PREDICTION OF

BIKE RENTAL COUNT

By

Kishore Kumar Nakka

Advisor

Jiming Wu

SUMMER 2020 – BAN 693

CALIFORNIA STATE UNIVERSITY – EAST BAY

BUSINESS ANALYTICS (MSBA)

Table of Contents

Chapter 1

Introduction

1.1 Problem Statement

The aim of this project is to predict the bike rental count on a particular day along with season, weather setting, and temperature. The advantage of predicting bike rental count will be scope for the management to maintain exact number of bikes according to the seasons weather conditions without losing the customers lack of bikes.

1.2 Data

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

Available at: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Our task is to build the regression model upon the training data and verify using the test data. Given below is the sample of data.

Bike Rental Count (Columns: 1-10)

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp
1	2011-01-01	1	0	1	0	6	0	2	0.344167
2	2011-01-02	1	0	1	0	0	0	2	0.363478
3	2011-01-03	1	0	1	0	1	1	1	0.196364
4	2011-01-04	1	0	1	0	2	1	1	0.200000
5	2011-01-05	1	0	1	0	3	1	1	0.226957
6	2011-01-06	1	0	1	0	4	1	1	0.204348

Bike Rental Count (Columns: 11-16)

atemp	hum	windspeed	casual	registered	cnt
0.363625	0.805833	0.1604460	331	654	985
0.353739	0.696087	0.2485390	131	670	801
0.189405	0.437273	0.2483090	120	1229	1349
0.212122	0.590435	0.1602960	108	1454	1562

S.No.	Predictor	Description
1	instant	Record index
2	dteday	Date
3	season	Season (1:spring, 2:summer, 3:fall, 4:winter)
4	yr	Year
5	mnth	Month (1 to 12)
6	holiday	weather day is holiday or not
7	weekday	Day of the week
8	workingday	if day is neither weekend nor holiday is 1, otherwise is 0.
9	weathersit	Listing of how the weather is (clear, cloudy, snow, rain, etc)
10	temp	Normalized temperature in Celsius. The values are divided to 41 (max)
11	atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (max)
12	hum	Normalized humidity. The values are divided to 100 (max)
13	windspeed	Normalized wind speed. The values are divided to 67 (max)
14	casual	count of casual users
15	registered	count of registered users
16	cnt	count of total rental bikes including both casual and registered

As you can see in the table below we have the following 16 variables, using which we have to correctly predict the count of bike rental

Here are the variables and description accordingly:

Except the 16th variable 'cnt' the other 15 will be our predictor variables.

Chapter 2

Methodology

2.1 Pre Processing

Any predictive model requires that we look at the data before we start modeling. However, in data mining terms, *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data, as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process, we will first try and look at all the probability distributions of the variables. Most analysis like regression require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

Converting data into required format

```
day$season = as.factor( as.character( day$season))

day$yr = as.factor( as.character( day$yr))

day$holiday = as.factor( as.character( day$holiday))

day$workingday = as.factor( as.character( day$workingday))

day$mnth = as.factor( as.character( day$mnth))

day$weekday = as.factor( as.character( day$weekday))

day$weathersit = as.factor( as.character( day$weathersit))
```

Missing value Anaysis

Checking data whether there are any missing values in the data.

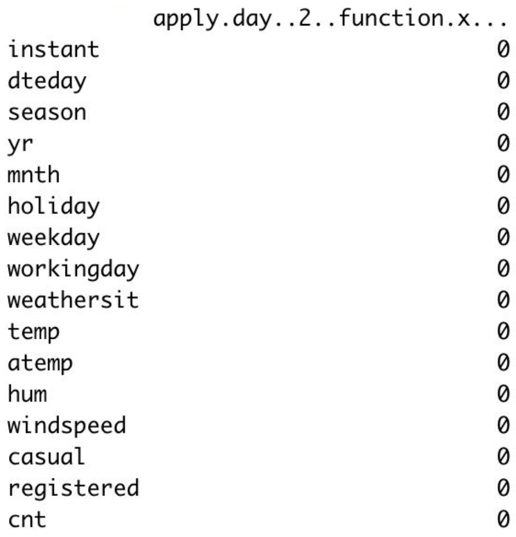
```
#missing value analysis

missing_val = data.frame( apply( day, 2, function(x) {sum( is.na(x))}))

missing_val

###no missing values  found in the data
```

Output:

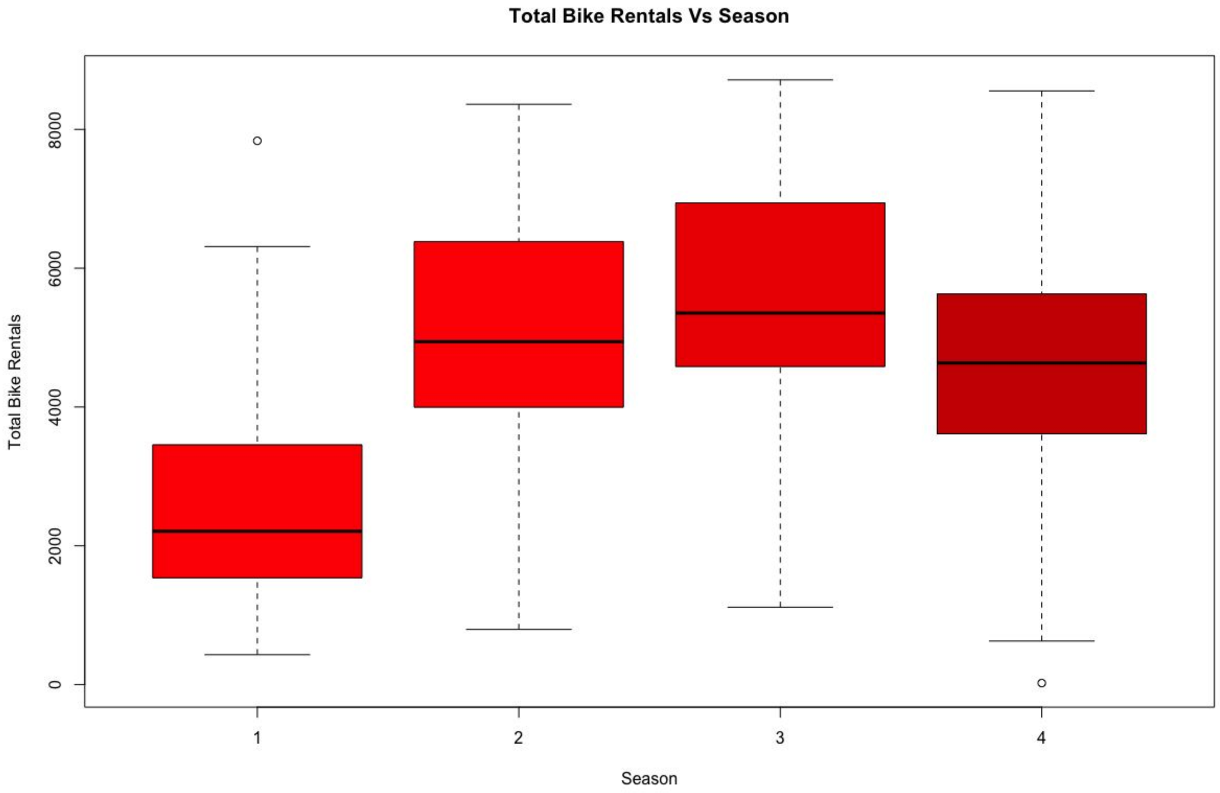


```
           apply.day..2..function.x...
instant                              0
dteday                               0
season                               0
yr                                   0
mnth                                 0
holiday                              0
weekday                              0
workingday                           0
weathersit                           0
temp                                 0
atemp                                0
hum                                  0
windspeed                            0
casual                               0
registered                           0
cnt                                  0
```

2.1.1 Outlier Analysis

Outliers are those data points which are away from the normal data. Outliers cause a skewness in the data and make model inconsistent. Outliers can be identified using boxplots and can be removed from the data. There is a need to remove outliers in the data.

