Predicting Bike Usage at UofT St. George with Bayesian Poisson Regression*

Usage grows by one every 20 days, drops to nearly zero in winter, with 79% occurring between 8 AM and 8 PM

Haowei Fan

November 28, 2024

University of Toronto's St. George campus students and staff frequently encounter challenges in locating available bike-sharing parking spots upon arrival and bikes upon leaving (El-Assi, Mahmoud, and Habib 2017), underscoring the necessity for accurate predictions of future bike-sharing demand to optimize campus commuting infrastructure. This study employs Bayesian Poisson regression to predict the utilization of bike-sharing stations at 27 locations across campus for specific years, months, and four-hour intervals of the day. The results indicate that bike-sharing usage across campus is experiencing rapid growth, with a sharp decline during the winter months and 79.8% of peak usage consistently observed between 8:00 AM and 8:00 PM daily. Specifically, on September 26, 2025, between 4:00 and 8:00 AM, the usage of 8 stations is projected to increase by 3 bikes, 14 stations by 2 bikes, 3 stations by 1 bike, while 1 station will remain unchanged, and 1 station will see a decrease by 1 bike, compared to the same timeframe in 2024.

Table of Contents

1 Introduction **1**
- 1.1 Overview . . . 1
- 1.2 Estimand . . . 3

2 Data **3**
- 2.1 Overview . . . 3
- 2.2 Measurement . . . 4

*Code and data are available at: https://github.com/HaoweiFan0912/Bikeshare-Forecast.git

2.3 Variables 5

3 Model **8**
3.1 Model descriptions 8
3.2 Validation 10

4 Results **12**

5 Discussion **13**
5.1 What the Study Brought to the World 13
5.2 Limitations 14
5.3 Suggestions for Future Research 15

Appendix **16**

A Data collection methodology **16**
A.1 Overview 16
A.2 Evaluation 16

B Idealized methodology **17**
B.1 Ideal Dataset Framework 18

References **22**

1 Introduction

1.1 Overview

Bike-sharing services have experienced rapid growth in the Toronto region, with ridership rising dramatically from approximately 665,000 trips in 2015 to over 4.5 million in 2022, according to a study by the University of Toronto's School of Cities(Liu and Allen 2023). This rapid increase highlights the critical need for data-driven analyses to optimize bike-sharing systems and guide policy decisions that can further enhance urban mobility(Zhang, Gu, and Zhao 2024). While city-wide studies provide valuable insights, research from Hangzhou, China, has demonstrated that bike-sharing usage patterns can vary significantly between urban areas and university campuses(Tang et al. 2020). These differences suggest that general urban studies conducted in the Toronto region may not fully capture the unique dynamics of campus-based bike-sharing systems. Therefore, to address this gap in the literature, this study focuses on the University of Toronto St. George Campus as a case study, analyzing bike-sharing usage within the campus in Toronto to provide a nuanced understanding of its role in campus mobility and inform policies tailored to similar settings.

The core of this study is to extract a sample of 27 bike-sharing stations within the University of Toronto campus from all bike-sharing usage data between January 1, 2017, and September 30, 2024. The total usage of bike-sharing at each station is calculated for each 4-hour interval. A Bayesian Poisson regression model is then employed to predict the usage based on the station, year, month, date, and the specific 4-hour interval of the day.

This study found that although the usage at all stations within the University of Toronto campus has increased rapidly each year, there are significant seasonal differences. During winter, usage at all stations, regardless of the year, remains close to zero. Additionally, from January to September each year, the usage shows an upward trend, which then gradually declines over the following months. There are also significant differences in usage across different times of the day. The period from 8 am to 8 pm accounts for an average of 79.8% of daily usage, while the period from midnight to 8 am accounts for only 6.8%. It is noteworthy that this project forecasts an average increase of 2 uses across 27 stations between 4:00 AM and 8:00 AM on September 26, 2025, compared to the same time period in 2024. Specifically, usage at 8 stations is expected to increase by 3, 14 stations by 2, and 3 stations by 1, while 1 station will remain unchanged and another will decrease by 1.

Compared to the general statistics on bike-sharing usage in the Greater Toronto Area, these findings offer more detailed recommendations specifically for transportation planning and policy adjustments within the University of Toronto campus. They also provide guidance on how to better allocate, deploy, or store shared bikes.

** ToBeMoreSpecific** The paper proceeds as follows: Section 2 outlines the data sources, processing techniques, and the variables employed in the study. Section 3 describes the Bayesian Poisson regression model used for prediction, including its formulation and validation. Section 4 discusses the results, emphasizing temporal and station-specific patterns in bike-sharing usage. Section 5 concludes with a discussion of the study's contributions, limitations, and recommendations for future research.

1.2 Estimand

This study aims to estimate the usage of a specific bike-sharing station within the University of Toronto campus during a particular 4-hour interval. By converting time into year, month, day, and the specific 4-hour interval of the day, it accounts for the overall trend, seasonal effects, and hourly variation in station usage. The core objective is to explore the temporal changes in the usage of 27 bike-sharing stations within the campus, thereby providing policymakers with recommendations for bike allocation to improve commuting efficiency on campus.

2 Data

2.1 Overview

The dataset used in this study comes from opendatatoronto (Gelfand 2022), uploaded by Toronto Parking Authority(Toronto Parking Authority, n.d.) and collected by Bike Share Toronto(Bike Share Toronto, n.d.). It records every bike-sharing usage in the Toronto area from 2015 to September 30, 2024, with a total of 28,017,329 records. The variables included in the data differ across years, but they all contain the following variables: Trip ID, Trip Duration, Trip Start Station ID, Trip Start Time, Trip Start Station Location, Trip End Station ID, Trip End Time, Trip End Station Location, Bike ID, and User Type.

