Predictive Analysis on Bike Sharing Registrations

Minh Thu Bui, Julia Gallini, Bolin Chen, Ziyi Shao, Xizhan Tan, Ze Li

Section 3

December 12, 2022

Abstract

Bike sharing has become popular in major cities around the world due to its convenience and positive effect on the environment. The present study aims to build a prediction model for volume of future bike rentals based on the investigation of key factors that influenced this outcome in 2011 and 2012. The data were collected on a daily basis along with climate information such as temperature, weather, season, day of the week, humidity, wind speed. We fit four multiple linear regression models using temperature, weather, and season as predictors and total number of bike rentals for casual weekday users, registered weekday users, casual weekend users, and registered weekend users as our four outcomes. Our models predicted bike rentals in 2012 relatively well though with room for improvement.

1 Introduction

Bike sharing systems are automated shared transportation systems in which individuals can rent bicycles for a short term with relatively low cost. Bike sharing systems can be either node-like or node-free form. The node-like form enables the system to be highly organized since users are required to pick up and return the shared bicycles at specified docks within the system (e.g. Bluebikes in the greater Boston area), and the node-free form offers more convenience for users because users can return the bicycle at any place in the city (e.g. Ofobikes in mainland China). Currently, there are over 500 bike-sharing programs around the world using a total of over 500,000 bicycles [1]. Both forms of bike sharing systems create social benefits like transportation flexibility, reductions in vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals [2]. However, there are issues related to node-free bike sharing such as oversupply, resulting in traffic congestion and waste in resources. Therefore, it is essential to analyze the bike usage data in the past and find the relationship of the usage to the environmental and seasonal settings to identify the optimal supply number of bikes available at various times of the year.

Our main interest is to estimate the total number of bike users in different circumstances, focusing on daily casual bike users and registered users. In considering predictors of bike usage, it is obvious that weather information has a strong influence on the number of bike rentals and the frequency of usage [3]. Temperature, weather, humidity, and wind speed were collected as a part of this data set and each of them evaluates a distinct component of a day's general climate. Moreover, bike users may have varying usage patterns depending on the days of the weeks, so we investigate this relationship as well in this study. Lastly, the paper also aims to investigate the relationship between season and the number of rentals since riders might have different behaviors depending on the time of year separate from the weather on a given day.

This paper will be organized as follows. In section 2, we introduce the background of our research, including data description. In section 3, we generate our models and conduct statistical analysis on the graphs and variables. In section 4, we make predictions for the level of bicycle rental in 2012 based on our training set. Additionally, section 5 includes our discussion on our analysis and present conclusions, while section 6 contains extra figures and references that we used to make the analysis.

Author contributions: J.G, Z.L: data preprocessing, initial analysis, modeling on training data; B.C, X.T: model analysis and diagnostics; Z.L, B.C, X.T: validation of prediction and diagnostics; Z.S, X.T: abstract, introduction, and background; J.G: edited the paper; M.T.B wrote and edited the paper.

2 Background

Due to the increasing trend in the usage of bike sharing systems in metropolitan cities around the world, real-time data regarding the bike sharing systems are readily available to be used for analysis. Our data set focuses on the number of casual and registered users in bike rentals. Other variables collected include seasons, rentals on weekdays and weekends, temperature, wind speed, etc. The data regarding weather information is gathered from multiple sources from the Washington D.C. area in 2011 and 2012 [4], [5], [6]. Using these data we aim to determine key factors in predicting bike rental usage so that bike sharing companies may accurately balance their inventory with customer demand and maximize profits.

3 Modeling and Analysis

We split our data into two data sets: data from 2011 were used to train our models and data from 2012 were used to test and validate these models. Regarding data pre-processing, our overall data set contained no null values and all values are input in appropriate numerical formats consistently for each variable. To begin the modeling process on the training data it is useful to fully understand the univariate relationships between each variable, i.e. potential predictors and outcomes, using a correlation matrix. After examining the list of all possible predictors in our data set, we decided to investigate the predictive strength of following variables: temperature, season, holiday, weekday, working day, weather, humidity, and wind speed. We also decided to examine these relationships stratified by casual users and registered users since we suspected the effects of our predictors may vary based on these variables. We display only the correlation matrix for casual users here for conciseness as the two matrices were quite similar. Since scatter plots are not as useful for visualizing the effect of categorical predictors, we also examined bar plots for key categorical predictors stratified by casual and registered users.

