Analysis of ttc-subway-delay in September,2021*

Vincent Yu

02/06/2022

Abstract

Toronto subway is one of the most effective subway systems in the world. However, due to its age and other reasons, the amount of time it delays is dramatic. TTC subway delay data is a public data posted by the Toronto Transit Commission on The City of Toronto's Open Data Portal for everyone to consume and analyze from 2017. From this report, I generate that the highest delay frequency happened on line1, Vaughan mc station, southbound and Thursday. "No Operator Immediately Available" is the highest reason why it was delayed provided by TTC. 5 minutes spare is enough for a common delay and Saturday is the best day to take subway. I hope this report could help passengers save time and the TTC company be aware of the delay's reason.

Introduction

As we all know, Toronto has one of the oldest subway systems globally and is run by the Toronto Transit Commission (TTC). TTC is the public transport agency that operates most of the public transit in the Toronto, Peel and York regions. It was established 101 years ago and it is the largest urban mass transit system in Canada and the third-largest in North America till now. In 1954, the TTC opened Canada's first underground rail line, Line1. At that moment, the total length of the route was only 7.4 kilometers. The line is also called Line-Yonge–University because it was under Yonge Street and the line was southwards to Union station under University Avenue. As of 2018, the whole network has four lines, 75 stations and 76.9 kilometers of route. Among them, Line 1 is still the longest route. After the latest extension from Sheppard West to Vaughan Metropolitan Center, its length became 38.8km. In general, Line 1 has two northern terminals with a U shape. Line 2 is nearly a straight line going east-west. The other 2 are extended lines connected on these two main routes. (Wikimedia_Foundation (2022)). However, as Toronto citizens, we all know that many questions happen every day on the subway. Although the staff tries to solve them as quickly as possible, the subway still can not arrive on time and ruins daily life frequently. Because of the vital role of public transit, the public transit service delay could cause serious results for the urban mass transit system and citizens' life.

In general, this report will provide frequency and reason for subway delays in everybody's daily life. For this report, I will use open-access data from the Toronto Transit Commission to analyze data and get some more information about the TTC Subway delay. First of all, I will introduce the methodology and data collection. Later, the 3 main topic is the number, reason and days of delays. There will be 2 kinds of summaries later: Numerical and Graphical, and it will start with a figure of the delay. Then, I will analyze some other relevance and get: the highest delay frequency happened on line1, Vaughan mc station, southbound and Thursday. "No Operator Immediately Available" is the highest reason why it was delayed provided by TTC. 5 minutes spare is enough for a typical delay and Saturday is the best day to take the subway.

The dataset will be processed and analyzed in R(R Core Team (2021)). The tidyverse (Hadley Wickham (2019)) dplyr (Wickham et al. (2021)) and Car (Fox and Weisberg (2019)) packages are used to generate

*Code and data are available at:https://github.com/YN7666/sta304-

data. Figures and tables will be created by ggplot2 (Wickham (2016)) and kableextra (Zhu (2021)). The packages knitr (Xie (2021a)), and tinytex (Xie (2021b)) and rmarkdown(Allaire et al. (2021)) are used to generate the R markdown report.

