AI

SUPERVISED ML REGRESSION

CAPSTONE PROJECT-2

BIKE SHARING DEMAND PREDICTION

BY

T. OMPRIYA SUBUDHI

CONTENTS

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 SUMMARY

	Date	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)	Seasons	Holiday	Functioning Day
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes
8756	30/11/2018	764	20	3.4	37	2.3	2000	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8757	30/11/2018	694	21	2.6	39	0.3	1968	-9.9	0.0	0.0	0.0	Autumn	No Holiday	Yes
8758	30/11/2018	712	22	2.1	41	1.0	1859	-9.8	0.0	0.0	0.0	Autumn	No Holiday	Yes
8759	30/11/2018	584	23	1.9	43	1.3	1909	-9.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

- This Dataset contain 8760 rows and 14 columns.
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- One Datetime column ‘Date’.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which shows the environmental conditions for that particular hour of the day.

DATA SUMMARY

- There are No Missing Values present
- There are No Duplicate values present
- There are No null values.
- The dependent variable is 'rented bike count' which we need to make predictions on.
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November 2018) (365 days).
- We changed the name of some features for our convenience, they are as follows , ‘date’, 'Bike_Count', 'Hour’, ‘temp’, ‘humidity', ‘wind', ‘visibility’, ‘dew_temp’, ‘sunlight’, rain', ‘snow', ‘seasons', ‘holiday', ‘functioning_day’.

ATTRIBUTES OF EACH VARIABLE

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)

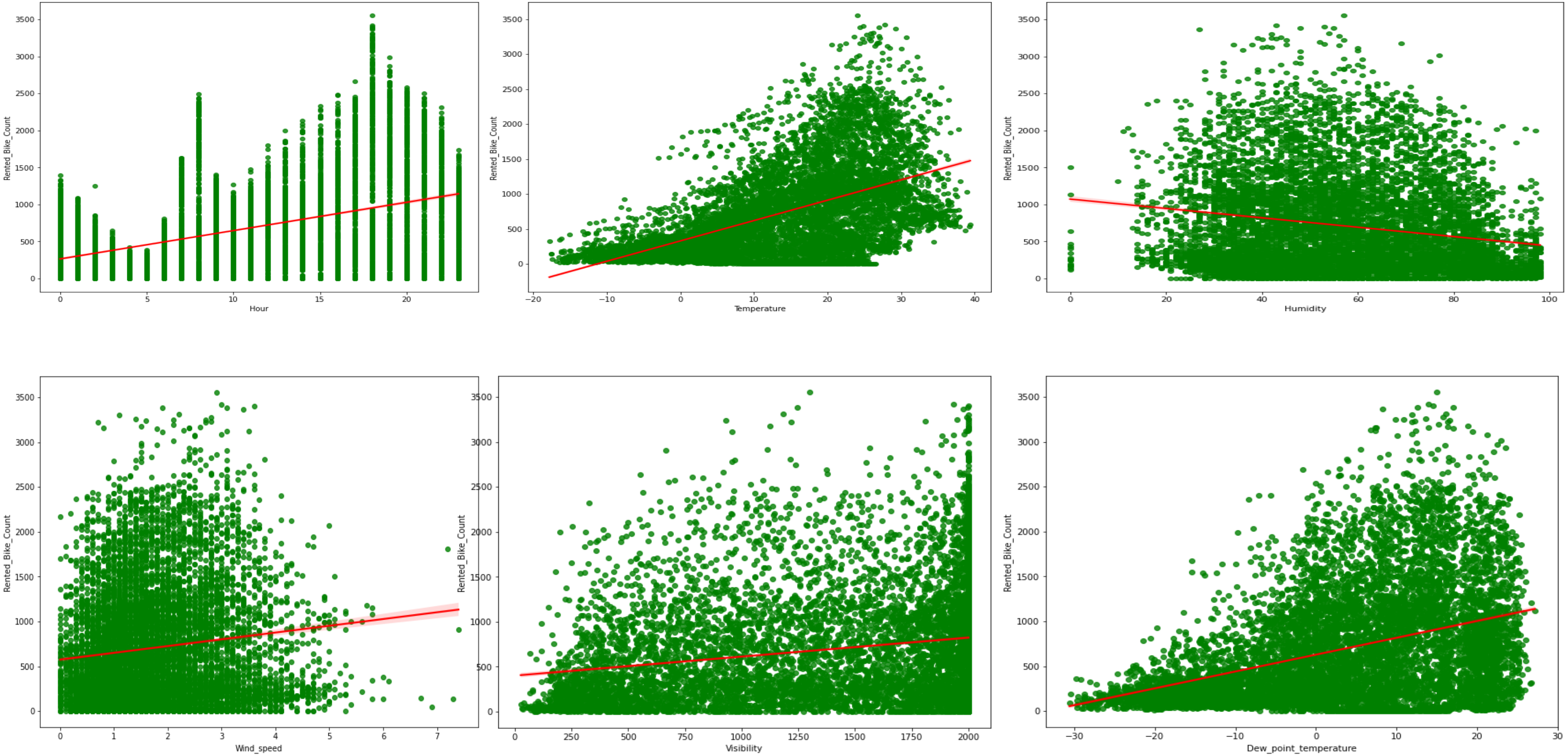
Snowfall: Snowfall (cm)

