Delays in Toronto's Bus Network: Identifying Critical Factors and Weekly Patterns*

Incident Types and Weekday Trends as Key Contributors to Transit Inefficiencies

Chendong Fei

December 4, 2024

This paper examines delays in Toronto's bus network to identify the primary factors contributing to service disruptions and how they vary across the week. Using a dataset of over 42,000 recorded incidents, we find that delays are most commonly caused by specific types of incidents, such as security issues and general operational delays, with significant variations in their impact. Furthermore, average delays tend to peak on certain days of the week, revealing patterns in operational inefficiencies. These findings provide actionable insights for transit authorities, highlighting areas where targeted interventions can improve service reliability and commuter experience.

1 Introduction

Public transportation systems play a vital role in urban mobility, enabling access to jobs, education, and services while alleviating traffic congestion and reducing environmental emissions. Toronto's bus network, a core component of the city's public transit system, serves thousands of commuters daily. However, delays in bus services disrupt schedules, create stress for passengers, and pose challenges for transit operations, leading to decreased confidence in the system. Understanding the causes and patterns of these delays is essential to improving service reliability and ensuring smoother operations.

This study focuses on analyzing the causes and characteristics of delays in Toronto's bus network. It examines factors such as the type of incident, time of occurrence, and location to determine their relationship to delays. While many studies on transit performance describe these factors, few apply predictive models to understand their combined impact. This paper employs logistic regression to estimate the likelihood of delays under various circumstances,

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

enabling a more structured analysis of delay contributors. The findings identify specific patterns in the factors influencing delays, such as the role of certain incident types (e.g., security concerns and mechanical problems) and differences across days of the week. These patterns highlight opportunities to improve transit reliability through targeted adjustments to operations and scheduling. The analysis offers practical recommendations for transit authorities to better address common delay scenarios and reduce the frequency of disruptions.

The study's significance lies in its potential to guide improvements in public transit systems, enhancing the experience of commuters and reducing operational inefficiencies. By focusing on the specific factors that contribute to delays, the study aims to provide a practical foundation for improving Toronto's bus services. This study aims to estimate the effect of key operational and temporal factors, including incident type (e.g., security issues, mechanical failures), time of day (Time interval), day of the week (Monday through Sunday), route characteristics (specific bus routes and their traffic patterns), and travel direction (North, South, East, West), on the occurrence of delays in Toronto's bus network. Delays are measured as a binary outcome, indicating whether a bus was delayed during a given incident.

This paper is organized as follows: Section 2 describes the dataset and its preparation. Section 3 explains the methodology, including the development and validation of the logistic regression model. Section 4 presents the results, focusing on trends and patterns in the data. Section 5 discusses the implications of these findings for managing transit delays and proposes steps for improvement.

2 Data

2.1 Overview

The data analyzed in this study comes fromth e Toronto Transit Commission (TTC) bus delay records, made publicly available through the City of Toronto Open Data Portal(Commission, n.d.). The analysis is conducted using the statistical programming language R (R Core Team 2023), leveraging several key packages for data manipulation and visualization. These include `readr`(Wickham, Hester, and Bryan 2024) for reading of CSV or other text formats, `tidyverse` (Wickham et al. 2019) for data wrangling, `ggplot2` (Wickham 2016) for data visualization, `dplyr`(Wickham et al. 2023) for data fame operations such as filtering and sorting,`knitr`(Xie 2014) for generating dynamic reports, `arrow`(Richardson et al. 2024) for reading and writing data in efficient formats like Parquet, and `caret` (Kuhn and Max 2008) for modeling.

The dataset contains over 42,000 observations of bus delay incidents, capturing variables such as the date, time, location, route, delay duration, and type of incident. This granular data enables an in-depth examination of the factors contributing to delays, from operational issues to external disruptions. Additionally, it provides temporal and spatial information, allowing for the exploration of patterns across days of the week and bus routes.

While similar datasets for other transit modes, such as subways or streetcars, are available, this dataset was selected due to its focus on bus operations, which are more susceptible to traffic and environmental disruptions. Its rich detail and wide scope make it particularly well-suited for building predictive models and identifying actionable trends to improve service reliability.

2.2 Measurement

The dataset captures real-world bus delays in Toronto's transit system, transforming observed disruptions into structured data entries that facilitate analysis. These delays arise from diverse causes, including mechanical failures, security incidents, road blockages, weather conditions, and operational inefficiencies. Each delay event is observed and recorded by the Toronto Transit Commission (TTC) through a combination of manual reporting by operators and automated logging systems. The data collection process involves systematically documenting key attributes of each delay, including the date and time of the incident, the affected route, the location (e.g., intersections, stations, or areas), the type of incident (e.g., "Mechanical Issue," "Security," "General Delay"), and the quantitative impact of the delay, represented by Min Delay (duration of delay in minutes) and Min Gap (the service gap caused by the disruption in minutes).

Once recorded, this raw data undergoes a validation and standardization process by the TTC to ensure reliability and consistency. This includes verifying the accuracy of reported delays, standardizing incident categories, and removing duplicate or erroneous entries. Each real-world disruption is ultimately represented as a structured row in the dataset, with each column capturing specific details about the event. For example, a mechanical failure causing a 20-minute delay on Route 89 at Keele and Glenlake is recorded with entries for Date ("2024-01-01"), Time ("02:08"), Route ("89"), Location ("Keele and Glenlake"), Incident ("Mechanical Issue"), Min Delay ("20"), and Min Gap ("30").

In preparation for analysis, additional variables were constructed to enhance the dataset's analytical utility. For instance, the day of the week was derived from the "Date" field to investigate patterns across weekdays, while the time of day was categorized into intervals (e.g., morning, afternoon, evening) for temporal trend analysis. This structured measurement process ensures that real-world phenomena are faithfully represented in the dataset, while the derived variables provide a framework for deeper exploration of patterns and relationships in bus delays. By meticulously capturing both the quantitative and categorical aspects of each disruption, the dataset offers a detailed foundation for studying the factors contributing to transit inefficiencies and their broader impacts on Toronto's bus network.

2.3 Data Cleaning

The raw dataset was cleaned and refined to focus on variables relevant to the analysis of bus delays in Toronto. Key variables retained include Date, Time, Route, Location, Incident Type, Direction, Vehicle, and a newly created binary variable, Delay, which indicates whether a delay occurred. The Delay variable was derived from the "Min Delay" column, where values greater than 0 were coded as 1 (delay), and values equal to 0 were coded as 0 (no delay). This transformation enabled the binary classification required for logistic regression analysis.

