To: Capital BikeShare

From: Jacob Igel

Date: May 13th, 2021

Subject: Analysis Memo of Bike Sharing Rentals

I am writing this memo to analyze Bike Sharing Rentals for Capital BikeShare. After analyzing the data, I believe I can point out the strengths and weaknesses of these particular datasets. In this memo, I will touch on several regressions and models that will all be shown in the appendix pages at the conclusion of this memo as well as in the R file attached.

**All Users**

While looking through the data through preliminary analysis, I was able to find some simple observations. For all users, as the weather gets worse (from 1 – 4), there are less bikes counted which is something that is expected since less users want to ride in bad weather. For casual users, we lose 11 at level two on average, 25 on average at level three, and 37 at level four. For registered users it is a loss of 20 at level two, 68 at level three, and 88 at level four. These are about the same due to the large difference in casual user’s vs registered in the dataset. I also came to the conclusion that there is a large correlation between atemp and temp, and cnt and registered (multicollinearity) Going off of that, I wanted to add some more variables in the mix so we can move away from the obvious.

After narrowing down 4 separate candidates, I found a model that explains about 27.9% of variation given the variables in the regression model (*can4*). It was tough to narrow down a fair model that would explain the general user since the model is split between the casual and registered. Another issue is the large number of degrees of freedom in the model that is also bringing the explained variability down. However, this model can be replicated with new data since our p-value is equal to 0%. In this model, we can see that there is a lot of interaction between the temperature and atemp, atemp and windspeed, and hum and atemp. These factors put together can give us a good idea on why riders might not like riding during these days. For example, if the atemp was 20, we can see that the number of riders will change by 800.82 - 666.14(actual humidity). This is an example of a significant predictor within this model. We can also see from plot1 that there is a small right skew for residuals.

**Casual**

When putting casual into the mix, I was able to explain about 51.03% of variation given this specific model (*regcas*). Given all of the variables within this regression, we can see that for everyone bike taken, the casual user count increases by 2 users. This is given we account for the temperature, humidity, actual temperature, and windspeed. All of which are significant predictors in this regression. We can also see with plot1 that the plot, for the most part, is normal until we get to the right side of the graph. This looks like that most of the model accounts for the lower half of the count/skewed right. All and all, however, our p-value with this model is low that it is usable for future data.

Before adding casual into the regression, it had an explanation of 27% much like the general model. With Casual added in, it barely brings down the of number of degrees of freedom so that is another concern with this model to keep in mind. This model accounts for residuals having a variance of 123.2 (RMSE) which is not awful but could be better.

**Registered**

Much like for the casual user regression, there is more explained variance as we add registered users. For our specific model (*regregist*), we can explain 95.5% of registered users given the other variables (temperature, humidity, actual temperature, and windspeed). The residuals also have a variance of 37.55 (RMSE) making this objectively better than the casual user model. Much like the two other regression models, it is also statistically significant (low p-value) meaning that this model can also be used for future data.

The downsides with this graph, much like the other two have the issue of a high degree of freedom count. Another issue that arises with this particular model is the low F-statistic. While looking at plot3, we can also see that much like plot1 for all users and the plot2 for casual users, the graph is skewed right.

**Appendix**

can4 = lm(formula = cnt ~ temp + hum + atemp + windspeed +

+temp:atemp + atemp:windspeed + hum:atemp, data = bike\_sample)

**Can4 Summary**

Text

Description automatically generated

R2 – .2773

ADJ-R2 – .2768

RMSE – 149.8

**Plot1**

![Chart, line chart

Description automatically generated]()

regcas = lm(cnt ~ temp + hum + atemp + windspeed + casual+ temp:atemp +

atemp:windspeed + hum:windspeed + hum:atemp, data = bike\_sample)

**RegCas Summary**

Graphical user interface, text

Description automatically generated

R2 – .5108

ADJ-R2 – .5103

RMSE – 123.2

**Plot2**

![Chart, line chart

Description automatically generated]()

regregist = lm(cnt ~ temp + hum + atemp + windspeed + registered + temp:atemp +

atemp:windspeed + hum:windspeed + hum:atemp, data = bike\_sample)

**RegRegist Summary**

Graphical user interface, text

Description automatically generated

R2 –.9551

ADJ-R2 –.955

RMSE – 37.35

**Plot3**

![Chart, line chart

Description automatically generated]()