Analyzing Patterns and Impacts of TTC Service Disruptions (January 1st 2023 - January 31st 2025)

Introduction

The Toronto Transit Commission (TTC) operates a large and multimodal transit network composed of subways, streetcars, and buses, with each mode facing its own operational challenges, including delays that vary in frequency, cause, and impact. While TTC delay data is published separately for each mode, this project takes an integrated approach to explore system-wide patterns of service disruptions from January 2023 to January 2025.

The objective is to identify common and mode-specific causes of delays, understand how incidents vary by location and time of day, and assess the overall reliability of each transit mode. By combining visual analysis with aggregated insights, the project aims to find operational vulnerability points and find ways for future data improvements or system planning.

Data Description

Datasets

The analysis focused on TTC datasets (2023, 2024 and January 2025) publicly available through Toronto Open Data Portal:

1. TTC Subway Delay Data
    - Format: XLSX (2023, 2024), CSV (2025) ~24,000 records/year
    - Key fields: Date/Time, Line, Station, Code, Min Delay, Min Gap
    - Updated monthly, last refresh: January 2025
    - Source: https://open.toronto.ca/dataset/ttc-subway-delay-data/
2. TTC Streetcar Delay Data
    - Format: XLSX (2023, 2024), CSV (2025) ~13,000 records/year
    - Key fields: Date/Time, Line, Location, Incident, Min Delay, Min Gap
    - Updated monthly, last refresh: January 2025
    - Source: https://open.toronto.ca/dataset/ttc-streetcar-delay-data/
3. TTC Bus Delay Data
    - Format: XLSX (2023, 2024), CSV (2025) ~54,000 records/year
    - Key fields: Date/Time, Route, Location, Incident, Min Delay, Min Gap
    - Updated monthly, last refresh: January 2025
    - Source: https://open.toronto.ca/dataset/ttc-bus-delay-data

Min Delay indicates the delay, in minutes, to the schedule for the following but/streetcar/train and Min Gap indicates the occurring gap, in minutes, between busses/streetcars/trains.

Data Preprocessing

During preprocessing, a significant number of missing values were observed in the Min Delay and Min Gap columns, particularly in the 2025 subway and streetcar datasets. These missing entries appeared in the original files as empty cells and were parsed as NaN (Not a Number) by pandas. Upon manual inspection in spreadsheet software, these blanks were displayed as "None", which further indicated the absence of an actual delay record rather than data corruption or entry error.

Given the context of delay reporting, I interpreted these missing values as representing instances where no delay or gap occurred. This is consistent with the presence of explicit 0 values in similar rows and aligns with typical operational reporting logic — i.e., delay values are only filled when a disruption occurs.

Therefore, missing values in Min Delay and Min Gap were imputed as 0, indicating a lack of disruption, to ensure consistent numeric representation and avoid skewing summary statistics or visualizations due to the gaps in the data.

In addition to handling these core numeric fields, early-stage preprocessing also involved removing rows with missing values in essential structural columns: Route (for bus data), Line (for subway and streetcar), and Location (for streetcar). These columns are critical for grouping, mapping, and summarizing events, so retaining incomplete records would have introduced inconsistencies in later stages of analysis.

Remaining missing values were mostly found in optional fields such as Bound, Direction, and Vehicle. These columns were kept in the dataset for completeness, but since they are not as critical to the analysis as the delay and location fields, their missing values were not treated as a major issue.

To make the subway incident codes easier to understand, I added a new column with human-readable descriptions using a separate mapping file (subway_delay_codes_adjusted.csv). This makes it much easier to interpret the types of delays and helps with clearer analysis and visualizations.

That said, it turns out the TTC didn't include every code in their original description file. After double-checking the source, I noticed that some codes couldn't be matched:

- Subway 2023: 531 missing descriptions out of 22,901 rows
  (8 unique codes not present in the mapping)
- Subway 2024: 993 missing out of 26,423 rows
  (8 unique codes not present in the mapping)
- Subway 2025: 92 missing out of 2,099 rows
  (5 unique codes not present in the mapping)
- All Years Combined: 1,616 rows with missing descriptions out of 51,423 total
  (9 unique codes not included in the mapping)

These unmatched codes are likely newer or less commonly used and just weren't documented. I still included them in the analysis, they just just show up with a missing or NaN description.

Several inconsistencies were also identified in the Line fields for both streetcar and subway datasets. In the streetcar data, line values included floats (e.g., "501.0"), internal run codes (e.g., "RAD"), and malformed entries. To standardize these, numeric line values were extracted, converted to string format, and entries with fewer than 50 occurrences were removed as likely noise.
The subway data had similar inconsistencies, including non-standard labels like "YU / BD", "BD/YU", "LINE 1", and even station names or unrelated codes appearing in the Line field. In addition, incident codes included many rarely used or invalid entries. To simplify and focus the analysis, only the top 5 most frequent incident codes were retained, and the rest were excluded due to low occurrence and likely entry error.

As a final step, the Route column in the bus dataset and the Station column in the subway dataset were cleaned by standardizing formatting (e.g., trimming whitespace and converting to uppercase) to support basic inspection and identify inconsistencies in naming, as well as removing outliers by keeping only those stations across each transit type that are responsible for 99% of entries (surprisingly, it brought the number of unique Station entries for Subway from 651 to 194).

