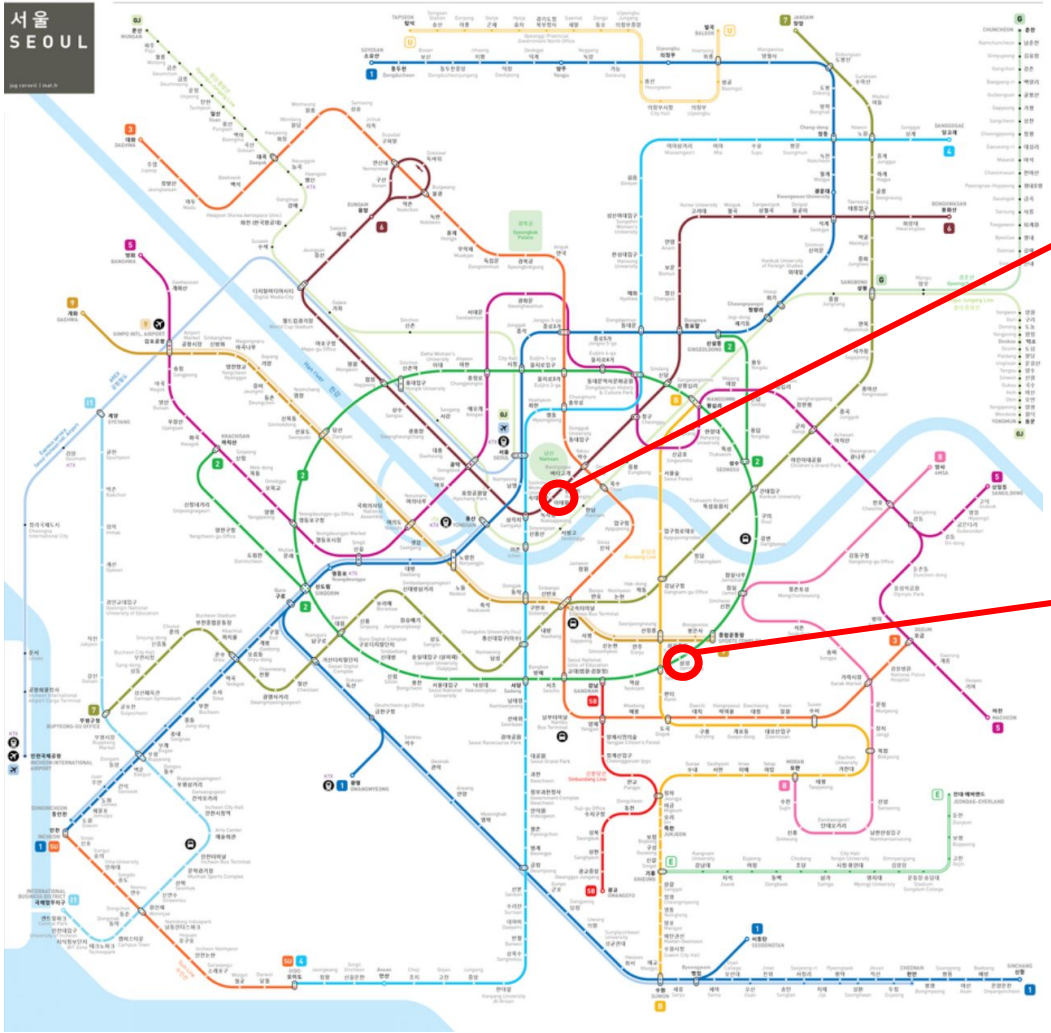
Seoul Bike Sharing

Final Presentation for 2023 Spring Cloud Computing

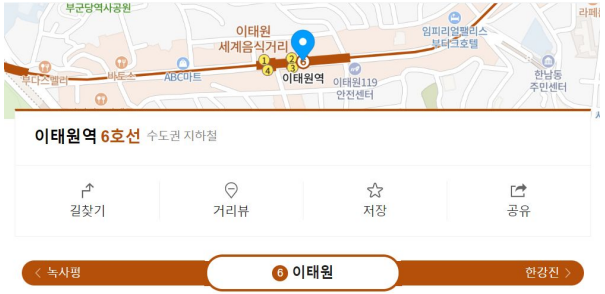
HaeLee Kim, Kyuri Kim, Nayaeun Kwon



[A Complete Guide to Attractions in Seoul | The Official Travel Guide to Seoul (visitseoul.net)]