Many bike-sharing service providers worldwide have made their operational data publicly available for research purposes. For instance, the Shenzhen Municipal Transport Bureau in China released a dataset containing partial bike-sharing order information from January to August 2021. This dataset includes eight variables: user ID, start time, start longitude, start latitude, end time, end longitude, end latitude, and company ID, with a total of approximately 240 million records (Shenzhen Open Data Platform 2021). While this dataset offers a rich variety of variables, it has limitations that make it unsuitable for my research. The data only covers an eight-month period, which is insufficient to capture long-term trends in bike-sharing usage. Additionally, university campuses in China are typically closed environments, unlike the open urban settings of Toronto, making the data less generalizable for my research context. Similar issues are present in other bike-sharing datasets from different regions. These include limited temporal coverage, which restricts long-term analysis, strong regional characteristics influenced by local cultural habits, and significant variations in data quality and structure due to fierce market competition among bike-sharing companies. These factors make it challenging to integrate data across companies or regions. Therefore, while these datasets may hold value for other studies, they do not meet the requirements of my research.

This study follows the workflow of Telling Stories with Data (Alexander 2023), using its initial folder structure and part of its code. Data downloading, cleaning, modeling, and visualization were carried out using R (R Core Team 2023). The following R libraries were also used alongside R:

tidyverse(Wickham et al. 2019): Used for data wrangling, cleaning, and analysis, integrating multiple data manipulation packages.

dplyr(Wickham et al. 2023): Used for data frame operations such as filtering, sorting, and aggregation.

arrow(Richardson et al. 2024): Used for reading and writing data in efficient formats like Parquet to speed up processing.

stringr(Wickham 2023): Used for handling and manipulating strings, allowing for text data cleaning and transformation.

readr(Wickham, Hester, and Bryan 2024): Used for fast reading of CSV and other text formats.

lubridate(Grolemund and Wickham 2011): Used for handling and converting date-time data, simplifying time-related data analysis.

brms(Bürkner 2017): Used for building and estimating Bayesian regression models for flexible data analysis.

rstanarm(Goodrich et al. 2022): Used for Bayesian regression modeling, helping to understand uncertainty better.

bayesplot(Gabry et al. 2019): Used for visualizing posterior distributions and diagnostic plots of Bayesian models.

ggplot2(Wickham 2016): Used for creating various types of plots, supporting exploratory data analysis and result presentation.

RColorBrewer(Neuwirth 2022): Used for generating color palettes to create visually distinct and appealing plots.

knitr(Xie 2014): Used for generating dynamic reports, integrating analysis results into documents.

kableExtra(Zhu 2024): Used for enhancing the presentation of tables, making tables in reports more visually appealing.

2.2 Measurement

To predict the relationship between bike-sharing usage and time within the University of Toronto's St. George campus, this study requires a time series to describe the changes in usage at different stations over time. Therefore, the dataset published by opendatatoronto (Gelfand 2022), which records every instance of bike-sharing usage in the Toronto area since 2015, is ideal for this study. Additionally, this data is ideal because of its reliability—due to the commercial nature of bike-sharing, the specific time and location of each bike's use and return are accurately recorded. However, the raw data cannot be used directly in this project; sophisticated data cleaning is required, with specific steps and reasons as follows:

First, data from 2015-2016 was excluded because the collection and recording methods for those years differ significantly from later years and do not include the time and station variables required for this study. Next, data for stations located within the University of Toronto's St. George campus was extracted from all remaining samples, and the usage of each station was calculated for every four-hour interval. Finally, the variables of interest were extracted, and the data format was standardized for subsequent analysis. The cleaned data is presented in the format of Table 1.

Table 1: Samples of the dataset used for analysis

station_name	time	count
Willcocks St / St. George St	2024-09-29 12:00:00	1
Willcocks St / St. George St	2024-09-29 16:00:00	1
Willcocks St / St. George St	2024-09-30 04:00:00	1
Willcocks St / St. George St	2024-09-30 08:00:00	1
Willcocks St / St. George St	2024-09-30 12:00:00	7
Willcocks St / St. George St	2024-09-30 16:00:00	4

2.3 Variables

This study focuses on the following variables:

- `count`: The dependent variable of the study, a non-negative integer. It describes the total usage of bike-sharing at a particular station during a specific 4-hour interval.
- `time`: An independent variable representing the time interval. For example, "2024-09-29 12:00:00" represents the time interval from 12 pm to 4 pm on September 29, 2024, in a 24-hour format. The earliest `time` is "2017-01-01 00:00:00," and the latest is "2024-09-30 24:00:00."
- `station_name`: An independent variable representing the unique name of one of the 27 stations within the University of Toronto's St. George campus.

Table 2 is the summary of varibales. It is evident that the dataset includes 27 unique station names and spans nearly eight years (from January 1, 2017, to September 30, 2024), providing opportunities to analyze long-term trends and seasonal variations. In the count column, the minimum and median values are both 1, the mean is 2, the first quartile is 1, the third quartile is 3, and the maximum reaches 31, showing a significant right-skewed distribution. This indicates that most of the records have low count values, but there are a few extreme high values.

Table 2: Variables' summary

station_name	time	count
Type: Character	Earliest: 2017-01-01 04:00:00	min: 1
Number of Unique Values: 27	Latest: 2024-09-30 20:00:00	1st Qu: 1
NA	NA	median: 1
NA	NA	mean: 2
NA	NA	3rd Qu: 3
NA	NA	max: 31

Figure 1 shows the daily usage totals of 27 shared bicycle stations from January 1, 2017, to September 30, 2024. It can be observed that the overall usage exhibits a significant upward trend, particularly after 2021, where usage fluctuations increased noticeably. In 2024, several peaks in daily usage reached historical highs, indicating a substantial growth in user demand. Additionally, the data displays some seasonal variations, with noticeable declines during the winter months and increases during spring and summer.

