Capstone Project 2

Seoul Bike Sharing Demand Prediction

Kiran Ahire

Points for Discussion

- Business Objective
- Data Summary
- Feature Summary
- Data Preprocessing
- Exploratory Data Analysis
- Corelation between data
- Algorithms implementation
- Conclusion

Business Objective

- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. The client is Seoul Bike, which participates in bike share program in Seoul, South Korea. An accurate prediction of bike count is critical to the success of Seoul Bike share program. It is important to make rental bike available and accessible to the public at right time as it lessens the waiting time.
- Rented bikes are mostly used by the people having no personal vehicles or those people who wants to avoid congested public transport. Eventually providing a city with stable supply of motor bikes become a major concern.
- The goal of the project is to predict number of rental bikes required at each hour for stable supply of rental bikes.

Data Summary

- This dataset contains 8760 rows and 14 columns.
- Three categorical features “Seasons ”,”Holiday” and “Functioning Day”
- One datetime feature “Date”
- Following columns are of numerical data
    1. Temperature
    2. Humidity
    3. Wind speed
    4. Visibility
    5. Dew point Temperature
    6. Rainfall
    7. Solar radiation
    8. Snowfall

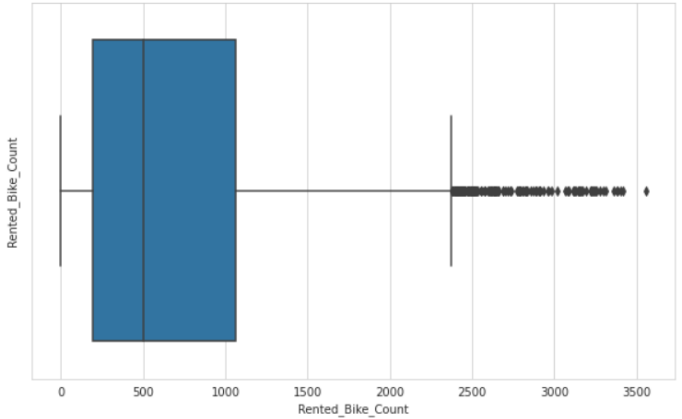
Feature Summary

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

Data Preprocessing

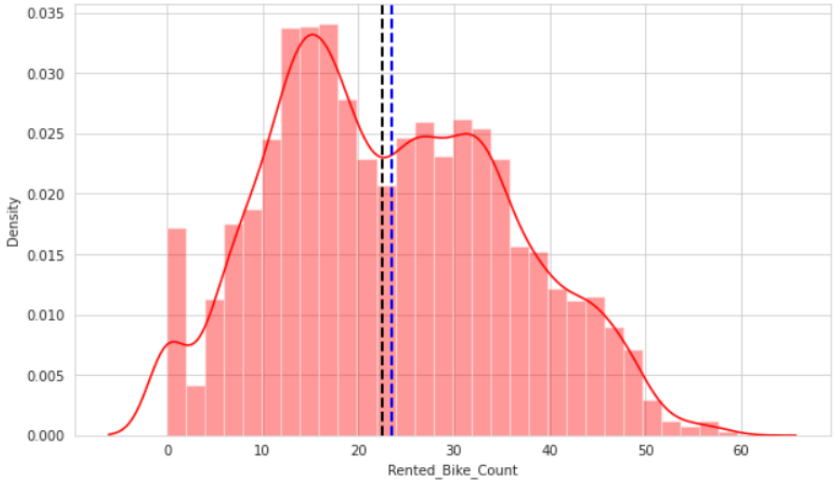
- There are no missing values present
- There are no duplicate values present
- There are no null values present
- ‘Rented bike count’ is the dependent variable which we need to predict for our new observations
- The dataset is hourly rental data for one year so that I converted ‘Date’ feature into new features namely ‘Day’ , ‘Month’ , ‘Year’
- In our dataset"Hour","month","weekdays_weekend" column are integer data type but it should be category data type. so I change their data type.
- I rename some of the features for convenience.

