Capstone Project – 2
Seoul Bike Sharing Demand Prediction

Puneet Suthar

What are we talking about ?

Content :

- Introduction to Bike rentals
- Problem Statement
- General overview of the dataset
- Data cleaning
- Exploratory Data Analysis
- Web application
- Conclusion

Introduction

Bike rentals have experienced a surge in popularity in recent years, with people using the service more frequently due to its relatively affordable rates and convenience of pick:up and drop:off at their own discretion. Ensuring the availability and accessibility of rental bikes to the public at the appropriate times reduces waiting time and ultimately provides a steady supply of bikes to the city. The objective of this project is to develop a machine learning model capable of forecasting the demand for rental bikes in Seoul.

Problem Statement

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

How our dataset Look like ?

General overview of the dataset

14 Columns

&

8760 Rows

General overview of the dataset

14 Columns

1.	Date	: year:month:day
2.	Rented Bike count	: Count of bikes rented at each hour
3.	Hour	: Hour of the day
4.	Temperature	: Temperature in Celsius
5.	Humidity	: %
6.	Wind Speed	: m/s
7.	Visibility	: 10m
8.	Dew point temperature	: Celsius
9.	Solar radiation	: MJ/m2
10.	Rainfall	: mm

General overview of the dataset

14 Columns

11. Snowfall : cm
12. Seasons : Winter, Spring, Summer, Autumn
13. Holiday : Holiday/No holiday
14. Functional Day : No(Non Functional Hours), Yes(Functional hours)

General overview of the dataset

- Some columns have very high No. of Zeros :-
  a. Solar Radiation (MJ/m2) – 4300
  b. Rainfall(mm) - 8232
  c. Snowfall (cm) – 8317

- In a day we have 24 hours and we have 365 days a year so 365 X 24 = 8760, which are the no.of lines in the given dataset.

- 'Seasons','Holiday', 'Functioning Day','Hour' are the categorical columns.

- No null values

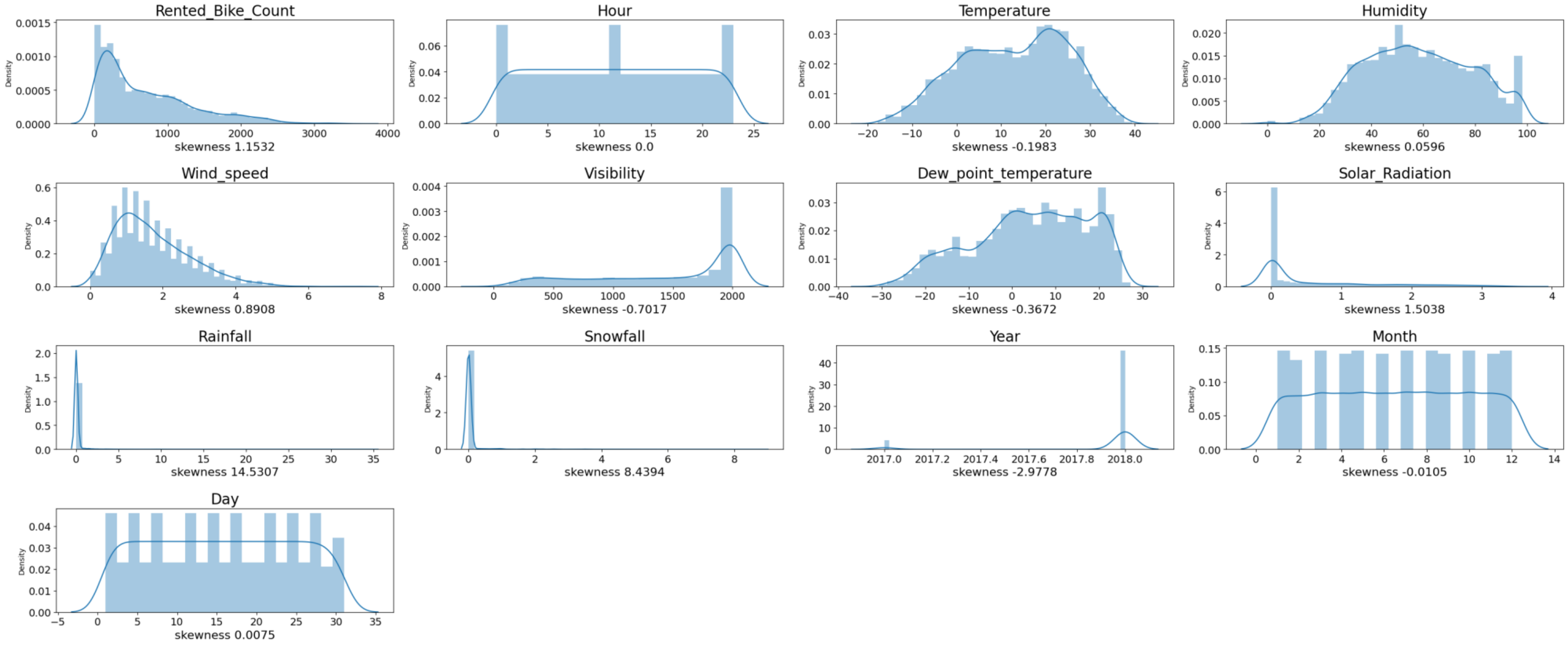
- No duplicated value.

