Predicting Bike Rental Count

By :- Rahul Prasad Sah

Contents:-

Introduction

Problem Statement:-

The objective of the project is to predict bike rental on daily, based on environmental and seasonal setting.

Data:-

The details of data attribute in the dataset as follows :-

- 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)
- hr: Hour (0 to 23)
- holiday: weather day is holiday or not (extracted fromHoliday Schedule)
- weekday: Day of the week
- workingday: If day is neither weekend nor holiday is 1, otherwise is 0.
- 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

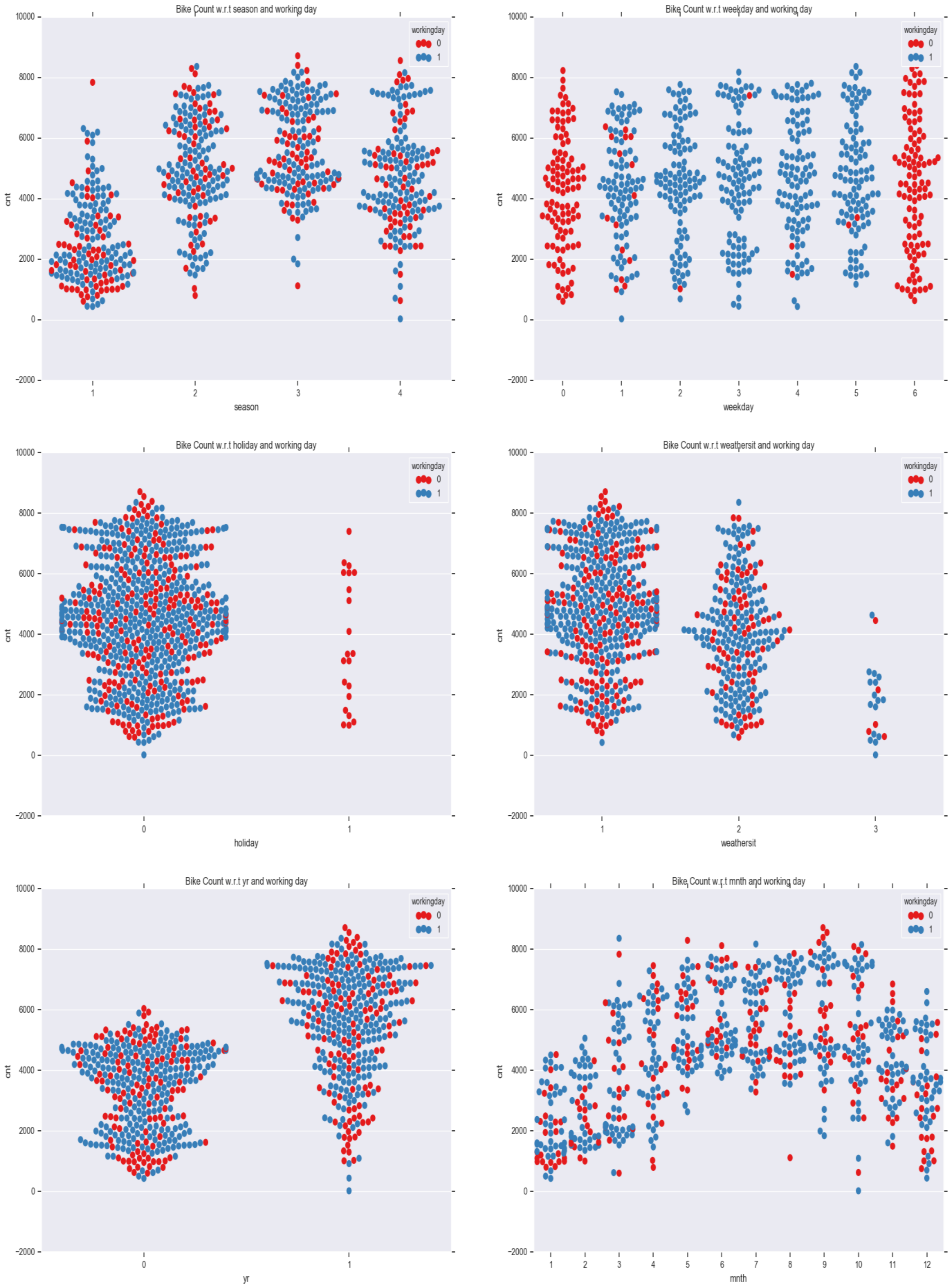
- temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min),t_min=-8, t_max=+39 (only in hourly scale).
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_maxt_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized 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: count of registered users
- cnt: count of total rental bikes including both casual and registered.

Methodology

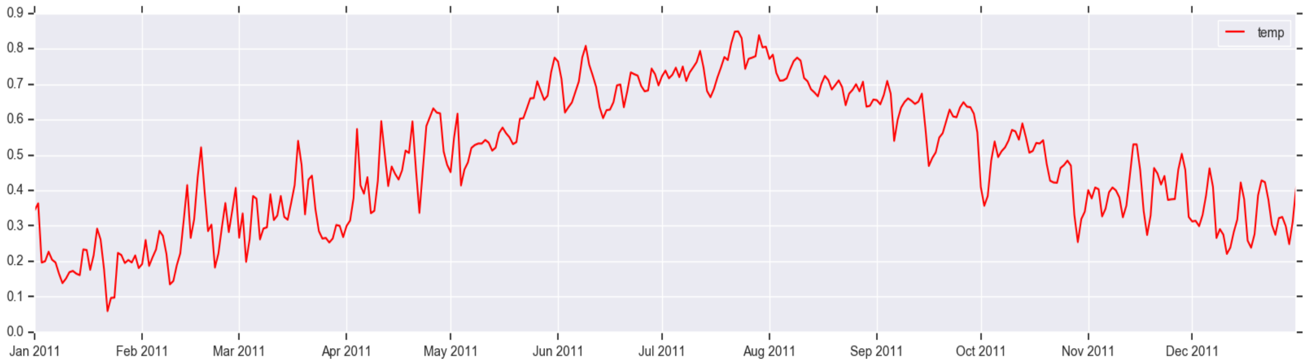
- Data Pre-Preprocessing :-
- Data pre-processing is the first stage of any type of project. We do this by looking at plots of independent variables vs target variables. If the data is messy, we try to improve it by sorting deleting extra rows and columns. This stage is called as Exploratory Data Analysis. This stage generally involves data cleaning, merging, sorting, looking for outlier analysis, looking for missing values in the data, Imputing missing values if found by various methods such as mean, median, mode, KNN imputation, etc.

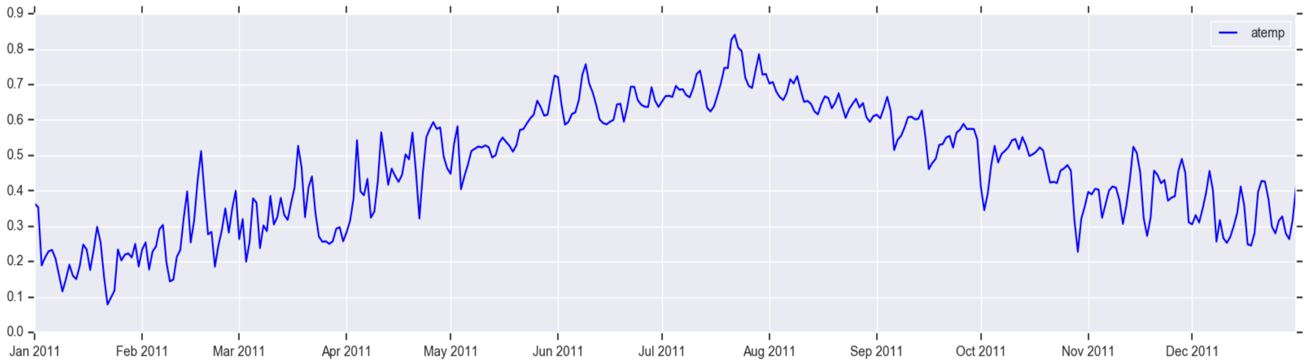
- Analysis of data via visualization:-

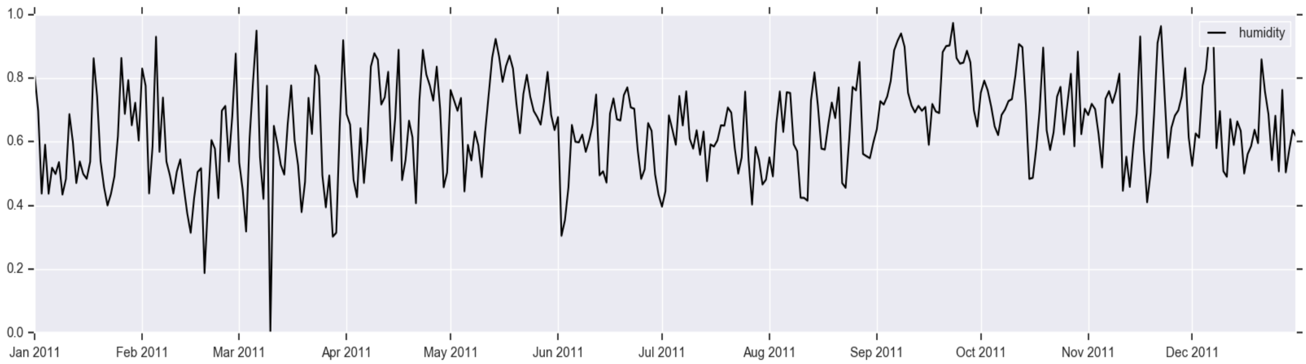
1. Bee Swarmplots because no binning Bias unlike histogram:

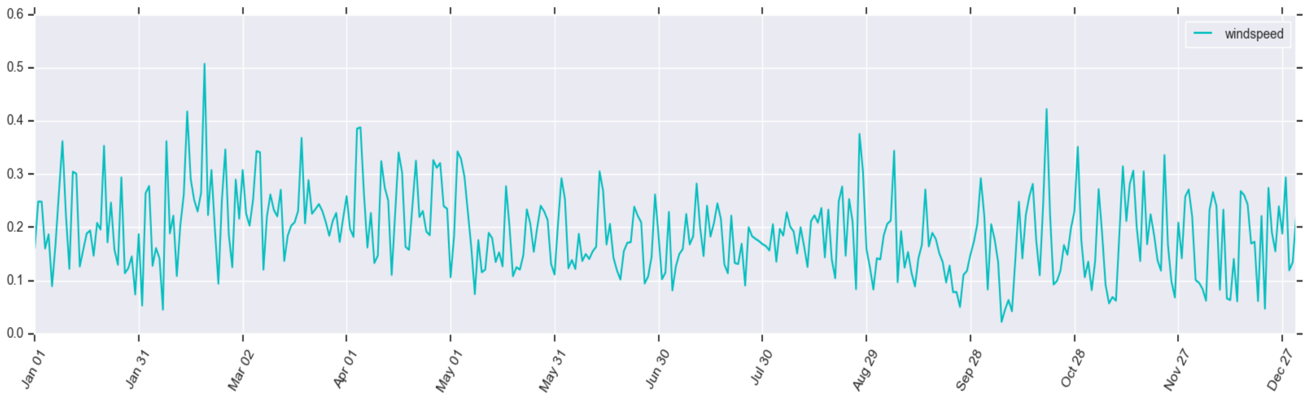

Bike Count w.r.t season and working day
workingday
0
1
cnt
10000
8000
6000
4000
2000
0
−2000
1
2
3
4
season
Bike Count w.r.t weekday and working day
workingday
0
1
cnt
0
1
2
3
4
5
6
weekday
Bike Count w.r.t holiday and working day
workingday
0
1
cnt
0
1
holiday
Bike Count w.r.t weathersit and working day
workingday
0
1
cnt
1
2
3
weathersit
Bike Count w.r.t yr and working day
workingday
0
1
cnt
0
1
yr
Bike Count w.r.t mnth and working day
workingday
0
1
cnt
1
2
3
4
5
6
7
8
9
10
11
12
mnth

