Project Report

Bike Renting

NITISH ROHILLA

5TH April 2019

Contents

1. Introduction

1.1 Problem Statement 3

1.2 Data 3

2. Methodologies

2.1 Pre-processing 3

2.1.1 Missing Value Analysis 7

2.1.2 Outlier Analysis 8

2.1.3 Feature Selection 9

2.2. Modelling 11

2.2.1 Decision Tree. 11

2.2.2 Random Forest 11

2.2.3 Linear Regression 11

3. Conclusion

3.1 Model Evaluation 13

3.2 Model Selection 13

3.3 Answers of asked questions 13

1.1 PROBLEM STATMENT

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

1.2 DATA

Our task is to build regression models which will predict the count of the bikes rented based on various factors. Given below is the sample of dataset using to predict the cnt

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

1.instant: Record index
2.dteday: Date
3.season: Season (1:springer, 2:summer, 3:fall, 4:winter)
4.yr: Year (0: 2011, 1:2012)
5.mnth: Month (1 to 12
6.holiday: weather day is holiday or not (extracted fromHoliday Schedule)
7.weekday: Day of the week
8 workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
9.weathersit: (extracted fromFreemeteo)
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 Pallets + Thunderstorm + Mist, Snow + Fog
10.temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=- t_max=+39 (only in hourly scale)
11.atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max- t_min), t_min=-16, t_max=+50 (only in hourly scale)
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

2.Methodologies

Hypothesis creating- g Before exploring data, I spend some of the time with the data to understand the relationship between the variables to understand the domain knowledge and gaining experience of problem

I created some of the hypothesis

- Registered demand users demand more bikes than casual users
- Traffic can be related with bike demand

- Due to rains the bike rental count might get lower
- Temp have -ve correlation with count

2.1 Pre-processing

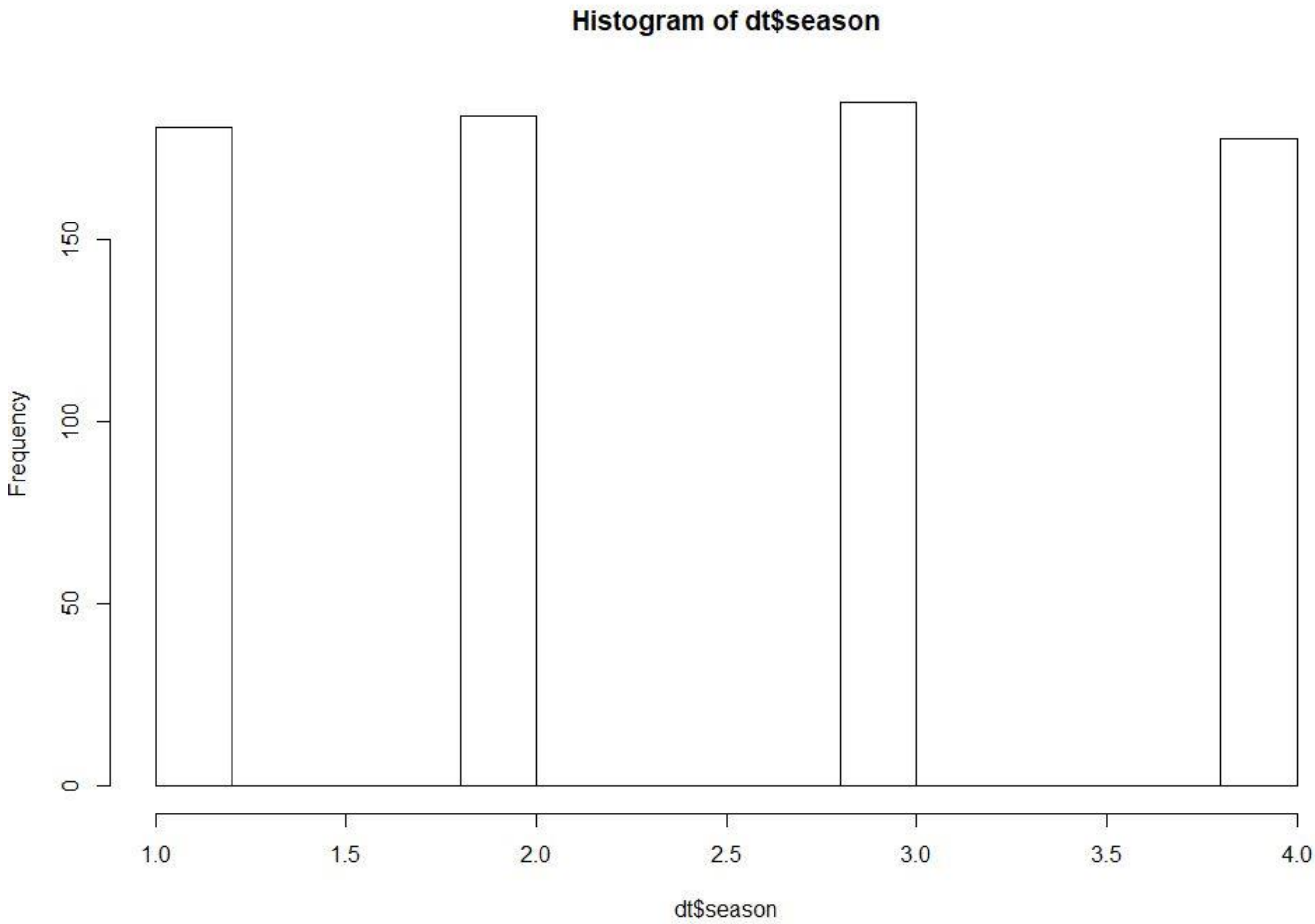
Exploratory Data Analysis is analysing the data sets to extract their characterstics. It is much more than just looking at the data but also to analyse, clean and to visualize through graphs and plots .

As we first look into the unique value consisted in each variable of the data as below

