

# What’s Delaying Toronto’s Subways?\*

## A Dive into Causes and Patterns: Uncovering the Trends Behind TTC Subway Delays

Weiyang Li

September 22, 2024

This paper explores the underlying patterns of subway delays in Toronto, revealing that subway delays are more frequent during peak hours and tend to be concentrated around major stations and high-demand subway lines. Notably, while most delays are brief, the data also highlights some extreme cases of disruptions lasting several hours. The analysis shows that facility failures and extreme weather can lead to longer delays. These insights suggest that targeted improvements to key stations and infrastructure can significantly reduce overall delay times and improve service reliability.

## 1 Introduction

Urban public transportation systems are an important part of urban construction and development, connecting millions of commuters to their daily destinations and underpinning the vitality of city life. Compared to traditional bus systems, subway systems are emerging as one of the most efficient transportation networks, thanks to advantages such as shorter commute times across regions, relief from surface traffic, and independence from the weather. However, even in large cities like Toronto, subway systems, while an integral part of the daily commute, still face operational challenges, the most common of which are delays. Based on feedback from a random sample of 1,936 transit users surveyed by Forum Poll™, about one-third of TTC (Toronto Transit Commission) regulars report being dissatisfied with its service, with almost half of riders mentioning that they often experience unexpected delays while riding the TTC (Research 2019).

Recent analyses show that the TTC’s ridership recovery after the pandemic remains only at 70% of pre-pandemic levels, exacerbating financial strain and leading to increased delays, particularly during peak hours (Elliott 2023). This “transportation death spiral” – reduced

---

\*Code and data are available at: <https://github.com/doravmony/TTC-Delay>

ridership leading to fewer resources and thus more delays – has placed the system in a critical state, threatening the reliability of the TTC network (Elliott 2023). Therefore, an in-depth understanding of the causes and patterns of delays is important for improving service reliability. By identifying which factors cause more frequent or longer delays and whether certain stations or lines are more prone to disruptions, it is critical to target interventions. Urban transit managers can target operational improvements to reduce unnecessary delays and improve the overall passenger experience.

Although previous research has identified commuter dissatisfaction and financial strain related to public transit delays (Elliott 2023), there has been little rigorous analysis to determine specific patterns in the duration and frequency of these delays. To address this gap, we conducted an in-depth analysis of the 2024 TTC subway delay dataset, focusing on delay trends at the station and line level, as well as across different dates. Through exploratory data analysis and visualization, our study examines which stations and lines experience the most delays, the times of day when issues are most severe, and the primary causes behind the longest service disruptions. Our findings reveal that certain stations and lines are disproportionately affected by delays, and that disruptions caused by mechanical failures tend to last significantly longer than those caused by other factors. These insights are crucial as they provide a foundation for policy changes, resource allocation, and targeted interventions to improve the reliability of TTC services.

**In this paper, we analyze the 2024 TTC subway delay dataset, made publicly available through Open Data Toronto, to examine the frequency, duration, and causes of delays across various subway stations and lines. Our analysis investigates patterns based on factors such as time of day, location, and specific causes of delays. We also focus on identifying which stations or lines are more prone to disruptions and what factors contribute to longer or more frequent delays.** The remainder of the paper is structured as follows: Section 2 introduces the raw dataset and describes the cleaned datasets, along with a preliminary analysis through numerical summaries and visualizations. Section 3 explores key findings from our analysis. Lastly, Section 4 addresses the limitations of the analysis and offers recommendations for reducing delays and improving subway service reliability.

## 2 Data

### 2.1 Raw Data

The dataset used for this analysis is the 2024 TTC Subway Delay Data, accessed from the Open Data Toronto portal and read into this paper via the `opendatatoronto` library (Gelfand 2022). This dataset, released by the Toronto Transit Commission, comprehensively documents 17,517 subway delays in Toronto’s public transit system during the year 2024 (City of Toronto 2024). Each entry in the dataset documents a delay event. Data is updated monthly and is

as of August 31, 2024 for use in this paper. All the data analysis was conducted using R (R Core Team 2023) with the support of the following packages: tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2023), readxl (Wickham and Bryan 2023), janitor (Firke 2023), ggplot2 (Wickham 2016), knitr (Xie 2023), kableExtra (Zhu 2024), and gridExtra (Auguie 2017).

