EXPLORATIVE ANALYSIS OF BIKE SHARING DATASET

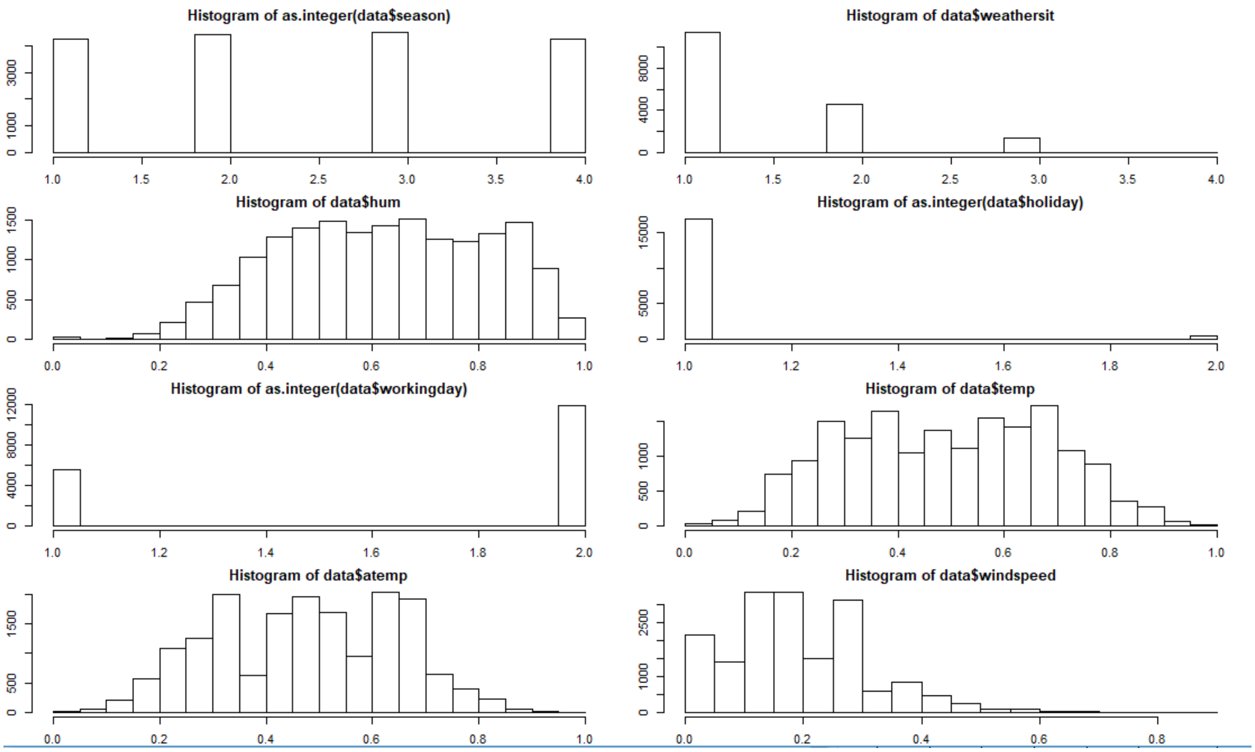
All analysis has been done in R. All the exploratory analysis is done through visualizations which is been described below.

Few inferences can be made from the below histograms-

- Season has four categories of almost equal distribution

```
   Spring    Summer      Fall    Winter 
0.2440877 0.2536970 0.2587030 0.2435123 
```

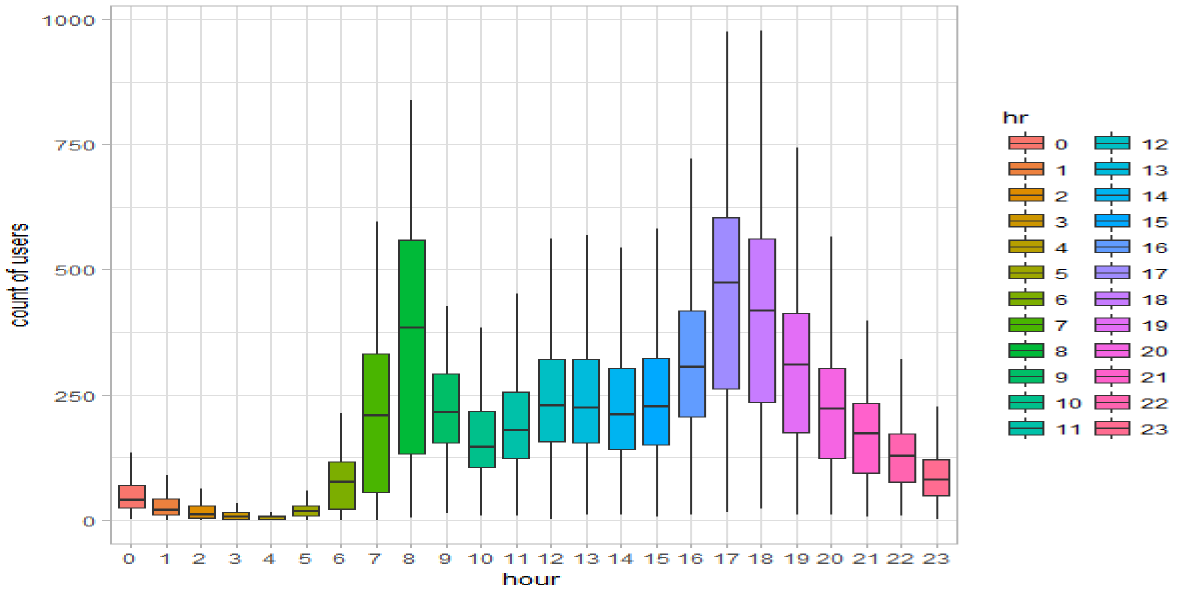
- Weathersit 1 has higher contribution that signifies clear weather
- Having a holiday signifies more contribution
- Variables temp, atemp, humidity and windspeed looks naturally distributed

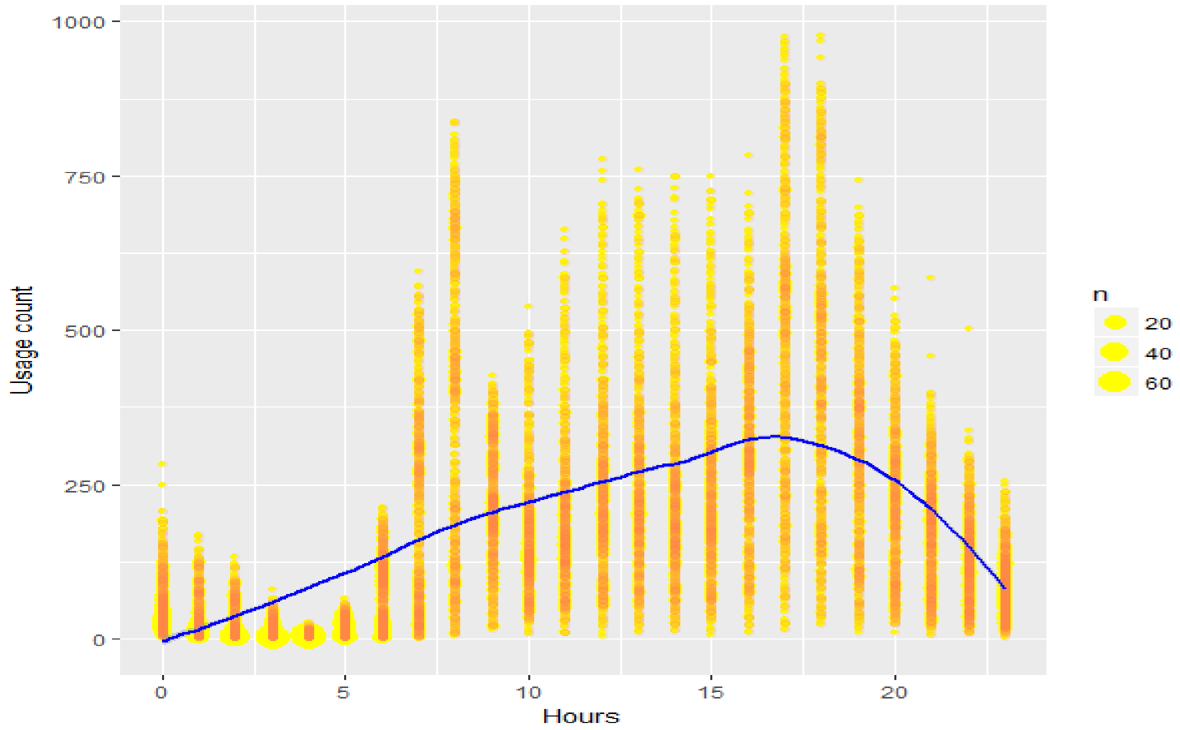


Below we can see the trend of bike demand over hours.

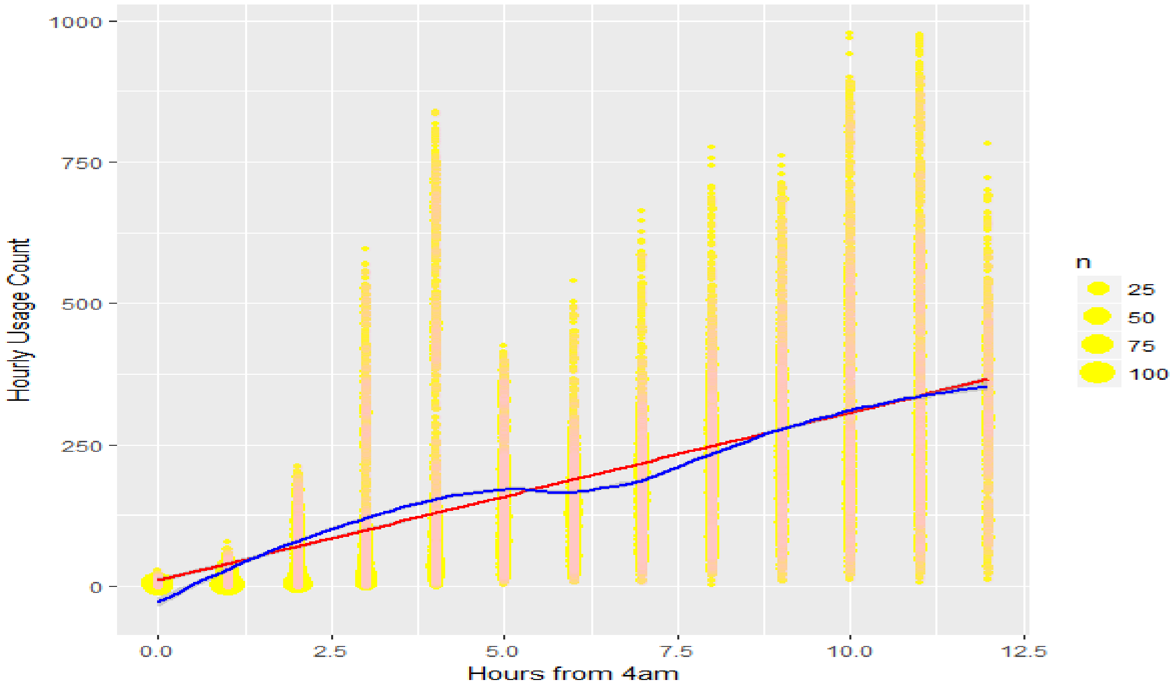
- High : 7-9 and 17-19 hours
- Average : 10-16 hours

