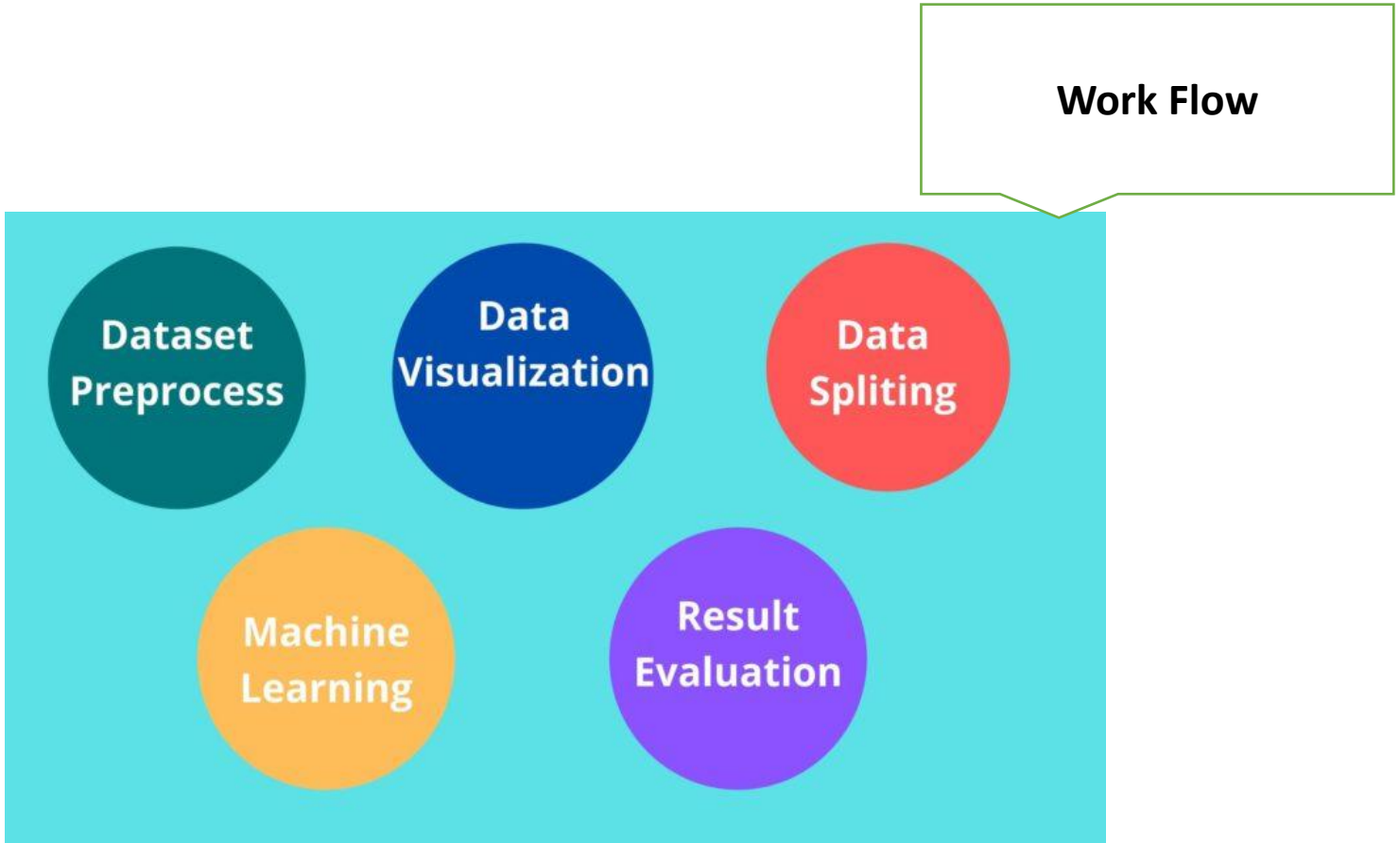
Seoul Bike Sharing Demand Prediction

By Soniya Kumawat

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.

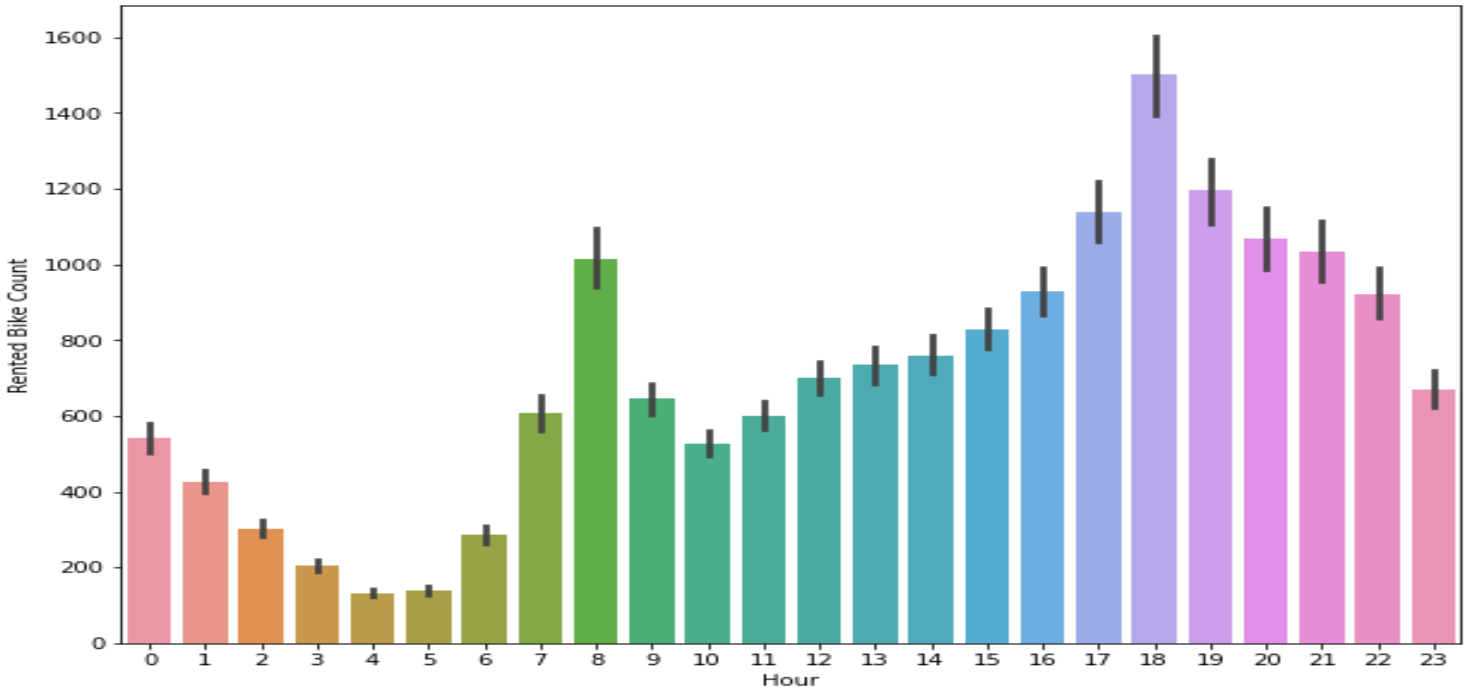


Data understanding

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- 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)

