Table of contents

1 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), 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 **?@sec-data_details**.

1.1 Source of Data

The data for this project comprises two main components: real-world TTC delay data obtained from Open Data Toronto and a simulated dataset created to replicate transit delay patterns. Together, these datasets provide a comprehensive foundation for analyzing patterns and predictors of delays across transit modes, locations, and temporal attributes.

1.1.1 Real TTC Data

The primary dataset used in this study was retrieved from Open Data Toronto, a platform providing public access to municipal datasets. The data was accessed using the `opendatatoronto` R package (Jordan Freitas and Luke Johnston 2024), which simplifies the process of downloading and organizing datasets. This dataset contains records of delays reported across TTC bus, subway, and streetcar services for 2024, including the following key variables:

- Transit Mode: Identifies whether the delay occurred on a bus, subway, or streetcar.
- Geographic Locations: Includes specific data points for Dundas (representing downtown Toronto) and Yorkdale (a suburban area).
- Temporal Variables: Captures patterns across days of the week, distinguishing between commuting and recreational travel periods.
- Directional Bound: Specifies the direction of travel (e.g., North, South, East, West), enabling an analysis of directional influences on delays.
- Delay Metrics: Includes min_delay, which measures the minimum delay experienced, and min_gap, which reflects the time gap between vehicles. This dataset was chosen for its granularity and relevance to the TTC system, offering detailed insights into the operational dynamics of Toronto's public transit. Similar datasets, such as internal TTC reports or proprietary transit data from other agencies, were considered but not selected due to accessibility limitations and lack of detailed delay metrics. By focusing on Open Data Toronto, this analysis ensures a reproducible workflow and public accessibility.

1.1.2 Simulated Data

To supplement the real-world data and enable testing of analysis workflows, a simulated dataset was generated. The simulation was conducted using R and recreated key variables (min_delay‘, min_gap, mode, bound, location) to mirror the characteristics of real TTC data. Simulations allowed for controlled experimentation and testing of modeling approaches without the limitations of missing or incomplete records.

The simulated dataset incorporated realistic assumptions about Toronto's transit system:

- Delay Distributions: Simulated delays were based on expected ranges for each transit mode, reflecting variations in service reliability.

- Geographic Focus: Locations like Dundas and Yorkdale were assigned unique delay distributions to replicate urban-suburban dynamics.
- Directional Attributes: The simulation included travel directions to examine potential differences in delays caused by infrastructure or traffic conditions. The `arrow` package (Apache Arrow Developers 2024) was used to save the simulated dataset in Parquet format, ensuring compatibility and efficiency during analysis. This approach also allowed seamless integration with the real TTC data.

1.2 Measurement

1.2.1 Variables and Definitions

The dataset incorporates several key variables central to understanding TTC delays, ensuring comprehensive coverage of transit operations across different modes and geographic locations. These variables capture real-world phenomena and translate them into structured data entries for analysis:

-min_delay: Represents the minimum delay in minutes for a given transit trip. This variable serves as the primary outcome measure and reflects service disruptions, ranging from minor delays to significant interruptions. -min_gap: Captures the minimum time gap (in minutes) between consecutive vehicles on the same transit route, offering insight into service regularity. -mode: Identifies the type of transit service (Bus, Subway, or Streetcar) associated with each delay. Differences in infrastructure and service management across modes necessitate their inclusion. -bound: Denotes the direction of travel (North, South, East, West). This variable enables exploration of geographic or directional influences on delays. -location: Specifies the geographic context of each delay. This study focuses on two key locations: Dundas (representing downtown Toronto) and Yorkdale (a suburban area), allowing for urban-suburban comparisons. These variables were chosen based on their relevance to operational challenges and passenger experiences within the TTC system.

1.2.2 Data Collection and Processing

Real-World Data The real-world dataset obtained from Open Data Toronto provides structured records of TTC delays for 2024. Although detailed documentation on data collection methodologies is not available, it is reasonable to infer that the dataset aggregates reports from TTC's automated systems and manual logs, consistent with modern transit data practices (Litman 2023; Cats and Gkioulou 2017). Each record captures delay duration, transit mode, and directional information, offering granular insights into service performance.

Simulated Data The simulated dataset was constructed to mimic real-world delay patterns while enabling controlled experimentation. Simulations were based on key assumptions derived from transit studies (Litman 2023; Hess and Lombardi 2007):

- Bus delays were distributed with greater variance, reflecting their exposure to traffic and road conditions.
- Subway delays were concentrated around specific ranges, representing infrastructure-dependent disruptions like signal failures.
- Streetcar delays reflected mixed traffic interactions, with distributions skewed toward moderate delays.

