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

Daily usage grows by one every 20 days, drops to between 0 and 3 in winter, with 79% occurring between 8 AM and 8 PM

Haowei Fan

December 2, 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 show that daily bike-sharing usage on campus increases by one every 20 days but decreases to between 0 and 3 in winter months, and 79% 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

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

1 Introduction

1.1 Overview

Bike-sharing services have experienced 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 increase highlights the need for 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 suggestions, 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 unexplored areas 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 detailed understanding of its role in campus mobility and inform policies tailored to similar settings.

The estimand of this study aims to predict the usage of a specific bike-sharing station on the University of Toronto campus during a particular 4-hour interval. By transforming time into components such as year, month, day, and the specific 4-hour interval of the day, the analysis incorporates overall trends, seasonal effects, and hourly variations in station usage. The primary objective is to investigate temporal changes in the usage of 27 bike-sharing stations on campus, thereby providing policymakers with recommendations for optimizing bike allocation to enhance commuting efficiency. To achieve this estimand, the study extracts a sample of 27 bike-sharing stations within the University of Toronto campus from all bike-sharing usage data recorded between January 1, 2017, and September 30, 2024. A Bayesian Poisson regression model is then employed to predict station usage based on factors such as station location, year, month, day, 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 by an average of one user every 20 days each year, there are significant seasonal differences. During winter, usage at all stations, regardless of the year, remains close to zero between 0 to 3. 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% of daily usage, while the period from midnight to 8 am accounts for only 6%. 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 St. George campus. They also provide guidance on how to better allocate, deploy, or store shared bikes.

The structure of this paper is as follows: Section 2 outlines the data sources, evaluates the data measurements, introduces key variables along with their visualizations, presents similar datasets, provides a high-level description of the data cleaning process, and details the technical support utilized in this project. Section 3 provides an in-depth explanation of the model types used in this study and the rationale behind the selection. It introduces the target variable, predictors, and all components of the model. This section also includes model validation and discusses an alternative model. Section 4 uses a specific time frame in 2025 as an example to make predictions using the model, comparing the results with 2024 data. Section 5

explores the implications of this study for the world, acknowledges its limitations, and offers recommendations for future research.

2 Data

2.1 Overview

The dataset used in this study is named "Bike Share Toronto Ridership Data" (Toronto Parking Authority 2024). It 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.

There are many similar datasets available, but they pose significant limitations in the context of this study. 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 this 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 this study's 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 this 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.

scales(Wickham, Pedersen, and Seidel 2023): Used for formatting large numbers.

reshape2(Wickham 2007): Used for reshaping redata, particularly converting between wide and long formats

Additionally, this project utilized OpenAI's ChatGPT-4 (OpenAI 2024) to assist in writing parts of the code.

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 "Bike Share Toronto Ridership Data" (Toronto Parking Authority 2024), which spans over eight years to date and documents every instance of bike-sharing usage, including the time and location of each ride, is ideal for this study.

This data is particularly ideal for the study also due to its high reliability, which stems from the commercial nature of bike-sharing operations. The bike-sharing system ensures accurate data collection and management by automatically recording ride information. Each bike and docking station is equipped with sensors and communication devices, enabling the capture of detailed trip data.

However, the raw data cannot be used directly in this project; sophisticated data cleaning is required, with general steps as follows: Data prior to 2017 was excluded, and the dataset was filtered to focus on 27 locations within the St. George campus of the University of Toronto. Essential variables (`trip_start_time` and `from_station_name`) were retained, missing values were removed, and `trip_start_time` was standardized and transformed into 4-hour intervals for temporal analysis. The total usage count for each station during these intervals from January 1, 2017, to September 30, 2024, was calculated. Finally, variables were renamed (`trip_start_time` to `time` and `from_station_name` to `station_name`). The cleaned data is presented in the format shown in Table 1. And a more detailed data cleaning process and reasons can be found in Appendix Section C.

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. This

is a constructed variable derived by calculating the total number of uses at a specific station within a given four-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". This is a constructed variable designed to indicate which four-hour interval the exact trip start time, recorded to the second in the raw data, falls into.
- `station_name`: An independent variable representing the unique name of one of the 27 stations within the University of Toronto's St. George campus.

It is important to note that there may be a relationship between the two independent variables, `time` and `station_name`, as the operating times of different stations might vary. However, this does not affect the validity of the study because the analysis focuses on the usage patterns within the available operational hours for each station, rather than comparing stations with differing start times directly.

