# Capstone Project – 2
# Bike Sharing Demand Prediction

By – Shri Prakash Yadav

# Points of Discussion

- Problem Statement
- Data Summary
- Bivariate Analysis
- Univariate Analysis
- Correlation
- Model Creation
- Feature Importance
- Conclusion
- Challenges

# 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, pr oviding 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. In this capstone project I have tried to predict the rental bikes count required at any hour of the day.

# Data Summary

This data set consists of 8760 rows and 14 features which are listed below:

**Numeric:-**

- **Rented Bike Count** – Number of rented bikes.
- **Hour** – Hour of the day.
- **Temperature( Celsius )** – Temperature at each hour of the day.
- **Humidity( % )** – Humidity at each hour of the day.
- **Wind Speed ( m/s )** – Wind speed at each hour of the day.
- **Visibility ( 10m** ) – Visibility index.
- **Dew point temperature** – Dew point temperature at each hour of the day.

# Data Summary

- **Solar Radiation (MJ/m2)** – solar radiation index.
- **Rainfall(mm)** – Total rainfall in millimetres.
- **Snowfall (cm)** – Total snowfall in centimetres.

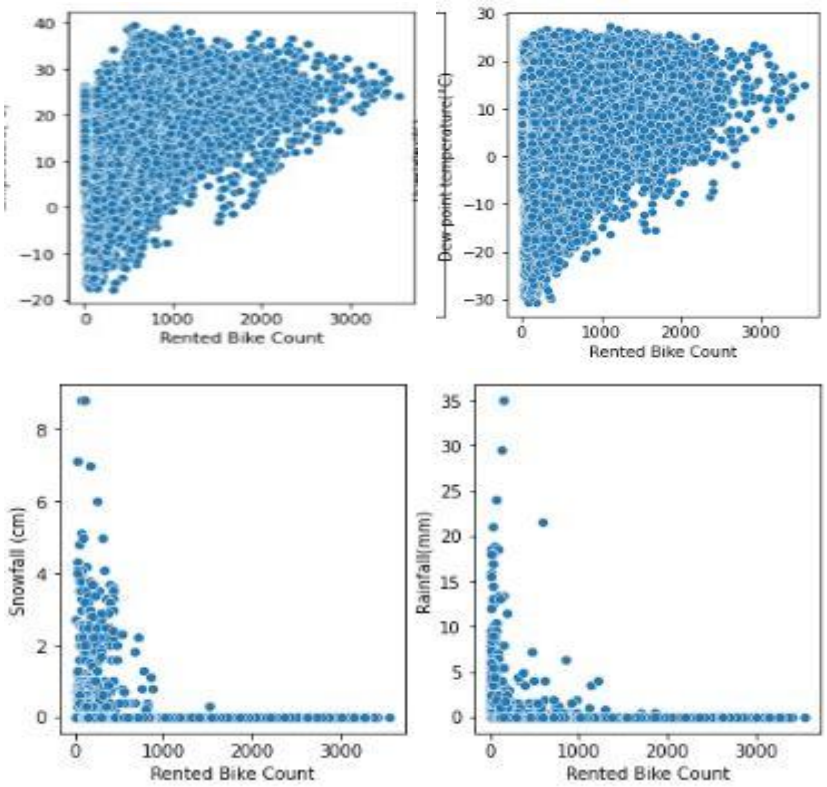
- **DateTime object:-**
- **Date** – date to which corresponding data point belongs.

- **String objects:-**
- **Seasons** – season to which corresponding data point belongs.
- **Functioning Day** – Whether the day was functioning or not
- **Holiday** – Whether the day was holiday or not.

