Capstone Project

Bike Sharing Demand Prediction

By – Amogha K

CONTENTS

- Introduction
- Problem Statement
- Data Summary
- Exploratory Data analysis
- Implementing algorithms
- Conclusion

INTRODUCTION

- A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short-term basis for a price.
- Bike share systems allow people to borrow a bike from a "dock," which is usually computer-controlled where in the user enters the payment information, and the system unlocks it.
- This bike can then be returned to another dock belonging to the same system.
- Rental Bike Sharing is the process of renting bicycles on an hourly, weekly, or membership-based basis.
- People who have no personal vehicles and also to avoid congested public transport. That's why people prefer rental bikes.
- Our project goal is a pre-planned set of bike count values that can be a handy solution to meet all demands.

PROMBLEM STATEMENT

The contents of the data came from a city called Seoul. The data had variables such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day, and rented bike count. The problem statement was to build a machine learning model that could predict the rented bike count required for an hour, given other variables.

DATA SUMMARY

1. Date :year-month-day
2. Rented Bike count -Count of bikes rented at each hour
3. Hour -Hour of the day
4. Temperature-Temperature in Celsius
5. Humidity -%
6. Windspeed -m/s (meter per second)
7. Visibility -10m (meter)
8. Dew point temperature -Celsius
9. Solar radiation -MJ/m2
10. Rainfall -mm (millimeter)
11. Snowfall -cm (centimeter)
12. Seasons-Winter, Spring, Summer ,Autumn
13. Holiday -Holiday/No holiday
14. Functional Day –No Func(Non Functional Hours),Fun(Functional hours)

EXPLORATORY DATA ANALYSIS

Head of Data

	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

Tail of Data

	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

Info of Data

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          8760 non-null   int64
 2   Hour                       8760 non-null   int64
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object
 12  Holiday                    8760 non-null   object
 13  Functioning Day            8760 non-null   object
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

Null Values of Data

```
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
```

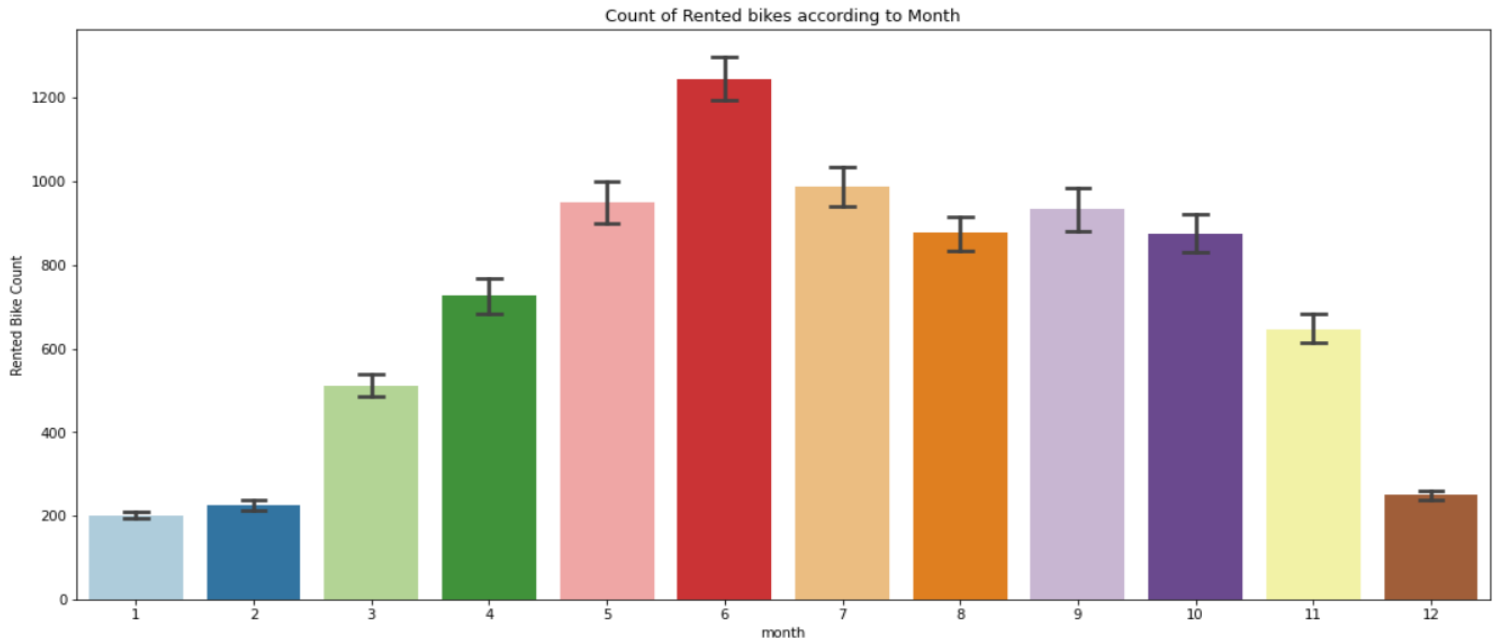
- ❑ Dataset contains 8760 lines and 14 columns.
- ❑ There were no Null values or Duplicate values in dataset.
- ❑ The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data.
- ❑ We have 'rented bike count' variable which we need to predict for new observations.

Describe of Data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Date	8760	365	01/12/2017	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rented Bike Count	8760.0	NaN	NaN	NaN	704.602055	644.997468	0.0	191.0	504.5	1065.25	3556.0
Hour	8760.0	NaN	NaN	NaN	11.5	6.922582	0.0	5.75	11.5	17.25	23.0
Temperature(°C)	8760.0	NaN	NaN	NaN	12.882922	11.944825	-17.8	3.5	13.7	22.5	39.4
Humidity(%)	8760.0	NaN	NaN	NaN	58.226256	20.362413	0.0	42.0	57.0	74.0	98.0
Wind speed (m/s)	8760.0	NaN	NaN	NaN	1.724909	1.0363	0.0	0.9	1.5	2.3	7.4
Visibility (10m)	8760.0	NaN	NaN	NaN	1436.825799	608.298712	27.0	940.0	1698.0	2000.0	2000.0
Dew point temperature(°C)	8760.0	NaN	NaN	NaN	4.073813	13.060369	-30.6	-4.7	5.1	14.8	27.2
Solar Radiation (MJ/m2)	8760.0	NaN	NaN	NaN	0.569111	0.868746	0.0	0.0	0.01	0.93	3.52
Rainfall(mm)	8760.0	NaN	NaN	NaN	0.148687	1.128193	0.0	0.0	0.0	0.0	35.0
Snowfall (cm)	8760.0	NaN	NaN	NaN	0.075068	0.436746	0.0	0.0	0.0	0.0	8.8
Seasons	8760	4	Spring	2208	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Holiday	8760	2	No Holiday	8328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Functioning Day	8760	2	Yes	8465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