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: Summary of Key Variables

station_name	time	count
Type: Character	Type: Character	min: 1
Number of Unique Values: 27	Earliest: 2017-01-01 04:00:00	1st Qu: 1
	Latest: 2024-09-30 20:00:00	median: 1
		mean: 2
		3rd Qu: 3
		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 usage of the 27 bike-sharing stations at the University of Toronto's St. George campus has shown a steady upward trend, with an average increase of 1 use every 20 days. At the same time, there are clear seasonal variations: usage is higher in the spring and summer months (April to September) and significantly lower in the fall and winter months (October to March). Additionally, there was an abnormal surge in usage from June to September 2024, which can be attributed to two main reasons. First, historical data indicates that usage typically spikes in August and September each year, likely

due to increased demand from students returning to campus for the start of the academic year. Second, during the same period in 2024, Bike Share Toronto added 52 new stations, most of which were located near the St. George campus, significantly enhancing the network's coverage and convenience(Bike Share Toronto, n.d.). These two factors combined led to the appearance of anomalies in August and September 2024.

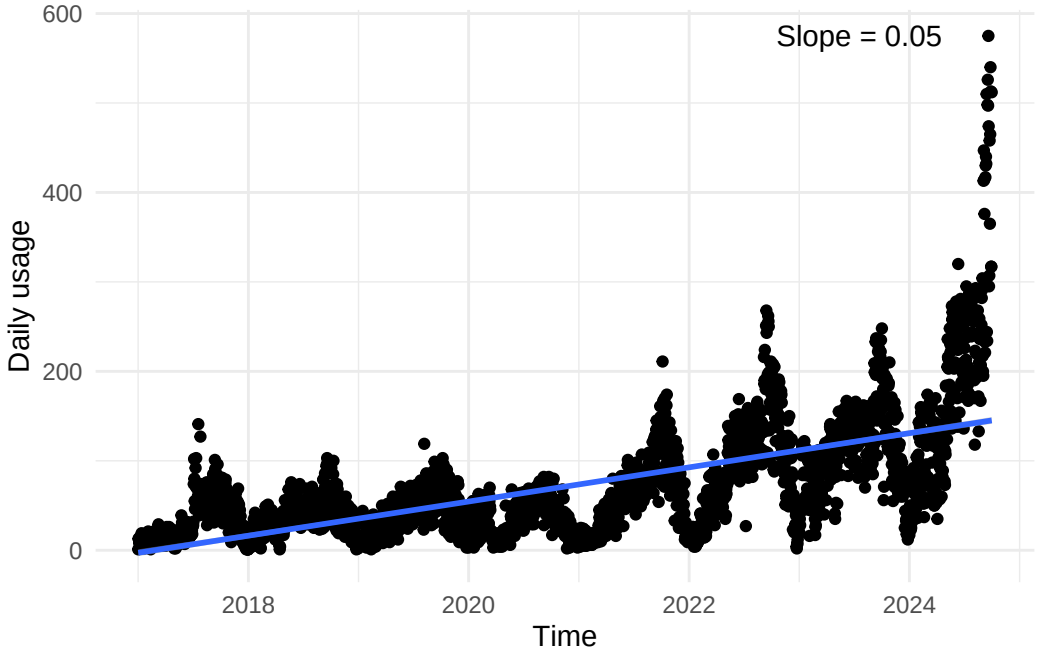

Figure 1: Daily usage from January 2017 to September 2024

Figure 2 shows 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%).

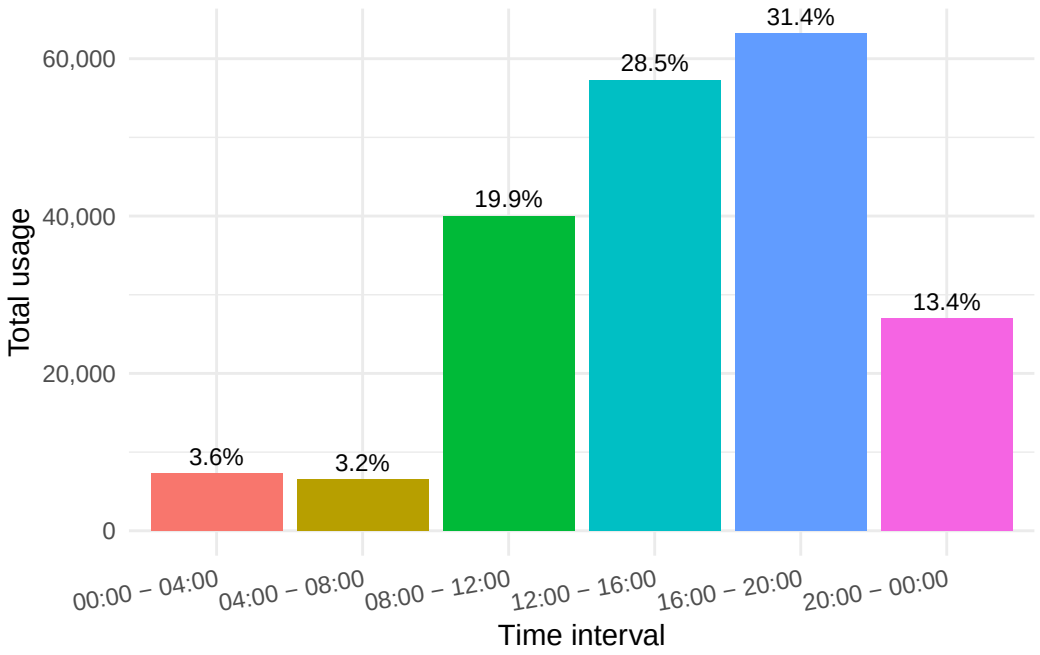

Figure 2: Total usage during different time periods from January 2017 to September 2024