# Temperature and Dew Point Temperature

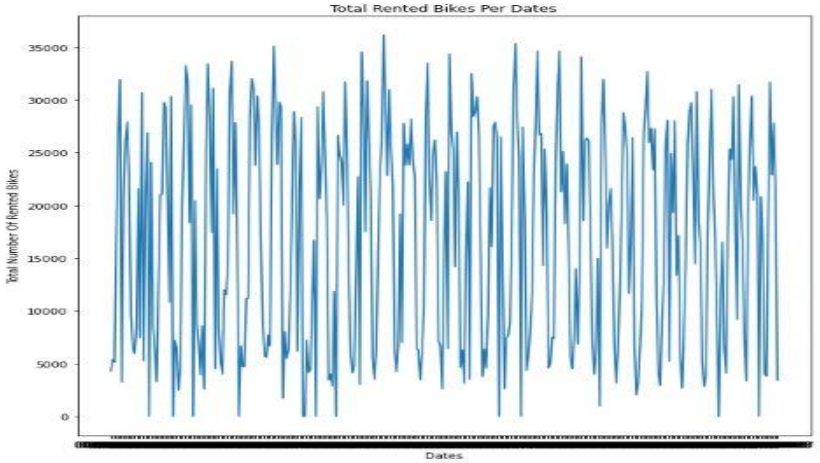
- As temperature and Dew point temperature increases Rented Bike count increases but the dependency is non linear.

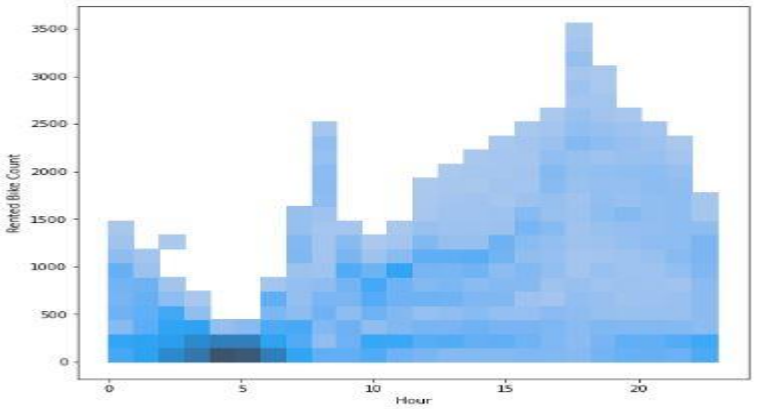
- As rainfall and snowfall increases rented bike count decreases but here too the dependency is non linear.



# Date and Hours

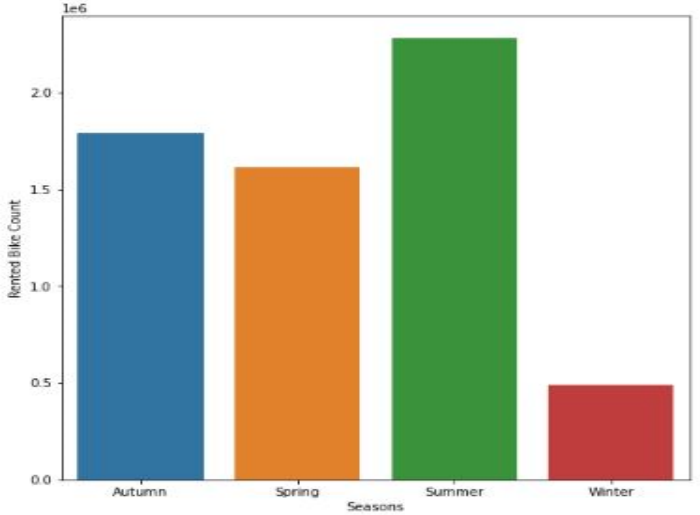
- Bike rental count is randomly distributed over dates.
- At 8th , 18th and 19th hour rented bike count is maximum.
- From 10th to 18th hour bike rental count is continuously increasing and then gradually decreases from 18th to 23rd hour.

