SUPERVISED ML REGRESSION CAPSTONE PROJECT

BIKE SHARING DEMAND PREDICTION

ROADMAP

- ➢ Introduction
- ➢ Problem Statement
- ➢ Data Summary
- ➢ Exploratory Data Analysis
- ➢ Modeling Overview
- ➢ Feature Importance
- ➢ Residual Analysis
- ➢ Conclusion

INTRODUCTION

A bike rental or bike hire business rents out motorcycles for short periods of time, Usually for a few hours. Most rentals are provided by bike shops
as a sideline to their main businesses of sales and service, but some shops specialize in rentals.

As with car rental, bicycle rental shops primarily serve people who do not have access to vehicles, typically travelers and particularly tourists.

Bike rental shops rent by the day or week as well as by the hour, and these provide an excellent opportunity for those who would like to avoid shipping their own bikes but would like to do a multi-day bike tour of a particular area.

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 Analysis Steps

Imported Libraries

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

Statistical Summary

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this told mean, median, standard deviation

Missing Value Imputation

We will now check for missing values in our dataset. after checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

Graphical Representation

We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models driving the Demand for bikes.

Data Summary

Date: Date in year-month-day format

Rented Bike Count: Count of bikes rented at each hour

Hour: Hour of the Day

Temperature: Temperature in Celsius

Humidity: Humidity in %

Windspeed: Speed of wind in m/s

Visibility (10m): Visibility

Dew point temperature: Dew Point Temp (Celsius)

Solar radiation: Radiation in MJ/m2

Rainfall: Rainfall (mm)

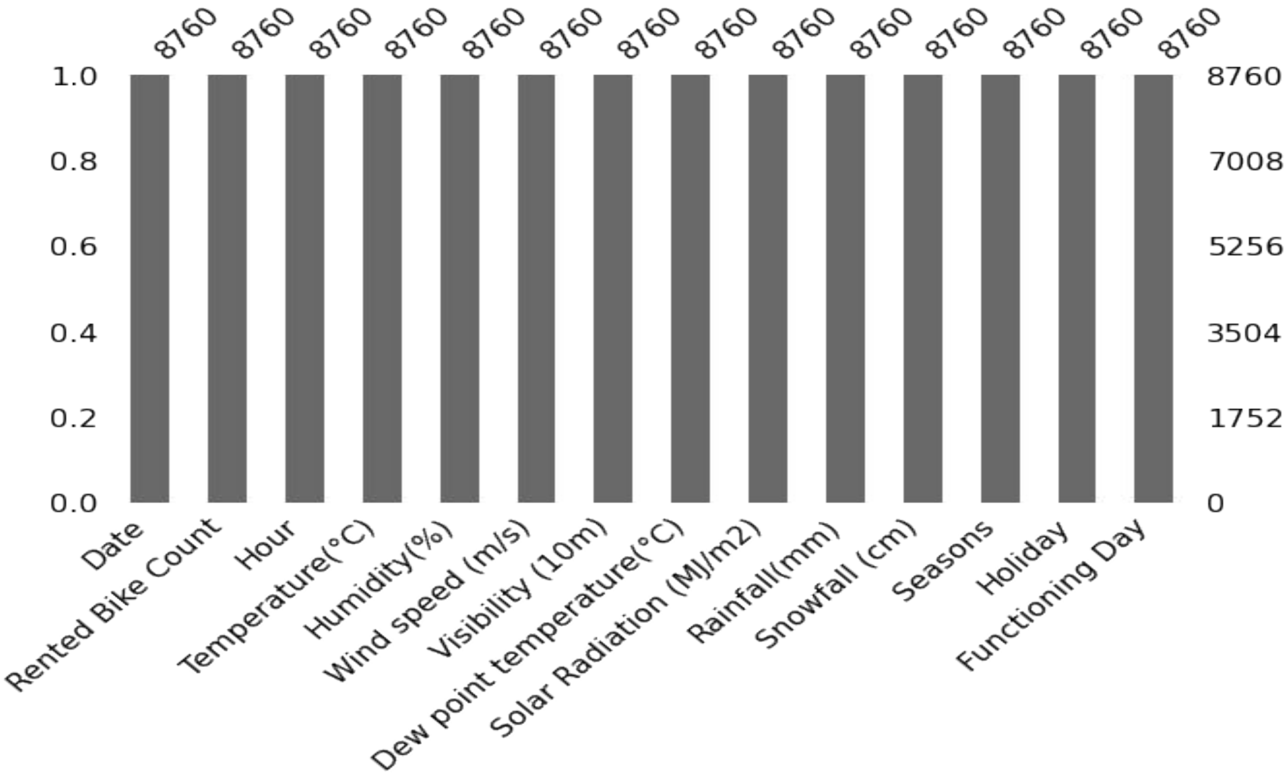
Snowfall: Snowfall (cm)

Seasons: Winter, Spring, Summer, Autumn

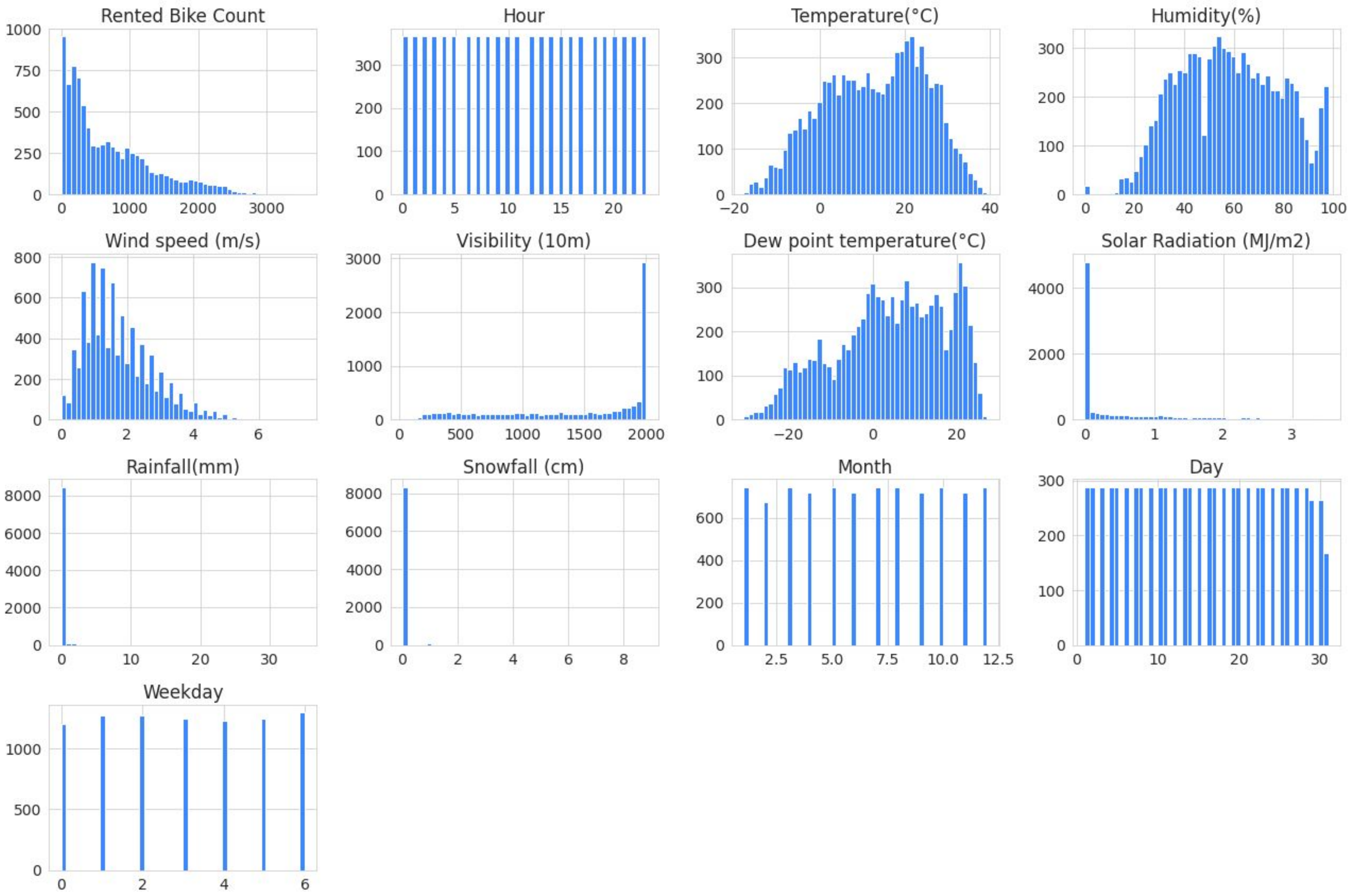
Holiday: Holiday/No holiday

Functioning Day: if the day is neither weekend, holiday.

