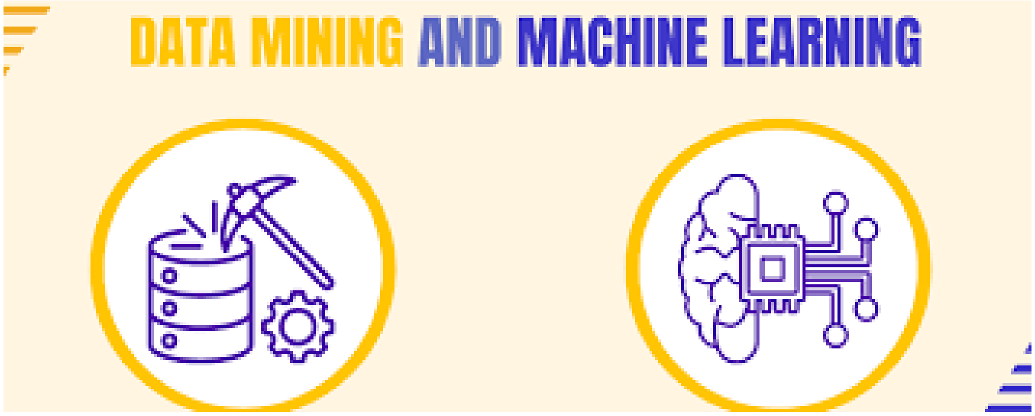


# MACHINE LEARNING & DATA MINING: BIKE SHARING INSIGHTS IN SEOUL

A Data Analysis & Visualization Journey

By:

Uzma Naeem

# TABLE OF CONTENTS

## EXECUTIVE SUMMARY

This report delves into the bike sharing demand prediction project, which aims to explore and model urban transportation patterns to enhance the efficiency of bike sharing systems. Central to our analysis is a comprehensive dataset that integrates rental bike counts with weather conditions and holiday information, offering a holistic perspective on urban mobility.

The dataset provides a multifaceted view of how environmental factors and societal events influence bike rental demand. By thoroughly examining this data, we uncover significant insights into the complex interplay between weather patterns, holiday seasons, and bike sharing usage. These insights are not merely statistical; they offer practical implications for urban planners and researchers striving to optimize transportation systems and promote sustainable urban development.

Our findings illuminate the potential for transforming urban mobility through predictive modeling. By accurately forecasting bike sharing demand, cities can better allocate resources, enhance user satisfaction, and reduce environmental impacts. This report presents a detailed account of our methodologies, analyses, and results, contributing to the broader discourse on sustainable urban transportation solutions.

## DESCRIPTION OF THE DATA

The dataset utilized for this project is the "Seoul Bike Sharing Demand" dataset, donated on February 29, 2020. It offers a detailed hourly account of public bicycle rentals in Seoul, combined with relevant weather and holiday information. Designed for regression tasks, this multivariate dataset comprises both integer and real-valued features, making it suitable for comprehensive analysis.

Covering a full year, the dataset includes **8,760 instances** and **13 features**. It captures critical aspects such as rental bike counts, weather conditions, and date-specific information, providing a holistic view of urban mobility.

**Key features** of the dataset include the date (year-month-day), the number of bikes rented each hour, and the hour of the day (0-23). Weather-related variables encompass temperature (in Celsius), humidity (percentage), wind speed (meters per second), visibility (decameters), dew point temperature (Celsius), solar radiation (megajoules per square meter), rainfall (millimeters), and snowfall (centimeters). Additionally, the dataset categorizes seasons (Winter, Spring, Summer, Autumn), identifies holidays with a binary indicator, and specifies whether the hour is functional or non-functional.

Notably, the dataset contains **no missing values**, ensuring data completeness for accurate analysis. Each variable is meticulously recorded to reflect the dynamic factors influencing bike rental demand, from environmental conditions to temporal and special day indicators. This dataset serves as a crucial tool for understanding the interplay between these variables and for developing predictive models aimed at optimizing bike-sharing systems in urban environments.

## PROBLEM STATEMENT

The objective of this project is to develop a robust predictive model capable of accurately forecasting bike-sharing demand based on comprehensive data analysis. This model aims to enhance the operational efficiency of bike-sharing systems by identifying key features that influence rental patterns. By understanding seasonal trends and determining feature importance, the insights gained can inform marketing strategies to increase system usage and optimize resource allocation. This endeavor seeks to provide a data-driven foundation for improving business performance and promoting the sustainable growth of bike-sharing services.

