W271 Group Lab

Bike share demand

Marisa Lenci Ahsin Saleem Justin Sterling Scott Stossel Lyn Wang

1 Introduction

In Seoul, South Korea, there exist a wide variety of factors that determine bike rental demand. The time of day, whether it is a weekday or weekend, and the weather among other factors all can impact demand. In order to maximize profit, bike rental businesses should use these variables to optimize the availability of rental bikes for the public. Tasks such as maintenance and logistics can be scheduled during periods of low demand so that busier times do not experience any shortages of supply.

We hypothesize that time of day, season, day of the week, and weather all influence the usage and demand of rental bikes, with the most important being the first three. While weather plays an important role in rental bike usage, time of day, season, and day of the week are consistent events that businesses can reliably plan around. Understanding how these factors influence demand can help businesses best optimize their bike availability based on known usage patterns.

H0: Time of day, season, day of the week, and temperature have no effect on the number of bikes rented.

Ha: != 0

2 Description of Data

The data set comes from the Seoul Bike Sharing System, also known as Ddareungi, with corresponding weather related data along with holiday information. Each row corresponds to one hour of the following information in the format: Date (DD/MM/YYYY), Rented Bike Count (number of bikes rented that hour), Hour (0-23), Temperature (℃), Humidity (%), Wind speed (m/s), Visibility (10 m), Dew point temperature (℃), Solar Radiation (MJ/m2), Rainfall (mm), Snowfall (cm), Seasons (Winter, Spring, Summer, or Autumn), Holiday (No Holiday or Holiday), and Functioning Day (Yes or No).

PLACEHOLDER FOR DATA DESCRIPTION TABLE

The Seoul Bike Sharing System automatically records information upon any rental at every one of their stations. This dataset spans every hour from December 2017 to November 2018. Since every rental event is recorded this data represents a city-wide census of bike sharing information rather than a sample. The sampling is not strictly identically distributed as different seasons exhibit varying distributions of bike rentals. In addition, the samples are not exactly independent due to temporal patterns.

3 EDA

The dataset contains no missing values, and each row is uniquely defined by a combination of date and hour. We filtered out non-functioning days to focus only on operational periods. The distribution of bike Rentals during each hour is highly right-skewed, using a log transformation on our dependent variable might be appropriate as it improves model interpretability and reduces the influence of extreme values (Figure 1).

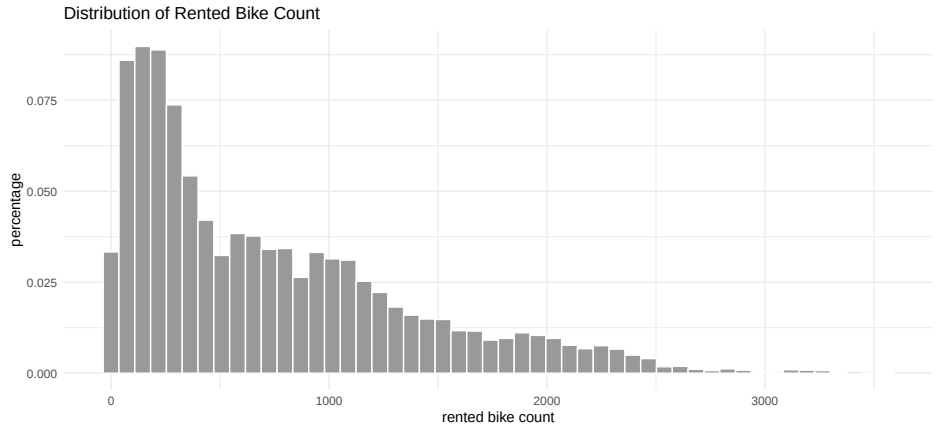

Figure 1: Distribution of Bike Count

For data transformation, we extracted 'is_weekday' to distinguish weekday vs weekends, and created two variables based on whether the hour falls in a rush period 'is_rush' and whether it is daytime hours 'is_day' to help us understand the usage behavior during the day. Specifically, Rush Hour is defined as 6-9 AM and 5-7 PM, which is aimed at capturing the outlier demand. Defining day vs night helps to differentiate day vs night behavior.

To test the core part of our hypothesis—that time-related variables explain bike demand—we began by plotting average rentals by hour, segmented by weekday/weekend. We observe a clear surge in demand during rush hours on weekdays, suggesting strong commuter usage (Figure 2 RHS). Outside of these rush hours, demand gradually increases from early morning (~5 AM) until ~7 PM on both weekdays and weekends, showing a generally consistent usage pattern. (Figure 2 LHS) We further confirmed this with a bar plot that compares demand across: Rush vs. non-rush hours and Weekdays vs. weekends. The effect of rush hour is visibly stronger than the day-of-week difference, supporting our hypothesis that time of day is a primary driver of demand.