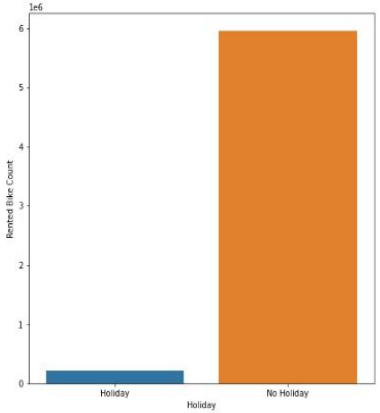


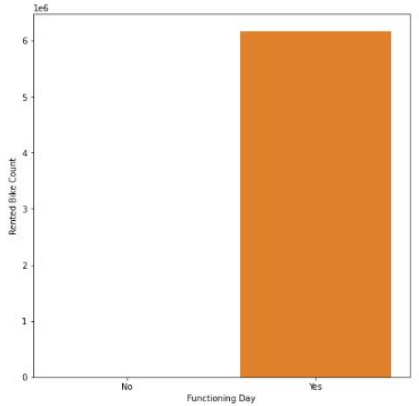



# Seasons, Functioning days and Holidays

- Approx 37 percent of total rental counts have taken place in summer.
- Minimum accros 8 percent of rentals have took place in winter
- Approx 97 percent of total rentals have took place on non-holidays.
- All of the rentals have took place on functioning days.

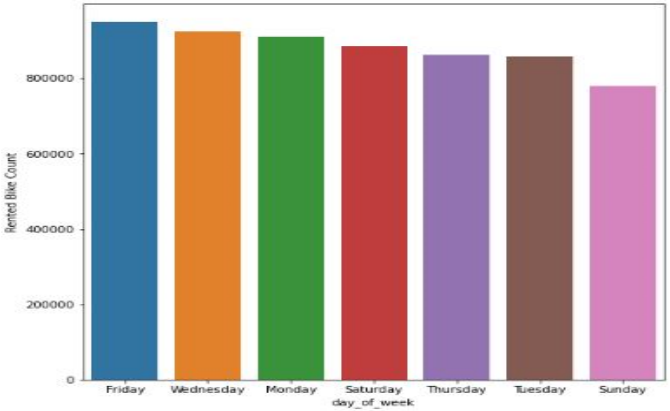


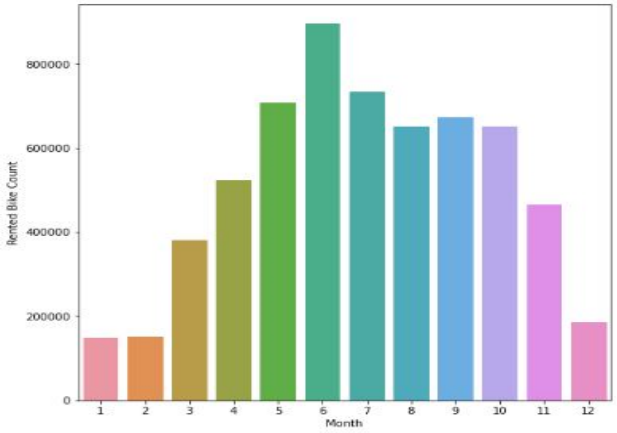



# Weekdays and Months

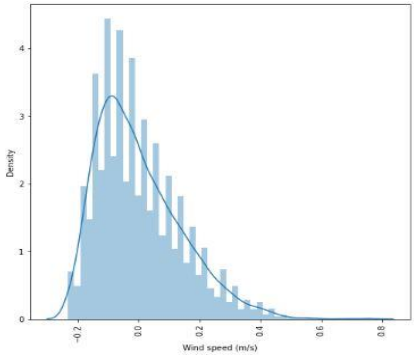
- There is not much difference between rented bike counts on different weekdays.
- In the month of june rented bike count is maximum.
- In the month of january and february rented bike count is minimum.

