Capstone Project
Bike Sharing Demand Prediction

By Ayush Kumar

Points for Discussion

- Problem Statement
- Data Description
- Data Preparation and Cleaning
- Exploratory Data Analysis
- Hypothesis Testing
- Feature Engineering
- Modelling
- Model Interpretation
- Conclusion

Problem Statement

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. The business problem is to ensure a stable supply of rental bikes in urban cities by predicting the demand for bikes at each hour. By providing a stable supply of rental bikes, the system can enhance mobility comfort for the public and reduce waiting time, leading to greater customer satisfaction.

Data Description

The **Seoul 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.

This dataset contains more than 8700 rows and 14 columns of data.

Data Description

- **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

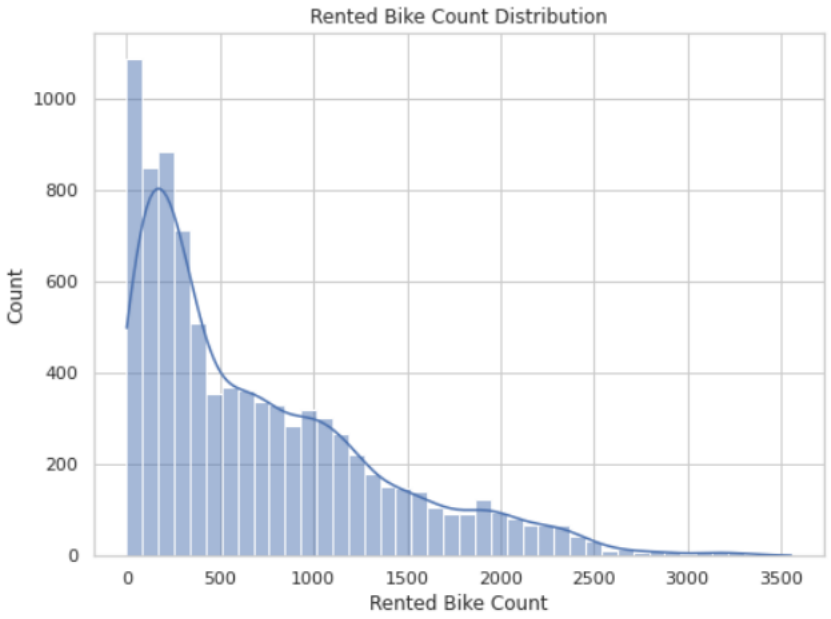
1. **Year**, **Month**, **Day**, **Weekday** from Date
2. **Temperature Bin** from Temperature(°C)

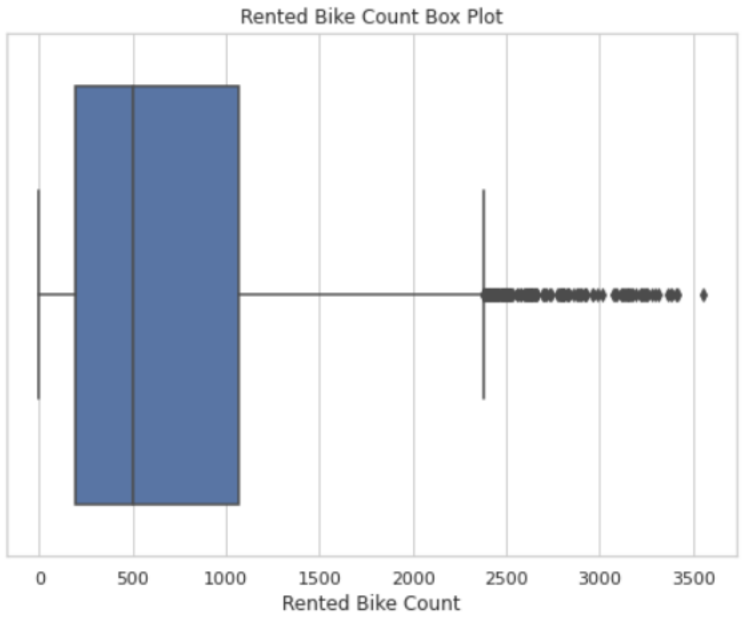
```
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
df['weekday'] = df['Date'].dt.day_name()
```

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

Exploratory Data Analysis

Rented Bike Count Distribution



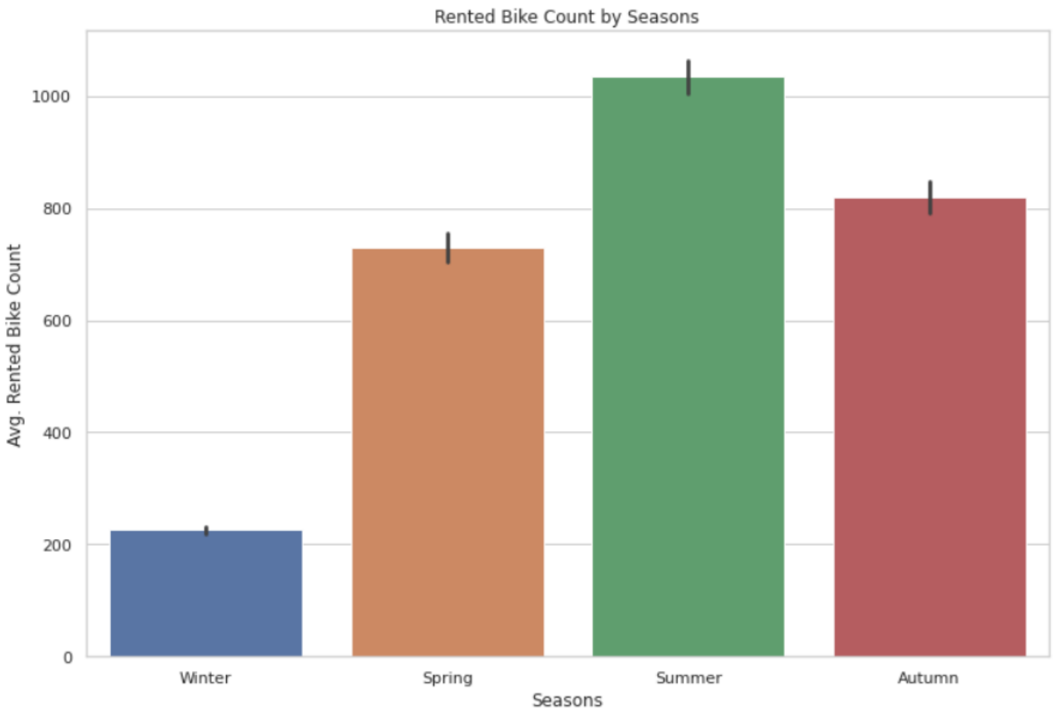


Exploratory Data Analysis

Rented Bike Count by Seasons

Rental Bike demand in winter season is significantly lower than other months.

