**Prediction Model for Bike Rentals in 2012**

****

By: Claire Kim, Jiabin Zheng, Ethan Biddle, Frederick Nusetor

MA 575 - C2

Group 3

Professor Judith Lok

Apr 12, 2023

**Contents**

I. Abstract ....……………………………………………………………………………... 2

II. Introduction .……………………………………………….……………..................... 2

III. Background ...……………………………………....……………………...…...…..... 2

IV. Modeling and Analysis .. …….……………………………..…............................. 2 - 6

V. Prediction ………………………………………….……………….………………6 - 7

VI. Discussion …………..…….………………………….…….…….…………….... 7 - 8

VII. Contributions …………………………………………………….…………………. 8

VIII. Appendix ……………………………………………………….……….…….. 9 - 12

IX. References …………………………………………….……………....……..…….. 13

1. **Abstract**

With the increasing concern for the world’s environment and health, the government is promoting the use of eco-friendly transportation. Bike sharing has been one of the methods that have been initiated to tackle this problem and have grown to such a degree that several bike rentals can be easily found today. To keep promoting the bike-sharing industry, we are interested in predicting when people are most likely to rent bikes and how many would be renting in the future. In this experiment, we tried to predict the number of people that would be using bikes in the future by determining the most influential factors that would promote or discourage peoples’ usage of bike rentals in 2011, using those results to predict how many people would rent bikes in 2012, and then finally compare those results to the actual number of users that rented bikes in 2012.

1. **Introduction**

Providing an accurate prediction of the number of people that would be using bikes in the next year, as well as finding what factors would be most significant in determining bike usage are crucial aspects for the bike industry, and hence were our main goals for this experiment. With the data collected from 2011, we want to build a model using the data to predict the number of bike users for the following year and compare those results with the true 2012 data that was also included in the initial dataset. We also want to figure out which factor was the most significant in determining the number of bike users. Our initial hypothesis is that average temperature will be the most decisive factor, as people are likely to refrain from riding bikes at times that are too cold, and in turn are more likely to rent and ride bikes when the weather is warm. Alongside average temperature, we also plan to consider other factors such as humidity, wind speed, the given season, different weather conditions, and whether it was a working day or a holiday.

1. **Background**

This dataset is data collected by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto INESC Porto, Campus da FE. The bike-sharing system is a new generation of bike rental where registered people can rent bikes at one location and return them at another. Apart from interesting real-world applications of bike-sharing systems, the characteristics of data generated by these systems make them attractive for research. This feature turns the bike-sharing system into a virtual sensor network that can be used for sensing mobility in the city. This leads us back to our initial goal in that we want to know the most important factors that could influence bike usage daily by using the data and developing a significant model for predicting daily bike usage for future years.

1. **Modeling and Analysis**

After splitting the dataset into two separate ones for 2011 and 2012, we produced a correlation matrix to see which of our continuous variables are most correlated with the number of bike rentals. As evidenced in Figure 1 below, we found that temperature and average temperature were highly correlated to the number of bike rentals. In turn, they are also highly correlated with each other, suggesting we should only use one of the two predictors to avoid multicollinearity in our model. We also can see some correlation between windspeed and bike rentals, suggesting that it could also be useful within our model. Our final variable, humidity, does not appear to have much correlation with our data and does not necessarily have to be in our model at this time.

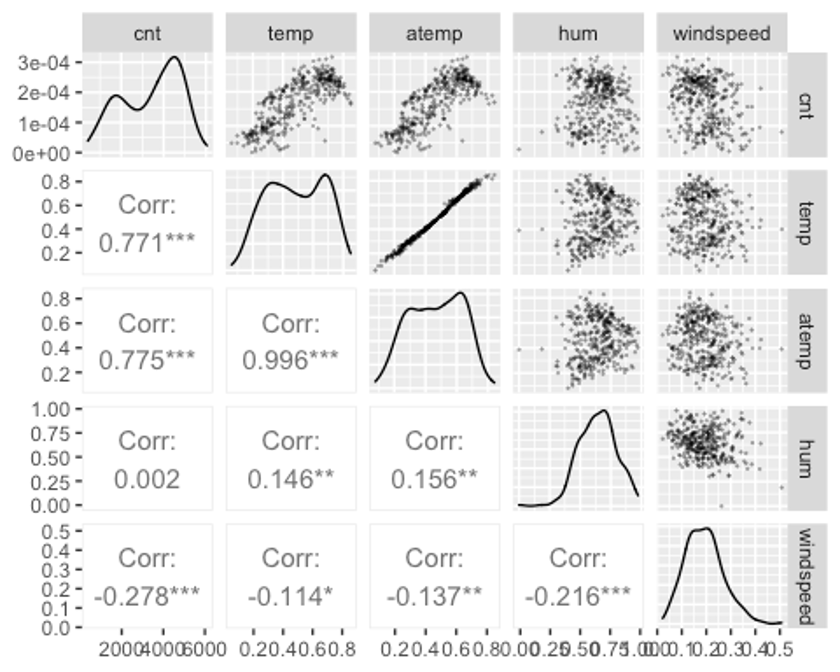


Figure 1. This is a correlation matrix directly from R, depicting the correlation between not only our response variable and predictors but also the correlation between the predictors themselves. Correlation coefficients are on a scale from -1 to 1, with the polar ends of the spectrum signifying high negative or positive correlation respectively.