EDA

(Exploratory Data Analysis)

Distribution plot & skewness

EDA
(Exploratory Data Analysis)



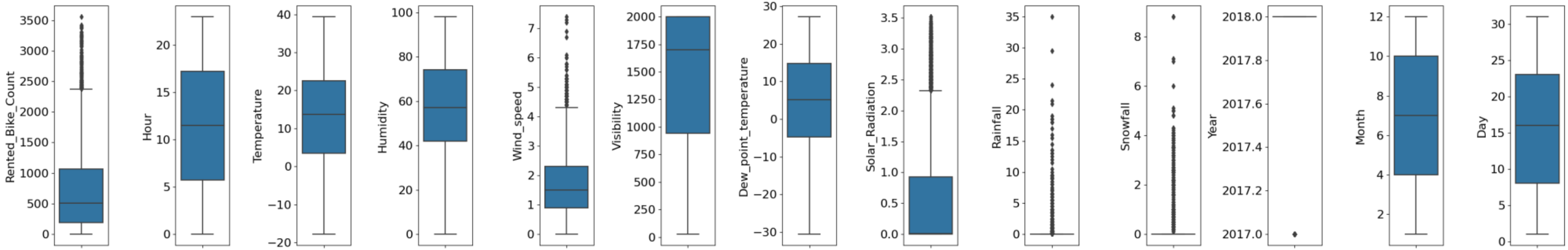
Outliers

EDA
(Exploratory Data Analysis)

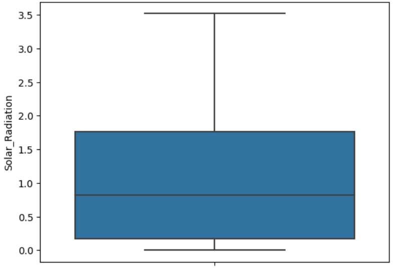
```
Total No. of rows in dataframe                    - 8760
Total rows containing outliers in data            - 1838
Percentage of total rows containing outliers in data - 20.98%
```

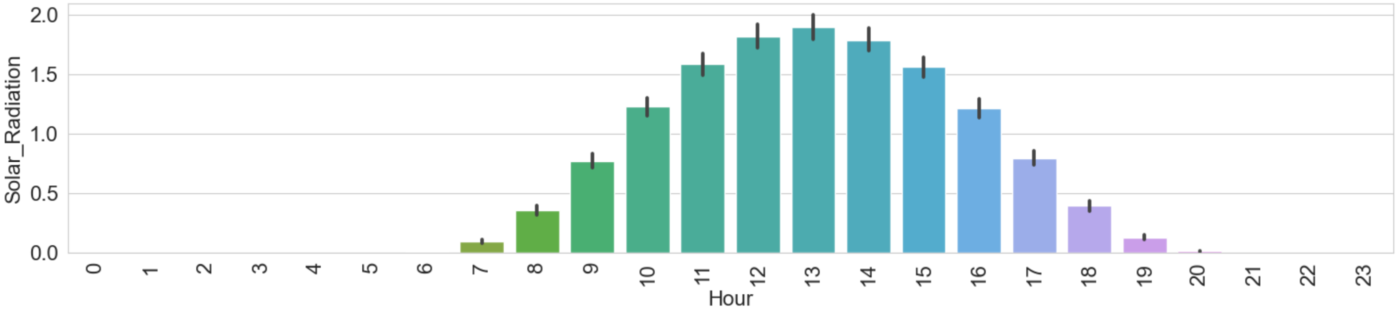
feature	upper_whisker	data_above_upper_whisker	lower_whisker	data_below_lower_whisker	outlier_per
Rented_Bike_Count	2376.625	158	-1120.375	0	1.80
Hour	34.500	0	-11.500	0	0.00
Temperature	51.000	0	-25.000	0	0.00
Humidity	122.000	0	-6.000	0	0.00
Wind_speed	4.400	161	-1.200	0	1.84
Visibility	3590.000	0	-650.000	0	0.00
Dew_point_temperature	44.050	0	-33.950	0	0.00
Solar_Radiation	2.325	641	-1.395	0	7.32
Rainfall	0.000	528	0.000	0	6.03
Snowfall	0.000	443	0.000	0	5.06