Demand is highest in Summer

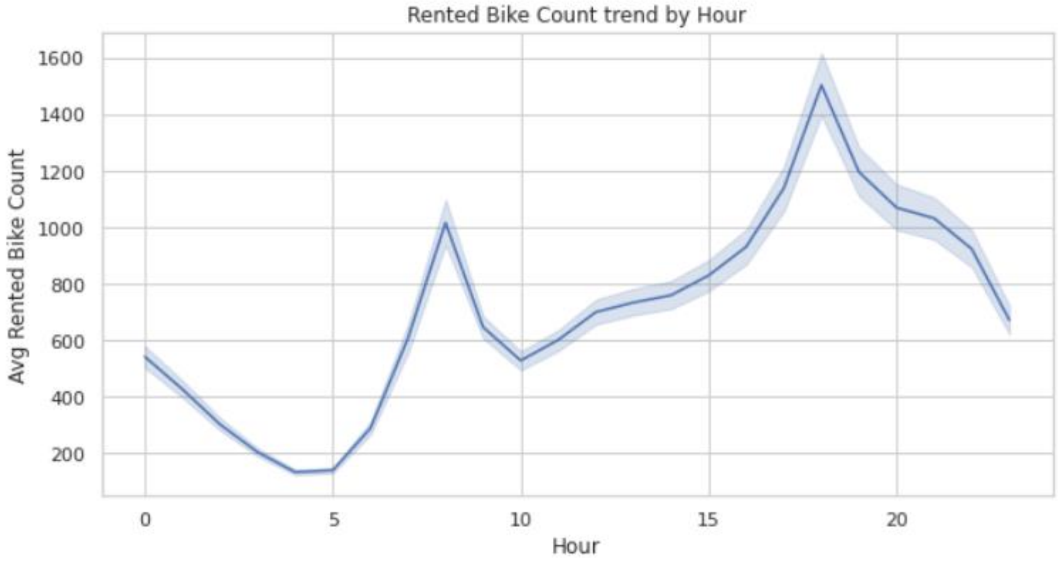


Exploratory Data Analysis

Rented Bike Count by Hour

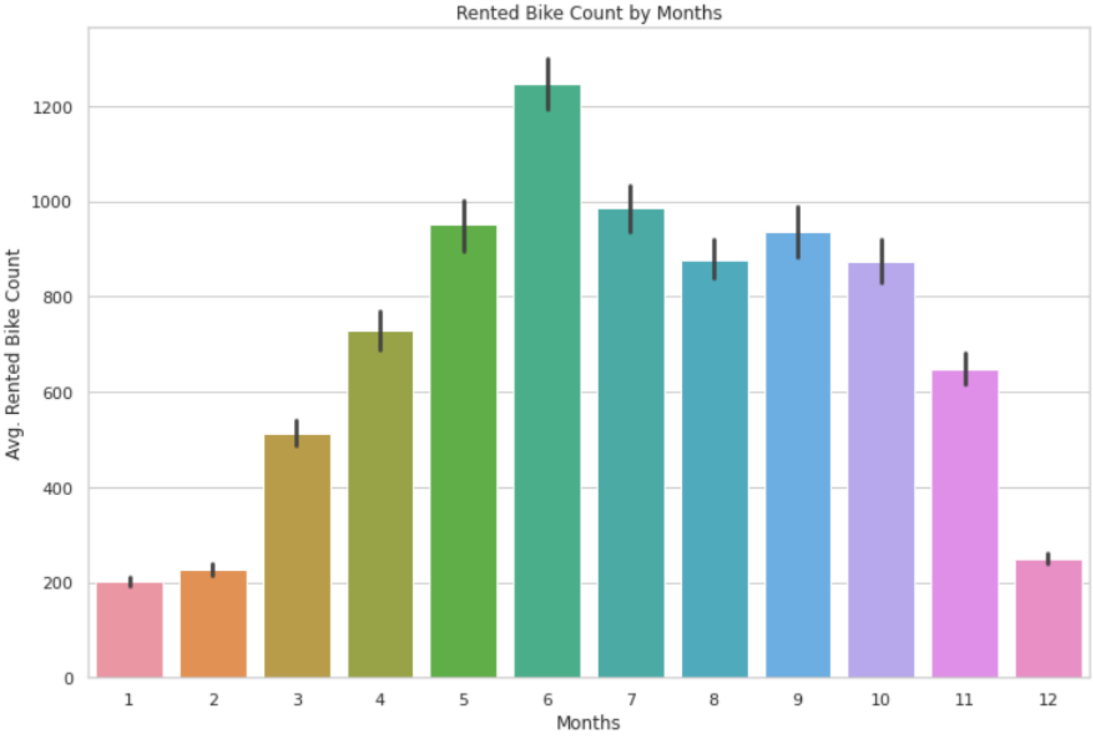
We can see demand peaks during rush hours of the day.

Rush hour is generally around 8AM in the morning and 6PM in the evening.



Exploratory Data Analysis

Rented Bike Count by Months



Similar to what we saw with seasons, demand decreases significantly during winter months like Dec, Jan, Feb etc.

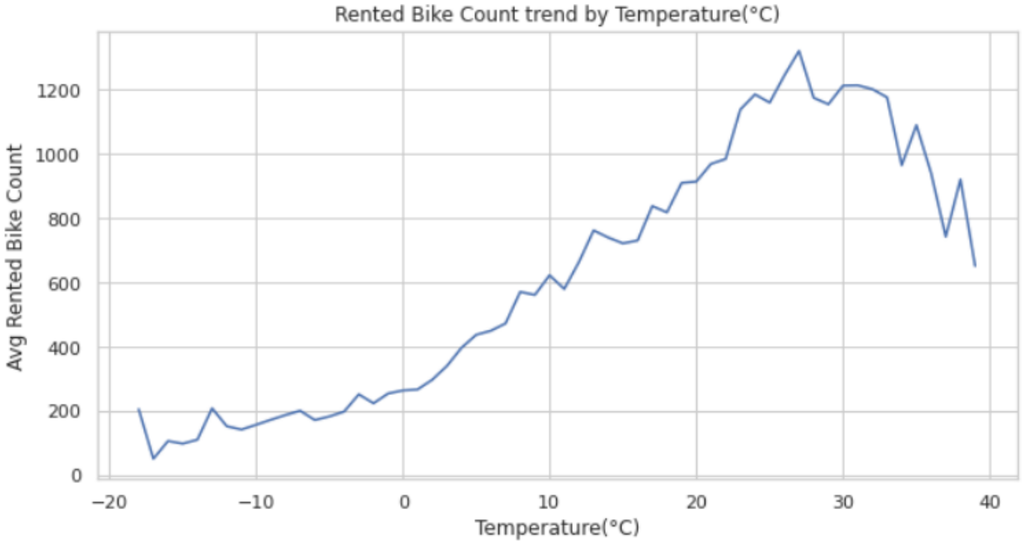
Demand peaks at summer months like May, June July etc.

Exploratory Data Analysis

Rented Bike Count by Temperature

The Bike rental demand increases as the temperature increases.

Although too high temperature leads to decrease in demand again.

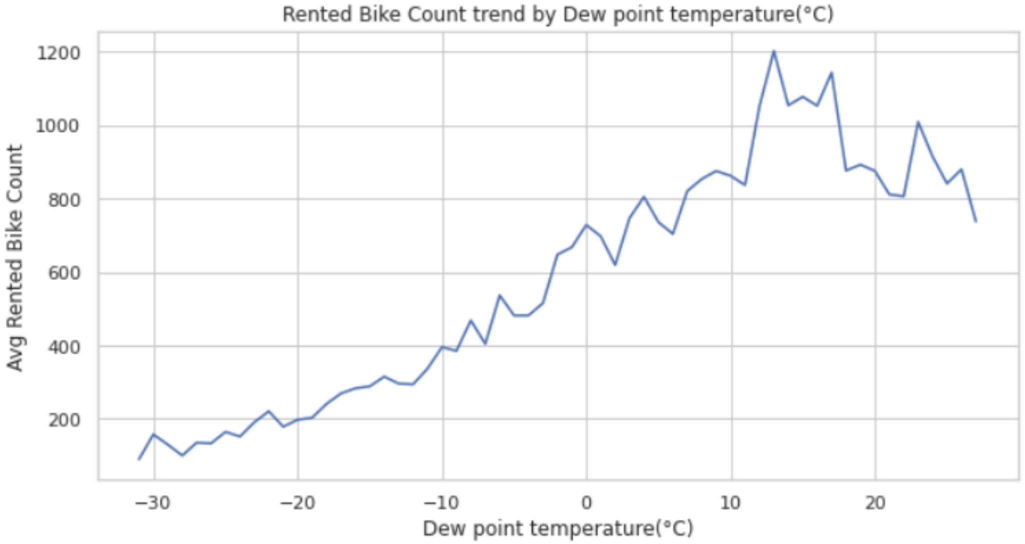


Exploratory Data Analysis

Rented Bike Count by Dew Point Temperature

Similar trend for dew point temperature as well i.e., The Bike rental demand increases as the temperature increases.