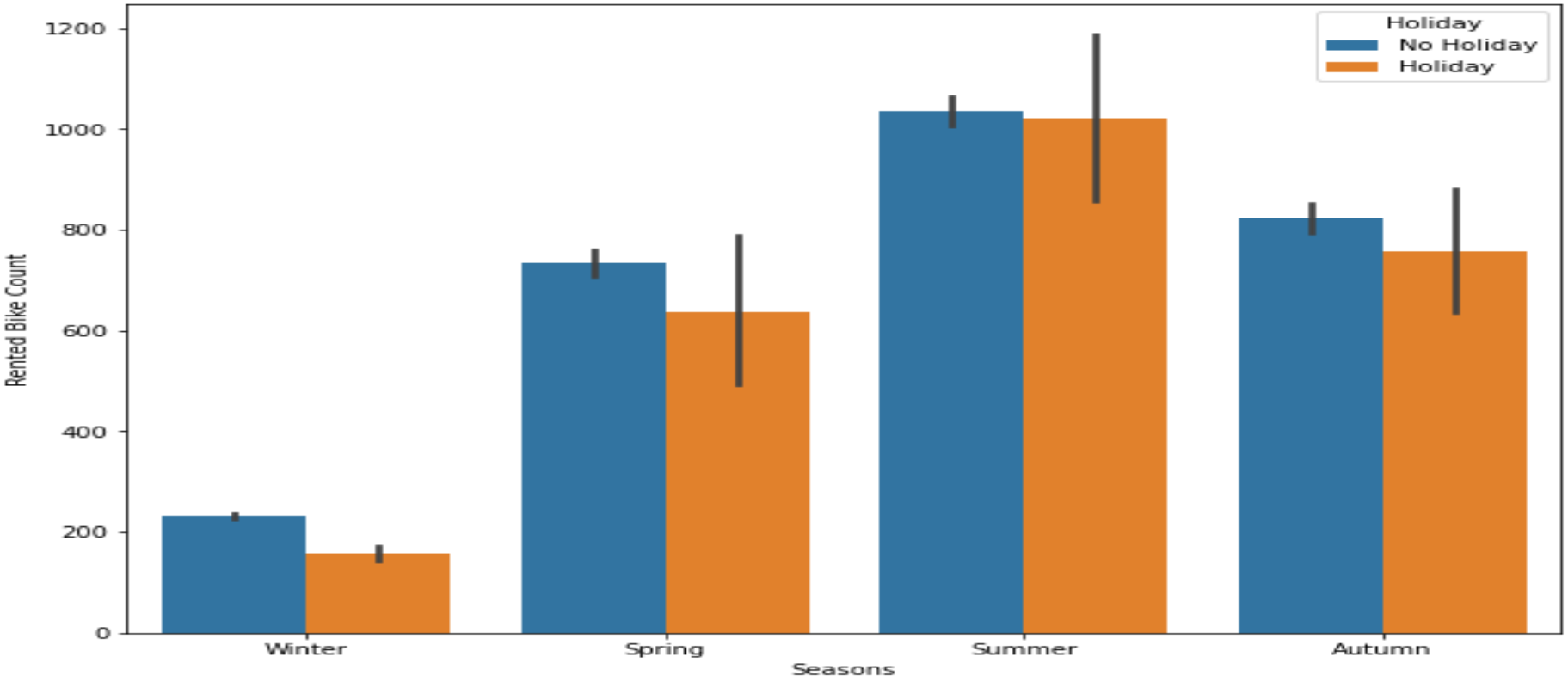
Exploratory data analysis

Hour and Rented count graph



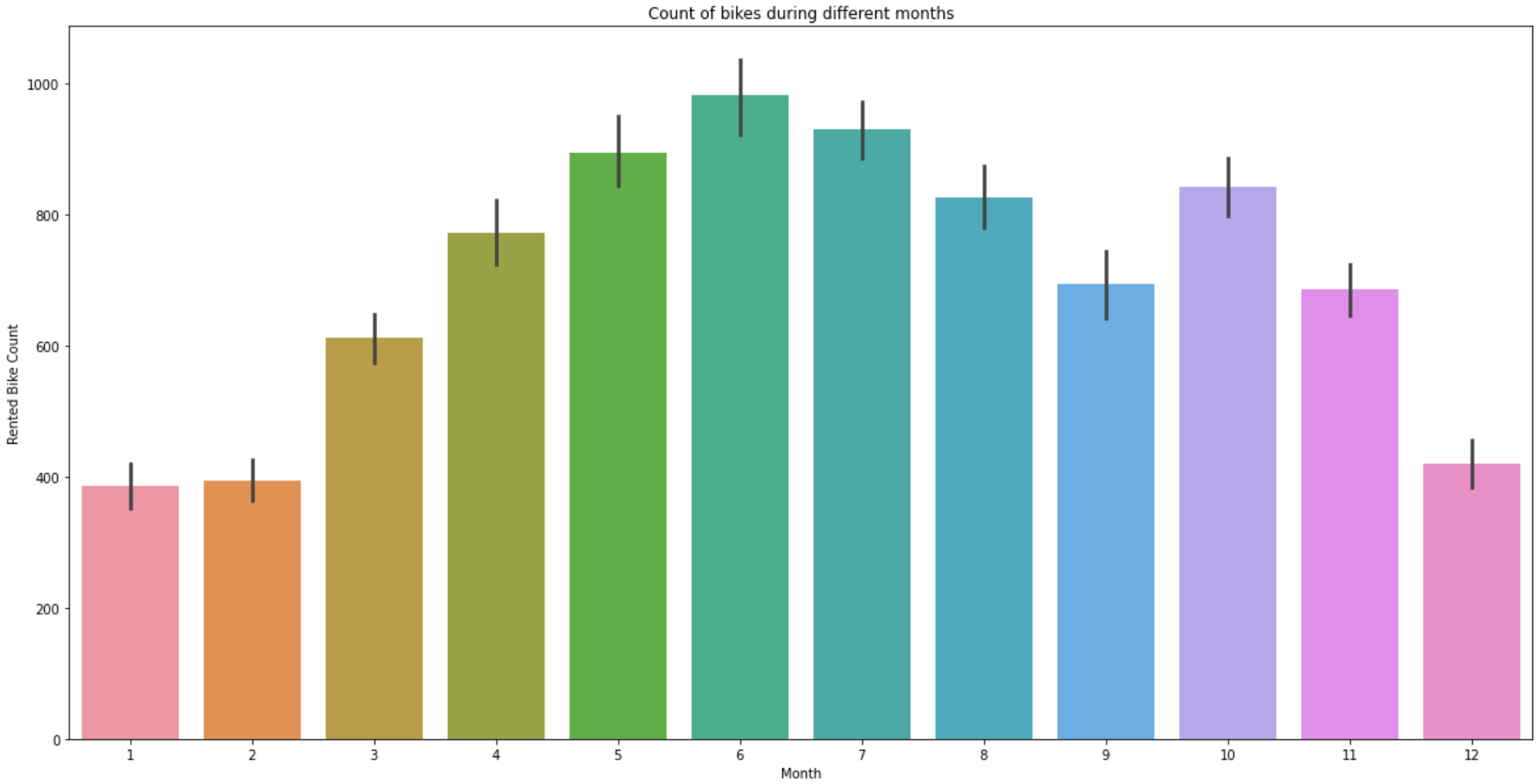
1) bike is rented out for 18 hours mostly
2) people are renting bikes for hours ranging 10 to 24 the most might be they are using bike for offices.