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.

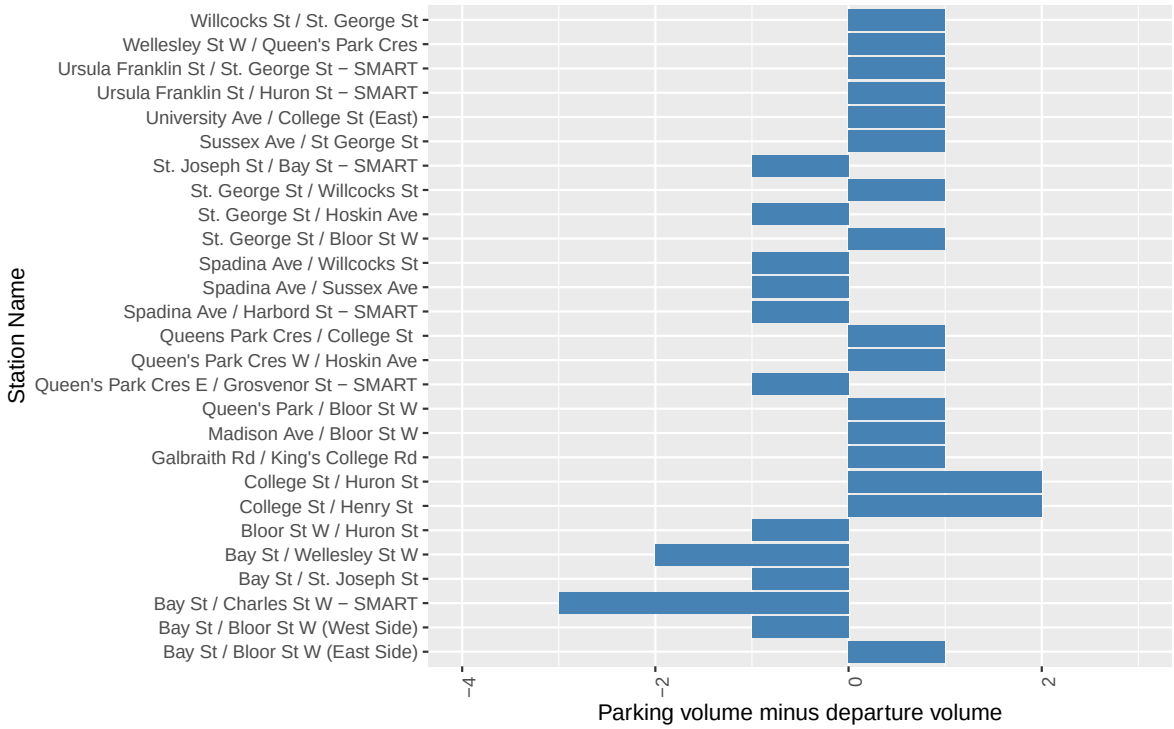

Figure 3: Difference of daily average of bicycle parking and departure volumes for each station

3 Model

In the process of model development, this study utilized R (R Core Team 2023) and the following R packages:brms(Bürkner 2017) and rstanarm(Goodrich et al. 2022) were used for model construction. bayesplot(Gabry et al. 2019), tidyverse(Wickham et al. 2019), ggplot2(Wickham 2016), knitr(Xie 2014), and kableExtra(Zhu 2024) were used for verifying model assumptions and model validation.

3.1 Model Descriptions

Equation1 is the model used in this study:

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

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 λ represents the average number of times a specific station is used every four hours, and $\log(\lambda)$ is the logarithmic transformation of λ. 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.

All predictors in this study are treated as categorical variables rather than time-based or numerical, aiming to reduce potential latent correlations between samples caused by temporal continuity.

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 ARIMA (autoregressive integrated 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. Although the ARIMA model has advantages in time series analysis, allowing for more accurate short-term predictions with parameters that have clearer interpretability. Bike-sharing usage exhibits significant seasonal patterns, time-of-day effects, and count-based characteristics, while ARIMA models are primarily suited for stationary time series with linear autocorrelations. ARIMA 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 ideas into parameter uncertainty and variability, which are important for policy decisions. Therefore, Bayesian Poisson regression is more suitable for this study than the ARIMA model.

