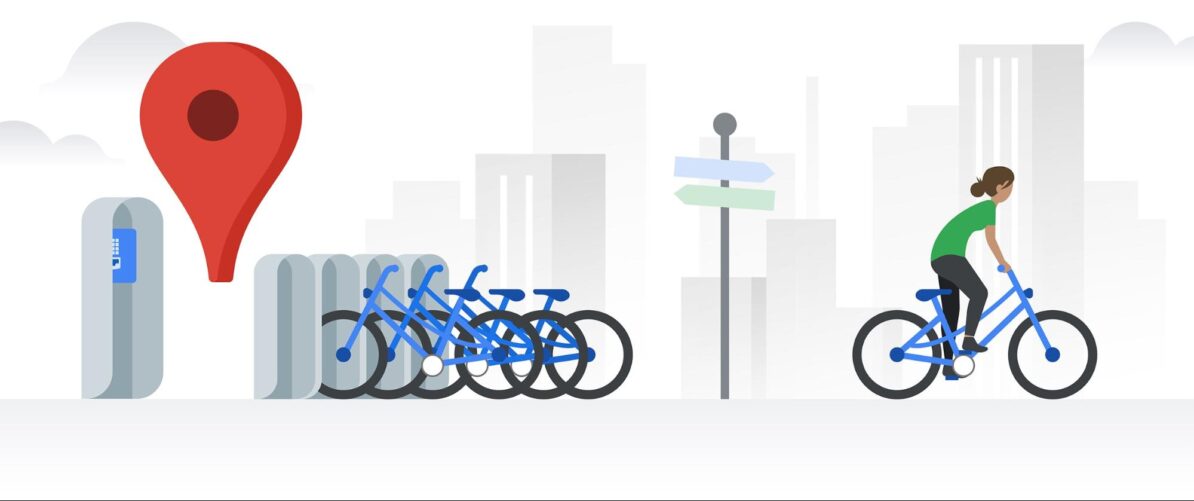
**Capstone Project-2 Submission**

**Seoul Bike Sharing Demand Prediction**

**Sonika Baheti**

**Data science trainee**



**AlmaBetter, Bangalore**

**Cohort: Chicago**

---------------------------------------------------------\*\*\*-------------------------------------------------------------

**Sonika Baheti** – [bahetisonika50@gmail.com](mailto:bahetisonika50@gmail.com)

**GitHub Link~**

**Sonika Baheti:** - https://github.com/sonika-07/Bike-Sharing-Demand-Prediction

**Abstract-** This research paper presents a rule-based regression predictive model for bike sharing demand prediction. A bike-sharing system provides people with a sustainable mode of transportation and has beneficial effects for both the environment and the user. In recent days, Pubic rental bike sharing is becoming popular because of is increased comfortableness and environmental sustainability. Data used include Seoul Bike and Capital Bikeshare program data. Data have weather data associated with it for each hour. For the dataset, I am using linear regression model and tree based models with hyperparameters tuning and testing set is used for evaluation. Multiple evaluation indices such as 𝑅2, adjusted R2, Root Mean Square error are used to measure the prediction performance of the regression models.

***Key Words:*** Rented bike sharing, Exploratory Data Analysis, Modelling, Regression Analysis, Visualization.

# OBJECTIVE

This project focuses on predicting the rented bike count in the city of Seoul given weather conditions. **Data is Money, Information is Power.** I tried to gain power using Exploratory Data Analysis by unfolding it through various graphs, charts and tables.

The goal of this project is to combine the historical bike usage patterns with the weather data to forecast bike rental demand. The data set consists of hourly rental data of one year.

The main objective is to find out the best model which can predict the rented bike count with maximum accuracy.

# 2. INTRODUCTION

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

   !

"! " # $"" !% 

 !

!"#

!&%

 !  #   !  

 !   !   !# 

%

' !  !   

! !#  !     "

! !      ! !   

 ( "

%

   !

"! " # $"" !% 

 !

!"#

!&%

A bike rental or bike hiring business rents out motorcycles for short period of time, usually for few hours.. These provides an excellent opportunity for those who would like to avoid shipping their own bikes but like to do multi day bike tour of a particular area.

# 3. PROBLEM STATEMENT

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.

# 4. data description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

Date : year-month-day

Rented Bike count - Count of bikes rented at each hour

Hour - Hour of he day

Temperature-Temperature in Celsius

Humidity - %

Windspeed - m/s

Visibility - 10m

Dew point temperature - Celsius

Solar radiation - MJ/m2

Rainfall - mm

Snowfall – cm

Seasons - Winter, Spring, Summer, Autumn

Holiday - Holiday/No holiday

Functional Day – No Func(Non Functional Hours), Fun(Functional hours)

# 5. STEPS INVOLVED

* **Data Overview**

As a first step we take the overview of data, where we specially made our focus on understanding what each column means. So that we can be clear from what perspective we have to analyze our data. After understanding different column, we mark that some of the features are numerical while some are categorical features**.**

* **Breakdown of Dataset**

Before proceeding to data visualization, we need to perform the following steps:

1. Importing required packages for future analysis.

2. Mounting drive and reading data files from Google drive.

3. Removing future warning seaborn plots.

4. Visualizing all the columns of the respective data frame.

5. Viewing all data information.

6. Checking duplicates if any then drop.

7. Checking unique values, null count, datatypes.

8. Segregation of numerical and categorical data.

9. See the unique values for categorical columns and value count of same.

* **Data Wrangling**
* Changed the datatype to ‘Date’ column to ‘datetime’.
* Created New columns ‘Month’ and ‘weekdays\_weekend’ to have more deeper analysis.
* Dropped the column ‘Date’ because I won’t be using this.

Here I’m considering the following columns as categorical columns : ***['Seasons’,'Holiday', 'Functioning Day', 'month', 'weekdays\_weekend’]***

And others as numerical columns: ***['Rented Bike Count', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)’,  ‘Visibility (10m)', 'Dew point temperature(°C)','Solar Radiation (MJ/m2)’,'Rainfall(mm)', 'Snowfall (cm)']***

* **Exploratory Data Analysis**

After establishing a good sense of each feature, I have performed EDA. EDA is the process of trying to understand data in the ways possible in order to derive insights from it. Use exploratory data analysis to understand important factors or characteristics such as Average, Std Mean Deviations, and also check for missing or null values and outliers. Exploratory data analysis is the process of looking at available data sets to identify patterns and anomalies, test hypotheses, and validate assumptions using statistical means. Using Python in exploratory data analysis processes and visual comparisons between variables is easy to understand and insightful.

In EDA, I took out **Univariate, analysis, Bivariate Analysis and Multivariate Analysis**. Also I have seen relations among variables to bring insightful information from the dataset.

Throughout the analysis, I tried to answer questions that help us understand the factors determining the data trends.

1. **Univariate Analysis:**

**a)** Distribution plot of numerical Variables:



INFERENCE:

Here it is clearly visible that some of the columns are not normally distributed. These are: ['Rented Bike Count', 'Wind speed (m/s)', 'Visibility (10m)','Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)']. These columns have skewed distribution.

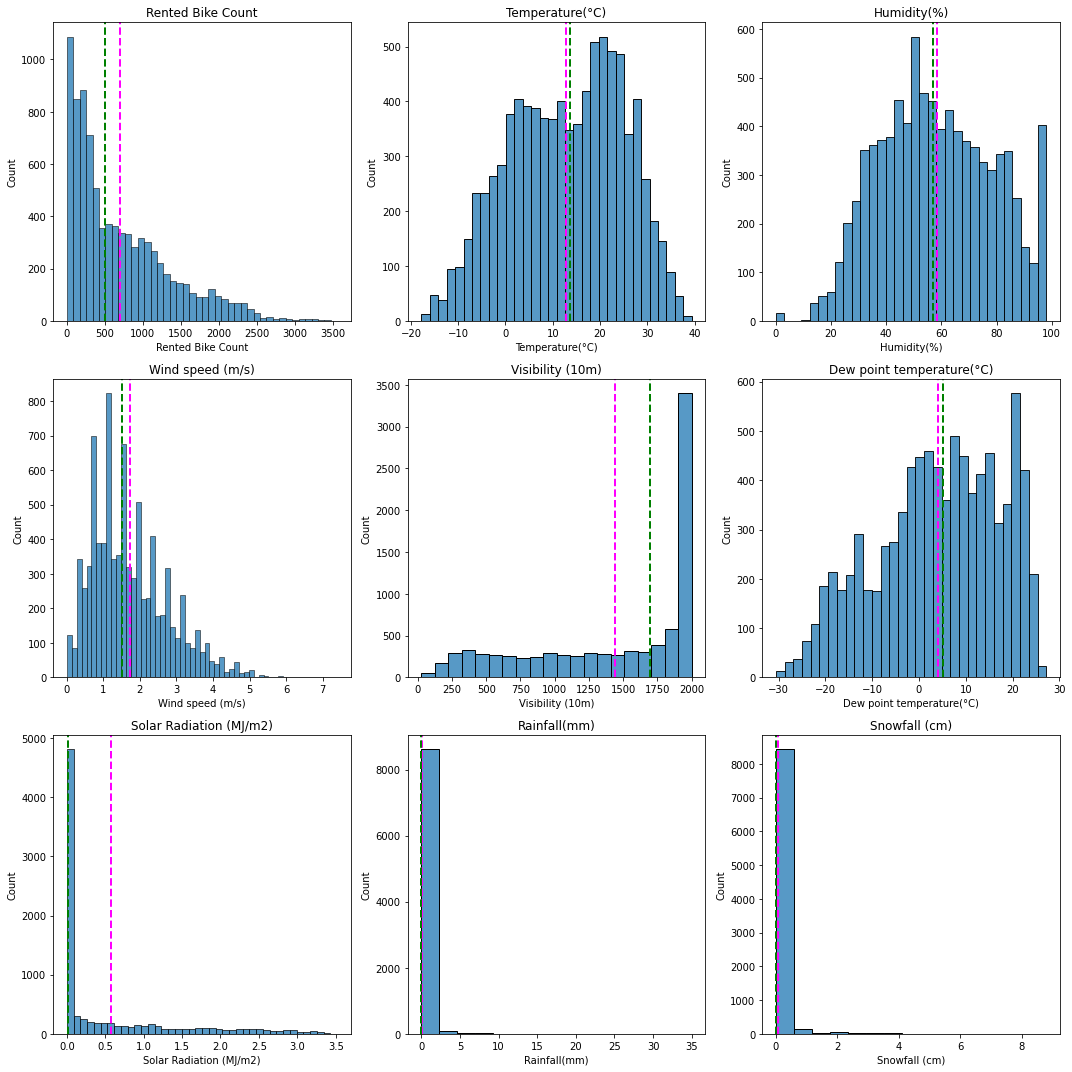
Approximately normally distributed: ***Temperature, Dew point temperature & Humidity***

Right skewed distributed : ***Wind speed, Solar Radiation, Rainfall, Snowfall***

Left skewed: ***Visibility***

Let's check if mean and median are on same axis or not.

b) Histogram of numerical variables



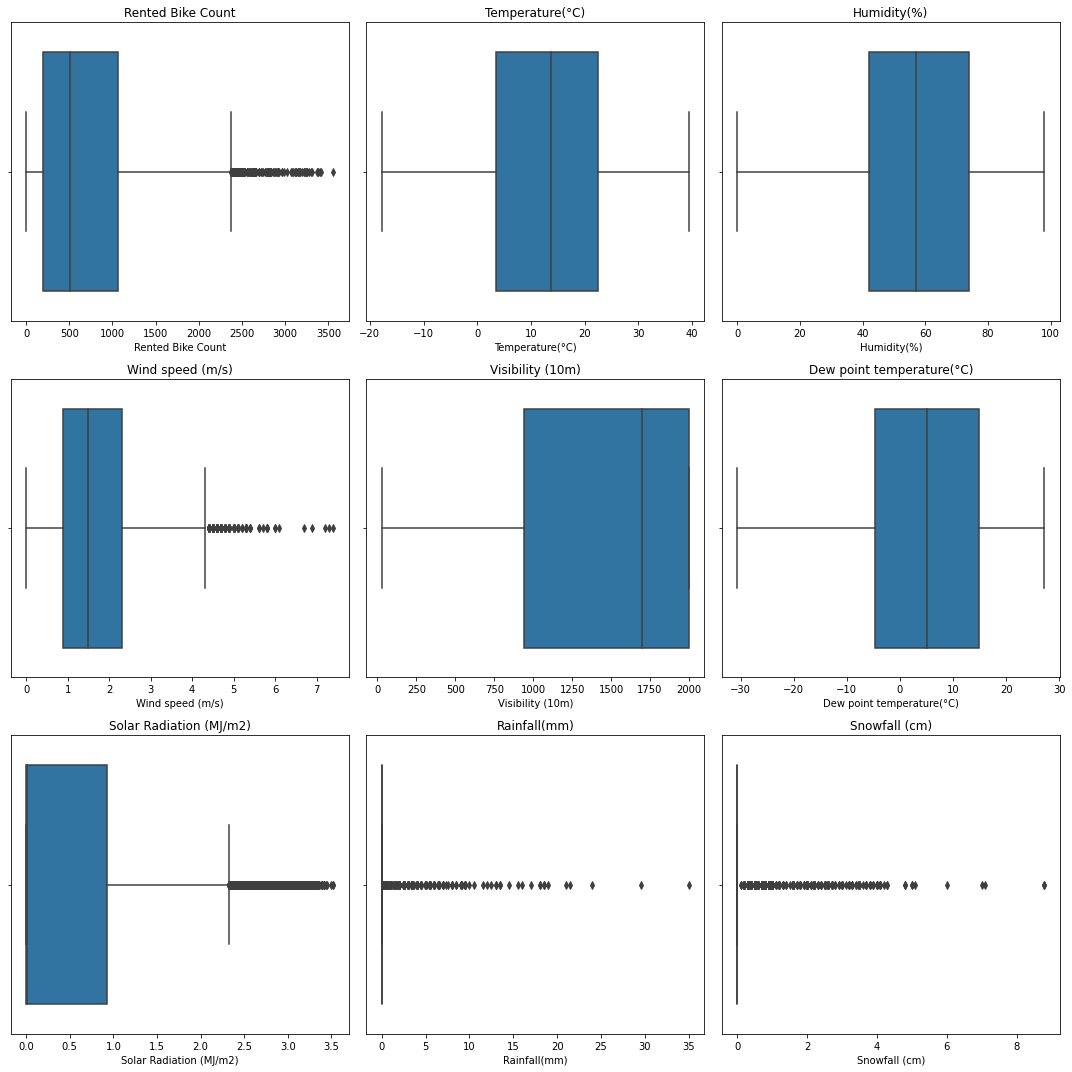
INFERENCE:

Here ‘Rented bike count’, ‘wind speed’, ‘visibility’ and ‘solar radiation’ columns do not have mean and median on same axes.