Data

Data Collection Process

This dataset is provided by The City of Toronto's Open Data Portal. The City of Toronto Open Data Portal is an open-source delivery tool to provide and collect Toronto's data with people's lives together. This is proposed and maintained by the City of Toronto government. The TTC had posted some data on the City of Toronto's Open Data site for anyone to consume and analyze since 2017. To better understand TTC Subway & SRT Train Service Delay is the primary purpose of this data analysis. The report is supported by the R package opendatatoronto (Gelfand (2020)) and the dataset was last updated on 26th October 2021 and openly available to the public at any time.
The dataset contains information of TTC Subway & SRT Train Service Delay in September-2021. One of the reasons I choose this one is the latest TTC dataset on the Open Data Portal is this one. The delay counts were calculated by reviewing all of the TTC Alerts for service interruptions on the four subways. A TTC Alert is the information about the current status of TTC service including Line & route alerts, Accessibility alerts and General alerts. In this case, Line & route alerts are counted for the event of a TTC Subway service delay. However, there is specific instruction on how they collect the alert. I assume it is by both automatically and manually. There is an article that pointed out the reason why the delay happened in Hongkong may be the following factors: power cable failure, signal cable failure, turnout communication disruption and crashes involving a casualty. Also, a longer subway operation incident delay is higher. (JinxianWeng (2014)). However, since we are far more north than Hongkong the reason may be different, and here we only consider the most simple ones. Besides, the CBC has done an analysis of major reasons that caused TTC subway delay 6 years ago. It concludes that the most common reason is caused by passengers. In 2016, 66 percent of delays were caused by emergencies and misbehavior triggered by passengers(CBC_News (2017)). Last but not least, due to the covid, the day has some difference compared to the previous day.
But, We all know some delays never make it to an alert. One of the problems is: some drivers may lie about the delay reason since they could get fired.Also, studies of transit service disruptions duration are also less than highway delay concerning inspecting the effects of non-causal variables on the delay duration (JacobLouie (2017)). Less attention means less investment which could lead to inattentive personnel and equipment in poor condition and aging and finally cause misconduct in the alert. Last but not least, due to the technology issue, some period of the data is missing or miss counted for example delay in 1 min is hard to detect. This is not some issue we can fix. This problem will make the average length of delay underfitted So the delay referred by this data is not accurate enough. All in all, this is the best we have for published information right now.

Data Summary

This dataset is called TTC Subway & SRT Train Service Delay in September-2021. The total population of the dataset is 1433 observations, which represent the overall Subway delay in the last September. There will be ten variables Date, Time, Day, Station, Code, Min Delay, Min Gap, Bound, Line, vehicle. The first variable is the time variable and the other 2 are categorical variables that indicate when the subway is late. Station means the subway is delayed to arrive at this station. Code is the reason why the subway is delayed. By matching the codes to a TTC document "ttc-subway-delay_codes," we can translate it into sentences. For example, MUSC - Miscellaneous Speed Control, MUPAA - Passenger Assistance Alarm Activated - No Trouble Found, TUSC - Operator Overspeeding and SUDP - Disorderly Patron. min_delay and min_gap are the duration and the frequency of delay count by minutes. Bound and line are two common indicators

of the subway. the vehicle is the number representing each train.
The very first step here is data cleaning. It is necessary for us to clean the data in order to make it cleaner, more accurate and more precise. Filter and rename are the two steps here. Filter every missing value(NA) in every column in the data. Rename `Min Delay` into min_Delay and `Min Gap` into min_Gap to avoid the ugly comma.
After these steps, there are 1104 data left. A sample view of the dataset is displayed below.

Table 1: ttc-subway-delay in September,2021

Date	Time	Day	Station	Code	min_delay	min_gap	Bound	Line	Vehicle
2021-09-01	06:32	Wednesday	NORTH YORK CTR STATION	EUSC	0	0	S	YU	5726
2021-09-01	08:08	Wednesday	ROSEDALE STATION	MUSC	3	6	N	YU	5846
2021-09-01	09:26	Wednesday	GREENWOOD STATION	MUD	3	7	W	BD	5253
2021-09-01	11:10	Wednesday	SPADINA BD STATION	MUDD	3	7	W	BD	5255
2021-09-01	11:28	Wednesday	DUNDAS STATION	MUPAA	0	0	S	YU	5776
2021-09-01	12:17	Wednesday	ROSEDALE STATION	MUSC	0	0	N	YU	5466