Boxplot of season and bike rentals:

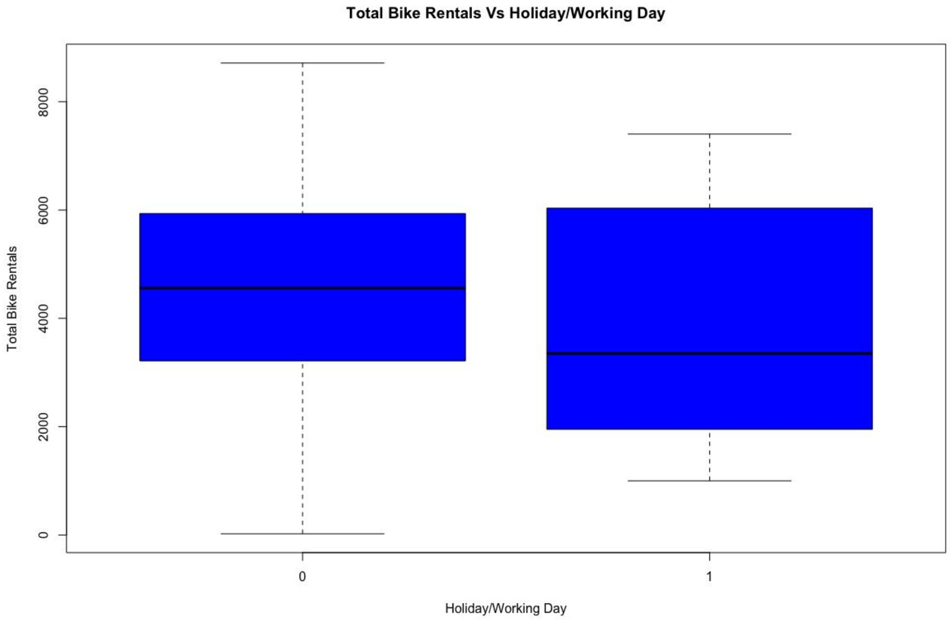


From here, we can see that the the bike rentals were more in the bracket of summer and fall comparetively. Highest rentals were during fall. Where as lowest in spring.

Also, we have outliers in the 1 and 4.

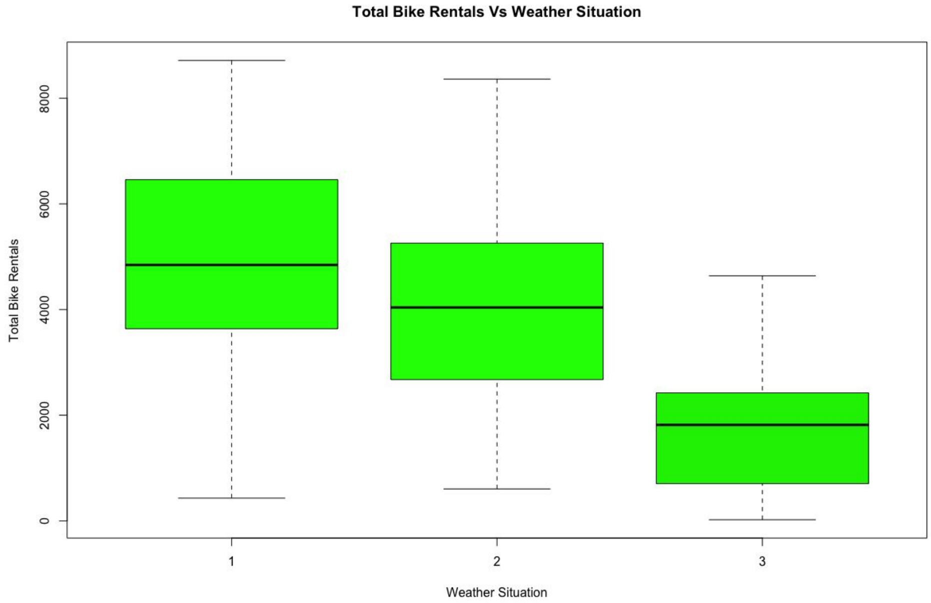
1 is considered spring, 2 is considered summer, 3 is considered fall, 4 is considered winter.

BoxPlot Bikerentals vs Workingday:



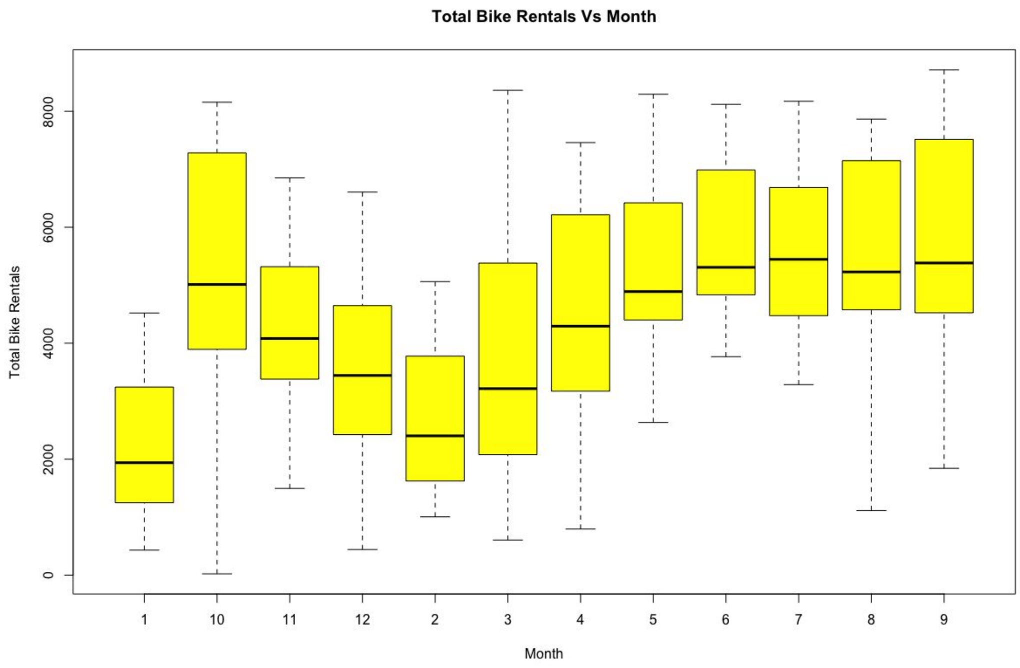
From the graph below, the comparision says the average rentals were more on holidays than the working days. But the total usage is more in the working days than the holidays.

