**Bike Renting**

**Stuti Ghildiyal**

**August 2019**

# Contents

1. **[Introduction](#_bookmark0) 3**
   1. [ProblemStatement](#_bookmark1) 3
   2. [Data](#_bookmark2) 3
2. [**Methodology**](#_bookmark3) **5**
   1. [PreProcessing](#_bookmark4) 5
      1. Variable identification 6
      2. Visualization 6
3. Univariate Analysis 6
4. Bivariate Analysis 7
   * 1. Missing Value Analysis 11
     2. Outlier Analysis 12
     3. Feature Selection 15
5. Correlation 15
6. Anova 17
   * 1. Feature Scaling 19
   1. Modelling 19
      1. ModelSelection 19
7. Decision Tree 19
8. Random Forest 20
9. Linear Regression 22
   * 1. Visualizing models 24
10. Prediction Plots 24
11. **Conclusion 27**
    1. ModelEvaluation 27
       1. Mean AbsoluteError(MAE) 27
       2. Mean SquaredError(MSE) 27
    2. ModelSelection 27

**Appendix A -Imp Plots 29**

**Chapter 1**

1. **Introduction** 
   1. **Problem Statement:** The objective of this Project is to predict the count of the bikes to be rented on daily basis. This count will take environmental and seasonal settings from the historical data in account while forecasting the daily demand.
   2. **Data:** As the dataset given has dependent and independent values, it will come under supervised Machine learning. Our task is to build a Regression models which will help us predicting the count of bikes rented depending on the factors provided.

Given below is a sample of the data set that we are using for our prediction.

This dataset contains the rental count in between year 2011 and 2012 based on seasonal and environment etc.

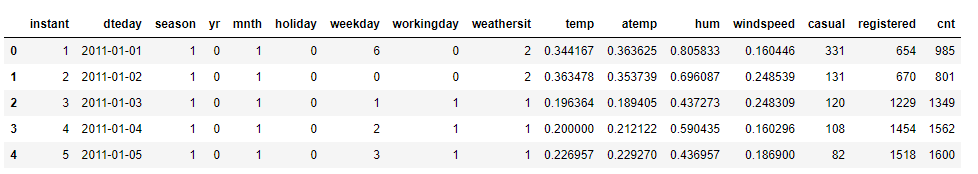
The data given is having 731 rows and 16 columns/variables.

Out of these 16 variables two variables which is **casual**, **registered** are the target variables while their sum constitutes to be another target variable named as **cnt.**

The casual target variable contains total bikes acquired by the customers who are not already registered means at random they hired the bike.

While registered variable represents the hired number of bikes only by the persons who are already registered and they are historical customers.

Thus, combining both the counts that is casual and registered builds “cnt” and it gives the total count of rented bikes on each particular date of given month and year.

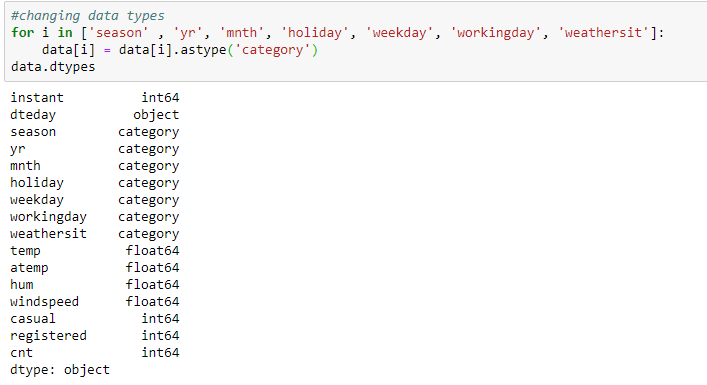
****

|  |  |
| --- | --- |
| **Variable** | **Explanation** |
| instant | Daily customer index |
| dteday | Date index for both the years |
| season | Season (1:springer, 2:summer, 3:fall, 4:winter) |
| yr | Year (0: 2011, 1:2012) |
| mnth | Month (1 to 12) |
| holiday | weather day is holiday or not (extracted from Holiday Schedule) |
| weekday | Day of the week |
| workingday | If day is neither weekend nor holiday is 1, otherwise is 0. |
| weathersit | (extracted from Freemeteo) 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 + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| temp | Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) |
| atemp | Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale) |
| hum (Humidity) | Normalized humidity. The values are divided to 100 (max) |
| windspeed | Normalized wind speed. The values are divided to 67 (max) |
| casual | count of casual users |
| registered | The number of registered users at a given day |
| cnt (Count) | Total Rentals with both casual and registered users |

