MQM Stats Final Project: Predicting Bike Rentals Demand

Section B, Team #30
Avani Bhargava, Chu-Ru (Ruth) Cheng, Roshni Shahani, Xinyan (Sarah) Wang, Yuchen (Vinnie) Zhang

Introduction

The dataset analyzed is called "Bike Sharing Dataset Data Set"[1], adopted from UCI Machine Learning Repository. Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, users are able to easily rent a bike from a particular position and return back at another position. The dataset we obtained contains hourly and daily records of rental bikes between the years 2011 and 2012 from Capital Bikeshare system, Washington D.C. with the corresponding weather and seasonal information.

Business Understanding

We attempted to use the dataset to predict the bike rental demands under different environmental conditions and seasonal settings. We expected that the bike renting company needs to decide how much bikes to put on reserve, so it will neither impose too much cost because they're keeping too many bikes, nor earn less because there are not enough bikes. We tried to identify variables that can potentially have a correlation to the number of bikes being rented on a certain day. Possible attributes include weather, season, whether a certain day is a holiday and etc. In essence, we are helping the bike rental company to predict the demand of bike rentals and decide how many bikes to be reserved on a certain day for customer usage to help them maximize their profit.

Data Understanding

The bike sharing dataset consists of 731 observations and has 17 variables. Among the 17 variables, the first two are column index and respective date, which we excluded from the modeling process. The variable "cnt", which stands for the total count of rented bike, is the sum of two other variables, which account for two customer groups, registered users and casual users. The analysis of the difference between these two customer groups is not included in this report, so we excluded variables "casual" and "registered". We also excluded hourly data, because it is more reasonable to predict daily demand, as changing bike reserves on an hourly basis is not feasible. There are 7 categorical and 4 numeric variables that categorize seasonal and environmental conditions on a daily basis. [For full definition of variables please refer to appendix 1.]

---

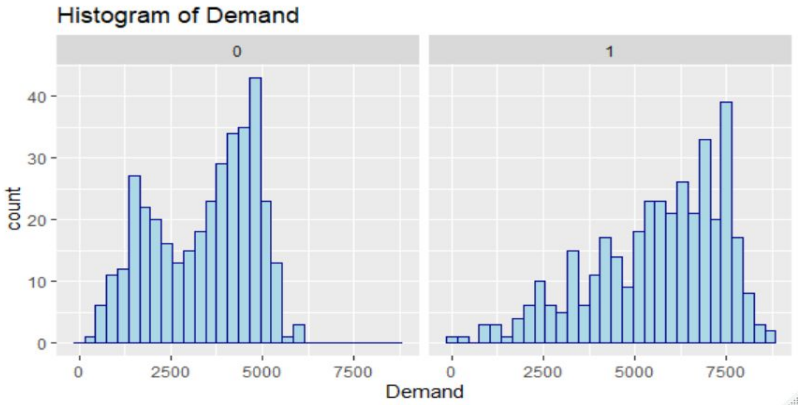
[1] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg

Data Cleaning

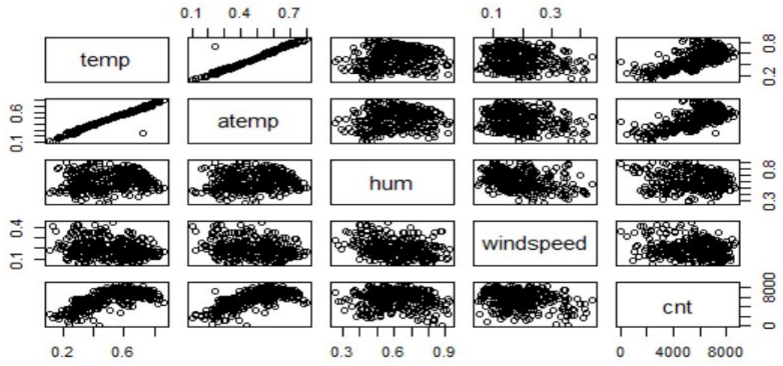
We explored the dataset for potential issues. We first used sum(is.na()) to identify missing values and found none. All categorical variables were defined numerically so we transformed them into factors for recognition in R. The independent numeric variables have already been normalized to values between 0 and 1, but count of bike rental has not. Further exploration identifies one observation with 0 humidity, which we excluded from the modeling process.

