Project 2

Liam Graham

Section 1 - Introduction

Bikes are not only efficient ways of getting exercise, but are efficient alternatives to fuel-consuming vehicles. The effectivity of these bikes, especially in light of the increased awareness of cars' impact on global warming and climate change, makes them a hot commodity, and cities are happy to offer them at a rental cost for intercity travel. However, though bikes have their benefits, they have their flaws too. It is often dangerous, impossible, or just not as enjoyable to use bikes if the weather conditions are too extreme, such as in cases of snow, rain, and winds. Temperature also has an impact on bike usage; extreme heat and cold are not good environments to use bikes.

With this information in mind, rental companies must think about the demand for bikes with changes in weather conditions. We are assigned the task of predicting the variable count, which in this case is the total bikes rented during a given hour, for a company named Capital Bikeshare Company. We are evaluating the change in rentals per hour based on the changes in weather conditions, specifically changes in precipitation, changes in temperature, humidity, and the month in the year.

Section 2 - Data

Table 1: Table 1: Table of Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	149.566907	58.9355881	2.5378029	0.0114613
rain_1h	-84.084983	29.9343111	-2.8089834	0.0051668
snow_1h	-56.903942	169.0234978	-0.3366629	0.7365140
temp	13.380258	1.4979255	8.9325252	0.0000000
humidity	-0.237067	0.8349276	-0.2839371	0.7765775
month	2.835888	4.3651579	0.6496644	0.5162110

Table 1 contains the model coefficients pertaining to the count of bike rentals. This table was made without interaction. We observe that rain has the biggest impact on the number of bikes rented in one hour, while humidity has the smallest. Regardless of the size of the impact, I decided to include all of these variables in future tests.

Table 2: Table 2: Table of Model Coefficients with Transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.1770326	0.2758747	15.1410508	0.0000000
rain_1h	-0.3787650	0.1401211	-2.7031264	0.0071057
snow_1h	-0.3009210	0.7911910	-0.3803393	0.7038570
temp	0.0469158	0.0070117	6.6910624	0.0000000
humidity	-0.0009507	0.0039083	-0.2432616	0.8079037
month	0.0426453	0.0204331	2.0870686	0.0373939

This table shows the transformed data from this dataset. The rain_1h coefficient appears to have the largest impact on bike rentals, while the humidity has the lowest impact, similarly to Table 1.

Table 3: Table 3: Table of Model Coefficients with Interaction

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.3320493	155.2789253	0.3434597	0.7314037
rain_1h	-231.6036335	2291.8386208	-0.1010558	0.9195485
snow_1h	4978.4591560	9118.6343449	0.5459654	0.5853444
temp	8.8759689	14.3544942	0.6183408	0.5366453
humidity	1.6759033	2.4789526	0.6760530	0.4993340
month	1.1418744	27.4534934	0.0415930	0.9668405
rain_1h:temp	-23.2459850	139.2902837	-0.1668888	0.8675281
snow_1h:temp	1717.5632398	3351.4516636	0.5124834	0.6085495
rain_1h:humidity	1.9888052	26.4496435	0.0751921	0.9400933
snow_1h:humidity	-65.2841790	110.0922230	-0.5929954	0.5534647
temp:humidity	0.0346062	0.2299746	0.1504786	0.8804505
rain_1h:month	136.6495467	350.6198920	0.3897370	0.6969046
snow_1h:month	278.1216740	428.7021349	0.6487527	0.5168098
temp:month	1.8507027	2.2332493	0.8287040	0.4076855
humidity:month	-0.0285883	0.4148869	-0.0689062	0.9450931
rain_1h:temp:humidity	0.1227400	1.6367600	0.0749896	0.9402543
snow_1h:temp:humidity	-35.0659348	58.6745494	-0.5976345	0.5503669
rain_1h:temp:month	-0.9561958	20.0538738	-0.0476814	0.9619901
snow_1h:temp:month	652.7693145	1034.0920766	0.6312487	0.5281796
rain_1h:humidity:month	-1.6091355	3.9866385	-0.4036322	0.6866637
temp:humidity:month	-0.0232118	0.0344852	-0.6730941	0.5012127
rain_1h:temp:humidity:month	0.0299083	0.2320490	0.1288880	0.8975005

This summary table is like Table 1, but includes an interaction term for the variables. Since there are dozens of coefficients, it does not make sense to use this table over the previous two tables. Many of the coefficients values display "NA," making this data unreliable and illogical to use as well.

