

Data Cleaning Report

Project Info

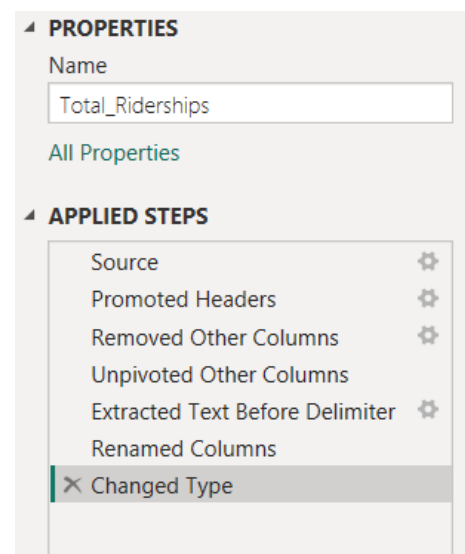
- **Project Title:** Analyzing and Forecasting Public Transportation Ridership Trends in New York (MTA)
 - **Team Members:**
 - Gaser Ahmed Mohammed Saad
 - Mohamed Abdelfattah Ibrahim Elwa
 - Abdulrahman Mohamed Safaa Ismail
 - Basant Abdelhalim Mohamed Safan
 - Doha Medhat Elsayed Salem
 - **Team Leader:** Gaser Ahmed Mohammed Saad
 - **Supervisor:** Eng. Sherihan Ali
 - **Date Submitted:** 03/10/2025
 - **Dataset Used:** MTA Daily Ridership Dataset (2020–2024)
-

Checking the Data (Assumptions & Data Quality)

| What You Checked | What You Found | What You Did |
|-----------------------|--|--|
| Missing Values | 0 | Already Cleaned |
| Duplicates | 0 | Already Cleaned |
| Outliers | 0 | Already Cleaned |
| Column Types | "Date" was text, not a date "Total Ridership" was text not a whole number %Pre-Pandemic was text, not a decimal number | Changed column type to "Date" in Power Query Changed column type to "Whole number" in Power Query Changed column type to "Whole number" in Power Query |

Steps You Did in Power Query

1. **Duplicated the original dataset** into two separate queries.
 - Query 1: kept only **Total Ridership columns** (removed % columns).
 - Query 2: kept only **% of Comparable Pre-Pandemic columns** (removed Totals).
2. **Promoted headers** to use the first row as column names.
3. **Removed unnecessary columns** that were not required for analysis.
4. **Unpivoted columns** to convert the wide format into a normalized long format (easier for modeling and visualization).
5. **Extracted text before delimiter** to simplify column names and make them consistent (e.g., "Subways: Total Estimated Ridership" → "Subways").
6. **Renamed columns** to short, clear, and consistent names.
7. **Changed data types:**
 - Date → Date
 - Ridership (Totals) → Whole Number
 - Percentage columns → Decimal Number



8. **Merged the two queries** (Totals + Percentages) on the **Date** field to create one clean, structured dataset containing both actual ridership and pre-pandemic comparison percentages.

9. **Created a separate Dimension Table for Transport Types:**

- Duplicated the merged table.
- Kept only the column containing transport type names.
- Removed duplicates to get a unique list of transport modes.
- Final result = **7 unique transport types** (Subways, Buses, LIRR, Metro-North, Access-A-Ride, Bridges & Tunnels, Staten Island Railway).

The screenshot displays the Microsoft Power BI Desktop interface. The main view shows a data table with the following columns: Date, Transport_Name, Total_Riderships, and %Prepandemic. The table contains 999+ rows of data, with the first 28 rows visible. The data is organized by date, showing ridership trends over time for various transport types.

| Date | Transport_Name | Total_Riderships | %Prepandemic |
|----------|-----------------------|------------------|--------------|
| 3/1/2020 | Subways | 2212965 | 97 |
| 3/1/2020 | Buses | 984908 | 99 |
| 3/1/2020 | LIRR | 86790 | 100 |
| 3/1/2020 | Metro-North | 55825 | 59 |
| 3/1/2020 | Access-A-Ride | 19922 | 113 |
| 3/1/2020 | Bridges and Tunnels | 786960 | 98 |
| 3/1/2020 | Staten Island Railway | 1636 | 52 |
| 3/2/2020 | Subways | 5329915 | 96 |
| 3/2/2020 | Buses | 2209066 | 99 |
| 3/2/2020 | LIRR | 321569 | 103 |
| 3/2/2020 | Metro-North | 180701 | 66 |
| 3/2/2020 | Access-A-Ride | 30338 | 102 |
| 3/2/2020 | Bridges and Tunnels | 874619 | 95 |
| 3/2/2020 | Staten Island Railway | 17140 | 107 |
| 3/3/2020 | Subways | 5481103 | 98 |
| 3/3/2020 | Buses | 2228608 | 99 |
| 3/3/2020 | LIRR | 319727 | 102 |
| 3/3/2020 | Metro-North | 190648 | 69 |
| 3/3/2020 | Access-A-Ride | 32767 | 110 |
| 3/3/2020 | Bridges and Tunnels | 882175 | 96 |
| 3/3/2020 | Staten Island Railway | 17453 | 109 |
| 3/4/2020 | Subways | 5498809 | 99 |
| 3/4/2020 | Buses | 2177165 | 97 |
| 3/4/2020 | LIRR | 311662 | 99 |
| 3/4/2020 | Metro-North | 192689 | 70 |
| 3/4/2020 | Access-A-Ride | 34297 | 115 |
| 3/4/2020 | Bridges and Tunnels | 905558 | 98 |
| 3/4/2020 | Staten Island Railway | 17136 | 107 |

