# Analyzing NY Transit Ridership Trends

## Using PySpark and MongoDB for Data Processing and Visualization

Presenter:Carina Yan (yy3838), Xiaozhou Wen (xw3759)

Date: August 12, 2025

# Agenda

- Introduction to the Project
- Data Sources and Overview
- Analysis and Visualization
- Key Insights and Conclusions
- Q&A

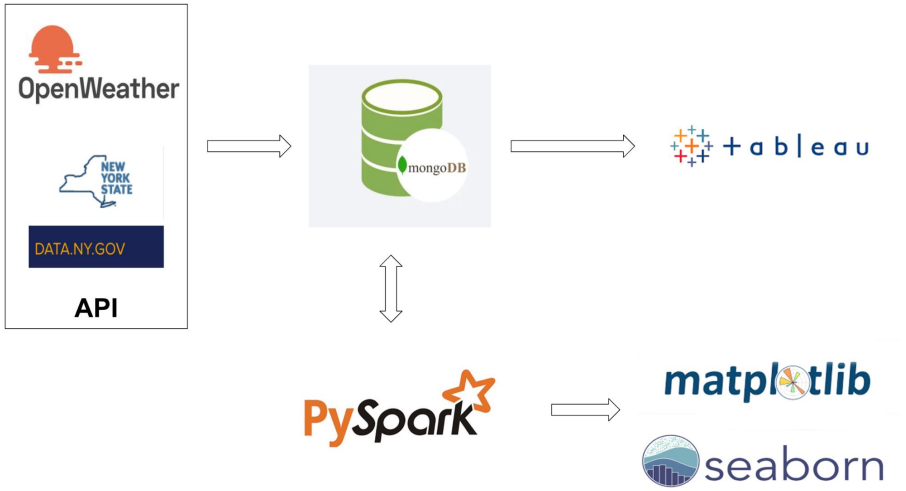
# Project Introduction

**Objective**: Analyze NYC transit ridership patterns. What time fo the day people uses subways? Is Monday ridership pattern same as Friday? Is November ridership pattern same as December? How weather affects the ridership?

**Why This Matters**: Helps metropolitan transportation authority optimize operations, predict demand, and understand external factors like weather.

**Tools Used**:

- MongoDB: Data storage
- PySpark: Scalable processing for large datasets
- Matplotlib: Visualization
- Tableau: Interactive Visualization



# Data Sources and Overview

## MTA Subway Hourly Ridership: 2020-2024

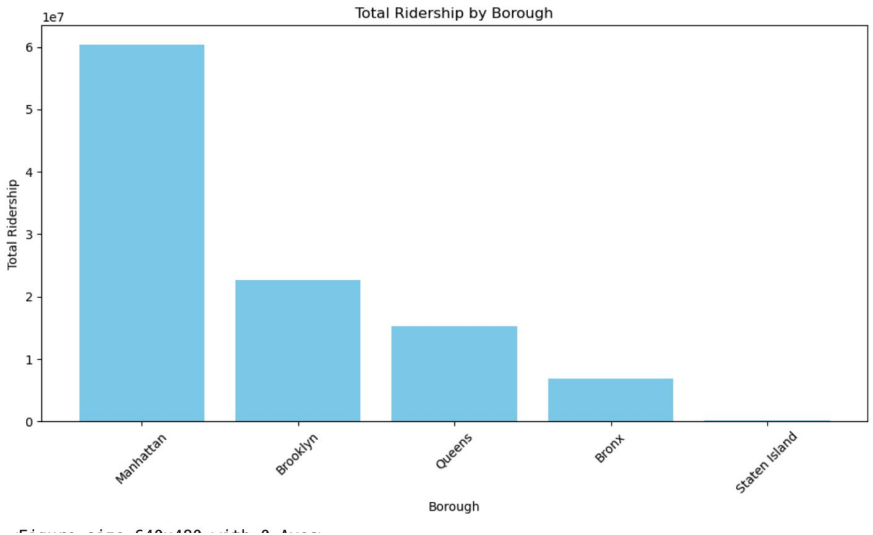
- This dataset provides subway ridership estimates on an hourly basis by subway station complex and class of fare payment.
- 2.4 M records per months

