Capstone Project-2

Bike Sharing Demand Prediction

Team Members:

1. Aditya Kumar Saw
2. Parijat Krishna

List of contents

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.

How do we solved this?

We have Seoul Bike Sharing data, which we can use to predict the number of bikes that might be in demand for rentals. We need to follow some steps in order to get there.

1. Setting the ultimate Goal.
2. Understanding the dataset.
3. EDA & Feature Engineering.
4. Preparing data for modelling.
5. Applying models.
6. Model Validation & Selection.

Data Collection and Understanding:

- For analysis and model building we are have the **Seoul Bike Data**.
- The Dataset contain **8760 rows** of observations and **14 attributes**.

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** - %
- **Wind Speed** - m/s

Contd..

Attributes Information

- **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 Hours), Fun(Functional hours)

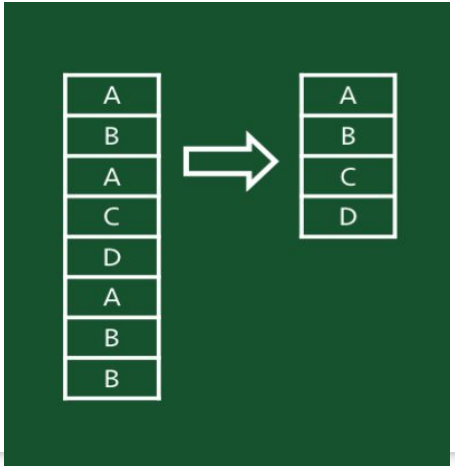
Data Wrangling

Why do we need to handle missing values?

- The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.that's why we check missing values first.

- As we can see above there are no missing value presents thankfully.

Duplicate values



Why is it important to remove duplicate records from my data?

"Duplication" means that you have repeated data in your dataset. This could be due to things like data entry errors or data collection methods. by removing duplication in our data set, Time and money are saved by not sending identical communications multiple times to the same person.

```
[ ]  # Checking Duplicate Values
     value=len(bike_df[bike_df.duplicated()])
     print("The number of duplicate values in the data set is = ",value)
```

```
The number of duplicate values in the data set is =  0
```

- *In the above data after count the missing and duplicate value we came to know that there are no missing and duplicate value present.*

- *Some of the columns name in the dataset are too large and clumsy so we change the the into some simple name, and it don't affect our end results.*

Contd..

- In dataset we don’t have any missing or null value.
- We have zero duplicates.
- For feature engineering we changed data type of Date column form ‘object’ to ‘datetime64[ns]’.
- We creating new columns ‘Day’ and ‘Month’ form Data for further EDA.

EDA & Feature Engineering :

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. This is where we have spent most of the time.

Univariate Analysis

- The key objective of Univariate analysis is to simply describe the data to find patterns within the data.

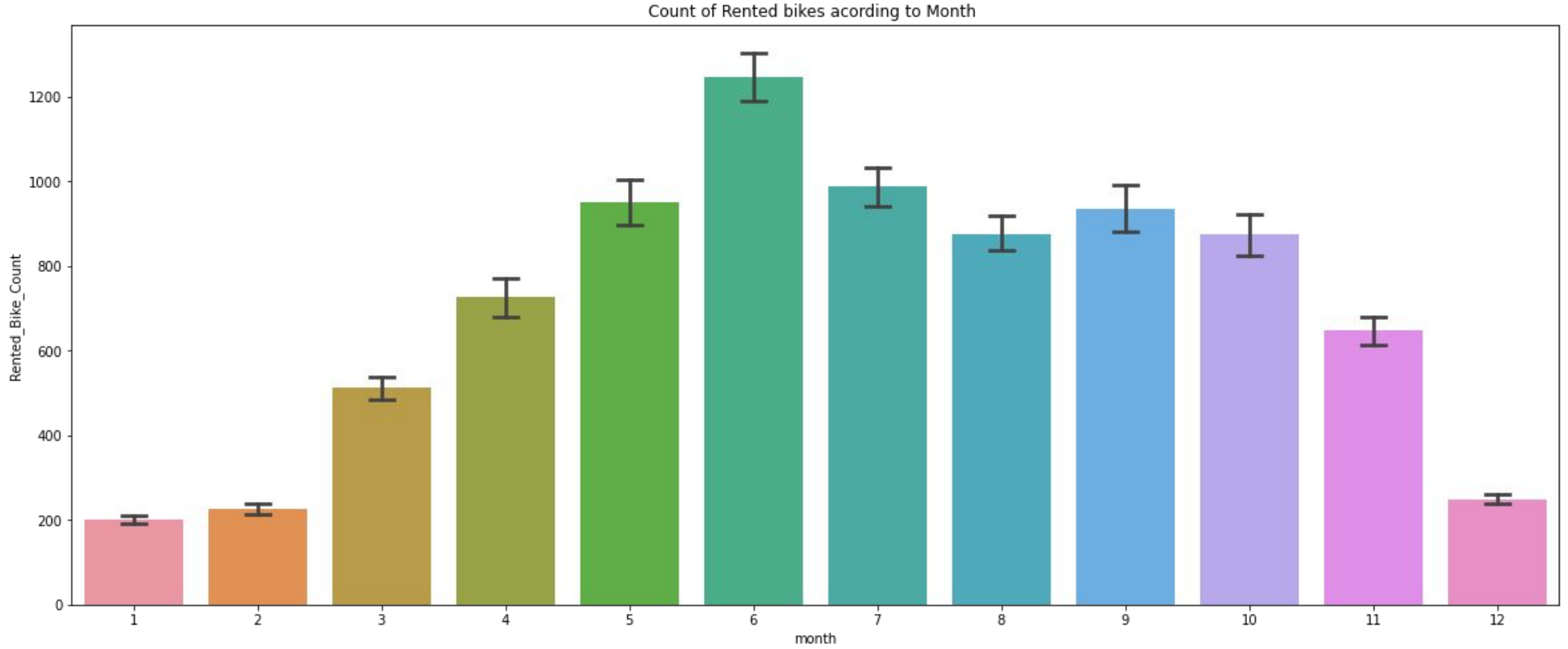
Dependent variable in data analysis

- We analysed our dependent variable A, dependent variable is a variable whose value will change depending on the value of another variable.

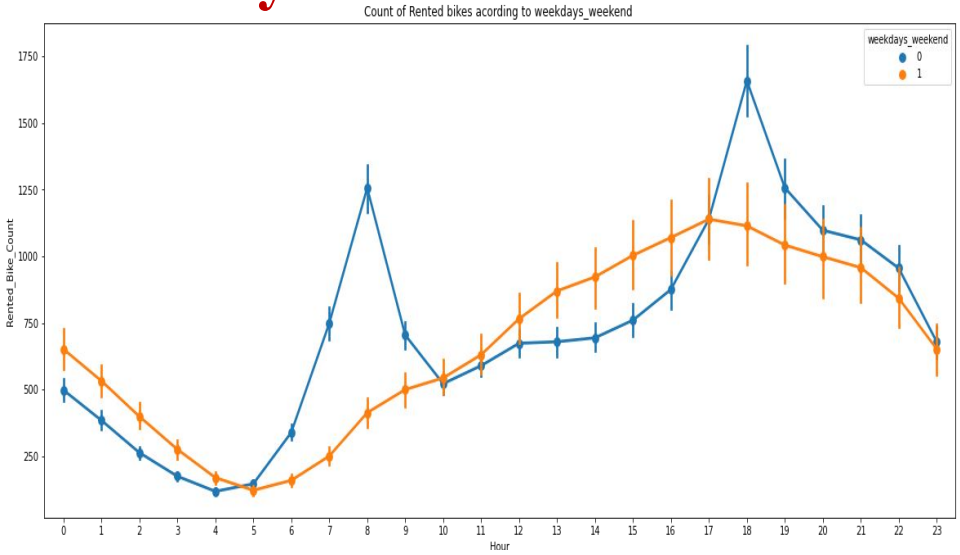
Analysation of categorical variables

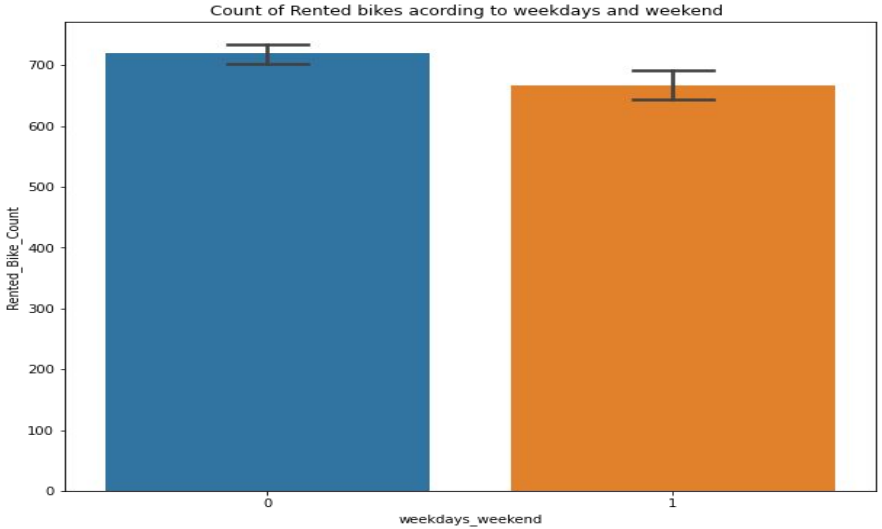
- Our dependent variable is "Rented Bike Count" so we need to analysis this column with the other columns by using some visualisation plot.first we analyze the category data tyep then we proceed with the numerical data type.

