Predicting Seoul Bike Rental Demand Based on Weather and Temporal Factors

Miles A. Peña

Bellevue University

June 1, 2025

Table of Contents

Abstract

This paper presents a machine learning approach to predicting bike rental demand in Seoul, South Korea. By using publicly available data on hourly bike rentals along with environmental and temporal variables, this project aims to identify key patterns and train predictive models to improve resource allocation in bike-sharing systems. The methods employed include exploratory data analysis (EDA), regression-based modeling, time-series analysis, and feature engineering. Initial results indicate significant correlations between temperature, time of day, and bike usage, with machine learning models demonstrating strong predictive performance. The paper concludes with recommendations for practical deployment and future use cases.

Introduction

Bike-sharing programs have become an essential part of modern, sustainable urban transportation systems. In Seoul, accurately predicting hourly rental demand can minimize idle inventory, boost availability, and enhance rider satisfaction. Traditional planning methods often fail to account for rapidly changing weather or commuting patterns. This project employs a data-driven approach to analyze how weather conditions and time-based trends impact rental volumes and utilizes machine learning to forecast future demand.

Business Problem

Seoul's bike-sharing system struggles with operational inefficiencies due to demand fluctuations affected by temperature, precipitation, holidays, and time of day. Over- or under-supplying bikes at various times and locations can diminish customer experience and operational cost efficiency. This project aims to establish a reliable predictive model to support real-time and scheduled resource allocation decisions.

Background / History

The Seoul Bike Sharing dataset offers a comprehensive view of hourly rental counts from 2017 to 2018, along with corresponding weather and date-related factors. Similar studies have demonstrated the value of predictive analytics in transportation planning but often lack hyper-localized weather detail or temporal depth. By leveraging hourly granularity and diverse features, this project provides an opportunity to create targeted forecasts for urban planning and operations.

Data Explanation

The primary dataset includes hourly rental counts and weather data for Seoul. The target variable is the Rented Bike Count. The dataset consists of date and time information, including the Date and Hour columns, as well as environmental features such as Temperature, Humidity, Wind Speed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, and Snowfall. Additionally, it contains temporal attributes like Seasons, Holiday, and Functioning Day.

Data preprocessing involved converting the Date column to day-first formatting to align with the dataset's structure. The Hour column was parsed into integer values. Missing values were addressed by filling them with the mean of each column. Weather variables were normalized using the StandardScaler from scikit-learn to ensure equal weight across features.

Feature engineering involved creating an AvgHourlyRental variable that represents the average number of rentals per hour, along with a TempBand variable that categorizes temperature values into labeled ranges from Very Cold to Hot.

Methods

This project utilized a standard machine learning pipeline. Initially, exploratory data analysis (EDA) was performed using line plots, heatmaps, and summary statistics to unveil seasonal and hourly rental patterns. The subsequent step involved training predictive models, including Random Forest, Gradient Boosting, and Linear Regression. Additionally, a time-series model, ARIMA, was trained to forecast overall trends. Models were assessed using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to compare accuracy.

Analysis

EDA indicated that rentals increased significantly during commute hours, particularly in the morning and late afternoon. Rentals were also higher on mild and warm days compared to cold or extremely hot days. Holidays showed a correlation with reduced rentals. Boxplots, scatter plots, and heatmaps illustrated that temperature, time of day, and type of day meaningfully influence rental patterns. Among the models tested, Random Forest and Gradient Boosting outperformed Linear Regression in terms of RMSE and MAE. The ARIMA model effectively captured overall temporal trends but lacked short-term sensitivity.

Conclusion

The final models showed strong predictive accuracy, confirming that weather and temporal features are valuable inputs for demand forecasting. These predictions can assist operators in redistributing bikes more effectively, anticipating demand spikes, and improving plans for adverse weather conditions.

Assumptions

This project assumes that weather readings reflect conditions throughout Seoul. It also assumes that demand patterns remain relatively stable from year to year and that the Functioning Day and Holiday columns are accurate and reliable.

Limitations

The dataset lacks geographic information, which would allow for station-level predictions. It only includes data from 2017 to 2018, limiting long-term trend analysis. Moreover, there is no distinction between commuter and leisure riders, which could enhance modeling.