To ensure the dataset's quality, we addressed missing values by removing rows where critical variables such as Route or Direction were missing, as these are essential for understanding spatial and operational patterns. Additionally, redundant variables, such as "Unnamed" columns or identifiers unrelated to the analysis, were removed. The Date and Time variables were standardized, with "Date" converted to a proper date format and "Time" retained for temporal pattern analysis. From these, new variables such as Day of the Week were derived to explore weekday versus weekend trends.

This process ensured the dataset's relevance and quality for examining the factors influencing bus delays in Toronto. Detailed steps of the data cleaning process are outlined in Appendix A.

Table 1: A Sample of the cleaned TTC Bus Delay Data

Date	Route	Time	Day	Incident	Direction	Vehicle	delay
2024/1/14	40	20:58:00	Sunday	Mechanical	W	8078	1
2024/7/21	91	09:11:00	Sunday	Utilized Off Route	N	8854	1
2024/2/28	102	15:37:00	Wednesday	Mechanical	S	3401	1
2024/1/11	21	15:31:00	Thursday	Vision	S	8412	1
2024/8/30	902	16:38:00	Friday	Operations - Operator	N	9054	1

2.4 Outcome variables

The outcome variable for this analysis is `delay`, a binary variable that indicates whether a bus delay occurred during a given incident. This variable is derived from the Min Delay column in the raw dataset, which records the duration of delays in minutes. To simplify the analysis and align with the objectives of binary classification, the delay variable is constructed as follows:

- delay = 1 if Min Delay is greater than 0, indicating that a delay occurred.
- delay = 0 if Min Delay equals 0, indicating no delay occurred.

The binary nature of the `delay` variable makes it suitable for logistic regression modeling, where the objective is to predict the likelihood of a delay based on various predictors, such as incident type, time of day, day of the week, and route characteristics. This transformation not only simplifies the data but also focuses the analysis on identifying the factors associated with the occurrence of delays rather than their specific durations. By modeling delay, this analysis aims to uncover patterns and relationships that can help improve the reliability of Toronto's bus network.

2.5 Predictor variables

The predictor variables in this analysis were chosen based on their potential impact on bus delays in Toronto's transit network. Each predictor reflects operational characteristics, temporal patterns, or geographic factors that contribute to the likelihood of a delay occurring. These variables aim to capture the underlying causes of delays and provide actionable insights for improving transit reliability. Key predictor variables include:

- **Incident**: This categorical variable identifies the type of incident causing the delay, such as "Security," "Mechanical Issue," or "General Delay." The coefficient for each incident type reflects its relative contribution to the likelihood of a delay.
- **Date**: This categorical variable indicates the day on which the incident occurred. The model captures variations in delay likelihood between weekdays and weekends, reflecting differences in traffic and operational conditions.
- **Time** : This variable categorizes the time of an incident into intervals (morning, afternoon, evening, or late night).
- **Route**: A categorical variable representing the affected bus route (e.g., Route 39 or Route 113).
- **Direction**: Indicates the direction of the bus (North, South, East, or West).
- **Vehicle**: A numerical variable representing the unique identifier of the bus involved in the incident.

2.6 Data Visulization

2.6.1 Delay Distribution by Incident Type

Figure 1 illustrates the relationship between different incident types and the likelihood of delays occurring. The x-axis represents categories such as "General Delay," "Collision - TTC," and "Security," while the y-axis shows whether a delay occurred (0 for no delay, 1 for delay). The width of each violin reflects the distribution of delay outcomes for each incident type, with wider sections indicating a higher frequency of those outcomes.

Incidents like "General Delay" show a higher density near 1, indicating that they are more frequently associated with delays. In contrast, categories such as "Vision" and "Utilized Off Route" are concentrated around 0, meaning they are less likely to cause delays. For some incidents, such as "Collision - TTC," the distribution is more evenly spread between 0 and 1, reflecting a mix of delay outcomes. The boxplots within each violin highlight the median and range of outcomes for each category, with most medians near 0, suggesting that delays are relatively uncommon for most incident types.

This graph helps identify which incidents are more strongly linked to delays, providing a clearer understanding of how different categories contribute to overall system performance.

Figure 1: The distribution of probabilities across categories. The width of each violin represents the density of observations, illustrating variations in likelihood associated with different categories.

2.6.2 Hourly Delay Patterns

The trend of delays across different hours of the day, represented as the average proportion of delays per hour, is shown in Figure 2. The X-axis spans the 24-hour period (0 to 23), capturing the hourly breakdown of delay trends, while the Y-axis represents the average delay proportion. The pattern indicates that delays are relatively high around midnight, drop significantly in the early morning hours (around 4-5 AM), and then rise sharply as the day progresses, peaking in the early to mid-morning (around 6-9 AM). Throughout the rest of the day, delay proportions

remain relatively stable, with slight fluctuations. This trend suggests that delays are influenced by factors such as traffic density, operational schedules, or service demands, which vary across different times of the day.

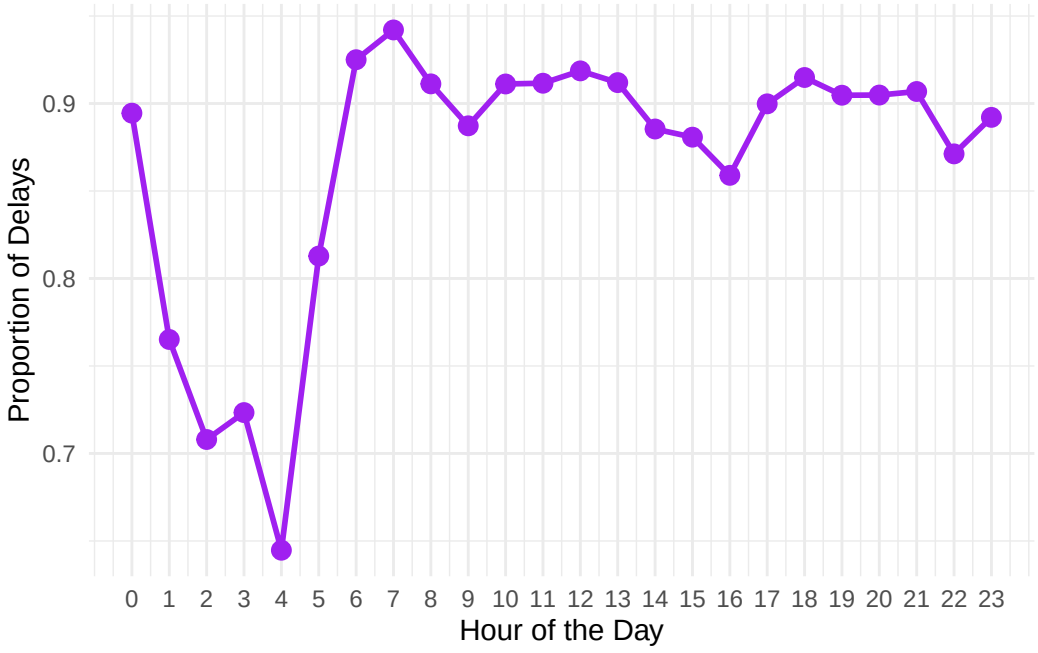

Figure 2: The proportion trends across different time intervals. The curve highlights variations over the period, emphasizing peaks and troughs in the observed proportions.

