YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

CSE464 Case Study 1

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

MEHMET ÖNAL

20200702081

YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

I. Seoul Bike Sharing Demand Analysis:

This case study analyzes Seoul's bike sharing demand data to understand influencing factors and predict future rentals. Using descriptive analytics and visualizations, we identified key relationships between rentals and variables like weather, seasonality, and time of day. These insights allow for better resource allocation, optimized pricing strategies, and improved service availability for Seoul's bike sharing system.

II. Background/Introduction

- **About the Client/Company**: A bike-sharing company operating in Seoul, South Korea. The company aims to improve its operational efficiency and customer satisfaction.
- **Industry Context:** The bike-sharing industry is a growing market in urban areas, providing an eco-friendly and convenient alternative transportation method. Accurate demand prediction is crucial for managing resources, optimizing pricing, and ensuring user satisfaction.
- **Goals and Objectives:** The primary goal is to predict the "Rented Bike Count" based on available data. This prediction will enable the company to optimize bike availability at different stations and times, minimize user wait times, and potentially increase revenue.

III. Challenge/Problem Statement

The Problem: The bike-sharing company faces the challenge of accurately predicting bike rental demand, which leads to inefficient bike allocation and potentially dissatisfied customers.

Why It Matters: Accurate predictions are crucial for optimizing bike availability at different stations and times. This can lead to increased customer satisfaction, improved operational efficiency, and maximized revenue potential.

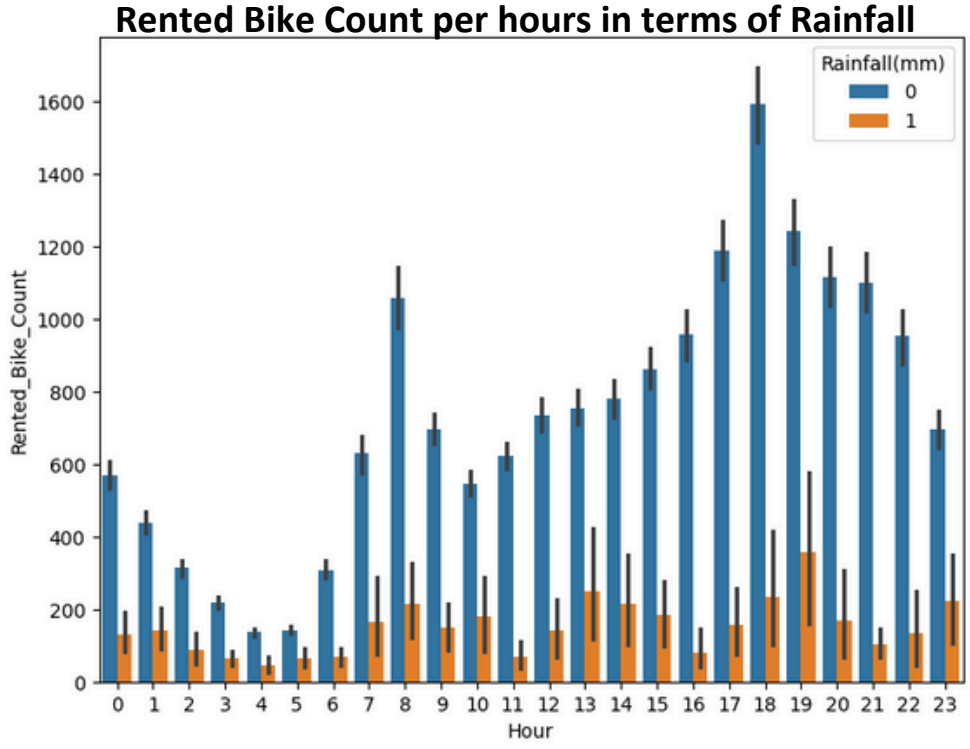
IV. Solution/Approach

A descriptive analytics approach is employed, focusing on understanding historical data and identifying patterns to predict future demand. The provided Python code demonstrates the following steps:

