Capstone Project - II

Supervised ML- Regression

Bike Sharing Demand Prediction

BY

Yashwant B. Raul
(yashwantraul24@gmail.com)
Mayur S. Marathe
(marathemayu1990@gmail.com
Sanket Gawali
(sanketgawali23@gmail.com)

Content

- Project Overview
- Feature Summary
- Feature Analysis
- Exploratory Data Analysis
- Implementing Algorithm
- Conclusion

Project Overview

The contents of the data came from a city called Seoul. It is the capital city of South Korea and has a population of around 9.7 million people. It was the 4th largest metropolitan economy in 2014. It has a humid continental climate influenced by monsoons.

Quick transportation is a big need in most cities. Ola & Uber are providing good transportation but nowadays the rates are getting high over this the rented bike is a good option for transportation which is quick & cheap. The aim of this project is to check the availability of the bike for a particular time in the day

Feature 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.

Attribute Information:

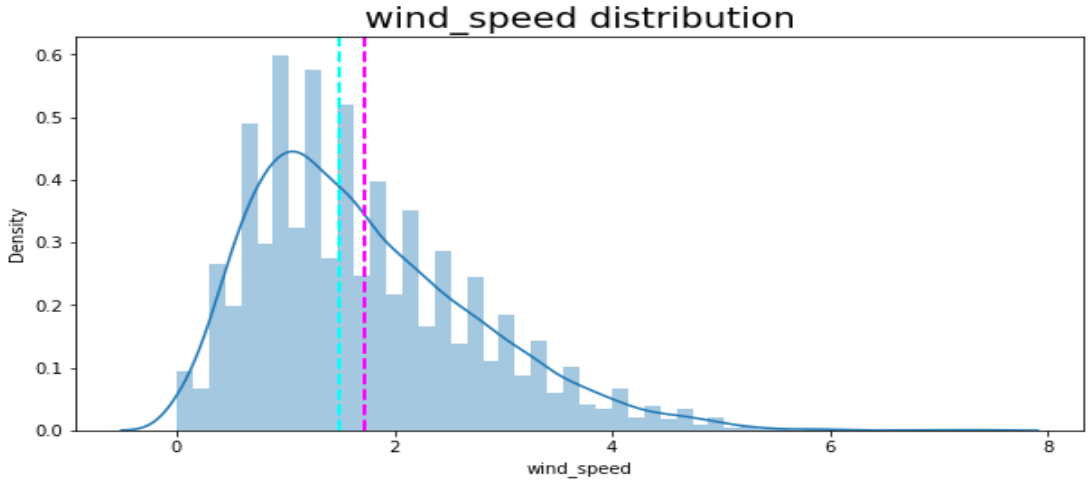
- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - Humidity in the air in %
- Wind speed - Speed of the wind in m/s
- Visibility - Visibility in m, 10m
- Dew point temperature - Dew point temperature in Celsius
- Solar radiation - Energy radiated by Sun in MJ/m2
- Rainfall - Amount of raining in mm
- Snowfall - Amount of snowing in cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours

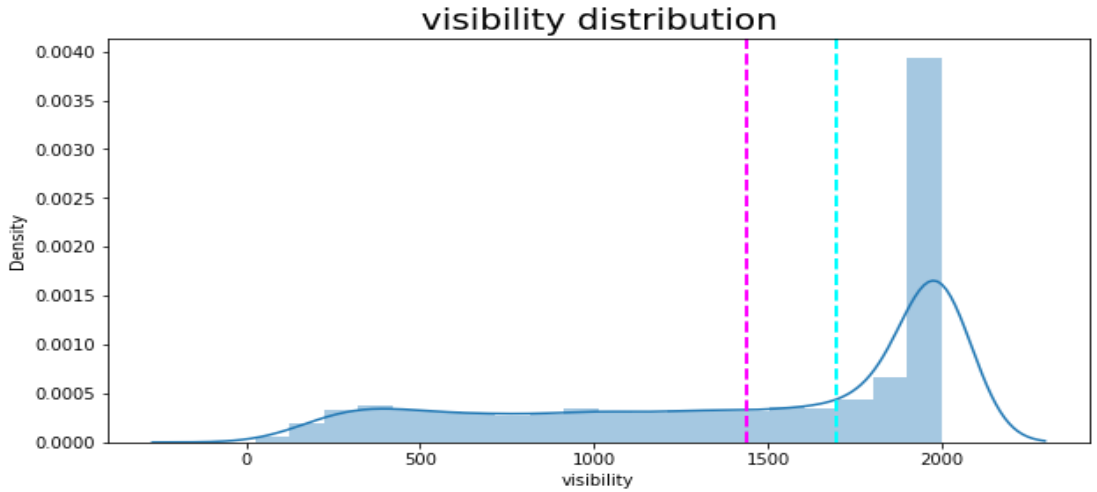
Insight Of Dataset

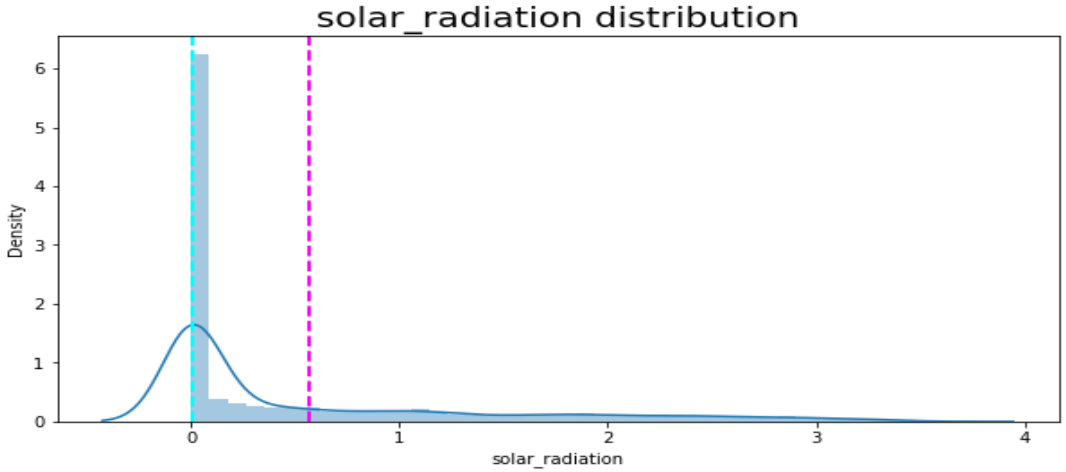
- There are No Missing Values present
- There are No Duplicate values present
- There are No null values.
- We have 'rented bike count' variable which we need to predict for new observations
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data
- We convert the "date" column into 3 different column i.e "year","month","day".
- 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'

