Capstone Project - 2

Bike Sharing Demand Prediction

Submitted by

Aadarsh Pandey

Ankita Hanamshet

Darpan Agrawal

Vandana Pattnaik

Vinay Kulkarni

Data Science Trainees, Almabetter

Content

- Introduction
- Data Summary
- Data Description
- Preprocessing of the data
- Exploratory Data Analysis (EDA)
- Correlation Matrix
- Models
- Challenges faced and Conclusions

Introduction

- 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.
- The main goal of project is to maximize the availability of bikes to the customer and minimize the time of waiting to get a bike on rent.

Data Summary

- ➢The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.
- ➢This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Capital bike share system with the corresponding weather and seasonal information.
- ➢The dataset contains 8760 rows (every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration).
- ➢ One Datetime features 'Date'.
- ➢Three categorical features 'Seasons', 'Holiday', & 'Functioning Day'.

Data Description

Dependent variable:

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

Independent variables:

·Date : year-month-day
·Hour - Hour of he day
·Temperature-Temperature in Celsius
·Humidity - %
·Wind speed - 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 - NoFunc(Non Functional Hours), Fun(Functional hours)

Preprocess Data

- Records=8760 and Features=14
- No data is Missing in given dataset.
- There are No Duplicate values present
- There are No null values.
- And finally we have 'rented bike count' variable which we need to predict for new observations.
- 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'

Feature Engineering

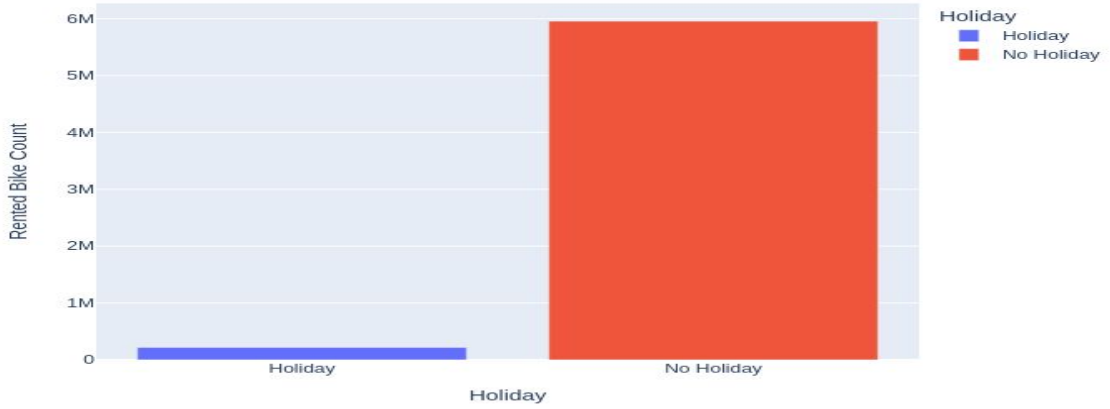
- In addition to existing independent variables, we will create new variables to improve the prediction power of model.

- Here, “Date” column can be splitted into multiple independent variables such as date, month and year.

- This will help to classify the bike rented rate day wise, month wise and year wise.

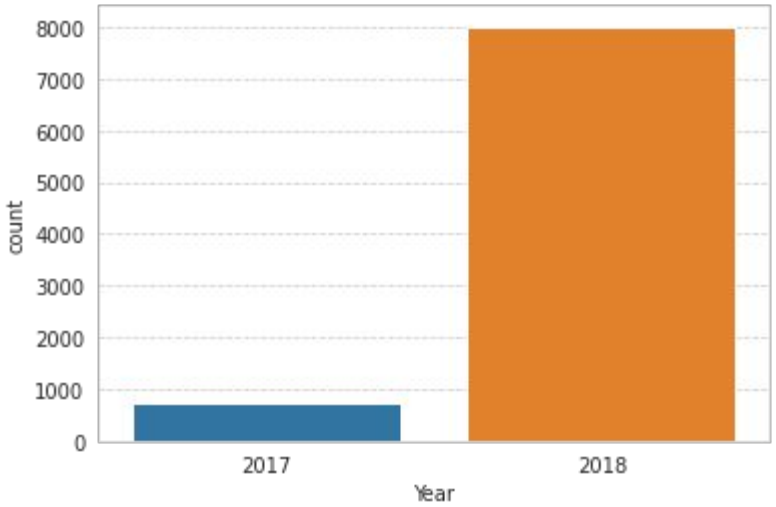
Exploratory Data Analysis (EDA)

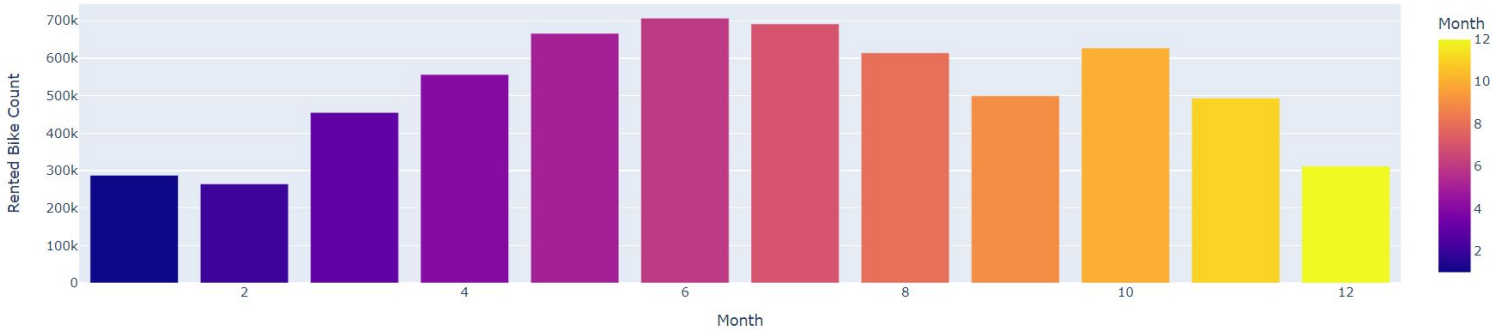
- This **Bar Plot** indicated difference between **Rented Bike Count** on **Holiday** and on **No Holiday.**
- According to the analysis of the below given bar plot we get that People have used most number of bike on **No Holiday** compare to **Holiday.**



EDA (Contd.)