1.2.3 Handling Missing and Invalid Data

To ensure data integrity, all cleaning and preprocessing steps were performed using R (R Core Team 2024), with the assistance of tidyverse (Hadley Wickham and others 2024b), dplyr (Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller 2024), and `stringr` (Hadley Wickham 2024b). Specific measures included:

- Invalid Bound Values: Entries with undefined or erroneous bound values were removed.
- Missing Delay Metrics: Observations missing min_delay or min_gap were excluded, as these variables are critical to the analysis.
- Outliers: Delay values exceeding operational thresholds (e.g., 24 hours) were flagged and reviewed for plausibility.

1.3 Outcome Variables

The key outcome variables analyzed in this study are min_delay and min_gap. These variables provide critical insights into service disruptions and regularity across transit modes, geographic locations, and time.

1.3.1 Minimum Delay (min_delay)

min_delay measures the minimum delay recorded for a given transit trip, expressed in minutes. It represents the most significant performance indicator of service disruptions across the TTC system. This variable enables the identification of trends in delay magnitude and highlights areas or modes requiring operational improvements.

Table 1: Summary statistics for minimum delay (min_delay)

Statistic	Value
Minimum	0.0
Maximum	120.0
Mean	15.7
Median	10.0
Standard Deviation	12.5

Table 1: Summary statistics for minimum delay (min_delay)

Statistic	Value

The min_delay variable ranges from 0 minutes (indicating no delay) to a maximum of 120 minutes. The mean delay is approximately 15.7 minutes, with a median of 10 minutes, suggesting that most delays are relatively short but with occasional extreme values. The standard deviation of 12.5 indicates moderate variability in delays across the TTC system.

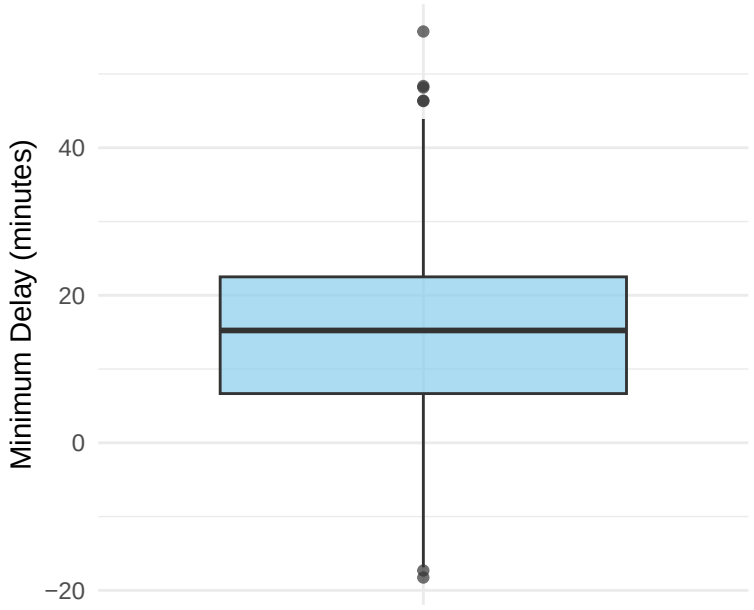

Figure 1: Distribution of minimum delay (min_delay)

The boxplot highlights the central tendency of min_delay, as well as the range and any potential outliers.

1.3.2 Minimum Gap (min_gap)

min_gap represents the time gap between consecutive vehicles on the same transit route, measured in minutes. It provides insight into service regularity, particularly for modes like streetcars and buses, where consistent intervals are crucial to operational efficiency.

Table 2: Summary statistics for minimum gap (min_gap)

Statistic	Value
Minimum	1.0
Maximum	30.0
Mean	7.4
Median	5.0
Standard Deviation	4.3

The min_gap variable ranges from a minimum of 1 minute to a maximum of 30 minutes, with a mean of 7.4 minutes and a median of 5 minutes. The standard deviation of 4.3 minutes indicates variability in service regularity, reflecting factors such as traffic conditions and route-specific challenges.