Figure 1: Total Frequency of Bike Rentals

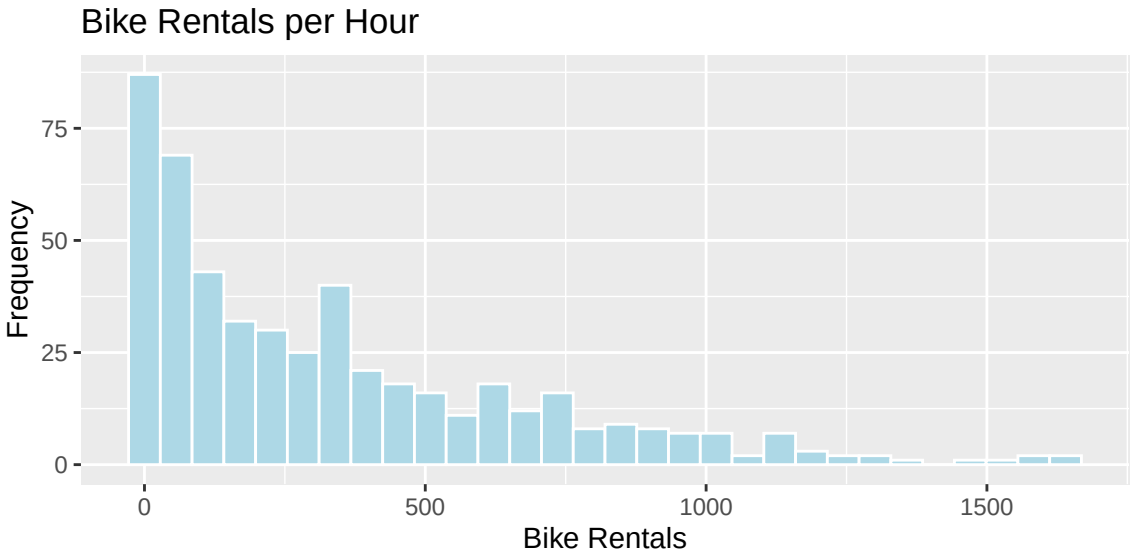


Figure 1 displays the total number of times that a bike was rented, divided up by the specific number of

rentals. The higher the bike rentals, the lower the occurrences of that specific number of rentals. Hours with zero rentals had the highest frequency. There were only three instances where the nunber of bike rentals exceeded 1500 in one hour.

Figure 2: Correlation Plot of Bike Rentals

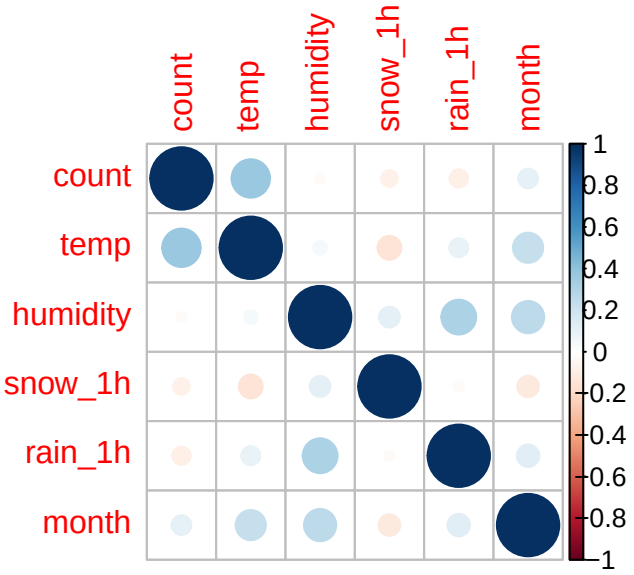
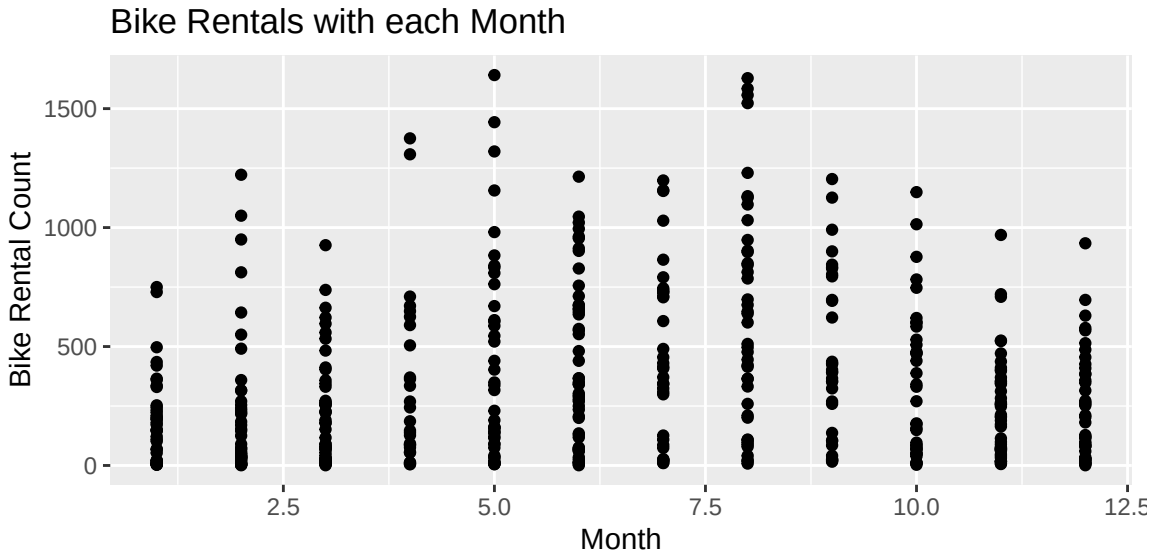