In the real world, subway delays are caused by a variety of operational or external factors, such as mechanical issues, weather conditions, or accidents. Each time a delay occurs, the TTC records detailed information, including the location, time, and cause of the delay. These recorded incidents are compiled into the TTC subway delay dataset, where each entry represents a specific delay event. In the data section, we trace this process—how real-world delays are documented, transformed into structured data, and ultimately used for our analysis. In this way, the TTC dataset is well suited for our analysis and specificity allows for a thorough analysis of Toronto’s transportation challenges, whereas other datasets such as New York (MTA) or London (Underground) may not be as feasible. The essential details included in the dataset that allow for an in-depth examination of the temporal and spatial distribution of delays, as well as an exploration of the underlying causes contributing to service interruptions. The key variables are as follows:

| Variable  | Description  |
|-----------|--|
| Date      | The date when the delay occurred.  |
| Time      | The time when the delay occurred (24-hour format).                                   |
| Station   | The specific subway station affected by the delay.                                   |
| Line      | The subway line involved in the delay (e.g., Yonge-University).                      |
| Bound     | The direction of travel (e.g., Northbound.).   |
| Code      | The cause of the delay, categorized into different codes (e.g., mechanical failure). |
| Min Delay | The duration of the delay in minutes.  |
| Min Gap   | The time gap between trains during the delay event, in minutes.                      |
| Vehicle   | The train number associated with the delayed subway car.                             |

Figure 1: Key Variables in the TTC Subway Delay Data

## 2.2 Cleaned Data

In this paper, we utilize two cleaned datasets for analysis. In the first dataset ([ttc dataset](#)), missing values in the station and line variables were removed, as the focus is on analyzing delays at specific stations and on particular lines. In the second dataset ([delay dataset](#)), we retained only records with actual delays, ensuring that our analysis is centered on significant delay events. Detailed data cleaning is included in the [Appendix A.1](#).

We analyzed the numerical summary using the cleaned TTC dataset containing complete records. The mean delay duration of 2.84 minutes suggests that subway delays are generally

| Statistic          | Delay Duration | Time Gap |
|--------------------|----------------|----------|
| Min                | 0.00           | 0.00     |
| 1st Qu.            | 0.00           | 0.00     |
| Median             | 0.00           | 0.00     |
| Mean               | 2.84           | 4.20     |
| 3rd Qu.            | 4.00           | 8.00     |
| Max                | 716.00         | 719.00   |
| Standard Deviation | 11.06          | 11.80    |
| Variance           | 122.42         | 139.22   |

Figure 2: Summary Statistics for Delay Duration and Time Gap

brief. Also, with a median of 0 minutes, more than half of the recorded events involved no actual delay. The third quartile value of 4 minutes indicates that 75% of the delays lasted less than 4 minutes, implying that longer delays are relatively uncommon. Most of the delays can therefore be considered minor. Furthermore, the standard deviation and variance reveal that the delays varied significantly, with many being short, some were considerably longer. Notably, we observed a maximum delay of 716 minutes, a substantial outlier. Upon closer inspection, it was determined that this outlier should not be discarded, as it resulted from an exceptional event rather than a data entry error.

The average gap between trains during delays is 4.20 minutes. Like the delay duration, the median gap is 0 minutes, indicating that most events in the record do not lead to significant gaps between trains. The maximum gap recorded is 719 minutes, representing an extreme case likely associated with rare or exceptional incidents causing major service disruptions. Notably, 75% of the gaps are under 8 minutes, suggesting that while delays do occur, they generally do not significantly disrupt the overall subway schedule.