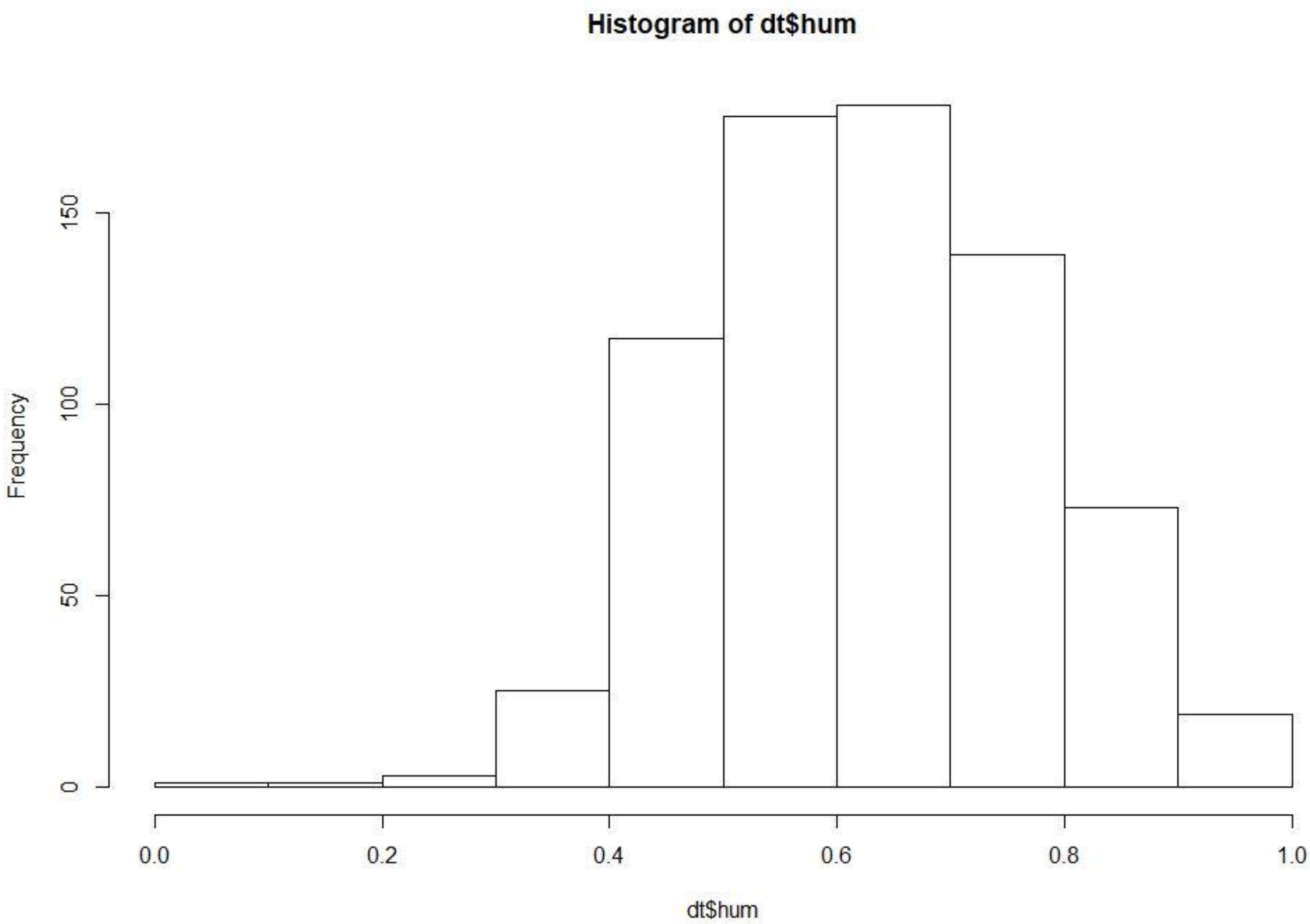
```
dteday          731
season            4
yr                2
mnth             12
holiday           2
weekday           7
workingday        2
weathersit        3
temp            499
atemp           690
hum             595
windspeed       650
casual          606
registered      679
cnt             696
dtype: int64
```

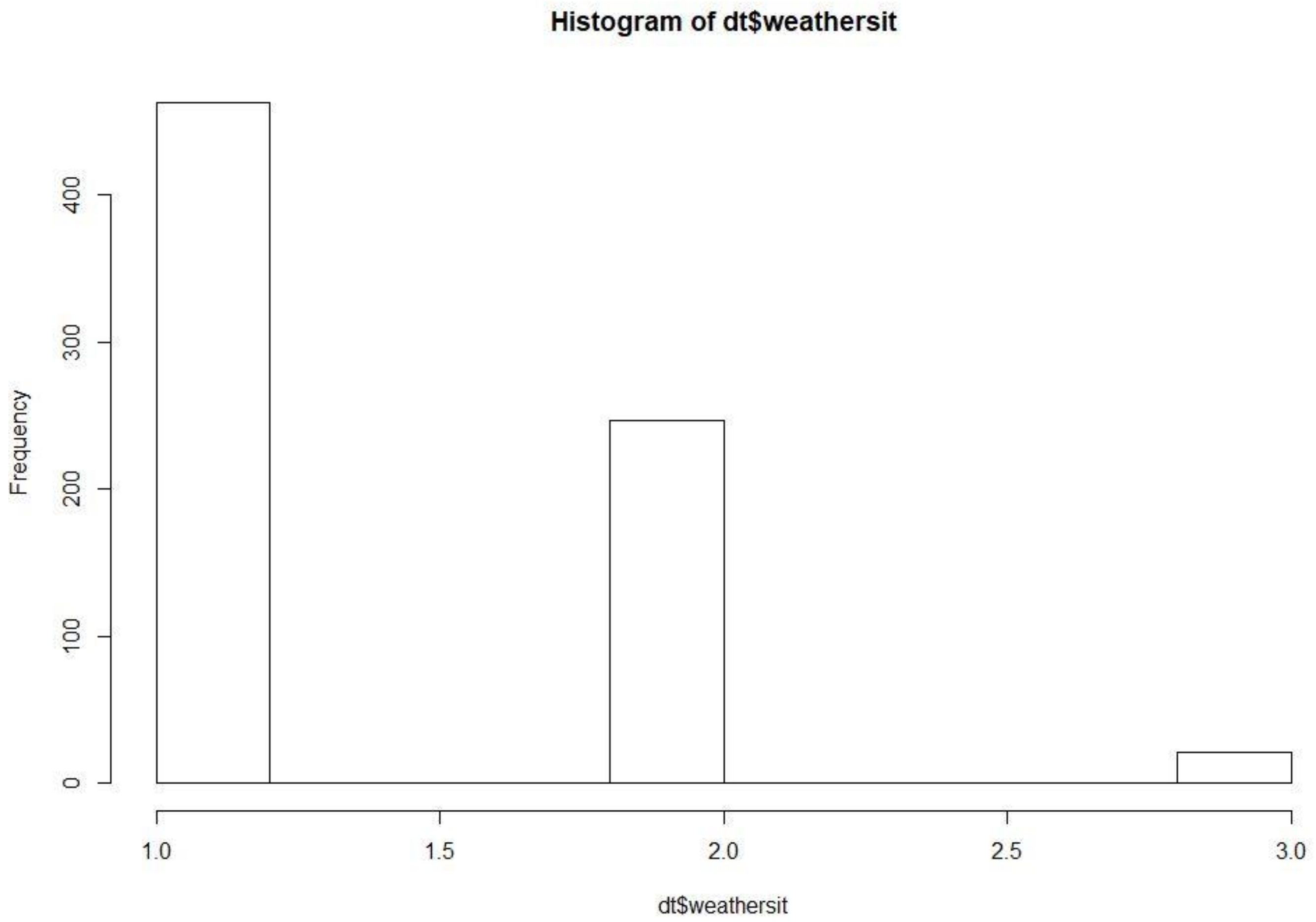
While looking at the table it was very much clear that there were two kind of variables present in the data set. Categorical and Numerical. Out of all variables there were 3 dependent variable Casual, dependent and cnt. It is these we need to predict on the basis of other dependent variables.