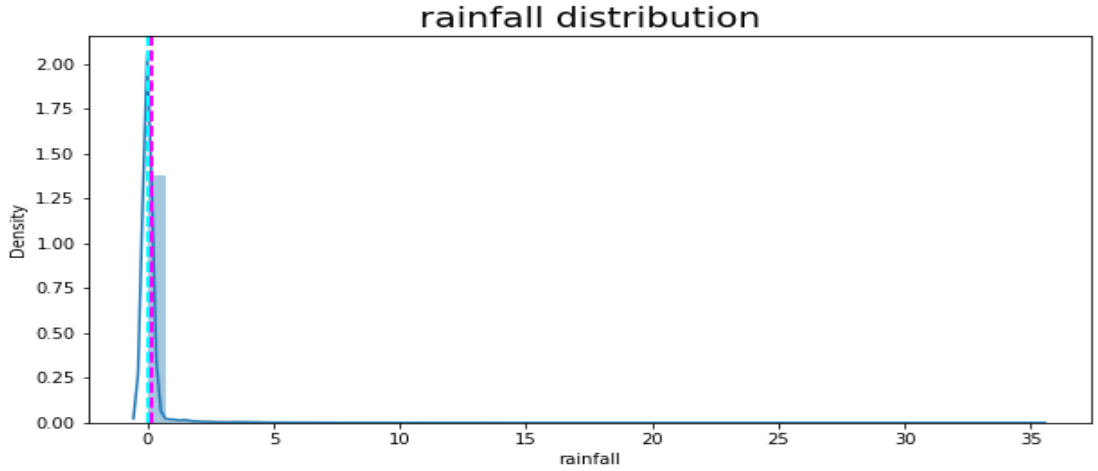
Exploratory Data Analysis

Distribution Of Data

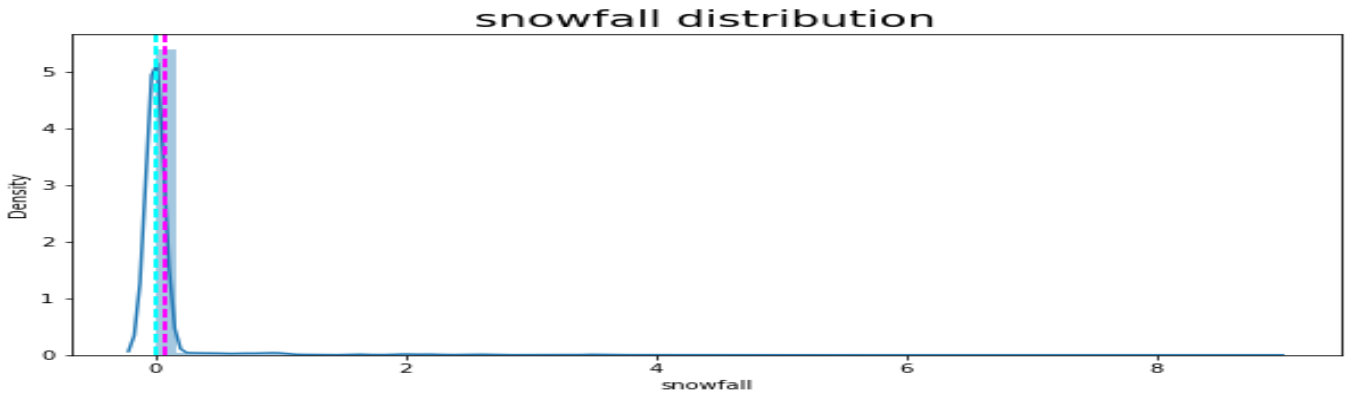


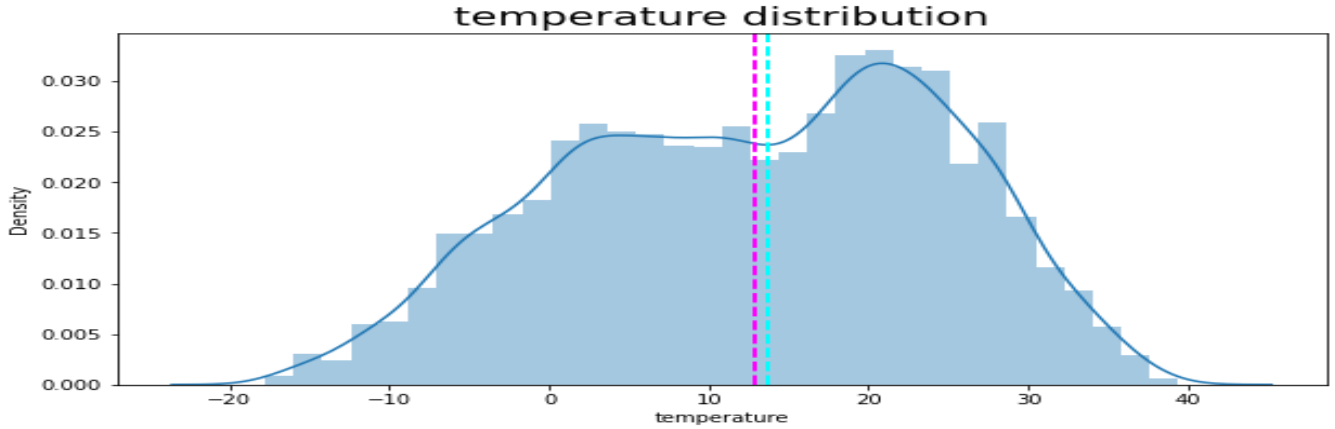


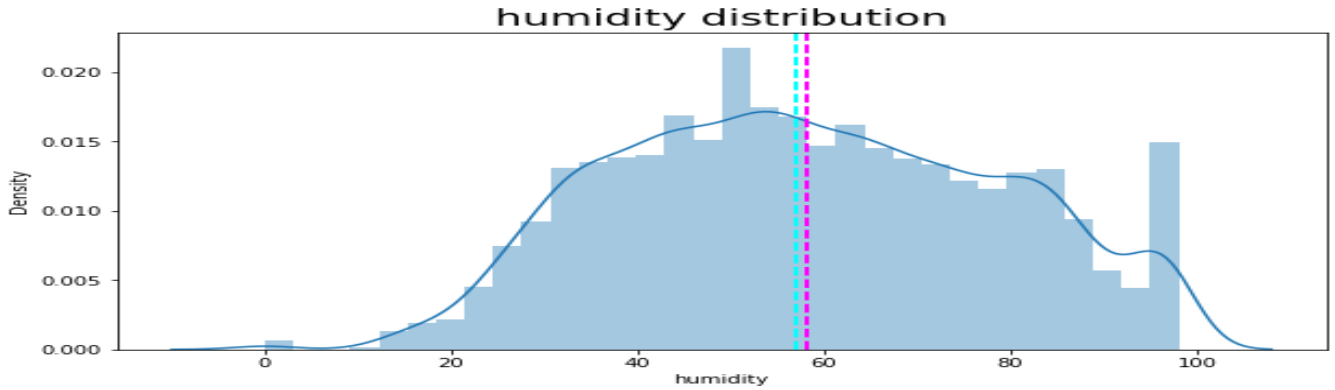




Distribution Of Data







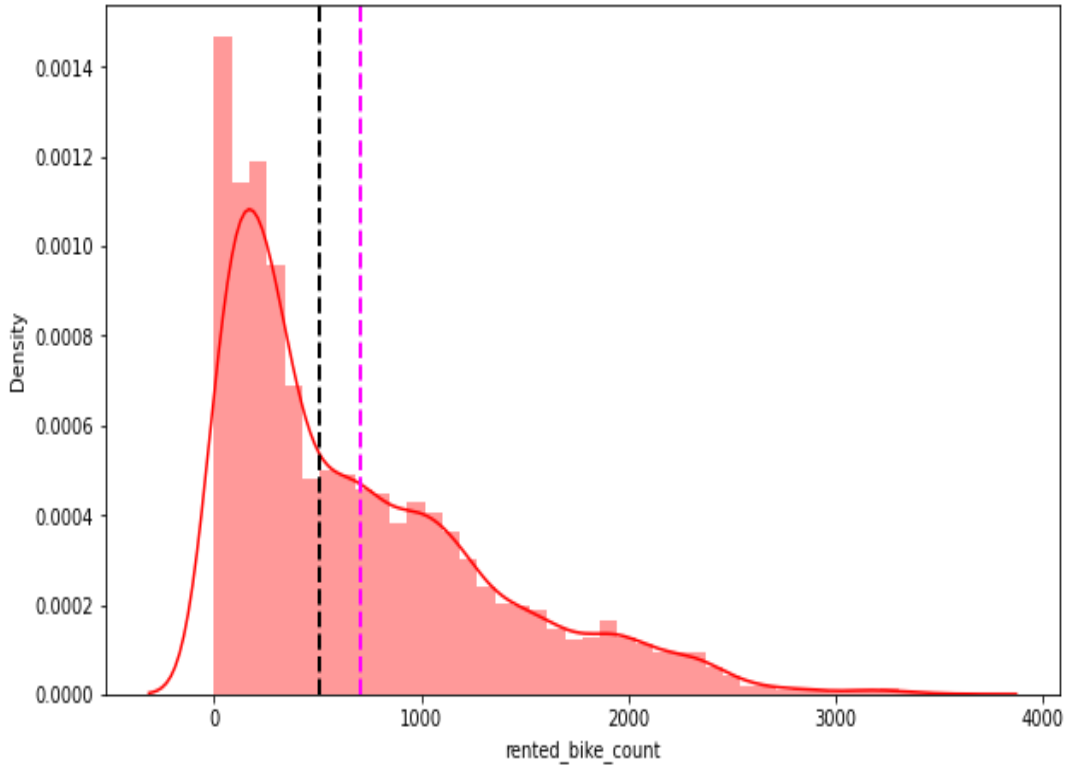
Normally distributed attributes: temperature, humidity.

Positively skewed attributes: wind, solar_radiation, snowfall, rainfall.

