**Seoul Bike Sharing Demand Prediction**

**Arunav Goswami, Nayanjoythi Sharma,**

**Mohammed Saad Pasha**

**Data science trainees,**

**AlmaBetter, Bangalore.**

**Abstract:**

In recent days, Pubic rental bike sharing is becoming popular because of is increased comfortableness and environmental sustainability. Dataset provided to us is of a bike sharing system of Seoul city in South Korea which contains weather data for one hour impacting the supply of bikes.

We used different machine learning models like Linear regression, Decision Tress, Random Forest, Gradient Boosting using repeated cross validation approach for predicting the bike count required at each hour for the stable supply of rental bikes. Multiple evaluation indices such as R2, Root Mean Squared Error, Mean Absolute Error are used to measure the prediction performance of the regression models.

**1.Problem Statement**

Bicycle system provides user to rent a bike from one docking station, where user can ride and then return in another docking station. Weather conditions heavily impact the supply of bikes for which the people have to face problems.

Our objective is to build a predictive model so that the company can predict the minimum bicycles required at each hour for stable operation.

The detailed description of the dataset is as follows:

Date: year-month-day

Rented Bike count - Count of bikes rented at each hour

Hour – Hours of the day

Temperature-Temperature in Celsius

Humidity - relative humidity in %

Windspeed – wind speed in m/s

Visibility - 10m

Dew point temperature - Celsius

Solar radiation - MJ/m2

Rainfall – rainfall in mm

Snowfall – snowfall in cm

Seasons - Winter, Spring, Summer, Autumn

Holiday - whether the day is considered a holiday

Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

**2. Introduction**

Traffic transportation is an essential component of a smart city. In certain cities, the taxi demand count is equivalent to the passenger count of certain public transportation modalities.

For example, in Paris 2007, ”velibs” were introduced, and in Amsterdam, there are more bikes than cars. The goal is to facilitate commuting in the city and reduce the amount of cars and the pollution. Indeed, the development of the way to commute reduced the use of cars to go to work and visit the city. It is important to make the rental bike available and accessible to the public, as it provides many alternatives to commuters in metropolises. There are a lot of advantages to bike rents, it is convenient because it permits people not to keep the bike all day long, whether it is at work or at school. Furthermore, it is the healthiest way to travel and it has environmental benefits. The studied dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), The dataset presents the company’s data between December the 1st of 2017 and finishes one year later.

On the dataset, we have 1 year of data. It begins at 2017-12-01 and it ends one year later, 2018- 11-30. The time series has a step of 1 hour, so we have 24 lines of data per day in our DataFrame. The target is the rented bike count per hour and the features are mainly weather features like temperature, wind, rain, snow but also the importance of the studied day

**3**. **Steps involved:**

i) Pre-processing the dataset: - Checking for Missing values, Duplicate values etc.

ii) Exploratory Data Analysis: - Analysing the dependent variable, categorical and numerical variables individually.

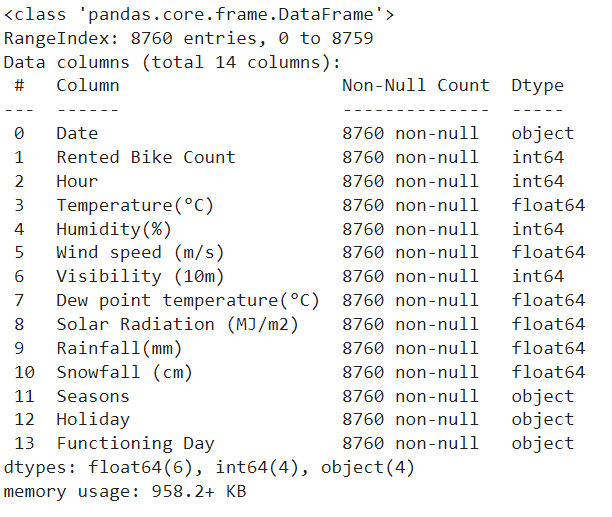
iii) Feature Engineering: - Creation of new features according to our need, dropping of unnecessary data points or features by checking correlation, VIF etc., handling of outliers, Standardization and normalization of features.

iv) Fitting of Machine Learning models with training dataset and evaluating with testing dataset: - like Linear regression, Decision Tress, Random Forest, Gradient Boosting

v) Hypertuning and Explanatory(SHAP) model.

**4. Pre-processing:**

After loading the dataset, we first check concise summary of the data frame:



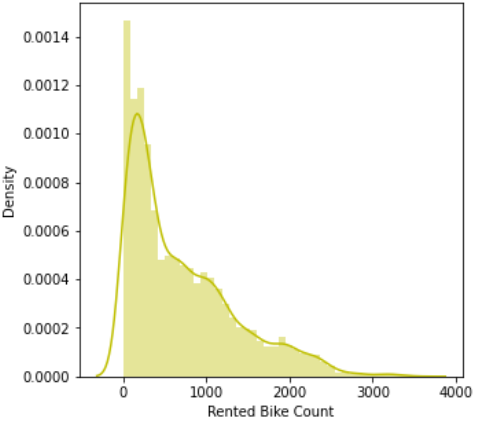
From the dataset information, we can observe that there are 8760 rows (observations) and 14 columns(features). Apparently, there are no null values on the data set. Each row contains weather related data of each hour.

Two types of variables are present: i) Categorical- Date, Seasons, Holiday, Functioning Day.

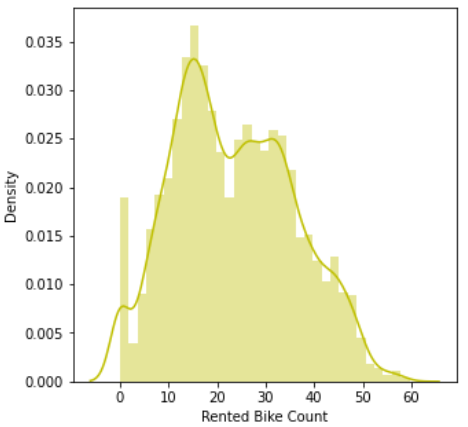
ii) Continuous: Rented Bike Count, Hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar Radiation, Rainfall, Snowfall.

