**BIKE SHARING PREDICTION DEMAND**

**Swapnil Bhagawat, Sumit Zoting,**

**Bhagyashree Bhoyer**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Our study used one dataset consisted of weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

My experiment can help you understand what could be the best performing model for our project.

1. **PROBLEM STATEMENT:**

The contents of the data came from a city called Seoul. A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short-term basis for a price or free.

The data had variables such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count. The problem statement was to build a machine learning model that could predict the rented bikes count required for an hour, given other variables.

* 1. **Seoul Bike Data DF:**

The contents present in Seoul\_Bike\_Data Dataset are:

• **Date** - year-month-day

• **Rented Bike count** - Count of bikes rented at each hour

• **Hour** - Hour of the day

•**Temperature**-Temperature in Celsius

• **Humidity** -Humidity in %

• **Windspeed** –Windspeed in m/s

• **Visibility** – Visibility in10m

• **Dew point temperature** - Dew point temperature in °C

• **Solar radiation** -Solar radiation in MJ/m2

• **Rainfall** - Rainfallin mm

• **Snowfall** -Snowfall in cm

• **Seasons** - Winter, Spring,

Summer, Autumn

• **Holiday** - Holiday/No holiday

•**Functional Day** – NoFunc (Non-Functional Hours), Fun (Functional Hours)

1. **INTRODUCTION:**

Exploratory Data Analysis (EDA) is for understanding and analyzing the datasets given to get better insights into the data and can be used for taking important business decisions. We can summarize the main characteristics of the datasets and can plot the data visually by using Python libraries. There are various plots like Heatmap, Bar plot, Point Plot, Box Plot, etc. by using libraries Matplotlib, Seaborn. We can store our data in data frames by using the Pandas library from Python.

1. **STEPS INVOLVED:**
   1. **Data Exploration:**

There are data sets which were given Seoul\_Bike\_Data dataset and user review dataset in which one dataset was having 8760 rows and 14 columns.

* 1. **Data Cleaning:**
* **Null Values/ NaN Values Treatment / Removing Duplicate Values:**

In the Seoul\_Bike\_Data, there were no null values present and no duplicate values present in dataset.

* **Datatype Correction:**

**Convert the "date" column into 3 different columns i.e., "year", "month", "day". The "year" column in our data set is basically contain the 2 unique number contains the details of from 2017 December to 2018 November so if i consider this is a one year then we don't need the "year" column so we drop it. The other column "day", it contains the details about each day of the month, for our relevance we don't need each day of each month data but we need the data about, if a day is a weekday or a weekend so we convert it into this format and drop the "day" column.**

* **Outlier-Detection:**

An outlier is a point or set of points that are different from other points. Sometimes they can be very high or very low. It’s often a good idea to detect and remove the outliers. Because outliers are one of the primary reasons for resulting in a less accurate model. Hence, it’s a good idea to remove them

* 1. **Data Analysis:**

It is a systematic way of describing and illustrating, condensing, and evaluating data by systematically using statistical or logical methods.

From the given data I have analysed on the basis of Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall etc. to get insights from the data.

* 1. **Data Visualization:**

Python's libraries provide plenty of features with which users can create highly customized, elegant, and interactive plots, making data visualization with Python one of the most widely used features in today's data science environment.

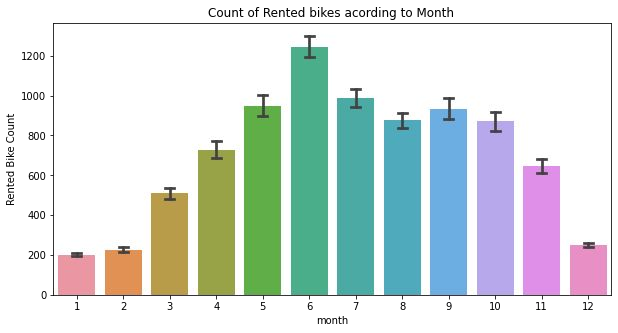
Data Visualization libraries in Python are:

* **Matplotlib:** With a Python library, you can visualize arrays in 2D. NumPy is used to create Matplotlib, which is written in Python. There is a wide variety of plots in Matplotlib, including line, bar, scatter, histogram, etc. that can help us understand trends, patterns, correlations.
* **Seaborn:** A Python library for representing statistics with datasets, Seaborn is a dataset-oriented library. This library is built on matplotlib and creates a variety of visual representations. Pandas data structures are incorporated into it. To create informative visuals, the library performs mapping and aggregation internally

