Final Group Project Report



# Group 6 - Team members and Individual Contribution

|  |  |
| --- | --- |
| NAME | INDIVIDUAL CONTRIBUTION |
| Giacomo Bizzotto | Led the overall project management, overseeing all stages from data preparation to final delivery. Contributed significantly to goals 1, 2, and 5 by developing the corresponding forecasting models, writing the reports, and synthesizing key findings. Designed and structured the final dashboard, coordinated team responsibilities, and ensured alignment with the company's requirements. |
| Rishitha Thatipally | Contributed to the creation of the final presentation and dashboard visualizations. |
| Irfan Shaik | Focused on goal 3 by developing the forecasting model for individual carriers using the Prophet algorithm. Prepared and analyzed the relevant dataset, ensuring model accuracy and completeness for carrier-specific projections. |
| Errolla Vivek | Assisted with data collection |
| Shraddha Shakya | Contributed to goal 4 by applying the ARIMA model for identifying peak usage periods across facilities. Enhanced the dashboard’s visual appeal and usability, making the results more accessible for stakeholders. |
| Sampath Sai Raghav Allaboyina | Collaborated on the final presentation and contributed to the recommendation section. Provided technical and organizational support to group members and assisted in the development of dashboards and forecasts across multiple project goals. |

# Introduction

This report provides a comprehensive summary of the work conducted to date for the Port Authority of New York and New Jersey, supporting data-driven infrastructure planning and operational decision-making for the period 2025 to 2030. While this document offers a structured overview of our methodologies and findings, it is not intended to replace the detailed analyses and technical explanations presented in previous reports. Rather, it serves as a concise, professional, and pragmatic synthesis of our work, aligned with the organization’s five core business questions.

Each question is addressed through a combination of advanced forecasting models, rigorous data integration processes, and targeted visualization strategies. For every question, our response is organized into four key components: (1) identification of the most relevant variables and the selected dependent variable, (2) practical recommendations related to data quality and the inclusion of additional variables, (3) a summary of the models and analytical techniques employed, and (4) the tools and technologies used throughout the project.

# Question 1: Forecasting Total Passenger Volumes (2025–2030)

### *1. Key Variables and Dependent Variable Selection*

To forecast total passenger volumes for the Port Authority bus terminals from 2025 to 2030, we considered a comprehensive set of variables. Temporal variables—such as year and month—were foundational in capturing recurring patterns and long-term ridership trends. The dependent variable selected was the total number of passengers forecasted per facility. This was calculated by first forecasting the number of buses using the ETS model, then multiplying this forecast by the average number of passengers per bus. The average was derived using SQL joins between the MBT\_Passenger\_Departures and MBT\_Bus\_Departures tables from the Unit\_571\_Database. External weather conditions were included as secondary predictors to enrich the model and add context. These included average wind speed (AWND), precipitation (PRCP), snowfall (SNOW), snow depth (SNWD), and temperature extremes (TMAX, TMIN). This combination of internal and external variables allowed for facility-level granularity in forecasting passenger demand.

### *2. Data Quality and Additional Considerations*

Data cleaning and integration were critical given the size and complexity of the datasets—especially the TBT\_Traffic\_Database\_PA-1, which contains over 5 million records. We handled missing values through linear interpolation for time series data and mean or median imputation for numerical fields, depending on the skewness of the distribution. To avoid bias from pandemic-related anomalies, data from March to June and November to December 2020 were excluded. We used SQL Server Management Studio (SSMS) to efficiently join and manage data across multiple tables, ensuring referential integrity and preparing the structure for modeling in R. Further improvements could be made by integrating economic indicators such as employment rates, regional GDP, or telecommuting rates, which would help explain longer-term shifts in transit demand. Additionally, continued attention to data normalization and outlier management was necessary to preserve model reliability.

### *3. Models and Techniques Applied*

We used the ETS (Error, Trend, Seasonality) model in R to generate the passenger forecasts. ETS was selected for its proven ability to model time series data with strong seasonality and trend components, and for its flexibility in adapting to structural breaks—particularly important post-COVID. In some facilities, an ETS+X strategy was employed to incorporate weather variables as external regressors, which added contextual nuance, especially in months with extreme weather fluctuations. This enabled more dynamic short-term predictions without sacrificing long-range trend stability.

### *4. Tools and Technologies Used*

The analysis for this question utilized a combination of R (for statistical modeling and forecasting), SQL Server Management Studio (SSMS) (for data extraction, transformation, and integration), Excel (for formatting and exploratory review), and Power BI (to create interactive dashboards visualizing forecasted trends by facility, year, and month).

# Question 2: Identifying Key Predictors of Passenger Volume

### *1. Key Variables and Dependent Variable Selection*

The dependent variable in this case was monthly passenger volume. The most significant predictors for this were temporal features such as year, month, and lagged passenger values. Historical ridership patterns captured through these time-based variables demonstrated strong predictive power. These features allowed the model to reflect long-term growth trends, seasonal fluctuations, and cyclical commuting behavior. Additional weather variables—namely average wind speed (AWND), precipitation (PRCP), snowfall (SNOW), snow depth (SNWD), and max/min temperatures (TMAX and TMIN)—were considered for their potential short-term influence on ridership behavior.

