Paving the Way for Smart Mobility Solutions

IMPROVING MTA FOR NEW YORKERS

The MTA Subway system is a vital public transportation that is critical for daily commutes, economic activity, and environmental sustainability. ***With the booming population****, it would be really difficult to have the same setup & strategy followed at present. It would be the perfect time to utilize the data to further optimize the existing strategy. It would be beneficial to ease the burden that MTA could face in the future.*

Adding New Express Lines

New York's subway system is a huge network that connects all boroughs in the New York City area. With massive amount of people and numerous stations, maintaing all of them would be a huge challenge. By looking at local stations (stations without an express line) that are mostly crowded (high volume of people tapping in), we can decide which stations from which line needs an extra express line.

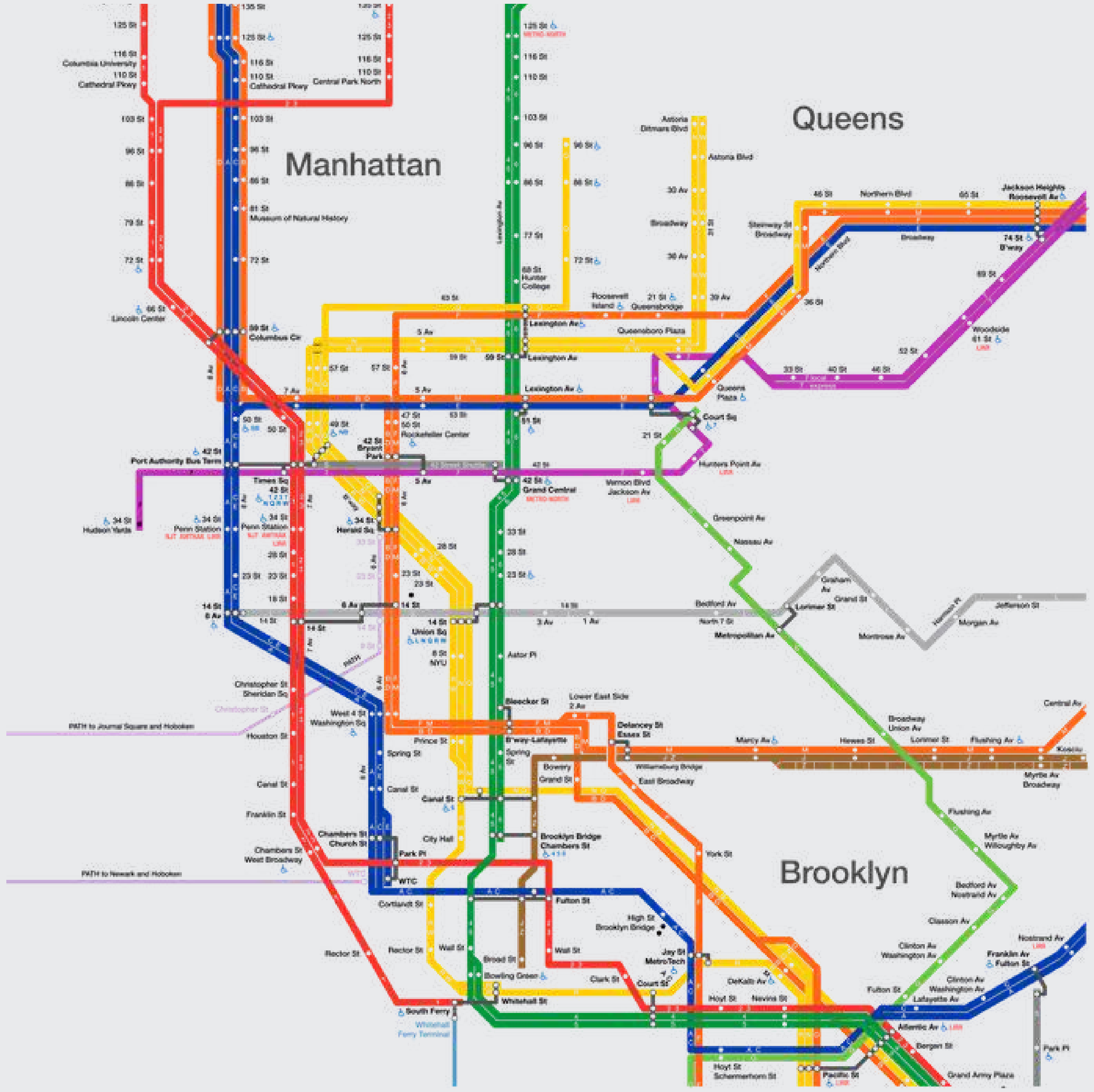
ADA Accessible Infrastructure

Only 23% of MTA subway stations are fully ADA accessible. While the MTA is committed to fund accessibility to 70 new stations, careful planning is needed so that it's funded for the right stations (Zoning for Transit Accessibility, 2019). Analyzing this data can determine which stations to focus on accessibility funding based on the number of people going into the station using the senior & disability fair class card.

Maintaining Station Security

Delays in MTA subway are unavoidable and the causes can vary such as track issue, subway car issue, signal issue, incidents, structural maintenance, and many more. By identifying delays in track lines, we can determine which lines have the worst performance in terms of delays and we can recommend those lines to have extra security to mitigate delays better.

DATA SOURCE AND PROCUREMENT



Data Source

Primary Data: The main dataset is the MTA Subway Hourly Ridership data, accessible via MTA Subway Hourly Ridership since February 2022. It provides detailed hourly ridership patterns across the subway system.

