SDSS Datathon
Team Members: Yulia Miyajima, Lucie Yang, Jing Yu, Maggie Zhang
March 2, 2025

Modelling TTC Delays in Buses, Streetcars and the Subway

Introduction

In this project, we aim to investigate past Toronto Transit Commission (TTC) buses, streetcars, and subway lines to predict future occurrences of delays. Customers are often dissatisfied with the high amount of TTC delays, making them less likely to use the TTC again if they could drive, bike, walk, or Uber instead. The TTC operates at a loss and could significantly benefit from higher rates of ridership by improving their delay response strategy. The majority of bus delays in 2024 were attributed to mechanical issues. To combat this, the TTC could use our model to predict where bus delays are most likely to occur to have repairmen on standby before rush hour around 3 p.m. For streetcars, operational issues led to the most delays but are unspecified in the dataset and unavailable through internet search, so further inquiry with the TTC would be required to find solutions. Lastly, unruly patrons were the largest cause of subway delays in 2024, so greater security efforts would decrease response time and help riders feel safer.

Using our model, the TTC can predict how the seasons, weather, or time of day will impact delays and preemptively adjust scheduling of services or make preparations for the most common causes of delays. For instance, the TTC could establish a business relationship with local Toronto weatherproofing businesses to ensure that vehicles are prepared for snow in the winter and other inclement weather. In addition, the TTC could perform further analysis on the causes of the high number of operational delays in preparation for peak delays during rush hours.

Data Sources

We used external data from Esri Canada Education to find TTC station coordinates to create a map of the stations in R. To expand the original dataset on TTC delay times, we used historical data from Environment and Climate Change Canada to include the daily measurements of minimum and maximum temperature, precipitation, and amount of snow throughout 2024. In addition, we filtered the delays by rush hour times (6 a.m. to 9 a.m. and 3 p.m. to 7 p.m.) defined by the TTC website.

The data from Environment and Climate Change Canada may be inaccurate as stated in their data quality disclosure due to difficulty performing quality control on measurements

collected by automated weather machines. In addition, the weather data is from one station in Toronto City, which may not track weather for all locations serviced by the TTC.

Methodologies

To clean data, key categorizations were added to support analysis. Delays were labeled as occurring on weekdays or weekends, as well as during peak hours (morning and evening rush). Seasonal tags were assigned based on the month, and a list of 2024 Toronto holidays was used to flag delays on those dates. Missing values were handled by removing records with critical gaps, such as missing dates or locations. The dataset was further enriched by merging it with daily weather data, allowing for analysis of how temperature, precipitation, and snow affected transit performance. For subway delays, only those occurring at a single station were kept to ensure clarity in station-level analysis. These preprocessing steps made the data cleaner and more structured for business insights.

Then, we conducted exploratory data analysis to find patterns and identify key variables for our model. Based on those findings, we then began experimenting with machine learning models. The TTC bus delay dataset was prepared for machine learning analysis by combining the date and time columns into a single datetime column and extracting key time-based features such as the hour of the day, the day of the week, and the month. Categorical variables like location, incident, route, direction, day_type, rush_type, and season were converted into numerical values. A set of features including time-based attributes, incident-related details, and external factors such as holidays and weather conditions like temperature, precipitation, and snow accumulation were defined to be used for modeling.

Three models for predicting bus, streetcar, and subway delays respectively were made. Using a Random Forest Classifier, a model for delays of each transportation type was made that creates a binary target variable, delay_occurrence, which is set to 1 if the delay is greater than zero and 0 otherwise. The dataset is then split into training and testing sets using an 80-20 split, with X representing the selected features and y representing the target variable. Next, a Random Forest Classifier is initialized with a fixed random seed for reproducibility and trained on the training data. After training, the model makes predictions on the test set, which are then evaluated using an accuracy score and a classification report.

Then, we conducted a deeper investigation on the incidence to make actionable recommendations for the future.

Results

Many of the visualizations are dynamic or interactive. Please go to https://lucieyang1.github.io/Datathon-TTC/ for the full experience!

With the goal of analyzing contributing factors to delays on TTC buses, streetcars and the subway, we decided to conduct exploratory data analysis on the geographic hotspots of subway delays, the top incidence types, the peak times for delays, and the distribution of delay durations for each transport type.

