A Modest Proposal for Developing Better Experience with TTC Subway*

Terry Tu

January 25, 2024

This study analyzes the 2023 TTC subway delay data from OpenDataToronto to uncover delay patterns across different days and lines. Various tests are conducted to analyze the patterns in delay durations and frequencies. The analysis highlights the need for targeted strategies to improve subway efficiency, particularly on the most affected lines. The aim of this study is to lead to better subway service, making everyday travel smoother for many people in Toronto.

Table of contents

*Code and data supporting this analysis are available at: https://github.com/TEJMaster/TTC-Subway-Delay-Analysis-2023.git

1 Introduction

Public transportation, like the TTC (Toronto Transit Commission), is essential for city life. In Toronto, many people rely on the TTC every day. However, it's not always smooth sailing. A survey in May 2019 showed mixed feelings about the TTC: about two-thirds of riders are happy with things like how clean it is and the cost, but nearly half of the people are often late because of unexpected delays, especially during busy hours (Vuong 2019). These delays can really disrupt daily routines and make people wonder how well the TTC is being run.

This project digs into the TTC subway delay data for 2023 from OpenDataToronto using a data set called "subway_delay_data_2023.csv". We want to find patterns in the delays to help understand and fix the problems. This is especially important as more and more people are living in Toronto, and they all need a reliable subway system. Even with the issues, a good number of users, 77%, still think the TTC is dependable, which shows that the system has a strong base to improve from (Vuong 2019).

Here is a brief summary of steps I used to conduct this study. First, I downloaded the data from OpenDataToronto (Gelfand 2022). Then, I cleaned up the data to make sure everything was accurate and ready for analysis. After the data cleaning, I used R (R Core Team 2022), which is a tool for doing statistics, to analyze the data. I looked at which subway lines have the most delays, and I checked to see if there were certain days or times when delays happened more often. This is useful to understand the patterns and figure out where the TTC might need to make some changes.

2 Data

Data for this study were meticulously curated from the Open Data Toronto Portal via the opendatatoronto package (Gelfand 2022). The primary dataset utilized is the TTC subway delay data for 2023. This dataset offers a granular view into each recorded delay within the TTC subway system, encapsulating critical attributes such as the date, time, duration, and affected subway line.

2.1 TTC Subway Delay Data

This dataset, provided by the Toronto Transit Commission via the Open Data Toronto Portal, captures comprehensive information about delays occurring within the TTC subway network throughout 2023. As of the data retrieval date, the dataset includes detailed records of each delay event, structured with several key fields to offer insights into the nature and impact of these delays. Initial inspection of the data revealed fields such as 'Date', 'Time', 'Station', 'Line', 'Min Delay', and 'Reason for Delay', among others. However, it was noted that not all records were complete, necessitating a rigorous data cleaning process.

2.2 Data Cleaning and Initial Observations

The raw data was initially fetched and then subjected to a systematic cleaning process to ensure data integrity for analysis (Please refer to appendix for details). After-cleaning, an initial exploration of the data was conducted below, providing some insight about how the dataset is looked like.

Table 1: Sample of the Cleaned TTC Subway Delay Data

Date	Time	Day	Station	Code	Min Delay	Min Gap	Bound	Line	Vehicle
2023-01-01	02:22:00	Sunday	MUSEUM STATION	MUPAA	3	9	S	YU	5931
2023-01-01	02:30:00	Sunday	KIPLING STATION	MUIS	0	0	E	BD	5341
2023-01-01	02:33:00	Sunday	WARDEN STATION	SUO	0	0	W	BD	0
2023-01-01	03:17:00	Sunday	KEELE STATION	MUIS	0	0	NA	BD	0
2023-01-01	07:16:00	Sunday	BATHURST STATION	MUIS	0	0	NA	BD	0
2023-01-01	07:44:00	Sunday	JANE STATION	MUNCA	0	0	NA	BD	0

Table 1 presents a snippet of the cleaned dataset, showcasing the first few rows after data cleaning. The table provides a glimpse into the structured format of the data, ready for in-depth analysis.