| transit_timestamp | transit_mode | station_complex_id | station_complex | borough | payment_method | fare_class_category | ridership | transfers | latitude | longitude |
|---|---|---|---|---|---|---|---|---|---|---|
| 12/1/24 0:00 | subway | 627 | Franklin Av (C,S) | Brooklyn | omny | OMNY - Full Fare | 15 | 0 | 40.68138 | -73.95685 |
| 12/1/24 0:00 | subway | 427 | West Farms Sq-E Tremont Av (2,5) | Bronx | omny | OMNY - Full Fare | 7 | 1 | 40.840294 | -73.88005 |
| 12/1/24 0:00 | subway | 240 | 7 Av (F,G) | Brooklyn | omny | OMNY - Full Fare | 29 | 0 | 40.66627 | -73.98031 |

## Historical Hourly NYC Weather Data

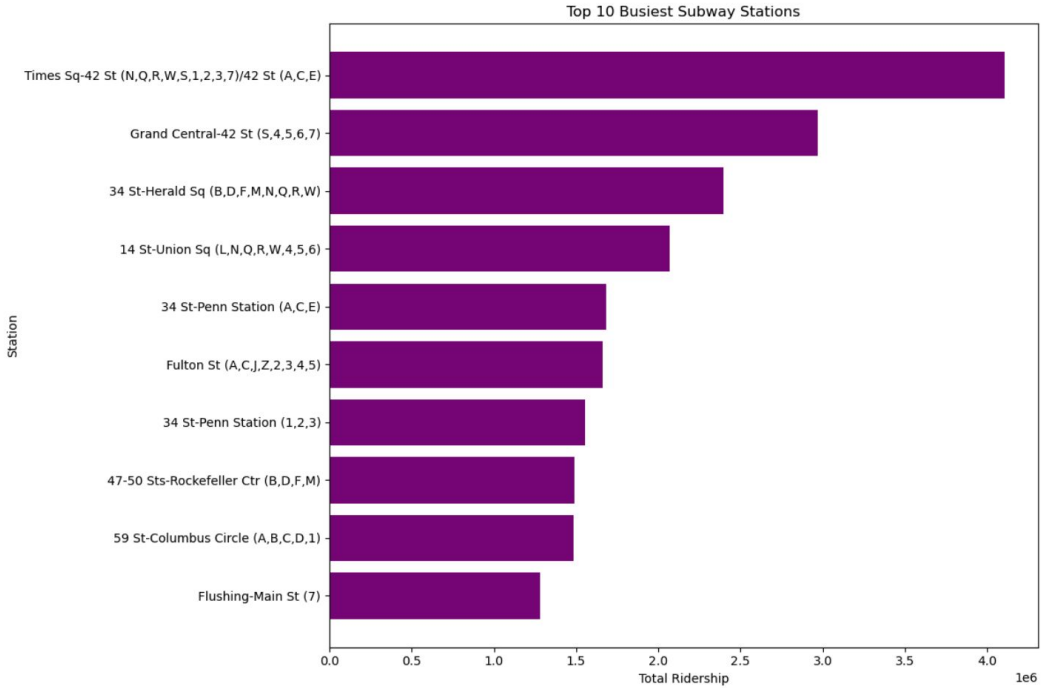
| dt_nyc | dt | id | main | description | icon | main.temp | main.feels_like | main.pressure | main.humidity | wind.speed | wind.deg | wind.gust |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024-09-01 00:00:00 EDT | 1725163200 | 802 | Clouds | scattered clouds | 03n | 295.84 | 296.41 | 1014 | 86 | 3.6 | 180 | NaN |
| 2024-09-01 01:00:00 EDT | 1725166800 | 800 | Clear | clear sky | 01n | 295.7 | 296.26 | 1014 | 86 | 3.09 | 180 | NaN |
| 2024-09-01 02:00:00 EDT | 1725170400 | 802 | Clouds | scattered clouds | 03n | 295.47 | 296.03 | 1014 | 87 | 3.09 | 190 | NaN |
| 2024-09-01 03:00:00 EDT | 1725174000 | 804 | Clouds | overcast clouds | 04n | 295.64 | 296.16 | 1013 | 85 | 4.12 | 210 | NaN |
| 2024-09-01 04:00:00 EDT | 1725177600 | 804 | Clouds | overcast clouds | 04n | 295.66 | 296.19 | 1013 | 85 | 3.6 | 190 | NaN |

