CAPSTONE PROJECT

Bike Sharing Demand Prediction

By

Auti Rupali .

CONTENT

BUSINESS UNDERSTANDING

Bike rentals have became a popular service in recent years 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 locations, to fulfil the demand.

Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

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
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

This Dataset contains 8760 lines and 14 columns.

Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.

One Datetime features ‘Date’.

We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.

FEATURE SUMMARY

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/m2

Rainfall -mm

Snowfall -cm

Seasons -Winter, Spring, Summer, Autumn

Holiday -Holiday/No Holiday

Functional Day - NoFunc(Non Functional Hrs),Fun(Functional Hrs)

INSIGHTS FROM OUR DATASET

There are No Missing Values present

There are No Duplicate values present

There are No null values.

And finally we have 'rented bike count' variable which we need to predict for new observations

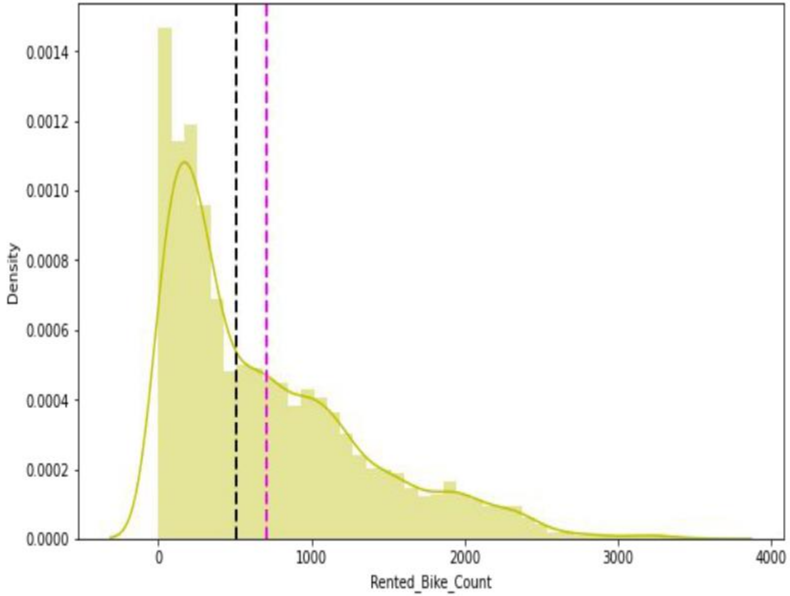
The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data

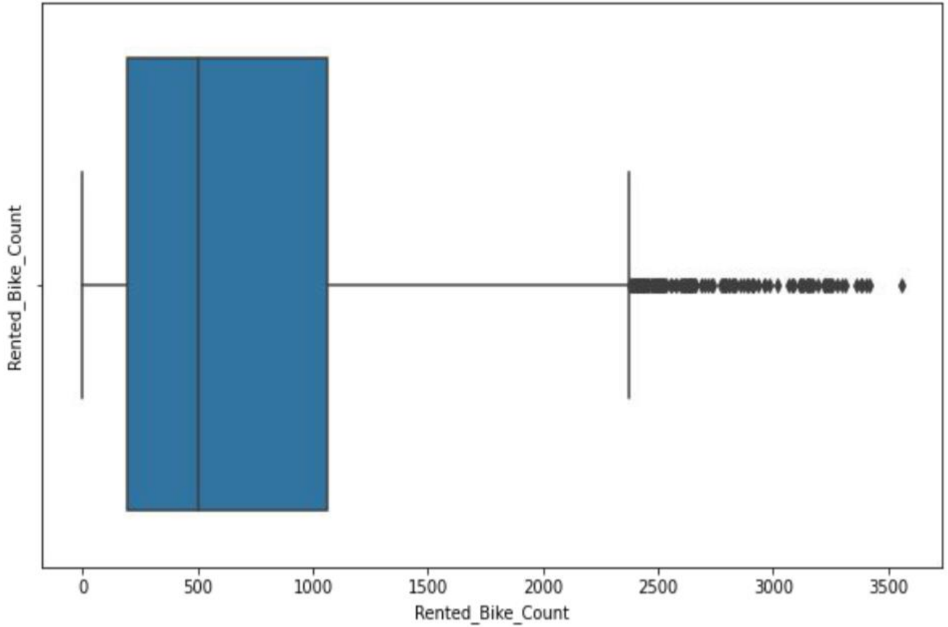
So we convert the "date" column into 3 different column i.e "year","month","day".

We change the name of some features for our convenience , they are as below

'Rented_Bike_Count', 'Hour', 'Temperature', 'Humidity', 'Wind_speed', 'Visibility', ‘Dew_point_temperature', 'Solar_Radiation', 'Rainfall', 'Snowfall', 'Seasons', 'Holiday', 'Functioning_Day', 'month','weekdays_weekend'

ANALYSIS OF RENTED BIKE COLUMN



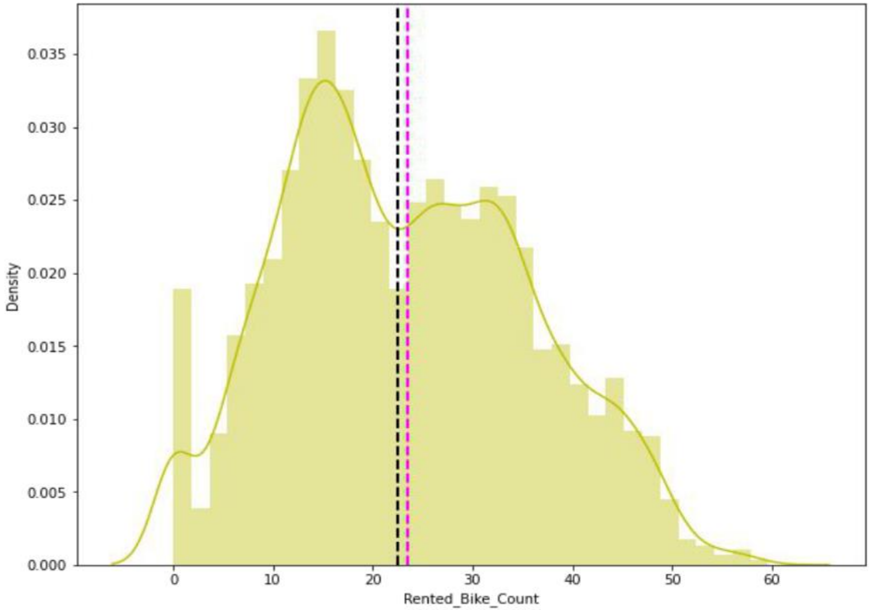


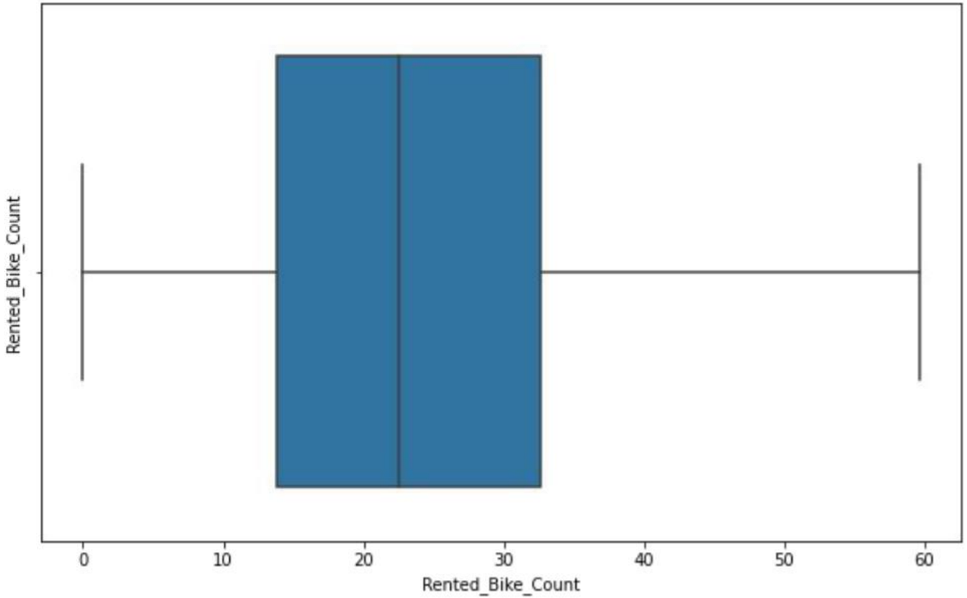
The above graph shows that Rented Bike Count has moderate right skewness.

