***Predicting Bike for Rent***

Aishwarya Kapoor

31st March 2019

**Content**

1. Introduction

1.1 Problem Statement …………………………………………… 3

1.2 Data …………………………………………………………… 3

2. Methodology

2.1 Pre-processing ………………………………………………… 6

2.2 Feature Selection ……………………………………………... 8

2.3 Modeling

2.3.1 Model selection ……………………………………………… 9

3. Conclusion

3.1 Conclusion …………………………………………………….. 12

4. Appendix 1 – Extra Figures ……………………………………... 14

5. Appendix B - R Code……………………………………………... 17

6. Reference …………………………………………………………. 23

**Chapter 1**

Introduction

* 1. Problem Statement

The objective of this Case - Predication of bike rental count on daily based on the environmental and seasonal settings. The aim of this project is to forecast the number of bikes that are going on rent on weekdays, weekends or holidays. Accordingly numbers of bikes will be available for the people who wants to rent the bike.

This help our client to reduce to cost of maintaining, reduce the cost of warehouse and happy customers.

* 1. Data

Our task is to build a forecasting model which will predicts the numbers of bikes are going for rent on weekdays, weekend or holidays. Given below is the sample of data which we are going to use for forecasting the data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | mnth | holiday |
| 1 | 1/1/2011 | 1 | 0 | 1 | 0 |
| 2 | 1/2/2011 | 1 | 0 | 1 | 0 |
| 3 | 1/3/2011 | 1 | 0 | 1 | 0 |
| 4 | 1/4/2011 | 1 | 0 | 1 | 0 |
| 5 | 1/5/2011 | 1 | 0 | 1 | 0 |
| 6 | 1/6/2011 | 1 | 0 | 1 | 0 |

Table 1.1 Bike Rental Sampling Data Year 2011 (columns 1 to 6)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| weekday | workingday | weathersit | Temp | Atemp |
| 6 | 0 | 2 | 0.344167 | 0.363625 |
| 0 | 0 | 2 | 0.363478 | 0.353739 |
| 1 | 1 | 1 | 0.196364 | 0.189405 |
| 2 | 1 | 1 | 0.2 | 0.212122 |
| 3 | 1 | 1 | 0.226957 | 0.22927 |
| 4 | 1 | 1 | 0.204348 | 0.233209 |

Table 1.2 Bike Rental Sampling Data Year 2011 (columns 7 to 11)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| hum | windspeed | Casual | Registered | Cnt |
| 0.805833 | 0.160446 | 331 | 654 | 985 |
| 0.696087 | 0.248539 | 131 | 670 | 801 |
| 0.437273 | 0.248309 | 120 | 1229 | 1349 |
| 0.590435 | 0.160296 | 108 | 1454 | 1562 |
| 0.436957 | 0.1869 | 82 | 1518 | 1600 |
| 0.518261 | 0.089565 | 88 | 1518 | 1606 |

Table 1.3 Bike Rental Sampling Data Year 2011 (columns 12 to 16)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | mnth | holiday |
| 366 | 1/1/2012 | 1 | 1 | 1 | 0 |
| 367 | 1/2/2012 | 1 | 1 | 1 | 1 |
| 368 | 1/3/2012 | 1 | 1 | 1 | 0 |
| 369 | 1/4/2012 | 1 | 1 | 1 | 0 |
| 370 | 1/5/2012 | 1 | 1 | 1 | 0 |
| 371 | 1/6/2012 | 1 | 1 | 1 | 0 |

Table 1.4 Bike Rental Sampling Data Year 2012 (columns 1 to 6)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| weekday | workingday | weathersit | temp | atemp |
| 0 | 0 | 1 | 0.37 | 0.375621 |
| 1 | 0 | 1 | 0.273043 | 0.252304 |
| 2 | 1 | 1 | 0.15 | 0.126275 |
| 3 | 1 | 2 | 0.1075 | 0.119337 |
| 4 | 1 | 1 | 0.265833 | 0.278412 |
| 5 | 1 | 1 | 0.334167 | 0.340267 |

Table 1.5 Bike Rental Sampling Data Year 2012 (columns 7 to 12)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| hum | windspeed | Casual | registered | Cnt |
| 0.6925 | 0.192167 | 686 | 1608 | 2294 |
| 0.381304 | 0.329665 | 244 | 1707 | 1951 |
| 0.44125 | 0.365671 | 89 | 2147 | 2236 |
| 0.414583 | 0.1847 | 95 | 2273 | 2368 |
| 0.524167 | 0.129987 | 140 | 3132 | 3272 |
| 0.542083 | 0.167908 | 307 | 3791 | 4098 |

Table 1.6 Bike Rental Sampling Data Year 2012 (columns 12 to 16)

There are total 16 Variables out which 15 are Independent variables and one is our dependent variable that is – “cnt”

Below is a list of all independent variables:

|  |  |
| --- | --- |
| S.No | Predictors |
| 1 | instant |
| 2 | dteday |
| 3 | season |
| 4 | Yr |
| 5 | mnth |
| 6 | holiday |
| 7 | weekday |
| 8 | workingday |
| 9 | weathersit |
| 10 | temp |
| 11 | atemp |
| 12 | Hum |
| 13 | windspeed |
| 14 | casual |
| 15 | registered |

