TTC Subway Delay Analysis

Omina Nematova

2025-03-17

Project GitHub: https://github.com/omina26/TTC-Subway-Delay-Analysis

Introduction

As businesses increasingly encourage employees to return to in-office work, reliable transportation is becoming more important than ever for commuters. Many workers rely on subway systems for their daily commutes. Subway delays can disrupt schedules, increase travel time uncertainty, and even influence long-term decisions about where people choose to live or work. While some delays may be unavoidable, understanding when and where these delays occur, as well as how external factors like weather conditions and population density contribute to service disruptions can help both commuters and policymakers make informed decisions.

This study investigates when and where subway delays occur within the Toronto Transit Commission (TTC) system and analyzes the impact of weather conditions and service population density on these delays using a merged dataset from multiple sources. Subway delay reports (see reference) were obtained from the City of Toronto Open Data Catalogue, providing details on location, time, and cause of delays across the TTC subway network. Service population estimates were derived from Toronto Neighbourhood Profiles (see reference), also sourced from the Open Data Catalogue, while spatial data for neighbourhoods and station locations was incorporated using the Toronto Neighbourhoods shapefile (see reference), TTC Subway and Streetcar Map (see reference) and OpenStreetMap API (see reference). Additionally, hourly weather data from 2024 was retrieved using the Open-meteo API (see reference), allowing for an assessment of temperature, precipitation, and other environmental factors at the time of each recorded delay.

Methodology

To analyze the occurrence of subway delays and the potential impact of weather conditions and population density, this study integrates multiple datasets, combining transit delay reports, weather data, and demographic information. The TTC Subway Delay Data from 2024, acquired from the City of Toronto Open Data Catalogue, consists of 26,467 entries with details on the date, time, day of the week, station, delay code (cause of delay), minutes of delay, minutes of gap between trains, train direction (bound), subway line, and vehicle ID. For this analysis, all columns were retained except for minutes of gap between trains and vehicle ID. To account for population density, Toronto Neighbourhood Profiles data from 2021, also sourced from the City of Toronto Open Data Catalogue, was incorporated. This dataset includes 158 entries with demographic information, from which only neighbourhood ID and total population were used to estimate the service population surrounding each station. Additionally, Toronto Neighbourhoods shapefile data, obtained from the City of Toronto Open Data Catalogue, contained 158 entries and 12 columns and was used for geographic mapping.

To further enrich the analysis, a custom subway station dataframe was created by manually collecting the names of all 70 subway stations from the TTC Subway and Streetcar Map and retrieving station longitude and latitude coordinates using the OpenStreetMap API to allow for spatial integration with other dataframes. Weather conditions were incorporated using hourly weather data from 2024, retrieved from the Open-Meteo API, which contained 8,784 entries with columns detailing the date, temperature, relative humidity, apparent temperature, precipitation, rain, snowfall, snow depth, cloud cover, wind speed, wind direction, and wind gusts. All columns from this dataset were used to assess the impact of different weather variables on subway delays.

Once collected, the datasets underwent extensive cleaning and preprocessing to ensure consistency and usability. The TTC Subway Delay Data, originally containing 10 columns, was filtered to remove minutes of gap between trains and vehicle ID. Columns were renamed for clarity, and duplicate rows or rows with null values in the bound or line columns were removed. Subway station names were standardized, and a new hour column was created to facilitate merging with the weather dataframe. The day, station, code, bound, and line columns were converted to categorical types to improve usability during analysis. Most importantly, only observations with delays greater than zero minutes and less than or equal to 30 minutes were retained, reducing the dataset to 9,038 rows. The population dataframe was created by filtering the Toronto Neighbourhood Profiles data to retain only neighbourhood ID and population estimates. In the weather dataset, null values in the snow depth column were filled with zero, and an hour column was generated from the date column.

Data wrangling was performed to integrate these datasets into a unified structure. First, the Toronto Neighbourhoods shapefile was merged with the population dataframe using neighbourhood IDs. Next, the stations dataframe was spatially joined with the neighbourhoods shapefile to identify neighborhoods within 0.5 km of a station, allowing service population estimates to be merged into the stations dataframe. The delay dataframe and weather dataframe were then merged on date and hour to create a comprehensive dataset. Finally, service population counts for each station were merged into this final dataset using station names as the key identifier.

To understand patterns in subway delays, summary statistics were computed for all numerical variables in the dataset. A histogram was generated to visualize the distribution of delay durations, highlighting the frequency of different delay times. The most common delay codes were identified, and their frequency was plotted to determine the predominant causes of disruptions. Additionally, delays were analyzed across subway lines, train directions (bounds), and stations to assess variations in delay occurrence by location.

Temporal analysis was conducted by examining delay frequencies and average delay durations by hour of the day. A bar plot illustrated the number of delays per hour, while a line plot showed fluctuations in average delay durations throughout the day. Spatial trends were further explored by identifying the top three stations per subway line with the highest delay occurrences and comparing their average delay durations. A regression analysis was performed to evaluate the relationship between service population and delay duration, providing insights into whether higher-density areas experience more frequent or prolonged delays.

A correlation heatmap was generated to assess relationships between weather variables and subway delays. Finally, an ANOVA test was conducted to determine whether delay durations significantly differed across days of the week, stations, subway lines, and subway bounds, offering statistical validation of observed trends.

