Toronto Subway Delays Vary Based on Specific Ridership Patterns*

Alyssa Schleifer

6 February 2022

Abstract

While the Toronto subway system might be one of the most widely-used and convenient modes of transportation in the city, it is no secret to commuters that subway service has been steadily declining in recent years. As the number of subway delays continues to rise, the overall reliability of the subway is called into question. Data is a critical tool for evaluating and understanding common delay causes and patterns among delay occurrences. In our analysis, we noted a strong correlation between delay frequency and peak ridership hours, and established how some of the most common causes of delays such as passenger-related illness and operator error vary among stations. Finally, we found some inconsistencies in the data in terms of differences in reporting standards between stations. These inconsistencies could hinder the efforts of policymakers who require a clear picture of the problem to address these issues.

1 Introduction

The Toronto subway system serves over 1.5 million people daily, providing a level of efficiency and convenience that other modes of transportation lack (TTC 2019). It continues to grow and expand to meet the evolving needs of the city it supports. Since its opening in 1954, the subway has grown from only twelve stations to 75 stations spread over four lines spanning two cities (City of Toronto Archives 1954). With ongoing efforts in place to continue to expand the subway system, it is important to consider aspects of the subway that could benefit from improvement and try to better understand any major downfalls. Potentially one of the biggest complaints regarding the Toronto subway is the number of delays the subway experiences on a daily basis. For many individuals, these frequent and unpredictable delays compromise the reliability of the subway, especially for those commuting to work and school. In fact, estimates of subway use show evidence of a gradual decline in subway popularity over the past few decades (Nowlan and Stewart 2007). However, much of the previous research fails to offer an in depth examination regarding the reasons for the subway's decline in popularity, or how the ongoing delays might affect the perceived reliability of the subway, and consequently lead to a decrease in subway ridership.

Data regarding subway operations is an extremely important tool in understanding possible flaws or areas of improvement. With delays being one of the most prevalent issues, data focused on recording delays as well as different variables pertaining to these delays could provide extremely valuable insight into some of their leading causes as well as potential patterns in delay prevalence. Since so many trains are operating at any given time, even a brief delay at one station will inevitably have a ripple effect on the rest of the line. This is a trade-off for the convenience the subway offers, as running individual trains so close together is the only way to provide ongoing access to the trains. Thus, being able to minimize the frequency of delays would have a huge impact on the overall service and reliability of the trains.

In the following paper, I will analyze TTC subway delay data in an attempt to shed light onto some of the common causes of subway delays, as well as investigate some common patterns regarding delay frequency. I will do this by first looking into how subway delays fluctuate based on day of the week as well as time of day. In addition, I will look for relationships between delays and train line as well as direction of travel.

*Code and data is available at: github.com/alyssaschleifer/ttc-subway-delay-analysis

Finally, I will examine some of the most common reported causes of delays and their incidences and impact at some of the most frequently delayed stations. This analysis will be carried out in `R` (R Core Team 2021), using the `dplyr` (Wickham et al. 2021), `knitr` (Xie 2021), `bookdown` (Xie 2021), and `tidyverse` (Wickham et al. 2019) packages. All figures in the report are generated using `ggplot2` (Wickham 2016) and tables are created with `kableExtra` (Zhu 2021).

2 Data

2.1 Source and Data Collection

The following report uses the TTC Subway Delay Data (Toronto Open Data 2021) obtained from the City of Toronto's Open Data Portal. This dataset was accessed through R using the `opendatatoronto` package (Gelfand 2020). The data is published by the Toronto Transit Commission on a monthly basis since 2017, and was last updated on October 26, 2021. Although not explicitly stated, it is likely that this data is collected through reports made directly by the TTC as the data contains variables such as ambiguous alphabetical codes which correspond to the reason for the delay.

Due to the methodology surrounding the data collection process, there are a few inconsistencies that could alter the accuracy of the data. For example, some stations such as Eglinton Station record the delays that take place while at the station, but also keep a separate record of the delays that occur on trains that are approaching as well as leaving the station. Most stations only keep record of delays that happen on trains already at the respective station, but do not make note of the delays that occur on trains that are approaching or leaving. This has the potential to alter the data as it could lead to an under-representation of train delays at a particular station if that station does not keep record of trains that are delayed while approaching or leaving the station. If this is the case, there would also be an overall issue regarding the under-representation of train delays as a whole, as many subway delays can occur between stations. However, in the data there is no specific attribute that corresponds to between-station delays, and only a few stations make note of approaching and leaving train delays. Thus, it is unclear how between-station delays are accounted for at the majority of stations or if these delays are simply omitted from the data.

Other issues regarding the collection of data and the potential for bias include how the reasons for the delays are recorded. Since 2016, TTC subway trains have begun transitioning from two-person crews to only a single operator per train (Spurr 2016). In the event of an issue on a train that might cause a delay, the operator would likely be the one responsible for reporting the delay. Reports from recent years indicate that only eight percent of the service hours lost to subway delays were the result of an issue caused by an operator, such as failing to correctly operate the doors (Ouellet 2016). It is unclear if the operator is the sole person responsible for reporting the reason for a delay, however if this were the case it would potentially indicate a source of bias which would lead to an under-representation of delays caused by an operator error, as well as an over-representation of delays attributed to other issues such as mechanical problems or technical malfunctions.

2.2 Looking at the Data

The TTC subway delay dataset I will be using for this analysis is a combination of TTC subway delay data from January to June of 2021. This data contains 7010 samples, across eight variables. These eight relevant variables used to classify information pertaining to the subway delays are Date, Time, Day of the Week, Station, Code, Code Description, Bound (north, south, east, or west), and Line (ex. Yonge-University or Bloor-Danforth).

Working with these variables can provide a great deal of information in terms of categorizing subway delays to gain a better understanding of when and where delays occur, for example. Figure 1[1] shows the total number of subway delays broken down by (1) day of the week and (2) time of day.

[1]Hue palettes used for both graphs in Figure 1 are from the `scales` package (Wickham 2020); ggplots were combined into the same figure using the `patchwork` package (Pederson 2020).

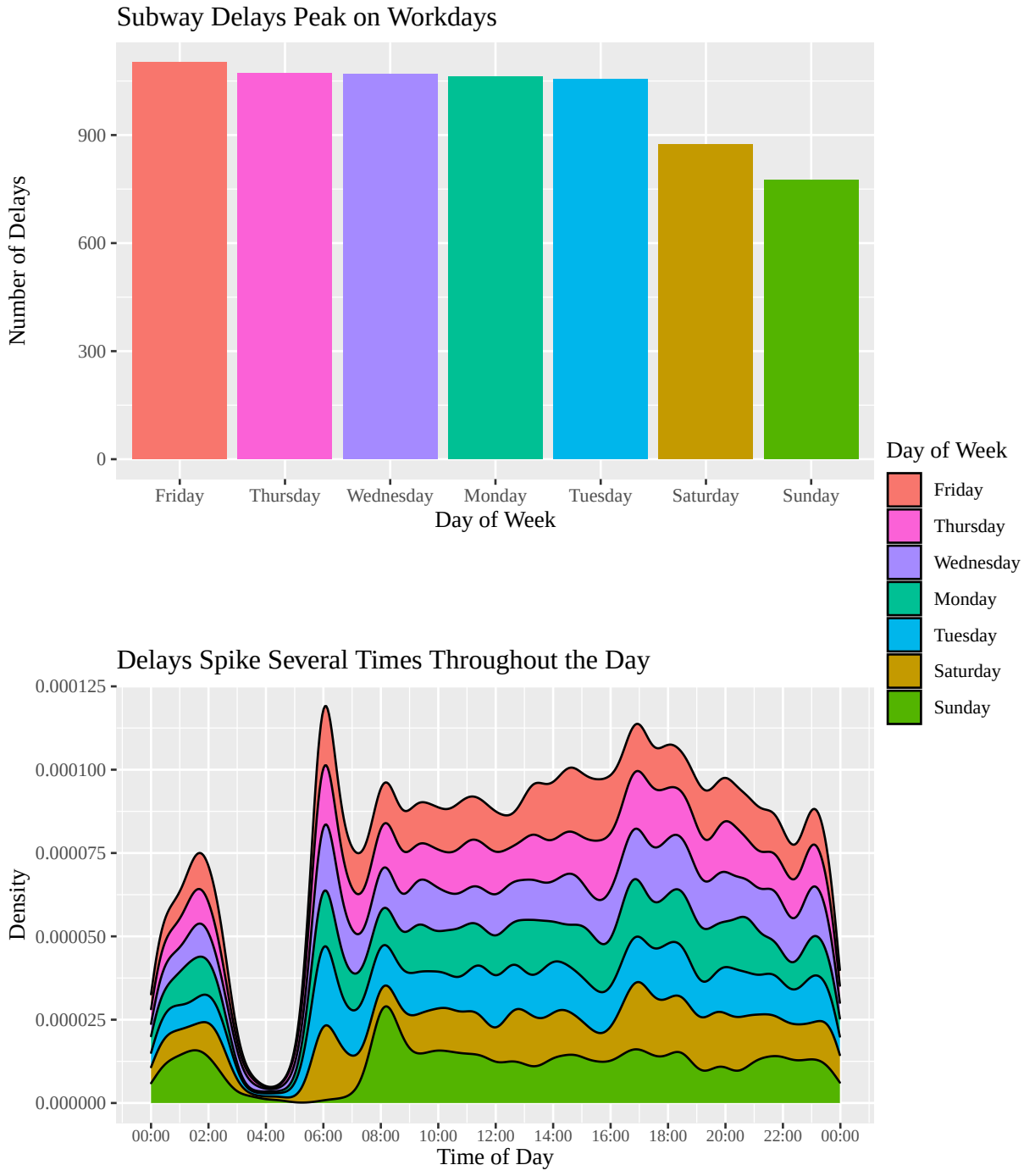

Figure 1: Subway Delays by Day of the Week and Time of Day

Based on these results, we can see delays are much less likely to occur on Saturdays and Sundays, whereas subway delays are the highest at the end of the workweek on Fridays. However, it should be noted that the number of delays recorded from Monday to Friday are not significantly different when compared to the drastic drop in delays occurring on weekends. In order to look into the delay pattern with respect to time of day, we can use a kernel density estimation plot to visualize the probability density of the time of day variable, or more simply, the likelihood of subway delays at any point in a 24 hour period. Based on the peaks in the plot, we can see that the estimated likelihood of subway delays spikes at 6:00am and drops nearly to 0 just after 2:00am, which follows from the fact that the subway does not operate at all between 2:00am and 6:00am (TTC 2022). The exception is Sunday, as the first peak does not occur until 8:00am. This corresponds to the TTC's modified Sunday schedule; trains are not operational between 2:00am and 8:00am. 8:00am peaks can also be observed on the other six days of the weak, as well as another common peak which occurs from around just after 4:00pm until just past 6:00pm.

The days of the week and times of day that most frequently experience TTC subway delays have extremely strong correlation to the typical workweek hours of 9:00am-5:00pm between Monday to Friday. These are the hours and days when the subway would be the busiest due to the volume of people from all over Toronto and the Greater Toronto Area trying to commute to and from work at the same time. This observation could be cause for concern when considering the fact that the times when the greatest amount of people are relying on the subway to get to work are also the times when the subway is the least reliable.

Likewise, Figure 2[2] indicates that the direction of the train also has some influence over the number of delays.

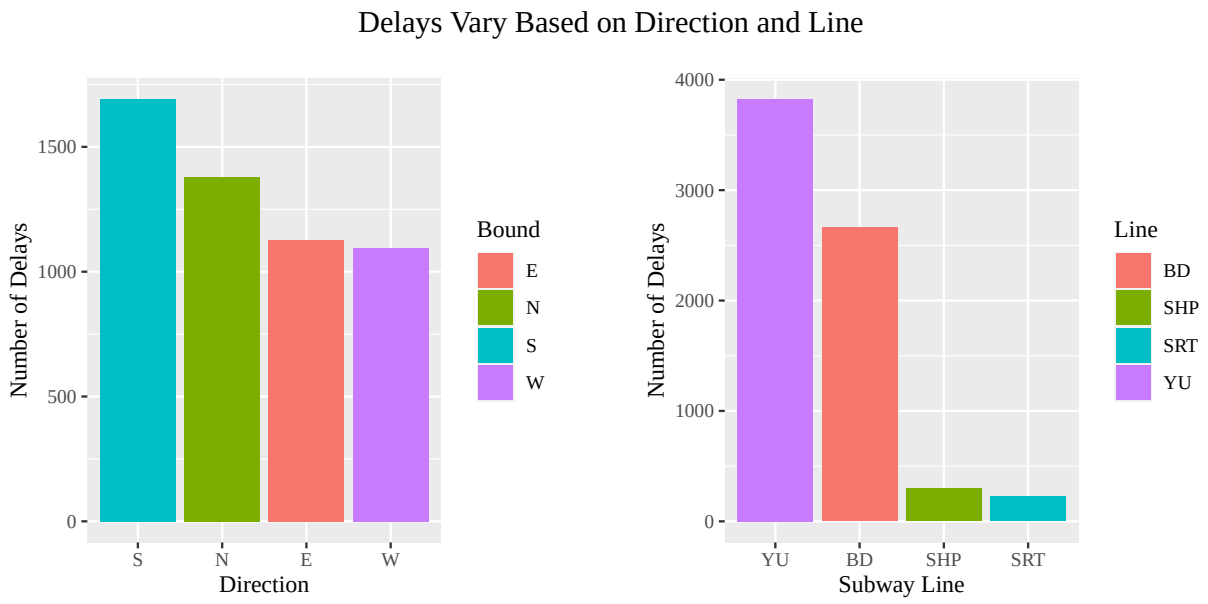

Figure 2: Subway Delays by Bound and Line

Trains going southbound (towards downtown Toronto) encounter more delays that trains traveling in any other direction. This could potentially be due to the increase in population density as you travel more south and farther into the city (ex: from Vaughan to Toronto). This same reasoning can be applied when looking at the specific lines in terms of delays. The Yonge-University subway line experiences the most delays when compared to the other three lines. This line is also the only line that operates within the downtown core, which is the location of approximately 600 000 jobs (City of Toronto 2020), 250 000 residents (Fox 2016), and almost 100 000 university students.

[2]Colours for both graphs in Figure 2 are from the `scales` package (Wickham 2020); ggplots were combined into the same figure using the `patchwork` package (Pederson 2020).

The Code Description variable provides valuable insight into the reason for these delays. Table 1[3] outlines some of most common causes of delay. In order to determine the prevalence of the different delay causes, the three stations with the highest number of reported delays for each of the two major subway lines were established and subsequently analyzed to determine the top 12 causes of delays among these frequently disrupted stations.

Table 1: Commonly Reported Delay Causes At Most Frequently Delayed Stations

	Yonge-University Line			**Bloor-Danforth Line**		
Reason for Delay	**Finch**	**Vaughan MC**	**Rosedale**	**Kennedy**	**Bloor**	**Kipling**
Passenger Related Issues						
Injured or Ill Customer	67	45	4	63	68	54
Disorderly Patron	23	27	12	27	64	36
Assault / Patron Involved	9	2	6	9	11	2
Passenger Assistance Alarm Activated	14	6	3	10	37	6
Operator Related Issues						
Operator Not In Position	28	29	0	3	1	3
Operator Overspeeding	23	0	13	1	0	3
Operator Violated Signal	11	0	2	12	0	19
Mechanical Problems / Signal Malfunction						
Track Switch Failure	2	0	0	2	0	0
Speed Control Equipment	15	0	6	1	0	2
Brakes	5	1	2	3	0	2
Signals or Related Components Failure	4	1	1	2	0	0
Door Problems	2	4	3	11	2	6

