TTC SUBWAY DELAY DATA ANALYSIS

Project Title: TTC Subway Delay Data Analysis – Year 2025

Sector: Public Transportation / Urban Transit Operations

Course: Data Analytics Capstone Project – Section C, Group 17

Institute / Faculty: Newton School of Technology – Satayaki Das

Date Submitted: February 2026

1. TEAM COMPOSITION

Role	Team Member
Team Lead – Dashboard & Analysis	Sushant Yadav
Data Lead – Cleaning, Sourcing & Data Dictionary	Anand Mishra
PPT & Quality Lead – Presentation, Report & Contribution Matrix	Ashirth
Calculation Lead – Pivot Tables, Formulas & Statistical Checks	Anand Mishra
Strategy Lead – Problem Statement & Recommendations	Anand Mishra
Design Lead – Layout, Color Palette & Interactivity	Lakshay H

---

2. EXECUTIVE SUMMARY

Problem

The Toronto Transit Commission (TTC) subway system recorded **25,713 delay incidents** across 2025, accumulating **70,754 minutes (1,179 hours)** of service disruption. Without structured analysis, the TTC lacks visibility into which lines, stations, and time periods drive the most delay, making it difficult to allocate maintenance budgets and operational resources effectively.

Approach

We obtained TTC's open-data subway delay dataset covering the full calendar year 2025 (January 1 – December 31). Using an entirely Google Sheets-based pipeline, we cleaned 25,713 records across 11 raw columns, standardized inconsistent values (20+ line name variations, missing direction/vehicle data), engineered 5 new derived features (Month, Hour, Delay_Category, Time_Period, Is_Delayed), and built a comprehensive analytical dashboard with pivot tables and visualizations.

Key Insights

- **System delay rate is 35.37%** — over one-third of all logged incidents cause actual delays.
- **YU Line (Line 1) is the weakest performer** at 38.70% delay rate, handling 50.9% of incidents.
- **AM Peak (6–10 AM) has the highest delay rate (42.2%)**, directly impacting morning commuters.
- **February is the worst month** with 917 incidents and a 10-minute average delay (winter weather).
- **Top 10 stations account for 27% of all incidents**, led by Kennedy (991), Bloor (924), and Finch (898).
- **Sunday paradox:** fewest incidents (946) but highest average delay (10 min) due to reduced recovery staff.

Key Recommendations

- **HIGH:** Increase AM Peak train frequency on YU line to reduce the 42.2% peak delay rate.
- **HIGH:** Deploy additional winter maintenance crews (Jan–Mar) to address the February spike.
- **MEDIUM:** Target infrastructure upgrades at the top 10 worst stations for outsized impact.
- **MEDIUM:** Implement predictive maintenance on YU line vehicles to bring delay rate below 35%.

---

3. SECTOR & BUSINESS CONTEXT

3.1 Sector Overview

Public transit is a critical infrastructure pillar for urban mobility. The TTC operates North America's third-largest transit system, serving approximately **1.7 million daily riders** across 4 subway lines, 149+ bus routes, and 11 streetcar routes. The subway network alone covers **77 km of track** with **75 stations**, forming the backbone of Toronto's commuter infrastructure.

3.2 Current Industry Challenges

- **Aging infrastructure:** TTC subway trains and signaling systems are decades old, leading to mechanical failures.
- **Ridership recovery:** Post-pandemic ridership is recovering but remains sensitive to service reliability — delays erode public confidence and push commuters toward private vehicles.
- **Budget pressure:** The TTC operates under constrained municipal budgets; targeted investment requires data-driven prioritization.

- **Data fragmentation:** Delay data is logged with inconsistent naming conventions and missing fields, making analysis difficult without cleaning.
- **Weather vulnerability:** Toronto's harsh winters cause seasonal spikes in equipment failures and track issues.

