2019

PavanKumar BL

5/5/2019

**Bike Rental Count**



Contents

[Chapter 1 2](#_Toc7960164)

[Problem Statement 2](#_Toc7960165)

[Data 2](#_Toc7960166)

[Chapter 2 4](#_Toc7960167)

[Methodology 4](#_Toc7960168)

[Pre-Processing 4](#_Toc7960169)

[Missing Value Analysis 6](#_Toc7960170)

[Data Understanding 6](#_Toc7960171)

[Number Summary of the Bike Rental Count 'cnt' Feature 6](#_Toc7960172)

[Quantitative Features vs. Rental Counts 7](#_Toc7960173)

[Let’s Explore on Categorical Variable 8](#_Toc7960174)

[Outlier Analysis 14](#_Toc7960175)

[Feature Selection 18](#_Toc7960176)

[Feature Scaling 19](#_Toc7960177)

[Dummy Variables 21](#_Toc7960178)

[Chapter 3 22](#_Toc7960179)

[Modeling 22](#_Toc7960180)

[Model Selection 22](#_Toc7960181)

[Methodology: 22](#_Toc7960182)

[Hold-Out Method 22](#_Toc7960183)

[R2 score 22](#_Toc7960184)

[Model Building 23](#_Toc7960185)

[Decision Tree 23](#_Toc7960186)

[Random Forest 23](#_Toc7960187)

[Linear Regression 23](#_Toc7960188)

[Gradient boosting 24](#_Toc7960189)

[Chapter 4 24](#_Toc7960190)

[Conclusion 24](#_Toc7960191)

[Model Evaluation 24](#_Toc7960192)

[Model Selection 25](#_Toc7960193)

[Appendix A: Extra Figures 26](#_Toc7960194)

## Chapter 1

## Problem Statement

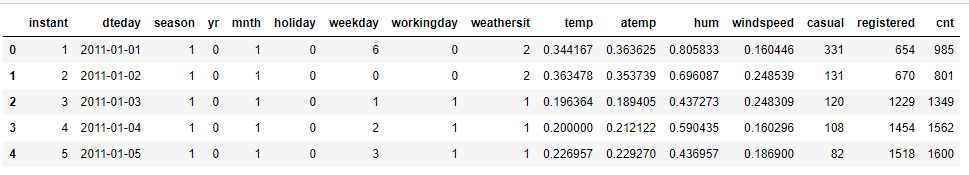
The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

## Data

**Bike Rental** dataset was provided for analysis. Data contains 15 predictor variables and 1 target variable.

|  |  |
| --- | --- |
| Variables | Description |
| Instant | Record Index |
| Dteday | Date |
| 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) |
| Temp | Normalized temperature in Celsius |
| Atemp | Normalized feeling temperature in Celsius. |
| Hum | Normalized humidity |
| Windspeed | Normalized wind speed |
| Casual | count of casual users |
| Registered | count of registered users |
| Cnt | count of total rental bikes including both casual and registered |

**Size of Dataset Provided**: - 731 rows, 16 Columns



## Chapter 2

## Methodology

Bike rental Count is a Project where we extract the total number of people who rent a Bike daily based on Weather condition.

**Exploratory Data Analysis (EDA)-** It includes following steps

Looking into the data and analyzing all variables

* Visualization
* Missing Value Analysis
* Outlier Analysis
* Correlation analysis
* Feature Scaling
* Dummy data creation
* Feature Sampling.

**Basic Modeling-** Trying different models over preprocessed data

* Decision Tree
* Random forest
* Linear regression
* Gradient Boosting

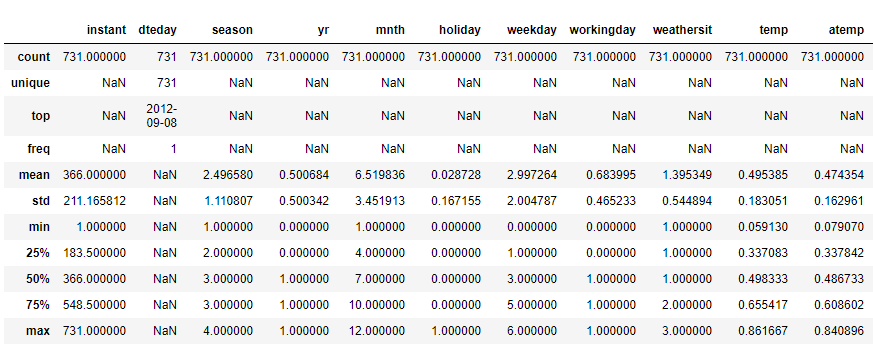
**Model Evaluation & Optimization-** Evaluating model performances and then selecting the best model fit for our data, optimizing hyper parameters tuning and cost effectiveness of model. This step is optional. We may or may not involve it. It is basically done to avoid a scenario where the selected approach works very well with training data but fails to support out test data in similar way.

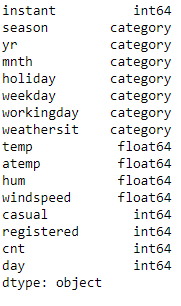
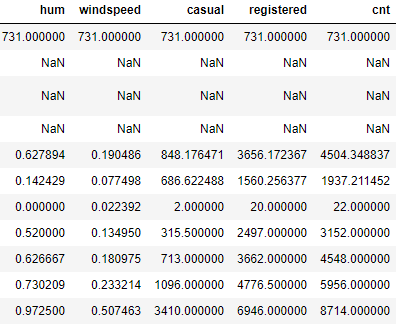
**Implementation model on Final test data and saving the results**

## Pre-Processing

Any predictive modeling 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.





From above details we can confirm that

* Data looks fine.
* From Dteday attribute we will have to extract the day
* Instant variable can be discared from processing since it convey no info.
* Atrributes are converted to proper data types.

## Missing Value Analysis

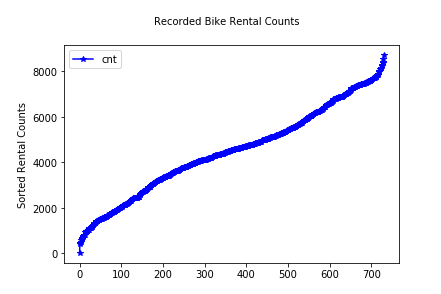
In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value either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

|  |  |
| --- | --- |
| **Variables** | **Missing values** |
| **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 |

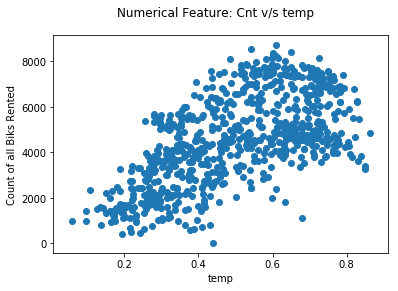
## Data Understanding

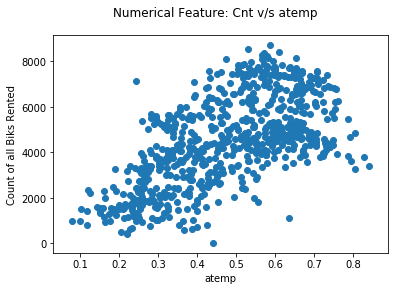
In order to get further insight and understand the data set and to see how different features interact with each other and the target. First the amount of bike rental counts for each day of the week is analyzed.

