# STAT 408 Final Paper - Bike Sharing

**By Rolando Santos, Sathvik Maridasana Nagaraj and Aaron Myrold**

## Introduction

### Dataset Source

Our data was found from https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset. This data provides 2 files for data, we will be using the 'day.csv' since every entry is an individual day compared to every hour in the 'hour.csv'.

### Background

The dataset contains the hourly and daily count of rental bikes between the years 2011 and 2012 in a Capital bikeshare system with the corresponding weather and seasonal information.

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, users are able to easily rent a bike from a particular position and return back at another position. As of August 2021, there are about over 3000 bike-sharing programs around the world which is composed of over 10 million bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns the bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

With this dataset we want to predict the number of rental bikes would be needed in a given day based on environmental and calendar factors. This would allow the Capital bike sharing system to optimize expected bike maintenance and traffic.

### Research Question

Can we create a model that will allow us to predict the number of bikes needed on a specific day based on calendar and environmental factors?

**Used Variables**

| Name | Description | Type | Notes |
|---|---|---|---|
| season | Season | Categorical | 1: Winter, 2: Spring, 3: Summer, 4: Fall |
| year | Year 2011 or 2012 | Binary | 0: 2011, 1: 2012 |
| mnth | Month | Categorical | 1: Jan - 12: Dec |
| holiday | Whether day is holiday or not | Binary | 0: No, 1: Yes |
| weekday | Day of the week | Categorical | 0: Sunday - 6: Saturday |
| workinday | Whether day is neither weekend nor holiday | Binary | 0: No, 1: Yes |
| weathersit | Weather Situation | Categorical | 1: Clear, 2: Misty, 3: Snow/Rain |
| temp | Normalized Temperature in Celcius | Numeric (Float) | Between 0 and 1 |
| atemp | Normalized Feeling Temperature in Celcius | Numeric (Float) | Between 0 and 1 |
| windspeed | Normalized Wind Speed | Numeric (Float) | Between 0 and 1 |
| cnt | Count of total rental bikes including both casual and registered users | Numeric (Integer) | Response Variable |

**Unused Variables**

| Name | Description | Type | Notes |
|---|---|---|---|
| instant | Record Index | Numeric (Integer) | |
| dteday | Date | DateTime | |
| casual | Count of Casual Users | Numeric (Integer) | |
| registered | Count of Registered Users | Numeric (Integer) | |

The reason why `casual` and `registered` are not used is because they are too closely related to the response `cnt` due to `cnt` being the sum of both variables and can cause issues with our model due to the extremely high correlation.

Each group member will try to fit several models with the price as the response, and a combination of the remaining variables as predictors. The best model fit by each member will be listed below, and will be discussed in the `Results` and `Discussion` sections that follow.

# Methods

The dataset was loaded, there were 731 observations and 16 variables present. The 731 observations each represented a day from the year 2011 (365 days) and 2012 (366 days). The dataset was also confirmed to have not contained any missing variables. The variables `instant`, `dteday`, `casual` and `registered` were removed, leaving the dataset with 12 variables total, 11 predictors and 1 response variable.

