TTC Bus Delay Data for 2022*

Eshta Bhardwaj

February 4 2023

This paper conducts an analysis and visualization of TTC bus delays in 2022. The focus of the analysis is on the impact of temporal data types and incident types on number and severity of delays to help understand causes of delay in order to mitigate them. The study found that key areas of improvement include mitigating risks that occur in peak commute hours, on Fridays, and for mechanical and operations delays.

Introduction

The Toronto Transit Commission (TTC) provides transportation services for commuters in the Toronto region. In 2021, the TTC service made a total of 197, 842, 000 trips which included 156 bus routes that serviced 111, 979, 186 passengers (Corporate Communications Department 2021).

Given the extensive bus service of the TTC along with its large ridership, it is important to reflect on how the service can be improved. One primary source of improvement is through the reduction of delays. In this paper, I will be processing, analyzing, and visualizing TTC bus delay data from 2022 to demonstrate the effects of temporal variables and incident types on the duration and frequency of delays. It was found that Fridays have the most frequent and longest delays while Sundays have the fewest and shorter delays regardless of the cause of the delay. The most common delay types were mechanical and operations delays which should be investigated for opportunities of improvement. Furthermore, most delays occurred during morning and evening commute hours.

The remainder of the paper is structured as follows: the next section discusses the exploratory review of the dataset to establish questions of interest. This section also contains a brief overview of the dataset. The subsequent section discusses the variables in the dataset further and presents visualizations based on the established scope of analysis.

*Code and data available at: https://github.com/eshtab/paper_1

Exploratory Review

Overview of Dataset

The TTC Bus Delay dataset (Toronto Transit Commission 2023) has 10 variables detailing various aspects of information regarding each delay in 2022. However, an exploration of the dataset was required to understand each variable type (eg: continuous, categorical, etc.) and assess its data quality. This, in turn, allowed a scope to be established for the analysis and summarization of the delay data. For the exploration, preprocessing, and visualization of the data, the R statistical programming language was used (R Core Team 2020).

Scope of Analysis

In order to establish scope, I first read the data from the Open Data Toronto portal using the package opendatatoronto (Gelfand 2022) which imported data from the Open Data Portal directly. I then saved the dataset using the write_csv function which is part of the tidyverse package (Wickham et al. 2019). The tidyverse package contains various data related functions. I then used the function clean_names to change the names of the loaded dataset easier to type. This function is part of the janitor package which contains functions for cleaning data.

The last and key step of the exploratory review was using the unique function to check the quality and nature (i.e. type) of each variable in the dataset. This allowed some key insights to come forward. For example, the variables route, location, direction, and vehicle contained fairly unique values denoting specifics about the bus and or location of the accident. Summary statistics would not be appropriate for these variables because of their varied nature which would not lend itself to aggregation. Some of these variables also contained blank or null values that could not be filled in or be interpolated. On the other hand, variables like date of delay, day of week of delay, time of delay, length of delay in minutes, and reason for delay (i.e. incident) had perfect completeness and were consistently entered in the correct format. The exploration of the variables also revealed that the time of delay and length of delay variables were continuous and not categorical such as 0-5 minute delay, 6-10 minute delay, etc. This provided insights into the questions that could be explored and the types of graphs that would appropriately visualize the results.

The following questions were selected for analysis:

1. How many delays were there in total?
2. What was the cumulative delay time in minutes?
3. What are the impacts of the delay reason (i.e. incident) on the frequency and severity of delays?
4. What are the impacts of temporal variables on the frequency of delays?
5. Which month of the year had the most severe delays (where severity is indicated by delay length)?

