Capstone Project

Seoul Bike Sharing Demand Prediction

By

Ravi Kumar

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 bike count required at each hour for the stable supply of rental bikes.

Data Description

AI

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**

Data Description

AI

- 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)

Insights from the Problem and Data

AI

The following elements may have an impact on the supply and demand for rental bikes:

Weather Condition:

In warmer climates with moderate sunlight, low humidity, quiet winds, and dry roads, people prefer to ride bike. People may prefer to drive a car or other means of transportation than ride a bike if the weather is poor.

Demand for Rental Bikes:

People who don't own bike and those who migrate to cities temporarily need a bike and its an easy means of transportation. Those who possess their own bike may not choose to rent one.

Insights from the Problem and Data

Availability of Rental Bikes:

Bike rental prices and availability at off-peak and rush hour times, as well as on weekdays and weekends.

Stable Supply of Rental Bike:

The weather conditions, demand, and supply sides may all have an impact on the stable supply of rental bikes.

Analysing Dataset

The dataset contains 8760 rows and 14 columns

Check the column data types:

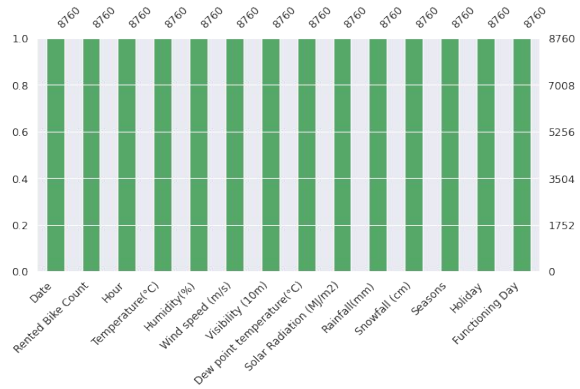
- **From the dataset information, 10 columns are integer types and 4 columns are object types**

Checking for missing values and duplicates:

- **You can receive a quick visual assessment of how complete your dataset is using the customizable and user-friendly missing data visualisations by Missingno.**
- **There are no missing values and duplicates found in the dataset.**

```
df.info()
```

```
<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
```



Split the Dataset into three Categories

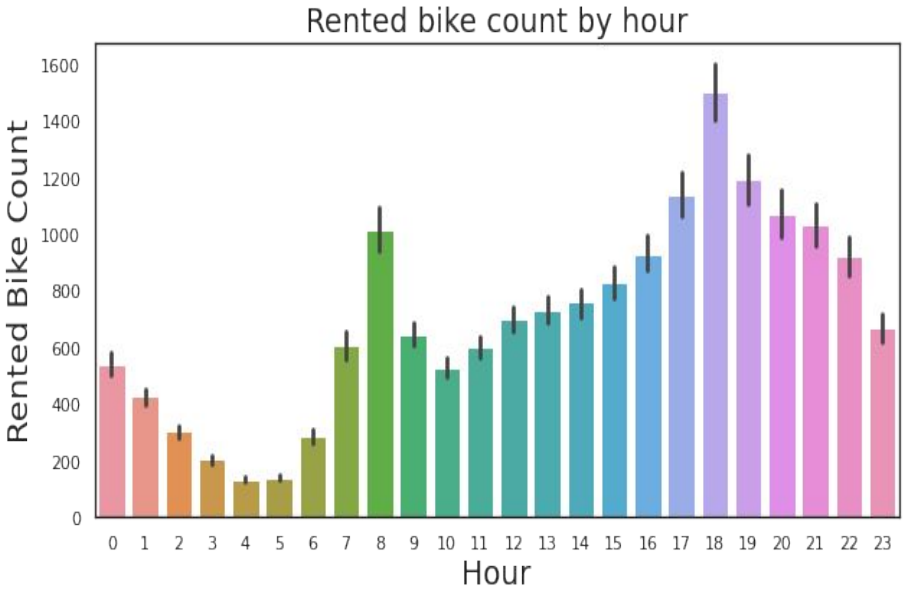
AI

- Date and Time Variables
- Numerical Variables
- Categorical Variables

Date and Time Variables

Barplot is being used to represent hour versus rented bikes.

