CSDA1010WM18 - Basic Methods of Data Analysis

Course Project

Examining the Driving Factors Of Demand in Bike Share Rentals

Team Name:
Wecka Flocka Flame

Members:
Devin Woods
Michael Baranov
Franck Boutaud
Ansar Mek Koodathil
Joseph Hirmiz

Business Problem

For many years bicycle (hereinafter referred to as "bike") rental shops have provided a service to customers looking for a temporary mode of transportation. Following the advent of the bike, rental shops began servicing the public's desire for short-term access to these two-wheeled, non-motorized vehicles. Bike rental shops are typically located nearby popular outdoor tourist destinations, where travellers can rent a bike to explore the surrounding area, providing a much needed alternative for those who choose not to travel with their own bike, or for those who simply do not own one themselves. It would be difficult to accurately assess the total number of bike rental shops throughout North America at any given time, as many are owned by sole proprietors ("mom-and-pop" stores), but one thing is for certain: there is a continued, and ever-present, need for bike rental shops.

Bike rental shops—at least the kind previously described—are less likely to succeed in a large metropolitan area, where brick-and-mortar is far more expensive. The breakeven point, as a result of these large monthly fixed expenditures, can prove to be out-of-reach for many who attempt to compete in the space. Seasonality and climate also carry unique challenges to owner-operators in the industry.

In place of, or at the very least adjunct to, the archetypal brick-and-mortar bike rental shop is the bike share company. In Seattle, Washington, there are currently three companies—LimeBike, Spin, and Ofo—that provide bike share services across the city. Each company's respective fleet of bikes is colour-coded (LimeBike bikes are painted a lime colour, for example), and there is no present, or past, restriction over who can operate where. Customers need only own an iOS or Android smartphone device in order to participate in the service. Customers can obtain a bike from one designated location and return it to another. In cities like Seattle, bike share companies largely target commuters, or tourists seeking to explore the city for a day. Said tourists may not want to have to worry about peddling all the way back from a destination of their choosing simply to return a bike, for instance.

The three bike share companies discussed above all operate under the following model: registered riders pay a monthly membership fee for unlimited rides, and non-registered riders pay on an as-needed, per-ride basis.

Team Wecka Flocka Flame obtained the hourly records of bike share usage data for Seattle, Washington, and aims to determine what factors influence the consumption of this service. We will provide a linear regression model, as well as randomForest model, to the three aforementioned bike share companies so that they may better understand the drivers of demand in their industry, as well as predict the bike share rental count. Additionally, the insights gleamed from our analysis will shed light on when—that being the time of day and day of week—each company should stock the most bikes. Our intent is to, with a high level of certainty, determine what non-controllable attributes, such as weather, contribute the most to the count of bike share rentals. To accomplish this, we will utilize the statistical modelling software R to perform data mining procedures on the dataset obtained, and run it through a linear regression model.

Data Understanding

We obtained the dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset).

The data was split into two datasets: the first detailing usage numbers on a daily basis, and the second on an hourly basis. We chose to examine the hourly dataset.

The hourly dataset—*hourly.csv*—contains 17,379 records. Each row in the dataset reflects a specific hour, of a specific date, of consumption data, spanning two full years: January 1, 2011 to December 31, 2012. More recent data was unavailable. The dataset contains 15 potential independent variables and 1 dependent variable. The dependent variable chosen for our analysis is the "*cnt*"—that being the number of bike share rentals. The dataset does not specify the number of bike share rentals by company; therefore, our analysis will not attempt to determine, in any way, which of the three companies hold what proportion of Seattle, Washington's bike

share rental market. We will only seek to understand the underlying drivers of demand based on the attributes contained the dataset.

Dataset Dimensions:

```
> dim(hour)
[1] 17379    17
```

The dataset attributes are described below:

Table 1—Data Dictionary

Attribute Name	Attribute Description
dteday	Date of record
season	Season (1: spring; 2: summer; 3: fall; 4: winter)
yr	Year (0: 2011; 1: 2012)
mnth	Month (1: January; … 12: December)
hr	Hour (0 to 23)
holiday	Holiday or Not Holiday
weekday	Day of the week
workingday	If day is neither a weekend or a holiday, then 1; otherwise 0.
weathersit	1: Clear, Few Clouds, Partly Cloudy 2: Mist + Cloudy, Mist + Broken Clouds, Mist + Few Clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered Clouds, Light Rain 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Normalized temperature in Celsius—values are divided to 41 (max)
atemp	Normalized feeling temperature in Celsius—values are divided to 50 (max)
hum	Normalized humidity—values are divided to 100 (max)
windspeed	Normalized wind speed—values are divided to 67 (max)
casual	Count of casual users (not a registered rider who pays a monthly membership fee)
registered	Count of registered users (riders who pay a monthly membership fee)
cnt	**DEPENDENT VARIABLE**—Count of total bike share rentals