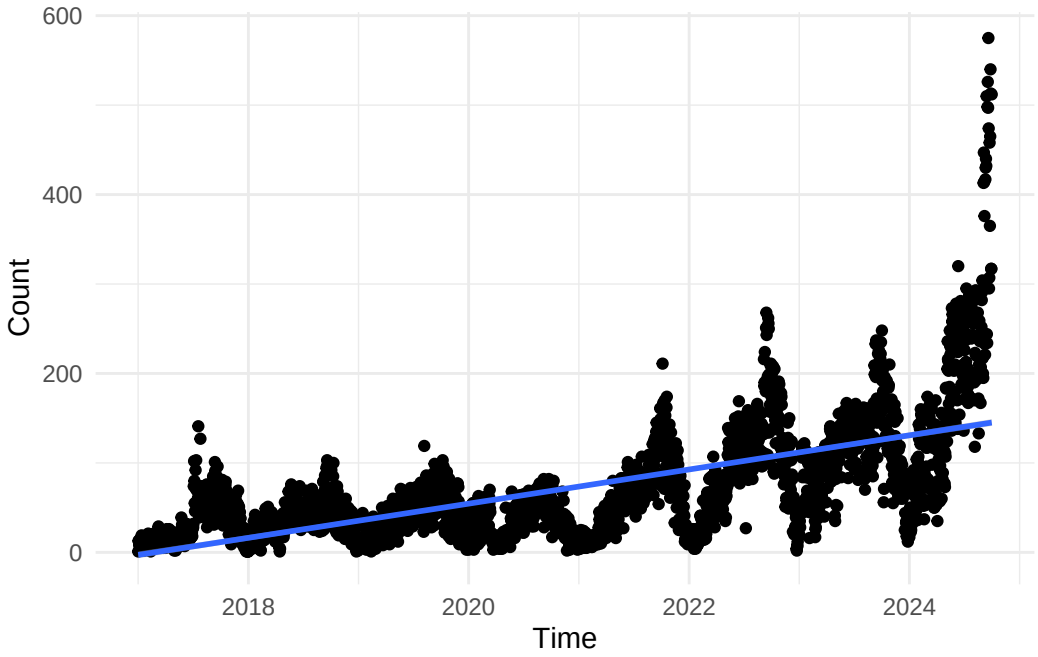

Figure 1: Daily usage from January 2017 to September 2024

hows the total usage and percentage of shared bicycles during different time intervals in a day from January 1, 2017, to September 30, 2024. It can be observed that the peak usage occurs between 12:00 to 16:00 (28.5%) and 16:00 to 20:00 (31.4%). The usage between 08:00 to 12:00 accounts for 19.9%, while the usage from 20:00 to 00:00 accounts for 13.4%. The time intervals with the lowest usage are 00:00 to 04:00 (3.6%) and 04:00 to 08:00 (3.2%).

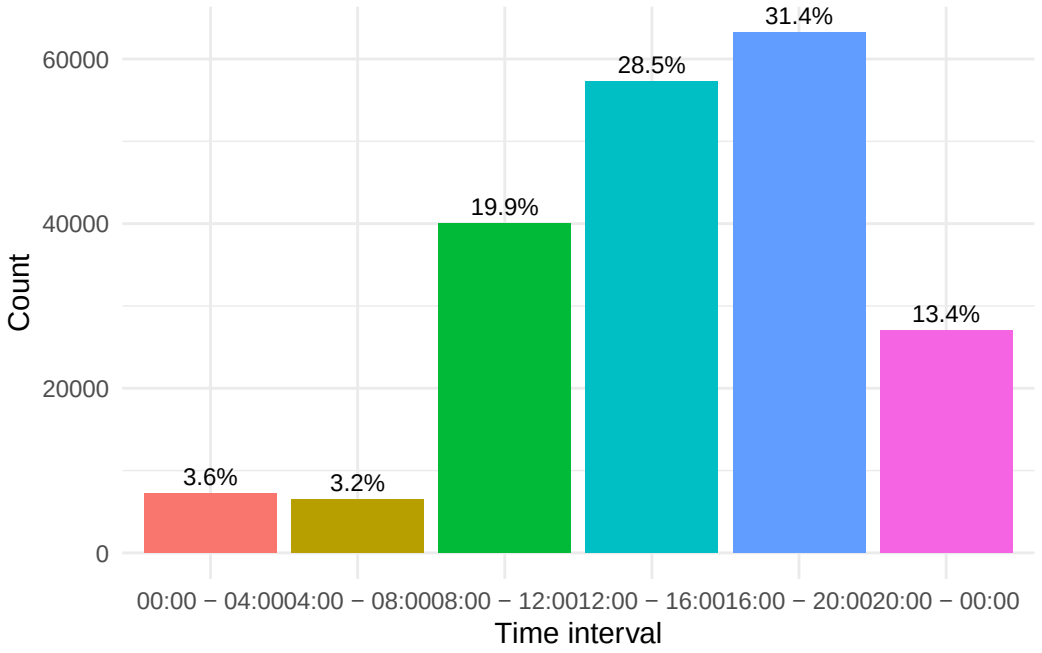

Figure 2: Total usage during different time periods from January 2017 to September 2024

Figure 3 shows the result of the average daily parking volume minus the departure volume at each station. It can be observed that most stations have balanced parking and departure volumes, but some stations show significant supply-demand differences. Specifically, stations such as Bay St / Charles St W – SMART and Bay St / Wellesley St W have noticeably higher departure volumes than parking, while College St / Huron St and College St / Henry St have higher parking volumes than departure, indicating an imbalance in supply and demand at these locations.

