There were four main steps I followed in order to do a complete analysis of the training data set and determine the predictions of the test data set. The first step was Data Tidying. The second was Data Exploration where I discovered the most significant predictors to use in my models. The second step was the actual analysis where I fit several models to the data and compared each model's test error. The final step was where I determined the best model and used it to obtain predictions of the test data.

Upon loading the dataset in, I executed several data tidying techniques in order to make the data cleaner and more interpretable. I utilized the na.omit function to remove NA observations in the dataset. This made the data cleaner and less likely for the predictions to be affected by unnecessary observations. Next, I noticed that the Date data was not in an ideal format. Each date was in day/month/year format, but this would cause issues in the future because each date would be treated as a separate categorical data value. Furthermore, when conducting any analysis on the data, each unique date value would be treated as a single categorical predictor. I realized that this wouldn't be efficient and would result in too high of a number of predictors. Additionally, I also realized that these date values are not useful in their current format and that no trends or correlations can currently be made with them. Taking both of these points into account, I used the separate function to split the single Date variable into three variables-Day, Month, Year. Then, I used the as.numeric function on each of these three predictors to turn the data values from qualitative to quantitative. Converting the data to numerical in this manner allowed me to do best subset selection more efficiently without each unique Date value being treated as a separate categorical predictor. Upon doing so, I also recognized that the ID variable is simply a unique identifier for each observation and doesn't

have any influence on the actual Counts. Taking this into account, I removed the ID variable from the training and testing data sets.

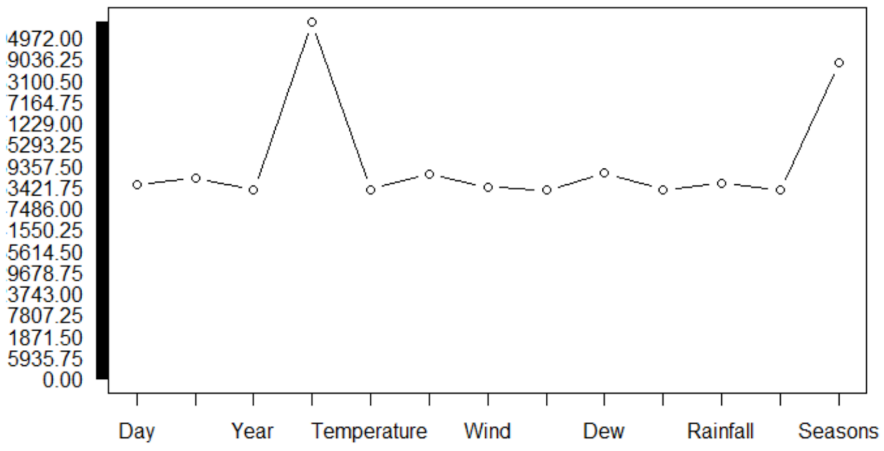
The second step done was Data Exploration. This step primarily involved using Best Subset Selection in order to determine the best set of predictors to use for linear models. Upon carrying out this analysis, I discovered that all of the variables except “Dew”, “Visibility” (and “ID”), were considered to be the subset of predictors which minimized Mallow’s cP. Consequently, I identified these predictors as the most significant. Although, I carried out best subset selection to find the subset of predictors which minimized RSS, CP, *and* BIC, I only decided to use the set of predictors which minimized the cP for my future analysis because this measure of error takes into account the model accuracy as well as the model complexity/interpretability. Furthermore, it is important to take into account model complexity and interpretability so that we can ensure that the model would be able to generalize well to our test observations we use the model to predict.