Now analysing some relevant variables visually

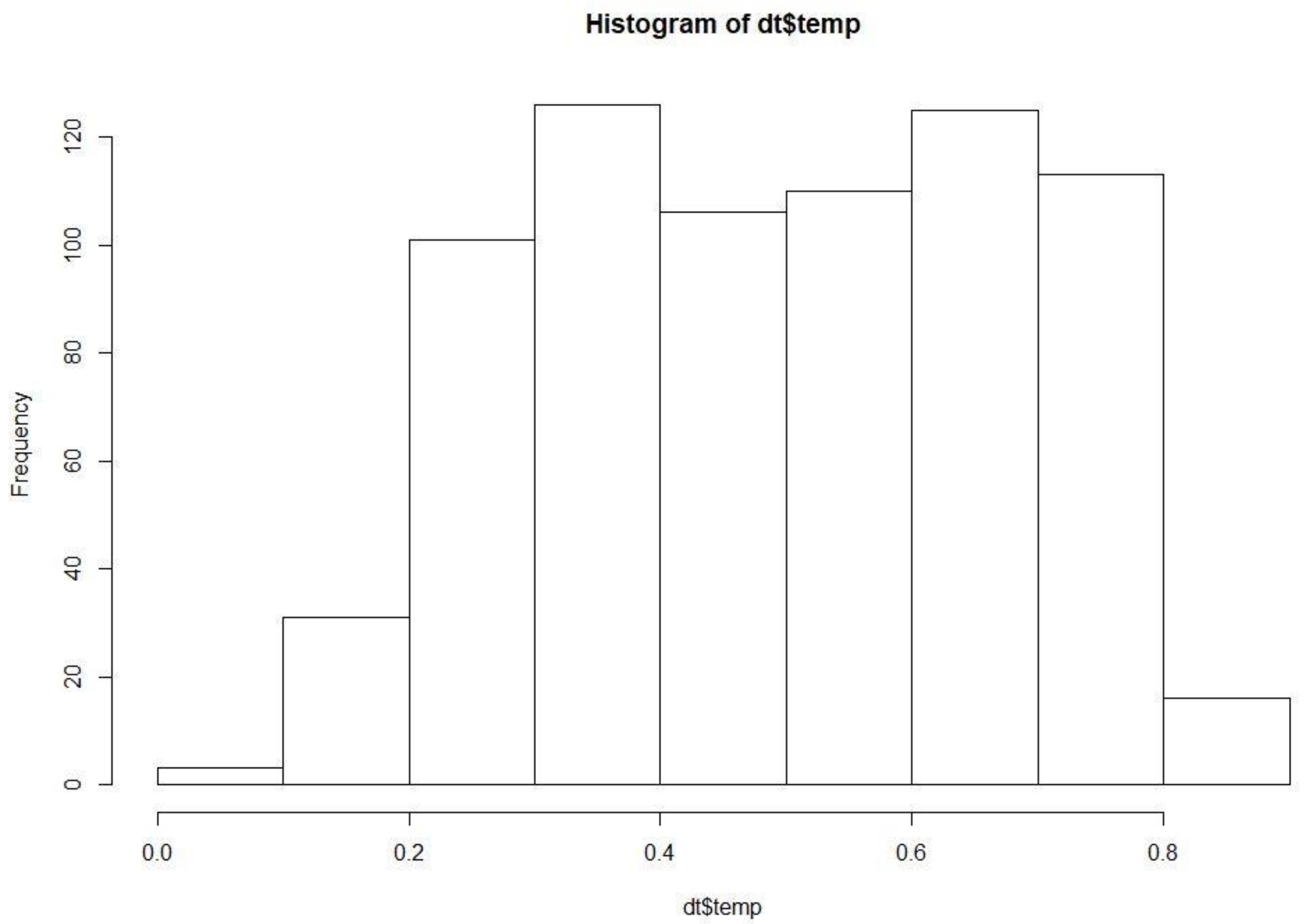


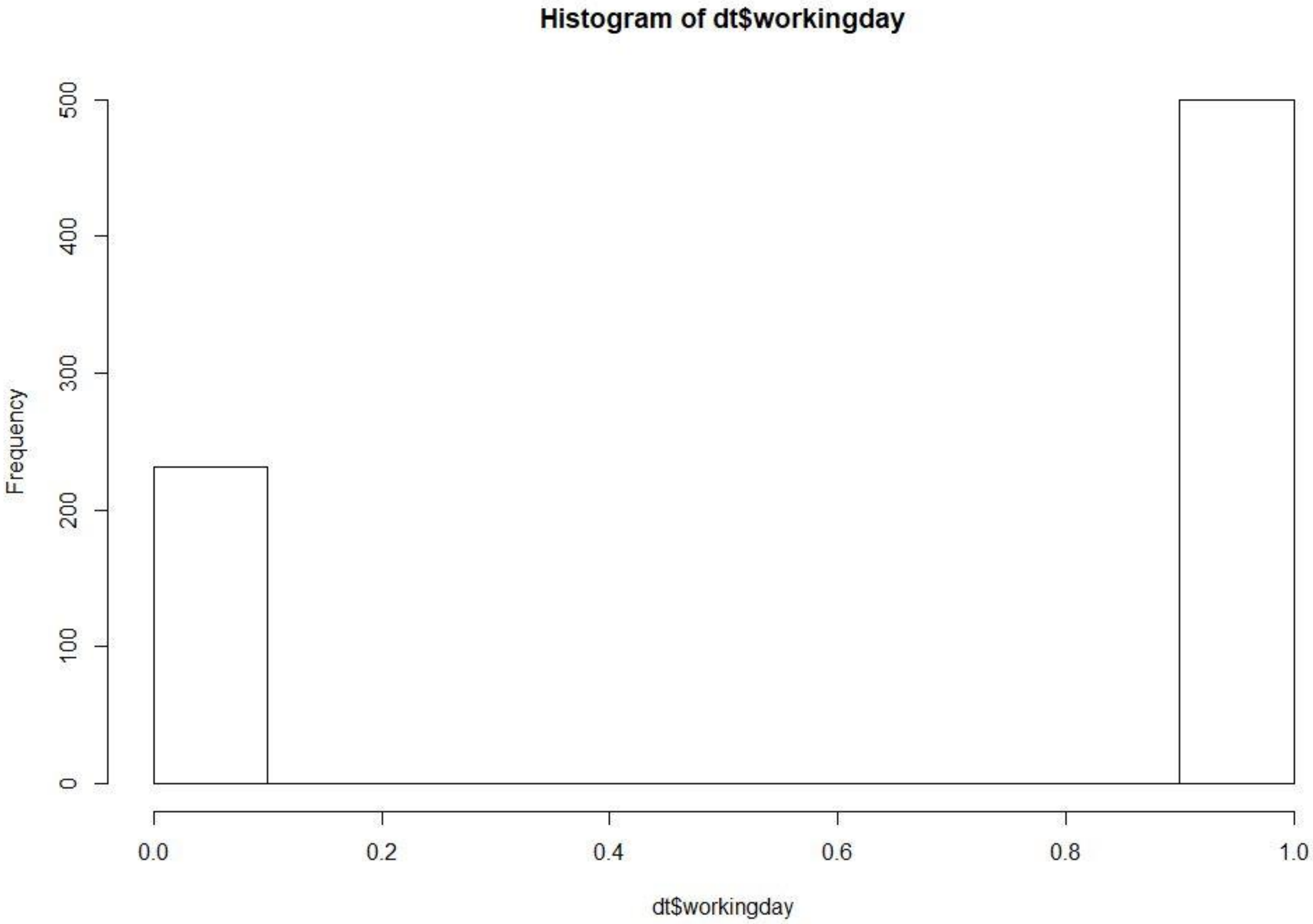
Season variable has 4 categories and all of them have almost equal distribution.