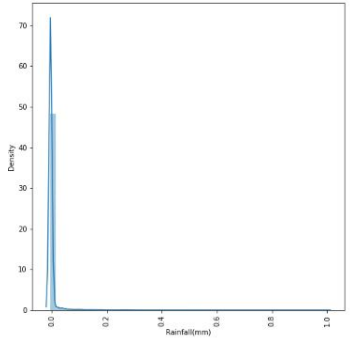


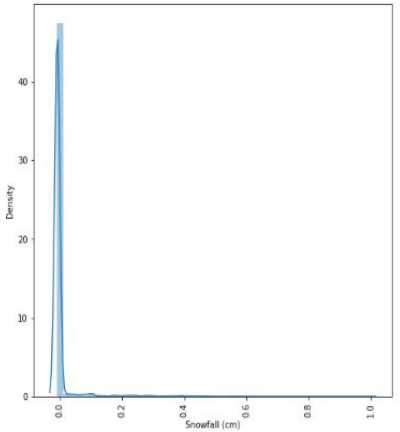



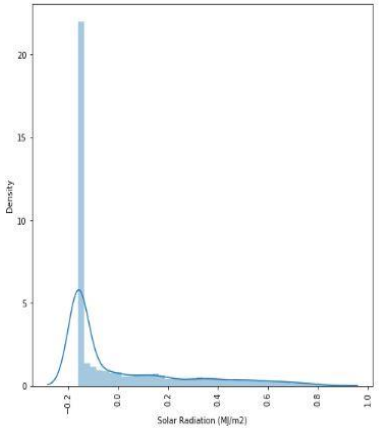
# Numerical Features

- From the pdf distribution we could say that all the features wind speed, solar radiation, snowfall and rainfall are positively skewed.
- These features have huge difference between mean values and max values.
- Snowfall , rainfall, solar radiation are mean. centered whereas wind speed has more spread.