## SIGNIFICANCE & RELEVANCE

The significance and relevance of this project with the Data Mining Course makes it **interesting**, which is because of the following factors:

### Addressing Urban Mobility Challenges

Urban mobility is a pressing issue in modern cities, with traffic congestion and environmental concerns being major challenges. Bike sharing systems offer a sustainable solution to these challenges by providing eco-friendly transportation alternatives. By accurately predicting bike-sharing demand, this project contributes to the optimization of these systems, making them more efficient and accessible.

### Practical Application of Course Concepts

By applying predictive analysis techniques to real-world data from the Seoul Bike Sharing System, this project demonstrates the practical application of concepts learned in the data mining course.

This hands-on approach enabled us to translate theoretical knowledge into actionable insights with tangible real-world impact.

**Optimizing Bike-sharing Systems**

Accurate demand prediction enables bike-sharing system operators to better allocate resources, such as bikes and docking stations, to meet user demand effectively. This optimization leads to improved operational efficiency, reduced wait times for users, and enhanced overall user experience.

**Promoting Eco-friendly Transportation**

Bike sharing systems play a crucial role in promoting eco-friendly transportation modes and reducing carbon emissions. By optimizing bike-sharing systems, this project contributes to the promotion of sustainable transportation options, aligning with broader efforts to combat climate change and reduce environmental impact.

**Informing Policy Decisions and Urban Planning**

Understanding the factors influencing bike rental counts provides valuable insights for policymakers and urban planners. These insights can inform policy decisions related to transportation infrastructure, bike lane development, and urban planning efforts aimed at creating bike-friendly cities.

## LITERATURE REVIEW

To gain insights into the problem at hand, we conducted a literature review. One notable study is given below:

### Predicting Bike Sharing Demand using Recurrent Neural Networks

A seminal research paper titled "Predicting Bike Sharing Demand using Recurrent Neural Networks" by Yan Pan, Ray Chen Zheng, Jiaxi Zhang, and Xin Yao addresses the critical challenge of forecasting bike-sharing demand. Efficient bike allocation is paramount for customer satisfaction in bike-sharing systems, yet predicting demand accurately remains challenging due to the dynamic influence of factors such as time, events, and weather conditions. This research was conducted at The High School Affiliated to Renmin University of China and the Institute of Remote Sensing and Geographical Information Systems at Peking University.

### Background & Importance

The efficient allocation of bikes in a sharing system is vital for maintaining high levels of customer satisfaction. However, the demand for bikes is influenced by various fluctuating factors, including temporal patterns, special events, and weather conditions. These complexities necessitate sophisticated predictive models capable of handling such dynamic variables.

### Methodology

To tackle these prediction challenges, the researchers employed a Deep Long Short-Term Memory Recurrent Neural Network (LSTM RNN). Unlike traditional neural networks, LSTM RNNs are particularly adept at learning long-term dependencies in sequential data, effectively mitigating the vanishing gradient problem. This capability makes them ideal for analyzing time-series data, such as bike-sharing demand.

### Data Processing

The dataset used for training and evaluating the LSTM RNN model was sourced from the Citi Bike System. It was divided into training and test sets to validate the model's performance. The dataset included not only bike rental counts but also historical weather data and time-related features, providing a comprehensive foundation for accurate demand forecasting.

### Experimentation & Results

Through rigorous experimentation, the study compared the performance of various deep learning models. LSTM RNN emerged as the most effective model, demonstrating superior forecasting accuracy. This conclusion was substantiated by lower Root Mean Squared Error (RMSE) values, indicating the model's proficiency in making precise demand predictions.

### Key Take Aways & Future Applications of The Research

The study underscores the significant benefits of accurate demand prediction facilitated by LSTM RNNs. These benefits include enhanced efficiency in bike allocation, potential optimization of distribution networks, and reduced operational costs. Moreover, the adaptability of the LSTM RNN model extends beyond the scope of the Citi Bike System, suggesting its applicability to other bike-sharing systems globally. This research highlights the promising role of LSTM RNNs in addressing the complex challenges of bike-sharing demand prediction, offering valuable insights for the optimization of urban mobility systems.