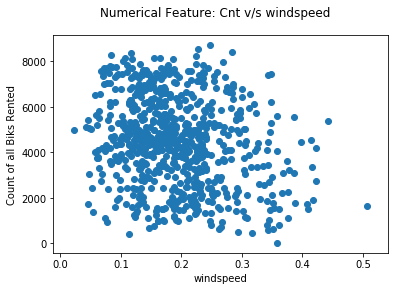
## Number Summary of the Bike Rental Count 'cnt' Feature



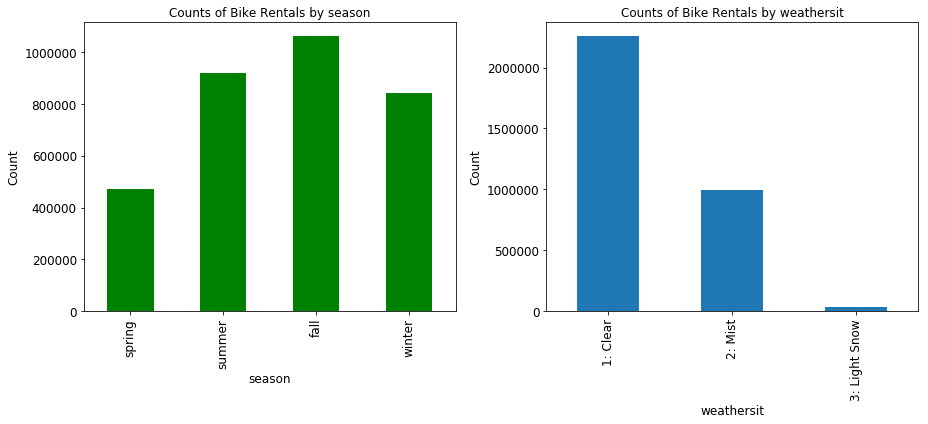
## Quantitative Features vs. Rental Counts



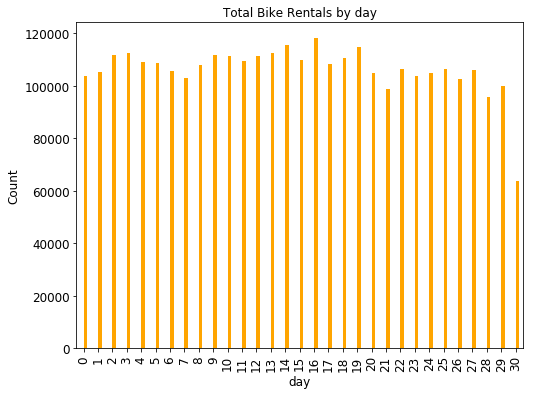


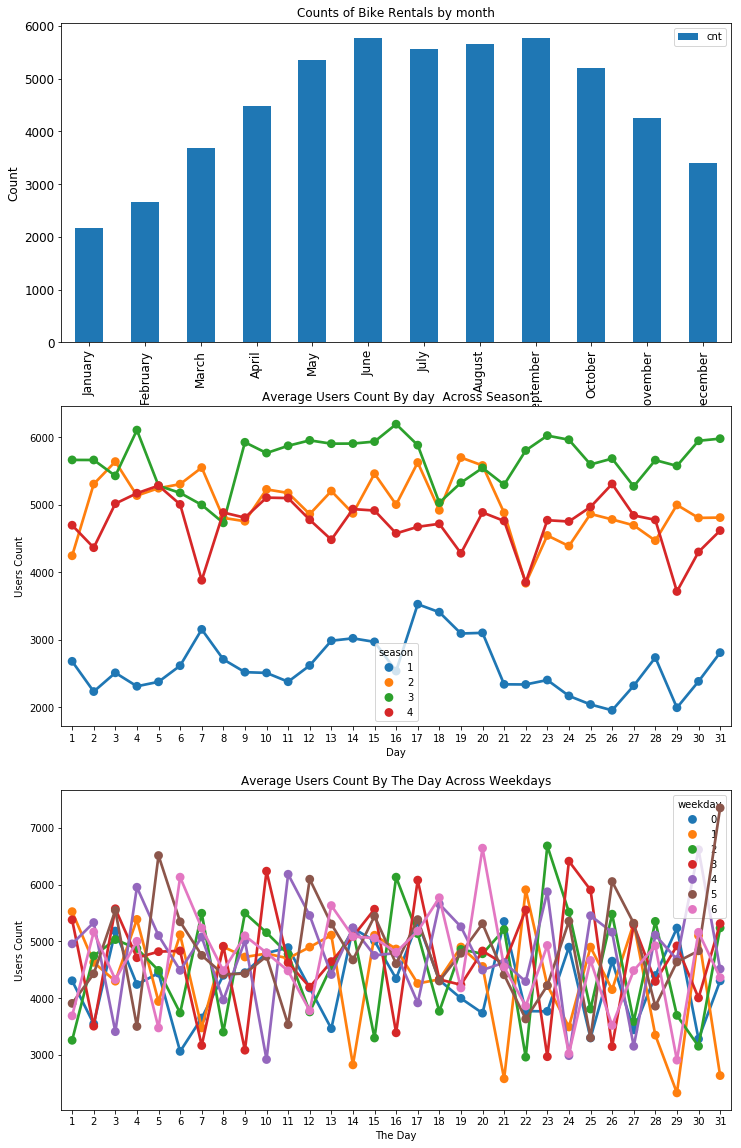


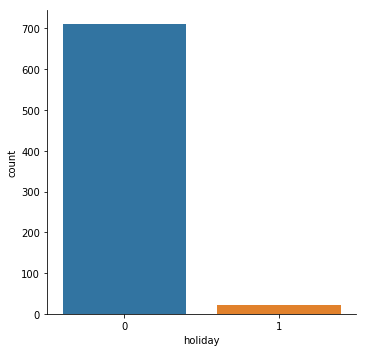
## Let’s Explore on Categorical Variable

