CAPSTONE PROJECT 2

Bike Sharing Demand Prediction

By - Wasim Ahmad

Let’s Ride the Bike

- Introduction
- Business Context
- Defining Problem Statement
- Data Summary
- Exploratory Data Analysis
- Data Preprocessing
- Data Visualisation
- Models and Evaluation
- Conclusion and Inference

A person checks a bike from Seoul City's bike-sharing service, Ttareungyi, at a station near City Hall in central Seoul, June 22.
Newsis

Introduction

- Ddareungi is Seoul's bike sharing system, which was set up in 2015. It is also named Seoul Bike in English.
- Ddareungi was first introduced in Seoul in October 2015 in select areas of the right bank of the Han river. After a few months, the number of stations reached 150 and 1500 bikes were made available.
- Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system. Rental Bike Sharing is the process by which bicycles are procured on several basis - hourly, weekly, membership-wise, etc.

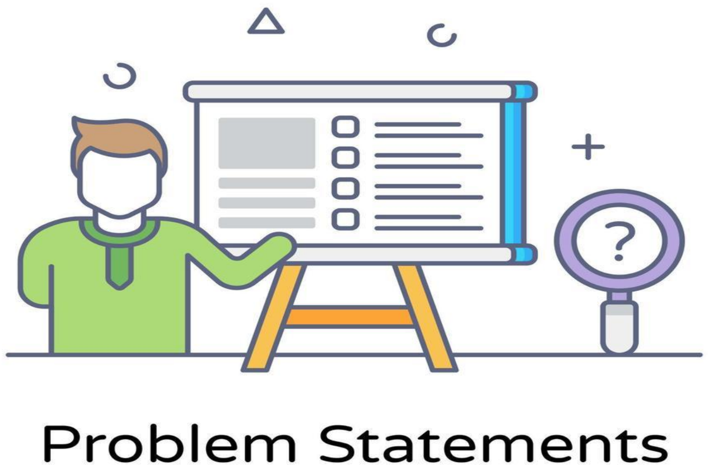
Business Context

- Right now Rental bikes are presented in numerous metropolitan urban communities for the improvement of mobility comfort.
- It is important to make the rental bikes accessible and open to people in general brilliantly as it decreases the holding up time. At last, furnishing the city with a steady availability of rental bike turns into a central issue.

Defining Problem Statement

Build a model that predict the number of bikes required at each hour for the stable supply of rental bikes.

Predict the factors affecting the demand for rental bikes with the help of data provided



Data Summary

Seoul bike data has 8760 rows and 14 columns. The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), Total hours bikes rented for, holiday, Functional day and date information.

The detailed description are as follows.

Date: This column contains the date of the day given from 01/12/2017 to 30/11/2018 its data type is object.

Rented Bike Count: This column contains Number of rented bikes per hour which is our dependent variable and we will predict it.

Data Summary Continued….

Hour: The hour of the day, starting from 0-23

Temperature(°C): Temperature of weather in Celsius

Humidity(%): This column has Humidity in the air in %

Wind speed (m/s): Speed of the wind given in this column in m/s.

Visibility (10m): It contains Visibility in m.

Dew point temperature(°C): Temperature at the beginning of the day its data type is Float

Solar Radiation (MJ/m2): Solar radiation outside

Data Summary Continued….

Rainfall(mm): Rainfall in mm

Snowfall (cm): Amount of snowfall in cm

Seasons: This column has Season of the year (ie. summer, winter, autumn, rain)

Holiday: It consist the two category of data that is holiday and no holiday showing weather the day is holiday or not.

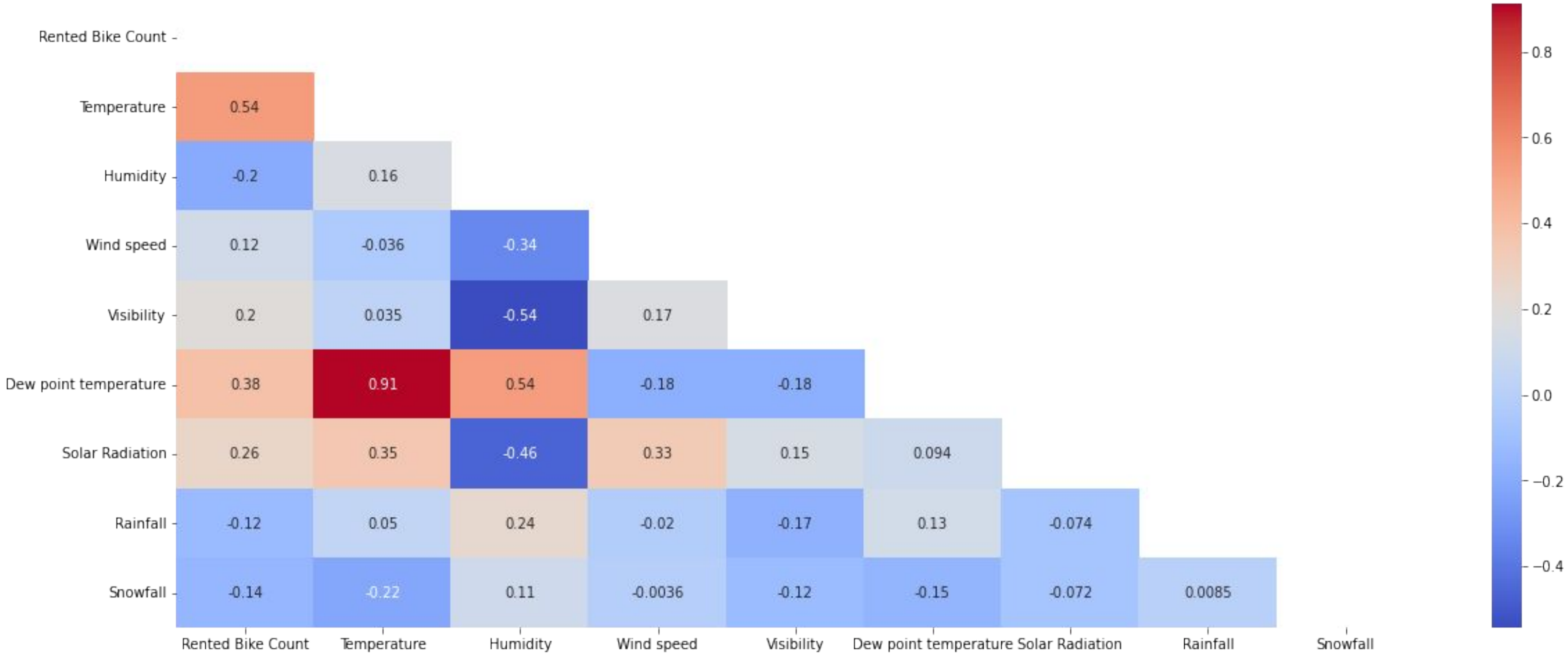
Functioning Day: It also consist the two category of data that If the day is a Functioning Day or not

