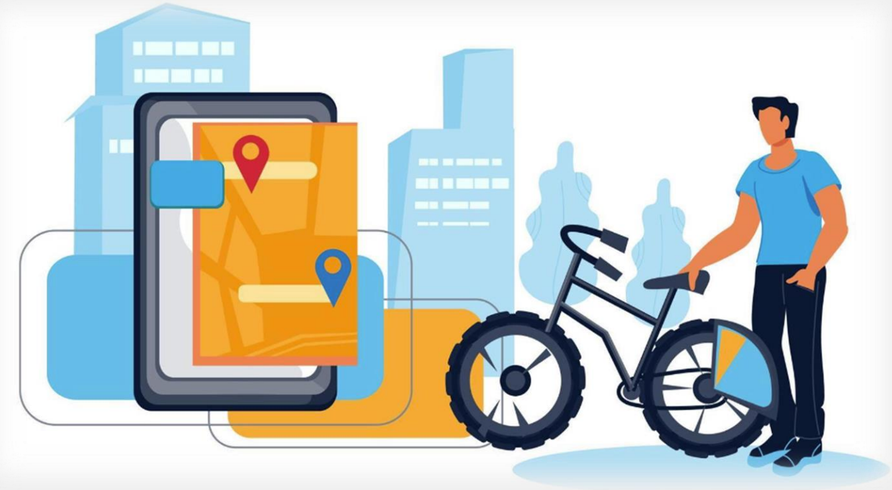
MINI PROJECT

BIKE SHARING DEMAND PREDICTION



Mrudula A P

POINTS FOR DISCUSSION

- Problem Statement
- Data Description
- Data Preparation and Cleaning

- Exploratory Data Analysis
- Hypothesis Testing
- Feature Engineering
- Model Implementation
- Model Interpretation
- Conclusion

PROBLEM STATEMENT

Unleashing the Power of Bike Sharing with Precision Demand Predictions

The introduction of rental bikes in various urban areas has notably enhanced mobility comfort. It is imperative to ensure the timely availability and accessibility of rental bikes to minimize waiting times. The imperative of sustaining a stable supply of rental bikes in the city is a pressing concern. Accurately gauging bike demand at specific times is especially vital for bike rental companies.

Our primary goal is to address this challenge by creating a solution that forecasts bike demand, considering diverse factors such as city weather conditions and different times of the day.

DATA DESCRIPTION

- The Bike Sharing Demand Dataset contains information about bike rentals in Seoul from Dec 2017 to Nov 2018. It includes hourly observations of bike rentals, such as the date, time, number of rented bikes, weather conditions, and other factors that may influence bike rental demand.
- The dataset that we are working with contains 8,760 observations and 14 features.

	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
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

DATA DICTIONARY

- **Date**: The date of the observation.
- **Rented Bike Count**: The number of bikes rented during the observation period.
- **Hour**: The hour of the day when the observation was taken.
- **Temperature(°C)**: The temperature in Celsius at the time of observation.
- **Humidity(%)**: The percentage of humidity at the time of observation.
- **Wind speed (m/s)**: The wind speed in meters per second at the time of observation.
- **Visibility (10m)**: The visibility in meters at the time of observation.
- **Dew point temperature(°C)**: The dew point temperature in Celsius at the time of observation.
- **Solar Radiation (MJ/m2)**: The amount of solar radiation in mega-joules per square meter at the time of observation.
- **Rainfall(mm)**: The amount of rainfall in millimeters during the observation period.
- **Snowfall(cm)**: The amount of snowfall in centimeters during the observation period.
- **Seasons**: The season of the year when the observation was taken.
- **Holiday**: Whether the observation was taken on a holiday or not.
- **Functioning Day**: Whether the bike sharing system was operating normally or not during the observation period.

DATA PREPARATION & CLEANING

- There were no duplicate rows in the dataset.
- There were no missing values in the dataset.
- Changed datatype of Date to datetime.
- Created new columns for better visualize the data

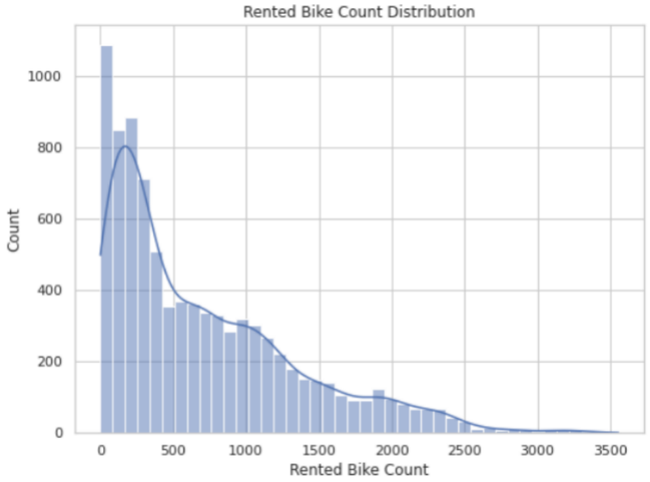
```
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
df['weekday'] = df['Date'].dt.day_name()
```