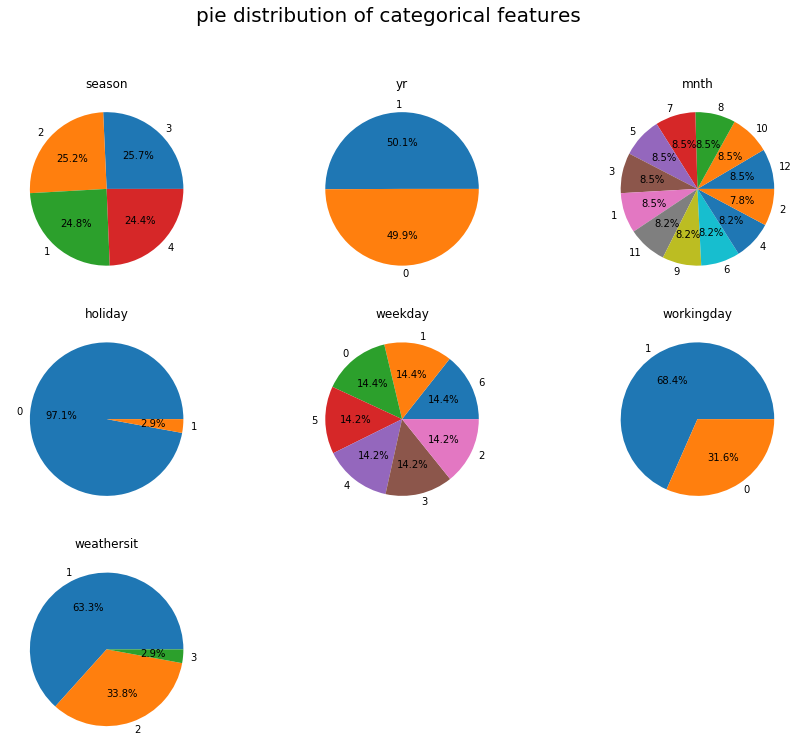


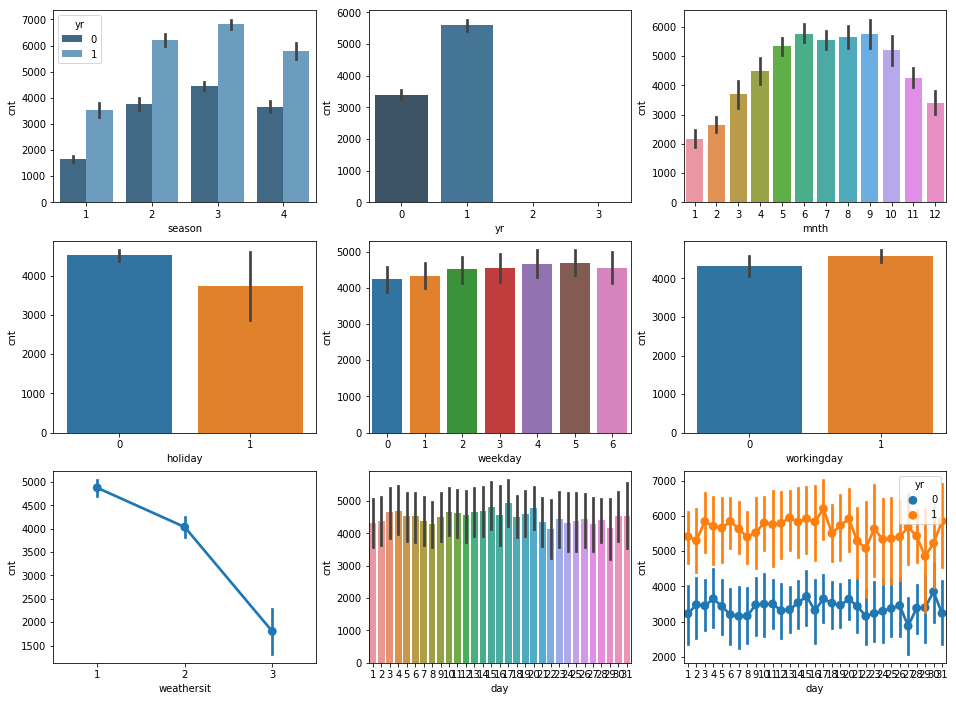
its observed from above plot that Bike rent count is high in Fall season and clear weather