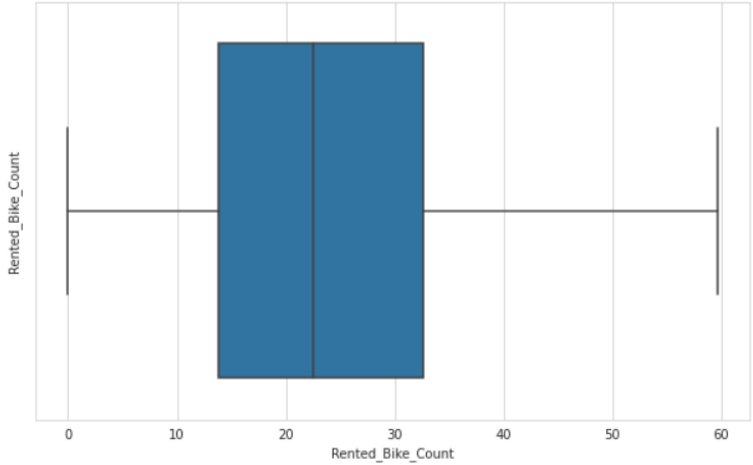
Analysis of Rented Bike count



- The above graph shows that Rented Bike Count has moderate right skewness.
- The above boxplot has detected outliers in Rented_Bike_Count
- Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform some operation to make it normal. To reduce the right skewness we will apply Square root method.