We then produced boxplots on the rest of our predictors which were non-continuous variables. These boxplots, much like the correlation matrix above, allowed us to see which predictors were influential on the data. Our results showed that both variables about which type of day it was were not significant predictors, as there is very minimal change between the number of bike rentals whether it be a working day or non-working day, as well as if it was a holiday, or not one. These results can be found in the Appendix as Figures A and B.

While the previous two plots showed two variables that were not influential on bike usage, the following two did. First, we saw that change in climate was impacting bike rental numbers, as understandably so, people rented bikes less frequently when it was cloudy, and significantly less when it was raining in comparison to when the sky was clear. A similar trend can be found in Figure 3, where we see a notable difference in bike rentals during winter months as opposed to any other season of the year. This boxplot of summer also showed potential outliers in our data, which will be revisited later in our experiment once our model is finalized.

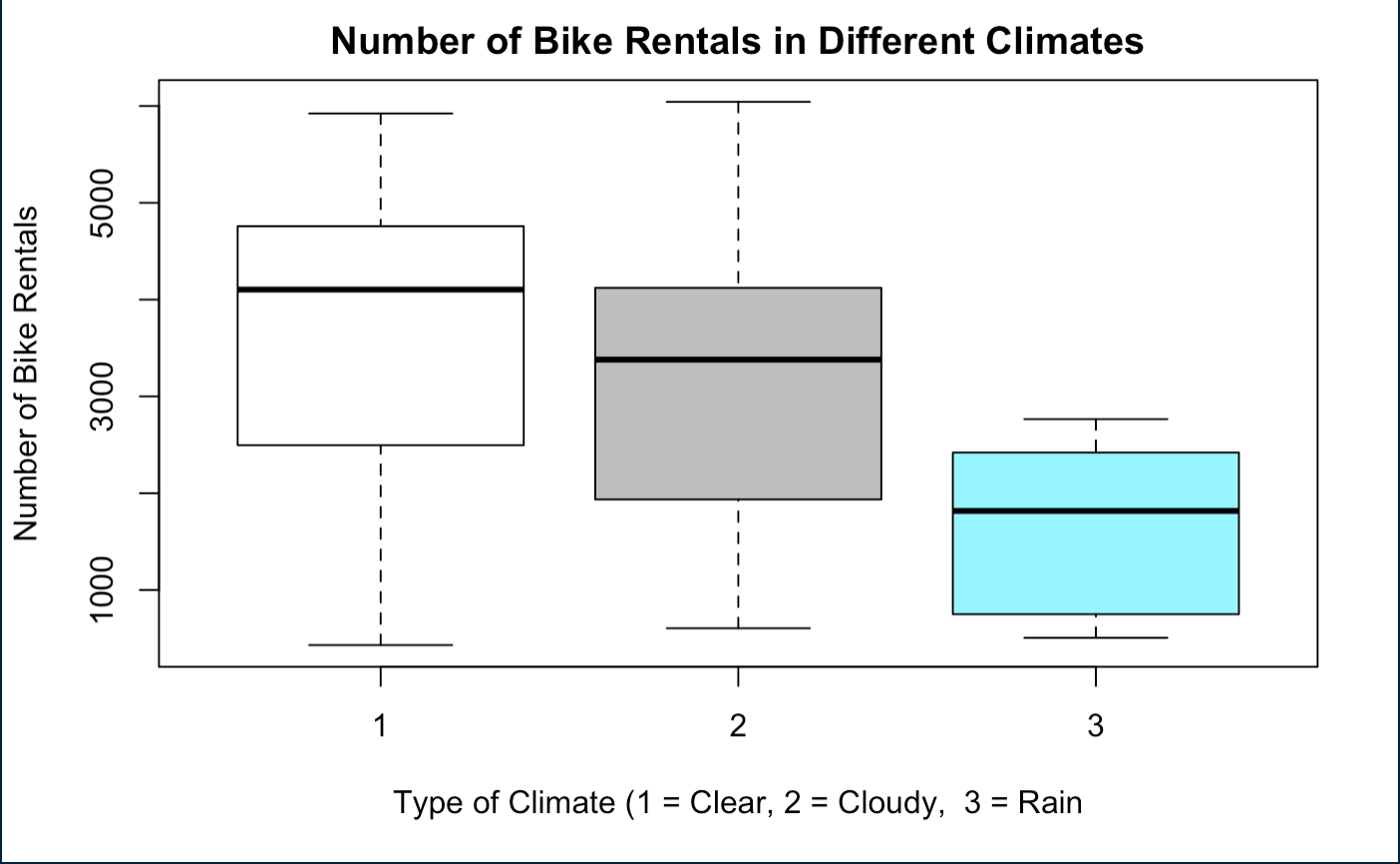


Figure 2. This is a boxplot directly from R, showing how many bikes were rented during different climates. Based on the plot, it can be seen that the difference in type of climate has a substantial influence on bike rentals, with most bikes being rented on clear days, and the least being rented on rainy days. This suggests that the variable “weathersit” could be included in our model.

