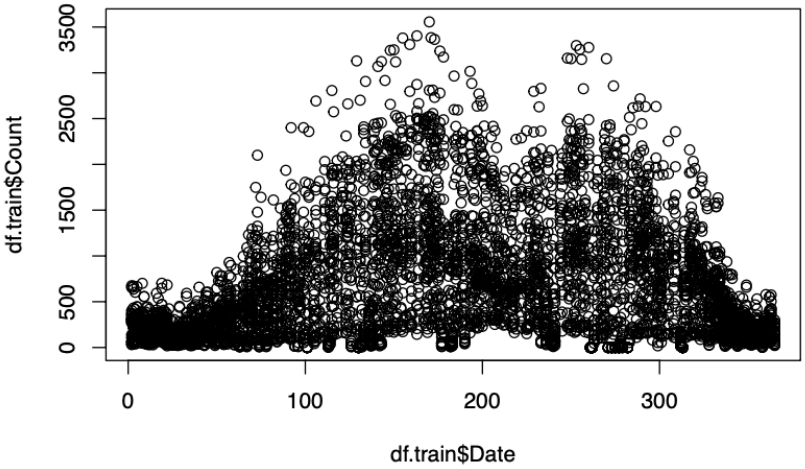
When initially tasked with handling this project, I first decided it was necessary to examine the dataset and think of the big picture before tackling any data analysis. With 14 predictors and 1 quantitative response, I knew there was going to be some variables that were more important than others. Upon first examination of the csv file, I was confused as to what the Functioning variable may represent. After scrolling through the dataset, I determined that it marked days for which the bike system was down, and upon further examination I saw that this only occurred about 10 days a year. To avoid skewing the data with days that had drastically less bikes rented than the mean (around 700), I decided to remove this predictor from the dataset altogether. I also removed the 'ID' predictor, as it was nothing more than a way to differentiate between hours of different days. Another glaring aspect I noticed was that the Date predictor was in character type, which I thought was going to cause problems down the road when attempting to calculate which days had an impact on bike count. I converted the predictor to days of the year instead (for example, January 1st is 1 and December 31st is 365). Once I had modified the training data set to my liking, I broke the data into a training and test set with an 80-20 split, and proceeded to take an overall view of the data by running a linear model on all the predictors.

Although the model was undoubtedly full of extra noise and unnecessary variables, it did give me a bit of a better idea of what predictors were going to be important throughout the course of the project. Hour, Solar, Rainfall, and Winter Season all had incredibly small p-values, with Humidity, Holiday and Temperature also proving extremely significant. I thought these results made sense, as my personal decision to bike would depend heavily on these factors. I would assume that most people like to ride bikes when it's relatively sunny and warm out in the afternoon, and not on rainy, miserable days. One aspect of the model that I found curious was the extremely large intercept value of 539. This value was also significant and would be a common theme throughout my discoveries that I never truly understood (more on that later). I took the test MSE of this model and got a large value around 580,000 to act as a baseline for future models. While I couldn't plot Count on all these variables, I was interested in its distribution over Date specifically.

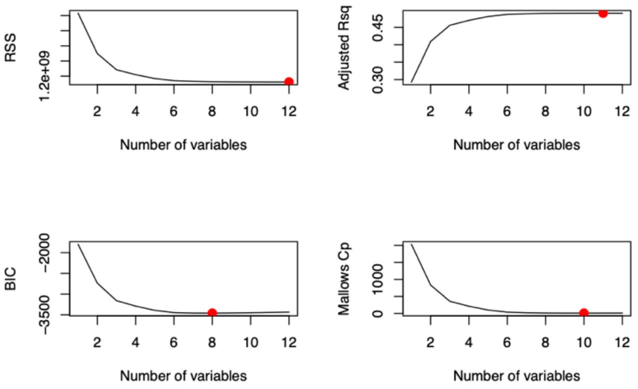


This plot only furthered my initial thoughts that people tend to ride bikes more in the warmer months towards the middle of the year.