Graph Between rented hour count and Seasons

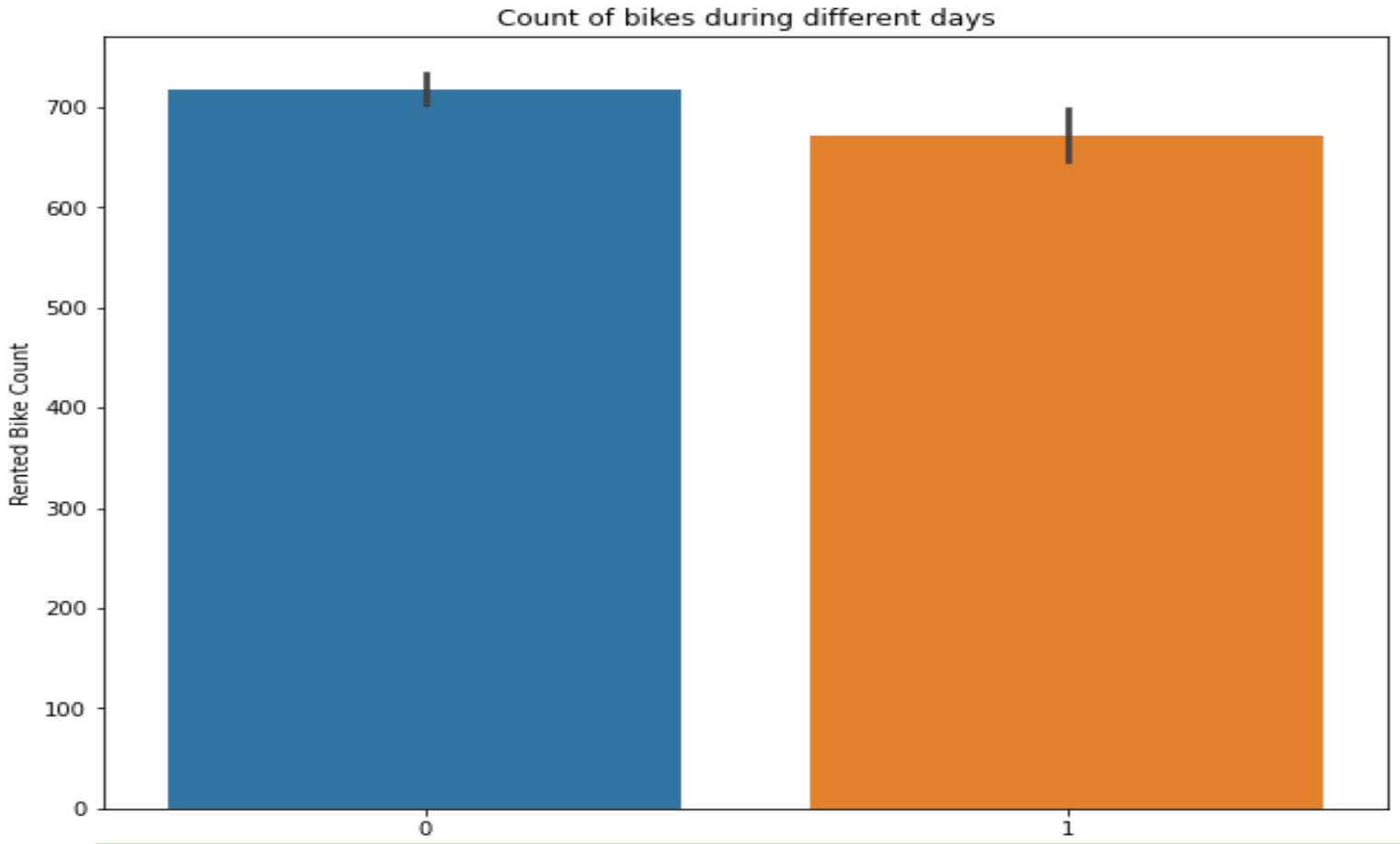


1) People are using bike on holidays as well as no holiday equally
2) people are using bike mostly in summer season and autumn