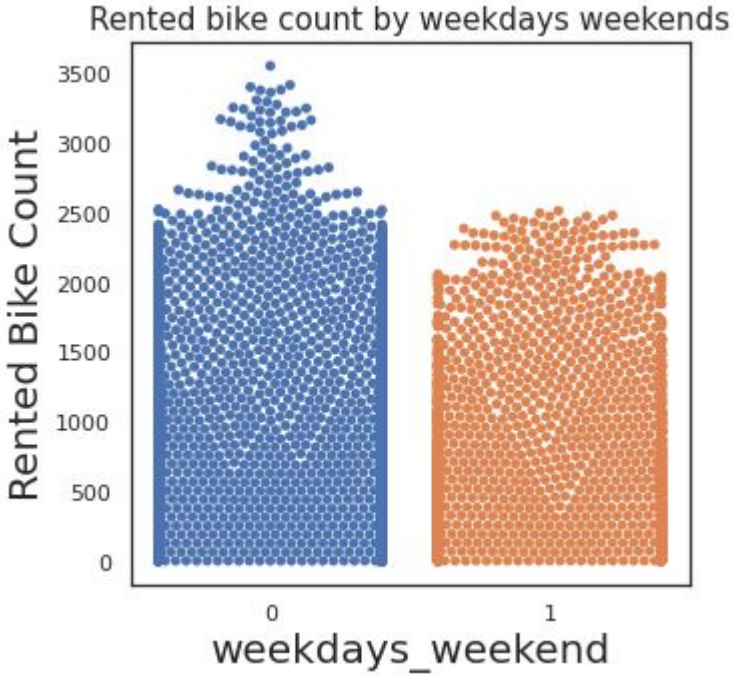
- **From the figure, it is clear, Between 8:00 a.m. and 9:00 p.m. hours, people prefer to ride bikes, which increases demand for bike rentals.**



Date and Time Variables

Swarmplot is being used to represent weekdays_weekend versus rented bikes.

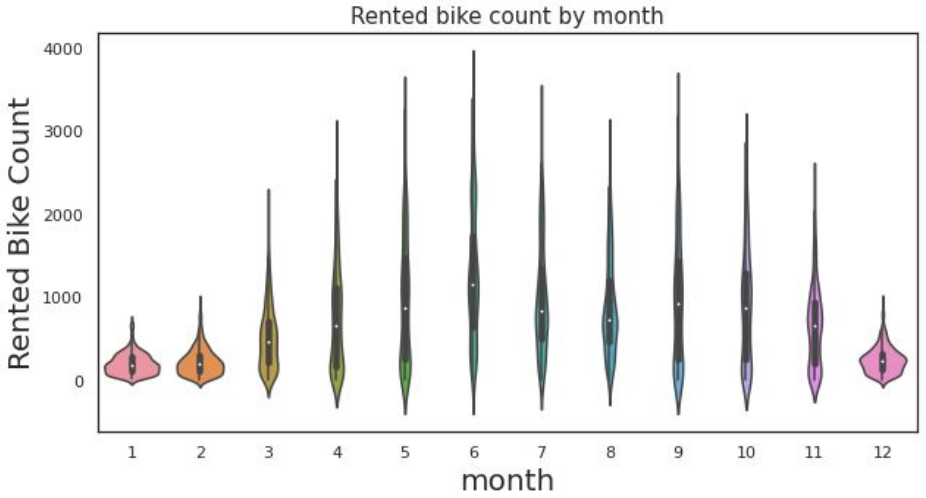
- **From the figure it is clear, weekdays rented more bikes than weekends.**



Date and Time Variables

Violinplot is being used to represent month versus rented bikes.

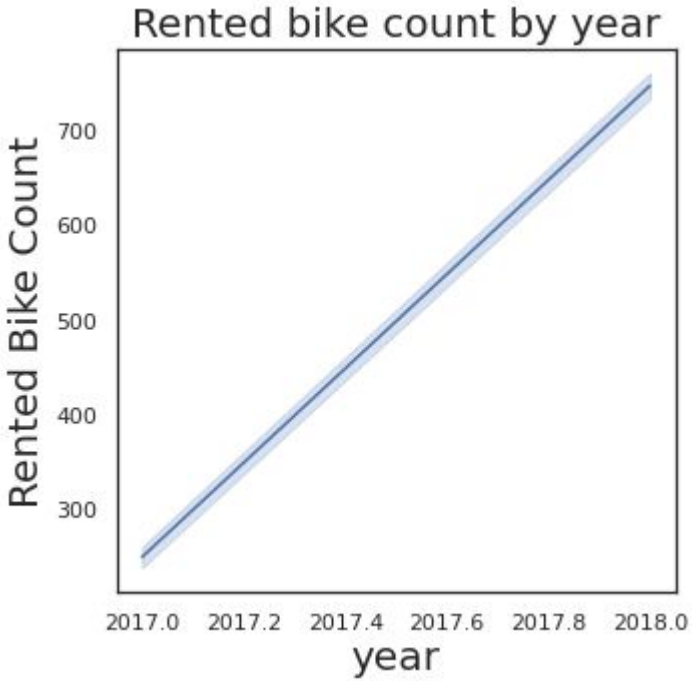
- **From the figure it is clear, In the sixth month, bike rentals are more and took a ride.**