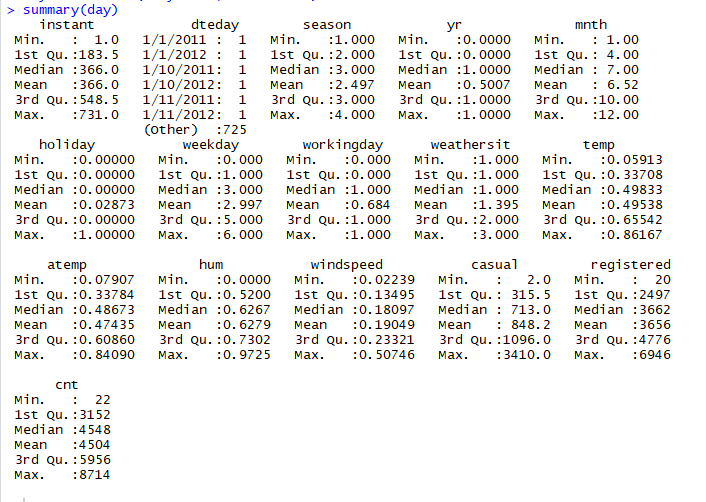
Table 1.7 Bike Rental Predictors Names

**Chapter 2**

2.1 Methodology

Before going further, we need to clean our data. Pre-processing of the data requires - study of the data, study about the types of variables are present, which all variables are important for the model, selection of the variables, check the NA’s in the data, check the outliers in the data after cleaning the data we will start to build the model.

Summary of our Data in which mean, medium and range of the each variable is present:



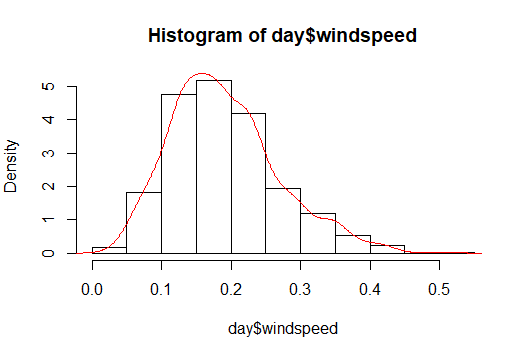
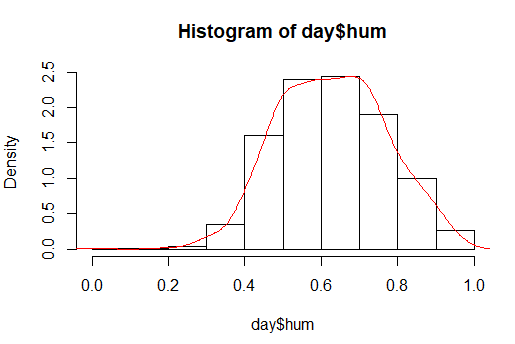
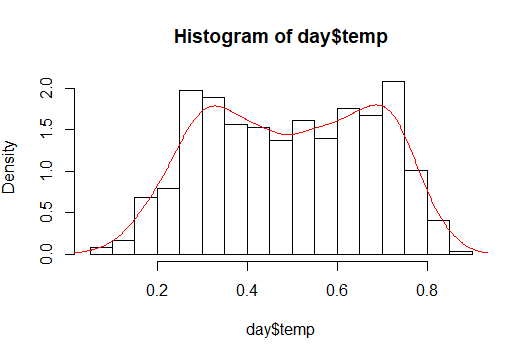
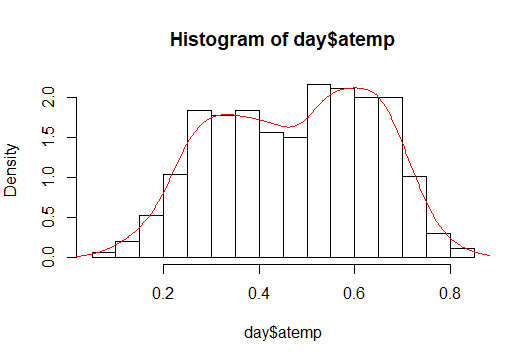
While pre-processing we have removed 2 variables from our data(“instant” and “dtdate”) because instant is the index number and dtdate is date of the each day. Also there are very few NA’s in the data and also very few outliers due to which are processing the data as is it(because to remove these 2 we need to require a minimum %).