The above boxplot shows that we have detect outliers in Rented Bike Count column

Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform Square root operation to make it normal

ANALYSIS OF RENTED BIKE COLUMN

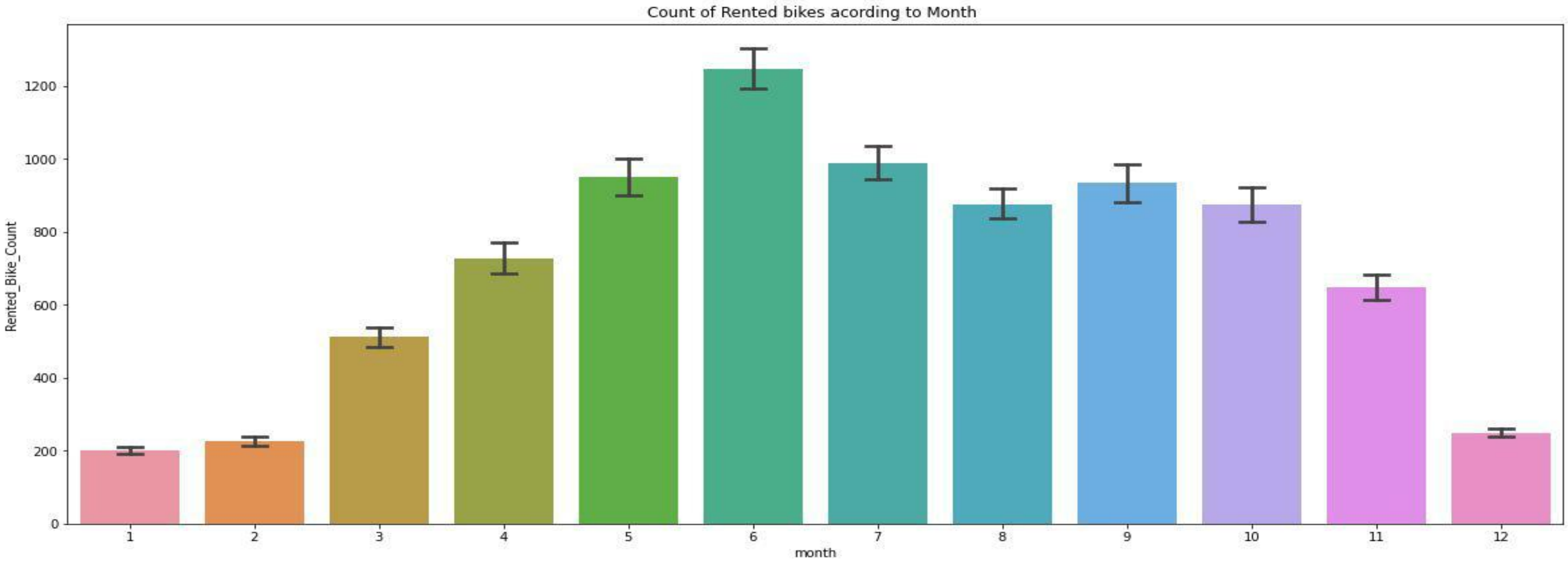




After applying Square root to the skewed Rented Bike Count, here we get almost normal distribution.

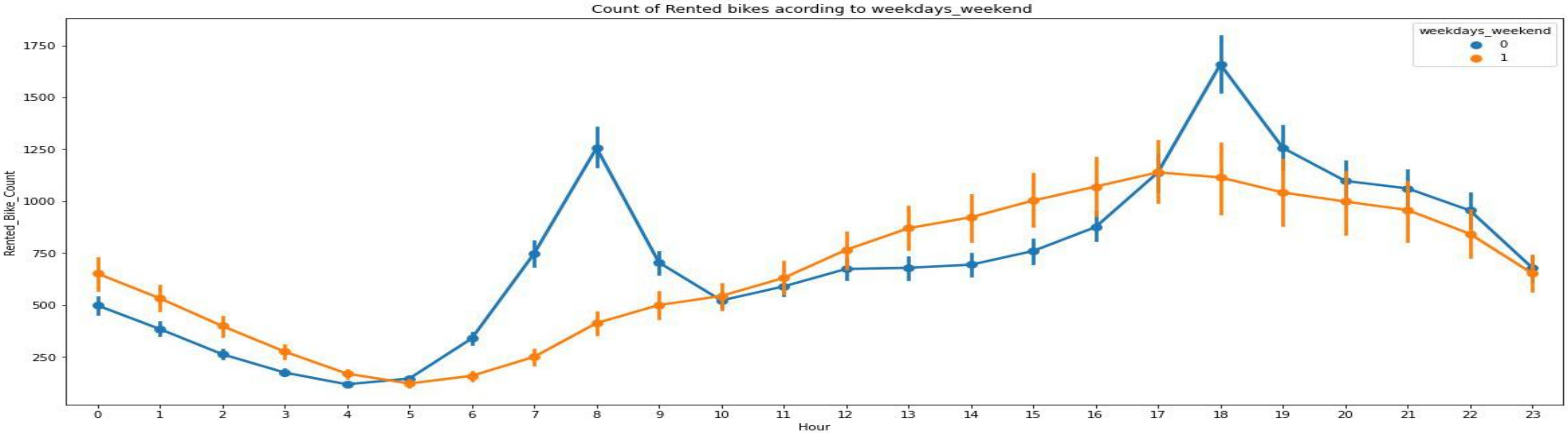
After applying Square root to the Rented Bike Count column, we find that there is no outliers present