Data Exploration

We first analyzed the effect of year on rental bike demand. From the Histogram of Demand, we can safely conclude that there is a significant increase in the demand of bikes from year 0 to year 1. However, further research reveals that such increase in demand is due to the fact that year 0 is the year of inception of the company. Since we didn't expect to see such growth in the years to come, we excluded year 0 from our dataset.

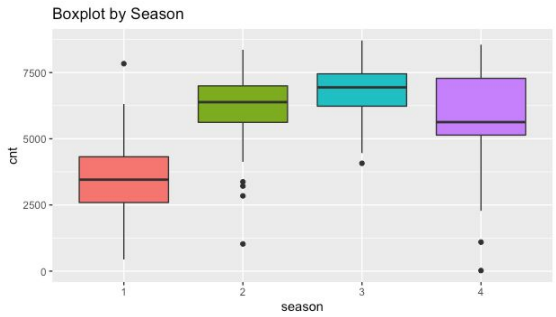


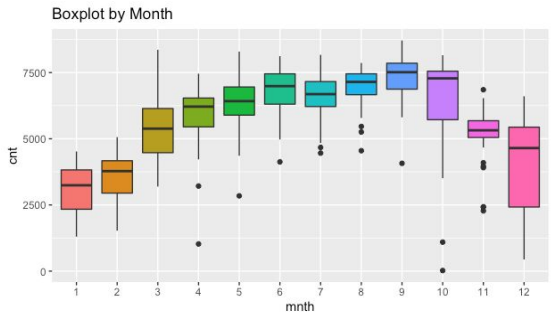
Plotting a correlation matrix on numeric variables helped us in detecting collinearity among variables. From the analysis, the variable "atemp" showed a strong correlation with variable "temp" and so, we removed "atemp" from all of our further analysis.

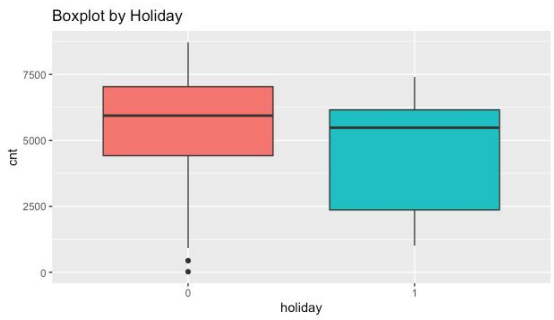


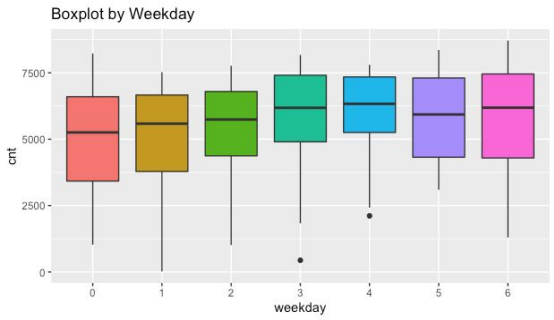
We then plotted boxplots of all categorical variables against the dependent variable, and histograms for all the numeric variables. The most prominent factor that can be observed from

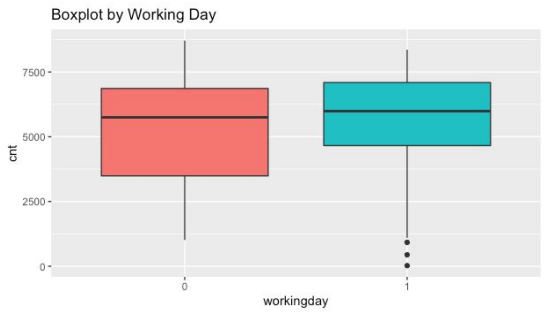
the boxplots below is that the count of the bikes rented daily is the most influenced by season, month and weather situation (weathersit). The other variables like holiday and working day have a moderate influence on the bikes rented daily.

