# Capstone Project - 2

# Bike Sharing Demand Prediciton


**Team Member**

**Vinay Kumar**

**Nikhil Lamje**

# Let’s Predict the Demand

1. Defining Problem Statement
2. Exploratory Data Analysis
3. Feature Engineering
4. Preparation of Dataset for Modeling
5. Applying Model
6. Model Validation and Selection

# What is the need

- Bikes have long been an important part of city transportation. As a result, bike sharing has received a lot of attention recently all around the world. Customers who utilise bike-sharing programmes want to be able to get a bike as soon as they need one.

- Because there are so many underlying variables, such as time of day, day of week, weather, and contributor correlation, predicting bike demand is difficult.

- As a result, bike rental companies must distribute bikes effectively depending on demand.

# Data Summary

Our DataFrame has 8750 records and 14 columns.

## Time Series

1. Date : timestamp

## Categorical Data

1. Season : 1 = spring, 2 = summer, 3 = autumn, 4 = winter
2. holiday : whether the day is considered a holiday
3. Functioning day : whether the day is neither a weekend nor a holiday

# Data Summary

## Numerical Data

1. **Humidity (%)** : Relative humidity
2. **Windspeed (m/s)** : Wind speed in meter per second
3. **Visibility (10m)** : Visibility on the roads in winter and foggy days.
4. **Dew Point Temperature (C)** : Temperature at which air cannot hold water.
5. **Solar Radiation (MJ/m2)** : Solar radiation favours the sunny day which is good weather for bike.
6. **Rainfall (mm)** : Rainfall
7. **Snowfall (cm)** : Snowfall
8. **Rented Bike Count** : number of bikes has been rented at some specific hour.

# Features

**Independent Features**

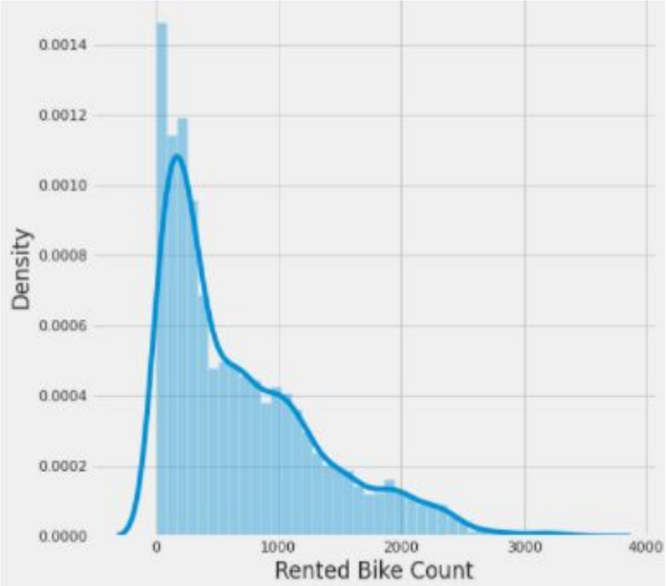
1. Humidity (%)
2. Windspeed (m/s)
3. Visibility (10m)
4. Rainfall (mm)
5. Snowfall (cm)
6. Hour
7. Seasons