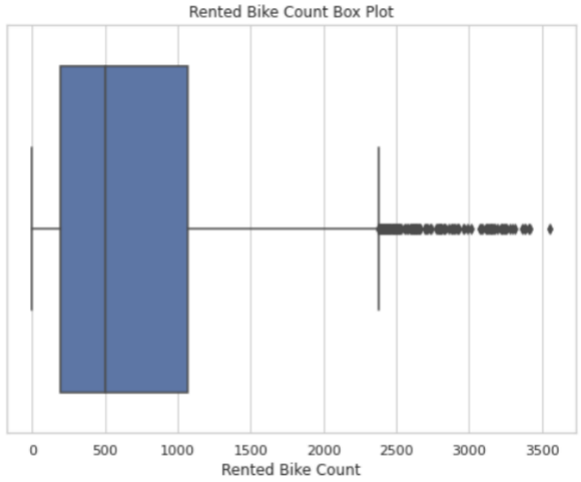
1. Year, Month, Day, Weekday from Date
2. Temperature Bin from Temperature(°C)

- Changed Data types of numerical columns which represents categories like Year, Month, Day to categorical data type

EXPLORATORY DATA ANALYSIS

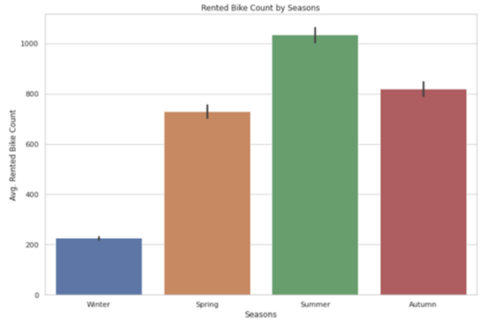
Rented Bike Count Distribution





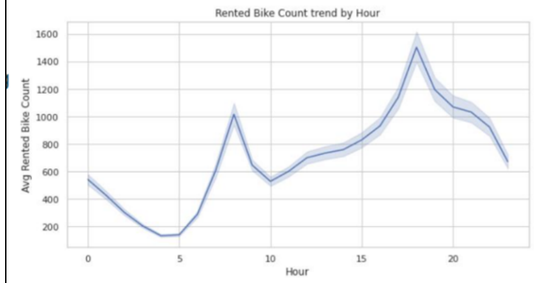
Rented Bike Count by Seasons

- Rental Bike demand in winter season is significantly lower than other months.
- Demand is highest in Summer



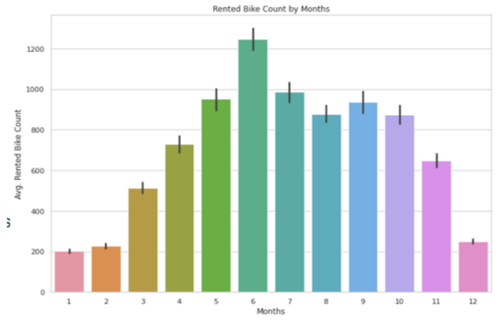
Rented Bike Count by Hour

- Rental Bike demand peaks during rush hours of the day. Rush hour is generally around 8AM in the morning and 6PM in the evening.



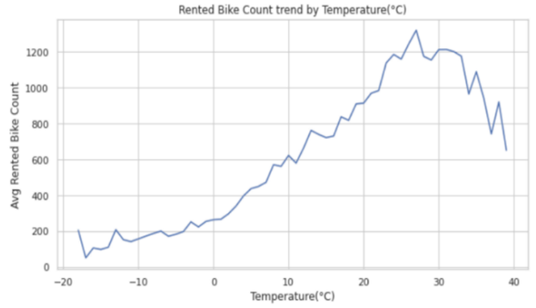
Rented Bike Count by Months

- Rented Bike demand decreases significantly during winter months like Dec, Jan, Feb etc.
- Demand peaks at summer months like May, June July etc.



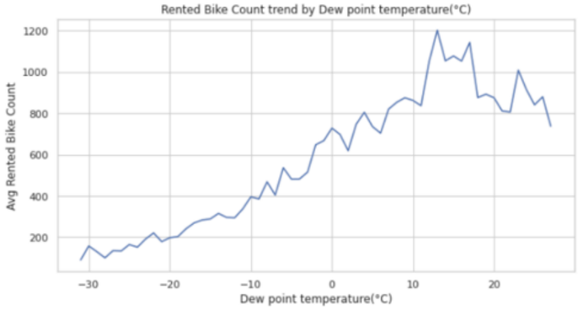
Rented Bike Count by Temperature

- Rented Bike demand increases as the temperature increases.
- Although too high temperature leads to decrease in demand again.



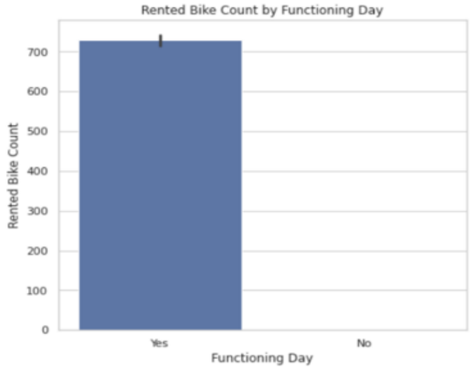
Rented Bike Count by Dew Point Temperature

The demand increases as the temperature increases. Although too high dew point temperature leads to decrease in demand again.



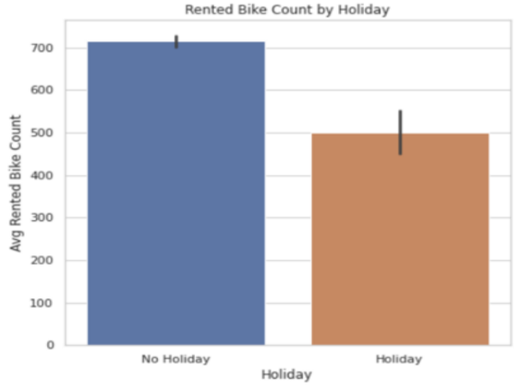
Rented Bike Count by Functioning Day

Obviously on non functioning day i.e., when the bike renting service was not operating, there was zero bikes rented.