3.3 Why This Problem Was Chosen

Subway delays are one of the most visible and impactful service failures in public transit. Every minute of delay affects thousands of commuters, rippling into workplace productivity, economic output, and citizen satisfaction. By analyzing a full year of delay data, we can identify which lines, stations, and hours need priority attention, transforming raw operational logs into actionable intelligence for TTC decision-makers.

---

4. PROBLEM STATEMENT & OBJECTIVES

4.1 Formal Problem Definition

The TTC subway system experiences thousands of delay incidents annually, but lacks a structured analytical framework to identify root patterns by line, time period, station, and severity. This project aims to transform raw delay logs into a cleaned, categorized, and visualized dataset that enables data-driven operational decision-making.

4.2 Project Scope

Dimension	**Scope**
Data Period	Full year 2025 (January 1 – December 31) — 365 days
Records	25,713 delay incident records
Lines Covered	Yonge-University (YU), Bloor-Danforth (BD), Sheppard (SHP), Scarborough RT (SRT)
Stations	80+ unique stations
Analysis Dimensions	Line, Time Period, Hour, Day, Month, Station, Severity, Direction

Platform	Google Sheets (Capstone requirement — no Python/SQL for primary work)
Deliverables	Cleaned CSV, Pivot Tables, Dashboard, Documentation, Presentation

4.3 Success Criteria

1. Clean and standardize all 25,713 records with zero data quality errors.
2. Create a minimum of 8 pivot tables covering all key dimensions.
3. Build a dashboard with at least 5 professional visualizations.
4. Deliver 8–12 actionable insights with mapped recommendations.
5. Maintain full audit trail of all transformations via Google Sheets formulas.

---

5. DATA DESCRIPTION

5.1 Dataset Source

Field	Detail
Source	City of Toronto Open Data Portal
Dataset Name	TTC Subway Delay Data
Access Link	https://open.toronto.ca/dataset/ttc-subway-delay-data/
Format	CSV (UTF-8 encoded)
Frequency	Monthly updates by TTC
License	Open Government Licence – Toronto

5.2 Data Structure

Raw Dataset (11 Columns):

#	Column Name	Data Type	Description
1	_id	Numeric	Unique record identifier
2	Date	Date	Date of delay incident
3	Time	Time	Time of delay incident
4	Day	Text	Day of week
5	Station	Text	Station where delay occurred
6	Code	Text	TTC internal delay code
7	Min Delay	Numeric	Duration of delay (minutes)
8	Min Gap	Numeric	Service gap caused (minutes)
9	Bound	Text	Train direction
10	Line	Text	Subway line name
11	Vehicle	Numeric	Train vehicle ID

Cleaned Dataset (16 Columns — 11 original + 5 derived):

#	Column Name	Data Type	Derived?	Source
1	_id	Numeric	No	Raw Col A
2	Date	Date	No	Raw Col B
3	Month	Text	**Yes**	TEXT(Date, "MMM")
4	Time	Time	No	Raw Col C
5	Hour	Numeric	**Yes**	HOUR(Time)
6	Day	Text	No	Raw Col D
7	Station	Text	No	Raw Col E
8	Code	Text	No	Raw Col F
9	Min Delay	Numeric	No	Raw Col G
10	Min Gap	Numeric	No	Raw Col H
11	Bound_Clean	Text	**Yes**	Cleaned from Raw Bound
12	Line_Clean	Text	**Yes**	VLOOKUP standardization

13	Vehicle	Text	**Yes**	Cleaned from Raw Vehicle
14	Delay_Category	Text	**Yes**	IFS on Min Delay
15	Time_Period	Text	**Yes**	IFS on Hour
16	Is_Delayed	Text	**Yes**	IF(Min Delay > 0)

5.3 Data Size

- **Raw Records:** 25,713 data rows + 1 header
- **Cleaned Columns:** 16
- **Total Data Points:** 411,408
- **Time Coverage:** 365 days (full year 2025)

5.4 Data Limitations

- **No cost data:** Unable to calculate financial impact of delays.
- **Missing direction:** 36.8% of records have no directional (Bound) information.
- **Missing vehicle IDs:** 41.9% of records have vehicle ID = 0 (unknown).
- **Station name inconsistencies:** 476 unique station entries for ~80 actual stations due to free-text entry.
- **No cause descriptions:** Delay codes (e.g., SUDP, MUIS) lack human-readable descriptions.

