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

Final Progress Report

Prepared By:

Aastha Kale
Ayesha Kabir
Meghana Lakshminarayana Swamy
Rajyalakshmi Nelakurthi
Venkata Sai Phanindra Namburu

Submission Date- May 8, 2025

👥 Team Member Contributions

This project was a collaborative effort that combined technical data analysis, modeling, dashboard development, and presentation design. Each team member contributed meaningfully to different components of the project. The breakdown of contributions is as follows:

🔹 Aastha Kale and Meghana Lakshminarayana Swamy

- Led the **final project report creation**, ensuring every business question was answered with data-backed insights and professional language.
- Coordinated the structure and layout of the **PowerPoint presentation** for delivery clarity and storytelling.
- Worked extensively on the **Power BI dashboard**, including creating visuals for forecasting, failure classification, and cluster-based risk mapping.
- Contributed heavily to **data integration and model building**, especially for regression and classification models using R.
- Played a key role in designing and populating the **Power BI dashboard**, ensuring alignment with the project's business questions.
- Collaborated with Aastha to clean and **merge multiple datasets**, including traffic, mechanical failure, and weather data.
- Supported the **development of clustering and ARIMA forecasting models**, and participated in interpreting the results.
- Helped write critical parts of the report, particularly the modeling rationale and tool usage sections.

🔹 Rajyalakshmi Nelakurthi

- Took charge of **PowerPoint presentation visuals**, focusing on clear and engaging storytelling for the audience.
- Assisted in summarizing findings related to **carrier-level analysis** and **borough-level risk zones**.
- Provided content and layout suggestions for the dashboard narrative and contributed to quality review before final submission.

🔹 Venkata Sai Phanindra Namburu

- Took charge of **PowerPoint presentation visuals**, focusing on clear and engaging storytelling for the audience.
- Scripted and organized the **presentation flow**, making sure each business question was addressed in a cohesive sequence.
- Supported development of the **forecasting logic** and helped narrate insights for future planning during the final presentation.

🔹 Ayesha Kabir

- Assisted in **data validation**, ensuring integrity during the cleaning and merging process.
- Reviewed dashboard visuals provided peer feedback during design refinements, and helped validate the final visual outputs.
- Supported formatting and layout of the project documentation during report compilation.

.

📘 Project Overview

The Port Authority Bus Terminal Passenger Prediction project is designed to provide the Port Authority of New York with data-driven insights to improve terminal operations, resource allocation, and long-term infrastructure planning from 2025 to 2030. By leveraging historical traffic, mechanical failure, and weather datasets, the project addresses critical questions related to forecasting, carrier usage, failure risk, and peak time periods.

This report presents structured answers to five business questions, with each response including:

1. **Key variables considered**, including the dependent variable and rationale behind its selection.
2. **Strategic recommendations** addressing missing data, additional variables to consider, and any critical data concerns.
3. **A concise summary of the models and analytical techniques used**, such as regression, classification, clustering, and time series forecasting.
4. **A brief list of tools and technologies** used, including Power BI, SQL Server, Excel, Python, and R.

🧩 1. The Most Important Variables & The Dependent Variable Used (Per Question)

Q1: Predicting Monthly MDBF (Mean Distance Between Failures)

The dependent variable selected is Monthly_MDBF, which reflects the average distance a bus travels before experiencing a mechanical failure. This is a standard reliability metric used by public transit agencies such as MTA and NJ Transit. It's crucial for assessing vehicle performance and planning maintenance. The most important independent variables include Monthly_Miles, Monthly_Road_Call_Count, and selected weather variables such as AvgMaxTemp, AvgMinTemp, AvgWind, TotalPrecipitation, and TotalSnow. These were chosen based on their logical relationship with mechanical performance and were validated using regression model outputs, where Monthly_Miles and Road_Call_Count showed high statistical significance.

Q2: Classifying Failure Risk Using Logistic Regression

For classification, the dependent variable is a derived binary label High_Failure, where a month is labeled “1” if the Monthly_MDBF is less than the dataset’s median. This allows us to classify borough-months as high or low risk. Independent variables used include Monthly_Miles, AvgWind, TotalPrecipitation, and AvgMaxTemp. These were selected based on domain knowledge (e.g., heavy rain and high usage may increase breakdowns) and tested through logistic regression. Monthly_Miles was found to have a strong negative correlation with failure risk, highlighting the operational strain's role in system reliability.

Q3: Clustering Boroughs Based on Failure and Weather Trends

In clustering, no dependent variable exists as it is an unsupervised task. The clustering algorithm (K-Means) grouped boroughs using variables like Monthly_MDBF, Monthly_Miles, Road_Call_Count, AvgWind, TotalPrecipitation, AvgMaxTemp, and a binary High_Failure tag. These inputs helped identify boroughs with similar usage patterns and mechanical issues. For instance, Bronx and Manhattan formed a "High Failure Urban Core" cluster, while Staten Island was categorized as a "Low Risk Zone". This helps in zoning and maintenance prioritization.