## 2.3 Exploratory Data Analysis

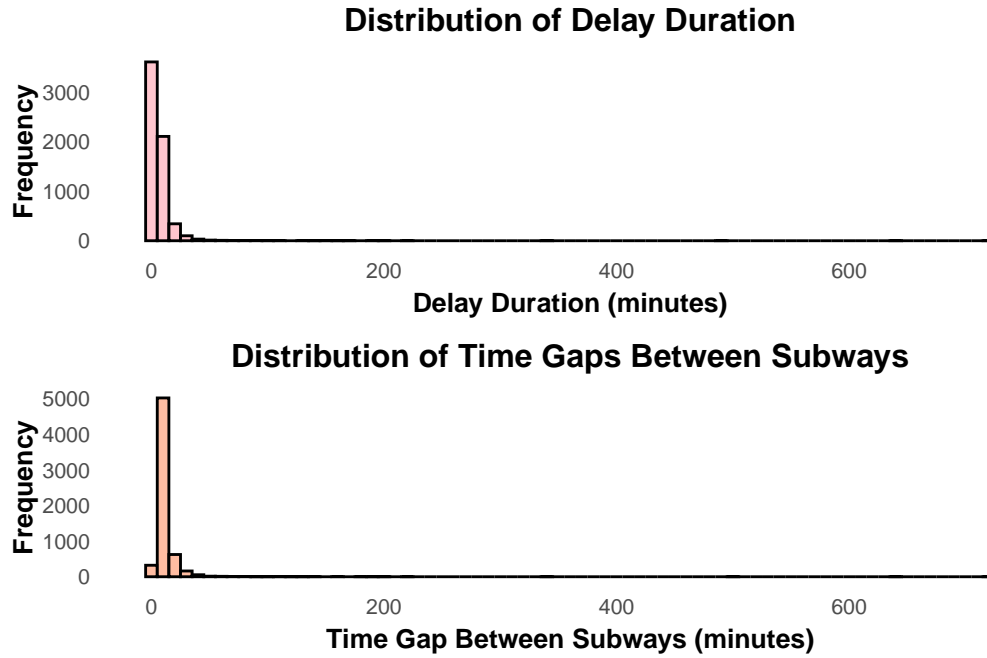


Figure 3: Histogram of Delay Duration and Time Gap

The histograms (Figure 3) for delay duration and time gaps for actual delay dataset align with our analysis from the numerical summary (Figure 2). The vast majority of delays and time gaps are concentrated around very short durations, with most data points clustered near small values. Both distributions exhibit extreme outliers, resulting in long right tails that reflect rare but significant disruptions. While these outlier events are infrequent, their magnitude is substantial, as highlighted by the extended tails of the distributions.