- Low : 0-6 and 20-24 hour

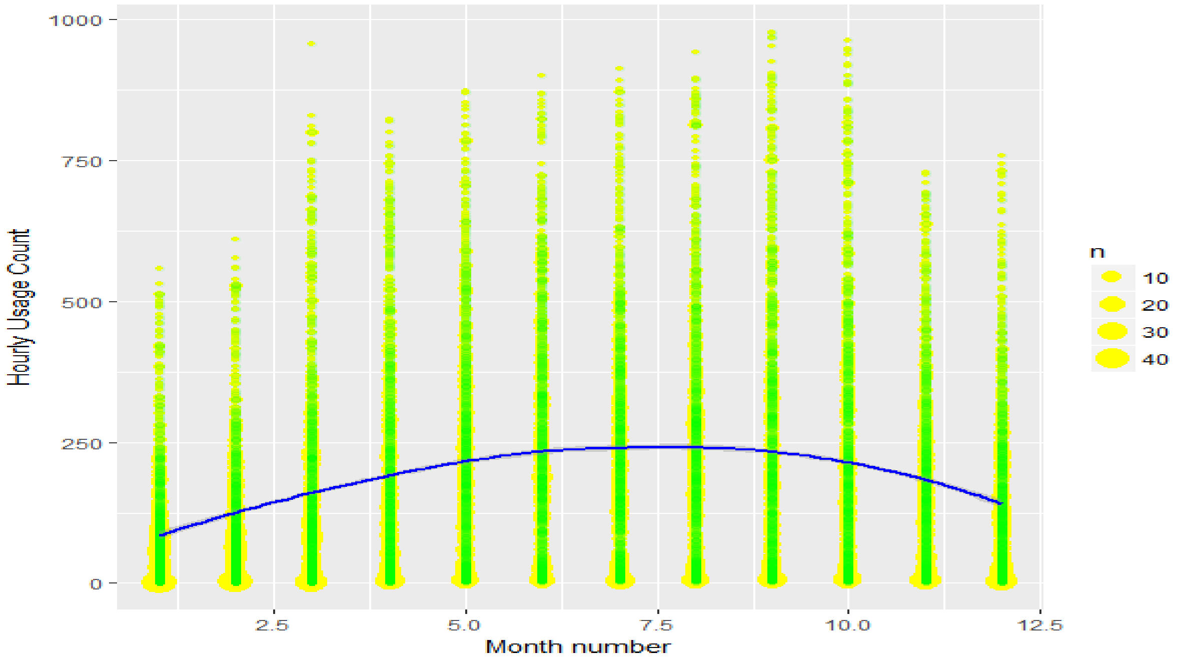


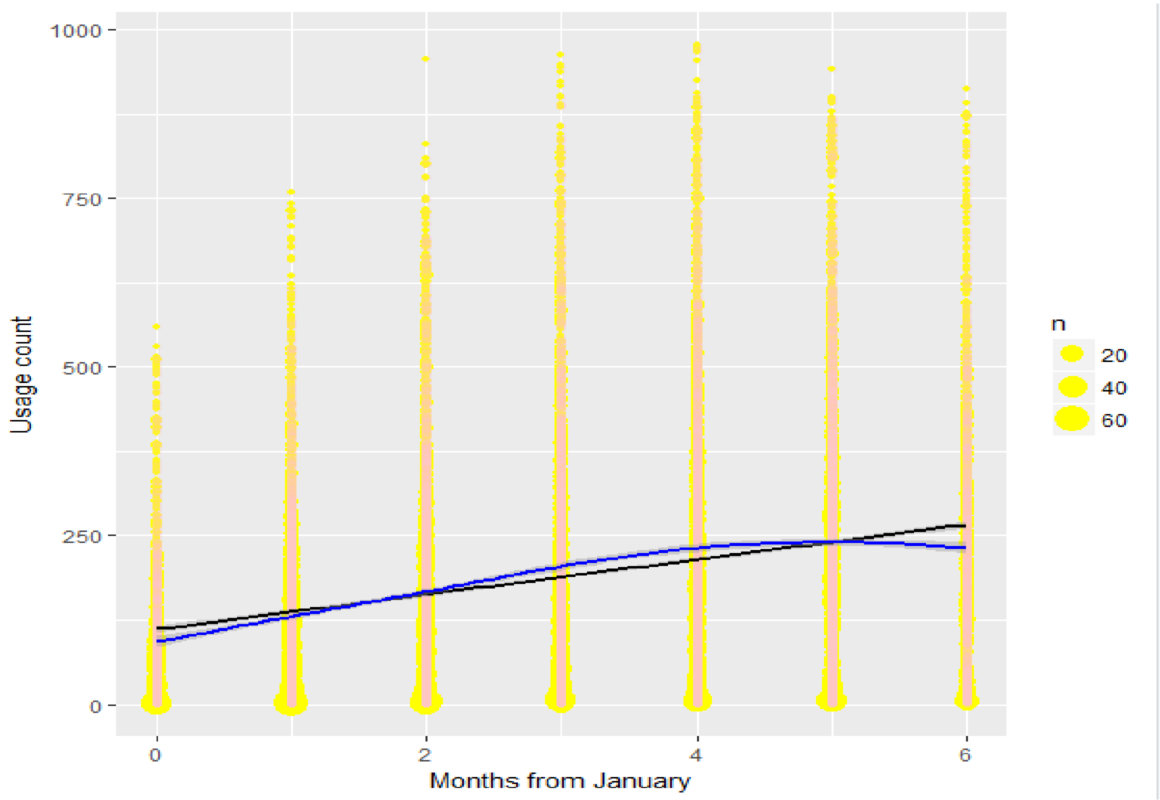


The fit in above graph is far from linear, we can change this to represent the usage rate based on the temporal distance to 4 am

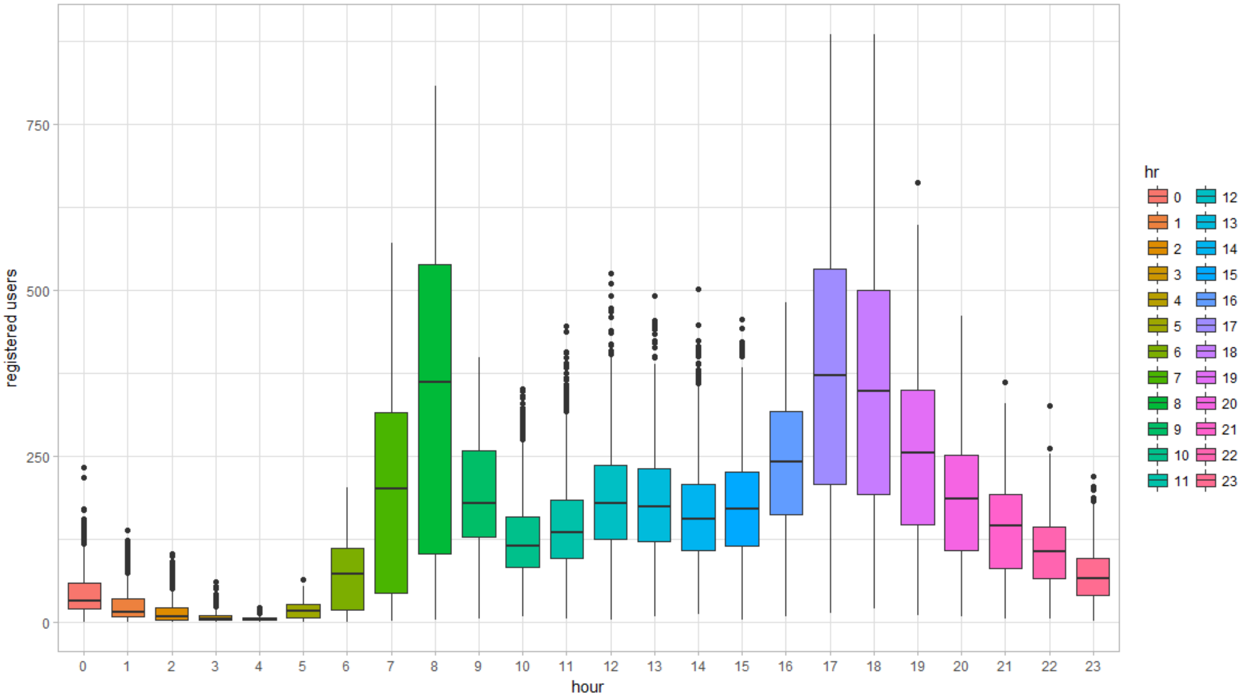


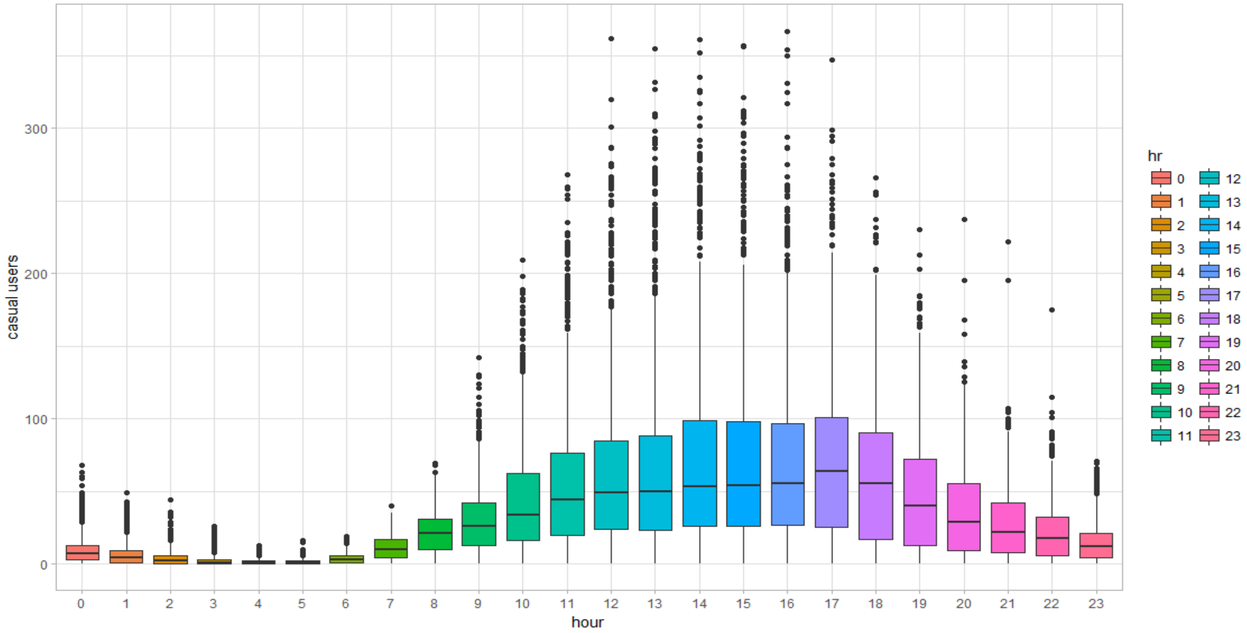
With some manipulation similar to the previous plot, this data can also used to represent usage based on the temporal distance to the month of January. This correlation, however, is not as strong as that for the manipulated time graph.



