
University of New Haven

Pompea College of Business

**Course: BANL-6430-03 – Database Management for Business Analytics**

**Instructor: Dr. Pindaro Demertzoglou**

# Project Title:

## Port Authority Bus Terminal Passenger Prediction

## Project Progress Report-2

**Prepared By:**

**Aastha Kale**

**Ayesha Kabir**

**Meghana Lakshminarayana Swamy**

**Rajyalakshmi Nelakurthi**

**Venkata Sai Phanindra Namburu**

**Submission Date- March 17, 2025**

# Project Progress Report:

## 1. Data Sources and Integration Report

### 1. Available Data Sources

The project incorporates multiple datasets, both internal and external, to analyze the impact of various factors on passenger flow and mechanical failures.

#### Company-Provided Datasets

1. **Weather Data (Tbl_Weather)**

- Tracks weather conditions that could impact bus operations.
- Fields include: Date, PRCP (Precipitation), TMAX (Max Temperature), TMIN (Min Temperature), SNOW (Snowfall).

#### External Datasets

1. **MTA Bus Mechanical Failure Data**
    - Provides breakdown patterns over time to analyze their effect on passenger experience and ridership.
    - Used to assess service reliability and identify maintenance improvement areas.
      Available at: Data Catalog – Port Authority Bus Terminal

## 2. Data Preparation, Integration, and Cleaning

To ensure accurate analysis, the data is cleaned, formatted, and integrated using **Microsoft SQL**.

### Data Cleaning

1. **Checking for Missing Values**:

```
SELECT * FROM MTA_Bus WHERE MonthlyMiles IS NULL OR MonthlyRoadCallCount IS NULL OR MonthlyMDBF IS NULL;
SELECT * FROM Tbl_Weather WHERE PRCP IS NULL OR TMAX IS NULL OR TMIN IS NULL;
```

2. **Handling Missing Data**:
    - Replacing missing values in the weather dataset using **averages from nearby dates**:

```
UPDATE Tbl_Weather
SET AWND = (SELECT AVG(AWND) FROM Tbl_Weather
        WHERE DATE BETWEEN DATEADD(day, -3, DATE) AND DATEADD(day, 3, DATE))
WHERE AWND IS NULL;
```

**Standardizing Data Formats**:

- Converting Month field into a **DATE format** for compatibility:

```
ALTER TABLE MTA_Bus ADD MonthDate DATE;
UPDATE MTA_Bus
SET MonthDate = CONVERT(DATE, Month + '/01/' + RIGHT(Month, 4), 101);
```

**Data Integration**

- **Aggregating Weather Data to Monthly Level**:

```
SELECT DATEPART(year, DATE) AS Year, DATEPART(month, DATE) AS Month,
    AVG(TMAX) AS AvgTMAX, AVG(TMIN) AS AvgTMIN,
    SUM(PRCP) AS TotalPRCP, SUM(SNOW) AS TotalSNOW
INTO MonthlyWeather
FROM Tbl_Weather
GROUP BY DATEPART(year, DATE), DATEPART(month, DATE);
```

- **Joining Datasets (MTA Bus + Weather Data)**:

```
SELECT MTA.MonthDate, MTA.Borough, MTA.MonthlyMiles, MTA.MonthlyRoadCallCount,
MTA.MonthlyMDBF,
    W.AvgTMAX, W.AvgTMIN, W.TotalPRCP, W.TotalSNOW
INTO FinalDataset
FROM MTA_Bus MTA
LEFT JOIN MonthlyWeather W
ON DATEPART(year, MTA.MonthDate) = W.Year AND DATEPART(month, MTA.MonthDate) =
W.Month;
```

**Ensuring Data Quality**

- **Checking Row Consistency**:

```
SELECT COUNT(*) FROM MTA_Bus;
SELECT COUNT(*) FROM FinalDataset;
```

- This ensures that the integrated dataset does not lose records during merging.

## 3. Use of External Datasets

Yes, in addition to the company's provided datasets, we are incorporating an **external dataset: MTA Bus Mechanical Failure Data** to analyze the impact of mechanical breakdowns on **passenger flow, service reliability, and demand**.