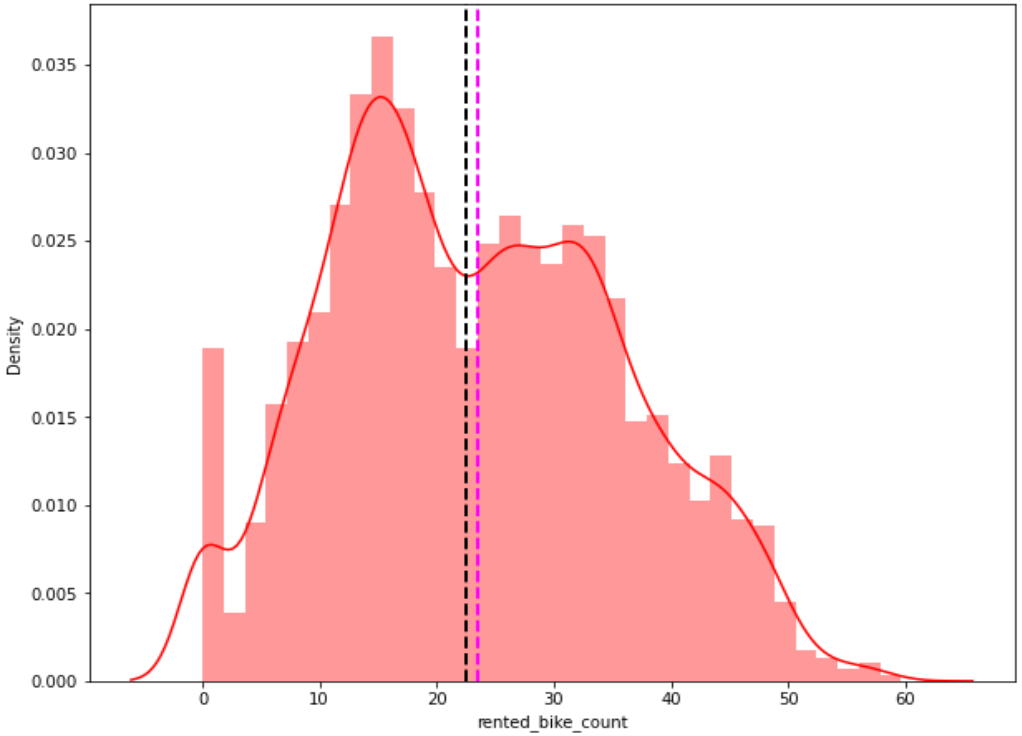
Negatively skewed attributes: visibility.

Analysis Of Dependent Variable: Rented_bike _count

Available distritribustion of data



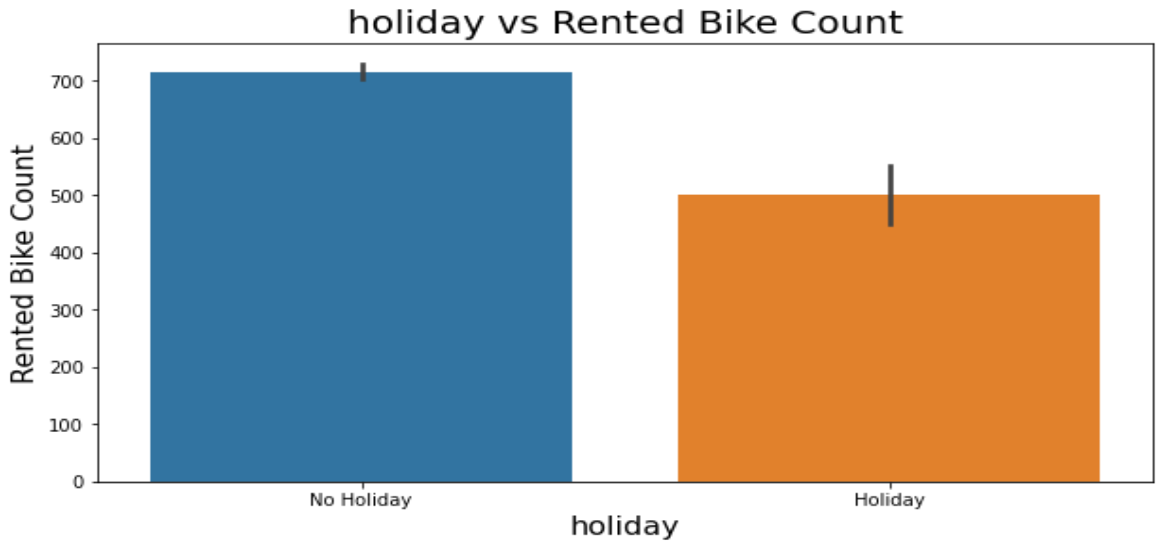
After Normalisation

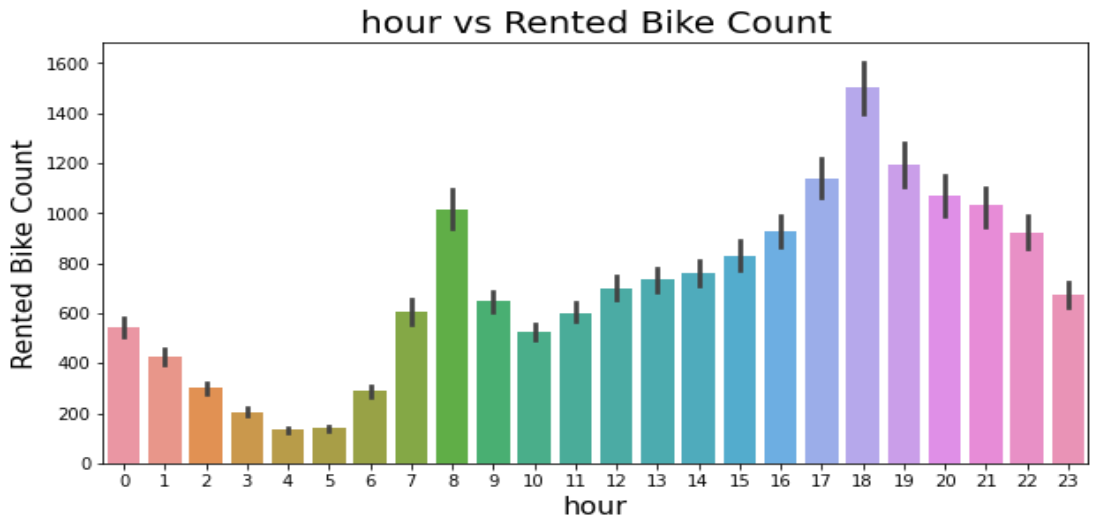


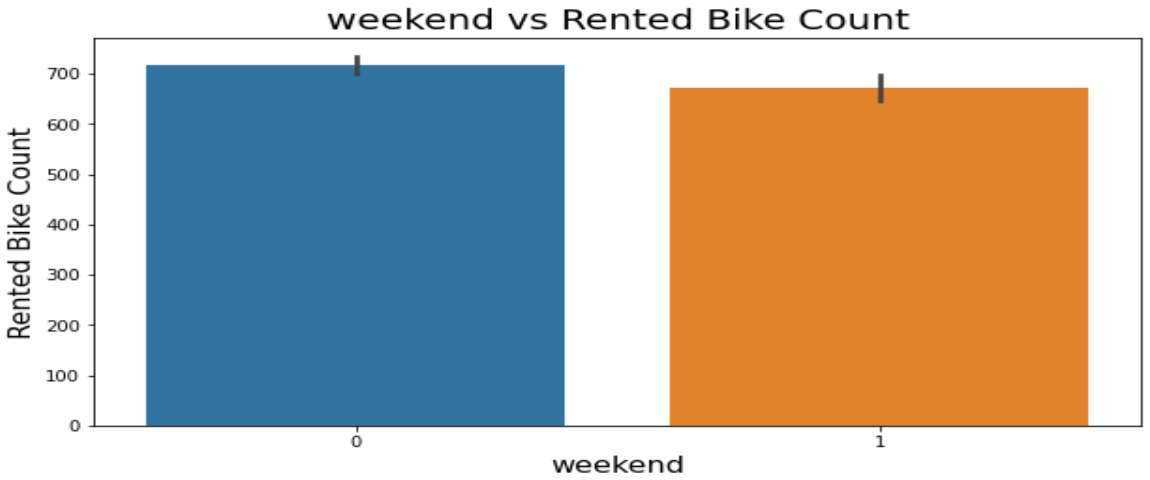
Above LHS graph shows that Rented Bike Count has moderate right skewness.
Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we perform Square root operation to make it normal.