After reading the *hourly.csv* file into R Studio, we ran a summary report:

Figure 1 – Summary of Data Frame Statistics

```
> summary(hour)
       dteday           season           yr              mnth           hr
 2011-01-01:   24   Min.   :1.000   Min.   :0.0000   5      :1488   16     :  730
 2011-01-08:   24   1st Qu.:2.000   1st Qu.:0.0000   7      :1488   17     :  730
 2011-01-09:   24   Median :3.000   Median :1.0000   12     :1483   13     :  729
 2011-01-10:   24   Mean   :2.502   Mean   :0.5026   8      :1475   14     :  729
 2011-01-13:   24   3rd Qu.:3.000   3rd Qu.:1.0000   3      :1473   15     :  729
 2011-01-15:   24   Max.   :4.000   Max.   :1.0000   10     :1451   12     :  728
 (Other)   :17235                                   (Other):8521   (Other):13004
    holiday           weekday      workingday weathersit     casual         registered
 Min.   :0.00000   Min.   :0.000   0: 5514    1:11413    Min.   :  0.00   Min.   :  0.0
 1st Qu.:0.00000   1st Qu.:1.000   1:11865    2: 4544    1st Qu.:  4.00   1st Qu.: 34.0
 Median :0.00000   Median :3.000              3: 1419    Median : 17.00   Median :115.0
 Mean   :0.02877   Mean   :3.004              4:    3    Mean   : 35.68   Mean   :153.8
 3rd Qu.:0.00000   3rd Qu.:5.000                         3rd Qu.: 48.00   3rd Qu.:220.0
 Max.   :1.00000   Max.   :6.000                         Max.   :367.00   Max.   :886.0

      cnt           raw.temp       raw.atemp     raw.windspeed       raw.hum
 Min.   :  1.0   Min.   : 0.82   Min.   : 0.00   Min.   : 0.000   Min.   :  0.00
 1st Qu.: 40.0   1st Qu.:13.94   1st Qu.:16.66   1st Qu.: 7.002   1st Qu.: 48.00
 Median :142.0   Median :20.50   Median :24.24   Median :12.998   Median : 63.00
 Mean   :189.5   Mean   :20.38   Mean   :23.79   Mean   :12.737   Mean   : 62.72
 3rd Qu.:281.0   3rd Qu.:27.06   3rd Qu.:31.06   3rd Qu.:16.998   3rd Qu.: 78.00
 Max.   :977.0   Max.   :41.00   Max.   :50.00   Max.   :56.997   Max.   :100.00

      rush
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2513
 3rd Qu.:1.0000
 Max.   :1.0000
```

On the next page is a view of the data structure and the characteristics of each of the dependent variables and independent variable:

Figure 2 – Summary of Data Frame Structure

```
> str(hour)
'data.frame':   17379 obs. of  17 variables:
 $ dteday       : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ season       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ yr           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ mnth         : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ hr           : Factor w/ 24 levels "0","1","2","3",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ holiday      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ weekday      : int  6 6 6 6 6 6 6 6 6 6 ...
 $ workingday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ weathersit   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 2 1 1 1 1 ...
 $ casual       : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered   : int  13 32 27 10 1 1 0 2 7 6 ...
 $ cnt          : int  16 40 32 13 1 1 2 3 8 14 ...
 $ raw.temp     : num  9.84 9.02 9.02 9.84 9.84 ...
 $ raw.atemp    : num  14.4 13.6 13.6 14.4 14.4 ...
 $ raw.windspeed: num  0 0 0 0 0 ...
 $ raw.hum      : num  81 80 80 75 75 75 80 86 75 76 ...
```

We then visualized and examined the attributes of the dataset that we found to be significant in making our predictions. We explored these variables primarily through bivariate analysis by comparing them to the output variable, "*cnt*," as this would give us a better understanding of how each feature affects total rentals. This is in contrast to univariate analysis, which we did not use for this dataset because viewing the distribution of these attributes on their own would not yield informative results (i.e., displaying the distribution of weekdays, seasons, temperatures, or years).

Data Preparation

The *hourly.csv* dataset, upon close inspection, appears to be very clean, with 16 attributes, including the dependent variable (but excluding the index column).

