Capstone Project 2

Machine Learning –Regression Project

Bike Sharing Demand Prediction

By

Anjali A Khillare

https://github.com/Anjalikhillare/Bike-Sharing-Demand-Prediction

AlmaBetter

Table of Content

Introduction

An inventive form of transportation is the bike sharing system, which lends out bikes to people in short periods of time . Internationally, the use of bike-sharing programs has significantly increased during the past few decades. This is due to the fact that it is a practical and affordable method of enhancing urban mobility. Additionally, by encouraging better practices among its users, this system also helps to lower fuel use.

Project's objectives are:

- Recognize data trends and important influences on the hourly demand for rental bikes.
- Create a suitable regression model to predict how many rental bikes will be needed per hour.

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 Description

Dependent Variable : Rented Bike Count

Independent Variable:

- Date - The date of each observation in the format 'year-month-day'
- Hour - Hour of the day
- Temperature - Temperature recorded in the city in Celsius (°C).
- Humidity - Relative humidity in %
- Wind speed - Speed of the wind in m/s
- Visibility - Measure of distance at which object or light can be clearly discerned in units of 10m
- Dew point temperature - Temperature recorded in the beginning of the day in Celsius(°C).
- Solar radiation - Intensity of sunlight in MJ/m^2
- Rainfall - Amount of rainfall received in mm
- Snowfall - Amount of snowfall received in cm
- Seasons - Season of the year (Winter, Spring, Summer, Autumn)
- Holiday - Whether the day is a Holiday or not (Holiday/No holiday)
- Functional Day -Whether the rental service is available (Yes-Functional hours) or not (No-Non functional hours)

Steps Involved

- Data preprocessing: Corrected data type and examination for outliers, erroneous values, missing values, and duplicates.
- Feature Extraction: From the Date column, new columns were created, including Day, Month, Year, Days of Week, and Weekend.
- Exploratory Data Analysis: To better understand the distribution of features and their correlations, analysis was carried out using a variety of graphs and plots. Dew point temperature was removed since they were substantially associated with other independent features after checking the VIF value (a measure of multicollinearity).Seasons, Holidays, Weekends, and Functioning Days were dummified as categorical variables in the dataset.
- Feature scaling: Using MinmaxScaler, features were brought to a comparable range.

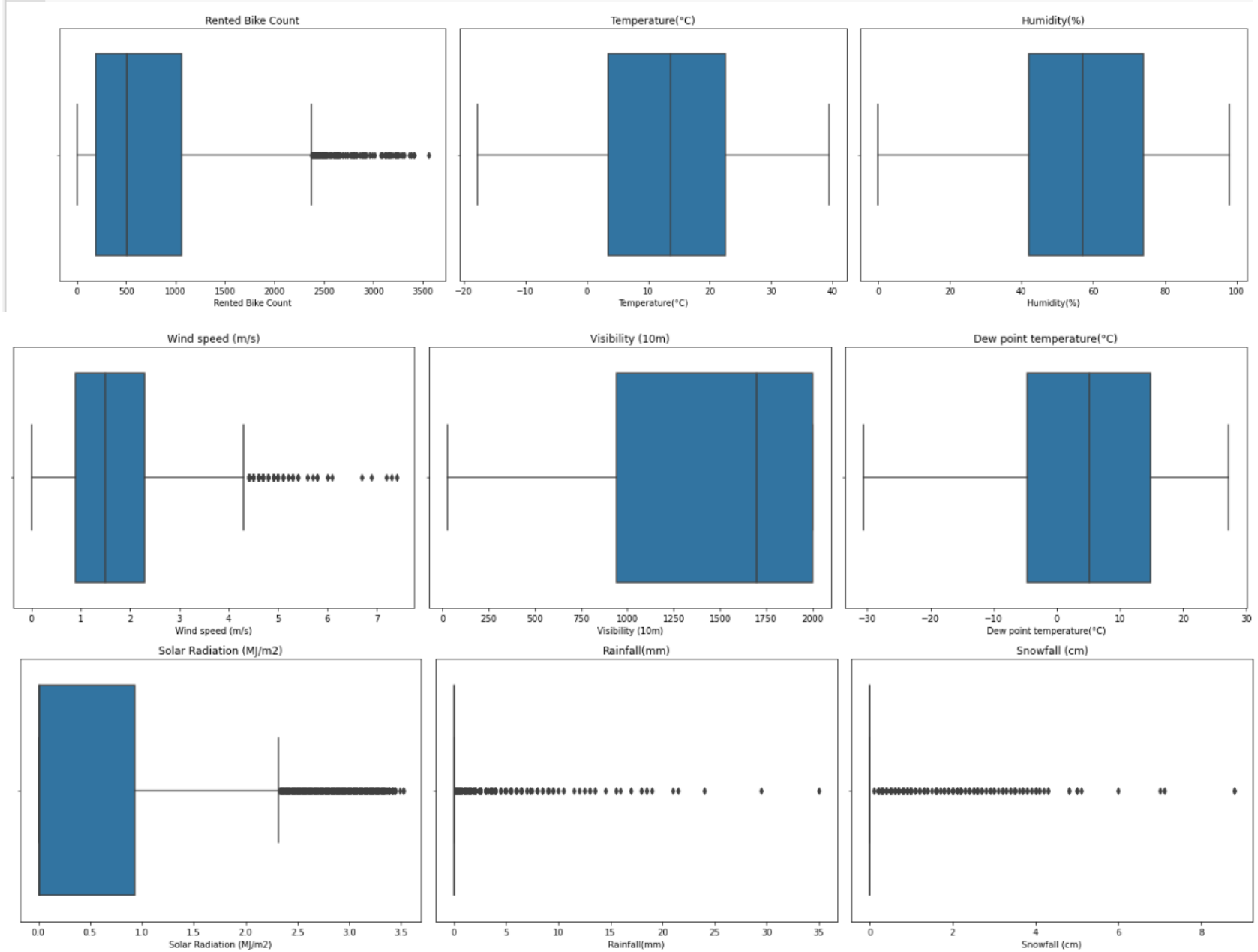
Data Wrangling and feature engineering

We had 8760 observations and 14 features.
Categorical Features: Seasons, Holiday and Functioning day.
Numerical Columns:
Date, Hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar radiation, Rainfall,Snowfall, Rented Bike count .

Rename Columns: Columns were renamed because they had units mentioned in brackets and was difficult to copythe feature name while working.

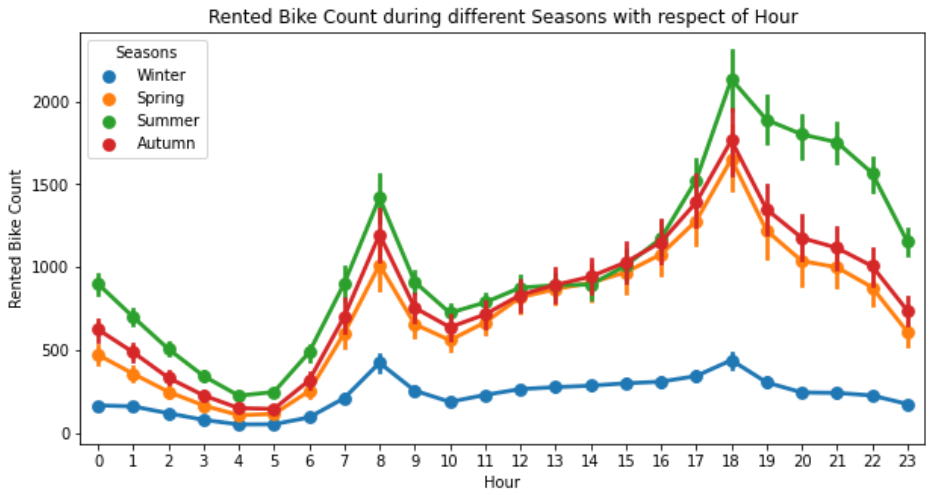
It zero null values in our dataset.
Zero Duplicate entries found.
The data type of Date column was changed from ‘object’ to ‘datetime’. This was done for feature engineering.
Two new columns were created with the help of Date column ‘Month’ and ‘Day’ which was further used for EDA and later we dropped Date column.

