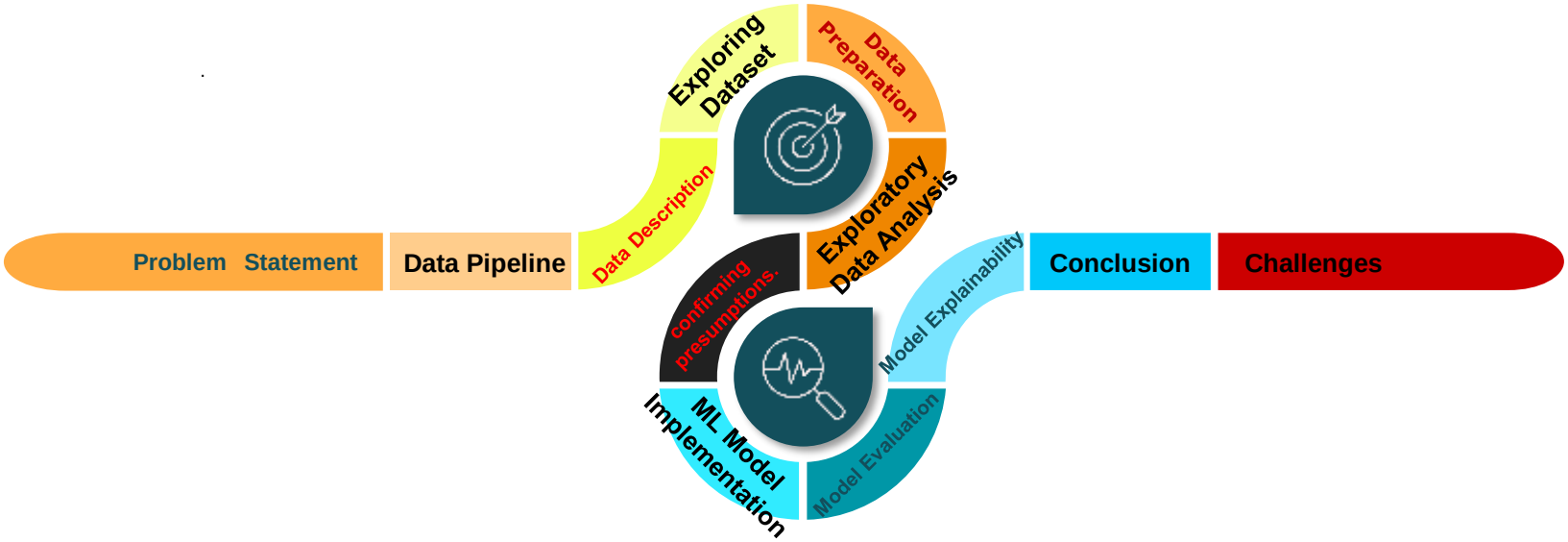
# Capstone Project

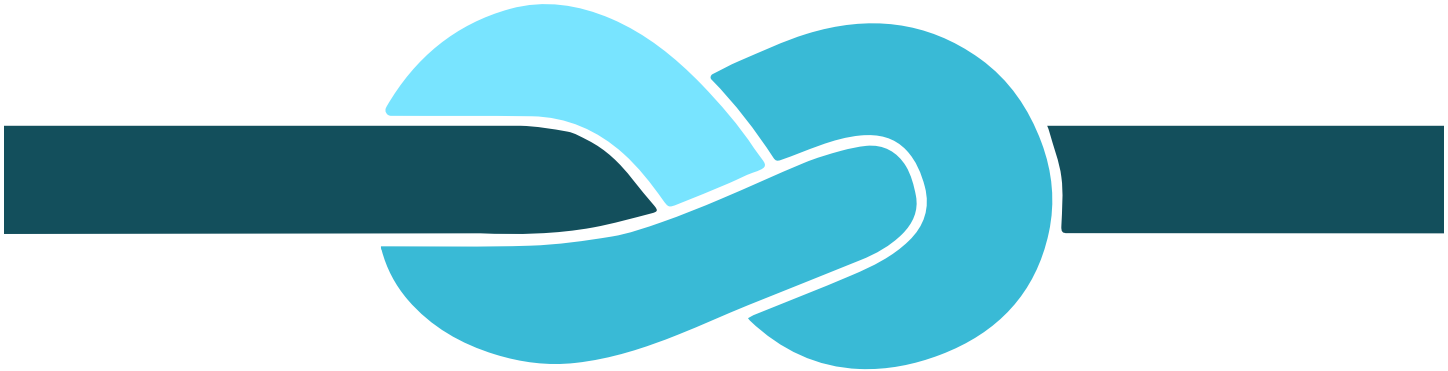
## Seoul Bike Sharing Demand Prediction

**Ajit Sharad Mane**
**(ajitmane36@gmail.com)**

# Index

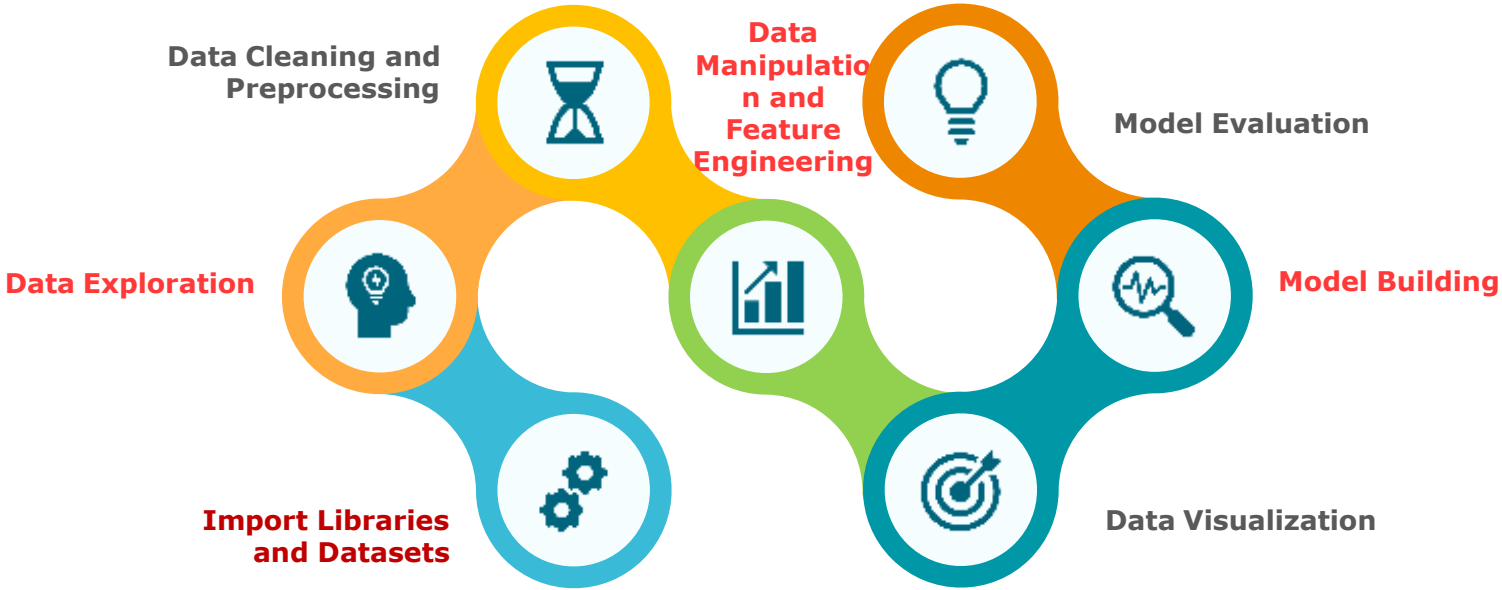

Problem Statement
Data Pipeline
Data Description
Exploring Dataset
Data Preparation
Exploratory Data Analysis
confirming presumptions.
Model Explainability
Conclusion
Challenges
ML Model Implementation
Model Evaluation

# 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 the bike count required at each hour for the stable supply of rental bikes.

# Data Pipeline

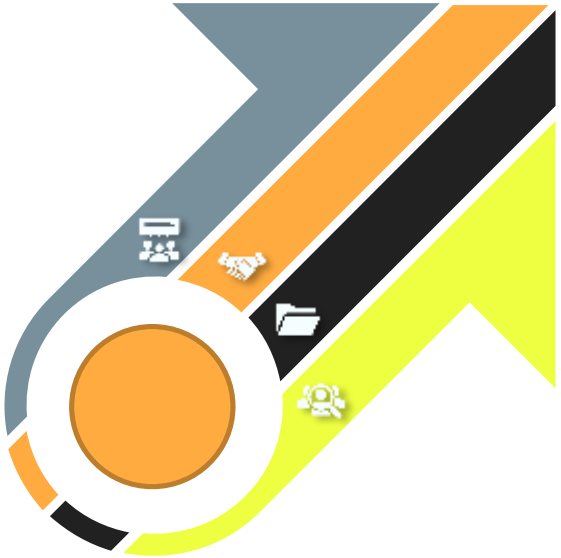


# Data Description

| Sr.No. | Feature Name | Description |
|---|---|---|
| 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 |
| 7 | Visibility | 10m |
| 8 | Dew point temperature | Celsius |
| 9 | Solar radiation | MJ/m2 |
| 10 | Rainfall | mm |
| 11 | Snowfall | cm |
| 12 | Seasons | Winter, Spring, Summer, Autumn |
| 13 | Holiday | Holiday/No holiday |
| 14 | Functional Day | NoFunc(Non Functional Hours), Fun(Functional hours) |

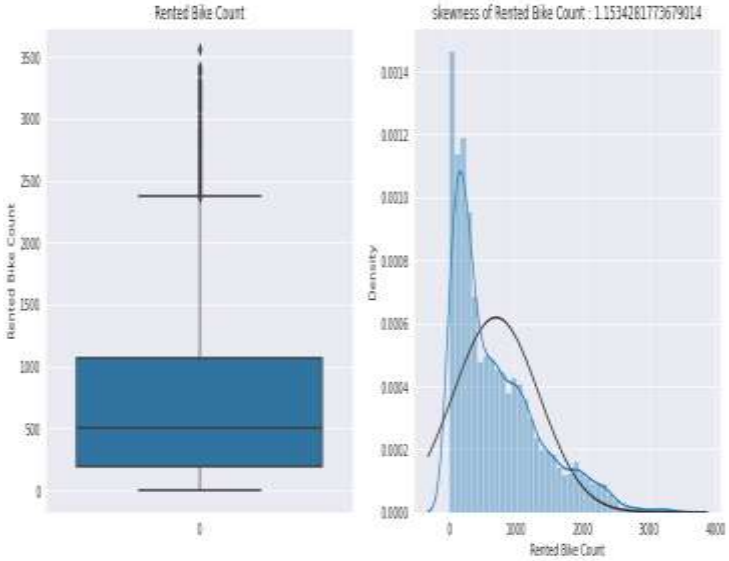
# Data Exploration

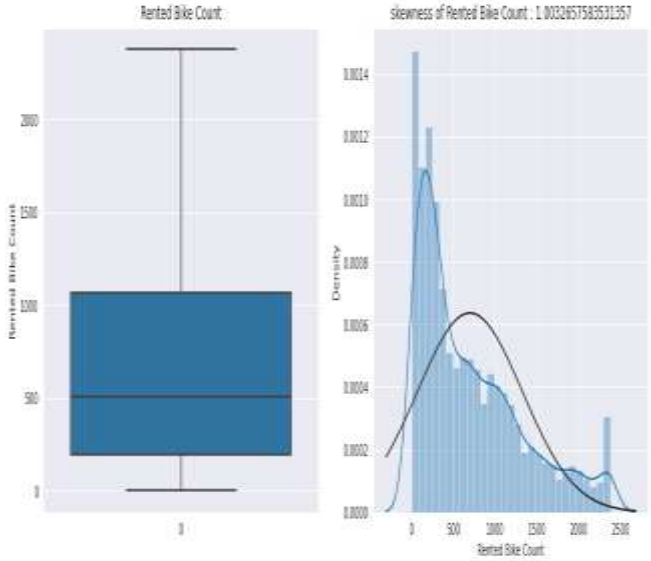
- There are 8760 observations and 14 features in this dataset.
- data set containing variables of the datatypes text, int64, and float64.
- Variable Date is the incorrect datatype in a string.
- No duplicate and null values are in the dataset.

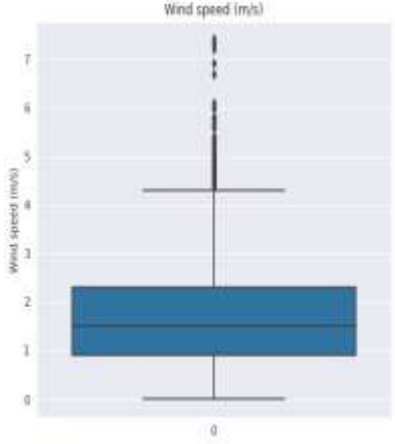


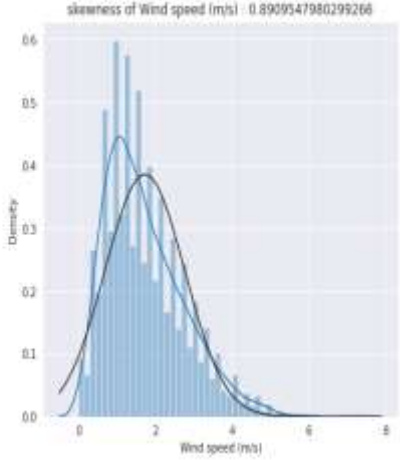
# Data Preparation

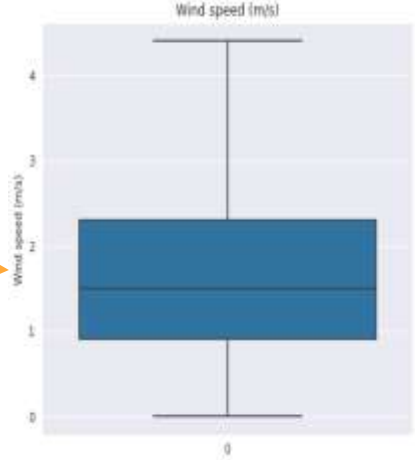
- The dataset does not have any duplicated and null values.
- The variables Rented Bike Count, Wind speed (m/s), and Solar Radiation (MJ/m2) have outliers.