**Helper Functions**

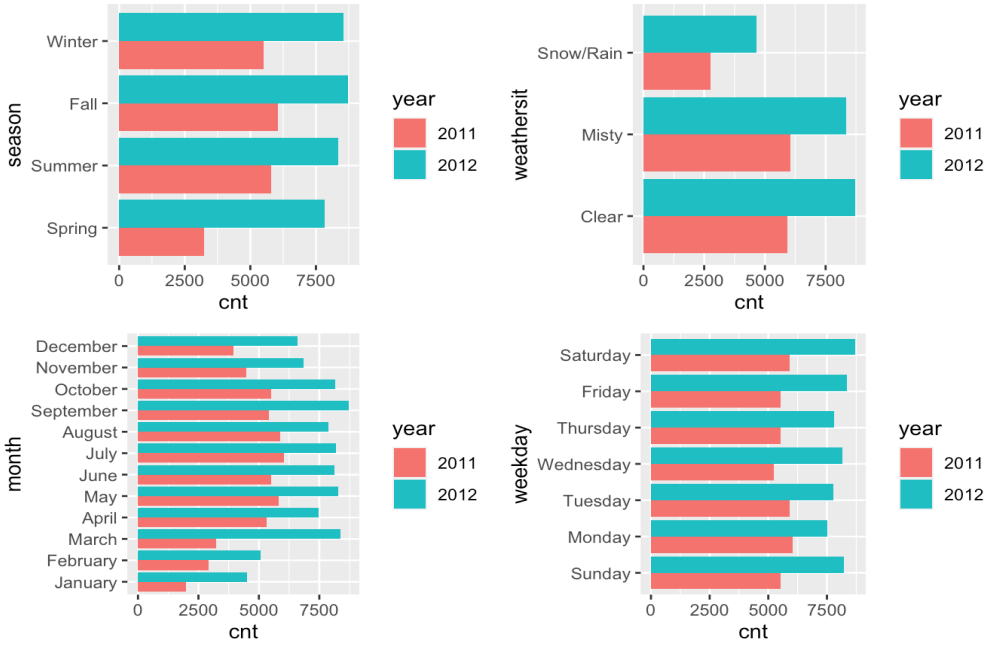
The following are helper functions used repeatedly throughout the report:

- plot_diagnostics: Creates a side-by-side fitted vs. residuals plot and a q-q plot with a linear model as the input
- create_bp: Creates a bar plot using the bike sharing dataset, with the y-value being `cnt` and a provided categorical variable as the x-value. The color of the bar plots can be set as well, with `year` as the default value.

- create_sp: Creates a scatter plot using the bike sharing dataset, with the y-value being `cnt` and a provided numeric variable as the x-value. The color of the scatter plots can be set as well, with `year` as the default value.
- get_nrmse: Given a model and a dataset, the nrmse will be calculated using the `cnt` response values from the dataset and the predicted values using the model
- get_r_squared: Given a model, fetch the r-squared value
- get_max_vif: Given a model, fetch the highest VIF value
- get_f_score: Given two models, fetch the anova f-test score
- get_p_value: Given two models, fetch the anova p-value

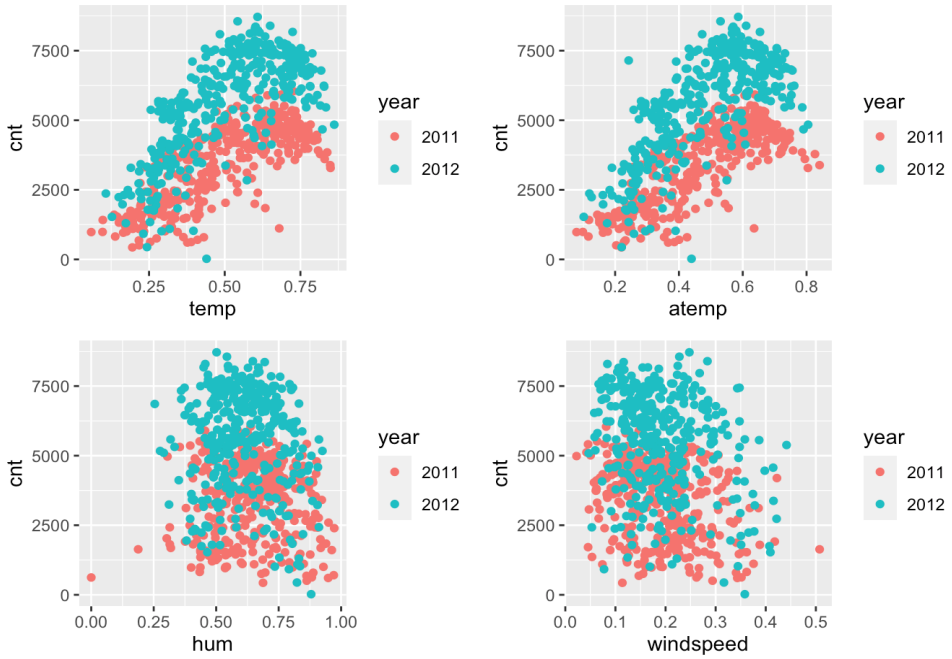
**Investigation and Variable Relationships**

First we will map our categorical variables to their respective descriptions for clarity. The following plots were created to show the relationship between some of our variables and the response.




