SUPERVISED ML REGRESSION CAPSTONE PROJECT

Prediction of demand for shared bikes

Contents

- Introduction
- Problem Statement
- Points for discussion
- Data Summary
- Analyzing Exploratory Data
- Overview of Modeling
- Important Feature
- Conclusion

Introduction

Motorcycles are rented out by a bike rental or bike hiring company for brief periods of time, typically a few hours. Most bike shops offer rentals as a sideline to their primary businesses of sales and service; however, some stores specialize in this service.

Like car rental businesses, bicycle rental businesses generally cater to those without access to vehicles, usually travelers and especially tourists.

For people who want to do a multi-day bike tour of a specific location but don't want to ship their own bikes, bike rental shops that rent by the day or week as well as by the hour are a great option.

For people who want to do a multi-d.

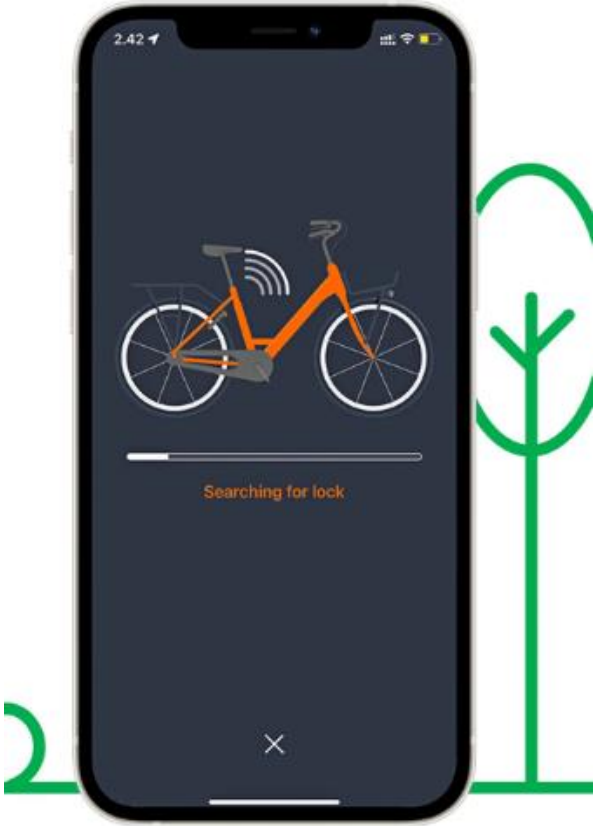
Problem Statement

Rental bikes are currently being introduced in many big locations to improve mobility comfort.

It is crucial to make the rental bikes accessible and available to the general public at the appropriate time since it reduces waiting.

Eventually, maintaining a steady supply of rental bikes for the city emerges as a top priority.

Predicting the number of bikes needed to maintain a steady supply of rental bikes at each hour's interval is essential.



Points for discussion

- Booking bikes throughout the year, on working days, weekends, and holidays.
- Comparing the number of rented bikes with the numerical data columns.
- Examining the linear relationship between the numerical data columns and the count of rented bikes.
- Climate Impact on Bike Sharing During Different Seasons.
- Correlation Map OR Heat Map
- Grid Search CV for Hyperparameter Tuning, Linear Regression Analysis, Lasso Regression Analysis
- XG boost, Random Forest Analysis, and Decision Tree Analysis
- Importance of the Feature

Data Summary

Imported Libraries

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

Descriptive Statistics

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this told mean, median, standard deviation

Missing Value Imputation

We will now check for missing values in our dataset. after checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

Graphical Representation

We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models driving the Demand for bikes.

Analyzing Exploratory Data

Date: Date in year-month-day format

Rented Bike Count: Count of bikes rented at each hour

Hour: Hour of the Day

Temperature: Temperature in Celsius

Humidity: Humidity in %

Windspeed: Speed of wind in m/s

Visibility (10m): Visibility

Dew point temperature: Dew Point Temp (Celsius)

Solar radiation: Radiation in MJ/m2

Rainfall: Rainfall (mm)

Snowfall: Snowfall (cm)