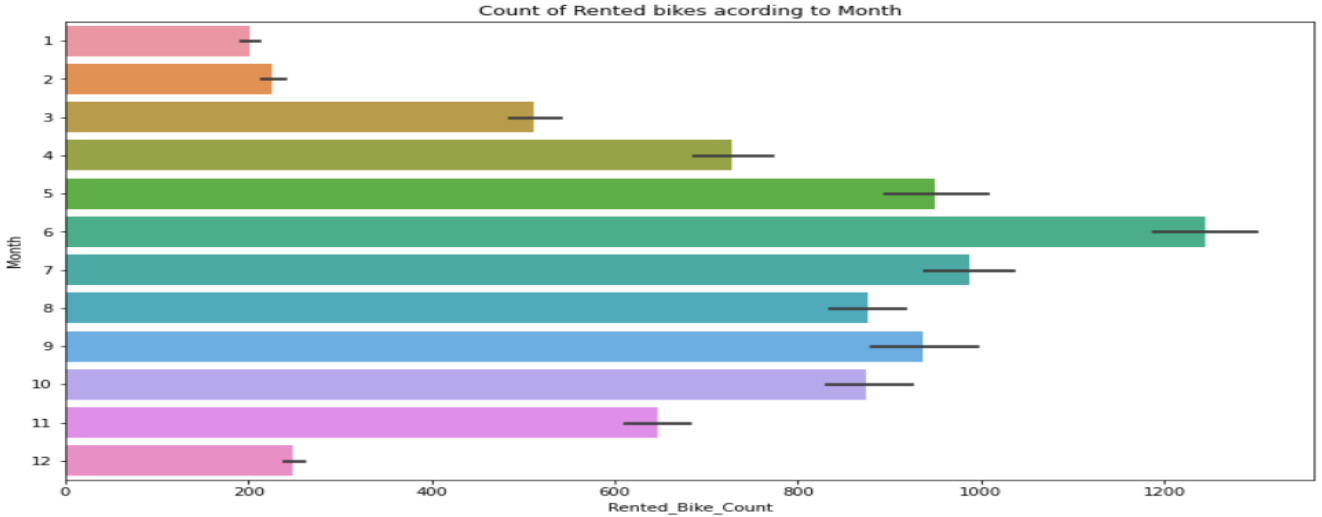
Analysis of Rented Bike count





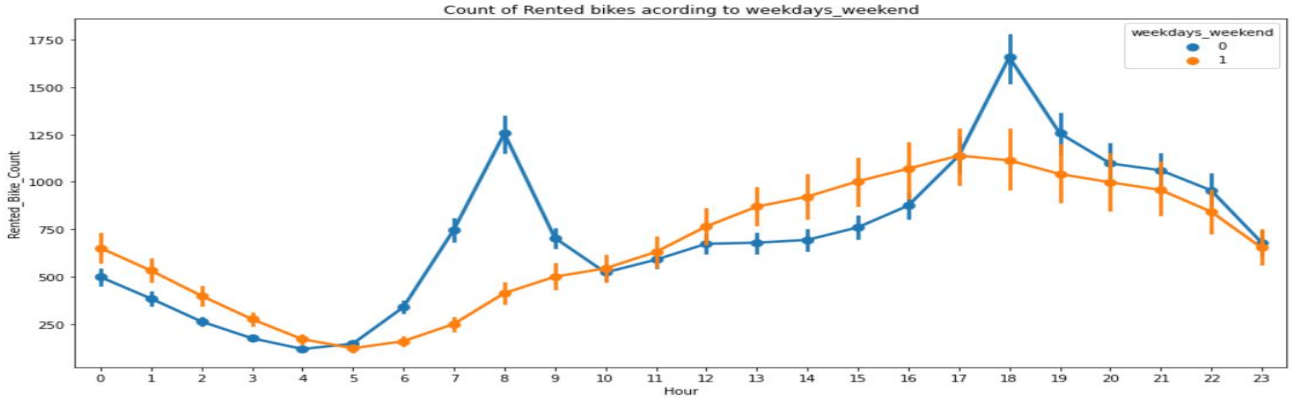
- From the above plot we can say that after applying Square Root method, Rented_Bike_Count column now is in Normal Distribution.
- After applying Square root to the Rented Bike Count column, we find that there is no outliers present.

Analysis of Month



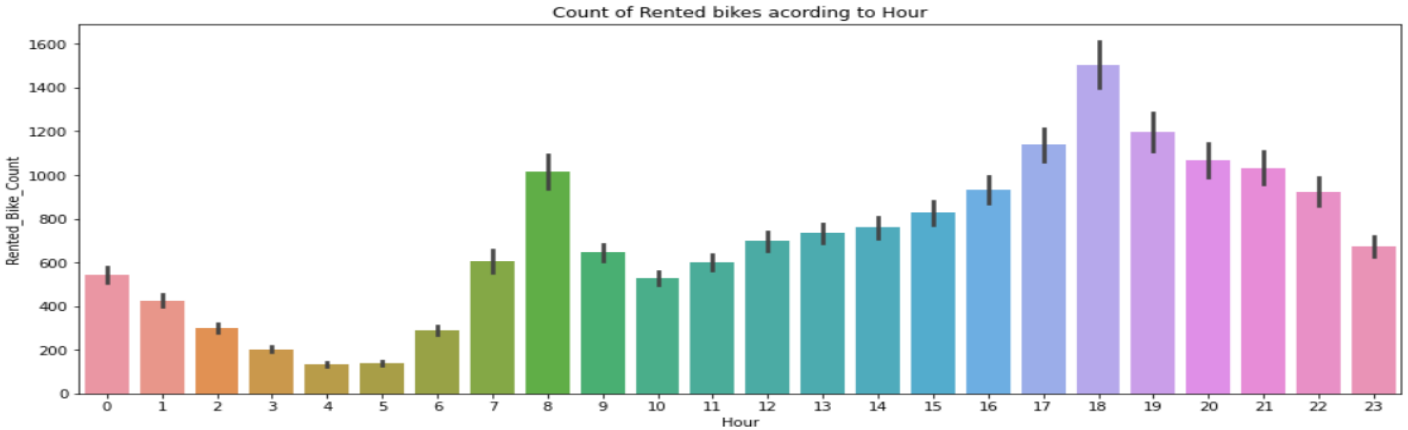
- From the above graph we can say that from the month 5 to 10 the demand of the rented bike is high as compare to other months. These months fall in the summer season.