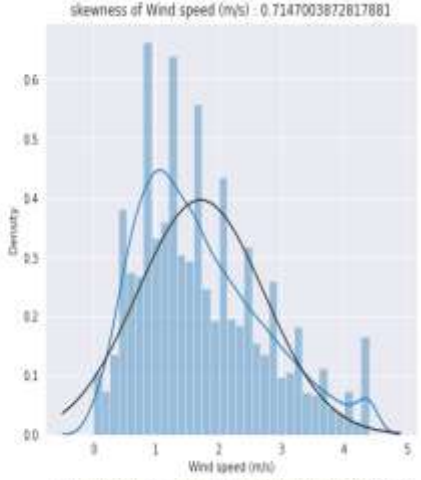


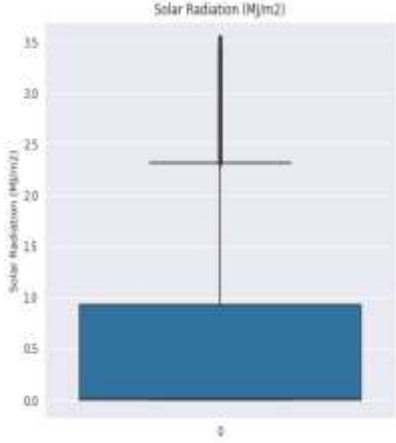



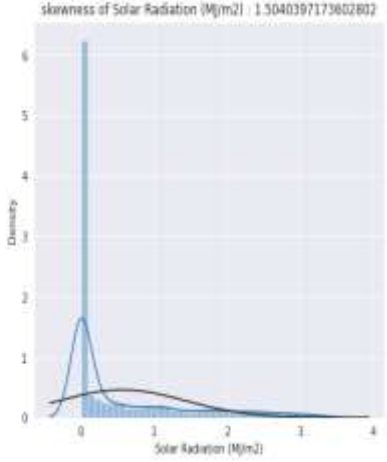

Wind speed (m/s)
Wind speed (m/s)
0

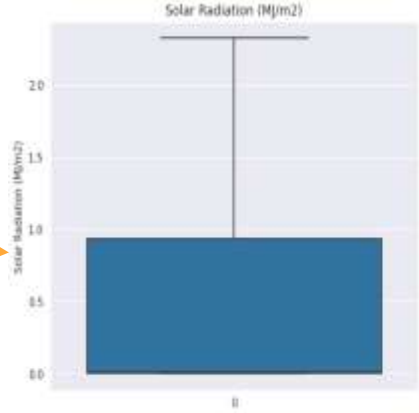

skewness of Wind speed (m/s) : 0.8909547980299266
Density
Wind speed (m/s)

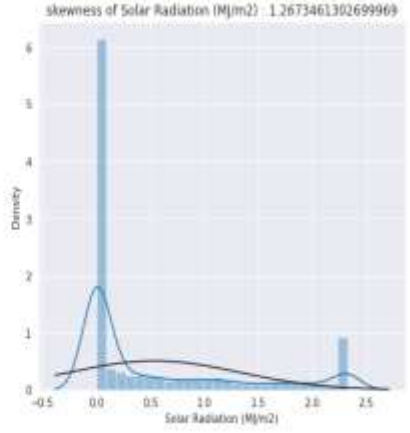

Wind speed (m/s)
Wind speed (m/s)
0


skewness of Wind speed (m/s) : 0.7147003872817881
Density
Wind speed (m/s)


Solar Radiation (MJ/m2)
Solar Radiation (MJ/m2)
0

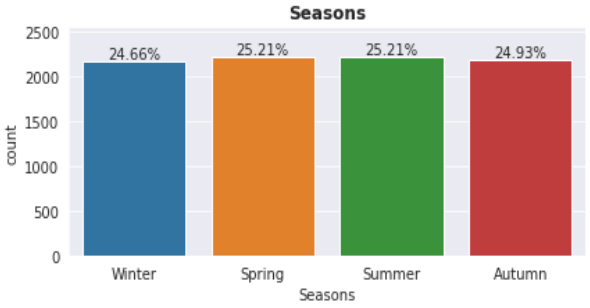

skewness of Solar Radiation (MJ/m2) : 1.5040397173602802
Density
Solar Radiation (MJ/m2)

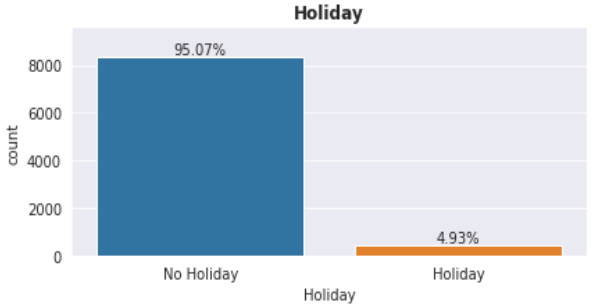

Solar Radiation (MJ/m2)
Solar Radiation (MJ/m2)
0

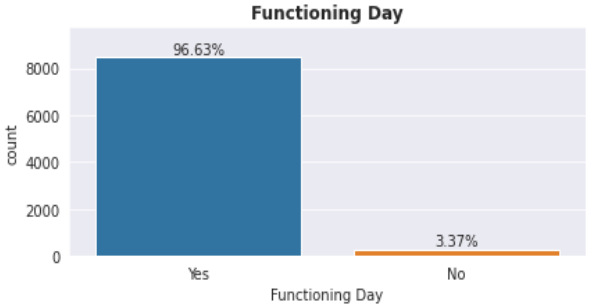

skewness of Solar Radiation (MJ/m2) : 1.2673461302699969
Density
Solar Radiation (MJ/m2)

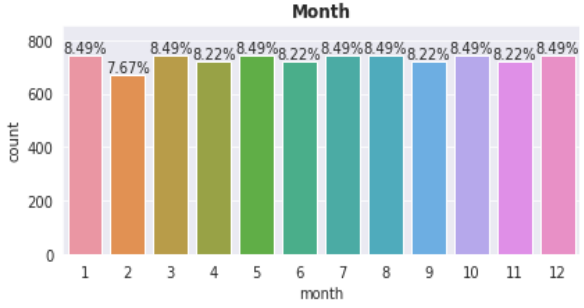
# Exploratory Data Analysis

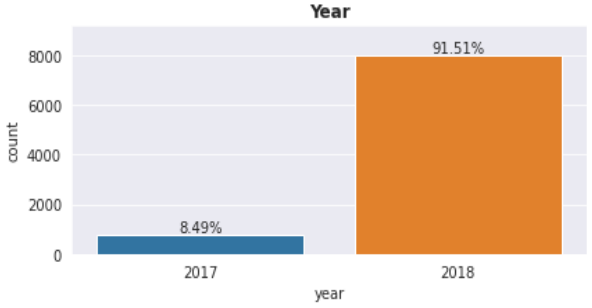
## Univariate Analysis



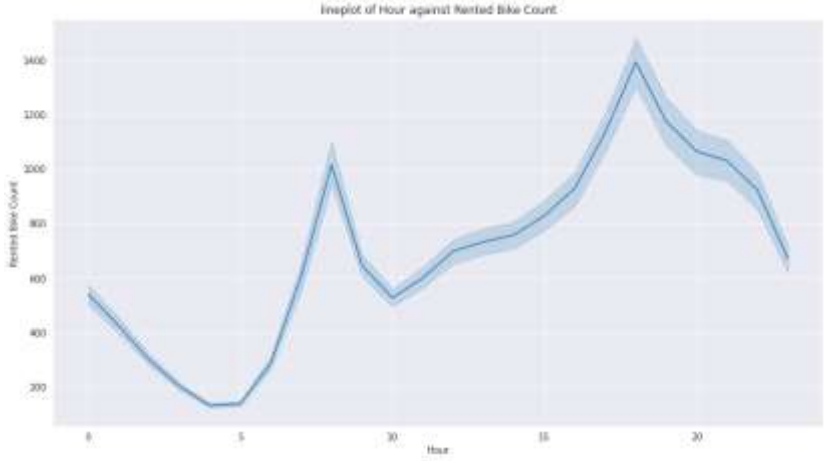



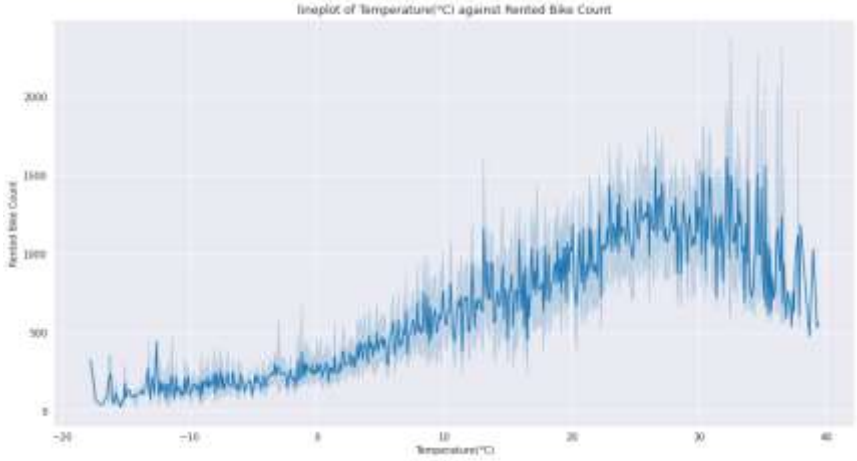



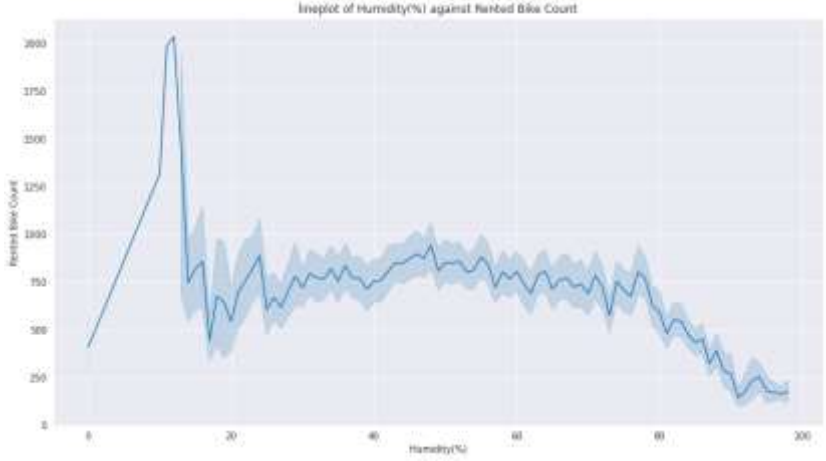



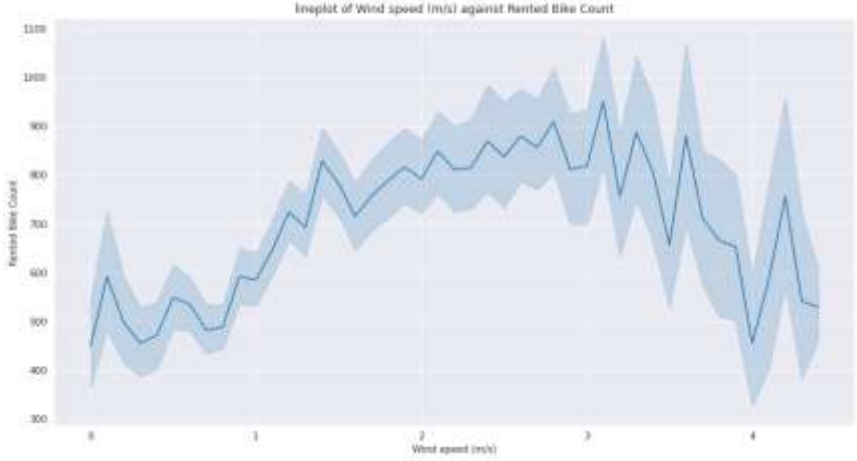

- Customers favour rental motorcycles equally in all seasons.
- When there are no holidays, customers choose to rent motorcycles. Customers hardly ever use the bikes they rent while traveling on holiday.
- Nearly all consumers preferred to rent bikes during functional hours.
- Bicycle rentals are popular all month long.
- Renting bicycles was not very popular in 2017, but it increased by 83.02 percent in 2018.