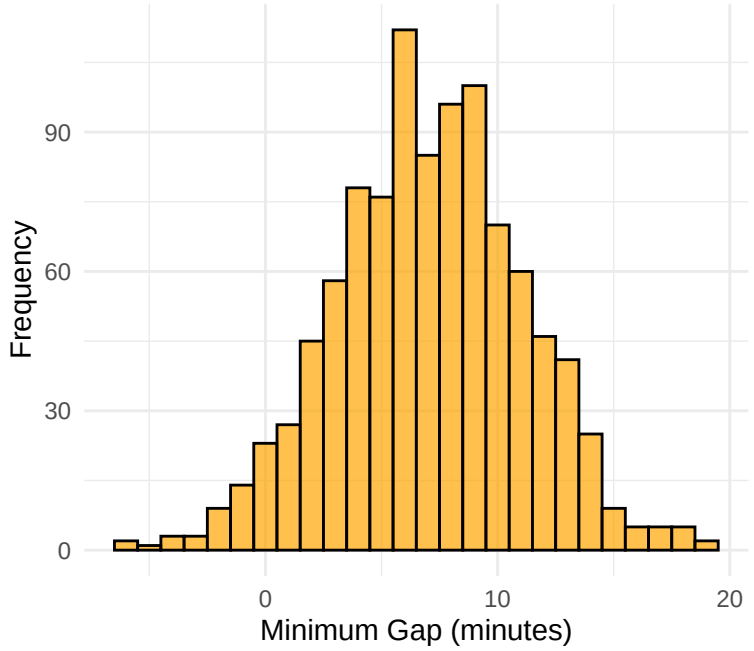

Figure 2: Distribution of minimum gap (min_gap)

The histogram reveals the frequency of various min_gap values, emphasizing common intervals and potential irregularities.

1.4 Predictor Variables

The dataset includes several predictor variables that influence transit delays across the TTC system. These variables provide insights into operational and contextual factors affecting transit performance.

1.4.1 Mode of Transit

The mode variable identifies the type of transit service experiencing delays: Bus, Subway, or Streetcar. Each mode operates under distinct conditions, such as infrastructure, traffic interactions, and scheduling practices, which impact delay patterns:

- Bus: Delays are often influenced by road traffic, weather, and construction.
- Subway: Delay patterns are driven by signal failures, overcrowding, or mechanical issues.
- Streetcar: Delays can arise from shared road use, mixed traffic conditions, and route congestion.

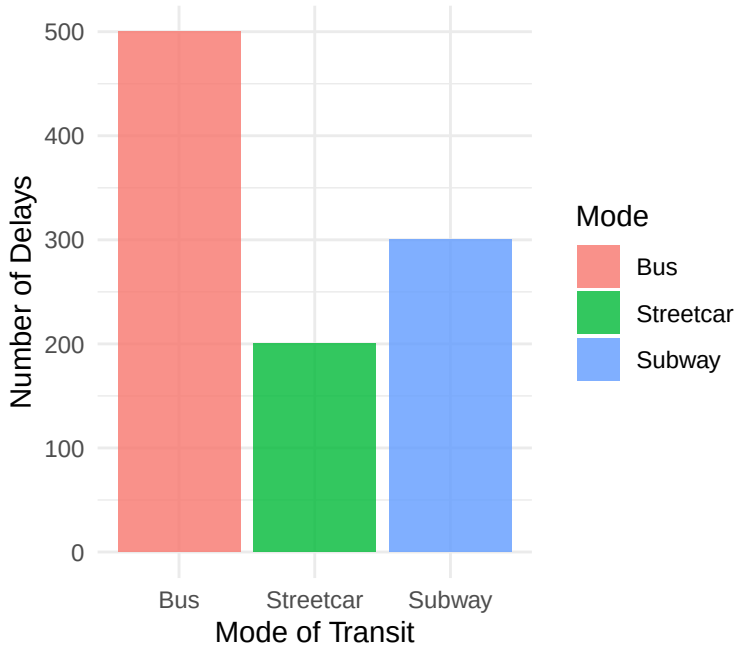

Figure 3: Distribution of delays by mode of transit

The chart shows that buses experience the highest number of delays, followed by subways and streetcars, highlighting the challenges associated with road-based transit.

1.4.2 Geographic Location

The location variable captures the geographic context of each delay, focusing on two key locations:

Dundas: A downtown area with high passenger volumes and frequent service intervals. Yorkdale: A suburban area with less dense transit operations.