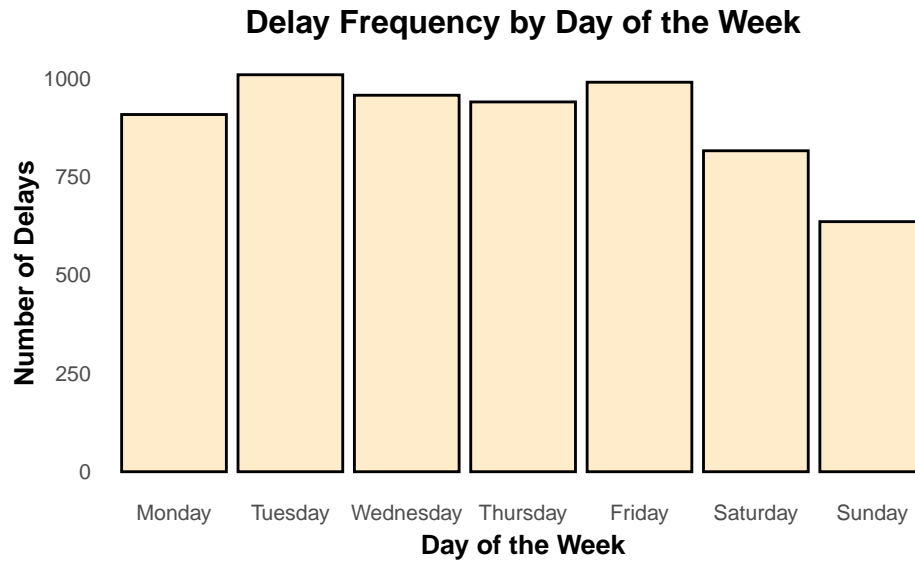


Figure 4: Bar Plot of Delay Frequency by Day of the Week

The bar plot (Figure 4) illustrates the frequency of subway delays by day of the week, considering only delays where the duration exceeds zero. The data reveals that delays are most frequent on Tuesdays, followed closely by Fridays and Wednesdays, indicating a peak in delays during the middle of the workweek. In contrast, Sundays experience the fewest delays, likely due to reduced subway usage and service during weekends. Overall, the pattern suggests that the subway system experiences more frequent delays during weekdays, particularly in the middle of the week, when commuter traffic and operational demands are higher.

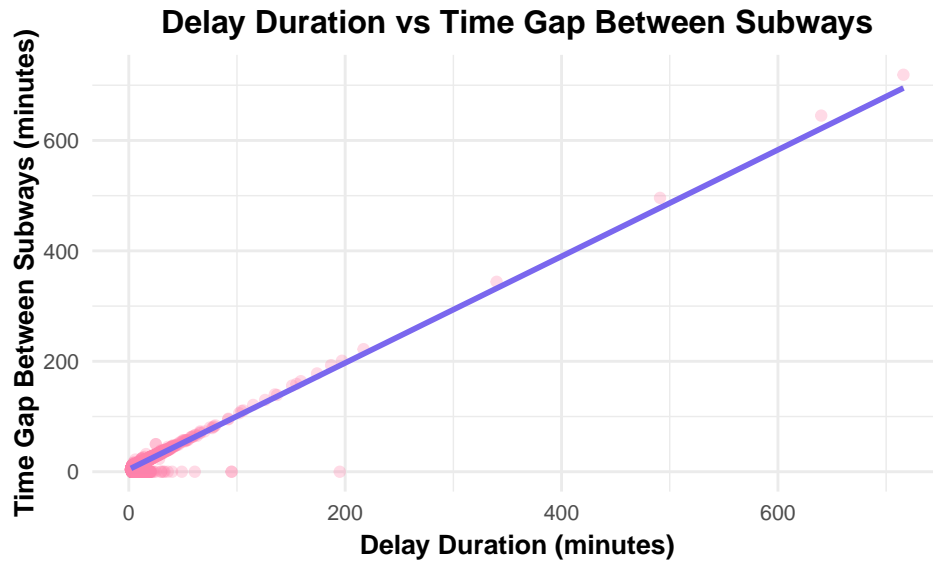


Figure 5: Scatter Plot between Delay Duration and Time Gap Between Subways

The scatter plot (Figure 5) shows a clear positive linear relationship between delay duration and the time gap between subways. As the delay duration increases, the time gap between trains also rises proportionally. This strong correlation indicates that longer delays directly lead to larger gaps between subway services, likely causing significant disruptions. This relationship highlights the operational challenges of managing longer delays, as they not only prolong the delay duration but also increase the intervals between trains, contributing to greater service inefficiencies.

