Capstone Project - 2

Bike Sharing Demand Prediction

Team

Radhika R Menon

Rohit Raj

Ghanal Kaushik

Jayaprakash Kunduru

Contents

Introduction

Bike sharing system is a shared transportation service that provides individuals with bikes for their common use on a short-term basis for a price or for free. Over the last few decades, there has been a significant increase in the popularity of bike-sharing systems all over the world.

Advantages

- Environmentally sustainable
- Reduces traffic congestion
- Physical health benefits to the users
- Economical
- Fast and easy accessibility

Problem statement

Currently, Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bikes 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 Overview

Rented Bike count - Count of bikes rented each hour

Hour - Hour of the day

Temperature - Temperature recorded in the city in *Celsius* (°C).

Humidity - Relative humidity in %

Wind-speed - Speed of the wind in *m/s*

Visibility - measure of distance at which object or light can be clearly discerned in units of 10*m*

Dew point temperature - Temperature recorded in the beginning of the day in *Celsius*(°C).

Data Overview

Date - The date of each observation in the format '***year-month-day***'

Solar radiation - Intensity of sunlight in ***MJ/m^2***

Rainfall - Amount of rainfall received in ***mm***

Snowfall - Amount of snowfall received in ***cm***

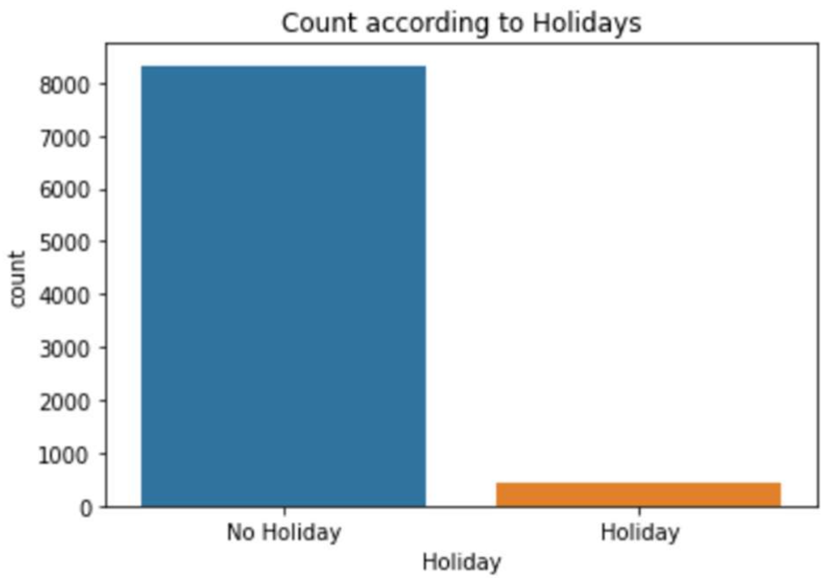
Seasons - Season of the year (***Winter, Spring, Summer, Autumn***)

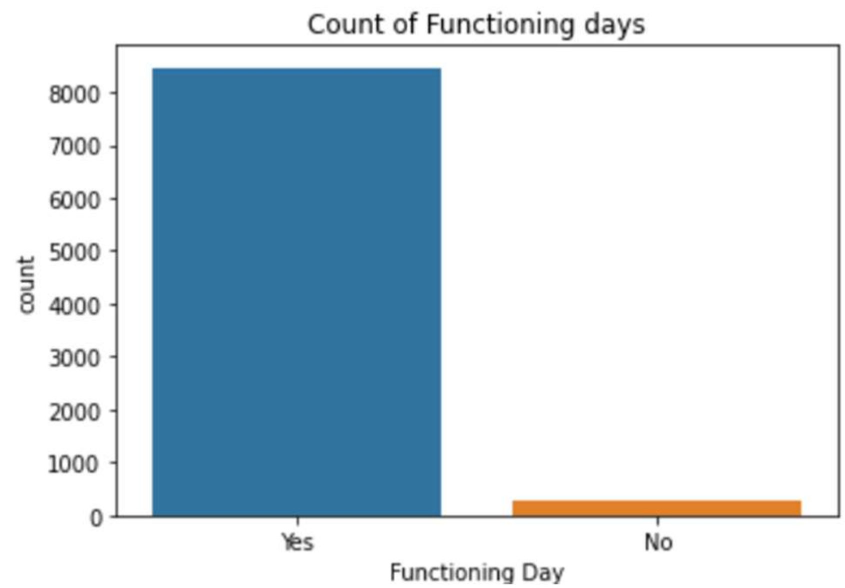
Holiday - Whether the day is a Holiday or not (***Holiday/No holiday***)

Functional Day -Whether the rental service is available (***Yes***-Functional hours) or not (***No***-Non functional hours)

