Capstone Project

Bike Sharing Demand Prediction

Gajanan kale

Problem statement

The contents of the data came from a city called Seoul. A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. The dataset variables are such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count.

The problem statement is to build a machine learning model that could predict the rented bikes count required per hour, given the other variables.

The crucial part is the prediction of bike count required at each hour for the stable supply of renta l bikes.

Points to discuss

- Data description and summary
- Analysis of categorical variable
- Analysis of numerical variable
- Handling outliers
- Regression plot
- Machine learning algorithms
- conclusion

Data description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

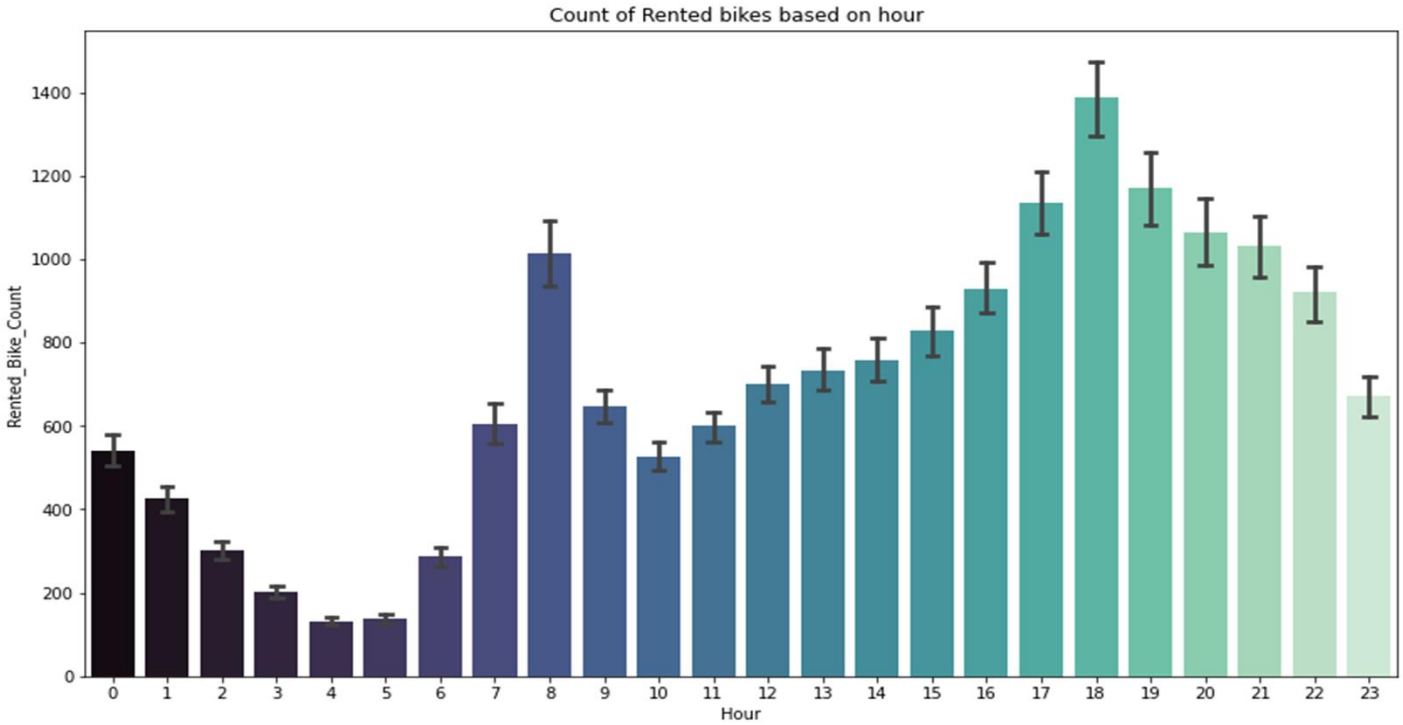
- Date : year-month-day
- Rented Bike count -Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm

Data description(cont,.)

- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day -NoFunc(Non Functional Hours), Fun(Functional hours)

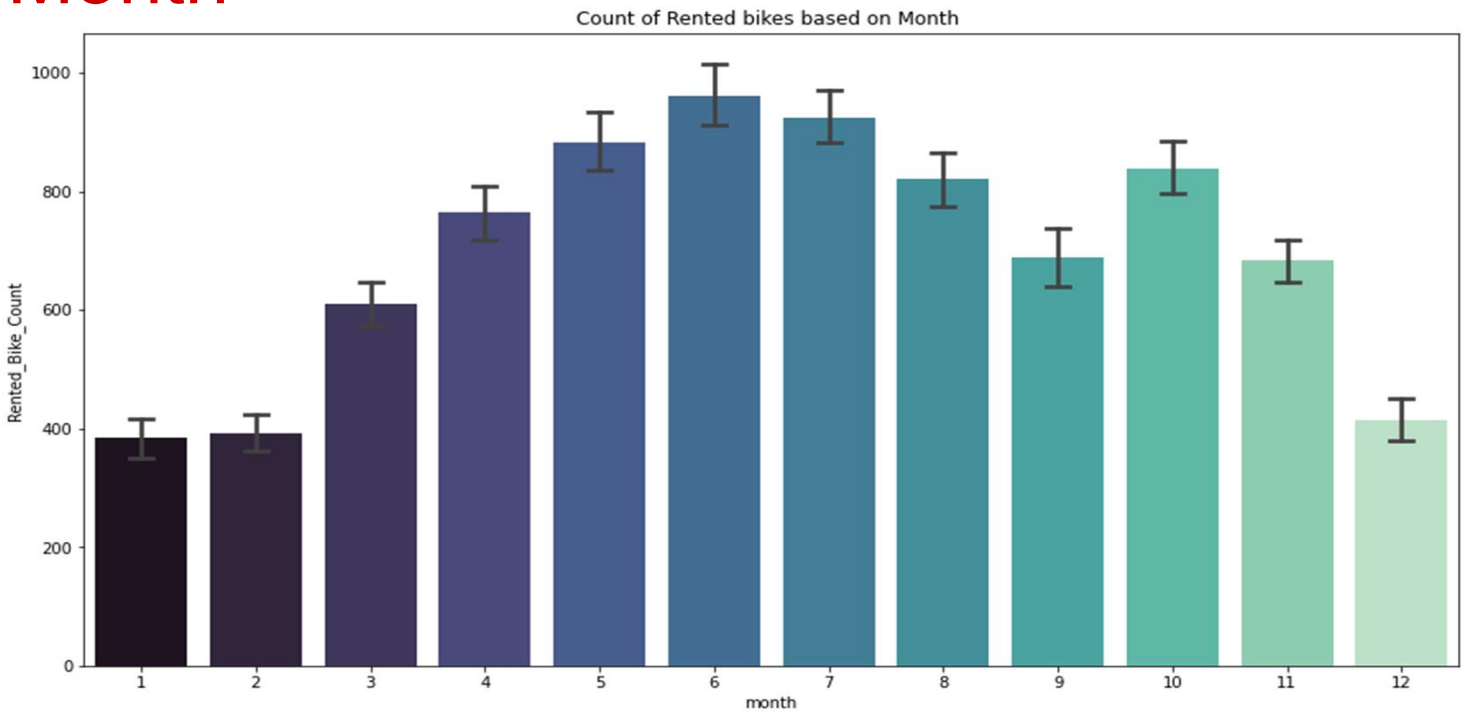
1. This dataset contains 8760 lines and 14 columns
2. Numerical variables -temperature, humidity,wind,visibility,dew point temp, solar radiation,rainfall,snowfall
3. Categorical variables -seasons,holiday and functioning day
4. Rented bike count -which we need to predict for new observations

Hour



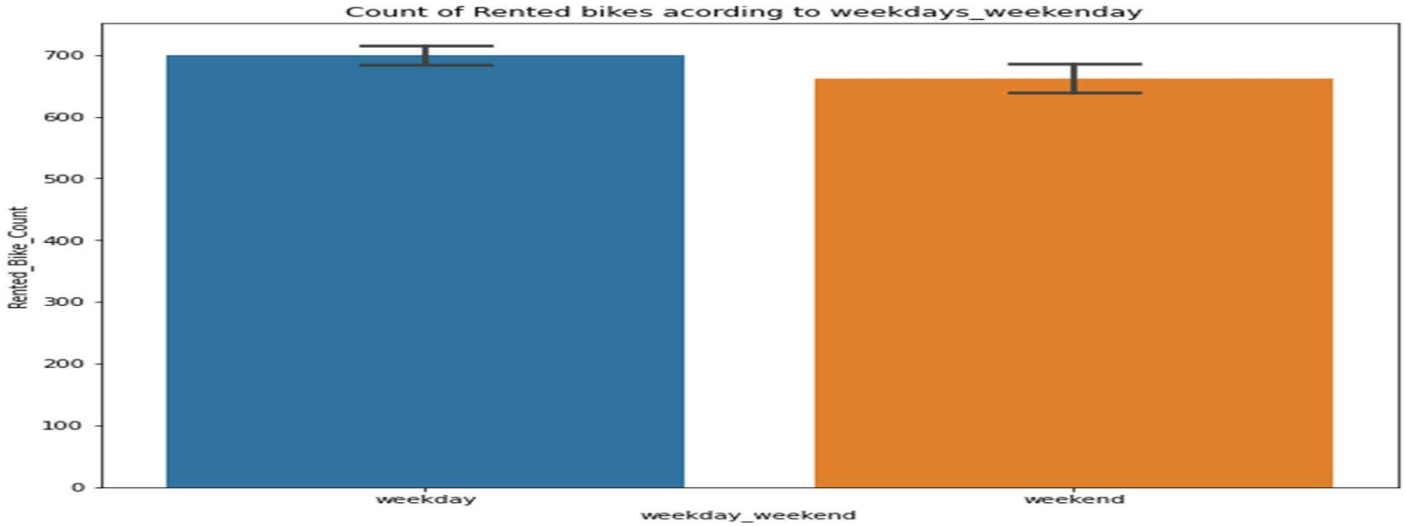
The demand of the rented bike is high between 7 am to 9 am and in evening 5 pm to 8 pm.

Month



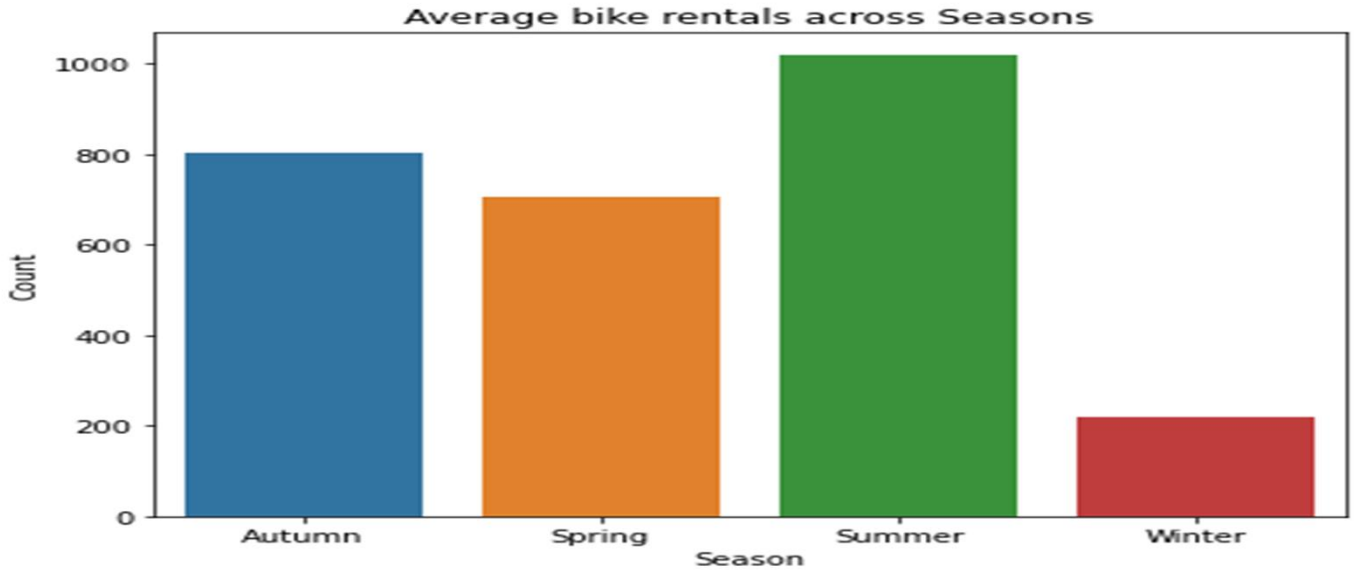
The demand of the rented bike is high from month 5th –May to 11th –November.

weekday



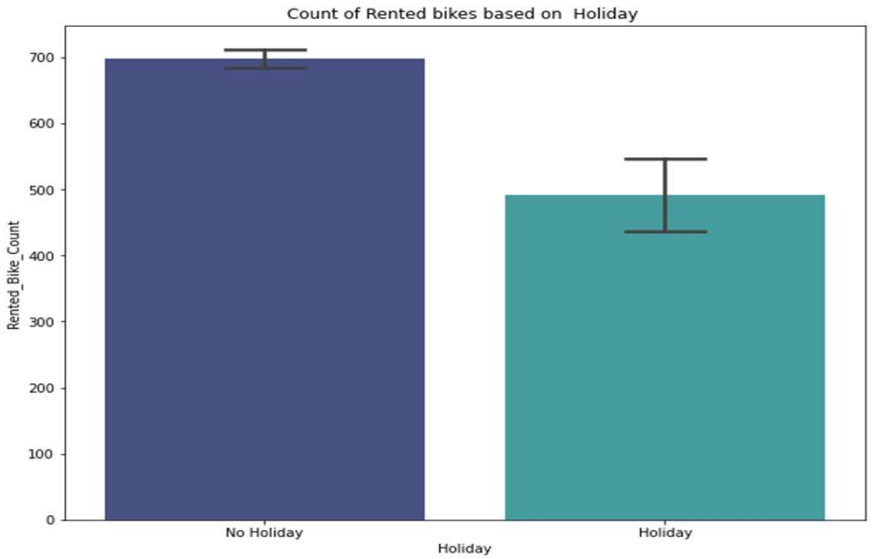
From the above bar plot we can say that in the week days which represent in blue colur show that the demand of the bike higher because of the office. Peak Time are 7 am to 9 am and 5 pm to 7 pm. The orange colour represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

seasons



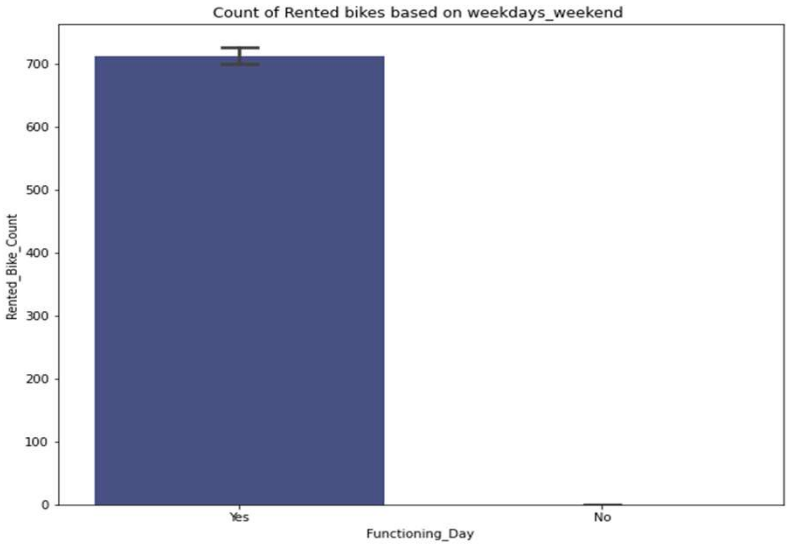
In the bar plot it shows the use of rented bike in in four different seasons, and it clearly shows that, In summer season the use of rented bike is high In winter season the use of rented bike is very low because of snowfall.

Holiday



From above plot we conclude that rented bikes are more used on no holiday compaired to holiday.

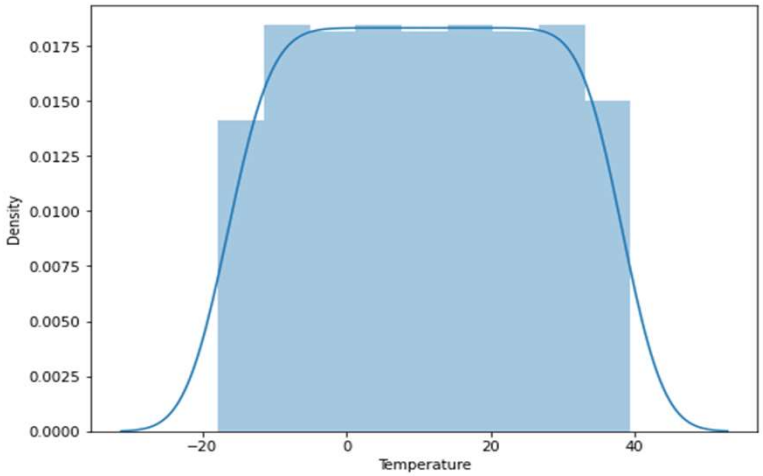
Functioning day



In the above Bar plot which shows the use of rented bike in functioning day or not, and it clearly shows that, Peoples dont use reneted bikes in no functioning day

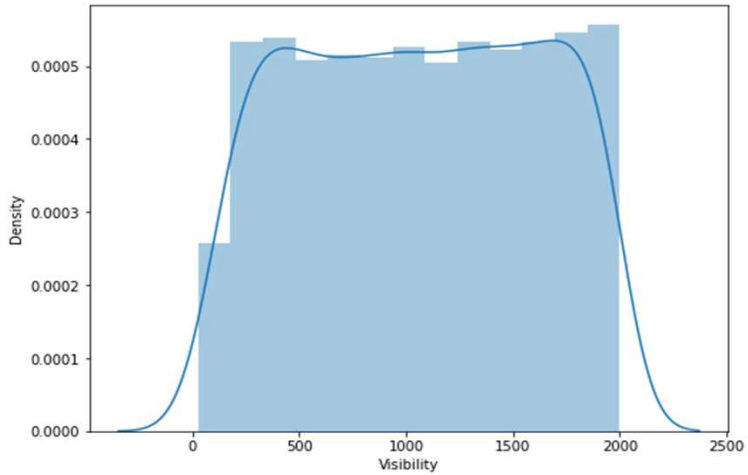
Numerical variables

Temperature



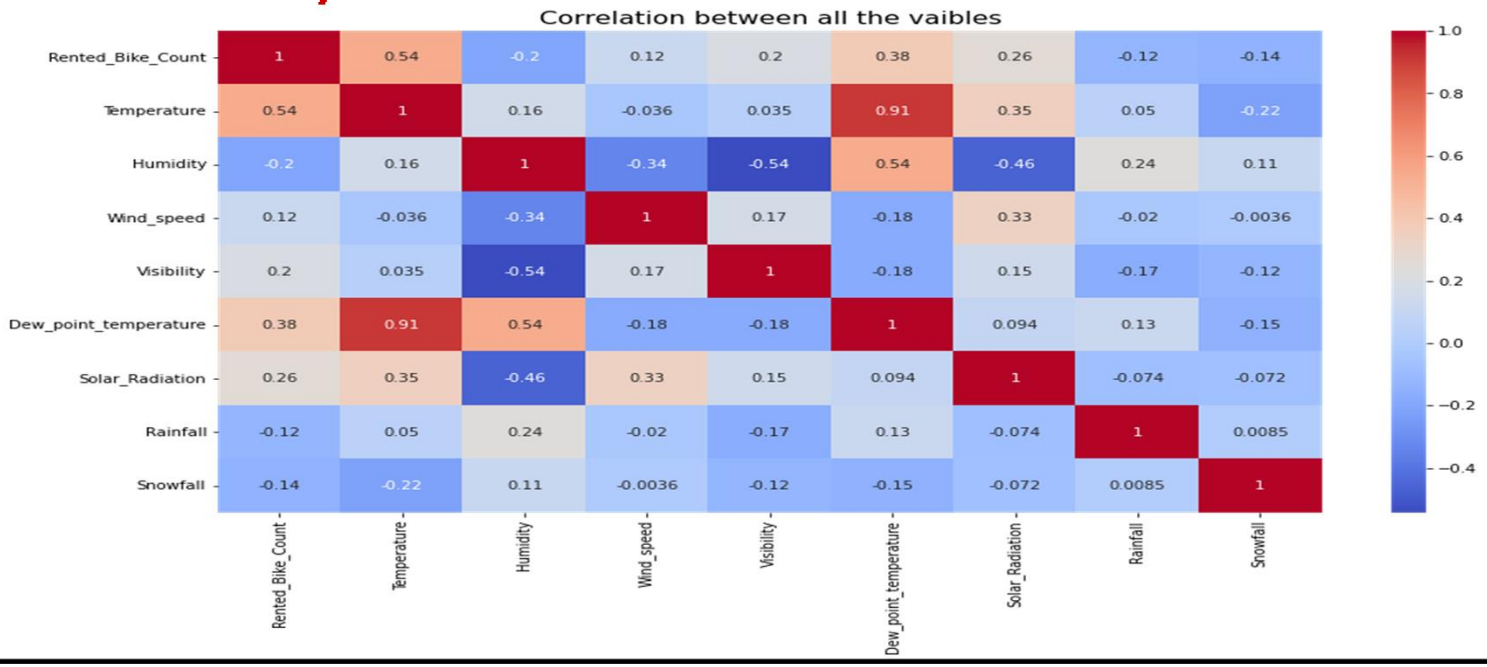
From above plot we conclude that when the temperature is between -5 to 25 degrees bikes are rented more.

visibility



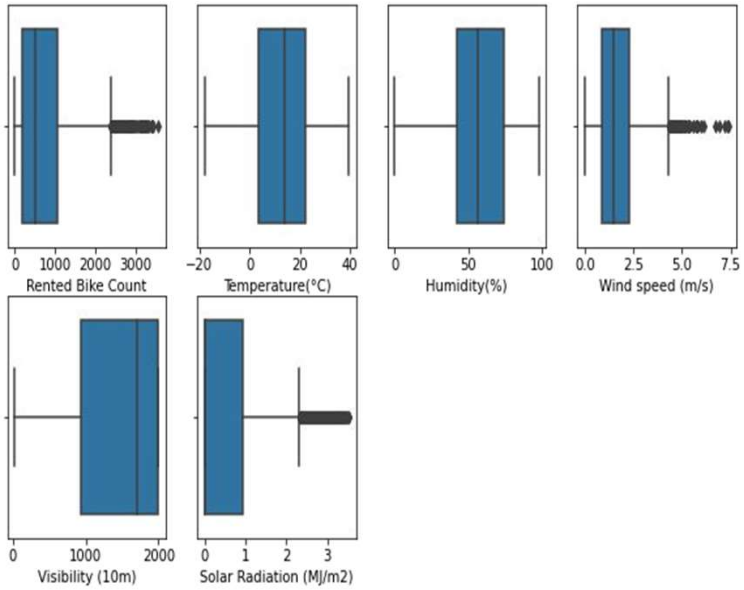
Above plot shows that people tend to rent bikes when the visibility is between 300 to 1700

Heat map



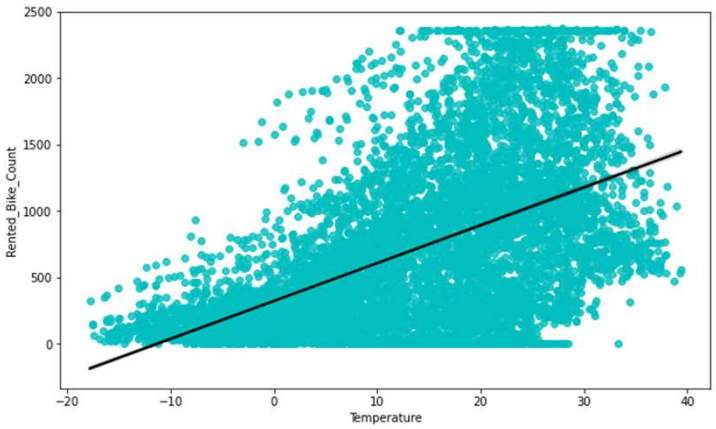
From the above heat map we conclude that temperature and dew point temperature are highly correlated so we need to drop one of them in order to give accurate prediction.

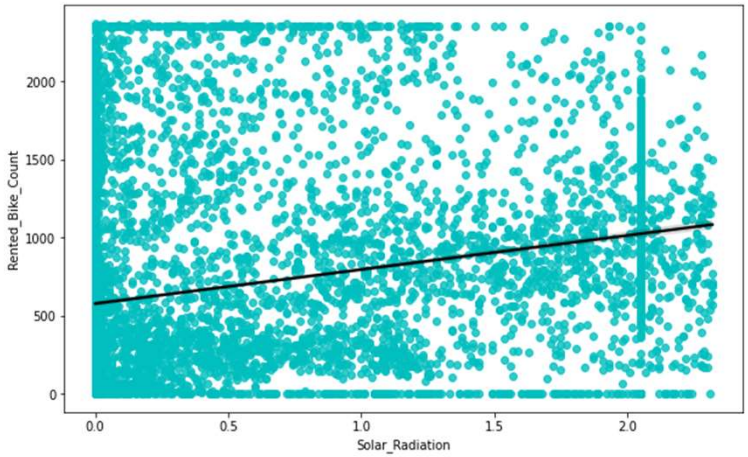
Handling outliers

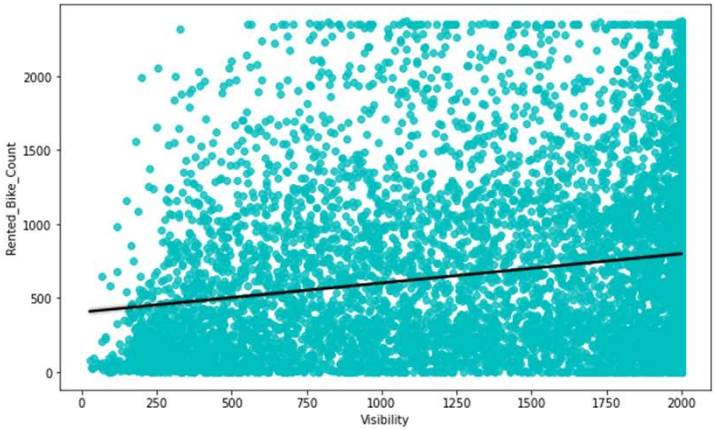


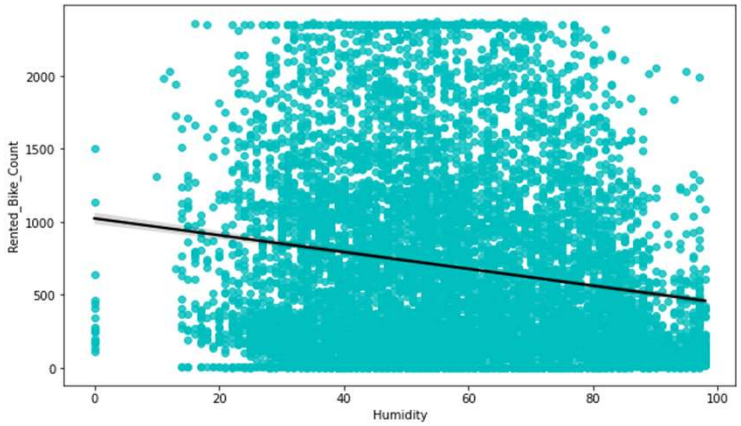
An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects,The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outlier

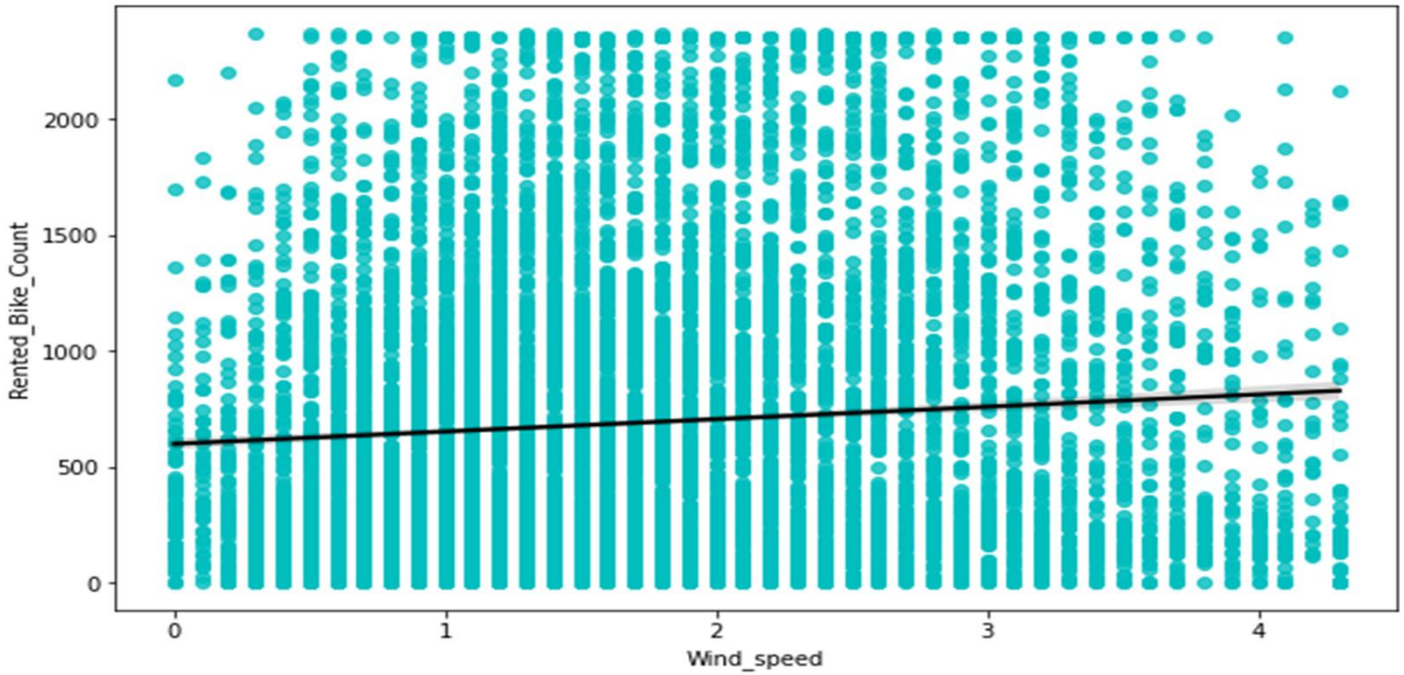
Regression plot











Temperature, Solar radiation, Visibity, Wind speed are positively correlated with target variable. Humidity is negatively correlated with target variable.
The rented bike count increases with increase in these features.

ML algorithms

1. Linear regression
2. Lasso regression
3. Ridge regression
4. Elastic net
5. Decision tree
6. Random Forest Regressor
7. XG boost Regressor

- # Linear regression

MSE : 0.3654626478005656

RMSE : 0.6045350674696759

MAE : 0.4248468842906076

R2 : 0.8630960604363349

Adjusted R2 : 0.8593199541220788

Ridge regularization

MSE : 0.3654639337955513
RMSE : 0.6045361310918904
MAE : 0.42484925264286016
R2 : 0.8630955786968716
Adjusted R2 : 0.8593194590952008

Decision tree

- MSE : 0.525793770695954
- RMSE : 0.725116384241836
- MAE : 0.5219409878865798
- R2 : 0.8030353059621238
- Adjusted R2 : 0.7976025943307974

Random Forest Regressor

MSE : 0.19504578269859701
RMSE : 0.44163987897221985
MAE : 0.2636480113398128
R2 : 0.9269349789713991
Adjusted R2 : 0.9249196878984272

XG Boosting Regressor with GridSearchCV

MSE :
0.17631659628695023
RMSE :
0.4199006981262954
MAE :
0.25232283149995133
R2 :
0.9339510158222456
Adjusted R2 :
0.9321292421976244

		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	**0**	Linear regression	0.431	0.385	0.621	0.842	0.84
	1	Lasso regression	0.559	0.560	0.748	0.770	0.76
	2	Lasso regression with CV	0.431	0.385	0.621	0.842	0.84
	3	Ridge regression	0.431	0.385	0.621	0.842	0.84
	4	Ridge regression with CV	0.431	0.385	0.621	0.842	0.84
	5	Elasticnet regression	0.434	0.388	0.623	0.841	0.84
	6	Decision tree regression	0.494	0.444	0.666	0.818	0.81
	7	Random forest regression	0.101	0.028	0.168	0.988	0.99
	8	XG Boost Regression	0.321	0.201	0.448	0.918	0.92
	9	XG boost regg GridserachCV	0.094	0.019	0.136	0.992	0.99

		Model	MAE	MSE	RMSE	R2	Adj_R2
Test set	**0**	Linear regression	0.425	0.365	0.605	0.863	0.86
	1	Lasso regression	0.553	0.550	0.742	0.794	0.79
	2	Lasso regression with cv	0.425	0.365	0.605	0.863	0.86
	3	Ridge regression	0.425	0.365	0.605	0.863	0.86
	4	Ridge regression with cv	0.425	0.366	0.605	0.863	0.86
	5	Elasticnet regression	0.428	0.369	0.607	0.862	0.86
	6	Decision tree regression	0.522	0.526	0.725	0.803	0.80
	7	Random forest regression	0.264	0.195	0.442	0.927	0.92
	8	XG Boost Regression	0.346	0.246	0.496	0.908	0.91
	9	XG boost regg GridserachCV	0.252	0.176	0.420	0.934	0.93

Challenges

- Treating the outliers in numerical features.
- Generation of new features which need to be added in the model.
- Choosing the right features for modelling.
- Choosing the right models to get the best scores.

Conclusion

- Hour of the day holds most importance among all the features for prediction of dataset
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
- the top important features of our dataset are: Season_winter, Temperature, Hour, Radiation, Humidity, Visibility
- Peoples don't use rented bikes in no functioning day
- people tend to rent bikes when the temperature is between -5 to 25 degrees
- people tend to rent bikes when the visibility is between 300 to 1700
- for all the above experiments we can conclude that random forest Regressor and XG Boost Regressor with using hyperparameters we got the best results.


thank you