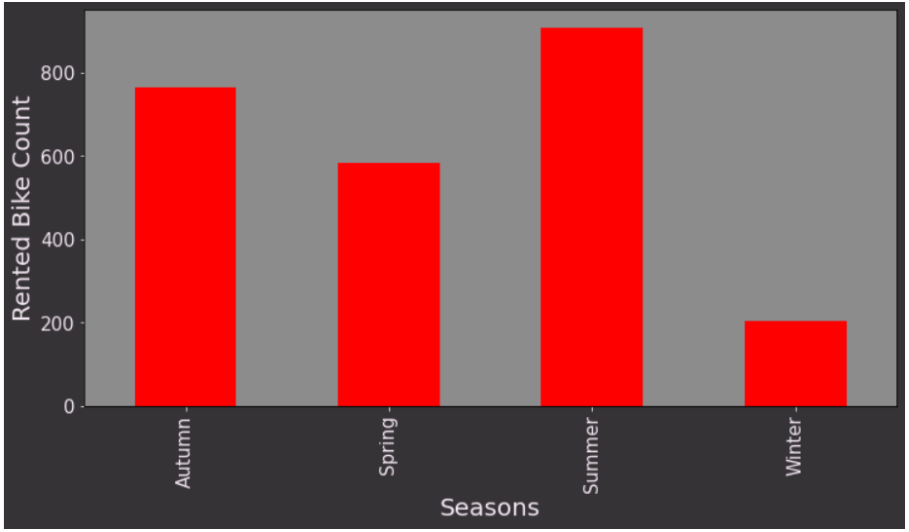
Seasons: Winter, Spring, Summer, Autumn

Holiday: Holiday/No holiday

Functioning Day: if the day is neither weekend, holiday than 1 else 0

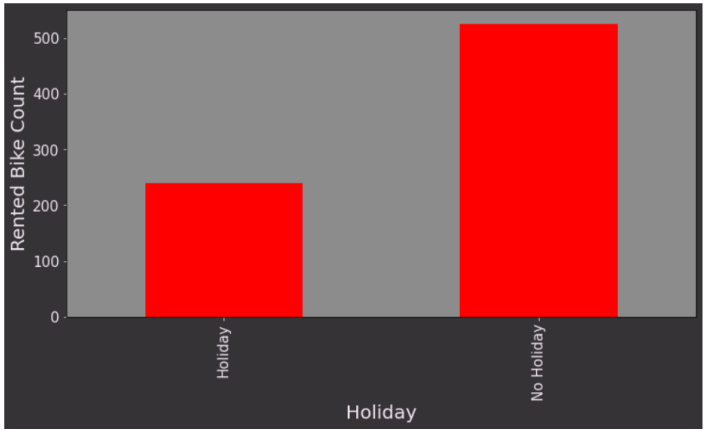
Bikes Rented per Season

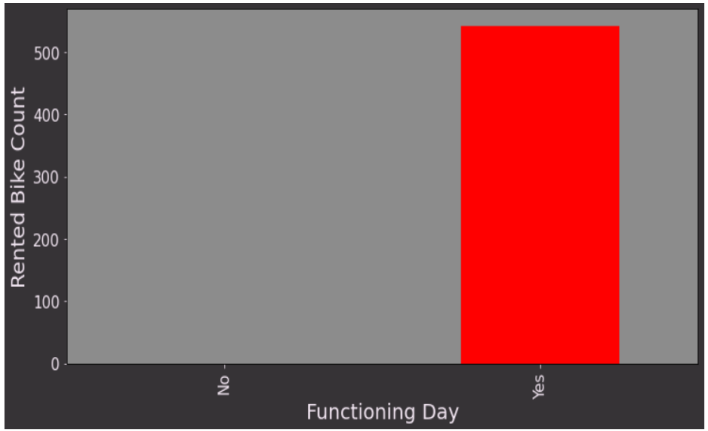
- The summer season saw the most bike rentals. 2.28 million bicycles were rented in total over the summer.
- The second-highest number of bikes were rented in the autumn, at roughly 1.79 million, and were then rented in the spring, at 1.6 million.
- The least frequent season to rent bikes seems to be winter. Only 487K bikes were hired during the winter.
- The severe winter weather in Seoul may contribute to the low winter demand for bicycles.



Bike Renting trend on holidays, Functioning days

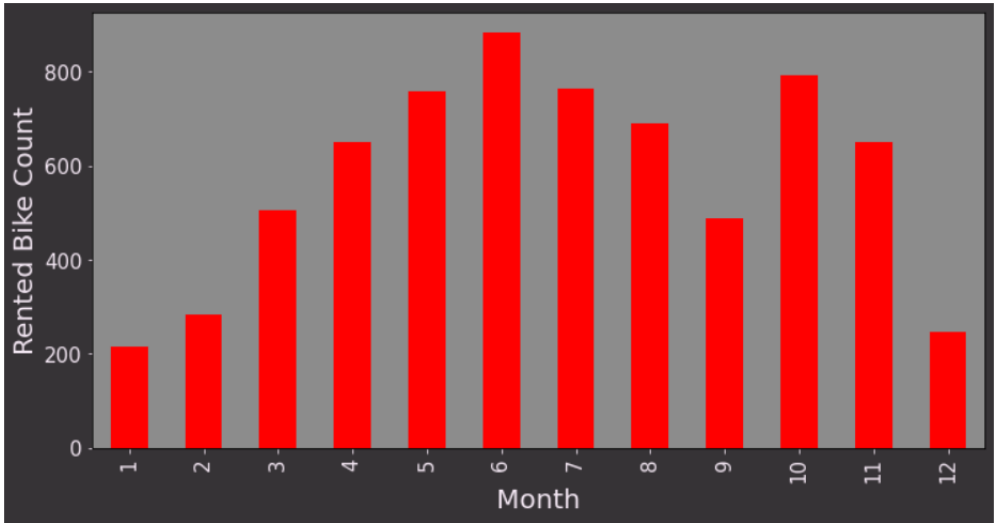
- **Compared to holidays, non-holiday days are when people prefer to ride their bikes.**
- **Only 215K bikes were rented during holidays, while 5.9 million bikes were rented on non-holiday days.**
- **It is safe to assume that the working class of Seoul makes up the bulk of the customers in the bike rental industry.**
- **Every bike that was hired was out on a working day.**





