Analyzing & Predicting Bike Sharing Rentals

Hana Shehadi, Bennett Solomon, Gia Nguyen, Khanh Nguyen, Yimin Wu
Lab Section C1, Group 5

Abstract

The goal of this report is to predict total bike rentals for upcoming years. To achieve this goal, this report includes evaluations and analyses of the factors that impact the volume of bike rentals. Following the modeling and analysis conducted using RStudio, including correlation matrices, summary statistics, diagnostic analyses, and training and validation, this report employs data from the 2011 bike rental trends to predict the trends for 2012 bike rentals. The findings indicate that temperature, season, weather, and wind speed are the most influential factors determining whether a user will rent a bike or not. The model created with these findings turns out to be a pretty good predictor of bike rental trends.

1. Introduction

Bike sharing systems, such as Capital Bikeshare, have become the new normal in metropolitan areas. Users are able to rent bikes from one station and drop it off at any other station in the city. There's a great interest in the expansion and success of such bike-rental services for a variety of economic and environmental reasons, as well as a greater desire to relieve problems with urban development, such as traffic congestion and air pollution. Our group's problem is rooted in the proliferation of bike sharing services and the factors that contribute to people choosing to rent bikes. We find our interests aligned with those of the bike vendors, and would want our findings to be utilized for the purpose of increasing the total amount of bike rentals.

2. Background Information

Our data set comes from a study conducted at the Laboratory of Artificial Intelligence and Decision Support (LIAAD) at the University of Porto, in Portugal. This data was collected as part of a study on automated event labeling. The authors utilized the automated records of Washington DC's *Capital Bikeshare* rentals from 2011 and 2012, and collected information ranging from the time of rental to the weather at time rented. In whole, the data set includes 8 categorical variables and 4 numerical variables. The categorical variables comprise of *instance*, which logged an index of all the collected data points; *date*; *season* (spring, summer, fall, winter); *year*; *month*; *weekday,* which indicates the day of the week; *holiday*, whether or not the date of rental was a holiday; *workingday*, a logical variable that indicated whether or not the date was one where people would go to work (ie. not a weekend or a holiday); and *weathersit*, which was variable indicating whether or not it was a sunny/clear day (notated by a 1), overcast or very cloudy day (notated by a 2), if there was light rain/snow (notated by a 3), or if there was heavy rain/precipitation (notated by a 4). The four numerical and continuous variables were *hum*, which logged the humidity (normalized by dividing to 100 as a maximum); *windspeed*, logged in miles per hour and normalized by dividing to 67; *temp*, in celsius and normalized by dividing to 50; *atemp*, the feeling-temperature in celsius and normalized by dividing the initial values to 41; *registered*, the count of rentals made by registered users; *casual*, the count of rentals made by non-registered customers; and *cnt*, the count of total number of bike rentals for the day.

3. Modeling & Analysis

Before beginning our analysis, we inputted the data into RStudio and organized the dataset. Then we split the data into two groups: the year 2011 for training, and the year 2012 for validation purposes. The following section only analyzes the training dataset, the year 2011.

3.1. Correlation Matrix

Since the goal of this research is to predict the daily bike rentals for the following year, we assigned *cnt* as the response variable, which is the total count of bike rentals per day, from both casual and registered users. We presumed that the number of bike rentals is affected by the weather and other environmental factors, so we plotted the correlation matrix against *total rentals* and the four continuous variables, *temp*, *feeling temp*, *humidity*, and *windspeed.*

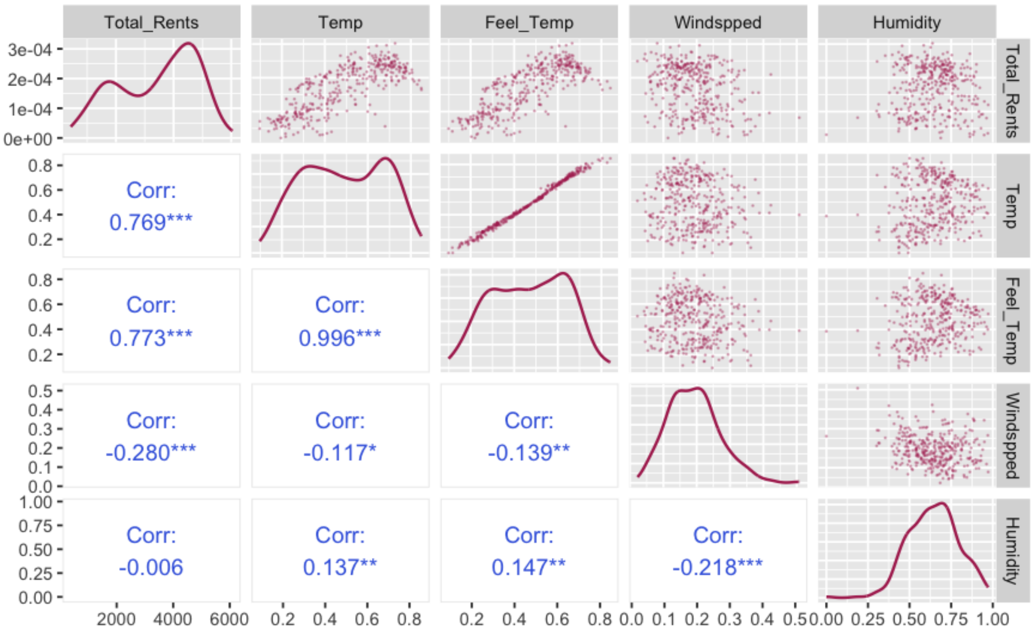


Figure 1: Scatter plot and correlation matrix against all continuous variables,*total rentals, temp, feeling temp, humidity,* and *windspeed* in 2011

As shown in Figure 1 above, temperature, feeling temperature, and windspeed have high correlations with total count of rentals. Humidity, on the other hand, does not have much influence on the total counts. Therefore, we decided to omit humidity from the model. However, a high correlation (0.996) was observed between two predictors *temp* and *feel_temp*, indicating a sign of multicollinearity. In order to fix that, we only included the true temperature, *temp***,** rather than *feel_temp* in our model as we think people generally would pay more attention to the number reported on the weather forecast than the feeling temperature on a particular day when they go out. We will include the categorical variables in later steps.

3.2. Initial Model M1

Our initial model **(M1)** consists of the two continuous variables, *temp* and *windspeed* (coefficients from Figure 3A):

$$Cnt = 1414.6 + 5448.5 \cdot Temp - 3450.7 \cdot Windspeed \qquad \textbf{(M1)}$$

After running the analysis on **M1**, we see that the p-values for both coefficients are well below 0.05, indicating they are significant to the model. However, the adjusted R-squared is only 0.6293 and the residual graph in Figure 2 below shows a non-random pattern, maybe even a quadratic or cubic pattern, as evident in the fitted line. Additionally, the Normal Q-Q plot shows that the tails of our model are a bit skewed.

