Transit Delay Patterns in 2024 Toronto: A Comparative Analysis of Geographic, Modal, and Temporal Factors*

Uncovering Operational Challenges and Opportunities for Improving Toronto's Transit System

Jerry Xia

December 3, 2024

This study investigates delays in the Toronto Transit Commission (TTC), focusing on buses, subways, and streetcars, to understand factors affecting transit reliability. Using real-world data from Open Data Toronto and a random forest model, it identifies key predictors of delays, including transit mode, geographic location, and temporal patterns. The analysis reveals that buses experience the most frequent and longest delays, particularly in urban areas like Dundas, while suburban locations such as Yorkdale face fewer disruptions. These findings highlight the operational challenges unique to transit modes and locations, providing actionable insights to improve public transit efficiency. This research contributes to understanding how transit delays vary across contexts, offering strategies to enhance service reliability for commuters and transit planners.

Table of contents

*Code and data are available at: https://github.com/Jerryx2020/TTC_Delay_Analysis

1 Introduction

The performance of public transit systems is a cornerstone of urban mobility, shaping commuter experiences and the broader dynamics of city life. In Toronto, the Toronto Transit Commission (TTC) operates a multimodal network of buses, subways, and streetcars, each of which faces distinct operational challenges. Transit delays remain a persistent issue, disrupting schedules, inconveniencing passengers, and reducing the overall efficiency of the city's transportation infrastructure. Delays can also erode public trust in transit systems, which is critical for cities aiming to increase public transit use as part of sustainable urban mobility strategies (Litman 2023). Understanding the underlying factors contributing to these delays is essential for designing interventions that enhance system reliability and commuter satisfaction.

This study investigates TTC delays across buses, subways, and streetcars, emphasizing patterns across geographic locations and temporal contexts. Two specific locations—Dundas, in Toronto's urban core, and Yorkdale, a suburban area—serve as focal points for comparative analysis. Dundas represents a dense, high-demand transit environment often subject to congestion and frequent service interruptions, consistent with findings in other high-density urban transit systems (Cats and Gkioulou 2017). In contrast, Yorkdale provides a suburban context with less frequent service and lower traffic volumes, which are typically associated with fewer delays and more predictable transit operations (Hess and Lombardi 2007). By exploring these contrasting environments, this research aims to uncover delay dynamics shaped by urban density, transit demand, and local infrastructure.

Previous research has examined broad operational inefficiencies, such as traffic congestion and service disruptions, as primary contributors to delays (Litman 2023). However, the complex interactions between transit mode, geographic factors, and temporal patterns remain underexplored (Currie and Sarvi 2012). This study addresses that gap by leveraging detailed datasets from Open Data Toronto to model delay durations and identify key predictors, including transit mode, travel direction (bound), and day of the week. The findings align with global research highlighting the importance of mode-specific challenges, such as the susceptibility of buses to traffic congestion and the relative reliability of rail-based systems (Cats and Gkioulou 2017; Hess and Lombardi 2007).

The estimand of this study is delay duration (min_delay), defined as the recorded delay for individual transit trips. Using a Random Forest algorithm, this study models delay duration with predictors such as mode, location, temporal factors, and service regularity (min_gap). The analysis reveals significant variations in delays across transit modes and geographic locations, underscoring the challenges faced by buses in urban settings compared to the more consistent performance of subways. Temporal patterns, such as higher delays during weekdays, further highlight the role of commuter-driven demand in shaping transit reliability.

The structure of this paper is as follows: Section Section 2 details the datasets, cleaning procedures, and key variables used in the analysis. Section Section 3 outlines the exploratory data analysis and the modeling framework, explaining the rationale for selecting Random

Forests. Section Section 4 presents the primary findings, including visualizations and statistical summaries of delay patterns. Finally, Section Section 5 interprets the results, explores their implications for TTC operations, and identifies limitations and opportunities for future research.

2 Data

The analysis was conducted using R (R Core Team 2024), leveraging the following R packages to facilitate data cleaning, visualization, and analysis: `tidyverse` (Hadley Wickham and others 2024b), `ggplot2` (Hadley Wickham 2024a), `dplyr` (Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller 2024), `here` (Kirill Müller and Jennifer Bryan 2024), `arrow` (Apache Arrow Developers 2024), `caret` (Max Kuhn 2024), `randomForest` (Andy Liaw and Matthew Wiener 2024), `stringr` (Hadley Wickham 2024b), `knitr` (Xie, Yihui 2023), `testthat` (Hadley Wickham, Jim Hester, and others 2024) and `readr` (Hadley Wickham and others 2024a).

Additional guidance for this project was drawn from Telling Stories with Data by Rohan Alexander (Alexander 2023), which shaped the analytical approach. Details about data cleaning and variable transformations are provided in Section C.

2.1 Source of Data

The dataset for this research was obtained from Open Data Toronto, a publicly accessible platform providing detailed municipal datasets. This particular dataset focuses on transit delays across the Toronto Transit Commission (TTC) network, including buses, subways, and streetcars, for the year 2024. The dataset offers a granular view of transit disruptions, capturing key operational metrics and contextual factors. It serves as the primary resource for analyzing delay patterns and identifying the influences of transit mode, geographic location, and temporal factors on service reliability.

The data encompasses several critical variables. 'mode' categorizes the transit type—bus, subway, or streetcar—each with distinct operational contexts. The 'location' variable specifies where delays occurred, with a focus on Dundas, representing Toronto's urban core, and Yorkdale, a suburban area. These locations allow for a comparative analysis of transit dynamics across contrasting environments. Temporal variables, such as 'day_of_week,' facilitate an understanding of how delays vary between weekdays and weekends, reflecting commuter versus recreational patterns. Directionality, captured in the 'bound' variable, enables an exploration of the geographic and directional influences on delays. Finally, the numerical variables 'min_delay,' measuring the shortest delay recorded, and 'min_gap,' reflecting the time gap between consecutive vehicles, are central to understanding service disruptions and regularity.

This dataset was selected for its level of detail and public accessibility, which align with the research goal of reproducibility and transparency. Alternative datasets, such as internal TTC

reports or proprietary transit data, were considered but were unavailable due to licensing restrictions and limited public access. Open Data Toronto, combined with R packages like `opendatatoronto` (Jordan Freitas and Luke Johnston 2024) and `tidyverse` (Hadley Wickham and others 2024b), enabled seamless integration and analysis.

2.2 Measurement

The TTC delay dataset translates real-world transit disruptions into structured data entries to facilitate an analysis of operational patterns and predictors of service reliability. Key variables include min_delay, min_gap, mode, bound, and location, each capturing distinct dimensions of transit performance. These variables collectively enable a comprehensive examination of delay dynamics across Toronto's public transit system.

The primary outcome variable, min_delay, represents the delay duration in minutes for each transit trip. This measure reflects the extent of service disruptions, ranging from brief delays to significant interruptions. It serves as a critical metric for assessing the impact of delays on the TTC system's overall reliability. The companion variable, min_gap, quantifies the time gap between consecutive vehicles on the same route, providing insights into service regularity and scheduling efficiency.

The mode variable categorizes the type of transit service—bus, subway, or streetcar—capturing operational differences shaped by infrastructure, traffic interactions, and service management. The bound variable specifies the direction of travel, such as North, South, East, or West, enabling an exploration of geographic and directional influences on delays. Finally, the location variable focuses on key geographic contexts, highlighting patterns in urban (Dundas) and suburban (Yorkdale) settings. This geographic distinction is central to understanding variations in delay dynamics across different transit environments.

Although the dataset does not provide explicit documentation of its collection methods, it is reasonable to infer that the TTC aggregates data through automated monitoring systems, supplemented by manual reporting. Automated data collection likely includes feeds from vehicle tracking systems, GPS technology, and onboard sensors, while manual logs may capture incidents that are not automatically detected. This dual approach ensures a detailed record of service performance, consistent with contemporary transit data practices (Litman 2023; Cats and Gkioulou 2017).

The structured nature of the dataset allows for nuanced analyses of delay patterns and their operational drivers. Each variable was selected to capture essential aspects of transit performance, transforming complex real-world phenomena into analyzable components. This structured data foundation enables a systematic exploration of factors influencing TTC delays and offers insights to inform transit planning and operational improvements.

A datasheet is available for viewing through the repository this paper is associated with.

2.3 Outcome Variables

The outcome variable analyzed in this study, min_delay, represents the delay experienced by transit vehicles in minutes. This variable is critical for understanding service disruptions in the Toronto Transit Commission (TTC) system. By modeling min_delay, the analysis aims to quantify delay patterns and identify key factors influencing transit reliability across different modes, geographic locations, and temporal contexts.

2.3.1 Delay Duration (min_delay)

The min_delay variable measures the recorded delay for a transit trip, expressed in minutes. It serves as a key indicator of service performance, enabling the identification of delay trends and areas requiring operational improvements. As shown in Table 1, the summary statistics for min_delay provide insights into the central tendency and variability of transit delays across the dataset.