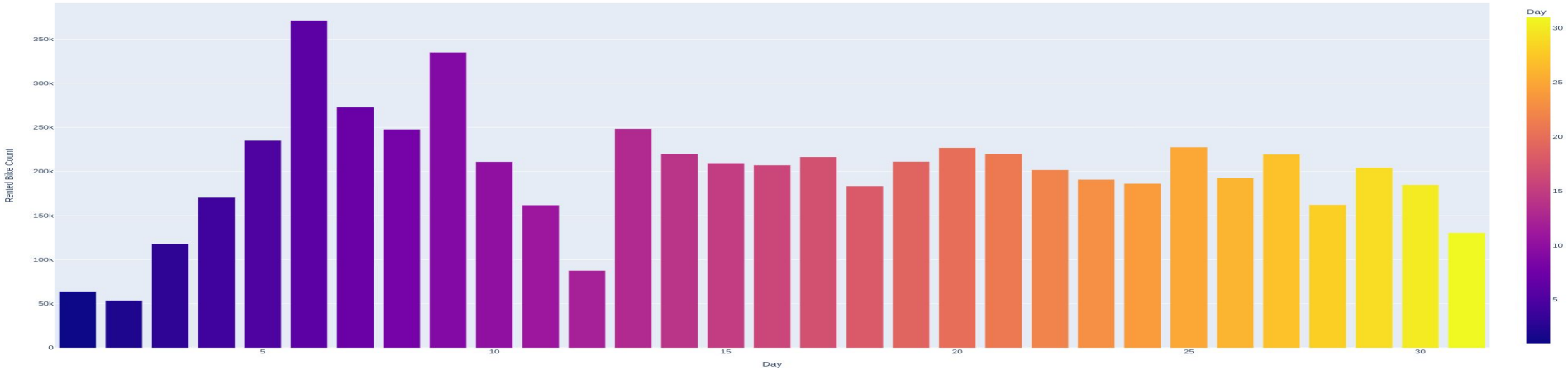
- Highest demand - **June**
- Lowest demand - **January**
- 2018 has the highest demand if we compare it with 2017





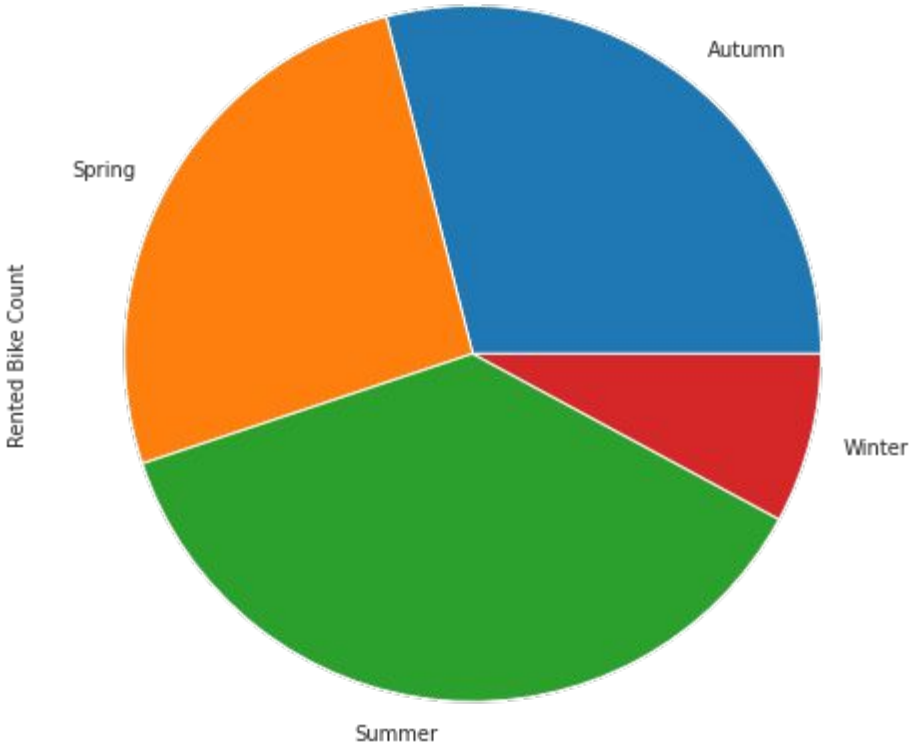
EDA (Contd.)

- According to this day analysis of rented bike count we get that, between day 5 to day 10 rented bike count is at highest demand in the whole month.



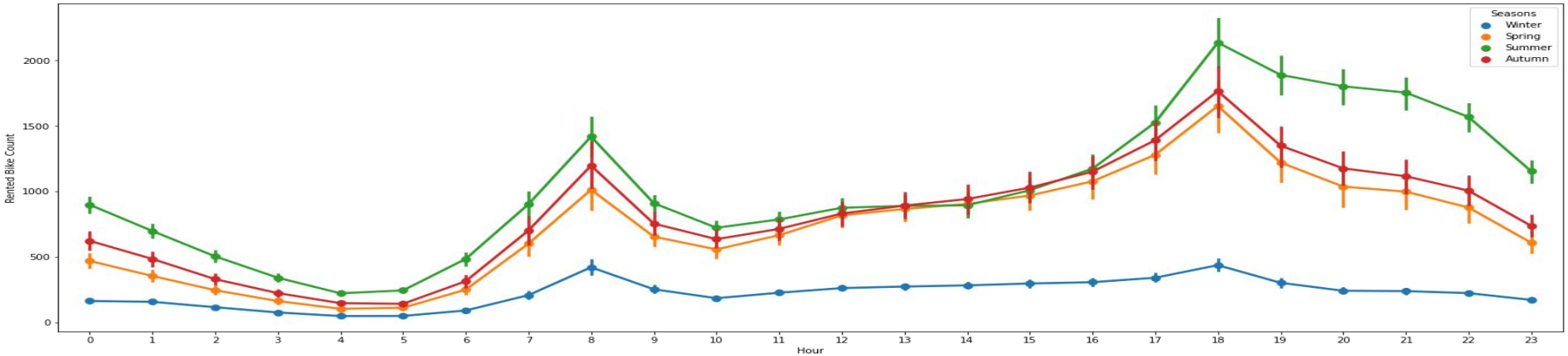
EDA (Contd.)

- Lowest demand - **Winter**
- Highest demand - **Summer**
- In autumn and spring, the demand on average is similar throughout the day



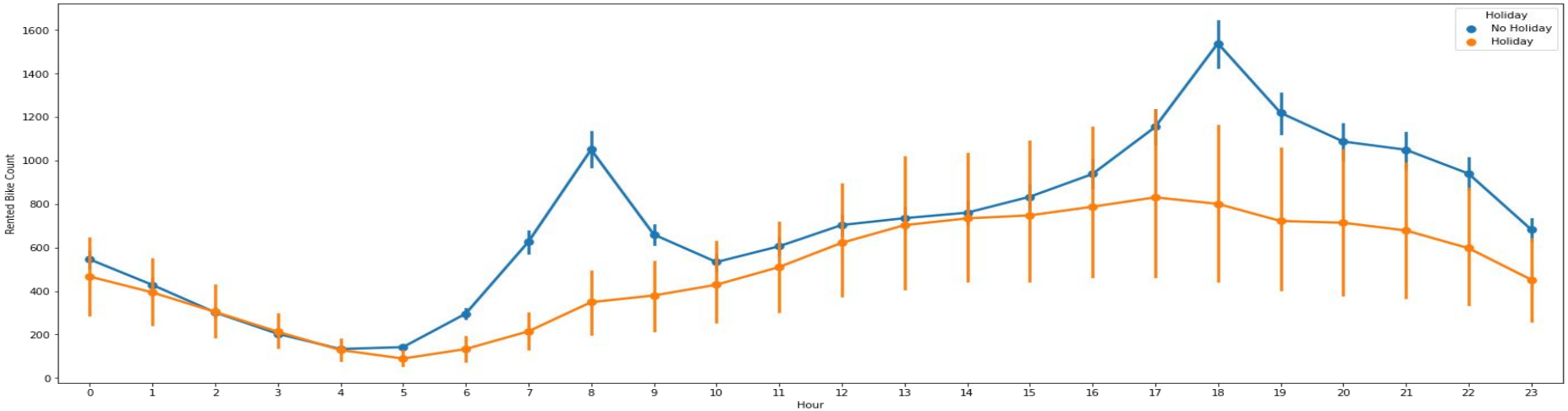
EDA (Contd.)