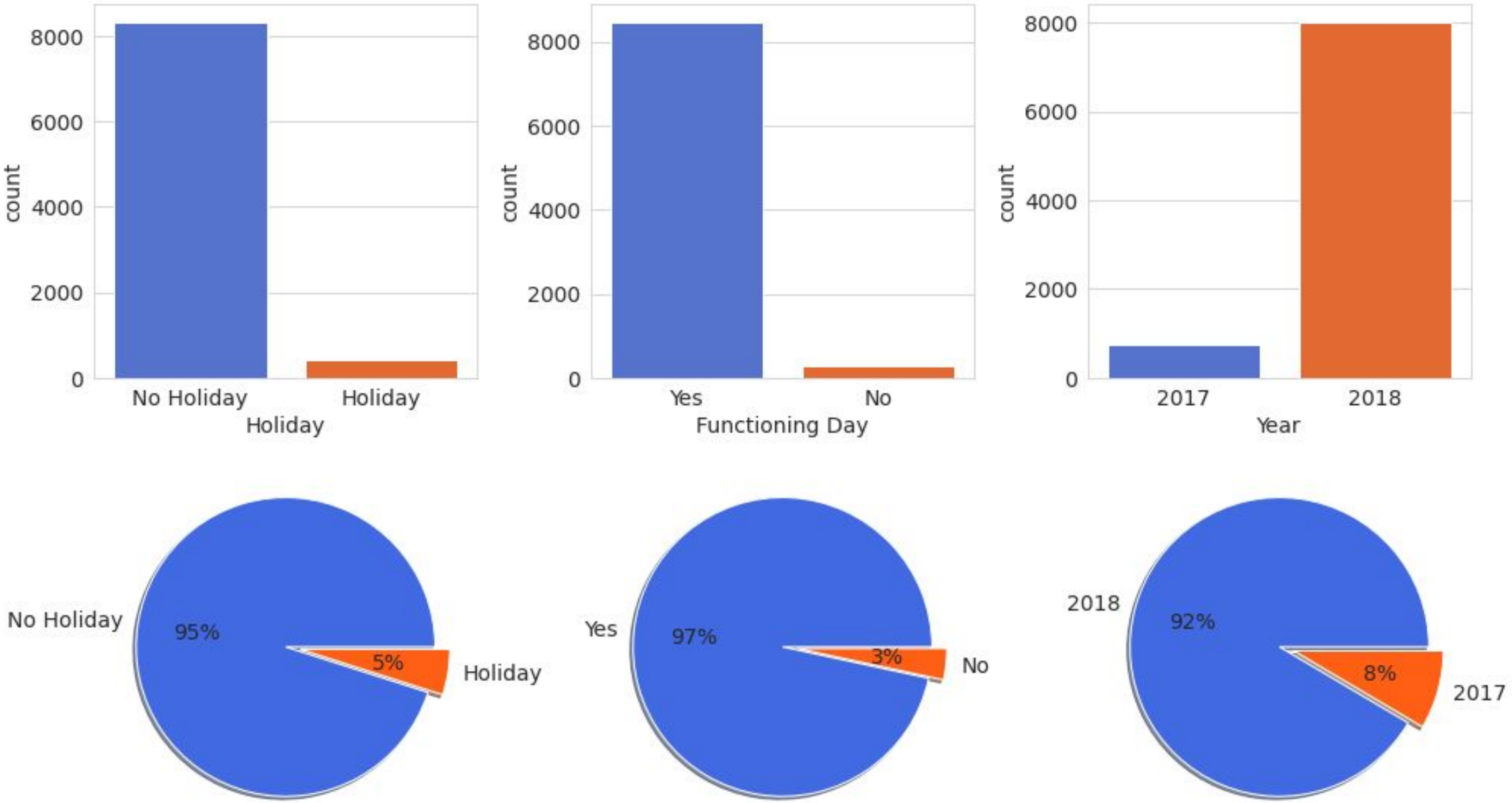
No missing values



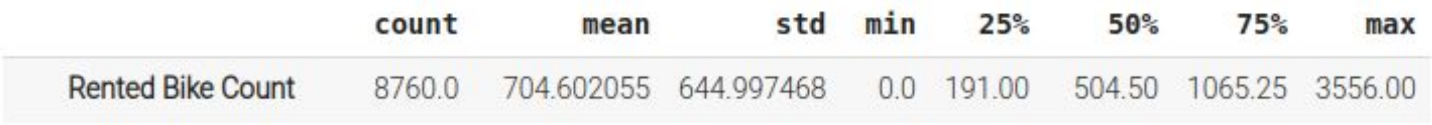
A quick look of Histogram of each numerical attributes



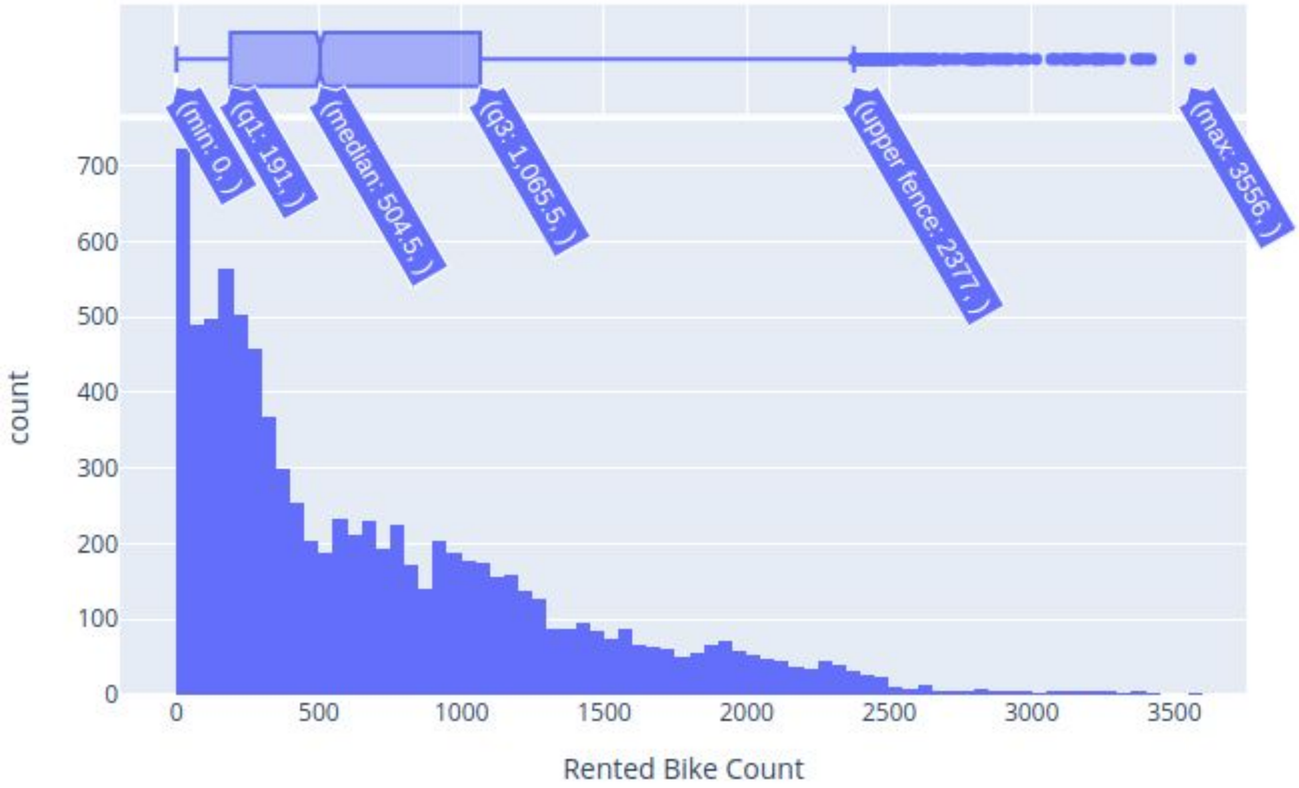
BINARY FEATURES



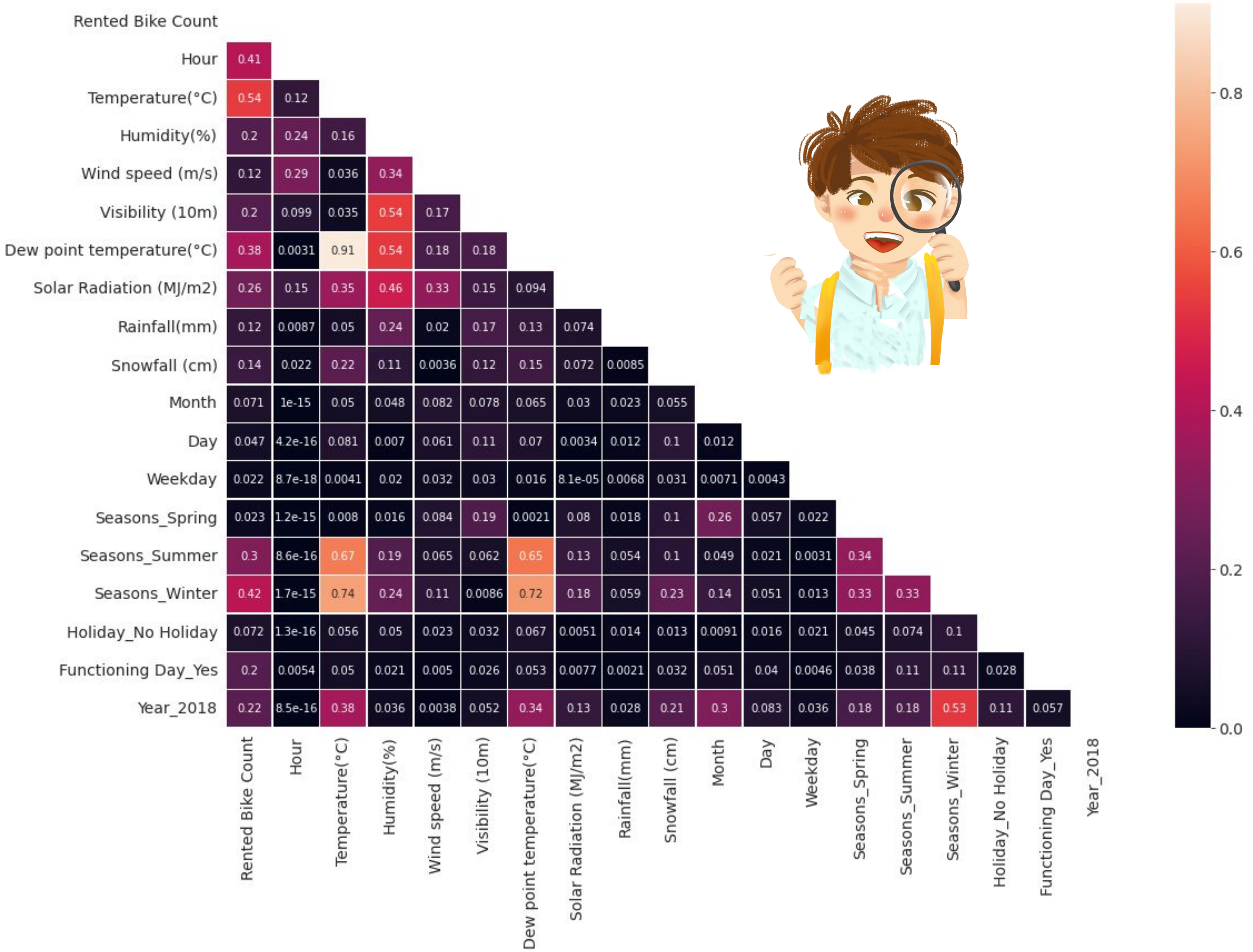
Histogram and Summary of Rented Bike Count



