Exploring Subway Delays and Delay Times*

A study of TTC's subway delays throughout the year 2022

Syed Hassan

04 February 2023

Subway delays are quite common and might throw us off our schedule if not anticipated. The paper draws insights from a large dataset of total subway delays in 2022. My data analysis and visualization not only enable us to see the bigger picture but also provide information that means something to us as subway users. I found that the start/end stations of the subway routes face the most delays, and the average and most common delay time is under five minutes.

Introduction

The city of Toronto has a great commuting system that includes buses, streetcars, and most importantly, subway lines (TTC, n.d.c). Road traffic might delay buses and streetcars, but delays are not limited to road transport. In fact, subway and rail delays are equally frequent as well. To an everyday user, the delay times are annoying and frustrating. Access to meaningful data about the delays will help us plan and foresee our commute and schedule. The daily users of the subway are over 1.8 million on average (APTA 2022). These numbers include me and might include you and the people around you.

In the following analysis, I aim to illustrate the number of delays, the stations which suffer the most, the average delay times on the different subway lines, and the average on each day of the week. The utilized dataset comes from TTC's published database available at *Open Data Toronto* (Gelfand 2022). Firstly, the paper demonstrates the frequency of delays and then lists the stations where most delays occur. Subsequently, the analysis summarizes the delay times per minute to provide data that concerns human users.

The report was created using *R* (R Core Team 2020) and *R Studio* (RStudio Team 2020) with *Quarto* (Quarto, n.d.) – a new version of *R Markdown* (RStudio, n.d.). The main library utilized for this purpose is *Tidyverse* (Wickham et al. 2019). Its used sub-packages include

*Code and data are available at: https://github.com/saiyedgh/ttc-delay-times

dplyr (Wickham et al. 2022) to enable query-like syntax, and *ggplot* (Wickham 2016) to create graphs and charts. Other packages and tools include *here* (Müller 2020), *janitor* (Firke 2021), *knitr* (Xie 2022), *kableExtra* (Zhu 2021), and *scales* (Wickham and Seidel 2022). Their respective function is to find *CSV* files, clean data, generate reports, create tables, and enable logarithmic axes.

Data

The data of interest comes from a credible source, the city of Toronto. Toronto Transit Commission frequently updates and exports its data at Open Data Toronto (Gelfand 2022), owned by the city of Toronto. However, TTC has shared no details on the website about how it collects the data. Our analysis assumes the data is complete, accurate, and non-manipulated. It contains all registered subway delays during the year 2022 (TTC 2023).

Table 1: An overview of the dataset.

date	time	day	station	code	min delay	min gap	bound	line	vehicle
2022-01-01	07:43:00	Saturday	WILSON STATION	TUATC	10	0	S	Line 01 Yellow	5896
2022-01-01	08:12:00	Saturday	FINCH STATION	TUNOA	6	12	S	Line 01 Yellow	0
2022-01-01	08:28:00	Saturday	GREENWOOD STATION	TUO	5	10	E	Line 02 Green	5091
2022-01-01	09:45:00	Saturday	KENNEDY BD STATION	TUO	5	10	W	Line 02 Green	5153
2022-01-01	09:51:00	Saturday	FINCH STATION	TUNOA	6	12	S	Line 01 Yellow	0
2022-01-01	11:17:00	Saturday	ST GEORGE YUS STATION	TUO	5	10	S	Line 01 Yellow	5936

The dataset includes details about the date and time when the delay occurs. Mainly, it records the station, the duration of delay in minutes, the train line, and its direction [Table 1]. The different columns include valuable information that will enable us to extract key insights[1]. The "line" column with "bound" gives us four train lines and two directions for each, giving us eight possible user routes. Similarly, the station information with line direction can narrow down the larger data to precise chunks of meaningful information for a daily user. But before going into summaries and specific inferences, we can analyze the extent of all delays to understand the bigger picture.

Furthermore, the delay codes can be decoded to evaluate the causes using the published information by TTC itself (TTC, n.d.a). However, that analysis has already been published in another report by Alyssa Schleifer (Schleifer 2022).

[1]The dataset also includes the gap in minutes that I have intentionally overlooked in this report. According to the published *readme*, the *min_gap* column refers to the "Time length (in minutes) between trains" (TTC, n.d.b).