Seasons: Winter, Spring, Summer, Autumn

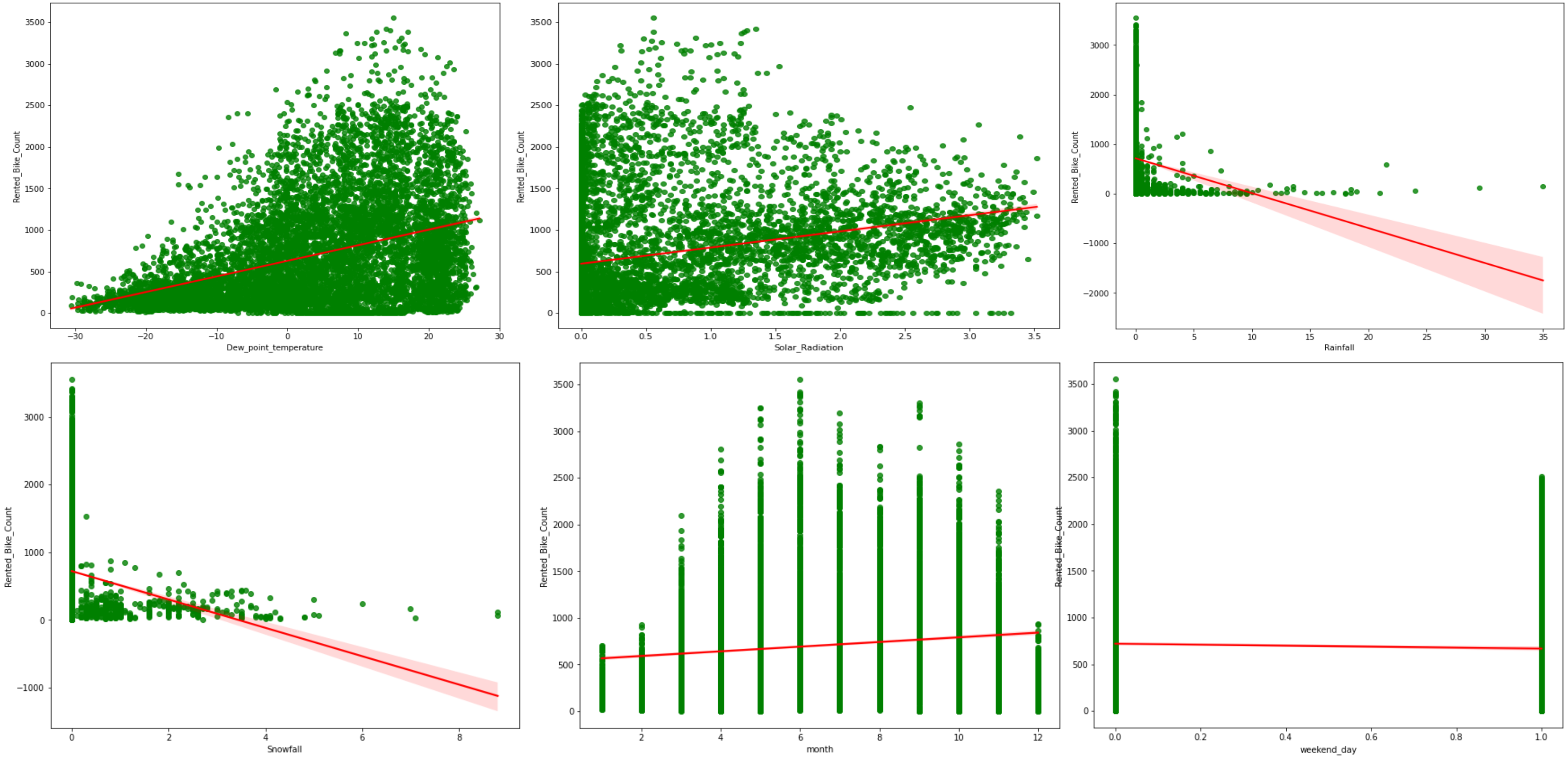
Holiday: Holiday/No holiday

Functioning Day: if the day is neither weekend, holiday than 1 else 0

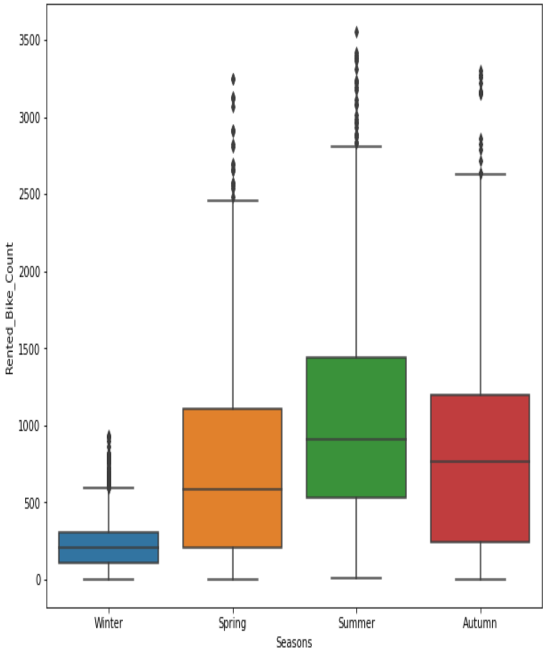
CHECKING LINEARITY IN THE DATA

