

COMPUTER ENGINEERING DEPARTMENT

CSE464 Case Study

**INTRODUCTION TO DATA SCIENCE & BIG DATA ANALYTICS
(AUTUMN SEMESTER 2024)**

**MEHMET ÖNAL
20200702081**

1. Short Story of Your Startup's Business Idea

The startup idea pertains to Seoul's bike-sharing system in line with fluctuating demands for rentals. Seoul being the metropolitan hub has massive population density and faces traffic jams as a major challenge to urban growth. The bike-sharing system represents a non-polluting, reasonably affordable, and highly convenient way of travel for short distances. It serves directly towards the aims of the city for sustainable urban growth with less congestion and healthy environment.

This startup deploys advanced data analytics to enhance the operational efficiency of the bike-sharing system. It analyzes historical data for patterns and trends in bike rentals, including seasonal changes, weather conditions, and time-specific demand fluctuations. The main goal is to offer a more reliable and user-friendly service by ensuring that bicycles are available when and where they are most needed.

Besides improving the management of resources, this further helps in contributing to a smart city concept. All this integrates predictive modeling for real-time insights into an integrated transportation experience for all citizens and tourists. The project does not solve only logistic challenges but also contributes to a better quality of urban living, meeting the modern trends of smart mobility solutions.

2. Problem Summary/Definition

Among the critical challenges which most bike-sharing systems around the world are facing, including Seoul, is to efficiently manage demand fluctuations. Various factors influence these fluctuations, such as weather conditions, time of day, seasons, and public holidays. For example, demands are usually high during warm months and rush hours but drop tremendously during adverse weather conditions or on public holidays. This variability complicates resource allocation and can lead to inefficiencies in operations.

Not making enough data-driven decisions is further exacerbating the problem. Without proper demand forecasting, bikes may be overstocked in low-demand areas or vice-versa. This leads to an imbalance whereby resources are either underutilized or the users' needs are not met, impacting customer satisfaction. In a metropolitan city like Seoul, relying highly on transport systems, such inefficiencies may lead to massive disruptions.

It is a problem that requires an understanding of the pattern of demand through data analysis and predictive modeling. Effective demand forecasting will make the system proactive in resource allocation for operational efficiency and customer satisfaction. Thus, the integration of advanced analytics into the bike-sharing system is quite important to overcome these challenges and achieve its full potential.

3. Justification of the Problem Using SWOT Analysis

Strengths:

The bike-sharing system has a few intrinsic strengths, such as being ecologically friendly and inexpensive to use. The system contributes less to urban traffic congestion and helps in leading a healthy lifestyle through bicycling. It also aligns with Seoul's sustainability objectives by connecting seamlessly with public transportation within the city for traveling short distances.

Weaknesses:

Despite the advantages, variability in demand ails the system. Seasonal changes, weather conditions, and public holidays make a huge difference in the behavior of the users of the system, hence causing inefficiency. For instance, the peak demand during summer months is in sharp contrast to the low utilization rates in winter. Besides, the system relies heavily on operational staff for manual redistribution of bikes, which may be resource-intensive and error-prone without accurate forecasting.

Opportunities:

With increased awareness environmentally and sustainable urban mobility, therefore, environmental awareness grows day by day. Thus, both factors are very strong for bike-sharing system growth. Using data analytics and machine learning to predict and understand user behaviors internal to the system so it could do the resource allocation more effective and also can be further developed with real-time monitoring and/or mobile application in creating values for users in attracting customers.

Threats:

The system also faces competition from other types of micro-mobility, like electric scooters and ride-sharing, all very convenient. The infrastructural challenges, such as weather conditions or lack of bike lanes, may further drive the customers away. Besides that, high investment in advanced technologies creates some financial risks.

This SWOT analysis highlights the critical need for data-driven solutions to address weaknesses and threats while capitalizing on strengths and opportunities to ensure the system's success.

4. Alternative Solutions/Recommendations/Decisions

It is further recommended that these challenges need to be addressed by descriptive analytics combined with predictive modeling, as the former uncovers the historical demand patterns whereas the latter predicts future demand for proactive decisions. This will make this system both reactive to currently surfacing trends and ready to handle the future ones.

One possible solution is the dynamic resource allocation strategy based on demand predictions. For example, bicycles can be redistributed in real time to high-demand areas during rush hours or peak seasons. This strategy minimizes the instances of bike shortages or excess inventory, hence improving operational efficiency. Moreover, weather-sensitive adjustments can be made, such as increasing bike availability on sunny days when demand is likely to be higher.

Other recommendations involve the increase in user engagement through technology. For example, mobile applications can display the real-time availability of bikes and reserve them in advance. Gamification, such as rewards for frequent usage, can also be used to incentivize users. Integrating customer feedback into the system would also help operators to constantly improve service quality and promptly address user concerns.

5. Implementation Strategy/Plan

The solution implementation process included the following steps:

1. Data Collection and Preprocessing

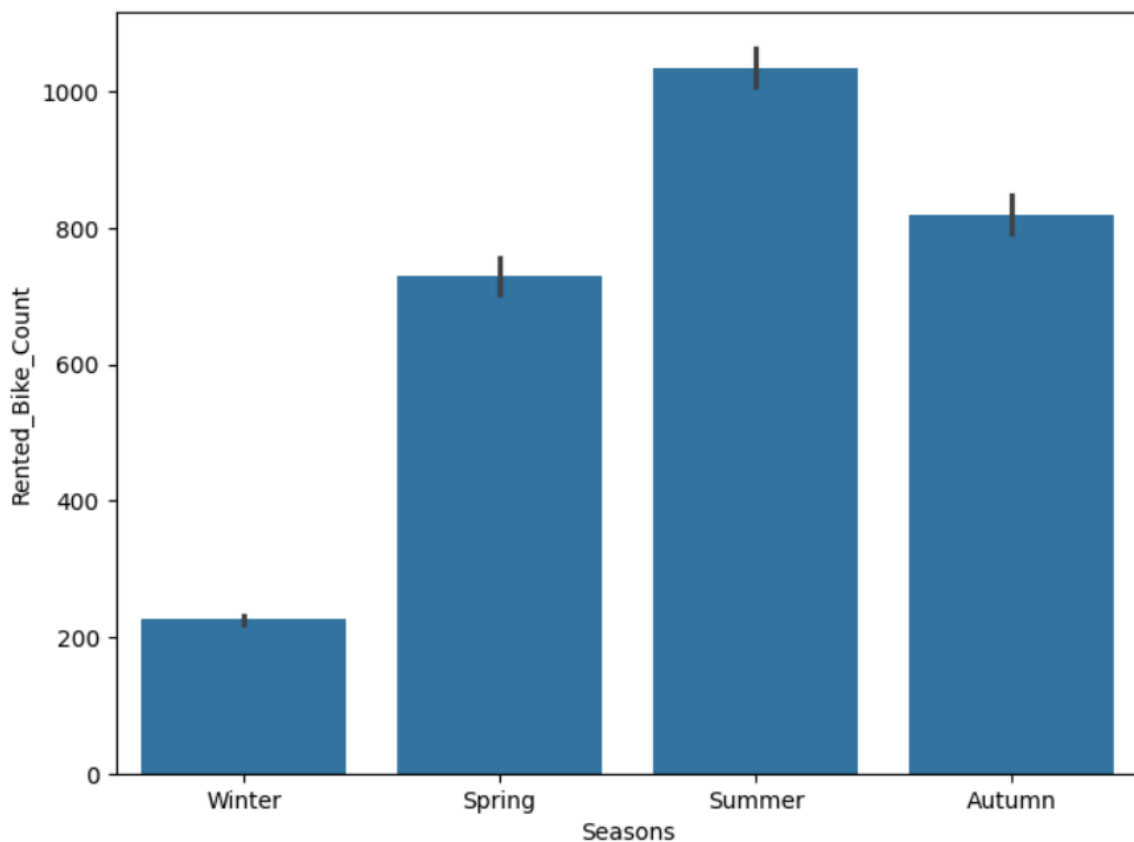
- The Seoul bike-sharing system dataset was loaded, and missing or erroneous values were identified and cleaned using appropriate methods.
- Categorical variables (e.g., seasons) were converted into numerical values.
- Features such as month, year, and day were extracted from the date column.
- Based on histogram analyses, skewness was observed in some numerical data, and transformations such as log1p, square root, and cube root were applied to approximate normal distribution.

2. Descriptive Analysis

- Basic statistics of the dataset were calculated.
- Seasonal, monthly, and hourly rental trends were visualized.
- The effects of weather variables on rental demand were analyzed.
- The distribution of data was examined using histograms.

Seasonal Analysis

To visualize and understand how bike rental demand changes across seasons, a bar chart showing the number of rented bicycles per season was used. This chart revealed seasonal trends, with significantly higher rental numbers during summer and the lowest demand in winter. This analysis highlights the considerable impact of seasonal changes on bike rental demand, suggesting that the system should be planned based on seasonal demand variations.



COMPUTER ENGINEERING DEPARTMENT

Histograms

Histograms were used to examine the distributions of numerical variables. These visualizations illustrated the frequency distributions, skewness, and spread of the data. Specifically, the target variable, "Number of Rented Bicycles," exhibited a right-skewed distribution, suggesting the potential need for data transformation during modeling.

Hourly Analysis

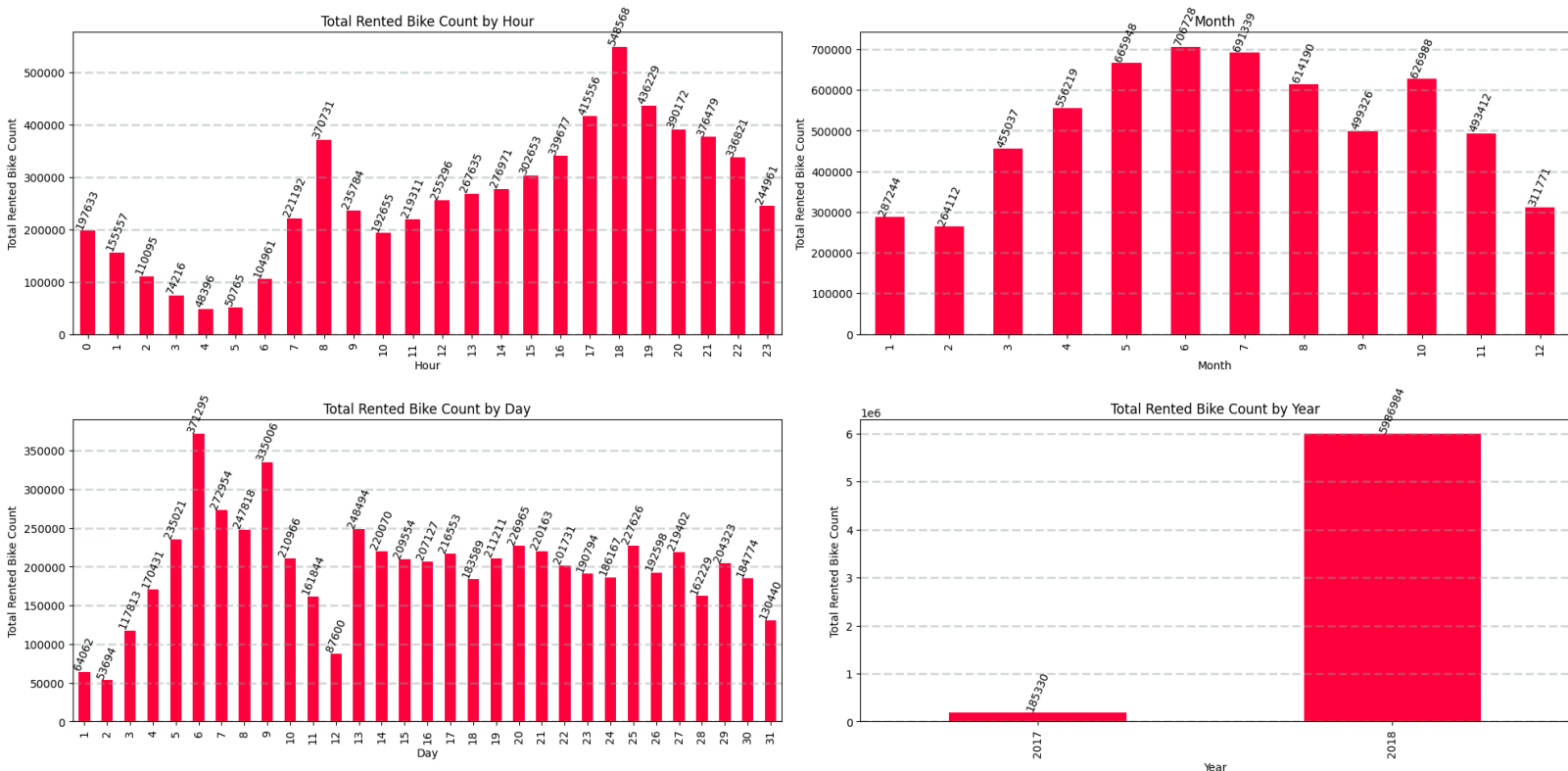
A bar chart was created to display hourly rental counts, revealing how demand changes throughout the day. The analysis showed peak demand between 5:00 PM and 7:00 PM (rush hours). This indicates the need for optimizing the bike-sharing system during these hours.