2.6.3 Route and Vehicle

Figure 3 illustrates the relationship between routes and vehicle counts, incorporating delay information for better analysis. The vertical spread of points within each route indicates variability in vehicle deployment, with routes showing greater spread handling more vehicles. Delays are visually represented through color differentiation, where blue signifies delayed vehicles and red indicates no delays.

2.6.4 The distribution of delay with weekdays

Figure 4 shows variations in average delays across the week. Weekdays often have higher delays, likely due to increased demand, while weekends may reflect event-driven or reduced schedules. Outliers on certain days suggest operational challenges or disruptions, indicating opportunities for improved scheduling and resource distribution to reduce delays.

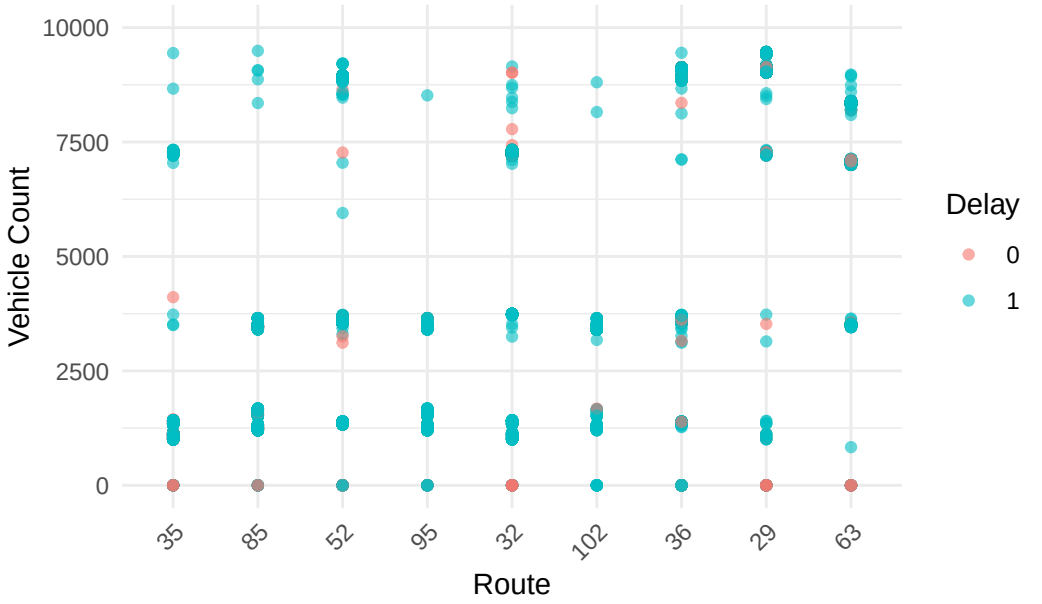

Figure 3: The relationship between categories and counts, with color-coded markers indicating delays (blue) and no delays (red).

3 Model

The goal of our modeling is to identify the factors influencing whether a delay occurs in transit operations. This includes examining predictors such as the route number, time of day, type of incident, and direction of travel. We seek to quantify the relationships between these predictors and the likelihood of a delay occurring, focusing on key contributors like the nature of incidents and temporal patterns.

The model set up by using logistic model(Cox 1958) and a Gradient Boosting Machine (GBM) model(others 2020), with the results summarized using the `modelsummary` package(Arel-Bundock 2022) for model output, the `broom` package(Robinson et al. 2023) for tidying model results, and the `pROC` package(Robin et al. 2023) for evaluating classification performance through metrics such as the ROC curve and AUC score.

Here, we describe the logistic regression model used to investigate these relationships. Logistic regression was chosen due to its ability to estimate the probability of binary outcomes (e.g., delay or no delay) while providing interpretable coefficients for each predictor. Background details, including diagnostic checks, are provided in Appendix B.

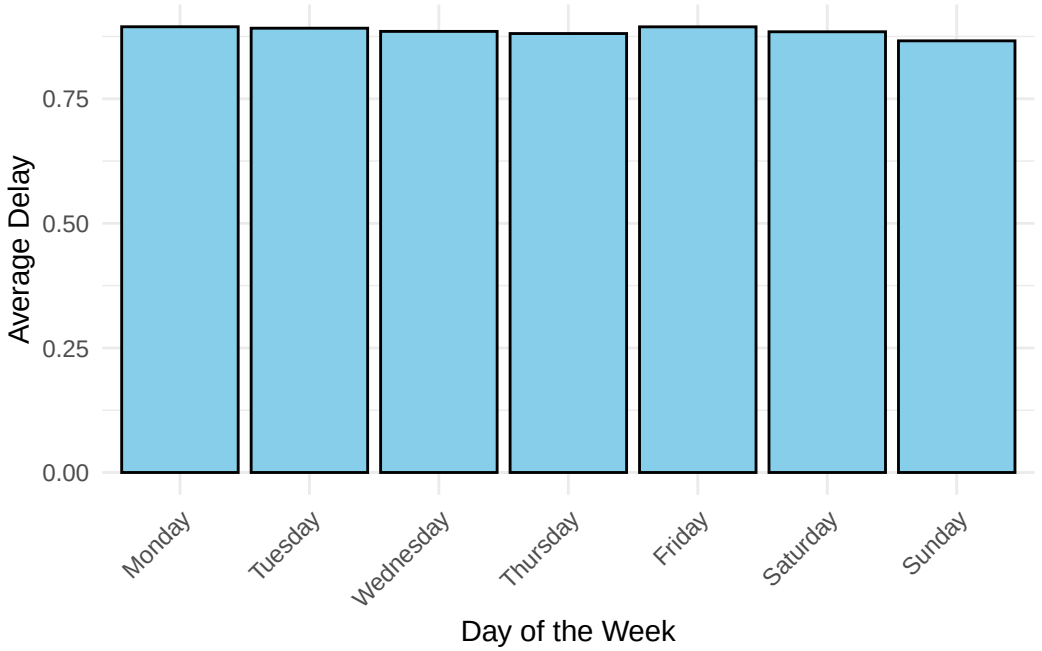

Figure 4: illustrating the average delay observed across different days of the week. Each bar reflects the proportion of delays, highlighting patterns in how delays vary throughout the week.