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

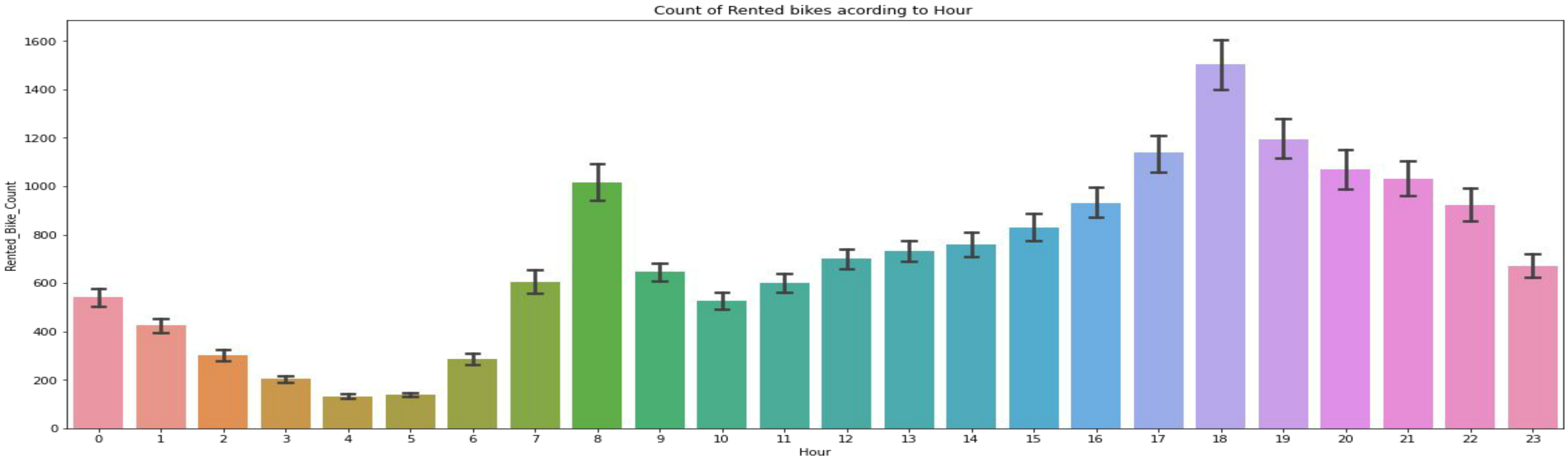


From the above point plot and bar plot we can say that in the weekdays which represent in blue colour show that the demand of the bike higher because of the office.

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 especially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

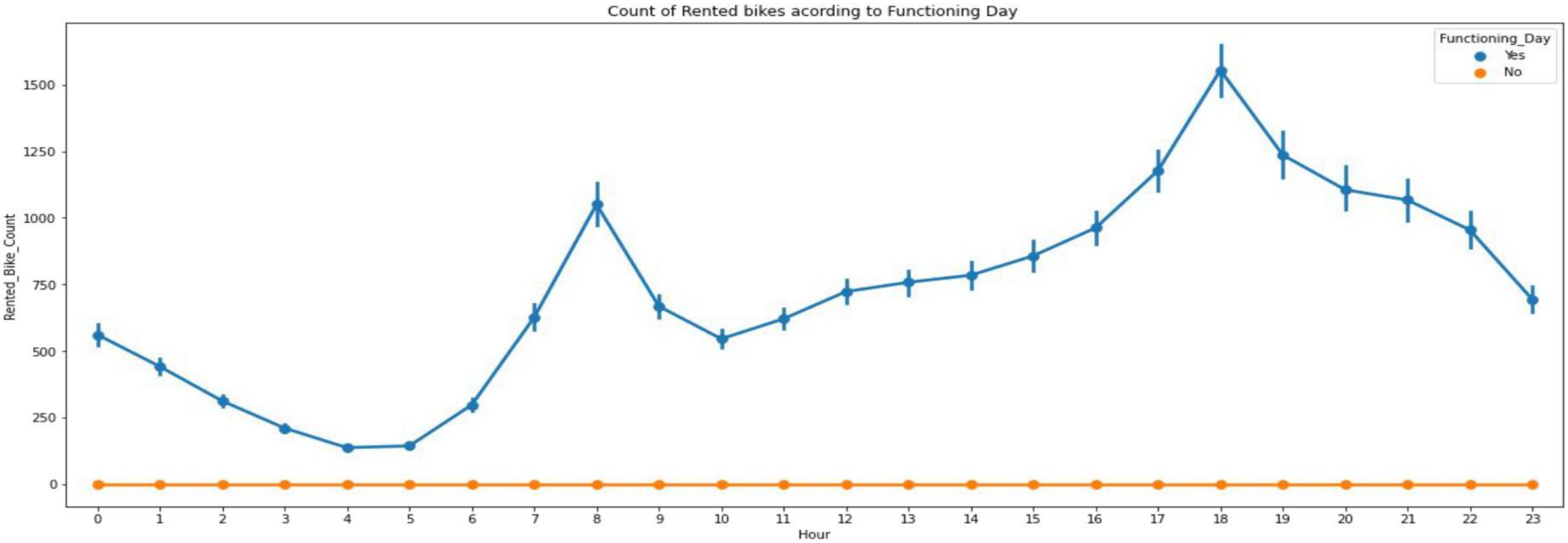
ANALYSIS OF HOUR VARIABLE



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.

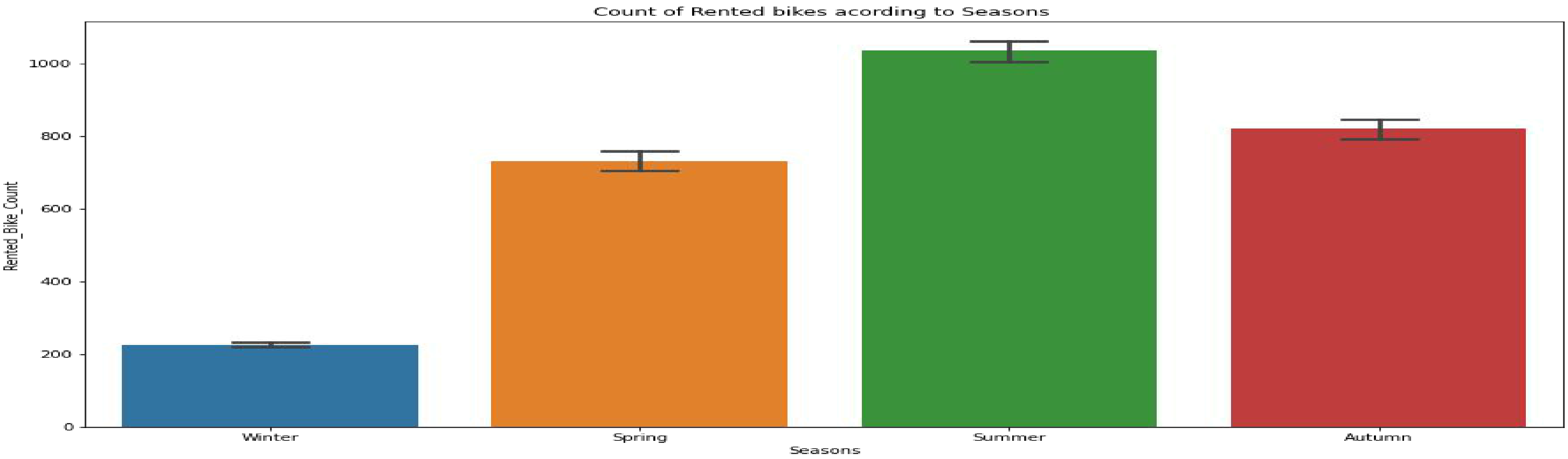
ANALYSIS OF FUNCTIONING DAY VARIABLE



In the above point plot which shows the use of rented bike in functioning daya or not, and it clearly shows that,

Peoples dont use rented bikes in no functioning day.