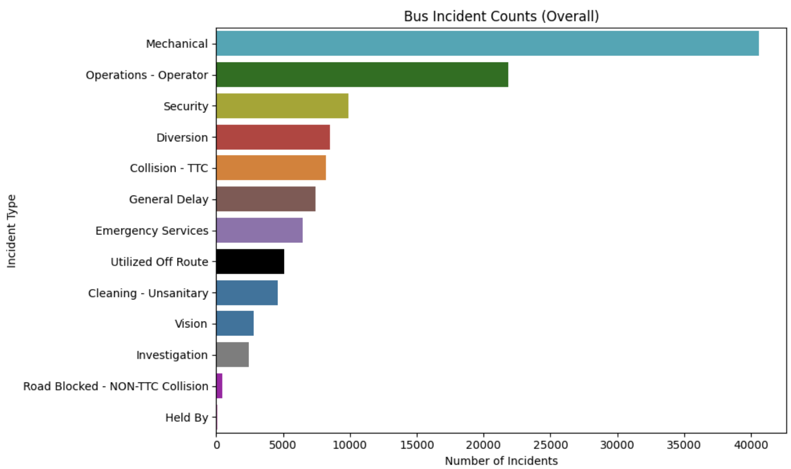
While significant effort has gone into cleaning and standardizing key fields, the data still contains inconsistencies. For example, after cleaning, there are still 194 unique subway station names, despite the TTC operating only 70 stations (with 60 more under construction). Many of these additional values are likely due to typos, alternate spellings, or inconsistent formatting.

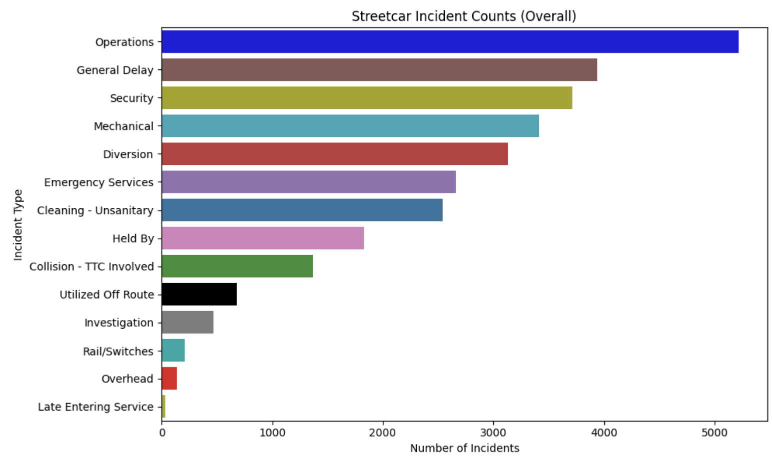
Given the scope of the project and time constraints, no further cleaning was performed. These remaining inconsistencies are acknowledged, and further refinement could improve location-based analysis in the future.

Exploratory Data Analysis (EDA)

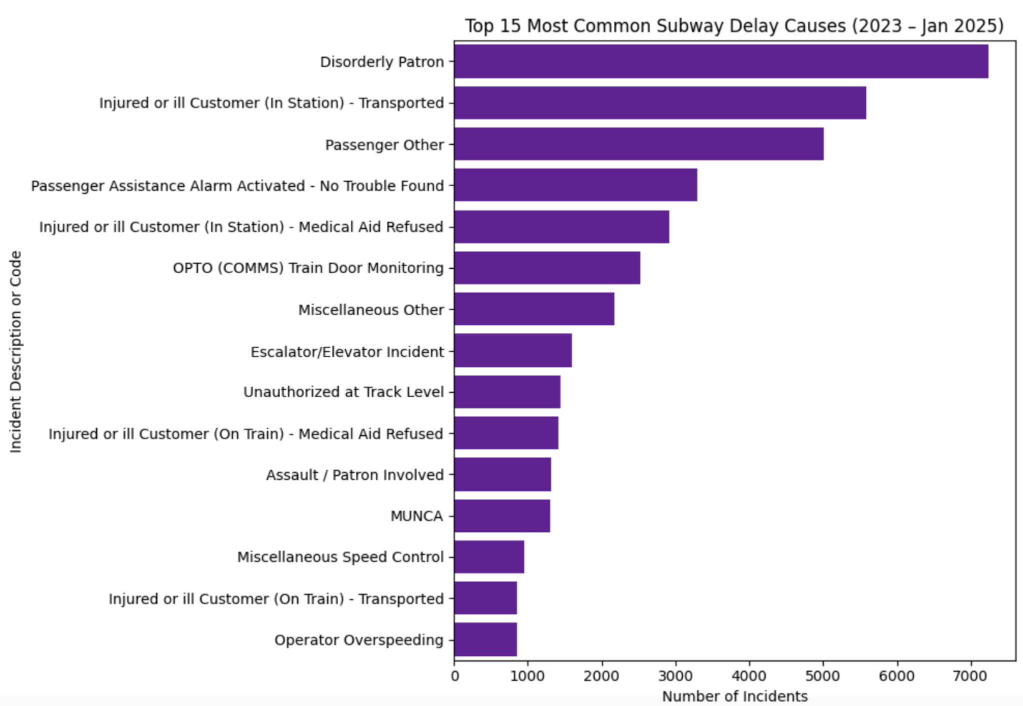
One of the first things to notice is that the number of incident types is much larger for Subways (180) than for Streetcars (14) and Buses (13). Since 9 incident types overlap between streetcars and buses, I chose to use a shared colour scheme for their histograms to reflect this overlap.