Additionally, any question at the intersection of temporal variables, severity of delay, and incident type would yield interesting results. For this reason, the following question was selected for analysis (as a sample among other feasible questions):

6. How do the most common incident types differ in frequency and severity of delays when analyzed by day of week?

Data

Overview of Variables

The 2022 TTC Bus Delay dataset contained 10 variables including:

- Date: the date (yyyy-mm-dd) the delay occurred
- Route: the bus route where the delay occurred
- Time: the time (hh:mm) the delay occurred
- Day: the day of week the delay occurred
- Location: the intersection where the delay occurred
- Incident: the reason for the delay
- Min Delay: the delay time, in minutes, to the schedule for the following bus
- Min Gap: the time, in minutes, from the bus ahead of the following bus
- Direction: the direction of the bus route
- Vehicle: the unique identifier of the bus causing the delay

Processing, Analysis, and Visualization

Prior to starting the visualization of the questions proposed, three processing steps were conducted. The first step was to remove variables that would not be used for analysis, therefore only the date, time, day, incident reason, and delay time variables were kept. The next step was to change the day variable into a factor variable so that a non-numeric ordering could be applied.

The third step involved looking at the data quality of the delay time variable. Based on the exploratory review, the delay time variable contained various delay occurrences of over 120 minutes. A decision was taken to remove any delay record with delay time over 120 minutes as this indicated an error.

How many delays were there in total? What was the cumulative delay?

The total delays were calculated by summing the number of rows in the data using the nrow function, while the total delay time required use of the sum function on the delay time variable. The total delay time of 851, 999 minutes over 57, 625 delay incidents indicates that on average each delay was approximately 15 minutes long.

```
The total delay time in 2022 was:  851999  minutes.
```

```
The total number of delays in 2022 were:  57625 .
```

What are the impacts of the delay reason on the frequency and severity of delays?

To visualize the most common delay reason based on the frequency of delays, I used the ggplot function to create a bar chart with incidents as the x-axis and number of delays as the y-axis. The results in Figure 1 show that the 3 most common incidents (in order) are operations delays, mechanical delays, and collisions.

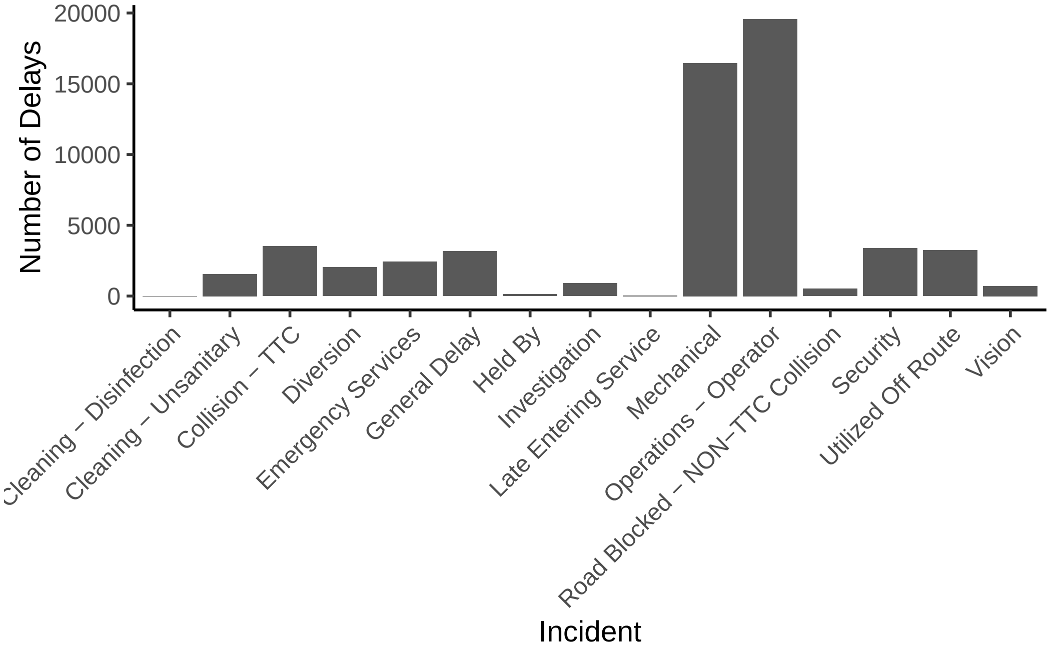

Figure 1: Most Common TTC Delay Reason.