Secondary Data:

- The additional dataset relates to the on-time performance of trains, available through MTA's Train On-time Performance since 2020. This dataset includes performance metrics and adherence to schedules for subway services. Based on subway line.
- Another dataset used to track the cause of delays and low on-time performance are the major incidents on the subway tracks. The MTA's Subway Incidents dataset is a supplemental source for the On-time Performance.
- The 4th Data set is the MTA's ADA Station dataset. This includes whether or not the subway station has an accesibility service

Procurement

- ***Procurement and Integration:***
  - Extraction: Both datasets will be programmatically accessed through their respective APIs, ensuring real-time data updates and integration into the ETL pipeline. The APIs provide flexibility in data extraction, allowing for specific queries and reducing the load on the system.
  - Integration: The project will involve integrating these two datasets to provide a comprehensive view of subway ridership. While the hourly ridership dataset offers a broad overview, the turnstile data adds depth, enabling more detailed analyses such as station-specific congestion and the effectiveness of station layouts.
- ***Data Management Considerations:***
  - Data Quality and Consistency: Initial efforts will focus on assessing the quality and consistency of both datasets, ensuring that discrepancies, missing values, and anomalies are addressed to maintain the integrity of the analysis.
  - Data Privacy and Compliance: Although the datasets are publicly available, our team will adhere to data privacy standards and ethical guidelines, particularly in handling any data that could potentially be linked to individual behaviors or patterns.

TECHNOLOGIES USED

Socrata®

We are using Socrata for accessing the dataset via the API. This was provided by the NYS Government website with a step-by-step guideline on how to access the data.

mongoDB

This helped us extract the data from Socrata and store it from the API. We used the hourly ridership data and the station data as we extracted a total of 3.36 GB.

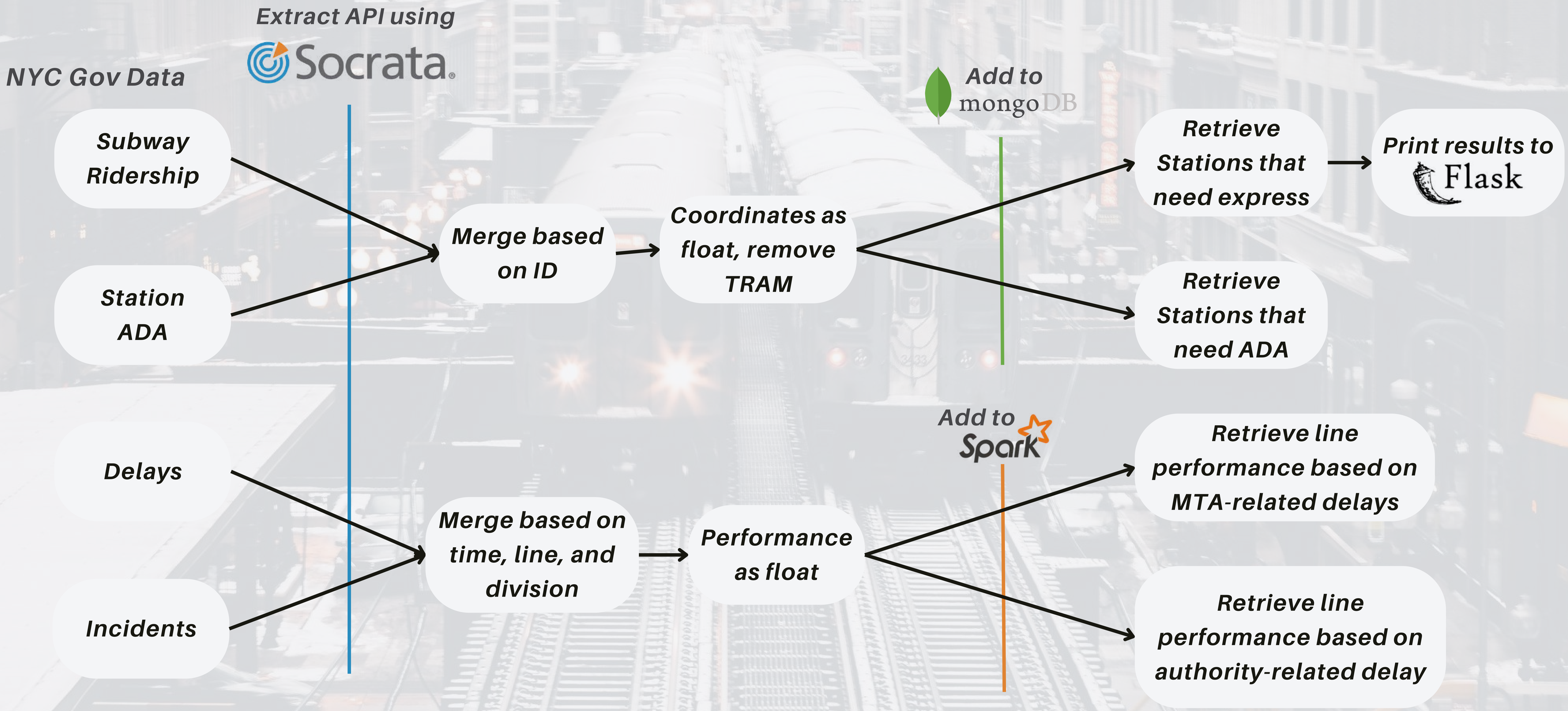
Spark

Spark was used to merge the incidents and the delays dataset based on the exact combination of Month, Division, Line, and Day-Type . We were able to extract the on-time-performance by station based on the incidental category.