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.



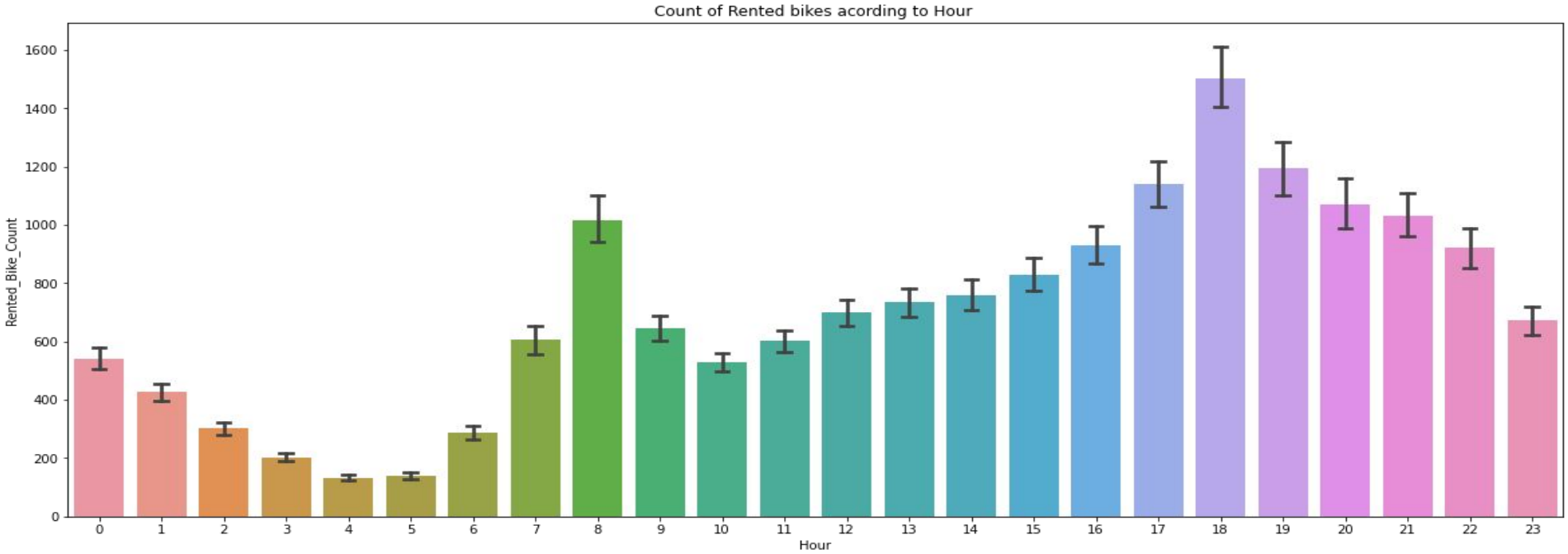
Weekdays v/s Weekend





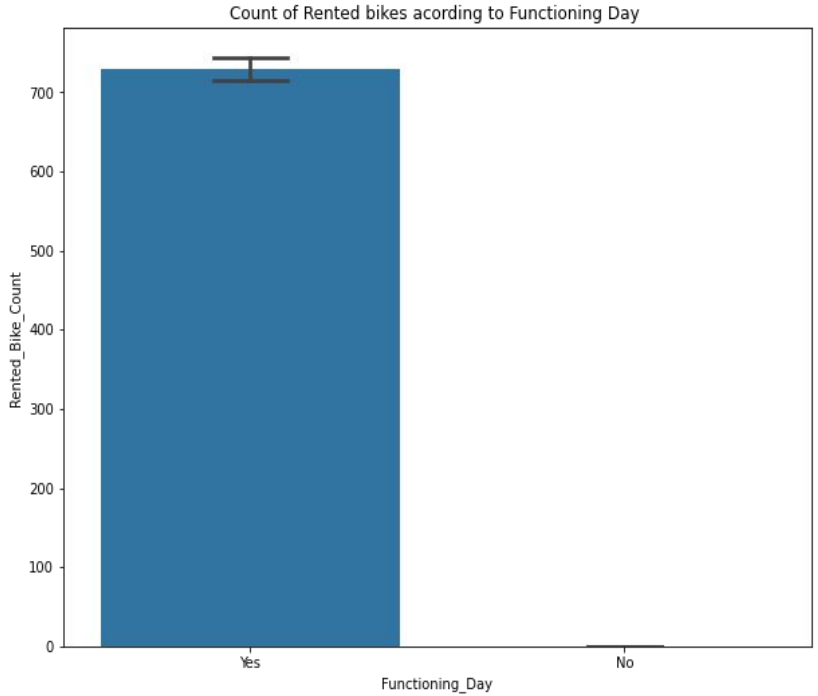
- From the above point plot and bar plot we can say that in the week days which represent in blue colur show that the demand of the bike higher because of the office.
- Peak Time are 7 am to 9 am and 5 pm to 7 pm
- The orange colur represent the weekend days, and it show that the demand of rented bikes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

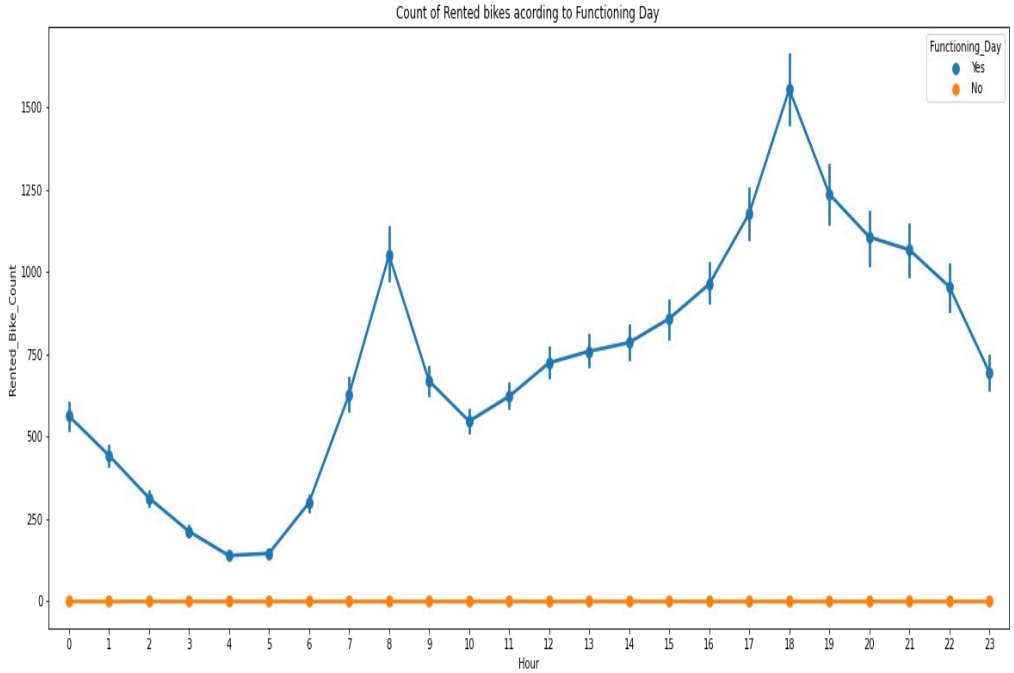
Hour



- 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.