Monthly Analysis

A bar chart was used to examine the monthly variation in bike rental demand. The chart displayed the average rental numbers for each month, helping identify monthly trends. It was found that May and June experienced the highest demand, while January showed the lowest. This analysis emphasizes the need to allocate resources according to the demand peaks in specific months.

Yearly Analysis

A bar chart was generated to examine yearly rental counts and observe changes over time. The analysis showed a significant increase in 2018 compared to 2017, indicating a rise in system usage over the years.



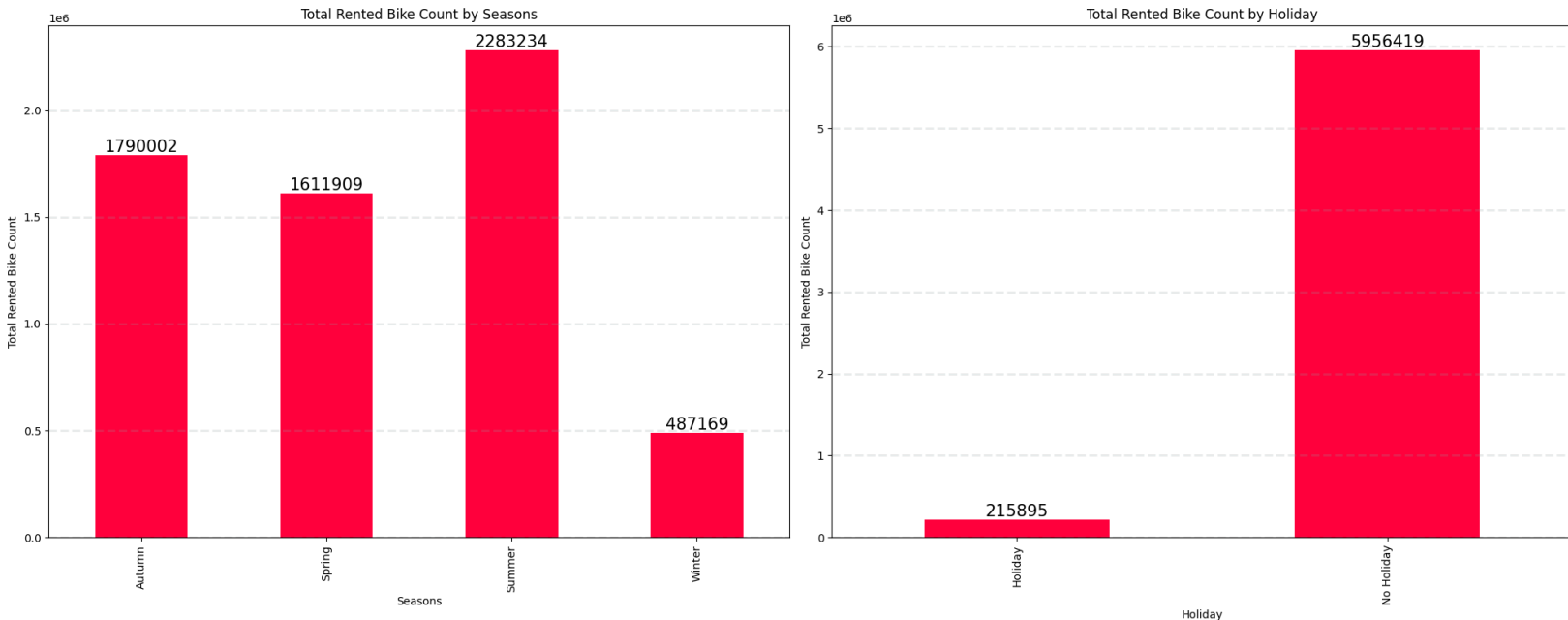
COMPUTER ENGINEERING DEPARTMENT

Weather Impact

A correlation matrix was generated to analyze the impact of weather conditions on bike rentals. The matrix displayed correlations among numerical variables, identifying which weather variables were most associated with rental demand. Notably, temperature and dew point temperature were found to have a positive relationship with the number of rented bicycles. This analysis demonstrates that weather conditions significantly affect bike usage, making them crucial features for forecasting models.

Holiday Analysis

To analyze the impact of holidays on bike rental demand, a bar chart comparing rentals on holidays and regular days was created. This analysis showed that bike rentals were lower on holidays compared to regular days, highlighting the need for special planning for holidays.



