SDS292 Project 1 Report

Adriana Soldat

This project analyzes daily ridership data from the Metropolitan Transportation Authority in New York City from 2020 to 2025. The dataset contains multiple modes of transportation, including Subway, Bus, LIRR, Metro-North, Staten Island Railway, Access-A-Ride, and Bridges & Tunnels. Each row represents a single day's ridership, with variables such as:

- **Date**: The date of observation.
- **Transport Mode**: Type of transportation (Subway, Bus, etc.).
- **Total Ridership**: Estimated total ridership for that day.
- **Usage Ratio**: Ridership as a percentage of a comparable pre-pandemic day.
- **Pandemic Phase**: Categorized as Pre-Pandemic, Peak-Lockdown, or Recovery.

The dataset includes daily records for approximately six years, resulting in over 2,000 observations. Missing values were checked and appropriately handled. The primary question addressed in this analysis is: *How did the COVID-19 pandemic impact daily ridership across different MTA transport modes, and what trends can we observe over time?*

Analysis

Distribution of Daily Ridership (Plot 1)

The first plot examines the distribution of daily ridership for the Subway and Bus systems. The histogram reveals that subway ridership is generally higher and more variable than bus ridership, with some days reaching extreme peaks. In contrast, bus ridership is more stable with a narrower range of values. This difference likely reflects the subway's role as the backbone of NYC commuting, while buses serve more localized or less densely populated routes.

Observation: The overall shape of the distribution appears to be bimodal, with two distinct peaks. The histogram shows how daily ridership varies, highlighting the subway's higher passenger volume. The wide spread indicates fluctuations likely influenced by weekdays, holidays, and pandemic restrictions. For buses, the distribution is between 0 and 2 million ridership, while for the subway, it is between 0 and more than 4 million.

Average Ridership by Pandemic Phase (Plot 2)

The second plot presents the average daily ridership for all transport modes grouped by pandemic phase. The results show a sharp decline in ridership during the Peak-Lockdown phase (March–June 2020) across all modes. Subways and buses experienced the largest declines, while other modes like Access-A-Ride and Bridges & Tunnels saw smaller decreases. During the Recovery phase, ridership gradually increased but had not fully returned to pre-pandemic levels by 2025.

Observation: The plot clearly illustrates the pandemic's immediate and prolonged effects on transportation usage. Subway and bus systems dominate total ridership, and their drop during lockdown reflects behavioral changes such as remote work and reduced travel. We can notice that the transportation usage didn't completely return to the pre-pandemic statistics, specifically for bus and subway, which suggests that these transportation modes lost some users, potentially indicating in larger amount of personal cars or alternative transportation.

Ridership Distribution by Pandemic Phase (Plot 3)

The third visualization uses boxplots to show the distribution of daily ridership for Subway and Bus across pandemic phases. Outliers during the Peak-Lockdown phase suggest some days with atypically high travel, possibly related to essential workers or partial reopenings. The median ridership and interquartile range were lowest during lockdown, indicating a broad reduction in passenger numbers. Pre-Pandemic and Recovery phases show higher medians and wider variability.

Observation: This plot complements the previous bar chart by providing insights into data spread, extremes, and consistency. It highlights not only mean trends but also variability and unusual events during lockdowns. The mean value for ridership for the subway, for example, before the pandemic was around 4.8 mil, during the lockdown it dropped to 0.5 mil, and recovered to 2.5 mil

Monthly Trends in Ridership (Plot 4)

The fourth plot tracks monthly average ridership for each mode over 2020–2024 (excluding 2025 due to incomplete data). Seasonal fluctuations are apparent, with ridership generally increasing during fall and winter months. Each transport mode exhibits distinct trends: subways and buses follow similar seasonal patterns, while commuter rail lines (LIRR, Metro-North) show smaller peaks and dips. The overall trajectory shows a sharp drop in early 2020, slow recovery through 2021–2022, and gradual stabilization afterward.

Observation: This time series analysis emphasizes how external factors—pandemic phases, seasonal patterns, and weekdays—affect transportation usage. It also provides a baseline for forecasting future ridership under similar conditions. For example, we see that the ridership decreases over summer months, which might be connected to schools having breaks and increases in September when more students might be coming back.

Conclusion

The analysis of MTA ridership data from 2020 to 2025 demonstrates the profound impact of the COVID-19 pandemic on public transportation in New York City. Subway and bus ridership were most affected, showing dramatic decreases during the lockdown and gradual recovery afterward. Monthly trends reveal both seasonal variation and the long-term recovery trajectory across modes.

Pitfalls and Limitations:

- Data for 2025 may be incomplete, potentially skewing trends for the latest year.

- Ridership estimates for some modes may have measurement errors or missing values.

- Pandemic phases were broadly defined and may not capture regional or policy-specific variations in mobility restrictions.
- The three pandemic phases—Pre-Pandemic, Peak-Lockdown, and Recovery—cover different lengths of time. The Pre-Pandemic phase spans approximately two and a half months, the Peak-Lockdown phase lasts about three and a half months, and the Recovery phase extends over multiple years. This uneven duration affects the batch sizes included in each phase. As a result, statistical summaries like averages, medians, and boxplots can be biased: longer phases with more data points may dilute extreme values, while shorter phases may exaggerate variability. For example, the sharp decline observed during the Peak-Lockdown phase may appear more extreme relative to other phases simply because it is measured over fewer days, while the Recovery phase's trends might appear smoother due to its larger sample size. For future improvements, I should consider these differences when comparing phases, as they influence the interpretation of ridership patterns and variability.