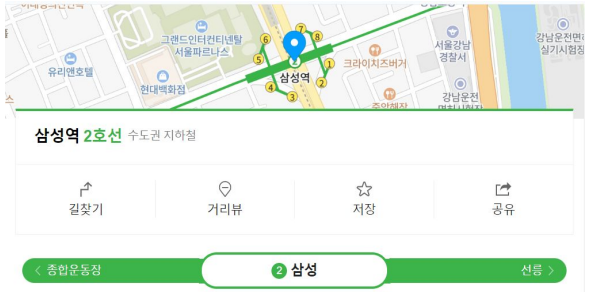
Itaewon

https://english.visitseoul.net/itaewon?curPage=1



Samsung - COEX

https://english.visitseoul.net/gangnamarea



Bike Sharing System in Seoul bikeseoul.com

city government system

20,000+ bikes

24/7 operation

stations strategically located

web/mobile registration

GPS tracking

Dataset Information

Number of Instances : 8760

Number of Features : 14

Types of Features : Integer, Float, and Character

Date Range : 12/1/2017 - 11/30/2018

Date Donated : 3/1/2020

Data Source URL: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#

Original Datasets URL: http://data.seoul.go.kr/ and publicholidays.go.kr

The Implemented Features

Date : year-month-day

Hour : Hour of the day

Rented Bike count : Count of bikes rented at each hour

Temperature : Temperature in Celsius

Humidity : %

Wind speed : m/s

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)

Expected Outcomes

Goal

Predict the bike count required at each hour and season for the stable supply of rental bikes

Target

Rented bike count

Hypothesis 1

During rush hours, the bike rental count will be higher compared to non-rush hours.

Hypothesis 2

Bike rental count will decrease when the temperature drops below 0°C (32°F) or rises above 30°C (86°F).

ML Models

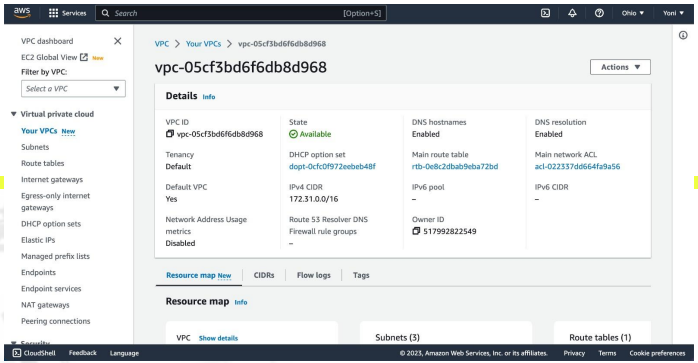
Regression Analysis, Random Forest Analysis, KNN, and XGBoost

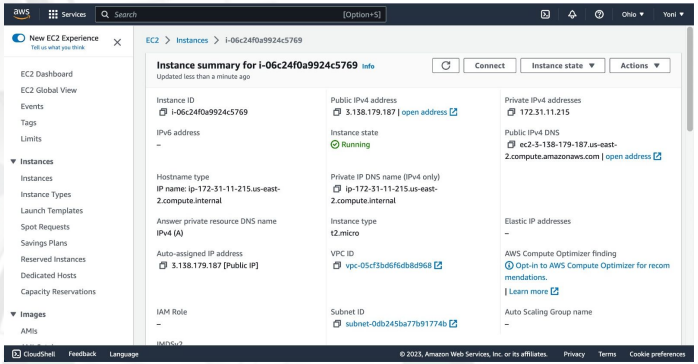
AWS cloud services in action

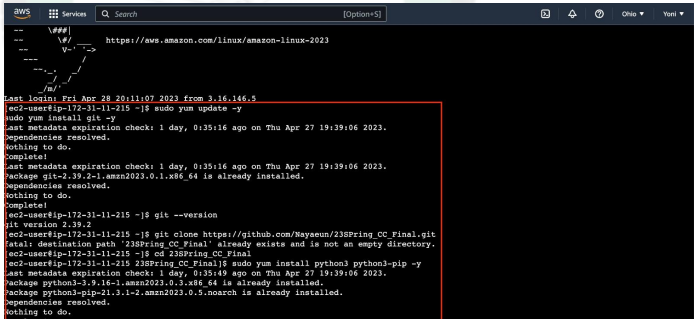
AWS cloud services used are VPC & EC2.