**5. EDA and Feature Engineering:**

**5.A. Analyzing Dependent Variable Rented Bike Count:**

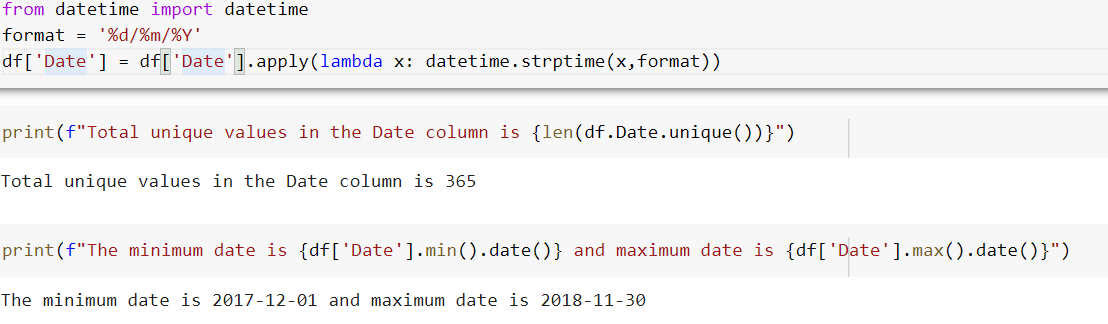


From the histogram, it can be seen that the spread of the data is positively skewed. So, we will use square root normalization technique to normalize the data:



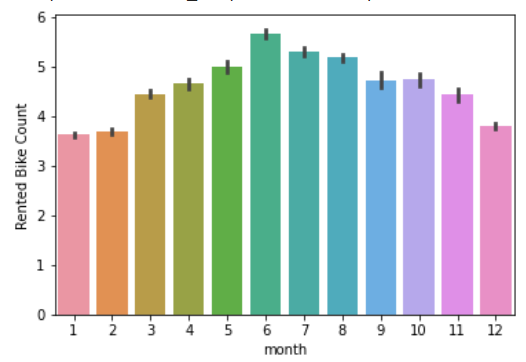
**5.B. Analyzing each categorical variable**:

**i) Date:**



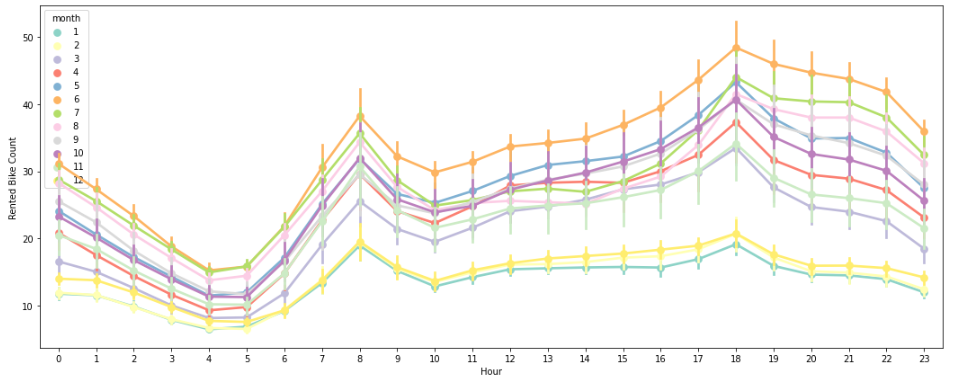
It can be seen that there are 365 days and for each day there are 24 entries (each hour) giving a total of 8760 i.e total no. of rows. But from each day, we cannot get useful information, so instead we create another feature **month**. Since starting date is 01/12/2017 and end date is 30/11/2018, we can neglect the year.

**ii) Month:**



We observe that rental bike count is highest in the month of June.

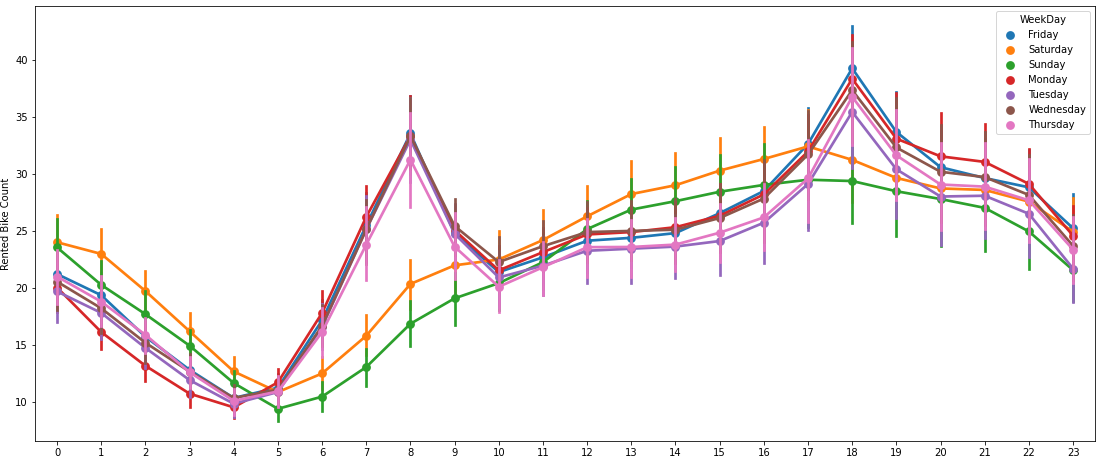
We plot a point plot graph to check the different trends of months with respect to bike count and hour.



From the point plot graph, we can clearly see a **different trend** for the months from December to February (winter season). Thus we can conclude that the usage of bike is lesser in these months than then months.

**iii) WeekDay:**

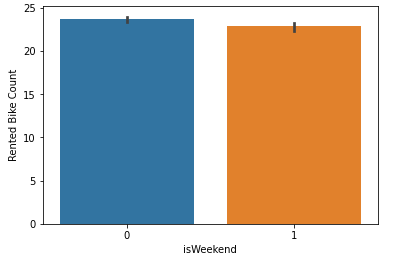
A new feature ‘Weekday’ is created to check bike count per day in a week:



From the point plot graph we can observe that there is a different trend for Saturday & Sunday compared to others. So we can create a new categorical feature where we consider Saturday & Sunday as weekend.