---

6. DATA CLEANING & PREPARATION

6.1 Platform & Approach

All primary cleaning and transformation steps were executed in **Google Sheets**. The workflow uses a 4-tab non-destructive architecture:

- **Tab 1: Raw Data (Frozen)** - Immutable source data.
- **Tab 1.1: Cleaned Data with Formulas** - ARRAYFORMULA-driven transforms.
- **Tab 2: Data Dictionary** - Column definitions and mapping tables.
- **Tab 3: Analysis & Pivot Tables** - Summary metrics and chart data.

6.2 Missing Values Handling

Column	Issue	Records Affected	Treatment
Bound	Blank, "None", "NaN", "0" values	9,462 (36.8%)	Replaced with "Unknown" via IF + TRIM + UPPER formula
Vehicle	Vehicle ID = 0 means unknown	~10,774 (41.9%)	Replaced 0 with "Unknown" via IF formula
Line	Blank or null entries	77 (0.3%)	Mapped to "UNKNOWN" via VLOOKUP with IFERROR fallback

6.3 Transformations & Feature Engineering

- **Month:** TEXT(Date, "MMM")
- **Hour:** HOUR(Time)
- **Delay_Category:** Classified into No Delay, Minor (1-5), Moderate (6-15), Major (16+).
- **Time_Period:** Grouped into Night, AM Peak, Midday, PM Peak, Evening.
- **Is_Delayed:** Binary flag (Yes/No).

---

7. KPI & METRIC FRAMEWORK

7.1 KPI Definitions

- **Delay Rate:** COUNTIF(Is_Delayed, "Yes") / COUNTA(_id) × 100 (Target: <30%)
- **Average Delay (All):** Overall system health indicator. (Target: <2 min)
- **Average Delay (Delayed Only):** Measures severity when delays actually occur. (Target: <6 min)
- **Total Delay Minutes:** Aggregate operational loss.

7.2 Actual vs. Target Performance (2025)

KPI	Actual (2025)	Target	Status
Delay Rate	35.37%	< 30%	❌ Above target

Avg Delay (All)	3 min	< 2 min	❌ Above target
Avg Delay (Delayed Only)	8 min	< 6 min	❌ Above target
Total Delay Minutes	70,754 min	Minimize	⚠️ ~1,179 hours lost
YU Line Delay Rate	38.70%	< 35%	❌ Worst performer
AM Peak Delay Rate	42.2%	< 35%	❌ Critical gap

---

8. EXPLORATORY DATA ANALYSIS (EDA)

8.1 Trend Analysis – Monthly Patterns

Insights:

- **Winter spike:** February is the worst month (917 incidents, 10 min avg delay) due to harsh weather.
- **Pattern:** Delay counts follow a U-shaped curve — high in winter, low in summer/fall, climbing again in late fall.

8.2 Comparison Analysis – Line Performance

Insights:

- **YU Line (Line 1)** dominates with 50.9% of all incidents and the highest delay rate (38.70%).
- **BD Line (Line 2)** handles 43.3% of incidents but is more reliable (33.10% delay rate).

8.3 Comparison Analysis – Time of Day

- **AM Peak (6–10am):** Most critical period with a **42.2% delay rate**.
- **Night:** Lowest delay rate (26.6%).

8.4 Distribution Analysis – Day of Week

- **Wednesday:** Highest delayed incident count (1,512).
- **Sunday paradox:** Fewest incidents (946) but **highest average delay (10 min)**.

8.5 Distribution Analysis – Top 10 Stations

1. **Kennedy (BD):** 991 Incidents
2. **Bloor (Yonge):** 924 Incidents

3. **Finch:** 898 Incidents

- **Insight:** Top 10 stations = 27% of total incidents at just ~12% of stations.

