Capstone Project

Bike Sharing Demand Prediction

By - Arindam Paul
(aripaul05@gmail.com)

Points for Discussion

01
Problem Statement

02
Data Description

03
Data Preparation and Cleaning

04
EDA (Exploratory Data Analysis)

Points for Discussion

05
Hypothesis Testing

06
Feature Engineering

07
Modelling

08
Model Interpretation

09
Conclusion

Problem Statement

Currently **Rental bikes** are introduced in many urban cities. 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 rental in Seoul from **2017-2018**. It includes **hourly observations** of **14 columns**, such as the **date**, **time**, **number of rented bikes**, **weather conditions**, and other factors that may influence **bike rental demand**.

This dataset contains **8760 rows** and **14 columns** of the data.

Data Description

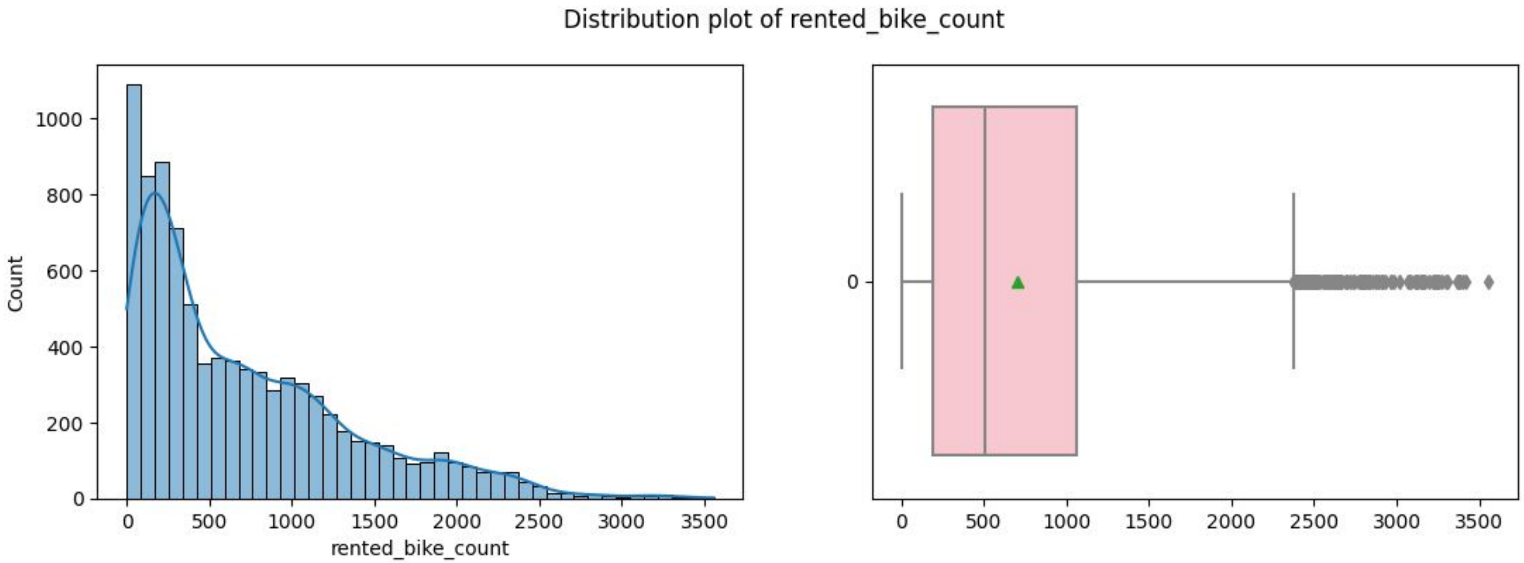
- **Date** : The date of the observation.
- **Rented Bike Count** : The number of bikes rented during the observation period.
- **Hour** : The hour of the day when the observation was taken.
- **Temperature(°C)** : The temperature in Celsius at the time of observation.
- **Humidity(%)** : The percentage 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 are **no duplicate rows** in the dataset.
- There are **no missing values** or **Null values** in the dataset.
- Change **datatype** of **Date** to **datetime**.
- From the **Date** column, **'month'** and **'day of the week'** columns are created.
- From the **'day of the week'** column, **'weekend'** column is created where **6** and **7** are the **weekends** (**Saturday** and **Sunday**).
- Change Data types of **numerical columns** which represents categories like **Month**, **Day of the Week**, **Weekend** to **categorical** data type.