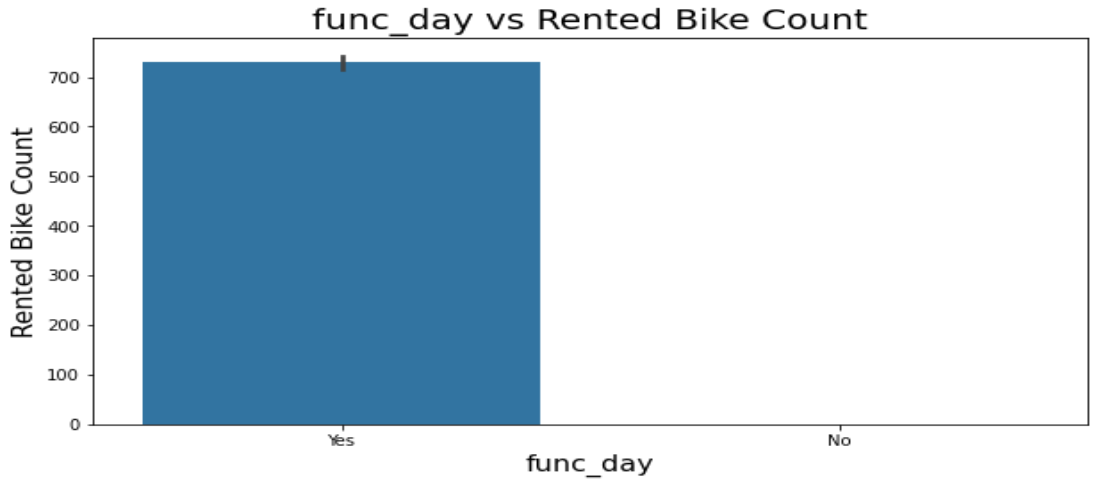
Analysis Of Categorical Variable

Categorical Variable = hour, seasons, holiday, func_day, month, day_of_week, weekend