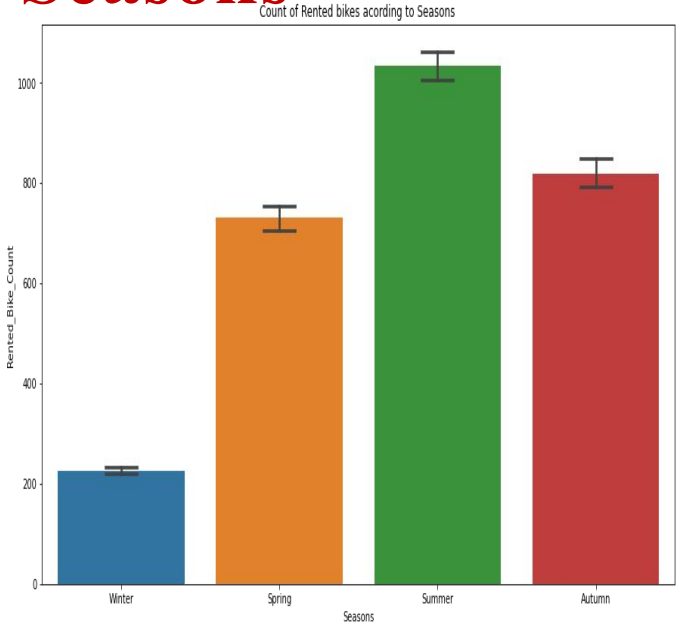
Functioning Day

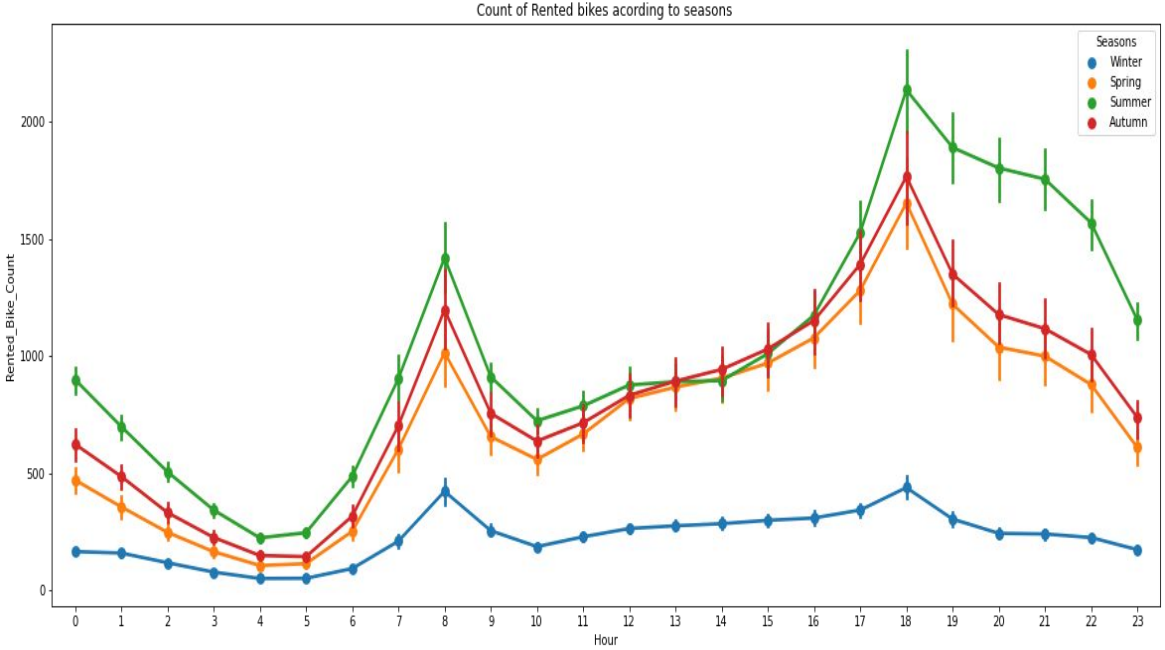




- In the above bar plot and point plot which shows the use of rented bike in functioning daya or not, and it clearly shows that,
- Peoples dont use rented bikes in no functioning day.

Seasons

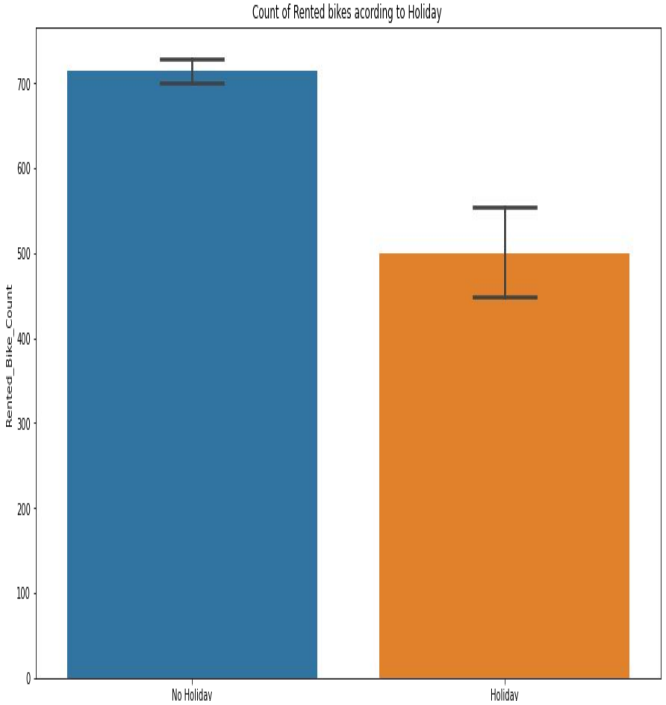


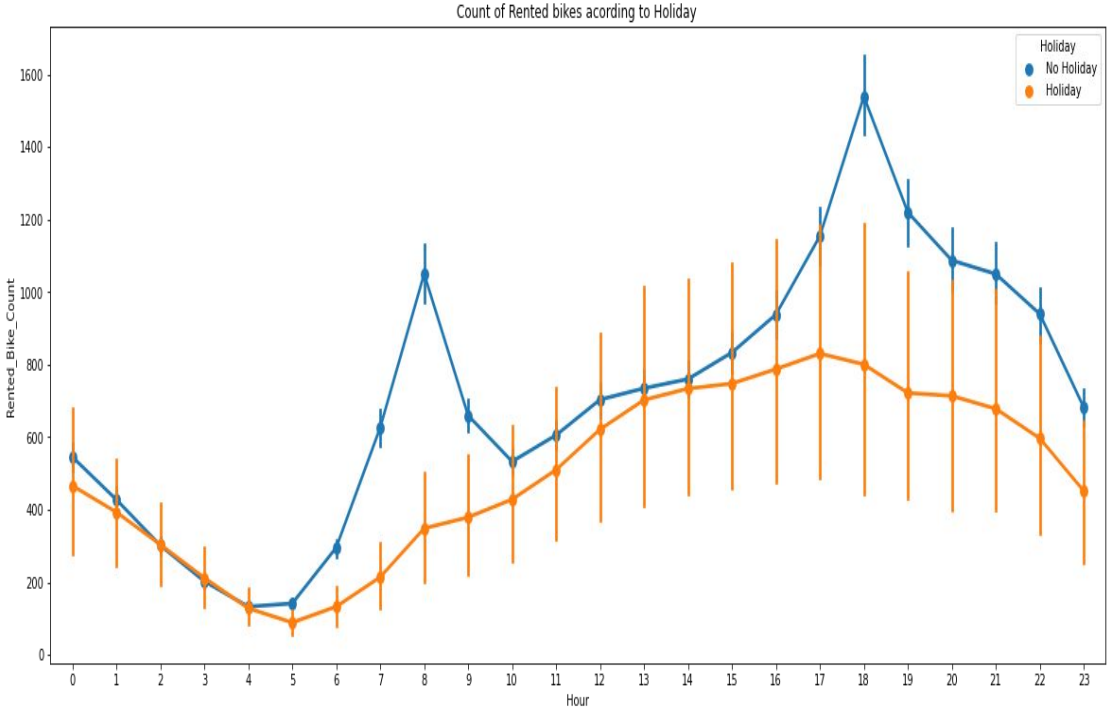


In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,

- 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.

