Capstone Project- 2

Supervised ML-Regression
(Bike Sharing Demand Prediction)

By **Team GodSpeed**

Members:

Akash Jaiswal

Jagdish Mahato

Abhishek Rana

Kanika Raj

CONTENTS

- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATORY DATA ANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING ALGORITHMS
- CHALLENGES
- CONCLUSIONS

BUSINESS UNDERSTANDING

- **Bike rentals have become a popular service in recent years, and it seems like people are using them more often. With relatively cheaper rates and the ease of pick-up and drop-off at people's convenience, this is what makes this business thrive.**

- **This service is mostly used by those who have no personal vehicles.**

- **And also, some people prefer rental bikes to avoid congested public transport.**

- **Therefore, for the business to strive and profit more, it has to always be ready to supply an adequate number of bikes at various locations to meet the demand.**

- **Our project goal is to pre-plan the prediction of bike count values that can be a handy solution to meet all the demands.**

Problem Statement

Currently Rental bikes have been introduced in many urban cities for the enhancement of mobility and comfort. It is important to make the rental bikes 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 the bike count required at each hour for the stable supply of rental bikes.

Data Pipeline

- **Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.**

- **Data Preprocessing: In this part we went through each attributes and encoded the categorical features.**

- **Model Creation: Finally in this part we created the various models.**

- **These various models are being analysed and we tried to study various models to get the best performing model for our project.**

Data Description:

Independent variables:

- Date : year-month-day
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day – No Func(Non Functional Hours), Fun(Functional hours)

Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

DATA SUMMARY

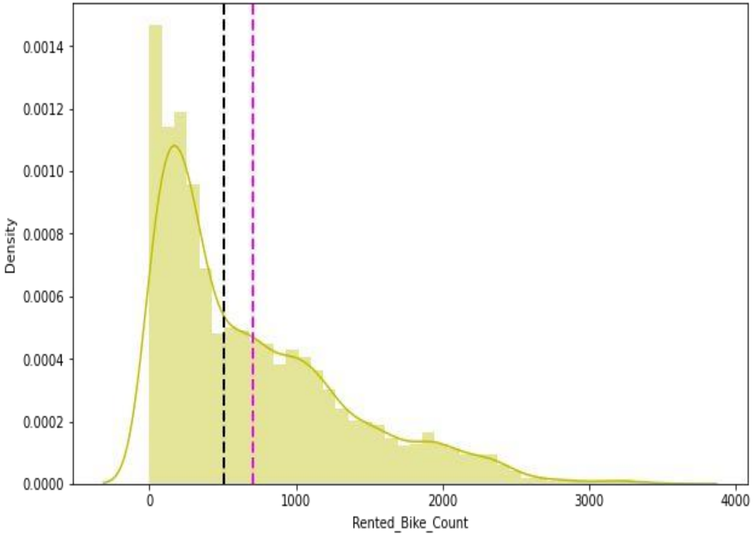
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	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

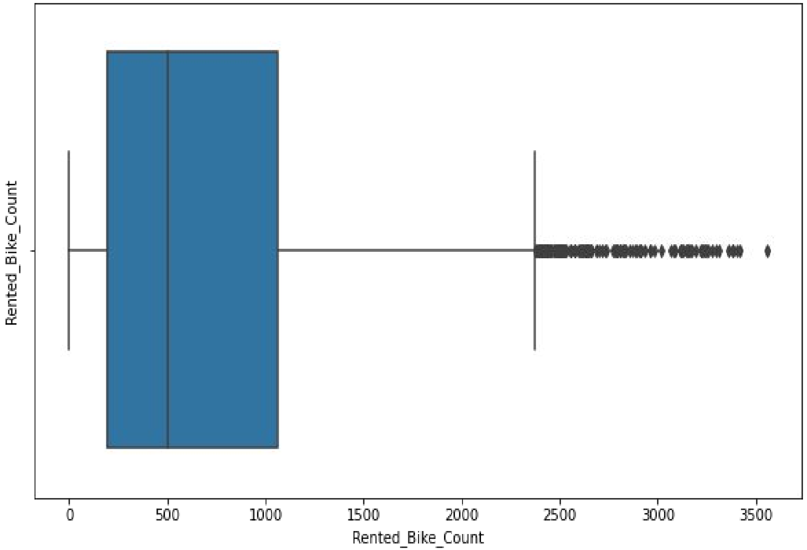
- **This Dataset contains 8760 lines and 14 columns.**
- **Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.**
- **One Datetime features ‘Date’.**
- **We have some numerical variables such as temperature, humidity, wind, visibility, dew point temperature, solar radiation, rainfall, snowfall, which describe the environmental conditions at that particular hour of the day.**

INSIGHTS FROM OUR DATASET

- **There are No Missing, Duplicate and null Values present in the SeoulBike Dataset.**
- **And finally we have 'rented bike count' variable which we need to predict for new observations.**
- **The dataset SeoulBike shows hourly rental data for one year i.e. 1 Dec 2017 to 31 Nov 2018 (365 days).we consider this as a single year data.**
- **So we convert the "date" col. into 3 different columns 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’.**