Missing Values

After examining the dataset for possible missing values we confirmed the existence of none. The lack of missing values is due, in whole, to the automated nature of bike share usage records. All data was captured via sensors fixed to kiosks located at the pick-up and drop-off points.

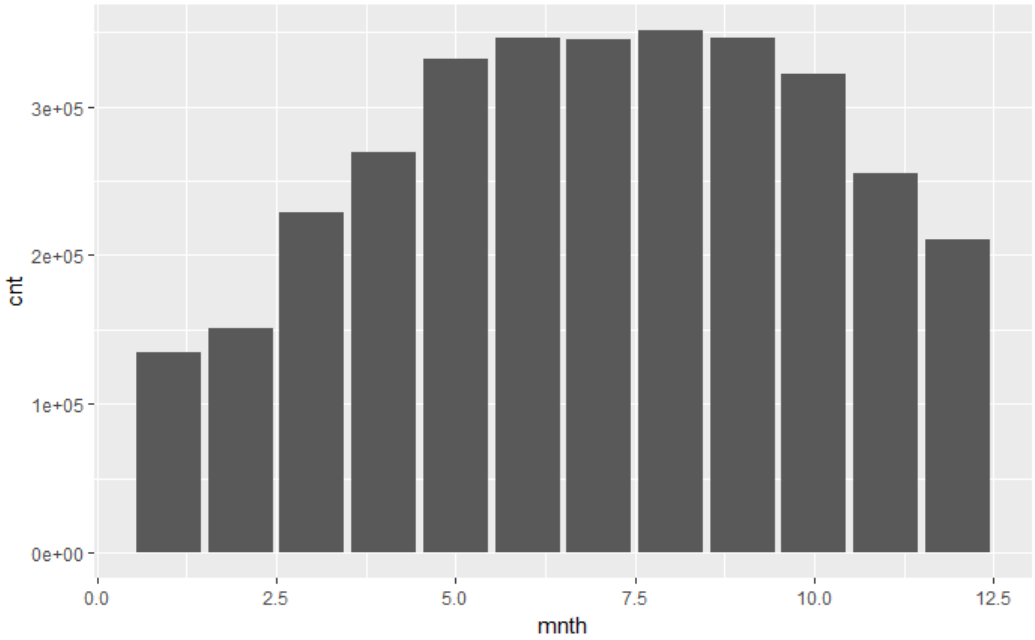
Outliers

We identified approximately 50 records—all of which were clustered across two consecutive days in October 2012 in the dataset—considered to be outliers. We discovered, through additional research, that these two days coincided with the peak of Hurricane Sandy. As a result of extreme weather conditions, the count of bike share rentals plummeted during this 48-hour period. These hourly records were scrubbed from the dataset so as not to skew our analysis.

Data Exploration and Transformation

After cleaning the dataset, we then performed statistical analyses on the various independent variables to obtain an understanding of how they correlate to the dependent variable, "*cnt*."

Figure 3 – Distribution of Bike Share Rentals by Month



In the above histogram, which shows the count of bike share rentals by month (January being 1; December being 12), we can see that the colder months—October through to April—yield a lower number of bike share rentals. In comparison, the warmer summer months—May through to September—have the highest number of bike share rentals.