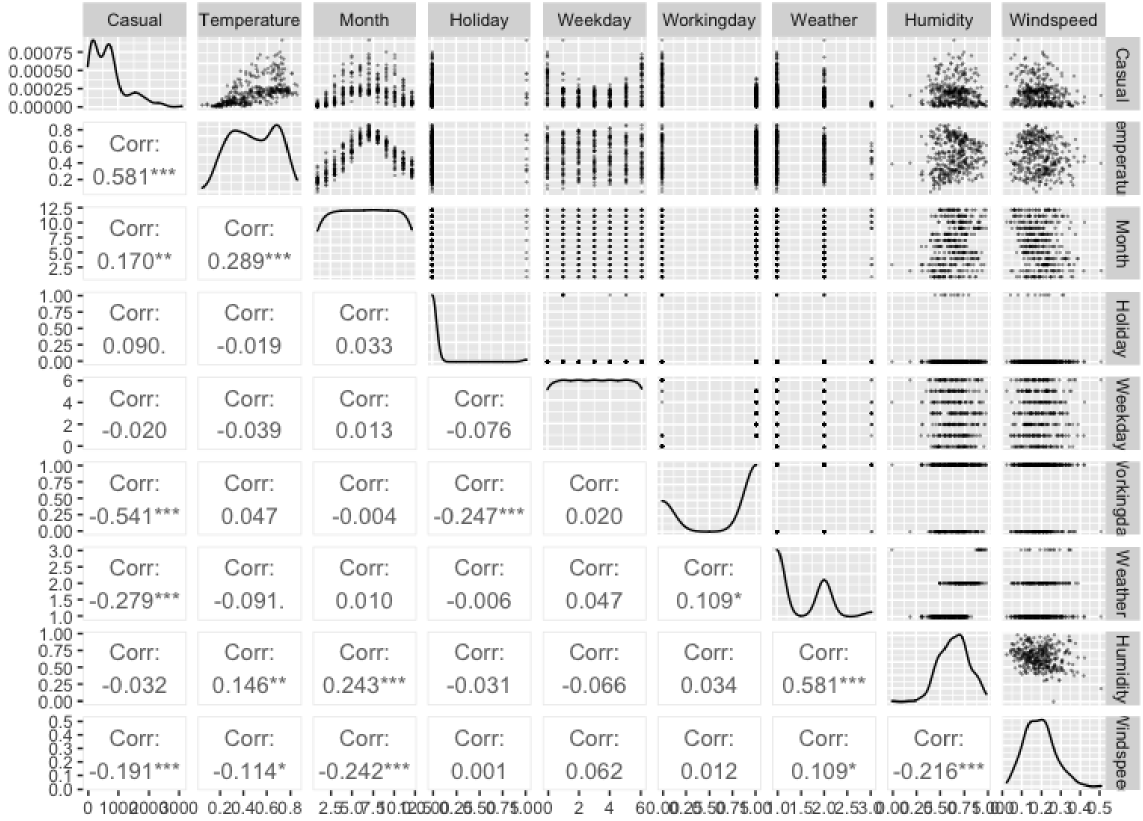

Figure 1: Correlation Matrix with Casual Rentals

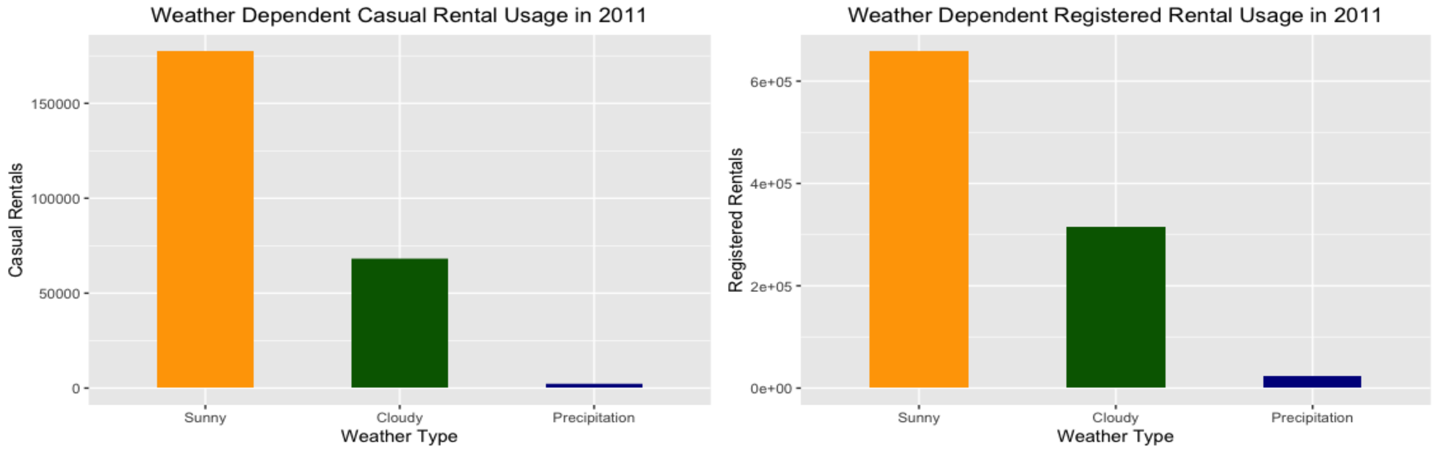

Figure 2: Casual (left) and Registered (right) Rentals in 2011 with respect to Weather Type