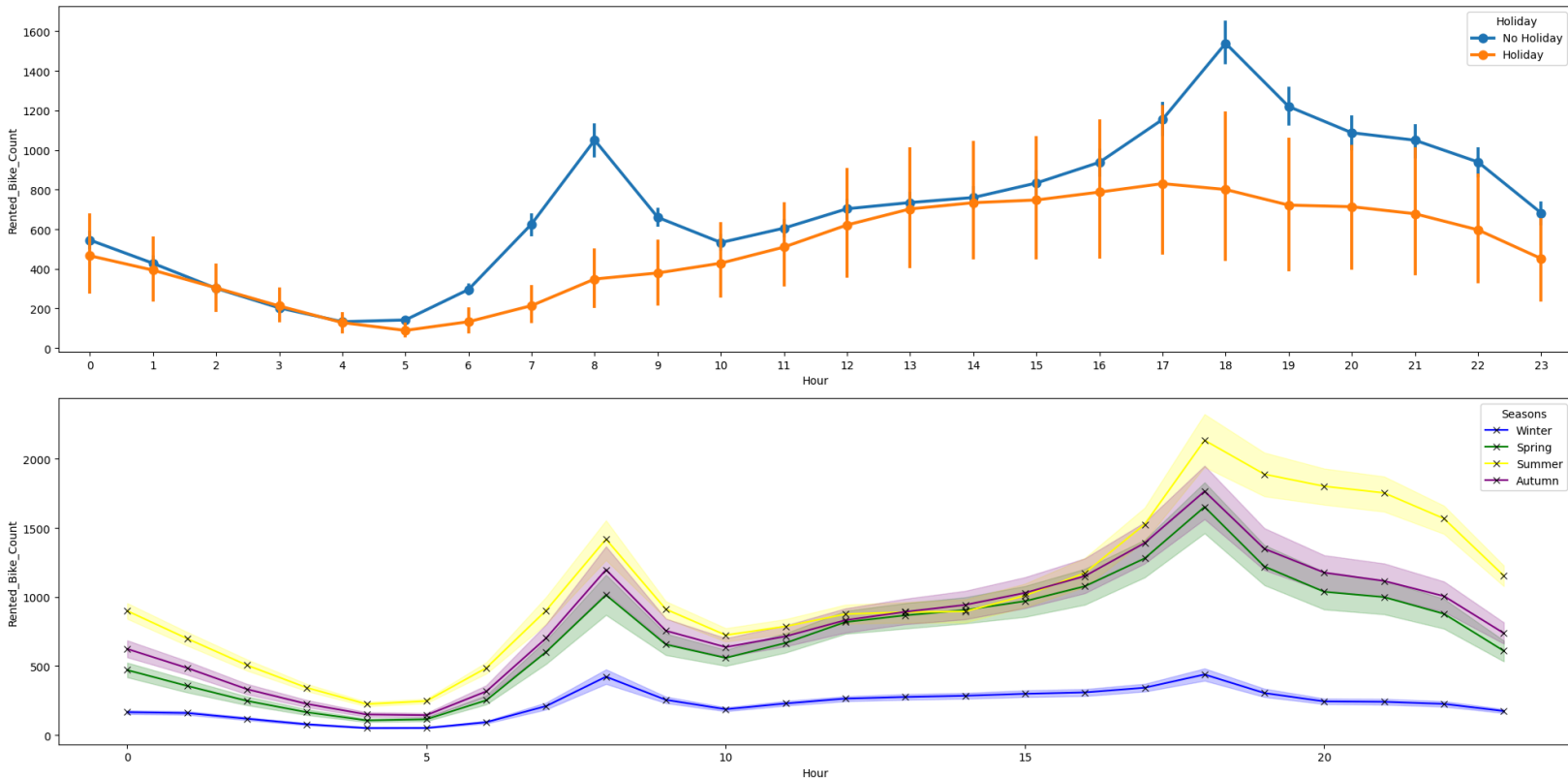
COMPUTER ENGINEERING DEPARTMENT

Hourly and Holiday Analysis

A line graph was created comparing hourly bike rentals on holidays and regular days. This visualization showed that rental activity was lower on holidays than on regular days. Also, on non-holiday days, rush hours (07:00-09:00) and 17:00-19:00 were quite busy.

Seasonal and Hourly Analysis

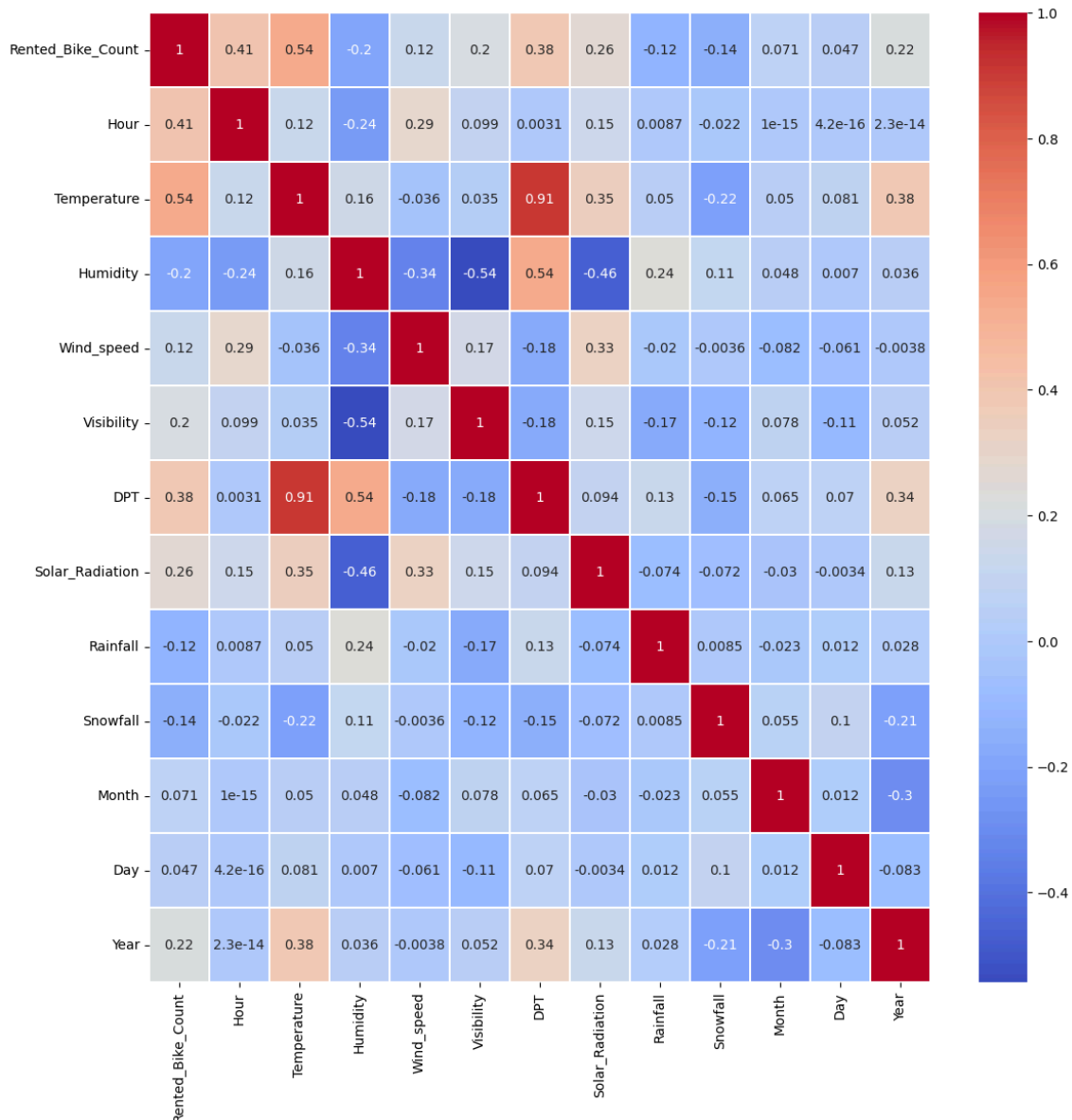
To compare hourly bike rentals across seasons, a line chart was created. This visualization revealed that rental demand during summer was higher throughout the day compared to other seasons.