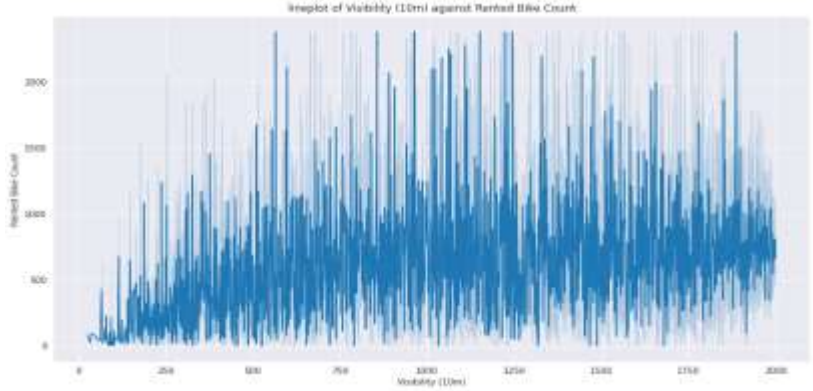
# Bivariate Analysis

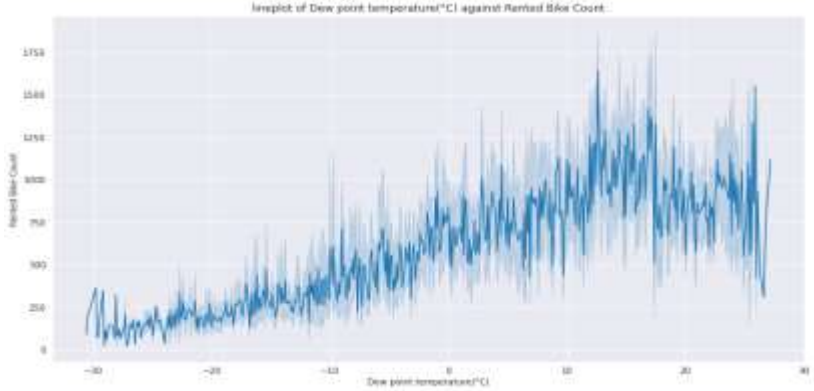


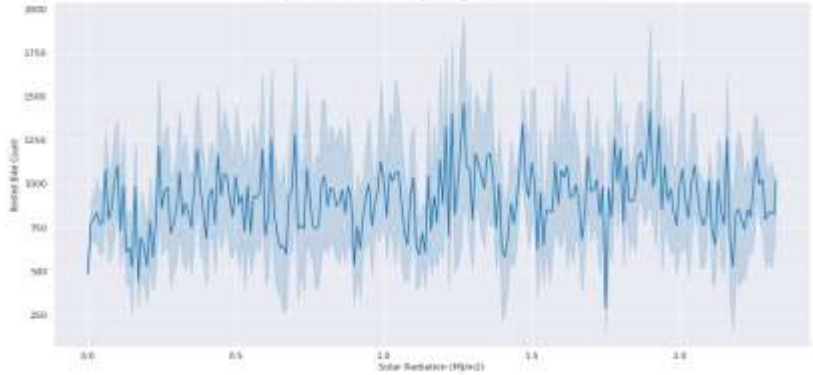



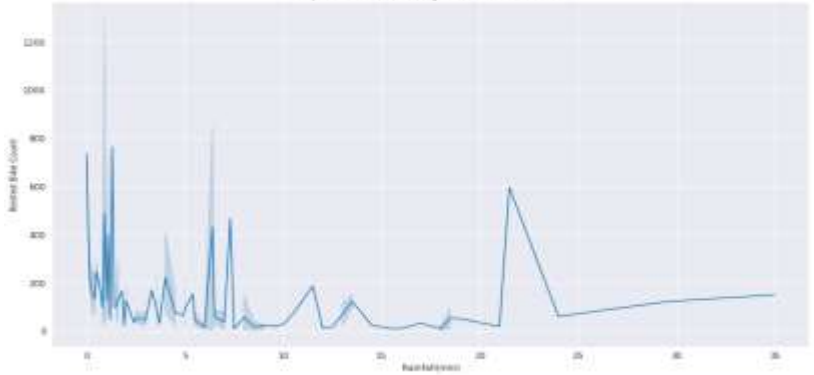



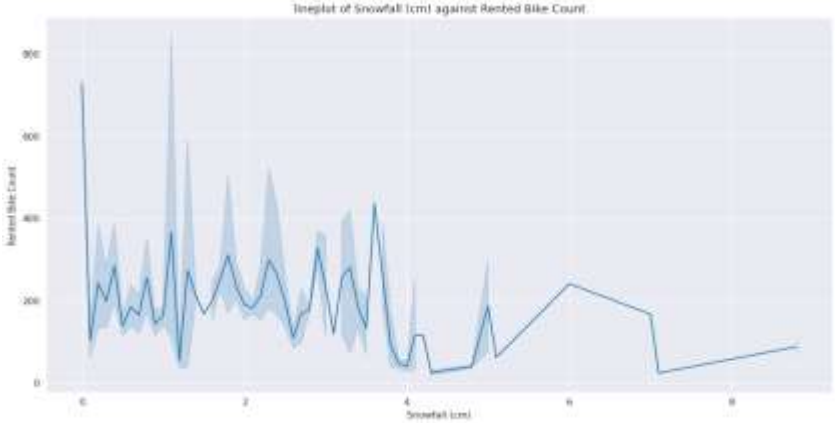



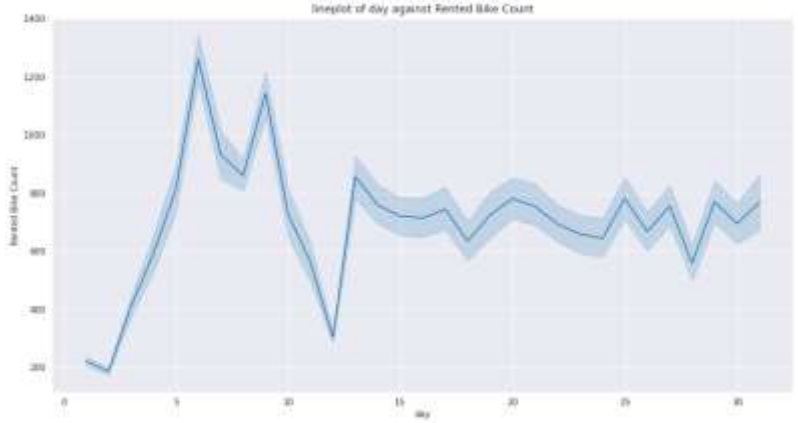

lineplot of Visibility (10m) against Rented Bike Count
Rented Bike Count
Visibility (10m)

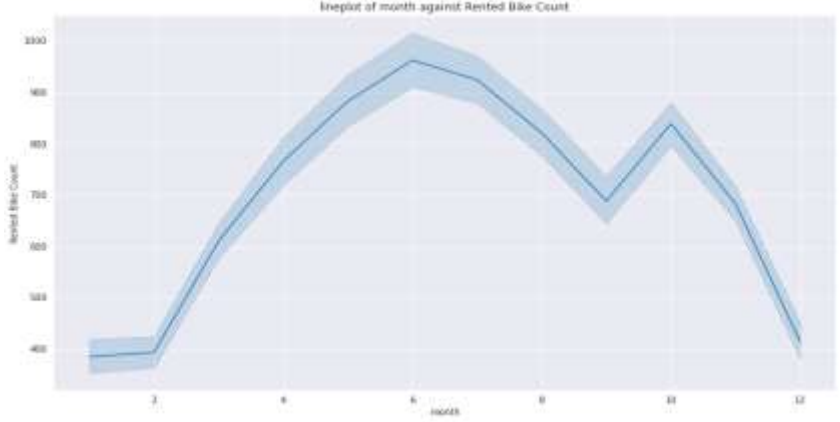

lineplot of Dew point temperature(°C) against Rented Bike Count
Rented Bike Count
Dew point temperature(°C)


lineplot of Solar Radiation (MJ/m2) against Rented Bike Count
Rented Bike Count
Solar Radiation (MJ/m2)


lineplot of Rainfall(mm) against Rented Bike Count
Rented Bike Count

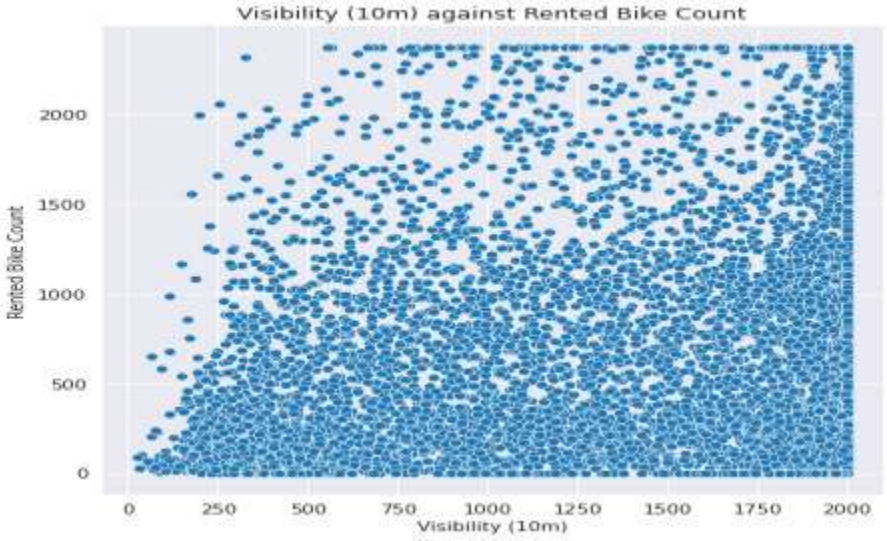

lineplot of Snowfall (cm) against Rented Bike Count
Rented Bike Count
Snowfall (cm)


lineplot of day against Rented Bike Count
Rented Bike Count
day


lineplot of month against Rented Bike Count
Rented Bike Count
month

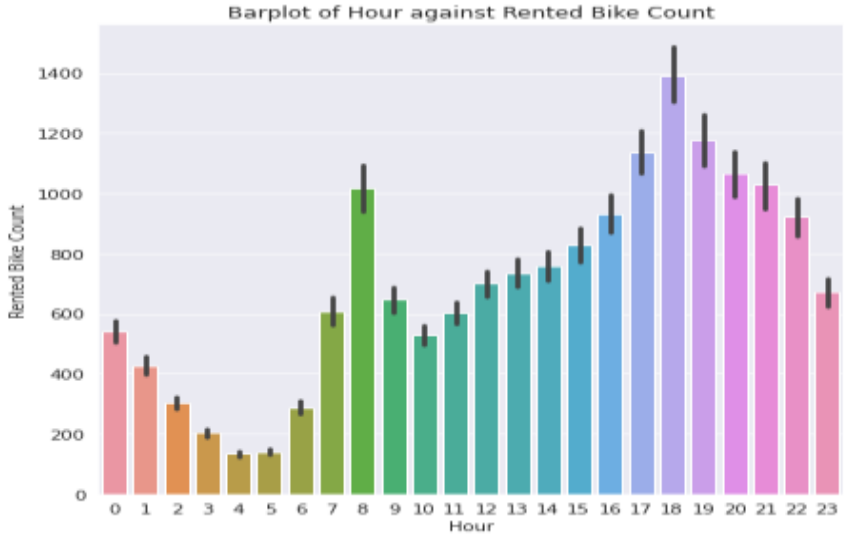
# Observations :

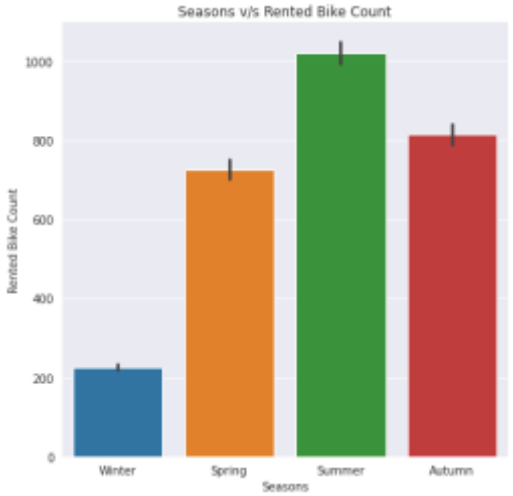
- Customers who travel most commonly use rented bikes in the morning at 8 A.M. and in the evening at 6 P.M.
- When the humidity level is between 10% and 18%, people prefer to rent bikes.
- When wind speed is between 2 m/s and 3.5 m/s, people consistently use rented bikes, and it is at its peak when wind speed is normal, which is 3.2 m/s.
- Renting a bike is the best option for customers in dew point temperatures ranging from 12°C to 18°C. The use of a rented bike increases with increasing dew point temperatures, but it still reaches normal dew point temperatures.
- According to the graph, solar radiation has no effect on customer use of rented bikes.
- When it's not raining, people prefer rental bikes the most.
- When there is no snowfall, most people opt to rent bikes. However, the majority of customers prefer to rent bikes when it snows up to 4 cm.
- In the first 10 days of the month, most rented bikes are used by customers. Customers consistently use rented bikes in the last 15 days of the month.
- In June, most rented bikes are used through the year, followed by October. Customers' use of rent bikes is at its peak from April to September.