## METHODS & TECHNIQUES USED

In our project, various machine learning algorithms were employed, including linear regression, decision trees, random forests, lasso regression, and ridge regression. After preprocessing the data, feature engineering was performed, and hyperparameter tuning was conducted to optimize model performance. Additionally, seasonal trends were explored, feature importance was analyzed, and model accuracy was evaluated through metrics such as RMSE and R-squared.

### Data Collection & Exploration

- The dataset from the Seoul Bike Sharing System was obtained and loaded using pandas' read_csv function.
- Initial exploration of the data involved examining its structure, checking for missing values using the isnull() function, and visualizing key features such as seasonal trends in bike rental demand through bar plots.

### Data Preprocessing

- Categorical variables like 'Seasons', 'Holiday', and 'Functioning Day' were identified and encoded into numerical values using conditional statements.
- Features irrelevant to the prediction task, such as the 'Date' column, were dropped from the dataset using the drop() function.
- The data was initially divided into training and testing sets, with 80% of the data allocated for training and 20% for testing. This ensured an adequate amount of data for model training while retaining a separate portion for evaluation.
- Furthermore, a finer granularity was achieved by splitting the data for each month independently into training and testing subsets. This approach facilitated a more comprehensive assessment of the predictive performance across different time periods, allowing for insights into potential variations and seasonal trends. By partitioning the data in this manner, the model's robustness and generalization capabilities were enhanced, ensuring a thorough evaluation of its efficacy in diverse temporal contexts.
- To address potential issues of heteroscedasticity and negative predictions, a log transformation was applied to the target variable. This transformation not only stabilized

the variance but also ensured the practicality of predictions by restricting them to non-negative values.

- Standardization was applied to scale the features for better model performance using scikit-learn's StandardScaler.

**Model Implementation**

- Various regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, Lasso Regression, and Ridge Regression were implemented using scikit-learn and log transformation to find the best model with highest accuracy for the data.
- Hyperparameter tuning for each model was performed using GridSearchCV to find the optimal parameters that minimize the root mean squared error (RMSE) and maximize the R-squared (R2) score.

**Model Evaluation:**

- The performance of each regression model was evaluated using RMSE and R2 scores on the test dataset.
- Results were compared across different models to select the best-performing one for predicting bike-sharing demand, which was visualized using bar plots.

## LEARNINGS

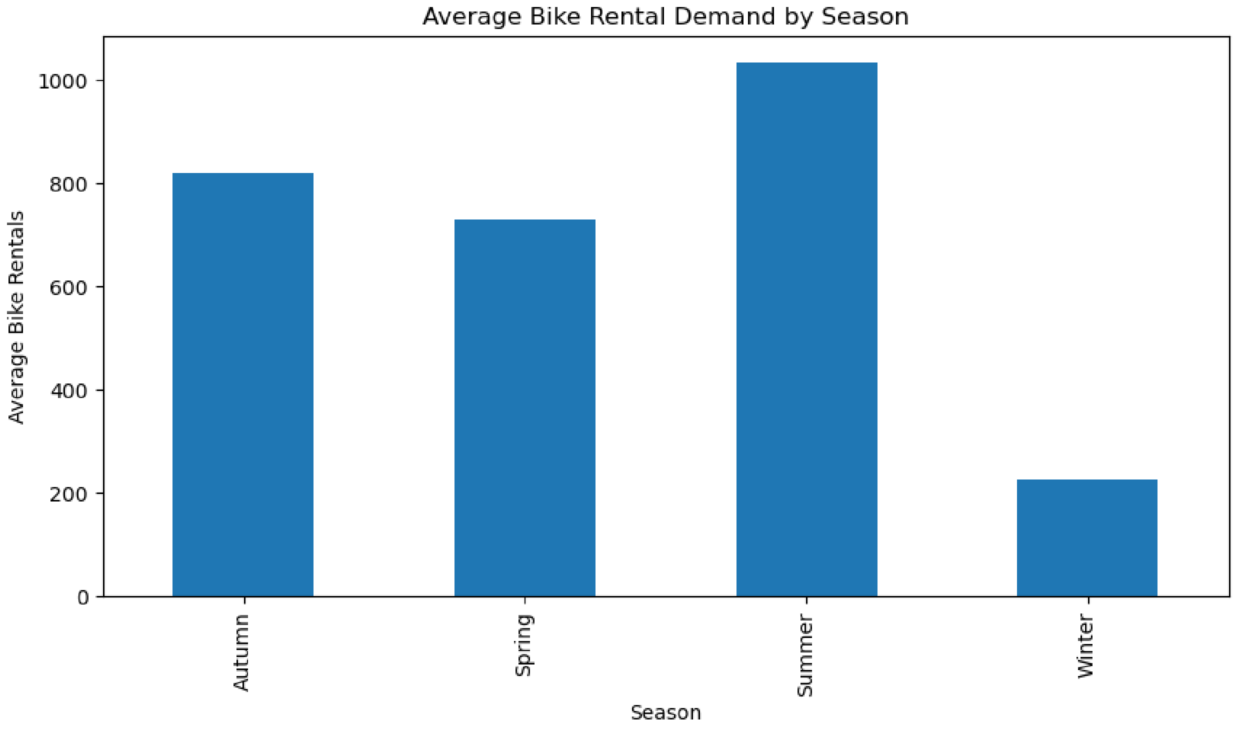
Throughout this project, we gained extensive knowledge in the application of various machine learning algorithms and models, including Linear Regression, Decision Trees, Random Forests, Lasso Regression, and Ridge Regression. This involved understanding the purpose of each model, implementing them effectively, and interpreting their results.