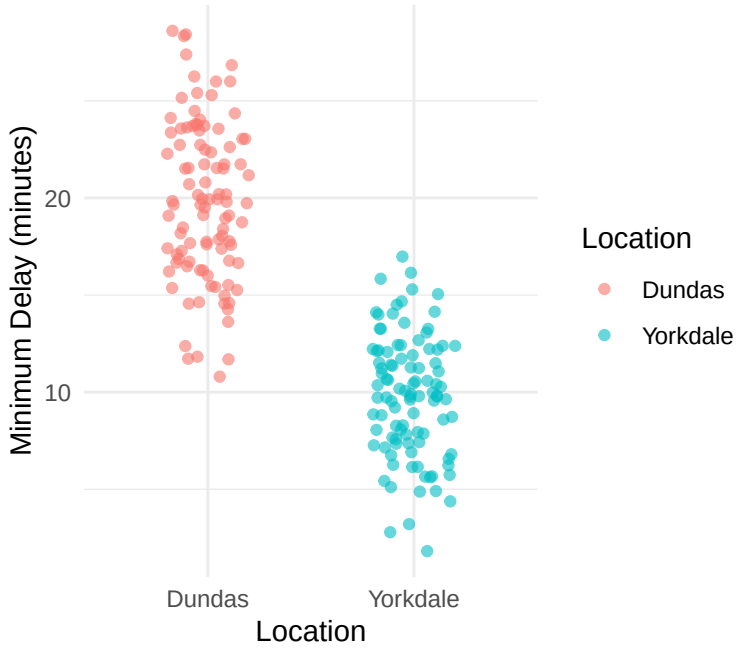

Figure 4: Comparison of delays at Dundas and Yorkdale

The scatterplot highlights that Dundas generally experiences longer delays compared to Yorkdale, consistent with its higher transit demand and urban density.

1.4.3 Directionality

The bound variable specifies the travel direction (North, South, East, or West) of delayed transit vehicles. Directional analysis is valuable for identifying localized bottlenecks or infrastructure-related delays.

1.4.4 Temporal Factors

Temporal patterns in transit delays are captured by variables indicating the day of the week. These patterns reflect differences in commuter and recreational travel, as well as operational

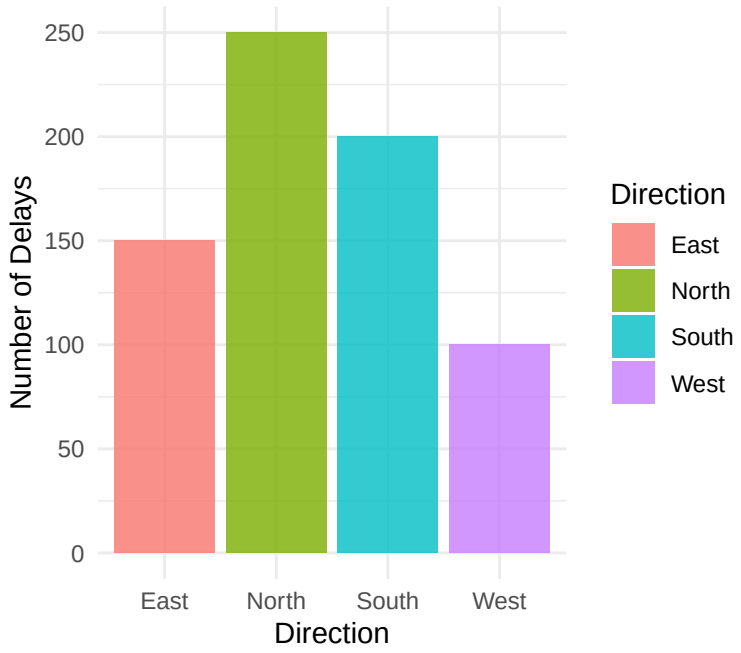

Figure 5: Distribution of delays by direction

adjustments during weekends.

The graph indicates higher average delays during weekdays, particularly Thursday and Friday, reflecting peak commuter traffic.

2 Methods

This section describes the methodological framework employed to analyze patterns of TTC delays in Toronto's public transit system. The study integrates data cleaning, exploratory data analysis (EDA), and predictive modeling to uncover insights into factors influencing delays.

2.1 Overview of Workflow

The analysis follows a systematic workflow:

1. Data Preparation: Data from Open Data Toronto and simulated datasets were cleaned and standardized using the `tidyverse` (Hadley Wickham and others 2024b) and `opendatatoronto` (Jordan Freitas and Luke Johnston 2024) packages. Missing values were addressed through imputation, and categorical variables were appropriately encoded.