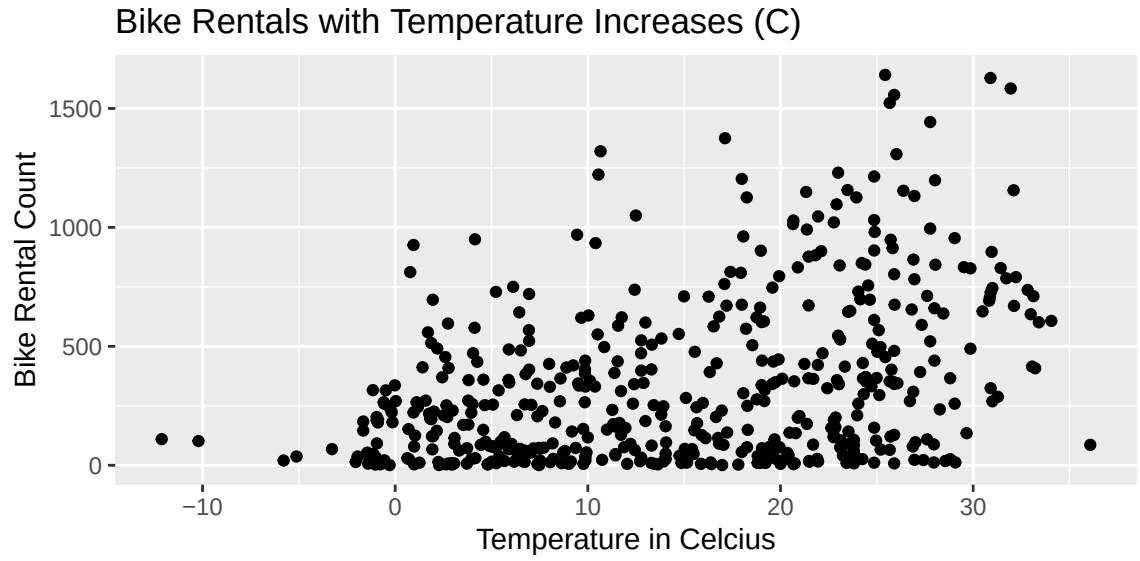
Figure 2 displays the correlation between the chosen variables. The correlations are fairly weak for all of the variables. Count is most closely associated with temperature, and is least associated with rain_1h. The other explanatory variables do not appear to be very correlated with each other whatsoever, which is good for our data, because explanatory variables that are highly correlated with each other can create interpretation problems.

Figure 3: Bike Rentals per Month



According to the scatterplot, there appears to be more rentals in the months that are in the middle of the plot (the data has the months as numbers; in this case, the months with the higher rental trend would be May, June, July, and August). The months on the edge of the data do not have as many rentals. The month with the least amount of rentals is January, followed by December and November.

Figure 4: Bike Rentals with Temperature Increase



While the increase isn't totally linear, Figure 4 appears to demonstrate a trend in which the number of bike rentals increases when the temperature increases. There were only two times when the temperature was less than -10 degrees Celcius, and very few rentals on that day. There was also only one time where the temperature was above 35 degrees Celcius, and there were very few rentals this day as well. Generally, though, as temperature increased, so did bike rentals, although there were many recorded hours that had temperature increases but still had zero rentals.