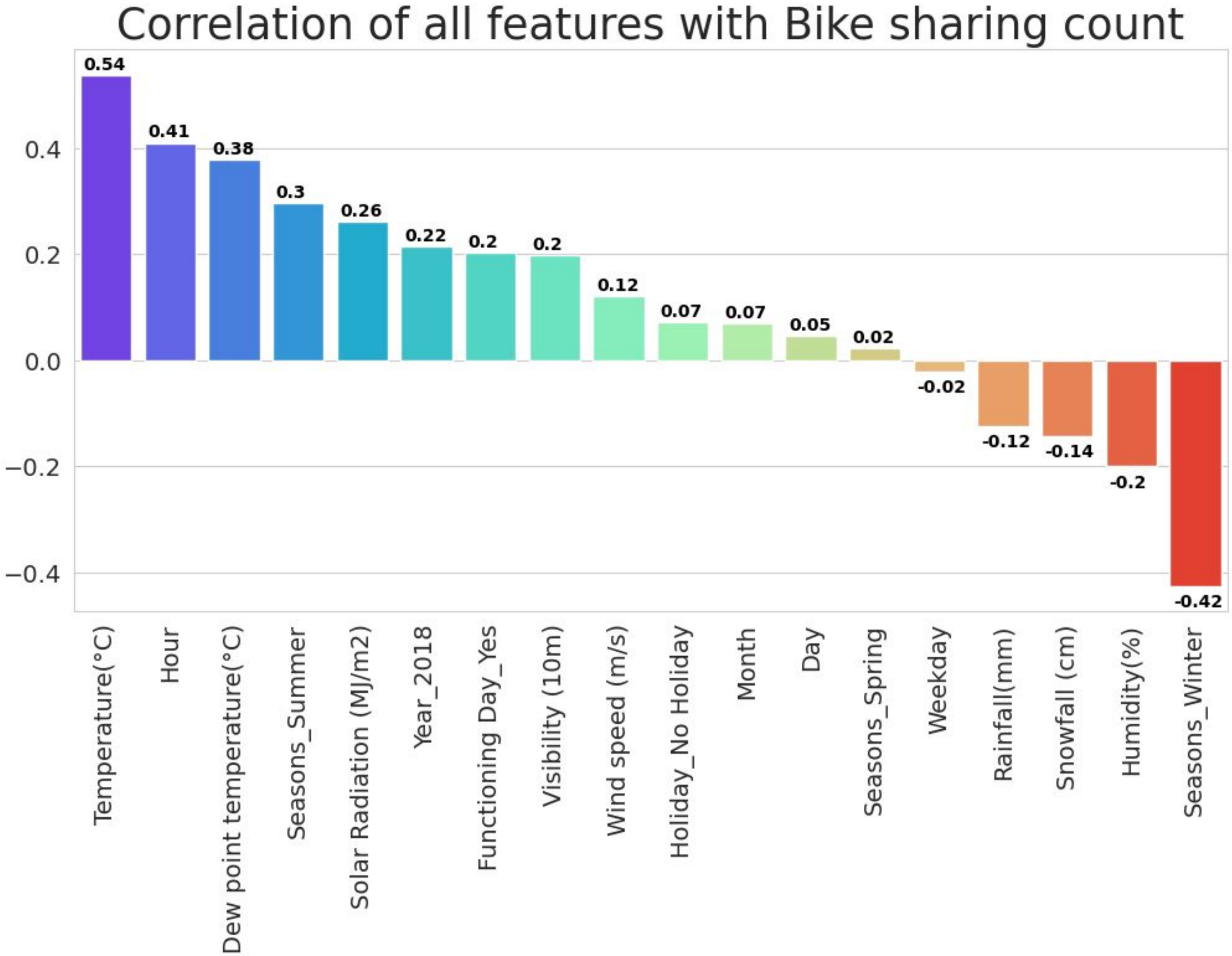
	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00



Display Heat Map of correlation matrix

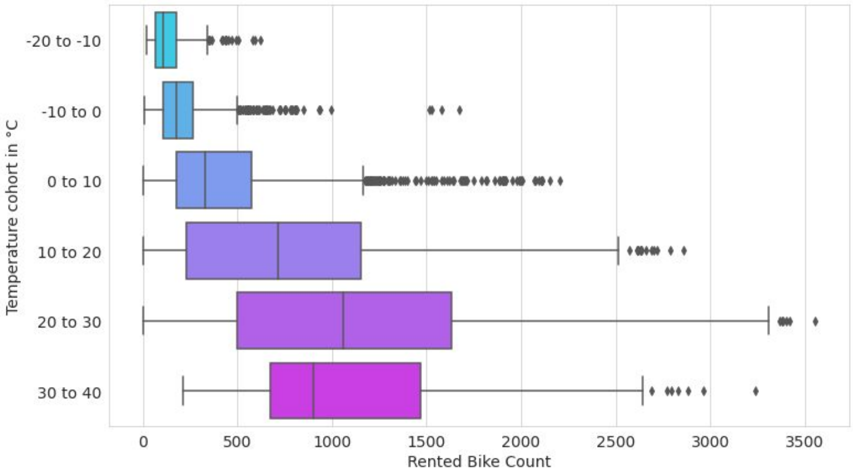


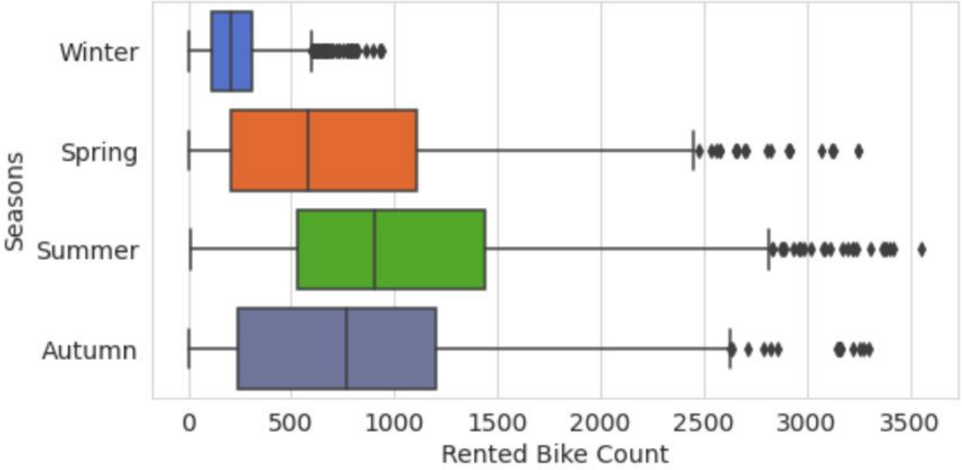
	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Month	Day	Weekday	Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_No Holiday	Functioning Day_Yes	Year_2018
Rented Bike Count																			
Hour	0.41																		
Temperature(°C)	0.54	0.12																	
Humidity(%)	0.2	0.24	0.16																
Wind speed (m/s)	0.12	0.29	0.036	0.34															
Visibility (10m)	0.2	0.099	0.035	0.54	0.17														
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18													
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094												
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074											
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085										
Month	0.071	1e-15	0.05	0.048	0.082	0.078	0.065	0.03	0.023	0.055									
Day	0.047	4.2e-16	0.081	0.007	0.061	0.11	0.07	0.0034	0.012	0.1	0.012								
Weekday	0.022	8.7e-18	0.0041	0.02	0.032	0.03	0.016	8.1e-05	0.0068	0.031	0.0071	0.0043							
Seasons_Spring	0.023	1.2e-15	0.008	0.016	0.084	0.19	0.0021	0.08	0.018	0.1	0.26	0.057	0.022						
Seasons_Summer	0.3	8.6e-16	0.67	0.19	0.065	0.062	0.65	0.13	0.054	0.1	0.049	0.021	0.0031	0.34					
Seasons_Winter	0.42	1.7e-15	0.74	0.24	0.11	0.0086	0.72	0.18	0.059	0.23	0.14	0.051	0.013	0.33	0.33				
Holiday_No Holiday	0.072	1.3e-16	0.056	0.05	0.023	0.032	0.067	0.0051	0.014	0.013	0.0091	0.016	0.021	0.045	0.074	0.1			
Functioning Day_Yes	0.2	0.0054	0.05	0.021	0.005	0.026	0.053	0.0077	0.0021	0.032	0.051	0.04	0.0046	0.038	0.11	0.11	0.028		
Year_2018	0.22	8.5e-16	0.38	0.036	0.0038	0.052	0.34	0.13	0.028	0.21	0.3	0.083	0.036	0.18	0.18	0.53	0.11	0.057	



Correlation of all features with Bike sharing count

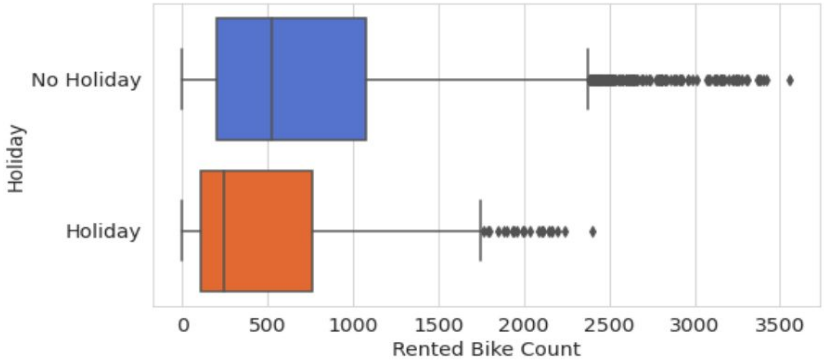
Bike Rent count with Temperature cohort And Seasons

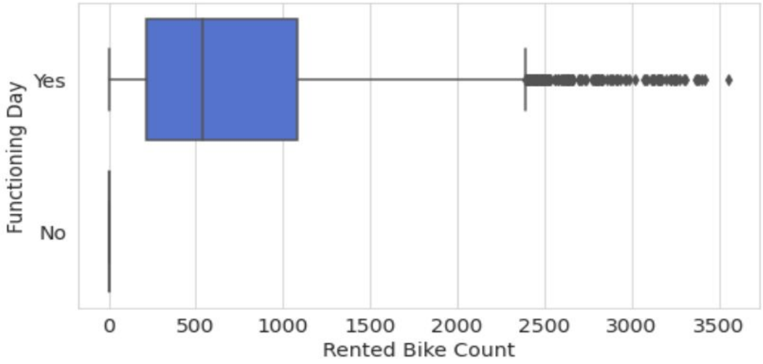


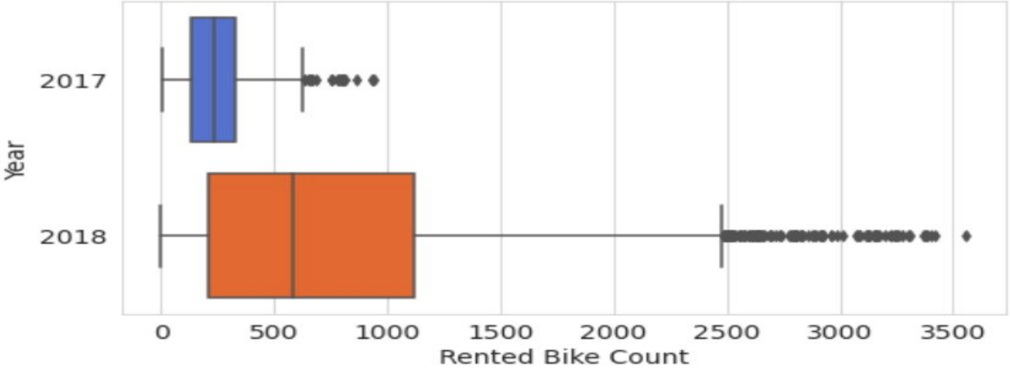


- ➢ Summer is the most preferred season to rent a bike and winter is the least.
- ➢ We see that temperatures between 20 and 30 are most favorable for renting, 30 to 40 in second and 10 to 20 in third. People do not like to rent bikes in extreme cold. And we've seen this pattern in this season's month as well.