Challenges

A primary challenge was correcting the non-standard date formats in the dataset. Ensuring numerical stability during feature scaling was also critical, especially given the varying scales of environmental features. Feature engineering, such as calculating hourly averages, had to be carefully constructed to avoid data leakage.

Future Uses / Additional Applications

This forecasting framework can be applied to other urban areas with similar datasets. Future enhancements may include the integration of real-time weather APIs, longer-term datasets, and geospatial details to facilitate neighborhood-level planning.

Recommendations

The predictions should be integrated into a daily dashboard utilized by dispatch and operations teams. These outputs could be paired with heatmaps for visual insights. It is advisable to retrain models weekly to include recent rental and weather data.

Implementation Plan

Implementing this would involve training and validating models using historical data, automating the intake of weather and rental data, and deploying the model through an interface such as Tableau or Streamlit.

Ethical Assessment

The dataset is anonymized and does not contain any personally identifiable information. All predictions are made at an aggregate level. Care must be taken to ensure that model-driven decisions do not unintentionally restrict services in low-income or underserved areas.

References

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

NewsAPI. (2024). *NewsAPI.org: JSON News Data API for Live News and Blog Articles*. Retrieved from https://newsapi.org

Ran, A. (2019). *yfinance: Yahoo! Finance market data downloader* [Python package]. Retrieved from https://pypi.org/project/yfinance/