Logical architecture

1. Created a VPC - Public Subnet
2. Launched EC2
3. Set up Git Repository
4. Collaborated the project in Git
5. Ran ML models



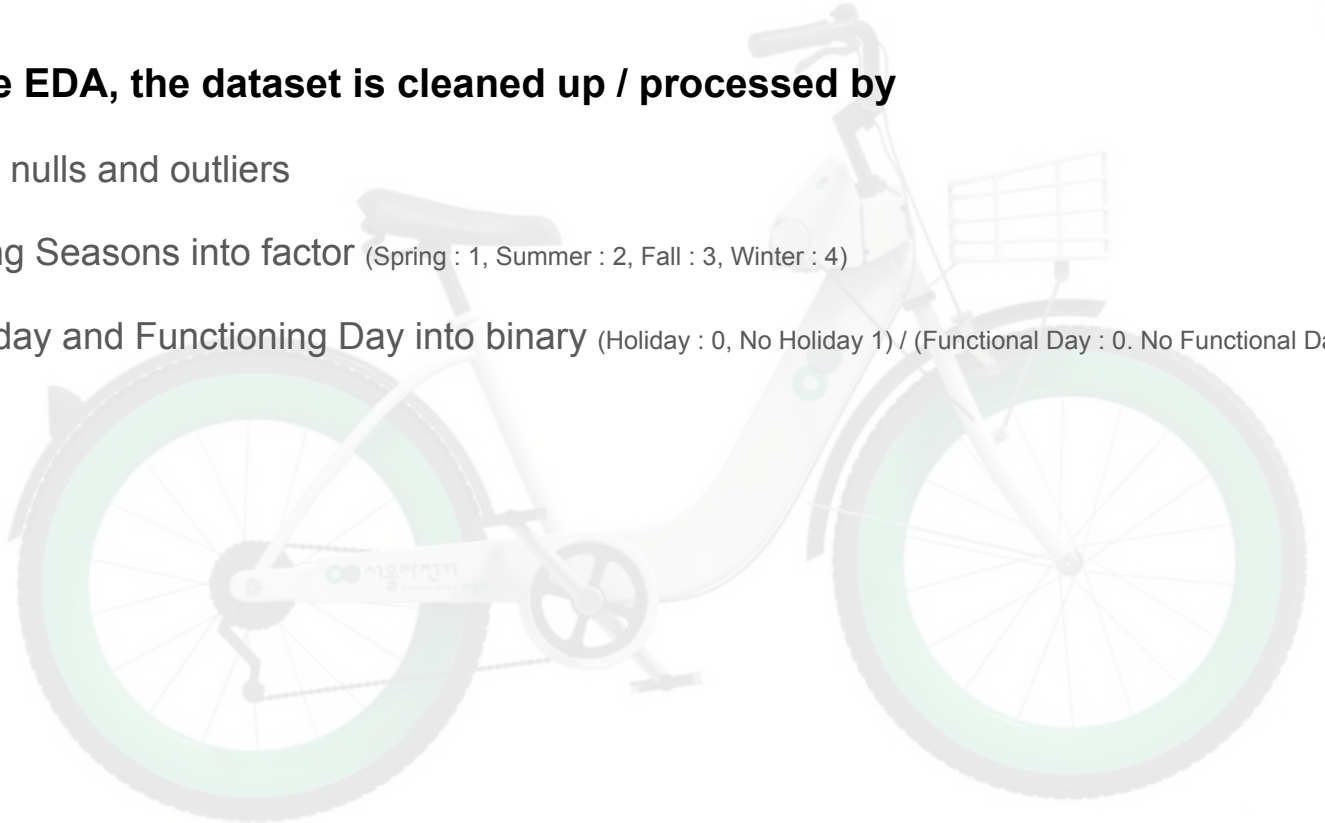




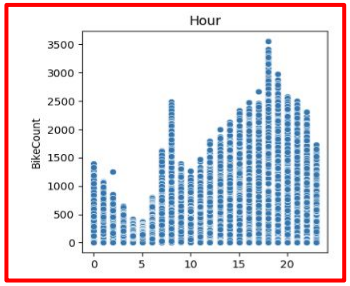
Data Cleaning

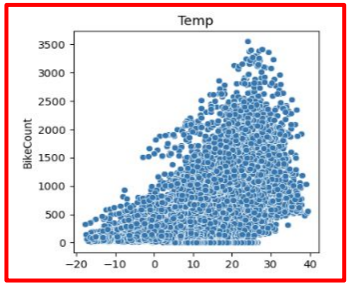
To prepare EDA, the dataset is cleaned up / processed by