Flask

We used Flask to generate an interactive output so the user can select the combination of borough and train line

PROJECT WORKFLOW



NYC Gov Data

Extract API using Socrata

Subway Ridership

Station ADA

Merge based on ID

Coordinates as float, remove TRAM

Add to mongoDB

Retrieve Stations that need express

Print results to Flask

Retrieve Stations that need ADA

Delays

Incidents

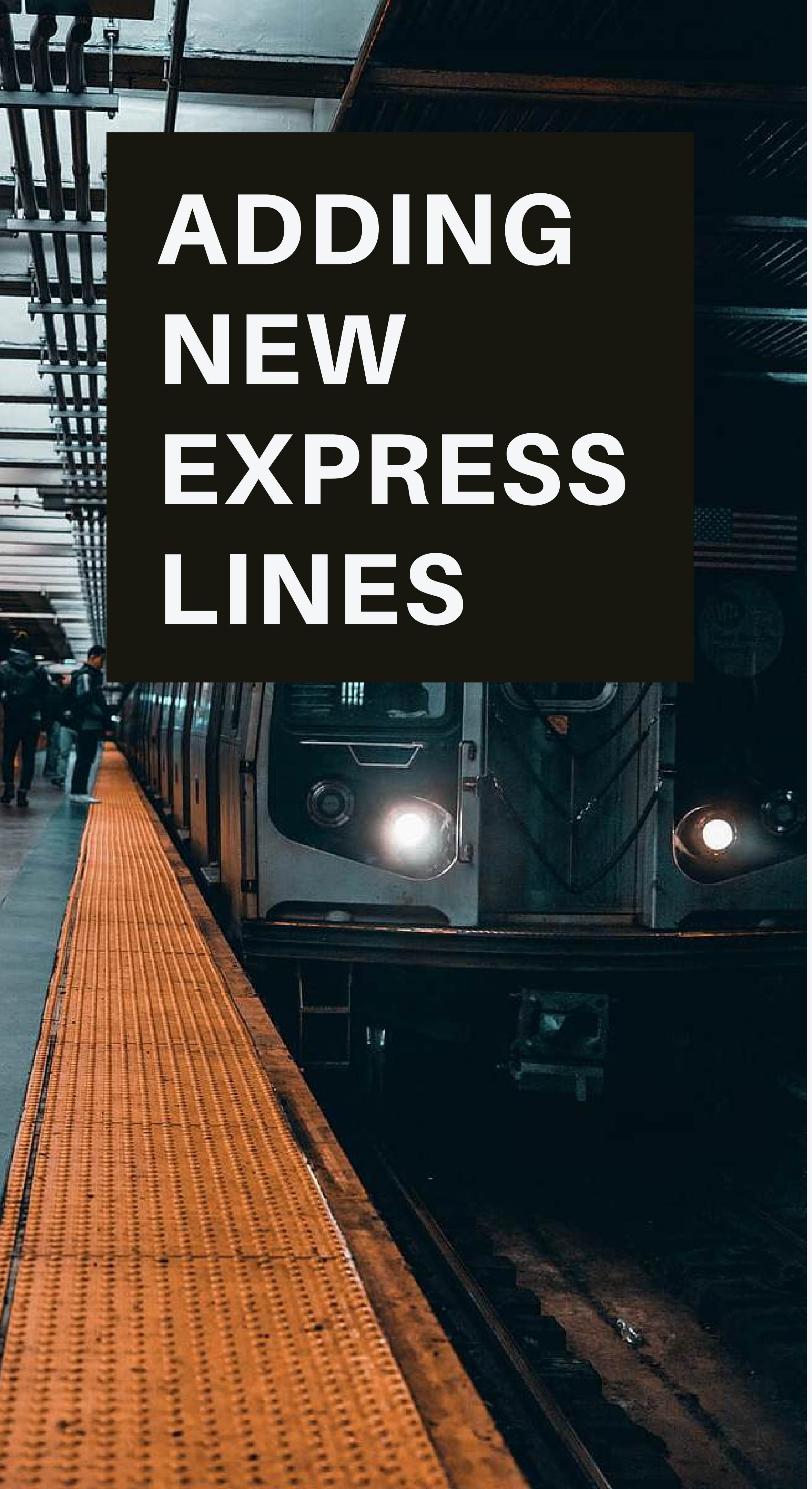
Merge based on time, line, and division

Performance as float

Add to Spark

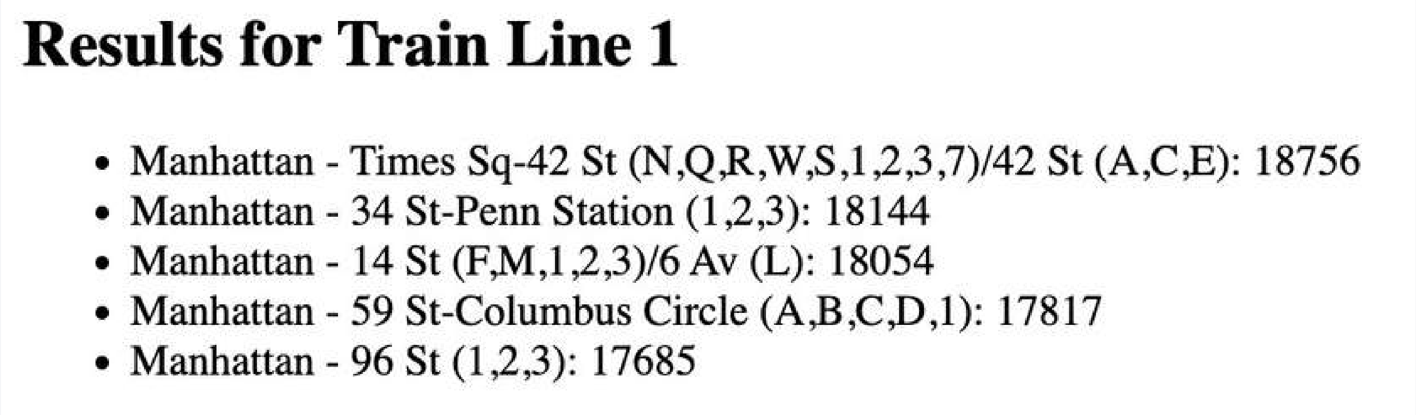
Retrieve line performance based on MTA-related delays