Results

To begin the analysis, summary statistics were computed for key numerical variables, including subway delay durations, weather conditions, and service population.

Table 1: Table 1: Summary Statistics of Merged Data

	Count	Mean	Std Dev	Min	Q1	Median	Q3	Max
Min Delay	9028.00	6.56	4.79	2.00	4.00	5.00	7.00	30.00
Hour	9028.00	12.96	6.02	0.00	8.00	13.00	18.00	23.00
Temperature	9028.00	9.66	9.46	-16.01	1.94	9.44	17.79	30.49
Relative Humidity	9028.00	73.68	15.41	21.45	62.80	75.17	86.09	100.00
Apparent Temperature	9028.00	7.17	11.79	-21.34	-2.51	6.09	17.18	33.97
Precipitation	9028.00	0.12	0.56	0.00	0.00	0.00	0.00	14.40
Rain	9028.00	0.11	0.55	0.00	0.00	0.00	0.00	14.40
Snowfall	9028.00	0.01	0.06	0.00	0.00	0.00	0.00	1.89
Snow Depth	9028.00	0.01	0.02	0.00	0.00	0.00	0.00	0.10
Cloud Cover	9028.00	62.90	41.79	0.00	15.00	90.00	100.00	100.00
Wind Speed	9028.00	13.45	6.73	0.00	8.43	12.40	17.69	39.12
Wind Gusts	9028.00	27.61	12.75	1.44	18.00	25.92	35.64	80.28
Service Population	9028.00	53200.92	21773.92	0.00	36845.00	53450.00	71680.00	105725.00

In Table 1, we see that the average delay duration is around 6.56 minutes, with weather factors like temperature (avg. 9.66°C), relative humidity (73.68%), and wind speed (13.45 km/h) included. Precipitation and snowfall are minimal, while service population varies widely (avg. 52,123).

Understanding the distribution of delay durations is necessary in assessing how frequently subway delays occur and their typical length. A histogram was generated to visualize the frequency distribution of delay durations

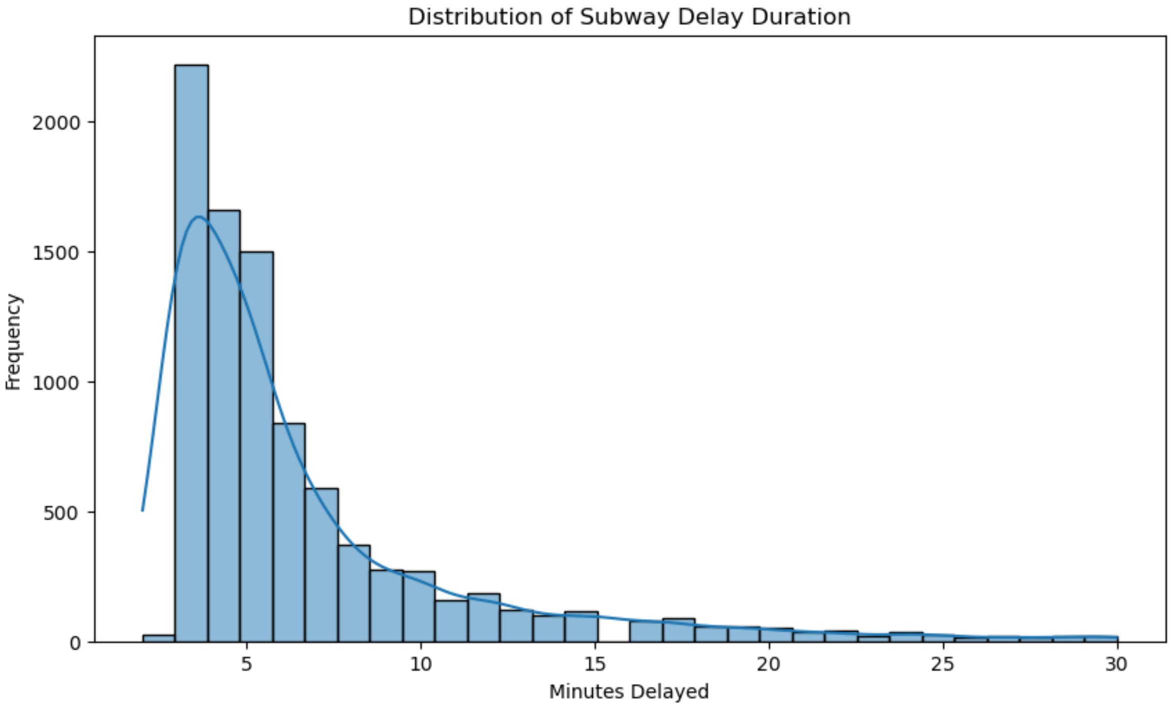

Figure 1: This plot shows the distribution of subway delay durations.

In Figure 1,it appears the distribution of subway delay durations is left skewed with center around 5. This suggests most subway delays don't last longer than a few minutes.

To further explore the nature of subway delays, the most common delay codes were identified and visualized in a bar chart.