The first two charts show the most common types of bus and streetcar incidents recorded between 2023 and early 2025:

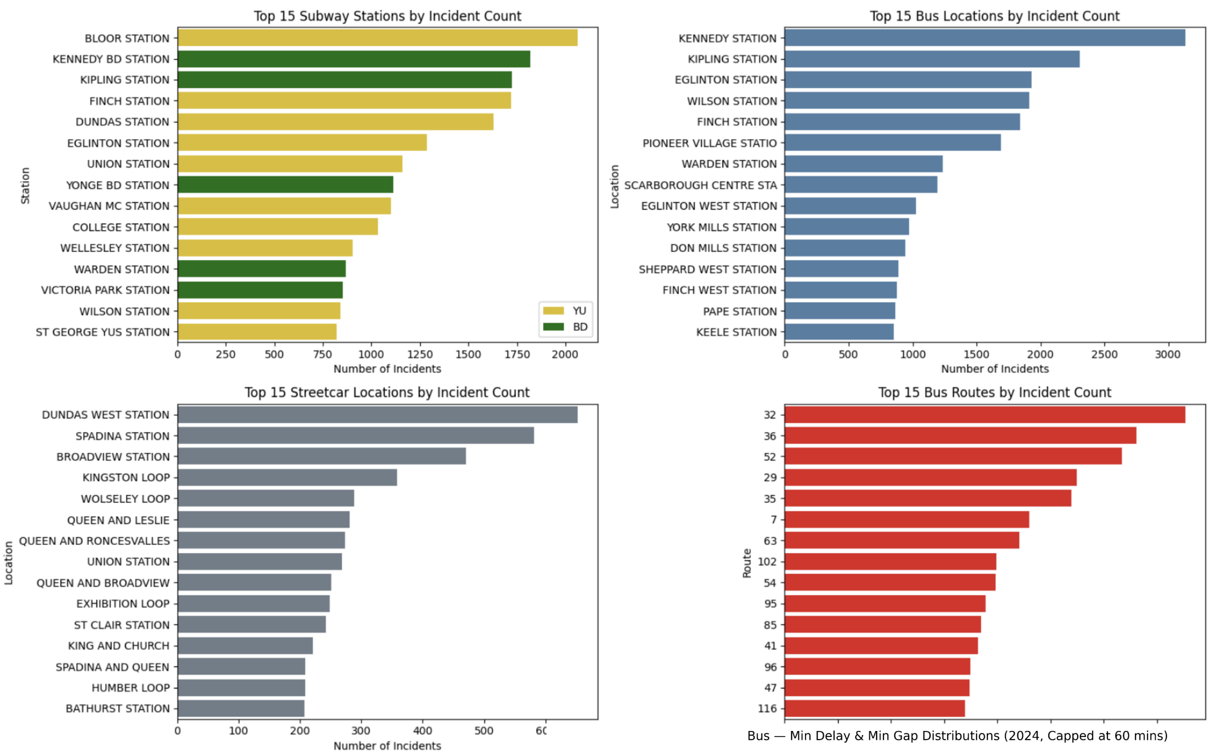




For the subway, only the top 15 most frequent incident types are plotted to keep the visualization readable:

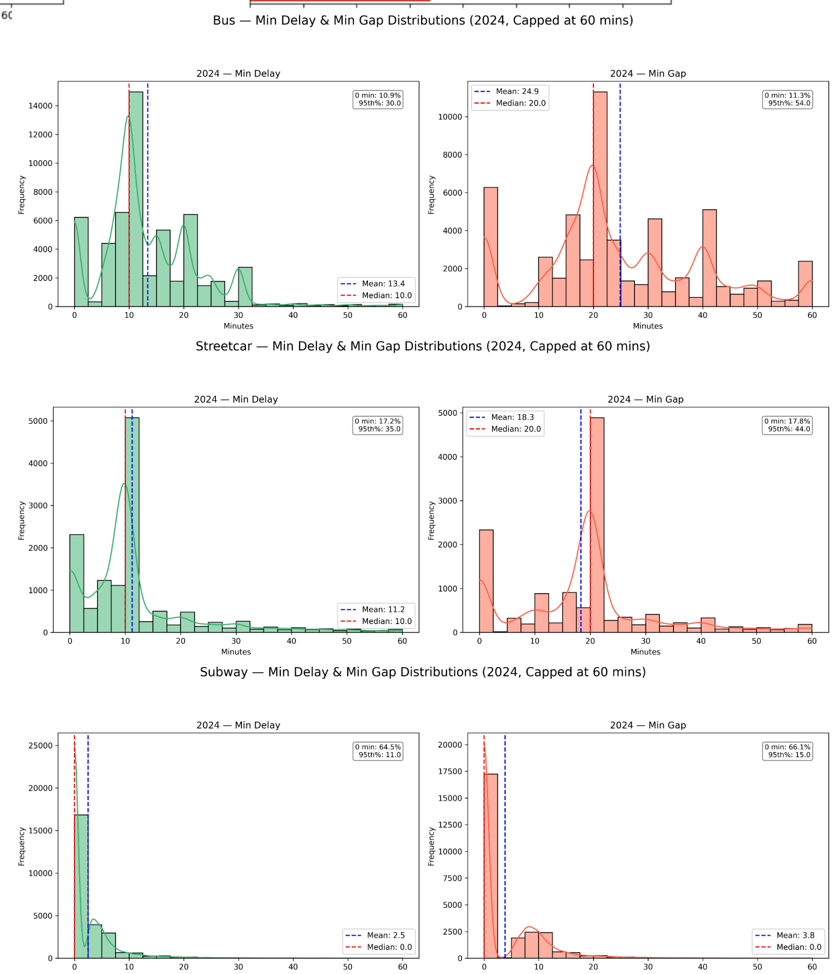


While the previous charts focused on what caused delays, this next visualization shifts focus to where those incidents were most concentrated, by looking at station-specific incident counts (with Subway stations assigned colour of their corresponding Line):

