Toronto Transit Commission (TTC) Delay Analysis

Statistical Analysis on the Frequency and Duration of TTC Subway Delays

Group 15:

Alagoa Ayebainasuoton

Arpana Sriram

Maria Rodriguez

Sathya Krishnakumar

Timothy Lam

Timothy Leung

Submitted for: University of Toronto: 3250 Foundations of Data Science - Group Assignment

December 13, 2020

Table of Contents

Executive Summary

It is believed that train delay prediction systems, in general, do not fully exploit historical data about train information. Additionally, traditional railway infrastructure systems that rely on static rules are believed to be using classic statistical algorithms. Delayed trains affect commuters all over the world and could be caused by numerous factors that are interrelated, making it hard to assess the effects and devise solutions, using traditional methods.

Today, however, in transportation industry, data scientists employ innovative techniques to identify root cause of the issue and can be applied to help transit commissions improve their services by enabling intelligent decision-making for minimizing train delays.

The purpose of our project is to build a train delay prediction assessment that exploits data science techniques. We have attempted to derive a conclusion based on a data-driven model, based primarily on the historical data about train movements along with various related information collected over few years, by Toronto Transit Commission (TTC).

We utilized the City of Toronto's open data portal to extract the data sets relating to TTC subway trains for 3 complete years. The dataset documented TTC delays on various attributes including: station affected, delay code, duration of delay, duration between trains, and subway line affected. This allowed us to dissect the data on various dimensions to investigate many questions including: 1) the pattern of delays on a yearly basis 2) impact of time (e.g., month, day of the week) on delay duration and type 3) subway lines experiencing the most delays. To accurately address these questions, we created new variables, imputed missing values and/or inconsistent entries, and dropped irrelevant information from the dataset.

Based on our analysis, we determined that there are more delays in the morning (6-8am) and evening (4-6pm) rush hours, with delays mostly caused by passenger and mechanical failures/issues. Mechanical issues seemed to have occurred more frequently on the weekdays compared to weekends. More delays occurred during the weekdays than the weekends and in the months of January and in the summer from May to August. Delays happened typically in stations at/near the terminals of the subway line (i.e., Kennedy, Kipling, and Finch).

We recommend that the TTC consider addressing passenger-related delays by developing procedures for rapid medical support to help distressed customers and enhanced surveillance to prevent hazardous conditions (e.g., customers on track-level). To address mechanical-related issues, we recommend the TTC institute regularly scheduled maintenance of subway trains to reduce the likelihood of these delays. We also suggest the TTC consider a public awareness campaign encouraging riders to: 1) stagger their commute time by leaving their house earlier or using the subway during off-peak hours to reduce the frequency of delays during typical rush hours (6-8am and 4-6pm); and 2) consider alternate routes (e.g., bus routes) with their subway commute to avoid delays at terminal stations.

The findings from this project can be further enhanced by including data from other sources, particularly the total TTC service capacity per day, public events information, and weather records. This would contextualize the frequency, severity, and potential reasons for the delays, and ultimately provide insight into strategies that will improve the subway experience for Torontonians.

1.0 Introduction

If you've ever had to commute to work either by driving or taking Toronto transit, you know that it can be quite a painful experience. In 2018, a study by U.K. based business solution company, Expert Market, found that Torontonians spend on average, 96 minutes commuting every day[1]. Of the 74 cities (with populations greater than 300,000) that the study ranked, Toronto had the "worst commute" of any major city in North America[1]. Toronto also had the second longest commute time, only behind Bogota at 97 minutes[1].

Part of the issue may be attributed to the local public transit service provide, the Toronto Transit Commission (TTC). Ask any commuter within Toronto and more likely than not, they will attest to the inefficiencies and delays that they have experienced while riding the "red rocket". It was reported by City News in 2019, TTC experiences more than 1 delay of 15 minutes[2] or greater per day and these delays were caused by numerous different events such as signal problems, rail issues, weather-related issues, passenger-related issues and the list goes on. The report also noted the longest delay recorded in 2019 was 455 minutes[2] (approximately 7.5 hours) between Wilson and Sheppard West stations. The TTC delays impact approximately a quarter of Toronto's workforce as it was reported by Statistics Canada that 24.7%[3] of workers' mode of transportation was public transit.

It is apparent that TTC has a lot of room to improve regarding the amount and duration of delays. The objective of this study is to investigate the TTC subway delay data published by the City of Toronto's Open Data Portal; and to provide insight into the reasons of delays and potential solutions to reduce the amount and/or duration of the delays.

Our group hypothesize that there will be a strong correlation or relationship with the amount and duration of delays with the following factors:

a. Month
b. Day
c. Hour
d. Station (location of delay)
e. Line

The study will aim to address the following set of questions:

1. How does the time of the day relate with the number and duration of delays?
2. Do certain types of delay occur more or less during specific hours of the day?
3. Are there relationship(s) between the day of the week and the number and duration of delays?
4. Are there relationship(s) between the day of the week and the type of delays?
5. Are there relationship(s) between month and number and duration of delays?
6. Are there relationship(s) between month and type of delays?
7. Which of the lines are causing more delays over the years?
8. Which stations are experiencing delays and what type?
9. Can the number and duration of delays be predicted?