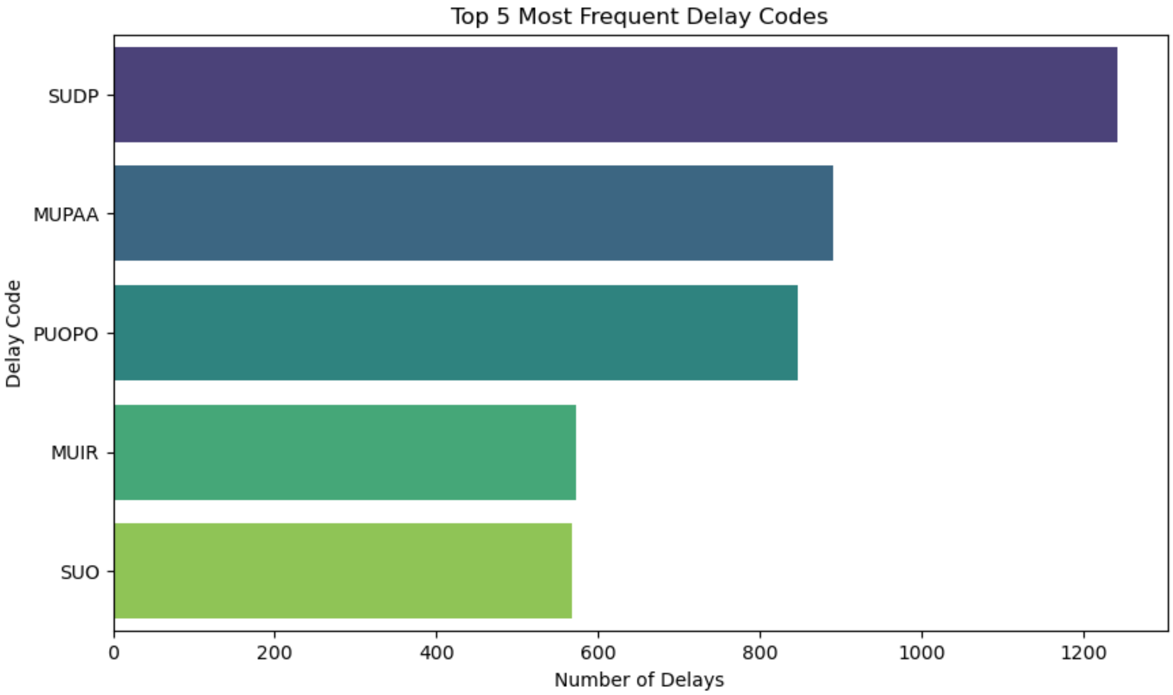

Figure 2: This plot shows the most frequent delay codes.

From Figure 2, we see the most frequent delay codes in our data. According to the documentation provided in the TTC Subway Delay data (see reference), 'SUDP' is the delay code for 'Disorderly Patron', 'MUPAA' is the delay code for 'Passenger Assistance Alarm Activated - No Trouble Found', 'SUO' is the delay code for 'Passenger Other', 'PUOPO' is the delay code for 'OPTO (COMMS) Train Door Monitoring', and 'MUIR' is the delay code for 'Injured or ill Customer (On Train) - Medical Aid Refused'. All of these delay codes are possible when stations are busy and full of commuters.

Given that different subway lines may experience varying levels of congestion and maintenance challenges, a boxplot was used to compare the distribution of delays across lines.

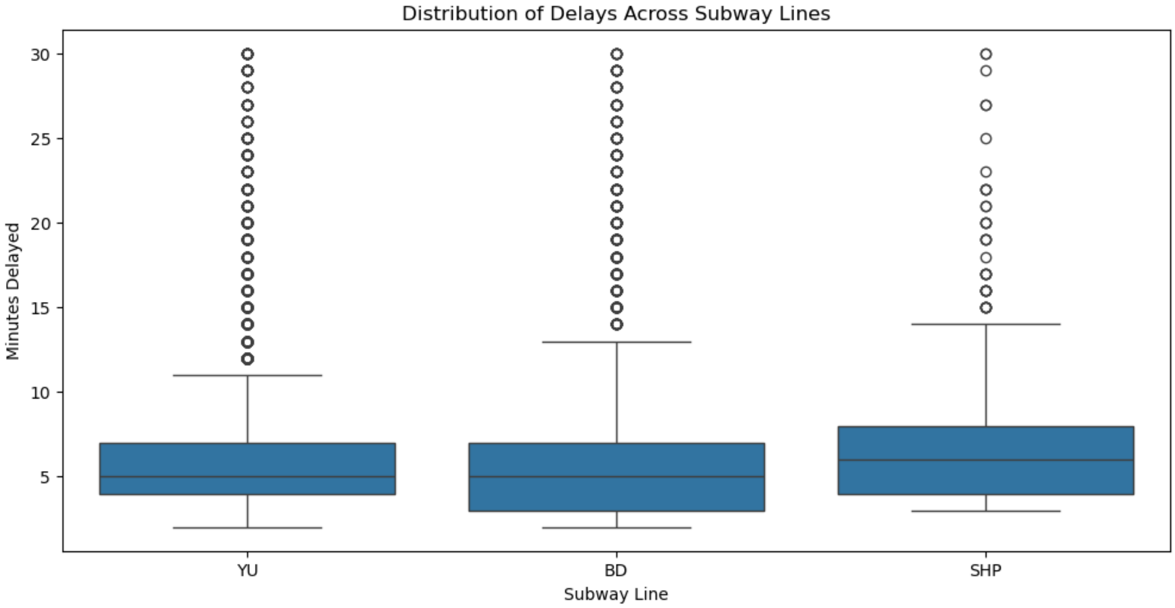

Figure 3: This plot shows the distribution of delays accross subway lines.