Also, we see same trend in WeekDay, so we can drop the column **WeekDay.**

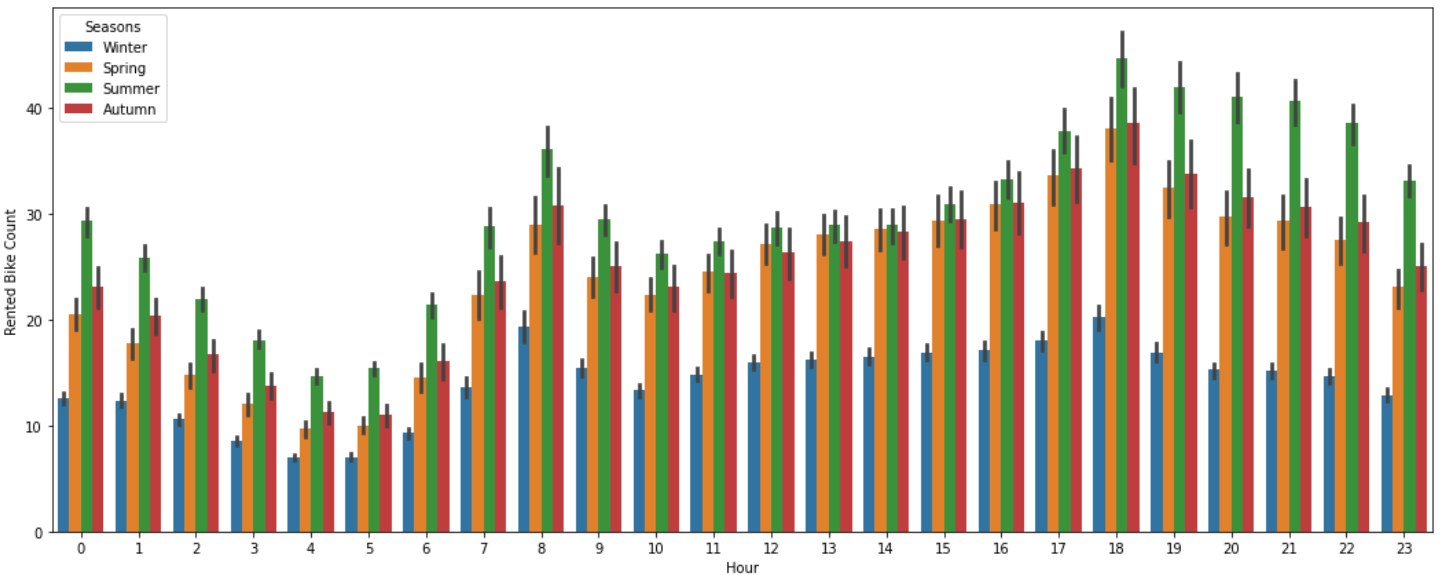
**iv) isWeekend**



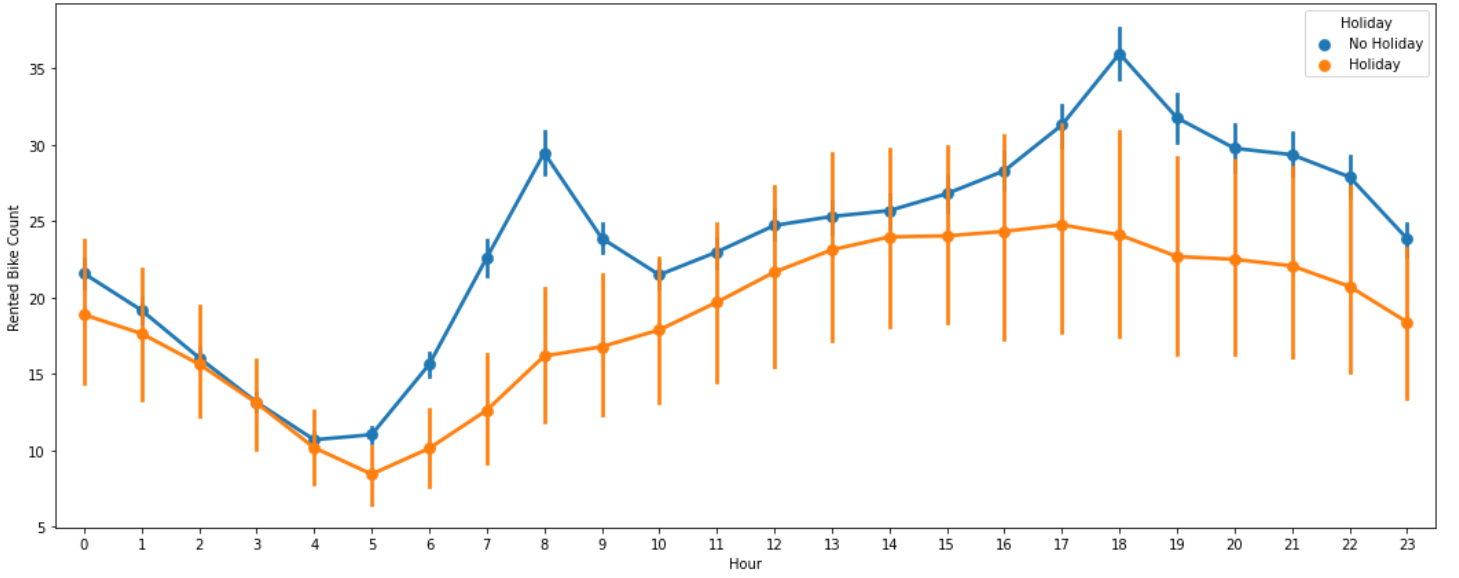
We have considered Saturday and Sunday as **1**  and normal weekdays as **0.**

**v) Seasons**

Visualizing the season wise trend of bike count for each hour:

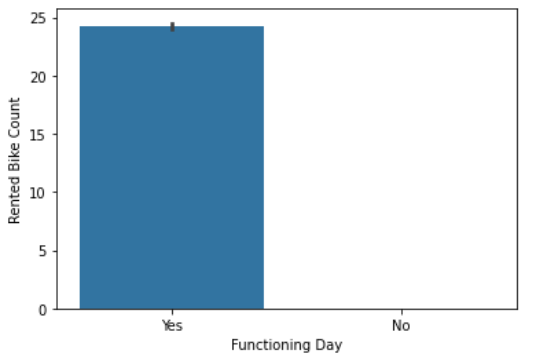
 We can observe that the rental bike count is lowest in winter and highest in summer.

**vi) Holiday**

 From the above plot we can see that rented bike count is lower in holiday compared to working day.