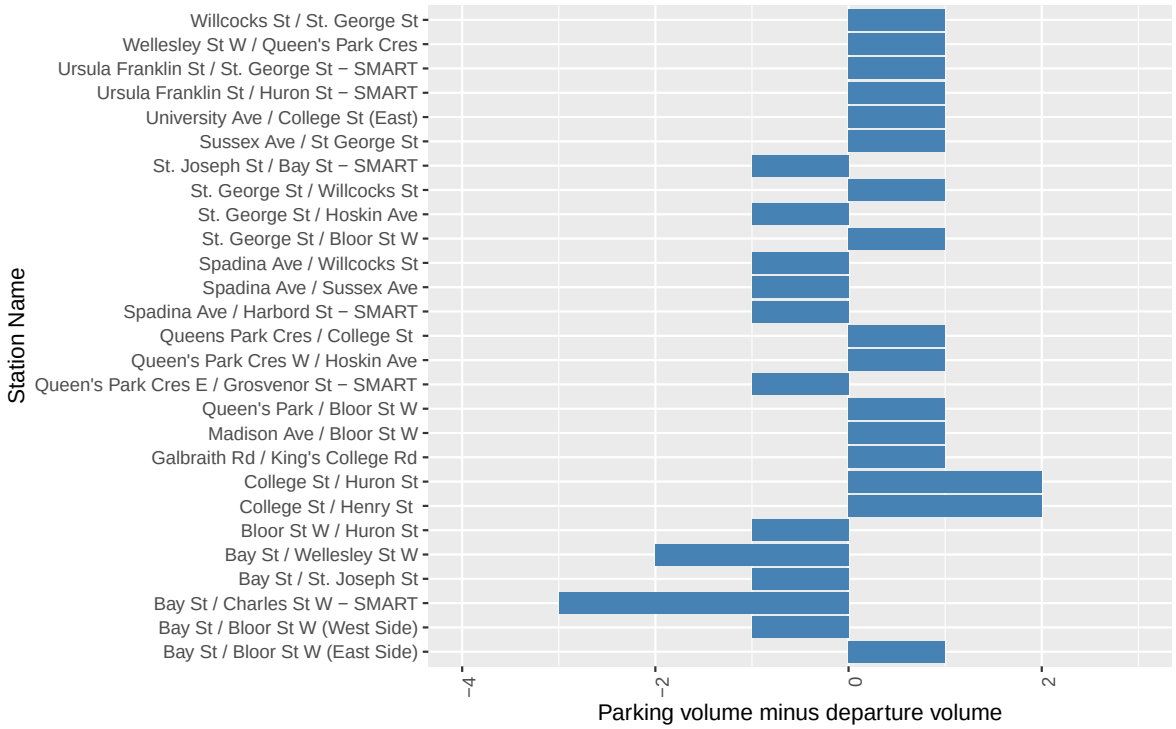

Figure 3: Difference of daily average of bicycle parking and departure volumes for each station

3 Model

3.1 Model descriptions

The model used in this study is as follows:

$$
\begin{aligned}
\text{count} &\sim \text{Poisson}(\lambda) \\
\log(\lambda) &= \beta_0 + \beta_1 \times \text{hour} + \beta_2 \times \text{day} + \beta_3 \times \text{month} + \beta_4 \times \text{year} \\
\beta_0 &\sim \mathcal{N}(0, 5) \\
\beta_j &\sim \mathcal{N}(0, 2) \quad \text{for } j = 1, 2, 3, 4
\end{aligned}
$$

This study employs a Bayesian linear Poisson model to predict the variable `count`, which represents the expected number of bike-sharing usage events at a specific station on the St. George campus of the University of Toronto within a future 4-hour interval. The `count` is assumed to follow a Poisson(λ) distribution, where $\log(\lambda)$ is the logarithmic transformation of the count. This transformation provides a simpler and more stable way to describe the relationship between the `count` and its influencing factors.

The predictors used in this model are as follows:

- year: The year of the time being predicted.
- month: The month of the time being predicted.
- day: The day of the month being predicted.
- hour: The time interval within the day being predicted. For example, a value of 12 represents the 4-hour interval from the 12th hour to the 16th hour in a 24-hour day.

The intercept β_0 represents the baseline level of the response variable `count` when all predictors (`year`, `month`, `day`, and `hour`) are zero. A relatively broad prior distribution, $\mathcal{N}(0, 5)$, is assigned to β_0 due to the lack of strong prior knowledge about the baseline demand. This allows the model to flexibly learn the actual baseline demand from the data. Additionally, given the potential for significant variability in baseline usage levels at different times, the broad prior ensures that the model can account for this uncertainty effectively.

For the coefficients $\beta_1, \beta_2, \beta_3, \beta_4$, which correspond to the predictors `hour`, `day`, `month`, and `year`, a narrower prior distribution, $\mathcal{N}(0, 2)$, is used. This reflects the assumption that the effects of these variables on `count` are likely to be moderate and within a reasonable range. The smaller standard deviation imposes stronger constraints on these effects, preventing them from causing unrealistic shifts in the predictions. At the same time, the prior remains flexible enough to allow the model to learn the actual effects of these time-related factors on demand patterns from the data.

This study employs the Poisson model, which assumes that the events under investigation are discrete and countable, occurring within a fixed time or spatial interval. Furthermore, it assumes that the occurrences of events are independent of each other, and the average rate of occurrence (λ) remains constant over the specified interval. Lastly, it is assumed that at most one event can occur within an extremely short time or spatial interval. These assumptions enable the Poisson distribution to effectively describe the probabilistic characteristics of the events.

The model was chosen based on the nature of the variable *count*, as this study hypothesizes that a Poisson distribution may better explain its discrete and non-negative characteristics. The selection of predictors is informed by the data structure: *year* is included to capture the overall upward trend over time, *month* and *day* account for seasonal variations, and *hour* reflects daily usage differences. This approach ensures that the model effectively addresses the temporal and seasonal dynamics inherent in the data.

This study considered using the ARMA (AutoRegressive Moving Average) model as an alternative but ultimately decided against it due to its limitations in capturing the specific characteristics of bike-sharing usage data. Bike-sharing usage exhibits significant seasonal patterns, time-of-day effects, and count-based characteristics, while ARMA models are primarily suited for stationary time series with linear autocorrelations. ARMA struggles to effectively address non-stationary trends and overdispersion, such as the marked decrease in usage during winter.

Although extensions like SARIMA (Seasonal ARIMA) can account for seasonality, they still cannot directly handle the count-based nature of the data. In contrast, Bayesian Poisson regression is specifically designed for count data, allowing for the flexible integration of variables such as year, month, and specific four-hour intervals. Additionally, the Bayesian framework provides posterior distributions, offering deeper insights into parameter uncertainty and variability, which are critical for policy decisions. Therefore, Bayesian Poisson regression is more suitable for this study than the ARMA model.

3.2 Validation

Figure 4 is the residual plot. It demonstrates several strengths of the model. First, the residuals are primarily distributed around the zero line, indicating that the model does not exhibit significant systematic bias in its predictions. This suggests that the model is generally effective at capturing the central trends of the data. Second, the residual distribution shows a consistent linear pattern, reflecting the model's stability across different levels of fitted values. This consistency implies that the model performs reliably in its predictions for various ranges of the response variable. Third, the model successfully captures localized patterns in the data, as evidenced by the structured yet controlled behavior of the residuals.