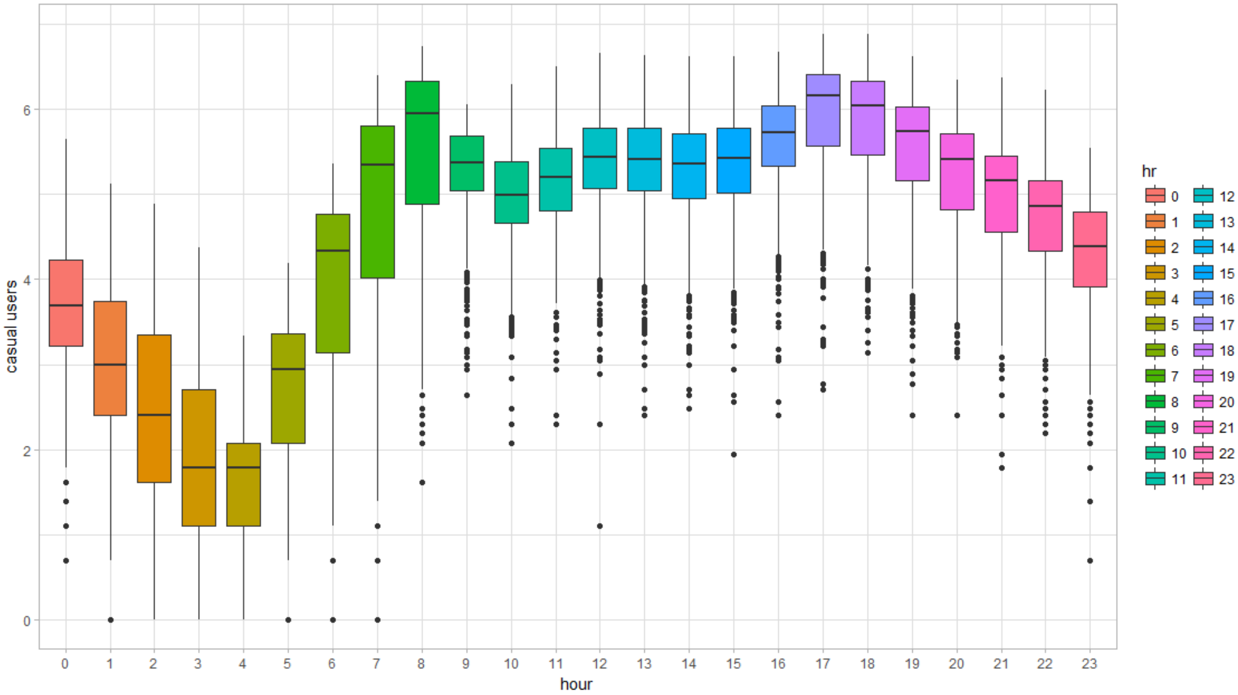


Now looking at the distribution of casual and registered users separately with respect to count variable. Below we can see that registered users have similar trend as count. Whereas, casual users have different trend. Thus, we can say that 'hour' is significant variable.

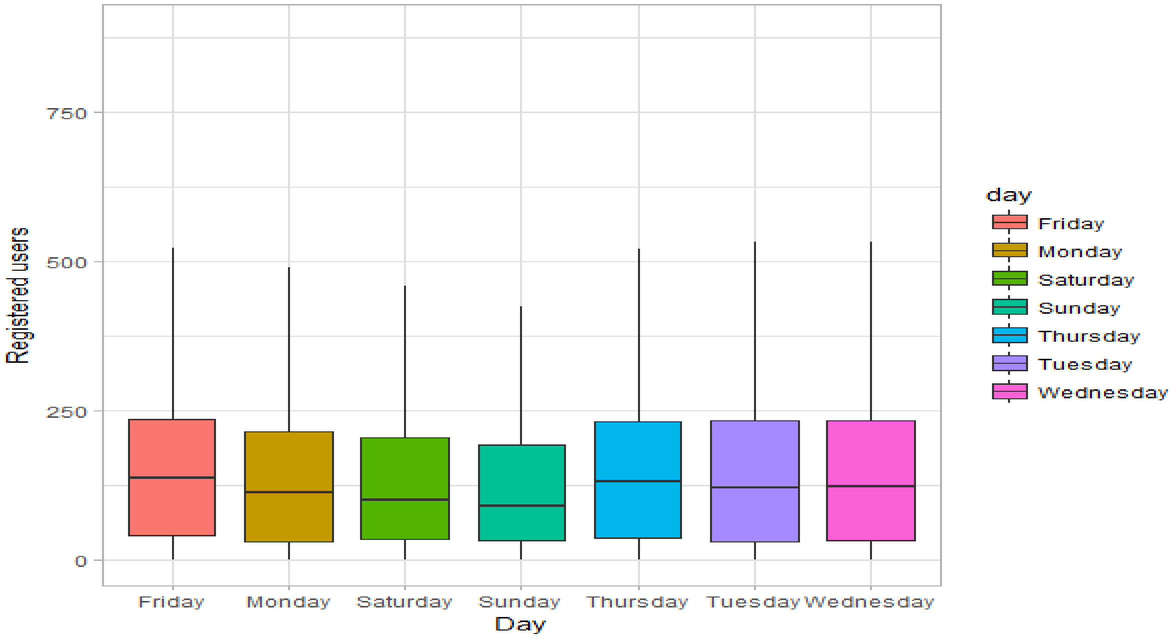


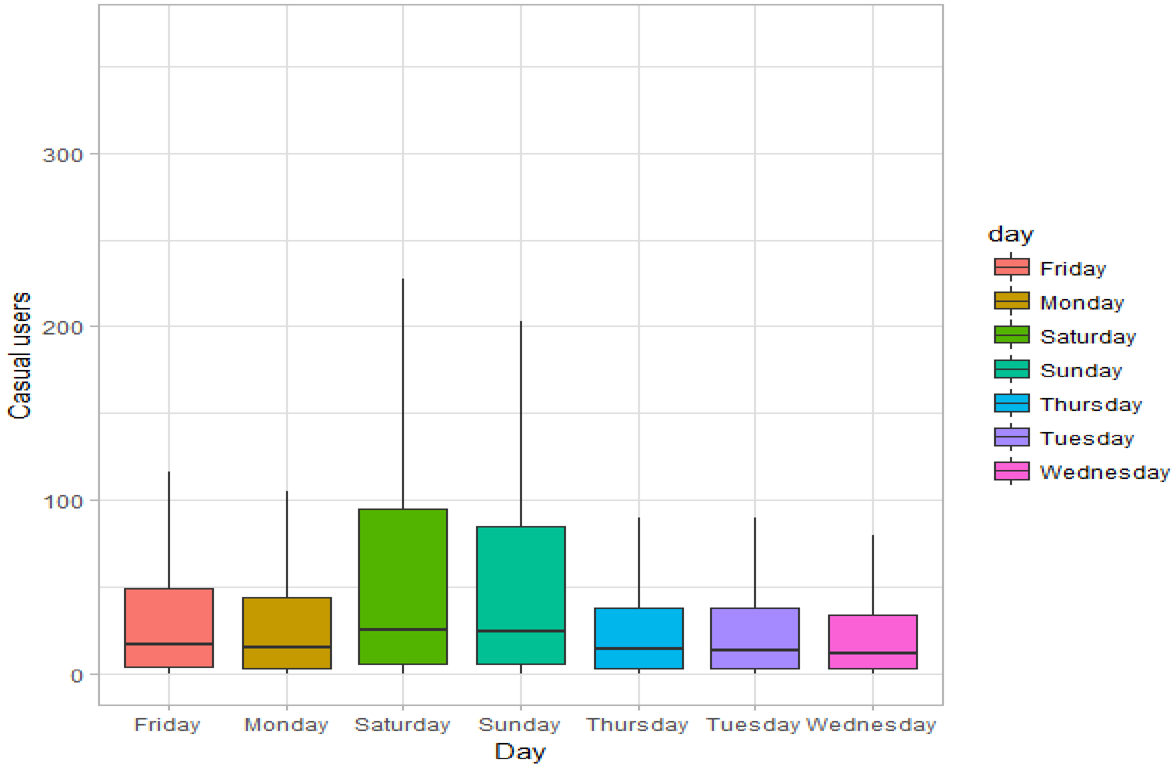


To treat such outliers, we will use logarithm transformation. Let's look at the similar plot after log transformation.

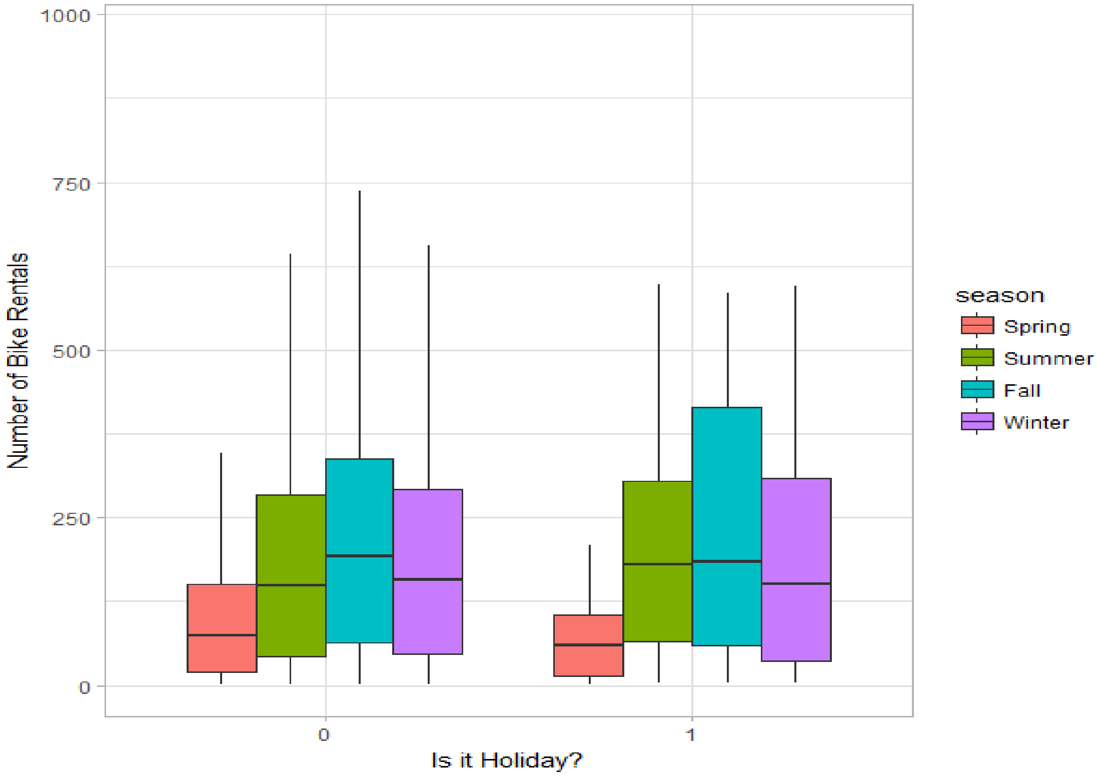


Plot below shows registered and casual users' demand over days. Observed demand of casual users increases over weekend.

