Rush hour subways surprisingly excel at punctuality: a closer look at TTC Delays*

Insight into when delays disproportionately tend to happen and related statistics

Andrew Goh

September 27, 2024

This paper uses data from Open Data Toronto on TTC subway and bus delays to analyze how the delays are generally distributed w.r.t. the time of day. Mean, total, and adjusted statistics were graphed, leading to the conclusions that subways are almost always more punctual than buses, peak hour leads to the fewest delays, and leaving 10-15 minutes earlier is a safe margin of error for delays. The conclusions of this paper provide a practical, actionable set of guidelines on how citizens can ensure/maximize punctuality when using the TTC to go about the city, as well as potential insight into where resources can be invested in to increase TTC efficiency the most.

1 Introduction

The TTC, as Toronto's primary public transport network, is how 1.2 million people (Toronto Transit Commission (2024a)) get about the city every day. With its network of subways, the average TTC rider will be less affected by traffic than the average driver is, but unfortunately, delays act as public transport's traffic counterpart. And while individuals have their own ways to consistently remain punctual, there remains a lack of objective analysis regarding TTC delays. To fill this gap, we look into TTC data to determine just when and how delays happen, as well as how individuals can best make use of this information.

Data on TTC delays for both subways and buses was obtained from opendatatoronto, then analysed and graphed. The delays were graphed both by mean delay time and total delay time (minutes) against the time of day (hour) to more intuitively see how the two explanatory variables, namely the type of transportation (subway/bus) and time of day (hour), affected

*Code and data are available at: https://github.com/mushroomcarbon/TTCDelays.

various aspects of the delay experienced. Data was then adjusted for the difference in subway frequencies during peak vs. off hours and then graphed.

From the graphs, we conclude that the mean delay times of buses is significantly higher than that of subways, whereas both transportation types have similar mean delays w.r.t. the time of day. Total delay times are higher during peak hours, but when accounting for the ~doubled amount of subways running during that time, peak hour subways actually end up being the least delayed ones on average. The significance of insight into public transportation delays lies both individually and socially: individuals can learn to better plan their trips to remain punctual in the event of delays, and considering how TTC delays can impact productivity as a whole on a large scale, governments and corporations alike can utilise this information to better manage resources to increase efficiency.

The paper begins by introducing and viewing the dataset used, as well as analysing collection methodologies, larger contexts, as well as how the data was processed in Section 2. We then continue by graphing the data and discussing some preliminary insights regarding the visuals. In Section 3, we discuss the results of the data in a more in-depth fashion, as well as highlighting some key insights extrapolated from the previous analysis. The section then concludes by acknowledging the limitations of the study and discussing ways to further the research.

2 Data

The R programming language (R Core Team (2023)), dplyr (Wickham et al. (2023)), opendatatoronto(Gelfand (2022)), and tidyverse(Wickham et al. (2019)) were used to download, modify, and analyse data obtained from Open Data Toronto. Styler (Müller and Walthert (2023)) was used to style the code. Data used in this paper comes from the "TTC Subway Delays" (Toronto Transit Commission (2024d)) and "TTC Bus Delays" (Toronto Transit Commission (2024c)) data sets from the Open Data Toronto database, published by the TTC.

The two datasets are similar, providing the same key information: when delays happened (date and time), and how long they were delayed by. Cleaning the data involved removing N/A rows and rows where the delay was 0, then adding a new variable for the hour the delay happened in simply by extracting the hour from the more detailed time information provided. This sums up to 4 variables in the analysis dataset - Date, Time, and Hour describe when the delay happened, and Min Delay describes how long the delay was for, i.e. the difference between the time that the vehicle arrived and the time that the vehicle should have arrived. Table 1 provides a short sample of the cleaned data, which is in the same format for both subway and bus datasets. Raw data for both datasets contained many more descriptors which were not used in this paper and is therefore shown in the Appendix in Table 2 and Table 3.

Table 1: Sample of cleaned analysis data

Time	Date	Min. Delay	Hour
5H 41M 0S	2022-01-01	4	5
6H 3M 0S	2022-01-01	3	6
6H 6M 0S	2022-01-01	10	6
6H 12M 0S	2022-01-01	11	6
6H 17M 0S	2022-01-01	38	6

Data was taken from 2022-2023 due to the recency - data before 2021 was left out due to COVID-19 affecting statistics and data before COVID too dated for either to be relevant to 2024. There are no similar datasets that could have been used to analyze TTC delays - other TTC datasets don't describe the same thing and therefore can't be used either.