Exploratory Data Analysis

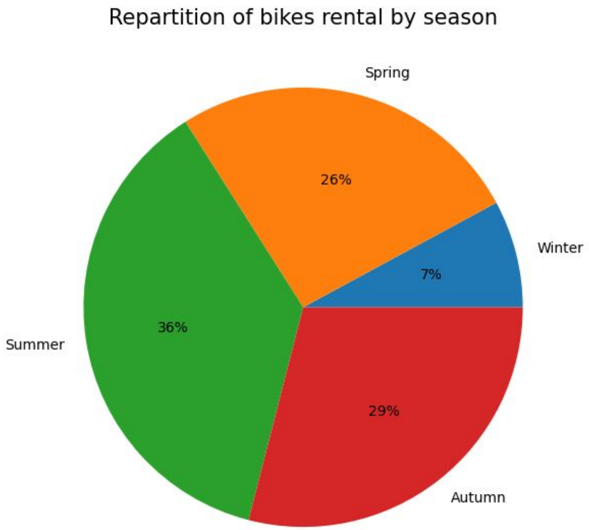
➢ Rented Bike Count Distribution

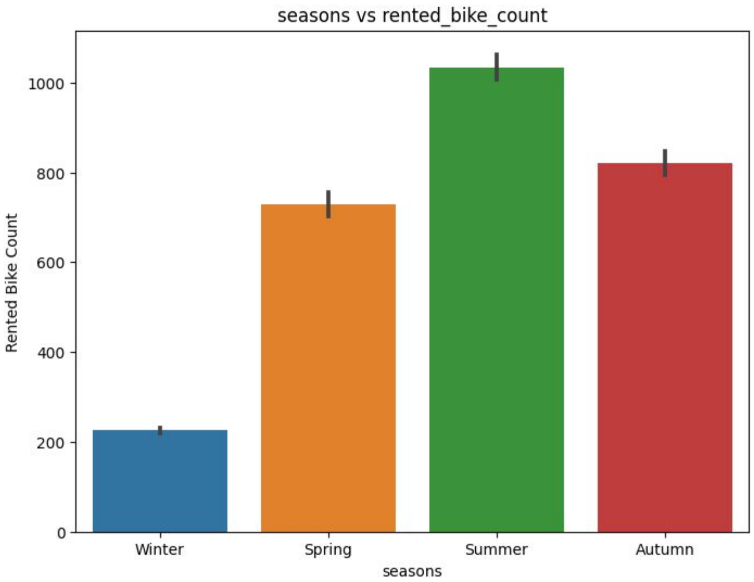


Exploratory Data Analysis

Rented Bike Count by Seasons

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

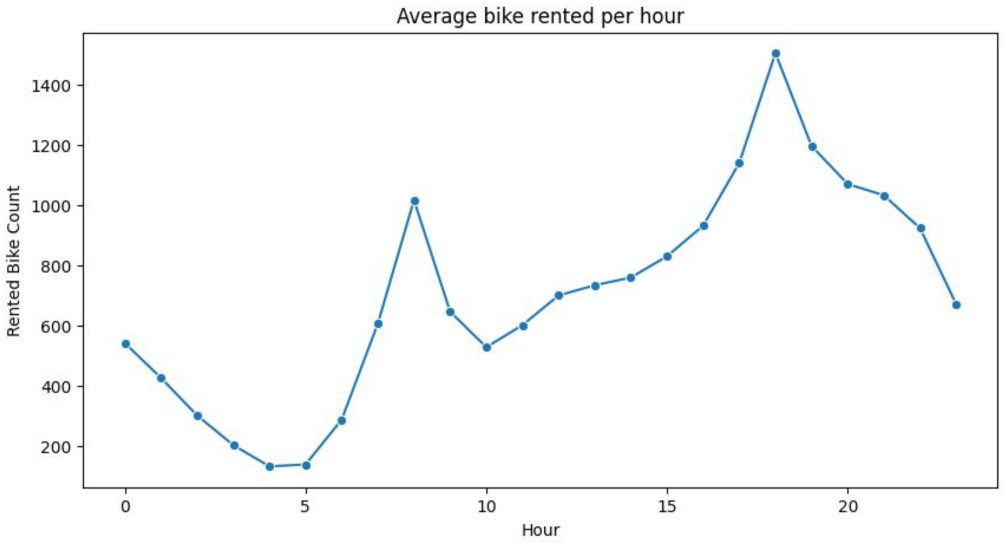




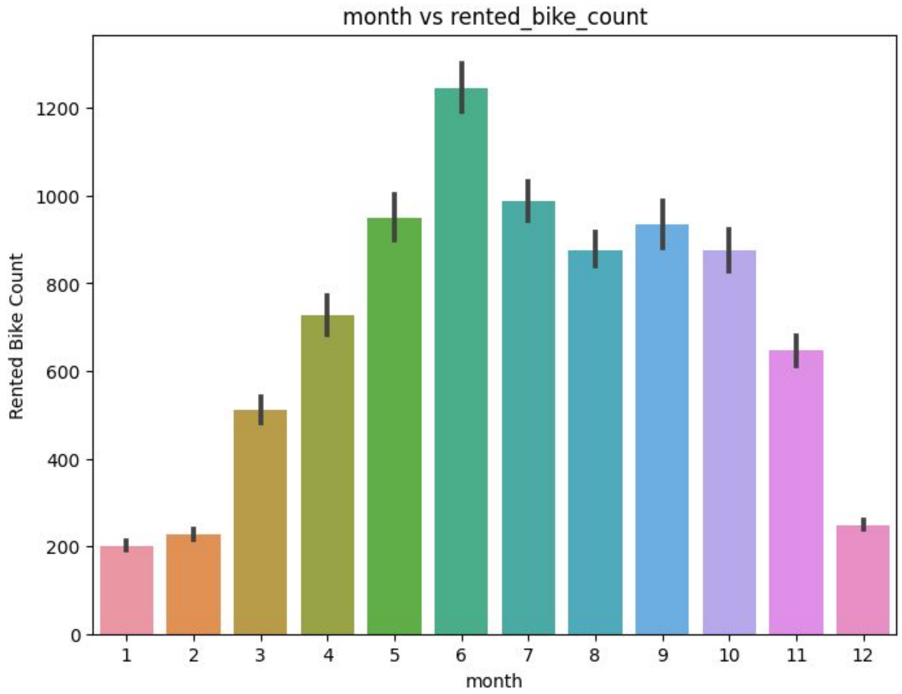
Exploratory Data Analysis

➢ 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



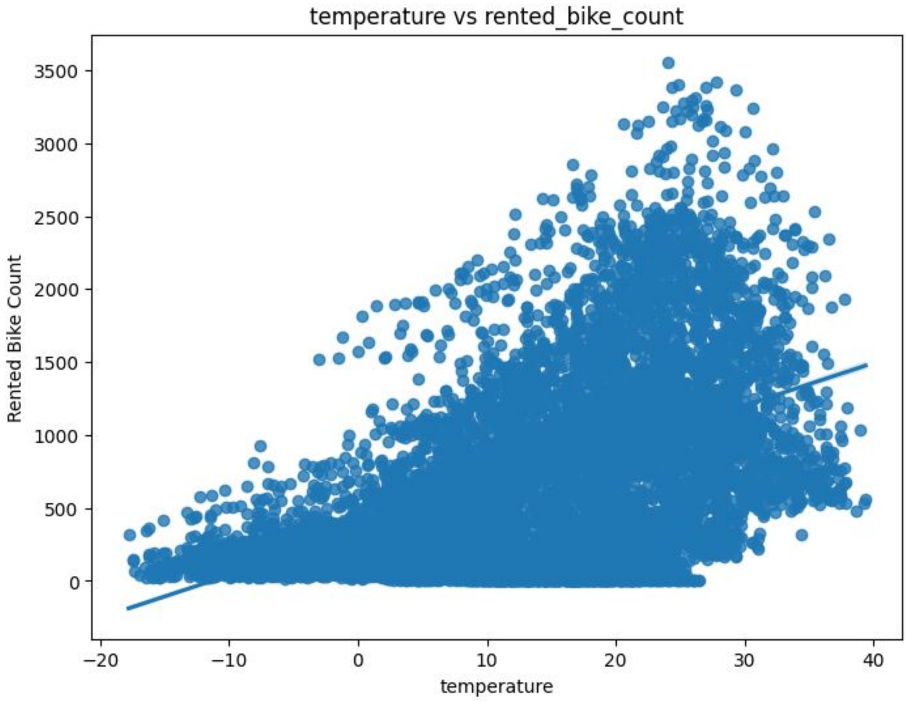
➢ Rented Bike Count by Months

- Similar to what we saw with **seasons**, **demand decreases** significantly during **winter** months like **Dec**, **Jan**, **Feb** etc.
- **Demand peaks** at the **summer** months like **May**, **June**, **July** etc.

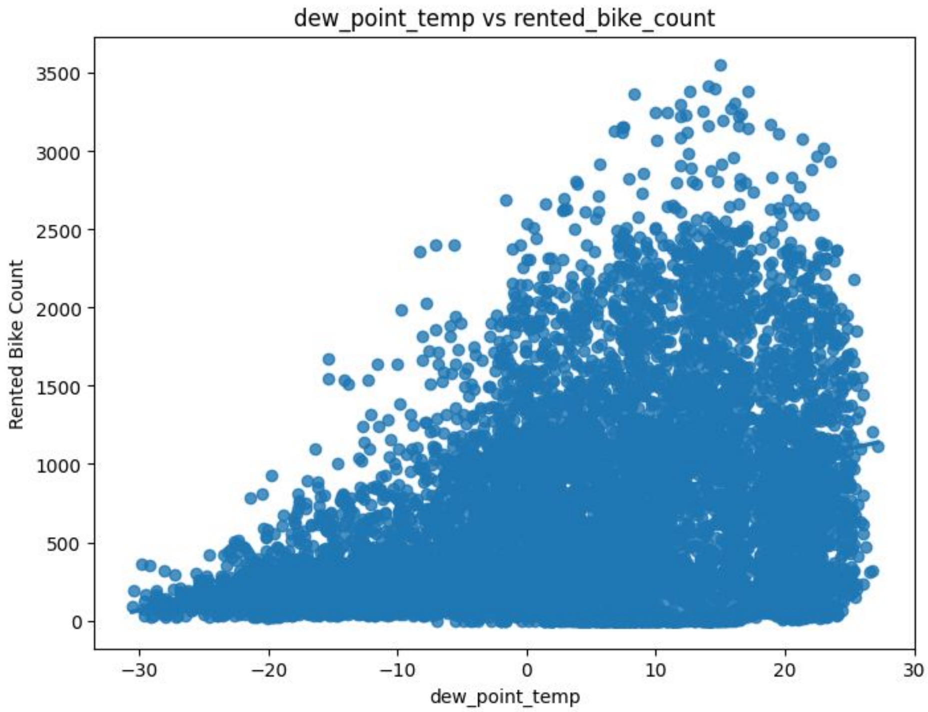
Exploratory Data Analysis

➢ 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



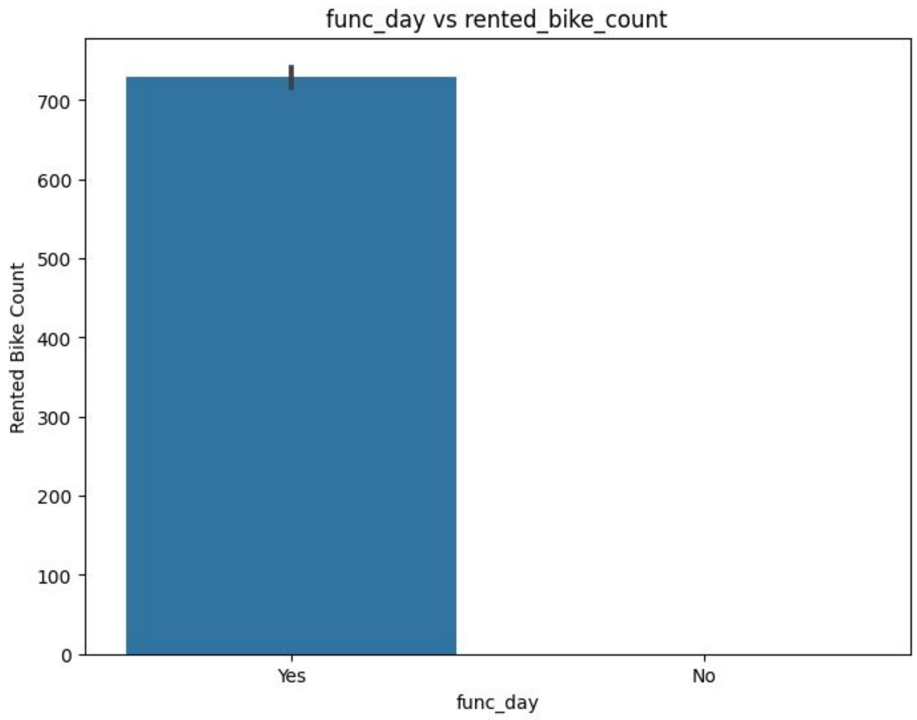
➢ Rented Bike Count by Dew Point Temperature

- 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.