1. **Visualizations:**
   1. **Count of rented bikes according to Month**

* **Bar Plot:**

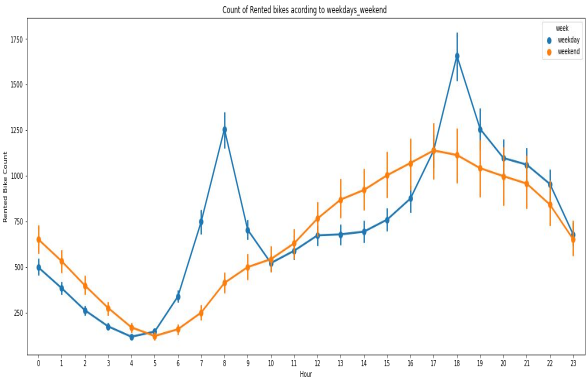
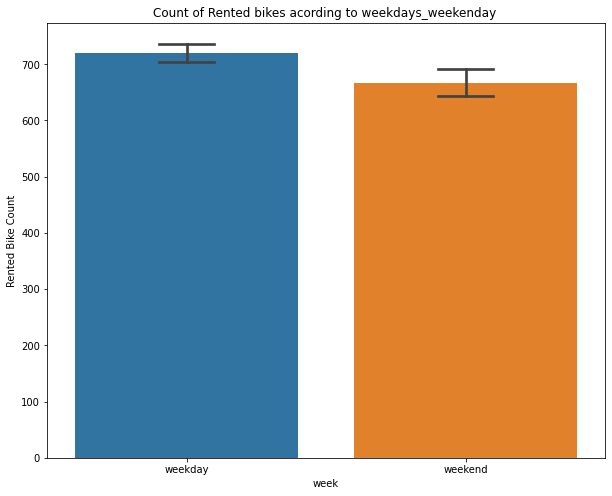
The demand of the rented bike is high from the month 5 to 10



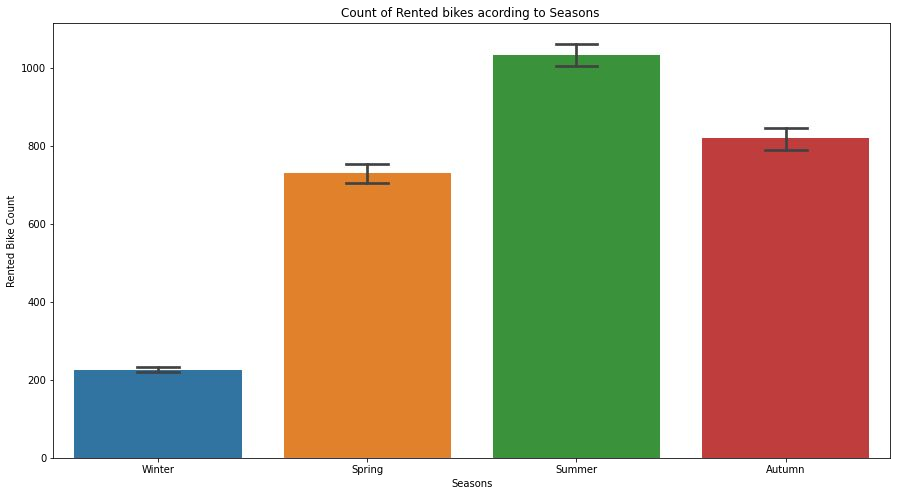
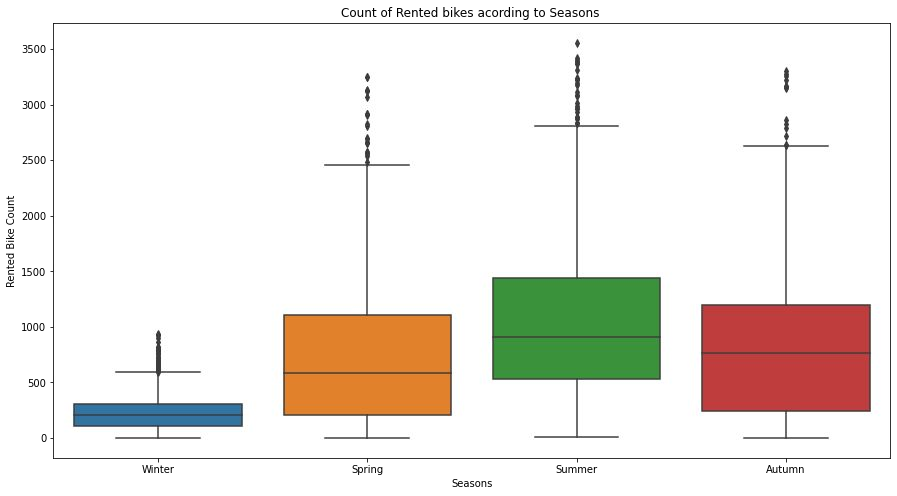
* 1. **Count of rented bikes according to weekdays weekend**
* **Bar Plot & Point Plot:**