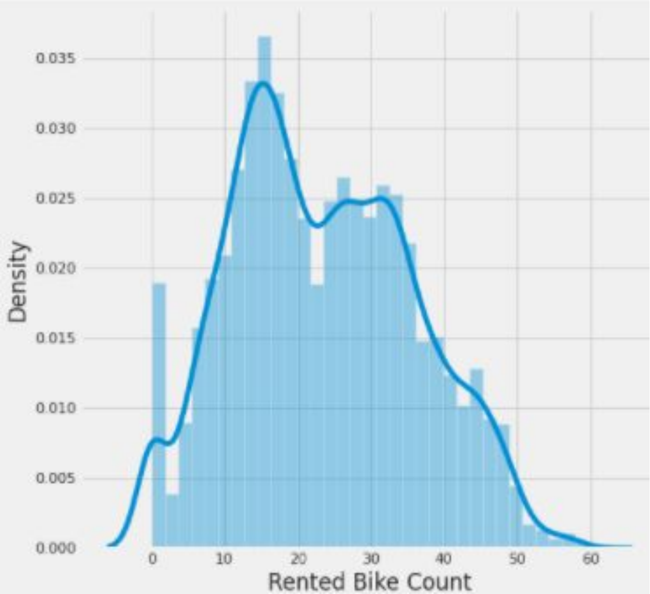
**Dependent Feature**

1. Rented Bike Count

# Dependent Variable

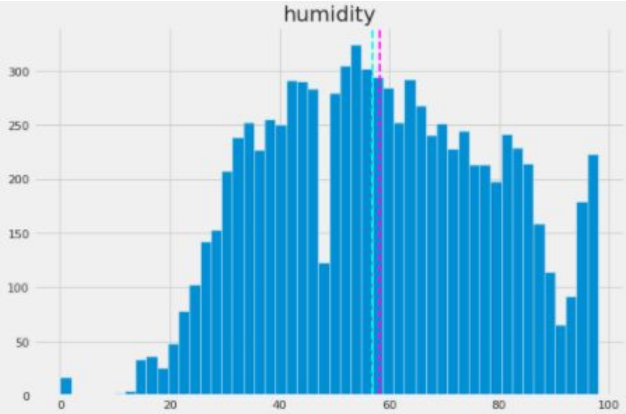
## Square root transformation

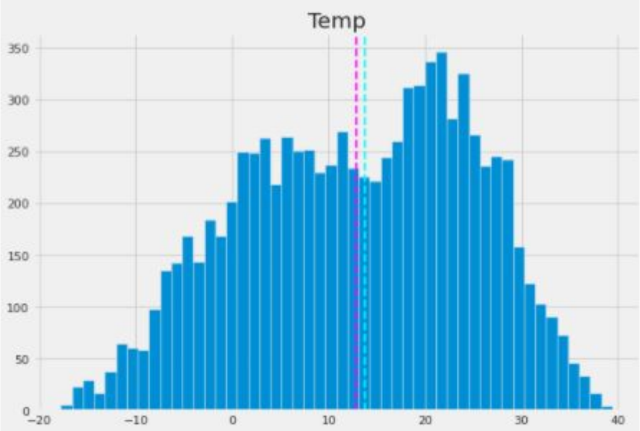


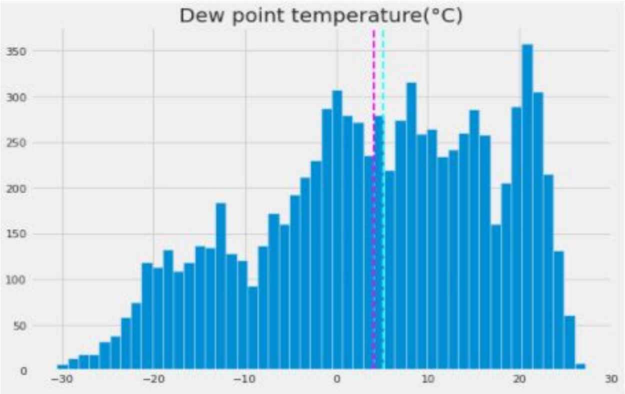



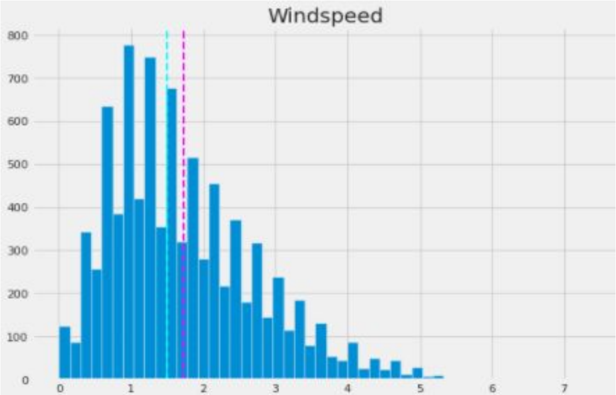

- **Normalizing a skewed distribution**

# EDA

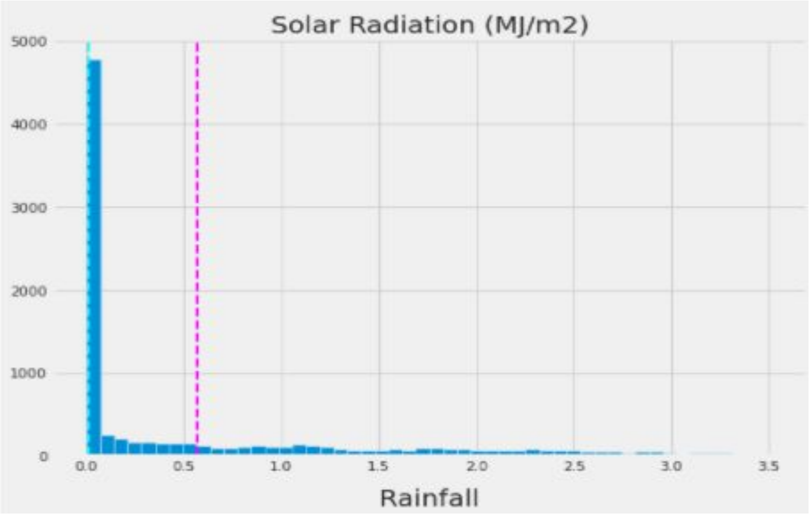

humidity
0
50
100
150
200
250
300
0
20
40
60
80
100


Temp
0
50
100
150
200
250
300
350
−20
−10
0
10
20
30
40

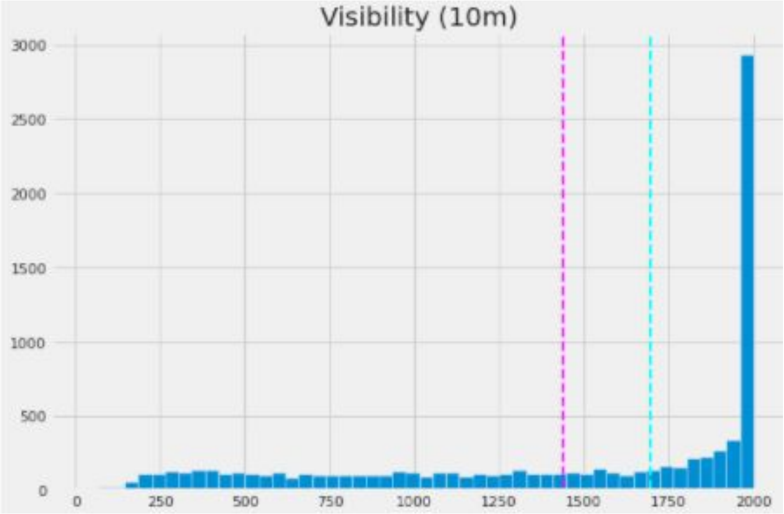

Dew point temperature(°C)
0
50
100
150
200
250
300
350
−30
−20
−10
0
10
20
30


Windspeed
0
100
200
300
400
500
600
700
800
0
1
2
3
4
5
6
7

# EDA

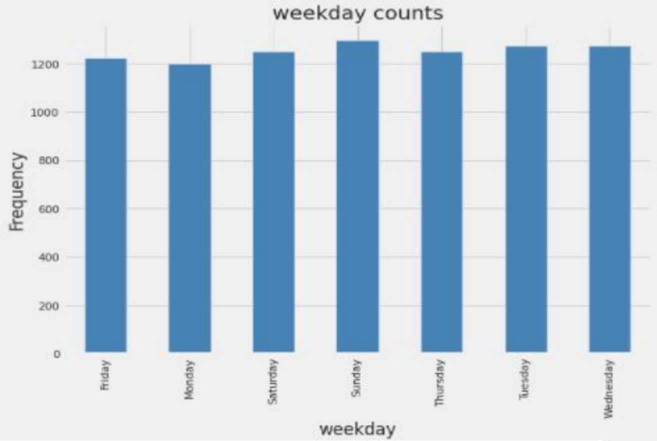


Solar Radiation (MJ/m2)
5000
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Rainfall