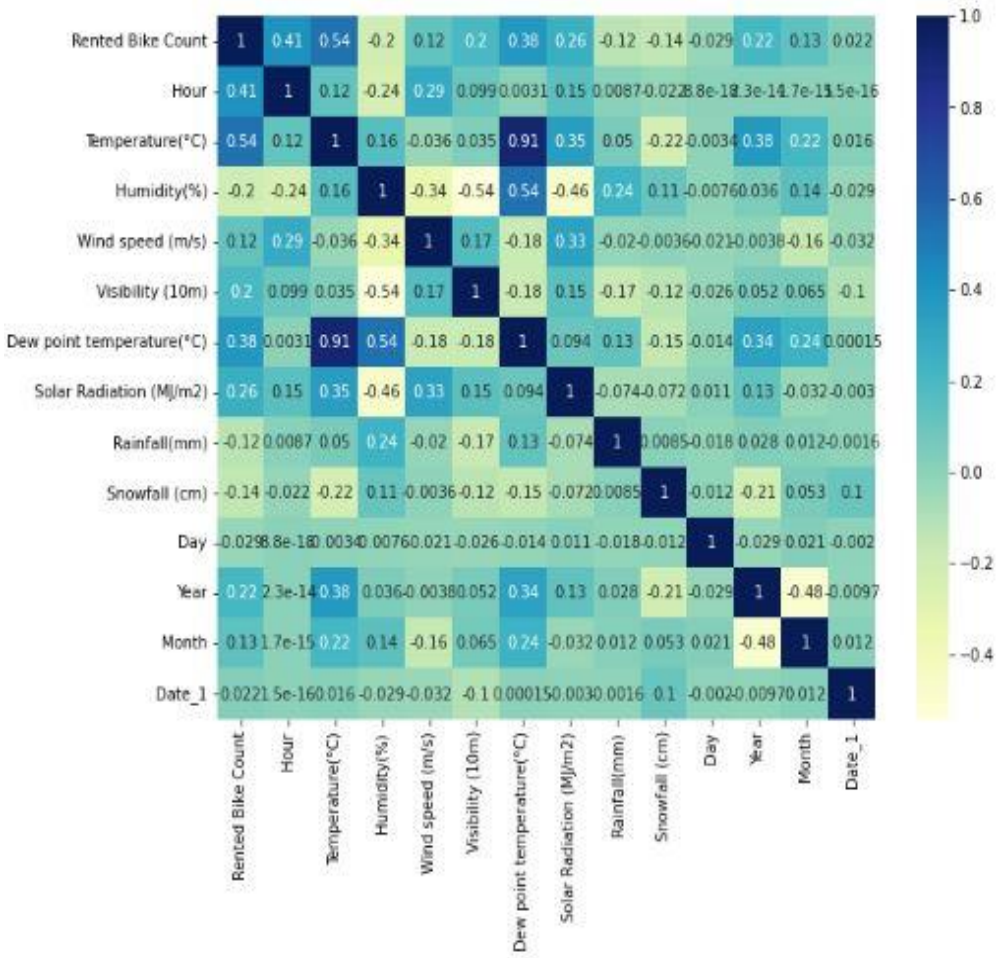



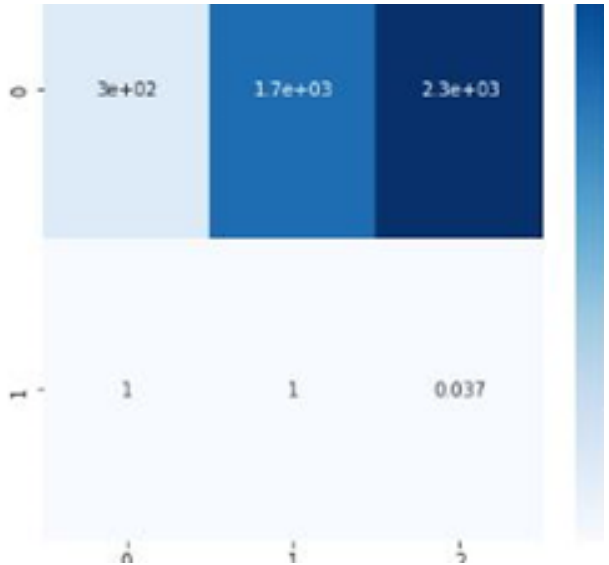
# Correlation

- Bike counts increases as Temperature, dew point temperature and hour increases.
- Bike count is also proportional to visibility, wind speed and solar radiation but very weak proportionality.
- Bike count decreases as snowfall, rainfall and humidity increases.



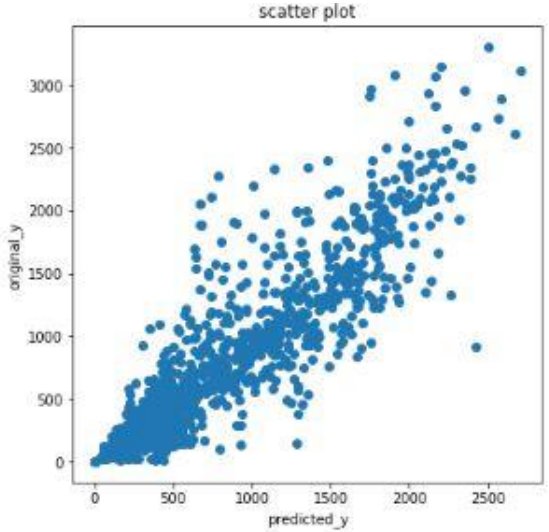
# Correlation

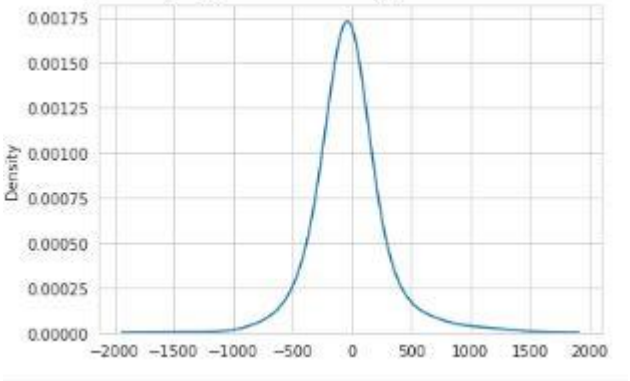
- For seasons and rented bike count p value is 0.037.
- For holiday and functioning day p value is 1.