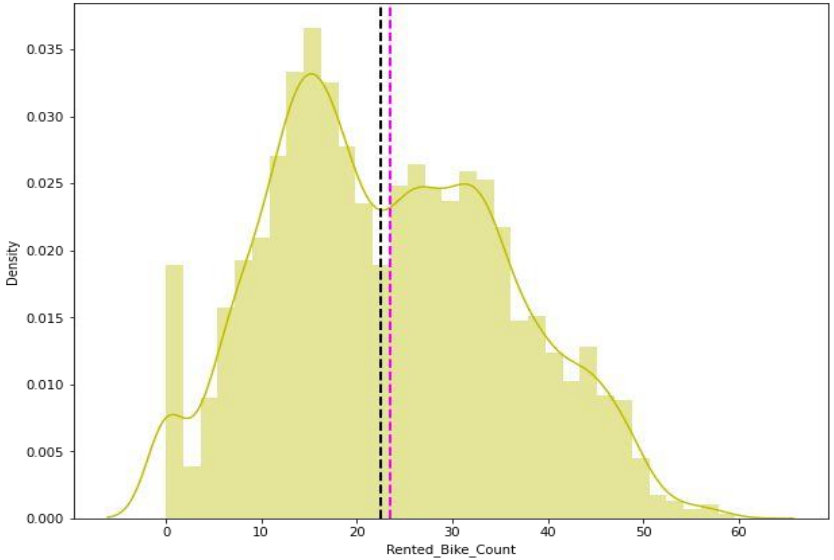
ANALYSIS OF RENTED BIKE COLUMN

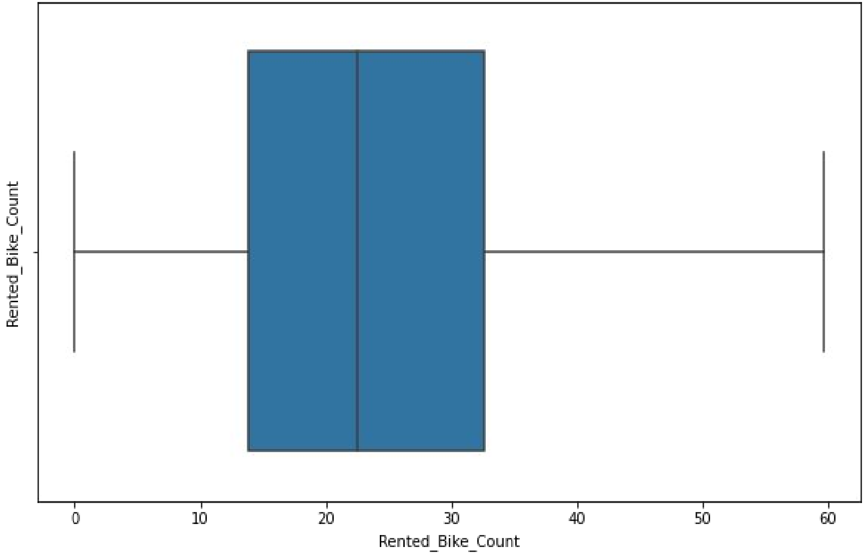




- The above graph shows that Rented Bike Count has **moderate right skewness.**
- The above right side boxplot shows that we have detected **outliers** in Rented Bike Count column.
- Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should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

ANALYSIS OF RENTED BIKE COLUMN

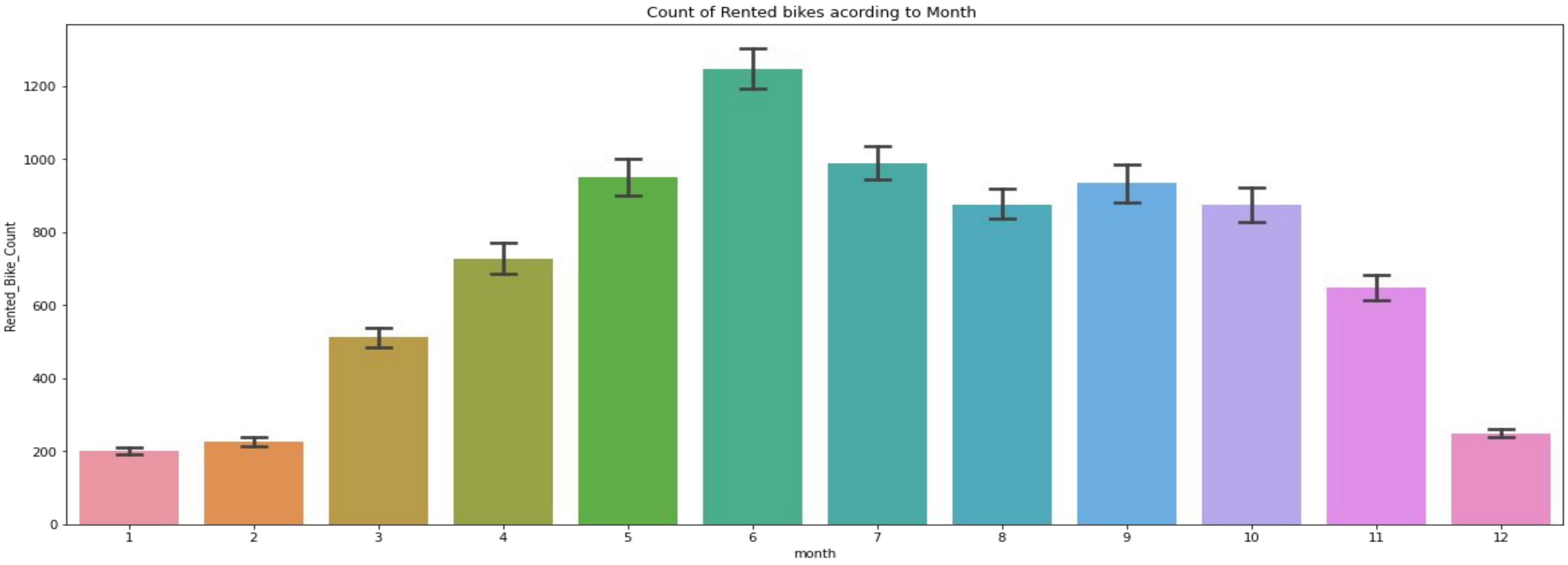




- **As we can see in the histogram graph, after applying Square root to the skewed Rented Bike Count, we get an almost normal distribution.**

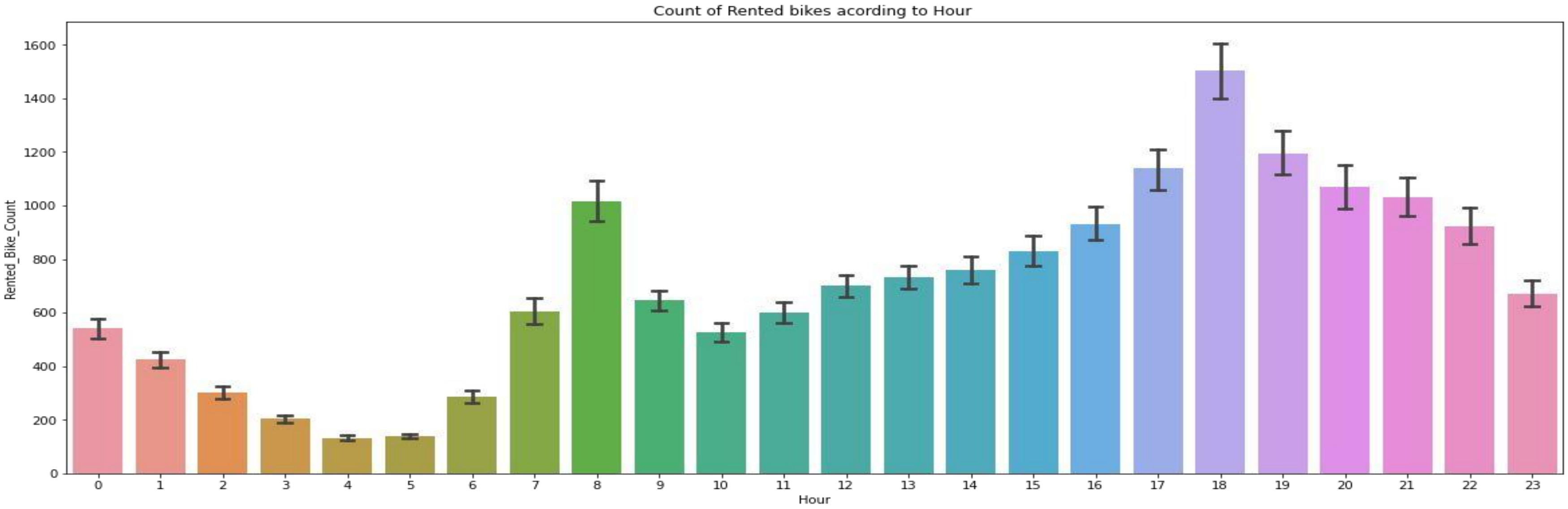
- **After applying Square root operation, there are no outliers present in the Rented Bike Count column as shown above in right-side box plot.**