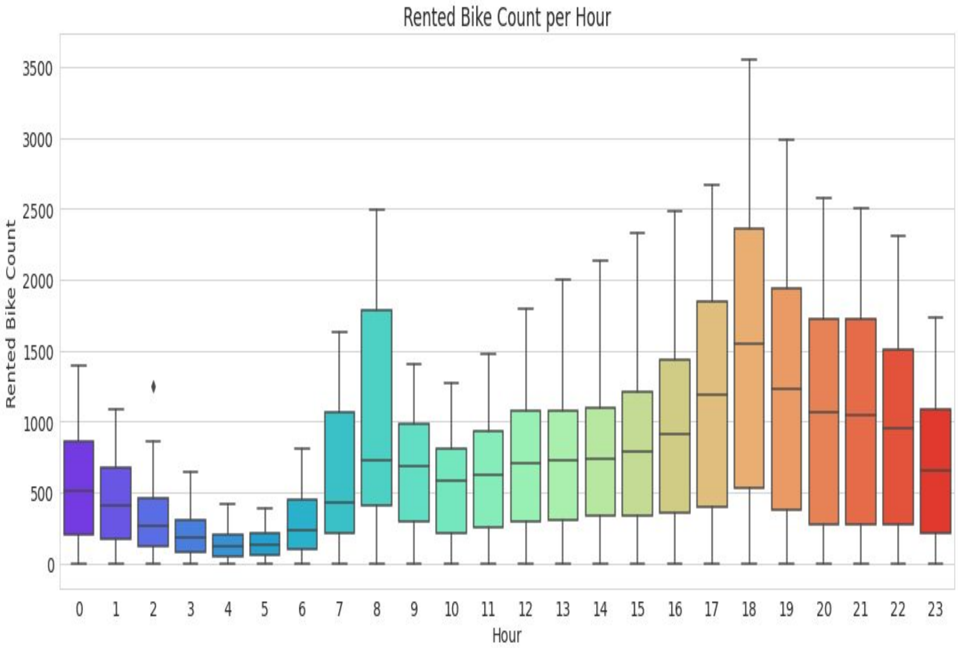
Bikes Rented with Categorical Features

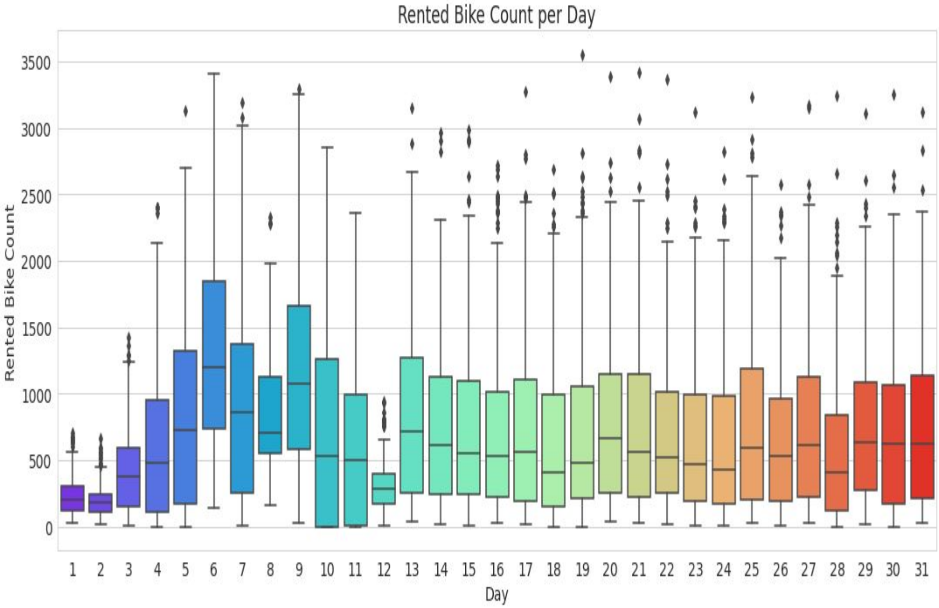




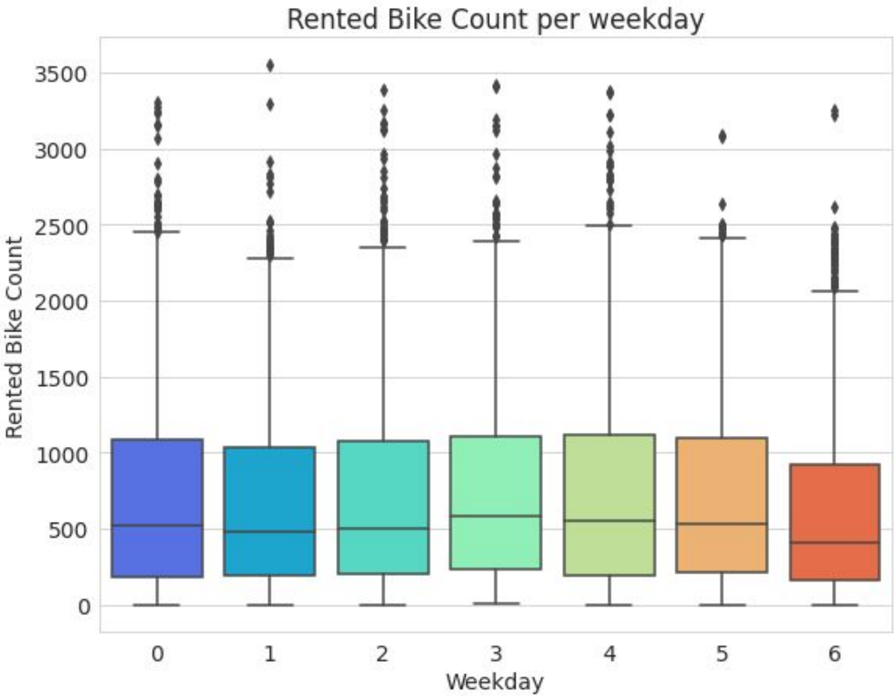


- ➢ People prefer to book bikes on working days rather than holidays.
- ➢ All bikes were hired on working day, Most of the customers are from the working class.
- ➢ People booked more bikes in 2018 than in 2017, the company may not be very popular in the year 2017.

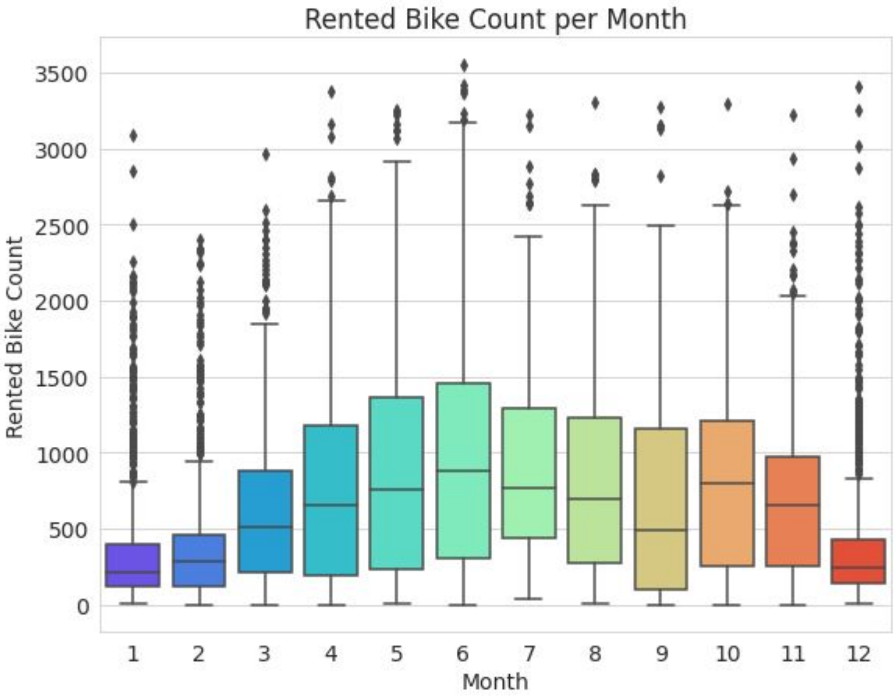




- Bike sharing is at its peak at 6pm and high between 4pm to 10pm, also at 8 am (may be due to office time) demand is also high.
- Bike sharing is least between 4am-6am.
- 1, 2 and 12 date is least preferred day and 6 and 9 is the most.



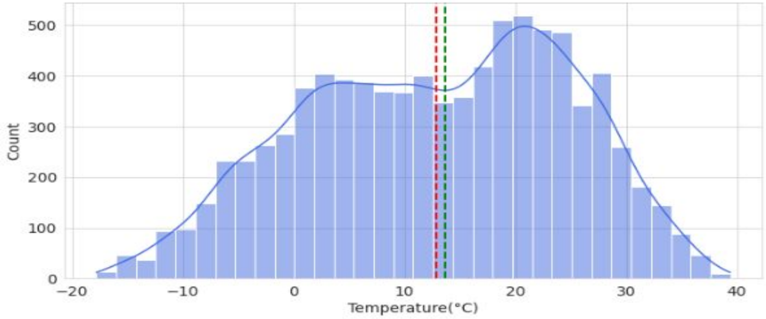
Rented Bike Count per weekday

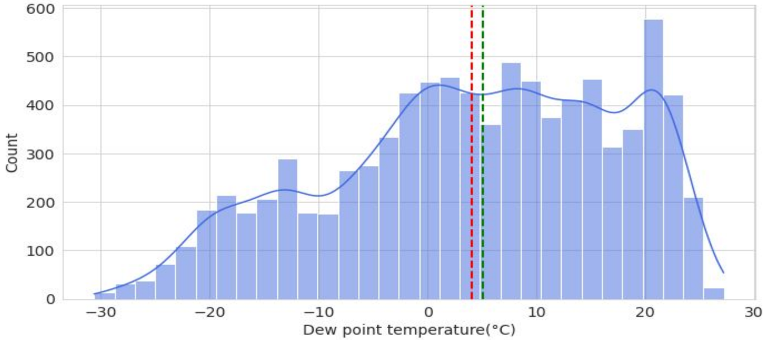


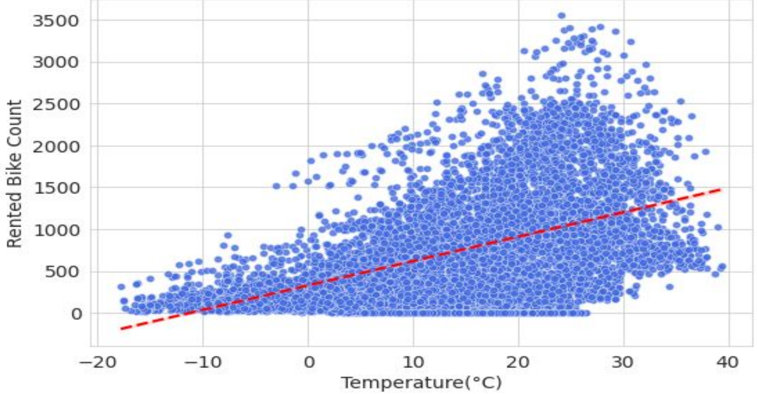
Rented Bike Count per Month

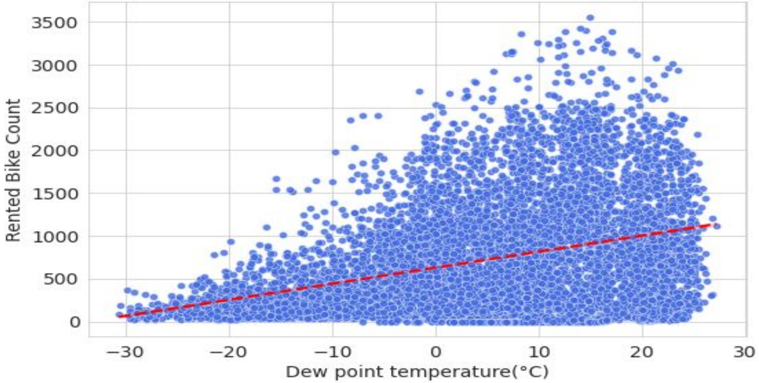
- ➢ June is the most preferred Month for bike sharing, july and May are the second best and least in December and January.
- ➢ Weekday have no impact on bike sharing count.(except sunday)