# KNN Regressor

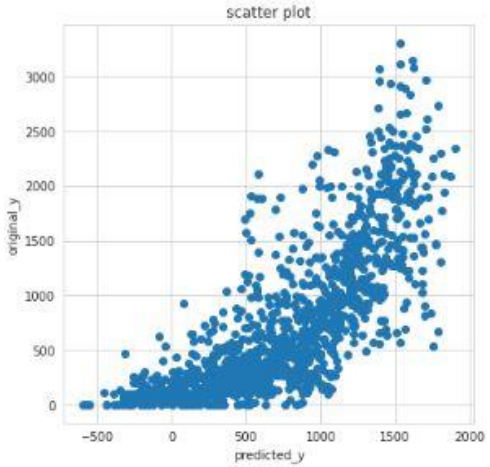
- KNN regressor gives best result when n_neighbour value is 11 and using weighted distance.
- KNN regressor gave mean_sq_error of 284.
- R_sq value for KNN regressor was 0.82.
- Mean_sq_log_error was 0.266

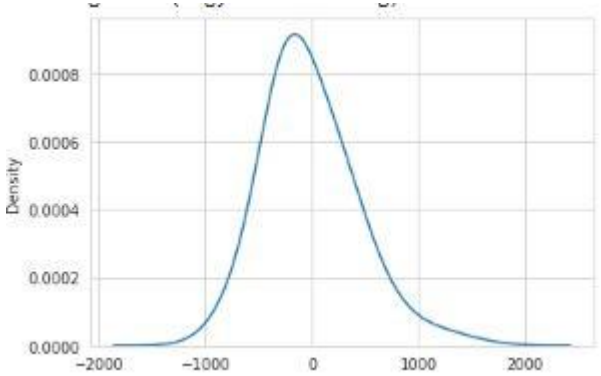


# Linear Regression

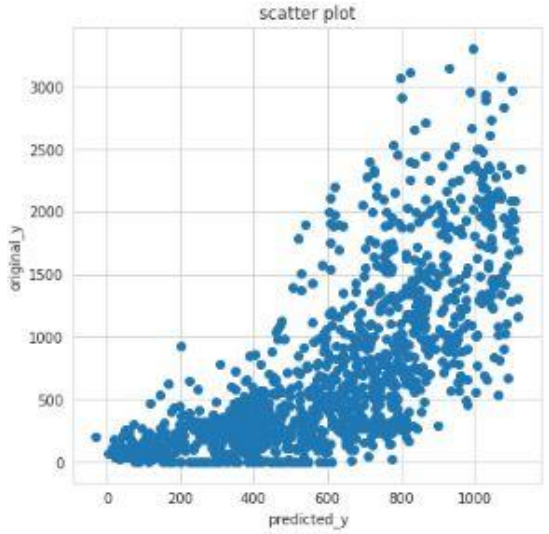
- Linear regression is working bad than KNN regressor.
- Mean_sq_error for linear regression model was 431.
- R_sq value was approx 0.6
- The variance of error distribution was higher than KNN regressor.

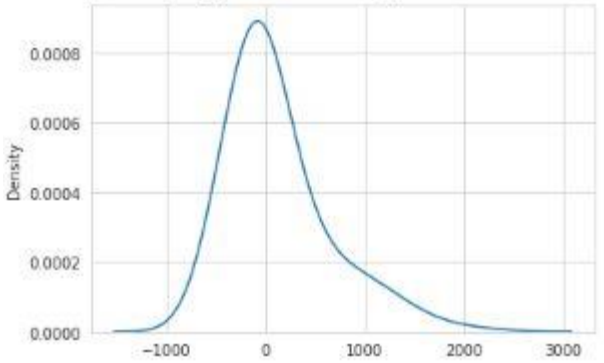


# Support Vector Regressor

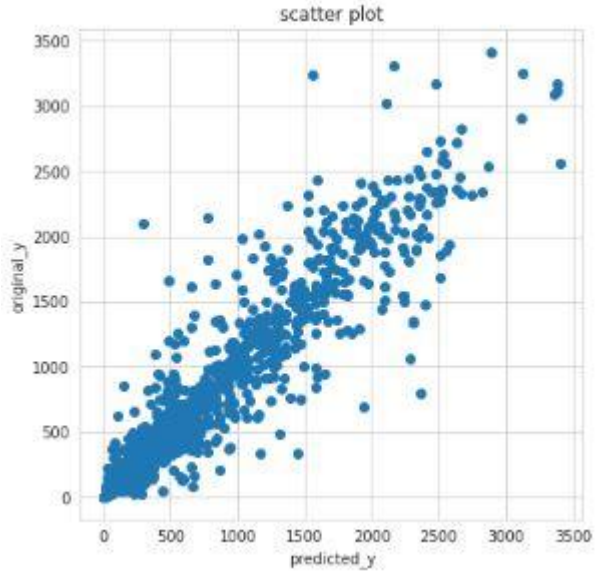
- For support vector regressor the error values are even higher than linear regression model.
- Mean_sq_error for svr was 529.
- R_sq value for svr was 0.51 slightly better than mean models.
- There are many points for whom error value is even greater than 2000