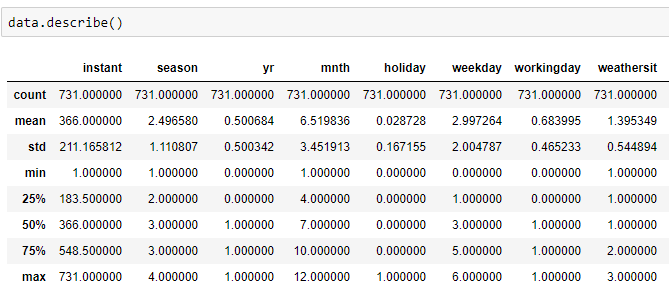
**Chapter 2**

1. **Methodology**
   1. **Pre Processing**: Before we proceed to create our model on top of the provided data. It is necessary to do Exploratory Data Analysis. Exploratory Data Analysis (EDA) is an approach to analysing data sets to summarize their main characteristics. As the result depends on the data, EDA makes sure the quality of input data is high which will lead to high quality results. We can perform EDA as follows:
      1. **Variable Identification:** In Order to understand the data, we need to first, Identifying Predictor (Input) and Target (output) variables. Then, Identifying the data type and category of the variables.
2. **Types of Variable:** Our Target Variable is ‘CNT’ , and Predictor variables are (dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,hum, windspeed, casual, registered) .
3. **DataTypes:** Character(dteday),Numeric(instant,season,yr,mnth,holiday,weekday,workingday,weathersit,casual,registered,cnt ) ,factor(temp,atemp,windspeed).

We need to change the datatype of our variables before starting anything as there are variables that are of category type but present in numeric type.

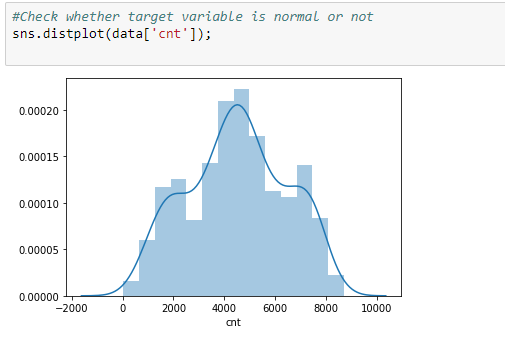


1. **Variable Categories:**Categorical (season, yr, mnth, holiday, weekday, workingday, weathersit), Continuous (temp, atemp, hum, windspeed, casual, registered)

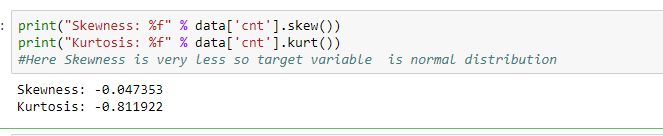


**2.1.2 Visualization:** Exploring Variables one by one to understand central tendency, spread of the variable, distribution of each category, association and disassociation between variables at a predefined significance level.

### Univariate Analysis: The analysis of univariate data is the simplest form of analysis since the information deals with only one quantity that changes. Here we will try to describe and find pattern in our target value. Checking the distribution of individual variables.

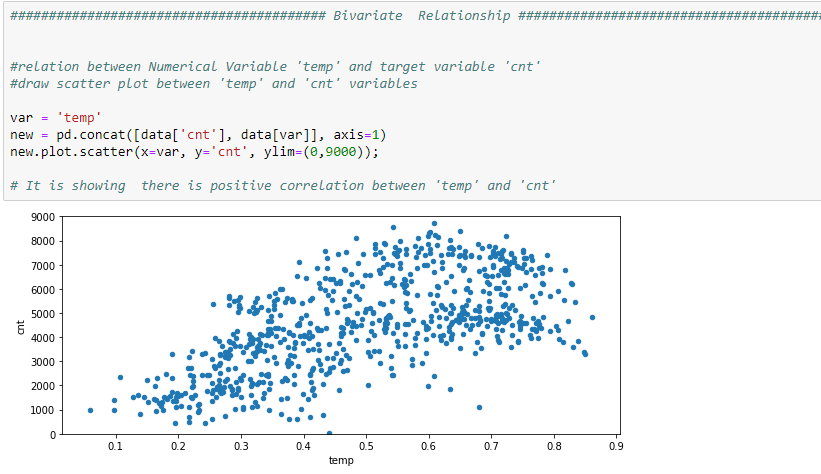


Checking for Skewness and Kurtosis: Skewness is usually described as a measure of a dataset’s symmetry – or lack of symmetry.   A perfectly symmetrical data set will have a skewness of 0. If the skewness is between -0.5 and 0.5, the data are fairly symmetrical, if the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed. If the skewness is less than -1 or greater than 1, the data are highly skewed.

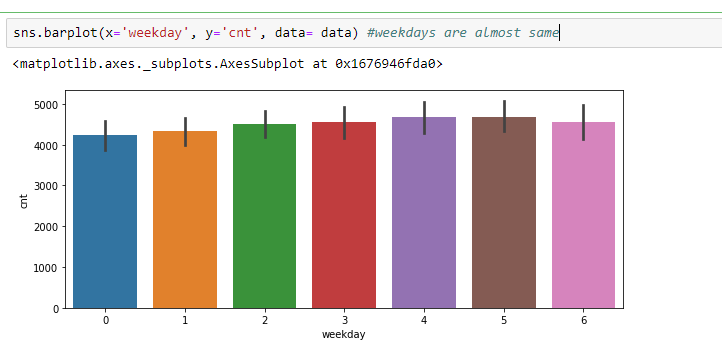
****

Here Skewness is very less so target variable is normal distribution.