Boxplot for Total bike rental vs weathersit :



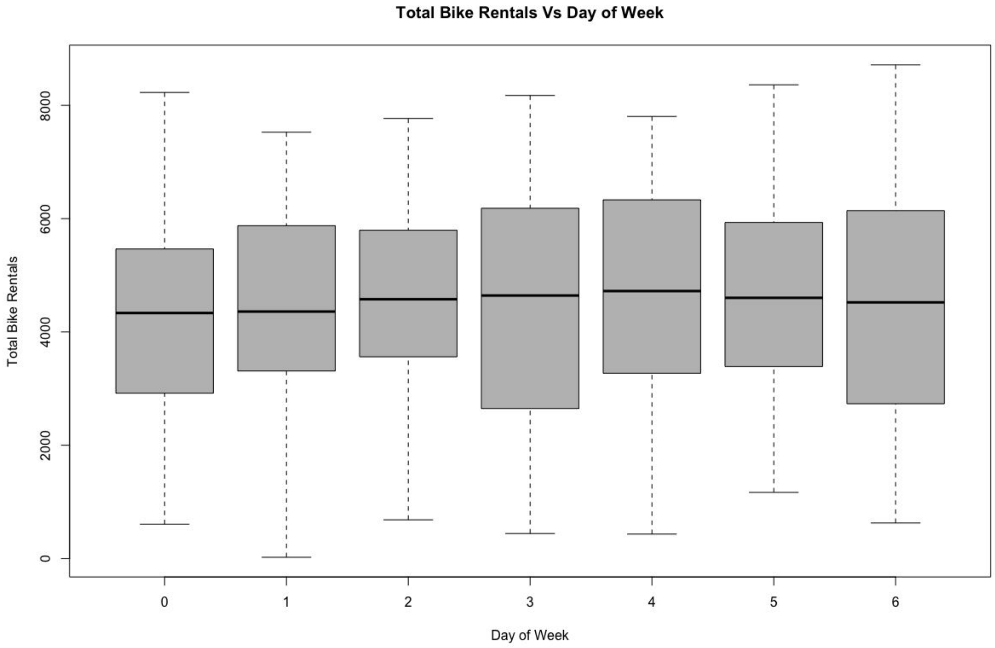
We can see from the above plots that it is the lowest on the weather situation 3. So, temperature also plays a major role.

Boxplot for Bike Rental vs Month:



By the graph, we can say the months from June (6) – October (10) are high.

Boxplot for BikeRental vs weekday:

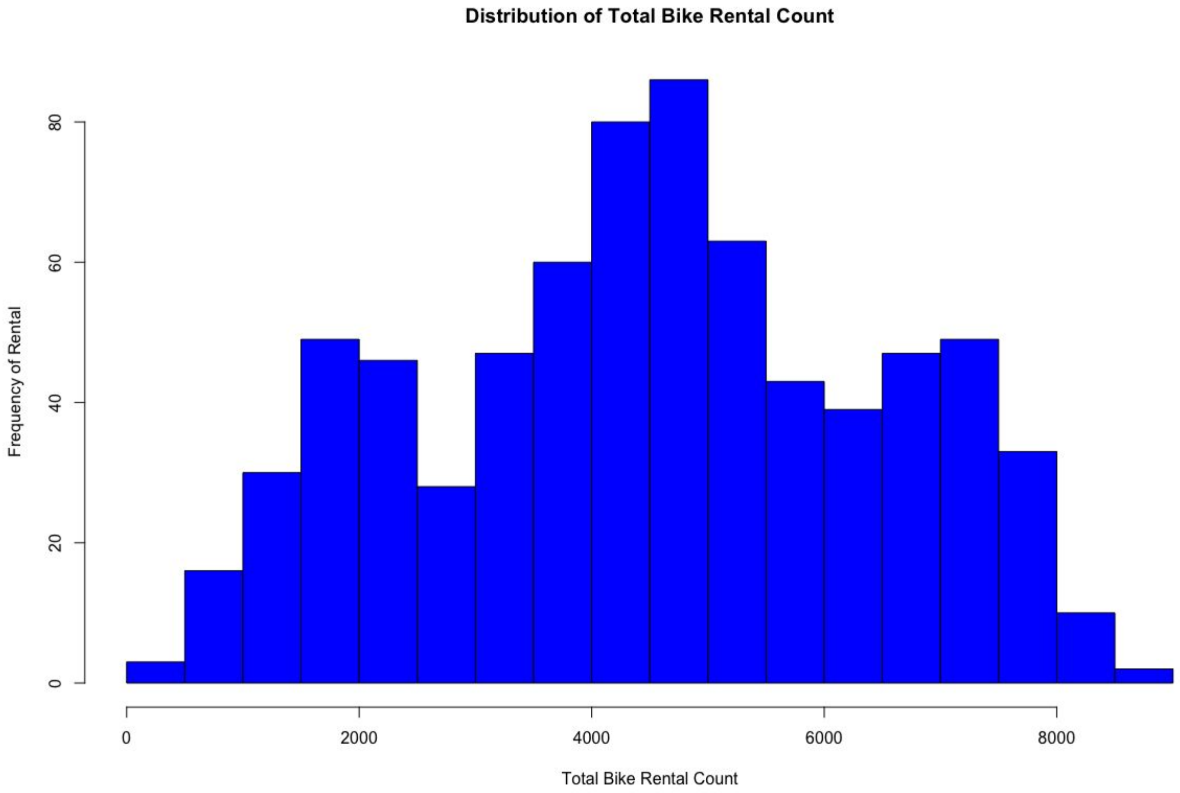


We can say that almost everyday of the week has similar number of users.

Now, we will be analyzing the data by using visualizations.

The following figures explain Data distribution of contiuous variables using different chart and graph plots