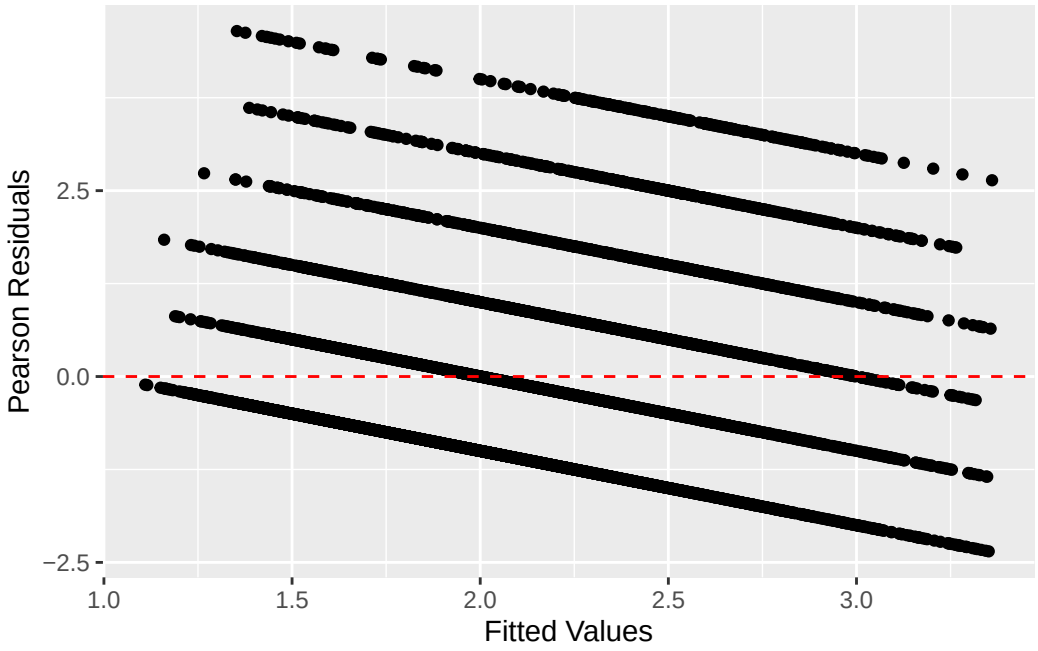

Figure 4: Pearson Residuals vs Fitted Values

Figure 5 is the posterior predictive check plot. The posterior predictive check plot highlights

several strengths of the model. The predicted distribution (y_{rep}) closely aligns with the observed distribution (y), particularly at the primary density peak near 0 and the secondary peaks within the 2-5 range, indicating that the model effectively captures the overall data distribution. Additionally, the model successfully reflects the multimodal nature of the data, accurately identifying multiple peaks, which demonstrates its capability to handle complex data patterns. The tail behavior of the predicted distribution is also consistent with the observed distribution, showing that the model performs well in accounting for rare or extreme events. Furthermore, the uncertainty of the predicted distribution is appropriately moderate, neither too narrow to indicate overconfidence nor too wide to suggest excessive variability.

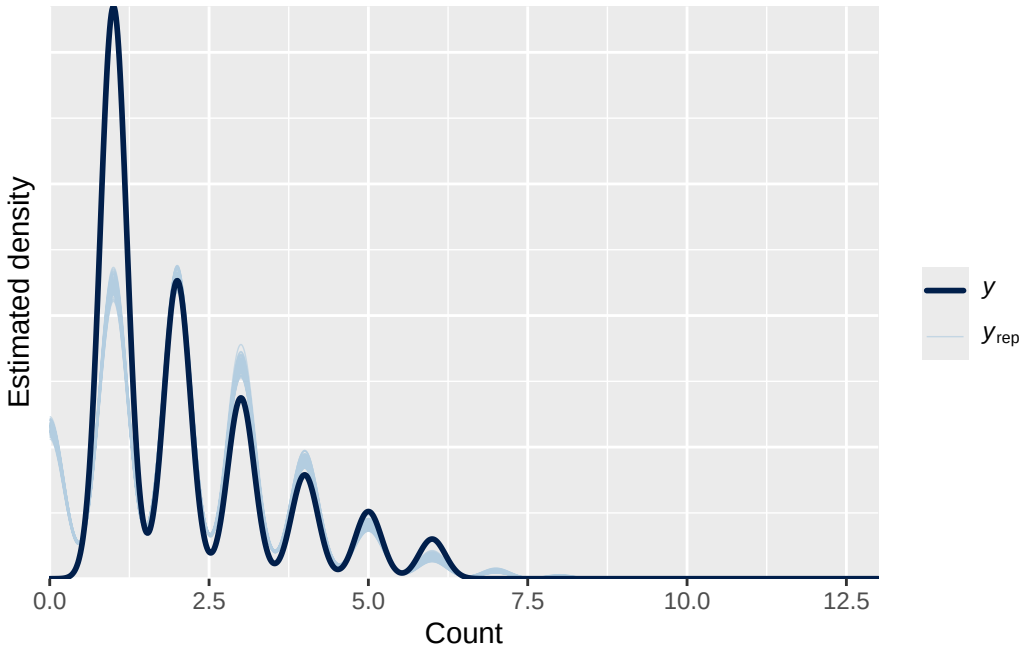

Figure 5: Posterior Predictive Checks

The credible intervals shown in Table 3 highlight several strengths of the model. Key variables such as hour, month, and year have narrow credible intervals that do not include zero, indicating their significant and consistent influence on the response variable. This demonstrates the model's ability to effectively capture critical trends and relationships in the data, particularly those related to time and seasonal effects. Additionally, the relatively small width of these intervals reflects the stability and reliability of parameter estimates, suggesting that the model is well-calibrated and robust in its predictions. The intercept, while having a broader credible interval, provides a reasonable baseline estimate, reflecting the overall level of the response variable.

Table 3: Credible intervals

	2.5%	97.5%
(Intercept)	-194.34770387	-164.75862551
hour	0.01472614	0.02020263
day	-0.00300555	0.00050100
month	0.02355016	0.03368799
year	0.08172812	0.09633039

4 Results

This study ultimately predicted the usage volume of 27 shared bicycle stations within the University of Toronto's St. George campus during a 4-hour time window from 4:00 AM to 8:00 AM on September 26, 2025. The prediction was then compared to the usage volume recorded during the same period in 2024. Figure 6 shows the results.

