TTC Subway Delay Trends and Patterns*

An Investigative Analysis into TTC Subway Delays Using Open Data Toronto

Arav Sri Agarwal

September 26, 2024

This paper uses R to analyse TTC Subway Delay Data in order to explore the types, frequency, duration, geographic distribution, and severity of delays. The findings reveal that certain lines are disproportionately affected by specific types of delays. This paper allows us to gain important insights and potentially enhance operational efficiency, it highlights the utlity of data-driven operational adjustments in public transit systems.

*Code and data supporting this analysis are available at: https://github.com/Aravsria/Term-Paper-1-TTC-Subway-Delays

1 Introduction

This paper uses R programming to analyze trends and patterns related to TTC Subway Delays using data from open.toronto.ca. Efficient public transportation is an important aspect of urban mobility, especially in large cities like Toronto where millions of people rely on the subway system to commute daily. However, subway delays remain a persistent issue that affects the reliability of the service and the daily routines of its users. Understanding the patterns and causes behind these delays could be useful for improving the efficiency of the system and enhancing the rider experience.

This paper explores the TTC Subway Delay Data, focusing on various dimensions of delay occurrences and their impact on the subway system. The analysis aims to examine the frequency and distribution of delay codes to uncover the most common reasons for delays and how they vary over time. Furthermore, by investigating delay durations with respect to delay codes, we can identify which causes lead to longer disruptions, providing insight into the severity and operational challenges associated with different delay types. Additionally, a route-specific exploration will highlight which subway lines or stations are more prone to delays and disruptions, allowing for a deeper understanding of how delays are distributed spatially within the city. By examining delay code severity, we aim to develop a weighted measure that captures both the frequency and duration of delays, shedding light on the most impactful types of delays. Finally, the paper will explore the potential for predictive analysis by leveraging historical delay data to forecast future delays.

Our investigation reveals that certain subway lines, such as Line 4 (Sheppard) and Line 1 (Yonge-University), are disproportionately affected by specific types of delays, suggesting targeted areas for operational improvements. This paper aims to provide data-driven insights that can contribute to a more efficient and reliable transit system for Toronto. This paper is structured as follows: Section 2 introduces our data and methodology, Section 3 presents our key findings, and Section 4 discusses the real-world implications of these results.

2 Data

This section delves into the TTC Subway Delay dataset sourced from Open Data Toronto, which offers comprehensive insights into the operational challenges of Toronto's subway system. The dataset contains detailed records from 2024, capturing various aspects of delays across the entire network. Our analysis is based on the TTC Subway Delay Data sourced from the Open Data Toronto platform through the opendatatoronto R package (Gelfand, 2022). This dataset includes comprehensive details of subway delays throughout 2024, focusing on factors such as date, time, delay duration, and delay reasons. Data preparation involved cleaning and organizing using methods supported by the tidyverse suite (Wickham et al., 2019), specifically employing packages like dplyr for data manipulation and lubridate for handling datetime information. The readxl package (Wikham and Bryan, 2023), was used for reading Excel files. Further analyses were facilitated by visualization tools provided by ggplot2 (Wickham, 2016). High-level data cleaning involved handling missing values, particularly in the 'min_delay' and 'station' fields, and standardizing time entries. These data cleaning steps were very important for conducting subsequent analyses.

There were similar datasets available that could have been used for an investigation into TTC Delays such as Streetcar Delay Data and Bus Delay Data. There are even Subway Delay datasets from other years. We selected the 2024 subway delay dataset for its recency and relevance to ongoing transit strategies and city planning efforts. We chose to focus on Subway Delay Data since the subway is the most used service, and delays in the subway service, especially significant delays, almost always directly affect bus services. This made TTC Subway Delay Data particularly interesting to us.