Retrieve line performance based on authority-related delay



ADDING NEW EXPRESS LINES

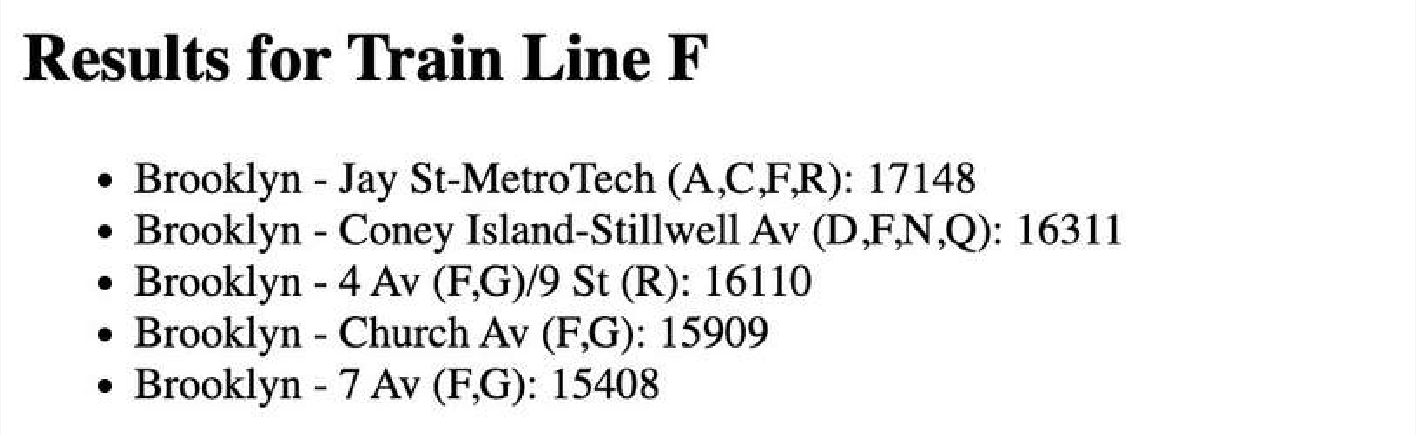
1 Line, Manhattan



Results for Train Line 1

- Manhattan - Times Sq-42 St (N,Q,R,W,S,1,2,3,7)/42 St (A,C,E): 18756
- Manhattan - 34 St-Penn Station (1,2,3): 18144
- Manhattan - 14 St (F,M,1,2,3)/6 Av (L): 18054
- Manhattan - 59 St-Columbus Circle (A,B,C,D,1): 17817
- Manhattan - 96 St (1,2,3): 17685

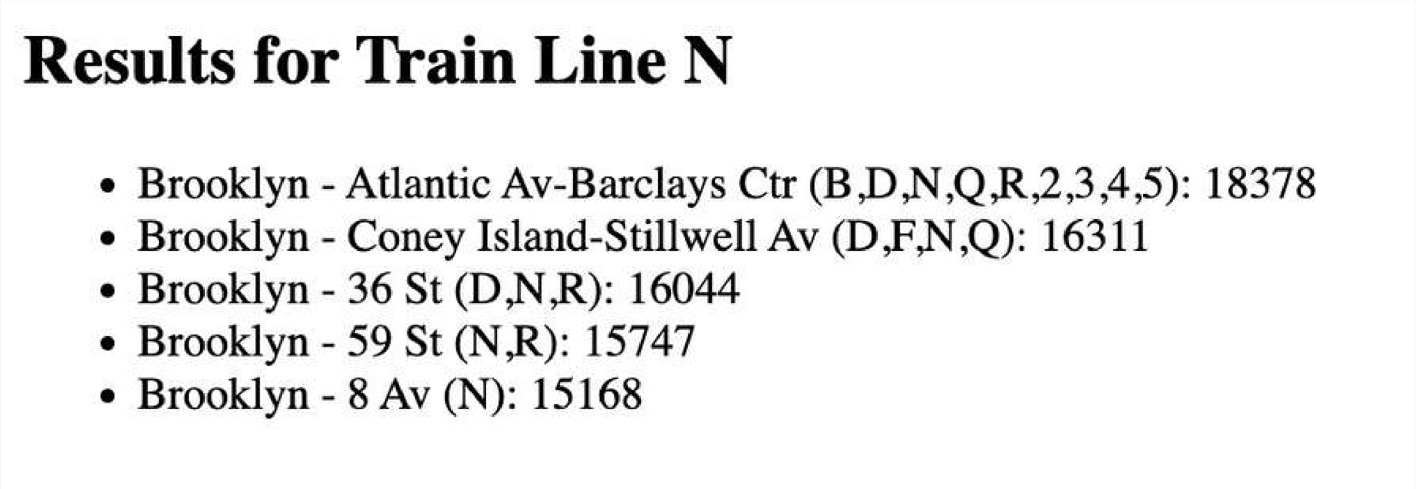
F Line, Brooklyn



Results for Train Line F

- Brooklyn - Jay St-MetroTech (A,C,F,R): 17148
- Brooklyn - Coney Island-Stillwell Av (D,F,N,Q): 16311
- Brooklyn - 4 Av (F,G)/9 St (R): 16110
- Brooklyn - Church Av (F,G): 15909
- Brooklyn - 7 Av (F,G): 15408