Sources

1. Expert Market Study – The Best and Worst Cities for Commuting, 2018 - https://www.expertmarket.co.uk/vehicle-tracking/best-and-worst-cities-for-commuting
2. City News – 345 TTC Delays of 15 Mins, 2020 - https://toronto.citynews.ca/2020/01/03/ttc-delays-2019/
3. Statistics Canada – Commuting within Canada's Largest Cities, 2019 - https://www150.statcan.gc.ca/n1/en/pub/75-006-x/2019001/article/00008-eng.pdf?st=HQYNQ7uP

2.0 Data and Analysis Approach

2.1 Data source and file preparation

Data on the TTC subway delays was retrieved from the City of Toronto's Open Data Portal. We selected the data for the three most recently completed years from the time of writing for this report (2017-2019). Data from multiple years allow us to determine whether certain relationships are observed across multiple years. We also retrieved two reference documents from the Open Data Portal. The first was a file that described each variable in the dataset and the second was a description of each possible delay code. The data and reference documents were downloaded from: https://open.toronto.ca/dataset/ttc-subway-delay-data/.

The City of Toronto's Open Data Portal presents the subway delay data in Microsoft Excel (.xslx) format. Data for each month from 2017 to 2019 was first downloaded from the Open Data Portal, concatenated in Excel, and then converted into a Comma Separated Values (.csv) file ready for use in Jupyter Notebook.

The subway delay data is comprised of the following variables:

- Date: Date of the subway delay
- Time: Start time of the subway delay
- Day: Day of the week when the subway delay occurred (e.g., Monday)
- Station: TTC subway station(s) where the delay occurred
- Code: TTC delay code documenting the reason for the delay incident
- Min Delay: Duration of delay in subway service (minutes)
- Min Gap: Time length in between subway trains (minutes)
- Bound: Direction of subway train (e.g., north or south)
- Line: Subway line(s) where delay occurred
    - 4 lines: Yonge-University, Bloor-Danforth, Sheppard, and Scarborough Rail Train
- Vehicle: Subway train number that was directly affected by the delay

2.2 Data analysis software and libraries

All Python code was written in Jupyter Notebook (version 6.1.4). We used Pandas to view the data, impute missing values, and perform data analyses. Matplotlib and Seaborn were used to visualize the results.

2.3 Preparation of dataset

Creation of new variables

Many variables, such as Time and Code, were very detailed, however, this made it difficult to analyze with expansive lists of possible entries. Therefore, we created new variables by collapsing certain variables in the raw dataset and dropping the information that we hypothesized to have minimal influence on the subway delays. The five variables we created were: 1) Year; 2) Month; 3) Hour; 4) Delay Group; and 5) Delay Type.

Using the Date variable, we extracted the Year and Month data as two variables to include in the analyses. We did not create a variable for Day Number since we did not expect certain dates in a month (e.g., the 15th day of each month) would contribute to the number or duration of subway delays.

The Time variable was provided in a 24-hour clock format in hours and minutes. We extracted the Hour information as a variable for the time of day. We excluded the minute information since we did not expect certain minutes in an hour would contribute to the number or duration of subway delays. For Delay Group, we created bins based on the Min Delay information since categorical variables are better suited for grouping

analysis. For the Min Delay variable, 67% of entries had a value of '0'. We interpreted these entries as delays that were less than a minute. Given the large presence of '0' in the dataset, we created a '0' bin for all zero entries. We then created bins for each 5-minute interval up to an hour. For the remaining Min Delay data, we created a bin for delays between 1-2 hours and delays over 2 hours.

For Delay Type, we derived this variable using the Code information provided in the raw dataset. Using the reference material for the Code descriptions, we classified each code into one of 9 categories to explain the type of delay: 1) Mechanical; 2) Employee; 3) Signal; 4) Passenger; 5) Track; 6) Fire; 7) Weather; 8) Other; and 9) Unknown (i.e., codes that were not described in the reference document). Although there are 200 different codes to describe a delay incident, many codes could be classified into the same category. For example, the codes "MUI" and "MUIR" both refer to an injured or ill customer on a subway train, yet the former is for a customer who was transported whereas the latter is for a customer who refused treatment. In this case, we classified both codes as passenger-related delays. This approach enabled us to determine if there were specific type(s) of delays that were more frequent, as each delay code by itself may not appear very often.

Missing and inconsistent data

We imputed missing and inconsistent values for the Line variable. The Line variable contained 6 possible entries: 1) Yonge-University (YU); 2) Bloor-Danforth (BD); 3) Sheppard (SHP); 4) Scarborough Rail Train (SRT); 5) Multiple lines simultaneously (YU/BD/SHP); and 6) Other. We first handled the twenty entries that did not relate to any subway line. These were the names of bus or streetcar lines, such as "11 Bayview" and "510 Spadina", which we classified as "Other". Next, we replaced inconsistent entries (e.g., "B/D" or "BD Line") to match one of the six aforementioned possibilities. For the remaining 191 entries with missing Line information (0.3% of data entries), we classified them based on the mode in the dataset (Yonge-University line).

Dropped variables

Min Gap was removed from the analysis since the gap in between trains did not seem to hold distinct information about the delay. Furthermore, Min Gap showed a strong positive correlation with Min Delay ($r = 0.93$). Thus, the Min Gap information is already indirectly captured in the Min Delay variable and hence dropping the Min Gap variable would not limit our ability to analyze or draw conclusions from the data.

Bound was also removed from the analysis due to a large amount (22%) of missing values that could not be accurately imputed with the available information in the dataset. For example, if the delay occurred on the Yonge-University line, we could reduce the possibilities to either "Northbound" or "Southbound" since these are the only two directions applicable to trains travelling on the Yonge-University line. However, we could not determine whether the train was moving "Northbound" *versus* "Southbound".