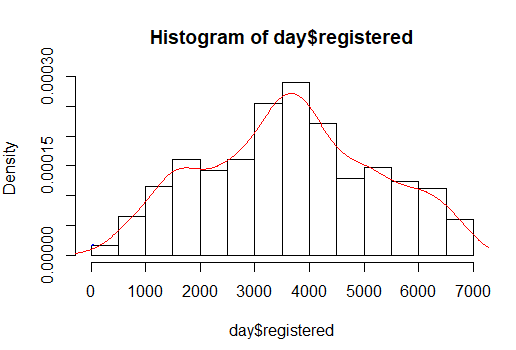
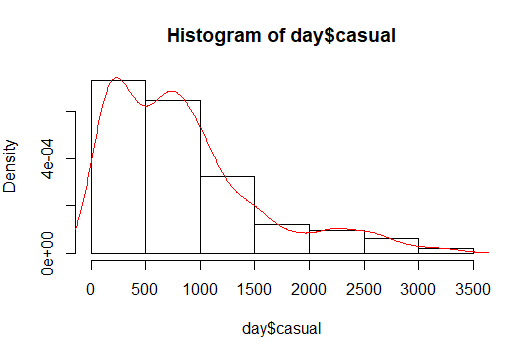
We have convert the following variables into categorical variable:

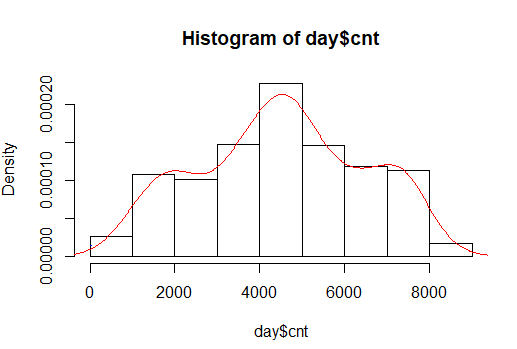
* season: Season (1:springer, 2:summer, 3:fall, 4:winter)
* yr: Year (0: 2011, 1:2012)
* mnth: Month (1 to 12)
* hr: Hour (0 to 23) holiday
* weather day is holiday or not (extracted from Holiday Schedule)
* weekday: Day of the week working day: If day is neither weekend nor holiday is 1,otherwise is 0.

we have plotted the probability density functions of all the numeric factors which we have available

in the data as well as the dependent variable.







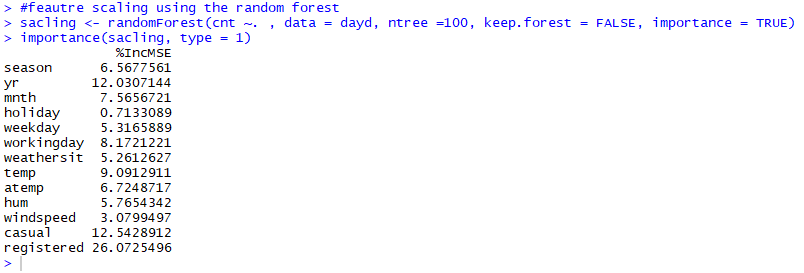
2.2 Feature Selection

Before performing any modeling 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

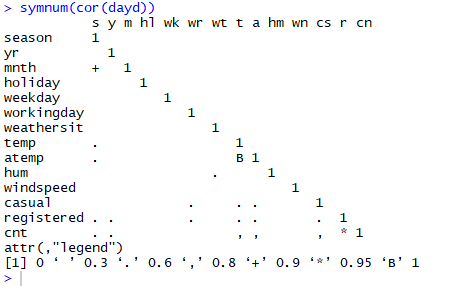
class prediction. There are several methods of doing that. Below we have used “Random Forest”to perform

features selection.



Using the random forest for scaling features give as a clear picture of the predictor that “registered” variable have the highest power for predicting the model. If, the number of the registered variable will increase the number renting bike will also increase which is our dependent variable “cnt”.

Now, to get the more detail about the variable we need to run one more test that is “correlation”. Below is the simplest way to know about the correlation. In the day data, registered variable is highly correlated with cnt by 0.95.



2.3 Modeling

2.3.1 Model selection

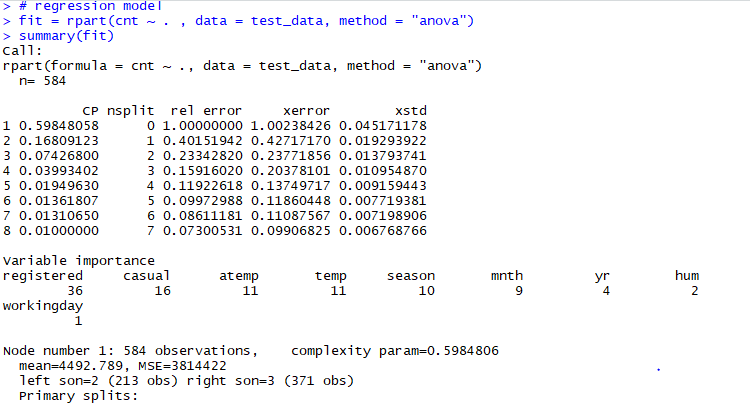
In the pre-processing stage we saw that the “cnt” is our dependent variable and it is highly correlated with the “registered” variable. Our client want to forecast the number of the rental bike in the future. We will use the “liner regression model” for the forecasting.

We will create the sample data and the test data which we will divide in 80 by 20 ratio .