2.3 Discription for variable used in this study

- Date: The date on which the delay occurred, providing a chronological context to the incident.
- Time: The exact time at which the delay was recorded, which is crucial for identifying peak delay periods throughout the day.
- Day: The day of the week, offering insights into how delays might fluctuate on weekdays versus weekends.
- Min Delay: The reported duration of the delay in minutes, reflecting the severity of the incident.
- Line: The subway line on which the delay occurred, essential for recognizing which lines are most frequent to delays.

3 Data Analysis & Results

The data analysis was performed using R (R Core Team 2022), a powerful open-source statistical programming language. Key packages from the tidyverse collection (Wickham et al. 2019) were employed to streamline data manipulation, visualization, and analysis processes. These packages include ggplot2 (Wickham 2016) for creating advanced graphics, dplyr (Wickham et al. 2022) for data manipulation, readr (Wickham, Hester, and Bryan 2022) for its robust data reading functionalities, lubridate (Grolemund and Wickham 2011) for handling date-time data, and knitr (Xie 2014) for dynamic report generation.

3.1 Analysis on subway delay based on day of the week

From Figure 1, we can clearly see that weekdays experience significantly longer cumulative delay duration as compared to weekends. Notably, Monday stands out with the highest aggregate minutes of delay, suggesting a peak in delay occurrences at the start of the workweek. Conversely, Saturday is the day with the least total delay time, indicating a smoother operational flow during weekend services.

From Figure 2 the frequency analysis of delays corroborates the trend observed in total delay duration. Weekdays generally have a higher incidence of delays, with Friday leading in the number of reported delays. On the other hand, during Sunday the plot shows the least frequency of delays, aligning with the expected reduced demand for travel on the last day of the weekend.

These insights could serve as a cornerstone for targeted strategies aimed at improving service efficiency, such as enhanced resource allocation during identified peak times. Commuters might also benefit from planning their travel schedules around these insights to avoid potential delays.