Graph Between rented hour count and Months



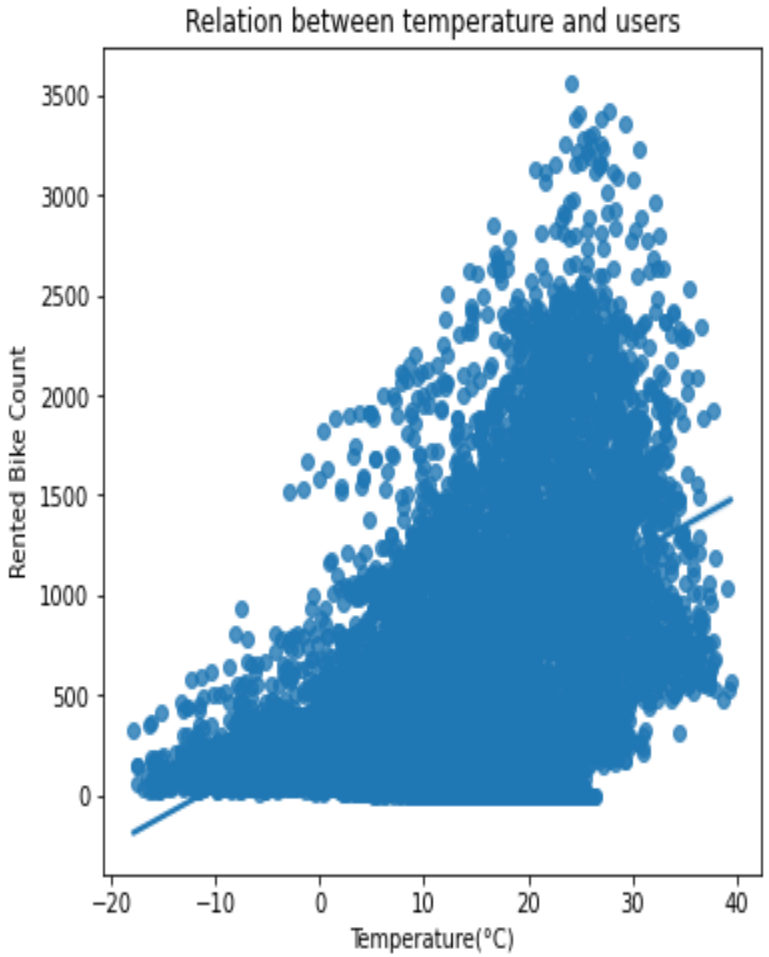
Jan, Feb December months have very less count might be because of winter season