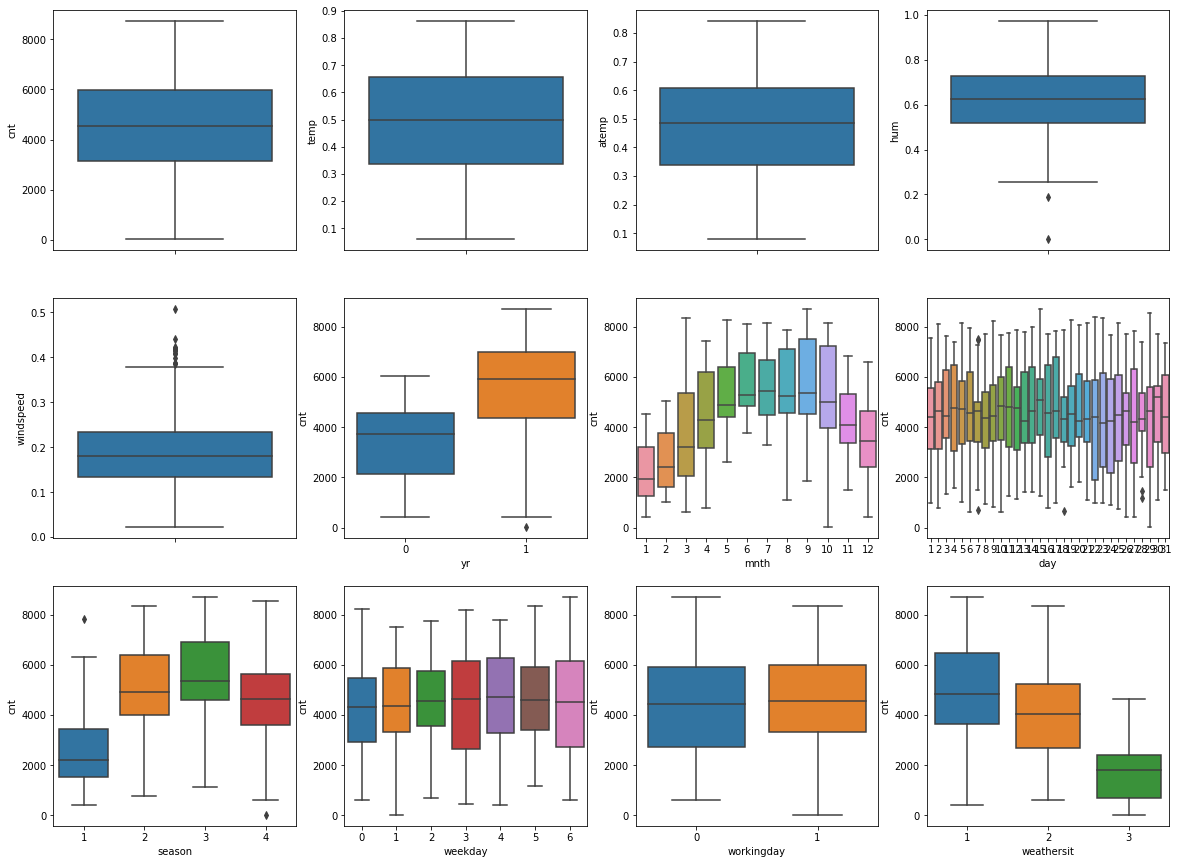


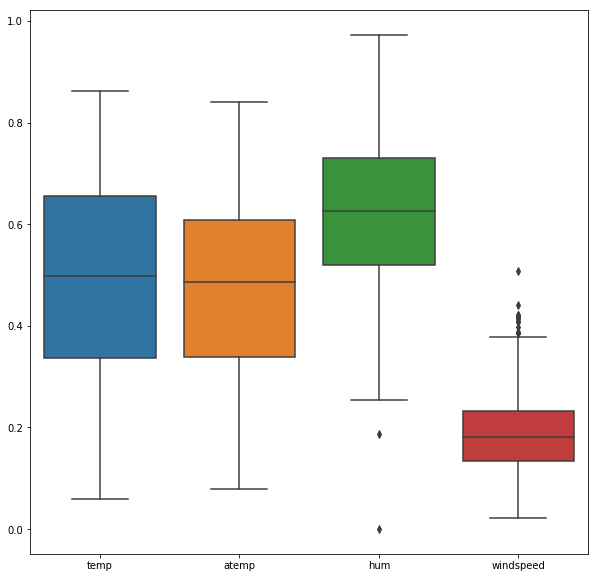


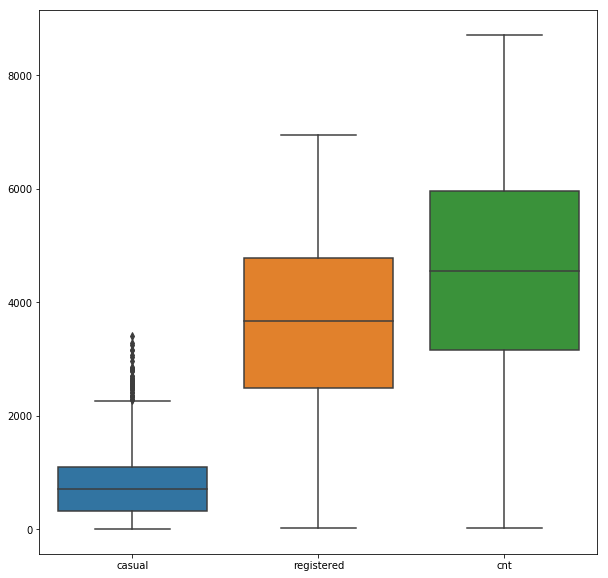
## Outlier Analysis

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

* We visualize the outliers using boxplots.
* It is observed that variables **Casual, Hum and Windspeed has Outliers.**
* **Capping** is done for Outliers Treatment.





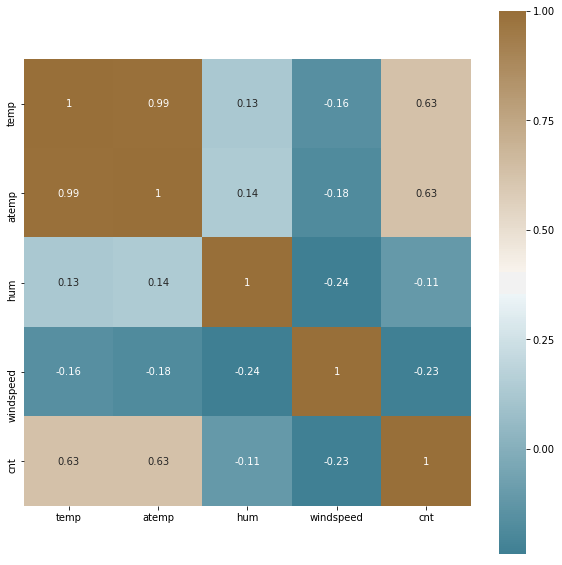


## Feature Selection

Before performing any type of 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. Selecting subset of relevant columns for the model construction is known as **Feature Selection**. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead, we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis & VIF** for numerical variable and **ANOVA** (Analysis of variance) for categorical variable.

**VIF**

In statistics, the variance inflation factor (VIF) is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least square’s regression analysis.



From correlation analysis and ANOVA test we have found that

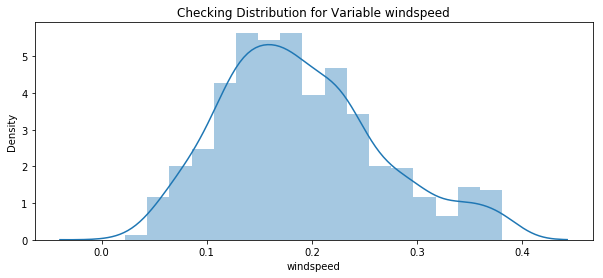
* ‘**temp’** and ‘**atemp’** have high correlation (>0.7), so we have excluded the **atemp** column.
* **'holiday'**, **'weekday'** and **'workingday'** have p>0.05 and hence were excluded.

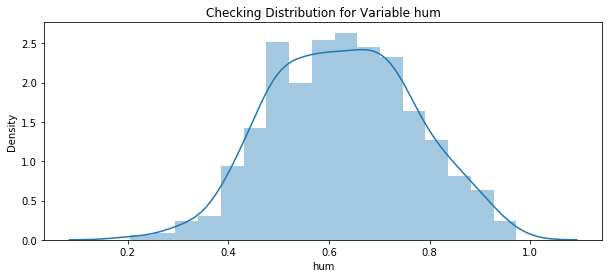
## Feature Scaling

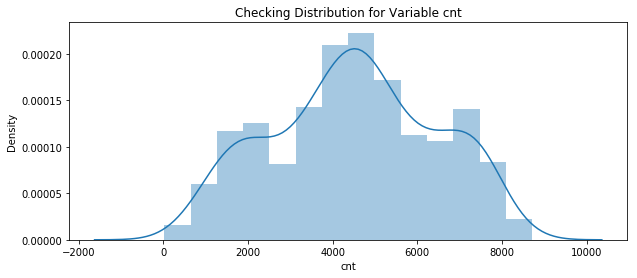
**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine

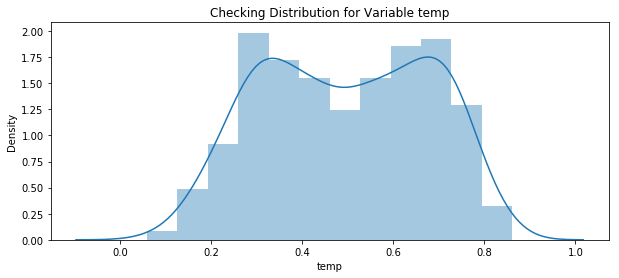
learning algorithms, objective functions will not work properly without normalization. For example,

most 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 feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.









Since our data is uniformly distributed we will use Standardization in this step as Feature Scaling Method.

## Dummy Variables

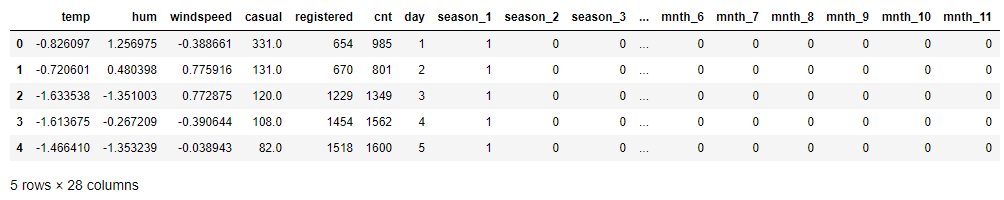
A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. Dummies are any variables that are either one or zero for each observation. pd.get\_dummies when applied to a column of categories where we

have one category per observation will produce a new column (variable) for each unique categorical value. It will place a one in the column corresponding to the categorical value present for that observation.

This is equivalent to one hot encoding.

One-hot encoding is characterized by having only one per set of categorical values per observation.