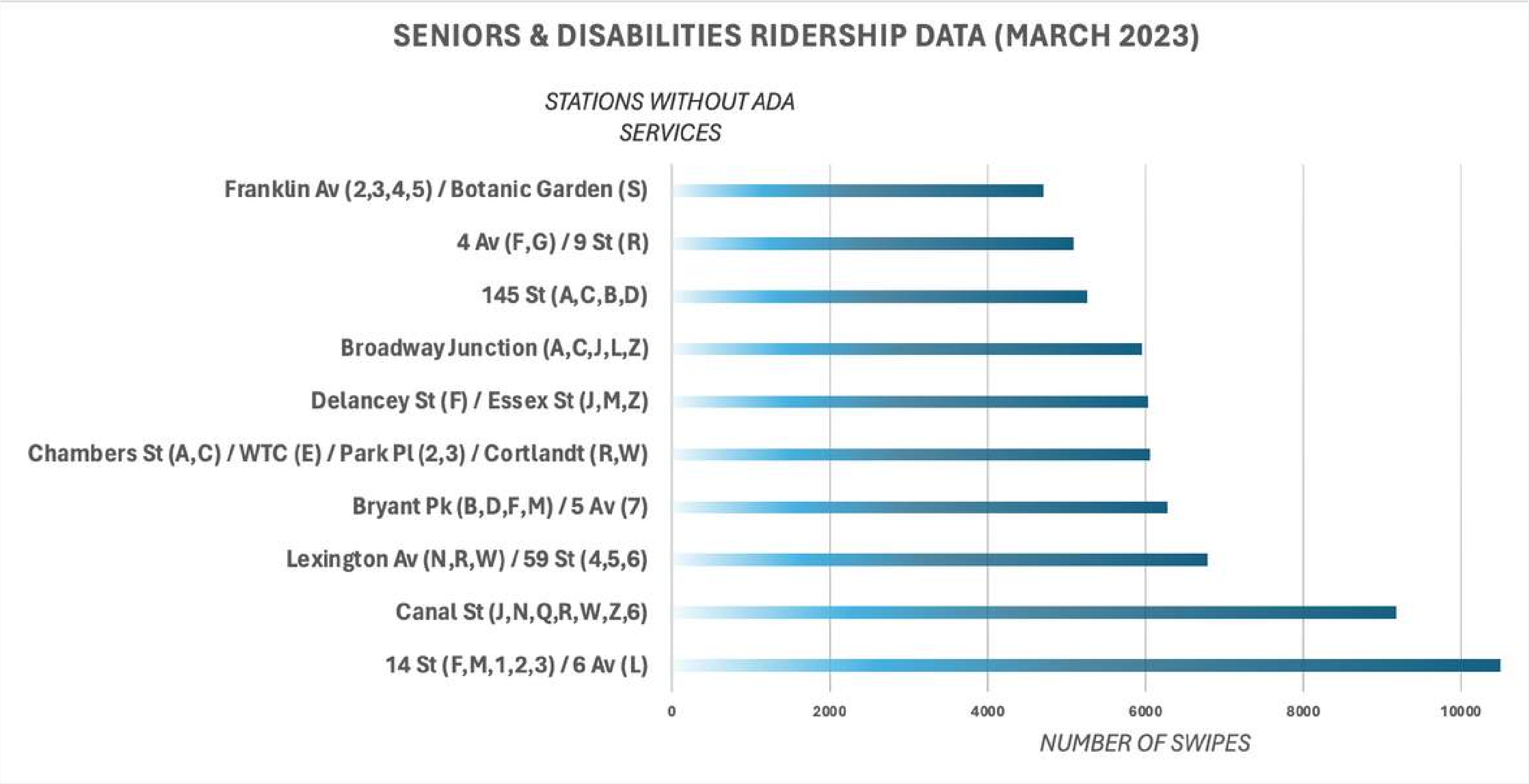
N Line, Brooklyn



Results for Train Line N

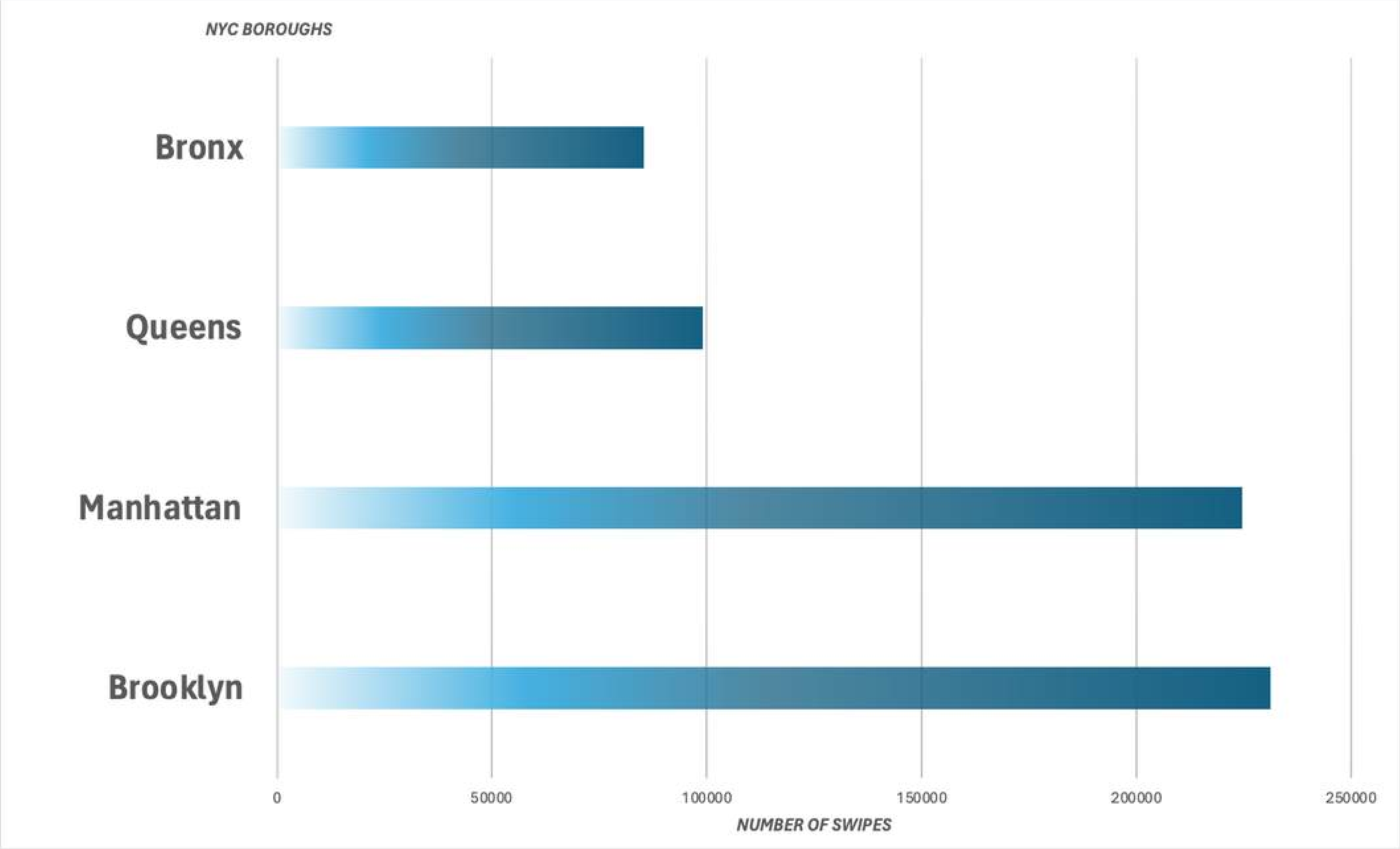
- Brooklyn - Atlantic Av-Barclays Ctr (B,D,N,Q,R,2,3,4,5): 18378
- Brooklyn - Coney Island-Stillwell Av (D,F,N,Q): 16311
- Brooklyn - 36 St (D,N,R): 16044
- Brooklyn - 59 St (N,R): 15747
- Brooklyn - 8 Av (N): 15168