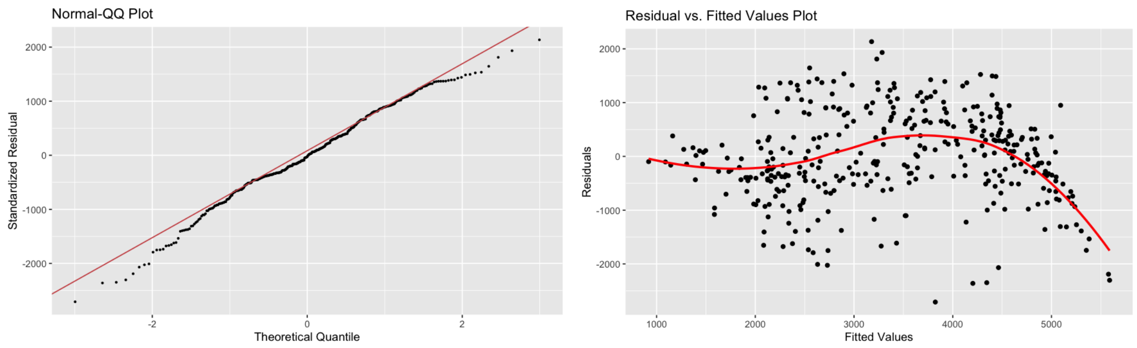


Figure 2: Residuals and Normal Q-Q plot for model **M1**

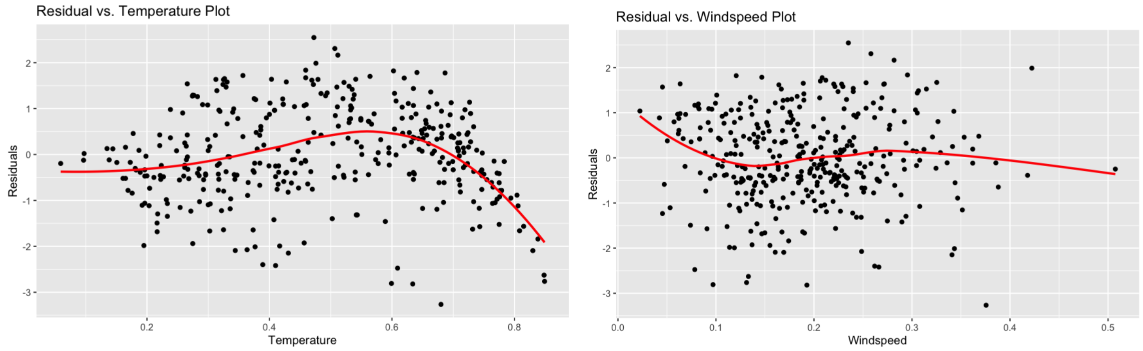


Figure 3: Standardized residuals plots against *temp* and *windspeed* for model **M1**

Since the normality assumption for model **M1** was violated, as is clear through the deviation at the edges in the Normal Q-Q plot in Figure 2, we decided to take a look into the standardized residuals plots against *temp* and *windspeed* (Figure 3), from which we observed a non-random pattern with *temp*, indicating room for improvement in the model.

3.3. Adding Higher Power Continuous Variables

Since we observed a non-linear relationship in our standardized residual graph with temperature in Figure 3, the next step was to investigate the relationship between total count of bike rentals and these higher power continuous variables: $I(temp^2),\ I(temp^3)$. After producing a new correlation matrix (Figure 1A), we picked the variables with a linear relationship to *total rentals* to be included in our next model – $temp$, $temp^2$, $temp^3$, $windspeed.$

3.4. Pattern Recognition for Categorical Variables

In addition to the continuous variables in this next model, we wanted to see what categorical variables should be added. The boxplots in Figure 4 show that there is only an observed significance difference in means for the variables *season* and *weathersit* (weather conditions), so they will be added to our next model, **M2**.

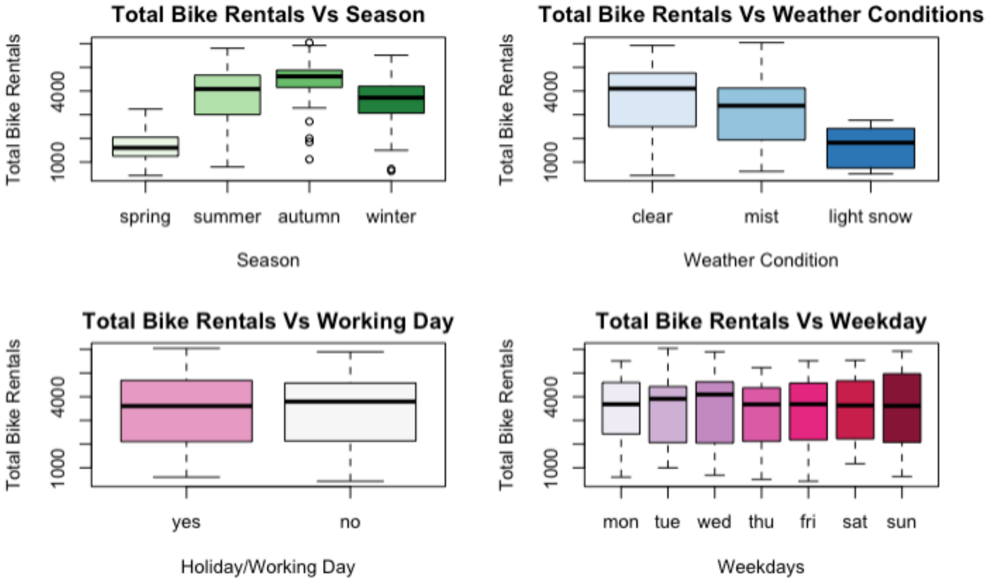


Figure 4: Boxplots for categorical variables *season*, *workingday*, *weathersit*, *weekday*

3.5. Model M2:

This leads us to our next model (**M2)** (coefficients from Figure 2A):

$$Cnt = 2571.18 - 10089.50 \cdot Temp + 41880.17 \cdot temp^2 - 34435.29 \cdot temp^3 + 823.68 \cdot summer + 981.48 \cdot autumn + 1196.54 \cdot winter - 542.67 \cdot mist - 2063.96 \cdot light\,rain - 2092.59 \cdot windspeed \qquad \textbf{(M2)}$$

3.6. Variable Selection with AIC and BIC

Now, with model **M2**, we want to verify the significance of each predictor through a series of variable selection techniques. We begin by using AIC and BIC forward selection to select the best predictors. Table 1 below shows the results (calculations are provided in Figure 10A):

Predictors	AIC **forward**	BIC **forward**
temp	4878.91	4886.68
temp + as.factor(season)	4774.4	4793.84
temp + as.factor(season) + as.factor(weathersit)	4630	4657.23
temp + as.factor(season) + as.factor(weathersit) + I(temp^3)	4581.93	4613.04
temp + as.factor(season) + as.factor(weathersit) + I(temp^3) + I(temp^2)	4526.84	4561.84
temp + as.factor(season) + as.factor(weathersit) + I(temp^3) + I(temp^2) + windspeed	4502.43	4541.32

Table 1: Values of AIC and BIC forward selection for best subset of each size

Predictors	AIC **backward**	BIC **backward**
temp + I(temp^2) + I(temp^3) + as.factor(season) + as.factor(weathersit) + windspeed	4502.43	4541.32
None		

Table 2: Values of AIC and BIC backward selection for the best subsets of each size

From Tables 1 & 2, we see that the forward and backward selection AIC and BIC agree that all predictors in model **M2** are important. Hence, we will use these predictors for our final model.