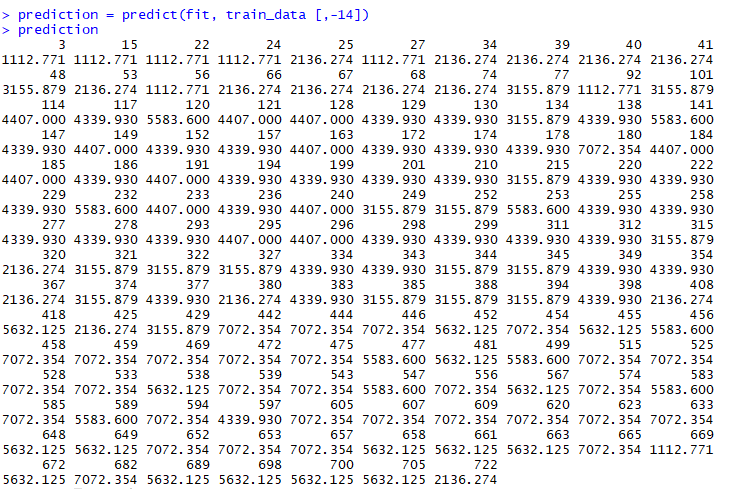
80% will be our test data on which we will run our model.

20% will be our train data on which we will do prediction.

We using the “anova” test which will help us to know whether there is any statistically significant differences between the means of independent (unrelated) groups.



After this, we will predict the rest 20% on the same model into “fit” to check that model is working fine and the rest other factors of the model.



We are using “Linear regression model” because this is not a normal case. This is a case of linear equation and in this we need to find the number of the rental bikes. That renting decision is depend on varies factor like, whether , speed of the wind, temperature , holiday or not but it is highly correlated with “registered” variable that means if people will registered online for bike there is a high chance that they will rent a bike.

**Chapter 3**

3.1 Conclusion

Now, we have run the few models we are most likely to know how accurate is the prediction.

Predictive performance can be measured by comparing Predictions of the models with real values of the target

variable, and calculating some average error measure.

**Mean Absolute Error (MAE)**

MAE 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.

> #calculating the MAPE

> mape = function(y, yhat){

+ mean(abs((y-yhat)/y))

+ }

> mape(train\_data[,14],prediction)

[1] 0.124693

**MSE**, **RMSE**, **MAE** can be obtained as follows:

> #Calculatig the regr.eval

> regr.eval(train\_data[,14], prediction, stats = c('mae' ,'rmse' ,'mape' , 'mse'))

mae rmse mape mse

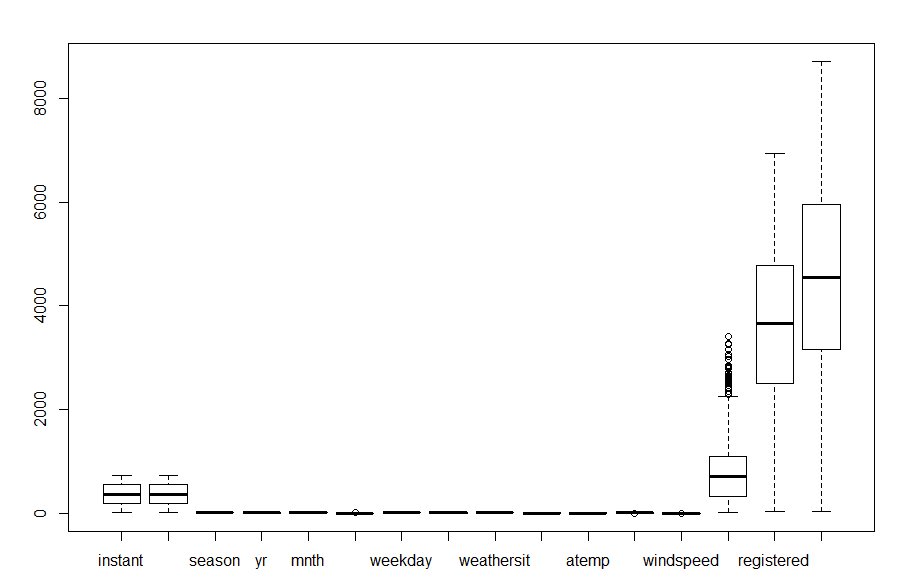
4.593207e+02 5.971043e+02 1.246930e-01 3.565335e+05