- This line plot describes about in which season and in which hour the demand of the bike is high and according to is we have summer which has the highest demand for the bikes.



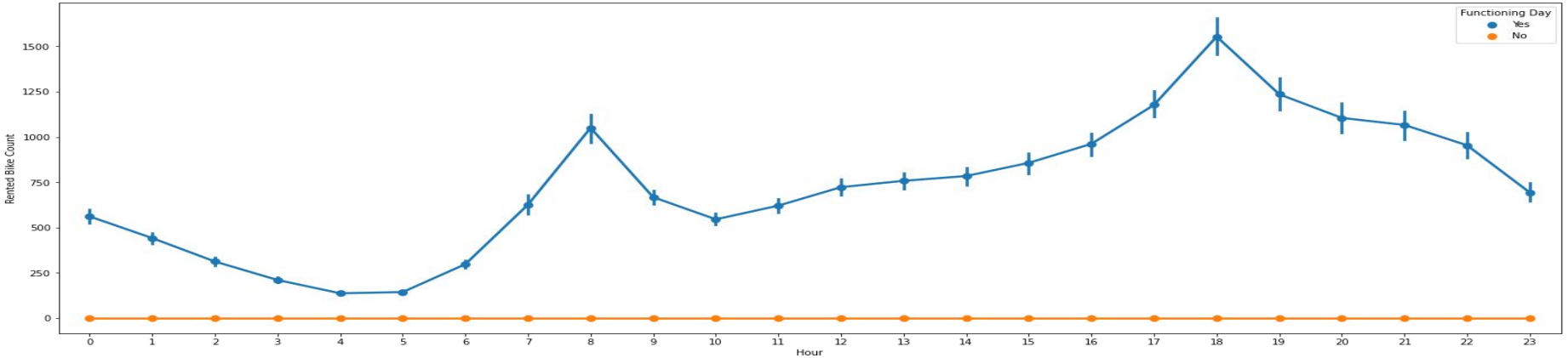
EDA (Contd.)

- This Line Plot tells us about in which hour the bike demand is at highest on NO Holidays and on Holidays.



EDA (Contd.)

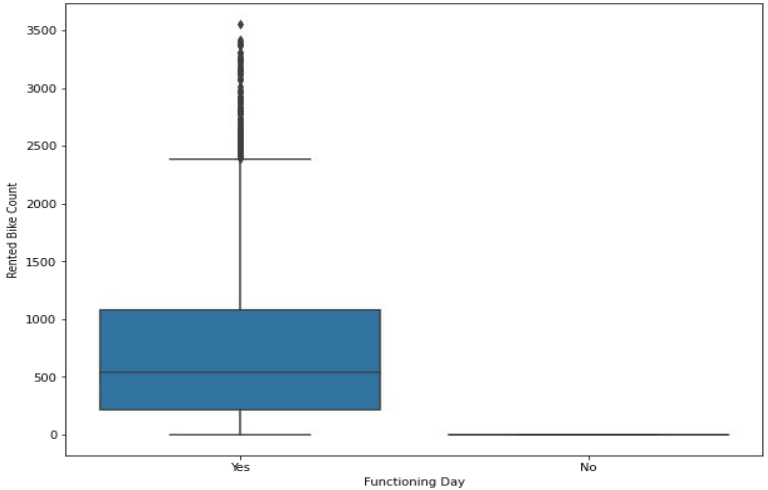
- No Function day has no demand for the bikes.
- Whereas Function day has the highest demand at 18:00 PM

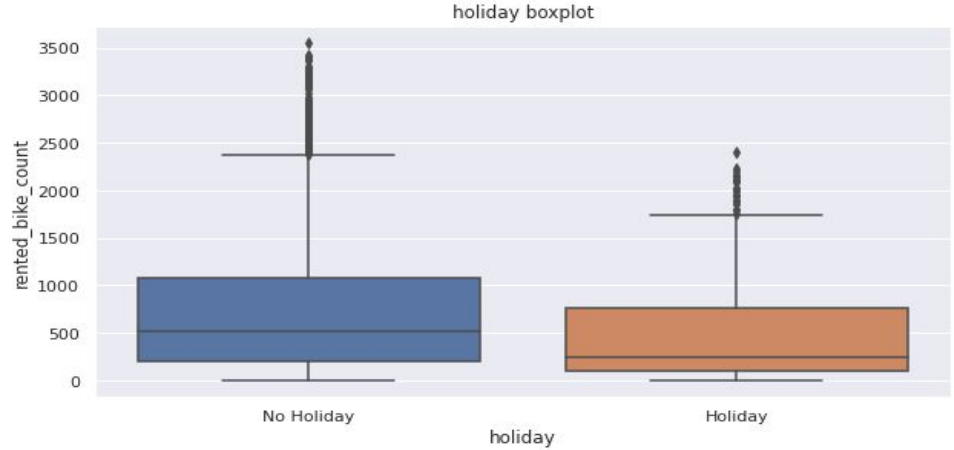


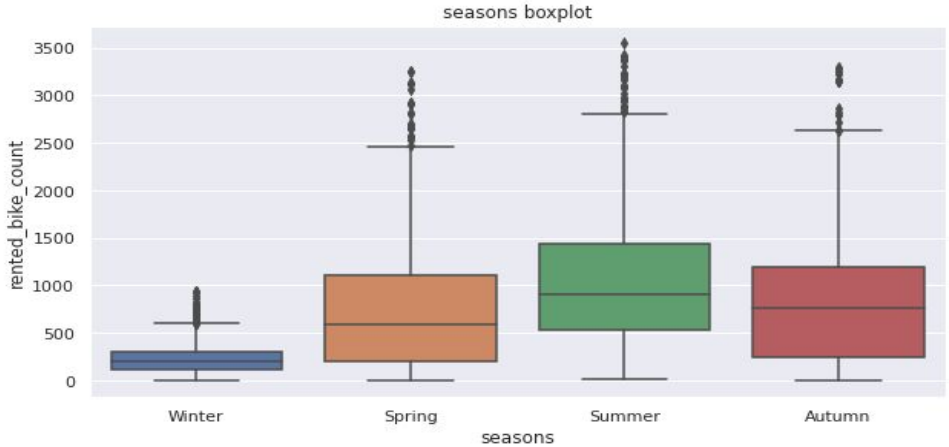
EDA (Contd.)