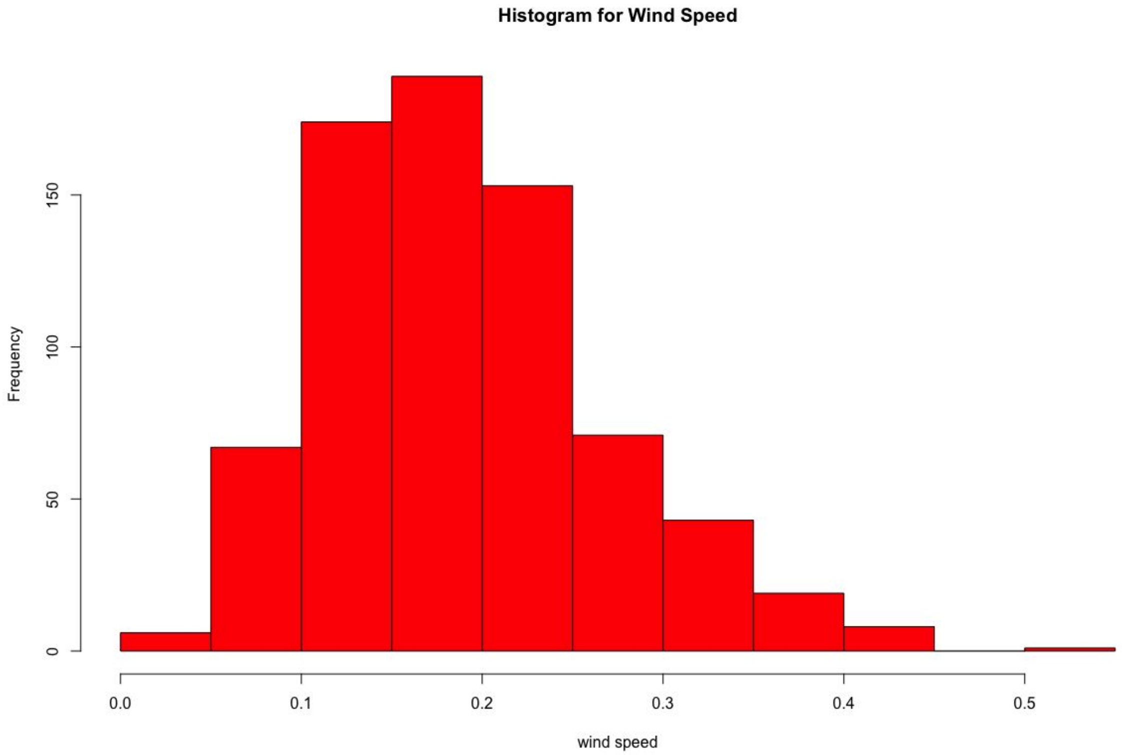
Distribution of total bike rentals

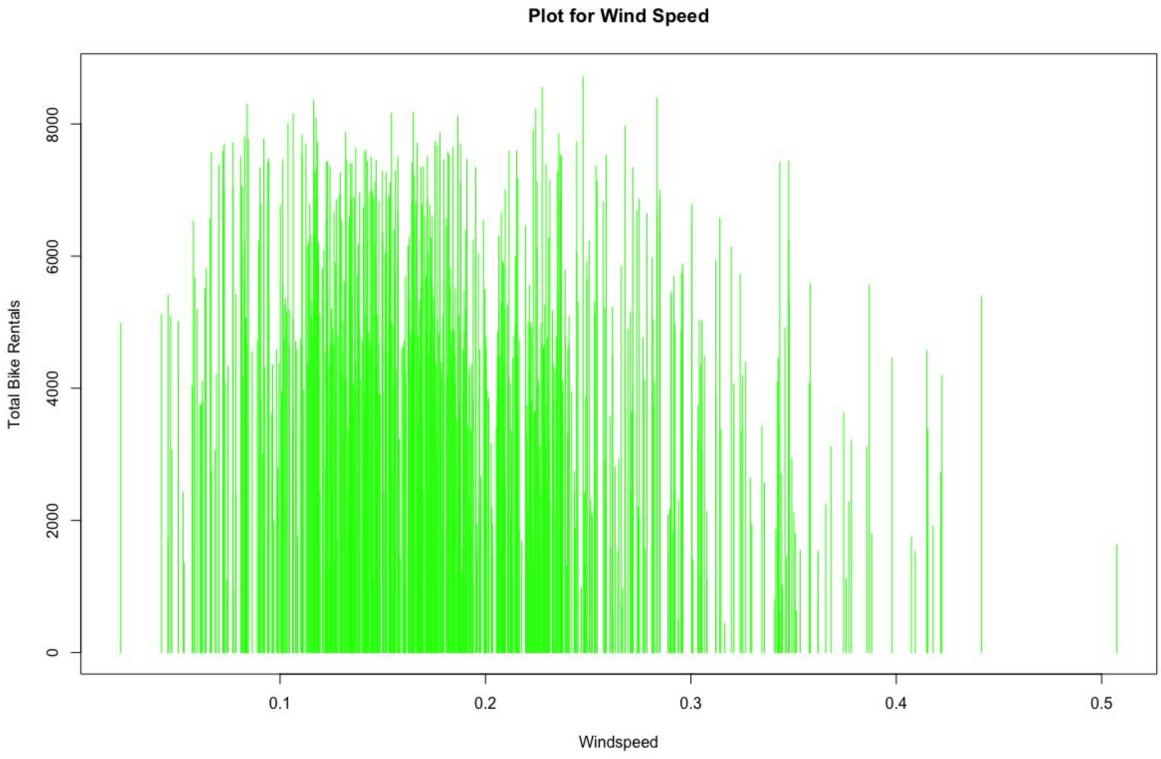


We can see from the above image that the total bike rental count is distributed unbiasedly which is symmetric.

So, we can say that the data is somewhat evenly distributed.

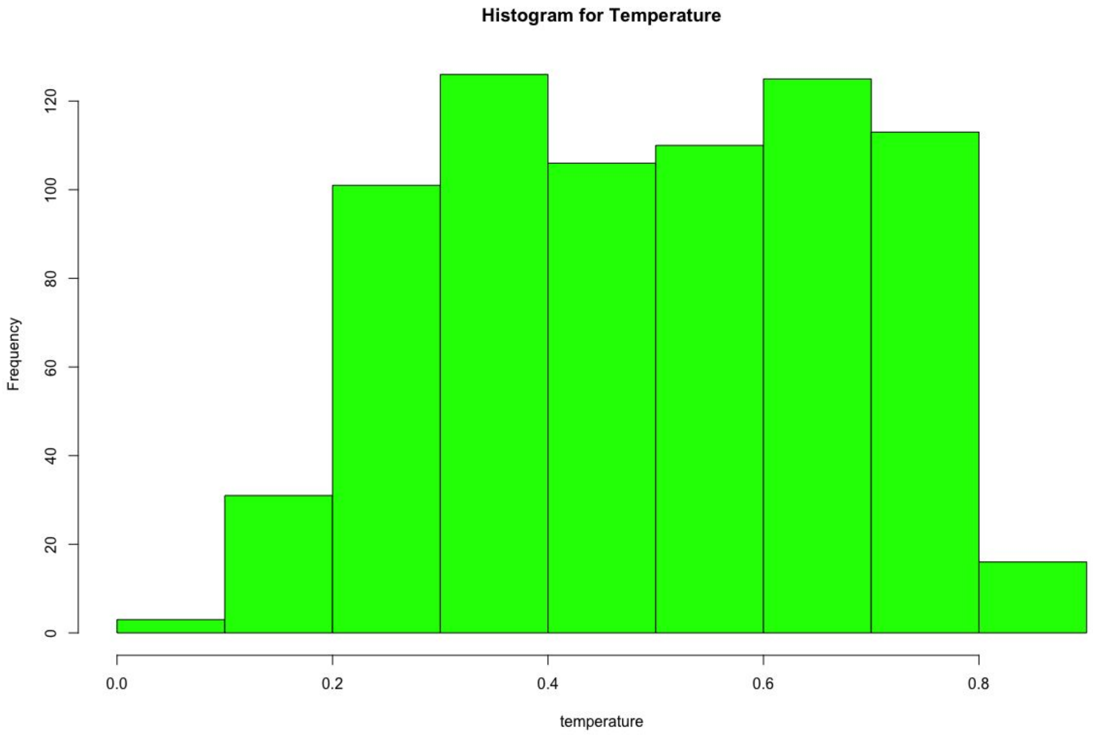
Rentals according to Wind Speed

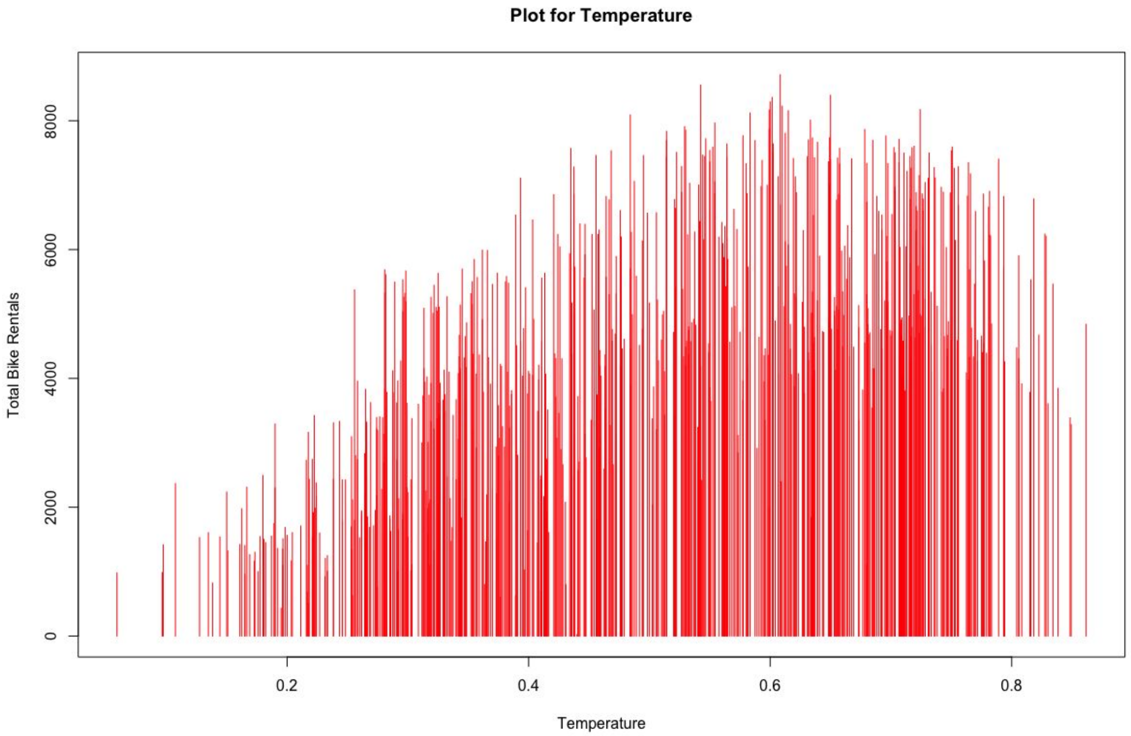




From the image above, we can see that the distribution is moderately right skewed. So, we can say by the axis that, most people would more likely rent when the wind speed is relatively lower.