3.2 Model Selection

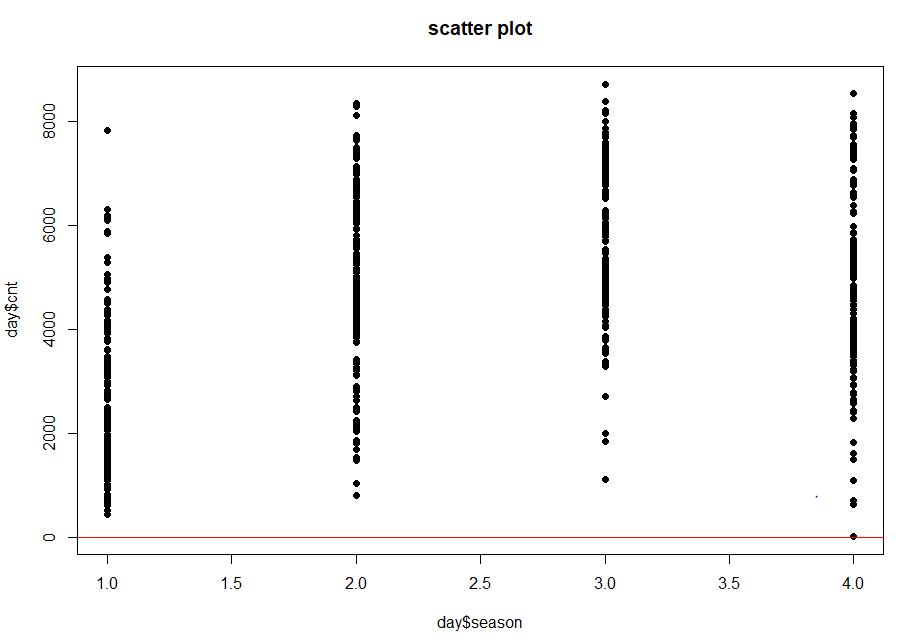
We can see that Linear regression model perform comparatively on average and therefore we can select either of the two modes without any loss of information.

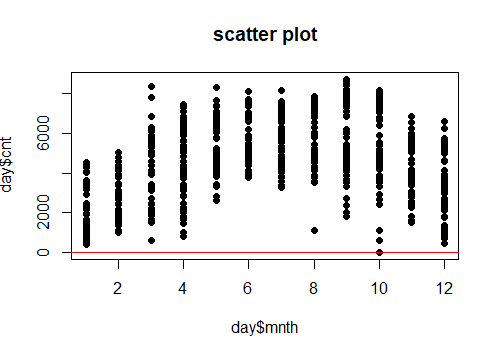
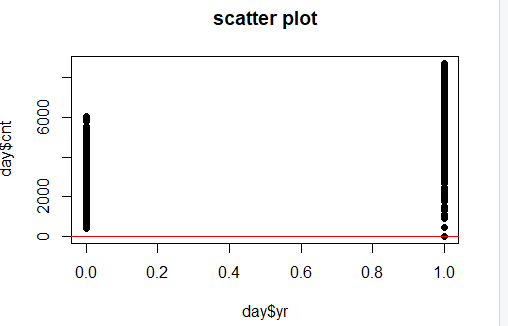
**Appendix 1 – Extra Figures**

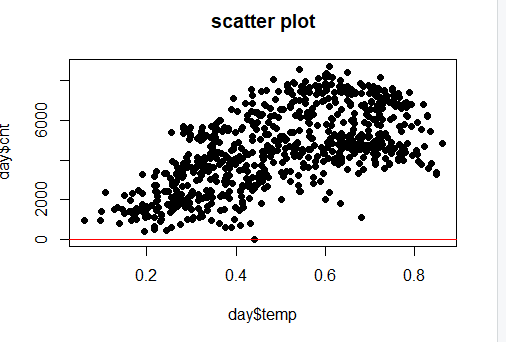
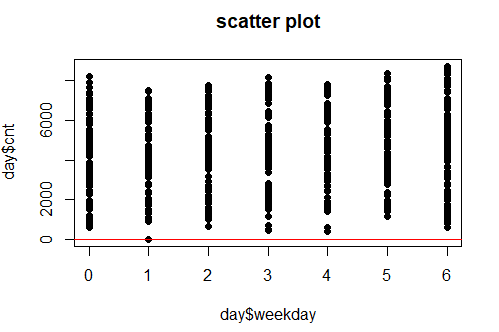
**#Box plot with all the variables**

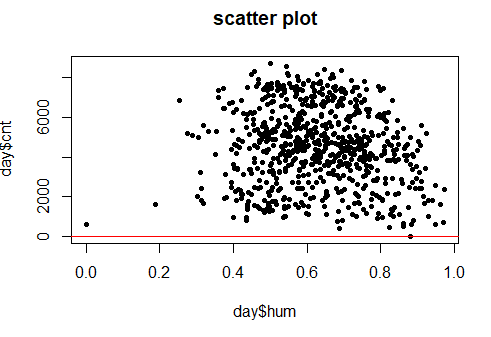
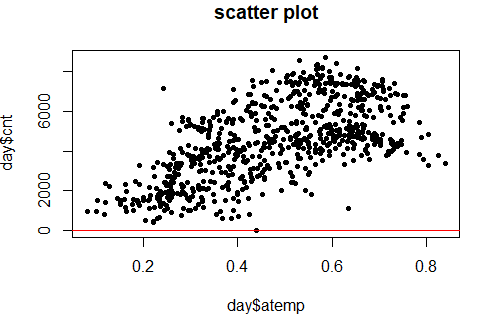


