

# A Modest Proposal for Developing Better Experience with TTC Subway\*

Terry Tu

January 23, 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 by 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 for this study is to lead a better subway service, making everyday travel smoother for lots of people in Toronto.

## Table of contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Data</b>	<b>2</b>
2.1	TTC Subway Delay Data . . . . .	2
2.2	Data Cleaning and Initial Observations . . . . .	3
2.3	Discription for variable used in this study . . . . .	3
<b>3</b>	<b>Data Analysis &amp; Results</b>	<b>4</b>
3.1	Analysis on subway delay based on day of the week . . . . .	4
3.2	Analysis on subway delay based subway line . . . . .	4
3.3	Analysis on subway delay based on time of the day . . . . .	7
<b>4</b>	<b>Conclusion</b>	<b>7</b>
<b>5</b>	<b>Appendix</b>	<b>8</b>
5.1	Data Cleaning Process . . . . .	8
	Reference . . . . .	8

---

\*Code and data supporting this analysis is 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 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.

#### 3.2 Analysis on subway delay based subway line

The visualized data indicates a huge contrast in delay frequencies across different subway lines for the year 2023. Notably, certain lines such as ‘999’ and ‘BD Line 2’ registered no delays, suggesting efficient operation or a lack of reporting. In contrast, other lines like ‘SHP’ and ‘SRT’ experienced a minimal number of delays, which could indicate isolated incidents or effective management of potential disruptions.

However, a significant concern is observed with the ‘BD’ (Bloor-Danforth) and ‘YU’ (Yonge-University) lines, which exhibited a substantially higher frequency of delays. This pattern highlights a critical area for operational focus. As a TTC subway manager, this insight necessitates an in-depth analysis of the causes behind the delays on these lines. Moreover, people