Function day has highest number of outliers and No Holiday has the highest outlier compare to Holiday boxplot

Whereas summer has the highest outlier among all the seasons

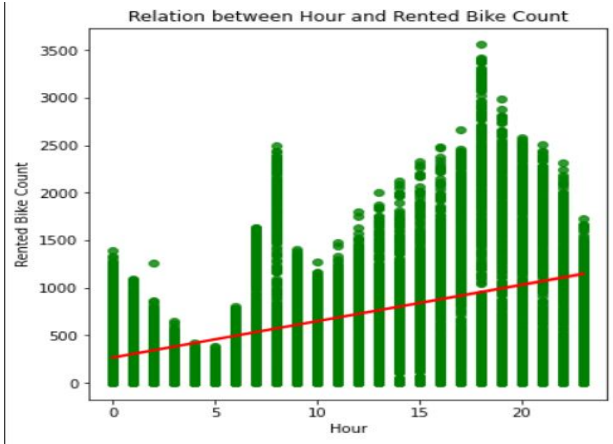


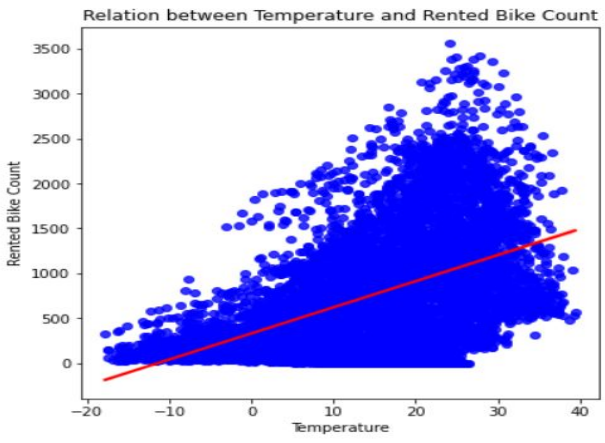


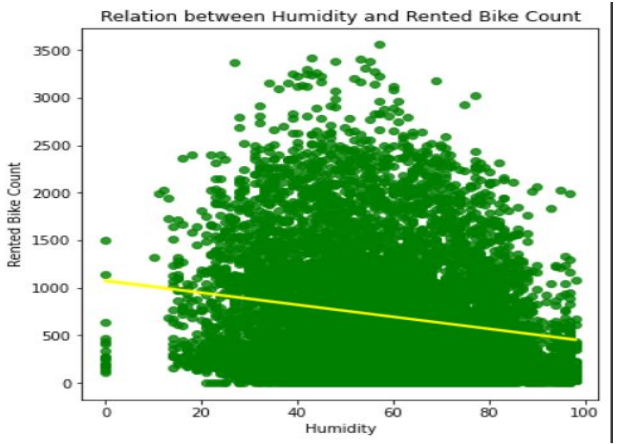


EDA (Contd.)

This is Reg plot for checking the colinearity of all the features and according to that we see Hour and Temperature are positively correlated whereas Humidity is negatively correlated to Rented Bike Counts.

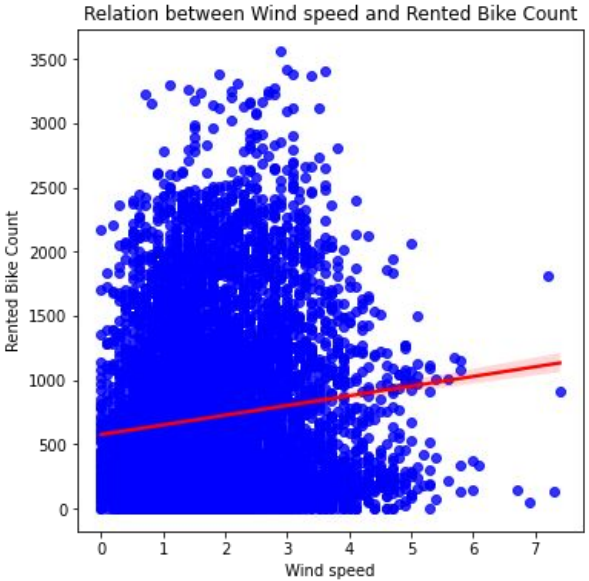


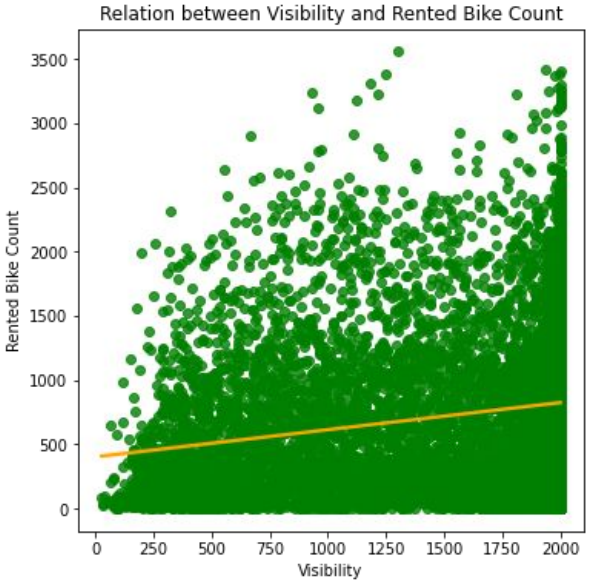


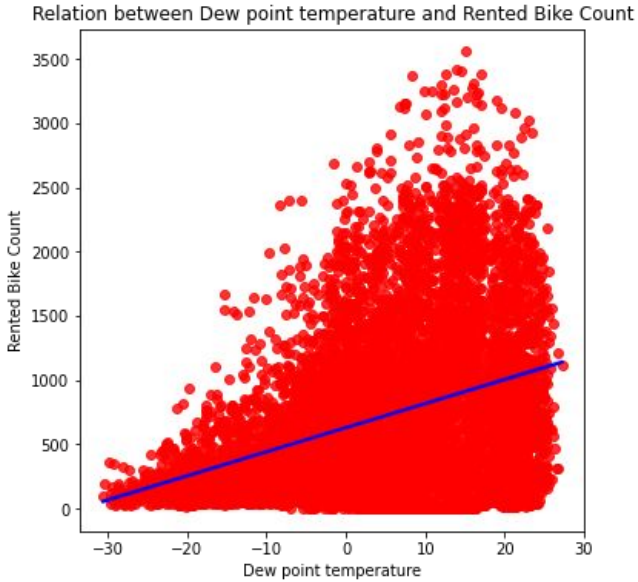


EDA (Contd.)

Here we can clearly see that Wind speed visibility and Dew point Temperature is positively correlated to Rented Bike Counts.