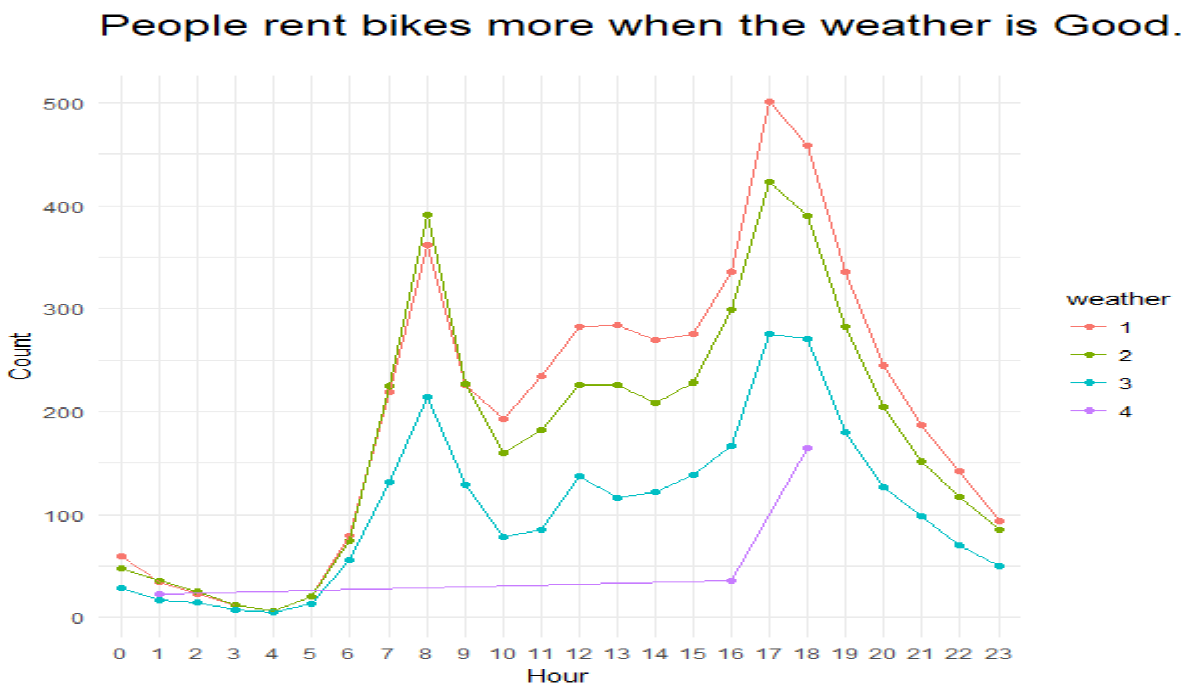




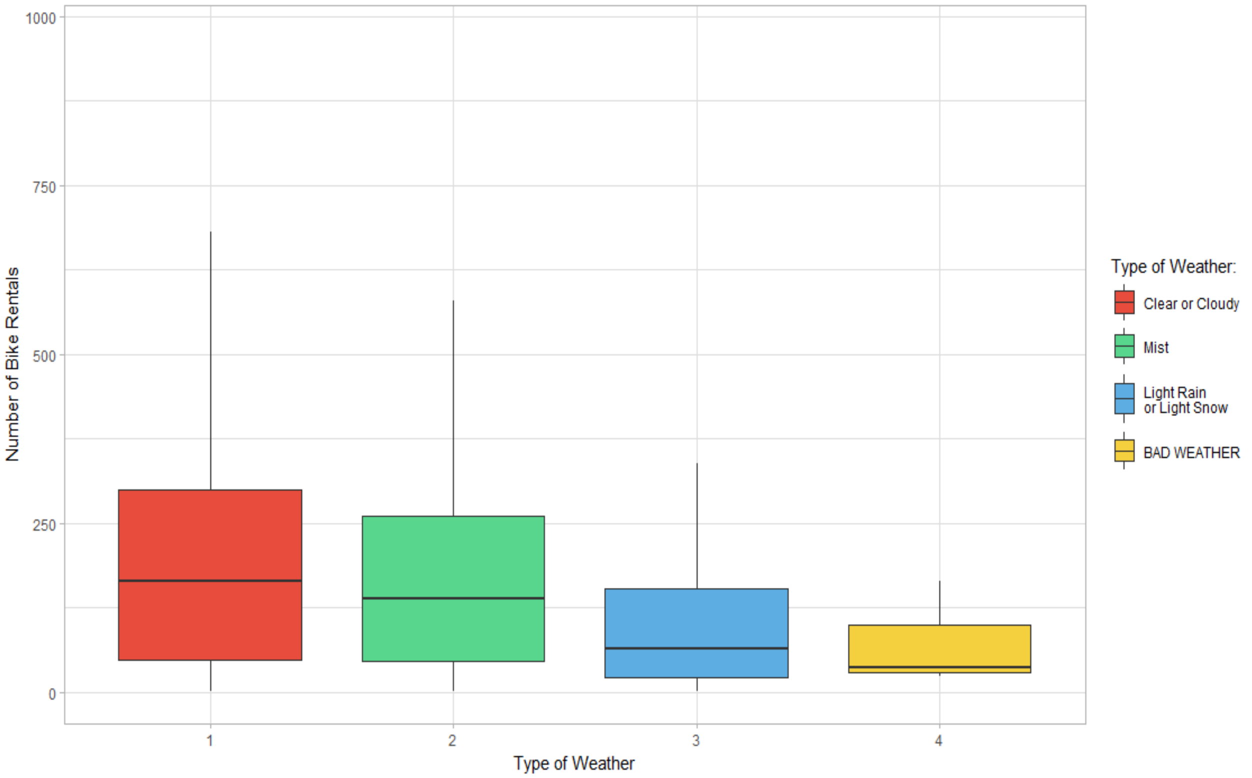
It can be observed that having a holiday has little impact on count of bikes rented.

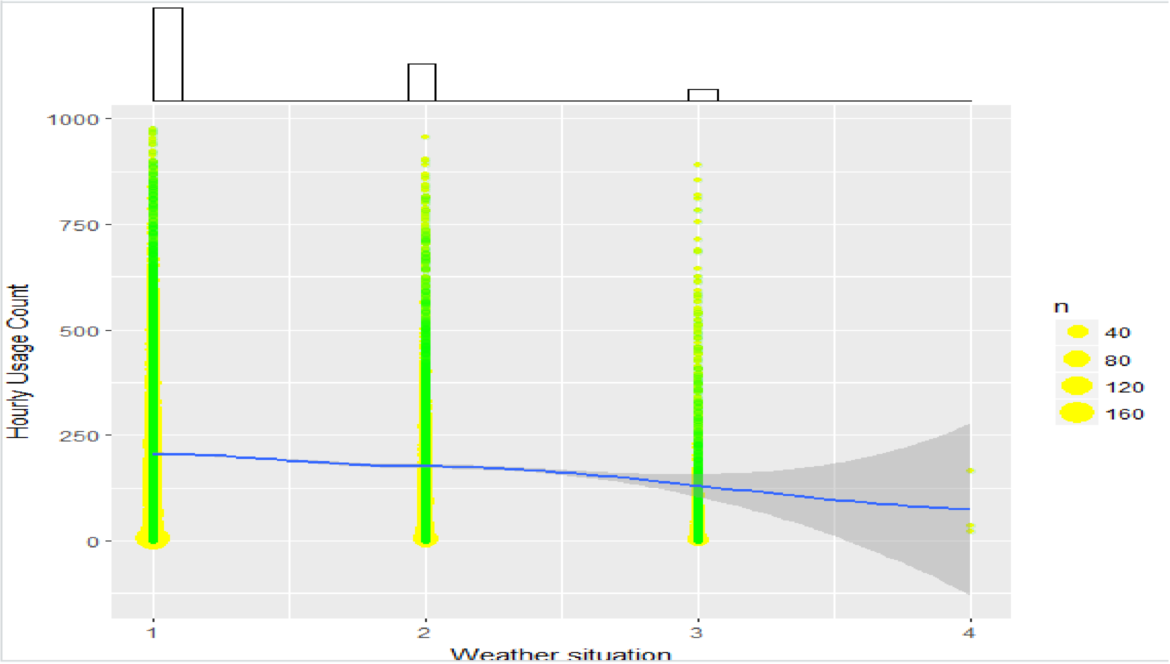


Also , observed bike rentals increases when weather is clear and good.

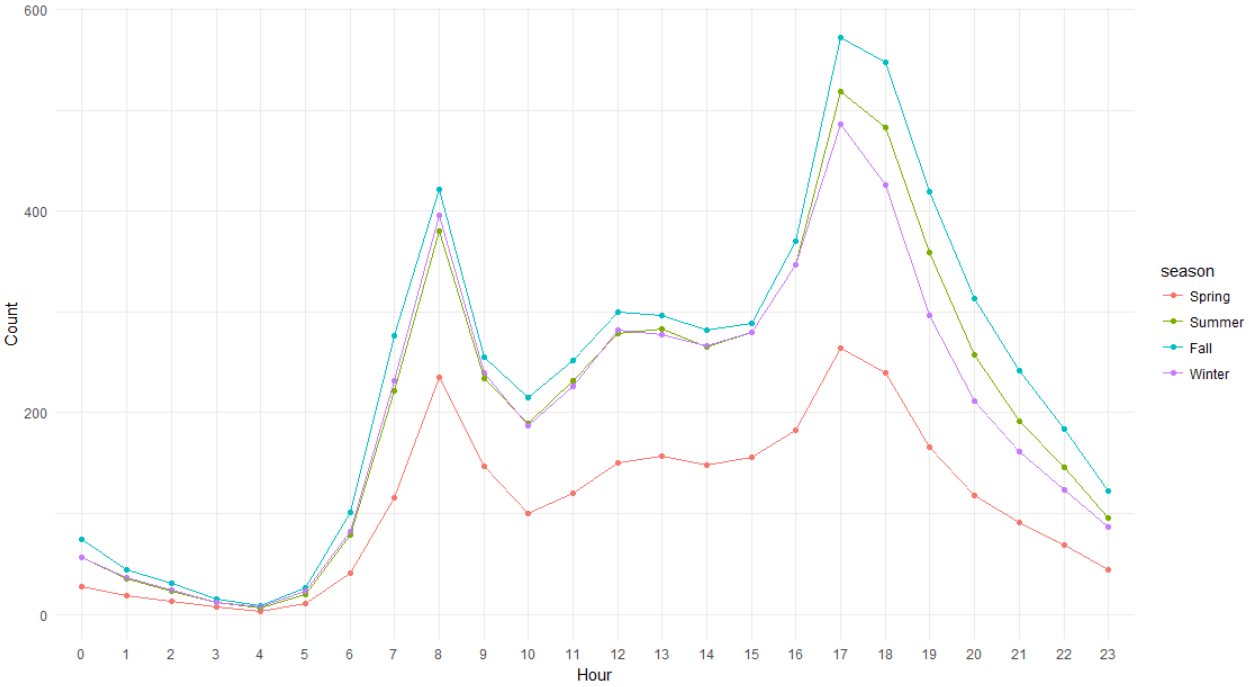


People rent bikes more when the weather is Good.