Figure 5: Bike Rentals with the Presence of Snowfall in the Previous Hour

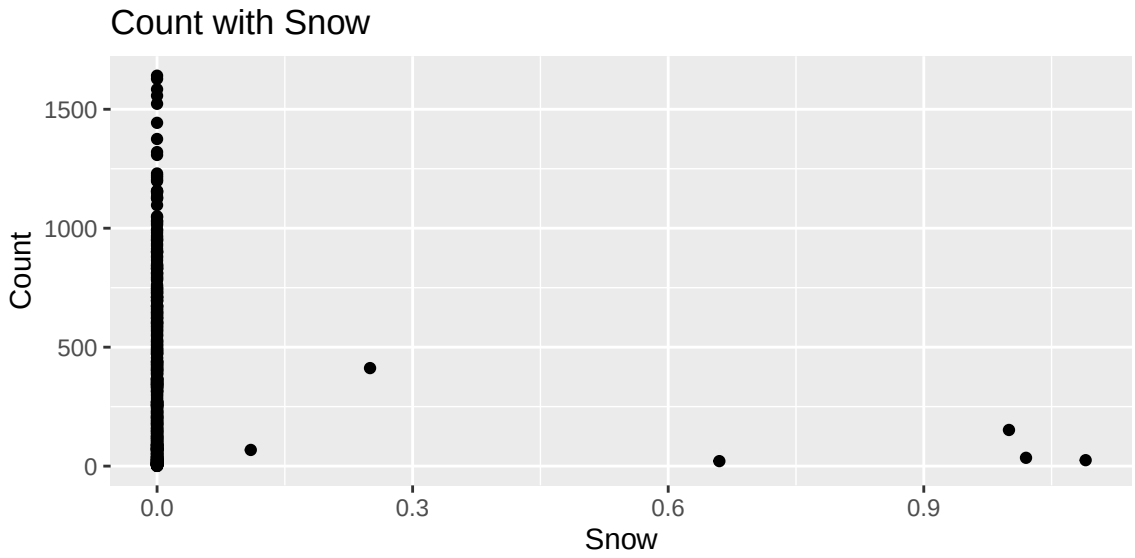
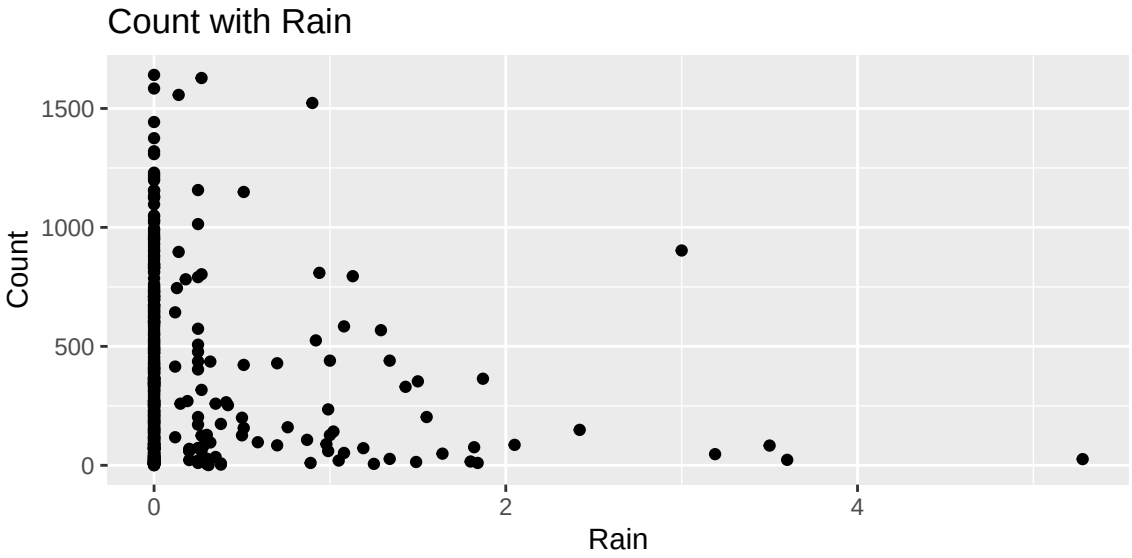


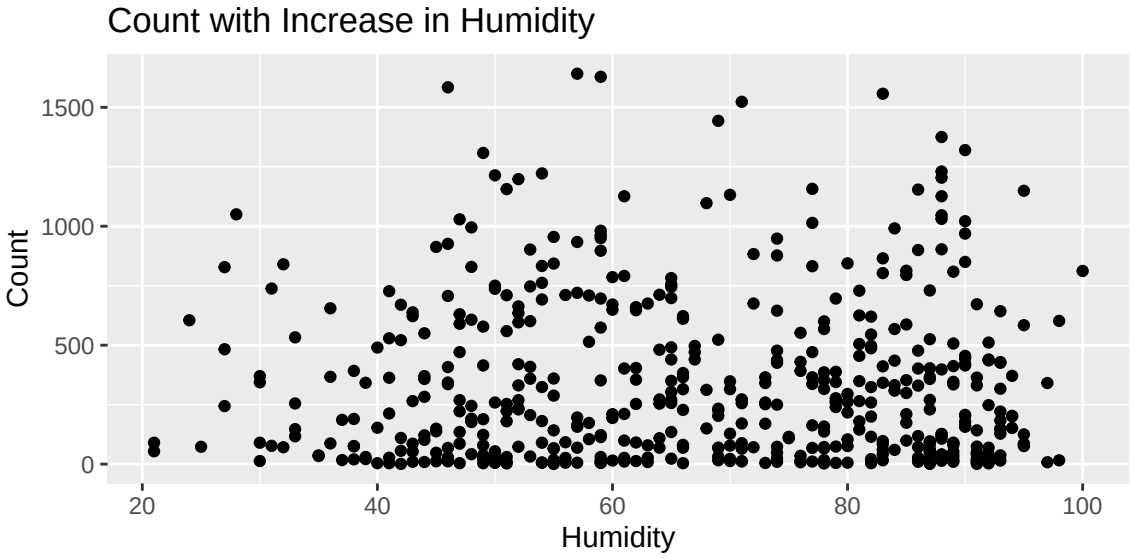
Figure 5 demonstrates the clear trend that the presence of snowfall has a negative effect on bike rentals. There were very few rentals when snow had fallen within the previous hour, and only six occurrences where snow had fallen and people had rented bikes. Only one of these occurrences yielded more than 250 rentals; the rest were well under. No bikes were rented after over 1.2 inches of snow had fallen in the previous hour. This would indicate that the presence of snow has a negative correlation with hourly bike rentals.

Figure 6: Bike Rentals with the Presence of Rain in the Previous Hour



Somewhat similarly to Figure 5, Figure 6 demonstrates a negative correlation associated with the presence of rain within the previous hour. There are far more rentals than there had been with snow, but still significantly less so than there were when there was no rain. There were many instances with very few rentals; there were also some occurrences where there were many rentals despite the rain. There is only one datapoint past the 5 inch mark, and it displays a very small number of rentals. This would indicate that there is a negative correlation between rain and bike rentals, like there was with snow.