A crucial aspect of this project prior to the application of models was data preparation. We learned the importance of data preprocessing techniques, which included checking for missing values, encoding categorical variables, and applying log transformations to stabilize variance and prevent negative predictions. These steps ensured the dataset was clean and suitable for model training and evaluation.

## KEY FINDINGS & INSIGHTS

### Seasonal Trends

The seasonal analysis of bike rental demand provided additional insights. By grouping the dataset by 'Seasons' and calculating the mean 'Rented Bike Count,' the analysis highlighted clear seasonal trends:



Summer: Peak bike rental demand.

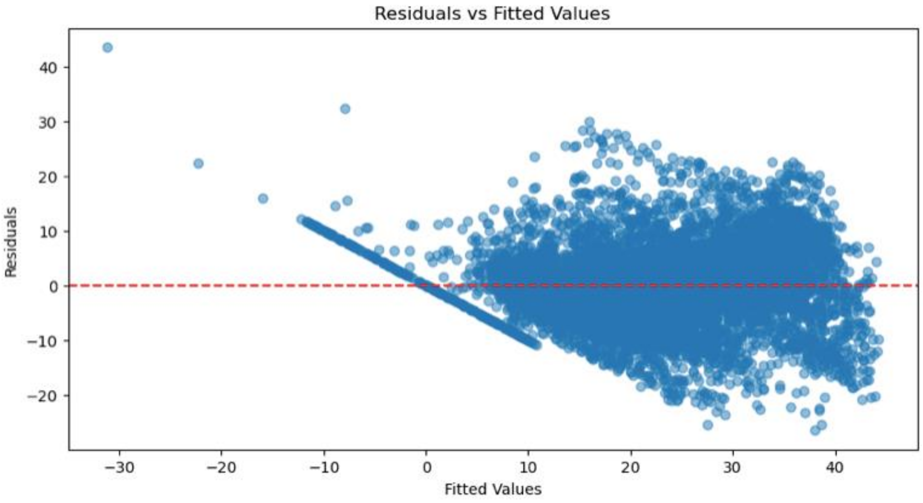
Autumn and Spring: Moderate demand.

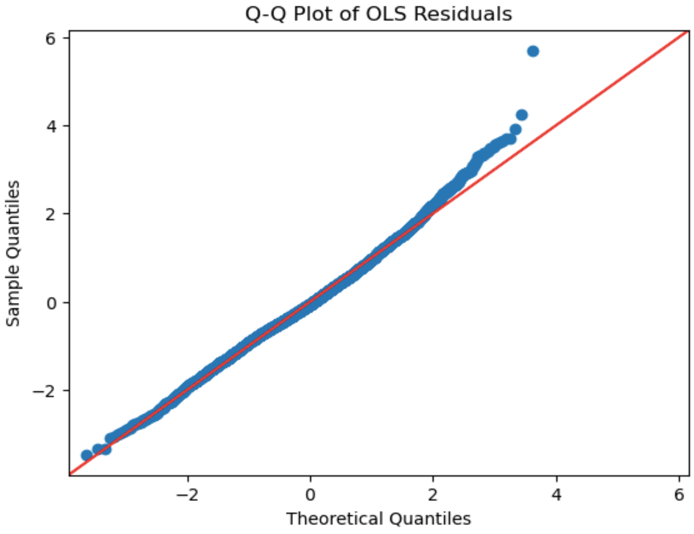
Winter: Lowest demand.