EDA - Univariate Analysis

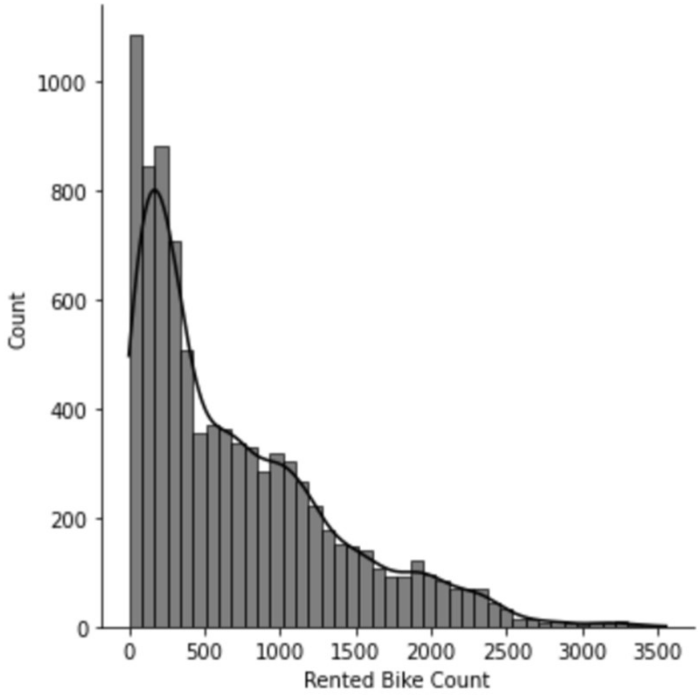
Description of Functioning days & Holidays from data