Figure 4 – Boxplot Showing Distribution of Total Bike Share Rentals by Month

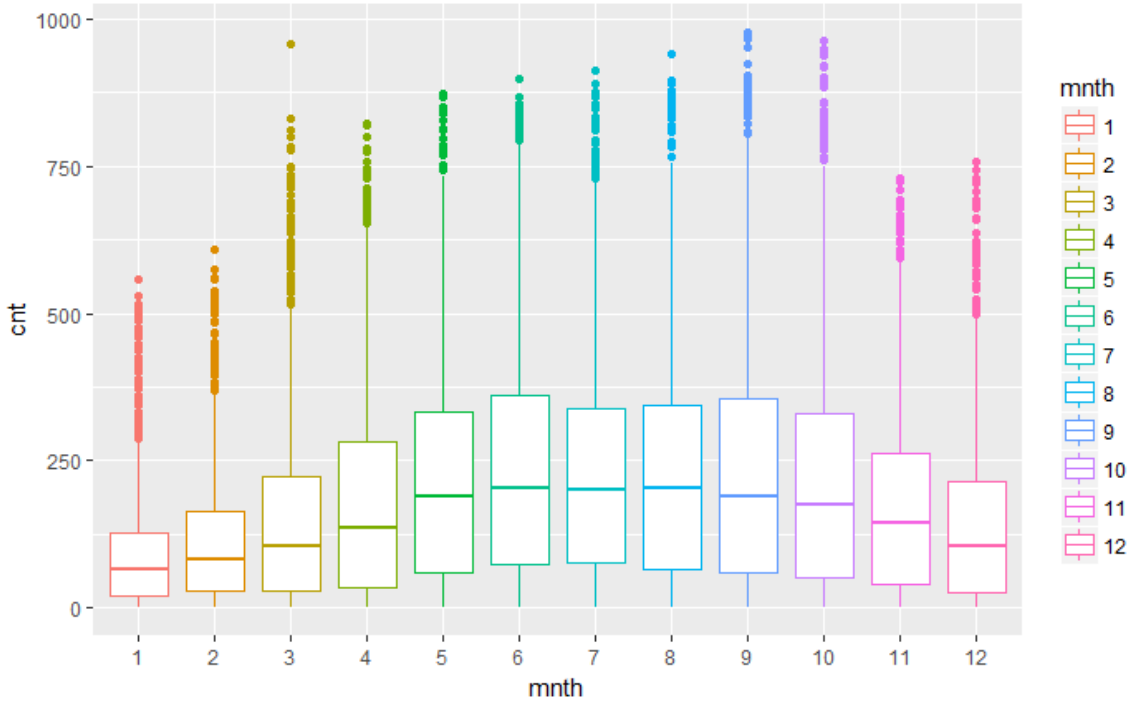
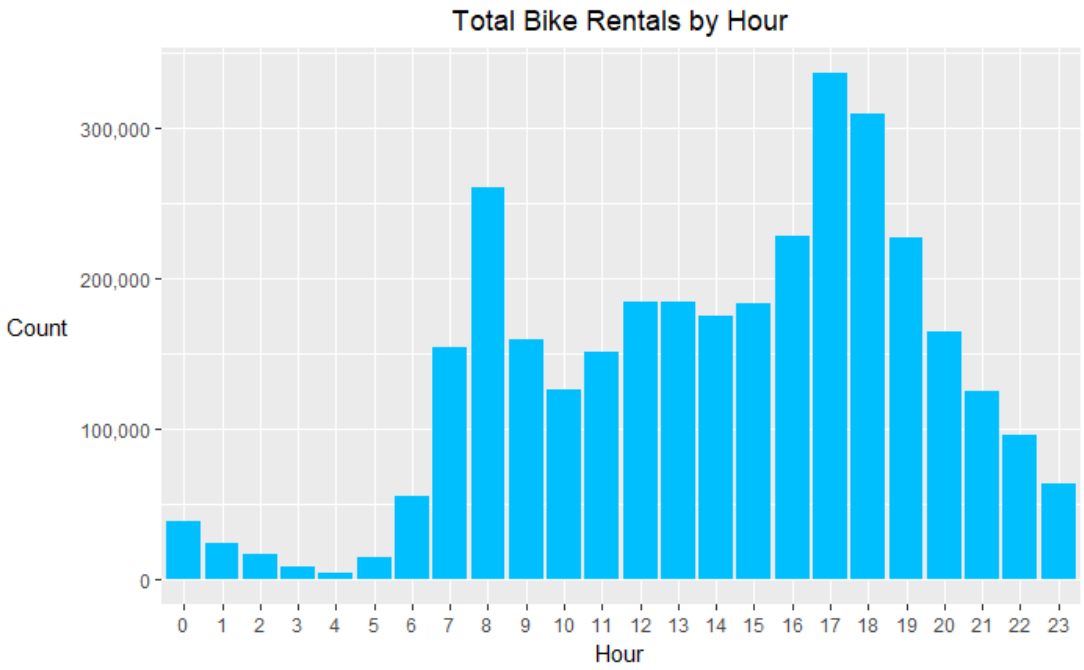


Figure 4 shows a similar distribution but visualizes all of the outliers in the dataset. However, we chose not to consider these true outliers that would need to be removed from the dataset because we found that these points exist due to a large spike in rentals around 8am and 5pm (see Figure 5 to follow), and reoccur each day of the work week, explaining the prevalence of these outliers. The outliers do not fit within the whisker of the boxplot, which extends only 1.5 times the Interquartile Range, but they are common occurrences.

Figure 5 – Total Bike Rentals by Hour of Day



In addition to viewing the rental count by month, we were interested in seeing if there is more or less activity during particular hours of the day. Figure 5 above shows, again, that there are spikes in bike share usage around 8am and 5pm. This is explained by the nature of individuals commuting to and from work in the city as it aligns with the typical person's work day.