Bike Booking Monthly Trend

- **June** is the most preferred month for bike booking around **896K** bikes were rented in June.
- **July** and **May** are the second and third best.**734K** bikes were booked in **July**, and **707K** were booked in **May**.
- Demand for bikes was **least** in **Jan**, followed by **Feb** and **Dec**. **150K** bikes were rented in **Jan**, **151k** in **Feb**, and **185K** in **Dec.**

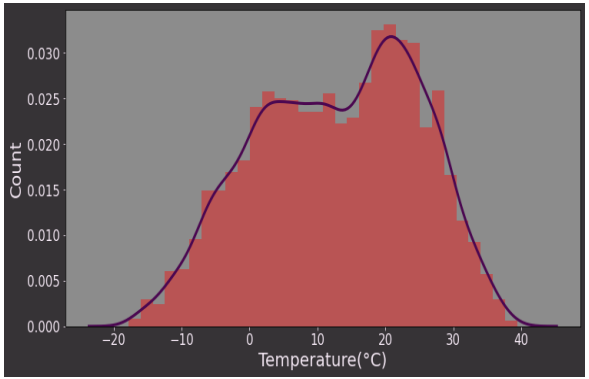


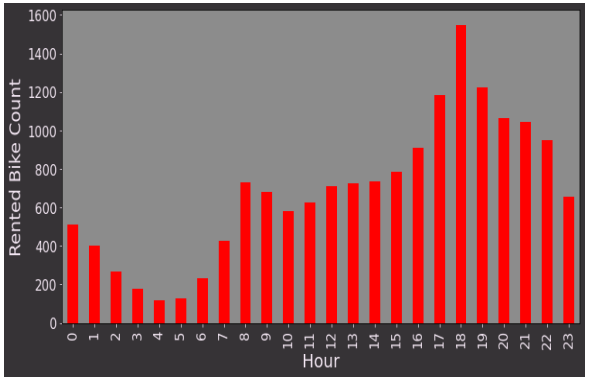
Rented Bike Count Against Numerical Data

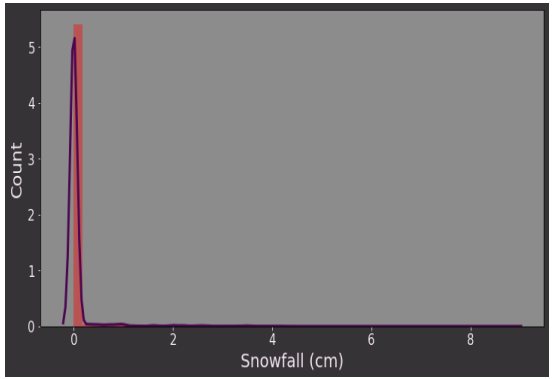
- **Most preferred** bike-sharing **temperature** is **20- 30** degrees Celsius. Bike renting is **minimal** when the **temperature** is **>35 or <5** degrees Celsius.

- Bike sharing is at its **peak between 4 pm to 8 pm**. Bike-sharing is at **least between 2 am to 6 am**, it **increases from 6 am onwards until 8 am**.

- **Snowfall** is **least favorable** for the bike renting Business.