ANALYSIS OF MONTH VARIABLE



➢ From the above **bar plot**, we can clearly show that the demand for the **rented bike** is high from the **5th** to the **10th month** as compared to other months, and these months represent the **summer season**.

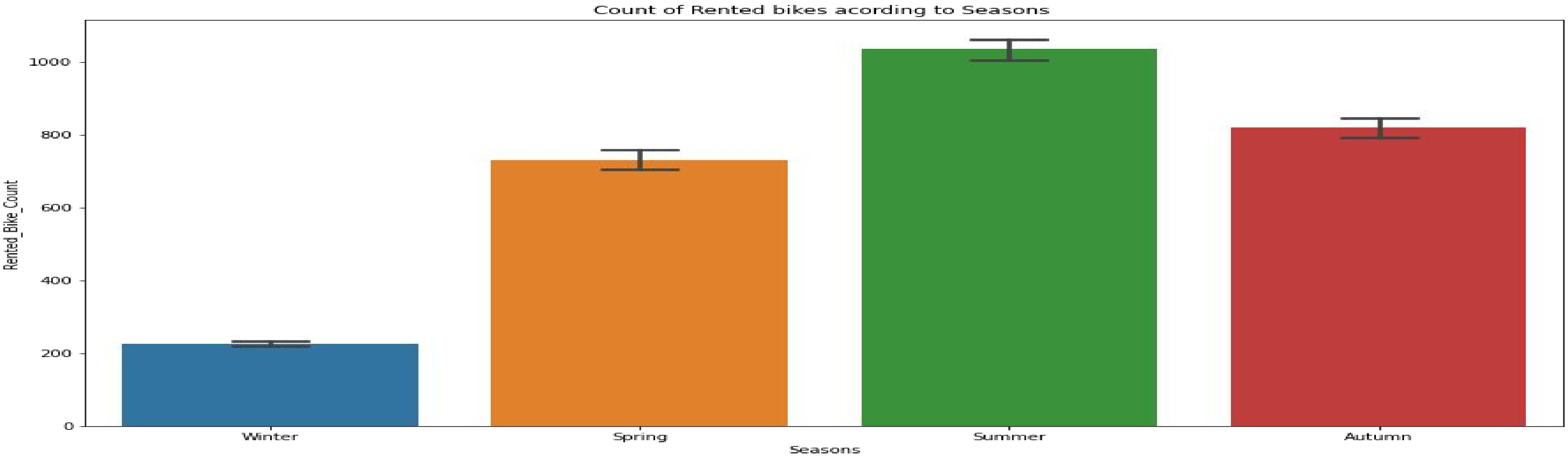
ANALYSIS OF HOUR VARIABLE



- The above **bar plot** describes the use of **rented bike count** according to the **'hour'** and the data shown for a year or 365 days.

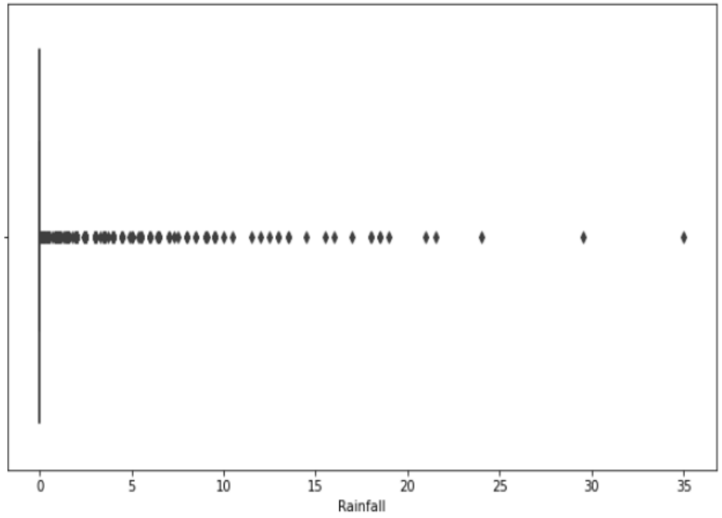
- It is clear from the graph that generally **people** use rented bikes during their **working hours** from 7 am to 9 am and 5 pm to 7 pm.

ANALYSIS OF SEASON VARIABLE

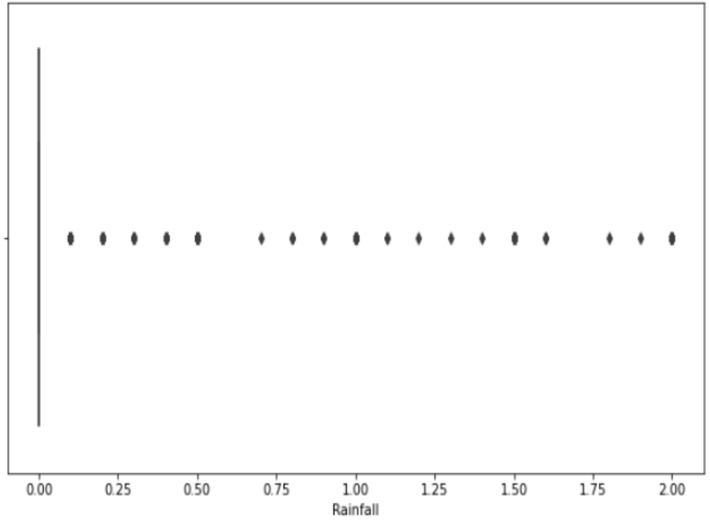


- The above **bar plot** shows the distribution of rented bike counts **seasonally.**
- And we can see that **most people** prefer to ride bikes in the **summer** and **autumn** seasons.
- Conversely, the **winter season** has the **lowest** number of rented bikes, and it may be because of **heavy snowfall.**

Detection and Treatment of outliers on numerical columns.