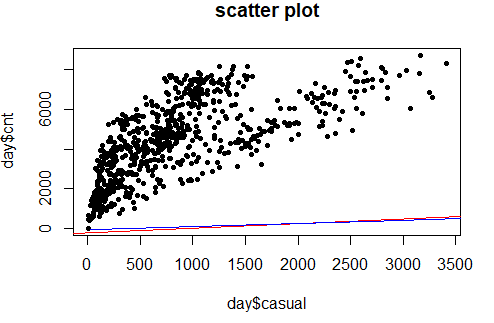
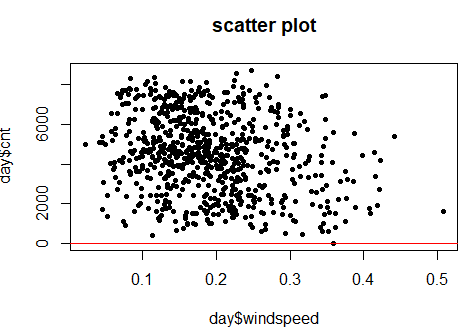
**#scatter plot**

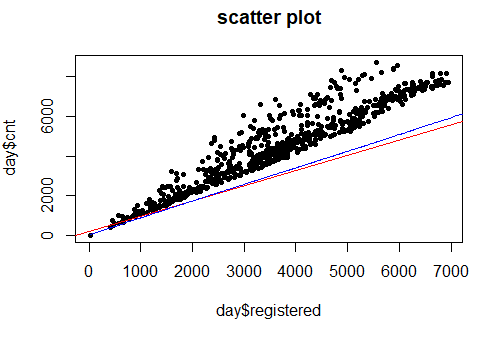




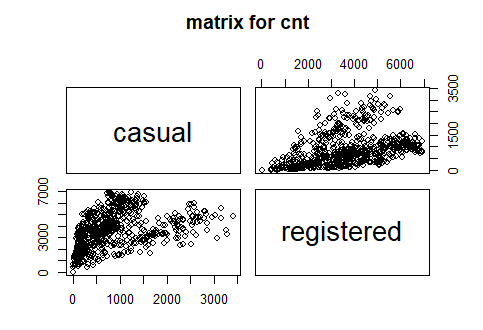




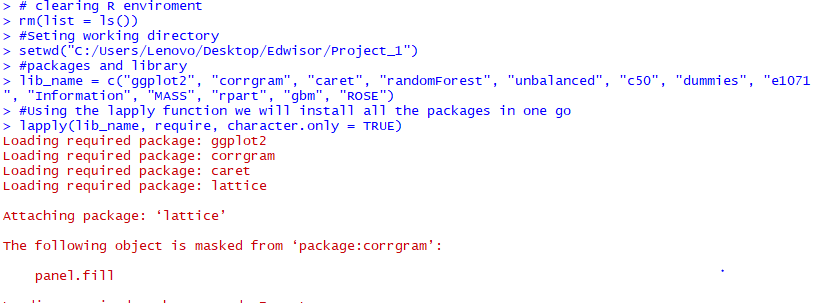
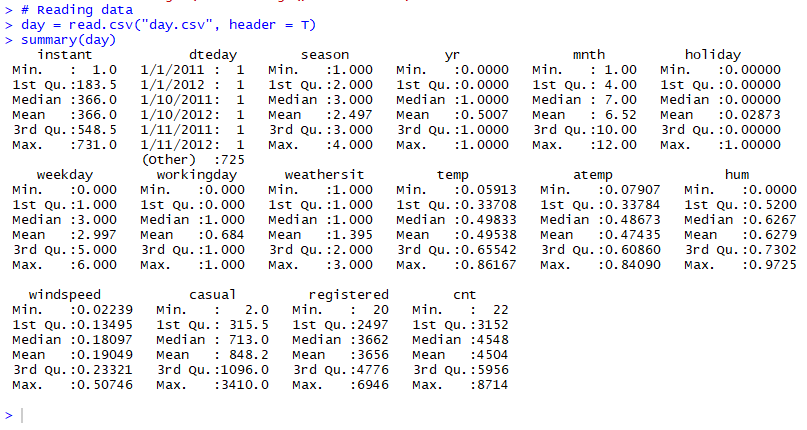
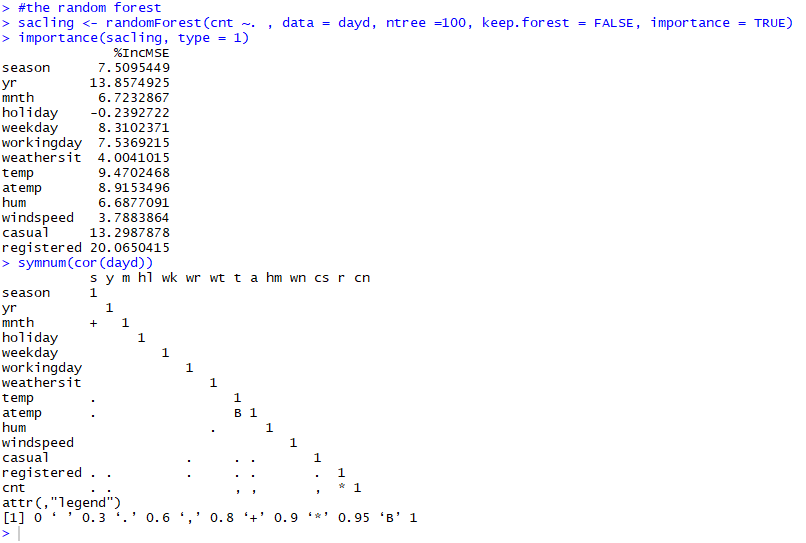