The data was collected via the CIS (communications information system), the TTC's vehicle monitoring system ("Methodology for Analysis of TTC's Vehicle Tracking Data" (2024)). This is done via GPS tracking on latitude-longitude data updated every 20 seconds, with erroneous data being autocorrected by the system. The TTC's method of measurement is reliable due to its high accuracy and error-tolerant nature, lending to the dataset's strong credibility.

To have a broad understanding of the data, we first graph the summary statistics: Figure 1 shows the mean delay w.r.t. the hour, and Figure 2 shows the standard deviations.

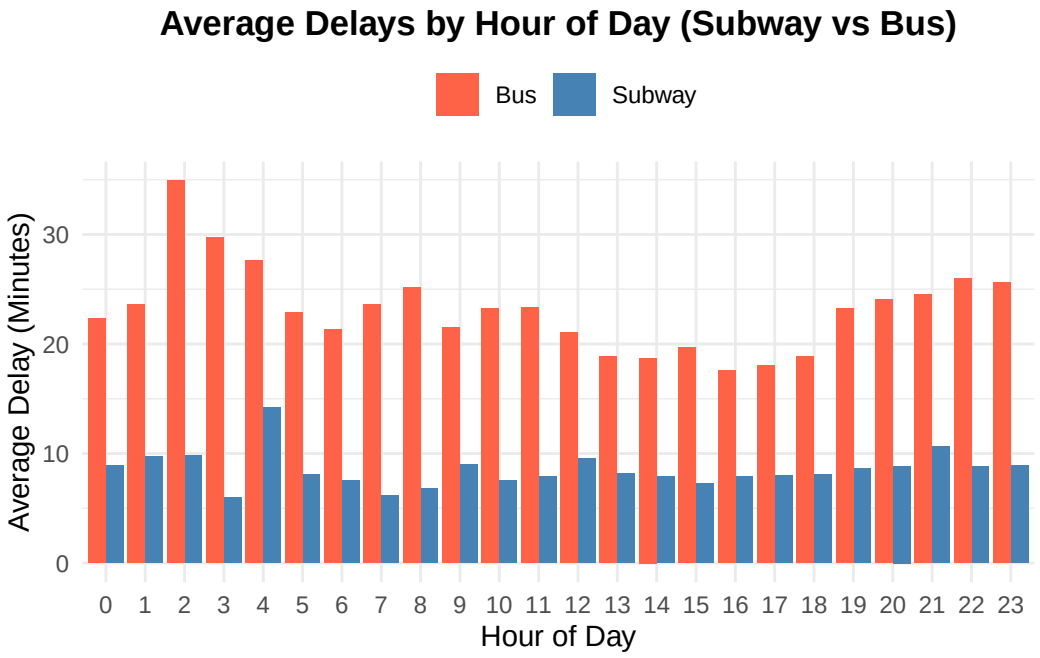

Figure 1: Subway delays w.r.t. time of day

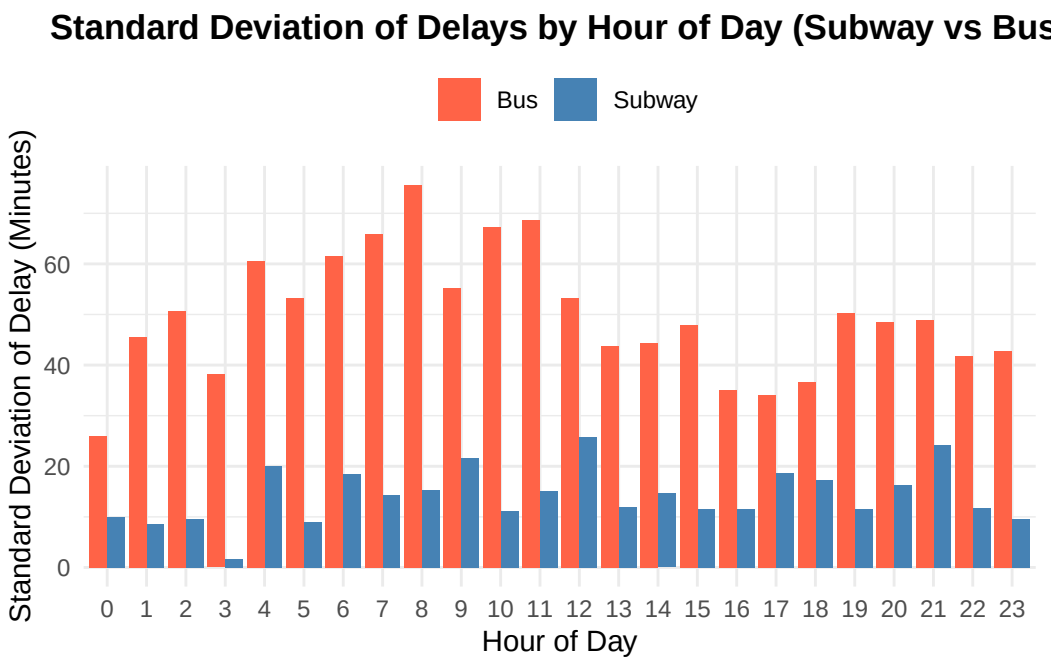

Figure 2: Standard deviation of subway and bus delays w.r.t. time of day

From the mean graph (Figure 1) we can see that 1. the average magnitude of delays is generally homogeneous within transportation categories, and that 2. the average delay times for buses is around 4-5 times more than that of subways. This confirms the intuition that subways are generally more time-reliable than buses, and we will therefore focus on subways from hereonafter to find the ideal, delay-minimising way of transport. The standard deviation graph (Figure 2) shows that the standard deviation of both bus and subway delays are all larger than the means of the respective measurements. This suggests that the mean itself, while still an accurate estimate of typical delays, holds less practical significance as a benchmark for normalisation (i.e. leaving x minutes early to offset delay).

While there is no data regarding the total number of subway trips and therefore cannot calculate an "average delay per ride" statistic, it is still possible to sum up the delay at each hour of day to obtain an estimator for the relative amount of delay encountered during each respective hour, as shown in Figure 3.

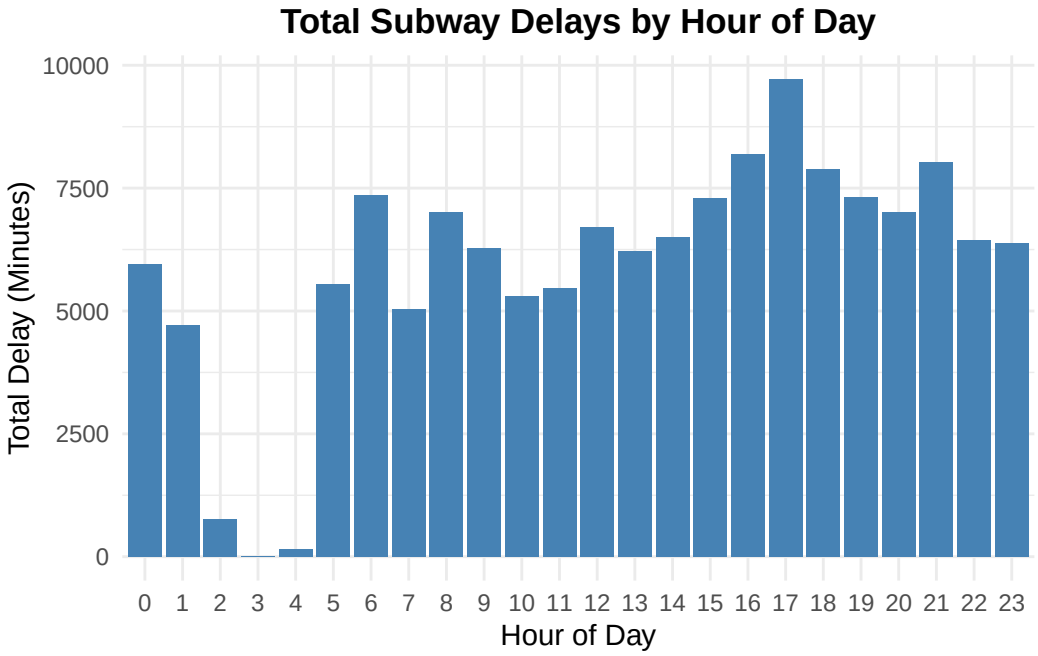

Figure 3: Total delay times of the TTC subway system per hour

The data in Figure 3 singles out 6-9am and 4-6pm to be the worst offenders, suggesting the conclusion that peak hour riders are the most susceptible to TTC-induced tardiness. According to the Toronto Transit Commission (2024e), though, subways run at approximately double the frequency during these rush hours, accounting for the increase in total delay times by sheer volume. Adjusting for this by halving the relative delay times for the aforementioned rush hours (6-9am, 3-7pm), we obtain Figure 4, the adjusted total delay graph.