Some Importants Notes :

Feature Engineering:

- Month and day names were extracted from the date column.
- The Date column was converted to datetime format.
- Categorical columns like Holiday, Seasons, Functioning_Day, and Day_Name were copied into a separate DataFrame.
- Day, month, and year information were extracted from the Date column to create new columns.
- Unnecessary columns, Date and DPT, were removed from the dataset.
- Correlations between all numerical columns were calculated using `df.corr()` and visualized with `sns.heatmap`.
- DPT and Temperature were highly correlated. Therefore, they could lead to multicollinearity problem.
- DPT was deleted because it was less related to the target value.



COMPUTER ENGINEERING DEPARTMENT

Label Encoding:

- Categorical variables were transformed into numerical values using LabelEncoder.
- Unnecessary columns such as Functioning_Day, Date and DPT were removed. Outlier Analysis:
- Outliers were visualized using sns.boxplot. Target Variable Transformation:

```
[ ] from sklearn.preprocessing import LabelEncoder
label_encoded = df.apply(LabelEncoder().fit_transform)
label_encoded.drop(columns=['Functioning_Day', 'Date', 'DPT'], inplace = True)
```

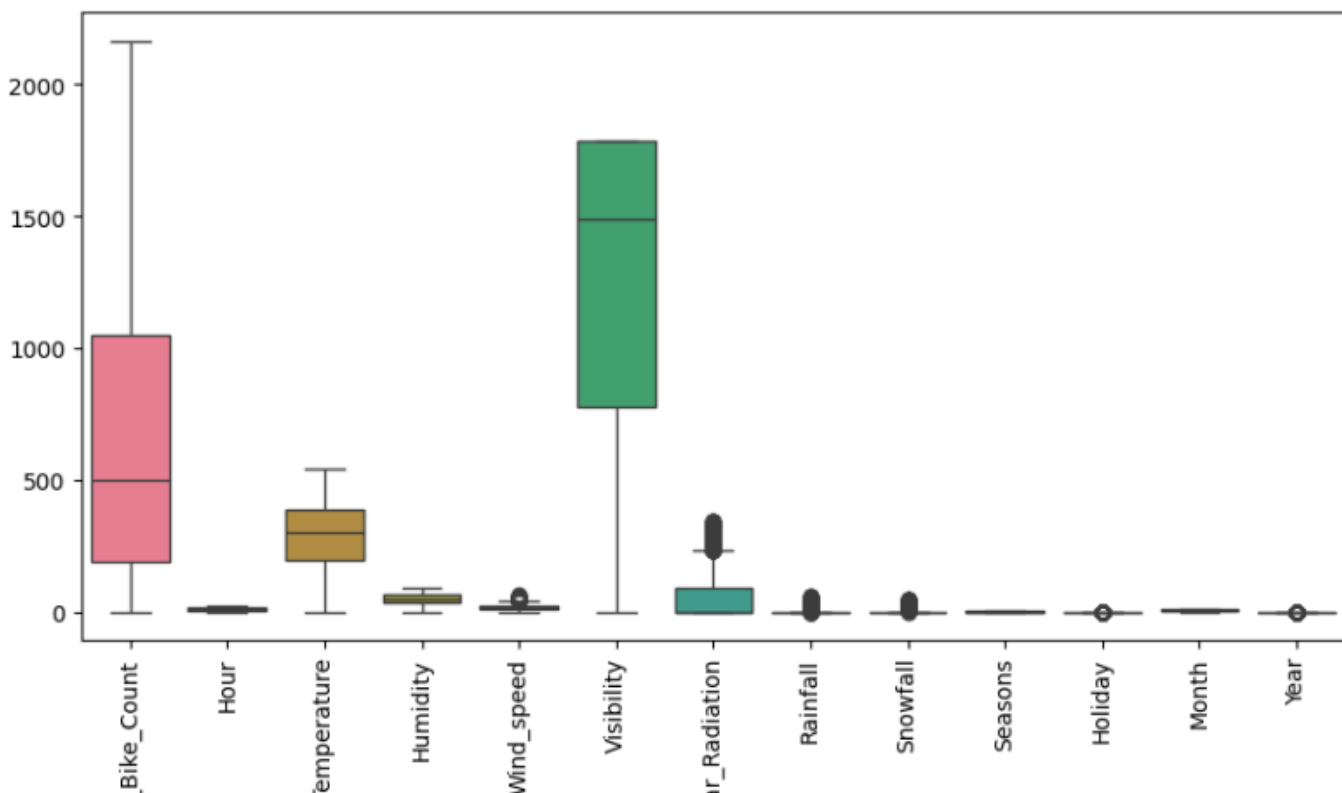
```
[ ] label_encoded.groupby('Seasons').count()
```



Seasons	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed
0	2184	2184	2184	2184	2184
1	2208	2208	2208	2208	2208
2	2208	2208	2208	2208	2208
3	2160	2160	2160	2160	2160