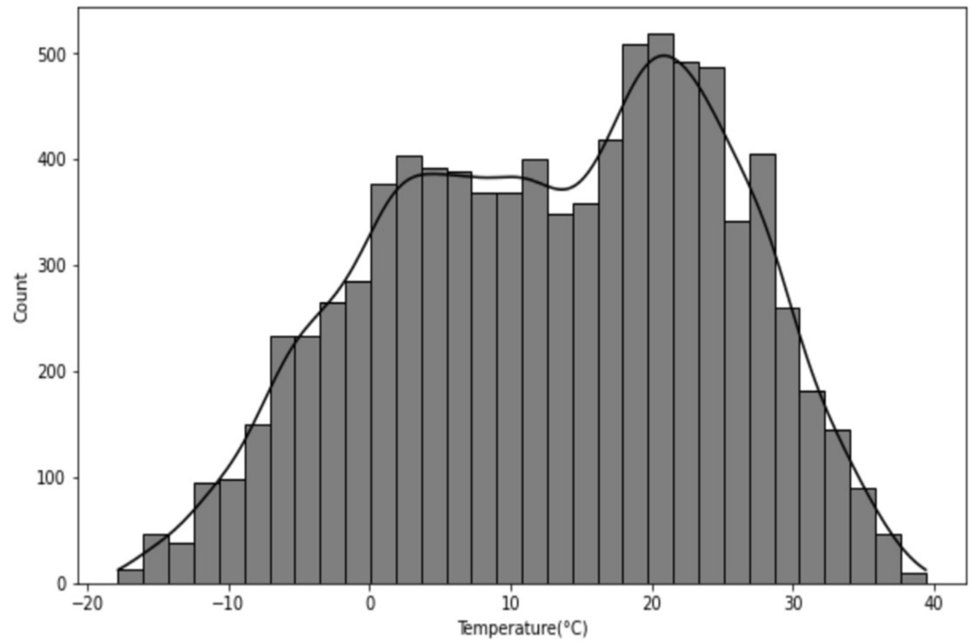


EDA - Univariate Analysis

Bike count distribution

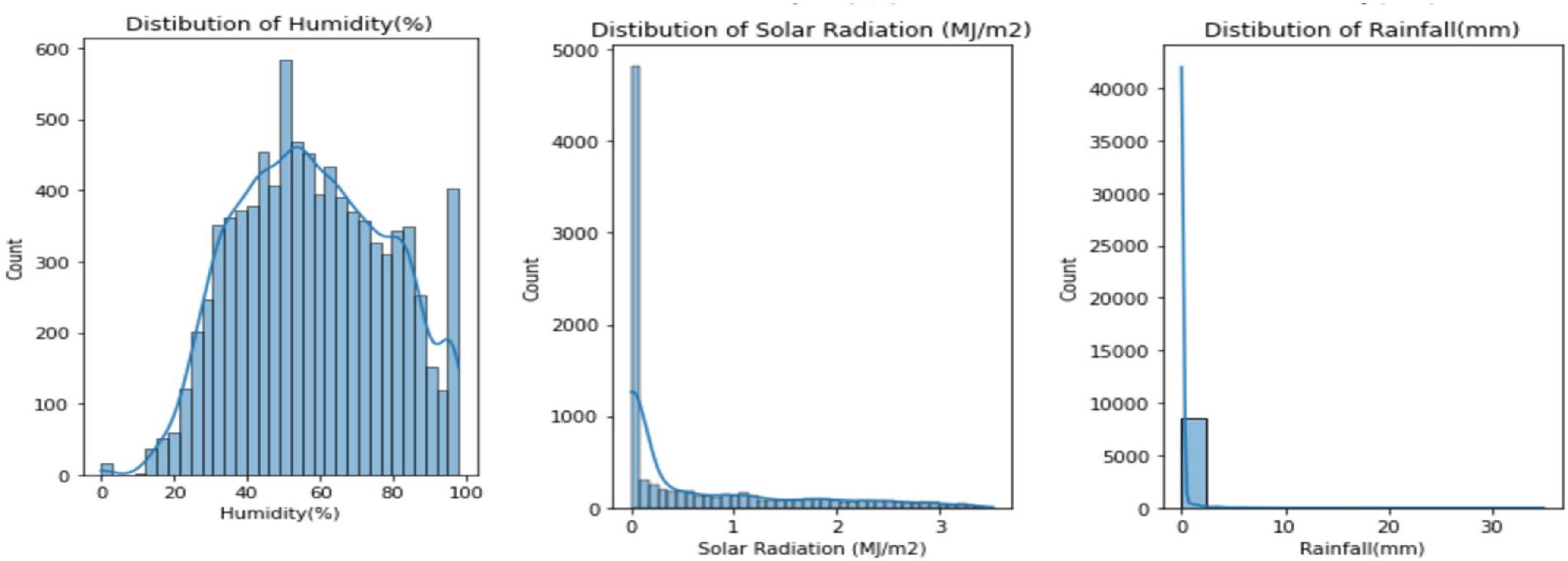


Temperature distribution



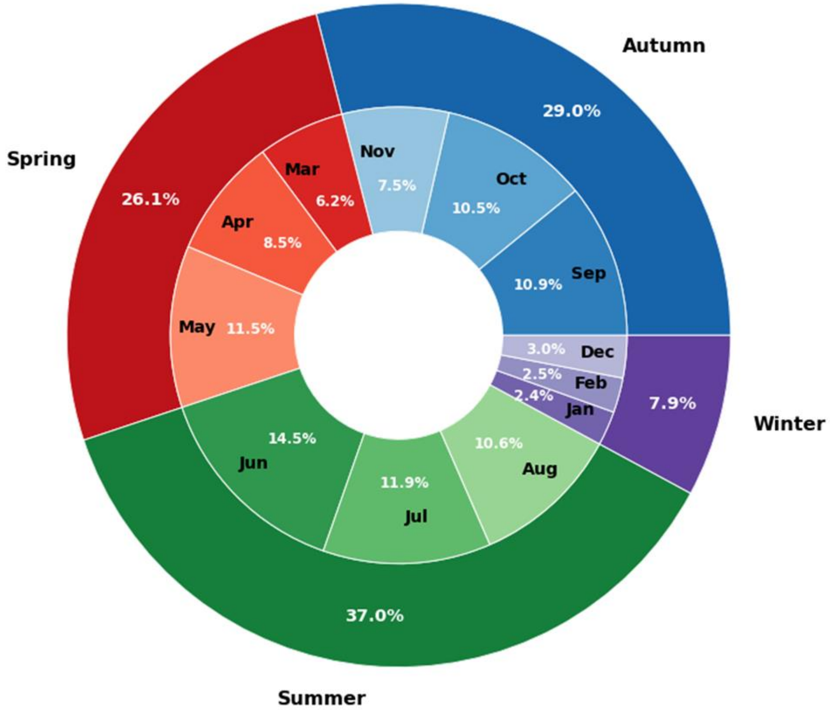
EDA - Univariate Analysis

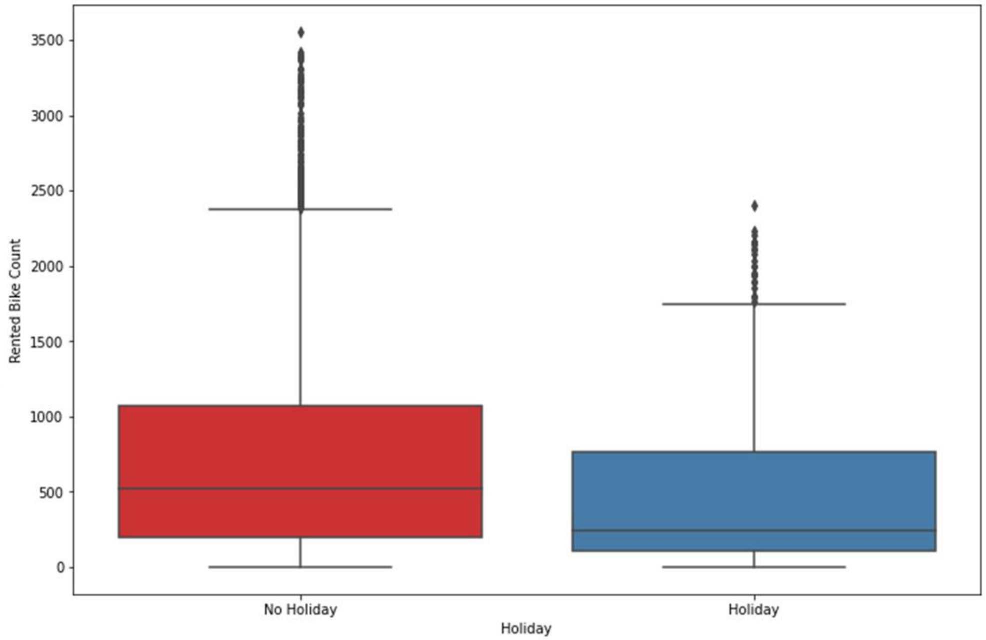
Description of Humidity, Solar Radiation & Rainfall