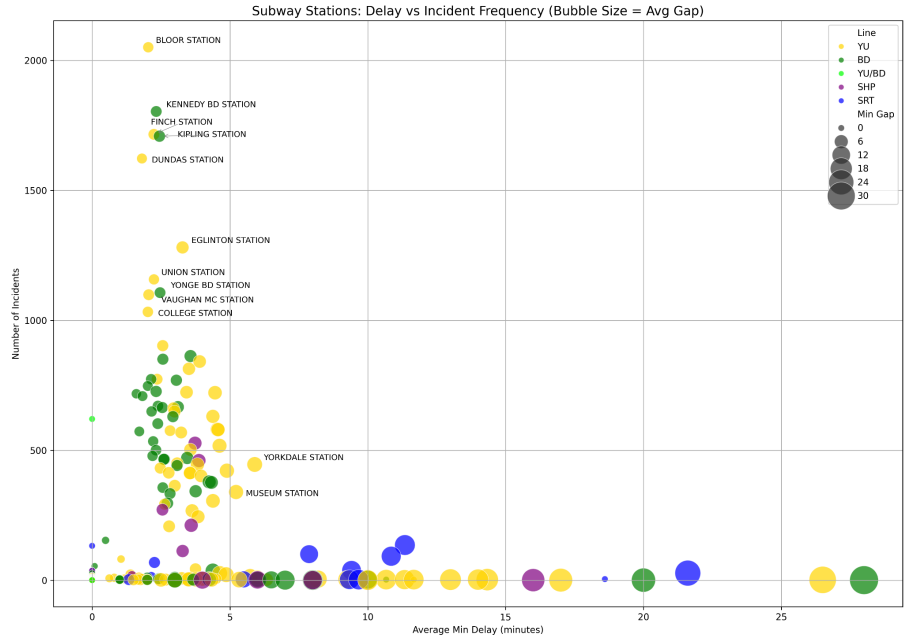


To better understand the patterns of service disruptions across different TTC transit modes, distributions of Min Delay and Min Gap were analyzed for buses, streetcars, and subways. Initially, the raw data showed extreme outliers, with some delays exceeding several hundred minutes. These long-tail values made the charts difficult to interpret, compressing the majority of shorter delays near the origin. To address this, I limited the maximum displayed delays/gaps to 60 minutes. This adjustment allowed for clearer comparison between the most frequent delay durations without discarding significant data.

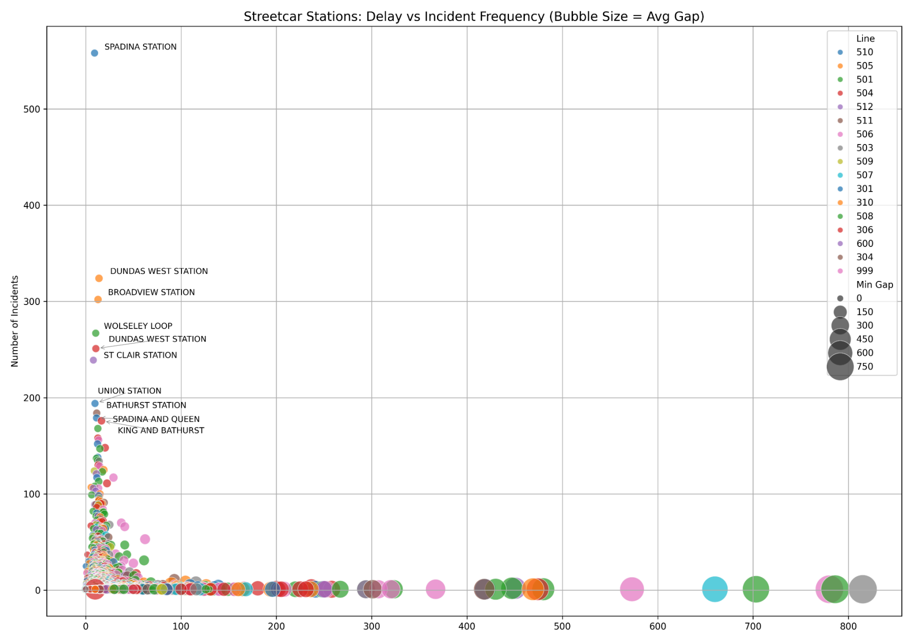
On the right hand side you can see the resulting distribution histograms for the year 2024 across all 3 types of public transportation.



After exploring basic distribution of the most important data features through simple bar charts and histograms, I wanted to explore the data further by using visualization techniques that can show more features of data at the same time.

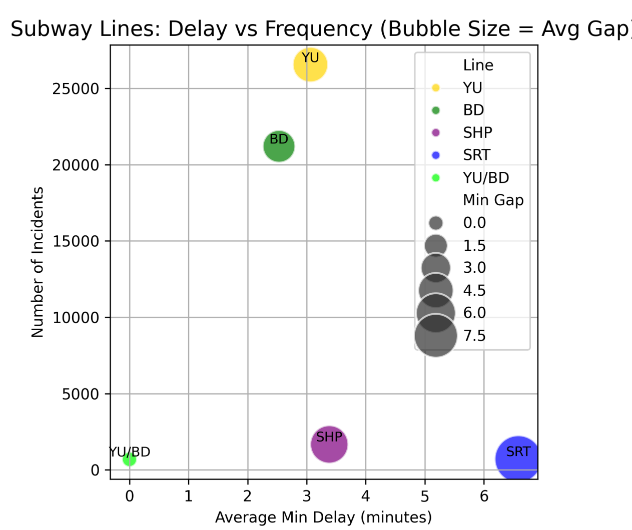


To visualize multiple aspects of station-level disruptions at once, I created bubble scatter plot where the x-axis represents the average minimum delay per station, the y-axis shows the total number of recorded incidents, the size of each bubble reflects the average minimum service gap (in minutes), and the colour indicates the subway/streetcar line each station belongs to.