After solidifying our model with AIC and BIC variable selection, we can conclude that **M2** is our final model. Additionally, based on the R output for model **M2** (Figure 2A), we can confirm that all the coefficients are significant and the adjusted R-squared also improved to 0.8395. Next, we begin regression diagnostics to check for potential problems with our model and for potential violations of multiple regression assumptions.

Figure 4A contains plots of the standardized residuals against each predictor and the fitted values for **M2**. Each of the scatter plots shows a random pattern. Thus, **M2** appears to be a valid model for the data and does not violate the constant variance assumption for multiple regression. Figure 5A shows the diagnostic plots for **M2** - these plots also confirm that our model is valid. Figure 6A contains the recommended marginal model plots for **M2**. Notice that the nonparametric estimates of each pair-wise relationship are marked as solid curves, while the smooths of the fitted values are marked as dashed curves. The two curves in each plot match very well, thus providing further evidence that **M2** is a valid model. Figure 7A shows the added-variable plots - all continuous predictors are statistically significant. We also checked potential issues due to time correlation in our model, but they were not helpful and did not improve our model understanding. With that, we have great confidence in our final model **M2**.

4. Growth rate

After observing a significant discrepancy between our prediction and the actual values for 2012 in Figure 9A, we decided to take another look at the data and see if there were any trends that may help improve the model.

In Figure 5, it is clear that there is a great increase in the amount of rentals from the beginning of the year to June (around 150 on the x-axis). After this increase, the number of rentals stayed relatively higher than at the beginning of the year. This general increase in rentals implies an increase in registered users, rather than just rentals - also confirmed by Figure 6. In Figure 6 it is clear that there is an increase in the number of rentals made by registered users in comparison to casual users. This not only explains the reason for the general increase in total counts of bike rentals (Figure 5), but it also explains the reason behind the discrepancy in our predicted model in Figure 9A - the business was growing in user numbers. Therefore, this growth needed to be factored into our model.

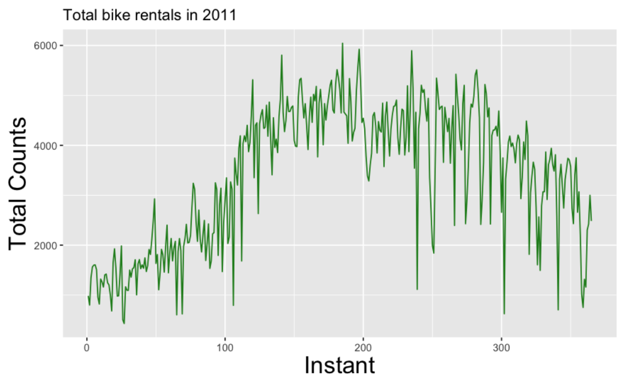


Figure 5: Plot of total bike rentals in 2011

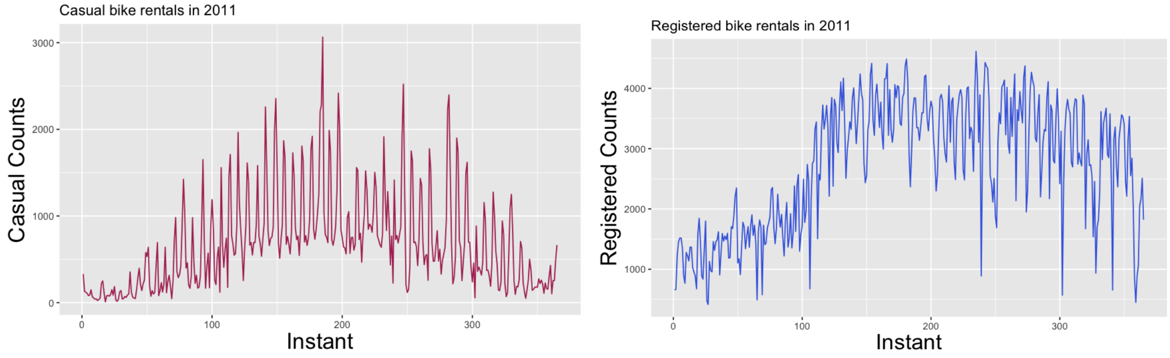


Figure 6: Plots of casual and registered users in 2011

Utilizing the fact that the growth in rentals is highly correlated with the growth in registered users with the correlation of 0.928, and to compensate for the fluctuation throughout the pre-growth and post-growth periods in 2011, we estimated the annual growth rate of 2011 as follows:

$$Growth\ rate\ =\ \frac{Average\ total\ count\ of\ customers\ from\ June\ to\ December\ of\ 2011}{Average\ total\ count\ of\ customers\ from\ January\ to\ May\ of\ 2011}\ \approx\ 1.520485$$

Since the belief that the customer base will either increase or stay relatively constant after growth, we use the calculated growth rate of 1.520485 to perform our predictions for 2012.

5. Prediction

The 2011 data and model **M2** will now be used to predict bike rental counts for the year 2012. The predictions will be compared to the true data from 2012 in order to validate our model's prediction power. Predictions in 2012 are calculated by taking the 2012 fitted values multiplied by the growth rate. The code for calculating the following metrics for training and validation are provided in Figure 8A:

MSE for training: 296667.9	MSE for validation: 789941.5
Root MSE for training: 544.6723	Root MSE for validation: 888.7866
Relative MSE for training: 0.0224777	Relative MSE for validation: 0.0228638

The RMSE for the training dataset holds for the validation dataset, and this indicates that the model **M2** is a reasonable fit and the growth factor corrected for the previous discrepancies. Figure 7 below shows the newly predicted values compared to the true values in 2012. The trends are similar in both datasets when it comes to increasing and decreasing counts, even in the extreme peaks and troughs, as shown more clearly in Figure 8.