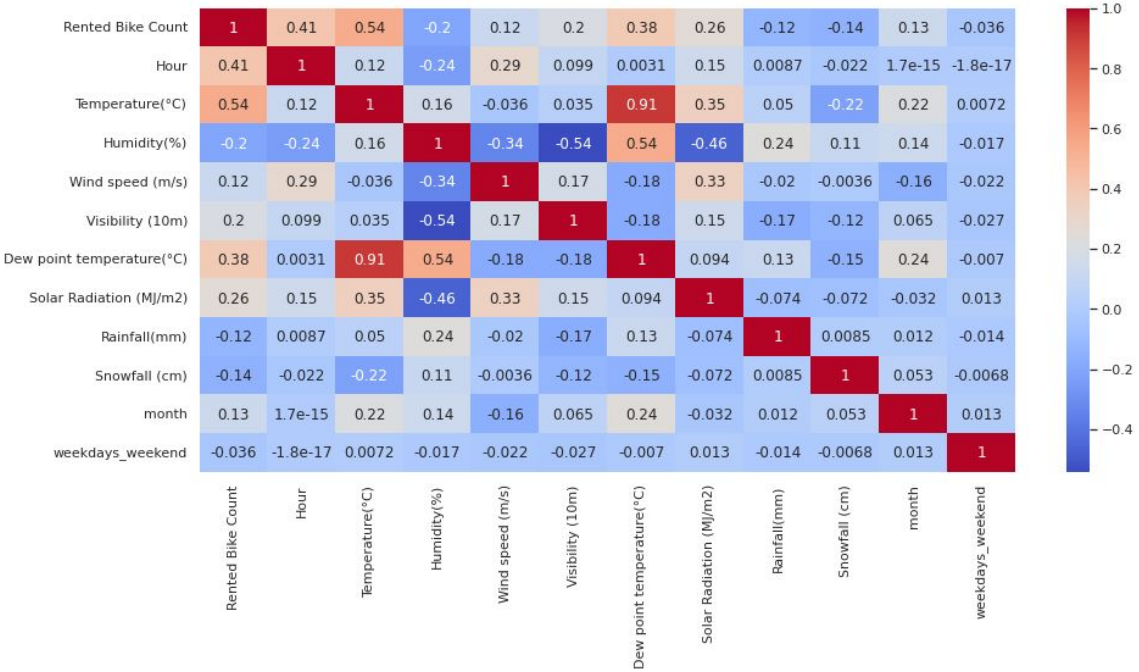
Date and Time Variables

Lineplot is being used to represent year versus rented bikes.

- **From the figure it is clear, The number of bike rentals increased from 2017 to 2018 has the business grows.**



Correlation Analysis



Correlation Analysis

AI

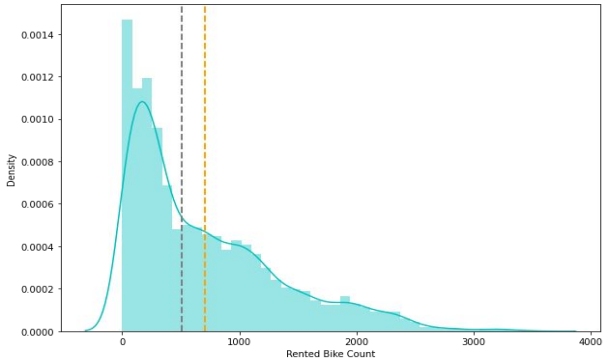
Correlation: You can only determine how much two variables are linearly reliant on one another by looking at their correlations.

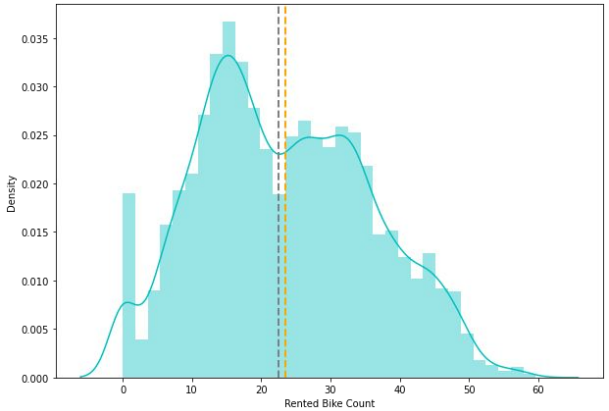
- **The variables are said to be positively connected if the value is positive.**
- **The variables are said to be negatively linked if the value is negative.**
- **It is said that there is no correlation between the variables if the value is zero or very near to it.**
- **Dew point temperature(C) and temperature is highly positive correlation, as temperature increases dew point temperature also increases. Hence, we remove this column, our analysis' results are unaffected.**

Analysis of Dependent Variable

Checking the skewness of the rental bike count distribution

- **We use distplot to check distribution for the output variables in the dataset and if it is skewed, we should do transformation to make it normal distributed.**
- **We can determine the direction of the skewness. The tail of a distribution curve has a longer right side when there is a positive skew and if it is left side, it is negatively skewed.**
- **We reduce the skewness, by square root.**