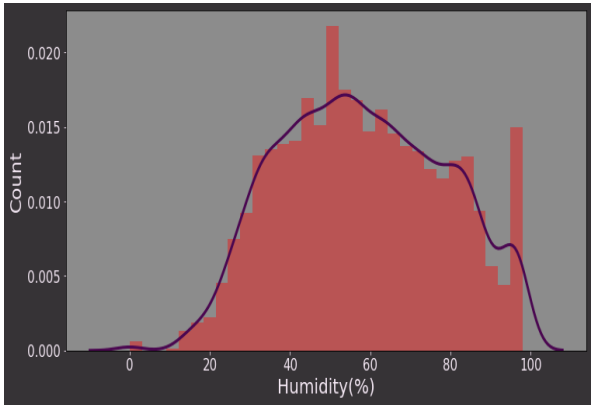


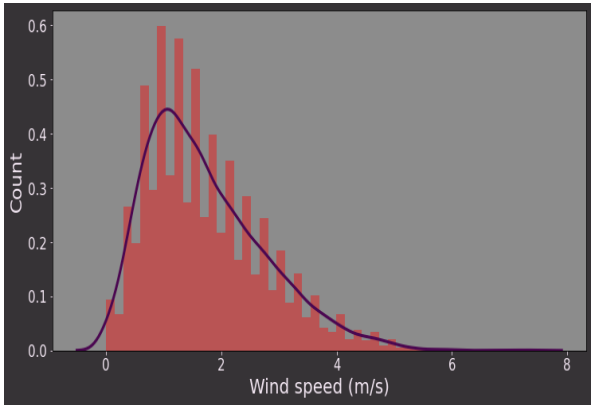


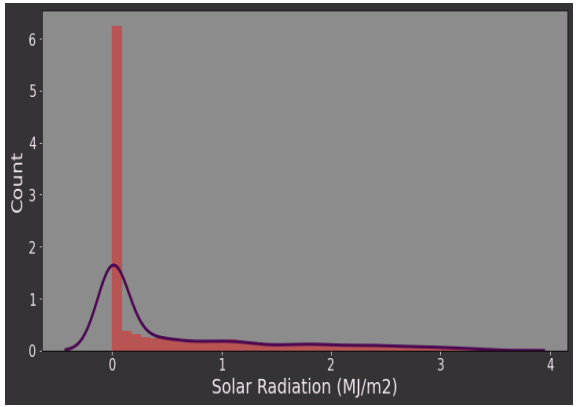


Rented Bike Count Against Numerical Data

- Bike renting is at its **peak** when the **humidity is 40%- 60%**. People avoid bikes when the climate is too humid or too dry.
- Favorable wind speed for Bike sharing is 1m/s -2 m/s as wind speed goes beyond 2m/s the count of bike-sharing starts dropping reaching minimal when the **speed > 5m/s.**
- Bike sharing is at its **peak** when the **radiation is minimal**.

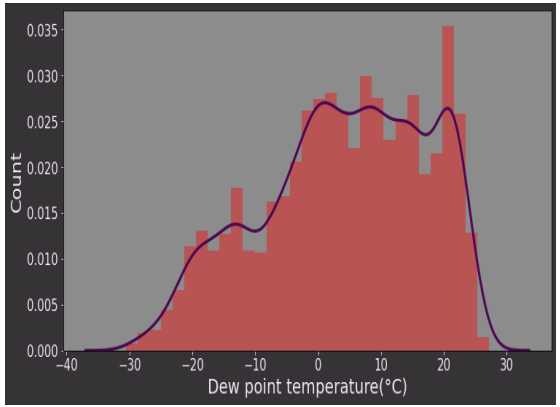


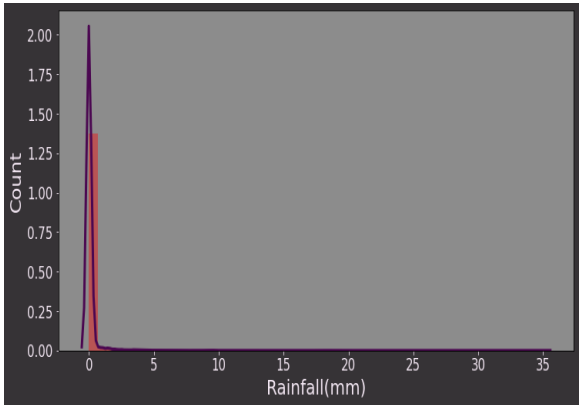


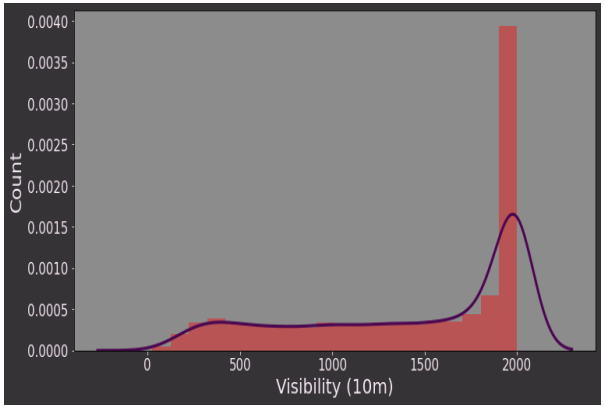


Rented Bike Count Against Numerical Data

- Dew point temperature between **5-25 Degrees** is **most favorable** for Bike sharing.
- Demand for bikes **dwindles** in case of **rainfall**.
- **Visibility** is an important factor for bike riders, bike sharing is at its **peak** when the **visibility is maximum**