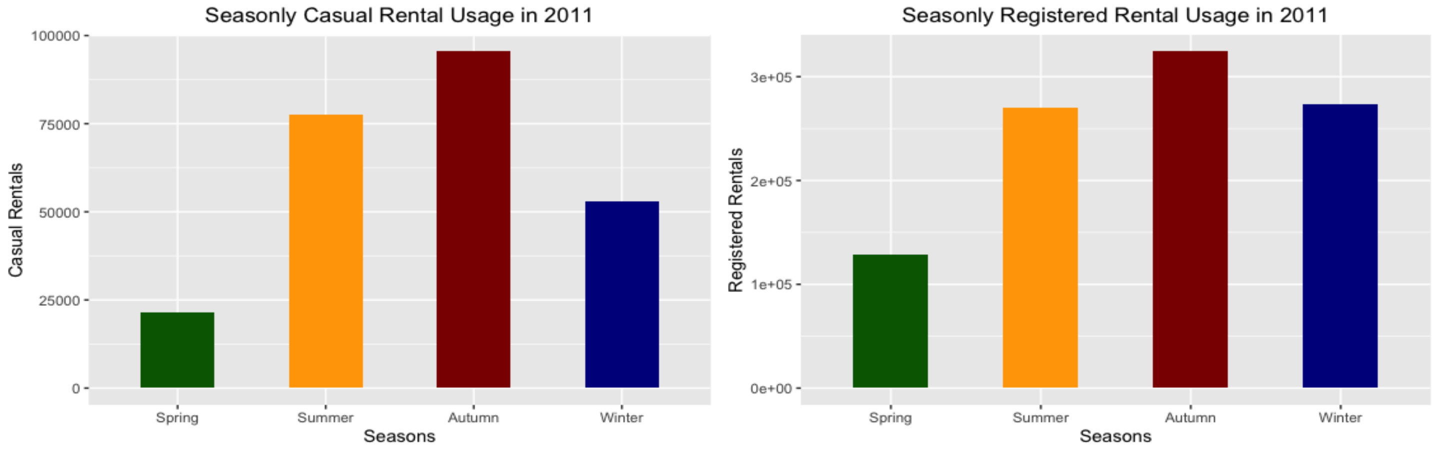

Figure 3: Casual (left) and Registered (right) Rentals in 2011 with respect to Seasons

Based on Figures 1-3, *Temperature, Season*, and *Weather* were promising candidates for the predictions of rentals based on climate factors. Specifically, temperature appeared to have reasonably strong relationships with both outcomes based on the correlation matrix. Meanwhile, *Holiday, Humidity*, and *Windspeed* did not appear to be particularly related to either outcome so we did not consider those variables for predictors in our final models. Moreover, *Working day* appeared to have an especially strong relationship in opposite directions for casual and registered users. We can observe that the number of rentals, for both casual and registered categories, is significantly higher in sunny weather than that in cloudy and precipitation conditions. Intuitively, we can infer from this trend that people tend to use bikes as a means of transportation when the weather is more favorable. From Figure 3, we observed that the number of rentals were higher in the summer and autumn compared to in the spring and winter. We suspected this trend happened due to the same reason as mentioned above for weather types, as well as temperature preferences. Generally, the temperatures in the summer and autumn are higher than in the spring and winter in Washington D.C., and since people are likely to prefer high temperatures over low temperatures when biking that could factor into their renting decision. To further understand the relationships between different types of ridership and daily usage, we then decided to create four different models each with a different outcome: total daily bike rentals for casual weekend users, casual weekday users, registered weekend users, and registered weekday users. Moreover, based on Figure 3 we suspected the prediction of bike rentals using temperature could be improved using a quadratic term $Temperature^2$, plotted stratified by casual and registered users in Figure 4.

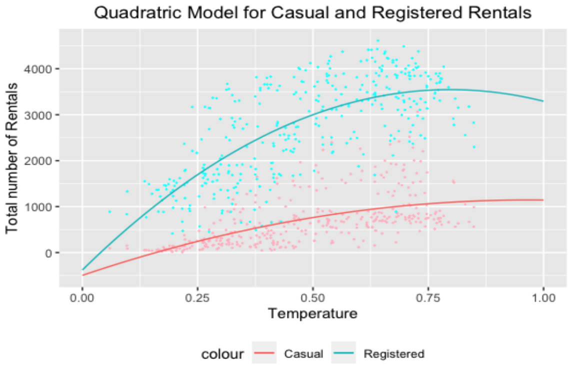

Figure 4: Model Performance with $Temperature^2$ Quadratic Term

Visually the quadratic prediction for rentals shows an upward trend as the curve maps better to the tends at higher temperature values, supporting the use of a quadratic term as opposed to only a linear term.

Overall, there is a similar univariate trend in both casual and registered rentals for all three weather types (sunny, cloudy, and precipitation). The number of rentals reaches the highest during sunny weather while it reduces dramatically during precipitation (Figure 2). We observed that the number of rentals in both categories follow a similar trend with respect to season, specifically the maximum rentals reach the highest in the fall and least during the spring. On the other hand, the number of registered rentals in the winter does not drop significantly compared to that of casual rentals in 2011 (Figure 3). We considered log-transforming the data, though this did not improve the normality of our outcomes at all, so we abandoned this approach for these data. After the visual univariate analysis of the relationships between temperature, weather types, and seasons on the casual and registered rentals in 2011, we explored the these relationships in four multivariate models using the four outcomes described previously. We included predictors: $Temperature^2$, $Temperature$, $Weather$, and $Season$ in each of the four models based on the univariate analyses. Moving forward we will refer to the casual weekday model as Model 1, the registered weekday model as Model 2, the casual weekend model as Model 3, and and the registered weekend model as Model 4 for simplicity. Displayed below is the result from the fit of Model 1.