**In working days from 7-9 AM and 5-7 PM there is a sudden spike.**

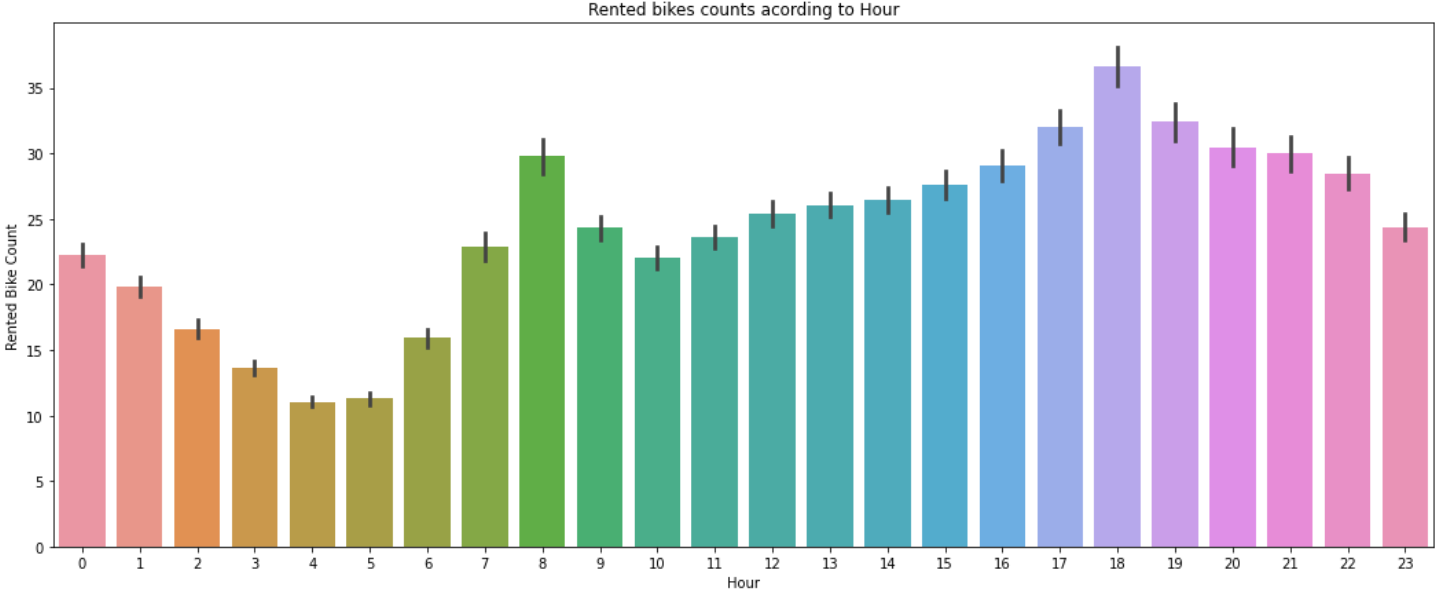
**vii) Functioning Day**

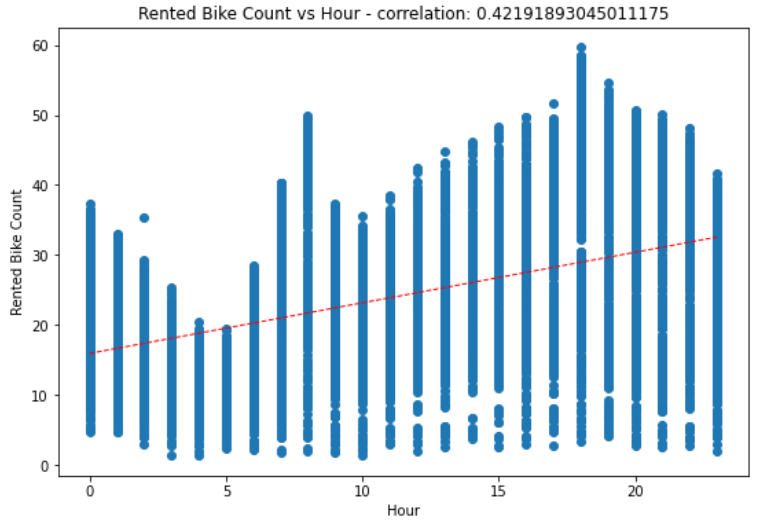


It can be clearly seen that the rented bike count is 0 for a non functioning day. It means that in each hour for that day, the service is discontinued that is why the total rented bike count is 0. So, we choose to remove the rows in which value is "No" and since remaining data points has same value (i.e. "Yes"), we choose to drop the "Functioning Day" column.

**5.C. Analyzing each Numerical variable**:

**i) Hour**





From the regression plot, we see there is a linear correlation between Hour and Rented bike count.

Also we can observe that there is a sudden spike in bike count from 7-9 AM and 5-7 PM. So we create a new categorical feature where we take 7AM to 7PM as **working hour**.

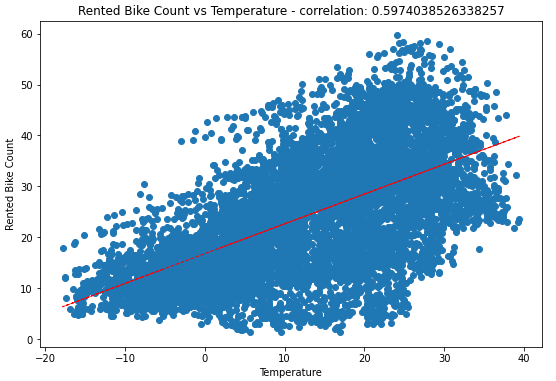
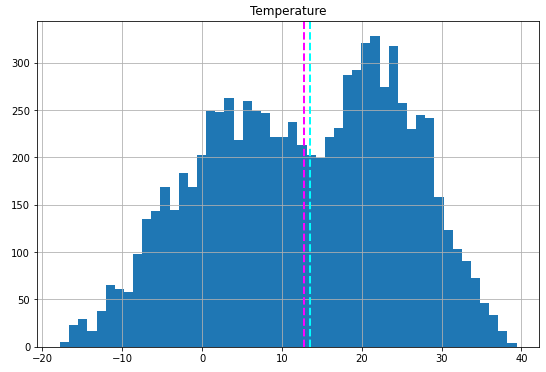
**ii) isWorkingHour:**

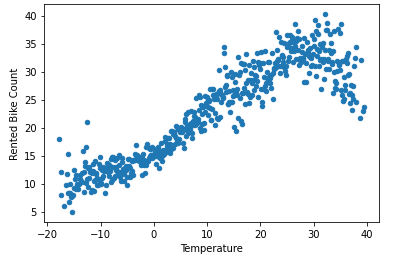


We have taken the working hours i.e. from 7 AM to 7 PM as **1**, and rest non-working hours as **0**.