Figure 3 suggests that Sheppard (SHP) line, is more prone to prolonged delays. Building on the subway line analysis, a bar chart was created to highlight the three most delay-prone stations per line.

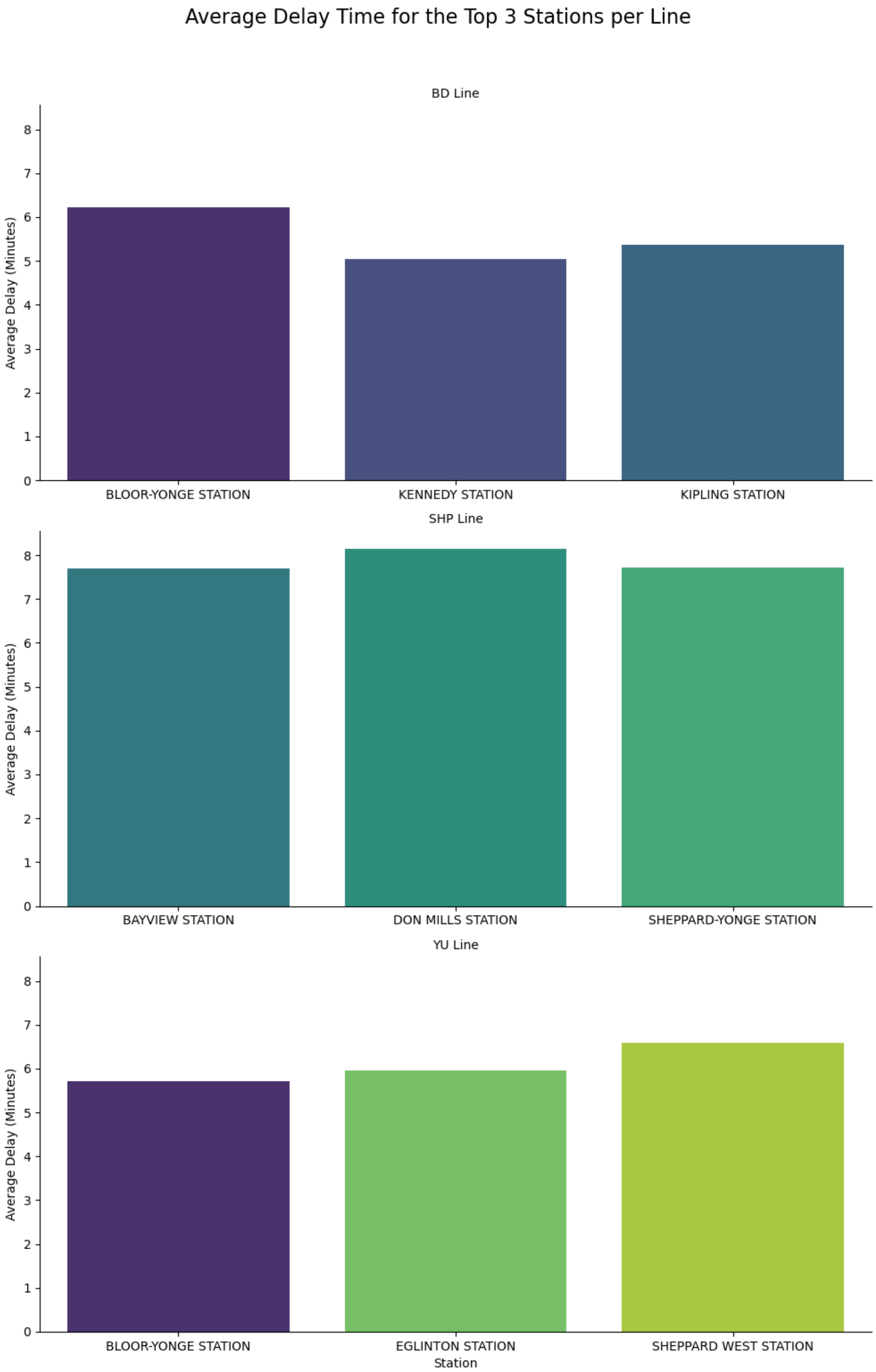

Figure 4: This plot shows the average delay time for the top 3 most delayed stations per subway line.

According to Figure 4, the most delayed stations on the Bloor-Danforth (BD) line are Bloor-Yonge, Kennedy, and Kipling stations. Further, the most delayed stations on the Sheppard (SHP) line are Bayview, Don Mills, and Sheppard-Yonge stations. Lastly, the most delated stations on the Yonge-University line are Bloor-Yone, Eglinton, and Sheppard West stations. All of these stations average a delay duration of over 5 minutes.

Another dimension of analysis involves train direction, as certain routes may be more susceptible to delays due to traffic flow or infrastructure constraints. A boxplot was used to compare the distribution of delays across train directions (northbound, southbound, eastbound, westbound), identifying whether specific bounds experience more disruptions.