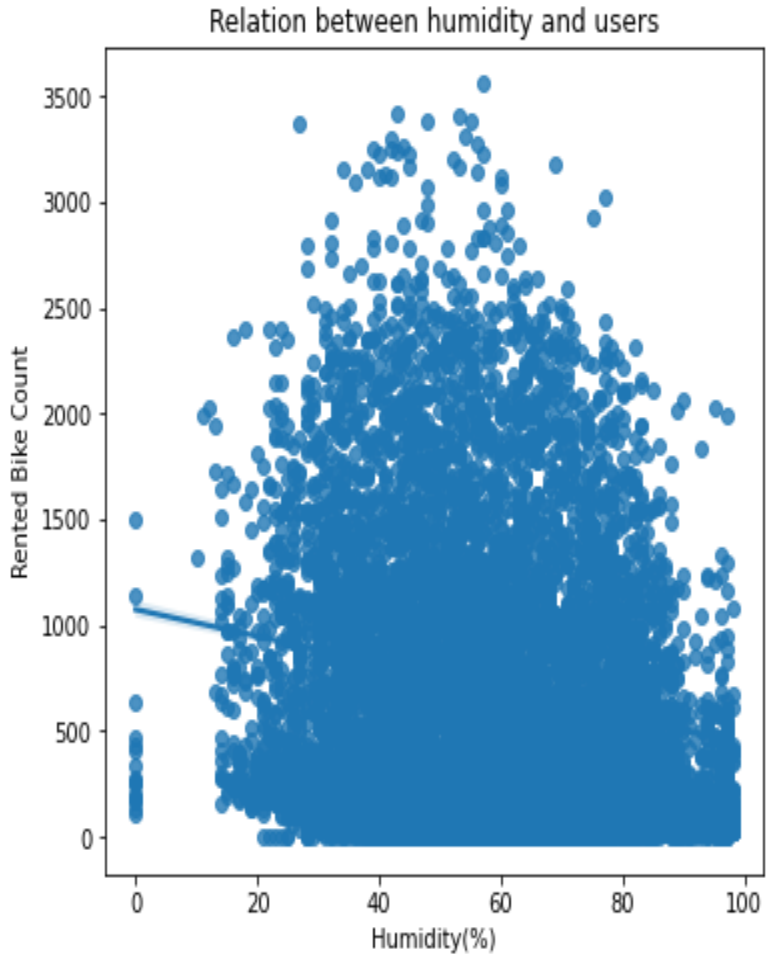


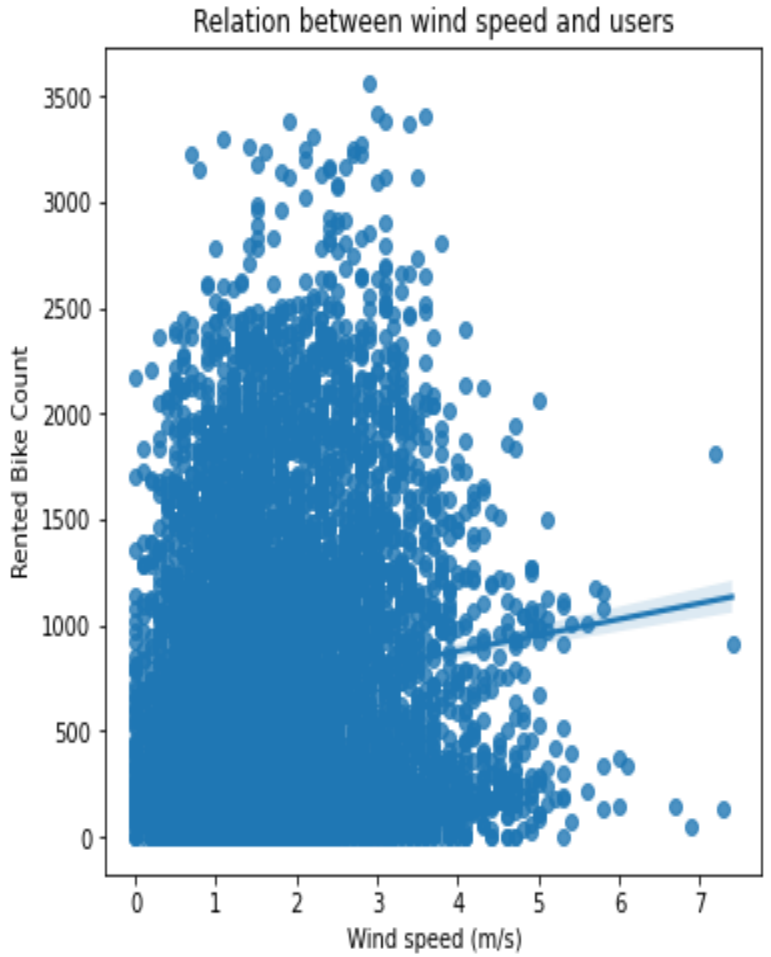
1=weekends

0= weekdays

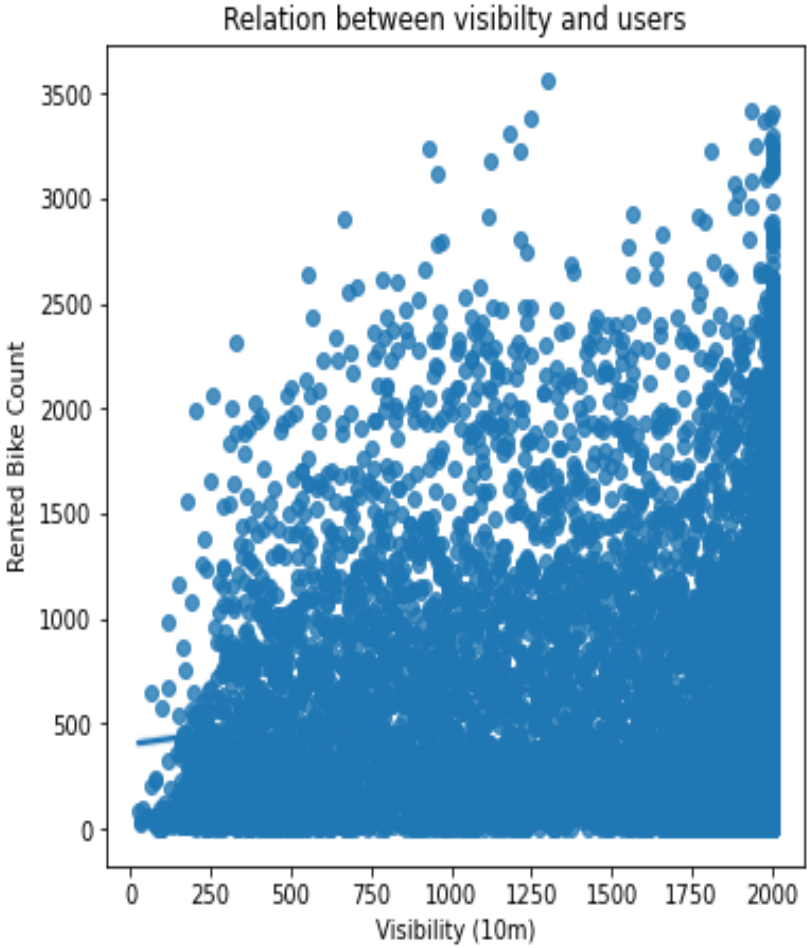
weekdays and weekends in both people are using bike almost equally