EDA and Data Preprocessing

These are **key aspect of data analysis**. Without spending significant time on understanding the data and its patterns one cannot expect to see useful insight build efficient predictive model

- We have **Rented Bike Count** which is our dependent variable and we need predict it for unseen data.
- There is no null and duplicate values present.
- Convert Date column into datetime format then we split it into three column i.e 'year', 'month', 'day' as a category data type as we need to analyze on the basis of day, month etc
- Changed data type of some column and dropped some as required

EDA and Data Preprocessing continued…



'Temperature' and 'Dew point temperature' is highly correlated i.e. 0.91 so dropped 'Dew point temperature'

EDA and Data Preprocessing continued…

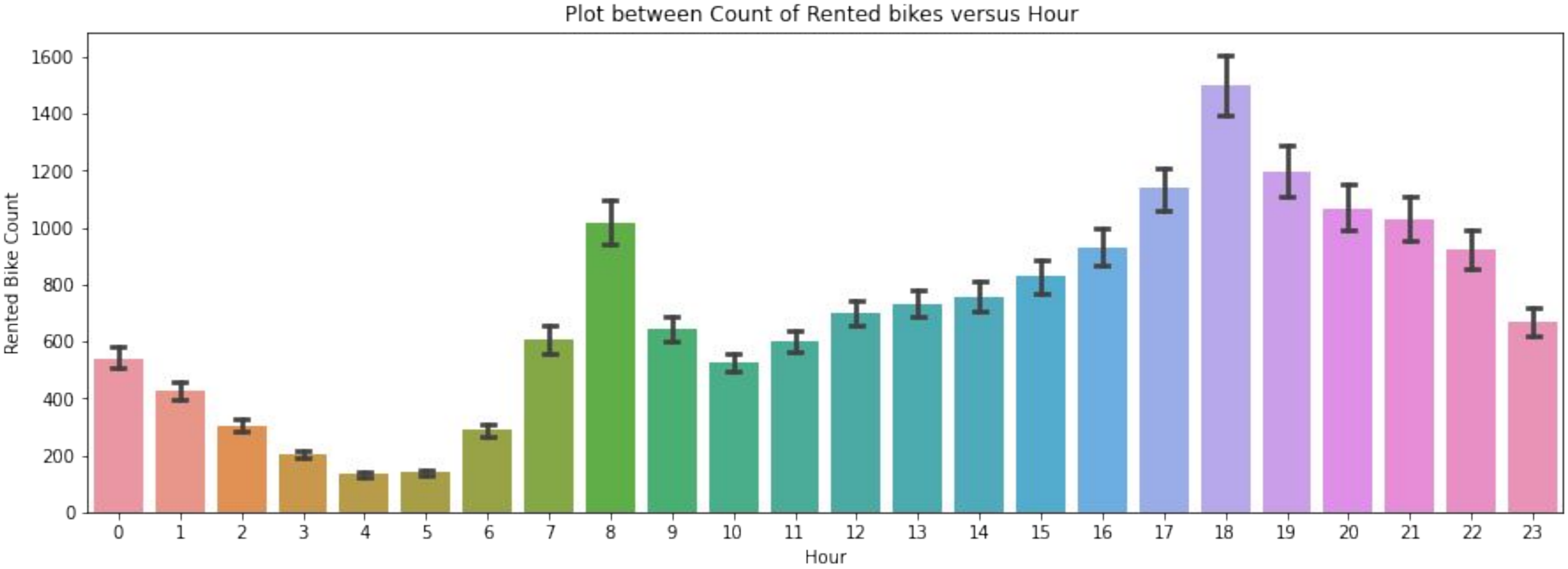
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Date                       8760 non-null   object 