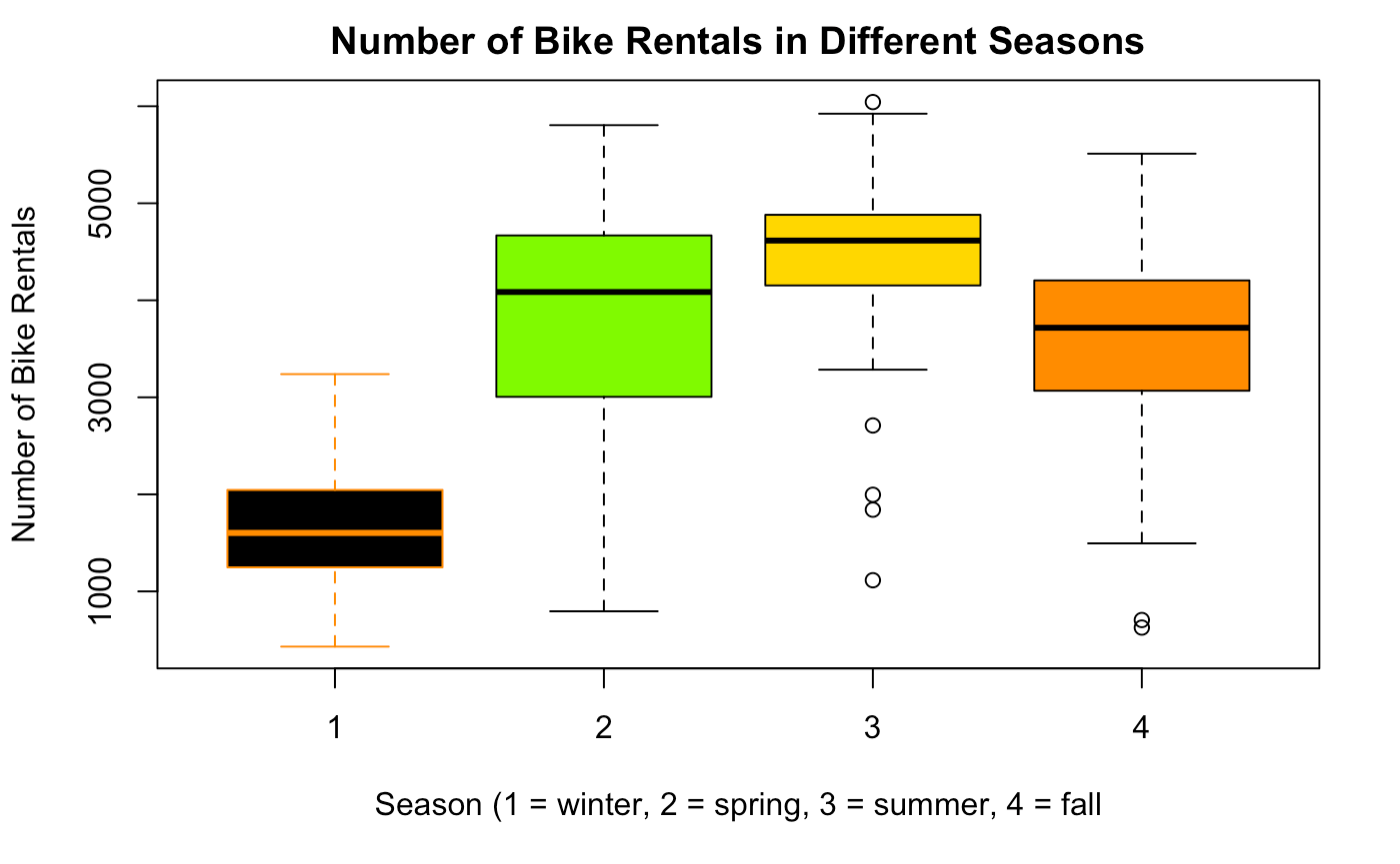


Figure 3. This is a boxplot directly from R, showing how many bikes were rented during different seasons of the year. Based on the plot, it can be seen that the difference in season has a substantial influence on bike rentals, with most bikes being rented during the summer and spring, and the least being rented during the winter. This suggests that the variable “season” could be included in our model.

Based on these results, we knew our model had to include average temperature or temperature, season, and climate. However, after conducting a Chi-squared test – Figure 4 – on season and climate, we found that the two variables were not significantly different and in turn highly correlated with each other. So much like average temperature and temperature, we had to include just one of the two choices.

**Pearson's Chi-squared test  
  
data: work\_hol  
X-squared = 11.699, df = 6, p-value = 0.06902**

Figure 4. This is a Chi-squared test directly produced from R. This test produced a p-value of 0.06902, which is greater than our critical value of 0.05 when working with 95% confidence. This suggests that the variables “weathersit” and “season” are strongly correlated, and in turn should not be in the model together to prevent overfitting.

We next decided to revisit our predictor of wind speed. As noted earlier, windspeed had some correlation with bike rentals, and in turn could be used in our model. To determine its validity within our model though, we conducted a hypothesis test, resulting in a p-value of 6.673e-8, which is much smaller than our .05 alpha value used for 95% confidence on the data, and in turn say that windspeed is indeed a significant predictor of bike rentals. The full R output for this experiment can be found in the Appendix as Figure C.