Lastly, Vehicle was removed from the analysis since 26% of the data was given a value of '0' which we interpreted as missing values. The dataset does not contain sufficient information to reliably impute the vehicle information. Furthermore, the most common vehicle number involved in the delays only occurred in 0.3% of the total entries which suggests that delays may not be limited to specific subway trains.

2.4 Assumptions in dataset

The dataset only provides a record of the subway delays that have occurred. We do not have information on the overall TTC subway operations for each day, such as the total number of subway trips made per day, which would contextualize the severity of these delays. We are making the assumption that each day is equal in ridership and service levels. Yet in reality, many factors influence TTC operations. Furthermore, the residual

effects of a subway delay are not accounted for in the dataset since long delays would likely lead to fewer subway trips made for a given day, and hence, a lower number of subway delays that have a *chance* to occur.

2.5 Analyses and graphs

To answer the questions we defined for our project, we primarily determined the count and percentage of delays that matched our list of criteria with the "groupby" function in Pandas. We used a combination of data plots, such as bar charts, line graphs, and scatterplots to display the main findings. We also performed autocorrelations and a linear regression analysis to determine if the duration and frequency of TTC delays could be predicted from historical data.

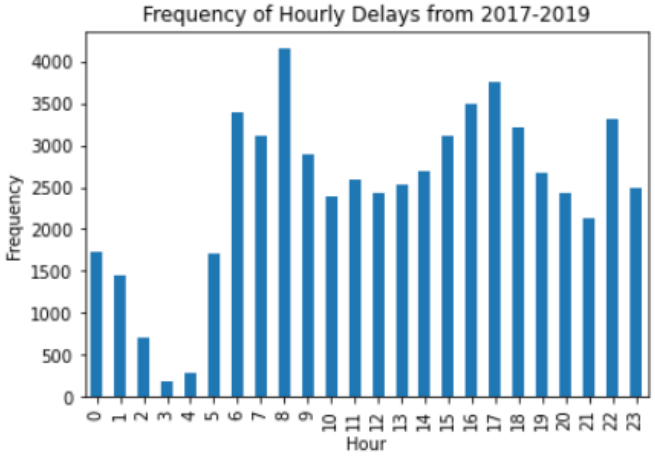
2.6 Limitations and challenges with the dataset

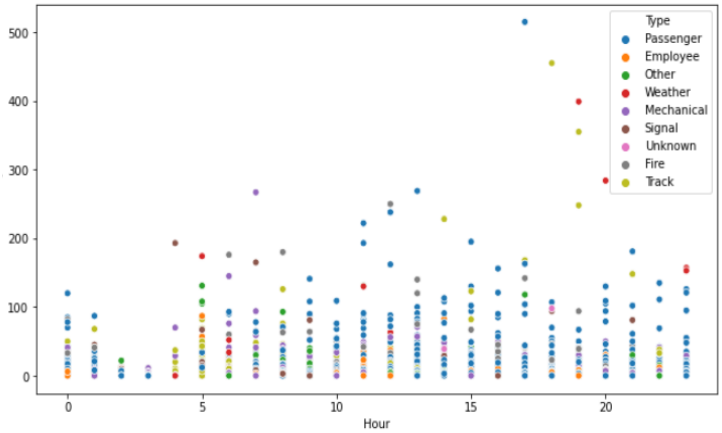
Given that the dataset does not include information on service levels, ridership numbers, or other contextual factors (e.g., public events or police investigations affecting subway service), we cannot draw conclusions on the impact that these delays have on "normal" subway service. Furthermore, most variables were categorical (both ordinal and nominal) rather than continuous. While categorical variables are useful for grouping analyses to determine the count and percentage of delays that meet specific criteria, this presented a challenge in performing more advanced analyses, such as regressions, that are more suited for continuous and ordinal variables. It did not seem appropriate to dummy code nominal variables, such as Line or Delay Type, since findings with dummy-coded nominal variables would be difficult to interpret.

2.7 Overall impressions

The quality of the data was satisfactory since most variables provided information with high granularity that in fact had to be collapsed for the purposes of this project. Given the size of the raw dataset (58,844 entries), there were relatively few missing values and it was limited to three variables (Bound, Line, and Vehicle). However, the dataset would benefit with supplemental information, including targeted service and ridership levels, since it would provide context on the severity and frequency of subway delays.