 1   Rented Bike Count          8760 non-null   int64  
 2   Hour                       8760 non-null   int64  
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64  
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64  
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object 
 12  Holiday                    8760 non-null   object 
 13  Functioning Day            8760 non-null   object 
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype   
---  ------             --------------  -----   
 0   Rented Bike Count  8760 non-null   int64   
 1   Hour               8760 non-null   category
 2   Temperature        8760 non-null   float64 
 3   Humidity           8760 non-null   int64   
 4   Wind speed         8760 non-null   float64 
 5   Visibility         8760 non-null   int64   
 6   Solar Radiation    8760 non-null   float64 
 7   Rainfall           8760 non-null   float64 
 8   Snowfall           8760 non-null   float64 
 9   Seasons            8760 non-null   object  
 10  Holiday            8760 non-null   object  
 11  Functioning Day    8760 non-null   object  
 12  month              8760 non-null   category
 13  Weekend            8760 non-null   category
dtypes: category(3), float64(5), int64(3), object(3)
memory usage: 779.8+ KB
```

Data Visualisation

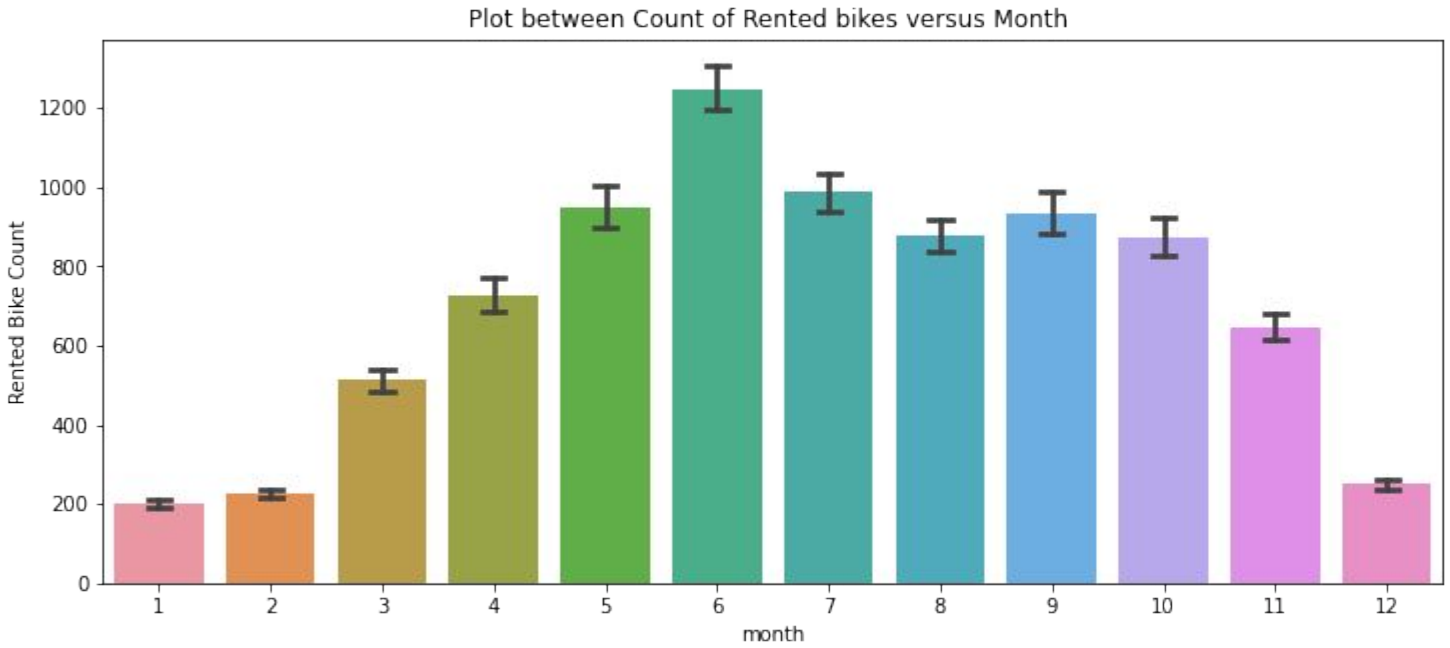
Data visualization is **the graphical representation of information and data**. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Analysis of Demand for Rented Bike versus per hour



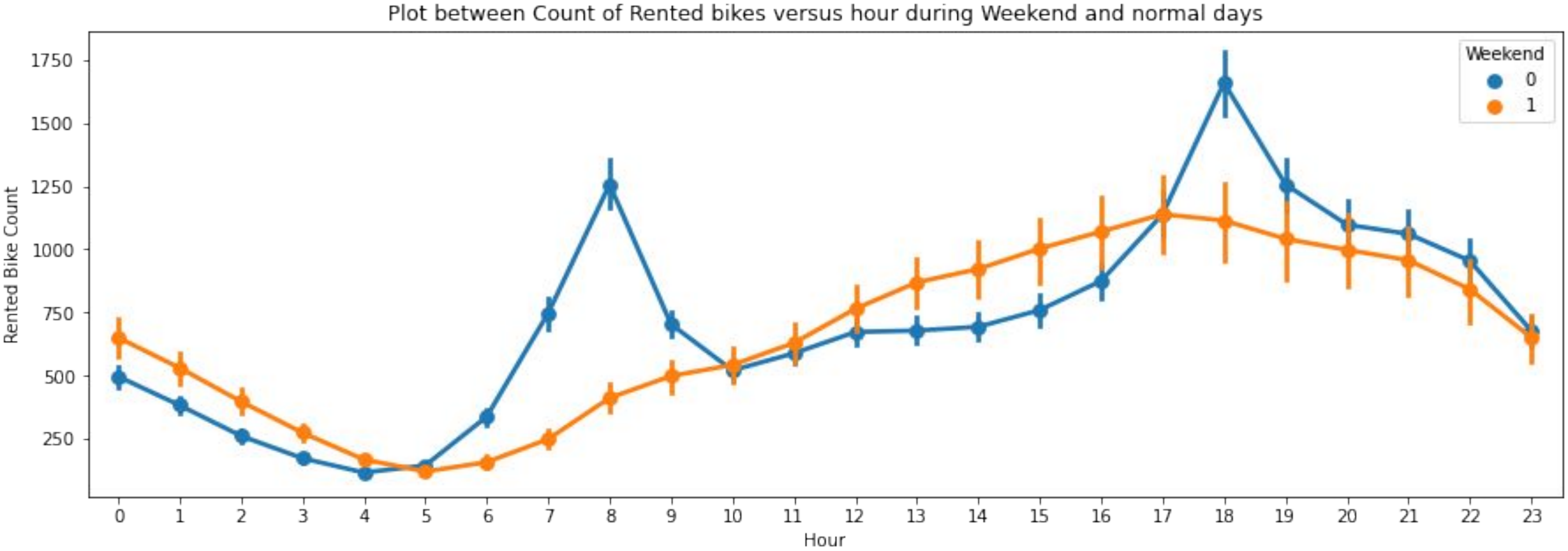
Above bar plot shows that the demand of rented bikes are high during the working hours from 7am to 9am in the morning and 5pm to 7pm in the evening throughout the year.

Analysis of Demand for Rented Bike versus Month