To visualize the most common delay reason based on the frequency of delays, I used the ggplot function to create a bar chart with incidents as the x-axis and length of delay time in minutes as the y-axis. Figure 2 shows that based on delay time, the 3 types of incidents that cause the longest delays are diversions, operations delays, and mechanical delays.

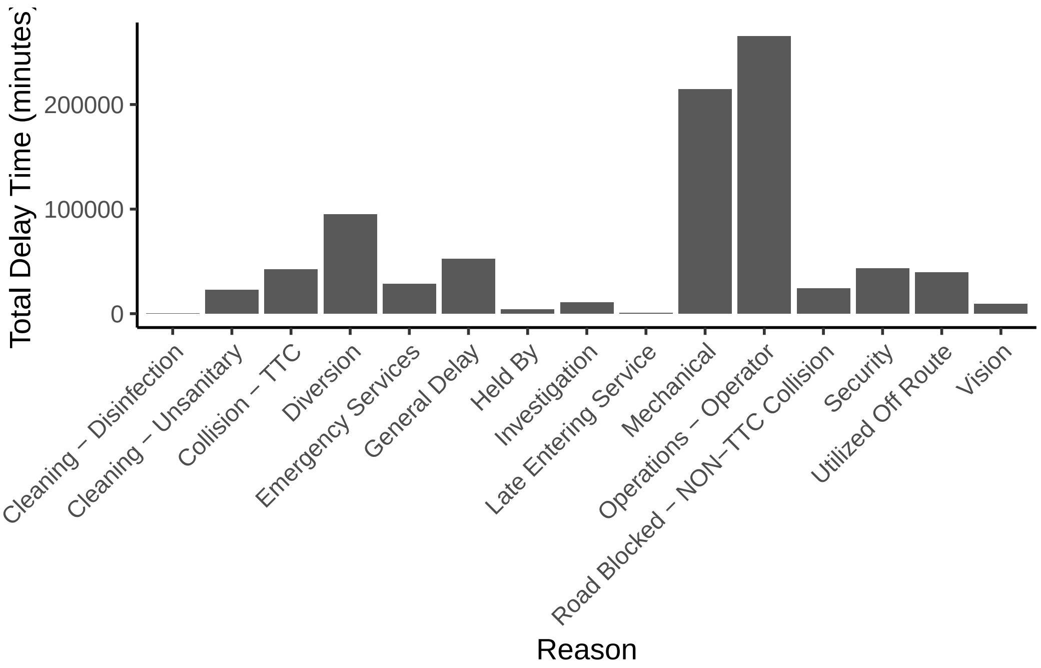

Figure 2: Incident Type with Total Delay Caused.

What are the impacts of temporal variables on the frequency of delays?

To be able to view the impacts of temporal variables like month, I used one of the dplyr functions (separate) to break the date column into year, month, and date. Based on this I created a bar chart using ggplot with month as the x axis and the number of delays as the y axis. Based on Figure 3, it can be seen that the months of July, August, and January had the most delays. However all months had approximately 4000 delay incidents and the highest number of delays in July and August was approximately 6000. This may indicate that the month of the year does not have a large impact on the number of delays.