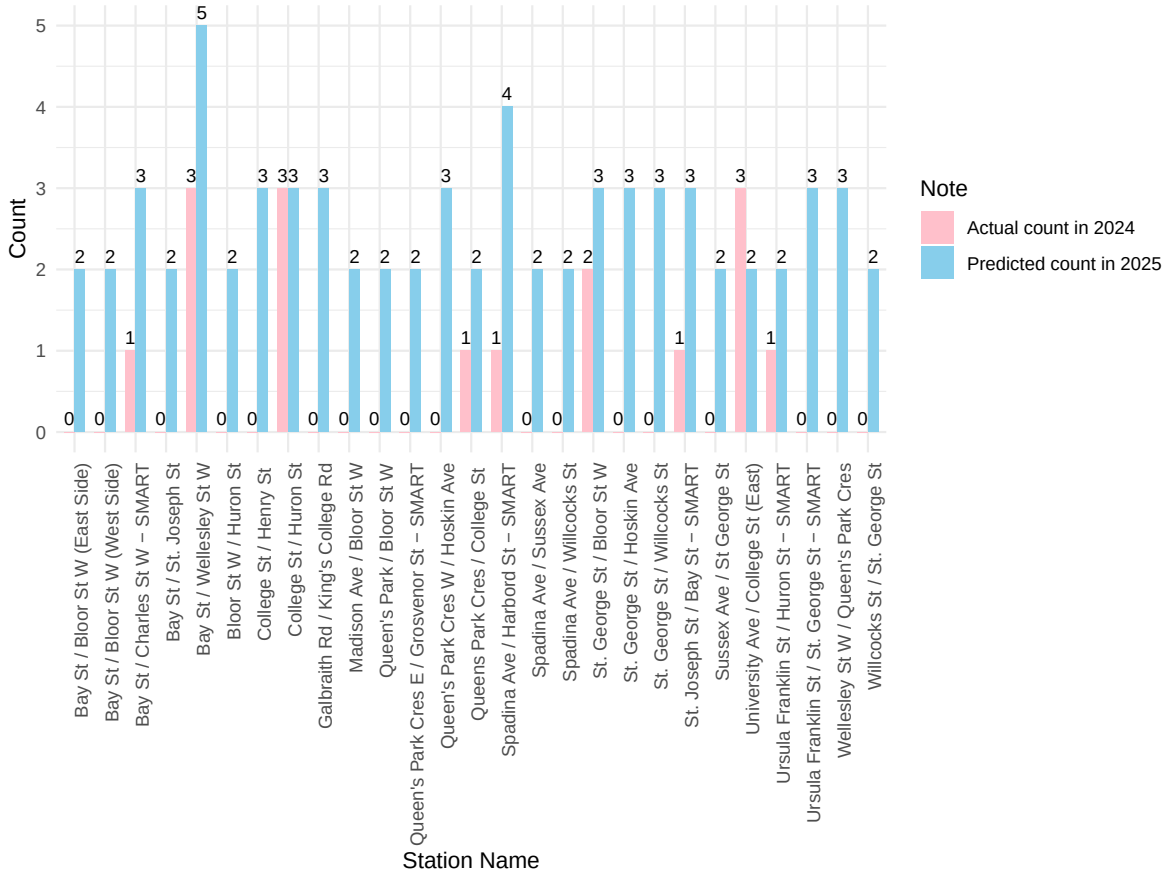

Figure 6: The actual count from 4:00 AM to 8:00 AM on September 26, 2024 vs the predicted count for the same time interval in 2025.

The prediction indicates an average increase of 2 uses per station across the 27 shared bike stations between 4:00 AM and 8:00 AM on September 26, 2025, compared to the same period in 2024. Specifically, 8 stations are projected to see an increase of 3 uses, 14 stations an increase of 2 uses, and 3 stations an increase of 1 use. Meanwhile, one station is expected to experience no change, and another is forecasted to see a decrease of 1 use.

5 Discussion

5.1 What the Study Brought to the World

The study reveals a few key insights about bike-sharing usage on university campuses. Specifically, we learn that bike-sharing usage on the University of Toronto's St. George campus

exhibits significant temporal patterns, influenced by season and time of day. For instance, usage drops significantly during winter months and peaks from 8 AM to 8 PM, with little demand between midnight and 8 AM. This indicates that bike-sharing systems need to adapt dynamically to changing demand patterns to be more effective, with an emphasis on considering seasonal and daily time variations for better resource allocation.

The findings provide valuable lessons for urban planners and campus management on improving bike-sharing infrastructure by optimizing the number and distribution of bikes according to the predicted demand in different seasons and time slots. This also underscores the broader principle that effective urban transport planning, even at a small scale like a university campus, requires a data-driven approach to match supply with temporal demand.

5.2 Limitations

This study employs a Bayesian Poisson regression model, which, while advantageous for handling count data, has significant limitations due to its assumptions. The Poisson model assumes that bike usage events are independent of one another, ignoring interactions among user behaviors, such as multiple users arriving at a station simultaneously during peak hours. Additionally, the model presumes a constant usage rate within each time period, which fails to capture dynamic fluctuations in reality, such as those caused by weather changes or campus events.

The study only considers temporal variables (e.g., year, month, day, and hour) and excludes other critical external factors, significantly limiting the accuracy of its predictions. Weather conditions, for instance, have a substantial impact on bike usage; rainy, snowy, or extremely hot or cold days see a marked decline in usage frequency. Moreover, special events on campus and holidays directly affect bike demand. Neglecting these external factors prevents the model from accurately capturing fluctuations in actual demand.

The handling of temporal variables in the model is relatively simplistic, overlooking potential complex interactions among these variables, which restricts the predictive performance. Poisson regression struggles to adequately address the high concentration of demand during peak periods, potentially underestimating peak usage. Consequently, more sophisticated and flexible models, such as random forests or deep learning approaches, may be better suited to capturing these intricate relationships and improving prediction accuracy.

The study also falls short in validating the model's generalization capability. Although residual analysis and posterior predictive checks were used to evaluate model fit, there was no assessment of the model's predictive performance on unseen data, such as cross-validation or evaluation in different scenarios. Broader validation would enhance the model's robustness and reliability in practical applications.

Additionally, choices made during the data cleaning process may affect the accuracy of the study's results. Excluding data from 2015–2016 improved data consistency but may have overlooked long-term trends in the evolution of the bike-sharing system. Furthermore, the way outliers and missing values were handled during data cleaning could significantly influence the model's outcomes. A lack of systematic and precise data standardization could lead to inconsistent model performance across data from different years, thereby affecting the reliability and accuracy of the predictions.