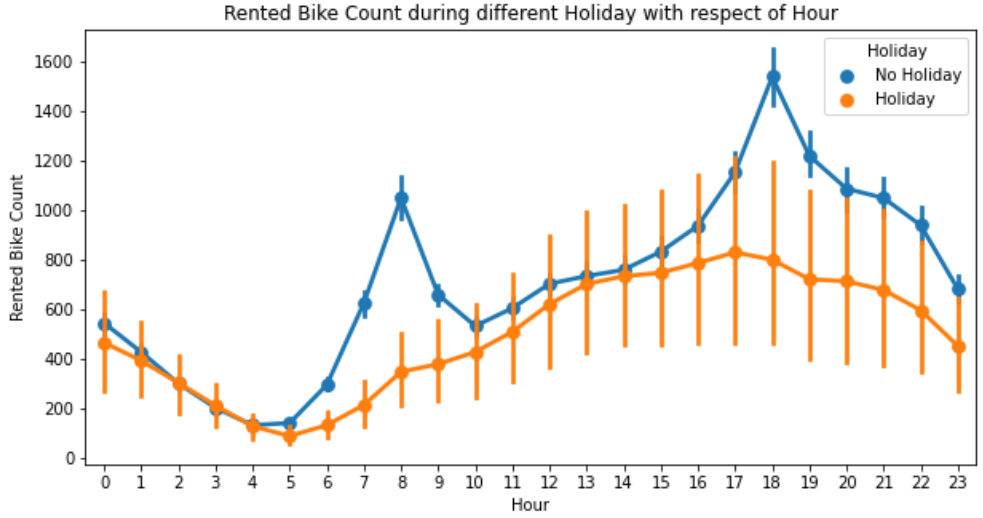
Outliers - Boxplot



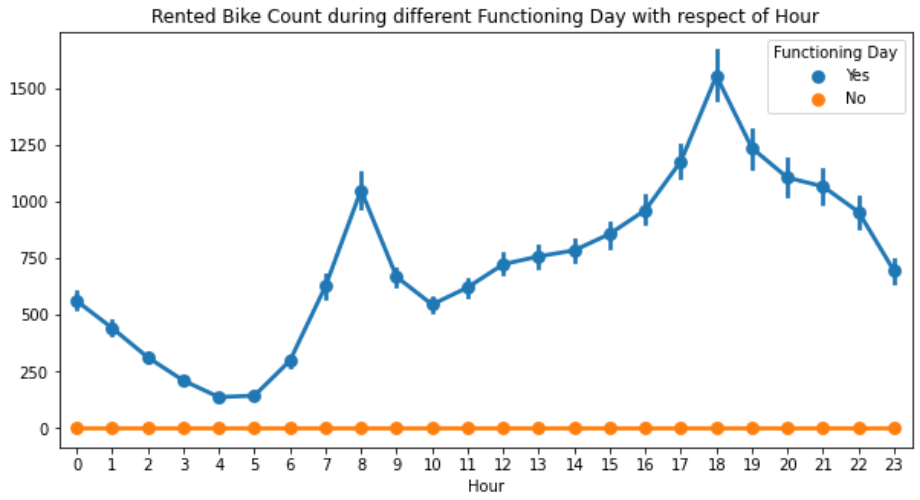
Exploring categorical feature

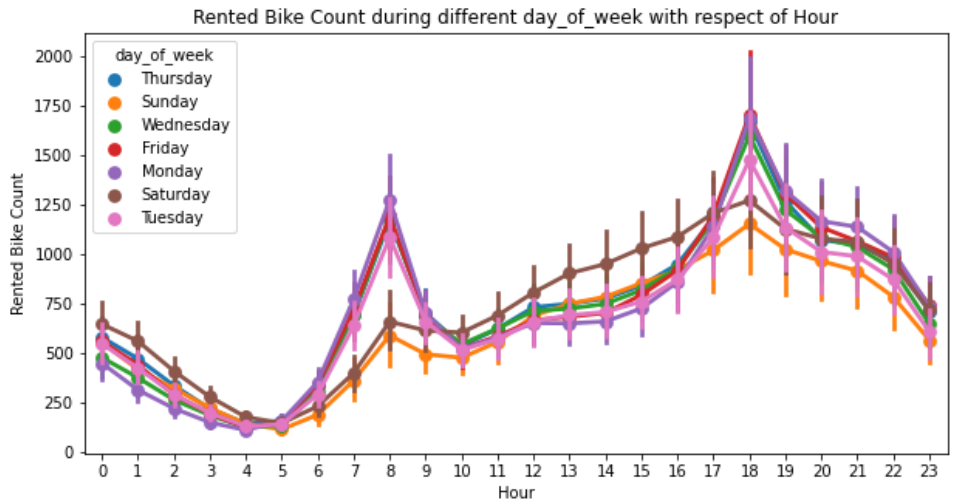
Point plots with Rented Bike Count during different categorical features with respect of Hour



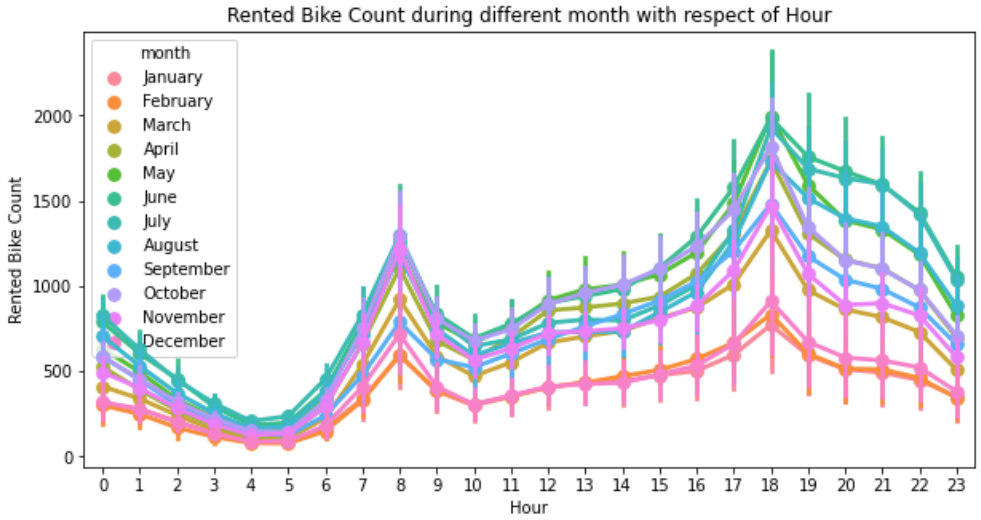


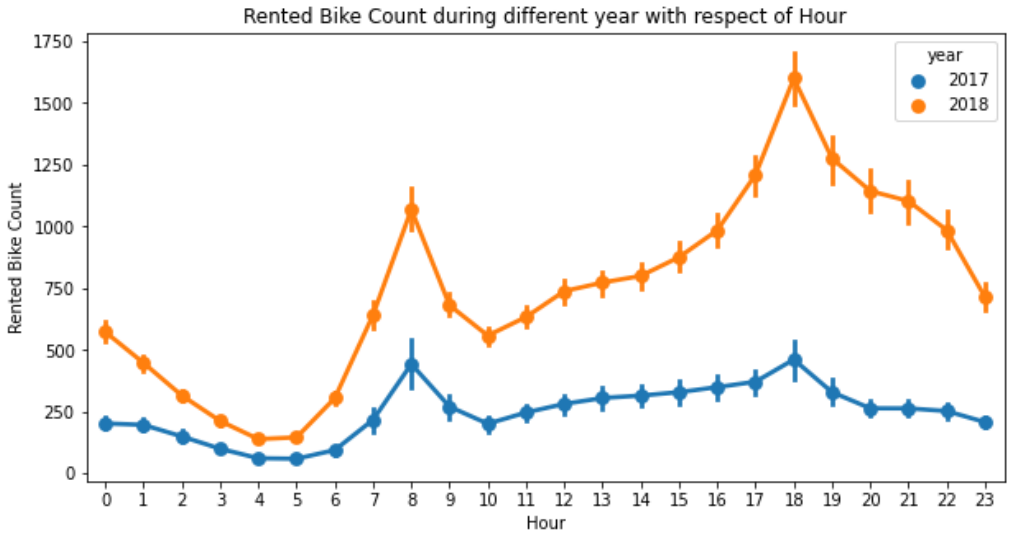
Season -In the season column, it is observed that the demand is low in the winter season.
Holiday-In the Holiday column, demand is low during holidays but strong during non-holiday times. This might be because many people commute to work on bicycles.





Functioning Day: If there is no entry in the Functioning Day column, then there is no demand.
Days of the week- In the Days of the week column , it is seen that the demand patterns on weekends are distinct from those on weekdays because they peak in the afternoon.