1. **Data Loading and Preprocessing:** The Seoul Bike dataset (SeoulBikeData.csv) is loaded using pandas. The 'Date' column is converted to datetime objects and used to extract 'Month', 'Year', and 'Day_Name'. Categorical features like 'Seasons', 'Holiday', and 'Functioning_Day' are converted to numerical representations. 'Rainfall(mm)', 'Snowfall(cm)', 'Visibility', and 'Solar Radiation(MJ/m2)' are thresholded and converted into binary categories (0 and 1).
2. **Exploratory Data Analysis (EDA):** Several bar plots are generated using seaborn to visualize relationships between 'Rented Bike Count' and various features like 'Hour', 'Seasons', 'Solar Radiation', 'Snowfall', 'Visibility', 'Rainfall', 'Functioning_Day', 'Day_Name', and 'Month'.
3. **Feature Engineering:** One-hot encoding is applied to categorical features such as 'Seasons' and 'Day_Name' to prepare them for the machine learning models.
4. **Model Building and Evaluation:** The dataset is split into training and testing sets. Three linear models—Linear Regression, Lasso, and Ridge Regression—are trained on the standardized training data. Model performance is evaluated using R^2, Adjusted R^2, and RMSE. The predict function calculates these metrics and visualizes the actual vs. predicted values using a scatter plot.

V. Results/Outcomes

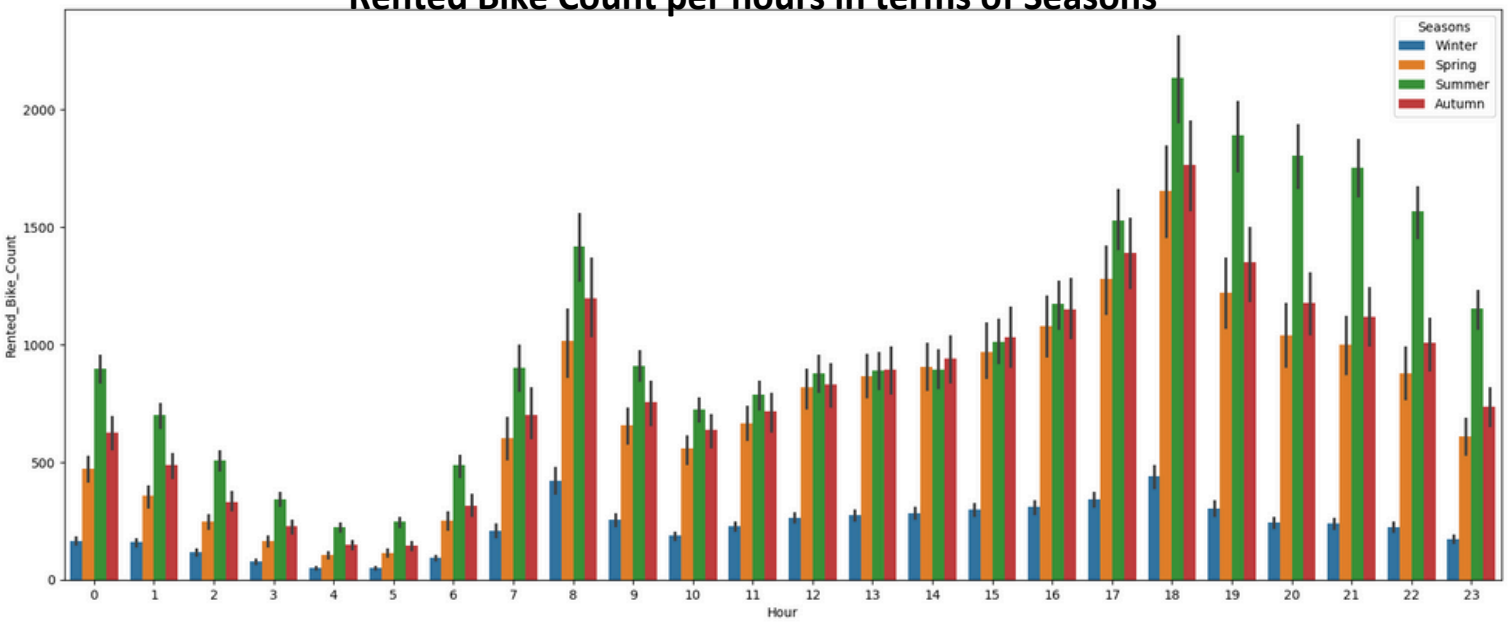
Quantifiable Results: The provided screenshots show the performance of the three models. All three achieve similar results:
R^2: Around 0.56 for all three models.
Adjusted R^2: Approximately 0.55 for all three.
RMSE: Around 426 for all three models.
Qualitative Impact: The EDA provides valuable insights into bike rental patterns. For instance, rentals are highest during peak commute hours (8 am and 6 pm), in the summer, and on functioning, non-rainy days. This information can guide operational and marketing strategies. The scatter plots of actual vs. predicted values show a reasonable correlation but also highlight the limitations of the linear models in capturing the full complexity of the data.