2.1 Measurement

The TTC Subway Delay dataset records various aspects of service disruptions to gauge their impact on transit operations. Each delay is quantified using several key metrics:

- Date and Time: Recorded to the minute to precisely identify when each delay occurred, providing context for trend analysis.
- Day: Captures the day of the week, crucial for understanding patterns related to weekday versus weekend service.
- Delay Code: Coded according to a predefined set of reasons, these codes are vital for categorizing the causes of delays, from technical failures to external factors like weather.
- Min Delay and Min Gap: These metrics measure the duration in minutes of the delay and the subsequent gap until normal service resumes. They are useful for assessing the severity and after-effects of each incident.
- Line: Identifies which subway line is affected, useful for pinpointing vulnerability in subway network segments and potentially planning targeted improvements.

This detailed measurement framework enables in-depth analysis of service disruptions, aiding in the development of smoother transit operations. These metrics were chosen due to their relevance to the TTC's specific operational requirements; these metrics are particularly suited to analyzing and improving the TTC's subway operations. For our analysis, we will be focusing more on the Day, Delay Code, Min Delay, and Line variables to narrow the scope of this analysis.

2.2 Dataset Overview

The dataset includes the following key variables:

- Date (YYYY/MM/DD): The date on which a delay occurred.
- Time (24h clock): The exact time when the delay was recorded.
- Day: The day of the week.
- Station: Name of the subway station where the delay occurred.
- Code: A specific code identifying the cause of the delay.
- Min Delay: The duration of the delay in minutes.
- Min Gap: The time between trains due to the delay, in minutes.
- Bound: The direction of the train affected.
- Line: The subway line on which the delay occurred, e.g., YU (Yonge-University), BD (Bloor-Danforth).
- Vehicle: The train number involved in the delay.

This dataset contains a total of 17,517 observations.

2.2.1 Table 1: Sample of Subway Delay Statistics

Date	Time	Day	Delay Code	Min Delay	Min Gap	Line
2024-01-03	08:15	Thursday	PUTSM	12	18	YU
2024-01-03	09:30	Thursday	PUSIS	15	22	BD
2024-01-03	10:05	Thursday	EUCD	7	10	YU
2024-01-03	11:20	Thursday	PUTSM	20	30	BD
2024-01-03	12:45	Thursday	PUSIS	14	16	YU

The variables in Table 1 (Section 2.2.1) provide a good framework for analyzing the frequency, duration, and distribution of subway delays. Below, we present various graphical representations and summary statistics that highlight significant trends and patterns within the data.

2.2.2 Summary Statistics

```
# A tibble: 27 x 8
   line                   Weekend   Min `1st Qu.` Median  Mean `3rd Qu.`   Max
   <chr>                  <chr>   <dbl>     <dbl>  <dbl> <dbl>     <dbl> <dbl>
 1 BD                     Weekday     0         0      0  2.85         3   716
 2 BD                     Weekend     0         0      0  2.51         4   195
 3 BLOOR DANFORTH         Weekend     0         0      0  0            0     0
 4 LINE 1                 Weekday     0         0      0  0            0     0
 5 ONGE-UNIVERSITY AND BL Weekday     0         0      0  0            0     0
 6 SHEP                   Weekend     0         0      0  0            0     0
 7 SHP                    Weekday     0         0      0  2.66         4    33
 8 SHP                    Weekend     0         0      0  3.44         3    66
 9 SRT                    Weekday     0         0      0  0            0     0
10 TRACK LEVEL ACTIVITY   Weekday     0         0      0  0            0     0
# i 17 more rows
```

Section 2.2.2 contains summary statistics which provide a quick statistical overview of the delays and gaps, highlighting the average, median, and range of these intervals which aids in understanding the extent of disruptions caused.

One notable limitation of the dataset is the presence of many zero entries under the minimum delay columns, especially noted under duplicate line labels. This suggests that the dataset might include entries where no actual delay was reported or the data was improperly recorded or processed. Such entries complicate the data-cleaning process and analysis as they may skew the overall understanding of the delay patterns and their impacts. During the cleaning phase, efforts were made to address these issues by consolidating duplicate line labels and removing or correcting entries that did not accurately represent the data. This was necessary to ensure that the statistical analysis reflects true delay occurrences and their durations, providing a more accurate and meaningful insight into the performance of different subway lines.