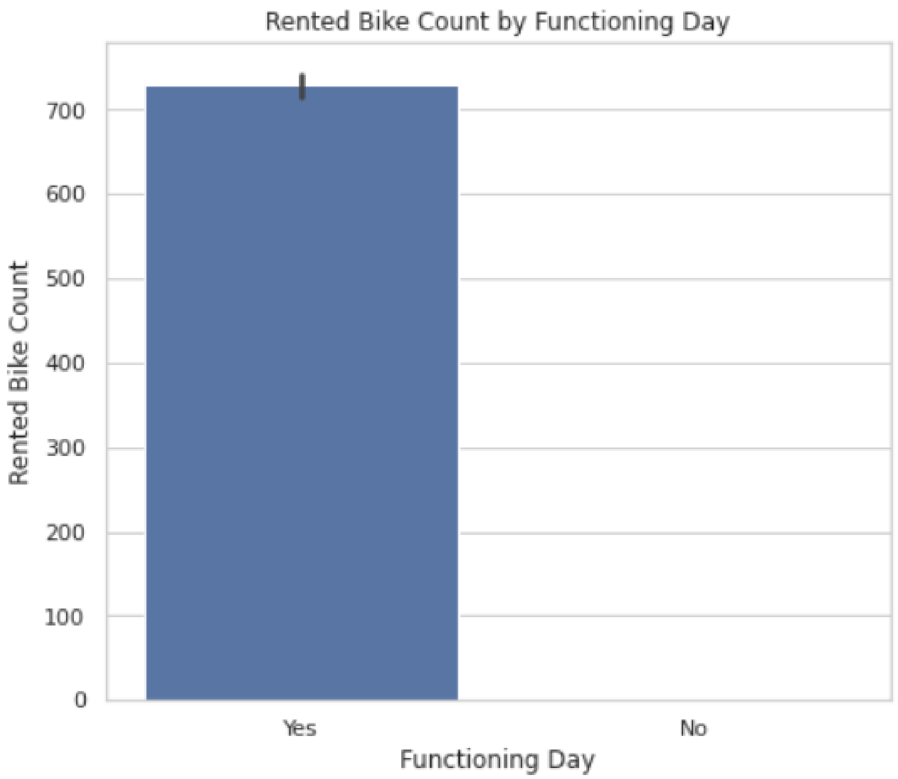
Although too high dew point temperature leads to decrease in demand again.



Exploratory Data Analysis

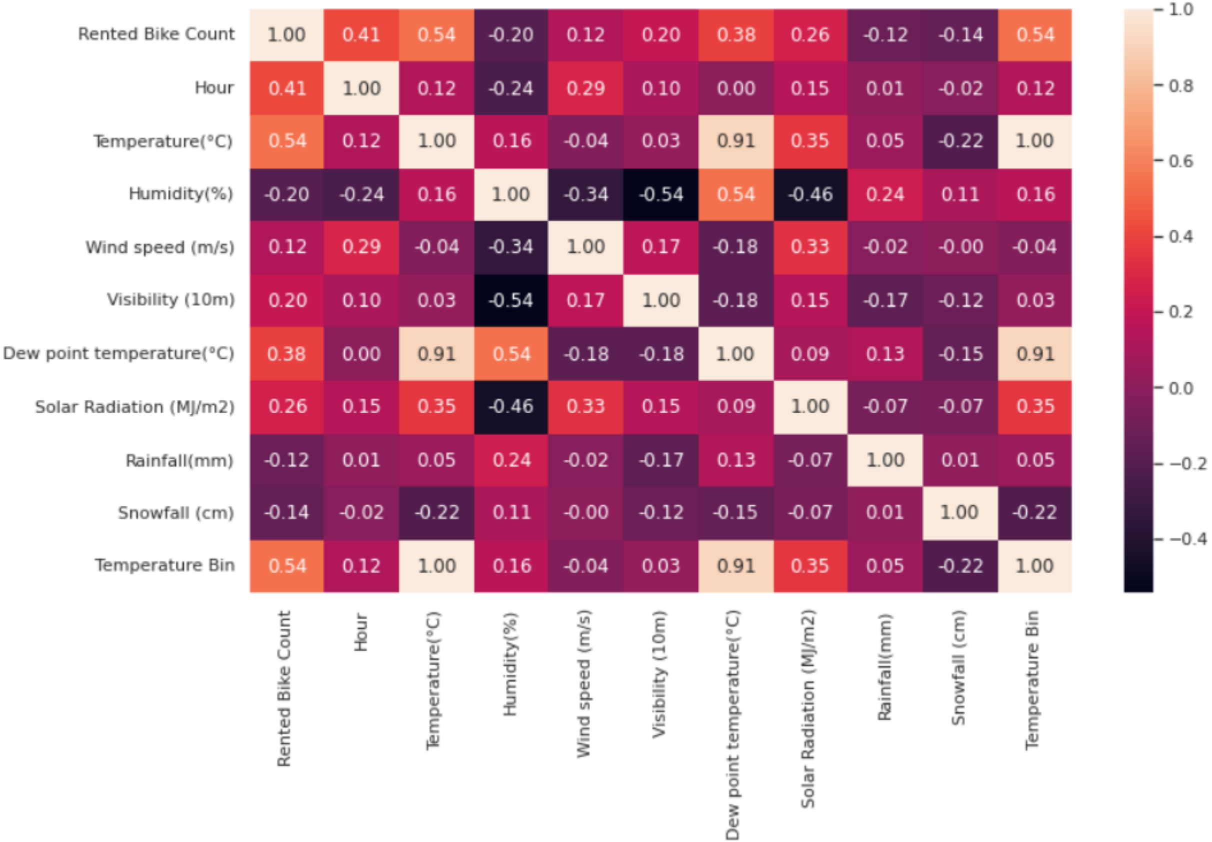
Rented Bike Count by Functioning Day

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



Exploratory Data Analysis

Correlation of features



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

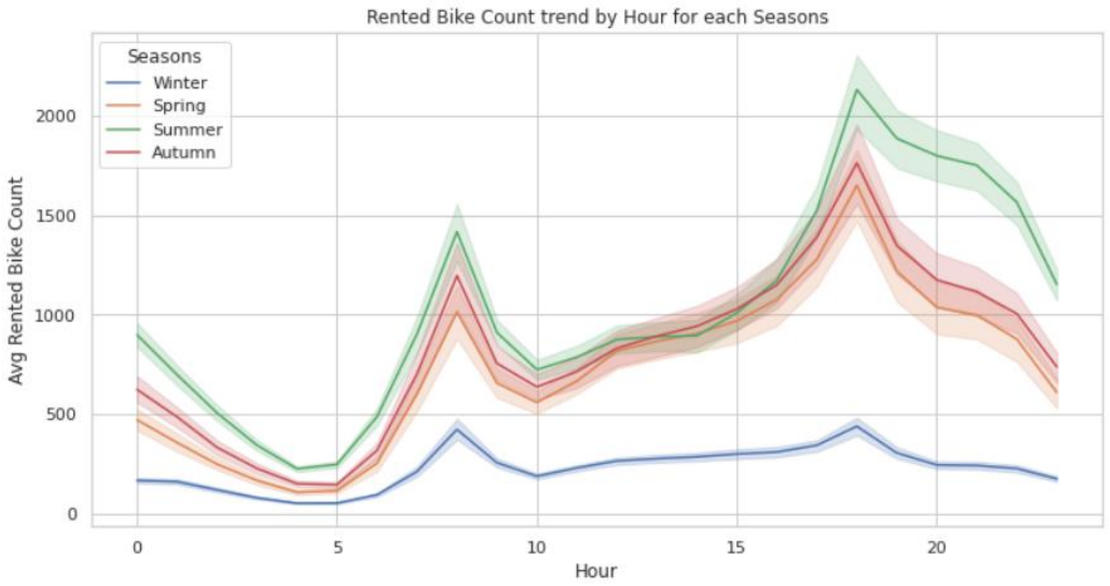
Hence will be dropping Dew Point Temperature later before modelling

Exploratory Data Analysis

Rented Bike Count by Hour for each Season

We can see demand peaks during rush hours of the day.