3.1 Model set-up

The logistic regression model estimates the probability of the binary response variable y (Delay or No Delay) as:

$$P(y = 1|X) = \frac{1}{1 + e^{-z}}$$

$$z = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where:

- $P(y = 1|X)$: Probability of delay given the predictors.
- z: Linear combination of the features (x_j) and coefficients β_j.
- x_j: Predictors derived from the dataset, such as Route, Time, Day, Incident, and Direction.
- β_0: Intercept of the model, representing the baseline log-odds of delay when all predictors are 0.
- β_j: Coefficient of the j-th predictor, representing its impact on the log-odds of delay.

3.1.1 Model justification

The logistic regression model was selected for this analysis because it is well-suited for binary classification tasks, particularly for predicting whether a delay occurs($y = 1$) or not ($y = 0$). This model aligns with the study's objectives as it provides a transparent and interpretable framework to understand the relationships between predictors and the likelihood of a delay. Logistic regression models the log-odds of the response variable as a linear combination of predictors, enabling the estimation of probabilities for delays while offering clear insights into how each variable impacts the likelihood of an event. Predictors in the model include Route, which captures route-specific trends such as traffic or operational challenges; Time, extracted as the hour of the day, which accounts for temporal variations like peak commute times; and Day, a categorical variable representing the day of the week, which captures weekly patterns in transit delays. Additionally, Incident categorizes disruptions (e.g., "Security" or "General Delay") to quantify their impact, while Direction accounts for geographic or directional effects on delays. The predictor variables used in the model include:

- **Route**: This numeric variable captures the route number. By including this feature, the model can identify route-specific trends and their relationship with delays, such as whether specific routes are more prone to delays due to traffic congestion or operational constraints.

- **Time**: Extracted as the hour of the day, this continuous variable helps capture temporal patterns, such as delays being more frequent during peak hours or late-night operational challenges.
- **Day**: This categorical variable indicates the day of the week. By incorporating this predictor, the model accounts for weekly variations, such as differences in transit performance between weekdays and weekends.
- **Incident**: This categorical variable describes the type of incident (e.g., "General Delay," "Security," or "Vision"). Incidents are critical predictors, as they often directly trigger delays. The model quantifies how much each incident type increases or decreases the likelihood of a delay.
- **Direction**: This categorical variable indicates the cardinal direction (e.g., North, South). Including direction accounts for route-specific traffic dynamics or geographic factors that might influence delays.

The choice of logistic regression was driven by its simplicity, efficiency, and ability to provide probabilistic outputs, which allow for nuanced decision-making beyond binary classifications.Logistic regression offers interpretable coefficients that quantify the influence of each predictor, making it easier to derive actionable insights. Modeling decisions were tailored to the dataset's structure: categorical variables were label-encoded to retain their information, missing values in Direction were imputed using the most frequent category, and class weighting was applied to address potential imbalance between delayed and non-delayed cases. Although logistic regression assumes linearity in the log-odds space and may not capture complex interactions between predictors, it remains an appropriate choice for this study. The model effectively balances interpretability and predictive power, providing a robust framework for analyzing the factors that contribute to transit delays.

3.1.2 Model validation

In our analysis, the ROC curve is a critical tool for model validation as it visually represents the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate) across various decision thresholds. In model validation, the ROC curve helps assess how well the model discriminates between classes—in this case, delayed versus non-delayed observations.

The Area Under the Curve (AUC) derived from the ROC provides a single metric summarizing the model's performance. An AUC closer to 1 indicates excellent discriminatory ability, while an AUC of 0.5 suggests no better performance than random guessing. For our dataset, the AUC enables us to compare models objectively and determine how well the logistic regression captures the patterns in the data.

By evaluating the ROC curve alongside other metrics like accuracy, precision, and F1-score, we ensure a comprehensive validation of the model's ability to generalize to unseen data, thus confirming its suitability for real-world applications.

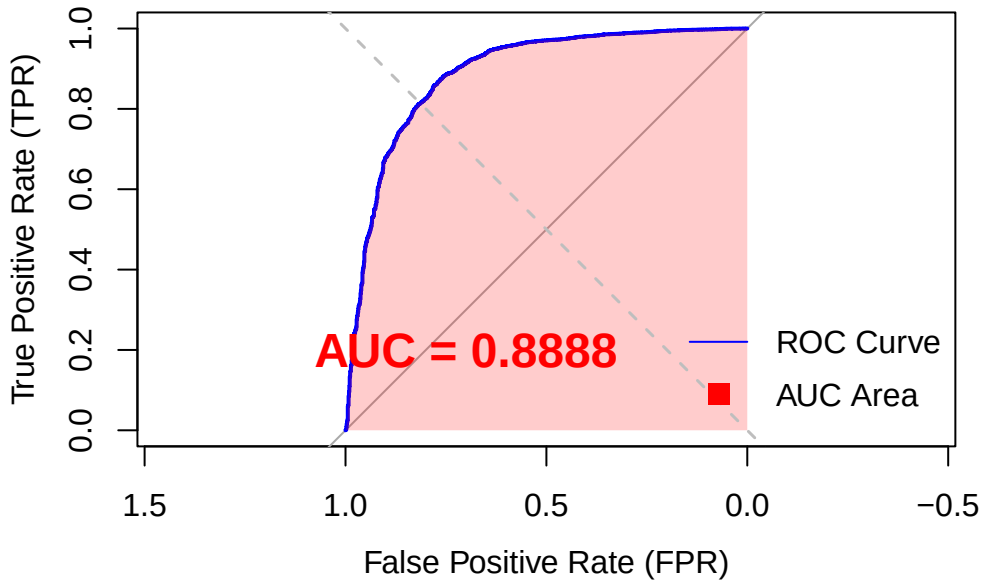

Figure 5: ROC curve for the logistic regression model with an AUC of 0.8888, showing strong predictive performance.The blue line represents sensitivity-specificity trade-offs, with the shaded red area highlighting the AUC. The dashed diagonal line indicates random classification.