This pattern suggests that warmer weather and longer daylight hours in summer significantly boost bike rental usage, while colder winter conditions lead to a substantial drop in demand. Understanding these seasonal trends is crucial for optimizing bike-sharing operations and planning resource allocation effectively.

### OLS Model

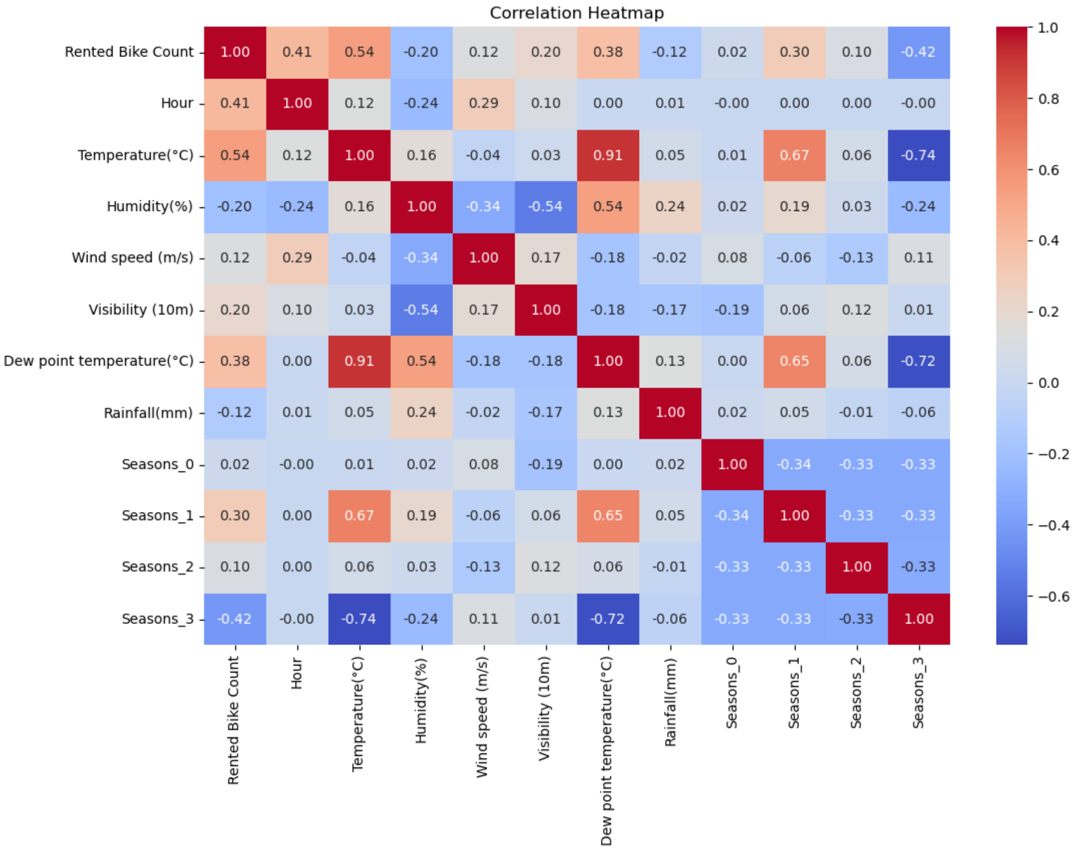
In this analysis, we aimed to predict the number of bike rentals in Seoul using an Ordinary Least Squares (OLS) regression model, addressing data preprocessing and feature selection issues. We first identified and encoded categorical variables, split the dataset into training and testing sets, and scaled the features. Using GridSearchCV, we optimized a Linear Regression model and evaluated its performance with RMSE and R² scores. To improve the model, we applied log and square root transformations to the target variable to address skewness and heteroscedasticity. The log-transformed model showed the best performance with the highest R² value and balanced residual plots. To address multicollinearity, we implemented backward elimination by iteratively removing features with the highest p-values, starting with 'Solar Radiation' and then 'Snowfall'. Despite these adjustments, residual and Q-Q plots indicated persistent issues, which are shown in the graphs below. Recognizing the limitations of backward elimination, we concluded that applying regularization models like Lasso or Ridge Regression might better manage multicollinearity and improve predictive power. This comprehensive approach provided a robust framework for predicting bike rentals with enhanced accuracy and interpretability.