From our graphs we can see that there was a large increase in bike sharing users in 2012 overall. Fall appears to be the most popular season for bike sharing, whereas spring appears to be the least popular. Understandably, clear and misty weather was the preferred weather over snow/rain. January and February are the least popular months to ride with July being the most popular in

2011 and September being the most popular in 2012. There doesn't appear to be any significant differences between the number of users based on the day of the week.




From our scatterplots, there appears to be a linear relationship between `temp` and `cnt` as well as between `atemp` and `cnt`. However, there is no clear relationship between `hum` and `cnt`, nor `windspeed` and `cnt`.

To prepare our dataset, we converted our categorical variables to dummy variables. To ensure collinearity would not be an issue, the first dummy variable was removed, i.e. `season_Spring` was removed and `season_Summer`, `season_Fall` and `season_Winter` remained. We then performed an 80:20 split on our dataset using random sampling, with 80% of our data used for training our models and 20% used for testing and result metrics.

# Models

### Base Model

### Model 1: AIC

Our first attempt at improving the base model was the use of AIC. After making some slight changes to the base model, we used the step() function to determine which predictors to remove from further analysis. The predictors that were deemed insignificant were: atemp, month_February, month_August, month_November, and month_December. First, atemp makes the most sense as it is highly correlated with temp – we only need one of them. Next, the various months are ones that seem fairly representative of their seasons. For the most part, any variation

in ridership could likely be quantified and modeled from the information in other variables like season_XXX or weathersit_XXX.

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1701.47     256.10   6.644 7.26e-11 ***
## temp                      4749.56     374.63  12.678  < 2e-16 ***
## hum                      -1750.02     317.32  -5.515 5.33e-08 ***
## windspeed                -3071.66     450.37  -6.820 2.36e-11 ***
## season_Summer              793.07     181.55   4.368 1.49e-05 ***
## season_Fall               1033.30     171.77   6.016 3.24e-09 ***
## season_Winter             1431.39     116.21  12.317  < 2e-16 ***
## year_2012                 2034.93      64.25  31.673  < 2e-16 ***
## month_March                479.04     138.25   3.465 0.000571 ***
## month_April                467.47     212.44   2.200 0.028183 *
## month_May                  738.25     213.68   3.455 0.000592 ***
## month_June                 289.79     189.35   1.530 0.126477
## month_July                -356.25     149.41  -2.384 0.017444 *
## month_September            786.46     129.93   6.053 2.60e-09 ***
## month_October              518.24     137.73   3.763 0.000186 ***
## holiday_Yes               -636.17     204.51  -3.111 0.001962 **
## weekday_Monday             295.26     126.00   2.343 0.019458 *
## weekday_Tuesday            290.19     118.24   2.454 0.014421 *
## weekday_Wednesday          359.63     118.35   3.039 0.002487 **
## weekday_Thursday           349.24     119.46   2.924 0.003601 **
## weekday_Friday             372.52     119.14   3.127 0.001860 **
## weekday_Saturday           500.18     117.11   4.271 2.29e-05 ***
## weathersit_Misty          -477.94      85.20  -5.610 3.19e-08 ***
## `weathersit_Snow/Rain` -1959.99     211.08  -9.285  < 2e-16 ***
## ---
```

**Model 2: Transformation**

For the 2nd model we choose to apply transformation. Logarithmic transformations are applied to the temperature (log_temp), humidity (log_hum), and windspeed (log_windspeed) variables to capture potential non-linear relationships between these predictors and the response variable.This can lead to a more accurate representation of the underlying patterns in the data.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.2358 -0.0860  0.0277  0.1270  0.9224

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           8.44642    0.15161  55.713  < 2e-16 ***
log_temp              0.66686    0.06340  10.518  < 2e-16 ***
log(1 + hum)         -0.79031    0.19268  -4.102 4.72e-05 ***
log_windspeed        -0.15972    0.03048  -5.240 2.28e-07 ***
season_Summer         0.24688    0.07518   3.284 0.001089 **
season_Fall           0.39490    0.09066   4.356 1.58e-05 ***
season_Winter         0.53907    0.07886   6.836 2.15e-11 ***
year_2012             0.45215    0.02468  18.323  < 2e-16 ***
month_February        0.04392    0.06230   0.705 0.481105
month_March           0.05749    0.07164   0.802 0.422650
month_April           0.02650    0.10601   0.250 0.802722
month_May             0.11220    0.10990   1.021 0.307743
month_June           -0.03887    0.11419  -0.340 0.733705
month_July           -0.23107    0.12675  -1.823 0.068826 .
month_August         -0.18595    0.12370  -1.503 0.133357
month_September      -0.01581    0.11165  -0.142 0.887445
month_October        -0.17327    0.10423  -1.662 0.096993 .
month_November       -0.11567    0.10087  -1.147 0.251991
month_December       -0.11919    0.08051  -1.480 0.139317
holiday_Yes          -0.17924    0.07884  -2.273 0.023379 *
weekday_Monday        0.05433    0.04816   1.128 0.259808
weekday_Tuesday       0.07759    0.04532   1.712 0.087430 .
weekday_Wednesday     0.08519    0.04531   1.880 0.060618 .
weekday_Thursday      0.09754    0.04581   2.129 0.033659 *
weekday_Friday        0.10484    0.04551   2.304 0.021603 *
weekday_Saturday      0.11572    0.04475   2.586 0.009971 **
weathersit_Misty     -0.11629    0.03220  -3.612 0.000331 ***
`weathersit_Snow/Rain` -1.06505    0.07876 -13.522  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2903 on 556 degrees of freedom
Multiple R-squared:  0.7624,    Adjusted R-squared:  0.7509
F-statistic: 66.09 on 27 and 556 DF,  p-value: < 2.2e-16
```