Rented bike count which is dependent variable is positively skewed. 'Wind speed (m/s)','Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)' which is independent variables are also positively skewed and 'Visibility (10m)' is also IDV which is negatively skewed.

c) Boxplot

To check for outliers in the dataset.



INFERENCE

Outliers present in some columns. These are:

[ 'Rented Bike Count' , 'Wind speed (m/s)' , 'Solar Radiation (MJ/m2)' , 'Rainfall(mm)' , 'Snowfall (cm)']

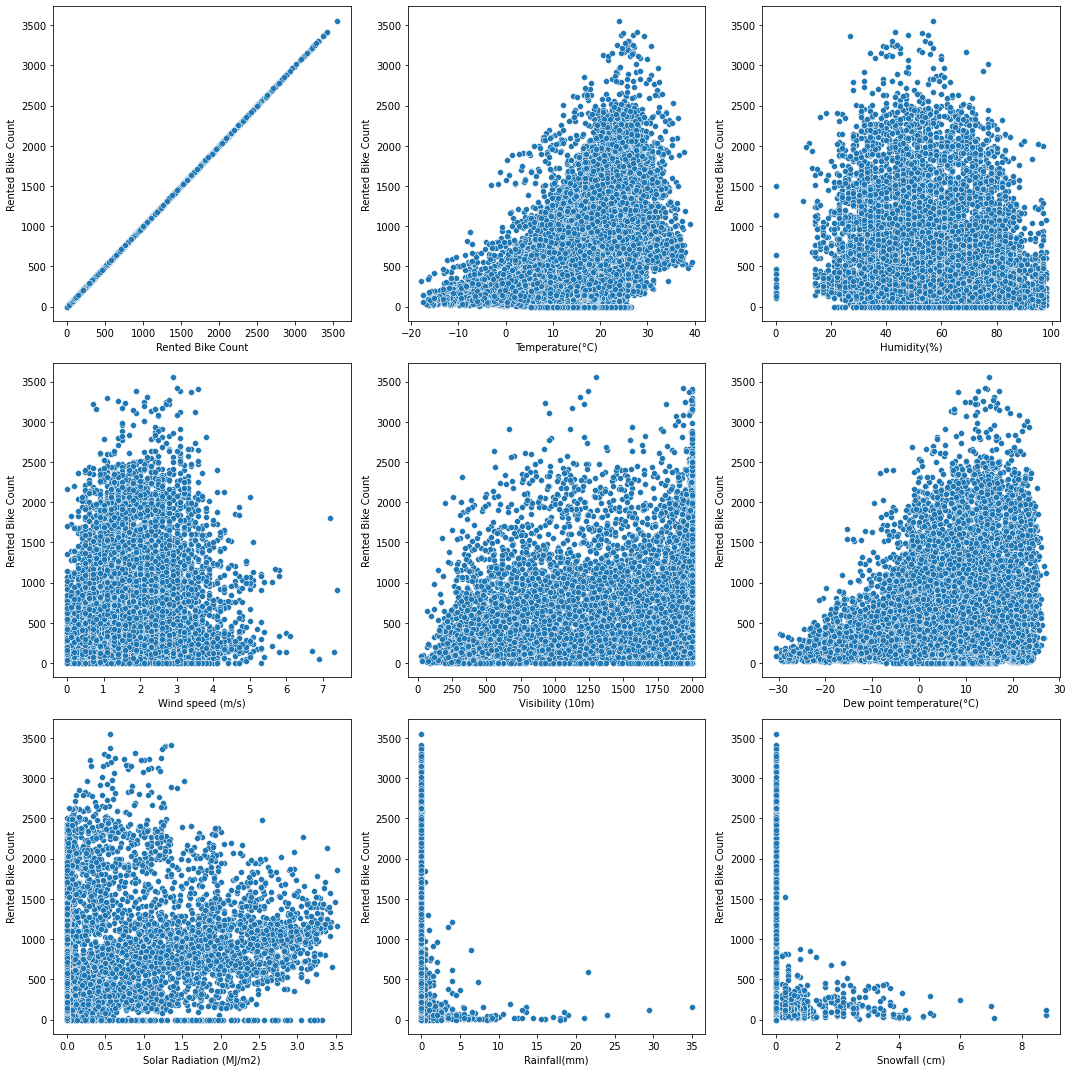
* The data points for Rented bike Count, Wind speed, Solar radiations, Rainfall and Snowfall lies at the left.
* Visibility is distributed towards right.
* Temperature, Humidity and dew point temperature are normally distributed at the center.

Outliers must be carefully handled during modelling.

**2. Bivariate Analysis:**

d) Scatter plot

Scatter plots are plotted to see the relationship between ‘Rented Bike count’ and other independent numerical features.

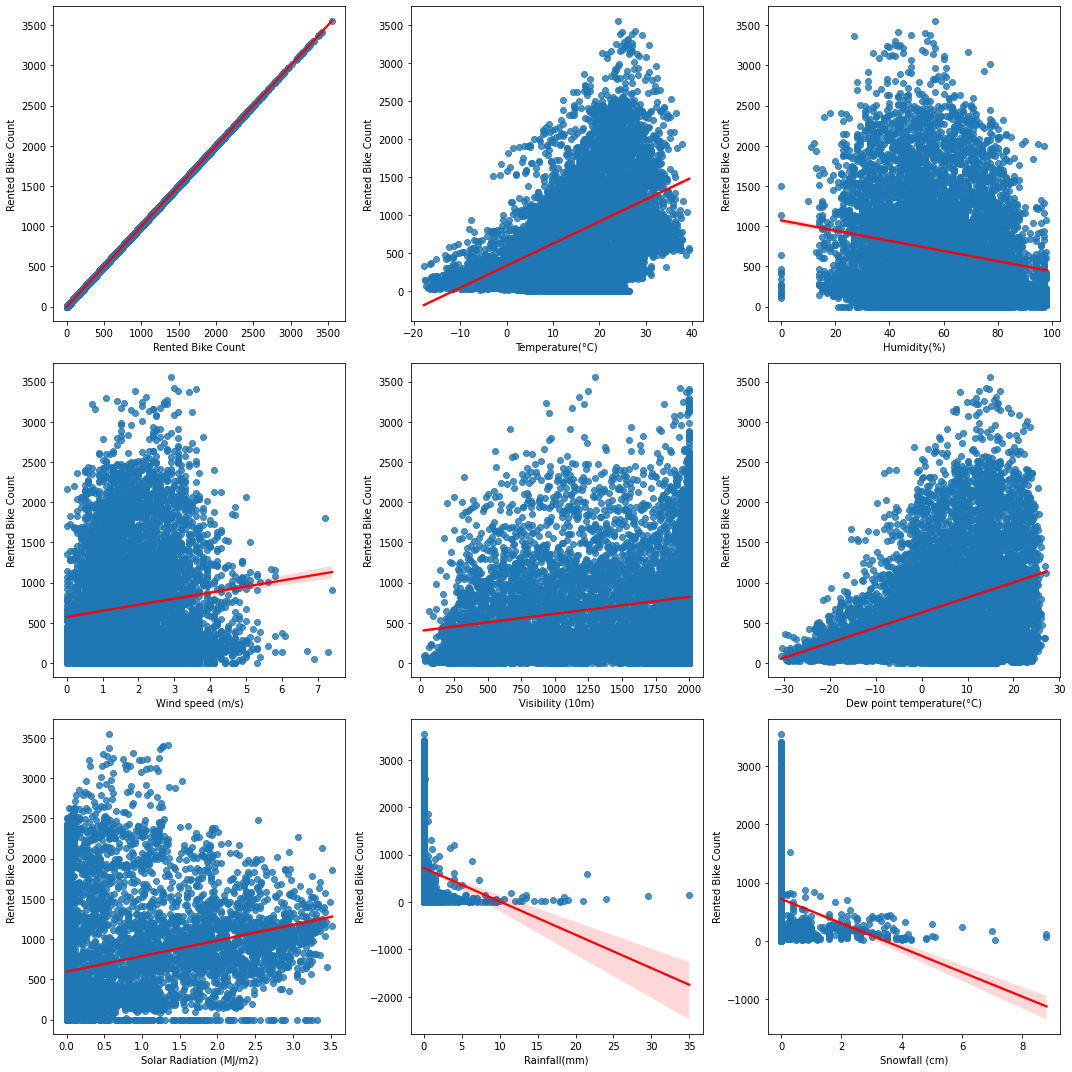


INFERENCE:

* Any linear relationship is not seen among dependent and independent variables.
* Bike count increases with increase in temperature and dew point temperature.
* Highest bike count when there is no rainfall and snowfall.
* Greater the wind speed, Lower is rented bike count.
* Solar radiations doesn’t seems to have much impact on bike count.
* When there is good visibility, there is rise in bike count.
* Lower humidity have almost no rented bikes.

e) Linear regression line

To see the linear relationship is positive or negative towards Dependent variable.

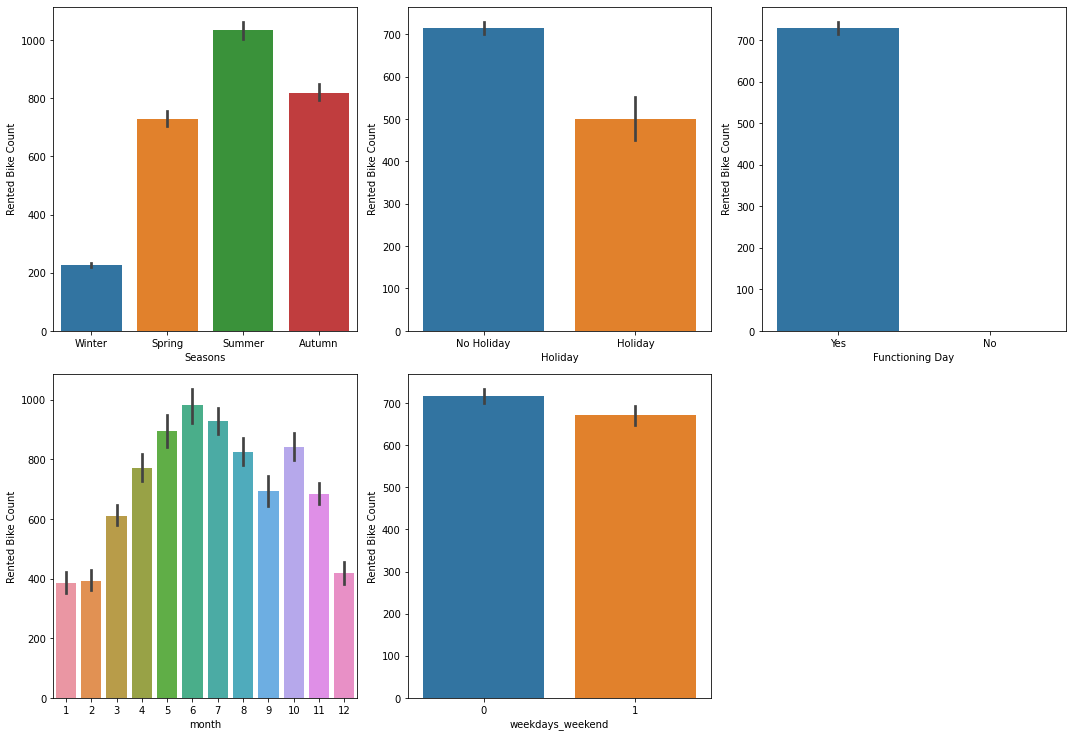


INFERENCE:

* Here Temperature, Wind Speed, Visibility, Dew point temperature, Solar radiations have slightly positive effect on dependent variable.
* And Humidity, Rainfall and Snowfall have negative impact on rented bike count.

f) Bar plot of Categorical columns

We have categorical columns : 'Seasons', 'Holiday', 'Functioning Day', 'month', 'weekdays\_weekend'.



INFERENCE

1.Seasons : Of all the four seasons Rented bikes are mostly used in summer season followed by Autumn season. Which means that people used to ride bike in pleasant weather.

2.Holiday : Count of rented bikes is more during non Holiday. Hence bikes are used for work or office purpose.

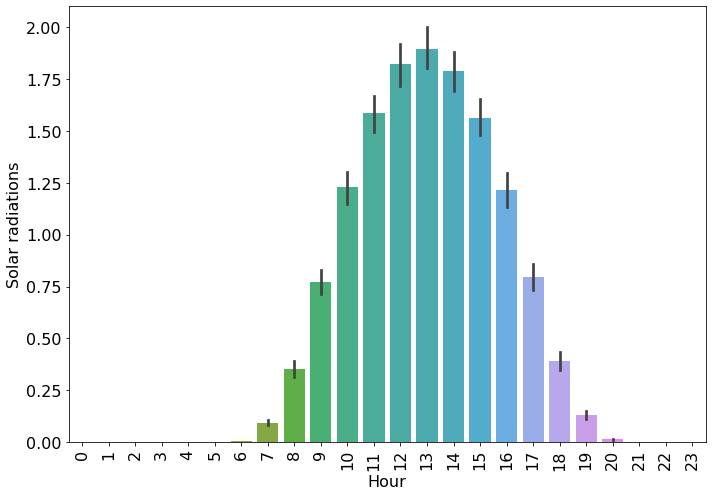
3.Functioning Day : Rented bikes are mostly used during functioning days. Hence used mostly for office or work purpose.

4.Month : Rented bikes are less used during December, January & February ,i.e., Winters. and mostly used in the month of May, June, July & October.

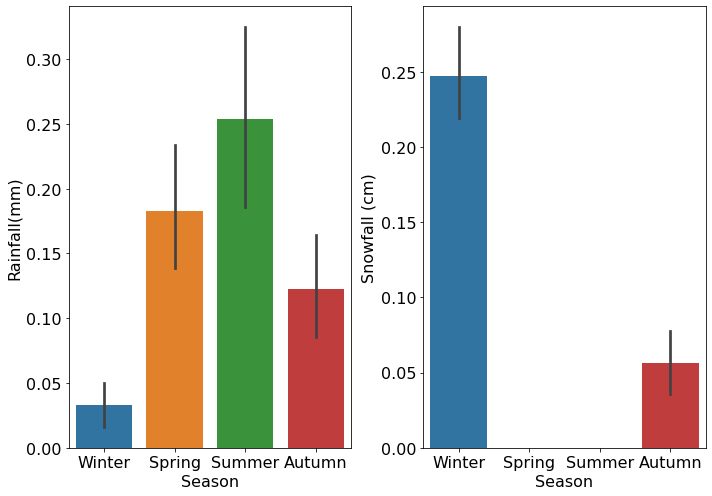
5.Weekends : During weekends bikes are comparatively less used than weekdays.

g) Solar radiations and hour:

We know that Sun shines during the day, and solar radiations are peak in the day-time. Let’s check if this data is correct or not.