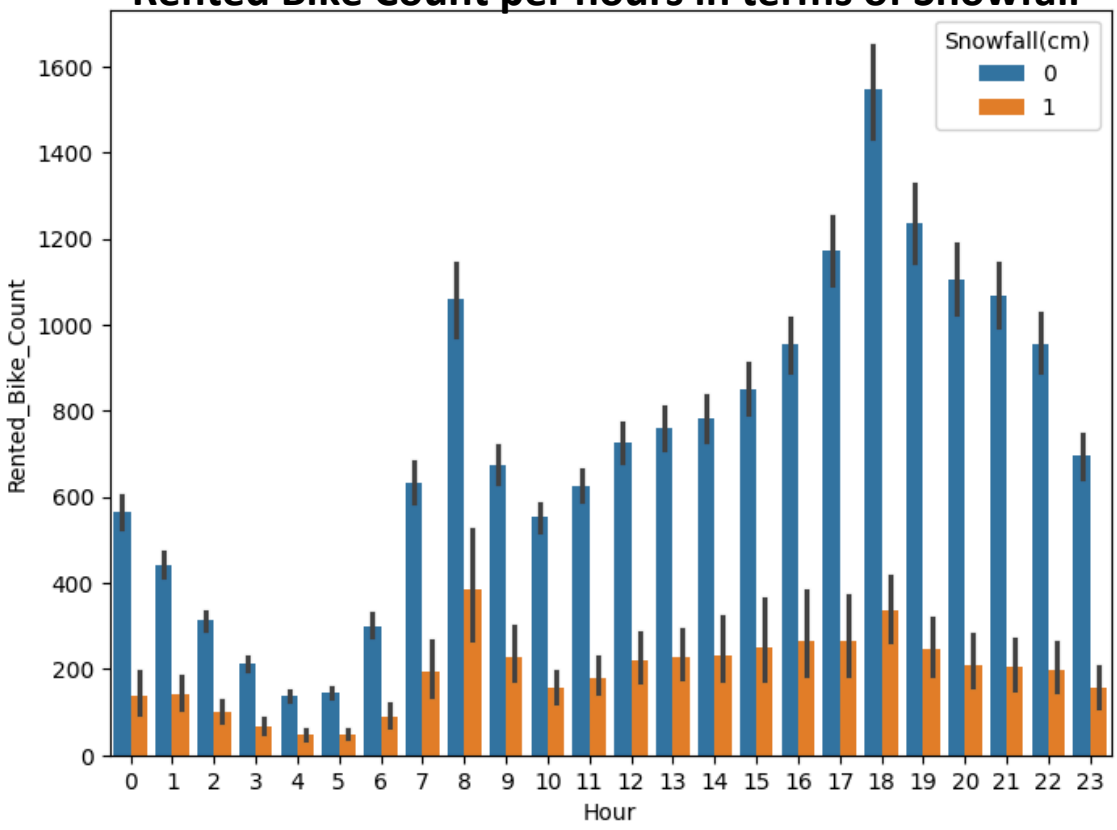
YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