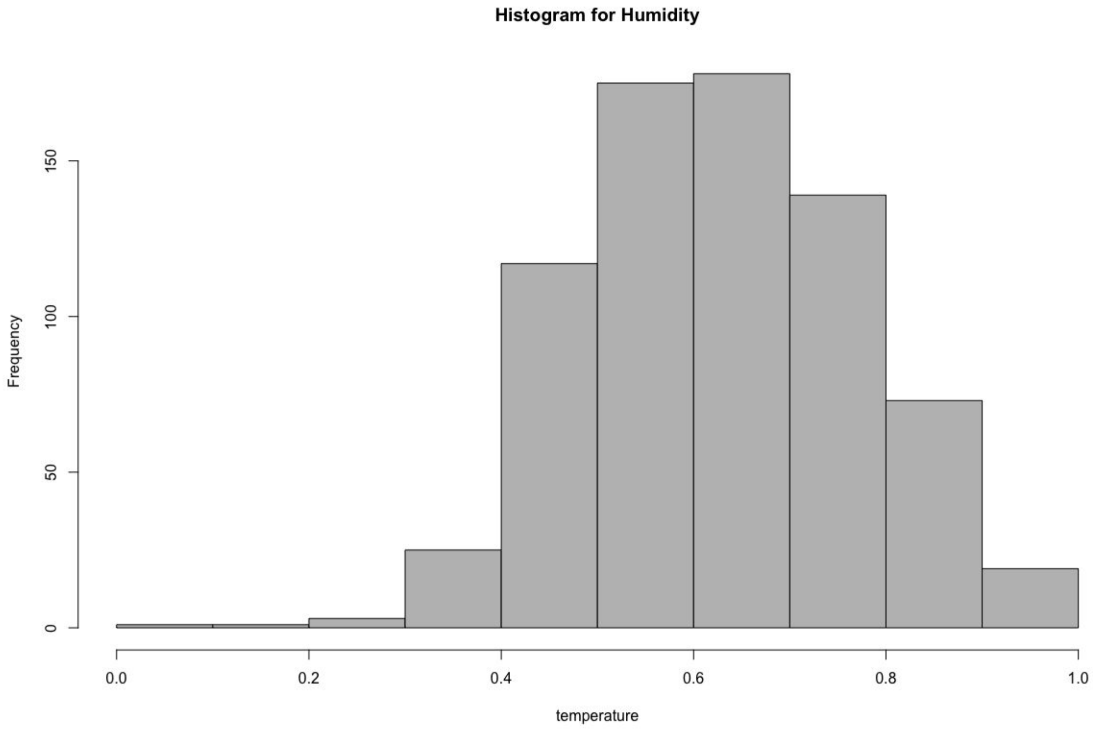
Rentals according to Temperature

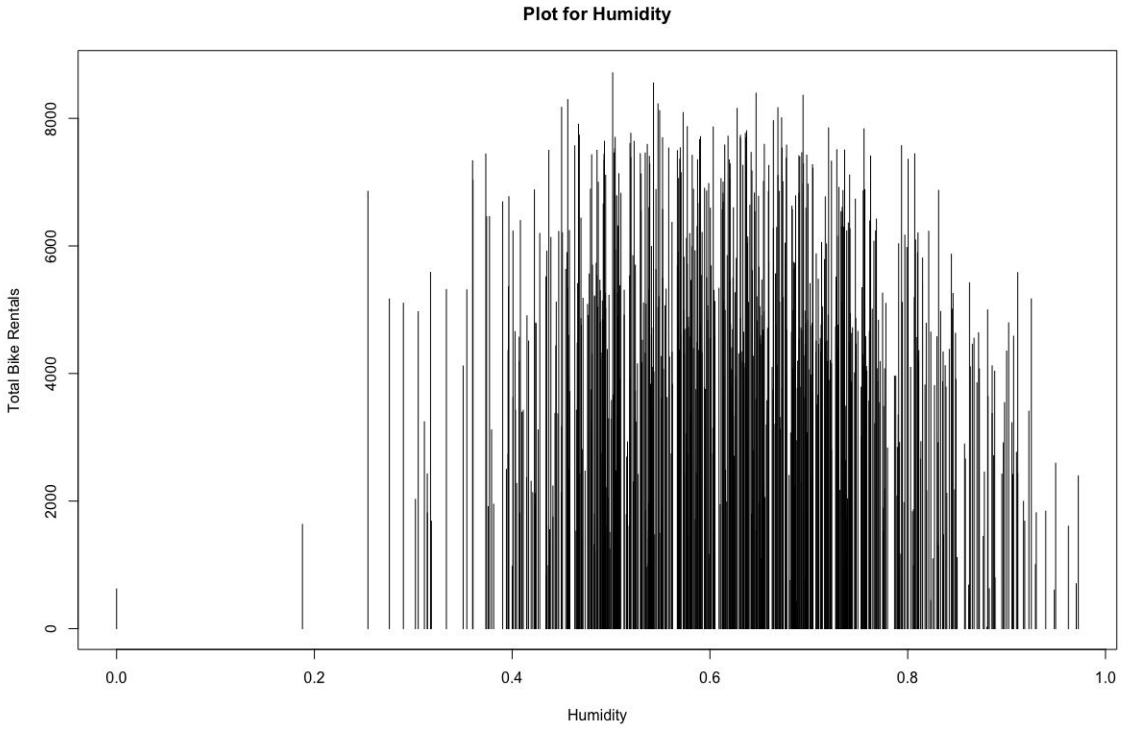




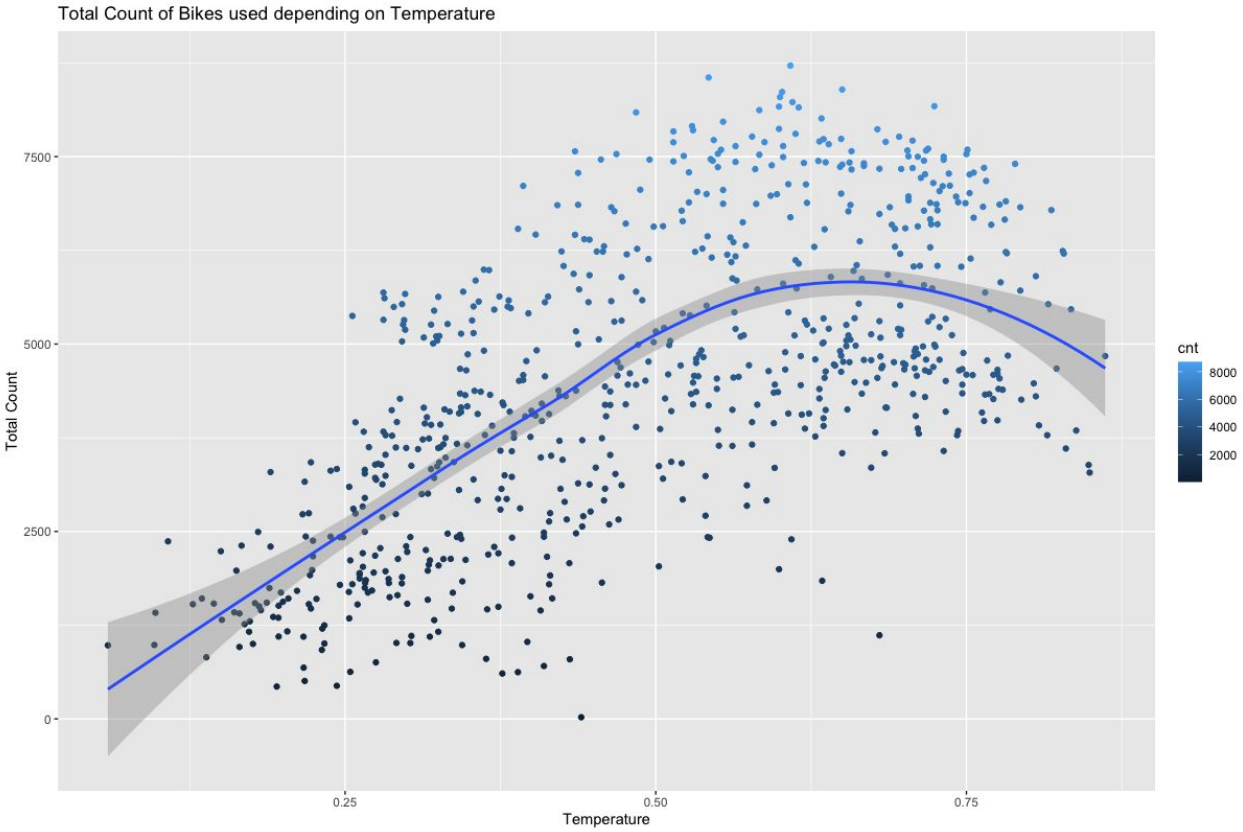
From the images, we can say that less rentals were made when the temperature is too low.