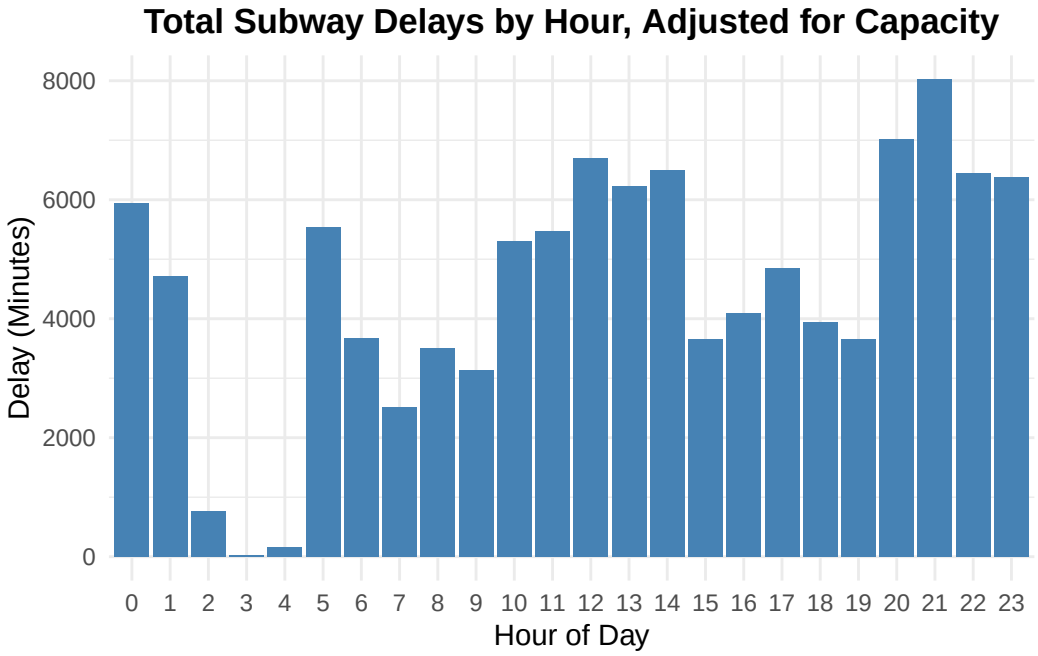

Figure 4: Delay time adjusted for the approximate subway load

3 Discussion

Looking at Figure 4, peak hours are, surprisingly, the least affected by delays on average. Combining this with previous conclusions, we get that peak-hour subways are the most time-reliable vehicles of the TTC. A potential reason for this could be the higher-stress situation leading to tighter schedules, the lack of shift changes, or even the general alertness of drivers, among others - this, however, is not the focus nor concern of our current investigation.

The simple conclusion of this is that students and workers regularly making use of the peak hour subways have no need to change their TTC usage - individuals making more flexible plans, however, are incentivised to head out during peak hours instead of avoiding them, if speed is the key concern.

A key implication of this that ties into the broader context is that with the limited resources of the TTC (Toronto Transit Commission (2024b)), focusing more resources into peak hours rather than spreading them out evenly is the most economic decision, barring other considerations: peak hour subways manage the highest efficiency of all lines currently, whether it be with respect to passenger load or time to arrival; therefore, increasing the frequency of peak-hour subways would do the most to reduce passengers' times in transit and increase the city's functionality as a whole.

A key question that remains, however, is how individuals can best use this data to inform their own travels. As mentioned in Section 2, a high standard deviation suggests that a simple

mean offset is insufficient/nonideal to offset the effect of delays. Considering how delays are measured, however, notice how a single subway would rack up delays as it operates - if its arrival at station 1 is delayed by 5 minutes, for example, its successive arrivals to future stations will all also reflect this "carried-over" delay. Subway delays, therefore, follow a roughly monotononically increasing pattern, which would largely explain the disproportionately large standard deviation.