2. Time Series Visualization:





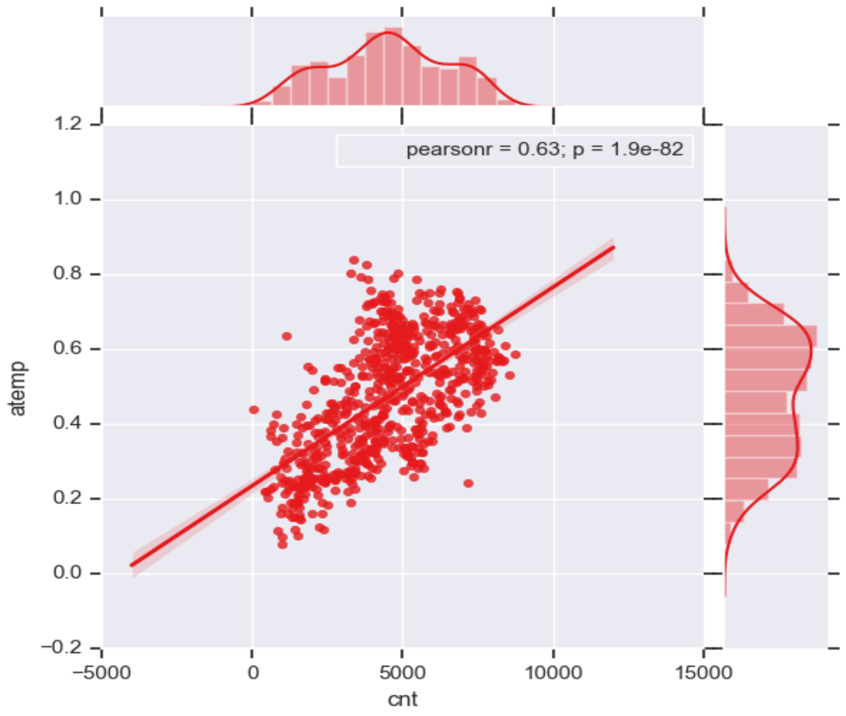


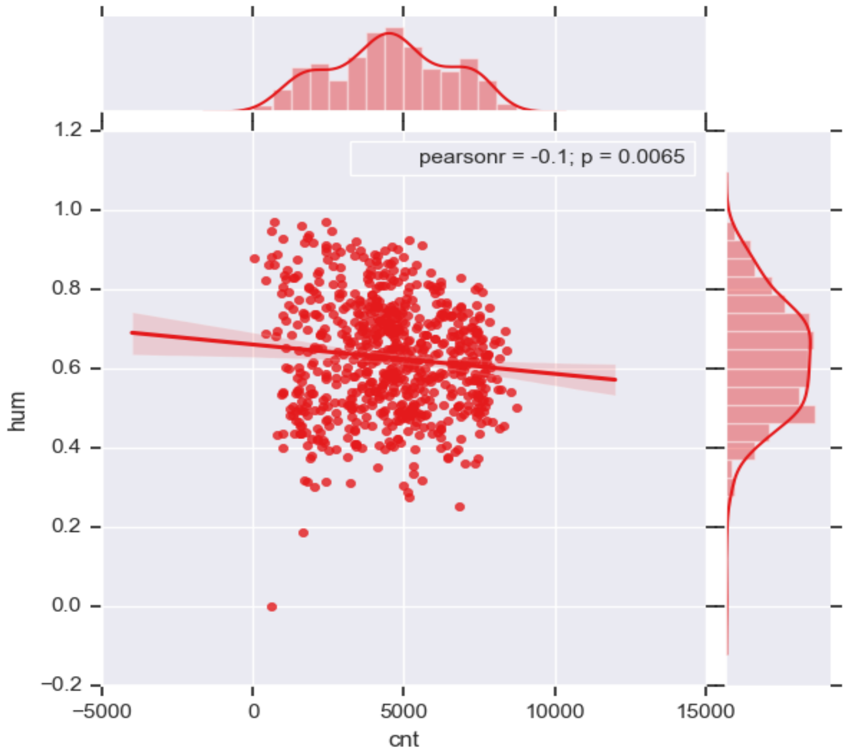


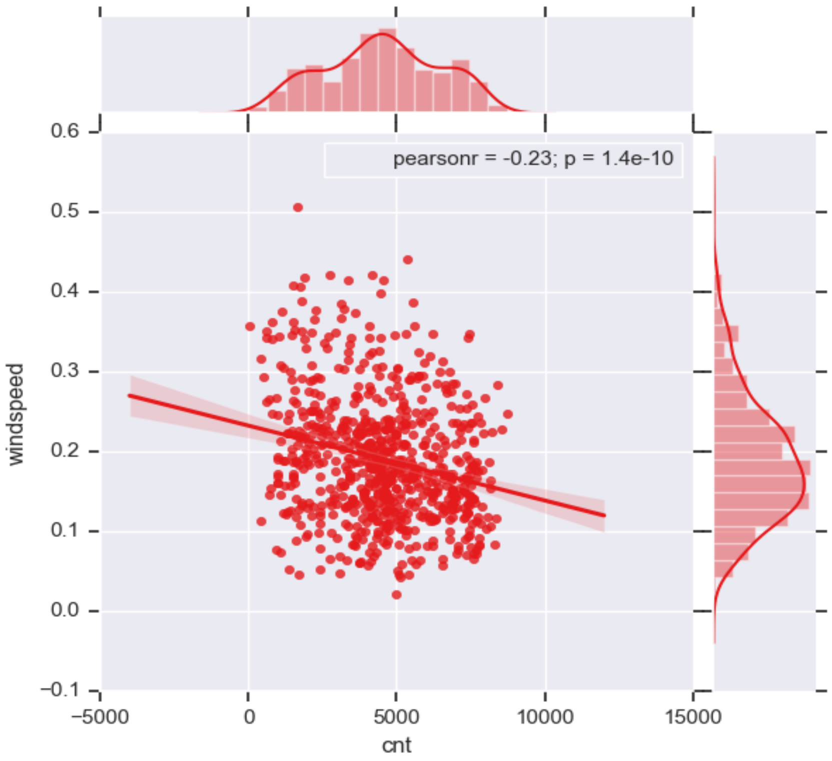
3. Jointplots:-

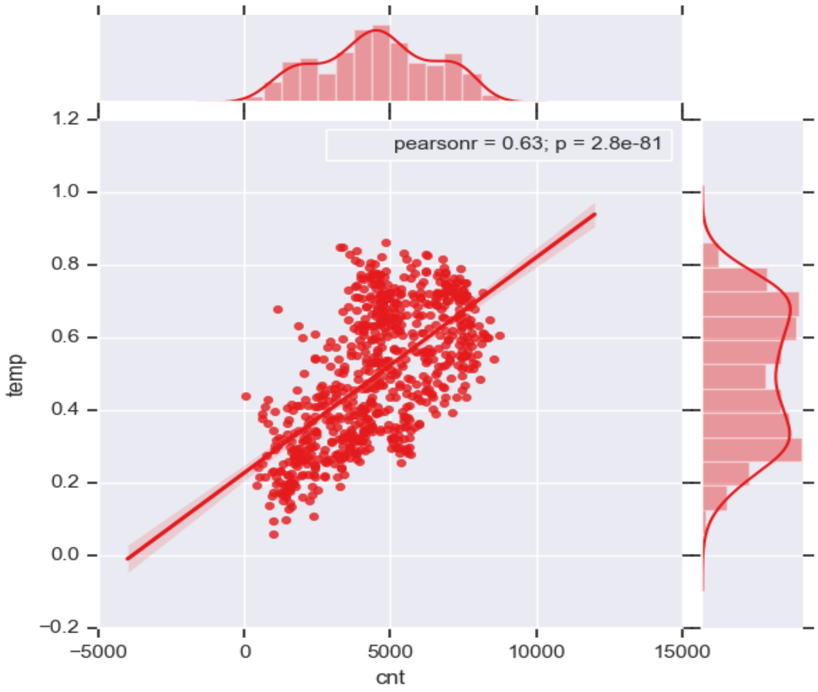
- They are used for Bivariate Analysis.
  Here we have plotted Scatter plot with Regression line between 2 variables along with separate Bar plots of both variables.

- Also, we have annotated Pearson correlation coefficient and p value.
- Plotted only for numerical/continuous variables
- Each numerical variable at a time Vs ‘cnt’ which is Target Variable.

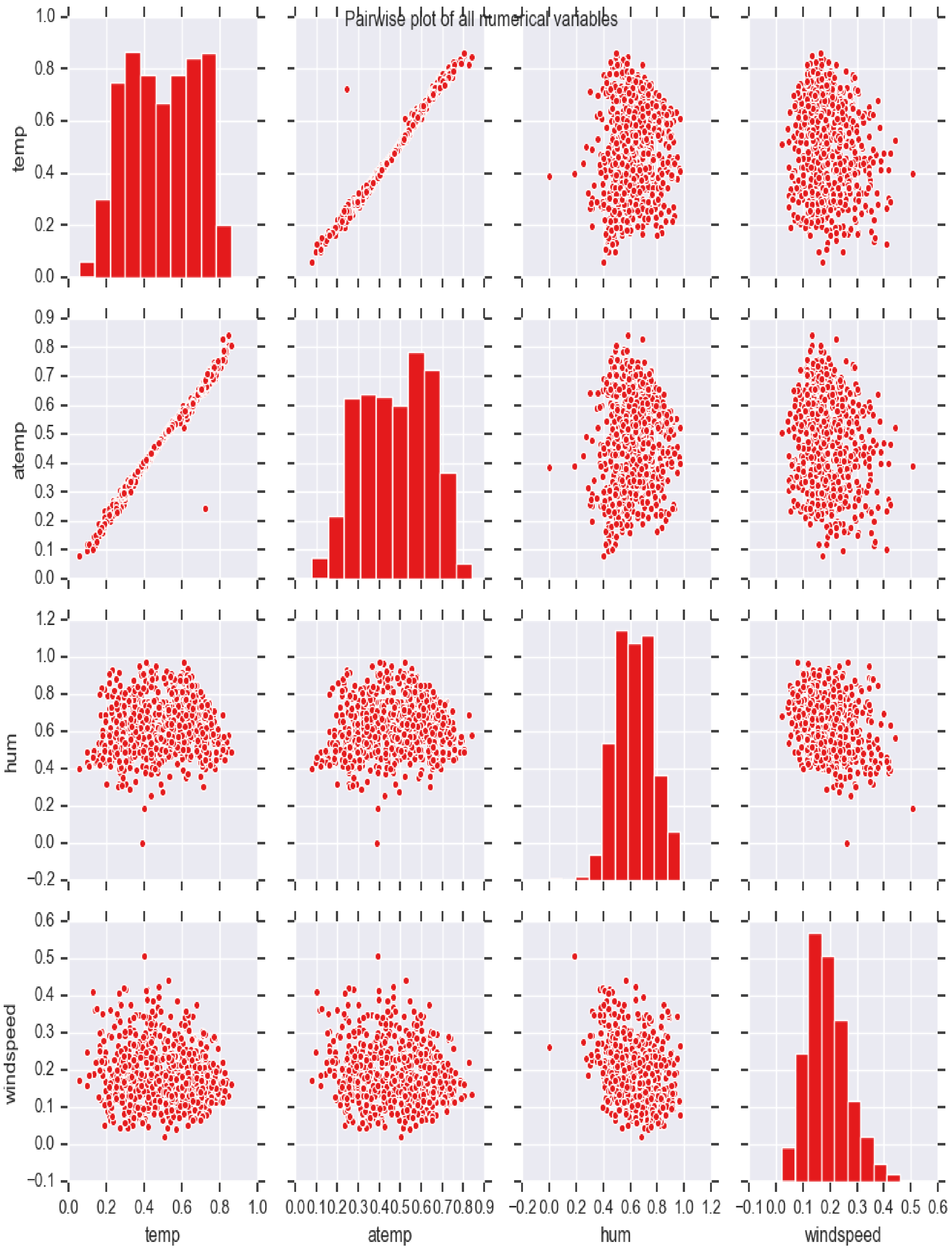



pearsonr = -0.1; p = 0.0065
hum
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−5000
0
5000
10000
15000
cnt


pearsonr = -0.23; p = 1.4e-10
windspeed
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−5000
0
5000
10000
15000
cnt



4. Pairwise Plots for all Numerical variables:


Pairwise plot of all numerical variables
temp
atemp
hum
windspeed
1.0
0.8
0.6
0.4
0.2
0.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0.0 0.2 0.4 0.6 0.8 1.0
0.00.10.20.30.40.50.60.70.80.9
−0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2
−0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6
temp
atemp
hum
windspeed

- ## Missing value Analysis:-
- In this step we look for missing values in the dataset like empty row column cell which was left after removing special characters and punctuation marks.

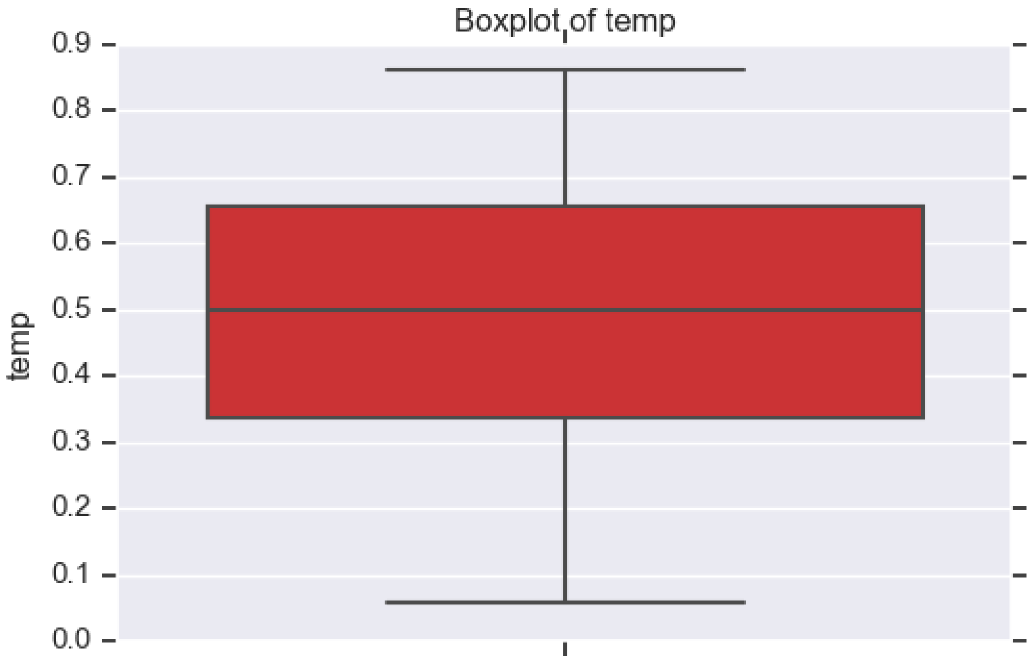
- Some missing values are in form of NA. missing values left behind after outlier analysis; missing values can be in any form. Unfortunately, in this dataset we haven't found any missing values. Therefore, we will continue to next step.