From the pie chart, we can observe that rented bike count is high(58%) for working hours as compared to non- working hours(42%).

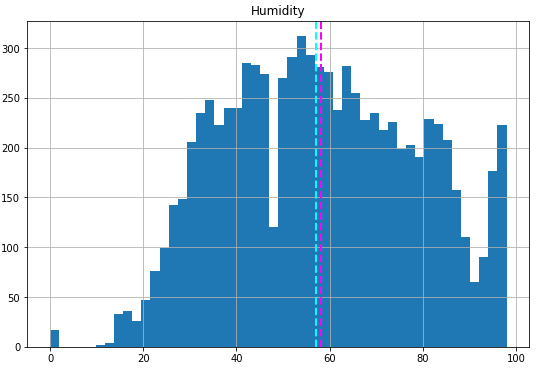
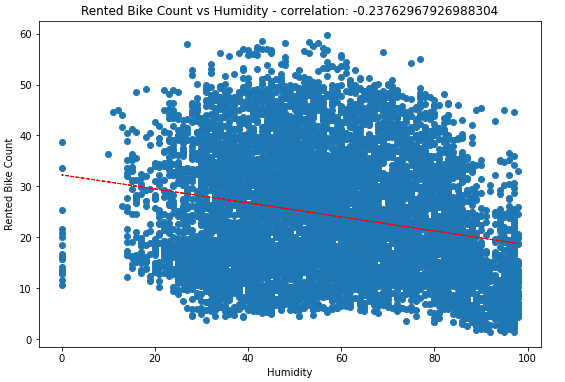
**iii) Temperature:**

We Observe that the feature Temperature follows a normal distribution. Also there is linear trend between Temperature vs Rented bike count. We see that between 20 and 30 degree Celsius, there is a sudden spike



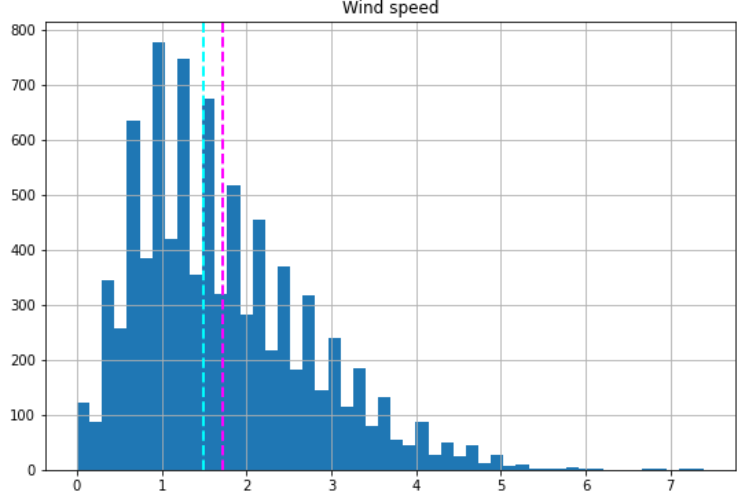
It seems that between 20-30 °C rented bike count increases.

**iv) Humidity:**

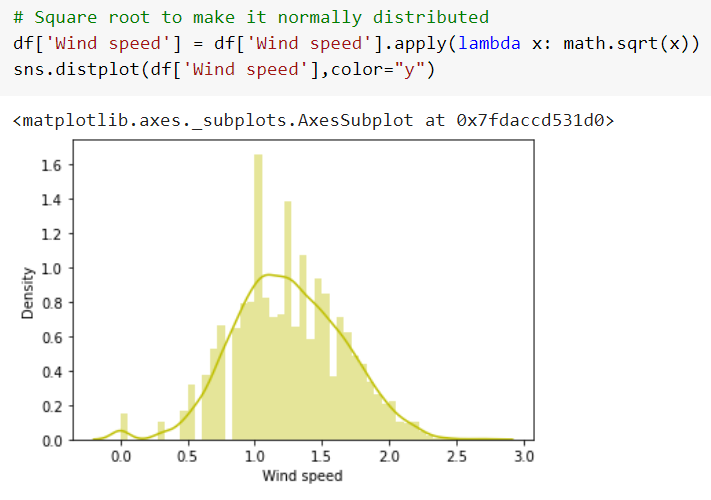
 

There is a slight negative linear relation between Humidity vs Rented bike count. So we can say, increase in humidity leads to decrease in bike count.

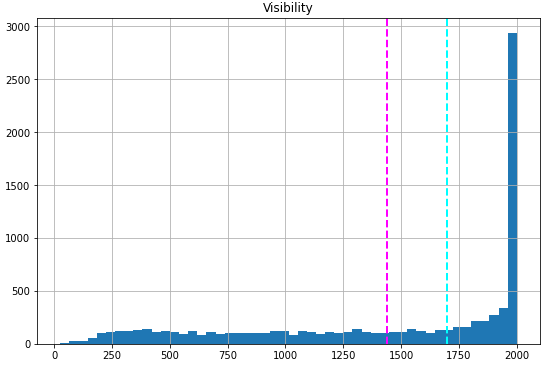
**v) Wind speed**



Wind speed is positively skewed. We use Square root technique to make it normally distributed:



**v) Visibility**



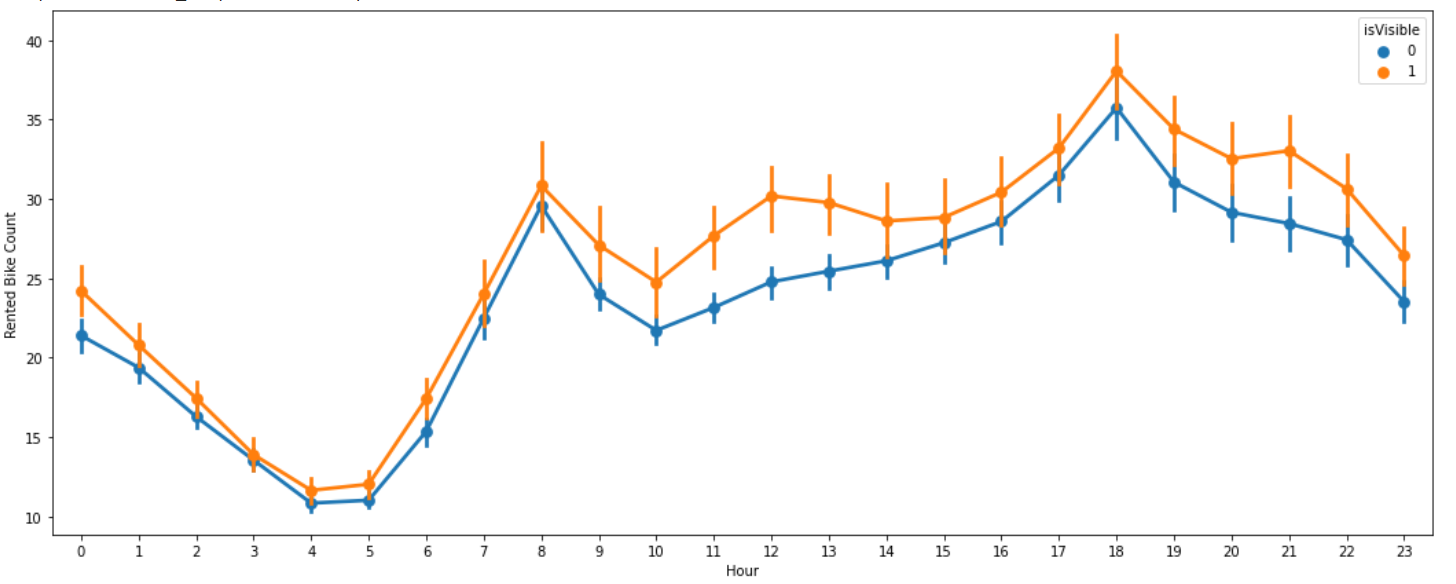
Applying square root transformation on feature Visibility:



Even after applying square root transformation, the values are not normally distributed. It may be because values are very much spreaded and density of 2000 value is much higher. Thus we chose to create a new feature **isVisible.**

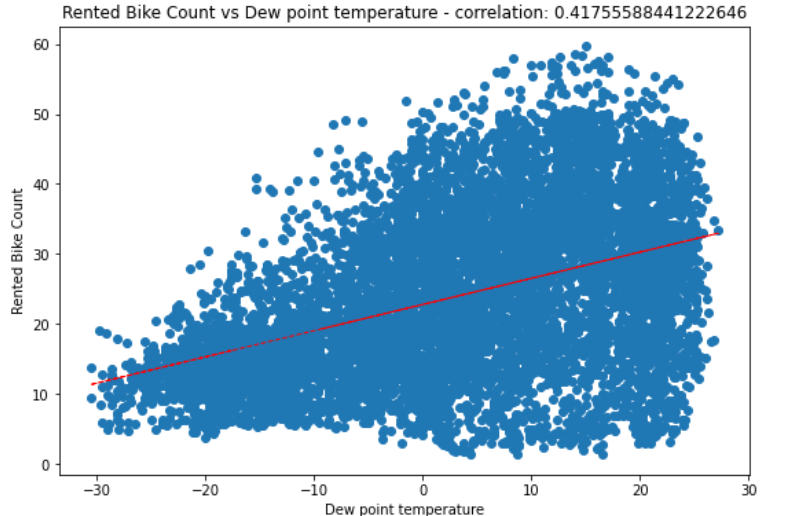
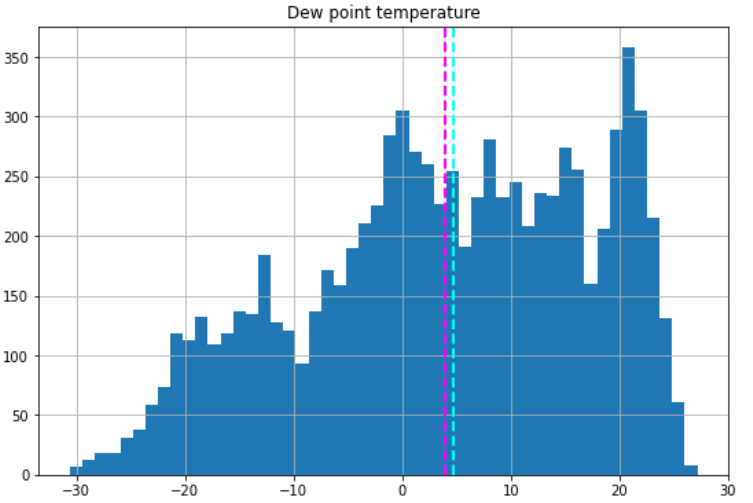
**vi) isVisible:**

Here we have taken the value 2000 as **1** and other values of Visibility feature as **0.**



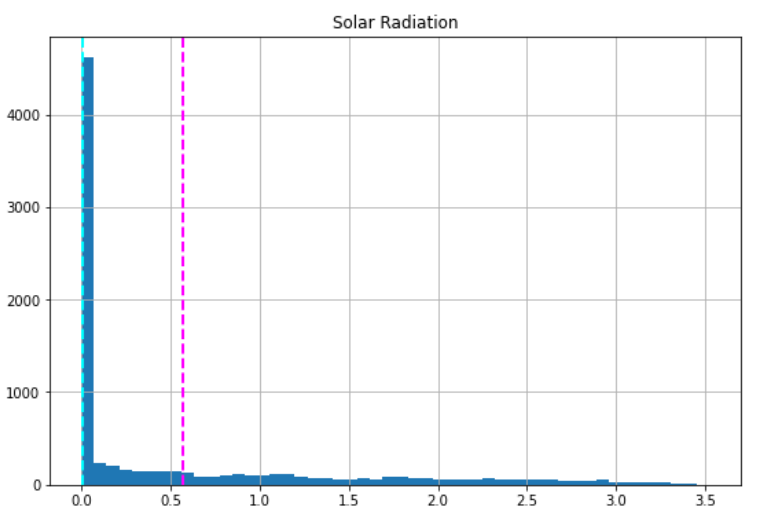
It can be seen that for isVisibility 1 rented bike count is slightly higher than 0.