Figure 5 shows the ROC curve, paired with the highlighted AUC area, serves as a visual tool for evaluating the logistic regression model's predictive performance in the context of the dataset. The ROC curve illustrates the model's trade-off between sensitivity (True Positive Rate, TPR) and specificity (False Positive Rate, FPR) across various classification thresholds. The curve's proximity to the upper left corner of the plot indicates that the model achieves a high sensitivity and specificity, effectively differentiating between delayed and non-delayed trips. The shaded AUC area, prominently displayed as 0.8888, quantitatively summarizes the model's overall discriminatory ability. An AUC value close to 1 signifies excellent model performance, confirming that the logistic regression model is well-suited for capturing the key relationships within the dataset. This value indicates that the predictors, such as `Route`, `Time`, `Day`, `Incident`, and `Direction`, contribute meaningfully to the model's ability to classify trips accurately.This validates the model's capacity to generalize and highlights the dataset's features as key contributors to accurate delay predictions.

Table 2 shows the model achieved an accuracy of 92.06%, indicating that it correctly predicts whether a delay occurs in the vast majority of cases. The precision of 94.22% shows that when the model predicts a delay, it is accurate in its prediction most of the time, minimizing false alarms. Additionally, the recall of 97.08% highlights the model's ability to identify the majority of actual delays, ensuring that very few delays go undetected. The F1-Score, a balanced metric combining precision and recall, is 95.63%, reflecting the model's overall effectiveness in handling both delayed and non-delayed cases. Furthermore, the ROC-AUC score of 88.88% demonstrates the model's ability to distinguish between delays and non-delays across various probability thresholds, showcasing strong discriminatory power.

These metrics collectively indicate that the predictors in the dataset, such as `Incident` and `Direction`, provide valuable insights into the occurrence of delays. The high precision and recall suggest that the model performs well despite any potential class imbalance in the data. The robust ROC-AUC score further confirms that the logistic regression model effectively captures the relationship between the predictors and the target variable. Overall, the model demonstrates strong predictive capability and generalizability, making it a reliable tool for understanding and predicting transit delays.

Table 2: Performance metrics for the logistic regression model, including accuracy, precision, recall, F1-Score, and ROC-AUC, to evaluate the model's classification ability on the test dataset.

Metric	Value
Accuracy	0.9206959
Precision	0.9422613
Recall	0.9708211
F1-Score	0.9563280
ROC-AUC	0.8888290

3.1.3 Model diagnostics

Residual plots are essential for evaluating the performance of the logistic regression model developed to analyze the relationship between transit delays and predictors such as Route, Time, Day, Incident, and Direction. These plots provide insights into how well the model fits the data, validates assumptions, and highlights potential areas for improvement. In the context of this dataset, residual plots were used to identify patterns, outliers, and potential deficiencies in the model.The residual plots allow us to visually assess whether the model adequately captures the relationships between the predictors and the target variable delay. For example, the residuals vs. Route plot shows a random scatter around zero, indicating that the model accounts for the variations in delays across different routes effectively. Similarly, the residuals vs. Day plot demonstrates that the residuals are centered around zero for all days of the week, suggesting no systemic bias related to specific days. A detailed analysis of the provided residual plots is shown in Appendix B.

3.1.4 Alternative Models Considered

Gradient Boosting, were considered but deemed less appropriate for this analysis. This model, while powerful, lack the interpretability required to identify and quantify the effects of individual predictors on delay likelihood. The Gradient Boosting model predicts the probability of a delay (y) based on the features in your dataset (x)including Route, Time, Day, Incident, and Direction):

$$\hat{F}(x) = \sum_{m=1}^{M} \nu \cdot h_m(x),$$

where:

- $\hat{F}(x)$ represents the predicted log-odds of a delay (or probability after applying a sigmoid function).
- $h_m(x)$ represents a weak learner (e.g., decision tree) added in the m-th iteration.
- ν represents Learning rate, a small constant to control the contribution of each weak learner.
- M: Total number of iterations (trees).

In contrast, logistic regression provides clear, actionable insights into how features such as Time, Route, and Incident affect the likelihood of delays, which is critical for making informed decisions and communicating results to stakeholders. Additionally, logistic regression is computationally efficient and easier to validate, making it the more practical choice for this dataset and analysis. The detailed comparison between two models shown in Appendix B.

4 Results

Our results are summarized in Table 3. The results of the logistic regression analysis provide a detailed assessment of the factors influencing the likelihood of delays, with incident type and travel direction emerging as the most influential predictors. Table 3 summarize the relationship between predictors and the probability of delays. "General Delay" has the most substantial negative coefficient, indicating its significant role in influencing delay outcomes. Other factors, such as "Emergency Services" and "Collision – TTC," also show strong negative associations, reflecting their distinct effects on delay probabilities.

Directional predictors, such as "Direction N," show slight variations, with northbound routes demonstrating a marginally negative relationship with delays. Meanwhile, weekday predictors, including "Monday" and "Saturday," exhibit modest negative coefficients, suggesting a lower likelihood of delays on these days compared to the baseline.

Confidence intervals vary across predictors, with narrower intervals for factors like "DirectionN" indicating higher certainty in these estimates, while broader intervals for predictors like "Utilized Off Route" suggest greater variability in their effects.

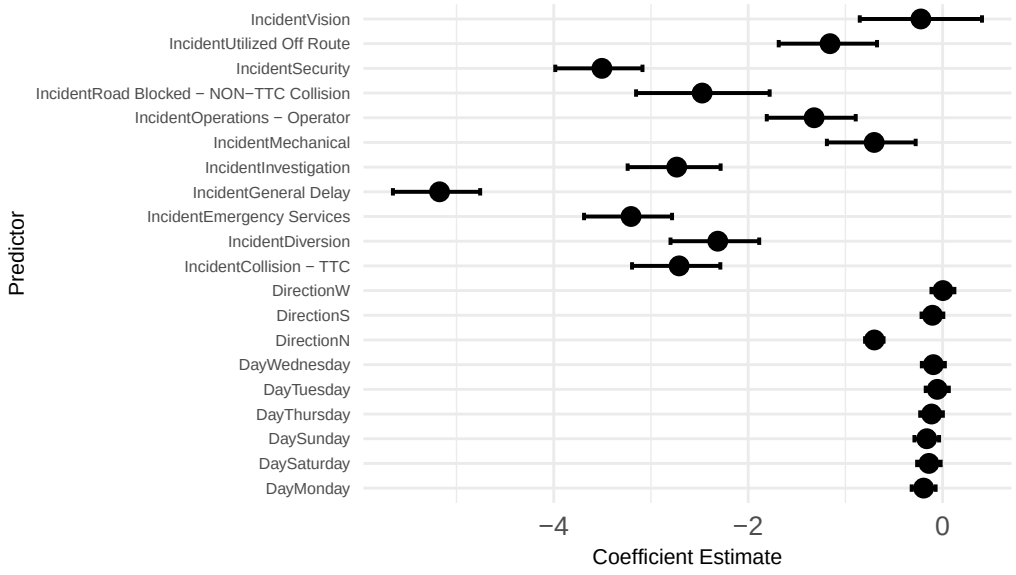

Figure 6: Coefficient Estimates with 95% Confidence Intervals from Logistic Regression Model.