First, we plot the number of delays reported at each subway station in Toronto in 2024, in **Figure 1.** As seen below, it seems like more delays occur at the endpoints of each line, namely at the Kennedy, Kipling, and Finch stations. Moreover, the figure suggests that the highest number of delays occurs at Bloor station, with 1008, although the average delay is only 1.97 minutes. Interacting with the full map, it suggests that disorderly patrons make up a substantial amount of the reasons for delays.

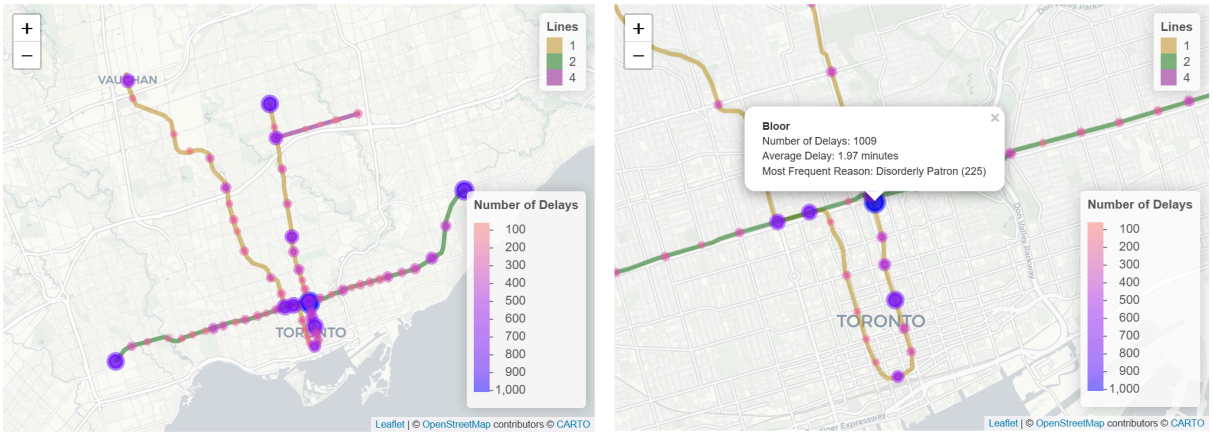

Figure 1. Geographic hotspots of subway delays in Toronto, 2024

Hence, we decide to investigate the most frequent reasons for delays for each transport type, as in **Figure 2**. This is vital to bring insight into how future delays might be mitigated and specific solutions to target. This plot suggests that mechanical reasons are the top reason for delays in buses, whereas operations and security are the top reasons for streetcars. Interestingly, the top reasons for delays in the subway are passenger-related, including disorderly patrons, illness and injury, matching what was previously suggested.