Table 1: Summary statistics for delay duration (min_delay, in minutes)

Statistic	Value
Shortest Delay (mins)	0.000
Longest Delay (mins)	975.000
Average Delay (mins)	15.614
Median Delay (mins)	10.000
Standard Deviation (mins)	44.856

The table indicates that min_delay ranges from 0 minutes (indicating no delay) to a maximum of 975 minutes, with a mean delay of 15.614 minutes and a median of 10 minutes. The standard deviation of 44.856 reflects substantial variability in delay durations, suggesting that while many delays are relatively short, there are significant outliers with extreme delay durations.

Figure 1 visualizes the distribution of min_delay using a histogram, highlighting the variable's central tendency and spread.

The histogram visualizes the distribution of min_delay after removing extreme outliers (values above 100 minutes). As shown in Figure 1, the majority of delay durations are concentrated within the range of 0–30 minutes, with a notable peak around the mean and median of approximately 10 minutes. Extreme delays above 100 minutes were excluded for clarity, as these cases are rare and could skew the visualization.

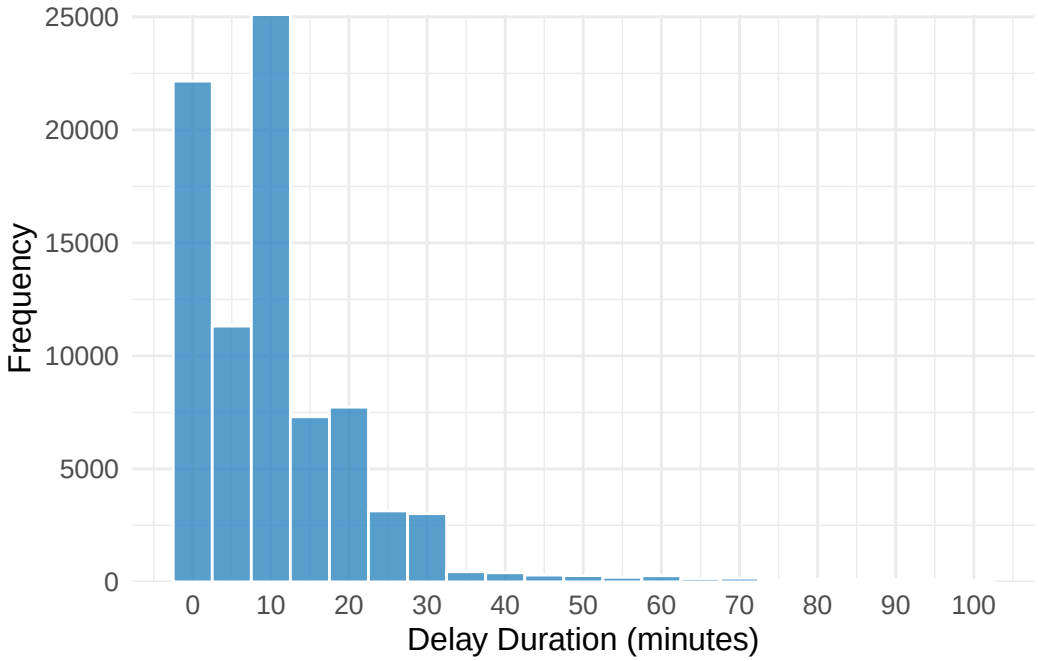

Figure 1: Distribution of delay durations (min_delay, in minutes, capped at 100)

2.4 Predictor Variables

The predictor variables include mode, location, bound, day_of_week, and min_gap. Mode captures the type of transit service—bus, streetcar, or subway—each with distinct operational dynamics influencing delay patterns. Location identifies where delays occur, focusing on DUNDAS, YORKDALE, and OTHER, to explore differences between urban and suburban transit environments. The bound variable specifies the direction of travel—north, south, east, or west—highlighting geographic and directional influences on delays. Temporal factors are represented by day_of_week, allowing an analysis of delay patterns across weekdays and weekends. Finally, min_gap measures the time gap between consecutive vehicles on the same route, reflecting service regularity and its impact on operational performance.

2.4.1 Mode of Transit

The mode variable captures the type of transit service (Bus, Subway, or Streetcar) and reflects the distinct operational challenges associated with each mode. Buses contend with external factors such as road traffic, weather, and construction, leading to frequent and highly variable delays. Subways are typically influenced by infrastructure issues like signal failures and mechanical disruptions, resulting in relatively shorter and more concentrated delays. Streetcars, operating in mixed traffic, encounter delays from route congestion and shared road usage.

Table 2 presents a summary of delay statistics for each transit mode, including average, median, maximum, and minimum delays, along with standard deviations.

Table 2: Summary statistics for delays by transit mode (in minutes)

Transit Mode	Average Delay	Median Delay	Max Delay	Min Delay	Std Dev
Bus	21.273	11	975	0	54.510
Streetcar	15.477	10	898	0	32.713
Subway	2.806	0	716	0	10.495

As shown in Table 2, buses experience the highest average delay (21.273 minutes) and the greatest variability, with a standard deviation of 54.51 minutes. Streetcars follow with an average delay of 15.477 minutes and a slightly lower standard deviation of 32.713 minutes. Subways show the lowest average delay of just 2.806 minutes, underscoring the relative reliability of rail-based transit compared to road-based services. These findings highlight the disproportionate impact of external factors like road congestion on buses and streetcars.

2.4.2 Geographic Location

The location variable provides a geographic lens for analyzing delays, focusing on two key areas: Dundas (an urban, high-traffic area) and Yorkdale (a suburban area with less frequent service). This variable allows for a comparison of transit performance in urban versus suburban contexts, offering insights into how population density and service frequency affect delays.

Table 3 summarizes delay statistics for each location, capturing key metrics like average delay, median delay, and variability.

Table 3: Summary statistics for the predictor variable (location)

Location	Average Delay	Median Delay	Max Delay	Min Delay	Std Dev
OTHER	15.717	10	975	0	45.364
DUNDAS	13.491	10	640	0	31.966
YORKDALE	11.521	9	303	0	23.821

According to Table 3, Dundas exhibits slightly higher average delays (13.491 minutes) than Yorkdale (11.521 minutes), with a standard deviation of 31.966 minutes compared to Yorkdale's 23.821 minutes. This suggests that urban environments like Dundas experience more variable and longer delays due to higher transit demand and traffic density. In contrast, Yorkdale benefits from lower demand and more predictable traffic conditions.

2.4.3 Directionality

The bound variable identifies the travel direction (North, South, East, or West) of delayed transit vehicles. This dimension helps uncover localized bottlenecks or infrastructure-related issues that may disproportionately affect certain routes.

Table 4 provides a detailed breakdown of delay statistics by direction.

Table 4: Summary statistics for delays by directional bound (in minutes)

Direction	Average Delay	Median Delay	Max Delay	Min Delay	Std Dev
East	16.591	10	975	0	43.019
North	17.252	10	950	0	49.128
South	14.742	10	967	0	36.895
West	16.450	10	950	0	41.879

As shown in Table 4, delays are relatively consistent across directions, with average delays ranging from 14.742 minutes (South) to 17.252 minutes (North). However, the standard deviation varies across directions, with the Northbound direction showing the greatest variability (49.128 minutes). These findings may reflect differences in route infrastructure, traffic bottlenecks, or operational scheduling.

2.4.4 Temporal Factors

Temporal factors, captured by the day of the week, provide insights into how commuter and recreational travel patterns influence transit delays. Weekday delays typically reflect peak

commuter traffic, while weekend delays may indicate reduced service levels or unique traffic patterns tied to recreational activities.

Table 5 presents the summary statistics for delays by day of the week, highlighting differences in average delay, variability, and extreme values.

Table 5: Summary statistics for delays by day of the week (in minutes)

Day of Week	Average Delay	Median Delay	Max Delay	Min Delay	Std Dev
Monday	15.916	10	942	0	48.686
Tuesday	15.793	10	816	0	44.651
Wednesday	15.822	10	950	0	48.737
Thursday	15.789	10	940	0	46.387
Friday	15.649	10	949	0	45.056
Saturday	14.748	10	967	0	39.019
Sunday	15.505	10	975	0	38.623

According to Table 5, average delays are relatively consistent across weekdays, peaking slightly on Monday (15.916 minutes) and Friday (15.649 minutes). The standard deviations for weekdays, such as Monday (48.686 minutes) and Friday (45.056 minutes), indicate substantial variability in delays during peak commuter periods. Weekend delays are slightly lower on average, with Saturday showing the lowest average delay (14.748 minutes) and reduced variability (39.019 minutes). These patterns align with the expected reduction in transit demand and traffic congestion during weekends.

2.4.5 Time Gap Between Vehicles (min_gap)

The min_gap variable captures the time gap (in minutes) between consecutive vehicles on the same route, providing insight into service regularity. Maintaining consistent gaps is crucial for operational efficiency and minimizing passenger wait times. Table 6 presents the summary statistics for min_gap.

Table 6: Summary statistics for time gaps between vehicles (min_gap, in minutes)

Statistic	Value
Shortest Gap (mins)	0.000
Longest Gap (mins)	997.000
Average Gap (mins)	24.438
Median Gap (mins)	19.000
Standard Deviation (mins)	47.659

The data indicates that min_gap ranges from 0 to 997 minutes, with an average time gap of 24.438 minutes and a median of 19 minutes. A standard deviation of 47.659 reflects substantial variability, likely influenced by external factors such as traffic congestion, irregular vehicle arrivals, and operational inefficiencies. The presence of extremely high values, such as 997 minutes, suggests rare but significant disruptions that may warrant further investigation or exclusion in detailed analyses.