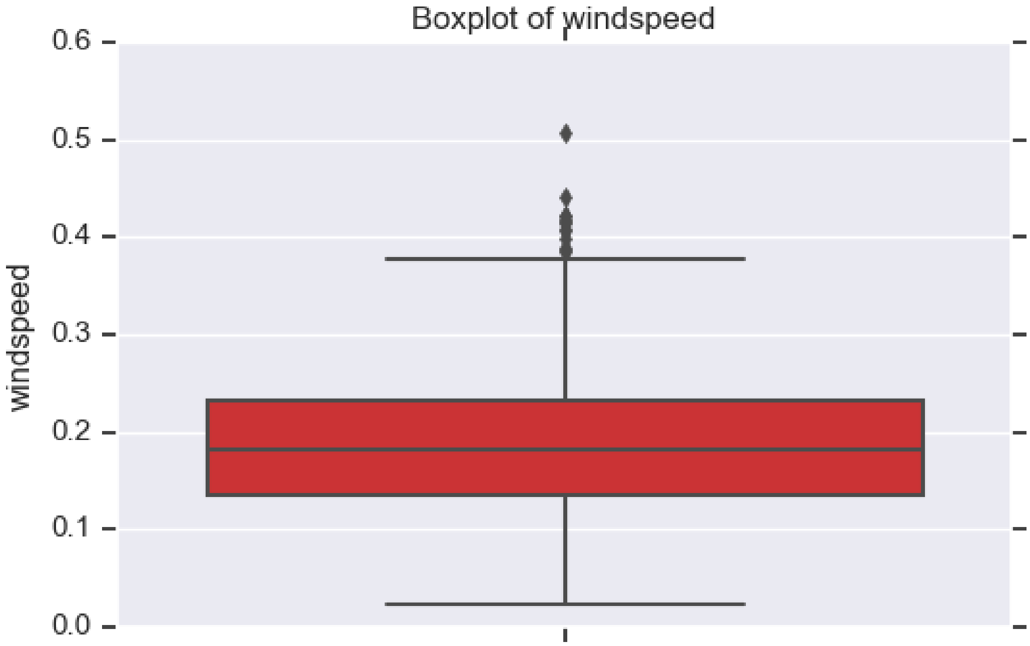
- ## Outlier Analysis:-

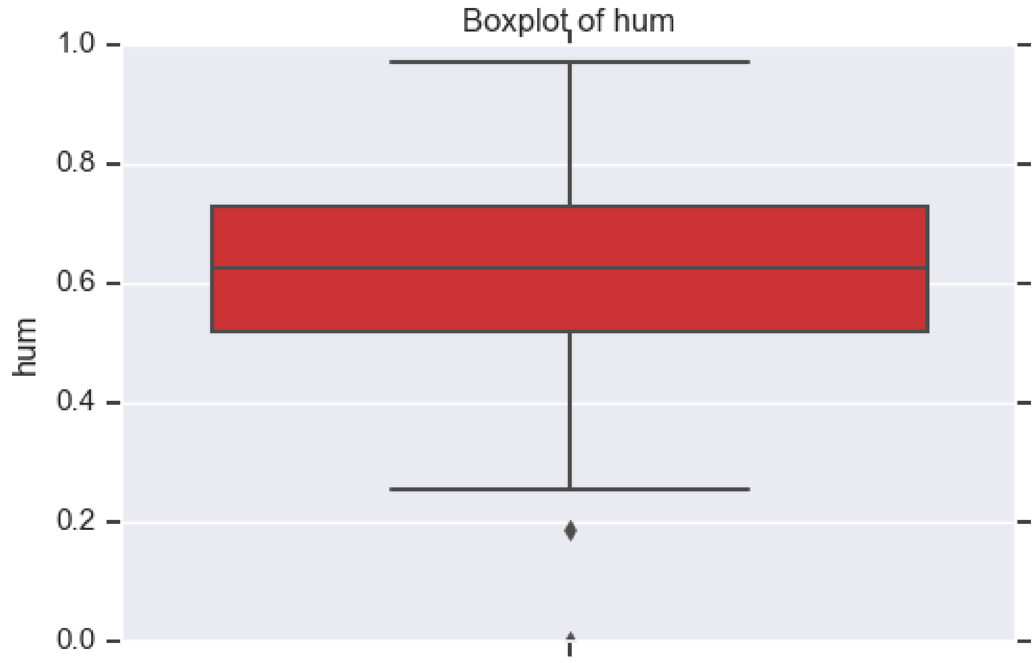
This is how we done,

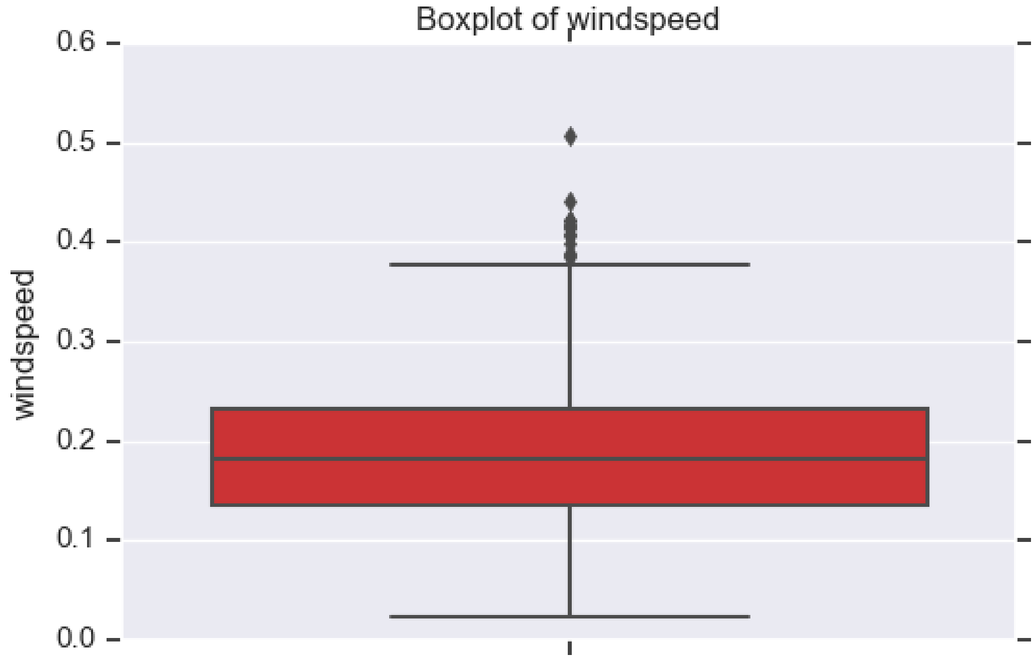
I. We replaced them with Nan values or we can say created missing values.
II. Then we imputed those missing values with KNN method.
III. We tried three methods to impute the missing value: mean, median, KNN. But KNN method outperformed mean and median methods.
IV. We checked the performance of each method by checking Standard Deviation of that variable which has outliers before imputation and after imputation

- Univariate Boxplots: Boxplots for all Numerical Variables also for target variable:-