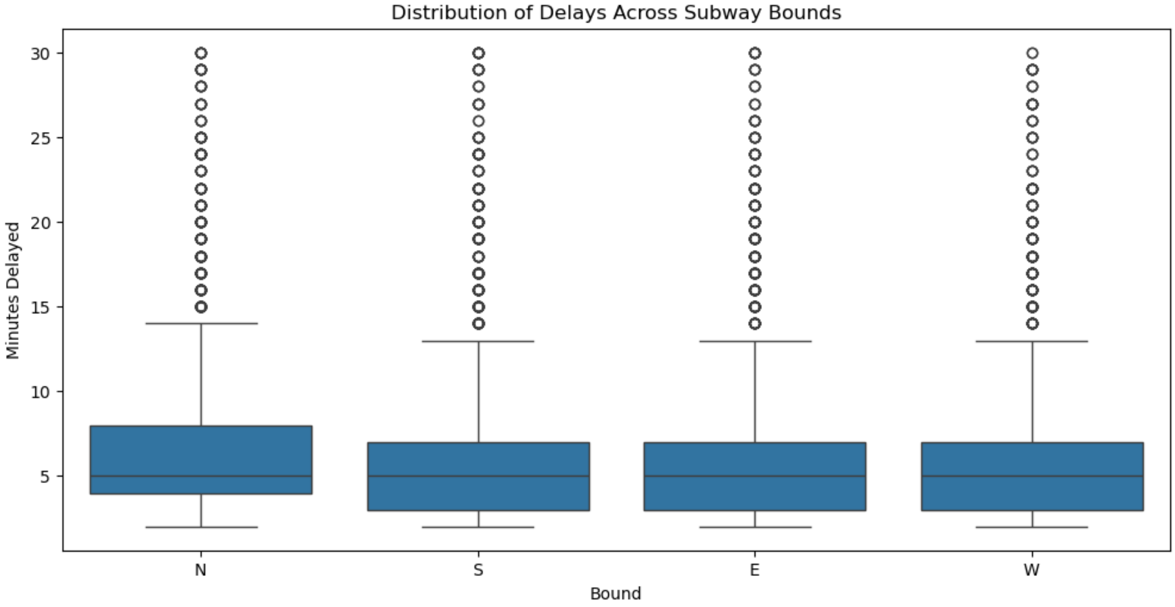

Figure 5: This plot shows the distribution of delays accross subway bounds.

In Figure 5, it appears northbound subways are more prone to longer disruptions with the other three having similar means and ranges.

Since subway usage fluctuates throughout the day, analyzing delays by hour can reveal peak periods of service disruption. A bar chart was used to visualize the number of delays occurring at each hour, helping to pinpoint when delays are most frequent. While delay frequency is important, understanding the average delay duration at different times of the day provides additional insight. A line plot was also generated to track fluctuations in average delay durations, highlighting whether peak travel hours correspond to longer delays.

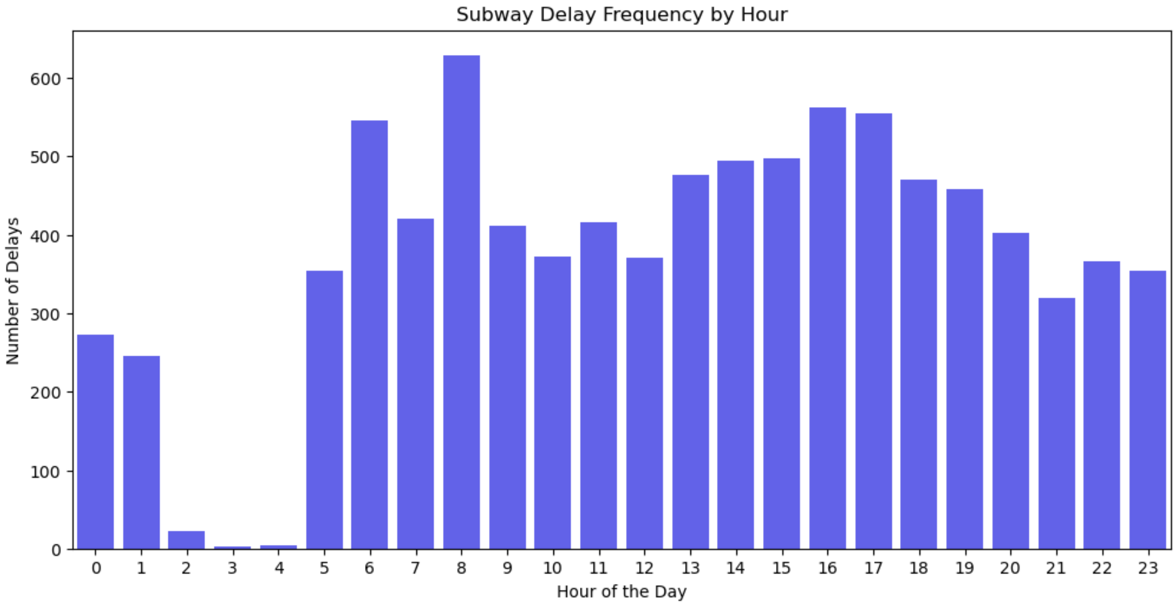

Figure 6: This plot shows the frequency of delays by hour.