Analysis of weekdays_weekend



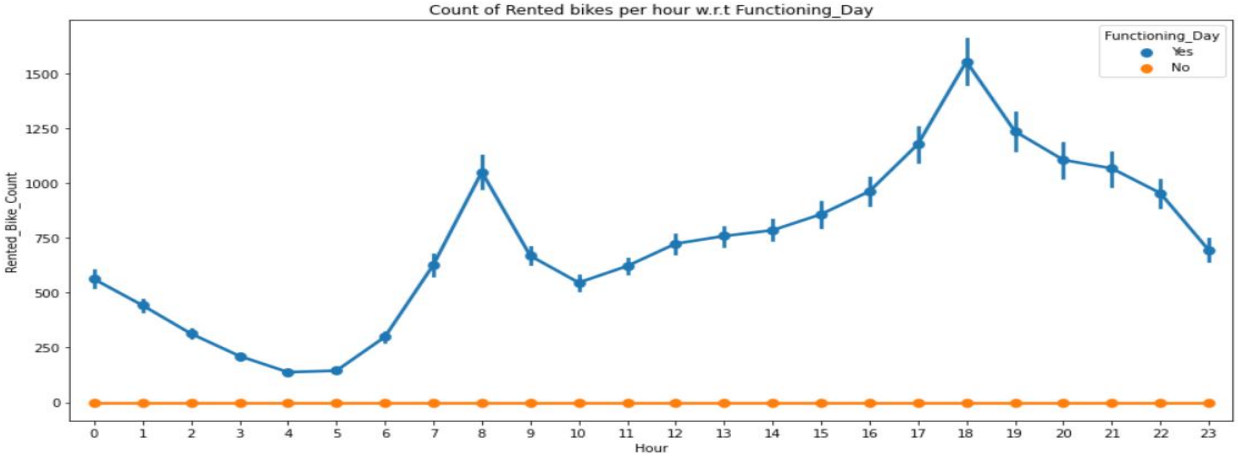
- From the above point plot and bar plot we conclude that in the week days which represent in blue color show that the demand of the bike higher during 7am to 9am and 5pm to 7pm.
- The orange color represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour.

Analysis of Hour



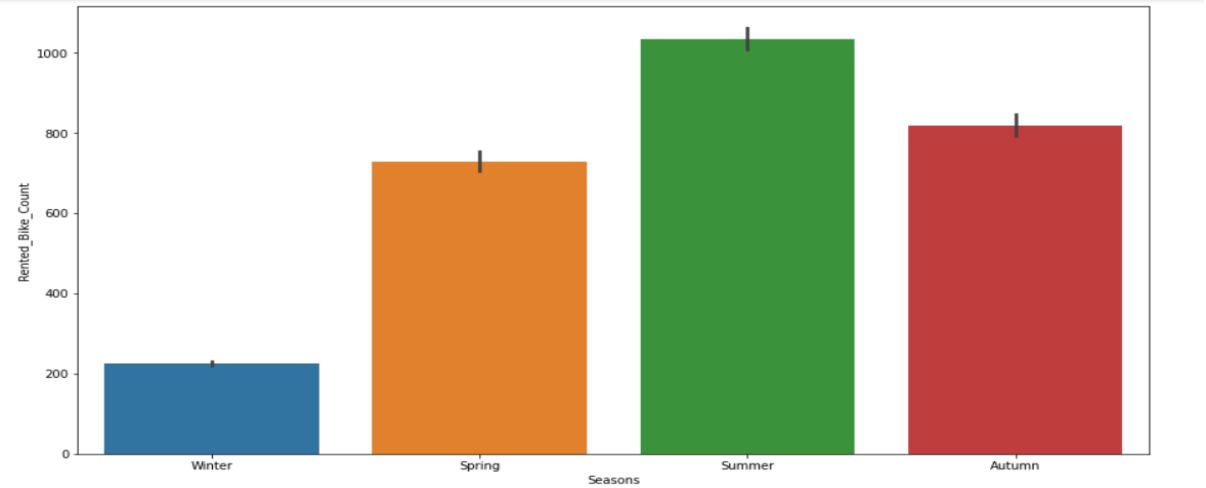
- The above plot shows the use of rented bike according the hours.
- From the above plot we conclude that people generally use rented bikes during working hours from 7am to 9am and 5pm to 7pm