Rented Bike Count Against Numerical Features





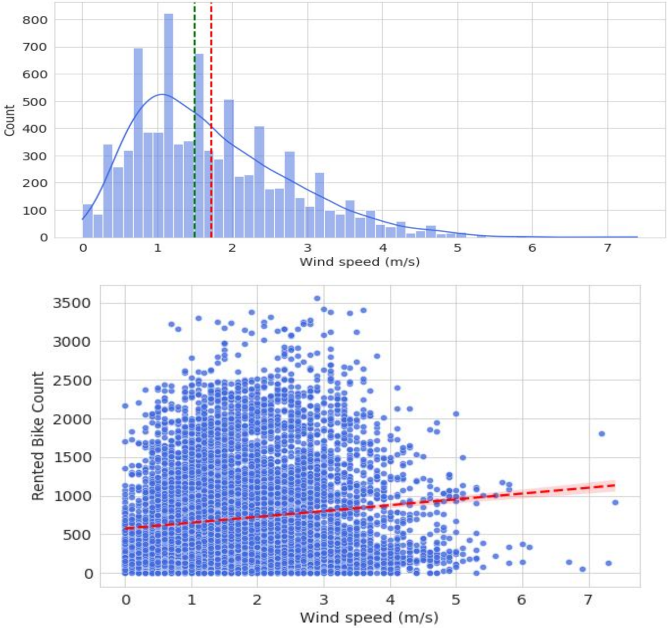




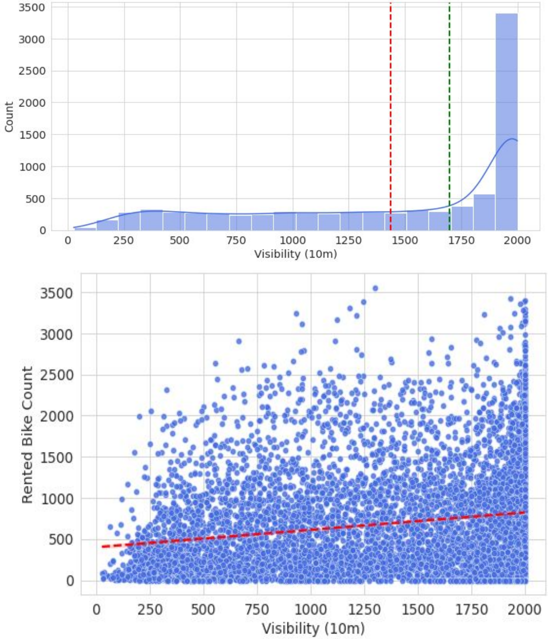
0.54

0.38

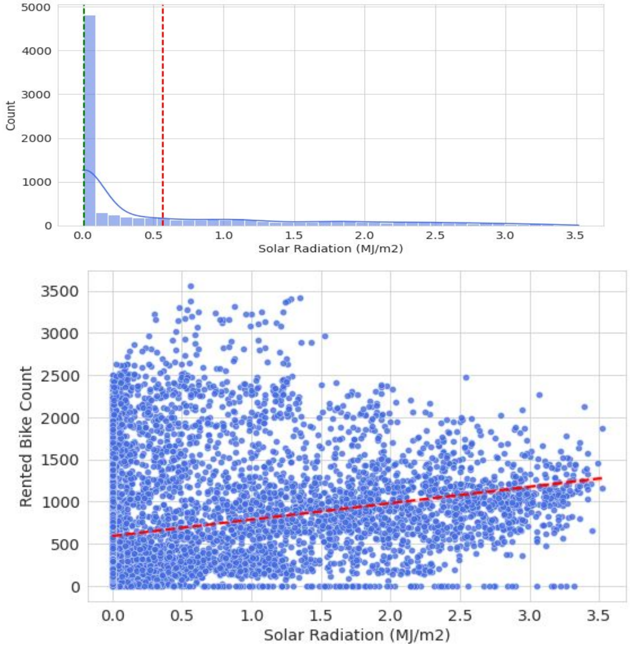
Rented Bike Count Against Numerical Features



0.12

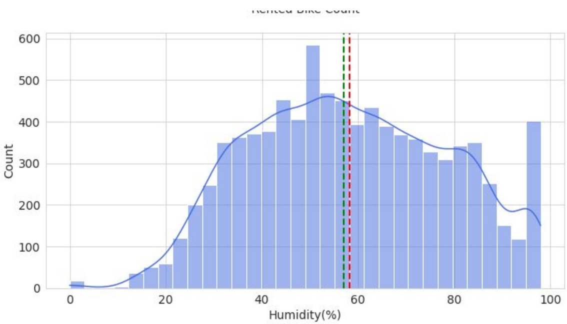


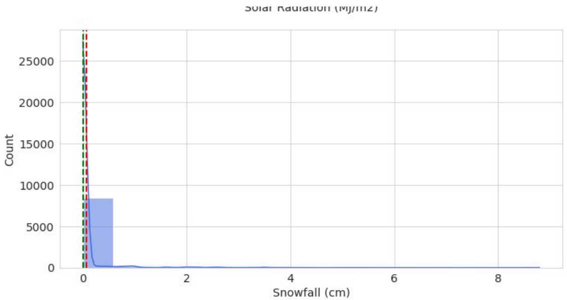
0.2

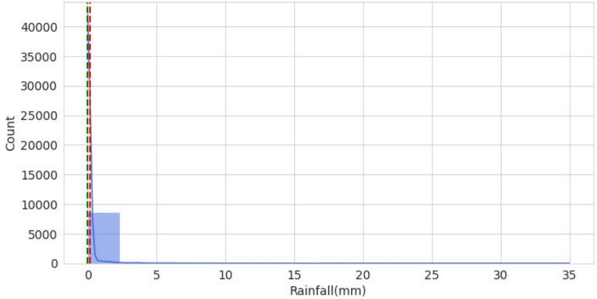


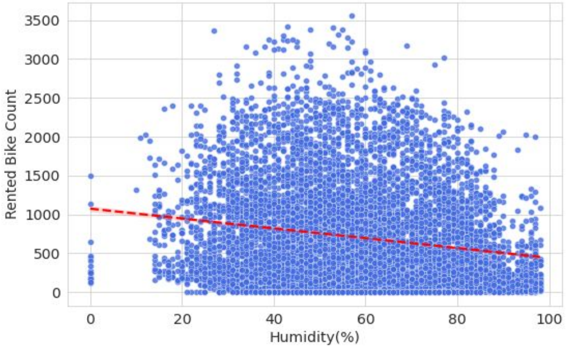
0.26

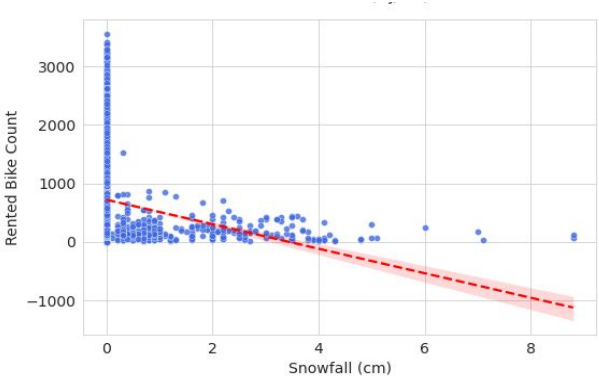
Rented Bike Count Against Numerical Features

