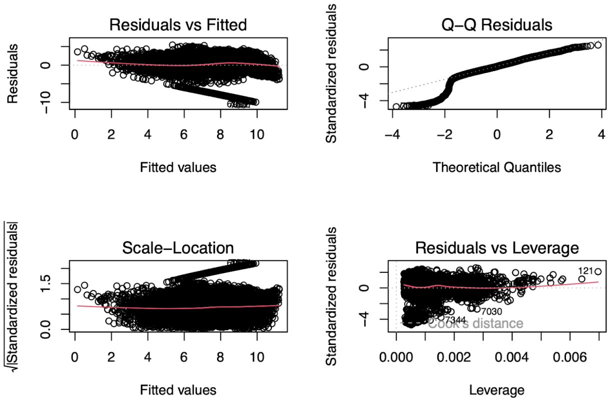
Hour	For every hour progression throughout the day, the rented bike count is predicted to increase by 27.27.
Temperature	For every degree Celsius increase in temperature, the rented bike count is predicted to increase by 31.67.
Humidity	For every unit increase in humidity, the rented bike count is predicted to decrease by 7.44.
Wind Speed	For every unit increase in wind speed, the rented bike count is predicted to increase by 6.63.
Visibility	For every unit increase in visibility, the rented bike count is predicted to increase by 0.0223
Solar Radiation	For every unit increase in solar radiation, the rented bike count is predicted to decrease by 81.69
Rainfall	For every unit increase in rainfall, the rented bike count is predicted to decrease by 59.59
Snowfall	For every unit increase in snowfall, the rented bike count is predicted to increase by 20.02.

Now, we want to test a transformed model to see if it yields improvements to the model assumptions that we discussed.

Therefore, we created a transformed model using PowerTransform (Box-Cox) to try to get the best powers for each of the variables, transforming the response variable and predictors simultaneously. After the PowerTransform function, we applied the suggested lambdas to the variables in the dataset and got a different result. Here is the summary & diagnostic plots:

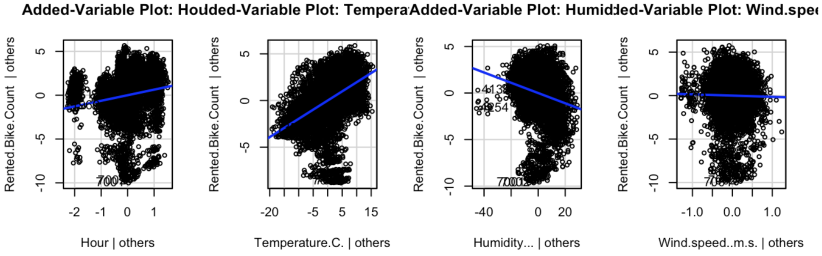


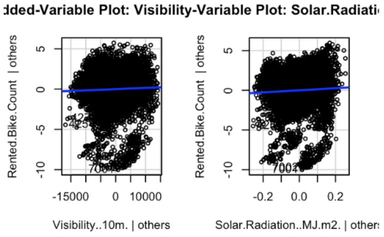
```
## Call:
## lm(formula = Rented.Bike.Count ~ ., data = sbd_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9292 -0.9325  0.1195  1.2699  5.4275
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1.792e+00  3.242e-01  -5.528 3.34e-08 ***
## Hour                     6.738e-01  3.381e-02  19.930  < 2e-16 ***
## Temperature.C.           1.937e-01  3.823e-03  50.661  < 2e-16 ***
## Humidity...             -2.953e-02  2.842e-03 -10.391  < 2e-16 ***
## Wind.speed..m.s.         3.567e-03  7.495e-02   0.048  0.96204
## Visibility..10m.         1.125e-05  4.168e-06   2.700  0.00695 **
## Solar.Radiation..MJ.m2.  1.761e+00  2.341e-01   7.524 5.84e-14 ***
## Rainfall.mm.             1.823e-07  6.544e-09  27.854  < 2e-16 ***
## Snowfall..cm.            8.139e-11  3.339e-10   0.244  0.80741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.105 on 8751 degrees of freedom
## Multiple R-squared:  0.4222, Adjusted R-squared:  0.4217
## F-statistic: 799.5 on 8 and 8751 DF,  p-value: < 2.2e-16
```