ANALYSIS OF SEASON VARIABLE

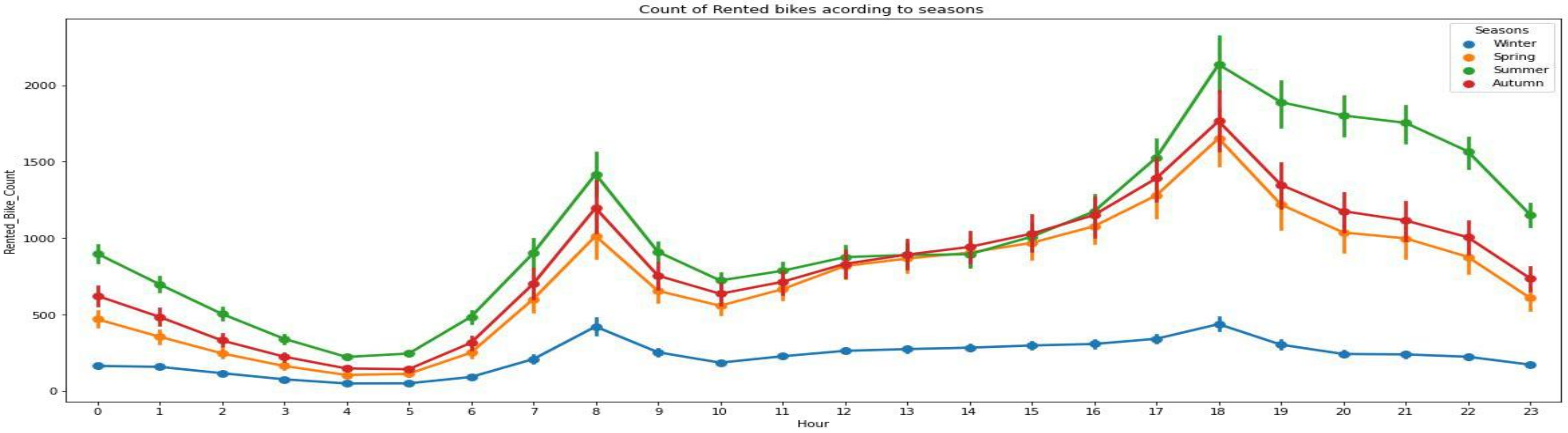


This above bar plot shows the distribution of rented bike count season wise

And we can clearly see that that peoples love to ride bike in summer seasons and autumn season

But in winter season people don't take any rented bike due to because of snowfall

ANALYSIS OF SEASON VARIABLE

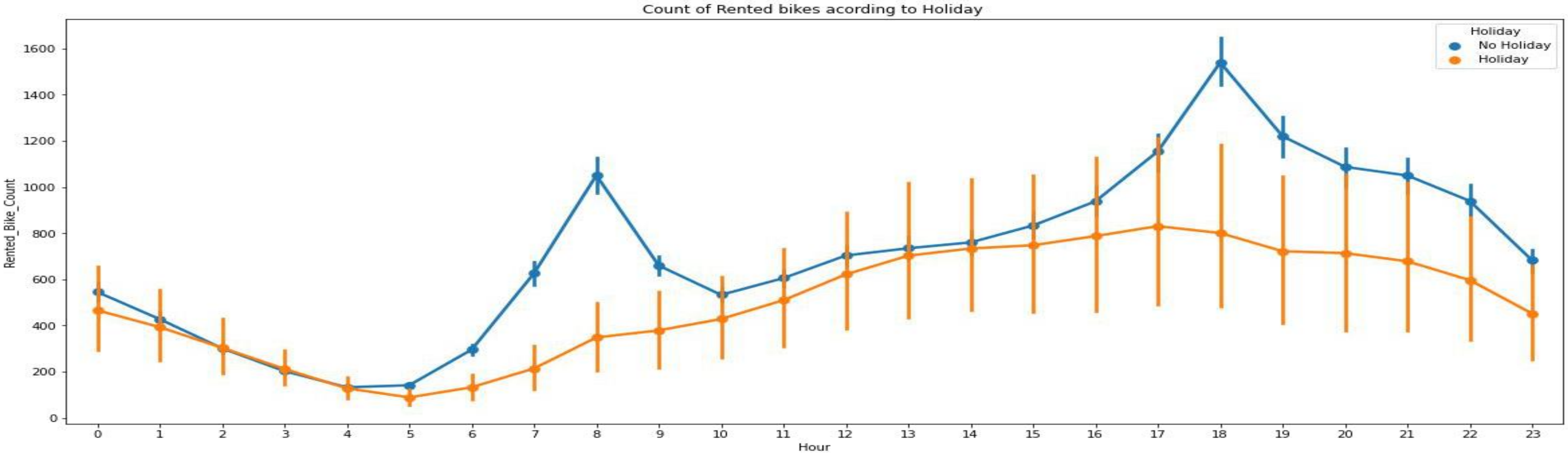


In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,

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

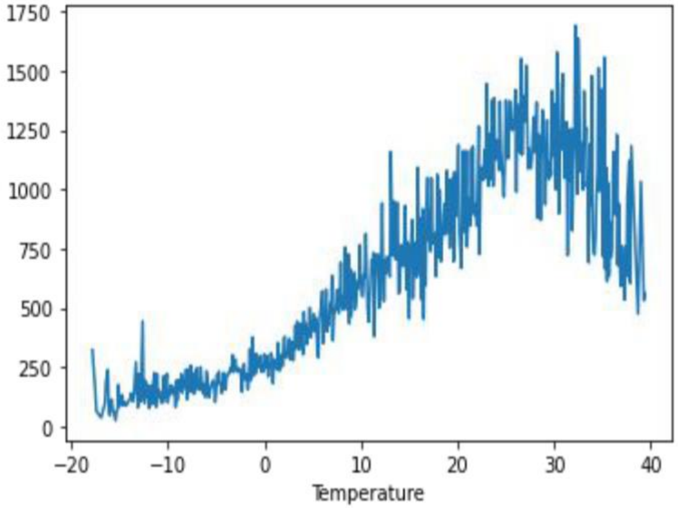
ANALYSIS OF HOLIDAY VARIABLE

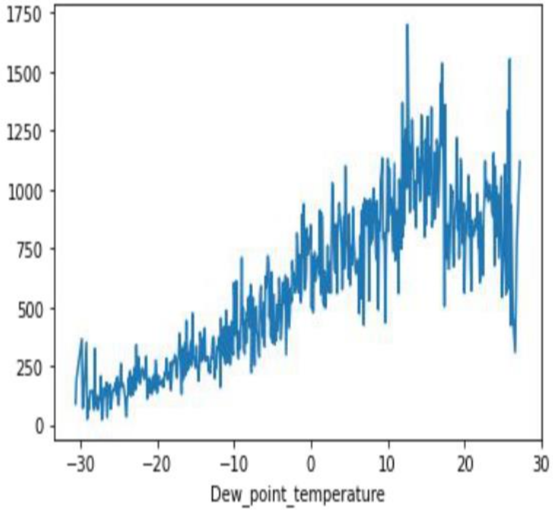


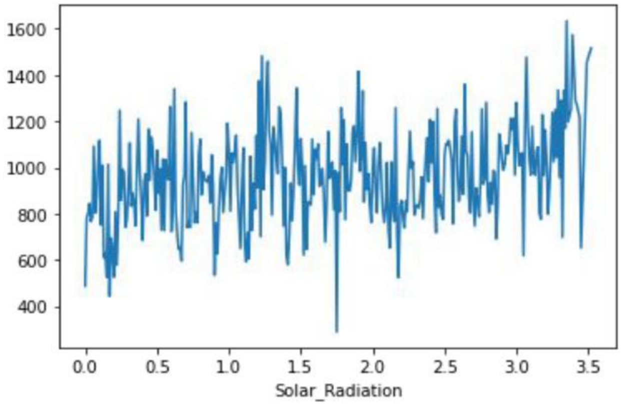
In the above bar plot and point plot which shows the use of rented bike in a holiday, and it clearly shows that,