Viewing data after adding dummy variables:



## Chapter 3

## Modeling

After a thorough preprocessing we will use some regression models on our processed data to predict the target variable.

## Model Selection

It has been noted in previous stages of our analysis that for different combinations of the independent variables, the count is different. The dependent variable is a continuous variable and hence the type model of model that would be developed for this problem is a regression model.

## Methodology:

Model Evaluation is an integral part of the model development process as it helps us find the best model for representing our data. It also helps to evaluate as to how it would on new data. In order to develop an efficient and accurate model to predict our target variable we shall use a combination of three different methods, the three different methods that can be used are given below.

* Hold-Out Method
* R2 Score

## Hold-Out Method

As evaluating model performance on training data set may lead to develop an over fitted model. Due to this is required to test the model on a separate data set. Hence the original data set is split into training and testing data. The training data set is used to build a predictive model and the testing data is used to evaluate the model performance.

## R2 score

R2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data.

## Model Building

We Start building our model by using the following models-

## 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. Split of decision tree is seen in the below tree. The MAPE, RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **R** | **PYTHON** |
| **MAPE** | 16.0869 | 43.4090739035 |
| **RMSE** | 583.8309749 | 151.39529301 |
| **R^2** | 0.9179334 | 0.9374478583 |

## 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. The MAPE, RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **R** | **PYTHON** |
| **MAPE** | 5.845628 | 0.566347094345 |
| **RMSE** | 226.92656 | 4.743890095112 |
| **R^2** | 0.9900862 | 0.999938583 |

## Linear 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. The MAPE, RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **R** | **PYTHON** |
| **MAPE** | 0.883809 | 9.19850423710725e-14 |
| **RMSE** | 71.5507857 | 3.1389314998647157e-13 |
| **R^2** | 0.9987668 | 1.0 |

## Gradient boosting

**Gradient boosting** is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. The MAPE, RMSE value and R^2 value for our project in R and Python are –

|  |  |  |
| --- | --- | --- |
| **Gradient boosting** | **R** | **PYTHON** |
| **MAPE** | 2.826727 | 0.978090059671626 |
| **RMSE** | 117.0507324 | 5.802180672253785 |
| **R^2** | 0.9966773 | 0.999908124453814 |

## Chapter 4

## Conclusion

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

## Model Evaluation

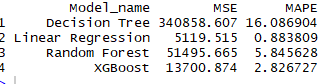
In the previous chapter we have seen the **Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE)** and **R-Squared** Value of different models.

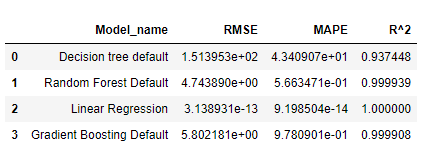
**Root Mean Square Error** (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

Whereas **R-squared** is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable.

The **mean absolute percentage error (MAPE)**, also known as **mean absolute percentage deviation (MAPD)**, is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a Loss function for regression problems in Machine Learning.

Lower values of **RMSE** and **MAPE** and higher value of **R-Squared** Value indicate better fit.