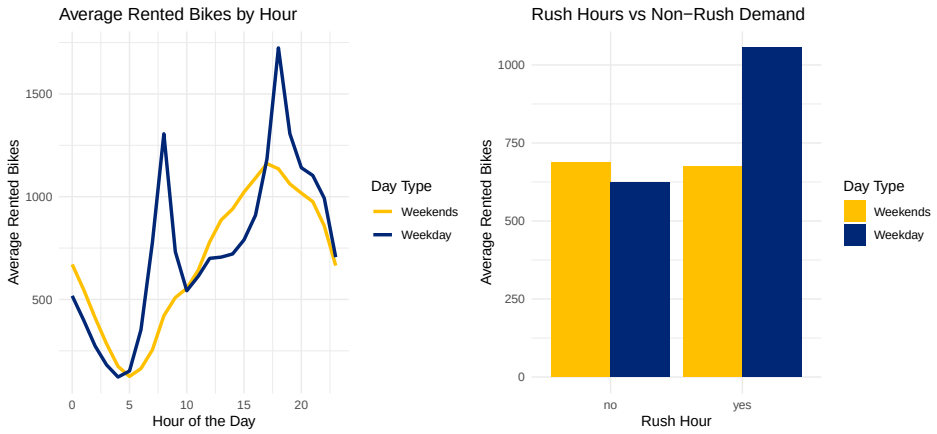

Figure 2: Rush Hour Bike Demand

We then examined seasonal trends, which are also important for business planning (e.g., maintenance scheduling). As we can see from this bar chart, average demand is much higher in Autumn, Spring, and Summer, compared to the average demand in Winter. We included 'is_holiday' as a control variable to ensure that holiday-driven fluctuations aren't misattributed to seasonal effects (Figure 4) This confirms that season is an important—but stable—factor for planning purposes.

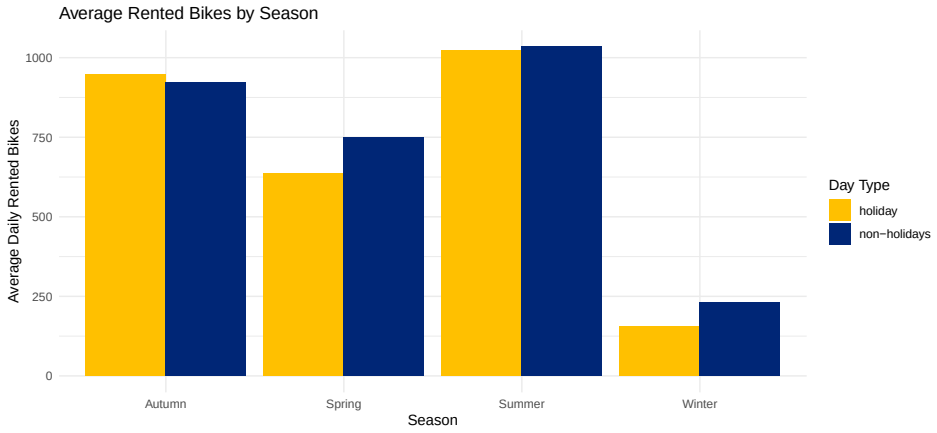

Figure 3: Seasonal Factor

Lastly, we explored weather variables including: Temperature, Rainfall (mm) and Snowfall (cm). These are expected to correlate with seasonal patterns and could have short-term effects on demand. While weather can strongly influence bike rentals (e.g., very low demand on rainy days), it's also less predictable than time or calendar-based variables. Therefore, it plays a secondary role in our business-centric planning framework. An interesting observation is the nonlinear relationship between temperature and bike demand. As temperature increases, demand initially rises—likely due to more comfortable riding conditions—but after reaching a certain threshold, demand drops off sharply, possibly due to heat-related discomfort (Figure 4).