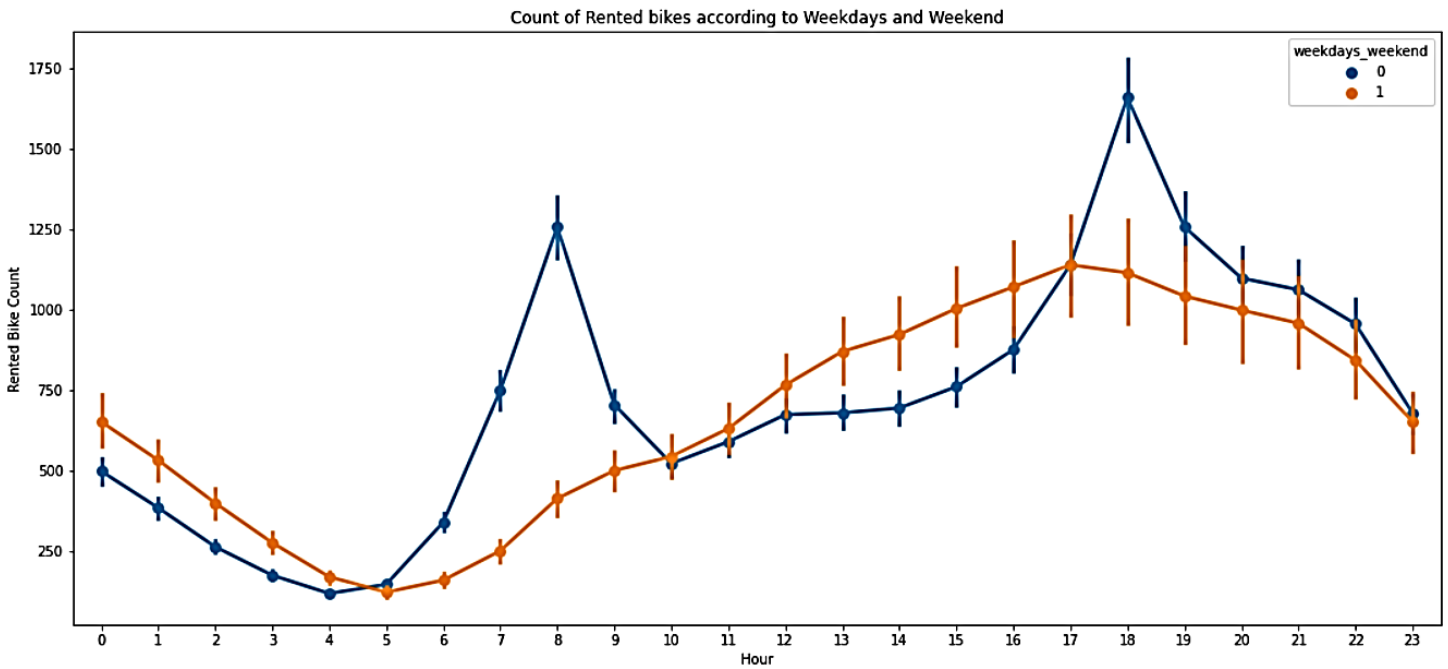
- We convert the "date" column into 3 different column i.e "year","month","day".
- We changed some columns name as Rented Bike Count, Hour ,Temperature, Humidity, wind speed, Visibility, Dew point temperature, Solar Radiation, Rainfall, Snowfall, Seasons ,Holiday ,Functioning Day ,month, weekdays_weekend.
- Contains three categorical features ‘Seasons’, ‘Holiday’, and ‘Functioning day’, Has one date column(contains date,year,month).
- Remaining 10 columns are numeric types.

Count of Rented Bike according to Month



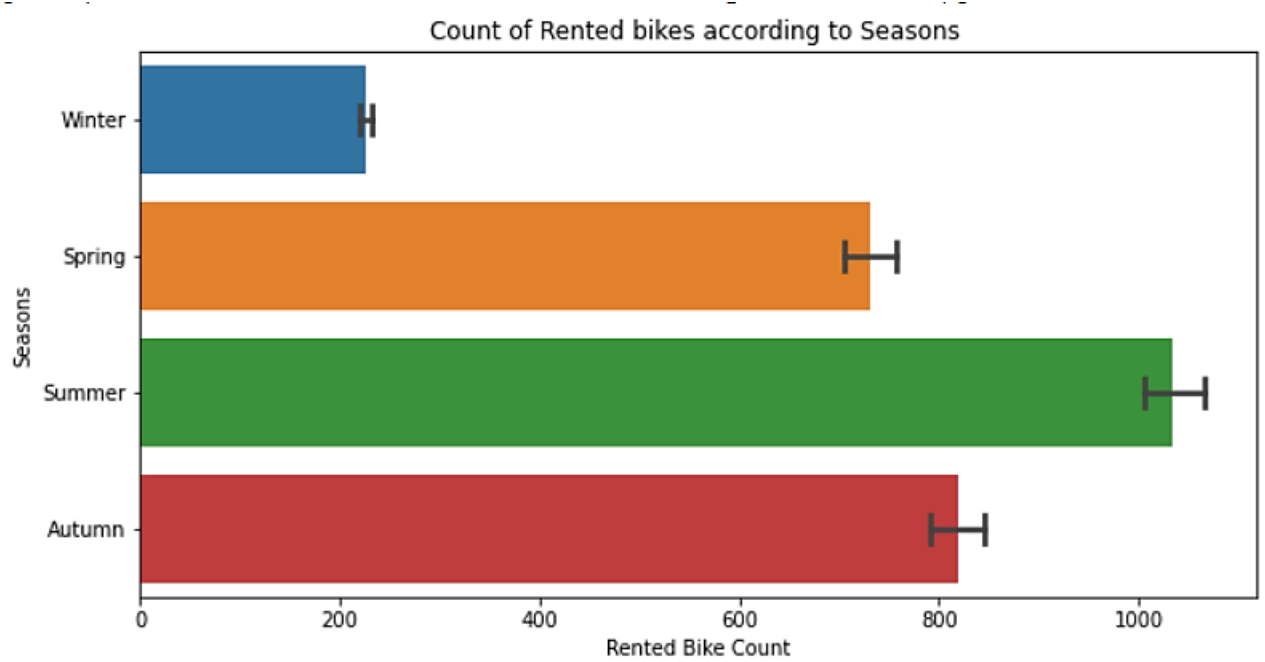
- The demand of the rented bike is high from the month 5 to 10 that is from may to October.
- The highest demand will be on summer season.

Count of Rented Bikes according to Weekdays and Weekend



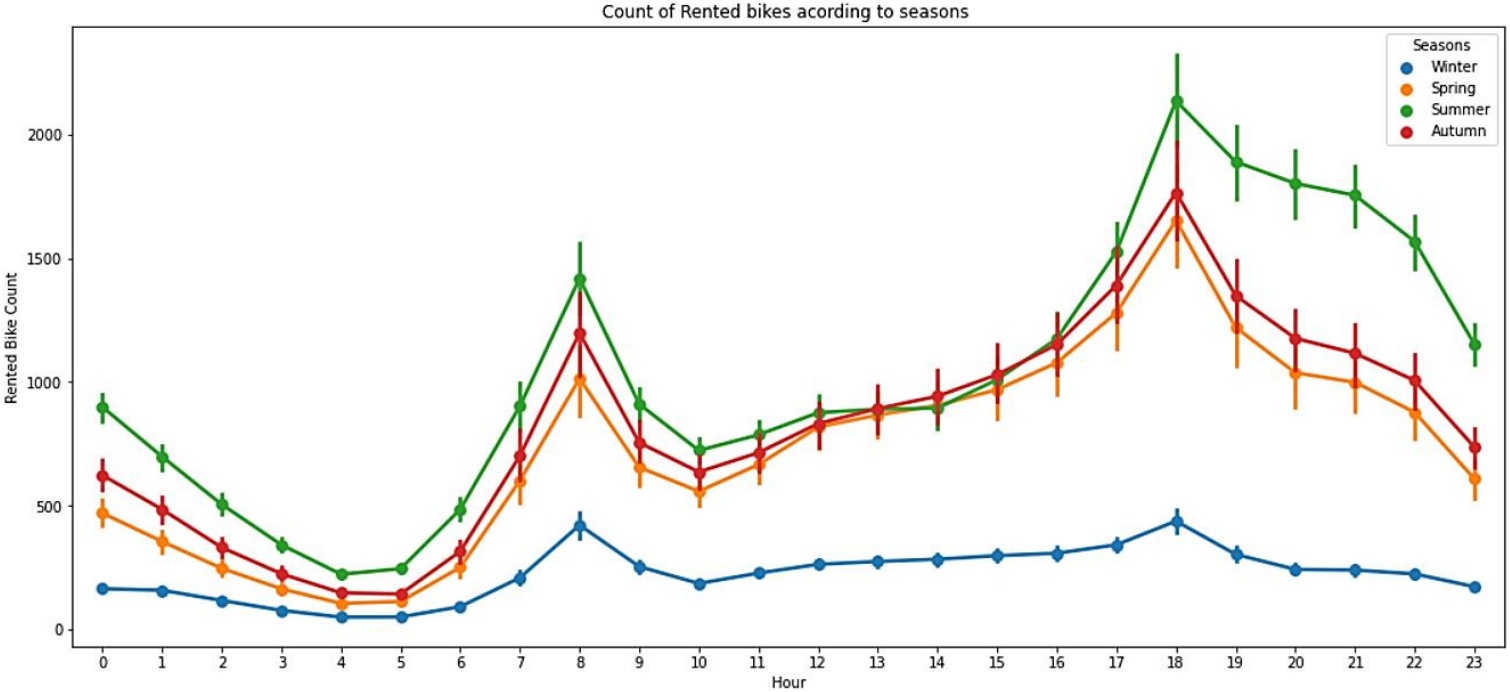
- The blue line(0) represents the weekdays and the orange line(1) represents the weekend.
- From this graph we observe that people are using rented bikes between 7AM - 9AM and 5PM- 7PM. That means they use bikes for reaching their office.
- There will be more demand during weekdays compared to weekends.