### 3 Results

#### 3.1 Do longer and more frequent subway delays occur at specific times of the day?

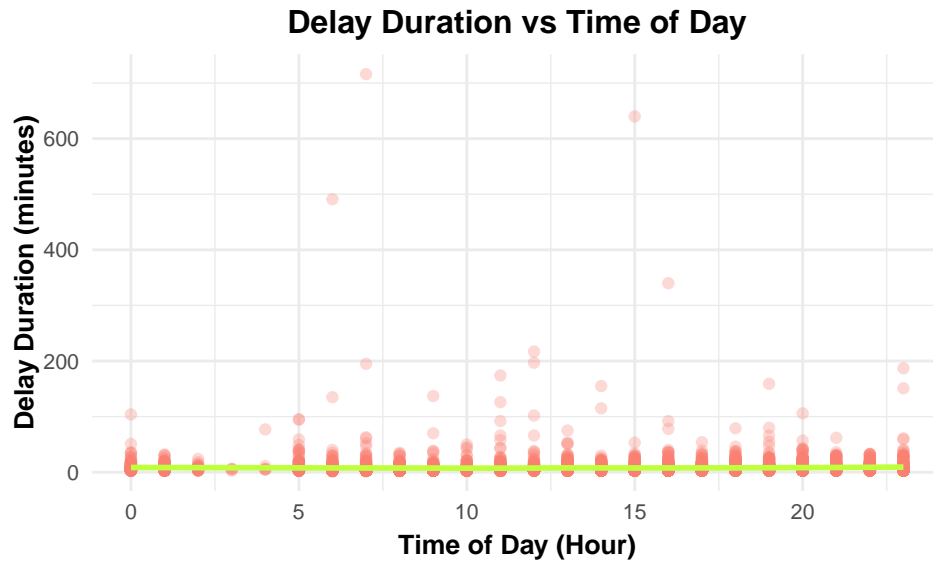


Figure 6: Scatter Plot of Delay Duration and Time of Day

The scatter plot of delay duration versus time of day (Figure 6) reveals that the majority of delays are clustered near zero, indicating that most recorded delays are relatively short and have minimal impact on service. These brief delays occur consistently throughout the day, with no significant variation during specific times, such as the morning or evening rush hours. However, the plot also highlights a few extreme outliers, where delay durations exceed 200 minutes, and in some cases, even 600 minutes. These outliers are scattered across different times of the day, suggesting that such long delays are likely caused by special or rare events, rather than being tied to a particular time of day.



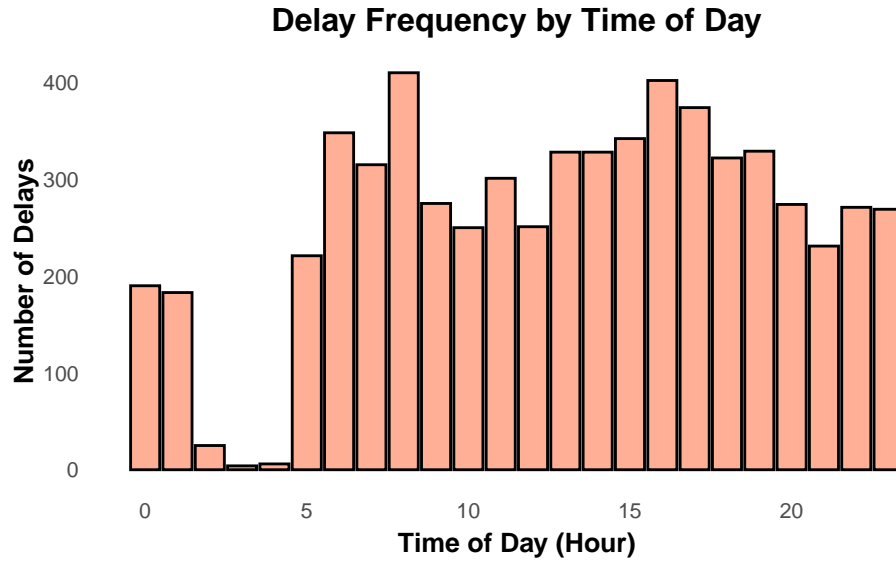


Figure 7: Bar Plot of Delay Frequency by Time of Day

The bar plot (Figure 7) illustrates the frequency of delays across different hours of the day, filtered to include only events where the delay duration is greater than zero. The data reveal that delays are more common during certain periods of the day. Notably, there is a significant rise in delays during the morning hours between 7 AM and 10 AM, coinciding with the morning rush hour. Similarly, another peak is observed between 4 PM and 6 PM, which corresponds to the afternoon and early evening, potentially linked to the afternoon rush. Conversely, delays are less frequent during late-night and early morning hours (12 AM to 4 AM), when subway service is likely reduced. This pattern suggests that subway delays are more likely to occur during high-traffic periods of the day, aligning with expected increases in train frequency and passenger volume during rush hours. These findings indicate a clear temporal pattern in the occurrence of subway delays, which could be attributed to the increased operational pressure on the system during peak times.