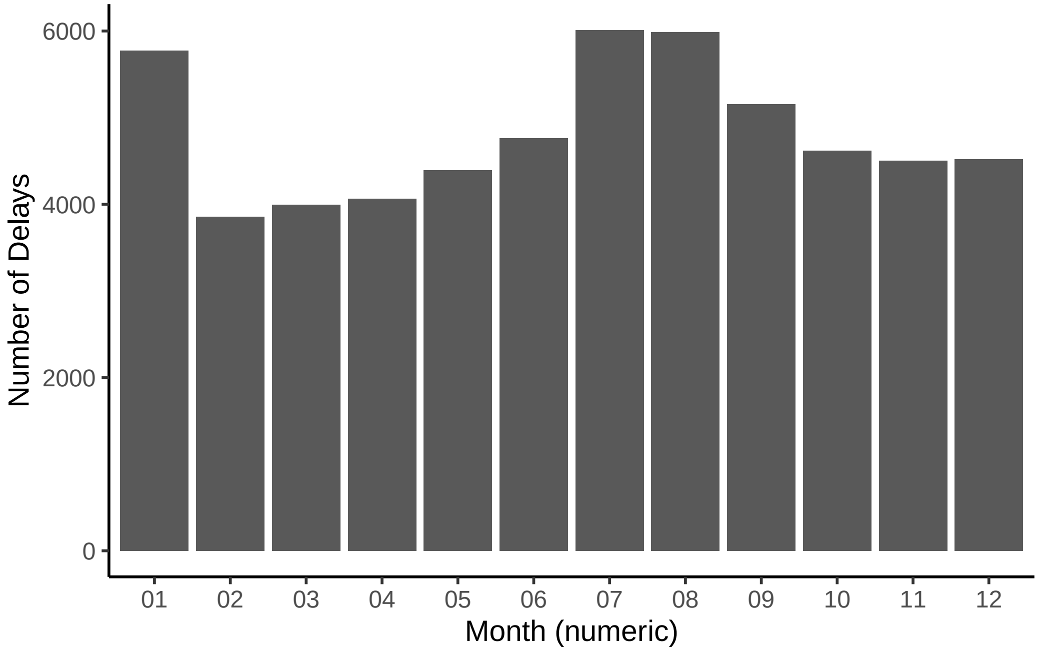

Figure 3: Number of Delays by Month.

A bar chart looking at the frequency of delay incidents by day of week was also created using a similar method as previous figures. In Figure 4, we can see that the most number of delays happened on Friday closely followed by the remaining days of the week with the least delays on Sunday. This may be attributed to the decreased TTC service on Sundays.