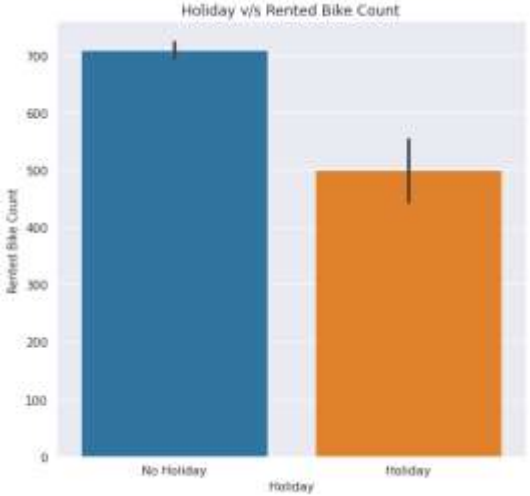


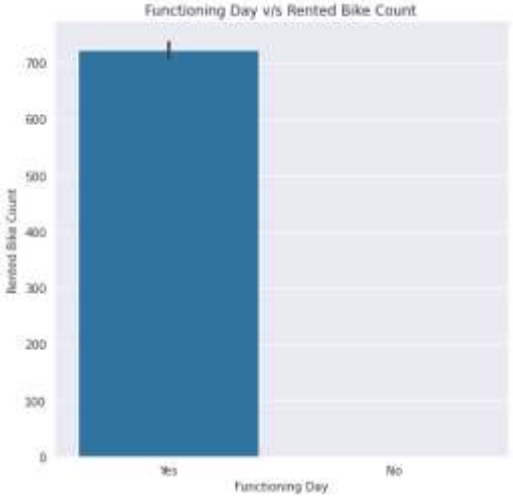

- The count of rented bikes on that day is unaffected by the day's visibility, but when visibility exceeds 1750, the use of rented bikes increases more than usual.

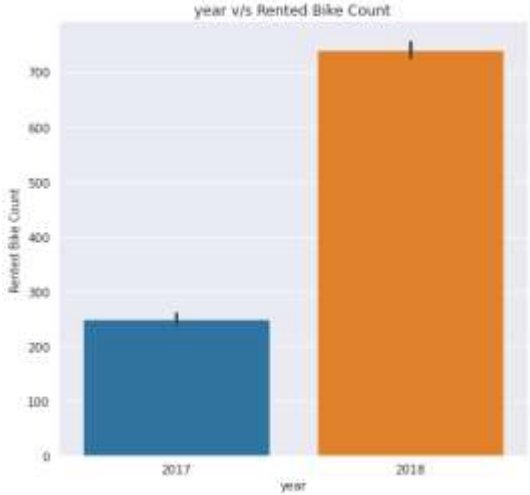
- At night, customers do not prefer to use rented bikes.
- Customers do not prefer rented bikes in the mornings 4 and 5, but from 7, 8, and 9, the use of rented bikes increases, possibly due to working people going to the office, and it is the same in the evenings 5, 6, and 7, because people are traveling from the office to home. Overall, the rented bike was the most frequently used during office in and out times.
- Customers mostly use rented bikes for transportation in the evening.





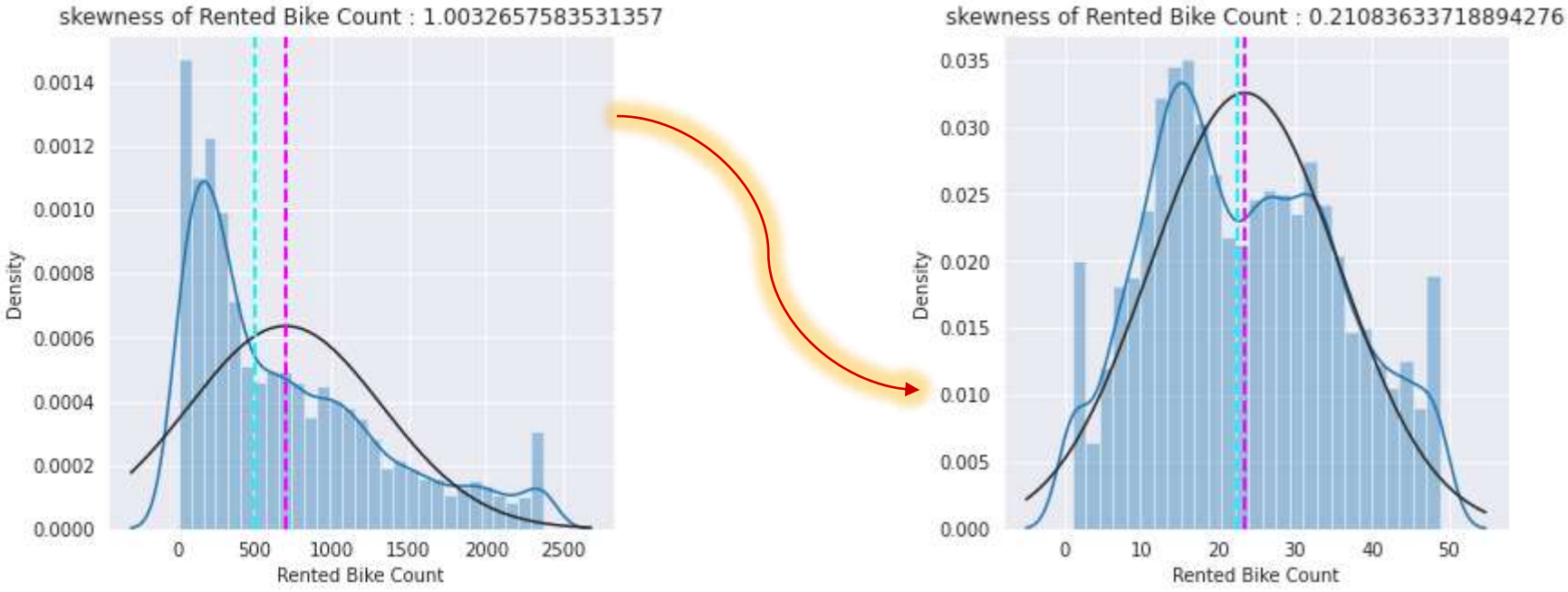



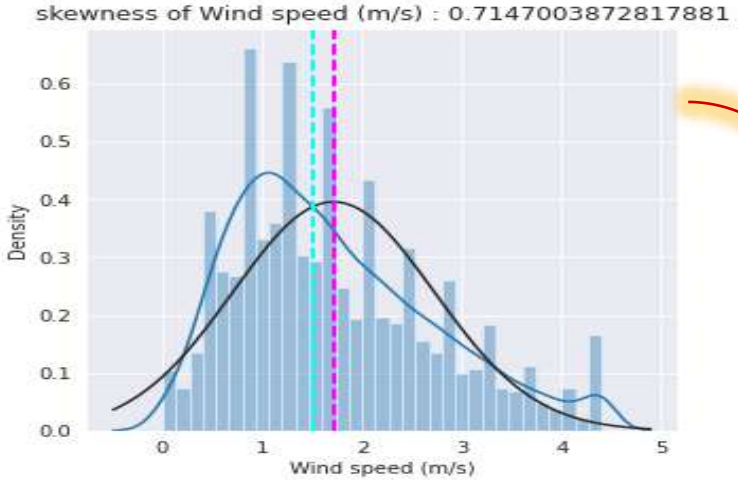

- **During the summer and autumn seasons, most people rent bikes. During the winter, fewer people choose to rent bikes.**
- **Even when there is no holiday other than a holiday, people rent bikes. The use of rented bikes on holidays is lower than on non-holiday days.**
- **Almost every rented bike is used during its functional hours.**
- **The use of rented bikes increased by three times in 2018 compared to 2017.**

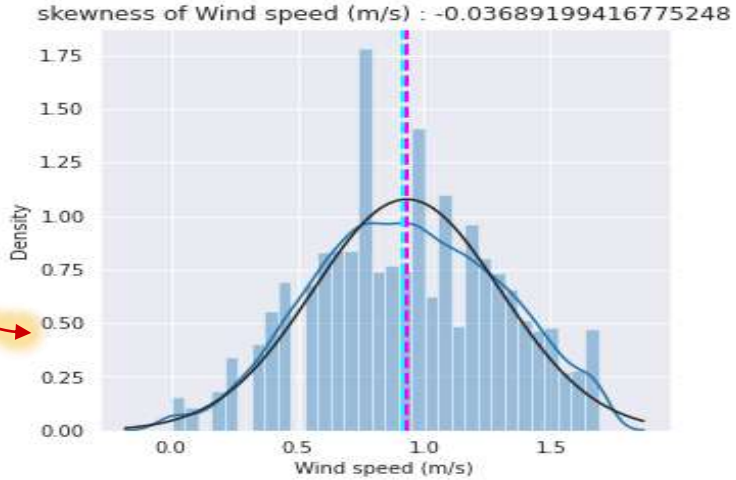
# Confirming Presumptions

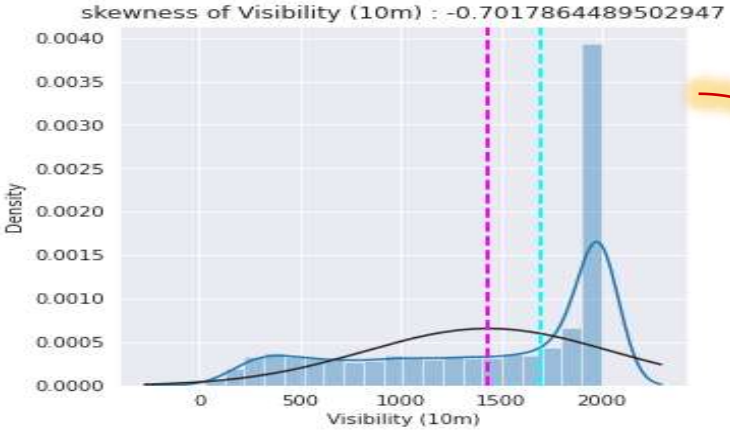
- *Checking Distribution of each feature and transform it to normal distribution*

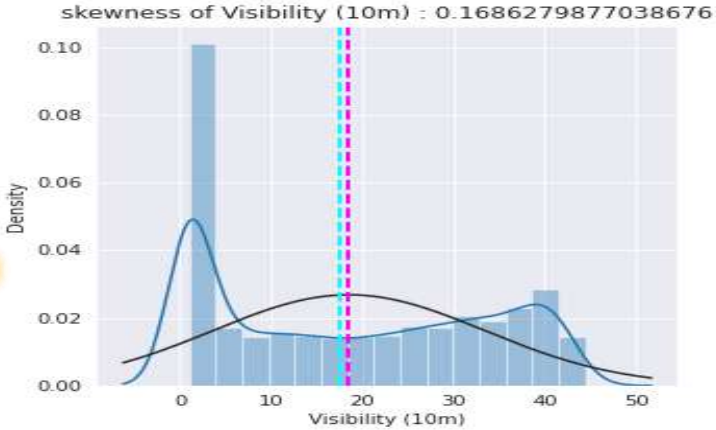
**Using log and square root transformation, the distribution of the features Rented Bike Count, Wind speed (m/s), Solar Radiation (MJ/m2), Visibility (10 m), Rainfall (mm), and Snowfall (cm) was returned to normal.**




skewness of Wind speed (m/s) : 0.7147003872817881
Density
Wind speed (m/s)

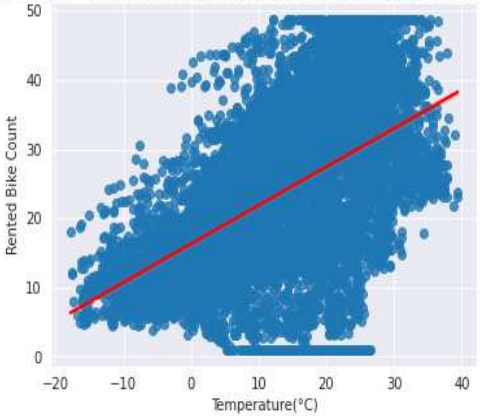

skewness of Wind speed (m/s) : -0.03689199416775248
Density
Wind speed (m/s)


skewness of Visibility (10m) : -0.7017864489502947
Density
Visibility (10m)

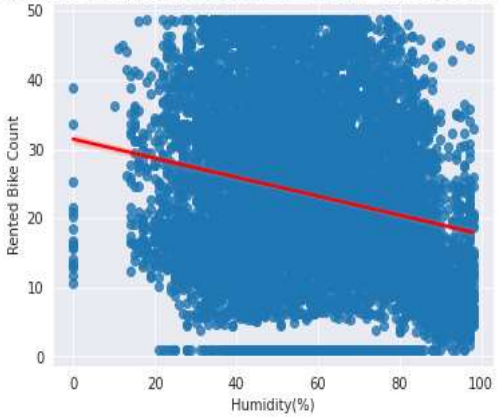

skewness of Visibility (10m) : 0.1686279877038676
Density
Visibility (10m)

## ❑ *Checking relationship between independent and dependent variables is linear*

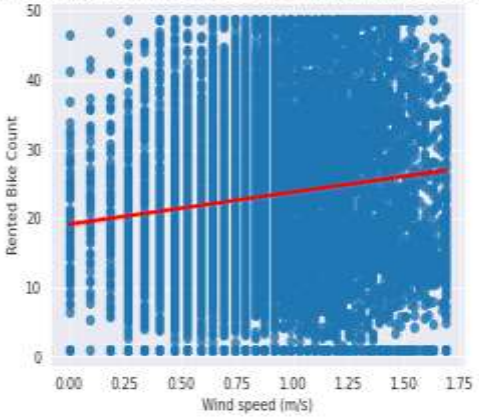
Temperature(°C) v/s Rented Bike Count correlation : 0.5437684239651075