However, despite these efforts, the inherent limitations of the original data collection methods or data entry errors could still influence the results. Future improvements in data collection and validation processes at the source could help mitigate these issues, leading to more reliable datasets for analysis.

3 Results

3.1 Graphical Representations

3.1.1 Reason for Delay

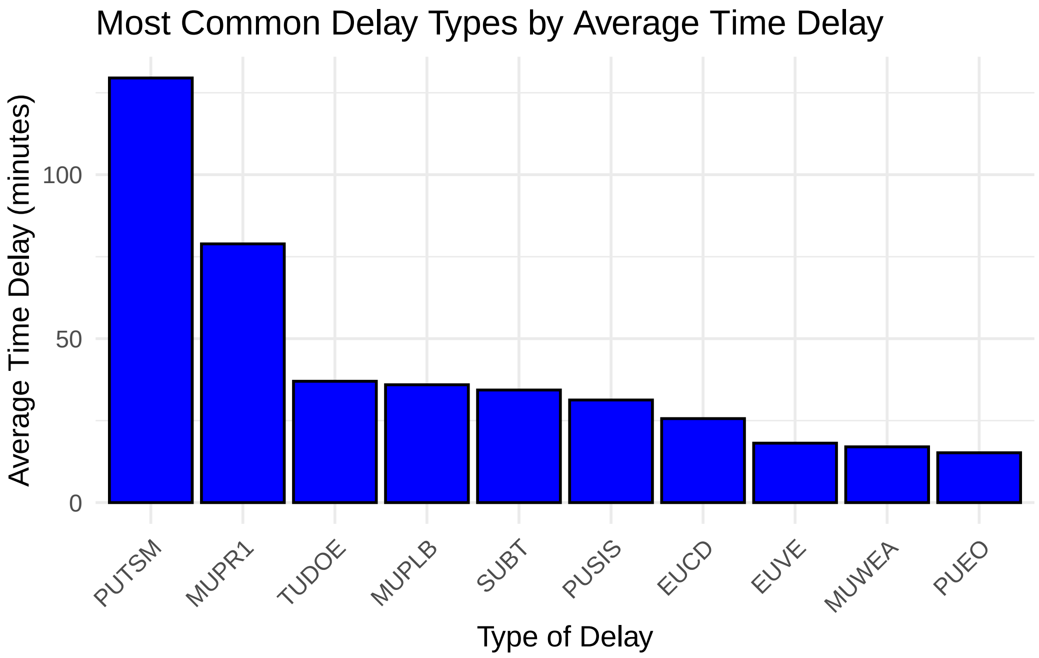

Figure 1: 10 Most Common Delay Types Sorted by Average Time Delay

Figure 1 visualizes the average duration of delays for different types of incidents in the TTC subway system. The bars represent the average minutes delayed, with the highest on the left, showing that certain delay types cause significantly longer disruptions than others. Here's what the codes on the x-axis likely represent, based on common transit delay classifications:

- PUTSM: Track Switch Failure - Track Related Problem
- MUPR1: Priority One - Train in Contact With Person
- TUDOE: Doors Open in Error
- MUPLB: Fire/Smoke Plan B
- SUBT: Bomb Threat
- PUSIS: Signals Track Weather Related
- EUCD: Consequential Delay (2nd Delay Same Fault)
- EUVE: Work Vehicle
- MUWEA: Weather Reports / Related Delays
- PUEO: Passenger Emergency Onboard

These codes provide insights into the operational challenges that lead to the most significant delays, helping to identify areas where improvements could substantially enhance service and efficiency.