Figure 7: Bike Rentals with Humidity Conditions



There does not appear to be a clear trend between bike rentals and humidity conditions. The data does not go in one direction, and has rentals at every humidity. It appears as if there is no relationship between the two. There are no observations below 20 units of humidity, and no observations above 100 units of humidity. The weak correlation indicates that humidity does not have a concrete effect on the rentals of bikes within one hour.

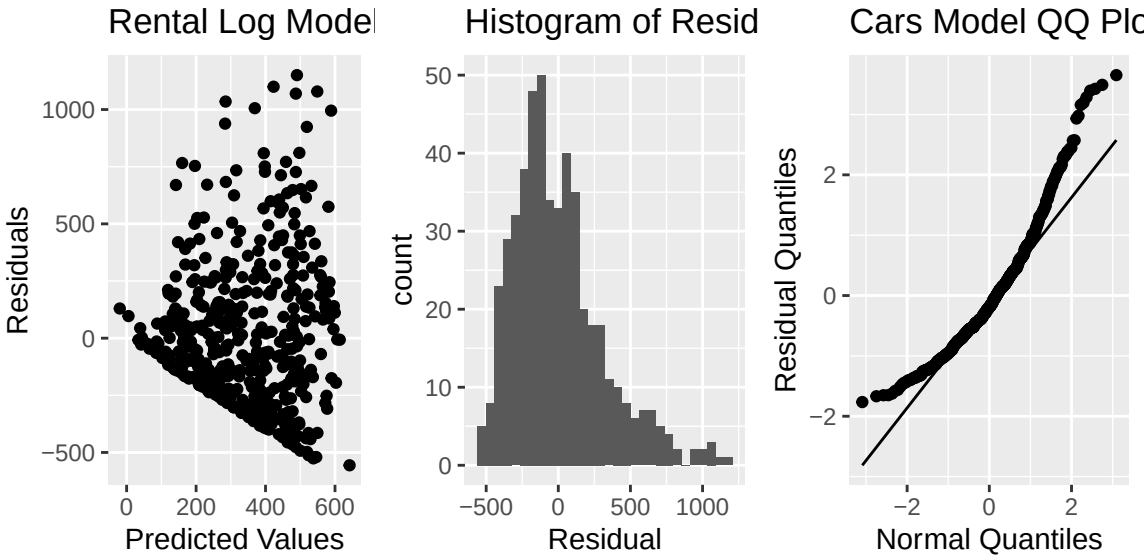
Section 3 - Model for Interpretation

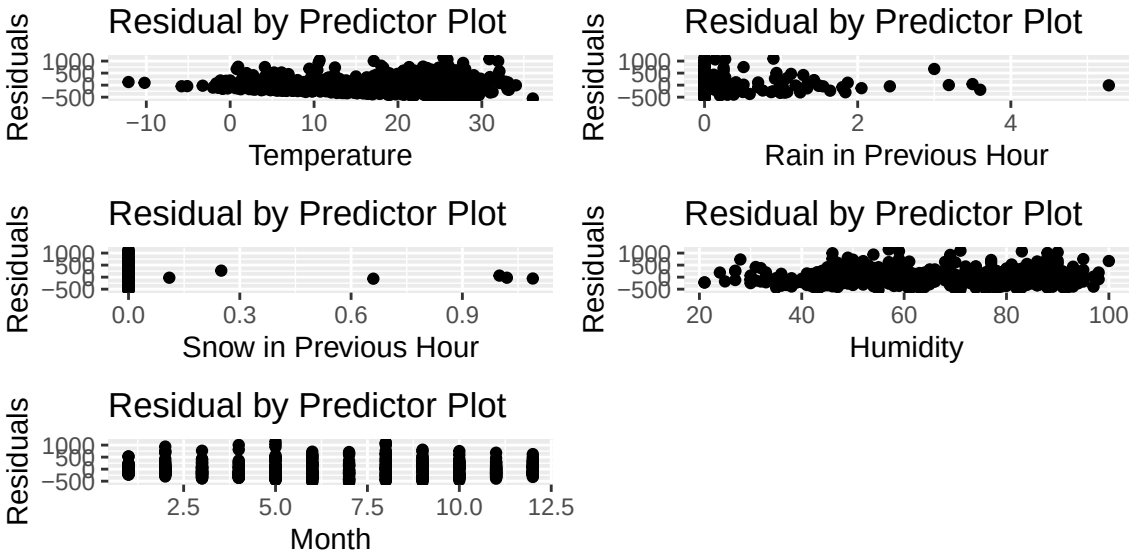
I decided to use Table 1, Table 2, Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6. In this case, I did not think that using interaction nor transformations were necessary, but decided to use transformed data

because it could possibly be relevant. The model with interaction was far too lengthy and overly complicated, so I deemed it unnecessary to use. I considered other variables, such as holidays, feels_like, and wind_speed, but deemed them inferior to my chosen variables for different reasons. Holidays would cause an unpredictable variation in bike usage, because of the different weather conditions of each holiday, and because of the ways people spend their holidays, so the data might not be relevant. What the temperature feels like MAY be relevant, but I thought that using the actual temperature would be more relevant because it does not vary and it is the ACTUAL temperature. Finally, with wind speed, I just did not see fit to use it over a similar variable such as humidity, which I believed would have more of a visible influence (my thought process was that the heavier the air, the less likely people are going to want to hop on a bike). I still do not want to use humidity, however, because of the lack of a visible correlation, but will include it regardless.