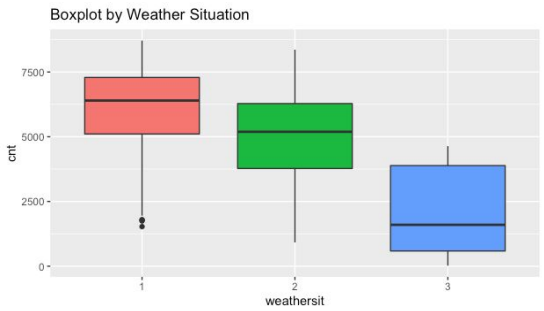


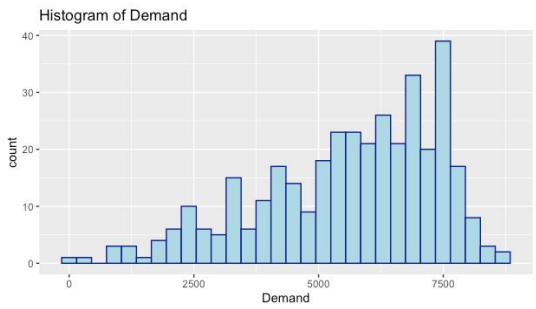


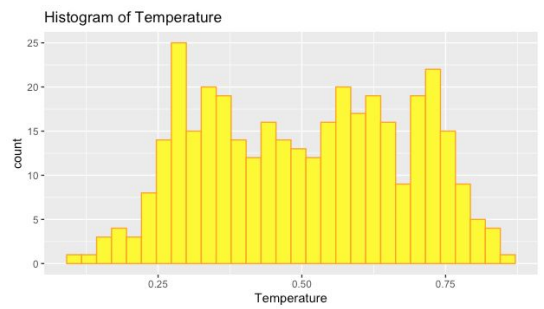


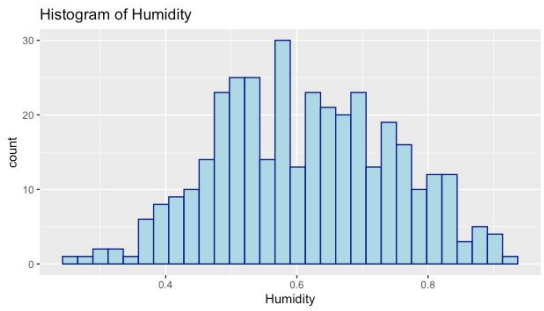


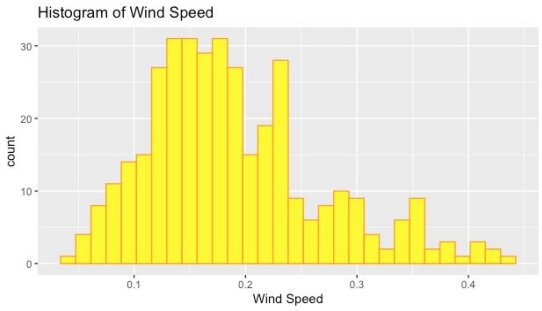








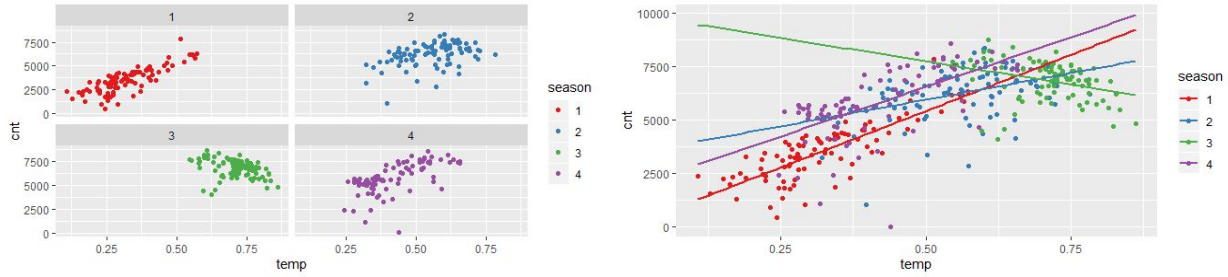




Interaction Exploration

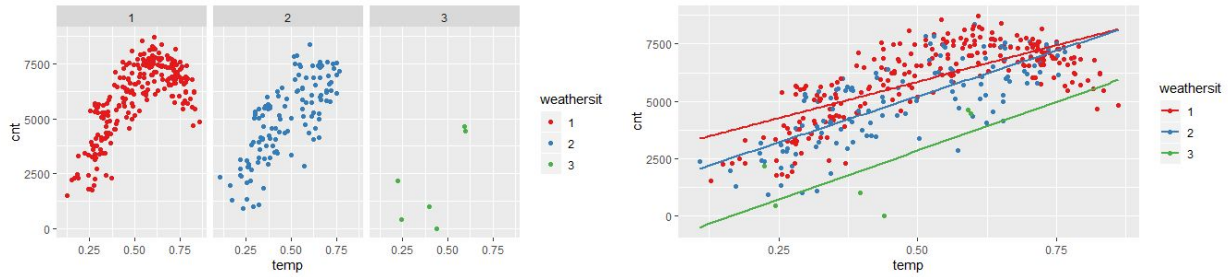
Here we conducted explorative analysis on variable interactions. We plotted the relationship between temperature (temp) and the bike rental count (cnt), and grouped the data by seasons, weather, and whether or not the day is a holiday.

Effects of Season on Relationship between Temperature and Rental Count



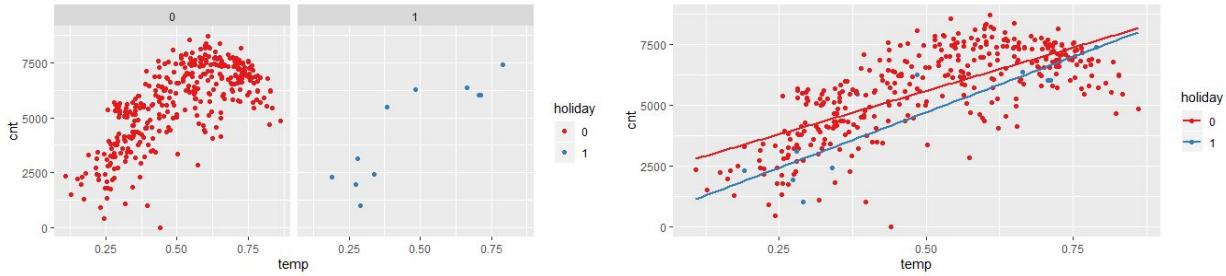
The relationship between temperature and rental counts can be observed from the plot . In spring (season=1) and in winter (season=4), the bike rental count increases the fastest with an increase in temperature; however, on average, at the same temperature, people rent more bikes in winter (season=4) than in spring (season=1). In summer (season=2), the slope of the linear model is flatter, indicating that, on average, the change in temperature has slightly less effect on the number of bikes rented. In fall (season=3), the relationship between temperature and the number of bikes rented is different from other seasons: on average, the hotter it gets, the lesser bikes were rented.

Effects of Weather on Relationship between Temperature and Rental Count



From the plot, we can see that the relationship between the temperature and the number of bikes rented are similar between different weather types. The most different aspect is that, on average, days with good weather (weathersit=1) has the highest bike rent count, days with misty weather (weathersit=2) has the second highest bike rent count, and days with light rain/snow (weathersit=3) has the lowest bike rent count. Also, the bike rental counts appears to decrease when the temperature is too high on the days with good weather, implying that there is an optimal temperature for biking.

Effects of Holiday on Relationship between Temperature and Rental Count



The plot shows that, on average, the rental counts on the same temperatures is higher on a non-holiday (holiday=0) than on a holiday (holiday=1). Similar to the days with good weather, there is also a dip in bike rental counts above a certain temperature on non-holidays; however, for holidays, the trend is not as obvious.