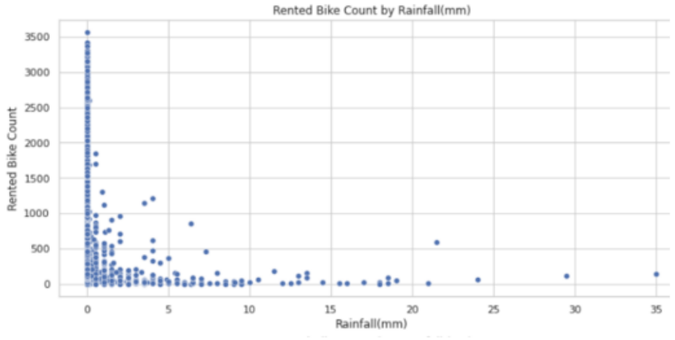
Rented Bike Count by Holiday

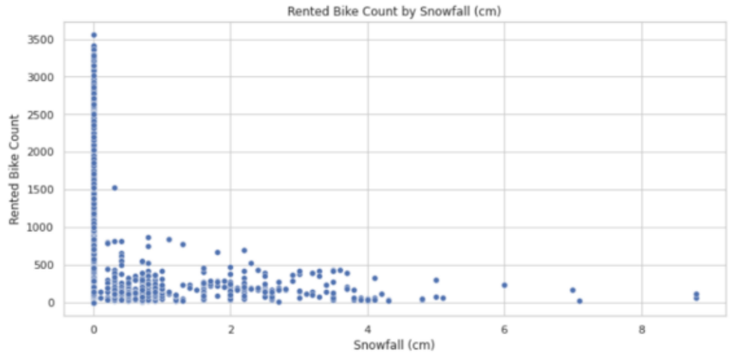
Rental Bike demand is higher on non holiday compared to holiday. Possible reason could be that a lot of people uses rental bike to go to offices or schools/colleges on non holiday.



Rented Bike Count by Rainfall and Snowfall

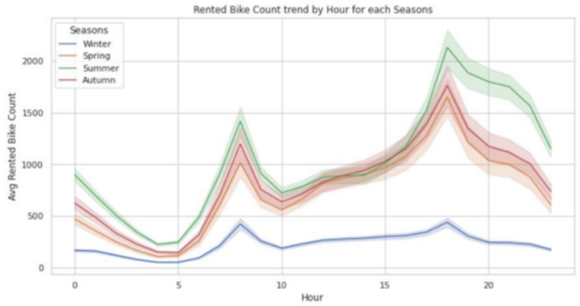
Bike rental demand tends to decline during periods of rainfall and snowfall, as people typically prefer not to use bikes in inclement weather unless it is absolutely necessary due to safety and comfort considerations.





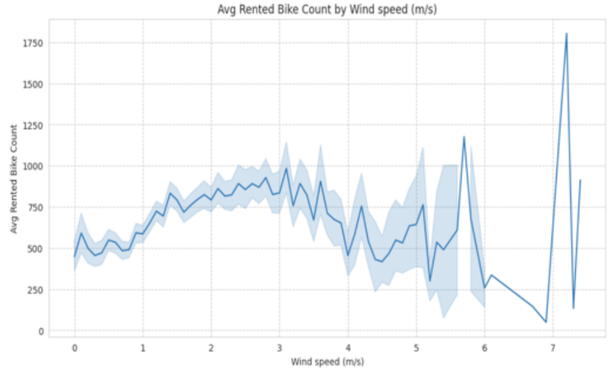
Rented Bike Count by Hour for each Season

The demand for rented bike peaks during rush hours of the day and each season has similar hourly pattern only levels are different



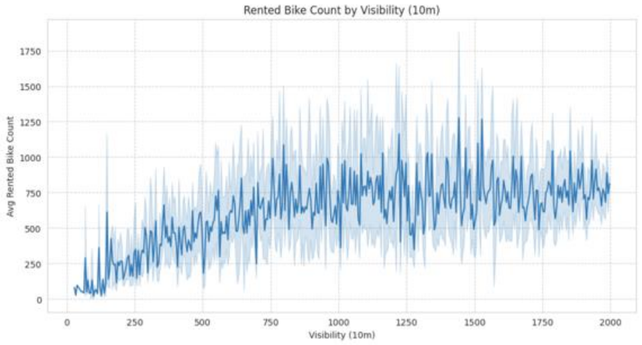
Rented Bike Count by Windspeed

- Initially, increased wind speed boosts rental bike demand, enhancing weather conditions.
- Yet, excessive wind speeds causing storms result in a decline due to adverse and unsafe conditions, highlighting the delicate balance in demand.



Rented Bike Count by Visibility (10mm)

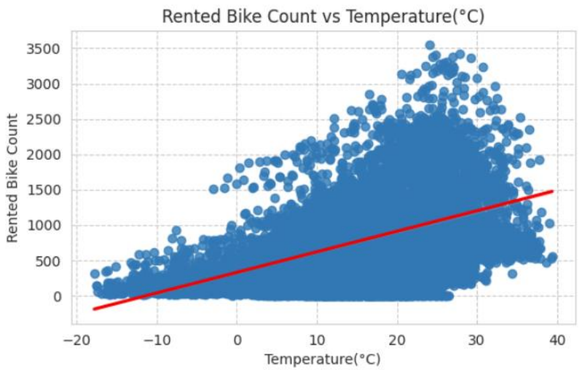
Bike demand rises with increased visibility, evident as people prefer clear days. Yet, beyond a certain point, visibility growth doesn't affect demand.