# Total Ridership By Borough



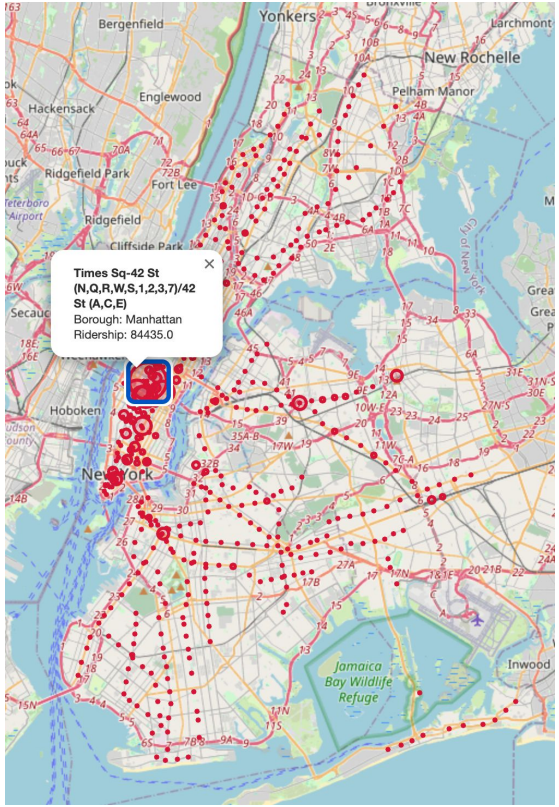

- Manhattan generally has the highest subway ridership among the five boroughs of New York City
- The combined ridership of all other boroughs is nearly equal to that of Manhattan, indicating that most people are commuting to and from other boroughs.