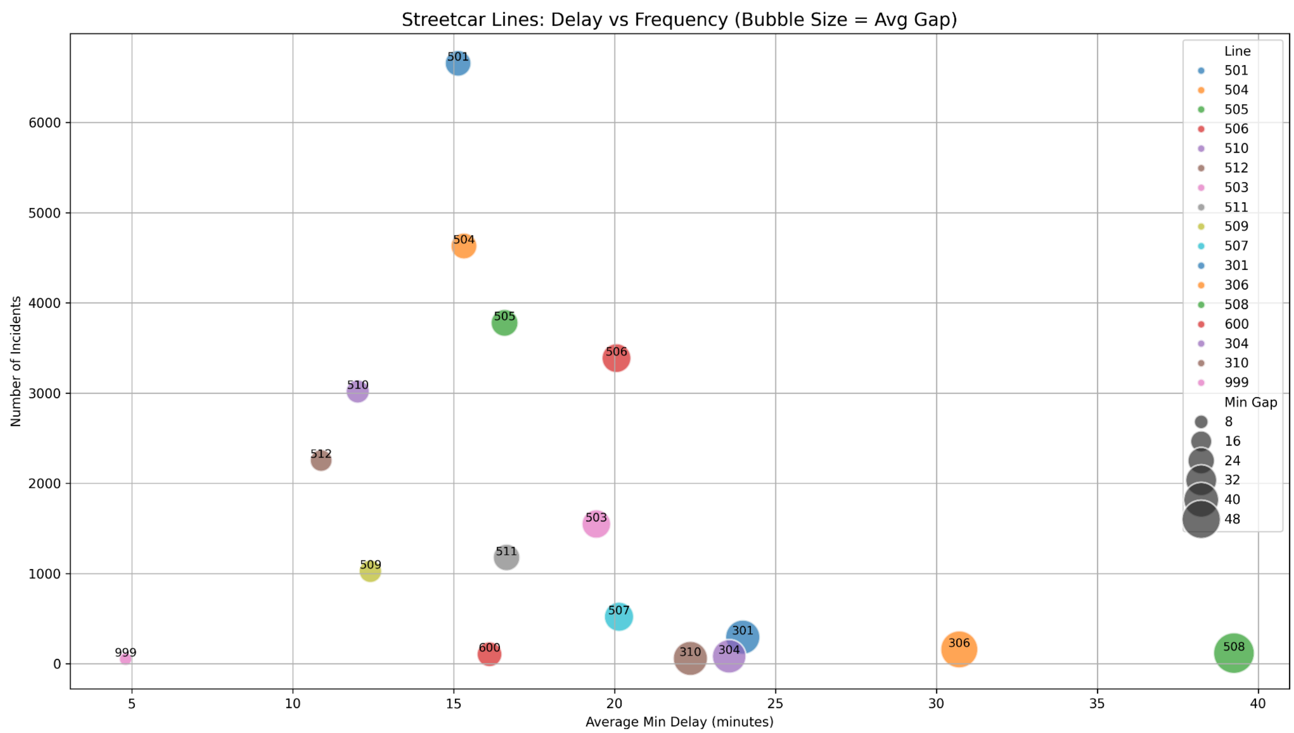


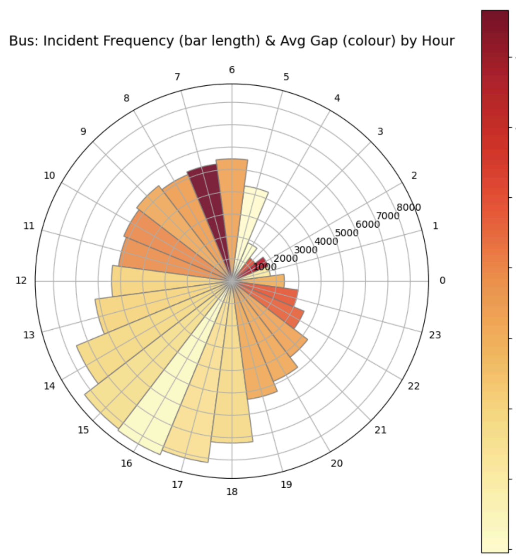
This type of visualization works particularly well for subways and streetcars, where the number of key locations is manageable, but is less effective for buses due to the significantly higher number of stops and routes, which can lead to overcrowding and visual clutter.

A key difference of frequency in these 2 charts compared to the horizontal bar charts before is that these scatter plots break down incidents by both station and transit line, whereas the bar charts counted total incidents per station regardless of line. This means stations with incidents across multiple lines are split into separate entries in the scatter plot, potentially changing which stations appear as the most incident-prone.



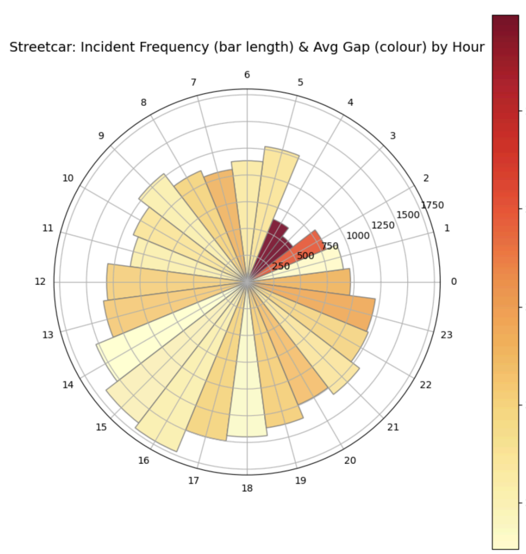
I also created similar bubble plots for both subway and streetcar lines by aggregating incidents across all stations.