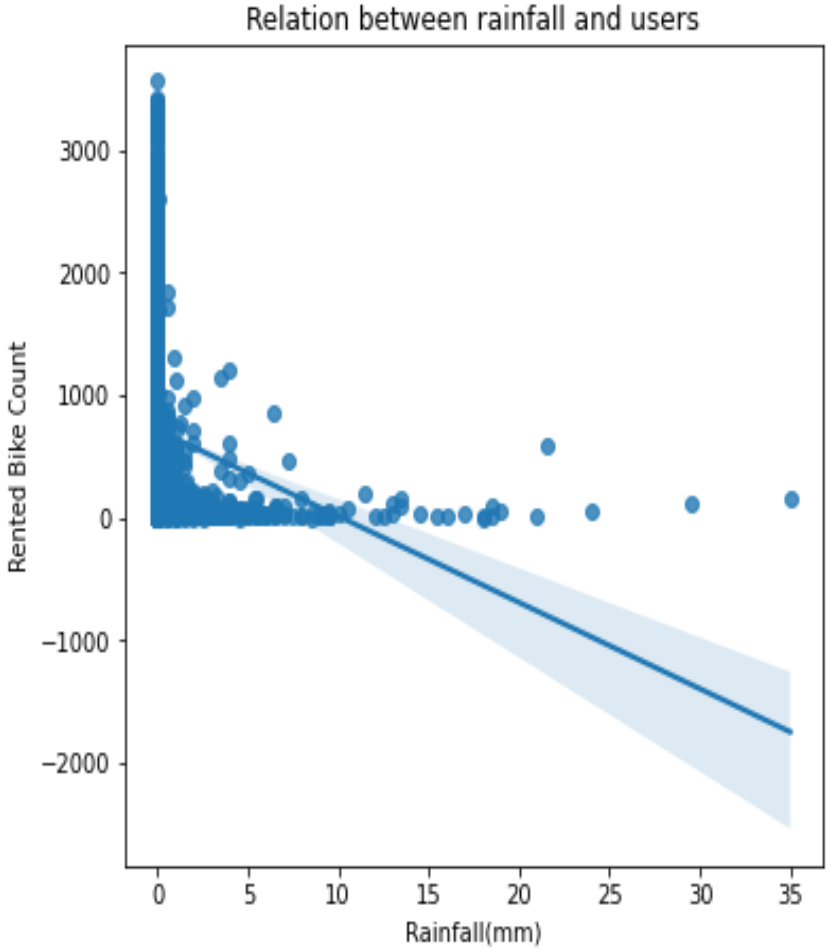


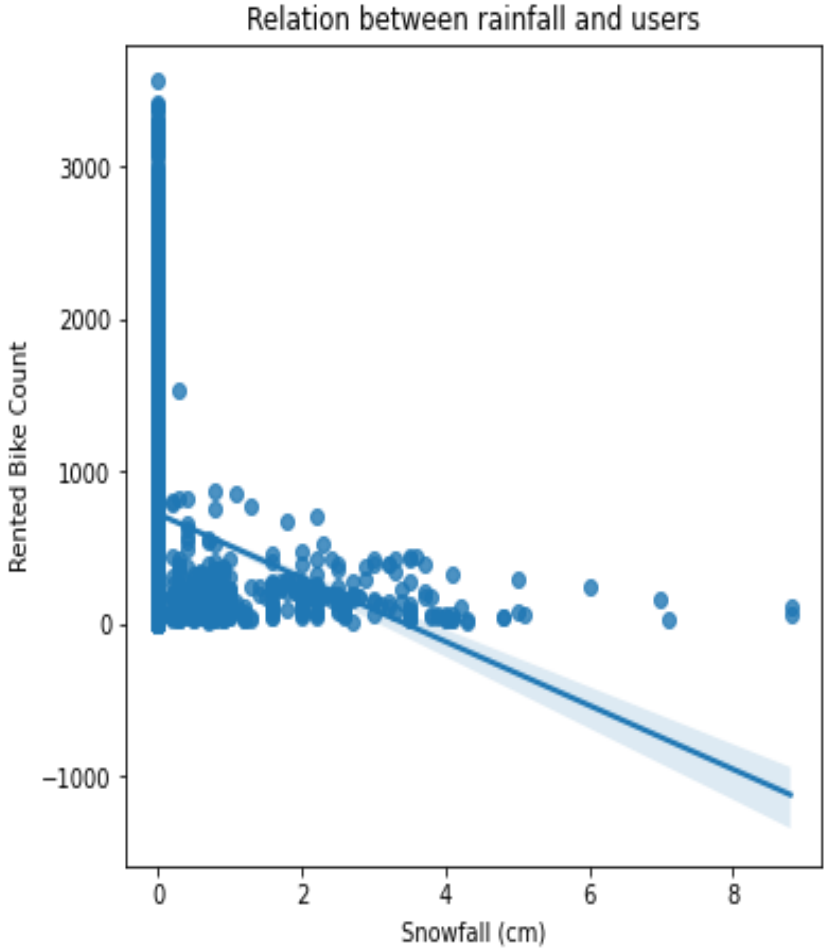




As the temperature and humidity increases users are increasing vice versa with wind speed







- ❑ Visibility increases user are increasing
- ❑ Rain and snow increasing user are decreasing

We see that there is a hyper correlation between columns 'Temperature(°C)' and 'Dew point temperature(°C)' so we can drop the column 'Dew point temperature(°C)'. And they have the same variations.

Conclusion

- Bike is rented out for 18 hours mostly
- people are rented bike for hours ranging 10 to 24 the most might be because people are renting bikes for office work
- As the temperature and humidity increases users are increasing
- people are preferring to ride bike when its little windy but not too much windy
- Visibility increases user are increasing
- rain and snow increasing user are decreasing
- functioning days are very much as compared to non functioning days no people can rent a bike on non functioning days
- weekdays and weekends in both people are using bike almost equally

AI

Training data for machine learning implementation

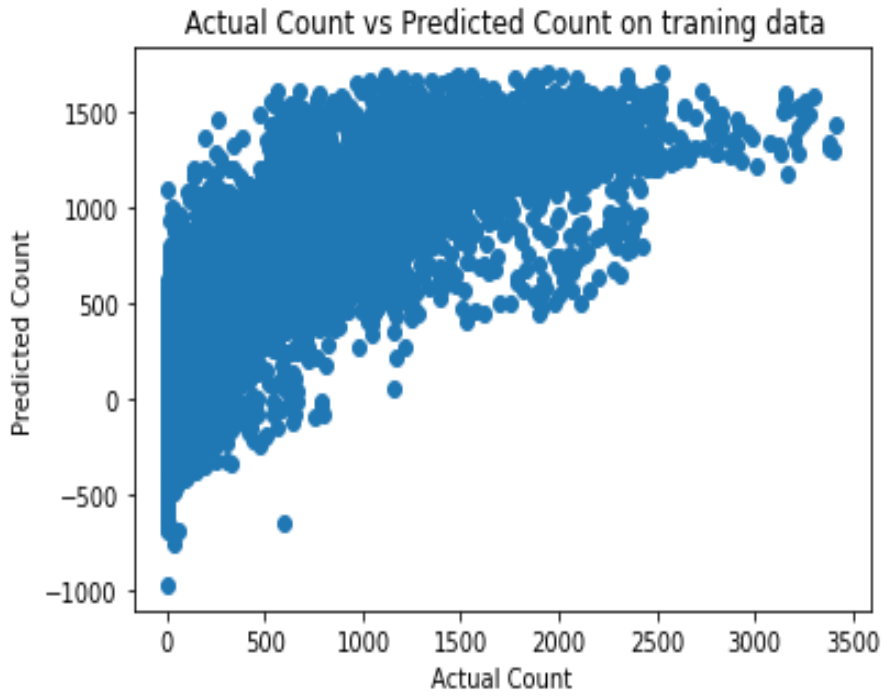
- Our dataset does not contain any null values
- Our dataset does not contain any duplicate values
- Changing all the categorical values to numeric type
- For categorical variables where no such ordinal relationship exists, there we will use one hot encoding
- Splitting Data into train and test
- Train Dataset: Used to fit the machine learning model.
- Test Dataset: Used to evaluate the fit machine learning model

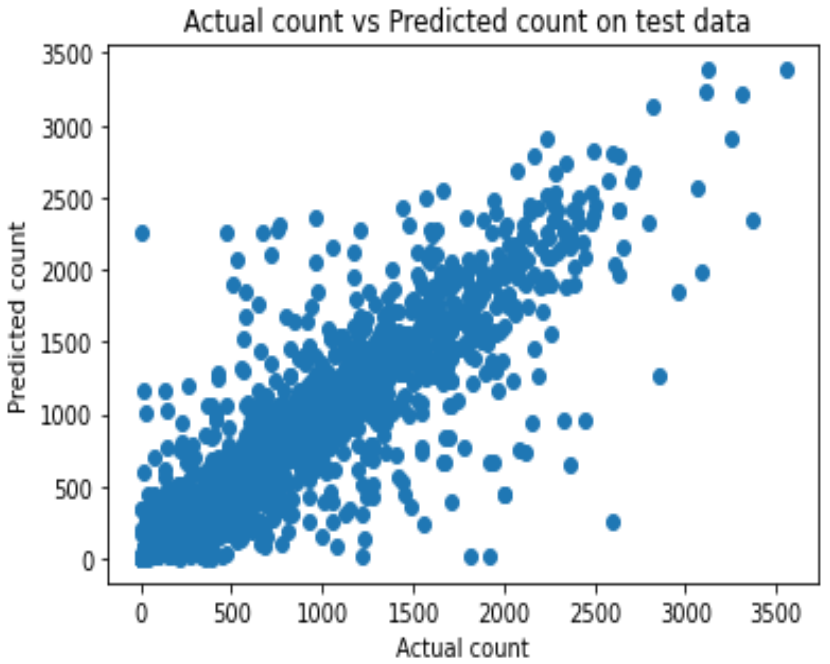
Model traning and prediction

- Linear Regression
- Lasso Regression
- Random Forest Regressor
- Decision Tree Regressor