# Top 10 Busiest Subway Stations

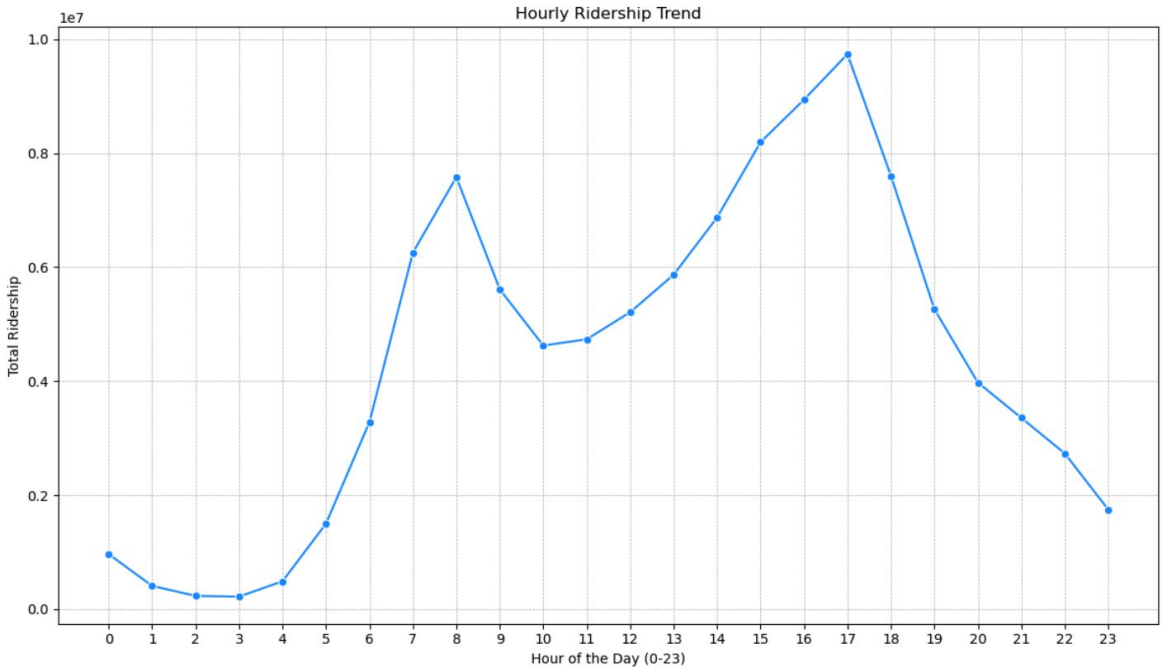



- Time Square is the busiest Subway station
- 9 stations are from Manhattan and 1 station is from Queens
- Most of the busiest subway stations are those that offer transfers to other subway lines.