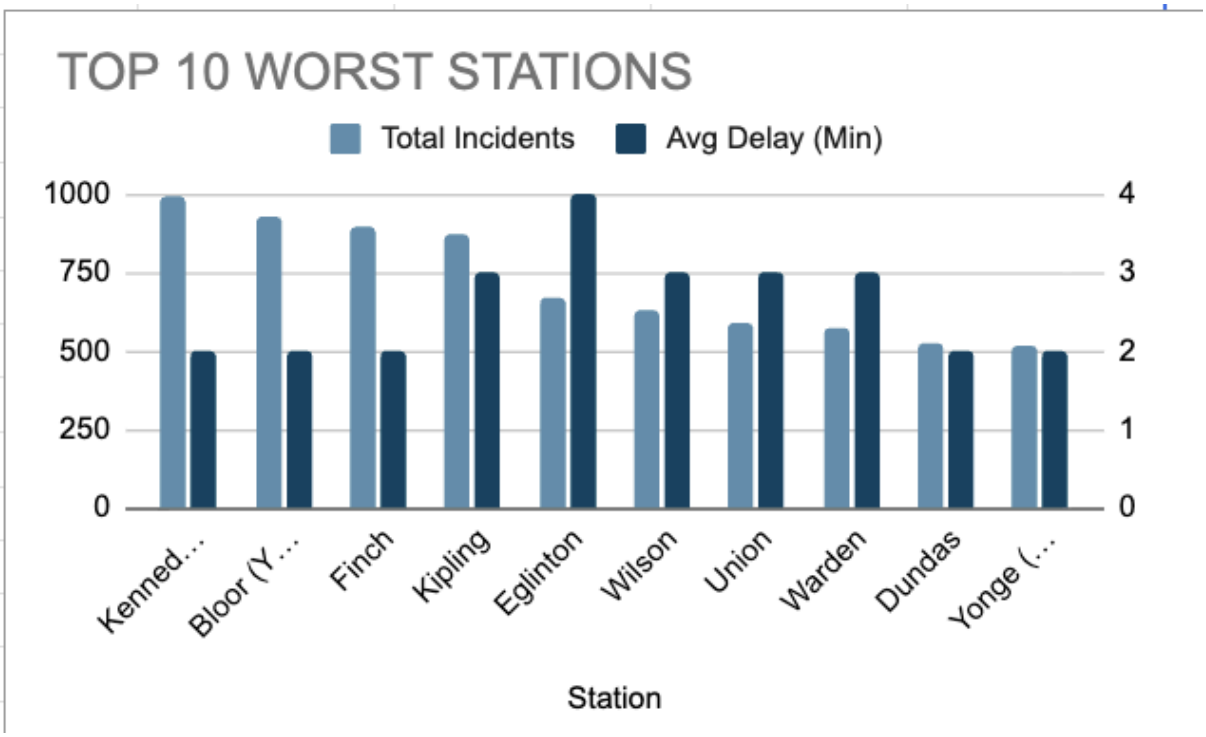
---

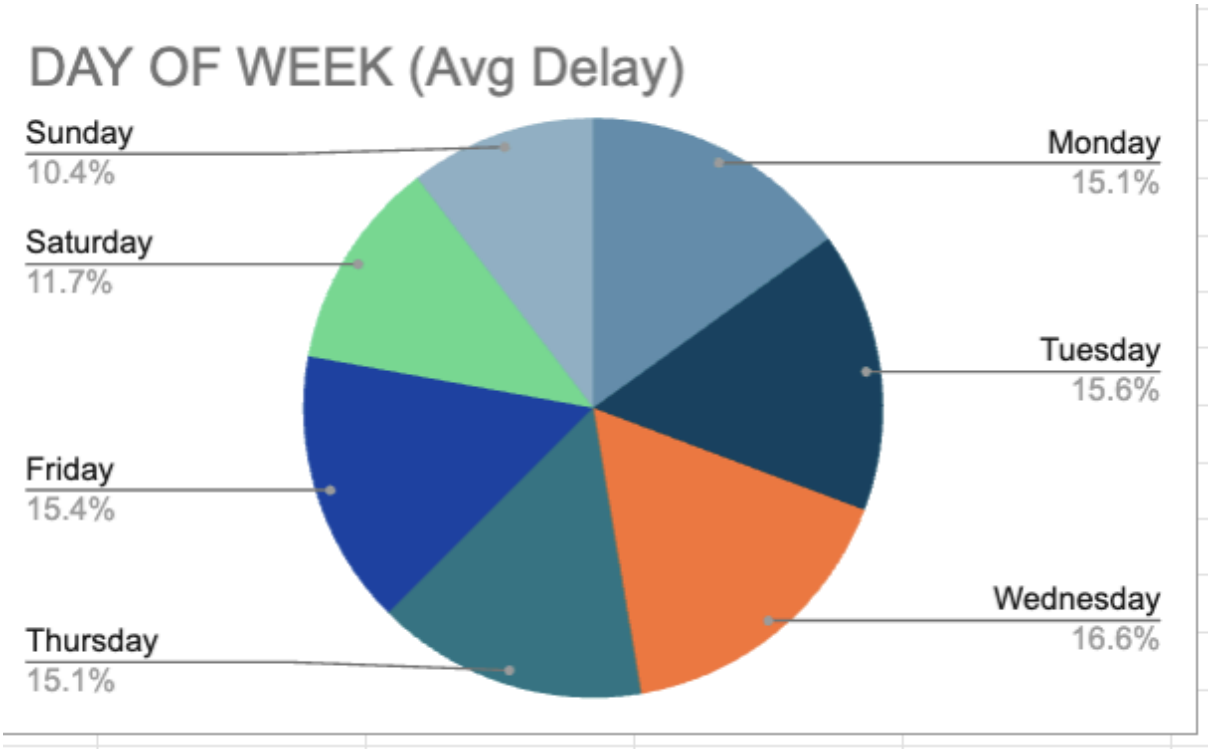
9. DASHBOARD DESIGN