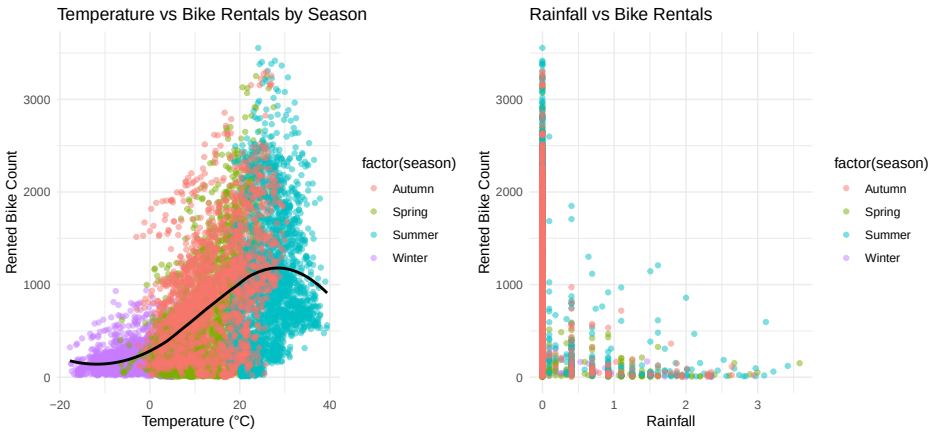

Figure 4: Weather Factors

In contrast to temperature, rainfall shows a consistently negative association with bike usage. Demand tends to decrease sharply as rainfall increases, with the exception of a few summer hours where short bursts of rain may not significantly deter riders. This highlights that while moderate temperatures promote ridership, extreme weather—either too hot or too wet—discourages usage.

4 Model Development

4.1 Poisson regression

[!h]

Table 1: Model 1 Confidence Intervals

	Estimate	CI_Lower	CI_Upper
(Intercept)	122.563	122.118	123.009
is_weekdayyes	1.111	1.109	1.113
hour	1.052	1.052	1.052
temperature_c	1.032	1.032	1.033
log1p(rainfall_mm)	0.150	0.148	0.151
log1p(snowfall_cm)	0.700	0.694	0.705
seasonAutumn	2.368	2.359	2.378
seasonSpring	2.008	2.000	2.016
seasonSummer	1.820	1.811	1.829

The Poisson regression model1 was specified to predict bike rental counts using weekday status, hour of day, temperature, rainfall, snowfall, and season. All variables in the model were found to be highly statistically significant, with p-values less than 0.001 and strong likelihood ratio chi-squared values, indicating each contributes meaningfully to the model. Interpreting the coefficients in multiplicative terms and holding other variables constant, we see that being a weekday increases expected rentals by approximately 11%, each additional hour in the day increases rentals by about 5.2%, and each one-degree Celsius increase in temperature boosts rentals by roughly 3.25%. In contrast, a 1% increase in rainfall leads to about a 1.9% drop in expected bike rentals. A 1% increase in snowfall leads to about a 0.36% drop in expected bike rentals. Even small increases in rain or snow can cause

noticeable declines in bike usage. Seasonal effects are also strong: compared to winter, expected rentals in autumn, spring, and summer are higher by 137%, 101%, and 82%, respectively. These results align with practical expectations about biking behavior, such as higher demand during good weather and workdays.

Furthermore, none of the coefficients have confidence intervals that include zero, which supports their statistical significance. This conclusion is based on the extremely small p-values and large z-statistics observed. Similar calculations for all other variables confirm the same, reinforcing the reliability of each predictor. Overall, the model demonstrates both strong statistical fit and meaningful practical insights. However, to make the model more robust, we should include interaction terms to capture how certain variables influence each other. For example, while temperature currently shows a positive overall effect on bike rentals, this relationship likely varies by season. In summer, higher temperatures can make biking uncomfortable, potentially reducing rentals, whereas in winter, warmer temperatures might encourage more people to ride. Including interactions like these would provide a more nuanced understanding of rider behavior.

4.2 Model Comparison

[!h]

Table 2: Model 3 Confidence Intervals

	Estimate	CI_Lower	CI_Upper
(Intercept)	148.327	147.142	149.521
is_weekdayyes	1.020	1.017	1.024
hour	1.049	1.049	1.049
is_dayyes	0.907	0.903	0.911
is_rush_houryes	1.087	1.083	1.092
temperature_c	1.050	1.049	1.051
log1p(rainfall_mm)	0.194	0.192	0.196
log1p(snowfall_cm)	0.836	0.830	0.843
seasonAutumn	2.453	2.442	2.465
seasonSpring	1.399	1.392	1.406
seasonSummer	9.157	9.076	9.239
humidity	0.992	0.992	0.992
wind_speed_m_s	1.007	1.006	1.008
visibility_10m	1.000	1.000	1.000
solar_radiation_mj_m2	1.076	1.074	1.077
holidayNo Holiday	1.221	1.216	1.227
is_weekdayyes:is_dayyes	0.866	0.862	0.870
is_weekdayyes:is_rush_houryes	1.851	1.843	1.860
temperature_c:seasonAutumn	0.983	0.982	0.983
temperature_c:seasonSpring	1.005	1.004	1.006
temperature_c:seasonSummer	0.929	0.928	0.929