Appendix A: Visualizations

Figure 1. Bike Rentals by Hour

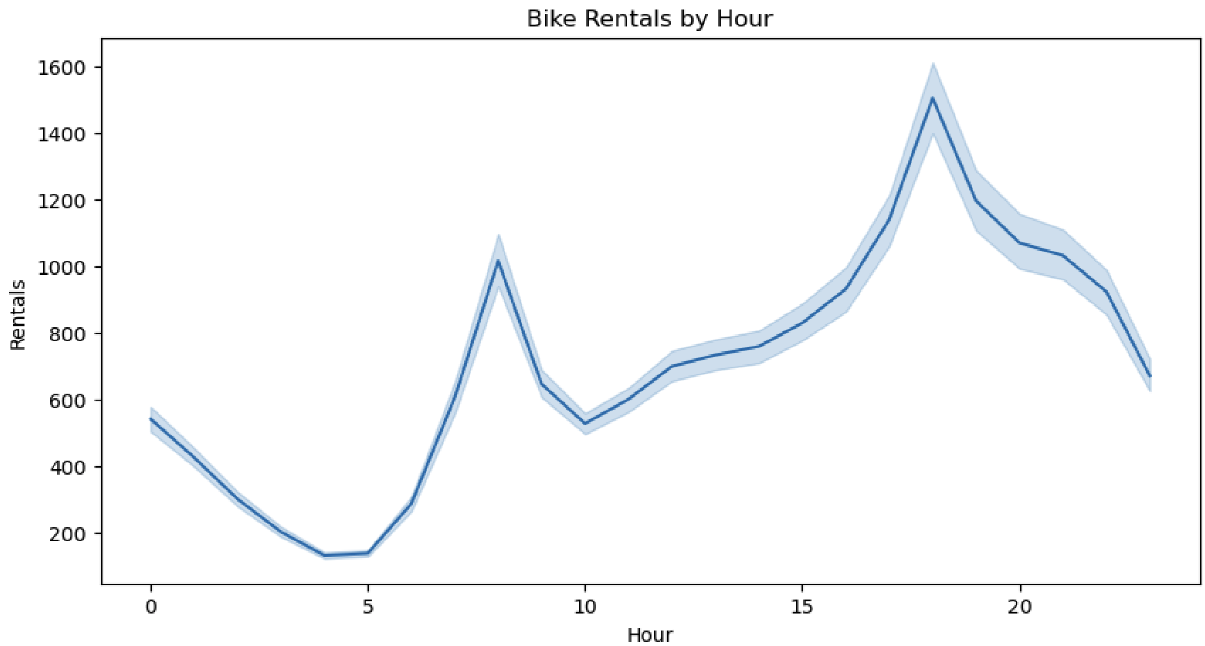


Figure 2. Heatmap of Bike Rentals

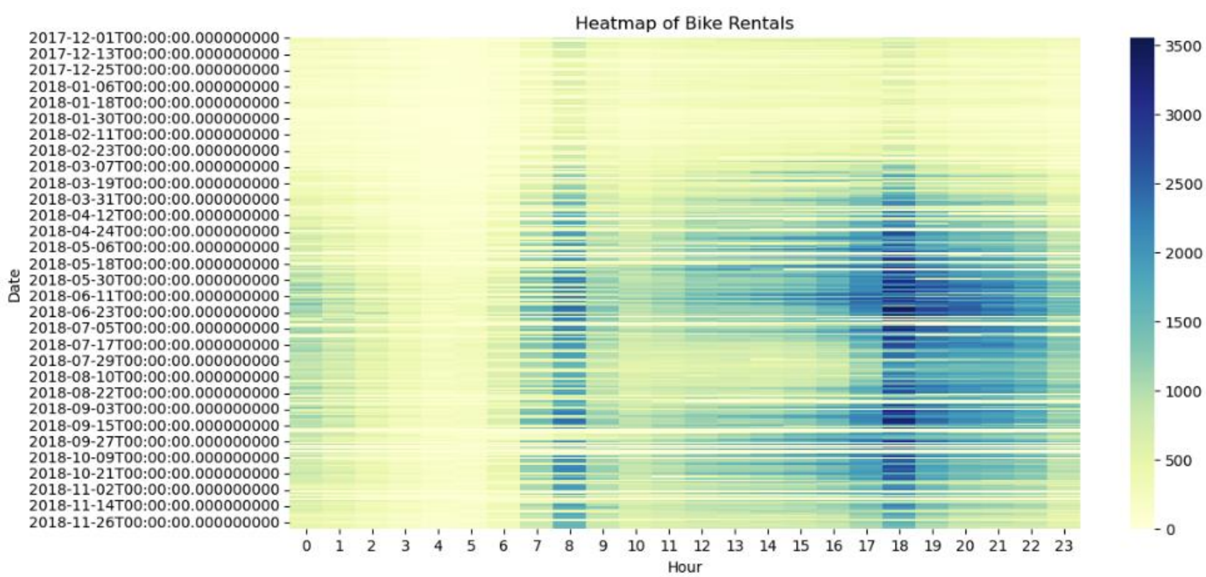


Figure 3. Seasonal Trend of Bike Rentals

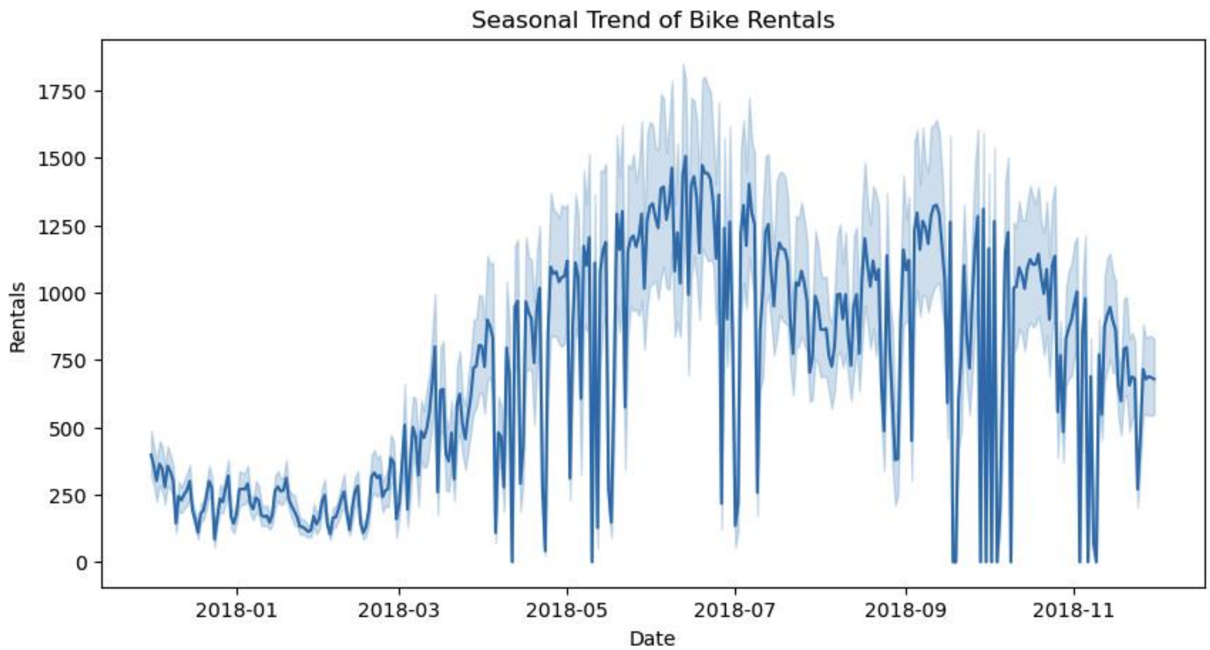