Base Model and Log Transformation

$$\widehat{Bike\ Count} = \beta_0 + \beta_1 * season + \beta_2 * month + \beta_3 * holiday + \beta_4 * Weekday + \beta_5 * Weathersit + \beta_6 * Temp + \beta_7 * Hum + \beta_8 * windspeed$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3266.5	395.5	8.259	3.33e-15	***
season2	1242.6	285.1	4.358	1.74e-05	***
season3	583.5	347.8	1.678	0.094270	.
season4	1828.7	291.1	6.281	1.03e-09	***
mnth2	198.4	224.4	0.884	0.377115	
mnth3	827.5	267.6	3.093	0.002149	**
mnth4	325.0	389.2	0.835	0.404298	
mnth5	263.2	417.6	0.630	0.528928	
mnth6	273.3	427.9	0.639	0.523465	
mnth7	261.4	470.0	0.556	0.578405	
mnth8	825.3	451.4	1.828	0.068416	.
mnth9	1394.2	396.1	3.519	0.000492	***
mnth10	542.6	374.1	1.450	0.147877	
mnth11	-371.8	359.4	-1.034	0.301656	
mnth12	-314.8	285.9	-1.101	0.271623	
holiday1	-822.4	279.3	-2.945	0.003458	**
weekday1	370.3	171.7	2.157	0.031727	*
weekday2	461.8	167.6	2.755	0.006193	**
weekday3	569.6	168.2	3.387	0.000788	***
weekday4	617.6	168.9	3.656	0.000297	***
weekday5	662.2	169.1	3.916	0.000109	***
weekday6	702.9	167.7	4.191	3.55e-05	***
workingday1	NA	NA	NA	NA	
weathersit2	-578.2	125.0	-4.625	5.35e-06	***
weathersit3	-2529.6	403.4	-6.270	1.10e-09	***
temp	5043.4	727.0	6.937	2.03e-11	***
hum	-1723.5	499.1	-3.453	0.000624	***
windspeed	-3345.6	637.6	-5.247	2.73e-07	***

$$\widehat{Log(Bike\ Count)} =$$
$$\beta_0 + \beta_1 * season + \beta_2 * month + \beta_3 * holiday + \beta_4 * Weekday + \beta_5 Weathersit + \beta_6 * Temp + \beta_7 * Hum + \beta_8 * windspeed$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.07469	0.14428	55.966	< 2e-16	***
season2	0.32210	0.10401	3.097	0.00212	**
season3	0.26378	0.12686	2.079	0.03835	*
season4	0.62388	0.10621	5.874	1.02e-08	***
mnth2	0.10013	0.08185	1.223	0.22205	
mnth3	0.17545	0.09761	1.797	0.07316	.
mnth4	0.02143	0.14199	0.151	0.88010	
mnth5	-0.06452	0.15235	-0.423	0.67221	
mnth6	-0.13460	0.15609	-0.862	0.38910	
mnth7	-0.17373	0.17144	-1.013	0.31160	
mnth8	-0.09359	0.16468	-0.568	0.57022	
mnth9	0.01163	0.14451	0.080	0.93589	
mnth10	-0.27459	0.13646	-2.012	0.04498	*
mnth11	-0.25991	0.13110	-1.982	0.04823	*
mnth12	-0.26839	0.10430	-2.573	0.01050	*
holiday1	-0.20852	0.10189	-2.047	0.04147	*
weekday1	0.02365	0.06263	0.378	0.70595	
weekday2	0.11264	0.06115	1.842	0.06633	.
weekday3	0.10421	0.06134	1.699	0.09027	.
weekday4	0.11924	0.06163	1.935	0.05384	.
weekday5	0.13872	0.06168	2.249	0.02515	*
weekday6	0.15031	0.06118	2.457	0.01452	*
workingday1	NA	NA	NA	NA	
weathersit2	-0.08590	0.04561	-1.883	0.06052	.
weathersit3	-1.45798	0.14718	-9.906	< 2e-16	***
temp	1.40153	0.26520	5.285	2.26e-07	***
hum	-0.50190	0.18206	-2.757	0.00615	**
windspeed	-0.94773	0.23259	-4.075	5.74e-05	***

This model is very useful to study how different factors affect the customer preferences in renting bikes. For an average buyer, we can use this model as a quick way to predict the bike count. The model is easy to use, and some variables are normalized.