plot shows that in holiday people uses the rented bike from 2pm-8pm

NUMERICAL VS.RENTED BIKE COUNT





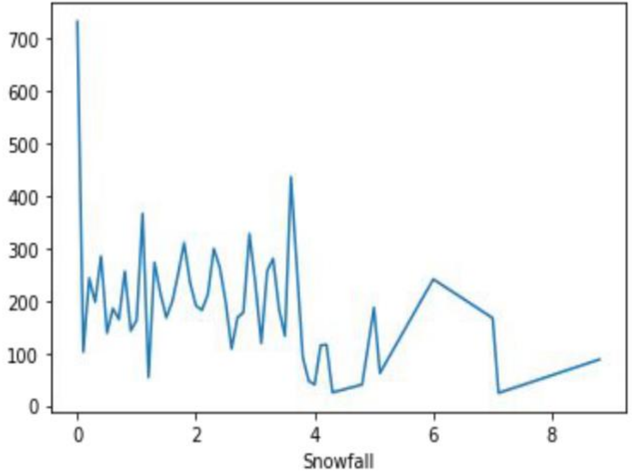


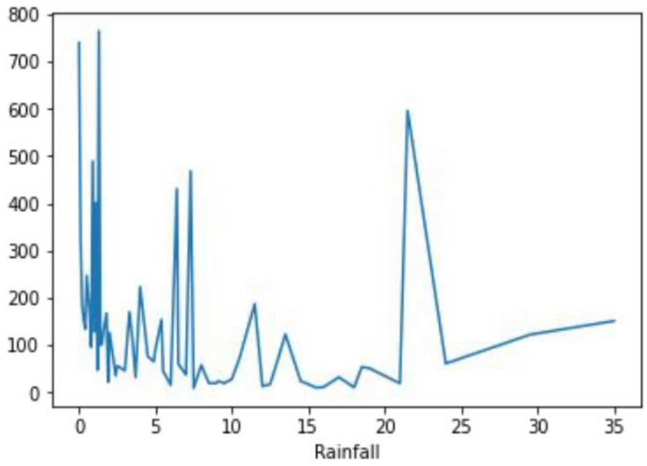
From the above plot we see that people like to ride bikes when it is pretty hot around 25°C in average

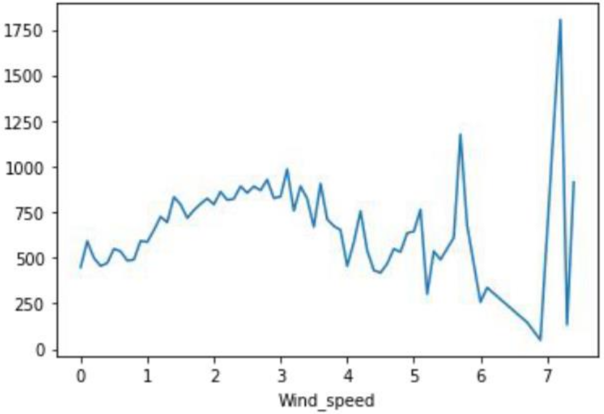
From the above plot of "Dew_point_temperature' is almost same as the 'temperature' there is some similarity present we can check it in our next step

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

NUMERICAL VS.RENTED BIKE COUNT





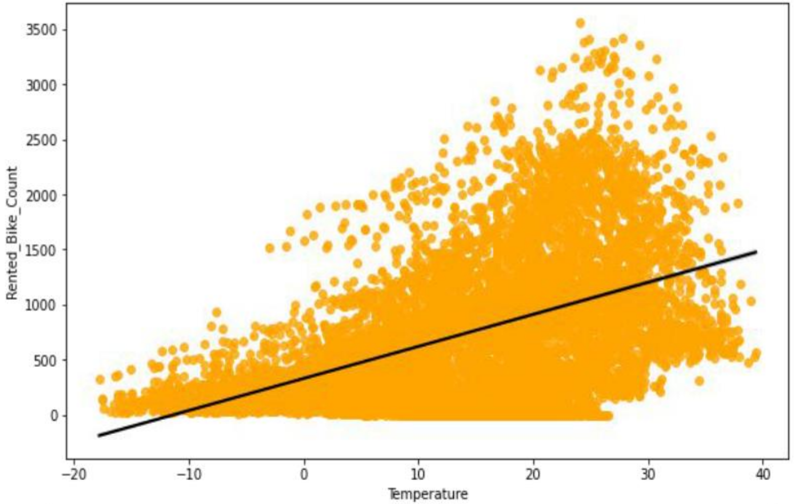


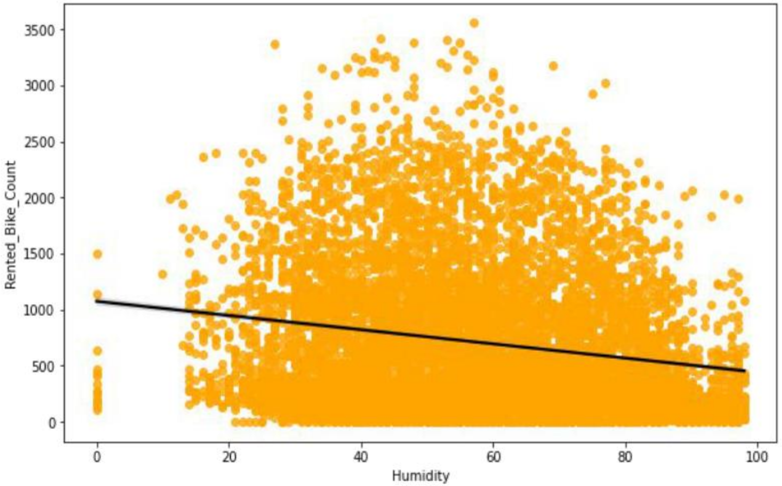
In snowfall plot, on the y-axis, the amount of rented bike is very low When we have more than 4 cm of snow, the bike rents is much lower

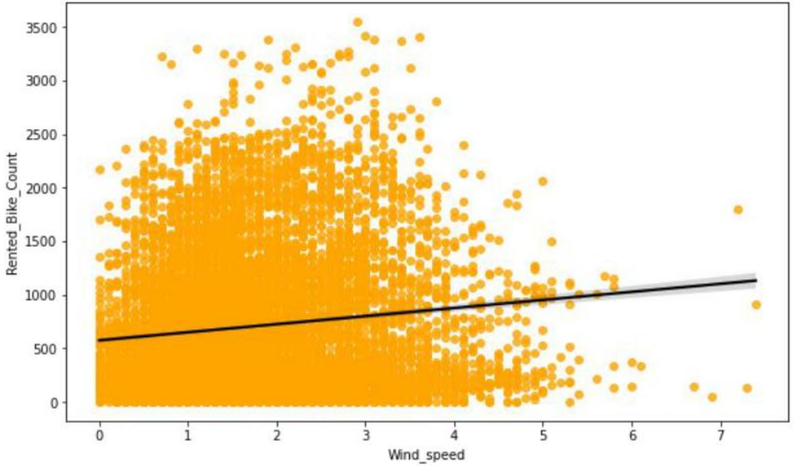
In rainfall plot if it rains a lot the demand of of rent bikes is not decreasing, here for example even if we have 20 mm of rain there is a big peak of rented bikes