My assumption is correct. And hence Solar radiations are not outliers. It is related with hour of the day.

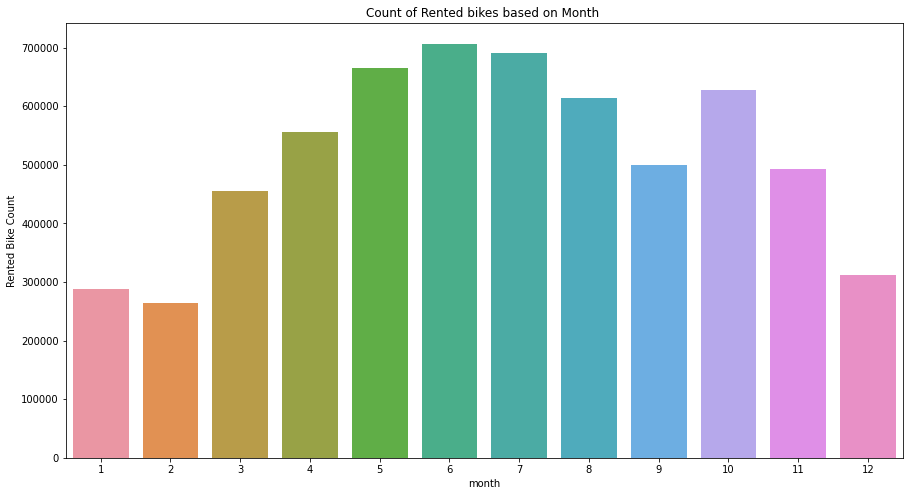
h) Does Rainfall & Snowfall depends upon Season?



INFERENCE: Clearly, Rainfall & Snowfall is dependent upon season. Because very little rainfall in winters is observed. And there is almost no snowfall in Spring & Summer.

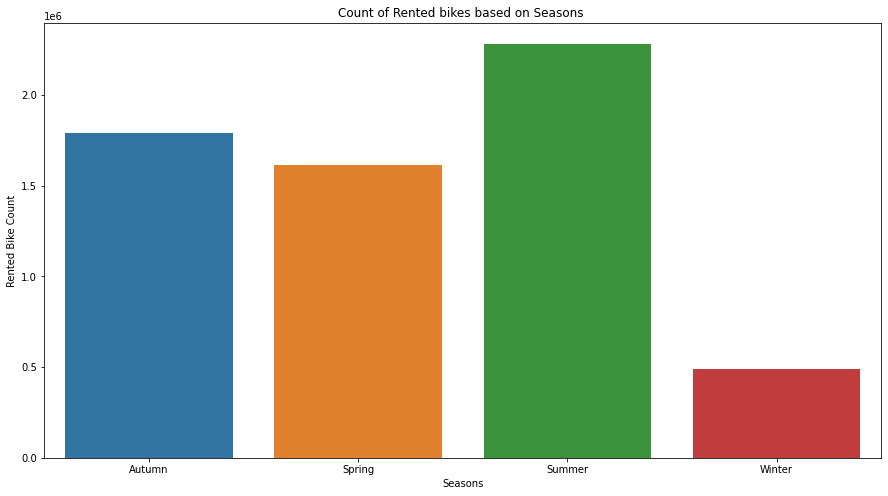
Hence these columns are not outliers.

i) Which month have the most demand for rented bike count?



The demand for rented bike count is highest in the month of May, June, July and lowest in the month of December, January, February.

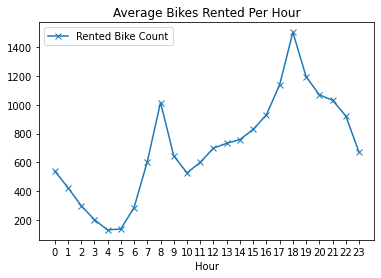
j) Demand of rented bikes for different seasons.



The demand for rented bike is high in summer season while the demand for rented bike is low in winter season.

It means that people like to have bike ride in convincing season.

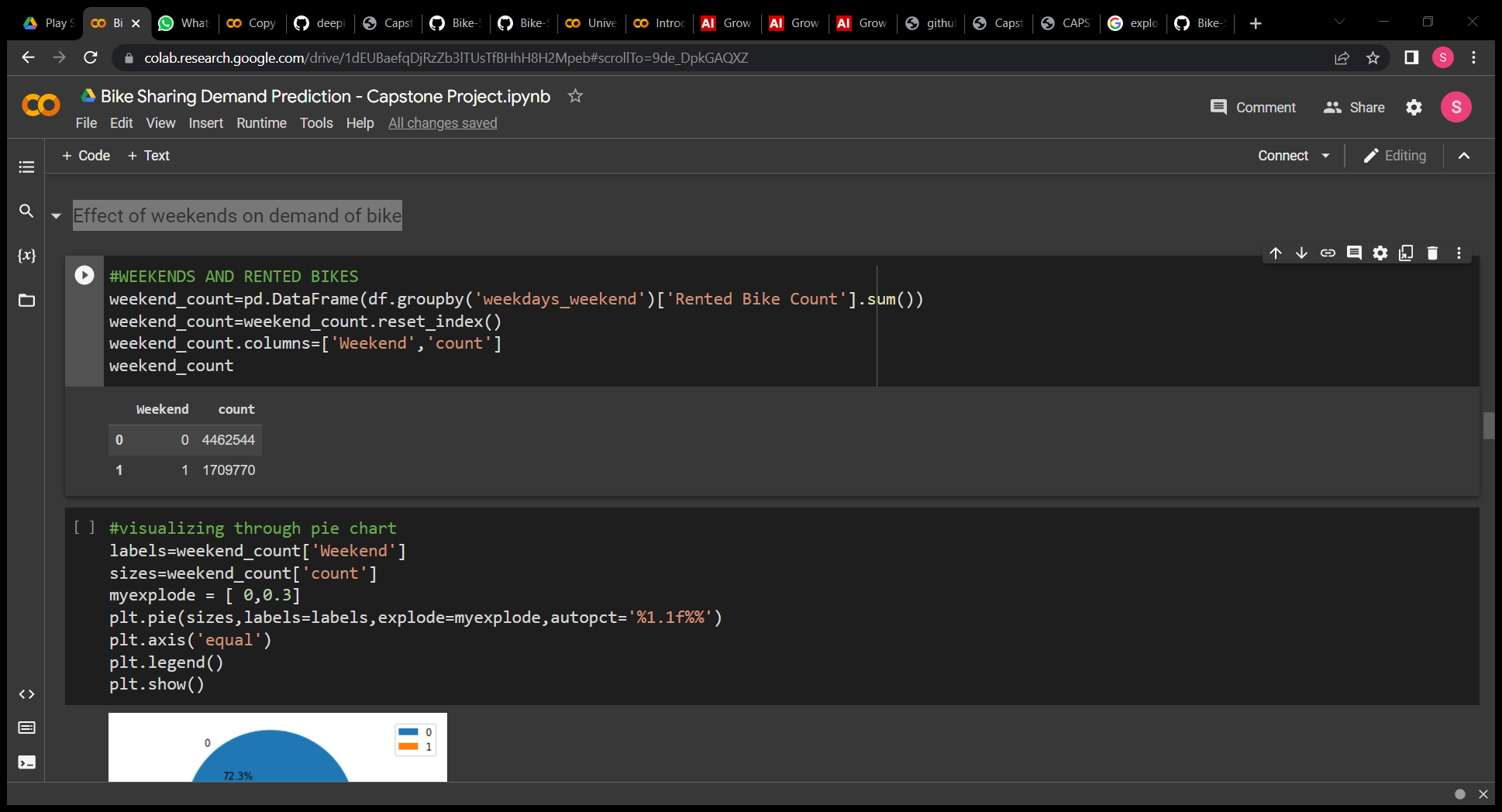
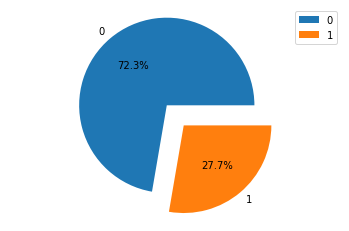
k) Average count of rented bike in each hour.



INFERENCE:

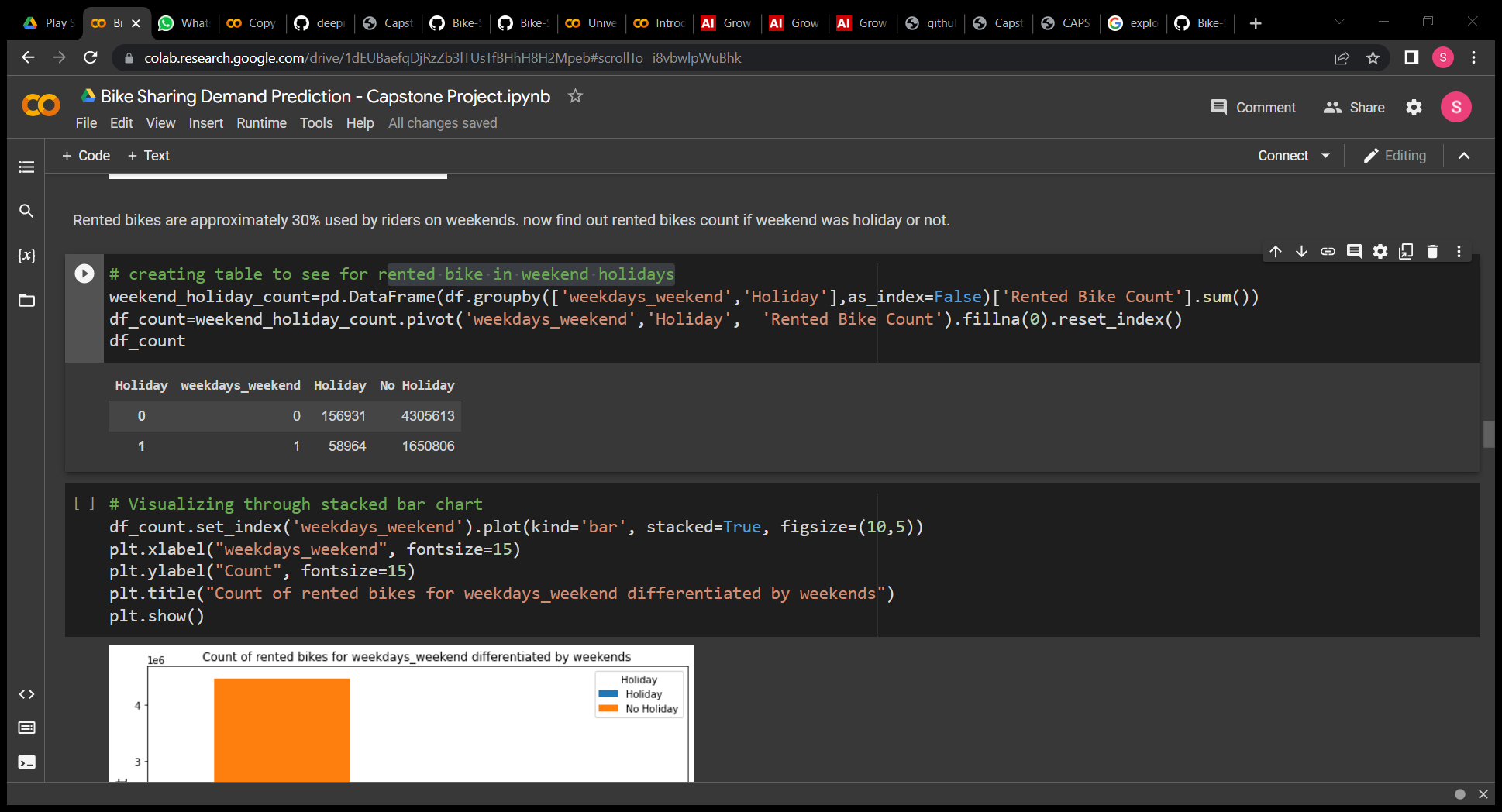
High rise of Rented Bikes from 8:00a.m to 9:00p.m means people prefer rented bike during rush hour.

l) Effect of weekends on demand of bike



Rented bikes are approximately 30% used by riders on weekends. now find out rented bikes count if weekend was holiday or not.

m) Rented bike in weekend holidays



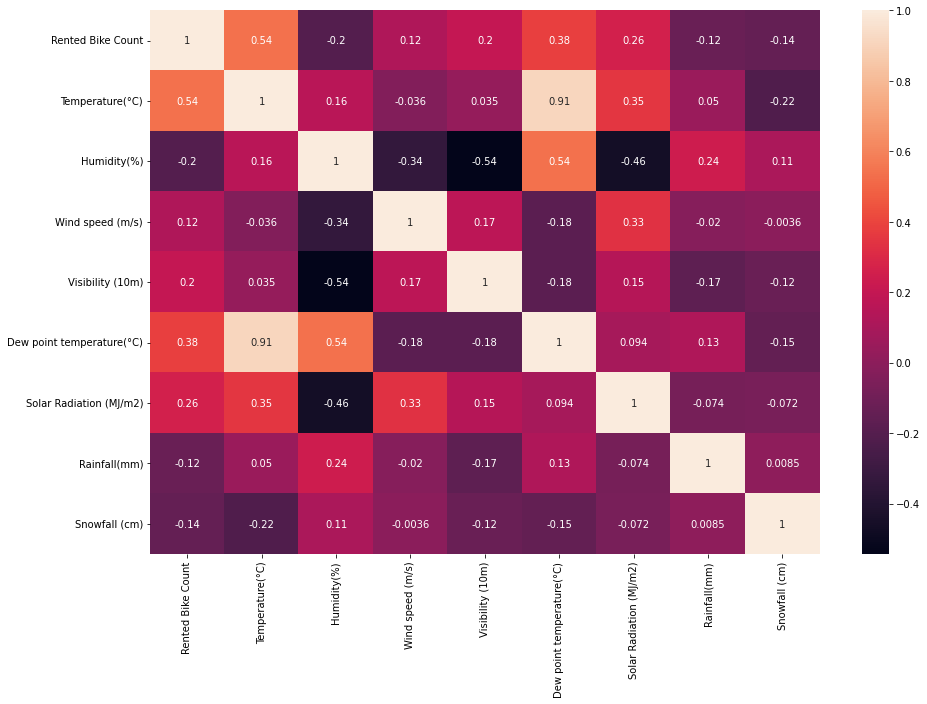


Rented bikes are mostly used in non-weekend. During holidays on weekends people generally do not use much rented bikes.

Which means that people generally use rented bike for travelling to office.