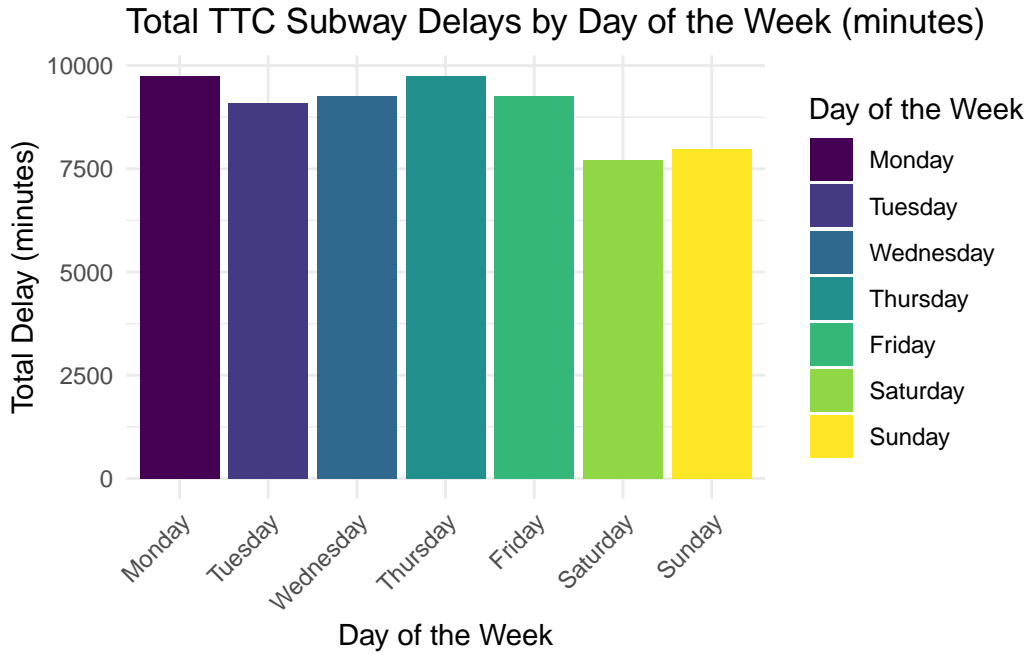


Figure 1: Total minutes of subway delays categorized by days of the week

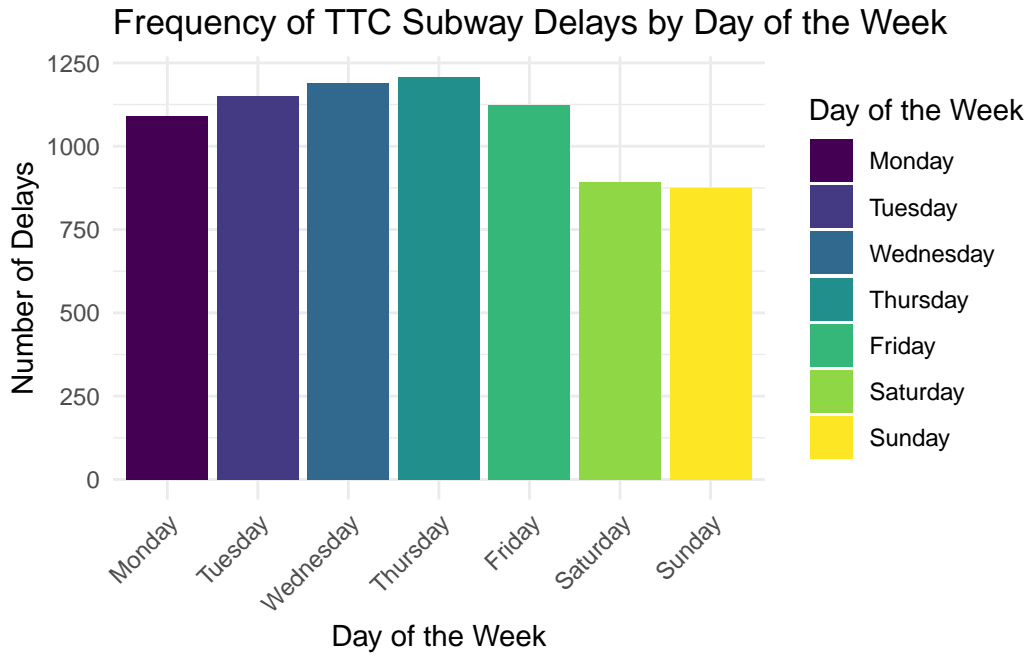


Figure 2: Frequency of subway delays categorized by days of the week

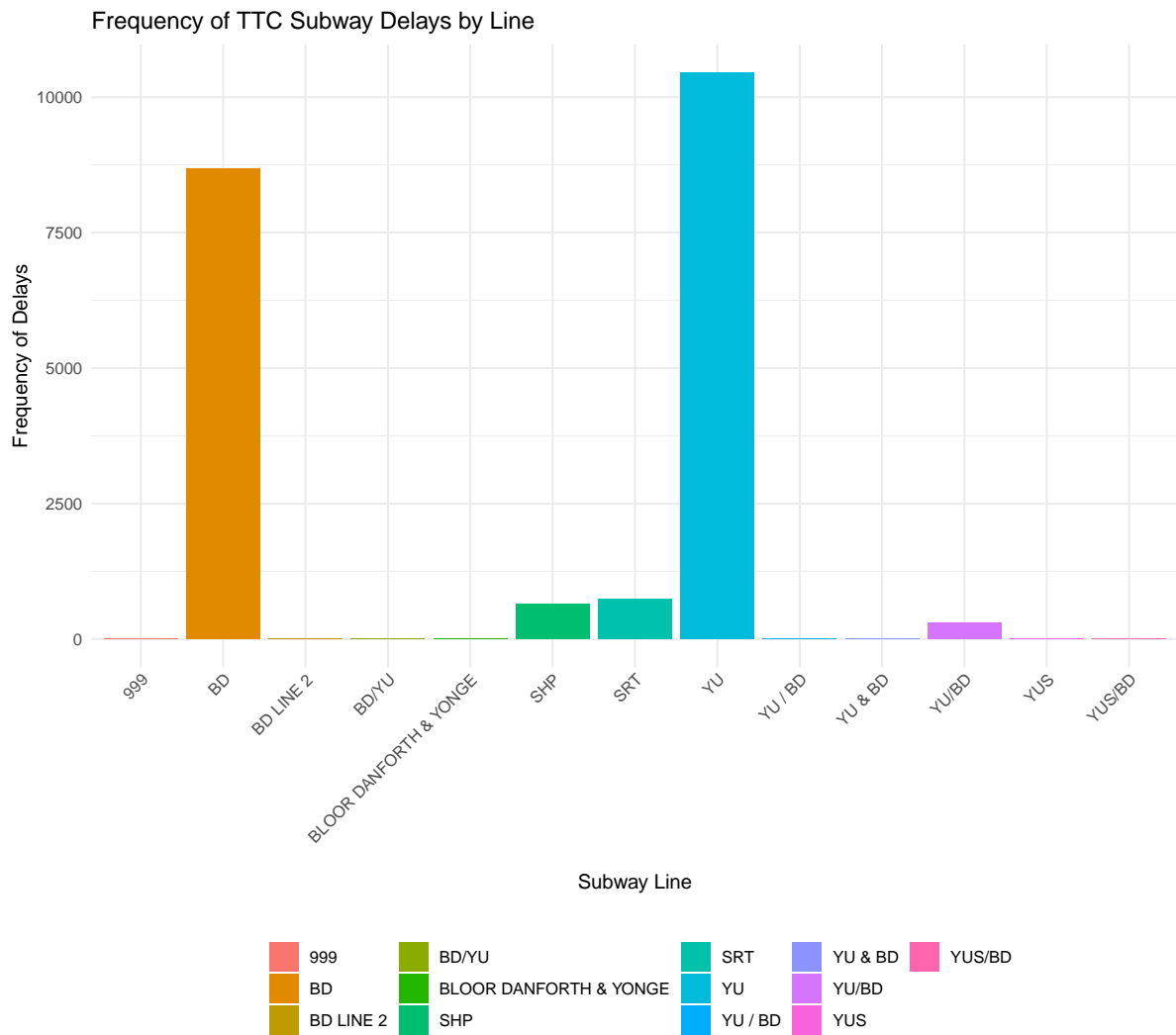


Figure 3: Frequency of subway delays categorized by subway line

taking these two lines are highly recommended to plan with extra time allowance in anticipation of the higher frequency of delays encountered on these routes.

### 3.3 Analysis on subway delay based on time of the day

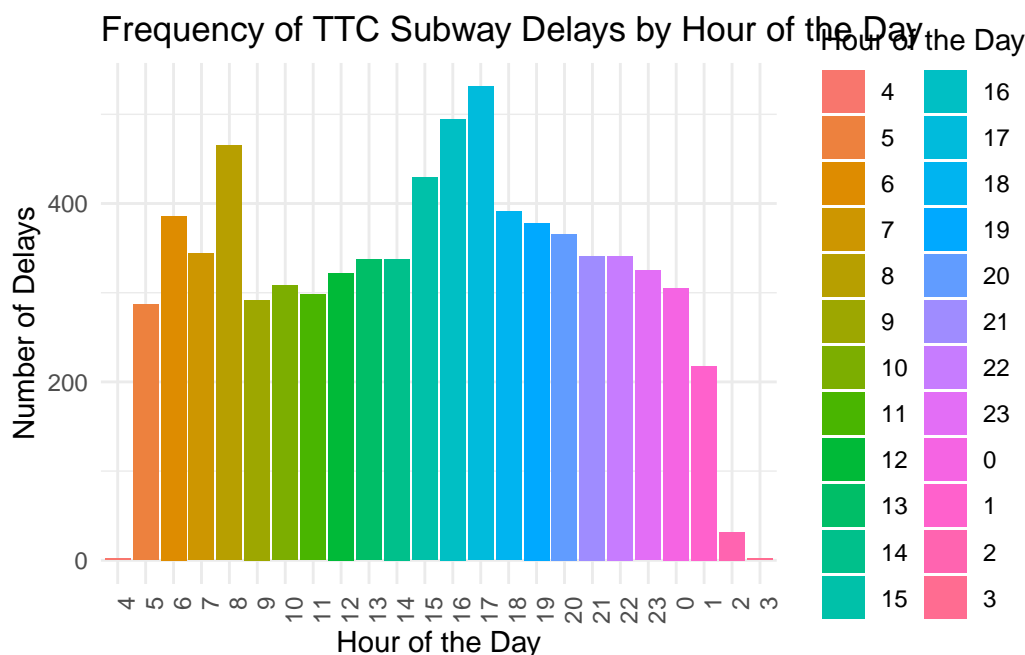


Figure 4: Frequency of subway delays categorized by time of the day

## 4 Conclusion

Based on the data from “ttc-subway-delay-data-2023”, I have provided 2 suggestions. For commuters, especially those frequenting the more affected lines, and on weekdays it is prudent to anticipate potential delays and allocate additional time for travel. On the management side, the insights from this report should catalyze a thorough review of operational procedures and the implementation of strategic improvements to address the root causes of frequent delays.

Ultimately, the goal is to achieve a level of service that ensures efficiency, satisfaction, and the continued trust of the commuting public in the TTC subway system. This analysis lays the groundwork for the continuous improvement of Toronto’s subway operations and the enhancement of the city’s public transportation ecosystem.

## 5 Appendix

### 5.1 Data Cleaning Process

In preparing the data set for analysis, the data is been cleaned by removing all lines with ‘na’ for their ‘Line’ column. 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. This script outlines the systematic approach taken to refine the dataset, making it suitable for the comprehensive analysis that follows.

### 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/>.
- Vuong, Oriana. 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>.