Holiday



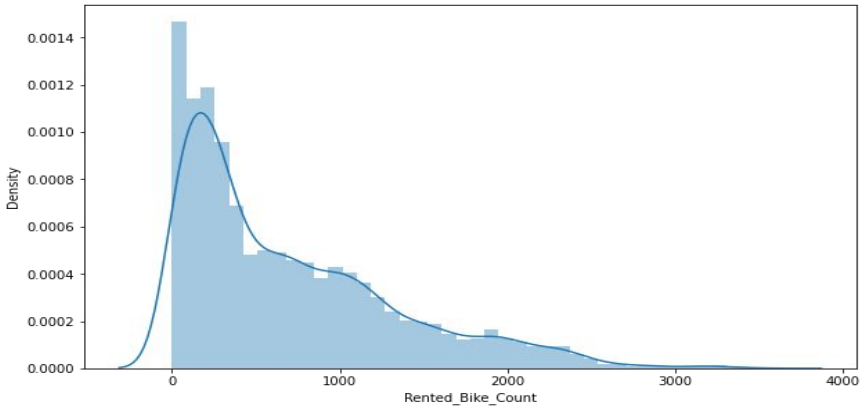


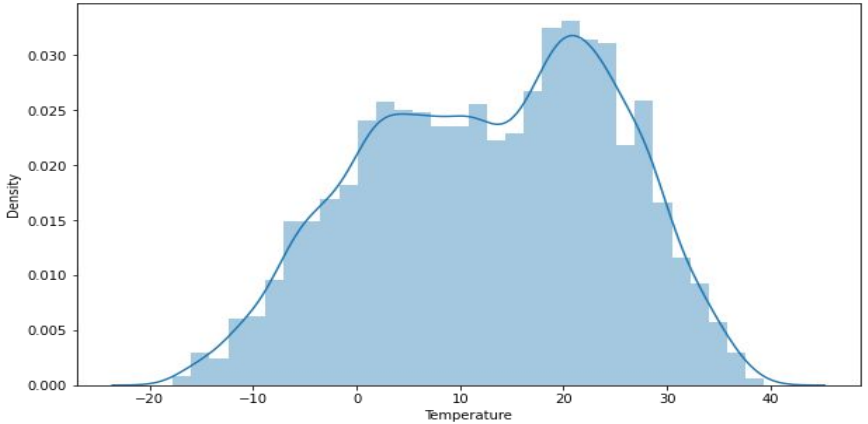
In the above bar plot and point plot which shows the use of rented bike in a holiday, and it clearly shows that,

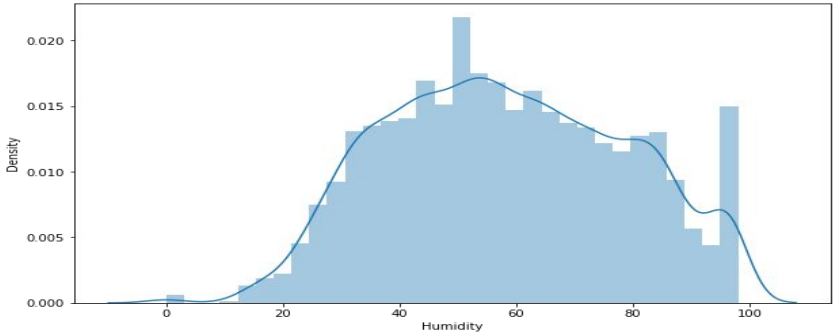
- Plot shows that in holiday people uses the rented bike from 2pm-8pm.

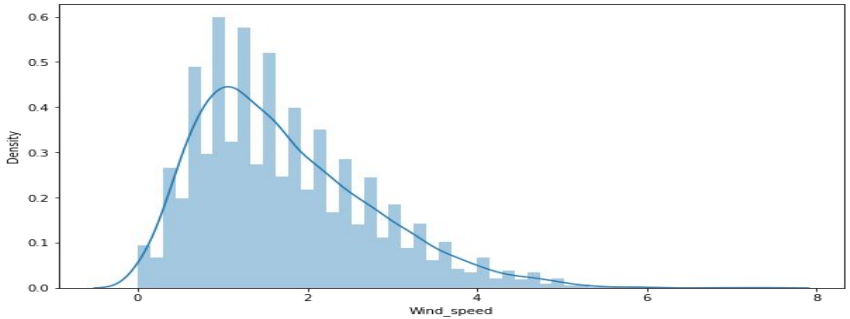
Analysis of Numerical variables

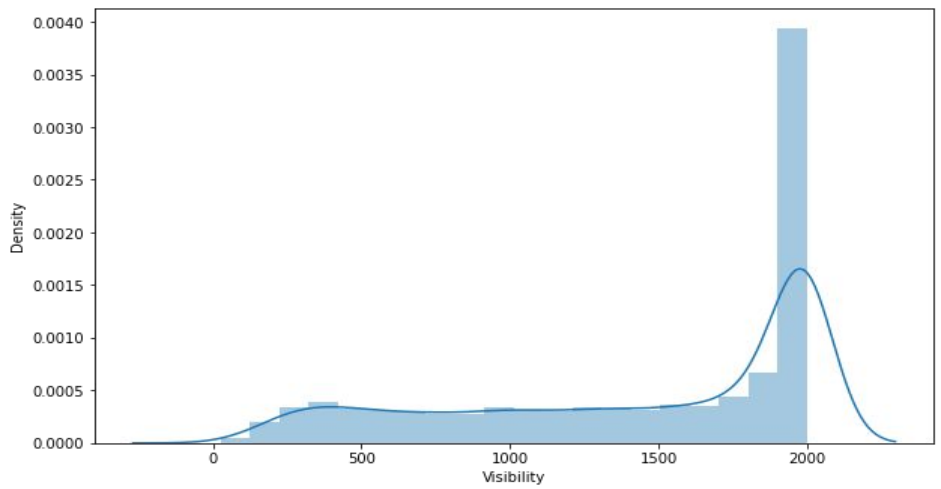
- Numerical data is a data type expressed in numbers, rather than natural language description. Sometimes called quantitative data, numerical data is always collected in number form. Numerical data differentiates itself from other number form data types with its ability to carry out arithmetic operations with these numbers.

