Data Analysis Report

Yeuk Yu (Tiffany) Lee
December, 2016

Abstract:

This report summarizes possible factors that could affect number of registered riders in a given hour in a bike sharing system. The dataset used in this study is gathered from bike sharing stations in VA/MD area, drawn over 509 hourly records in year 2011 and 2012. A multi-linear regression model is used in modeling the relationship between number of registered rider per hour, and year, hour of day, day of week, month, holiday, workday, weather, wind speed, humidity, temperature, and feel like temperature. The result of analysis has shown that a peak hour in a workday during April through November, with lower humidity and higher number of casual user, tends to lead to a higher number of registered riders. The finding dismisses the notion that temperature and wind speed play a significant role in determining bike usage.

I. Introduction

In recent years, bike sharing systems in cities have became increasingly prevalent. Besides from being a popular mean of transportation, bike sharing provides accurate usage data such as trip duration and date, which are believed to be useful in identifying changes in mobility in cities. To identify factors that are influential to bike usage, this study aims to uncover the impact of weather and time factor on number of registered bike user per hour. In addition, we would like to explore few hypothesis related to bike usage. First, we would investigate whether an increase in casual rider would lead to a decrease in registered rider. Then we would explore whether the number of registered users and weather are dependent on holiday. In addition, we are also interested in determining whether or not "feel like" temperature is a more significant predictor in registered user than other weather variables. Lastly, we would also examine whether the effect associated with the particular time of day changes depending on whether or not a given day is a weekend.

II. Explanatory Data Analysis

The bike usage data used in this study is a set of hourly data drawn from Washington D.C./ Arlington, VA/MD area. In this dataset, there are in total 509 hourly records, collected over year 2011 and 2012. There are in total 8 variables in the dataset: number of registered user, date of record (which is further broken down into Year, Month, Day of Week, and Hour), indicator of whether or not a given day is a holiday or workday, temperature, feels like temperature, humidity, wind speed, weather, and number of casual users in the given hour. Since the date variable is accounted for by year, month, and day of week variable, date will be exempted in this analysis. In this study, our response variable is the number of registered user in a given hour. In this particular sample, the number of registered users ranges from 0 to 779, with a mean of 150.5 users per hour and standard deviation of 146.898. The distribution of number of registered user appears to be heavily right skewed, with a peak at 100 users per hour. Most of the hours have below 200 registered users per hour and no obvious outlier is observed.

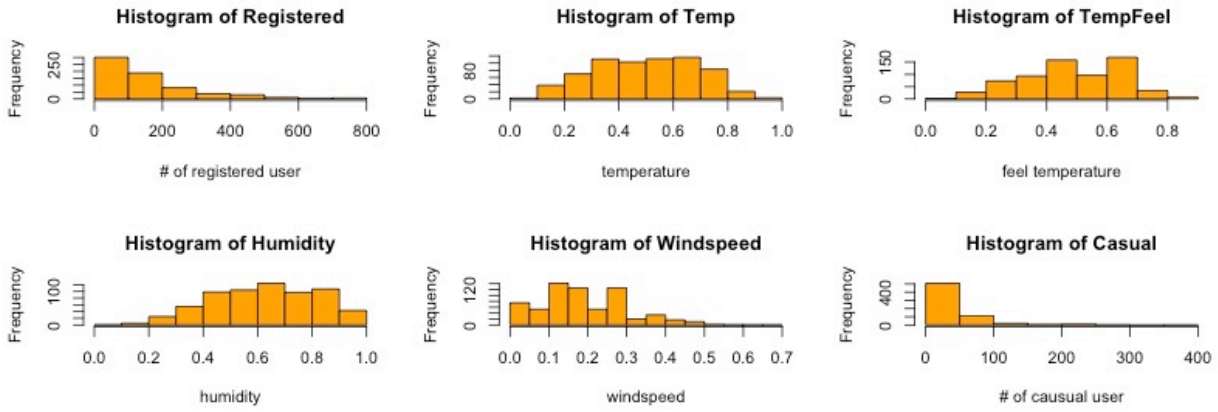

Figure 1: Histogram of number of registered riders, temperature, feel like temperature, humidity, wind speed, and number of casual riders

As for the general characteristics of the recorded hours, the number of causal user in given hour seemed to be similar to registered users. The distribution of causal user appears heavily right skewed, with a peak of below 50 users per hour. The computed average casual user is 33.53 per hour with SD (standard deviation) of 45.5.

Regarding weather factors, temperature seemed to follow a roughly symmetric, unimodal distribution with a mean of 0.501 degrees Celsius and a SD of 0.187. The ranges for both temperature and feel like temperature are from 0 to 1.0; however, the latter variable follows a slightly left skewed, bimodal distribution, with two peaks at 0.5 and 0.7 degrees Celsius. The reported mean and SD are respectively 0.484 and 0.165. The distribution of humidity also slightly tails off to the left, with only one observed peak, a mean of 0.192 and standard deviation of 0.630. Wind speed, on the other hand, appears slightly right skewed, with three modes at 0.0, 0.1, and 0.25 kilometers per hour. The distribution is asymmetric, with mean of 0.189 and SD of 0.118 km per hour. One notable feature is that weather is encoded on a scale from 1 to 4, so that 1 indicates clear to partially cloudy sky, and 4 meaning weather with heavy precipitation (i.e. snow, rain, hail, thunderstorm). In this particular dataset, we recorded 439 hours (66.6%) of clear and lightly cloudy sky, with ***no hour*** being weather 4 (heavy precipitation).

As for recorded date, most hours are recorded in Wednesday and Thursday, with about 107 (16.2%) and 104 (15.7%) hours. The least amount of hours is recorded in Monday, with only about 84 observations (12.7%). The

recorded months are evenly spread out throughout year, with each month capturing about 7% to 10% of total observations. We also observe that the recorded hours are evenly throughout the day, with lowest of 3.1% at 6p.m. and highest in 7a.m. and 10 p.m. Out of the 509 recorded hours, about 465 (70.5%) of them are workdays, and only about 14 (2.12%) are holidays. About 51.5% of the data are drawn from 2011 and the rest from 2012. Additional summary statistics can be found in table 1.

Figure 2 shows the bivariate relationships between registered users, temperature, feel like temperature, humidity, wind speed, and number of causal users. From the pairs plot, it appears that temperature, feel like temperature, humidity, and number of casual riders are significant predictors of number of registered users, with correlation coefficient of 0.3, 0.29, -0.31, and 0.47 respectively, all significant at confidence level of 99%. Among the predictors, we observe that all pairs of predictor but wind speed have moderately to strong linear relationship, all statistically significant at 95% confidence level. In particular, temperature and feel like temperature yields an r = 0.99, which indicate a very strong correlation between two, raising a flag on potential multicolinearity problem. In addition, winds speed seemed to be free of such speculation except humidity, with which the correlation is -0.26 and is significant at 99% confidence level.