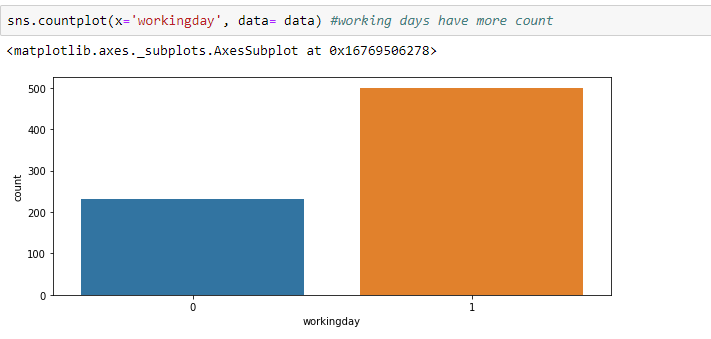
1. **Bi-variate Analysis:** We are checking the relationship of variables with our target variable.



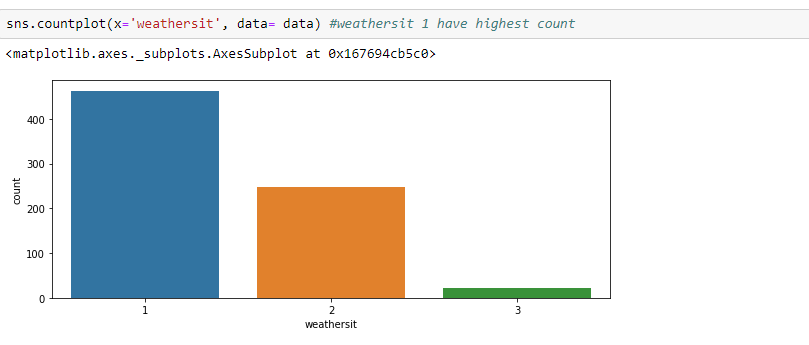
From this we can say that there is positive correlation between temp and cnt. This means as the value of temp increase keeping all other variables constant, there is an increase in cnt variable also.



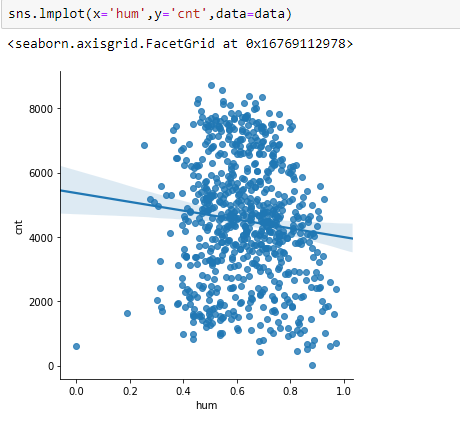
We can clearly see that there is almost no difference in rentals on each day; it means that all the days have almost same rental counts.

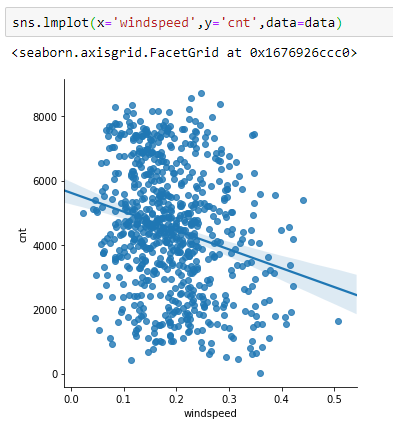


We can infer that people rent bikes more on working days as there may be chances of rental requirements to commute rather than on non-working days.

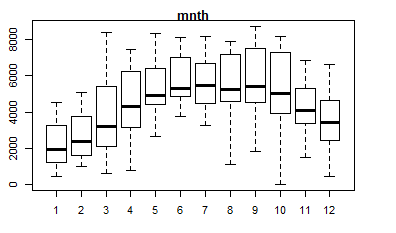


We can infer that on weatersit1 (clear weather), people prefer more to rent bikes rather on weathersit3 (heavy rain).





From Linear plot of humidity vs. count and linear plot of windspeed vs. count, it is clear that humidity and windspeed are inversely proportion with count, i.e. with increase in humidity, the rental counts decrease. Same with Wind speed.

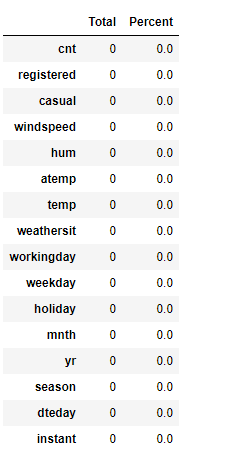


Bike rental is more between March till October.