**Importance of this dataset:**

- **Impact on Ridership**: Frequent bus breakdowns can cause delays, leading passengers to shift to alternative transport like subways or ride-sharing.
    - *Example:* If a route has high breakdowns, a decline in ticket sales may be observed.
- **Correlation with Service Availability**: More failures mean **fewer operational buses**, leading to overcrowding.
    - *Example:* If maintenance issues reduce bus availability during peak hours, wait times increase.
- **Maintenance Planning**: Identifying **seasonal breakdown patterns** allows for better fleet maintenance scheduling.
    - *Example:* If failures peak in winter, preventive maintenance can be **scheduled in advance**.
- **Pre/Post-COVID Comparison**: Analyzing trends before and after 2020 helps **assess the impact of aging fleets**.
    - *Example:* If failure rates increased post-pandemic, fleet upgrades may be necessary.

This dataset enhances the study by **correlating mechanical failures with passenger trends**, improving reliability and planning.

**4. Managing Large Datasets Efficiently**

After integrating the **MTA Bus Performance Data** and **Weather Data**, a large volume of records will be generated. To ensure efficient data management, we will use **SQL Server Management Studio (SSMS)** and apply the **GROUP BY** function to aggregate key metrics.

**Steps to Create a Manageable Dataset:**

- **Merging Data by Month:** The datasets will be joined on the **month and borough**, summarizing key variables such as:
    - **Breakdowns** (road call count)
    - **Miles Driven** (total monthly mileage)
    - **Weather Metrics** (average precipitation, temperature, snowfall)

- **Eliminating Redundancy & Incomplete Entries:** By grouping similar data, we will filter out **duplicate or missing records**, ensuring only relevant information is retained.
- **Optimizing for Quick Queries:** The final dataset will be structured **by month and borough**, allowing for **faster retrieval and analysis**.

The large dataset resulting from the joins, **SQL Server Management Studio (SSMS)** is used to efficiently manage and summarize the data.

- **Grouping the data by month and borough**:

```
SELECT MonthDate, Borough,
    SUM(MonthlyMiles) AS TotalMiles,
    SUM(MonthlyRoadCallCount) AS TotalRoadCalls,
    AVG(MonthlyMDBF) AS AvgMDBF,
    AVG(AvgTMAX) AS AvgTMAX, AVG(AvgTMIN) AS AvgTMIN,
    SUM(TotalPRCP) AS TotalPRCP, SUM(TotalSNOW) AS TotalSNOW
INTO ManageableDataset
FROM FinalDataset
GROUP BY MonthDate, Borough;
```

- **Benefits of this approach**:
    - Reduces redundancy.
    - Optimizes query performance.
    - Ensures quick access to data insights.

**Outcome:**

This approach **reduces data complexity, improves query performance, and creates a streamlined dataset** optimized for **trend analysis and forecasting**.

## 5. Final Dataset Showcase with SQL Query

The final dataset is structured to provide **comprehensive insights** into bus performance and weather impact across **various NYC boroughs**.

**Dataset Summary**

This **SQL-based dataset** tracks:

- **Bus Service Performance:**

    - Monthly **miles traveled** per borough.
    - **Road call counts** (mechanical failures).
    - **Mean Distance Between Failures (MDBF)**, a key reliability metric.
- **Weather Conditions:**
    - **Average Maximum Temperature**: [avgtmax]°F
    - **Average Minimum Temperature**: [avgtmin]°F
    - **Total Precipitation**: [totalprcp] inches
    - **Total Snowfall**: [totalsnow] inches

**Key Insights Enabled by the Dataset**

- **Breakdown Trends:** Helps in analyzing fleet performance and identifying high-failure routes.
- **Maintenance Planning:** Identifies peak failure months to optimize **preventive maintenance schedules**.
- **Climate Impact on Bus Performance:** Correlates **weather conditions with bus failures**, supporting transit resilience strategies.
- **Optimized Transit Operations:** Enables data-driven decision-making for **resource allocation and scheduling improvements**.

**SQL Query Used to Generate the Final Dataset**

To generate the final dataset, the following **SQL query** is used:

```
SELECT
    MTA_Bus.Month AS ReportMonth,
    MTA_Bus.Borough,
    MTA_Bus.MonthlyMiles,
    MTA_Bus.MonthlyRoadCallCount,
    MTA_Bus.MonthlyMDBF,
    Weather.AvgTMAX,
    Weather.AvgTMIN,
    Weather.TotalPRCP,
    Weather.TotalSNOW
INTO FinalDataset
FROM MTA_Bus
LEFT JOIN Weather
```

```
ON DATEPART(year, MTA_Bus.Month) = Weather.Year
AND DATEPART(month, MTA_Bus.Month) = Weather.Month
ORDER BY ReportMonth, Borough;
```

## Conclusion

This project integrates **MTA Bus Performance Data** and **Weather Data** to analyze the impact of mechanical failures on **passenger flow, service reliability, and maintenance planning**. By leveraging **SQL-based data cleaning, integration, and aggregation techniques**, we have developed a **comprehensive dataset** that provides **insights into transit operations** across NYC boroughs.