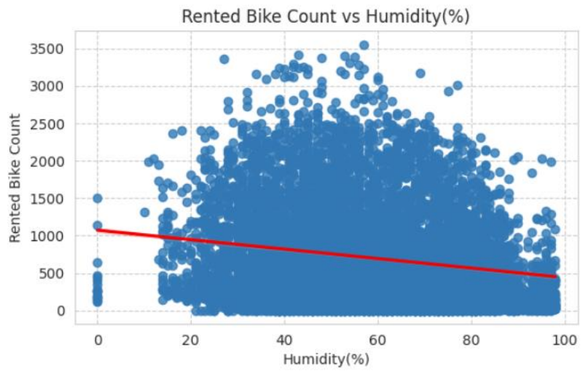


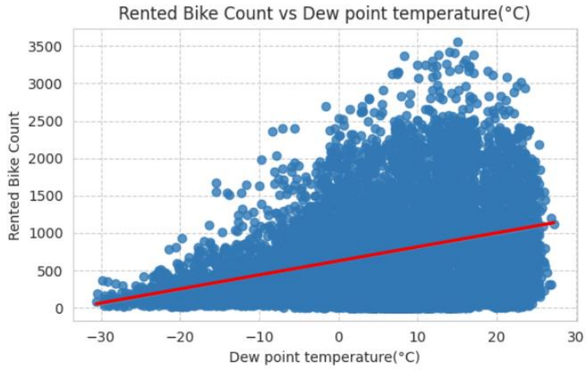
Heatmap : Correlation of features

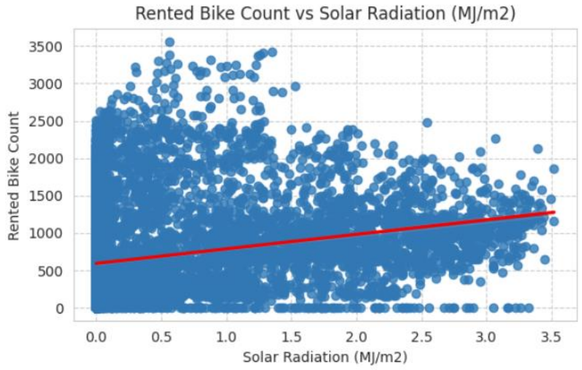
Temperature and Dew Point Temperature are highly correlated which can create problem while doing model interpretation.

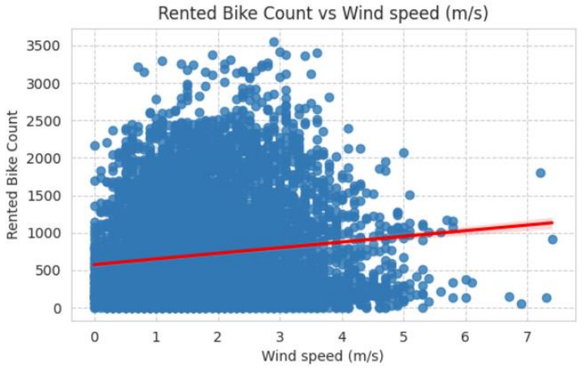
Regression Plot : Numerical features vs Rented Bike Count

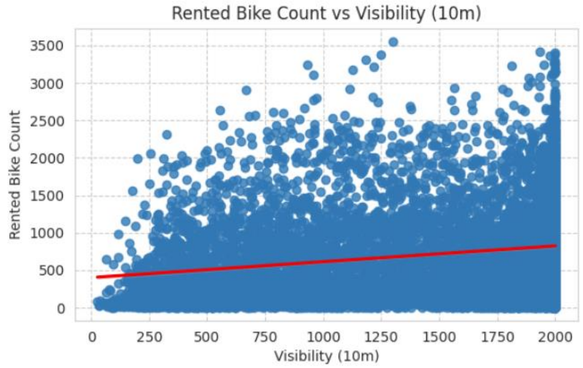


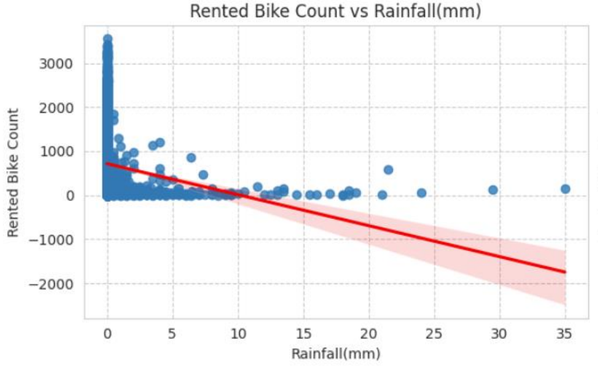


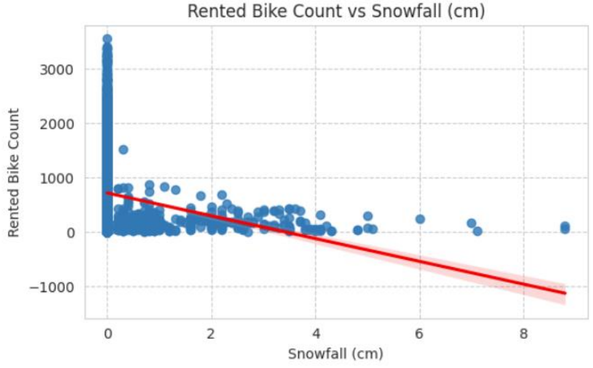












HYPOTHESIS TESTING

1. Rented Bike Demand in hot weather is higher compared to demand in cold weather.

Null Hypothesis: $H_o : \mu_{cold} = \mu_{hot}$

Alternate Hypothesis : $H_1 : \mu_{cold} \neq \mu_{hot}$

Test Type: Two-sample t-test

T - Statistic : 42.27607
p-value : 0.0

Since p-value is less than 0.05, we reject the null hypothesis.ie, Rented Bike Demand in hot weather is higher compared to demand in cold weather.