From the base model, we see that season, holiday, temperature, weather, humidity, wind speed are all statistically significant. For our business, in winter (September 22nd to December 20th) the demand for bikes peaks, and is lowest in spring (December 21st to March 30th). We suspect, on an average, the number of bikes rented decreases in cold months post December. We can also interpret, all else held constant, a one degree increase in temperature will increase the demand by 5043/41 = 123 bikes. Also, for holidays on an average the demand is less on holidays as compared to non-holidays when people might be riding to work. People are less likely to use rental bikes when the weather condition is cloudy or foggy, and do not use rental bikes when there is heavy rains and ice.

We also attempted this regression by log transforming (cnt). The log transformed base model resulted in a decreased R^2 from 0.7718 to 0.6211. Thus, going forward we will not apply the log transformation in our regression.

Intuitive Variable Selection

We built a model based on variables that have a strong correlation with bike count and especially can intuitively be explained in the business setting as independent variables. With our former evaluations, we selected season, holiday, weather situation and temperature as the four most intuitively explanatory variables. The intuitive linear regression model is given by :

$$\hat{bike\,count} = \beta_0 + \beta_1 * season + \beta_2 * holiday + \beta_3 * weathersit + \beta_4 * temp$$

Summary of the model is as follows

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1937.2      208.9   9.275  < 2e-16 ***
season2       1138.1      191.4   5.946 6.54e-09 ***
season3        717.1      253.7   2.826 0.004970 **
season4       1668.8      155.7  10.720  < 2e-16 ***
holiday1     -1075.5      301.9  -3.562 0.000417 ***
weathersit2   -755.7      110.6  -6.833 3.59e-11 ***
weathersit3  -3251.2      407.7  -7.974 2.09e-14 ***
temp          6201.9      536.4  11.562  < 2e-16 ***
```

This model, which is based on business understanding, delivers an *R-squared* value of 0.7042 and an AIC of 6091.93. All variables show a statistical significance at the level of 95%.
In this model, spring (December 21st to March 30th) is when bike demand is the lowest, and winter (September 22nd to December 20th) is when bike demand peaks. Bike demand on average is less on holidays relative to non-holidays. People tend to be less likely to use rental bikes when it is cloudy or foggy, and are very unlikely to use rental bikes when there is rain. Moreover, when the temperature is increased by 1 degree, holding other factors constant the demand will increase by approximately 6201 / 41 = 150 bikes.
This intuitive model can be useful in that the inputs are, in general, readily accessible. Even though it might not generate the best performance, the results are still adequately correct as we can see from the *R-squared* value. From a business standpoint, this model is relatively cost-saving and time-efficient, which can be crucial under the competitive business setting. Later on we will examine model with more predictive variables to improve *R-squared* and AIC further.

Backward Elimination

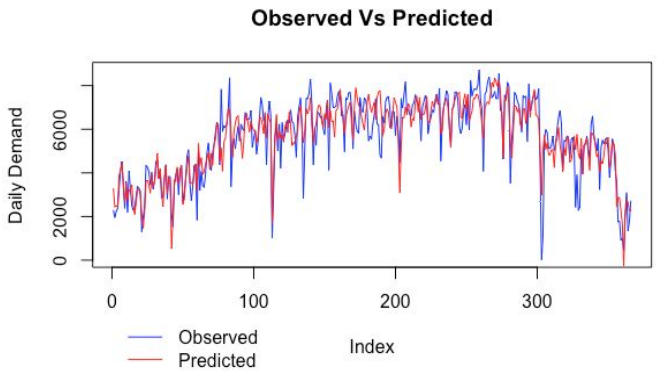
We now use the step() function in R to automatically generate a backward elimination process to select appropriate variables for building the linear regression model. We started with the basic model of all 9 remaining independent variables and eliminate insignificant ones based on respective AIC value. The final model excluded "workday" from the set of independent variables, and is given by

$$\hat{bike\,count} = \beta_0 + \beta_1 * season + \beta_2 * month + \beta_3 * holiday + \beta_4 * weekday + \beta_5\, weathersit + \beta_6 * temp + \beta_7 * hum + \beta_8 * windspeed$$

The summary of the model is as follows,

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3266.5      395.5   8.259 3.33e-15 ***
season2       1242.6      285.1   4.358 1.74e-05 ***
season3        583.5      347.8   1.678 0.094270 .
season4       1828.7      291.1   6.281 1.03e-09 ***
mnth2          198.4      224.4   0.884 0.377115
mnth3          827.5      267.6   3.093 0.002149 **
mnth4          325.0      389.2   0.835 0.404298
mnth5          263.2      417.6   0.630 0.528928
mnth6          273.3      427.9   0.639 0.523465
mnth7          261.4      470.0   0.556 0.578405
mnth8          825.3      451.4   1.828 0.068416 .
mnth9         1394.2      396.1   3.519 0.000492 ***
mnth10         542.6      374.1   1.450 0.147877
mnth11        -371.8      359.4  -1.034 0.301656
mnth12        -314.8      285.9  -1.101 0.271623
holiday1      -822.4      279.3  -2.945 0.003458 **
weekday1       370.3      171.7   2.157 0.031727 *
weekday2       461.8      167.6   2.755 0.006193 **
weekday3       569.6      168.2   3.387 0.000788 ***
weekday4       617.6      168.9   3.656 0.000297 ***
weekday5       662.2      169.1   3.916 0.000109 ***
weekday6       702.9      167.7   4.191 3.55e-05 ***
weathersit2   -578.2      125.0  -4.625 5.35e-06 ***
weathersit3  -2529.6      403.4  -6.270 1.10e-09 ***
temp          5043.4      727.0   6.937 2.03e-11 ***
hum          -1723.5      499.1  -3.453 0.000624 ***
windspeed    -3345.6      637.6  -5.247 2.73e-07 ***
```