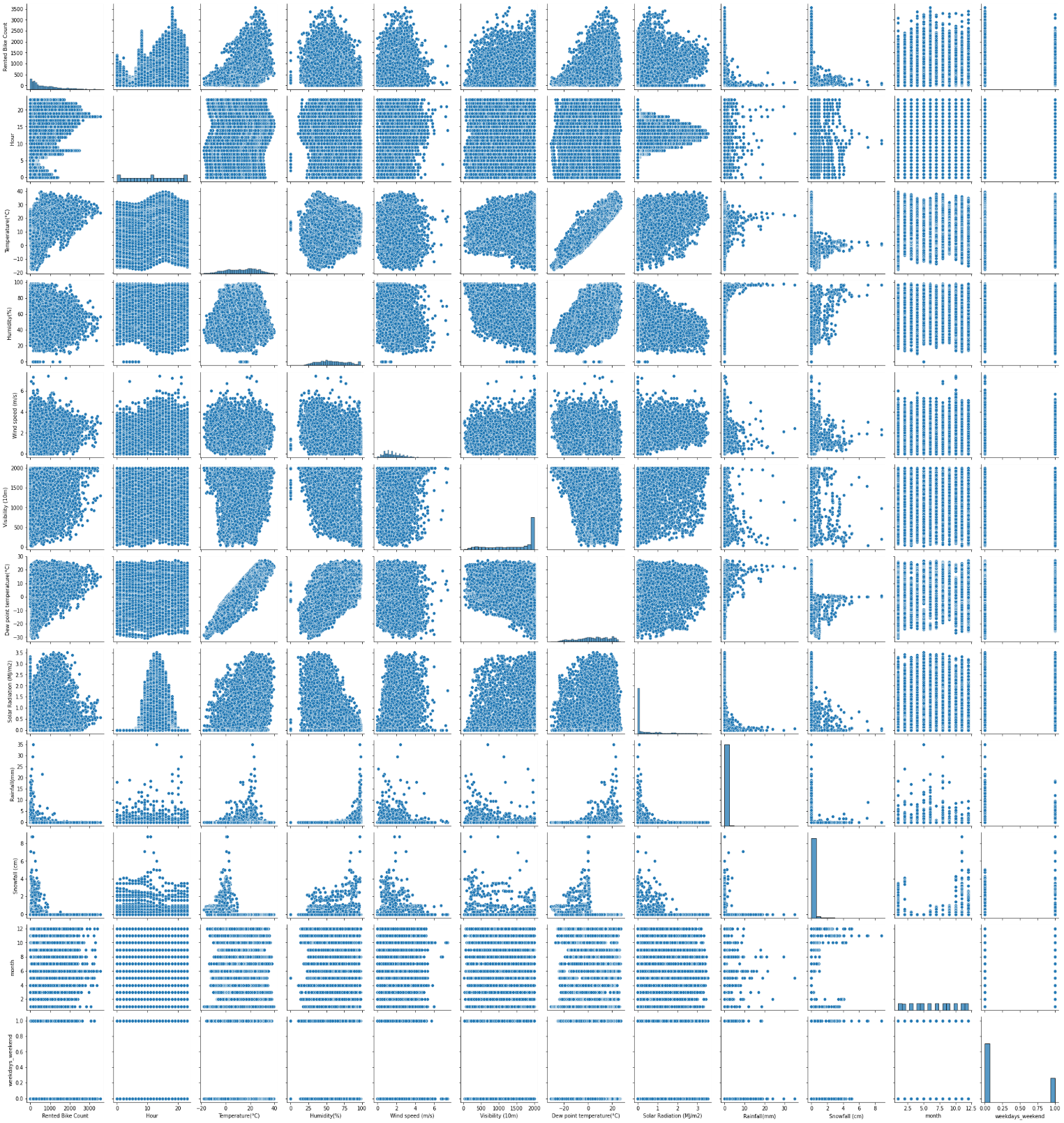
**3. Multivariate Analysis:**

n) Correlation



There is high correlation between Dew point temperature and Temperature, i.e., 0.91. The other independent variables are very low related.

o) Pairplot



Here we can see all the relations of dependent and independent variables.

~**Conclusion (From EDA)~**

* Some of the distributions are positively skewed while some are negatively skewed.

Right skewed distributed : ***Wind speed, Solar Radiation, Rainfall, Snowfall***

Left skewed: ***Visibility***

* Mean and median are not found on same axis of these columns.
* Outliers are seen in 'Rented Bike Count' , 'Wind speed (m/s)' , 'Solar Radiation (MJ/m2)' , 'Rainfall(mm)' , 'Snowfall (cm)', But it is proved that , 'Solar Radiation (MJ/m2)' , 'Rainfall(mm)' , 'Snowfall (cm)' are not outliers.
* 'Temperature(°C)','Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)','Solar Radiation (MJ/m2)' have positive effect while , 'Humidity(%)', 'Rainfall(mm)', 'Snowfall (cm)' have negative effect on ‘rented bike count’.
* Bikes are mostly rented during summer season and least rented during Winter season.
* Bikes are mostly rented during non-holidays and off-weekends and functioning days.
* High rise of Rented Bikes from 8:00a.m to 9:00p.m means people prefer rented bike during rush hour
* There is high correlation between Dew point temperature and Temperature, i.e., 0.91, as seen from heatmap and correlation matrix.

Hence I tried to figure out some of the important analysis that must be carried out before modelling. We have seen certain relations among variables and now let’s see if there is any multicollinearity in the data.

* **Data Preparation for Evaluating Models**

**For linear models:**

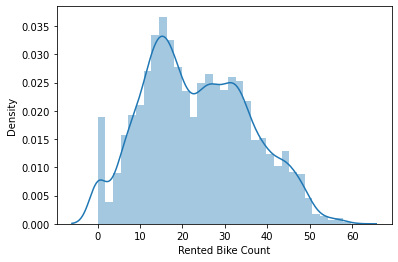
* Checking for Multi-collinearity:

Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity is a problem because it will make the statistical inferences less reliable.

* Using variance inflation factor and found high multicollinearity between ‘Temperature’ and ‘Dew point temperature’. So dropped ‘Dew point temperature’.
* Transformation:

Since dependent variable, Rented Bike Count, is positively skewed, it has some outliers. Outliers is a data point in the dataset that differs significantly from the other data or observation. The thing to remember that, not all outliers are the same. Some have a strong influence, some not at all. Some are valid and important data values. Some are simply errors or noise. But it needs to be transformed. So we can apply some transformation to make it gaussian distribution, for example, log transformation, square root transformation, cube root transformation etc.

* So I applied **Square root transformation** of Dependent variable to normalize it.



**Fig: After square root transformation**

* Standard Scale transformation of independent variables due to presence of outliers.

NOTE:

(*i) These transformations are not applied on tree based models because tree based models are not affected by multicollinearity. Because these follow the non parametric approach. As the decision at each node of the tree is made based on the single feature; Mutlicollinearity doesn't affect in decision trees.*

*(ii) Decision trees are also not sensitive to outliers since the partitioning happens based on the proportion of samples within the split ranges and not on absolute values.*

**For linear and tree based models:**

One hot encoding:

One hot Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning

* One hot encoding of categorical columns ‘Seasons’, ‘Holiday’, ‘Functioning Day’.
* Splitting data:

Train – test split divides the data into two subsets, training and testing. The training set is used to fit the model, while testing set is used to make predictions and compare with the given values. So testing set is used to evaluate the fitted model.

* I Splitted data to train and test with test size of 0.25.
* **Modelling Approach**

We have dataset which contains:

* Non-linear relationship among variables.
* Outliers
* Has multicollinearity.
* Categorical variables.

Hence models must be chosen such that:

* It gives high adjusted R2 score for test dataset.
* Root mean square is minimized.
* Easy to explain and interpret.
* **Models Implementation**

1. **LINEAR MODELS**

1. Multiple Linear Regression

2. Ridge Regression

3. Lasso Regression

4. Elastic net Regression

5. Polynomial Regression

1. **TREE BASED MODELS**

1. Decision Tree Regressor

2. Random Forest Regressor

3. Extreme Gradient Boosting

Since there was no linear dependency among dependent and independent variables So I have used Tree based models but also fitted linear models to see the performance and check for errors.

**FITTING DIFFERENT MODELS ALGORITHMS**

**A. LINEAR MODELS**

**1. Multiple Linear Regression:** Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. LR makes prediction for continuous as well as numeric variables

The model for linear regression is:

yi=β0+β1Xi1+β2Xi2+⋯+βpXip+εi, i=1,⋯,n,

where yi=Dependent variable

Xij = Independent variables

*βk* is the *k*th coefficient, where *β*0 is the constant term/intercept in the model.

Result obtained:

coefficients: [ 3.37670425 5.86152375 -3.27521585 0.08411524 0.12166761 -0.75073829

-1.70356276 0.02804395 0.30036678 -0.34981986 -1.07585807 -1.25423192

-3.22809512 0.64132045 4.92075262]

intercept: 23.534127039443128

-----------------------------------------------------------------------------------------------------------------

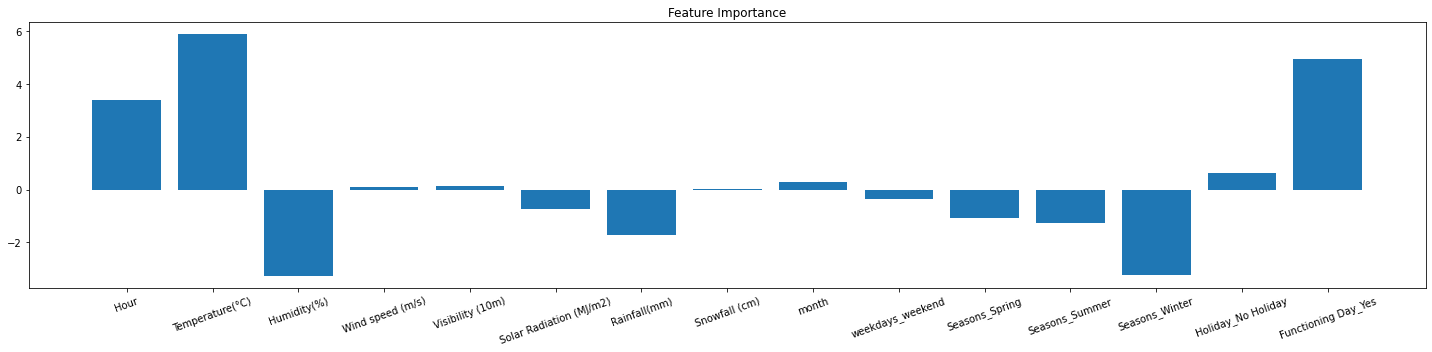
MSE: 173906.40891493318

RMSE : 417.02087347629634

R2 : 0.5830679795134978

MAE : 280.15776610071026

adj\_r2 : 0.5801912636407758

Train adj\_r2 : 0.5814971215796222 ****

**Figure: Feature Importance of linear regression**

**Inference:** Temperature and functioning day are taking the bike count towards positive while Humidity and winter season is taking it down.

These are important features in determining the prediction of rented bike count..

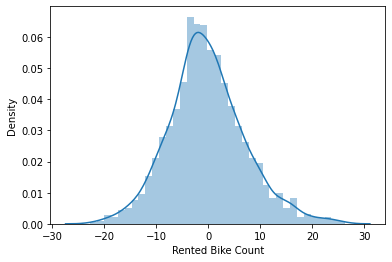
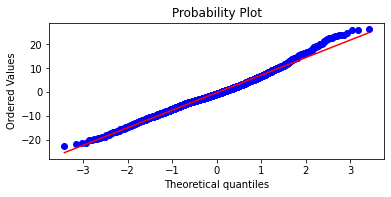
**Checking for assumptions:**

i) No multicollinearity

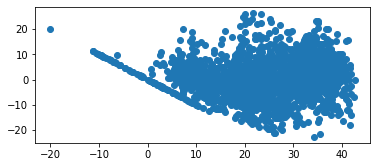
ii) Dependent variable is normally distributed

iii) Residuals are normally distributed.

From Distplot of residual and probability plot we get know that residuals are normally distributed.

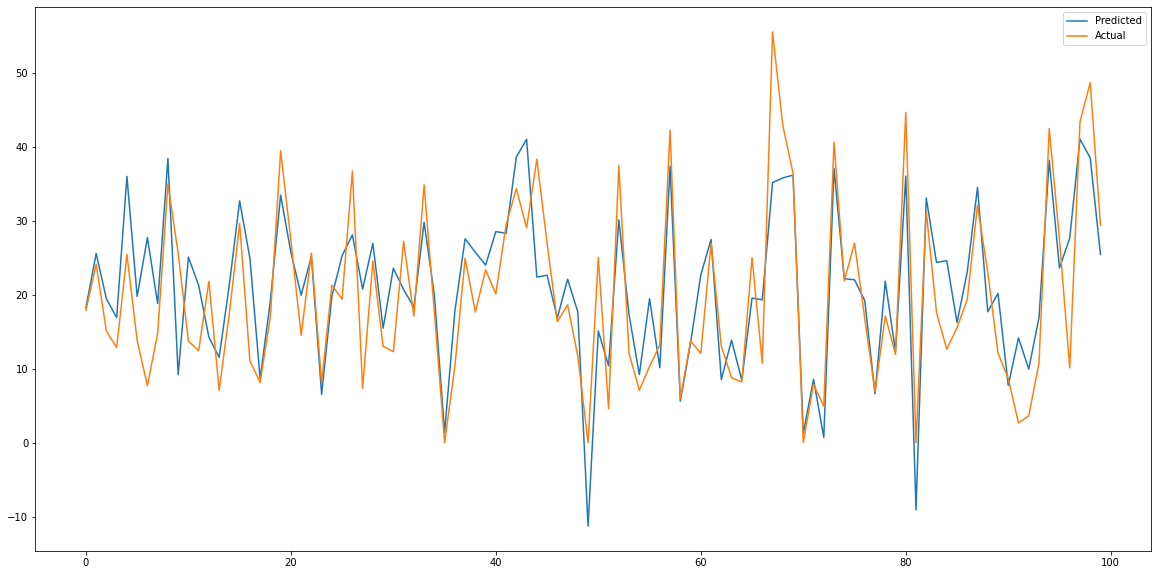
iv)Homoscedacity



Since we cannot see any bell shape or fan shape curve. Hence there is no heteroscedacidity.