Each season has similar hourly pattern only levels are different.

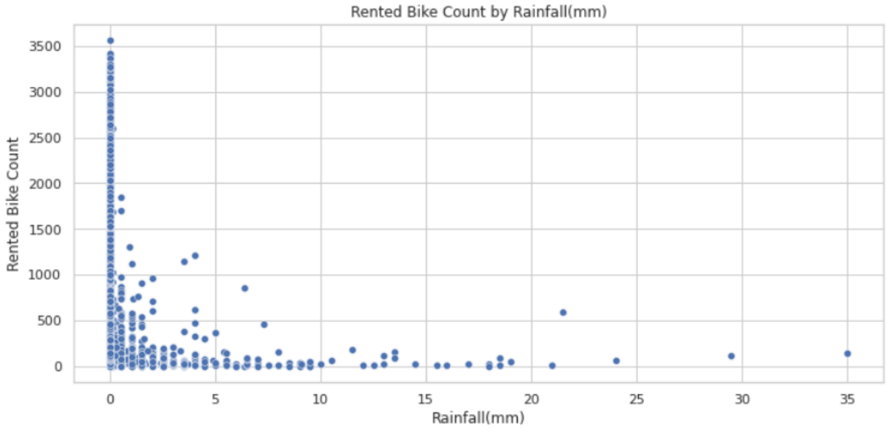


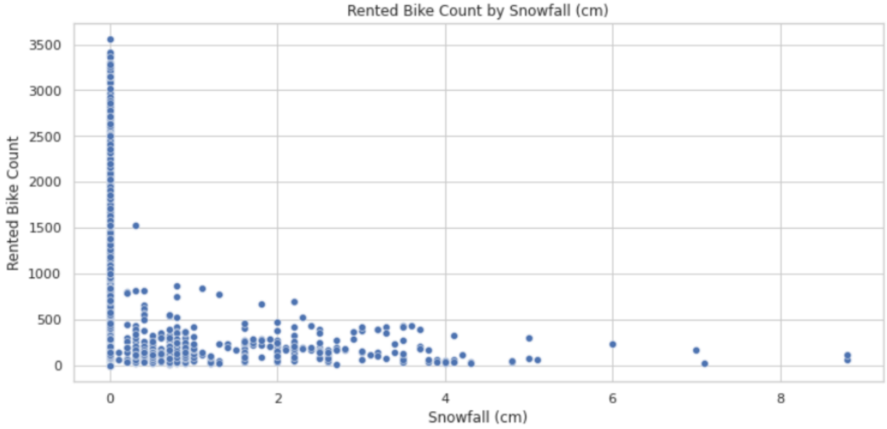
Exploratory Data Analysis

Rented Bike Count by Rainfall and Snowfall

Rainfall and Snowfall both leads to decrease in the demand in bike rentals.

Which is obvious because people do not want to go out on a bike when it is raining or snowing unless it is emergency.



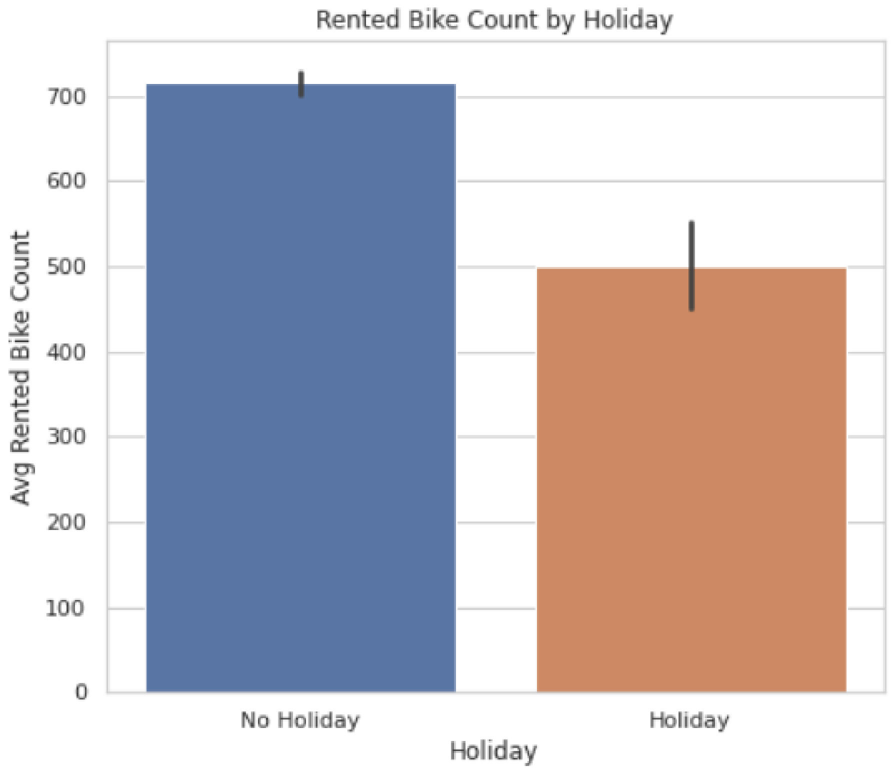


Exploratory Data Analysis

Rented Bike Count by Seasons

Rental Bike demand is higher on non holiday compared to holiday.

Possible reason for this can be that a lot of people uses rental bike to go to offices or schools/colleges on non holiday.

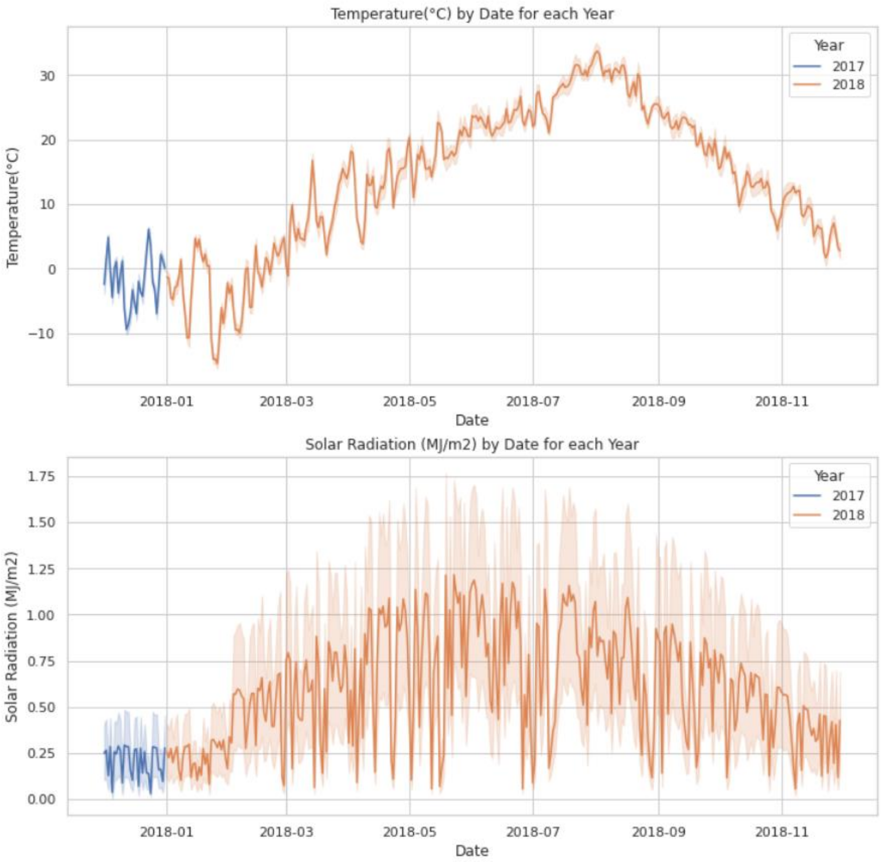


Exploratory Data Analysis

Temperature and Solar Radiation over time

As expected temperature rises during summer months like May, June, July etc. and decreases during months like December, January etc.