**vii) Dew point temperature**



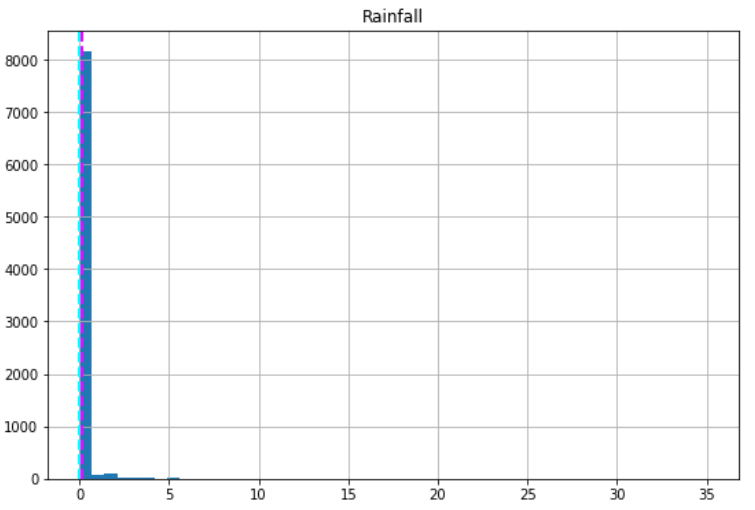
From the above plot, we can see a linear trend between Dew point temperature vs Rented bike count. In other words, we can say that when Dew point temperature increases, bike count increases.

**vii) Solar Radiation**



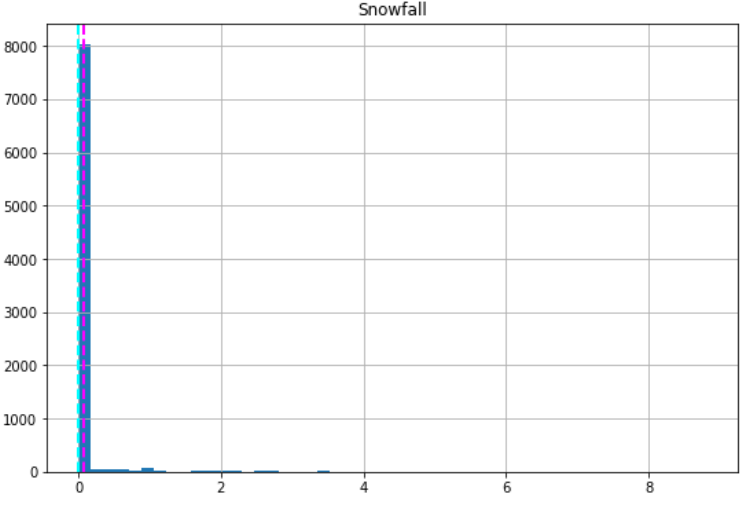
Around 50% of values of the feature ‘Solar Radiation’ are at 0.

**viii) Rainfall**



Around 90% of the values of the feature ‘Rainfall’ are at 0.

**ix) Snowfall**



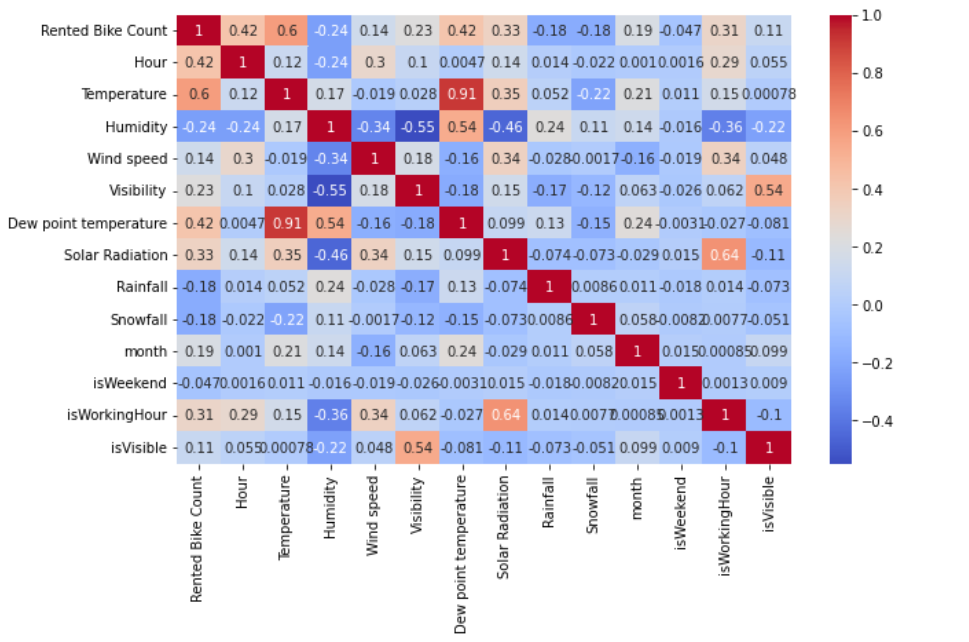
Around 90% of values of the feature ‘Snowfall’ are at 0.

**5.D. Encoding:**

There are 2 categorical features Holiday & Seasons which are needed to be encoded. We use Label Encoding here:



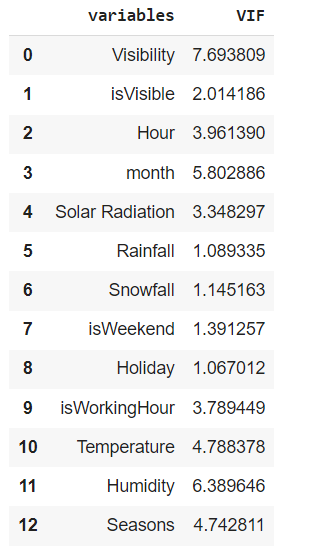
**5.E. Feature Selection using Correlation:**



From the heatmap, we can clearly see that Temperature & Dew point temperature is strongly correlated. So, even if we drop any one of them, it will not affect the data. We drop Dew point temperature.