3.0 Analysis and Findings



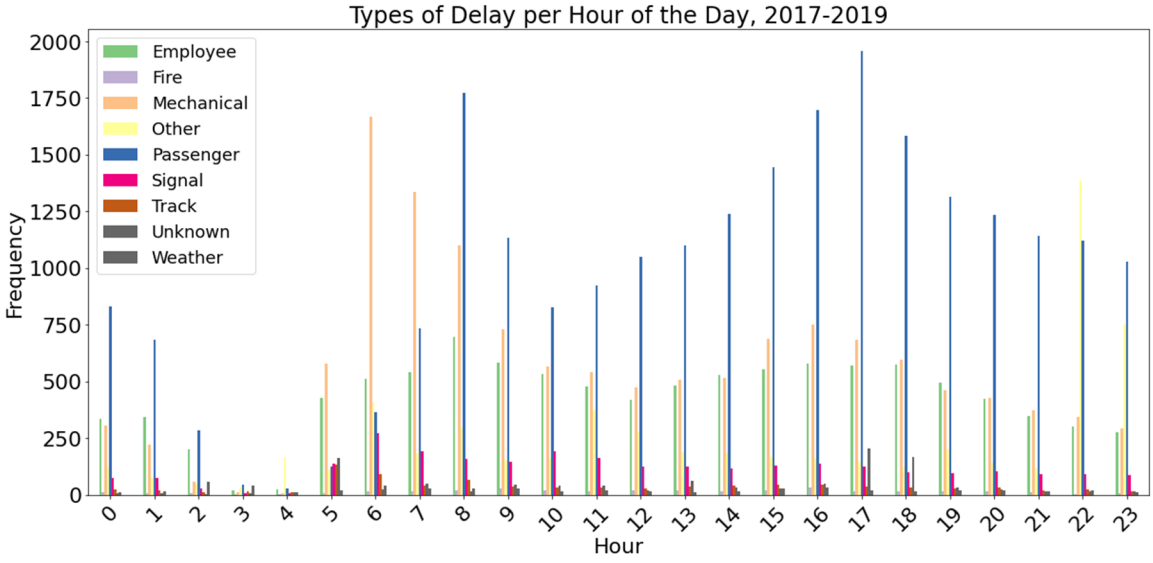


1. How does the time of the day impact the number and duration of delays?

Left plot: More delays occur in the morning (6-9am) and afternoon (3-6pm) rush hours. This can be explained by the fact that ridership is likely higher during these times as people are going to/from work. There is also a spike in delays at 10pm which might be related to system-wide reset/maintenance. In contrast, few delays occur between 3-4am likely since there is minimal subway service and riders.

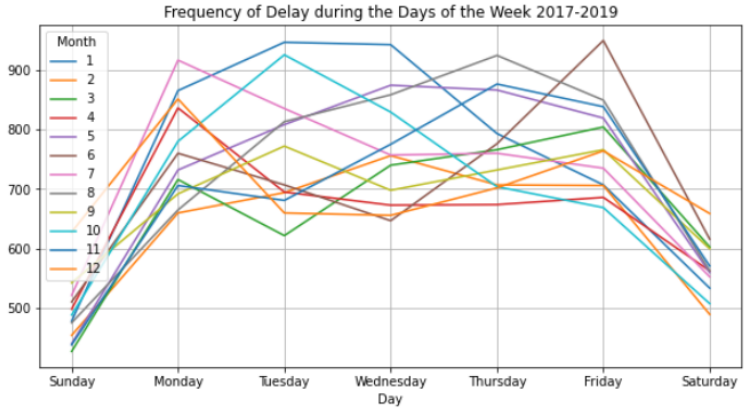
Right plot: With nearly 96% of delays lasting less than 10 minutes, this plot reinforces the fact that rush hour delays, while more frequent, are not necessarily longer than delay during non-peak hours. For delays (0.08%) lasting over 120 minutes, there is a cluster of passenger-related delays during the daytime (11am to 5pm) along with track and fire-related delays in the early evening (6-8pm).

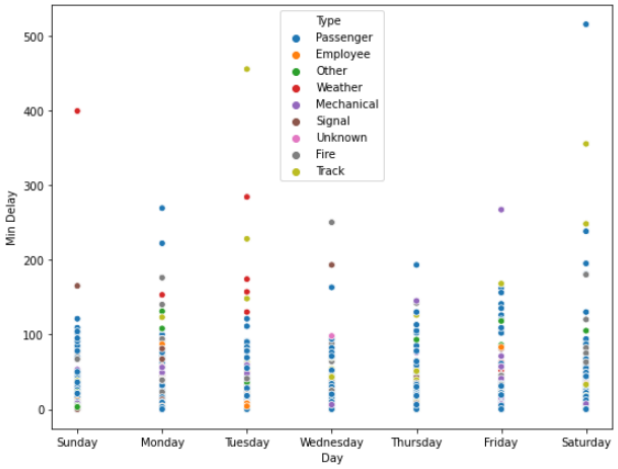
2. Do certain types of delay occur more or less during specific hours of the day?



A passenger-related delay is the most frequent delay type during the 24-hour period. This is most profound during the morning (8am) and afternoon (4-6pm) rush hours, likely due to peak ridership. Mechanical delays are more common in the morning (5-7am) when subways are starting service for the day.

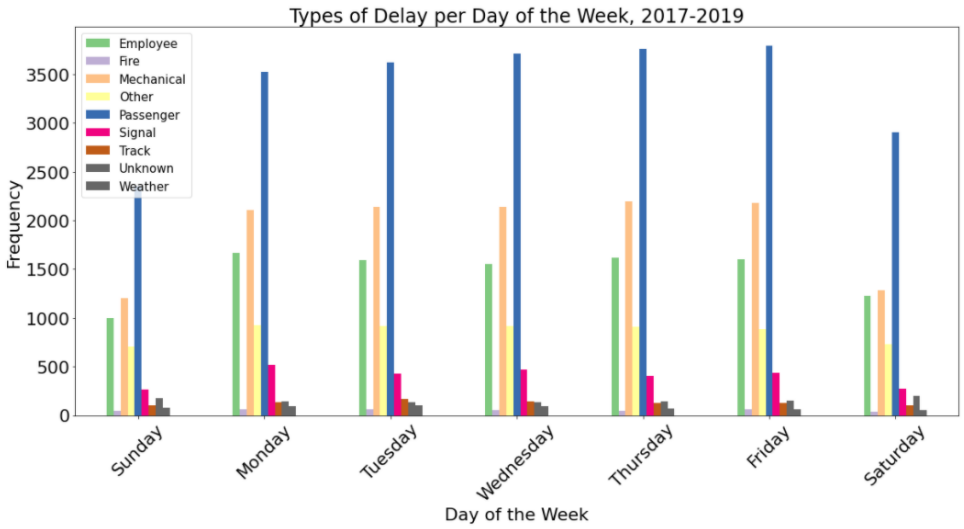
3. Are there relationship(s) between the day of the week and the number and duration of delays?