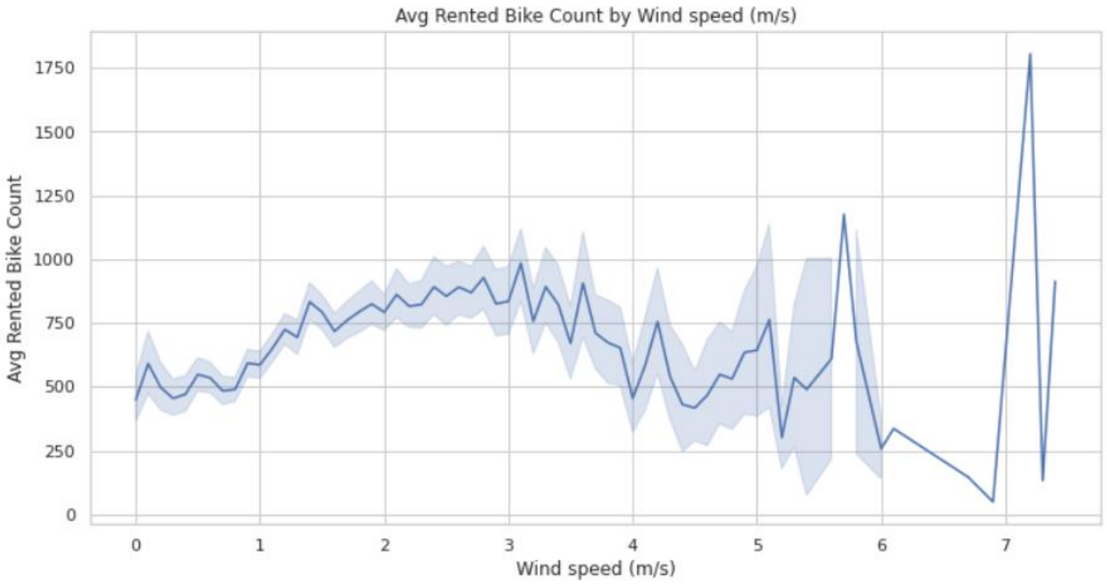
Similar trend for solar radiation as well, but one thing to observe that there are huge fluctuations in the value, it may be because of day-night cycle, as there is no sunlight at night time.



Exploratory Data Analysis

Rented Bike Count by Wind Speed (m/s)

There is a slight increase in demand as wind speed increases but too much wind speed leads to slight decreases in demand.

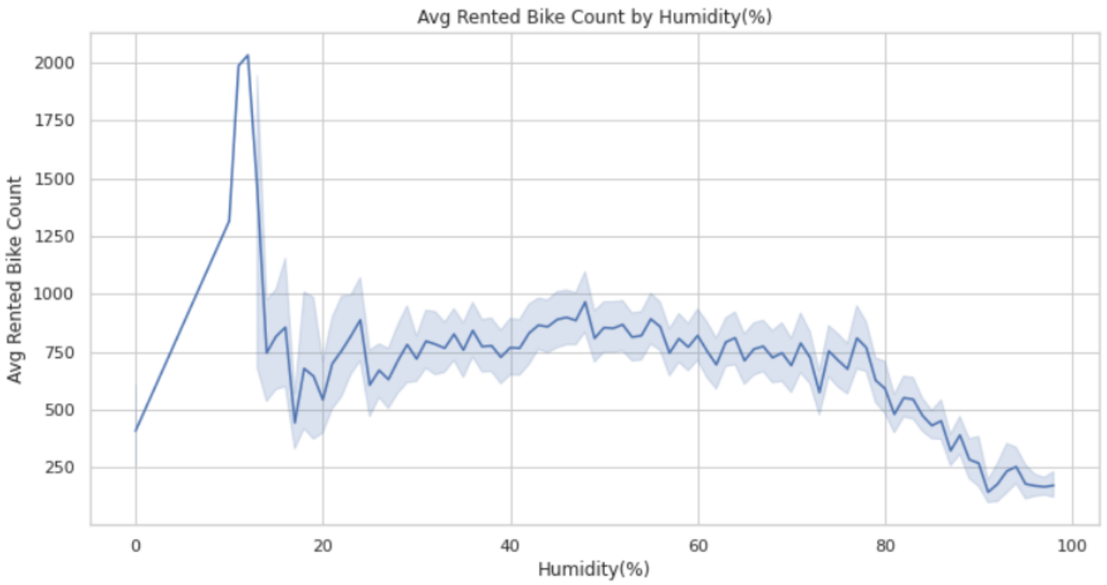


Exploratory Data Analysis

Rented Bike Count by Humidity

The demand is consistent for humidity till 75% but after that it starts decreasing.

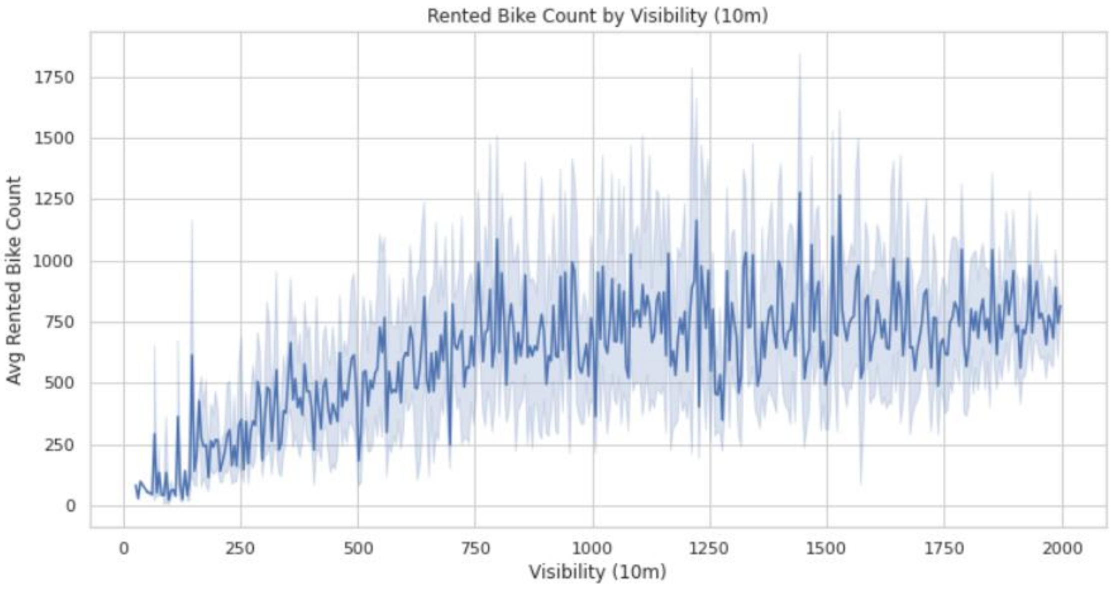
One reason for such high humidity can be rain and we already saw rain causes decrease in demand.



Exploratory Data Analysis

Rented Bike Count by Visibility

As visibility increases the demand increases till around 7500m after that it remains consistent.