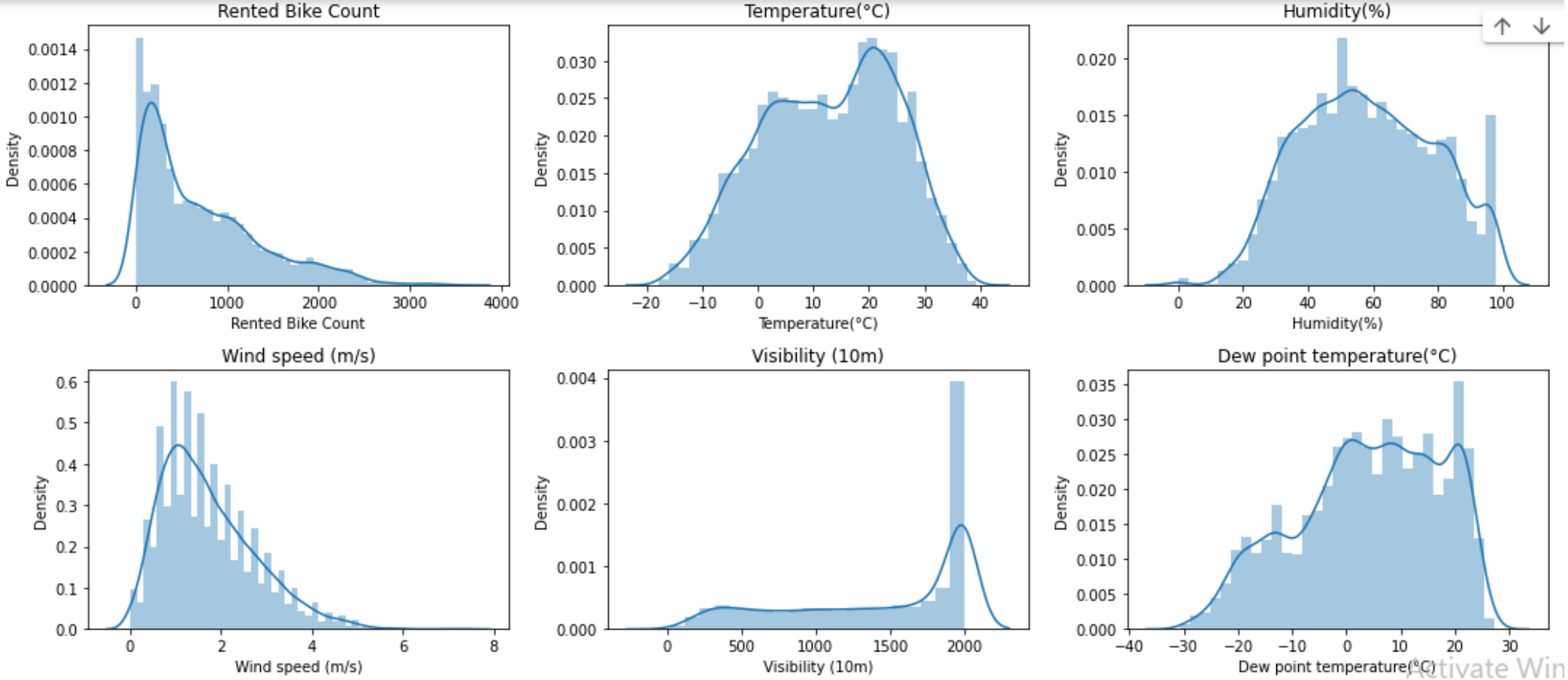


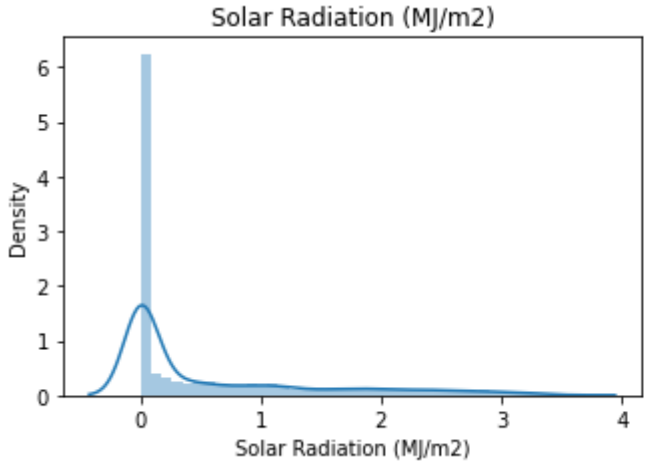
Month - In the month column, demand is lowest in December, January, and February because these are chilly months, as we already saw in the season column.

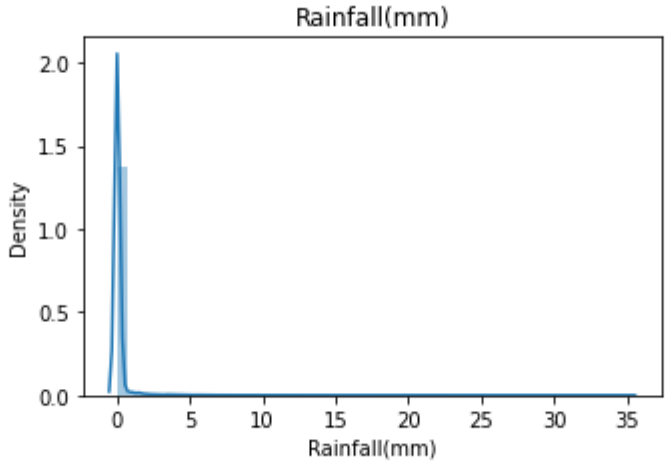
Year -Because it was new in 2017 and fewer people were aware of it, the demand was lower in 2017 and increased in 2018

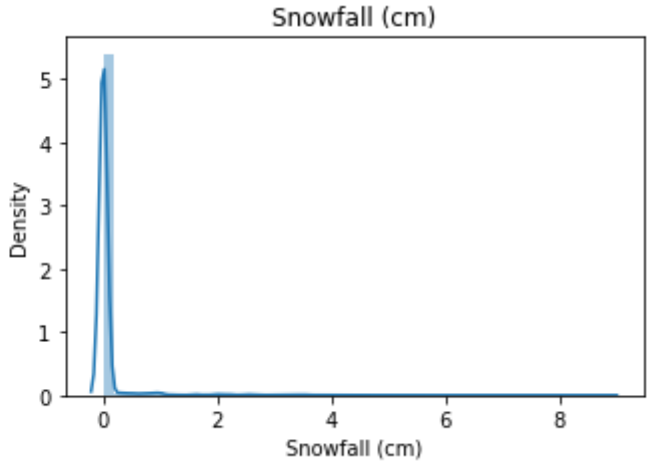
Now we can clearly see the pattern which shows that the demand is high in the afternoon on the weekend. While there is more demand during office hours in weekdays

Data Distribution









In these plots it is observed that some of the columns are right skewed and some are left skewed. Skewness is taken into consideration while applying algorithms.

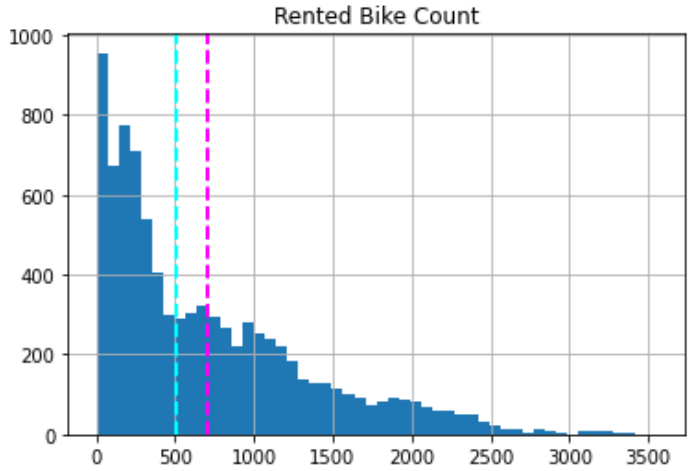
Right skewed columns:

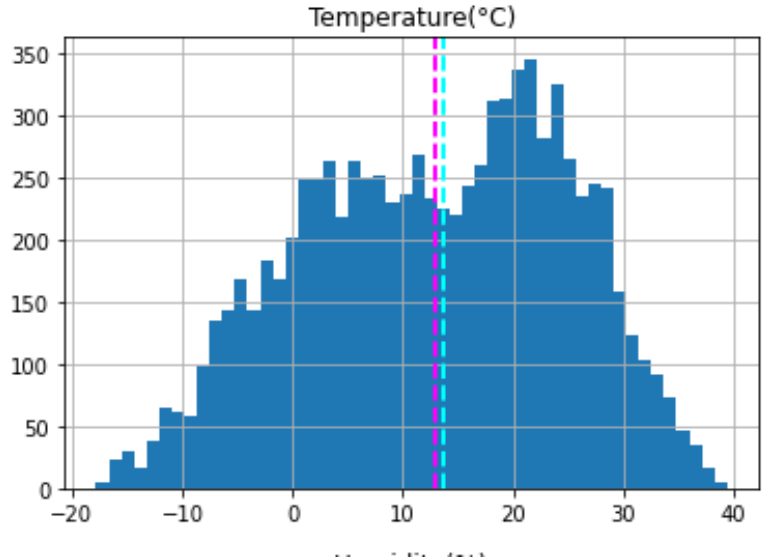
Rented Bike Count (Its also our Dependent variable), Wind speed (m/s), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm),

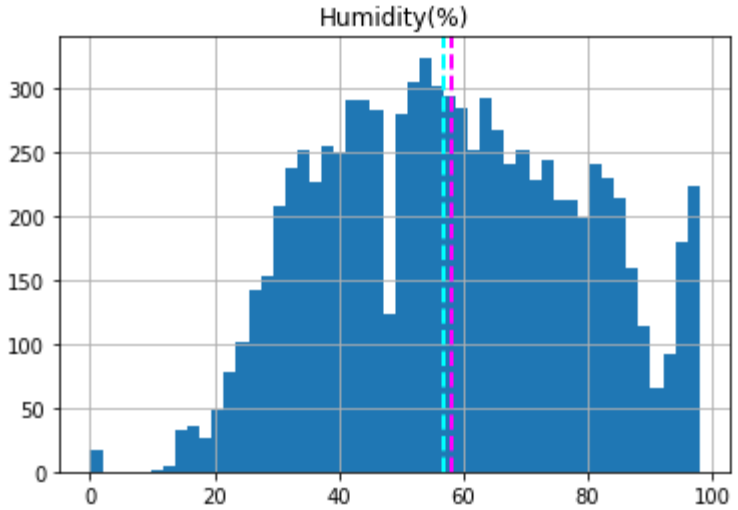
Left skewed columns:

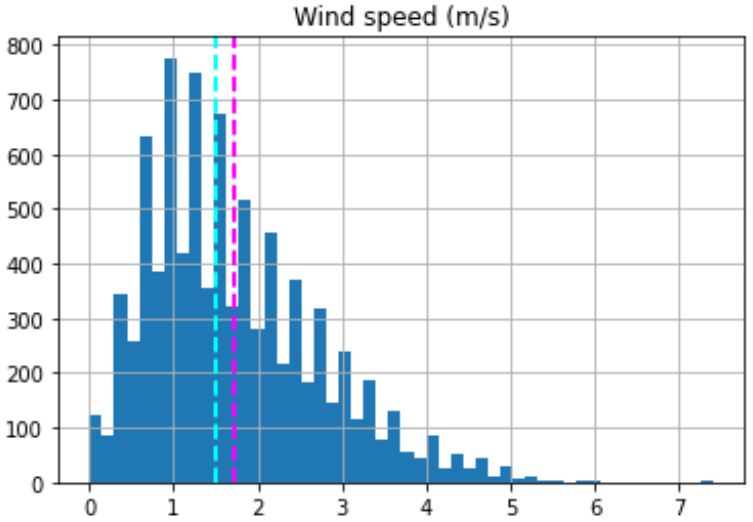
Visibility (10m), Dew point temperature(°C)

