PROJECT TITLE:

UNDERSTANDING RENTAL URBAN MOBILITY PATTERNS

Table of Contents

Abstract

Bike-sharing services have become an important component of modern urban transportation by providing accessible, flexible, sustainable, and environmentally friendly mobility options. Despite their popularity, bike-sharing systems often experience fluctuations in demand due to a range of environmental and temporal factors. This study examines the determinants of bike-sharing usage in Seoul by analysing how weather conditions, simple calendar indicators, and temporal patterns relate to rental counts using the Seoul Bike Sharing Demand Dataset. Because the response variable is a discrete count, a count-based model is required. The Negative Binomial Regression model was selected after comparing it with Poisson Distribution Model, as it appropriately accounts for overdispersion and provides a suitable structure for explanatory analysis. The explanatory approach used in this study differs from many existing works that focus primarily on prediction. Instead, the goal is to understand how specific variables influence ridership and to identify patterns that may confirm or challenge intuitive assumptions. Variables such as temperature, humidity, wind speed, rainfall, snowfall, solar radiation, day type, and seasons were examined. Exploratory data analysis showed that weather and temporal patterns are strongly associated with bike rental counts, with temperature and calendar-related factors emerging as key variables. The Negative Binomial model further demonstrated their statistical significance and effect magnitude. By combining environmental and operational factors within an explanatory framework, this study contributes to filling gaps observed in the literature, particularly the limited attention previously given to calendar-based effects. The insights derived from the analysis can support bike-sharing companies in enhancing operational efficiency, improving user experience, managing demand fluctuations, optimizing fleet distribution, and making more informed decisions for urban mobility planning.

Introduction

In recent years, bike sharing services have gained significant popularity as an affordable, sustainable, and convenient mode of urban transportation. Stations and docking points are strategically placed around neighbourhoods, making bicycles easily accessible to the public. These services cater to a wide range of users and purposes, including daily commuting to work or school, short-distance travel, grocery runs, leisure activities, and other personal errands. Their flexibility and accessibility have made bike sharing systems an essential component of modern urban mobility, supporting both environmental sustainability and citywide connectivity.

Despite their widespread adoption, there remain gaps in understanding the factors that drive or hinder bike sharing usage. Even in cities with well-developed infrastructure, usage patterns fluctuate considerably depending on various environmental and temporal conditions. Analysing these variations is crucial for helping operators align service availability with user demand, maximize ridership, and enhance system efficiency.

The motivation behind this study is to identify and understand the factors associated with the frequency of bike sharing usage. By examining how different variables such as weather conditions, day of the week, temperature, humidity, and seasonal effects influence the number of bike rentals, this research aims to uncover data-driven insights that explain user behaviour. These findings can be translated into practical recommendations for bike sharing companies to optimize operations, improve user experience, and address challenges such as demand fluctuations, fleet management, and station placement efficiency.

The objective of this project is to develop an explanatory regression model that quantifies how weather, calendar, and temporal factors relate to bike rental counts. While some relationships may appear intuitive, this study seeks to test these assumptions empirically, assess their statistical significance, and determine the direction and magnitude of their effects. The findings will provide empirical evidence to support data-driven decision-making and contribute to strategies that enhance user engagement and increase overall bike sharing adoption.

Problem Statement and Objectives

Although bike sharing systems have become increasingly popular in major urban centres, demand patterns remain inconsistent and difficult to predict. Factors such as weather, time, and operational conditions often interact in complex ways that influence demand and ridership. Operators frequently face challenges in maintaining optimal fleet availability, managing redistribution logistics. As a result, mismatches between supply and user demand can lead to station congestion, underutilization of bicycles, and reduced service reliability.

The core problem addressed in this study is the limited understanding of how environmental and temporal factors jointly influence bike sharing usage in Seoul. Existing operational decisions are often based on assumptions or historical averages rather than statistically validated models. There is a need for an analytical framework that not only identifies the key determinants of bike rental demand but also quantifies their relative impacts through modelling.

The primary objective of this project is to develop an explanatory regression model that explore the relationship between climatic, temporal, and operational variables and the demand for bike rentals.

Specifically, the study aims to:

Study the dataset to evaluate and determine which regression modelling approach best fits the response variable.

1. Conduct exploratory data analysis (EDA) to identify usage patterns.
2. Evaluate correlations and multicollinearity among explanatory variables.
3. Develop and compare regression models based on their statistical fit.
4. Identify statistically significant predictors influencing bike sharing demand.

5. Interpret the model results to identify statistically significant factors influencing demand.
6. Translate analytical insights into practical recommendations for improving operational efficiency and user engagement.