Before outlier removal



Box plot of solar radiation between 5 am to 7pm

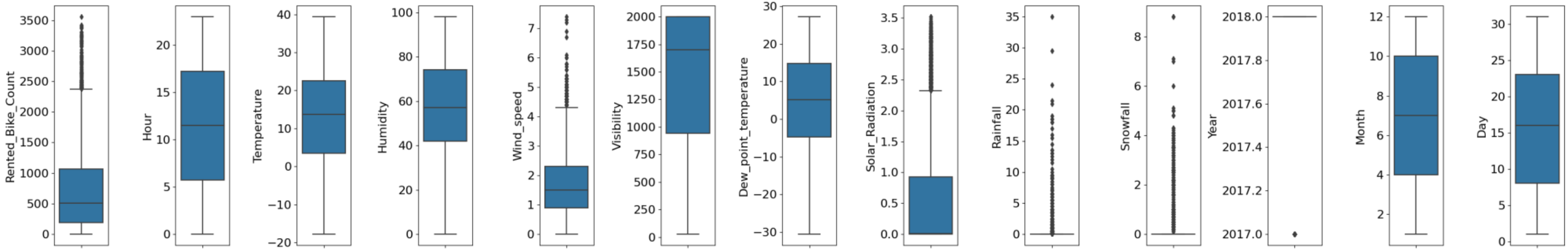




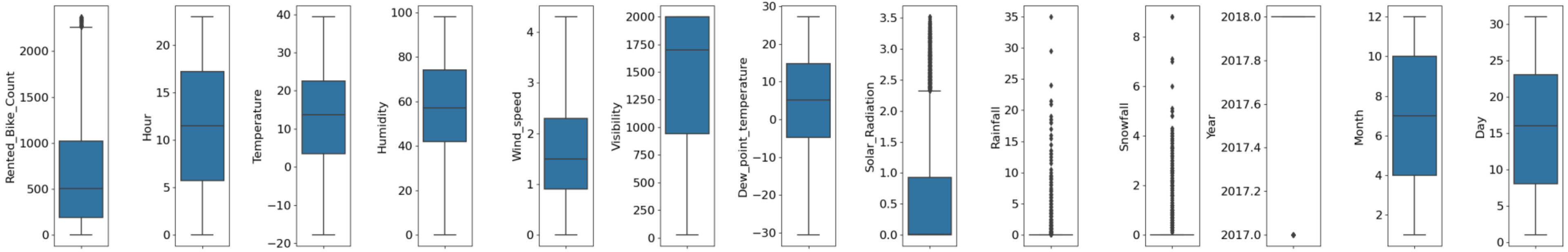
Box Plot

EDA
(Exploratory Data Analysis)

Before outlier removal



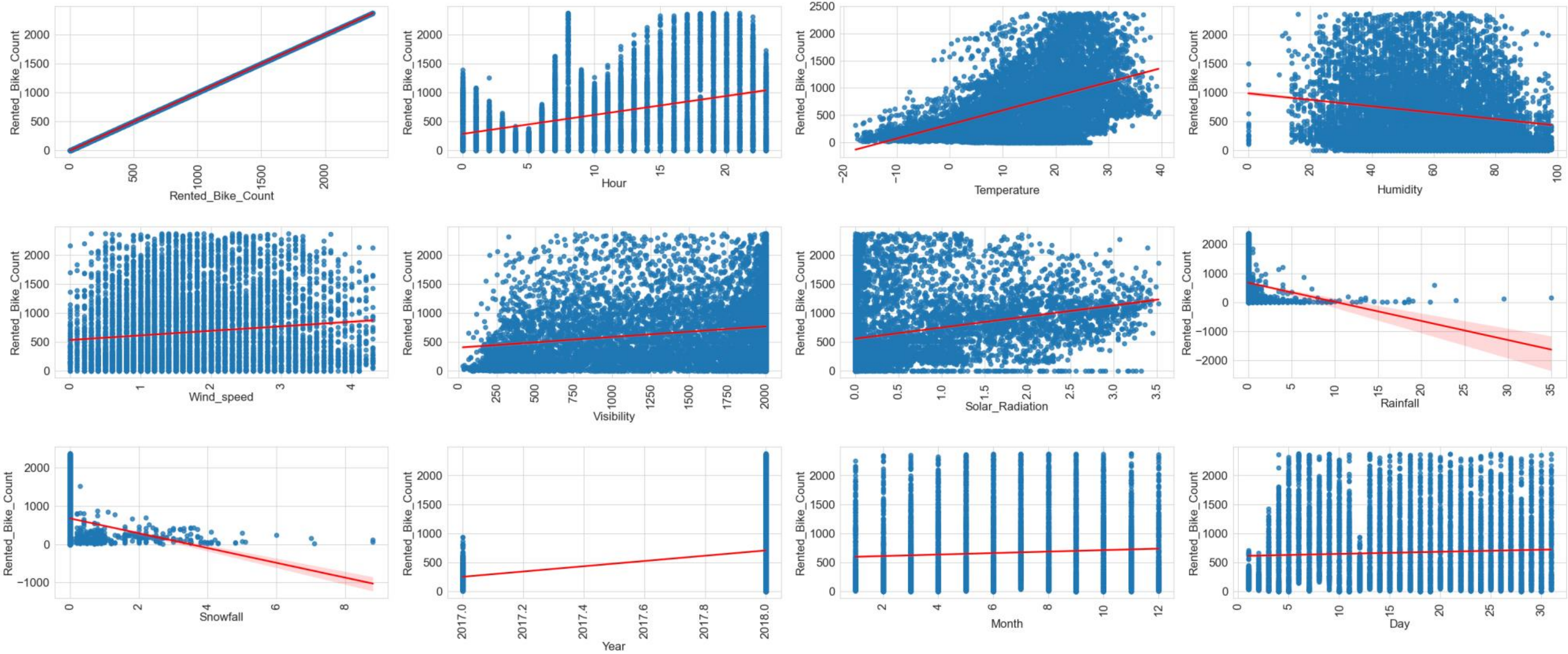
After replacing the outlier by median



Regression Plot

EDA
(Exploratory Data Analysis)