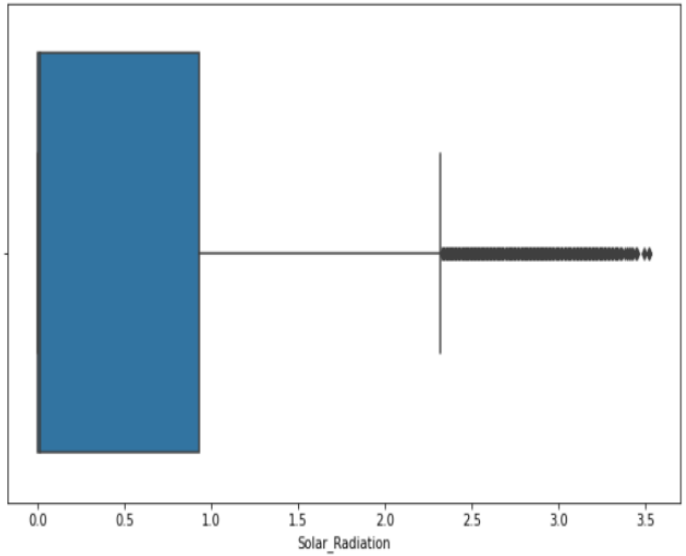
After Trimming



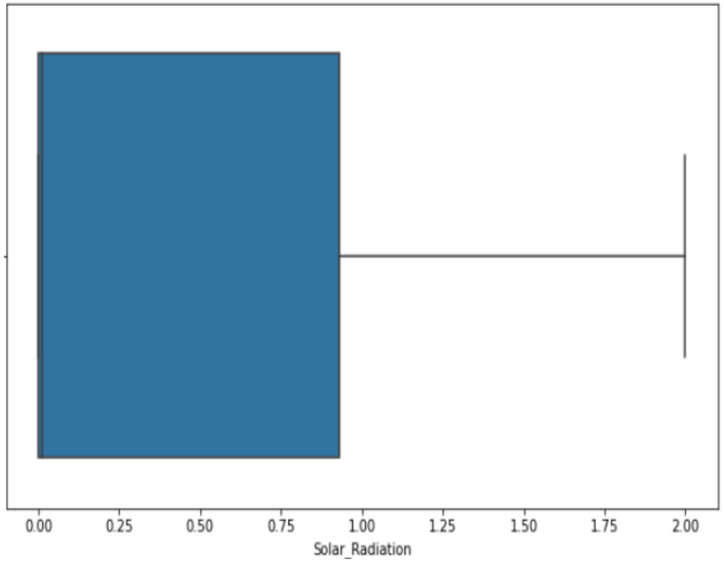
- **Trimming is one of the techniques for treating outliers in which we trim the outliers equal to the normal values of the column.**

- **In the above plot, we can see that column 'Rainfall' has many outliers reaching a maximum point of 35, but after trimming them we can see that the outliers' maximum range is capped to 2.**

Detection and Treatment of outliers on numerical columns.

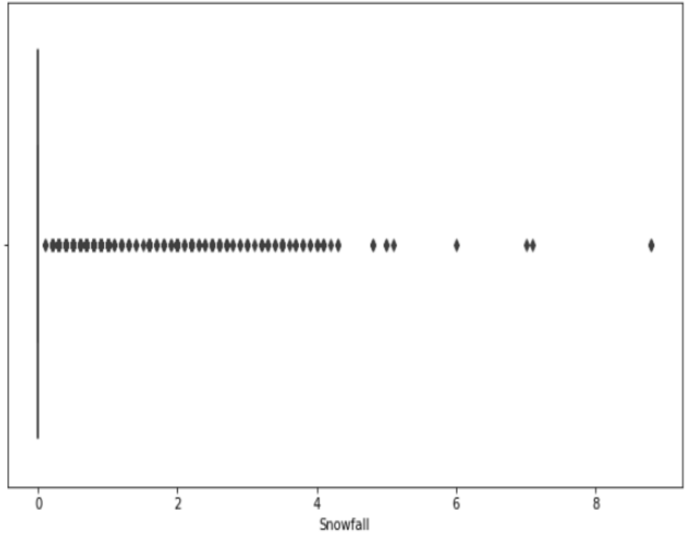


After Trimming

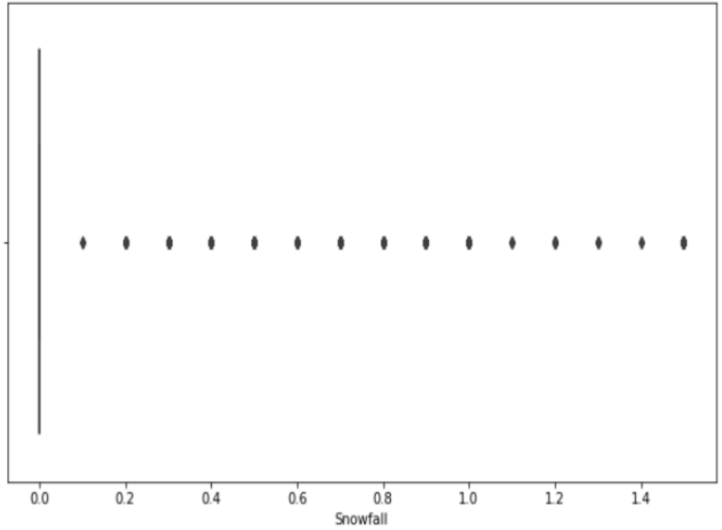


- **Trimming technique on 'Solar_Radiation' column.**
- **In the above plot, we can see that column 'Solar_Radiation' has many outliers reaching a maximum point of 3.5, but after trimming them we can see that there are no more outliers.**

Detection and Treatment of outliers on numerical columns.