Count of Rented bikes according to Seasons



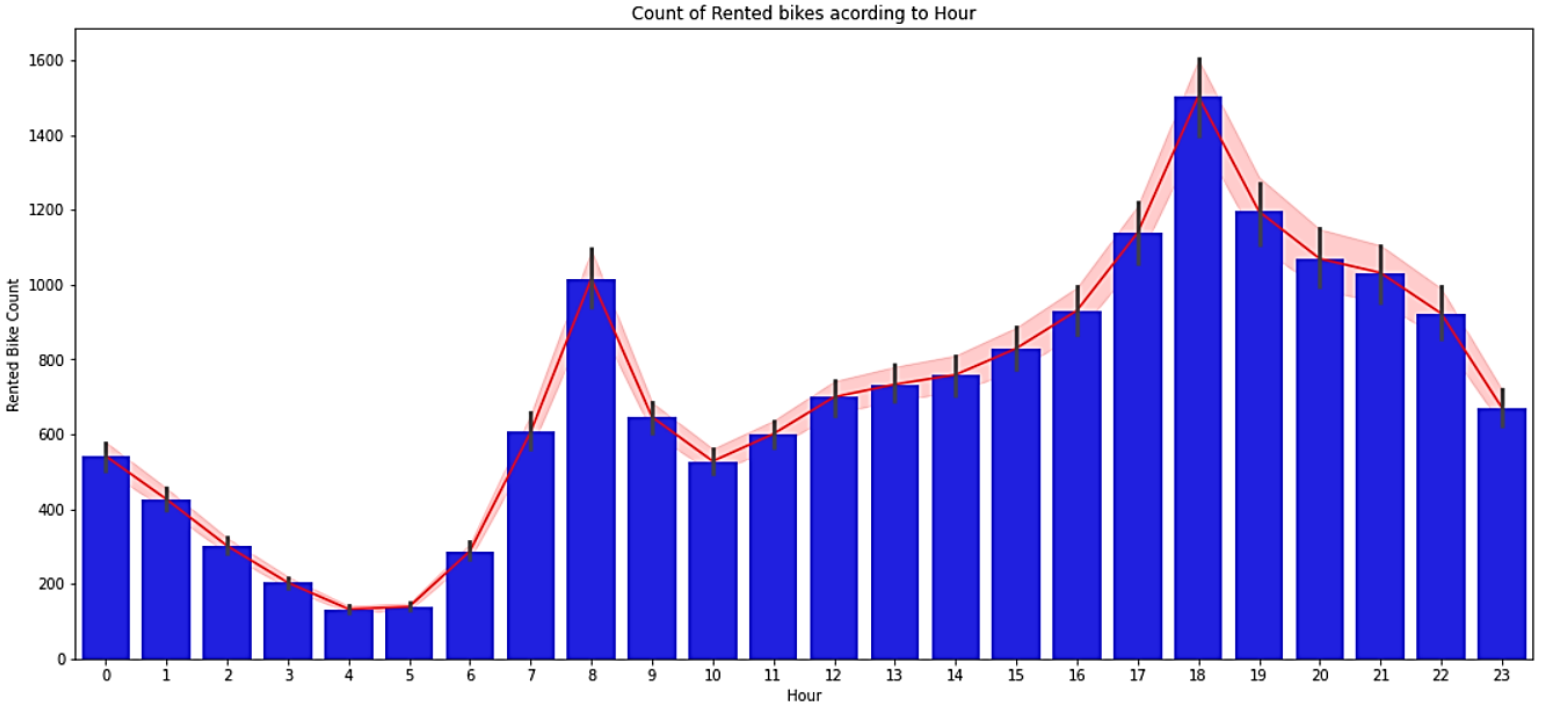
- ✓ People use rented bikes heavily in the summer season and less in the winter season because of snowfall.
- ✓ In the summer and autumn seasons, people use the rental bikes more compared to other seasons.

Count of Rented bikes According to Seasons with Hour



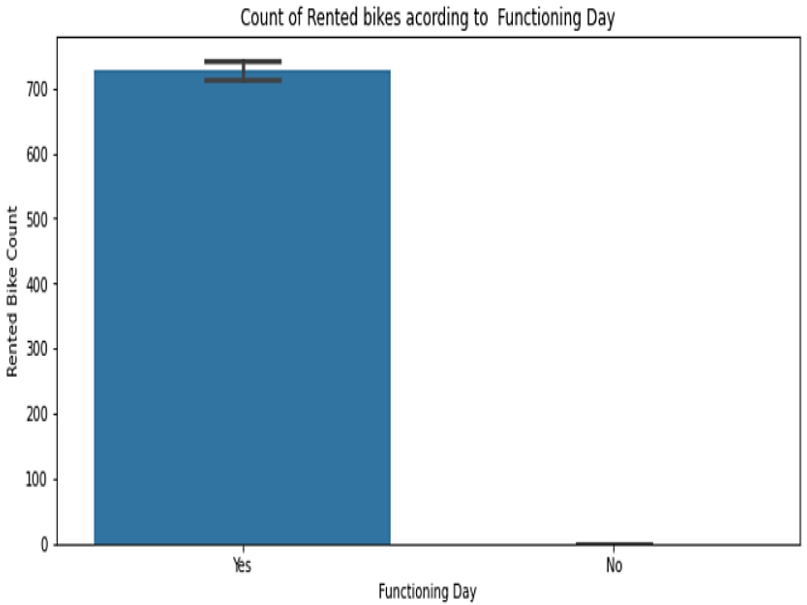
- By the above analysis, we can say that people use high rented bikes normally from 7 AM–9 AM and 5 PM–7 PM in all the seasons because of office hours.
- In winter season the use of rented bike is very low because of snowfall.

Count of Rented Bikes according to Hour

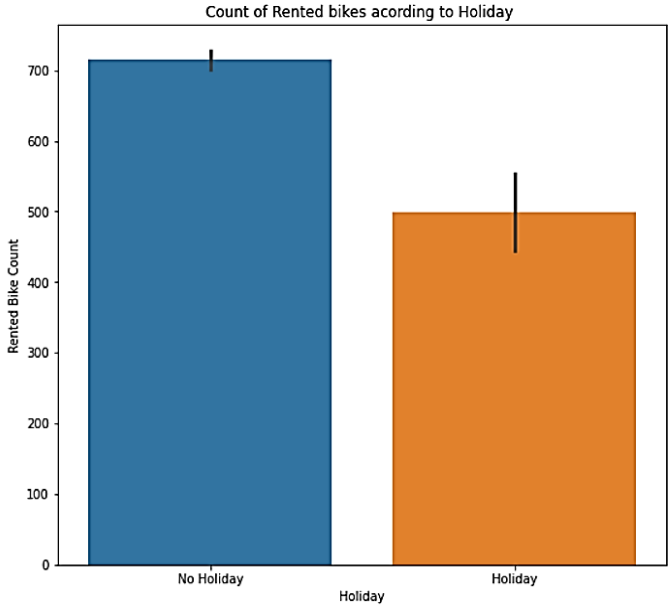


- Generally people use rented Bikes mostly for transits from 7am to 9am and 5pm to 7pm.