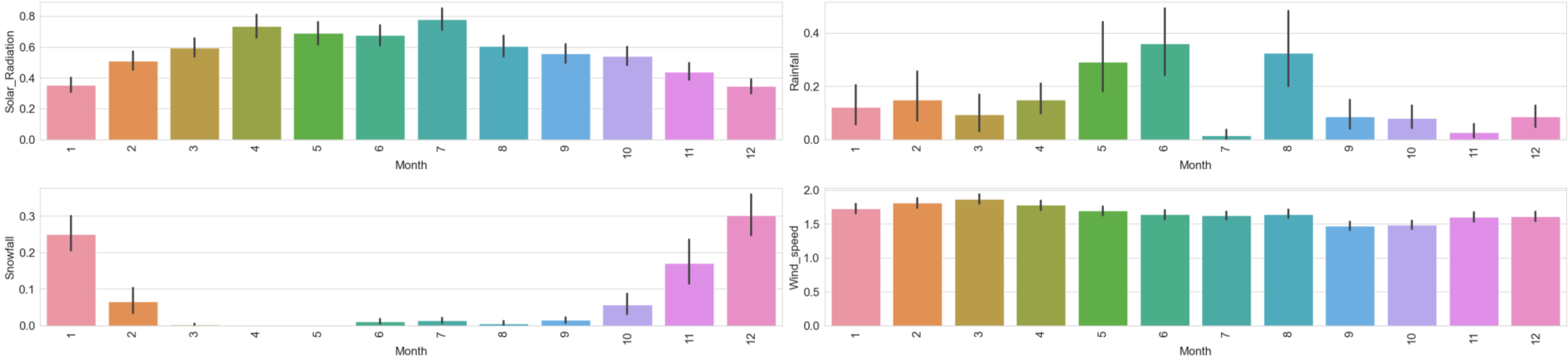
AI



Most of the features Doesn't show any linear relationship with target column
Only solar radiation and temperature shows a bit

Bar Plot

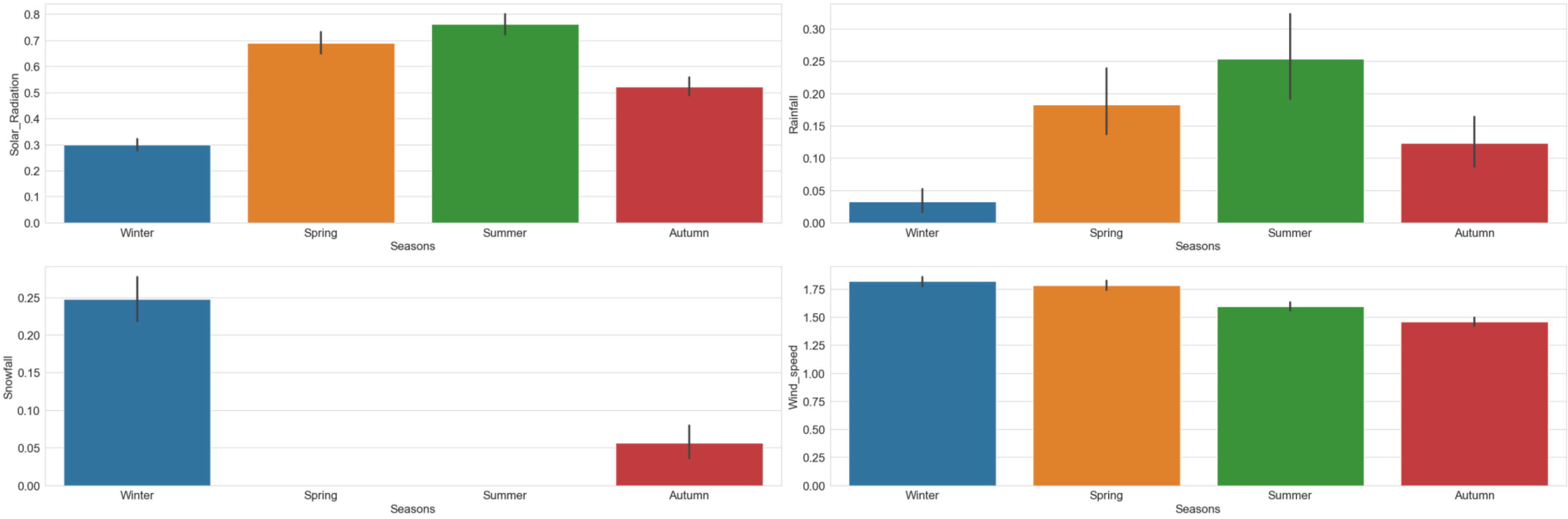
EDA
(Exploratory Data Analysis)



There is a huge impact on Solar_Radiation, Rainfall and Snowfall in different months