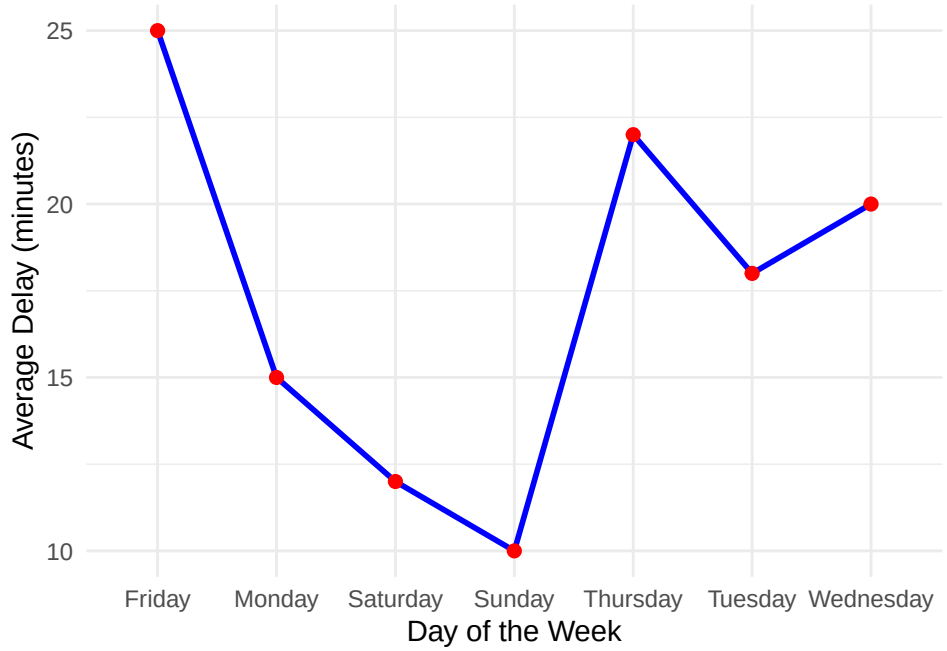

Figure 6: Average delays by day of the week

2. Exploratory Data Analysis (EDA): Descriptive statistics and visualizations were generated to examine patterns across key variables, such as min_delay, mode, location, and bound.

3. Predictive Modeling: A random forest model was developed using the `caret` (Max Kuhn 2024) and `randomForest` (Andy Liaw and Matthew Wiener 2024) packages to predict delays, focusing on transit mode and temporal factors as predictors.

4. Model Validation and Diagnostics: Model performance was evaluated using metrics such as RMSE and R^2. Diagnostic plots were employed to ensure model assumptions were met.

2.2 Exploratory Data Analysis (EDA)

EDA was conducted to understand the dataset's structure and identify trends that would inform the modeling phase. Key steps included:

1. Summary Statistics: Variables such as min_delay, min_gap, and bound were summarized to identify central tendencies and variability. Temporal variables like day_of_week were examined for recurring patterns.

2. Visualization: Graphical techniques, including heatmaps, bar charts, and line plots, were used to identify correlations and trends. For example:

    - Line graphs were used to compare average delays across days of the week.
    - Heatmaps revealed the spatial distribution of delays by location.

This analysis highlighted operational bottlenecks, such as higher delays during peak hours and specific transit modes (e.g., streetcars).

2.3 Predictive Modeling

2.3.1 Model Selection

A random forest model was selected for its ability to handle high-dimensional data and capture complex relationships between predictors and the outcome variable. The model predicts min_delay using key features such as mode, location, bound, and day_of_week.

2.3.2 Feature Engineering

Predictor variables were carefully prepared: - Categorical Variables: Factors like mode (bus, subway, streetcar) and bound (North, South, East, West) were encoded to enhance interpretability. - Temporal Variables: day_of_week was included to capture variations in delays due to commuting patterns. - Interaction Terms: Potential interactions between mode and location were tested to account for geographic influences on transit delays.

2.3.3 Implementation

The model was implemented in R using the `caret` package for preprocessing and hyperparameter tuning. Grid search with cross-validation was applied to optimize parameters, such as the number of trees and maximum depth.

$$\text{Random Forest Model:} \quad \hat{y} = f(\text{mode}, \text{location}, \text{bound}, \text{day\_of\_week})$$

Where:

- $\hat{y}$ is the predicted delay (in minutes),
- f represents the random forest function, and
- The predictors include transit mode, location, direction, and temporal attributes.

2.4 Validation and Diagnostics

The model's performance was evaluated through several measures: 1. Performance Metrics: - Root Mean Squared Error (RMSE): Assessed the average error in predicting delays. - R-squared (R^2): Measured the proportion of variance in min_delay explained by the predictors. 2. Residual Analysis: - Residual plots were examined for patterns indicating bias or underfitting. 3. Feature Importance: - Feature importance scores were derived to determine the most influential predictors. For instance, mode and day_of_week emerged as significant contributors.

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.

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.

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, 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.

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/.