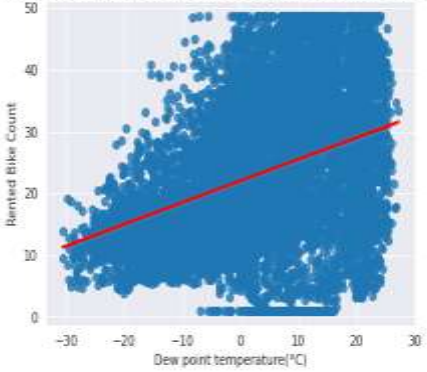

Humidity(%) v/s Rented Bike Count correlation : -0.22825659554920055



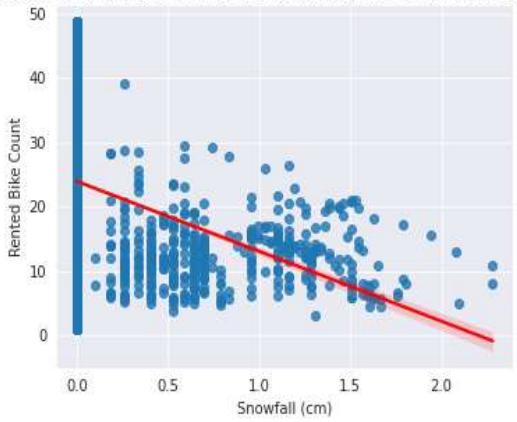

Wind speed (m/s) v/s Rented Bike Count correlation : 0.14011511273696708



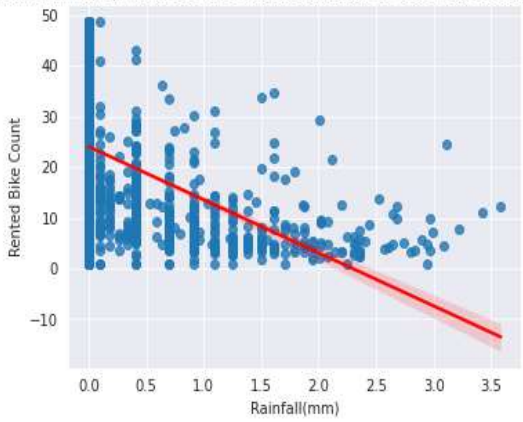

Dew point temperature(°C) v/s Rented Bike Count correlation : 0.37432235058492314



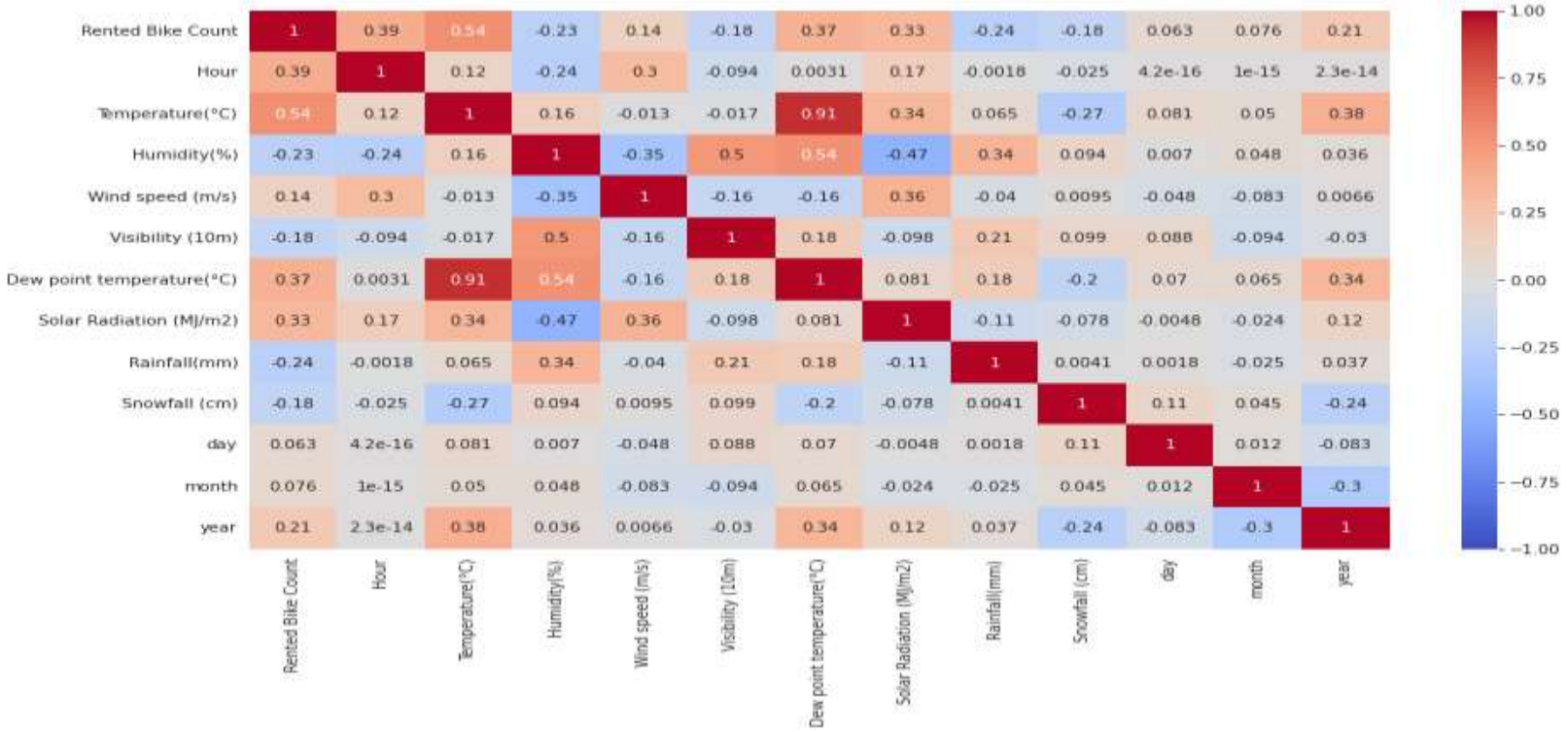

Snowfall (cm) v/s Rented Bike Count correlation : -0.1794573992096429




Rainfall(mm) v/s Rented Bike Count correlation : -0.23957068086384628



## ❑ *Checking multicollinearity in independent variables*

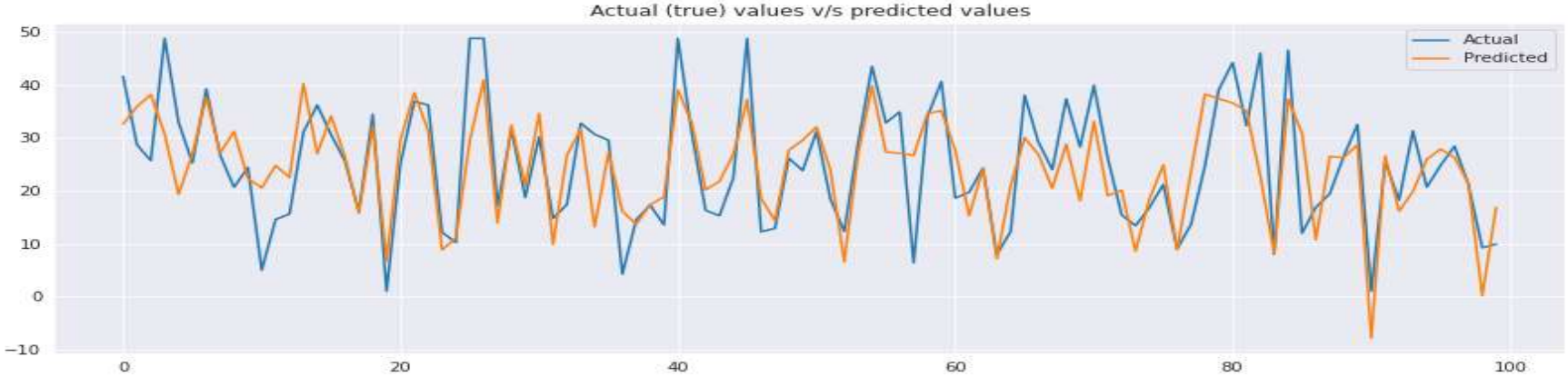


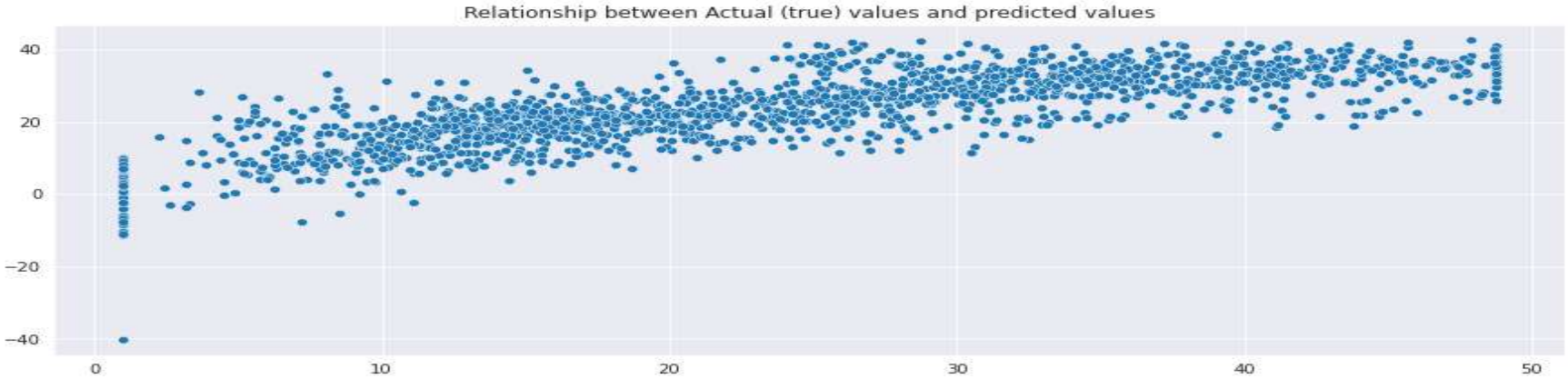

❑ **Dew point temperature (°C) and temperature (°C) have a strong correlation. A moderate correlation exists between "humidity (%)" and "dew point temperature (°C)". The variables "year" and "Dew point temperature (°C)" have a weak correlation.**

# Machine Learning Model Implementation

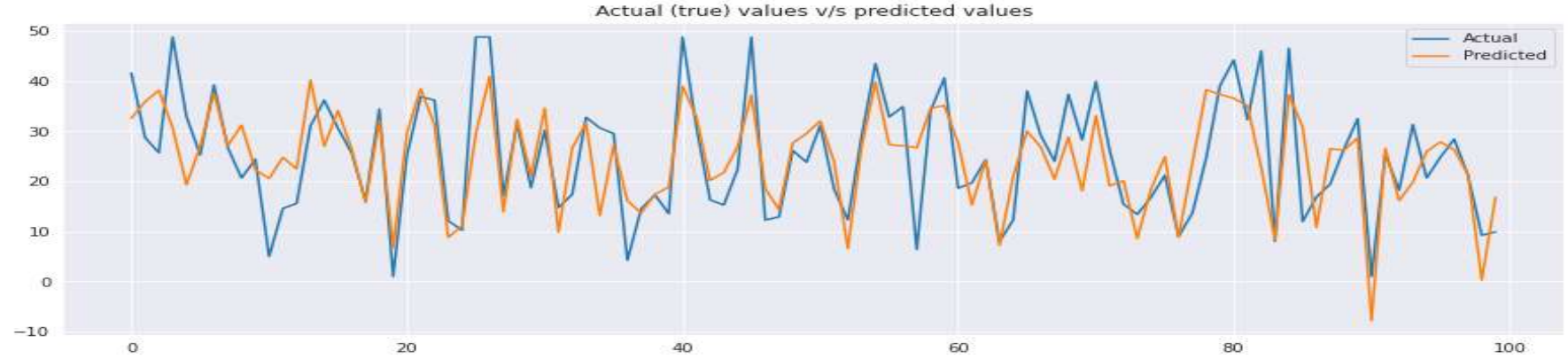
1. • **Linear Regression**
2. • **Lasso (L1) Regression**
3. • **Ridge (L2) Regression**
4. • **ElasticNet Regression**
5. • **Decision Tree**
6. • **Random Forest**
7. • **XGBoost**