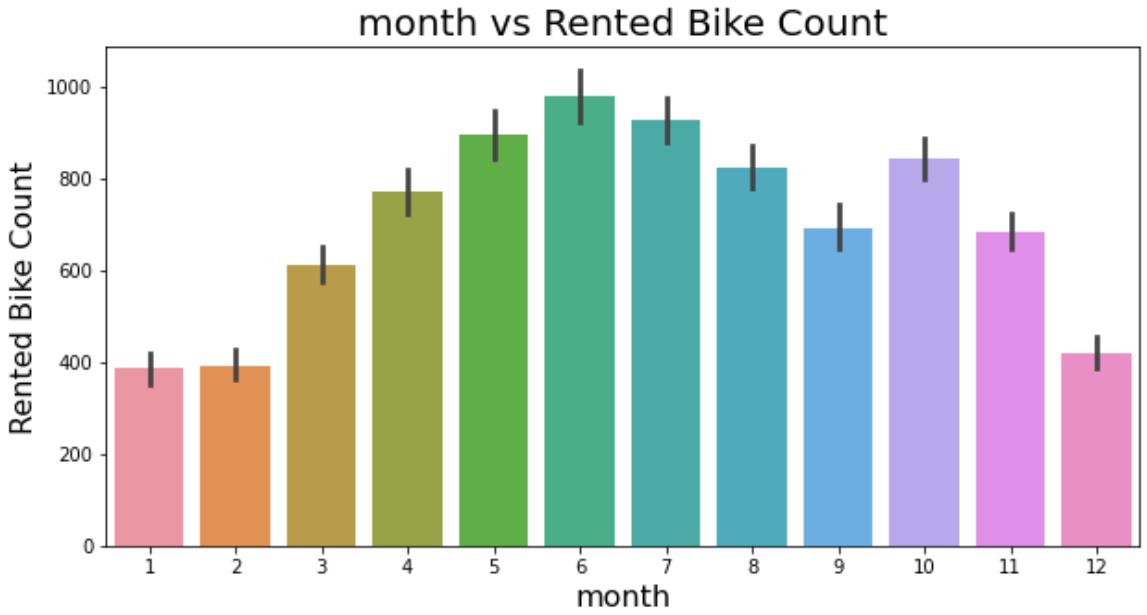


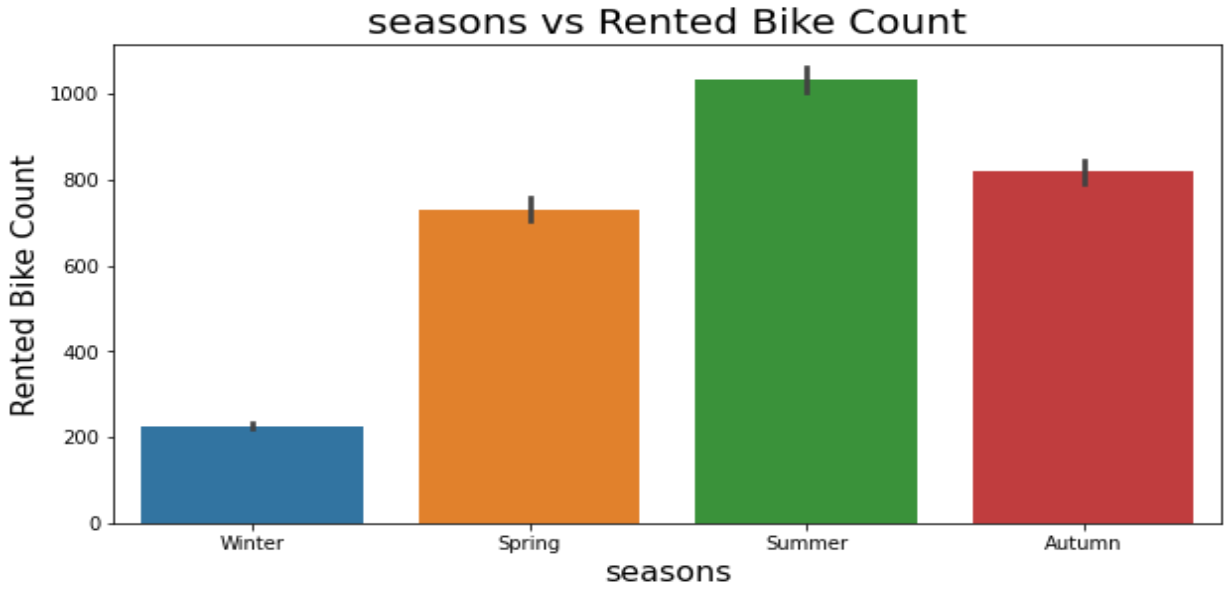






Analysis Of Categorical Variable



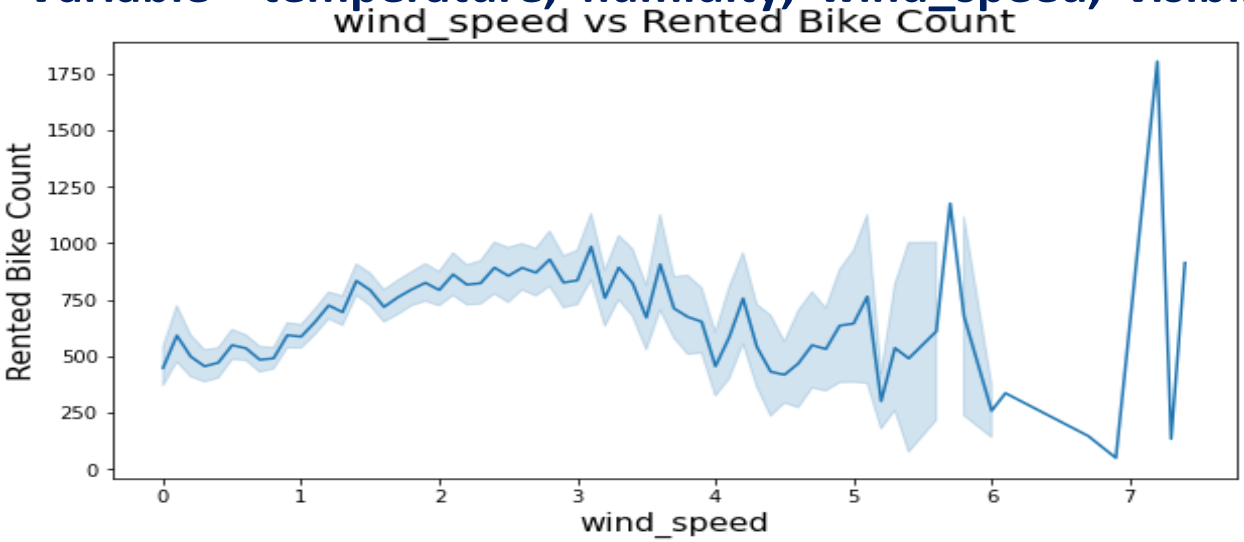


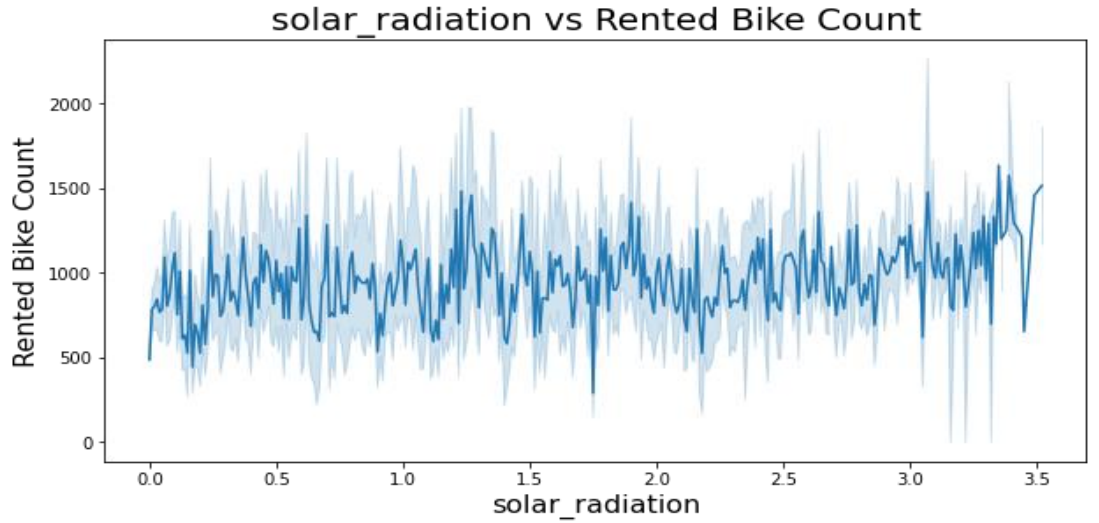
- The number of bikes rented is on average higher during the rush hours.
- The rented bike counts is higher during the summer and lowest during the winter.
- The rented bike count is higher on working days than on non working days.
- On a non functioning day, no bikes are rented in all the instances of the data.
- The number of bikes rented on average remains constant throughout Monday - Saturday, it dips on Sunday, and on average, the rented bike counts is lower on weenends than on weekdays.

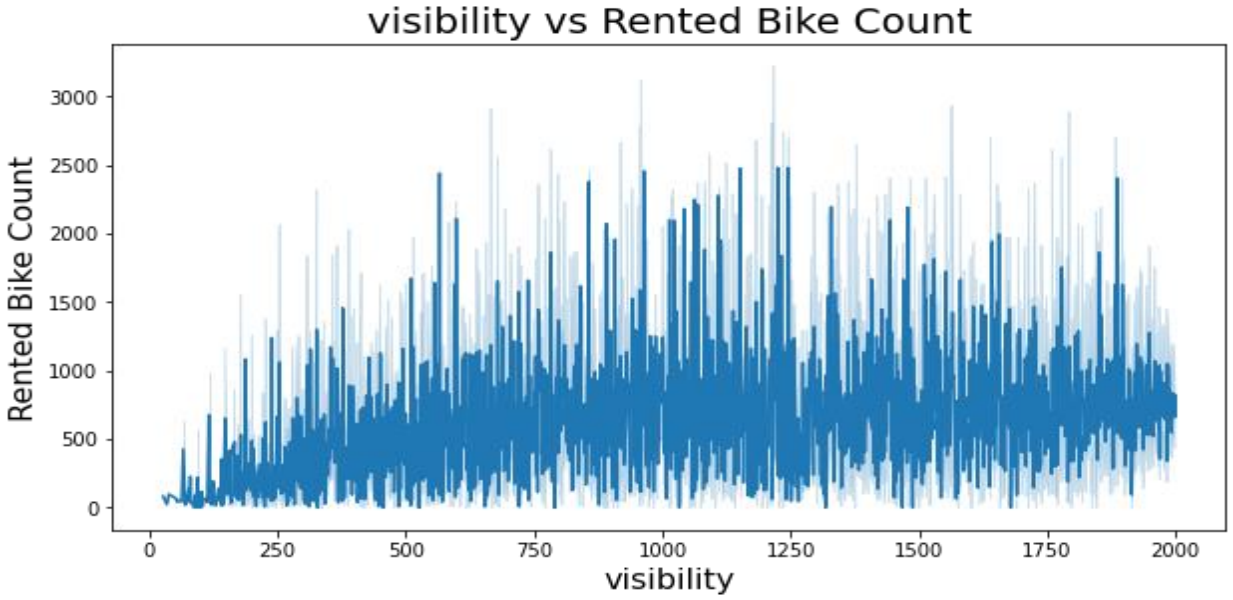
Analysis Of Continious Variable

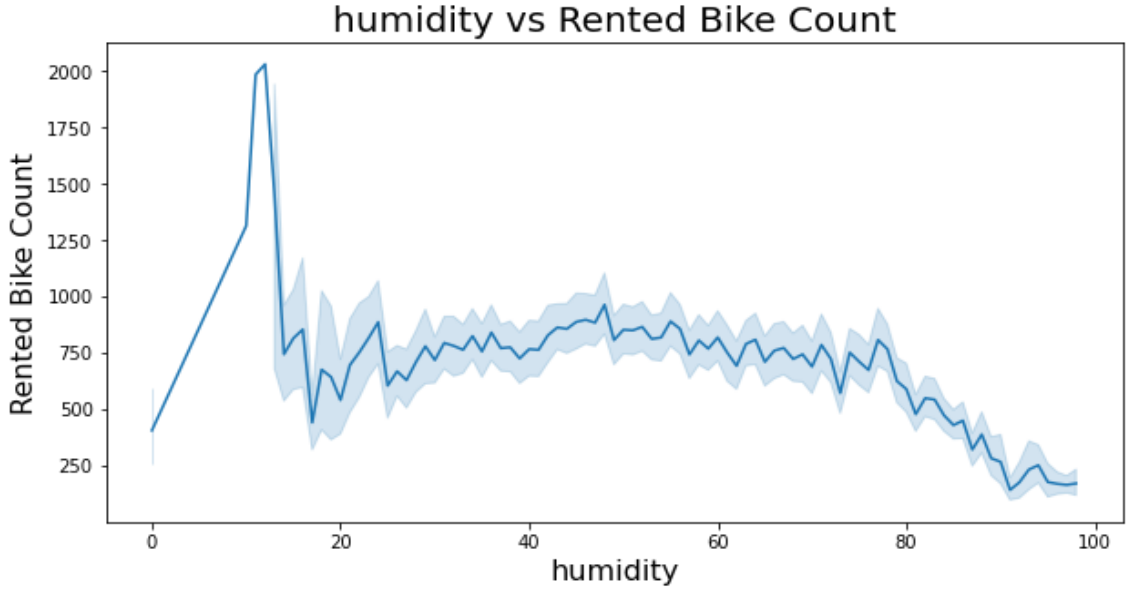
Continuous

Variable = temperature, humidity, wind_speed, visibility, solar_radiation, rainfall, snowfall