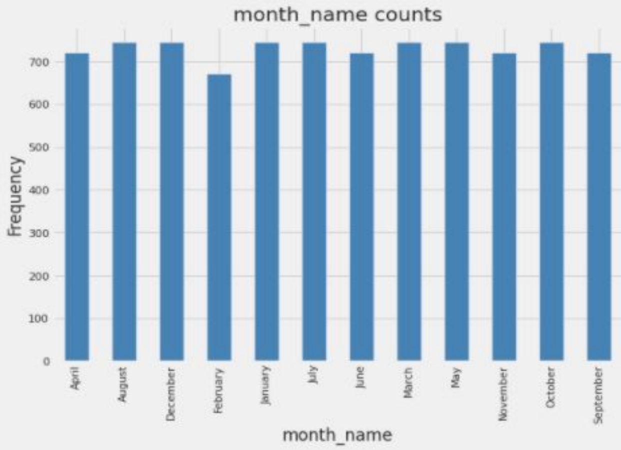



Visibility (10m)
3000
2500
2000
1500
1000
500
0
0
250
500
750
1000
1250
1500
1750
2000

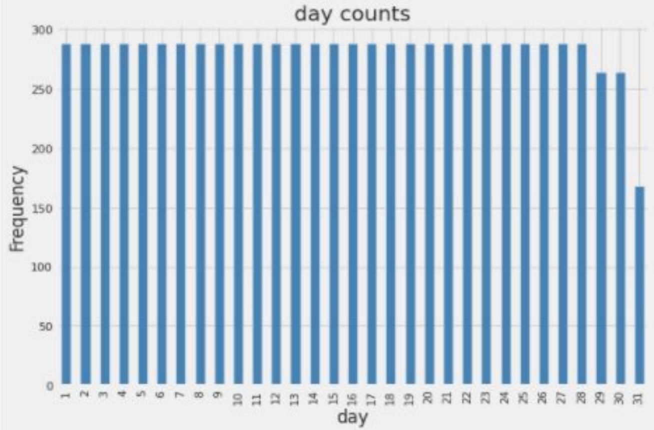
# EDA



weekday counts
Frequency
0
200
400
600
800
1000
1200
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
weekday

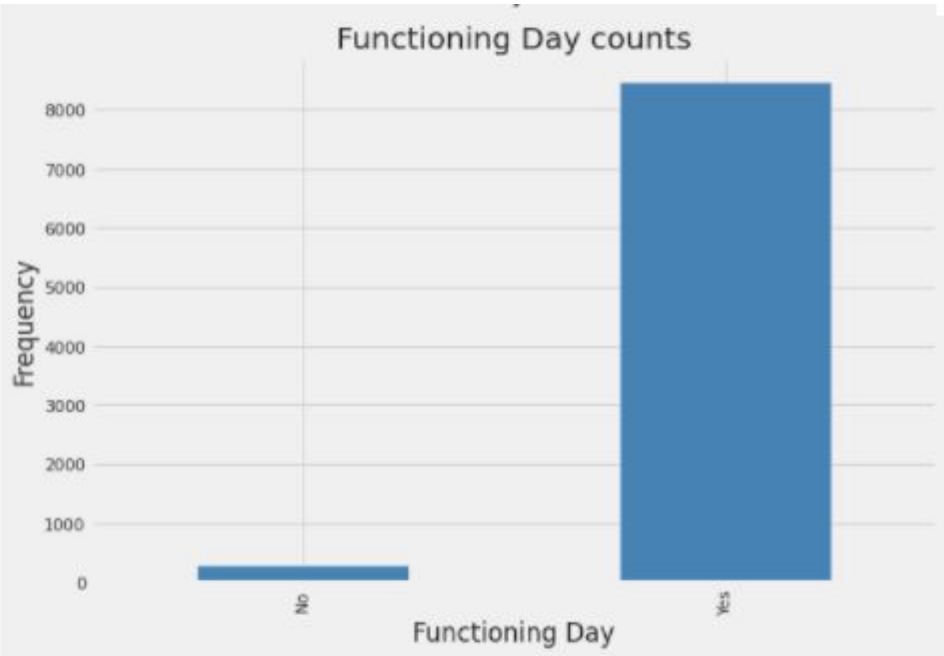



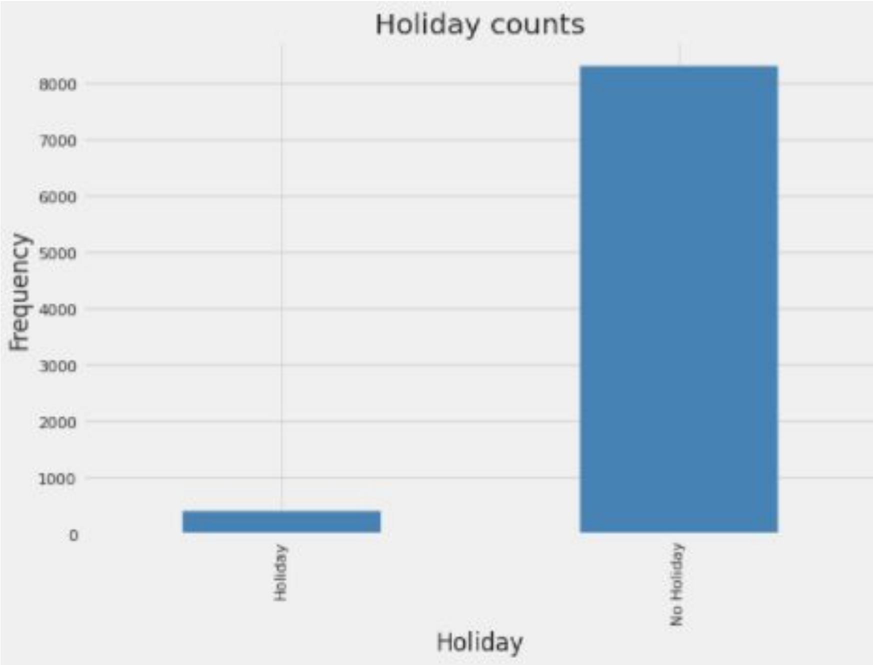
month_name counts
Frequency
0
100
200
300
400
500
600
700
April
August
December
February
January
July
June
March
May
November
October
September
month_name

day counts
Frequency
0
50
100
150
200
250
300
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day