Literature Review

Research on bike-sharing usage has increasingly focused on understanding how environmental and temporal factors shape ridership patterns. Jaber and Csonka (2023) compared Poisson, Negative Binomial Regression, Seasonal ARIMA, and Random Forest models to evaluate temporal differences in bike usage, showing that count-based models are suitable when the response variable is discrete counts by nature. Their research highlights the importance of accounting for overdispersion and non-linear patterns in bike-sharing data. Wen et al. (2025) expanded this study by examining spatiotemporal flows and peak-hour movements using Citi Bike data, demonstrating how hourly and seasonal dynamics influence travel demand. Similarly, Chun et al. (2024) emphasized the role of environmental and built-environment factors, showing strong associations between urban context, weather conditions, and ridership variation.

Although these studies offer valuable insights, there are several limitations. Much of the existing work emphasizes predictive performance rather than explanatory models, making it difficult to interpret how individual variables influence demand. Many studies focus primarily on weather variables such as temperature and rainfall, while giving less attention to temporal or calendar-based variables such as holidays, non-holidays, and seasons. These gaps restrict understanding of how temporal factors might link to bike rental beyond simple weekdays/weekend or hourly effects.

This project builds on prior research by integrating both environmental and calendar variables into an explanatory model using the Negative Binomial Regression model. By incorporating holidays, non-holidays, and seasonal variation, the study addresses an underexplored dimension in existing literature. The goal is to explore the association between factors and bike-sharing demand in Seoul and to identify predictors that may confirm or challenge intuitive assumptions about bike sharing usage. This approach contributes to a more complete understanding of ridership dynamics and provides interpretable insights that support operational decision-making and sustainable transport planning.

Data and Methods

Data Overview and Initial Analysis

The dataset used in this study is the Seoul Bike Sharing Demand Dataset, which records daily and hourly rental counts across different time periods and weather conditions. It includes 14 columns and 8,760 observations, and variables such as temperature, humidity, wind speed, rainfall, snowfall, visibility, solar radiation, season, and functioning day indicators. These variables allow for comprehensive analysis of both environmental and temporal influences on rental frequency. The response variable that we are using is the bike rental count, a discrete count variable, the response variable will be the deciding factor on which model to use.

Table 1. Dataset

Variables	**Unit of Measurement**	**Description**	**Data Type**	**Purpose**
Date	NA	Calendar day for each record	year-month-day	Explanatory variable, Creation of derived variable*
Rented Bike Count	Rental Hours	Number of rented bikes	Discrete	Response variable
Hour	NA	Hour of the day (0-23)	Continuous	Explanatory variable
Temperature	°C	Temperature in Celsius	Continuous	Explanatory variable
Humidity	%	Humidity Percentage	Continuous	Explanatory variable
Wind speed	m/s	Wind speed in meters per second	Continuous	Explanatory variable
Visibility	10m	Visibility distance in units of 10 meters	Continuous	Explanatory variable
Dew point temperature	°C	Dew point temperature in Celsius	Continuous	Explanatory variable
Solar Radiation	MJ/m^2	Solar radiation in megajoules per square meter	Continuous	Explanatory variable
Rainfall	mm	Amount of precipitation in millimetres	Continuous	Explanatory variable
Snowfall	cm	Depth of snow in centimetres	Continuous	Explanatory variable
Seasons	NA	Autumn, Spring, Summer, Winter	Categorical	Explanatory variable
Holiday	NA	Holiday or no holiday	Categorical	Explanatory variable
Functioning Day	NA	Bike sharing app operating	Categorical	Explanatory variable

Excel will first be used to perform a quick preliminary analysis before importing the Seoul Bike Sharing dataset into SAS Viya for data preparation. Summary statistics will then be generated in

SAS Studio using the Summary task to identify missing or invalid values. Duplicate records will be detected and removed using Data Step.

All data quality issues will be carefully reviewed to determine whether to retain or exclude affected records. Subsequently, data preparation will be conducted in Data Step to remove missing values, invalid entries, and duplicates. Feature engineering will then be applied to create additional explanatory variables such as day of the week, month, and weather-related groupings to capture potential behavioural effects.