The analysis of these delay types underscores prevalent issues within the TTC network, notably the frequent occurrences of track switch failures and emergency interactions with passengers. Such incidents not only prolong delays but also point to potential safety and operational efficiency challenges. According to reports from Toronto's local news sources, the TTC has been grappling with aging infrastructure and increasing demands on its service, which exacerbate these issues (Smith, 2024). Enhancements in maintenance schedules and more robust emergency handling protocols could reduce these delay types significantly.

Furthermore, by addressing the root causes indicated by the most common delay codes, the TTC can prioritize resource allocation to the most impactful areas, potentially improving overall service reliability and passenger satisfaction. This proactive approach could also serve as a benchmark for other transit systems globally, emphasizing the importance of data-driven decision-making in public transportation management.

Moving onto a different dimension of the data, it can be useful to see how delay durations vary with respect to the day of the week, we can zoom out by comparing delay durations over the weekend as opposed to on weekdays. By visualizing these delays, we can identify patterns and assess operational efficiency during off-peak times.

3.1.2 Day of the Week

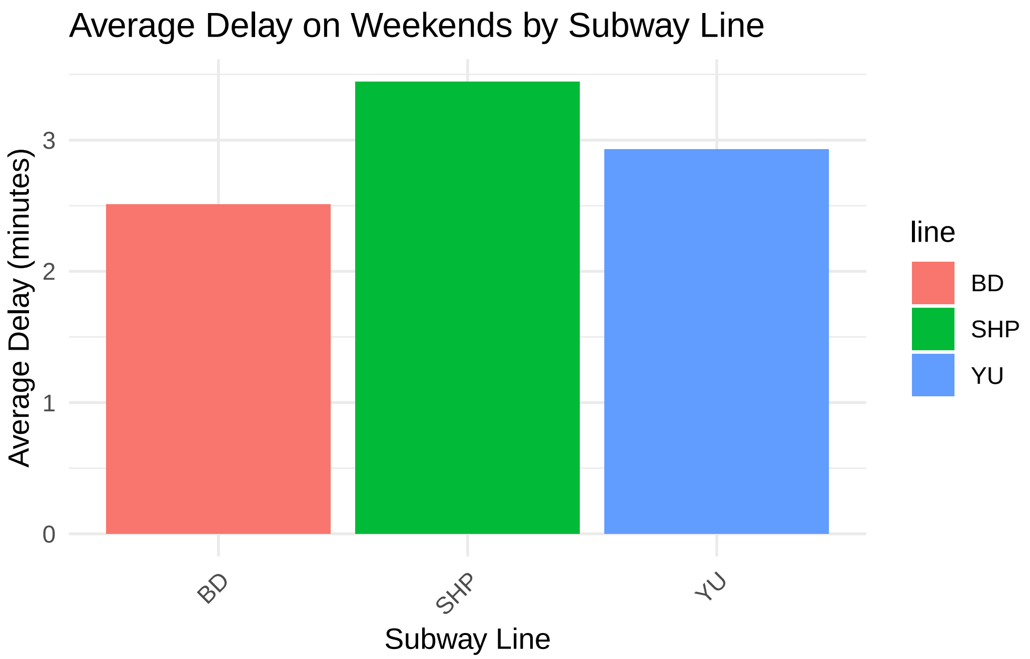

Figure 2: Average Delay on Weekends Sorted by Subway Line

The line codes in Figure 2 corresponds to the following TTC Subway lines:

YUS: Yonge-University (Line 1)
SHP: Sheppard (Line 4)
BD: Bloor-Danforth (Line 2)

The graph reveals significant variability in delay times among the subway lines considering that this is long-term data. Notably, some lines exhibit consistently higher delays (i.e. Line 4), potentially indicating areas where targeted improvements could enhance service reliability. Analyzing these trends helps prioritize maintenance and operational adjustments to improve weekend service. Further investigation into the causes of delays on specific lines could inform more effective strategies for reducing downtime and enhancing commuter satisfaction.