**Model 3: Correlation**

In the model 3 we first identified the highly Correlated pairs. The pairs identified are between atemp and temp. We excluded the highly correlated variable (atemp).The formula for the model excludes the workingday_Yes variable, which might be a conscious choice based on prior analysis.
The response variable is cnt, and the predictors include all other variables in the dataset.
This process is an attempt to improve the model by addressing multicollinearity (correlation between predictors), specifically by excluding one variable from a highly correlated pair (atemp).

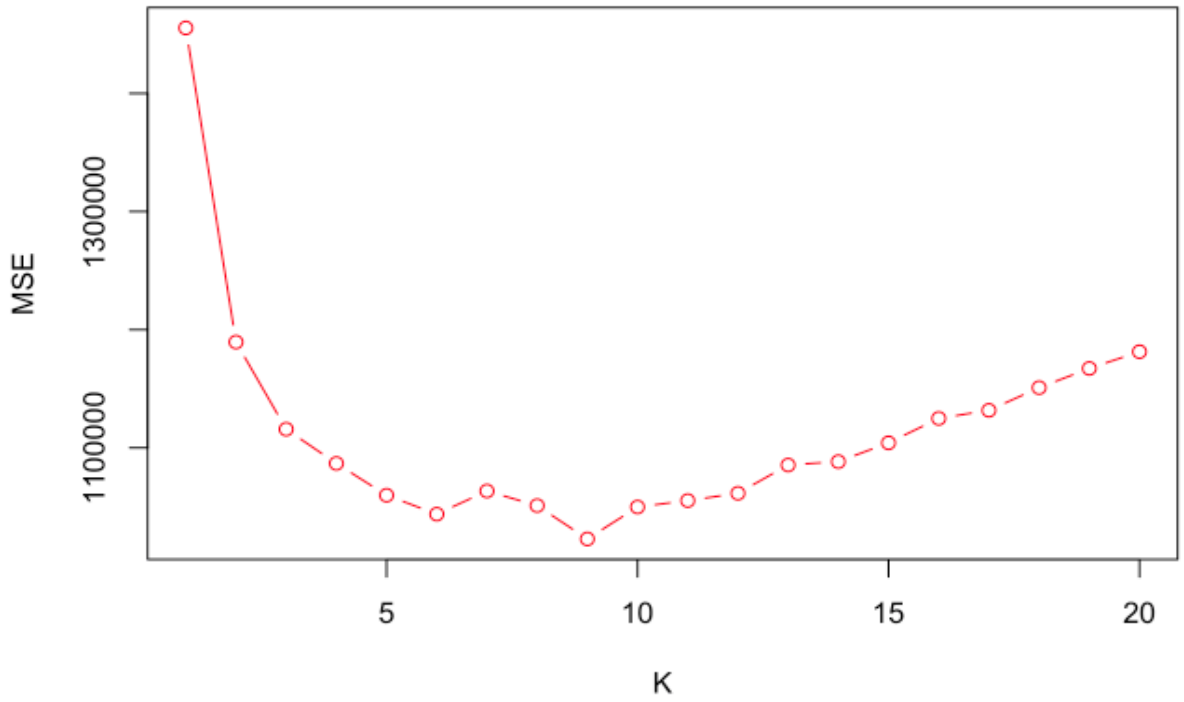
But, The adjusted R-squared is comparable to the previous models, suggesting that the exclusion did not significantly impact the model's explanatory power.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3842.2  -354.0    88.5   443.0  2841.3