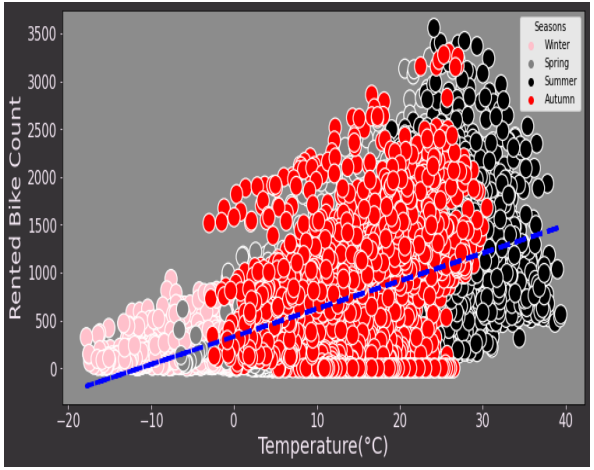


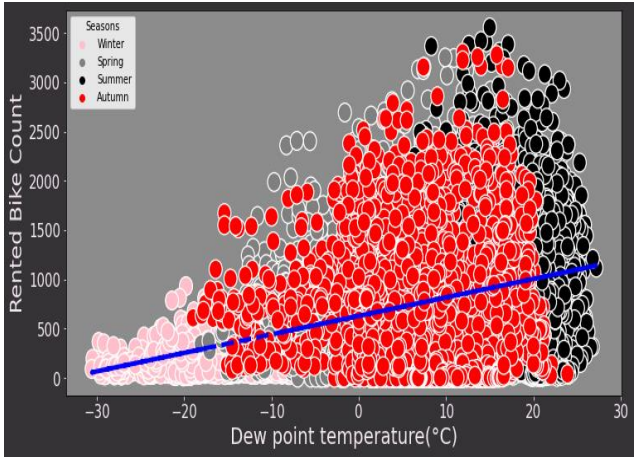


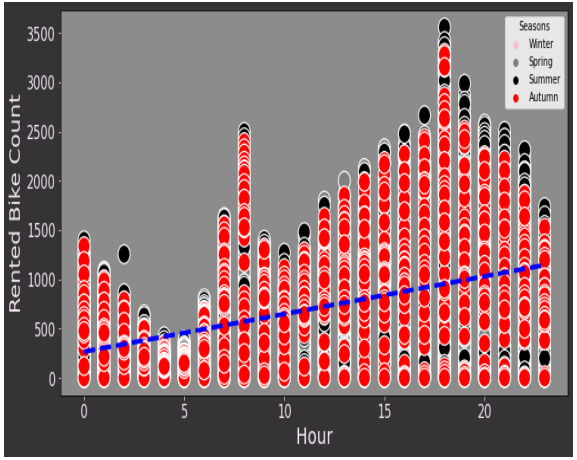


Co-relation: Rented bike count vs Temp, Dew point Temp, Hour

- Bike sharing is positively co-related to temperature and Dew point Temperature as the temperature approaches **30 degrees**.
- Though one thing to notice the positive co-relation is applicable only because the temperature in Seoul rarely crosses **40 Degrees**.
- Bike sharing count is positively co-related to hours as the Hours Progress from 0 (12 am) to 20 (8 pm) the bike-sharing count increases.

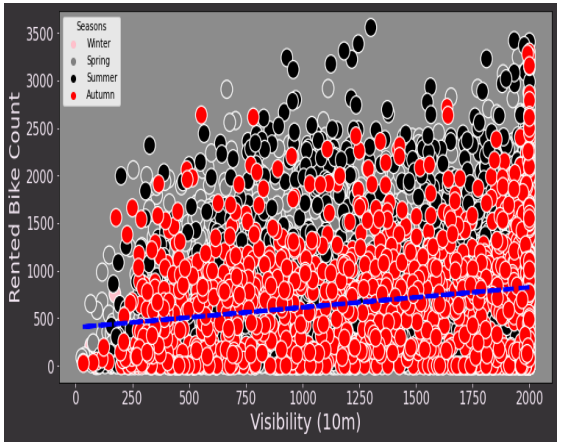


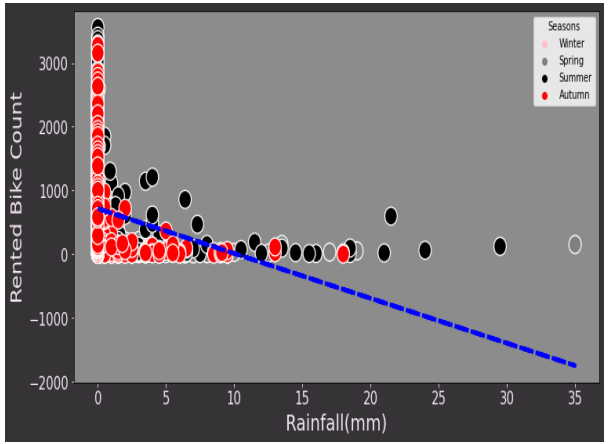


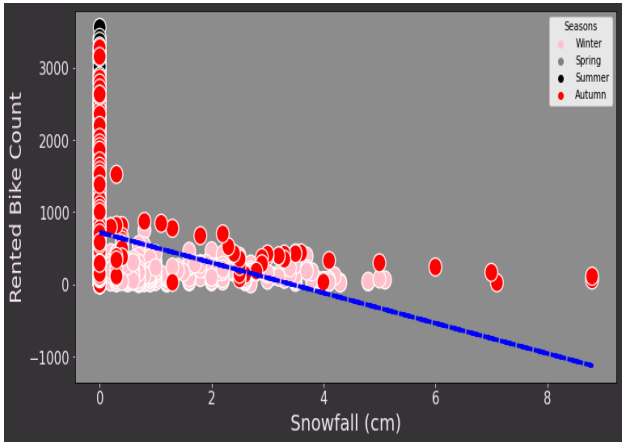


Co-relation: Rented bike count vs Visibility, Rainfall, Snowfall

- Additionally, bike reservations and visibility have a minor positive correlation.
- Snowfall and rain are negatively correlated with the number of bikes leased.