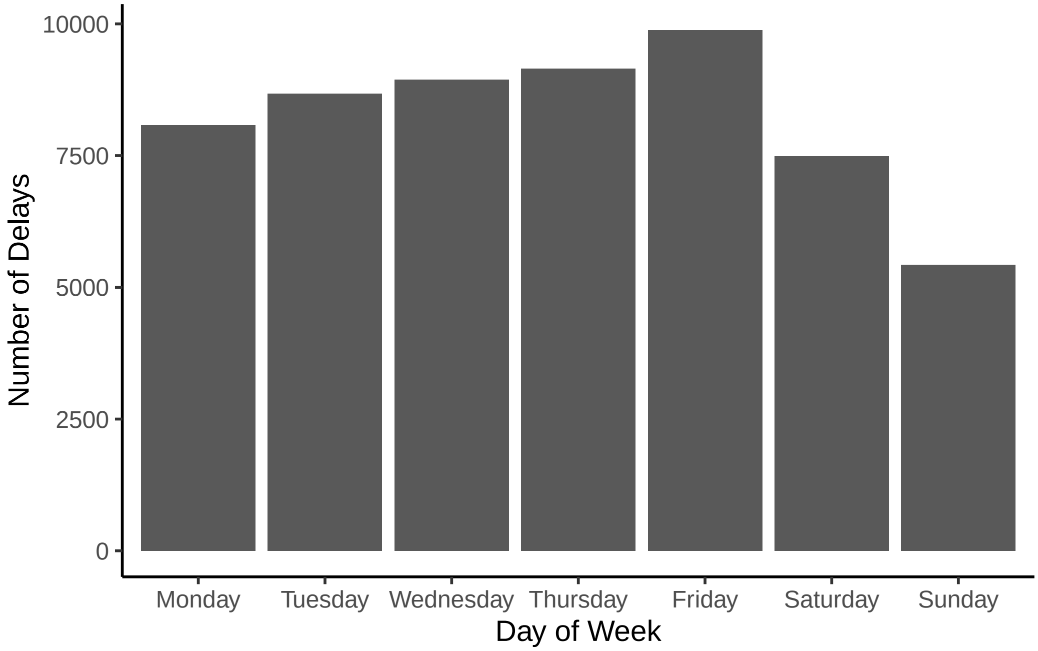

Figure 4: Number of Delays by Day.

In Figure 5, we can notice a few peaks. Firstly there are a few peaks starting at the 7am. These peaks persist consistently until approximately 3pm where higher peaks start to form. We then these higher peaks into the evening until approximately 6pm. This aligns with morning and evening commute hours which indicates that a greater number of vehicles on the road may cause more transit delays.

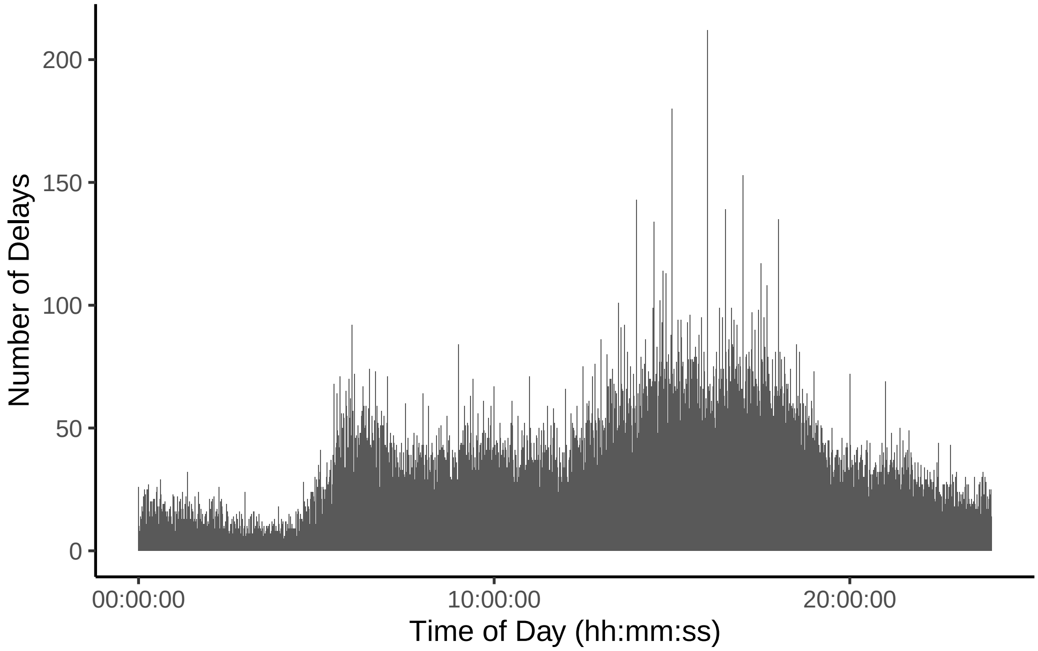

Figure 5: Number of Delays by Hour.

Which month of the year had the most severe delays?

A bar chart was also created to visualize the month of the year with the longest delays using ggplot. In Figure 6, the longest delay times are attributed to January, with other months following closely behind. The shortest delay time is experienced in March.