In the month 5 (May) to 10 (October), during summer season the demand of the rented bike is high as compare to other months and in June it is highest

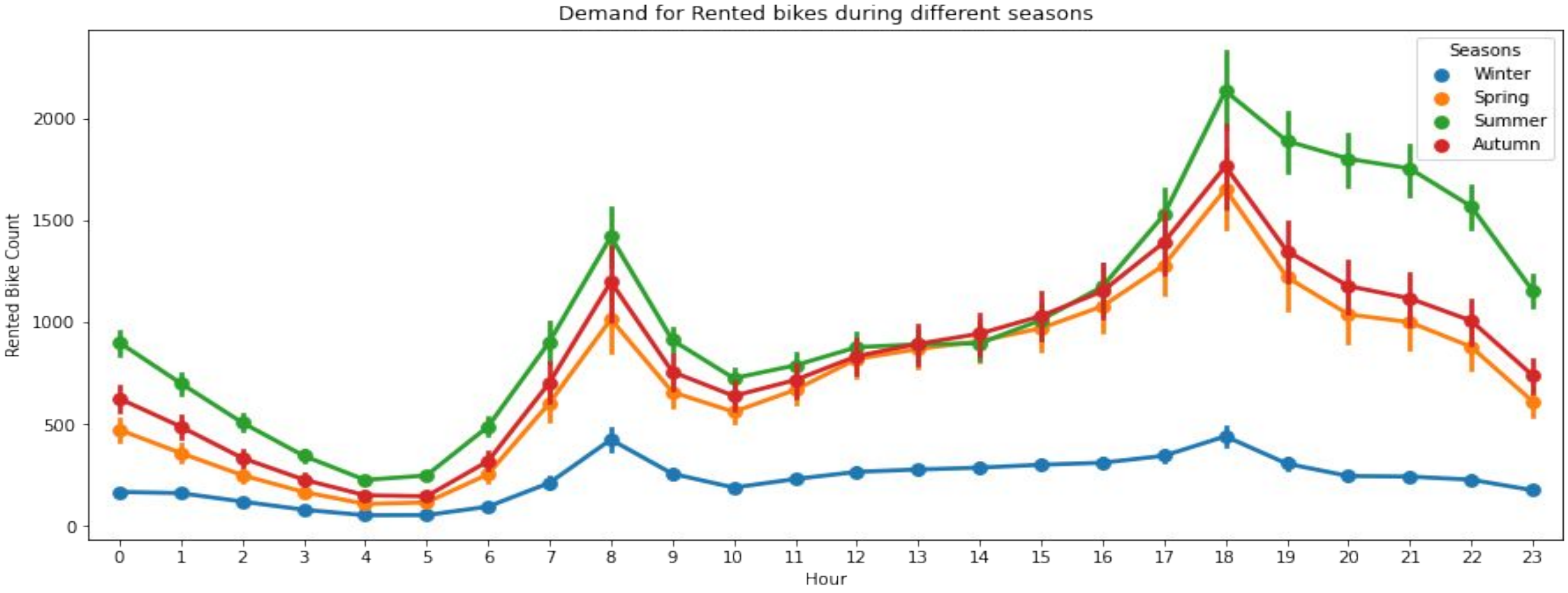
Analysis of Demand for Rented Bike versus weekend



Normal working days the demand of rented bikes are high between morning (7 am to 9 am) and evening (5 pm to 7 pm) . It shows the office opening and colsing time the demand is higher.

On **weekend days** the demand of rented bike is very low in the morning hour but it increases gradually and after 5 pm decreases as well.

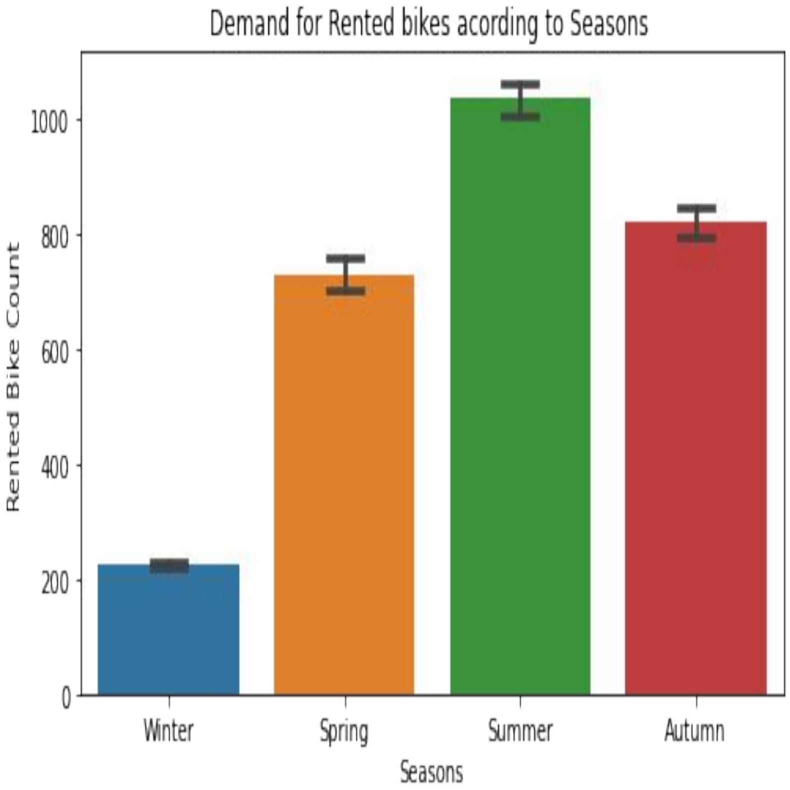
Analysis of Demand for Rented Bike versus Season



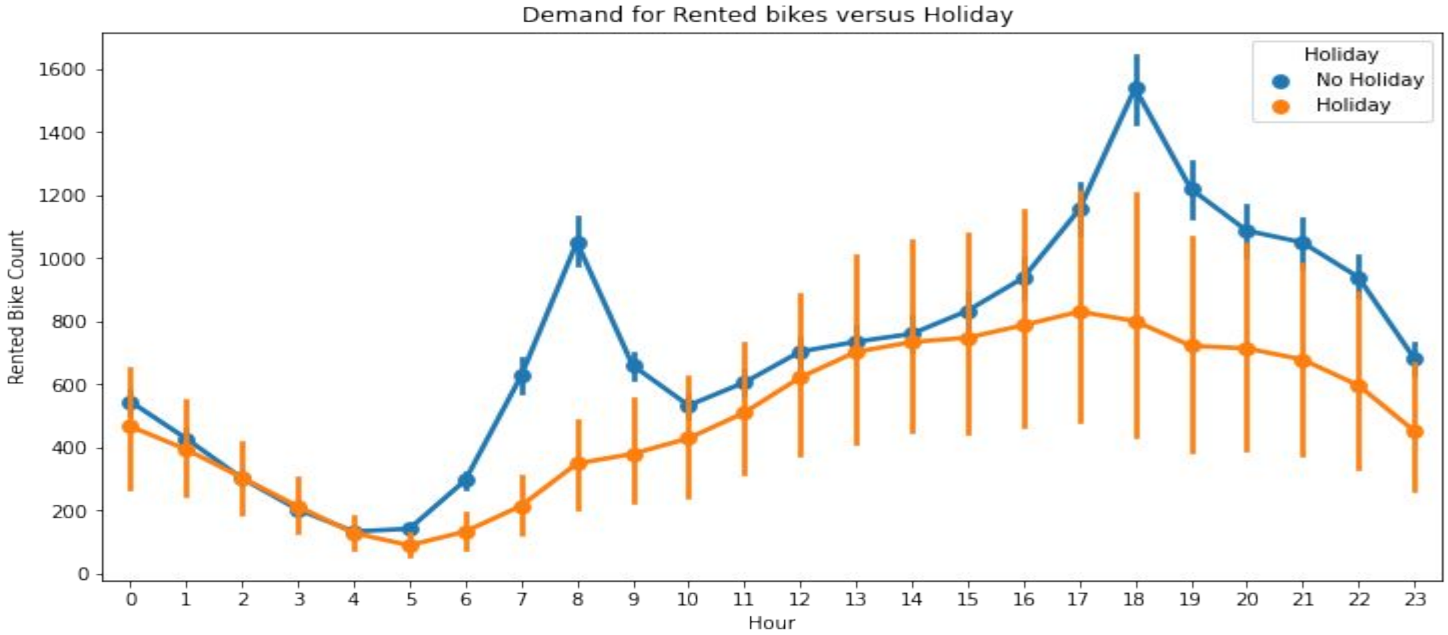
In every season the demand for rented bike has a peak at 7am-9am and 5pm-7pm.

Analysis of Demand for Rented Bike versus different Seasons

- The demand of rented bikes is comparatively very low in **winter season**, we may say so due to snowfall.