Linear Regression

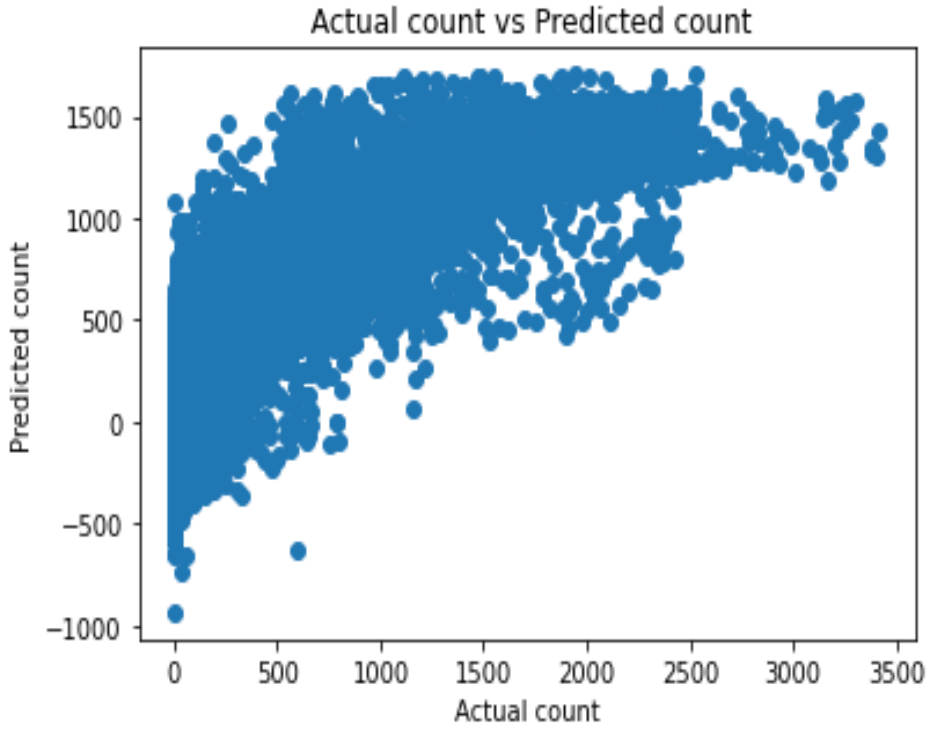
- R squared Error on training data id **0.55** and on test data is **0.53**