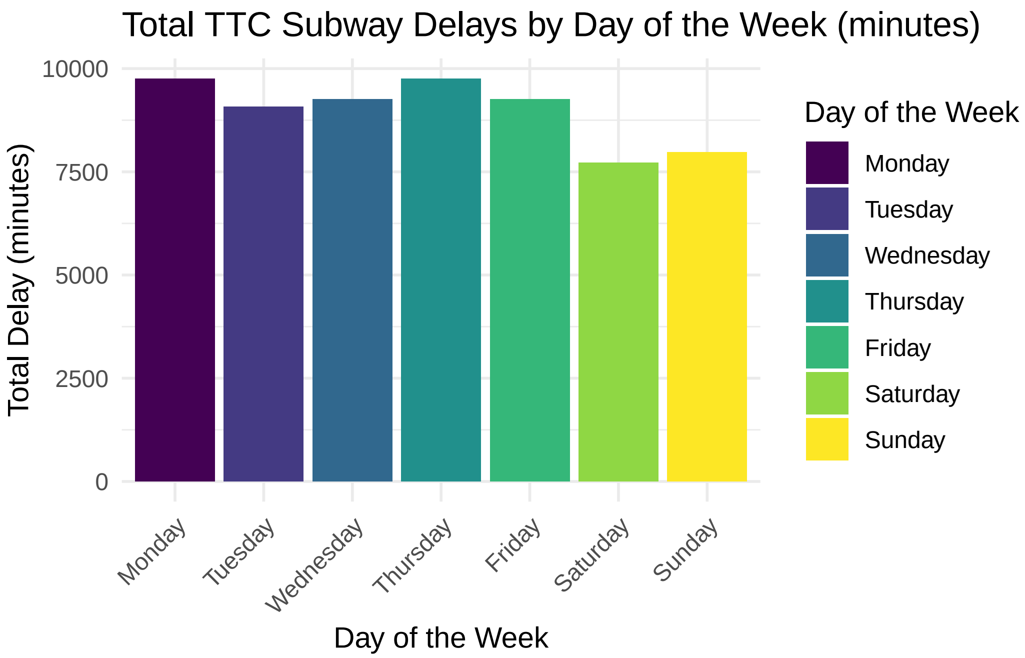

Figure 1: Total minutes of subway delays categorized by days of the week

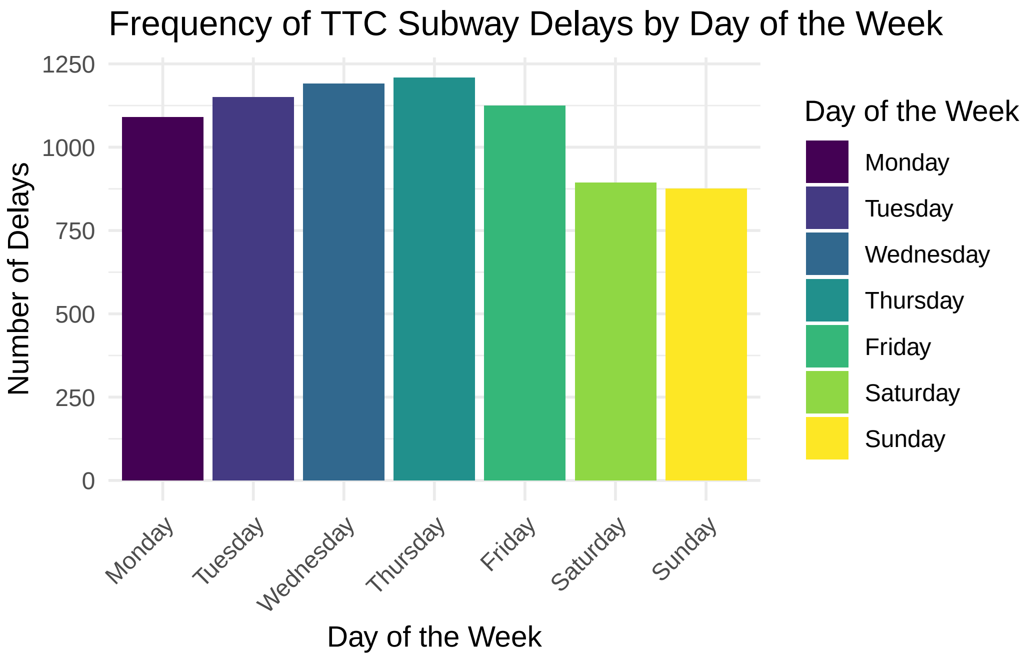

Figure 2: Frequency of subway delays categorized by days of the week

3.2 Analysis on subway delay based subway line

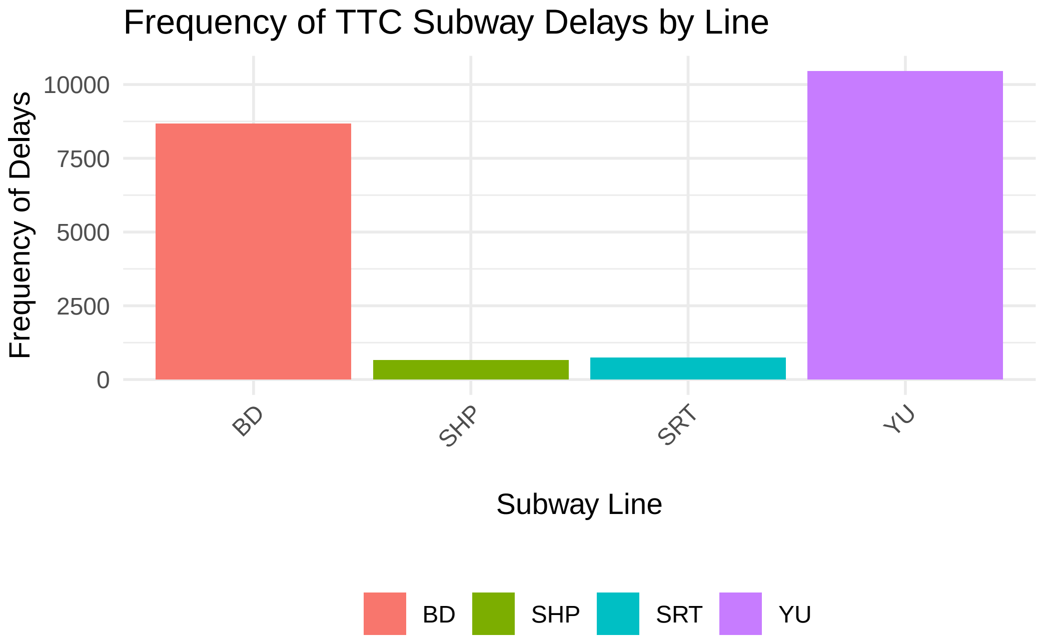

Figure 3: Frequency of subway delays categorized by subway line

Figure 3, indicates a substantial contrast in delay frequencies across different subway lines for the year 2023. Lines labeled 'SHP' (Line 4) and 'SRT' (Line 3) are notable for having much less delays, possibly reflecting efficient operations. Conversely, lines such as 'BD' (line 2) and 'YU' (line 1) report a high frequency of delays. This observed pattern calls for targeted operational focus. As a manager of TTC's subway system, these findings warrant a comprehensive analysis to understand the underlying causes of these delays. Furthermore, commuters who frequent these lines may wish to factor in additional travel time in anticipation of potential delays.

3.2.1 List of TTC Subway Line Codes and Their Corresponding Full Names:

- **YU**: Yonge-University Line (Line 1)
- **BD**: Bloor-Danforth Line (Line 2)
- **SRT**: Scarborough RT Line (Line 3)
- **SHP**: Sheppard Line (Line 4)