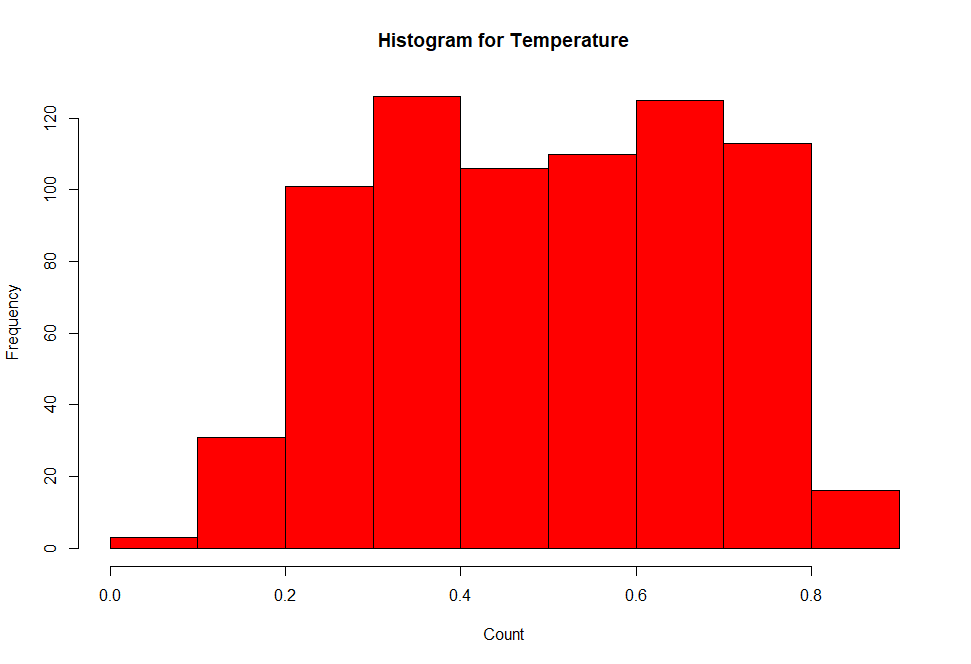


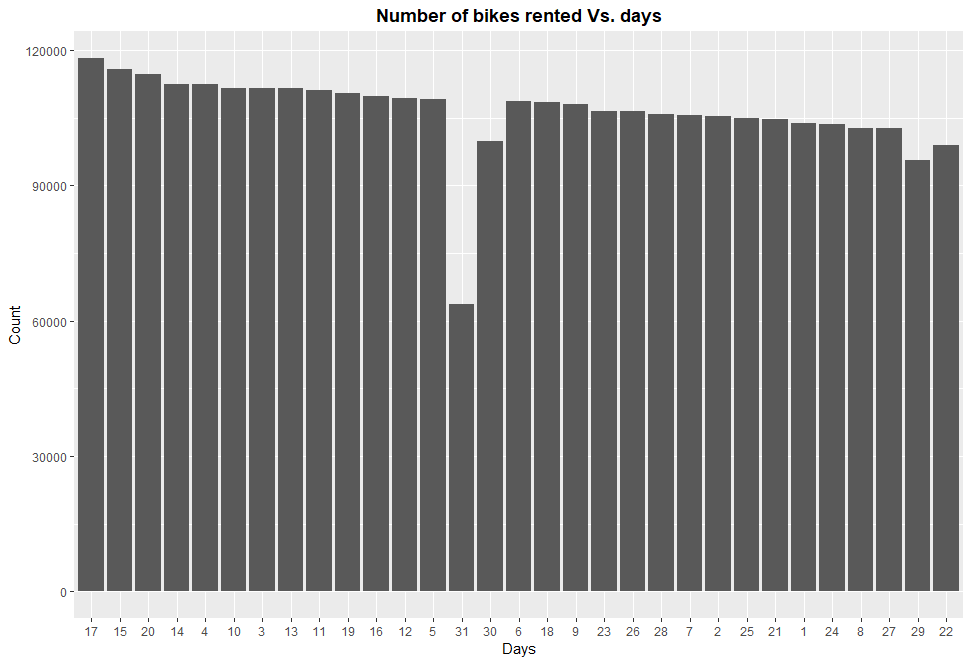
## Model Selection

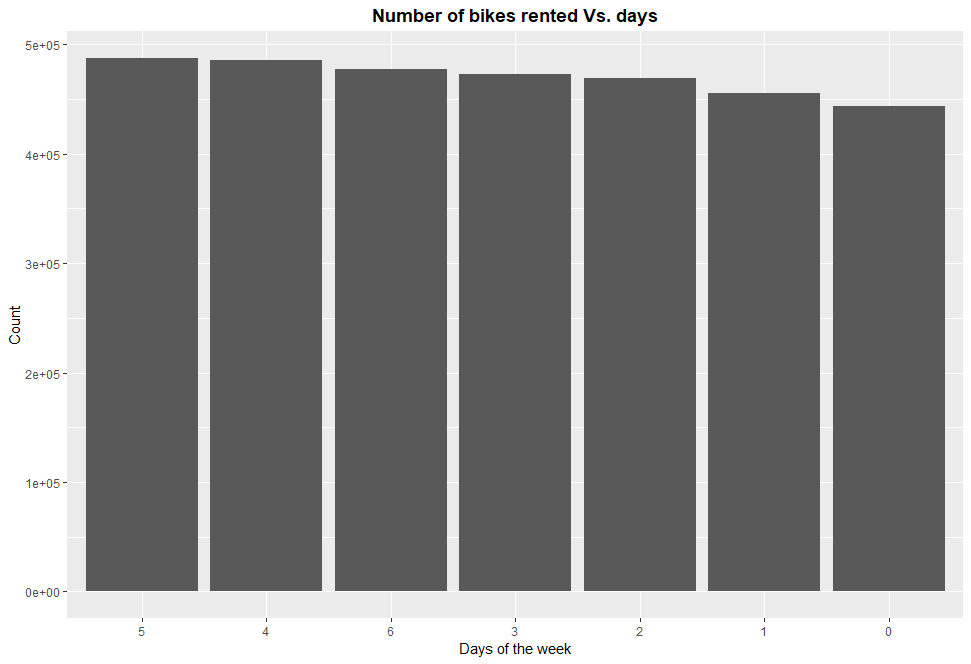
From the observation of all **MAPE, MSE Value** and **R-Squared** Value we have concluded that, Both the models- **Gradient Boosting Default ,Linear regression** and **Random Forest** perform comparatively well while comparing their MSE, R-Squared value and MAPE.

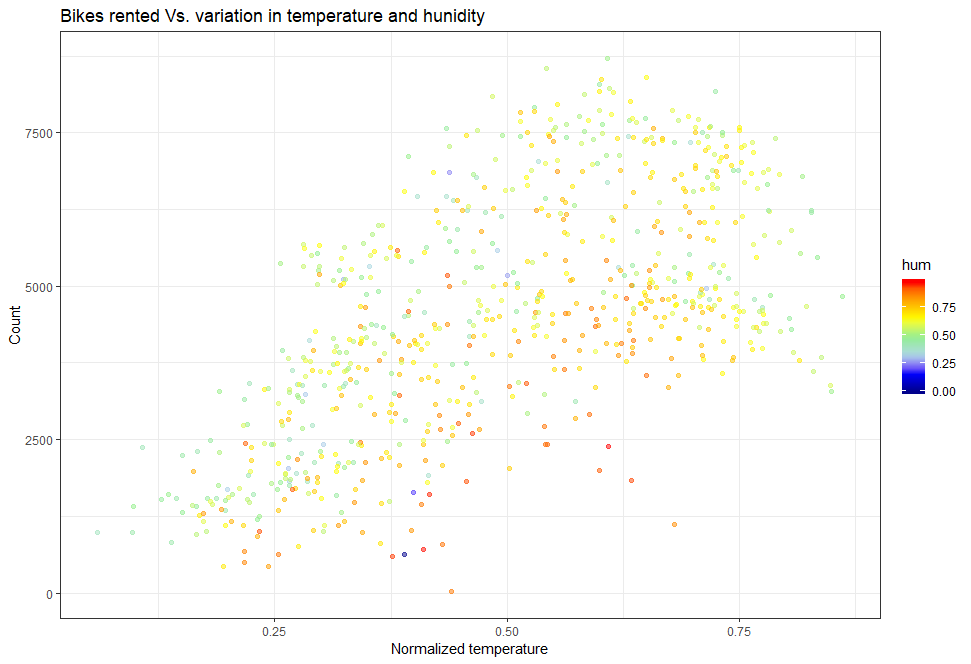
After this, I chose **Linear regression** as a method based on the R2 Score.