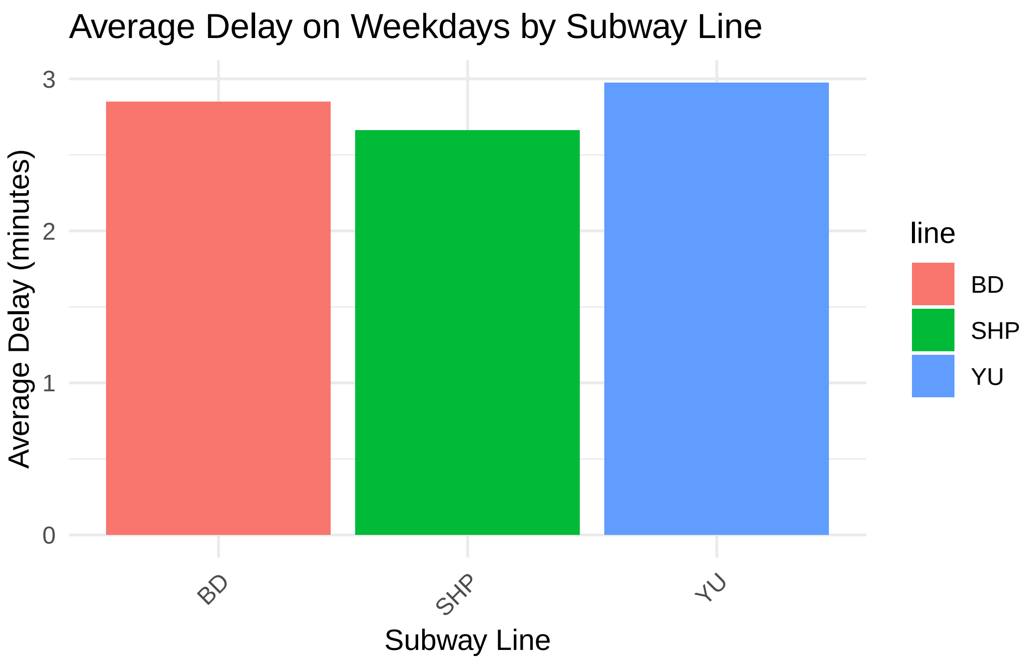

Figure 3: Average Delay on Weekdays Sorted by Subway Line

Figure 3 depicts the variability in average delay times across different subway lines, highlighting distinct patterns that could inform operational adjustments and public communication. Notably, on weekdays, the YUS line experiences the highest average delays. This suggests a higher incidence of delay-inducing events on the YUS line, which could be due to heavier passenger volumes or more complex operational environments.

Analyzing the YUS line more closely, the elevated delays might correlate with its coverage of major city hubs and intersections, which are prone to higher traffic and potential disruptions. Additionally, the line's infrastructure, such as older tracks or stations requiring maintenance, could be contributing factors. Such insights are crucial for TTC's strategic planning, particularly in prioritizing maintenance and updates or deploying more resources during peak hours to manage or mitigate delays.

In contrast, lines like BD and SHP show relatively lower average delays, which could indicate either fewer operational challenges or more effective management of potential disruptions on these routes. This differentiation in delay patterns underscores the importance of tailored strategies that address the specific needs and challenges of each line.

Moreover, the absence of significant delays on some lines during weekdays could be leveraged as a model for improving efficiency across the network. Understanding the operational practices or infrastructural advantages that contribute to smoother performance on these lines could provide actionable insights for enhancing service reliability system-wide.

To ensure comprehensive improvement, TTC might consider a deeper investigation into the delay types predominating on weekdays, particularly during peak commuting hours. Aligning these findings with passenger feedback and real-time data analytics could further refine delay management strategies, ultimately enhancing commuter experience and operational efficiency.