Summary Statistic

Variable Summary

Obs	Variable name	Number of missing values	Minimum numeric value	Maximum numeric value	Mean	Standard deviation
1	Temperature(°C)	0	-17.8	39.4	12.882922374	11.94482523
2	Humidity(%)	0	0	98	58.226255708	20.362413302
3	Wind speed (m/s)	0	0	7.4	1.7249086758	1.0362999934
4	Visibility (10m)	0	27	2000	1436.8257991	608.29871198
5	Dew point temperature(°C)	0	-30.6	27.2	4.0738127854	13.060369338
6	Solar Radiation (MJ/m2)	0	0	3.52	0.5691107306	0.8687462422
7	Rainfall(mm)	0	0	35	0.1486872146	1.1281929687
8	Snowfall (cm)	0	0	8.8	0.0750684932	0.4367461811
9	Holiday	0	.	.	.	.
10	Rented Bike Count	0	0	3556	704.60205479	644.99746774

Figure 1. Summary statistics table

Before beginning any data preparation, a summary statistics table was generated to examine the key explanatory variables in the dataset. As shown in the figure 1, all variables reported zero missing values, indicating that the dataset was fully complete and did not require imputation at this stage.

Data Preparation

Data preparation is performed via SAS Studio, including the creation of new variables by binning certain the measurements. Firstly, the original date-time values were parsed (Figure 2) into distinct day, month, and year components using a scan function, enabling more precise temporal analysis and simplifying the aggregation or filtering of records by specific calendar periods.

```
length date_part time_part $ 20;
    date_part = scan(Date, 1, ' ');
    day_str = scan(Date, 1, '/');
    month_str = scan(Date, 2, '/');
    year_str = scan(Date, 3, '/');
```

Figure 2. Parsing date

Entries with 'Functioning Day' marked as 'No' were dropped from the dataset, as they correspond to days with zero counts. (Figure 3)

```
rename Hour_Time = Hour;
where 'Functioning Day'n = 'Yes';
```

Figure 3. Dropping certain rows

Feature engineering was conducted on the date variables by extracting month strings and converting them to integer and descriptive month name formats, as well as deriving the day of week for each record (Figure 4). This information was then used to classify each entry as either a weekday or weekend, enabling richer temporal analysis and supporting downstream modeling of seasonal and weekly trends in bike-sharing demand.

```
month = input(month_str, best.);
    length day_of_week_name $ 5;
    day_of_week_name = put(mdy( input(month_str, best.),
        input(day_str, best.), input(year_str, best.)), downame3.);
    length weekend_or_day $ 20;
    if day_of_week_name in ('Mon', 'Tue', 'Wed', 'Thu', 'Fri') then
        weekend_or_day = 'weekday';
    else
        weekend_or_day = 'weekend';

length month_name $ 9;

    if month='1' then month_name='January';
    else if month='2' then month_name='February';
    else if month='3' then month_name='March';
    else if month='4' then month_name='April';
    else if month='5' then month_name='May';
    else if month='6' then month_name='June';
    else if month='7' then month_name='July';
    else if month='8' then month_name='August';
    else if month='9' then month_name='September';
    else if month='10' then month_name='October';
    else if month='11' then month_name='November';
    else if month='12' then month_name='December';
```

Figure 4. Feature engineering for day and month

Checking for Multicollinearity

After completing data preparation, we assessed multicollinearity among the explanatory variables. A correlation matrix was generated to identify any strongly correlated predictors that might distort the regression results.

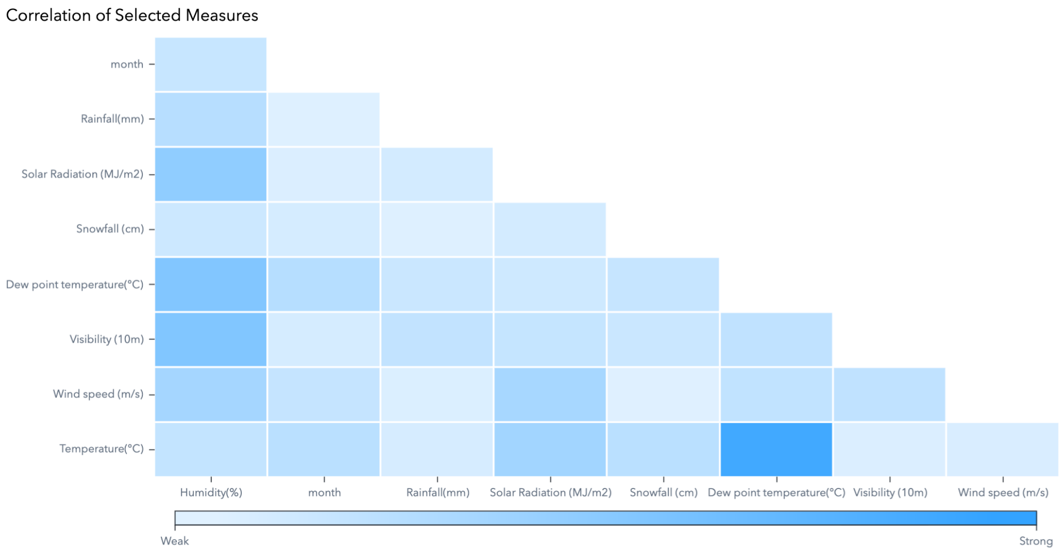

Figure 5. Correlation Matrix

As shown in Figure 5, correlations among most variables fall below the 0.8 threshold, indicating acceptable levels of multicollinearity. However, *Dew Point Temperature* and *Temperature* exhibit a correlation of 0.9145, which is too high for reliable regression analysis. Therefore, *Dew Point Temperature* was excluded from the model, and a revised correlation matrix was generated using the remaining variables.