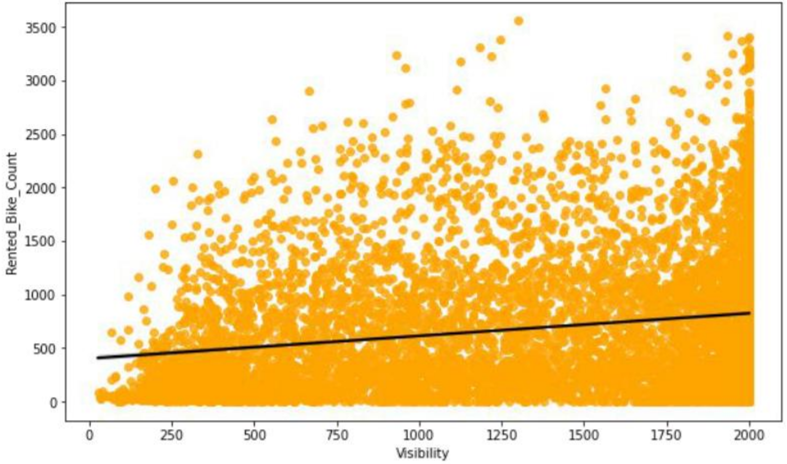
In wind speed plot that the demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of bike also increase that clearly means peoples love to ride bikes when its little windy

REGRESSION PLOT FOR NUMERICAL VARIABLE

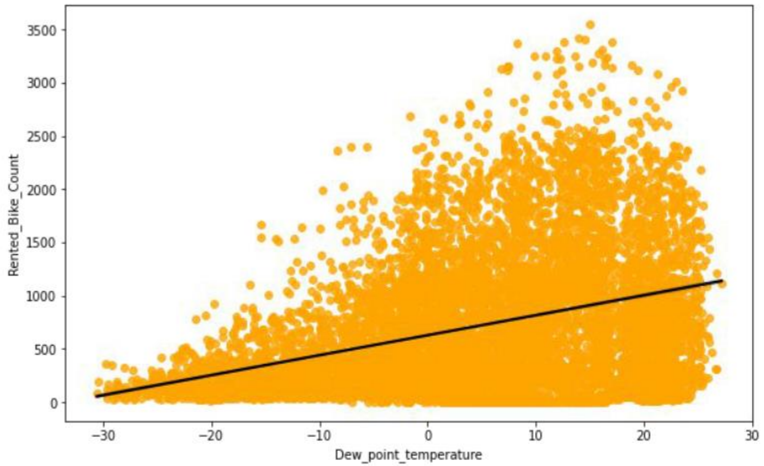


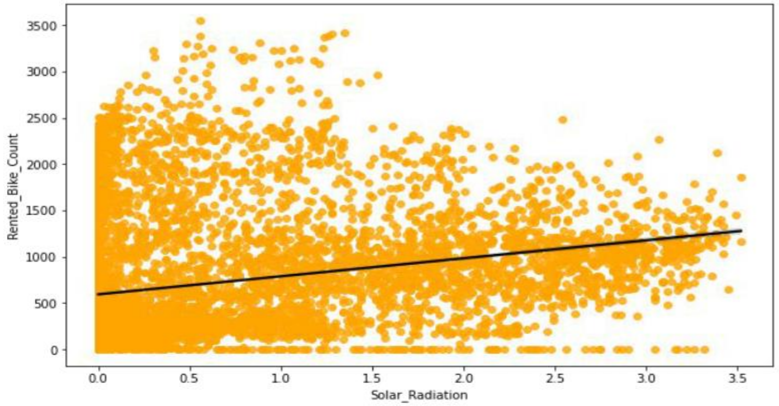


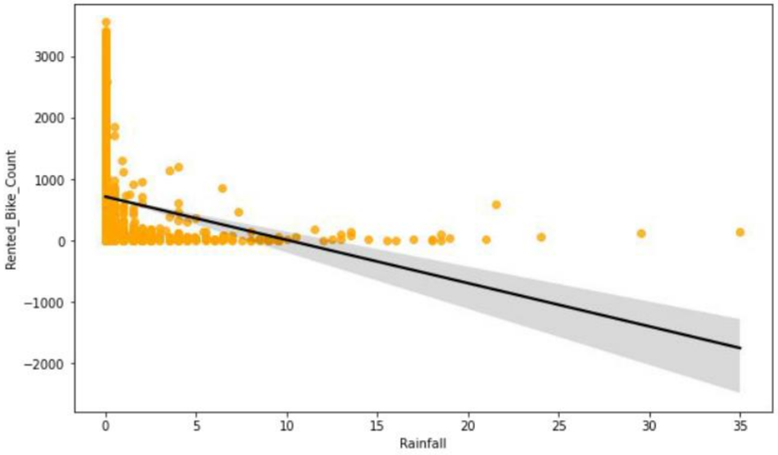


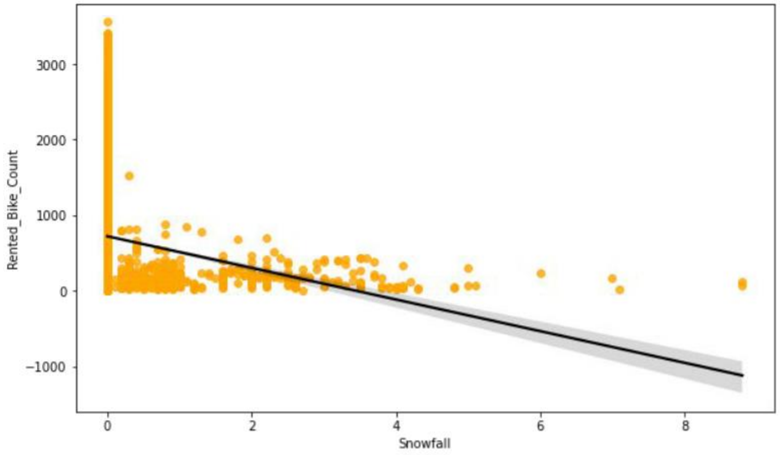


REGRESSION PLOT FOR NUMERICAL VARIABLE









REGRESSION PLOT FOR NUMERICAL VARIABLE

From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable.

which means the rented bike count increases with increase of these features.

'Rainfall','Snowfall','Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

OLS REGRESSION MODEL

R square and Adj Square are near to each other. 40% of variance in the Rented Bike count is explained by the model.

P value of dew point temp and visibility are very high and they are not significant.