3.2 Validation

In this section, the study selects "Galbraith Rd / King's College Rd" as the most representative station to validate the model. This station was chosen for several reasons: it is centrally located within the University of Toronto's St. George campus, it ranks among the top five stations in terms of data volume, and it has maintained the shortest average maintenance time every six months since its establishment(Bike Share Toronto, n.d.).

Figure 4 illustrates the comparison between Pearson residuals and fitted values for the model of the "Galbraith Rd / King's College Rd" station. 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.

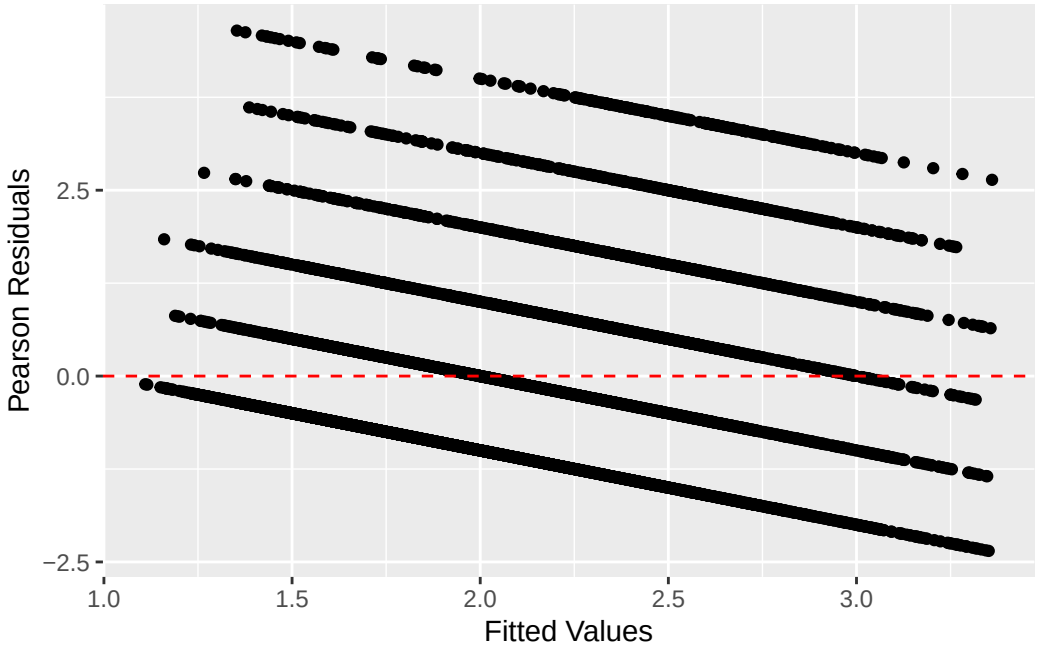

Figure 4: Pearson Residuals vs Fitted Values for "Galbraith Rd / King's College Rd" Model

Figure 5 is the posterior predictive check plot of the model of the "Galbraith Rd / King's College Rd" station. 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.

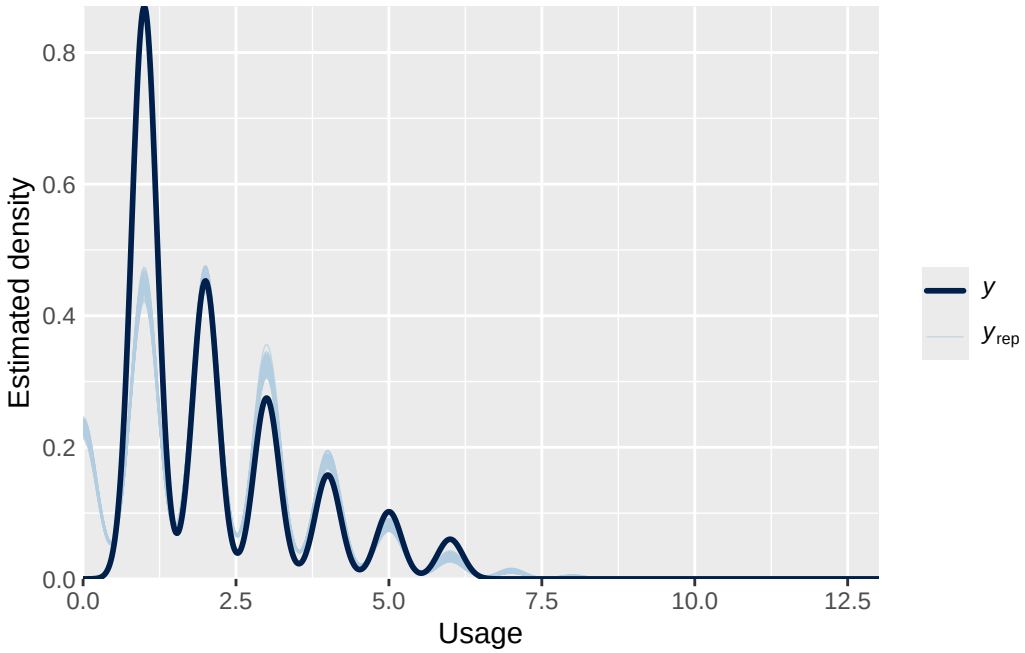

Figure 5: Posterior Predictive Checks for "Galbraith Rd / King's College Rd" Model