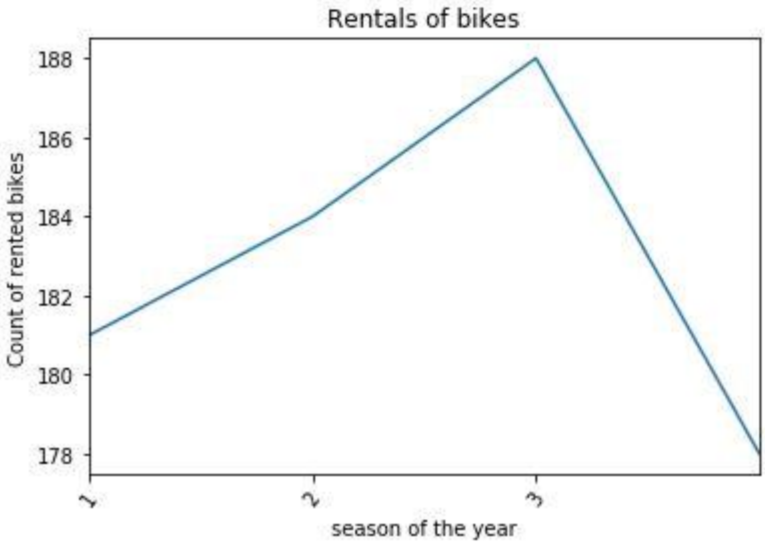


During weather 1 i.e mostly clear has the maximum contribution towards the count.

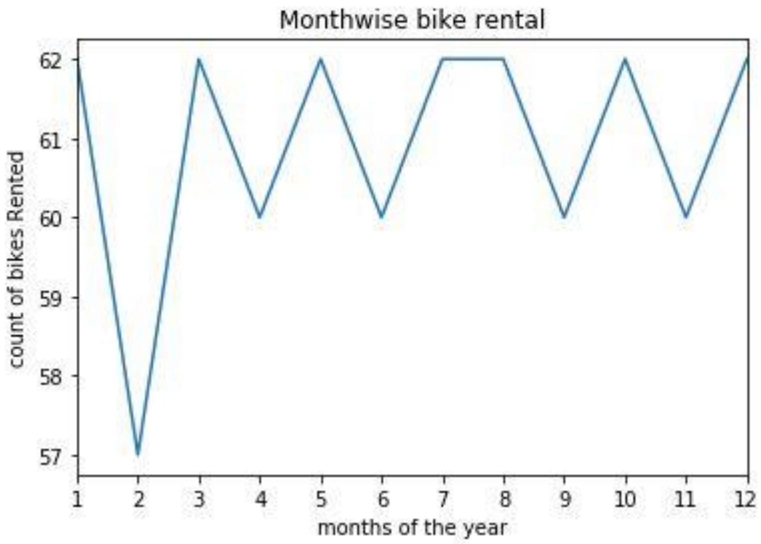




Mostly working days contributes more than the holidays.



The season 3 i.e fall has max booking of the bikes by the people.



Sudden increase can be seen during the march, the climatic condition could be the reason.

2.2.1 Missing Value Analysis

As we saw from the EDA that there were some missing values in between the observations of the variables, it may be due to human error, or just didn't have the information or etc. To calculate the missing value percentage of each variable , we do so that for us it is important to calculate the missing value if it exceeds 30%. To analyse this on the data set we created the following plot to understand it better. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

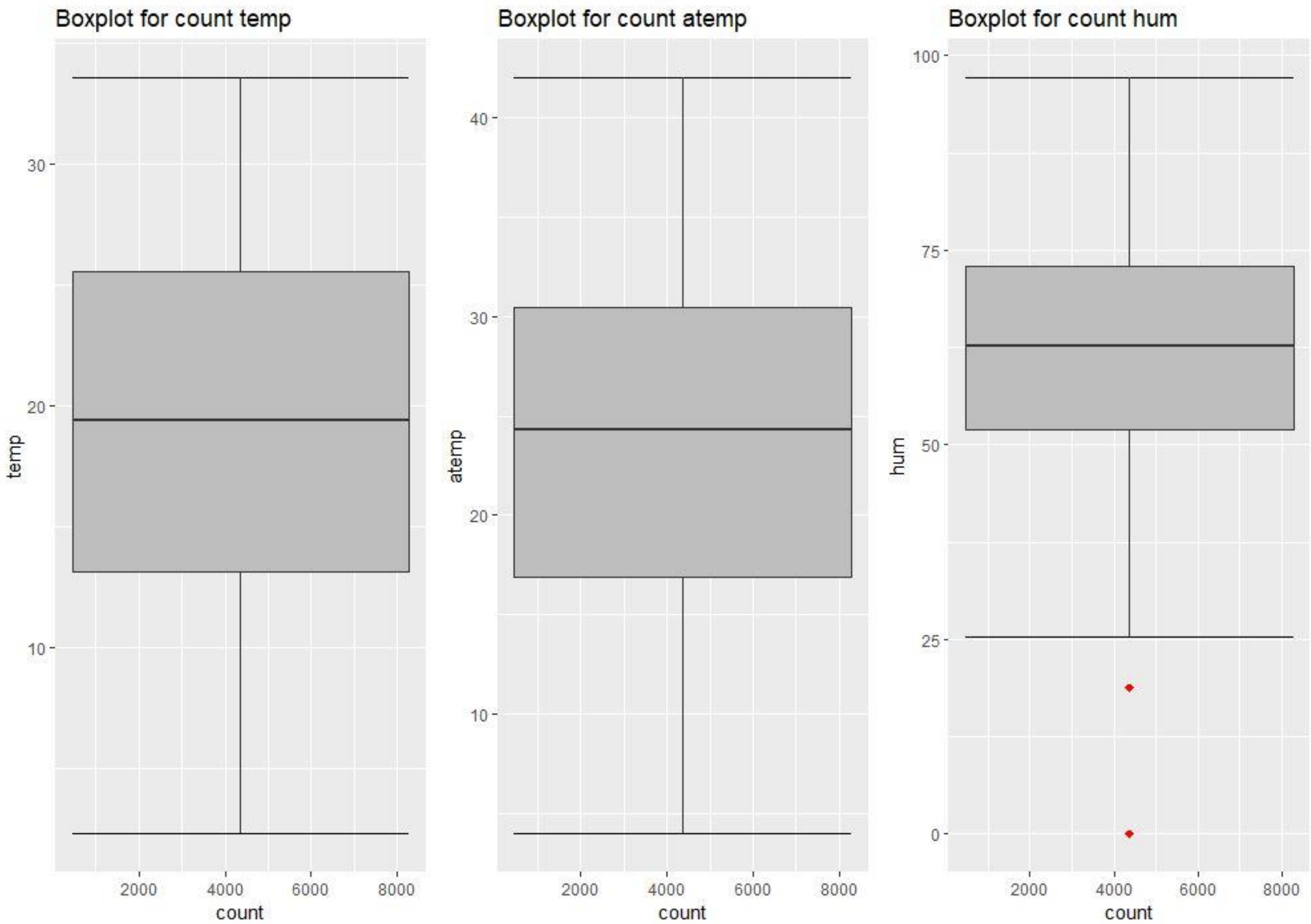
	Variables	Missed_val_percentage
0	dteday	0.0
1	season	0.0
2	yr	0.0
3	mnth	0.0
4	holiday	0.0
5	weekday	0.0
6	workingday	0.0
7	weathersit	0.0
8	temp	0.0
9	atemp	0.0
10	hum	0.0
11	windspeed	0.0
12	casual	0.0
13	registered	0.0
14	cnt	0.0