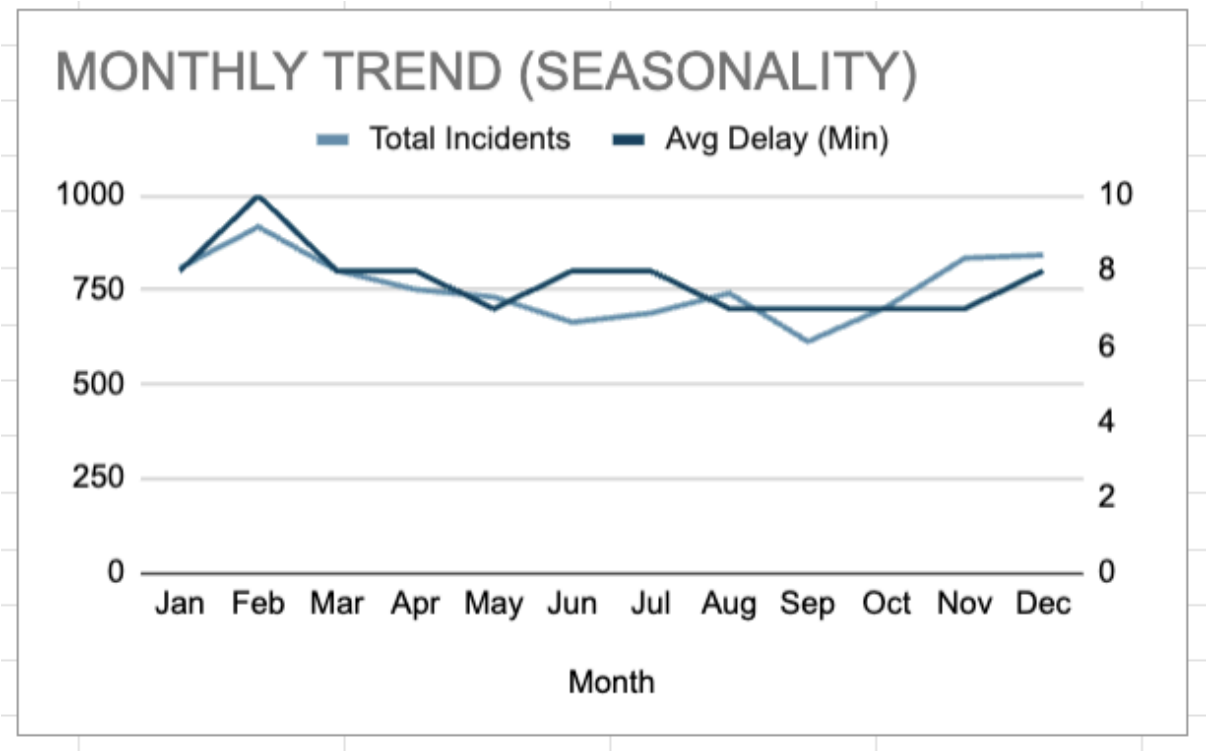
Platform: Google Sheets (No external BI tools).