Outliers

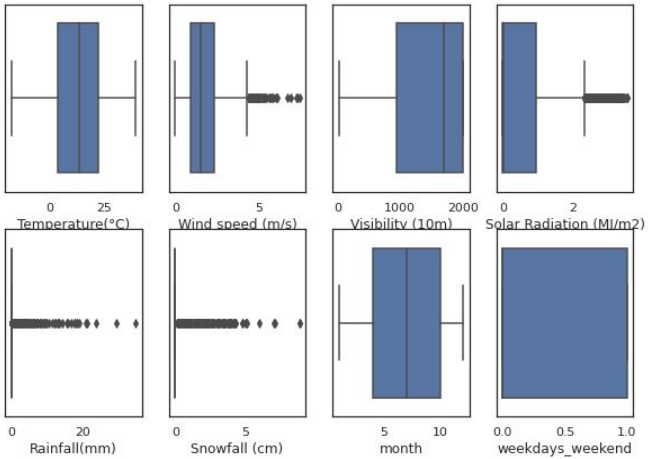
AI

Outlier represent errors in measurement, bad data collection.Outlier detection is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analysed once identified. Machine learning models learn from data to understand the trends and relationship between data points. Outliers can skew results, and impact overall model effectiveness.

- **For identifying outliers, scatter plots and box plots are the most used visualisation techniques.Here we use boxplot.**

Outliers

- Here we can see that the columns that contain outliers are Rainfall, Snowfall, Wind Speed and Solar Radiation
- We have removed the outliers through IQR

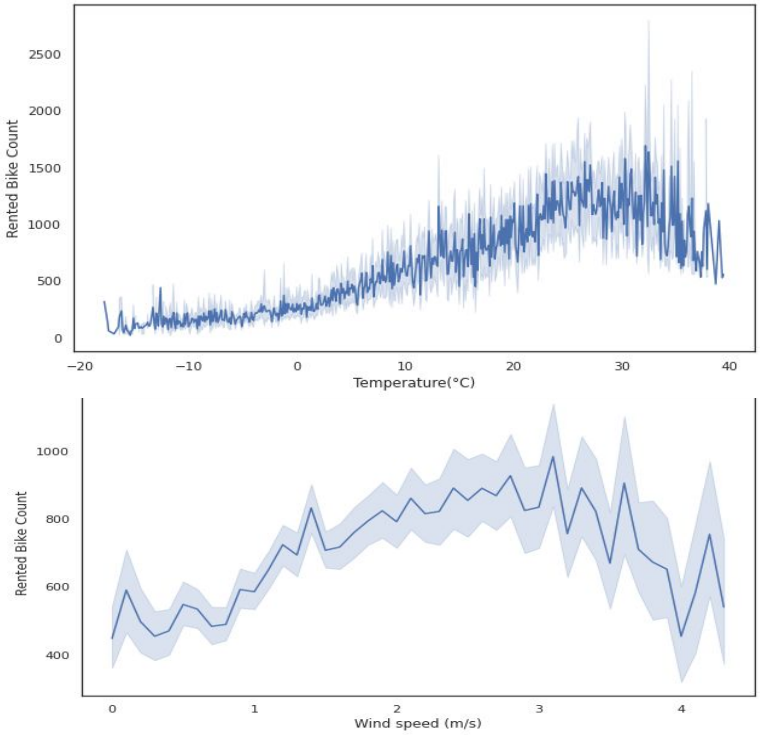


Handling Outliers: Through the removal of outliers, some null values have been added to these four columns. At this point, we have two options: either we can infer meaningful complete values to the observations with null values, or we can remove them. In this instance, we will impute them using the median value of each column.

Numerical Variables

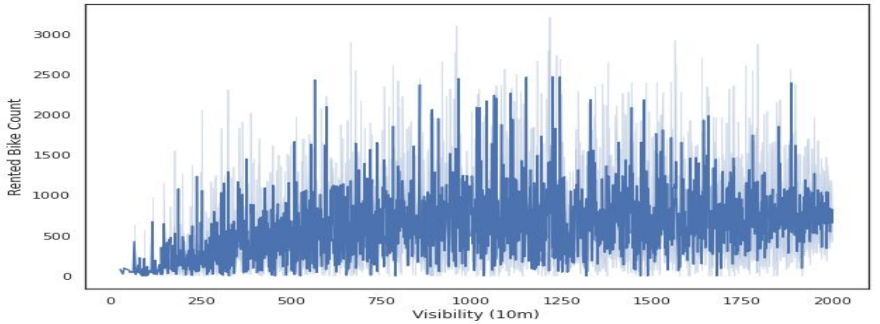
From the lineplot, the distribution between the numerical variable and the dependent variable shows:

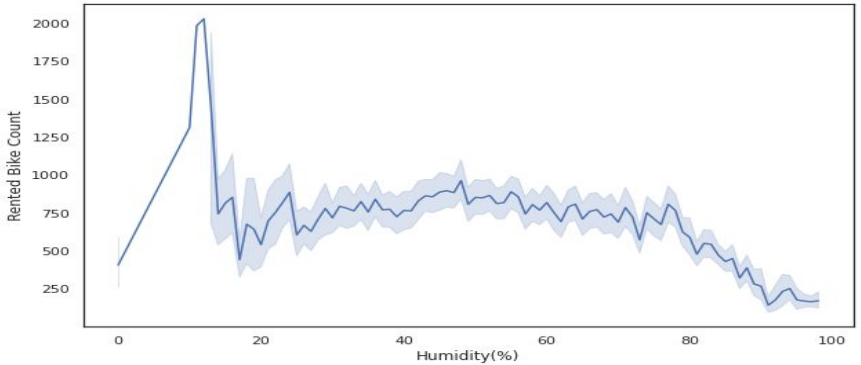
- **People enjoy riding bikes in warmer temperature.**
- **As the wind speed increases between 2 and 3 (m/s), people prefer to hire and ride bike.**