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1703.38     268.46   6.345 4.61e-10 ***
temp                 4599.16     460.97   9.977  < 2e-16 ***
hum                 -1699.37     322.45  -5.270 1.95e-07 ***
windspeed           -3033.74     453.01  -6.697 5.22e-11 ***
season_Summer         803.13     197.30   4.071 5.37e-05 ***
season_Fall          1062.03     238.52   4.453 1.03e-05 ***
season_Winter        1593.33     206.81   7.704 6.08e-14 ***
year_2012            2039.18      64.65  31.541  < 2e-16 ***
month_February         62.48     161.30   0.387  0.69866
month_March           496.31     181.13   2.740  0.00634 **
month_April           486.34     272.93   1.782  0.07530 .
month_May             774.13     289.61   2.673  0.00774 **
month_June            338.35     310.84   1.088  0.27685
month_July           -311.90     349.74  -0.892  0.37289
month_August           60.76     334.93   0.181  0.85611
month_September       770.24     294.09   2.619  0.00906 **
month_October         384.27     270.08   1.423  0.15536
month_November       -171.25     261.46  -0.655  0.51276
month_December       -159.57     208.86  -0.764  0.44519
holiday_Yes          -608.52     207.00  -2.940  0.00342 **
weekday_Monday        293.98     126.44   2.325  0.02043 *
weekday_Tuesday       291.30     118.95   2.449  0.01463 *
weekday_Wednesday     360.66     118.99   3.031  0.00255 **
weekday_Thursday      352.20     120.32   2.927  0.00356 **
weekday_Friday        374.12     119.48   3.131  0.00183 **
weekday_Saturday      499.01     117.46   4.248 2.53e-05 ***
weathersit_Misty     -487.66      86.18  -5.659 2.45e-08 ***
`weathersit_Snow/Rain` -1961.24     211.97  -9.253  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 761.9 on 556 degrees of freedom
Multiple R-squared:  0.8489,    Adjusted R-squared:  0.8416
