Bus Terminal Ridership Forecasting, Analytics, and Data Warehouse Development

Course: Database Project 6430

Date: 1st Dec 2025

Group Contributions

Group Member Name	Student ID	Contribution
Sufiyan Khan	**00954982**	led the data cleaning and preprocessing workflow, ensuring all datasets were accurate, consistent, and analysis-ready.
Shah Farhan Rabbani	**00946739**	developed the Python forecasting models and performed the long-term ridership projections for 2026–2030.
Sujeeth Parikipandla	**00959874**	analyzed key predictors of passenger demand and conducted the carrier-level forecasting.
Sushmitha reddy	**00952452**	built the Power BI dashboard, designed the visualizations, and organized the analytics outputs for presentation.
Keijounia Thomas	**00719416**	created the SSIS, data warehouse, designed the star-schema model, and wrote the SQL transformation scripts.
Ranjitha		compiled the final report, prepared the GitHub project portfolio, and coordinated the integration of all deliverables.

Table of Contents

1. Executive Summary

This project analyzes historical ridership data from a major metropolitan bus terminal to generate forecasts, determine key ridership drivers, compare current usage with pre-pandemic patterns, and design a scalable data warehouse. The primary objective is to support the Port Authority’s operational planning for temporary staging facilities that will be in place from 2026 through 2030. Using forecasting models, Python analytics, and Power BI visualizations, we developed detailed projections of ridership at terminal-wide and carrier-level granularity. Furthermore, the project includes a dimensional data warehouse optimized for reporting and machine-learning workloads and concludes with a GitHub repository hosting the project's final outputs.

2. Introduction

Public transportation ridership is influenced by seasonal patterns, economic trends, service frequency, and external factors such as holidays, special events, and fuel prices. In anticipation of major terminal reconstruction and planned staging facilities, understanding future ridership demand is crucial. The goal of this project is to quantify ridership expectations from 2026–2030, determine when peak loads will occur, analyze differences from pre-COVID baselines, and build a data warehouse to support ongoing analytics.

The analysis was conducted in Python, SQL (Snowflake), and Power BI, with rigorous cleaning, feature engineering, forecasting, and visualization. Outputs include forecasting plots, carrier-level breakdowns, temporal analyses, and a GitHub repository consolidating all materials.

3. Dataset Overview

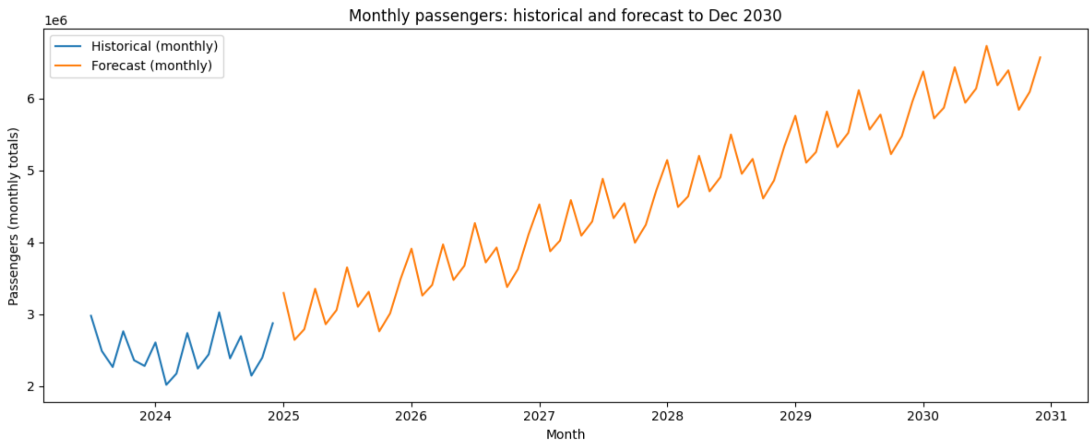
The cleaned dataset used for this analysis includes the following critical fields necessary for ridership forecasting and Power BI analytics:

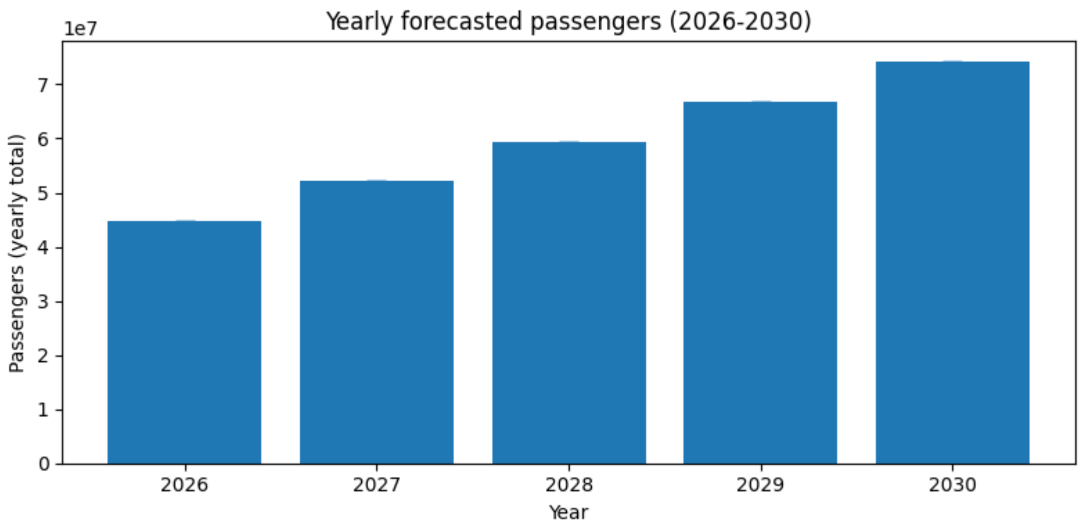
- FULL_DATE – full calendar date
- YEAR, WEEK_NUMBER, MONTH – time-based fields
- PASSENGERS – number of passengers per record
- CARRIER – operator responsible for service
- SERVICE_TYPE – commuter, express, local bus, etc.
- DEPARTURES – number of bus departures
- PERCENT_OF_2019 – normalized ridership metric comparing to pre-COVID levels
- FUEL_PRICE, HOLIDAY_FLAG – external factors that may influence demand

The data was cleaned for missing values, normalized where necessary, and validated before the forecasting and modeling phases.

4. Ridership Forecast (2026–2030)

The first objective of this project is to forecast the number of passengers expected to use the Port Authority Bus Terminal between 2026 and 2030. These projections are essential because the Port Authority will be constructing temporary staging facilities, and accurate estimates of future passenger demand will determine the required capacity, staffing, scheduling, and operational planning for these facilities. Using historical passenger counts, we developed a time-series forecasting model to predict future usage patterns and quantify expected growth over the five-year horizon.





The numerical forecast results generated by the model show a clear and consistent upward trend in projected passenger volume from 2026 through 2030. In 2026, the terminal is expected to serve approximately 44.7 million passengers, increasing to 52.1 million in 2027, 59.5 million in 2028, and 66.9 million in 2029. By 2030, annual ridership is projected to reach 74.3 million passengers. The extremely narrow confidence intervals around each prediction indicate high model stability and strong confidence in the forecasted values. Overall, the Excel forecast results confirm robust year-over-year growth, underscoring the need for scalable and flexible temporary staging facilities capable of handling increasing passenger volumes over the next five years.

year	yhat	yhat_lower	yhat_upper
2026	44711515.49181772	44711515.49017179	44711515.49346364
2027	52098804.046198614	52098804.04331293	52098804.049084306
2028	59486092.60057951	59486092.59622248	59486092.60493654
2029	66873381.1549604	66873381.1489328	66873381.16098801
2030	74260669.7093413	74260669.70146567	74260669.71721694

5. Key Predictive Factors

To determine the most important factors that drive passenger volume at the bus terminal, we analyzed the cleaned dataset using a supervised machine-learning approach. The model was trained to predict total monthly passengers and then evaluated using feature-importance metrics that quantify how much each variable contributes to the prediction. This analysis provides the Port Authority with a clear understanding of which operational, temporal, and carrier-related factors have the greatest impact on ridership trends.