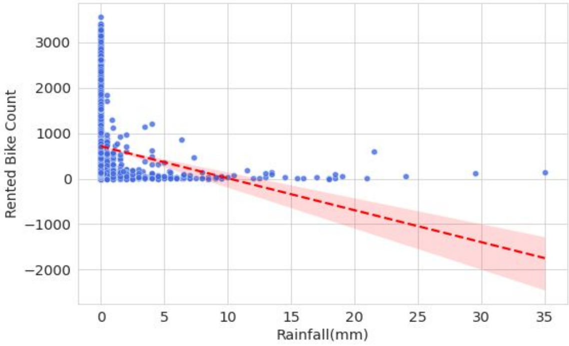










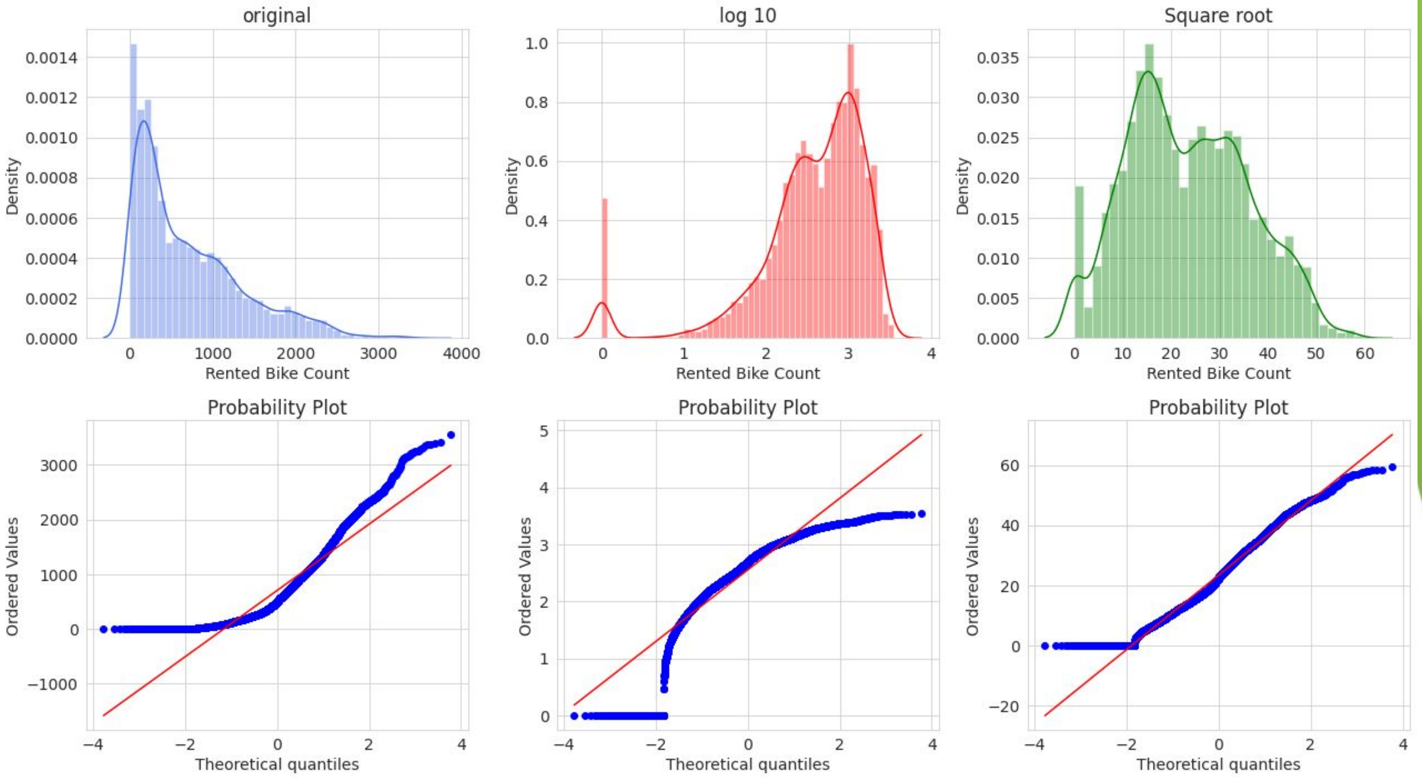


-0.2

-0.14

-0.12

Transformation of dependent variable to reduce its skewness



Prepare the Dataset for Training

Stratified Shuffle split

This technique consists of forcing the distribution of the target variable(s) among the different splits to be the same. This small change will result in training on the same population in which it is being evaluated, achieving better predictions.

Min Max Scaler

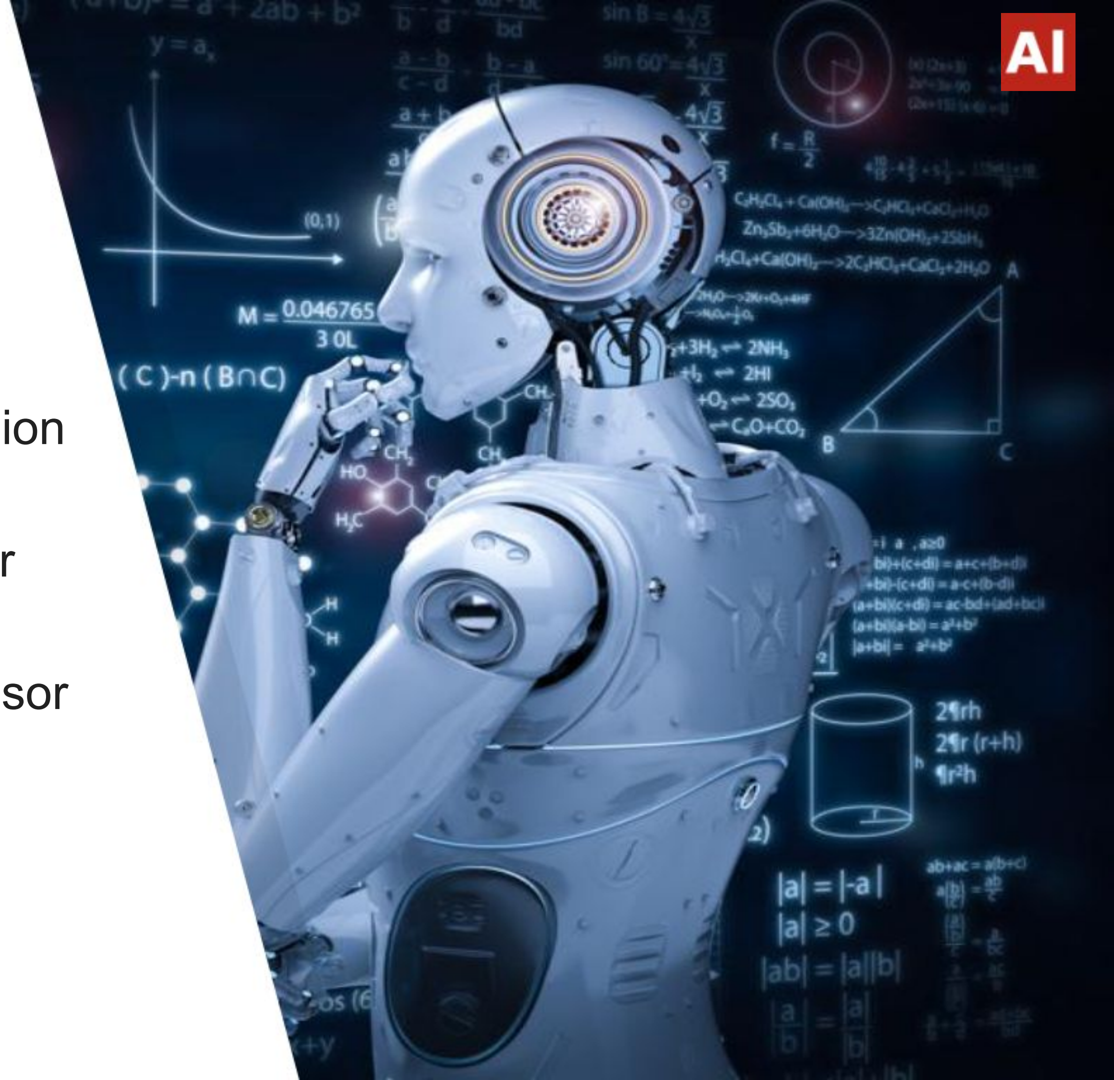
This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

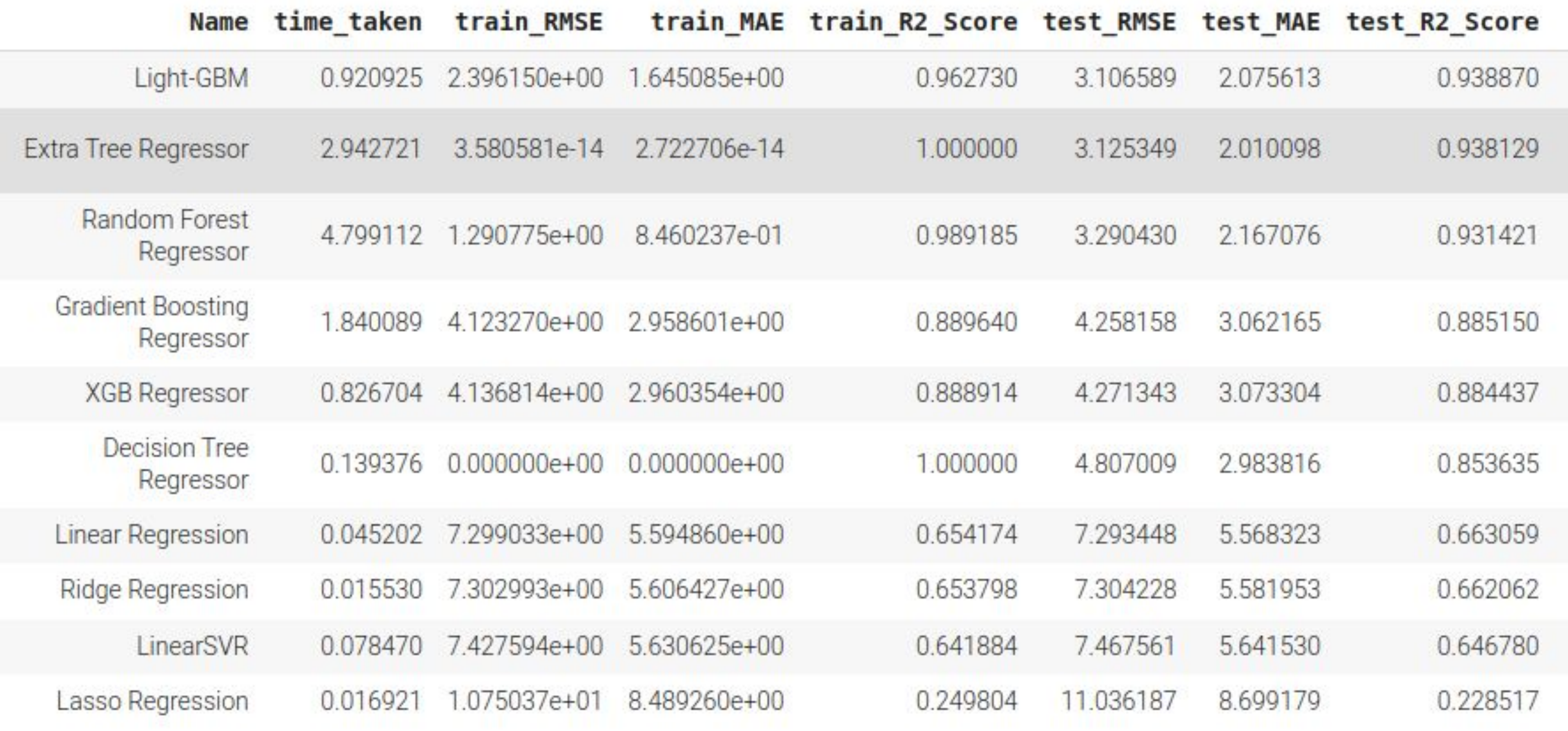
Applying models

I am applying these ML models

1. Linear Regression
2. Ridge and Lasso Regression
3. Decision Tree Regressor
4. Random Forest Regressor
5. Extra Tree Regressor
6. Gradient Boosting Regressor
7. XGB Regressor
8. Light-GBM
9. LinearSVR