Figure 4. Correlation Matrix of Numerical Features

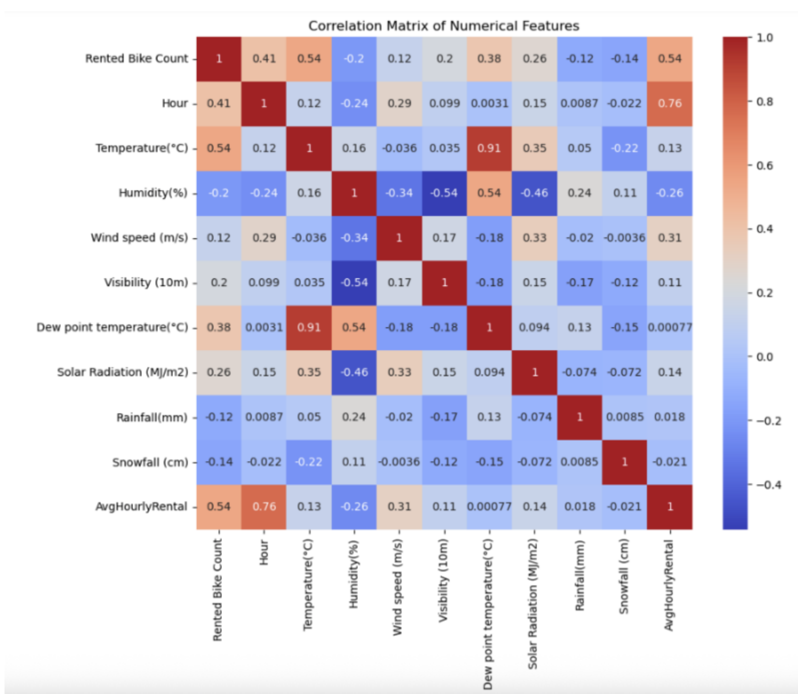


Figure 5. Bike Rentals by Temperature Band

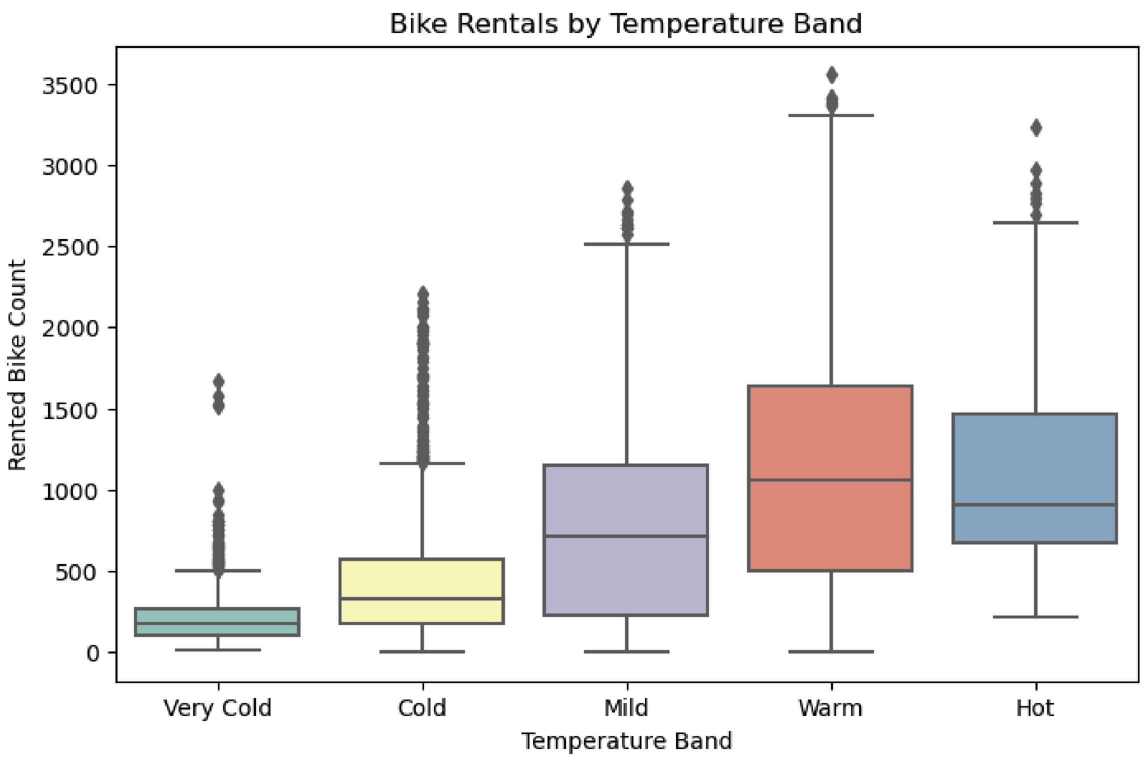


Figure 6. Average Bike Rentals: Holiday vs Non-Holiday

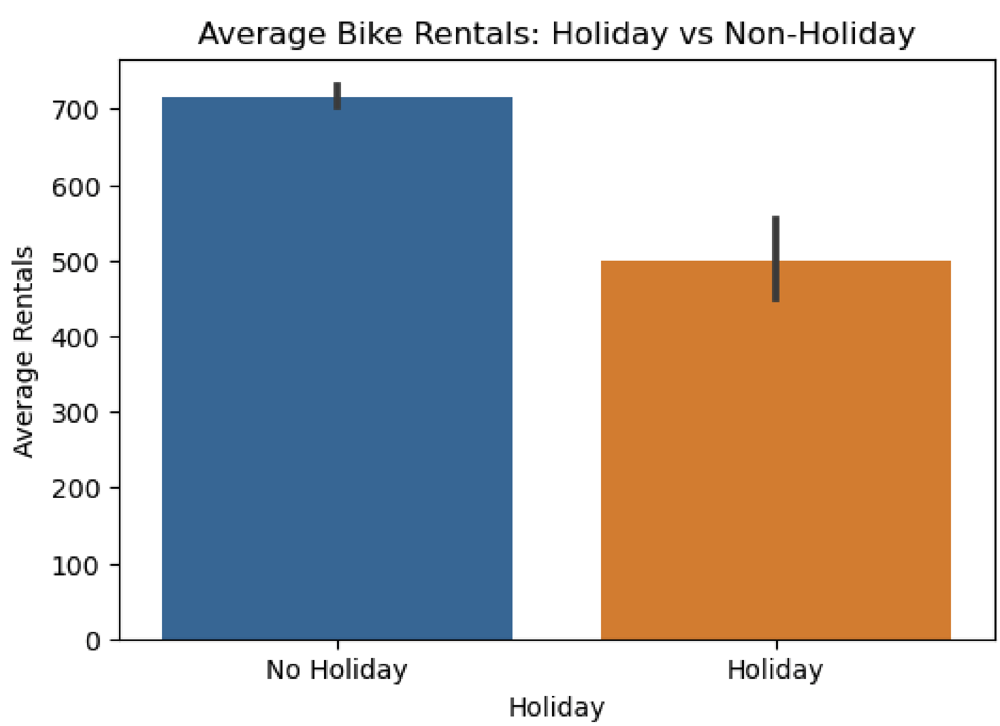


Figure 7. Average Rentals by Day of the Week