There does appear to be violations of the all of the assumptions. The residual plot makes a funnel shape, the histogram is heavily skewed to the left, and the points on the line in the QQ plot deviate from the diagonal significantly, so we should be concerned about the linearity, constance variance, and normal assumptions. We cannot graph the independence assumption, so we cannot say for sure whether or not this assumption has been violated. To evaluate the indepdence assumption, we can think about unseen variables like current events. The data cannot account for city-wide events, such as marathons, so bike rentals could vary based on the different things happening in the city. Additionally, it may be relevant to think about the pandemic in the context of this data; some of our data was taken from the year 2020, which is when COVID-19 was just beginning to spread. If the data includes months in which COVID was present, we need to think about the impact this had on the bike rentals.

For the plots of explanatory variables vs residuals, there do not appear to be any abnormalities. All of the data appears to be non-linear, aside from temperature, but in the context of the problem, this makes sense; the 'month' plot has 12 lines of data because there are 12 months, our rain and snow plots display information that we've confirmed from earlier, temperature appears to have a somewhat linear relationship with bike rentals, and humidity, while not displaying any linear relationship, does not appear to be different from what we expected.





Section 4: Interpretations

Table 4: Table 1: Table of Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	149.566907	58.9355881	2.5378029	0.0114613
rain_1h	-84.084983	29.9343111	-2.8089834	0.0051668
snow_1h	-56.903942	169.0234978	-0.3366629	0.7365140
temp	13.380258	1.4979255	8.9325252	0.0000000
humidity	-0.237067	0.8349276	-0.2839371	0.7765775
month	2.835888	4.3651579	0.6496644	0.5162110

Interpretation:

The intercept coefficient represents the number of bike rentals when rain_1h, snow_1h, temp, humidity, and month are all 0. In this case, it does not make sense to interpret this coefficient, because this scenario is implausible; there are no observations from the data where humidity is 0, and there is no month 0, so this cannot be the case. The "rain_1h" coefficient means that for every 1 unit increase in rain from the previous hour, bike rentals for the hour decrease by 84 units. The "snow_1h" coefficient means that for every 1 unit increase in snow from the previous hour, bike rentals for the hour decrease by 56 units. The "temp" coefficient means that for every 1 unit increase in temperature, bike rentals for the hour increase by 13 units. The "humidity" coefficient means that for every 1 unit increase in humidity, bike rentals for the hour decrease by 0.2 units. Finally, the "month" coefficient represents the increase in bike rentals for every additional month. For every one month increase, hourly bike rentals are expected to increase by 2 units. It makes sense to interpret all of these coefficients except month. As we saw previously, months have different rentals, which are primarily based on the weather conditions and temperature of the month itself. Therefore, it is not logical to assume that every change in months will mean that there are more bike rentals.

The p-value associated with this table is very, very small, meaning that any conclusions that we glean from hypothesis tests can be considered accurate.

Equation for this coefficients table:

$$\widehat{\text{Bike Rentals}} \text{ per Hour} = 145.57 - 84.09 * \text{Rain(1hr)} - 56.9 * \text{Snow(1hr)} - 13.38 * \text{Temperature} - 0.24 * \text{Humidity} + 2.84 * \text{Month}$$

Table 5: Table 2: Table of Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.1770326	0.2758747	15.1410508	0.0000000
rain_1h	-0.3787650	0.1401211	-2.7031264	0.0071057
snow_1h	-0.3009210	0.7911910	-0.3803393	0.7038570
temp	0.0469158	0.0070117	6.6910624	0.0000000
humidity	-0.0009507	0.0039083	-0.2432616	0.8079037
month	0.0426453	0.0204331	2.0870686	0.0373939

Interpretation:

The 'intercept' coefficient, e^4.17 represents the number of bike rentals if every other variable was e^0. In this case, it is not logical to interpret the coefficient, because we had zero observations in which humidity was 0, and there is no month 0. The 'rain_1h' coefficient indicates that for every 1 unit increase in rain in the previous hour, the number of bike rentals will multiply by a factor of e^-0.379. The 'snow_1h' coefficient indicates that for every additional 1 unit increase in snow in the previous hour, the number of bike rentals will multiply by a factor of e^-0.3. The 'temp' coefficient indicates that for every 1 unit increase in temperature, the number of bike rentals will multiply by a factor of e^0.047. The 'humidity' coefficient indicates that for ever 1 unit increase in humidity, bike rentals will multiply by a factor of e^-0.001. Finally, the 'month' coefficient indicates that for each additional month, the number of bike rentals will increase by a factor of e^0.043. Like the previous table, it makes sense to interpret all of these coefficients besides month, since changing from one month to the next will not have the same impact each time.