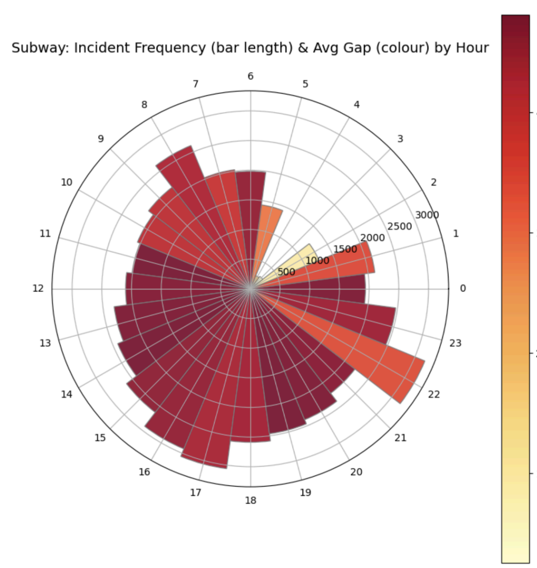




To get a better sense of when disruptions tend to occur, I visualized incident frequency and average service gaps by hour of day using radial charts, where bar length shows how often delays happen and colour represents the average gap.

I chose to visualize only the average service gap, as the correlation between delay and gap ranged from 0.95 to 0.96 across the different transit modes, making the inclusion of both features somewhat redundant.