Histogram of numeric features

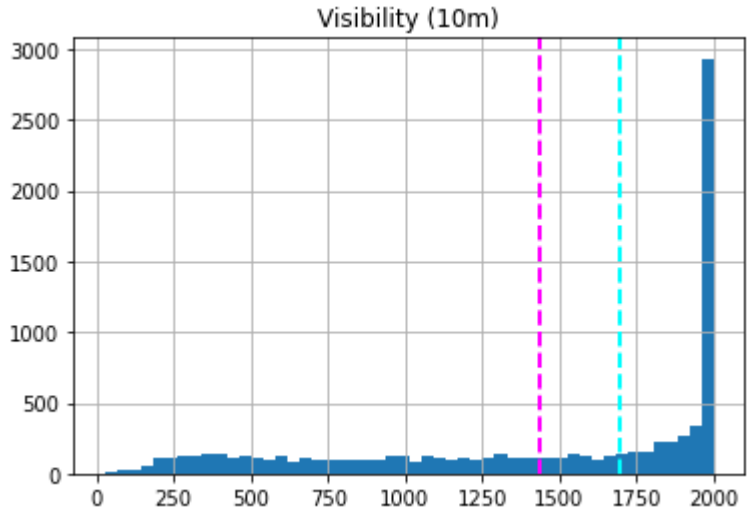


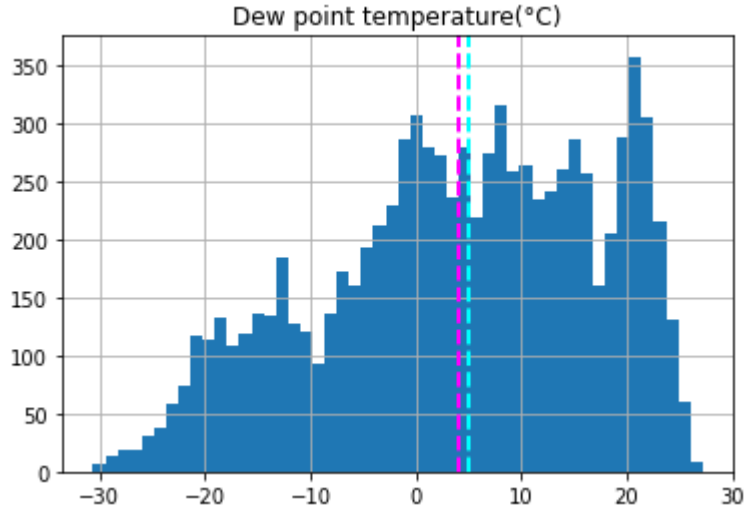


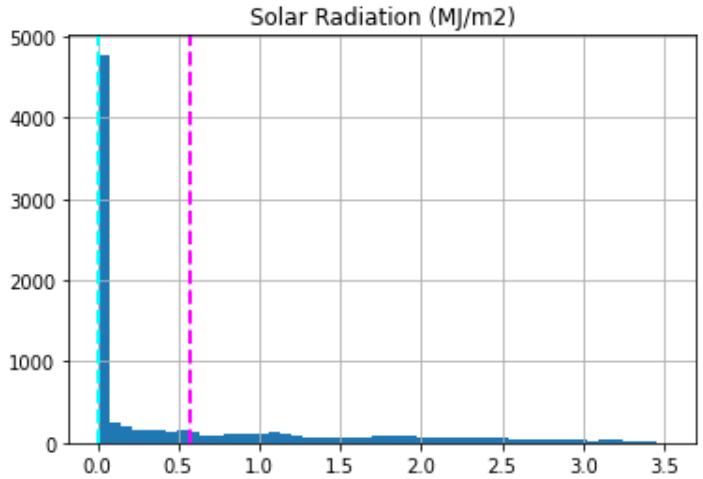


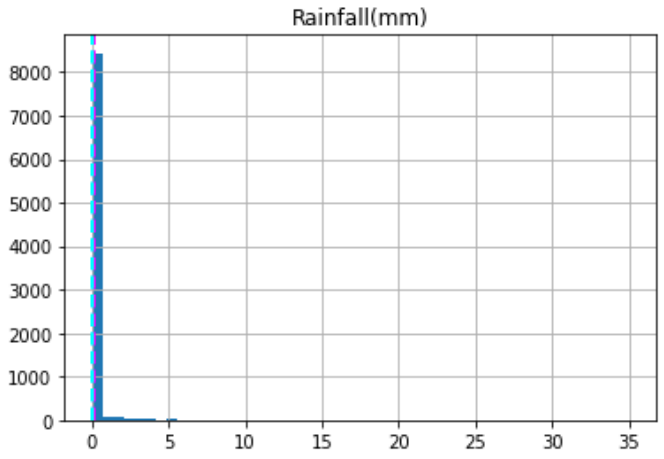


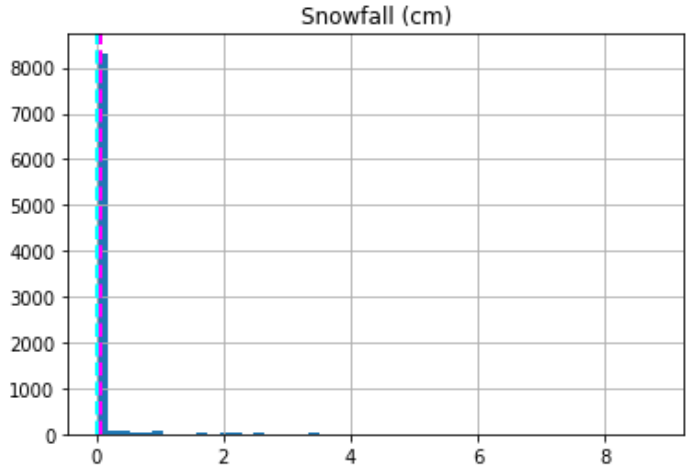
AlmaBetter


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


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


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

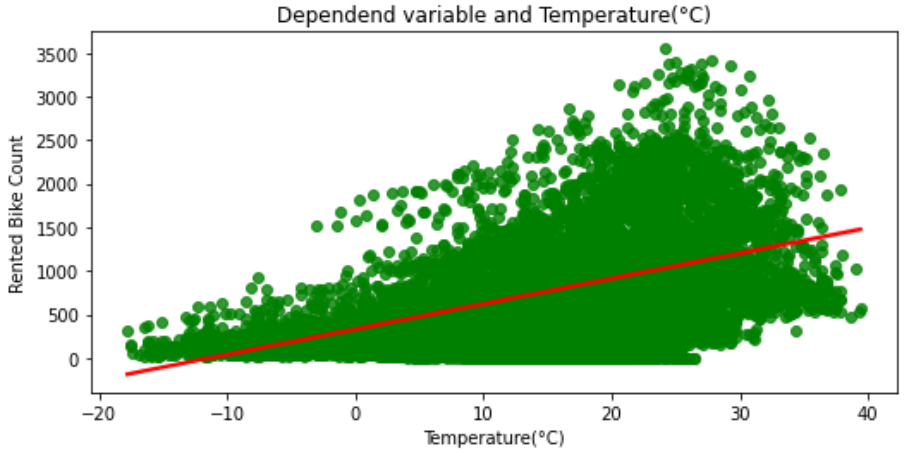

Rainfall(mm)
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
35

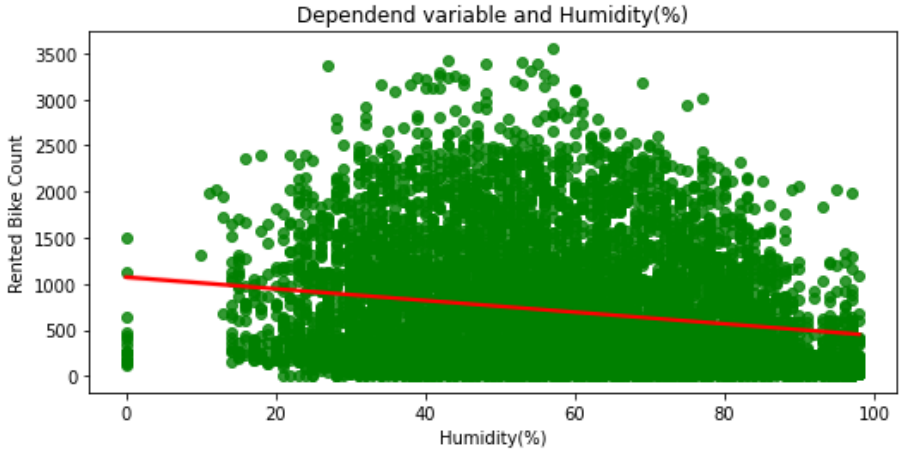

Snowfall (cm)
8000
7000
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8