F-statistic: 115.7 on 27 and 556 DF,  p-value: < 2.2e-16
```

**Model 4: KNN + Cross Validation**

For our 4th model, we decided to go with KNN. The predictors we used were the same as the most significant ones determined in our AIC model. We began by fitting a preliminary model

with k=4. Afterwards, we added cross-validation and produced the following graph:

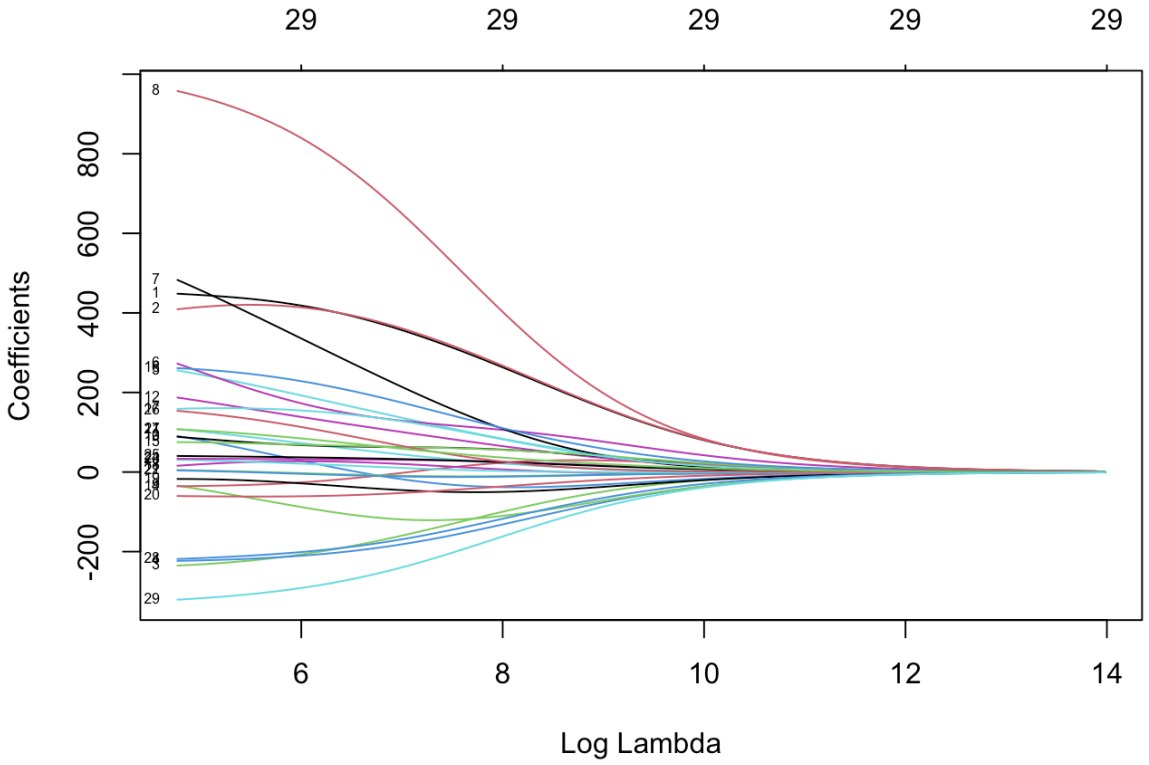



As we can see, the MSE improves up until around k=6 and stabilizes until k=10 when it slowly begins to worsen. We ended up using k=9 in our final analysis where the resulting test NRMSE was 0.199 and training NRMSE was 0.224. This model ended up being the worst performing out of the 5 we tried.

**Model 5: Ridge Regression**

Our final model was Ridge Regression. Unlike KNN, we trained this model using all predictors. We wanted to see how collinearity affected all the predictors and figured that the ridge regression would be able to handle the high dimensionality. We created this graph to show

the relative coefficients at the best lambda:

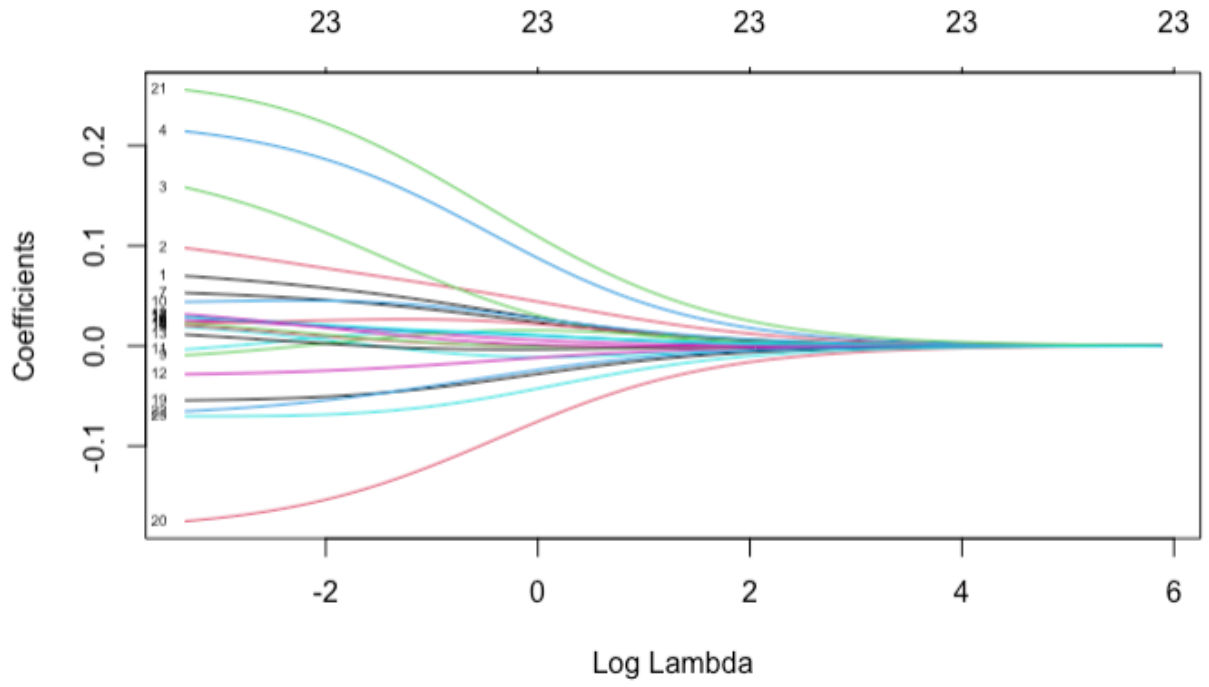



The best lamba reported by the model was 117.96 which falls between 5-6 on the Log(Lambda) axis. Here we can see that the two coefficients with the highest magnitude are year_2012 (957.5) and weather_sit (-320.8). These make intuitive sense, as in our exploratory analysis of the data, we saw significant differences between the number of riders in 2011 and 2012 as well as significant decreases in riders during rainy and snowy days. Overall, this model did a good job relative to KNN, however, given the use of all predictors it still performed worse than our first three models.

**Model 6: Transformation + Ridge Regression**

For our last model, we added transformation of y and some predictors (temp, hum, and windspeed) and additionally selected only predictors from AIC before using Ridge Regression.