Analysis of Functioning_Day



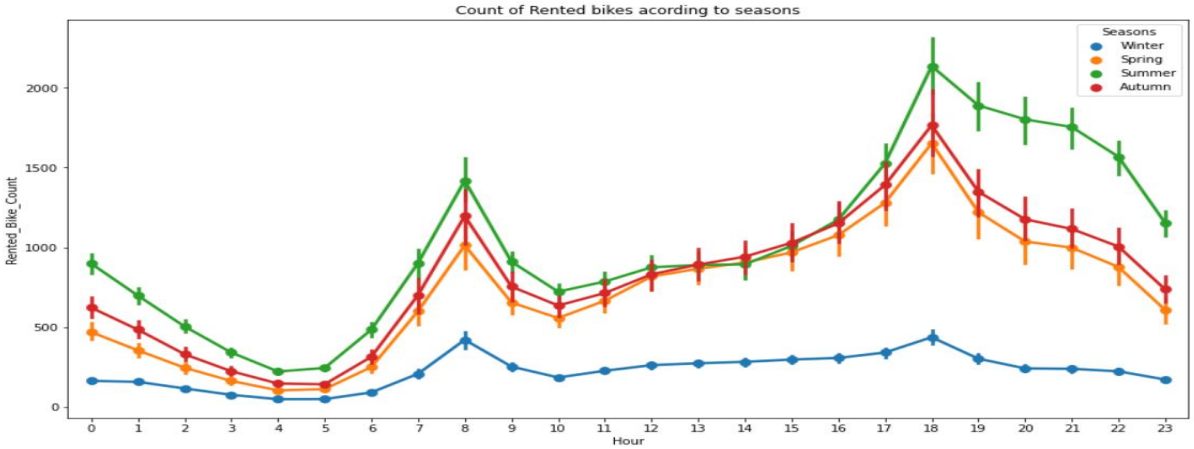
- From the above point plot we can conclude that people do not use rented bike on no functional day.

Analysis of Season



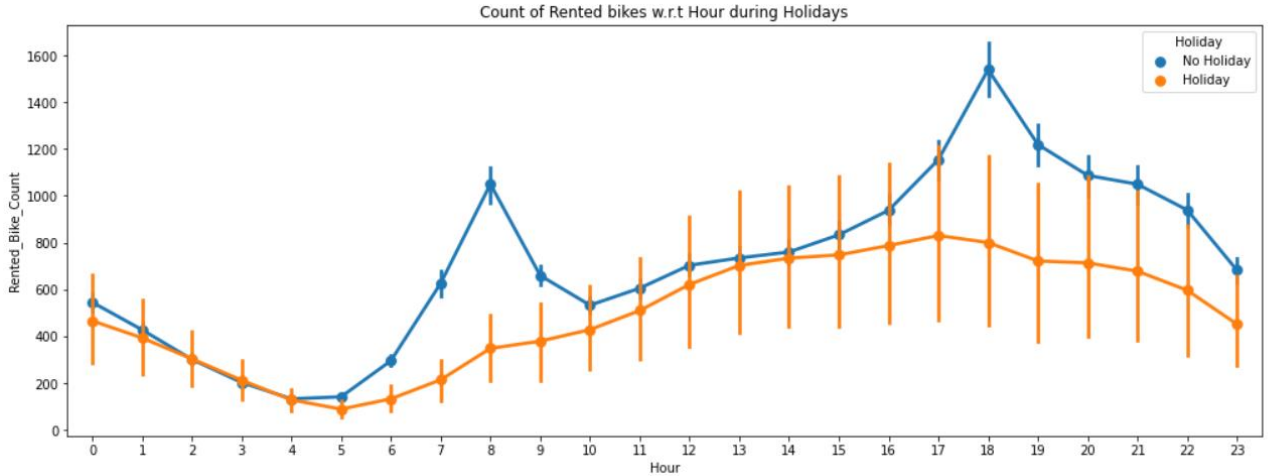
- The above bar plot shows the distribution of rented bike count season wise.
- From the above graph we can say that people love to ride bikes in Summer season followed by Autumn.
- People avoid riding bikes in winter season.