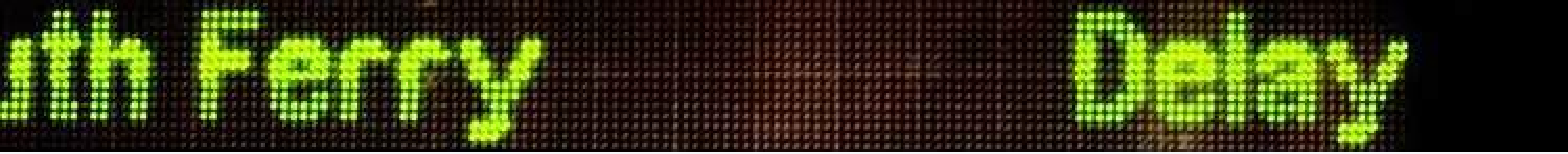
ADA ACCESSIBLE INFRASTRUCTURE



14th St / 6 Av station becomes the most important station to add fully ADA accessible infrastructure where it is the most number of swipes coming from an elderly ticket. The next station is for Canal St Station where fully ADA accessible infrastructure is needed.

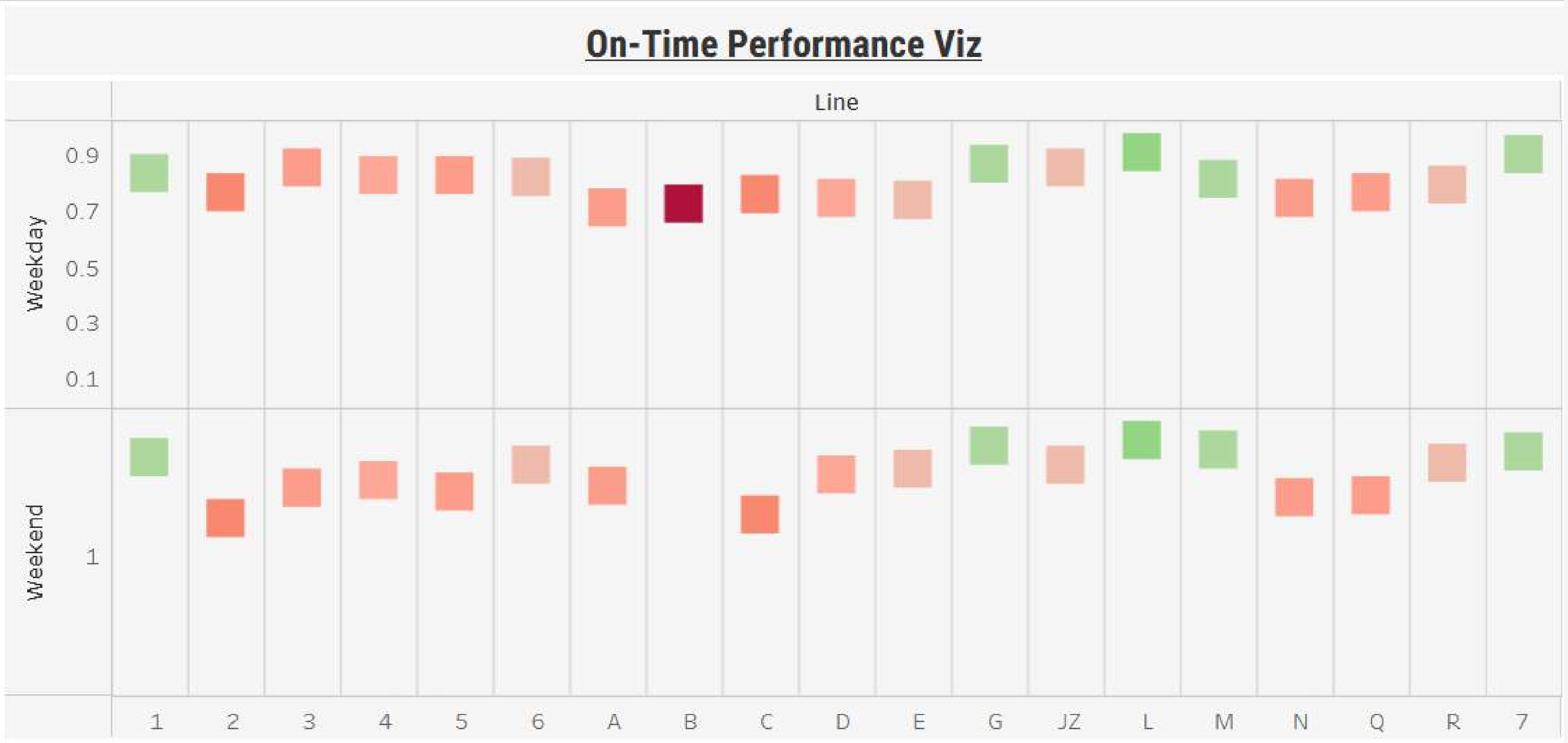
Brooklyn and Manhattan are the two most important boroughs to focus ADA accessible infrastructure on