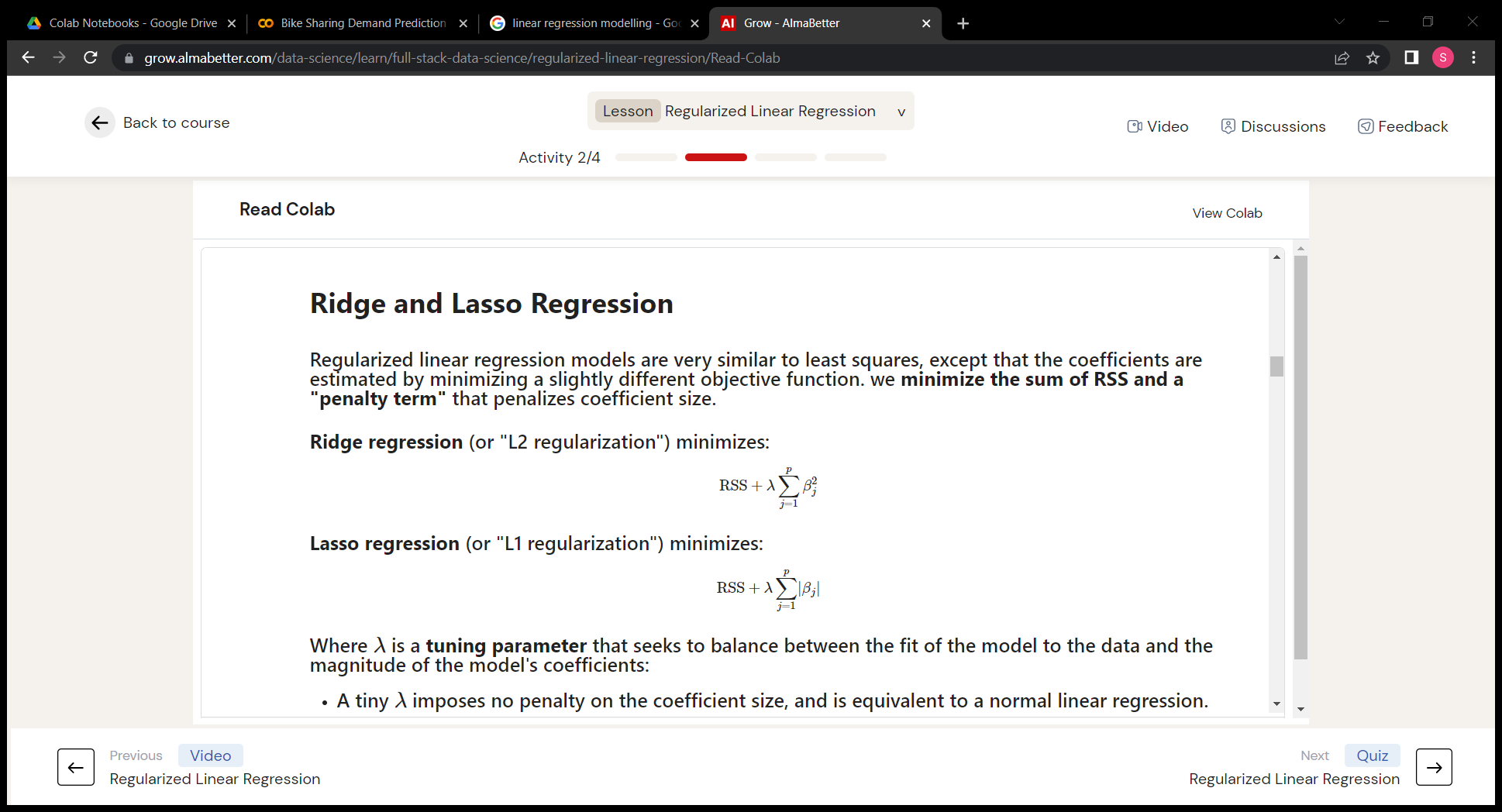
**But linear dependency is not satisfied.**

**Actual v/s predicted graph:**



**2. Ridge Regression**: It is regularized linear regression in which the Residual Sum of squares is minimized along with penalty term and then coefficient term is estimated. To overcome overfitting, I used ridge regression with hyperparameter tuning taking different parameters for alpha:

**Ridge regression** (or "L2 regularization") minimizes:



Where λ is a tuning parameter that seeks to balance between the fit of the model to the data and the magnitude of the model's coefficients.

Results obtained:

coefficients: [ 3.37717887 5.84287951 -3.26640503 0.08326942 0.12574371 -0.74207401

-1.70280053 0.02617297 0.30108058 -0.34970214 -1.07087821 -1.24227473

-3.228962 0.64056617 4.9151317 ]

intercept: 23.534127039443128

----------------------------------------------------------------------------------------------------------------

MSE: 173975.63167416563

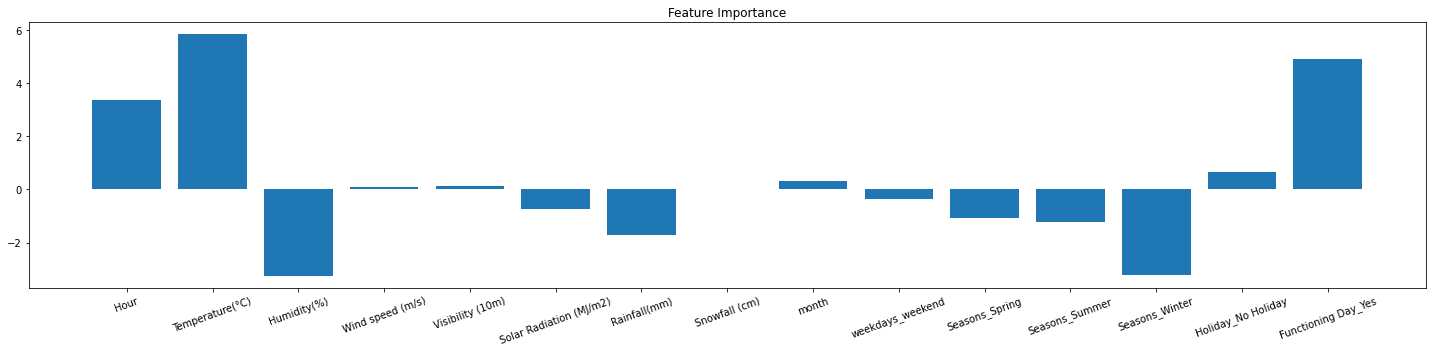
RMSE : 417.1038619746473

R2 : 0.582902021369399

MAE : 280.19439167375475

adj\_r2 : 0.5800241604312854

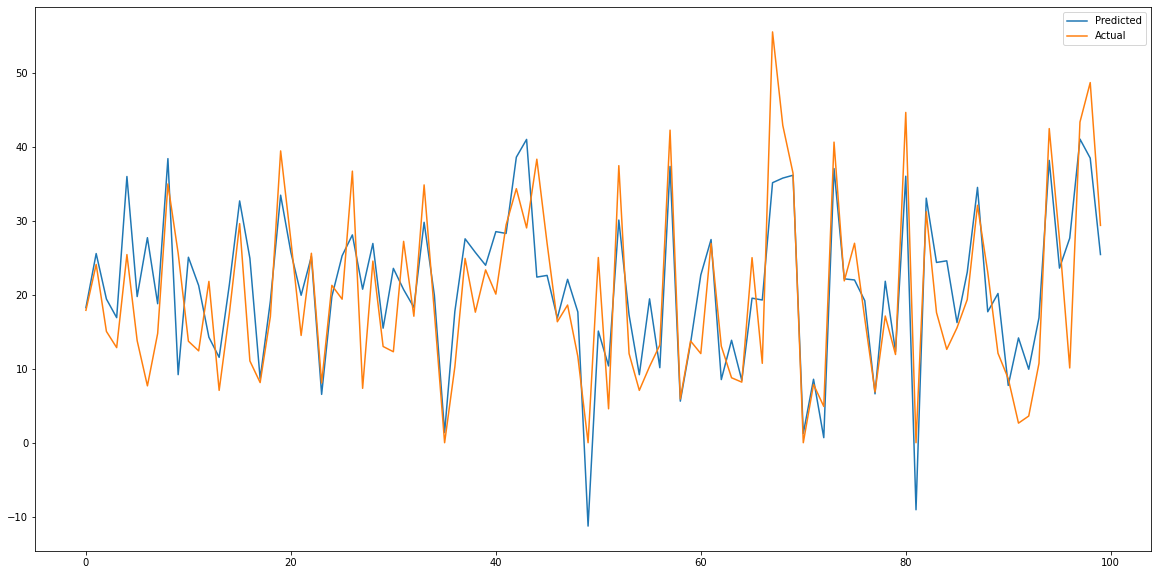
Train adj\_r2 : 0.5813401580195601



**Figure: Feature Importance of ridge regression**

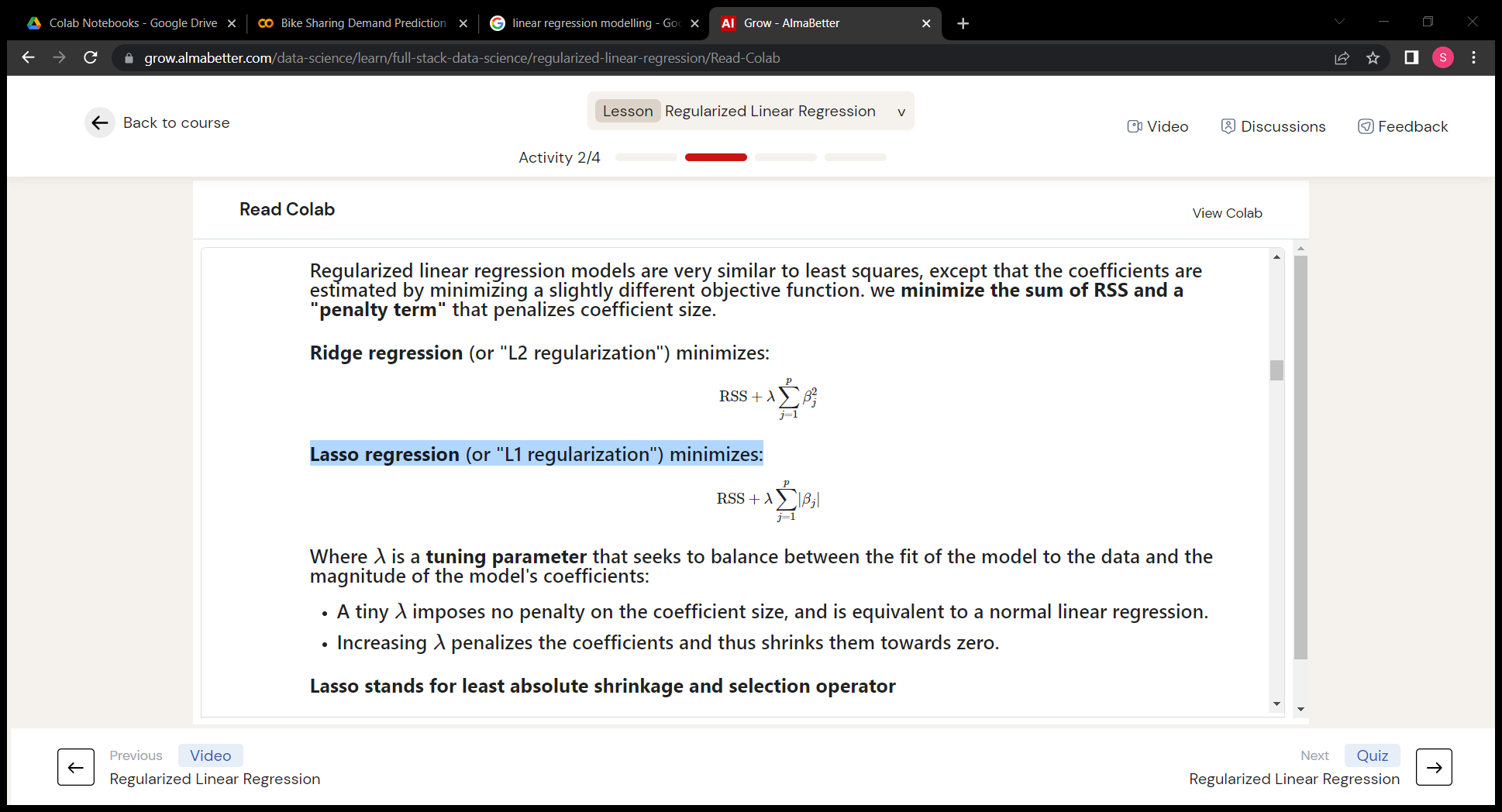
**Inference:**Temperature and functioning day are taking the bike count towards positive while Humidity and winter season is taking it down.These are important features in determining the prediction of rented bike count..

**Actual v/s predicted graph**



**3. Lasso Regression:** It is regularized linear regression in which the Residual Sum of squares is minimized along with penalty term and then coefficient term is estimated. To overcome overfitting, I used ridge regression with hyperparameter tuning taking different parameters for alpha:

**Lasso regression** (or "L1 regularization") minimizes:



Where λ is a tuning parameter that seeks to balance between the fit of the model to the data and the magnitude of the model's coefficients.

Results obtained:

coefficients: [ 3.37702794 5.85531884 -3.27222612 0.08227259 0.12223209 -0.7464786

-1.70280341 0.02641101 0.30025817 -0.34887271 -1.07269735 -1.24876798

-3.22665601 0.64017414 4.91881517]

intercept: 23.534127039443128

----------------------------------------------------------------------------------------------------------------