OLS Regression Results

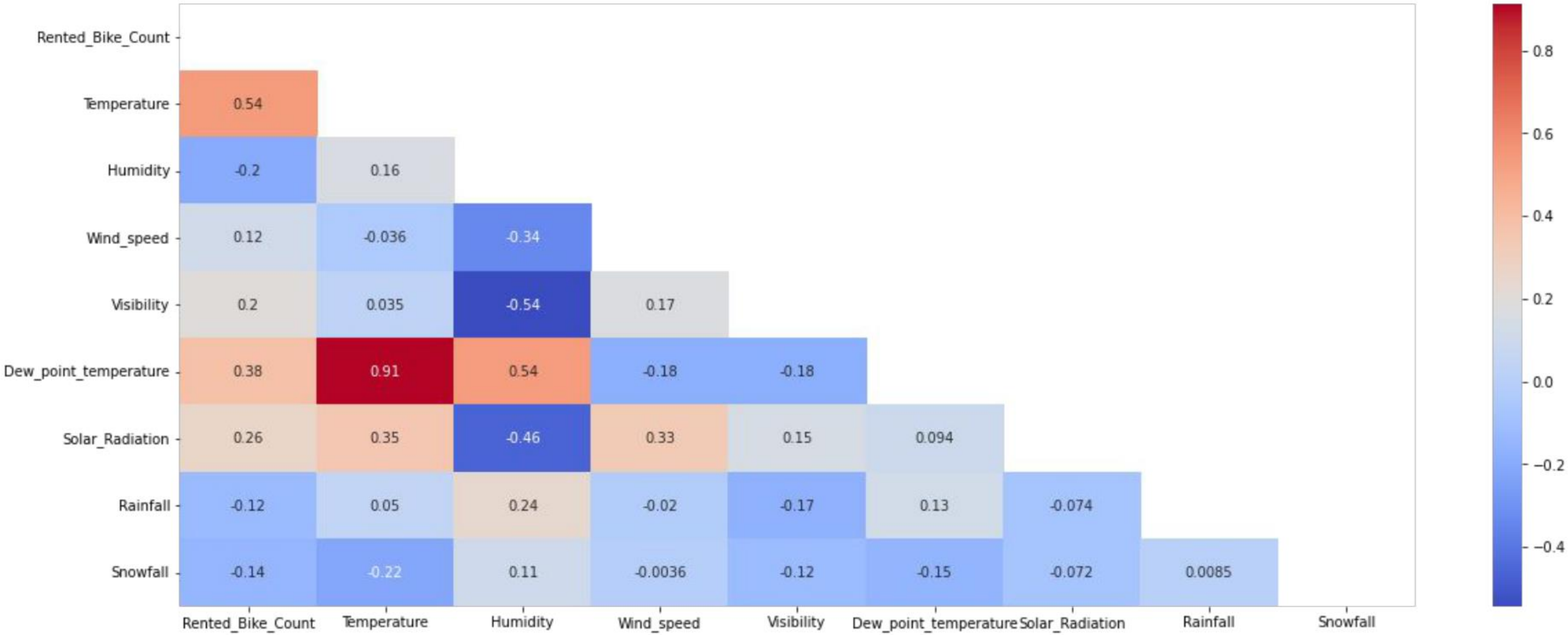
Dep. Variable:	Rented_Bike_Count	**R-squared:**	0.398
Model:	OLS	**Adj. R-squared:**	0.397
Method:	Least Squares	**F-statistic:**	723.1
Date:	Sat, 23 Oct 2021	**Prob (F-statistic):**	0.00
Time:	01:38:36	**Log-Likelihood:**	-66877.
No. Observations:	8760	**AIC:**	1.338e+05
Df Residuals:	8751	**BIC:**	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



Variables like Dew Point Temperature, and Temperature are highly correlated.

MODEL BUILDING

LINEAR REGRESSION

LASSO REGRESSION

RIDGE REGRESSION

DECISION TREES REGRESSOR

RANDOM FOREST REGRESSOR

GRADIENT BOOSTED REGRESSOR

GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

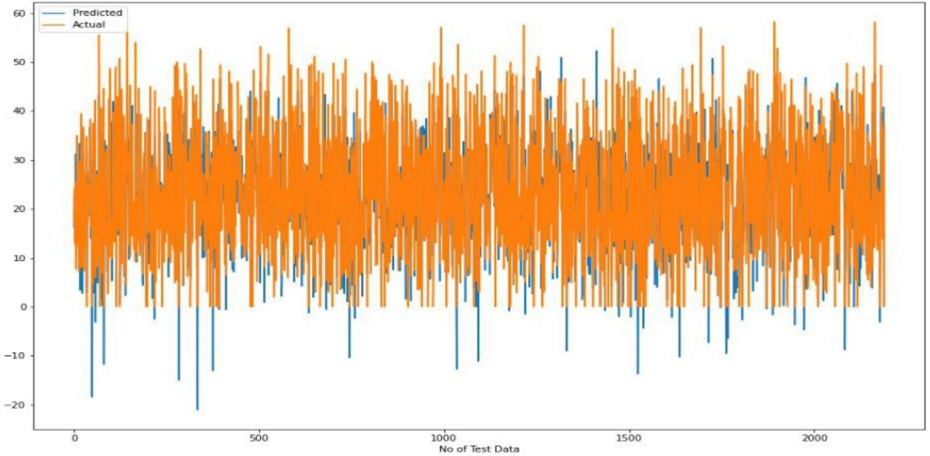
LINEAR REGRESSION

Train Set Results

MSE : 35.07751288189293
RMSE : 5.9226271942350825
MAE : 4.474024092996787
R2 : 0.7722101548255267
Adjusted R2 : 0.7672119649454145

Test Set Results

MSE : 33.27533089591926
RMSE : 5.76847734639907
MAE : 4.410178475318181
R2 : 0.7893518482962683
Adjusted R2 : 0.7847297833429184



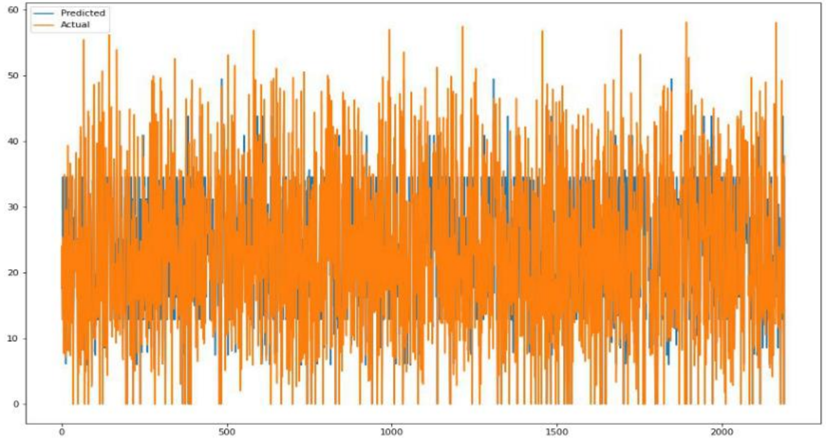
DECISION TREE

Train Set Results

Model Score: 0.6981559464575622
MSE : 46.48117069638428
RMSE : 6.817710077172854
MAE : 5.0257571131963195
R2 : 0.6981559464575622
Adjusted R2 : 0.6915328509783397

Test Set Results

MSE : 55.5974089986712
RMSE : 7.456367010727892
MAE : 5.445066995469023
R2 : 0.6480428254276878
Adjusted R2 : 0.6403201423254943



LASSO REGRESSION

Train Set Results

```
MSE : 91.59423336097032
RMSE : 9.570487623991283
MAE : 7.255041571454952
R2 : 0.40519624904934015
Adjusted R2 : 0.3921449996120475
```

Test Set Results

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

RIDGE REGRESSION

Train Set Results

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

Test Set Results

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

ELASTIC NET REGRESSION

Train Set Results

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

Test Set Results

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

RANDOM FOREST

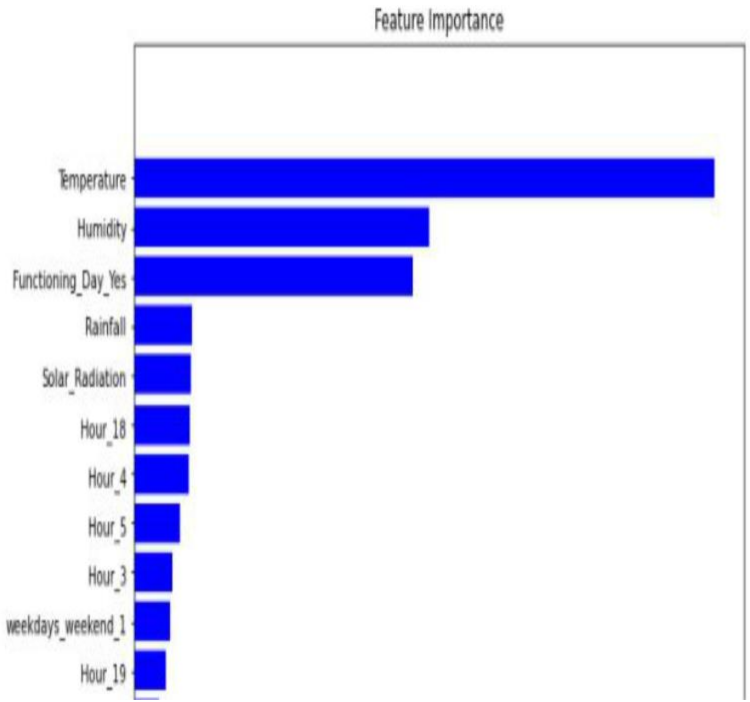
Train Set Results

```
Model Score: 0.9897105868043214
MSE : 1.5844737224439709
RMSE : 1.258758802330284
MAE : 0.7946856648569603
R2 : 0.9897105868043214
Adjusted R2 : 0.989484815366321
```