### Key Findings & Insights

1. **Mechanical Failures Impact Ridership**
    - Frequent breakdowns correlate with **longer wait times, overcrowding, and potential ridership declines** as commuters shift to alternative transport options.
2. **Service Availability & Fleet Management**
    - Increased **road call counts** result in fewer operational buses, leading to **higher congestion at bus stops and route inefficiencies**.
    - **Pre/Post-COVID comparison** highlights the need for **fleet upgrades and preventive maintenance**.
3. **Weather Conditions Affect Bus Performance**
    - Harsh weather (e.g., **snow and heavy precipitation**) contributes to **higher mechanical failures and reduced service efficiency**.
    - Understanding these trends allows for **better scheduling and resource allocation**.
4. **Data Optimization for Decision-Making**
    - The **final dataset is structured efficiently**, ensuring **quick query execution and improved analysis**.
    - SQL-based **grouping, filtering, and summarization** have enabled **a streamlined, manageable dataset** for forecasting.

### Dataset 1:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | DATE | AWND | PRCP | SNOW | SNWD | TMAX | TMIN |
| 2 | 2024-03-08 00:00:00 | 5.14 | 0 | 0 | 0 | 57 | 40 |
| 3 | 2024-03-09 00:00:00 | 7.38 | 1.53 | 0 | 0 | 49 | 41 |
| 4 | 2024-03-10 00:00:00 | 9.62 | 0.03 | 0 | 0 | 51 | 37 |
| 5 | 2024-03-11 00:00:00 | 12.75 | 0 | 0 | 0 | 52 | 35 |
| 6 | 2024-03-12 00:00:00 | 6.04 | 0 | 0 | 0 | 66 | 43 |
| 7 | 2024-03-13 00:00:00 | 3.58 | 0 | 0 | 0 | 62 | 48 |
| 8 | 2024-03-14 00:00:00 | 2.46 | 0 | 0 | 0 | 74 | 46 |
| 9 | 2024-03-15 00:00:00 | 6.71 | 0 | 0 | 0 | 73 | 51 |
| 10 | 2024-03-16 00:00:00 | 4.47 | 0 | 0 | 0 | 61 | 47 |
| 11 | 2024-03-17 00:00:00 | 6.71 | 0 | 0 | 0 | 63 | 48 |
| 12 | 2024-03-18 00:00:00 | 7.38 | 0 | 0 | 0 | 51 | 38 |
| 13 | 2024-03-19 00:00:00 | 7.61 | 0 | 0 | 0 | 48 | 36 |
| 14 | 2024-03-20 00:00:00 | 6.71 | 0.01 | 0 | 0 | 57 | 34 |
| 15 | 2024-03-21 00:00:00 | 9.17 | 0 | 0 | 0 | 43 | 30 |
| 16 | 2024-03-22 00:00:00 | 4.92 | 0 | 0 | 0 | 46 | 29 |
| 17 | 2024-03-23 00:00:00 | 7.16 | 3.66 | 0 | 0 | 50 | 35 |
| 18 | 2024-03-24 00:00:00 | 7.83 | 0 | 0 | 0 | 48 | 31 |
| 19 | 2024-03-25 00:00:00 | 8.72 | 0 | 0 | 0 | 53 | 35 |
| 20 | 2024-03-26 00:00:00 | 6.04 | 0 | 0 | 0 | 53 | 39 |
| 21 | 2024-03-27 00:00:00 | 2.68 | 0.06 | 0 | 0 | 52 | 42 |
| 22 | 2024-03-28 00:00:00 | 4.7 | 0.7 | 0 | 0 | 51 | 44 |
| 23 | 2024-03-29 00:00:00 | 11.18 | 0 | 0 | 0 | 56 | 42 |
| 24 | 2024-03-30 00:00:00 | 8.5 | 0.04 | 0 | 0 | 62 | 41 |
| 25 | 2024-03-31 00:00:00 | 4.47 | 0.02 | 0 | 0 | 62 | 47 |
| 26 | 2024-04-01 00:00:00 | 2.91 | 0.01 | 0 | 0 | 54 | 49 |
| 27 | 2024-04-02 00:00:00 | 6.93 | 0.87 | 0 | 0 | 51 | 44 |
| 28 | 2024-04-03 00:00:00 | 14.09 | 1.55 | 0 | 0 | 45 | 41 |
| 29 | 2024-04-04 00:00:00 | 6.71 | 0.21 | 0 | 0 | 53 | 37 |
| 30 | 2024-04-05 00:00:00 | 8.05 | 0 | 0 | 0 | 50 | 38 |
| 31 | 2024-04-06 00:00:00 | 7.61 | 0 | 0 | 0 | 54 | 42 |
| 32 | 2024-04-07 00:00:00 | 6.49 | 0 | 0 | 0 | 60 | 43 |
| 33 | 2024-04-08 00:00:00 | 2.24 | 0 | 0 | 0 | 69 | 43 |