MSE: 173937.32639803702

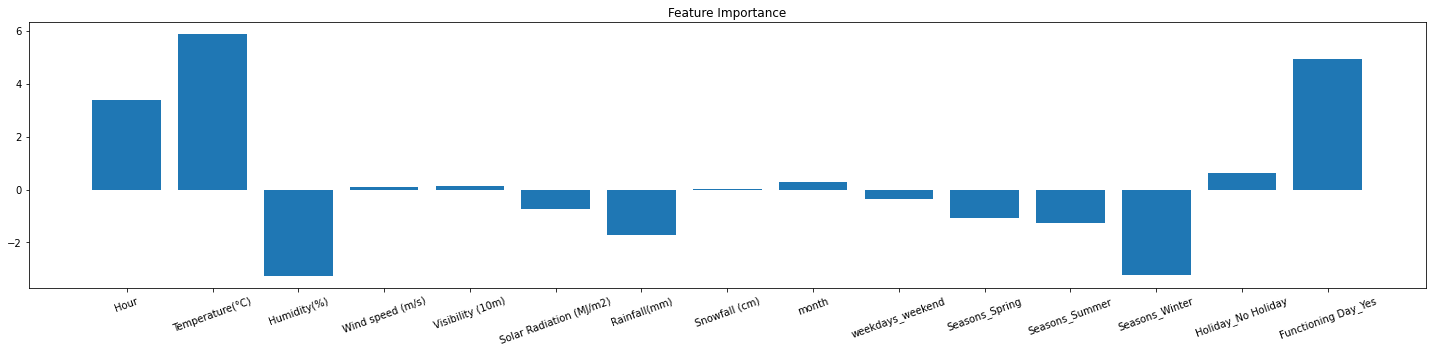
RMSE : 417.0579412959751

R2 : 0.5829938563757753

MAE : 280.17363848818394

adj\_r2 : 0.5801166290738602

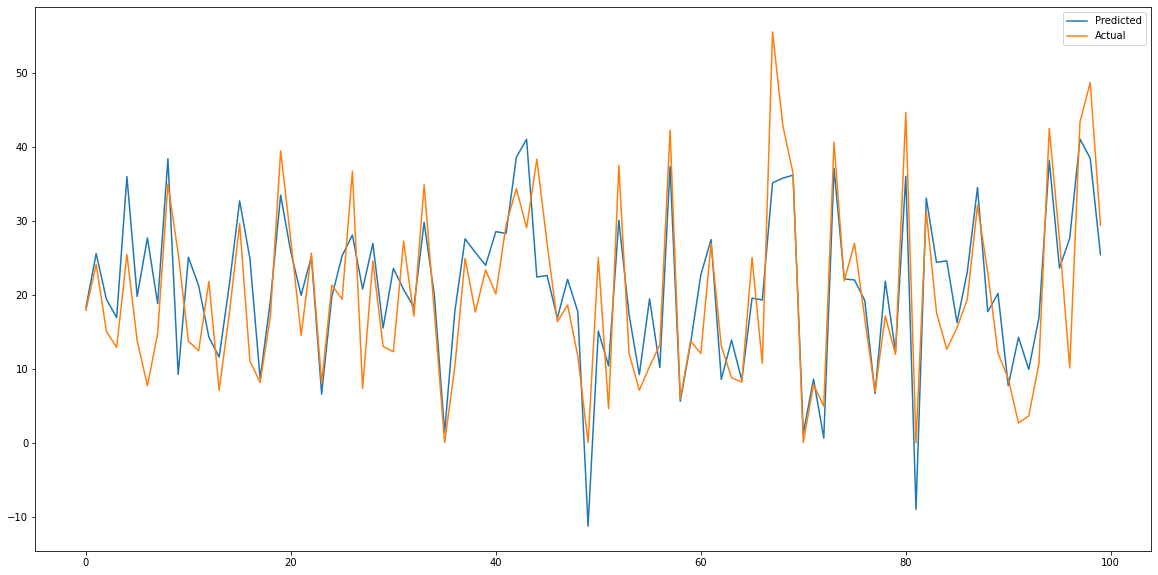
Train adj\_r2 : 0.5814224402440751



**Figure: Feature Importance of lasso regression**

**Inference:** Temperature and functioning day are taking the bike count towards positive while Humidity and winter season is taking it down. These are important features in determining the prediction of rented bike count.

**Actual v/s predicted graph**

****

### 4. Elastic Net Regression: It is a combination of Lasso regression and Ridge regression in a certain ratio. I applied elastic net regression with hyperparameter tuning on Alpha and L1 ratio using GridSearchCV.

Results obtained:

coefficients: [ 3.37729367 5.8388821 -3.26450579 0.08293289 0.12654024 -0.74007683

-1.70258503 0.02566995 0.30120025 -0.34958593 -1.06965628 -1.23956046

-3.2289672 0.64031501 4.91391449]

intercept: 23.534127039443128

----------------------------------------------------------------------------------------------------------------

MSE: 173991.44613014164

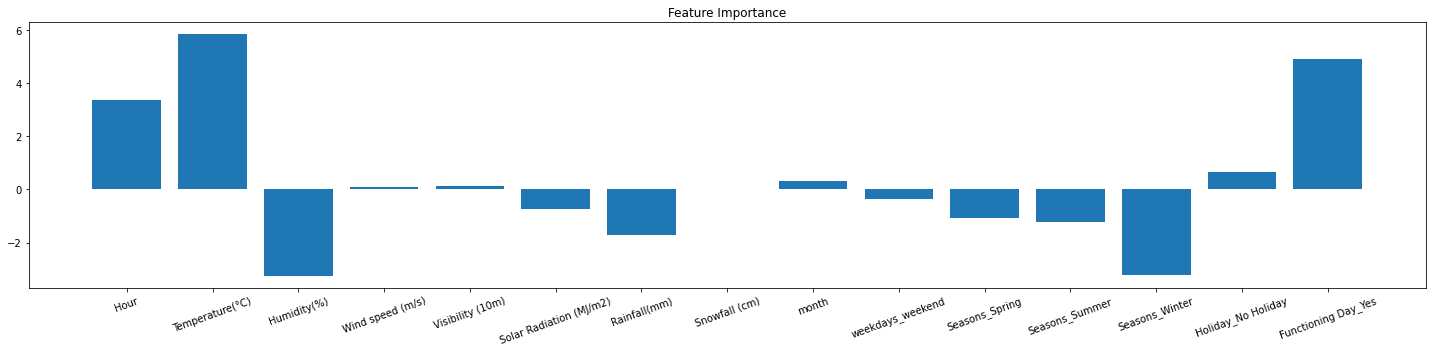
RMSE : 417.12281899956236

R2 : 0.582864106992786

MAE : 280.20264686626666

adj\_r2 : 0.5799859844559376

Train adj\_r2 : 0.5813038407360959



**Figure: Feature Importance of lasso regression**

**Inference:**Temperature and functioning day are taking the bike count towards positive while Humidity and winter season is taking it down.These are important features in determining the prediction of rented bike count.

**Actual v/s predicted graph**

****

**5. Polynomial Regression:** (With degree 2)

Models the relationship between a dependent(y) and independent variable(x) as 2nd degree polynomial.

Results obtained:

coefficients: [ 2.73946690e-10 1.36500861e-02 6.94794789e-02 6.90939539e-01

4.04483050e+00 -3.30332091e-03 3.91336344e+00 -2.63790886e+01

5.35247997e+00 -2.43223202e-01 -7.99136962e-01 -2.89788123e+00

6.00642977e+00 -3.10865894e+00 -5.39184224e+00 1.35141514e+01

-1.40542637e-03 1.90703489e-02 -7.98238815e-03 5.61487562e-04

1.70402736e-06 5.69753958e-02 -7.36595214e-02 3.78789467e-02

1.10186432e-02 -1.15174707e-01 -4.09525360e-02 5.26070522e-02

-1.07560990e-01 1.89934076e-01 6.16295026e-01 -4.25444151e-05

-3.72877312e-03 4.56565355e-02 5.35052812e-05 -2.21856919e-01

-1.30158102e-01 -2.39260249e-01 4.13913697e-03 9.21674027e-02

2.85114393e-01 -7.69534974e-01 -1.08211816e-01 -1.48849683e-01

4.33401989e-01 -4.42056257e-03 -4.12147683e-02 8.15669366e-07

3.04226856e-02 2.88840148e-01 2.16233130e-02 2.38091898e-03

2.74016685e-02 4.29784972e-02 3.55402140e-02 5.59191884e-02

5.91761522e-02 -2.45051826e-01 -9.99156444e-02 -6.05687317e-04

-8.52393351e-01 -4.93747495e-02 1.43083570e-01 -5.33147239e-02

6.60368683e-01 7.90098492e-02 5.96929873e-01 1.02693740e-01

3.25990883e-02 -4.77768105e-01 9.10581190e-08 7.42224745e-04

-1.79047730e-04 -7.92749552e-04 2.55020007e-05 8.88412039e-04

6.16164543e-04 -2.34136872e-03 6.14019121e-04 2.53853162e-03

4.72936192e-04 -6.37814602e-01 -4.19512372e+00 1.18453195e-01

-2.36032866e-02 1.05596401e+00 9.33919381e-01 1.60561908e+00

-2.36700534e+00 -2.93432819e-01 -7.85876453e-01 1.22140709e-01

-4.44970290e-01 -1.07457048e-01 -8.82718898e-02 -1.32236699e+00

-5.13217721e-01 5.44807689e-01 -1.86270256e-01 4.76051832e-01

-2.14939669e-01 -1.42666268e-01 -4.47301144e+00 0.00000000e+00

-6.21724894e-15 -6.49942277e+00 -4.96083766e+00 5.35247997e+00

-2.43971077e-02 -3.23178801e-02 1.04932445e-01 1.52365179e-02

3.30002303e-01 3.00729620e-01 -7.65782776e-02 -7.99136962e-01

-3.28335366e-01 -1.45106332e+00 1.89441746e+00 -1.44872416e+00

-2.03417279e+00 -2.89788123e+00 0.00000000e+00 0.00000000e+00

-1.68930133e+00 -2.86830551e+00 6.00642977e+00 0.00000000e+00

-1.05613357e+00 6.00642977e+00 -3.10865894e+00 -2.24943740e+00

-3.10865894e+00 -5.39184224e+00 5.98839714e+00 1.35141514e+01]

intercept: -17.120177398933418

----------------------------------------------------------------------------------------------------------------

MSE: 114327.60766682056

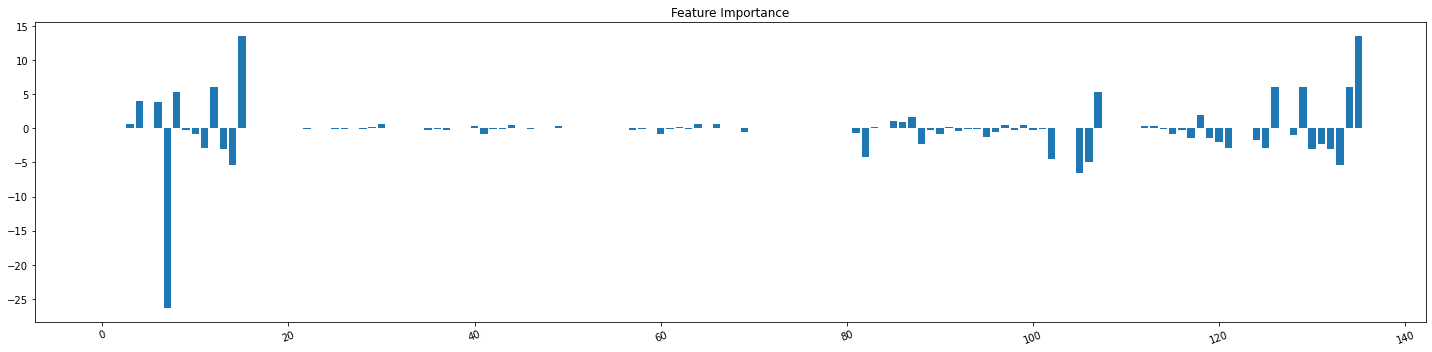
RMSE : 338.1236573604701

R2 : 0.7197445326584553

MAE : 223.5170029889027

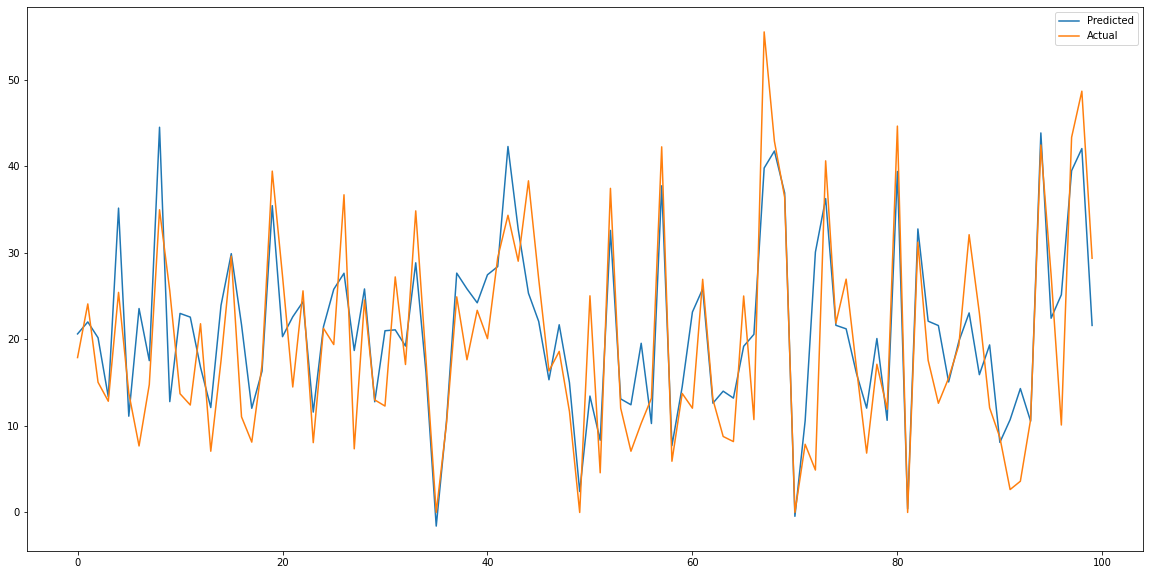
adj\_r2 : 0.7044435517036379

Train adj\_r2 : 0.7225983539806552



**Figure: Feature Importance of polynmial regression**

**Actual v/s predicted value graph**

****

**B. TREE BASED MODELS**

It is non-linear regression model. I used these models because:

* lack of linear dependency between dependent and independent variables.
* Multicollinearity among dependent variables which is very sensitive to linear regression.
* Outliers present in the dataset which in cause will affect the linear model.

**1. Decision Tree Regressor**: It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes. I used Decision Tree regressor with hyperparameter tuning on ‘splitter’, ’max\_depth’, ‘min\_samples\_split and the results are as followed:



**Figure: Feature Importance of decision tree regression**

Here Hour is the most important column which is followed by season Winter.