Figure 2 depicts the distribution of min_gap, using a histogram to illustrate common intervals and variability.

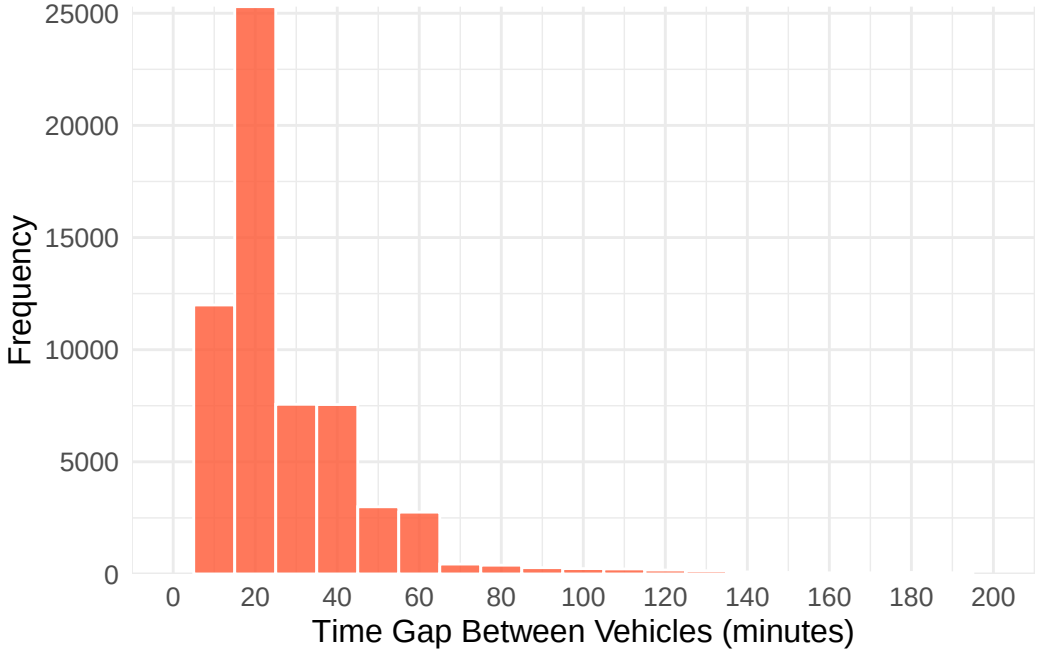

Figure 2: Distribution of time gaps between vehicles (min_gap, in minutes)

The histogram illustrates that the majority of min_gap values fall within the 15–30 minute range, with a noticeable peak around 20 minutes. While most values are clustered within this interval, there are occasional larger gaps that extend up to 100 minutes, reflecting variability in service intervals. These deviations may highlight operational inconsistencies or external disruptions affecting transit schedules.

3 Methods

This section outlines the methodology employed to analyze patterns in Toronto Transit Commission (TTC) delays (min_delay) using a Random Forest model. By leveraging an ensemble approach, the Random Forest effectively handles nonlinear relationships and interactions

between predictors, providing robust predictions for delay durations across diverse transit contexts.

3.1 Model Overview

The goal of the model is to predict the delay duration (min_delay) in minutes for TTC transit vehicles using the following predictors:

$$\text{min\_delay} \sim \text{mode} + \text{location} + \text{bound} + \text{day\_of\_week} + \text{min\_gap} \tag{1}$$

Where:

- **mode**: The transit mode (Bus, Streetcar, or Subway).
- **location**: Geographic location (Dundas, Yorkdale, or Other).
- **bound**: Travel direction (North, South, East, or West).
- **day_of_week**: Day of the week (Monday through Sunday).
- **min_gap**: Time gap between vehicles in minutes, representing service regularity.

3.2 Training and Validation

To ensure robust performance, the dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain the distribution of min_delay min_delay. The training set was used to build the Random Forest model, while the testing set evaluated its predictive accuracy.

The model was implemented in R using the randomForest package, with the number of trees (`ntree`) set to 500. This parameter balances computational efficiency and model stability by ensuring sufficient averaging across trees.

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{x}_i) \tag{2}$$

Where:

- $\hat{y}_i$: Predicted delay for observation i.
- T: Total number of trees ($T = 500$).
- $f_t(\mathbf{x}_i)$: Prediction from tree t based on predictors $\mathbf{x}_i$.

The model hyperparameters, including the default number of predictors sampled at each split ($\sqrt{p}$, where p is the number of predictors), were kept unchanged, as they offered a balanced trade-off between bias and variance.

3.3 Evaluation Metrics

The model's performance was assessed using:

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{3}$$

  Where n is the number of observations, y_i is the observed delay, and $\hat{y}_i$ is the predicted delay.

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4}$$

  This measures the proportion of variance in the observed data explained by the model. Values close to 1 indicate strong predictive performance.

3.4 Model Justification, Assumptions, and Limitations

The Random Forest regression model was selected for its ability to handle the diverse and nonlinear relationships within the dataset, making it the most suitable choice for analyzing TTC delays. This model is particularly effective at capturing complex interactions between predictors, such as transit mode and day of the week, which influence delay durations.

The justification for using Random Forest lies in its ability to model nonlinear relationships without assuming specific functional forms. Unlike linear regression, which oversimplifies these relationships, Random Forest accommodates interactions between predictors and provides robust predictions even in the presence of multicollinearity. For example, the influence of min_gap on delays is not consistent across different transit modes, a pattern that Random Forest can identify and model effectively. Additionally, the algorithm provides feature importance metrics, offering interpretable insights into the drivers of delays, such as mode, location, and day_of_week.

Alternative models, such as linear regression, were considered but found to be insufficient for this analysis. Linear regression assumes linear relationships and independence between predictors, leading to poor performance metrics, including lower R^2 values and higher RMSE scores, when applied to this dataset. These results highlighted the need for a more flexible model, prompting the selection of Random Forest.

Key predictors were evaluated for their importance using the Mean Decrease in Accuracy (MDA) metric, which measures the reduction in predictive accuracy when a variable is excluded.

This analysis, performed using the `randomForest` package (Andy Liaw and Matthew Wiener 2024), revealed that min_gap (time between consecutive vehicles) and mode (transit type) were the most significant predictors. The prominence of min_gap underscores its role in service regularity, while the importance of mode reflects the operational differences among buses, subways, and streetcars. Other influential predictors included day_of_week and location, capturing temporal and geographic dimensions of delays.

While Random Forest excelled in its predictive capabilities, the model operates under several assumptions and faces certain limitations. First, although it handles outliers better than simpler models, extreme values, such as exceptionally high min_gap durations, can still affect individual tree predictions. These outliers were mitigated during preprocessing (see Section C). Second, the model is less interpretable compared to linear regression, as it does not provide straightforward coefficients for predictors. Feature importance metrics were used to partially address this limitation. Third, Random Forest can be sensitive to imbalances in predictor distributions, such as fewer observations for specific locations or directional bounds. Careful data cleaning and preprocessing steps were implemented to reduce this bias.

Despite its strengths, the model has limitations in its applicability. For example, it may struggle with extreme imbalances in predictor distributions or scenarios requiring highly interpretable predictions. Future work could explore ensemble methods that integrate Random Forest with more interpretable models to address these challenges. The comprehensive evaluation and selection process affirm the Random Forest model as the best fit for understanding and predicting TTC delays.

See Section B for additional details.

4 Results

4.1 Visualizing Delays

Figure 3 illustrates the average delay duration across three key Toronto locations: Dundas, Yorkdale, and "Other." The figure shows distinct variations in delay durations, with Other locations experiencing the longest delays, averaging 15.7 minutes, compared to 13.5 minutes at Dundas and 11.5 minutes at Yorkdale. These results suggest that factors specific to "Other" locations, such as less efficient transit infrastructure or increased congestion, contribute to higher delays. Conversely, Yorkdale exhibits relatively shorter delays, which could reflect better-managed transit operations or reduced traffic volume in the area.