Rentals according to Humidity





By the histogram of the above image, we can see it is skewed to the left. So, more rentals are likely to be made when there is humidity.



The following inferences can be drawn from the data:

It was seen that the number of registered users was overall higher when compared to that of casual users. Specifically when classified by working days, it was found that more number of registered bikes that were rented on the days when it was neither a weekend nor a holiday as compared to casual bike rentals which were more during situations when it could have been a weekend or a holiday. This possibly helps us to have an understanding of purpose and type of bike rentals. Registered users might use bikes on a daily basis, example for work or other day to day activities where as casual bike rentals are associated with holiday and leisure.

The highest number of bike rentals were between the months of June and August, whereas the lowest number of bike rentals were between the months November and February. This gives us an indication about the role of corresponding seasons associated with these months. Furthermore, I also checked for total count of bike rentals across seasons which confirms that the highest rentals were made during fall, followed by summer, winter and lastly in spring.

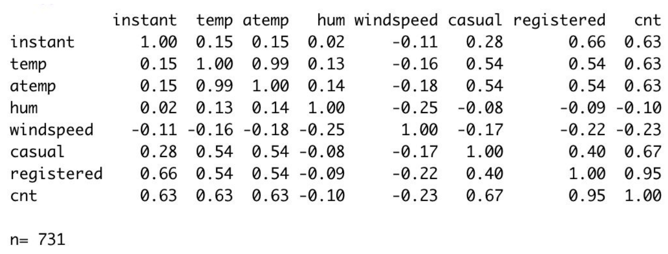
Another important factor to understand the bike rental trends is temperature. There was not a significant difference between real temperature and feeling of the temperature. Total Temperature was significantly correlated with total bike rentals. Irrespective of the type of bike rentals, the temperature was equally correlated with both casual and registered rentals. However, one important thing to mention again is that the registered users were higher overall compared to casual users. Hence the results could be biased or misleading, but overall it seems like temperature plays an important role for total count. Furthermore, weather situation was found to be significant predictor of bike rentals. However,the impact of holidays is not significant. One reason for holidays not being significant could be that there were probably a very few holidays during the years for it to have a significant impact.

2.1.2 Feature Selection

Before performing any type of modelling, we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of regression. There are several methods of doing that. Below we have used correlation and anova tests for feature selection.

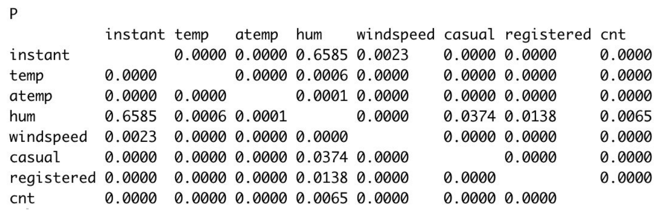
Another step of Exploratory Data Analysis is to look for highly correlated variables in the data. A very simple way of looking at correlations in the data is shown below.

res = rcorr (as.matrix (num_data))



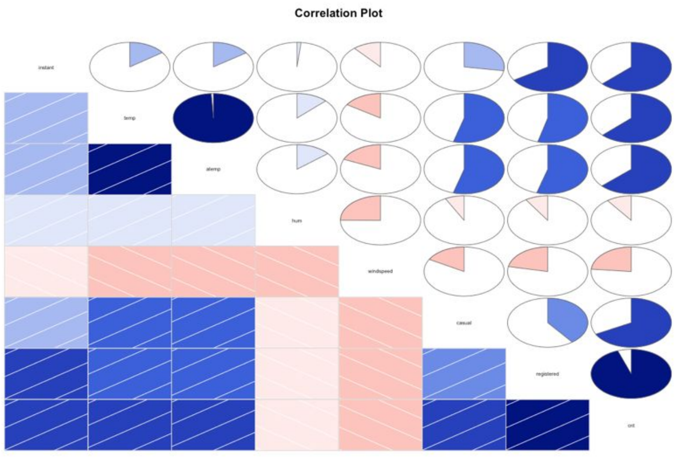
```
           instant  temp atemp   hum windspeed casual registered   cnt
instant       1.00  0.15  0.15  0.02     -0.11   0.28       0.66  0.63
temp          0.15  1.00  0.99  0.13     -0.16   0.54       0.54  0.63
atemp         0.15  0.99  1.00  0.14     -0.18   0.54       0.54  0.63
hum           0.02  0.13  0.14  1.00     -0.25  -0.08      -0.09 -0.10
windspeed    -0.11 -0.16 -0.18 -0.25      1.00  -0.17      -0.22 -0.23
casual        0.28  0.54  0.54 -0.08     -0.17   1.00       0.40  0.67
registered    0.66  0.54  0.54 -0.09     -0.22   0.40       1.00  0.95
cnt           0.63  0.63  0.63 -0.10     -0.23   0.67       0.95  1.00

n= 731
```