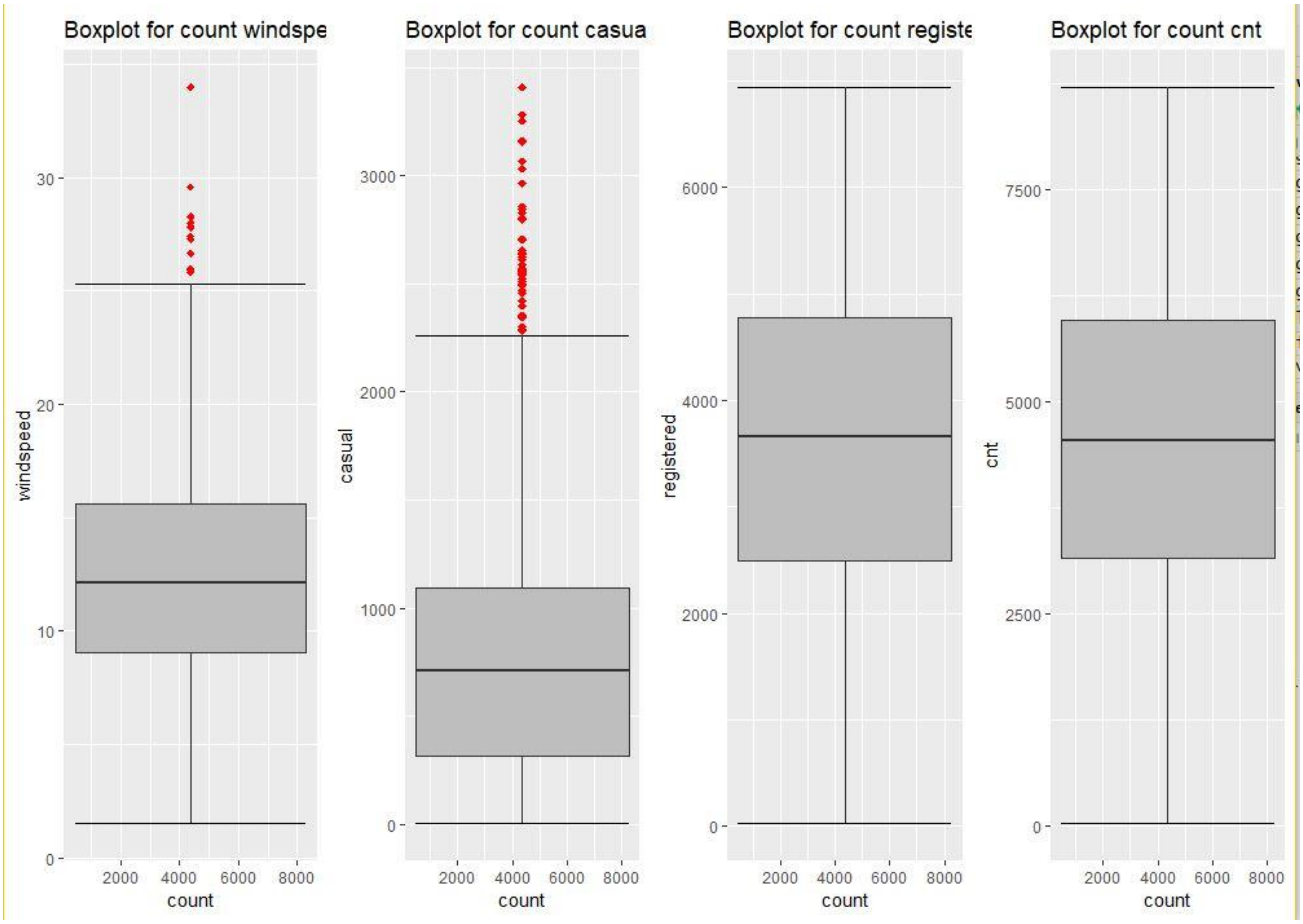
Now, as we had no missing value in our model we moved further.

2.1.2 <u>Outlier Analysis</u>

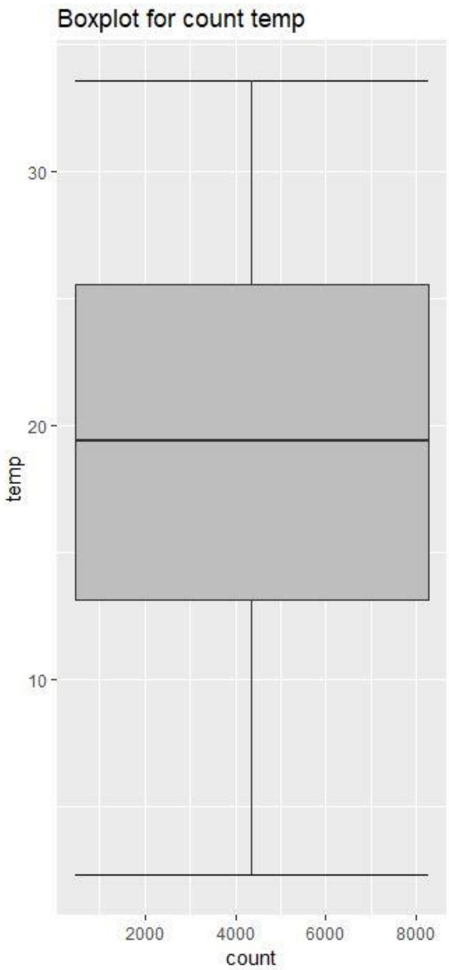
Presence of outliers in the dataset can lead to the biased outcome of the model, which may effect the accuracy of the model developed. This is mostly explained by the presence of the extreme values, this could be solved by two approaches. These approached can be , eliminating the outliers out of the data set and Second is to replace the outliers with NA and then performing the imputation. Outliers are very important because they affect the mean and median which in turn affects the error (absolute and mean) in any data set. When you plot the error you might get big deviations if outliers are in the data set. ... Or sometimes, outliers are so large that they should broken down and analysed separately.

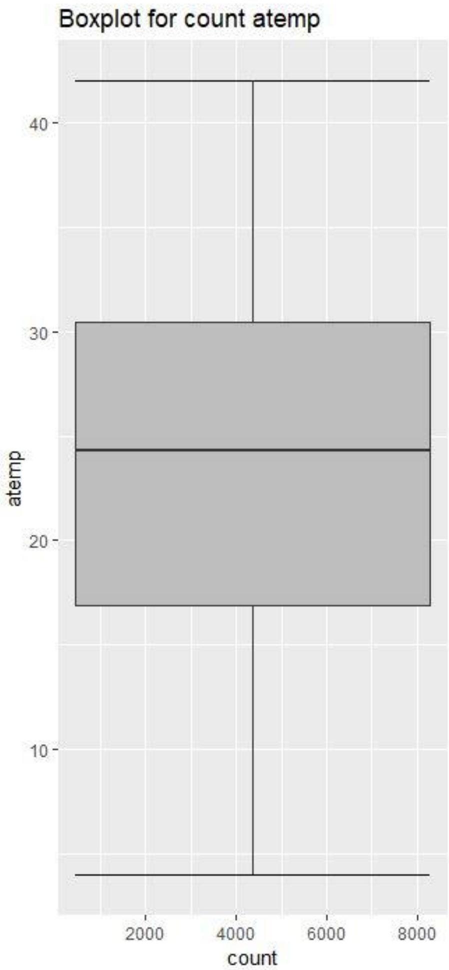
As you will see first we created the box plots for the variables with outliers.