Figure 4 highlights the temporal trends in average TTC delays across the days of the week. Mondays show the highest average delays (15.9 minutes), potentially due to increased traffic demand as the workweek begins. Saturdays see the lowest average delays (14.7 minutes), aligning with reduced commuter traffic during weekends. While the variations in delays across

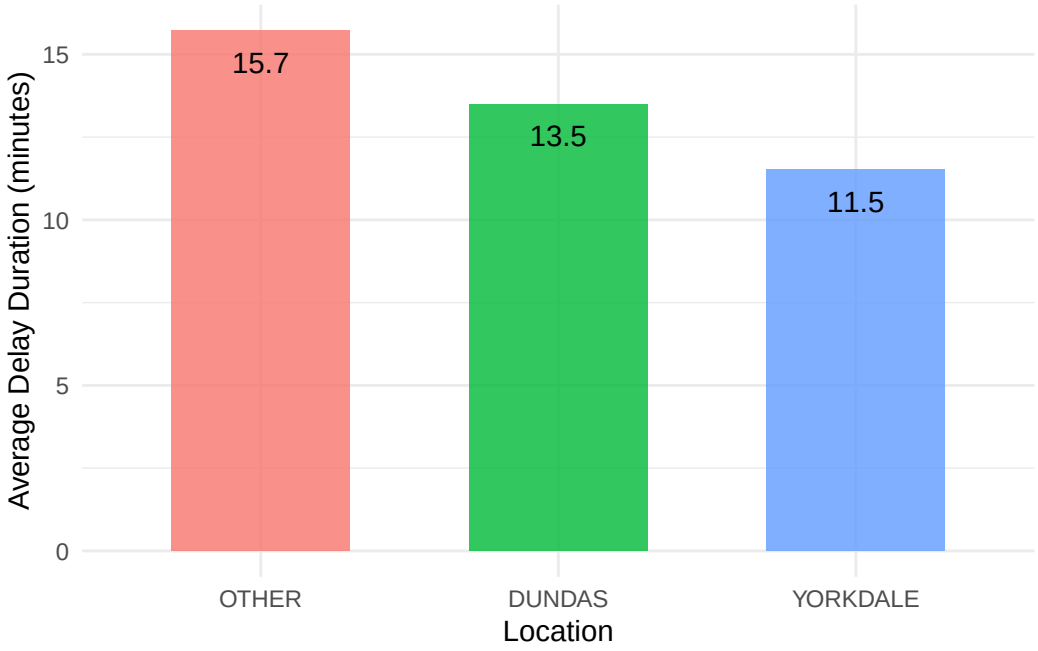

Figure 3: Average delays across Toronto locations, with average delay values annotated.

days are relatively small, these findings emphasize the subtle but consistent influence of weekly transit demand patterns on delay durations.

The distribution of delays across different transit modes—bus, streetcar, and subway—is presented in Figure 5. Buses exhibit the most variability, with delays ranging from a few minutes to over 750 minutes, likely due to their exposure to road traffic and weather conditions. Streetcars show moderate delays, reflecting their shared use of road infrastructure but also benefiting from dedicated tracks in certain areas. Subways demonstrate the most consistent performance, with shorter and less variable delays, likely due to their isolated underground operation.

Figure 6 examines the average delays based on the travel direction: East, West, North, and South. Northbound trips have the longest average delays (17.3 minutes), while southbound trips experience the shortest delays (14.7 minutes). East and westbound trips have comparable delays, averaging 16.6 minutes and 16.4 minutes, respectively. These findings may reflect route-specific operational challenges or variations in traffic and transit infrastructure across different directions.

The scatterplot in Figure 7 illustrates the relationship between the time gap between consecutive transit vehicles and the resulting delay duration. The trend line indicates a positive correlation: larger time gaps correspond to longer delays. This relationship highlights the importance of maintaining consistent headways to minimize delays and optimize transit performance. The scatterplot also shows a wide range of delays for small time gaps, suggesting

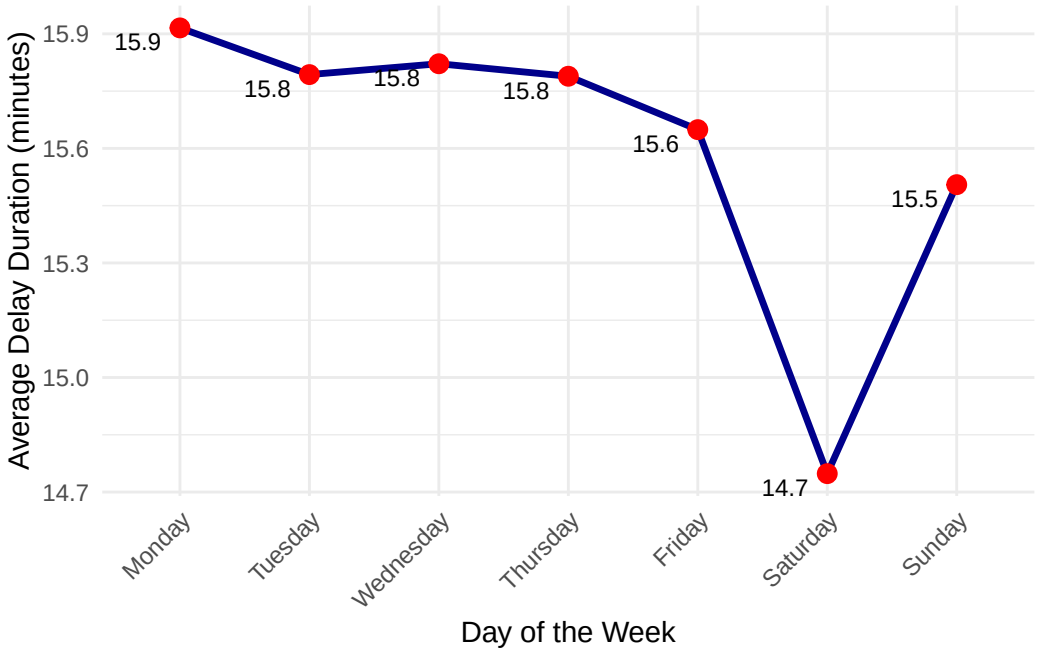

Figure 4: Temporal trends in TTC delays, with average delay values annotated.

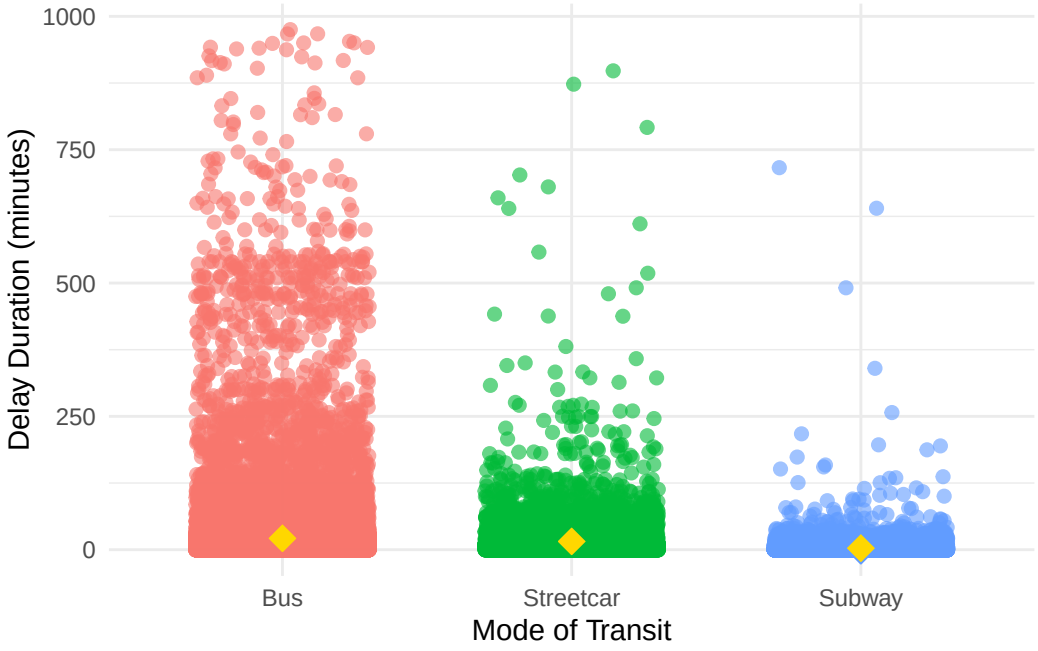

Figure 5: Distribution of delays by mode of transit

Figure 6: Average delays by travel direction.

that other factors, such as operational issues or traffic conditions, play a significant role in influencing delays.

A closer examination of streetcar lines 501 and 510, shown in Figure 8, reveals notable differences in their delay durations. Line 501 experiences longer average delays (16.1 minutes) compared to line 510 (10.7 minutes). These disparities may be attributed to differences in route length, traffic congestion, or infrastructure quality. Line 501's longer delays suggest the need for targeted interventions to improve reliability along this route.

Figure 9 explores the interplay between daily delays and the day of the week over time. The temporal trends reveal significant variations, with some months exhibiting consistent spikes in delays on specific weekdays. For instance, delays are particularly high during April, potentially reflecting seasonal disruptions such as weather-related challenges or scheduled maintenance. These patterns emphasize the need for adaptive scheduling and operational strategies to address temporal fluctuations in delay durations.