The credible intervals shown in Table 3 highlight several strengths of the model of the "Galbraith Rd / King's College Rd" station. 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 important 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 of model coefficients for "Galbraith Rd / King's College Rd" Model

	2.5%	97.5%
(Intercept)	-194.35	-164.76
hour	0.01	0.02
day	0.00	0.00
month	0.02	0.03
year	0.08	0.10

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.

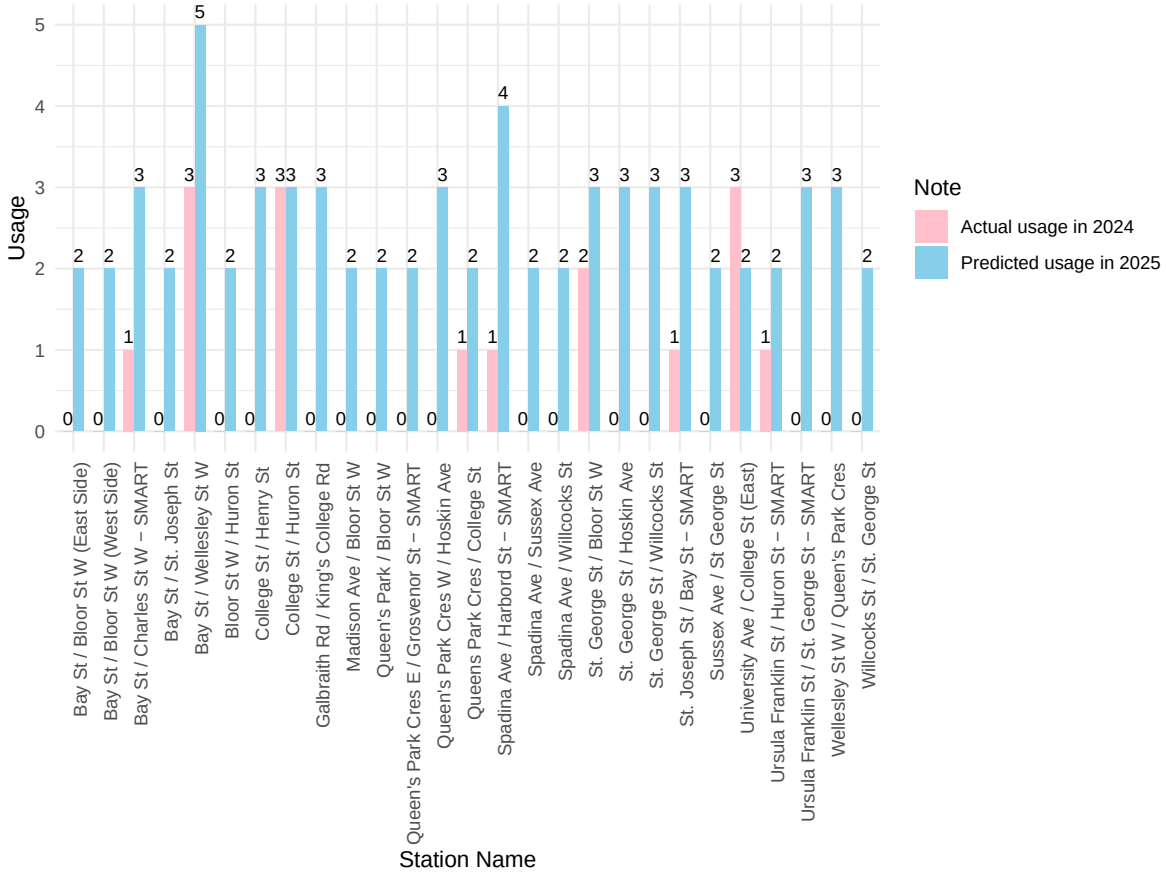

Figure 6: The actual usage from 4:00 AM to 8:00 AM on September 26, 2024 vs the predicted usage for the same time interval in 2025.