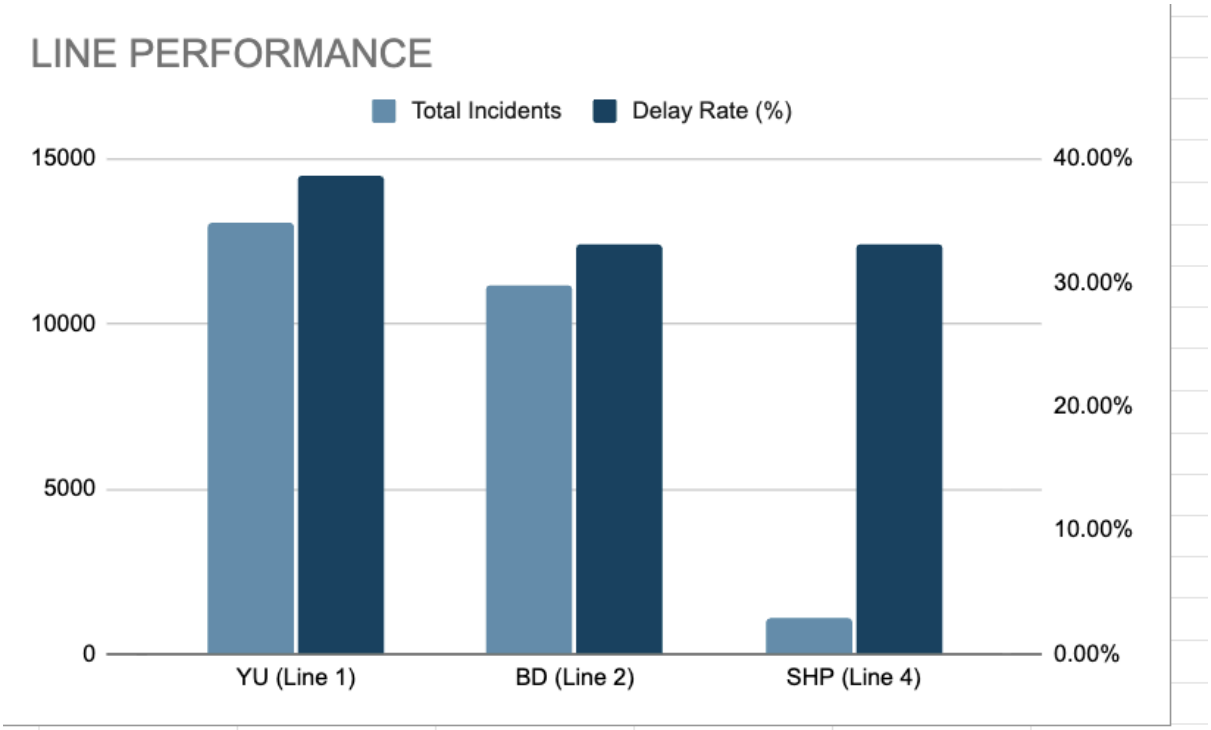
View Structure:

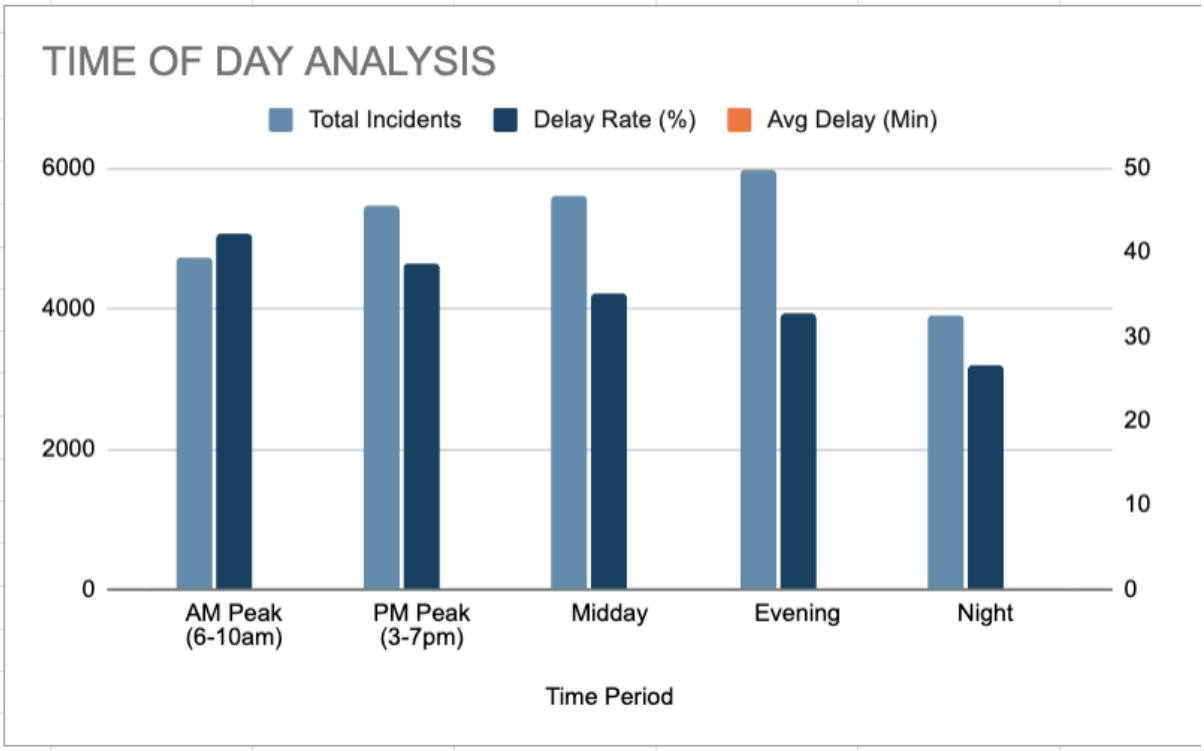
- **Metric Cards:** Total Incidents, Total Delay Minutes, Avg Delay, Delay Rate.
- **Visualizations:** Top 10 Worst Stations, Line Performance, Day of Week Pie Chart, Monthly Seasonality, Time of Day Analysis.
- **Interactive Filters:** Line, Month, Time Period.