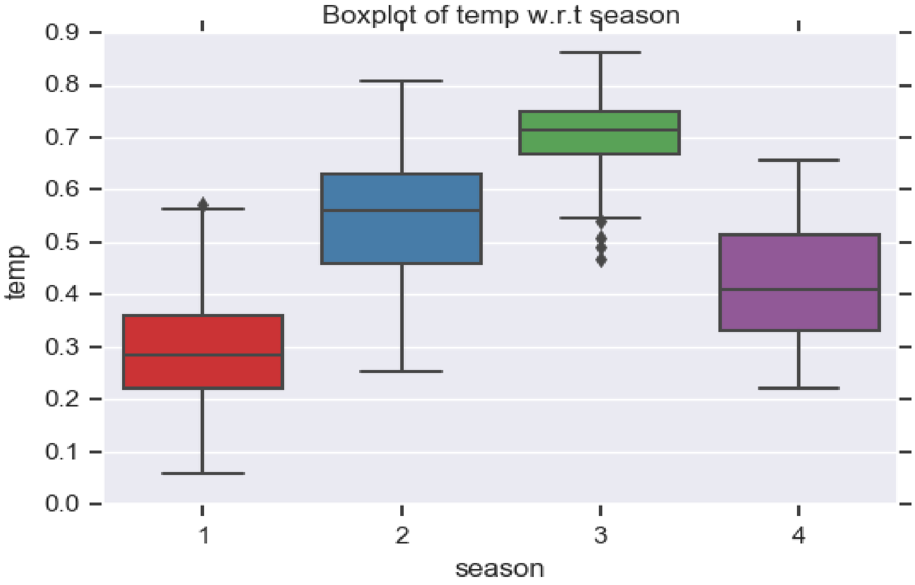

Boxplot of temp
temp
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

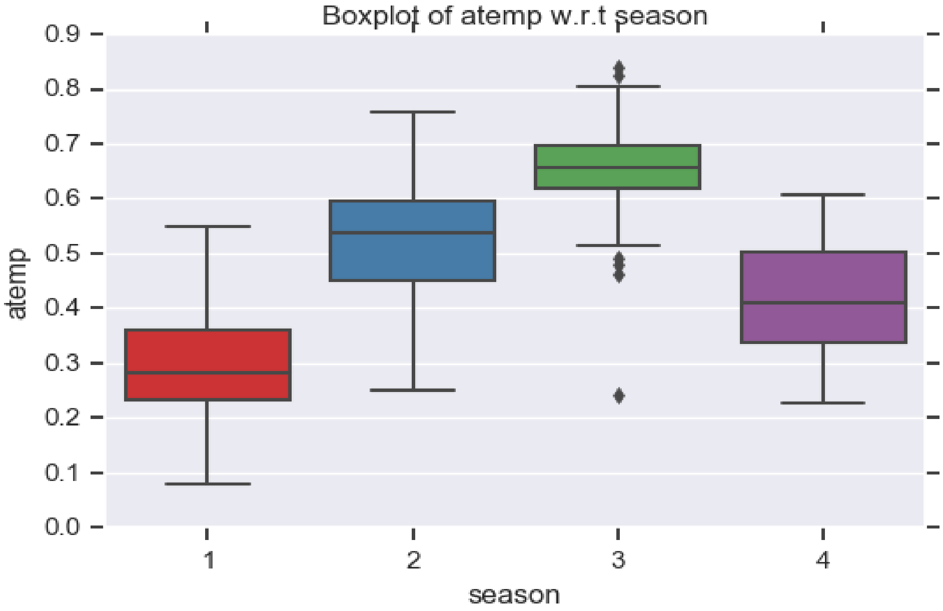

Boxplot of windspeed
windspeed
0.6
0.5
0.4
0.3
0.2
0.1
0.0

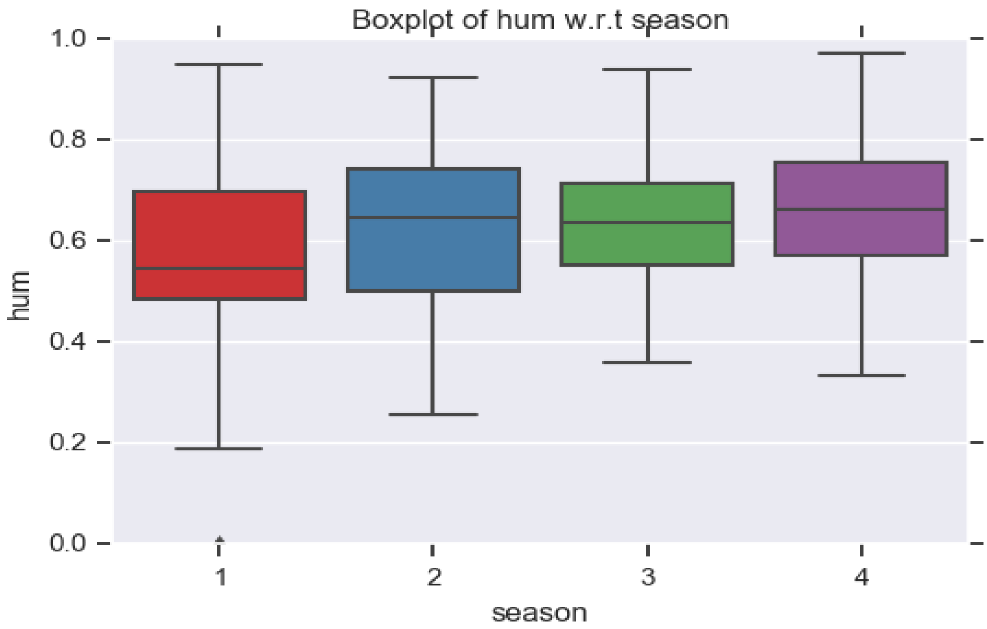

Boxplot of hum
1.0
0.8
0.6
0.4
0.2
0.0
hum

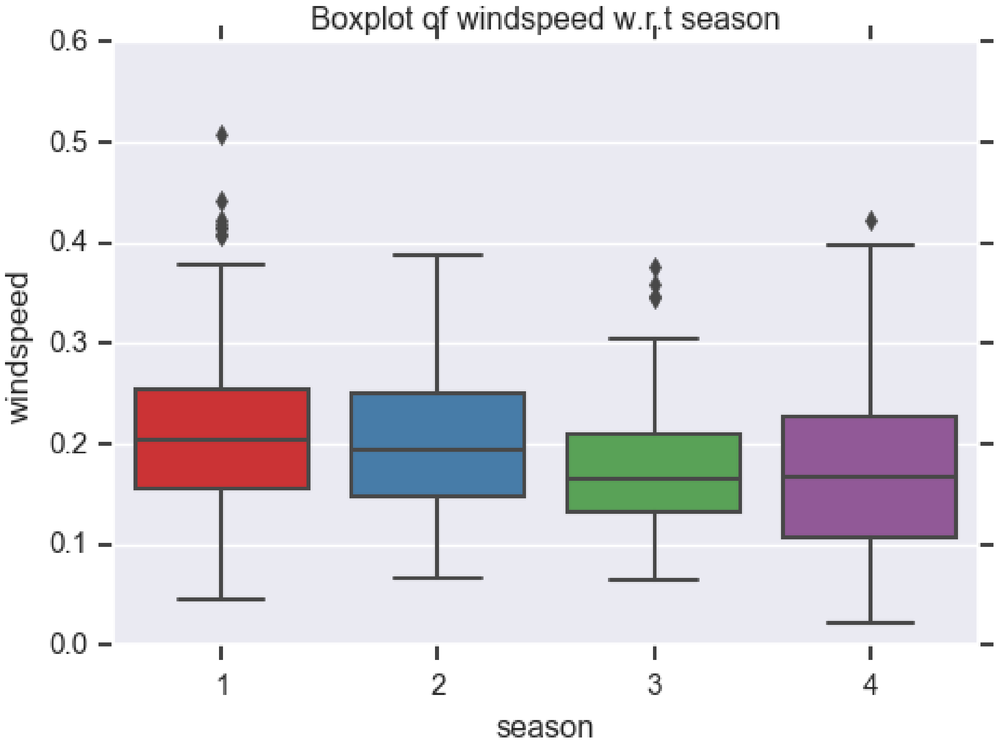

Boxplot of windspeed
0.6
0.5
0.4
0.3
0.2
0.1
0.0
windspeed

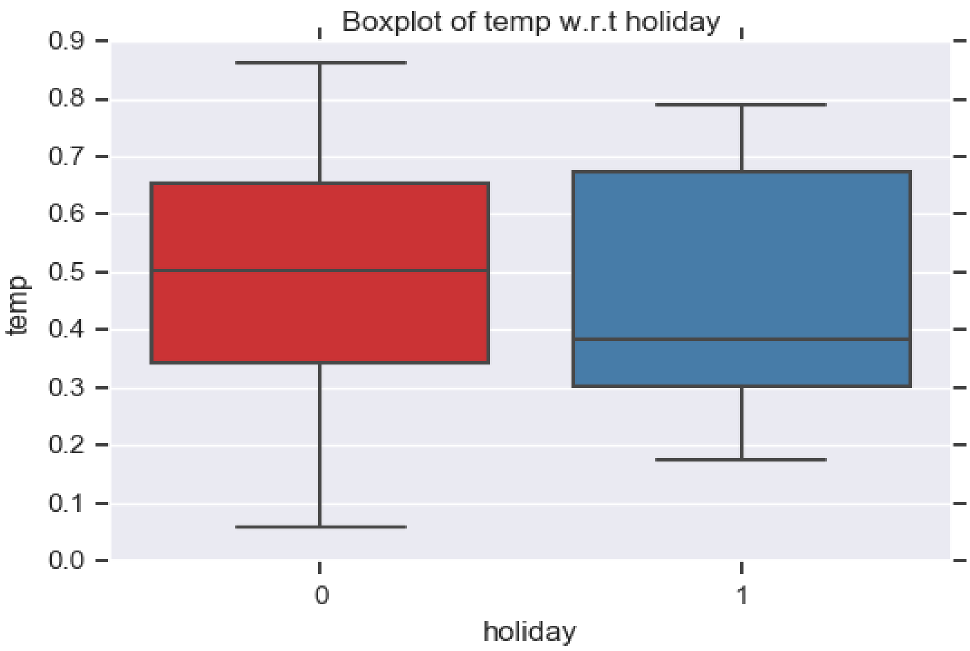
- Bivariate Boxplots: Boxplots for all Numerical Variables Vs all Categorical Variables :-

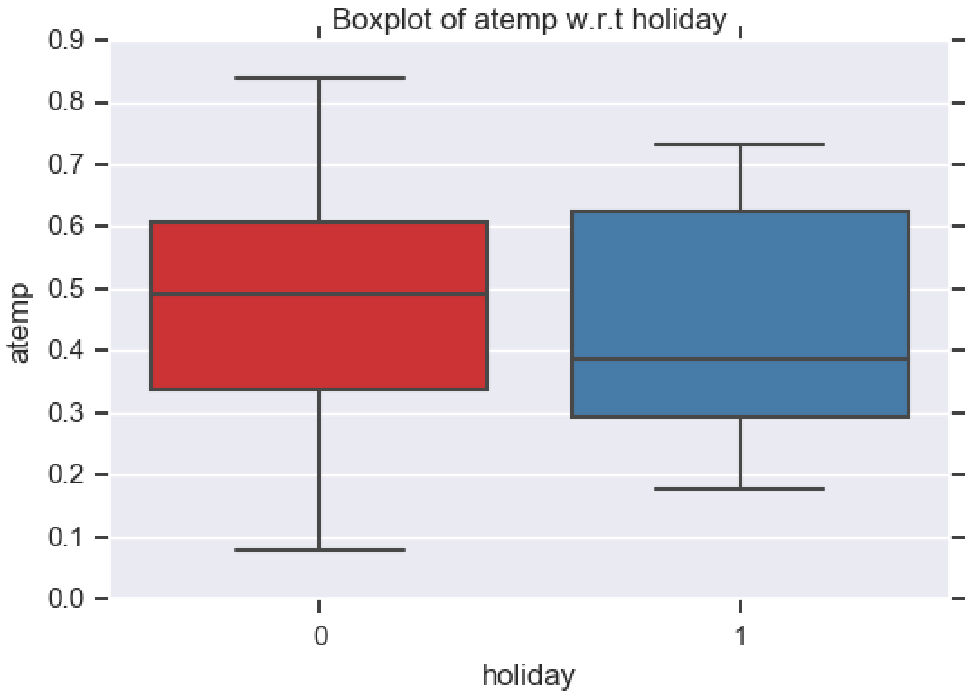


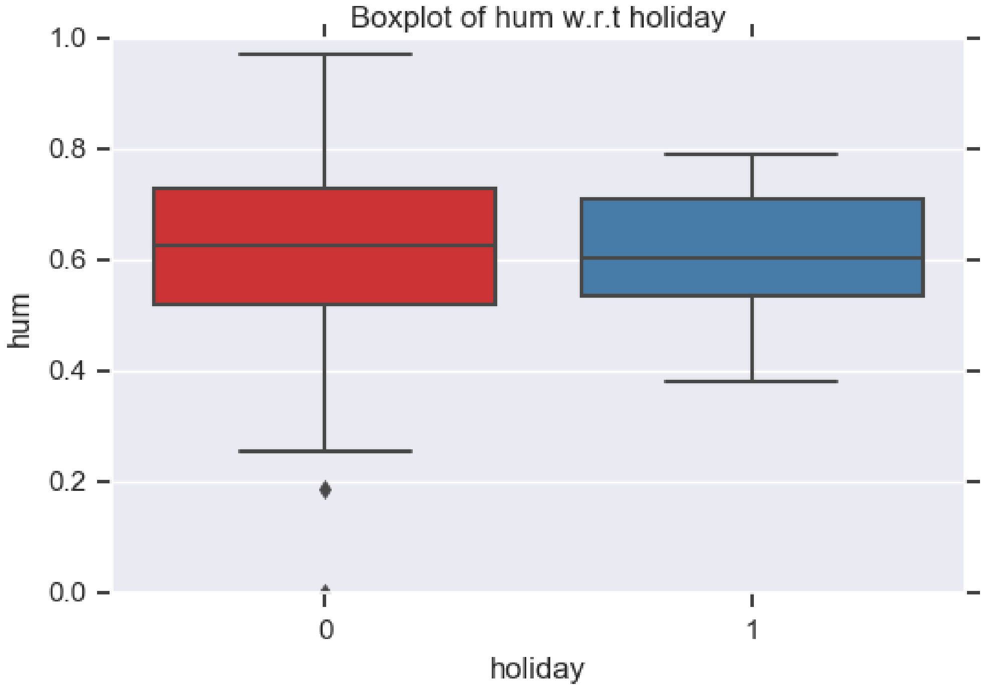


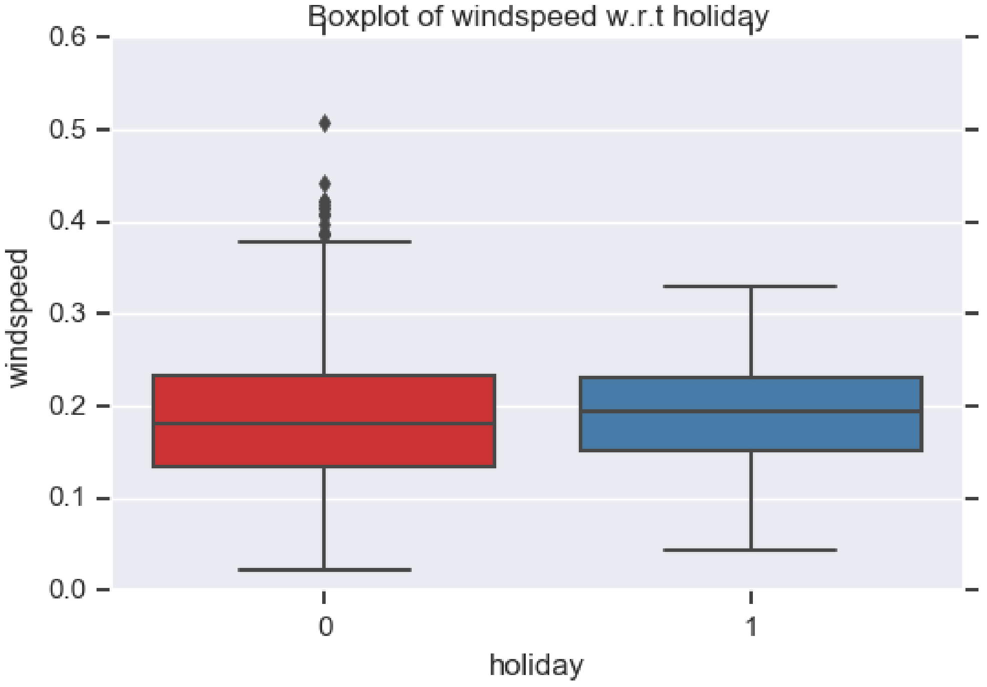

Boxplot of hum w.r.t season
hum
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
season