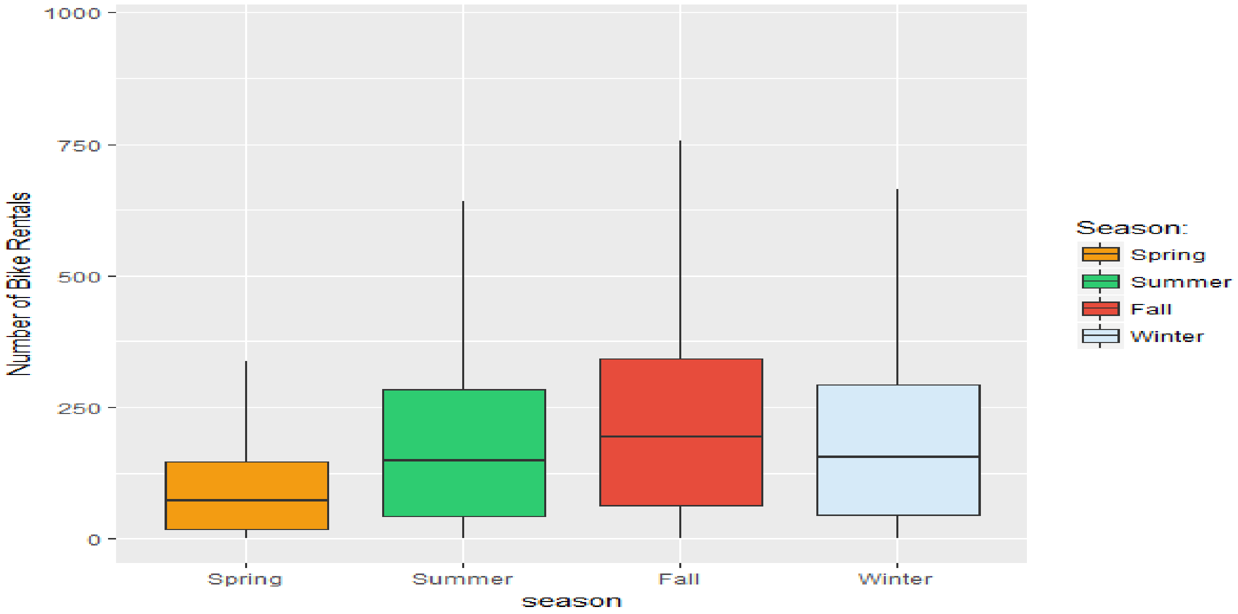




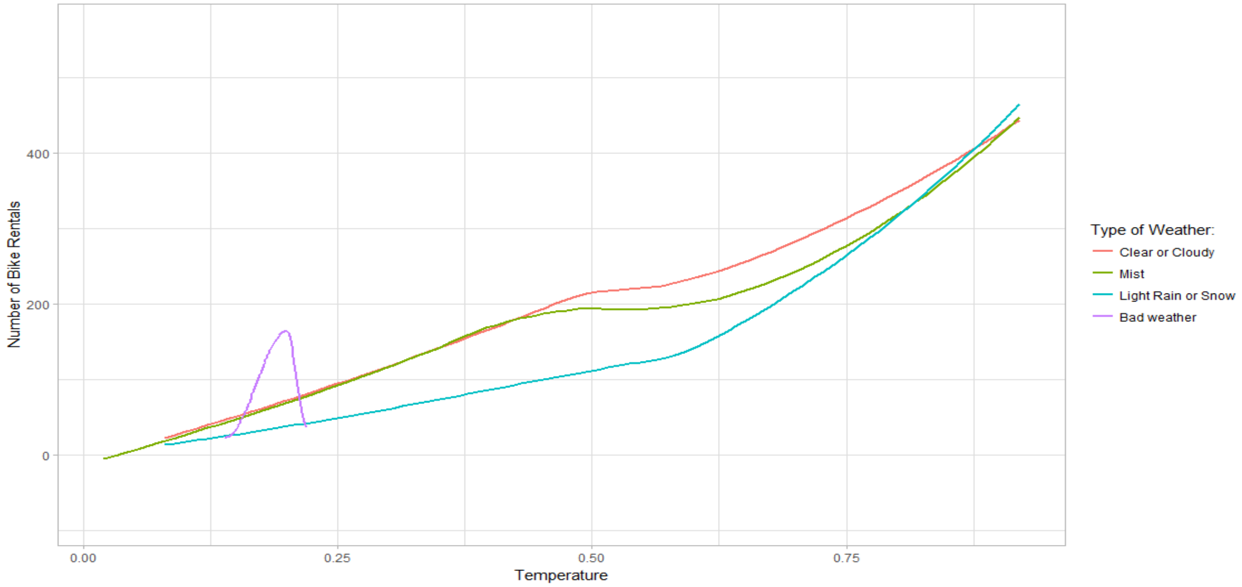
People rent bikes more in Fall, and much less in Spring.



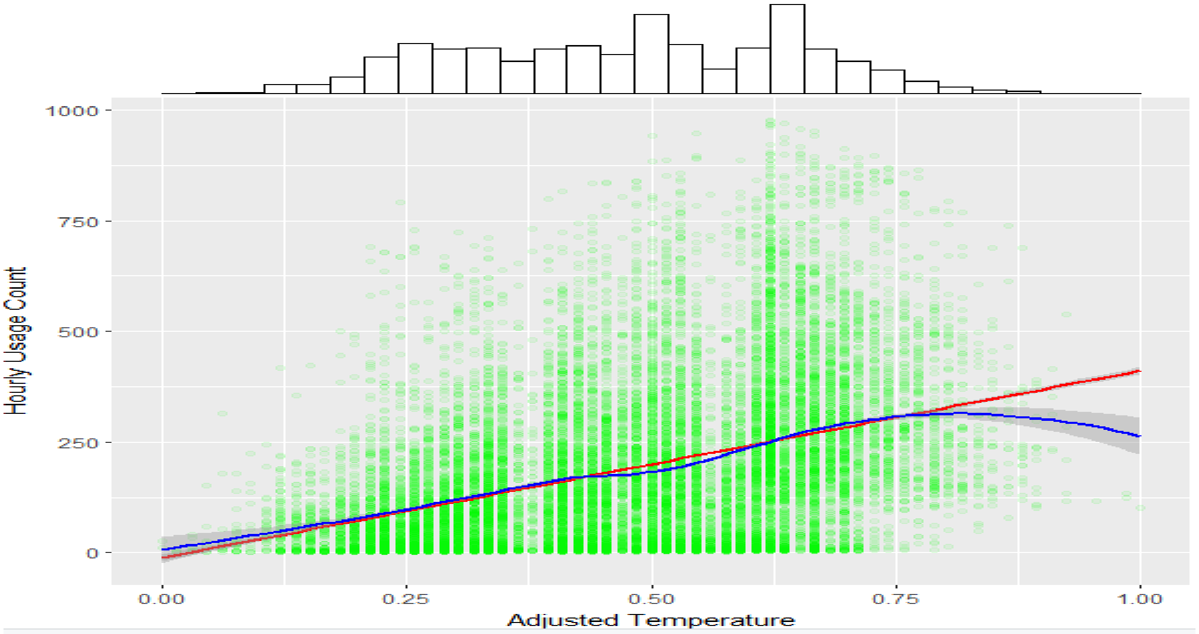
The boxplot of different seasons against bike rental count reveals that there is a seasonal trend with the rental count. Season can be one of the determining factors that affects bike rental count

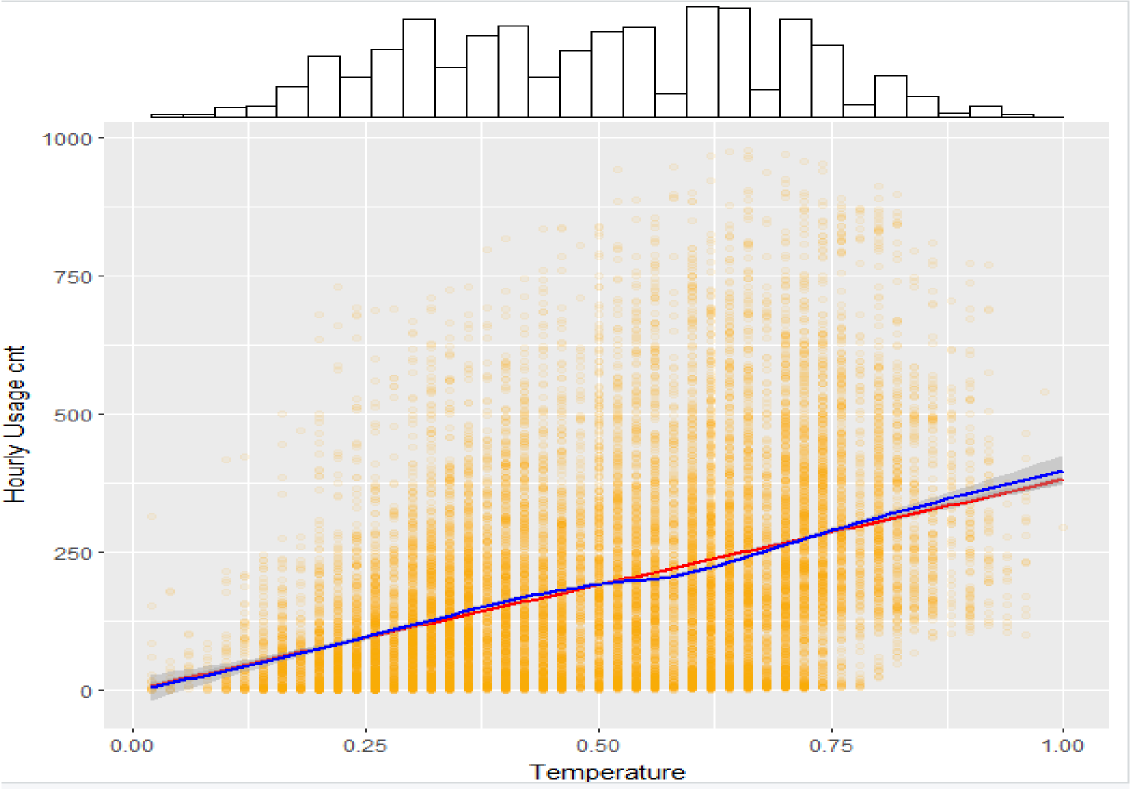


The temperature plot shows that generally, the warmer the temperature, the higher bike rental demand.