**2.1.3 Missing values treatment:**

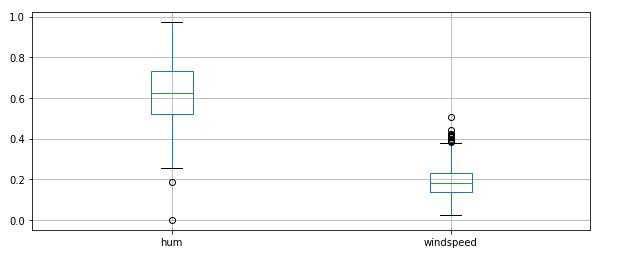
In statistics, missing data*,* ormissing values*,* occurs when there is no data value stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data*.* If a column has more than 30% of data as missing value then either we can ignore the entire column or we ignore those observations.

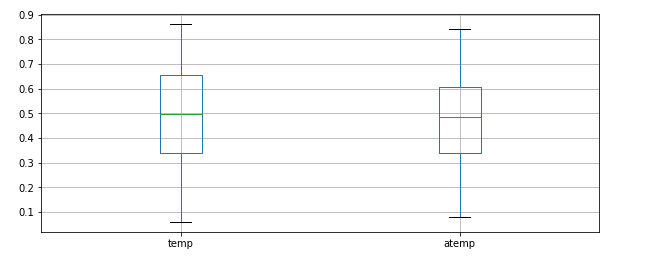
Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. We checked for missing values in our data and came to know that there are no missing values

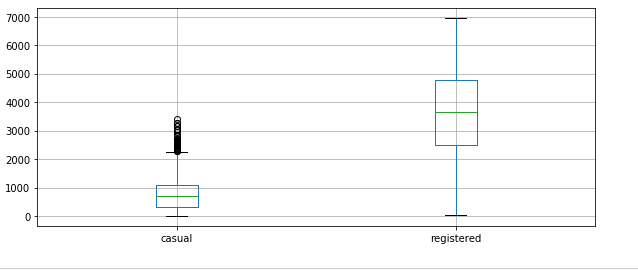


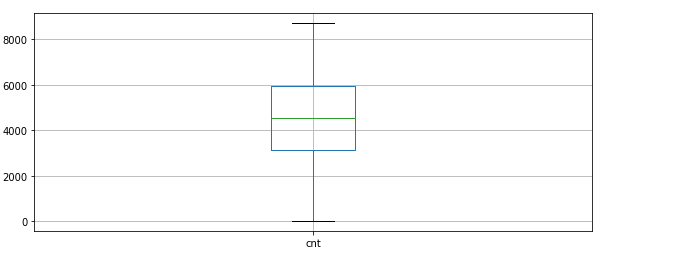
**2.1.4 Outlier treatment:** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers can drastically change the results of the data analysis and statistical modelling. Outliers are the unwanted abnormal values that may get generated due to rough handling of data or few values emerging as out of the range value in which most of the data lies. These outside range data is also known as anomalies.

There are numerous unfavourable impacts of outliers in the data set. It increases the error variance and reduces the power of statistical tests. If the outliers are non-randomly distributed, they can decrease normality. They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions. The box plot for our data could be seen as follows:









In order to remove the outliers we can either make these outliers as NA and can impute as missing value using methods as KNN or median or mean or mode or we can delete the entire row which contains outliers.

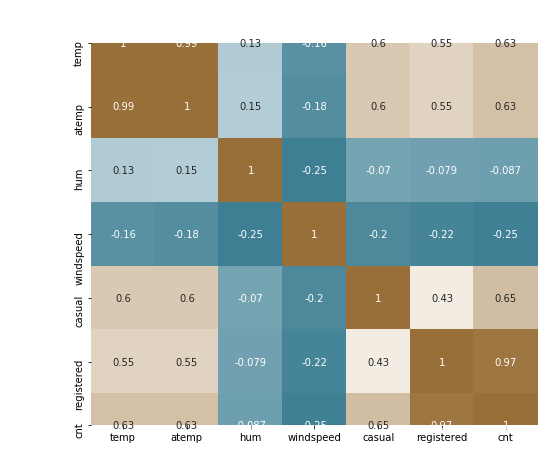
The box plot helps us to identify outliers in each column. In our data, outliers are found in humidity, windspeed and casual columns. Outliers in humidity were imputed with mean of that column (Only 2 outliers). Outliers in windspeed were imputed with median of that column.

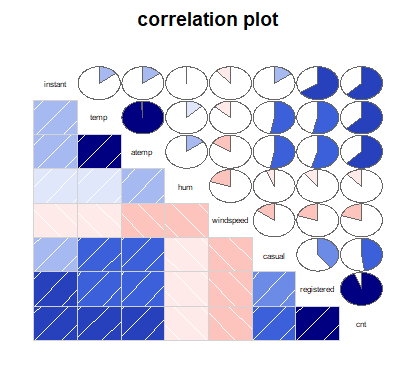
Outliers in casual were not going even after computation also, so I preferred to drop the records with outliers in this variable.

**2.1.5 Feature Selection:** Variable selection is an important aspect of model building. It helps in building predictive models free from correlated variables, biases and unwanted noise. It helps in selecting a subset of relevant features (variables, predictors) for use in model construction and subset of a learning algorithm’s input variables upon which it should focus attention, while ignoring the rest.