Hypothesis Testing

- **Rented Bike Demand in hot weather is higher compared to demand in cold weather.**
- Assumed threshold as 20°C for hot and cold.
- The **two sample t-test** is used to determine if there is a significant difference between the means of two groups.
- Also we know from previous charts that Rented Bike Count is right skewed with large sample sizes (i.e., $n_{hot} = 2928$ & $n_{cold} = 5832$) and we don't know σ_p

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

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

Test Type: Two-sample t-test

```
Since p-value (0.0) is less than 0.05, we reject null hypothesis.
Hence, There is a significant difference in mean bike rentals between the 'hot' and 'cold' temperature groups.
```

Hypothesis Testing

- **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

- The **two sample t-test** is used to determine if there is a significant difference between the means of two groups.
- Also we know from previous charts that Rented Bike Count is right skewed with large sample sizes (i.e., $n_{rush} = 2190$ & $n_{non\text{-}rush} = 6570$) and we don't know σ_p

```
Since p-value (9.381784283723713e-104) is less than 0.05, we reject null hypothesis.
Hence, There is a significant difference in mean bike rentals between the 'rush hour' and 'non-rush hour' times of day.
```

Hypothesis Testing

- **Rented Bike Demand is different in different seasons with highest in summer and lowest in winter.**
- The **one-way ANOVA** test is used to determine if there is a significant difference between the means of more than two groups.
- Also We know from previous charts that Rented Bike Count is right skewed with large sample sizes (i.e., $n_{autumn} = 2184$, $n_{spring} = 2208$, $n_{summer} = 2208$, $n_{winter} = 2160$).

```
F-statistic: 776.4678149879506
p-value: 9.381784283723713e-104

  Multiple Comparison of Means - Tukey HSD, FWER=0.05
============================================================
group1 group2  meandiff p-adj   lower      upper    reject
------------------------------------------------------------
Autumn Spring   -89.5667   0.0 -134.0266  -45.1069   True
Autumn Summer   214.4754   0.0  170.0156  258.9352   True
Autumn Winter -594.0568   0.0 -638.7616  -549.352   True
Spring Summer   304.0421   0.0  259.7039  348.3803   True
Spring Winter    -504.49   0.0 -549.0739 -459.9062   True
Summer Winter -808.5322   0.0  -853.116 -763.9483   True
------------------------------------------------------------
```

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.

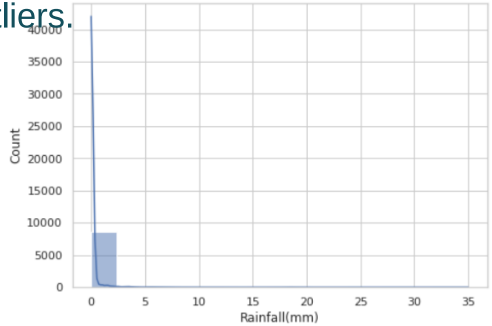
Test Type: One-way ANOVA test

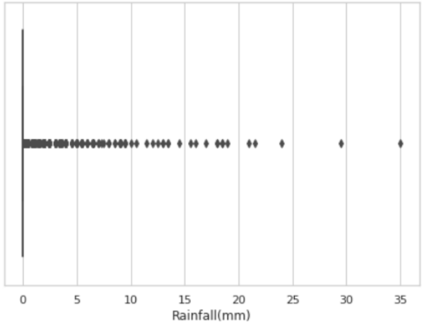
Feature Engineering

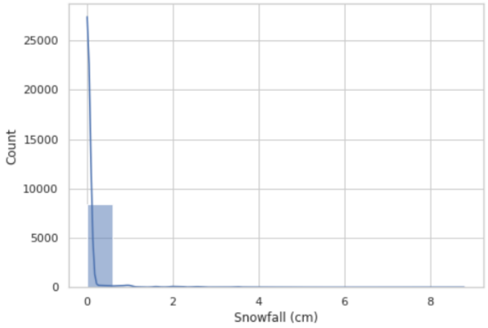
From distribution plots of different features, I got to know that **Wind speed (m/s)**, **Solar Radiation (MJ/m2)**, **Rainfall(mm)**, **Snowfall (cm)** columns have outliers.

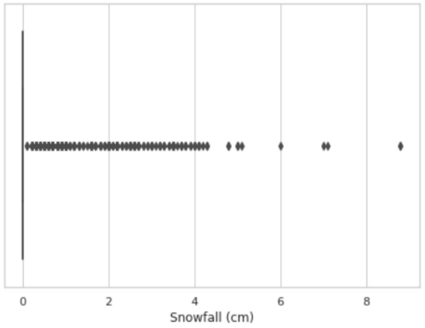
In **Rainfall(mm)** and **Snowfall(cm)** column , we see that that most of the values are zero and few are non zero which is understandable as we don't see rain and snow everyday. Given the nature of data, it is unlikely that the non-zero values represent outliers. However value that is significantly higher can be treated as outlier.

Hence used **99th quantile** for capping outliers.









Feature Engineering

- The **Wind speed(m/s)** and **Solar Radiation(MJ/m2)** column , the values are right skewed. Hence used **IQR method** for capping outliers.

- Converted **Seasons**, **Holiday**, **Month**, **weekday** columns to one-hot encoding as they represent categorical values.

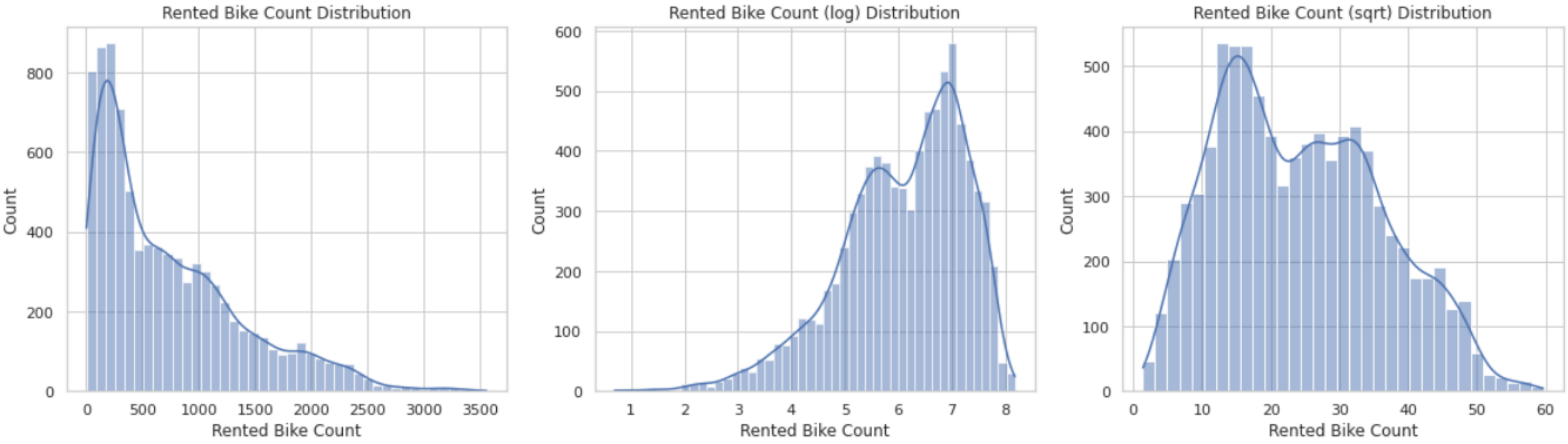
- All the values in target variable (**Rented Bike Count**) were zero for non functioning day hence removed those rows as we need to predict demand on functioning day only.

- Used VIF for checking multicollinearity, also we already saw before Dew Point Temperature is highly correlated to Temperature hence dropped Dew Point Temperature column.

Feature Engineering

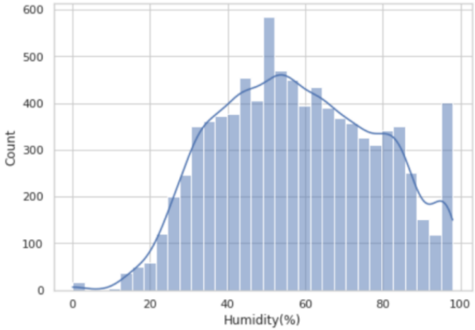
The **Rented Bike Count** was right skewed, and to train a robust model we can transform it to normal.

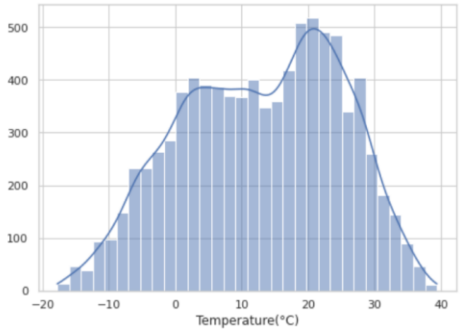
Applied **square root** to transform it to normal.