Figure 6 illustrates the distribution of delays for various incident types. The highest concentration of delays occurs in "General Delay," indicating this category's significant impact on overall delay frequencies. Incidents like "Collision – TTC" and "Emergency Services" also

Table 3: Summary of Coefficient Estimates from Logistic Regression Model.

	(1)
(Intercept)	5.029 (0.232)
DayMonday	−0.195 (0.063)
DaySaturday	−0.141 (0.062)
DaySunday	−0.163 (0.065)
DayThursday	−0.114 (0.059)
DayTuesday	−0.055 (0.061)
DayWednesday	−0.096 (0.060)
IncidentCollision - TTC	−2.710 (0.230)
IncidentDiversion	−2.313 (0.231)
IncidentEmergency Services	−3.205 (0.230)
IncidentGeneral Delay	−5.175 (0.228)
IncidentInvestigation	−2.735 (0.243)
IncidentMechanical	−0.704 (0.232)
IncidentOperations - Operator	−1.322 (0.232)
IncidentRoad Blocked - NON-TTC Collision	−2.474 (0.348)
IncidentSecurity	−3.505 (0.228)
IncidentUtilized Off Route	−1.158 (0.257)
IncidentVision	−0.224 (0.318)
DirectionN	−0.702 (0.049)
DirectionS	−0.105 (0.057)
DirectionW	0.003 (0.060)

show noticeable delay patterns, though with less variation, implying more consistent durations for these events.

Conversely, categories such as “Mechanical” and “Operations – Operator” are associated with smaller and more localized delay distributions. Other incident types, such as “Vision” and “Utilized Off Route,” have minimal contributions to the delay distribution, reflecting their limited role in the overall delay occurrences. These variations highlight the different degrees to which incident types influence delay durations and frequencies, suggesting targeted measures could address delays in high-impact categories.

The findings emphasize the need to focus on specific incident types, such as “General Delay” and “Emergency Services,” to reduce delays. Operational adjustments and resource allocation tailored to high-impact categories could significantly enhance system performance. The directional and weekday trends provide further opportunities for optimizing scheduling and resources to mitigate delays during predictable time frames. These results point to actionable pathways for improving the reliability and efficiency of transportation networks.

5 Discussion

5.1 Delays Across Transportation Networks: Patterns and Key Observations

This paper investigated the distribution and underlying causes of delays in urban transportation systems, identifying important patterns across routes, vehicles, and operational dynamics. Notably, delays were highly concentrated along specific high-traffic routes, suggesting vulnerabilities in how these routes are managed. These findings emphasize the need to assess route-specific factors such as infrastructure quality, traffic load, and scheduling effectiveness. The analysis further highlighted how certain routes, such as those with higher commuter density, disproportionately contribute to delay frequencies, pointing to an imbalance in resource allocation across the network.

Moreover, the study found that vehicle-related factors, including type and frequency of usage, play a significant role in delay occurrences. Routes with older or high-maintenance vehicles exhibited significantly higher delay proportions, indicating a pressing need for fleet upgrades. These results illustrate that delays in urban transportation systems cannot be viewed in isolation but rather as an intersection of multiple, interconnected factors, each influencing the system's overall efficiency.

5.2 Operational Incidents and Their Role in Delay Dynamics

One of the most significant findings of this study is the substantial impact of operational incidents on delay occurrences. Mechanical failures, security problems, and other disruptions were shown to increase delays, especially on already vulnerable routes. For example, mechanical issues resulted in extended delays during peak hours when demand for public transit is highest. This escalation highlights weaknesses in routine maintenance schedules and the system's capacity to manage unexpected events.

Additionally, the analysis revealed that routes frequently impacted by specific types of incidents often experienced cascading effects, where delays on one route triggered disruptions across the broader network. This underscores the fragility of the current system and its insufficient redundancy. A clear example is delays caused by vehicle breakdowns, which disproportionately affected central routes connecting major commuter hubs, creating ripple effects on peripheral routes. Addressing these bottlenecks calls for targeted strategies, such as implementing real-time monitoring systems to identify and respond to incidents before they intensify.

5.3 Temporal and Spatial Trends in Delays

The temporal analysis identified a strong correlation between delays and specific time intervals, with delays peaking during rush hours. This reflects the increased strain on the system during high-demand periods, where even minor disruptions have magnified effects. Morning delays

were often longer, potentially due to unresolved overnight maintenance issues, while evening delays were linked to higher passenger volumes and extended service times.

Spatially, delays were more significant on routes serving dense urban centers and high-demand residential areas. This pattern emphasizes the importance of revisiting route planning and prioritizing consistently delayed routes. Additionally, certain regions showed a higher likelihood of specific types of delays, such as security-related incidents in commercial districts and mechanical failures on suburban routes. Recognizing these spatial patterns helps city planners allocate resources more strategically, ensuring that the most vulnerable areas of the network are adequately supported.

5.4 Limitations and Areas for Improvement

Despite the thorough nature of the analysis, several limitations must be acknowledged. One significant limitation was the absence of detailed data on external factors such as weather conditions, road quality, or construction activities, all of which are known to strongly influence delays. Including these variables could provide a more detailed understanding of the causes behind delays.

Another limitation is related to the modeling approach. While the logistic regression model used in this analysis offered meaningful observations, it may not fully account for complex interactions between variables, such as how the type of incident interacts with route features to impact delays. Additionally, the analysis assumes that the dataset is complete and accurate, which might not be true given the possibility of underreporting or misclassification of certain incidents.

The temporal scope of the data also poses a limitation. The study analyzed a relatively short period, restricting the ability to identify long-term trends or seasonal variations. For example, the effects of special events, policy adjustments, or economic changes could not be evaluated due to these temporal constraints.

5.5 Recommendations for Future Research

Future research should aim to address these limitations by incorporating additional datasets, such as traffic density, weather patterns, and socio-economic characteristics of neighborhoods served by each route. This would allow for a more detailed understanding of the factors contributing to delays and provide a broader perspective on external influences.

Additionally, applying alternative modeling techniques, such as ensemble methods or machine learning algorithms, could enhance predictive accuracy and capture non-linear relationships between variables. For example, decision trees or random forests could identify complex interactions among incidents, vehicle attributes, and routes that may not be adequately represented by linear models.