# Station Ridership on the map

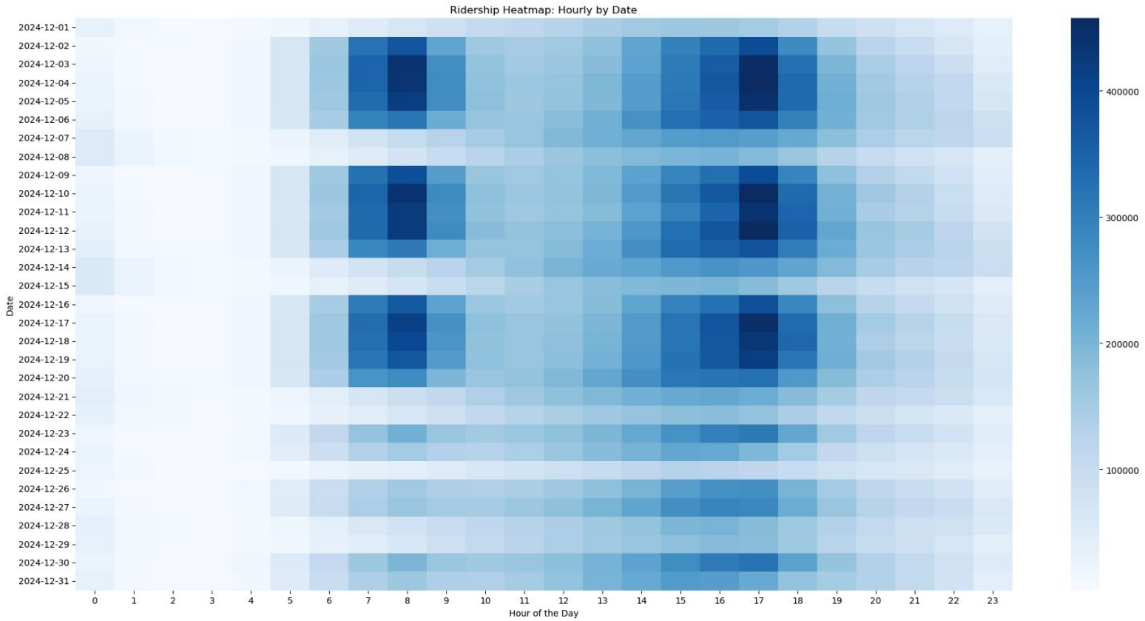


# Hourly Ridership Trend



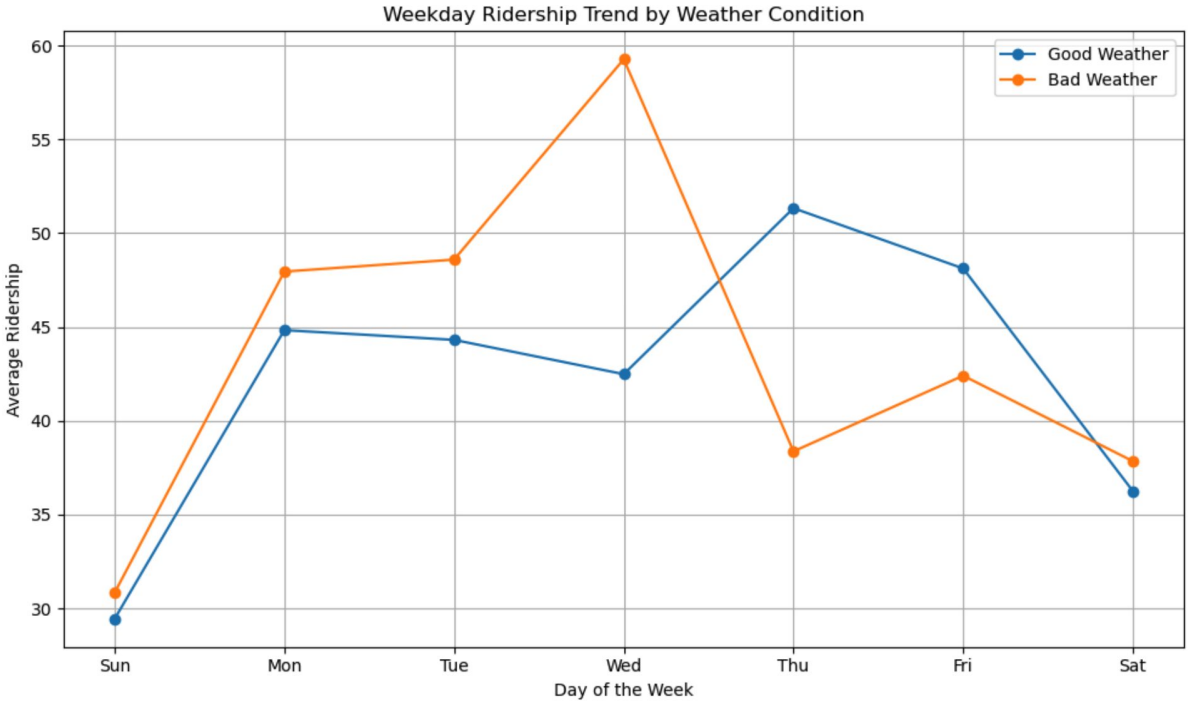

- Morning Rush hour is between 7am - 9am
- Evening Rush hour is between 3pm - 6pm

# Ridership Heatmap: Hourly by Date




- This is my favorite Chart. Easily compare hourly travel pattern between dates.
- Toward the end of the year, during the big holiday season, it’s hardly noticeable that there’s a rush hour on the subways—likely because most people choose to take vacation at that time.

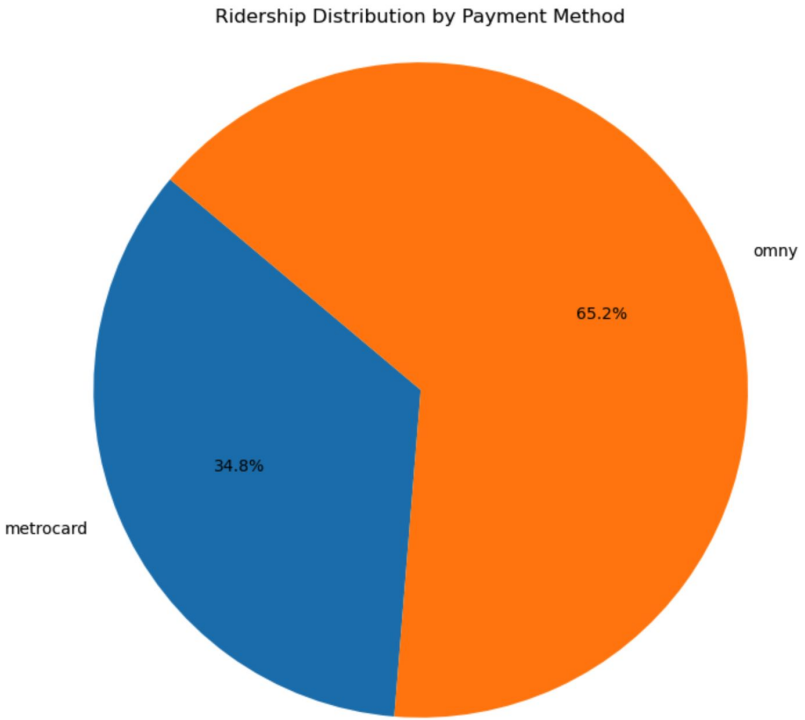
# Ridership Trend by Weather Condition




good_weather = ['Clear', 'Clouds']