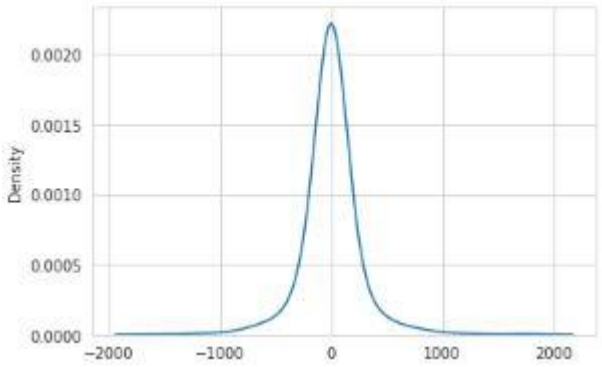


# Decision Tree Regressor

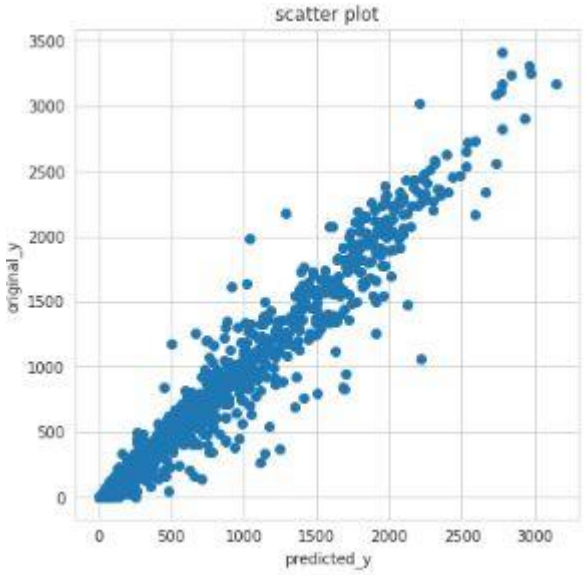
- DT regressor model is working better than all distance minimisation models.
- Mean_sq_error for DT regressor was 244.
- R_sq value for DT regressor was 0.87
- The variance is low and errors are mostly mean centered.