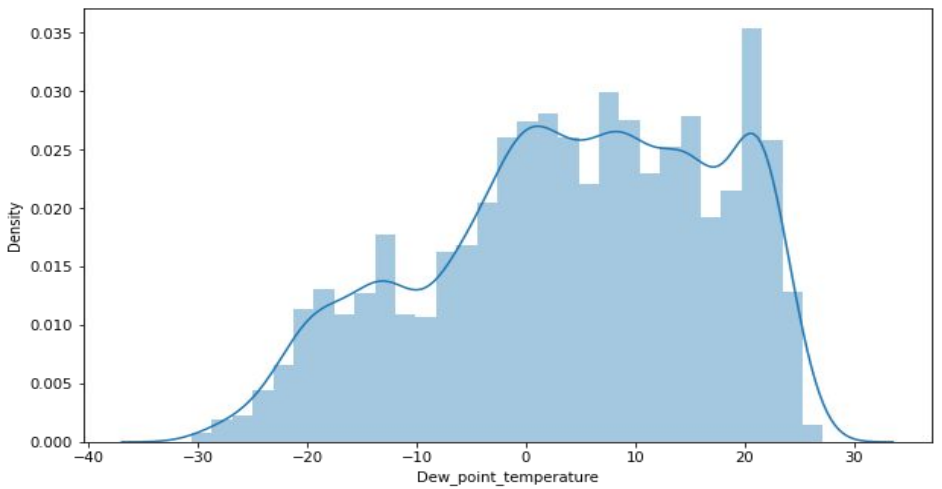


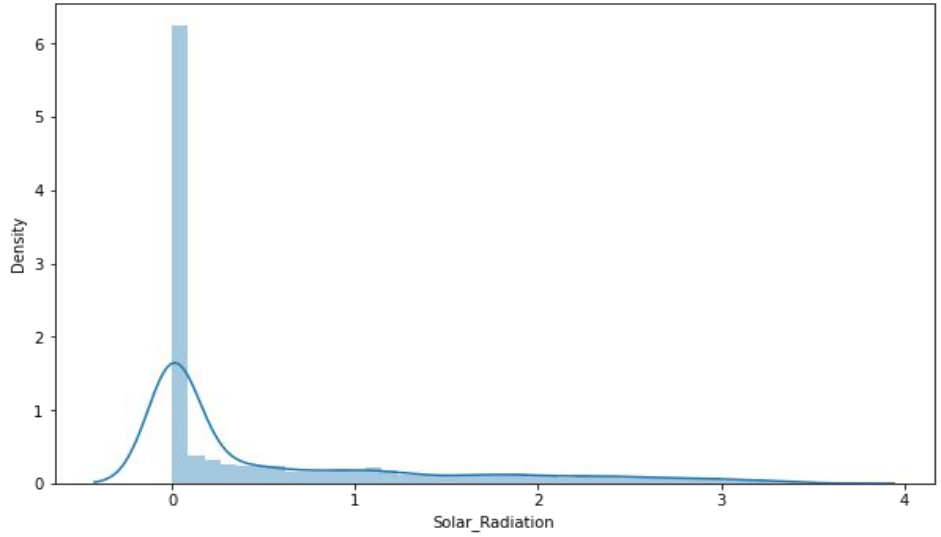


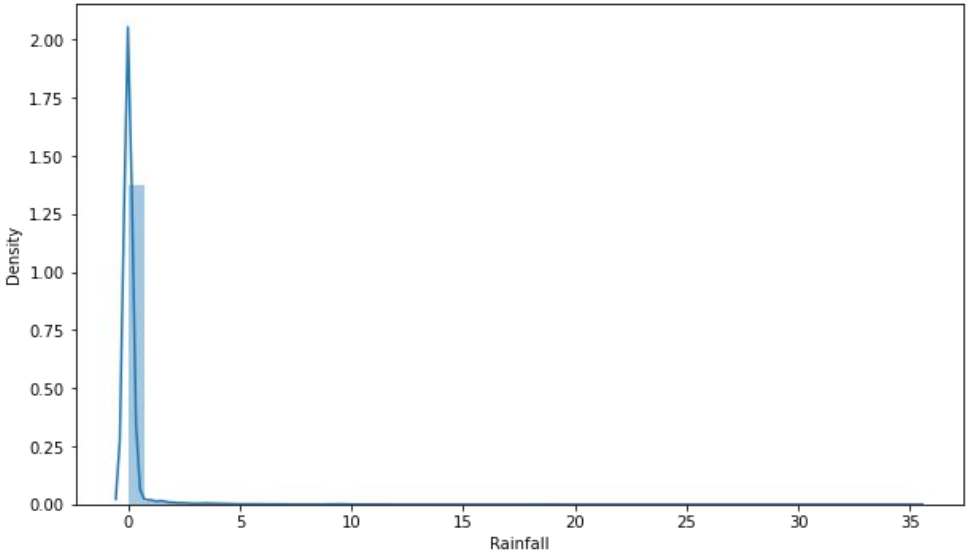


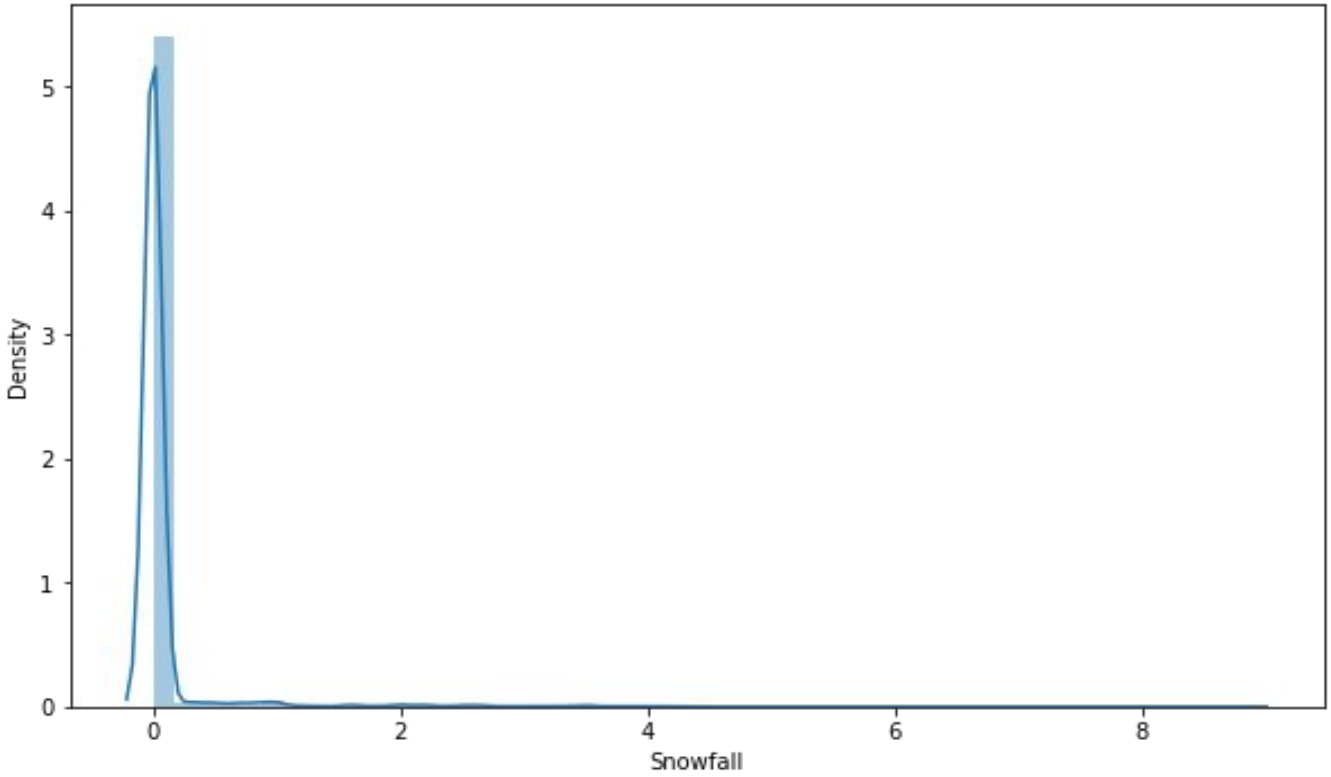



Density
Visibility

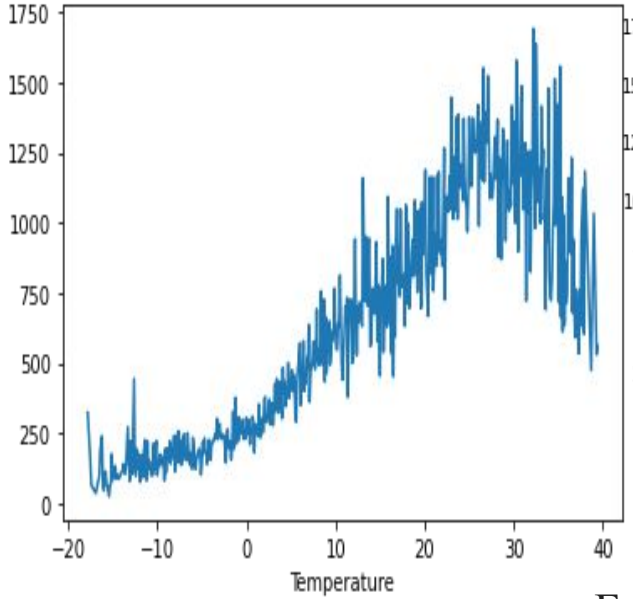

Density
Dew_point_temperature


Density
Solar_Radiation

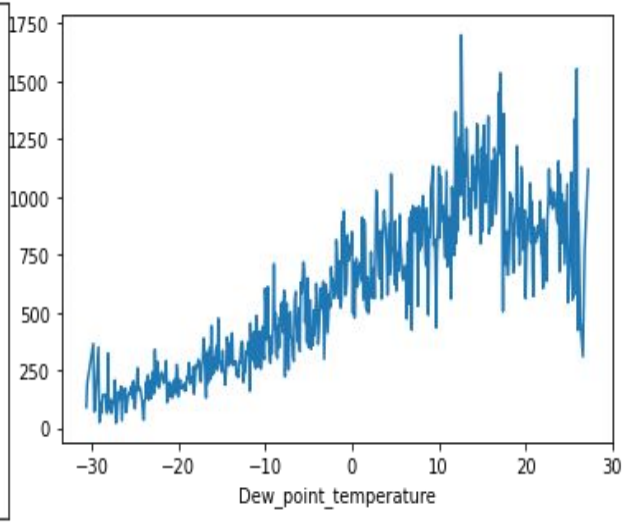

Density
Rainfall


Density
0
1
2
3
4
5
0
2
4
6
8
Snowfall

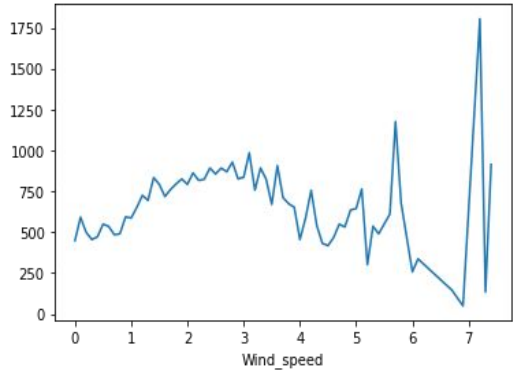
Numerical v/s Rented Bike Count



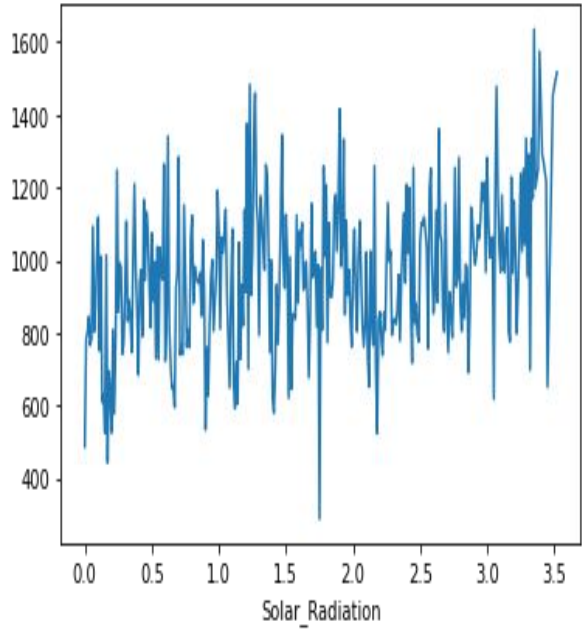
From the above plot we see that people like to ride bikes when it is pretty hot around 25°C in average



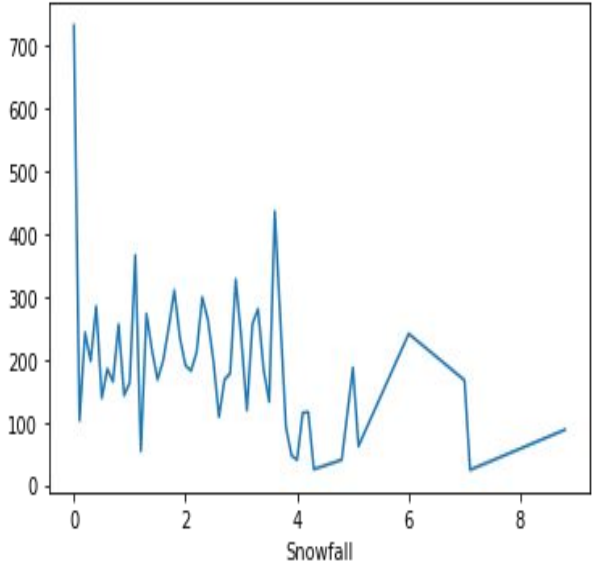
From the above plot of "Dew_point_temperature' is almost same as the 'temperature' there is some similarity present we can check it in our next step.



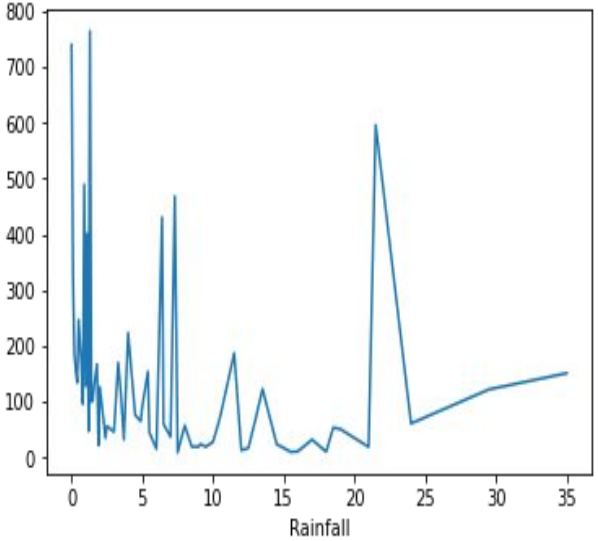