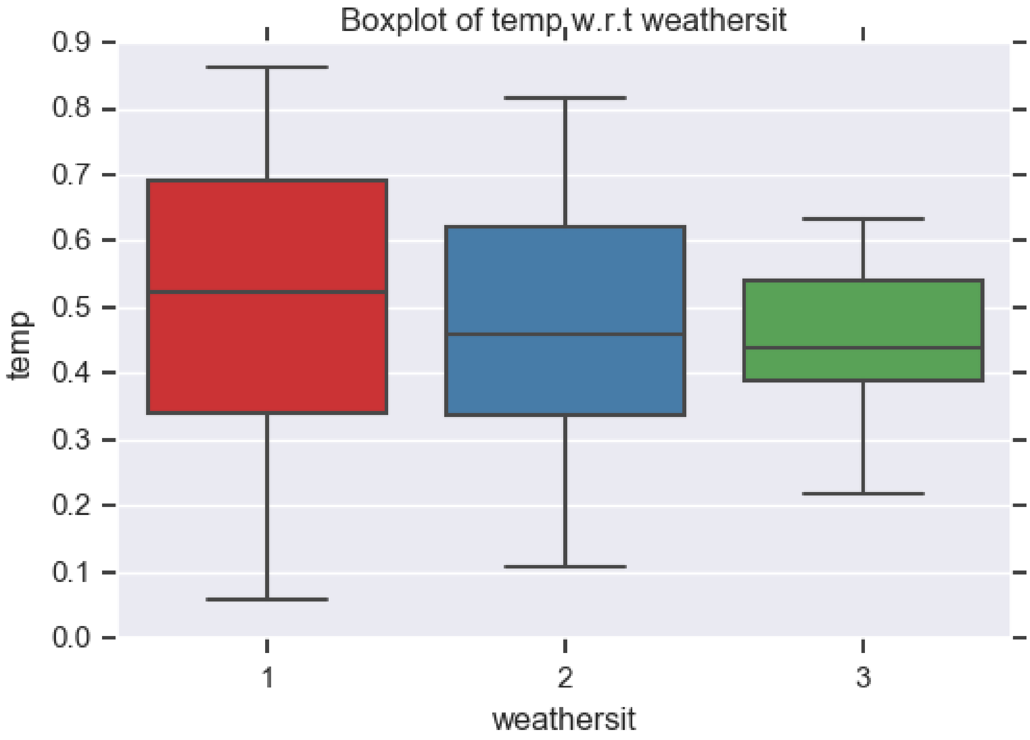

Boxplot of windspeed w.r.t season
windspeed
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
season

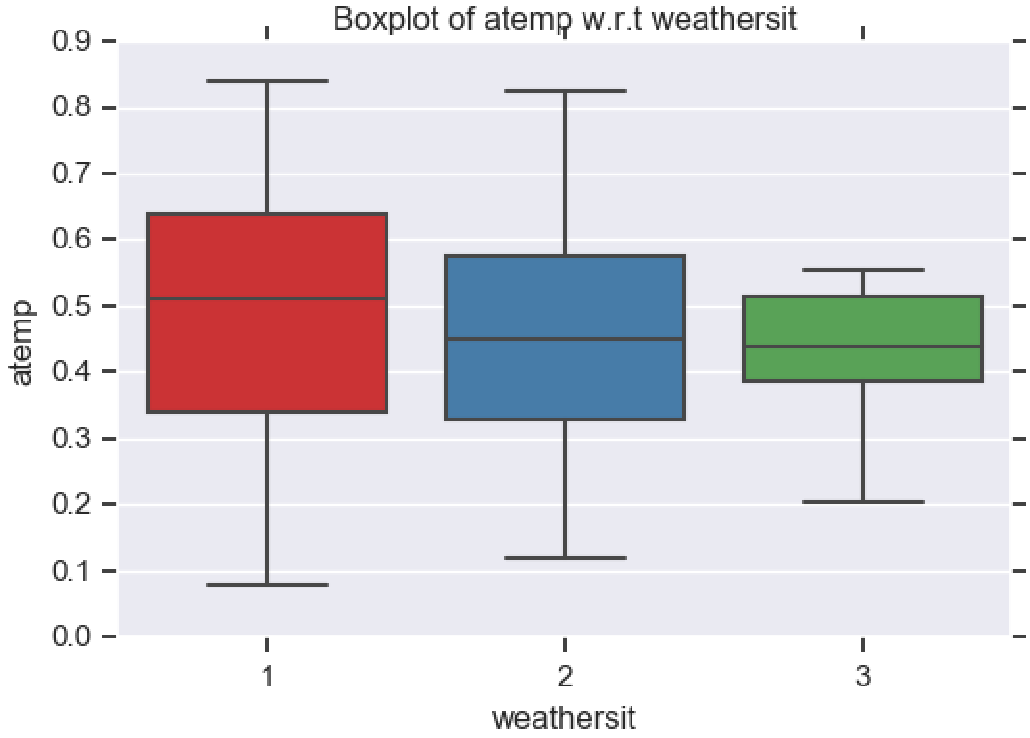

Boxplot of temp w.r.t holiday
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
temp
0
1
holiday

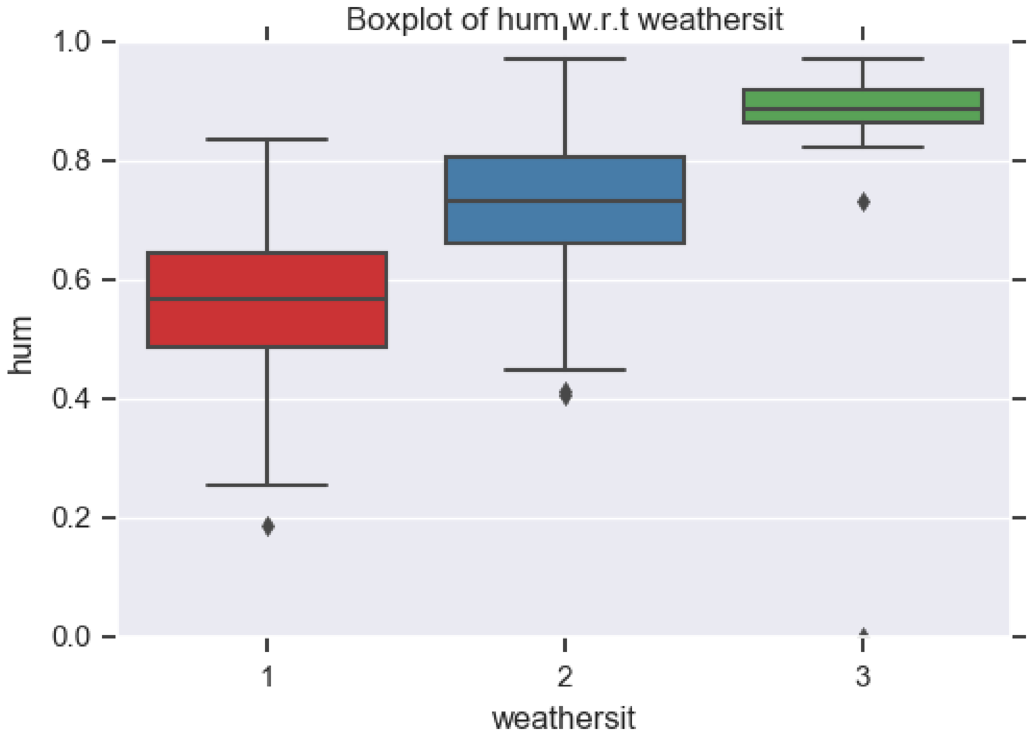

Boxplot of atemp w.r.t holiday
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
atemp
0
1
holiday

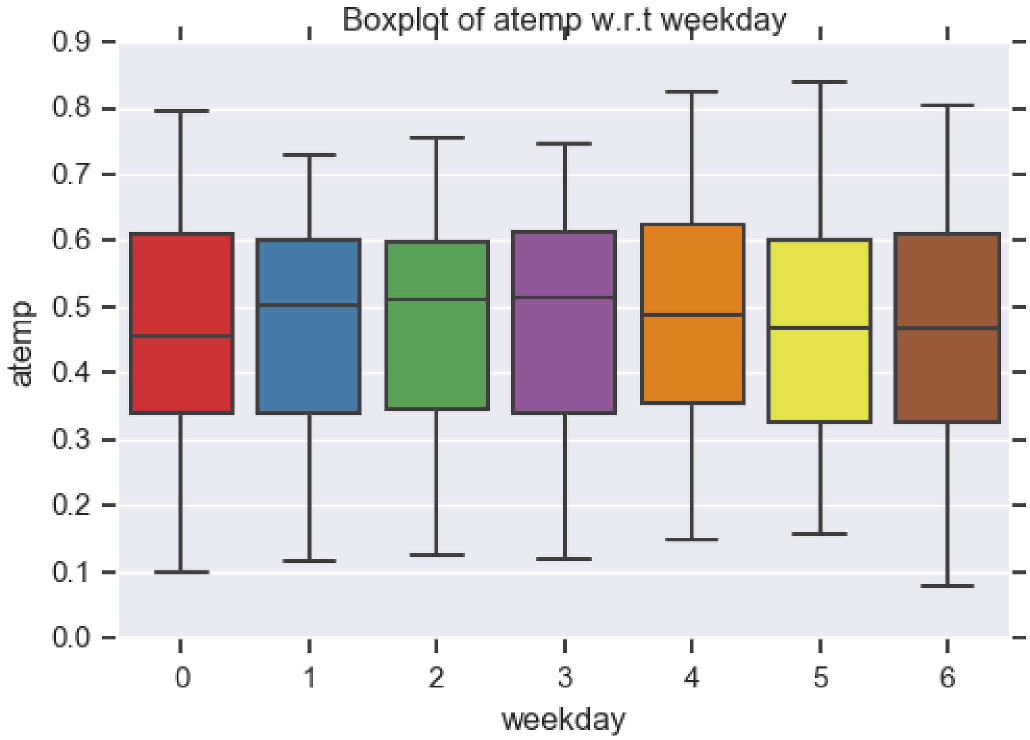

Boxplot of hum w.r.t holiday
hum
1.0
0.8
0.6
0.4
0.2
0.0
0
1
holiday

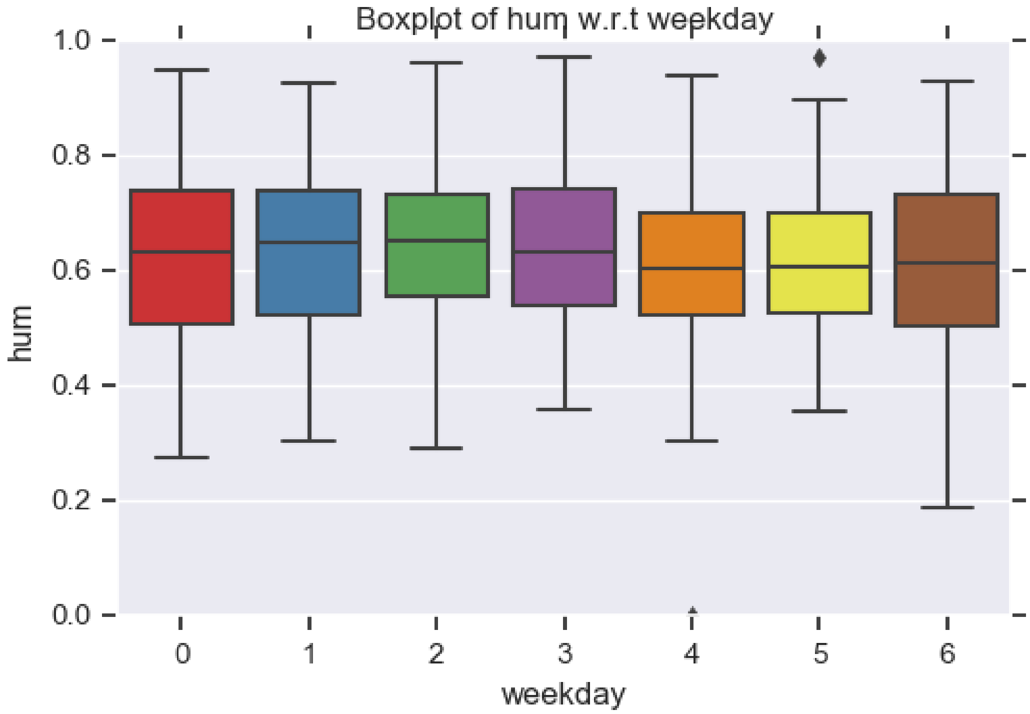

Boxplot of windspeed w.r.t holiday
windspeed
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
holiday