### 3.2 Do certain stations or lines experience more frequent delays?

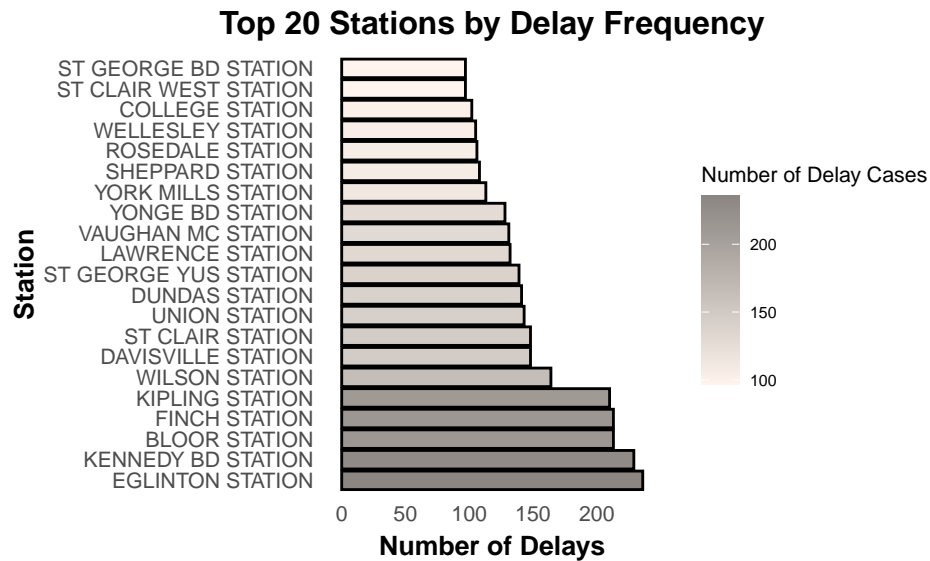


Figure 8: Bar plot for Delay Frequency by Top 20 Stations

The bar plot (Figure 8) shows the 20 subway stations with the most frequent delays. The data indicates that certain stations experience a significantly higher number of delays compared to others. Eglinton Station and Kennedy BD Station top the list, each with over 200 recorded delays. These stations are followed closely by Bloor Station, Finch Station, and Kipling Station, which all have more than 150 delays. The frequency of delays appears to be more concentrated in central and high-traffic stations, such as Union Station and St. George YUS Station, which serve as major transfer points for commuters. These findings suggest that stations along high-demand subway lines, particularly those that serve multiple routes or experience heavy passenger traffic, are more prone to frequent delays. It is likely that operational pressures, maintenance, or higher passenger volumes contribute to the higher number of delays at these locations.