Evaluation matrics obtained:

Mean squared error: 68759.58724562923

Root mean squared error: 262.2204935652994

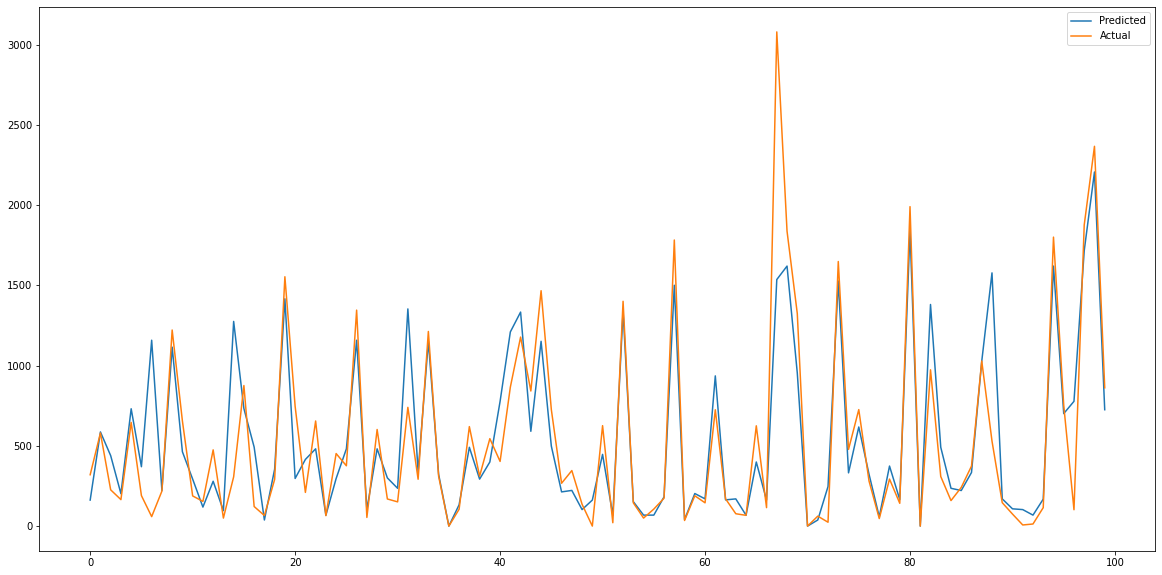
mean absolute error: 153.24269442632456

R2: 0.8351522878483381

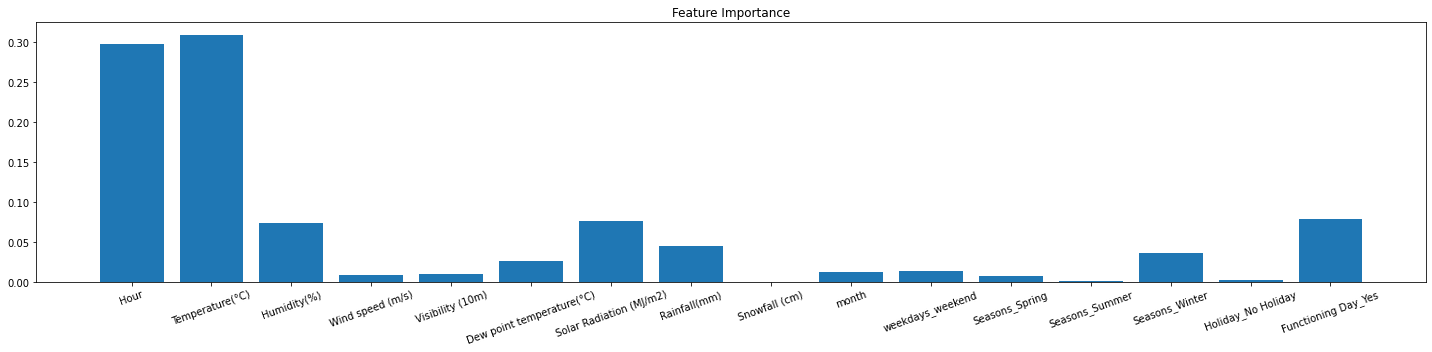
ADJUSTED R2 0.8339384988955416

Train\_adj\_R2: 0.9439298755168863

**Actual v/s predicted value graph**



**2. Random Forest Regressor**: This is an ensemble of Decision trees. The greater number of trees in the forest leads to higher r2 and prevents the problem of overfitting. I used random forest regressor with hyper parameter tuning on the parameters ‘n\_estimators’, ‘max\_depth’, ‘min\_sample\_split’, ‘max\_features’. The results obtained is:



**Figure: Feature Importance of random forest regression**

Hour and temperature are the most important columns in determining the rented bike count.

Evaluation matrics obtained:

Mean squared error: 45296.83512475205

Root mean squared error: 212.8305314675318

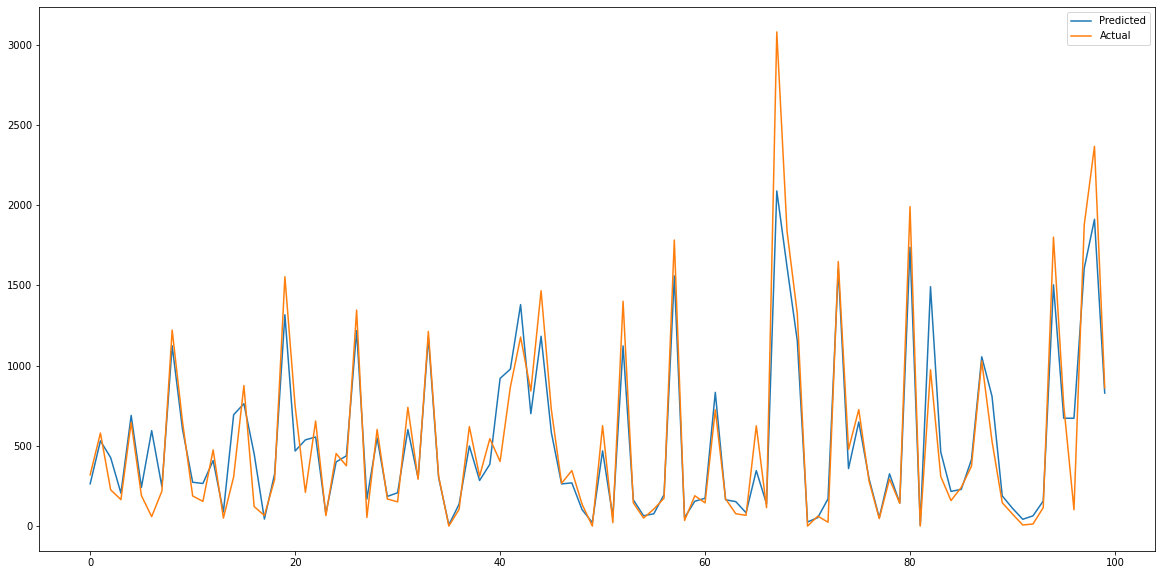
mean absolute error: 128.1672195366369

R2: 0.8914030764706042

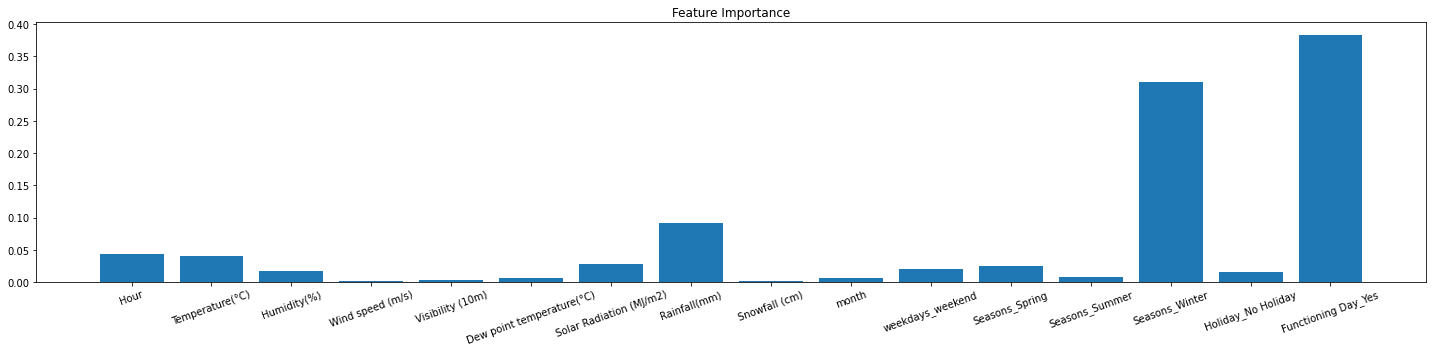
ADJUSTED R2 0.8906034672775668

Train\_adj\_R2: 0.9656327626816767

**Actual v/s predicted value graph**

****

**3. Extreme Gradient Boosting**: XGBoost is an implementation of Gradient Boosted decision trees. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. I used XGB regressor with hyper parameter tuning on the parameters ‘n\_estimators’, ‘max\_depth’, ‘learning\_rate’. The results obtained is:



**Figure: Feature Importance of XGB regression**

Evaluation matrics obtained:

Mean squared error: 35952.87945042624

Root mean squared error: 189.61244539962624

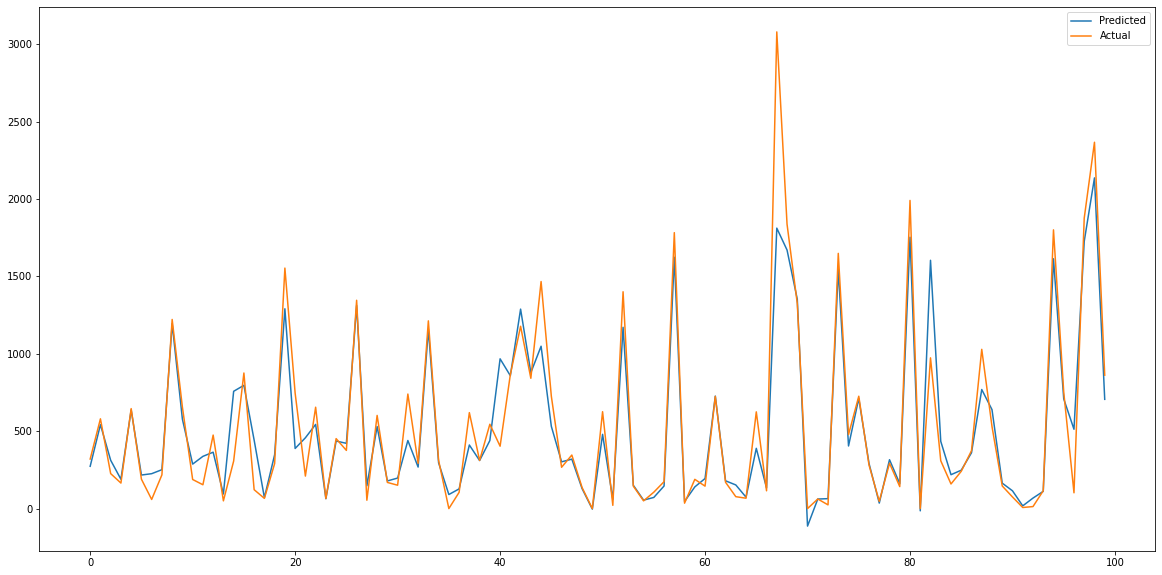
mean absolute error: 111.0026615084034

R2: 0.9138047483982824

ADJUSTED R2 0.9131700847877774

Train\_adj\_R2: 0.9994469904497246

**Actual v/s predicted value graph**



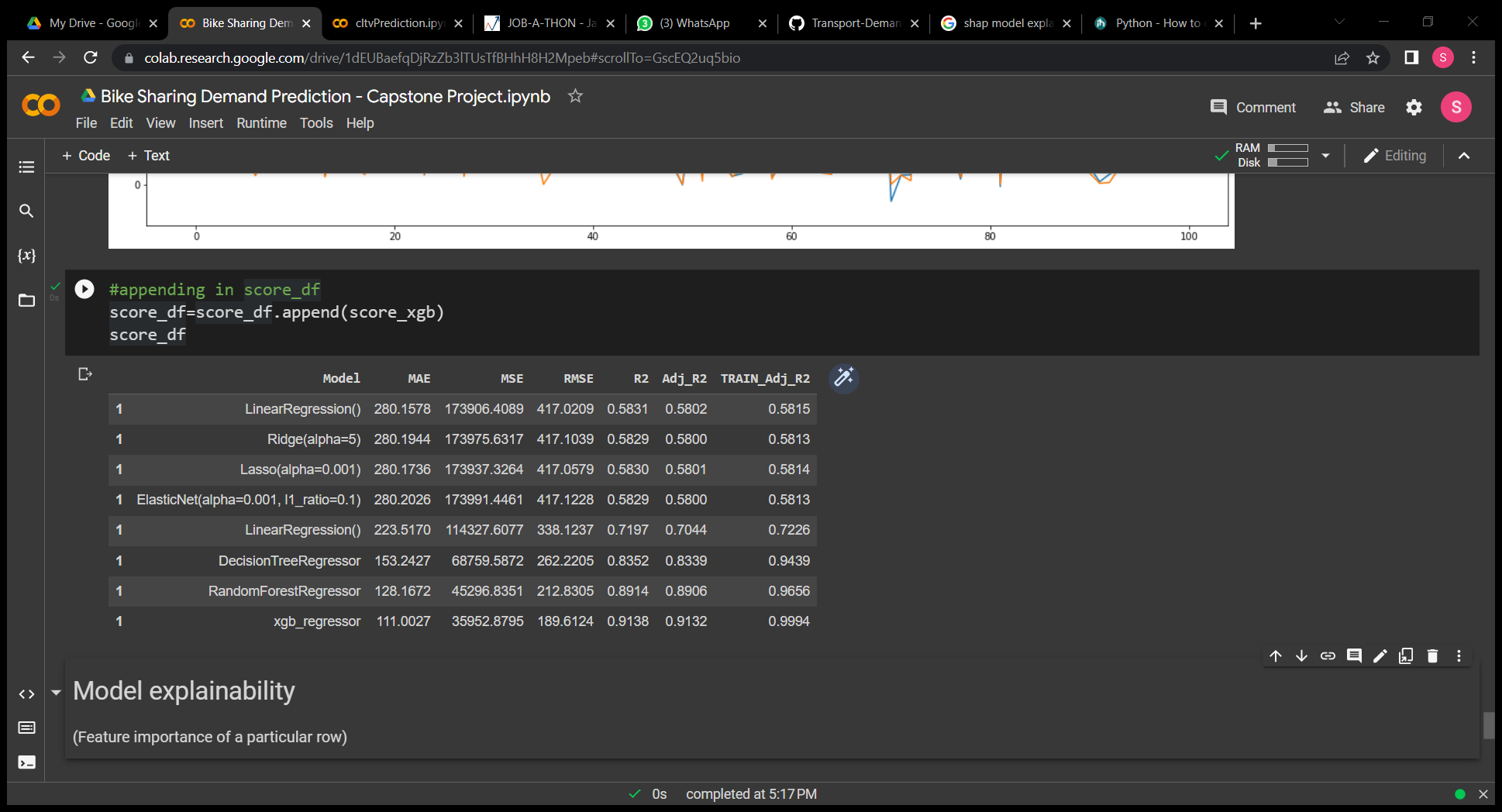
**INFERENCE:**

* Temperature and hour are the most important features observed in Decision Tree and Random Forest Regressor.
* XGB model shows that Season\_winter and Functioning day are most important variables in predicting the rented bike count.

What does different matrices imply?

1. **MAE (Mean Absolute Error):** It is the average of absolute differences of actual and predicted value.
2. **MSE (Mean Squared Error):** It is the average of square root of difference of actual and predicted value.
3. **RMSE** (**Root mean squared error):** It is the square root of mean squared error.
4. **R2 (R square) :** It determines the proportion of variance in the dependent variable that can be explained by the independent variable. It is also called as the coefficient of determination.
5. **Adjusted R2 :** Adjusted R Square determines the extent of the variance of the dependent variable, which the independent variable can explain.