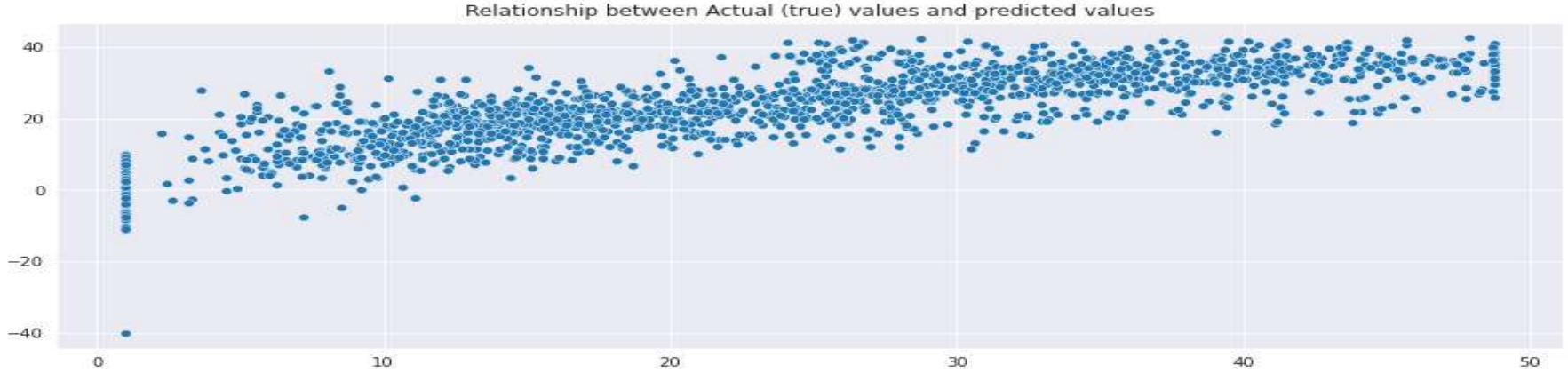
# ❑ Linear Regression



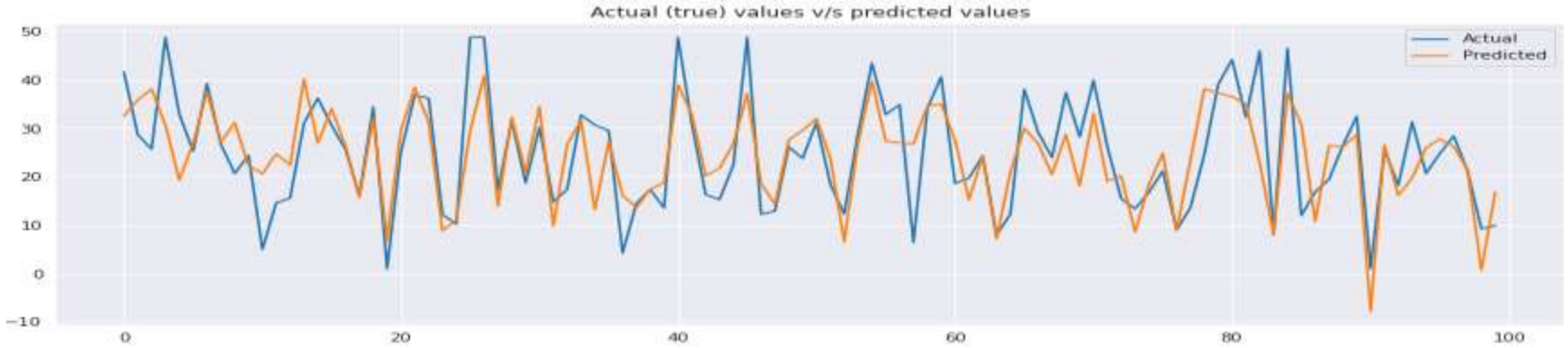



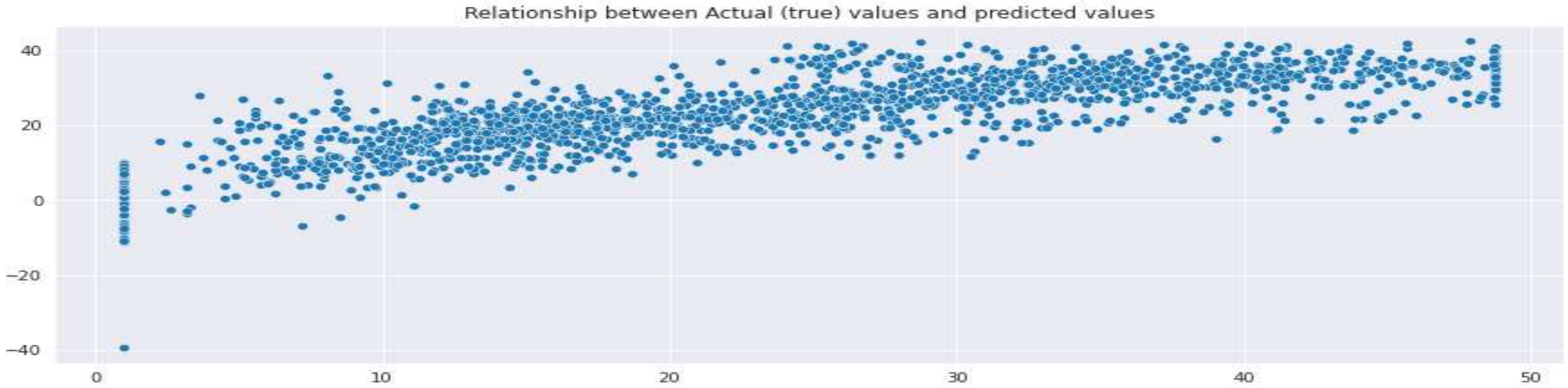
# ❑ Lasso (L1)


Actual (true) values v/s predicted values
Actual
Predicted
50
40
30
20
10
0
−10
0
20
40
60
80
100

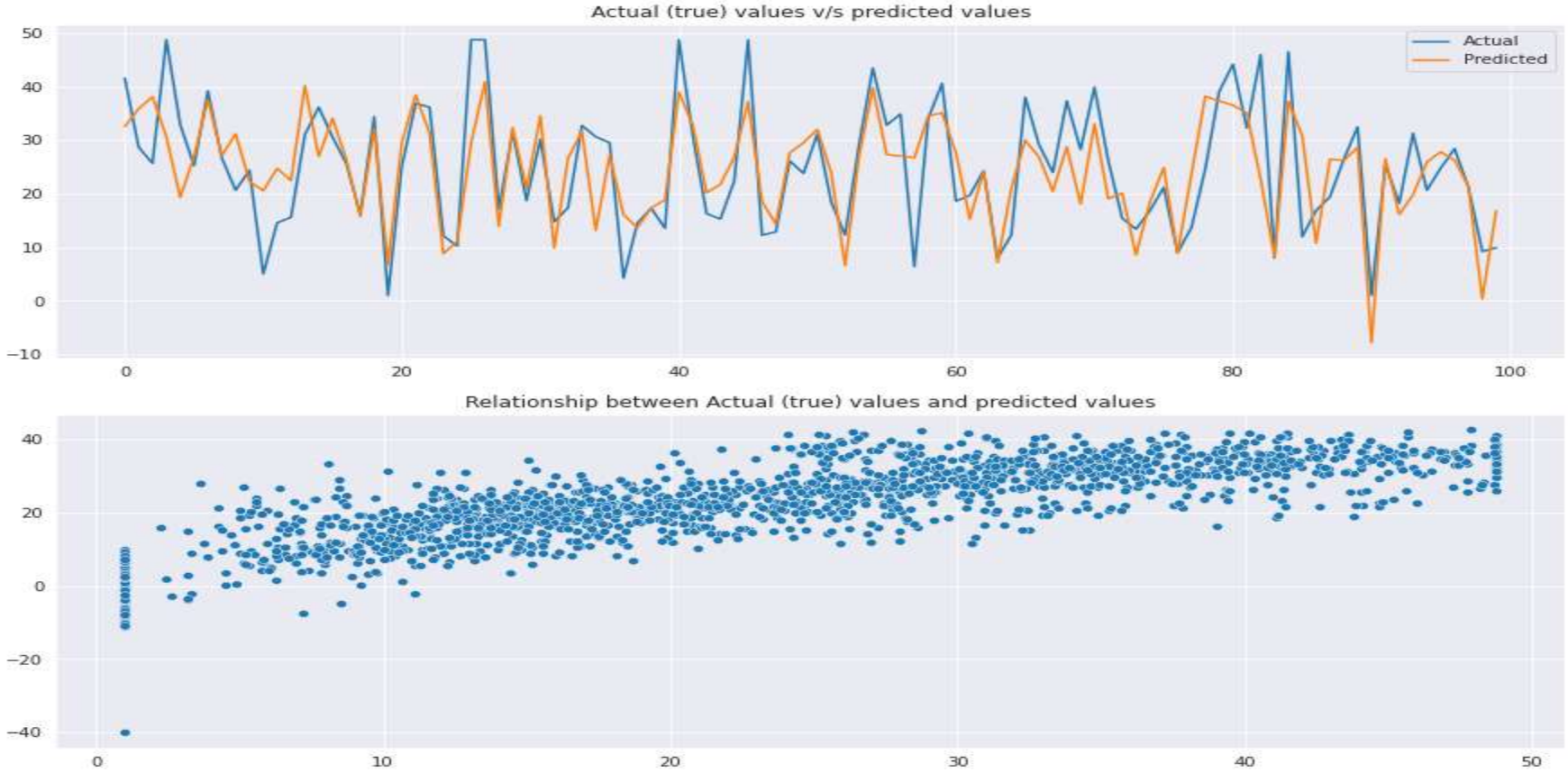

Relationship between Actual (true) values and predicted values
40
20
0
−20
−40
0
10
20
30
40
50