The interface also shows a ribbon with various data manipulation tools, a right-hand pane for query settings, and a bottom status bar indicating 4 columns and 999+ rows.

Final Clean Dataset (Before & After)

Before Cleaning

1706 Rows
15 columns

After Cleaning

11942 (Unpivot)
4 Columns (Unpivot)
None left

Created new table from Date: Year, Month# , Month Name , Quarter, Week#, Day Name, Date Type, Season, Covid Period.

Before

The screenshot displays the Power Query Editor interface. The main area shows a table with 15 columns and 1706 rows. The columns are: Date, Subways: Total Estimated Ridership, Subways: % of Comparable Pre-Pandemic Day, Buses: Total Estimated Ridership, and Buses: % of Comparable Pre-Pandemic Day. The data is organized into rows, with the first row showing data for 3/1/2020 and the last row for 3/27/2020. The right-hand pane shows the 'Query Settings' for 'MTA_Daily_Ridership', including the 'Name' and 'Applied Steps' (Source, Promoted Headers, Changed Type). The bottom status bar indicates '15 COLUMNS, 999+ ROWS' and 'Column profiling based on top 1000 rows'.

| Date | Subways: Total Estimated Ridership | Subways: % of Comparable Pre-Pandemic Day | Buses: Total Estimated Ridership | Buses: % of Comparable Pre-Pandemic Day |
|-----------|------------------------------------|---|----------------------------------|---|
| 3/1/2020 | 2212965 | 97 | 984908 | |
| 3/2/2020 | 3329915 | 96 | 2209066 | |
| 3/3/2020 | 5481103 | 98 | 2228608 | |
| 3/4/2020 | 5498809 | 99 | 2177165 | |
| 3/5/2020 | 5496453 | 99 | 2244515 | |
| 3/6/2020 | 5189447 | 93 | 2066743 | |
| 3/7/2020 | 2814637 | 92 | 1249085 | |
| 3/8/2020 | 2120656 | 93 | 957163 | |
| 3/9/2020 | 4973513 | 89 | 2124770 | |
| 3/10/2020 | 4867818 | 87 | 2111989 | |
| 3/11/2020 | 4697122 | 84 | 2112967 | |
| 3/12/2020 | 4149505 | 75 | 1938424 | |
| 3/13/2020 | 3484996 | 63 | 1715737 | |
| 3/14/2020 | 1670665 | 54 | 593287 | |
| 3/15/2020 | 1157711 | 51 | 711555 | |
| 3/16/2020 | 2178555 | 39 | 1237309 | |
| 3/17/2020 | 1788786 | 32 | 1094949 | |
| 3/18/2020 | 1625280 | 29 | 1059502 | |
| 3/19/2020 | 1422112 | 26 | 933602 | |
| 3/20/2020 | 1309125 | 24 | 868602 | |
| 3/21/2020 | 619618 | 20 | 411491 | |
| 3/22/2020 | 408723 | 18 | 73517 | |
| 3/23/2020 | 709499 | 13 | 59321 | |
| 3/24/2020 | 741587 | 13 | 60334 | |
| 3/25/2020 | 690032 | 12 | 51769 | |
| 3/26/2020 | 680360 | 12 | 49970 | |
| 3/27/2020 | 656817 | 12 | 45514 | |

After

| Transport Name | Transport Type |
|-----------------------|----------------|
| Subways | Rail |
| Buses | Road |
| LIRR | Rail |
| Metro-North | Rail |
| Access-A-Ride | Paratransit |
| Bridges and Tunnels | Road |
| Staten Island Railway | Rail |