5.3 Suggestions for Future Research

While this study successfully predicts bike-sharing usage at the University of Toronto's St. George campus, several areas warrant further exploration to enhance the reliability and applicability of the results.

First, expanding the dataset to include additional environmental variables could provide deeper insights. Incorporating factors such as weather conditions, traffic congestion data, and socioeconomic variables could improve the model's accuracy, especially for predicting daily or hourly usage variations. Future studies could also include other campus transportation modes to enable a more comprehensive analysis of mobility patterns within the campus.

Additionally, this research is limited to employing Bayesian Poisson regression. Exploring more advanced machine learning methods, such as random forests or neural networks, could improve prediction accuracy by capturing the nonlinear relationships among factors influencing bike-sharing usage.

Another potential avenue for future research is extending the study beyond the campus. Comparing usage patterns across multiple university campuses could uncover broader trends and differences in bike-sharing demand. Such comparative analyses would help researchers and policymakers determine whether the usage patterns observed at the University of Toronto are unique or represent generalizable phenomena across other campuses.

Appendix

A Data collection methodology

A.1 Overview

Bike Share Toronto is the largest bike-sharing service provider in the Greater Toronto Area, committed to promoting gas-free transportation since its establishment. Starting with 79 stations in 2014, the system has expanded rapidly over the past decade. As of September 2024, the number of stations has grown to 855, capturing the entire bike-sharing market in Toronto (Liu and Allen 2023).

Bike Share Toronto uses its advanced bike-sharing system to automatically collect and manage ride data. Each bike and docking station is equipped with sensors and communication devices to record trip data in detail. This data is transmitted in real time to a central database via a wireless network, used for system monitoring, operational maintenance, and user behavior analysis. To support academic research and urban planning, Bike Share Toronto also anonymizes the data and makes it publicly available. This dataset includes all ride records since the system's inception and is therefore considered a population rather than a sample.

With ongoing system and technological updates, the data collected and published each year varies. Data from 2014-2015 primarily captured hourly usage on weekdays and weekends. Starting in 2016, more detailed data was collected, including trip start and end stations, start and end times, trip duration, trip ID, and user type. From 2017 to 2023, the data was further refined with the addition of unique identifiers for each station (station ID). In 2024, information on bike models was added. The data release cycle also varied across different periods, with some data released quarterly and others monthly. Despite the consistency in the recorded content each year, variable names differ.

A.2 Evaluation

In evaluating the data collection methods, I focused on accuracy, completeness, consistency, and relevance to ensure the credibility and scientific quality of the research. Below is a detailed assessment of the Bike Share Toronto dataset:

Accuracy
Bike Share Toronto records ride data in real time through sensors and communication devices, an automated collection method that effectively reduces errors associated with manual recording. Additionally, data is directly stored in a central database and updated in real time, enhancing accuracy. However, sensor devices may be subject to hardware failures or communication interruptions, leading to minor data loss or errors. Although such occurrences are rare, they need to be addressed during data preprocessing.

Completeness
The dataset covers all ride records since 2014, with data variables expanding as the system and technology improved. This historical and comprehensive nature allows the data to reflect long-term trends in Toronto's bike-sharing system. However, the limited variables in early data (e.g., only hourly usage for 2014-2015) require researchers to be mindful of inconsistencies across years, and early data may need supplementation or transformation to meet research needs.

Consistency
Although the Bike Share Toronto dataset has expanded annually with new variables, changes in variable names can pose challenges for data analysis. For instance, some variables have been named differently across different years, requiring standardization during data cleaning. Additionally, the variation in data release cycles (quarterly or monthly) could affect the continuity of time series analysis, necessitating data resampling or aggregation to ensure consistency.

Relevance
The detailed variables in the dataset (such as station ID, ride duration, and user type) provide opportunities for multidimensional analysis, enabling researchers to explore geographic distribution, temporal patterns, and user behavior. Since the dataset represents a population rather than a sample, researchers have greater flexibility in their study focus, ranging from micro-level user behavior to macro-level trend analysis. However, this comprehensiveness also increases the difficulty of data cleaning, such as dealing with redundant data, outliers, and the computational burden of large-scale data processing. Moreover, anonymization protects user privacy, it may limit certain studies focusing on micro-level user behavior.

B Idealized methodology

In an ideal research methodology, the design of data collection should emphasize comprehensiveness, systematic approach, and relevance, ensuring that the collected data can adequately reflect the multidimensional characteristics of the research subject. Specifically, in addition to core data from the bike-sharing system (such as ride duration, start and end stations, riding time, etc.), external environmental factors should also be considered, including weather conditions (temperature, precipitation, wind speed), holiday information (whether it is a public holiday, behavioral changes before and after holidays), traffic flow (interactions between shared bikes, motor vehicles, and public transportation), and socio-economic factors of the region (income levels, population density, and urbanization degree of different communities). These external variables provide richer contextual information for data analysis, thereby supporting more precise research conclusions.

Moreover, the comprehensiveness of data collection should not only reflect the diversity of variables but also consider continuity in temporal and spatial dimensions. Ideally, the data should cover the full-year operation of all stations in the bike-sharing system, with high-frequency recording (e.g., per minute or second) of trip dynamic data to avoid biases or misinterpretations due to data loss. A unified data format and standardized naming conventions should

be adopted to ensure compatibility between different time periods and various data sources, thereby achieving seamless data integration.

To address potential information gaps caused by facility malfunctions, several solutions can be implemented: firstly, multiple data collection devices should be integrated into the system to minimize the impact of a single device failure on overall data integrity. For instance, sensors on bicycles can serve as backups for station records, ensuring that even if one side fails, the other can still provide critical data. Secondly, the central database can be designed to automatically fill in missing data using mechanisms such as utilizing historical data, average values from neighboring stations, or similar time periods, thus reducing the impact of information gaps on the analysis results.

B.1 Ideal Dataset Framework

A well-designed ideal dataset framework should be comprehensive, flexible, and structured, enabling multidimensional analysis and dynamic scalability. Below is the proposed framework for a bike-sharing system dataset:

Temporal Dimensions

- Date: Year/Month/Day, supporting time-series analysis.
- Timestamp: Precise to the second, enabling high-precision analysis.
- Time Period: Categorized into morning peak, evening peak, weekdays/weekends, etc., for behavioral pattern analysis.