Numerical Variables

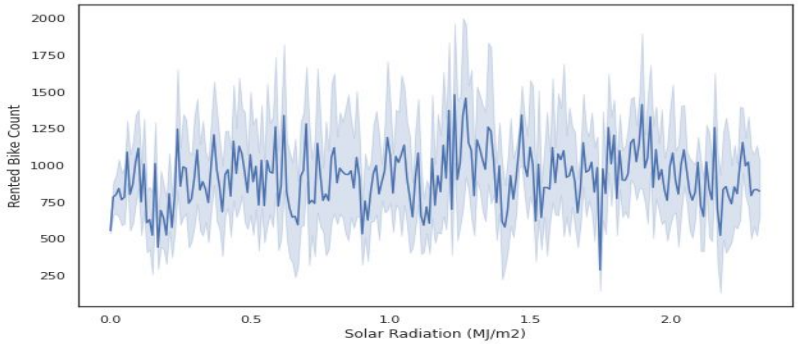
- People favour biking in areas with great visibility.
- People prefer to ride bikes where humidity below 20%.

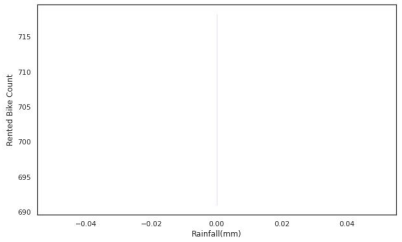


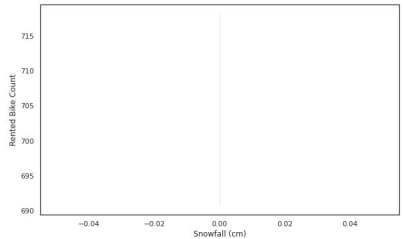


Numerical Variables

- When there is solar radiation, there are a lot of leased bikes—the rental counter is approximately 1000.
- People prefer to hire and ride bikes when there is no rainfall and snowfall.

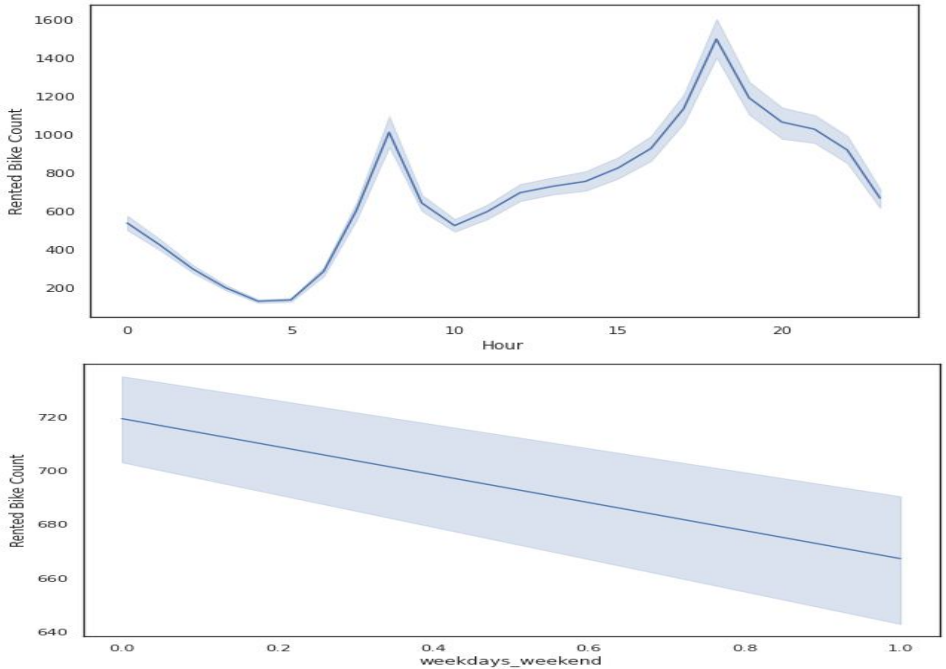






Numerical Variables

- people like to ride a bike in the 8th and 18th hour.
- Weekdays are the preferred days for people to hire and ride bike than weekends.