Analysis of Season



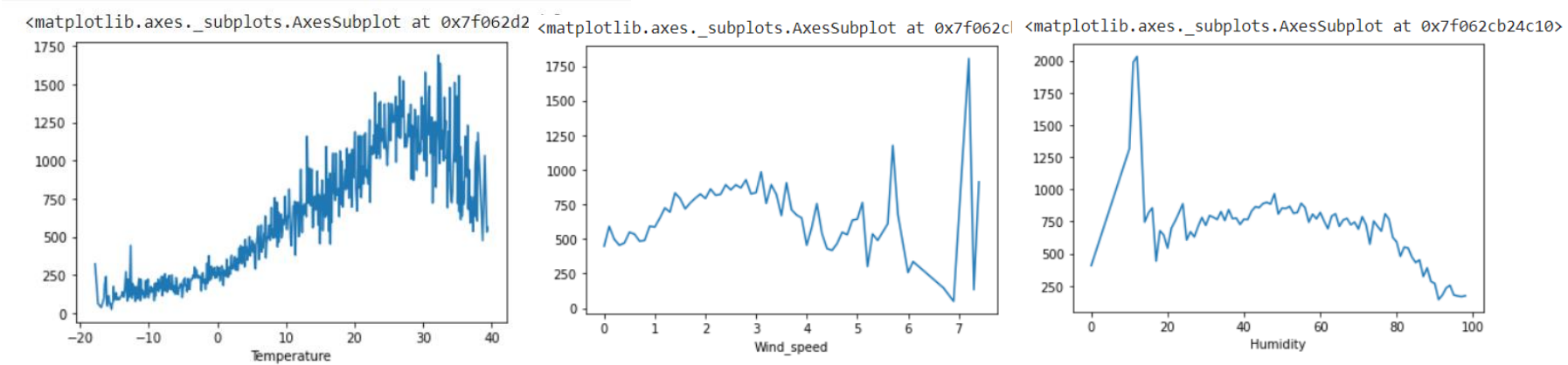
- The above graph shows the use of rented bike in different seasons, from the above graph we can say that in summer season the use of rented bike is high and peak time is 7am-9am and 7pm-5pm.
- In winter season the use of rented bike is very low because of snowfall.

Analysis of Holiday



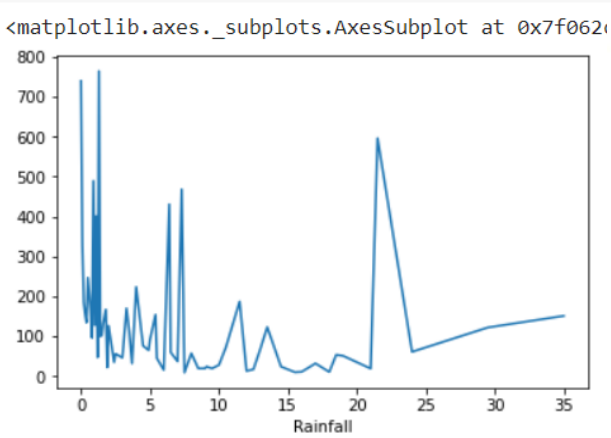
- The above plot shows the use of Rented bikes during holidays and from above plots we can say that people generally use rented bikes between 2pm to 8pm on holidays.

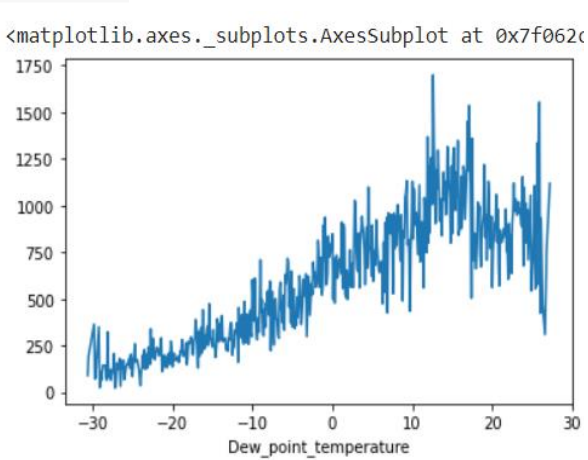
Numerical vs Rented_Bike_Count

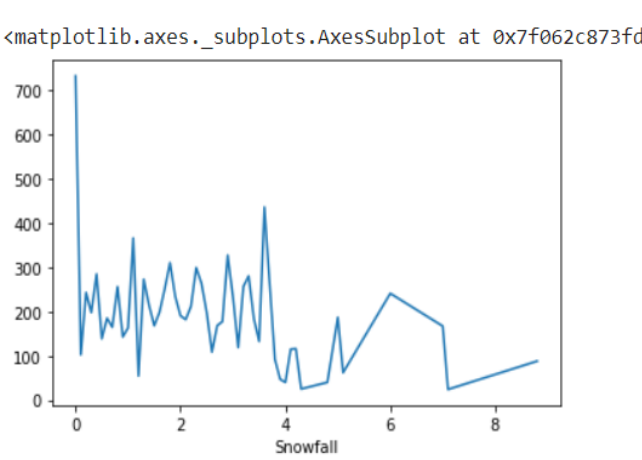


- From above graph we conclude that people like to ride bikes when the temperature is around 25°C
- From above plot we can say that Wind_speed is uniformely distributed but when Wind_speed is around 7 m/s demand for bikes is higher so we can say that people likes to ride bikes when its little windy.
- From the above graph we can say that demand for rented bikes is higher when humidity is around 10.