**Correlation Analysis**

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. Here each numerical variable’s correlation is mapped with each other’s in a matrix which has been plotted in the following heat map.





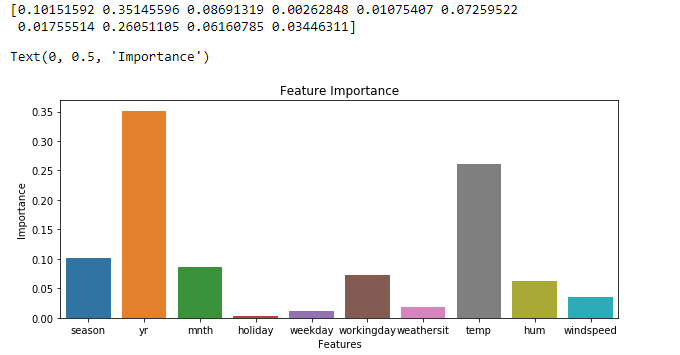
The correlation plot shows how the variables are correlated. Here we can observe that **temp** and **atemp** are highly positively correlated. Also, variable **registered** and **cnt** are highly positively correlated. Thus they can induce Multicollinearity in the model if these variables happens to feed into the model.

Hence it is required to remove the highly correlated variables. I am deleting atemp, registeredas well as casual. It can be observed that casual is not correlated with any variable still I am deleting it, the reason is casual and registered both are target variable which sums up to make an ultimate target variable **cnt**. So I am considering cntas target variable for this problem dataset instead of deriving casual and registered and doing sum to get my result.

So, we are dropping variables('instant','atemp','casual','registered'). Instant is unique for all observations hence has no significance, atemp is strongly correlated with temp, cnt is sum of casual and registration.

* **Feature Importance:** The concept is really straightforward: We measure the importance of a feature by calculating the increase in the model’s prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.
* **Checking via Tree**

Here, we can see the importance of Holiday (0.00262848) and weekday (0.01075407) is extremelylow . We can remove these variables for our dataset as its importance is not much to be considered.



1. **ANOVA** (Analysis of Variances)

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. As our target variable is numerical we will use ANOVA for feature selection technique to see whether any categorical variable is related to target variable. The higher the variance between the variables, the less likely that they are related (or correlated). The result of anova is as follows:

The prprobability value generated by ANOVA test is observed to select whether we should keep a particular variable in our model input or not. Those variables which are having p value less than 0.05 are considered as important predictors and can influence the target variable so they are being selected.

1. For season:

df sum\_sq mean\_sq F PR(>F)

season 1.0 4.517974e+08 4.517974e+08 143.967653 2.133997e-30

Residual 729.0 2.287738e+09 3.138187e+06 NaN NaN

1. For year:

df sum\_sq mean\_sq F PR(>F)

yr 1.0 8.798289e+08 8.798289e+08 344.890586 2.483540e-63

Residual 729.0 1.859706e+09 2.551038e+06 NaN NaN

1. For month:

df sum\_sq mean\_sq F PR(>F)

mnth 1.0 2.147445e+08 2.147445e+08 62.004625 1.243112e-14

Residual 729.0 2.524791e+09 3.463362e+06 NaN NaN

1. For Holiday:

df sum\_sq mean\_sq F PR(>F)

holiday 1.0 1.279749e+07 1.279749e+07 3.421441 0.064759

Residual 729.0 2.726738e+09 3.740381e+06 NaN NaN

1. For Weekday:

df sum\_sq mean\_sq F PR(>F)

weekday 1.0 1.246109e+07 1.246109e+07 3.331091 0.068391

Residual 729.0 2.727074e+09 3.740843e+06 NaN NaN

1. For Working day:

df sum\_sq mean\_sq F PR(>F)

workingday 1.0 1.024604e+07 1.024604e+07 2.736742 0.098495

Residual 729.0 2.729289e+09 3.743881e+06 NaN NaN

1. For Weathersit:

df sum\_sq mean\_sq F PR(>F)

weathersit 1.0 2.422888e+08 2.422888e+08 70.729298 2.150976e-16

Residual 729.0 2.497247e+09 3.425578e+06 NaN NaN

Ho = Target variable is Independent from Categorical variable

Ha = Target variable is Dependent on the Categorical variable

From the above result we can see that only four variables are very much related to target variable hence we delete all the other variables.

Therefore from both the correlation analysis and ANOVA we got some variable which we shouldn’t consider for further processing. The variables that could be deleted are as

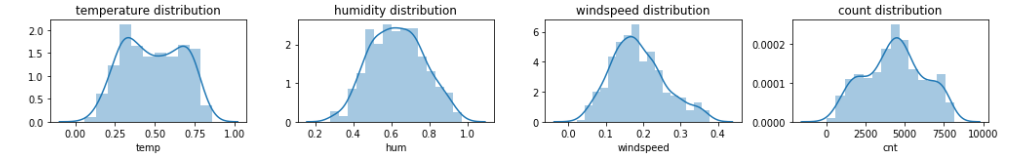
Numerical: Instant, atemp, casual, registered

Categorical: dteday, holiday, workingday, weekday

Hence, we have deleted these Variables from our data set.