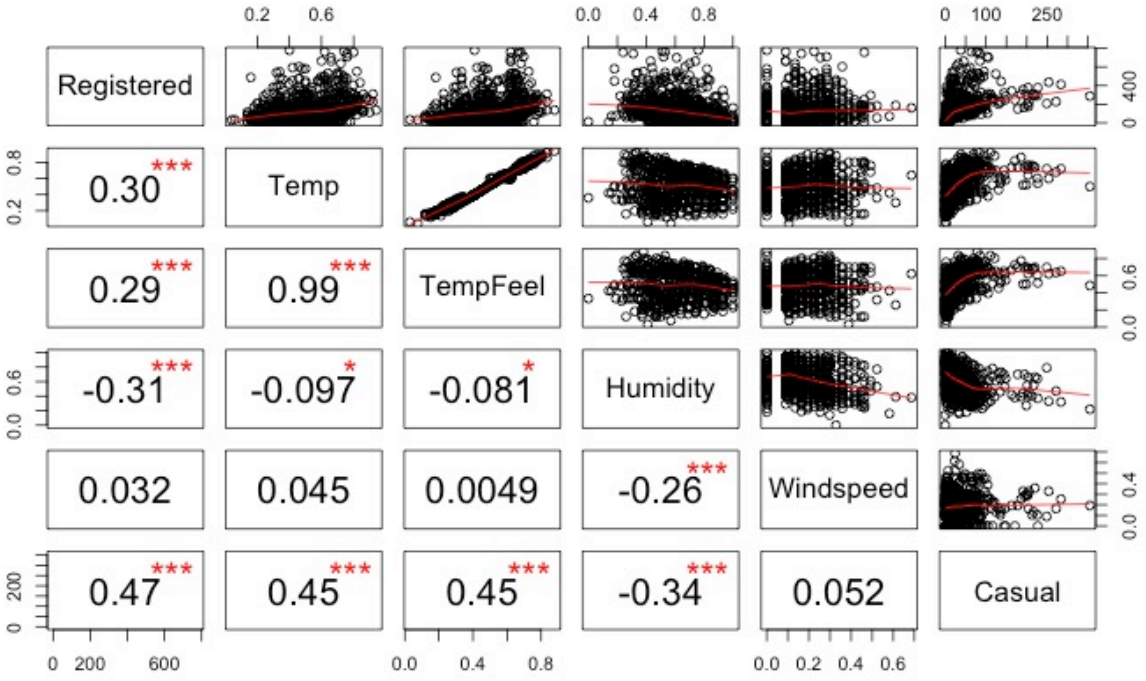

Figure 2 Pairs plot of continuous variables, with correlation between variables plotted at corresponding row and column.

Table 1: General Bike Rental Characteristics

Weather	Count	Proportion	Hour	Count	Proportion	Day of Week	Count	Proportion
1	439	0.66616085	0	30	0.04552352	0	85	0.128983308
2	162	0.245827011	1	31	0.047040971	1	84	0.127465857
3	58	0.08801214	2	29	0.04400607	2	85	0.128983308
			3	26	0.039453718	3	107	0.162367223
Holiday	**Count**	**Proportion**	4	28	0.042488619	4	104	0.157814871
0	645	0.97875569	5	21	0.031866464	5	99	0.150227618
1	14	0.02124431	6	26	0.039453718	6	95	0.144157815
			7	33	0.050075873			
Workday	**Count**	**Proportion**	8	22	0.033383915	**Month**	**Count**	**Proportion**
0	194	0.294385432	9	27	0.040971168	1	52	0.078907436
1	465	0.705614568	10	25	0.037936267	2	46	0.069802731
			11	23	0.034901366	3	62	0.094081942
Year	**Count**	**Proportion**	12	24	0.036418816	4	43	0.065250379
2011	339	0.514415781	13	27	0.040971168	5	46	0.069802731
2012	320	0.485584219	14	25	0.037936267	6	44	0.06676783
			15	38	0.057663126	7	59	0.08952959
			16	29	0.04400607	8	70	0.106221548
			17	28	0.042488619	9	60	0.091047041
			18	21	0.031866464	10	72	0.109256449
			19	25	0.037936267	11	57	0.086494689
			20	28	0.042488619	12	48	0.072837633
			21	31	0.047040971			
			22	33	0.050075873			
			23	29	0.04400607			

We then further examine the conditional distribution of number of registered user given day of week, month of year, and hour of day, workday, holiday, and weather of a given hour. From table 2, we observe that there are more riders during April to November, with about 150 to 195 users per hour on average,

while there is only about 78 to 122 riders in months December through March. There is most number of users during weekdays, with a mean in between 147 to 177 riders per hour Monday thru Friday, while the mean number of riders during weekend is only about 104 to 130 riders per hour. We also observe that the spread of weekend is smaller than weekdays, the standard deviation (SD) of riders during weekend is about 100 while it is about 150 for weekdays. As for hours of the day, the observed mean number of riders is higher in 7-9am and 7pm, with about 208 to 341 riders per hour, when other hours have about 50 to 180 riders. We observe that weather 3 has slightly lower mean and SD of about 92 riders per hour and 89, when compared to weather 1 and 2, of which the mean and SD are above 140 riders. As for holiday factor, we observe a lower mean of 91 riders per hour than an hour that is not a holiday (about 95 riders). A similar pattern is observed in workday, where an hour in workday has higher average number of registered rider (about 165) than a non-workday hour. In addition, we also observe that an hour in 2011 has lower mean (about 110 users) than an hour in 2012 (about193 riders per hour) but a smaller SD (about 105 to 170). By performing a Chi-square test on across all the categorical variables, we notice a significant correlation between workday variable and day of week, and also between workday and holiday. This result is unsurprising since whether or not a day is a workday or holiday is dependent on the fact that what day of week that is. Additional summary statistics can be found in table 1.

III. Initial modeling

Before proceeding to initial modeling, we further explore the possibility of collapsing some of the categorical variables, in particular day of week, hour, month, and weather. Since workday, holiday, and year recorded are binary variables, they are not subjected to possibility of combining into groups. The examination of whether or not these predictors are significant will be done using t test during statistical inference and modeling (section V).

The initial estimates of ***day of week*** without combining groups are included in table 2, with a reference group of Sunday. The model yields a ($r^2 = 0.02475, r^2_{adj} = 0.01578$). We can see that estimates from Tuesday through Saturday are not significant predictors in modeling number of riders per hour, adding them to the model does not increase model's predictability. The conditional boxplot of day published and summary measure table (table 2) gives evidence of fairly similar mean and standard deviation across all ***weekdays***. Collapsing all weekdays therefore seemed reasonable.