	Estimates	Standard Error	t-value	Pr > ($\|t\|$)
Intercept	-331.668	69.030	-4.805	2.72e-06
Temperature squared	-1680.657	359.169	-4.679	4.79e-06
Temperature	2599.024	353.670	7.349	3.06e-12
Weather (Sunny)	Reference			
Weather (Cloudy)	-147.848	20.516	-7.207	7.26e-12
Weather (Precipitation)	-386.514	44.499	-8.686	5.70e-16
Season (Spring)	Reference			
Season(Summer)	107.202	37.482	2.860	0.00461
Season(Fall)	136.453	46.914	2.909	0.00397
Season(Winter)	9.418	34.547	0.273	0.78539

Table 1: Summary of Model 1

The F-statistic of model 1 is 112.1 with p-value less than 2.2e-16, which means that the overall model is significant (at least one beta value is not 0). The R-squared is 0.7643, which says that 76.43% of the variance in the training set is explained by the model. Each variable has a p-value less than 0.05 except for the winter level of season. This could indicate that there is not a significant difference in the number of casual weekday users between winter and spring after adjusting for temperature and weather. We believe the T statistics and F statistic to be valid for this model since we observed noticeable visual patterns in the univariate relationships between each of these predictors and the outcome. Next, it is essential to check constant variance and normality of the model, using standardized residuals and $\hat{y}$ for fitted values. The results are illustrated in Figure 6 below.

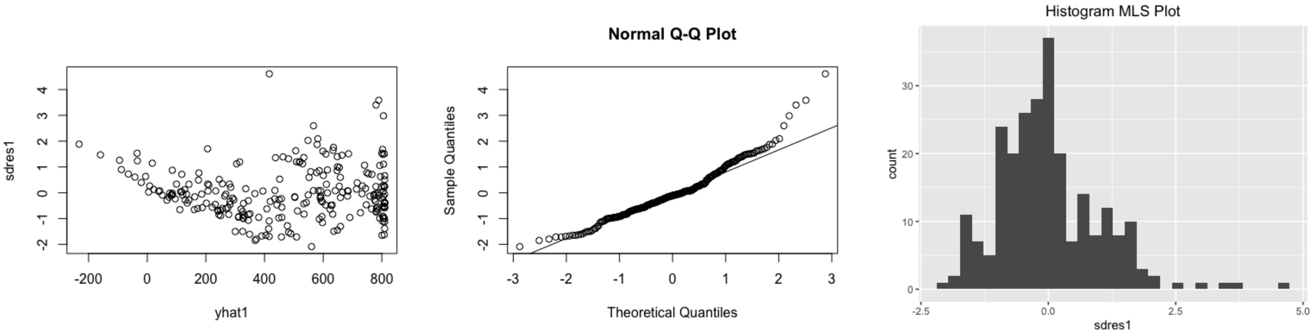

Figure 5: Constant Variance (left) and Normality check (center and right) of model 1

The constant variance assumption is violated since the standardized residuals display a fanning pattern in Figure 5. We observe minor violations of the normality assumption based on the QQ plot and the histogram in Figure 5, though we previously determined that log-transforming this data did not improve the normality. Similarly, for model 2 that focuses on registered rentals during the workdays, all predictor variables are significant since their p-values are under 0.05 but the the normality assumption is violated because the histogram is also highly-skewed. Meanwhile, in model 3 for casual rentals on the weekends, all variables are statistically significant except for variable that represents the winter season. This could indicate that the winter season has no significant impact on the number of casual rentals on the weekends, as seen in model 1. On the other hand, model 4 for registered rentals on the weekends appeared to several assumption violations with two outliers having sample quantiles of around -4 as well as a nearly bimodal outcome distribution. Each predictor variable in this model has a p-value less than 0.05 except for the quadratic term of the temperature. This could suggest that we should drop the quadratic term in this model specifically. Code snippets for step-by-step modelling and analysis are in section A of the appendix. Graphs and plots for constant variance and normality check for model 2, 3, and 4 are included in section B of the appendix.

4 Prediction

In section 3 we fit our models on the training data set and assessed the performance of each model. In this section, with the validation data set, we attempt to make predictions and determine how generalizable our models are using all the variables we have used above, $Temperature^2$, $Temperature$, $Weather$, and $Season$. To quantify the performance of each model on the validation data set, Mean Squared Errors on the training set (MSE Training) and validation set (MSE Validation) are calculated. In addition, Relative Mean Square Error (RMSE) is considered so as to normalize MSE. We also included the visualization of Model 1's performance in fitting validation data with the model.

	MSE Training	MSE Validation	RMSE
Model 1	21502.15	171077.2	0.5424973
Model 2	217478.9	5241667	0.5348757
Model 3	143526.9	811033.6	0.4633653
Model 4	195725.8	3048447	0.5184678