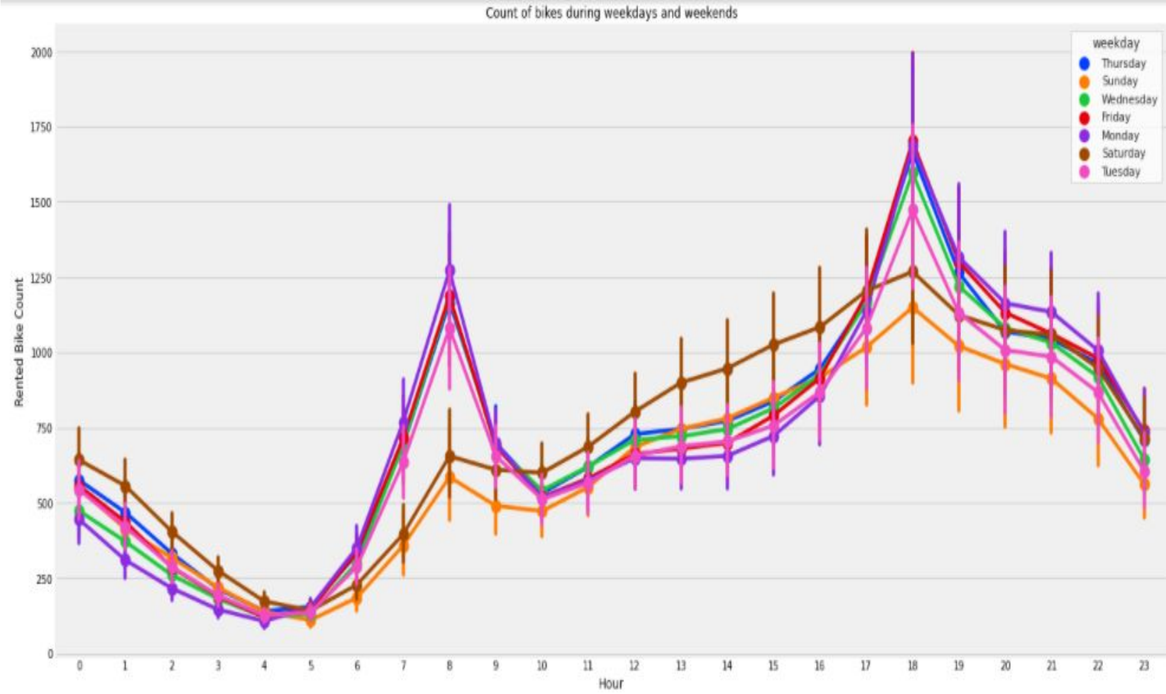
# EDA


Functioning Day counts
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
No
Yes
Functioning Day

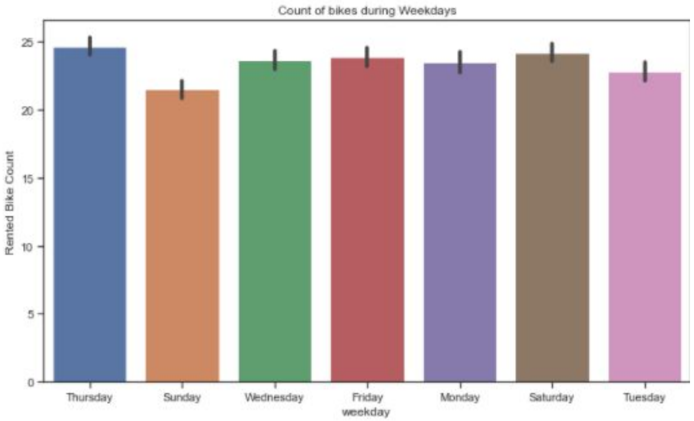

Holiday counts
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
Holiday
No Holiday
Holiday

# EDA

* Monday, Tuesday, Wednesday, Thursday, and Friday all follow the same pattern. Between 0600 hours and 1000 hours, and between 1700 hours and 2100 hours, there were a lot of rented bikes. Due to office hours, there is a hurry and increase in frequency.