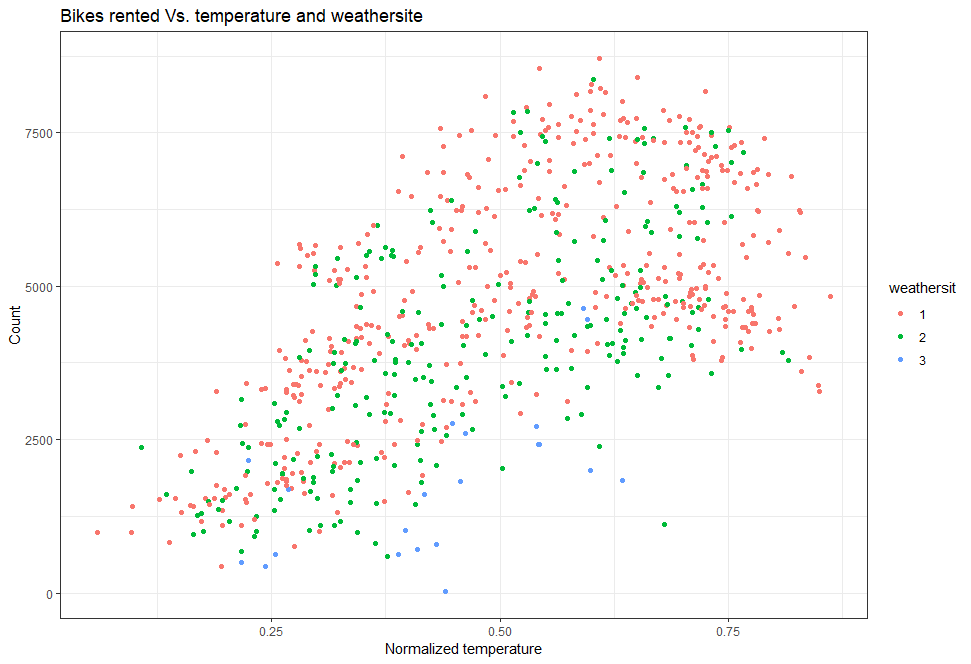
## Appendix A: Extra Figures

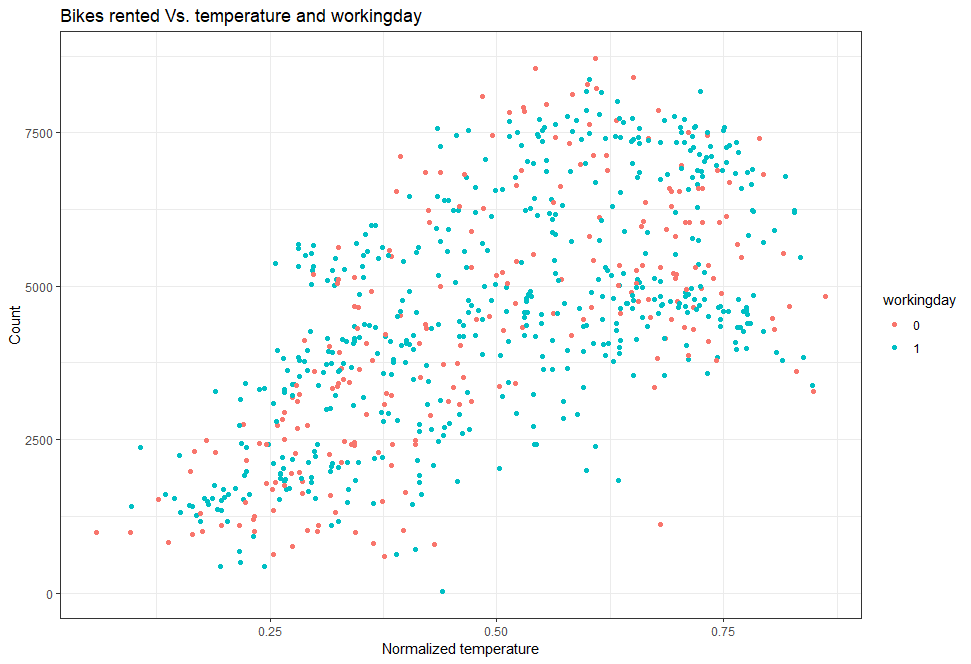


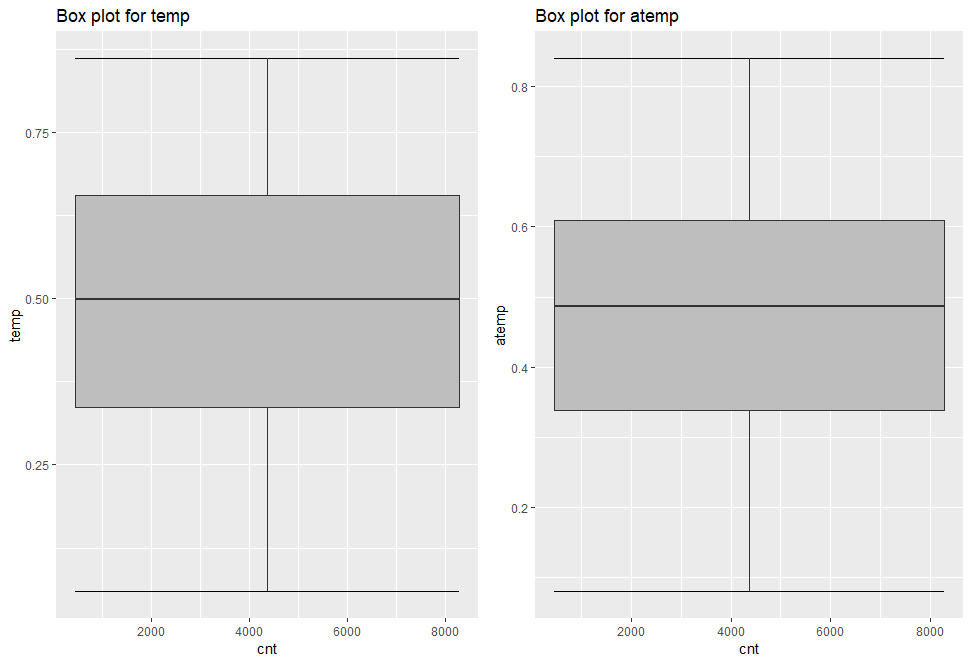


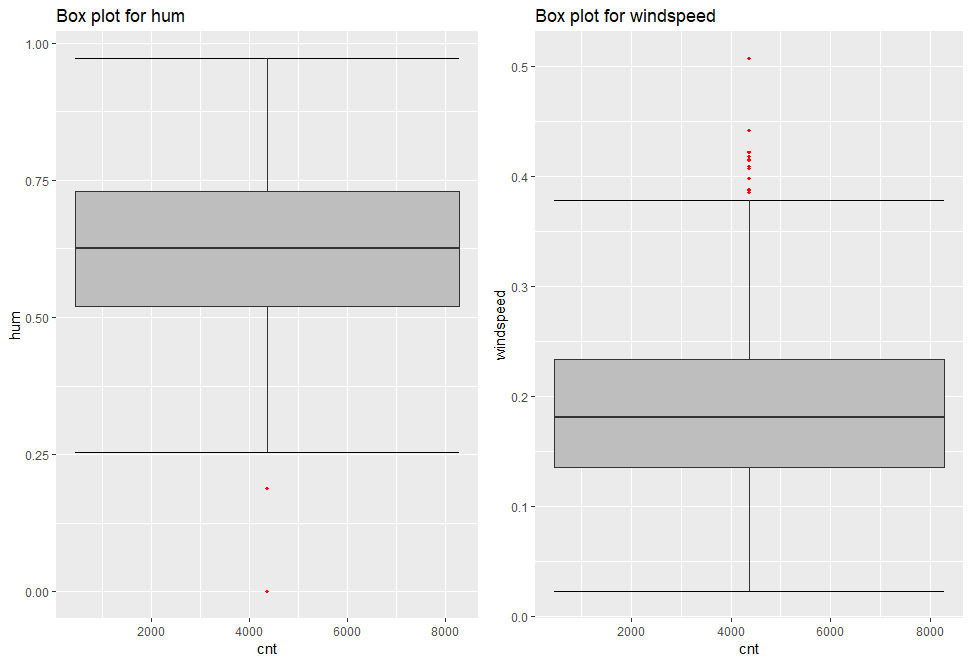