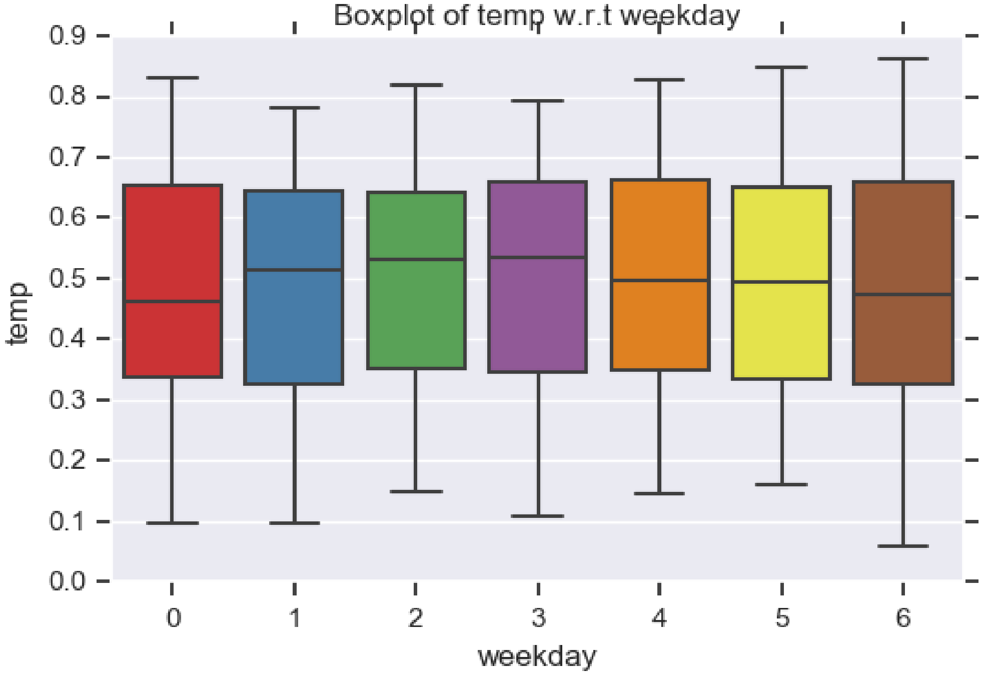

Boxplot of temp w.r.t weathersit
temp
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
weathersit

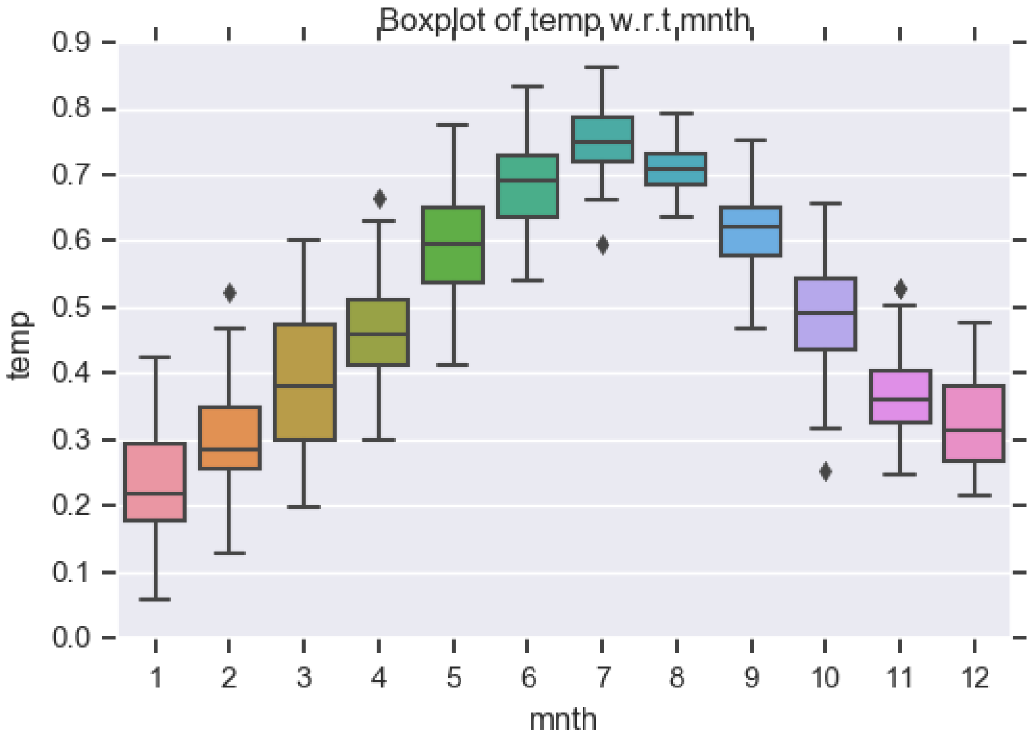

Boxplot of atemp w.r.t weathersit
atemp
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
weathersit

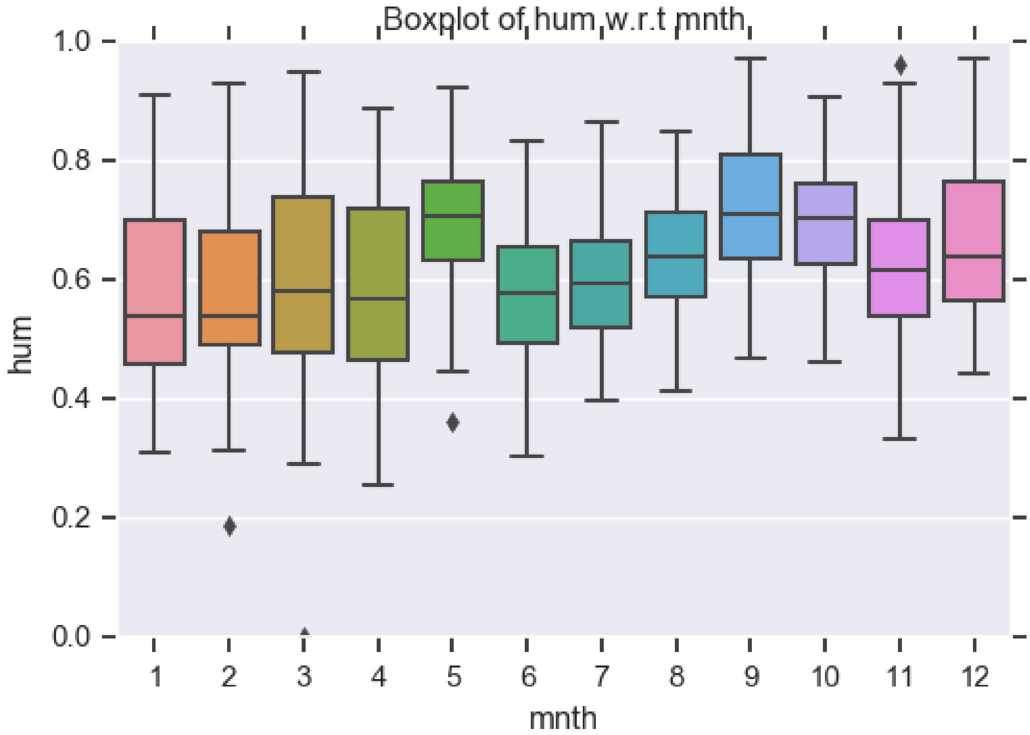

Boxplot of hum w.r.t weathersit
hum
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
weathersit

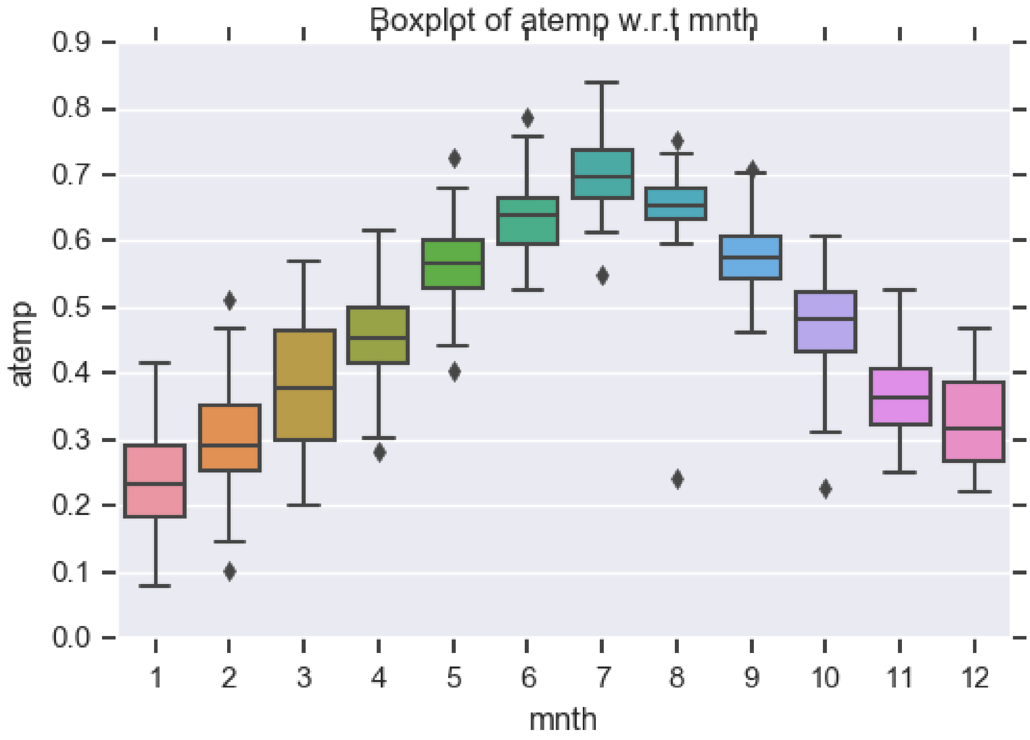

Boxplot of atemp w.r.t weekday
atemp
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
weekday

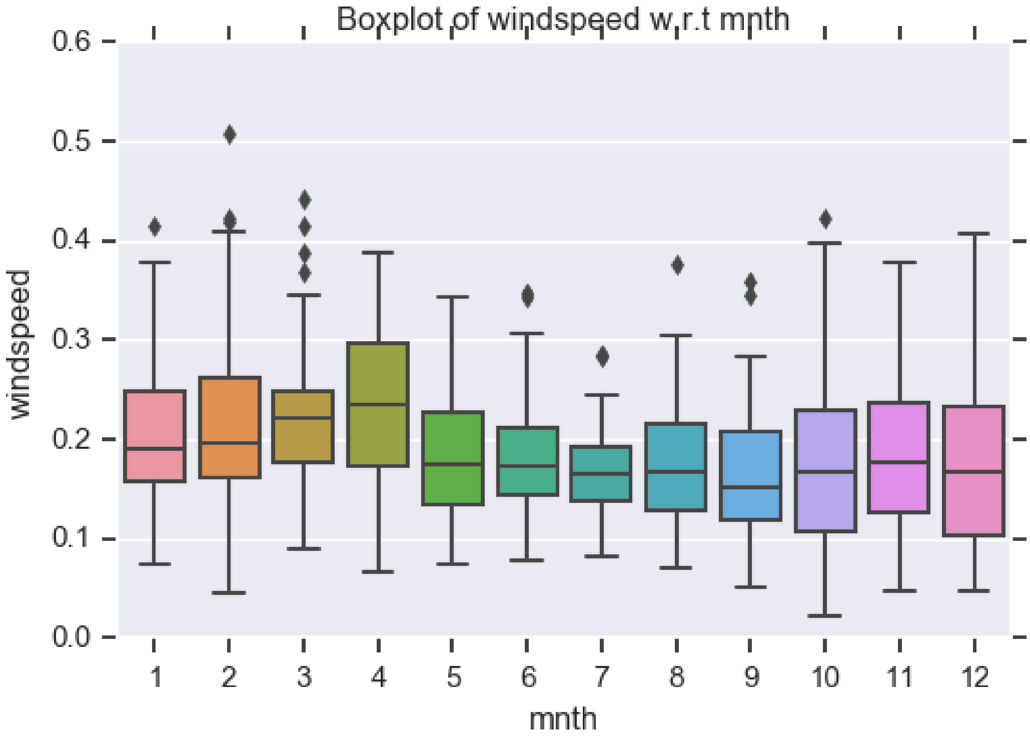

Boxplot of hum w.r.t weekday
hum
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
weekday


Boxplot of temp w.r.t weekday
temp
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
weekday


Boxplot of temp w.r.t mnth
temp
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
4
5
6
7
8
9
10
11
12
mnth


Boxplot of hum w.r.t mnth
hum
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
11
12
mnth


Boxplot of atemp w.r.t mnth
atemp
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
2
3
4
5
6
7
8
9
10
11
12
mnth


Boxplot of windspeed w.r.t mnth
windspeed
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
11
12
mnth

- From above Boxplots we see that only ‘hum’ and ‘windspeed’ have outliers in them.
  ‘hum’ has 2 outliers and ‘windspeed’ has 13 outliers.

Feature Selection:-

- In this dataset we have to predict the Count based on environmental and seasonal features, features which excludes this list is – instant.

- **Correlation analysis** – This requires only numerical variables. Therefore, we will filter out only numerical variables and feed it to correlation analysis. We do this by plotting correlation plot for all numerical variables. There should be no correlation between independent variables but there should be high correlation between independent variable and dependent variable. So, we plot the correlation plot. we can see that in correlation plot faded colour like skin colour indicates that two variables are highly correlated with each other.