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 Optimization Suggestions for Campus Bike-Sharing Systems

This study identifies significant patterns in the use of bike-sharing systems on the University of Toronto's St. George campus, including temporal, seasonal, and annual trends. The findings shows that 79.8% of bike usage occurs between 8 AM and 8 PM, while demand between midnight and 8 AM is minimal. This highlights the strong alignment between bike usage and users' daily routines. To improve commuting efficiency, it is recommended that the university collaborates with the bike-sharing provider to increase bike availability during peak hours (e.g., before and after class times) and optimize station layouts, particularly in high-demand areas such as dormitories, academic buildings, and transit hubs. During low-demand hours (e.g., early morning), maintenance and redistribution efforts can be scheduled to ensure the availability of bikes during peak periods.

Seasonal data also show that bike usage significantly decreases during winter, approaching zero, while increasing steadily during spring and summer, peaking in August and September during the start of the academic year. This suggests the need for proactive and dynamic adjustments in bike supply. During the start of the academic year, expanding station capacity and offering user-friendly navigation services can help new and returning students quickly familiarize themselves with the system. In winter, efforts should focus on improving station infrastructure (e.g., anti-freeze facilities) and maintaining safety under cold conditions. Additionally, by utilizing predictive analytics to forecast station-specific demand fluctuations, resources can be optimized, minimizing inconveniences caused by bike shortages or overcrowded parking during peak times.

5.2 A New Research Methodology for Shared Bikes in Semi-Closed Areas

This study provides actionable recommendations for improving user experience in semi-closed environments like university campuses. Unlike urban open systems, bike usage in campuses is strongly influenced by fixed schedules such as class timings and campus events. This approach can be further extended to other similar semi-closed areas, such as industrial parks, manufacturing facilities, and residential complexes, where movement patterns are often dictated by predictable schedules, shifts, or local routines.

In these environments, bike-sharing systems can be optimized to align with peak demand periods, such as shift changes in factories, scheduled breaks in office parks, or recreational times in residential areas. By leveraging predictive analytics administrators can strategically allocate bikes and docking stations to match demand patterns, ensuring users have access when and where they need it most.

Moreover, this methodology is cost-effective. By targeting specific time windows and locations with higher demand, resources can be distributed more efficiently, reducing waste and operational costs. This reduces the need for large-scale deployment of bikes and infrastructure, which is often necessary in open urban systems. Instead, investments can be focused on key high-demand areas within these semi-closed environments, maximizing return on investment while enhancing user satisfaction.

5.3 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 important 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 complicated 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.

5.4 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 research to enhance the reliability and applicability of the results.

First, expanding the dataset to include additional environmental variables could provide deeper understandings. 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 detailed analysis of mobility patterns within the campus.

Additionally, this research is limited to employing Bayesian Poisson regression. Exploring more 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 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 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 detailed 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 exhaustiveness 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

B.1 Overview

This study proposes an ideal data collection approach for bike-sharing systems, emphasizing key aspects like data diversity, reliable data collection, anomaly detection and correction, and ethical considerations.

B.2 Data Diversity

Diversity in data collection is important for enhancing research flexibility and ensuring reliable results. A well-rounded dataset for bike-sharing systems should consider several key aspects:

First, **data variable diversity** is fundamental. Beyond basic cycling metrics such as frequency, duration, distance, and station information, data should include geographic details (e.g., station latitude, longitude, topographical features), temporal information (e.g., timestamps, weekday or weekend classification, seasonal changes), environmental factors (e.g.,

weather, air quality, traffic), and socioeconomic indicators (e.g., regional income, population density). These dimensions help capture the range of factors that influence bike-sharing usage, enabling a detailed analysis of user behavior and external influences.

Spatial and temporal coverage also play a significant role in data representativeness. Spatially, data should include all bike stations, from central urban to suburban areas, to analyze varying usage patterns. Temporally, data should span several years to support trend analysis and should have high temporal resolution (e.g., minute-level precision) to capture short-term variations.

Ensuring **compatibility of data formats and platforms** is important for effective use of diverse datasets. Collected data should support structured formats (e.g., rental records), unstructured formats (e.g., user feedback), and geospatial information (e.g., GIS formats). Standardized formats like CSV, JSON, or Parquet along with metadata should be used for easy integration across platforms such as GIS systems, machine learning tools, and statistical analysis tools.

B.3 Reliable Data Collection Processes

To ensure reliability, bike-sharing data collection should involve a well-structured plan covering multiple aspects.

Automated data collection systems are essential for improving efficiency and accuracy. IoT-enabled sensors can be installed on bikes and docking stations to automatically record data such as trip start and end times, duration, and station occupancy. GPS and accelerometers on bicycles enable accurate tracking of movement patterns, distance, and speed. Wireless data transmission reduces manual errors and minimizes the risk of data loss during storage or transfer.

