

# CAPSTONE PROJECT

## Bike Sharing Demand Prediction

### TEAM MEMBERS

Aditya Tadas

Nikhil Machave

Aishwarya Methe



# CONTENT

- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- FEATURE ENGINEERING
- EXPLORATORY DATA ANALYSIS
- DATA CLEANING
- DATA PREPROCESSING
- IMPLIMENTING ALGORITHEMS
- HYPERPARAMETER TUNING OF BEST FIT ALGORITHEMS
- CONCLUSIONS



# BUSINESS UNDERSTANDING

- Bike rentals services are often use by the peoples now a days hence this business becomes more popular having relatively cheaper rates also ease to pick up and drop at own convenience is one of the best thing to make business popular and profitable.
- It is mostly use by the people having no vehicles and having schedule to travel daily for there work and other related things.
- Bike rentals services are also often use for vacations day to travel for a long distance also minimize time and avoid congested public transport .
- For that reasons it is essential to have supply of no. of bikes at different locations,to fulfill the demand.
- Our goal for these project is to predict supply of bike count according to certain conditions to meet the demand of the bikes for business Profitability.

# DATA SUMMARY

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

- Here We have Data Summary of our data Which shoes that these dataset contains 8760 Rows and 14 Columns.
- Three Categorical Features 'Seasons', 'Holiday', and 'Functioning Day'.
- One Date Time Feature "Date".
- Also we have some numerical variables such as Temperature ,Humidity, Wind Speed, Visibility, Solar Radiation, Rainfall, Snowfall which tells us how environmental coditions affects Rental Bike Count.

# Feature Analysis

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m<sup>2</sup>
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

# INSIGHTS FROM OUR DATASET

- There are No missing values present in our dataset.
- There are No duplicate values present in our dataset.
- There are No Null values in our dataset.

```
# finding how many null value are present in each column  
bike_df.isnull().sum()
```

```
Date                0  
Rented Bike Count    0  
Hour                0  
Temperature(°C)      0  
Humidity(%)          0  
Wind speed (m/s)     0  
Visibility (10m)     0  
Dew point temperature(°C) 0  
Solar Radiation (MJ/m2) 0  
Rainfall(mm)         0  
Snowfall (cm)        0  
Seasons              0  
Holiday              0  
Functioning Day      0  
dtype: int64
```

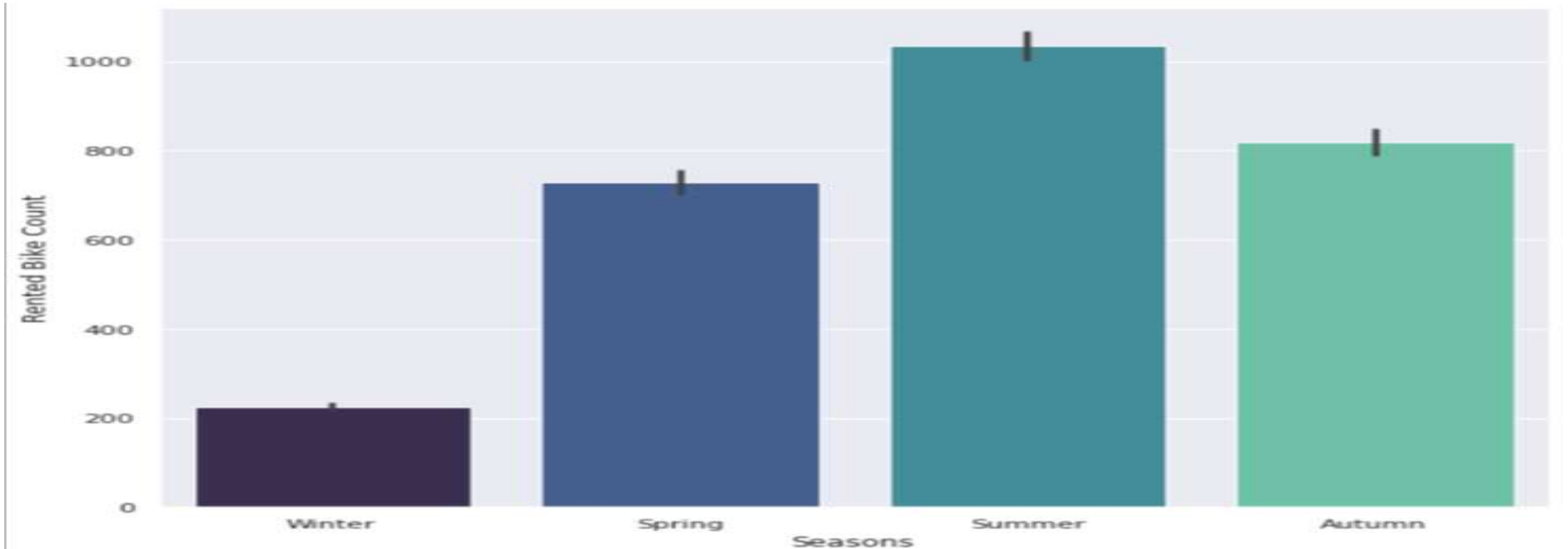
```
# Finding how many duplicate rows in our dataset  
bike_df.duplicated().sum()
```

```
0
```

# FEATURE ENGINEERING

- We convert Date column into three different column i.e 'year', 'month,' 'Week day.
- We replace month number in words for understanding.  
i.e 'Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'
- We replace Day of week into into numbers i.e  
'Mon','Tues','Wed','Thur','Fri','Sat','Sun'

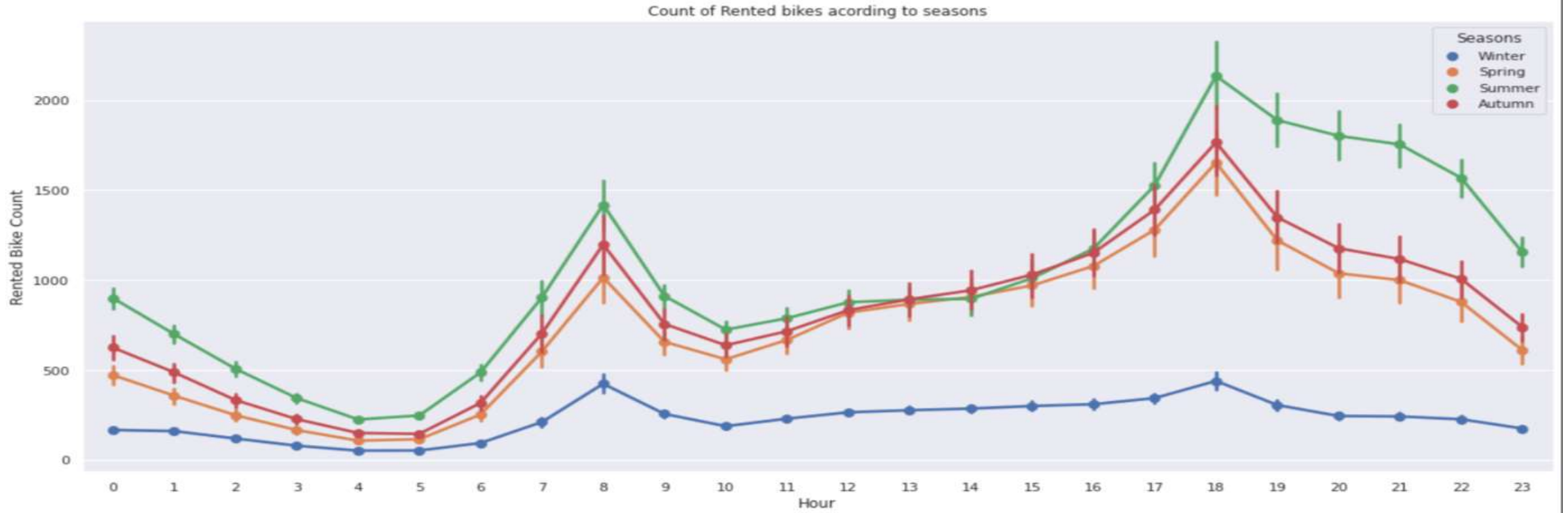
# ANALYSIS OF SEASONS VARIABLE



- Above Bar plot shows that season wise distribution of rented bike counts.
- We can see that rented bike counts are maximum in summer season and it is minimum in winter season.
- The spring and Autumn season has same rented bike counts.