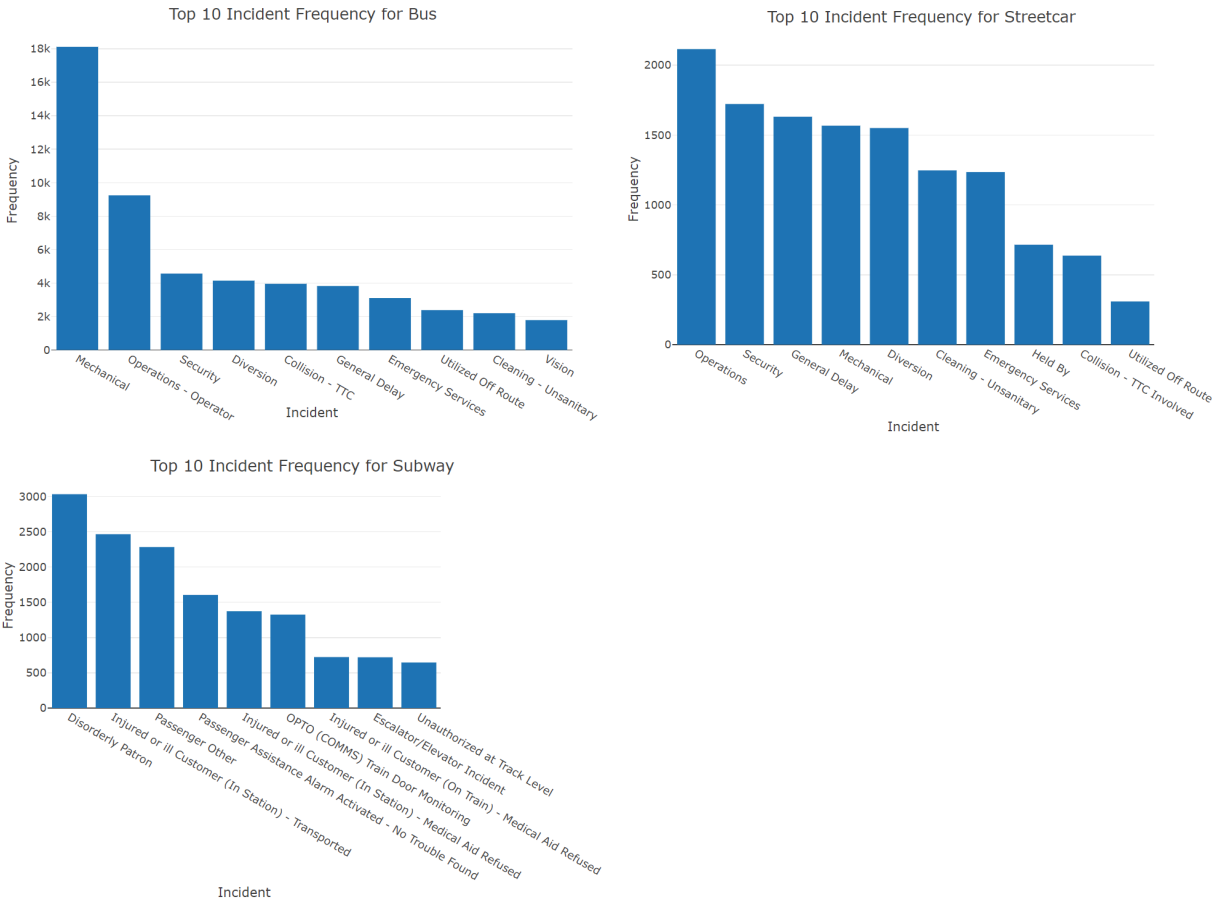

Figure 2. Top 10 most frequent incident types for each transport type in the TTC, 2024

Then, we investigate the peak times for delays in each of these transport types. As shown in **Figure 3,** the frequency of delays for the subway and streetcars seems relatively consistent over time, except from around 2am to 5am, when it is usually closed and with slight peaks at 8am and 4pm, at peak hour. This led us to include a variable to indicate rush hour for our model. Moreover, the bus has a significant peak at 5pm, supporting this decision.