3.3 Analysis on subway delay based on time of the day

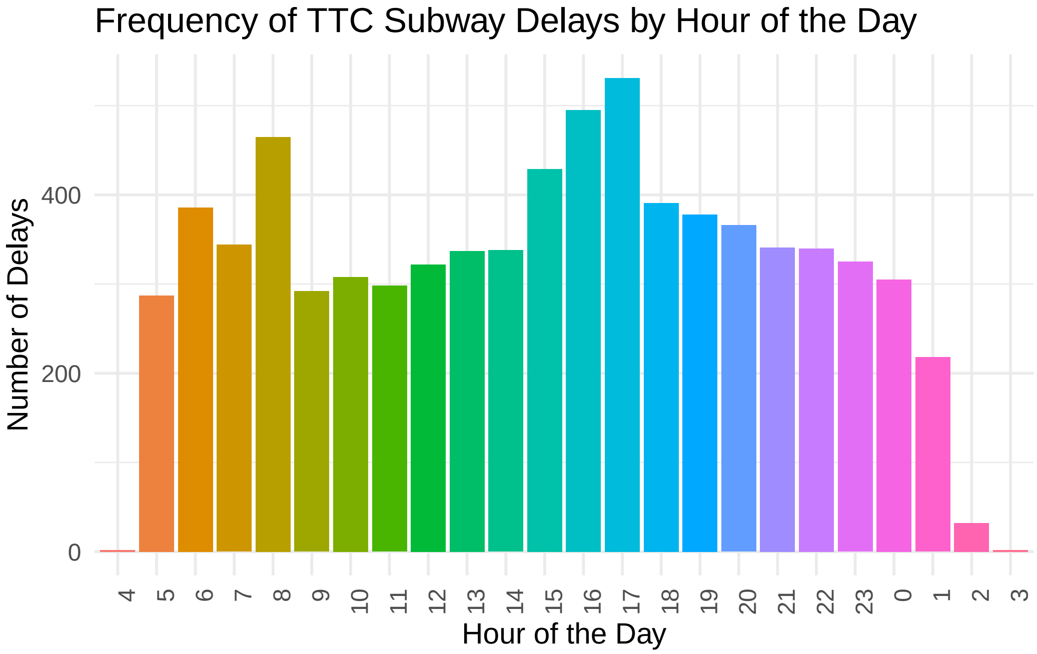

Figure 4: Frequency of subway delays categorized by time of the day

The data illustrated in Figure 4 showcases the distribution of TTC subway delays throughout the day. Notably, there are significant peaks in the number of delays around 8 a.m. and again between 4 p.m. and 5 p.m., reflecting the typical rush hour periods when the subway system is under the most strain from high passenger volumes. Moreover, the number of delays in the evening rush hour is much higher compared to moring rush hour, suggesting a greater crowd level. Outside of these peak times, the frequency of delays appears relatively consistent, averaging approximately 300 instances per hour over the course of the year. It is important to note that the TTC subway's operational hours are from 6 a.m. to 2 a.m., as stated by the official TTC schedule (TTC, n.d.). The occurrence of delays recorded outside these operational hours, specifically at 2 a.m. and 5 a.m. to 6 a.m., may be attributed to late-night service extensions, early morning starts, or ongoing service from the previous day that has not been concluded after the end of service time.

4 Conclusion

Based on the comprehensive analysis of the 2023 TTC subway delay data, this report has led to two primary suggestions. For commuters, particularly those who regularly use the heavily impacted lines during peak weekday hours, it is advisable to plan for potential delays by allowing extra travel time. Extra time are highly recommended especially for people taking "line 1" and "line 2" since those 2 lines are more likely to get delays. Such foresight can mitigate the inconvenience caused by unexpected waiting times.

From an operational perspective, the data presented herein should prompt an extensive evaluation of current TTC practices. There is a clear opportunity for management to refine operational strategies, focusing on peak times that have been identified as hot-spots for delays (i.e. at 5 p.m. for line 1). Addressing the fundamental issues contributing to frequent delays could significantly enhance service reliability.

The overarching aim is to elevate the level of service to one that not only meets the expectations for punctuality and efficiency but also fosters confidence among Toronto's commuters. By harnessing the insights from this analysis, there is potential to drive progressive changes within the TTC's subway operations, ultimately benefiting the broader public transportation network in Toronto. This study, therefore, not only sheds light on existing patterns of delay but also serves as a catalyst for ongoing enhancements to the city's transit infrastructure.

5 Appendix

5.1 Rough Sketch about Brainstorming

Please check the `/input/misc/Plan-Sketch.pdf` for my sketch of my brainstorming steps. This document contains a rough sketch including a sample data set and an example of plot that can be useful to answer the research questions.

5.2 TTC Subway Map

Figure 5 below shows the map of subway lines in Toronto from TTC (TTC, n.d.).