Visualizing Data

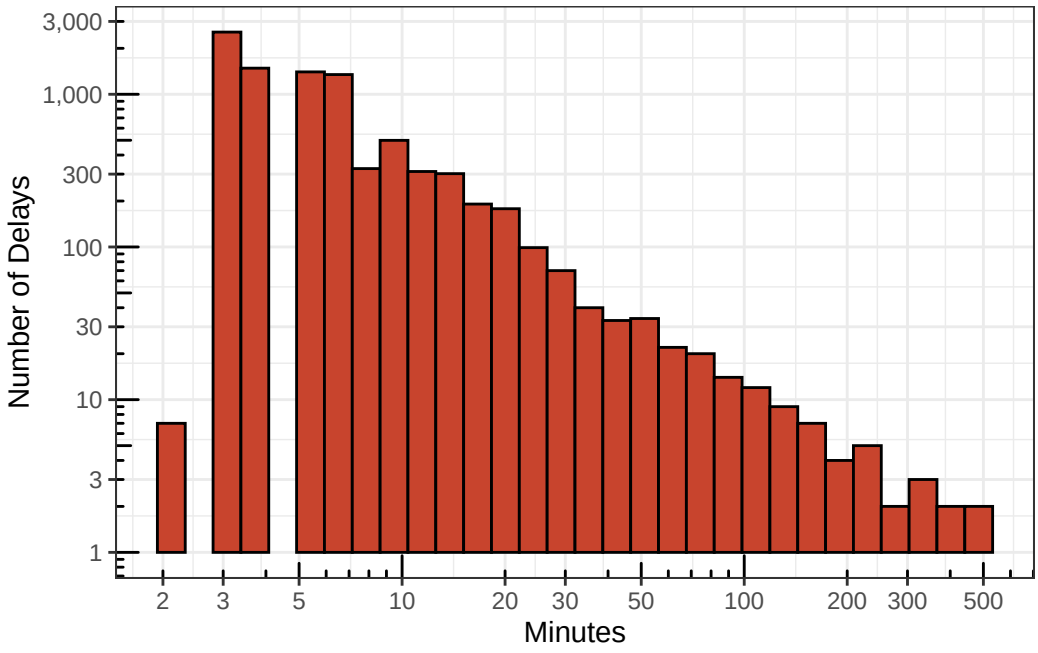

Figure 1: Number of subway delays in minutes.

The above histogram [Figure 1] illustrates the number of delays on all subway lines. It plots the numbers on a log scale which grows exponentially. The breakpoints of the y axis, on the left, start from 1 and increase drastically to 3,000. Similarly, the x axis range at the bottom starts from two minutes and overreaches 500 minutes, which is roughly more than 8 hours. However, the delay duration of most of the delays is less than 30 minutes.

According to Figure 1, the most number of delays are around the three-minute mark reaching almost 3,000 in number. Also, delays at the four and five-minute marks stand at around 1,000+ each. That explains that TTC faced around 14 delays of five minutes or less every day.

Table 2: Total number of delays according to the four subway lines.

Subway Line	**Delays**
Line 01 Yellow	5182
Line 02 Green	2869
Line 03 Scarborough	532
Line 04 Sheppard	390

As per Table 2, the total delays of all trains is 8,973, which is around 24-25 delays per day.

Moving on to a more meaningful inference, let's look at the stations that faced the most delays during 2022.

Table 3: Top 10 stations with most number of delays and average delay time.

#	Station	Line	Bound	Delays	Avg Delay Time
1	FINCH STATION	Line 01 Yellow	S	518	4.61
2	EGLINTON STATION	Line 01 Yellow	S	318	5.12
3	VAUGHAN MC STATION	Line 01 Yellow	S	260	4.71
4	KENNEDY BD STATION	Line 02 Green	W	242	6.04
5	KIPLING STATION	Line 02 Green	E	189	5.21
6	ST GEORGE YUS STATION	Line 01 Yellow	S	180	6.54
7	EGLINTON STATION	Line 01 Yellow	N	144	6.72
8	HIGHWAY 407 STATION	Line 01 Yellow	S	132	7.47
9	WILSON STATION	Line 01 Yellow	S	132	6.45
10	BLOOR STATION	Line 01 Yellow	N	128	7.59

According to Table 3, most delays occurred at *Finch Station* with 518 delays in total and an average delay of 4.6 minutes. Similarly, *Eglinton Station*, not only ranked number two but also number seven, with delays occurring on both routes, north, and south. The above information is beneficial for TTC users, which will help them anticipate delays in their schedule if they pass or use the above stations.

Transitioning back to the various train lines and their delays, let's visualize and compare their average delay times.