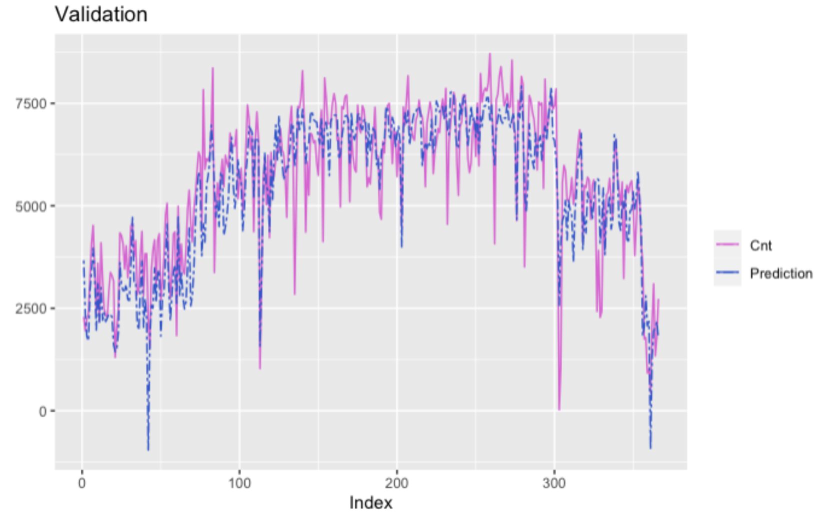


Figure 7: Plot of predicted values against true values from 2012

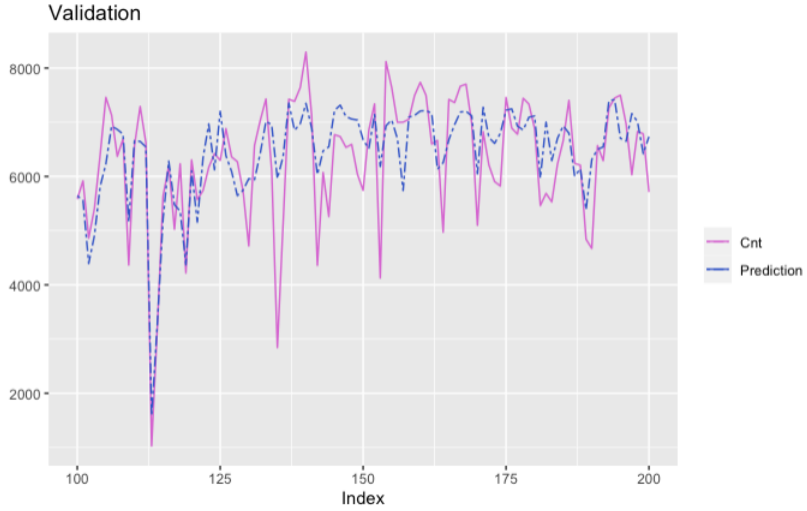


Figure 8: A zoomed in plot from Figure 7 in order to see the trends more clearly

6. Discussion

Given the fact that our model did a good job at predicting the number of bike rentals in 2012, there is now an opportunity for further analysis. With that, we investigated where the growth of total rentals comes from, or rather, how the trends within the casual and registered groups differ. Beginning with Figure 9, we see that there is a big difference between the number of bikes rented on weekdays for registered users compared to casual users.

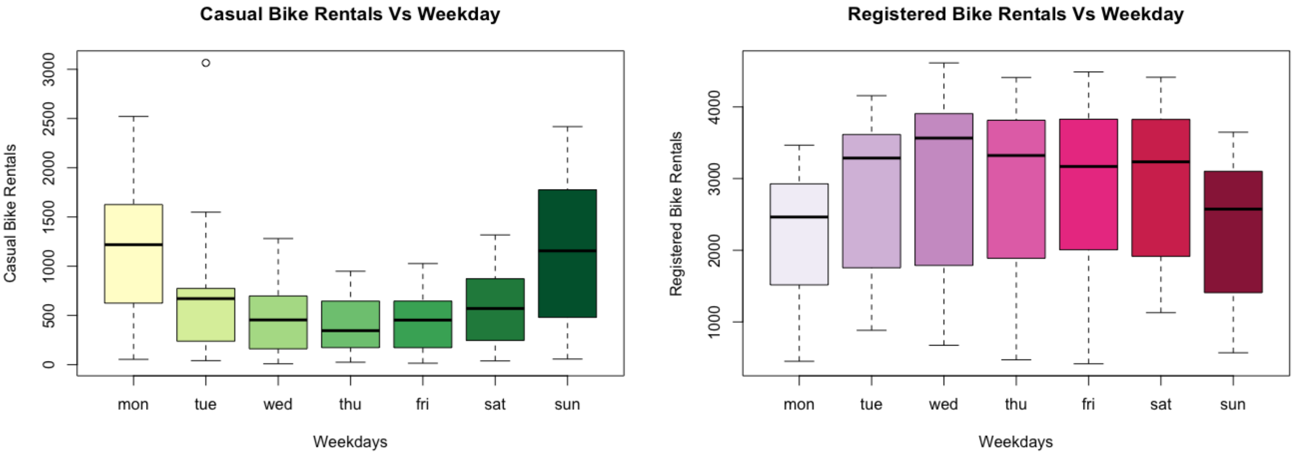


Figure 9: Boxplots of bike rentals on days of the week, comparing casual users and registered users

Supportingly, casual user rentals peak every 7 days as shown in Figure 10 - casual users tend to enjoy riding bikes exclusively on weekends.

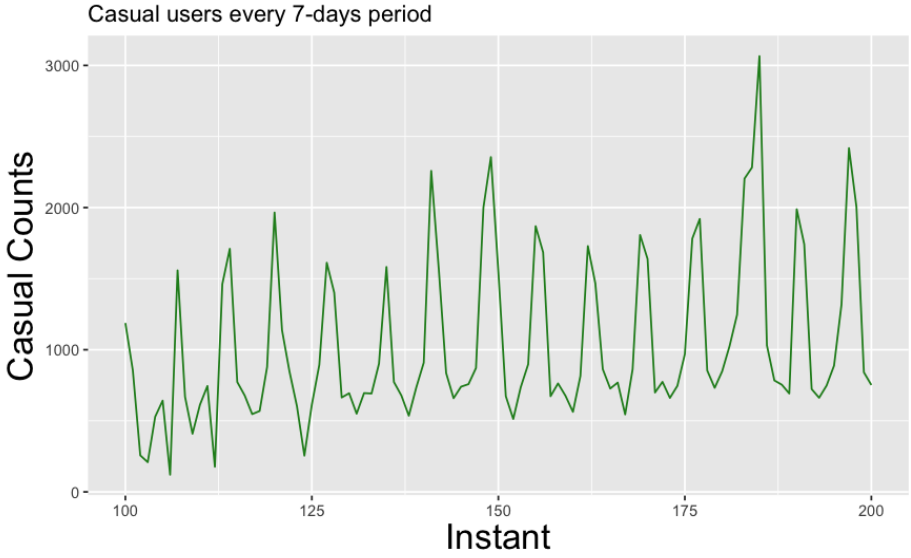


Figure 10: Instance plot for casual users, with every 7-day period represented by a peak

With these major differences in trends between casual users and registered users, it is clear that individual analysis of the two groups is necessary to truly understand patterns in rentals. Further data collection and analyses should lead to more information regarding which of the two groups should be prioritized going forward (regarding to which group has a greater potential for later growth in the marketplace).

In conclusion, the findings of all these analyses indicate that temperature, season, weather, and wind speed are the most influential factors determining whether a user will rent a bike or not and our model **M2** turns out to be a pretty good predictor of bike rental trends. With this information, our model could be utilized by Capital Bikeshare (and other bike rental services) to capitalize on seasonal/weather related trends, such as moving around kiosk locations based on the time of year (a practice already implemented by Boston's own *Blue Bikes*).

