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` (**citeDplyr?**), `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.

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.

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.

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.

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/.