From the point plot and bar plot we can say that in the week days which represent in blue color show that the demand of the bike higher because of the oﬃce. Peak Time are 7 am to 9 am and 5 pm to 7 pm

The orange color represents the weekend days, and it show that the demand of rented bikes is very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

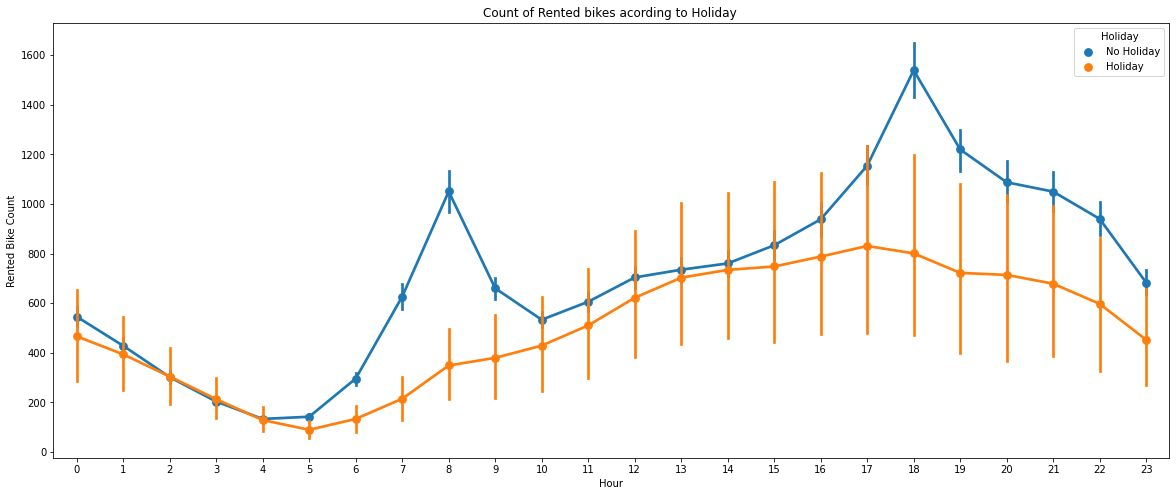


* 1. **Count of rented bikes according to seasons**
* **Box Plot & Bar Plot:**

In the box plot and bar plot which shows the use of rented bike in in four different seasons, and it clearly shows that, in summer season the use of rented bike is high in winter season the use of rented bike is very low because of snowfall.  