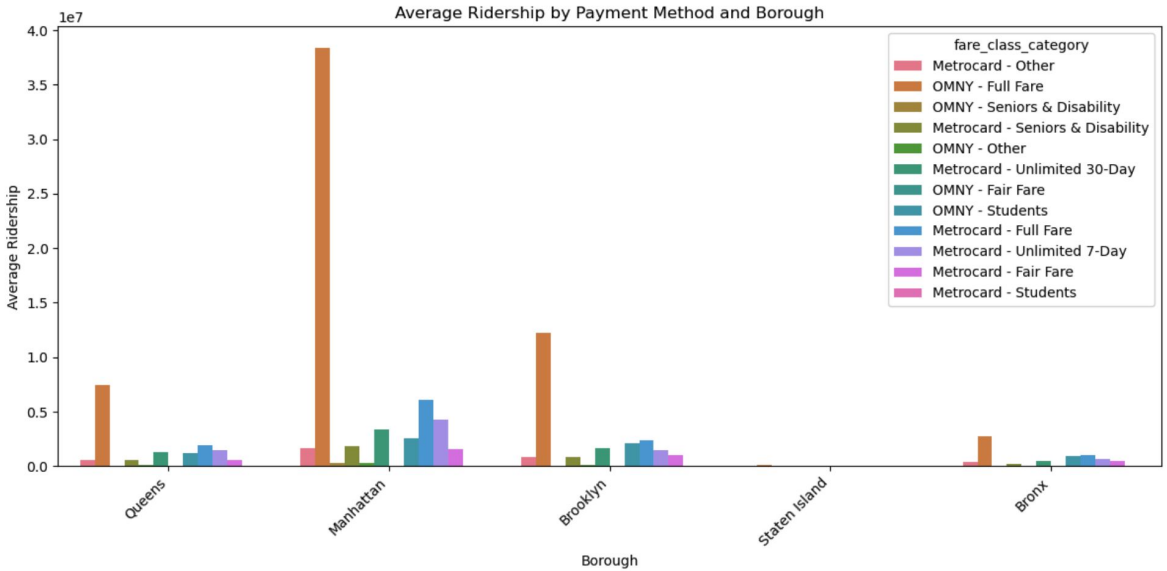
bad_weather = ['Rain', 'Thunderstorm', 'Snow', 'Drizzle','Fog', 'Mist', 'Haze', 'Smoke']