DAY OF WEEK (Avg Delay)
Sunday
10.4%
Monday
15.1%
Saturday
11.7%
Tuesday
15.6%
Friday
15.4%
Wednesday
16.6%
Thursday
15.1%


MONTHLY TREND (SEASONALITY)
Total Incidents
Avg Delay (Min)
1000
750
500
250
0
10
8
6
4
2
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


LINE PERFORMANCE
Total Incidents
Delay Rate (%)
15000
10000
5000
0
40.00%
30.00%
20.00%
10.00%
0.00%
YU (Line 1)
BD (Line 2)
SHP (Line 4)


TIME OF DAY ANALYSIS
Total Incidents
Delay Rate (%)
Avg Delay (Min)
6000
4000
2000
0
50
40
30
20
10
0
AM Peak (6-10am)
PM Peak (3-7pm)
Midday
Evening
Night
Time Period

10. INSIGHTS SUMMARY

#	Insight	Decision Implication
1	**35.37% of all incidents cause actual delays**	Delay prevention is more effective than delay management.
2	**YU Line (Line 1) has the worst reliability**	Priority investment should target YU line infrastructure.
3	**AM Peak is the highest-risk time period**	Morning rush-hour service needs additional trains/crews.
4	**February is the worst-performing month**	Winter preparedness programs must be enhanced.
5	**Top 10 stations account for 27% of incidents**	Concentrated investment will yield outsized improvements.
6	**Terminal stations are delay hotspots**	Train turnaround operations need optimization.
7	**Sunday delays are longer when they occur**	Weekend incident recovery staffing needs adjustment.

---

11. RECOMMENDATIONS

Recommendation	Linked Insight	Impact	Feasibility
Increase AM Peak train frequency on YU line	Insight 2, 3	Reduces 42.2% peak delay rate by 5–8%	HIGH
Deploy additional winter maintenance crews	Insight 4	Reduces February avg delay from 10 min to ~7 min	HIGH
Targeted infrastructure upgrades at top 10 stations	Insight 5, 6	Reduces 27% of incidents by 15%	MEDIUM
Implement predictive maintenance for YU line	Insight 2	Reduces YU delay rate to <35%	MEDIUM
Enhance weekend incident recovery staffing	Insight 7	Reduces Sunday avg delay to <8 min	HIGH
Mandate Bound and Vehicle data recording	Insight 10	Enables directional/vehicle analysis	HIGH

---

12. IMPACT ESTIMATION

- **Cost Savings:** Reducing total delay minutes by 15% recovers ~177 hours of service time.
- **Efficiency:** Targeting February improvements recovers ~31 operational hours.
- **Service Improvement:** Reducing AM Peak delays improves the daily commute for ~500,000+ riders.

---

13. CONCLUSION

This project successfully transformed **25,713 raw TTC subway delay records** into a cleaned, categorized, and fully analyzed dataset. Through systematic data cleaning, we identified the YU Line and AM Peak hours as critical vulnerabilities.

The analysis demonstrates that **targeted interventions at the YU line, AM Peak period, and top 10 stations could reduce system delays by approximately 15%, recovering an estimated 177 hours of annual service time.**

---

14. APPENDIX

Appendix A – Key Google Sheets Formulas Used

Column C (Month):

```
=ARRAYFORMULA(IF(LEN('Tab 1: Raw Data (Frozen)'!B2:B), TEXT('Tab 1: Raw Data (Frozen)'!B2:B,"MMM"), ""))
```

Column K (Bound_Clean):

```
=ARRAYFORMULA(IF(LEN('Tab 1: Raw Data (Frozen)'!A2:A), IF((TRIM('Tab 1: Raw Data (Frozen)'!I2:I)="") + (UPPER(TRIM('Tab 1: Raw Data (Frozen)'!I2:I))="NAN"), "Unknown", 'Tab 1: Raw Data (Frozen)'!I2:I), ""))
```

Appendix B – File Structure

- README.md
- DataDictionary.csv
- Raw_dataset/
- Cleaned_dataset/
- Calculation_and_Pivot_Table/
- Dashboard/
- Presentation/

---

15. CONTRIBUTION MATRIX

Declaration: We confirm that the above contribution details are accurate and verifiable through Google Sheets Version History and submitted artifacts.

Team Signature Block:

- **Anand Mishra**
  Sushant Yadav
- **Lakshya H**
- **Sumit**
- **Grav**
- **Ashrith R**