- In the **spring season** the demand for rented bikes is comparatively higher than winter season
- In the **autumn season** the use of rented bikes is higher than the spring seasons
- In the **summer season** the use of rented bikes is highest among all seasons.

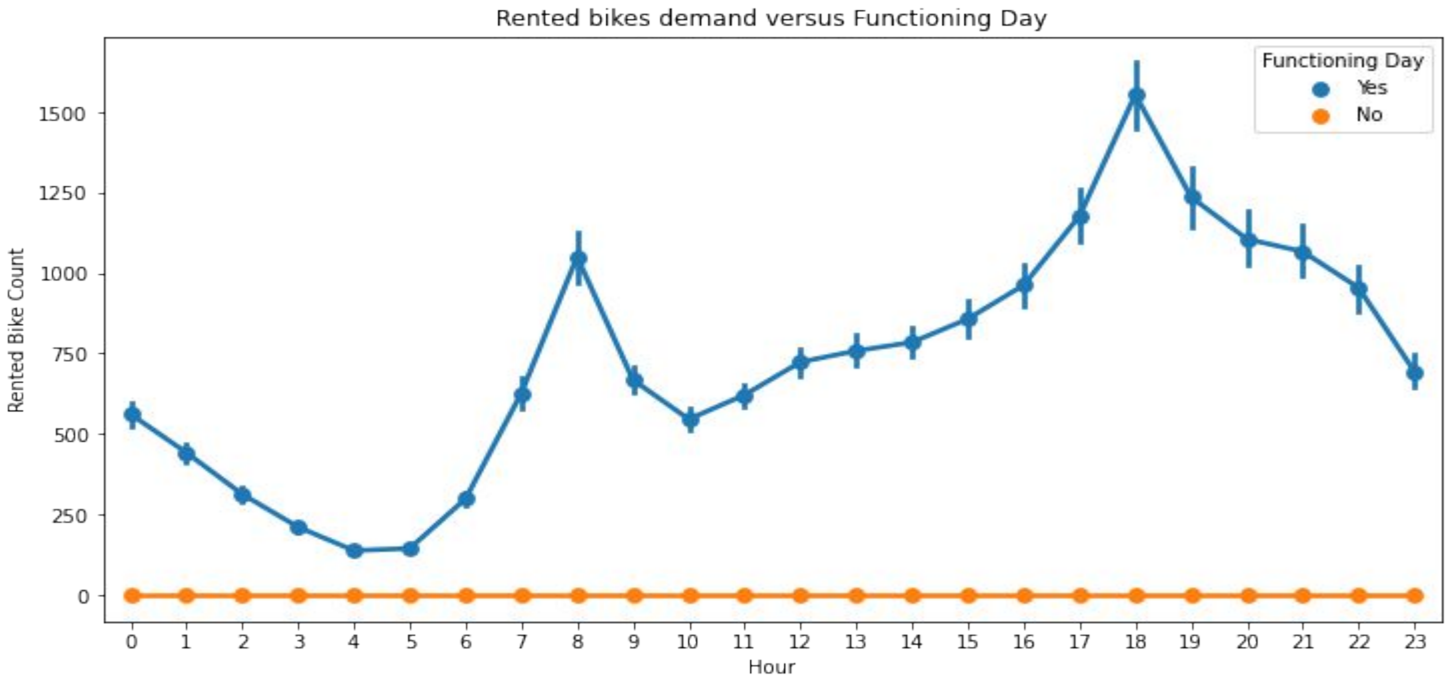


Analysis of Demand for Rented Bike versus Holiday



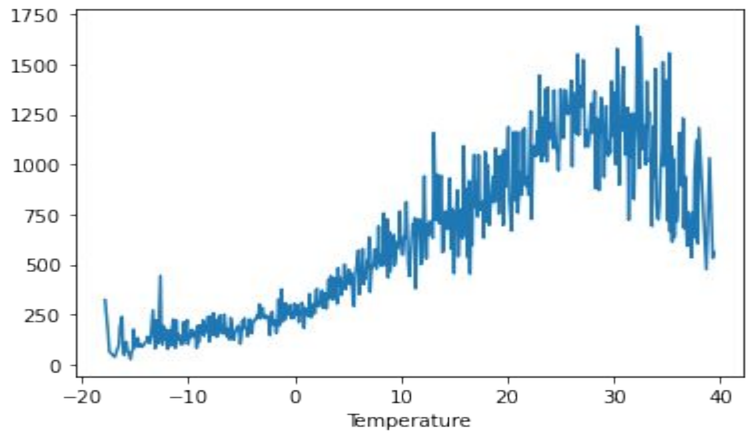
The demand for rented bikes starts from 5am and increases gradually till 5pm then again decreases gradually till 5am during holidays. Most people use rented bikes in the evening on holidays. When there is no holiday the demand of rented bikes are comparatively higher and has a peak at 7am-9am and 5pm-7pm

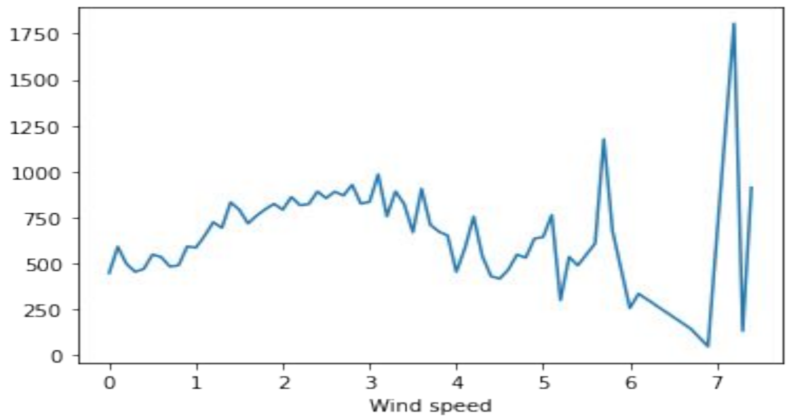
Analysis of Demand for Rented Bike versus Functioning Day

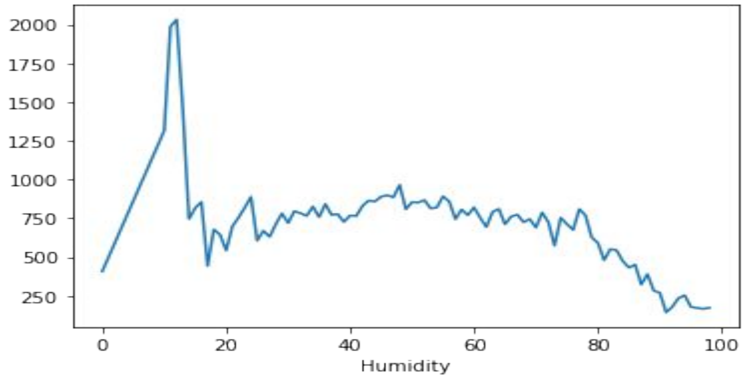


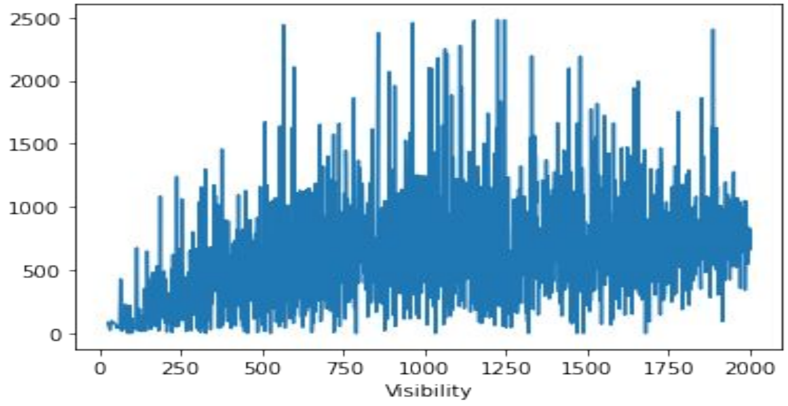
Above point plot clearly shows that the demand for rented bikes is only on functioning days Peoples don't use rented bikes on non functioning days.