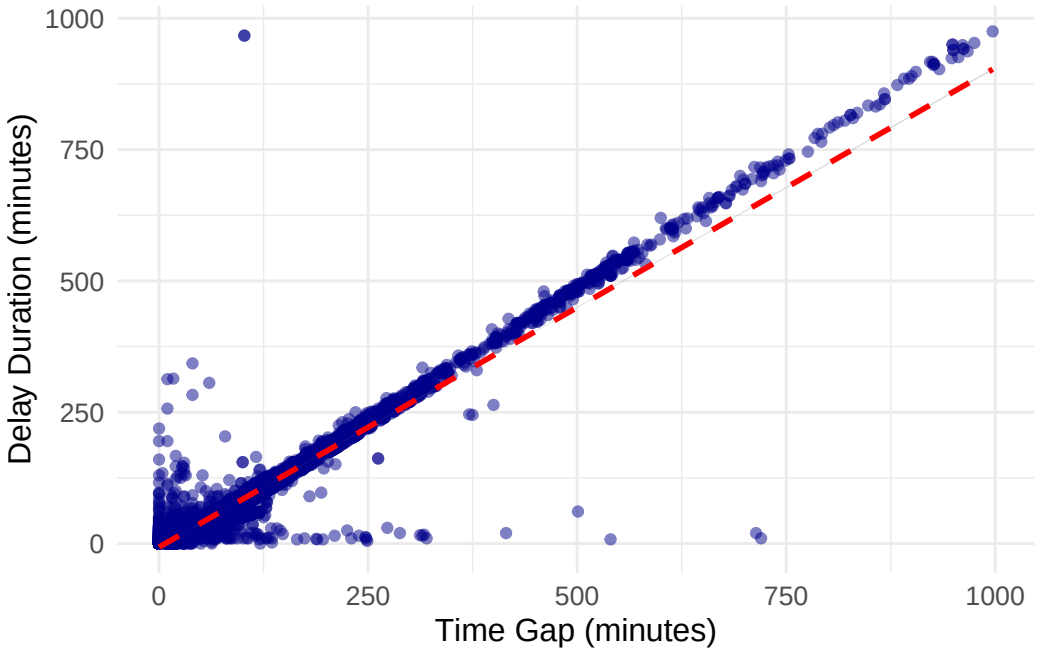

Figure 7: Relationship between time gap and delay duration.

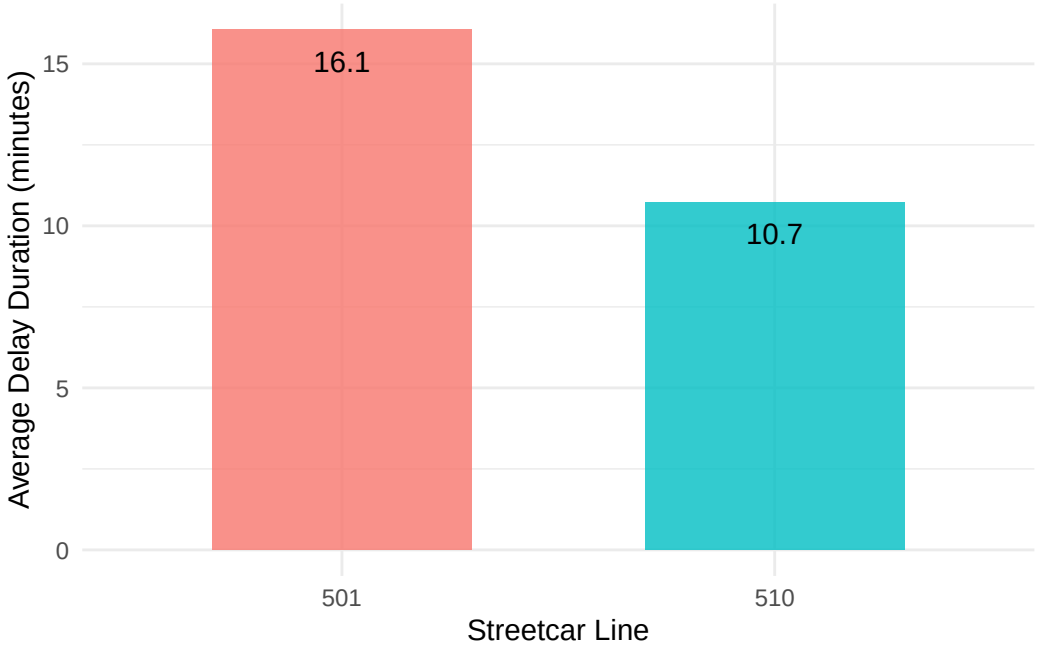

Figure 8: Average delays for Streetcar Lines 501 and 510.

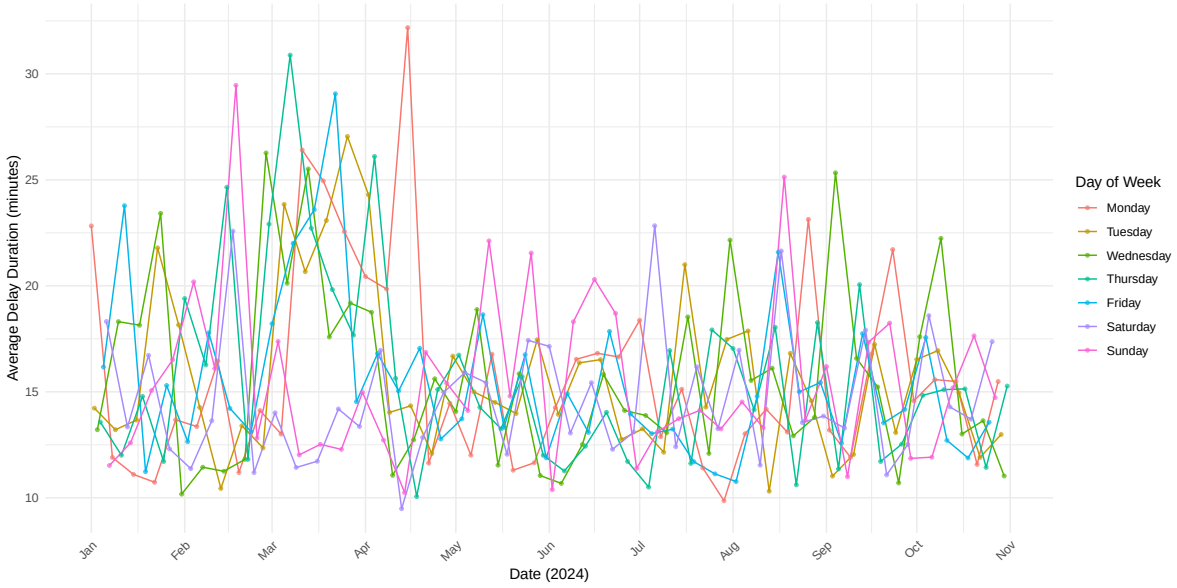

Figure 9: Temporal Trends in Delays by Day of the Week and Date.

4.2 Modeling Performance and Residual Analysis

The random forest model developed to predict TTC delays demonstrates strong predictive performance, as summarized in Table 7. Key metrics include:

- R-squared: 0.887, indicating the model explains 88.7% of the variance in delay duration.
- RMSE: 20.815 minutes, reflecting the typical error in delay predictions. These results suggest that the model effectively captures the key factors influencing delays, providing a reliable basis for forecasting and analysis.

Table 7: Random Forest Model Performance Metrics

Metric	Value
R-squared	0.887
RMSE	20.815

Figure 10 highlights the contribution of each feature to the random forest model's predictions. Features with greater importance, such as min_gap and mode, exert a stronger influence on delay predictions. This underscores the critical factors driving TTC delays and emphasizes the significance of leveraging detailed temporal, spatial, and operational data for accurate predictive modeling.

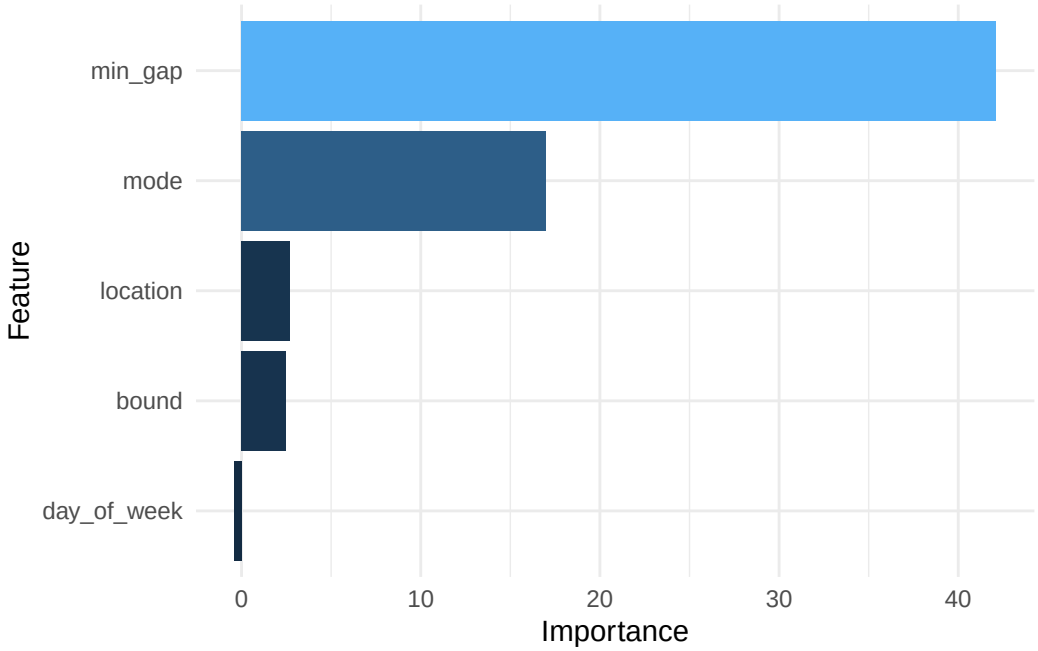

Figure 10: Feature importance in predicting TTC delays

The residual analysis, presented in Figure 11, examines the alignment between observed and predicted delay durations. The scatterplot shows that most predictions align closely with the observed values, particularly for shorter delays. Minor deviations are observed for extreme delays, indicating areas for further refinement in the model. Overall, the residual analysis supports the robustness of the model in accurately predicting TTC delays.