Concerning the month of year, there is a higher mean number of riders per hour between April and November (about 169 to 195 riders per hour) than the rest of the year (about 78 to 112 riders per hour). The model that includes only month with January as reference group yields an ($r^2 = 0.05235, r^2_{adj} = 0.03624$), with all month but February, March, and December as significant predictors. By manually inspecting the boxplot, we observe a fairly similar mean across months between April to November. The SD of these months are about 120 to 170 riders, which are fairly consistent. Contextual wise, it would also make sense to combine months between late spring to fall as category. We would therefore collapse these months in our model.

As for hour of day, we see that the spread of riders per hour tends to vary, but the IQR for each hour falls in between 100 to 250 riders per hours. The graph also shows that peak hours (in the case 9am-12noon, and 7pm) have a slightly higher median at about 180 riders per hour than other hours, which are about 20 riders short. Even though 12am midnight and 2am also have a high median, the mean is only about 52 and 14 riders per hour, suggesting that these hours have slightly smaller variability than peak hours. We would therefore only consider combining peak hours as one group, since it not only make sense contextual wise, but also have closer variability in categories.

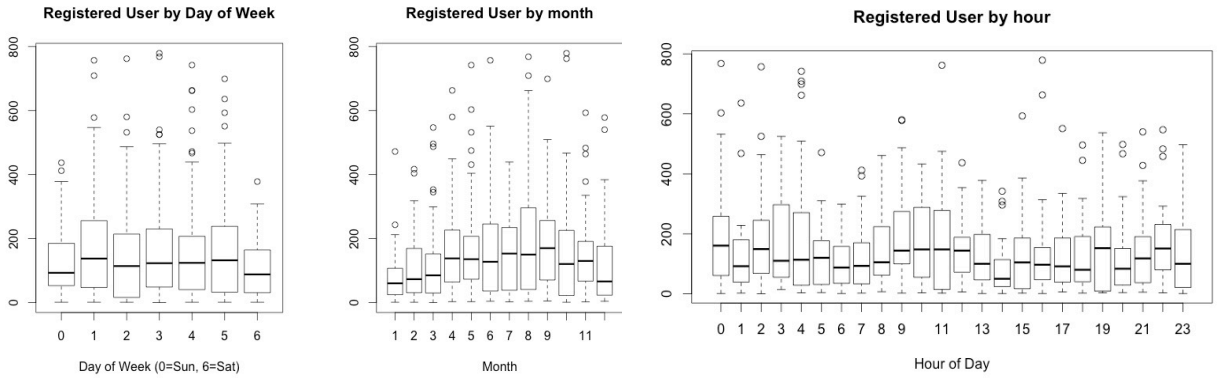

Figure 3: Graphs (from left) Distribution of registered user by day of week, month, and hour of day.

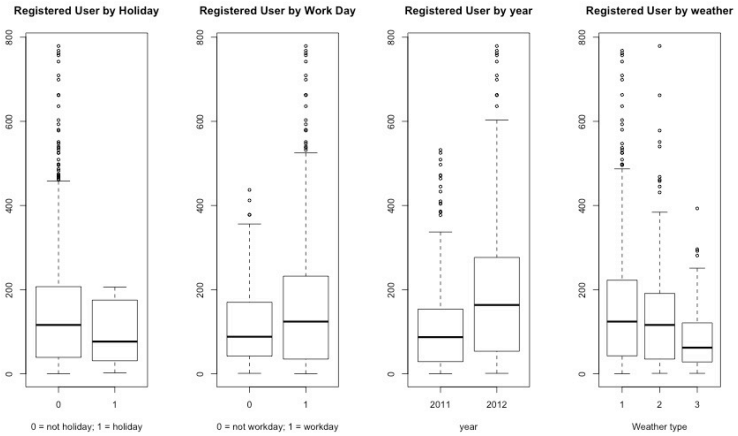

Figure 4: Distribution of registered users by holiday, work day, year, and weather.

For weather, we see that median between weather 1 and 2 are both at about 100 riders per hour; however, weather 3 gives an obviously lower median at about 80 riders per hour. From measure statistics table, we also observe that weather 3 has a smaller mean and SD than other weathers. The model with only weather predictor also confirms that weather 3 alone is a significant indicator in predicting registered user, while weather 2 is not. We will therefore proceed by encoding weather as a binary variable. All hours would be encoded as either belonged to weather 3 category or not.

We have now adjusted for potential confounding categorical variables. At this point of modeling, we have not excluded any variables in our model. We will move on to