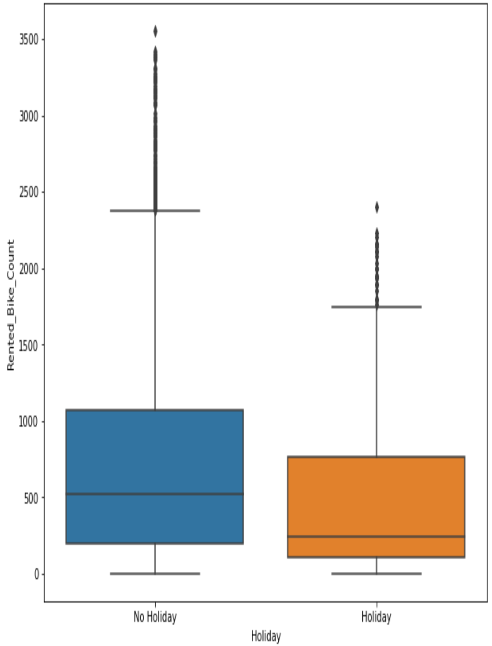


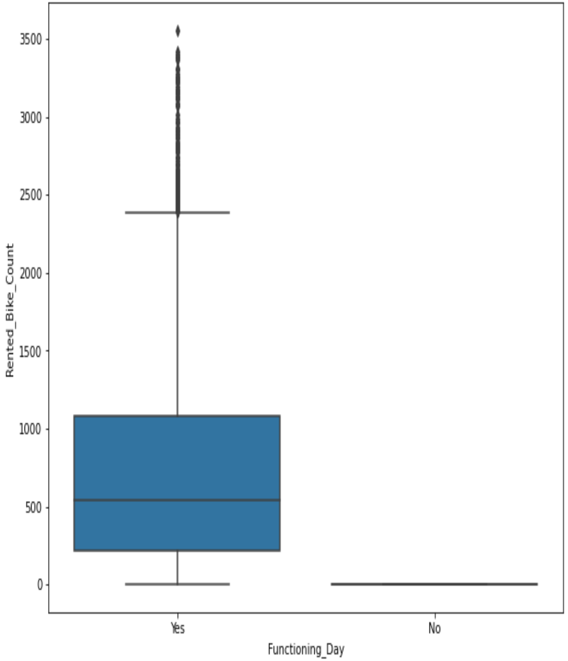
CHECKING LINEARITY IN THE DATA



EXPLORATORY DATA ANALYSIS







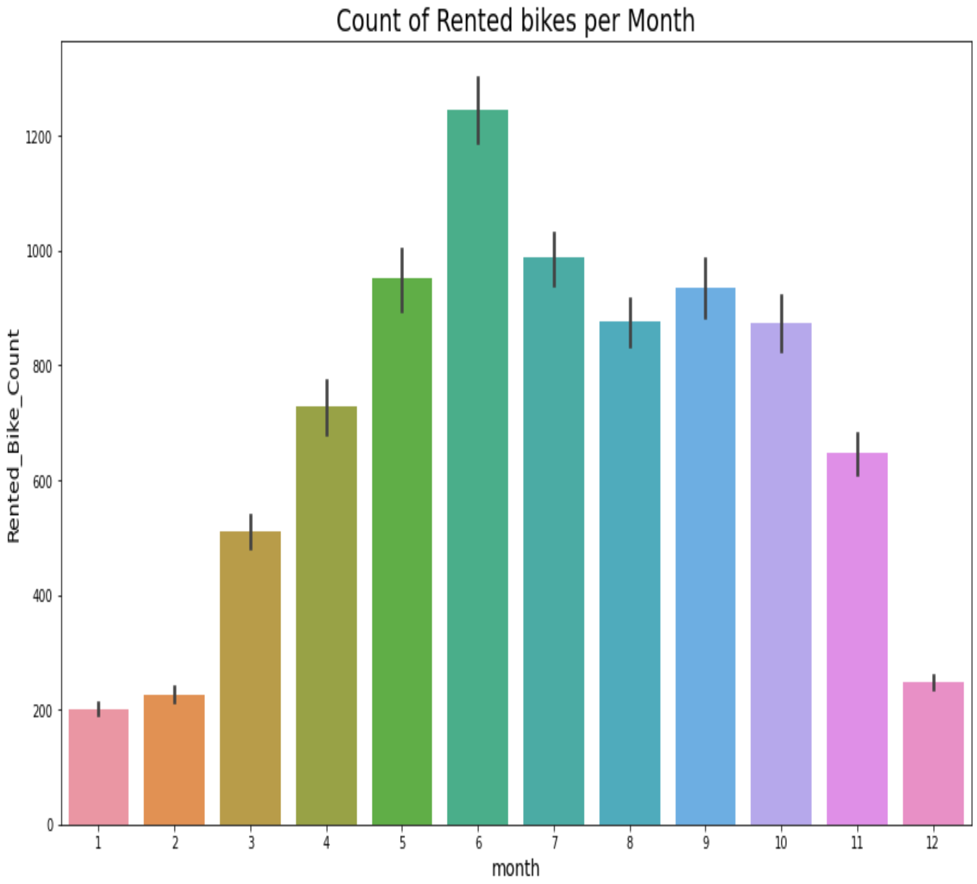
Seasons, Holiday & Functioning day vs Rented bike count