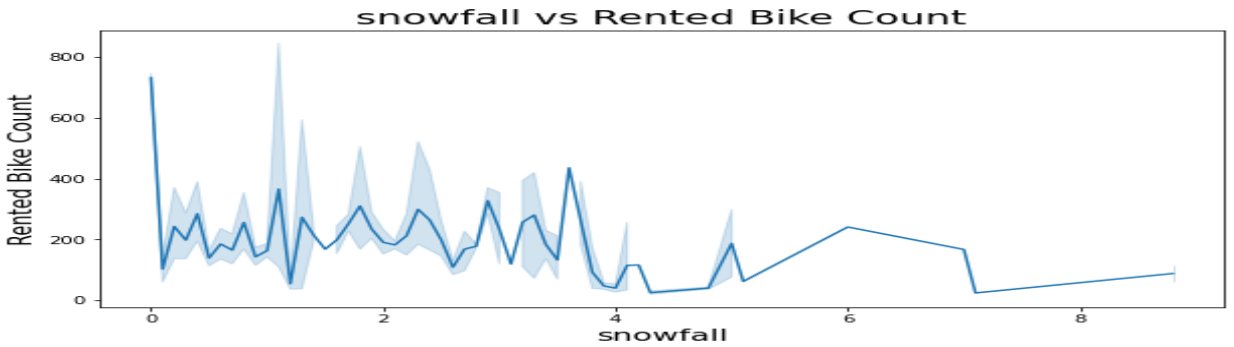


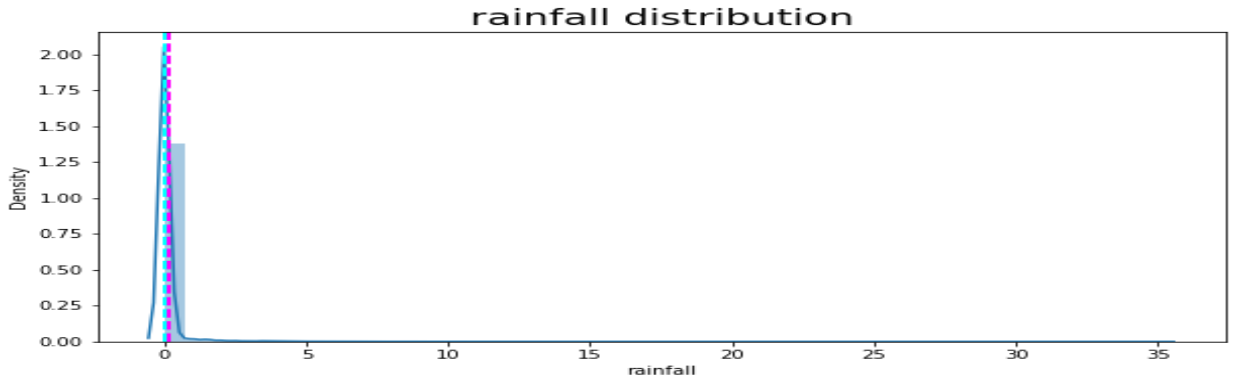


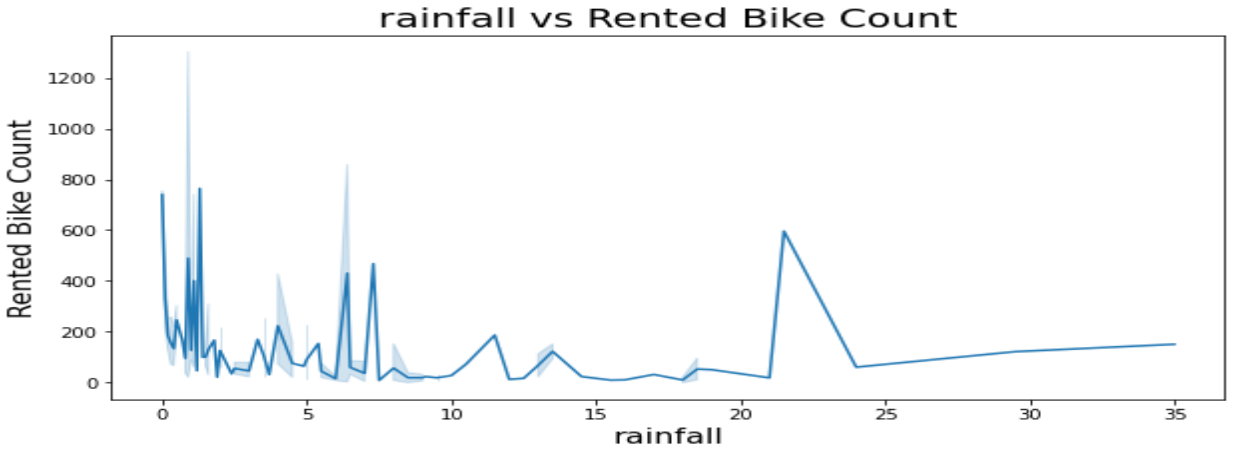




Analysis Of Continuous Variable



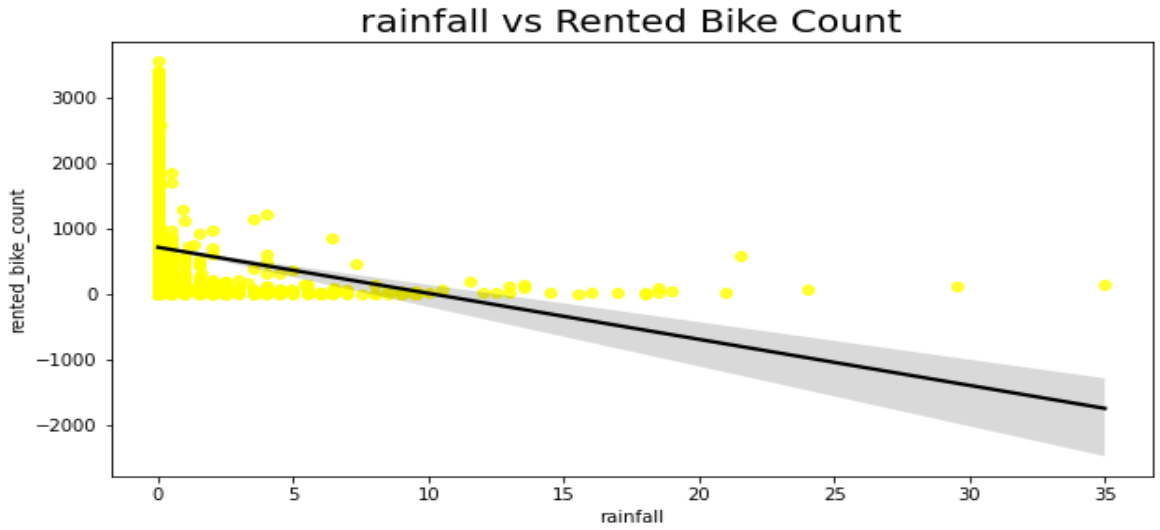


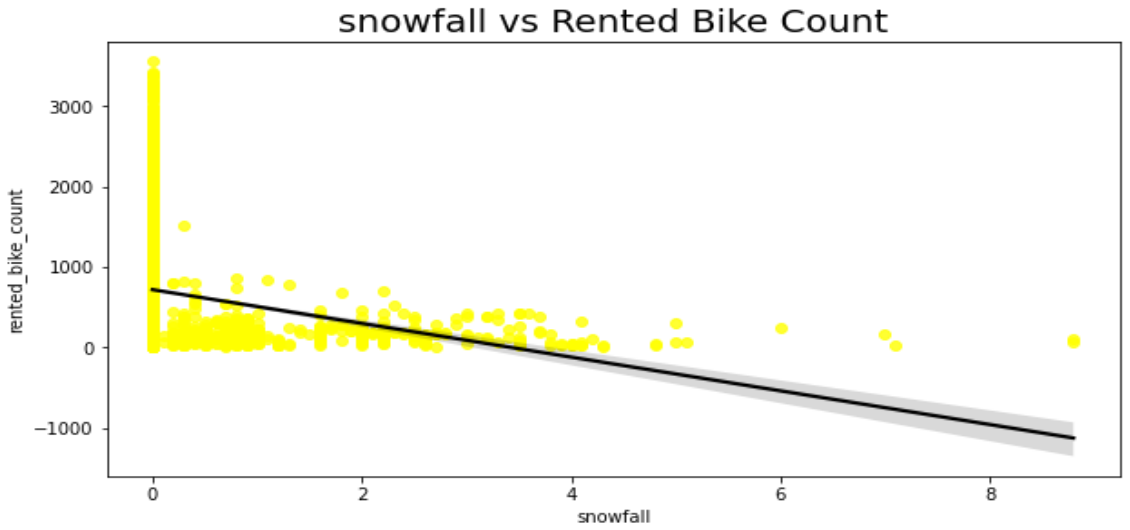


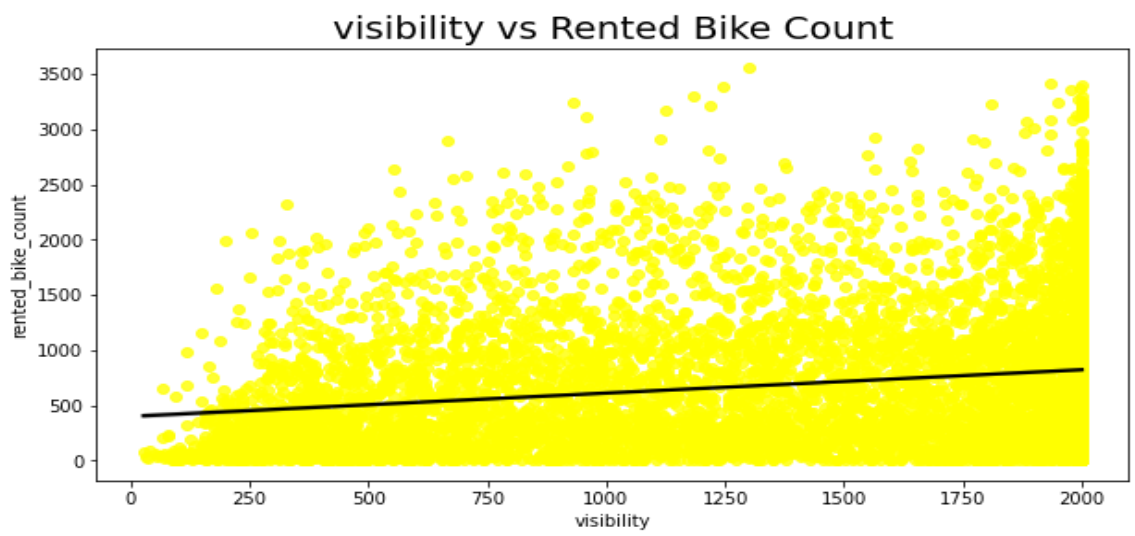
From the above related plot we seen that