Table 2: Comparison in Modeling

Comparison in Modeling Weather: Coef, SE, t-stats, p-value, significance											
	Estimate	Std. Error	t value	Pr(>|t|)	Significance		Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	160.736	6.959	23.098	< 2e-16	***	(Intercept)	156.34	5.95	26.277	< 2e-16	***
as.factor(bike$Weather)2	-16.297	13.404	-1.216	0.224464		weather.3	-64.27	20.05	-3.205	0.00142	**
as.factor(bike$Weather)3	-68.667	20.371	-3.371	0.000793	***	weather 1 and 2	Reference				
Comparison in Modeling Month: Coef, SE, t-stats, p-value, significance											
	Estimate	Std. Error	t value	Pr(>|t|)	Significance		Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	78.85	20	3.943	8.93E-05	***	(Intercept)	107.399	9.986	10.76	< 2e-16	***
as.factor(bike$Month)2	29.39	29.19	1.007	0.314287		apr.thru.nov	63.251	12.071	5.24	2.17E-07	***
as.factor(bike$Month)3	43.96	27.12	1.621	0.105455		Other months	Reference				
as.factor(bike$Month)4	91.01	29.72	3.062	0.002289	**						
as.factor(bike$Month)5	101.63	29.19	3.482	0.000531	***						
as.factor(bike$Month)6	94.88	29.54	3.212	0.001382	**						
as.factor(bike$Month)7	76.17	27.43	2.777	0.005643	**						
as.factor(bike$Month)8	116.3	26.4	4.405	1.24E-05	***						
as.factor(bike$Month)9	105.92	27.32	3.877	0.000117	***						
as.factor(bike$Month)10	77.22	26.24	2.943	0.00337	**						
as.factor(bike$Month)11	71.75	27.65	2.595	0.009681	**						
as.factor(bike$Month)12	38.78	28.86	1.344	0.179562							
January	Reference										
Comparison in Modeling Day of Week: Coef, SE, t-stats, p-value, significance											
	Estimate	Std. Error	t value	Pr(>|t|)	Significance		Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	130.68	15.81	8.268	7.64E-16	***	(Intercept)	163.355	6.649	24.567	< 2e-16	***
as.factor(bike$Day)1	47.16	22.42	2.104	0.0358	*	bike$Weekend	-46.383	12.723	-3.646	0.000288	***
as.factor(bike$Day)2	17.21	22.35	0.77	0.4416							
as.factor(bike$Day)3	33.61	21.17	1.587	0.1129							
as.factor(bike$Day)4	29.93	21.31	1.405	0.1605							
as.factor(bike$Day)5	35.52	21.55	1.648	0.0998	.						
as.factor(bike$Day)6	-25.98	21.76	-1.194	0.2329							
Comparison in Modeling Hour: Coef, SE, t-stats, p-value, significance											
	Estimate	Std. Error	t value	Pr(>|t|)	Significance		Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	52.63	18.54	2.838	0.004678	**	(Intercept)	141.056	6.296	22.404	< 2e-16	***
as.factor(bike$Hour)1	-25.08	26.01	-0.964	0.335224		peak.hr	51.178	14.514	3.526	0.000451	***
as.factor(bike$Hour)2	-37.98	26.45	-1.436	0.151523							
as.factor(bike$Hour)3	-43.02	27.21	-1.581	0.114436							
as.factor(bike$Hour)4	-47.95	26.69	-1.797	0.072833	.						
as.factor(bike$Hour)5	-31.59	28.9	-1.093	0.274794							
as.factor(bike$Hour)6	14.94	27.21	0.549	0.583118							
as.factor(bike$Hour)7	189.09	25.62	7.38	4.96E-13	***						
as.factor(bike$Hour)8	292.28	28.51	10.252	< 2e-16	***						
as.factor(bike$Hour)9	155.77	26.94	5.782	1.16E-08	***						
as.factor(bike$Hour)10	61.89	27.5	2.25	0.024784	*						
as.factor(bike$Hour)11	97.37	28.15	3.459	0.000578	***						
as.factor(bike$Hour)12	110.03	27.81	3.956	8.48E-05	***						
as.factor(bike$Hour)13	125.18	26.94	4.646	4.11E-06	***						
as.factor(bike$Hour)14	116.45	27.5	4.234	2.64E-05	***						
as.factor(bike$Hour)15	96.52	24.81	3.891	0.00011	***						
as.factor(bike$Hour)16	166.68	26.45	6.302	5.50E-10	***						
as.factor(bike$Hour)17	335.12	26.69	12.557	< 2e-16	***						
as.factor(bike$Hour)18	293.27	28.9	10.149	< 2e-16	***						
as.factor(bike$Hour)19	267.09	27.5	9.711	< 2e-16	***						
as.factor(bike$Hour)20	121.97	26.69	4.57	5.85E-06	***						
as.factor(bike$Hour)21	93.5	26.01	3.594	0.00035	***						
as.factor(bike$Hour)22	53.73	25.62	2.097	0.036378	*						
as.factor(bike$Hour)23	28.61	26.45	1.082	0.279827							

test confoundedness of continuous variables in t-test in model inference.

We are also interested in exploring that whether or not higher casual users could lead to higher registered user. The scatterplot of number of registered users by casual users (as included in figure 2 pairs plot) gives a slightly positive trend line, with a correlation coefficient of r = 0.47, suggesting a fairly strong linear relationship. We would further explore this relationship in section V by determining whether or not number of casual rider is a significant predictor of registered user using t-test. In addition, the examination of whether or not feels like temperature is a more important than the actual temperature will be also be done during initial modeling. We are also asked to explore whether holiday would affect the relationship between hours and number of registered user. The box plot in figure 5 do not suggest a high average number of users in holiday or vice versa. In particular, there is no record of weather 3 in a holiday. As for the affect of weekend on relationship of registered user number and hour, we observe from figure 6 a roughly consistent bike usage distribution in both weekend and weekdays, suggesting that an interaction term between weekend and user might not be significant. These two distributions both peak at 7pm, though the peak at 7am to 9am is not observed in weekend. Both interaction terms are included in the initial model.

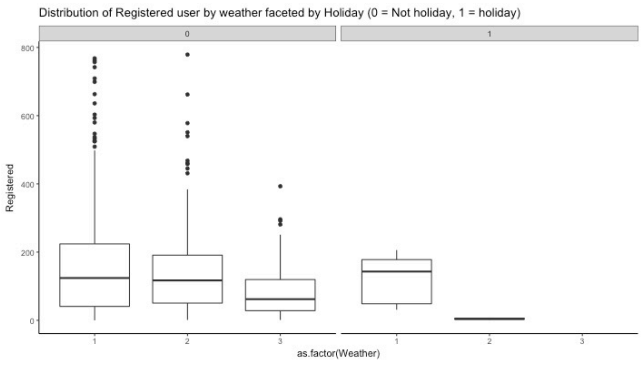

Figure 5: Distribution of users by weather, faceted by holiday

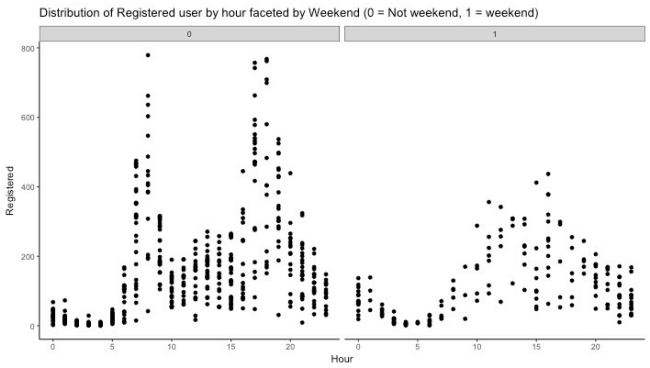

Figure 6: Distribution of users by hour, faceted by weekend

IV. Diagnostics

We now explore the possibility of removing possible outliers from the model. By using standardized residual technique, we observe that one particular hour with 288 registered user has a significantly higher residual than the bulk of the data (del.st.res = -3.143025), while the rest of the data points have a deleted standard residual between the range of -1.84 to 1.675. Using leverage on all predictor variables, we notice that the very same data point has a leverage of -3.121712, which is far from the next lowest leveraged point of -1.84. Though the critical score for outlier is (t(1-0.05/(2*659), 647) = 3.982694), this particular hour gives strong evidence as outlier and is therefore removed form the dataset. We therefore proceed to examine the normality assumption of our truncated dataset is met.

From the QQPlot in figure 6, we see that the distribution of registered is heavily right skewed and has a truncated tail at the left. The boxcox graph suggests a transformation of about lambda = 0.3. After transformation, the distribution of number of shares look approximately normal, and the boxcox transformation suggests that no further transformation is necessary, since lambda = 1 sits in the suggested range of transformation.