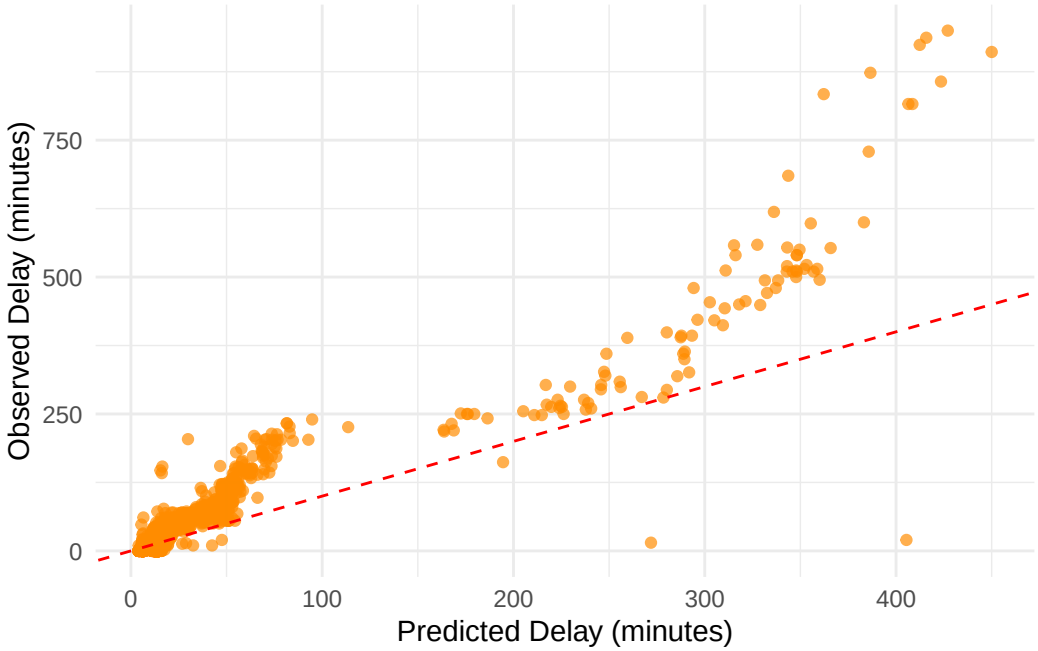

Figure 11: Residual analysis for model predictions vs. observed delays.

5 Discussion

5.1 Interpretation of Key Findings

This study uncovers critical factors influencing delays in the Toronto Transit Commission (TTC) system, revealing the interplay between geographic, modal, and temporal dimensions. The random forest model identified transit mode and service regularity (min_gap) as the most influential predictors, while geographic and temporal factors provided further explanatory power. These findings align with transit literature that emphasizes the multidimensional nature of delays, where operational, spatial, and temporal elements converge to shape commuter experiences (Currie and Sarvi 2012; Cats and Gkioulou 2017).

The analysis of delay patterns across locations (Figure 3) demonstrates that Dundas, situated in Toronto's urban core, experiences longer delays than suburban Yorkdale. This difference reflects the challenges of operating in high-demand, high-congestion environments, consistent with urban transit studies that identify population density and traffic bottlenecks as significant contributors to delays (Litman 2023). The variability in delays across modes, illustrated in Figure 5, highlights the unique challenges faced by buses, which are more exposed to traffic conditions and external disruptions compared to streetcars and subways. Temporal trends (Figure 4) further reveal that delays peak on weekdays, particularly on Mondays and Fridays, correlating with commuter-driven demand.

The strong performance of the random forest model, as summarized in Table 7, underscores the predictive power of incorporating multiple predictors, including min_gap and mode. The R-squared value of 0.887 and RMSE of 20.815 minutes indicate that the model captures most of the variance in delay durations while maintaining reasonable predictive accuracy.

5.2 Geographic and Modal Challenges

The geographic focus on Dundas and Yorkdale reveals stark contrasts in delay dynamics. Dundas exhibits higher average delays, likely due to high transit demand and frequent service intervals, as reflected in Table 3. These findings align with urban transit studies that highlight how congestion and dense service schedules exacerbate delays in city centers (Cats and Gkioulou 2017). In contrast, Yorkdale's suburban environment, characterized by lower demand and less frequent service, provides a more predictable operating context.

The analysis underscores the disproportionate challenges faced by buses, which exhibit the longest and most variable delays across all modes (Table 2). These findings mirror global patterns where buses are particularly susceptible to traffic congestion, road conditions, and external events (Hess and Lombardi 2007). Streetcars, operating in mixed-traffic conditions, experience moderate delays, reflecting their partial reliance on dedicated tracks. Subways emerge as the most reliable mode, benefiting from isolated infrastructure that minimizes external disruptions.

The scatterplot in Figure 7 highlights the positive correlation between larger time gaps (min_gap) and longer delays, emphasizing the importance of maintaining consistent headways for transit reliability. This aligns with findings in transit scheduling literature, which emphasize the critical role of regular intervals in minimizing delays and optimizing passenger wait times (Currie and Sarvi 2012).

5.3 Temporal Patterns and Scheduling Implications

Temporal variations in delays reveal significant insights into the influence of weekly commuter patterns. Delays peak on Mondays and Fridays, consistent with higher transit demand during the start and end of the workweek (Table 5). These trends underscore the need for dynamic scheduling that aligns service levels with commuter demand, as suggested by operational best practices (Litman 2023).

The analysis also identifies Saturdays as having the lowest average delays, reflecting reduced transit usage and operational stress during weekends. This finding is consistent with transit studies emphasizing the cyclical nature of weekday versus weekend demand patterns (Hess and Lombardi 2007).

5.4 Limitations and Weaknesses

Despite the robustness of the findings, several limitations warrant consideration. The study's geographic scope is limited to two locations, Dundas and Yorkdale, which restricts the generalizability of results to other parts of Toronto. Expanding the analysis to include a broader range of locations would provide a more comprehensive understanding of transit dynamics.

The exclusion of contextual variables, such as weather conditions, construction activity, and special events, represents another limitation. Previous studies highlight the impact of such factors on transit reliability (Hess and Lombardi 2007), suggesting that their inclusion could improve the model's explanatory power. Additionally, the focus on weekday versus weekend patterns, without accounting for finer-grained time-of-day trends, limits insights into peak versus off-peak delay dynamics.

While the random forest model provides strong predictive accuracy, its interpretability is limited compared to traditional regression models. Future research could explore hybrid approaches that combine the predictive strength of machine learning with the interpretability of linear models, addressing this trade-off (Breiman 2001).

5.5 Future Directions

Future research should expand the geographic scope to include additional locations across Toronto's transit network. This would enable a more nuanced analysis of spatial variations and their implications for system-wide operations. Comparative analyses between urban, suburban, and rural contexts could further enrich understanding.

Integrating contextual variables such as weather, road closures, and special events would provide a more comprehensive understanding of delay drivers. For example, extreme weather events are known to disrupt transit operations, increasing delay variability (Currie and Sarvi 2012). Including such factors could enhance both predictive accuracy and practical relevance.

The incorporation of real-time data into dynamic models could enable adaptive scheduling and operational decision-making. Additionally, integrating passenger metrics, such as satisfaction surveys or ridership data, would provide a holistic view of transit performance. This approach aligns with emerging trends in transit research, emphasizing the integration of operational and user-centered metrics (Hess and Lombardi 2007).

Future studies should explore alternative modeling techniques, such as mixed-effects models or ensemble approaches, to better capture hierarchical relationships and predictor interactions. These techniques could address some limitations of random forests while retaining predictive strength (Breiman 2001).

Appendix

A Idealized Methodology for a Survey on TTC Delays

A.1 Overview

To complement the quantitative analysis of TTC delays, an idealized survey methodology could capture commuter perceptions and experiences with transit delays across Toronto. The survey would aim to identify the factors influencing passenger satisfaction, the perceived causes of delays, and preferences for operational improvements. This approach would provide a deeper understanding of commuter behavior and the social implications of transit delays, enriching the findings presented in this study.

A.2 Sampling Approach

A stratified random sampling method would be employed to ensure representation across key demographic and transit-related variables. This approach would capture variations in delay experiences based on commuter profiles and transit contexts, facilitating detailed comparisons across geographic locations, transit modes, and demographic groups.