At this point, we had settled upon our model consisting of average temperature, season, and windspeed. This model, which can be found in the Appendix as Figure D, came with some flaws though, as our data was not completely random and normal. A slight curve in our residuals vs. fitted plot suggested the inclusion of a quadratic term in our model, which ultimately led to us including a squared term of average temperature.

With the inclusion of this quadratic term, our model, figure E in the Appendix, was much improved and did a better job of meeting the necessary normality and random assumptions for our data. While this model did a better job in that regard, it still didn’t have a substantial R-squared and low standard error, with values of 0.7017 and 757.2 respectively.

To improve those values, we opted to include humidity in our model. And while we found humidity to not be a significant predictor of bike rentals on its own, it actually was a significant predictor of the data as part of the model, as evidenced by its p-value of 2e-16, which is substantially lower than our desired alpha value of 0.05. Furthermore, our goal of increasing our R-squared and lowering our standard error was complete, as we now saw values of 0.7632 and 675.7 respectively. Lastly, our residuals vs. fitted plot was now completely random, and our Normal Q-Q plot had very little tailing, suggesting our assumptions of normality and independent random errors were now complete. The results and plots of this data can be found below in Figures 5 and 6.

**Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) 398.16 285.95 1.392 0.165   
 atemp 15048.97 1307.83 11.507 < 2e-16 \*\*\*  
 I(atemp^2) -10201.54 1374.31 -7.423 8.38e-13 \*\*\*  
 season 299.47 37.08 8.077 1.02e-14 \*\*\*  
 hum -2514.63 260.58 -9.650 < 2e-16 \*\*\*  
 windspeed -3313.34 485.57 -6.824 3.79e-11 \*\*\*  
 ---  
 Residual standard error: 675.7 on 359 degrees of freedom  
 Multiple R-squared: 0.7632, Adjusted R-squared: 0.7599  
 F-statistic: 231.4 on 5 and 359 DF, p-value: < 2.2e-16**

Figure 5. This is a summary of our model directly from R, showcasing our estimates for each parameter of our model, as well as their p-values, all of which are significant. We can also see our residual standard error and R-squared for the model, which is now much improved. Lastly, we see our F-statistic and p-value for the entire model, both of which suggest our model is significant under 95% confidence.

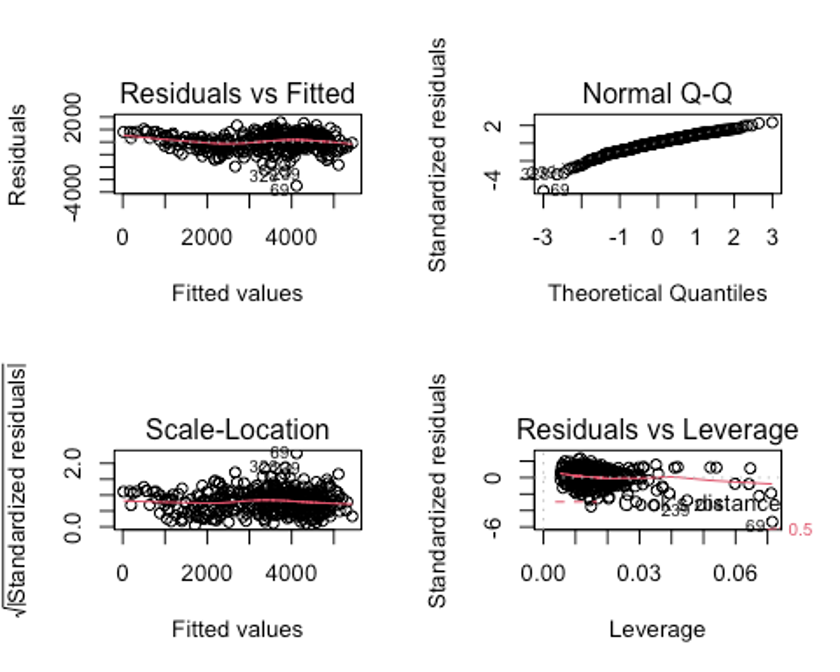


Figure 6. These plots are directly from R and show the validity of our model under the assumptions of multiple linear regression. Our residuals vs. fitted plot is random, implying that our errors are independent of one another. We also see little tailing in our Normal Q-Q plot, suggesting that our errors are distributed normally. We then can see our scale-location plot is random, satisfying homoscedasticity. Lastly, our leverage plot shows points that have high leverage and could be deemed outliers.

Our last step before proceeding with our prediction, was to check for outliers in our data. As seen in our leverage plot, we do have some points with high leverage. Using the boxplot function in R, we attempted to remove any outliers outside of the plots Inter Quartile Range, however no points were found, suggesting our data did not have any outliers. This R output can be found in the Appendix under Figure G.

1. **Prediction**

Based on the ideal model created in section (IV), we conducted a prediction experiment in R using a function titled “predict”. This function takes in our model’s intercept and estimated values of our parameters and then uses the bike rental data from 2012 in that equation to find not only a predicted number of users for a given day, but also an interval in which that value is likely to fall under 95% of the time. The prediction for the first six days of 2012 can be found in Figure 7.