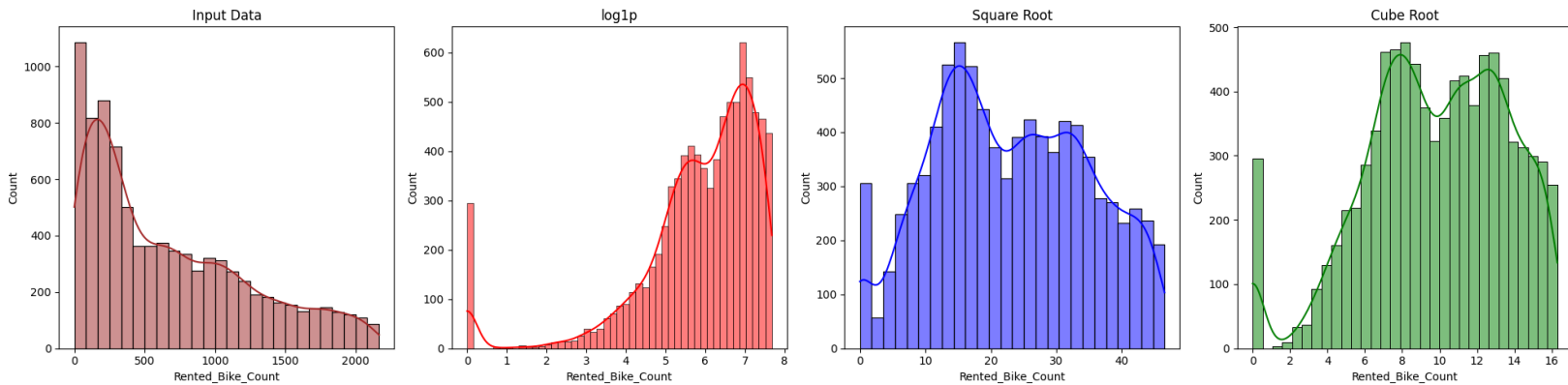
Show per page

```
plt.figure(figsize=(10,5))
plt.xticks(rotation=90)
sns.boxplot(data = label_encoded)
plt.show()
```



COMPUTER ENGINEERING DEPARTMENT

- The distribution of Rented_Bike_Count was analyzed using histograms and Q-Q plots.
- Log, square root, and cube root transformations were attempted.
- The cube root transformation was found to produce a distribution closer to normal, and Rented_Bike_Count was cube root transformed and stored in the RBC_qb column.
- The square root transformation was found to be better and was applied to the label_encoded1 DataFrame. Wind_speed Column Transformation: The same approach was used and Wind_speed column was transformed with square root.



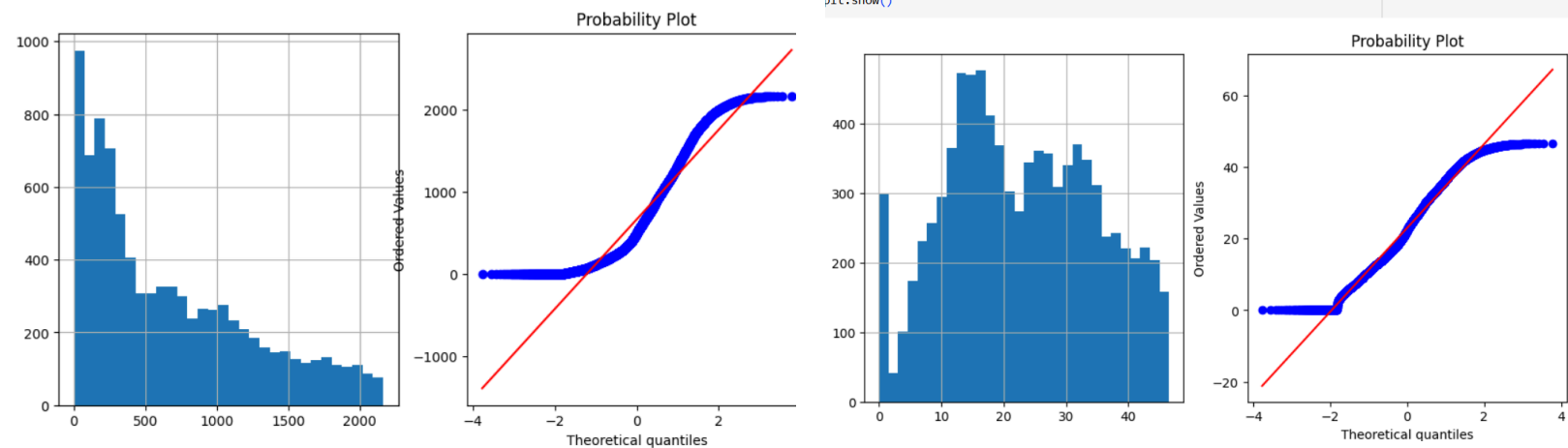
```
plt.figure(figsize=(10,5))
plt.subplot(1,2,1)
label_encoded['Rented_Bike_Count'].hist(bins = 30)

#QQ Plot
plt.subplot(1,2,2)
stats.probplot(label_encoded['Rented_Bike_Count'], dist = 'norm', plot = plt)
plt.show()
```

```
label_encoded1['Rented_Bike_Count'] = np.sqrt(label_encoded1['Rented_Bike_Count'])

plt.figure(figsize=(10,5))
plt.subplot(1,2,1)
label_encoded1['Rented_Bike_Count'].hist(bins = 30)

#QQ Plot
plt.subplot(1,2,2)
stats.probplot(label_encoded1['Rented_Bike_Count'], dist = 'norm', plot = plt)
plt.show()
```



COMPUTER ENGINEERING DEPARTMENT


3. Development of Predictive Models

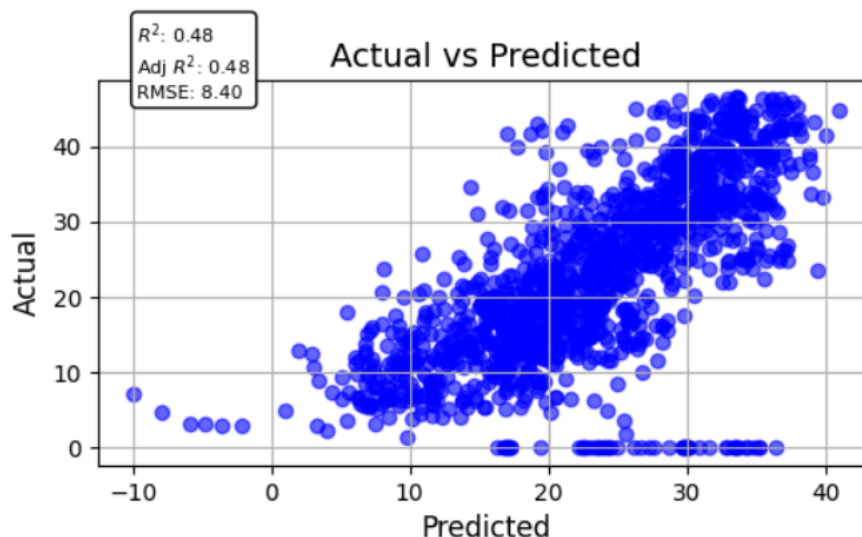
- The dataset was split into training and testing sets.
- Models such as Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor, and LightGBM were trained.
- R^2 and RMSE metrics were used to evaluate model performance.