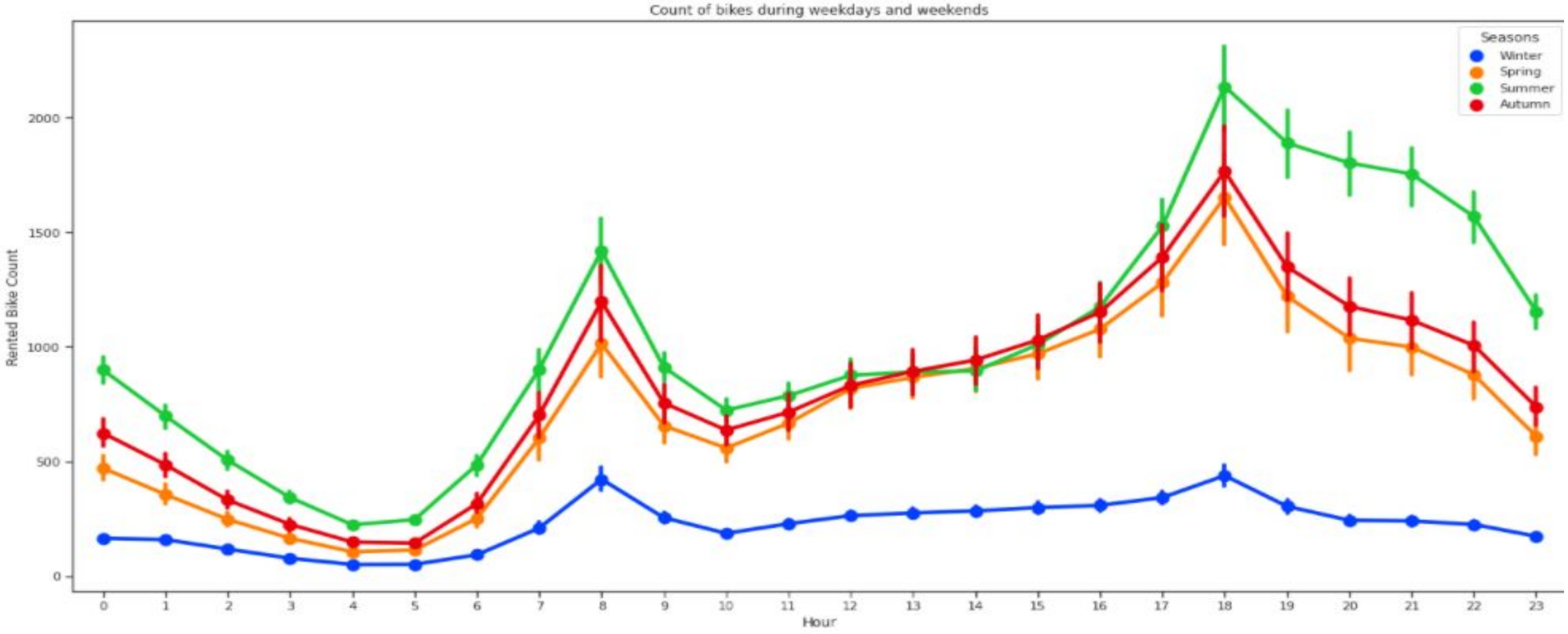



* We saw a different trend on Saturday and Sunday. The afternoon and early evening are the most popular times for people to leave their homes. That's why we witnessed a spike in the number of people during those hours.



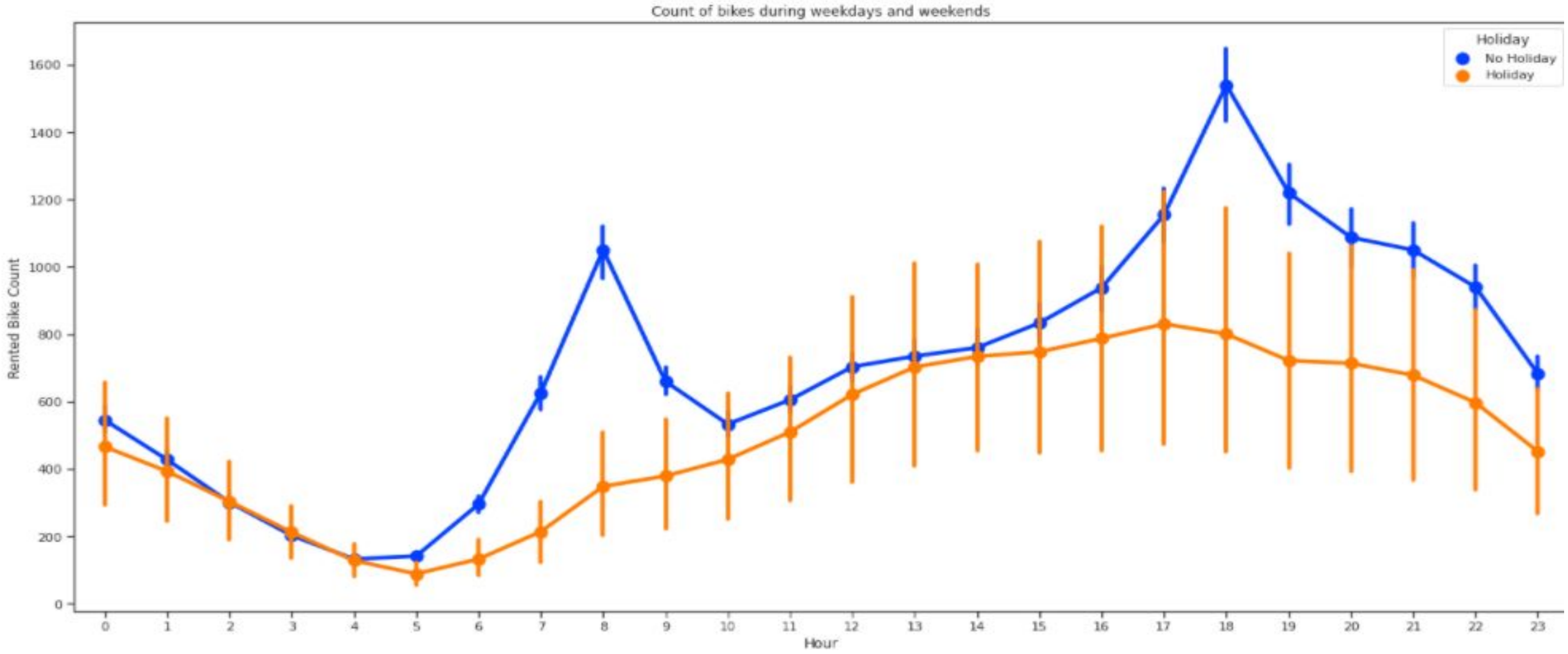
# EDA

We can see from the graph above that people do not prefer to ride bikes in the cold. Choosing to ride a bike during the snow may not be the most practical decision.