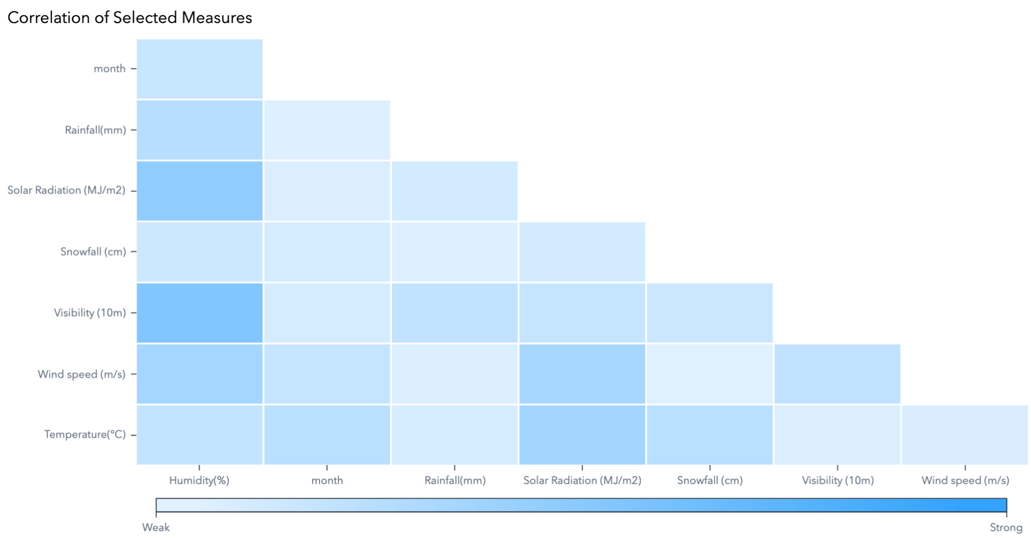

Figure 6. Correlation matrix with selected variables

In Figure 6, all correlations among the selected variables fall below the 0.8 threshold, indicating no multicollinearity concerns. With this confirmed, the variables are ready to be included in the regression model.