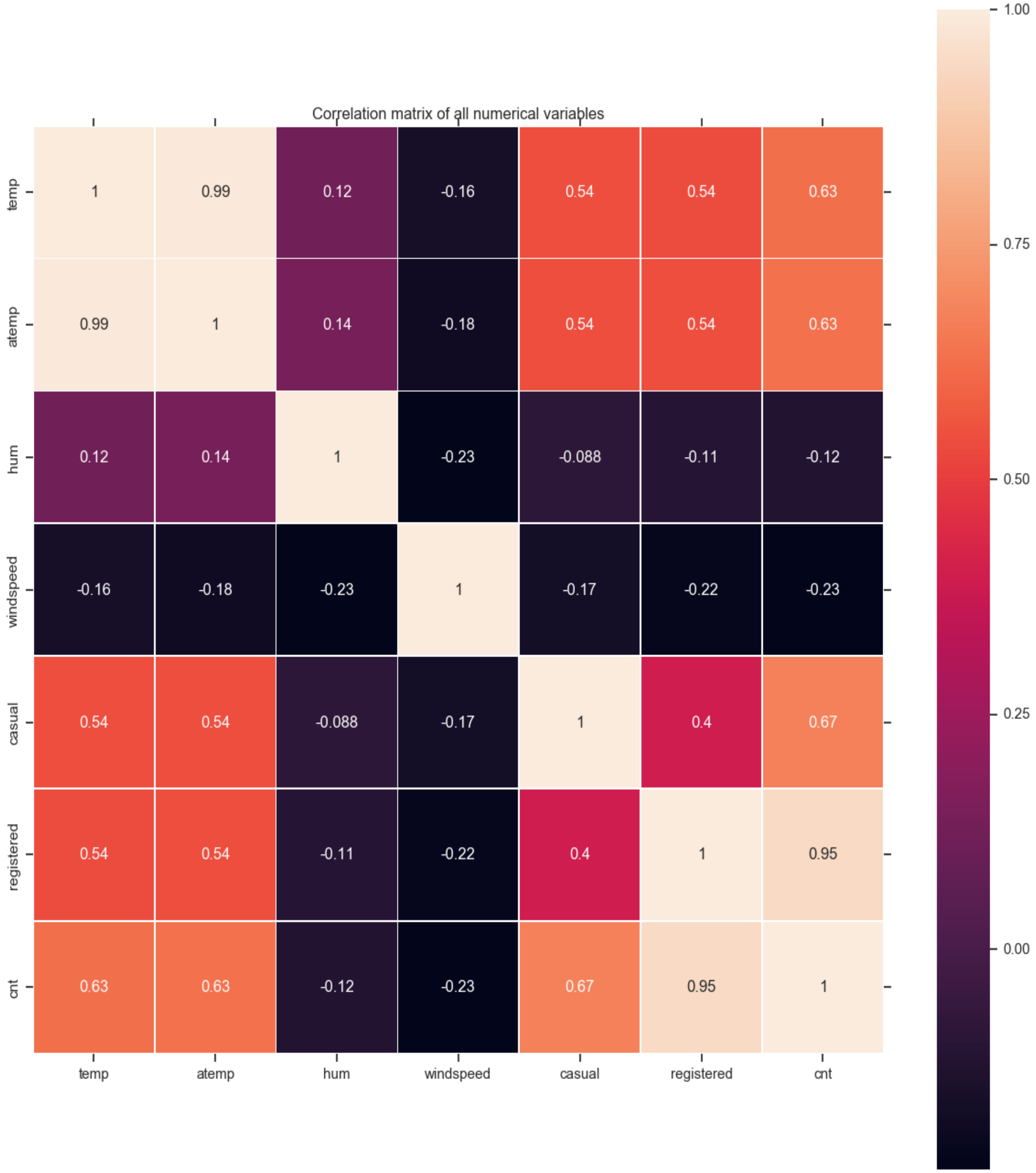
From below correlation plot we see that:

• 'temp' and 'atemp' are very highly correlated with each other.

• Similarly, 'registered' and 'cnt' are highly correlated with each other.

• We also came to know that--'cnt'='casual'+'registered' .

Correlation plot :-


Correlation matrix of all numerical variables

Chi-Square test of independence –
Unlike correlation analysis we will filter out only categorical variables and pass it to Chi-Square test. Chi-square test compares 2 categorical variables in a contingency table to see if they are related or not.

I. Assumption for chi-square test: Dependency between Independent variable and dependent variable should be high and there should be no dependency among independent variables.

II. Before proceeding to calculate chi-square statistic, we do the hypothesis testing:

Null hypothesis: 2 variables are independent.
Alternate hypothesis: 2 variables are not independent.

The interpretation of chi-square test:

I. For theorical or excel sheet purpose: If chi-square statistics is greater than critical value then reject the null hypothesis saying that 2 variables are dependent and if it's less, then accept the null hypothesis saying that 2 variables are independent.

II. While programming: If p-value is less than 0.05 then we reject the null hypothesis saying that 2 variables are dependent and if p-value is greater than 0.05 then we accept the null hypothesis saying that 2 variables are independent.

Here we did the test between categorical independent variables pairwise.

- If p-value<0.05 then remove the variable,
    - If p-value>0.05 then keep the variable

variables which are highly dependent on each other based on p-values are:

- season and weathersit
- season and month
- holiday and weekday
- hoilday and workingday
- weekday and holiday
- weekday and workingday
- workingday and holiday
- workingday and weekday
- weathersit and season
- weathersit and mnth
- mnth and season
- mnth and weathersit

- After analysing p value of all categorical features, we come to a conclusion that, Variables which we have removed and kept:

Removed: mnth, weekday, workingday, weathersit.
Kept: season, holiday, yr.

Analysis of Variance(Anova) Test –
I. It is carried out to compare between each group in a categorical variable.
II. ANOVA only lets us know the means for different groups are same or not. It doesn't help us identify which mean is different.

Hypothesis testing: -

- Null Hypothesis:
mean of all categories in a variable are same.

- Alternate Hypothesis:
mean of at least one category in a variable is different.

- If p-value is less than 0.05 then we reject the null hypothesis.
- And if p-value is greater than 0.05 then we accept the null hypothesis.

- **Multicollinearity**–

In regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other.

- Multicollinearity increases the standard errors of the coefficients.
- Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0.

- In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant.

- VIF is always greater or equal to 1.

if VIF is 1 --- Not correlated to any of the variables.
if VIF is between 1-5 --- Moderately correlated.
if VIF is above 5 --- Highly correlated.
If there are multiple variables with VIF greater than 5, only remove the variable with the highest VIF.

And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

We have checked for multicollinearity in our Dataset and VIF values for temp and atemp are above 5.

- ## Feature Scaling:-

- **Normalization**:
  Normalization refer to the dividing of a vector by its length. normalization normalizes the data in the range of 0 to 1. It is generally used when we are planning to use distance method for our model development purpose such as KNN. Normalizing the data improves convergence of such algorithms. Normalisation of data scales the data to a very small interval, where outliers can be loosed.

- **Standardization**:
  Standardization refers to the subtraction of mean from individual point and then dividing by its SD. Z is negative when the raw score is below the mean and Z

is positive when above mean. When the data is distributed normally you should go for standardization.

Splitting Train & Test Dataset

a. With the time series data, we will break up our train and test into continuous chunks.

b. The training data should be the earliest data and test data should be the latest data.

c. we will fit our model on the training data and test on the newest data, to understand how our model performs on new, unseen data.

d. we can't use sklearn's train_test_split bcoz it randomly shuffles the train and test data.
We have divided our data in 80-20% i.e. 80% in train and 20% in test.

Model Development

Our problem statement wants us to predict the Bike rental count. This is a Regression problem. So, we are going to build regression models on training data and predict it on test data. In this project I have built models using 5 Regression Algorithms:

I Linear Regression

II Decision Tree

III Random Forest

We will evaluate performance on test dataset generated using Sampling. We will deal with specific error metrics like – Regression metrics for our Models:

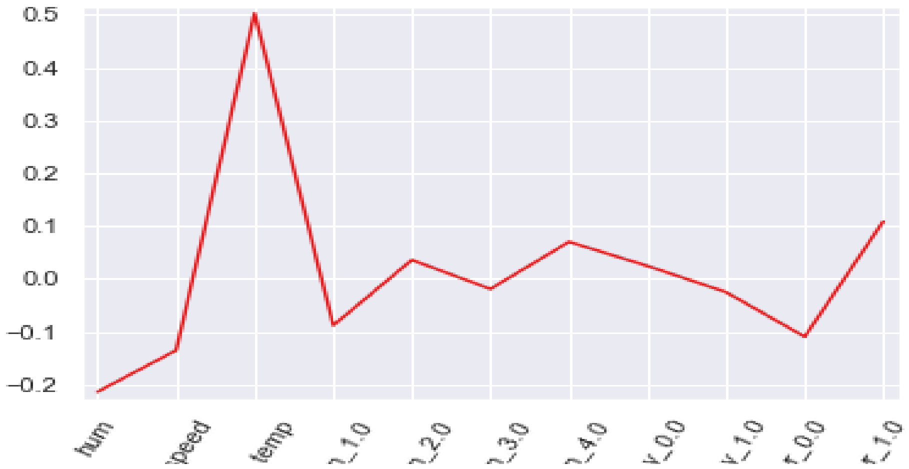
- r square
- Adjusted r square
- MAPE(Mean Absolute Percentage Error)
- MSE(Mean square Error)
- RMSE(Root Mean Square Error)
- RMSLE(Root Mean Squared Log Error)

- Model Performance:-

Linear Regression:

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.819	0.8163	18.88	0.0076	0.0872	0.058
Test	0.554	0.5177	Inf	0.020	0.144	0.093

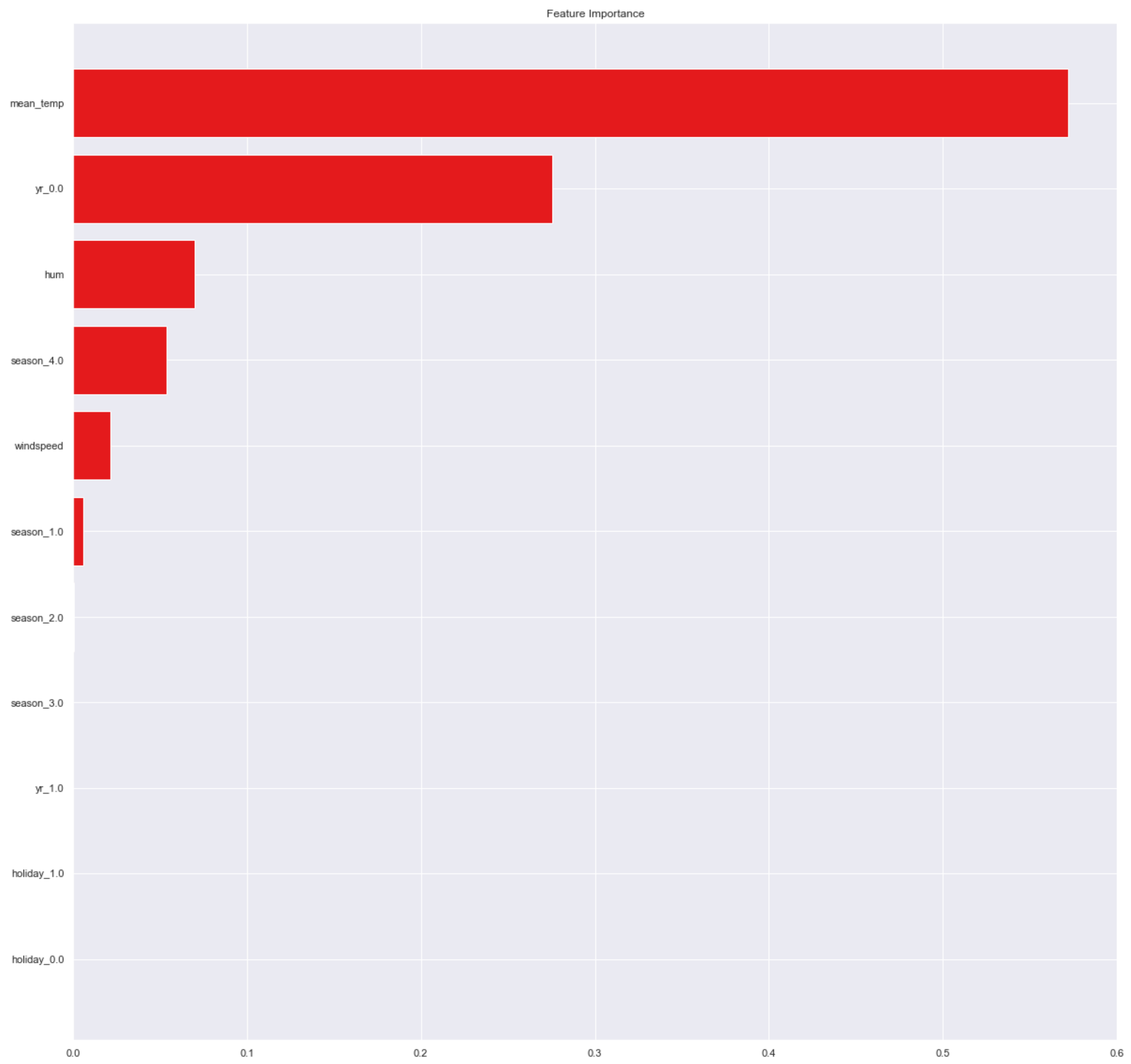
- Line Plot for Coefficients of Linear regression:-