Figure 6 – Bike Share Rental Count at Various 'Real' Temperatures

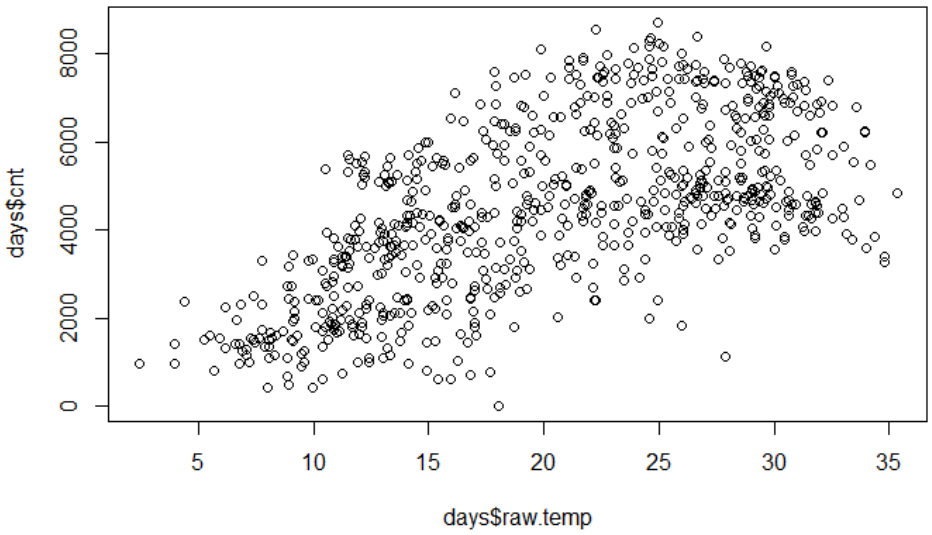


Figure 7 – Bike Share Rental Count at Various 'Feeling' Temperatures

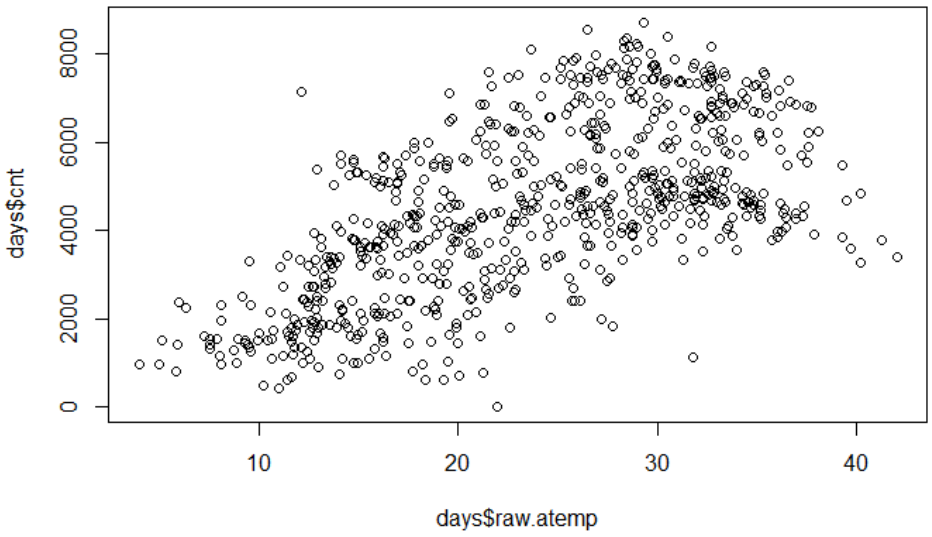


Figure 6 and Figure 7 display the next two independent variables of interest—those being real temperature ("*raw.temp*") and feeling temperature ("*raw.atemp*"). We examined both attributes because we were curious to see if the number of bike share rentals is associated with how hot or cold it is on any given day.

Figure 6 contains the real temperature, which excludes humidity and windspeed, and appears to be positively linearly correlated to a degree with the bike share rental count. Figure 7 contains the feeling temperature, which accounts for humidity and windspeed, and the graph, again, clearly shows a relationship to bike share rental count as seen previously with the real temperature graph.

Note that the real temperature and feeling temperature attributes used were not part of the original dataset. We created new columns with the temperatures converted to degrees Celsius by multiplying each value by a specific fixed number, as explained in the meta-data that was attached to the dataset. The dataset used to create the scatterplots for temperature was aggregated by day into a more condensed set called "days," because plotting the temperature by hour revealed a very cluttered and unreadable graph.