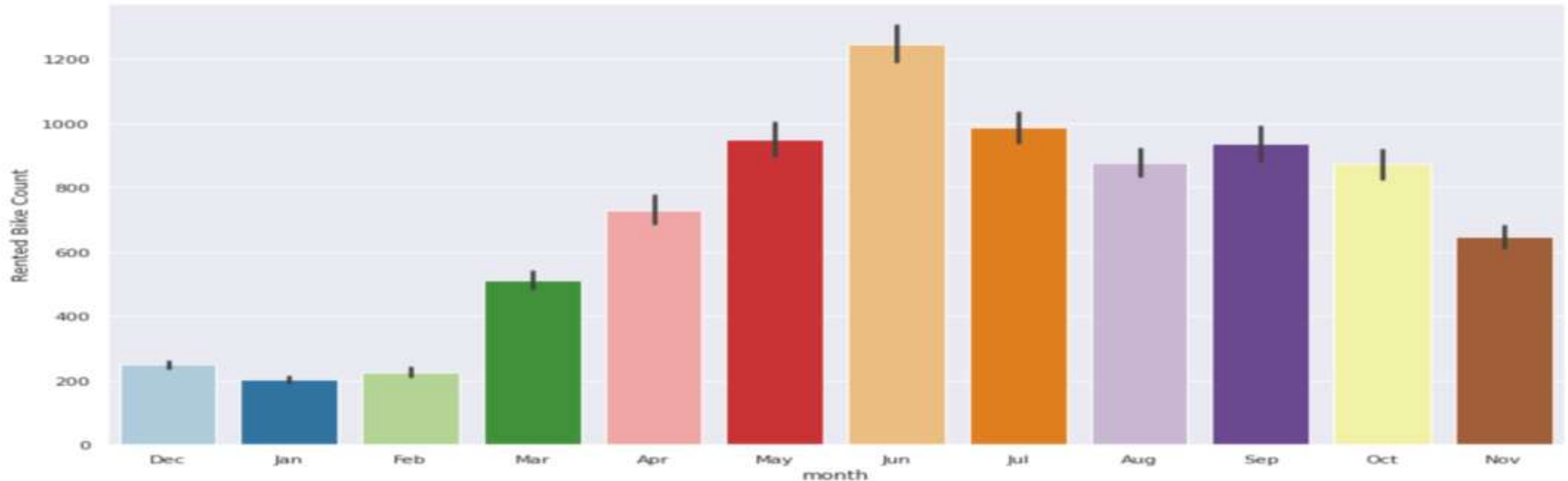


# ANALYSIS OF SEASONS VARIABLE WITH RESPECT TO HOUR VARIABLE



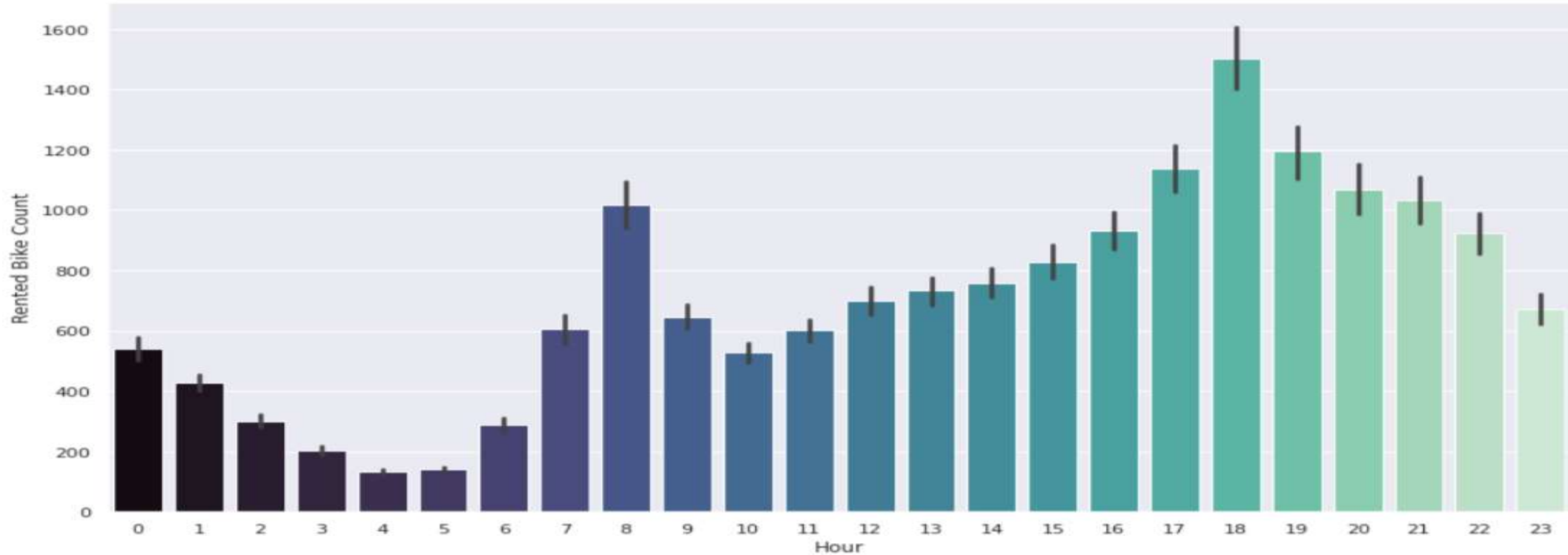
- Above line plot shows rented bike counts in seasons with respect to hours.
- From the above visualization we can clearly see that maximum rented bike counts have a high fluctuation in between 5 am to 8 am also in between 5 pm to 7 pm so most people have rented the bikes in between these two time intervals.

# ANALYSIS OF MONTH VARIABLE



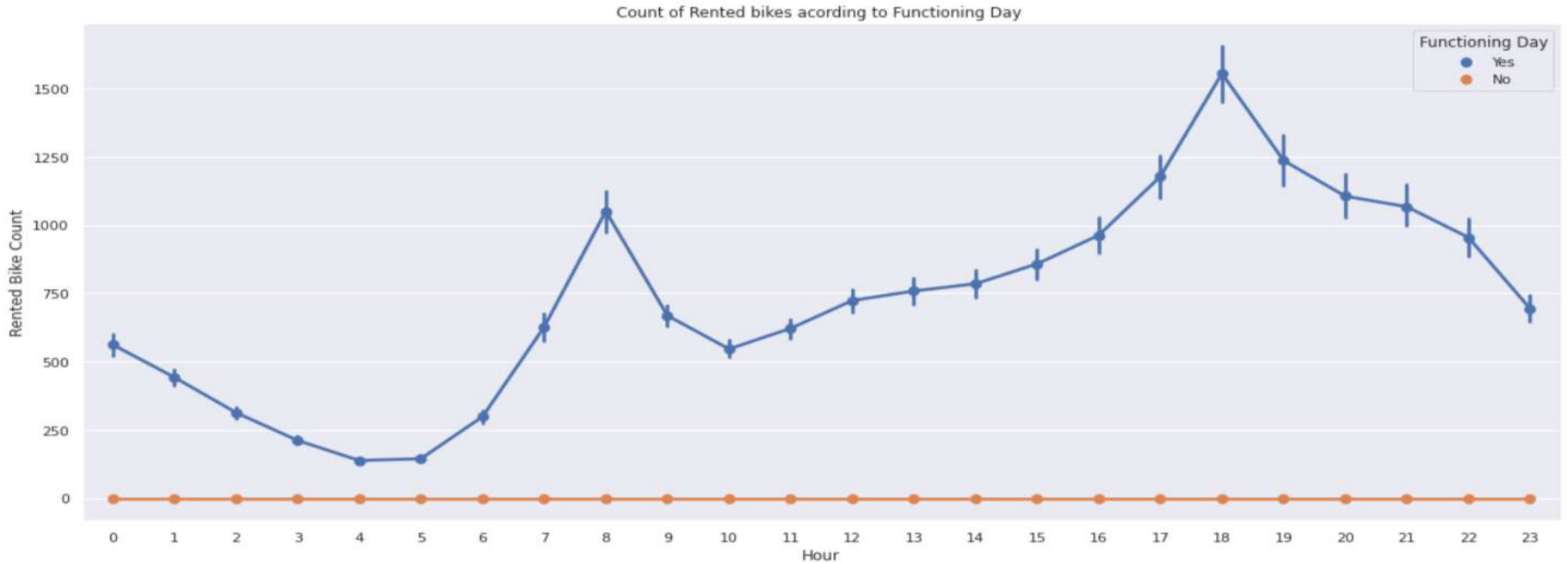
- In the above Bar plot as we can see that maximum bikes are rented in May , June , July months.
- Intermediate bikes are rented in August , September and Oct months.
- And there are comparatively minimum bikes are rented in December , January , Feb and November Months.