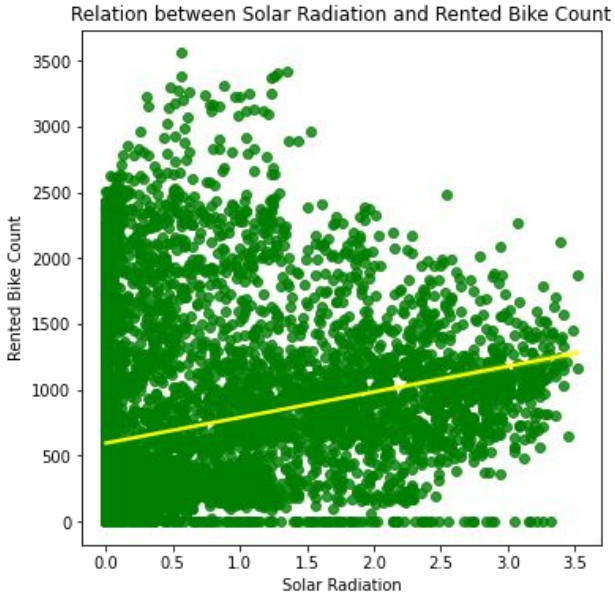


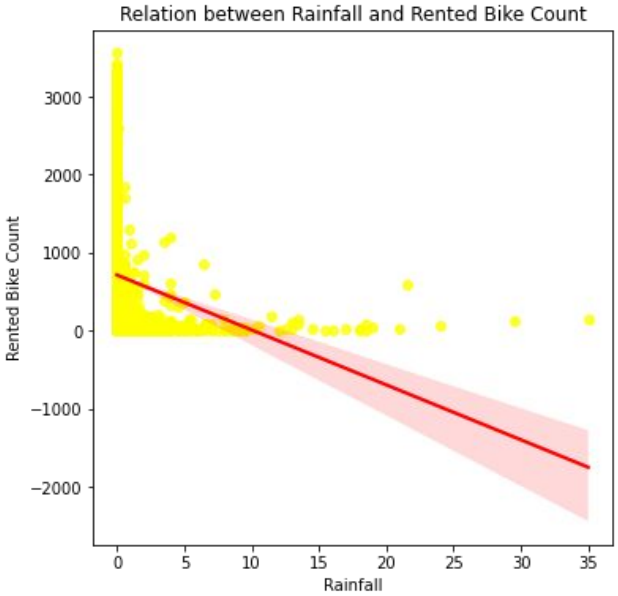


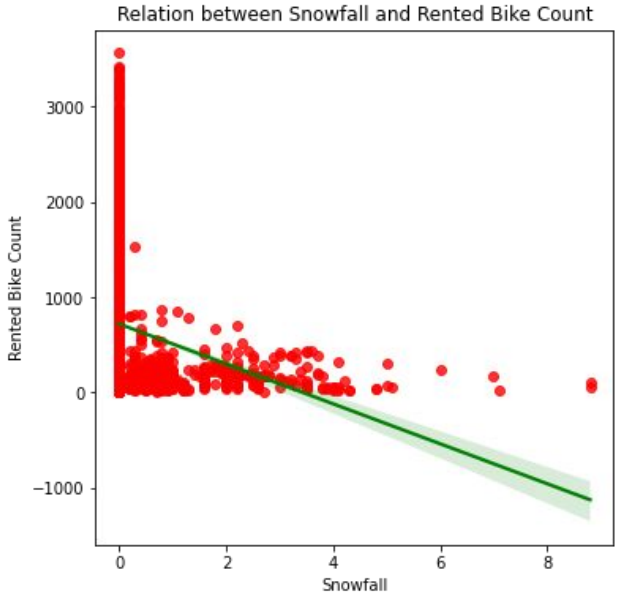


EDA (Contd.)

Solar Radiation, Rainfall and Snowfall all are highly Negatively Corelated to Rented Bike Counts.

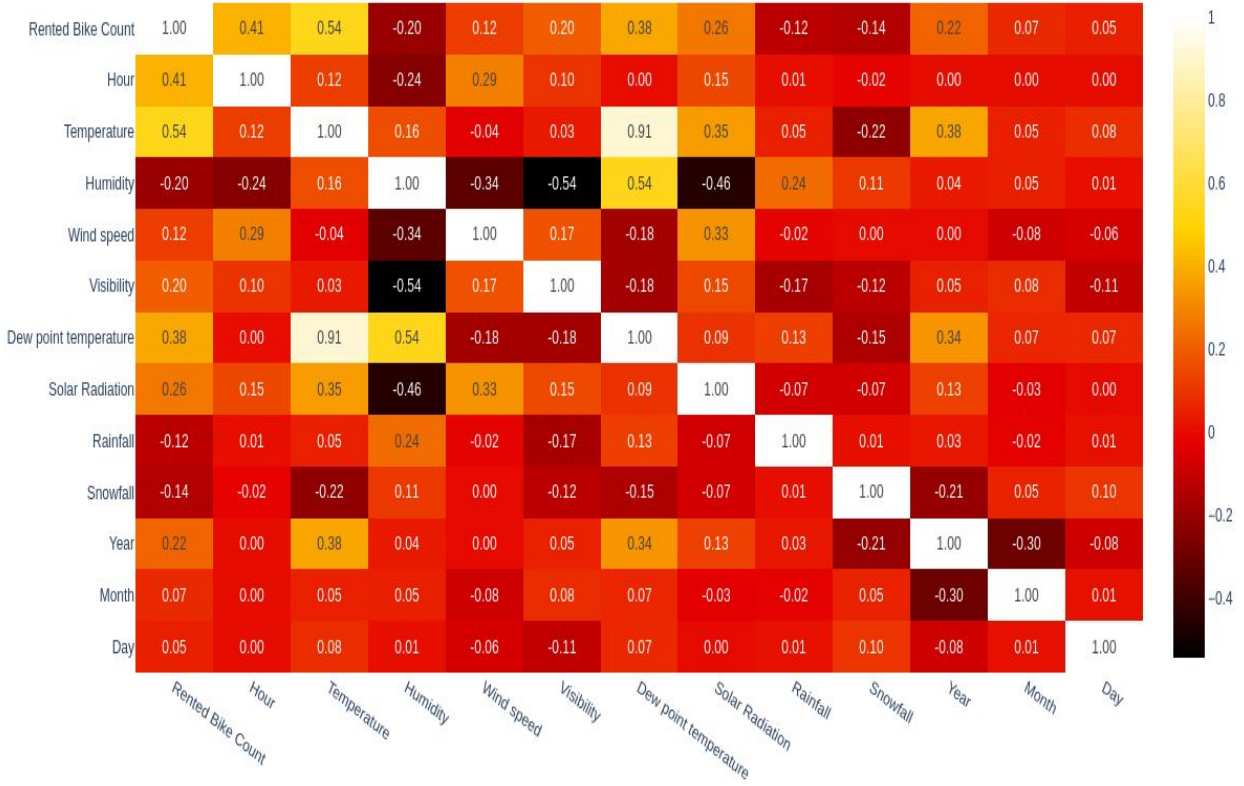






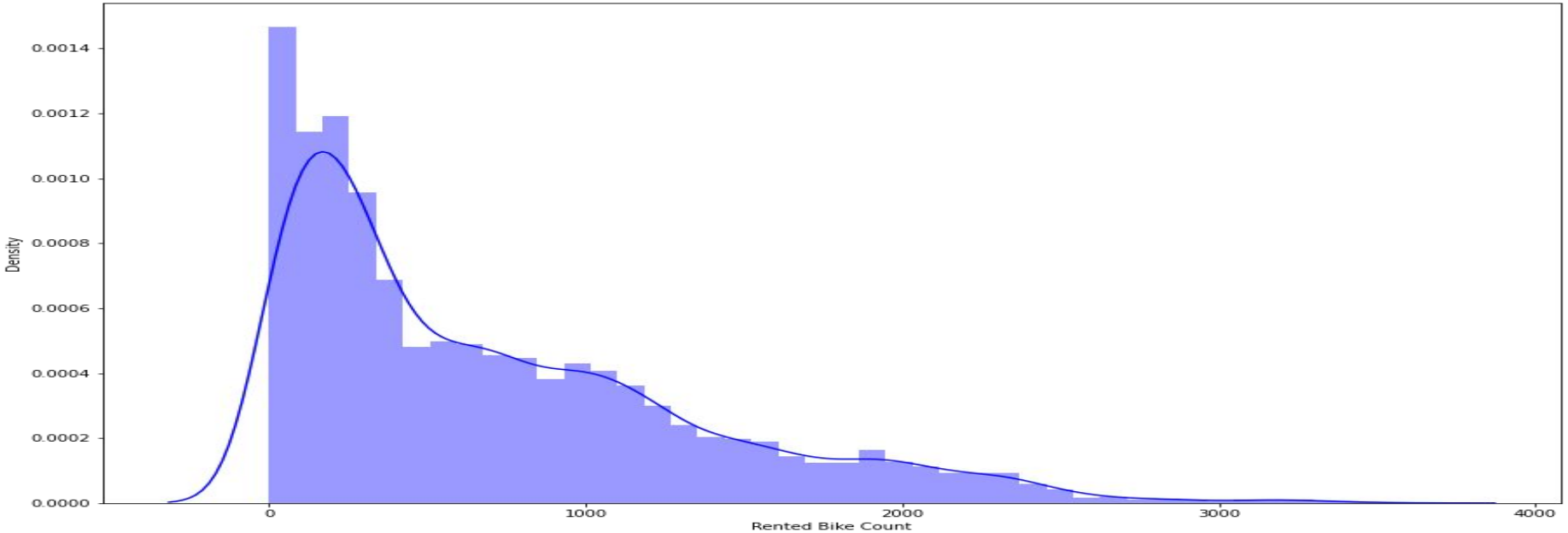
EDA (Contd.)