# EDA

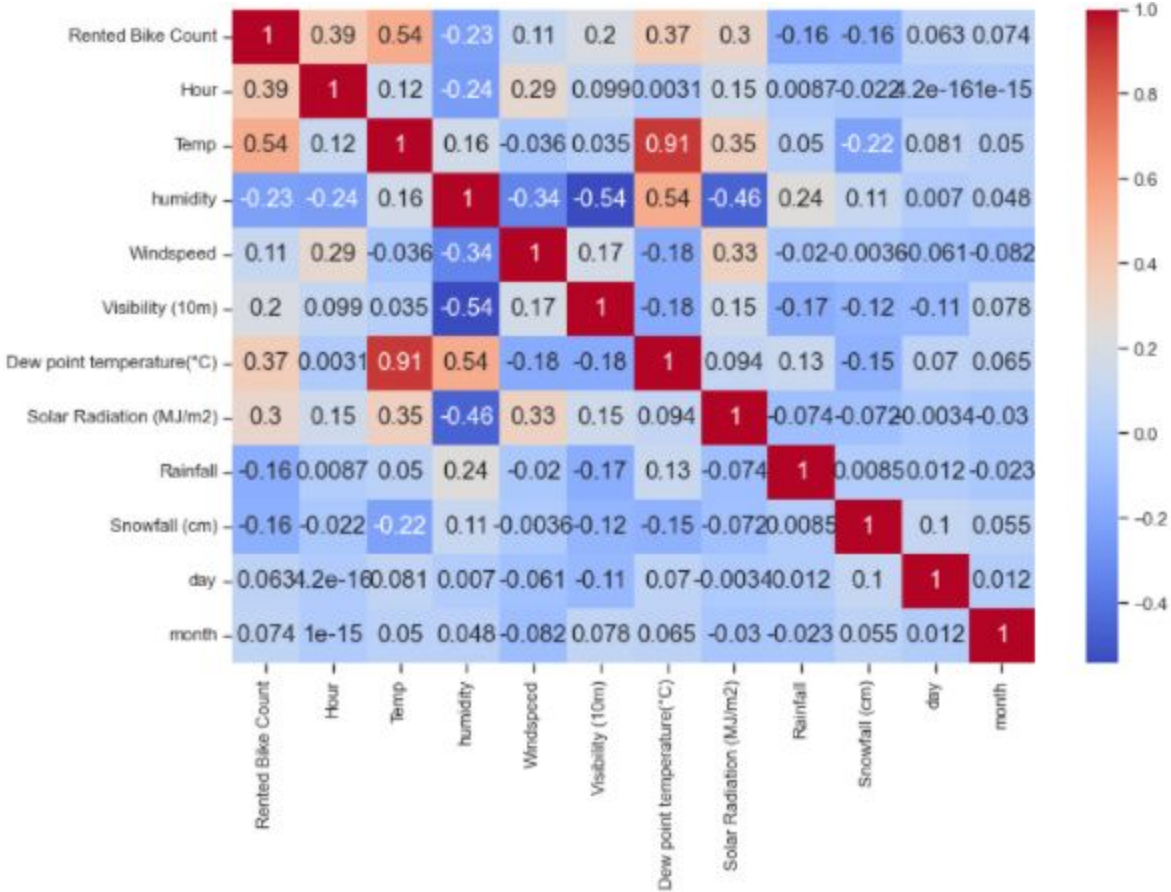
We're seeing a similar pattern here, such as Non-holidays are Monday through Friday, whereas holidays are Saturday and Sunday.