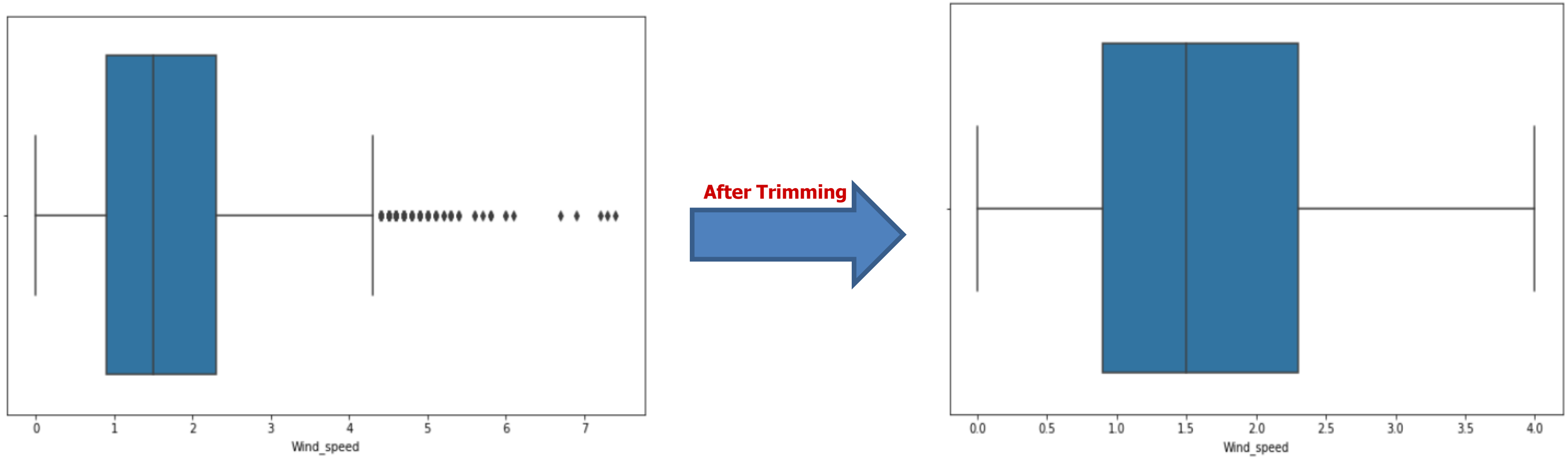


After Trimming



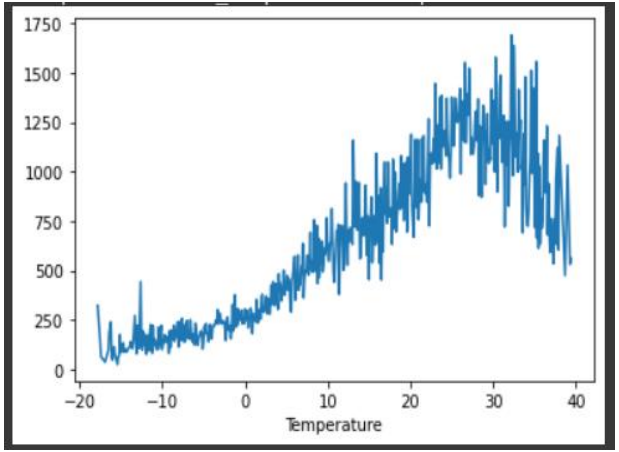
- **Trimming technique on 'Snowfall' column.**
- **In the above plot, we can see that column 'Snowfall' has many outliers reaching a maximum point of 8.5, but after trimming them we can see that the outliers are now capped to 1.5.**

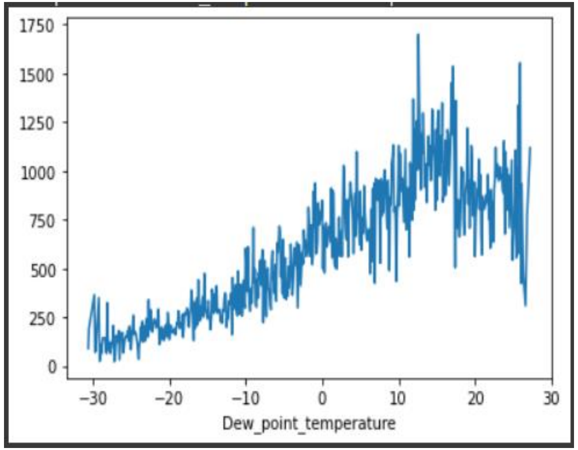
Detection and Treatment of outliers on numerical columns.

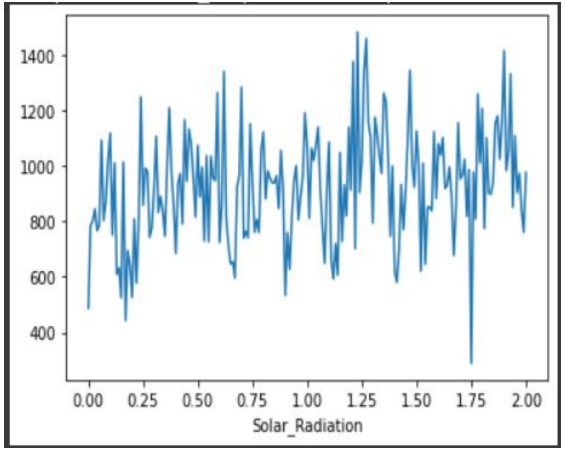


- ➢ **Trimming technique on 'Wind_speed' column.**
- ➢ **In the above plot, we can see that column 'Wind_speed' has many outliers reaching a maximum point of 7.5, but after trimming them we can see that there are no more outliers.**

NUMERICAL VS RENTED BIKE COUNT





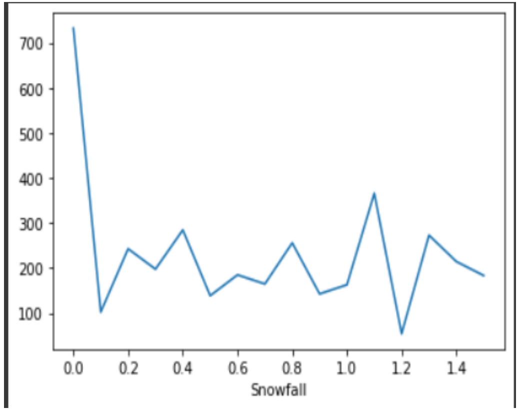


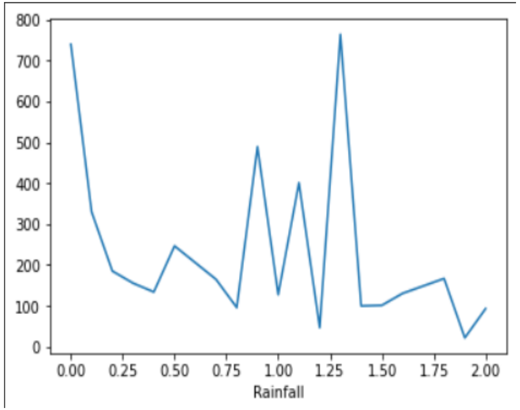
- **The above plot shows the comparison between 'Temperature', 'Dew_point_temperature', and 'Solar_Radiation' and it shows that people like to ride bikes when it is pretty hot, around 25°C to 30°C on average.**

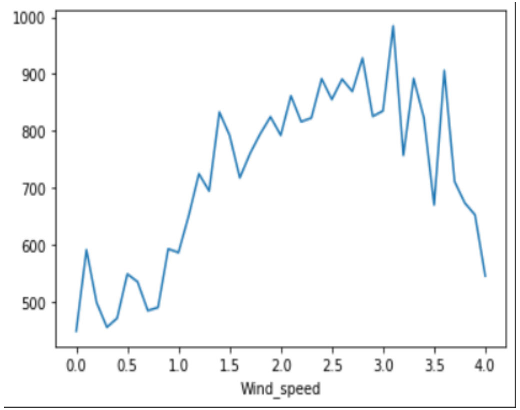
- **Also, 'Dew_point_temperature' is almost the same as the 'Temperature' because of the similar nature of the data.**