My next move was to construct several histograms of the varying quantitative variables to examine if any frequencies lined up between variables (although date and hour are quantitative, their frequencies are constant throughout the dataset). Count, Solar, Rainfall and Snowfall all had most of their values skewed to the right, while Visibility was the only predictor skewed to the left. Temperature, Humidity, and Dew all followed a rather normal distribution and once again, these graphs made sense for the most part. Humidity and dew occur simultaneously in nature and tend to follow temperature in these trends. Sunny, rainy, and snowy days also occur rare in comparison to the rest of the year, so I could understand why they were skewed to low values. At this point, my logical hypothesis was that Solar was going to be the most significant impactor from this group, as its distribution of nearly no sunny days matched with Count's small number of high-rental days.

To see if my prediction was correct, I ran a for loop on each predictor individually to see which variable would help explain Count the best. To my surprise, the categorial Holiday variable edged out Rainfall and Date for having the smallest test MSE at around 370,000. This was an improvement from the complete model, but my intuition told me otherwise. While it makes sense that Holidays impact whether people bike, it didn't explain all the variance in the data. I proceeded to examine more complicated models via Subset, Forward, and Backward selection.

For all three of these model selection methods, I produced four graphs that showed the optimal variable number for RSS, Adjusted R-squared, BIC and Mallows Cp.



Strangely enough, all these methods ended up producing nearly identical ideal models. I knew that RSS was going to be best at max complexity, so I decided to check out the model with 7 predictors and 10 predictors that produced the minimal Cp, BIC, and AR2 respectively. Both these models included Hour, Temperature, Humidity, Solar, Rainfall, Holiday, and Winter Season, with the 10-predictor model including Date, Snowfall and Wind as well. I was shocked to see that SeasonsWinter had the largest Coefficient value in both scenarios, and that date played practically no role despite it appearing to have a strong correlation in the plot. Even more surprising was that these models had even higher MSEs than the individual predictors. I decided to see if this was a coincidence or not by examining each of the models built during forward and backward selection (they were all the same somehow).

The MSEs of all the different models didn't tell me much (it said the model with one predictor – Temperature – had the lowest MSE), but it did show that there wasn't a whole great deal of accuracy being gained from moving past 7 predictors. Keeping this in mind, I shifted my focus to developing lasso and ridge regression models on all the predictors in hope they would surpass the Holiday model's MSE. To my delight, they both did by a large amount (MSE was now at 200,000), and lasso managed to shrink many variables to near nothing. Ridge kept more of the coefficients in tack, and thus I thought it was going to be easier to interpret and understand than keeping extremely small values. I then decided to try ridge regression on the 7 variable model produced by Forward selection with Hour, Temperature, Humidity, Solar, Rainfall, Seasons, and Holiday. This produced a barely smaller MSE than that of the full model, but it was much simpler and more interpretable. In this model, the largest coefficients were SeasonsWinter (-274), Holiday (139), and Solar (-74). I fit it to the test csv file, and after rounding the numbers to whole and increasing all negative values into 0, I produced my final results.

When investigating this dataset, I ran into a lot of dead ends and misleading information. I could not understand how the same models were being built with different selection methods, and I still do not understand the idea behind the intercept being such a large number. It doesn’t make sense in context of the problem (if it’s midnight on the first of the year and there’s no weather?), yet it is nearly at the mean of Count by itself. Having a mix of categorial and quantitative predictors also made it difficult, as coefficients were being made for different seasons but didn’t appear in the actual dataset – as a result I got a lot of length errors and models that didn’t’ line up. While I do feel that this model is the optimal one that I found, I am not entirely confident about its performance. The MSE is still high, even for a large data set, but with so many predictors collinearity is bound to be in play. I will conclude however that my initial predictions and thoughts on the problem were fairly accurate. Most of my findings were logical, and I could understand how each of the predictors in the model could contribute to predicting bike rental demand.