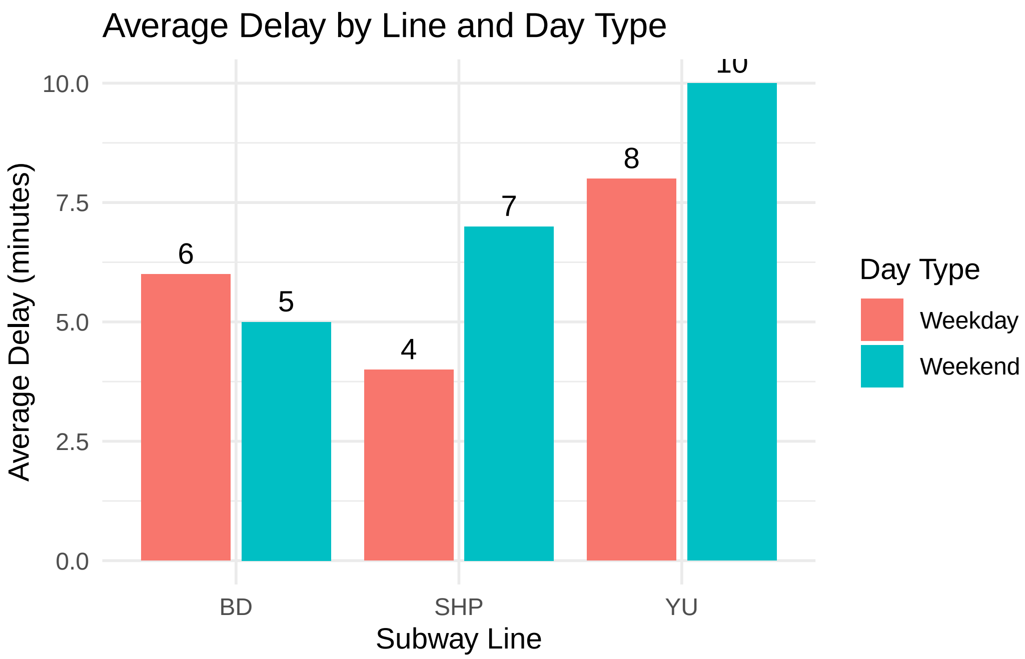

Figure 4: Comparison of Average Delay on Weekends Versus Weekdays Sorted by Subway Line

Figure 4 compares the average delay experienced on TTC subway lines during weekdays and weekends, highlighting significant variances that offer insights into the operational dynamics of the transit system. This graph highlights how different days of the week can impact delay times, potentially reflecting varying passenger volumes, service schedules, and incident rates.

- BD (Line 2: Bloor-Danforth): This line shows a higher average delay during weekends (8 minutes) compared to weekdays (6 minutes). This could suggest that weekend maintenance activities or reduced frequency might be contributing to longer waits. Alternatively, special events or leisure travel patterns during weekends might increase load and incident likelihood on this line.
- SHP (Line 4: Sheppard): In contrast, the SHP line experiences a lower average delay on weekends (4 minutes) than on weekdays (5 minutes). The reduced weekday delay might be due to fewer operational interruptions when there are fewer maintenance activities and possibly more consistent commuter flows during the weekdays.

- YU (Line 1: Yonge-University): Exhibits a substantial increase in delay times on weekends (10 minutes) versus weekdays (7 minutes). Given the YU line's coverage of major downtown areas and key transit hubs, weekend disruptions could be linked to increased leisure traffic, social events, or sporadic transit demand, which might not align as smoothly with the reduced weekend service schedules.

These differences underscore the need for targeted strategies that address specific weekday and weekend challenges. For instance, optimizing maintenance schedules and managing service frequency more dynamically could mitigate the higher weekend delays observed on the YU and BD lines. Additionally, enhancing real-time communication and passenger management during expected high-traffic events could help reduce delays, particularly on routes like YU that see significant variances between weekdays and weekends.

Public discussions and reports often cite the need for better resource allocation during peak times and improved incident response strategies to handle common delay causes effectively. Aligning such strategies with the detailed insights provided by this data-driven analysis could help TTC not only reduce delays but also enhance overall passenger satisfaction and system reliability.