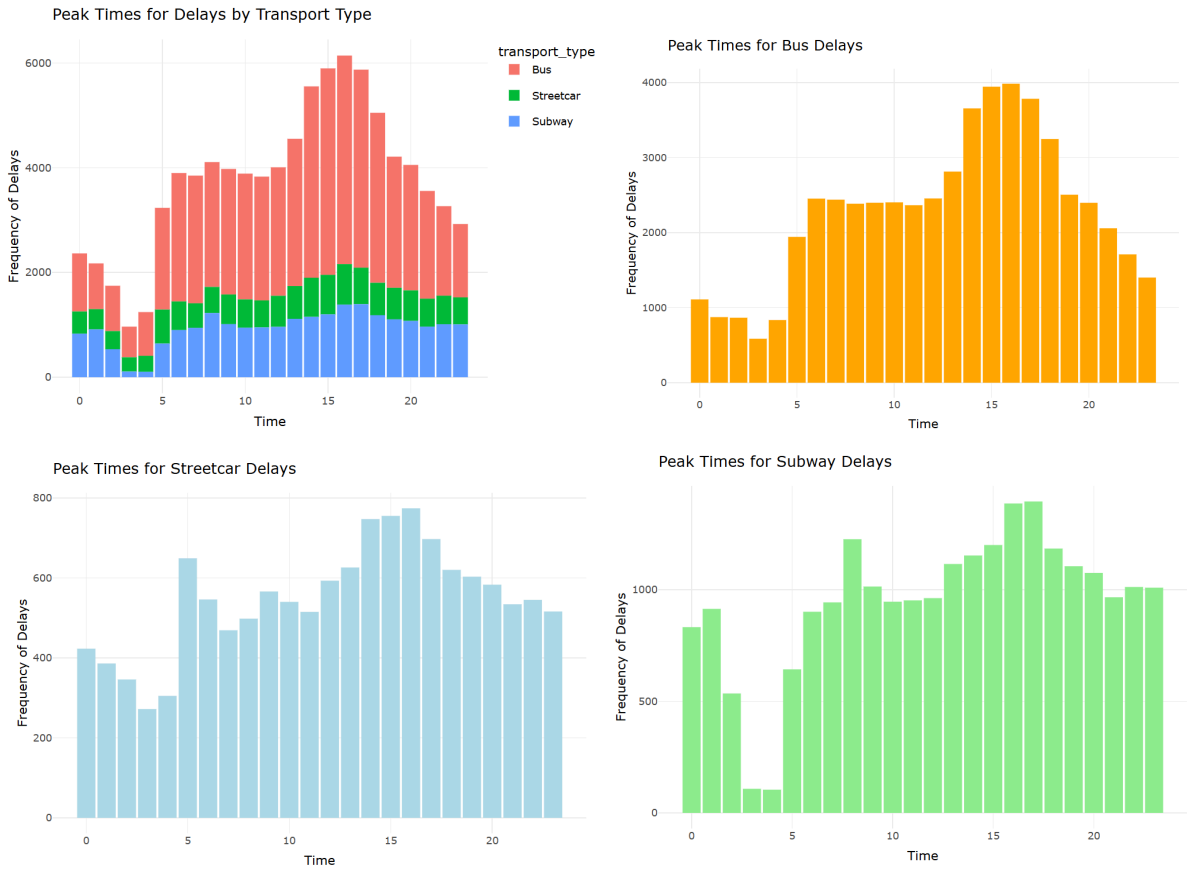

Figure 3. Peak times for delays, by transport type

After removing outliers from the data, we investigate how the distribution of the delay durations vary between the different transport types in **Figures 4** and **5**. The violin plots in **Figure 4** suggest that the delay durations for buses has a large spread, as compared to streetcars and the subway. The subway seems to have a significant number of values near 0, while the streetcar has peaks at around 0 and 10 minutes. **Figure 5** shows how this varies over months: it seems that the number of delays increases in the summer months, although the general pattern remains consistent over the months. This motivated us to include season as a variable of interest in our models.