Exploratory Data Analysis

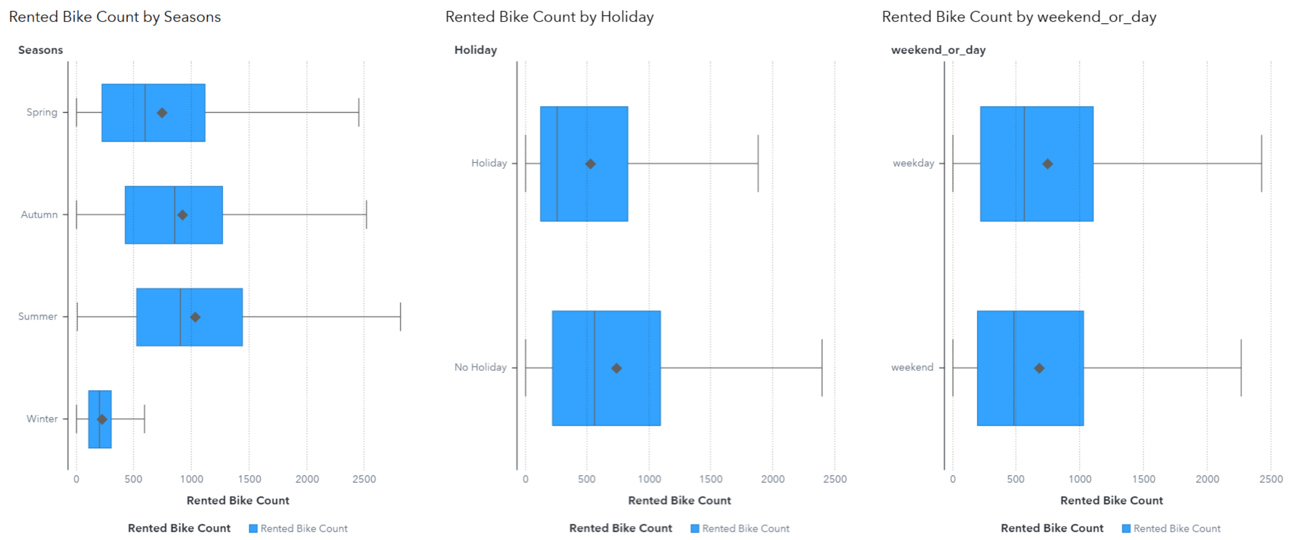

Figure 7. Box plots for weekday/weekend, season, holiday

As in figure 7, bike-sharing usage shows clear temporal patterns across weekdays, seasons, and holidays. Weekday rentals tend to be slightly higher than weekend rentals, with a wider spread of values, suggesting that weekday usage is more consistent and likely driven by routine commuting. Seasonal patterns reveal that ridership peaks during summer and autumn when weather conditions are more comfortable, while winter shows the lowest rental counts due to colder temperatures. Holiday effects also emerge, as average rentals are lower on holidays compared to non-holidays, indicating that most users rely on bike sharing for daily commuting rather than leisure travel.

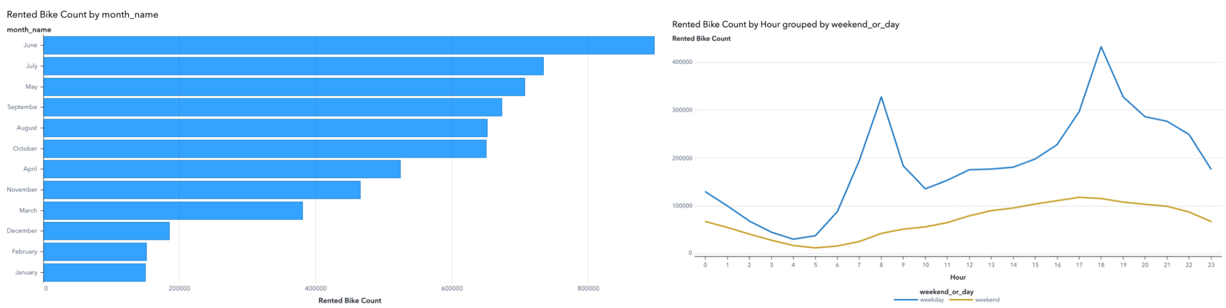

Figure 8. EDA for month and hours of the day

According to figure 8, monthly rental patterns align closely with weather seasonality, with June showing the highest usage due to its favourable temperatures, while winter months from December to February record the lowest counts as cold weather limits riding comfort. Hourly patterns differ between weekdays and weekends. On weekdays, rentals display two distinct peaks around 8am and 6pm, which correspond to common commuting periods. In contrast, weekend usage increases gradually throughout the day and peaks in the afternoon, reflecting more flexible and leisure-driven travel behaviour.

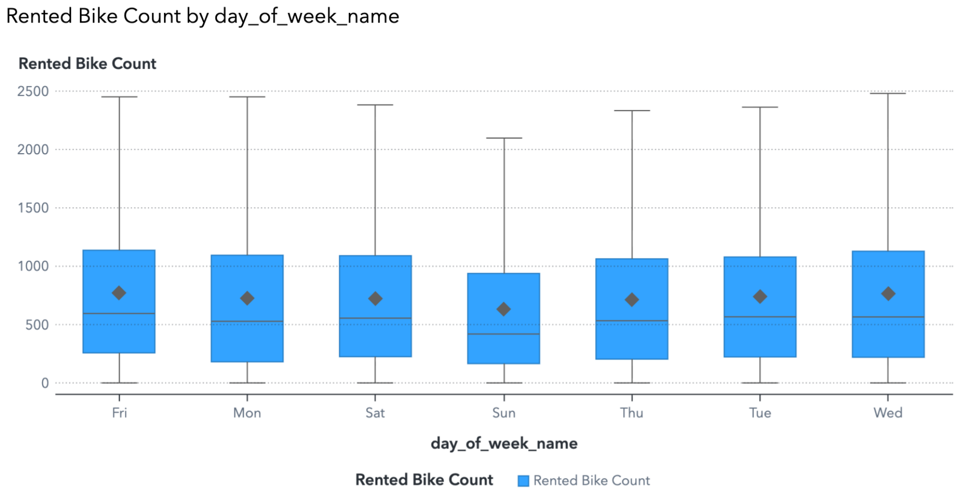

Figure 9. Box plot for days of the week

Bike rentals remain consistent across weekdays, reflecting strong commuter-driven demand. Counts increase slightly on Friday and Saturday, while Sunday shows a noticeable decline, likely due to reduced work-related travel. (Figure 9)

Model Comparison

For model comparison, generalized linear model is performed under SAS Studio tasks and fit in the distribution type, continuous variables, classification effects with response variable as *Rented Bike Count*.