Bar Plot

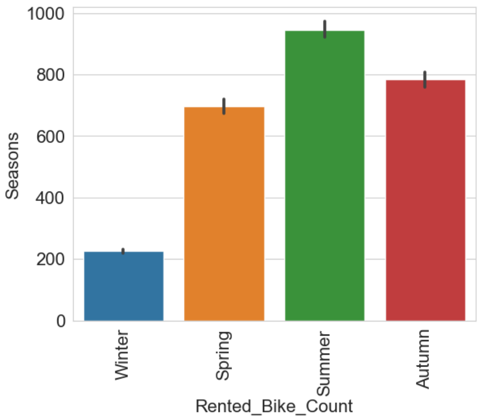
EDA
(Exploratory Data Analysis)

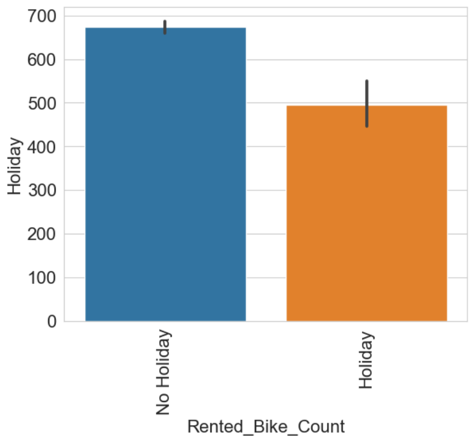


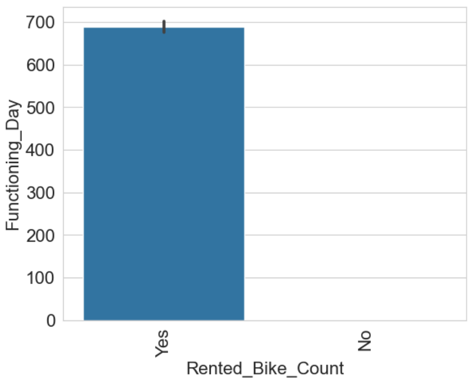
As we can see the snow fall and rainfall are highly related to the seasons, so these are the outliers, but the data is correct.

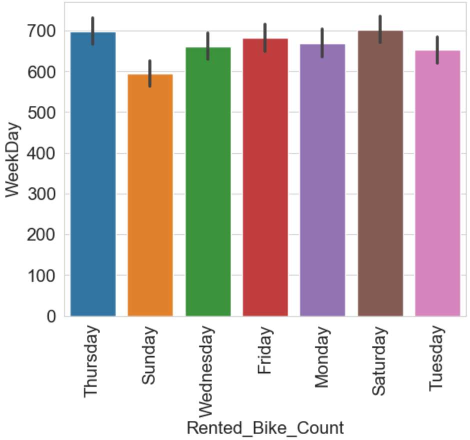
Bar Plot

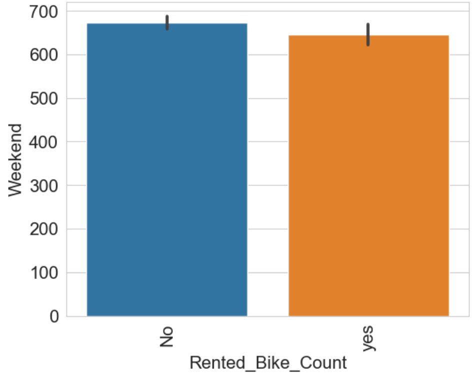
EDA
(Exploratory Data Analysis)