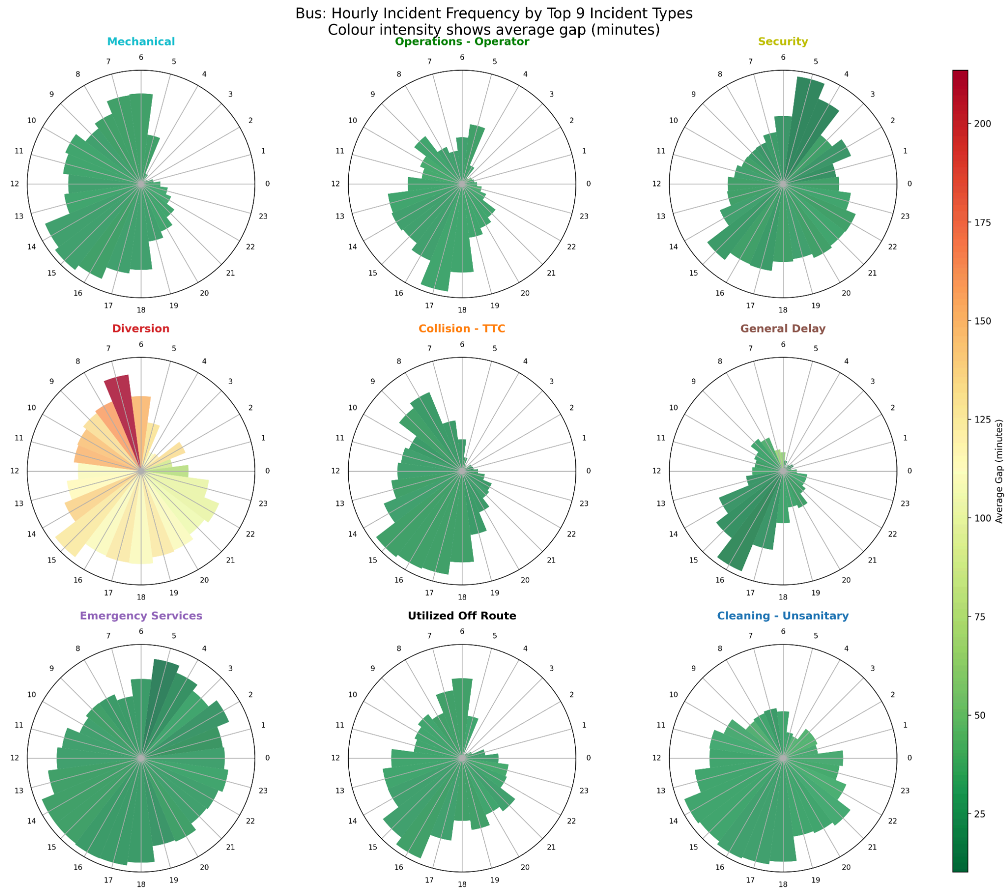


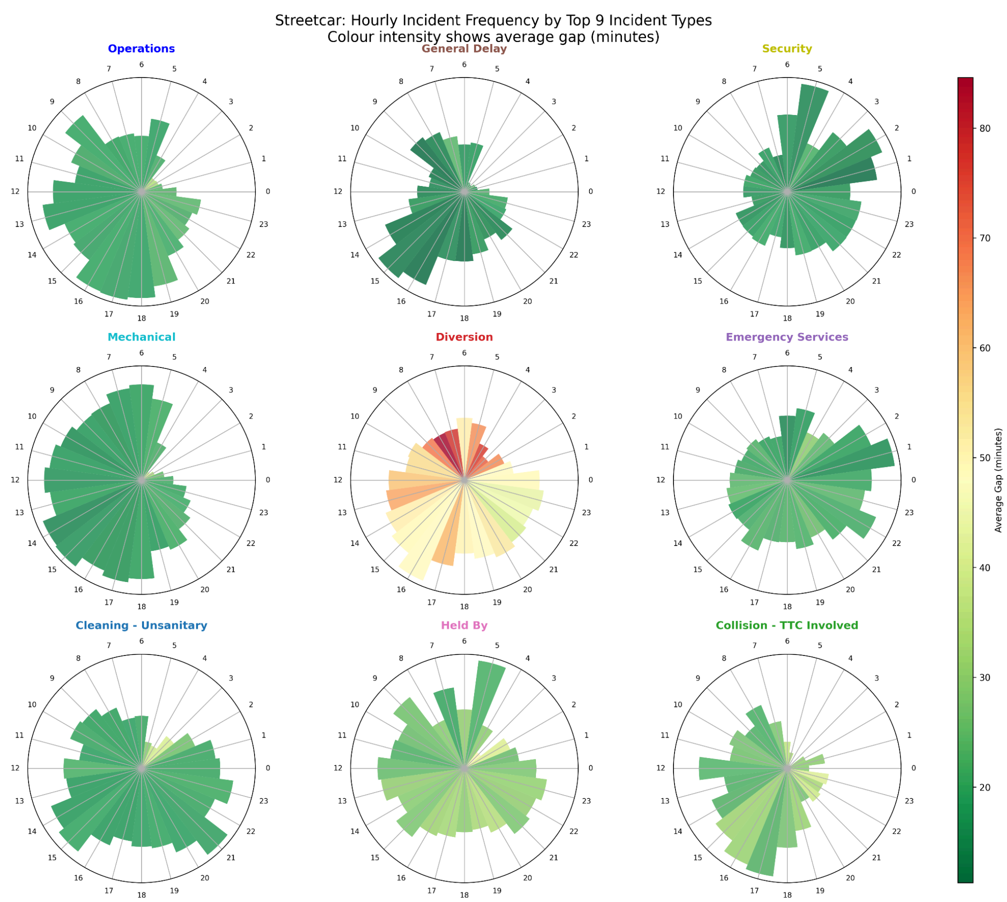


Finally, to further explore temporal patterns, I broke down the radial charts by incident type for both buses and streetcars.

Here, I reused the incident type colour scheme from the earlier horizontal bar charts for the title names.

The Summary section goes into more details of these graphs and highlights key trends and correlations uncovered through the EDA, focusing on the most common causes, locations, and time patterns of service disruptions across TTC transit modes.





Findings

The horizontal bar charts revealed clear differences in the most common delay causes across transit modes: bus delays were most often due to mechanical issues, followed by operator and security-related incidents; streetcar delays had a more diverse mix of causes, including infrastructure-specific problems; and subway delays were primarily linked to passenger behavior, such as disorderly patrons and false alarms.

The most common incident locations for buses are all connected to subway stations, most of which feature integrated bus terminals. For streetcars, hotspots include a mix of dedicated loops (such as Exhibition Loop and Kingston Loop) and subway-adjacent stations like Union and Spadina, reflecting the system's hybrid design. In the subway network, Bloor Station recorded the highest number of incidents, with three of the top five being end-of-line stations: Kennedy, Kipling, and Finch.

Interestingly, the two bus routes with the highest incident frequencies (routes 32 and 36) run along Eglinton West and Finch West, the very corridors that will be served by the upcoming Subway Lines 5 and 6, which are currently under construction.

The capped delay and gap distributions reveal distinct operational patterns across TTC transit modes. For both buses and streetcars, delay durations frequently occur in rounded increments such as 10, 15, or 20 minutes, resulting in a jagged, saw-like shape in the histograms. This pattern likely reflects how delays are typically estimated and recorded in practice. Gaps tend to cluster around 20 minutes, suggesting that service intervals often widen significantly during disruptions.

In contrast, subway service appears to be far more consistent, with over 60% of entries reporting zero delay or gap. When delays do occur, the resulting gap tends to be around 10 minutes. These differences highlight the relative reliability of subway operations compared to surface-level transit.

The station-level bubble scatter plot for subways offers a more nuanced view than the earlier bar chart, as it separates incidents by transit line. This highlights how Bloor Station rises even further above the rest, visually isolated from the next cluster of high-incident stations like Kennedy, Finch, and Kipling. Yorkdale Station stands out because of its combination of high average delay and substantial service gaps, pointing to fewer but more disruptive incidents at that location.

In the streetcar station scatter plot, Spadina Station is a clear outlier, recording the highest number of delays among all stops. Other hotspots such as Dundas West, Broadview, and Union also reappear, but none reach Spadina's incident frequency.

At the line level, the subway chart shows a clear separation between Lines 1 (Yonge–University) and 2 (Bloor), which carry the bulk of incidents but maintain relatively low average delays. In contrast, the now-decommissioned Scarborough RT line (closed in July 2023) shows the highest average delays, despite a relatively low incident count, reflecting the service reliability issues that contributed to its shutdown.

On the streetcar network, Lines 501 (Queen St), 504 (King St), and 505 (Dundas) record the highest incident volumes. In contrast, routes like 306 (Carlton) and 508 (Lake Shore) show fewer incidents but much higher delays and service gaps—suggesting infrequent but more disruptive service. A few routes such as 503(King St E/Kingston Rd) and 511 (Bathurst) perform relatively well across all metrics.

These observations, however, may be distorted by long-term service closures due to construction. For instance, Route 510 (Spadina) was replaced by buses for over half of 2024 due to major track and power upgrades. To better isolate true reliability patterns, future analysis should incorporate service availability data and visualize adjusted metrics that account for planned outages.