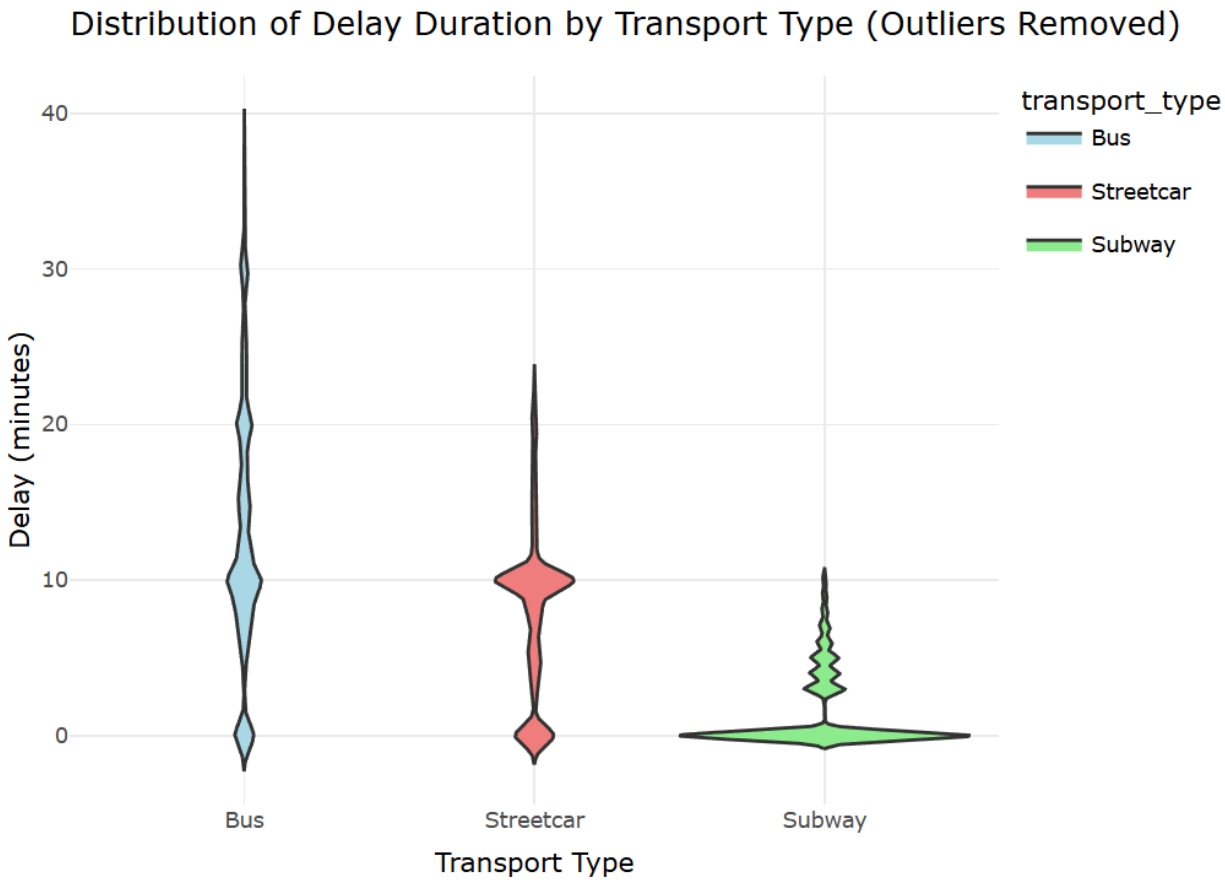

Figure 4. Distribution of delay durations by transport type, after removing outliers

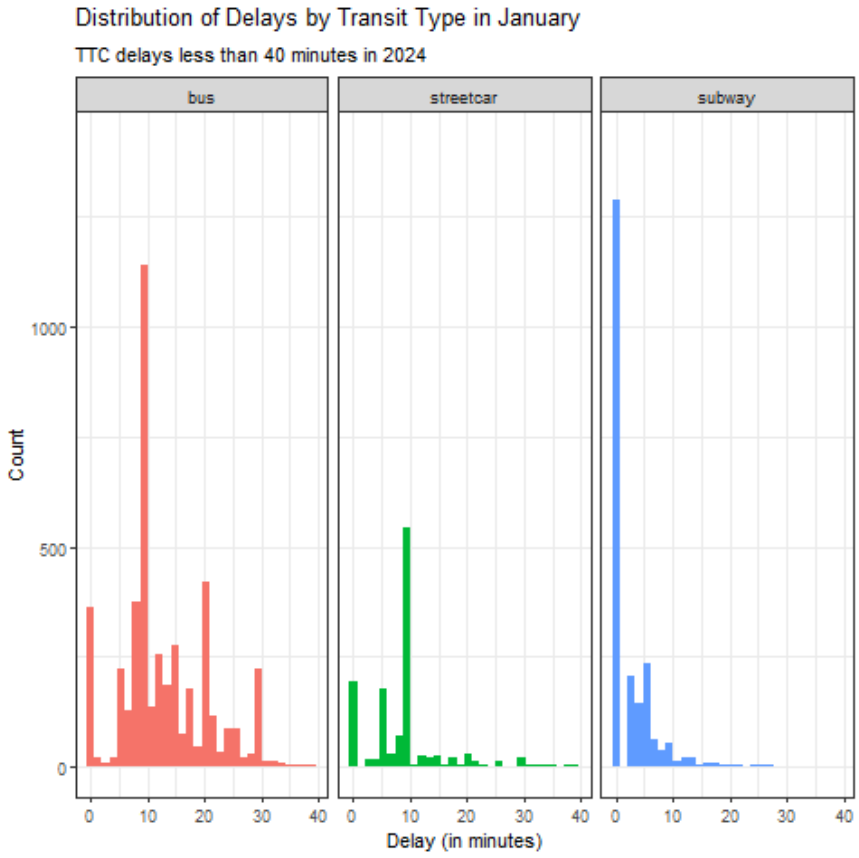

Figure 5. Distribution of delay durations by transport type, after removing outliers

From fitting high performing random forests on each transport type to investigate the effect of different predictors on both existence of delay and length of delay, with accuracies up to 92%, as shown below.