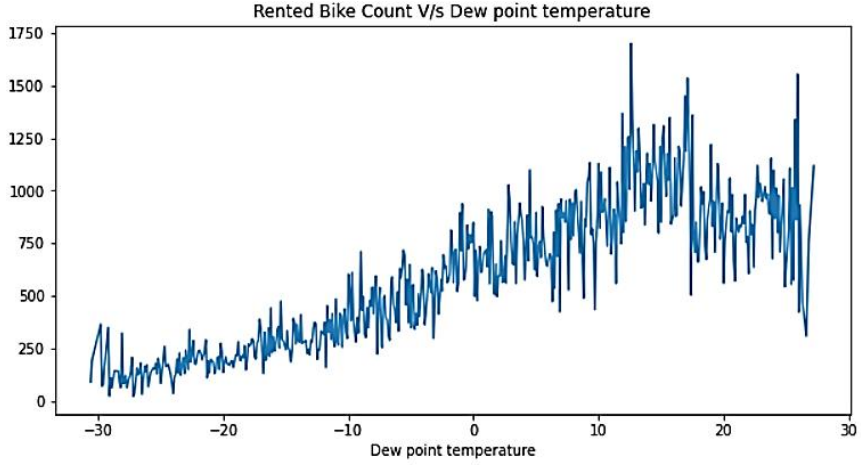
Count of Rented Bike V/s Functioning day

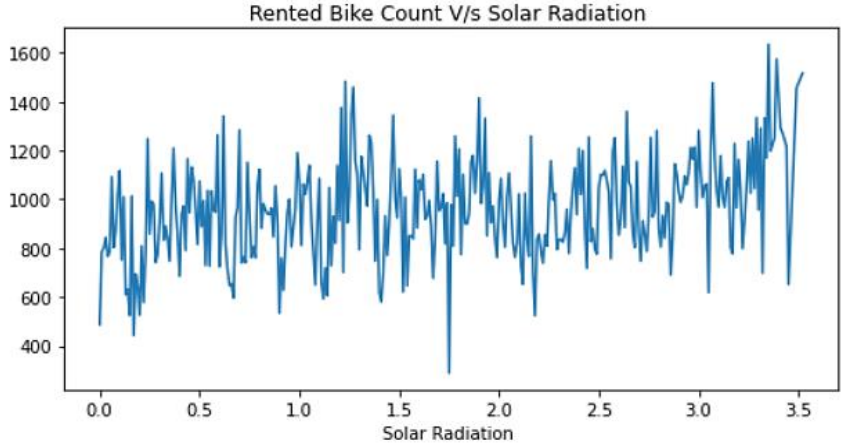


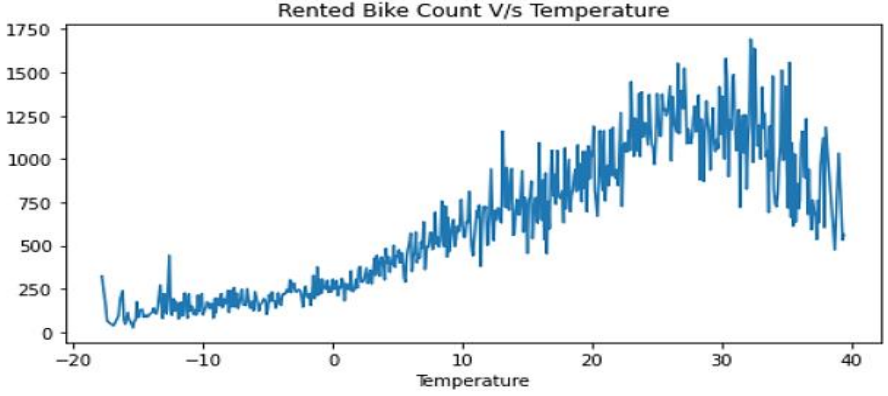
Count of Rented Bike V/s Holiday



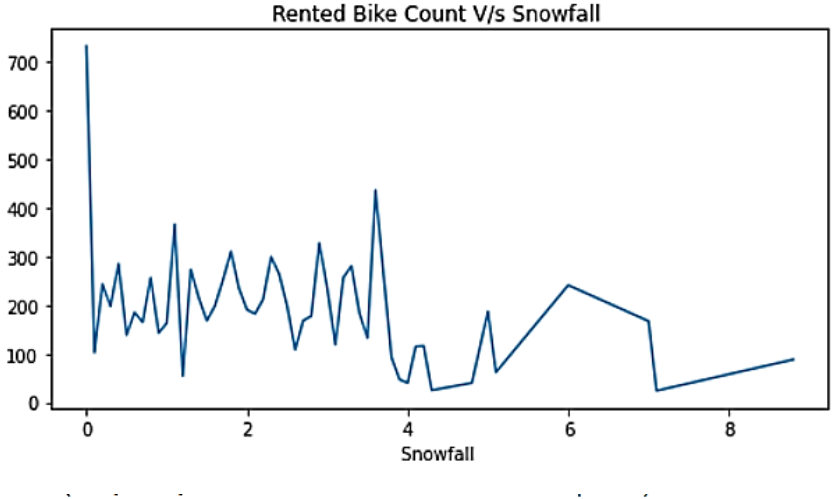
- The above graph shows clearly that people will use only rented bikes on functioning days and not on non-functioning days.
- People use the rented bikes while working and less on holiday days.

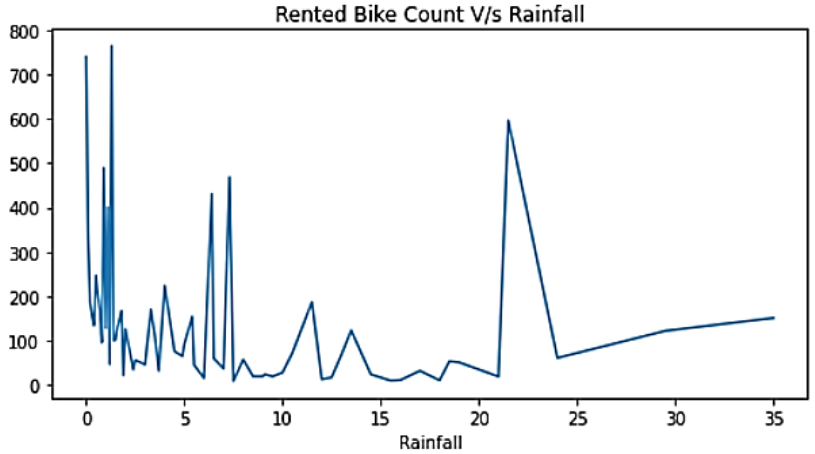


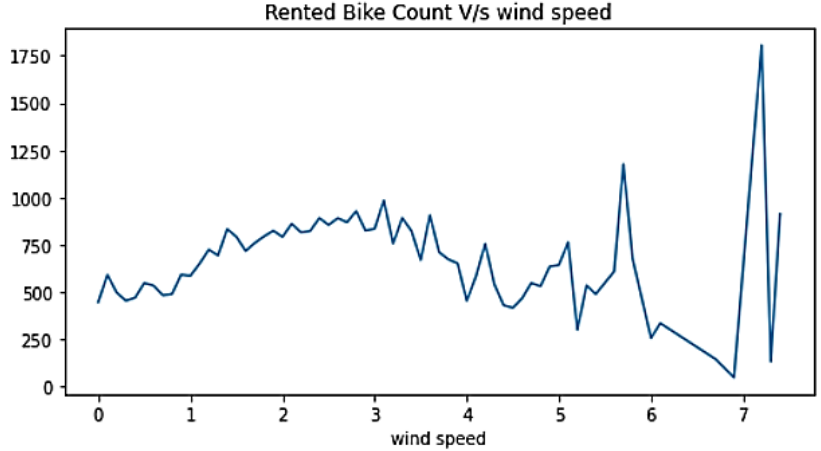




- People like to ride bikes when it is pretty hot around 25℃ in average.
- The amount of rented bikes is huge, when there is solar radiation, the counter of rents is around 1000