2. Rented Bike Demand during rush hour (7-9AM & 5-7PM) is higher compared to non- rush hour.

Null Hypothesis: $H_o : \mu_{rush} = \mu_{non-rush}$

Alternate Hypothesis : $H_1 : \mu_{rush} \neq \mu_{non-rush}$

Test Type: Two-sample t-test

T - Statistic : 22.54238
p-value : 9.381784283723713e-104

Since p-value is less than 0.05, we reject null hypothesis.ie, Rented Bike Demand during rush hour (7-9AM & 5-7PM) is higher compared to non- rush hour.

3. Rented Bike Demand is different in different seasons with highest in summer and lowest in winter.

Null Hypothesis: H_o : **No significant difference** between rented bike counts for different seasons.

Alternate Hypothesis : H_1 : **Significant difference** between rented bike counts for different seasons.

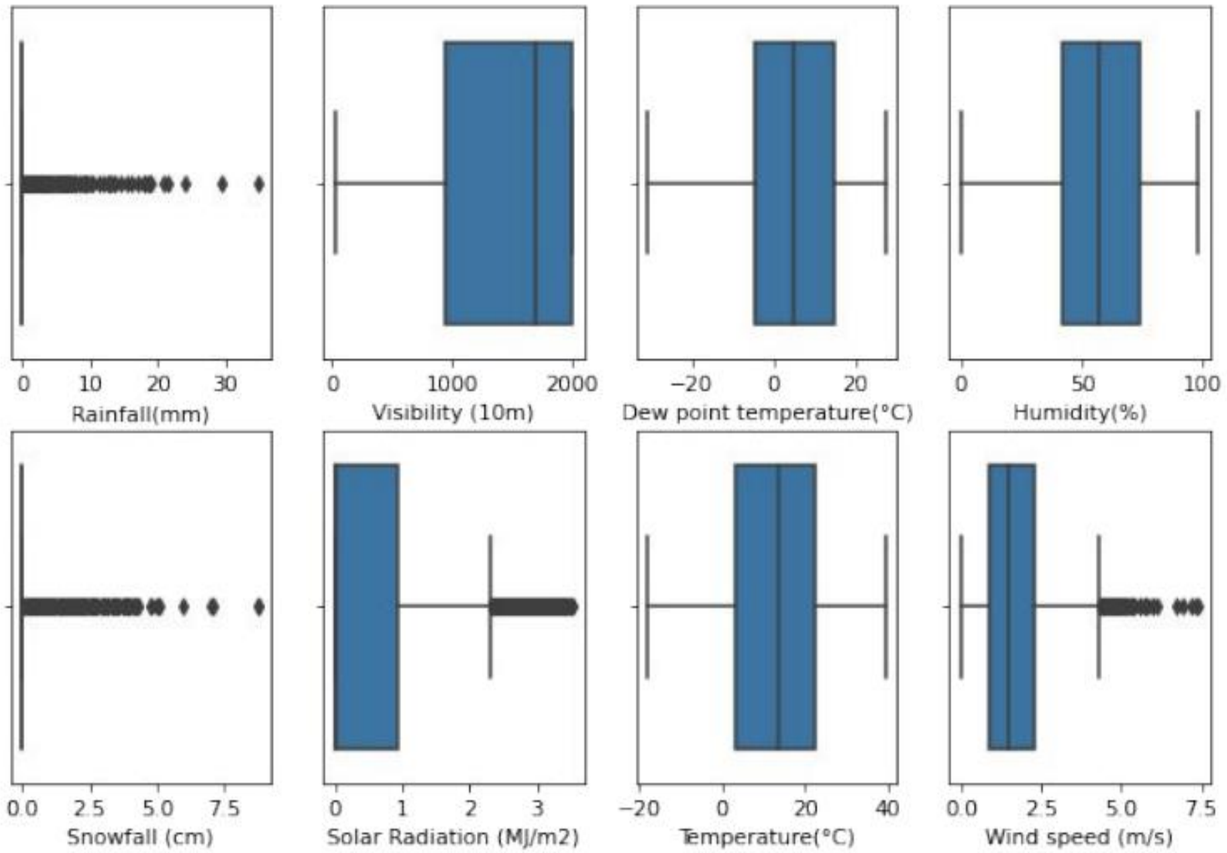
Test Type: One-way ANOVA test

F - Statistic : 776.467815
p-value : 0.0

Since p-value is less than 0.05, we reject null hypothesis.ie, Rented Bike Demand is different in different seasons with highest in summer and lowest in winter.

FEATURE ENGINEERING

OUTLIER TREATMENT



From distribution plots of different features, outliers are present in

- **Rainfall(mm), Snowfall (cm) columns**

Most of the values are zero and few are non zero which is understandable as we don't see rain and snow everyday. Based on the nature of data, it is unlikely that the nonzero values represent outliers. Hence used **99th quantile** for capping outliers.

- **Wind speed (m/s), Solar Radiation (MJ/m2) columns**

The distribution of values are right skewed. Hence used **IQR method** for capping outliers.

❖ The NULL values created in place of these outliers are imputed with the median value of that particular column or field.

FEATURE ENCODING

Machine learning models can only work with numerical values and therefore we have to turn the categorical columns to numerical columns, and this is achieved by feature encoding.