Numerical vs Rented_Bike_Count

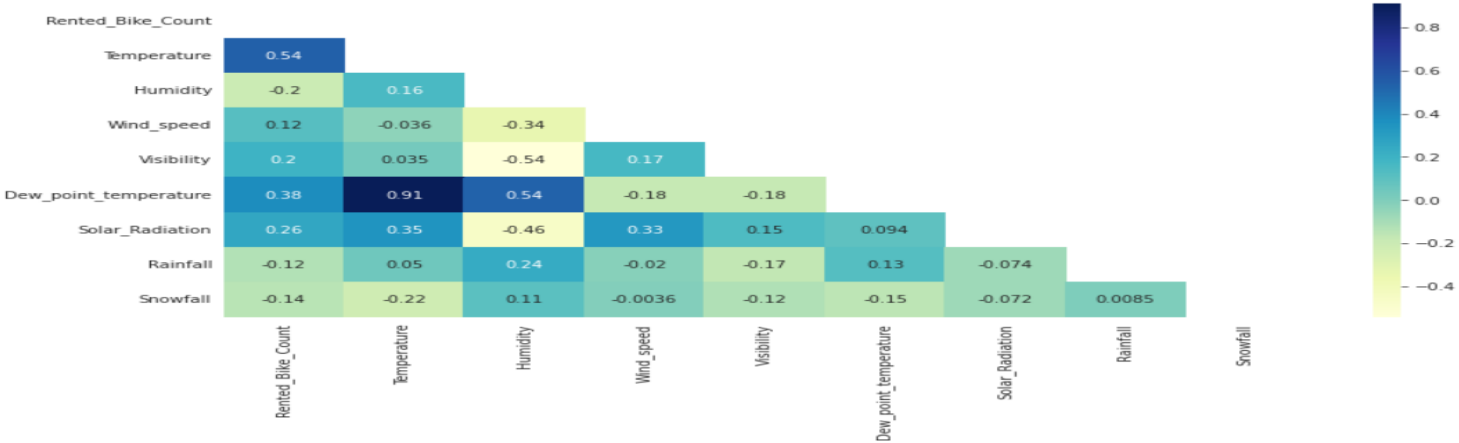






- We can see from the above plot that even if it rains a lot the demand of of rent bikes is not decreasing, here for example even if we have 20 mm of rain there is a big peak of rented bikes
- From the above plot we can say that demand for rented bikes is high when Dew point temperature is between 10 to 20.
- From this we can see when there was no snow then the demand is highest whereas when the snow is more than 4 cm then demand is decreasing.

Corelation between variables



- From the above Heatmap we can say that target variable Rented_Bike_Count is positively corelated with Tempreture , Dew_point_tempreture , Solar_Radiation and negatively corelated with Rainfall , Snowfall.
- From above graph we can say that columns 'Temperature' and 'Dew point temperature are highly corelated.

One Hot Encoding

- Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. This is required for both input and output variables that are categorical.
- One hot encoding is **one method of converting data to prepare it for an algorithm and get a better prediction**.
- With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.
- We have Categorical variables such as ‘Season’, ’Hour’, ’Month’ , ‘Holiday’
- So we applied One Hot Encoding on them.