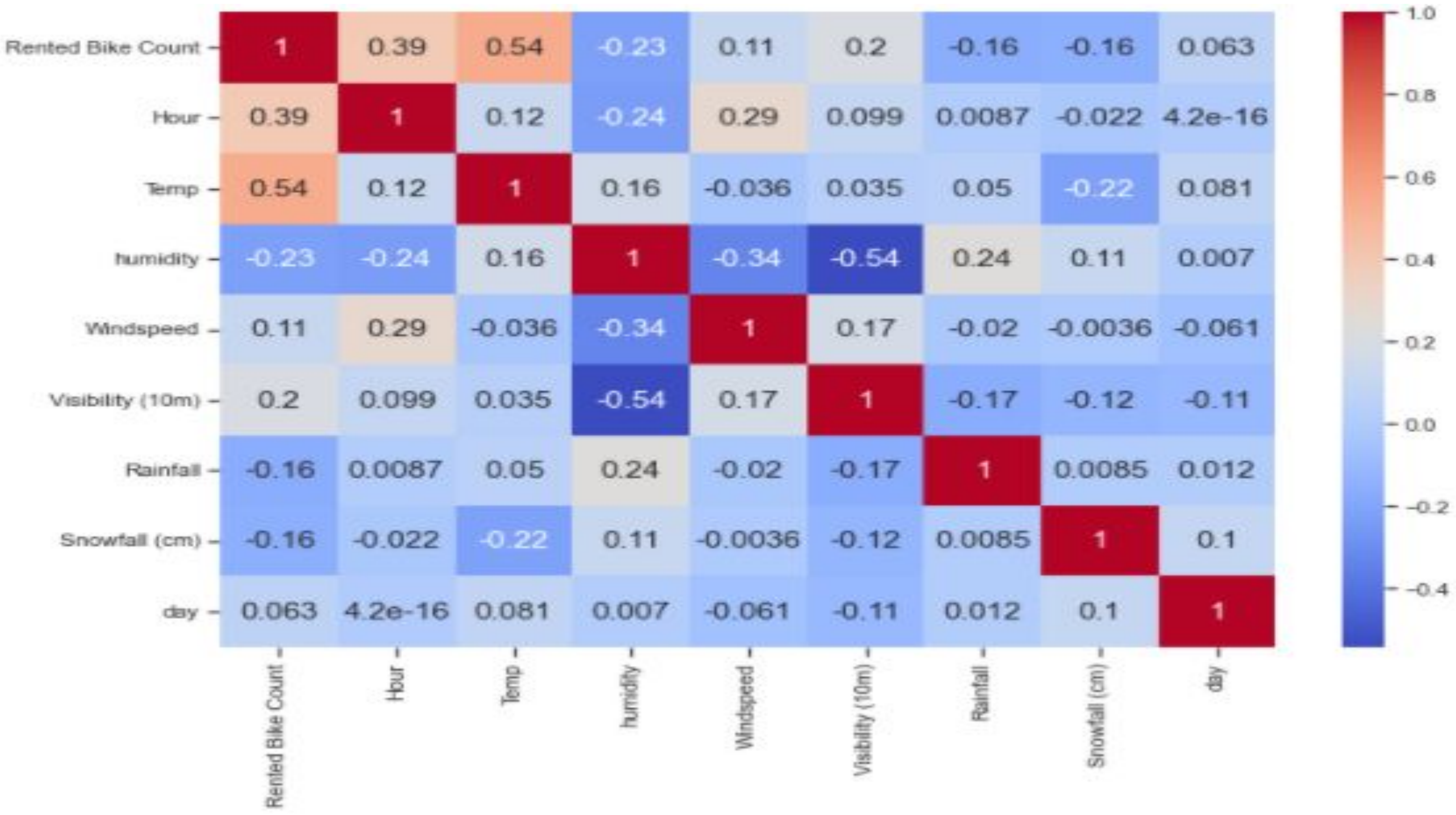
# Feature Selection

- Because there is a strong link between temperature and dew point temperature, we may ignore any of the columns, such as dew point temperature.

- Temperature and windspeed are also strongly associated to solar radiaiton. We'll have to drop that as well.



# Feature Selection



# Preparing Dataset for Modelling

**Train Set - (7008,53)**

**Test Set - (1752,53)**

| | Rented Bike Count | Hour | Temp | humidity | Windspeed | Visibility (10m) | Rainfall | Snowfall (cm) | Seasons_Spring | Seasons_Summer | Seasons_Winter | holiday_No Holiday |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15.937377 | 0 | -5.2 | 37 | 2.2 | 2000 | 0.0 | 0.0 | 0 | 0 | 1 | 1 |
| 1 | 14.282857 | 1 | -5.5 | 38 | 0.8 | 2000 | 0.0 | 0.0 | 0 | 0 | 1 | 1 |
| 2 | 13.152946 | 2 | -6.0 | 39 | 1.0 | 2000 | 0.0 | 0.0 | 0 | 0 | 1 | 1 |
| 3 | 10.344080 | 3 | -6.2 | 40 | 0.9 | 2000 | 0.0 | 0.0 | 0 | 0 | 1 | 1 |
| 4 | 8.831761 | 4 | -6.0 | 36 | 2.3 | 2000 | 0.0 | 0.0 | 0 | 0 | 1 | 1 |

# Applying Models

## Linear Regression

### Underfit

```
Model : LinearRegression(Test)
MAE : 29.98679738080441
MSE : 50.3643820657739
RMSE : 7.096786742306261
R2 : 0.6776946294181957
Adjusted R2 :  0.6676344500066318
```

```
Model :  LinearRegression(Train)
MAE : 29.54126553275913
MSE : 50.06618224296554
RMSE : 7.075746055573613
R2 : 0.676106087090919
Adjusted R2 :  0.673637525488362
```

## Decision tree

### Overfit

```
Model : DecisionTreeRegressor(Test)
MAE : 10.228640235377737
MSE : 25.25032330281588
RMSE : 5.024969980290019
R2 : 0.838411304266518
Adjusted R2 :  0.8333676052832198
```

```
Model : DecisionTreeRegressor(Train)
MAE : 0.0
MSE : 0.0
RMSE : 0.0
R2 : 1.0
Adjusted R2 :  1.0
```

## Randomforest

```
Model : RandomForestRegressor(Test)
MAE : 5.988899421341081
MSE : 14.259190996504975
RMSE : 3.776134398628441
R2 : 0.908748729760581 4
Adjusted R2 :  0.9059004863432144
```

```
Model : RandomForestRegressor(Train)
MAE : 0.872874278298 4237
MSE : 2.049641952322902
RMSE : 1.4316570651950495
R2 : 0.9867402201993593
Adjusted R2 :  0.9866391606179048
```

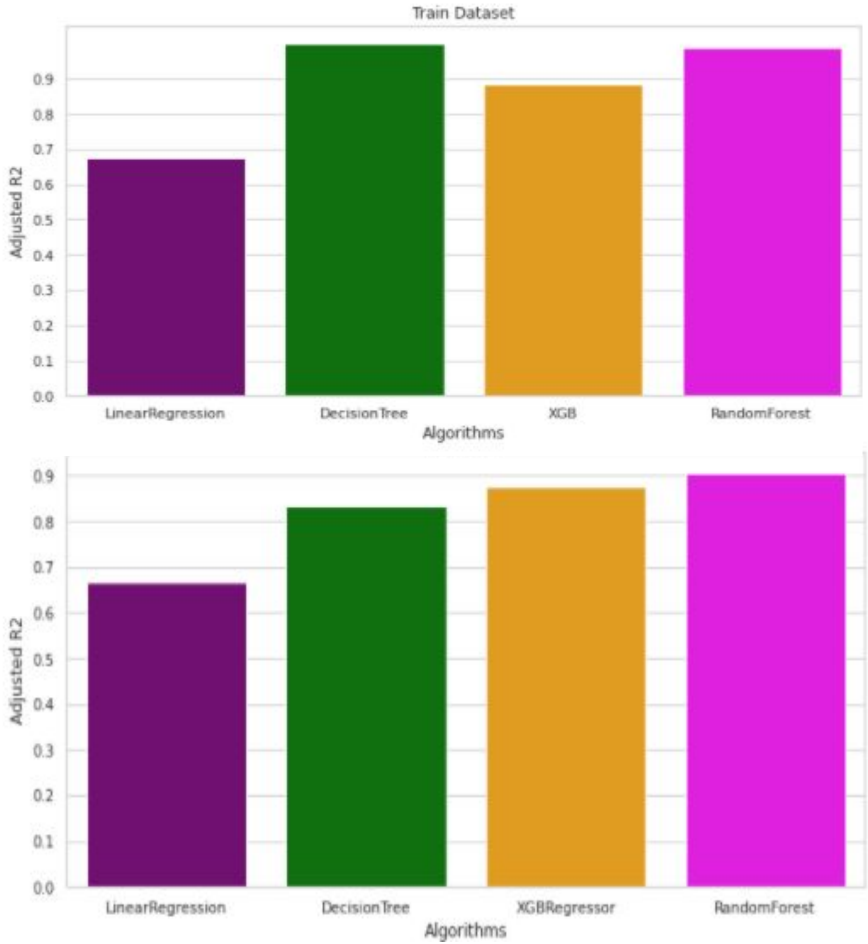
## XGBoost

```
Model : XGBRegressor(Test)
MAE : 9.823108748409625
MSE : 18.820259819834554
RMSE : 4.338232338157392
R2 : 0.8795603049838707
Adjusted R2 :  0.8758009976600457
```

```
[13:28:10] WARNING: /workspace/src
Model : XGBR(Train)
MAE : 9.3438932056 1393
MSE : 18.006181751521684
RMSE : 4.243369150983884
R2 : 0.883512335017877 7
Adjusted R2 :  0.882624522788362
```