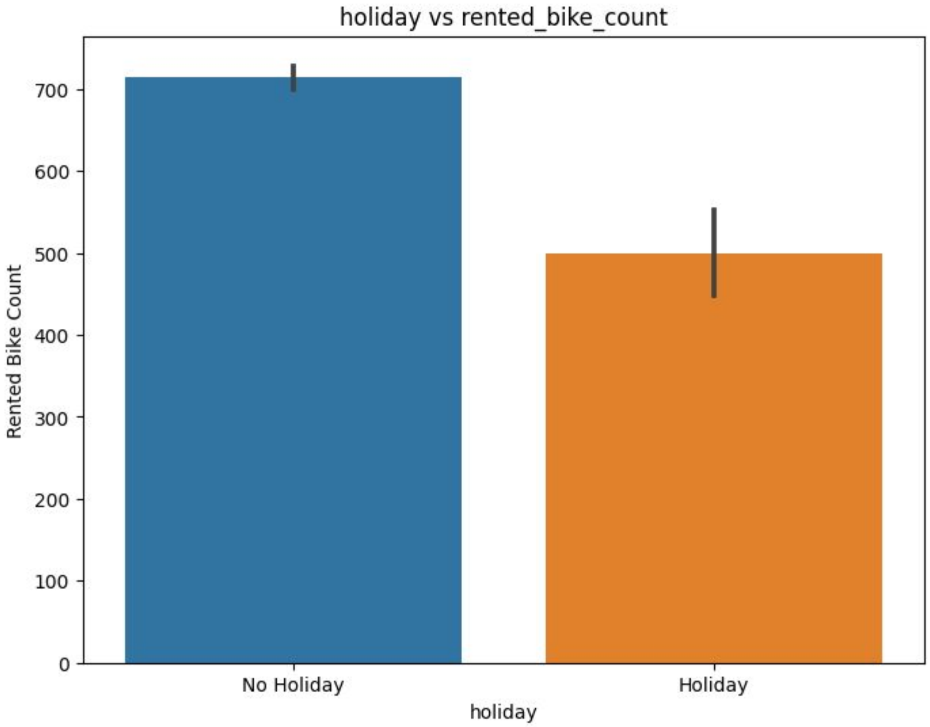
Exploratory Data Analysis

➢ 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



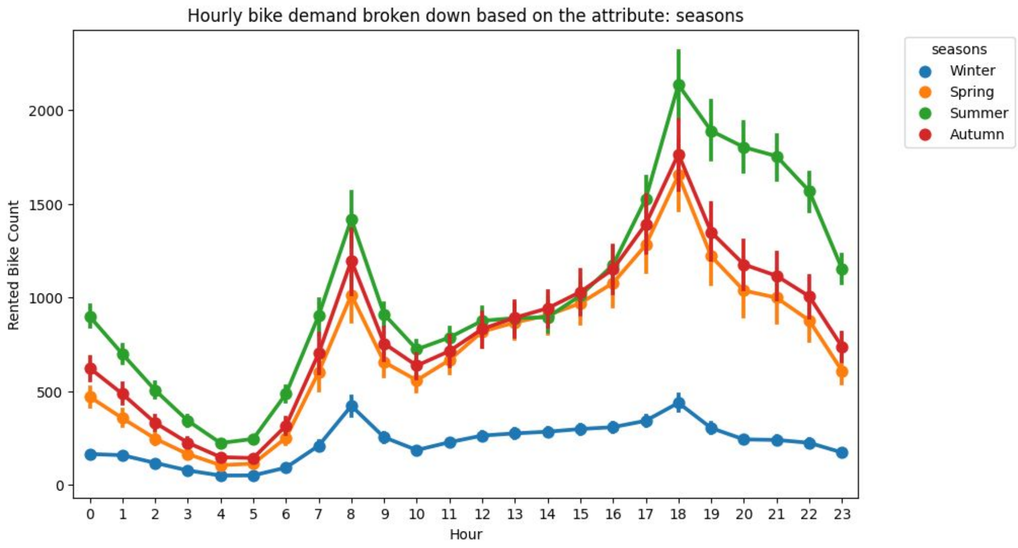
➢ Rented Bike Count by Holiday

- **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**.

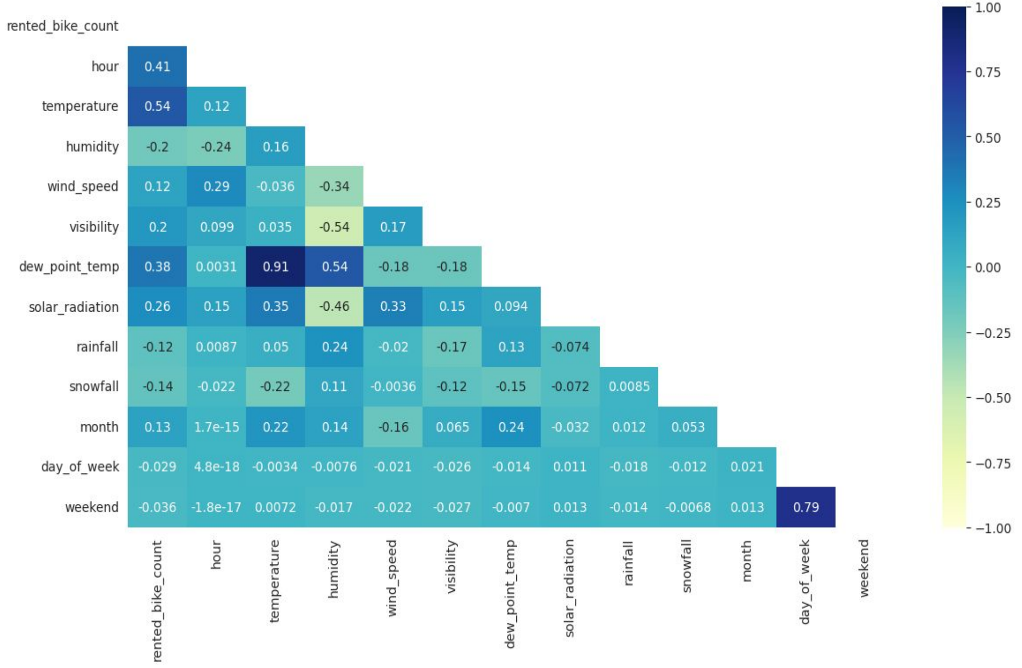
Exploratory Data Analysis

➢ Rented Bike Count by Hour by 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



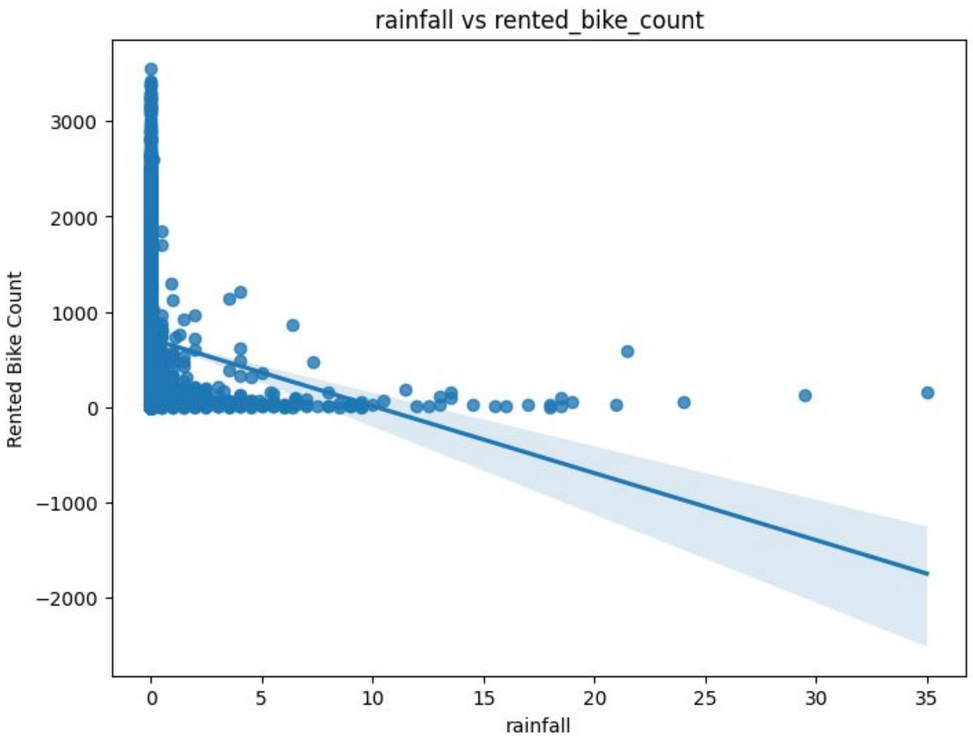
➢ Correlation of features

- **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.