Conducting longitudinal studies that track changes in delay patterns over multiple years would also be advantageous. Such studies could evaluate the effects of policy changes, infrastructure upgrades, or shifts in commuter behavior over time, offering actionable recommendations for long-term planning and decision-making.

5.6 Practical Implications for Urban Planning and Transportation Management

The findings of this study have important implications for urban planning and transportation management. First, addressing delays requires targeted strategies on high-delay routes, such as increasing the frequency of maintenance checks, deploying better-equipped vehicles, and adopting dynamic scheduling to manage demand across the network. Additionally, installing real-time monitoring systems can provide actionable information to address incidents as they occur, reducing their impact on the system.

Second, planners should prioritize enhancing the resilience of the network by incorporating redundancies, such as alternative routes or additional vehicles, to mitigate the cascading effects of delays. Investments in infrastructure, particularly in high-density urban areas, are also essential for addressing delays caused by traffic congestion and high passenger volumes.

Finally, public education campaigns promoting off-peak commuting could help ease the strain on the system during rush hours. Policies that encourage flexible working hours or telecommuting could further balance demand across the network, leading to more sustainable usage patterns.

A Appendix

A.1 Data Cleaning Notes

The data cleaning process began with loading the raw TTC Bus Delay dataset using the read_csv function in R, ensuring that the data structure was intact and ready for transformation. Unnamed columns (e.g., "…1"), which likely resulted from exporting or indexing during data collection, were removed using the `select(-starts_with("..."))` command. These columns were redundant and added no analytical value, so their removal helped streamline the dataset. Following this, the Min Delay and Min Gap columns, which initially had character string data types, were converted to numeric using mutate. This step ensured that these key variables were properly formatted for numerical analysis, enabling operations such as comparisons and aggregations.

To prepare the dataset for a binary classification approach, a new variable, `delay`, was created. This binary variable indicates whether a delay occurred: it is set to 1 if the value of `Min Delay` is greater than 0 and to 0 otherwise. This transformation simplified the data, focusing on the occurrence of delays rather than their exact durations, which aligns with the requirements for logistic regression modeling. After constructing the delay variable, the original `Min Delay` and `Min Gap` columns were removed using the select function. These columns, while initially useful for creating the delay variable, were no longer necessary in the final dataset as their information had been captured.

Further cleaning focused on the Direction column, which was critical for understanding the geographic aspects of the delays. Entries in this column were evaluated, and only valid directional codes ("N", "S", "E", "W") were retained. Invalid or unrecognized entries, which may have resulted from data entry errors or inconsistencies, were replaced with NA. This ensured that the Direction variable was accurate and free from noise, allowing for more reliable analysis of directional patterns in delays.

The final dataset was reduced to only the most relevant columns: `Date`, `Route`, `Time`, `Day`, `Incident`, `Direction`, `Vehicle`, and the newly constructed delay variable. These variables were cleaned and structured to enable a clear and focused analysis. The Date column was retained in a standard date format, while Time was kept as a separate variable for potential temporal trend exploration. The Day column represented the day of the week, extracted from the date, to facilitate analysis of weekly delay patterns. The Route and Location variables provided spatial context, and the Incident column described the nature of each delay. The cleaned dataset is now consistent, structured, and ready for logistic regression modeling, ensuring a robust foundation for analyzing the factors influencing delays in Toronto's bus network.

B Model details

B.1 Diagnostics

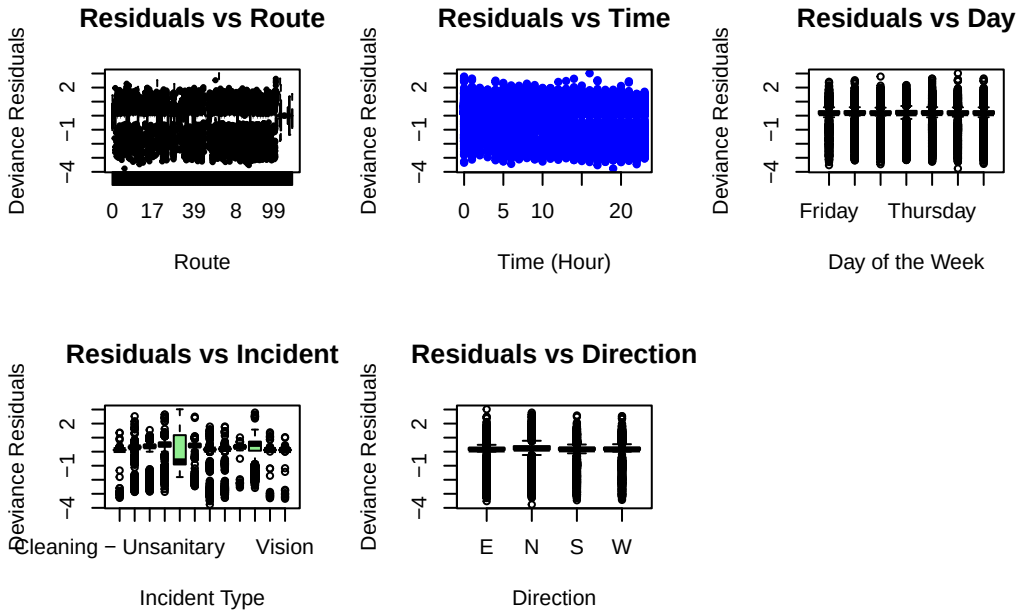

Figure 7: Deviance residual plots for the logistic regression model against predictors show a generally good model fit, with random scatter around zero and some outliers indicating areas for further refinement.

Figure 7 is a figure of residual plots, which providse a comprehensive diagnostic assessment of the logistic regression model's performance with respect to the predictors: `Route`, `Time`, `Day`, `Incident`, and `Direction`.

The residuals vs. `Route` plot shows a random scatter around zero, indicating that the model effectively captures variations in delays across routes without significant bias. Similarly, the residuals vs. `Time` plot demonstrates a uniform spread of residuals across the hours of the day, suggesting that the relationship between `Time` and delay is adequately modeled. The boxplots for `Day` reveal consistent distributions of residuals across days of the week, with median values close to zero, confirming the absence of systematic day-specific effects. However, the residuals vs. `Incident` plot shows higher variability for certain incident types, such as Cleaning - Unsanitary, implying that these incidents may not be fully captured by the model and might benefit from further exploration or additional predictors. Lastly, the residuals vs. `Direction` plot displays symmetric distributions across all directions, supporting the conclusion that directional effects are well-accounted for in the model. While the overall fit appears satisfactory, the

presence of outliers across the plots highlights specific observations or scenarios that warrant further investigation or refinement.