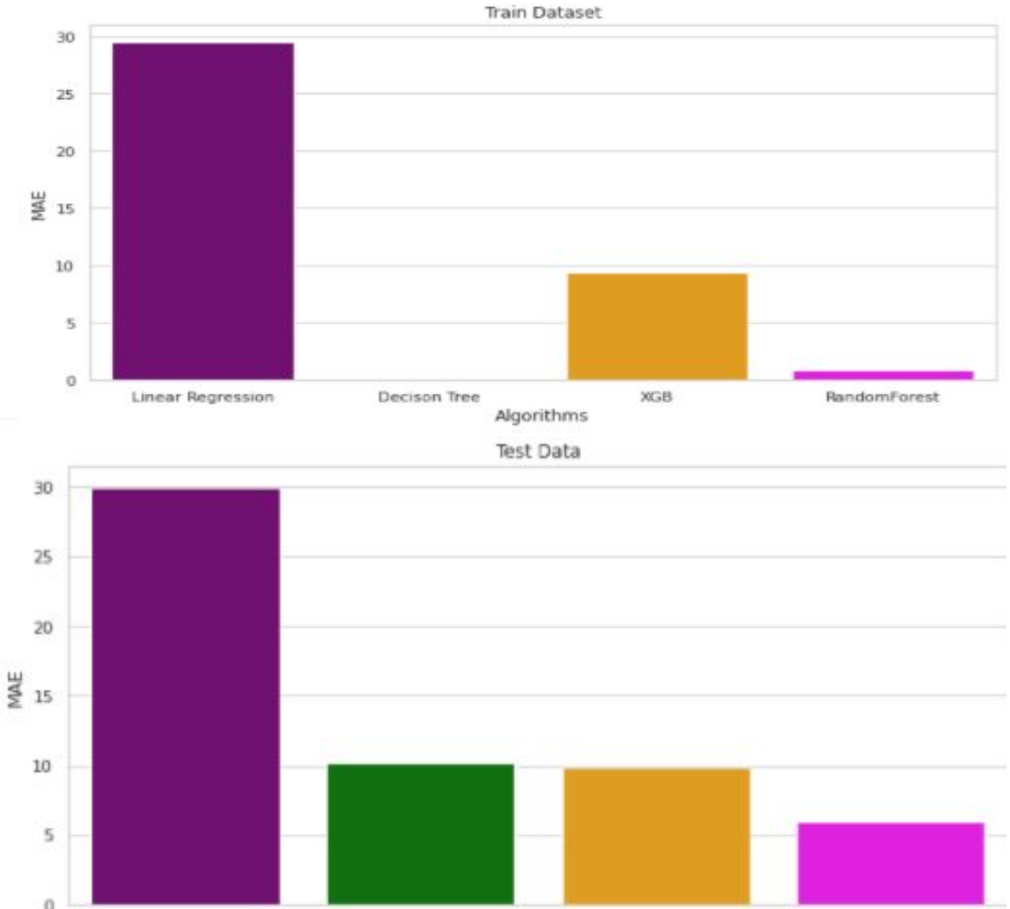
# Model Performance

- the R2 score in linear regression is 0.66, which is the lowest of all of the models.
- With an R2 score of 0.88 on training data and 0.87 on test data, XGBoost produces the best results.



# Model Performance

- We used MAE to evaluate the models before deciding on one.
- We get better results with the XGBoost regressor, with an MAE of 9.34 and 9.89 on the test and training datasets, respectively.
- On XGBoost, we'll now perform Hyperparameter optimization.



# Hyperparameter Optimization

```
Model : XGBRegressor Gridsearch
MAE : 4.680984255449909
MSE : 11.248557977113744
RMSE : 3.353886995280811
R2 : 0.9280151865540645
Adjusted R2 :  0.9257683107515706
```

```
{'learning_rate': 0.1,
 'max_depth': 7,
 'min_child_weight': 1,
 'n_estimators': 500,
 'objective': 'reg:squarederror'}
```

**Performance Metrics**

**Best Parameters**

- The best results were obtained utilising hyperparameter adjustments, with MAE as low as 4.68.

# Conclusion

- This study proposed the use machine learning techniques to identify the demands in a bike-sharing system.

- Indeed, selecting the appropriate features to get the intended outcome was crucial, with the hour and temperature features being the most closely related to the target variable.

- During the model development, we found that some models were overfit and underfit on the training data, whereas the XGBoost regressor appears to be a perfect fit on both test and training data.

- By tweaking the hyperparameters of the XGBoost regressor, we were able to achieve the desired outcomes, with an adjusted R2 score of 0.92 on the test data, up from 0.87 earlier.