## Correlation Analysis

We also wanted to investigate the correlation between our variables before running predictive models. The correlation matrix shown below provided valuable insights into the relationships between different variables and the target variable, 'Rented Bike Count.'



Key findings include:

Positively correlated with:

Temperature (0.54): Higher temperatures correspond to increased bike rentals.

Hour (0.41): Certain hours of the day show higher bike rental activity.

Dew Point Temperature (0.38): More humid air correlates with higher bike rentals.

Seasons_1 (0.30): Likely represents a season which is Summer with higher bike rental activity.

Negatively correlated with:

Seasons_3 (-0.42): Likely represents a season which is Winter with lower bike rental activity.

Humidity (-0.20): Higher humidity levels are associated with fewer bike rentals.

### Model Performance Comparison

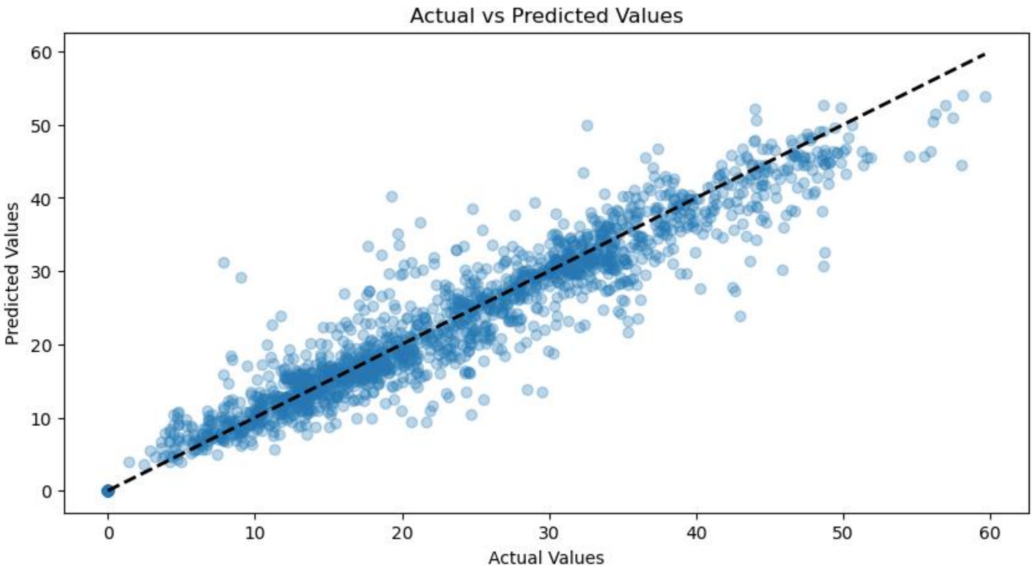
We compared the performance of different regression models using RMSE and R-squared metrics. The results are summarized as follows:

| Model | RMSE | R² |
|---|---|---|
| Linear Regression | 7.103216 | 0.671092 |
| Decision Tree | 4.793213 | 0.850232 |
| Random Forest | 3.804972 | 0.905623 |
| Lasso Regression | 7.066286 | 0.674503 |
| Ridge Regression | 7.027445 | 0.678071 |

The Random Forest Regression model outperformed other regression techniques in accurately predicting bike-sharing demand, demonstrating the efficacy of ensemble learning methods for complex predictive tasks. The systematic evaluation of model performance using RMSE and R-squared scores provided quantifiable metrics for assessing model efficacy and guiding future model refinement efforts.