- Used **ONE-HOT ENCODING** to convert categorical columns Seasons, Holiday, Month, Weekday columns to numerical columns.

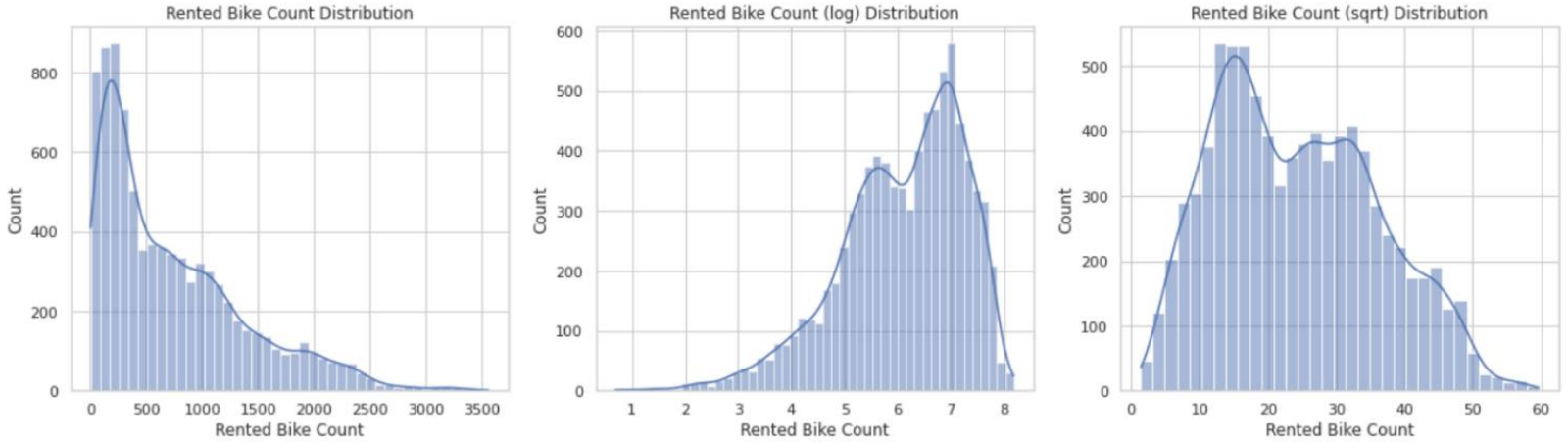
MULTI- COLLINEARITY CHECK FOR FEATURE REMOVAL

Multicollinearity is when two independent variables are highly correlated to each other.

- Used VIF for checking multicollinearity.
- Dew Point Temperature is highly correlated to Temperature. Hence dropped Dew Point Temperature column.

	feature	VIF
1	Temperature(°C)	33.678675
4	Dew point temperature(°C)	17.629784
3	Visibility (10m)	9.157289
2	Humidity(%)	5.803109
7	Wind speed (m/s)_capped	4.837381
0	Hour	4.419686
8	Solar Radiation (MJ/m2)_capped	2.866329
9	is_weekend	1.416439
5	Rainfall(mm)_capped	1.174577
6	Snowfall(cm)_capped	1.150386

- The distribution of Rented Bike Count (target variable) was right skewed, and to train a robust model we can transform it to normal. Applied **square root** to transform it to normal.

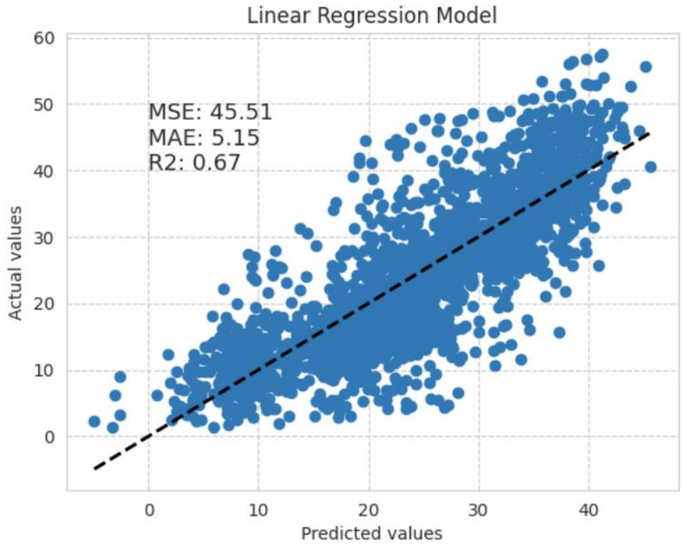


- Similarly applied square root to Wind Speed (m/s) to transform it to normal as it was originally skewed