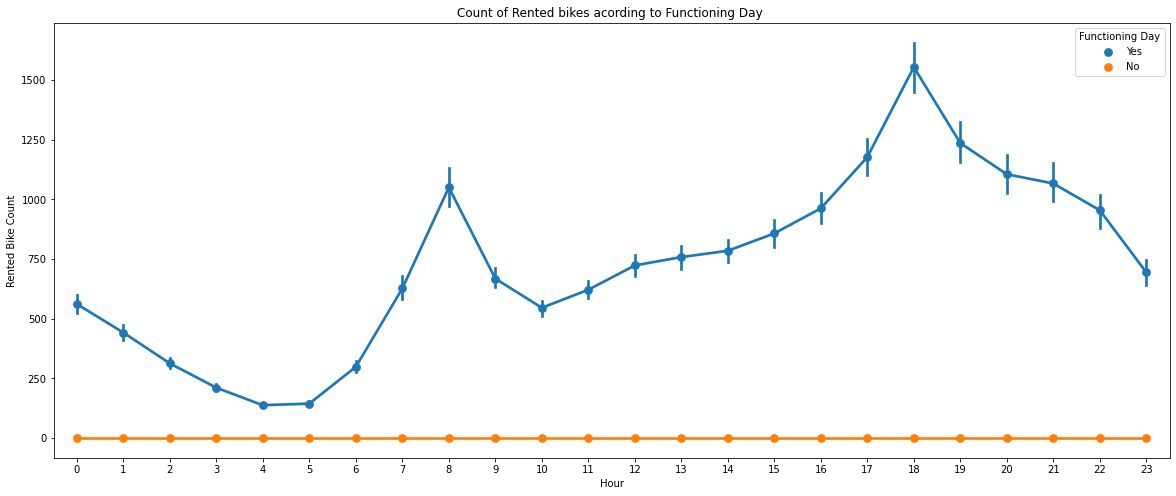
* 1. **Count of rented bikes according to Holiday**
* **Point Plot:**

In the point plot which shows the use of rented bike in a holiday, and it clearly shows that, plot shows that in holiday people uses the rented bike from 2pm-8pm

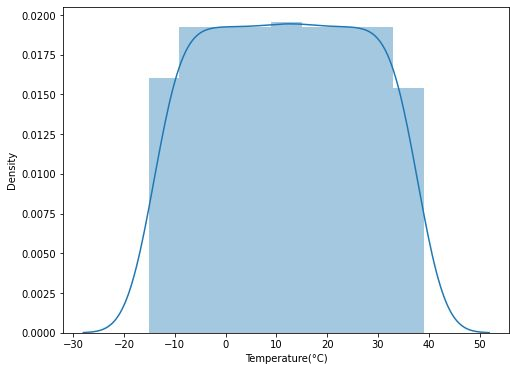


* 1. **Count of rented bikes according to functioning day**
* **Point Plot:**

In the point plot which shows the use of rented bike in functioning day or not, and it clearly shows that, Peoples don’t use rented bikes in no functioning day

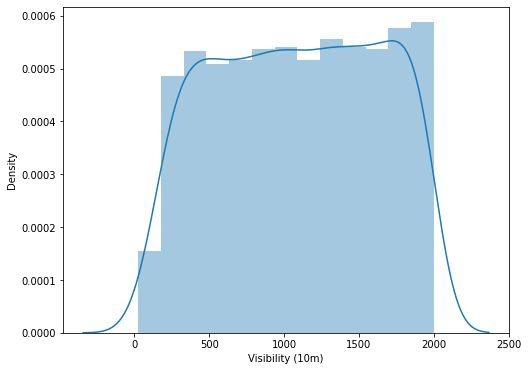


* 1. **Numerical Variables**
* **Temperature-**
* **DistPlot:**

****Plot shows that people tend to rent bikes when the temperature is between -5 to 25 degrees.

* **Visibility-**
* **DistPlot:**

Plot shows that people tend to rent bikes when the visibility is between 300 to 1700.