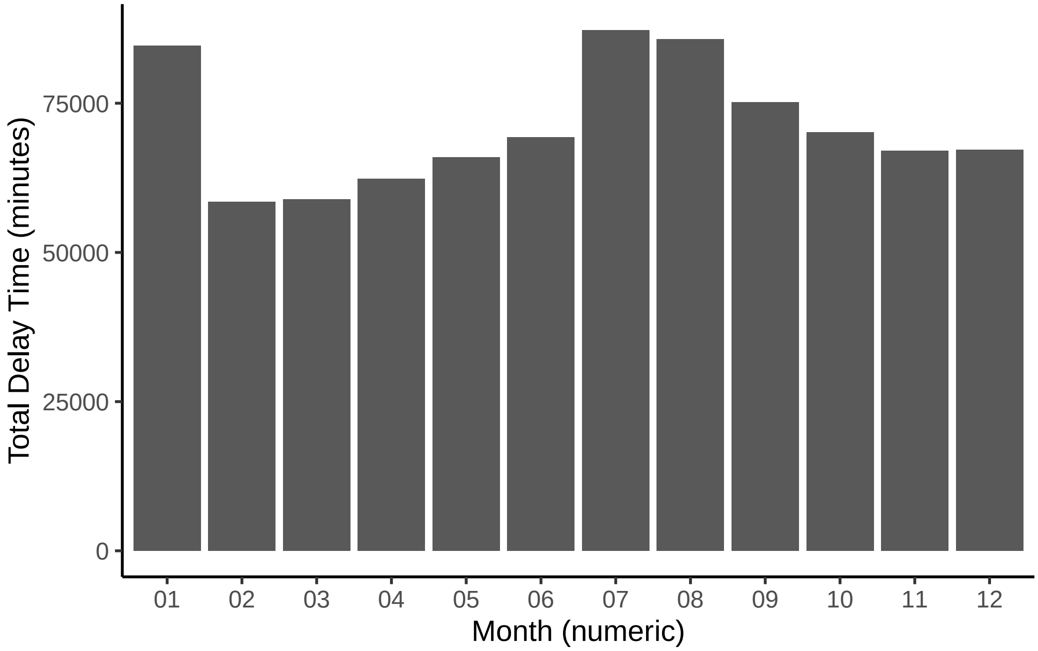

Figure 6: Month with Longest Delays.

How do the most common delay types differ in frequency and severity of delays when analyzed by day of week?

To visualize the number of delays for the three most frequent delay reasons according to the day of week, I used the function facet_wrap to make categorized subplots. This helps in recognizing any patterns across subsets of the data. In Figure 7, it can be seen that although Fridays have the most frequent delays while Sunday has the least frequent, the gap between the two is only pronounced for operations delays.