1. Less demand on winter seasons
2. Slightly Higher demand during Non holidays
3. Almost no demand on non functioning day

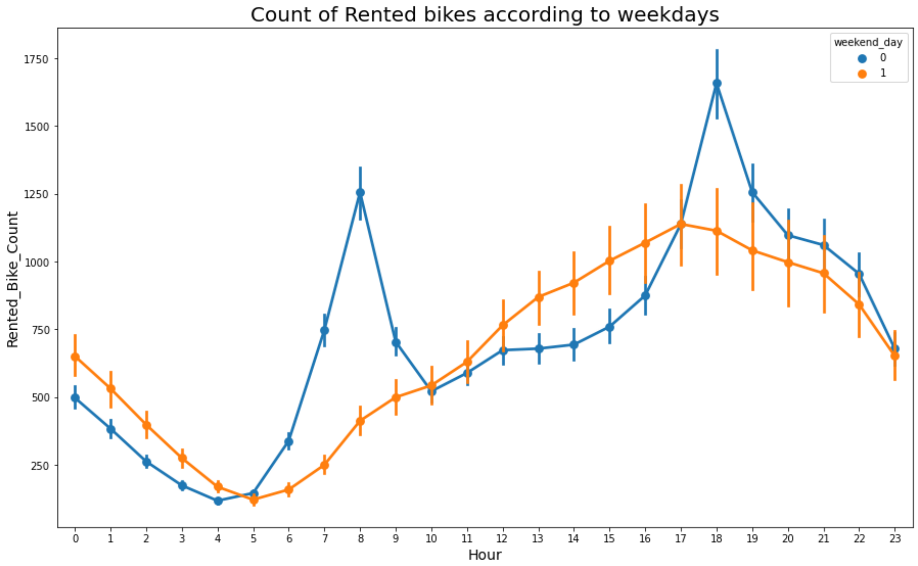
EXPLORATORY DATA ANALYSIS..

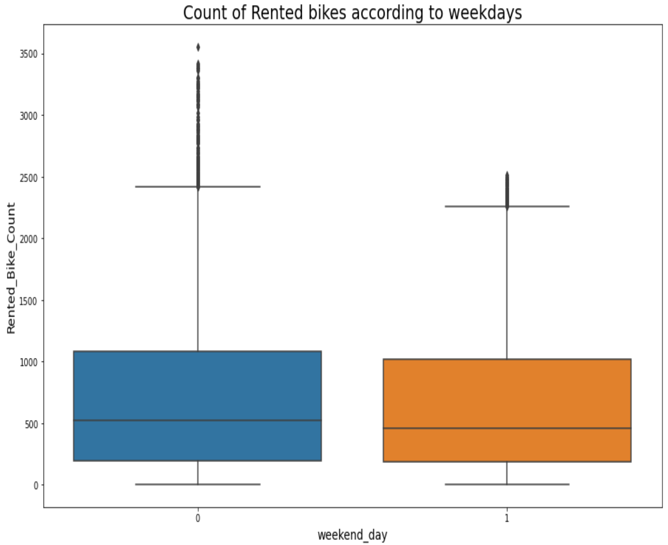
Month vs Rented bike count

1.From the above bar plot we can clearly say that the demand for rented bike from month 5 to 10 is higher than other months and these months fall in summer
2.The highest demand of rented bike is in the month of June.
3.The lowest demand of rented bike is in the month of January and February



EXPLORATORY DATA ANALYSIS..