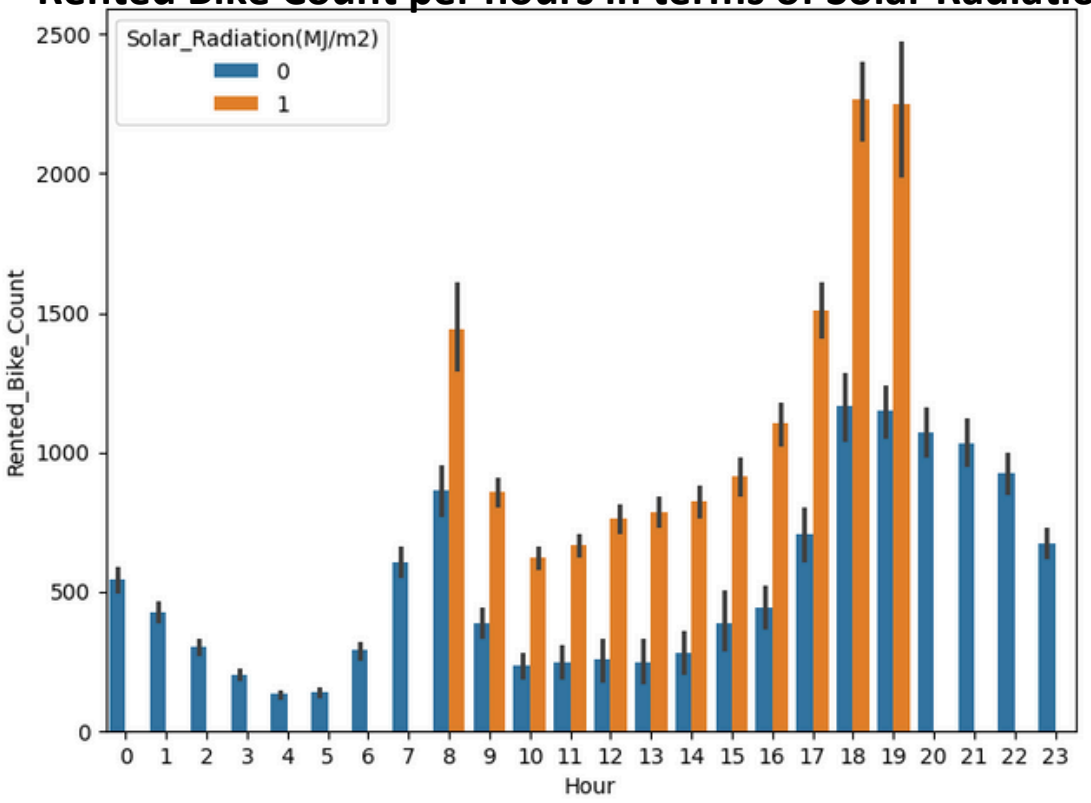
Rented Bike Count per hours in terms of Seasons



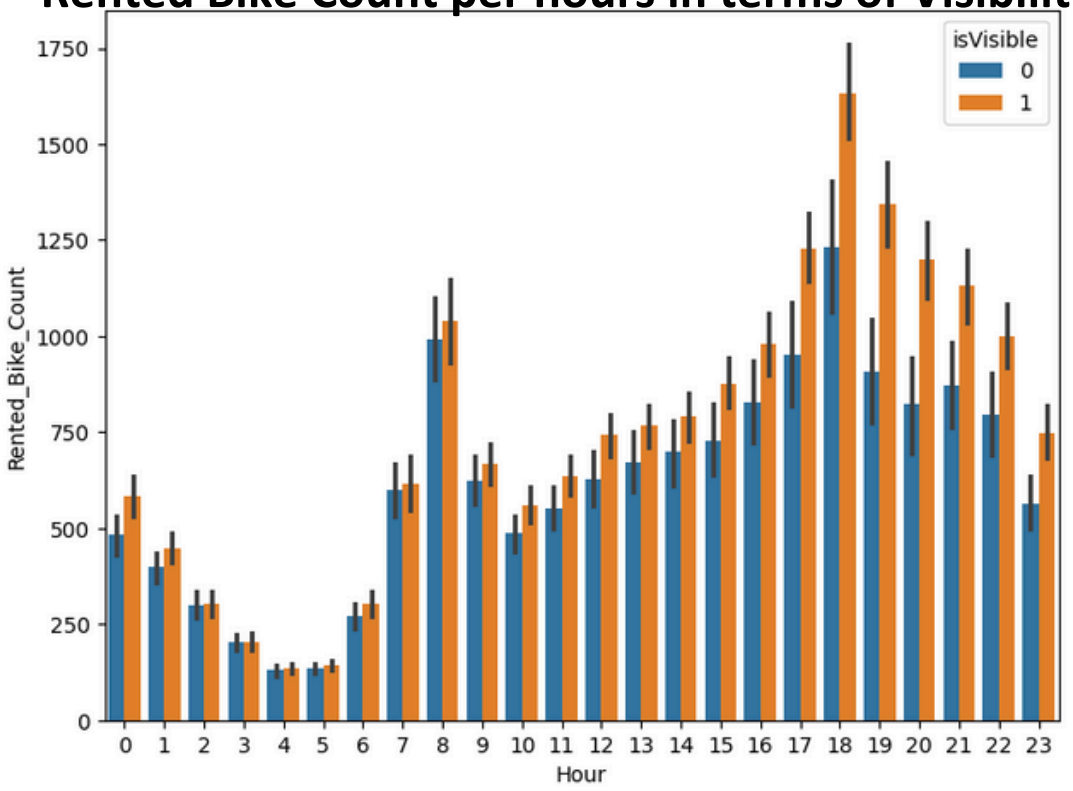
Rented Bike Count per hours in terms of Snowfall