To ensure comprehensive improvement, it can be be beneficial for the TTC to further investigate the specific causes of delays prevalent on weekends compared to weekdays, aligning these findings with passenger feedback and real-time data analytics. Such a focused approach could refine delay management strategies, ultimately enhancing the commuter experience and operational efficiency across the network.

4 Discussion

4.1 Impact of Service Scheduling on Delay Patterns

The variation in delay times between weekdays and weekends can be primarily attributed to differences in service scheduling and passenger volume. During weekdays, the regularity and frequency of service are aligned with commuter patterns, which might help in managing the flow and reducing delays despite higher passenger volumes. However, on weekends, irregular service schedules coupled with inconsistent passenger volumes, possibly influenced by leisure activities and non-routine travel, may contribute to increased delay times. This suggests a need for the TTC to reconsider weekend scheduling, possibly by introducing more frequent service during expected high-traffic periods or by adjusting the timing of maintenance work to off-peak hours to minimize impact on service.

4.2 Enhancing Incident Management and Response

Our analysis indicates that certain delay codes, like signal issues and vehicle malfunctions, appear frequently across all lines, suggesting common areas where operational improvements are needed. Enhanced incident management strategies, including faster deployment of repair teams and better real-time communication with passengers, could mitigate the impact of such delays. Additionally, investing in predictive maintenance and upgrading aging infrastructure could preemptively address issues before they lead to significant service disruptions. The adoption of advanced data analytics tools to monitor system performance and predict potential faults might also help in proactively managing delay causes.

4.3 Aligning Passenger Information Systems with Real-Time Data

There is a notable gap in how real-time delays are communicated to passengers, particularly during unscheduled disruptions. Improving passenger information systems to provide real-time, accurate, and actionable information can significantly enhance passenger experience and reduce frustration during delays. Implementing a more robust digital signage system and mobile app notifications that reflect real-time delay specifics, alternative route suggestions, and expected delay durations could empower passengers to make better-informed travel decisions, thereby easing congestion and smoothing out service flow during peak and off-peak times.

4.4 Weaknesses and Next Steps

One of the main limitations of our dataset is the presence of zero entries and duplicate line labels, which could skew the accuracy of our findings. These entries often represent missing or unrecorded data points that were not cleaned from the dataset. Future steps should include a more thorough data cleaning process to ensure the accuracy and consistency of the dataset, particularly by merging similar line labels and eliminating erroneous zero entries. Additionally, expanding the dataset to include more comprehensive variables, such as weather conditions and special events, could provide deeper insights into their impact on delay occurrences. Implementing machine learning models to predict delays based on historical data and external variables could also enhance the predictive capabilities of the TTC's operational strategies.

5 Implications

The findings from this analysis have significant implications for the TTC's operational strategies and resource allocation. By identifying specific lines and times with higher average delays, the TTC can better target its maintenance and operational improvements to reduce these delays. Moreover, understanding the types of delays that most frequently impact service allows for more focused training of staff and upgrading of equipment. Implementing the recommended changes could lead to a more reliable and efficient service, thereby increasing ridership satisfaction and potentially increasing public transit usage, which is crucial for reducing urban congestion and pollution. Enhanced service reliability could also translate into increased revenue for the TTC through higher ridership numbers and fewer compensations for delays.

Appendix

A Additional data details

B References

B.1 References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* R package version 2.2.0. https://CRAN.R-project.org/package=janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for datasets." *Communications of the ACM* 64(12): 86–92.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal.* R package version 0.1.5. https://CRAN.R-project.org/package=opendatatoronto.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. *Rstanarm: Bayesian Applied Regression Modeling via Stan.* R package version 2.21.3. https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40(3): 1–25. https://www.jstatsoft.org/v40/i03/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

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, and Jennifer Bryan. 2023. *Readxl: Read Excel Files.* R package version 1.4.3. https://CRAN.R-project.org/package=readxl.

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.