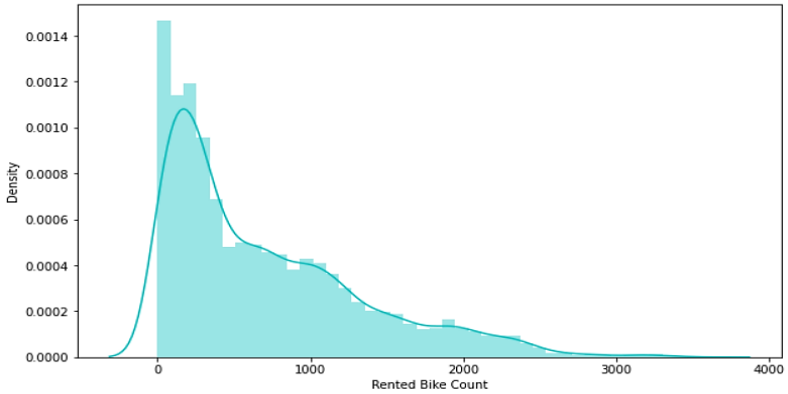


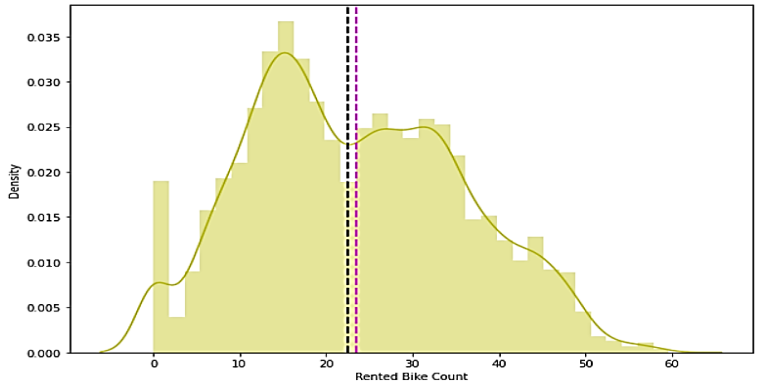


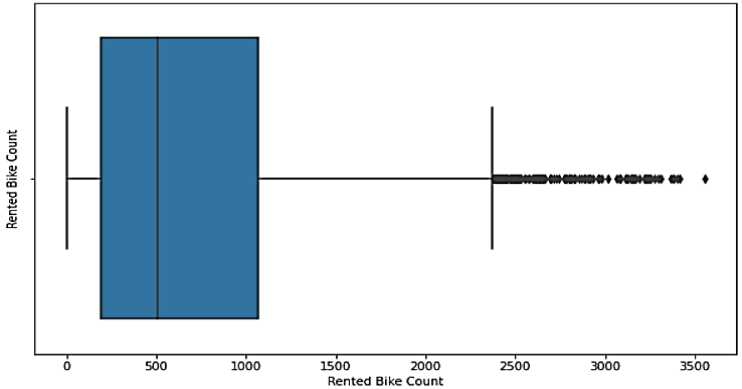


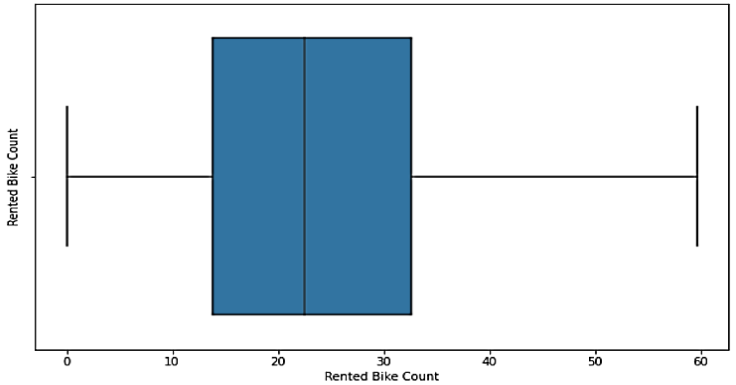
- In the snowfall season, people will not like to rent a bike, and we can see the total bike count in the graph, which is very low.
- In the rainfall graph, there is an increase in the bike count at 20 mm.
- In the wind speed graph, there is a constant bike count and there is an increase in 7 m/s.

Removing outliers



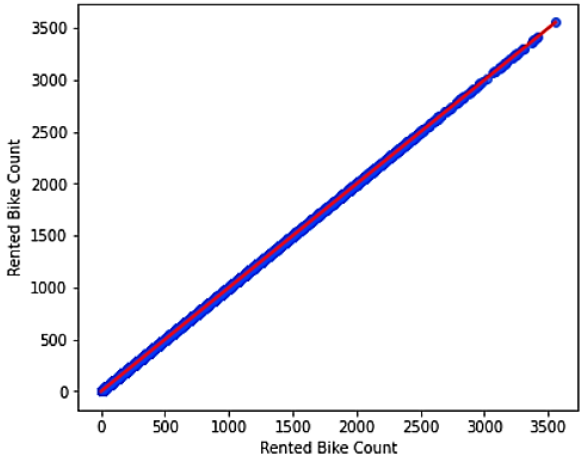


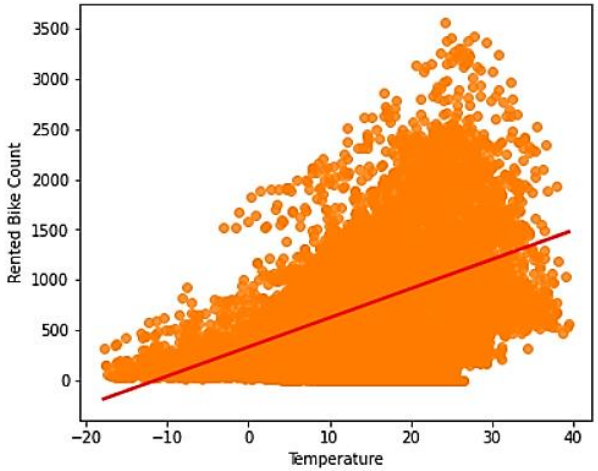


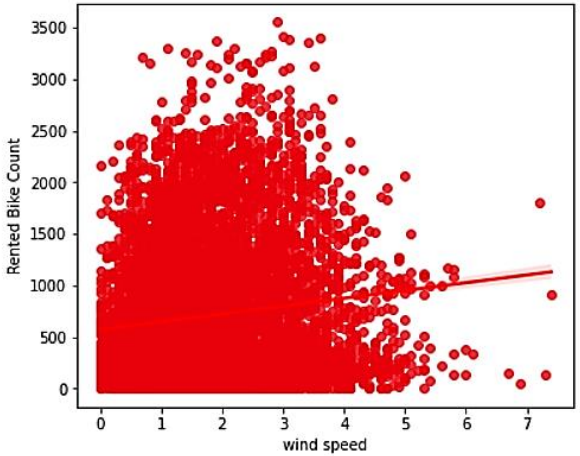


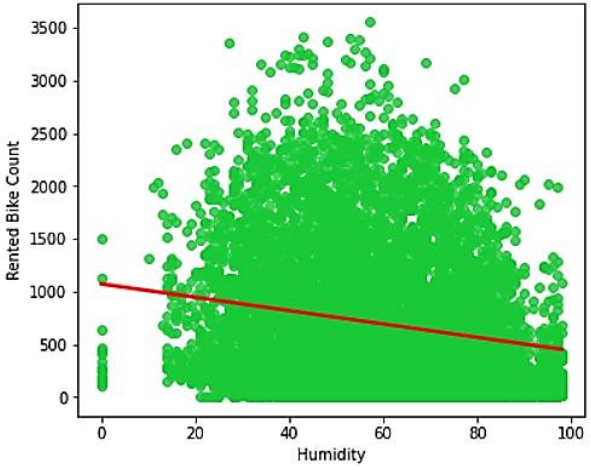
- By the above visualization, we can see there are many outliers. We have to remove it for further analysis.
- Applied square root to the skewed Rented Bike Count, here we get an almost normal distribution.
- After normalizing, the column outliers are also eliminated.