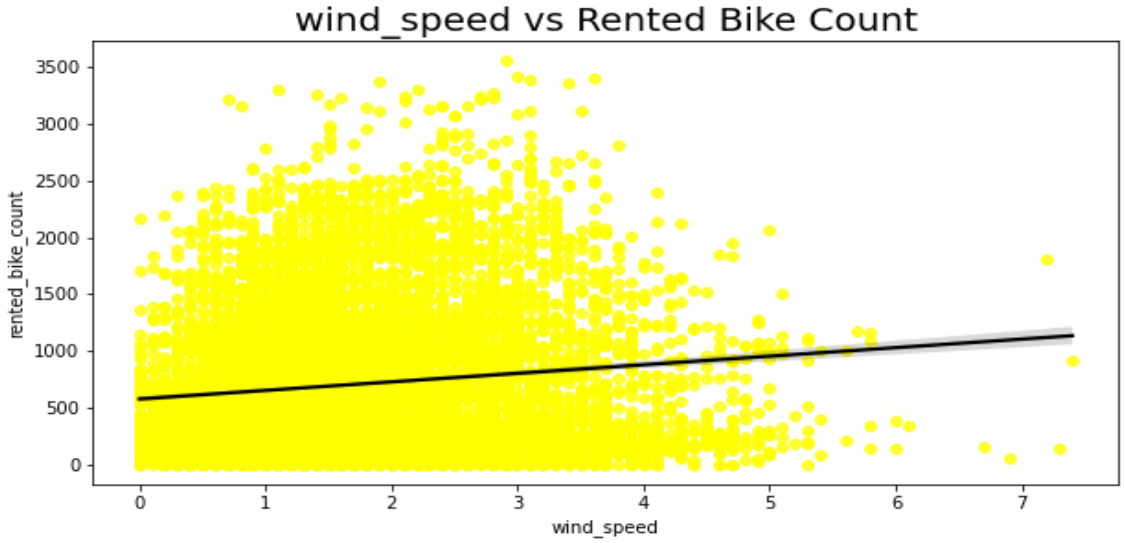
- ✔ People like to ride bikes when it is pretty hot around 25°C in average.
- ✔ The demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of bike also increase that clearly means peoples love to ride bikes when its little windy.
- ✔ The amount of rented bikes is huge, when there is solar radiation, the counter of rents is around 1000.
- ✔ On the y-axis, the amount of rented bike is very low When we have more than 4 cm of snow, the bike rents is much lower.
- ✔ Even if it rains a lot the demand of of rent bikes is not decreasing, here for example even if we have 20 mm of rain there is a big peak of rented bikes.