Table 2: MSE and RMSE of all four models

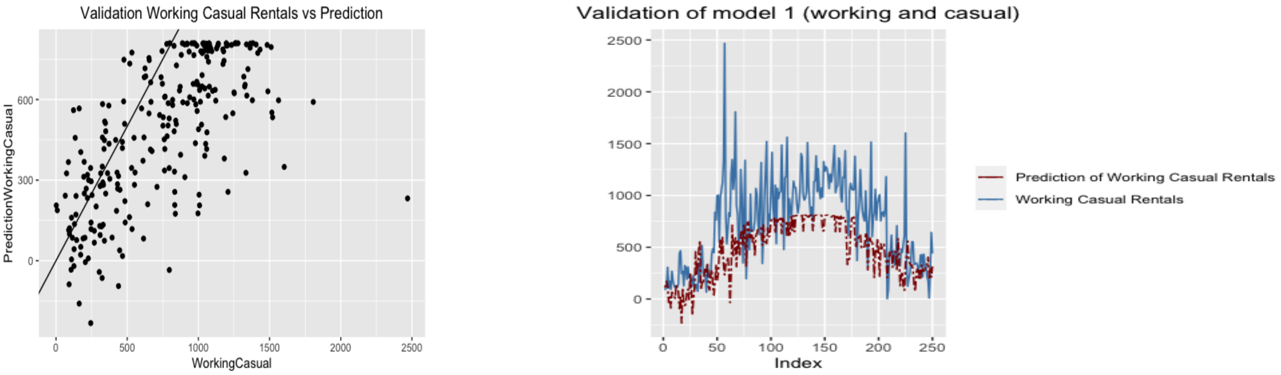

Figure 6: True and predicted casual rentals during the workdays in Model 1

In Model 1, the prediction illustrated on both graphs fits relatively well to the actual data, though our model slightly under-predicts total rentals for 2012. We attempt to address this through a time-series analysis in the appendix as the under-prediction problem persists across all four models. Specifically, in the second line plot in Figure 6, most of the extreme values are captured by our model, suggesting a pretty good fit. Moreover, the validation set has an MSE of 171077.2, and the RMSE is about 0.5. These values indicate that the model predicts the casual rentals on working days decently. Similarly, for model 2, the validation set has a MSE of 5241667, which is higher than that of the training set, and the RMSE has a value around 0.5. This could be explained by the fact that that the number of registered rentals on a working day is larger with higher variance. On the other hand, model 3's prediction did not perform as well. Although RMSE shows that the model performs well on this validation data set, the graph shows that some peaks in rentals were not captured by the prediction model. The validation of model 4 shows our model works well aside from the under-prediction since most of the extreme values are captured by our model. Our prediction for registered rentals on the weekend fits well with the actual data. Also, in further comparisons, the validation of model 4 shows our prediction model works well since most of the extreme values are captured by our model. Moreover, the validation set has an MSE of 3048447, which is not bad compared to that of the training set, and an RMSE of 0.5184678 indicated that the model performs really well on predicting the registered rentals on the weekend. Prediction figures for models 2, 3, and 4 are also included in section C of the appendix.

5 Discussion

Our data set contains weather and climate information for bike rentals, both casual and registered, in 2011 and 2012 in Washington D.C. With four predictors ($Temperature^2$, $Temperature$, $Weather$, and $Season$), the analysis focuses on casual and registered rental predictions during the weekdays and the weekends. Except in model 3, the quadratic temperature term proves to be a valuable variable in our predictions for general rental trends in both categories on all days of the week. Overall, from the graphical presentations above, the model performances for casual and registered rental predictions during the weekends are better than those during the weekdays. Based on these model diagnostic techniques we feel we have fit four relatively good models, though there is always room for improvement especially with the models on the weekdays. In the appendix, we explore more sophisticated methodology such as adding a time-series component to account for the consistent under-prediction of our models of the actual 2012 data, as well as an XGBoost model to explore machine learning prediction capabilities.

From a business perspective, the modeling and analysis has given us insights into the underlying patterns and characteristics of bike sharing systems in Washington D.C. in 2011 and 2012, which can be used for business decision-making in the future. The inferences on bike rentals based on weather, season and temperature have been carefully studied by both programming and visual illustrations in our paper. Our analyses can contribute to the improvement of bike sharing systems so as to maximize companies' profits and increase customer satisfaction. Based on our results businesses can modify their plans based on the current weather, temperature, and season. For example, to increase bike usage in non-peak seasons (spring and winter), a company can offer discounts on rental prices during those seasons, or a more flexible pricing plan depending on their pricing policy. Another example of a policy that could improve business performance is to perform maintenance activities during non-peak time periods such as spring or winter. Future directions of this research include expanding the modeling process to capture data from other cities, particularly those with different weather from Washington D.C. to see if the same conclusions hold and to improve the generalizability of these findings.

References

[1] H. Fanaee-T and J. Gama, “Event labeling combining ensemble detectors and background knowledge,” *Progress in Artificial Intelligence*, vol. 2, no. 2, pp. 113–127, 2014.

[2] A. Nikitas, “The global bike sharing boom–why cities love a cycling scheme,” 2016.

[3] E. Eren and V. E. Uz, “A review on bike-sharing: The factors affecting bike-sharing demand,” *Sustainable Cities and Society*, vol. 54, p. 101 882, 2020.

[4] *System data*, `https://ride.capitalbikeshare.com/system-data`, Accessed: 2022-11-20.

[5] *Holiday schedule*, `http://dchr.dc.gov/page/holiday-schedule`, Accessed: 2022-11-20.

[6] *Weather information*, `http://www.freemeteo.com`, Accessed: 2022-11-20.