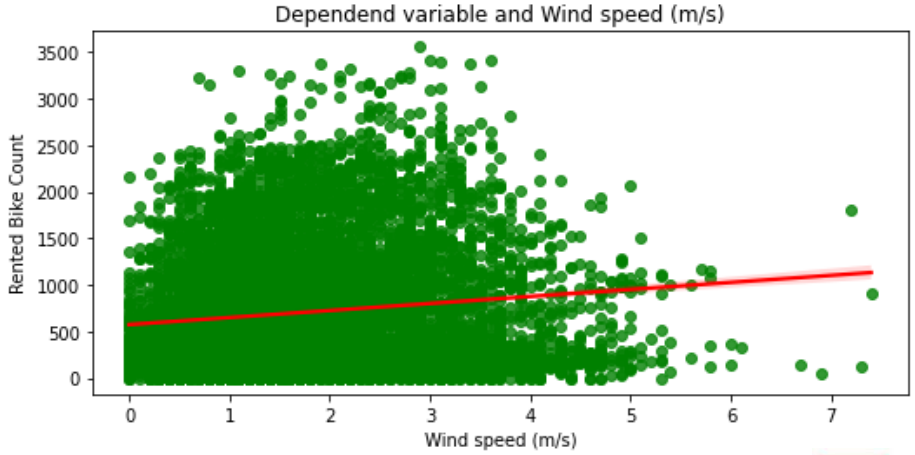
AlmaBetter

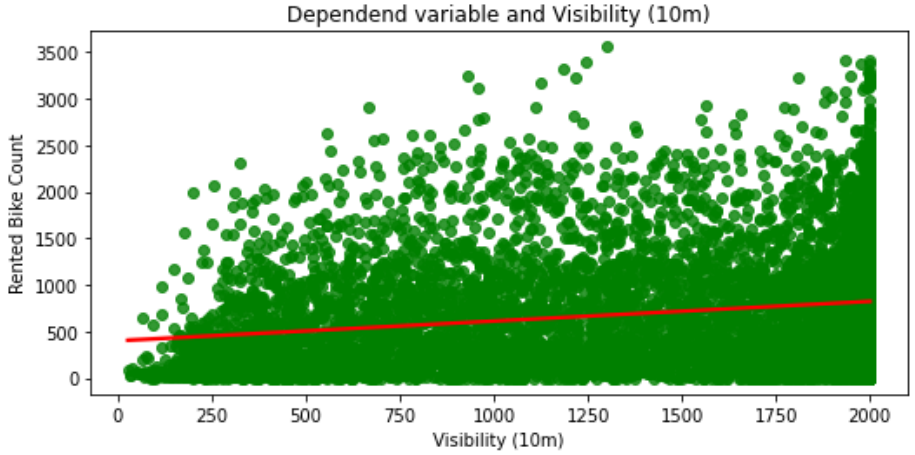
Relation of numerical features with dependent variable

Regression plot of dependent and independent variable

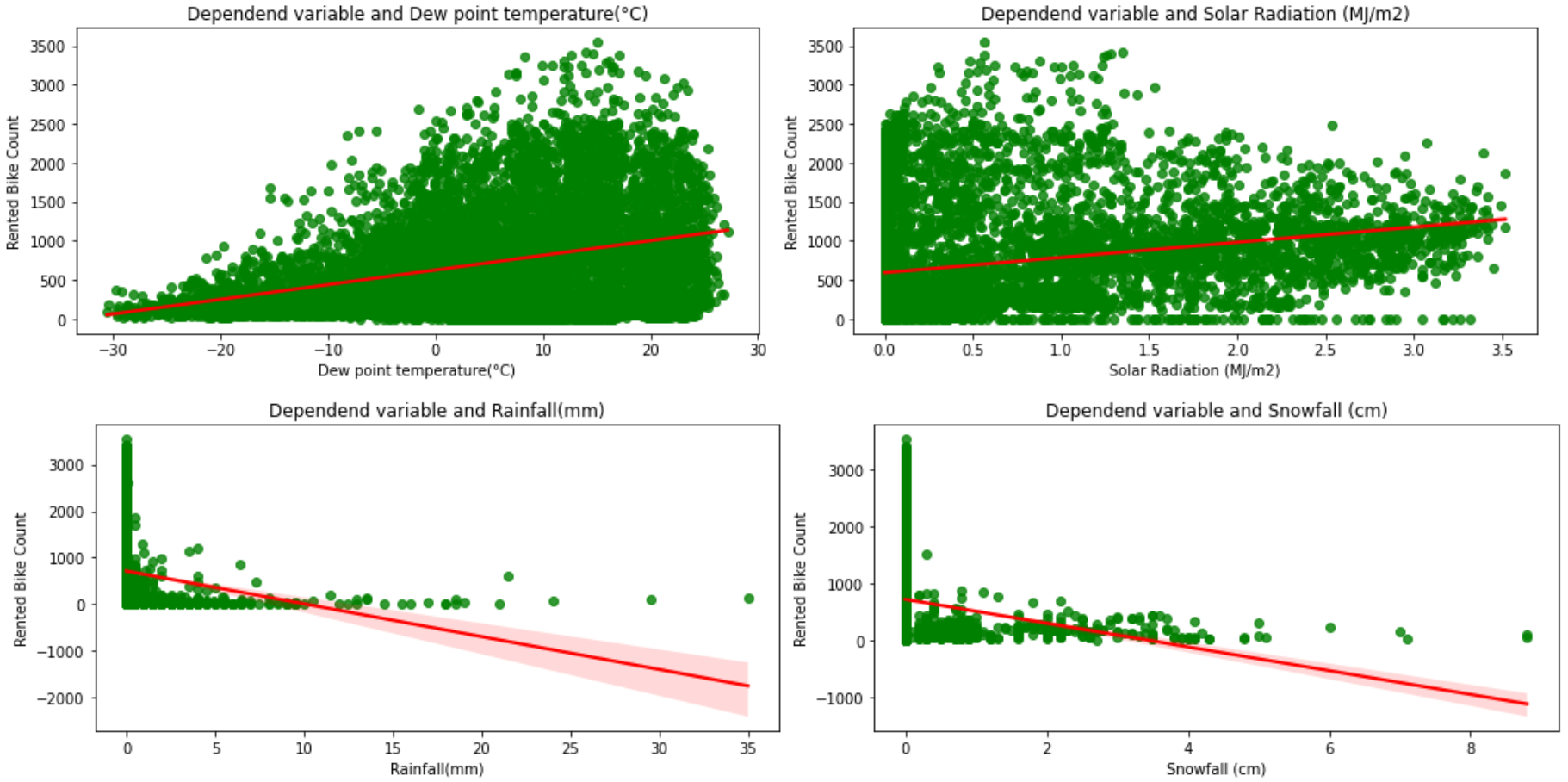






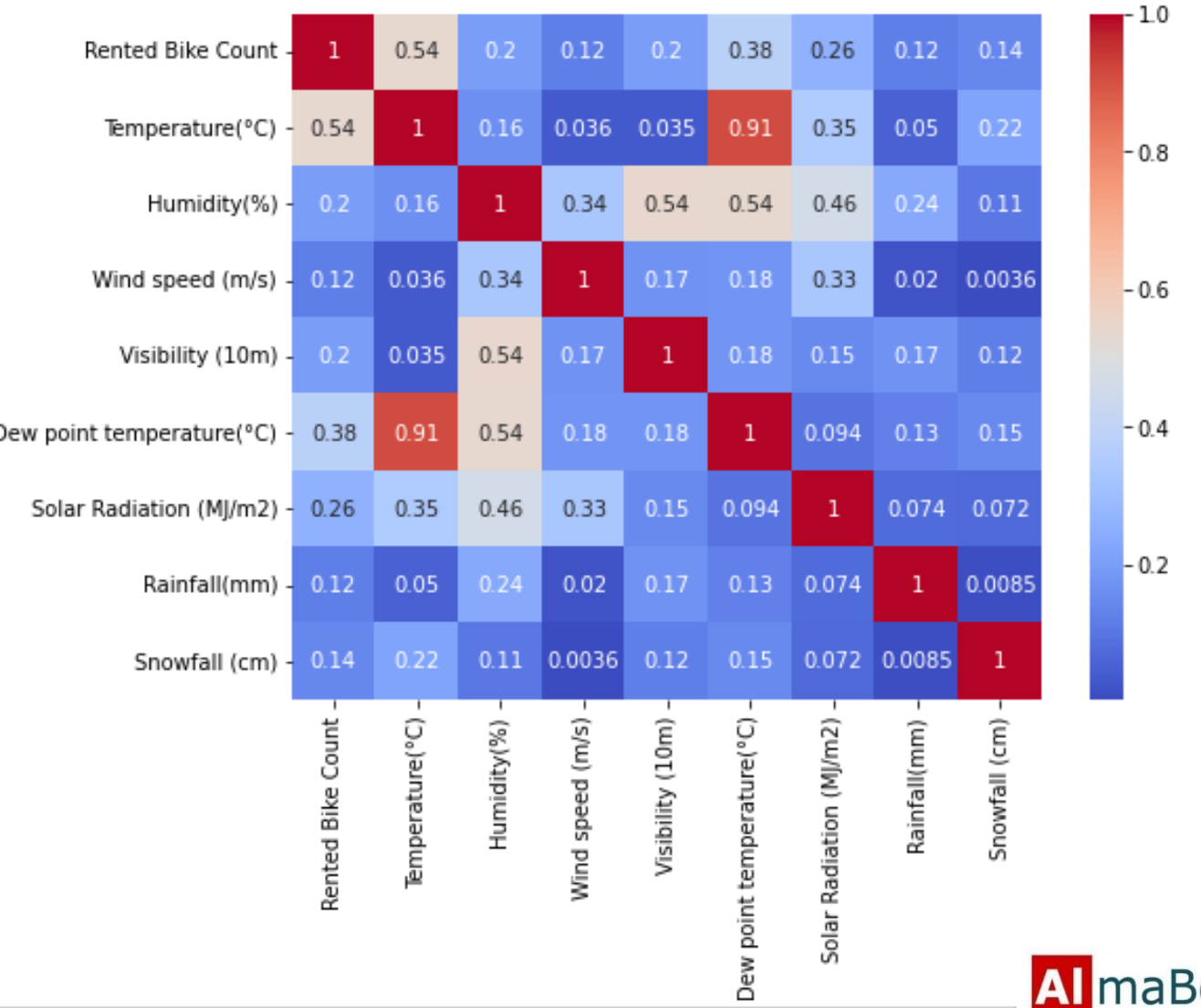


AlmaBetter



Plots shows that some of the features are positive linear and some are negative linear with respect to target variable. It is observed that the columns 'Temperature', 'Wind speed', 'Visibility', ‘Dew point temperature', 'Solar Radiation' are positively in relation to the target variable, which means the rented bike count increases with increase of these features.