B.2 Models Comparsion

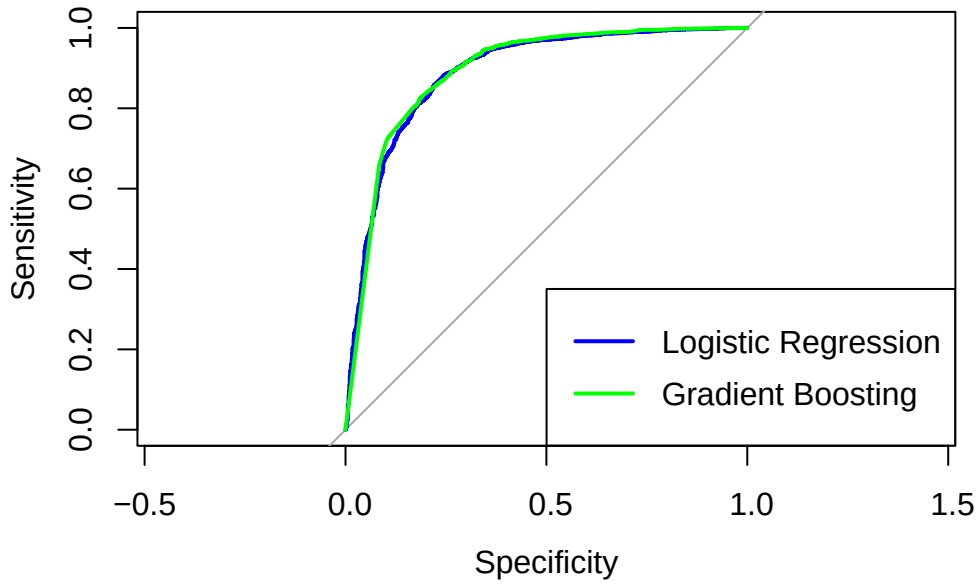

Figure 8: Comparison of ROC curves for Logistic Regression (blue) and Gradient Boosting (green) models, illustrating the trade-off between sensitivity and specificity. Both models demonstrate high predictive performance, with Gradient Boosting showing a slightly higher AUC.

Table 4: Comparison of AUC scores between Logistic Regression and Gradient Boosting models.

Model	AUC
Logistic Regression	0.8888290
Gradient Boosting	0.8913507

Figure 8 and Table 4 shows that both Logistic Regression and Gradient Boosting models effectively predict transit delays, with ROC-AUC scores of 0.8888 and 0.8913, respectively. The Gradient Boosting model marginally outperforms Logistic Regression, indicating its ability to capture complex, non-linear relationships in the dataset. However, Logistic Regression

provides clear and interpretable coefficients that directly link predictors such as `Route`, `Time`, and `Incident` to the probability of delays. This interpretability is important for actionable insights and operational decision-making, aligning with the primary goal of the analysis to understand and quantify the effects of key predictors on delay likelihood. While Gradient Boosting may be preferred for purely predictive tasks, Logistic Regression is favored in this study for its balance of accuracy and interpretability, making it more aligned with the study's objective of providing transparent and actionable results.

C Ideal Survey

The purpose of the survey is to gather insights to better understand factors contributing to delays in public transit, including incident types, vehicle behavior, and timing. The survey ensures inclusivity by engaging diverse stakeholders such as public transit operators, passengers, and industry experts.

C.1 Sampling Frame

The target population can be divided into the following categories:

- Transit Operators: Bus, train, or other vehicle operators responsible for service delivery. Include individuals familiar with route and traffic management.
- Passengers:Regular commuters across different routes and times. Represent diverse demographics (age, region, income levels).

The sample size:

- Total Respondents: ~10000
- Operators: ~1000
- Passengers: ~8000
- Experts: ~1000

C.2 Survey Distributing

The survey will be distributed using targeted approaches for each group of respondents. For operators and experts, email invitations will be sent through professional organizations or transit unions to ensure the survey reaches qualified individuals. Passengers will be engaged through public outreach on transit platforms, mobile apps, and in-station posters. To encourage participation, QR codes and online links will be provided for easy access. As an incentive, operators and experts will receive a small honorarium for their time, while passengers will have the chance to win transit credits or discounts. The survey will be conducted over a span of

four weeks, ensuring responses capture a range of operational conditions, including weekdays and weekends.

C.3 Sampling Methodology

The survey will use a stratified sampling methodology to ensure balanced representation across several variables. Stratification will account for route type (urban vs. suburban routes), travel timing (morning, afternoon, and evening), and demographic characteristics such as age, occupation, and dependence on transit systems. To refine the questions and improve clarity, a pilot test will be conducted with around 100 participants before rolling out the full survey.

C.4 Survey Develiery

Delivery of the survey will be managed through online platforms like Google Forms or SurveyMonkey, designed for accessibility on both mobile and desktop devices. To maintain ethical standards, all responses will remain anonymous, and participants will be provided with a consent statement explaining the purpose of the survey and the intended use of their responses.

C.5 Survey Questions

1. What is your primary role?

- Passenger
- Operator
- Expert

2. What is your age group??

- 18-25
- 26-40
- 41-60
- 60+

3. What is your usual route

- write down you route:___

4. How frequently do you use public transit?

- Daily
- Weekly
- Monthly

5. What time of day do you typically travel?

- Morning
- Afternoon
- Evening
- All day

6. If yes, what was the most common cause?

- General Delay
- Emergency Services
- Security
- Mechanical
- Cleaning

7. Rate the following factors contributing to delays: (Scale of 1-5)

- Vehicle breakdowns
- Road conditions
- Scheduling errors
- Incidents (e.g., security or collisions)

8. What improvements do you suggest to reduce delays? (Open-ended)

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Commission, Toronto Transit. n.d. "Open Data Dataset." *About TTC Bus Delay Data*. https://open.toronto.ca/dataset/ttc-bus-delay-data/.

Cox, David R. 1958. "The Regression Analysis of Binary Sequences." *Journal of the Royal Statistical Society: Series B (Methodological)* 20 (2): 215–32.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

others, Greg Ridgeway with contributions from. 2020. *gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm.

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.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédéric Lisacek, Jean-Charles Sanchez, and Markus Müller. 2023. *pROC: Display and Analyze ROC Curves*. https://CRAN.R-project.org/package=pROC.

Robinson, David, Alex Hayes, Simon Couch, Maxime Pelletier, Meagan Murthi, Leo Hermanson, Shannon Pileggi, et al. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

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

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.