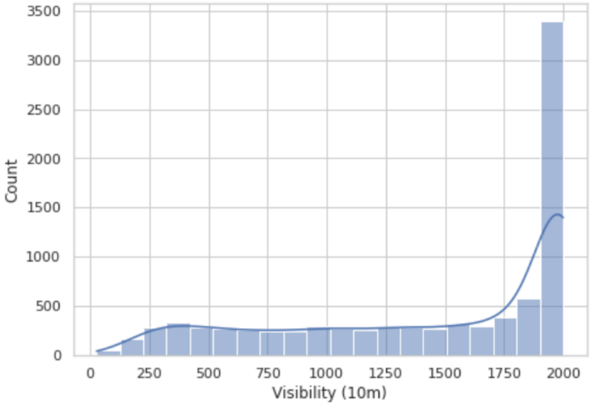


Feature Engineering

- Similarly applied square root to **Wind Speed (m/s)** to transform it to normal as it was originally skewed.
- All the columns was in similar scale except Temperature, Humidity and Visibility. Hence applied StandardScaler and MinMaxScaler to scale them.
- Splitted Data into train and test sets with ratio 75:25.







Modelling

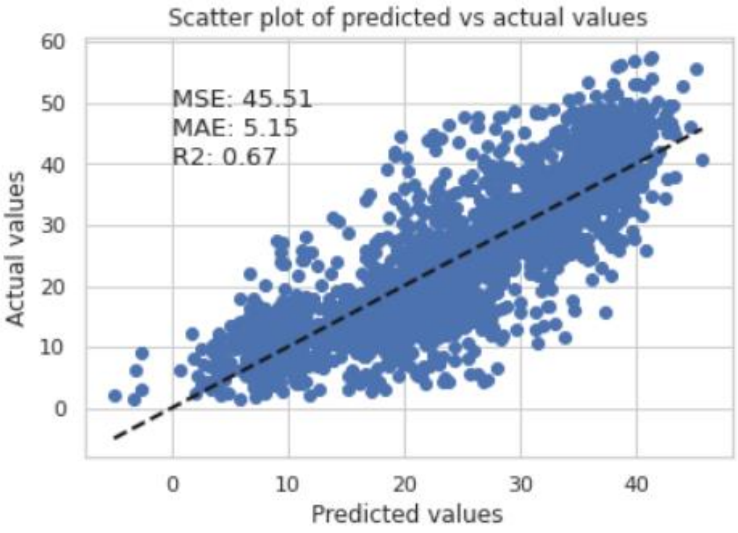
Since we’re trying to predict continuous variable, I trained various regression algorithms along with Hyper parameter tuning and cross validation to get he best model.

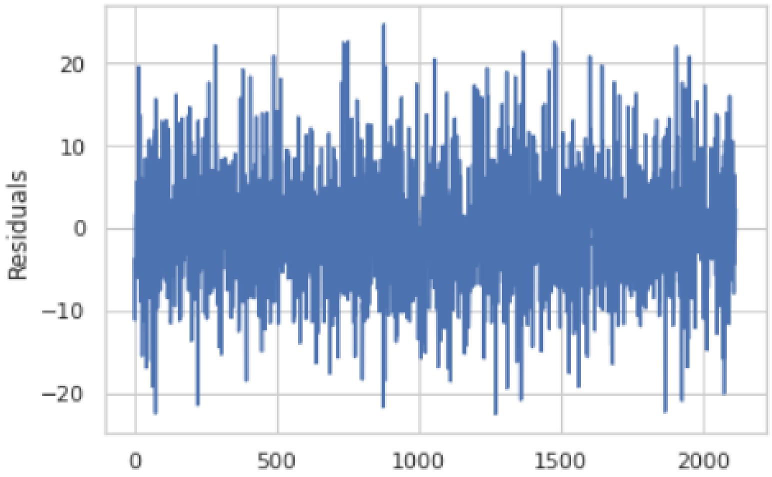
Algorithms used:

- Linear Regression
- Ridge Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor

Modelling

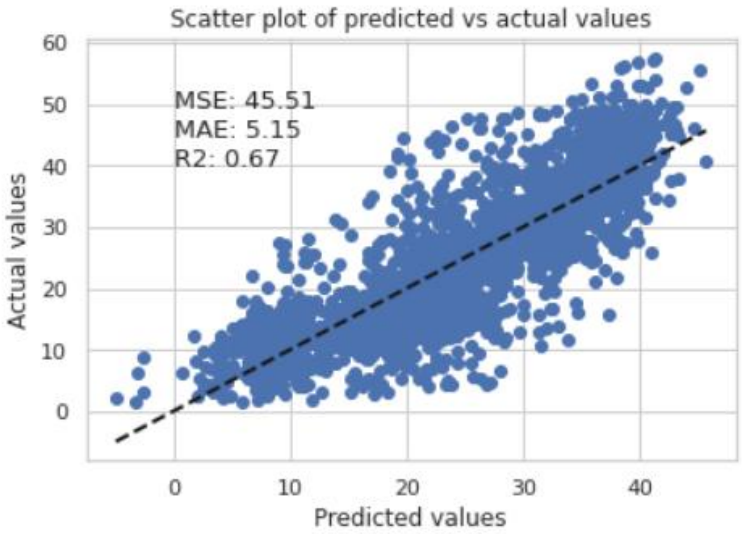
Linear Regression

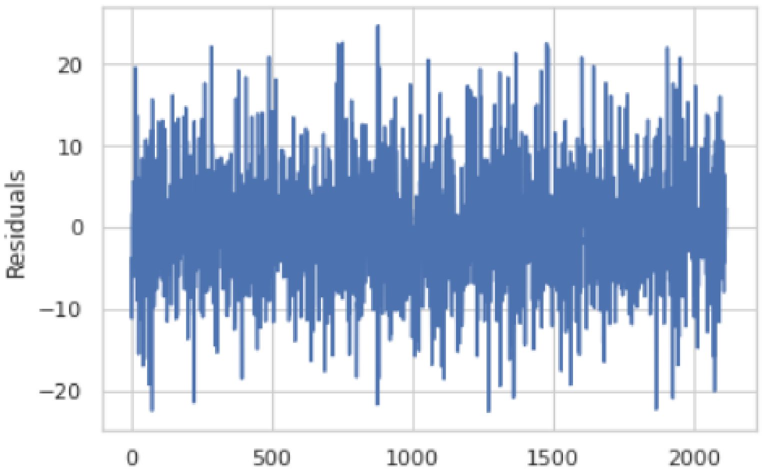




Modelling

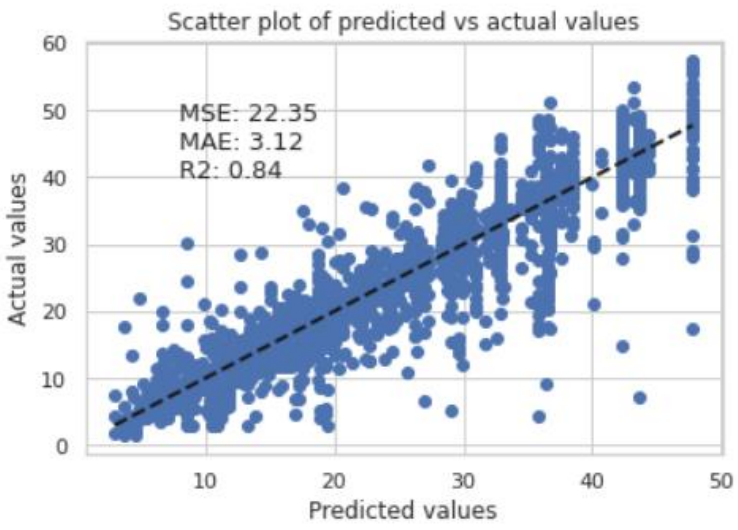
Ridge Regression



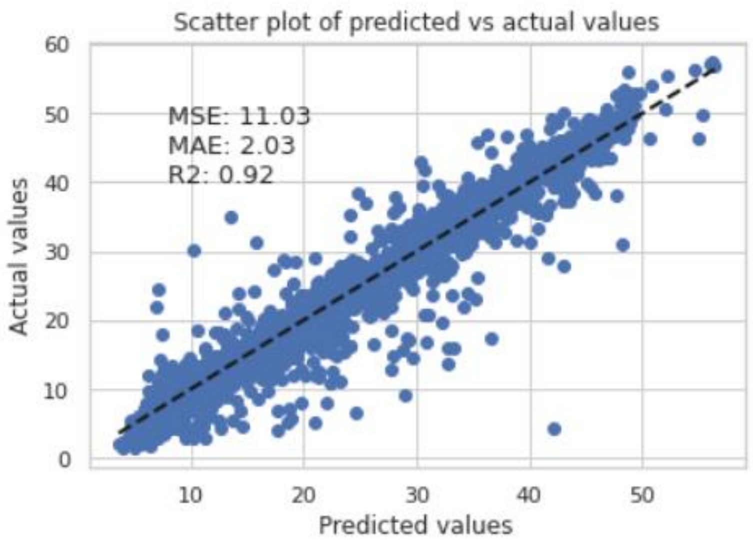


Modelling

Decision Tree Regressor

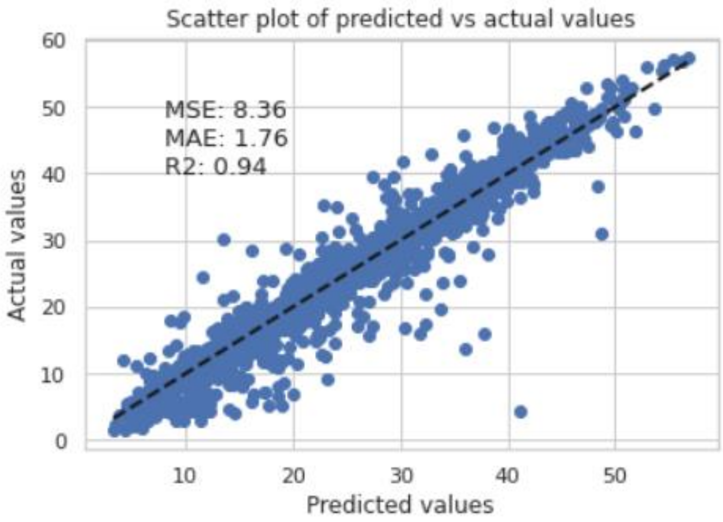


Random Forest Regressor



Modelling

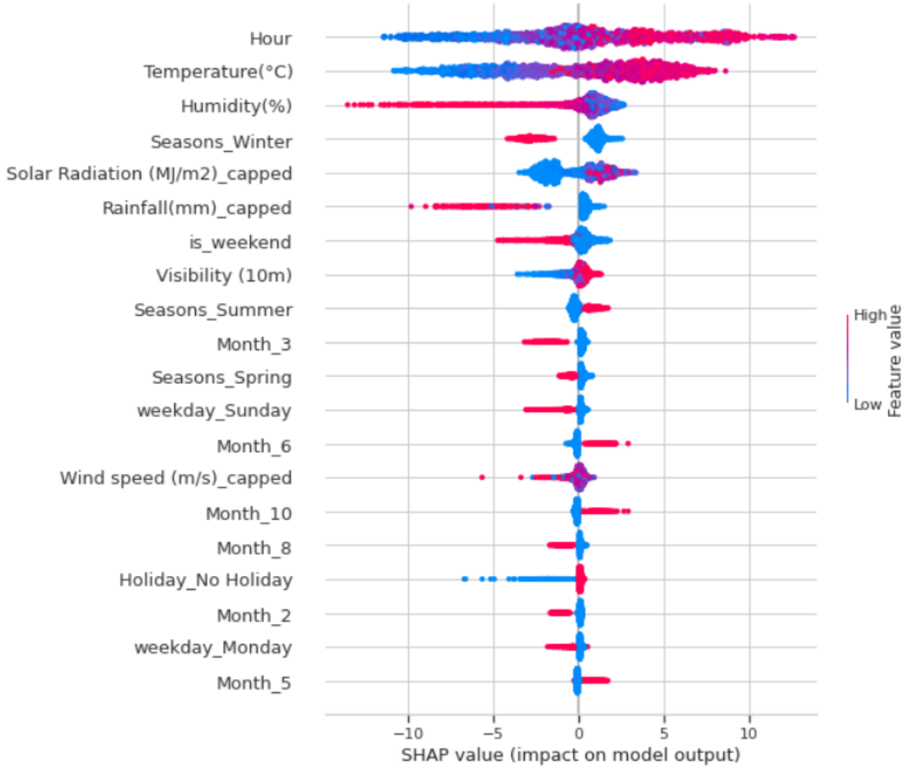
XGBoost Regressor

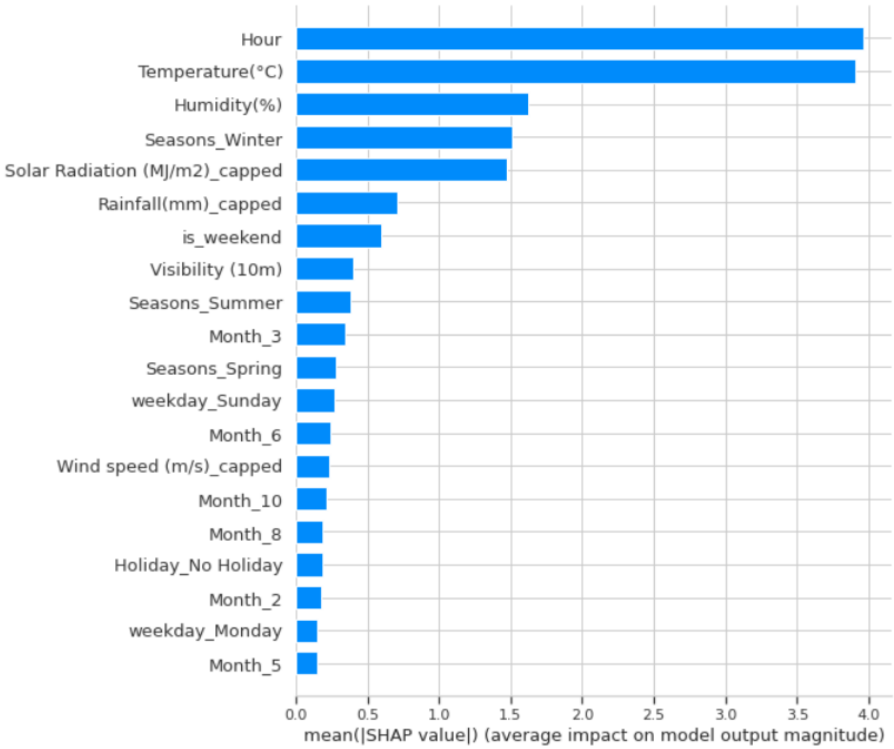


Performance Comparison

	R2	MSE	MAE
XGBoost CV	0.940	8.356	1.755
XGBoost	0.932	9.397	1.890
Random Forest	0.920	11.027	2.031
Random Forest CV	0.919	11.295	2.077
Decision Tree CV	0.839	22.350	3.123
Decision Tree	0.776	31.026	4.032
RidgeCV	0.672	45.500	5.148
Linear Regression	0.672	45.506	5.148
Linear Regression CV	0.672	45.506	5.148
Ridge	0.672	45.509	5.149

Model Interpretation





Conclusions

The XGBoost (Extreme Gradient Boosting) which gave the best result for predicting Rented Bike Count using several features on both on train and test data with R2 score of 0.94.

Thank You!