1. Transit Mode: The sample would include users of buses, streetcars, and subways to reflect the modal diversity within the TTC system. Bus users, who experience the highest variability in delays, would form a larger stratum, followed by streetcar and subway users. This stratification accounts for the operational differences among modes, as highlighted in Table 2.
2. Geographic Location: Respondents would be grouped based on their primary transit location, focusing on urban (e.g., Dundas) and suburban (e.g., Yorkdale) contexts. Urban commuters are likely to report longer delays and more frequent disruptions, as shown in Figure 3, whereas suburban commuters may face less frequent but potentially longer interruptions.
3. Commuter Profile: Respondents would be stratified by key demographic variables such as age, employment status, and commuting purpose. For example, working professionals may emphasize reliability during peak hours, while students may prioritize affordability. Stratifying by these variables ensures the inclusion of diverse perspectives on TTC service quality.
4. Frequency of Use: Respondents would be grouped into frequent users (daily commuters), occasional users (weekly or less), and infrequent users (sporadic riders). Frequent users may offer more detailed insights into recurring delay patterns, while occasional users may provide perspectives on specific disruptions.

A total sample size of approximately 1,500 respondents would be targeted, divided proportionally across the strata. This approach ensures meaningful representation and supports robust statistical analyses of commuter perceptions and preferences.

A.3 Recruitment Method

Participants would be recruited through multichannel outreach, leveraging digital platforms, in-transit recruitment, and community-based engagement.

Online Panels: Collaborating with consumer research platforms would provide access to a broad demographic range, ensuring representation of frequent and infrequent transit users. Online recruitment enables targeted outreach based on commuter profiles and geographic locations.

In-Transit Recruitment: QR codes placed on TTC vehicles and stations would invite commuters to participate in the survey. This method ensures real-time engagement with active transit users, capturing immediate perceptions of delays.

Community Outreach: Partnering with local organizations, community centers, and neighborhood associations would help reach underrepresented groups, including suburban commuters and culturally diverse populations. Social media campaigns targeting Toronto-based users would further enhance participation.

A.4 Survey Structure

The survey would include a mix of closed-ended and open-ended questions to balance quantitative insights with qualitative feedback.

A.4.1 Question Types

1. **Commuting Habits**

- Frequency of transit use (e.g., daily, weekly).
- Primary mode of transit and typical routes taken.
- Average time spent commuting.

2. **Delay Perceptions**

- Perceived frequency of delays (e.g., often, sometimes, rarely).
- Typical delay duration experienced (e.g., less than 10 minutes, 10–20 minutes).
- Satisfaction with delay communication (e.g., announcements, apps).

3. **Impact of Delays**

- How delays affect daily schedules (e.g., late to work, missed appointments).
- Strategies used to mitigate delays (e.g., leaving earlier, using alternative routes).
- Willingness to pay for reliability improvements (e.g., higher fares for faster service).

4. **Transit Satisfaction**

- Overall satisfaction with TTC services on a Likert scale.
- Perceptions of mode-specific reliability (bus vs. streetcar vs. subway).
- Desired operational improvements (e.g., more frequent service, real-time delay updates).

5. ***External Factors**

- Awareness of factors contributing to delays (e.g., traffic, weather, construction).
- Perceptions of TTC's efforts to address delays.
- Suggestions for improving commuter experiences during delays.

A.4.2 Sample Survey

Background and Purpose of the Study
This survey is part of a research study aimed at understanding transit delays within the Toronto Transit Commission (TTC). By participating, you will help identify key patterns in delays and provide insights to improve transit reliability and commuter experiences.

Participation Involvement
You are invited to complete a short survey about your experiences with TTC transit delays. The survey will take approximately 5–7 minutes and includes both multiple-choice and open-ended questions about your transit habits and perceptions.

Confidentiality
Your responses will remain strictly confidential. No identifying information will be shared in any reports or presentations. All responses will be anonymized, ensuring that individual participants cannot be identified in the analysis.

Usage of Data
The results of this study will contribute to research reports and presentations designed to better understand the factors influencing TTC transit delays. By consenting, your responses will be included in the analysis, providing critical data for this research.

Withdrawal from the Study
Participation in this survey is entirely voluntary. You may withdraw your consent and request the removal of your data within one month of survey completion by contacting the research team.

Informed Consent

After reading the details of the study, do you consent to have your anonymized survey responses included in the research?
- Yes, I consent to participate and have my responses included.
- No, I do not consent, but I would still like to complete the survey.

1. How often do you use TTC transit services?

- Daily
- Weekly
- Monthly
- Rarely

2. Which mode of transit do you primarily use?

- Bus
- Streetcar
- Subway

3. How frequently do you experience delays?

- Always
- Often
- Sometimes
- Rarely
- Never

4. What is the average delay duration you experience?

- Less than 5 minutes
- 5–10 minutes
- 10–20 minutes
- More than 20 minutes

5. How satisfied are you with the communication of delays by TTC?

- Very dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very satisfied

6. How do delays typically impact your daily routine?

- No impact
- Minor inconvenience
- Significant disruption

- Major disruption

7. Would you support increased fares for improved reliability?

- Strongly oppose
- Oppose
- Neutral
- Support
- Strongly support

8. What operational improvements would you prioritize?

- Increased service frequency
- Real-time delay updates
- Improved delay communication
- Other (please specify)

9. What factors do you believe most contribute to delays?

- Traffic congestion
- Weather conditions
- Infrastructure issues
- Other (please specify)

10. How satisfied are you overall with TTC services?

- Very dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very satisfied

11. How do you usually respond to transit delays? (Select all that apply)

- Wait for the delay to resolve
- Switch to another mode of transit (e.g., take a bus instead of a subway)
- Cancel or change your plans
- Other (please specify): _____

Open-Ended Feedback

11. Can you describe a specific instance where a TTC delay significantly impacted you? How did this affect your perception of TTC services?

12. In your opinion, what actions or improvements could the TTC take to minimize delays or improve commuter experiences?

End of Survey

Thank you for your time and participation! Your insights are invaluable to this research and will contribute to a better understanding of TTC transit delays and the development of strategies to enhance transit reliability. For further information or updates on this research, please feel free to contact the research team.

A.5 Linkage to Literature

The survey design leverages established principles in sampling methodology and survey research to ensure robust and representative data collection. Stratified random sampling is a foundational technique in survey methodology, recommended for achieving representative samples across key population subgroups (Lohr 2019). By dividing the population into meaningful strata—such as transit modes, geographic locations, and demographic categories—this approach improves precision and ensures that diverse perspectives are captured in the analysis (Groves et al. 2009).

To maximize participation and minimize selection bias, multichannel recruitment strategies are employed. Research emphasizes the importance of utilizing varied recruitment methods, including digital platforms, in-person outreach, and community engagement, to enhance inclusivity and reach underrepresented populations (Couper 2000; Dillman, Smyth, and Christian 2014). This multichannel approach aligns with best practices for public surveys, allowing researchers to overcome accessibility barriers and improve response rates.

Including questions on delay perceptions and their impacts aligns with broader survey design strategies that emphasize the collection of both subjective and objective data for comprehensive analysis (Fowler 2014). The focus on mode-specific reliability and geographic variations reflects principles from urban mobility studies, where localized and operational factors are critical to understanding transit system performance (Babbie 2020).

By integrating demographic, geographic, and behavioral data, this survey aims to provide a detailed and representative picture of the factors influencing TTC delays. This approach not only enhances the validity of the findings but also supports actionable insights for improving commuter experiences and operational planning.

B Model Details

B.1 Out-of-Bag Error Analysis

The Out-of-Bag (OOB) error rate, visualized in Figure 12, provides an internal estimate of the Random Forest model's predictive performance during training. The OOB error rate decreases rapidly as the number of trees grows, stabilizing after approximately 200 trees, indicating that the model has achieved an optimal balance between bias and variance.

Unlike the RMSE calculated on the test dataset, the OOB error reflects how well the model generalizes to unseen data during training. The stabilization of the OOB error suggests that additional trees provide diminishing improvements, ensuring robustness without overfitting. While the specific OOB error values depend on the dataset and feature selection, the curve highlights the model's iterative refinement, demonstrating that it effectively captures the patterns in the training data. This behavior underlines the reliability of the Random Forest approach for modeling TTC delays.

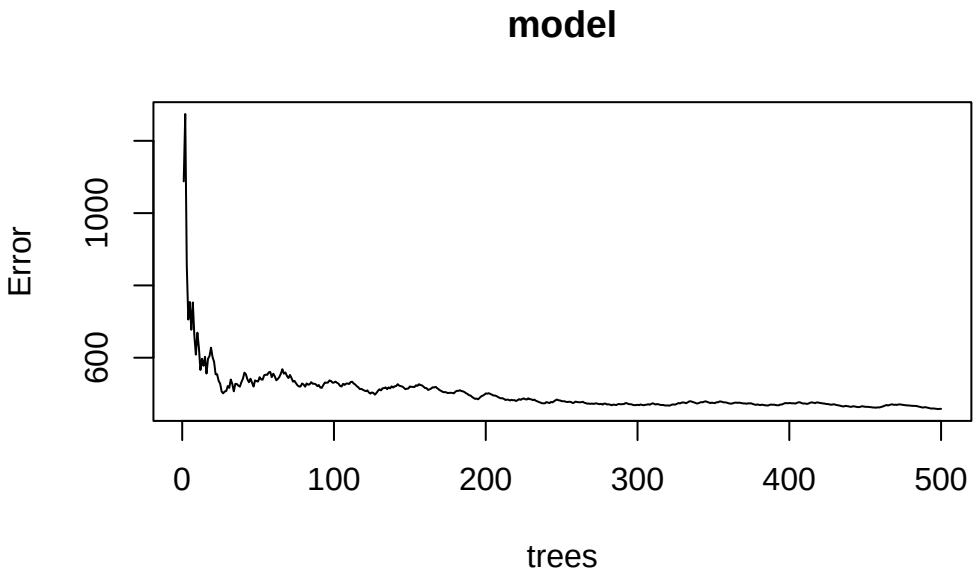

Figure 12: Out-of-Bag error for the Random Forest model.