Regression Plot

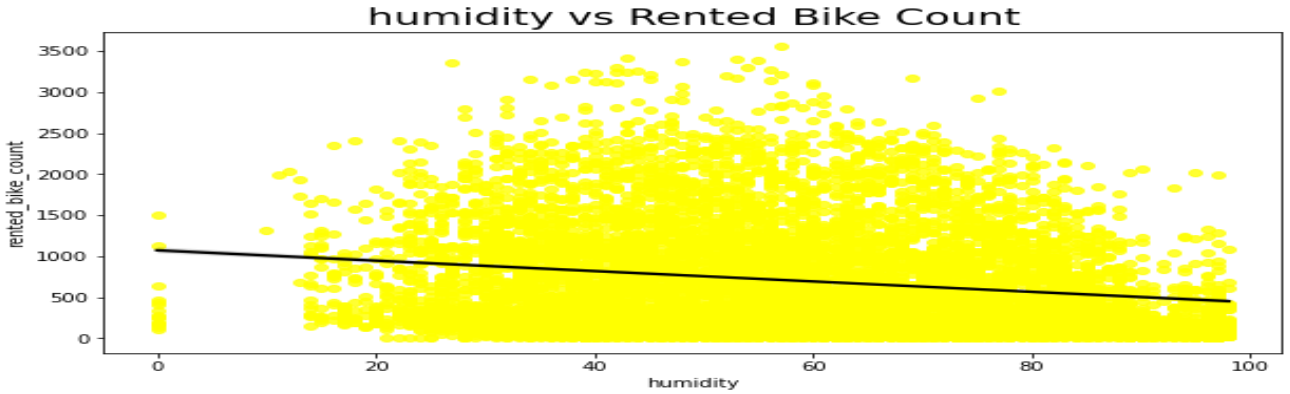


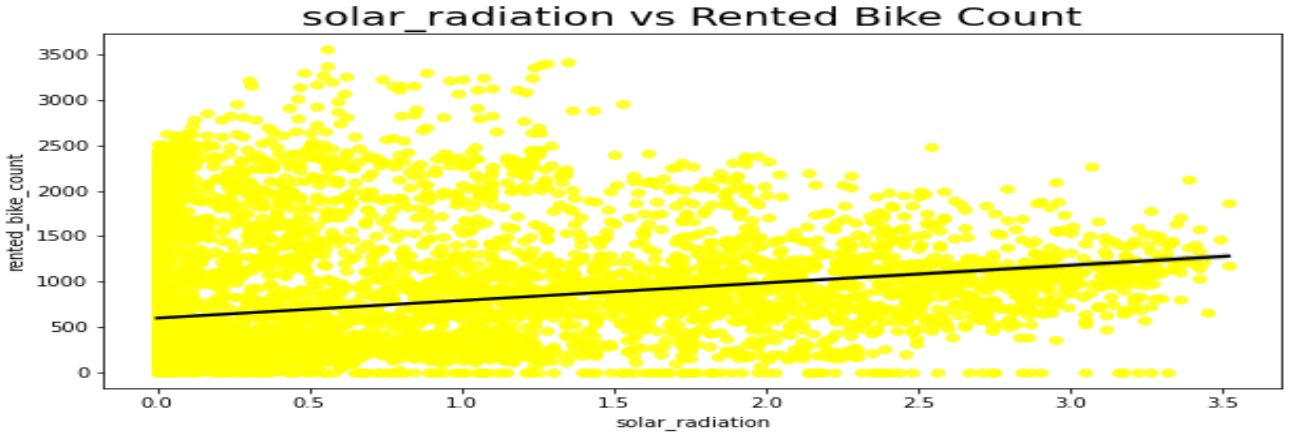


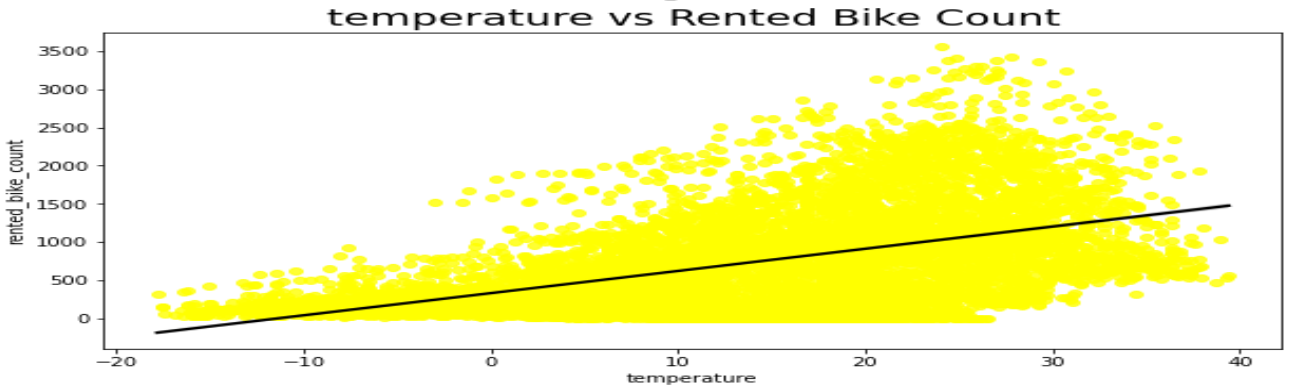




Regression Plot



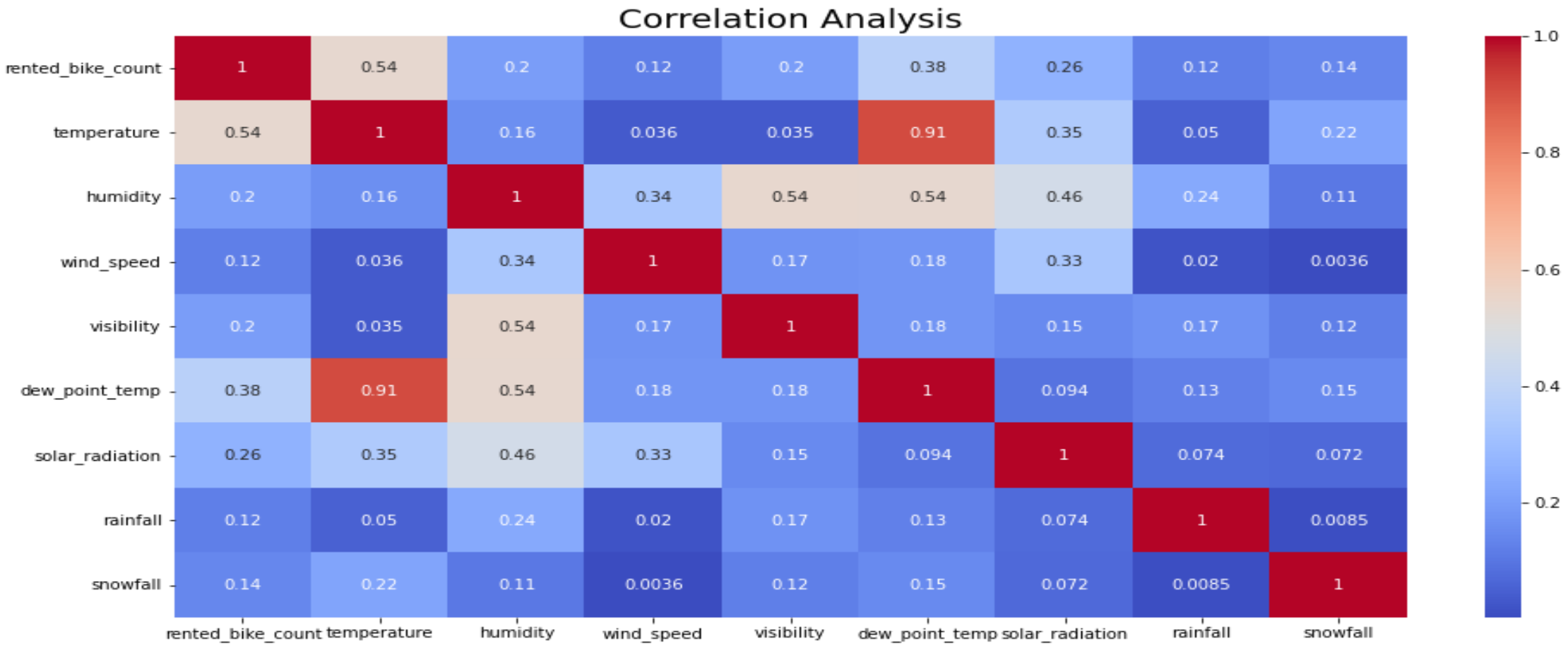




From the above regression plot we see that

- The columns 'Temperature', 'Wind_speed','Visibility', 'Solar_Radiation' are positively relation to the target variable, which means the rented bike count increases with increase of these features.
- Rainfall','Snowfall','Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

Correlation Analysis



- Variables like Dew Point Temperature, and Temperature was highly correlated. So we drop the Dew Point Temperature and check for multicollinearity
- There is no multicollinearity in the data.

Model Building

- LINEAR REGRESSION
- LASSO REGRESSION
- RIDGE REGRESSION
- DECISION TREES REGRESSOR
- RANDOM FOREST REGRESSOR
- GRADIENT BOOSTED REGRESSOR
- GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

Conclusion

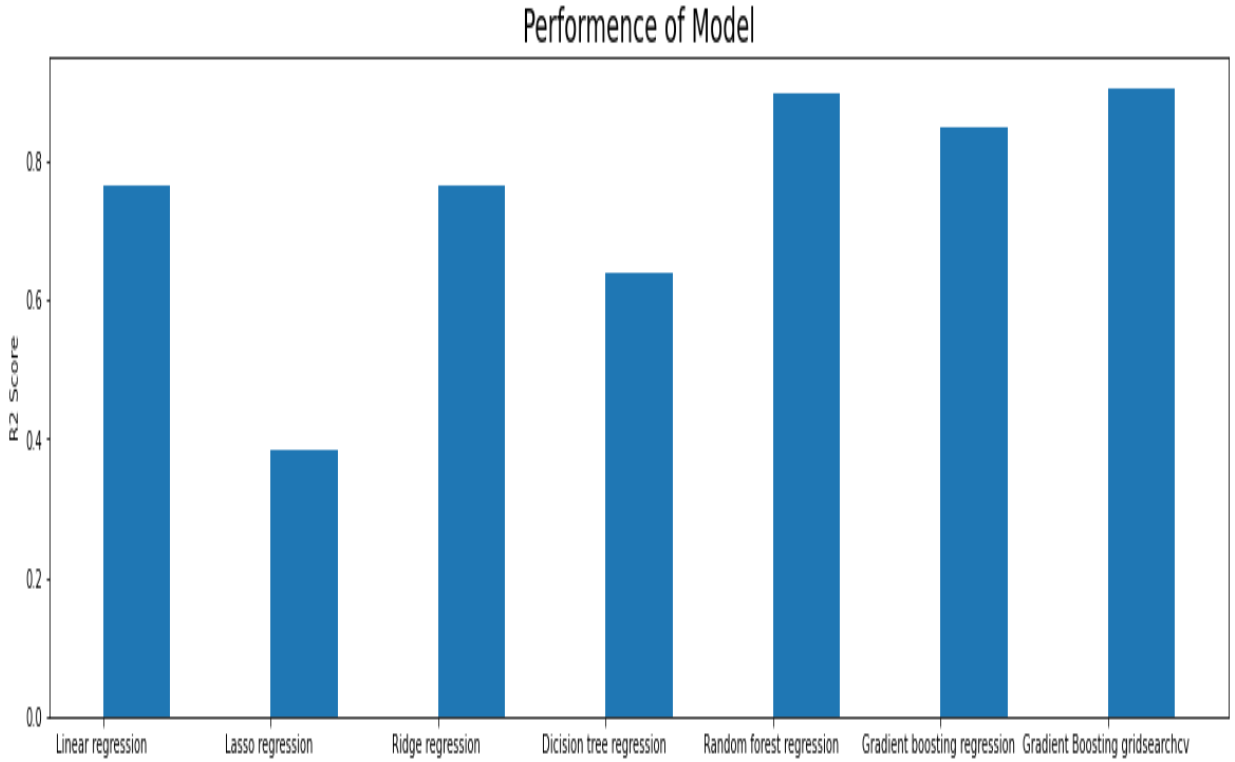
During the time of our analysis, we initially did EDA on all the features of our dataset. We first analyzed our dependent variable, 'Rented Bike Count' and also transformed it. Next we analyzed categorical variable and dropped the variable who had majority of one class, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable. We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.

Next we implemented 6 machine learning algorithms Linear Regression, Lasso, Ridge, Decission tree, Random Forest and GradientBoost. We did hyperparameter tuning to improve our model performance

Conclusion

The results of our evaluation

Model	MAE	MSE	RMSE	R2 Score	Adjusted R2
Linear Regression	4.66	37.13	6.09	0.76	0.76
Lasso Regression	7.443	97.08	9.8	0.38	0.37
Ridge Regression	4.66	37.13	6.09	0.76	0.79
Decision Tree	5.40	54.28	7.36	0.65	0.65
Random Forest	2.605	16.382	4.04	0.89	0.89
Gradient Boosting	3.62	23.65	4.86	0.85	0.85
Gradient Boosting using GridsearchCV	2.65	15.05	3.87	0.90	0.90



Random forest Regressor and **Gradient Boosting gridsearchcv** gives the highest **R2 score** of **89%** and **90%** respectively.

Feature Importance value for Random Forest and Gradient Boost are different.

We can use **Random forest Regressor** and **Gradient Boosting gridsearchcv** for predicting bike rented column on daily basis.