Implementing **redundant data collection mechanisms** is important for preventing data loss due to sensor malfunctions. Installing multiple sensors on bikes and stations serves as a backup system. For instance, sensors on bikes can complement the data from docking stations. A centralized monitoring system can cross-check data from multiple sources to detect inconsistencies or missing information, thereby improving data integrity and reliability.

To enhance data exhaustiveness, integrating **authoritative environmental and traffic data APIs** can be highly beneficial. For example, APIs from OpenWeatherMap or government meteorological services can provide real-time information on weather conditions, air quality, wind speed, and precipitation. Traffic data from relevant authorities can provide understandings into road congestion levels. Combining these variables with bike-sharing records can understand how external conditions affect user behavior, ultimately guiding bike allocation and station optimization.

B.4 Error Detection and Correction

Error detection and correction are vital for maintaining data quality and ensuring the stability of bike-sharing systems.

Real-time detection algorithms are the first line of defense against anomalies. Built on rule-based or machine learning techniques, these algorithms can quickly identify outliers. For instance, trip durations outside the typical range (e.g., less than 1 minute or more than 24 hours) should be flagged. Similarly, the number of available bikes at a station should always be between zero and the station's capacity, and any value outside this range should trigger an alert. GPS data should also be verified to confirm that bikes remain within designated operational areas.

Automated correction mechanisms are needed to address these anomalies. Three common correction methods include: (1) filling in missing data using historical trends—for example, estimating missing values based on average usage from similar periods; (2) referring to data from nearby stations to estimate missing information; and (3) using contextual data such as weather or traffic to adjust corrections, which ensures the data aligns with real-world scenarios. These methods help maintain accuracy while reducing the impact of missing or incorrect data.

Ensuring **data transparency** is important. All corrected records should be labeled with information about the sources, methods, and reasons for modification. This transparency helps maintain data integrity and fosters trust among researchers and system operators.

B.5 Ethical Considerations

Data collection in bike-sharing systems must take ethical considerations into account, particularly with respect to sustainability and user privacy.

For **sustainable data collection**, it is important to optimize both infrastructure and data management. Solar-powered equipment at docking stations can reduce dependence on traditional power sources by using clean energy for data collection. Dynamic data collection mechanisms that adjust frequency based on demand can help avoid unnecessary energy use. For example, reducing data collection during off-peak hours and increasing it during peak times can help balance operational needs while minimizing resource consumption. Sharing data collection infrastructure with other city services, such as traffic and environmental monitoring, can also reduce redundant investments.

In terms of **user privacy protection**, both technical and management measures are necessary. Anonymization techniques should replace identifiable information with pseudonyms, and differential privacy can be used to add noise to the dataset, preventing re-identification of users. Following the principle of data minimization, only essential data should be collected—for instance, avoiding the collection of detailed home addresses. Clear privacy policies should

inform users about data collection, use, and storage, while allowing them to access or delete their personal data. Regular third-party audits can assess privacy measures and identify areas for improvement, building user trust through transparency.

B.6 A Proposed Idealized Data Collection Checklist

This streamlined checklist ensures that the data collection process is efficient, ethical, and aligned with operational objectives.

1. Data Requirements

- ☐ **Clearly defined objectives:** Is the purpose of data collection (e.g., operational optimization, user behavior analysis) well-defined?
- ☐ **Key variables identified:** Are the data and variables necessary to meet the research requirements clearly defined and well-structured?
- ☐ **Data minimization:** Is only the data necessary to meet objectives being collected?

2. Ethical and Legal Considerations

- ☐ **User privacy protection:** Are anonymization techniques (e.g., pseudonymization, differential privacy) in place to protect user identities?
- ☐ **Informed consent:** Have users been informed about the data being collected and provided their consent?
- ☐ **Compliance with laws:** Is the process compliant with relevant regulations?
- ☐ **Transparency:** Are privacy policies clearly communicated and accessible to users?

3. Data Collection Methods

- ☐ **Functional devices:** Are IoT sensors, GPS, and accelerometers operating effectively for accurate data collection?
- ☐ **Optimized frequency:** Is the frequency of data collection adjusted dynamically based on operational needs (e.g., higher during peak hours)?
- ☐ **Integration with third-party sources:** Are APIs for weather, traffic, and environmental conditions incorporated?

4. Infrastructure and Resource Efficiency

- ☐ **Energy-efficient systems:** Are low-power devices and renewable energy sources (e.g., solar panels) used to minimize energy consumption?

- ☐ **Shared infrastructure:** Are existing city resources (e.g., traffic management systems) utilized to avoid redundancy?

5. Data Quality Assurance

- ☐ **Validation rules:** Are there checks for variable accuracy, such as valid GPS coordinates and positive trip durations?
- ☐ **Error detection:** Are algorithms in place to identify outliers, inconsistencies, and missing data?
- ☐ **Correction mechanisms:** Are processes for handling errors or filling missing data (e.g., using historical trends or nearby station data) established?