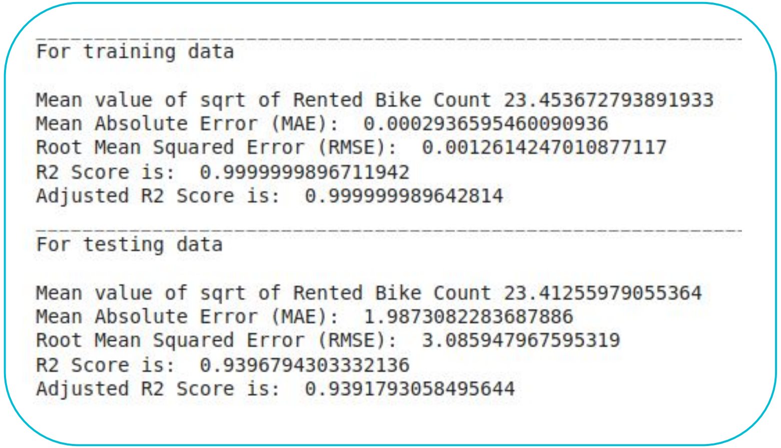


Result



Name	time_taken	train_RMSE	train_MAE	train_R2_Score	test_RMSE	test_MAE	test_R2_Score
Light-GBM	0.920925	2.396150e+00	1.645085e+00	0.962730	3.106589	2.075613	0.938870
Extra Tree Regressor	2.942721	3.580581e-14	2.722706e-14	1.000000	3.125349	2.010098	0.938129
Random Forest Regressor	4.799112	1.290775e+00	8.460237e-01	0.989185	3.290430	2.167076	0.931421
Gradient Boosting Regressor	1.840089	4.123270e+00	2.958601e+00	0.889640	4.258158	3.062165	0.885150
XGB Regressor	0.826704	4.136814e+00	2.960354e+00	0.888914	4.271343	3.073304	0.884437
Decision Tree Regressor	0.139376	0.000000e+00	0.000000e+00	1.000000	4.807009	2.983816	0.853635
Linear Regression	0.045202	7.299033e+00	5.594860e+00	0.654174	7.293448	5.568323	0.663059
Ridge Regression	0.015530	7.302993e+00	5.606427e+00	0.653798	7.304228	5.581953	0.662062
LinearSVR	0.078470	7.427594e+00	5.630625e+00	0.641884	7.467561	5.641530	0.646780
Lasso Regression	0.016921	1.075037e+01	8.489260e+00	0.249804	11.036187	8.699179	0.228517

Hyperparameter Tuning



```
-------------------------------------------------------------------------
For training data

Mean value of sqrt of Rented Bike Count 23.453672793891933
Mean Absolute Error (MAE):  0.00029365954600909936
Root Mean Squared Error (RMSE):  0.0012614247010877117
R2 Score is:  0.9999999896711942
Adjusted R2 Score is:  0.999999989642814

-------------------------------------------------------------------------
For testing data

Mean value of sqrt of Rented Bike Count 23.41255979055364
Mean Absolute Error (MAE):  1.9873082283687886
Root Mean Squared Error (RMSE):  3.085947967595319
R2 Score is:  0.9396794303332136
Adjusted R2 Score is:  0.9391793058495644
```

ExtraTreesRegressor

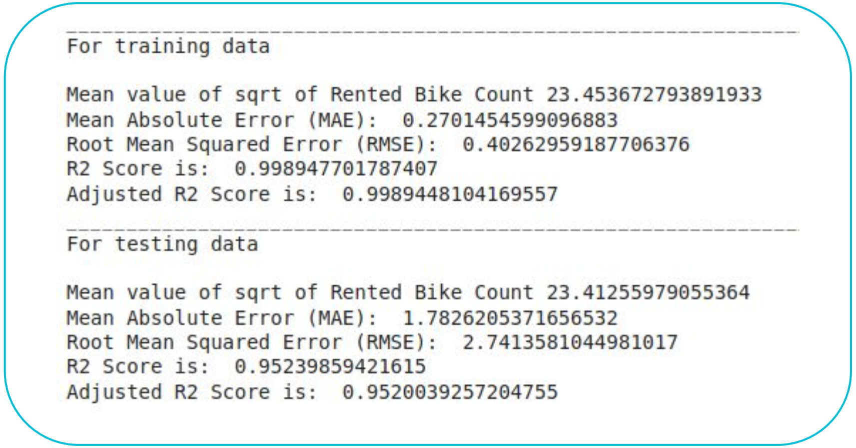
R2 score > 93.81 % to 93.91 %
RMSE > 3.12 to 3.08

RMSE is 13.15 % of 23.41

Light GBM

R2 score > 93.88 % to 95.23 %
RMSE > 3.10 to 2.74

RMSE is 11.70 % of 23.41



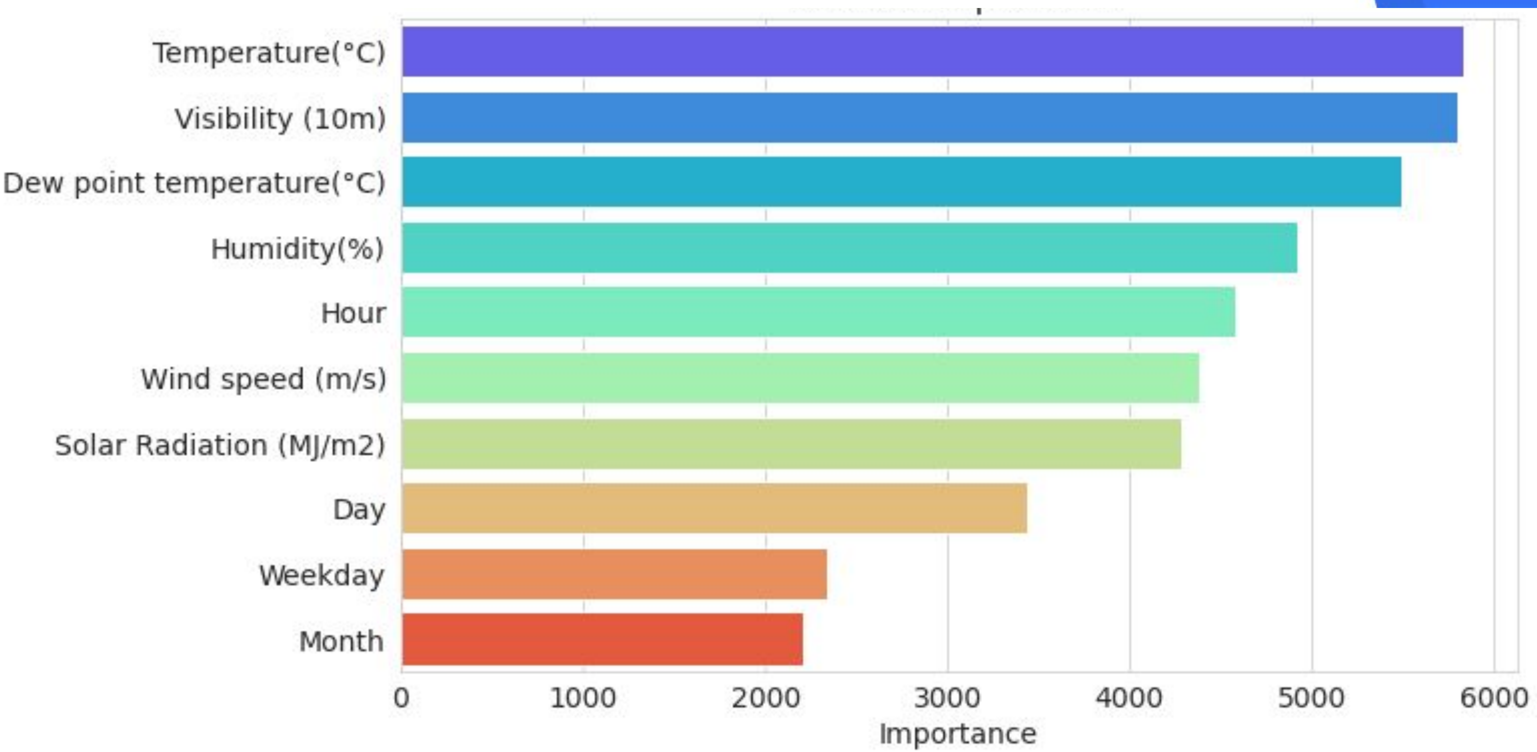
```
-------------------------------------------------------------------------
For training data

Mean value of sqrt of Rented Bike Count 23.453672793891933
Mean Absolute Error (MAE):  0.2701454599096883
Root Mean Squared Error (RMSE):  0.40262959187706376
R2 Score is:  0.998947701787407
Adjusted R2 Score is:  0.9989448104169557

-------------------------------------------------------------------------
For testing data

Mean value of sqrt of Rented Bike Count 23.41255979055364
Mean Absolute Error (MAE):  1.7826205371656532
Root Mean Squared Error (RMSE):  2.7413581044981017
R2 Score is:  0.95239859421615
Adjusted R2 Score is:  0.9520039257204755
```

Best Parameters (LGBMRegressor)

We had chosen LightGBM Regressor for our prediction and best hyperparameters obtained are as below.

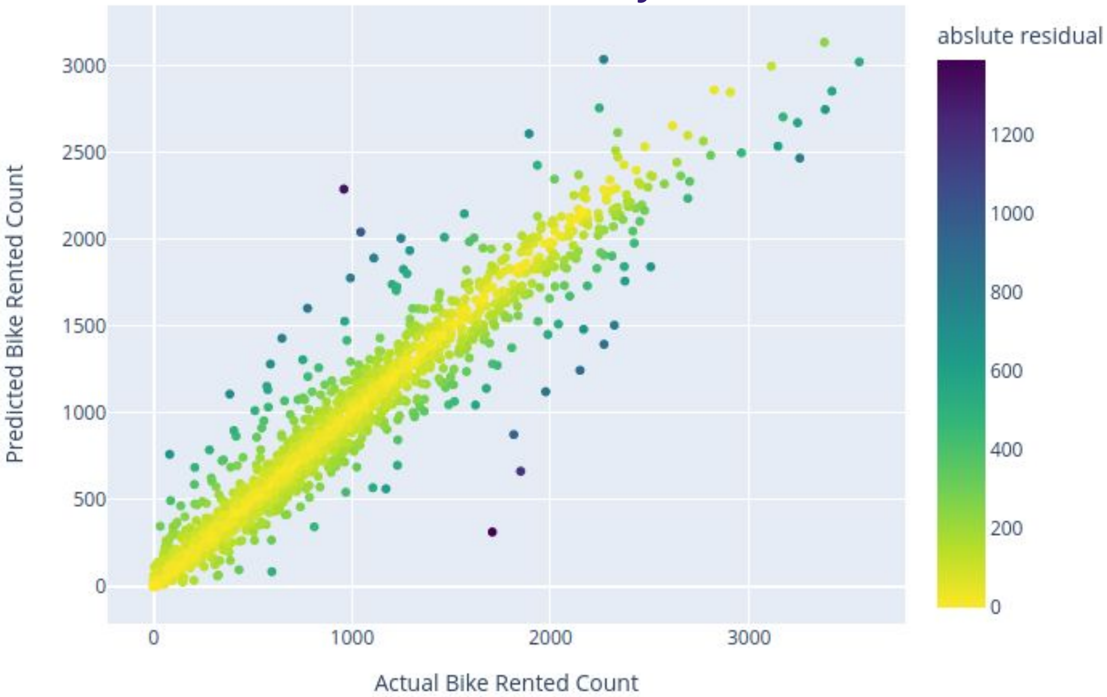
bootstrap = True
max_depth = 30
min_samples_leaf = 1
min_samples_split = 2
n_estimators = 1500
learning_rate = 0.1
oob_score = False