All three information criteria residual deviance, AIC, and the likelihood ratio test consistently indicate that Model 3 is the best performing model, followed by Model 2, and then Model 1. Model 3 has the lowest residual deviance (1,041,590) and the lowest AIC (1,108,725), suggesting it fits the data significantly better than the other two models. Model 2, which includes more predictors than Model 1, also shows clear improvement, with a residual deviance of 1,306,308 and an AIC of

1,373,434 compared to Model 1's values of 1,644,270 and 1,711,382, respectively. The likelihood ratio tests further support these findings, showing that both Model 2 and Model 3 offer statistically significant improvements over simpler models, with p-values well below 0.001. These results are consistent because each added variable and interaction term in the more complex models helps explain more of the variation in bike rental counts without leading to overfitting. Overall, Model 3 provides the best balance of fit and complexity and is therefore the most robust among the three. The results are consistent because including more variables helps reduce omitted variable bias, allowing the model to better account for important factors that influence bike rentals. Adding interaction terms improves the model by capturing how the effect of one variable may depend on another. The fact that both the AIC and residual deviance decrease as variables and interactions are added confirms that the model fit is improving.

Practical findings from our best model (Model 3), highlight several key factors driving bike rental demand. Rentals are significantly higher in summer, though the interaction between temperature and summer shows that extreme heat can actually reduce usage, suggesting that warmer weather only helps up to a point. Rainfall has a strong negative impact, with even small increases leading to steep drops in rentals. Weekday rush hour nearly doubles demand, highlighting commuting as a major use case. Sunlight also increases rentals, while humidity and snowfall slightly reduce them. Winter shows the lowest demand overall, making it a strategic time for major fleet maintenance and system upgrades with minimal disruption to riders. Overall, the model shows that both environmental conditions and time-related factors strongly influence biking behavior, and interactions between them like season and temperature are essential for capturing real-world patterns.

5 Model Assessment

An assumption required for using Poisson distribution was violated in the bike rental dataset. Poisson random variables are assumed to be taken from a distribution that has a mean equal to the variance, but rented_bikes_count has a log mean of 6.59 (log(729)) and a log variance of 12.93 (log(412,615)), so it does not seem appropriate to use this method. We run a chi-square goodness of fit test on rented_bikes_count for confirmation. At any reasonable level of significance, we reject the null hypothesis that the data came from a poisson distribution. Since the variability of the rented bikes is greater than the average level we expect from a poisson random variable (variance > mean), our models exhibit overdispersion, resulting in underestimated standard errors and in turn overly narrow confidence intervals. This likely explains why all coefficients in our poisson models were statistically significant, even though this seems unlikely.

To address this problem, we first try using robust standard errors. The tables below compare results from each of the three models with and without robust standard errors. It does not appear that using robust standard errors is enough to correct for overdispersion in our model, since the coefficients do not change.

ADDED METHODS, TABLES TO MARKDOWN, will double check

As a more extreme measure, we change our modeling approach to use negative binomial regression, which introduces a dispersion parameter. We run all three models using this approach and compare the results to the original poisson models.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Jun 09, 2025 - 02:48:57 AM

Table 3: Poisson vs NegBin Model 3

	Dependent variable:	
	Poisson	*coefficient test*
	Poisson	NegBinomial
	(1)	(2)
is_weekdayyes	0.020***	−0.065***
	(0.002)	(0.021)
hour	0.048***	0.042***
	(0.0001)	(0.001)
is_dayyes	−0.098***	−0.294***
	(0.002)	(0.030)
is_rush_houryes	0.084***	0.076**
	(0.002)	(0.030)
temperature_c	0.049***	0.057***
	(0.0003)	(0.002)
log1p(rainfall_mm)	−1.641***	−1.069***
	(0.005)	(0.023)
log1p(snowfall_cm)	−0.179***	−0.116***
	(0.004)	(0.031)
seasonAutumn	0.897***	0.805***
	(0.002)	(0.030)
seasonSpring	0.336***	0.136***
	(0.003)	(0.030)
seasonSummer	2.215***	1.995***
	(0.005)	(0.073)
humidity	−0.008***	−0.010***
	(0.00003)	(0.0005)
wind_speed_m_s	0.007***	0.004
	(0.0005)	(0.007)
visibility_10m	0.00001***	0.00002
	(0.00000)	(0.00001)
solar_radiation_mj_m2	0.073***	0.137***
	(0.001)	(0.012)
holidayNo Holiday	0.200***	0.282***
	(0.002)	(0.028)
is_weekdayyes:is_dayyes	−0.144***	−0.008
	(0.002)	(0.031)
is_weekdayyes:is_rush_houryes	0.616***	0.785***
	(0.002)	(0.033)
temperature_c:seasonAutumn	−0.018***	−0.024***
	(0.0003)	(0.003)
temperature_c:seasonSpring	0.005***	0.004
	(0.0003)	(0.003)
temperature_c:seasonSummer	−0.074***	−0.077***
	(0.0003)	(0.004)
Constant	4.999***	5.191***
	(0.004)	(0.054)
Observations	8,465	
Log Likelihood	−554,341.600	