## ❑ Ridge Regression (L2)



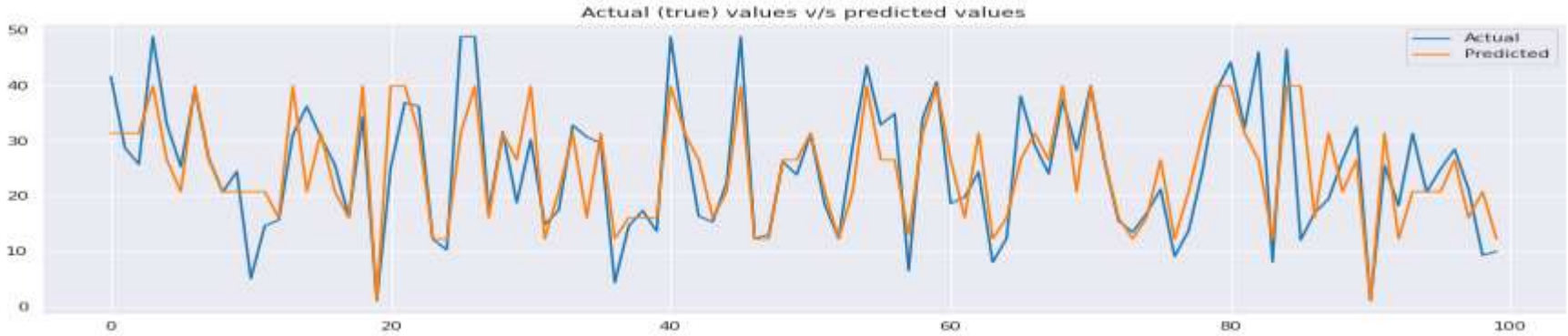



## ❑ ElasticNet Regression

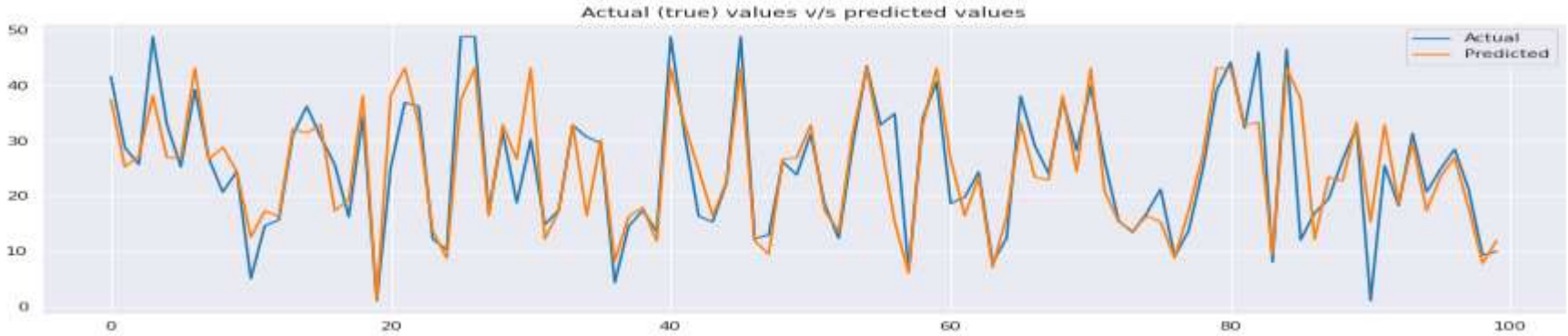


## ❑ Decision Tree

### Train_test_split :

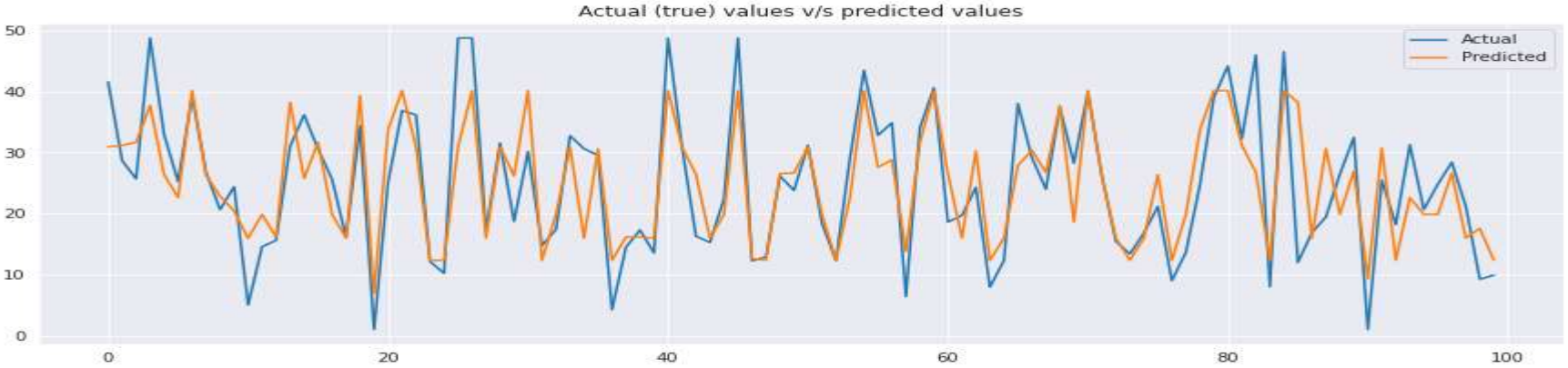


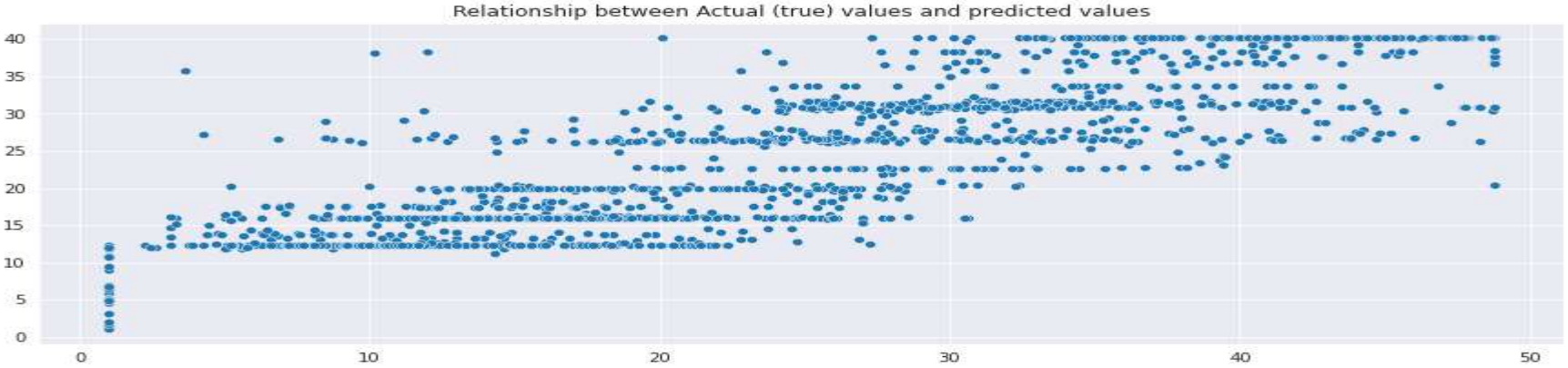

### Cross Validation :



## ❑ Random Forest

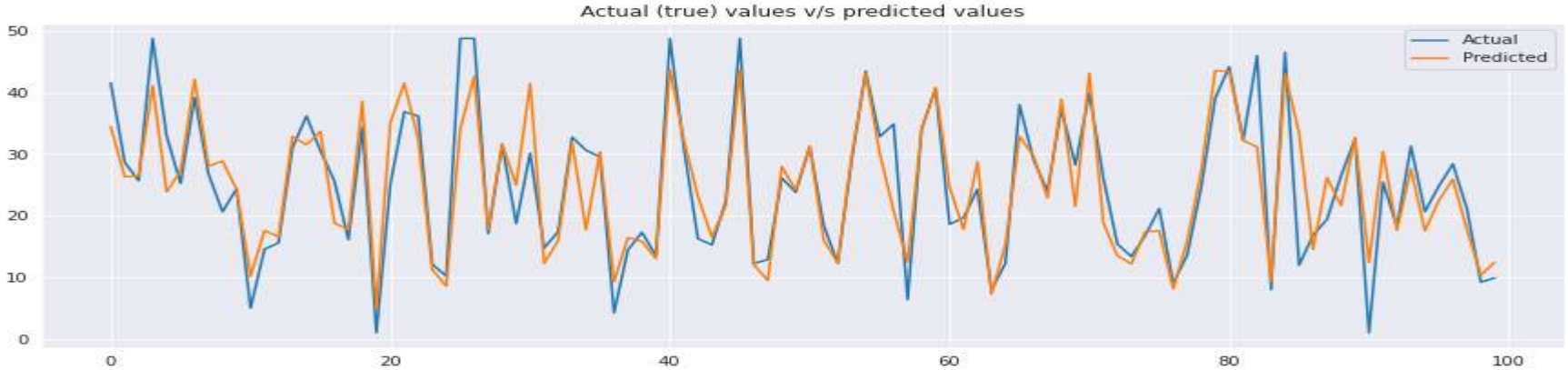
**train_test_split :**

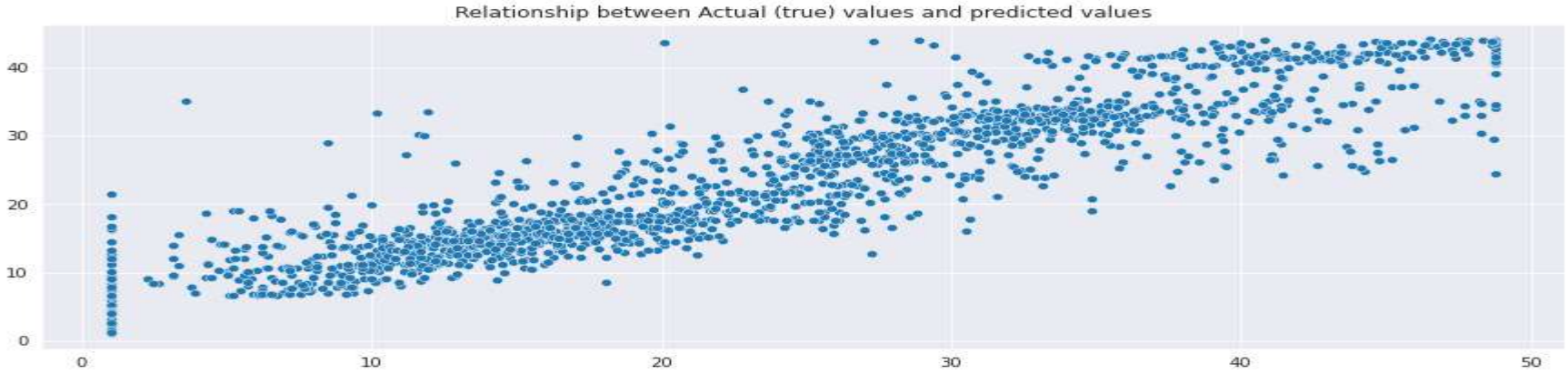


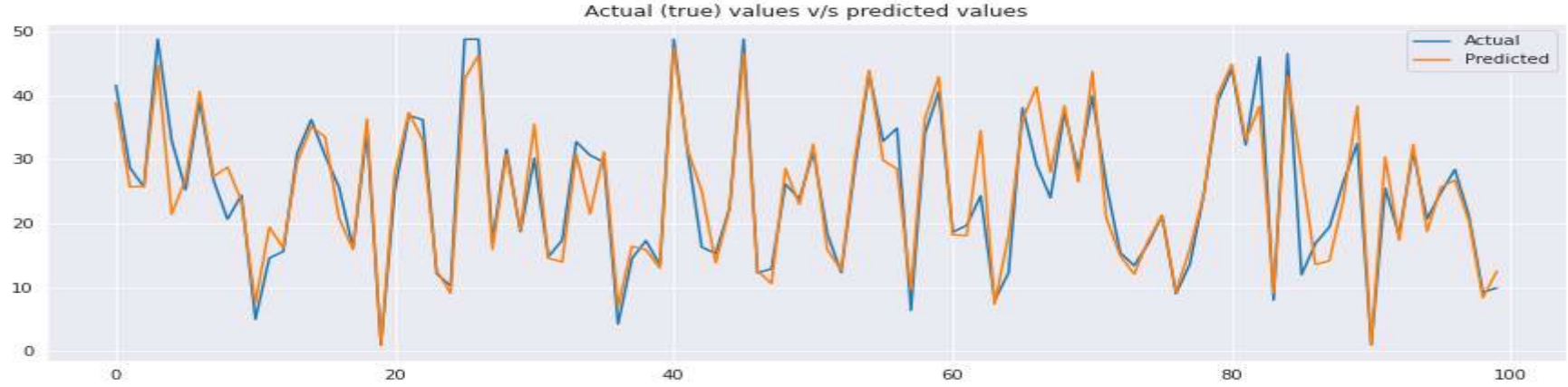
# ❑ Random Forest

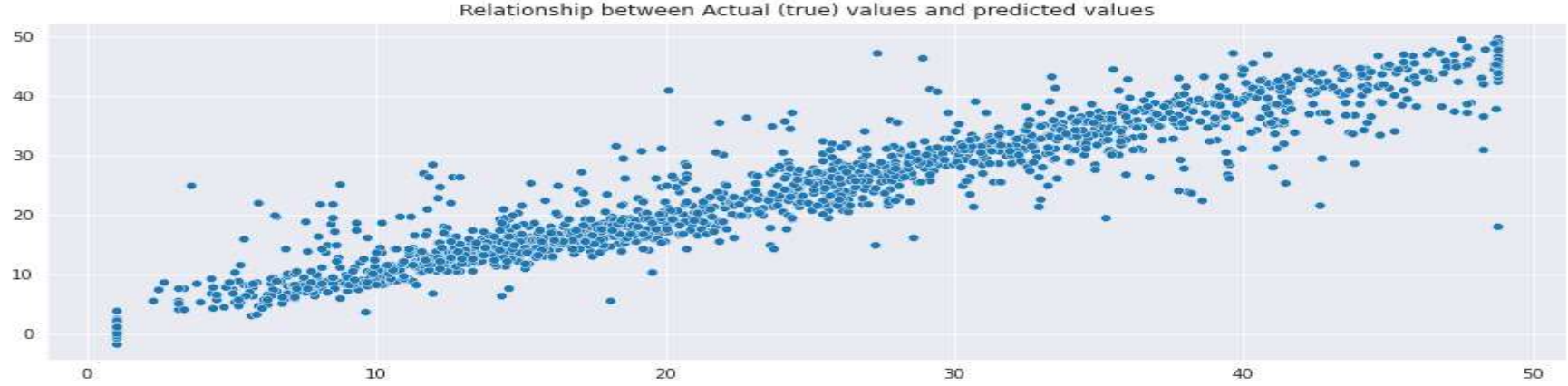
## Cross Validation :



## ❑ XGBoost


Actual (true) values v/s predicted values
Actual
Predicted
0
10
20
30
40
50
0
20
40
60
80
100