**head(predict(model, newdata = BikeData12, interval = 'confidence'))**

**day fit lwr upr  
 1/1/2012 2532.8863 2385.4059 2680.3668  
 1/2/2012 1794.0027 1608.3835 1979.6220  
 1/3/2012 114.0889 -181.8514 410.0291  
 1/4/2012 693.7396 425.4473 962.0320  
 1/5/2012 2347.9028 2187.7488 2508.0567  
 1/6/2012 2717.6629 2565.1952 2870.1306**

Figure 7. This is direct R output and showcases the fitted values of our prediction as well as a 95% confidence interval for those values. This section of the data showcases a prediction of the number of uses for the first six days of 2012 based on 2011 data.

After conducting this prediction, we wanted to compare it to the actual number of users that rented bikes in 2012. We did this by calculating the average of both our predicted number of users on a random day, as well as the actual number of daily users in 2012. This process can be shown in Figure 8 below.

**mean(predict(model, newdata = BikeData12))**

**Predicted Number of Daily Users for 2012 = 3610.471**

**mean(BikeData12$cnt)**

**Actual Number of Daily Users for 2012 = 5599.934**

Figure 8. This is direct R output showing the average number of daily users our model predicts for 2012, as well as the actual average number of daily users in 2012.

Based on the results of our prediction, we would predict that on average, 3,610 users would rent bikes on a random day in 2012. However, based on the actual 2012 data, it can be found that the average day saw 5,599 users rent bikes. This suggests that although our model is significant for 2011 data, it is not an accurate predictor for 2012.

One reason for this is the lack of inclusion of the variable “registered” in our model, which calculated the number of registered users that rented bikes on a given day. After further examination of the data, it can be found that while the variables of average temperature, humidity, season, and windspeed are all consistent from 2011 to 2012, a huge spike in registered users caused a stark increase in overall bike rentals between the two years.

This led us to experiment with the inclusion of registered users in our model, resulting in a much more accurate prediction of 2012 users. Although the prediction for 2012 users is now much more effective, it is a concern that including registered users in our model could result in overfitting, given that our response variable “cnt” is simply “registered” plus “casual”. In turn, we opted to stick with our previous model while continuing to acknowledge that an increase in registered users will have a great impact on bike rental numbers. The second model and prediction of 2012 users can be found in the appendix as Figure H.

1. **Discussion**

The goal of this project was to create a model that accurately predicts the number of bike rentals based on various factors such as temperature, humidity, season, and wind speed. The ideal model was created in section IV, and a prediction experiment was conducted using R. The results of the prediction experiment were compared to the actual number of bike rentals in 2012, which led to the discovery that the model, while significant for 2011, was not an accurate predictor for 2012.

Upon further examination of the data, it was discovered that the lack of inclusion of the variable "registered" in the model was a significant factor in its inability to accurately predict the number of bike rentals in 2012. However, including registered users in the model could lead to overfitting, which is why the original model was ultimately chosen.

In conclusion, while the model created in this project was successful being a significant predictor for bike rentals in 2011, it was not as effective for predicting rentals in 2012. This highlights the importance of continuously evaluating and updating models as new data becomes available, as well as acknowledging the impact of additional variables on the accuracy of the model. Future studies could focus on refining the model by including additional variables or exploring different machine-learning techniques to better predict bike rentals.