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Jun 09, 2025 - 02:48:57 AM

Table 4: Poisson vs NegBin Model 3

FALSE

Many of the coefficients are still significant in the negative binomial regression models, but each had to meet a higher threshold since this involved larger standard errors and wider confidence intervals. In general, it seems that negative binomial regression is a more appropriate model to use for the rented bike dataset since the variance of rented_bike_count substantially exceeds its mean.

6 Alternative Specification

The results from the OLS model indicate that the time of day, weather, and seasonality are strong predictors of bike demand - consistent with the direction of coefficients from Poisson models. Specifically, holding all else constant, on average each additional hour of the day is associated with approximately 29 more bike rentals, and being in a rush hour adds an additional 89 bike rentals on average. Temperature also has a significant effect: each 1 degree increase in temperature is associated with 15 additional bike rentals. We also find that bike demand is even more sensitive to temperature increases in the shoulder seasons (Autumn and Spring); in Autumn, each 1 degree increase leads to 18 more bike rentals, and in spring, about 27 more rentals. In contrast, in summer a 1 degree increase results in 24 fewer bike rentals, suggesting that in more extreme heat, bike rental demand falls. Rainfall shows a strong negative effect: a 1-unit increase in log-transformed rainfall is associated with about 406 fewer bike rentals, showing that rainfall strongly suppresses bike demand. The model's F-statistic is highly significant ($p<2.2e-16$), which suggests that overall the explanatory variables included contribute meaningfully to explaining bike rentals. The model yields an R2 value of 0.66, indicating that the linear model explains about 66% of the variance in bike rentals. The residual standard error of 373.6 tells us that on average, the linear predictions deviate from the actual bike rental counts by about 374 bikes.

Based on the fitted plot comparing the OLS and Poisson models, the Poisson regression is a more suitable model for this dataset. The linear regression model produces some negative predictions where bike demand is below 0, which is not possible. The fitted plot reveals that the linear model also fails to accurately predict bike rental demand particularly at higher counts, as observed by the flattening of OLS (blue) fitted values as actual bike rentals increase. This suggests that the assumptions of linearity and constant variance are violated. The Poisson model better represents the nature of count data in bike rental demand, as demonstrated by Poisson (yellow) fitted values always above 0. The Poisson model also predicts bike rentals more consistently across all counts of bike demand, as observed by the symmetric shape around the dotted line.

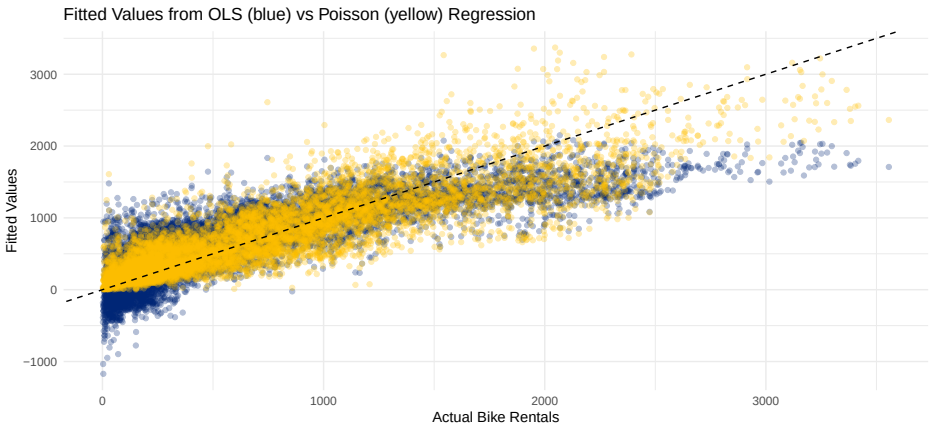

Figure 5: Poisson Model Compared to OLS

7 Conclusion

PLACEHOLDER