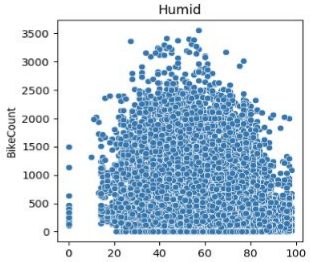
- dropping nulls and outliers
- converting Seasons into factor (Spring : 1, Summer : 2, Fall : 3, Winter : 4)
- and Holiday and Functioning Day into binary (Holiday : 0, No Holiday 1) / (Functional Day : 0. No Functional Day : 1)

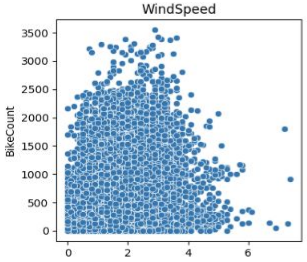


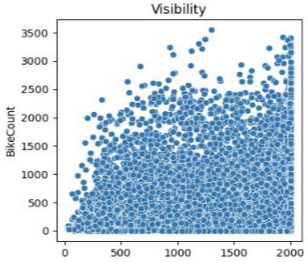
EDA

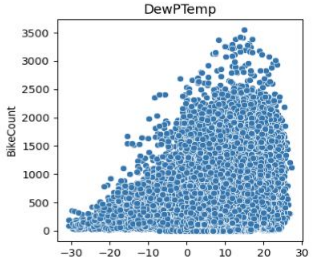

Hour
BikeCount

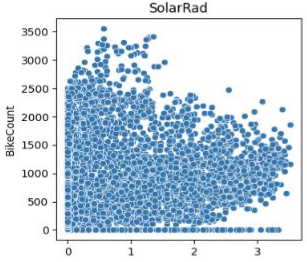

Temp
BikeCount

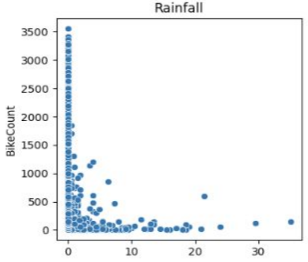

Humid
BikeCount

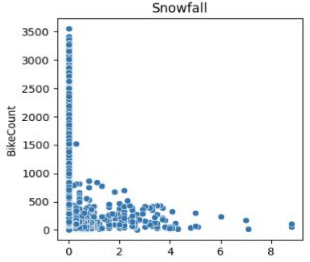

WindSpeed
BikeCount

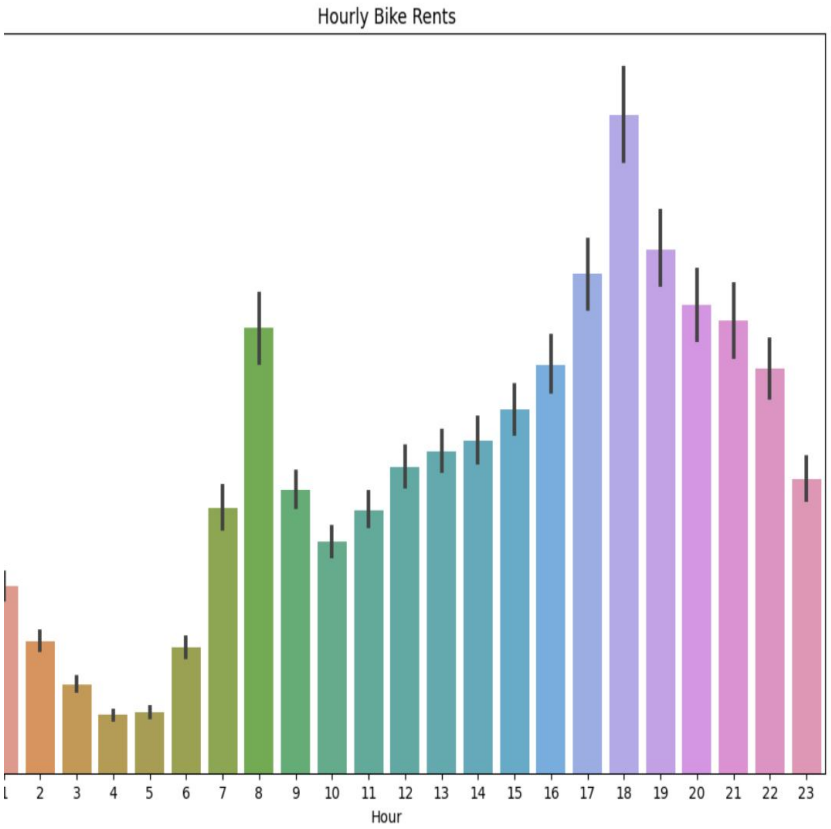

Visibility
BikeCount


DewPTemp
BikeCount

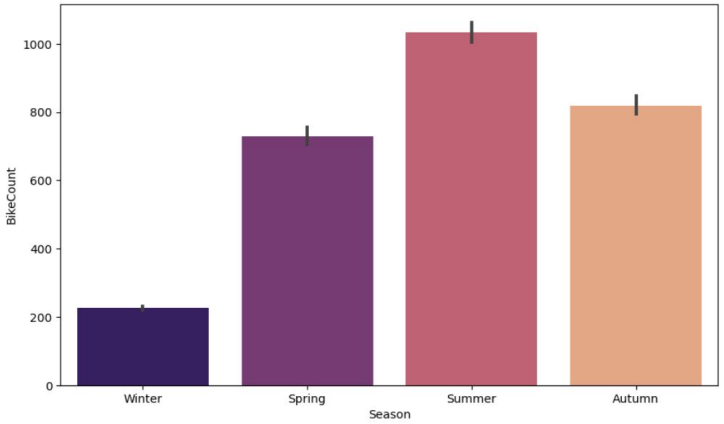

SolarRad
BikeCount

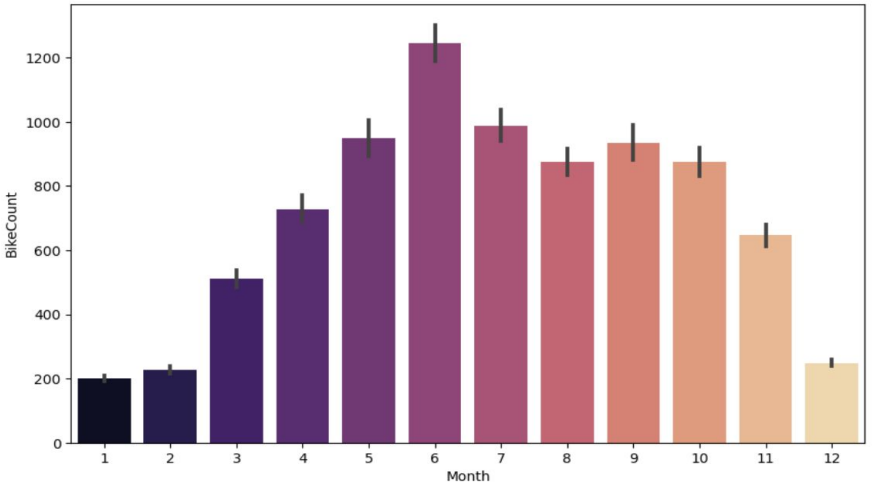

Rainfall
BikeCount

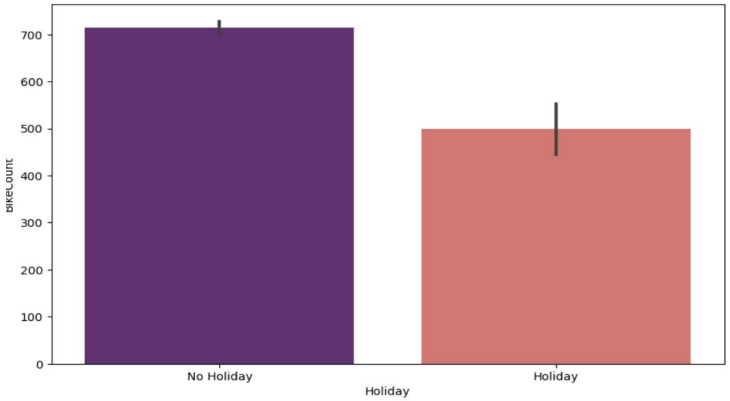

Snowfall
BikeCount

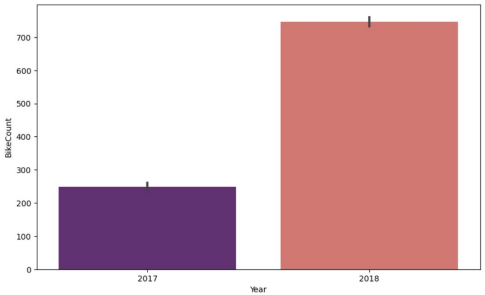

Hourly Bike Rents
Hour

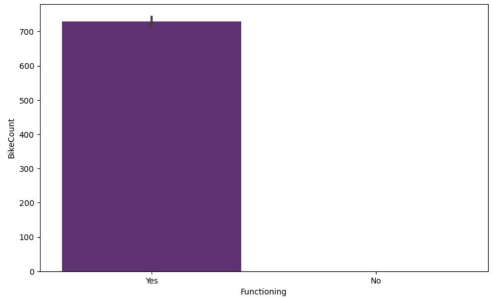
EDA


BikeCount
0
200
400
600
800
1000
Winter
Spring
Summer
Autumn
Season


BikeCount
0
200
400
600
800
1000
1200
1
2
3
4
5
6
7
8
9
10
11
12
Month


BikeCount
0
100