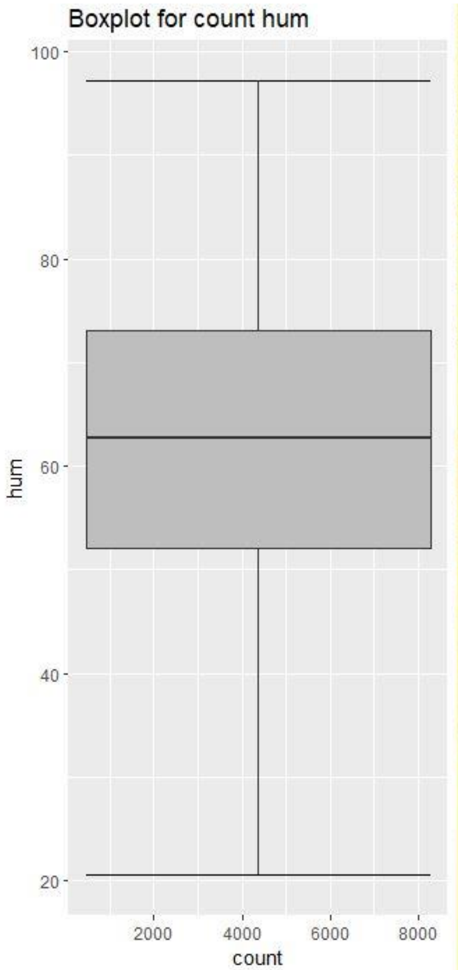


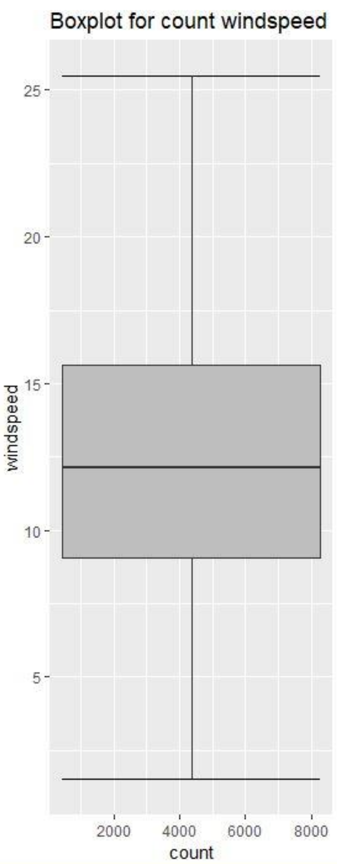


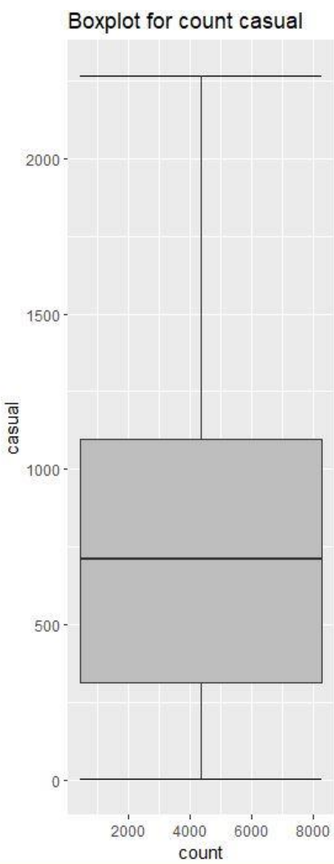
Now, forming the boxplots after replacing the outliers with NA and performing the imputation

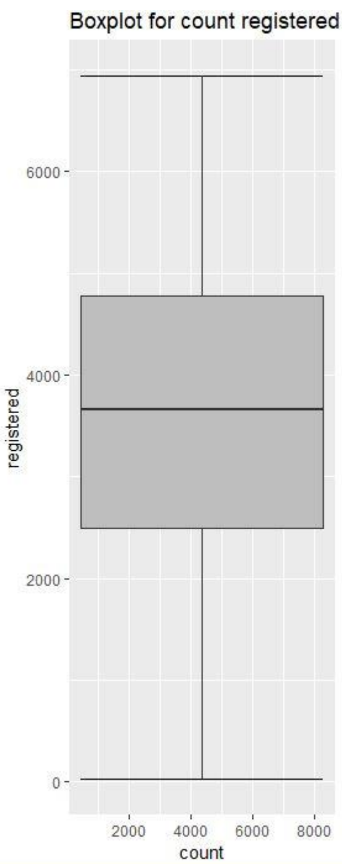

Boxplot for count temp
temp
30
20
10
2000
4000
6000
8000
count

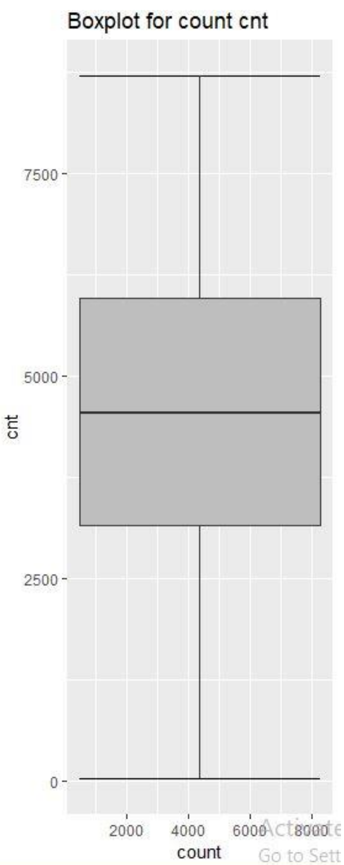

Boxplot for count atemp
atemp
40
30
20
10
2000
4000
6000
8000
count


Boxplot for count hum
hum
100
80
60
40
20
2000
4000
6000
8000
count


Boxplot for count windspeed
windspeed
25
20
15
10
5
2000
4000
6000
8000
count


Boxplot for count casual
casual
2000
1500
1000
500
0
2000
4000
6000
8000
count


Boxplot for count registered
registered
6000
4000
2000
0
2000
4000
6000
8000
count