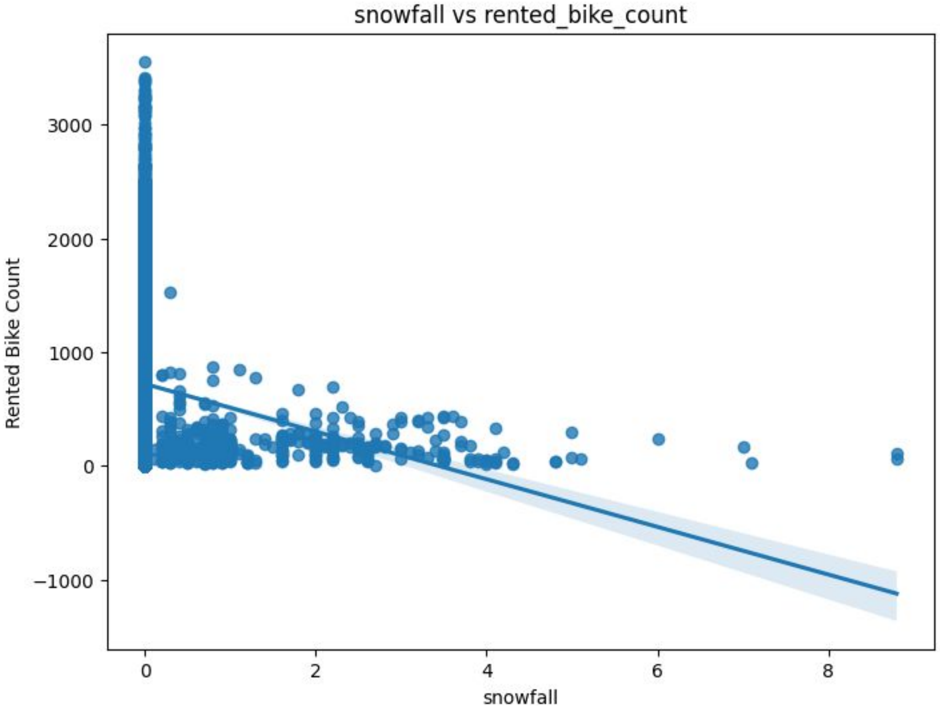
Exploratory Data Analysis

➢ Rented Bike Count by Rainfall

- **Rainfall** leads to **decrease** in the **demand** in **bike rentals**.
- This is obvious because **people do not** want to **go out** on a bike when it is **raining unless** it is **emergency**.



Exploratory Data Analysis



➢ Rented Bike Count by Snowfall

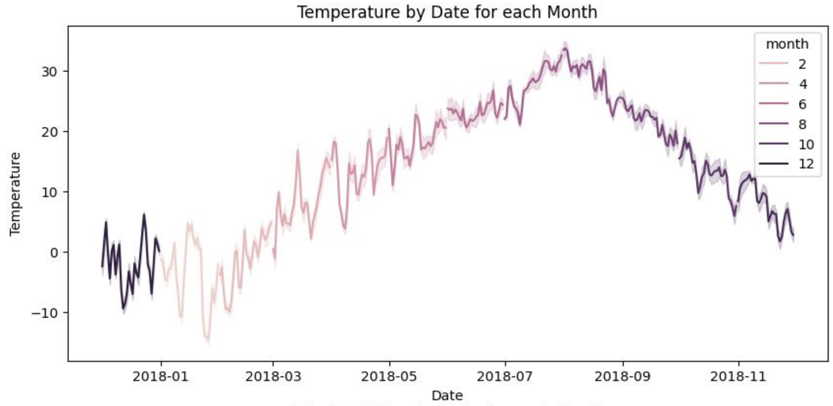
- Similarly **snowfall** leads to **decrease** in the **demand** in **bike rentals**.
- This is obvious because **people** also **do not** want to **go out** on a bike when it is **snowing unless** it is **emergency**.

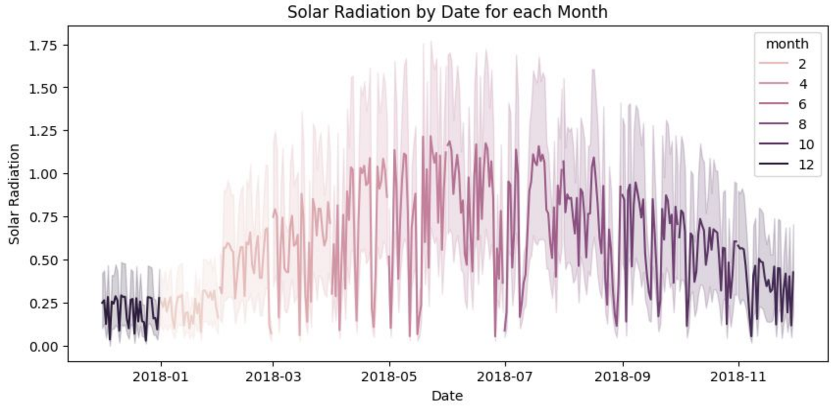
Exploratory Data Analysis

Temperature and Solar Radiation over time

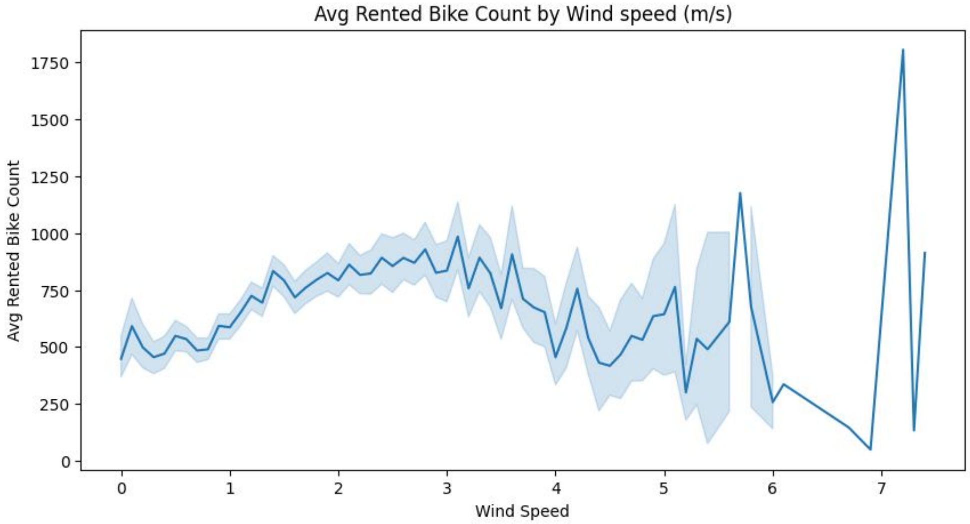
- As expected **temperature rises** during **summer months** like **May**, **June**, **July** etc. and **decreases** during months like **Dec**, **Jan** etc.

- 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



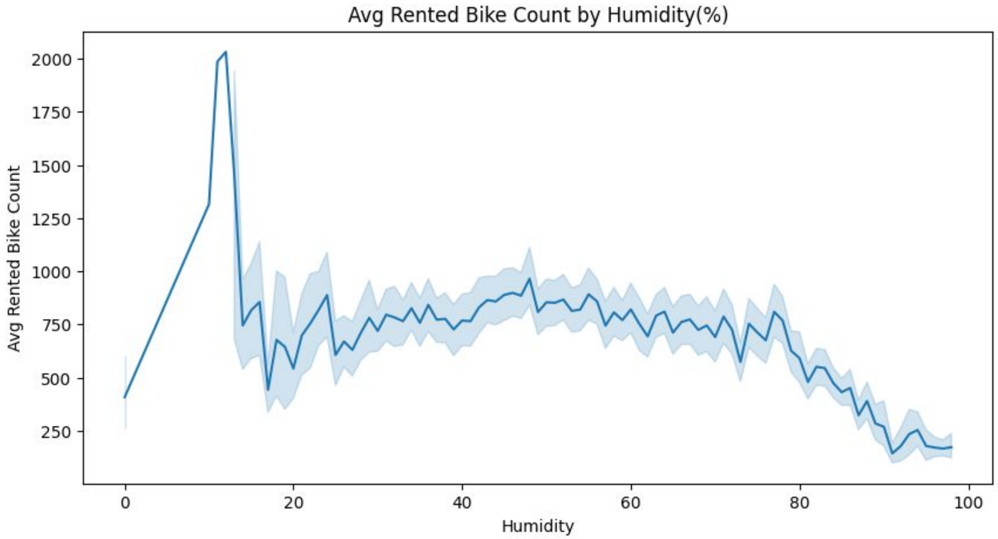
➢ Rented Bike Count by Wind Speed (m/s)

- 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 is starts **decreasing**.
- One reason for such **high humidity** can be **rain** and we already saw **rain causes decreases** in demand.



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., **nhot = 2928** & **ncold = 5832**) and we don't know **σp**.

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.**

- 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., **nrush = 2190** & **nnon−rush = 6570**) and we don't know **σp**.