Bus:

```
Accuracy: 0.9222669840688519
              precision    recall  f1-score   support

           0       0.77      0.50      0.61      1309
           1       0.93      0.98      0.96      9613

    accuracy                           0.92     10922
   macro avg       0.85      0.74      0.78     10922
weighted avg       0.92      0.92      0.91     10922

Mean Squared Error: 1676.6484673060688
Mean Absolute Error: 11.811802878201272
R-squared: 0.4015687445549774
```

Streetcar:

```
Accuracy: 0.8707093821510298
              precision    recall  f1-score   support

           0       0.80      0.40      0.53       484
           1       0.88      0.98      0.92      2138

    accuracy                           0.87      2622
   macro avg       0.84      0.69      0.73      2622
weighted avg       0.86      0.87      0.85      2622

Mean Squared Error: 1236.2823183225041
Mean Absolute Error: 13.226827393059942
R-squared: 0.032810975640353734
```

Subway:

```
Accuracy: 0.7735890652557319
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      2889
           1       0.68      0.72      0.70      1647

    accuracy                           0.77      4536
   macro avg       0.76      0.76      0.76      4536
weighted avg       0.78      0.77      0.77      4536

Mean Squared Error: 48.93389357331211
Mean Absolute Error: 2.6807790349164353
R-squared: 0.1014800466185879
```

Figure 6. Results of the fitted random forest models

Deeper investigation into the incidents

After we fit the model, we decided to do a deeper investigation into the incidences for each transport type, in order to make targeted recommendations. To do so, we fit logistic regression

models on whether or not the incident was the top incident for each transport type: mechanical issues for buses, operational issues for streetcars, and disorderly patrons for the subway.

Our findings highlight that rush type, season, and weather conditions significantly influence delays, with winter exacerbating mechanical issues in buses and disorderly patron incidents in subways, and summer leading to more operational issues in streetcars. Based on these insights, we recommend winter maintenance for buses, optimized scheduling for streetcars, and enhanced crowd management and safety measures for subways to address these preventable disruptions.

Usage Instructions

To use the delay prediction model, ensure you are using a cleaned dataset containing relevant features such as min_delay, time-based attributes, weather conditions, and operational details. The dataset should be loaded into a Pandas DataFrame (streetcar), and all categorical features should be converted into numerical values using LabelEncoder if necessary. Additionally, make sure that the feature columns used for training are present in the test dataset.

If you are using a pre-trained model, load it using joblib.load('rf_streetcar_model.pkl'). Otherwise, you can train a new model by running the provided training script. Once the model is ready, it can be used to predict delays on new data. To make predictions on a test dataset, use y_pred_occurrence = rf_occurrence.predict(X_test). If you need to predict delays for a single new observation, create a DataFrame with the required features and pass it to the predictmethod. The output will indicate whether a delay is expected (1 for delay, 0 for no delay).

To evaluate the model's performance, compare predictions with actual outcomes using accuracy and classification metrics. The accuracy_score and classification_report functions from sklearn.metrics can provide insights into precision, recall, and overall model effectiveness. If you plan to reuse the model, save it with joblib.dump(rf_occurrence, 'rf_streetcar_model.pkl'). For deployment, the model can be integrated into an API or a dashboard to provide real-time delay predictions. This approach helps transit agencies optimize scheduling, minimize delays, and improve passenger experience.

Bibliography and Acknowledgements

"Daily Data Report for 2024 ." Environment and Climate Change Canada, Government of Canada, 1 Oct. 2024, climate.weather.gc.ca/climate_data/daily_data_e.html?hlyRange=2002-06-04%7C2025-02-28&dlyRange=2002-06-04%7C2025-02-28&mlyRange=2003-07-01%7C2006-12-01&StationID=31688&Prov=ON&urlExtension=_e.html&searchType=stnProx&optLimit=specDate&Month=1&Day=6&StartYear=1840&EndYear=2018&Year=2024&selRowPerPage=25&Line=0&txtRadius=25&optProxType=navLink&txtLatDecDeg=43.6275&txtLongDecDeg=-79.396111111111&timeframe=2.

Davis, Stephen Spencer. "Who Broke the TTC? Inside Toronto's Public Transit Disaster." Toronto Life, 6 July 2023, torontolife.com/deep-dives/who-broke-the-ttc-inside-torontos-public-transit-disaster/.

Esri Canada Education. "Toronto Subway Stations." ArcGIS, www.arcgis.com/home/item.html?id=05200e06ff524319bde9f16e5955496b. Accessed 1 Mar. 2025.

"Find TTC Schedule." Toronto Transit Commission, www.ttc.ca/routes-and-schedules/1/1/13816. Accessed 1 Mar. 2025.

Sattler, Mayson. "The TTC & Rider Backlash: Grappling with Public Perception and Performance." Queen's Business Review, 13 Nov. 2024, www.queensbusinessreview.com/articles/the-ttc-and-rider-backlash-grappling-with-public-perception-and-performance.