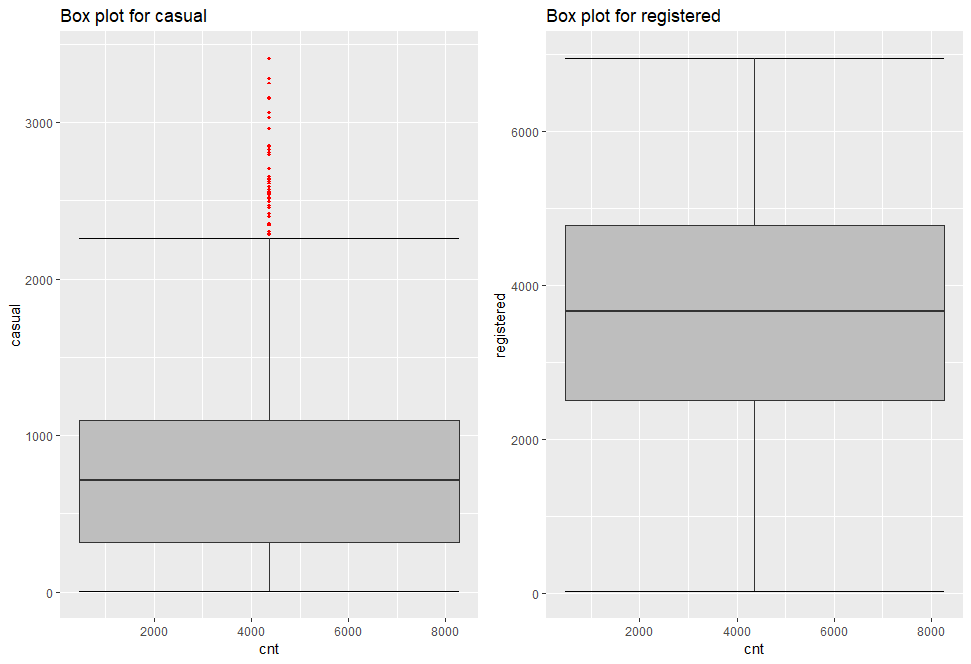


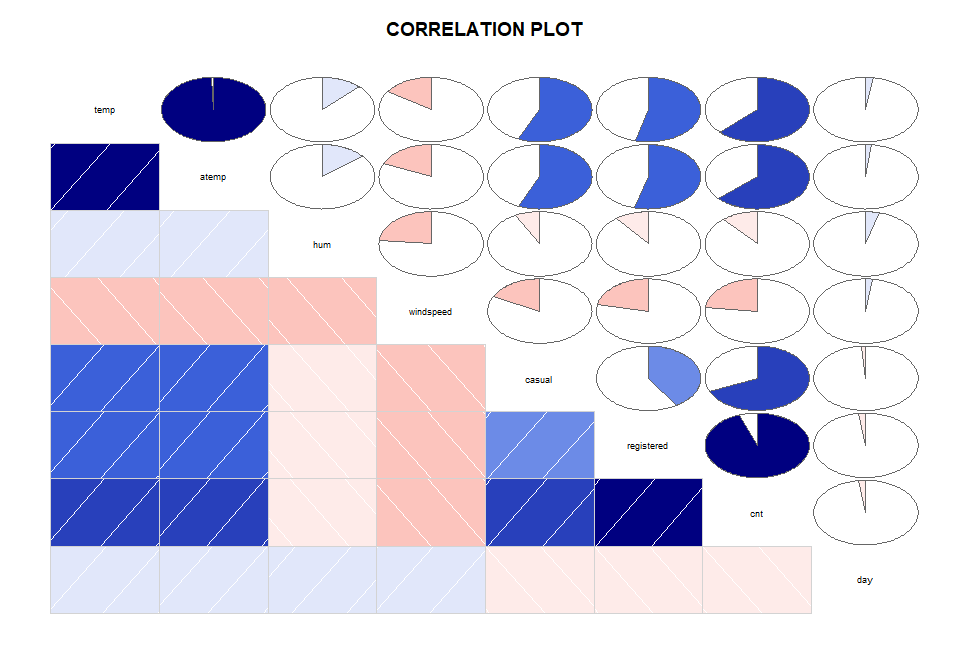


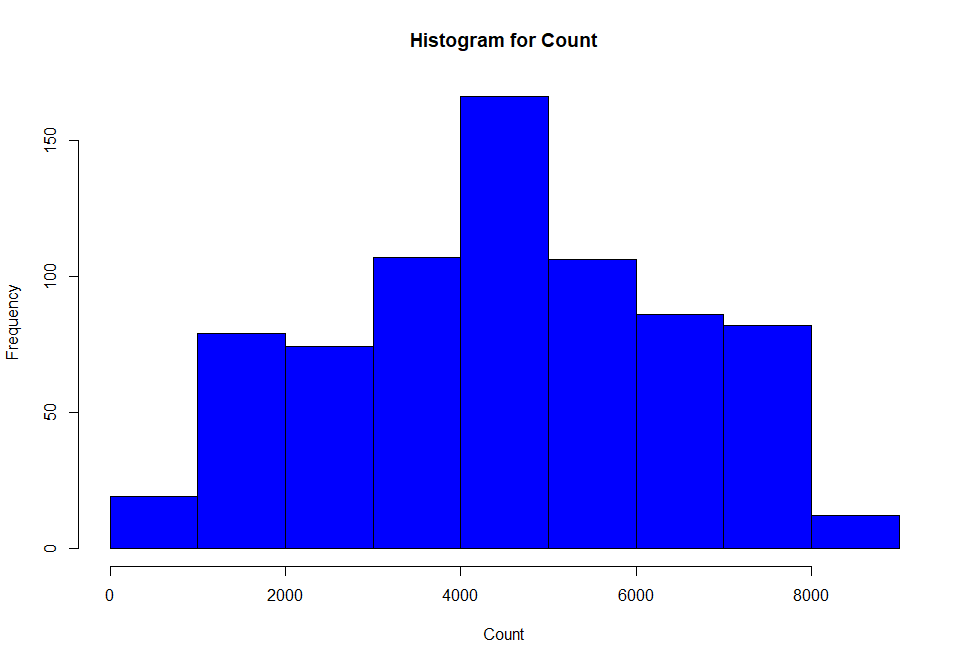


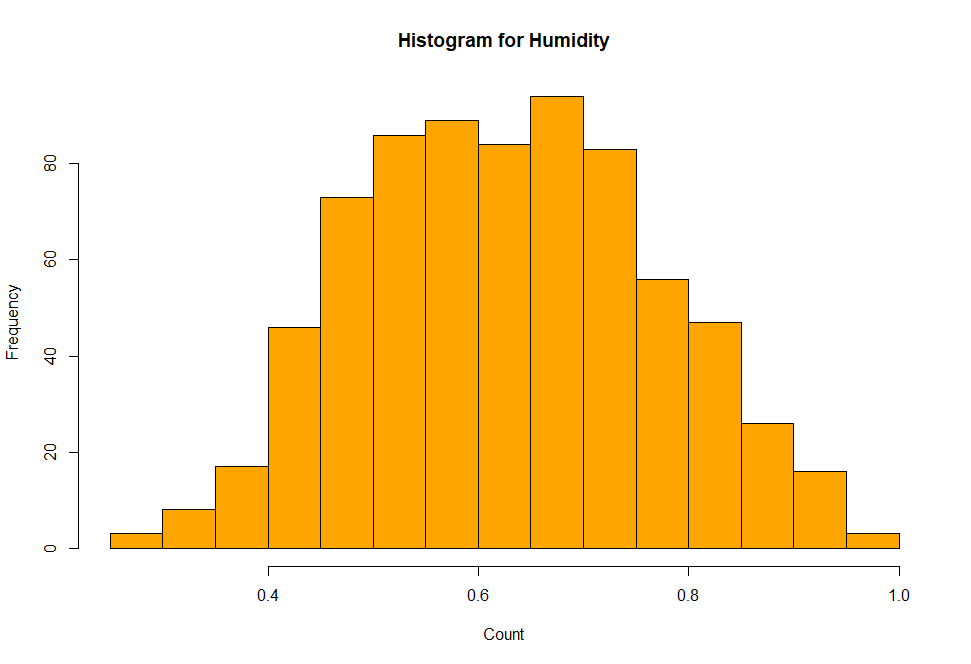












Feature selection