Model Evaluation


- The trained models were evaluated on test data.
- The LightGBM model was identified as the best-performing model.
- Model Selection and Training:
- Multiple regression models (LinearRegression, Lasso, Ridge, RandomForestRegressor, LGBMRegressor) were tested.
- A separate predict function was written for each model, which:
 - 1. Split the data into training and test sets.
 - 2. Standardized the data.
 - 3. Trained the model.
 - 4. Made predictions.
 - 5. Calculated and printed R^2 , Adjusted R^2 , and RMSE metrics.
 - 6. Plotted a scatter plot comparing actual and predicted values.


```
[ ] predict(LinearRegression(),X,y)
```

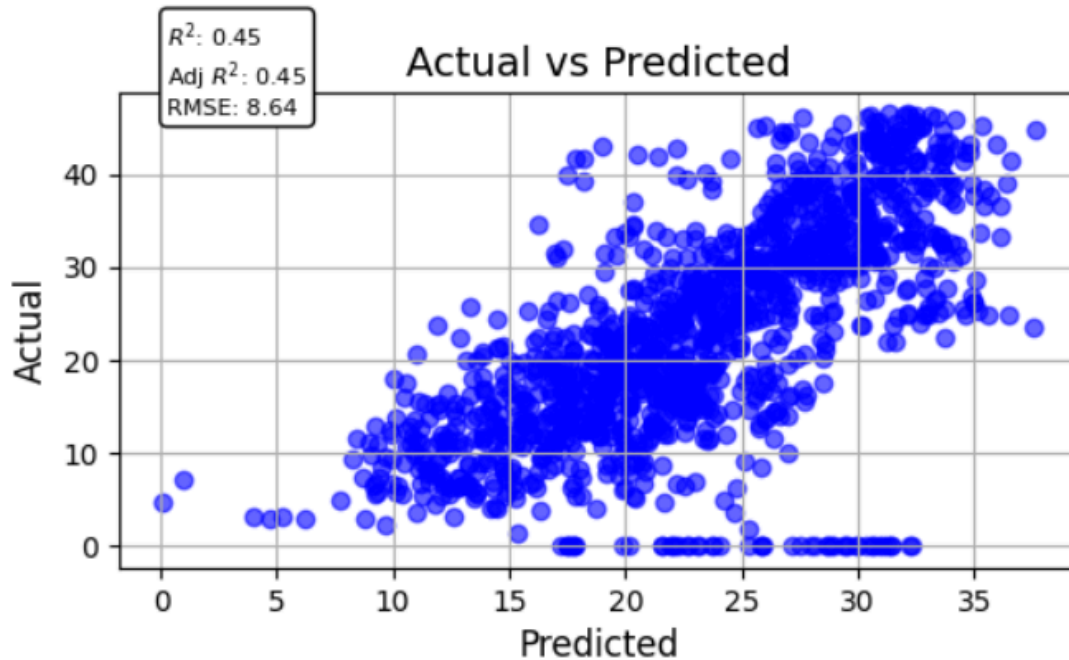
 R^2 : 0.48
Adjusted R^2 : 0.48
RMSE: 8.40