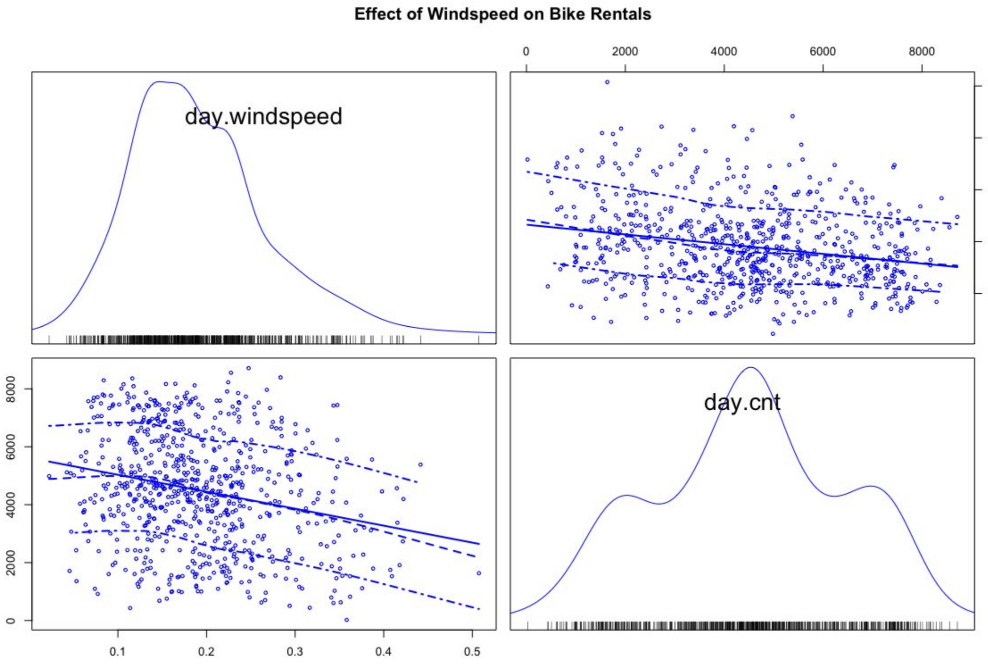


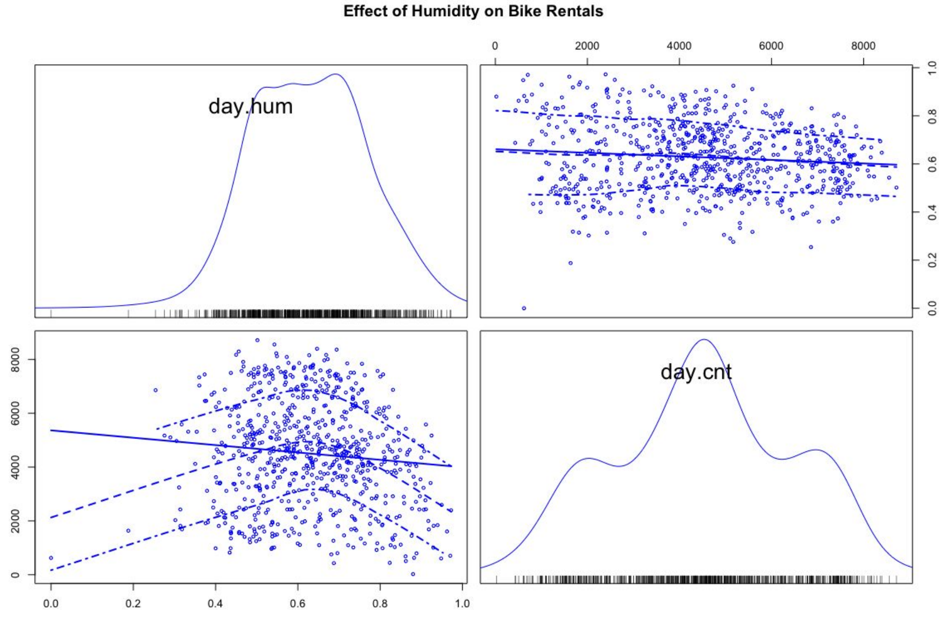
```
P
           instant temp   atemp  hum    windspeed casual registered cnt
instant            0.0000 0.0000 0.6585 0.0023    0.0000 0.0000     0.0000
temp       0.0000         0.0000 0.0006 0.0000    0.0000 0.0000     0.0000
atemp      0.0000  0.0000        0.0001 0.0000    0.0000 0.0000     0.0000
hum        0.6585  0.0006 0.0001        0.0000    0.0374 0.0138     0.0065
windspeed  0.0023  0.0000 0.0000 0.0000           0.0000 0.0000     0.0000
casual     0.0000  0.0000 0.0000 0.0374 0.0000           0.0000     0.0000
registered 0.0000  0.0000 0.0000 0.0138 0.0000    0.0000            0.0000
cnt        0.0000  0.0000 0.0000 0.0065 0.0000    0.0000 0.0000
```

corrgram (day[,num_index], order = F,upper.panel=panel.pie, text.panel=panel.txt, main = "Corr Plot")



Scatterplot for windspeed and humidity





From the above correlation plot and scatterplots it is clear that the variables windspeed and humidity have negative correlation. The slopes of the variables are also negative. This indicates these variables are negatively correlated.

Anova test for categorical variables

Anova test is used to analyse if there are any significant difference in groups of categorical variables. It compares means of dependent continuous variable among the groups of categorical variable and checks if differences are statistically significant.

Given below are the null and the alternative hypothesis:

Null Hypothesis: all group means are equal —> there is no relationship between categorical variable and dependent variable, which we can write as follows:

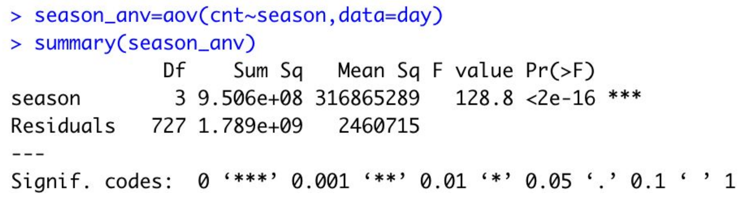
H0: all means are equal

Alternative Hypothesis: not all group means are equal —> there is a relationship between categorical variable and dependent variable.:

H1: not all means are equal

F statistics = Variation among sample means / Variation within groups

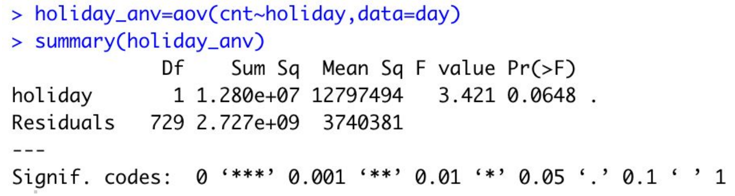
Through the F statistics, we can see if the variation among sample means dominates over the variation within groups, or not. In the first case we will have strong evidence against the null hypothesis (means are all equals), while in the second case we would have little evidence against the null hypothesis.



```
> season_anv=aov(cnt~season,data=day)
> summary(season_anv)
             Df    Sum Sq   Mean Sq F value Pr(>F)    
season        3 9.506e+08 316865289   128.8 <2e-16 ***
Residuals   727 1.789e+09   2460715                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our p-value is less than 0.05 (as suggested by normal scientific standard). Hence we can conclude that for our confidence interval we accept the alternative hypothesis H1 that there is a significant relationship between variables.

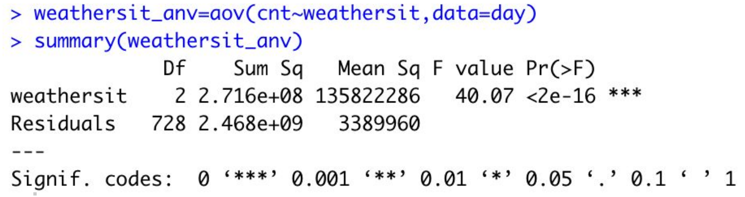
So, we can say that there is a change in count when there is a change in season. And so, the season variable is important.



```
> holiday_anv=aov(cnt~holiday,data=day)
> summary(holiday_anv)
             Df    Sum Sq  Mean Sq F value Pr(>F)  
holiday       1 1.280e+07 12797494   3.421 0.0648 .
Residuals   729 2.727e+09  3740381                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our p-value is more than 0.05 (as suggested by normal scientific standard). Hence we can conclude that for our confidence interval we accept the Null hypothesis H0 that there is a no significant relationship between variables.