Decision Tree Regression:

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.913	0.911	11.85	0.003	0.060	0.040
Test	0.519	0.480	Inf	0.022	0.149	0.097

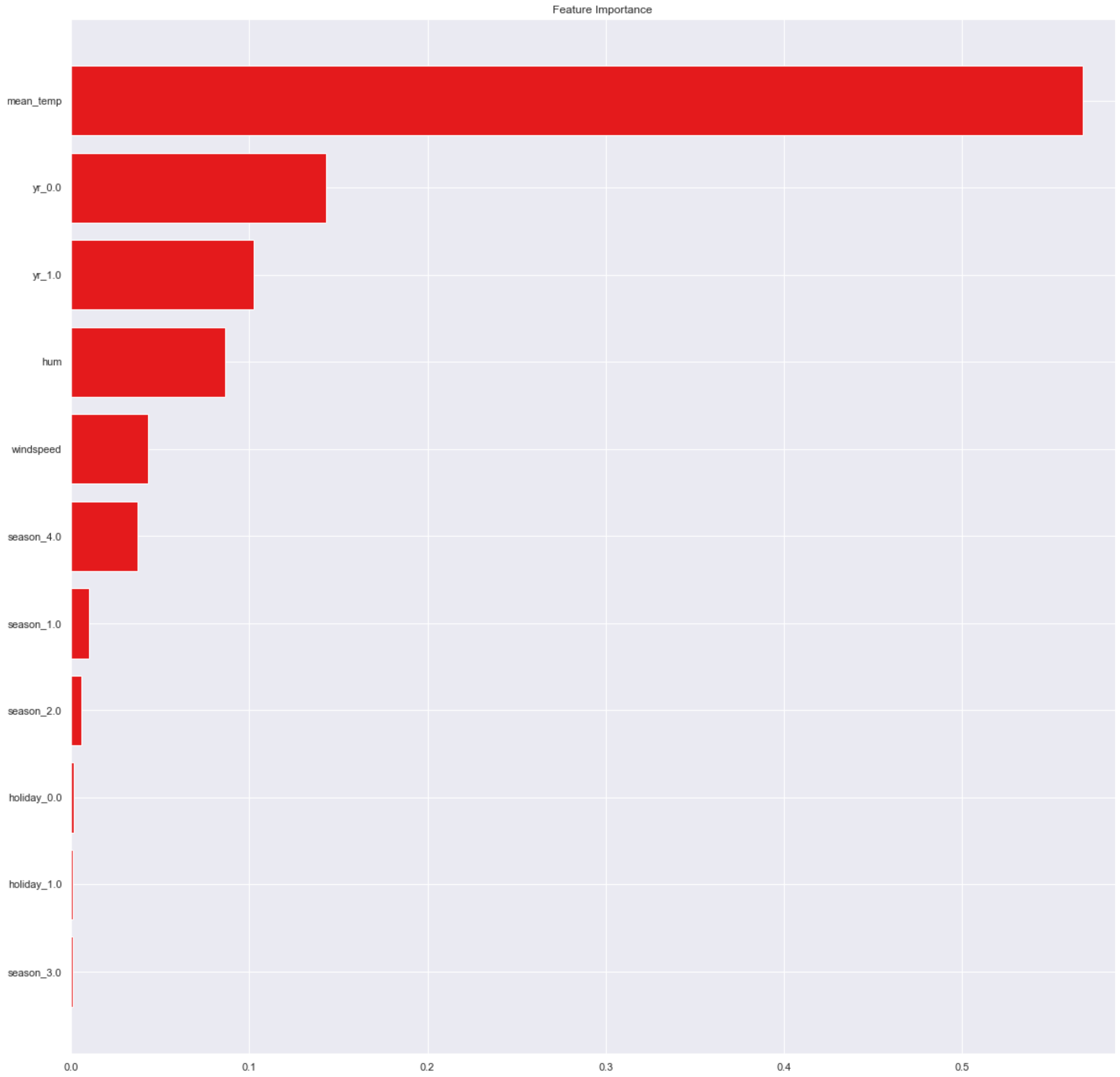
- Bar Plot of Decision tree Feature Importance:



Random Forest Regression:

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.981	0.98	6.29	0.0007	0.027	0.019
Test	0.551	0.515	Inf	0.020	0.144	0.093

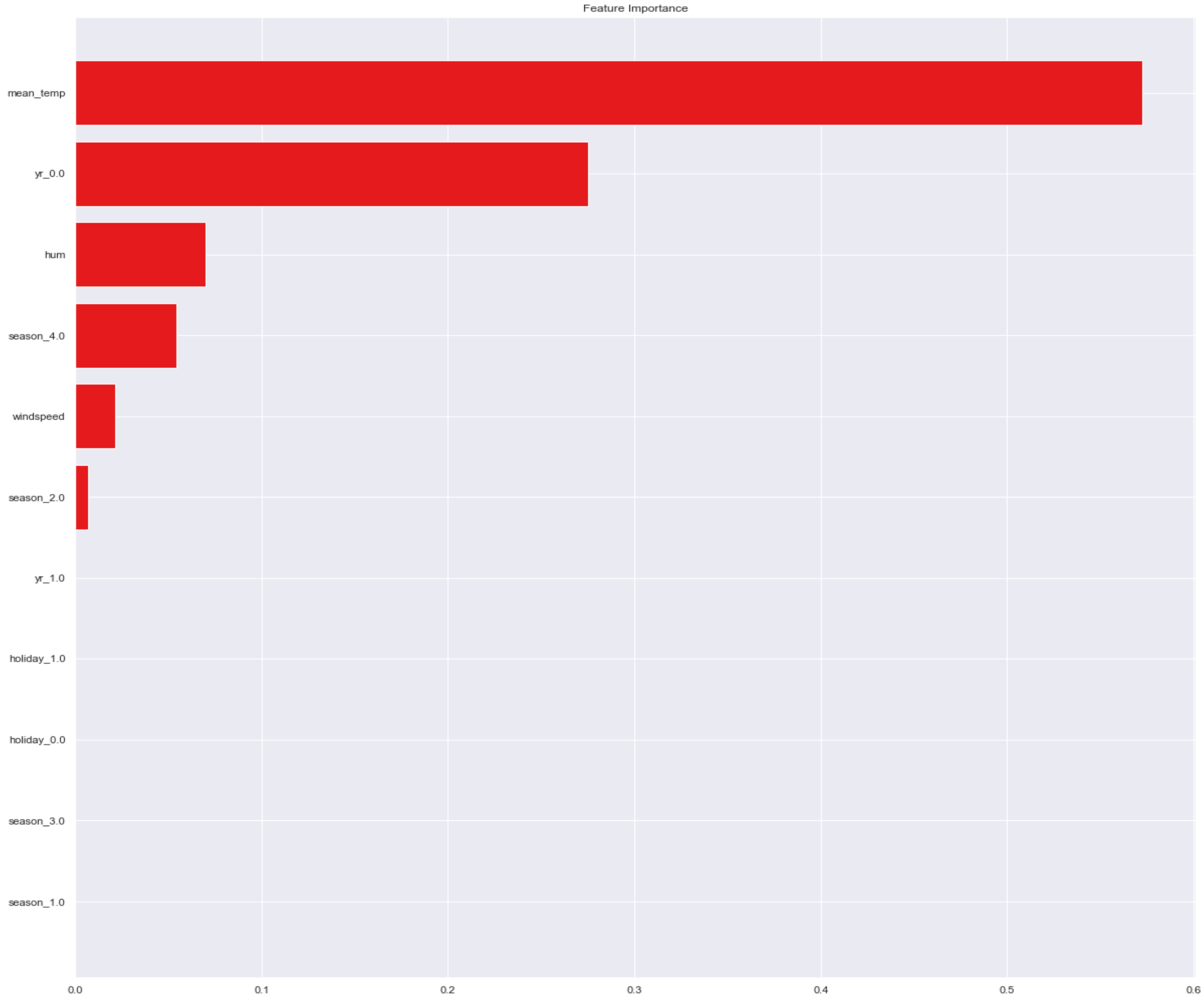
- Bar Plot of Random Forest Feature Importance:-



Decision Tree Regression:

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.913	0.911	11.73	276087.26	525.44	0.160
Test	0.511	0.471	167.54	1718405.67	1310.87	0.543

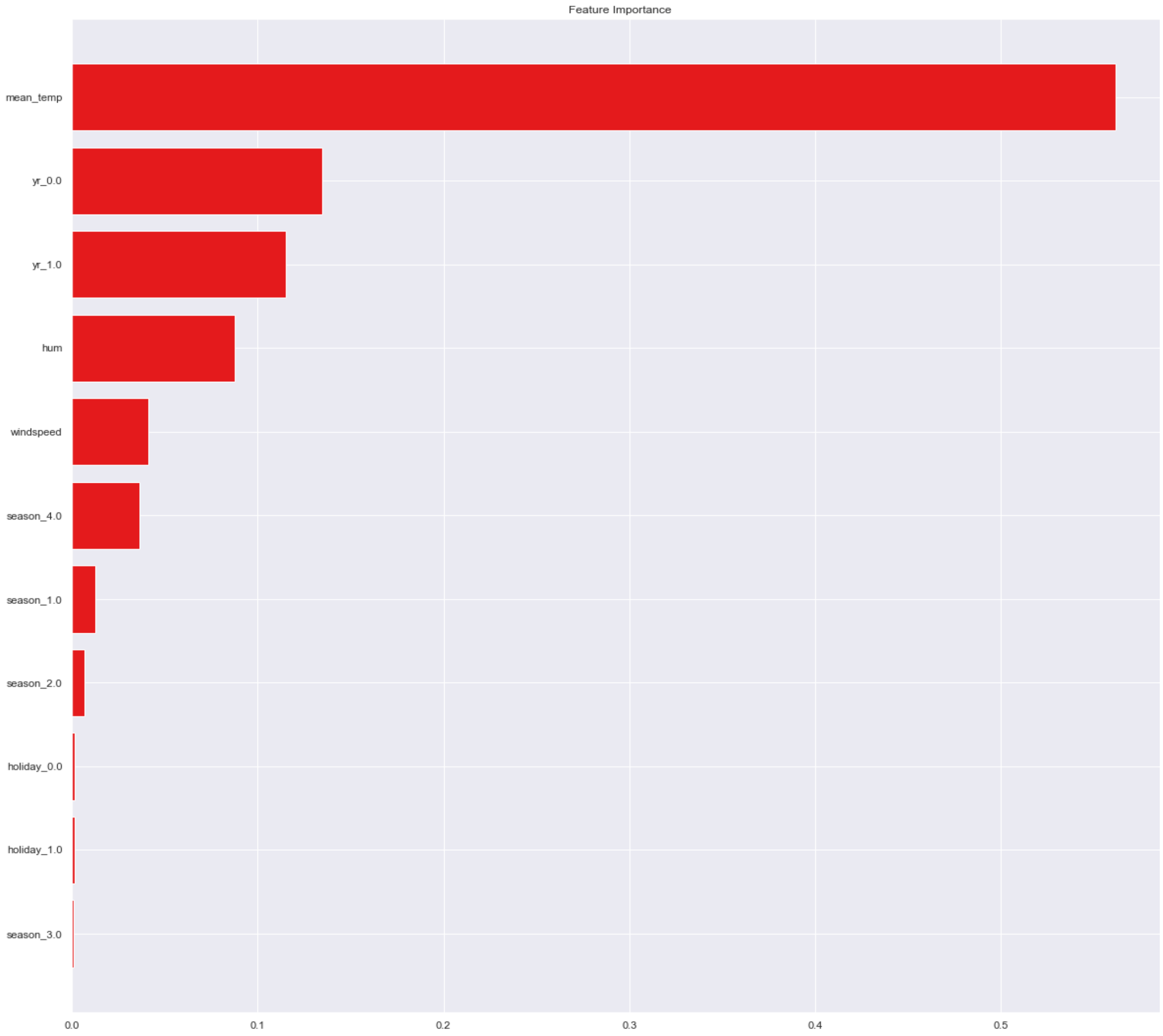
- Bar Plot of Decision tree Feature Importance:-



Random Forest Regression:

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.980	0.980	6.41	61368.633	247.72	0.110
Test	0.553	0.5167	162.20	1570322.70	1253.12	0.535

- Bar Plot of Random Forest Feature Importance:-



Conclusion :-

We have selected Random Forest Regression as our Best Model to Predict Bike Rental Count.

Error Metrics	r square	Adj r sq	MAPE	MSE	RMSE	RMSLE
Train	0.981	0.98	6.29	0.0007	0.027	0.019
Test	0.551	0.515	Inf	0.020	0.144	0.093

- Predicted Bike Rental Count By Random Forest Regression Model:-

	Cnt
dteday	
2012-08-7	7164.005139
2012-08-8	7216.232500
2012-08-9	6661.975841
2012-08-10	5995.545167
2012-08-11	6379.173818

>>>>>>>>>>>>>>>>>> End <<<<<<<<<<<<<<<<<<