Toronto Transit Commission (TTC) Delay Analysis – Summary

Introduction

Delayed public transportation affect all commuters all over the world and affect most daily aspects of living. These delays could be caused by numerous interrelated factors. Evolving techniques in data science that incorporate statistical knowledge, database concepts and new computing power could help understand the causes of the delays and propose a streamlined approach to improved systems.

Purpose

The purpose of the project is to come up with a delay prediction model that uses data science techniques, using data collected by the Toronto Transit Commission (TTC).

Approach

We utilized the City of Toronto's open data portal (https://open.toronto.ca/dataset/ttc-subway-delay-data/) 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 in relation to time (year, month, day of the week, hour),
2) the specific lines and stations that experience the most delays,
3) the common causes of delays.

To accurately address these questions, we created new variables, imputed missing values and/or inconsistent entries, and dropped irrelevant information from the dataset.

Data Quality

The quality of the database was sufficient, with a good number of data and variables, with few missing values. Supplemental information such as total service provided and ridership levels would improve the analysis by providing context on the severity of the delays when viewed against the totality of the system.

Limitations

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 were unable to draw conclusions on the impact that these delays have on "normal" subway service.
Furthermore, most variables were categorical which presented a challenge in performing more advanced analysis, such as regressions, that are more suited for continuous variables.

Analysis and Observations

Our analysis showed more frequent delays in the morning and evening rush hours on the weekdays, which were predominantly caused by passenger- and mechanical-related delays *(Figure 1)*. Delays on the weekends were less frequent compared to weekdays; however, the duration was longer. Delays tended to be more frequent during the summer and winter months. The overall duration of delay was in general short-lasting (<10 minutes) and were shorter during the summer and longer during the winter. The delays have a slight worsening trend through the 3-year observation period. The delays were more frequent on the Yonge-University and Bloor-Danforth Lines at the terminal stations of Kennedy, Kipling and Finch.

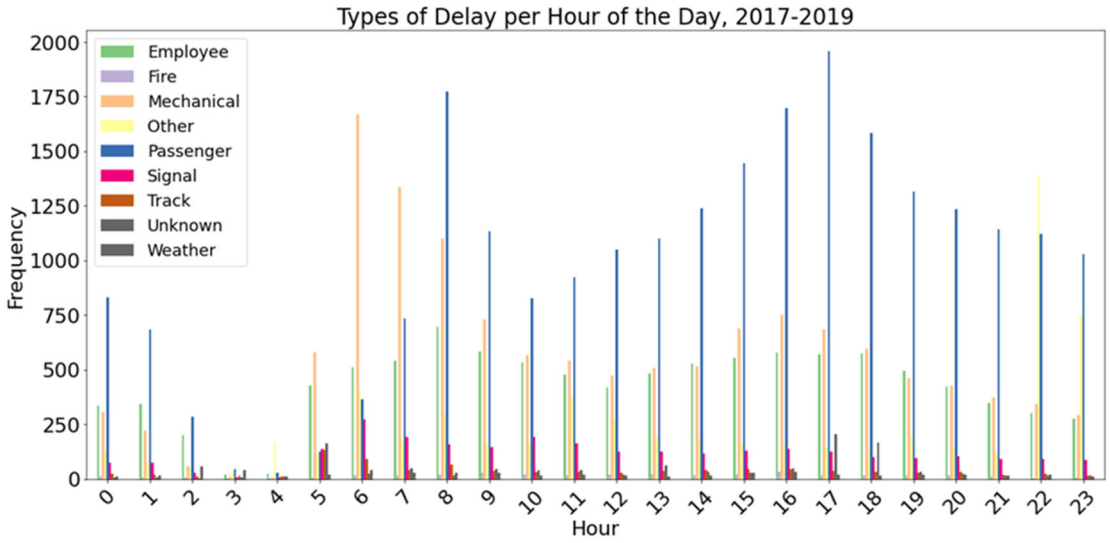

Figure 1. Types of Delay by Hour

An autocorrelation using different time frames and focusing on mechanical-types of delay (presumed to have a higher probability of predictability), showed a random pattern (*Figure 2*). A linear regression analysis was done to determine the presence of a relationship between the duration and frequency of delays (*Figure 3*). The model revealed a mild positive correlation with a coefficient of 78. Taken together, the models employed in this study indicate that the duration and frequency of delays cannot be reliably predicted, and other models should be sought.

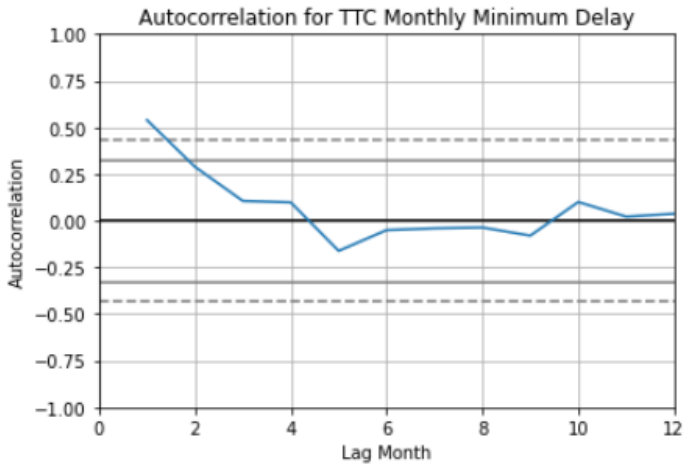

Figure 2. Autocorrelation - Monthly Delays

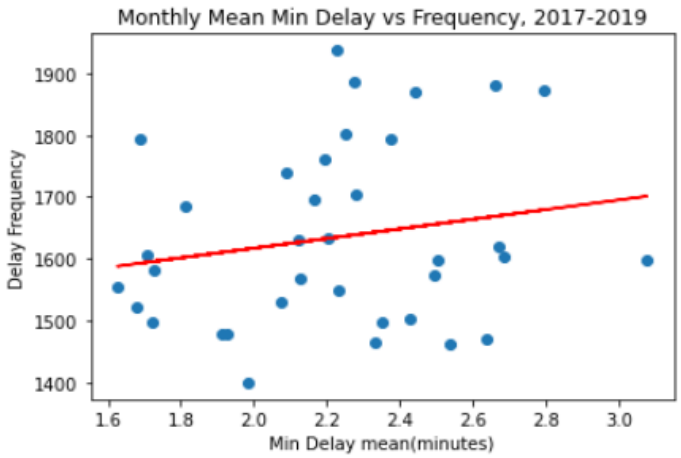

Figure 3. Linear Regression - Frequency and Duration of Delays

Recommendations

While our results indicate that delays cannot be reliably predicted, we recommend the TTC take a proactive approach. This includes 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 that the TTC intensify regularly-scheduled maintenance of subway trains. We also suggest that the TTC consider a public awareness campaign encouraging riders to stagger their commute time by leaving their homes earlier or using the subway during off-peak hours to reduce the frequency of delays during typical rush hours (6-8am and 4-6pm).

Continued enhancement of this analysis by including data from other sources, particularly the total TTC service capacity per day, public events information, and weather records 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.