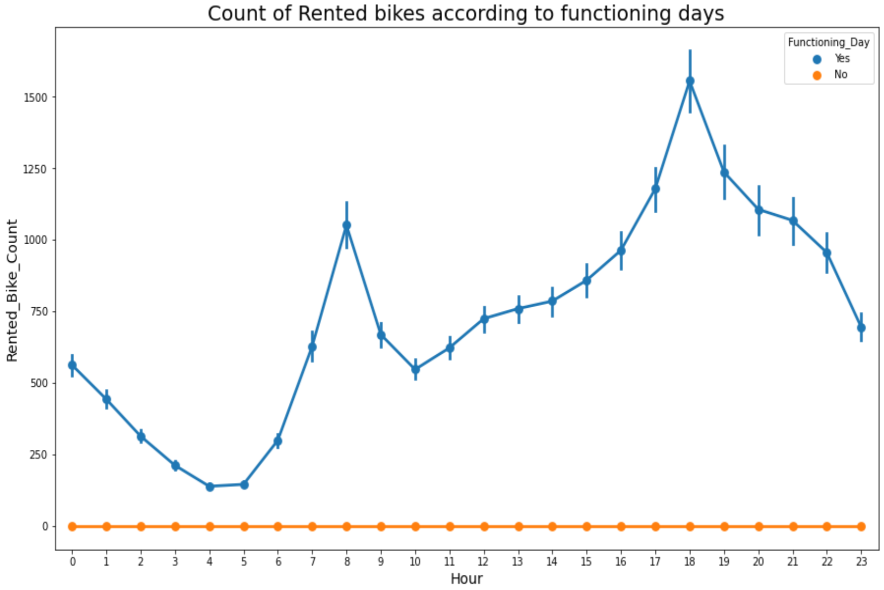


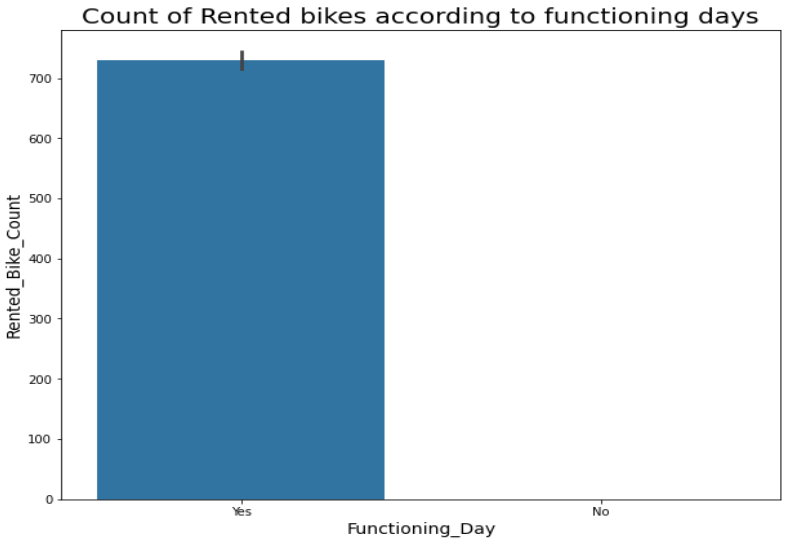
Weekdays Vs Rented Bike count

The Weekend day and Week off day based Analysis shows almost equal weightage on rented bike count.

The 1st plot shows that for weekends the rented bike counts remain in saddle condition whereas for weekdays rented bike count is peak at 8.00 A.M and in the evening at 6.00 P.M which may be the result of working and class time which rent off bikes during the day.

EXPLORATORY DATA ANALYSIS..