Model – Multiple Linear Regression

Our team chose to utilize a multiple linear regression model because our output variable, "*cnt*," is continuous and so are most of the independent variables. First, we transformed the data into training and testing sets, based on an assumed 80/20 split, respectively. The code used to accomplish this is shown below:

```
train.rows = createDataPartition(y= hour$cnt, p=0.8, list = FALSE)
train.data<- hour[train.rows,] # 80% data goes in here
test.data<- hour[-train.rows,] # 20% data goes in here
```

Next, we constructed our model fit based on the training set, which included all of the variables just to see how the model would fare:

```
test.lm = lm(formula = cnt ~., data = train.data)
summary(test.lm)
```

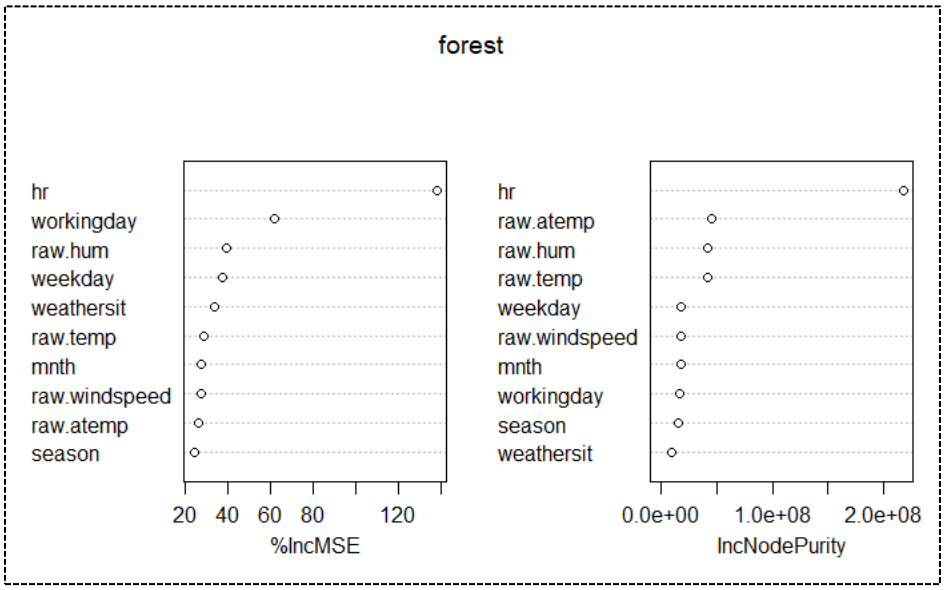
```
dteday2011-07-04  5.575e-12  1.001e-12  5.570e+00 2.60e-08 ***
dteday2011-07-05  5.895e-12  9.652e-13  6.107e+00 1.04e-09 ***
dteday2011-07-06  5.967e-12  9.370e-13  6.369e+00 1.97e-10 ***
dteday2011-07-07  5.793e-12  9.811e-13  5.904e+00 3.63e-09 ***
dteday2011-07-08  5.807e-12  9.540e-13  6.087e+00 1.18e-09 ***
dteday2011-07-09  5.855e-12  9.751e-13  6.005e+00 1.96e-09 ***
dteday2011-07-10  2.573e-12  9.642e-13  2.668e+00 0.007631 **
dteday2011-07-11  5.876e-12  9.704e-13  6.055e+00 1.44e-09 ***
dteday2011-07-12  6.239e-12  9.628e-13  6.480e+00 9.51e-11 ***
dteday2011-07-13  5.905e-12  9.798e-13  6.026e+00 1.72e-09 ***
dteday2011-07-14  5.823e-12  9.535e-13  6.107e+00 1.04e-09 ***
dteday2011-07-15  3.593e-12  9.772e-13  3.677e+00 0.000237 ***
dteday2011-07-16  1.222e-11  9.542e-13  1.281e+01  < 2e-16 ***
dteday2011-07-17  6.910e-12  9.390e-13  7.359e+00 1.96e-13 ***
dteday2011-07-18  5.699e-12  9.921e-13  5.745e+00 9.41e-09 ***
dteday2011-07-19  6.027e-12  9.963e-13  6.049e+00 1.49e-09 ***
 [ reached getOption("max.print") -- omitted 545 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.914e-12 on 13165 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.408e+28 on 738 and 13165 DF,  p-value: < 2.2e-16
```

We can see that including all of the variables in the model yields a very inaccurate and overfitted response. The R-squared is 1, and the P-value is miniscule. This inaccuracy is caused by including the variables "*registered*" and "*casual*," as the output variable, "*cnt*," is simply the sum of these two figures.