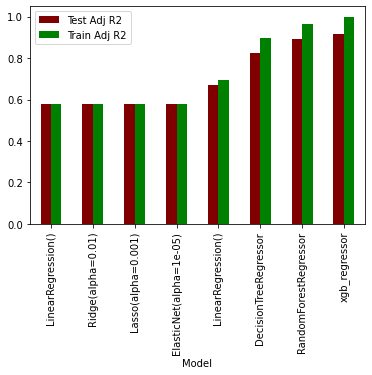
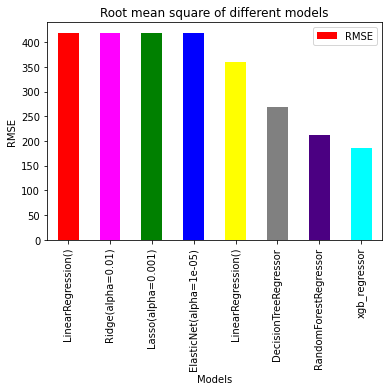
* **Models Evaluation**



Mean squared error is lowest in XGB regressor, also

Best adjusted R^2 value is obtained in extreme gradient boosting regressor.

It indicates that XGB is performing best among all the other models**.**

**Visual comparison of Test adjusted R^2 AND Train adjusted R^2****Visual comparison of Root Mean Squared Error of different models**

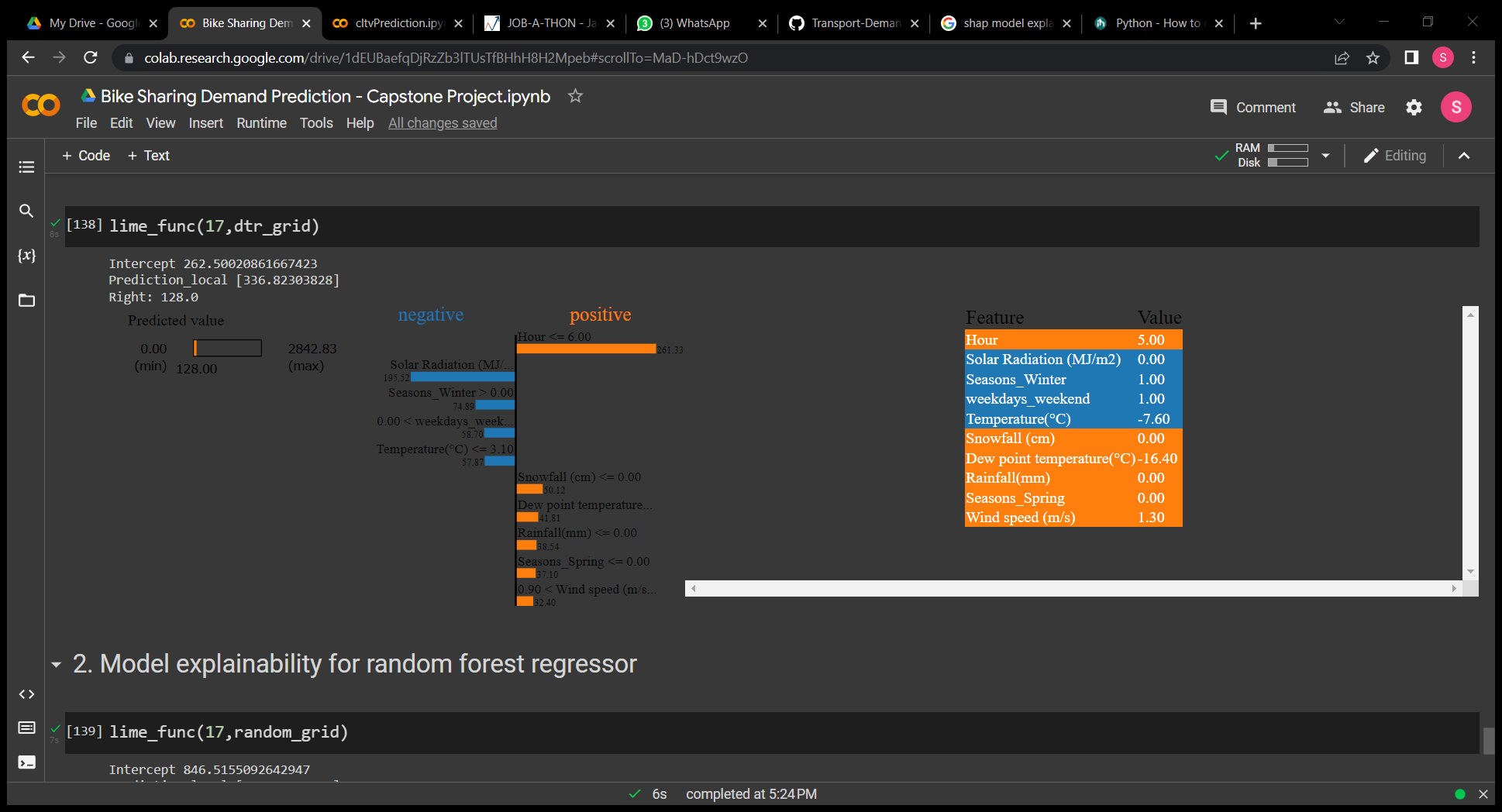
**INFERENCE:**

It is clearly visible that Adjusted r2 and Root mean squared value of XGB is lowest among all other models.

* **MODEL EXPLAINABILITY**

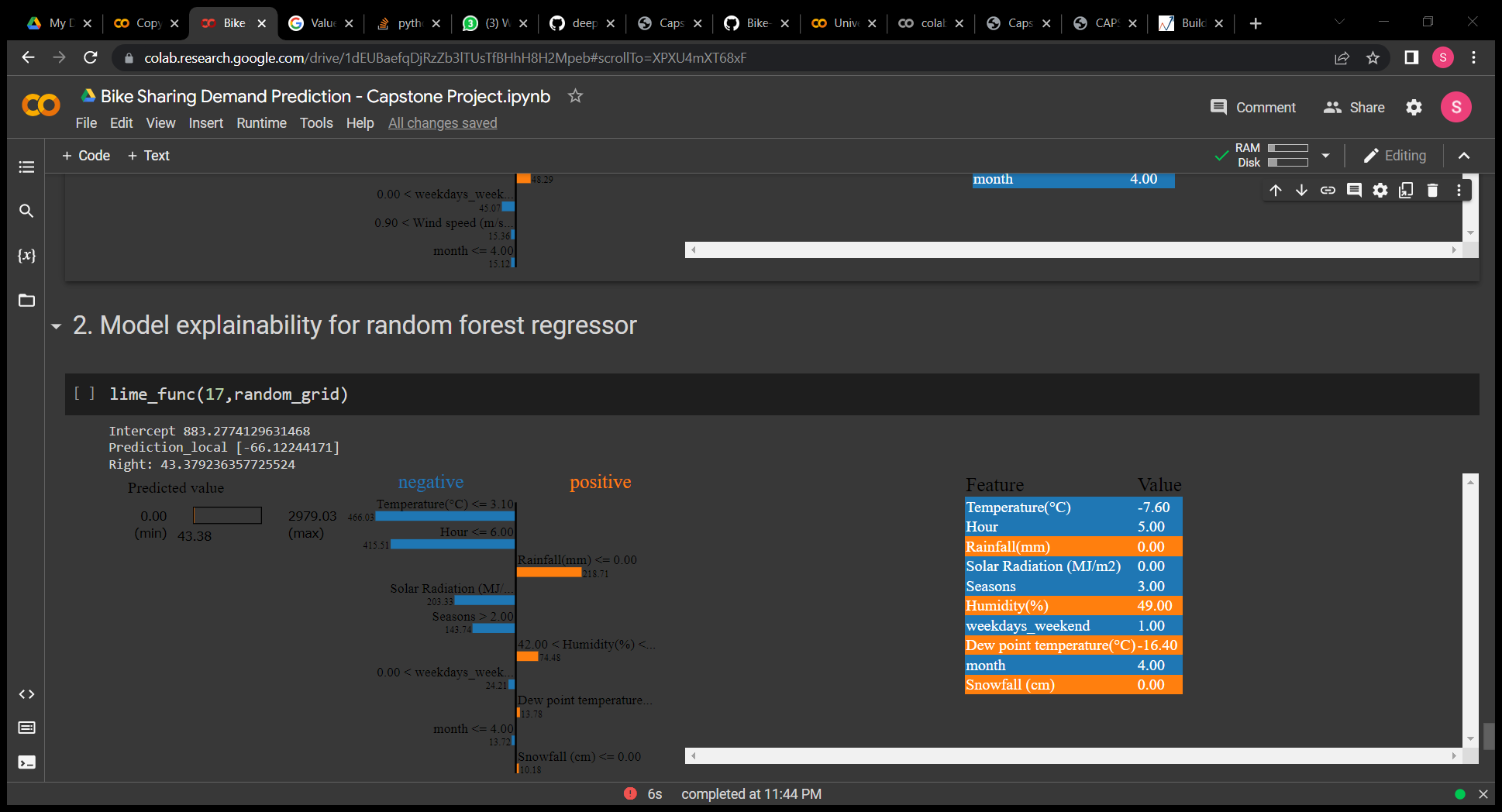
Model explainability refers to the ability to understand and interpret the decisions made by a machine learning model. This includes understanding how the model arrived at a particular prediction, as well as understanding the factors that contributed to that prediction.

1. **Decision tree regressor**:



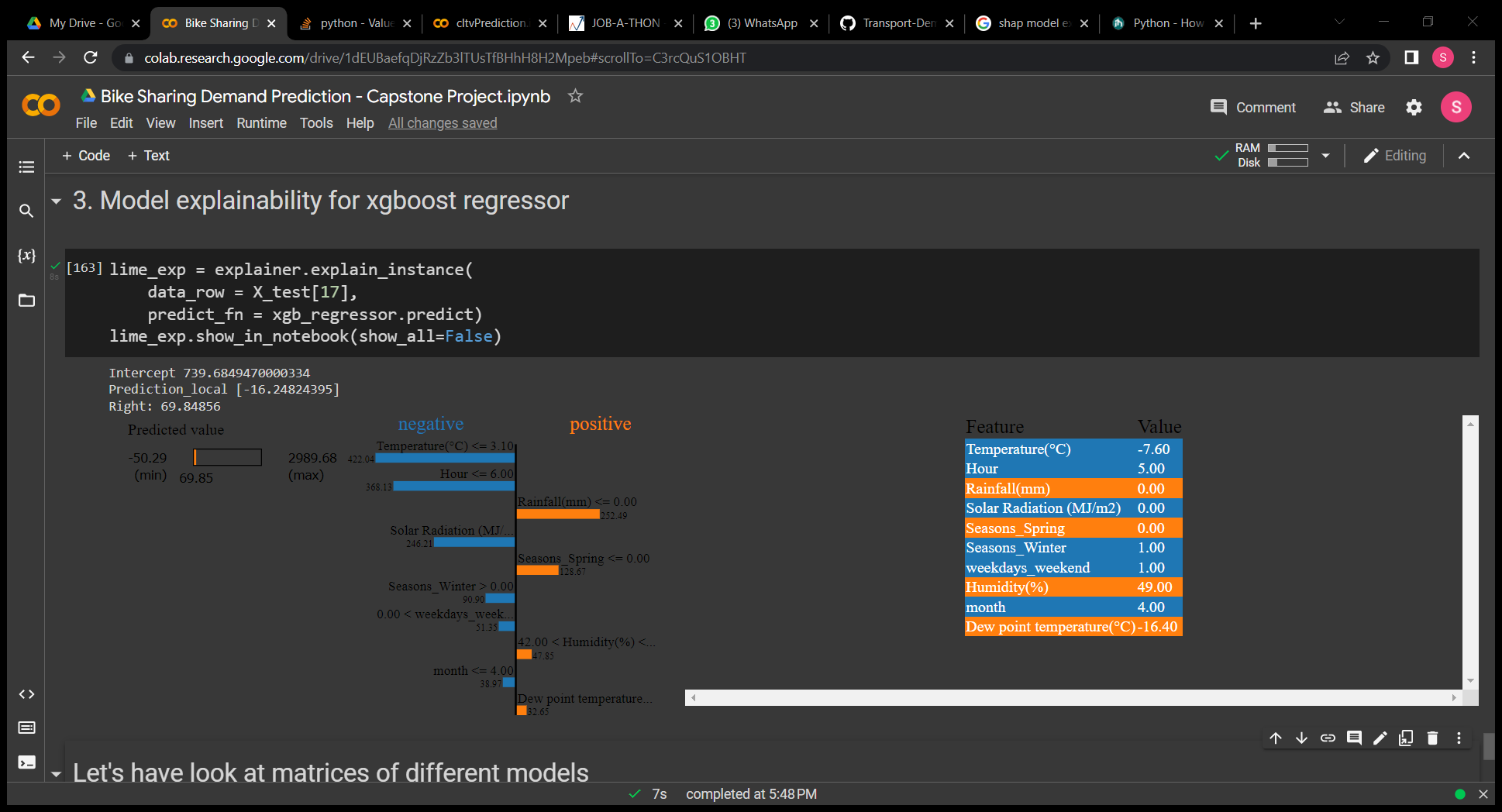
For the particular observation no 17, I used model explainability technique (LIME). In Decision tree, Hour is positively affecting the rented bike count while solar radiation is negatively affecting the rented bikes.

1. **Random Forest regressor**:



For the particular observation no 17, I used model explainability technique (LIME). In Random forest regressor, Temperature, Hour is negatively affecting the rented bike count while rainfall is positively affecting the rented bikes.

c. **Extreme Gradient boosting regressor**:



For the particular observation no 17, I used model explainability technique (LIME). In XGB regressor, Temperature, Hour is negatively affecting the rented bike count while rainfall is positively affecting the rented bikes. Solar radiations are also affecting negative for this observation.

# 6. CONCLUSION~

* Since there is no linear dependency among dependent and independent variables, hence Linear regression models are not giving the greater results.
* Hence tree based models are preferred over linear models to predict the count of rented bike.
* Among the tree based models, XGB is providing the best result
* In terms of adjusted R2 and Root mean squared error.
* one can prefer XGB for predicting the count of rented bikes given the weather condtion.

# 7. FUTURE WORK~

It will be beneficial for bike rental companies and city officials to better understand the patterns and trends of rented bike usage, and to make more informed decisions about where to place bikes and how to manage bike rental services. This can also help to optimize the fleet and the maintenance of the bikes. This will help them in predicting the rented bike to be needed by riders at each hour of the day given weather conditions.

**CHALLENGES FACED~**

* Comprehending the problem statement, and understanding the business implications.
* Feature engineering – deciding on which features to be dropped/kept/ transformed so that assumptions of linear models is satisfied.
* Choosing the best visualization to show the trends among different features clearly in the EDA phase.
* Deciding on how to handle outliers.
* Choosing the ML models to make predictions.
* High computational time in hyper parameter tuning for ensemble models.

**References~**

* GeeksforGeeks
* Analytics Vidhya
* Stackoverflow
* Towards data science
* Python libraries documentation

THANK YOU