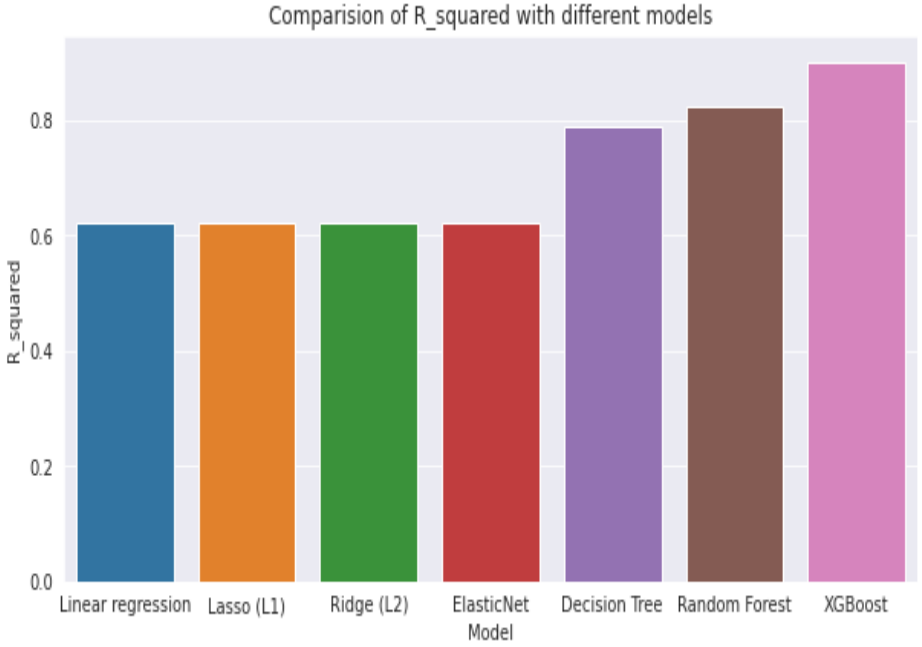

Relationship between Actual (true) values and predicted values
0
10
20
30
40
50

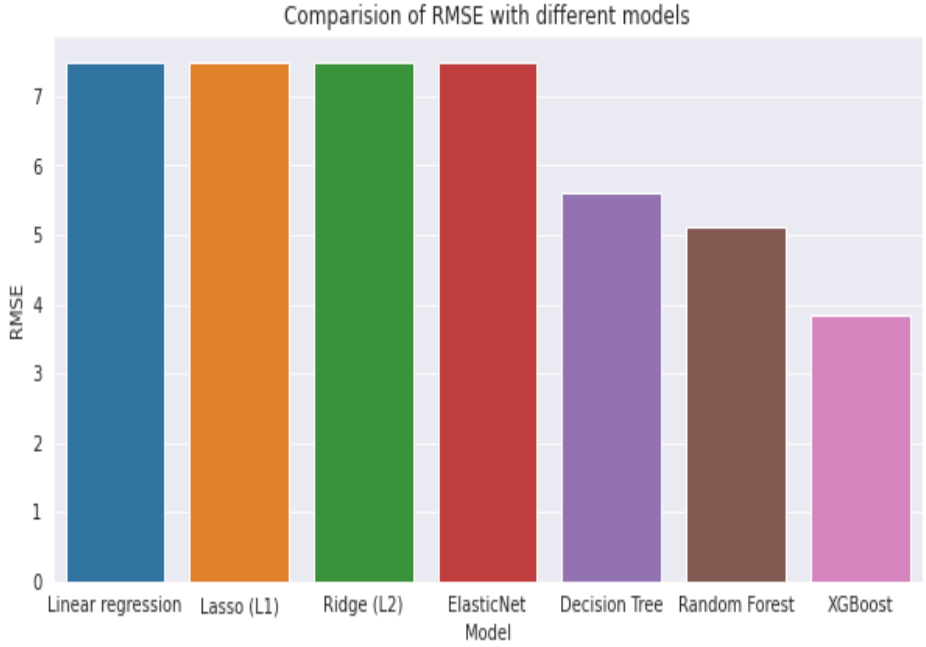
# Models Evaluation

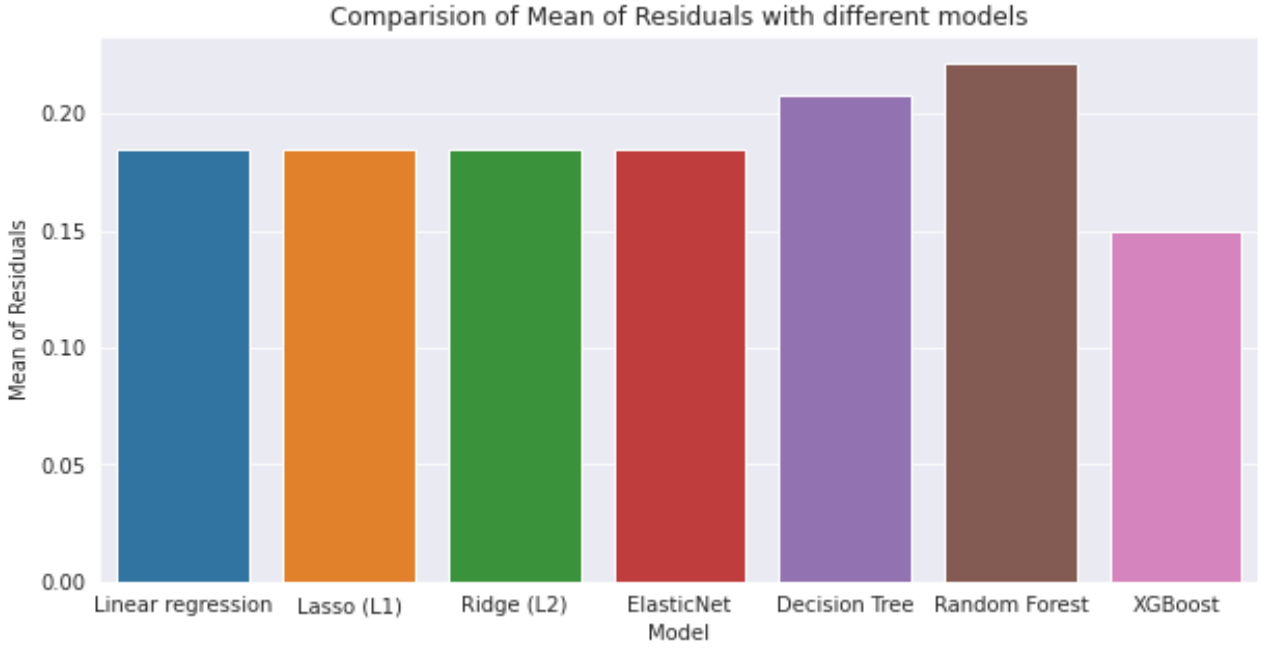
## ❑ Evaluation Metrics

| ML Model / Metrics | Linear regression | Lasso (L1) | Ridge (L2) | ElasticNet | Decision Tree | | Random Forest | | XGBoost |
|---|---|---|---|---|---|---|---|---|---|
| | TTS | CV | CV | CV | TTS | CV | TTS | CV | CV |
| MSE | 56.11 | 56.10 | 56.10 | 56.10 | 47.74 | 31.47 | 38.16 | 26.07 | 14.68 |
| RMSE | 7.49 | 7.49 | 7.49 | 7.49 | 6.53 | 5.61 | 6.17 | 5.10 | 3.83 |
| MAE | 5.79 | 5.79 | 5.79 | 5.79 | 4.91 | 3.85 | 4.74 | 3.62 | 2.43 |
| R-squared | 0.6231 | 0.6231 | 0.6231 | 0.6131 | 0.7128 | 0.7885 | 0.7436 | 0.8248 | 0.9013 |
| Ad. R-squared | 0.6205 | 0.6205 | 0.6205 | 0.6205 | 0.7109 | 0.7871 | 0.7419 | 0.8236 | 0.9007 |
| Mean of Residuals | 0.1847 | 0.1849 | 0.1847 | 0.1847 | 0.2473 | 0.2073 | 0.2471 | 0.2243 | 0.1494 |

# Comparison of Evaluation Metrics :



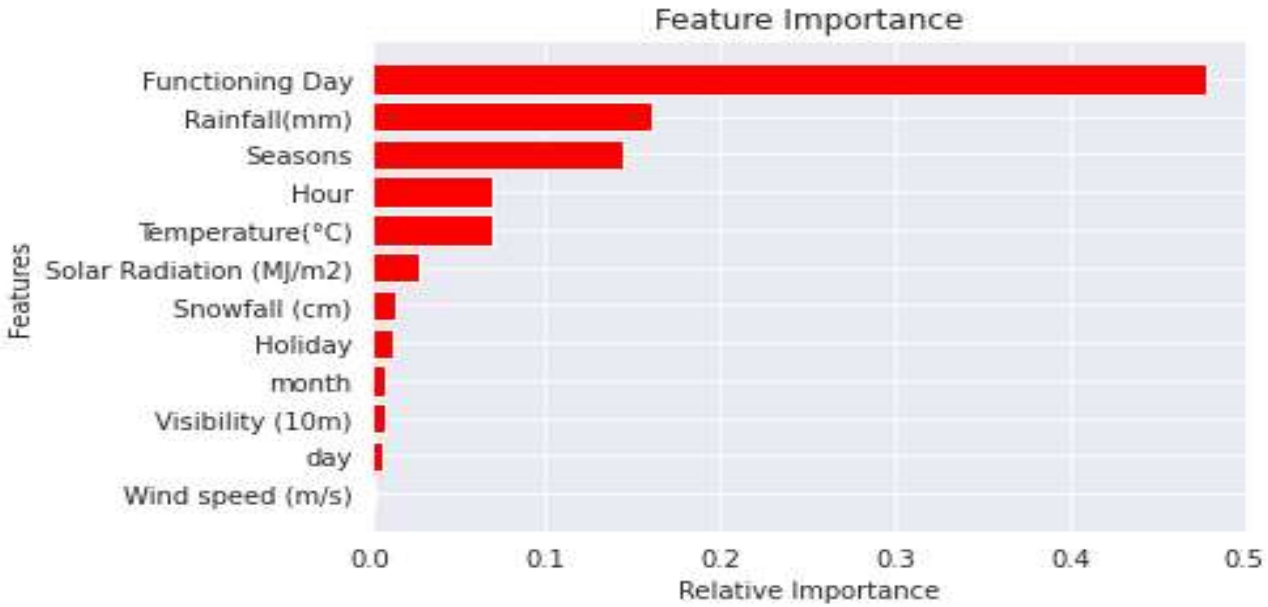






- ***The XGBoost regression model has the highest R-squared score, the lowest Root Mean Squared Error (RMSE), and has very close to having zero mean of residuals.***
- ***Therefore, the XGBoost regression model is the ideal model and well-trained for forecasting the number of rented bikes required per hour based on the model's high accuracy, low error, and zero mean of residuals.***

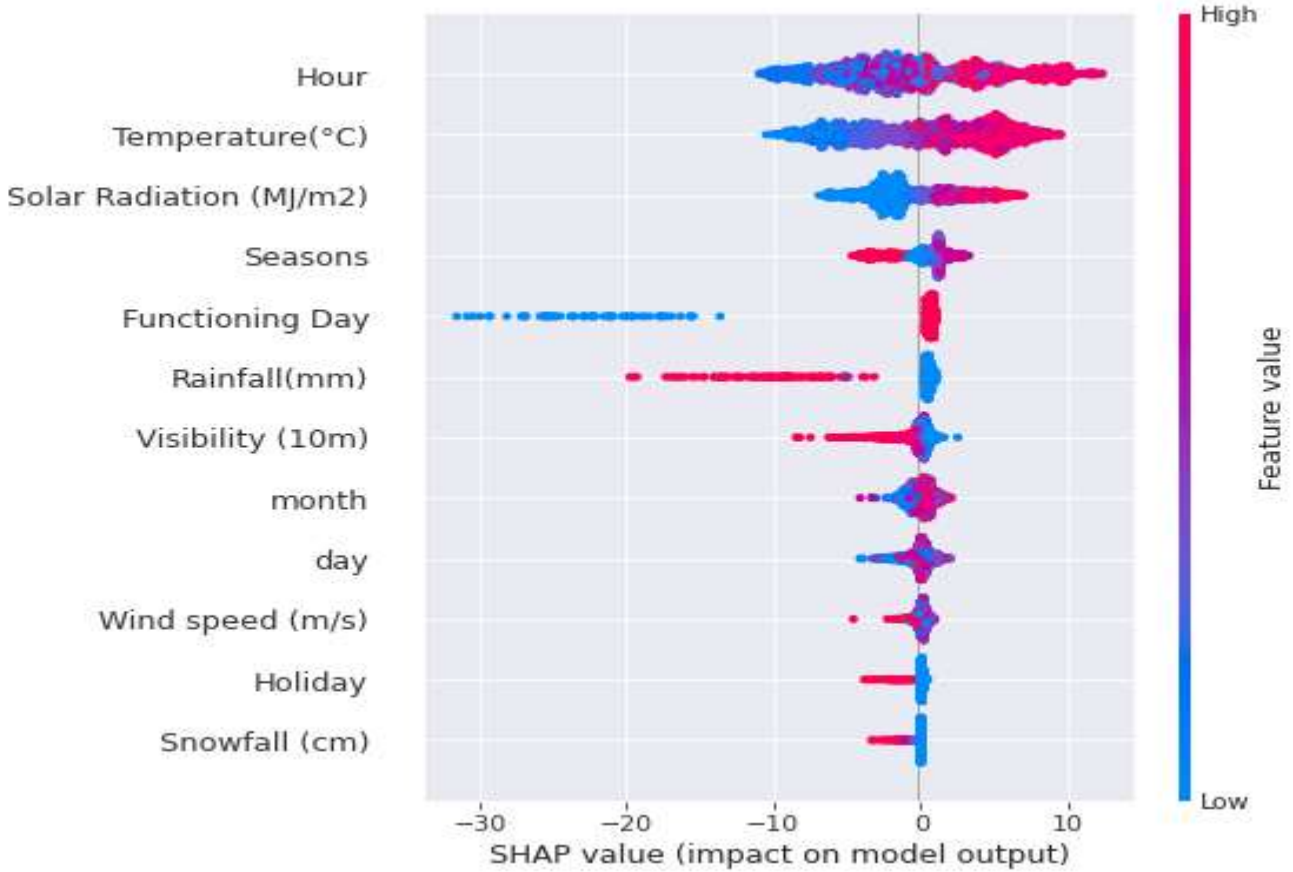
# Model Explainability

**Feature Importance :**




- **A higher feature importance score for features : Functioning Days, Rainfall (mm), and Seasons, respectively, indicates that those specific features have a larger influence on the model used to forecast a certain variable, respectively.**

**Model Explainability :**



## Observations :

- The model-impacting features (Importance) are listed in descending order.
- For that particular feature, the blue values are low and the red-colored values are high.
- High values from the Hour, Temperature (°C), Solar Radiation (MJ/m2), and Functioning Day features have a positive impact on the model, while low values have a negative impact.
- High values of the variables for Wind speed (m/s), Holidays, Snowfall (cm), Visibility (10m), Seasons, and Rainfall (mm) have a negative impact on the model, while low values have a positive impact.

# Conclusion

- **The XGBoost regression model has the highest R-squared score, and the lowest Root Mean Squared Error (RMSE), and has very close to having zero mean of residuals.**

- **As a result of the model's high accuracy, low error, and zero mean of residuals, the XGBoost regression model is the ideal and well-trained model for forecasting the number of rented bikes required per hour.**

# Challenges

AI

- **Data Preprocessing:**

Data preprocessing is an essential step in any machine learning project, and we faced difficulties identifying and fixing errors in the data.

- **Feature Engineering:**

we faced difficulties choosing the right features, but it was difficult to determine which ones were most important.

- **Algorithm Selection:**

Choosing the right algorithm is critical to the success of the model. It was difficult to determine which algorithm would be most effective for a particular problem.

- **Model Training:**

Training the model requires a lot of resources and can take a long time. It was important to choose the right parameters for the model in order to achieve the best performance.

- **Model Evaluation:**

Evaluating the performance of the model is essential to determining how well it is performing. It was important to choose the right metrics for evaluating the model and to interpret the results correctly.

# Thank You !