- Correlation magnitude
- There is no **multicollinearity** in the attributes
- Temperature has the **highest** correlation with the dew point temperature



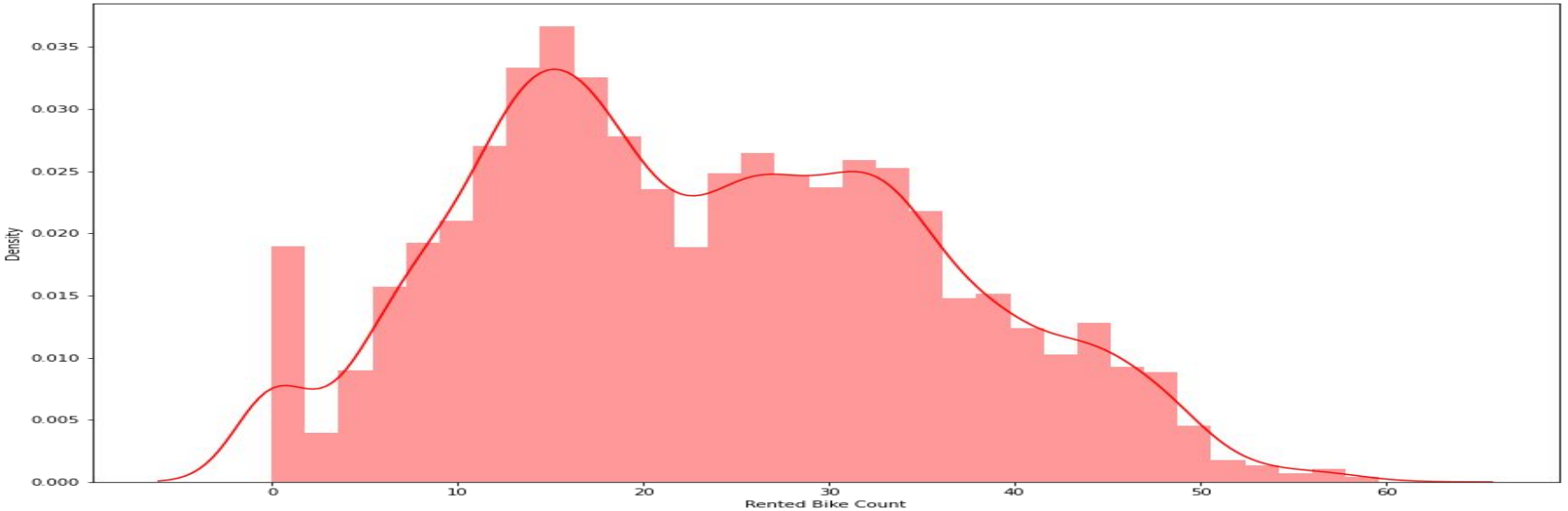
EDA (Contd.)

Here this graph is plotted for checking the skewness of the dependent variable and we can clearly see that Rented Bike Count is Rightly Skewed.



EDA (Contd.)

After normalizing the data we see that Rented Bike Count is symmetrically skewed.



EDA Summary

- The dependent variable - rented bike counts is **positively skewed**
- Demand for rental bikes is lowest in the winter and highest in summer.
- On regular days, there is a **surge** in demand for rental bikes during **rush** hours, this was absent during **holidays** and **No Function Day.**
- June is the month which has the highest demand for the bike.
- On the analysis of the daily basis we can see that 18 pm bike has highest demand.
- During summer Bike demand is highest compare to all other seasons.

Modelling Approach

- Since the data contains no **outliers**, and many **categorical** attributes. It would be wise to model in both linear models as well as tree models
- In our project we will be using Linear regression,Decision Tree Regressor, XG Boost, Gradient Boost Regressor, Random Forest, Lasso Regression, Ridge Regression and Elastic Net Regressor
- Final choice of model will depend on whether interpretability or accuracy is important to the stakeholders.

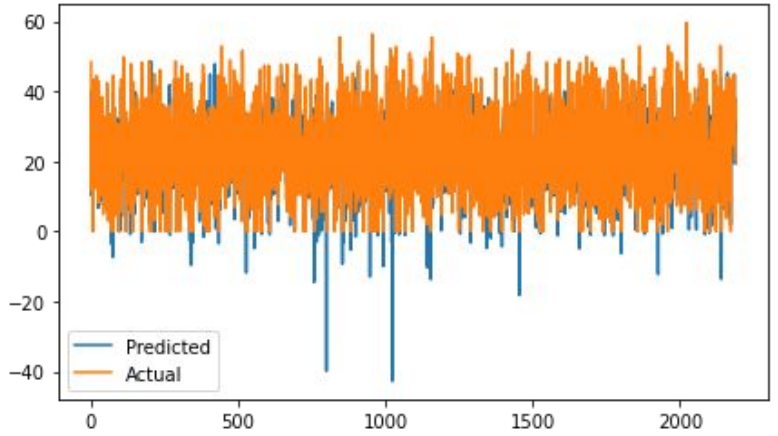
Modelling Approach (Contd.)