1. **Contributions**

Formatting - CK

Abstract - CK

Introduction - CK, JZ

Modeling & Analysis - EB

Outlier Testing - EB

Prediction Analysis - EB

Model Construction - EB

Discussion - JZ

Editing - CK, JZ, EB

1. **Appendix**

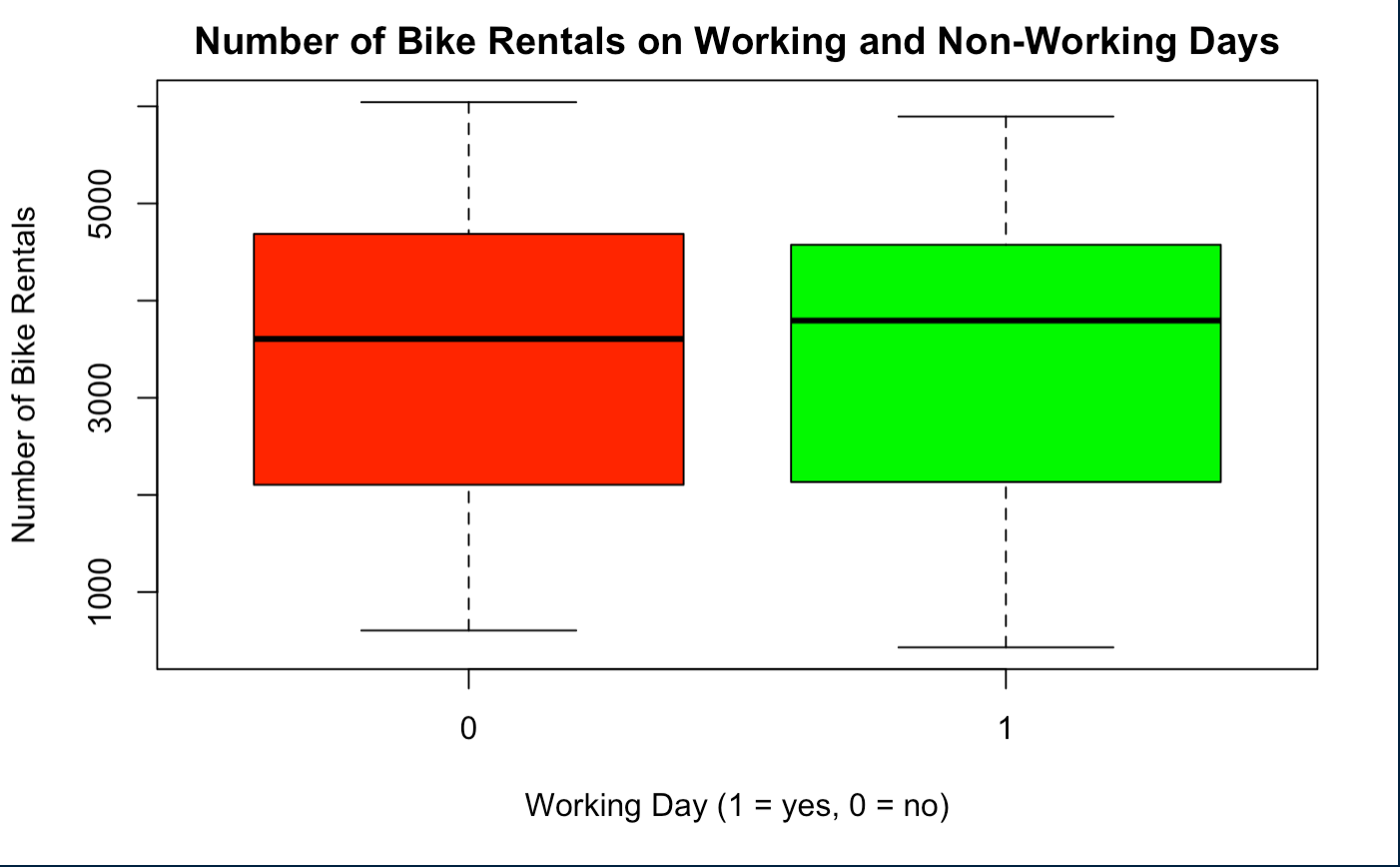


Figure A. This is a boxplot directly from R, showing how many bikes were rented on working and non-working days. Based on the plot, it can be seen that the difference in the type of day does not have a substantial influence on bike rentals.