Number of TTC Subway Delays

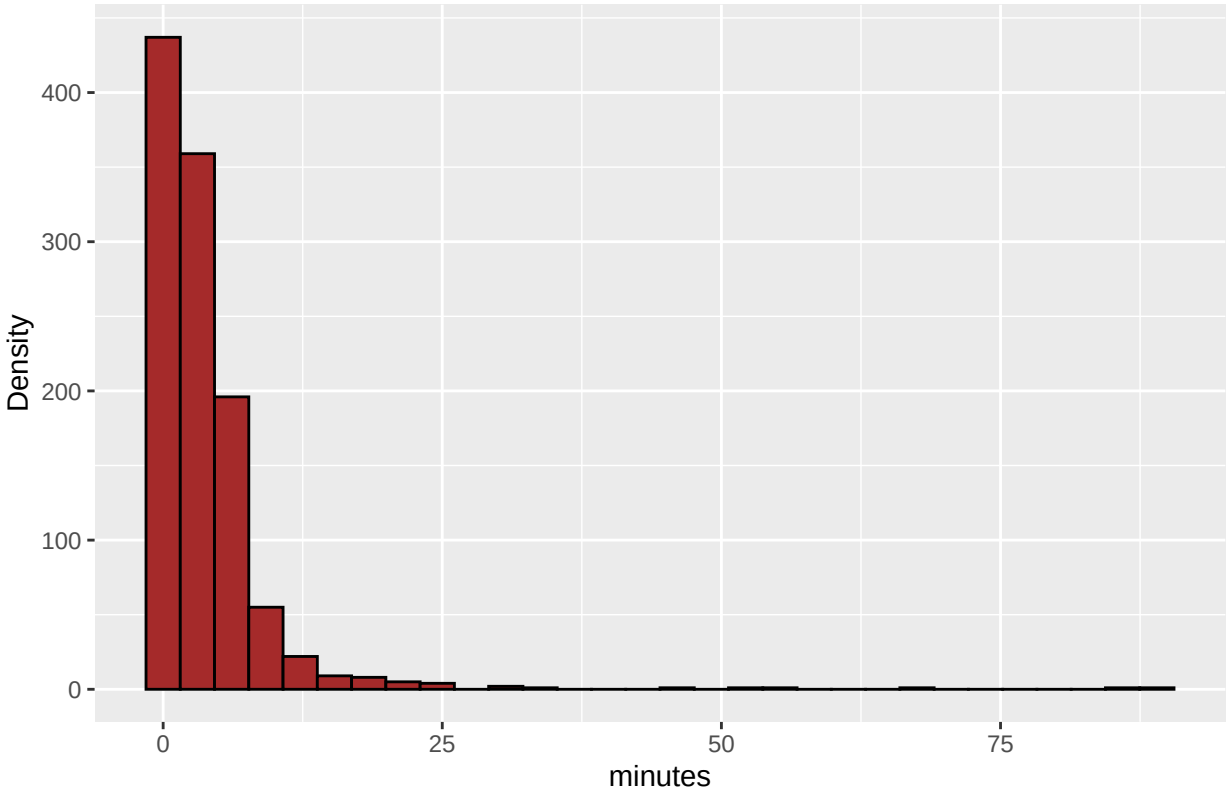
Here I will analyze some interesting variables which could give a clear view of how each variable is constituted. In this report, there will be three essential and interesting variables that I will mainly focus on: number, reason and days of delays. The most important variable is Min_Delay. This is used to measure the duration of each delay. Min_Delay. It is a numerical variable. min_gap is the time gap between two delay events. This is also a numerical variable that could represent the min_delay from the side. Table 2 demonstrates the data summaries of the Min_Delay.

Table 2: delay in September

total_delay	ave_delay	ave_gap	sd_ave_delay	Q1	median	Q3	Large_Outliers	max
1104	3.575181	6.008152	6.243311	0	3	5	39	89

First of all, the total number of observations in this database is 1104. On the basis of this summary, the average time of delayed subways in September 2020 is about 3.6 minutes, and the average time gap between two delay events in September 2020 is about 6 minutes. Q1 is the value of 25% data points of the total variables and Q3 is the value of 75% data points of the total variables. Since we know the Q1 is 0 and time could not be less than 0. We get the lowest amount of delayed time is 0 and there is no small outlier. The most significant delay is 89 minutes. Thus, the range of this data is from 0 minutes to 89 minutes. The median is 3 minutes which is a little smaller than the mean value. This means most of the delay data are gathered together. Besides, Q3 is 5 minutes long. So 75 % of the data is under 5 minutes. The standard deviation of the time subway delay is about 6.2, which is a moderate number. This means variables are not tightly clustered around the mean. The large outlier is calculated by Q3 + 1.5 times the difference between Q3 and Q1. Then we get 39 outliers on the right tails from the total 1104 variables. A clear view of the variable delay is here.