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

```
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., **nautumn = 2184**, **nspring = 2208**, **nsummer = 2208**, **nwinter = 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
-------------------------------------------------------
```

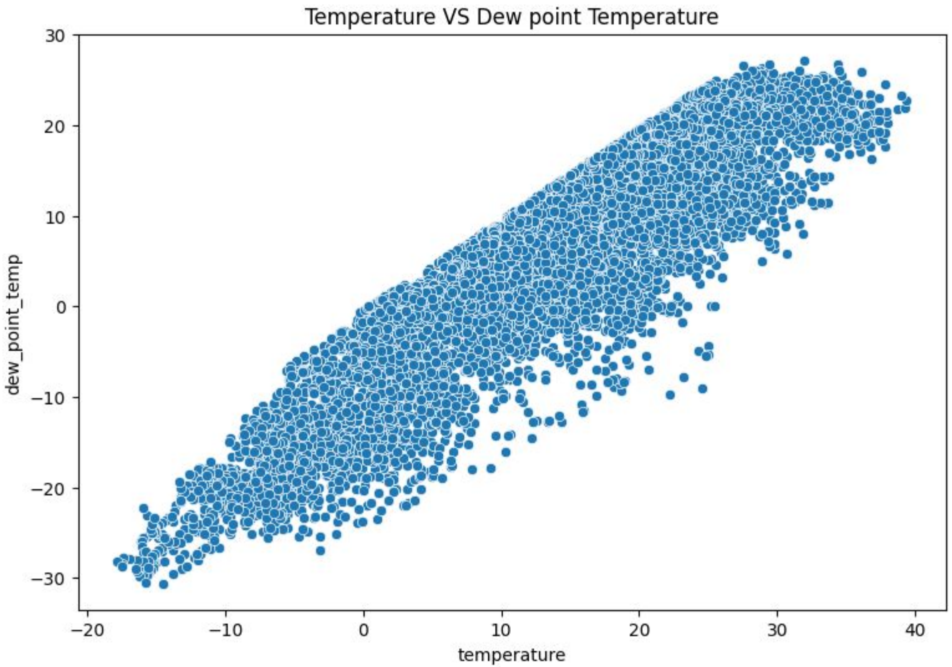
Null Hypothesis: H_o : **No significant difference** between rented bike counts for different seasons.

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

Test Type: One-way ANOVA test

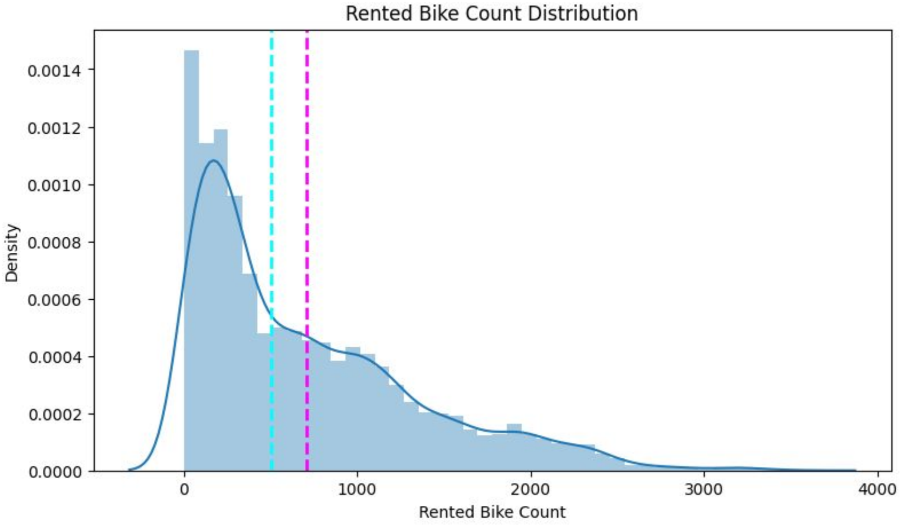
Feature Engineering

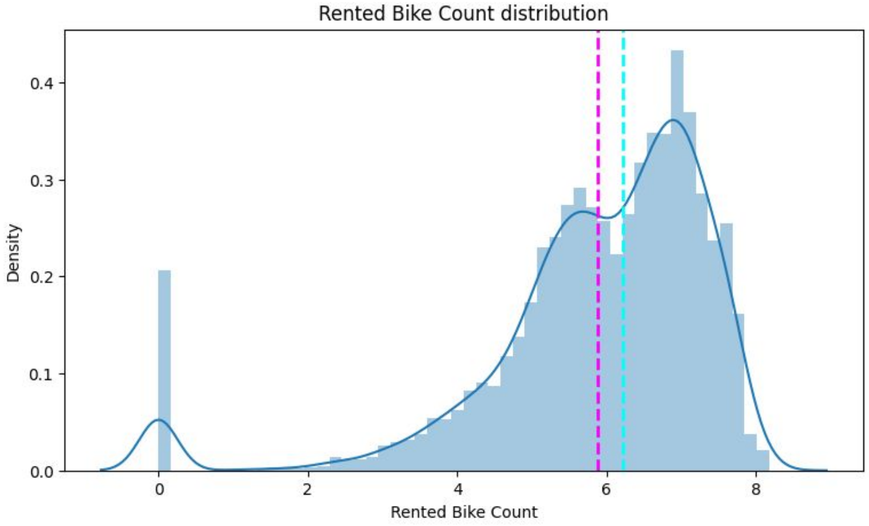
- I have used **pearson correlation coefficient** to check **correlation** between **variables** and also with **dependent variable**.
- And also i check the **multicollinearity** using **VIF** and **remove** those who are having **high VIF** value.
- Found that there is **high correlation** between **temperature** and **dew point temperature**. So, i take **50 %** of the both and **create** new variable **'temp'** by adding both of them.



Feature Engineering

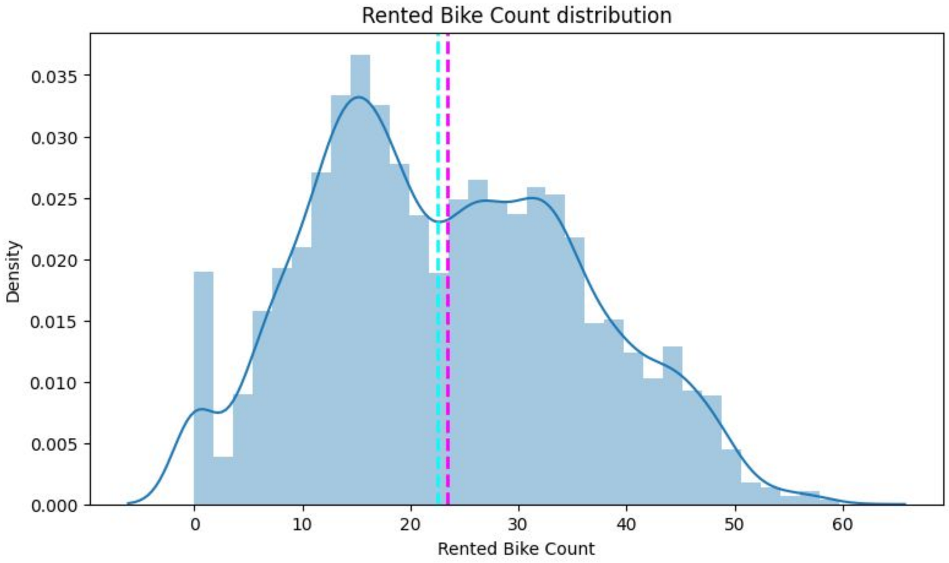
- The **Rented Bike Count** was **right skewed**, and to train a robust model **transform** it to **normal**.
- Applied **square root** to **transform** it to **normal**.





Feature Engineering

- I have different **independent features** of different scale so i have used **standard scalar** method to scale our independent features into one scale.
- **Splitted Data** into train and test sets with ratio **80:20**.



Modelling

- Since we're trying to predict **continuous variable**, I trained various **regression algorithms** along with **hyper parameter tuning** and **cross validation** to get the best model.

01 Linear Regression

02 Lasso Regression

03 Ridge Regression

04 Decision Tree

05 Random Forest

06 Gradient Boosting

07 Xtreme Gradient Boosting

Modelling

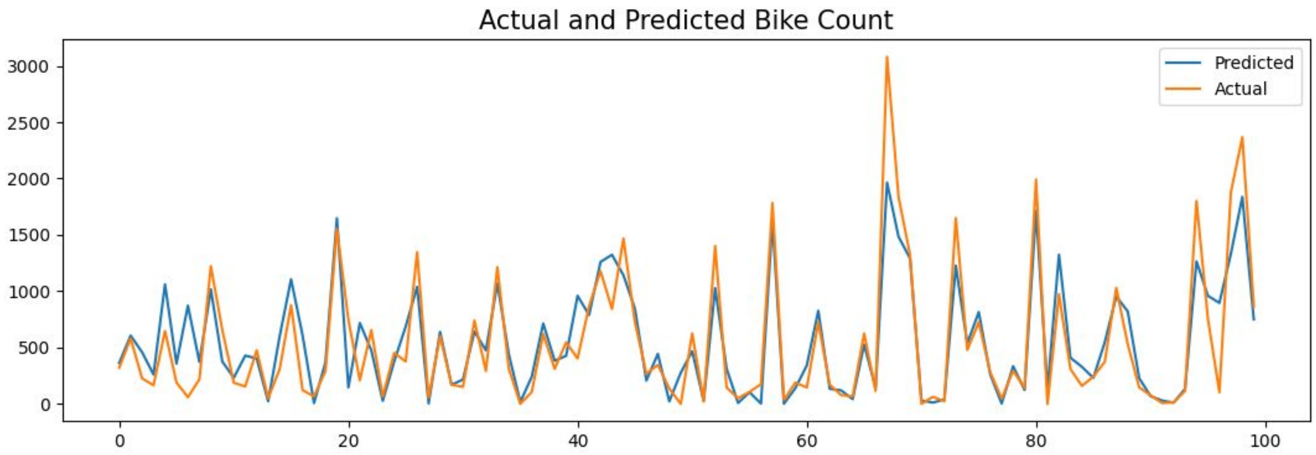
➢ Linear Regression

Performance (before tuning)

```
MSE : 88090.65909000415
RMSE : 296.80070601331823
MAE : 201.8068025396329
Train R2 : 0.784428200422006
Test R2 : 0.7895199410494631
Adjusted R2 :  0.7828222844888686
```