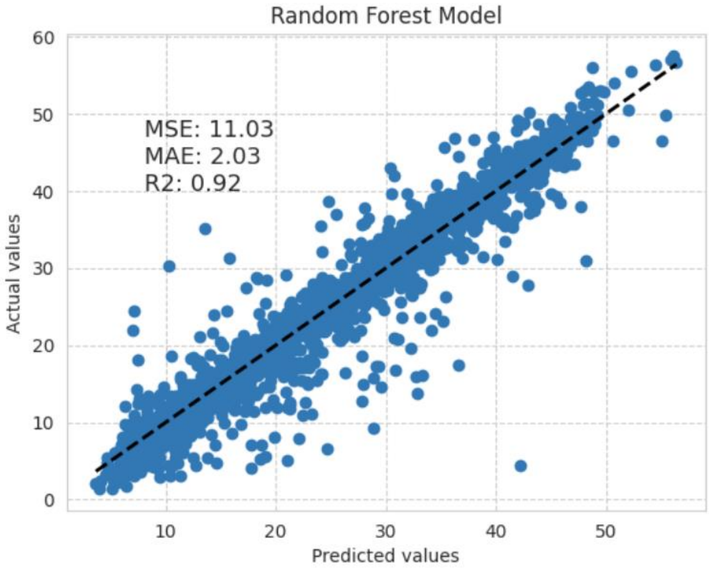
MODEL IMPLEMENTATION

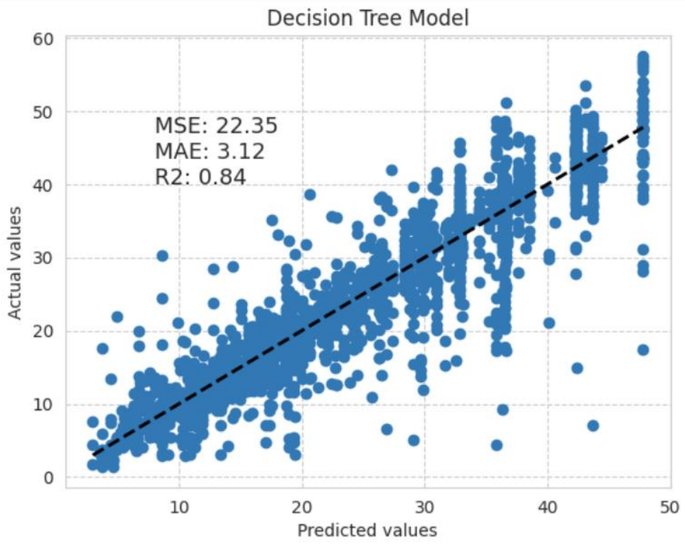
- The Rented Bike Count (target variable) is continuous in nature.
- Used various regression algorithms along with hyper parameter tuning and cross validation to get the best model.
- Algorithms used:
  - Linear Regression
  - XGBoost Regressor
  - Decision Tree Regressor
  - Random Forest Regressor

MODEL PERFORMANCE EVALUATION

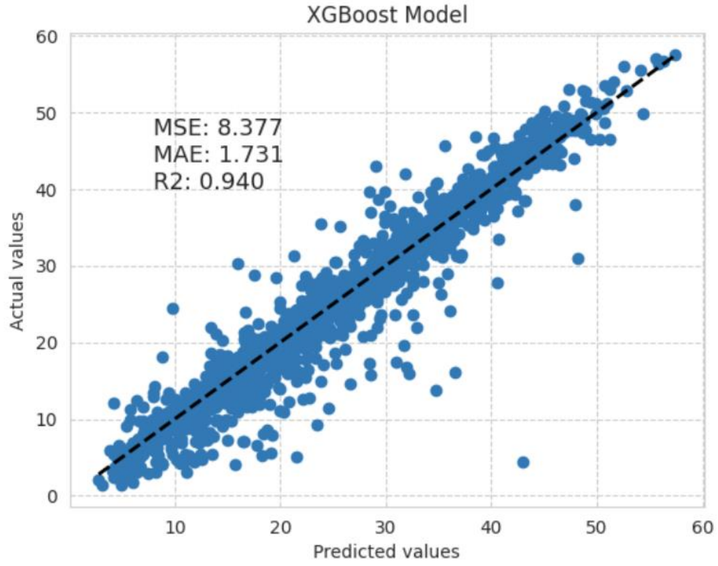


	R2	MSE	MAE
XGBoost	0.940	8.377	1.731
Random Forest	0.920	11.027	2.031
Decision Tree	0.839	22.350	3.123
Linear Regression	0.672	45.506	5.148