- **From the above graph, we see that the number of rented bikes is uniformly distributed when there is solar radiation.**

NUMERICAL VS RENTED BIKE COUNT







- **The above graph for 'Snowfall' represents that the number of rented bikes is very low, especially in the 4 cm range.**

- **The above graph for 'Rainfall' shows that during heavy rainfall, the demand for rented bikes does not decrease. Here, for example, even if we have 20 mm of rain, there is a big peak for rented bikes.**

- **In the 'Wind_speed' graph, the demand for rented bikes is uniformly distributed, but when the wind speed is 7 m/s, the demand for bikes increases.**

OLS REGRESSION MODEL

- **R-Squared and Adj. R-Squared are near each other. 40% of the variance in the rented bike count is explained by the model.**
- **The P values for 'Dew_point_temperature' and 'Visibility' are very high, hence they are not significant to the model.**
- **The OLS model concludes that the 'Dew_point_temperature' and 'Visibility' columns are not necessary for the model.**

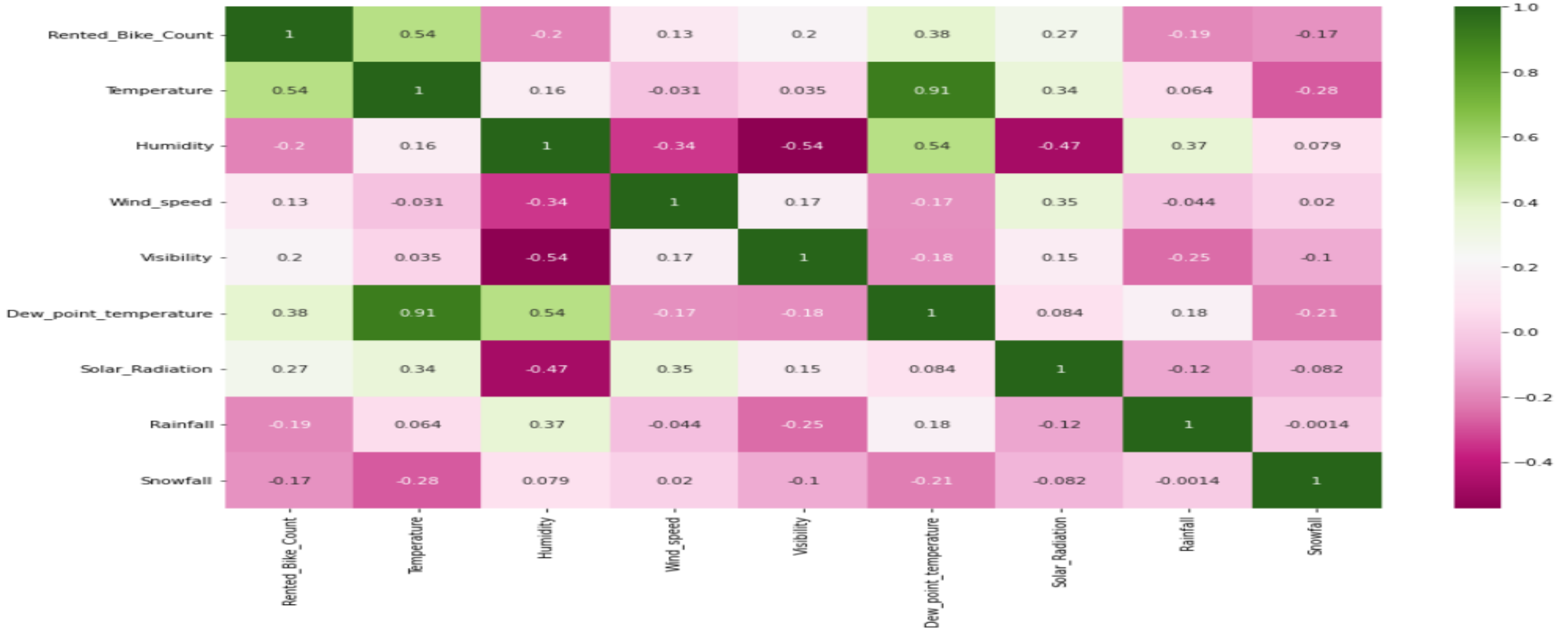
OLS Regression Results

Dep. Variable:	Rented_Bike_Count	R-squared:	0.405
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	745.9
Date:	Sat, 19 Mar 2022	Prob (F-statistic):	0.00
Time:	09:49:35	Log-Likelihood:	-66823.
No. Observations:	8760	AIC:	1.337e+05
Df Residuals:	8751	BIC:	1.337e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	650.1002	106.960	6.078	0.000	440.434	859.766
Temperature	40.1523	4.154	9.666	0.000	32.010	48.295
Humidity	-7.7780	1.200	-6.480	0.000	-10.131	-5.425
Wind_speed	61.8430	5.961	10.374	0.000	50.158	73.528
Visibility	-0.0100	0.011	-0.916	0.360	-0.031	0.011
Dew_point_temperature	-5.4120	4.393	-1.232	0.218	-14.024	3.200
Solar_Radiation	-122.2291	10.383	-11.772	0.000	-142.582	-101.876
Rainfall	-286.7998	17.903	-16.020	0.000	-321.894	-251.705
Snowfall	58.9518	23.449	2.514	0.012	12.987	104.916

Omnibus:	1000.333	Durbin-Watson:	0.348
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1702.173
Skew:	0.787	Prob(JB):	0.00
Kurtosis:	4.478	Cond. No.	3.15e+04

CORRELATION HEATMAP



- ➢ **Variables like Dew Point Temperature, and Temperature are highly correlated.**