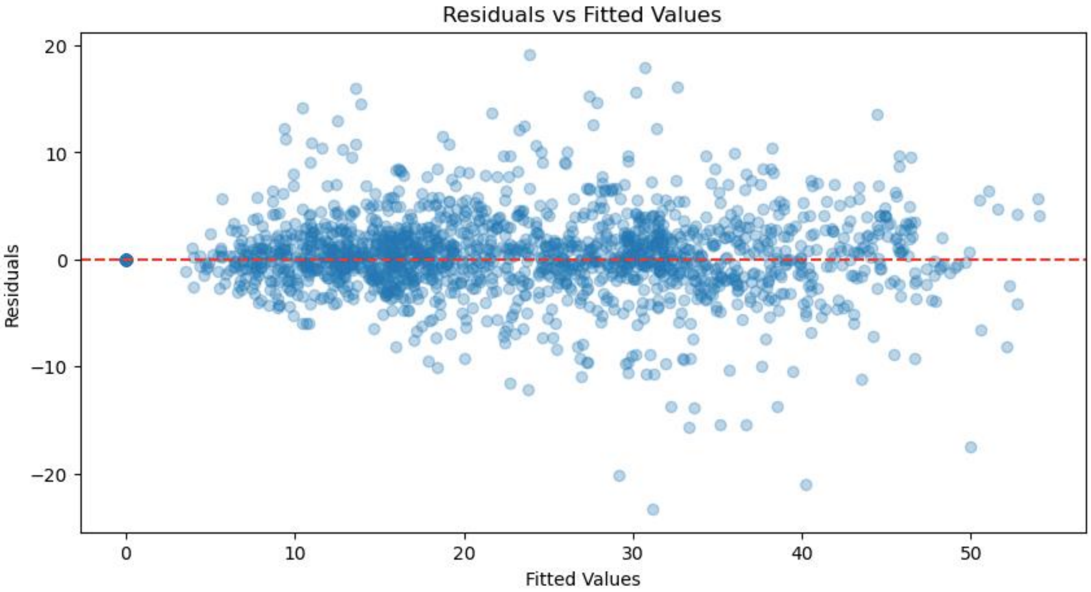
### Actual Versus Predicted – Random Forest

This scatter plot shows that the model generally predicts well, with most predicted values being close to the actual values. However, the prediction error appears to increase with larger actual values, and there are some notable outliers.



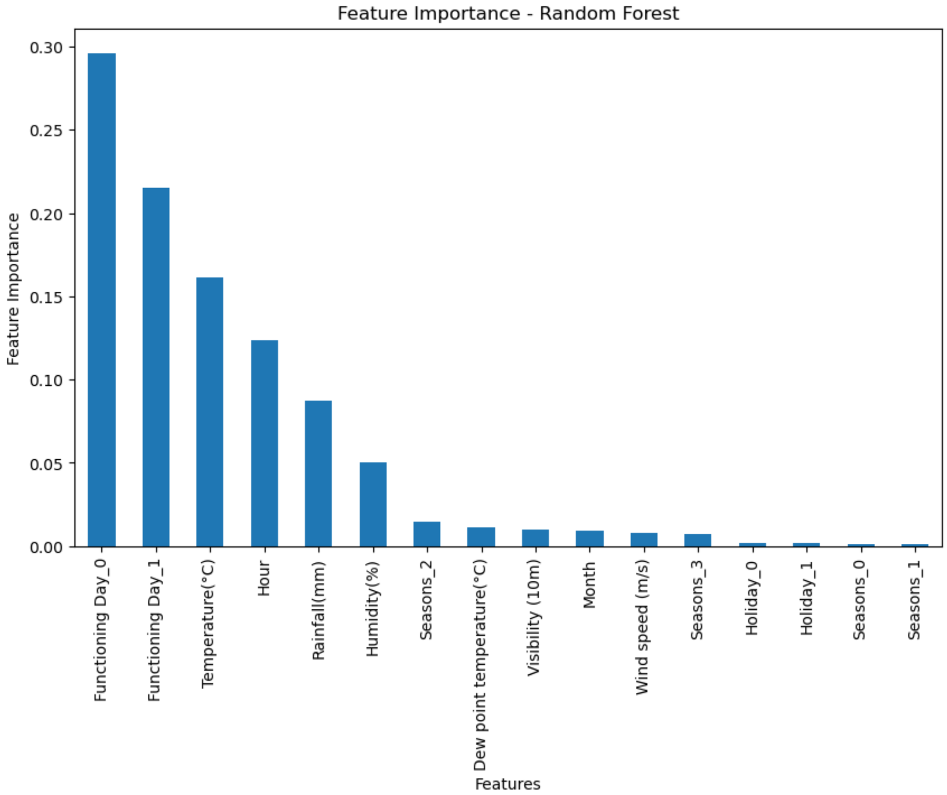
### Residual Versus Fitted Values - Random Forest

The plot suggests that the model's predictions are reasonably good. The residuals are mostly randomly dispersed and centered around zero with a constant spread, indicating that the model captures the relationship in the data well and does not suffer from major issues like heteroscedasticity or significant non-linearity. However, there is a presence of notable outliers.