Performance (after tuning)

```
MSE : 88090.65909000415
RMSE : 296.80070601331823
MAE : 201.8068025396329
Train R2 : 0.784428200422006
Test R2 : 0.7895199410494631
Adjusted R2 :  0.7828222844888686
```



Modelling

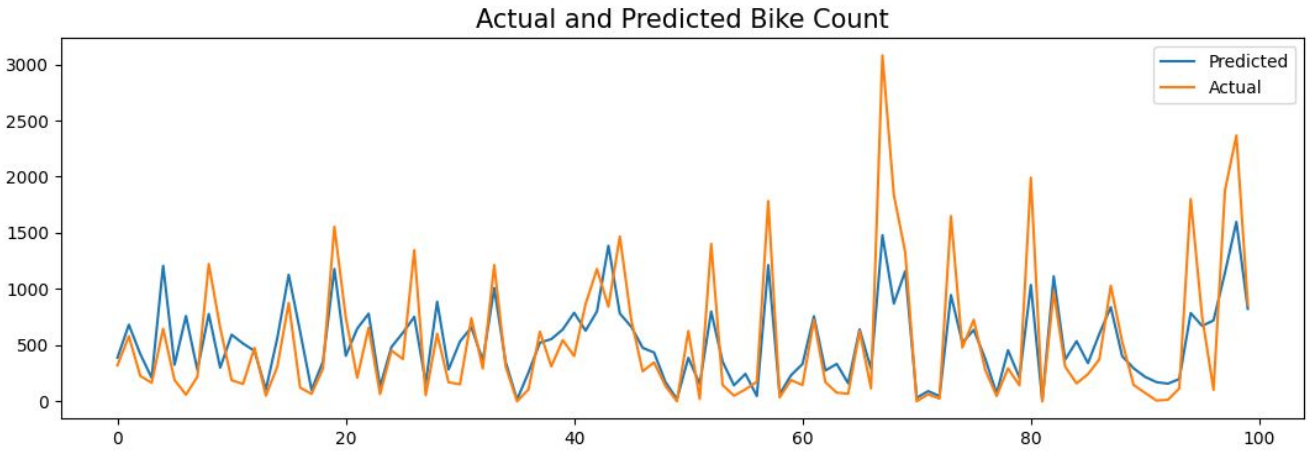
➢ Lasso Regression

Performance (before tuning)

```
MSE : 199251.13943499743
RMSE : 446.37555873389556
MAE : 303.758212165632
Train R2 : 0.5201240107717402
Test R2 : 0.5239178363804666
Adjusted R2 :  0.5087684923407172
```

Performance (after tuning)

```
MSE : 88358.33989461442
RMSE : 297.25130764155506
MAE : 201.70425481228804
Train R2 : 0.7834759534324424
Test R2 : 0.7888803559661375
Adjusted R2 :  0.7821623472579298
```



Modelling

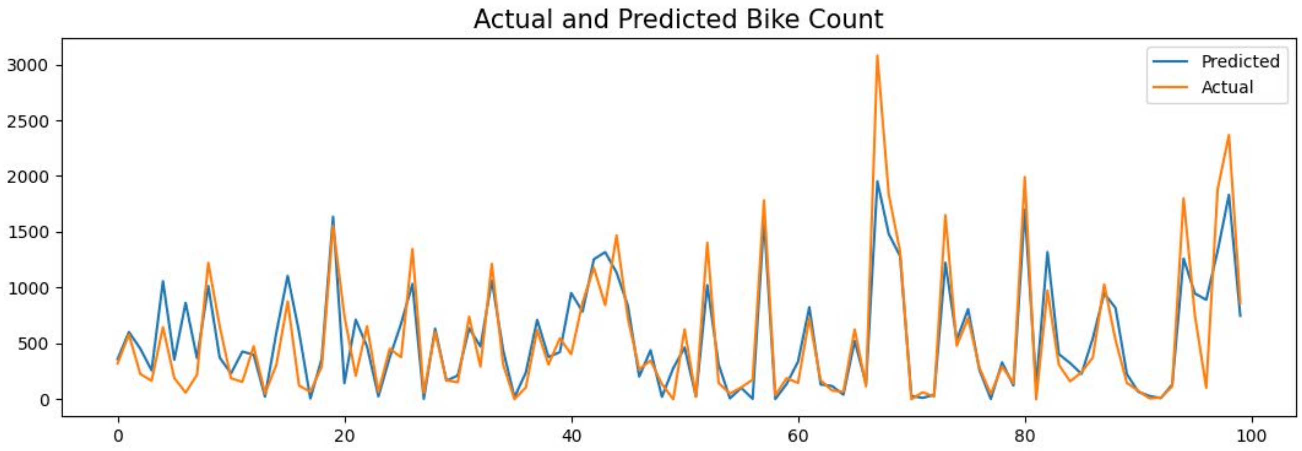
➢ Ridge Regression

Performance (before tuning)

```
MSE : 88365.68734453894
RMSE : 297.26366637135277
MAE : 201.717161005283
Train R2 : 0.7834601310784806
Test R2 : 0.788862800283058
Adjusted R2 :  0.7821442329379108
```

Performance (after tuning)

```
MSE : 88416.84817480511
RMSE : 297.3497068685374
MAE : 201.78332256985254
Train R2 : 0.7833529276131157
Test R2 : 0.7887405587800244
Adjusted R2 :  0.7820181016050811
```



Modelling

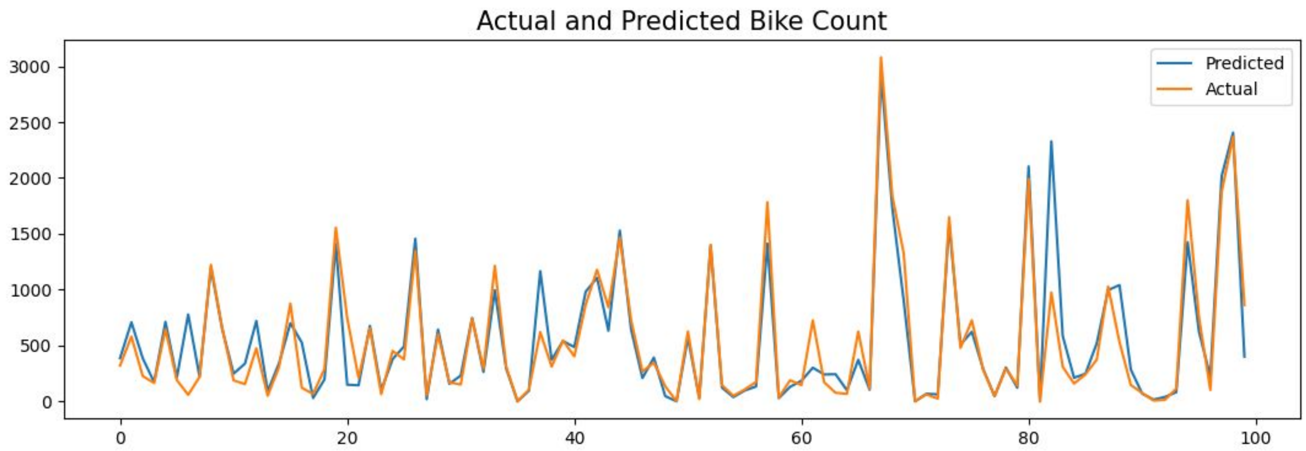
Decision Tree

Performance (before tuning)

```
MSE : 73491.68150684932
RMSE : 271.09349218830266
MAE : 152.0673515981735
Train R2 : 1.0
Test R2 : 0.824402114642698
Adjusted R2 :  0.8188144388564315
```

Performance (after tuning)

```
MSE : 89557.65707345174
RMSE : 299.2618536891258
MAE : 184.41649011155485
Train R2 : 0.8375934856837123
Test R2 : 0.7860147587154215
Adjusted R2 :  0.7792055642373029
```