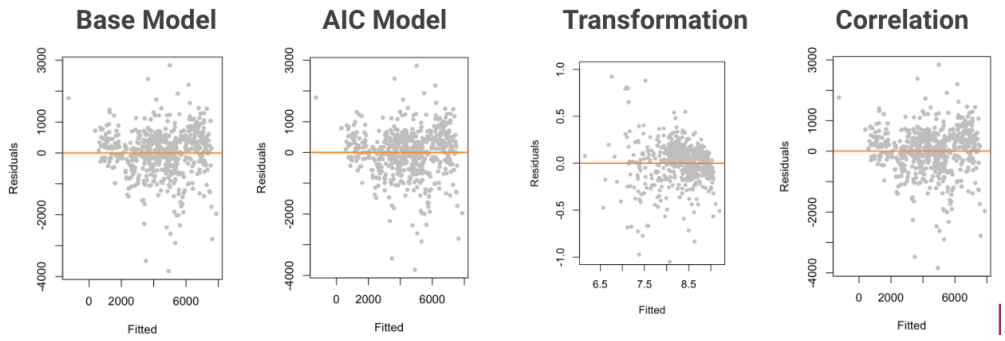
The resulting lambda graph is this:

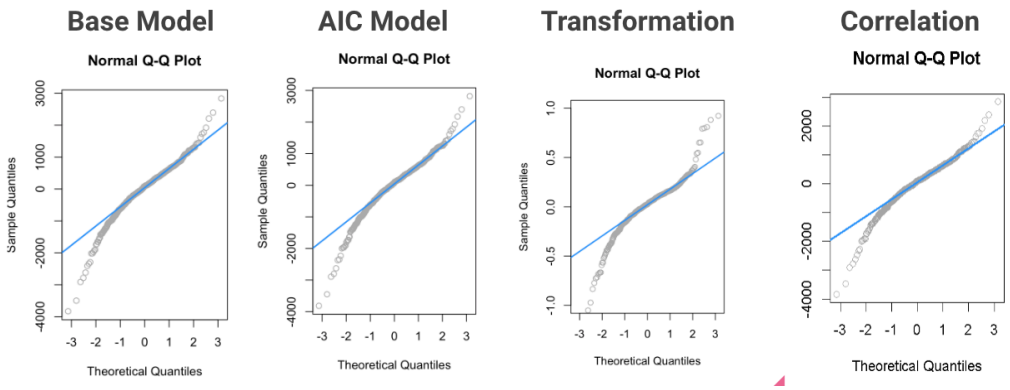



And, surprisingly, the ridge regression still performs worse than only linear regression against the transformed data. Regardless, transforming the “cnt” or “y” clearly made the most significant improvement out of all our models and tests. Using the transformed data took our ridge regression NRMSE from 0.22 to 0.04 – a large improvement.

# Results

### Constant Variance and Normality Check






Although this study intended to fit a good predictive model, the group members decided that constant variance and normality should be factored when considering the best model. Looking at the plot diagnostics, we notice that both the Q-Q Plots and the Fitted vs. Residual Plots were similar for the base model, AIC model and correlation model. The Fitted vs. Residuals plots indicated constant variance, however the Q-Q plots had a heavy tail skew which can indicate that there is less data at the center of the distribution. The exception appears to be the transformation model, where most of the points appear to be clumped on the right side. We can also see that there is some skewness on the both tails in the Q-Q plot for this model as well. In our KNN model we see that at k=3-4 we get an optimal MSE, however after k=9 we start seeing an increase in MSE again.

**NRMSE Comparison**

| **Model** | **Training NRMSE** | **Testing NRMSE** |
|---|---|---|
| Base Model | 0.1639217 | 0.1861117 |
| AIC Model | 0.1642360 | 0.1876196 |
| Transformation Model | 0.0341608 | 0.0319169 |
| Correlation Model | 0.1640145 | 0.1866556 |
| KNN Model | 0.1992266 | 0.2241404 |
| Ridge Regression | NA | 0.1923851 |