Appendix

A Code Snippets for Model and Analysis:

1. Scatter Matrix and Weather Impacts on Rentals: (To avoid repetition, we will not attach code for impact on registered rentals):

```
# Plot scatter matrix
data1 <- data.frame(Casual,Temperature,Month,Holiday,Weekday,Workingday,Weather,Humidity,Windspeed)
ggpairs(data1, upper = list(continuous = wrap("points", alpha = 0.3, size=0.1)),
        lower = list(continuous = wrap(’cor’, size = 4)))
# bar plot with Weather
ggplot(Train, aes(x=Weather,y=Casual)) +
  geom_bar(stat="identity") +
  scale_x_discrete(limit = c("Sunny","Cloudy", "Precipitation")) +
  labs(title="Weather Dependent Casual Rental Usage in 2011",
       x ="Weather Type", y = "Casual Rentals")
```

2. Quadractic Model for Casual and Registered Rentals:

```
bike_sharing_casual.quadls <- lm(casual~actual_temp + I(actual_temp**2),data=Train)
bike_sharing_registered.quadls <- lm(registered~actual_temp + I(actual_temp**2),data=Train)
XNew<-seq(0,1,len=length(Train$actual_temp))

predQuad1  = predict(bike_sharing_casual.quadls,newdata=data.frame(actual_temp=XNew));
predQuad2  = predict(bike_sharing_registered.quadls,newdata=data.frame(actual_temp=XNew));
```

3. Model 1 Setup and Summary: (for simplicity and conciseness, we will attach code for Model 1 only from now on)

```
Working <- Train[Train$workingday==1,]
Weekend <- Train[Train$workingday==0,]
m.mls1 <- lm(Working$casual ~ I(Working$actual_temp^2) + Working$actual_temp +  factor(Working$weathersit)
     + factor(Working$season))
```

4. Constant Variance and Normality Check for Model 1:

```
# Check the constant variance for model 1
sdres1 = rstandard(m.mls1)
yhat1 = m.mls1$fitted.values
plot(yhat1, sdres1)
# Check the normality for model 1
qqnorm(sdres1)
```

```
qqline(sdres1)
ggplot(data = data.frame(sdres1), aes(x = sdres1)) + geom_histogram(bins = 30) + ggtitle("Histogram MLS Plot")
```

5. Prediction Diagnostics on Model 1:

```
# Residuals for training data of model 1 (working and casual)
ResMLS1 <- resid(m.mls1)
# Change to Test Data of Working
Working2 <- Test[Test$workingday==1,]
# Prediction for validation data
output1<-predict(m.mls1, se.fit = TRUE, newdata=data.frame(Working2$casual,Working2$actual_temp,Working2$weathersit,Working2$season))
ResMLS1Validation <- Working2$casual - output1$fit
# Mean Square Error for training data
mean((ResMLS1)^2)
# Mean Square Error for validation data
mean((ResMLS1Validation)^2)
# Relative Mean Square Error for validation data # to normalize mean sq error
mean((ResMLS1Validation)^2) / mean((Working$casual)^2)
```

B Constant Variance and Normality Check:

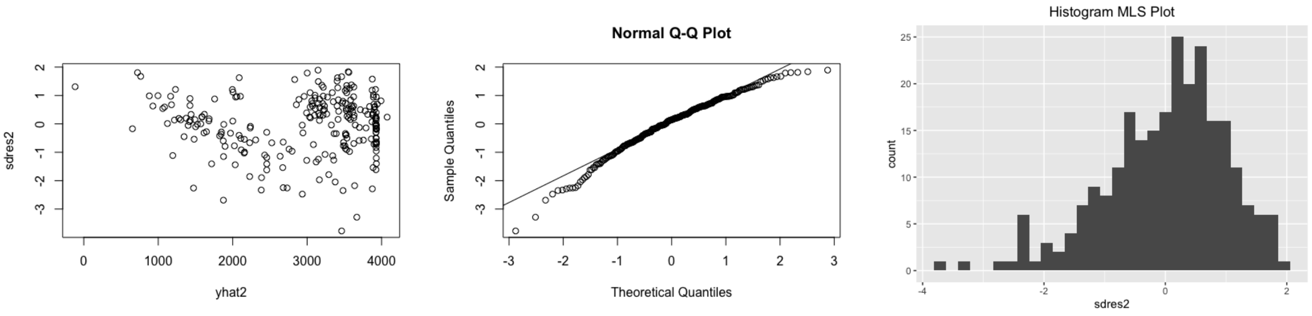

Figure 7: Constant Variance (left) and Normality check (center and right) of model 2

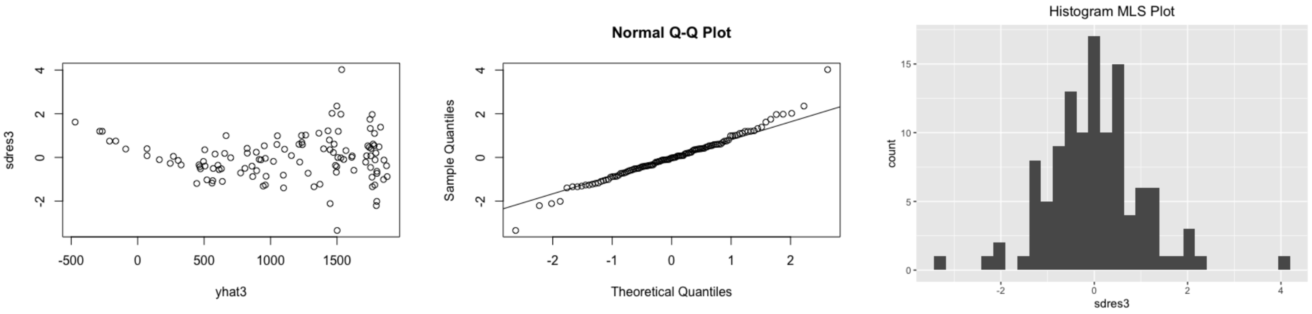

Figure 8: Constant Variance (left) and Normality check (center and right) of model 3