Fig.1: Number of TTC Subway Delays by minutes in September 2021



I also make a histogram to show the general distribution of the delay. This could provide a better exploration of the data. Based on figure 1, we can see the aspect of the distribution is right-skewed. Most of the data are assembled between 0 -25mins. There is a long tail on the right of the figure. Most of the bins are 0 and only a few of them have values like 1or 2. The highest bin is at 0 minutes. This is because every delay under 3 minutes is considered as 0 minutes. So, we can conclude that over 400 delays are under 3 minutes instead of 0 minutes. There are approximately 350 delays of around 3 minutes. Only 0 min to 12 mins have more than 50 times delays, which is significant. This graph proved the 3 minutes median and 3.6 minutes mean we get from table 2 is reasonable. All in all, whenever you want to take subway to go out, usually only 5 minutes are set aside and at most 25 minutes in total.

Table 3: general delay by line or bound or station

Bound	n	Line	n	Station	n
S	443	YU	696	VAUGHAN MC STATION	171
N	267	BD	328	ROSEDALE STATION	44
E	222	SHP	50	FINCH STATION	42
W	159	SRT	26	BLOOR STATION	34
B	13	YUS	4	WILSON STATION	29

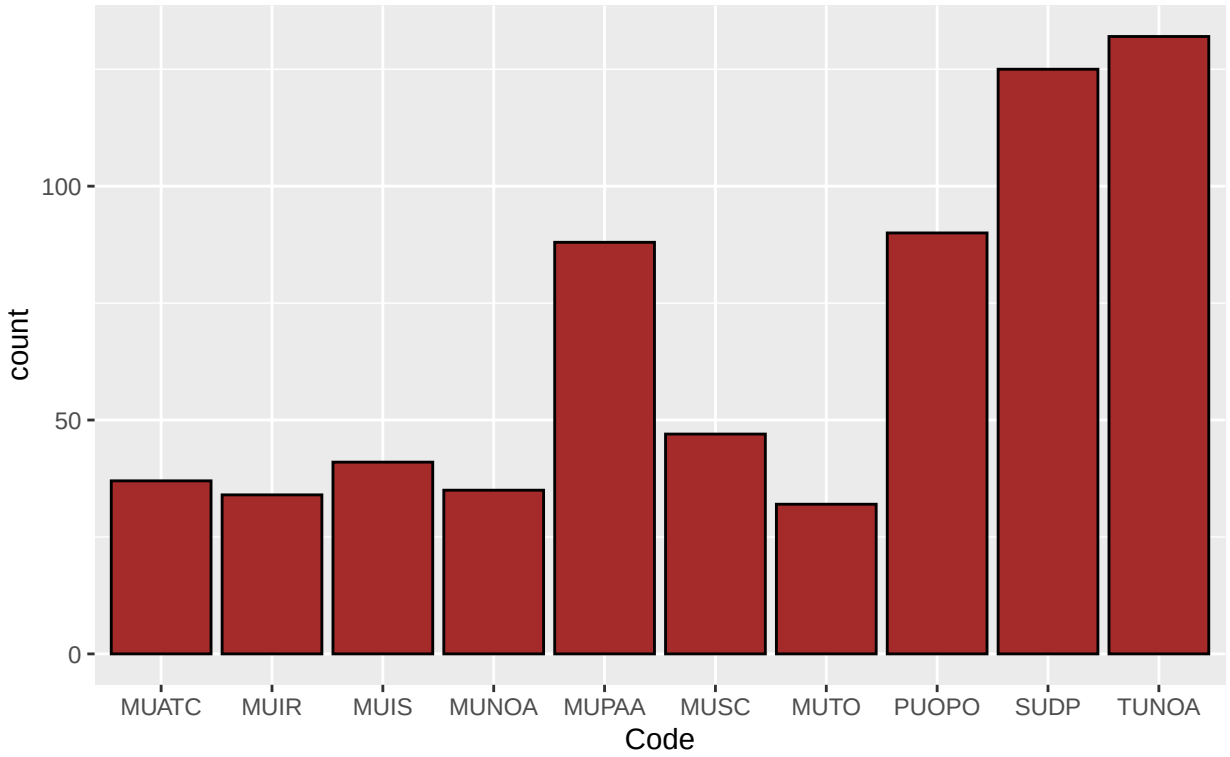
Here are several other variables in table 3. This dataset captures mostly categorical values such as station, bound, vehicle number and delay code. So here I list a table below, which contains all of them with the highest frequency of occurrence in the data. Bound is a categorical variable that indicates the direction of each subway. “SWEN” represent south, north, east, and west. However, instead of recording the direction of the end of the subway passing through this station, they consider the direction of this station only. Here, a southbound station could appear on an east-west line. B is the bound at Spadina, which is a particular