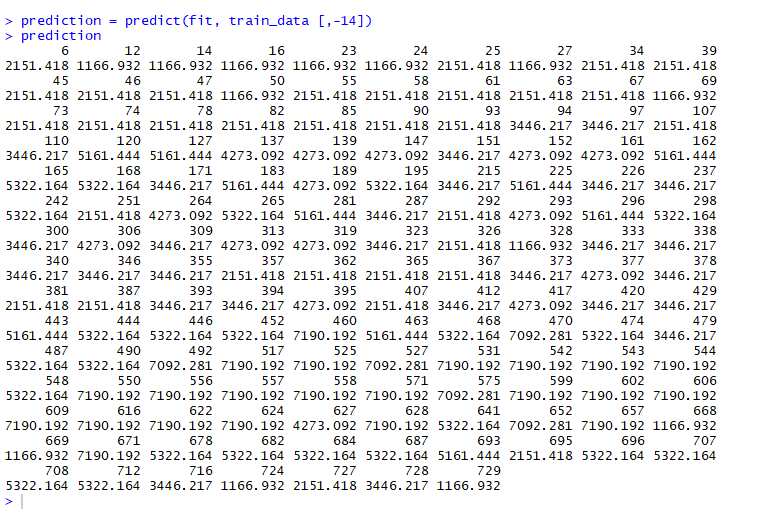
**#Matrix**



**Appendix B - R Code**

|  |
| --- |
| > # creating the sample data  > sample\_data = sample(1:nrow(dayd), 0.80 \* nrow(dayd))  > test\_data = dayd[sample\_data,]  > train\_data = dayd[-sample\_data,]  > # regression model  > fit = rpart(cnt ~ . , data = test\_data, method = "anova")  > summary(fit)  Call:  rpart(formula = cnt ~ ., data = test\_data, method = "anova")  n= 584  CP nsplit rel error xerror xstd  1 0.57999287 0 1.00000000 1.00335673 0.046559332  2 0.19942576 1 0.42000713 0.46970544 0.021224526  3 0.06162885 2 0.22058137 0.24040427 0.012445522  4 0.03700890 3 0.15895252 0.20841538 0.012494583  5 0.02726799 4 0.12194362 0.15830087 0.010472446  6 0.01446975 5 0.09467563 0.12887622 0.008465478  7 0.01212214 6 0.08020588 0.11494599 0.007430155  8 0.01000000 7 0.06808374 0.09512454 0.006079361  Variable importance  registered casual yr atemp temp season mnth workingday weekday  41 15 14 10 10 4 3 2 1  Node number 1: 584 observations, complexity param=0.5799929  mean=4603.95, MSE=3636063  left son=2 (337 obs) right son=3 (247 obs)  Primary splits:  registered < 3929.5 to the left, improve=0.5799929, (0 missing)  casual < 645.5 to the left, improve=0.4382784, (0 missing)  temp < 0.432373 to the left, improve=0.3935946, (0 missing)  atemp < 0.4308565 to the left, improve=0.3898811, (0 missing)  yr < 0.5 to the left, improve=0.3358901, (0 missing)  Surrogate splits:  yr < 0.5 to the left, agree=0.795, adj=0.514, (0 split)  casual < 734 to the left, agree=0.688, adj=0.263, (0 split)  temp < 0.46375 to the left, agree=0.670, adj=0.219, (0 split)  atemp < 0.4308565 to the left, agree=0.668, adj=0.215, (0 split)  season < 2.5 to the left, agree=0.592, adj=0.036, (0 split)  Node number 2: 337 observations, complexity param=0.1994258  mean=3360.694, MSE=1794963  left son=4 (111 obs) right son=5 (226 obs)  Primary splits:  registered < 2294 to the left, improve=0.7000677, (0 missing)  casual < 532.5 to the left, improve=0.4727075, (0 missing)  temp < 0.457917 to the left, improve=0.4089953, (0 missing)  atemp < 0.4305435 to the left, improve=0.4040404, (0 missing)  season < 1.5 to the left, improve=0.3341176, (0 missing)  Surrogate splits:  atemp < 0.2642065 to the left, agree=0.798, adj=0.387, (0 split)  casual < 172.5 to the left, agree=0.798, adj=0.387, (0 split)  temp < 0.264674 to the left, agree=0.792, adj=0.369, (0 split)  season < 1.5 to the left, agree=0.777, adj=0.324, (0 split)  mnth < 3.5 to the left, agree=0.772, adj=0.306, (0 split)  Node number 3: 247 observations, complexity param=0.06162885  mean=6300.215, MSE=1161806  left son=6 (148 obs) right son=7 (99 obs)  Primary splits:  registered < 5283.5 to the left, improve=0.4560347, (0 missing)  casual < 831 to the left, improve=0.4383200, (0 missing)  yr < 0.5 to the left, improve=0.2605385, (0 missing)  temp < 0.417083 to the left, improve=0.1658754, (0 missing)  atemp < 0.415708 to the left, improve=0.1554371, (0 missing)  Surrogate splits:  casual < 829.5 to the left, agree=0.660, adj=0.152, (0 split)  windspeed < 0.143044 to the right, agree=0.656, adj=0.141, (0 split)  atemp < 0.6524705 to the left, agree=0.636, adj=0.091, (0 split)  temp < 0.712083 to the left, agree=0.619, adj=0.051, (0 split)  Node number 4: 111 observations, complexity param=0.01212214  mean=1761.171, MSE=404282.4  left son=8 (44 obs) right son=9 (67 obs)  Primary splits:  registered < 1377.5 to the left, improve=0.5736089, (0 