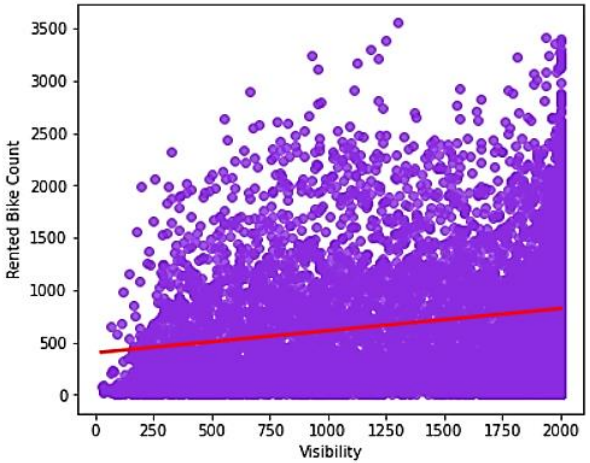
REGRESSION PLOT

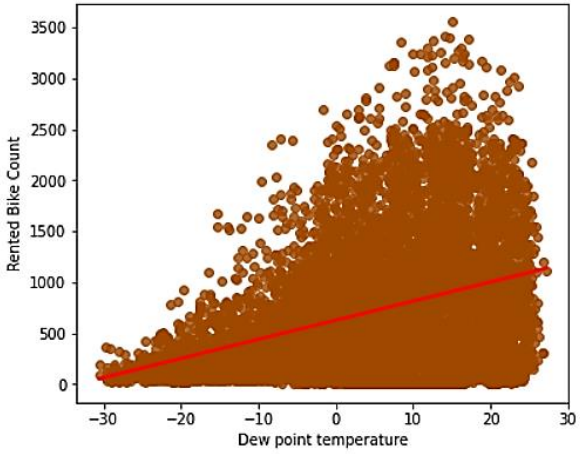


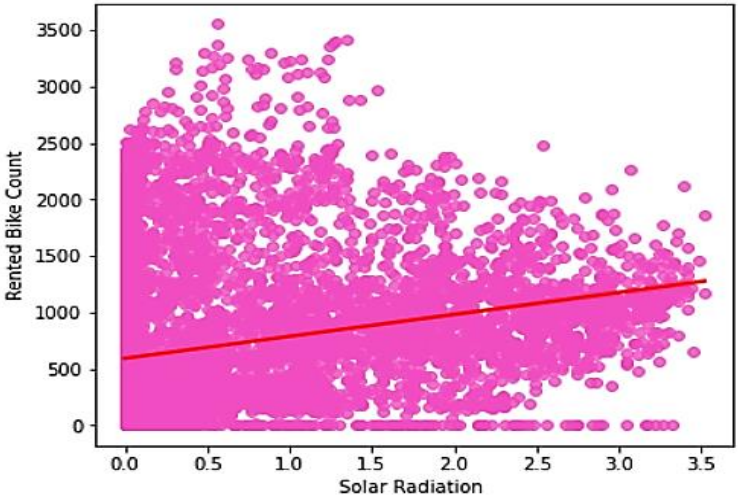


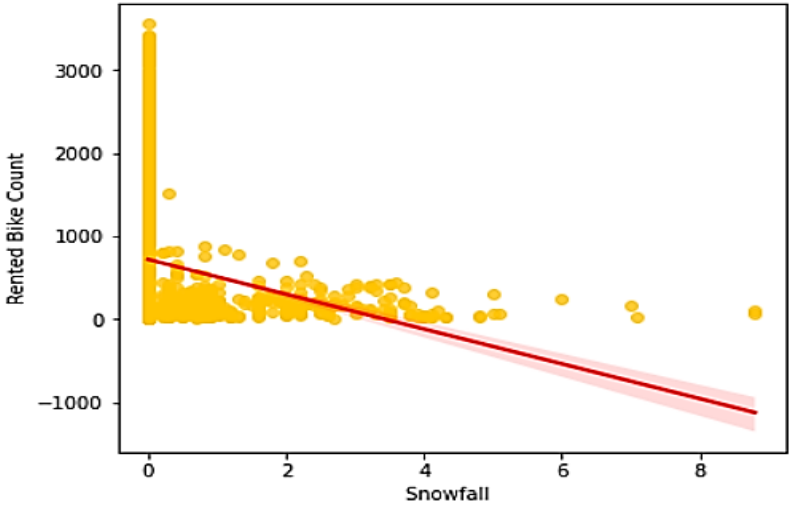


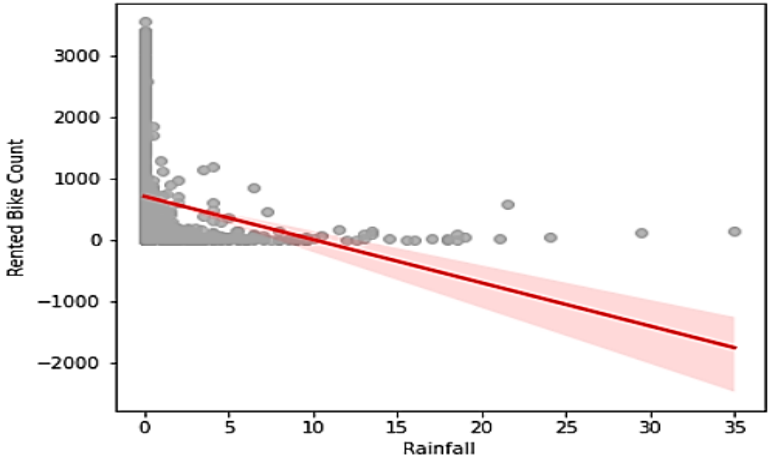






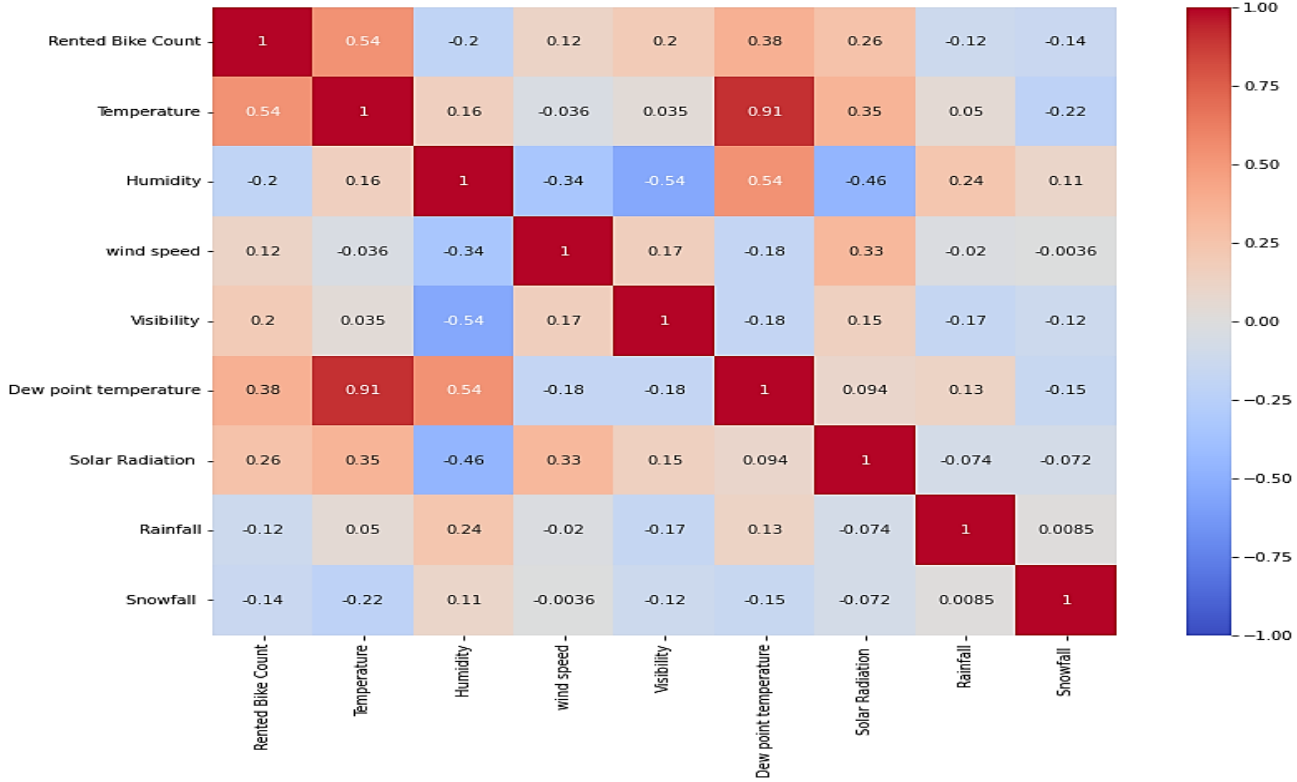






- Regression plot of a numerical feature 'Temperature' ,'Wind speed', 'Visibility', 'Dew point temperature', and 'Solar Radiation' are bike count increases with these features.
- 'Rainfall' ,'Snowfall', and 'Humidity'. These features are negatively related to the rented bike count, which means the rented bike count decreases when these features increase.

CORRELATION HEATMAP



- Variables like Dew Point Temperature, and Temperature was highly correlated. So we drop the Dew Point Temperature and check for multicollinearity.
- There is no multicollinearity in the data.

IMPLEMENING ALGORITHMS