This table also produces a very low p-value, indicating that the predictions that we make from hypothesis tests are reliable.

Coefficients Table Equation:

$$\text{Bike } \widehat{\text{Rentals}} \text{ per Hour} = e^{4.18} e^{-0.38 * \text{Rain(1hr)}} e^{-0.30 * \text{Snow(1hr)}} e^{0.05 * \text{Temperature}} e^{-0.00096 * \text{Humidity}} e^{0.04 * \text{Month}}$$

Because we have concerns about our model assumptions, we could use bootstrapping as a substitute to check the security of our data, because it does not rely on the assumptions, unlike the normal error model.

```
##        fit     lwr      upr
## 1 551.3953 496.462 606.3286
```

We are 95% confident that the average number of bike rentals during a given hour in June, with no rain or snow in the previous hour, at a temperature of 30 degrees Celcius and with a humidity of 70 will be between 496 and 606 rentals.

```
##         fit       lwr      upr
## 1 -122.9426 -1118.263 872.3781
```

We are 95% confident that the average number of bike rentals during a given hour in January, with 0.06 inches of rain and 3 inches of snow within the previous hour, a temperature of -7 degrees Celcius and humidity of 25, will be between 0 and 872 rentals.

```
##         fit       lwr      upr
## 1 -329.3259 -1080.982 422.3299
```

We are 95% confident that the number of bike rentals during an individual hour in September, with no snow but 7 inches of rain in the previous hour, a temperature of 7 degrees Celcius, and 40 units of humidity, will be between 0 and 422 rentals.

```
##        fit       lwr      upr
## 1 385.1552 -238.5926 1008.903
```

We are 95% confident that the number of bike rentals during an individual hour in April, with half an inch of rain and no snow, temperature of 21 degrees Celcius, and 62 units of humidity will be between 0 and 1008.

Our confidence intervals may be considered invalid because we have negative values within our intervals. In this case, we can say that there were 0 rentals instead of negative rentals, since one cannot negatively rent a bike, but it still not a good sign that the intervals contain negatives in this case. Another red flag when working with these confidence and prediction intervals is that the upper bound goes up when variables which we previously said would negatively affect rentals are added, such as rain_1h and snow_1h, which does not make sense. Logically, we know that the presence of snow means that rentals would be much less, since snow is associated with cold temperatures and slippery sidewalks. But when we add to the snow_1h variable, the upper bound continues to rise. This leads us to question the validity of our intervals.

Section 5: Model for Prediction

For this cross validation, we used 10 repeats of 10 folds to find the RMSPE of the data.

```
set.seed(10312022)
model1 <- train(data=Train_Data, count ~ `temp`,
                method="lm", trControl=control)
RMSPE1 <- sqrt(mean((model1$pred$obs-model1$pred$pred)^2))
```

```
## [1] 318.4878
```

```
set.seed(10312022)
model2 <- train(data=Train_Data, count ~ `temp` + `rain_1h` + `snow_1h`,
                method="lm", trControl=control)
RMSPE2 <- sqrt(mean((model2$pred$obs-model2$pred$pred)^2))
```

```
## [1] 316.1248
```

```
set.seed(10312022)
model3 <- train(data=Train_Data, count ~ `rain_1h` +
                  `snow_1h` + `temp` +
                  `humidity` + `month`,
                method="lm", trControl=control)
RMSPE3 <- sqrt(mean((model3$pred$obs-model3$pred$pred)^2))
```

```
## [1] 317.1055
```

```
set.seed(10312022)
model4 <- train(data=Train_Data, count ~ `rain_1h` +
                  `snow_1h` + `feels_like` +
                  `wind_speed` + `month` + `holiday` + `weather_main`,
                method="lm", trControl=control)
RMSPE4 <- sqrt(mean((model4$pred$obs-model4$pred$pred)^2))
```

```
## [1] 320.1844
```

```
set.seed(10312022)
model5 <- train(data=Train_Data, count ~ .,  method="lm", trControl=control)
RMSPE5 <- sqrt(mean((model5$pred$obs-model5$pred$pred)^2))
```

```
## [1] 293.6701
```

```
## [1] 318.4878
```

```
## [1] 316.1248
```

```
## [1] 317.1055
```

```
## [1] 320.1844
```

```
## [1] 293.6701
```

For the first model, I decided to include temperature as the only variable. In my opinion, this would have the greatest impact on bike rentals because generally, the colder the temperature, the less likely someone would want to spend time on a bike. This is also the variable that is the most prominent on someone's mind when they consider the idea of riding a bike.

For the second model, I chose to only include temperature, snow_1h, and rain_1h. I included temperature because of the impact I believe it has, and then also included the other two variables because previous conclusions allow us to believe that they are the most negatively correlated variables associated with hourly bike rentals. This makes them two very relevant variables.

The third model calculated the RMSPE using all of the variables that I had tested out so far: temperature, rain_1h, snow_1h, humidity, and month. I decided to choose these five variables because I had been using them for all models leading up to this point, and it made sense to do a test with them all included. All three of the RMSPEs from the models were similar, only varying by two or less from each other.