### *2. Data Quality and Additional Considerations*

Data preparation involved standard imputation techniques and the exclusion of pandemic-skewed months. Although environmental features showed limited improvement for long-term accuracy, they contributed to enhanced short-term responsiveness in the models. Using R's tslm() function, we evaluated the contribution of weather data through regression to decide whether to include those variables or not. For example, the adjusted R² reached 0.83 for Holland Tunnel, 0.68 for Outerbridge Crossing, and 0.55 for Lincoln Tunnel, indicating strong explanatory value in localized contexts. In future work, incorporating dynamic economic and policy-related variables (e.g., gas prices, remote work trends, school reopening schedules) could improve both interpretability and forecast robustness.

### *3. Models and Techniques Applied*

We used a combination of ETS and TSLM (Time Series Linear Models) in R. ETS formed the baseline forecasting model, while TSLM was applied to measure the incremental predictive value of external regressors like weather variables.

### *4. Tools and Technologies Used*

R for statistical modeling and evaluation, and SQL Server Management Studio (SSMS) for managing the relational database and structured queries.

# Question 3: Forecasting Ridership by Individual Carrier

### *1. Key Variables and Dependent Variable Selection*

This task aimed to forecast weekly ridership per bus operator using carrier-specific data from January 2020 onward. The dependent variable was passenger volume per carrier per week. Time (weeks) served as the main independent variable, capturing seasonal and recurring trends. The dataset included key carriers such as NJ Transit, Greyhound, Academy, Peter Pan, Martz, Coach USA, and others. Each carrier exhibited unique seasonality, holiday-driven surges, and ridership recovery patterns following COVID disruptions. These dynamics necessitated individualized forecasting strategies.

### *2. Data Quality and Additional Considerations*

We handled missing weekly data using median imputation and structured all datasets into uniform weekly intervals. The Prophet model’s flexibility in dealing with missing observations and its ability to automatically detect change points were crucial for producing high-fidelity forecasts for each carrier. However, including special events (e.g., sports games, public holidays, regional policy changes) could further explain carrier-specific fluctuations. The resulting forecasts were consolidated in Power BI dashboards with filters for date range and carrier name, enabling strategic comparisons.

### *3. Models and Techniques Applied*

The Prophet model was used for its capacity to handle multiple seasonal effects, incorporate holiday/event awareness, and provide robust forecasts in the presence of irregular patterns. Each bus operator’s time series was modeled independently to account for carrier-specific behaviors.

### *4. Tools and Technologies Used*

R (for Prophet modeling), Excel (for data structuring), and Power BI (for dynamic visualization by carrier and date).

# Question 4: Determining Busiest Periods for Staging Facilities

### *1. Key Variables and Dependent Variable Selection*

The objective of this question was to identify the busiest weeks, months, and years for staging facilities. The dependent variable was total traffic volume (buses, trucks, and cars), aggregated by facility and time period (week, month, year). The independent variable was time. Forecasts were built on historical traffic data from 2015 to 2024. A 5% uplift was applied to missing 2024 values by extrapolating from 2023 data. Weekly time series were generated for each facility to support high-resolution peak traffic forecasts.

### *2. Data Quality and Additional Considerations*

Traffic volumes were aggregated in SQL and cleaned in Excel before modeling. Incomplete 2024 data was completed through proportional year-over-year growth, preserving seasonal patterns. Although effective for capacity planning, accuracy could be improved by including context-aware variables like major construction projects, weather alerts, or congestion pricing policies. The forecast output was deployed in Power BI, enabling week-by-week facility-specific analysis through slicers and filters.

### *3. Models and Techniques Applied*

ARIMA models were used to forecast traffic volume. Weekly data was converted into time series for each facility, and auto.arima() in R selected optimal parameters based on historical structure and autocorrelation.

### *4. Tools and Technologies Used*

R (for ARIMA modeling), Excel (for cleaning and preparing data), SQL (for aggregation), and Power BI (for visual analysis).

# Question 5: Comparison of Current Usage to Pre-COVID Levels (2019)

### *1. Key Variables and Dependent Variable Selection*

This comparative analysis focused on total annual passenger volumes, using 2019 as a baseline year. Additional indicators such as passengers per bus and vehicle traffic volume (cars, trucks, and buses) were also reviewed to measure shifts in transportation behavior. The dependent variable was total ridership by year, with an emphasis on analyzing post-pandemic recovery trends relative to the pre-pandemic peak.

### *2. Data Quality and Additional Considerations*

Using outputs from ETS, Prophet, and ARIMA models, we found that total ridership is projected to stabilize at 145–165 million annually through 2030—about 40–50 million short of 2019 levels. This is in stark contrast to increasing vehicle traffic, forecasted to reach 108 million vehicles by 2030. The data suggests a structural shift toward private travel modes. Hybrid work adoption and health concerns appear to have permanently altered transit demand. Additional data sources such as mobile tracking, commuter surveys, and telework penetration rates would help refine future comparisons and policy strategies.

### *3. Models and Techniques Applied*

We used ETS and Prophet for ridership projections, and ARIMA for vehicle traffic forecasts. Comparative visualizations were created in Power BI to show multi-year trends side by side.

### *4. Tools and Technologies Used*

R (for time series modeling), SQL and Excel (for data preprocessing and transformation), and Power BI (for comparative visualization and presentation).

# Visuals from our Dashboard:

A screen shot of a graph

AI-generated content may be incorrect.

A graph on a yellow background

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A graph of blue lines

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.