Instead, we fitted the model, but included only the variables we found to be of significance by creating a randomForest for the fit.

```
forest = randomForest(cnt ~
raw.temp+raw.atemp+season+hr+mnth+weathersit+workingday+weekday+raw.windspeed+raw.hum, data=train.data, importance=TRUE, ntree=200)
```

Figure 8 – Output of randomForest Model



This forest concluded that the most important variables are the following:

- The hour of day;
- Whether it is a working day or not;
- Feeling temperature; and
- Weather condition.

For our revised model, we decided to use feeling temperature instead of real temperature because the tree placed the former at a higher level of significance. In addition, including both temperatures would skew the model since they are highly correlated with one another (cor = 0.9877). We also excluded humidity and windspeed from the model because the feeling temperature already takes these factors into account, and thus they are redundant.

The revised model code is as follows on the next page:

```
test.lm = lm(formula = cnt ~ hr+raw.atemp+weathersit+weekday, data = train.data)
summary(test.lm)
```

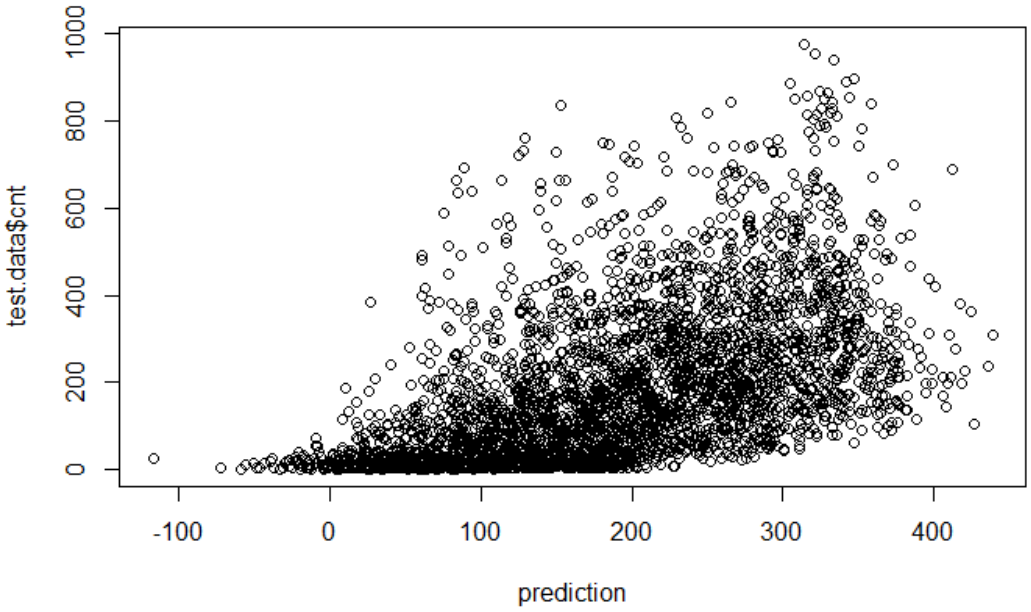
Figure 9 – Multiple Linear Regression Output

```
Call:
lm(formula = cnt ~ hr + raw.atemp + weathersit + weekday, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-340.33 -101.86  -30.19   56.78  699.68

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.2371     5.7338 -10.506  < 2e-16 ***
hr            9.0565     0.1906  47.509  < 2e-16 ***
raw.atemp     7.3781     0.1542  47.839  < 2e-16 ***
weathersit  -27.6442     2.0436 -13.527  < 2e-16 ***
weekday       3.1584     0.6506   4.855 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 154 on 13899 degrees of freedom
Multiple R-squared:  0.2896,    Adjusted R-squared:  0.2894
F-statistic:  1416 on 4 and 13899 DF,  p-value: < 2.2e-16
```

A much more realistic image is provided in the above output summary. As a result of the relatively low R-squared value of 0.2896, which means that only 28.9% of the data is fitted to the regression line, we chose to model the data using a second approach, which we discuss on the next page.

```
prediction = predict(test.lm, test.data)
plot(prediction, test.data$cnt)
```