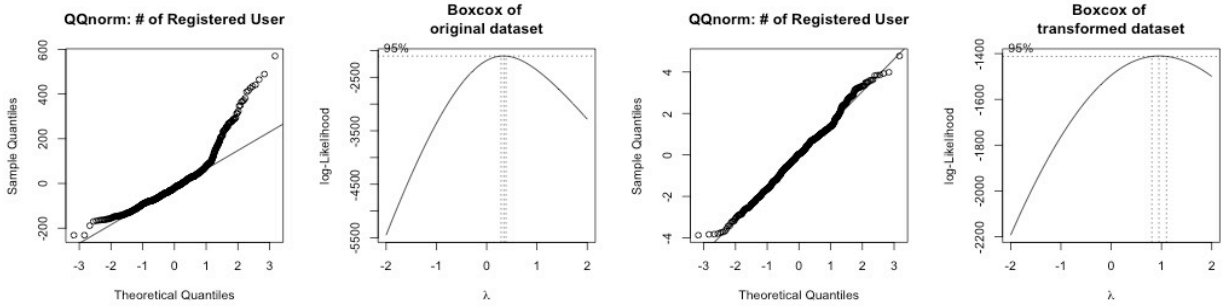

Figure 7: Linear regression diagnostic plots

We at last examine the residual plots of the variables against number of registered users for possible error assumption violation in figure 8. From the left figures, we see that the residual plot of registered users, temperature, feel like temperature, humidity, and wind speed appear to center around y = 0 line with a constant spread around 0.

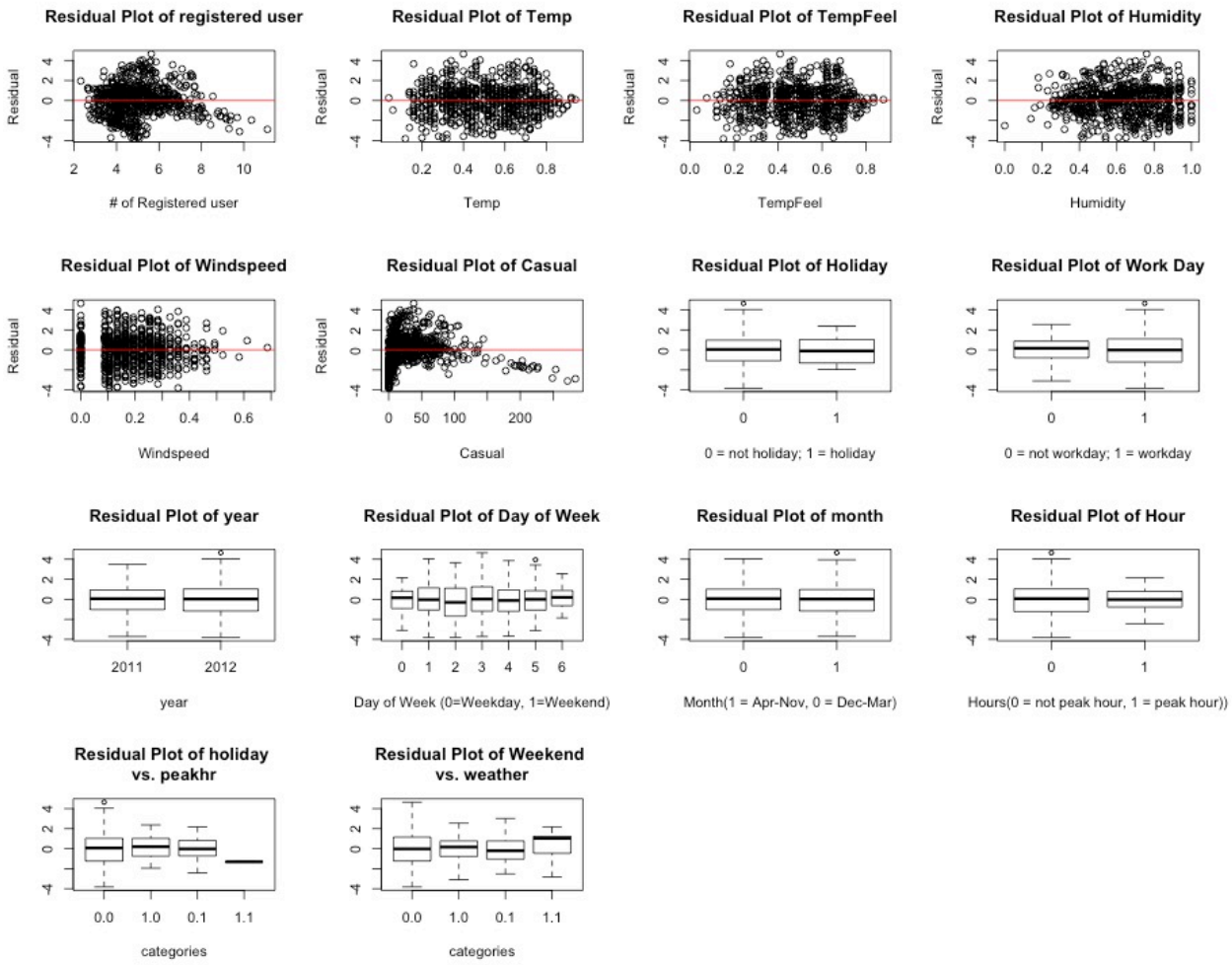

Figure 8: Residual Plots

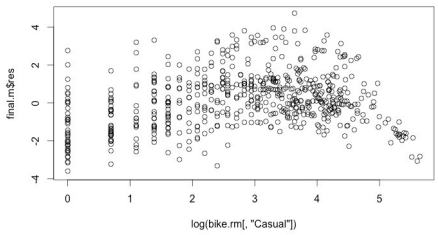

Figure 9: Residual plot of logarithmic transformed number of casual bikers

However, the residual plot of causal users appears to tightly clustered at lower predictor value, causing the constant variation assumption be violated. To remedy that, we perform log transformed on number of casual riders. Holiday, workday, year, weather, day of week, and month appear to have no violation of constant variance and residual centering at zero. The post-transformed residual plot of casual users is included in figure 9. Notice that there is still a noticeable upward curve in the residuals, but this is the best we can do to adjust for the constant variation violation.

V. Model Inference and Results

After transformation, the estimates of our final model are now displayed in table 5, alongside with our initial model. In the final model, we include year, holiday, workday, temperature, humidity, feel like temperature, wind speed, number of casual user, weather (weather 3 vs. weather 1 or 2), peak hour (7-9am, 7pm vs. other hours), months (April through November vs. other months), interaction term between peak hours and holiday, and interaction day between weather and weekend. In this model, the variable weekend is added as a variable to help define interaction term; as a result, workday is excluded since whether a given day is a workday or not can be inferred from holiday and weekend.

The following effects on number of registered users are all adjusted for the final model in transformed power of 0.35. We now try to remove one variable at a time to test the significance of the predictor in the model. The p-values of resulting models are included in table 5 as well. Notice that in this table, number of casual user is logarithmic transformed.