The final step-Determining the best performing Model- was the most significant step of the four. This step consisted of running almost every possible model and comparing the MSE of each model in order to find the best performing model. In total, I ran 12 different types of models and within each model type, I ran multiple models each with varying tuning parameters and predictors. Some of the models I ran included multiple linear regression, ridge, lasso, pcr, pls, regression splines, smoothing splines, trees, bagging, and random forests. From this analysis, I discovered that the Ridge model with all 15 predictors and the Lasso model with all 15 predictors (excluding ID) were the best performing models. Ridge was able to produce predictions with an MSE of 189448.4 while Lasso gave predictions with an MSE of 204254.3. One can make the prediction that because both these models use regularization and shrinkage, are high bias, and

low flexibility, the bike data may have a relatively low SNR or might have a medium to high amount of noise.

There were several methods I performed in order to gain insights on variable importance. In total, there were three main methods. The first method was previously described earlier. This was the method of best subset selection from which I discovered that Temperature was the most significant variable and that the variables - Day, Month, Year, Temperature, Rainfall, Snowfall, Seasons, Functioning, Holiday, Solar, Wind, Humidity- were the best subset of predictors to use in order to minimize the cP. The second method I utilized was Out of Bag Variable Importance using Random Forests by use of the "importance()" function. By this measure, Temperature and Hour were the most significant predictors. As can be seen by the table below, Temperature had the highest IncNodePurity of 658608257.9 while Hour had the highest %IncMSE of 212.669337. The last method I used to measure variable importance was permutation tests with random forest models. Furthermore, I carried out a permutation test where each of the 15 predictors were shuffled, a new random forest model was created with the shuffled predictor, and the MSE was calculated for each of these models. As can be seen in the plot below, the variables-Temperature and Seasons- were the most significant by this method because the MSE was at its peak when either the variables Temperature or Seasons were shuffled in the model. Across all three of these methods, the results were pretty consistent in that Temperature was viewed as one of the most significant predictors by all three measures.

```
               %IncMSE IncNodePurity
Day          29.271643    30284747.2
Month        62.761784    43053051.6
Year          8.247905      726863.1
Hour        212.669337   625685740.2
Temperature 162.476437   658608257.9
Humidity     62.859781   129868634.7
Wind         24.575276    30518017.0
Visibility   29.810278    29900590.9
Dew          57.504819    64985434.5
Solar        92.232178   180885092.0
Rainfall     61.644123   133514469.8
Snowfall     19.202067     2416454.1
Seasons      55.042055    50604459.5
Holiday      20.118755     4642319.7
Functioning 156.948385   201354096.0
```



One of the most challenging aspects of this dataset was handling all of the different variables, deciding what the best set of variables to use for each model would be, and identifying each situation as one where it would be best to use all predictors or only the most significant ones. I addressed this difficulty by identifying the best subset of predictors to use in the very beginning though best subset selection. And moving forward, rather than making a decision to use all or only the best subset of predictors for each model, I opted to try both ways for each model and simply compare the MSE's. By carrying out every model with multiple sets of predictors in this manner, I was able to better ensure that I would find the most optimal model with the lowest MSE. A more specific challenge I encountered was understanding how to handle

the Date variable. In the very beginning, I left the variable completely unchanged and attempted best subset selection with this variable included as one of the predictors. However, because the Date variable was initially categorical, each unique Date value was treated as a separate predictor which made the analysis very inefficient and confusing since the number of predictors was now in the thousands. To combat this issue, I decided to remove the Date variable from my analysis altogether. However, I soon realized that even though the variable is in this unusable format, I can still extract valuable information from it and change the formatting in order to make it usable. This is when I used data wrangling to separate the Date variable into the three variables-Day, Month, Year-, converted them to numeric variables and was then able to perform all the needed analysis while including the Date data.

To conclude, I do feel fairly confident that my model is the best or is close to the best model because I tested the performance of every possible model where each model was also tested with different conditions(different tuning parameters and predictors). Additionally, I created a null model where the MSE of my optimal model was much less than the MSE of the null model. In order to improve my predictions further, it would be helpful to know where exactly this data is coming from and how accurate the data is. This would help us understand the SNR of the data which would help in better understanding if a more flexible or a more biased model would perform better and/or if shrinkage or regularization would be beneficial.