We can see from the above plot that the demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of bike also increase that clearly means peoples love to ride bikes when its little windy.



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.

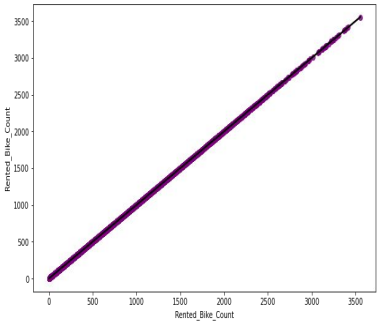


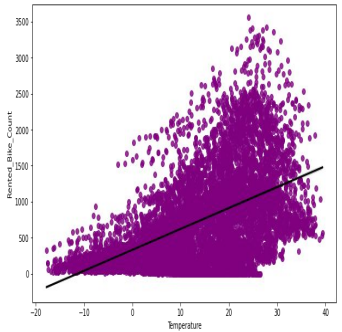
We can see from the plot that, on the y-axis, the amount of rented bike is very low When we have more than 4 cm of snow, the bike rents is much lower.

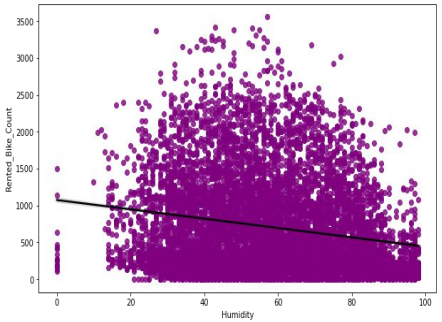


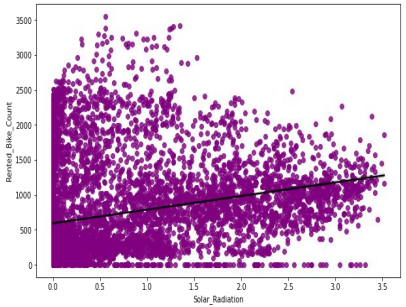
We can see from the above plot that even if it rains a lot the demand of of rent bikes is not decreasing, here for example even if we have 20 mm of rain there is a big peak of rented bikes.

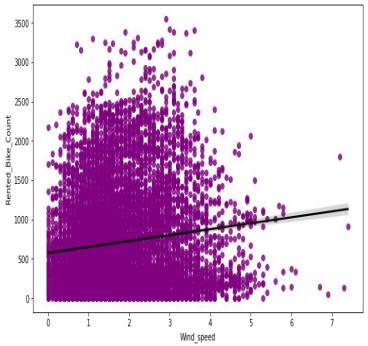
The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses. Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