From our NRMSE results we can see that the transformation has the lowest NRMSE score out of all the models by a large margin. All other models seem to be close to one another, with the KNN model having the highest test NRMSE. For our log transformation NRMSE, the test set responses were also transformed in a similar manner.

**Adjusted R-Squared, VIF & P-Value**

| Model | Adjusted R-Squared | Max VIF | F-Score | P-value |
|---|---|---|---|---|
| Base Model | 0.8414820 | 70.18607 | NA | NA |
| AIC Model | 0.8422943 | 6.07074 | 0.4261423 | 0.8305423 |
| Transformation Model | 0.7509001 | 11.19322 | 21.1412523 | 0.0000053 |
| Correlation Model | 0.8415880 | 11.25029 | 0.6284737 | 0.4282545 |

Of all the models, the transformation model appears to have the lowest adjusted r-squared, whereas our AIC model has the best. The base model appears to have the highest VIF value, indicating that there is some level of collinearity present in the model.

We performed ANOVA tests to view any significance between our base model and our other regression models we noticed that for the correlation model and AIC model that neither are significant at the $p < .05$ level. Our transformation model appears to have a low p-value indicating significance in this model at the 5% level.

# Conclusion

After looking at the model results, although our transformation model has the lowest NRMSE, the adjusted r-squared score and failing constant variance can be used to argue that this model can be improved. The same can be applied to the transformation + ridge regression model. The model that followed our linear assumptions, had a reasonable NRMSE and had the best adjusted r-squared ended up being the AIC model. Thus, the group decided that building an AIC model with our bike sharing dataset would be best to predict the number of bikes needed on any given day, based on our predictors.