Our final model delivered a higher *R-squared* of 0.7881 and a lower AIC of 4967.209. The automatic selection process left the model with independent variables that can most significantly influence the demand in rental bikes. In general the conclusions we have drawn from the result of our intuitive model still hold true in the final model. In addition, September seems to be the most popular month in terms of number of bikes demanded, possibly because of the starting of most schools' fall semesters, while November and December display a downward trend in demand when compared to January, which most likely can be explained by the decrease in weather and end-of-year holidays such as Thanksgiving and Christmas. Moreover, people on average have a less demand on rental bikes when it is more humid and when there is heavier wind. When we have more data on current environmental and seasonal conditions, our final model shows a better performance in generating a decent prediction on rental bike demand. The following is the observed vs. predicted comparison plot for the final model.



Evaluation

All the models arrive at the same conclusion. The variables that most significantly affect demand for rental bikes are:

1. Weather Situation
2. Season
3. Holiday
4. Temperature

The key takeaways to our report are listed in the following

- Boxplots and histograms show that the count of bikes rented is most influenced by season, month and weather situation (weathersit).
- Base model shows statistical significance in 6 variables out of 8, and has a good enough *R-squared* of 0.7718.
- Log transformation does not improve the fit of the model, as can be seen in the lowering of the *R-squared* (0.7718 to 0.6211).
- Intuitive variable selection is cost-saving and time-efficient method. Since it has a good enough *R-squared* value of 0.7042, it is a good approach.
- Backward elimination provided the final model of the highest accuracy, but is less intuitive than the above model. The use of numerous variables also imposes the danger of overfitting.

Deployment

Based on the results of these models and analysis, bike rental companies can have a whole picture of how many bikes to put on reserve each day with change in environmental and seasonal conditions. Due to the limitation of data size, firms should notice that seasonal conditions related to the bike rental behaviors may need further research. Also, the results are concluded based on the historical data back in 2012 when bike sharing systems were relatively new to the market. The bike sharing market may have changed in some ways today and the firms would be aware of that. In general, our model and results are a good exploration of people's bike rental behaviors under different situations and can help companies promote their bike supply strategy.

Appendix 1

- instant: record index
- dteday : date
- season : season (1:spring, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : whether day is a holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday it is 1, otherwise it is 0.
+ weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered