Capstone project

(SUPERVISED ML – REGRESSION)

Bike Sharing Demand Prediction

BY

Snehal D. Ramteke

Content

- Business Understanding
- Problem statement
- Introduction
- Data Summary
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Implementing Algorithms
- Challenges
- Conclusions

Business Understanding

- Bike rentals have become a popular service in recent years and it seems people are using it more often. Relatively cheaper rates and ease of pick up and drop at own convenience is what make this business thrive.
- Mostly used by people having no personal vehicles and also to avoid congested public transport which that's why they prefer rental bikes.
- Therefore, for the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfill the demand.
- Our project goal is a pre-planned set of bike count values that can be a handy solution to meet all demands.

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 for concern.

Introduction

- Prediction of bike sharing demand can help bike sharing companies to allocate bikes better and ensure more sufficient circulation of bikes for customers.
- This presentation proposes a real-time method for predicting bike renting based on historical data, weather data, and time data.
- This demand prediction model can provide a significant theoretical basis for management strategies and vehicle scheduling in the public bike rental system.

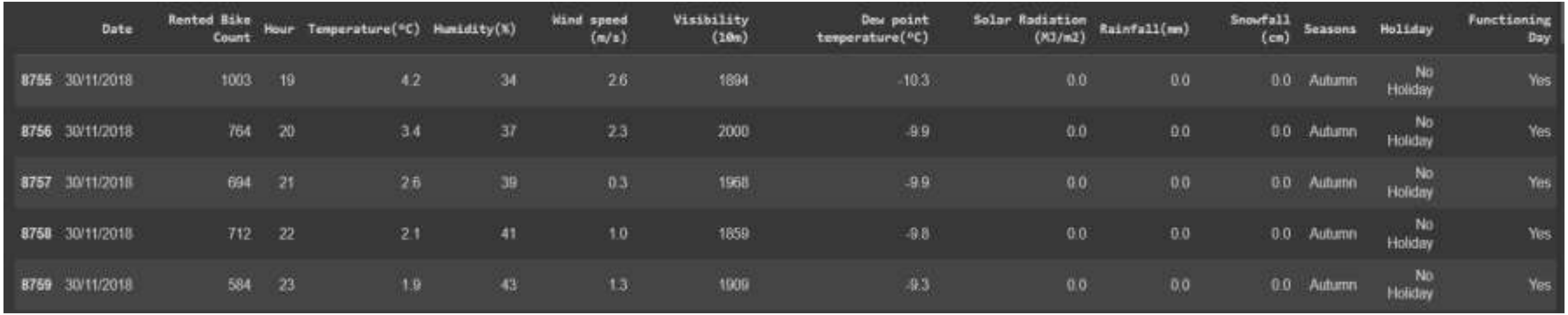
Data Summary

- The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information.
- Date: year-month-day
- Rented Bike count -Count of bikes rented at each hour
- Hour -Hour of the day
- Temperature-Temperature in Celsius
- Humidity -%
- Windspeed -m/s
- Visibility -10m

Data Summary (contd)

- Dew point temperature –Celsius
- Solar radiation -MJ/m2
- Rainfall -m
- Snowfall -cm
- Seasons -Winter, Spring, Summer, Autumn
- Holiday -Holiday/No holiday
- Functional Day -NoFunc(Non-Functional Hours), Fun(Functional hours)

Data Summary (contd)



	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 contains 8760 lines and 14 columns.
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘FunctioningDay’.
- One Datetime features a ‘Date’.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, and snowfall which tell the environmental conditions at that particular hour of the day.

Insight From our Dataset

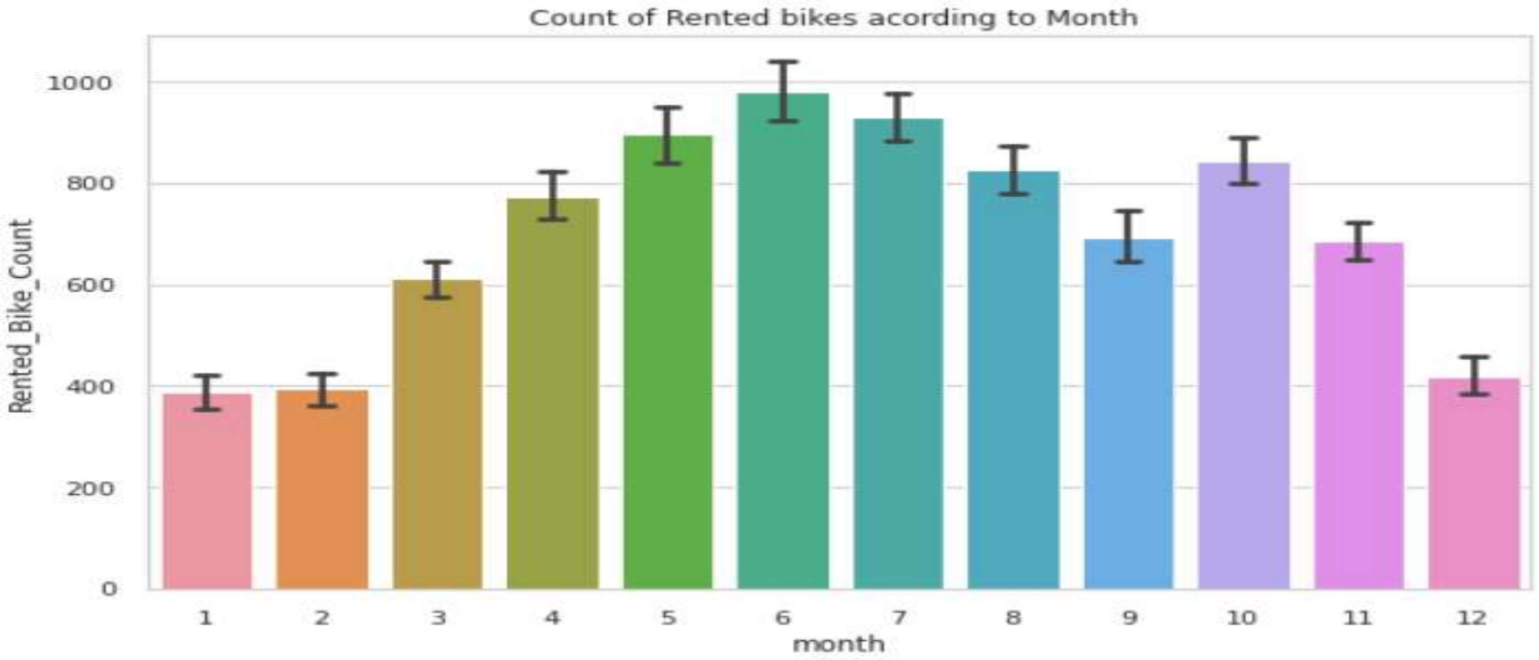
- There are No Missing Values present.
- There are No Duplicate values present.
- There are No null values.
- And finally we have the ‘rented bike count’ variable which we need to predict for new observations
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365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

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

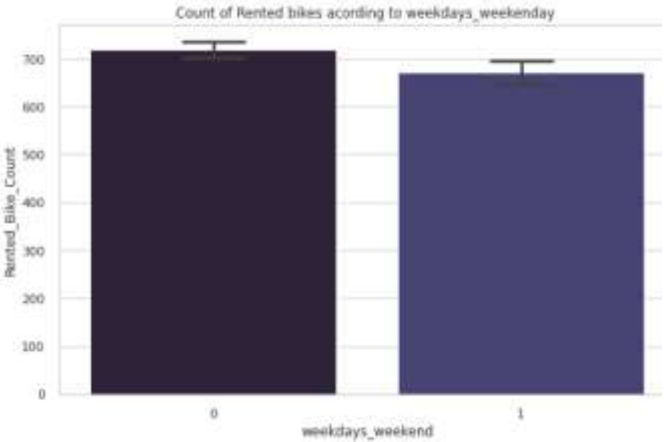
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis-testing task.

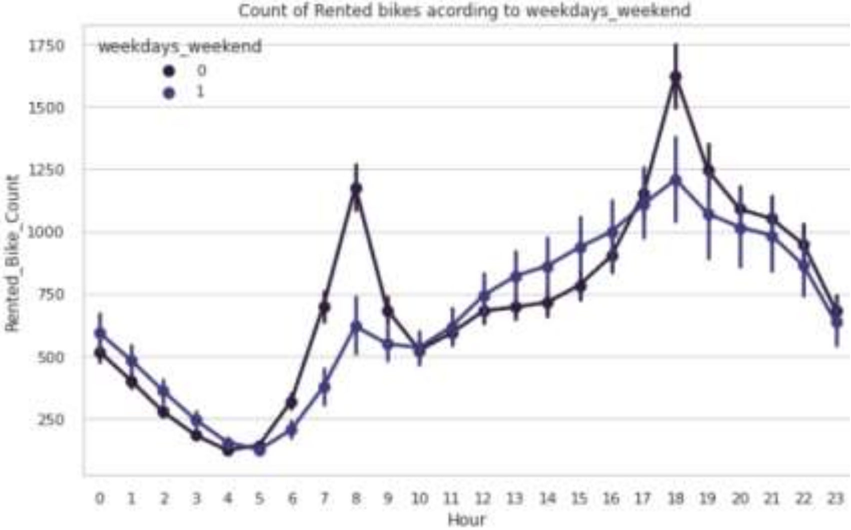
Analysis of Month Variable



- From the above bar plot we can clearly say that from the month 5 to 10 the demand for the rented bike is high as compared to other months. these months are comes during the summer season.

Analysis of weekdays_weekend variable



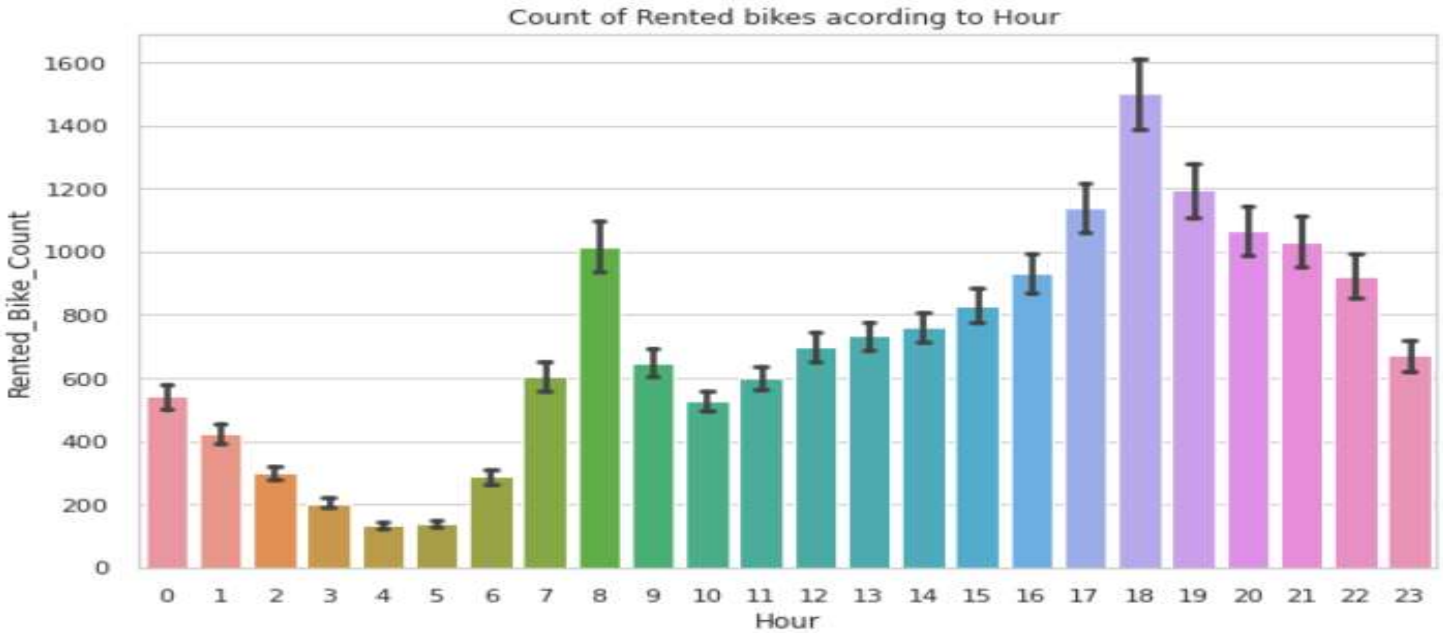


- From the above point plot and bar plot we can say that the weekdays which represent blue

The color shows that the demand for bikes is higher because of the office.

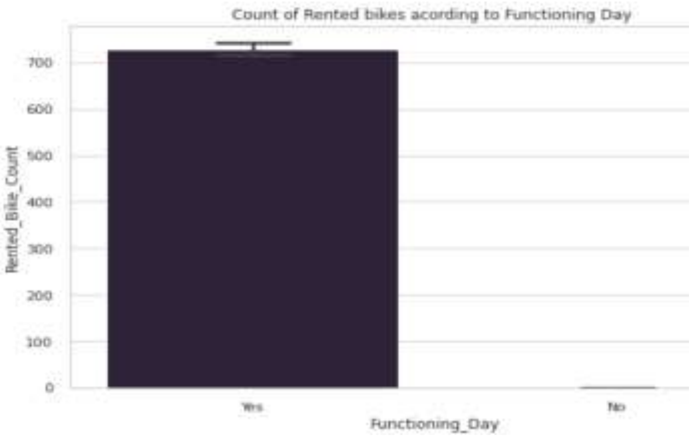
- Peak times are 7 am to 9 am and 5 pm to 7 pm.
- The orange color represents the weekend days, and it shows that the demand for rented bikes is very low, especially in the morning hour but when the evening starts from 4 pm to 8 pm the demand slightly increases.

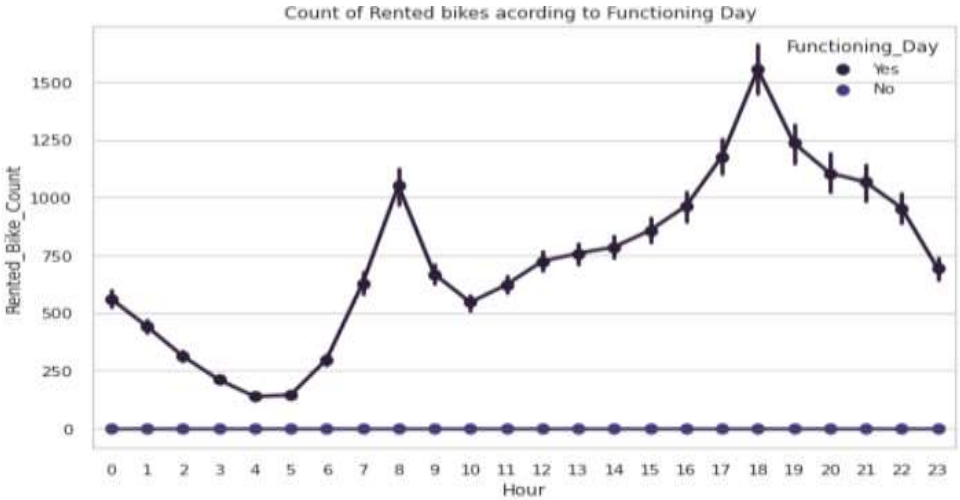
Analysis of Hour Variable



- In the above plot shows the use of rented bikes according to the hours and the data are from all over the year.
- generally people use rented bikes during their working hours from 7 am to 9 am and 5 pm to7pm.

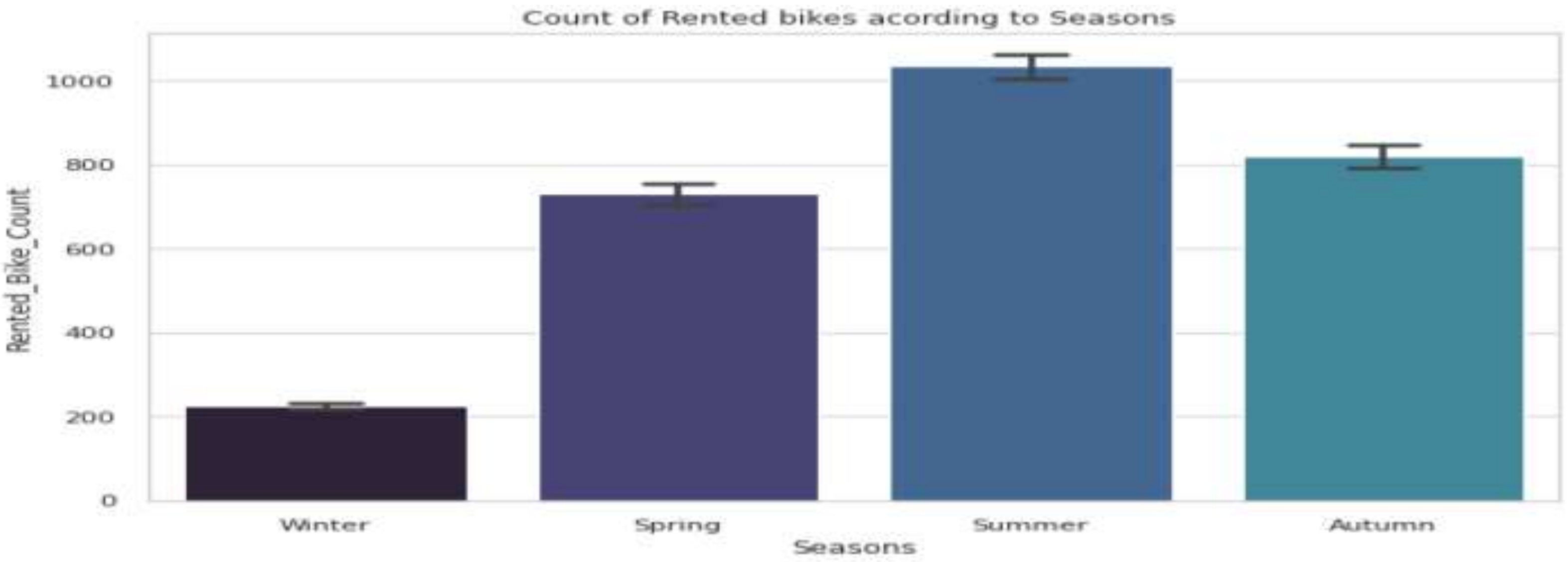
Analysis of Functioning Day Variable





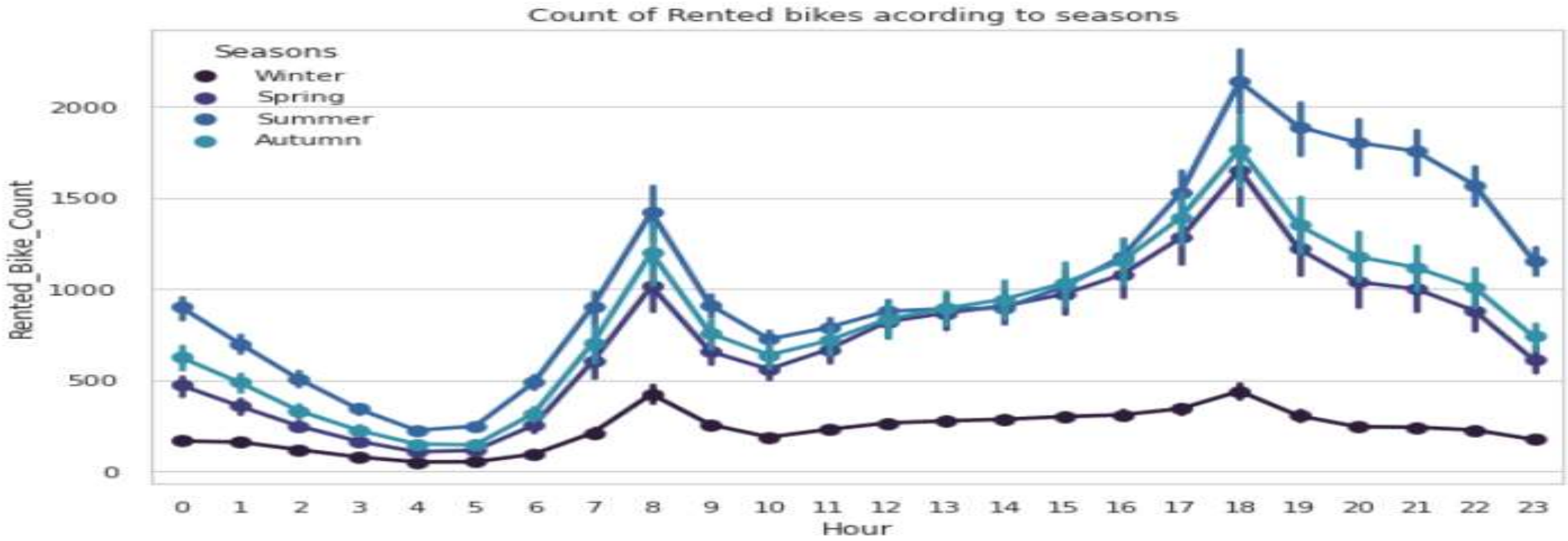
- In the above point plot shows the use of rented bikes on functioning days or not, and it clearly shows that.
- People don’t use rented bikes on the no-functioning day.

Analysis of Season Variable



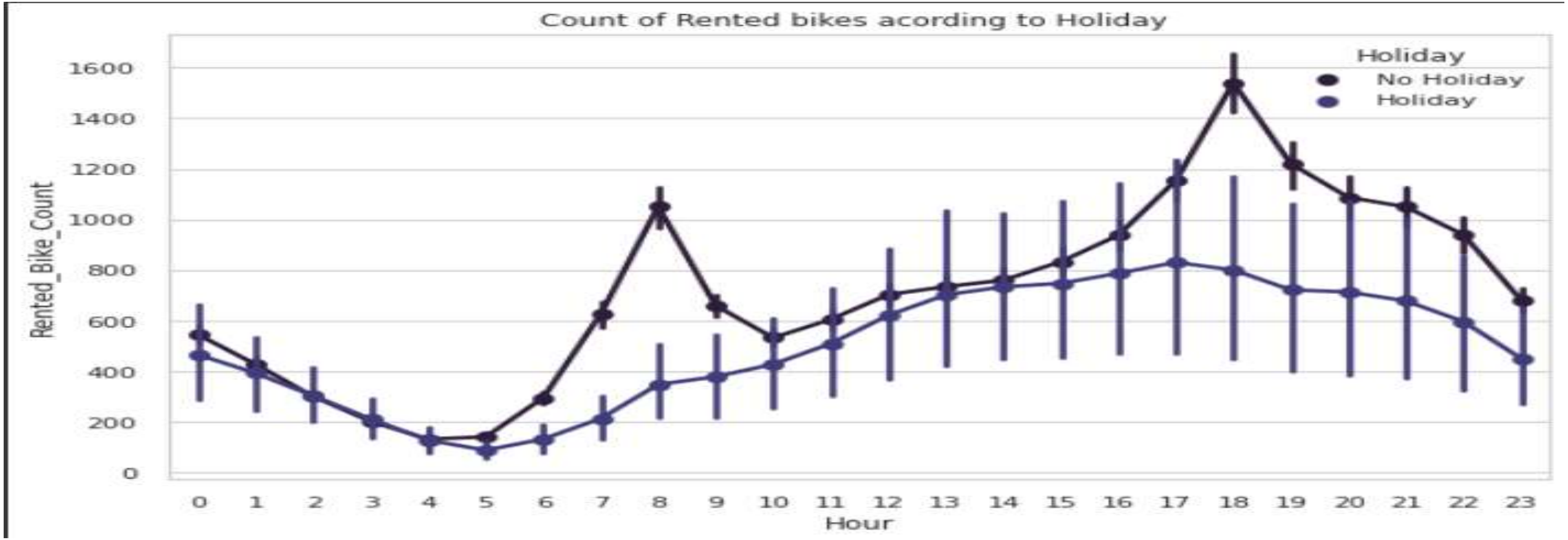
- This above bar plot shows the distribution of rented bike count season wise
- And we can clearly see that peoples love to ride bikes in the summer seasons and autumn seasons.
- But in the winter season people don't take any rented bikes due to because of snowfall

Season Variable(contd.)



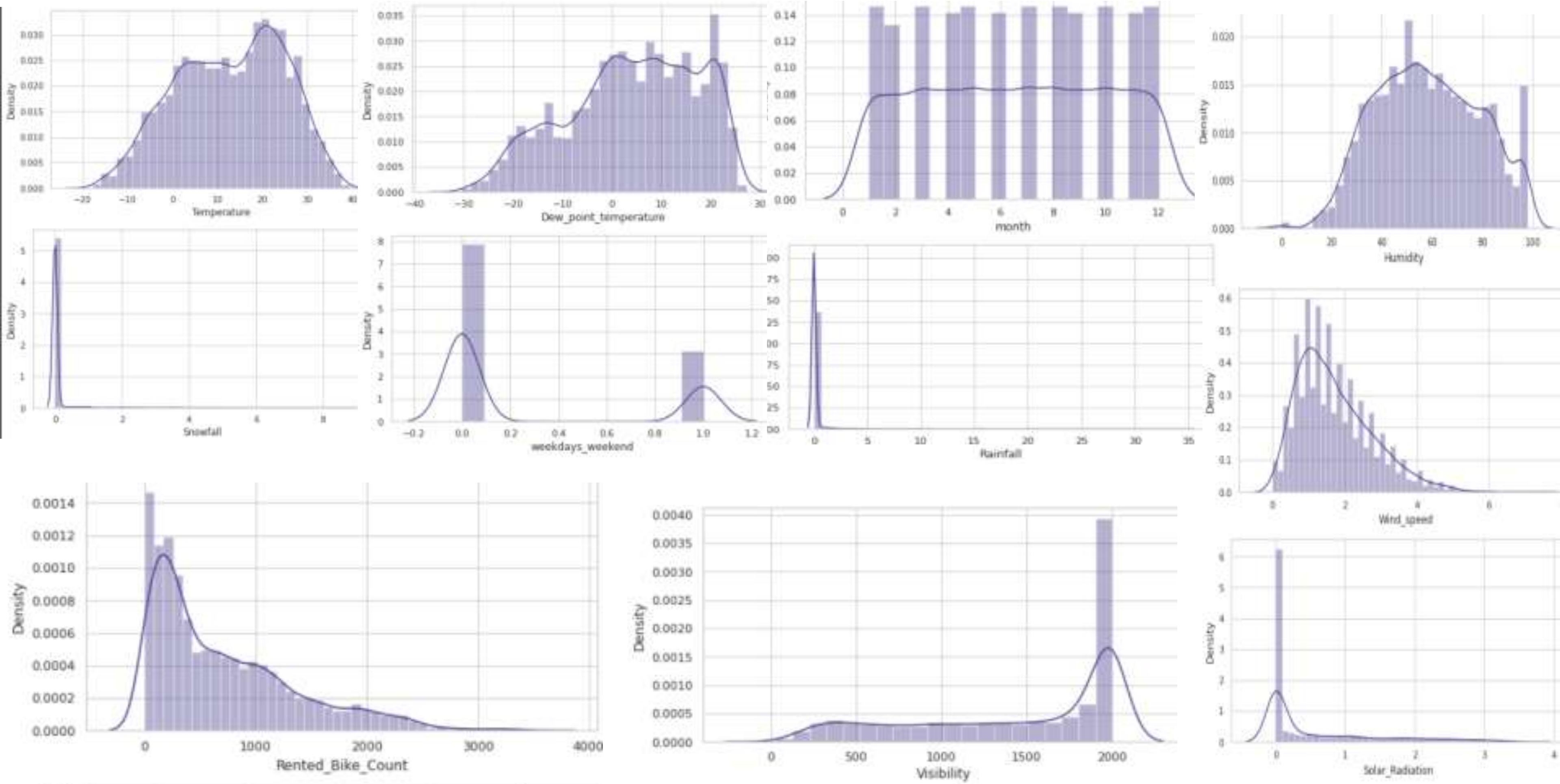
- In the above bar plot and point plot which shows the use of rented bikes in four different seasons, it clearly shows that,
- In the summer season the use of the rented bike is high and the peak time is 7 am-9 am and 7pm5pm.
- In the winter season the use of rented bikes is very low because of snowfall

Analysis of Holiday Variable



- In the above bar plot and point plot shows the use of a rented bike on a holiday, and it clearly shows that
- plot shows that on holiday people use the rented bike from 2pm-8pm.

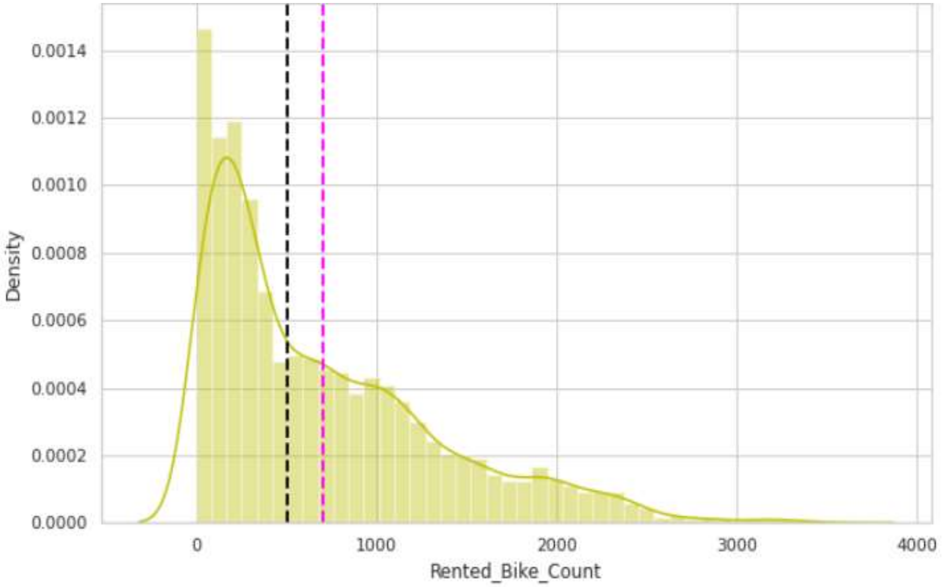
Visualizing Distribution

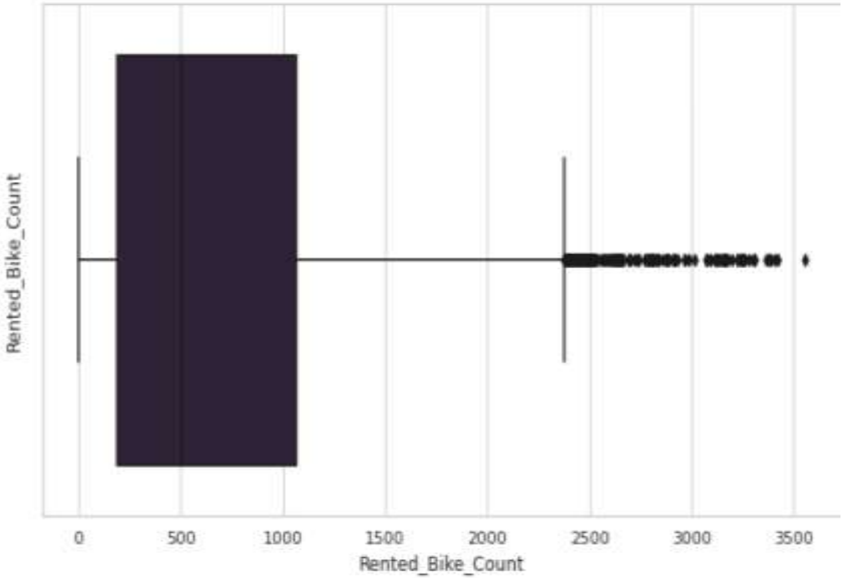


Analysis of Visualization distribution

- “Temperature”, “Hour”, “Month” and “Humidity” columns follow a uniform distribution.
- “Wind Speed”, “Solar Radiation”, “Rainfall” and “Snowfall” are having positively skewed distribution.
- “Visibility” column is negatively skewed

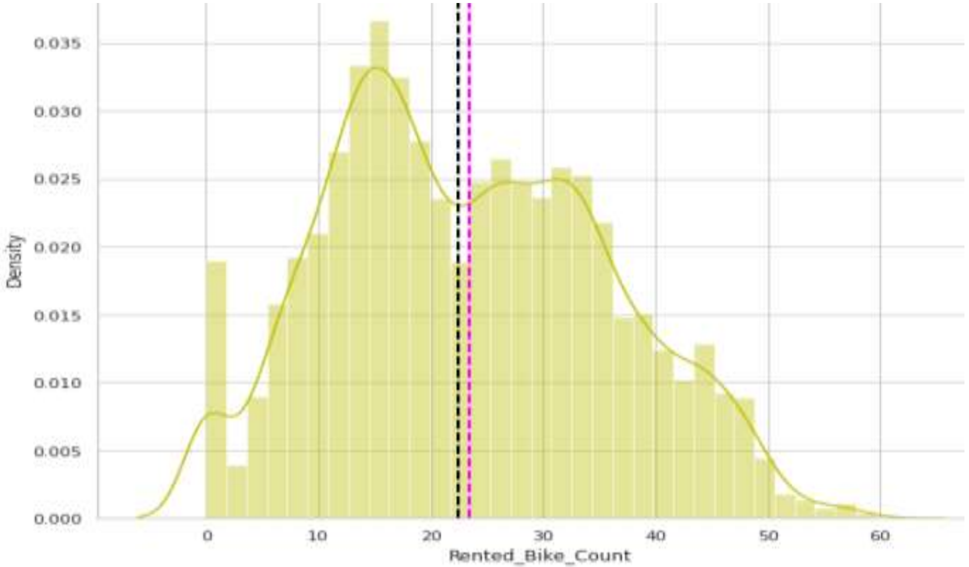
Analysis of Rented Bike Column

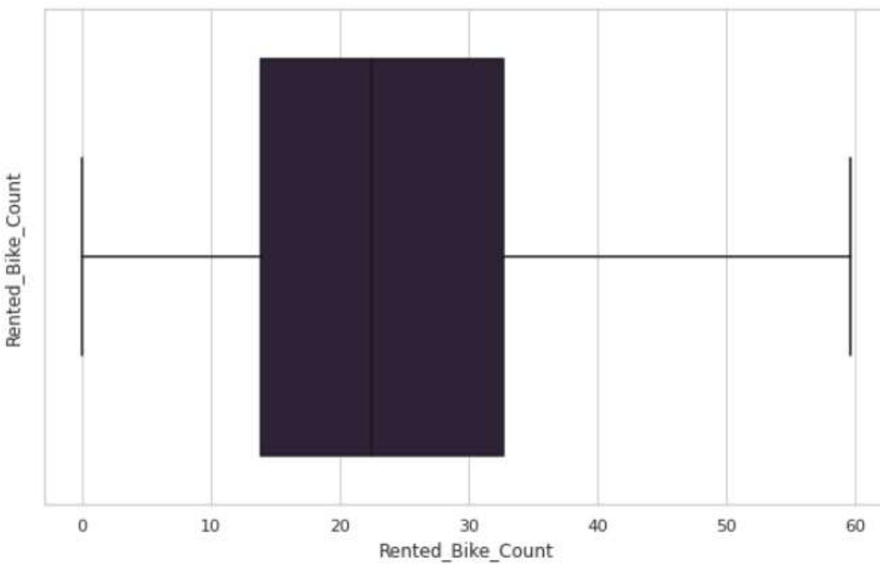




- The above graph shows that the Rented Bike Count has moderate right skewness.
- The above boxplot shows that we have detect outliers in Rented Bike Count column
- Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform Square root operation to make it normal

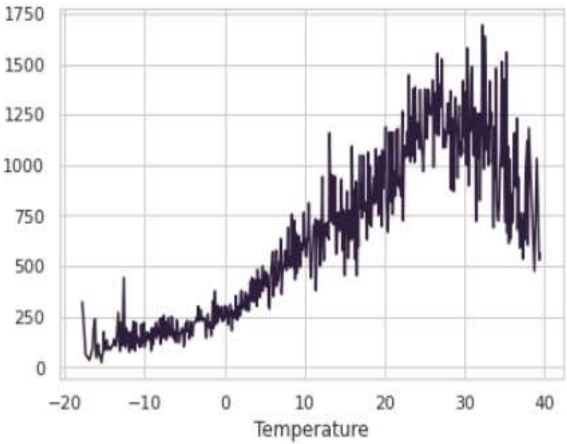
Analysis of Rented Bike Column

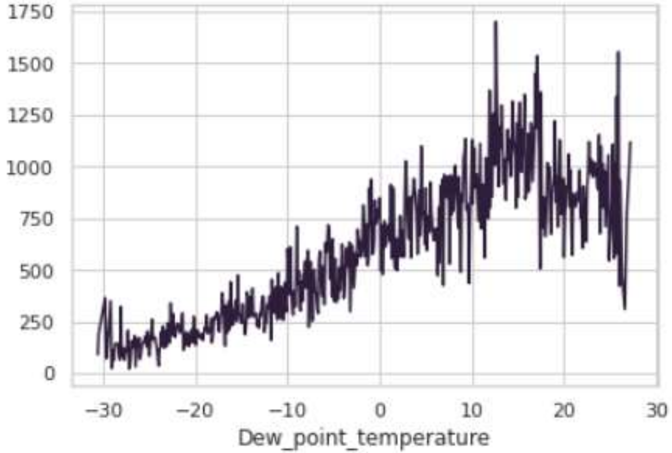


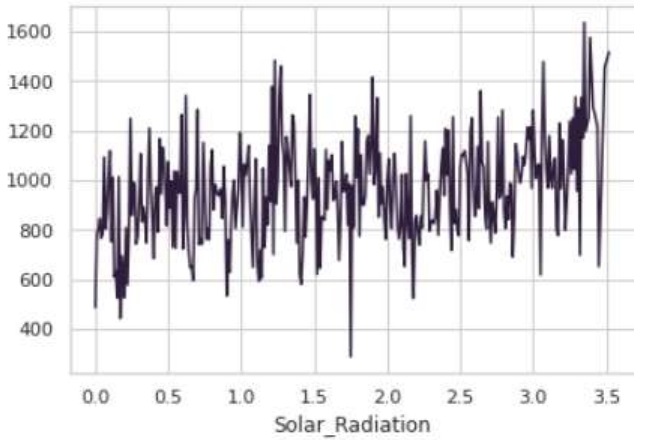


- After applying Square root to the skewed Rented Bike Count, here we get an almost normal distribution.
- After applying Square root to the Rented Bike Count column, we find that there are no outliers present.

Numerical vs Rented Bike Count

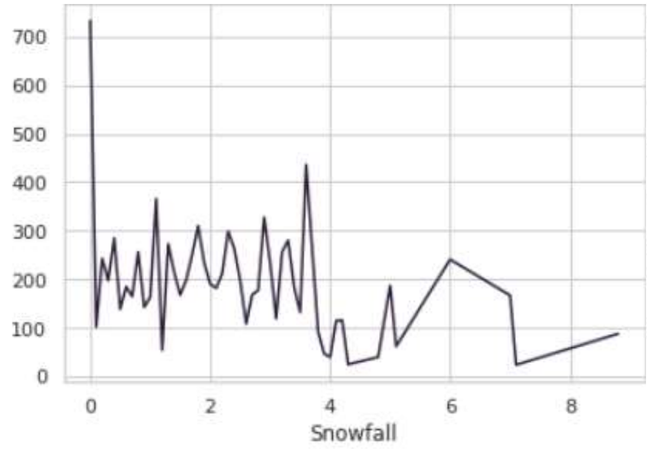


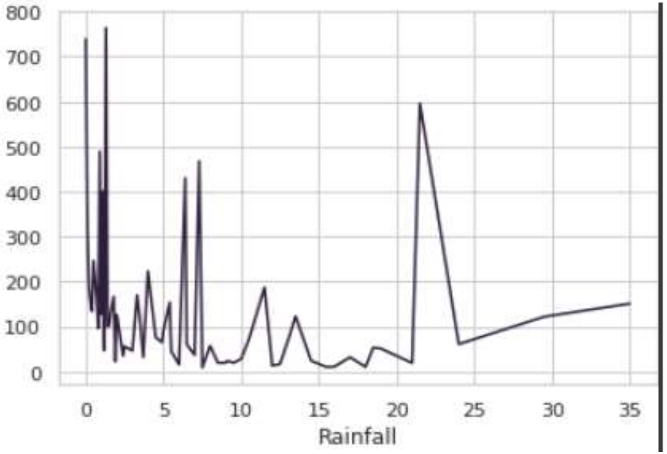


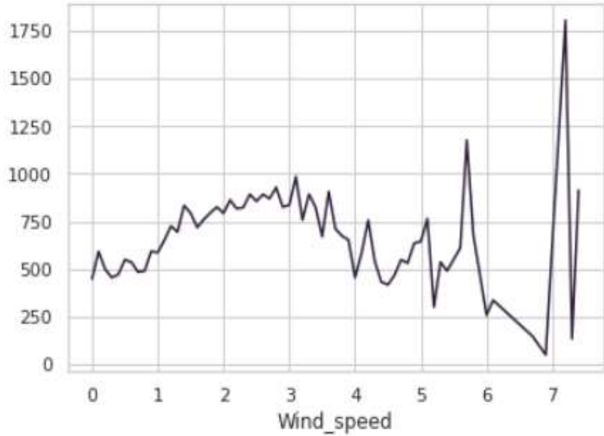


- From the above plot we see that people like to ride bikes when it is pretty hot around 25°C on average.
- From the above plot of "Dew_point_temperature' is almost the same as the ‘temperature there is some similarity present we can check it in our next step.
- From the above plot we see that, the amount of rented bikes is huge, when there is solar radiation, the counter of rents is around1000.

Numerical vs Rented Bike Count

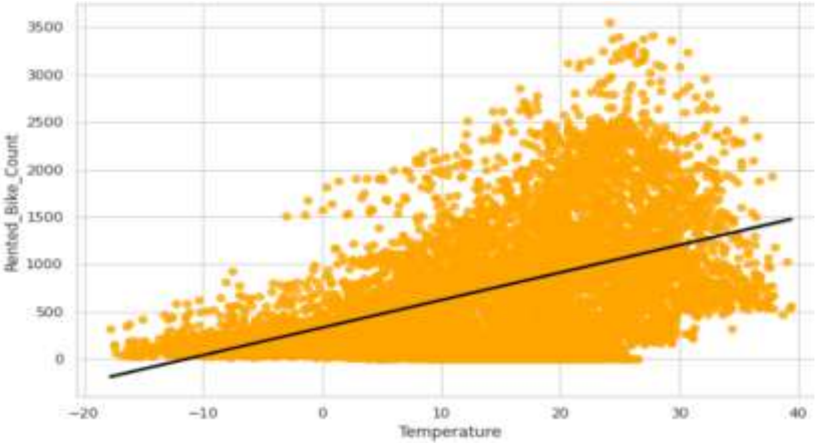


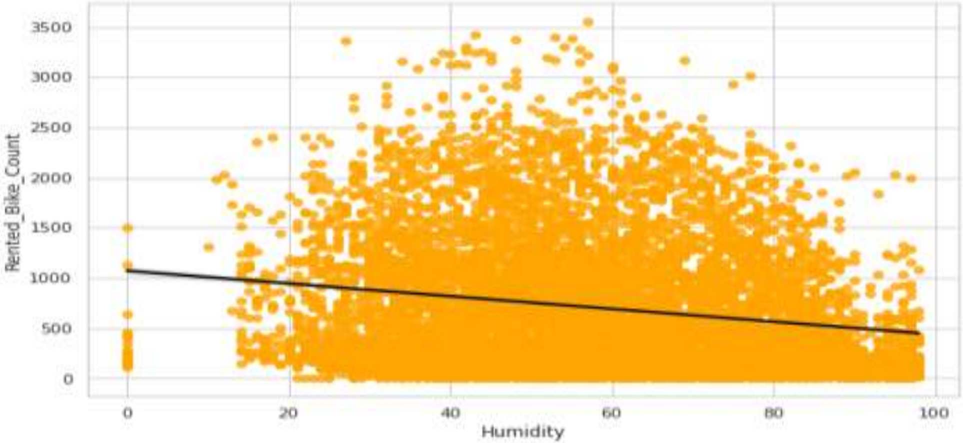


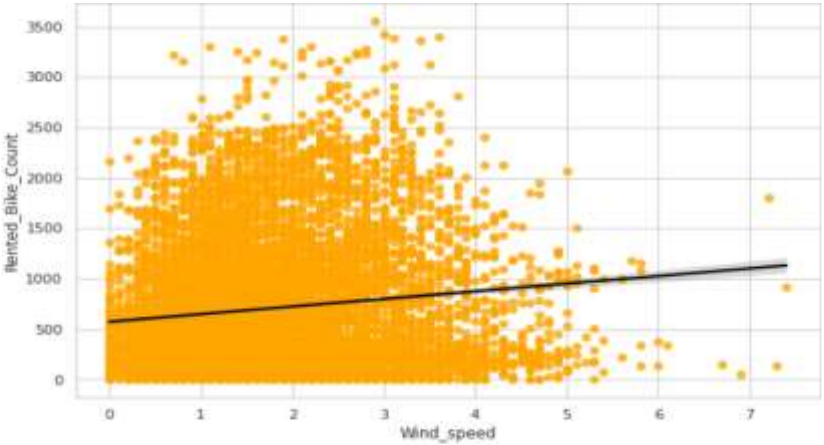


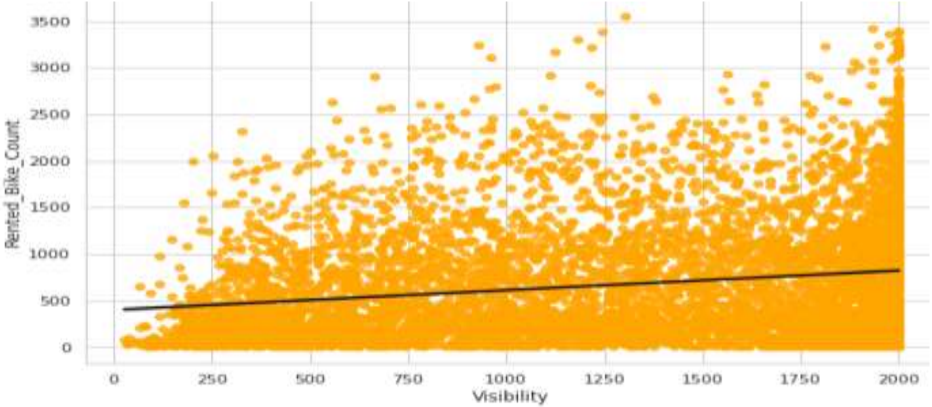
- n snowfall plot, on the y-axis, the number of rented bikes are very low When we have more than 4 cm of snow, the bike rents are much lower
- In rainfall plot if it rains a lot the demand for of rental bikes is not decreasing, here for example even if we have 20mm of rain there is a big peak of rented bikes
- In the wind speed plot that the demand for rented bikes is uniformly distributed despite wind speed but when the speed of the wind was 7 m/s then the demand for bikes also increase which clearly means peoples love to ride bikes when it’s a little windy.

Regression plot for Numerical Variable

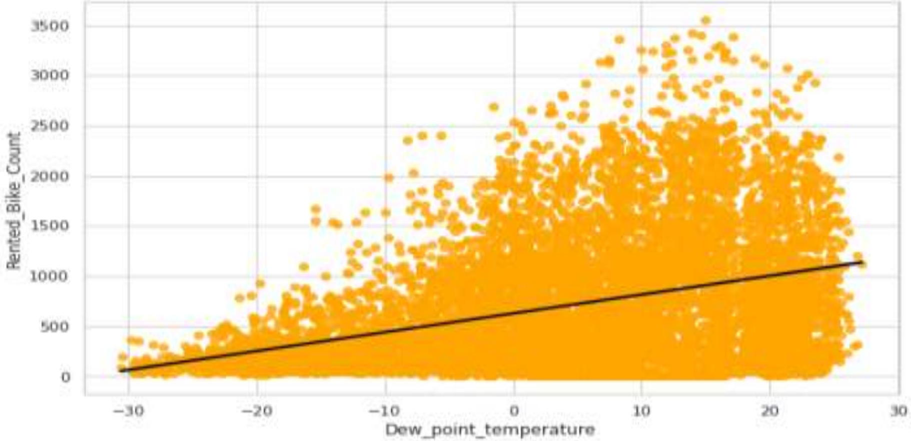


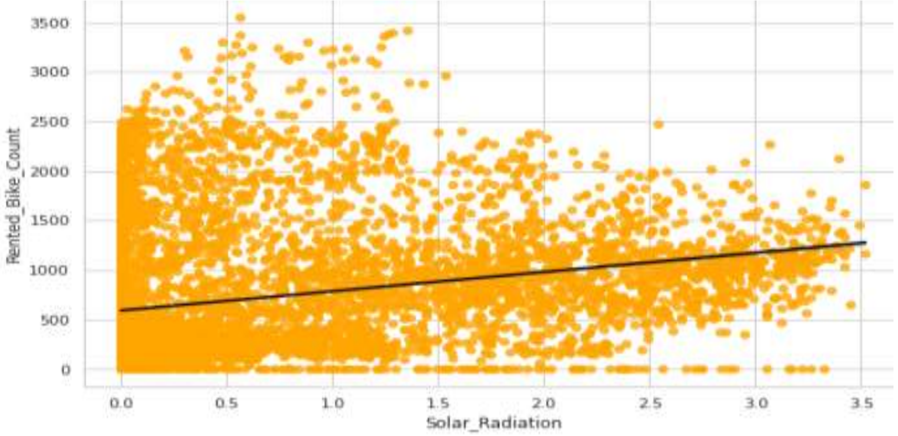


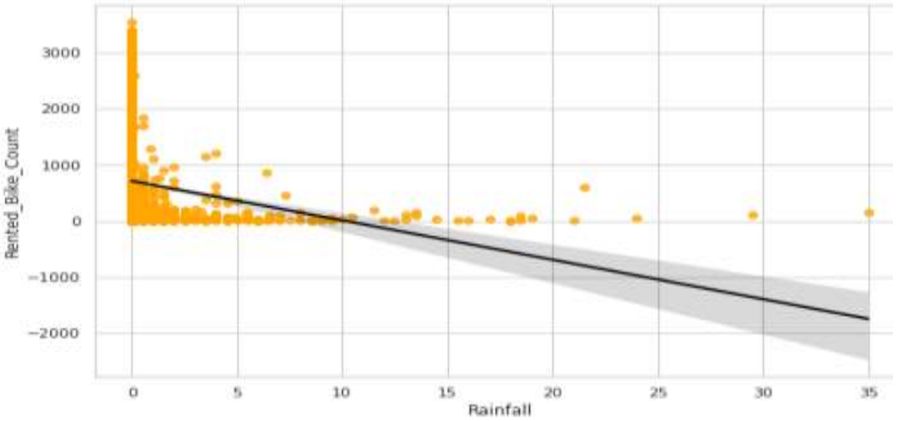


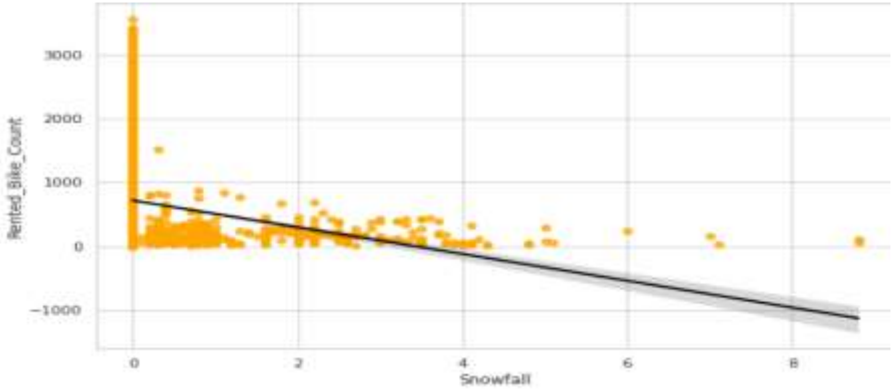


Regression plot for Numerical Variable





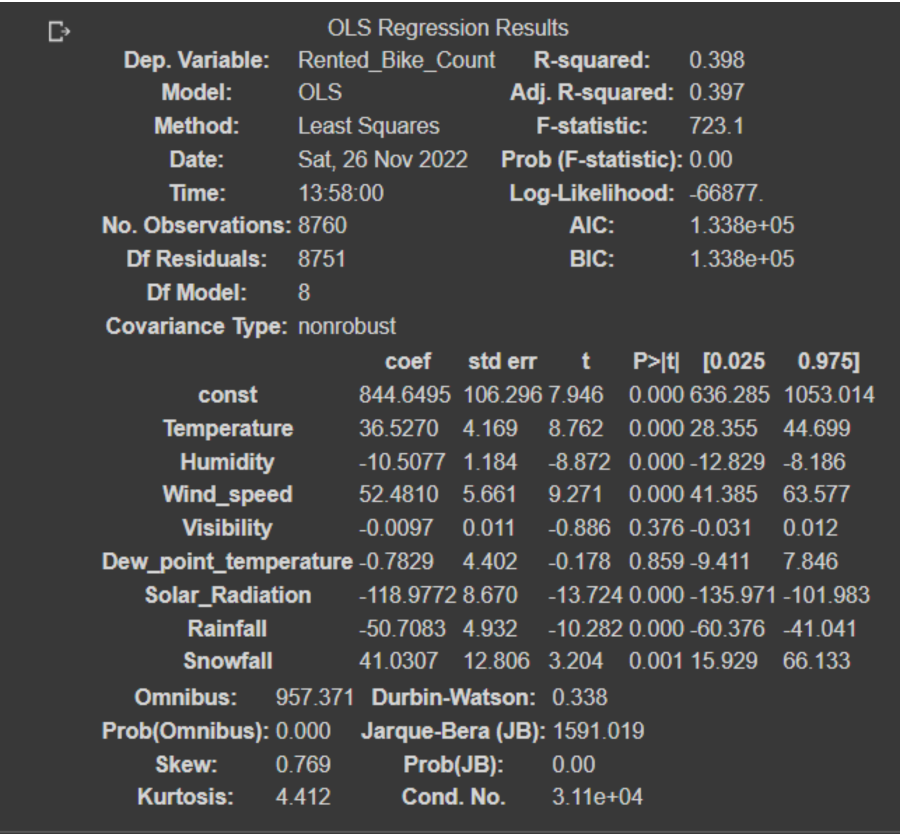




Regression plot for Numerical Variable

- From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed', 'Visibility', 'Dew_point_temperature', and 'Solar_Radiation' are positively related to the target variable.
- This means the rented bike count increases with the increase of these features.
- 'Rainfall', 'Snowfall', 'Humidity' these features are negatively related to the target variable which means the rented bike count decreases when these features increases.

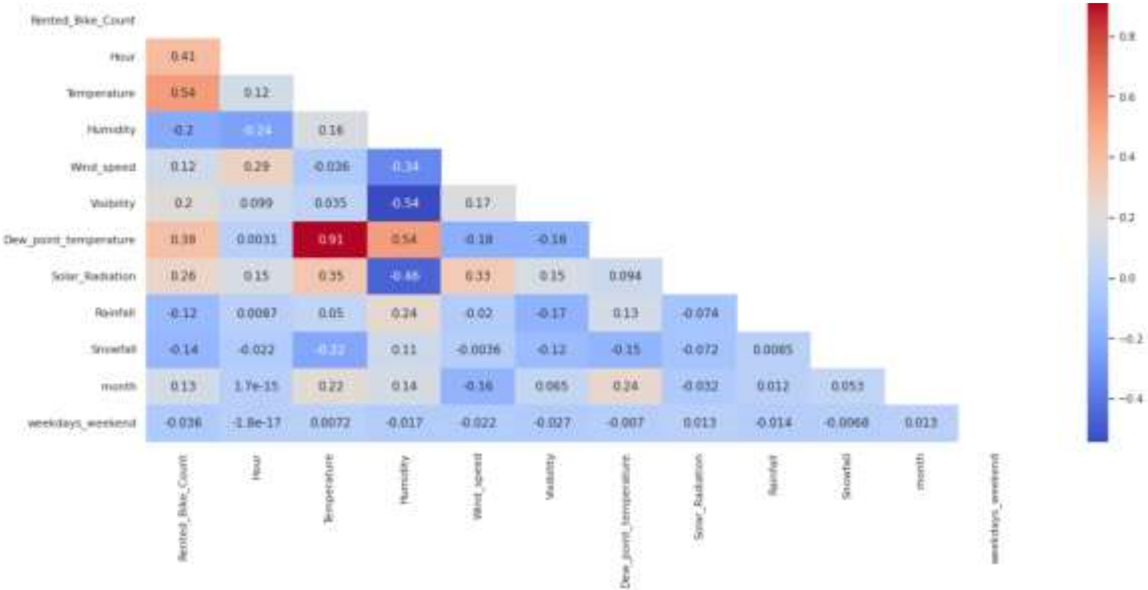
OLS Regression Model



```
                         OLS Regression Results
      Dep. Variable:    Rented_Bike_Count    R-squared:          0.398
          Model:        OLS                Adj. R-squared:       0.397
         Method:        Least Squares        F-statistic:        723.1
          Date:         Sat, 26 Nov 2022  Prob (F-statistic):    0.00
          Time:         13:58:00           Log-Likelihood:       -66877.
   No. Observations:    8760                     AIC:            1.338e+05
     Df Residuals:      8751                     BIC:            1.338e+05
       Df Model:        8
   Covariance Type:     nonrobust
                           coef     std err    t       P>|t|  [0.025    0.975]
         const           844.6495  106.296  7.946    0.000  636.285   1053.014
      Temperature         36.5270    4.169   8.762    0.000   28.355     44.699
       Humidity          -10.5077    1.184  -8.872    0.000  -12.829    -8.186
      Wind_speed          52.4810    5.661   9.271    0.000   41.385     63.577
       Visibility         -0.0097    0.011  -0.886    0.376   -0.031     0.012
  Dew_point_temperature   -0.7829    4.402  -0.178    0.859   -9.411     7.846
    Solar_Radiation     -118.9772    8.670  -13.724   0.000 -135.971  -101.983
        Rainfall         -50.7083    4.932  -10.282   0.000  -60.376   -41.041
        Snowfall          41.0307   12.806   3.204    0.001   15.929     66.133
     Omnibus:       957.371   Durbin-Watson:     0.338
  Prob(Omnibus):    0.000   Jarque-Bera (JB):   1591.019
       Skew:        0.769         Prob(JB):      0.00
     Kurtosis:      4.412        Cond. No.       3.11e+04
```

- R Square and Adj Square are near each other. 40% of the variance in the Rented Bike count is explained by the model.
- For the F statistic, the P value is less than 0.05 for a 5% level of significance.
- P value of dew point temp and visibility are very high and they are not significant.
- Omnibus tests the skewness and kurtosis of the residuals. Here the value of Omnibus is high., which shows we have skewness in our data.
- The condition number is large, 3.11e+04. This might indicate that there are strong multicollinearity or other numerical problems.
- Durbin-Watson tests for autocorrelation of the residuals. Here value is less than 0.5. We can say that there exists a positive autocorrelation among the variables.

Correlation Matrix



We can observe on the heatmap that on the target variable line the most positively correlated variables to the rent are :

- the temperature
- the dew point temperature
- the solar radiation

And most negatively correlated variables are:

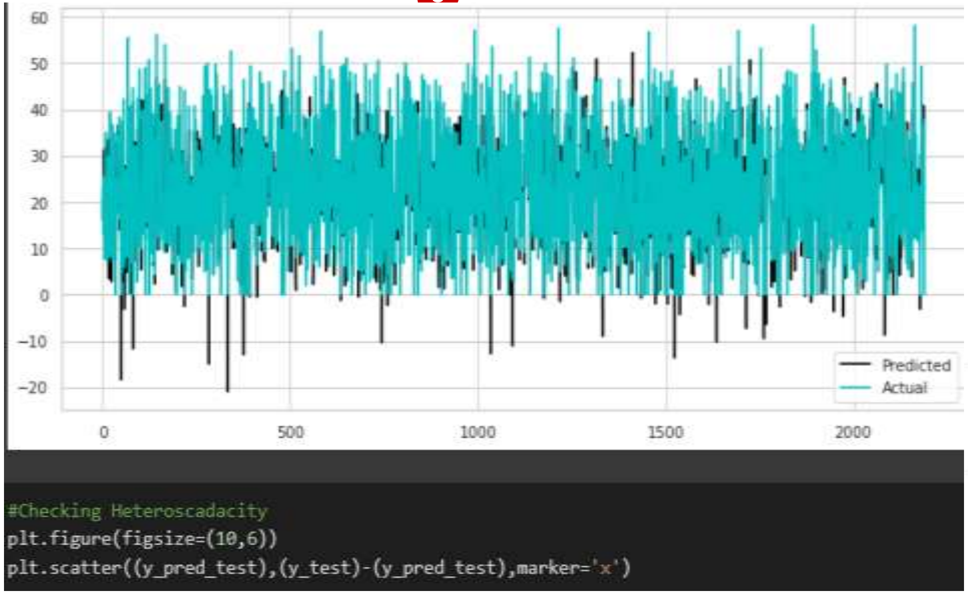
- Humidity
- Rainfall

- From the above correlation heatmap, We see that there is a positive correlation between columns 'Temperature' and 'Dew point temperature' i.e 0.91 so even if we drop this column then it doesn't affects the outcome of our analysis. And they have the same variations.. so we can drop the column 'Dew point temperature(°C)'.

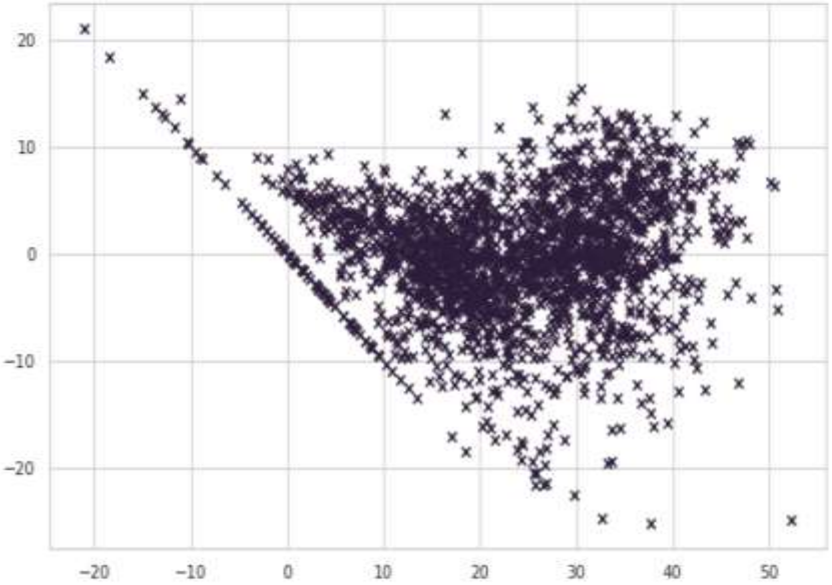
Fitting Various Model

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net Regression
5. Decision Tree Regression
6. Random Forest Regression
7. XG Boost Regression
8. XG boost Regressor with GridSearchCV

Linear Regression

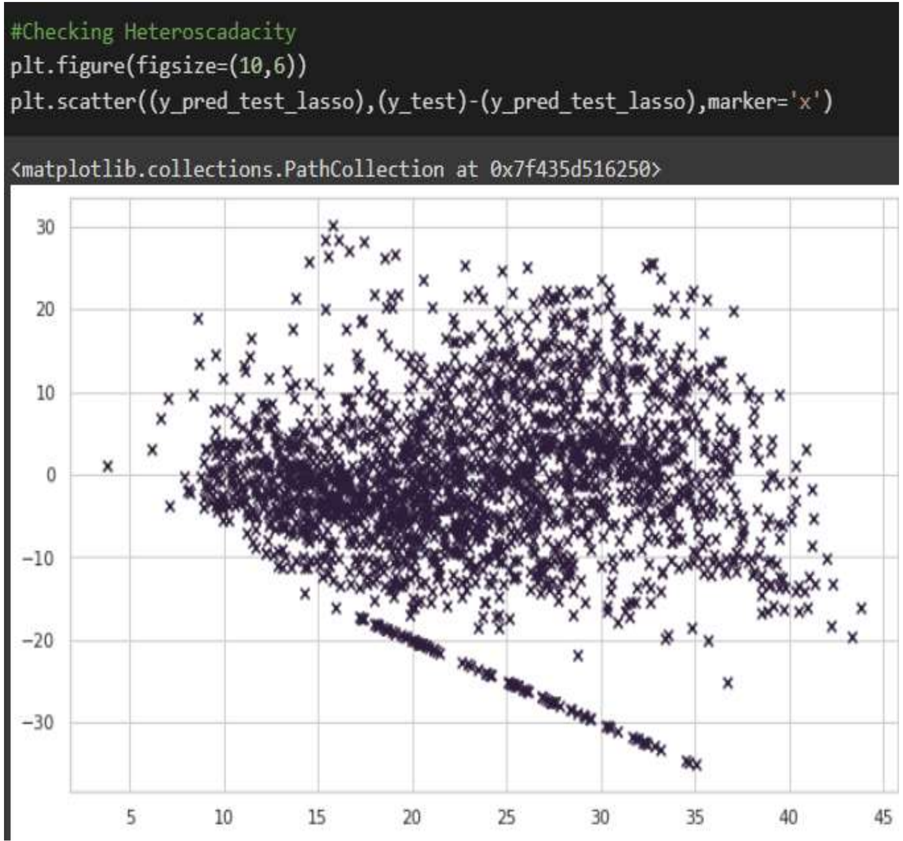


```
#Checking Heteroscadacity
plt.figure(figsize=(10,6))
plt.scatter((y_pred_test),(y_test)-(y_pred_test),marker='x')
```



- Looks like our r2 score value is 0.77 which means our model is able to capture most of the data variance.
- The r2_score for the test set is 0.78 which means our linear model is performing well on the data.

Lasso Regression

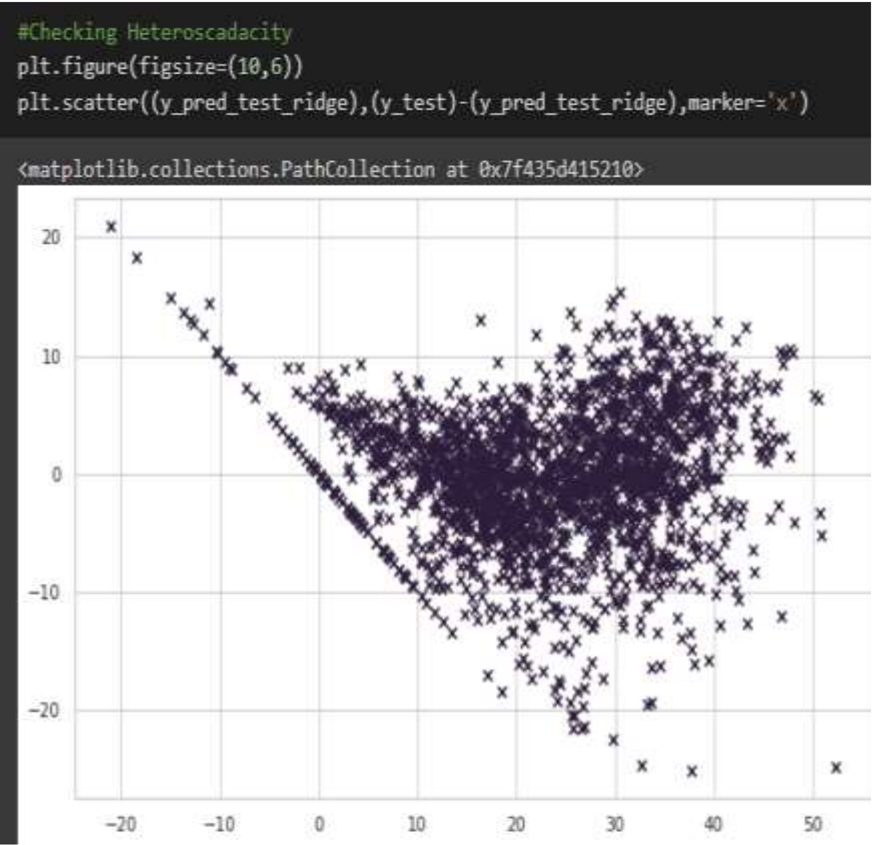


```
#Checking Heteroscadacity
plt.figure(figsize=(10,6))
plt.scatter((y_pred_test_lasso),(y_test)-(y_pred_test_lasso),marker='x')
```

```
<matplotlib.collections.PathCollection at 0x7f435d516250>
```

- Looks like our r2 score value is 0.40 which means our model is able to capture most of the data variance.
- The r2_score for the test set is 0.38 which means our linear model is performing well on the data.

Ridge Regression

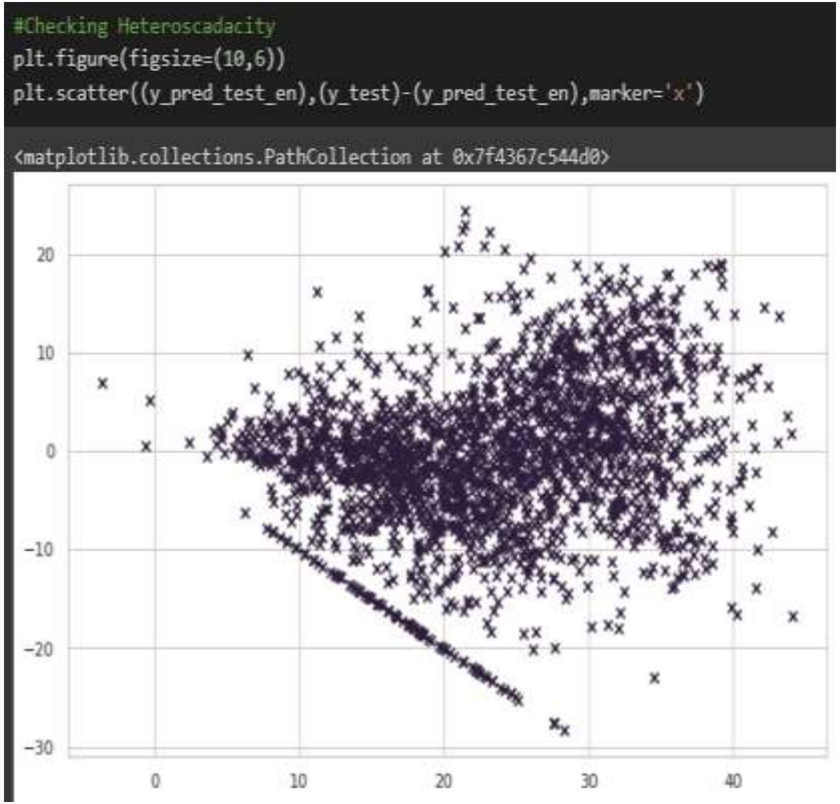


```
#Checking Heteroscadacity
plt.figure(figsize=(10,6))
plt.scatter((y_pred_test_ridge),(y_test)-(y_pred_test_ridge),marker='x')
```

```
<matplotlib.collections.PathCollection at 0x7f435d415210>
```

- Looks like our r2 score value is 0.77 which means our model is able to capture most of the data variance.
- The r2_score for the test set is 0.78 which means our linear model is performing well on the data.

Elastic Net Regression

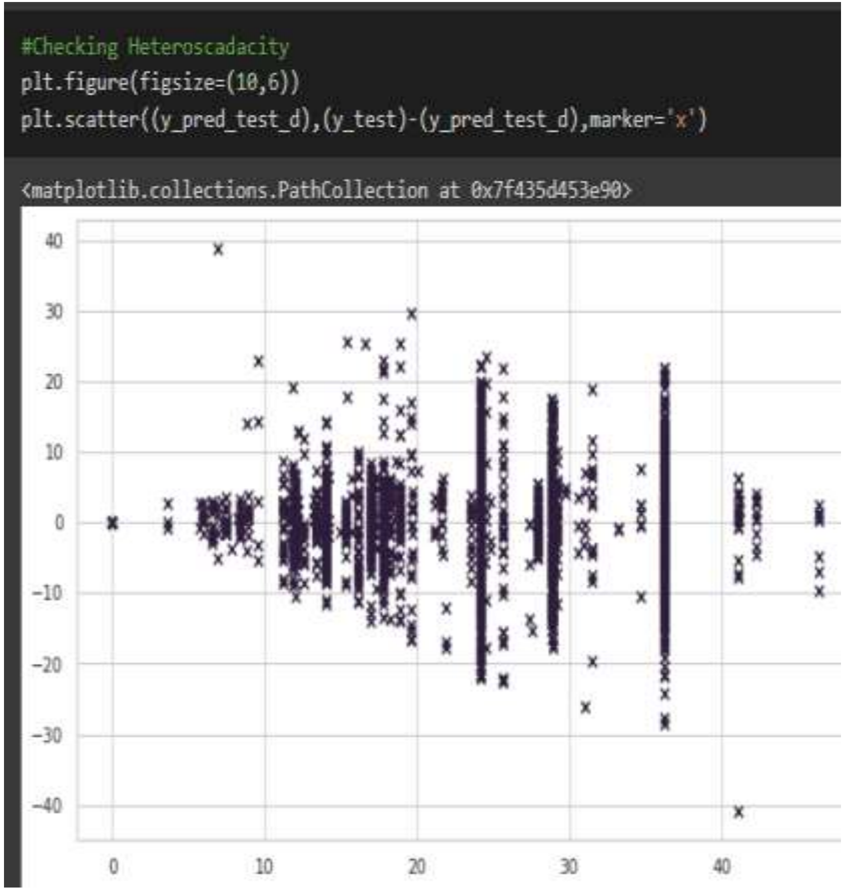


```
#Checking Heteroscadacity
plt.figure(figsize=(10,6))
plt.scatter((y_pred_test_en),(y_test)-(y_pred_test_en),marker='x')
```

```
<matplotlib.collections.PathCollection at 0x7f4367c544d0>
```

- Looks like our r2 score value is 0.62 which means our model is able to capture most of the data variance.
- The r2_score for the test set is 0.86 which means our linear model is performing well on the data.

Decision Tree Regression

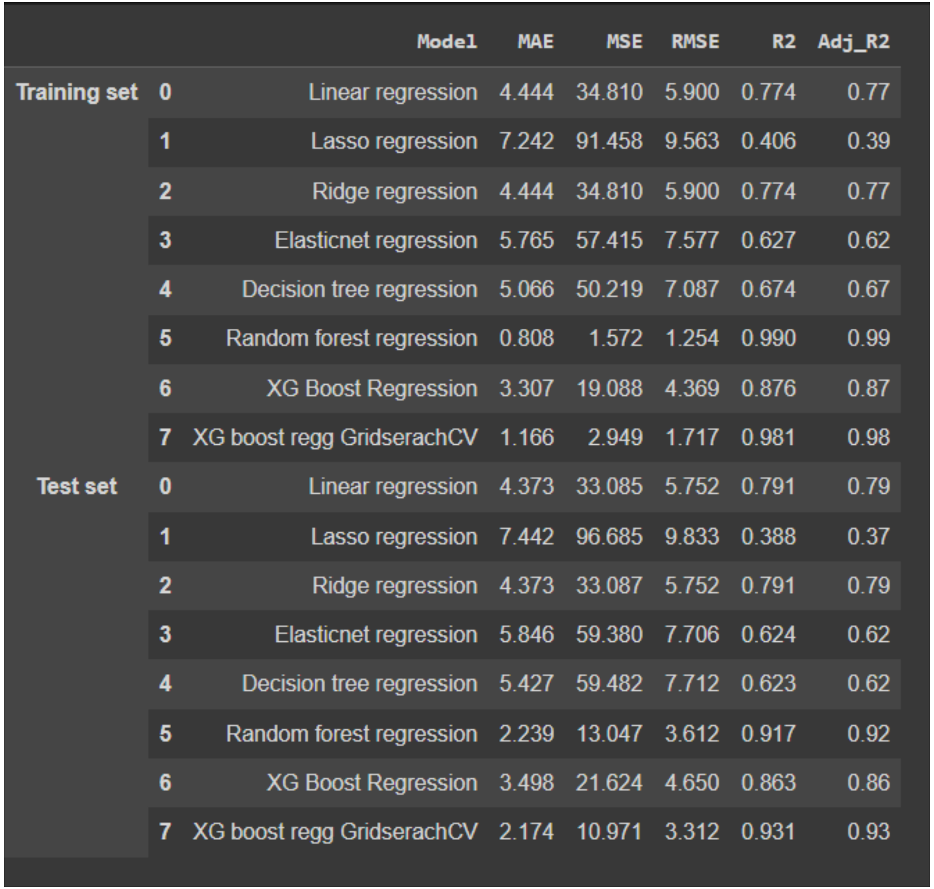


```
#Checking Heteroscadacity
plt.figure(figsize=(10,6))
plt.scatter((y_pred_test_d),(y_test)-(y_pred_test_d),marker='x')
```

```
<matplotlib.collections.PathCollection at 0x7f435d453e90>
```

- Looks like our r2 score value is 0.65 which means our model is able to capture most of the data variance.
- The r2_score for the test set is 0.63 which means our linear model is performing well on the data.

Evaluation of Model



		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	0	Linear regression	4.444	34.810	5.900	0.774	0.77
	1	Lasso regression	7.242	91.458	9.563	0.406	0.39
	2	Ridge regression	4.444	34.810	5.900	0.774	0.77
	3	Elasticnet regression	5.765	57.415	7.577	0.627	0.62
	4	Decision tree regression	5.066	50.219	7.087	0.674	0.67
	5	Random forest regression	0.808	1.572	1.254	0.990	0.99
	6	XG Boost Regression	3.307	19.088	4.369	0.876	0.87
	7	XG boost regg GridserachCV	1.166	2.949	1.717	0.981	0.98
Test set	0	Linear regression	4.373	33.085	5.752	0.791	0.79
	1	Lasso regression	7.442	96.685	9.833	0.388	0.37
	2	Ridge regression	4.373	33.087	5.752	0.791	0.79
	3	Elasticnet regression	5.846	59.380	7.706	0.624	0.62
	4	Decision tree regression	5.427	59.482	7.712	0.623	0.62
	5	Random forest regression	2.239	13.047	3.612	0.917	0.92
	6	XG Boost Regression	3.498	21.624	4.650	0.863	0.86
	7	XG boost regg GridserachCV	2.174	10.971	3.312	0.931	0.93

- ➢ Out of all the above models "Random forest Regressor" gives the highest Adj.R2 score of 99%.
- ➢ For the Train Set and "XG Boost Grid search CV" gives the highest Adj.R2 score of 91% for the Test set.
- ➢ No overfitting is seen

Challenges

- A huge amount of data needed to be dealt with while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- Needs to plot a lot of Graphs to analyze.
- Feature engineering.
- Feature selection.
- Model Training and performance improvement.
- As the dataset was quite big enough it led to more computation time.

Conclusion

1. ‘Hour’ of the day holds the most important feature.
2. Bike rental count is mostly correlated with the time of the day as speak at 10 am morning and 8 pm evening.
3. We observed that the bike rental count is high during working days than on nonworking days.
4. We see that people generally prefer to bike at moderate to high temperatures, and when a little windy
5. It is observed that the highest number of bike rentals counts in the Autumn & Summer seasons & the lowest in the winter season. We observed the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day. We observed that with increasing humidity, the number of bike rental counts decreases.

thank
you