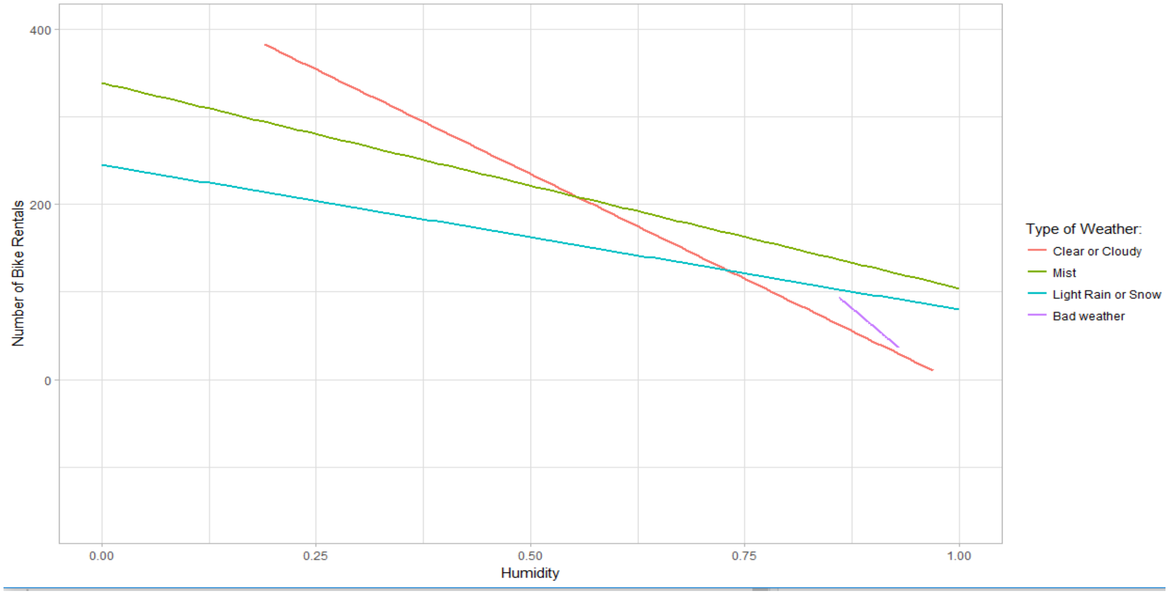


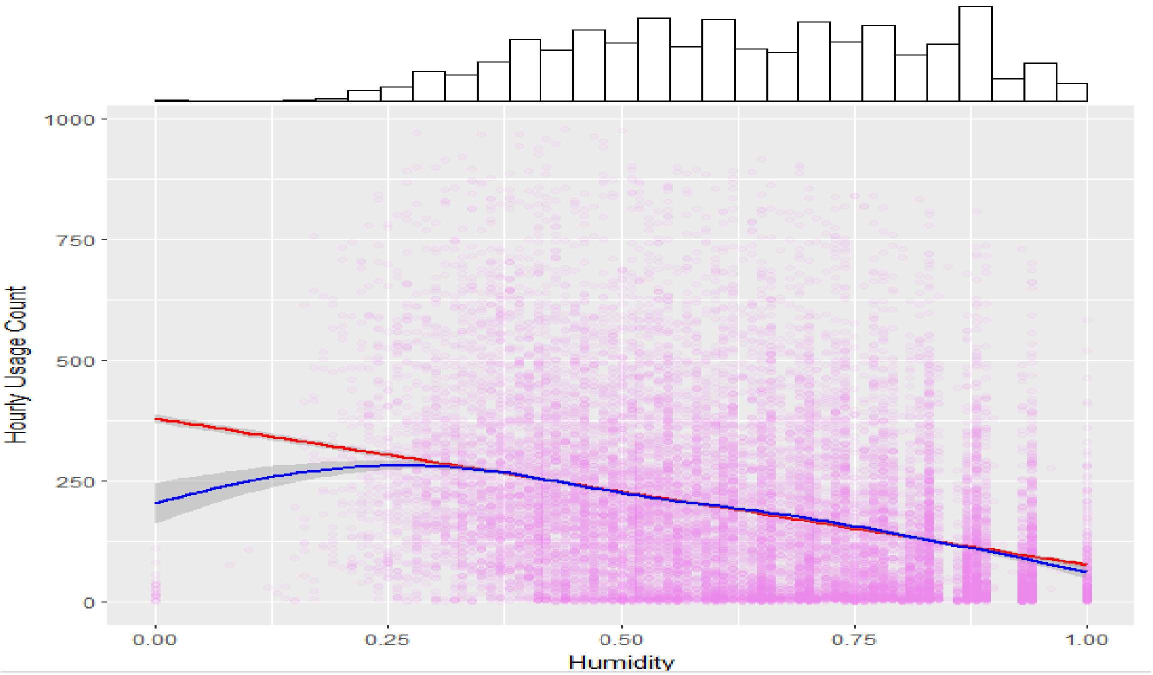
As seen from the scatter plots below, there is a positive correlation between both temperature-to-usage and adjusted-temperature-to-usage for most of the temperature range, and a linear fit isn't far from the best-fit curve. This should intuitively make sense, as people are not likely to bike outside in cold weather. For the maximum temperatures, which seem to be a small subset of the data, there is a dip in this curve. Once again, this should make sense as users may also be discouraged to bike when it’s too hot outside.



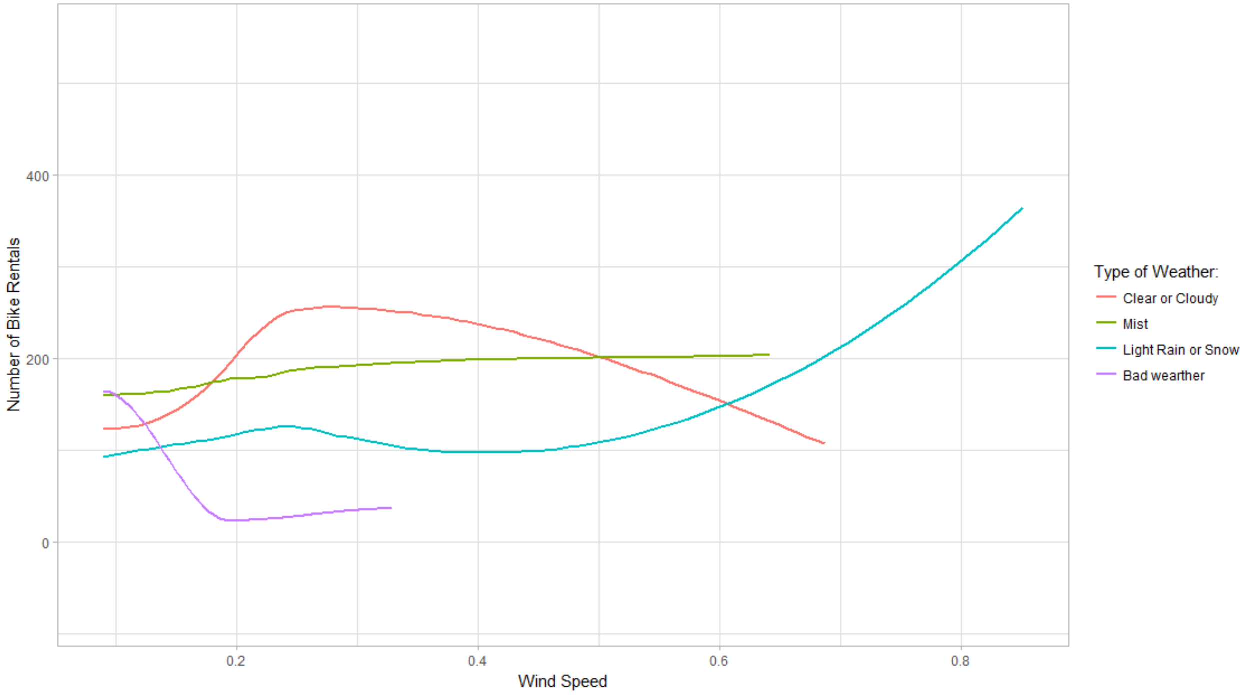


The humidity plot shows that generally, the higher the relative humidity, the lower bike rental demand