MODEL BUILDING

- LINEAR REGRESSION
- LASSO REGRESSION
- RIDGE REGRESSION

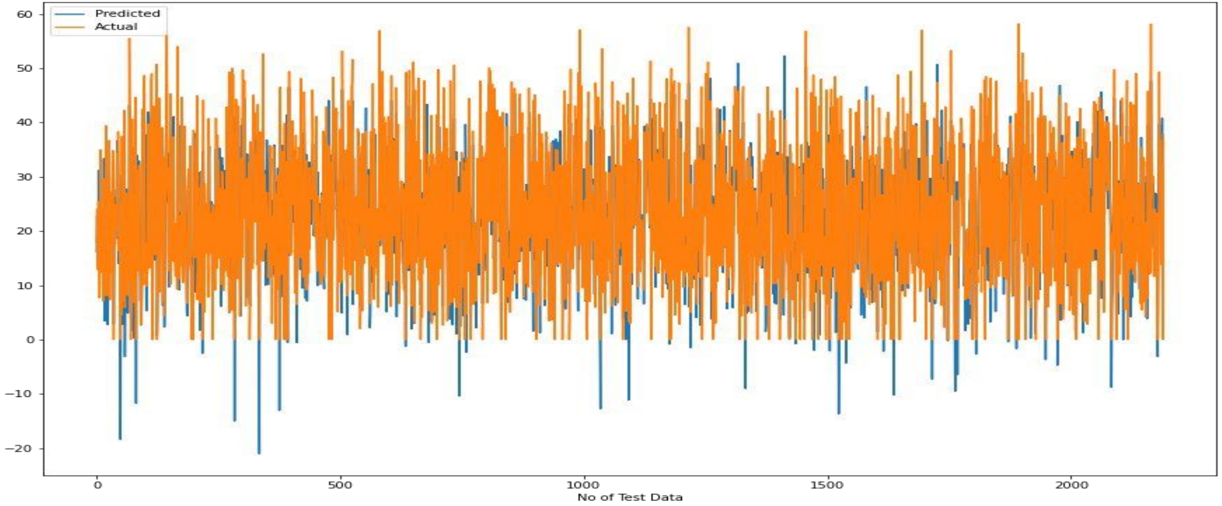
LINEAR REGRESSION

Train Set Results

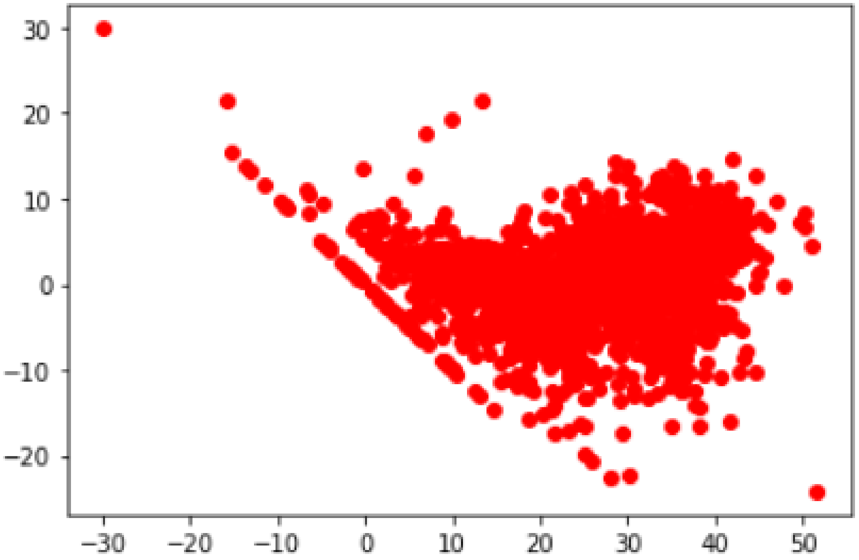
	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	4.238	30.145	5.490	0.802	0.80

Test Set Results

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	4.238	30.145	5.490	0.802	0.80



Heteroscedasticity



- In this **linear regression model**, we can see that the model is **able** to **capture** most of the data and variance, and hence the data is less and randomly scattered, and the errors are not forming any pattern.

- So, this model is **suitable** for the given dataset.

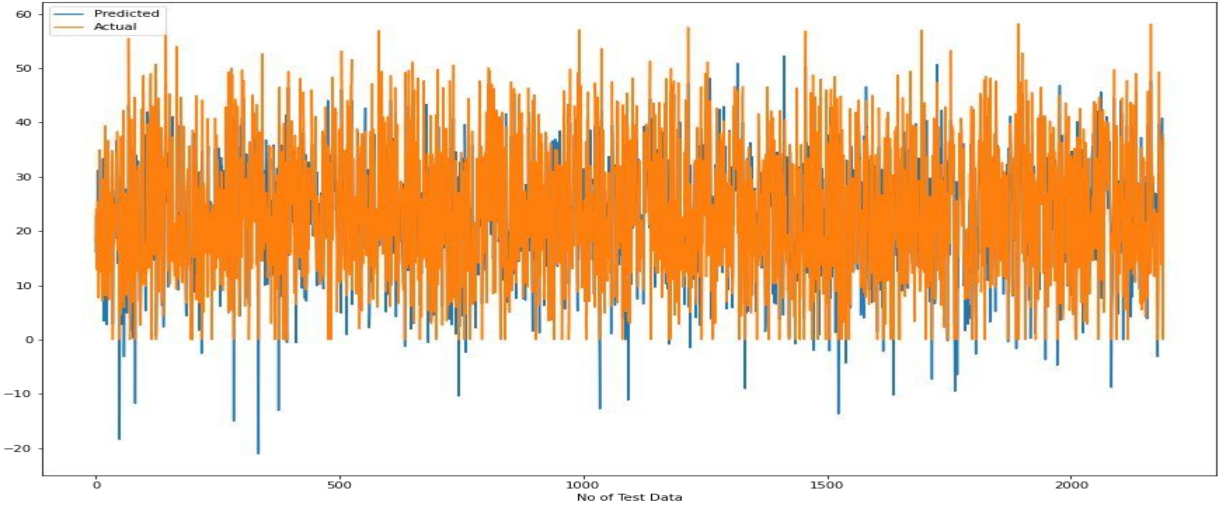
LASSO REGRESSION

Train Set Results

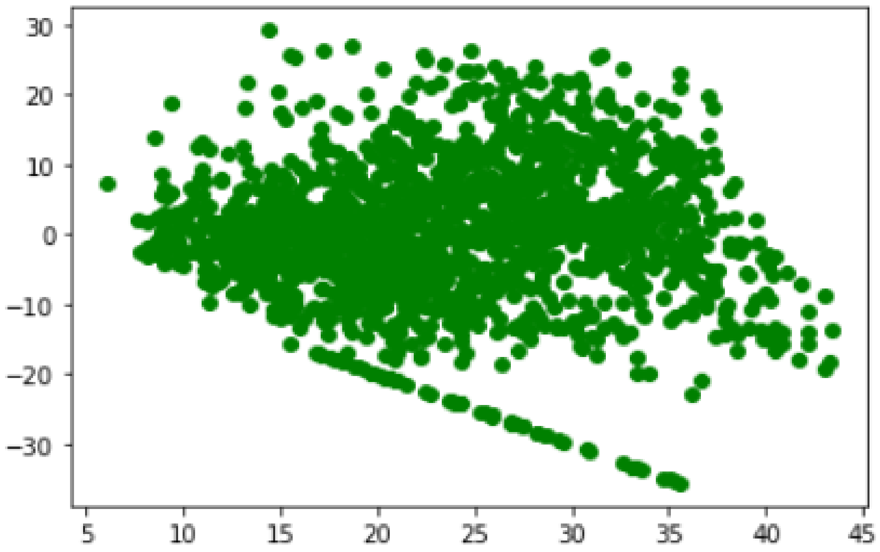
	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
1	Lasso regression	7.375	94.185	9.705	0.393	0.38