Figure 10 – Predicted Output Values As Compared to Actual Output Values



Model – randomForest

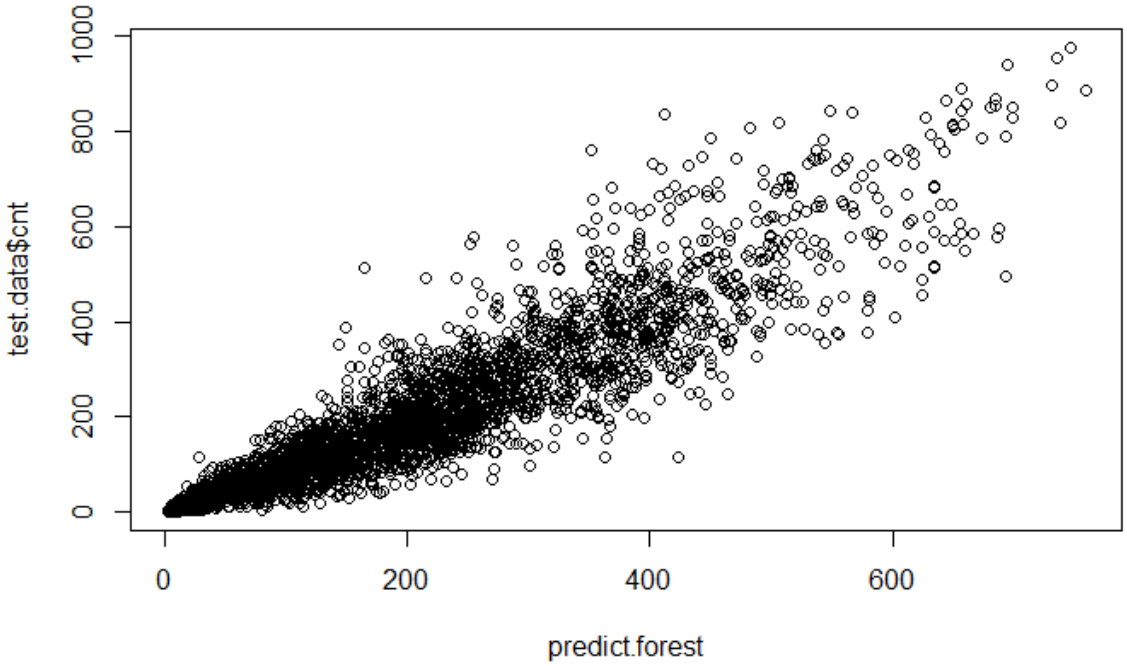
We utilized a randomForest model to compare to the multiple linear regression model, using the same tree as before:

```
forest = randomForest(cnt ~ raw.atemp+hr+weathersit+workingday+weekday, data=train.data,
importance=TRUE, ntree=200)
```

And our prediction was as follows:

```
predict.forest <- predict(forest, test.data)
plot(predict.forest, test.data$cnt)
```

Figure 11 – randomForest Predicted Values As Compared to Actual Output Values



Evaluation

From the randomForest we generated, and with the use of common sense, we found several attributes to be most significant in predicting the hourly bike rental count:

- *hr* – the hour of day
- *raw.atemp* – the feeling temperature at a given hour
- *weathersit* – the weather situation at a given hour, which ranges between 1 and 4 (1 being calm weather, and 4 being the most extreme)
- *workingday* – whether it is a working day (1=yes, 0=no)
- *weekday* – if it is a weekday (1=yes, 0=no)

The revised linear regression model we used had an adjusted R-squared value of only 0.2894, which means, as previously stated, that only 28.9% of our data fit to the regression line. After applying the model to the test set, we found the correlation between the vectors of predicted outcomes and actual outcomes to be 0.53745, which means that our model predicted the hourly

bike rental count correctly only about half the time. This is not a very good prediction outcome, so we compared it with the randomForest model. The randomForest yielded a correlation coefficient of 0.9262, meaning that the model was able to predict the bike rental count accurately approximately 92.6% of the time. This is a much better result than the linear regression model was able to provide, so we conclude that the randomForest was a more ideal fit for this predicting rentals in this case.

Conclusion

Based on our analysis, the factors that influence bike share rental consumption that most are the hour of the day, the feeling temperature, the weather situation, and whether it is a working day or not. The hour of day is primarily tied to an individual's work day, and, as we previously discussed, the daily usage peaks take place around 8am and 5pm. The feeling temperature drives demand, with bike share rentals being at their lowest when the temperature is too hot or too cold. The weather situation, too, influences demand as riders will typically not choose to utilize bike share services if the weather conditions are too extreme (for example, during a snow storm, or when there is a heavy rain).

The randomForest model that we constructed was capable of accurately predicting the results by identifying and heavily weighting the independent variables of the most significance. Our team will seek to deploy the prediction model to the three bike share rental companies located in Seattle. If, and when, they provide additional, and more up-to-date, bike share rental data, we will continue to further reinforce of the parameters of the model.