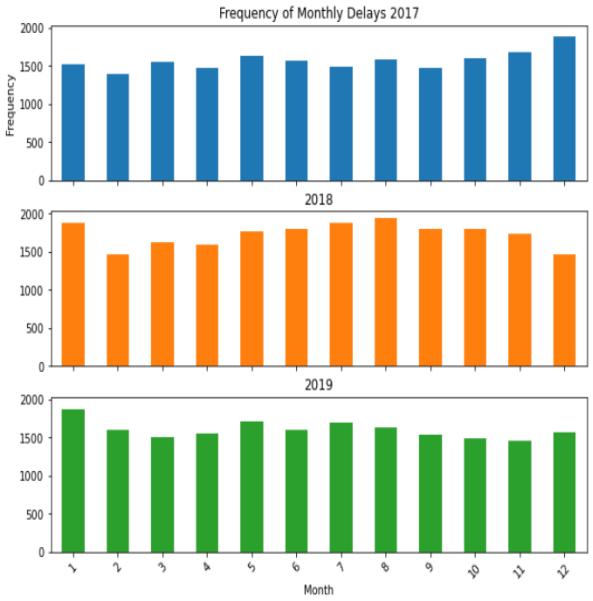
Left Plot: There are a greater number of delays during weekdays (Monday to Friday) relative to the weekends (Saturday to Sunday). This is likely attributed to the number of riders as the majority of the work force likely has their workweek fall within Monday to Friday.
Right Plot: Although the frequency of delays are higher on weekdays, the duration of delays do not appear to be any longer on weekdays or weekends. The duration of delays appears to be consistent throughout the week.

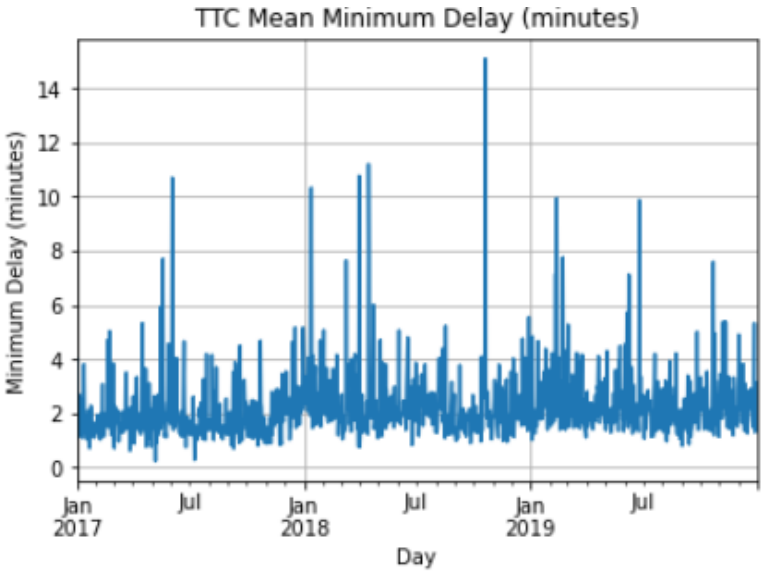
4. Are there relationship(s) between the day of the week and the type of delays?



Passenger-related delays are consistently higher than other types of delay throughout the week. Mechanical-related delays are the second most common type of delays followed by employee-related delays during the week. Mechanical-related delays appear to occur slightly more frequently on weekdays (11%-12%) than during the weekends (approx. 9%). This may be attributed to an increased number of trains running during Monday-Friday to accommodate the higher demand. Passenger-related delays appear to be the lowest on Sundays compared to the rest of the week.