Feature Importance



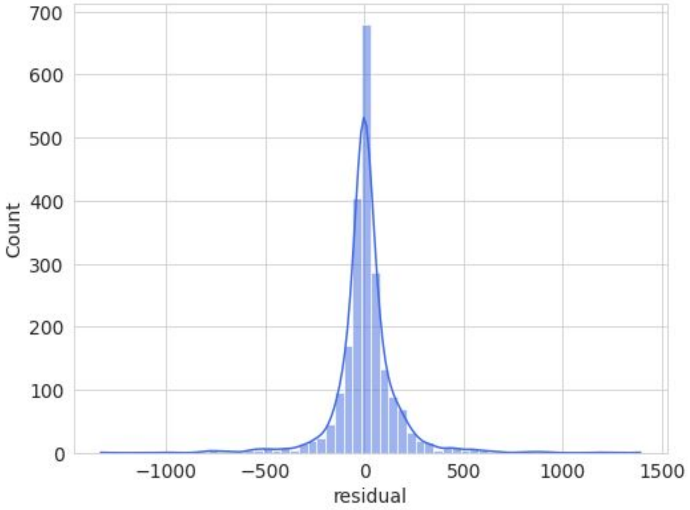
Feature importance refers to **technique that assigns a score to features based on how significant they are at predicting a target variable**

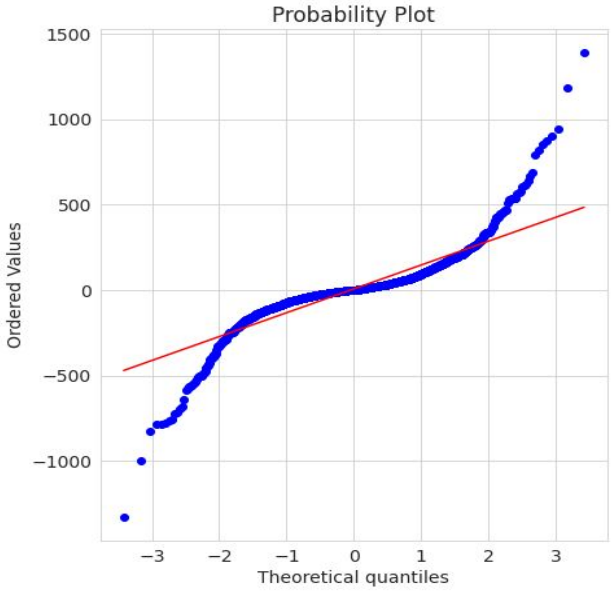
Residual Analysis



- ➢ Correlation between actual and predicted Bike Rented Count is 0.97
- ➢ Mean of Rented Bike Count : 704.60
- ➢ MAE of Rented Bike Count : 88.36, which is 12.54 % of Mean of Rented Bike Count
- ➢ RMSE of Rented Bike Count : 156.26, which is 22.18 % of Mean of Rented Bike Count

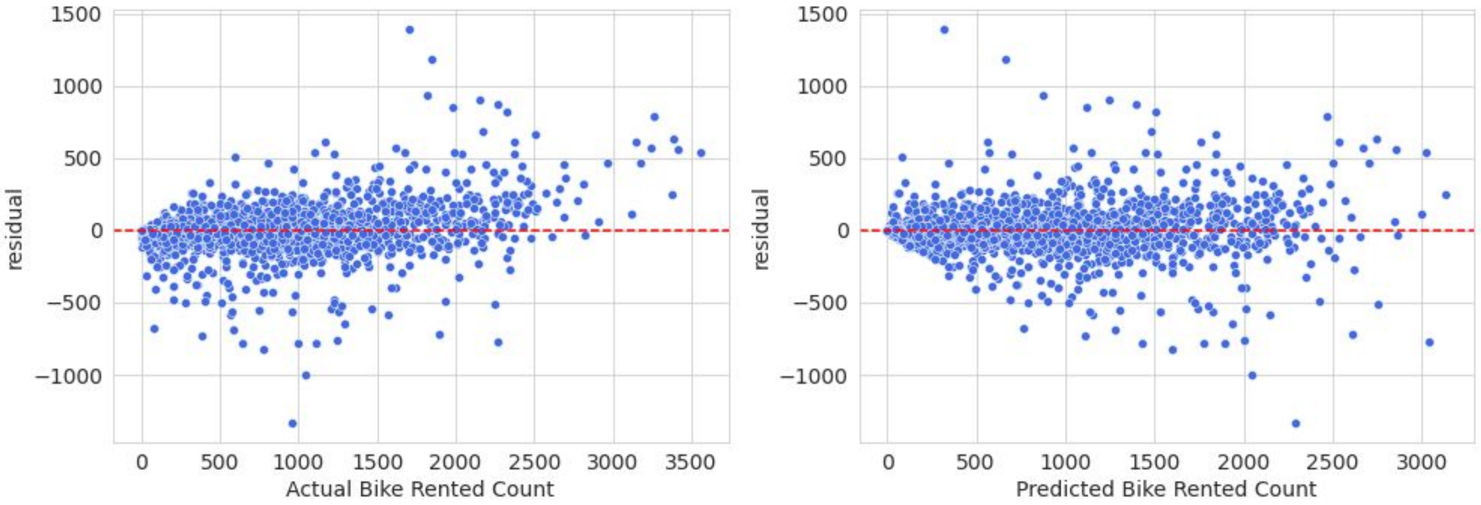
Residual Analysis



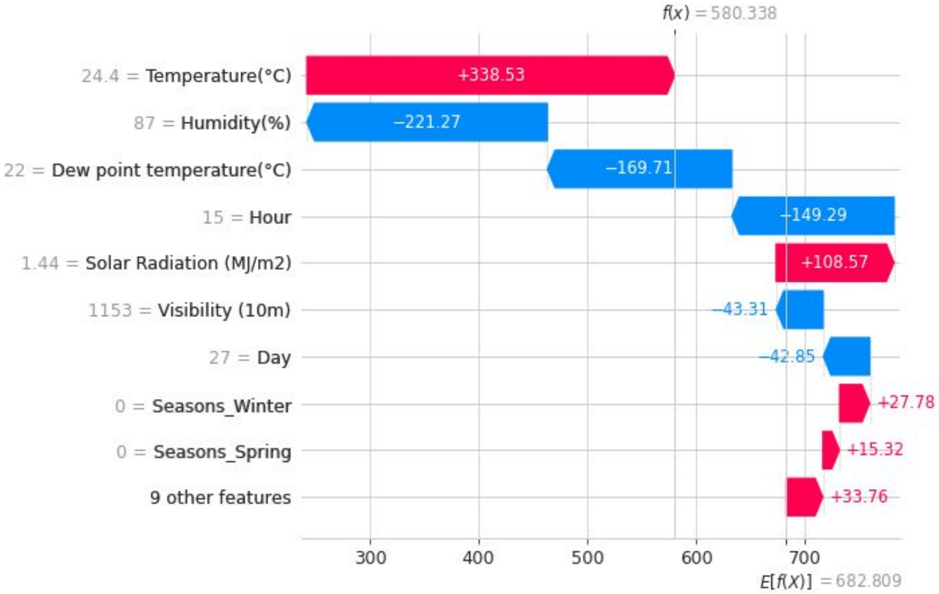


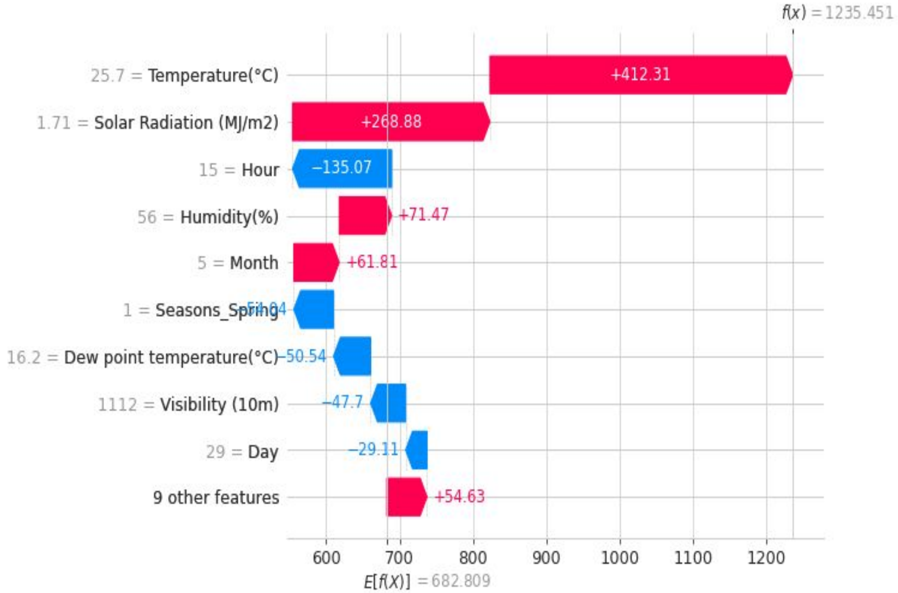
- ➢ For normal distribution, all blue points should lie on red lines.
- ➢ Points between -2 to 2 in Theoretical quantiles approximately follow normal distribution.
- ➢ points greater than 2 and less than -2 in Theoretical quantiles, show that residual plot have high skewness on both the side means it have heavy tail compared to normal distribution.

Homoscedasticity



Decoding the Black Box: SHAP





- For this observation , predicted value is 580.338 and base value is 682.809 (mean of test data).
- Features in red color features are responsible for increasing the value of prediction, while those in blue are said to decrease the value of prediction.

Overall Conclusion

- Most numbers of Bikes were rented in Summer, followed by Autumn, Spring, and Winter. May-July is the peak Bike renting Season, and Dec-Feb is the least preferred month for bike renting.
- Majority of the client in the bike rental sector belongs to the Working class. This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.
- Temperature of 20-30 Degrees, evening time 4 pm- 8 pm,Humidity between 40%-60% are the most favorable parameters where the Bike demand is at its peak.
- Temperature, Hour of the day, Solar radiation, and Humidity are major driving factors for the Bike rent demand.
- Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low. Best predictions are obtained with a LightGBM model with an R2 Score of 0.9523 means 95.23 percent of the variance in the dependent variable that is predictable from the independent variables.
- MAE of Rented Bike Count is 88.36, and root mean square error is 156.26, with overall mean of Rented Bike Count is 704.60.

Thank You!!!

Any questions?