case. There are 4 lines in this city as well. Among all 5 lines listed here, line Yu and line Yus are parts of line 1. So we do have 4 lines in total instead of 5. Combining these 2 data, we can reach many conclusions. First, We can see that the southbound has the highest frequency: 443 and it is the direction to downtown. And, the opposite direction northbound has the second-highest frequency, 267. We also know that Line1 has 696 + 4, which is 700 times of delays which is almost the same as the sum north-south direction. This is reasonable since line 1 is in a north-south direction and nearly none of the others have a north-south bound. Line 1 is also the longest line, which is sensible to have the highest times of delays (JinxianWeng (2014)). In the station part, I list the 5 popular stations with the highest delay times by sorting the delay from highest to lowest. The highest one is Vaughan mc station. 171 times the Subway delayed arriving at this station in that month. The other thing is, all 5 stations are from line 1 as well. So, line1 has the highest frequency of delay.

Reason of TTC Subway Delays

On the other hand, I would like to study the relationship between the reason why the subway is delayed(Code) and the number of delays it triggered. Due to the code being a categorical variable, in figure 2, we used a bar plot to show the top 10 reasons TTC Subway Delays in September 2021. The top 10 reasons contain 661 records which are 60% of the total data. Figure 2 is ordered by the alphabet. We can see the highest bin is "TUNOA" and "SUDP" is also over one hundred. The lowest bin of the top ten is "MUTO."

Fig.2: Top 10 reason TTC Subway Delays in September 2021



After matching the codes to the TTC document "ttc-subway-delay_codes"I generate the top 5 reasons and the number occurs here:

Code	reason	frequency
TUNOA	No Operator Immediately Available	numbers of occurs: 132
SUDP	Disorderly Patron	numbers of occurs: 125
PUOPO	OPTO (COMMS) Train Door Monitoring	numbers of occurs: 90
MUPAA	Passenger Assistance Alarm Activated - No Trouble Found	numbers of occurs: 88
MUSC	Miscellaneous Speed Control	numbers of occurs: 47

We could learn that "No Operator Immediately Available" is the most common error here. The top 4 reasons have a frequency of around 100 times. Meanwhile, the others are under 50. The graph shows a clear view that they have a considerable gap with the 5th one: "Miscellaneous Speed Control" and all other reasons. Compared to 2016, I think the main reason sticking on Toronto's subway is not caused by passengers anymore since the first, second and fifth reasons are not their fault directly.(CBC_News (2017))

Days of TTC Subway Delays

Table 5: delay on each day

Day	n
Thursday	196
Sunday	195
Wednesday	168
Tuesday	160
Friday	148
Monday	133
Saturday	104

As I mentioned above, the last one of the three important and interesting variables is the day. Day is a variable that represents the subways have more delays on which day of the week? In this section, I will also do both Graphical Summaries and Numerical Summaries to represent the number of TTC Subway Delays by minutes in September 2021 in the different week of days. From table 5, Thursday has the highest number of delays amoung all 7 days: 196. This could be caused not only because it is a busy day, but also by Wednesday and Thursday being repeated 5 times while others only had 4 times in September.From the graph, The Mean of delay is 3.6 minutes with an IQR from 0 to 5 minutes on Thursday. The longest delay is 33 minutes. Saturday has the lowest number of delays: 104. The Mean of delay is 2.6 minutes with an IQR from 0 to 5 minutes. The longest delay is 16 minutes which is also the lowest outlier. Besides, all of the boxes have a left bound equal to 0 and some outliers on the right tail. The most significant outlier is 89 minutes, which is on Friday. The lowest boxes(IQR) are Monday, Tuesday, and Wednesday: from 0 to 4 minutes. All the others are 0 to 5 minutes. The median of most boxes is 3 minutes, except Saturday, which is 0.