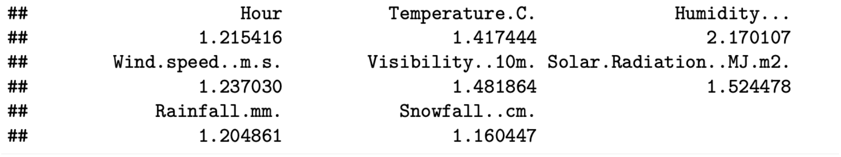


The model has Adjusted R^2 = 0.4217 which is lower than our non-transformed model, and shows that the model is a bad fit. Looking at the diagnostic plots, the Normal Q-Q plot has a heavy left tail and shows a non-normal distribution of the error terms. However, the other diagnostic plots improved with a random scatter of residual terms implying good linearity, less outliers, and a constant variance – an improvement from the non-transformed model. Overall, the transformed model seems to be mostly valid, with the only concern being the non-normality of the error terms. However, it should be noted that even after transformations, there is still a violation of the normality assumptions, suggesting that a different type of regression may be better for these predictors.

Now, with this transformed model, we shall proceed with variable selection – we noticed, from the original scatterplot matrix, that correlations existed among the predictor variables. In addition, from the added variable plots (shown below), some predictors, such as Wind speed and visibility and solar radiation, are not significant. So, although all the VIF's are less than 5 (shown below), we proceeded with backwards and forwards AIC stepwise regression.







```
##                    Hour          Temperature.C.             Humidity...
##                1.215416                1.417444                2.170107
##        Wind.speed..m.s.        Visibility..10m. Solar.Radiation..MJ.m2.
##                1.237030                1.481864                1.524478
##            Rainfall.mm.           Snowfall..cm.
##                1.204861                1.160447
```

Variable selection output:

Forwards AIC final output:

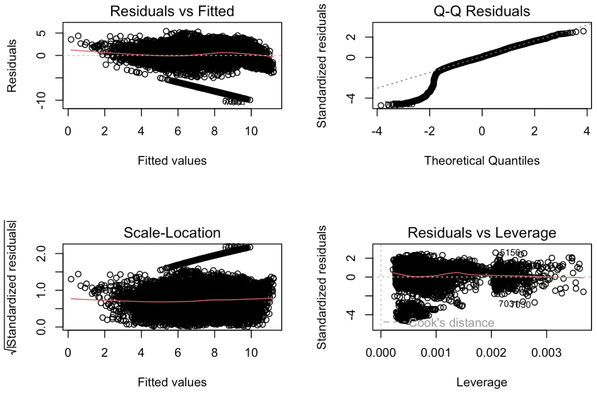
```
## Step:  AIC=13042.72
## Rented.Bike.Count ~ Temperature.C. + Rainfall.mm. + Hour + Humidity... +
##     Solar.Radiation..MJ.m2. + Visibility..10m.
##
##                    Df Sum of Sq   RSS   AIC
## <none>                          38764 13043
## + Snowfall..cm.     1  0.258620 38764 13045
## + Wind.speed..m.s.  1  0.005409 38764 13045
```

Backwards AIC final output:

```
## Step:  AIC=13042.72
## Rented.Bike.Count ~ Hour + Temperature.C. + Humidity... + Visibility..10m. +
##     Solar.Radiation..MJ.m2. + Rainfall.mm.
##
##                           Df Sum of Sq   RSS   AIC
## <none>                                 38764 13043
## - Visibility..10m.         1      32.3 38796 13048
## - Solar.Radiation..MJ.m2.  1     260.3 39024 13099
## - Humidity...              1     505.8 39270 13154
## - Hour                     1    1800.8 40565 13438
## - Rainfall.mm.             1    3454.4 42218 13788
## - Temperature.C.           1   13088.0 51852 15589
```

Thus, after performing these variable selection methods, it suggested we remove the Snowfall and Wind Speed variables. Let's take another look at the summary of our final transformed model after variable selection:

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ Hour + Temperature.C. + Humidity... +
##     Visibility..10m. + Solar.Radiation..MJ.m2. + Rainfall.mm.,
##     data = sbd_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9306 -0.9327  0.1199  1.2690  5.4252
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.762e+00  2.960e-01  -5.953 2.74e-09 ***
## Hour                    6.738e-01  3.341e-02  20.165  < 2e-16 ***
## Temperature.C.          1.940e-01  3.568e-03  54.363  < 2e-16 ***
## Humidity...            -2.963e-02  2.773e-03 -10.687  < 2e-16 ***
## Visibility..10m.        1.126e-05  4.165e-06   2.703  0.00689 **
## Solar.Radiation..MJ.m2.  1.759e+00  2.294e-01   7.666 1.97e-14 ***
## Rainfall.mm.            1.822e-07  6.524e-09  27.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.104 on 8753 degrees of freedom
## Multiple R-squared:  0.4222, Adjusted R-squared:  0.4218
## F-statistic:  1066 on 6 and 8753 DF,  p-value: < 2.2e-16
```



The diagnostic plots look pretty similar to prior variable selection–the normality assumption is still unfortunately violated, but the linearity and constant variance assumptions are satisfied, indicating an improvement. It is also worth noting that, because of the variable selection, all of the predictors are significant.

So, here is the regression equation of our final model:

$$
\begin{gathered}
\text{Rented.B}\hat{\text{i}}\text{ke.Count}^{0.32} = -1.762 + (6.738 \times 10^{-1} \times \text{Hour}^{0.36}) + (1.940 \times 10^{-1} \times \text{Temperature}^{0.88}) \\
+ (-2.953 \times 10^{-2} \times \text{Humidity}^{0.89}) + (1.126 \times 10^{-5} \times \text{Visibility}^{1.30}) \\
+ (1.759 \times \text{Solar Radiation}^{0.01}) + (1.822 \times 10^{-7} \times \text{Rainfall}^{-0.60})
\end{gathered}
$$

Interpretation of Slopes:

Hour	For every Hour$^{0.32}$ progression throughout the day, the rented bike count$^{0.32}$ is predicted to increase by 6.738e-1.
Temperature	For every degree Celsius$^{0.88}$ increase in temperature, the rented bike count$^{0.32}$ is predicted to increase by 1.940e-1.
Humidity	For every unit increase in Humidity$^{0.89}$, the rented bike count$^{0.32}$ is predicted to decrease by 2.953e-2.
Visibility	For every unit increase in Visibility$^{1.30}$, the rented bike count$^{0.32}$ is predicted to increase by 0.0223
Solar Radiation	For every unit increase in solar radiation$^{0.01}$, the rented bike count$^{0.32}$ is predicted to decrease by 81.69
Rainfall	For every unit increase in rainfall$^{-0.60}$, the rented bike count$^{0.32}$ is predicted to increase by 1.822e-7

Discussion

In this project, we aimed to investigate the relationship between demand of the Seoul Public Bike Sharing System and various weather conditions. Our final model is a regression transformed by the box-cox method, and additionally processed through backwards AIC variable selection in order to improve the many violated assumptions in the full model.

In a real world situation, our model could be useful for consumers hoping to use the Seoul public bike sharing system. They could use our model to assess the demand for bikes at the time they want to use one. If the predicted rented bike count is high, the consumer might want to wait until a better time when there will be bikes available. This model could also be used for managers of the bike sharing system, as well. They can use the model to understand at what times and in what conditions their system has the most use. During busier times, they might want to raise the price to properly align with demand. Similarly, when demand is low, they could lower the price so that more people use the system.

The biggest limitation of our model is, due to our time series data, our variables are not independent. This violates a major assumption of linear modeling, and can lead to the failure of other assumptions such as constant variance or normality. In addition, our final model violates the assumption of normality and has slight patterning in the residual diagnostic plots. This could be further investigated to determine what exactly is causing the abnormal left tail on the Normal Q-Q plot. Finally, the coefficients of the transformed model are difficult to interpret. The underlying meaning of the coefficients is unintuitive and difficult to understand. This could be improved by either not using a transformed model, or attempting to solely look at whether a variable leads to an increase or decrease in rented bike count. In summary, our project provided valuable insights into the relationship between the rented bike demand and various weather conditions in the Seoul Bike Sharing System.