- Choice of split is taken as **test train split**, because of the computational power available and to reduce overfitting
- Evaluation metrics is **MSE,RMSE,R2 and Adjusted R2** to punish outliers, and choose a model that is able to generalize the results for all points including outliers.
- Hyperparameter tuning is done to prevent overfitting, and the best parameters are chosen using **GridsearchCV**

Linear Regression

Evaluation metrics Test:

- MSE : 37.13314681383164
- RMSE : 6.093697302445506
- R2 : 0.762933419958126 1
- Adjusted R2 : 0.757618522320568 9



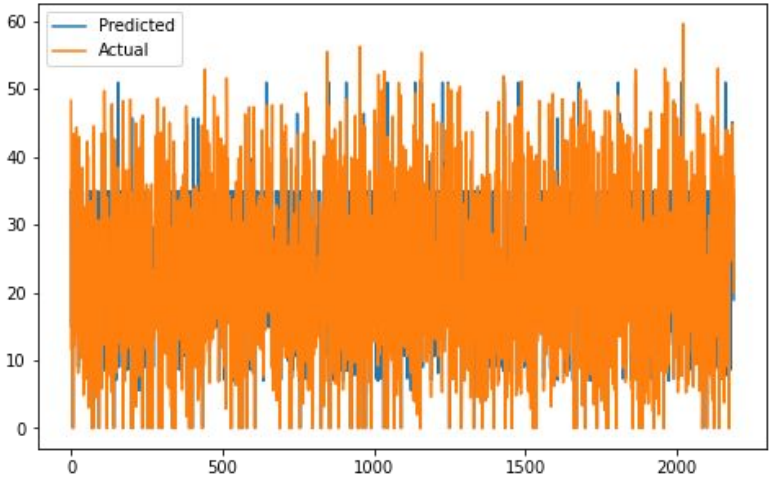
Evaluation metrics Actual vs Predicted:

- MSE : 37.273629840755824
- RMSE : 6.105213329438885
- R2 : 0.750827145595724
- Adjusted R2 : 0.745240832185446

Decision Tree

Evaluation metrics Test:

- Model Score: 0.795415208822012 8
- MSE : 32.04533125397695
- RMSE : 5.660859586138571
- R2 : 0.795415208822012 8
- Adjusted R2 : 0.790285343817777



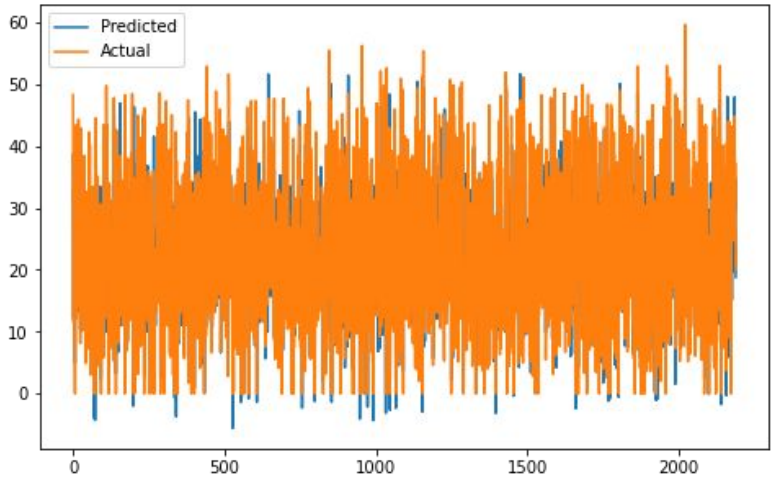
Evaluation metrics Actual vs Predicted:

- MSE : 34.876743604873255
- RMSE : 5.905653529023969
- R2 : 0.766850242558074 3
- Adjusted R2 : 0.761623157851296

XG Booster

Evaluation metrics Test:

- Model Score: 0.892549996227313
- MSE : 16.830532437483136
- RMSE : 4.102503191648135
- R2 : 0.892549996227313
- Adjusted R2 : 0.8901410283706623



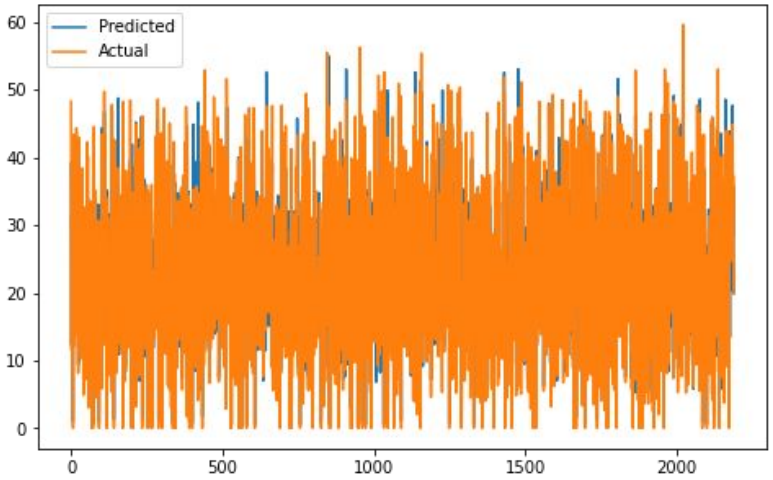
Evaluation metrics Actual vs Predicted:

- Model Score: 0.892549996227313
- MSE : 19.446866450146597
- RMSE : 4.409860139522181 5
- R2 : 0.8699984079011416
- Adjusted R2 : 0.8670838462847262

Gradient Boosting

Evaluation metrics Test:

- Model Score: 0.968691928550953 7
- MSE : 4.903969228265
- RMSE : 2.214490737904541
- R2 : 0.968691928550953 7
- Adjusted R2 : 0.967990019429256 2



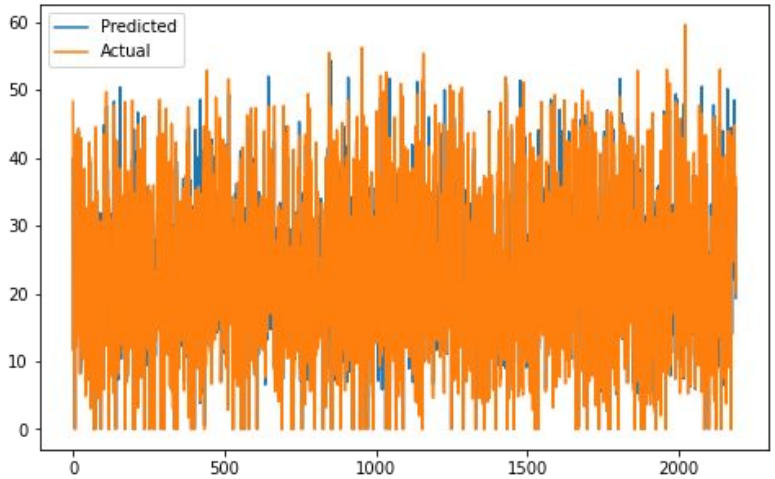
Evaluation metrics Actual vs Predicted:

- Model Score: 0.892375005071604 8
- MSE : 16.09956362287185
- RMSE : 4.012426151703212
- R2 : 0.892375005071604 8
- Adjusted R2 : 0.889962114012958

Random Forest Regressor

Evaluation metrics Test:

- Model Score: 0.985012541920559
- MSE : 2.3475745975317563
- RMSE : 1.532179688395508
- R2 : 0.985012541920559
- Adjusted R2 : 0.9846765316506789



Evaluation metrics Actual vs Predicted:

- Model Score: 0.8934260679169712
- MSE : 15.94233571162436
- RMSE : 3.992785457750561 5
- R2 : 0.8934260679169712
- Adjusted R2 : 0.8910367410883933