7. Contributions

All group members participated equally in the formation of this statistical report. We worked on a shared Google Document to write the report and input any information as we worked together, in-person or on Zoom, and shared opinions and ideas. Gia and Khanh worked primarily on the R code and outputs. While Hana and Bennett worked primarily on the write up and presentation of the report, along with Yimin who helped with the interpretations of the outputs. The group discussed any discrepancies on the iMessage group and worked to find an agreement on how to proceed. Everyone had equal input and the group worked smoothly.

Appendix

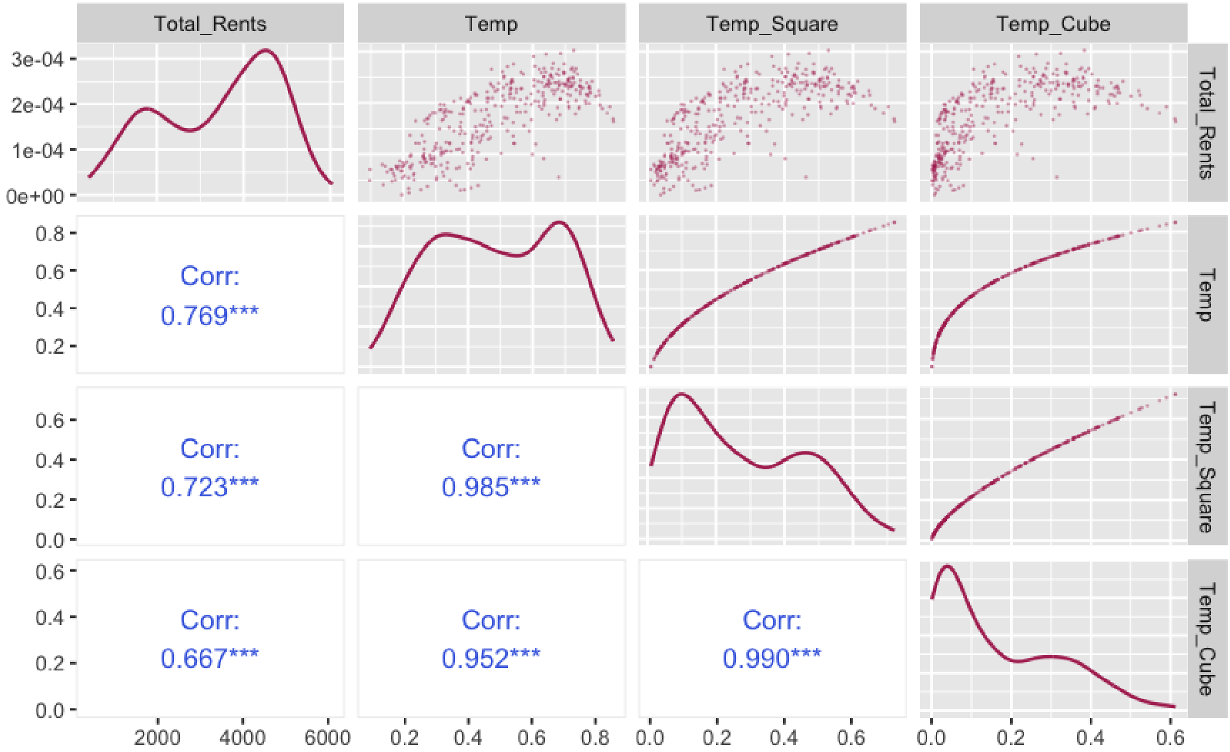
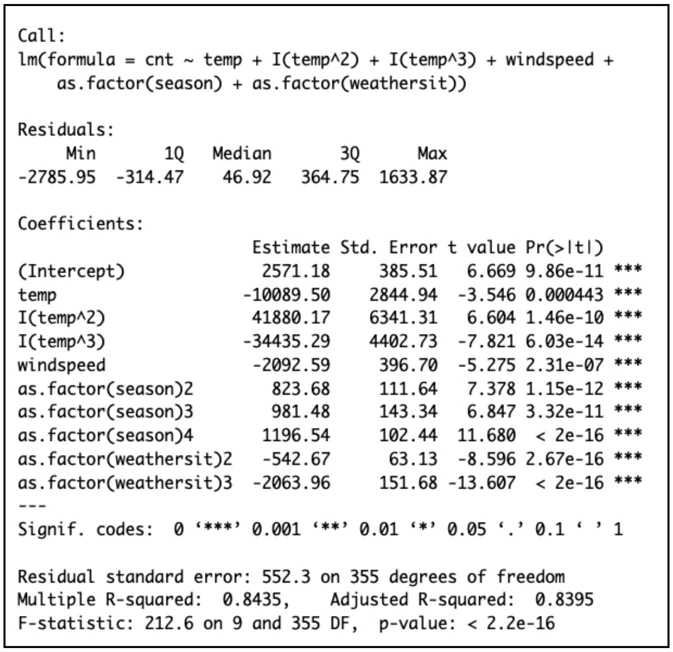


Figure 1A: correlation matrix against *total rentals* and higher power continuous variables $I(temp^2)$, $I(temp^3)$, $I(hum^2)$, $I(windspeed^2)$.



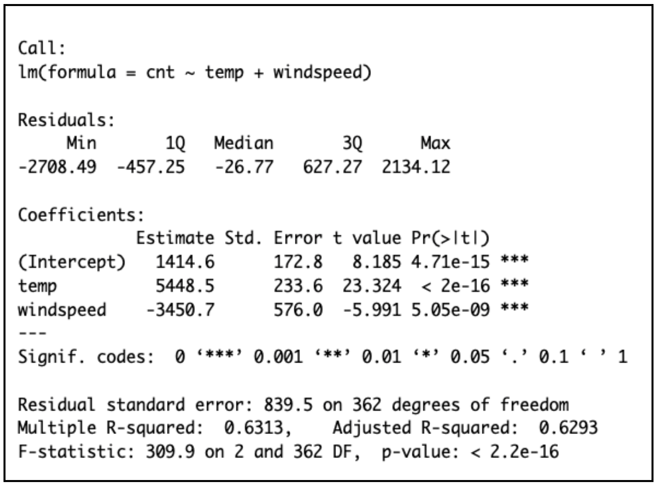
```
Call:
lm(formula = cnt ~ temp + I(temp^2) + I(temp^3) + windspeed +
    as.factor(season) + as.factor(weathersit))

Residuals:
     Min       1Q   Median       3Q      Max
-2785.95  -314.47    46.92   364.75  1633.87

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              2571.18     385.51   6.669 9.86e-11 ***
temp                   -10089.50    2844.94  -3.546 0.000443 ***
I(temp^2)               41880.17    6341.31   6.604 1.46e-10 ***
I(temp^3)              -34435.29    4402.73  -7.821 6.03e-14 ***
windspeed               -2092.59     396.70  -5.275 2.31e-07 ***
as.factor(season)2        823.68     111.64   7.378 1.15e-12 ***
as.factor(season)3        981.48     143.34   6.847 3.32e-11 ***
as.factor(season)4       1196.54     102.44  11.680  < 2e-16 ***
as.factor(weathersit)2   -542.67      63.13  -8.596 2.67e-16 ***
as.factor(weathersit)3  -2063.96     151.68 -13.607  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 552.3 on 355 degrees of freedom
Multiple R-squared:  0.8435,    Adjusted R-squared:  0.8395
F-statistic: 212.6 on 9 and 355 DF,  p-value: < 2.2e-16
```