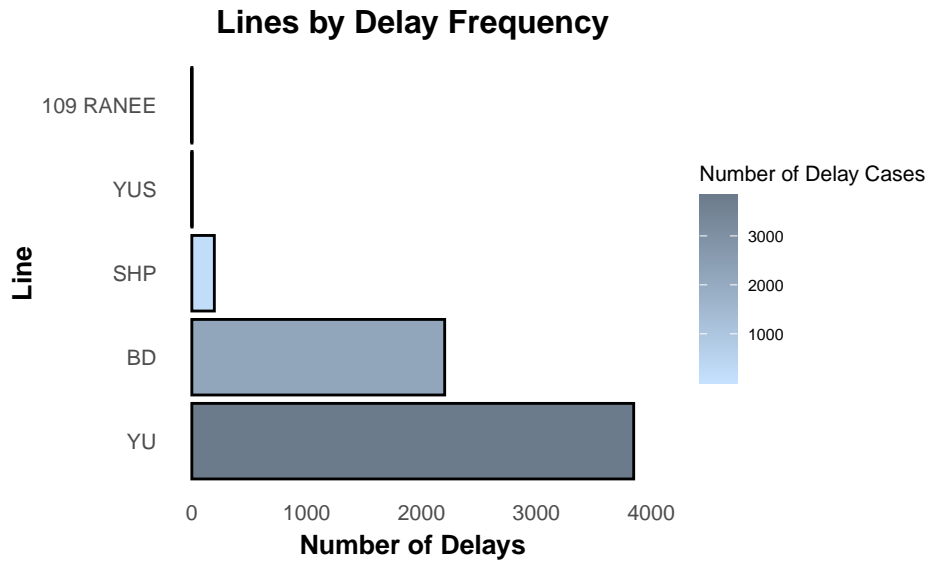


Figure 9: Bar plot for Delay Frequency by Lines

The bar plot (Figure 9) illustrates the delay frequency across different subway lines. The YU (Yonge-University) line experiences the highest number of delays, with nearly 4,000 recorded instances. This is followed by the BD (Bloor-Danforth) line, which also has a significant number of delays, close to 3,000. This distribution highlights that the most heavily used lines, particularly the YU and BD lines, tend to experience more frequent delays, likely due to their higher passenger volume and operational load. These lines serve as the backbone of the Toronto subway system, connecting major hubs and handling a significant proportion of daily commuters.

### 3.3 Do certain delay causes lead to longer delays than others?

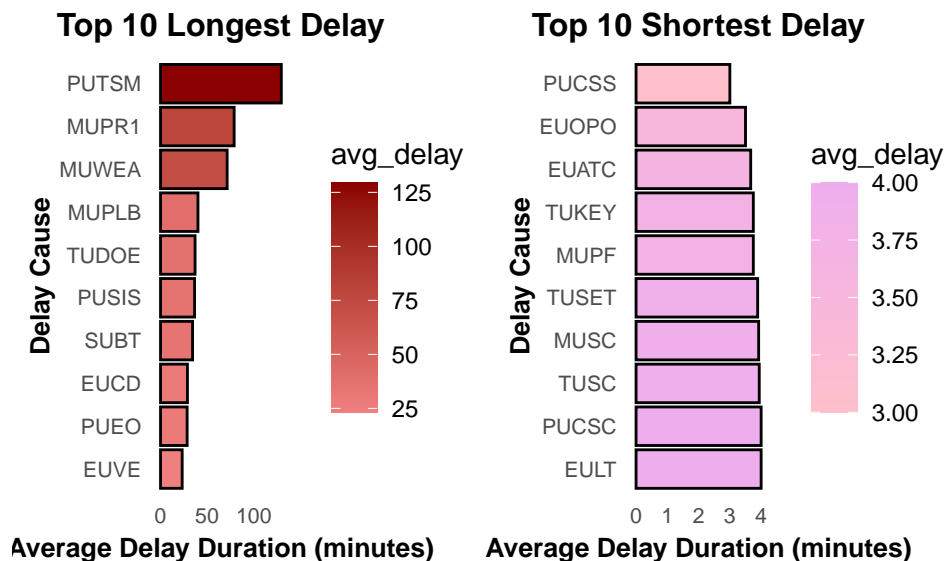


Figure 10: Top 10 Longest and Top 10 Shortest Delay Causes by Average Duration

The two bar plots (Figure 10) present the top 10 causes of the longest and shortest delays, highlighting significant variations in the impact of different delay causes on service duration. On the left, the Top 10 Longest Delay Causes plot shows that the PUTSM (Track Switch Failure) cause leads to the longest average delay, with an average delay duration of over 125 minutes. Track switch failures are critical infrastructure issues that directly affect the movement of trains, often requiring technical teams and considerable time to repair. Other notable causes include MUPR1 (Train in Contact With Person), which involves serious incidents requiring emergency response and safety protocols, and MUWEA (Weather Reports), where extreme weather conditions, such as snow or storms, disrupt service significantly. Both of these causes result in substantial delays, exceeding 75 minutes on average, due to the severity and complexity of resolving these issues, which often involve external factors and safety measures. These incidents require extensive coordination, leading to prolonged interruptions in service. On the right, the Top 10 Shortest Delay Causes plot demonstrates that the shortest delays are associated with causes such as PUCSS (Central Office Signalling System) and EUOPO (Non-Train Door Monitoring), with average delay durations of only about 3 to 4 minutes. These issues may be less severe, or more easily resolved, leading to minimal disruption in subway operations.