Poisson Regression Model

The model was fit by using seasons, day of the week and holiday as classification variables, and the selected variables from Figure 6.

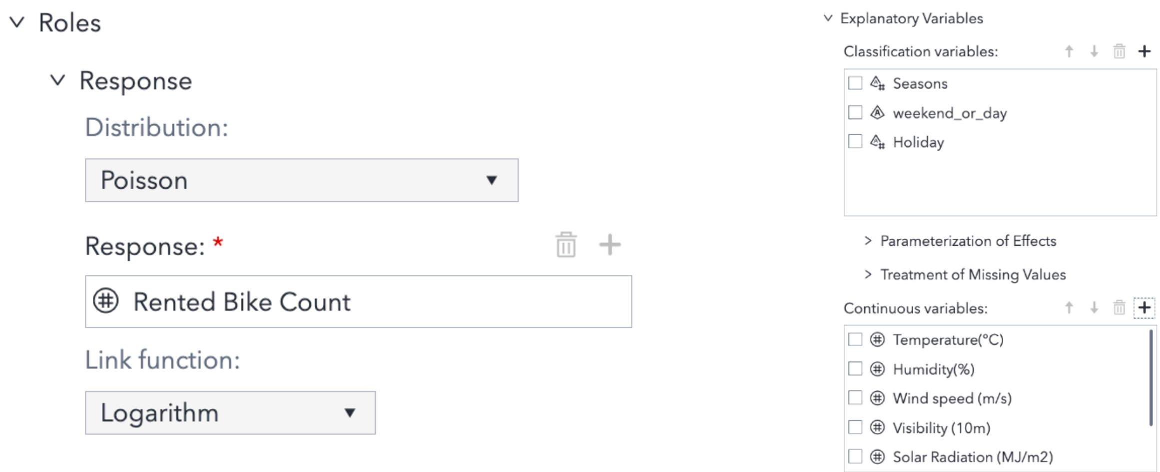

Figure 1010. Poisson Regression Model

According to figure 11, the table summarizes goodness-of-fit metrics for the Poisson Regression Model, including deviance, log likelihood, AIC, and BIC. Notably, the values are extremely large, while the model with the lowest BIC and AIC is preferred. (Jaber and Csonka, 2023). This indicates possible overdispersion and lack of model fit. Given these high values, the Poisson model is inadequate for the data, motivating the preference for Negative Binomial Regression, which better handles overdispersion.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8455	2345731.1978	277.4372
Scaled Deviance	8455	2345731.1978	277.4372
Pearson Chi-Square	8455	755700414.14	89379.1146
Scaled Pearson X2	8455	755700414.14	89379.1146
Log Likelihood		35620222.198	
Full Log Likelihood		-1206412.465	
AIC (smaller is better)		2412844.9295	
AICC (smaller is better)		2412844.9555	
BIC (smaller is better)		2412915.3664	

Figure 1111. Result table of Poisson Regression Model

Negative Binomial Regression

Since the Deviance with the Poisson Regression is overly dispersed with a value of 277, the Negative Binomial Regression is better model (variance > mean). The same variables were fit in the model again by using Negative Binomial distribution. (Figure 12)

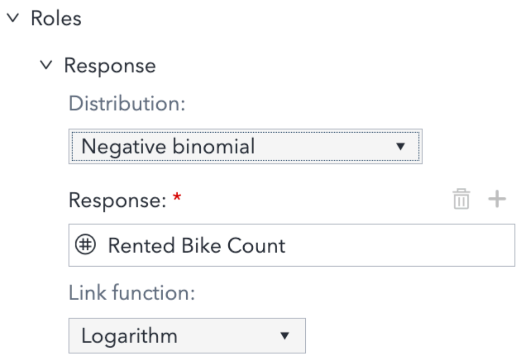

Figure 1212. Negative Binomial Regression

The lower AIB/BIC as well as the Log Likelihood became lower, and it concludes the Negative Binomial Regression being a better explanatory model.