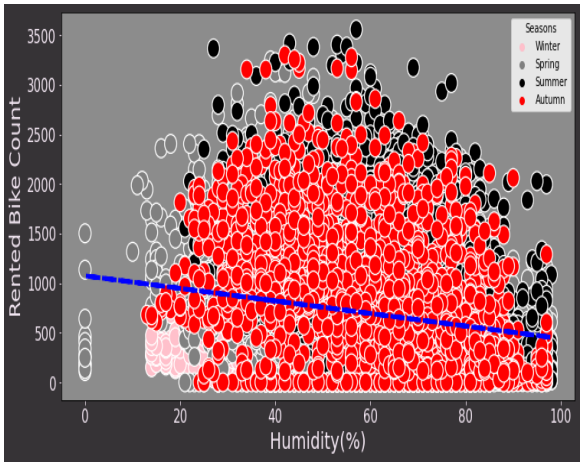


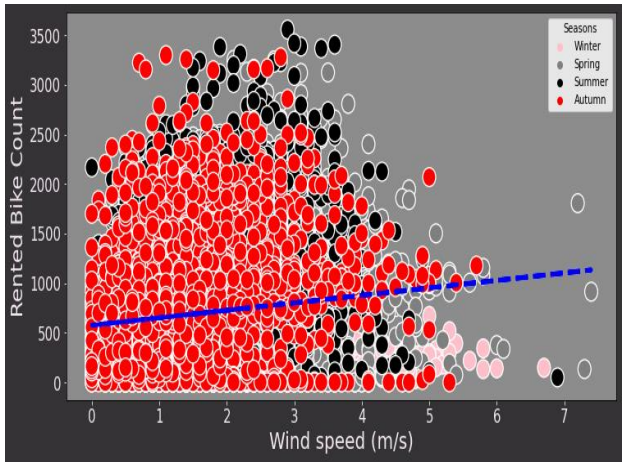


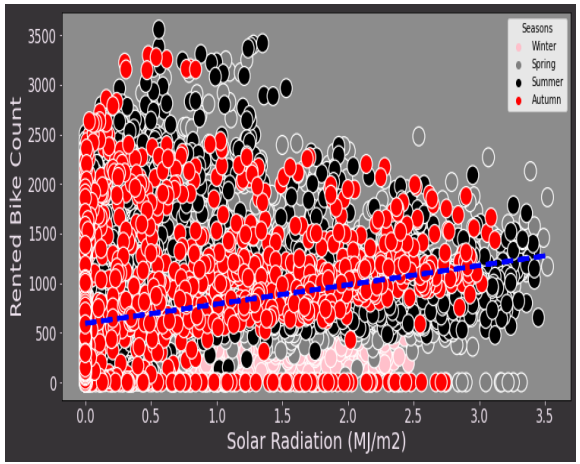


Co-relation: Rented bike count vs Humidity, Wind Speed, Radiation

- A minor inverse relationship exists between the number of shared bikes and humidity.
- Bike-sharing count and wind speed have a somewhat favourable relationship.







Correlation map

- The relationship between the number of rented bikes and the heat map's indicators of hour, temperature, dew point, and solar radiation is somewhat positive.
- Humidity, snowfall, and rain are all inversely correlated with the number of shared bicycles.
- Dew point temperature and temperature have a positive correlation.

Models List

In order to compare the final Root Mean Square Error and R2 score of these models, a total of twelve models were used in this project.

```
# List of models that we are going to use for this dataset
models = [
          ['LinearRegression: ',            LinearRegression()],
          ['Lasso: ',                       Lasso()],
          ['Ridge: ',                       Ridge()],
          ['KNeighborsRegressor: ',         neighbors.KNeighborsRegressor()],
          ['SVR:' ,                         SVR(kernel='rbf')],
          ['DecisionTree ',                 DecisionTreeRegressor(random_state=42)],
          ['RandomForest ',                 RandomForestRegressor(random_state=42)],
          ['ExtraTreeRegressor :',          ExtraTreesRegressor(random_state=42)],
          ['GradientBoostingRegressor: ',   GradientBoostingRegressor(random_state=42)],
          ['XGBRegressor: ',                xgb.XGBRegressor(random_state=42)] ,
          ['Light-GBM: ',                   lightgbm.LGBMRegressor(num_leaves=41, n_estimators=200,random_state=42)],
]
```