200
300
400
500
600
700
No Holiday
Holiday
Holiday


BikeCount
0
100
200
300
400
500
600
700
2017
2018
Year

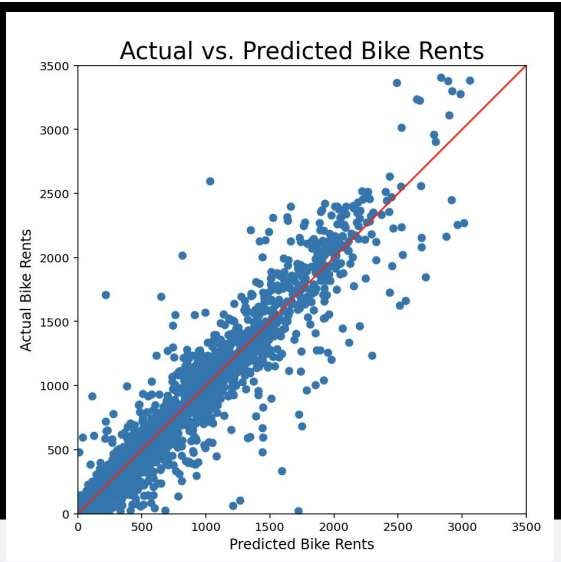

BikeCount
0
100
200
300
400
500
600
700
Yes
No
Functioning

ML Analyses

```
23SPring_CC_Final        CC_code.py           README.md                                SeoulBikeData.csv
'AWS + ML Project.pdf'   CC_code_xgboost.py  'Seoul Bike Sharing_CC Midterm Pj.pptx'
[ec2-user@ip-172-31-11-215 23SPring_CC_Final]$ python3 CC_code_xgboost.py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 17 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Date         8760 non-null   datetime64[ns]
 1   BikeCount    8760 non-null   int64
 2   Hour         8760 non-null   int64
 3   Temp         8760 non-null   float64
 4   Humid        8760 non-null   int64
 5   WindSpeed    8760 non-null   float64
 6   Visibility   8760 non-null   int64
 7   DewPTemp     8760 non-null   float64
 8   SolarRad     8760 non-null   float64
 9   Rainfall     8760 non-null   float64
 10  Snowfall     8760 non-null   float64
 11  Season       8760 non-null   object
 12  Holiday      8760 non-null   object
 13  Functioning  8760 non-null   object
 14  Day          8760 non-null   int32
 15  Month        8760 non-null   int32
 16  Year         8760 non-null   int32
dtypes: datetime64[ns](1), float64(6), int32(3), int64(4), object(3)
memory usage: 1.0+ MB
Linear Regression Mean R2: 0.5525259734396035
Random Forest Mean R2: 0.8788824781113111
K-Nearest Neighbors Mean R2: 0.5290049963959454
[ec2-user@ip-172-31-11-215 23SPring_CC_Final]$
```

```
Baseline XGBoost Mean R2 0.8847560843431364
```

```
Tuned XGBoost Mean R2 0.8905698632871903
```



Conclusion of the Project Implementation

- The AWS cloud services being used are VPC, public subnet, and EC2.
- Based on EDA and ML analysis were performed, and two hypotheses were supported.
- During rush hours, the bike rental count will be higher compared to non-rush hours.
- The data reveals a preference for increased bike usage in warmer seasons, especially summer. While our hypothesis might not be entirely accurate, temperature remains a crucial factor in bike usage patterns.
- XGBoost had the highest accuracy among the ML models used, and it was chosen as the final model.

Thank you!