Sheet1

**Dataset 2:**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Month | Borough | Monthly Miles | Monthly Road Call Count | Monthly MDBF | | |
| 2 | 2015-01-01 00:00:00 | Bronx | 2166371 | 434 | 4992 | | |
| 3 | 2015-01-01 00:00:00 | Brooklyn | 2901602 | 576 | 5038 | | |
| 4 | 2015-01-01 00:00:00 | Manhattan | 1283763 | 458 | 2803 | | |
| 5 | 2015-01-01 00:00:00 | Queens | 4032200 | 810 | 4978 | | |
| 6 | 2015-01-01 00:00:00 | Staten Island | 2029610 | 230 | 8824 | | |
| 7 | 2015-02-01 00:00:00 | Bronx | 1962258 | 518 | 3788 | | |
| 8 | 2015-02-01 00:00:00 | Brooklyn | 2767879 | 571 | 4847 | | |
| 9 | 2015-02-01 00:00:00 | Manhattan | 1200788 | 422 | 2845 | | |
| 10 | 2015-02-01 00:00:00 | Queens | 3675299 | 956 | 3844 | | |
| 11 | 2015-02-01 00:00:00 | Staten Island | 1906088 | 278 | 6856 | | |
| 12 | 2015-03-01 00:00:00 | Bronx | 2258765 | 609 | 3709 | | |
| 13 | 2015-03-01 00:00:00 | Brooklyn | 3117255 | 617 | 5052 | | |
| 14 | 2015-03-01 00:00:00 | Manhattan | 1351993 | 434 | 3115 | | |
| 15 | 2015-03-01 00:00:00 | Queens | 4310871 | 991 | 4350 | | |
| 16 | 2015-03-01 00:00:00 | Staten Island | 2184715 | 303 | 7210 | | |
| 17 | 2015-04-01 00:00:00 | Bronx | 2279624 | 511 | 4461 | | |
| 18 | 2015-04-01 00:00:00 | Brooklyn | 3044734 | 474 | 6423 | | |
| 19 | 2015-04-01 00:00:00 | Manhattan | 1316299 | 430 | 3061 | | |
| 20 | 2015-04-01 00:00:00 | Queens | 4270979 | 950 | 4496 | | |
| 21 | 2015-04-01 00:00:00 | Staten Island | 2167176 | 233 | 9301 | | |
| 22 | 2015-05-01 00:00:00 | Bronx | 2231238 | 558 | 3999 | | |
| 23 | 2015-05-01 00:00:00 | Brooklyn | 3075059 | 480 | 6406 | | |
| 24 | 2015-05-01 00:00:00 | Manhattan | 1301748 | 444 | 2932 | | |
| 25 | 2015-05-01 00:00:00 | Queens | 4164297 | 982 | 4241 | | |
| 26 | 2015-05-01 00:00:00 | Staten Island | 2086027 | 271 | 7698 | | |
| 27 | 2015-06-01 00:00:00 | Bronx | 2248911 | 641 | 3508 | | |
| 28 | 2015-06-01 00:00:00 | Brooklyn | 3019898 | 524 | 5763 | | |
| 29 | 2015-06-01 00:00:00 | Manhattan | 1301651 | 493 | 2640 | | |
| 30 | 2015-06-01 00:00:00 | Queens | 4218402 | 1088 | 3877 | | |
| 31 | 2015-06-01 00:00:00 | Staten Island | 2148061 | 328 | 6549 | | |
| 32 | 2015-07-01 00:00:00 | Bronx | 2324600 | 618 | 3761 | | |
| 33 | 2015-07-01 00:00:00 | Brooklyn | 3053172 | 515 | 5928 | | |

Sheet1

**This is the Screenshot for final dataset available for further analysis.**