EDA - Bivariate Analysis

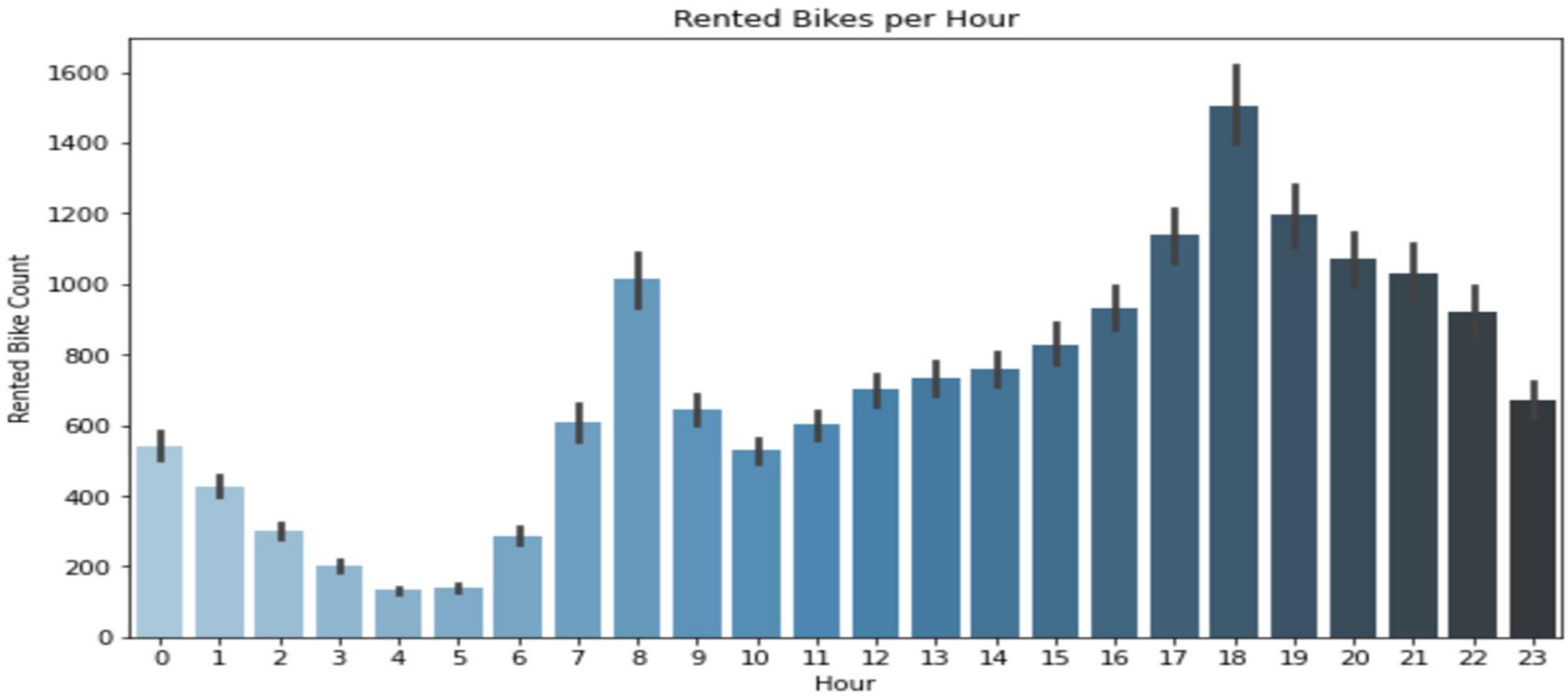
Bike rentals according to Seasons & Holidays