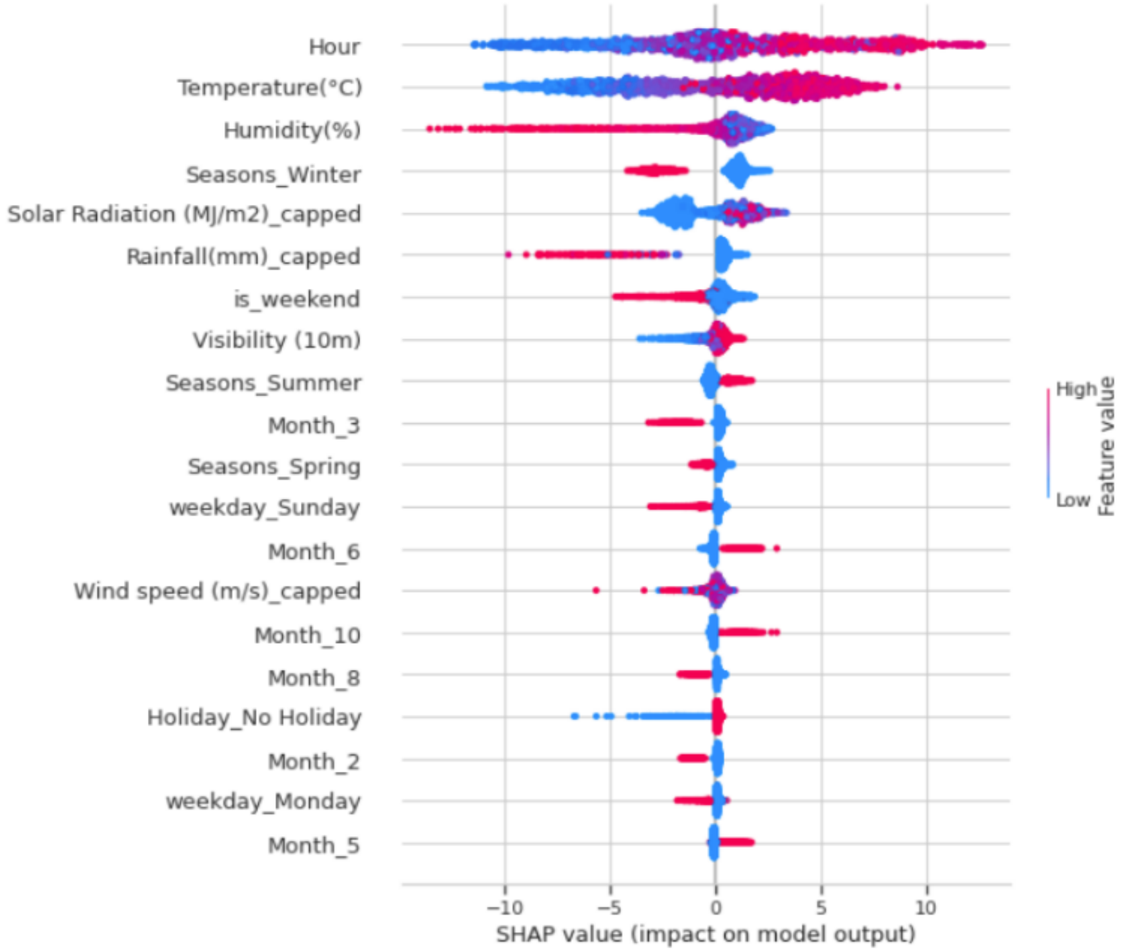


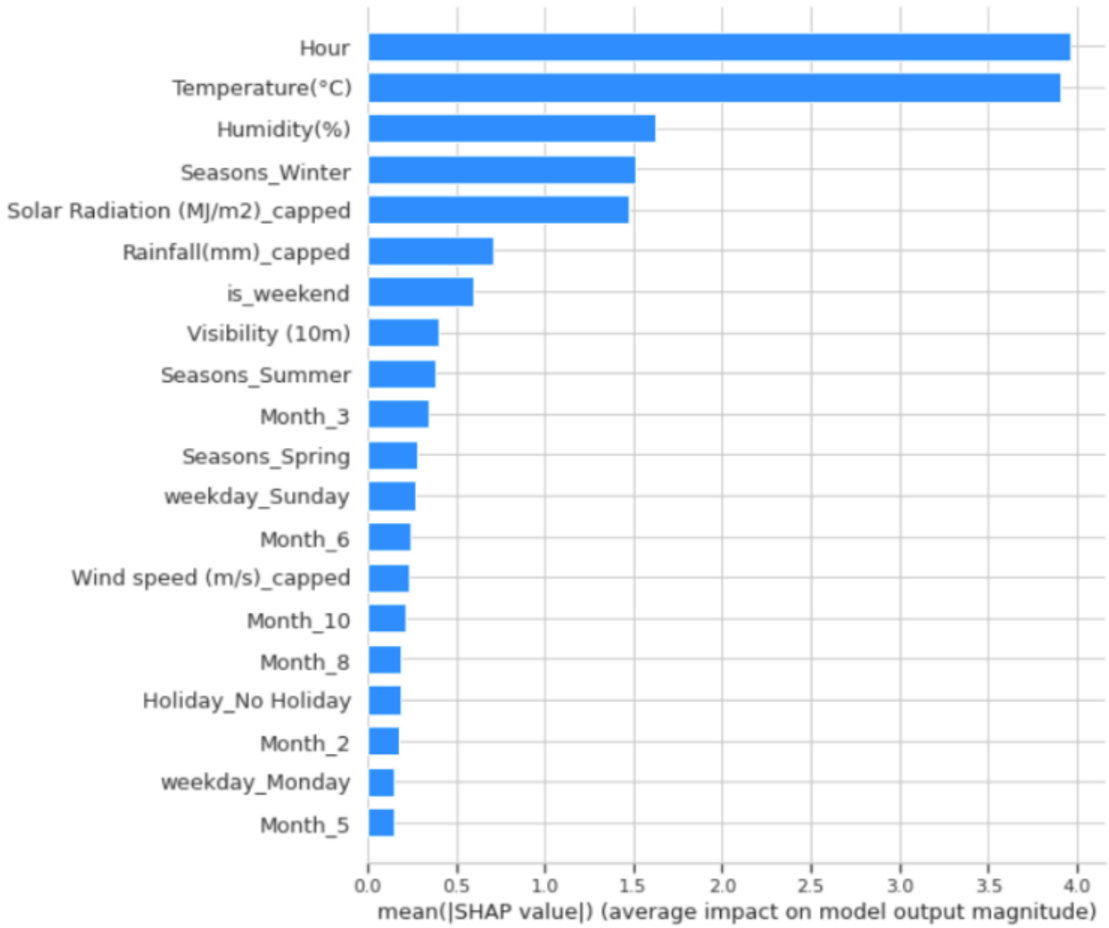


The XGBoost model is the best choice for this dataset as it has the highest R2 score, the lowest MSE, and the lowest MAE



MODEL INTERPRETATION





CONCLUSION

- Challenges faced:
  - ❖ Removing Outliers.
  - ❖ Encoding the categorical columns.
  - ❖ Removing Multicollinearity from the dataset.
  - ❖ Choosing Model explainability technique.

- XGBoost model seems to be the best choice for this dataset as it has the highest R2 score, the lowest MSE, and the lowest MAE. Linear Regression Model is the worst model.

- Hour and Temperature are the two most important feature to predict rental bike demand followed by Humidity, Solar Radiation, Rainfall etc.

Thank You…