BIKE SHARING DEMAND PREDICATION

DATA AVENGERS

ASHISH KUMAR PANDEY

TEAM MEMBER

VIVEK KUMAR SONI

CONTENT

- INDIVIDUAL INFORMATION
- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATARY DATA ANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING ALGORTHIMS
- CHALLENGES
- CONCLUSION

BUSINESS UNDERSTANDING

- **Bike Rentals have became a popular service in recent year and it seems people are using it more often with relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive.**
- **Mostly used by people having no personal vehicles and also to avoid Congested public transport which that's why they prefer rental bikes.**
- **Therefore , the business to strive and profit more , it has to be always ready and supply no of bikes at different location to supply and demand.**
- **Our project goal is pre planned set of bike count variable a hand solution to meet all demands.**

DATA SUMMARY

- **This Dataset contain 8760 lines and 14 columns.**
- **Three categorical features ‘season’, ’Holiday’,&’functioning’.**
- **One Date Time features Data table.**
- **We have some numerical type variables such as temperature , visibilty , dew point temp ,**
- **solar radiation , rainfall, snowfall, environment condition at that particular hour of the day .**

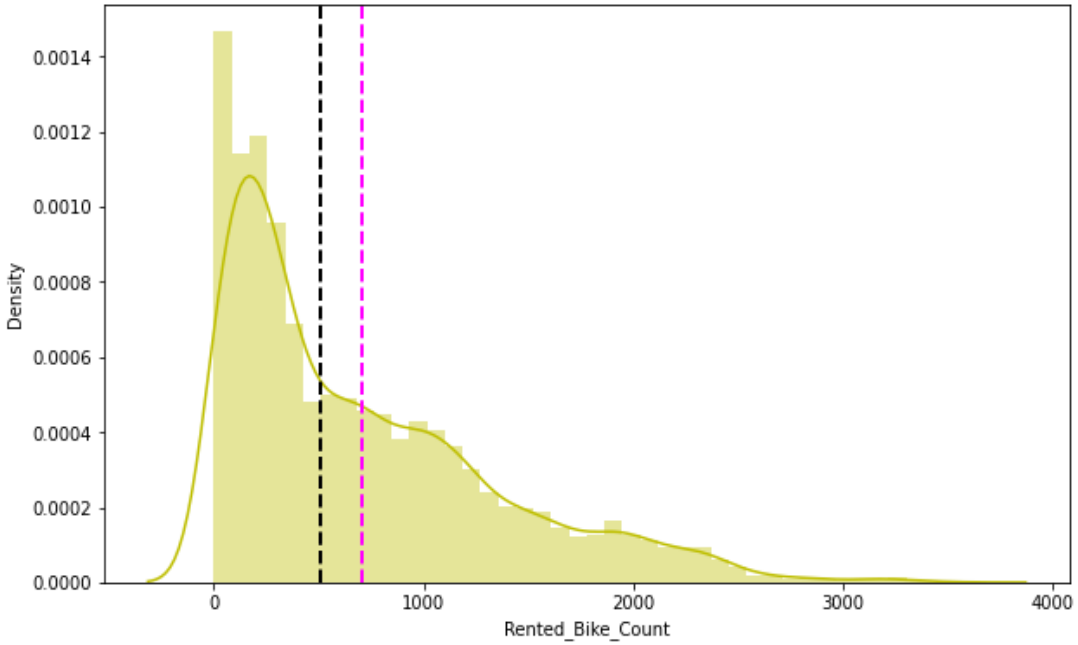
FEATURE SUMMARY

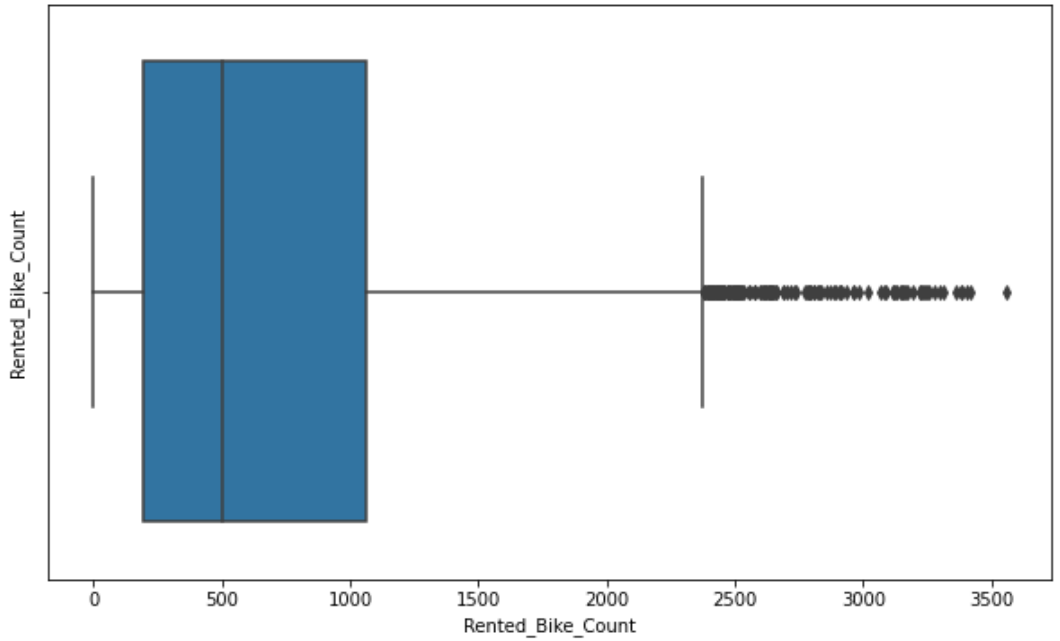
- Date :year-month-day
- Rental bike count- count of bikes rented at each hour
- Hour-hour of the day
- Temperature-Temperature in Celsius
- Humidity-%
- Wind speed-m\s
- Dew point temperature-MJ\m2
- Rainfall-mm
- Snowfall=cm
- Seasons-Summer , winter , Autumn , spring
- Holiday-Holiday/no holiday
- Functional Day-no Fun (non fun hrs),Fun(fun hrs)

INSIGHTS FORM OUR DATASET

- There is no Missing values present.
- There are no duplicate values present.
- There are no Null values.
- And Finally we have ‘rented bike counted’ variable which we need to predict for new observation.
- The dataset show hourly rental data for one year (1 dec 2017\30 nov 2018)(365 days), we consider this as a single year data.
- So we convert the data column into 3 different column year, month,day.
- We change the name of some features of our convenience,they are ‘rented_bike_count’,’hour’,’temperature’,’Humidity’,’Wind_speed’,’Dew_point_temperature’,’solar_Radiation’,’Rainfall’,’snowfall’,’Function_Day’,’Month’,’Weekdays_weekend.

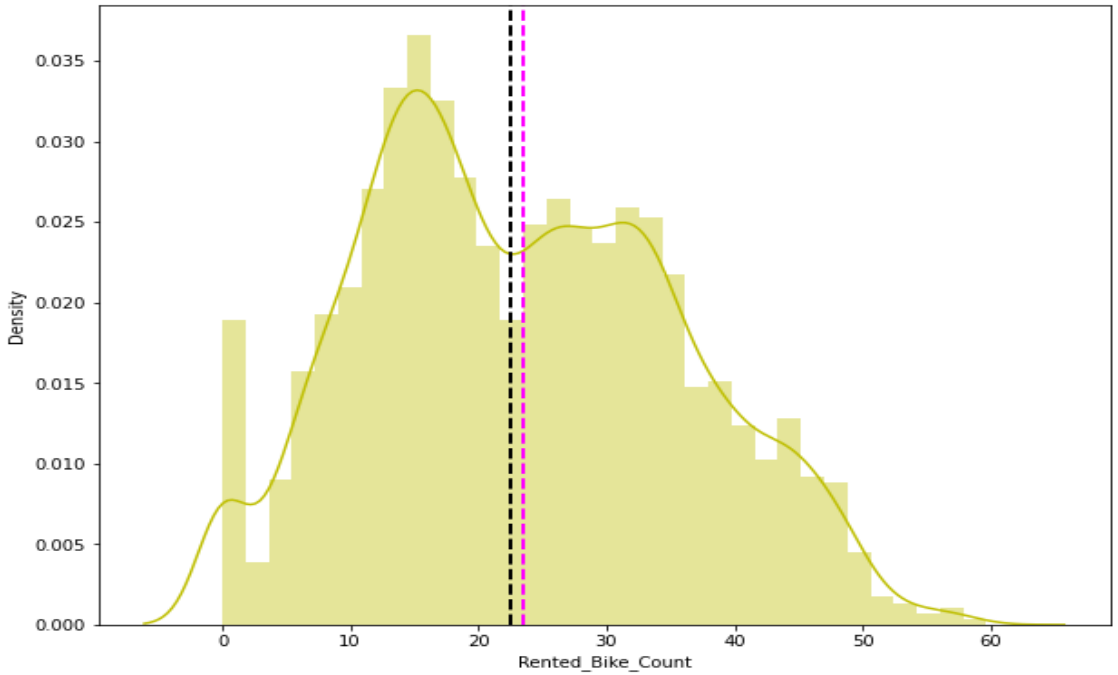
ANALYSIS OF RENTED BIKE COLUMN

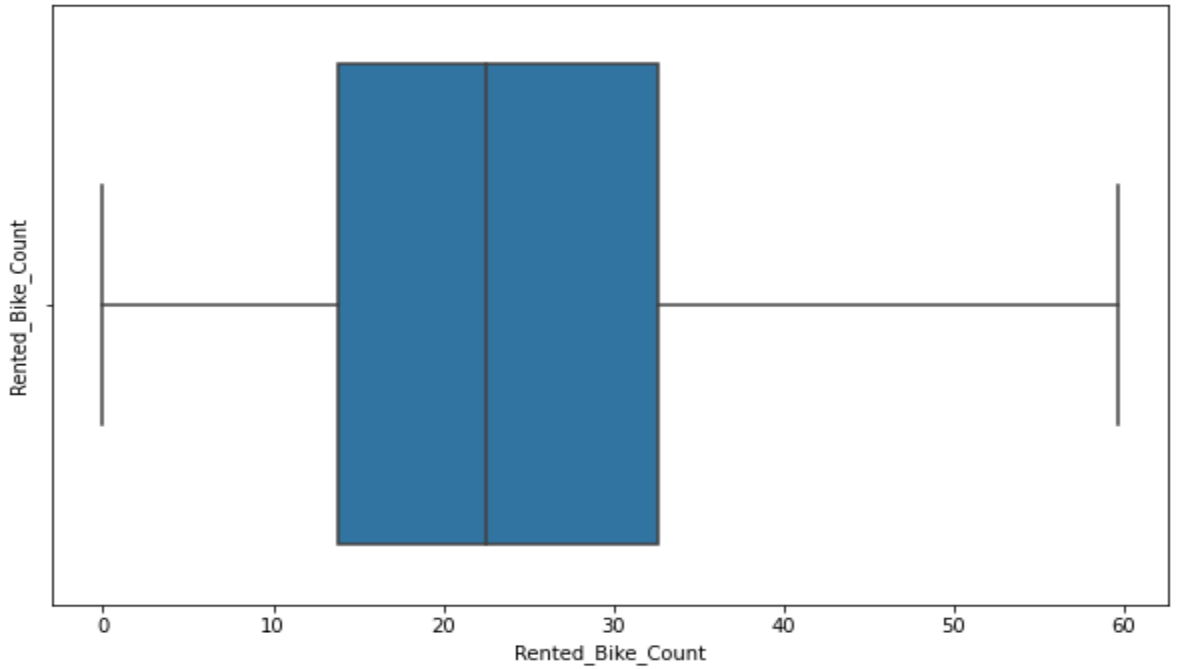




- ✓ **The above graph shows that Rented Bike Count has moderate right skewness.**
- ✓ **Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform some operation to make it normal.**
- ✓ **The above graph boxplot shows that we have detect outliers in rented bike .**

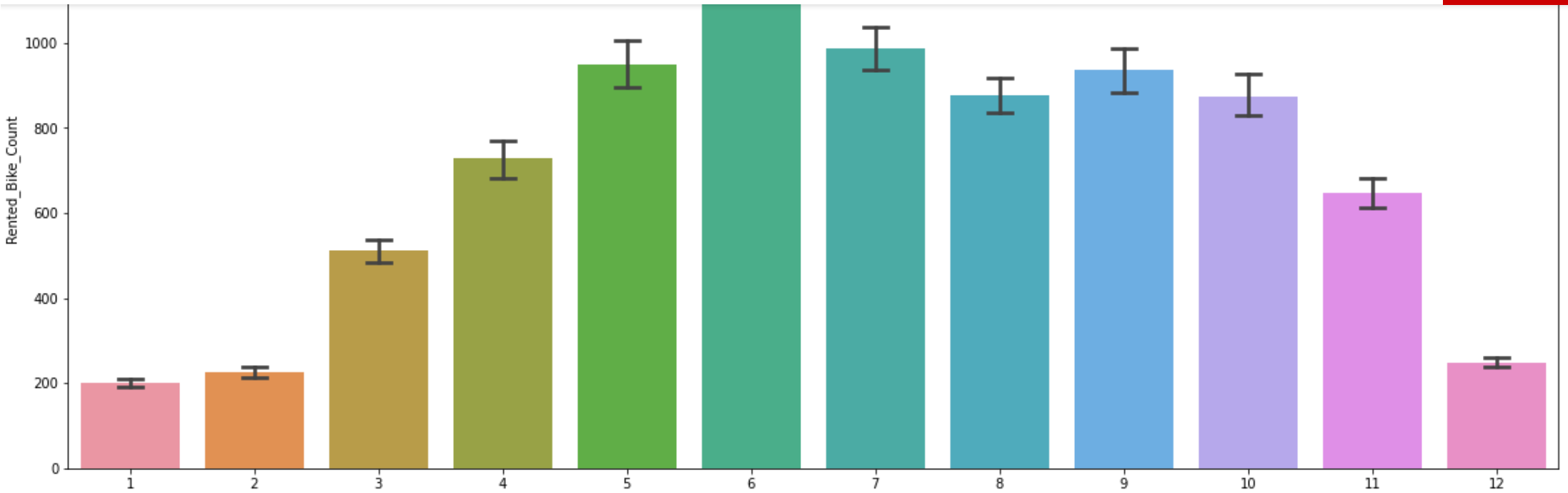
ANALYSIS OF RENTED BIKE COLUMN





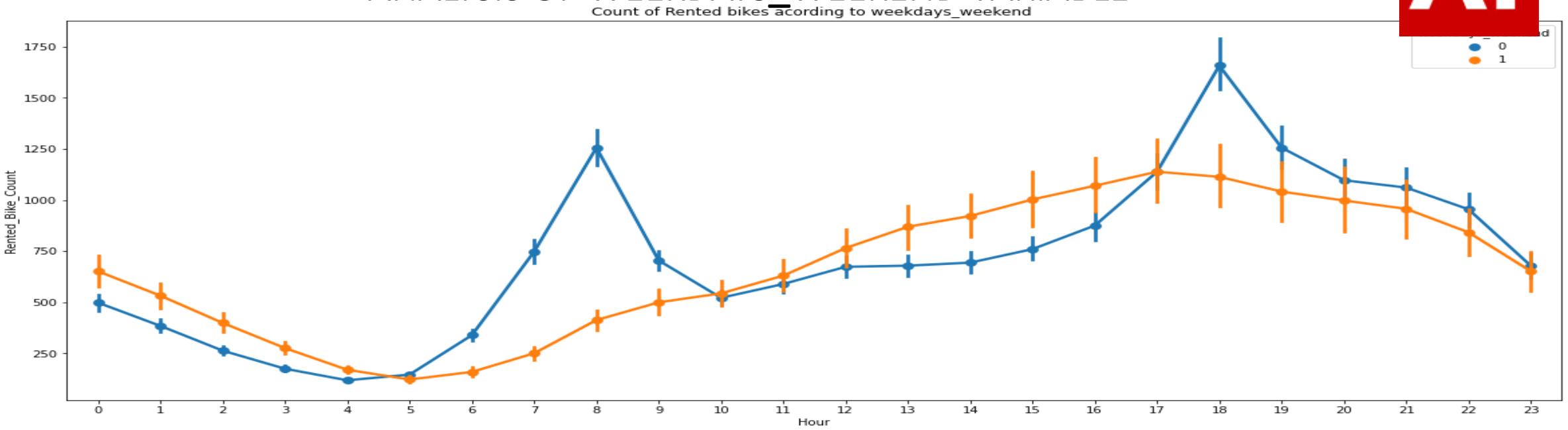
- ✓ **Since we have generic rule of applying Square root for the skewed variable in order to make it normal .After applying Square root to the skewed Rented Bike Count, here we get almost normal distribution.**
- ✓ **After applying Square root to the Rented Bike Count column, we find that there is no outliers present.**

ANALYSIS OF MONTH VARIABLE



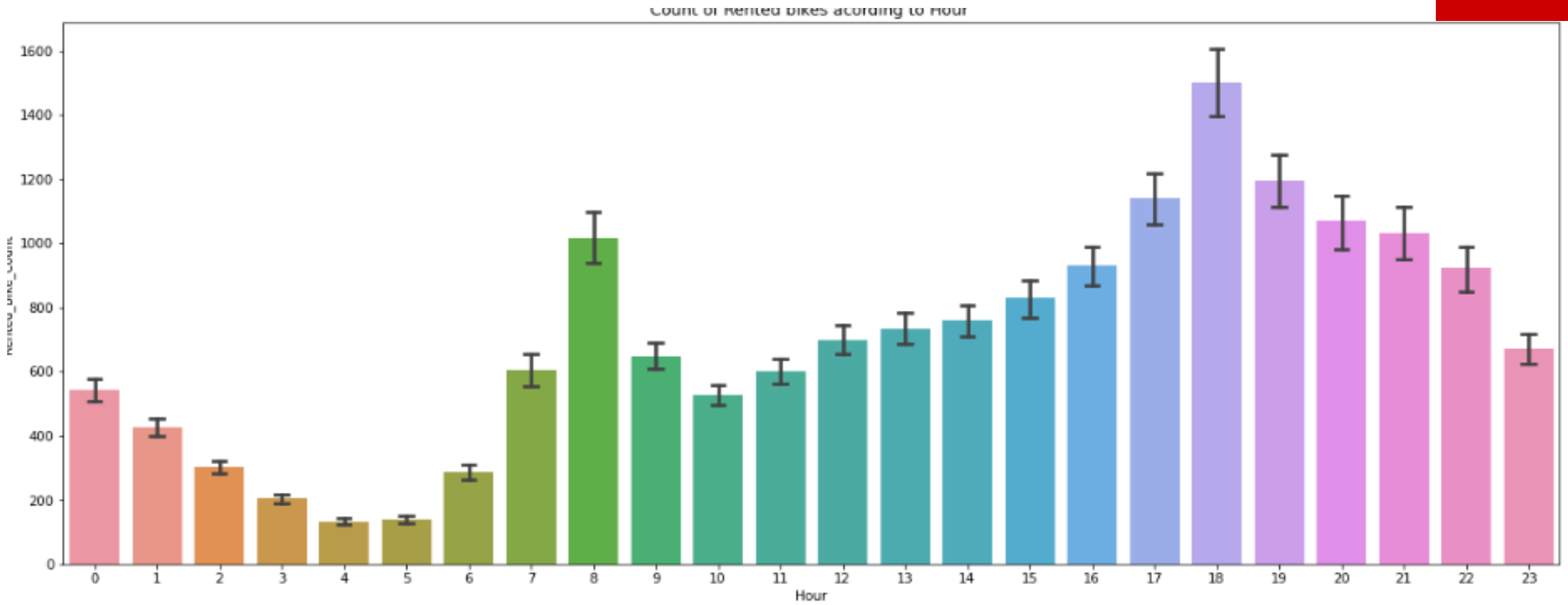
- ✓ From the above bar plot we can clearly say that from the month 5 to 10 the demand of the rented bike is high as compare to other months.these months are comes inside the summer season.

ANALYSIS OF WEEKDAYS_WEEKEND VARIABLE



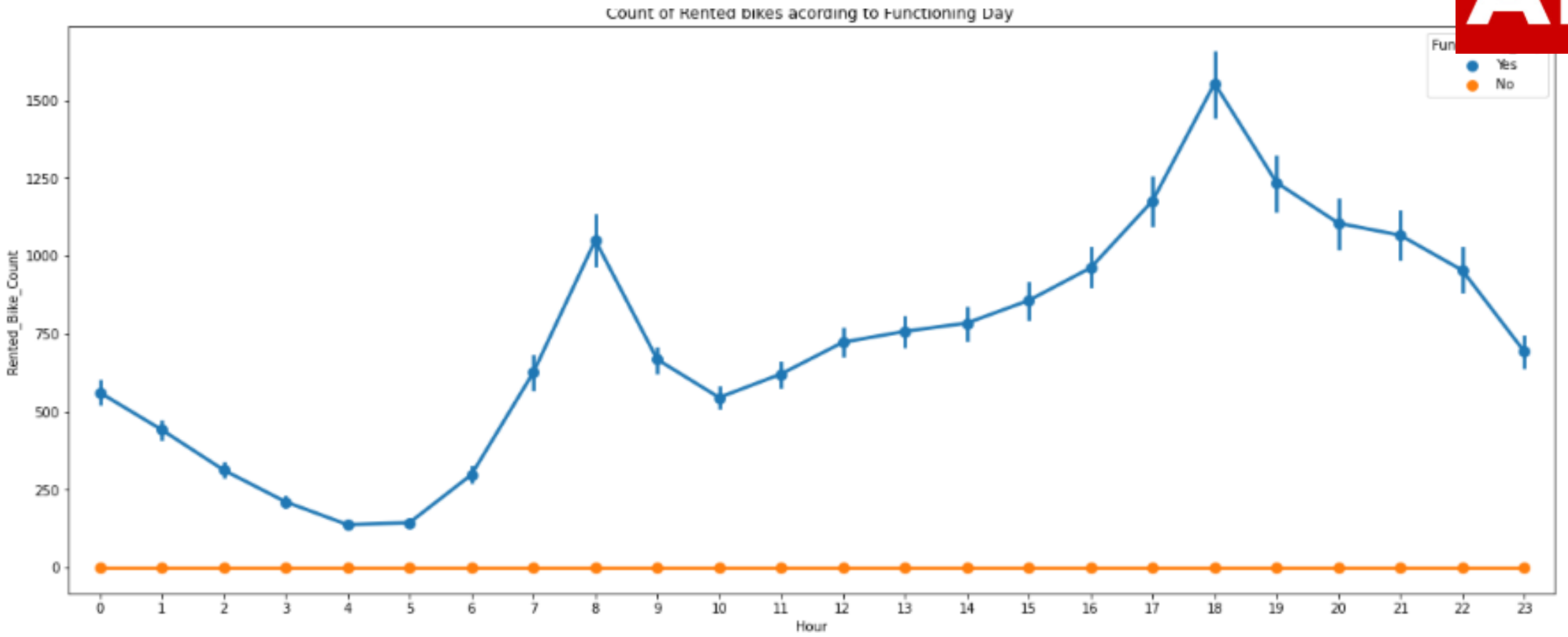
- ✓ **From the above point plot and bar plot we can say that in the week days which represent in blue color show that the demand of the bike higher because of the office.**
- ✓ **Peak Time are 7 am to 9 am and 5 pm to 7 pm**
- ✓ **The orange color represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.**

ANALYSIS OF HOUR VARIBLE



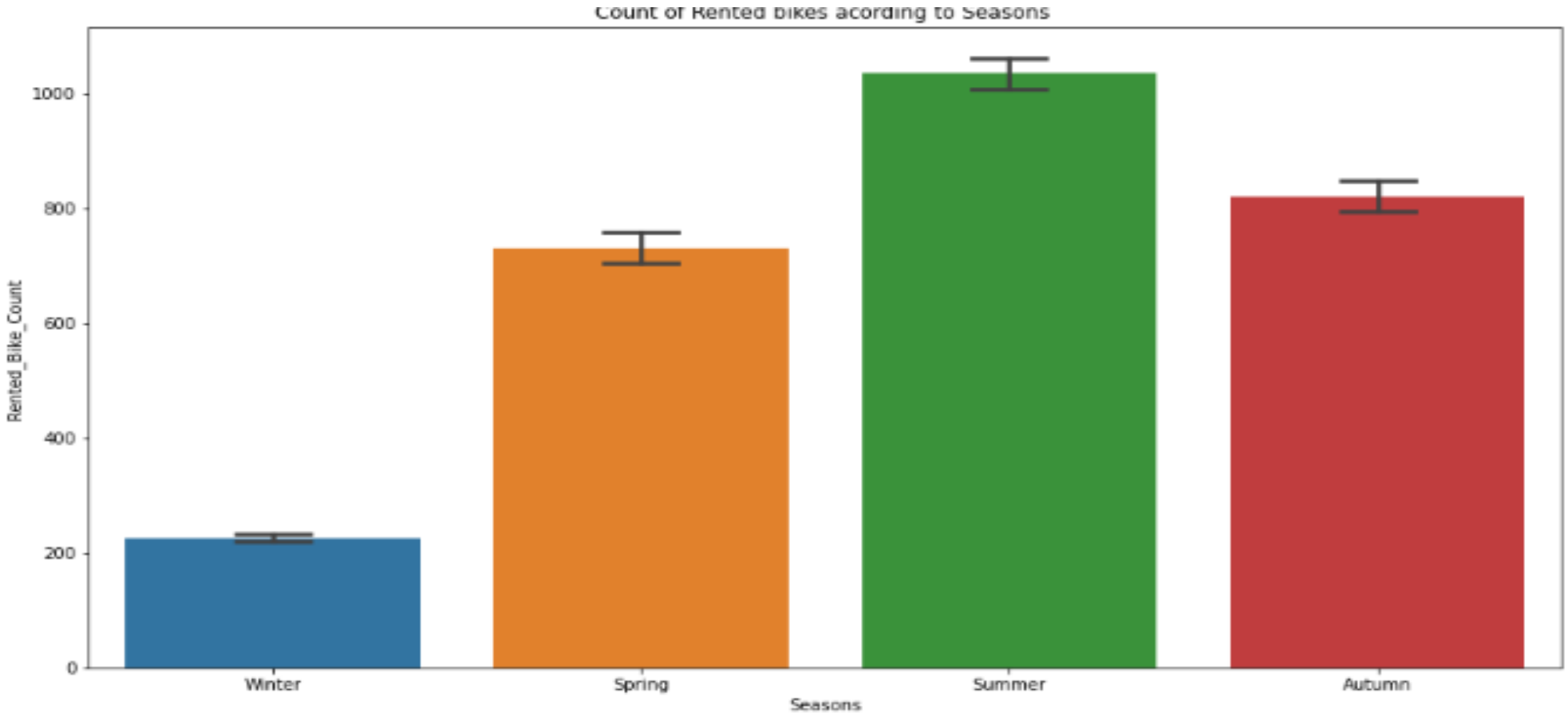
- ✓ In the above plot which shows the use of rented bike according the hours and the data are from all over the year.
- ✓ generally people use rented bikes during their working hour from 7am to 9am and 5pm to 7pm.

ANALYSIS OF FUNCTIONING DAY VARIABLE

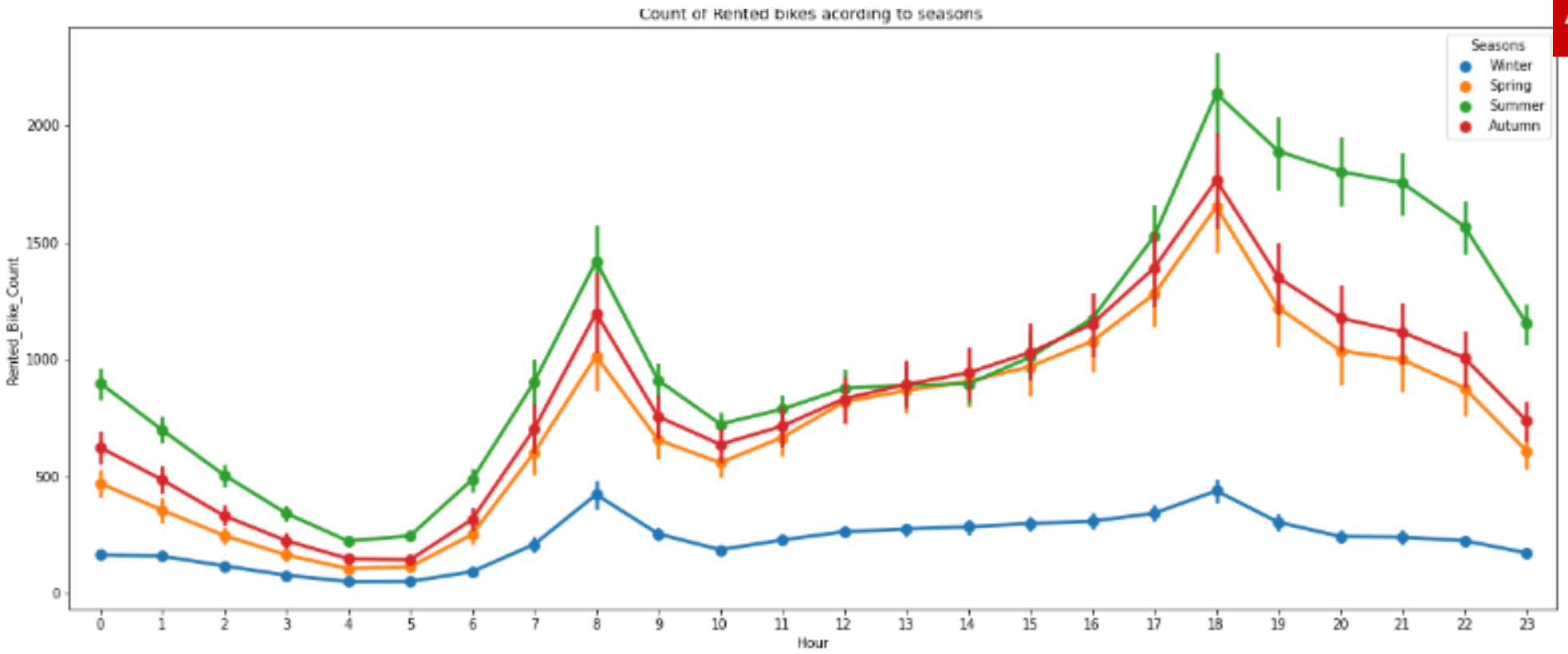


- ✓ **In the above bar plot and point plot which shows the use of rented bike in functioning days or not, and it clearly shows that,**
- ✓ **Peoples do not use rented bikes in no functioning day.**

ANALYSIS OF SEASON VARIABLE

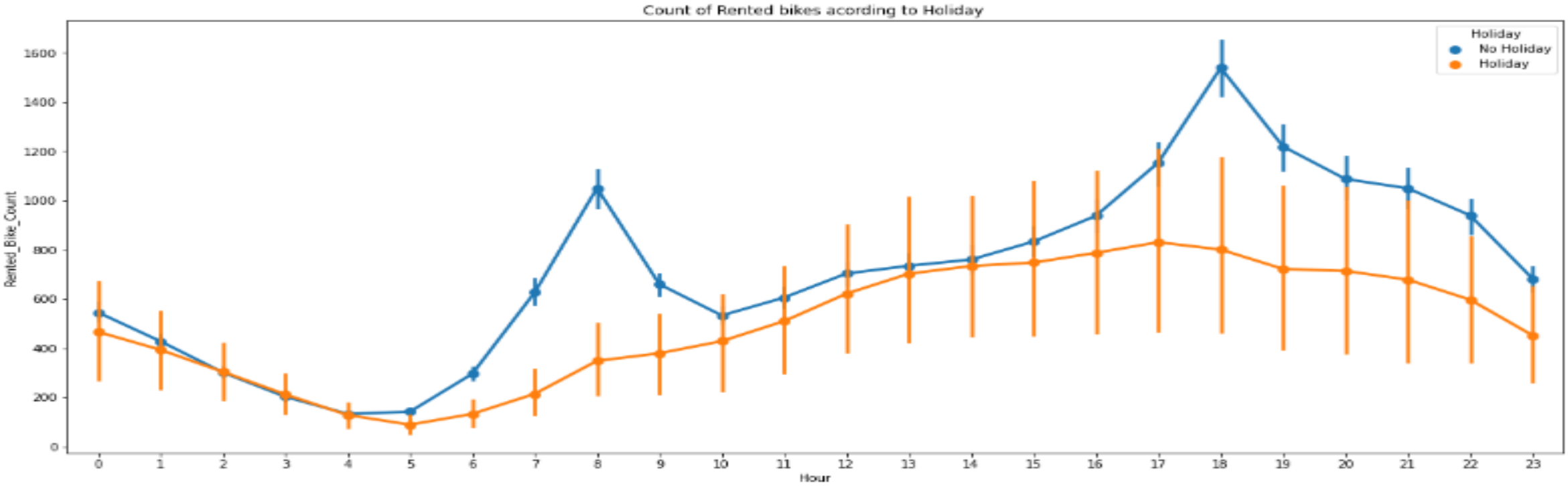


ANALYSIS OF SEASON VARIABLE



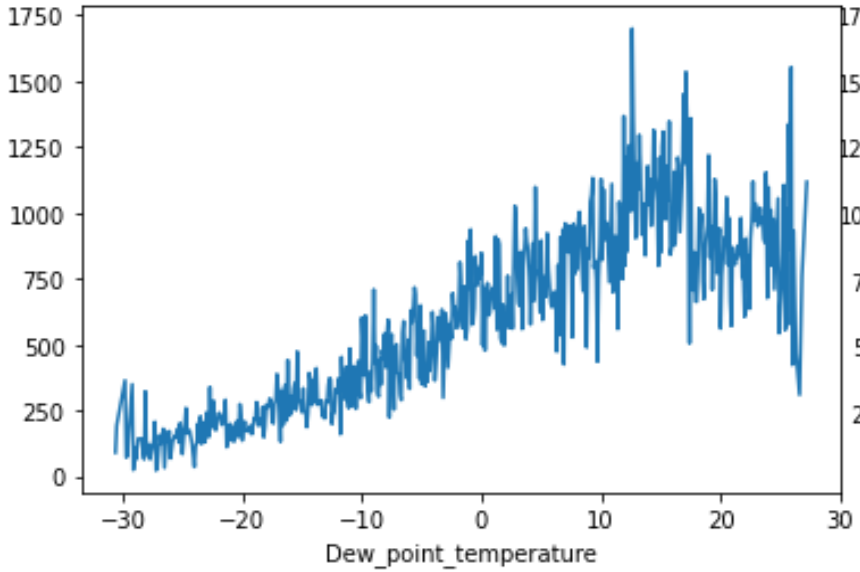
- ✓ **In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,**
- ✓ **in summer season the use of rented bike is high and peak time is 7am-9am and 7pm-5pm.**
- ✓ **In winter season the use of rented bike is very low because of snowfall.**

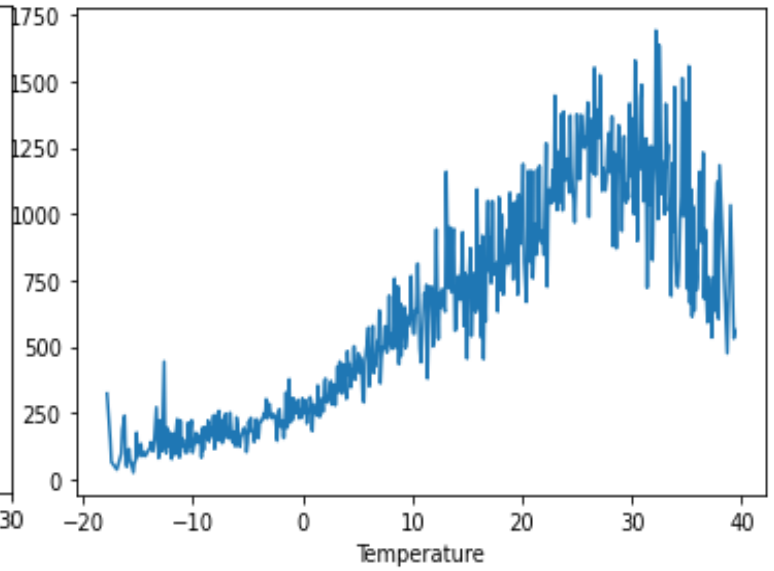
ANALYSIS OF HOLIDAY VARIABLE

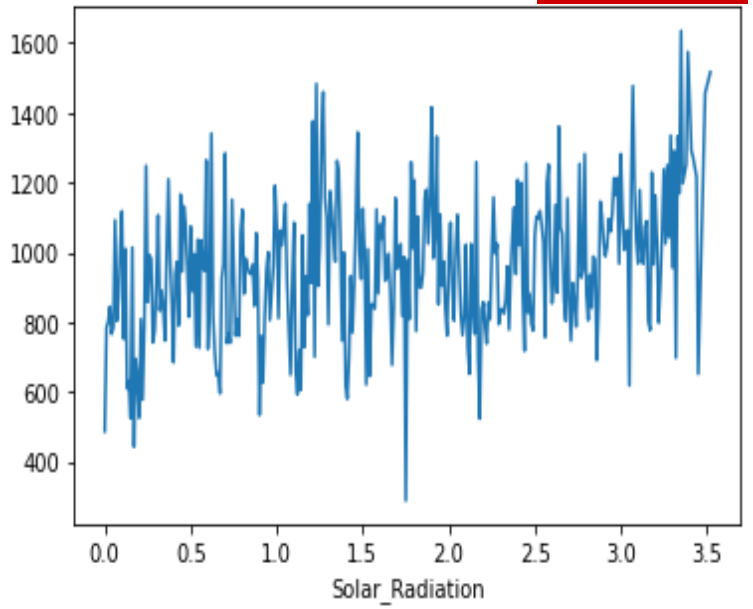


- ✓ **In the above bar plot and point plot which shows the use of rented bike in a holiday, and it clearly shows**
- ✓ **that,plot shows that in holiday people uses the rented bike from 2pm-8pm**

NUMERICAL AND RENTED BIKE COUNT

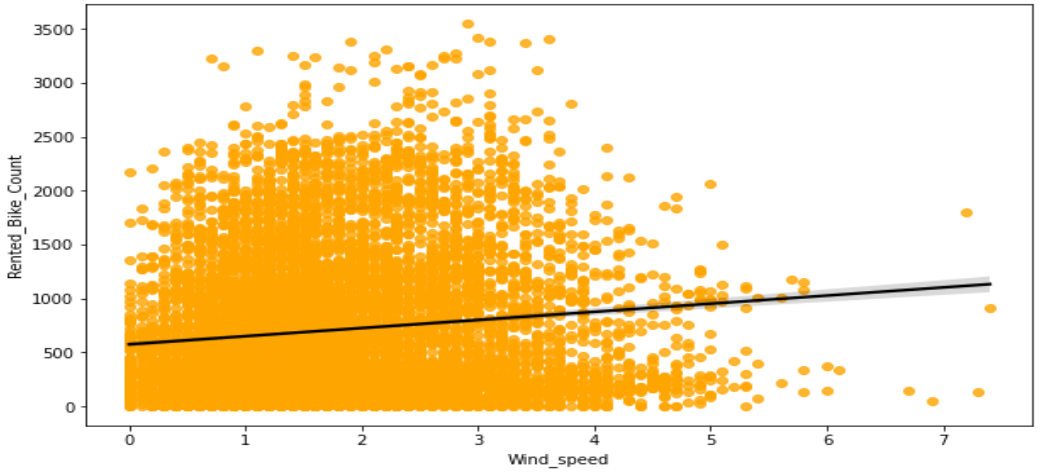


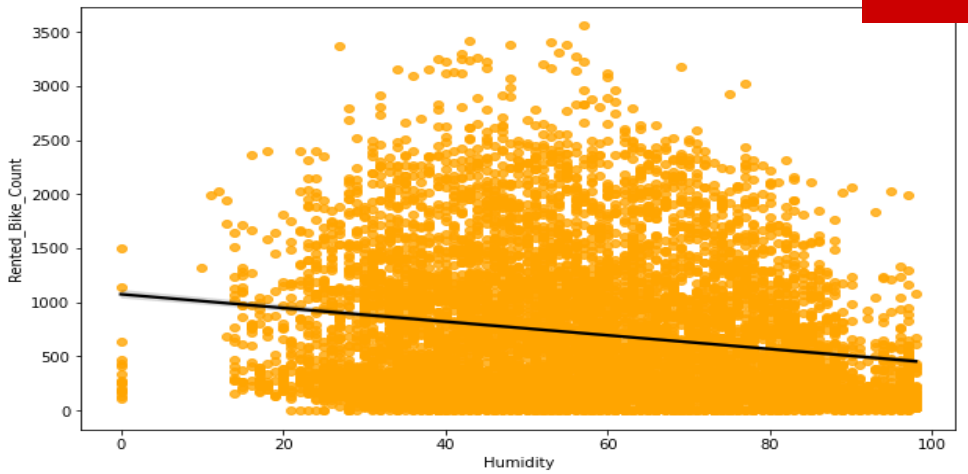


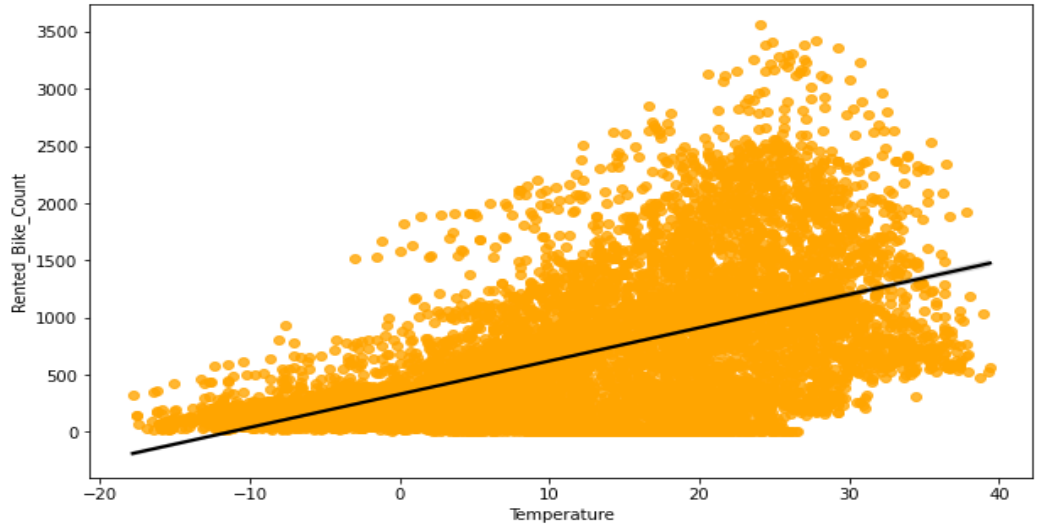


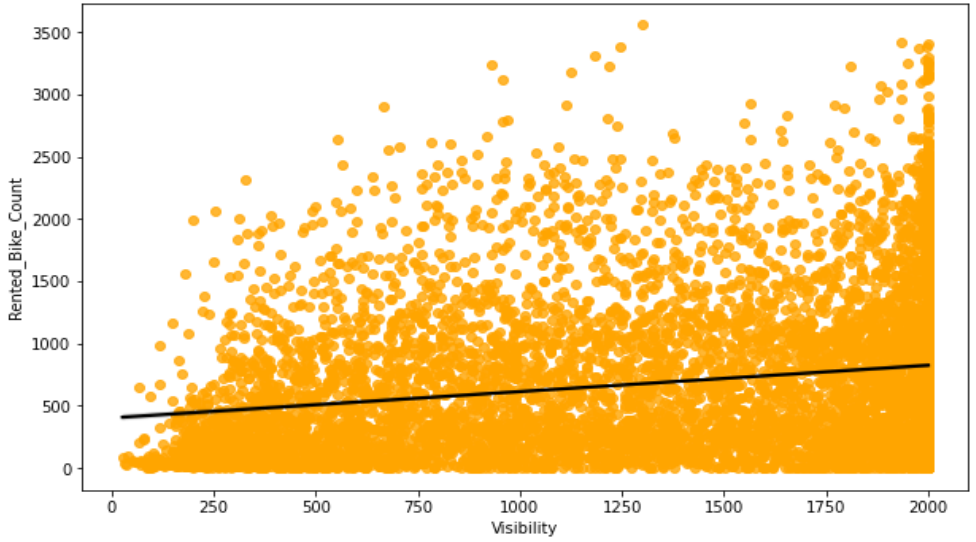
- ✓ **From the above plot we see that people like to ride bikes when it is pretty hot around 25°C in average.**
- ✓ **From the above plot of "Dew_point_temperature' is almost same as the 'temperature' there is some similarity present we can check it in our next step.**
- ✓ **from the above plot we see that, the amount of rented bikes is huge, when there is solar radiation, the counter of rents is around 1000.**

REGRESSION PLOT FOR NUMERICAL VARIABLE





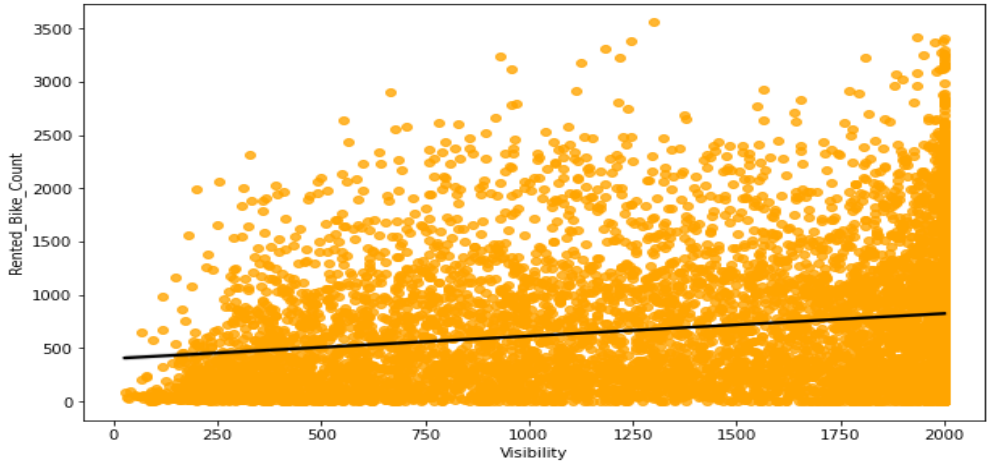


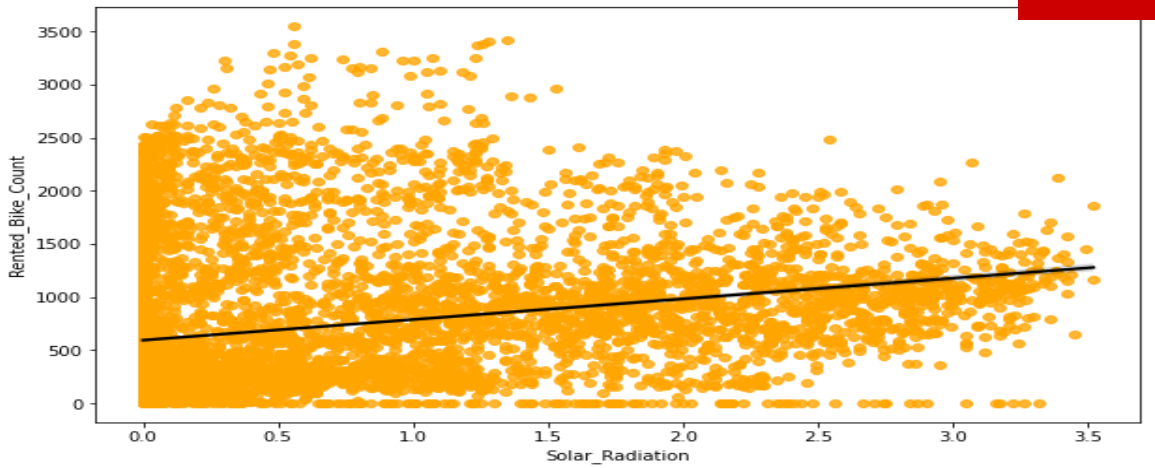


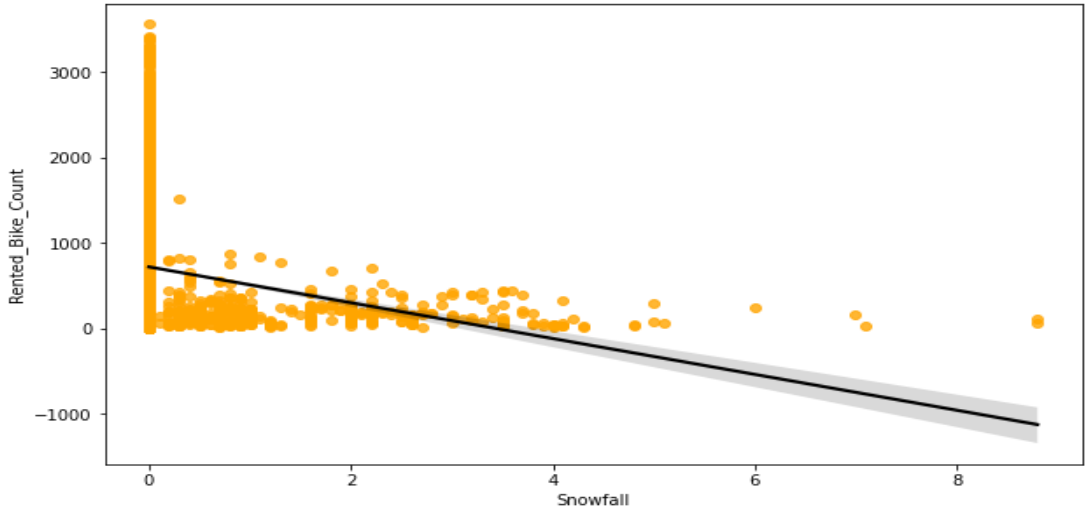
REGRESSION PLOT FOR NUMBERICAL VARIABLE

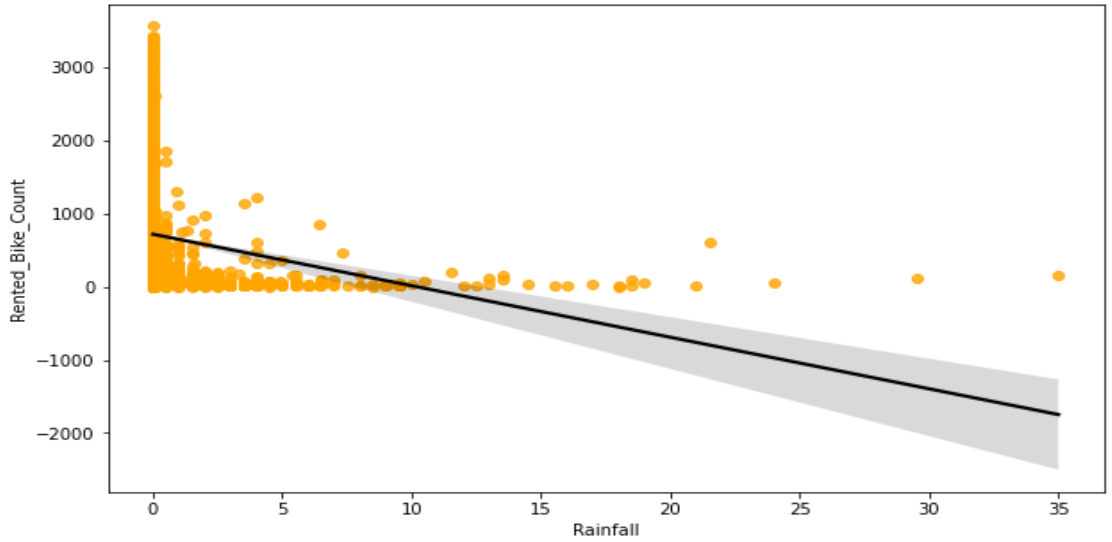
- From the above regression plot of all numerical feature we see that the column Temperature ’ , Wind_speed
- , Visibility, dew point temp, solar radiation are positively relation to the target variable.
- The Rented bike counted increases with increase of these features .
- Rainfall , snowfall , Humidity these features are neglected with the target variable means the rented bike counted decreases when these feature increase.

REGRESSION PLOT FOR NUMERICAL VARIABLE









OLS REGRESSION MODEL

- R square and adjacent square are near to each other .40% of variance on rented bike counted is explained by model.
- P value of dew point temp visibility are very high and they are not significant.

OLS Regression Results

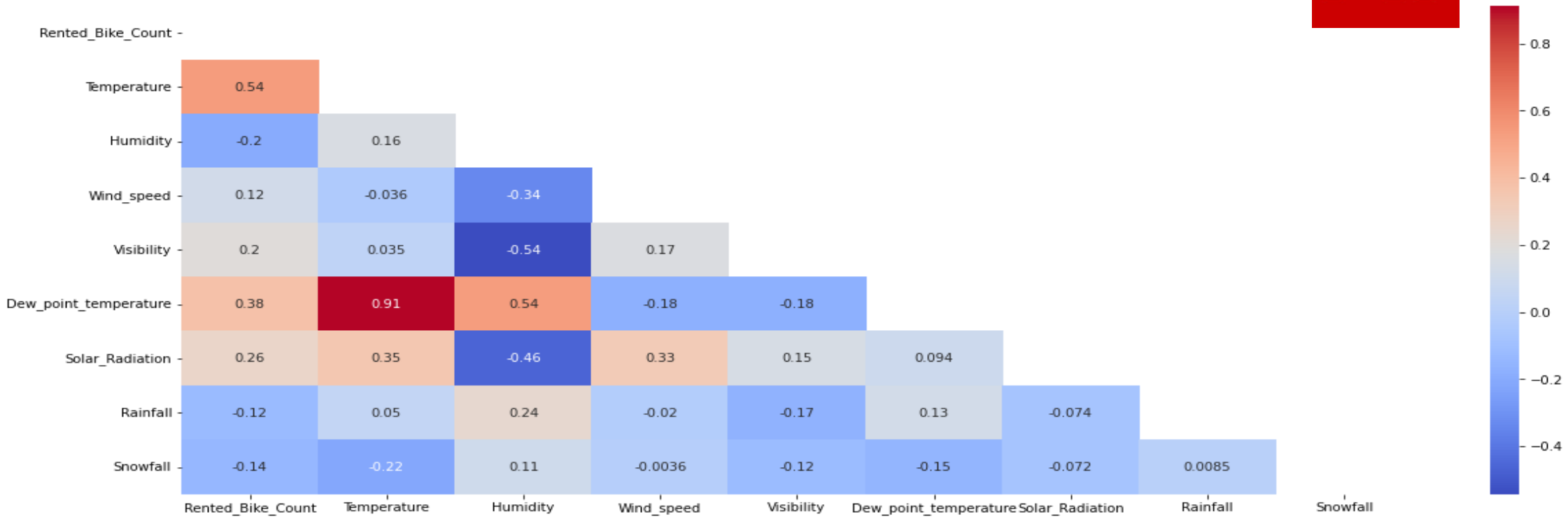
Dep. Variable:	Rented_Bike_Count	R-squared:	0.398
Model:	OLS	Adj. R-squared:	0.397
Method:	Least Squares	F-statistic:	723.1
Date:	Sun, 30 Jan 2022	Prob (F-statistic):	0.00
Time:	02:23:02	Log-Likelihood:	-66877.
No. Observations:	8760	AIC:	1.338e+05
Df Residuals:	8751	BIC:	1.338e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	844.6495	106.296	7.946	0.000	636.285	1053.014
Temperature	36.5270	4.169	8.762	0.000	28.355	44.699
Humidity	-10.5077	1.184	-8.872	0.000	-12.829	-8.186
Wind_speed	52.4810	5.661	9.271	0.000	41.385	63.577
Visibility	-0.0097	0.011	-0.886	0.376	-0.031	0.012
Dew_point_temperature	-0.7829	4.402	-0.178	0.859	-9.411	7.846
Solar_Radiation	-118.9772	8.670	-13.724	0.000	-135.971	-101.983
Rainfall	-50.7083	4.932	-10.282	0.000	-60.376	-41.041
Snowfall	41.0307	12.806	3.204	0.001	15.929	66.133

Omnibus:	957.371	Durbin-Watson:	0.338
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1591.019
Skew:	0.769	Prob(JB):	0.00
Kurtosis:	4.412	Cond. No.	3.11e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.11e+04. This might indicate that there are strong multicollinearity or other numerical problems.

CORRELATION MATRIX



- Variable like dew point temperature and temperature are correlated.

MODEL BUILDING

- LINEAR REGRESSION
- LASSO
- RIDGE
- DECISION TREE
- RANDOM FOREST
- GRADIENT BOOSTED
- GRAIDENT BOSSTING WITH GRIDSEARCHCV

LINEAR REGRESSION

TRAIN SET RESULT

MSE : 35.07751288189293

RMSE : 5.9226271942350825

MAE : 4.474024092996787

R2 : 0.7722101548255267

Adjusted R2 : 0.7672119649454145

TEST SET RESULT

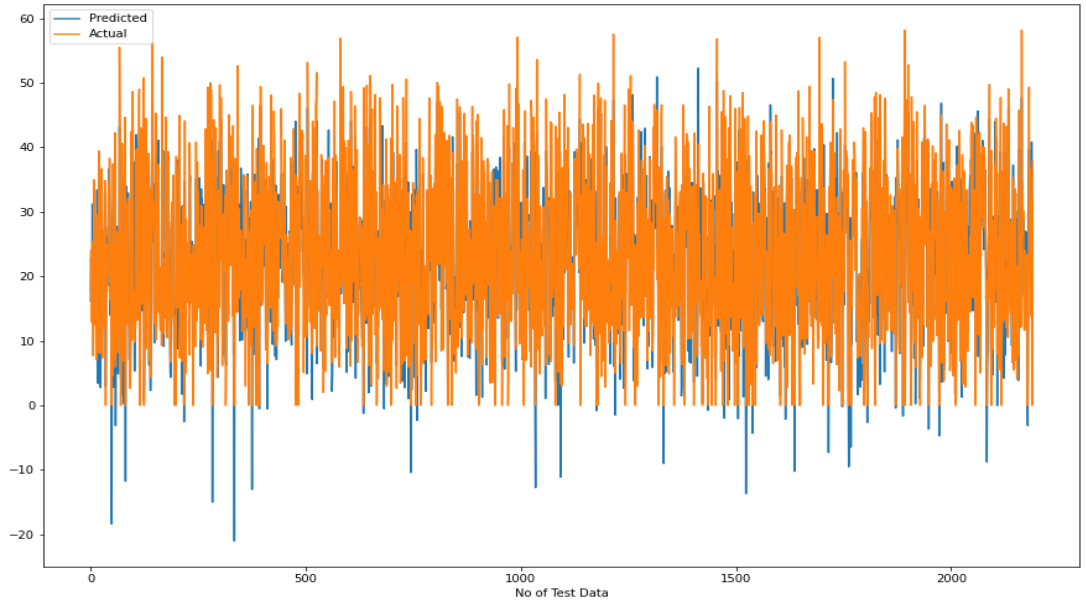
MSE : 33.27533089591926

RMSE : 5.76847734639907

MAE : 4.410178475318181

R2 : 0.7893518482962683

Adjusted R2 : 0.7847297833429184



DECISION TREE

TRAIN SET RESULT

MSE : 71.30709145763139

RMSE : 8.44435263697765

MAE : 6.260322435441764

R2 : 0.5485933087817594

Adjusted R2 : 0.5386884934282312

TEST SET RESULT

Model Score: 0.5620087311452759

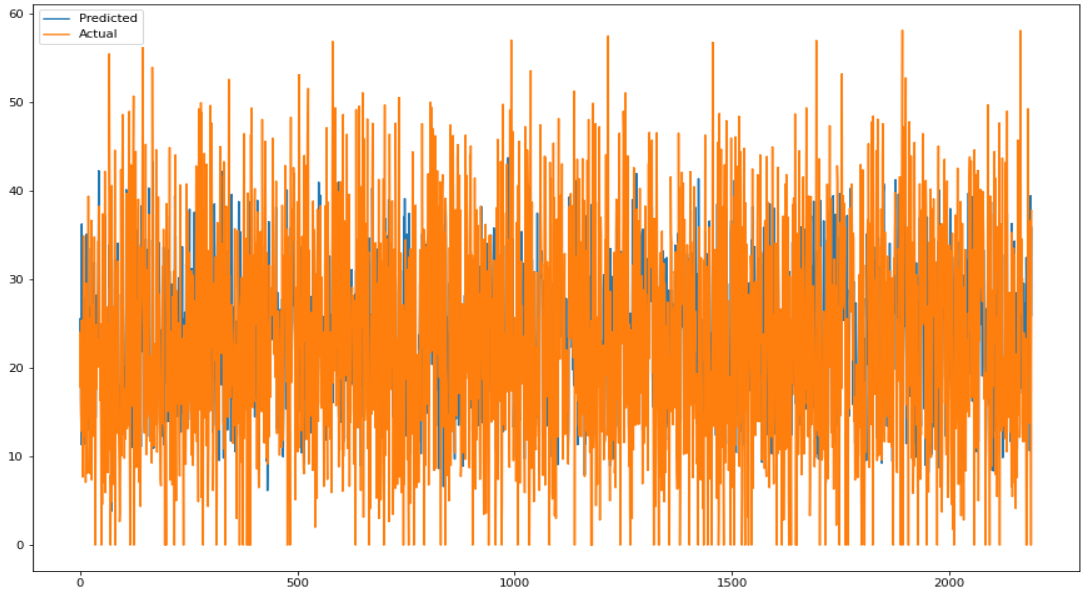
MSE : 67.44657279889095

RMSE : 8.212586243010843

MAE : 6.0732199531973015

R2 : 0.5620087311452759

Adjusted R2 : 0.5523982784673245



LASSO REGRESSION

TRAIN SET RESULT

MSE : 91.59423336097032
RMSE : 9.570487623991283
MAE : 7.25504157145495 2
R2 : 0.40519624904934015
Adjusted R2 : 0.3921449996120475

TEST SET RESULT

MSE : 96.7750714044618
RMSE : 9.837432155011886
MAE : 7.455895061963607
R2 : 0.3873692800799008
Adjusted R2 : 0.37392686932535146

RIDGE REGRESSION

TRAIN SET RESULT

MSE : 35.07752456136463
RMSE : 5.922628180239296
MAE : 4.474125776125378
R2 : 0.7722100789802107
Adjusted R2 : 0.7672118874358922

TEST SET RESULT

MSE : 33.27678426818438
RMSE : 5.768603320404722
MAE : 4.410414932539515
R2 : 0.7893426477812578
Adjusted R2 : 0.7847203809491939

ELASTIC NET REGRESSION

TRAIN SET RESULT

MSE : 57.5742035398887
RMSE : 7.587766703048315
MAE : 5.792276538970546
R2 : 0.6261189054494012
Adjusted R2 : 0.6179151652795234

TEST SET RESULT

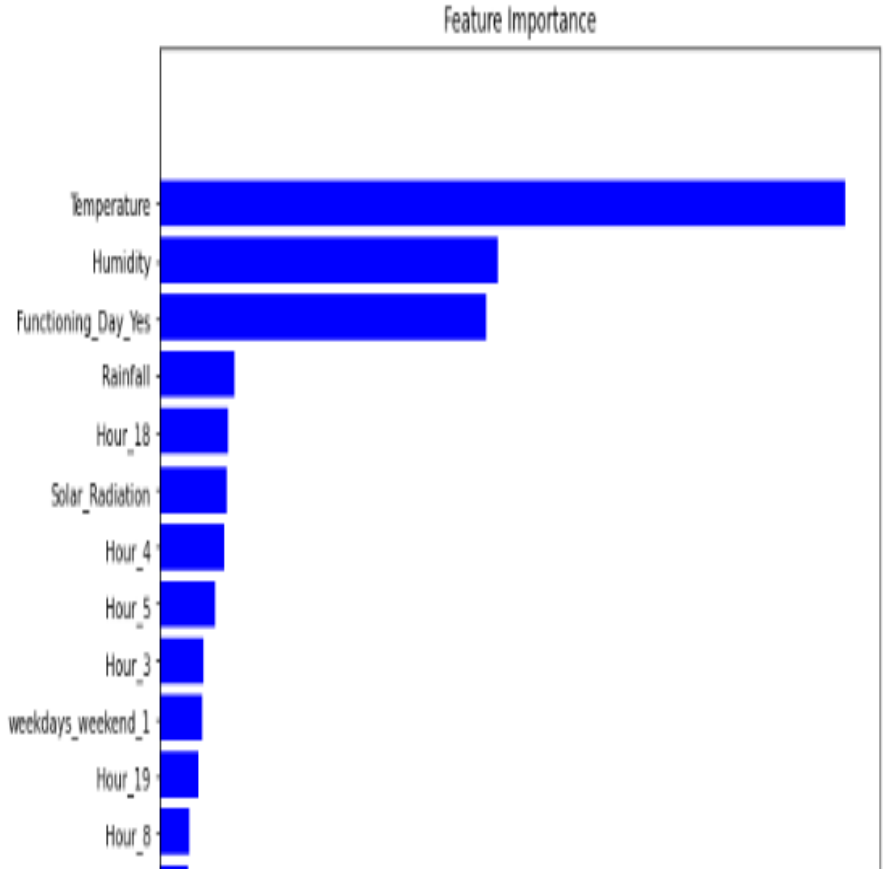
MSE : 59.45120536350042
RMSE : 7.710460775044538
MAE : 5.873612334800099
R2 : 0.6236465216363589
Adjusted R2 : 0.6153885321484546

RANDOM FOREST

```
Model Score: 0.9896599224004005
MSE : 1.5922755683946939
RMSE : 1.2618540202395419
MAE : 0.805947919494939
R2 : 0.9896599224004005
Adjusted R2 : 0.9894330392784672
```

```
MSE : 12.687827551655872
RMSE : 3.561997691135674
MAE : 2.2017468429561533
R2 : 0.9196802150141876
Adjusted R2 : 0.9179178294426035
```

	Feature	Feature Importance
0	Temperature	0.31
1	Humidity	0.16
34	Functioning_Day_Yes	0.15
10	Hour_4	0.03
4	Solar_Radiation	0.03
5	Rainfall	0.03
24	Hour_18	0.03
25	Hour_19	0.02
11	Hour_5	0.02
46	weekdays_weekend_1	0.02
9	Hour_3	0.02



GRADIENT BOOSTING REGRESSION WITH GRIDSEARCHCV

TRAIN TEST RESULT

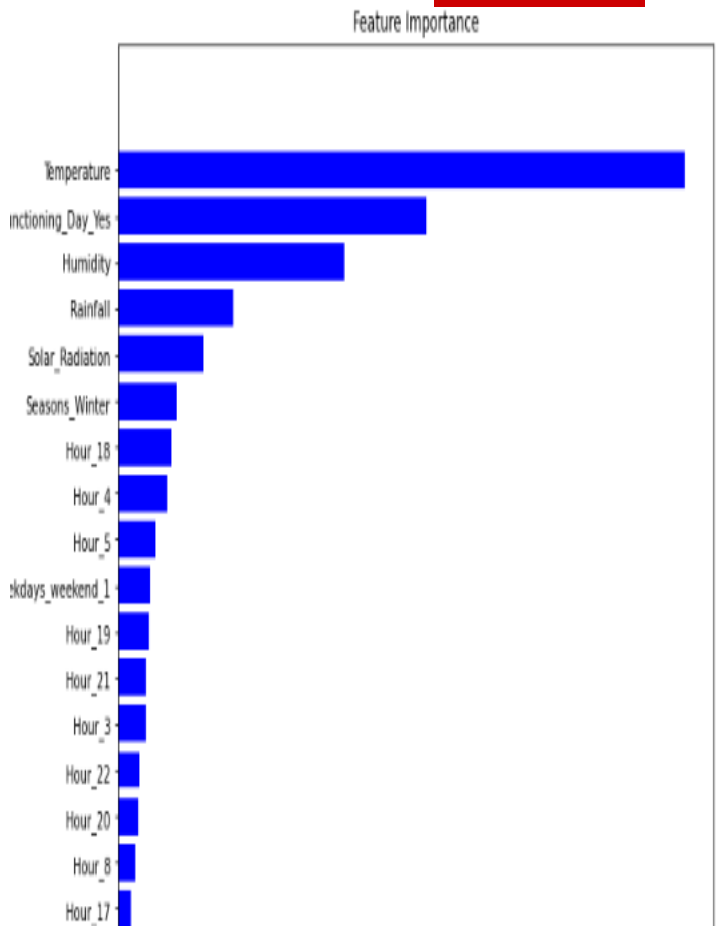
```
Model Score: 0.9515896672300013
MSE : 7.454740004128374
RMSE : 2.7303369762958516
MAE : 1.8489194833919358
R2 : 0.9515896672300013
Adjusted R2 : 0.9505274423746372
```

TEST SET RESULT

```
MSE : 12.392760556291103
RMSE : 3.520335290322657
MAE : 2.4005915565405354
R2 : 0.921548124829924
Adjusted R2 : 0.9198267251413182
```

	Feature	Feature Importance
0	Temperature	0.32
34	Functioning_Day_Yes	0.17
1	Humidity	0.13
5	Rainfall	0.07
4	Solar_Radiation	0.05

A hyper parameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyper parameter tuning.



CHALLANGES

- Large dataset to handle
- Needs to plot lot of graphs to analyse.
- Feature engineering
- Feature selection
- Optimising the model
- Carefully tuned Hyper parameters as it affect.

CONCLUSION

- ‘Hour’ of the day holds the most important feature.
- Bike rental count is more correlated with the time of the day as it is peak at 10 am morning and 8 pm in evening.
- We observed that bike rental count high during working days than the non working days.
- We see the people generally prefer to bikes at moderate to high temp and when little windy.

CONCLUSION

When we compare the root mean squared error and mean absolute error of all the models, random forest regressor and gradient boosting gridsearchcv gives the highest R2 score of 99% and 95% for train set and 92% for test set. So finally this model is best for predicating the bike rental count on daily basis.

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Dicision tree regression	6.073	67.447	8.213	0.562	0.55
	5	Random forest regression	0.806	1.592	1.262	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95

Test set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Dicision tree regression	6.260	71.307	8.444	0.549	0.54
	5	Random forest regression	2.202	12.688	3.562	0.920	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92