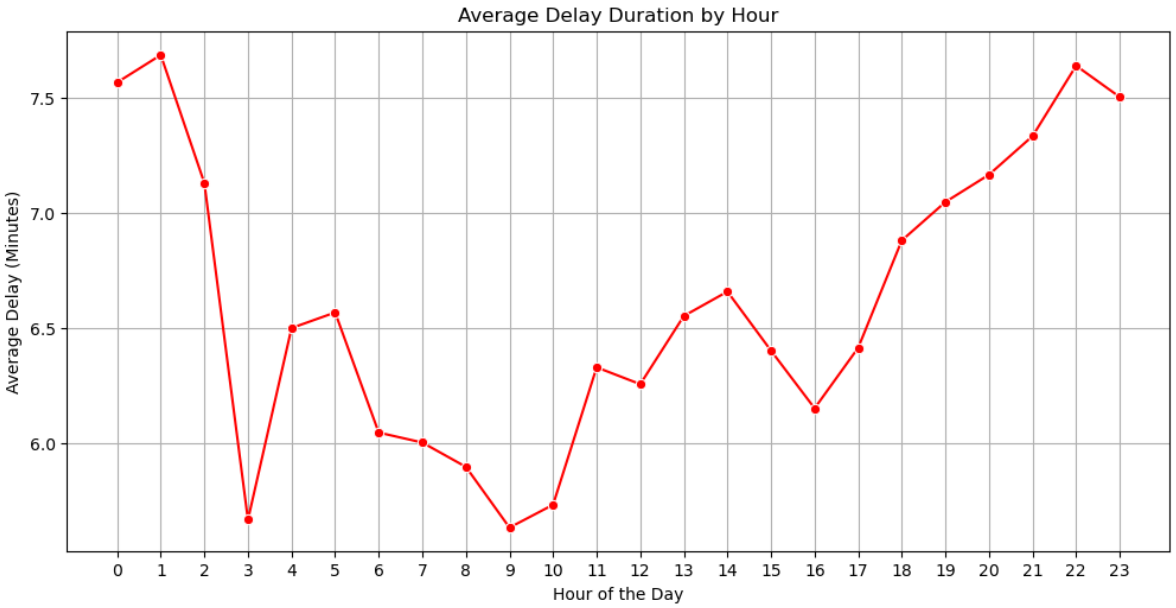

Figure 7: This plot shows the delay duration by hour.

Based on Figures 6 and 7, subway delays are most frequent during morning and afternoon rush hours. However, the average delay duration is shorter during these periods compared to non-rush hours. This aligns with expectations, as the TTC likely prioritizes rapid resolution of delays during peak commuting times to minimize disruptions for the highest volume of passengers.

Beyond daily variations, subway performance may also differ across days of the week. A boxplot was used to visualize delay distributions for each day, allowing for the identification of patterns that may be linked to weekday rush hours or weekend maintenance work.

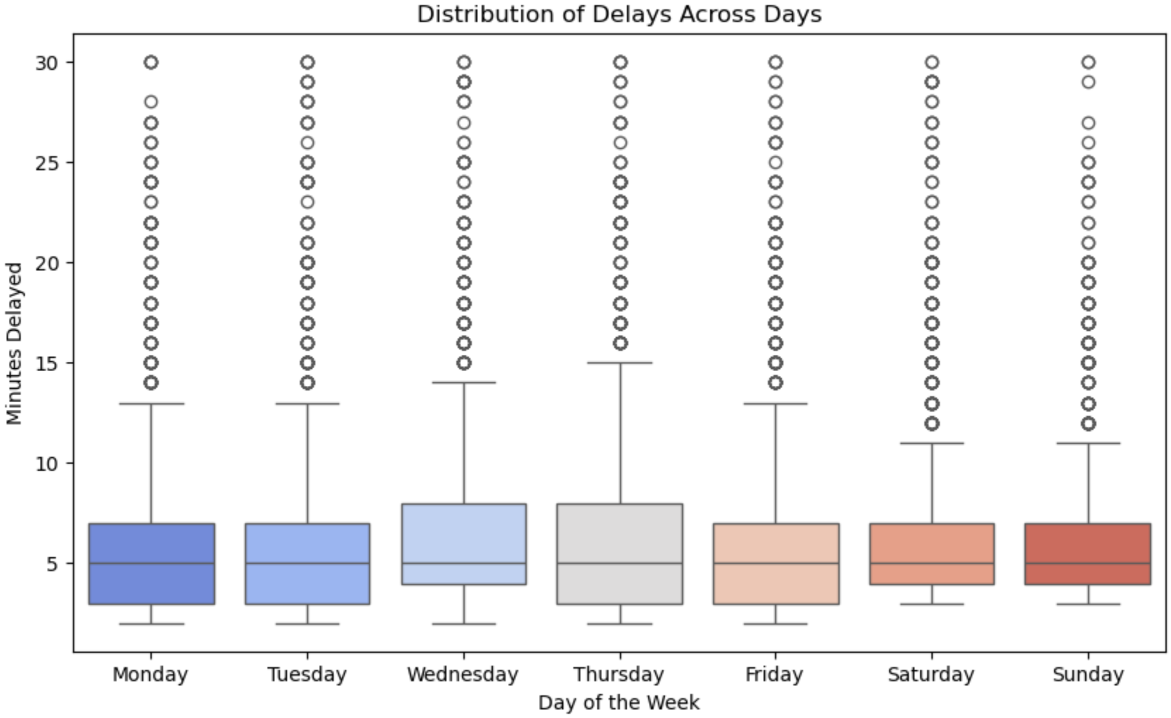

Figure 8: This plot shows the distribution of delays accross days of the week.