Overall, the data indicate that certain delay causes, particularly those related to significant operational issues, result in much longer service interruptions, while other, less impactful causes are resolved more quickly. This distinction is crucial for prioritizing resources and

interventions, as addressing the root causes of the most prolonged delays could significantly improve overall system efficiency.

## 4 Discussion

This study examines the patterns and causes of delays in the 2024 TTC subway system, focusing specifically on instances where actual delays occurred (i.e., when delay minutes exceeded zero). Through exploratory data analysis (Section 2.3), we observed that while most delays are brief and fall within an acceptable range, there are occasional major disruptions that significantly impact overall subway operations. Furthermore, the analysis reveals that delays are more frequent on weekdays compared to weekends, likely due to higher commuter traffic and increased operational demand during the workweek. Additionally, the data suggests that longer delays can increase the spacing between trains, resulting in a cascade of subsequent delays and further reducing service efficiency.

In the results section (Section 3), the analysis indicates that delays occur more frequently during peak hours, though there is no evidence suggesting that these delays are longer in duration compared to non-peak hours. This suggests that while delays are more common during busy periods, their severity remains consistent throughout the day. Additionally, the data reveals a significantly higher number of delays at key transit hubs, likely due to their role as major interchange points in the TTC network, where high traffic volumes and complex operations increase the likelihood of disruptions. The analysis also shows that the most heavily utilized lines experience more frequent delays, driven by higher ridership and operational demands. Moreover, mechanical failures are identified as a leading cause of major delays, in contrast to shorter delays caused by minor issues. Addressing these large-scale disruptions requires prioritizing investments in infrastructure and operational improvements to reduce the frequency and impact of prolonged delays.

### 4.1 Weaknesses and next steps

Our analysis has certain limitations, as it focuses solely on a snapshot of TTC operations from January to August 2024, which may not capture long-term trends in operational efficiency. In the future, we could incorporate the full set of TTC subway delay data from 2017 to 2023, available through Open Data Toronto, to examine whether subway delays have improved over a longer period, particularly around the pre- and post-pandemic periods (before and after 2019). With more comprehensive data, we could also further investigate whether large-scale delays caused by significant events have been mitigated through infrastructure upgrades or other measures. Additionally, a comparison of delays between the subway and bus systems could provide insights, such as whether subway delays are less severe than bus delays during extreme weather conditions.

Overall, addressing subway delays can enhance the reliability of TTC services. By focusing on infrastructure improvements and better management of peak-hour traffic, TTC can provide passengers with a more consistent and satisfactory experience.

## **A Appendix**

### **A.1 Data cleaning**

In the data cleaning process, we standardized the coding format of the station, line, and bound variables to ensure consistency for comparing delay events across stations and subway lines. We removed missing or invalid values from key fields such as station and line, as one of the key focuses of our analysis was to look at trends in delays at different stations and lines. Similarly, we removed entries with missing delay codes describing the cause of delays to ensure the reliability of our analysis of the causes of subway disruptions. We normalized the date and time columns to facilitate time-based analyses, such as identifying peak hours of delay. While we allowed for missing values in the constraint column (direction of travel) because it was not important to our investigation, we ensured the completeness and consistency of all other key information to enable robust exploration of the dataset.

## References

- Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. <https://CRAN.R-project.org/package=gridExtra>.
- City of Toronto. 2024. "TTC Subway Delay Data." <https://open.toronto.ca/dataset/ttc-subway-delay-data/>.
- Elliott, Matt. 2023. "The Rise of TTC Delays, Minute by Minute." <https://toronto.cityhallwatcher.com/p/chw257>.
- Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://CRAN.R-project.org/package=janitor>.
- Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. <https://CRAN.R-project.org/package=opendatatoronto>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Research, Forum. 2019. "Toronto Transit Poll, May 2019." <https://poll.forumresearch.com/post/2958/toronto-transit-may-2019/>.
- 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 Golemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. <https://CRAN.R-project.org/package=readxl>.
- 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>.
- Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.
- Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.