🛠️ 2. Recommendations: Data Issues, Additional Variables, and Considerations

During data preparation, the team encountered several challenges that required thoughtful handling. Notably, missing values in weather variables such as AWND, TMAX, and TMIN were addressed using interpolation and mean-based imputation in SQL. Fields like TotalTraffic were later removed due to consistently null values, especially in borough-level aggregations. Data standardization efforts included converting date strings into proper formats and aggregating daily weather data to the monthly level to match the MTA dataset granularity.

The team recommends that Port Authority consider integrating **passenger count data**, which is directly aligned with the business goal of predicting terminal usage. Currently, mechanical reliability is used as a proxy for service quality, but actual passenger movement data would refine predictions. Similarly, **event-based variables** such as holidays, concerts, or city-wide strikes could influence ridership and are valuable for future modeling. Lastly, tracking **fleet age**, **bus type**, or **maintenance history** would add another layer of granularity and help pinpoint root causes behind mechanical failures. These enhancements would improve model accuracy and operational insight.

📊 3. Models & Techniques Used

Predictive Models

1. **Multiple Linear Regression** – Used to predict Monthly_MDBF using bus miles, road calls, and weather. Chosen for its interpretability and strong industry usage in maintenance forecasting.
2. **Random Forest Regression** – Applied to capture non-linear relationships missed by linear models. Provided better accuracy and ranked Road_Call_Count as the top predictor.
3. **Logistic Regression** – Used for binary classification (High_Failure). This model helps identify months/boroughs that may require proactive maintenance scheduling.
4. **Decision Tree Classifier** – Provided interpretable rules (e.g., if road calls > 600, high risk). Useful for creating alert systems.
5. **K-Means Clustering** – Helped group boroughs based on failure and environment profiles. This supports targeted infrastructure upgrades.
6. **ARIMA Time Series** – Forecasted future MDBF trends by borough, especially for Manhattan, enabling proactive planning for seasonal spikes in failures.

Techniques:

- SQL-based data cleaning and integration (joins, GROUP BY, IS NULL filtering)
- Feature engineering (e.g., converting months, deriving High_Failure flags)
- Preprocessing: scaling, normalization, removing nulls
- Model performance metrics: R^2, RMSE, MAPE, confusion matrix, AIC/BIC
- Visualization: Power BI dashboards, cluster plots, ARIMA time series plots

---

💻 4. Tools Used

We utilized a diverse toolset to ensure accuracy, scalability, and ease of interpretation:

- **R Programming**: Main modeling tool (caret, forecast, randomForest, rpart). Used for regression, classification, clustering, and time series forecasting.
- **Python**: Employed early for exploratory data analysis (EDA), visualizations, and model prototyping using Pandas, Matplotlib, and Scikit-learn.
- **SQL Server Management Studio (SSMS)**: Used to clean and integrate datasets via SQL queries and views. Enabled joining weather, mechanical, and traffic datasets.
- **Power BI**: Built interactive dashboards to visually interpret MDBF trends, clusters, and forecasts for stakeholders.
- **Excel**: Used for early-stage data exploration and for capturing quick summary statistics and charts.

🧠 Final Summary

The Port Authority Bus Terminal Passenger Prediction project presents a comprehensive, data-driven framework to support decision-making for infrastructure and operational planning from 2025 to 2030. By integrating historical mechanical failure data, weather conditions, and traffic records, we were able to identify critical failure patterns, high-risk regions, and key operational cycles impacting service reliability.

Our models — ranging from linear regression to random forest, logistic classification, clustering, and ARIMA forecasting — allowed us to address all five business questions posed by the Port Authority. These models not only revealed strong

predictors such as Monthly Road Call Count and Total Miles Driven, but also highlighted seasonal patterns and geographic disparities in service reliability across New York City boroughs.

The Power BI dashboard, built on a thoroughly cleaned and engineered dataset, delivers intuitive, interactive insights across all focus areas — including peak usage months, risk zones, carrier-wise load distribution, and future reliability trends. It is supported by a reproducible analytics pipeline using R, Python, SQL, Excel, and Power BI, ensuring flexibility for future enhancements.

To maximize the impact of this work, we recommend bridging the data gap from 2017–2019, integrating real-time passenger counters, and tracking exogenous factors such as events or fleet upgrades. With these in place, the Port Authority will be well-equipped to ensure high-quality, reliable service and scalable operations through the next decade.

This report, together with the attached dashboard and dataset, provides the Port Authority with both strategic foresight and operational tools to elevate transit planning in New York City.