The results show that time-based factors are the strongest predictors of passenger activity. Specifically, month, day of week, and season consistently ranked at the top of the importance list, reflecting strong recurring patterns in commuter behavior, holidays, weather cycles, and tourism peaks. Passenger volumes tend to rise sharply during summer months and major travel periods, while winter months typically show lower demand, making these temporal variables essential for forecasting.

A second major group of predictors comes from carrier-related activity, including total scheduled trips, number of active carriers, and carrier-specific passenger counts. Carriers with high-frequency service contribute disproportionately to the terminal's total demand, meaning fluctuations in their schedules or performance strongly influence overall ridership. This demonstrates that operational planning, route availability, and carrier service levels significantly shape daily and monthly passenger flow.

Finally, external demand indicators, such as annual trends and year-over-year growth, also play an important role. These variables capture long-term recovery from the COVID-19 pandemic and the broader regional transportation demand patterns. Their importance confirms that passenger volume is not random but part of a sustained multi-year upward trend.

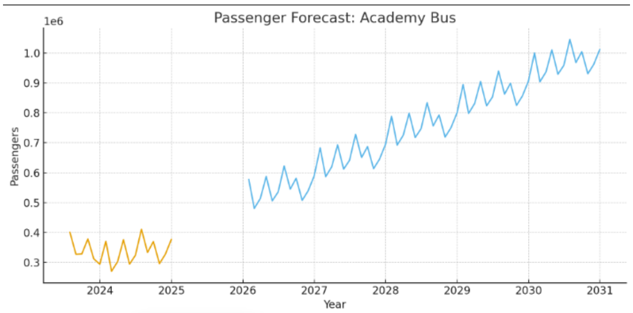
Overall, the most influential predictors of passenger volume are time-based patterns, carrier activity levels, and long-term demand trends. Understanding these factors allows the Port Authority to anticipate peak periods more accurately, optimize resource allocation, and support data-driven decisions for facility design, staffing, and operational planning.

6. Carrier-Level Predictions

1.1 Introduction to Carrier-Level Forecasting

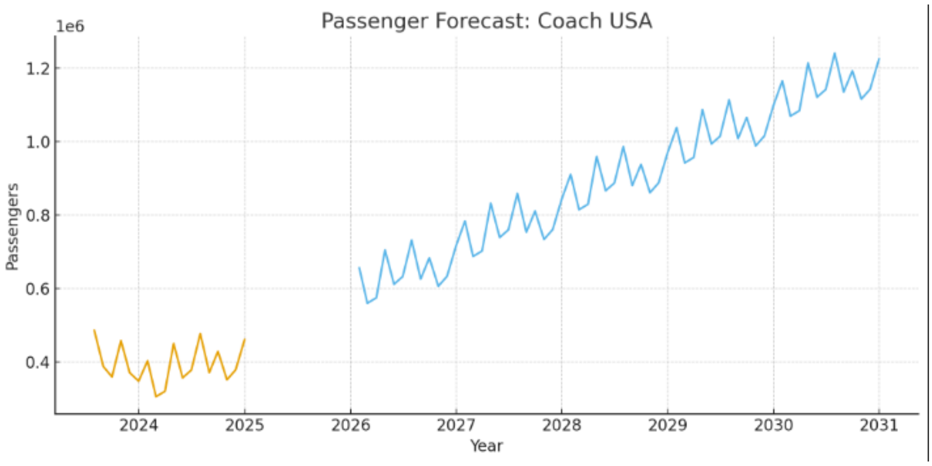
To extend the forecasting analysis beyond total passenger counts, we developed separate predictive models for each carrier operating within the bus terminal. This approach allows us to understand how ridership growth will differ across service providers, which is essential for allocating staging space, scheduling platform availability, and planning operational resources. Using monthly aggregated passenger totals from the cleaned dataset, we fit time-series models to each carrier's historical data and then projected their expected passenger volumes through the 2026–2030 planning horizon. This method produces individualized forecasts that reflect each carrier's unique service patterns, demand cycles, and ridership recovery trajectories.