5. Are there relationship(s) between month and number and duration of delays?

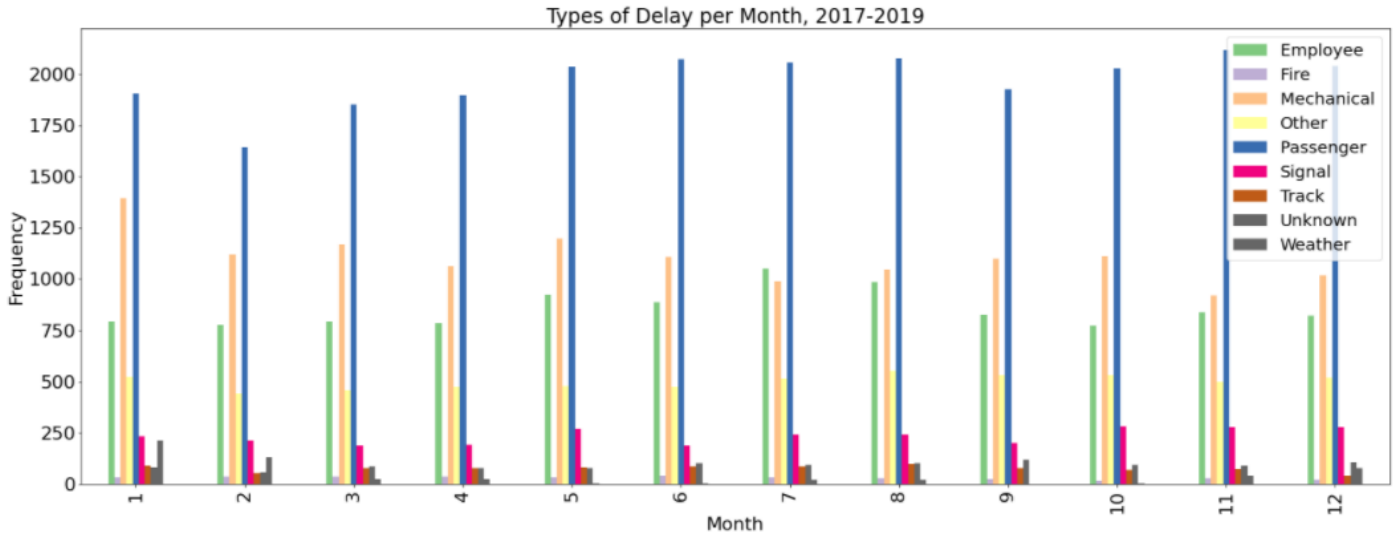




Left plot: There does not seem to be any significant relationship between the month and the number of delays, though we may cautiously infer that more delays tend to happen during the summer and winter months.

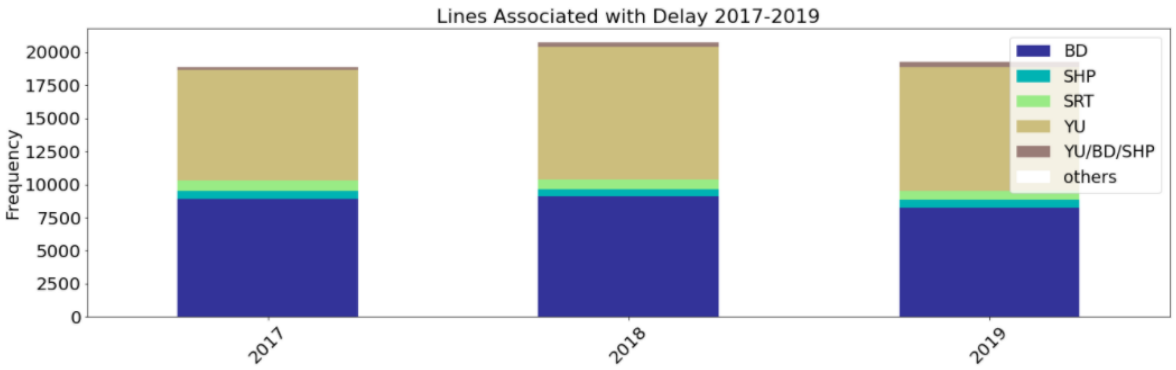
Right plot: We could see that the duration of delays seems to be more during the winter months of December and January. The month of May also seem to have higher duration of delays. It can also be seen that the months between July to October seem to have the lowest duration of delays.

6. Are there relationship(s) between month and type of delays?



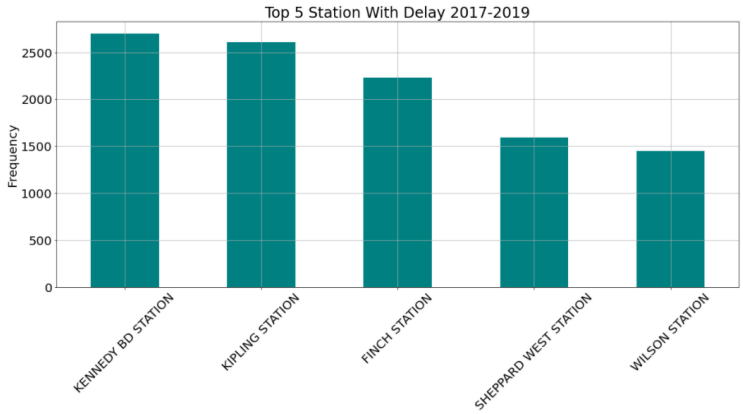
We can see that the passenger associated delays are consistently the major cause of delays throughout the year. The next major cause of delays throughout the year seems to be related to mechanical issues and are slightly significant in the winter months. Also, employee related delays were the third major cause of delays throughout the year.

7. Which of the lines are causing more delays over the years?



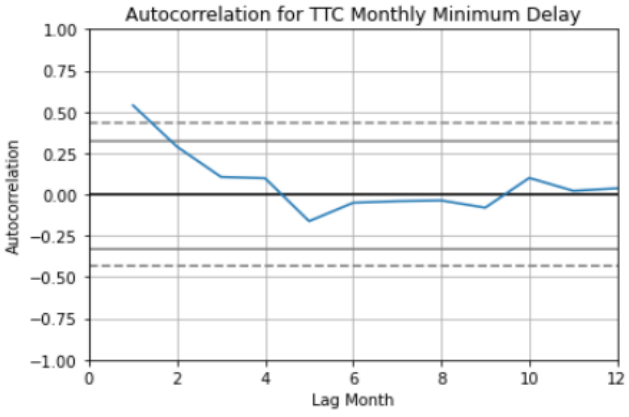
In terms of year over year study, the number of delays are relatively equal across the different subway lines. There appears to be an increasing trend in delays that affect multiple lines with each passing year. This could have been due to system-wide maintenance.