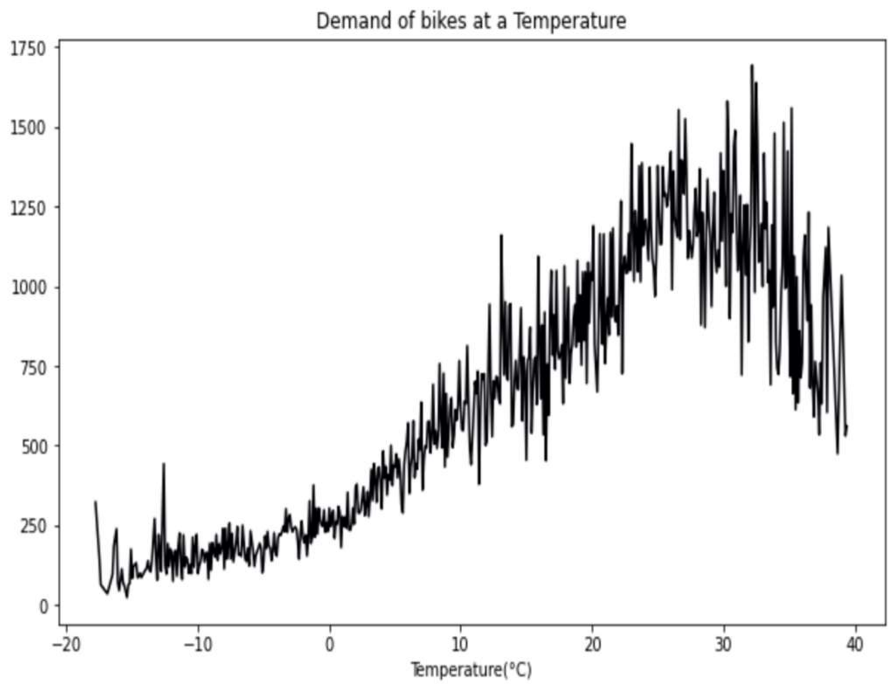
EDA - Bivariate Analysis

Bike rentals at different hours of the day



EDA - Bivariate Analysis

Bike rentals according to various temperatures

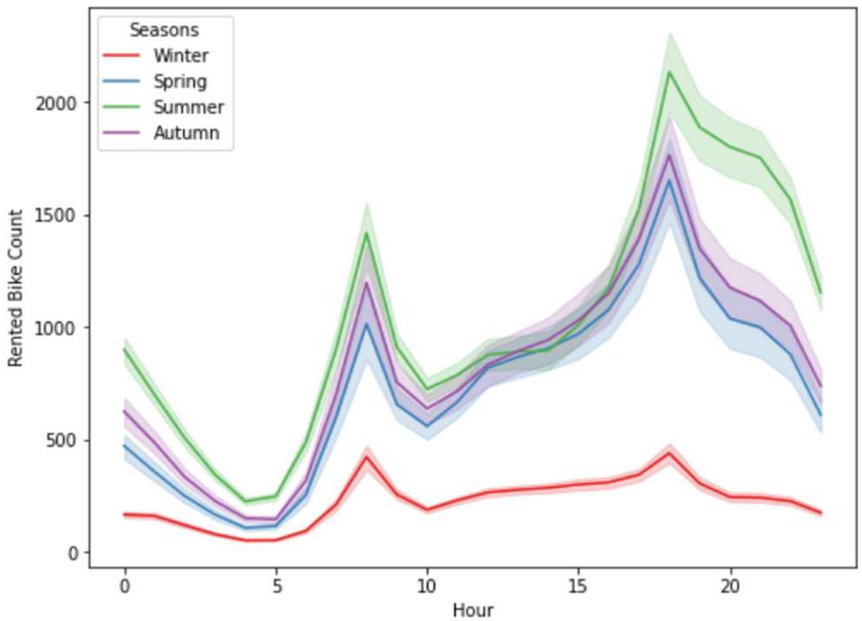


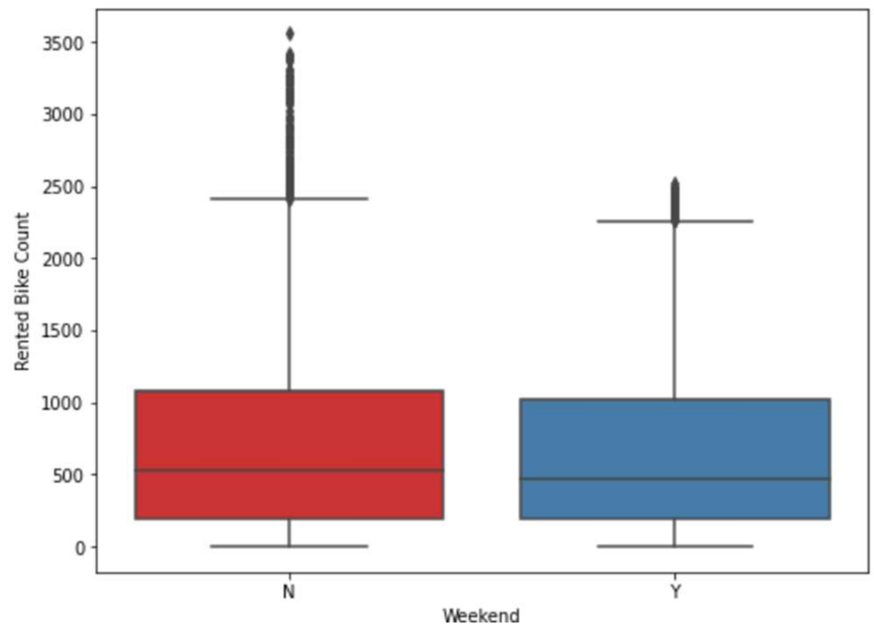
Rented Bike count Vs Functioning Day

Functioning Day	Holiday	bikerentalcounts
No	Holiday	0
	No Holiday	0
Yes	Holiday	215895
	No Holiday	5956419