Figure 2A: R output for model **M2**



```
Call:
lm(formula = cnt ~ temp + windspeed)

Residuals:
     Min       1Q   Median       3Q      Max
-2708.49  -457.25   -26.77   627.27  2134.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1414.6      172.8   8.185 4.71e-15 ***
temp          5448.5      233.6  23.324  < 2e-16 ***
windspeed    -3450.7      576.0  -5.991 5.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 839.5 on 362 degrees of freedom
Multiple R-squared:  0.6313,    Adjusted R-squared:  0.6293
F-statistic: 309.9 on 2 and 362 DF,  p-value: < 2.2e-16
```

Figure 3A: R output for model **M1**

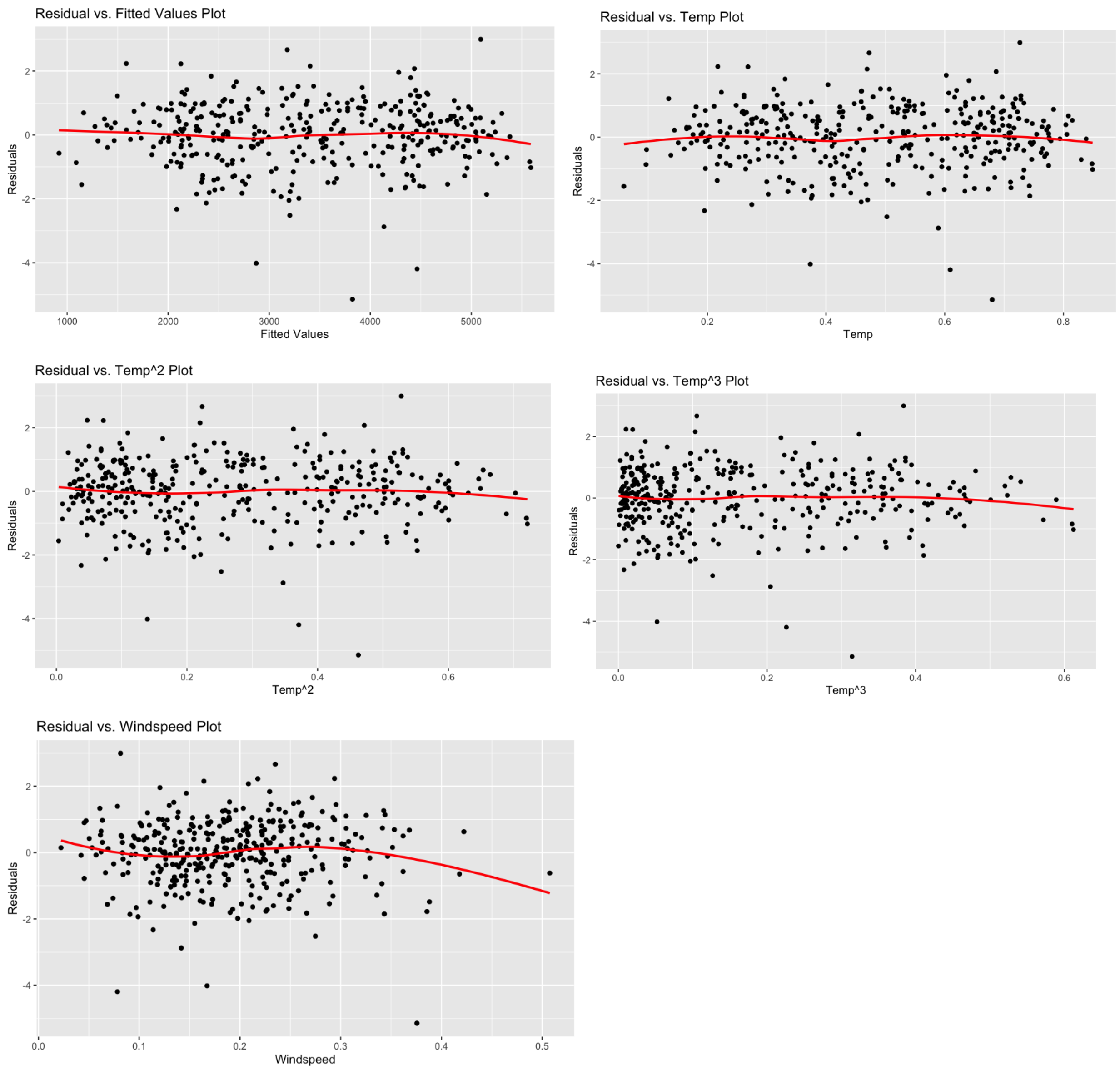


Figure 4A: plots of the standardized residuals from model **M2**

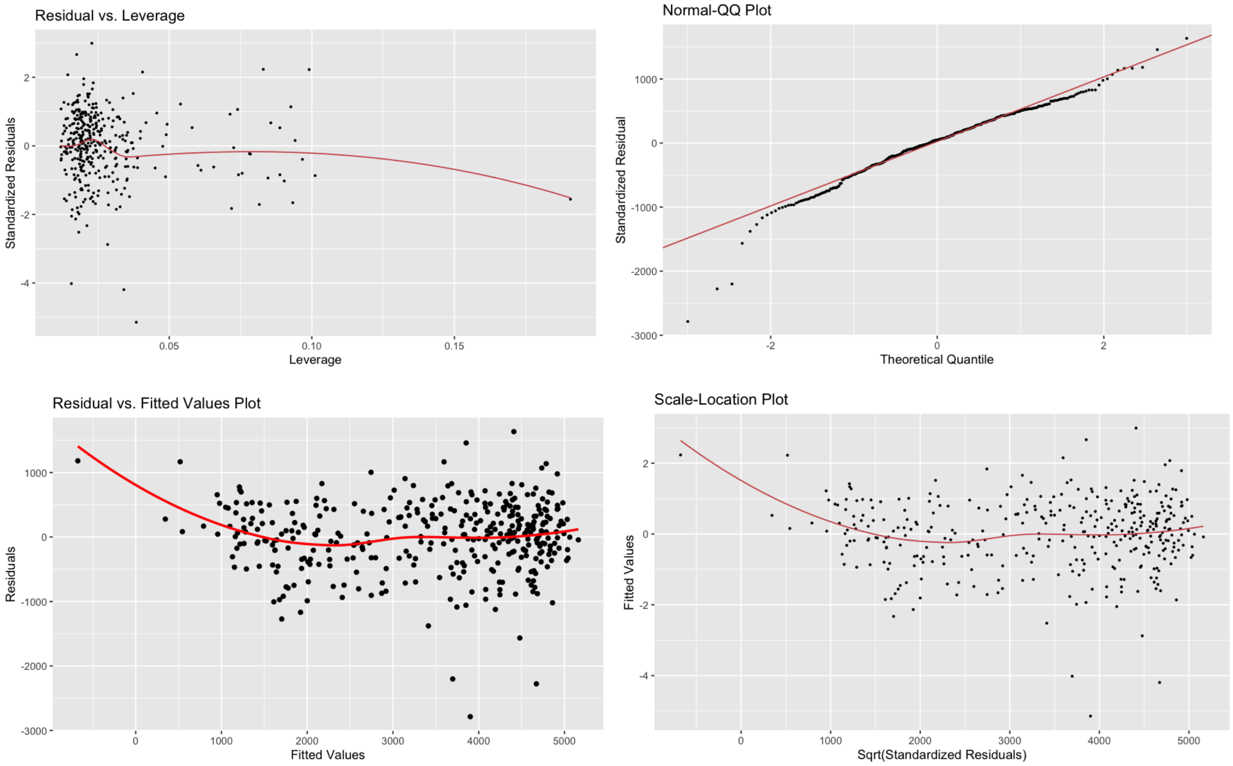
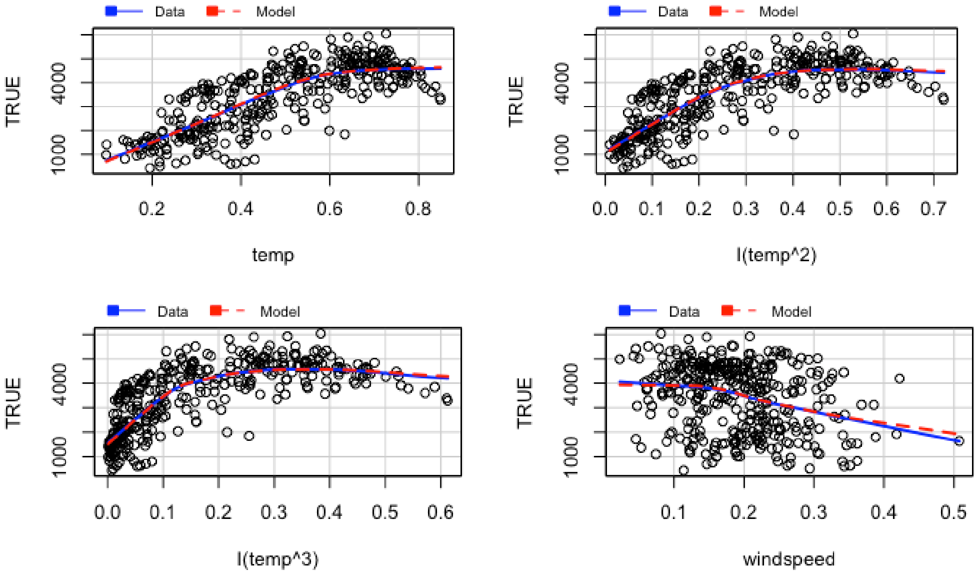


Figure 5A: diagnostics plots for model **M2**



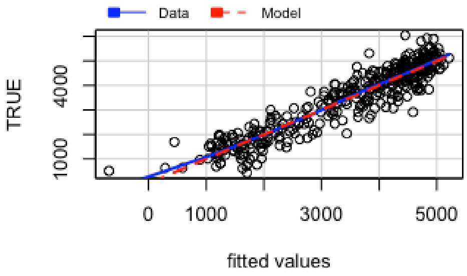


Figure 6A: marginal model plot for model **M2**

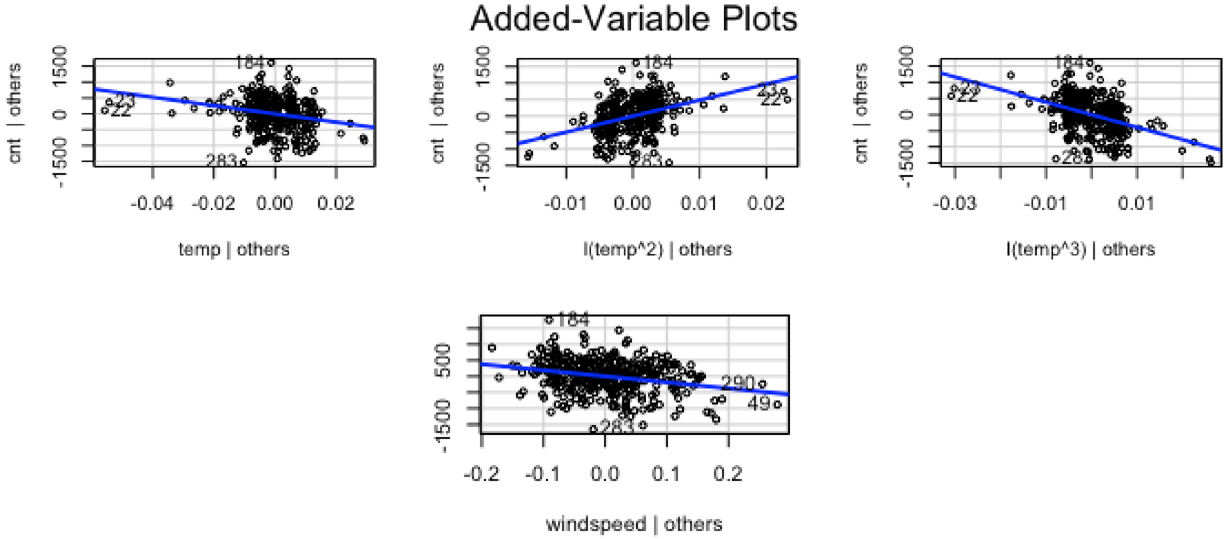
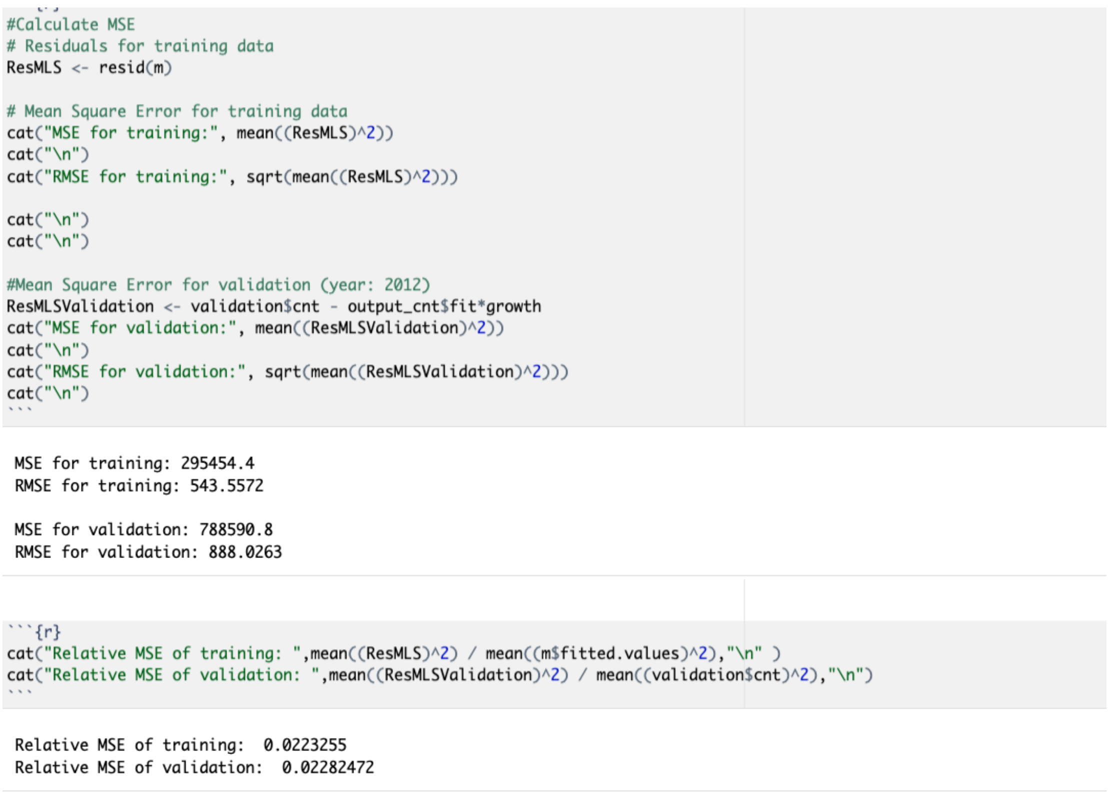