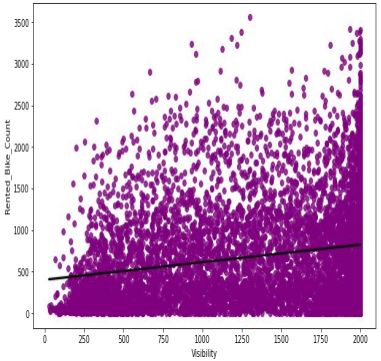


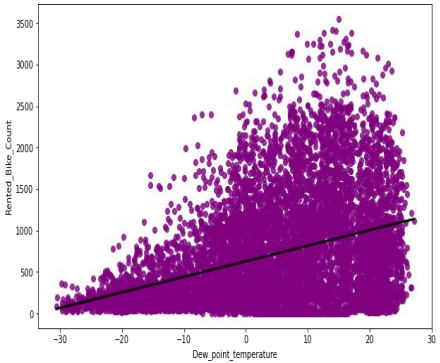


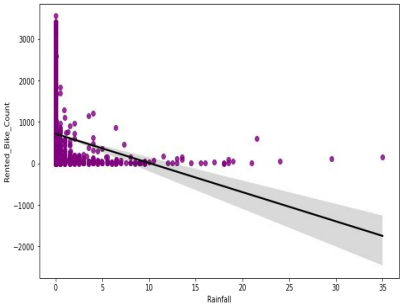


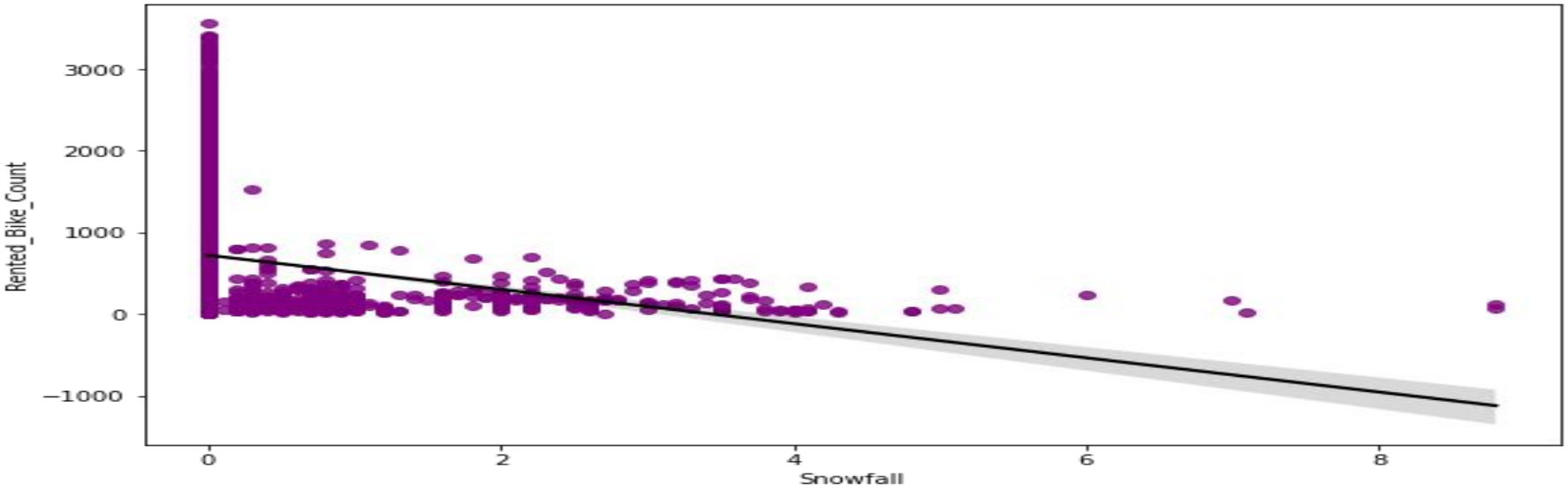








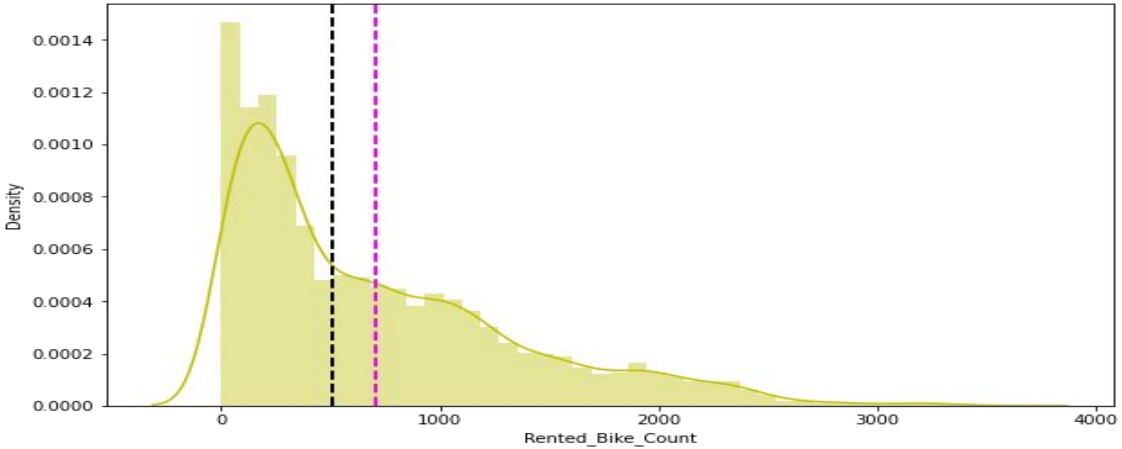


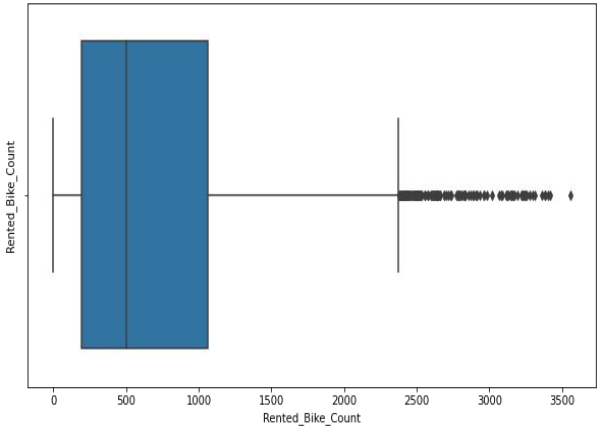


- From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable.
- which means the rented bike count increases with increase of these features.
- 'Rainfall','Snowfall','Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

Normalise Rented Bike Count Column Data

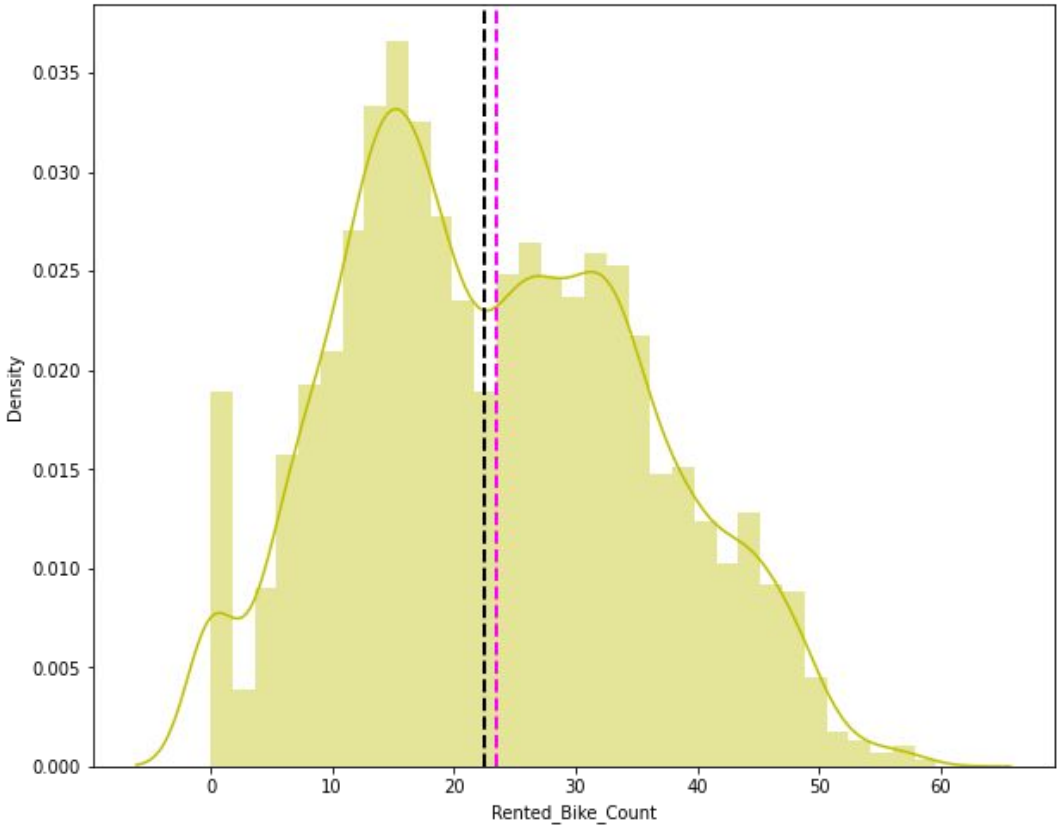
The data normalization (also referred to as data pre-processing) is a basic element of data mining. It means transforming the data, namely converting the source data in to another format that allows processing data effectively. The main purpose of data normalization is to minimize or even exclude duplicated data.

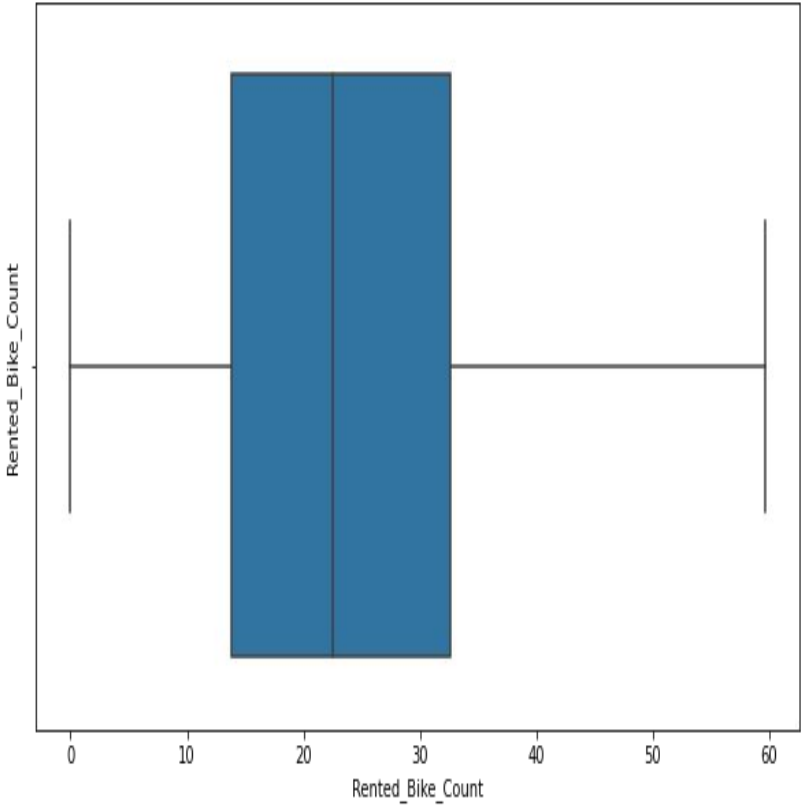




The above graph shows that Rented Bike Count has moderate right skewness. Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform some operation to make it normal.