Correlation of all the numerical features via heatmap



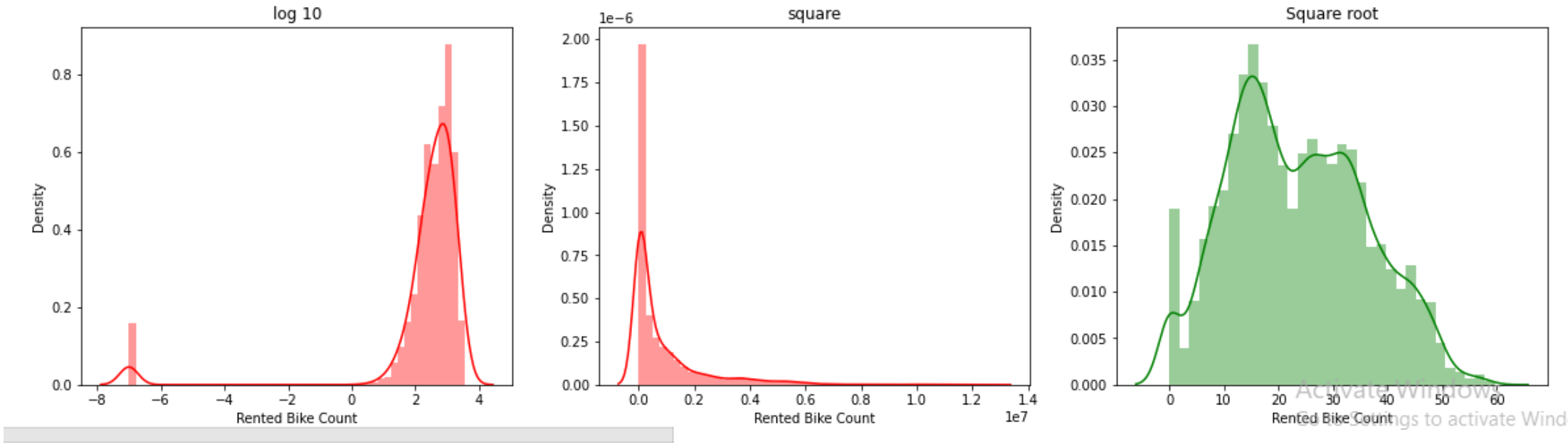
As we can see Temperature and Dew point temperature are 91 % correlated. So Dew point temperature was dropped because it has very low correlation with target variable as compared to temperature.

Preparation of data for model building

- With heatmap highly correlated variables were dropped. Thus data was prepared for model building.
- VIF was obtained to measure the strength of the correlation between the independent variables in regression analysis.

	variables	VIF
0	Temperature(°C)	3.166007
1	Humidity(%)	4.758651
2	Wind speed (m/s)	4.079926
3	Visibility (10m)	4.409448
4	Solar Radiation (MJ/m2)	2.246238
5	Rainfall(mm)	1.078501
6	Snowfall (cm)	1.118901

Normalization can be seen to some extent with square root on dependent variable



Square root transformation for normalization was selected to reduce heteroscedasticity of the residuals in linear regression.

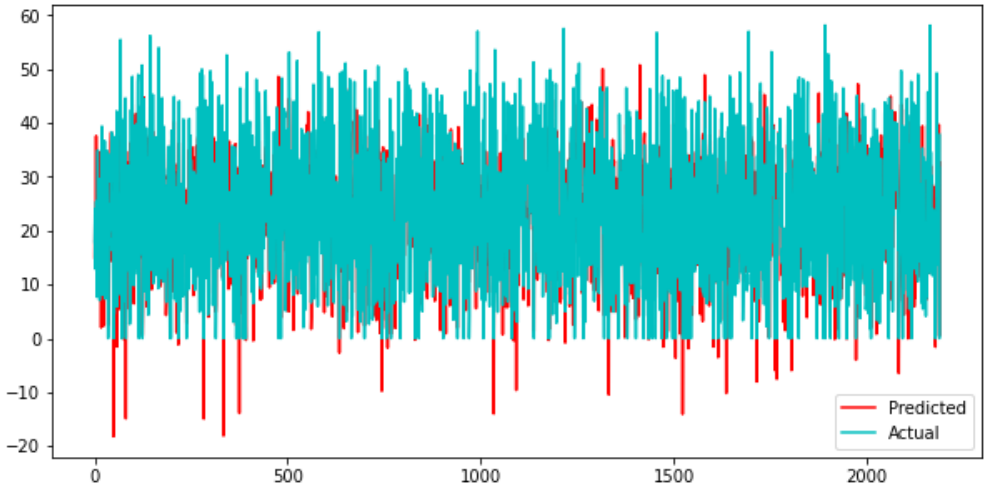
Model selection

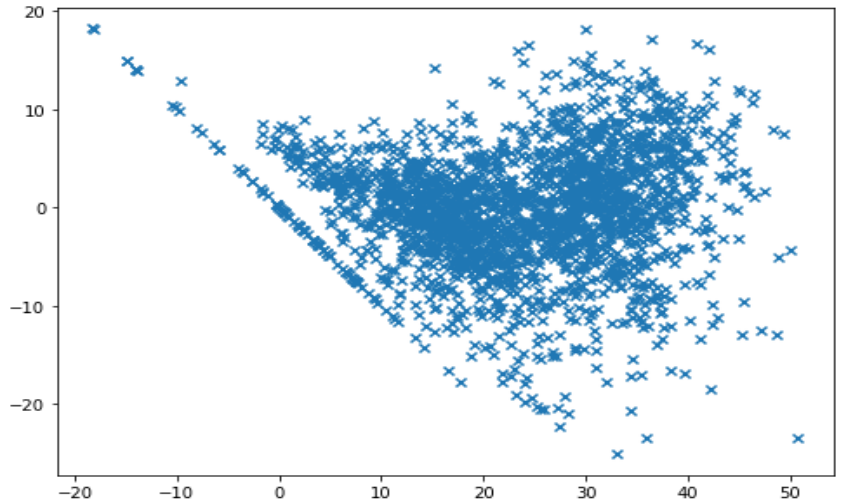
Regression models used

1. Linear Regression
2. Ridge Regression(regularized regression)
3. Lasso Regression (regularized regression)
4. ElasticNet Regression
5. Decision Tree Regression.
6. Random forest Regression

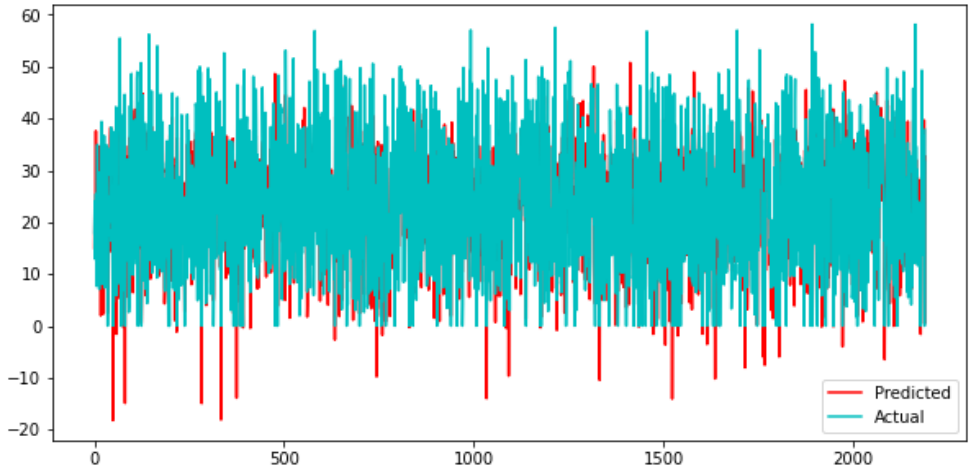
Before and after applying these models,regression assumptions were checked by scatter plot of actual and predicted values, removing multicollinearity among independent variables