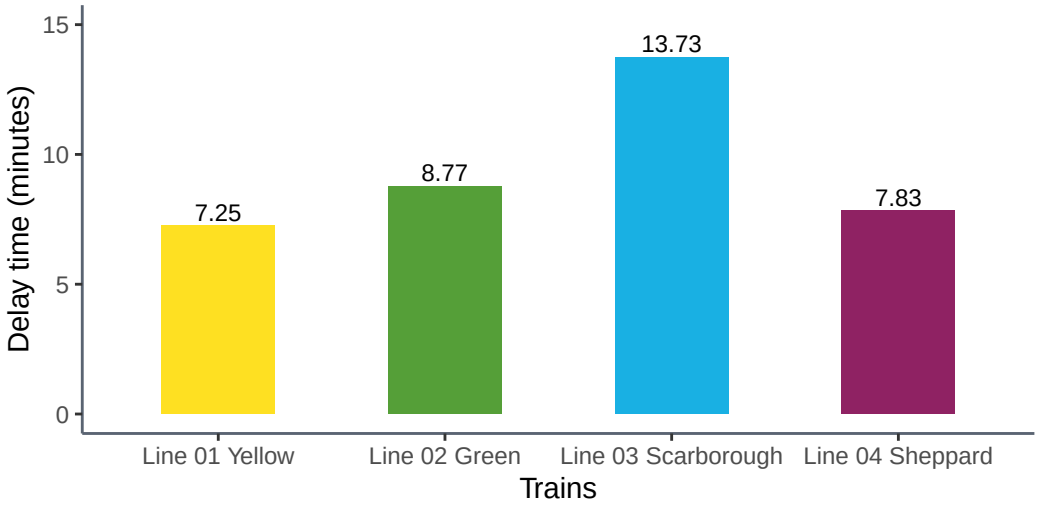

Figure 2: TTC subways line with average delay times.

Looking at Figure 2, we can see that the Scarborough or Blue line delays are the longest, averaging at around 14 minutes per delay. The number is lethargically long. In many cases, averages don't speak for themselves. And delays due to exceptional circumstances can distort the average. For this reason, let's move on towards visualizing the mode, or the most common delay occurrence.

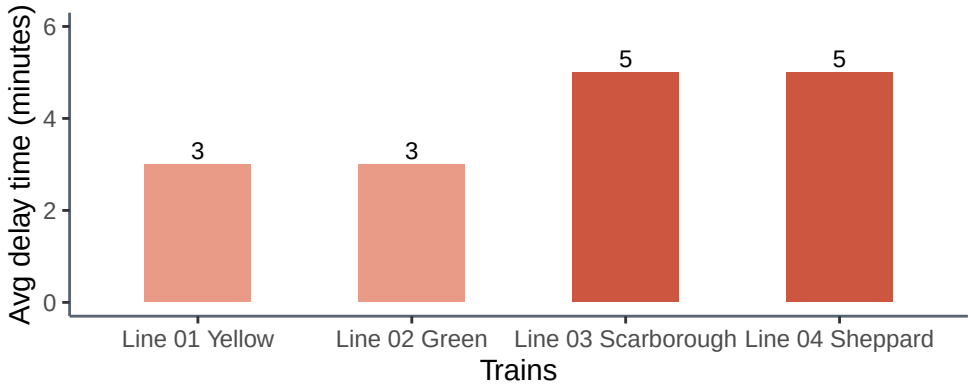

Figure 3: Most common delay duration in minutes.

The above chart [Figure 3] demonstrates that the most common delay is three minutes on the Yellow and Green lines, and five minutes on the Scarborough and Sheppard lines. The mode explains that there might be exceptional delays on the train lines that may be extending the average or mean delay time [Figure 2]. However, the everyday commuter can expect a five-minute delay or less as those are the most common duration.

Lastly, let's analyze the delays per day of the week, which will help us visualize and understand a weekly delay pattern.

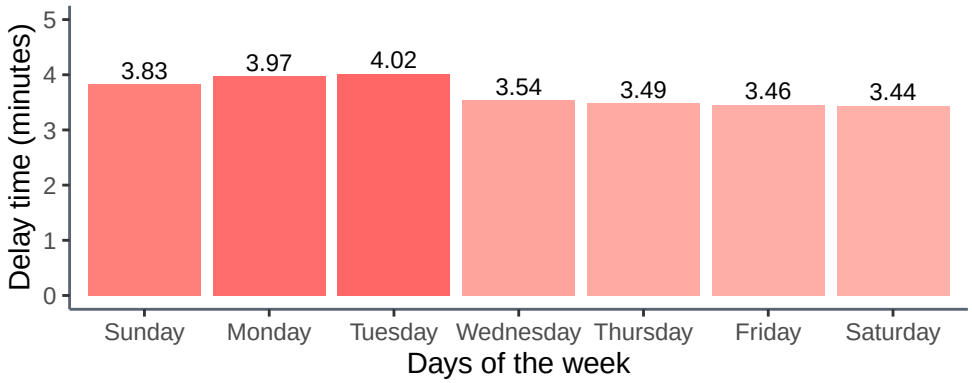

Figure 4: Subway delay average over days of the week.

We can visualize that Sunday, Monday, and Tuesday face more delays on average [Figure 4].

Discussion

The first and last stations face the most delays.

Except for *Wilson Station* and *Eglinton Station*, all the stations that made it to the top 10 list with the most delays are either the last station of the line or the transitioning stations [Table 2]. There can be many reasons which need further examination, but the pattern is quite obvious. The central and transitioning stations are the busiest. Hence, they bear the most burden as well as face delays. Lines 03 and 04 did not make it to the list because their frequency is less than the old lines 01 and 02. The delays at *Eglington Station* are also understandable due to the construction of line 05 (TTC 2015). *Wilson Station*, however, is an odd member within this pattern.

Averages are not always clear.

The averages did not speak clearly. Because the mean includes all delays, it adds unnecessary numbers to the list. Exceptional delays due to criminal incidents or weather conditions might be a part of the list. Even though these numbers are required to be included in data, they should be avoided when calculating analysis such as average delay times for everyday users [Figure 2]. Calculating the mode, however, bypasses such distortions of the mean or average. The most common delay record for each line solves our problem [Figure 3].

Analyzing different kinds of summaries might create a better picture.

Despite the murky picture sketched by averages, analyzing data summary from different perspectives fills the gaps. From Figure 2 we understand that Line 03 has an exceptional delay average despite its small route distance (blogTO 2019). We also know the most common delay times are three and five minutes [Figure 3]. However, Line 03 might have exceptionally long delays than Line 04, which increased its average delay in Figure 2.

Fourth discussion point

Lastly, the weekly delay average is around 3-4 minutes. However, we are more likely to face long delays on Sunday, Monday, and Tuesday. This summary does not explain any significant pattern by mere visuals. However, it paves the way to research the reason for the delay on one weekend and two weekdays. It could be due to TTC's work schedule or due to other external reasons.

The weekly average dilemma.

Even though this paper aimed to utilize data to create meaningful insights for a user, a more comprehensive analysis could have been more exploratory. The subsequent version of this paper will aim to analyze and visualize it.

The dataset is also limited to one year. Comparing data from previous years can help us understand the patterns even better. The next step would have data imported from the previous five years, from 2019-2022. Such analysis would also uncover more patterns, such as the effect of COVID-19 on subway delays.

References

APTA. 2022. *APTA Public Transportation Ridership Report*. https://www.apta.com/wp-content/uploads/2022-Q3-Ridership-APTA.pdf.

blogTO. 2019. "The Evolution of the TTC Subway Map." https://www.blogto.com/city/2013/10/the_evolution_of_the_ttc_subway_map/.

Firke, Sam. 2021. *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.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Quarto. n.d. *Quarto, an Open-Source Scientific and Technical Publishing System Built on Pandoc*. https://quarto.org.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio. n.d. *Your Data Tells a Story. Tell It with r Markdown*. https://rmarkdown.rstudio.com.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Schleifer, Alyssa. 2022. *Toronto Subway Delays Vary Based on Specific Ridership Patterns*. https://tellingstorieswithdata.com/inputs/pdfs/paper_one-2022-alyssa_schleifer.pdf.

TTC. 2015. "Line 05 Eglinton Station Names." https://web.archive.org/web/20210831143710/http://ttc.ca/Ab
.

———. 2023. *Opendatatoronto: TTC Subway Delay Data*. https://open.toronto.ca/dataset/ttc-subway-delay-data/.

———. n.d.a. *Opendatatoronto: TTC Subway Delay Codes*. https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/996cfe8d-fb35-40ce-b569-698d51fc683b/resource/3900e649-f31e-4b79-9f20-4731bbfd94f7/download/ttc-subway-delay-codes.xlsx.

———. n.d.b. *Opendatatoronto: TTC Subway Delay Readme*. https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/996cfe8d-fb35-40ce-b569-698d51fc683b/resource/ca43ac3d-3940-4315-889b-a9375e7b8aa4/download/ttc-subway-delay-data-readme.xlsx.

———. n.d.c. *TTC Website*. https://www.ttc.ca.

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, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*.

https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.