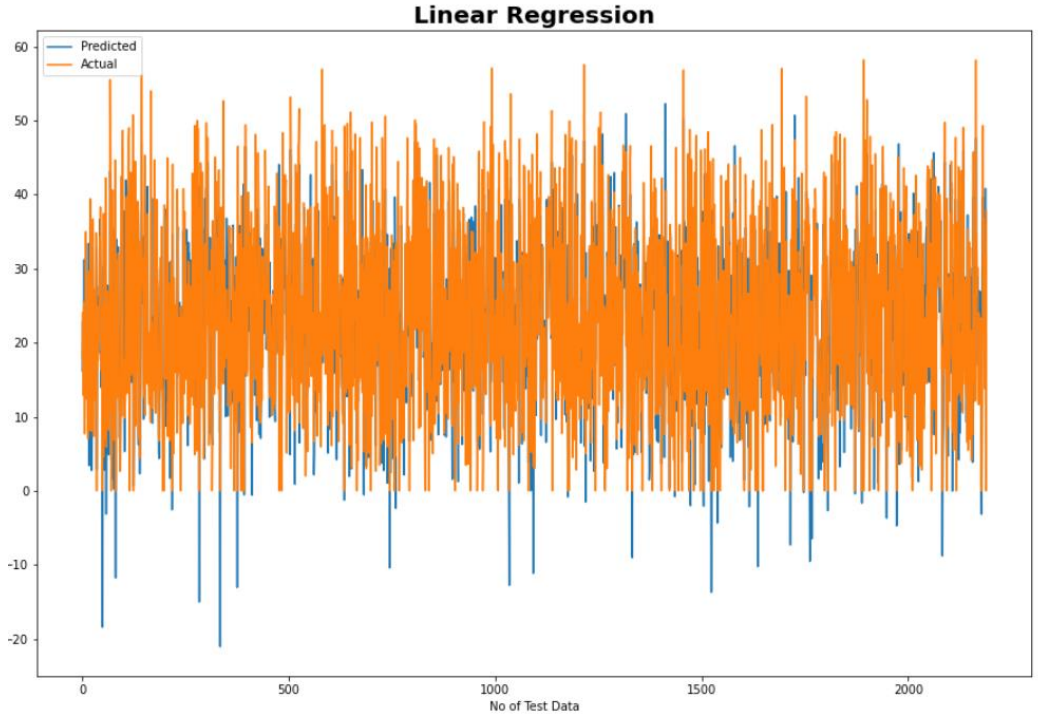
LINEAR REGRESSION

Evaluating the model on train set

```
MSE : 35.07751288189292
RMSE : 5.9226271942350825
MAE : 4.474024092996788
R2 : 0.7722101548255267
Adjusted R2 : 0.7672119649454145
```

Evaluating the model on test set

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



- Linear regression is performing well and the R2 score for the test set is 0.78.

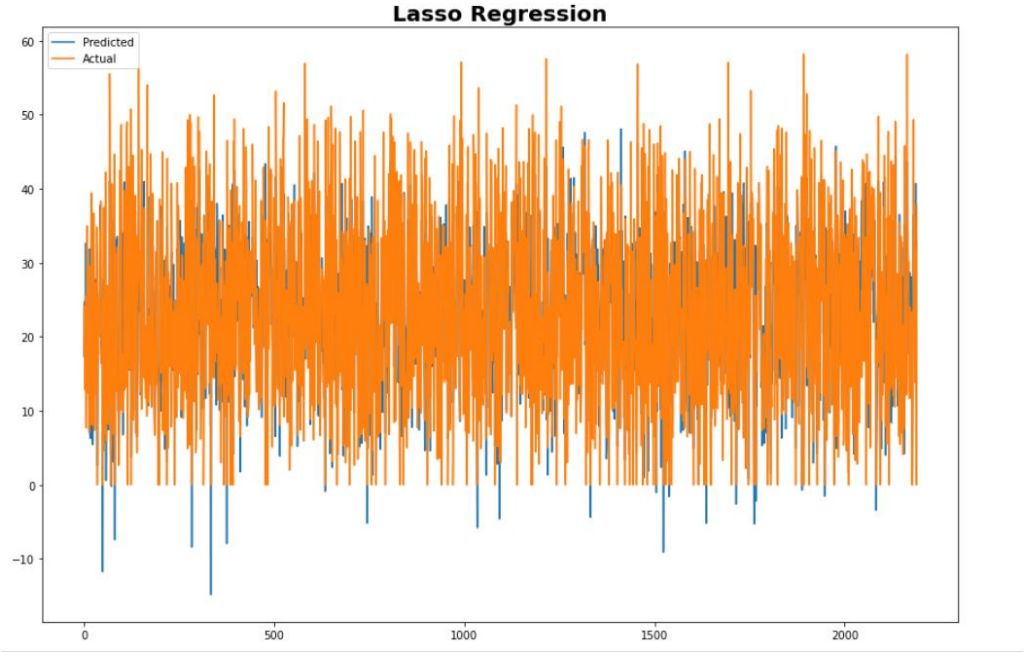
LASSO REGRESSION

Evaluating the model on train set

```
MSE : 41.48012492751929
RMSE : 6.440506573827815
MAE : 4.960430531038622
R2 : 0.7306322353334551
Adjusted R2 : 0.7247217381629008
```

Evaluating the model on test set

```
MSE : 39.91752283290745
RMSE : 6.318031563145871
MAE : 4.91263385826569
R2 : 0.7473037178309577
Adjusted R2 : 0.7417590281661841
```



- The R2 score for the test set is 0.74. This means our Lasso model is performing well on the data.