Finally, the radial charts reveal distinct time-based patterns in service disruptions. Bus and streetcar incidents peak during the afternoon rush hours (around 15:00-16:00), with bus service gaps also spiking during the morning period, suggesting operational strain during both commuting windows. Compared to buses, subways and streetcars show a more even distribution of incidents throughout the day. For the Subway, incident volume is especially high between 7 and 9 a.m., and again around 10 p.m., though average service gaps remain relatively low throughout

Breaking incidents down by type reveals both shared and mode-specific time patterns. Mechanical issues are consistently concentrated between 6 a.m. and 6 p.m. for both buses and streetcars, likely reflecting active service hours and prompt midday detection and resolution. General delays show a sharp peak during the afternoon rush (15:00 - 16:00), aligning with overall traffic strain. Diversions stand out for their extremely high service gaps despite lower frequency, especially around the 5–6 a.m. window, suggesting early morning disruptions with lasting impact. Security-related incidents in both modes also peak early in the morning, around 5 a.m., possibly tied to sheltering or loitering activity before full service begins. Emergency service interventions remain relatively low-volume across the day, but for streetcars, they show additional late-night peaks around 10 p.m., 1 a.m., and again near 6 a.m. Finally, unsanitary condition reports display divergent behaviors: they remain elevated into the early evening for buses, but drop off for streetcars before rising again around 9-11 p.m., a pattern that may reflect downtown service concentration and late-night crowding. These insights suggest that incident timing is influenced not only by ridership demand, but also by the geography, infrastructure, and social dynamics unique to each transit mode. Future visualizations, such as temporal heatmaps or geographic maps showing the spatial distribution of incident types, could help contextualize these patterns by comparing them against scheduled service levels and highlighting differences between the urban core and suburban areas.

Results

To clearly communicate the patterns uncovered through EDA, I implemented a range of targeted visualization techniques aligned with the structure of the data.

Horizontal bar charts were used to display the most frequent incident types across buses, streetcars, and subways. A shared colour scheme for buses and streetcars reflected overlapping categories, and for clarity, subway data was limited to the top 15 incident types. This colour scheme was later reused in the radial charts to maintain visual consistency across the report.

To support geographic and line-specific interpretation, I adopted the official TTC colour scheme when visualizing subway stations and lines, this was used consistently across horizontal bar charts and scatter plots to strengthen visual associations.

Bubble scatter plots were introduced to capture multiple dimensions: incident count, average delay, and service gap, at both the station and route level. These proved especially effective for subways and streetcars, where the number of stops is limited and visual clutter remained manageable.

Radial charts were used to explore hourly delay patterns, with bar length representing incident frequency and colour indicating average service gap. Given the strong correlation between delay and gap (0.95–0.96), only gap was visualized. These charts were further broken down by incident type.

Design decisions such as outlier capping, consistent colour encoding, and splitting plots by transit mode or line were made intentionally to support clarity and accessibility for both technical and general audiences.

Conclusions

This project revealed important contrasts in the nature and timing of service disruptions across TTC transit modes. Subways were generally more reliable, with delays often tied to in-station issues and passenger behavior. In contrast, buses and streetcars were more frequently impacted by mechanical failures, operator-related incidents, and infrastructure challenges. These surface-level-based modes of transportation also displayed unique patterns in delay timing and severity that reflect both how and where they operate across the city.

Interestingly, the two bus routes with the highest incident frequencies (routes 32 and 36) run along Eglinton West and Finch West, the very corridors that will be served by the upcoming Subway Lines 5 and 6, which are currently under construction. This points to a positive step in TTC's long-term planning: bringing higher-capacity rapid transit to corridors that are currently vulnerable to service interruptions.

While bus and streetcar networks share many of the same incident categories, this analysis revealed meaningful differences in how those categories are distributed and when they tend to occur. For instance, unsanitary condition reports show a divergent time-based pattern: they remain elevated into the early evening for buses, but drop off earlier for streetcars before rising again around 9–11 p.m. This could reflect the difference in geographic service areas—buses serving the broader city throughout the day, while streetcars are more concentrated in the urban core, with increased late-night activity. These kinds of overlaps and divergences point to operational realities shaped by the network's geography, infrastructure, and rider behavior.

Recommendations and Future Directions

The usefulness of the TTC delay data is limited by how inconsistent the key fields are, especially Line and Station. Because there are no enforced naming rules or dropdown selections for these fields, the entries contain typos, formatting differences, and sometimes completely unrelated values. For example, there originally were 658 unique Subway stations entries, even though TTC has a total of only 70 operating stations with an additional 60 under construction. This makes it very hard to match the delays to other sources, like Toronto Open Data's very own station coordinates dataset, which prevents mapping the disruptions on a city map.

A clear recommendation is to standardize these fields at the data entry stage. For example, having dropdown menus for selecting lines or stations would help prevent typos and duplicates. Additionally, incorporating semi-automated delay tracking, such as GPS-based time logs or onboard sensors, could improve the accuracy and consistency of delay by reducing the he rounded delay/gap durations (e.g., 10, 15, 20 minutes) that currently dominate the dataset, suggesting manual estimation.

Once these foundational data issues are addressed, more advanced location-based analysis becomes possible, such as building an interactive map of delays across the city to highlight network hotspots, identify underserved areas, and visualize differences between transit modes. With cleaner data, future projects could also compare disruptions to infrastructure upgrades, neighborhood demographics, or population density projections to better understand where interventions are most needed.