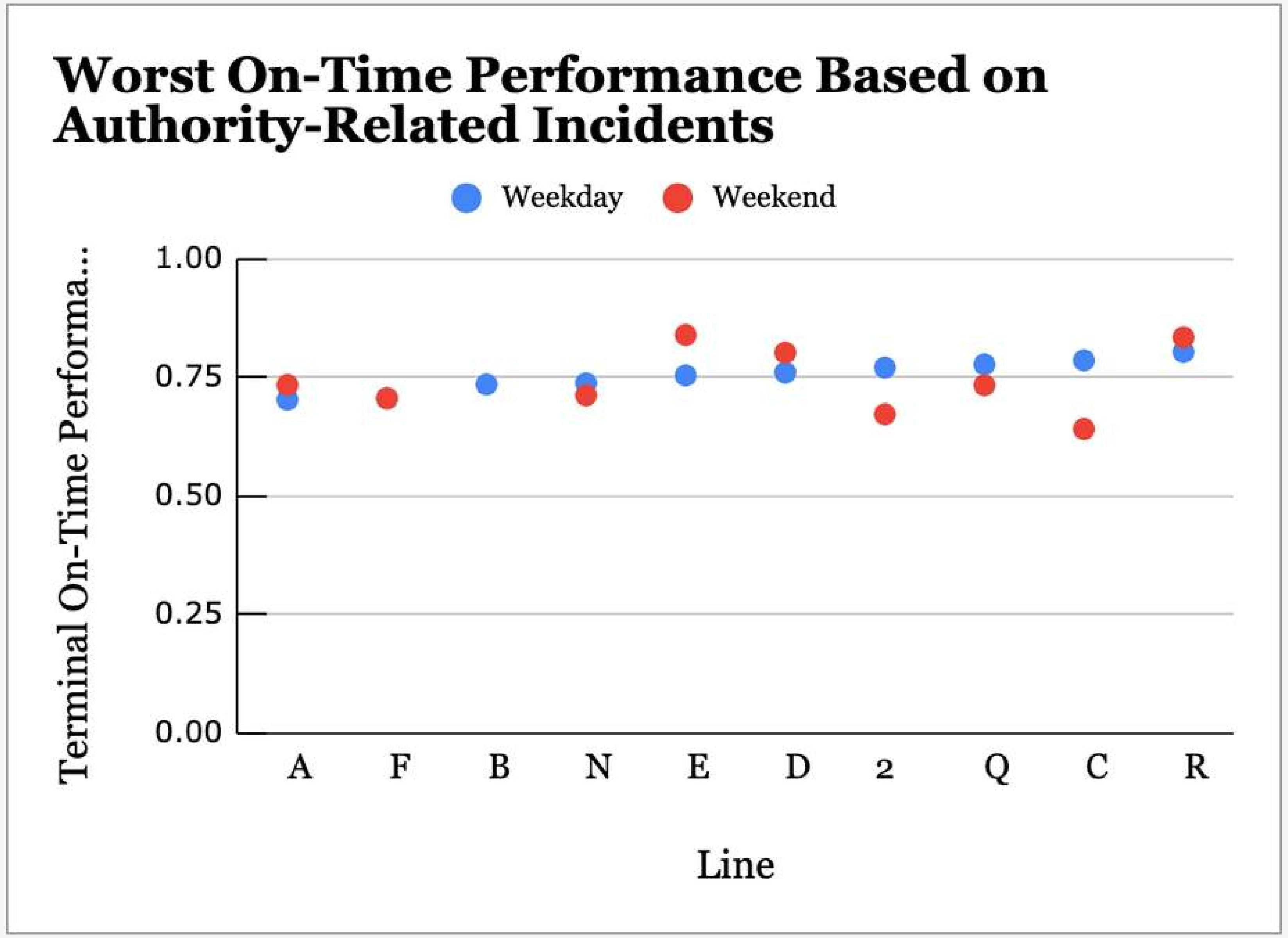
Terminal (Line) Performance

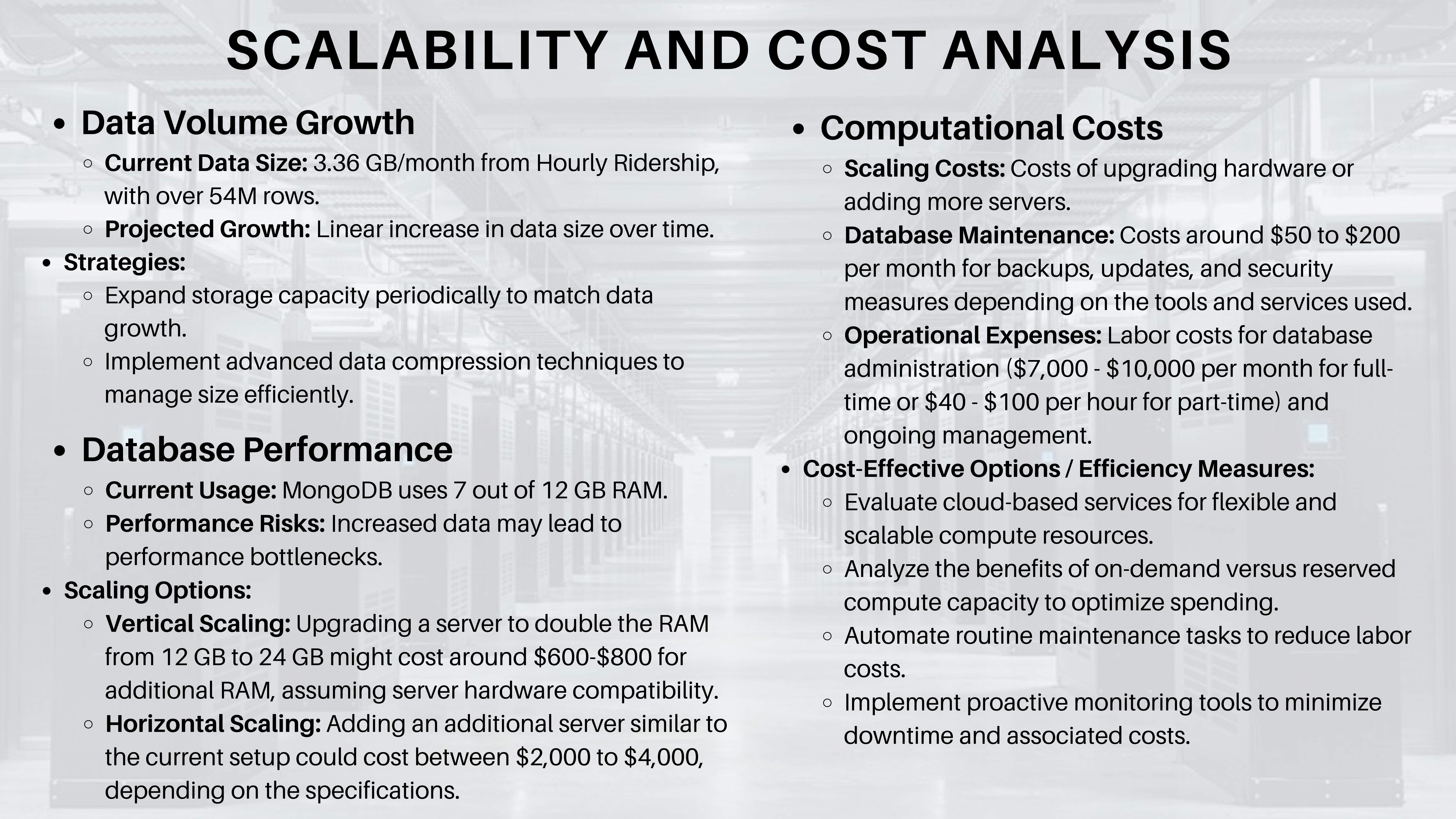
To figure which subway lines require maintenance and lines that need more safety



The A and F line have the worst overall on-time rating

‘A’ line has worst On-Time Performance on weekdays and ‘F’ has worst On-Time Performance during weekends due to Authority-Related Incidents. Terminals operating these lines should have stricter security protocols





SCALABILITY AND COST ANALYSIS

- ## Data Volume Growth
  - **Current Data Size:** 3.36 GB/month from Hourly Ridership, with over 54M rows.
  - **Projected Growth:** Linear increase in data size over time.
- **Strategies:**
  - Expand storage capacity periodically to match data growth.
  - Implement advanced data compression techniques to manage size efficiently.

- ## Database Performance
  - **Current Usage:** MongoDB uses 7 out of 12 GB RAM.
  - **Performance Risks:** Increased data may lead to performance bottlenecks.
- **Scaling Options:**
  - **Vertical Scaling:** Upgrading a server to double the RAM from 12 GB to 24 GB might cost around $600-$800 for additional RAM, assuming server hardware compatibility.
  - **Horizontal Scaling:** Adding an additional server similar to the current setup could cost between $2,000 to $4,000, depending on the specifications.

- ## Computational Costs
  - **Scaling Costs:** Costs of upgrading hardware or adding more servers.
  - **Database Maintenance:** Costs around $50 to $200 per month for backups, updates, and security measures depending on the tools and services used.
  - **Operational Expenses:** Labor costs for database administration ($7,000 - $10,000 per month for full-time or $40 - $100 per hour for part-time) and ongoing management.