Modelling

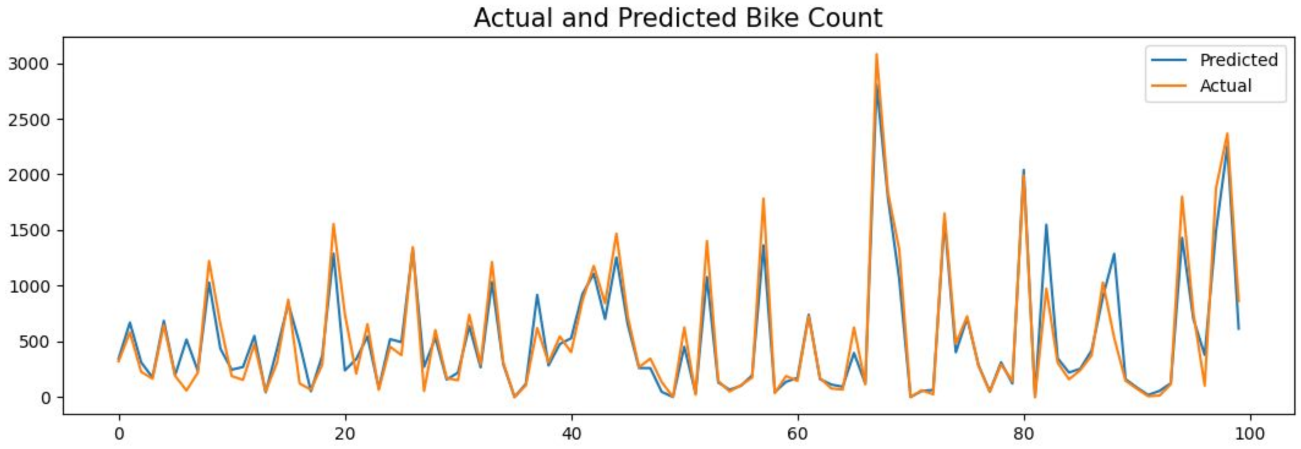
➢ Random Forest

Performance (before tuning)

```
MSE : 38747.939048529646
RMSE : 196.8449619587193
MAE : 112.34713100325216
Train R2 : 0.9881686099987611
Test R2 : 0.9074173291538947
Adjusted R2 :  0.9044712689148319
```

Performance (after tuning)

```
MSE : 75962.32375859405
RMSE : 275.6126335250147
MAE : 166.3139028404189
Train R2 : 0.8558829868398841
Test R2 : 0.8184988675542456
Adjusted R2 :  0.8127233453668145
```



Modelling

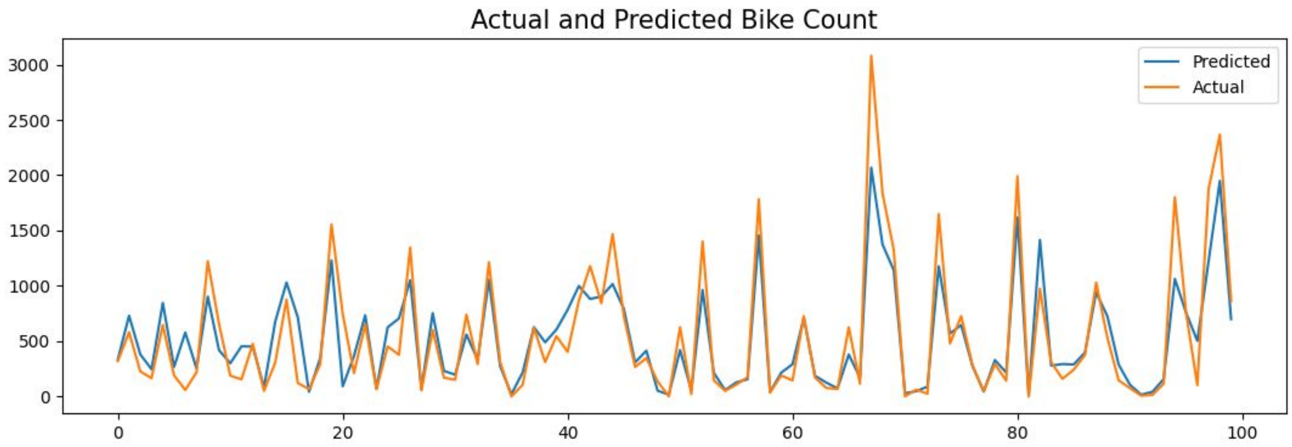
➢ Gradient Boosting

Performance (before tuning)

```
MSE : 77975.24920161186
RMSE : 279.2404863224741
MAE : 186.13005739507054
Train R2 : 0.8258252767878081
Test R2 : 0.8136892694619378
Adjusted R2 :  0.8077607017253112
```

Performance (after tuning)

```
MSE : 28399.67260989422
RMSE : 168.52202410929624
MAE : 97.43931235582568
Train R2 : 0.99476509675711
Test R2 : 0.9321430350634672
Adjusted R2 :  0.9299837680590047
```



Modelling

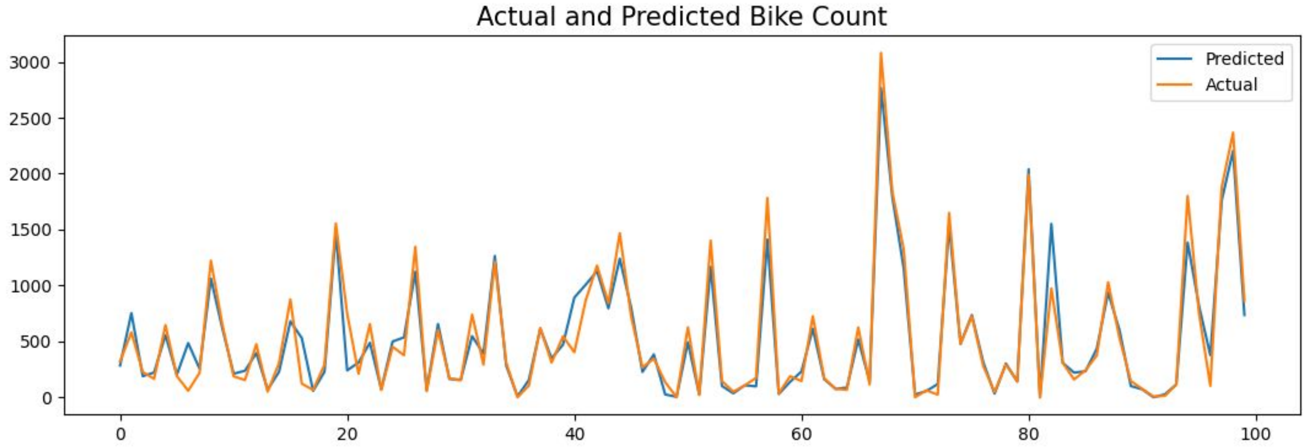
➢ Xtreme Gradient Boosting

Performance (before tuning)

```
MSE : 30509.675256575534
RMSE : 174.67018994830096
MAE : 103.70351631526435
Train R2 : 0.9759196537311359
Test R2 : 0.9271014848463719
Adjusted R2 :  0.9247817913765451
```

Performance (after tuning)

```
MSE : 27293.013392206205
RMSE : 165.2059726287346
MAE : 95.98635923067152
Train R2 : 0.9991356437190565
Test R2 : 0.934787239338737
Adjusted R2 :  0.9327121131892331
```