# ANALYSIS OF HOUR VARIABLE



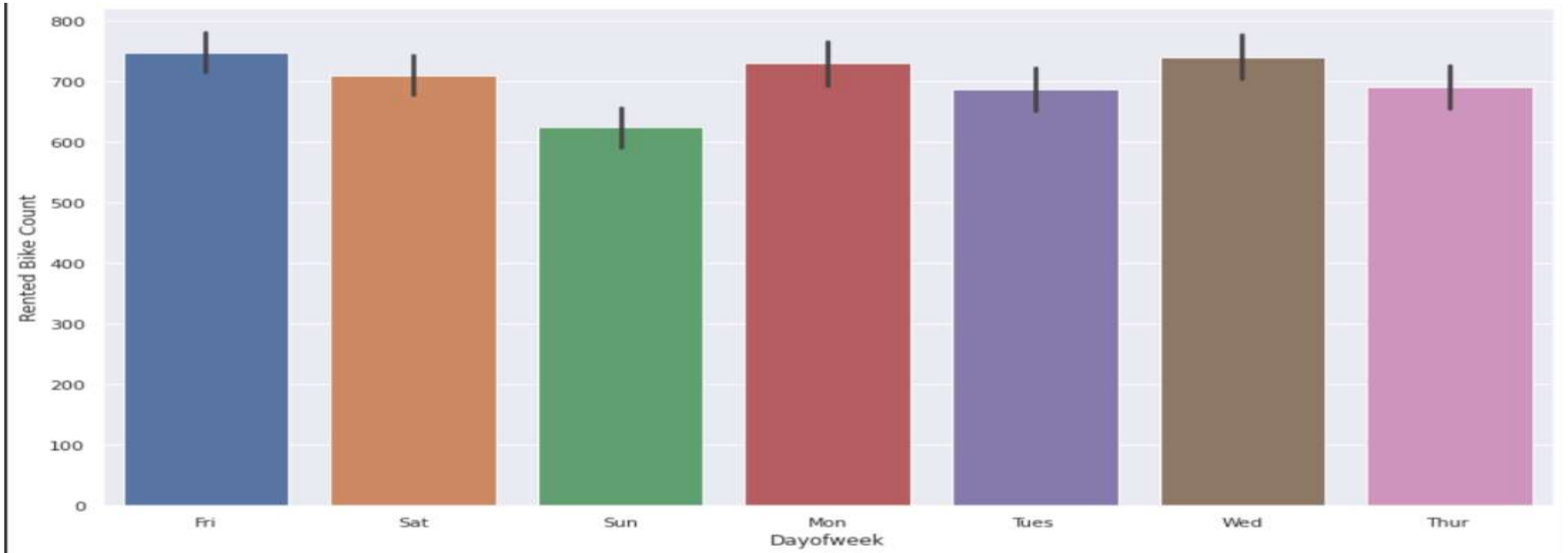
- As we can see in the following bar plot more bikes are rented in the time interval of 7 am to 9 am and 5 pm to 8 pm probably the people rented more bike to reach at office and to come to the residual place from the office
- Also we have minimum bikes are rented in between 1 am to 5 am .

# ANALYSIS OF FUNCTION DAY VARIABLE WITH RESPECT TO HOUR



- As we can see in the above line plot when there is no function day peoples are not prefer to rent bikes.
- Though when there is a function day we can see the high pick in between 6 am to 9 am as well as in between 5 pm to 8 pm.

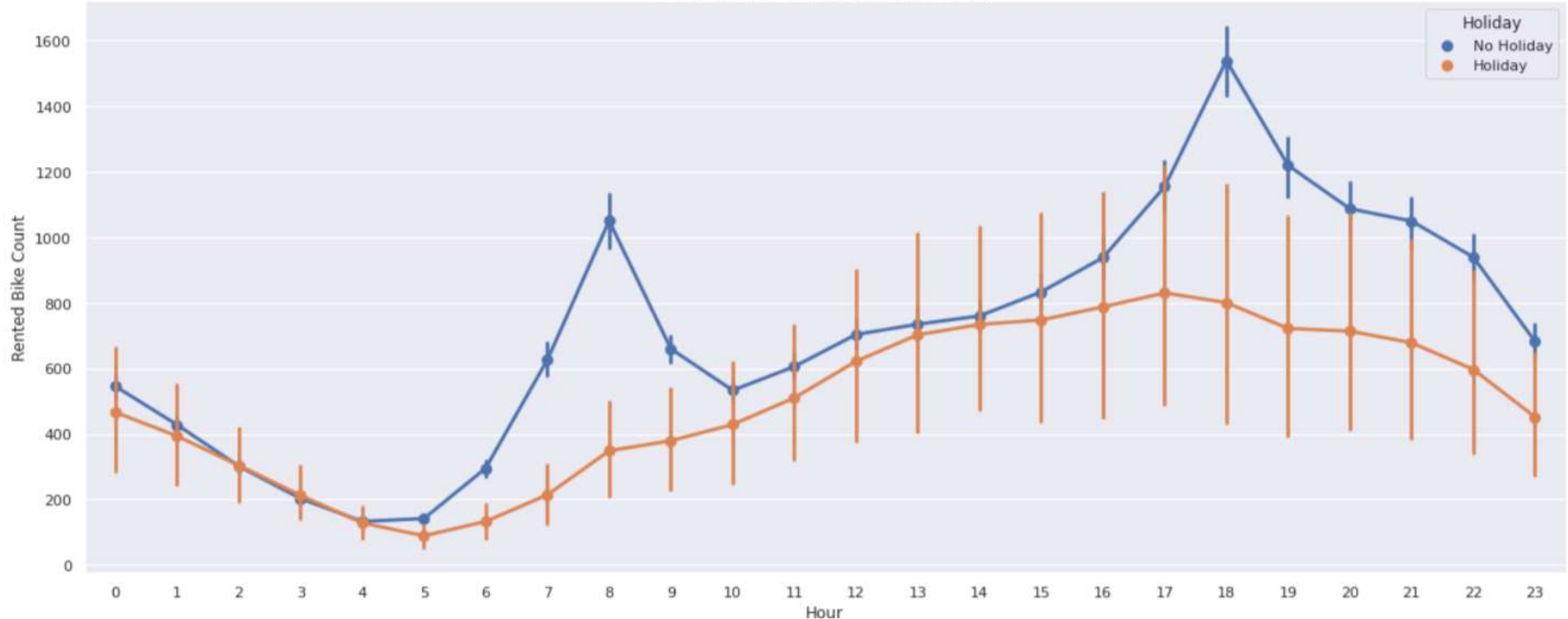
# ANALYSIS OF DAY OF WEEK VARIABLE



- As we can see in the following bar plot of day wise distribution of rented bike counts we can conclude that on the weekend day we have minimum bike rented as compared to weekday,

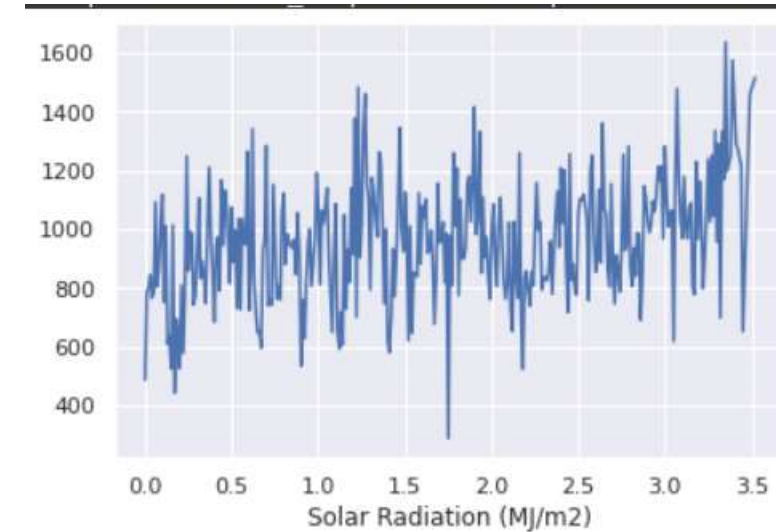
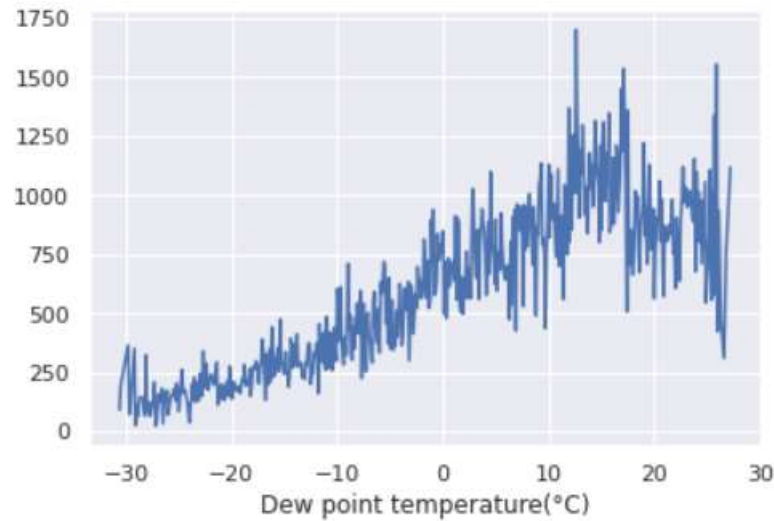
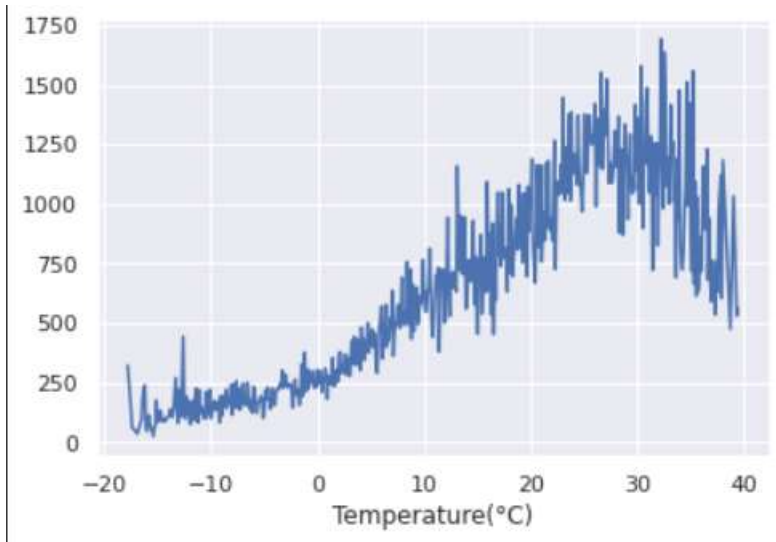
# ANALYSIS OF HOLIDAY VARIABLE WITH RESPECT TO HOUR

Count of Rented bikes according to holiday



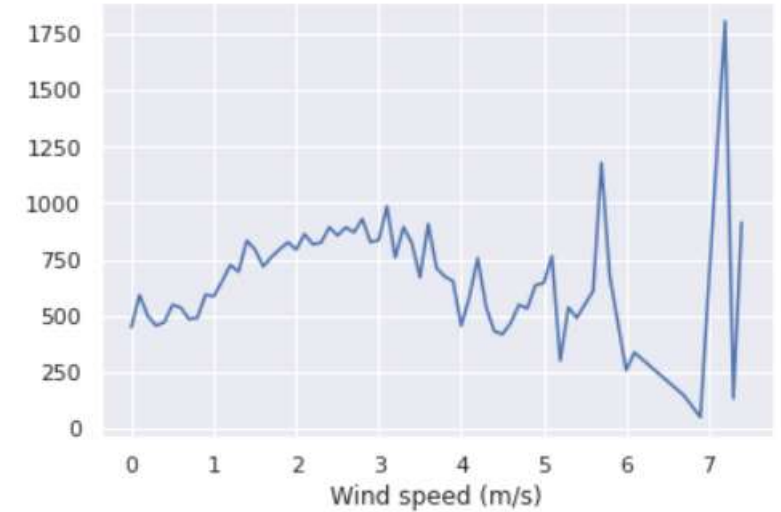
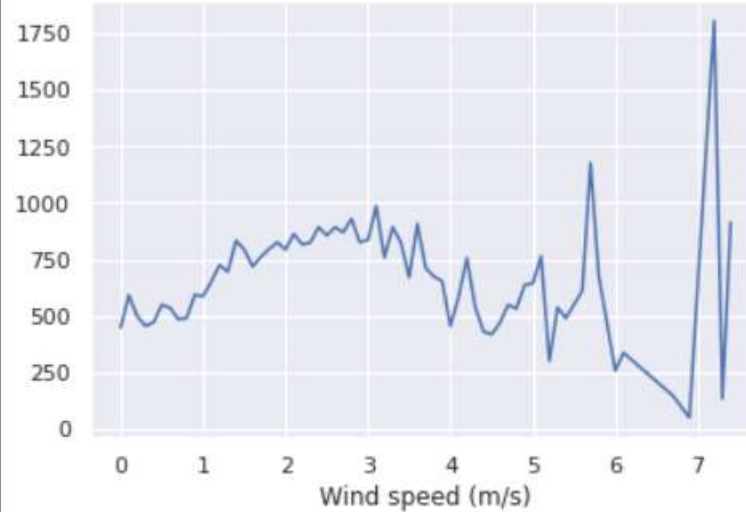
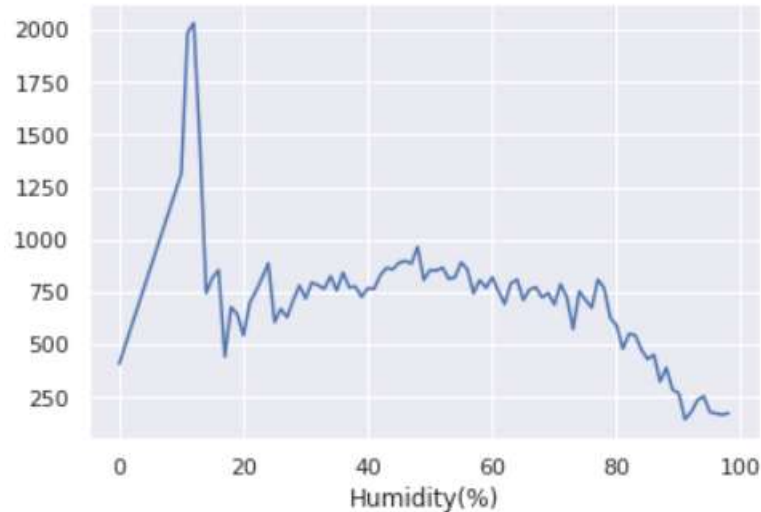
- In the following line plot we can see that when there is No Holiday peoples rented more bikes from 6 am to 9 am and 4 pm to 7 pm.
- But when there is Holiday most number of peoples have rented bikes in between 11 am to 4 pm.

# NUMERICAL VARIABLE VS RENTED BIKE COUNTS



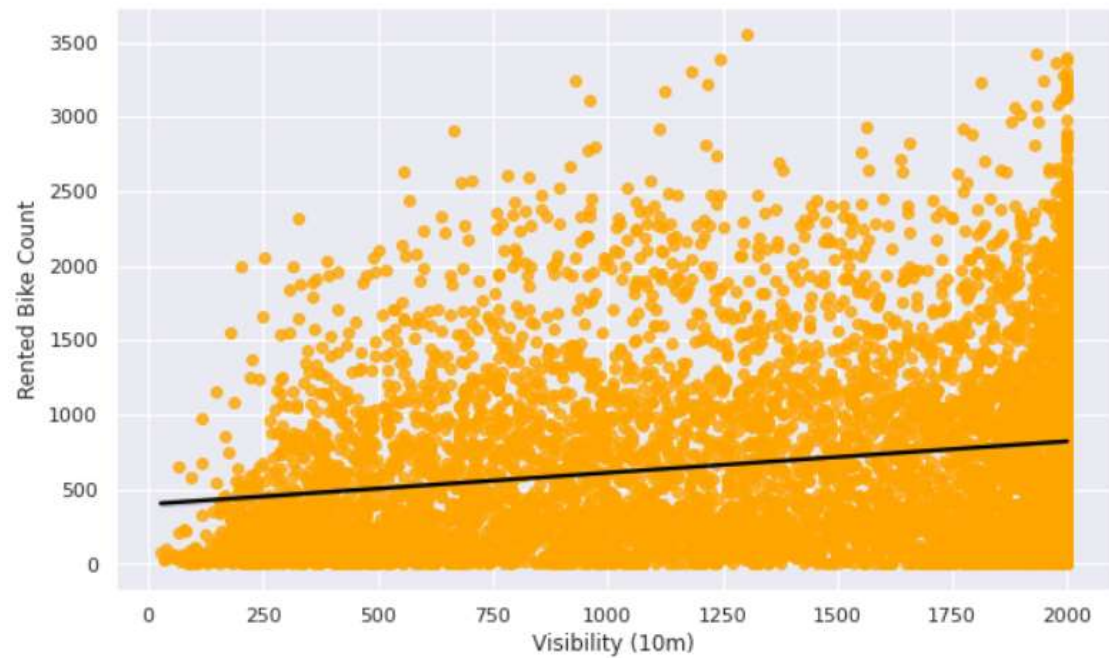
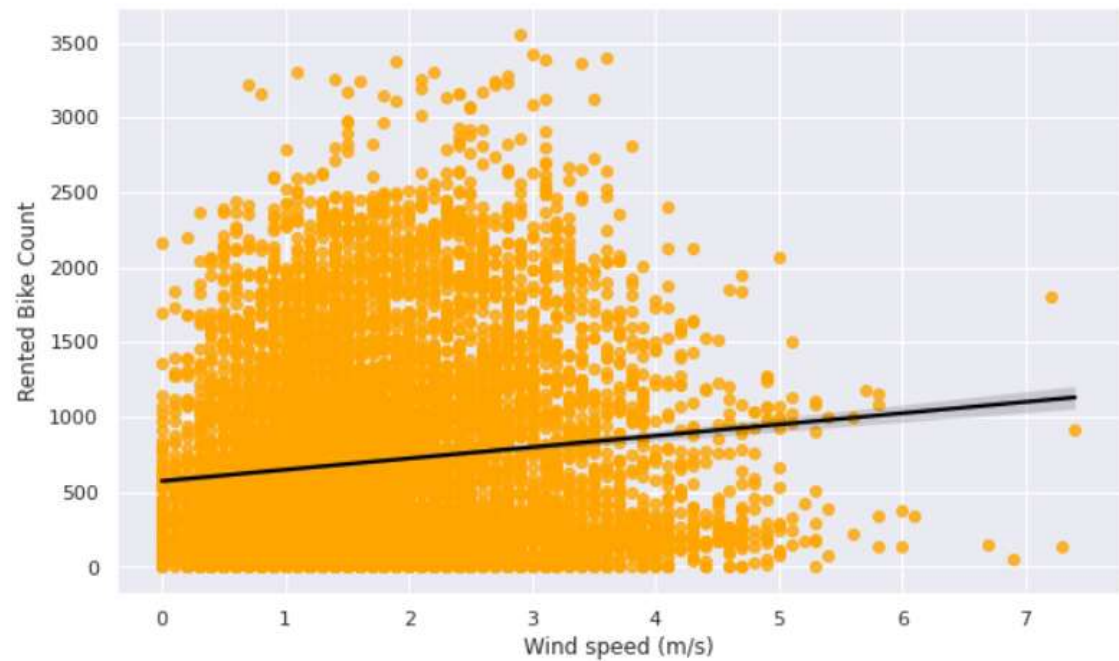
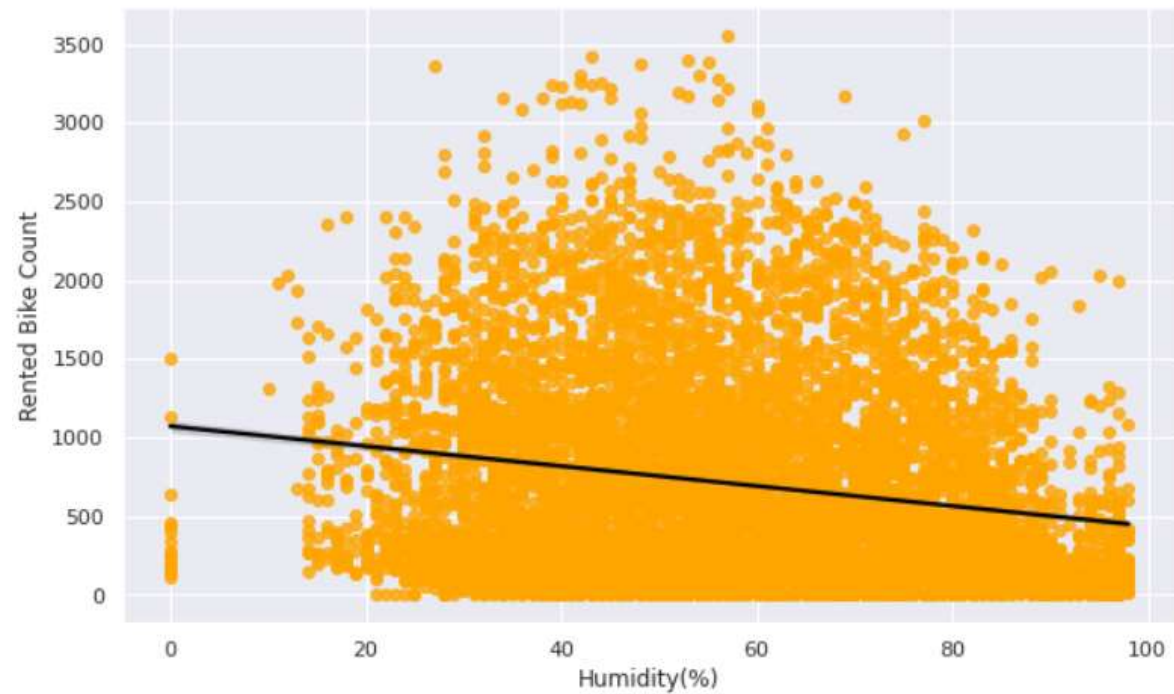
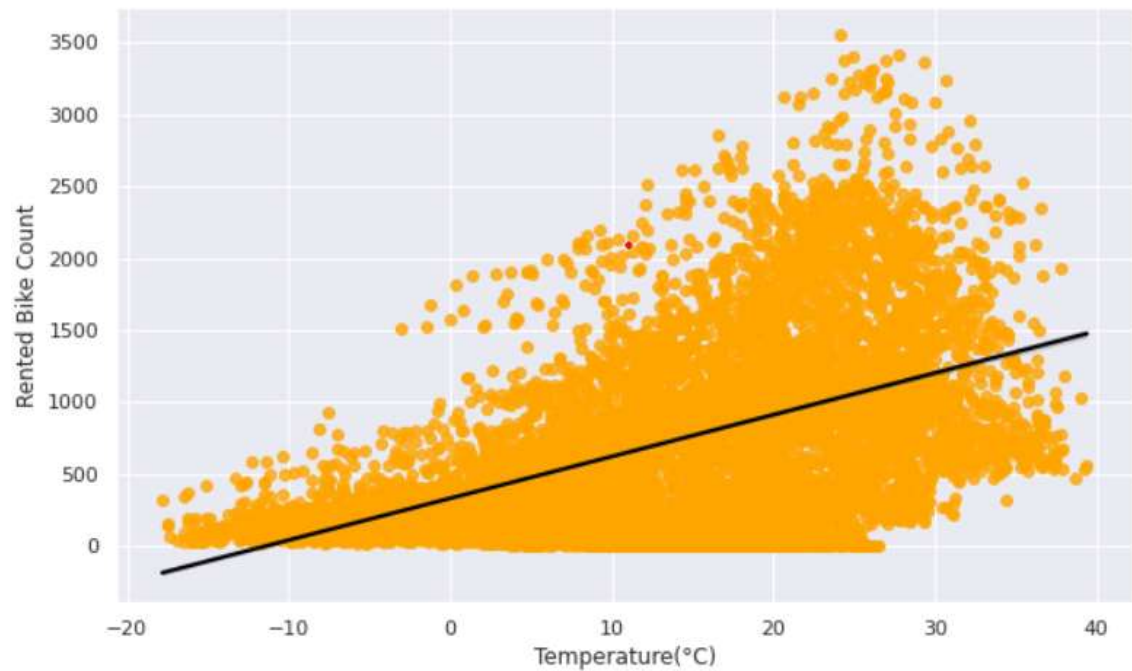
- As we can see that in the following graph No. of rented bike counts are gradually increase as the temperture increses when the temperature lies in between 25 to 35 peoples have rented more bikes soo we can say that people prefer to rent a bikes when the temperature is normal.
- From a graph of dew point temperature also number of rented bikes incresess when the Deu point temperature increases and the rented bike counts re high se we have a 10 to 20 celcius temperature so we can say that people prefers to rent bikes when Deu point temperature is normal.
- Also from the graph of solar radiation we can say that the amout of rented bikes is huge when there is solar radiation the counter of rent is around 1000

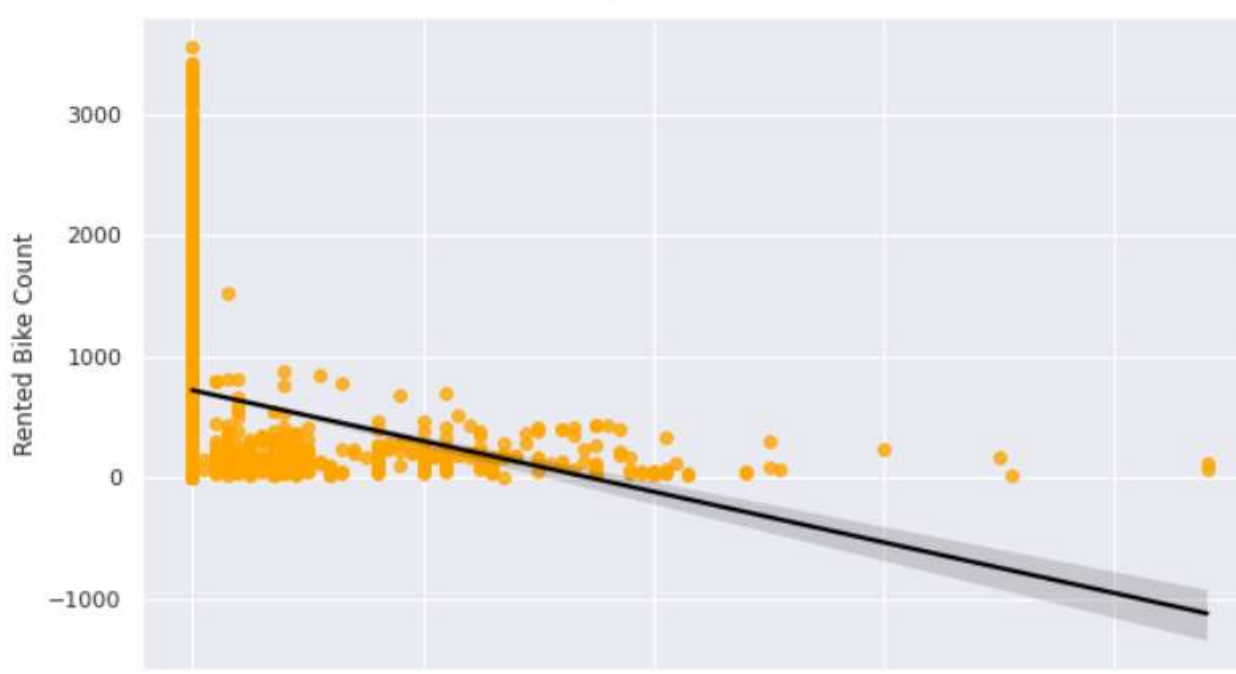
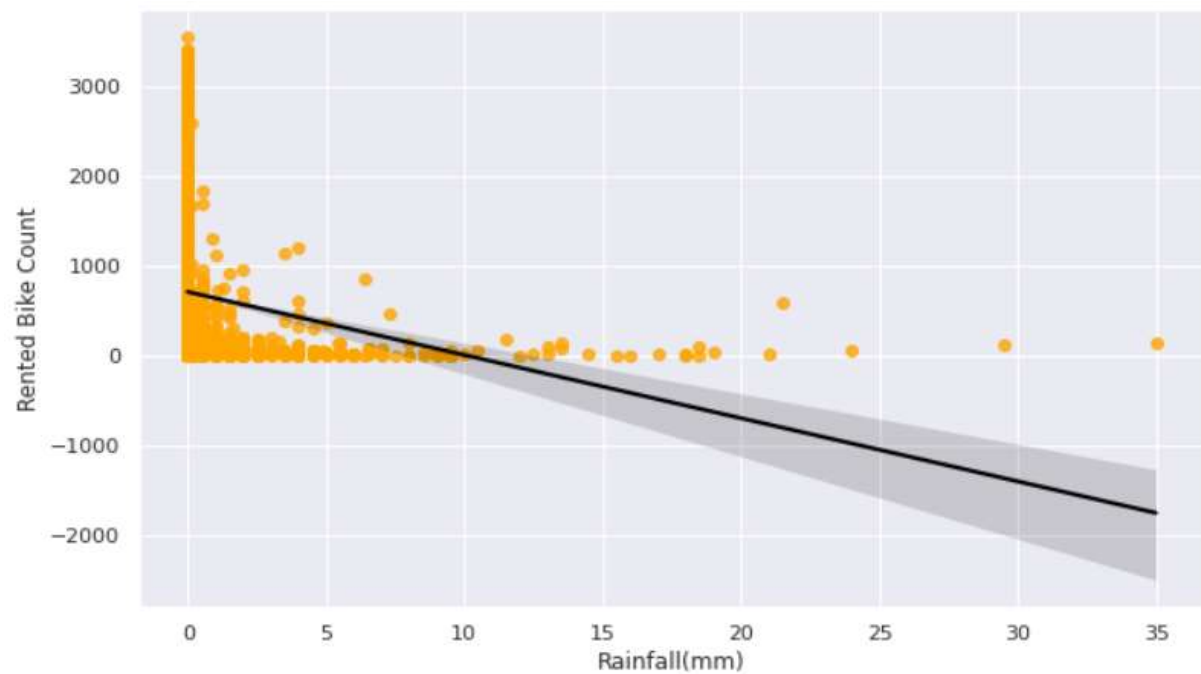
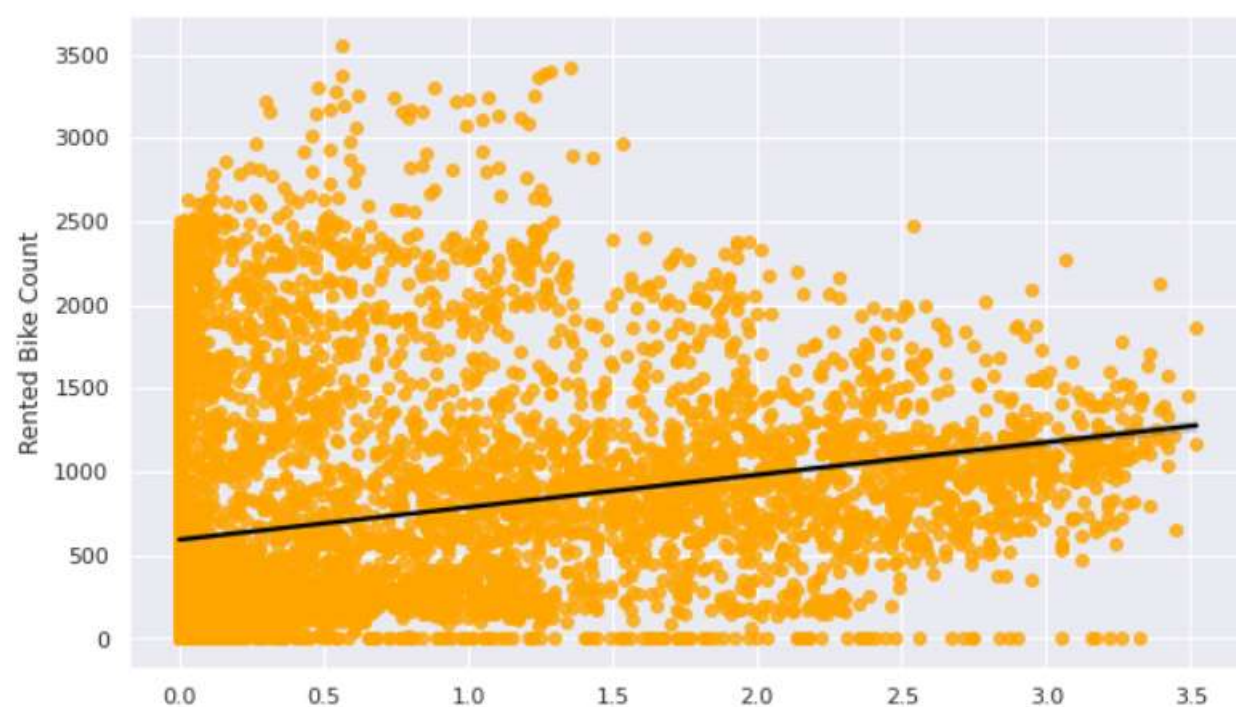
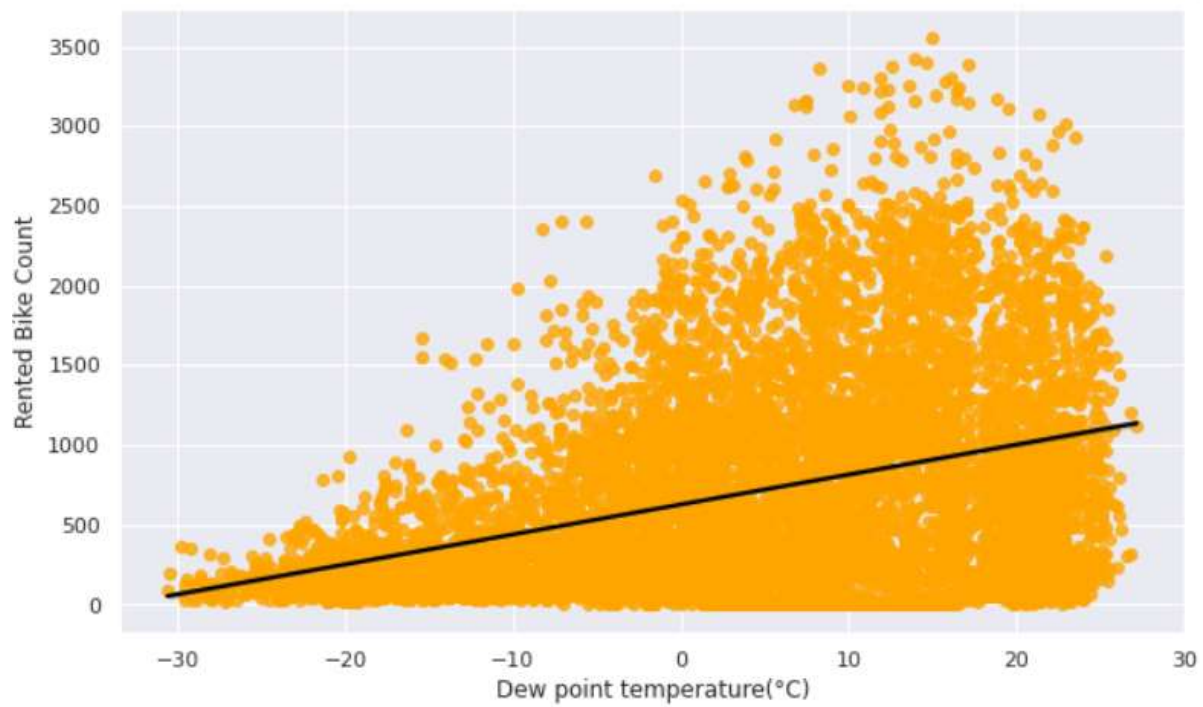
# NUMERICAL VARIABLE VS RENTED BIKE COUNTS



- Following graph of Humidity shows that when average humidity percentage is less 5 to 15 then most of the people rented bikes then as the average humidity percentage become increases number of rented bike count decreases therefore we can say that most of the people rented bikes when the humidity percentage is less.
- Then we have Wind Speed graph it shows that when wind speed is about 5 to 6 then the number of rented bikes increases
- Also when it becomes high upto 8 then we can see huge increase in number of rented bike counts therefore we can say that people prefer to rent a bike when the atmosphere is little windy.







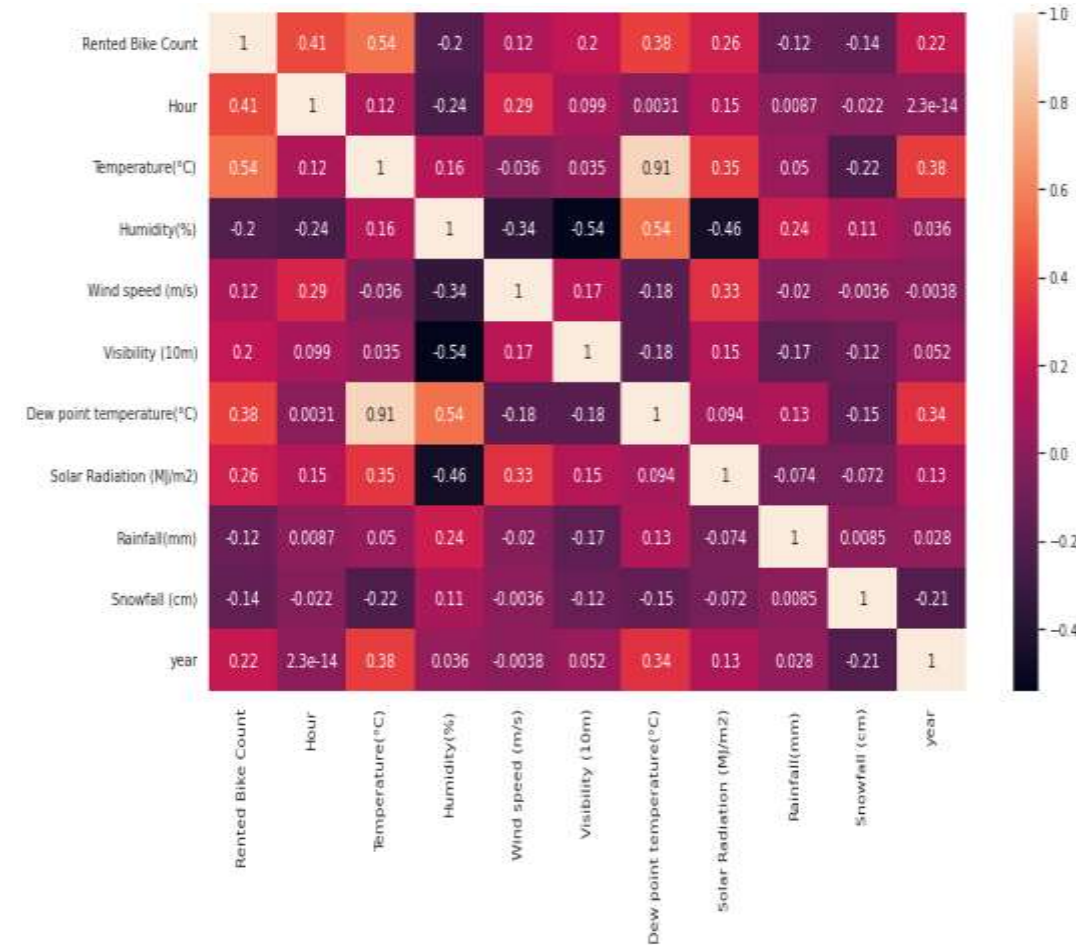
# REGRESSION PLOT FOR NUMERICAL VARIABLES

- From the above regression plots we can observe that numerical features i.e 'Temperature', 'Wind Speed', 'Visibility' , 'Dew Point Temperature' , 'Solar Radiation' having positive relation to the target Variable.
- Which means as these feature values are increases there will be a increase in Rented Bike counts.
- Also there are some features having Negative relation to the target variables these features are 'Rainfall' , 'Snow Fall' , 'Humidity' as these features values are increases Rented Bikes count Decreases.



# CORRELATION PLOT

- Some features having high correlation with our target variable  
I.e 'Temperature' , 'Hour' , ' Dew Point Temperature' these features more important while predicting rented bike counts.
- Some features having very low co-relation with rented bike counts  
i.e 'Rainfall' , 'Snow Fall' .
- 'Temperature' and 'Due Point Temperature' are the high correlated variables with each it has a correlation of 0.91 hence we need to drop one of these variable because It leads to acts like a duplicate while we are training our machine learning model.



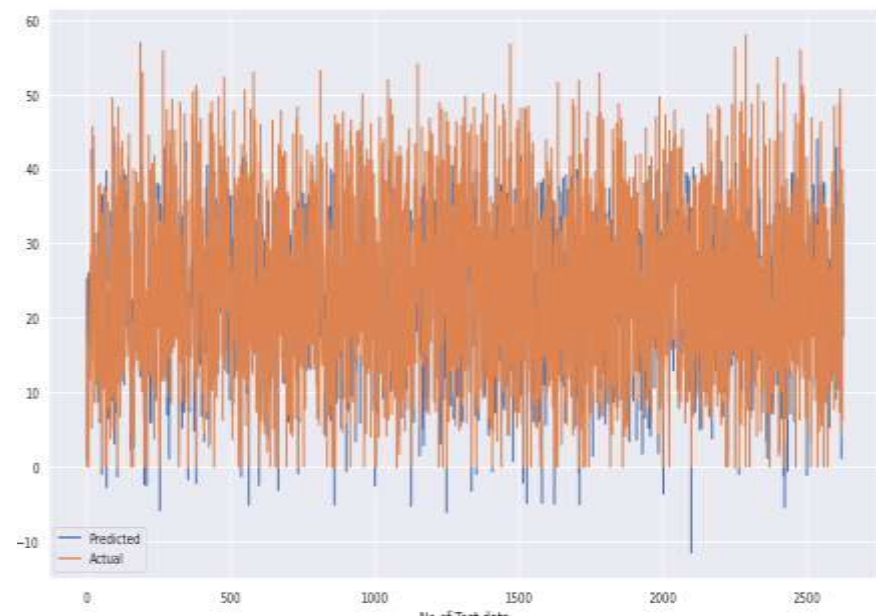
## LINEAR REGRESSOR

TRAIN SCORE

TEST SCORE

MSE : 52.19126735573527  
RMSE : 7.224352383136863  
MAE : 5.585665241559689  
R2 : 0.6669869851784311  
Adjusted R2 : 0.6622296563952659

MSE : 51.63085096806353  
RMSE : 7.185461082495926  
MAE : 5.451877045878587  
R2 : 0.657937923163272  
Adjusted R2 : 0.6530513220656045



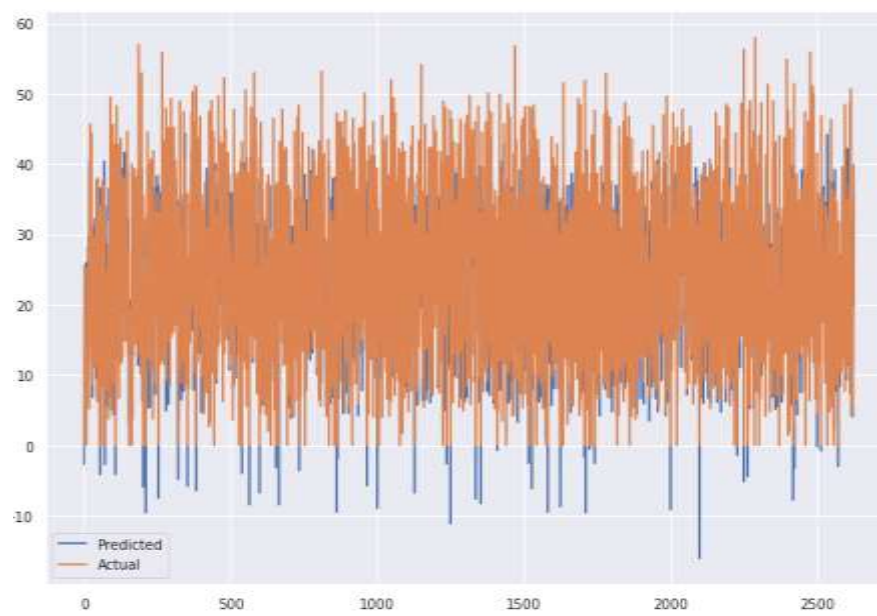
## LASSO

TRAIN SCORE

TEST SCORE

MSE : 49.13078267249149  
RMSE : 7.00933539449294  
MAE : 5.398256122706897  
R2 : 0.6865147966844348  
Adjusted R2 : 0.6820364366370697

MSE : 49.216837472871006  
RMSE : 7.015471293710139  
MAE : 5.316215850972862  
R2 : 0.6739311220781652  
Adjusted R2 : 0.6692729952507104



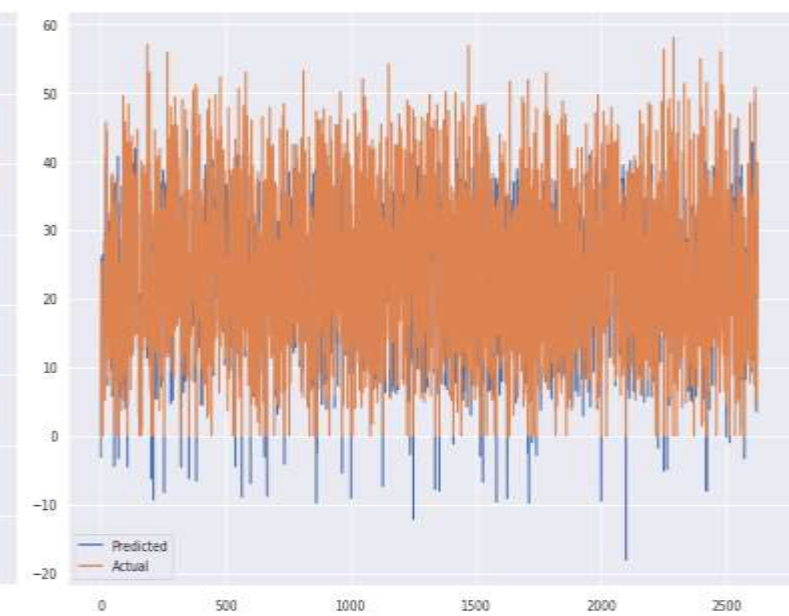
## RIDGE

TRAIN SCORE

TEST SCORE

MSE : 48.96559086601212  
RMSE : 6.997541773080895  
MAE : 5.38036528258554  
R2 : 0.6875688240013911  
Adjusted R2 : 0.6831055214871253

MSE : 49.13746708335218  
RMSE : 7.009812200291258  
MAE : 5.308803413677983  
R2 : 0.6744569627290378  
Adjusted R2 : 0.66980634791



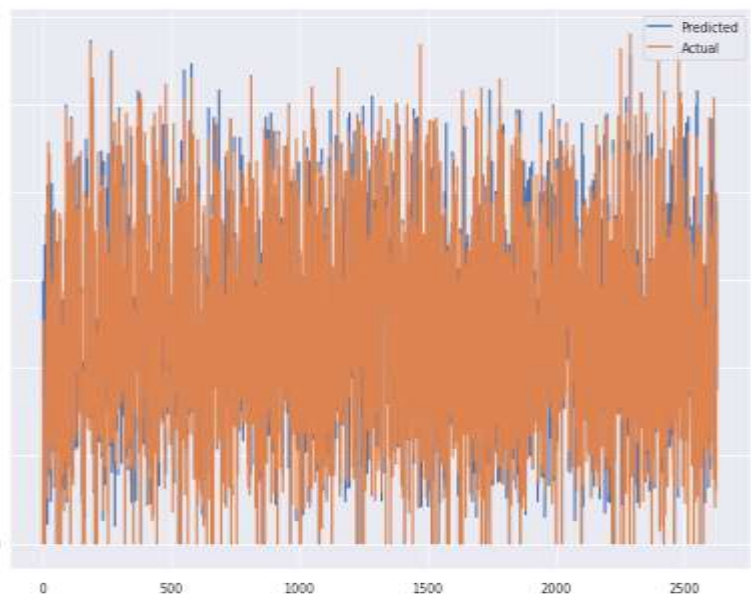
# DECISION TREE REGRESSOR

TRAIN SCORE

TEST SCORE

MSE : 0.0  
RMSE : 0.0  
MAE : 0.0  
R2 : 1.0  
Adjusted R2 : 1.0

MSE : 18.646347086154933  
RMSE : 4.318141624142837  
MAE : 2.6738531811617143  
R2 : 0.8764651736293502  
Adjusted R2 : 0.8747003903954838



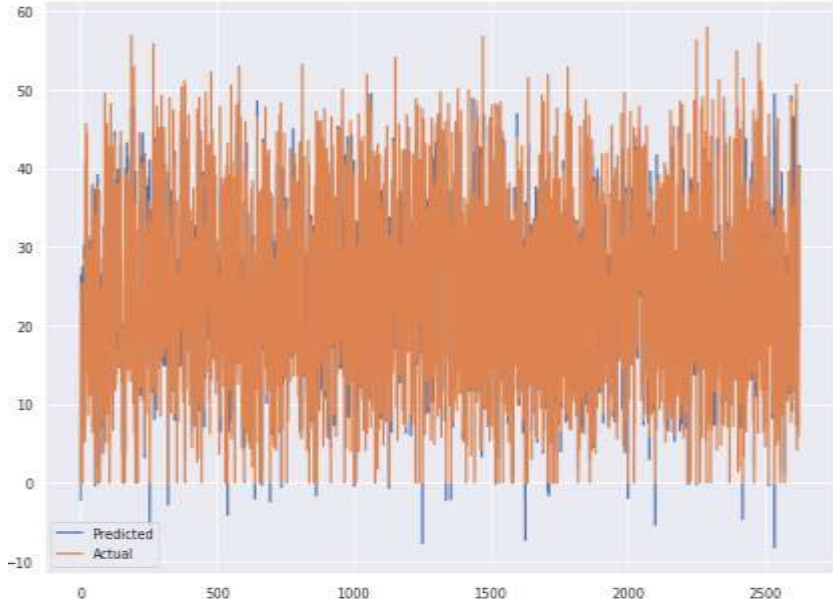
# XG BOOST REGRESSOR

TRAIN SCORE

TEST SCORE

MSE : 15.16744470681799  
RMSE : 3.894540371702159  
MAE : 2.810742713214145  
R2 : 0.9032221912809737  
Adjusted R2 : 0.9018396511564162

MSE : 17.632253196383164  
RMSE : 4.199077660199102  
MAE : 2.9907013044706305  
R2 : 0.8831836966739797  
Adjusted R2 : 0.8815148923407508



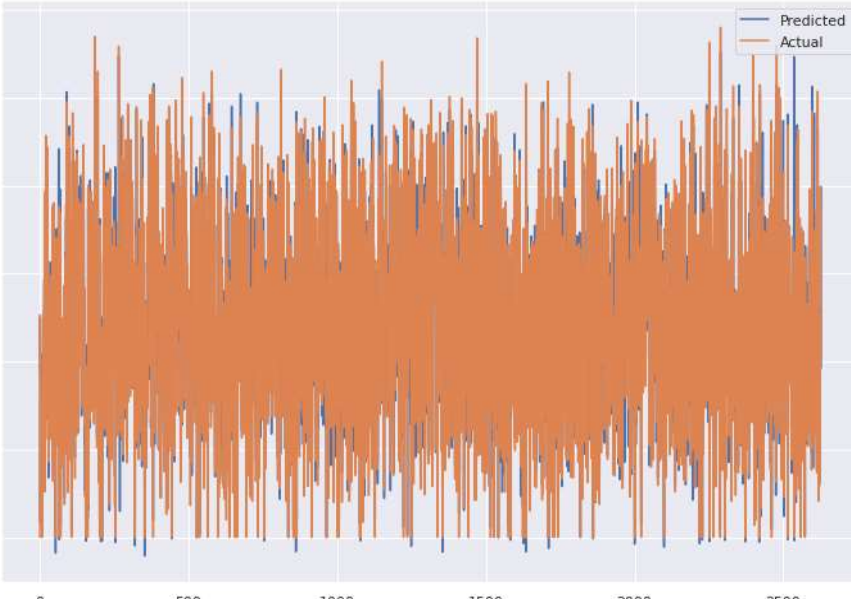
# HYPERPARAMETER TUNING OF XG BOOST

TRAIN SCORE

TEST SCORE

MSE : 0.08434638684563493  
RMSE : 0.2904244942246348  
MAE : 0.2135055042822931  
R2 : 0.9994618171583893  
Adjusted R2 : 0.9994541288320806

MSE : 10.183696207992677  
RMSE : 3.1911904061012524  
MAE : 2.0981942323331286  
R2 : 0.9325314960054601  
Adjusted R2 : 0.9315676602341096



# Model And r2 of Model

	model name	r2_train	r2_test
0	Linear regression	0.666987	0.657938
1	Lasso regression	0.686515	0.673931
2	Ridge regression	0.687569	0.674457
3	Decision Tree Regressor	1.000000	0.876465
4	Gradient Bossting regressor	0.903222	0.883184
5	Hyperparameter tuning of xg boost	0.999462	0.932531

# CONCLUSION

- EDA:

1. Demand for bikes got higher when the temperature and hour values were more.
2. Demand was high for low values of Humidity and solar radiation.
3. Demand was high during springs and summer and autumn and very low during winters.
4. Maximum bikes were rented in the year 2018
5. Count of rented bikes is high during no holiday and functioning day especially during office time.



# Model Fitting Conclusion

- **1. XG Boost regressor performs best in that dataset it gives the accuracy of 88 % also when we hypertune that model the accuracy increases by Around 5 % finally it gives the accuracy of 93 % on the testing dataset.**
  - **2. The most important features who had a major impact on the model predictions were; hour, temperature, Humidity, solar-radiation, and Winter.**
- The model performed well in this case but as the data is time dependent, values of temperature, wind-speed, solar radiation etc. will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time