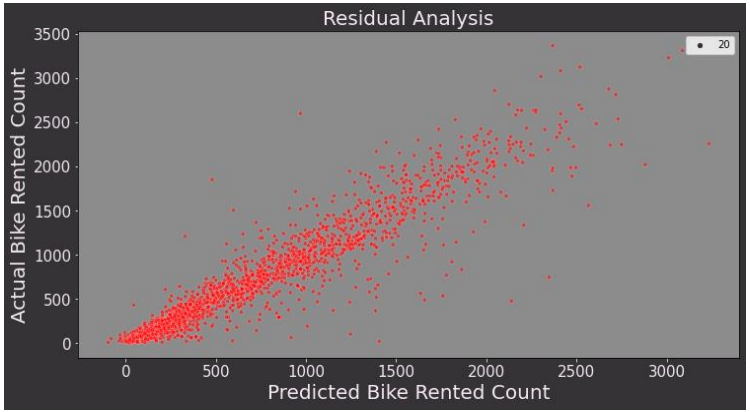
Result

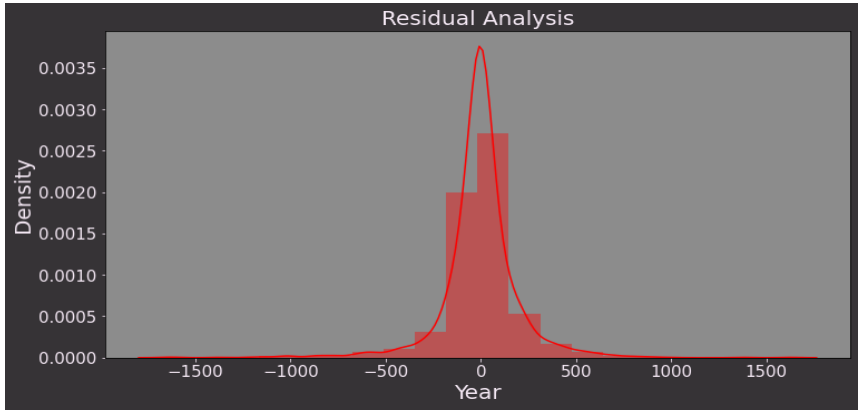
Out of the twelve models, Lightgbm, ExtraTreeRegressor, and RandomForestRegressor provide the highest R2 score and the lowest root mean square error on the test set, as is evident.

	Name	Train_Time	Train_R2_Score	Train_RMSE_Score	Test_R2_Score	Test_RMSE_Score
0	LinearRegression:	7.152557e-07	0.539158	439.103836	0.555399	419.033651
1	Lasso:	7.152557e-07	0.534718	441.213908	0.552075	420.596945
2	Ridge:	1.430511e-06	0.538927	439.213876	0.555500	418.986169
3	KNeighborsRegressor:	7.152557e-07	0.863473	239.001167	0.800327	280.816708
4	SVR:	1.192093e-06	0.263863	554.971348	0.289494	529.721550
5	DecisionTree	1.192093e-06	1.000000	0.000000	0.743764	318.114797
6	RandomForest	9.536743e-07	0.983017	84.294528	0.873439	223.570369
7	ExtraTreeRegressor :	4.768372e-07	1.000000	0.000000	0.878614	218.951240
8	GradientBoostingRegressor:	4.768372e-07	0.867964	235.037831	0.848602	244.525043
9	XGBRegressor:	1.192093e-06	0.866833	236.042007	0.848415	244.676517
10	Light-GBM:	9.536743e-07	0.974173	103.950667	0.888898	209.471137

Model-1 Extra Tree Regressor

- On the Test set, Extra tree considerably reduces the RMSE.
- The plot below shows that, in comparison to other models, the predicted and actual values are substantially closer.
- RMSE: 218.95 and an R-score of 0.878.For the Extra tree, residual values are dramatically decreased.
- Most Residual values are around 0 on the KDE plot, which is significantly leaner.





Hyperparameter Tuning of Extra Tree Regressor

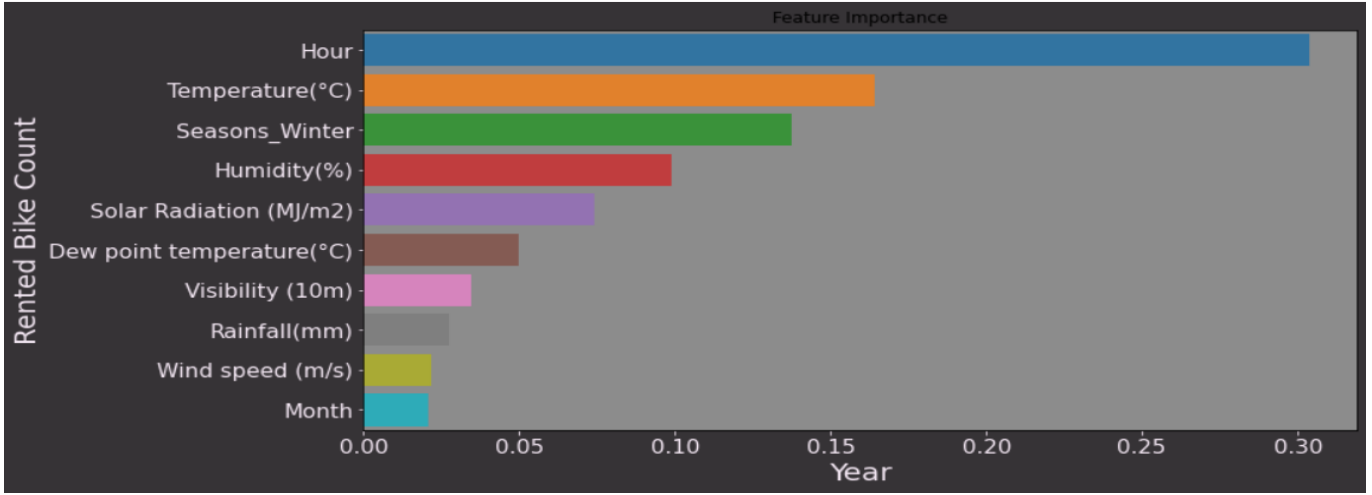
- We used the random search cv approach for hyperparameter tuning to identify the ideal hyperparameters for our model.
- Only R-Score.885 and RMSE: 212.78 increased after hyperparameter tuning by 0.76%.

```
# Create the random grid
random_grid ={'bootstrap': [True, False],
              'max_depth': [70, 80, 90, 100, None],
              'max_features': ['auto', 'sqrt'],
              'min_samples_leaf': [1, 2, 4],
              'min_samples_split': [2, 5, 10],
              'n_estimators': [800, 1000]}

RF = ExtraTreesRegressor(n_jobs=-1, random_state=42)

# Random search of parameters, using 3 fold cross validation,
random_search = RandomizedSearchCV(estimator = RF,
                                   param_distributions = random_grid,
                                   n_iter = 100, cv = 3, verbose=2)
# Fit the random search model
random_search.fit(train_inputs, train_targets)
```

```
print('Improvement of {:0.2f}%.'.format( 100 * (0.8853 - 0.87863) / 0.8773))
```

```
Improvement of 0.76%.
```