Test Set Results

```
MSE : 12.450659630923473
RMSE : 3.528549224670597
MAE : 2.1957334346668635
R2 : 0.921181597053091
Adjusted R2 : 0.9194521549716229
```

	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
11	Hour_5	0.03
25	Hour_19	0.02
46	weekdays_weekend_1	0.02
9	Hour_3	0.02



GRADIENT BOOSTING

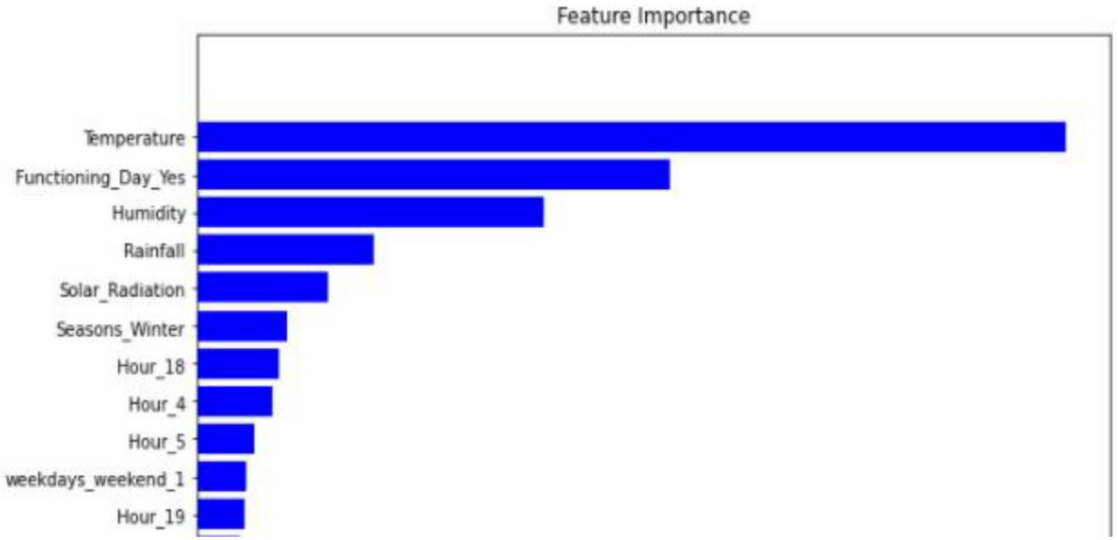
Train Set Results

```
Model Score: 0.8789016499095264
MSE : 18.648017131847947
RMSE : 4.318334995324928
MAE : 3.269035692731247
R2 : 0.8789016499095264
Adjusted R2 : 0.8762444965695393
```

Test Set Results

```
MSE : 21.28944184250869
RMSE : 4.6140483138463875
MAE : 3.4928587865599914
R2 : 0.8652280396863458
Adjusted R2 : 0.8622708584843188
```

	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
32	Seasons_Winter	0.03



GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

Train Set Results

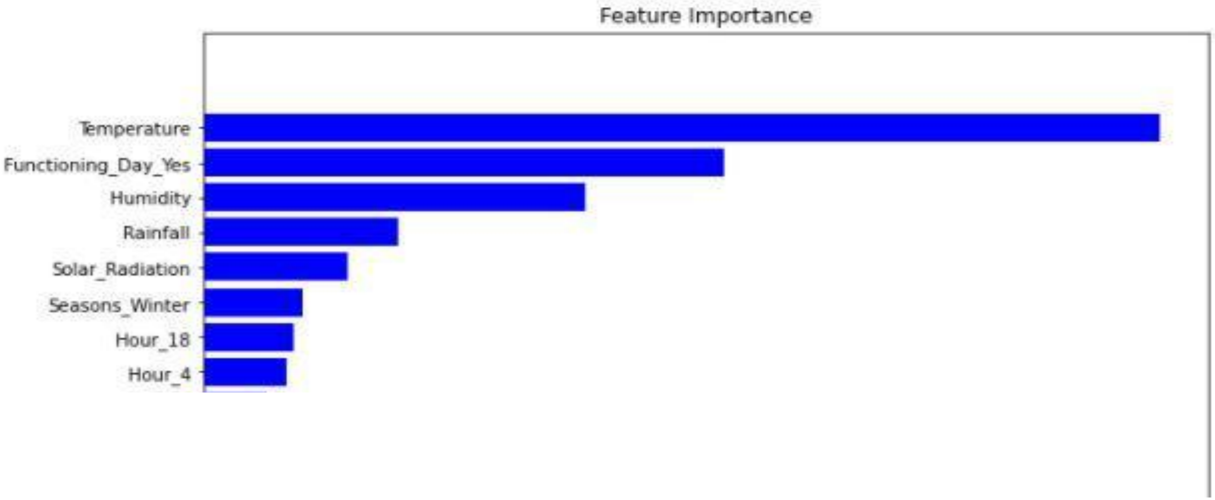
```
Model Score: 0.9515896672300013
MSE : 7.454740004128373
RMSE : 2.7303369762958516
MAE : 1.8489194833919358
R2 : 0.9515896672300013
Adjusted R2 : 0.9505274423746372
```

Test Set Results

```
MSE : 12.393403249345436
RMSE : 3.5204265720712646
MAE : 2.400740795687881
R2 : 0.921544056287242
Adjusted R2 : 0.9198225673262245
```

Hyper parameter

```
{'max_depth': 8,
 'min_samples_leaf': 40,
 'min_samples_split': 50,
 'n_estimators': 100}
```



	Feature	Feature Importance
0	Temperature	0.31
34	Functioning_Day_Yes	0.16
1	Humidity	0.15
4	Solar_Radiation	0.04
5	Rainfall	0.04

CHALLENGES

Large Dataset to handle.

Needs to plot lot of Graphs to analyse.

Feature engineering

Feature selection

Optimising the model

Carefully tuned Hyperparameters as it affects the R2 score.

CONCLUSION

‘Hour’ of the day holds the most important feature.

Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening.

We observed that bike rental count is high during working days than non working day.

We see that people generally prefer to bike at moderate to high temperatures, and when little windy

It is observed that highest number bike rentals counts in Autumn & Summer seasons & the lowest in winter season. We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day. We observed that with increasing humidity, the number of bike rental counts decreases.

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% respectively for Train Set and 92% for Test set. So, finally this model is best for predicting 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	5.026	46.481	6.818	0.698	0.69
	5	Random forest regression	0.795	1.584	1.259	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	5.445	55.597	7.456	0.648	0.64
	5	Random forest regression	2.196	12.451	3.529	0.921	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

THANK YOU