Table 3: Multivariate Regression Result for Predicting (# of Register User)^0.35

	Prior to Removal					After Removal (p-value)				
	Estimate	Std. Error	t value	Pr(>|t|)	Significance	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	5.065467	0.382107	13.257	<2.00E-16	***	5.077193	0.257418	19.724	<2.00E-16	***
as.factor(bike.trans$Year)	0.764036	0.125887	6.069	2.19E-09	***	0.757164	0.124657	6.074	2.12E-09	***
bike.trans$Temp	0.246031	2.878794	0.085	0.9319		2nd (0.9428)				
bike.trans$Humidity	-1.793544	0.387622	-4.627	4.49E-06	***	-1.857117	0.352175	-5.273	1.83E-07	***
bike.trans$TempFeel	-0.241274	3.251973	-0.074	0.9409		1st (0.9409)				
bike.trans$Windspeed	-0.054703	0.591336	-0.093	0.9263		3rd (0.9462)				
bike.trans$Casual	0.022912	0.001876	12.216	<2.00E-16	***	0.022977	0.001722	13.341	<2.00E-16	***
bike.trans$weather.3	-0.278085	0.267411	-1.04	0.2988		5th (0.5307)				
bike.trans$peak.hr	0.672798	0.164152	4.099	4.69E-05	***	0.632795	0.161169	3.926	9.55E-05	***
bike.trans$apr.thru.nov	0.322196	0.183822	1.753	0.0801	.	0.336617	0.142161	2.368	0.0182	*
bike.trans$Holiday	-0.837966	0.466642	-1.796	0.073	.	-1.045071	0.429753	-2.432	0.0153	*
bike.trans$Weekend	-1.032995	0.158917	-6.5	1.61E-10	***	-0.990525	0.150243	-6.593	8.94E-11	***
bike.trans$peak.hr:bike.tr	-1.599132	1.231308	-1.299	0.1945		6th (0.1909)				
bike.trans$weather.3:bike	0.52842	0.50126	1.054	0.2922		4th (0.2910)				

The final model after removal of variables yields an ($r^2 = 0.4279, r^2_{adj} = 0.4217$). In terms of research hypothesis, we conclude that number of causal users is indeed a significant predictor in predicting number of registered users. However, reading from the final model, for every extra casual rider in a given hour, we expect the transformed number of registered riders to rise by 0.0229, which refuted our initial speculation that more increase in casual riders would lead to a decrease in registered user.

The final model also suggests that year, humidity, peak hour, month, holiday, and weekend are significant predictors of number of registered riders. Compared to 2011, we expect the number of registered riders in 2012 to be an adjusted increase of 0.757. For every percentage increase of humidity, we expect its effect on the number of registered rider to be an adjusted decrease of 1.857. Compared to a normal hour, we expect the number riders in peak hour to be an adjusted increase of 0.632. Through out the year, we would expect an hour in April through November to have an adjusted increase in number of registered user to be 0.337 than an hour in December through March. We also conclude that the effect of an hour in a holiday to have an adjusted decrease of 1.046 riders than an hour in non-holiday. As for weekend, we would expect an adjusted decrease in number of registered riders in a weekend than a non-weekend.

With regard to our initial research hypothesis, we observe that the interaction terms for both between peak hour and weekend and between holiday and weather are dropped in the model as well because of insignificant predictive power. Therefore, the interaction affect on registered users between hour and weekend and also holiday and weather is dismissable. On the other hand, we observe that feel like temperature is excluded from the model due to insignificant predictive power. Table 3 further suggests that humidity is a more significant predictor than feel like temperature. However, given that the correlation between all the weather factors are significant (section II), the high dependency would have made it difficult to determine which is a more significant predictor. Therefore, the hypothesis that feel like temperature is a more significant predictor than temperature can be due to variation of dataset or inadequacy in study methodology.

VI. Conclusion/ Discussion

In the final model, we conclude that year, humidity, casual users, hour of day, month of year, and work day are significant factors in predicting number of registered users in a given hour. Surprisingly, temperature does not affect number of riders as much as humidity does. Another finding in this study is that there is a larger number of registered riders during weekdays than weekend, since registered users are mostly using bikes to travel to and forth work. In addition, a increase in casual bikers is correlated to an increase in registered biker, which undermines our initial speculation that more casual bikers is associated with less registered bikers. This result can be due to the fact that rise in number of casual bikers are related to the aggregate trend of bike usage. However, the real cause will be open up to future studies on related topic.