Figure 7A: added-variable plots for the model **M2**



~~~
#Calculate MSE
# Residuals for training data
ResMLS <- resid(m)

# Mean Square Error for training data
cat("MSE for training:", mean((ResMLS)^2))
cat("\n")
cat("RMSE for training:", sqrt(mean((ResMLS)^2)))

cat("\n")
cat("\n")

#Mean Square Error for validation (year: 2012)
ResMLSValidation <- validation$cnt - output_cnt$fit*growth
cat("MSE for validation:", mean((ResMLSValidation)^2))
cat("\n")
cat("RMSE for validation:", sqrt(mean((ResMLSValidation)^2)))
cat("\n")
```
~~~

```
MSE for training: 295454.4
RMSE for training: 543.5572

MSE for validation: 788590.8
RMSE for validation: 888.0263
```

~~~
```{r}
cat("Relative MSE of training: ",mean((ResMLS)^2) / mean((m$fitted.values)^2),"\n" )
cat("Relative MSE of validation: ",mean((ResMLSValidation)^2) / mean((validation$cnt)^2),"\n")
```
~~~

```
Relative MSE of training:  0.0223255
Relative MSE of validation:  0.02282472
```

Figure 8A: Validation metrics calculation

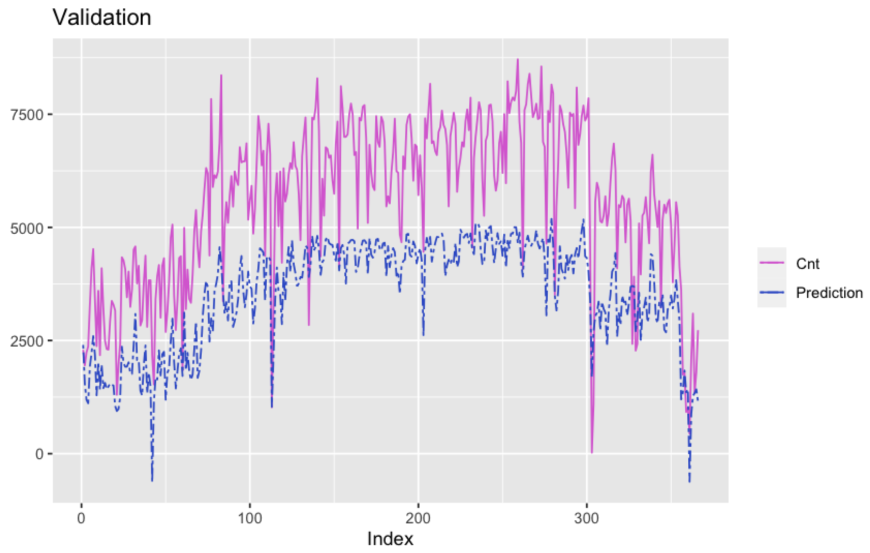
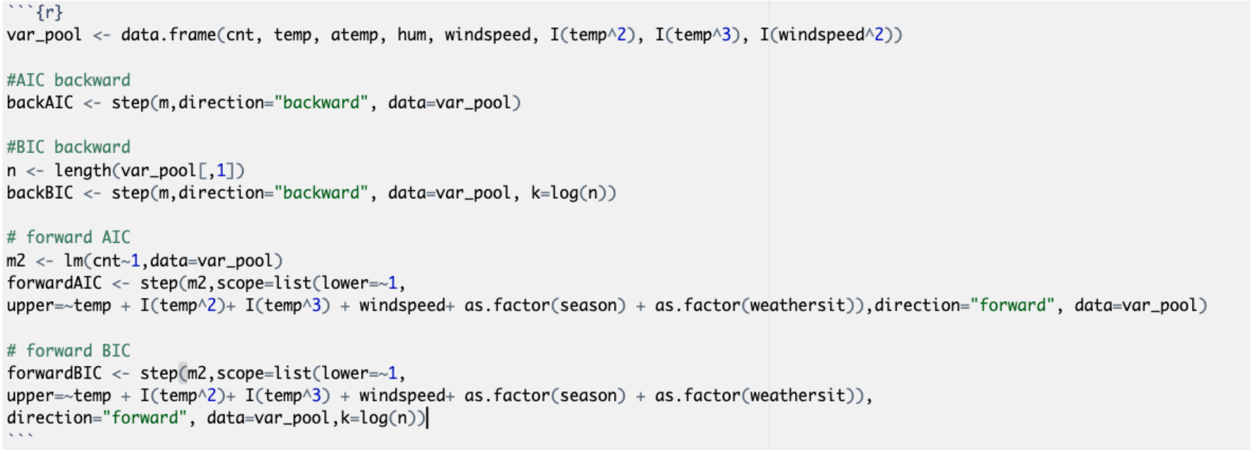


Figure 9A: Plot of predicted values against true values from 2012, before growth rate accounted for



~~~
```{r}
var_pool <- data.frame(cnt, temp, atemp, hum, windspeed, I(temp^2), I(temp^3), I(windspeed^2))

#AIC backward
backAIC <- step(m,direction="backward", data=var_pool)

#BIC backward
n <- length(var_pool[,1])
backBIC <- step(m,direction="backward", data=var_pool, k=log(n))

# forward AIC
m2 <- lm(cnt~1,data=var_pool)
forwardAIC <- step(m2,scope=list(lower=~1,
upper=~temp + I(temp^2)+ I(temp^3) + windspeed+ as.factor(season) + as.factor(weathersit)),direction="forward", data=var_pool)

# forward BIC
forwardBIC <- step(m2,scope=list(lower=~1,
upper=~temp + I(temp^2)+ I(temp^3) + windspeed+ as.factor(season) + as.factor(weathersit)),
direction="forward", data=var_pool,k=log(n))
```
~~~

Figure 10A: Variable Selection methods