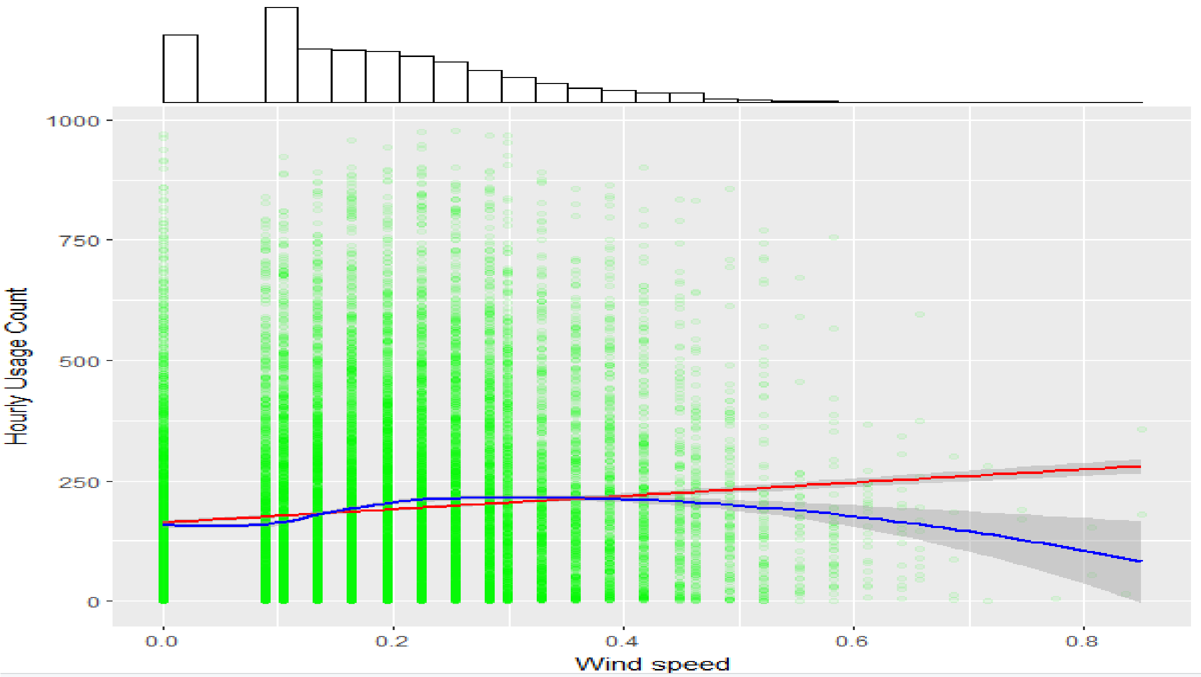




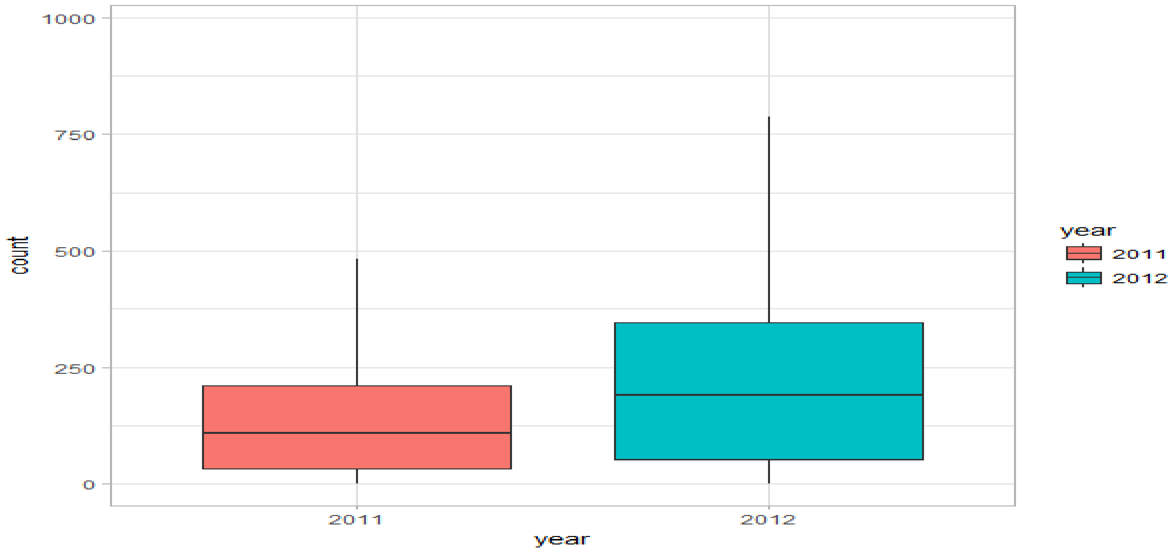
The wind speed plot shows that although people enjoy gentle breeze in good weathers, the bike rental demand is significantly lower no matter the wind speed in light rain or snow weathers.



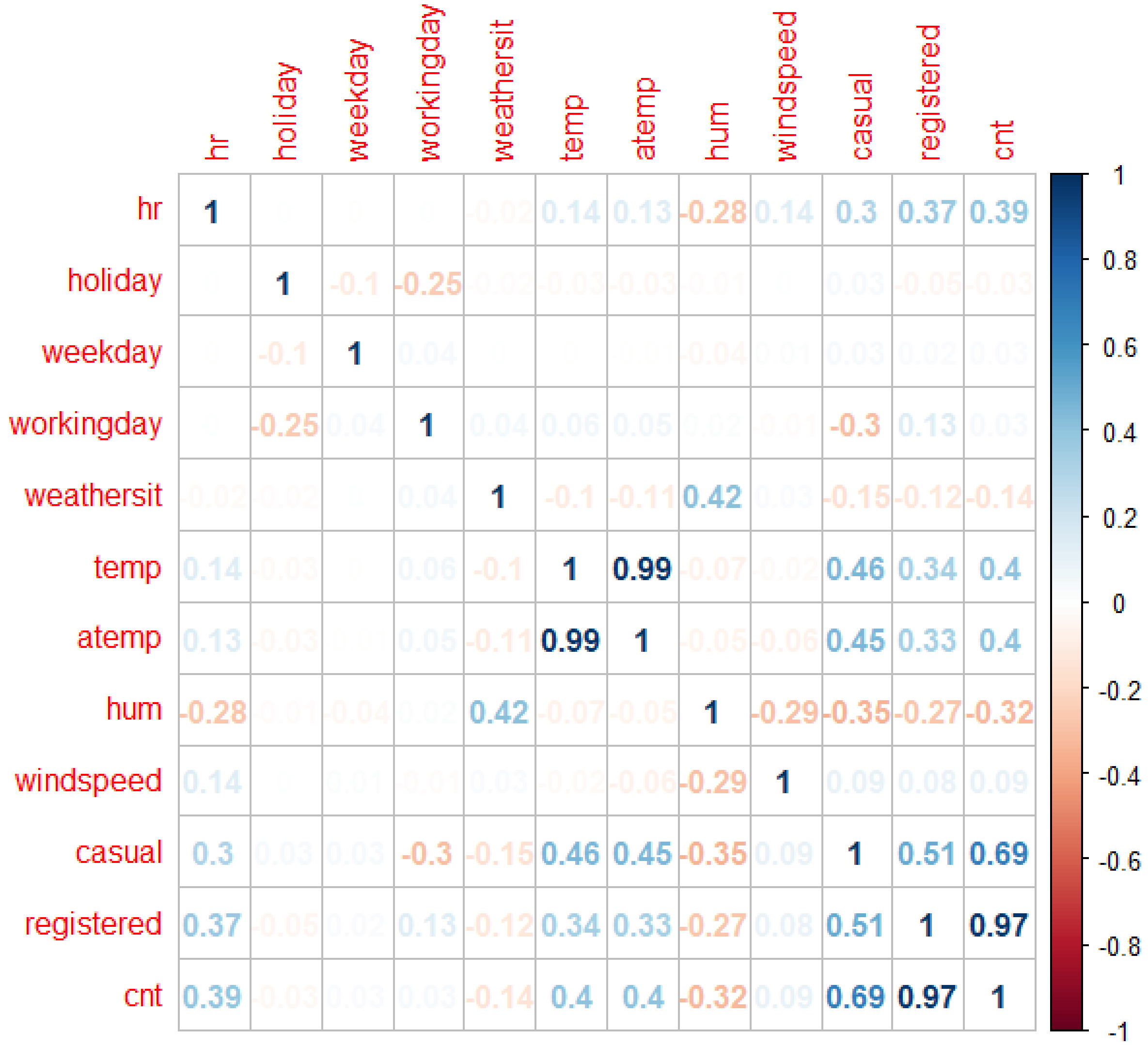
Looking at the wind speed data, however, doesn’t give us a clear interpretation of how it affects usage. The correlation between the two factors is weak at best.



Extracting year of each observation from the datetime column and see the trend of bike demand over year. You can see that 2012 has higher bike demand as compared to 2011.



Correlation Matrix -



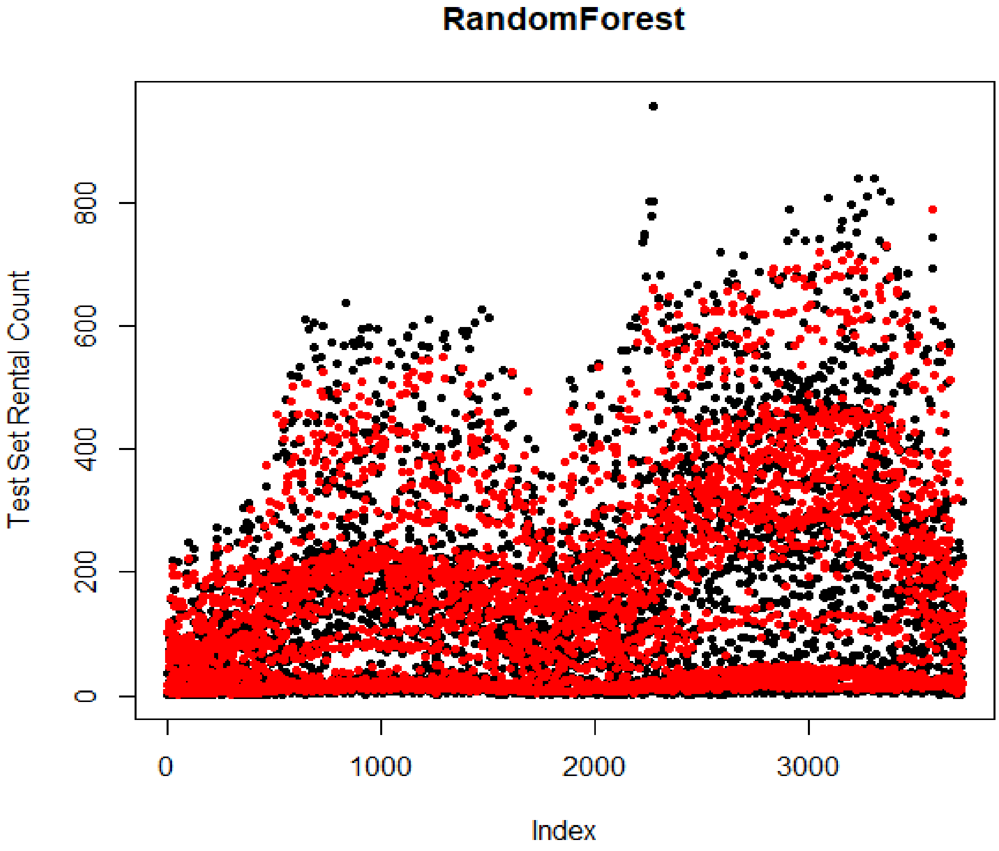
	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
hr	1	0	0	0	-0.02	0.14	0.13	-0.28	0.14	0.3	0.37	0.39
holiday	0	1	-0.1	-0.25	-0.02	-0.03	-0.03	-0.01	0	0.03	-0.05	-0.03
weekday	0	-0.1	1	0.04	0	0	-0.01	-0.04	0.01	0.03	0.02	0.03
workingday	0	-0.25	0.04	1	0.04	0.06	0.05	0.02	-0.01	-0.3	0.13	0.03
weathersit	-0.02	-0.02	0	0.04	1	-0.1	-0.11	0.42	0.03	-0.15	-0.12	-0.14
temp	0.14	-0.03	0	0.06	-0.1	1	0.99	-0.07	-0.02	0.46	0.34	0.4
atemp	0.13	-0.03	-0.01	0.05	-0.11	0.99	1	-0.05	-0.06	0.45	0.33	0.4
hum	-0.28	-0.01	-0.04	0.02	0.42	-0.07	-0.05	1	-0.29	-0.35	-0.27	-0.32
windspeed	0.14	0	0.01	-0.01	0.03	-0.02	-0.06	-0.29	1	0.09	0.08	0.09
casual	0.3	0.03	0.03	-0.3	-0.15	0.46	0.45	-0.35	0.09	1	0.51	0.69
registered	0.37	-0.05	0.02	0.13	-0.12	0.34	0.33	-0.27	0.08	0.51	1	0.97
cnt	0.39	-0.03	0.03	0.03	-0.14	0.4	0.4	-0.32	0.09	0.69	0.97	1