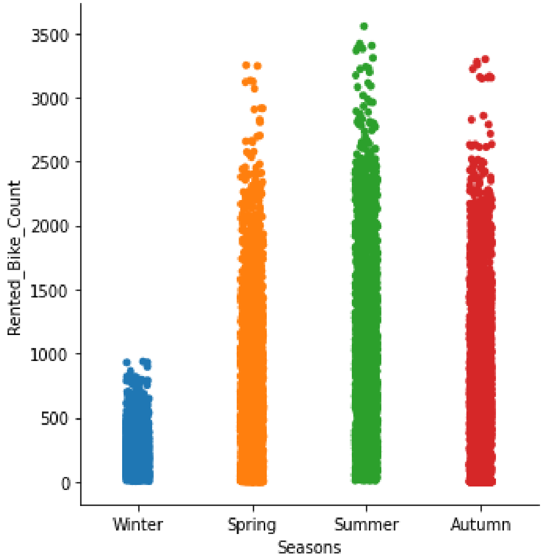


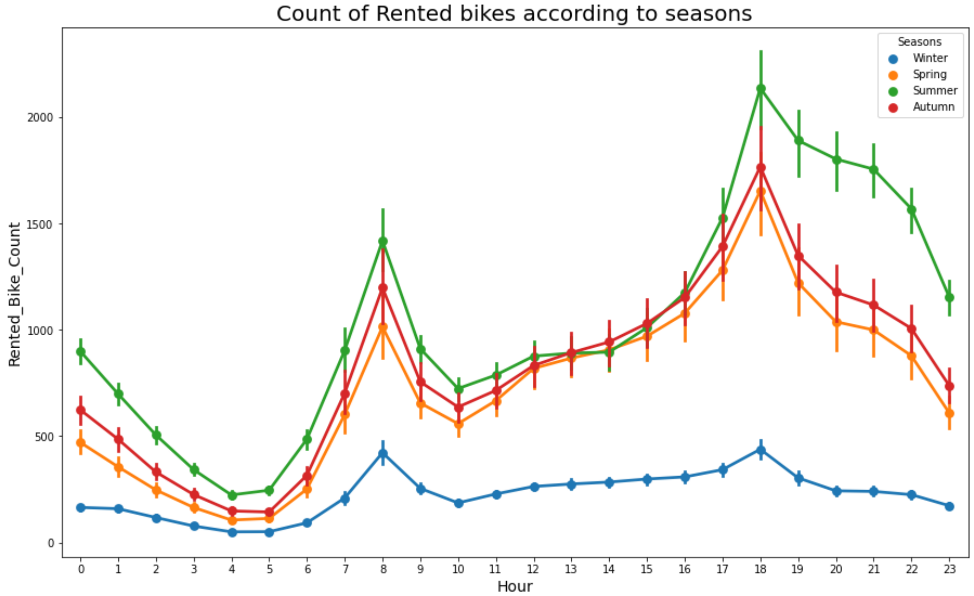


Function day vs Rented bike count

Here we can clearly see that the rented bike count directly proportional to only functioning days.

EXPLORATORY DATA ANALYSIS..





Seasons vs Rented bike count

From the above cat plot and point plot we can concluded that highest number of bike have rented in summer season.

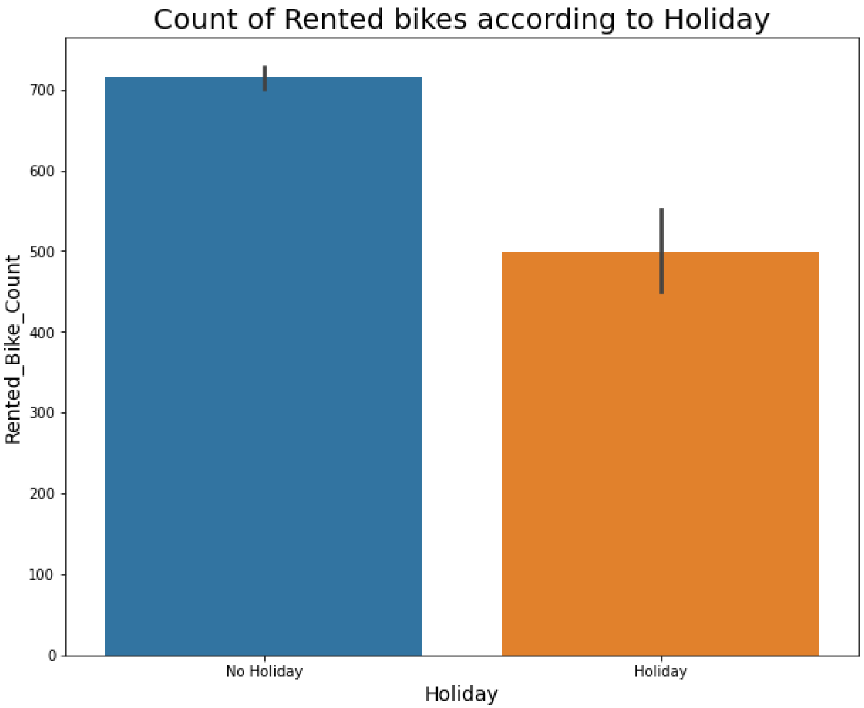
Less number of bike rented in winter season.

Almost equal percentage of bike rented in spring and autumn

EXPLORATORY DATA ANALYSIS..

Holiday Vs Rented bike count

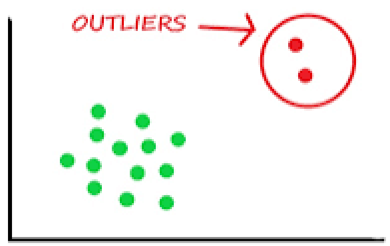
The box plot figure shows the relation of rented bike count on holidays. Since its values are unidirectional it may not be an important feature to predict bike sharing demand



OUTLIERS

Outliers are those **data points that are significantly different from the rest of the dataset**. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

Outliers brings skewness in the data. Thus decreasing the accuracy sometimes. So we will deal with this problem and make our distribution normal