Models Implemented

- **Linear Regression**

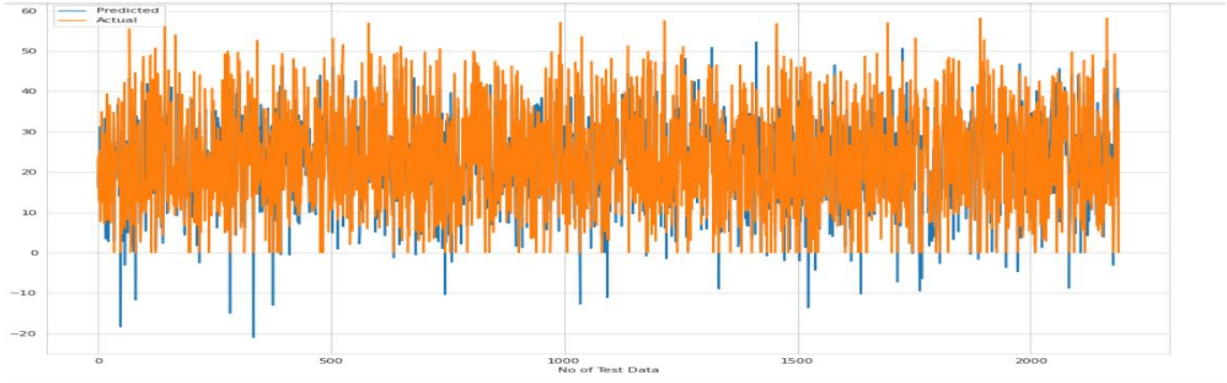
1. Linear Regression is a machine learning algorithm based on supervised learning
2. It performs a **regression task**
3. Regression models a target prediction value based on independent variables.

Train set results

```
MSE : 35.07751288189292
RMSE : 5.9226271942350825
MAE : 4.410178475318181
R2 : 0.7722101548255267
```

Test set results

```
MSE : 33.27533089591926
RMSE : 5.76847734639907
MAE : 4.410178475318181
R2 : 0.7893518482962683
```



Models Implemented

- **Lasso Regression**

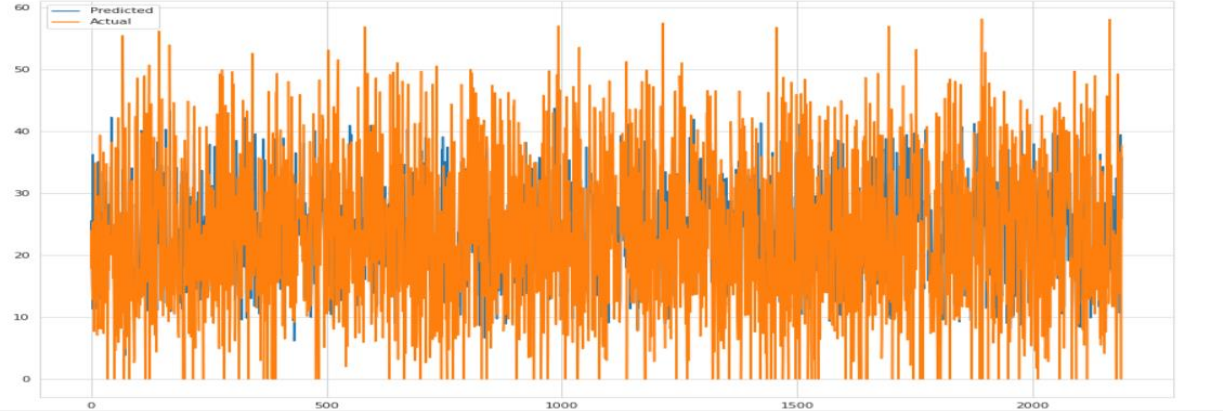
  Lasso Regression is an extension of linear regression that adds a regularization penalty to the loss function during training.

Train set results

```
MSE : 91.59423336097032
RMSE : 9.570487623991283
MAE : 7.255041571454952
R2 : 0.40519624904934015
```

Test set results

```
MSE : 96.7750714044618
RMSE : 9.837432155011886
MAE : 7.455895061963607
R2 : 0.3873692800799008
```



Conclusion

- Rented_bike _count which is our dependent variable is mostly corelated with time of the day.
- ‘Hour’ is important feature in the dataset.
- We observed that bike rental count is high during working days than non-working days.
- We observed that people generally preferred bikes when it is moderate to high temperature and when its little windy too.
- We can say that Rental bike count is high in Autumn and Summer season and low in Winter.
- In Linear Regression R2 score value is 0.77 that means our model is able to capture most of the data variance.
- In Lasso Regression R2 score value is 0.38 which is very poor as compared to Linear Regression so we can say LR is better model than Lasso.