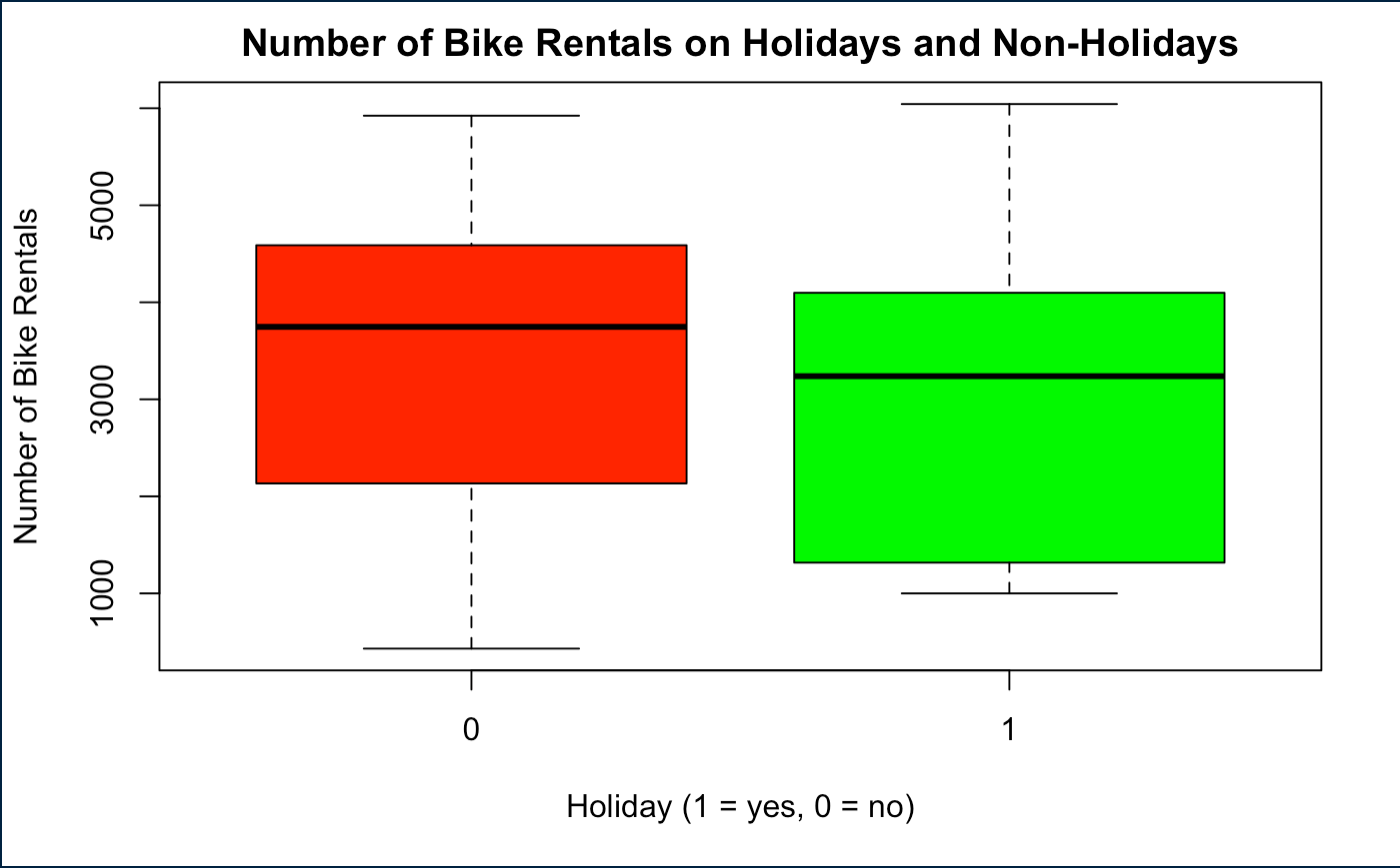


Figure B. This is a boxplot directly from R, showing how many bikes were rented on holidays and non-holidays. Based on the plot, the difference in the type of day does not have a substantial influence on bike rentals.

**Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) 4359.9 186.4 23.384 < 2e-16 \*\*\*  
 Windspeed -4985.0 904.1 -5.514 6.67e-08 \*\*\*  
 ---  
  
 Residual standard error: 1326 on 363 degrees of freedom  
 Multiple R-squared: 0.07728, Adjusted R-squared: 0.07474  
 F-statistic: 30.4 on 1 and 363 DF, p-value: 6.673e-08**

Figure C. This is a direct output from R, showcasing the significance of windspeed as a predictor for bike rentals. We used this information to eventually include windspeed as a covariate in our final model.