# Ridership Distribution by Payment Method



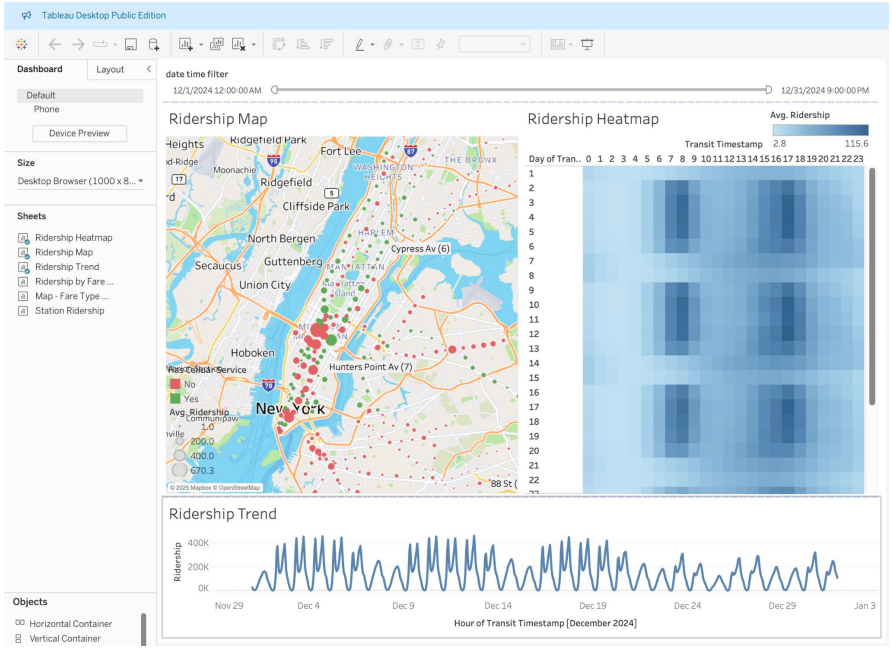

- Metrocard is replaced by OMNY by the end of 2025

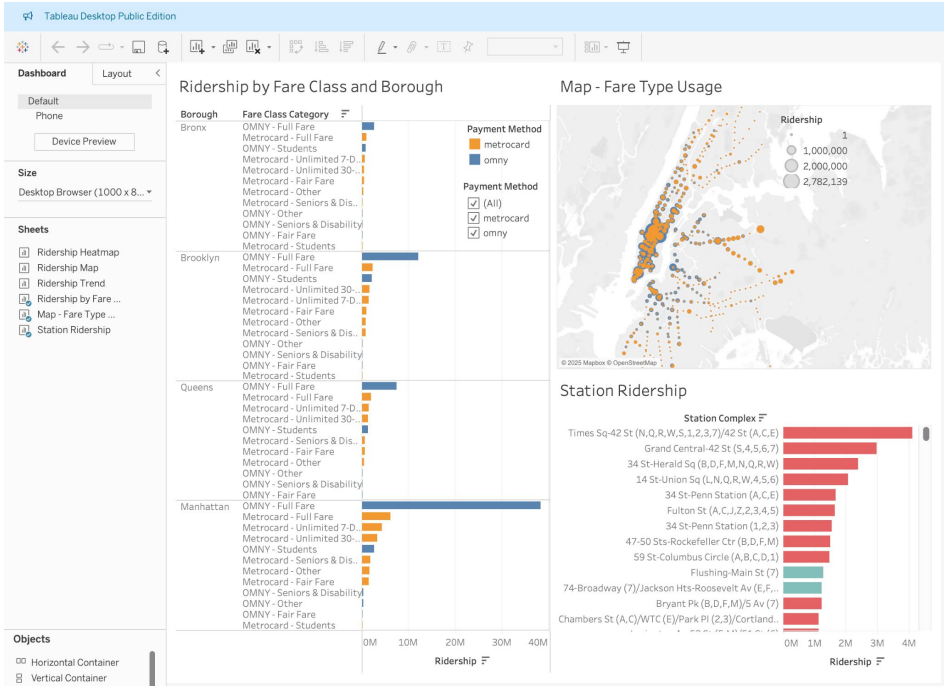
# Ridership by Payment method and Borough




- Manhattan has the highest adoption of OMNY
- Manhattan also has the highest ridership using metrocard

# Tableau - interactive dashboard



# Key Insights and Conclusions

This analysis of NYC transit ridership has revealed several key patterns and insights that are crucial for understanding and optimizing the city's transportation network.

- **Manhattan as the Epicenter of Ridership**: Manhattan dominates subway usage, boasting the highest overall ridership among all boroughs and home to nine of the ten busiest stations.
- **Predictable Daily Commuting Patterns**: Weekday ridership follows a clear and consistent pattern, with distinct morning rush hours between 7-9 AM and evening rush hours from 3-6 PM.
- **Holiday Season Alters Travel Behavior**: The typical rush hour patterns are significantly less pronounced during the major holiday seasons at the end of the year, likely due to an increase in vacation time.
- **OMNY is the Leading Payment Method**: The transition to the new OMNY payment system is well underway, accounting for 65.2% of ridership compared to MetroCard's 34.8%. Manhattan leads in both OMNY adoption and the remaining MetroCard usage.
- **Powerful Tools for Large-Scale Analysis**: The successful use of PySpark and MongoDB demonstrates the power of these tools for processing and analyzing large, complex datasets to extract meaningful transportation insights.