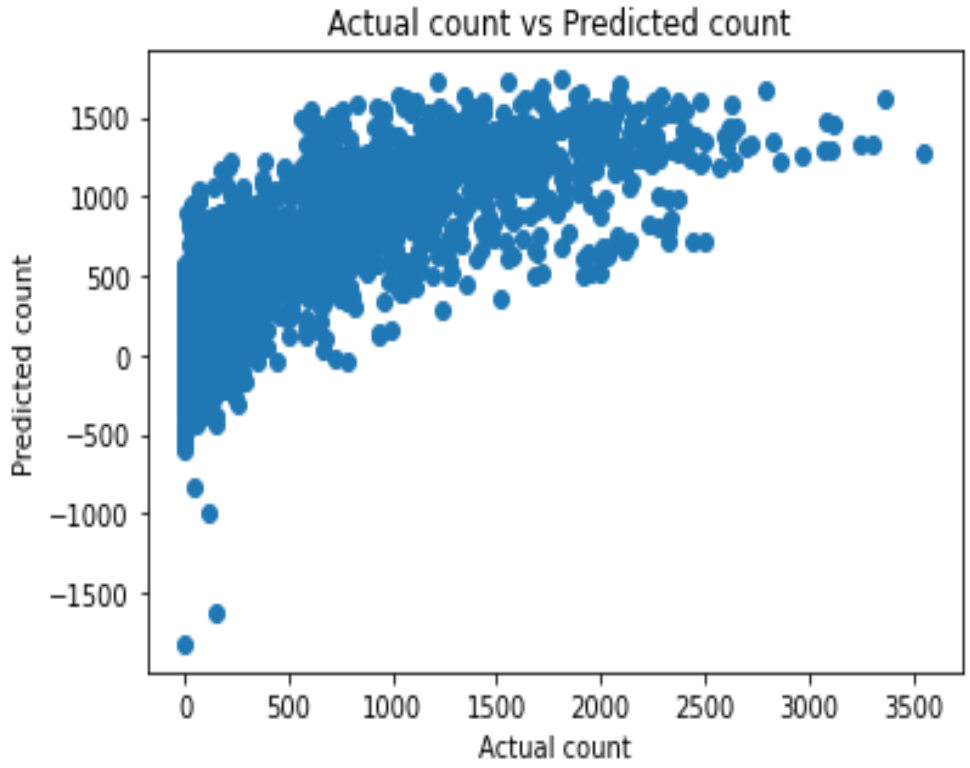




Lasso Regression

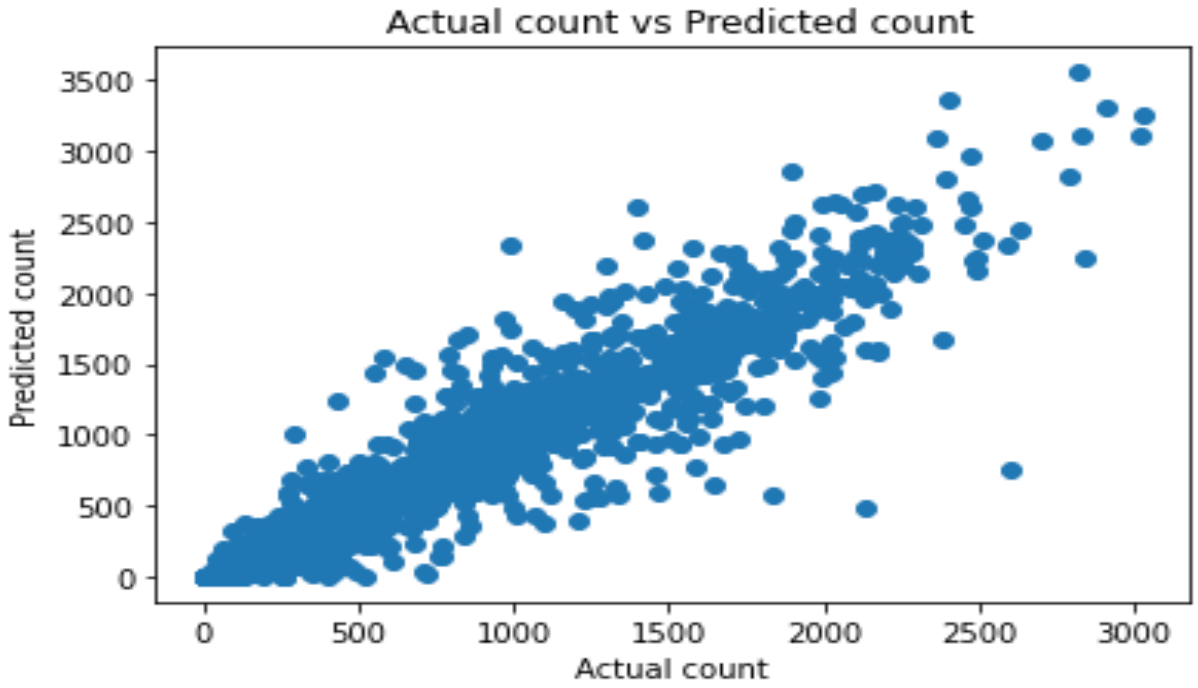
- Applied lasso regression on training as well as test data
- R squared Error on training data id **0.55** and on test data is **0.53** same as linear regression





Random Forest Regressor

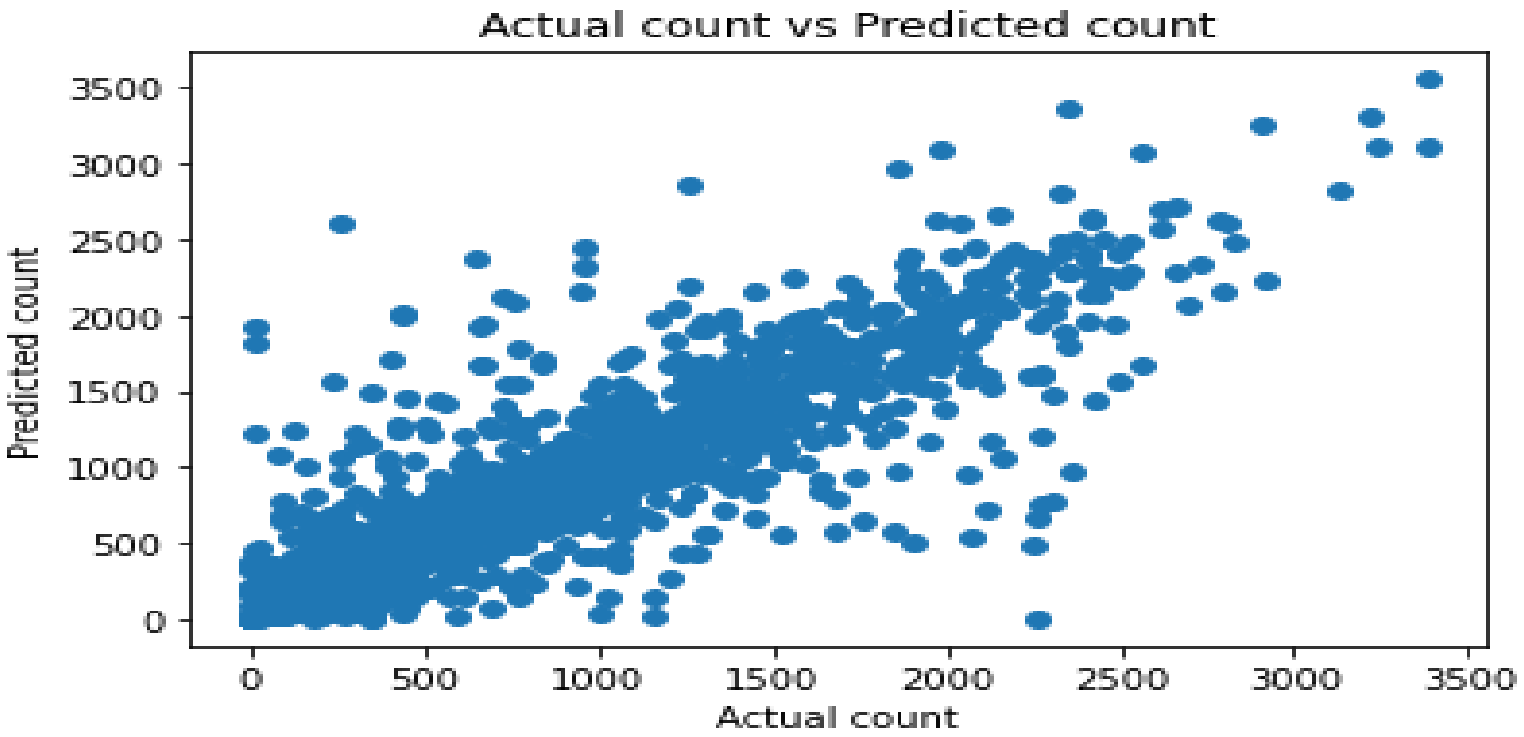
- **R squared Error of random forest regressor is- 0.89**



- **R squared Error of random forest regressor is- 0.89 which explains 89 percent of the variance of data so model is performing very good.**

Decision Trees (DTs)

- R squared Error on test data is- 0.79
- R squared Error of random forest regressor is- 0.79 which explains 79 percent of the variance of data so model is performing good.



Conclusion

- R squared Error of linear regression is-0.53
- R squared Error of lasso regression is-0.53
- **R squared Error of random forest regressor is- 0.89**
- R squared Error of Decision tree is-0.79
- **Random forest is performing very good as compared to another model**
- R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.
- **R squared Error of random forest regressor is- 0.89 which explains 89 percent of the variance of data so model is performing very good.**
- Root Mean Squared log error of random forest is 0.67 and Decision tree is 0.59


THANK YOU