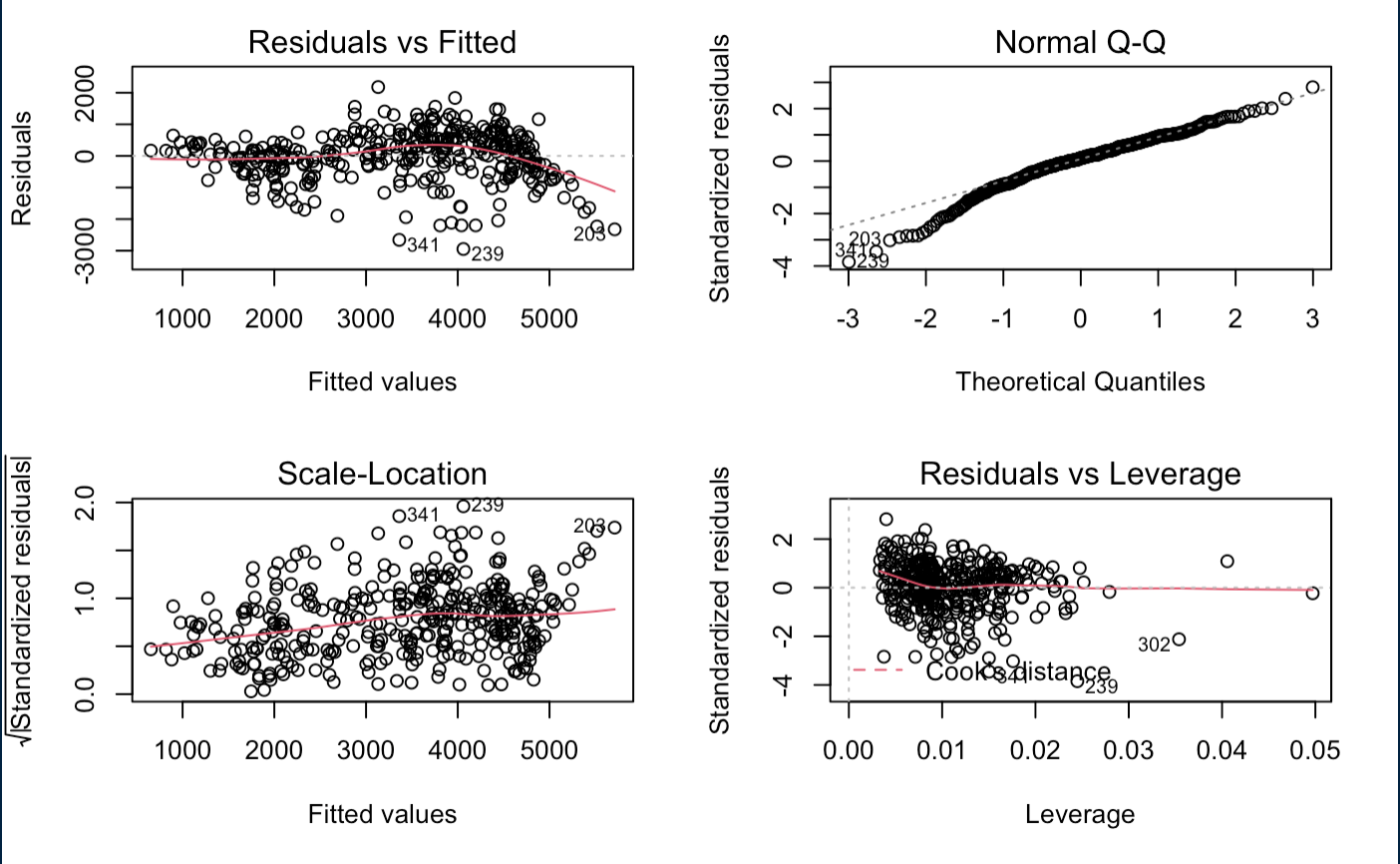


Figure D. These plots are directly from R, and are plots from our first model. Given the lack of randomness and tailing in our residuals vs. fitted and Normal Q-Q plots respectively, it can be concluded that the assumptions of independence and normality for our model’s errors were not being satisfied with this model.

**Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
 (Intercept) -530.24 301.77 -1.757 0.0798   
 atemp 11373.67 1402.12 8.112 7.94e-15 \*\*\*  
 I(atemp^2) -6414.88 1476.03 -4.346 1.81e-05 \*\*\*  
 season 268.21 41.39 6.479 3.03e-10 \*\*\*  
 windspeed -2421.12 534.21 -4.532 7.96e-06 \*\*\*  
 ---  
 Residual standard error: 757.2 on 360 degrees of freedom  
 Multiple R-squared: 0.7017, Adjusted R-squared: 0.6984  
 F-statistic: 211.7 on 4 and 360 DF, p-value: < 2.2e-16**

Figure E. This is direct R output for our second model with the quadratic term for average temperature. This model can be seen as significant based on the produced F-statistics and p-value, but a mediocre R-squared and standard error still could be improved upon.

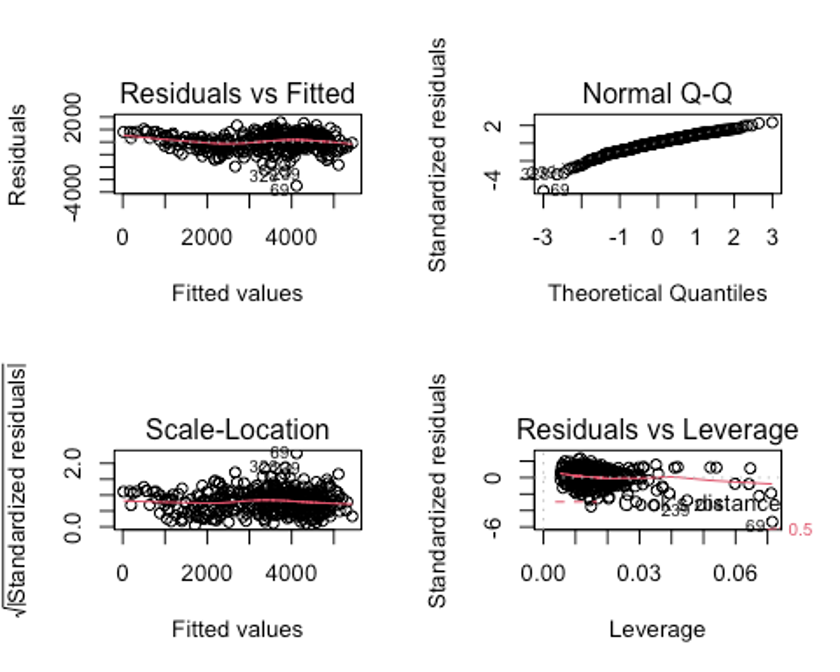


Figure F. These plots are directly from R and coincide with the model from Figure E. These plots have now satisfied our assumptions of independence and normality for our errors, as evidenced by the random spread and lack of tailing in our residuals vs. fitted and Normal Q-Q plots respectively.

**boxplot.stats(cnt)$out**

**integer(0)**

Figure G. This is output directly from R, showing that no data points were removed from our model when checking for outliers.

**model2 <- lm(cnt ~ atemp + I(atemp^2) + I(atemp^3) + hum + registered)**

**head(predict(model2, newdata = BikeData12, interval = 'confidence'))**

**day fit lwr upr  
1/1/2012 2146.765 2063.060 2230.469  
1/2/2012 2052.856 1949.665 2156.047  
1/3/2012 2181.191 1961.273 2401.109  
1/4/2012 2314.113 2077.438 2550.788  
1/5/2012 3300.983 3185.944 3416.022  
1/6/2012 4095.499 3968.764 4222.234**

**mean(predict(model2, newdata = BikeData12))**

**Predicted Number of Daily Users for 2012 = 5144.306**

**mean(BikeData12$cnt)**

**Actual Number of Daily Users for 2012 = 5599.934**

Figure H. This is direct R output displaying the second prediction we conducted for 2012 bike rental data. As evidenced by the predicted average and actual average being much closer together, this model was a much more accurate predictor for 2012 bike rental data, but also potentially suffers from overfitting within the data.

1. **References**

*Fanaee-T, Hadi, and Gama, Joao (2013). Bike Sharing Dataset [Data set]. University of California, Irvine (Porto Campus).*

*http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#*