COMPUTER ENGINEERING DEPARTMENT

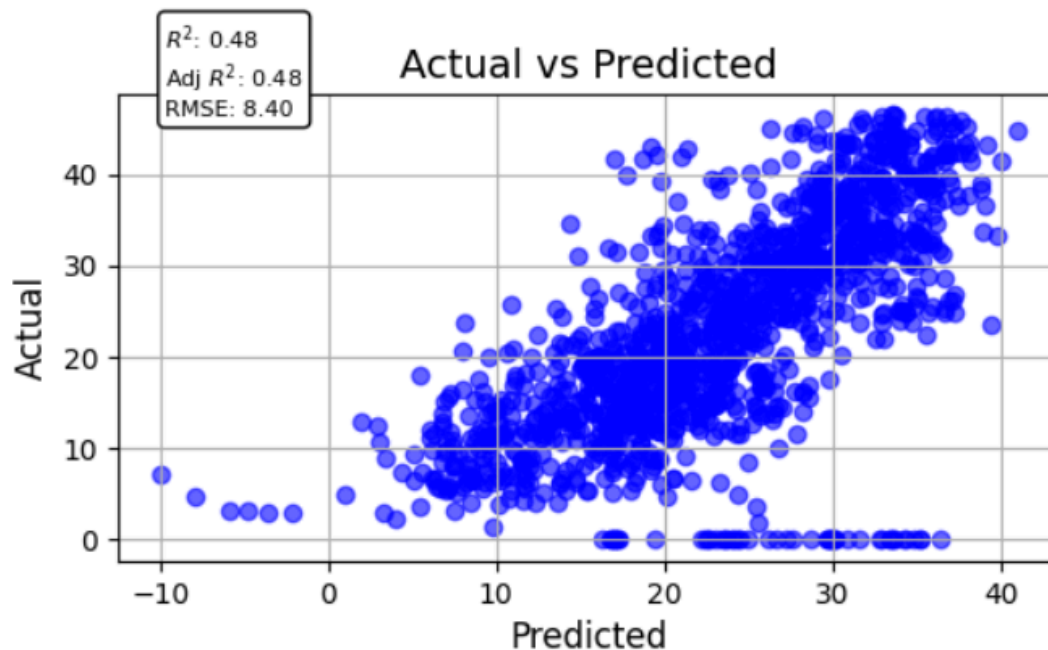
 `predict(Lasso(),X,y)`

 R^2 : 0.45
Adjusted R^2 : 0.45
RMSE: 8.64




 `predict(Ridge(),X,y)`

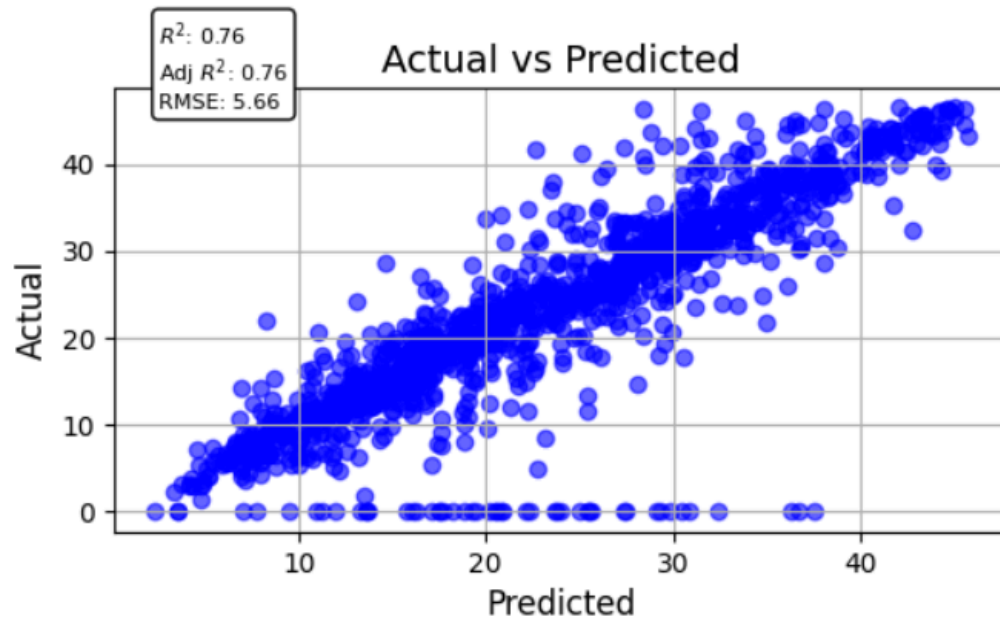
 R^2 : 0.48
Adjusted R^2 : 0.48
RMSE: 8.40



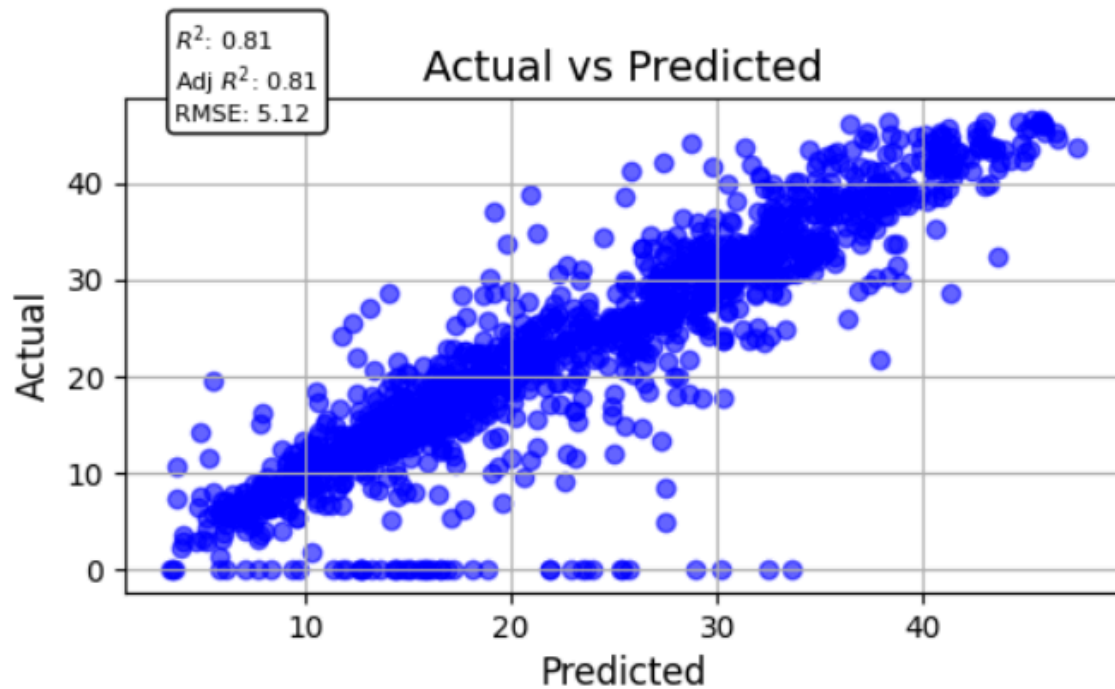
COMPUTER ENGINEERING DEPARTMENT

 `predict(RandomForestRegressor(),X,y)`

 R^2 : 0.76
Adjusted R^2 : 0.76
RMSE: 5.66



R^2 : 0.81
Adjusted R^2 : 0.81
RMSE: 5.12



Interpretation of Metrics:

R-squared (R^2): This metric indicates the proportion of the variance in the dependent variable (Rented_Bike_Count) that can be explained by the independent variables. A higher R^2 value suggests a better fit. From the results, we can say that:

- LinearRegression, Lasso, and Ridge models show about 45-48% of the variance in Rented_Bike_Count can be explained by the features.
- RandomForestRegressor can explain around 76% of the variance.
- LGBMRegressor performs best, explaining 81% of the variance. Adjusted R-squared: This metric is similar to R^2 , but it penalizes the model for adding irrelevant features. We see that for all the models the Adjusted R^2 is equal to R^2 , which indicates that there is no irrelevant feature and it is good for the model. Root Mean Squared Error (RMSE): This metric measures the average magnitude of the errors. Lower RMSE indicates a better model.

Results are:

- LinearRegression, Lasso, and Ridge models have a RMSE between 8.40 and 8.64.
- RandomForestRegressor shows an RMSE of 5.66
- LGBMRegressor has the lowest RMSE at 5.12.

6. Follow-Up & Evaluation Plan

The implementation plan will be considered successful if the performance monitoring is regularly undertaken along with user feedback. Some of the KPIs that will be tracked are bike utilization rates, user satisfaction scores, and operational efficiency metrics. Monthly reports summarizing these metrics will be shared with stakeholders to ensure transparency and accountability.

User feedback is collected through surveys and reviews from the mobile application. This shall then give insights into satisfaction and areas of improvement. The predictive models will also be retrained periodically with updated data for relevance and accuracy.

Follow-up will be done on strategies annually through an assessment meeting that looks at the general impact of the set strategies. It reviews the KPIs, challenges faced in that year, and further planning for improvement. With this kind of follow-up and evaluation plan, the bike-sharing system is bound to continuously improve towards the achievement of long-term success.

7. References And Links

Dataset : Seoul Bike Sharing Demand

Colab Work : CaseStudy2 SBSD.ipynb