So, we can say that there is almost no change in count when there is a change in holiday. So, we can say that the holiday variable is not so important



```
> weathersit_anv=aov(cnt~weathersit,data=day)
> summary(weathersit_anv)
             Df    Sum Sq   Mean Sq F value Pr(>F)    
weathersit    2 2.716e+08 135822286   40.07 <2e-16 ***
Residuals   728 2.468e+09   3389960                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our p-value is less than 0.05 (as suggested by normal scientific standard). Hence we can conclude that for our confidence interval we accept the alternative hypothesis H1 that there is a significant relationship between variables.

So, we can say that there is a change in count when there is a change in weather. And so, we can say that weather variable is important.

By these factor analysis, we can say that the count may change when weather or season changes but not the same with holiday.

2.2 Modeling

2.2.1 Model Selection

The dependent variable can fall in either of the four categories:

1. Nominal
2. Ordinal
3. Interval
4. Ratio

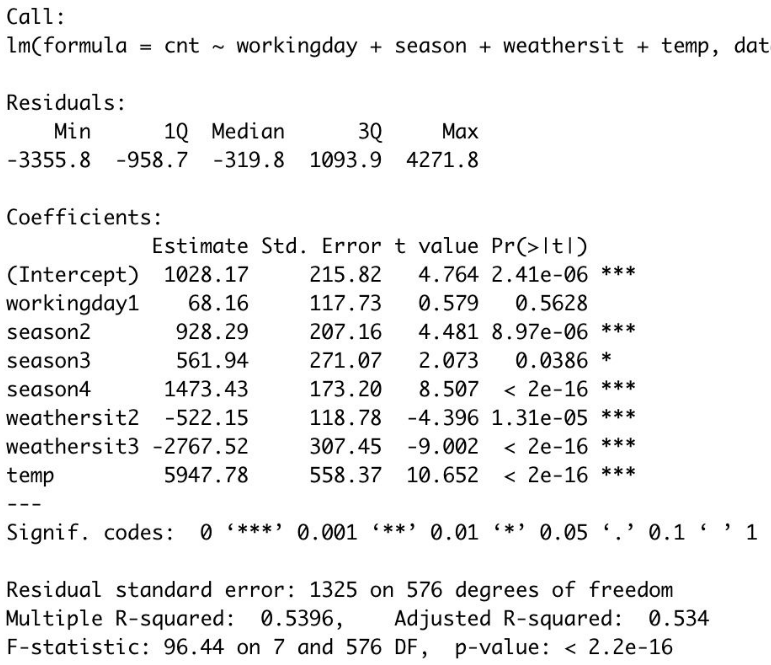
Based on the dependent of variable of your dataset we use a model accordingly. There are two types of supervised learning models. They are classification and regression. We choose one of them depending on the predicting variable. In this case, our depending variable is continuous variable. So, we use regression model. If the variable is categorical we go classification models. We try different regression models and analyse using error metrics choose the model which is optimal.

2.2.2 Linear Regression

lin.mod =lm(cnt ~ season + workingday+ weathersit +hum+ temp, data =train_data)

summary(lin.mod)

Results:



```
Call:
lm(formula = cnt ~ workingday + season + weathersit + temp, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3355.8  -958.7  -319.8  1093.9  4271.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1028.17     215.82   4.764 2.41e-06 ***
workingday1    68.16     117.73   0.579   0.5628    
season2       928.29     207.16   4.481 8.97e-06 ***
season3       561.94     271.07   2.073   0.0386 *  
season4      1473.43     173.20   8.507  < 2e-16 ***
weathersit2  -522.15     118.78  -4.396 1.31e-05 ***
weathersit3 -2767.52     307.45  -9.002  < 2e-16 ***
temp         5947.78     558.37  10.652  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1325 on 576 degrees of freedom
Multiple R-squared:  0.5396,	Adjusted R-squared:  0.534 
F-statistic: 96.44 on 7 and 576 DF,  p-value: < 2.2e-16
```

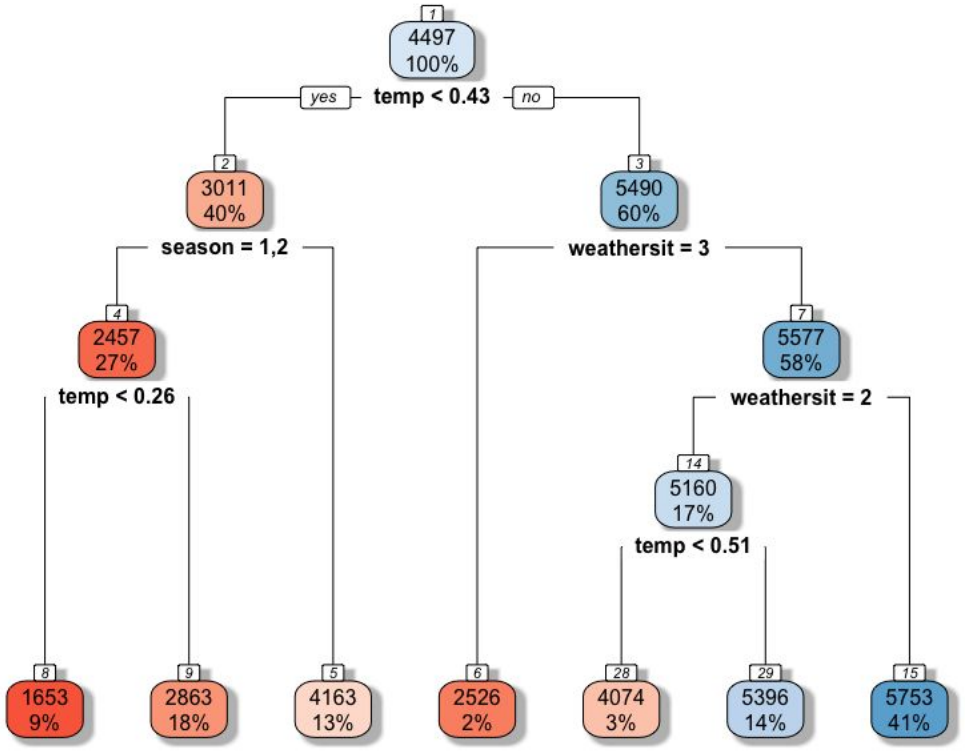
As you can see the *Adjusted R-squared* value, we can explain about 53.4% of the data using our linear regression model.

The mean absolute percentage error in this model 33.13. Adding or removing any other variables didn’t make much difference in the error.

2.2.3 Desicison Tree Regression

Now we will try and use a diferent regression model to predict our cnt variable. We will use a regression tree to predict the values of our target variable.

```
# decission tree regression
fit = rpart(cnt ~ season + workingday+ weathersit + temp, data = train_data, method =
"anova")
pr=predict(fit,test_data[,-16])
rpart.plot(fit,box.palette ="RdBu",shadow.col="gray",nn=TRUE)
```



From the above tree, we can say these following statements:

60% of rentals are made when temp greater than 0.43

41% of rentals were made when temp greater than 0.43 and weathersit is 2

18% of rentals were made when temp greater than 0.26 and season is 1,2

14% of rentals are when temp greater than 0.51

13% of rentals were made when temp less than 0.43 and and season is not 1,2

Chapter 3

Conclusion

3.1 Model Evaluation

Now that we have a few models to predict the target variable. We need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of our data, interpretability and computation efficiency, do not hold much significance. Therefore we will use Predictive performance as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the taret variables, and calculating some average error measure.

3.1.1 Mean Absolute Percentage Error (MAPE)

MAPE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section.

```
#linear regression
lin.mod = lm(cnt ~  season + workingday + weathersit + temp, data =train_data)
predictions = predict(lin.mod, test_data[,-16])
mapee = function(y, yhat){
  mean(abs((y - yhat)/y))*100
}
mapee(test_data[,16],predictions)
```

Results:

```
# 33.13643
```

```
#Decision tree
fit = rpart(cnt ~ season + workingday + weathersit + temp, data = train_data, method = "anova")
pr = predict(fit, test_data[,-16])
rpart.plot(fit, box.palette = "RdBu", shadow.col = "gray", nn = TRUE)
mapee(test_data[,16],pr)
```

Results:

```
# 30.24657
```

3.2 Model Selection

We can see that both models perform comparatively on average and therefore we can select either of the two models i.e., linear regression or decision tree without any loss of information.

References

https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Dr. Jiming Wu, Dr. Chongqi Wu, Balarama Rajan, Somak Paul and their lectures at California State University, East Bay