EDA - Bivariate Analysis

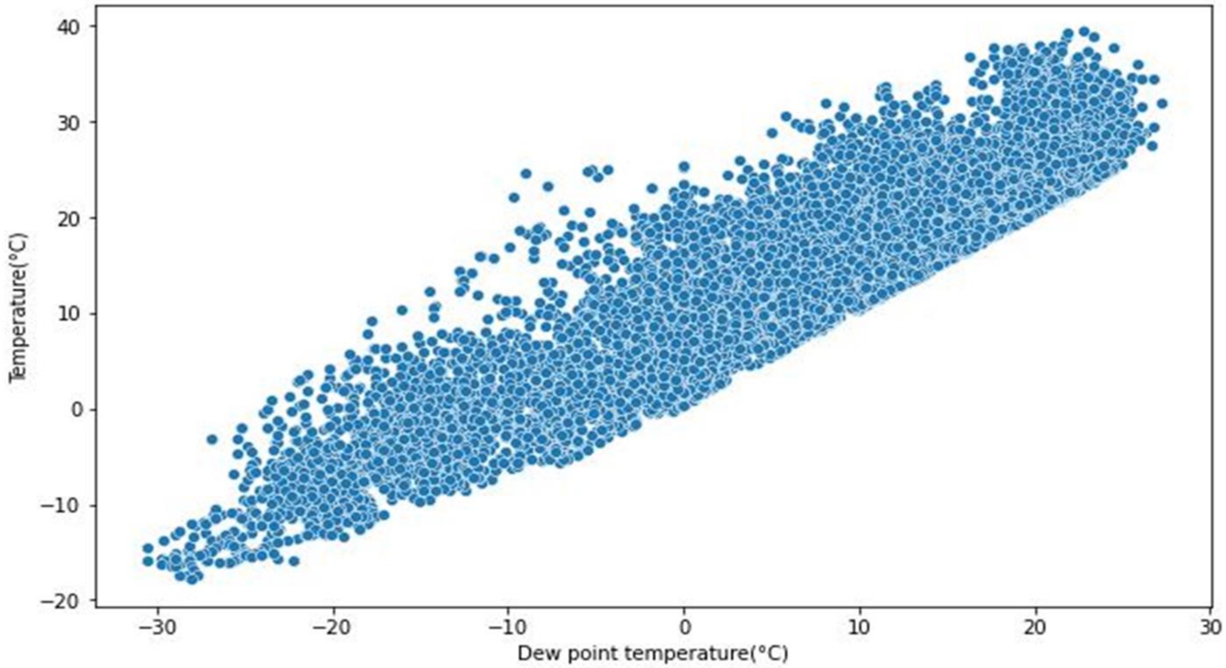
Bike rentals according to Seasons & Weekends





EDA - Bivariate Analysis

Dew Point Temperature and Temperature are highly positively correlated



EDA Conclusions

- Temperature and Hour have a strong correlation with the count of rented bikes.
- Dew point temperature is highly positively correlated to the Temperature.
- The demand for rental bikes is higher on Regular days(Non-Holidays) .
- There is more demand for rental bikes on Weekdays than on Weekends.
- The peak demands for rental bikes occur on the opening (8-9 AM) and closing times (6-7pm) of offices and institutions.
- There is a significant drop in the number of rented bikes during Winters(Dec-Feb) because it's freezing cold!
- The demand for bikes increases during warmer temperatures,which is why there's maximum count of rented bikes during the Summer season.

Feature Engineering

Checking with Multi-colinearity

	feature	VIF
0	Hour	4.458880
1	Temperature(°C)	188.666573
2	Humidity(%)	187.533688
3	Wind speed (m/s)	4.890096
4	Visibility (10m)	10.788995
5	Dew point temperature(°C)	126.954261
6	Solar Radiation (MJ/m2)	2.904971
7	Rainfall(mm)	1.103386
8	Snowfall (cm)	1.155412
9	Month	5.108772
10	Year	407.025112
11	Day	4.379818

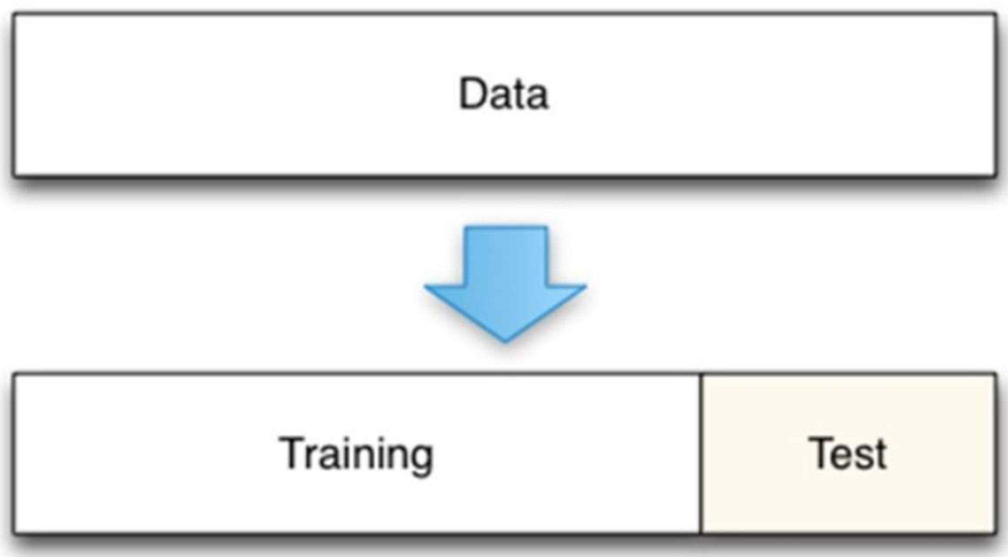
	feature	VIF
0	Hour	3.997641
1	Temperature(°C)	3.288024
2	Humidity(%)	6.802299
3	Wind speed (m/s)	4.667341
4	Visibility (10m)	5.471035
5	Solar Radiation (MJ/m2)	2.275006
6	Rainfall(mm)	1.080689
7	Snowfall (cm)	1.139759
8	Month	5.027060
9	Day	3.776455

Dropping dew-point temperature and year due to multi-colinearity problem

Dataset Splitting for Modelling

Train-data - (6132, 16)

Test-data : (2628,16)