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

	Rented_Bike_Count	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall
Rented_Bike_Count	1.000000	0.538558	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804
Temperature	0.538558	1.000000	0.159371	-0.036252	0.034794	0.912798	0.353505	0.050282	-0.218405
Humidity	-0.199780	0.159371	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183
Wind_speed	0.121108	-0.036252	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554
Visibility	0.199280	0.034794	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695
Dew_point_temperature	0.379788	0.912798	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887
Solar_Radiation	0.261837	0.353505	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301
Rainfall	-0.123074	0.050282	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500
Snowfall	-0.141804	-0.218405	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000

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

Checking of Correlation between variables

Checking in OLS Model

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable.

- R square and Adj Square are near to each other. 40% of variance in the Rented Bike count is explained by the model.
- For F statistic , P value is less than 0.05 for 5% level of significance.
- P value of dew point temp and visibility are very high and they are not significant.
- Omnibus tests the skewness and kurtosis of the residuals. Here the value of Omnibus is high., it shows we have skewness in our data.
- The condition number is large, 3.11e+04. This might indicate that there are strong multicollinearity or other numerical problems
- Durbin-Watson tests for autocorrelation of the residuals. Here value is less than 0.5. We can say that there exists a positive auto correlation among the variables.

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:	Mon, 12 Dec 2022	Prob (F-statistic):	0.00
Time:	03:16:41	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

Notes:
[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.

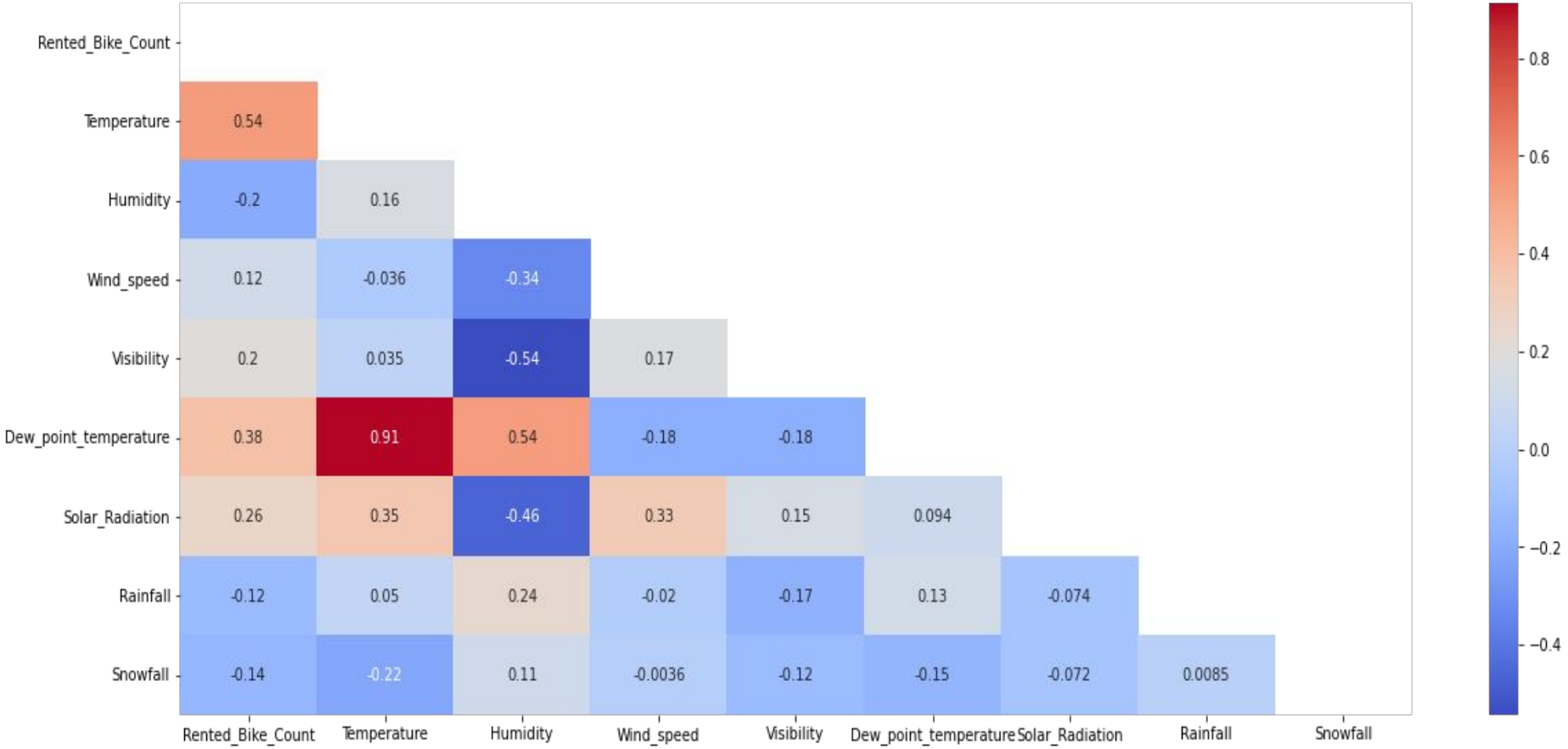
	const	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall
const	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Temperature	NaN	1.000000	0.159371	-0.036252	0.034794	0.912798	0.353505	0.050282	-0.218405
Humidity	NaN	0.159371	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183
Wind_speed	NaN	-0.036252	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554
Visibility	NaN	0.034794	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695
Dew_point_temperature	NaN	0.912798	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887
Solar_Radiation	NaN	0.353505	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301
Rainfall	NaN	0.050282	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500
Snowfall	NaN	-0.218405	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000

From the OLS model we find that the 'Temperature' and 'Dew_point_temperature' are highly correlated so we need to drop one of them.

for dropping the we check the (P>|t|) value from above table and we can see that the 'Dew_point_temperature' value is higher so we need to drop Dew_point_temperature column

For clarity, we use visualisation i.e heatmap in next step.

Correlation Heatmap



Applying Regression Models :

We applied the following baseline models and the results for the models were

Model
XGBoost Regressor
Gradient Boosting
Random Forest Regressor
Decision Tree Regression
Linear Regression
Ridge Regression
Elastic Net Regression
Knn Regressor
Lasso Regression

Metrics used for model evaluation :

- **R2 Score :** It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.
- **MSE Value :** Mean squared error (MSE) measures the amount of error in models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero.
- **RMSE Value :** Root mean square error or root mean square deviation shows how far predictions fall from measured true values using Euclidean distance.
- **Adjusted R2 Score :** The Adjusted Rsquared takes into account the number of independent variables used for predicting the target variable.

Scores (For Base Models)

		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	4.946	45.023	6.710	0.708	0.70
	5	Random forest regression	0.801	1.619	1.272	0.989	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	5.300	52.213	7.226	0.669	0.66
	5	Random forest regression	2.202	12.749	3.571	0.919	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

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.

Conclusion:

- During the time of our analysis, we initially did EDA on all the features of our datset. We first analysed our dependent variable, 'Rented Bike Count' and also transformed it. Next we analysed categorical variable and dropped the variable who had majority of one class, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable. We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.

- Next we implemented 7 machine learning algorithms Linear Regression,lasso,ridge,elastic net,decision tree, Random Forest and XGBoost. We did hyperparameter tuning to improve our model performance.

- No overfitting is seen.

Conclusion:

- 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.
- Feature Importance value for Random Forest and Gradient Boost are different.
- We can deploy this model.
- However, this is not the ultimate end. As this data is time dependent, the values for variables like temperature, windspeed, 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. Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future.

THANK YOU