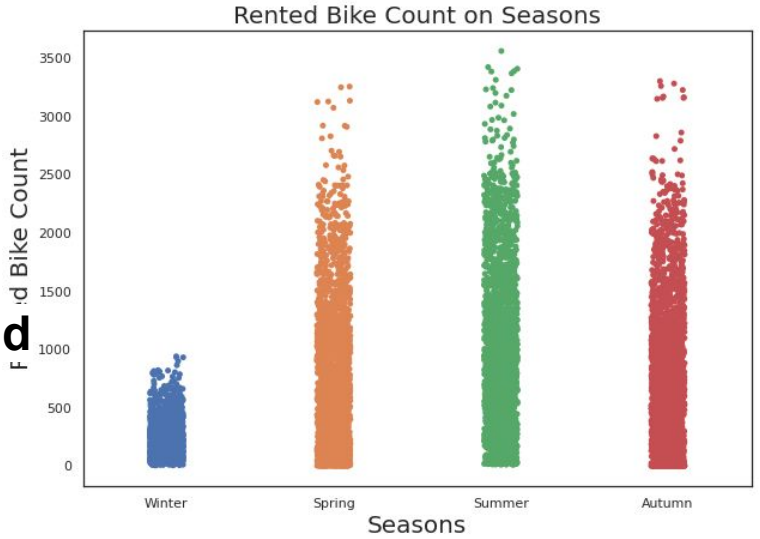


Categorical Variables

Three of the 14 columns—the seasons, holidays, and functioning days—are categorical data types.

From the plot between the categorical variable and the dependent variable shown:

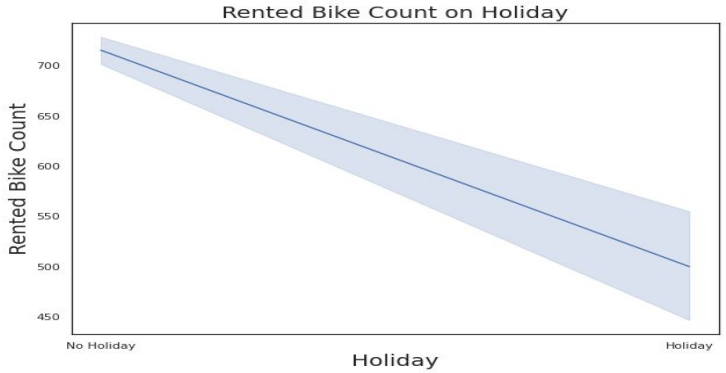
- Summer is the season with the most demand for bikes followed by Autumn, Spring, and Winter.
- Bike rentals are highest during the summer and are lowest during the winter seasons.

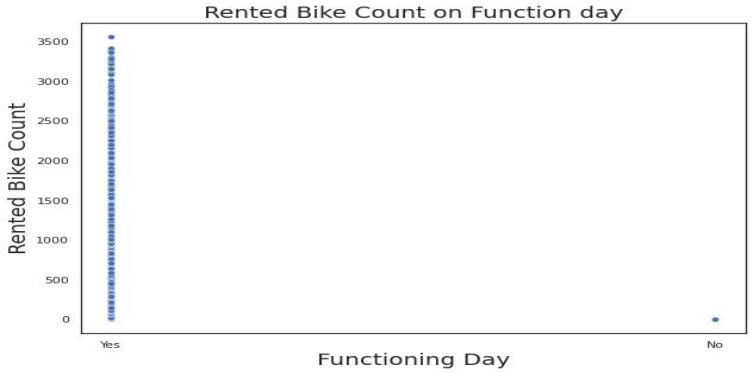


Categorical Variables

According to the plot:

- **More people hired bikes during non-holiday times than during holiday times.**
- **People leased bikes when the day was operational, but they didn't rent any on non-operational days.**

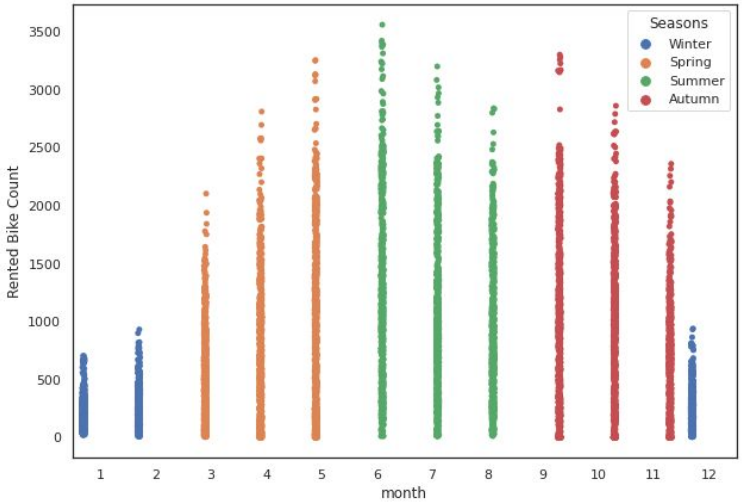




Categorical Variables

From the plot:

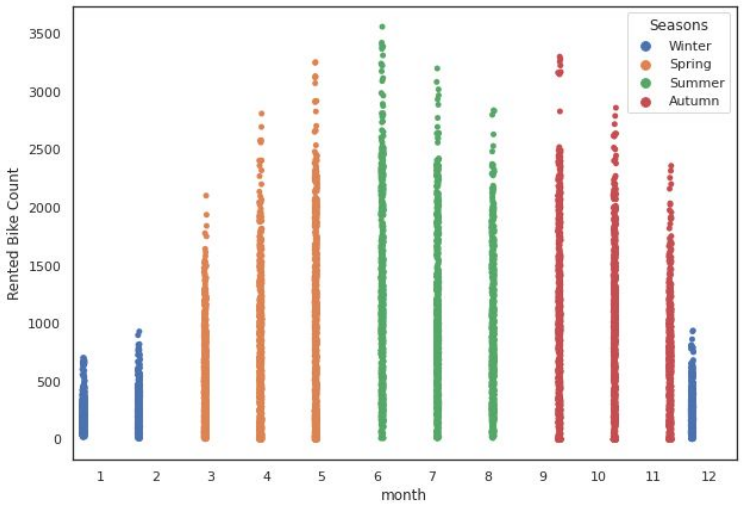
- **Most bikes were hired in the sixth, seventh, and eighth months of the summer.**
- **When compared to summer, the number of bikes hired in the ninth, tenth, and eleventh months of the autum season is quite low.**



Categorical Variables

- Comparatively fewer bikes are hired during the winter months of first, second, and third months than during the summer, autumn, and spring.

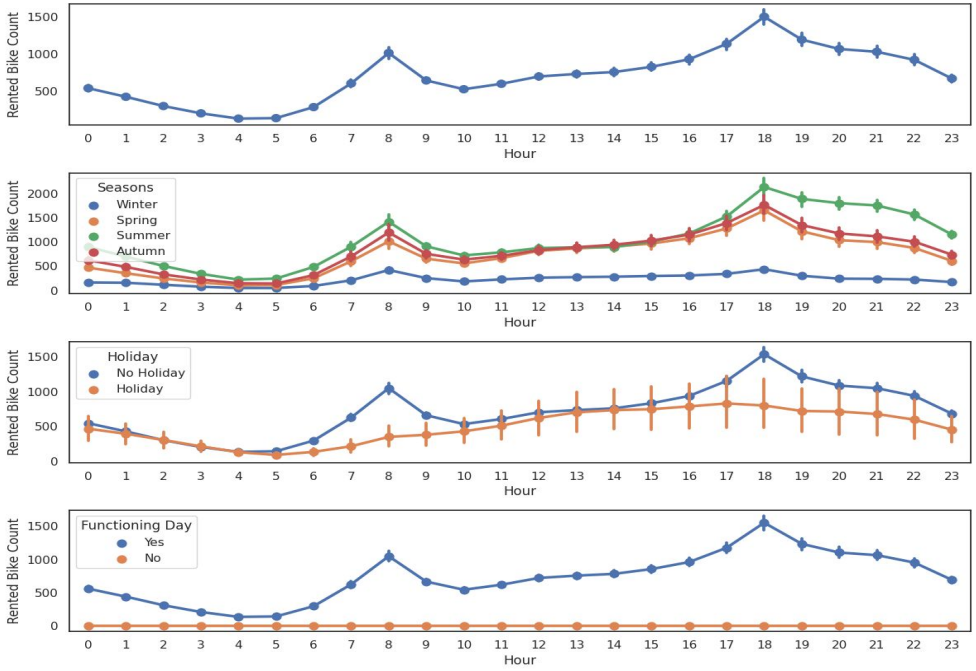
Summer Season is the peak season where most people hired a bike. There is a lot of demand for bikes in summer season than compare to other seasons



Pointplot Representation

From the plot:

- **In the summer, when there are no holidays and the day is still operating, people choose to rent and ride more bikes at that time.Bikes are in great demand and good supply at 18 hours during the summer.**



Feature Engineering

AI

Feature encoding:

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones.

- **The process of converting categorical data in a dataset into numerical data is called feature encoding.**
- **we have encoded Seasons, Holiday and Functioning Day.**

Model Performance

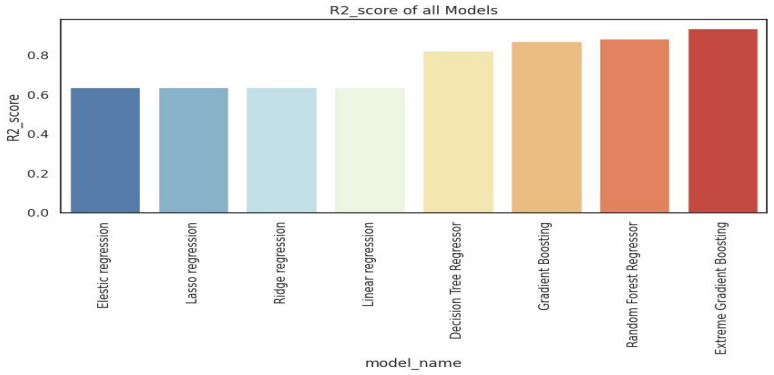
AI

- Linear regression
- Lasso regression
- Ridge regression
- Elastic regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Extreme Gradient Boosting