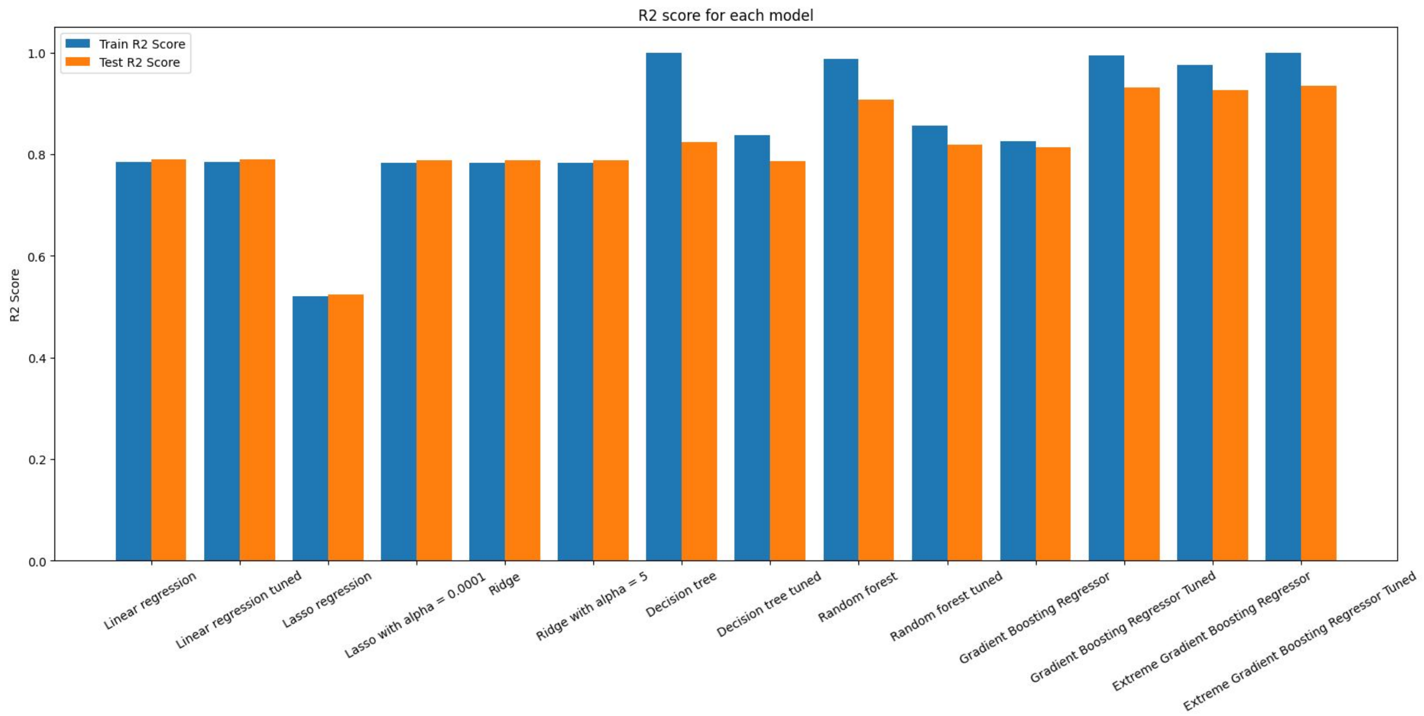
Modelling

➢ Performance Comparison

	Linear regression	Linear regression tuned	Lasso regression	Lasso with alpha = 0.0001	Ridge	Ridge with alpha = 5	Decision tree	Decision tree tuned	Random forest	Random forest tuned	Gradient Boosting Regressor	Gradient Boosting Regressor Tuned	Extreme Gradient Boosting Regressor	Extreme Gradient Boosting Regressor Tuned
MSE	88090.659090	88090.659090	199251.139435	88358.339895	88365.687345	88416.848175	73491.681507	89557.657073	38747.939049	75962.323759	77975.249202	28399.672610	30509.675257	27293.013392
RMSE	296.800706	296.800706	446.375559	297.251308	297.263666	297.349707	271.093492	299.261854	196.844962	275.612634	279.240486	168.522024	174.670190	165.205973
MAE	201.806803	201.806803	303.758212	201.704255	201.717161	201.783323	152.067352	184.416490	112.347131	166.313903	186.130057	97.439312	103.703516	95.986359
Train R2	0.784428	0.784428	0.520124	0.783476	0.783460	0.783353	1.000000	0.837593	0.988169	0.855883	0.825825	0.994765	0.975920	0.999136
Test R2	0.789520	0.789520	0.523918	0.788880	0.788863	0.788741	0.824402	0.786015	0.907417	0.818499	0.813689	0.932143	0.927101	0.934787
Adjusted R2	0.782822	0.782822	0.508768	0.782162	0.782144	0.782018	0.818814	0.779206	0.904471	0.812723	0.807761	0.929984	0.924782	0.932712

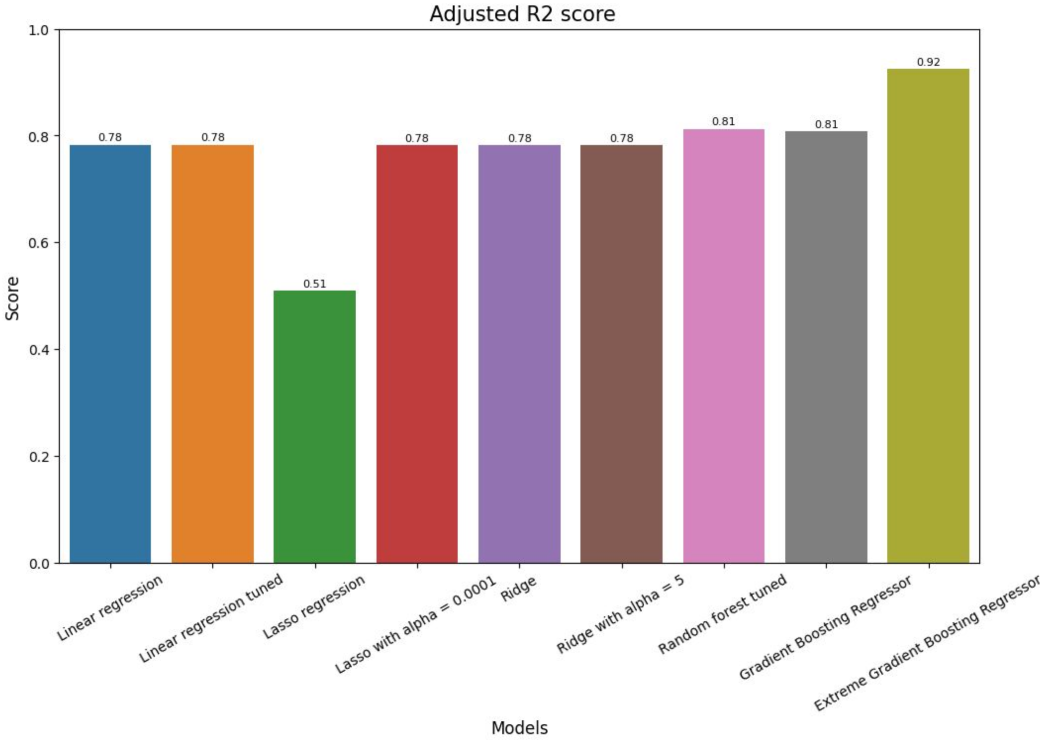
Modelling

➢ Plot of R2 score for each model

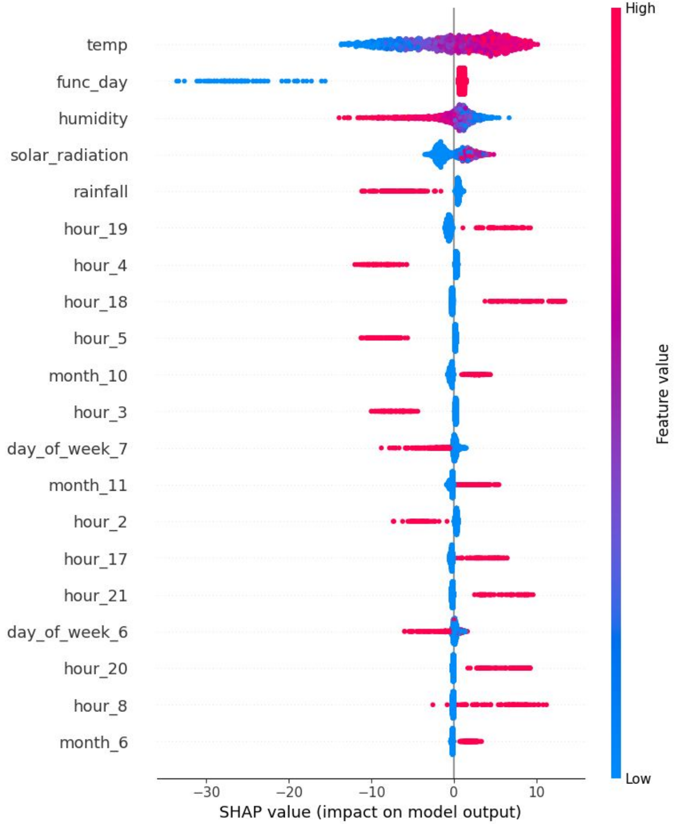


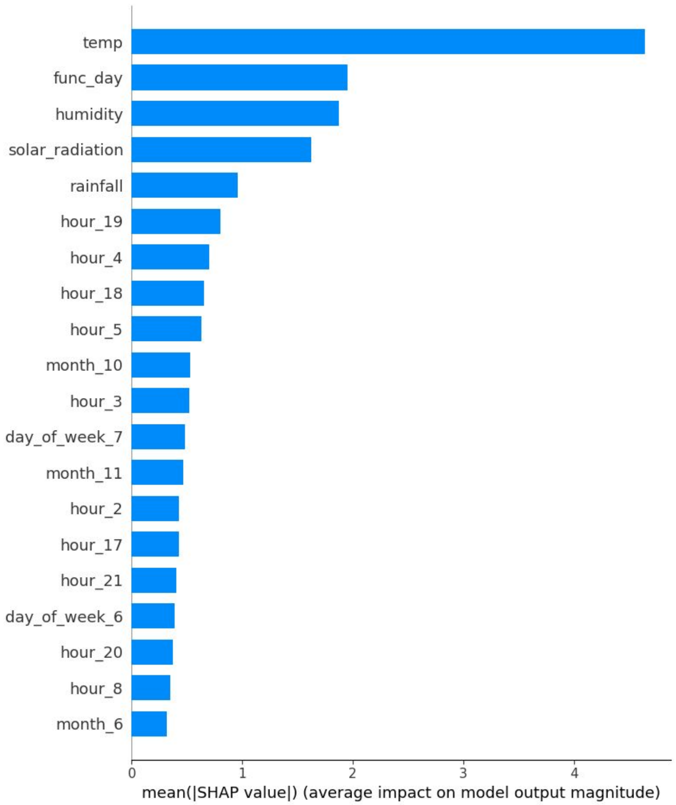
Modelling

➢ Plot of adjusted R2 score



Model Interpretation





Conclusion

- The **XGBoost** (**Extreme Gradient Boosting**) which gave the **best result** for **predicting Rented Bike Count** using several features on both **train** and **test** data with **R2 score** of **0.92**.
- There is **no use** of **removing outliers**, it **affects negatively** on model performance.

Thank You!