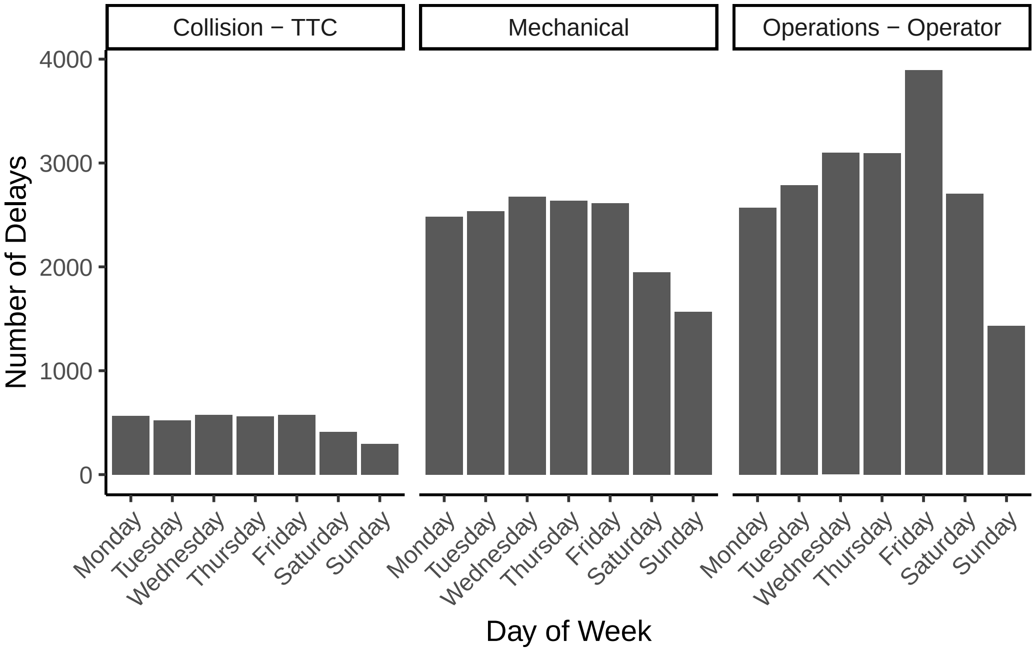

Figure 7: Number of Delays by Delay Reason and Day of Week.

Similarly for Figure 8, facet_wrap was used to create subplots but instead to visualize the severity (or length) of delays. The three most common incident types are different than in Figure 7 because they are chosen according to the delay reasons that caused the longest delays. As with Figure 7, Fridays also have the longest delay times and Sundays have the least delay times with this being most pronounced for operations delays.

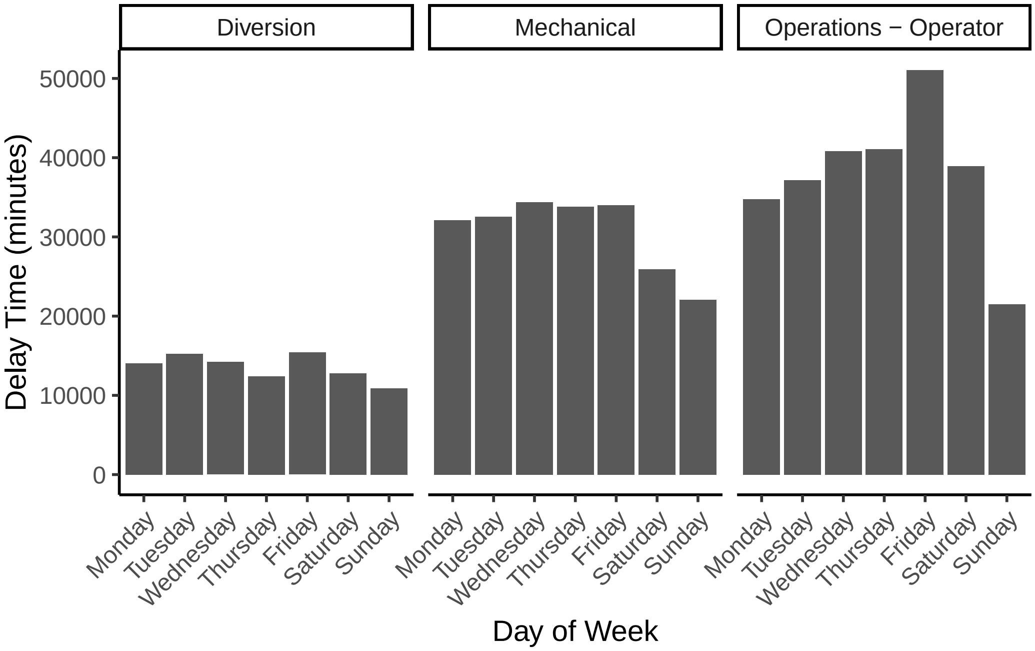

Figure 8: Cumulative Delay Time by Delay Reason and Day of Week.

References

Corporate Communications Department. 2021. *2021 Operating Statistics*. Toronto Transit Commission. https://www.ttc.ca/About_the_TTC/Operating_Statistics/2018/index.jsp.

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

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

Toronto Transit Commission. 2023. *TTC Bus Delay Data*. Open Data Toronto. https://open.toronto.ca/dataset/ttc-bus-delay-data/.

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.