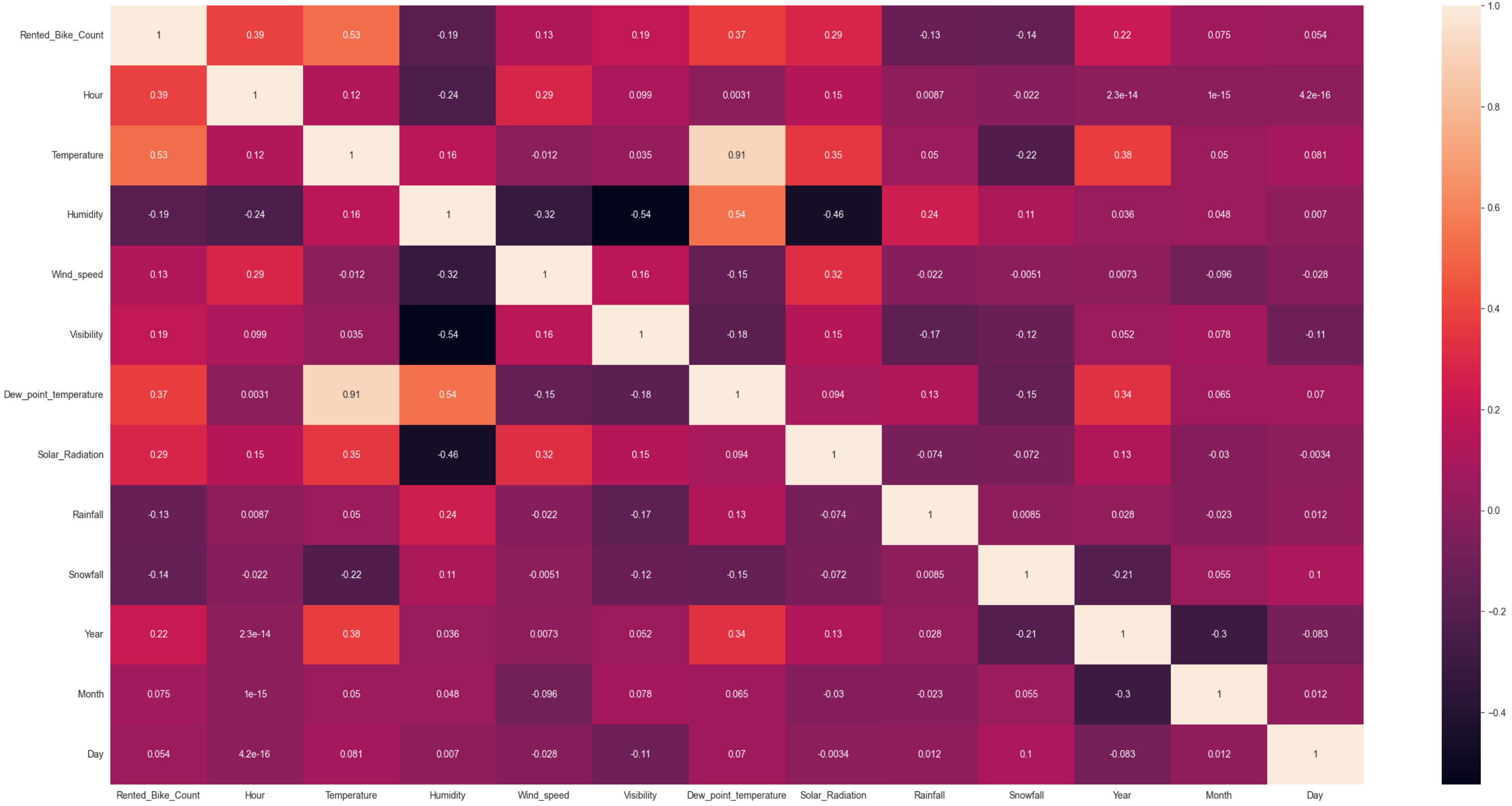




1. Bike demand is lower in winter.
2. Demand is higher in on non holiday.
3. Demand is lower on weekends, Sunday.

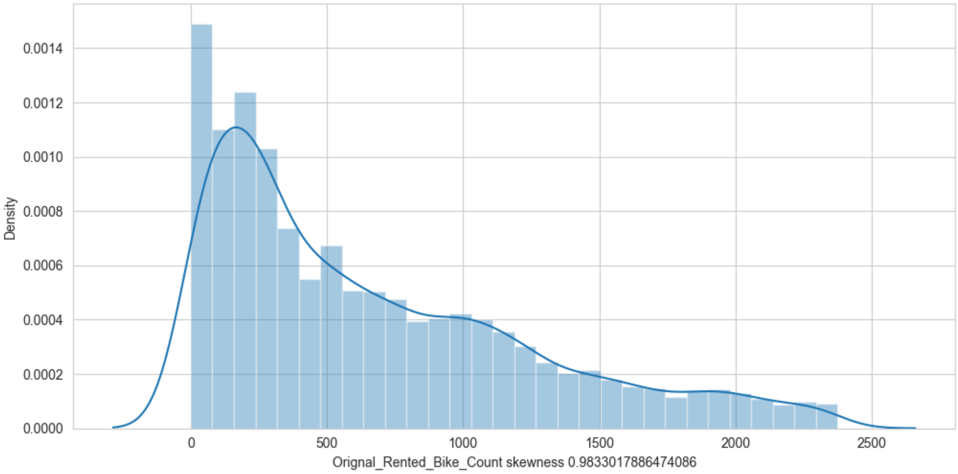
Correlation heatmap

EDA
(Exploratory Data Analysis)

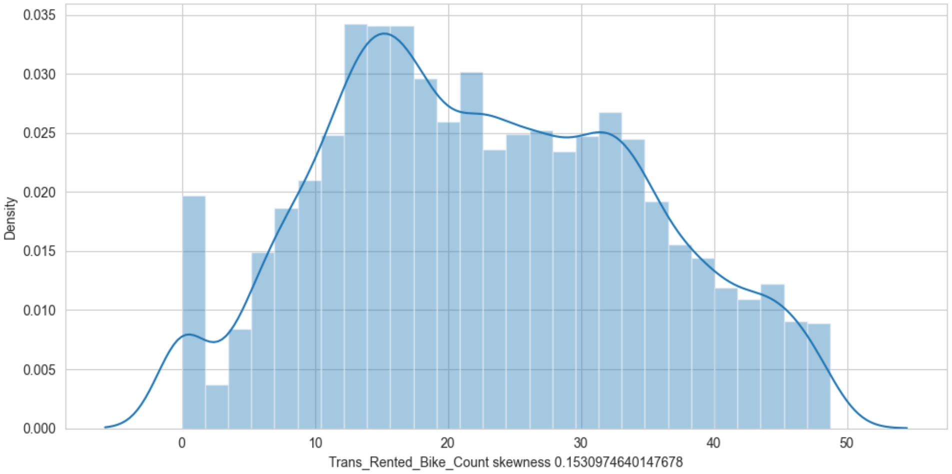


Handling skewed column

Applied square root transformation



Before



After

Model Building

Hyperparameter tuning

AI

```
models_list = [LR,La,R, En, DT, RF,GBR]
para_df = ModelSelection(X_data,y_data,models_list, model_hyperparameters,20,'neg_mean_absolute_error')
```