Fig.3: Number of TTC Subway Delays by minutes in September 2021 in different days

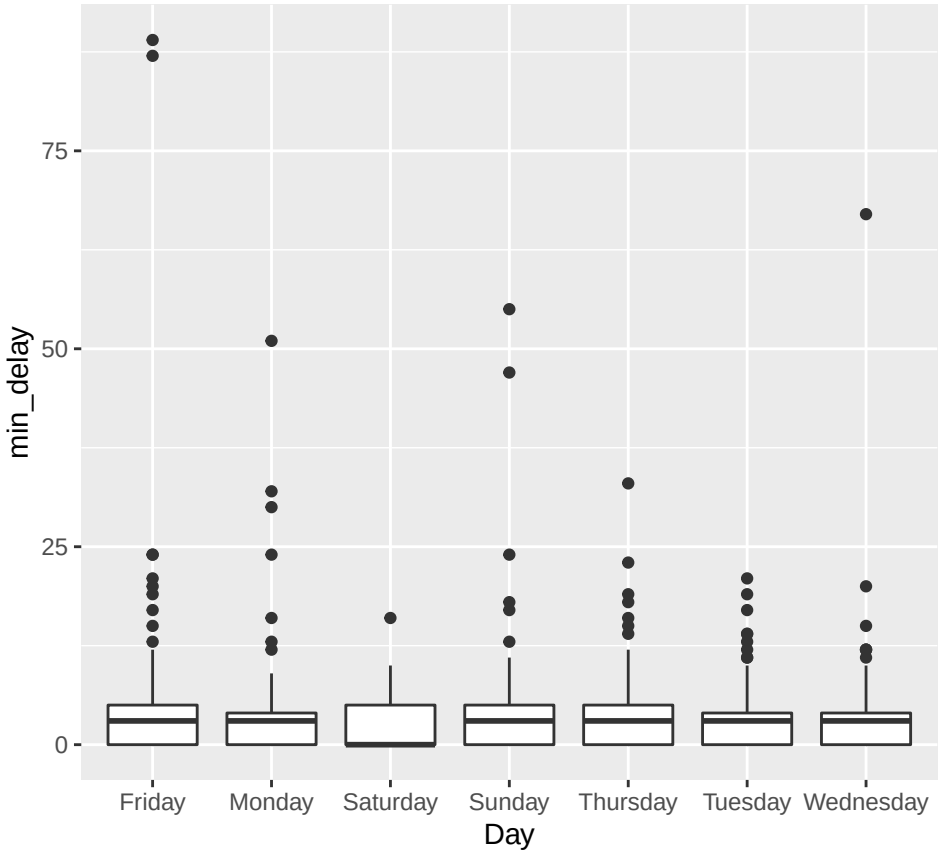


Table 6: outstanding delay times on every day

Day	delay_times	Day	delay_times	Day	delay_times
Monday	7	Tuesday	9	Wednesday	7

Day	delay_times	Day	delay_times	Day	delay_times	Day	delay_times
Thursday	7	Friday	10	Saturday	1	Sunday	6

Table 6 shows the outstanding delay times, which refers to the outliers that happened every different day. The outliers are calculated by the same method above. By table 6 and the plot, Saturday also has the lowest outstanding delay times: 1. Even though Friday has only 148 delay times, it has the highest rate(10/148) that something terrible happened and cost a much longer time than others. So my suggestion is: if you have any place to go by subway, Saturday is the best choice to schedule.

Reference

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

CBC_News. 2017. "66." https://www.cbc.ca/news/canada/toronto/ttc-subway-delays-1.4068358.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

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

Hadley Wickham, Jennifer Bryan, Mara Averick. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

JacobLouie, Shalaby, M.Eng.sAmer. 2017. "Modelling the Impact of Causal and Non Causal Factors on Disruption Duration for Toronto Subway System: An Exploratory Investigation Using Hazard Modelling." https://www.sciencedirect.com/science/article/abs/pii/S0001457516303694.

JinxianWeng, Xuedong, YangZheng. 2014. "Development of a Subway Operation Incident Delay Model Using Accelerated Failure Time Approaches." https://www.sciencedirect.com/science/article/abs/pii/S0001457514002322.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wikimedia_Foundation. 2022. "Toronto Subway." https://en.wikipedia.org/wiki/Toronto_Transit_Commission.

Xie, Yihui. 2021a. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2021b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. https://github.com/yihui/tinytex.

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