According to Figure 8, the average delay duration remains consistent across all days of the week, but the range of delays varies. Wednesdays and Thursdays show the widest range, while Saturdays and Sundays have the narrowest. This may be due to higher passenger volumes on weekends, potentially leading to longer but more consistent delays.

To further validate the observed trends, an ANOVA test was conducted to determine whether significant differences exist in delay durations across days, bounds, subway lines, and subway stations.

Table 2: Table 2: ANOVA Results on the Impact of Subway Line on Delay Time

	Degrees of Freedom	Sum of Squares	Mean Square	F-Statistic	p-Value
Day	6.0000	170.9064	28.4844	1.2722	0.2663
Bound	3.0000	364.6444	121.5481	5.4288	0.0010
Line	2.0000	269.9193	134.9596	6.0278	0.0024
Station	69.0000	5561.1966	80.5971	3.5998	0.0000
Residual	8947.0000	200319.2190	22.3895	nan	nan

Based on the results of our ANOVA test in Table 2, bound, subway line, and station were found to significantly affect the duration of delay as they all three has p-values less than 0.05. On the other hand, the day of the week was not found to significantly affect the duration of delay as it had a p-value greater than 0.05.

To assess whether population density influences subway delays, a scatter plot was created to examine the correlation between service population size and delay duration. This analysis helps determine whether higher-density areas experience longer or more frequent delays.

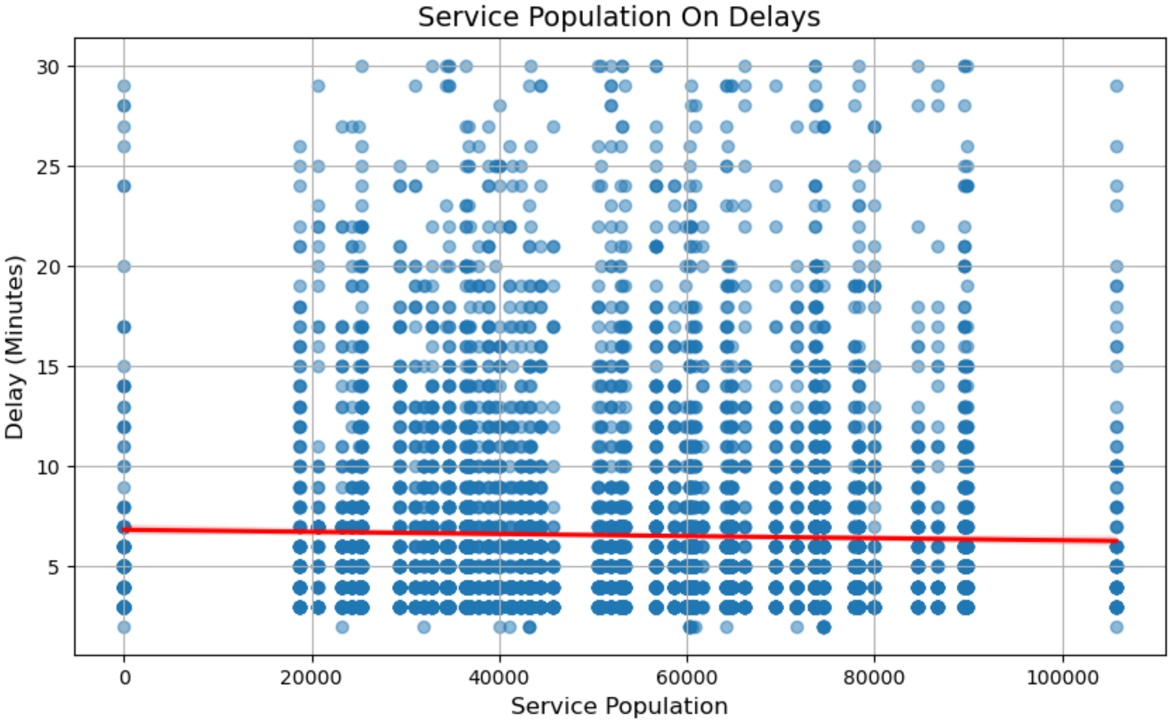

Figure 9: This plot shows the impact of service population size on the duration of a delay.

By Figure 9, the variation in delay duration appears consistent across both high- and low-density areas. However, delays seem less frequent in higher-density areas, as the data points are more concentrated on the left side of the plot.

Given that transit performance is influenced by multiple factors, weather conditions play a crucial role. To quantify this impact, a correlation heatmap was generated to examine the relationships between subway delays, weather conditions, and service population.