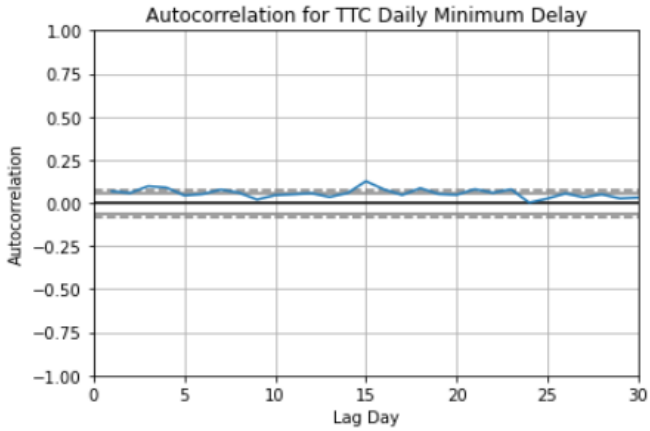
8. Which stations are experiencing delays and what type?

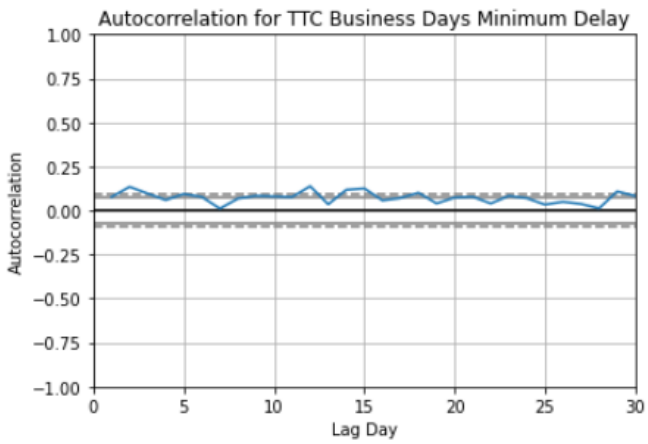


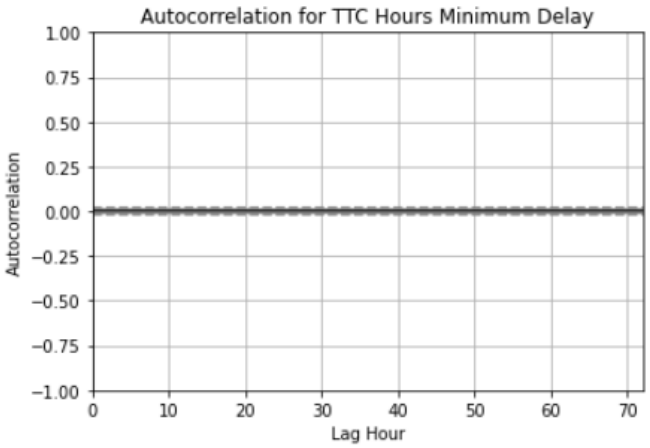
Top 3 stations with the most delays are Kennedy, Kipling, Finch. Typically, these stations are at/near the terminals of the subway line (i.e., Kennedy, Kipling, Finch). Most common type of delays are passenger, mechanical, and employee related.

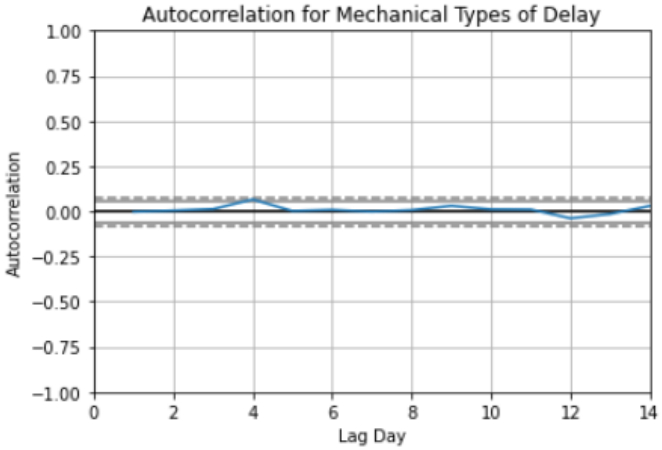
9. Can the number and duration of delays be predicted?