Model Implementation

These were the models taken into account

- **Linear Regression (with regularization)**
- **Polynomial Regression**
- **Decision Tree Regressor**
- **Random Forest Regressor**
- **XGBoost Regressor**
- **CatBoost Regressor**

Model Selection & Validation

Metrics of various models used

	Regression Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	r2 score	adjusted r2 score
0	Multiple Linear Regression	327.665790	188770.692536	434.477494	0.539348	0.536525
1	Lasso Regression(Tuned)	327.578790	188737.802791	434.439642	0.539428	0.536606
2	Ridge Regression (default)	327.657172	188767.250135	434.473532	0.539356	0.536533
3	Ridge Regression(Tuned)	327.657214	188767.267238	434.473552	0.539356	0.536533
4	Elastic Net Regression(default)	332.689114	206068.926029	453.948153	0.497135	0.494054
5	Elastic Net Regression(Tuned)	327.559625	188730.345723	434.431060	0.539446	0.536624
6	Polynomial Regression(Tuned)	241.675585	134304.471436	366.475745	0.672260	0.670252
7	Decision Tree Regression (Tuned)	131.377984	52751.836188	229.677679	0.871271	0.870482
8	Random Forest Regression(Tuned)	103.313872	32096.340024	179.154514	0.921676	0.921196
9	XGBoost Regression(default)	168.226186	66671.412597	258.208080	0.837304	0.836307
10	XGBoost Regression(Tuned)	86.542427	22418.539122	149.728218	0.945293	0.944957
11	Catboost Regression(default)	91.881711	24268.891334	155.784760	0.940777	0.940460
12	Catboost Regression(tuned)	86.315659	22706.282422	150.686039	0.944590	0.944294

Model Selection & Validation

Observation 1:

Linear & polynomial regression models are not performing well, but Tree based models are performing better

Observation 2:

RF regressor, XGBoost regressor & Cat Boost regressor are giving better performance

Model Selection & Validation

Observation 3:

By performing Hyper-parameter tuning(Gridsearch CV) on XGBoost,CatBoost & RF Regressor, their performances were further improved.

Model Selection & Validation

CatBoost

```
Training score:0.985898943747007
MAE : 86.31565900219194
MSE : 22706.282422074244
RMSE : 150.68603924078118
R2 : 0.9445904572284984
Adjusted R2 :  0.9442935825255513
```

XGBoost

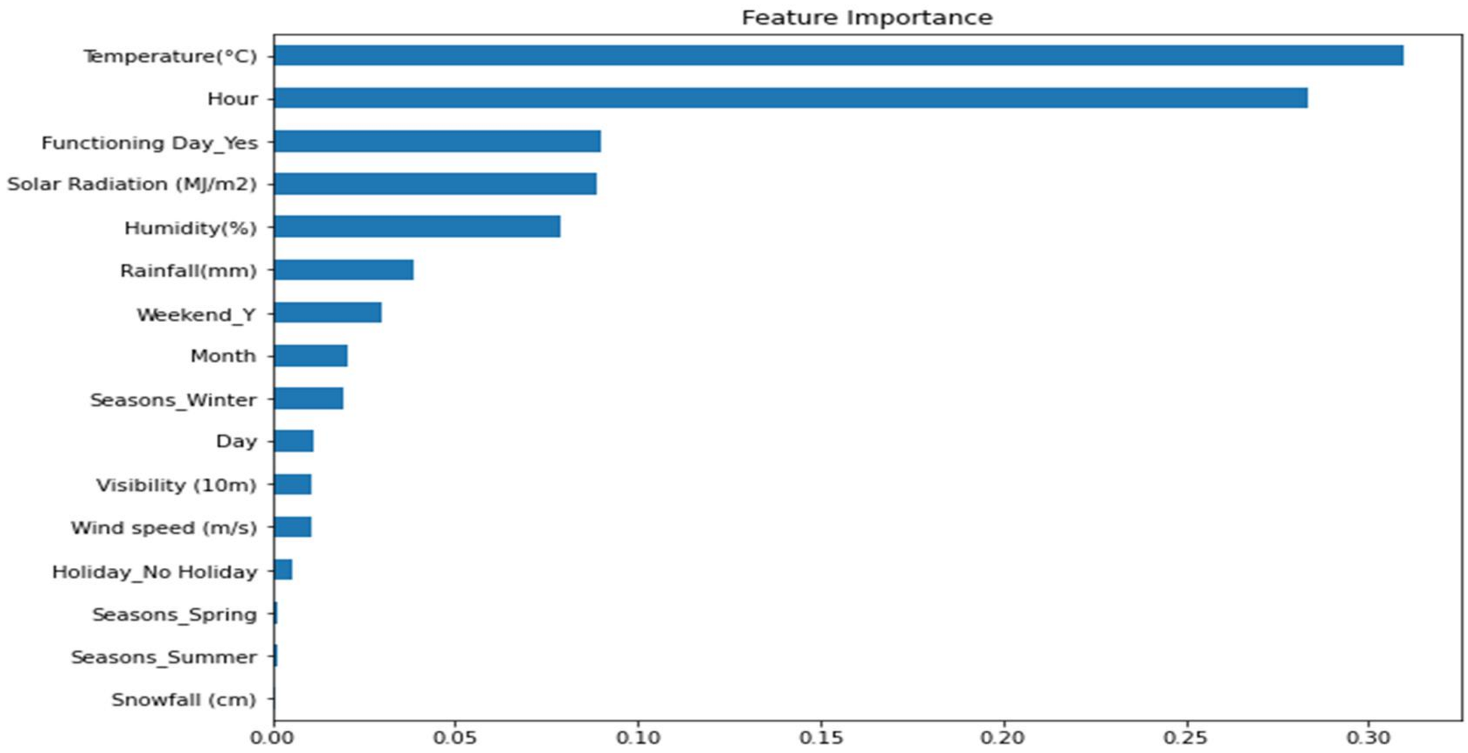
```
Training score:0.9951066432885016
MAE : 86.54242736425483
MSE : 22418.539122401522
RMSE : 149.72821752228776
R2 : 0.9452926296217621
Adjusted R2 :  0.9449573872142355
```