Rented Bike Count per hours in terms of Solar Radiation



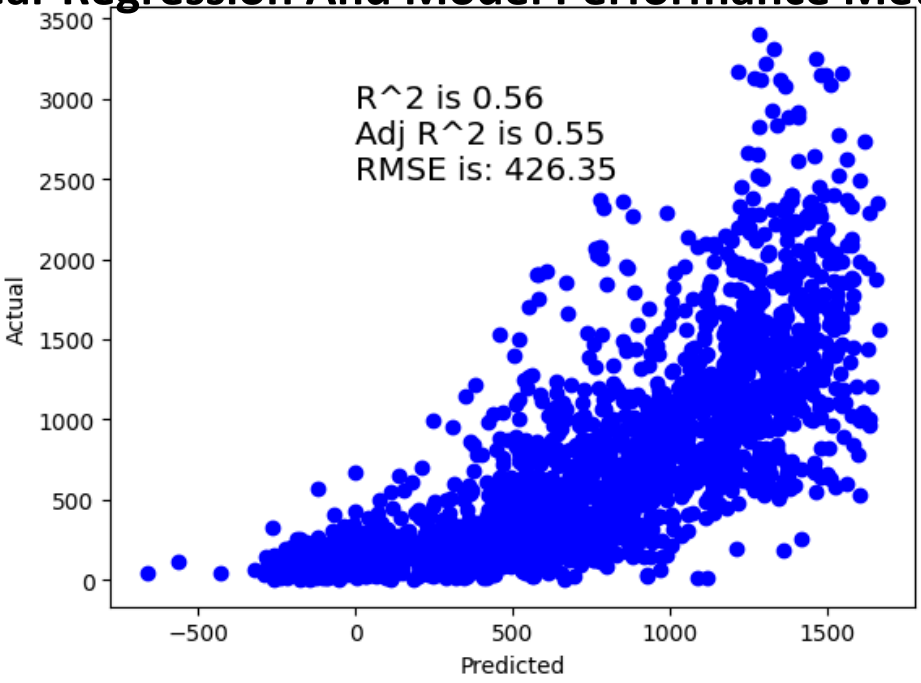
Rented Bike Count per hours in terms of Visibility



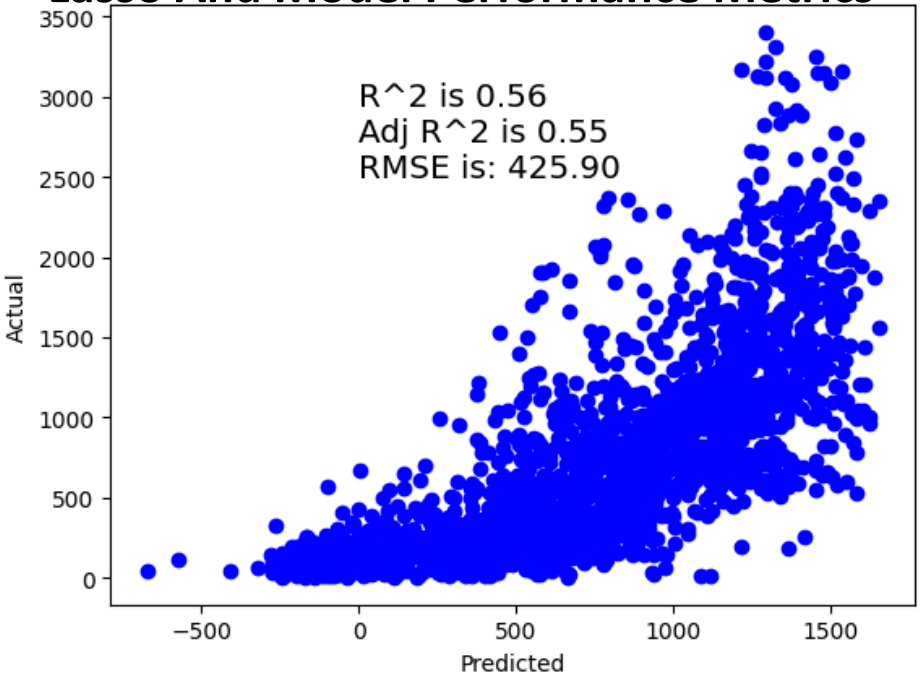
YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