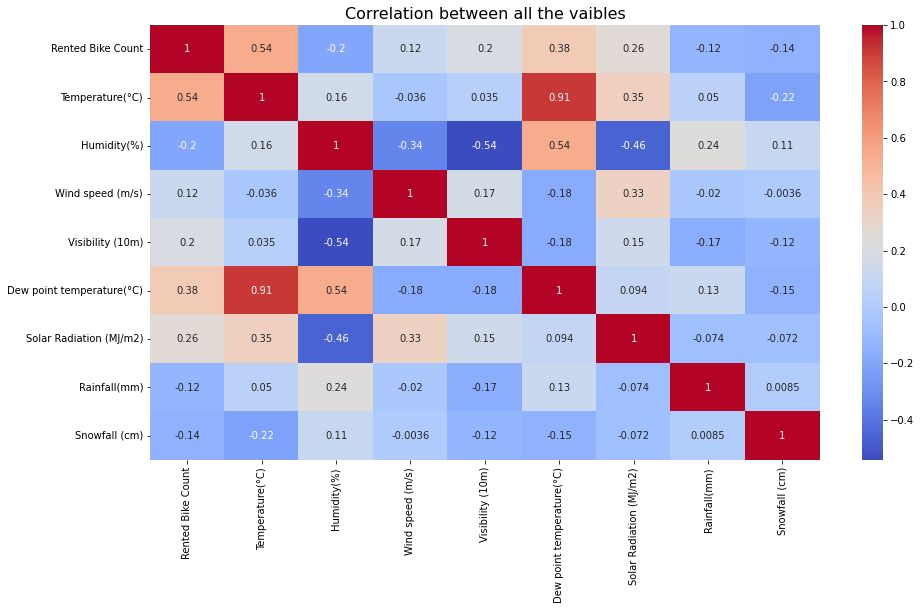


* 1. **Correlation**
* **HEAT MAP:**

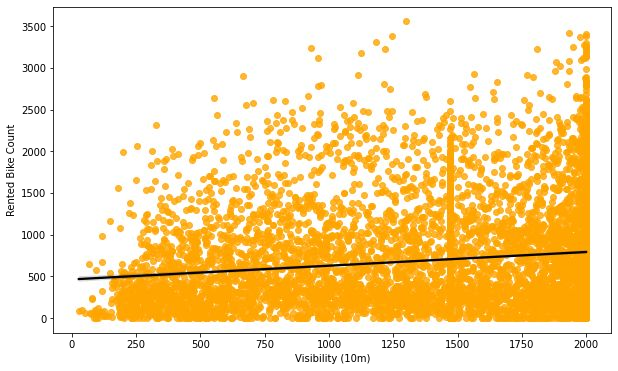
Heatmap is defined as a graphical representation of data using colors to visualize the value of the matrix. In this, to represent more common values or higher activities brighter colors basically, reddish colors are used, and to represent less common or activity values, darker colors are preferred.

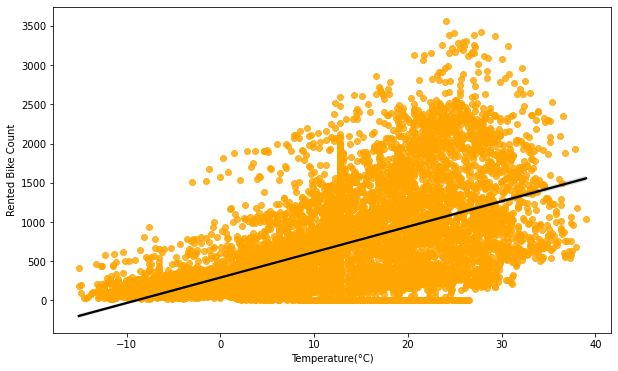
* **Corr() function:**

To determine the pairwise correlation of all columns in the data frame, Pandas dataframe.corr() is used.

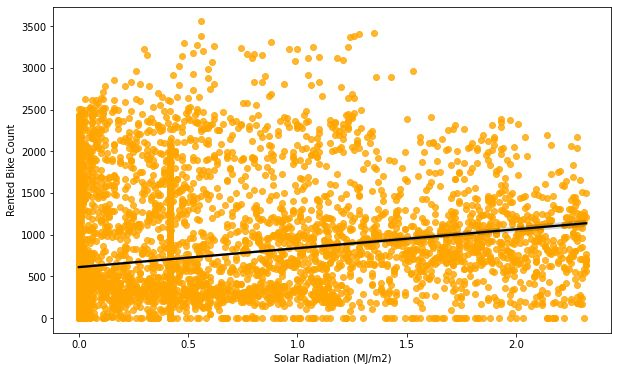


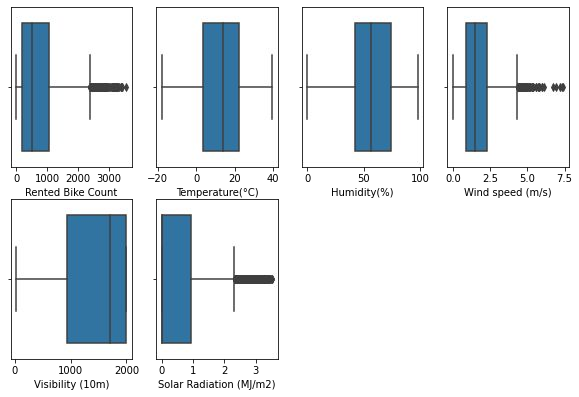
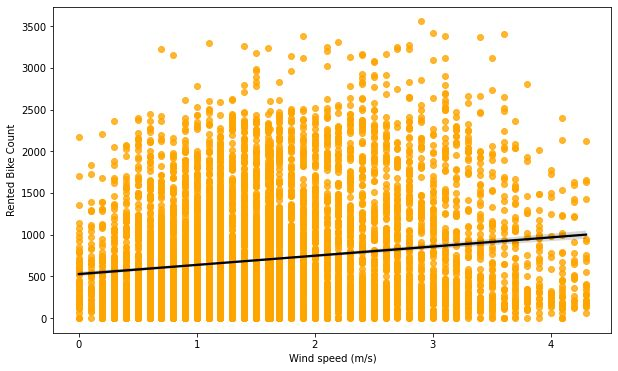
From the above heat map i can conclude that Temperature and Dew point temperature(°C) has the high correlation. we drop this column then it doesn’t affect the outcome of our analysis.