RF regressor

```
Training score:0.9902438072149231
MAE : 102.91840182648401
MSE : 31775.25514710807
RMSE : 178.2561503766646
R2 : 0.9224596820200995
Adjusted R2 :  0.9219845211286103
```

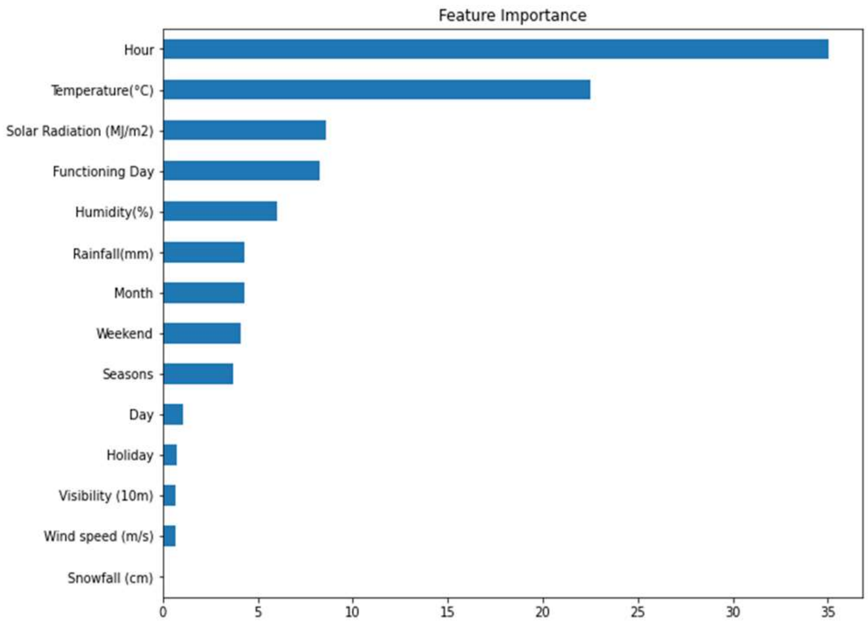
Feature Importances

Random Forest Regression

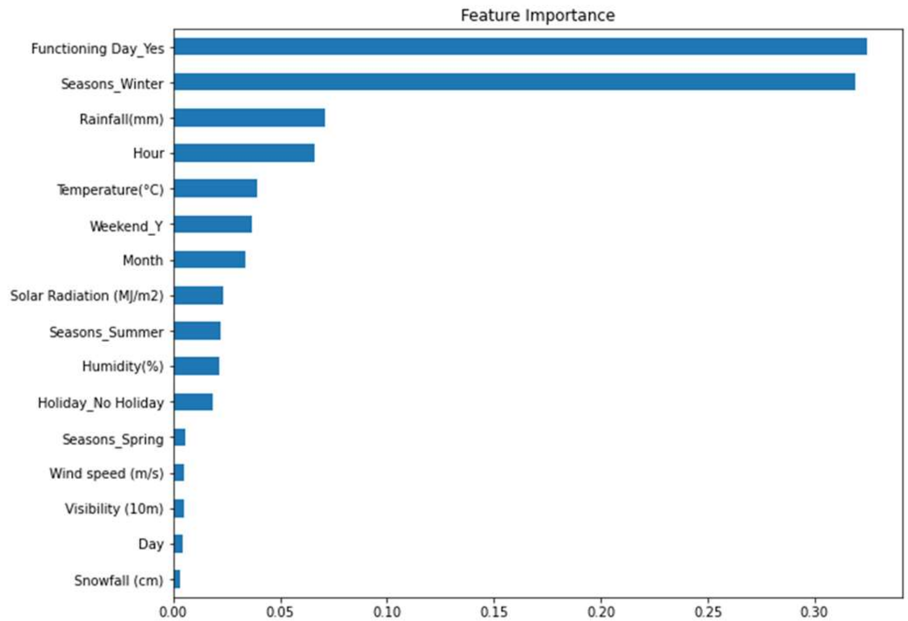


Feature Importances

Cat Boost



Xg Boost



Conclusions

- **Evaluating the performance metrics of the models has brought us to a conclusion that Decision tree based Ensemble models like XGBoost and CatBoost models are the most suitable for Predicting the number of bikes required on an hourly basis.**
- **The important features for prediction are : Hour & Temperature.**
- **Due to the lack of significant linear correlation between the independent variables and the count of Rented bikes,Linear regression and Polynomial regression are not good fit in this scenario.**

Thank You