Test Set Results

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
1	Lasso regression	7.375	94.185	9.705	0.393	0.38



Heteroscedasticity



- **In this Regularised lasso regression model, we can see that the model is not able to capture most of the data and variance, and hence the data is more randomly scattered and the errors are trying to form a pattern.**

- **So, this model is not suitable for this dataset.**

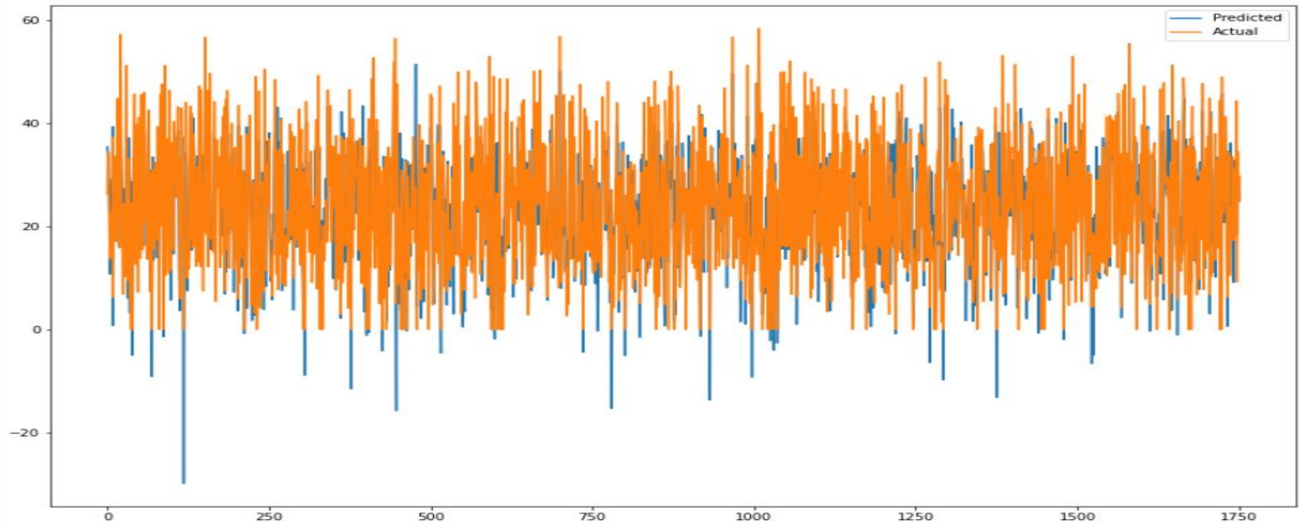
RIDGE REGRESSION

Train Set Results

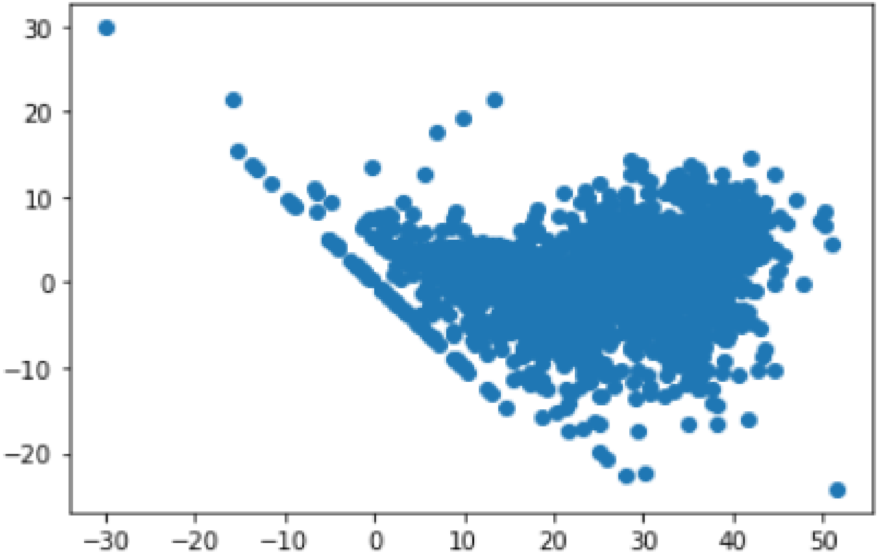
	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
2	Ridge regression	4.239	30.769	5.547	0.802	0.80

Test Set Results

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
2	Ridge regression	4.239	30.769	5.547	0.802	0.80



Heteroscedasticity



- **In this Regularised Ridge regression model, we can see that the model is able to capture most of the data and variance, and hence the data is less and randomly scattered, and the errors are not forming any pattern.**

- **So, this model is also suitable for the given dataset.**

CHALLENGES

- Large dataset to handle.
- Required to plot different graphs for better analysis.
- Outliers detection.
- Treatment of Outliers.
- Optimising the model for better accuracy.
- Understanding the evaluation matrix.

CONCLUSION

- The 'Hour' of a day holds the most important feature.
- The bike rental count is mostly correlated with the time of the day, as it peaks at 10 am in the morning and 8 pm in the evening.
- We observed that the bike rental count is higher during working days, especially between 7 am and 9 am and 5 pm and 7 pm, than on non-working days.
- We see that people prefer to ride in temperatures ranging from moderate to high, as well as in light winds.
- It is observed that the highest number of rental bikes is counted in the autumn and summer seasons, and the lowest in the winter season.
- We observed that the highest number of bike rentals was on a normal day, and the lowest on a snowy and rainy day.
- We observed that with increasing humidity, the number of bike rental counts decreases.

CONCLUSION CONT.

- When we compare the root mean squared error and the mean absolute error of all the models, we can see that we have very minimal errors. Finally, this model is best for predicting the bike rental count on a daily basis.

Train Set Results

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.238	30.145	5.490	0.802	0.80
	1	Lasso regression	7.375	94.185	9.705	0.393	0.38
	2	Ridge regression	4.239	30.769	5.547	0.802	0.80

TEST Set Results

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Test set	0	Linear regression	4.238	30.145	5.490	0.802	0.80
	1	Lasso regression	7.375	94.185	9.705	0.393	0.38
	2	Ridge regression	4.239	30.769	5.547	0.802	0.80

Thank You

<u>Team GodSpeed☺</u>