### Feature Importance Analysis

The analysis of feature importance using the Random Forest model revealed that features like 'Functioning Day' and 'Temperature' are highly predictive of bike rental demand, underscoring the importance of operational factors and weather conditions in driving user engagement. These insights can inform strategic decision-making and marketing efforts aimed at maximizing user satisfaction and service utilization.



In conclusion, the project's comprehensive analysis and implementation of various machine learning models provided actionable insights into factors influencing bike rental demand. These insights can guide stakeholders in making informed decisions to enhance the efficiency and user satisfaction of bike-sharing services.

## FUTURE SCOPE

Reflecting on the current achievements of our project, we identify several promising directions for future research and improvement, specifically tailored to the characteristics and scope of the Seoul Bike Sharing Demand dataset. With additional time and resources, the following avenues could be explored to enhance predictive accuracy and operational effectiveness:

### Incorporation of Additional Features

To refine the predictive models, we could explore incorporating additional external data sources, such as socioeconomic factors, traffic conditions, and special events. Although our current dataset includes comprehensive weather and temporal data, integrating these additional factors could provide a more nuanced understanding of demand drivers and improve model performance.

### Enhanced Predictive Models

Future work could focus on the continued refinement and optimization of machine learning models. Experimenting with advanced algorithms such as neural networks, or ensemble methods could enhance predictive accuracy. More extensive hyperparameter tuning could also be conducted to further optimize model configurations.

### Dynamic Demand Forecasting

Implementing real-time data integration could enable dynamic demand forecasting. By incorporating live weather updates, traffic conditions, and event schedules, our model could make more accurate and timely predictions. This real-time analysis would allow for more responsive and adaptive bike-sharing services, reducing waiting times and improving user satisfaction.

### Customer Segmentation

Applying clustering algorithms to identify distinct customer segments based on usage patterns and preferences could facilitate targeted marketing efforts and personalized services. Understanding different user groups, such as daily commuters, occasional riders, and tourists, could help tailor promotions, pricing strategies, and service improvements to better meet the needs of each segment.

### Integration of External Data

Incorporating additional external data sources, such as public event calendars, and real-time traffic patterns, could enhance prediction accuracy and provide deeper insights into demand fluctuations. This comprehensive data integration would support more informed decision-making for operational adjustments and strategic planning.

### Advanced Forecasting Techniques

Implementing advanced forecasting techniques, such as time series analysis and seasonal decomposition, could improve the model's ability to predict future demand patterns. Techniques like ARIMA (Auto-Regressive Integrated Moving Average) or LSTM (Long Short-Term Memory) networks, specifically designed for time series data, could be particularly beneficial.

### Expansion of Service Offerings

Exploring opportunities for expanding service offerings could significantly benefit the bike-sharing program. For example, introducing electric bikes could attract a broader user base, including those who may find traditional biking challenging. Additionally, partnerships with other transportation providers, such as public transit systems or ride-sharing services, could create a more integrated and convenient urban mobility solution.

### User Feedback Integration

Incorporating user feedback mechanisms into the system could provide valuable insights into user satisfaction and areas for improvement. Analyzing feedback data could help identify pain points and areas where service adjustments are needed, leading to a more user-centric approach to bike-sharing services.

By pursuing these avenues, future work can build upon the foundation laid by this project, driving continuous improvements in the accuracy and effectiveness of bike-sharing demand prediction models. These enhancements would not only optimize operational efficiency but also improve user satisfaction and the overall sustainability of bike-sharing programs.

## CONCLUSION

This bike-sharing demand prediction project highlights the transformative potential of data-driven insights in enhancing urban transportation systems. By analyzing the "Seoul Bike Sharing Demand" dataset, we identified key factors such as temperature, functional days, and seasons that influence bike rental patterns. Using various machine learning models, the Random Forest model emerged as the most effective for accurate forecasting. This project underscores the importance of data preprocessing and feature engineering and demonstrates how accurate demand prediction can optimize resource allocation, improve user satisfaction, and promote sustainable urban mobility. Future research can refine predictions with additional features, real-time data, and customer segmentation, ensuring bike-sharing systems remain vital in modern urban transportation.