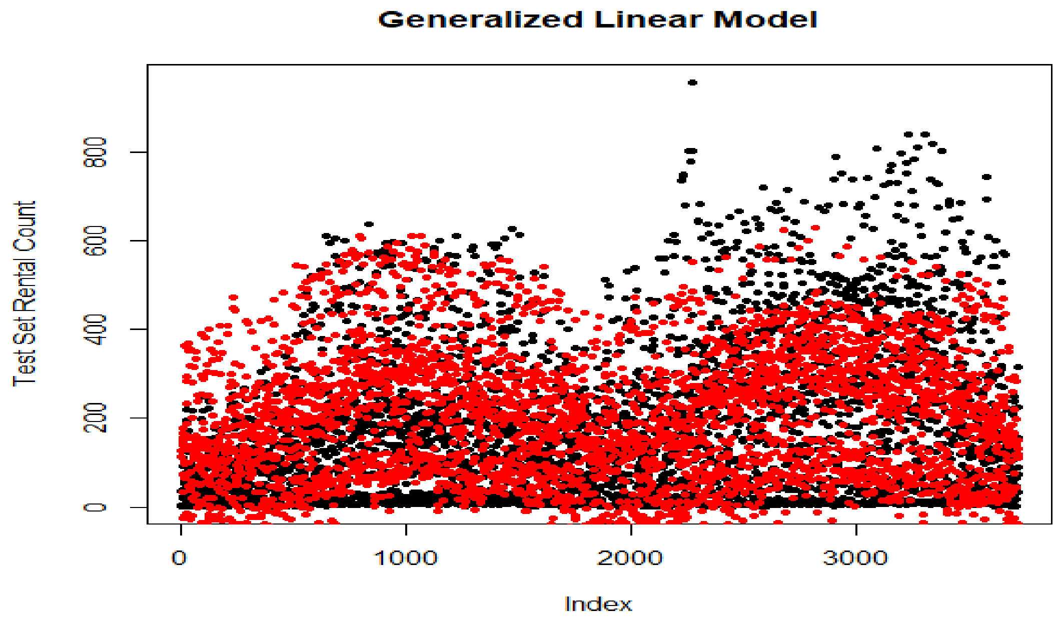
Few inferences which can be taken out from above table-

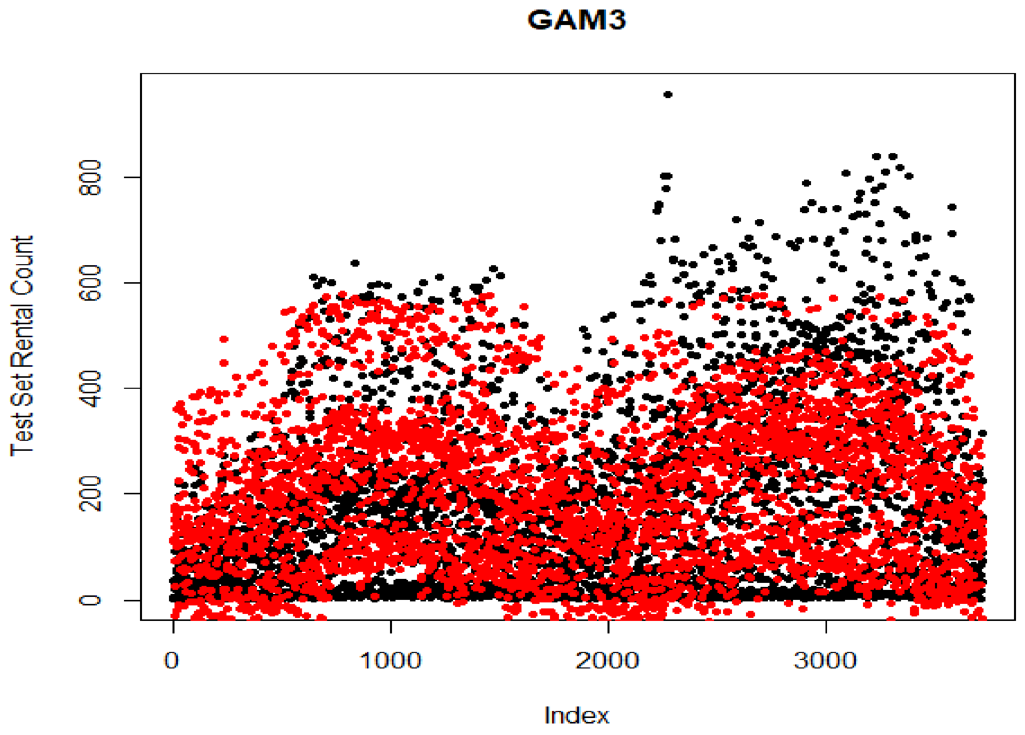
- Cnt variable is highly correlated to registered variable
- Other than registered, cnt variable is positively correlated with casual, temp ,hr and atemp variables
- Temp and atemp variables are strongly correlated

Model Building

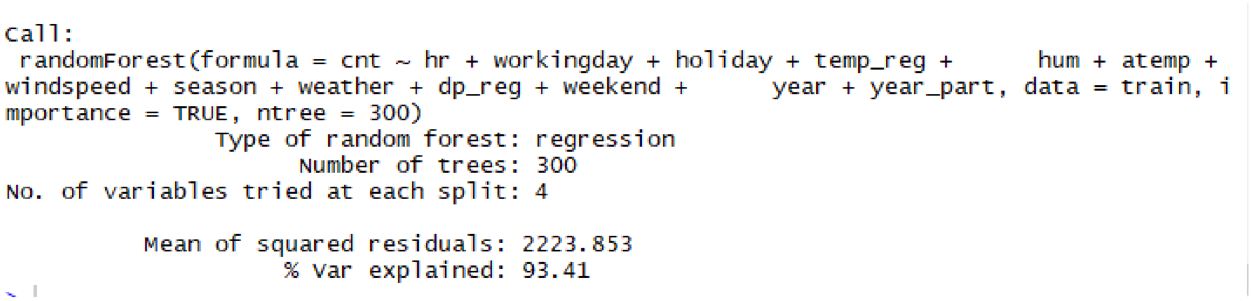
Out of all models, Random Forest is performing the best. Blow are the scatter plots of some models implemented. The below plots clearly show why Random Forest is chosen.





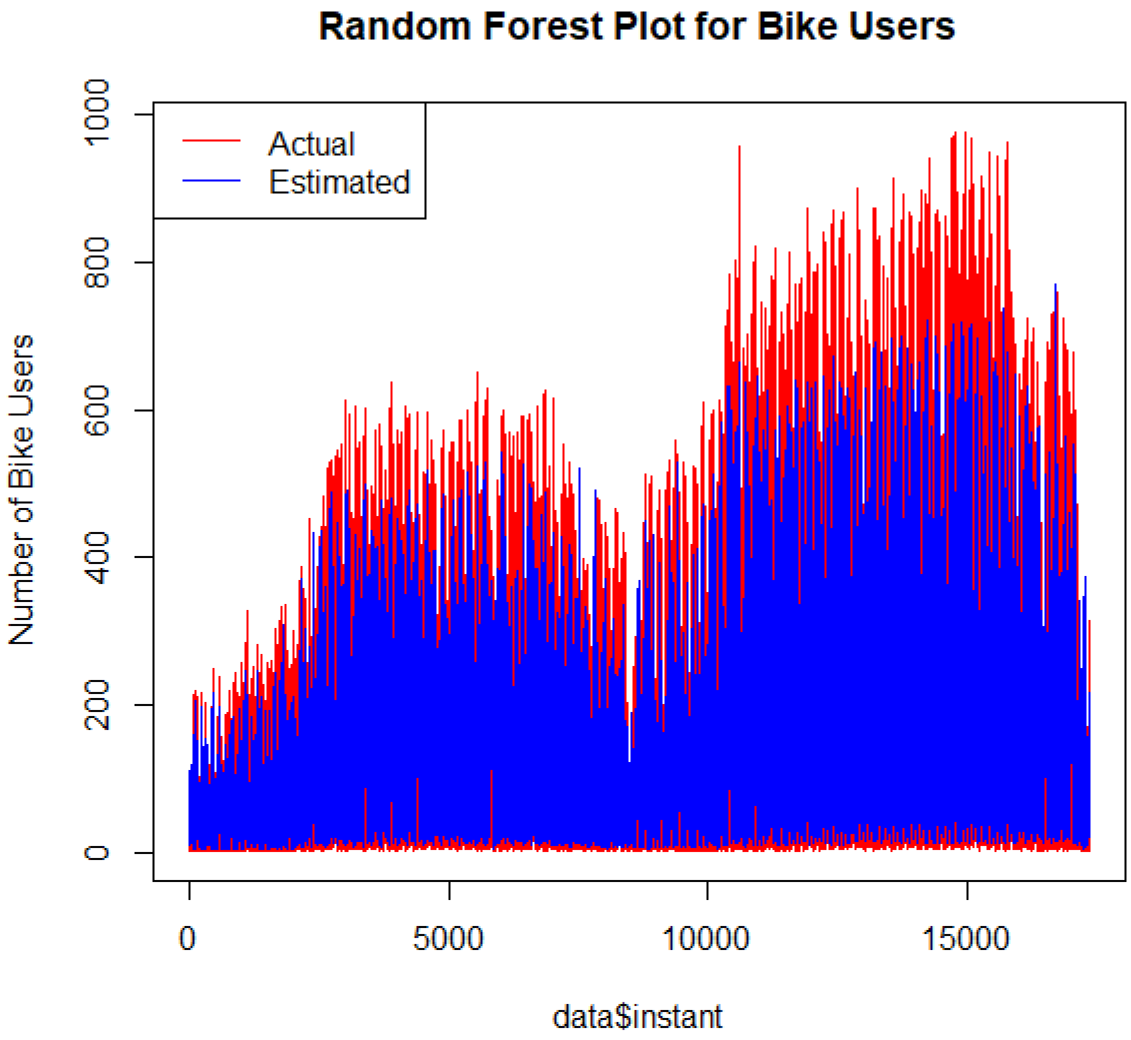


Random Forest Summary -



```
Call:
 randomForest(formula = cnt ~ hr + workingday + holiday + temp_reg +      hum + atemp +
windspeed + season + weather + dp_reg + weekend +      year + year_part, data = train, i
mportance = TRUE, ntree = 300)
               Type of random forest: regression
                     Number of trees: 300
No. of variables tried at each split: 4

          Mean of squared residuals: 2223.853
                    % Var explained: 93.41
```



2.

That means it should be integrated with some web service that takes in some parameters and gives the predictions right away. Also, before this , the model should be retrained to capture all the samples and deviations in the data to give the best results.

Real time streaming analysis can also be used. For Example -Running as as a service on Spark Cluster.