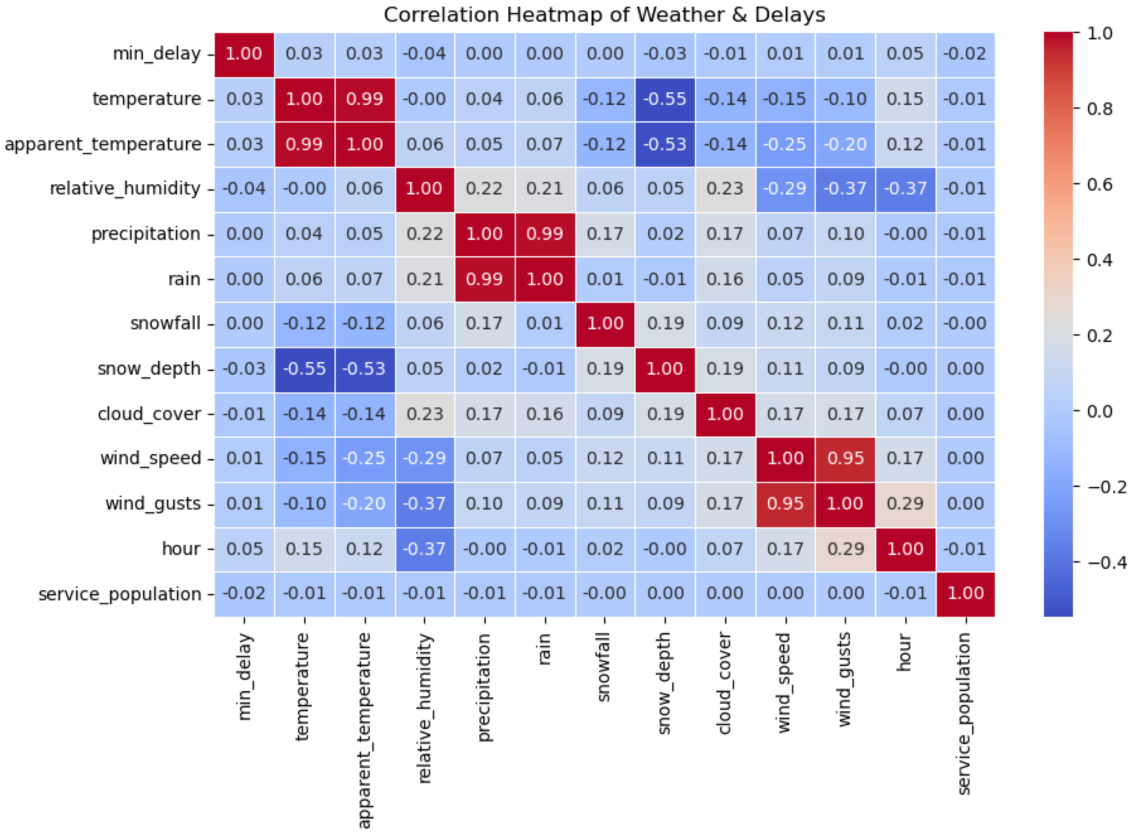

Figure 10: This plot shows a correlation matrix of weather conditions, service population, and delay duration.

The correlation heatmap, in Figure 10, reveals that weather conditions have minimal correlation with subway delay duration. Most weather variables show near-zero correlations with delay duration, while precipitation, rain, and snowfall show zero correlation, suggesting that weather may not be a strong predictor of delay length. Additionally, service population also has little correlation with delay duration, indicating that population density near stations does not significantly impact the length of delays. However, expected strong correlations are observed between related weather variables, such as temperature and apparent temperature, as well as precipitation and rain.

Summary

This study analyzed subway delays within the Toronto Transit Commission (TTC) system, examining their occurrence and potential influences from weather conditions and population density. Using a merged dataset of transit delay reports, hourly weather data, and demographic information, key patterns in delay frequency and duration were explored. The findings indicate that delays are most frequent during morning and afternoon rush hours, but their durations tend to be shorter compared to non-peak times. Passenger-related incidents were the most common causes of delays, and the Sheppard (SHP) Line experienced the longest delays. Directional patterns also emerged, with northbound trains facing more frequent disruptions. However, weather conditions and service population density showed little correlation with delay duration, suggesting other operational factors play a larger role.

To build on these findings, the next phase will involve developing machine learning models to predict subway delays. Both classification and regression models will be explored, including decision trees, logistic regression, gradient boosting models. Classification models will be used to determine whether a delay will occur based on historical conditions, while regression models will estimate the expected duration of a delay. By implementing predictive modeling, this study aims to create a data-driven framework for anticipating delays, helping transit planners optimize scheduling and commuters make informed travel decisions.

References

- Open-meteo documentation: https://open-meteo.com/en/docs/historical-weather-api
- OpenStreetMap API documentation: https://wiki.openstreetmap.org/wiki/API_v0.6
- Toronto Neighbourhood Profiles: https://open.toronto.ca/dataset/neighbourhood-profiles/
- Toronto Neighbourhoods: https://open.toronto.ca/dataset/neighbourhoods/
- TTC Subway and Streetcar Map: https://cdn.ttc.ca/-/media/Project/TTC/DevProto/Images/Home/Routes-and-Schedules/Landing-page-pdfs/TTC_SubwayStreetcarMap_2021-11.pdf?rev=909317034177450b8b09ba5b247e24bf
- TTC Subway Delay Data: https://open.toronto.ca/dataset/ttc-subway-delay-data/