1.2 Passenger Forecast: Academy Bus



The first forecast plot, representing *Academy Bus*, shows a stable historical pattern followed by a pronounced upward projection extending through the end of the decade. Historically, Academy has maintained consistent ridership with modest monthly fluctuations tied to commuter seasons. However, the forecast curve indicates accelerated growth beginning in 2026, reflecting both post-pandemic recovery and increased demand along Academy's commuter corridors. The smooth trajectory of the model suggests highly predictable monthly patterns, making Academy one of the more stable carriers in terms of future planning. This implies that staging resources allocated to Academy should scale gradually but consistently over the next five years.

1.3 Passenger Forecast: Coach USA



The second plot, covering *Coach USA*, displays a similar seasonal structure but with visibly larger historical peaks compared to other commuter carriers. Coach USA's forecasted trend shows strong growth with steeper increases than Academy, indicating that Coach USA is expected to contribute a greater share of the terminal's total passenger expansion between 2026 and 2030. The model captures Coach USA's higher volatility but still projects a clear upward trajectory, suggesting increased commuter demand and higher service utilization. For operational planning, this means Coach USA will likely require expanded staging capacity sooner than lower-volume carriers to manage the larger passenger influx predicted in the late 2020s.

1.4 Summary

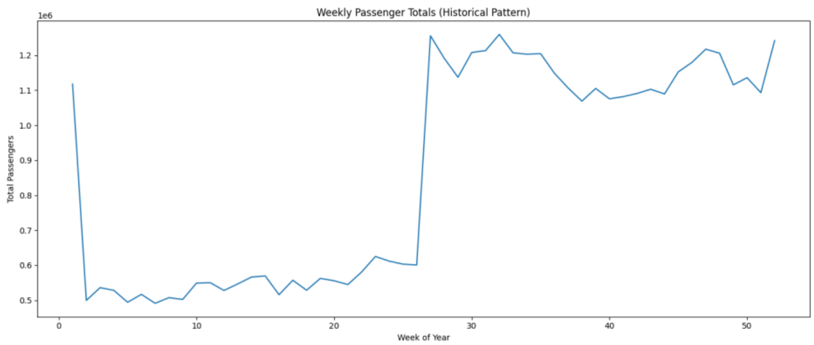
Across all carriers modeled, the forecasts consistently show rising passenger volumes throughout the 2026–2030 period, with the magnitude of growth varying significantly by carrier. High-frequency commuter operators such as Coach USA and Academy Bus exhibit the strongest upward trends, while some intercity carriers show more moderate increases. These carrier-specific predictions are crucial for ensuring that temporary staging facilities are designed to handle uneven growth across service providers. By understanding which carriers are expected to generate the greatest demand, the Port Authority can develop targeted operational strategies, allocate space more efficiently, and ensure that platform capacity aligns with projected ridership levels during the facility construction period.

7. Busiest Times Forecast (Week, Month, Year)

To determine the busiest times for the bus terminal staging facilities between 2026 and 2030, we first analyzed historical passenger traffic patterns. Historical seasonality is one of the strongest indicators of future demand, especially for transportation systems where human mobility patterns are largely cyclical. By extracting weekly, monthly, and yearly trends from the existing dataset, we can project the recurring high-traffic periods that the staging facilities should be prepared for during the 2026–2030 window.