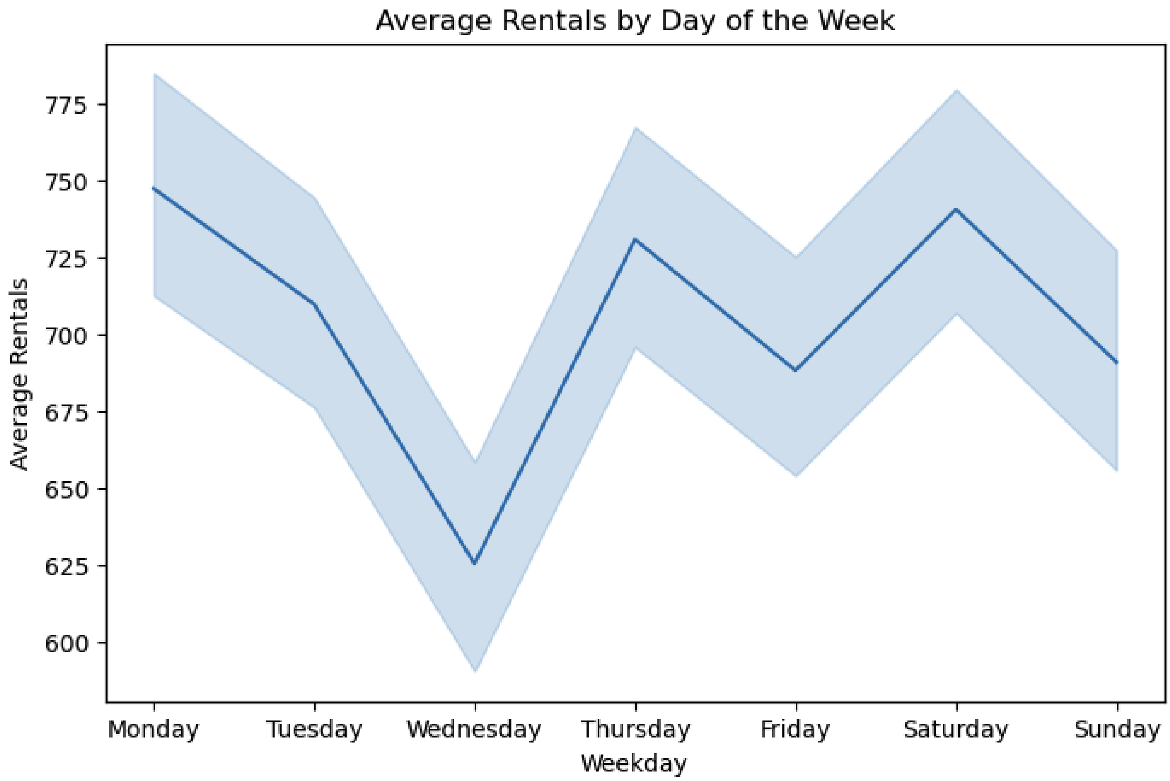


Figure 8. Rentals vs Temperature

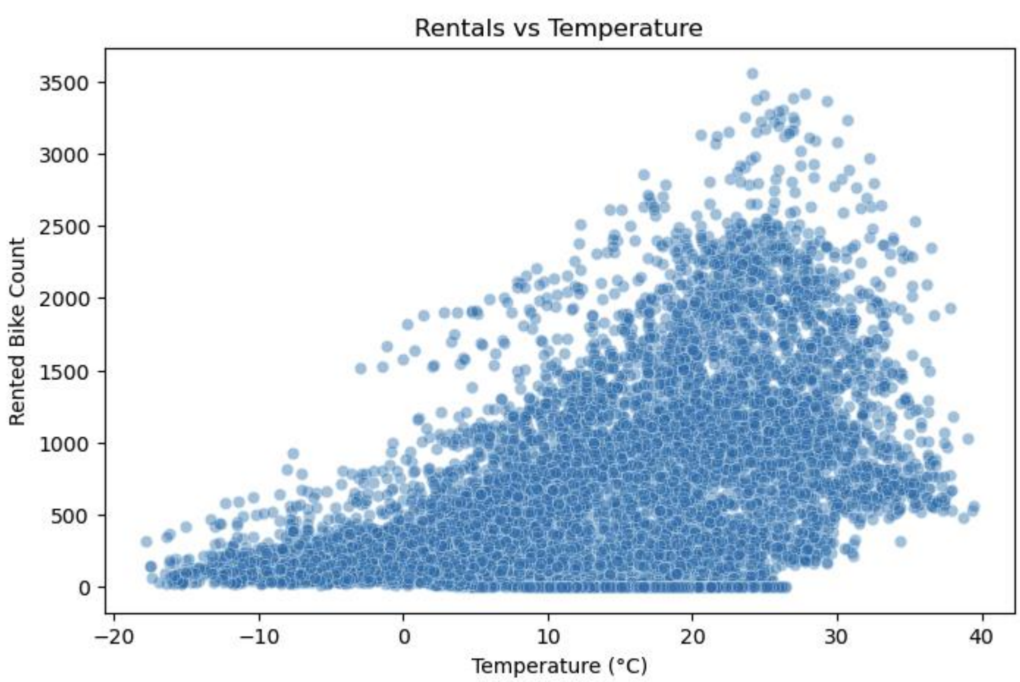
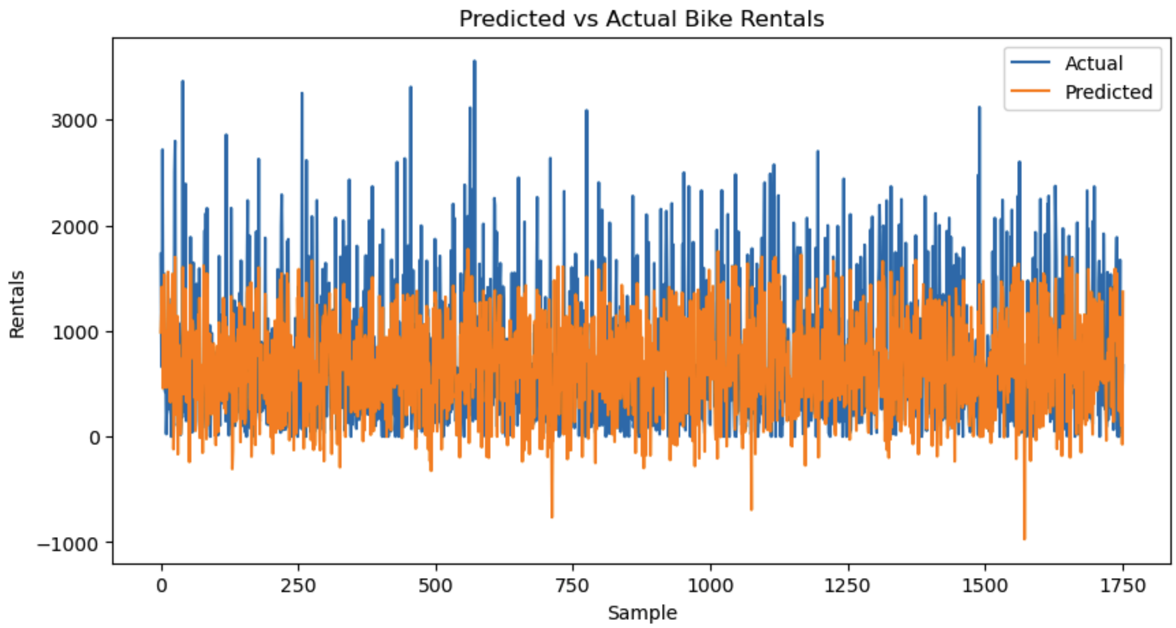


Figure 9. Predicted vs Actual Bike Rentals



Appendix B: Audience Questions

1. What is needed to transition from pilot to full implementation?
2. How does the model handle unexpected weather shifts?
3. Would geospatial data improve prediction accuracy?
4. How does holiday demand differ from weekends?
5. Can these forecasts integrate with mobile apps?
6. What insights can temperature bands provide to operators?
7. Could prediction errors affect user experience?
8. How frequently should the model be retrained?
9. Are the models explainable enough for non-technical users?
10. How can this project be expanded to real-time use?
11. What would be required to scale this citywide?