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8454	9164.4361	1.0840
Scaled Deviance	8454	9164.4361	1.0840
Pearson Chi-Square	8454	8792.3316	1.0400
Scaled Pearson X2	8454	8792.3316	1.0400
Log Likelihood		36765327.046	
Full Log Likelihood		-61307.6167	
AIC (smaller is better)		122639.2333	
AICC (smaller is better)		122639.2702	
BIC (smaller is better)		122723.7577	

Figure 1313. Result of Negative Binomial Regression

Negative Binomial Regression Model (Visual Analytics)

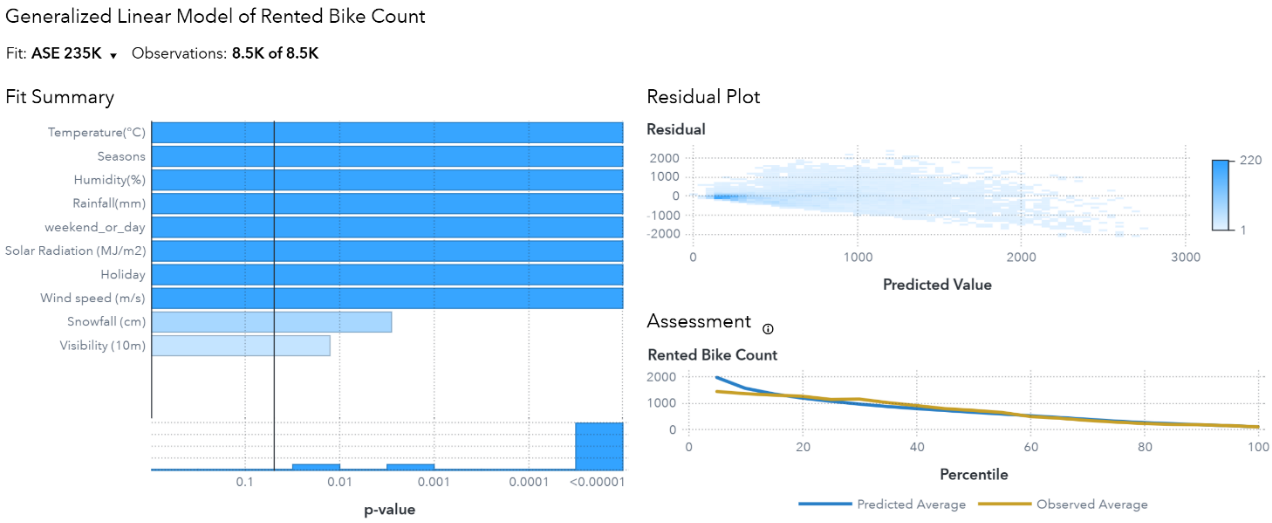

Figure 14. Visual Analytics of Negative Binomial Regression

Model Information | Dimensions | Iteration History | Convergence | Fit Statistics | Parameter Estimates | Type III Test | Assessment | Assessment Statistics | Assessment Summary

Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	6.213103	0.054157	13161.64	<0.00001
Holiday Holiday	-0.36018	0.035226	104.55	<0.00001
Holiday No Holiday	0	.	.	.
weekend_or_day weekday	0.212476	0.01659	164.0351	<0.00001
weekend_or_day weekend	0	.	.	.
Seasons Autumn	0.727268	0.030642	563.3311	<0.00001
Seasons Spring	0.446495	0.030012	221.3358	<0.00001
Seasons Summer	0.354874	0.04362	66.18608	<0.00001
Seasons Winter	0	.	.	.
Wind speed (m/s)	0.046349	0.008164	32.22972	<0.00001
Humidity(%)	-0.01662	0.0006	768.3936	<0.00001
Rainfall(mm)	-0.11294	0.004408	656.5058	<0.00001
Snowfall (cm)	-0.05504	0.018438	8.912861	0.00283
Solar Radiation (MJ/m2)	-0.12061	0.011723	105.8488	<0.00001
Temperature(°C)	0.051247	0.001363	1414.385	<0.00001
Visibility (10m)	-0.00004	0.000016	6.210051	0.01270
Dispersion	0.467726	0.006784	.	.

Figure 15. Parameter Estimates of NBR

The analysis was extended into Visual Analytics, where the Negative Binomial distribution was selected as the modelling approach. Predictor variables were refined using the correlation matrix, and those with p-values greater than 0.05 or deemed statistically insignificant were removed from the model.