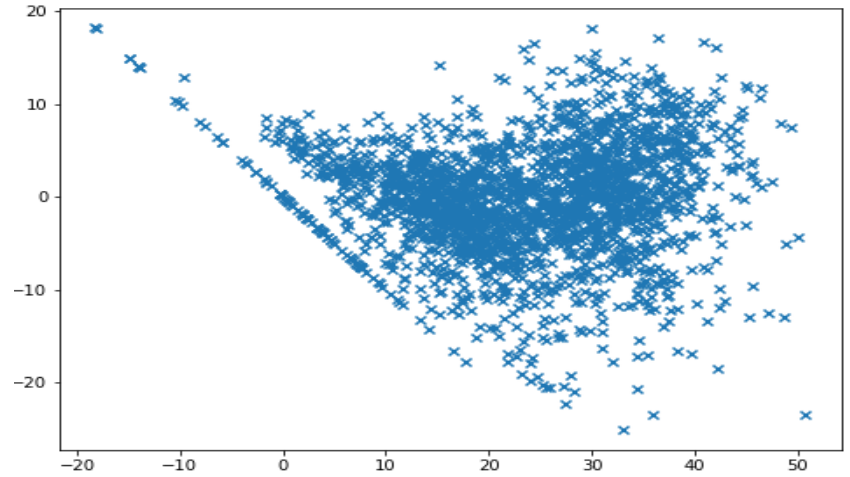
1.Linear Regression





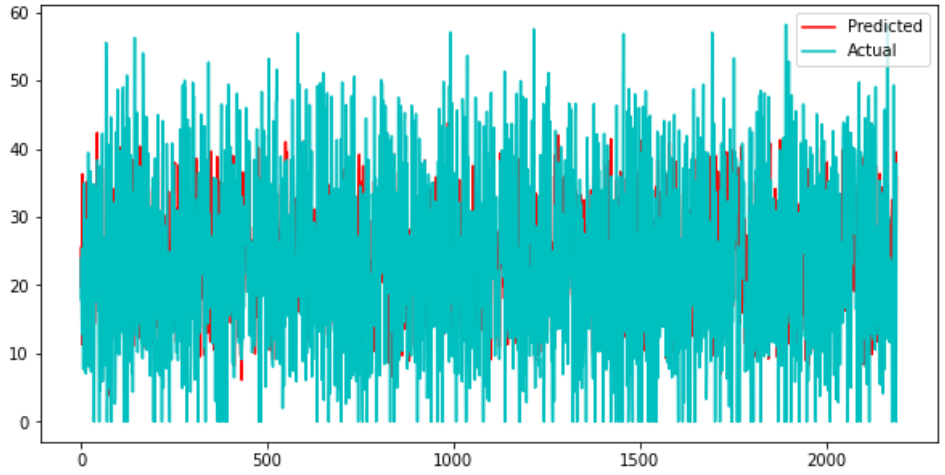
2.Ridge Regression

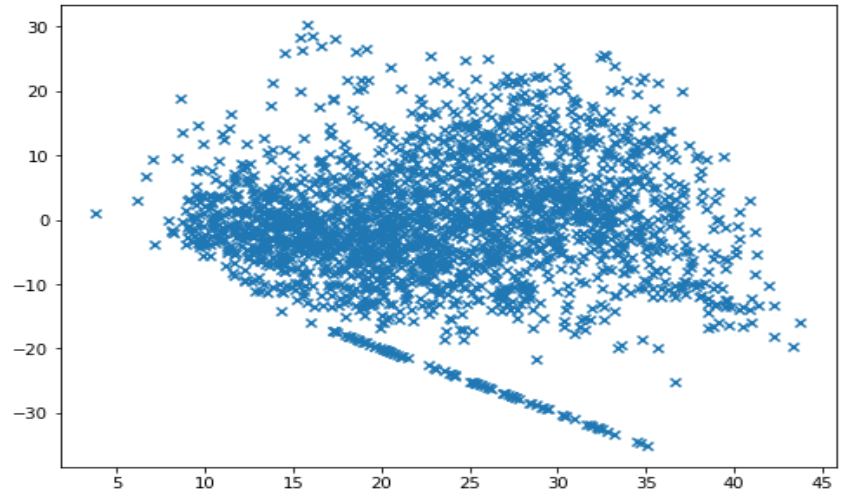




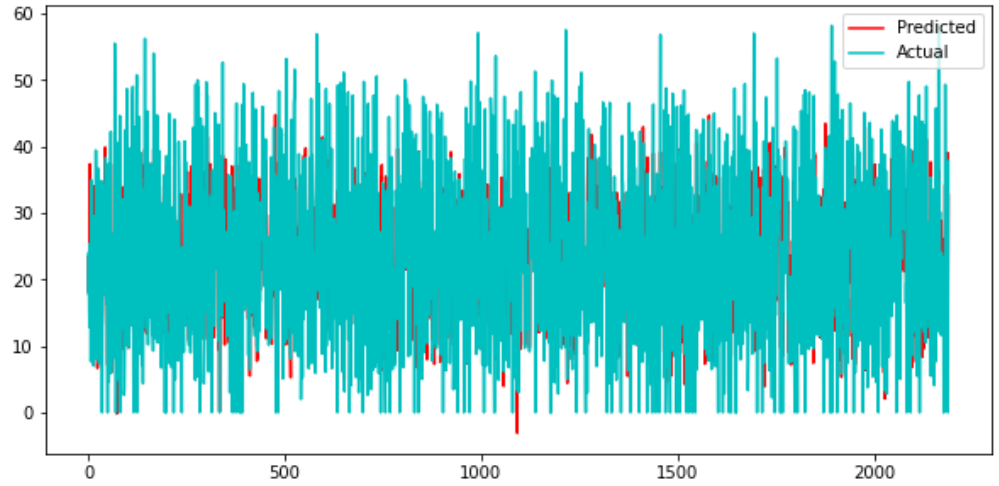
AlmaBetter

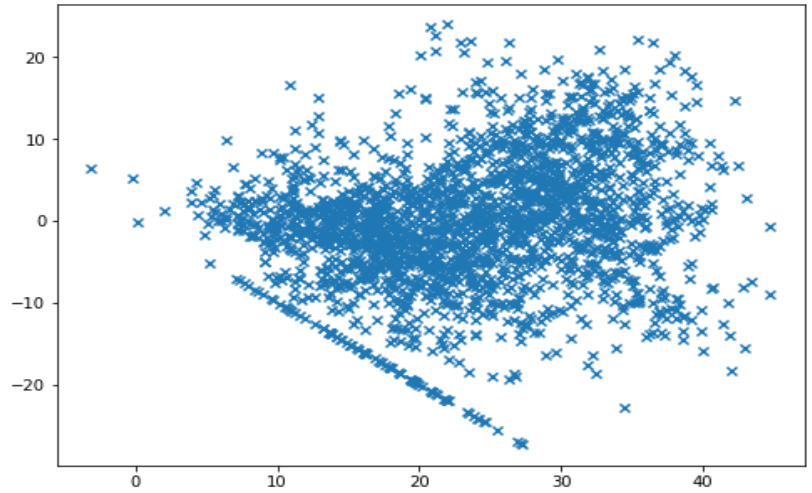
3.Lasso Regression



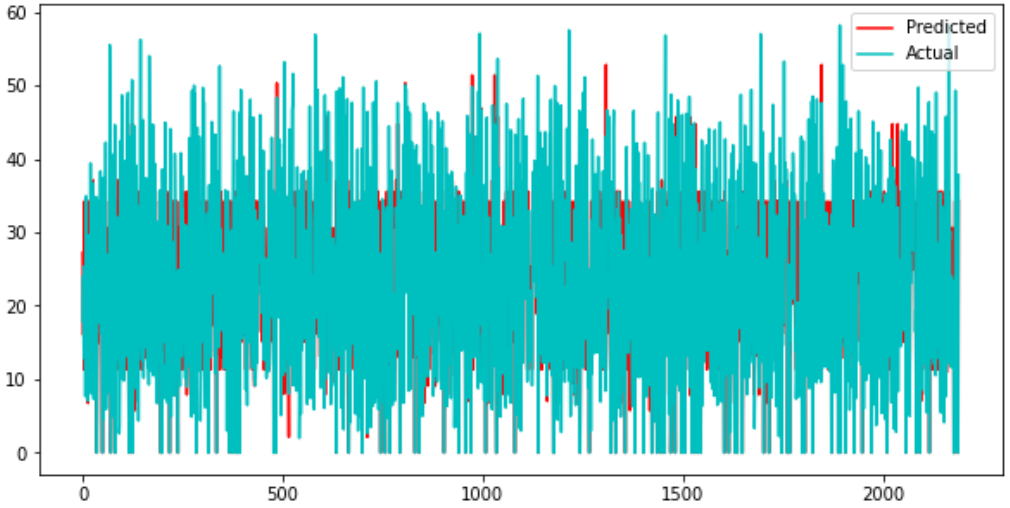


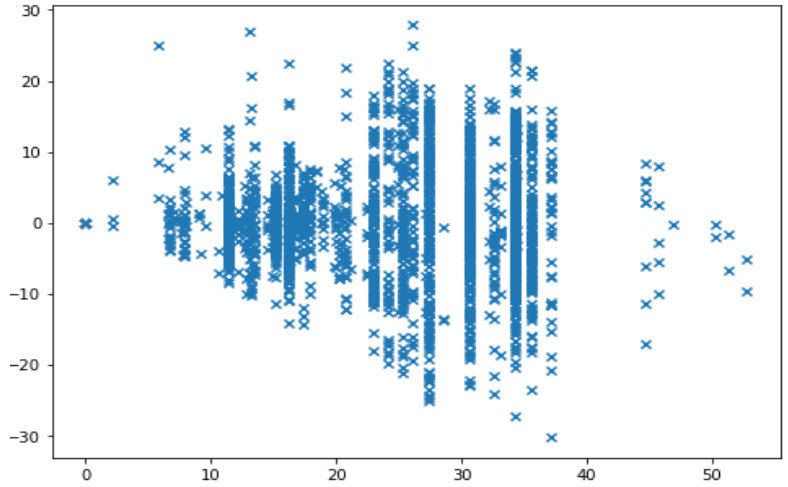
4.ElasticNet Regression

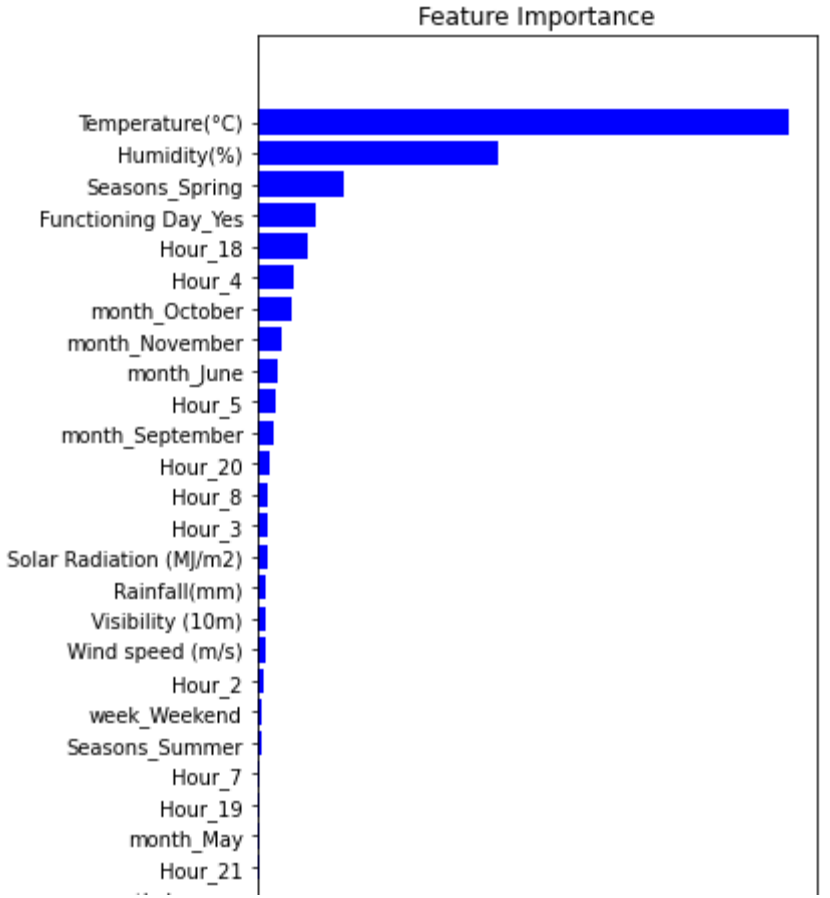




5.Decision Tree

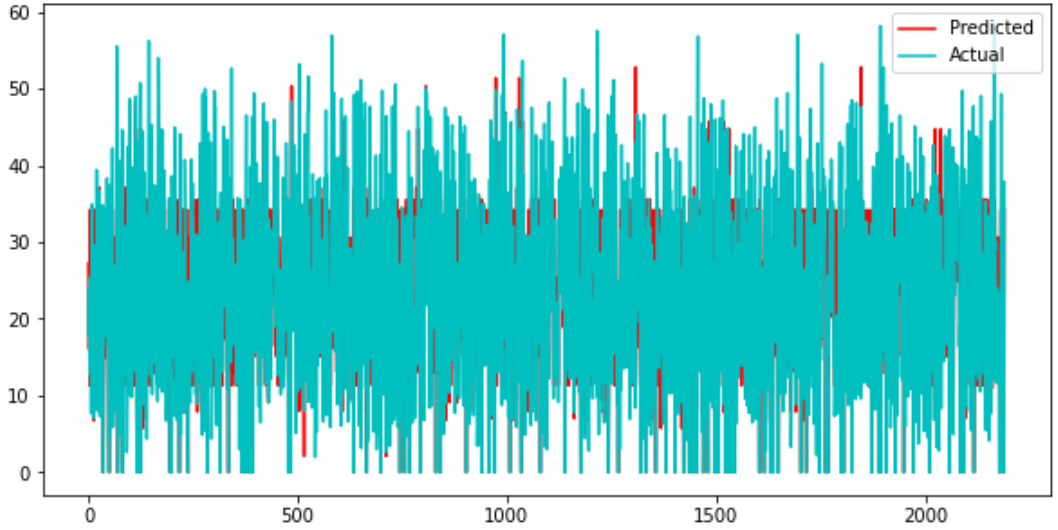


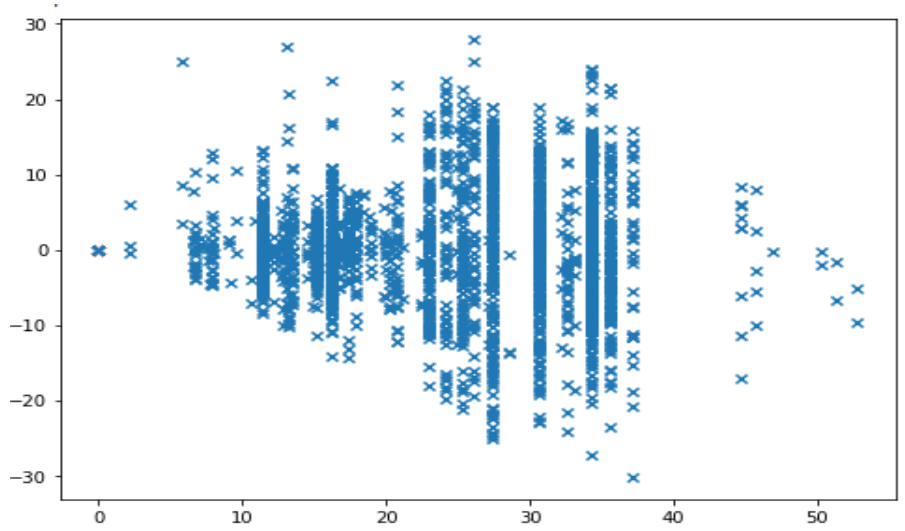


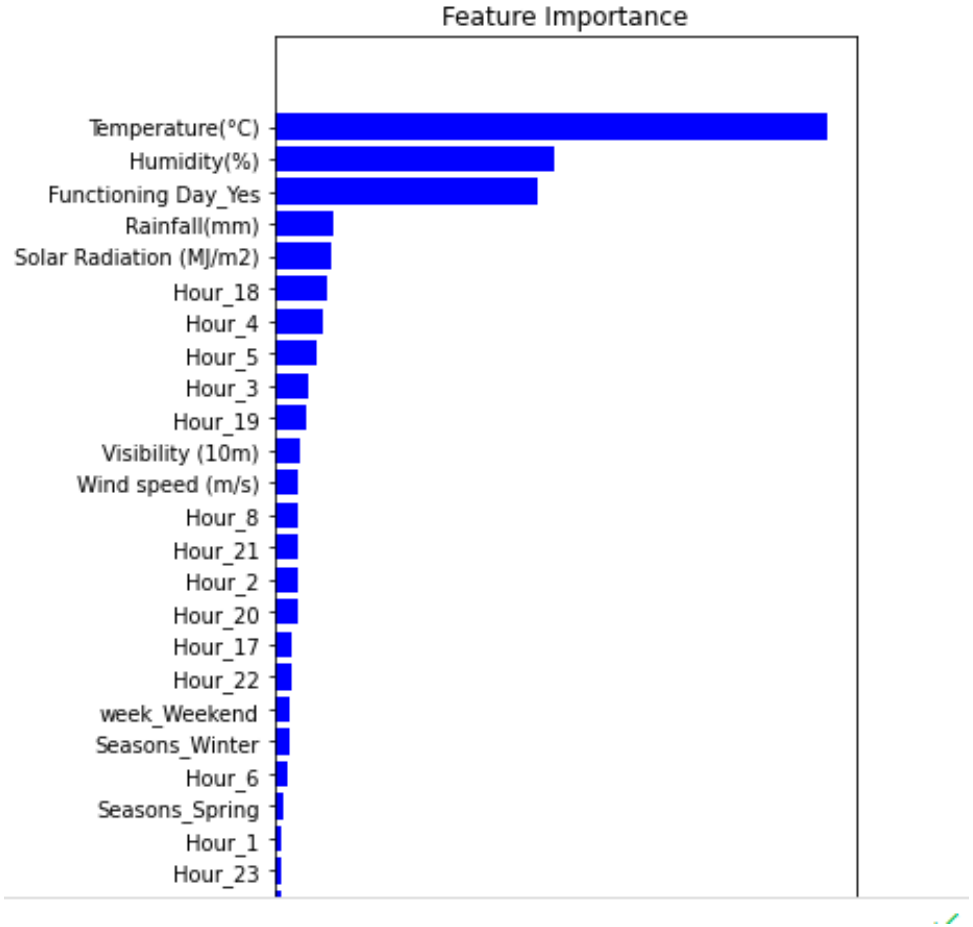


AlmaBetter

6.Random Forest







AImaBetter

Conclusion

		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	**0**	Linear regression	4.641	37.405	6.116	0.757	0.75
	1	Ridge regression	4.641	37.405	6.116	0.757	0.75
	2	Lasso regression	7.255	91.594	9.570	0.405	0.39
	3	Elasticnet regression	5.890	59.232	7.696	0.615	0.61
	4	Decision tree regression	5.760	61.997	7.874	0.597	0.59
	5	Random forest regression	0.941	2.038	1.428	0.987	0.99
Test set	**0**	Linear regression	4.638	36.419	6.035	0.769	0.76
	1	Ridge regression	4.638	36.421	6.035	0.769	0.76
	2	Lasso regression	7.456	96.775	9.837	0.387	0.37
	3	Elasticnet regression	6.009	61.634	7.851	0.610	0.60
	4	Decision tree regression	6.252	73.678	8.584	0.534	0.52
	5	Random forest regression	2.551	15.881	3.985	0.899	0.90

MAE,MSE,RMSE and R2 score for each model was calculated. Based on r2 score model performance and selection was done.

- Temperature holds as the most important feature for our selected Random Forest Regression model.
- Bike rental count is high during working days as compared to non working days.
- It is observed that people generally prefer bike during moderate to high temperature.
- Linear and Ridge: Linear and Ridge models have almost similar R2 scores(around 75%) on both training and test data.
- Lasso: Lasso proved to be the weakest model of all with R2 value of 0.405
- ElasticNet: The value of R2 for this model was 0.615
- Decision Tree Regression : On Decision tree regressor model, R2 value is 0.597 on training data and on test data it was very less with 0.534 as its R2 value.
- Random Forest: On Random Forest regressor model, R2 value is 0.987 on training data and around 90% on test data.
- The method that proved to be most effective for predicting the hourly demand for rental bikes was Random Forest. The result was an Adjusted R2 value of 0.99.The most significant factors in determining the hourly demand for rental bikes was found to be temperature.

Thank You

AlmaBetter