To better estimate the average delay amount given that a delay happens, we look at average ridership data from "The 2041 Regional Transportation Plan" (2024) and "What Would Fare by Distance Mean for Toronto?" (2024), stating that "local" trips take around 4.2km on average. Average distance between stations, calculated using data from the Toronto Transit Commission (2024b), is 70 stations / 70.1km for approximately 1km per station (on average) and a resulting average of around 4 stations per trip.

Accounting for the previously mentioned carry-over delay for the specific subway arriving at each of the 4 stations, we conclude that for the average trip during peak hour, leaving 10-15 minutes early is a fairly safe margin to ensure punctual arrival even in the case of delays, scaling this value as fit to account for extra/fewer stations (leave a 20-30 minutes margin for an 8-station trip, for example).

4 Limitations and further research

A limitation of this paper is how, as mentioned in Section 2, there is a lack of information on the absolute (as opposed to relative) frequency of subway trips. This means that average-case scenarios can only be calculated in the conditional case (i.e. "given that a delay happens"), rather than factoring for the total probability of delays happening. A further limitation is the lack of the most recent data - data from 2024 was omitted to avoid introducing bias to the model due to the year, as the data from 2024 is missing data from the final three months for obvious reasons. A repeat of the study, conducted on the complete 2024 dataset after september, could better reflect recent trends and patterns.

Future studies could focus on the omitted columns from the raw data, analysing how factors such as the station where delays happen, model of bus, and direction of travel affect the magnitude of delays. Adding more variables to the analysis would likely improve the insights obtained from the data - fitting a statistical model (ANOVA, KMeans, etc.) could also be a potential route to take to obtain some more precise, methodical conclusions.

5 Appendix

Table 2: Sample of raw data

Date	Time	Day	Station	Code
2022-01-01T00:00:00Z	15:59	Saturday	LAWRENCE EAST STATION	SRDP
2022-01-01T00:00:00Z	02:23	Saturday	SPADINA BD STATION	MUIS
2022-01-01T00:00:00Z	22:00	Saturday	KENNEDY SRT STATION TO	MRO
2022-01-01T00:00:00Z	02:28	Saturday	VAUGHAN MC STATION	MUIS
2022-01-01T00:00:00Z	02:34	Saturday	EGLINTON STATION	MUATC
2022-01-01T00:00:00Z	05:40	Saturday	QUEEN STATION	MUNCA
2022-01-01T00:00:00Z	06:56	Saturday	DAVISVILLE STATION	MUNCA
2022-01-01T00:00:00Z	06:58	Saturday	ST PATRICK STATION	MUNCA

Table 3: Sample of raw data (continued)

Min Delay	Min Gap	Bound	Line	Vehicle
0	0	N	SRT	3023
0	0	NA	BD	0
0	0	NA	SRT	0
0	0	NA	YU	0
0	0	S	YU	5981
0	0	NA	YU	0
0	0	NA	YU	0
0	0	NA	YU	0

References

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

"Methodology for Analysis of TTC's Vehicle Tracking Data." 2024. https://stevemunro.ca/methodology-for-analysis-of-ttcs-vehicle-tracking-data/.

Müller, Kirill, and Lorenz Walthert. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"The 2041 Regional Transportation Plan." 2024. https://www.metrolinx.com/en/projects-and-programs/regional-transportation-plan.

Toronto Transit Commission. 2024a. "2023 Operating Statistics." https://www.ttc.ca/transparency-and-accountability/Operating-Statistics/Operating-Statistics---2023.

———. 2024b. "2023 Operating Statistics - Conventional System." https://www.ttc.ca/transparency-and-accountability/Operating-Statistics/Operating-Statistics---2023/Conventional-System.

———. 2024c. *Open Data Toronto: TTC Bus Delays*. https://open.toronto.ca/dataset/ttc-bus-delay-data/.

———. 2024d. *Open Data Toronto: TTC Subway Delays*. https://open.toronto.ca/dataset/ttc-subway-delay-data/.

———. 2024e. *Subway Line 1 (Yonge-University) Schedule*. https://www.ttc.ca/routes-and-schedules/1/0/15664.

"What Would Fare by Distance Mean for Toronto?" 2024. https://stevemunro.ca/2016/02/14/what-would-fare-by-distance-mean-for-toronto/.

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, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.