The fourth model used seven variables: feels_like, rain_1h, snow_1h, wind_speed, month, and holiday. I thought that it would be relevant to mix up the variables and include the three variables I previously deemed inadequate for the project, which were holiday, feels_like, and wind_speed. I also included weather_main, because this has a large impact on the decision to ride a bike, and includes rain and snow, which we previously deemed meaningful. The RMSPE was similar to the previous 3 models, but was the highest.

For the fifth and final model, I combined all of the variables into one model, to see how the RMSPE would compare to the others. This model produced the lowest RMSPE out of the five models.

I based my decision regarding whether or not to include these variables on the idea that these are the variables that take up space in people's thoughts when considering whether or not to rent a bike in the city. Oftentimes, they consider weather conditions or the day, which covers most of the variables chosen for the models. Many people do not consciously think about months when deciding whether or not to rent a bike, but the weather of each month is ultimately a large factor in their decision-making process, and the weather varies by the month. People are not usually considering pressure and wind degree when they are deciding about riding a bike, so I chose to exclude those variables.

The results of our cross-validation are clear. The fifth and final model, which included all of the variables, produced the lowest RMSPE. The other four models yielded RMSPEs that all fell between 316 and 320. It appears that using only some of the variables is not nearly as effective as using all of them to predict bike rentals. It is difficult to determine which specific variables were unnecessary to include, since each RMSPE was so similar.

```
## # A tibble: 6 x 19
##   count holiday workingday  temp feels~1 temp_~2 temp_~3 press~4 humid~5 wind_~6
##   <lgl>   <dbl>      <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 NA          0          1  1.31    1.31   -0.96    2.06    1017      69    0.45
## 2 NA          0          0  8.3     8.3     7.54    8.85    1020      94    0
## 3 NA          0          0 21.1    20.0    19.6    22.5     1015      28    1.34
## 4 NA          0          1 18.5    18.6    16.7    19.5     1017      84    1
## 5 NA          0          1  1.09    1.09   -2.03    3.51    1029      84    0.5
## 6 NA          0          1 32.8    33.6    30.6    32.8     1017      41    2.06
## # ... with 9 more variables: wind_deg <dbl>, rain_1h <dbl>, snow_1h <dbl>,
## #   clouds_all <dbl>, weather_main <chr>, year <dbl>, month <dbl>, day <dbl>,
## #   hour <dbl>, and abbreviated variable names 1: feels_like, 2: temp_min,
## #   3: temp_max, 4: pressure, 5: humidity, 6: wind_speed
```

Section 6: Discussion and Conclusions

The models used for interpretation made it clear that each variable had different effects on the number of rentals per hour. The model coefficients tables indicated that the variables I chose did have some sort of impact on bike rentals. The coefficients were backed by low p-values for all three tables, though we decided to not use the interaction table. We saw through the correlation plot that no variables stood out as overly correlated, which incited some doubt about the results. The scatterplots involving month, rain_1h, snow_1h, and temperature all indicated that all four of these variables had a noticeable impact on hourly bike rentals. With colder months, as well as the presence of rain and snow, we saw a negative correlation, but with rising temperature and hotter months, we saw a positive correlation. However, we had violations of all three visible assumptions in our plots, which caused us to question the results of the models.

The models used for prediction had their flaws. The confidence intervals included negative values in the range, something that should not happen in this context. Additionally, some variables seemed to impact the intervals in the opposite way that they were expected, such as the increase in snow_1h causing an increase in the upper bound of any confidence interval it was included in. The RMSPEs calculated in our models were all similar in size, aside from the final model, which included all of the variables and had a slightly lower RMSPE. With this information in mind, we can say that the prediction and interpretation models yielded different findings.

In the interpretation models, temperature, rain_1h, and snow_1h appeared to have the greatest impact on bike rentals, according to their respective scatterplots. There were much less bike rentals with the presence of rain, and nearly no rentals when snow had been present, and we saw that temperature increases generally caused bike rental increases. However, all three of our model assumptions were violated by this data, causing us to question the data's validity. In the prediction models, it is not clear that any variables had more of an impact than others. When interpreting our confidence intervals, we found that higher temperatures and warmer months yielded higher rentals, while cold temperatures, the presence of precipitation, and colder months yielded lower rentals. However, when doing our cross-validation, we found that no specific variables had more of an impact than others, as the RMSPEs were all very similar for tests that used specifically named variables. The RMSPE was lowest with the model that contained all of the variables, so it is difficult to say which specific variable had a large impact, since none of them are listed specifically. I learned through this dataset that interpretation and prediction models can be drastically different, and that one may be more reliable than the other. I also learned that even if models can appear to be relevant and correct, with p-values to back them up, the linearity, constance variance, and normal assumptions may still be violated. Finally, I learned about the true impact of variables that cannot always be measured by models. For this data specifically, it takes observations from a time where COVID-19 caused the world to be shut down. This more than likely impacted the number of bike rentals DRASTICALLY, because in fear of sanitation, bike companies probably either reduced or completely pulled the number of bikes available from the city. It's also difficult to include the year coefficient because not all of 2020 involved shut-downs, so any conclusions we glean from that data may be considered inadequate.