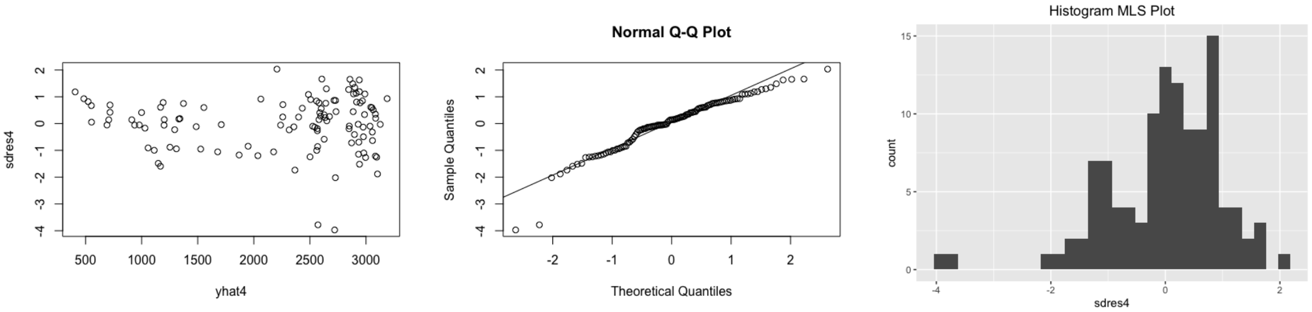

Figure 9: Constant Variance (left) and Normality check (center and right) of model 4

C Visual illustrations of model performances:

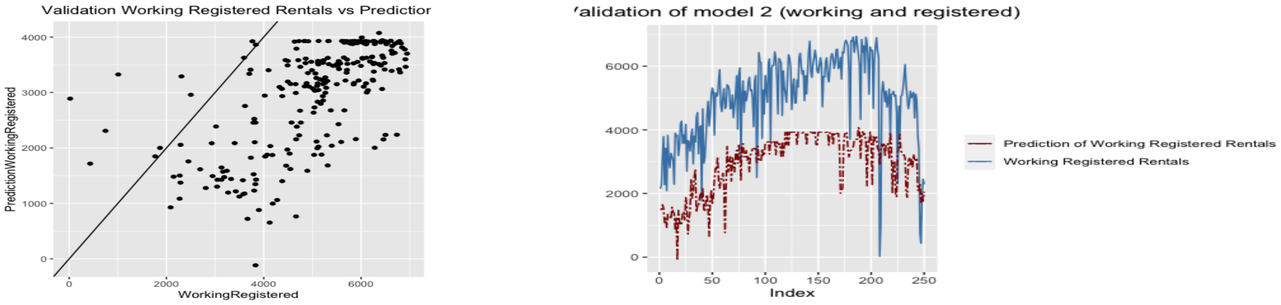

Figure 10: True and predicted registered rentals during the workdays in model 2

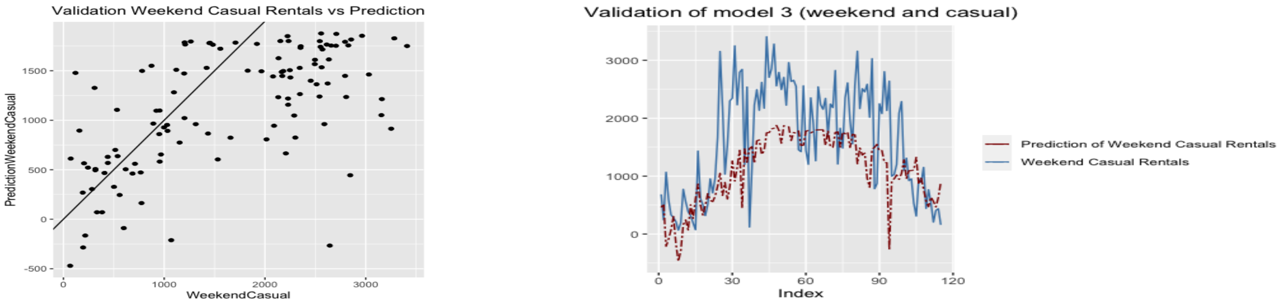

Figure 11: True and predicted casual rentals on the weekends in model 3

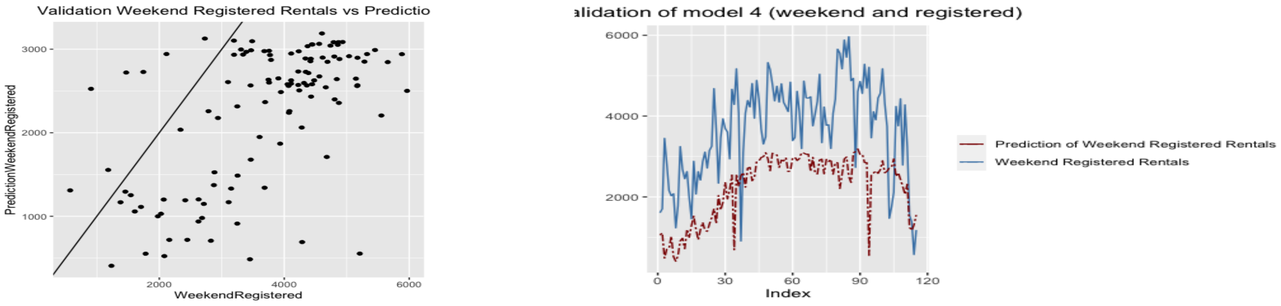

Figure 12: True and predicted registered rentals on the weekends in model 4

D Time Series Analysis:

We observed that all of our prediction models for 2012 underestimated the true total bike rentals throughout the year. To lower the risk of underestimation, we performed a time-series analysis of the data set. We separated the data into four groups and performed the correlation matrix for each of them. The correlation between time and casual rentals and registered rentals during the weekdays are 0.203 and 0.534, respectively. Meanwhile, the correlation for casual rentals on the weekend is 0.216, and the correlation for registered counts for the weekends is 0.464. All the coefficients are significant and the correlation between registered users and time is quite strong. For simplicity, we will attach the figures for model 1 only.