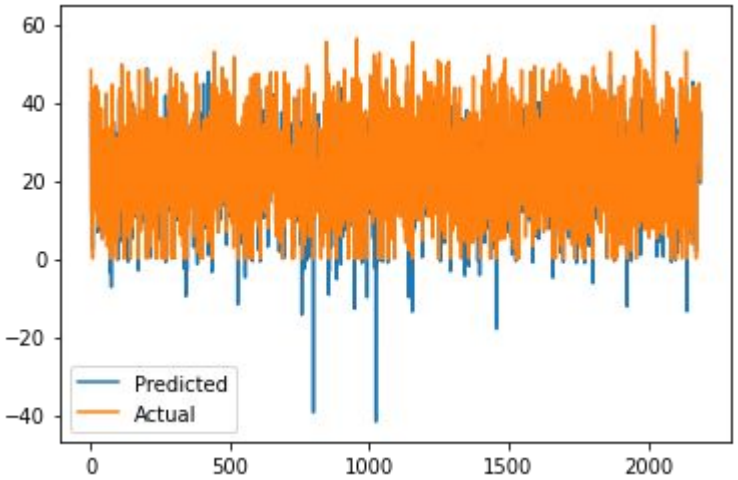
Lasso Regression

Evaluation metrics Test:

- MSE : 37.13314681383164
- RMSE : 6.093697302445506
- R2 : 0.7629334199581261
- Adjusted R2 : 0.757618522320568 9

Evaluation metrics Actual vs Predicted:

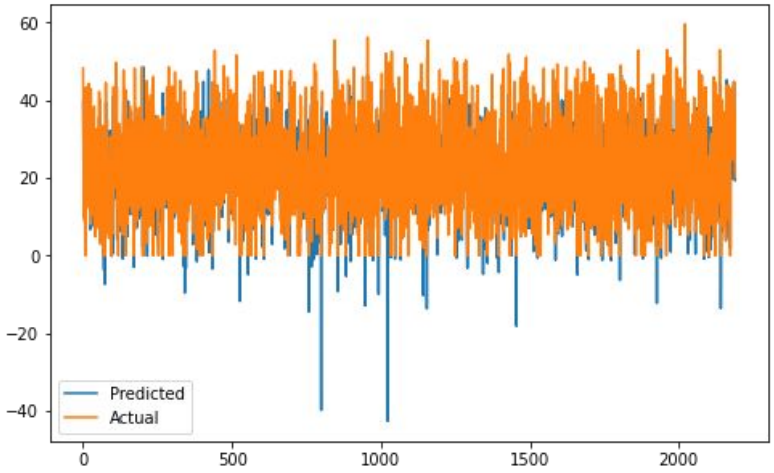
- MSE : 37.18201422405232
- RMSE : 6.097705652460794
- R2 : 0.7514395926479578
- Adjusted R2 : 0.7458670099516018



Ridge Regressor

Evaluation metrics Test:

- MSE : 37.13315291299508
- RMSE : 6.093697802893993
- R2 : 0.762933381019659
- Adjusted R2 : 0.7576184825091236



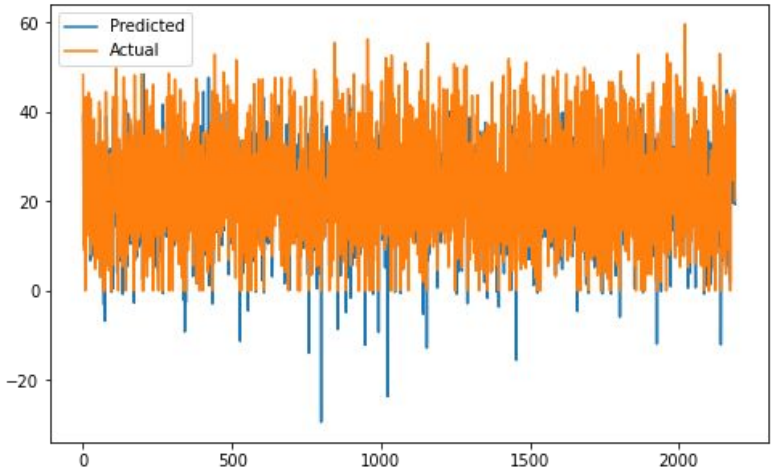
Evaluation metrics Actual vs Predicted:

- MSE : 37.26764694572257
- RMSE : 6.104723330808905
- R2 : 0.7508671410305504
- Adjusted R2 : 0.7452817242951307

Elastic Net Regressor

Evaluation metrics Test:

- MSE : 37.47983954768055
- RMSE : 6.122078041619573
- R2 : 0.7607200535243308
- Adjusted R2 : 0.7553555334725643



Evaluation metrics Actual vs Predicted:

- MSE : 36.46986831447373 6
- RMSE : 6.039028755890614
- R2 : 0.756200262048822
- Adjusted R2 : 0.7507344108476746

Model Comparison Train

	Train MSE	Train RMSE	Train R^2	Train Adjusted R^2
Linear regression	37.133147	6.093697	0.762933	0.757619
DecisionTreeRegressor	32.045331	5.660860	0.795415	0.790829
XGBoost	16.830532	4.102503	0.892550	0.890141
GradientBoostingRegressor	4.903969	2.214491	0.968692	0.967990
Random_forest	2.347575	1.532180	0.985013	0.984677
Lasso Regressor	37.133147	6.093697	0.762933	0.757619
Ridge regreson	37.133153	6.093698	0.762933	0.757618
ElasticNet	37.479840	6.122078	0.760720	0.755356

Model Comparison Test

	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
Linear regression	37.273630	6.105213	0.750827	0.745241
DecisionTreeRegressor	34.876744	5.905654	0.766850	0.761623
XGBoost	19.446866	4.409860	0.869998	0.867084
GradientBoostingRegressor	16.099564	4.012426	0.892375	0.889962
Random_forest	15.942336	3.992785	0.893426	0.891037
Lasso regression	37.182014	6.097706	0.751440	0.745867
Ridge regreson	37.267647	6.104723	0.750867	0.745282
ElasticNet	36.469868	6.039029	0.756200	0.750734

Model Comparison Inference

From all the models that we have trained we found following models to be more suited to our data in order of preference due to consistent favourable results in both test and train scenarios

1. Random Forest
2. Gradient Boosting Regressor
3. XGBoost

Challenges Faced

- Comprehending the problem statement, and understanding the business implications
- Feature engineering – deciding on which features to be dropped / kept / transformed
- Choosing the best visualization to show the trends among different features clearly in the EDA phase
- Deciding on how to handle outliers
- Choosing the ML models to make predictions
- Deciding the evaluation metric to evaluate the models
- Choosing the best hyperparameters, which prevents overfitting

Conclusion

- On non-holiday the rented bike count is much higher
- Highest demand was seen in 2018 compared to 2017 that too in the month of June
- Summer season has the highest demand for rented bikes
- Functioning days has the highest demand for bikes at around 18:00 PM
- Rainfall, Snowfall and Solar radiation were not correlated to rented bike count as they mostly had zero values by magnitude.
- There was no multicollinearity in the attributes but temperature showed the highest correlation with the dew point temperature
- Rented bike count vs Density curve is rightly skewed which required normalization for symmetrical skewness.

Conclusion(Contd.)

- We have successfully built predictive models that can predict the demand for rental bikes based on different weather conditions and other factors and, they were evaluated using RMSE
- The Random forest prediction model had the lowest RMSE
- The final choice of model for deployment depends on the business need; if high accuracy in results is necessary, we can deploy Random forest model
- If the model interpretability is important to the stakeholders, we can choose to deploy the decision tree model.
- Lasso, Ridge, Elastic Net and Linear regressions had almost the same R squared value since the alpha coefficient for Lasso, Ridge and Elastic Net was near to 0 making them concurrent with Linear regression model.

Thank You!