**2.1.6 FeatureScaling:Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

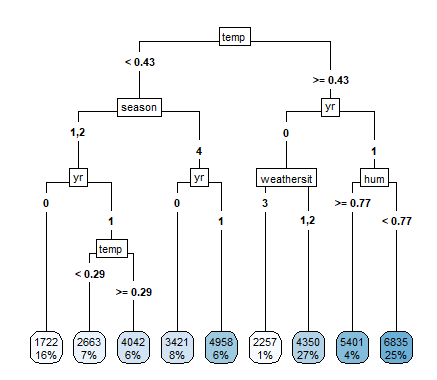
Normalization brings all of the variables into proportion with one another. It transforms data into a range between 0 and 1. All our continuous variables are already normalized except the target variable which we prefer not to scale because its variation is spread quite widely and after scaling, the difference between the numbers is diminishing.



**2.2 Modelling**

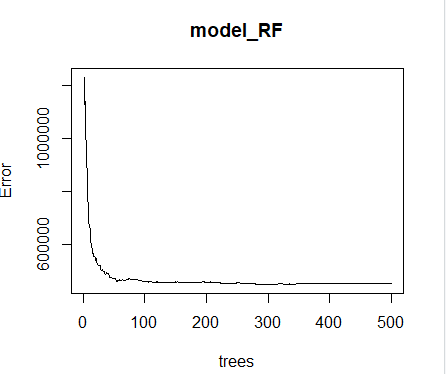
2.2.1 **ModelSelection:**For modelling, we are going to use some famous models to our data-set and will conclude the result according to it.

1. **Decision Tree:** Decision tree is a rule. 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. Split of decision tree is seen in the below tree. Decision tree regression is as follows



**2) Random Forest:**Random Forest or decision tree forests are an ensemble learning method for classification, regression and other tasks. It consists of an arbitrary number of simple trees, which are used to determine the final outcome. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases). The goal of using a large number of trees is to train enough that each feature has a chance to appear in several model.

As we increase the number of trees the error count decrease until a point and then becomes constant. Error vs number of trees to be used graph is as follows:



Call:

randomForest(formula = cnt ~ ., data = data\_train, importance = TRUE, ntree = 500)

Type of random forest: regression

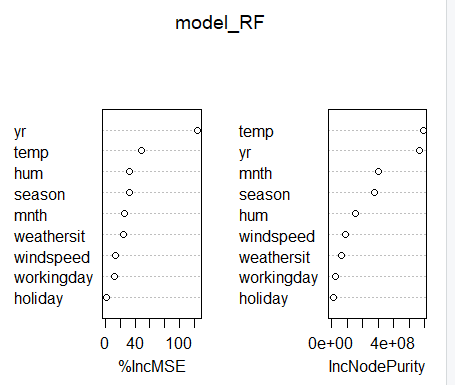
Number of trees: 500

No. of variables tried at each split: 3

Mean of squared residuals: 457475.8

% Var explained: 87.65

We can check the importance of our variables in Random Forest Model with (varImpPlot(MODELNAME))



The first graph shows that if a variable is assigned values by random permutation by how much will the MSE increase. Higher the value, higher the importance. On the other hand, node purity is measured by the Gini index which is the difference between before and after split on that variable.

**Linear Regression:**Linear regression is the most basic type of regression and commonly used predictive analysis. Linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Following is the summary of the Linear model:

Call:

lm(formula = cnt ~ ., data = data\_train)

Residuals:

Min 1Q Median 3Q Max

-2949.1 -405.0 70.6 475.6 3350.2

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1642.02 267.06 6.149 1.51e-09 \*\*\*

season2 958.20 205.00 4.674 3.72e-06 \*\*\*

season3 897.09 244.71 3.666 0.000270 \*\*\*

season4 1563.30 202.31 7.727 5.26e-14 \*\*\*

yr1 2031.97 67.14 30.265 < 2e-16 \*\*\*

mnth2 111.49 159.92 0.697 0.485987

mnth3 562.94 188.30 2.990 0.002918 \*\*

mnth4 392.67 280.57 1.400 0.162220

mnth5 725.57 301.90 2.403 0.016577 \*

mnth6 592.64 319.72 1.854 0.064335 .

mnth7 85.93 359.88 0.239 0.811364

mnth8 474.64 344.94 1.376 0.169380

mnth9 1109.30 300.15 3.696 0.000241 \*\*\*

mnth10 712.81 269.98 2.640 0.008521 \*\*

mnth11 -36.09 260.02 -0.139 0.889658

mnth12 42.90 203.46 0.211 0.833095

holiday1 -521.30 192.14 -2.713 0.006873 \*\*

workingday1 228.49 75.34 3.033 0.002537 \*\*

weathersit2 -439.11 89.20 -4.923 1.13e-06 \*\*\*

weathersit3 -1734.64 234.64 -7.393 5.39e-13 \*\*\*

temp 4232.31 463.16 9.138 < 2e-16 \*\*\*

hum -1626.23 350.58 -4.639 4.39e-06 \*\*\*

windspeed -2339.27 504.59 -4.636 4.44e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 779.2 on 549 degrees of freedom