B.2 Residuals Analysis

Table 8 provides a detailed summary of the observed, predicted, and residual values from the test dataset. The mean observed delay is 16.138 minutes, while the model's mean predicted

delay is 16.181 minutes, resulting in a minimal mean residual of -0.043 minutes. The Root Mean Squared Error (RMSE) is 20.815, reflecting the average deviation between predicted and actual values. This small residual error suggests that the model provides robust predictions across different transit scenarios.

Table 8: Summary of Model Predictions on Test Data

Statistic	Value
Mean Observed Delay	16.138
Mean Predicted Delay	16.181
Mean Residual	-0.043
RMSE	20.815

Residual Analysis by Transit Mode Residuals, which represent the difference between observed and predicted delay values, are analyzed by transit mode in Table 9. Key observations include:

- Buses: Mean residual of 0.864 minutes with a standard deviation of 24.378, reflecting higher variability due to external factors like traffic congestion.
- Streetcars: Slightly negative mean residual (-0.374 minutes) with lower variability (SD 17.967), indicating more consistent predictions.
- Subways: The lowest mean residual (-2.446 minutes) and variability (SD 6.122), aligning with the relatively stable performance of subway services. This analysis emphasizes the model's ability to capture delay dynamics specific to each mode while acknowledging the unique operational challenges buses face.

Table 9: Error Distribution by Transit Mode

Transit Mode	Mean Residual	Median Residual	Residual Std Dev
Bus	0.864	-3.211	24.378
Streetcar	-0.374	-3.821	17.967
Subway	-2.446	-3.547	6.122

Day-of-Week Residual Patterns Residuals are further analyzed by day of the week in Table 10. The findings reveal that residuals vary slightly across days, with the largest mean residual on Wednesday (0.678 minutes) and the lowest on Sunday (-0.858 minutes). These patterns likely reflect differences in transit demand and scheduling efficiency during weekdays compared to weekends.

- Weekdays: Higher variability, consistent with increased commuter traffic and peak-hour disruptions.

- Weekends: Lower residuals and variability, indicating better alignment between predicted and observed delays. This analysis supports the model's ability to adapt to temporal fluctuations in transit operations.

Table 10: Residuals by Day of the Week

Day of Week	Mean Residual	Median Residual	Residual Std Dev
Monday	-0.231	-3.506	20.371
Tuesday	-0.066	-3.652	18.046
Wednesday	0.678	-3.746	26.748
Thursday	0.666	-3.686	24.035
Friday	-0.432	-3.460	19.077
Saturday	-0.297	-3.902	19.924
Sunday	-0.858	-3.933	13.159

The diagnostic metrics demonstrate the robustness of the Random Forest model in capturing delay patterns across modes and temporal strata. The low RMSE, combined with minimal mean residuals, underscores the model's predictive accuracy. Additionally, the nuanced residual patterns by mode and day of the week highlight the importance of contextual factors, such as traffic congestion and scheduling practices, in influencing delay dynamics.

C Data Cleaning Details

The cleaning process for the TTC delay data was a critical step in preparing the raw datasets for meaningful analysis. This process focused on standardizing, harmonizing, and integrating raw data from bus, subway, and streetcar services into a single, comprehensive dataset. The cleaning script handled inconsistent formats across the datasets and ensured that relevant variables were appropriately structured. Three separate raw datasets were processed: raw_data_bus.csv, raw_data_subway.csv, and raw_data_streetcar.csv. Each dataset was preprocessed to address mode-specific nuances before being combined into a unified dataset.

The cleaning process began by loading the datasets and standardizing date formats to the common YYYY-MM-DD structure using the as.Date() function. Temporal variables such as the day of the week were derived from the standardized date column, enabling weekday-specific analysis. For the bus data, the Direction column was recoded to cardinal directions (North, South, East, West) using case_when(), with invalid entries marked as NA. Similarly, locations in the bus data were categorized into key landmarks such as DUNDAS and YORKDALE, with all other locations grouped under OTHER. Additional columns like Min Delay and Min Gap were retained as quantitative metrics of delay durations and gaps. Since buses do not have defined lines, a placeholder column for line was added to ensure compatibility with subway and streetcar data.

The subway data underwent a similar standardization process. The Bound column was cleaned to align with cardinal directions, and the Station column was grouped into predefined categories based on prominent landmarks or grouped under OTHER. As with the bus data, Min Delay and Min Gap metrics were preserved for further analysis, and a placeholder line column was included for consistency. The streetcar data required additional transformations to address its unique structure. Specific lines, such as 510 and 501, were explicitly retained, while other streetcar lines were grouped under OTHER. Locations and directions were cleaned following the same approach used for buses and subways.

Once cleaned, the datasets for all three transit modes were combined using bind_rows(), resulting in a unified dataset with consistent column names and data structures. This combined dataset included key attributes such as the mode of transit, date, day of the week, cardinal direction of travel, delay metrics (Min Delay and Min Gap), location, and line information. The cleaned dataset captures essential information while reducing noise from irrelevant or inconsistent raw data entries.

C.1 Combined Dataset

The final cleaned dataset integrates data from bus, subway, and streetcar services, creating a comprehensive resource for analyzing TTC delays. The dataset includes harmonized variables such as mode, date, day_of_week, bound, min_delay, min_gap, location, and line. This structure ensures compatibility across transit modes and provides the flexibility to perform

mode-specific or aggregate analyses. The first few rows of the cleaned dataset are shown in Table 11, highlighting the key variables retained after cleaning.

Table 11: Final Cleaned TTC Delay Dataset

Mode	Date	Day of Week	Bound	Min Delay	Min Gap	Location	Line
Bus	2024-01-01	Monday	North	10	20	OTHER	NA
Bus	2024-01-01	Monday	NA	20	40	OTHER	NA
Bus	2024-01-01	Monday	NA	0	0	OTHER	NA
Bus	2024-01-01	Monday	North	0	0	OTHER	NA
Bus	2024-01-01	Monday	NA	0	0	OTHER	NA
Bus	2024-01-01	Monday	North	8	16	OTHER	NA

To facilitate efficient storage and downstream analysis, the cleaned dataset was saved as a Parquet file using the arrow package. This format ensures data integrity and allows for faster read and write operations compared to traditional CSV files. The cleaned dataset was saved to the analysis_data directory for use in subsequent modeling and analysis.

References

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

Andy Liaw and Matthew Wiener. 2024. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. https://CRAN.R-project.org/package=randomForest.

Apache Arrow Developers. 2024. *arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

Babbie, Earl R. 2020. *The Practice of Social Research*. 15th ed. Boston, MA: Cengage Learning.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Cats, Oded, and Zafeira Gkioulou. 2017. "Modeling the Impacts of Public Transport Reliability and Travel Information on Passengers' Waiting-Time Uncertainty." *EURO Journal on Transportation and Logistics* 6 (3): 247–70. https://doi.org/10.1007/s13676-014-0070-4.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64 (4): 464–94. https://doi.org/10.1086/318641.

Currie, Graham, and Majid Sarvi. 2012. "New Model for Secondary Benefits of Transit Priority." *Transportation Research Record: Journal of the Transportation Research Board* 2276 (1): 1–8. https://doi.org/10.3141/2276-08.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. 4th ed. Hoboken, NJ: Wiley.

Fowler, Floyd J. 2014. *Survey Research Methods*. 5th ed. Thousand Oaks, CA: SAGE Publications.

Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. Hoboken, NJ: Wiley.

Hadley Wickham. 2024a. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

———. 2024b. *stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Hadley Wickham and others. 2024a. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

———. 2024b. *tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Hadley Wickham, Jim Hester, and others. 2024. *testthat: Unit Testing for R*. https://CRAN.R-project.org/package=testthat.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2024. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Hess, Daniel B., and Peter A. Lombardi. 2007. "Policy Support for and Barriers to Transit-Oriented Development in Canada: The Importance of Transit Accessibility." *Journal of Urban Planning and Development* 133 (1): 32–40. https://doi.org/10.3141/1887-04.

Jordan Freitas and Luke Johnston. 2024. *opendatatoronto: Access Open Data Toronto*

Datasets. https://CRAN.R-project.org/package=opendatatoronto.

Kirill Müller and Jennifer Bryan. 2024. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Litman, Todd. 2023. "Valuing Transit Service Quality Improvements." *Journal of Public Transportation* 25 (2): 1–17. http://doi.org/10.5038/2375-0901.11.2.3.

Lohr, Sharon L. 2019. *Sampling: Design and Analysis*. 3rd ed. Boca Raton, FL: CRC Press.

Max Kuhn. 2024. *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.