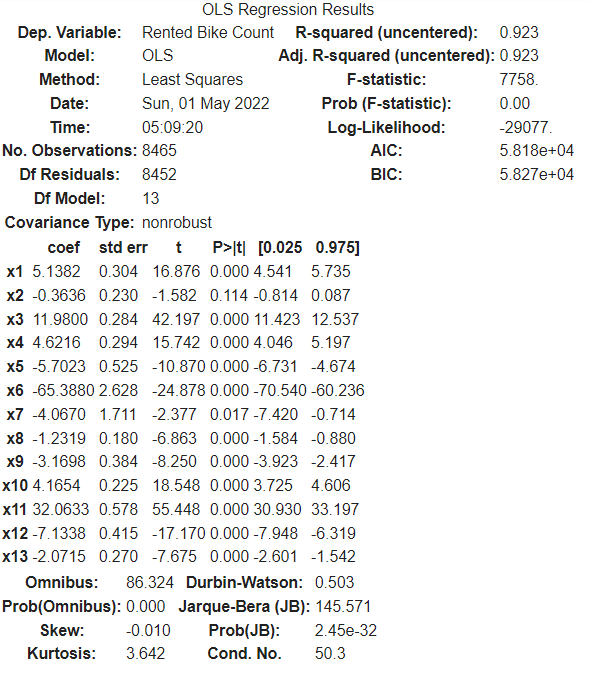
**5.F. Variance Inflation Factor (VIF):**

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Values of VIF that exceed 10 are often regarded as indicating multicollinearity.



Since the value of VIF for all variables is less than 10, we will take all of it.

**5.G. Statistical Test using OLS Method**

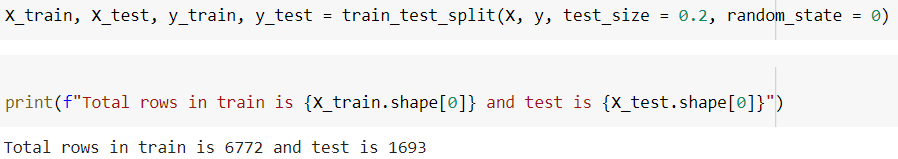


If the P-Value is less than the significance level (0.05) then we can say, our model fits the data well.

Here, for x2 (i.e ‘isVisible’), p > 0.05, so we can remove the isVisible’ feature.

**6. Fitting of Machine Learning Models:**

**6.A. Train Test split:**



The dataset is split into the Training set and Test set in 80:20 ratio.

**6.B. Transformation of features:**

We have used MinMaxScaler technique Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

**6.C. Algorithms:**

We will use various regression models to predict our target variable i.e. ‘Rented Bike Count per hour’:

**i) Linear Regression**: Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. And our main objective in this algorithm is to find this best fit line.

**ii) Decision Tree:**

Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represent the result of the test, and each leaf node represents the final decision or result. Decision trees are good at capturing non-linear interaction between the features and the target variable.

**iii) Random forest:**

The Random Forest regression is an ensemble learning method which combines multiple decision trees and predicts the final output based on the average of each tree output. The combined decision trees are called as base models. Random forest uses Bagging or Bootstrap Aggregation technique of ensemble learning in which aggregated decision tree runs in parallel and do not interact with each other.

**iv) Gradient Boosting:**

It utilizes a gradient descent algorithm that can optimize any differentiable loss function. An ensemble of trees is constructed individually, and individual trees are summed successively. The next tree tries to restore the loss (difference between actual and predicted values).

**7. Model Evaluation Metrics:**

We will use following regression model evaluation metrics to check the performance of our machine learning model:

**i) MSE (Mean Squared Error):** Mean Squared Error is one of the most preferred metrics for regression tasks. It is the average of the squared difference between the target value and the value predicted by the regression model.

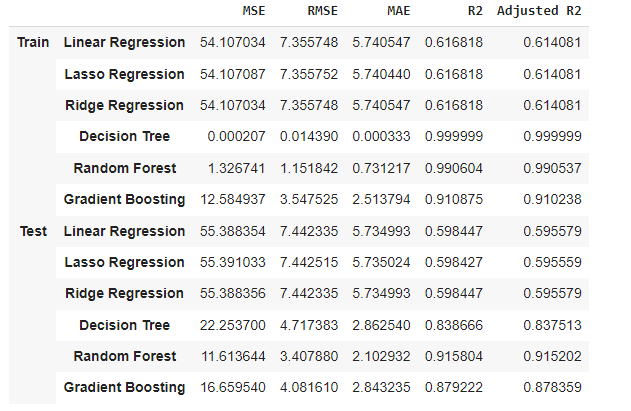
**ii) RMSE (Root Mean Squared Error):** RMSE is the square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred because the errors are first squared before averaging which poses a high penalty on large errors.

**ii) MAE (Mean Absolute Error):** MAE is the absolute difference between the target value and the value predicted by the model. The MAE is more robust to outliers and does not penalize the errors as extremely as MSE.

**iv) R2 (R – Squared):** Coefficient of Determination or R² is a metric which helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean.

**v) Adjusted R2:** Adjusted R² is always lower than R² as it adjusts for the increasing predictors and only shows improvement if there is a real improvement.

**After applying machine learning models and calculating evaluation metrics for each model, we find results as follows:-**



From the results, we see that Random Forest gives highest R2 and Adjusted R2 both for training and testing data with least difference i.e. model is neither overfit nor underfit. So we use Random Forest to find best hyperparameters and apply hyperparameter tuning.

**8. Hyper parameter Tuning and Cross Validation:**

Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins.

The Hyperparameters chosen by us are:

n\_estimators :- number of trees in the random forest

max\_features :- number of features in consideration at every split

max\_depth :- maximum number of levels allowed in each decision tree

min\_samples\_split :- minimum sample number to split a node

min\_samples\_leaf :- minimum sample number that can be stored in a leaf node

bootstrap :- method used to sample data points

Two best strategies for Hyperparameter tuning are:

* GridSearchCV
* RandomizedSearchCV

i) RandomizedSearchCV:- Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It is good in testing a wide range of values and normally it reaches a very good combination very fast, but the problem here is that it doesn’t guarantee to give the best parameters combination.

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. Thisapproach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

**9. Conclusion.**

i) We observed that the bike rental count is high on non holiday than on holiday.

ii) During weekdays at 7-9 AM and 5-7 PM, there are sudden spikes in bike count.

iii) The bike count is high at high temperatures.

iv) Rental bike count is highest in the month of June.

iv) In summer the bike count is the highest and in winter it is the lowest.

v) When we compare the RMSE and Adjusted R2 of all the models for test data, Random Forest gives the highest Score where the Adjusted R2 score is 0.91 and RMSE is 3.4. So, we can conclude that this model is the best for predicting the bike rental count on hourly basis.