Boxplot for count cnt
cnt
7500
5000
2500
0
2000
4000
6000
8000
count

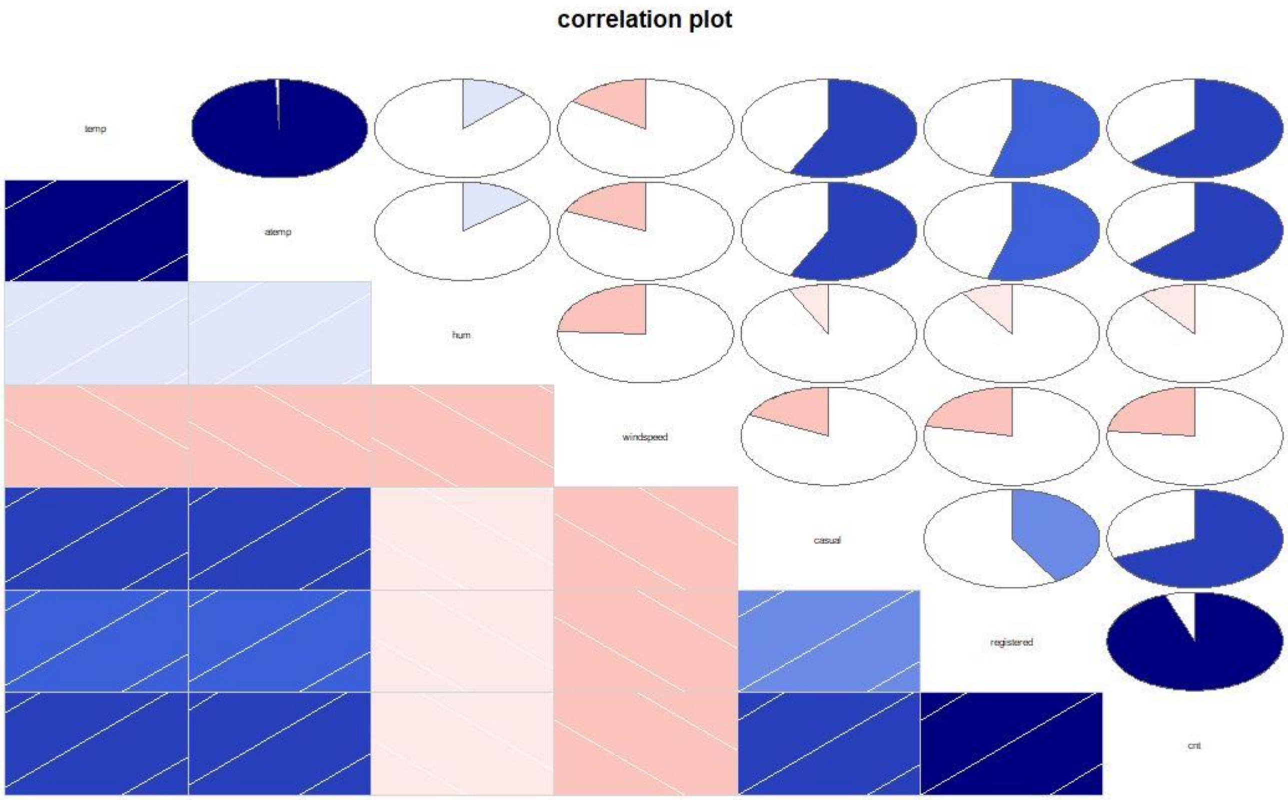
2.1.3 <u>Feature Selection</u>

This is the very important part of the pre-processing . It is very important to extract the meaningful components of the data that have been provided to increase the efficiency of the model also this may cost us redundency of data. There is no point of carrying the all the components along the way which provides same information to us which may cause increase in overhead. To deal it, we need to reduce the unnecessary components through selection technique of meaningful variables out of data given. To achieve this we have chosen correlation analysis for the numerical variables and chi-sq test for the categorical variables.

All this nuance in variable importance is worth negotiating, because feature selection has a multiplicative effect on the overall modeling process. Good variable importance and feature selection means: Less Data, so easier data layer. Simpler Models, so faster machine learning

For continuous variables it was clearly seen that atemp had correlation more that 0.95 , so elimination took place . and for the categorical variables. VIF value were considered

	temp	atemp	hum	windspeed	casual	registered	cnt
temp	1.0000000	0.9917016	0.12672159	-0.1569155	0.57379603	0.54001197	0.6274940
atemp	0.9917016	1.0000000	0.13992396	-0.1829480	0.57424494	0.54419176	0.6310657
hum	0.1267216	0.1399240	1.00000000	-0.2411599	-0.07511766	-0.09598498	-0.1056645
windspeed	-0.1569155	-0.1829480	-0.24115987	1.0000000	-0.17815527	-0.21692701	-0.2336573
casual	0.5737960	0.5742449	-0.07511766	-0.1781553	1.00000000	0.41491716	0.6845470
registered	0.5400120	0.5441918	-0.09598498	-0.2169270	0.41491716	1.00000000	0.9455169
cnt	0.6274940	0.6310657	-0.10566446	-0.2336573	0.68454699	0.94551692	1.0000000



Correlation among the continuous variables were <0.95 in most of the cases except the temp and a temp variable where it is more than 0.95 and hence we dropped the variable.

Noe checking for the categorical variable through chi-sq test

	season	yr	mnth	holiday	weekday	workingday	weathersit
season	0.0000000	9.999288e-01	0.00000000	6.831687e-01	1.000000e+00	8.865568e-01	2.117930e-02
yr	0.9999288	4.011854e-160	1.00000000	1.000000e+00	9.999996e-01	1.000000e+00	1.273794e-01
mnth	0.0000000	1.000000e+00	0.00000000	5.593083e-01	1.000000e+00	9.933495e-01	1.463711e-02
holiday	0.6831687	1.000000e+00	0.55930831	2.706945e-153	8.567055e-11	4.033371e-11	6.008572e-01
weekday	1.0000000	9.999996e-01	1.00000000	8.567055e-11	0.000000e+00	6.775031e-136	2.784593e-01
workingday	0.8865568	1.000000e+00	0.99334952	4.033371e-11	6.775031e-136	5.484935e-160	2.537640e-01
weathersit	0.0211793	1.273794e-01	0.01463711	6.008572e-01	2.784593e-01	2.537640e-01	2.484533e-315

We saw that the holiday variable is correlated to some extent and through anova test we checked that it is not contributing towards the dependent variable therefore we dropped the variable.

```
No variable from the 9 input variables has collinearity problem.

The linear correlation coefficients ranges between:
min correlation ( temp ~ weekday ):  -0.0001699624
max correlation ( mnth ~ season ):  0.8314401

---------- VIFs of the remained variables --------
   Variables      VIF
1     season 3.535209
2         yr 1.021548
3       mnth 3.326460
4    weekday 1.012165
5 workingday 1.010438
6 weathersit 1.785102
7       temp 1.217569
8        hum 1.947422
9  windspeed 1.163527
```

2.2 Modeling

After a thorough preprocessing we will be variable. Following are the models which we have built –

2.2.1 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users.

After building model the model on the Decision tree on the train data and predicting the dependent variable of the test data we calculated the efficiency of the model with the help of MAPE.

MAPE-22.5

After which we calculated the efficiency of the model

Accuracy -77.5

2.2.2 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

After building model the model on the Random Forest on the train data and predicting the dependent variable of the test data we calculated the efficiency of the model with the help of MAPE.

MAPE-14.2

After which we calculated the efficiency of the model

Accuracy -85.8

.

2.2.3 Liner Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

After building model the model on the Linear Regression on the train data and predicting the dependent variable of the test data we calculated the efficiency of the model with the help of MAPE.

MAPE-19.2
After which we calculated the efficiency of the model
Accuracy -80.8

.

Chapter 3
Conclusion

3.1 Model Evaluation

After Creating the model we analyzed the efficiency of the model by MAPE

What is MAPE?

Mean Absolute Percent Error (MAPE) is a very commonly used metric for forecast accuracy. ...
Since MAPE is a measure of error, high numbers are bad and low numbers are good. For reporting purposes, some companies will translate this to accuracy numbers by subtracting the MAPE from 100.