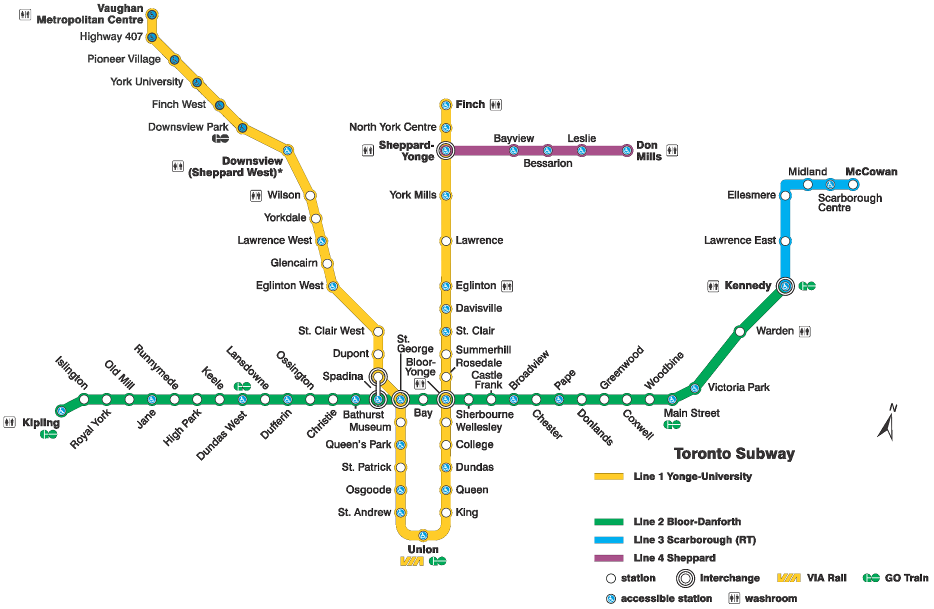

Figure 5: TTC Subway Map

5.3 Data Simulation

For details on the simulated data set creation, please check the script located at `Script/00-simulate-data_data.R`. This script generates a simulated data set consisting of 1000 entries, representing hypothetical subway delay instances for the year 2023. The data set is been saved to `\input\data\simulate_subway_data.csv`. The data set includes the following columns:

- Date: The date of the delay, formatted as a date within the year 2023.
- Day: The day of the week on which the delay occurred.
- SubwayID: A unique identifier for the subway vehicle involved in the delay.
- Time: The time of day when the delay took place, formatted as 'HH:MM'.
- Delay_Durations: The length of the delay in minutes.

5.4 Download Data from OpenDataToronto

For details about download the data set from OpenDataToronto, please check the script located at `Script/01_download_data.R`. This script retrieves the TTC Subway delay data for the year 2023 using the package called `opendatatoronto` (Gelfand 2022) and save to a file at `/inputs/data/subway_delay_data_2023.csv`.

5.5 Data Cleaning Process

In preparing the data set for analysis, the data is been cleaned by removing all rows with ‘na’ for their ‘Line’ column, and their ‘Line’ column not in [“BD”, “YU”, “SHP”, “SRT”] (the 4 subway lines in Toronto). This step ensures the accuracy and reliability of the subsequent analysis by excluding incomplete or irrelevant data points. For a detailed view of the cleaning procedure and the code used, please refer to the script located at `Script/02_cleanup_data.R` in the repository. The cleaned data set is saved to `/output/data/`

5.6 Data Set Validity Testing

Once the data set has been cleaned, its integrity can be validated by executing the script located at scripts/03-test-data-validity.R. This script performs three crucial checks to ensure the data set’s accuracy and consistency:

1. Year Consistency: Verifies that all entries in the ‘Date’ column correspond to the year 2023.
2. Minute Delay not Negative: Confirms that all values in the ‘Min Delay’ column are greater than or equal to zero, ensuring no negative delay times are recorded.
3. Line Validation: Ensures that the ‘Line’ column includes only the subway lines in Toronto: “BD” (Bloor-Danforth), “YU” (Yonge-University), “SHP” (Sheppard), and “SRT” (Scarborough RT).

After running this script, it’s expected that the variables `test_year_2023`, `test_min_delay_non_negative`, and `test_line_inclusion` should all return TRUE, indicating that the data set meets the specified criteria for each test. If any of these checks fail, it would suggest discrepancies within the data set that may require further investigation or correction.

Reference

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

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. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

TTC. n.d. *TTC.ca*. https://www.ttc.ca/.

Vuong, Oriena. 2019. "Most Transit Commuters Satisfied with TTC, but Many Affected by Unexpected Delays: Poll - Toronto." *Global News*. Global News. https://globalnews.ca/news/5338014/ttc-service-toronto-transit-commuters-poll/.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.