Appendix: R code

```
library(data.table)
library(dplyr)
library(MASS)
library(ggplot2)
library(MASS)

# Hypothesis
# 1. given that every rider is either a casual user or registered user,
# an increase in casual users is associated with a decrease in registered users.
# 2. It is also hypothesized that the relationship between the number of registered users and
# the weather is dependent on whether or not it is a holiday.
# 3. Someone suggests that among general weather conditions, the "feels like" temperature
# is more important that the actual temperature, windspeed, and humidity.
# 4. It is also believed that the effect associated with the particular time of day changes
# depending on whether or not it is a weekend.


################################################################################
############################## Reading in data #################################
################################################################################

# Reading in data
bike <- fread("Desktop/Academic/f16/36401/DA/Final\ DA/final-59.txt")

# Adding weekend columnn
bike$Weekend <- ifelse(bike$Holiday == 0 & bike$WorkDay == 0, 1, 0)

################################################################################
############################## Pairs Plot code #################################
################################################################################

# Professor Nugent's code
panel.smooth<- function (x, y, col = par("col"), bg = NA, pch = par("pch"),
                         cex = 1, col.smooth = "red", span = 2/3, iter = 3)
{
   points(x, y, pch = pch, col = col, bg = bg, cex = cex)
   ok <- is.finite(x) & is.finite(y)
   if (any(ok))
      lines(stats::lowess(x[ok], y[ok], f = span, iter = iter),
            col = col.smooth)
}

panel.hist<-function(x)
{
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(usr[1:2], 0, 1.5) )
   h <- hist(x, plot = FALSE)
   breaks <- h$breaks; nB <- length(breaks)
   y <- h$counts; y <- y/max(y)
   rect(breaks[-nB], 0, breaks[-1], y, col="cyan")
}

panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(0, 1, 0, 1))
   r <- cor(x, y)
   txt <- format(c(r, 0.123456789), digits=digits)[1]
   txt <- paste(prefix, txt, sep="")
   if(missing(cex.cor)) cex<-2  #cex <- 0.8/strwidth(txt)

   test <- cor.test(x,y)
   # borrowed from printCoefmat
   Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                    cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                    symbols = c("***", "**", "*", ".", " "))
```

```
   #text(0.5, 0.5, txt, cex = cex * abs(r))
   text(0.5,0.5,txt,cex=cex)
   text(.8, .8, Signif, cex=cex, col=2)
}
#############################################################################
############################## EDA ##########################################
#############################################################################

# histogram for all the cont variables
par(mfrow = c(2, 3))
bike.cont <- bike[c(1, 10, 11, 12, 13, 14)]
var.names <- names(bike.cont)
x.axis <- c("# of registered user", "temperature", "feel temperature", "humidity", "windspeed",
            "# of causual user")
for (var in 1:ncol(bike.cont)){
   hist(as.numeric(unlist(bike.cont[var])), main = paste("Histogram of", var.names[var]),
        xlab = x.axis[var], col = "orange")
}

# Pair plot of multicolinearity
pairs(bike.cont, lower.panel = panel.cor, upper.panel=panel.smooth)

# Table for categorical variable
table(bike$Weather)
table(bike$Month)
table(bike$Year)
table(bike$Hour)
table(bike$Holiday)
table(bike$WorkDay)
table(bike$Day)

# Compare between categories
bike.cat <- bike[-c(1, 10, 11, 12, 13, 14)]
par(mfrow = c(1, 4))
boxplot(bike$Registered ~ as.factor(bike$Holiday), main = "Registered User by Holiday",
        xlab = "0 = not holiday; 1 = holiday")
boxplot(bike$Registered ~ as.factor(bike$WorkDay), main = "Registered User by Work Day",
        xlab = "0 = not workday; 1 = workday")
boxplot(bike$Registered ~ as.factor(bike$Year), main = "Registered User by year", xlab = "year")
boxplot(bike$Registered ~ as.factor(bike$Weather), main = "Registered User by weather",
        xlab = "Weather type")
boxplot(bike$Registered ~ as.factor(sort(bike$Hour)), main = "Registered User by hour")

par(mfrow = c(1, 2))
boxplot(bike$Registered ~ as.factor(bike$Day), main = "Registered User by Day of Week",
       xlab = "Day of Week (0=Sun, 6=Sat)")
boxplot(bike$Registered ~ as.factor(bike$Month), main = "Registered User by month",
        xlab = "Month")

# Chisq Tests
chisq.test(bike$Year, bike$Month)
chisq.test(bike$Year, bike$Holiday)
chisq.test(bike$Year, bike$WorkDay)
chisq.test(bike$Year, bike$Day)
chisq.test(bike$Year, bike$Day)
chisq.test(bike$Year, bike$Day)

chisq.test(bike$Day, bike$WorkDay)
chisq.test(bike$WorkDay, bike$Holiday)
#############################################################################################
############################## Collapsing Categories ########################################
#############################################################################################

# Try using regression in all categories
summary(lm(bike$Registered ~ as.factor(bike$Weather)))
summary(lm(bike$Registered ~ as.factor(bike$Year)))
summary(lm(bike$Registered ~ as.factor(bike$Hour)))
summary(lm(bike$Registered ~ as.factor(bike$Holiday)))
summary(lm(bike$Registered ~ as.factor(bike$WorkDay)))
```

```
summary(lm(bike$Registered ~ as.factor(bike$Day)))
summary(lm(bike$Registered ~ as.factor(bike$Month)))


# Finding mean and sd of every category vs. Regstered
weather.df <- split(bike, bike$Weather)
weather.m <- sapply(weather.df, function(df){mean(df$Registered)})
weather.sd <- sapply(weather.df, function(df){sd(df$Registered)})

day.df <- split(bike, bike$Day)
day.m <- sapply(day.df, function(df){mean(df$Registered)})
day.sd <- sapply(day.df, function(df){sd(df$Registered)})

hour.df <- split(bike, bike$Hour)
hour.m <- sapply(hour.df, function(df){mean(df$Registered)})
hour.sd <- sapply(hour.df, function(df){sd(df$Registered)})

holiday.df <- split(bike, bike$Holiday)
holiday.m <- sapply(holiday.df, function(df){mean(df$Registered)})
holiday.sd <- sapply(holiday.df, function(df){sd(df$Registered)})

workday.df <- split(bike, bike$WorkDay)
workday.m <- sapply(workday.df, function(df){mean(df$Registered)})
workday.sd <- sapply(workday.df, function(df){sd(df$Registered)})

month.df <- split(bike, bike$Month)
month.m <- sapply(month.df, function(df){mean(df$Registered)})
month.sd <- sapply(month.df, function(df){sd(df$Registered)})

year.df <- split(bike, bike$Year)
year.m <- sapply(year.df, function(df){mean(df$Registered)})
year.sd <- sapply(year.df, function(df){sd(df$Registered)})


##############################################################################################
############################### Recoding Categories ##########################################
##############################################################################################

# Collapsing day -> weekends
summary(lm(bike$Registered ~ bike$Weekend))

# weather -> weather 3
weather.3 <- ifelse(bike$Weather == 3, 1, 0)
bike$weather.3 <- weather.3
summary(lm(bike$Registered ~ weather.3))

# Month -> april thru november
apr.thru.nov <- ifelse(bike$Month >= 4 & bike$Month <= 11, 1, 0)
bike$apr.thru.nov <- apr.thru.nov
summary(lm(bike$Registered ~ apr.thru.nov))