- **Cost-Effective Options / Efficiency Measures:**
  - Evaluate cloud-based services for flexible and scalable compute resources.
  - Analyze the benefits of on-demand versus reserved compute capacity to optimize spending.
  - Automate routine maintenance tasks to reduce labor costs.
  - Implement proactive monitoring tools to minimize downtime and associated costs.

Conclusion

1. Each line has overcrowded stations that can be accomodated by new express lines, this can be analyzed using our dashboard result.
2. Brooklyn and Manhattan need more attention regarding ADA accessible infrastructure, especially in transit stations that have Rush-Hour Express.
3. The ‘A’ and ‘F’ line show the highest delays and biggest security hazards hence additional safety and security measures should be implemented in stations operating for those lines.

Recommendations

1. Incorporating more data especially information on train lines and their users would lead to more robust conclusions and wider problems can be tackled by following a similar working schema.
2. The MTA should use “weight-sensing” data to record which trains are the busiest. This could give users a visual on what trains need to be in more demand.

Dataset References:

https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-February-202/wujg-7c2s/about_data

https://data.ny.gov/Transportation/MTA-Subway-Terminal-On-Time-Performance-Beginning-/vtvh-gimj/about_data

https://data.ny.gov/Transportation/MTA-Subway-Major-Incidents-Beginning-2020/j6d2-s8m2/about_data

https://data.ny.gov/Transportation/MTA-Subway-Stations/39hk-dx4f/about_data

https://map.mta.info/ - Interactive Map for station and train information