Analysis of Demand for Rented Bike versus numerical features

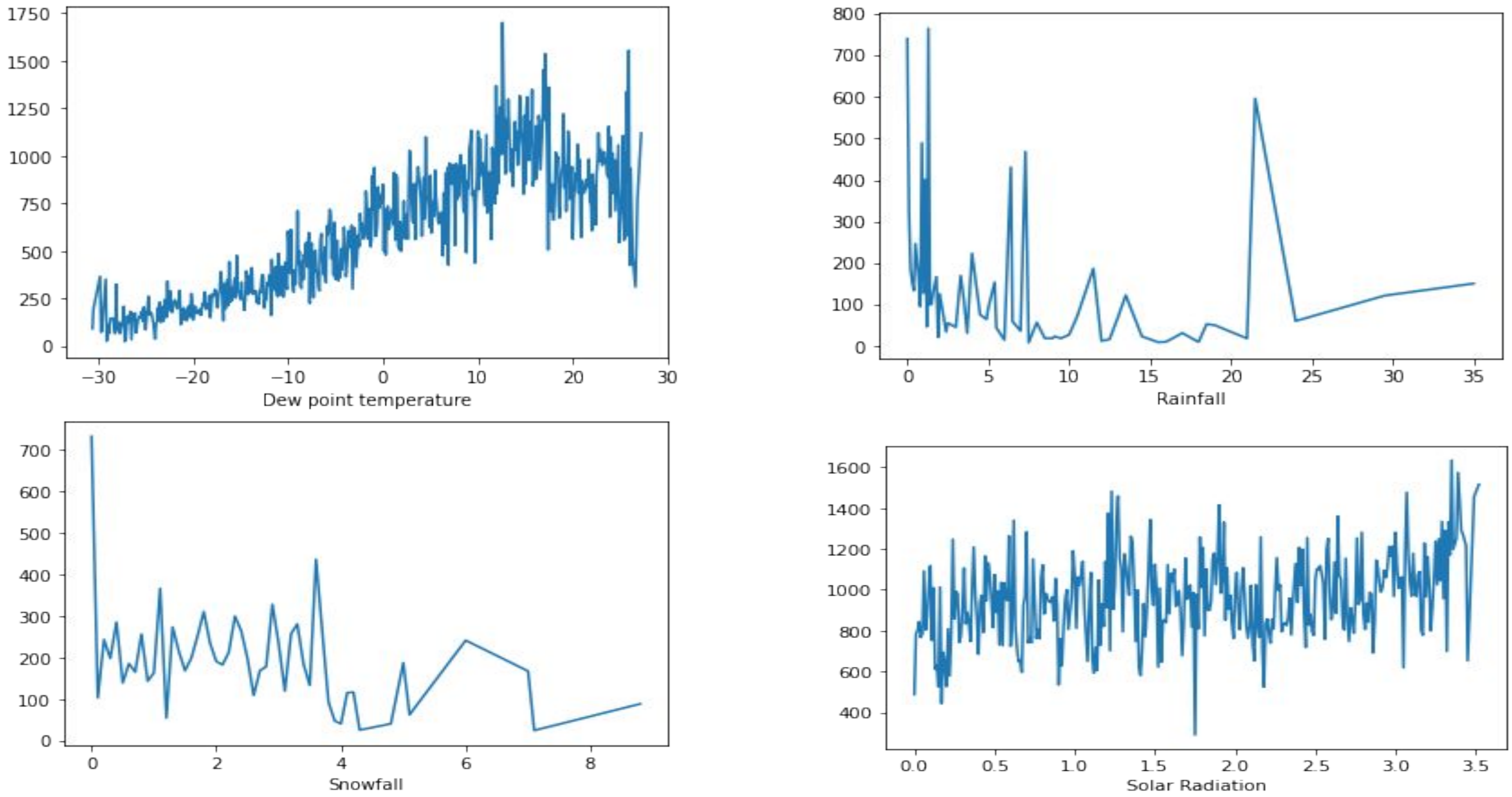








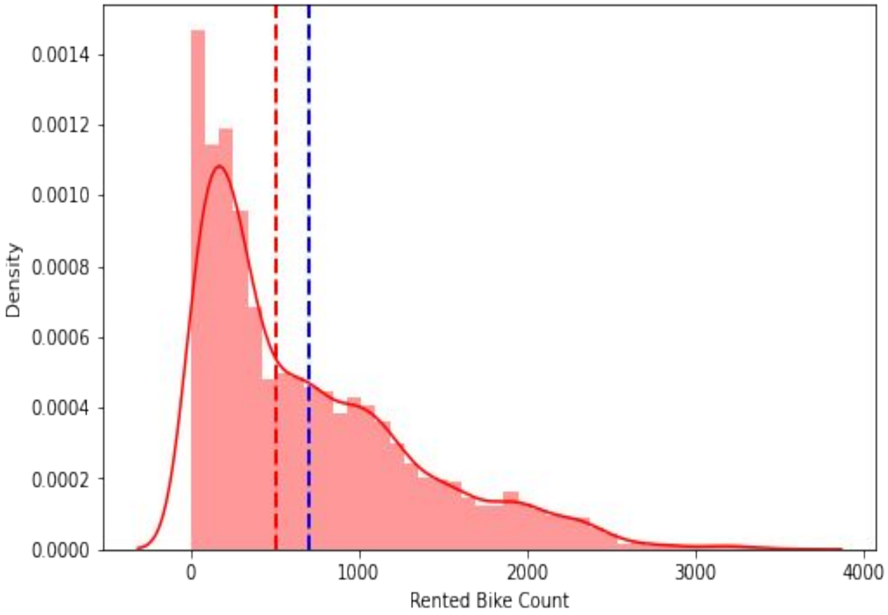
Analysis of Demand for Rented Bike versus numerical features

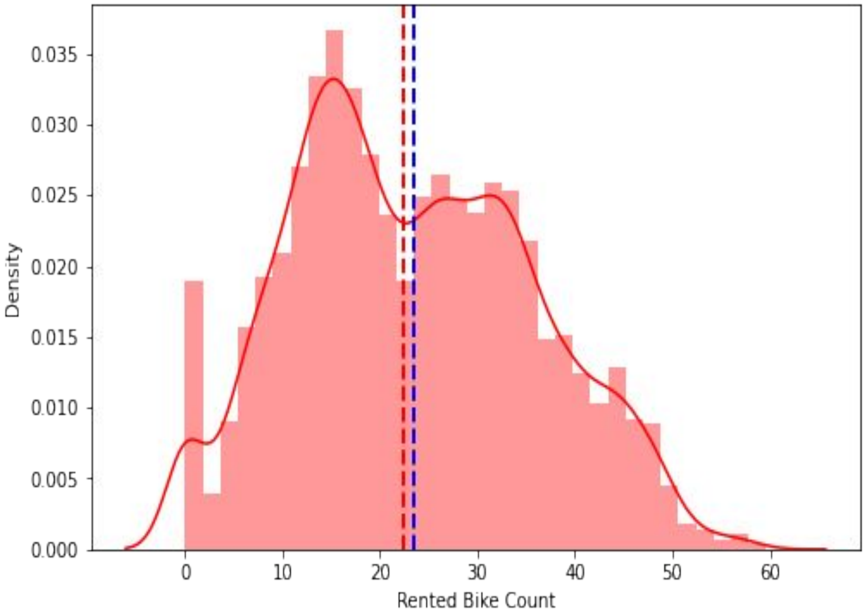


Analysis of Dependent Variable

In the first graph we can see that the Rented Bike Count has a right skewed.

In the second graph after applying Square root it became almost normal distribution.





Regression Model and Evaluation Metrics used

Linear Regression

Lasso Regression

Ridge Regression

Elastic Net Regression

Decision Trees

Random Forest

Gradient Boosting

MSE or Mean Squared Error

Root Mean Squared Error (RMSE)

MAE (Mean Absolute Error)

MAPE (Mean Absolute Percentage Error)

R2 (R – Squared)

Adjusted R2

Evaluation Metrics result of different models

Linear Regression

Train