missing)  casual < 207 to the left, improve=0.3289677, (0 missing)  atemp < 0.4261145 to the left, improve=0.1498984, (0 missing)  season < 1.5 to the left, improve=0.1329693, (0 missing)  temp < 0.433333 to the left, improve=0.1215148, (0 missing)  Surrogate splits:  casual < 70 to the left, agree=0.685, adj=0.205, (0 split)  hum < 0.82125 to the right, agree=0.658, adj=0.136, (0 split)  mnth < 1.5 to the left, agree=0.640, adj=0.091, (0 split)  weekday < 5.5 to the right, agree=0.640, adj=0.091, (0 split)  workingday < 0.5 to the left, agree=0.640, adj=0.091, (0 split)  Node number 5: 226 observations, complexity param=0.02726799  mean=4146.301, MSE=604223.6  left son=10 (181 obs) right son=11 (45 obs)  Primary splits:  casual < 1475 to the left, improve=0.4240247, (0 missing)  registered < 2915 to the left, improve=0.2872480, (0 missing)  temp < 0.5957425 to the left, improve=0.2817689, (0 missing)  atemp < 0.5590415 to the left, improve=0.2683115, (0 missing)  workingday < 0.5 to the right, improve=0.1297540, (0 missing)  Surrogate splits:  holiday < 0.5 to the left, agree=0.805, adj=0.022, (0 split)  weekday < 0.5 to the right, agree=0.805, adj=0.022, (0 split)  workingday < 0.5 to the right, agree=0.805, adj=0.022, (0 split)  windspeed < 0.0538021 to the right, agree=0.805, adj=0.022, (0 split)  Node number 6: 148 observations, complexity param=0.0370089  mean=5704.892, MSE=819098.1  left son=12 (116 obs) right son=13 (32 obs)  Primary splits:  casual < 1990 to the left, improve=0.6482653, (0 missing)  workingday < 0.5 to the right, improve=0.5394435, (0 missing)  weekday < 5.5 to the left, improve=0.2505940, (0 missing)  registered < 4283 to the left, improve=0.2462550, (0 missing)  yr < 0.5 to the left, improve=0.2129420, (0 missing)  Surrogate splits:  workingday < 0.5 to the right, agree=0.932, adj=0.688, (0 split)  weekday < 0.5 to the right, agree=0.878, adj=0.438, (0 split)  registered < 3958 to the right, agree=0.791, adj=0.031, (0 split)  Node number 7: 99 observations  mean=7190.192, MSE=352253.1  Node number 8: 44 observations  mean=1166.932, MSE=135840.5  Node number 9: 67 observations  mean=2151.418, MSE=196380.1  Node number 10: 181 observations, complexity param=0.01446975  mean=3893.917, MSE=346212.6  left son=20 (83 obs) right son=21 (98 obs)  Primary splits:  registered < 3334.5 to the left, improve=0.49032470, (0 missing)  casual < 587 to the left, improve=0.37137880, (0 missing)  temp < 0.5957425 to the left, improve=0.24605720, (0 missing)  atemp < 0.5805125 to the left, improve=0.24039720, (0 missing)  season < 1.5 to the left, improve=0.08570313, (0 missing)  Surrogate splits:  workingday < 0.5 to the left, agree=0.696, adj=0.337, (0 split)  atemp < 0.2812455 to the left, agree=0.635, adj=0.205, (0 split)  weekday < 5.5 to the right, agree=0.619, adj=0.169, (0 split)  temp < 0.2804165 to the left, agree=0.619, adj=0.169, (0 split)  mnth < 4.5 to the left, agree=0.613, adj=0.157, (0 split)  Node number 11: 45 observations  mean=5161.444, MSE=355278.7  Node number 12: 116 observations  mean=5322.164, MSE=282697.7  Node number 13: 32 observations  mean=7092.281, MSE=307707.3  Node number 20: 83 observations  mean=3446.217, MSE=236443.5  Node number 21: 98 observations  mean=4273.092, MSE=125650.3 |
|  |
| |  | | --- | | > | |



**References**

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical*

*Learning*. Vol. 6. Springer.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media.