After removing the Dew point temperature

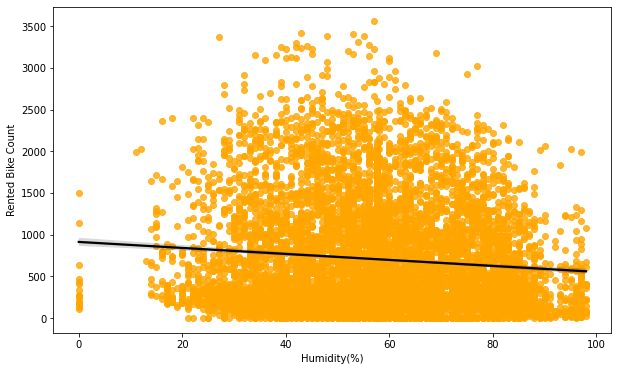
* 1. **Handling outliers**

An Outlier is a data-item/object that deviates signiﬁcantly from the rest of the (so-called normal) objects, The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outlier.



1. **Regression Plot**

Temperature, solar radiation, wind speed, visibility are positively related to target variable, the rented bike count increases with increase of these features.

Humidity is negatively related to target variable, the rented bike count decreases with increase of these features.

1. **ML Algorithms Results:**

**6.1 Linear regression:**

MSE: 61.15584375724336

RMSE: 7.820220185982192

MAE: 5.923260387972038

R2: 0.6028600936479797

Adjusted R2: 0.5979004370931673

**6.2 Ridge regression**

MSE: 60.83384712272495

RMSE: 7.799605574817546

R2: 0.6148937632652991

Adjusted R2: 0.6100843884309619

**6.3 Elastic net**

MSE: 61.36957118125049

RMSE: 7.833873319198523

MAE: 5.93382066577593

R2: 0.6014721692250582

Adjusted R2: 0.5964951796640391

**6.4 Decision tree**

* The r2 score of decision tree is 0.7619882672759375
* The r2 score of decision tree with hyper parameters tunning is 0.8013257980814106

**6.5 Random Forest regressor**

MSE= 7.548469131956553

RMSE= 2.747447748721812

R2\_Score\_train= 0.9509809996888273

MSE= 19.900384576287564

RMSE= 4.460984709264039

R2\_Score\_test= 0.8740214111679199

**6.6 SVR**

The MAE of training set = 4.312974309168605

The MSE of training set = 44.525880818380806

The R2\_score of training set = 0.7108534025195494

The MAE of test set = 4.674553587050071

The MSE of test set = 46.854036819585005

The R2\_score of test set = 0.703392393398722

**6.7 Gradient boosting**

R2 score of training data: 0.91%

R2 score of test data: 0.869495

1. **CONCLUSION:**

* Hour of the day holds most importance among all the features for prediction of dataset
* It is observed that highest number bike rentals count in Summer and Autumn Seasons and the lowest in Spring season.
* We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day
* The top 5 important features of our dataset are: Season winter, Temperature, Hour, Season autumn, Humidity
* Peoples don’t use rented bikes in no functioning day
* People tend to rent bikes when the temperature is between -5 to 25 degrees
* People tend to rent bikes when the visibility is between 300 to 1700
* For all the above experiments we can conclude that gradient boosting and random forest regressor with using hyperparameters we got the best results

1. **References:**

* https://www.analyticsvidhya.com
* https://www.kaggle.com
* https://www.geeksforgeeks.org/overview-of-data-science/