# Hour -> peak hour
peak.hrs <- c(9,10,11,12,19)
peak.hr <- ifelse(bike$Hour %in% peak.hrs, 1, 0)
bike$peak.hr <- peak.hr
summary(lm(bike$Registered ~ peak.hr))

##############################################################################################
############################### Explore Hypothesis ###########################################
##############################################################################################

# 1. Relationship between casual and registered bikers
plot(bike$Casual, bike$Registered, main = "Scatterplot of Registered by Casual",
     xlab = "# of Casual User", ylab = "# of Registed User")

# 2. It is also hypothesized that the relationship between the number of registered users and
# the weather is dependent on whether or not it is a holiday.
ggplot(data = bike) + geom_boxplot(aes(y = Registered, group = Weather)) +
                      facet_grid(Holiday ~ .)
```

```
# 4. It is also believed that the effect associated with the particular time of day changes
# depending on whether or not it is a weekend.
ggplot(data = bike) + geom_point(aes(group = Hour, y = Registered)) +
                      facet_grid(Weekend ~ .)

################################################################################
############################## Identify Outliers ###############################
################################################################################

init.m <- lm(bike$Registered ~ as.factor(bike$Year) + as.factor(bike$Holiday) +
            as.factor(bike$WorkDay) + bike$Temp +
               bike$Humidity + bike$TempFeel + bike$Windspeed +
               bike$Casual + weather.3 + peak.hr)
summary(init.m)

is2011 <- ifelse(bike$Year == 2011, 1, 0)
X <- cbind(1, is2011, as.factor(bike$Holiday), as.factor(bike$WorkDay),bike$Temp, bike$Humidity,
bike$TempFeel,
           bike$Windspeed, bike$Casual, weather.3, peak.hr)
H <- X %*% solve(t(X) %*% X) %*% t(X)

n <- nrow(X)
p <- ncol(X)
SSE <- sum(init.m$res^2)
res <- init.m$res
del.res <- res*sqrt((n-p-1)/(SSE*(1-diag(H))-res^2))
sort(del.res)[1:10]
sort(del.res[(n-10):n])
alpha <- 0.05
qt(1-alpha/(2*n), n-p-1)

MSE <- sum(init.m$res^2) / (n-p)
res <- init.m$res
r <- res/sqrt(MSE*(1-diag(H)))
sort(r)[1:10]; sort(r)[(n-9):n]

# identify outlier 356 in both x and y
bike[356, ]
mean(bike$Registered)
mean(bike$Casual)
out.inds <- 356  # residual -3.14 while the rest are above between -1.846 and 1.6757

# remove outliers
bike.rm <- bike[-356,]

################################################################################
############################## Diagnostics #####################################
################################################################################

rm.m <- lm(bike.rm$Registered ~ as.factor(bike.rm$Year) + bike.rm$peak.hr * bike.rm$WorkDay +
             as.factor(bike.rm$Year) + as.factor(bike.rm$Holiday) +
               as.factor(bike.rm$WorkDay) + bike.rm$Temp +
               bike.rm$Humidity + bike.rm$TempFeel + bike.rm$Windspeed + bike.rm$apr.thru.nov +
               bike.rm$Casual + bike.rm$weather.3 + bike.rm$peak.hr + bike.rm$peak.hrs * bike.rm$Weekday)
par(mfrow = c(1,4))
qqnorm(rm.m$res, main = "QQnorm: # of Registered User")
qqline(rm.m$res)

####### Response variable transformation

# shifting registered user data to apply boxcox
par(mfrow = x(1, 4))
Registered.shift <- bike.rm$Registered - min(bike.rm$Registered) + 1
rm.m <- lm(Registered.shift ~ as.factor(bike.rm$Year) + as.factor(bike.rm$Holiday) +
             as.factor(bike.rm$WorkDay) + bike.rm$Temp +
             bike.rm$Humidity + bike.rm$TempFeel + bike.rm$Windspeed + bike.rm$apr.thru.nov +
             bike.rm$Casual + bike.rm$weather.3 + bike.rm$peak.hr)
boxcox(rm.m)
```

```
title("Boxcox of \noriginal dataset")

# transform data as suggested in boxcox graph
final.m <- lm((bike.rm$Registered)^0.35 ~ as.factor(bike.rm$Year) + as.factor(bike.rm$Holiday) +
                as.factor(bike.rm$WorkDay) + bike.rm$Temp +
                bike.rm$Humidity + bike.rm$TempFeel + bike.rm$Windspeed + bike.rm$apr.thru.nov +
                bike.rm$Casual + bike.rm$weather.3 + bike.rm$peak.hr)
qqnorm(final.m$res, main = "QQnorm: # of Registered User")
qqline(final.m$res)
final.m <- lm(Registered.shift^0.35 ~ as.factor(bike.rm$Year) + as.factor(bike.rm$Holiday) +
              as.factor(bike.rm$WorkDay) + bike.rm$Temp +
              bike.rm$Humidity + bike.rm$TempFeel + bike.rm$Windspeed + bike.rm$apr.thru.nov +
              bike.rm$Casual + bike.rm$weather.3 + bike.rm$peak.hr)
boxcox(final.m)
title("Boxcox of \ntransformed dataset")

### Predictor transformation
bike.trans <- bike.rm
bike.trans$Registered <- bike.rm$Registered^0.35


# Boxplots
par(mfrow = c(4, 3))
plot(final.m$fit, final.m$res, ylab = "Residual", xlab = "# of Registered user")
abline(a = 0, b = 0, col = "red")
cont.inds <- c(10, 11, 12, 13, 14)
var.name <- colnames(bike[, cont.inds])
inds <- 1
for (i in cont.inds){
   plot(bike.trans[, i], final.m$res, ylab = "Residual", xlab = var.name[inds])
   abline(a = 0, b = 0, col = "red", main = paste("Residual Plot of", var.name[inds]))
   inds = inds + 1
}

attach(bike.trans)
boxplot(final.m$res ~ as.factor(Holiday), main = "Residual Plot of Holiday",
        xlab = "0 = not holiday; 1 = holiday")
boxplot(final.m$res ~ as.factor(WorkDay), main = "Residual Plot of Work Day",
        xlab = "0 = not workday; 1 = workday")
boxplot(final.m$res ~ as.factor(Year), main = "Residual Plot of year", xlab = "year")
boxplot(final.m$res ~ as.factor(weather.3), main = "Residual Plot of weather",
        xlab = "Weather (1 = weather 3, 0 = weather 1/2)")
boxplot(final.m$res ~ as.factor(peak.hrs), main = "Residual Plot of hour")
boxplot(final.m$res ~ as.factor(Day), main = "Residual Plot of Day of Week",
        xlab = "Day of Week (0=Weekday, 1=Weekend)")
boxplot(final.m$res ~ apr.thru.nov, main = "Residual Plot of month",
        xlab = "Month(1 = Apr-Nov, 0 = Dec-Mar)")
boxplot(final.m$res ~ WorkDay * bike.trans$peak.hr, main = "Residual Plot of month",
        xlab = "Month(1 = Apr-Nov, 0 = Dec-Mar)")


plot(bike.rm[, "Casual"], final.m$res)
plot(log(bike.rm[, "Casual"]), final.m$res)
bike.trans$Casual <- log(bike.rm$Casual)
for (i in 1:nrow(bike.trans)){
   if (is.infinite(bike.trans$Casual[i])){
      bike.trans$Casual[i] = 0
   }
}

#############################################################################
############################## Final Modeling ###############################
#############################################################################

# Stepw-wise removing variable
final.m <- lm(bike.trans$Registered ~ as.factor(bike.trans$Year) + as.factor(bike.trans$Holiday)+
                bike.trans$Temp +
                bike.trans$Humidity + bike.trans$TempFeel + bike.trans$Windspeed +
                bike.trans$Casual + bike.trans$weather.3 + bike.trans$peak.hr + bike.trans$apr.thru.nov +
```

```
                bike.trans$WorkDay * bike.trans$peak.hr)

final.m <- lm(bike.trans$Registered ~ as.factor(bike.trans$Year) +
                bike.trans$Humidity + as.factor(bike.trans$WorkDay) +
                bike.trans$Casual + bike.trans$peak.hr + bike.trans$apr.thru.nov)

final.m <- lm(bike.trans$Registered ~ as.factor(bike.trans$Year) +
                as.factor(bike.trans$WorkDay) +
                bike.trans$TempFeel +
                bike.trans$Casual)

summary(final.m)
```