[70]

```
LinearRegression()
--------------------------------
Lasso()
--------------------------------
Ridge()
--------------------------------
ElasticNet()
--------------------------------
DecisionTreeRegressor()
--------------------------------
RandomForestRegressor()
--------------------------------
GradientBoostingRegressor()
--------------------------------
```

	model used	highest neg_mean_absolute_error score	best hyperparameters
0	LinearRegression()	-5.687872	{'copy_X': True, 'fit_intercept': True, 'positive': False}
1	Lasso()	-5.687823	{'alpha': 0.0001, 'copy_X': True, 'fit_intercept': True, 'positive': False}
2	Ridge()	-5.688480	{'alpha': 0.1, 'fit_intercept': True, 'positive': False, 'solver': 'sag'}
3	ElasticNet()	-5.687823	{'alpha': 0.0001, 'copy_X': True, 'fit_intercept': True, 'l1_ratio': 1.0, 'positive': False}
4	DecisionTreeRegressor()	-4.300701	{'criterion': 'friedman_mse', 'max_depth': 16}
5	RandomForestRegressor()	-3.391525	{'max_features': None, 'n_estimators': 150}
6	GradientBoostingRegressor()	-3.315588	{'max_depth': 10, 'min_samples_leaf': 70, 'n_estimators': 150, 'random_state': 4}

Model comparision

```python
models = [LR,La,R, En, DT, RF,GBR]
results = CrossValidation_model_comparision(models,X_data,y_data,10,scaler)
```

[80]

Python

...

Models Compaired

	Models	Parameters
0	LinearRegression	()
1	Lasso	(alpha=0.0001)
2	Ridge	(alpha=0.1, solver='sag')
3	ElasticNet	(alpha=0.0001, l1_ratio=1.0)
4	DecisionTreeRegressor	(criterion='friedman_mse', max_depth=16)
5	RandomForestRegressor	(max_features=None, n_estimators=150)
6	GradientBoostingRegressor	(max_depth=10, min_samples_leaf=70, n_estimators=150, random_state=4)

Common Metrics Report After 10 Fold Cross Validation

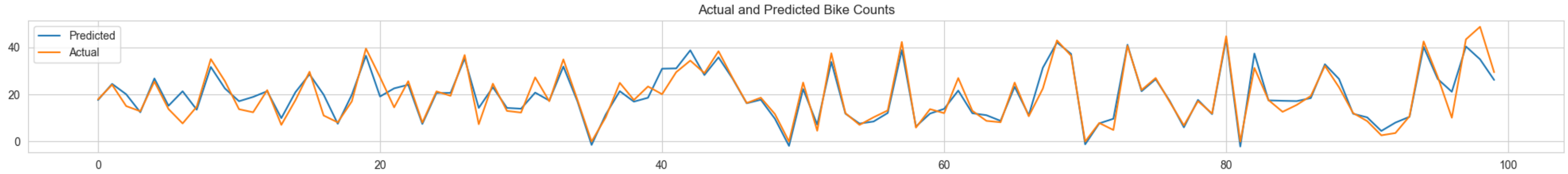
	Mean Absolute Error (MAE)			Mean Squared Error (MSE)			Root Mean Squared Error (RMSE)			r2_score		
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
LinearRegression	3.483169	8.144862	5.749263	19.148170	96.548506	55.128206	4.375862	9.825910	7.228983	-0.038652	0.726956	0.381590
Lasso	3.483713	8.144663	5.749254	19.150181	96.544308	55.125820	4.376092	9.825696	7.228818	-0.038761	0.726916	0.381629
Ridge	3.485228	8.145631	5.749882	19.162500	96.543674	55.124779	4.377499	9.825664	7.228817	-0.039429	0.726739	0.381608
ElasticNet	3.483713	8.144663	5.749254	19.150181	96.544308	55.125820	4.376092	9.825696	7.228818	-0.038761	0.726916	0.381629
DecisionTreeRegressor	3.212965	5.942551	4.683842	19.233462	73.825390	46.625189	4.385597	8.592170	6.716621	-0.326980	0.797834	0.434008
RandomForestRegressor	2.702963	5.432127	3.725001	12.175893	53.489935	27.430525	3.489397	7.313681	5.138280	0.141370	0.889928	0.660291
GradientBoostingRegressor	2.566331	5.448857	3.645469	10.698994	55.023106	26.249838	3.270932	7.417756	4.991712	0.390504	0.882203	0.695725