Result of all models

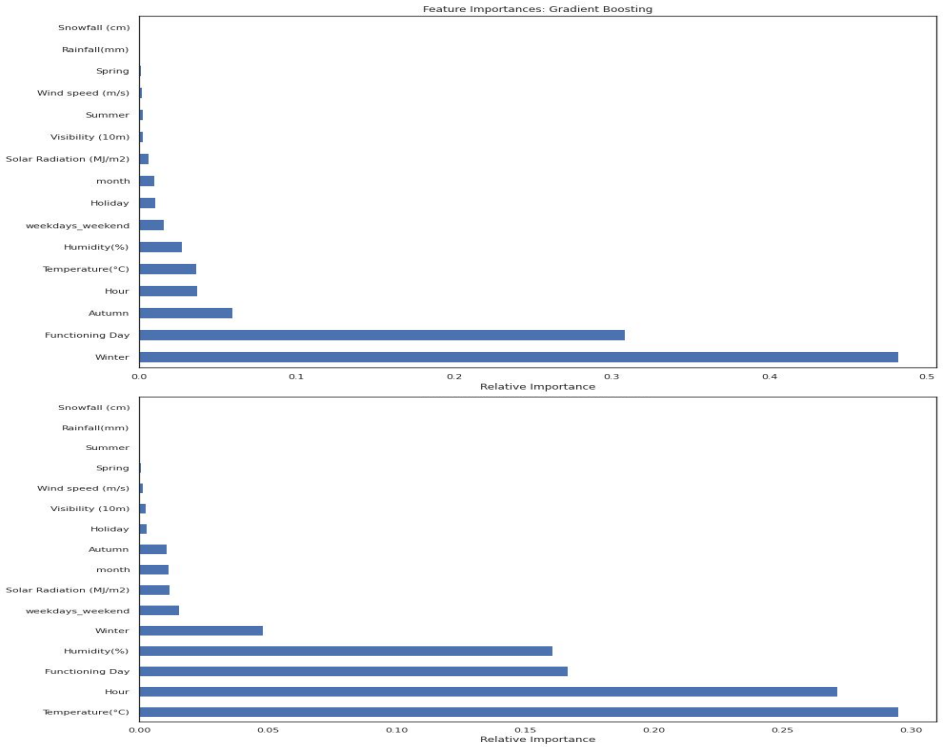
- Comparing the result of all the model, XGBoost and Random Forest gives the best R2 score.
- We can deploy this models.

model_name	MSE_score	RMSE_score	MAE_score	R2_score	Adjusted_R2_score
Linear regression	56.945815	7.546245	5.859794	0.639507	0.636852
Lasso regression	56.969414	7.547809	5.861282	0.639357	0.636702
Ridge regression	56.946640	7.546300	5.859868	0.639502	0.636847
Elastic regression	56.983494	7.548741	5.862294	0.639268	0.636612
Decision Tree Regressor	27.915300	5.283493	3.542101	0.823283	0.821982
Random Forest Regressor	18.127895	4.257687	2.957885	0.885242	0.884397
Gradient Boosting	20.139894	4.487749	3.185413	0.872505	0.871566
Extreme Gradient Boosting	10.165641	3.188360	1.997298	0.935647	0.935173



Feature Importances

- The most crucial feature taken into account when utilising a XGboost to make predictions is winter and Functioning day.
- The most crucial feature taken into account when utilising a random forest to make predictions is temperature and hour.



Summary and Conclusion:

- In the morning and evening, many people rent bikes between 8:00 a.m. and 9:00 p.m. which increases demand for bike rentals

People prefer to ride bikes:

- when the weather is warm (20° to 28°C),
- the wind speed is between 2 and 3 m/s,
- where humidity below 20%,
- when there is no rainfall and snowfall and there is adequate visibility.
- Weekdays rented more bikes than weekends.
- In the sixth month, people rented more bikes and took a ride.
- The number of bike rentals increased from 2017 to 2018 has the business grows.

Summary and Conclusion:

AI

- More people hired bikes during non-holiday times than during holiday times.
- People leased bikes when the day was operational, but they didn't rent any on non-operational days.
- Most bikes were hired in the sixth, seventh, and eighth months of the summer.
- When compared to summer, the number of bikes hired in the ninth, tenth, and eleventh months of the autumn season is quite low.
- Comparatively fewer bikes are leased during the third and fourth months of the spring season than during the summer and the Autumn.
- Summer Season is the peak season where most people hired a bike. There is a lot of demand for bikes in summer season than compare to other seasons

Summary and Conclusion:

AI

- With a r2 score of 93%, 88% , Extreme Gradient Boosting and Random Forest is the model that performs the best. We can deploy this models into production.
- Therefore, there will be high supply and demand for bikes when there is mild sunlight, low humidity, low wind, and dry roads. There will be a lot of demand and supply on both sides when the weather is pleasant.

Thank you