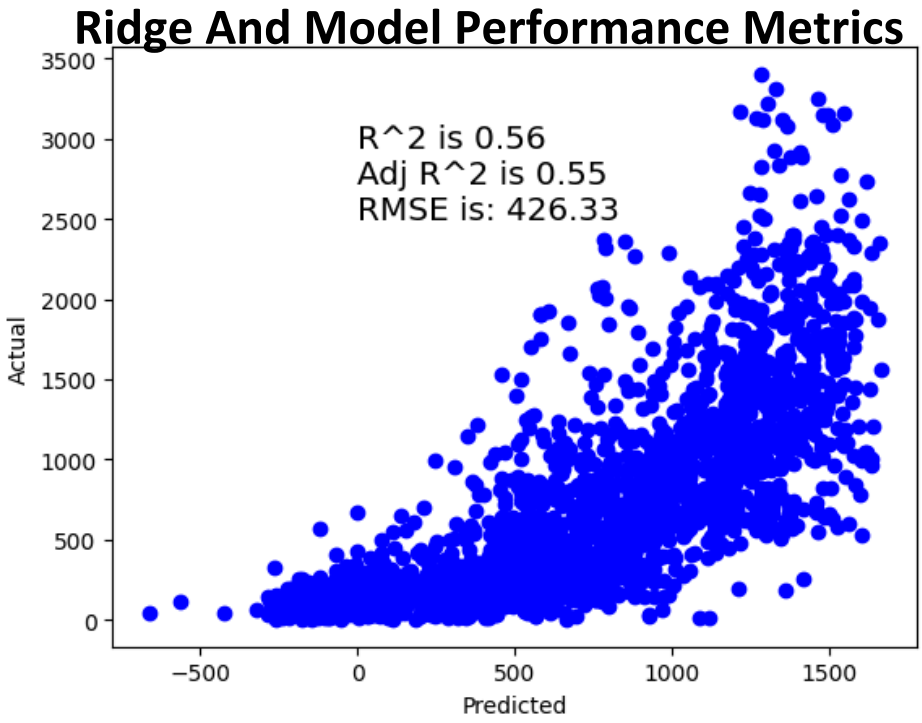
COMPUTER ENGINEERING DEPARTMENT

Linear Regression And Model Performance Metrics



Lasso And Model Performance Metrics





VI. Recommendations

H. Key Takeaways: Insights from the project that others can apply. İ.

- **Predictive Analytics for Operational Optimization:**

The use of predictive modeling can greatly enhance the efficiency of bike-sharing operations by forecasting demand patterns. By anticipating peak times and periods of high usage, the company can better allocate resources, reduce downtime, and improve customer satisfaction. Predictive insights allow for dynamic fleet management, ensuring that the availability of bikes aligns with user needs, especially during peak hours, seasons, or adverse weather conditions.

- **Impact of Feature Engineering on Model Accuracy:**

Thoughtful feature engineering played a pivotal role in boosting the model's performance and interpretability. Transforming complex weather data into binary features such as "isVisible" (based on visibility levels) and "Rainfall(mm)" (indicating rain presence) simplified the model's understanding of the weather's effect on bike demand. These engineered features improved the model's ability to capture key relationships, leading to more accurate and interpretable predictions.

- **Importance of Model Selection and Evaluation:**

Testing multiple models and thoroughly evaluating their performance was crucial in identifying the best-fitting approach. The close performance of Linear Regression, Ridge, and Lasso models highlights the importance of regularization techniques in controlling overfitting, while also showing that simpler models might sometimes provide comparable results. Careful model evaluation, including metrics like R^2 and RMSE, ensured the chosen model was robust and well-suited for predicting bike rental demand.

Challenges Overcome: How obstacles were addressed during the process.

- Data cleaning and handling missing values.
- Selecting the most appropriate machine learning model for the problem.
- Balancing model complexity and prediction accuracy.

This case study demonstrates the effectiveness of machine learning in predicting bike rentals. By incorporating these insights, bike-sharing companies can optimize their operations, improve customer experience, and contribute to a more sustainable transportation system in Seoul.

Data Quality and Preprocessing Challenges:

One of the initial hurdles was handling the dataset's missing values and categorical variables. Ensuring data consistency through imputation and applying transformations like one-hot encoding to categorical variables were essential steps. These actions significantly enhanced the quality of the dataset, making it suitable for effective machine learning. By removing irrelevant or redundant features and focusing on those most impactful to the prediction, the model's accuracy was further refined.

Model Selection:

The iterative process of model selection, combined with thoughtful parameter tuning and advanced preprocessing techniques, was key to achieving optimal performance. While initial models showed promising results, a deeper analysis and adjustments to feature transformations and regularization techniques (like Ridge and Lasso) allowed the models to capture the complexities of the data more effectively. This adaptive approach ensured that the final models were both accurate and reliable.