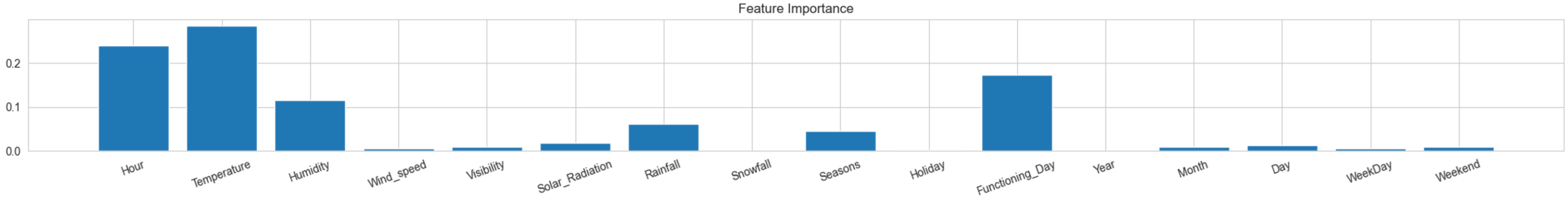
```
**Blue are highest
**Red are Lowest
```

Best model

```
GradientBoostingRegressor (max_depth          =10,
                           min_samples_leaf   =70,
                           n_estimators       =150,
                           random_state       =4)
```

	MSE	RMSE	MAE	Train_R2	Test_R2	Adjusted_R2
0	16.448202	4.055638	2.552573	0.916574	0.883052	0.881973





Web application

Rental bike count Prediction

Prediction Date

2023/02/24

Holiday or not

Yes

Functioning_Day or not

Yes

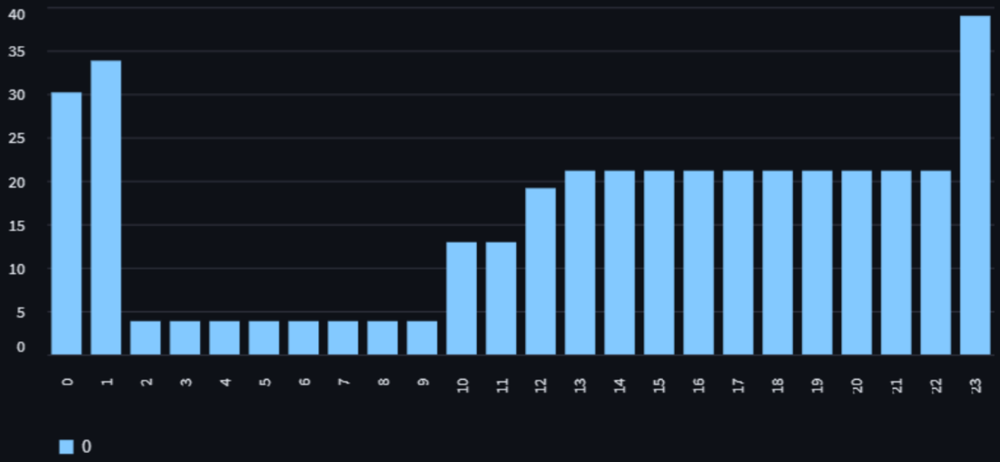
2023-02-24

	Hour	Temperature	Humidity	Wind_speed	Visibility	Solar_Radiation	Rainfall	Snowfall	Season
0	0	-1.3	93	3.7	24,140	29.5	0	0	Winter
1	1	0.4	80	3.3	24,140	210	0	0	Winter
2	2	2.7	64	5.8	24,140	376.5	0	0	Winter
3	3	4.7	50	7.1	24,140	482.6	0	0	Winter
4	4	5.6	42	9.2	24,140	533.4	0	0	Winter
5	5	5.9	38	10.5	24,140	521.5	0	0	Winter
6	6	6	31	11.4	24,140	452.9	0	0	Winter
7	7	5.5	30	11	24,140	328.2	0	0	Winter
8	8	4.7	29	9.6	24,140	168.7	0	0	Winter
9	9	3.6	30	8.3	24,140	36.1	0	0	Winter

Predict

2	2	2.7	64	5.8	24,140	376.5	0	0	Winter
3	3	4.7	50	7.1	24,140	482.6	0	0	Winter
4	4	5.6	42	9.2	24,140	533.4	0	0	Winter
5	5	5.9	38	10.5	24,140	521.5	0	0	Winter
6	6	6	31	11.4	24,140	452.9	0	0	Winter
7	7	5.5	30	11	24,140	328.2	0	0	Winter
8	8	4.7	29	9.6	24,140	168.7	0	0	Winter
9	9	3.6	30	8.3	24,140	36.1	0	0	Winter

Predict



Conclusion

Conclusion

We trained 7 models for our machine learning project to forecast bike rental demand based on weather conditions and other factors. We refined each model through hyperparameter tuning and found that the Gradient Boost model had the lowest RMSE, making it an excellent choice for accuracy-focused businesses. However, the decision tree model may be preferable for businesses that value model interpretability. Overall, our project developed accurate predictive models that can optimize rental operations and improve business outcomes.

Q & A