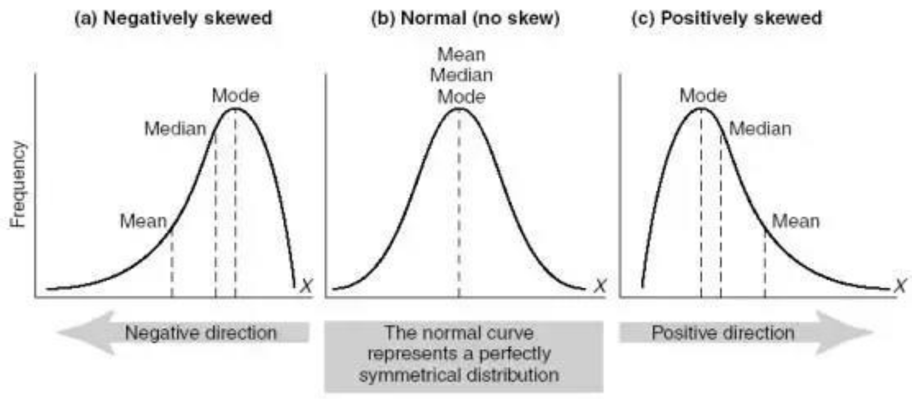
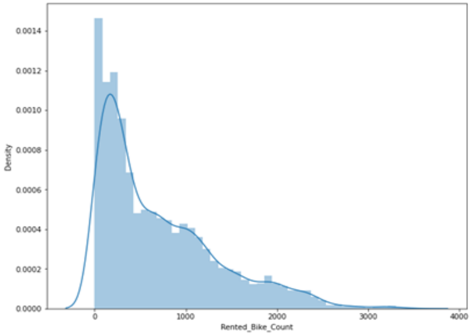
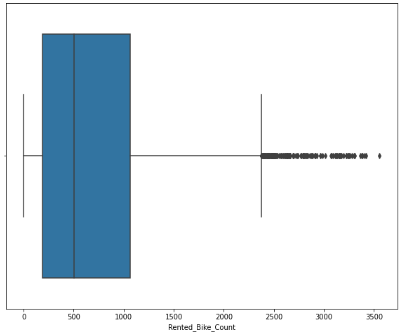


FIGURE 15.6 Examples of normal and skewed distributions

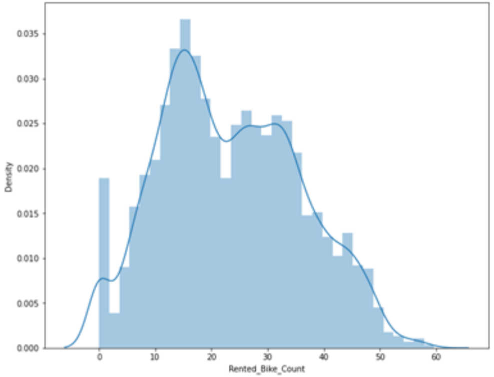
OUTLIERS (CONTINUED)

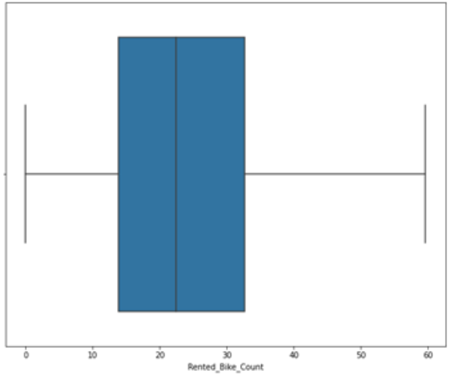




In the following graph plots you can see positive skewed histogram and its corresponding skewed probability plot because of outliers present.

To correct the skewness we have applied <u>square root transform</u> . To get the normal distribution from positive skewed data.