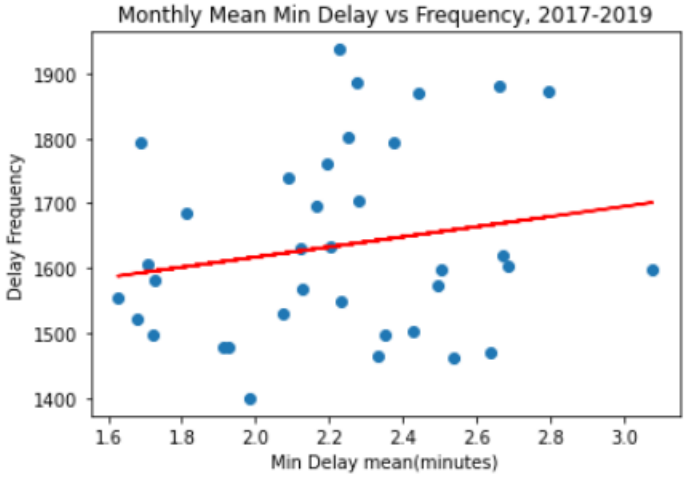








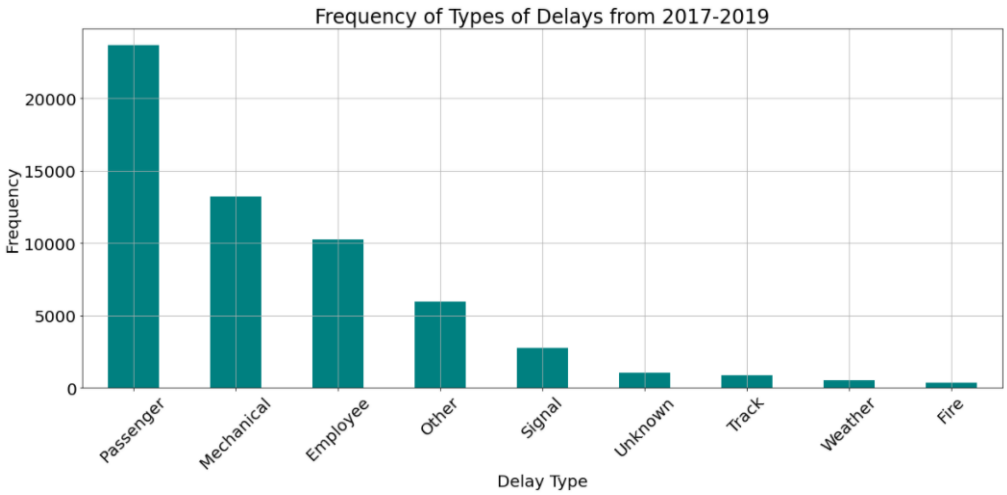
Autocorrelation was performed to determine if mean Min Delay can be forecasted on various time frames (monthly, daily, business daily and hourly). The data showed random behaviour, even with short lag periods. With the assumption that mechanical types of delays might be more reliably predicted because they are more affected by physical non-abstract variables, an autocorrelation determination was also done, but likewise did not show a predictable pattern.



A linear regression analysis was done to determine the presence of a relationship between the duration and frequency of delay. The model revealed a mild positive correlation with a coefficient of 78.

The models that the study has employed indicate that the duration and frequency of delays cannot be reliably predicted.

Other Findings:



Type/ Code

- the most common type of delay is passenger-associated, followed by mechanical type
- most passenger-related delays are caused by medical emergencies, false alarms and disorderly conduct, and occur during weekday rush hours
- the mechanical-related delays are associated with speed control and occur mostly during the early morning.
- the cause for prolonged delays are varied, including passenger, weather- and track-related causes

Min Delay

- majority of delays had a value of '0' which should be interpreted with caution
- almost 90% of the delays were of short duration, lasting for <= 5 minutes
- only a very small percentage (0.08%) lasted for >= 2 hours
- the leading causes of short duration delays are:
    - for passenger-related causes:
        - medical emergencies and disorderly conduct
    - for mechanical causes:
        - faulty doors
    - for employee causes:
        - operator not being in position or violating a signal
- the leading causes of delays >=2 hours are:
    - for passenger-related causes:
        - trains going in contact with people
- there is no significant autocorrelation of delays
- there is a weak correlation between frequency and duration of delay

4.0 Conclusion and Recommendations

4.1 Conclusion

Based on our analysis, we would like to make the following conclusion on each of the attributes:

Type/Code - The most common type of delay was passenger related followed by mechanical issues. We also noted that the most passenger related delays were caused by medical emergencies, false alarms and disorderly conduct by passenger which generally occurred during weekday rush hours. Regarding mechanical related delays, they were associated with speed control and occurred mostly during the early morning hours. The reasons which contributed to prolonged delays were passenger, weather and track-related issues.

Line - The line which was most associated with delays was the Yonge-University line which contributed to 47% of delays, followed closely by the Bloor-Danforth line at 44.5%. We could also see that the frequent prolonged duration of delays was associated with the Scarborough RT line.

Station - We could see that there was a trend of twice the number of delays on the terminal stations of Kennedy and Kipling (BD Line) and Finch (YU Line).

Year - We noted that the frequency of delays seemed to be stable through the 3 years we analyzed, however the duration of delays seemed to have an increasing trend.

Month - It may be cautiously inferred that more delays tended to happen during the summer and winter months, however the duration during winter was generally longer than in the summer months.

Day - There was an equivalent distribution of delays during the weekdays. There were less delays during the weekends however the duration of delays was as long, or even longer as compared to weekdays.

Hour - The delays seemed to be clustered around the rush hours of 6-9am (peaking at 8am) and 3-6pm (peaking at 5pm). There seems to be significantly decreased duration and number of delays during the dead hours of the night troughing at 3-4am. In general, there were less delays during the weekend hours except for Sunday which showed remarkable delays at 8-9am.

Min Delay - Many delays had a value of '0' which should be interpreted with caution as almost 90% of the delays were of short duration lasting for <= 5 minutes and only a very small percentage (0.08%) lasted for >= 2 hours. The leading reasons of short duration delays for passenger-related causes was disorderly conduct and medical emergencies, for mechanical issues was faulty doors and for employee causes was operator not being in position or violating a signal. The leading reason of delays for >= 2 hours were for passenger-related causes when train got in contact with person.

4.2 Recommendations

Based on our analysis, we would like to put forward below recommendations to reduce delays:

- ✓ Address passenger-related delays that can be prevented such as provision of quick medical support, instituting corrective measures for disorderly conduct or intensifying surveillance.

- ✓ Re-train personnel on speed control.

- ✓ Provide measures for early intervention on failing equipment such as train doors.

- ✓ Reanalyze delay data alongside TTC data on overall train operations to look for proportionality of delay, to better identify time/lines/stations that need to be addressed.

- ✓ Public awareness campaigns for riders to stagger their commute times to avoid rush hour delays and use alternate routes (e.g., bus, streetcar) with their subway commute to avoid the use of terminal stations.