Feature Importance



- The significance of the characteristics on our number of rented bikes is depicted in the accompanying graph.
- One of the main elements influencing the demand for motorcycles is the temperature and time of day.
- Other factors that affect demand for motorcycles include solar radiation, humidity, rainfall, and whether it is a working day.

Model 2 - Light GBM

A gradient boosting system called Light GBM makes use of tree-based learning techniques.
It has the following benefits and is distributed and effective by design:

1. Increased efficiency and training pace.
2. Reduced memory use.
3. More precise.
4. Assistance with distributed, parallel, and GPU learning.
5. Capable of managing massive amounts of data.

On the Test set, Light GBM greatly reduces the RMSE. The plot below shows that, in comparison to other models, the predicted and actual values are substantially closer. RMSE: 218.95 and an R-score of 0.878.

Hyperparameter Tuning of Light GBM

- We used the Optuna Library for hyperparameter tuning to identify the ideal hyperparameters for our model.Optuna uses a Bayesian approach to search.

```
import optuna
from optuna import Trial, visualization
from optuna.samplers import TPESampler
def objective(trial,data=data):


    param = {
        'metric': 'rmse',
        'random_state': 42,
        'n_estimators': 10000,
        'reg_alpha': trial.suggest_loguniform('reg_alpha', 1e-3, 10.0),
        'reg_lambda': trial.suggest_loguniform('reg_lambda', 1e-3, 10.0),
        'colsample_bytree': trial.suggest_categorical('colsample_bytree', [0.3,0.4,0.5,0.6,0.7,0.8,0.9, 1.0]),
        'subsample': trial.suggest_categorical('subsample', [0.4,0.5,0.6,0.7,0.8,1.0]),
        'learning_rate': trial.suggest_categorical('learning_rate', [0.006,0.008,0.01,0.014,0.017,0.02]),
        'max_depth': trial.suggest_categorical('max_depth', [10,20,100]),
        'num_leaves' : trial.suggest_int('num_leaves', 1, 1000),
        'min_child_samples': trial.suggest_int('min_child_samples', 1, 300),
        'cat_smooth' : trial.suggest_int('min_data_per_groups', 1, 100)
    }

    model = lightgbm.LGBMRegressor(**param)
    model.fit(train_inputs,train_targets,eval_set=[(val_inputs,val_targets)],early_stopping_rounds=100,verbose=False)

    preds = model.predict(val_inputs)
    rmse = metrics.mean_squared_error(val_targets, preds,squared=False)

    return rmse
```

```
print('Improvement of {:0.2f}%.'.format( 100 * (0.9037 - 0.8878) / 0.8878))
```

```
Improvement of 1.79%.
```

Conclusion

- Summer was the season with the highest number of bike rentals, followed by Autumn, Spring, and Winter. The busiest months for renting bikes are May through July, and December through February are the least popular months.

- The working class makes up the vast majority of customers in the bike rental industry. The EDA analysis shows that in Seoul, the demand for bikes is higher during the weekdays while people are at work.

- The best conditions are found in the evenings, between 4 and 8 p.m., when the temperature is between 20 and 30 degrees and the humidity is between 40 and 60 percent.

- **Major elements influencing the demand for rental bikes include temperature, daytime, solar radiation, humidity, and hour of the day.**

- The linear model's prediction was very low since there was a very weak linear relationship between the feature and the labels. Light GBM's best predictions have a r2 score of 0.894 and an RMSE of 203.91.

Thank You