| Date | Transport_Name | Total_Riderships | %Prepandemic |
|---------------------------|----------------|------------------|--------------|
| Sunday, March 1, 2020 | Subways | 2212965 | 97 |
| Monday, March 2, 2020 | Subways | 5329915 | 96 |
| Tuesday, March 3, 2020 | Subways | 5481103 | 98 |
| Wednesday, March 4, 2020 | Subways | 5498809 | 99 |
| Thursday, March 5, 2020 | Subways | 5496453 | 99 |
| Friday, March 6, 2020 | Subways | 5189447 | 93 |
| Saturday, March 7, 2020 | Subways | 2814637 | 92 |
| Sunday, March 8, 2020 | Subways | 2120656 | 93 |
| Monday, March 9, 2020 | Subways | 4973513 | 89 |
| Tuesday, March 10, 2020 | Subways | 4867818 | 87 |
| Wednesday, March 11, 2020 | Subways | 4697122 | 84 |
| Thursday, March 12, 2020 | Subways | 4149505 | 75 |
| Friday, March 13, 2020 | Subways | 3484996 | 63 |
| Saturday, March 14, 2020 | Subways | 1670665 | 54 |
| Sunday, March 15, 2020 | Subways | 1157711 | 51 |
| Monday, March 16, 2020 | Subways | 2178555 | 39 |
| Tuesday, March 17, 2020 | Subways | 1788786 | 32 |
| Wednesday, March 18, 2020 | Subways | 1625280 | 29 |
| Thursday, March 19, 2020 | Subways | 1422112 | 26 |
| Friday, March 20, 2020 | Subways | 1309125 | 24 |
| Saturday, March 21, 2020 | Subways | 619618 | 20 |
| Sunday, March 22, 2020 | Subways | 408723 | 18 |
| Monday, March 23, 2020 | Subways | 709499 | 13 |
| Tuesday, March 24, 2020 | Subways | 741587 | 13 |
| Wednesday, March 25, 2020 | Subways | 690032 | 12 |
| Thursday, March 26, 2020 | Subways | 680360 | 12 |
| Friday, March 27, 2020 | Subways | 656817 | 12 |
| Saturday, March 28, 2020 | Subways | 332393 | 11 |

Table: fRiderships (11,942 rows)

| Date | Month# | Month Name | Week# | Qtr | Year | Day Name | Day Type | Season | CovidPeriod |
|-----------------------|--------|------------|-------|-----|------|----------|----------|--------|-----------------|
| 7/1/2020 12:00:00 AM | 7 | Jul | 27 | Q3 | 2020 | Wed | Weekday | Summer | During-pandemic |
| 7/2/2020 12:00:00 AM | 7 | Jul | 27 | Q3 | 2020 | Thu | Weekday | Summer | During-pandemic |
| 7/3/2020 12:00:00 AM | 7 | Jul | 27 | Q3 | 2020 | Fri | Weekday | Summer | During-pandemic |
| 7/4/2020 12:00:00 AM | 7 | Jul | 27 | Q3 | 2020 | Sat | Weekend | Summer | During-pandemic |
| 7/5/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Sun | Weekend | Summer | During-pandemic |
| 7/6/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Mon | Weekday | Summer | During-pandemic |
| 7/7/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Tue | Weekday | Summer | During-pandemic |
| 7/8/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Wed | Weekday | Summer | During-pandemic |
| 7/9/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Thu | Weekday | Summer | During-pandemic |
| 7/10/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Fri | Weekday | Summer | During-pandemic |
| 7/11/2020 12:00:00 AM | 7 | Jul | 28 | Q3 | 2020 | Sat | Weekend | Summer | During-pandemic |
| 7/12/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Sun | Weekend | Summer | During-pandemic |
| 7/13/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Mon | Weekday | Summer | During-pandemic |
| 7/14/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Tue | Weekday | Summer | During-pandemic |
| 7/15/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Wed | Weekday | Summer | During-pandemic |
| 7/16/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Thu | Weekday | Summer | During-pandemic |
| 7/17/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Fri | Weekday | Summer | During-pandemic |
| 7/18/2020 12:00:00 AM | 7 | Jul | 29 | Q3 | 2020 | Sat | Weekend | Summer | During-pandemic |
| 7/19/2020 12:00:00 AM | 7 | Jul | 30 | Q3 | 2020 | Sun | Weekend | Summer | During-pandemic |
| 7/20/2020 12:00:00 AM | 7 | Jul | 30 | Q3 | 2020 | Mon | Weekday | Summer | During-pandemic |
| 7/21/2020 12:00:00 AM | 7 | Jul | 30 | Q3 | 2020 | Tue | Weekday | Summer | During-pandemic |

dCalendar (1,827 rows)

Problems You Faced & How You Solved Them

- **Problem:** Column names were very long and not user-friendly (e.g., "Subways: Total Estimated Ridership").
Solution: Renamed columns to shorter, consistent names (e.g., "Subway_Ridership") for easier use in Power query.
- **Problem:** Dataset was in wide format (each transport mode in separate columns), which made analysis and modeling difficult.
Solution: Unpivoted the dataset into a long format to normalize the data and simplify relationships in the data model.
- **Problem:** Some columns (ridership totals) were stored as Text instead of Numeric.
Solution: Changed column types to Whole Number/Decimal to allow aggregation and visualization.