Analysis and Results

Table 2. Parameter Estimates result from NBR Model

Parameter	Estimate Parameter	Standard Error	Chi-Square	Pr > ChiSq	exp(β)
Intercept	6.21	0.05	13,161.64	0.00000	499.25
Holiday Holiday	-0.36	0.04	104.55	0.00000	0.70
Holiday No Holiday	-				1.00
weekend_or_day weekday	0.21	0.02	164.04	0.00000	1.24
weekend_or_day weekend	-				1.00
Seasons Autumn	0.73	0.03	563.33	0.00000	2.07
Seasons Spring	0.45	0.03	221.34	0.00000	1.56
Seasons Summer	0.35	0.04	66.19	0.00000	1.43
Seasons Winter	-				1.00
Wind speed (m/s)	0.05	0.01	32.23	0.00000	1.05
Humidity(%)	-0.02	0.00	768.39	0.00000	0.98
Rainfall(mm)	-0.11	0.00	656.51	0.00000	0.89
Snowfall (cm)	-0.06	0.02	8.91	0.00283	0.95
Solar Radiation (MJ/m2)	-0.12	0.01	105.85	0.00000	0.89
Temperature(°C)	0.05	0.00	1,414.39	0.00000	1.05
Visibility (10m)	-0.00	0.00	6.21	0.01270	1.00
Dispersion	0.47	0.01			1.60

Negative Binomial Regression Model can be expressed as:

$$\log(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

All key predictors are statistically significant. The large Chi-square statistics indicate strong explanatory power, with temperature and seasonal weather conditions emerging as key drivers. overall, favourable weather and typical commuting periods lead to higher ridership, showing that both environmental and calendar factors play important roles in Seoul's bike-sharing demand. Bike usage follows clear temporal and environmental patterns where ridership is higher on weekdays and non-holidays, reflecting work related mobility. This aligns with our EDA findings, where bike counts peak at 8 a.m. and 6 p.m. on weekdays.

Weather conditions strongly influence behaviour. Rainfall, snowfall, and high humidity significantly reduce rentals, with rainfall having the most pronounced negative effect. According to Jaber & Csonka (2023), precipitation significantly affects only occasional users on the contrary to the response seen in our data. Seasonal patterns also matter where usage is observed to increases with warmer temperatures (up to around 30°C) and drops sharply during winter months. Jaber & Csonka (2023) also observed that temperature is insignificant for occasional users but negatively affects members.

Higher solar radiation is associated with reduced rentals, suggesting that excessive heat or strong sunlight discourages cycling. Visibility shows the weakest relationship with bike count, likely because most days observe consistently clear conditions near the maximum visibility range.

Wind speed is found to have a positive association with bike count in our model, contrary to the findings of Zhou et al. (2017), who report that wind speed negatively impacts usage across all user groups. This may be explained by the relatively steady wind conditions in Seoul, where stronger wind speeds are uncommon.

Discussion

This study found that both environmental and calendar variables significantly influence Seoul's bike-sharing demand. Higher temperatures and favourable weather conditions encouraged more rides, while rainfall, snowfall, and humidity discouraged usage. Weekdays and non-holidays recorded higher ridership, reflecting strong commuter-driven behaviour. These results align with Jaber and Csonka (2023), who found that weather and temporal factors shape ridership patterns, but this study extends their work by including holidays and seasons. The findings not only confirm common intuitions that good weather increases cycling but also challenge some assumptions by revealing that higher solar radiation and certain seasonal effects can reduce usage. This demonstrates the value of explanatory modelling in uncovering nuanced relationships that can guide policy, optimize operations, and improve long-term sustainability in bike-sharing systems.

Conclusion & Contributions

The Negative Binomial Regression Model effectively identified key factors influencing Seoul's bike-sharing demand, addressing overdispersion and providing reliable insights for count-based data. Results showed that higher temperatures and warmer seasons increased usage, while rainfall, snowfall, and high humidity reduced ridership. Weekdays and non-holidays recorded higher counts, reflecting strong commuter-driven demand patterns.

This study contributes to both theory and practice by integrating calendar variables such as holidays and seasons into explanatory modelling, an area often overlooked in previous research. Practically, the findings can help bike-sharing operators and city planners optimize fleet distribution, scheduling, and promotional strategies. More broadly, the results support urban sustainability efforts by emphasizing how environmental and temporal factors shape shared mobility usage.

References

Jaber, A., & Csonka, B. (2023). *Investigating the temporal differences among bike-sharing users through comparative analysis based on count, time series, and data mining models*. Alexandria Engineering Journal. https://doi.org/10.1016/j.aej.2023.06.087

Chun, B., Nguyen, A., Pan, Q., & Mirzaaghazadeh, E. (2024). *Spatial analysis of bike-sharing ridership for sustainable transportation in Houston, Texas*. Sustainability, 16(6), 2569. https://doi.org/10.3390/su16062569

Wen, Z., Tian, D., & Wu, N. (2025). *Modeling and analyzing the spatiotemporal travel patterns of bike sharing: A case study of Citi Bike in New York*. Sustainability, 17(1), 14. https://doi.org/10.3390/su17010014

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies*, *75*, 17-29. https://doi.org/10.1016/j.trc.2016.12.001