The reasoning for reported subway delays appears to vary greatly based on the type of delay as well as the station. Based on the data in Table 1, we can see some discrepancies which are likely due to differences in reporting standards which might vary between stations. However, this data still provides us with several useful insights into the causes behind these delays. For example, it appears that that most prevalent causes of delays among the heavily-delayed stations are related to passenger related issues, specifically injured or ill passengers or disorderly passengers. The term "disorderly" typically refers to passengers that force the train to stop due to disruptive and aggressive behaviour (Ouellet 2017). Earlier observations that noted an increase in delays when the subway was at its busiest align with this information that suggests most delays are due to passenger related issues such as injury, illness, and general disruption. In addition, an interesting observation can be made regarding the types of stations that are the most likely to experience delays. Finch and Vaughan MC, as well as Kennedy and Kipling are the terminal stations for the Yonge-University line and the Bloor-Danforth line, respectively. For each of these lines, both sets of terminal stations were included in the list of top three most delayed stations (noted in Table 1). This suggests that subway delays are much more likely to occur at the end of the line. Based on the data, we can also note that the terminal stations are more likely to experience operator related related issues. The terminal stations Vaughan Metropolitan Centre and Finch, for example, experience more delays due to operator error when compared to Rosedale, a non-terminal station. The same is true for stations on the Bloor-Danforth line. One could speculate that operators might be more prone to errors towards the beginning or end of their shifts, i.e. when trains are at either the beginning or end of the line.

In order to improve upon the current subway delay crisis, having access to complete and reliable data is

[3]Table 1 was created using the `kableExtra` package (Zhu 2021).

crucial in order to better understand the root of the problem and implement effective solutions. While this data does offer a number of valuable insights, inconsistencies in reporting standards and biases do hinder the reliability of the data. Regardless, the available information still offers numerous key insights into the prevalence of subway delays throughout Toronto and can be used as an effective starting point in identifying ways to address these issues.

Bibliography

City of Toronto. 2020. *Toronto Employment Survey* <https://www.toronto.ca/city-government/data-research-maps/research-reports/planning-development/toronto-employment-survey/#

City of Toronto Archives. 1954. *Canada's First Subway: Open for Business* https://www.toronto.ca/explore-enjoy/history-art-culture/online-exhibits/web-exhibits/web-exhibits-transportation/canadas-first-subway/canadas-first-subway-open-for-business/

Fox, Chris. 2016. *Downtown population will nearly double by 2041 amid building and baby boom.* https://www.cp24.com/news/downtown-population-will-nearly-double-by-2041-amid-building-and-baby-boom-keesmaat-1.2846605

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal.* https://CRAN.Rproject.org/package=opendatatoronto

Nowlan, David M, and Greg Stewart. 2007. "Downtown Population Growth and Commuting Trips: Recent Experience in Toronto." *Journal of the American Planning Association* 57(2): 165-182.

Ouellet, Valerie. 2017. *66% of subway delays are caused by passengers, CBC Toronto data analysis shows.* https://www.cbc.ca/news/canada/toronto/ttc-subway-delays-1.4068358>

Pederson, Thomas Lin. 2020. *patchwork: The Composer of Plots.* https://CRAN.R-project.org/package=patchwork

R Core Team. 2021. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spurr, Ben. 2016. *TTC forging ahead with one-operator subway trains.* https://www.thestar.com/news/gta/2016/10/02/ttc-forging-ahead-with-one-operator-subway-trains.html?rf

Toronto Open Data. 2021. "TTC Subway Delay Data." https://open.toronto.ca/dataset/ttc-subway-delay-data/

Toronto Transit Commission, 2019. *2019 Operating Statistics.* https://www.ttc.ca/transparency-and-accountability/Operating-Statistics/Operating-Statistics---2019

Toronto Transit Commission. 2022. *TTC Service Details.* https://www.ttc.ca/customer-service/TTC-Service-Details

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

Wickham, Hadley. 2020. *scales: Scale Functions for Visualization.* https://CRAN.R-project.org/package=scales

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. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *bookdown: Authoring Books and Technical Documents with R Markdown.* https://CRAN.R-project.org/package=bookdown

Xie, Yihui. 2021. *knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://CRAN.R-project.org/package=knitr

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