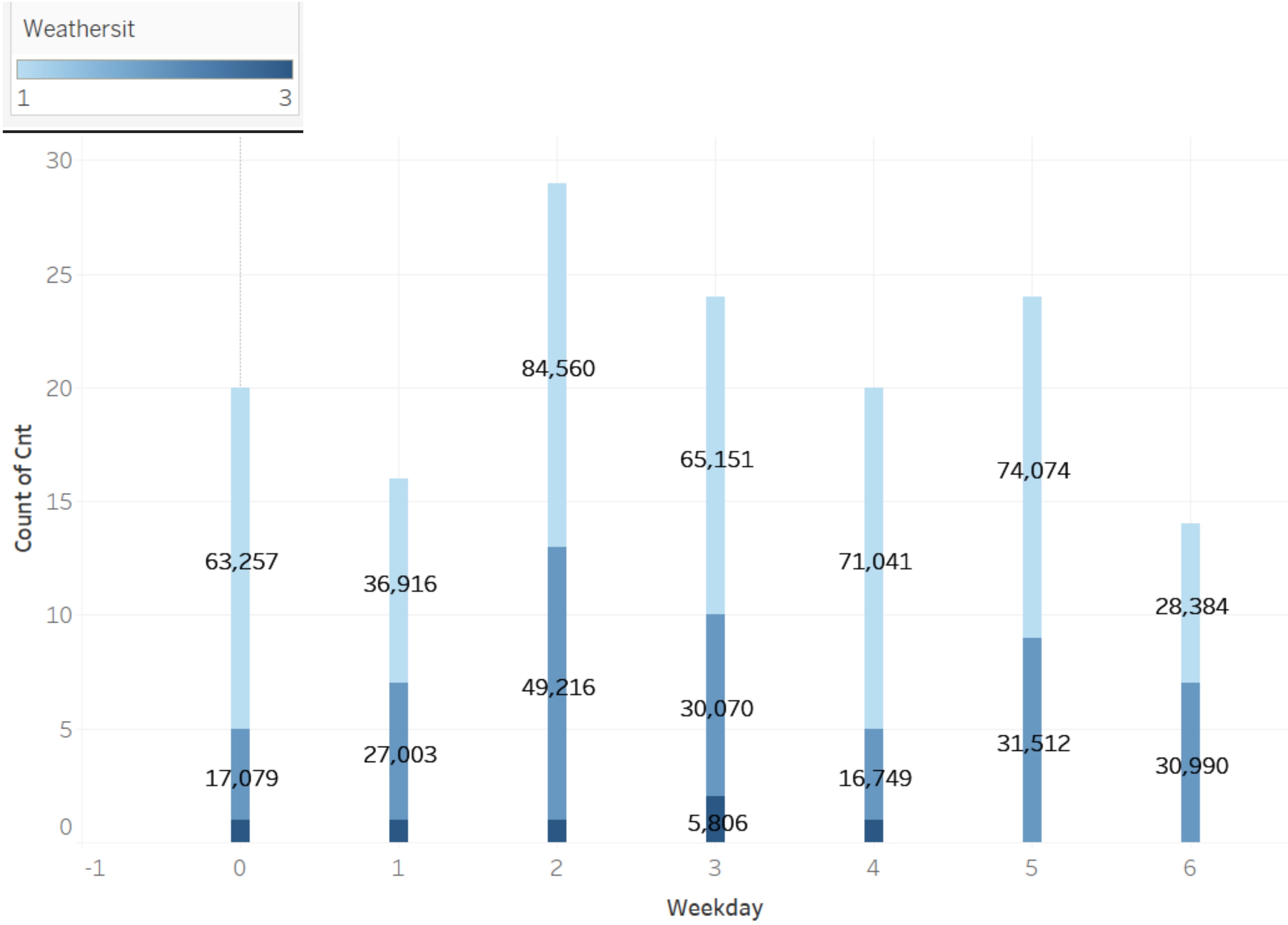
3.2 Model Selection

We conclude from the model evaluation that the best suited model for the data set was Random Forest, having the highest accuracy.

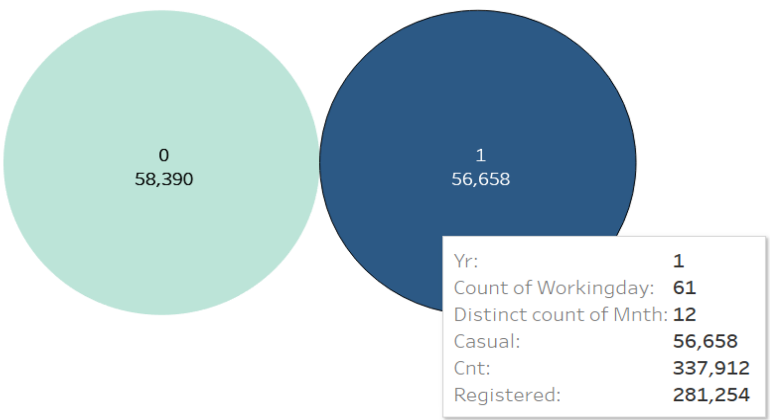
3.2 Answers of asked questions

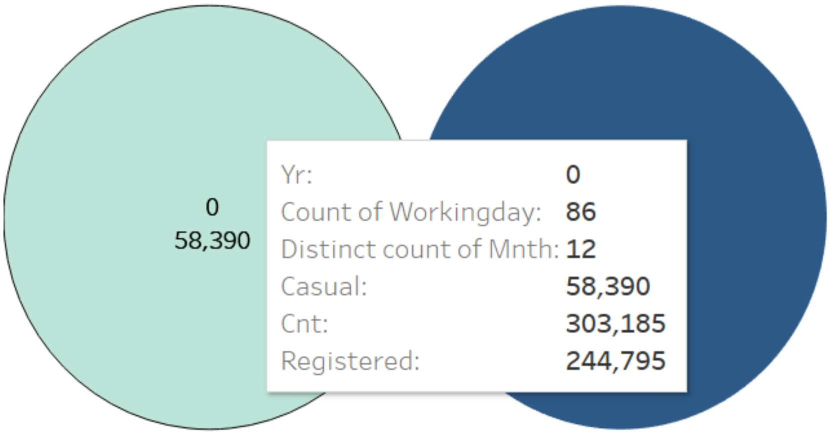
As we were asked to predict the bike renting on the daily basis , I generated the following graph using the Tableau Tool for better understanding

As we can see from the graph that weekdays were the days when the bike renting count was mostly maximum than the week days.
The count of bikes rented is high when the weather conditions is 1 i.e clear weather which was also one of the hypothesis that we made in the beginning



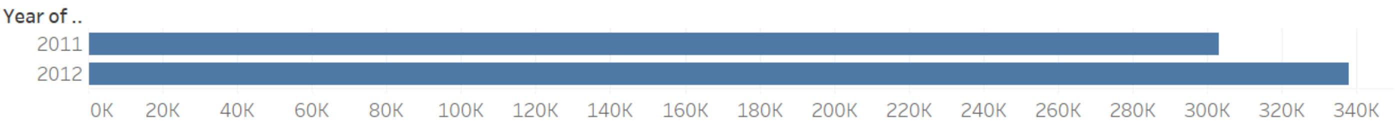
I created the packed bubbles for both the years with labels of total contribution by the casual customers.
The bubble 0 is for year 2011 where registered users total contribution is 244,795 for the whole year and the for the year 2012 it is 281,254





From the above packed bubble we also observed that there is decrease in the count of weekdays but count remains high.

Now Looking at the following graph we can say the there is time trend in the bike renting as the total count by the casual and the registered customers have increased in the following year



APPENDIX:

R code for fig

```
for(i in 1:length(var_cont))

{

 assign(paste0("gn",i),ggplot(aes_string(y=(var_cont[i]),x="cnt"),data=subset(dt))+

       stat_boxplot(geom="errorbar",width=0.5)+

       geom_boxplot(outlier.colour="red",fill="grey",outlier.shape=18,outlier.size=2,notch=FALSE)+

       theme(legend.position="bottom")+

       labs(y=var_cont[i],x="count")+
```

```
    ggtitle(paste("Boxplot for count",var_cont[i])))

}
#plotting together all the plots generated
gridExtra::grid.arrange(gn1,gn2,gn3,ncol=3)
gridExtra::grid.arrange(gn4,gn5,gn6,gn7,ncol=4)corr diag
corrgram(dt[,var_cont],order=F,upper.panel=panel.pie, text.panel=panel.txt,main="correlation plot")
hist(dt$season)
hist(dt$hum)
hist(dt$holiday)
hist(dt$workingday)
hist(dt$temp)
hist(dt$atemp)
hist(dt$windspeed)
hist(dt$weathersit)
```