STAT 230A Project Proposal
Predicting Hourly Bike Sharing Demand in Seoul

Yirong Huo, Hao Wang

April 10, 2025

1 Introduction

Urban bike-sharing systems are vital for sustainable mobility in Seoul, yet operational challenges persist despite 27% annual ridership growth (Seoul Metropolitan Government, 2022). A 15% dissatisfaction rate from station imbalances during peaks highlights the urgency of demand forecasting. Leveraging Seoul TOPIS's hourly dataset with weather integration, this study advances prior daily-scale analyses by developing a regression framework quantifying temperature, precipitation, temporal cycles, and holiday impacts on ridership. The model addresses three city priorities: (1) optimizing bike redistribution via peak-demand identification, (2) weather-resilient station placement strategies, and (3) climate adaptation planning for temperature extremes. Unlike black-box ML methods, linear regression explicitly quantifies marginal effects—e.g., translating 1°C temperature increases to fleet expansion costs.

2 Data

Our analysis is based on the seo (2020) dataset, which records hourly bike rental information across the Seoul metropolitan area. The data spans the full calendar year of 2017, comprising 8,760 observations—one for each hour. Each record includes both the number of bikes rented (our response variable) and a rich set of explanatory variables, making this dataset well-suited for regression analysis.

Response Variable:

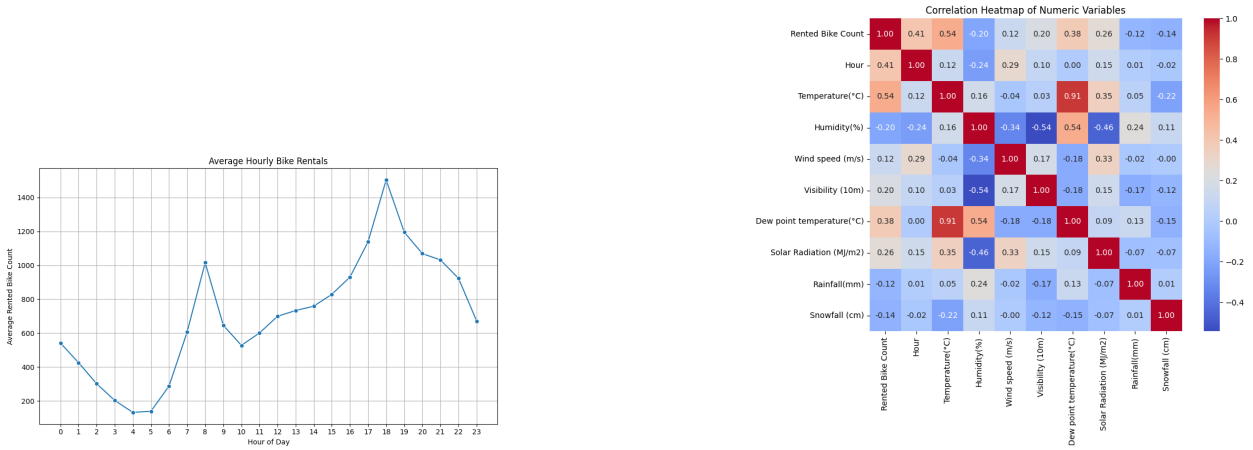
- **Rented Bike Count**: A continuous variable representing the total number of bikes rented across all stations in Seoul during a given hour.

Covariates:

- **Hour (0–23)**: Categorical indicator for the hour of the day, capturing daily usage cycles.
- **Seasons**: Categorical variable (Winter, Spring, Summer, Fall), reflecting long-term climate effects.
- **Holiday**: Binary variable indicating whether the day is a public holiday.
- **Functioning Day**: Binary indicator for whether the day is a regular working day.
- **Temperature (°C)**: Ambient temperature.
- **Humidity (%)**: Relative humidity percentage.
- **Wind Speed (m/s)**: Speed of wind at the time of recording.
- **Rainfall (mm)**: Amount of rain, an important factor in determining user willingness to ride.
- **Snowfall (cm)**: Accumulation of snow, particularly relevant in winter.
- **Solar Radiation (MJ/m^2)**: A proxy for sunlight exposure, which may influence outdoor activity levels.
- **Date**: Recorded date allows for further derivation of features such as day of the week or month if needed.

Preliminary exploratory analysis via `pandas` shows clear hourly demand cycles (Figure 1), with peaks during

commute times, and notable variation across seasons and weather conditions. For example, demand tends to decrease significantly during heavy rain or snowfall (Figure 2), while higher temperatures and solar radiation are associated with increased rentals. These patterns guide our variable selection and modeling strategy in the next section.



(a) Average Hourly Bike Rentals in Seoul

(b) Correlation Heatmap

Figure 1: Combined Visualization of Bike Rental Patterns

3 Regression Analysis Plan

The analytical approach centers on establishing interpretable relationships between bike rentals and multimodal predictors through parametric regression. The primary specification adopts multiple linear regression to quantify marginal effects, formally expressed as:

$$\text{Bike Count}_t = \beta_0 + \beta_1 \text{Temp}_t + \beta_2 \text{Hour}_t + \beta_3 \text{Rain}_t + \sum_{s=1}^{3} \gamma_s \text{Season}_{s,t} + \epsilon_t$$

where temporal effects are encoded through categorical hour indicators and seasonal dummies. Should residual diagnostics reveal count data characteristics (e.g., overdispersion), we will transition to Poisson regression with exposure-adjusted offsets. For temporal dependence concerns, an ARIMA-X framework incorporating lagged residuals and exogenous weather variables will be evaluated through ACF/PACF analysis.

Model integrity will be safeguarded through three diagnostic protocols: variance inflation factors (VIF) to detect multicollinearity with threshold-triggered ridge regularization, Breusch-Pagan testing for heteroscedasticity addressed via robust standard errors, and time-aware cross-validation splitting the 2017 timeline into training (Jan-Nov) and out-of-sample testing (Dec).

4 Anticipated Challenges

Methodological constraints arise from the data's operational nature. Zero-inflation during overnight hours challenges linear assumptions, requiring truncated models. Non-linear thermal effects (e.g., reduced ridership at temperature extremes) demand polynomial specifications. Furthermore, station maintenance closures induce censored demand observations that risk selection bias unless explicitly modeled. Each challenge will be addressed through targeted model extensions while preserving interpretability for policy applications.

References

(2020). Seoul Bike Sharing Demand. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5F62R.