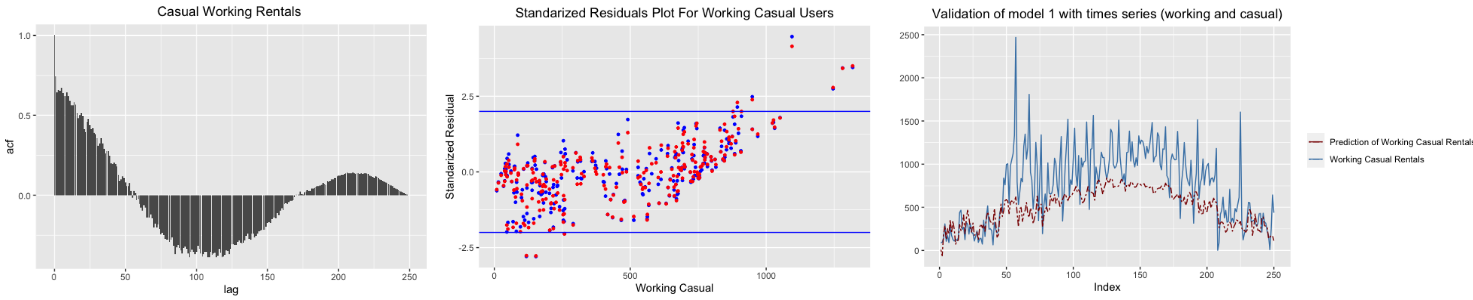

Figure 13: Autocorrelation (left), standardized residuals (middle), and model prediction (right) for model 1

Comparing the standardized residual plot of the quadratic and the GLS model from the second plot in Figure 13, we did not see a significant difference in the scattering patterns, which could imply that our predicting variables and model work well in predicting total bike rentals. One limitation to note with this model is that the AR(1) model is usually applied when there exists an exponential pattern in the ACF lag plot (far left). However, an AR(2) model would be a better model in this case as our plot shows negative ACF values. We also observe a cosine pattern annually, which is likely due to the seasonality of our data set. Despite these limitations, when compared with the original model 1 our model does a better job at adjusting for the increase in bike rentals in 2012 based on the prediction plot on the far right of Figure 13, particularly at the beginning and end of the year. There is still some under-prediction during the middle of the year, suggesting that having a few more years of summer data would be beneficial to build an even more accurate prediction model.

E Extreme Gradient Boosting (XGBoost):

Another possible model improvement method we considered was to apply a machine learning library called XGBoost to our model. This model gives us an MSE of 3.069127 for the training set and 839763.7 for the validation set. Our prediction is extremely good when applying the XGBoost regression model with a relative MSE score of 0.0243. We speculate the reason that the model predicts worse during summer and fall is that the model fails to predict the annual increase in rental usage, an issue we dealt with in all of our models of this data. Again, if we provided the model more data, we believe that the XGBoost regressor model would perform even better.

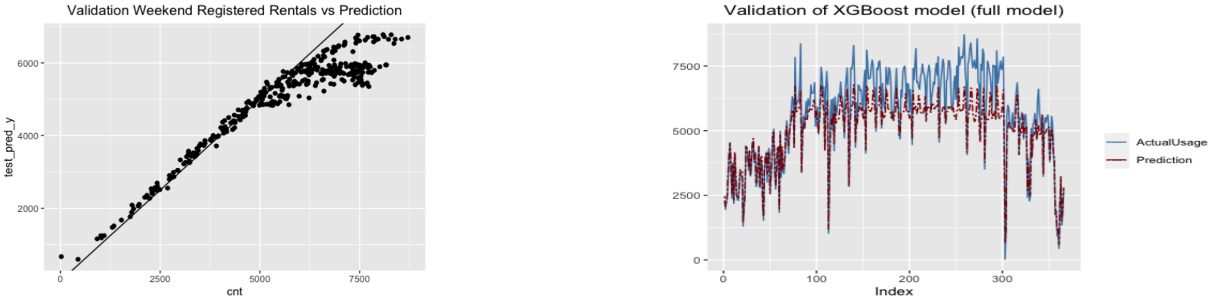

Figure 14: True and predicted rentals using XGBoost

Our prediction in the second graph of Figure 14 fits extremely well with the actual data, resulting in the most accurate prediction of all of the models fit in this paper. A possible improvement on this XGBoost algorithm is to incorporate a times series component as implemented above. However, a limitation of the XGBoost method is that machine learning models like this are a bit of a black box in terms of how the algorithm functions, so while the prediction is very strong it is hard for us to say exactly which predictors are the strongest contributors and why.