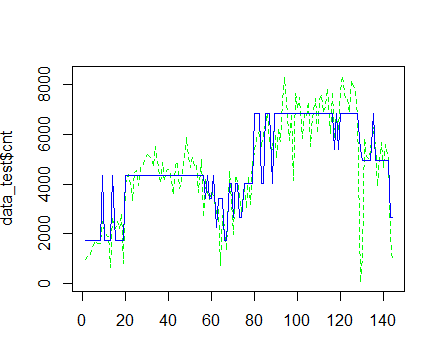
Multiple R-squared: 0.8427, Adjusted R-squared: 0.8364

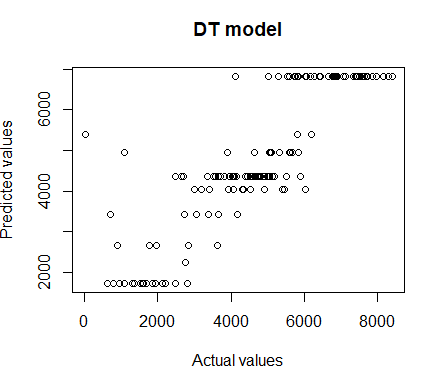
F-statistic: 133.7 on 22 and 549 DF, p-value: < 2.2e-16

**2.2.2Visualizing models** :We can see the plots of our predicted model to understand it better

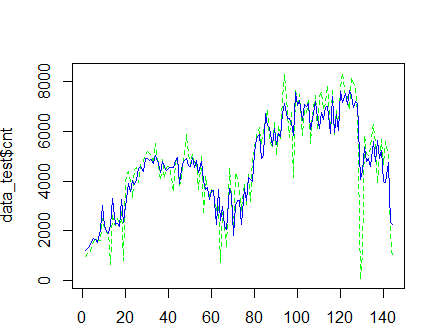
A) Prediction Plots:

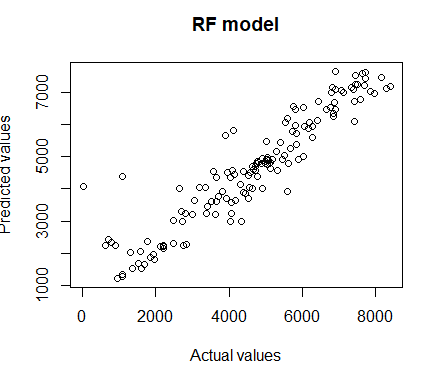
* **Decision Tree:**



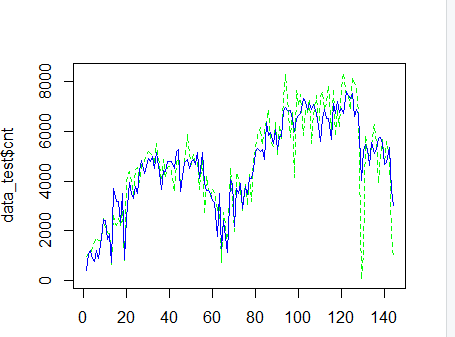


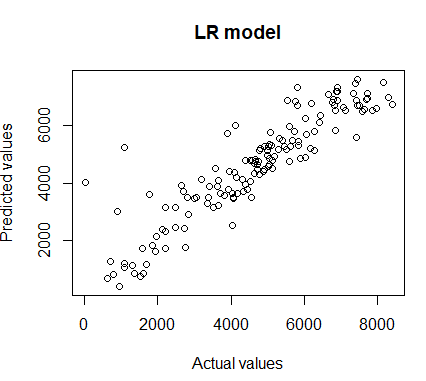
* **Random Forest:**





* **Linear regression:**

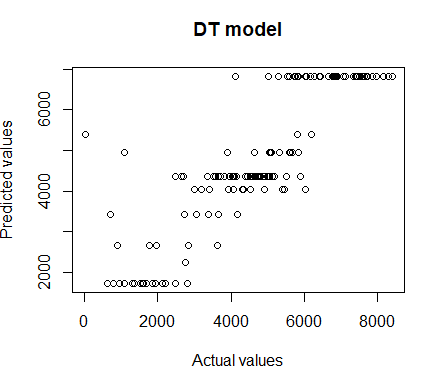


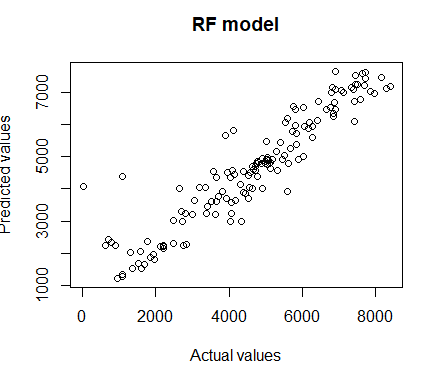


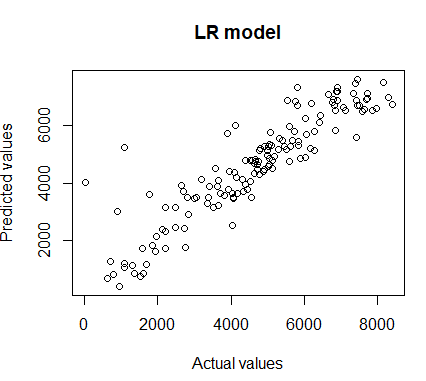
1. **Conclusion**
   1. **Model Evaluation:**Model evaluation is done on basis of evaluation metrics or error metrics. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results. Simply, building a predictive model is not our motive. But, creating and selecting a model which gives high accuracy on out of sample data. Hence, it is crucial to check accuracy or other metric of the model prior to computing predicted values. In our data as we applied regression models we have error metrics like Mean square error(MSE), MAPE, Root mean square error (RMSE), Mean absolute error (MAE).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Language/Model** |  | **Python** |  | |  | **R** | |  |  |
| **MODELS** | MSE | RMSE | MAPE | R-SQ | MAE | | RMSE | MAPE | R-SQ |
| **Decision Tree** | 744168 | 862 | 18.70 | 0.7529 | 703.46 | | 988.5 | 193 | 0.75 |
| **Random Forest** | 531050 | 728 | 16.05 | 0.8237 | 502.64 | | 755.97 | 142 | 0.86 |
| **Linear regression** | 715199 | 845 | 19.18 | 0.7620 | 574.70 | | 839.72 | 143 | 0.81 |

* 1. **Model Selection:**We can see that all models perform comparatively on average and therefore we select Random forest classifier models for better prediction.

******

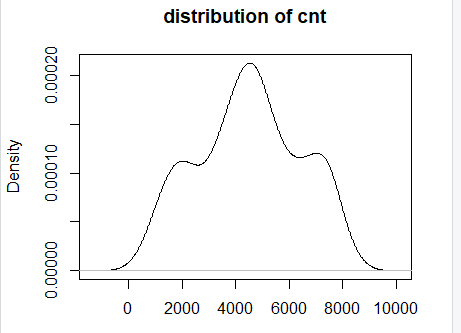




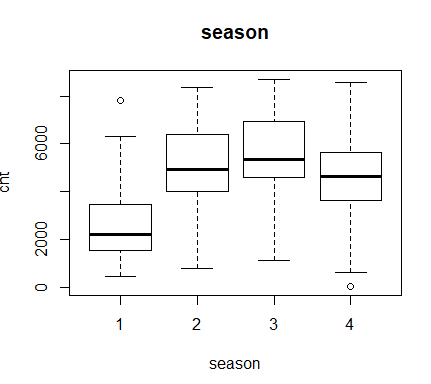
From the above plots of Actual Vs Predicted values, we can infer that values of Random forest falls on straight line indicating random forest fits better than the other three models. Also amongst the three models, Random forest has best R-sq. (Coef. of determination). Hence we’ll fix Random Forest as our model.

**Appendix A -Imp Plots**

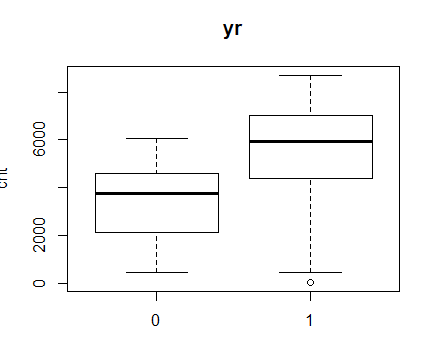
1. Distribution of Count:

****

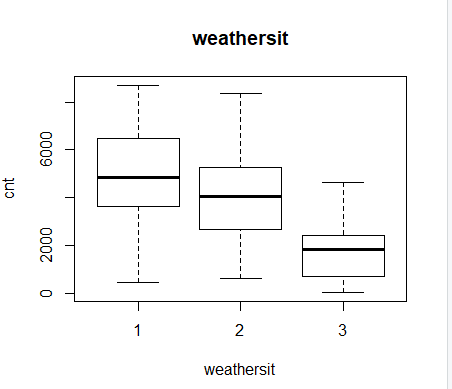
1. Box plot for Season Vs cnt



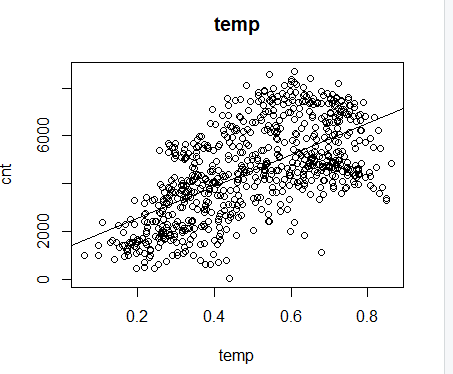
1. **Box plot for Year Vs Cnt**



1. Weathersit Vs Cnt



1. Temp Vs Cnt



1. Registered Vs cnt