6. Data Security and Storage

- ☐ **Secure protocols:** Is data encrypted during storage and transmission
- ☐ **Retention policy:** Is a clear data retention timeline defined and adhered to?

7. Monitoring and Maintenance

- ☐ **Regular calibration:** Are sensors regularly calibrated to maintain accuracy?
- ☐ **Real-time monitoring:** Is a system in place to track device health and data flow?
- ☐ **Proactive maintenance:** Are preventive maintenance plans implemented to reduce downtime?

8. Feedback and Review

- ☐ **Stakeholder review:** Has the data collection process been reviewed by all relevant teams (e.g., operations, legal, ethics)?
- ☐ **User feedback:** Are mechanisms available for users to provide feedback or report concerns about data practices?

B.7 A Proposed Idealized Data Framework

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

Timestamp	Time_Period	Start_Station_ID	Start_Station_Name
2024-12-02 04:58:14	Morning Peak	107	Station 154
2024-12-02 10:16:37	Weekday	113	Station 188

Table 4: Simulated Ideal Dataset Part 1

Timestamp	Time_Period	Start_Station_ID	Start_Station_Name
2024-12-02 12:43:43	Weekend	172	Station 118
2024-12-02 07:24:19	Weekday	188	Station 169
2024-12-02 02:09:38	Morning Peak	137	Station 179

Table 5: Simulated Ideal Dataset Part 2

End_Station_ID	End_Station_Name	Latitude	Longitude	Trip_ID
101	Station 169	43.64769	-79.39606	TRIP-8362
104	Station 154	43.66215	-79.30661	TRIP-2736
169	Station 183	43.65516	-79.38214	TRIP-3444
168	Station 178	43.66700	-79.31730	TRIP-9493
138	Station 142	43.60153	-79.35507	TRIP-1844

Table 6: Simulated Ideal Dataset Part 3

Trip_Duration_s	Trip_Distance_m	Bike_ID	User_Type
318	5487	BIKE-7679	Member
941	7811	BIKE-5803	Casual
1234	2733	BIKE-7290	Member
2208	7828	BIKE-2495	Member
2946	3037	BIKE-6856	Member

Table 7: Simulated Ideal Dataset Part 4

Weather_Conditions	Air_Quality_Index	Traffic_Conditions	Available_Bikes_at_Start_Station
Rainy	47	Low	19
Cloudy	51	Low	2
Cloudy	28	High	18
Sunny	73	High	20
Cloudy	62	Low	18

Table 8: Simulated Ideal Dataset Part 5

Available_Docks_at_End_Station	Device_Status	Maintenance_Records
9	Faulty	None
4	Operational	Minor Repair
19	Faulty	Minor Repair
18	Operational	None
1	Faulty	None

Table 9: Simulated Ideal Dataset Part 6

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

C Detailed Data Cleaning Process

To facilitate unified processing of all data, this study consolidated all CSV files into a single dataset by standardizing variable names, ensuring consistency across different data sources.

For consistency in the study, all data prior to 2017 was removed, as the Open Data Toronto website specifies that data before 2017 was provided by a different vendor and collected using different statistical methods (Gelfand 2022). Another reason for excluding data prior to 2017 is: the earlier data did not include variables of interest, specifically the start time and station of each trip.

To collect the target sample, the study filtered the dataset to include only data of interest related to 27 specific locations within the St. George campus of the University of Toronto, focusing the analysis on the relevant geographical and contextual scope.

To simplify the dataset's structure and reduce its size for efficient processing, the study extracted only the necessary variables, including `trip_start_time` and `from_station_name`, which are important for subsequent analysis.

Considering that missing values represented a negligible proportion of the overall sample and would not significantly impact the results, the study removed all rows with `NA` or `NULL` values to maintain data integrity and ensure analytical accuracy.

To standardize the temporal information for consistent analysis, the format of `trip_start_time` was unified as "year-month-day hour:minute:second," providing a uniform reference for time-based computations.

In order to calculate the total usage across different time intervals, `trip_start_time` was transformed into a categorical indicator representing the corresponding 4-hour interval of the day.

The study derived the target variable by calculating the total usage count for each station during every 4-hour interval from January 1, 2017, 00:00 to September 30, 2024, 24:00.

Subsequently, the variable `trip_start_time` was renamed to `time`, and `from_station_name` was renamed to `station_name`.

Finally, the file was saved in Parquet format for efficient storage and processing.

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.

OpenAI. 2024. "ChatGPT: OpenAI's Language Model." https://openai.com/chatgpt.

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. 2024. "Bike Share Toronto Ridership Data." https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/.

———. n.d. "About Us." https://parking.greenp.com/about/about-us/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 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.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

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.