Spatial Dimensions

- Start Station ID: A unique identifier for the starting station.
- Start Station Name: Readable names for better understanding.
- End Station ID: A unique identifier for the destination station.
- End Station Name: Helps with geographic visualization.
- Latitude/Longitude: Precise coordinates for spatial analysis.

Trip Information

- Trip ID: A unique identifier for each trip.
- Trip Duration: Duration of the trip in seconds, indicating the trip length.
- Trip Distance: Distance traveled in meters, providing spatial context.
- Bike ID: A unique identifier for each bike.

User Information

- User Type: Differentiates between members and casual users.
- User Age: Estimated age range of the user, supporting demographic analysis.
- User Gender: Categorized as male, female, or other/unspecified.

Environmental Factors

- Weather Conditions: Includes temperature, humidity, precipitation, and wind speed.
- Air Quality Index: Provides environmental health metrics.
- Traffic Conditions: Includes road congestion index or traffic volume.

System Status

- Available Bikes at Start Station: Number of bikes available at the starting station at trip initiation.
- Available Docks at End Station: Number of free docks at the destination station at trip conclusion.
- Device Status: Indicates whether the equipment (e.g., sensors) is functioning properly.

Operations and Management

- Maintenance Records: Includes repair times and reasons for bikes and station facilities.
- Anomalies: Records equipment failures, data abnormalities, etc.

Data Quality

- Missing Data Flag: Indicates whether data is missing or contains substituted values.
- Data Source: Records the source or method of data collection (e.g., station sensors, user applications).

Table 4, Table 5, Table 6, Table 7, Table 8, Table 8 show a simulation of proposed ideal dataset.

Table 4: Simulated Ideal Dataset Part 1

Date	Timestamp	Time_Period	Start_Station_ID	Start_Station_Name
2024-10-31	2024-11-28 15:43:45	Weekend	188	Station 169
2024-11-26	2024-11-28 02:39:12	Weekday	137	Station 179
2024-11-01	2024-11-28 07:57:41	Evening Peak	154	Station 101

Table 4: Simulated Ideal Dataset Part 1

Date	Timestamp	Time_Period	Start_Station_ID	Start_Station_Name
2024-11-25	2024-11-28 04:38:44	Weekend	188	Station 104
2024-11-07	2024-11-28 15:07:06	Morning Peak	118	Station 169

Table 5: Simulated Ideal Dataset Part 2

End_Station_ID	End_Station_Name	Latitude	Longitude	Trip_ID
168	Station 178	43.66700	-79.31730	TRIP-3444
138	Station 142	43.60153	-79.35507	TRIP-9493
169	Station 119	43.60394	-79.36377	TRIP-1844
154	Station 125	43.69338	-79.34735	TRIP-5114
183	Station 153	43.61786	-79.33301	TRIP-8871

Table 6: Simulated Ideal Dataset Part 3

Trip_Duration_s	Trip_Distance_m	Bike_ID	User_Type	User_Age
941	5487	BIKE-7679	Member	60
1234	7811	BIKE-5803	Casual	19
2208	2733	BIKE-7290	Member	57
2946	7828	BIKE-2495	Member	29
2111	3037	BIKE-6856	Member	18

Table 7: Simulated Ideal Dataset Part 4

Weather_Conditions	Air_Quality_Index	Traffic_Conditions	Available_Bikes_at_Start_Station
Cloudy	73	High	18
Rainy	62	Low	20
Cloudy	14	High	18
Cloudy	14	High	9
Rainy	68	Moderate	4

Table 8: Simulated Ideal Dataset Part 5

Available_Docks_at_End_Station	Device_Status	Maintenance_Records
19	Faulty	Minor Repair
18	Operational	Minor Repair
1	Faulty	None
19	Faulty	None
10	Operational	Minor Repair

Table 9: Simulated Ideal Dataset Part 6

User_Gender	Anomalies	Missing_Data_Flag	Data_Source
Male	Sensor Error	FALSE	Mobile App
Other	None	TRUE	Station Sensor
Male	Docking Failure	FALSE	System Log
Other	Sensor Error	TRUE	Mobile App
Male	None	TRUE	Station Sensor

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Bike Share Toronto. n.d. "How It Works." https://bikesharetoronto.com/how-it-works/.

Bürkner, Paul-Christian. 2017. *Brms: An r Package for Bayesian Multilevel Models Using Stan*. *Journal of Statistical Software*. Vol. 80. https://doi.org/10.18637/jss.v080.i01.

El-Assi, Wafic, Mohamed Salah Mahmoud, and Khandker Nurul Habib. 2017. "Effects of Built Environment and Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing in Toronto." https://doi.org/10.1007/s11116-015-9669-z.

Gabry, Jonah et al. 2019. *Bayesplot: Plotting for Bayesian Models*. *Journal of Statistical Software*. https://doi.org/doi:10.1111/rssa.12378.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Liu, Michael, and Jeff Allen. 2023. "Exploring Bike Share Growth in Toronto." https://schoolofcities.github.io/bike-share-toronto/growth.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Shenzhen Open Data Platform. 2021. "Shenzhen Public Bicycle Dataset." 2021. https://opendata.sz.gov.cn/data/dataSet/toDataDetails/29200_00403627.

Tang, Yang, Weiwei Liu, Chennan Zhang, Yihao He, Ning Ji, and Xinyao Chen. 2020. "Research on the Traveling Characteristics and Comparison of Bike Sharing in College Campus–a Case Study in Hangzhou." https://file.techscience.com/uploads/attached/file/20200916/20200916072445_75552.pdf.

Toronto Parking Authority. n.d. "About Us." https://parking.greenp.com/about/about-us/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. *Journal of Statistical Software*. Vol. 40. https://doi.org/10.18637/jss.v040.i06.

Zhang, Xuxilu, Lingqi Gu, and Nan Zhao. 2024. "Navigating the Congestion Maze: Geospatial Analysis and Travel Behavior Insights for Dockless Bike-Sharing Systems in Xiamen." *arXiv:2401.03987*. https://arxiv.org/abs/2401.03987.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.