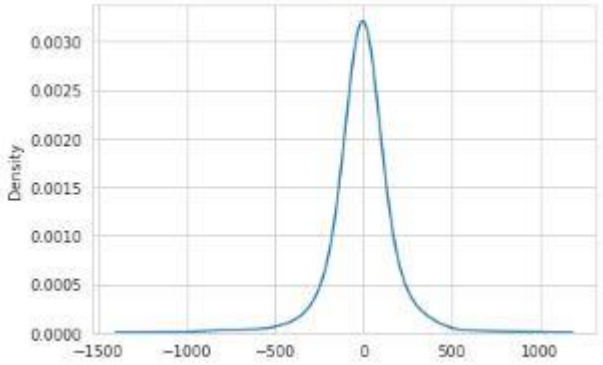


# Random Forest Regressor

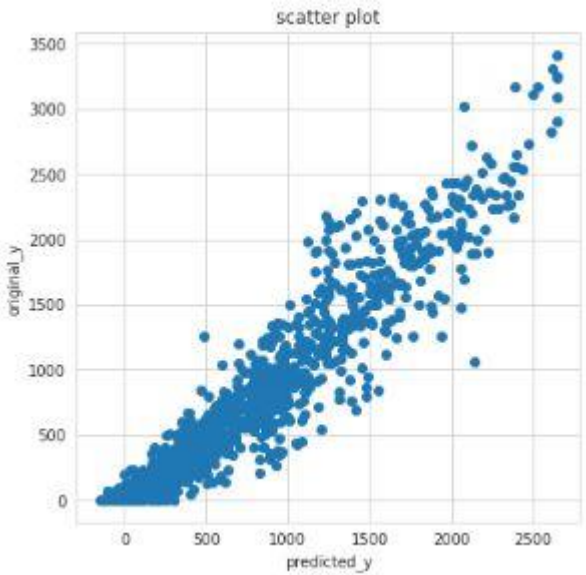
- We got thin and straight distribution.
- Mean_sq_error for Random Forest regressor was 162.
- R_sq value for RF regressor was 0.94.
- The variance of error is lower than DT Regressor model.

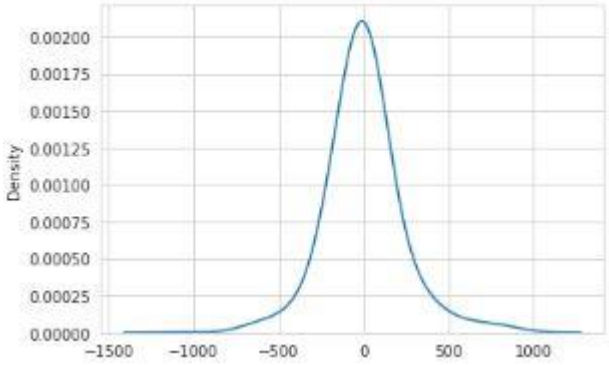


# Gradient Boosting Regressor

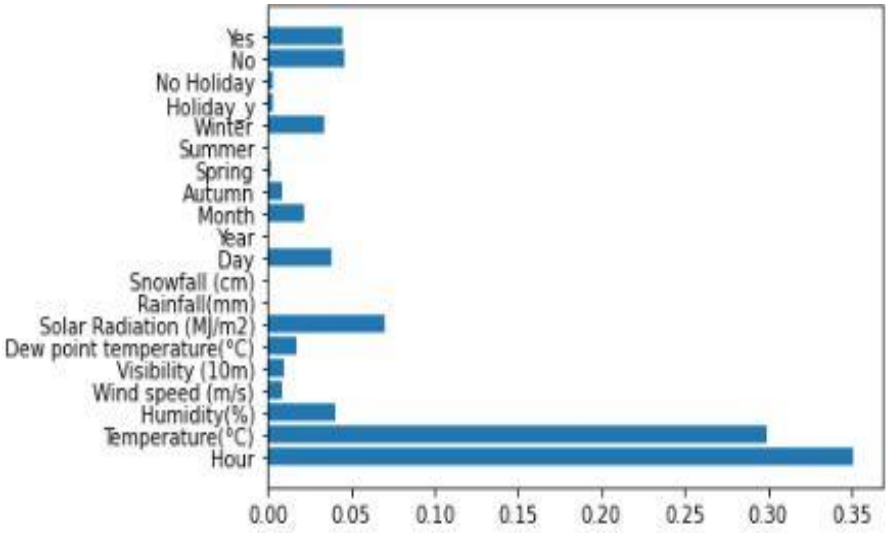
- The distribution is straight but thick.
- Mean_sq_error for GB regressor was 218 a little more than RF regressor.
- R_sq coefficient for GB regressor was 0.90.
- Variance of the error is higher than RF regressor.



# Feature Importance

- The most important features in determining rented bike counts are hour and temperature.
- The least important features are season, holiday, snowfall and rainfall.



# Conclusion

- Bike sharing demand dataset was very neat and clean. The rented bike counts was little or more correlated with all the features. The most important features in determining the rented bike counts were hour and temperature and the best prediction model which I could train was Random Forest Regressor whose mean_sq_error was 160 and r_sq coefficient was 0.94.

# Challenges

- As I said the dataset was very clean so performing data analysis and model creation was pretty much straight. The dataset also didn’t had any missing values. So, I didn’t faced any problem in this capstone project.

# THANK YOU