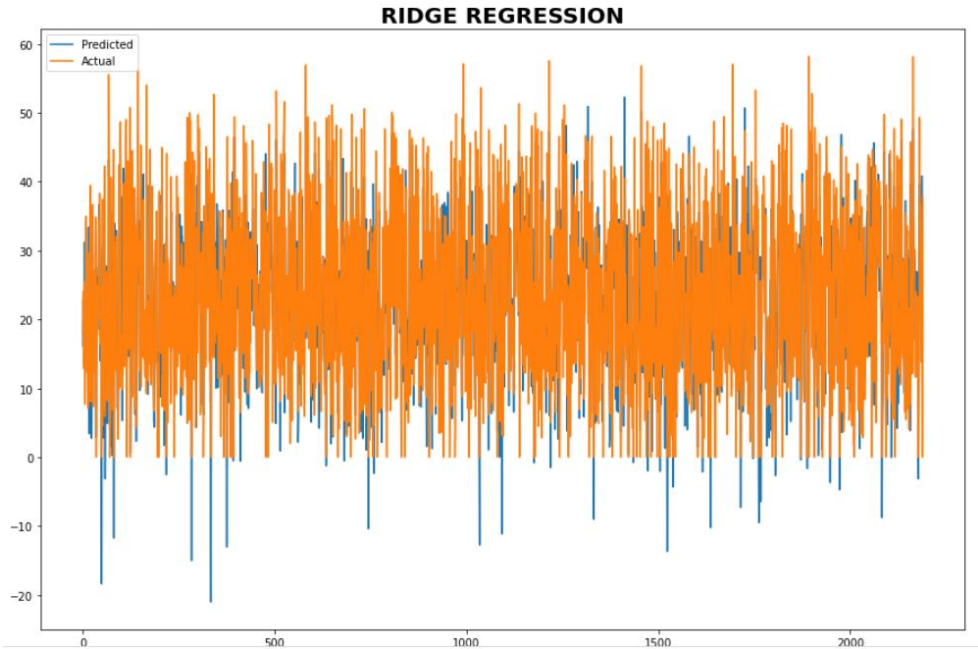
RIDGE REGRESSION

Evaluating the model on train set

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

Evaluating the model on test set

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



The R2 score is 0.78 for test data so that ridge regression is performing well on data.

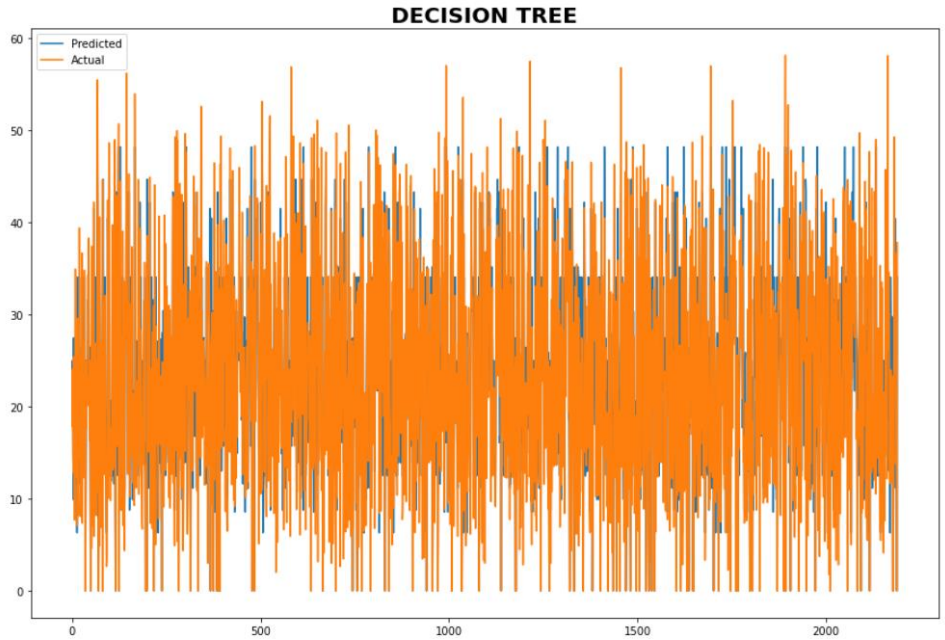
DECISION TREE

Evaluating the model on train set

```
MSE : 37.57505795046481
RMSE : 6.12984974942003
MAE : 4.54662831907701
R2 : 0.7559913480246622
Adjusted R2 : 0.727869464554237
```

Evaluating the model on test set

```
MSE : 42.06449807344292
RMSE : 6.485714923849407
MAE : 4.697733245173592
R2 : 0.7337123769187646
Adjusted R2 : 0.727869464554237
```



- The R2 score for the test set is 0.73. This means our decision model is performing well on the data.

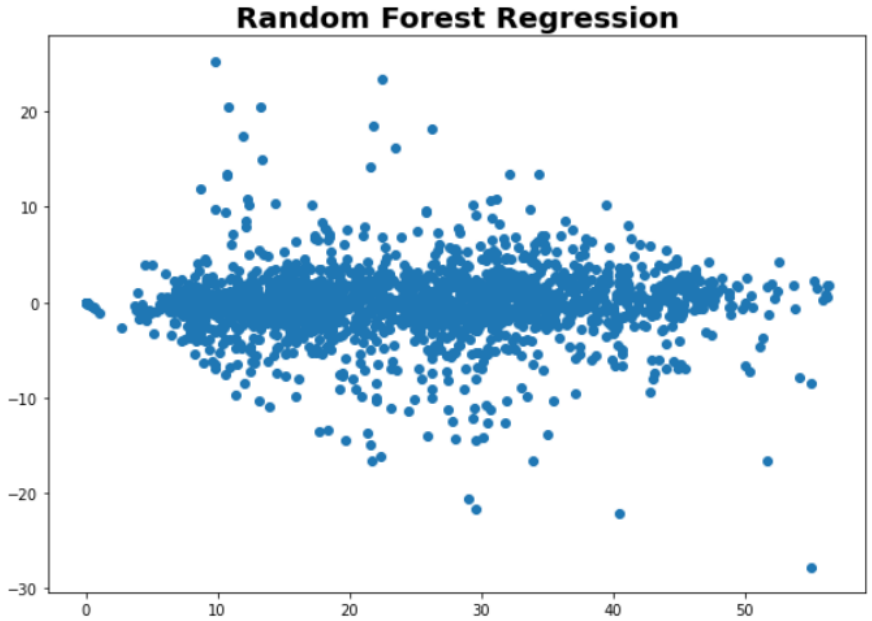
RANDOM FOREST

Evaluating the model on train set

```
MSE : 1.5852436605279605
RMSE : 1.259064597440481
MAE : 0.8001902708465634
R2 : 0.9897055869037431
Adjusted_R2 : 0.9894797057573733
```

Evaluating the model on test set

```
MSE : 12.909717427225228
RMSE : 3.5930095222842406
MAE : 2.2119105042842446
R2 : 0.9182755500292877
Adjusted R2 : 0.9164823431438426
```



•The R2 score for the test set is 0.92. This means our random forest model is performing well on the data and has more accuracy among all the models.

Comparing results of all the 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	4.960	41.480	6.441	0.731	0.72
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Dicision tree regression	4.547	37.575	6.130	0.756	0.73
	4	Random forest regression	0.800	1.585	1.259	0.990	0.99
Test set	**0**	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	4.913	39.918	6.318	0.747	0.74
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Decision tree test	4.698	42.064	6.486	0.734	0.73
	4	Random forest regression	2.212	12.910	3.593	0.918	0.92

❖ All our models are performing well on the data set and when compared to all, random forest regression has the highest R2 score of 0.92. So we can say that the Random Forest module is the best of all.

Conclusion

- Peak hours for rented bikes are between 7AM and 9 AM in the morning and 5 PM to 7PM in the evening, which suggests that the bikes are rented mostly by office-going people.
- People preferred to rent bikes in the morning rather than in the evening.
- Demand for rented bikes is high during no-holiday and functioning days. Bike demand was high during spring, summer, and autumn due to the beautiful weather.
- When the rainfall was less, people would have booked more bikes, except in a few cases.
- People seem to prefer biking in moderate to high temperatures and when it's a little windy.
- The temperature, hours, and humidity are the most important features that positively drive the total rented bike count.
- We used different types of regression algorithms to train our model, like: Linear Regression, Regularized linear regression (Ridge and Lasso), random forest regression, and decision tree ,where we tuned the parameters of the decision tree. Out Of them, random forest regression gave the best result.

Thank You!