```
MSE : 35.07751288189292
RMSE : 5.922627194235082 5
MAE : 4.474024092996788
R2 : 0.772210154825526 7
Adjusted R2 : 0.767211964945414 5
```

Test

```
MSE : 33.27533089591926
RMSE : 5.76847734639907
MAE : 4.410178475318181
R2 : 0.7893518482962683
Adjusted R2 : 0.7847297833429184
```

Lasso Regression

Train

```
MSE : 41.48012492751929
RMSE : 6.440506573827815
MAE : 4.960430531038622
R2 : 0.7306322353334551
Adjusted R2 : 0.7247217381629
```

Test

```
MSE : 39.91752283290745
RMSE : 6.318031563145871
MAE : 4.91263385826569
R2 : 0.7473037178309577
Adjusted R2 : 0.7417590281661
```

Ridge Regression

Train

```
MSE : 35.07752456136463
RMSE : 5.922628180239296
MAE : 4.474125776125378
R2 : 0.7722100789802107
Adjusted R2 : 0.767211887435892
```

Test

```
MSE : 33.27678426818438
RMSE : 5.768603320404722
MAE : 4.410414932539515
R2 : 0.7893426477812578
Adjusted R2 : 0.78472038094919
```

Elastic Net Regression

Train

```
MSE : 36.18743775720237 5
RMSE : 6.015599534310971
MAE : 4.571576500667136
R2 : 0.765002414175460 7
Adjusted R2 : 0.759846071255
```

Test

```
MSE : 34.8954576068473 4
RMSE : 5.907237730686597
MAE : 4.550377877058056
R2 : 0.779095700934419
Adjusted R2 : 0.774248594465659
```

Decision Tree

Train

MSE : 51.972784860024426
RMSE : 7.209215273524881
MAE : 5.238193841710909
R2 : 0.6624939557028167
Adjusted R2 : 0.655088360893308

Test

MSE : 58.32815297715247
RMSE : 7.637287016811171
MAE : 5.525496840397561
R2 : 0.6307559598622512
Adjusted R2 : 0.622653966451198

Random Forest

Train

MSE : 1.6116660853005755
RMSE : 1.2695141138642672
MAE : 0.7995415722249986
R2 : 0.9895340023313604
Adjusted R2 : 0.9893043562573987

Test

MSE : 12.626485560041926
RMSE : 3.5533766420183954
MAE : 2.2086036076326776
R2 : 0.9200685380393056
Adjusted R2 : 0.918314673094323

Gradient Boosting

Train

MSE : 18.64801713184794
RMSE : 4.3183349953249275
MAE : 3.2690035692731247
R2 : 0.8789016499095264
Adjusted R2 : 0.8762444965695393

Test

MSE : 21.28944184250869
RMSE : 4.6140483138463875
MAE : 3.4928587865599914
R2 : 0.8652280396863458
Adjusted R2 : 0.8622708584843188

While comparing evaluation metrics of all the models on training and test data, Random forest Regression gives the highest accuracy of 99% and 92% on train and test

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Evaluation Metrices of Train Data	**0**	Linear Regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso Regression	4.960	41.480	6.441	0.731	0.72
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	4.572	36.187	6.016	0.765	0.76
	4	Dicision tree regression	5.238	51.973	7.209	0.662	0.66
	5	Random forest regression	0.800	1.612	1.270	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
Evaluation Metrices of Test Data	**0**	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	4.913	39.918	6.318	0.747	0.74
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	4.550	34.895	5.907	0.779	0.77
	4	Dicision tree regression	5.525	58.328	7.637	0.631	0.62
	5	Random forest regression	2.209	12.626	3.553	0.920	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86

Conclusion and Inference

- Demand for rented bikes are high during the working hours from 7am to 9am in the morning and 5pm to 7pm in the evening throughout the year
- The most important features who had a major impact on the model predictions were; temperature, hour, wind-speed, solar-radiation, month and seasons.
- When there is no holiday the demand of rented bikes are comparatively higher and has a peak at 7am-9am and 5pm-7pm
- We observed that the demand for rented bikes is only on functioning days People don't use rented bikes when there is no functioning days.
- During summer season the demand of the rented bike is high as compare to other months especially in month of June it is highest, and it is comparatively very low in winter season we may say so due to snowfall.

Conclusion and Inference

- Demand for bikes got higher when the temperature and hour values were more.
- Demand was high for low values of wind-speed and solar radiation.
- People like to ride bikes when it is little windy and sunny day.

Let’s sum up the presentation with some key the point:

- We know that the temperature, wind speed, rainfall and snowfall generally not consistent for every year
- As the features of the given dataset is not always like every year
- So there may be a possibility of randomness according to this model prediction.
- So it need to continuous supervision and modification

Q & A

Thank You