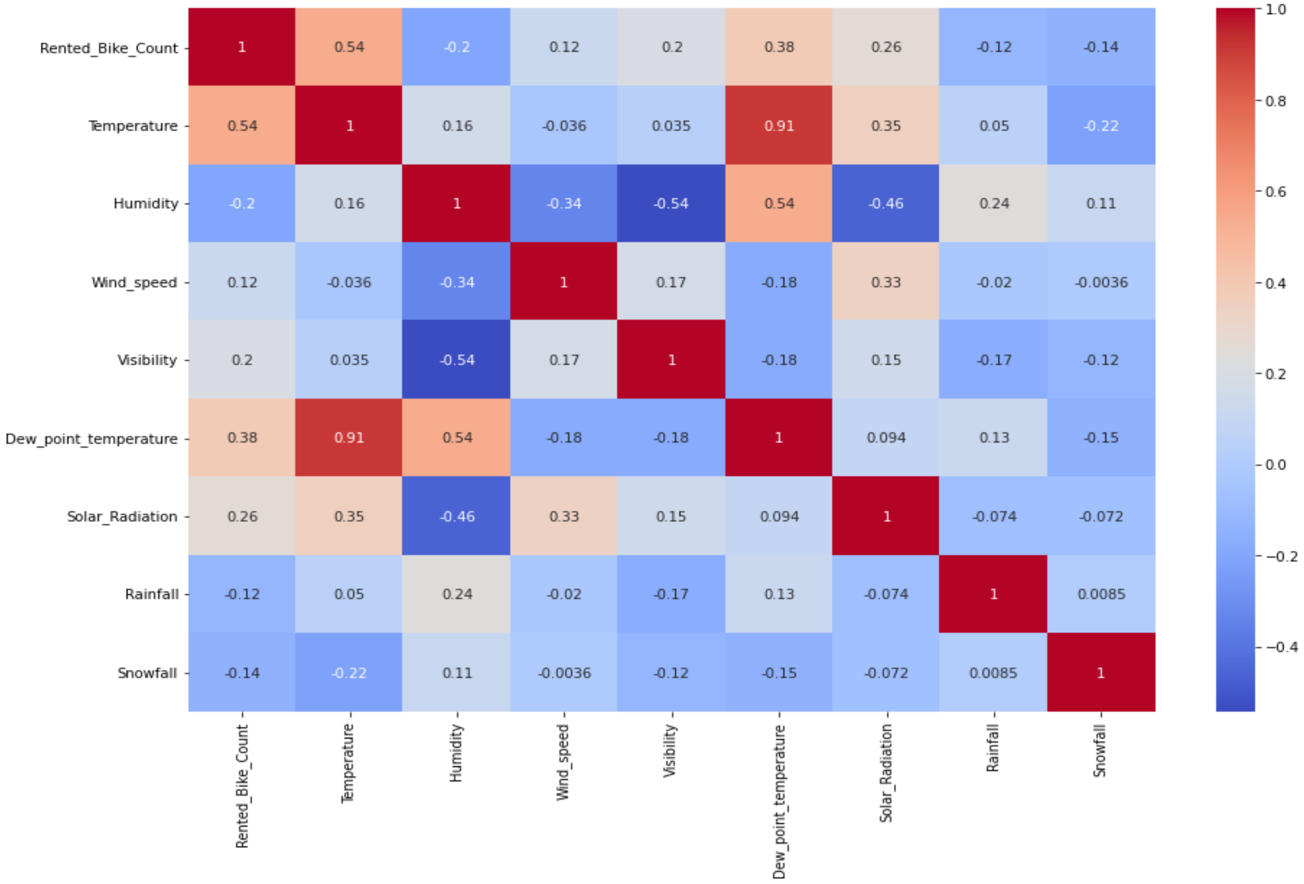


CORRELATION ANALYSIS

The correlation analysis has been done to get a better understanding of importance of other features on dependent variable.

The most critical factors for predicting the number of bikes needed per hour, is temperature and solar radiation.

By analyzing multicollinearity we can see temperature and dew point temperature are highly correlated , so as the value with rented bike count is less for dew temperature we may drop dew temperature.



MODEL BUILDING PREREQUISITES

Feature Scaling or Standardization: It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically **helps to normalize the data within a particular range**. Sometimes, it also helps in speeding up the calculations in an algorithm. Here we used standard scaler.

MODELS BUILT

Total 8 models have done

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Random forest
5. Polinomial Regression
6. Decision Tree Regression
7. Elastic Net Regularization
8. CV Elastic Net Regularization

MODEL ANALYSIS

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2 score
Train set	**0**	Linear Regression	5.582400	52.526900	7.247500	0.663800	0.661300
	1	Lasso Regression	5.582800	52.527100	7.247600	0.663800	0.661300
	2	Rigde Regression	5.587900	52.544100	7.248700	0.663700	0.661200
	3	Random Forest	2.564600	12.331100	3.511600	0.921100	0.920500
	4	ElasticNet Regularization	5.641900	53.262600	7.298100	0.659100	0.656600
	5	CV ElasticNet Regularization	5.590800	52.561600	7.249900	0.663600	0.661100
	6	Polynomial Regression	4.194200	32.120000	5.667400	0.794500	0.781900
	7	Decision Tree	2.626100	13.070700	3.615300	0.848600	0.91130

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2 score
Test set	**0**	Linear Regression	5.598800	54.960400	7.413500	0.636900	0.634200
	1	Lasso Regression	5.598800	54.955700	7.413200	0.636900	0.634200
	2	Ridge Regression	5.601400	54.957100	7.413300	0.636900	0.634200
	3	Random Forest	2.759000	14.984400	3.871000	0.901000	0.900300
	4	ElasticNet Regularization	5.623500	55.101200	7.423000	0.635900	0.633200
	5	CV ElasticNet Regularization	5.602700	54.947200	7.412600	0.636900	0.634300
	6	Polynomial Regression	4.385600	43.065500	6.562400	0.716500	0.699100
	7	Decision Tree	3.140809	22.996755	3.615339	0.848629	0.839332

CONCLUSION

- Rental bikes are in demand on holidays or non-holidays . We may say that the number of rental bikes is significantly higher on non-holidays than on holidays.
- 8AM and 6PM have high demand people go to their work at 8 am and return from work at 6 pm. The demand for rented motorcycles is most closely related to the number of working hours per day.
- People prefer bikes rented in the morning rather than in the evening.
- People have booked more bikes except in few cases when rain has subsided.
- After testing various feature combinations using linear regression, the model was found to be unsuitable. Because the data is so widely scattered, it became clear. Fitting a line didn't seem realistic.
- The most critical factors for predicting the number of bikes needed were hour, temperature and solar radiation.
- With good model performance and low RMSE , the random forest regressor outperforms than linear regression.
- Elastic Net regularization and Cross Validation on Elastic Net regularization are not fitted for this model because both has low r2 score.
- Polynomial regression performs better than Linear Regression
- Decision tree can be unstable because small variations in data might result in completely different tree being generated . We got adjusted r2 score as 0.91 & 0.83 for training and testing data respectively
- Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low.
- Best predictions are obtained with Random forest Regressor with applied hyperparameter tuning with r2 score of **0.9** and RMSE of **3.87**

THANK YOU