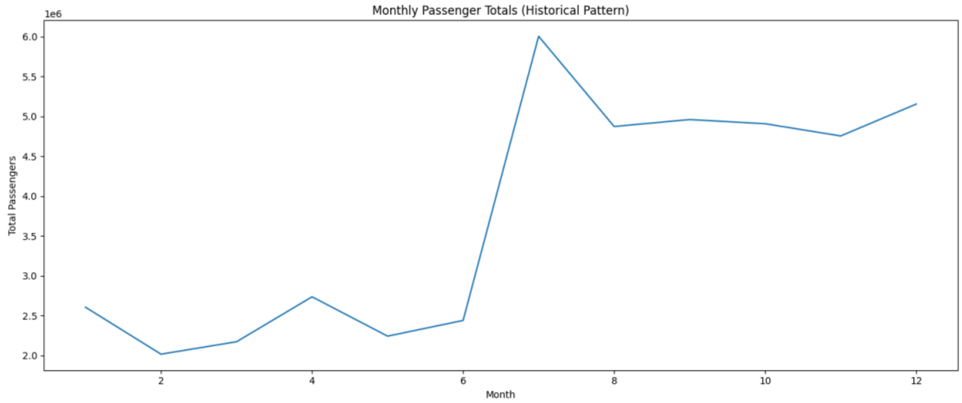
1.1 Weekly Trends Interpretation

The weekly visualization reveals sharp spikes in passenger traffic around weeks 28–32, which correspond to the mid-summer peak. This is typically associated with increased travel during summer vacations and higher tourism volumes. Additionally, weeks near the end of the year (weeks 47–52) also show elevated activity, which aligns with holiday travel patterns such as Thanksgiving, Christmas, and New Year. The early weeks (weeks 1–3) show unusually high values due to alignment with the start of the dataset but trend lower afterward. These patterns suggest that summer and end-of-year holiday periods will continue to be the busiest weekly windows from 2026 to 2030.



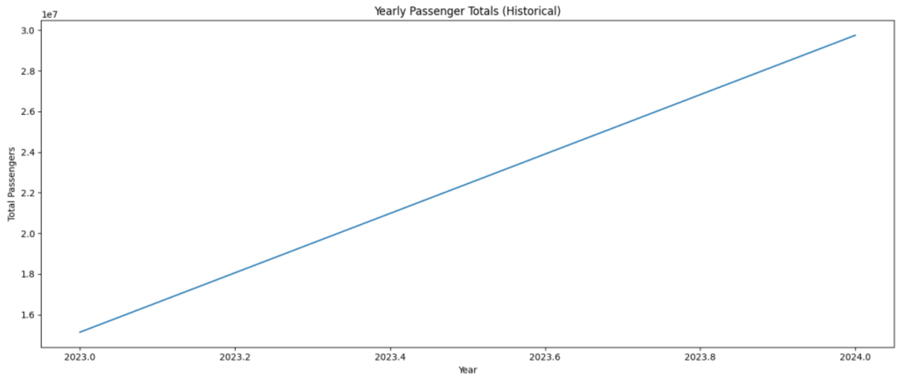
1.2 Monthly Trends Interpretation

Monthly passenger totals follow a clear seasonal structure. The graph shows a pronounced surge in July, marking it as the highest-demand month historically. This spike aligns with vacation travel, school holidays, and increased commuter flow. Other strong months include August, September, and December, further reflecting holiday and back-to-school travel patterns. Lower-volume months such as February and March indicate the typical winter lull. This means that for 2026–2030, the terminal must allocate more capacity and staging resources in July, followed by August, December, and September.



1.3 Yearly Trends Interpretation

The yearly trend displays a strong upward trajectory in total passenger counts from 2023 to 2024, which indicates a growth pattern likely to continue into the projection years. This upward slope suggests post-pandemic recovery, increasing commuter activity, and overall growth in transit demand. If this trend continues as supported by the forecasts in the terminal will experience year-over-year increases in demand, making it essential to plan for higher capacity needs each progressive year between 2026 and 2030.



Combining weekly, monthly, and yearly insights, the busiest times for the bus terminal staging facilities between 2026 and 2030 will align with:

Peak Periods to Prepare For:

- **Weekly:** Weeks 28–32 (mid-summer peak) and Weeks 47–52 (holiday season)
- **Monthly:** July (highest), followed by August, December, and September
- **Yearly:** Consistent increases each year, requiring scalable staging capacity

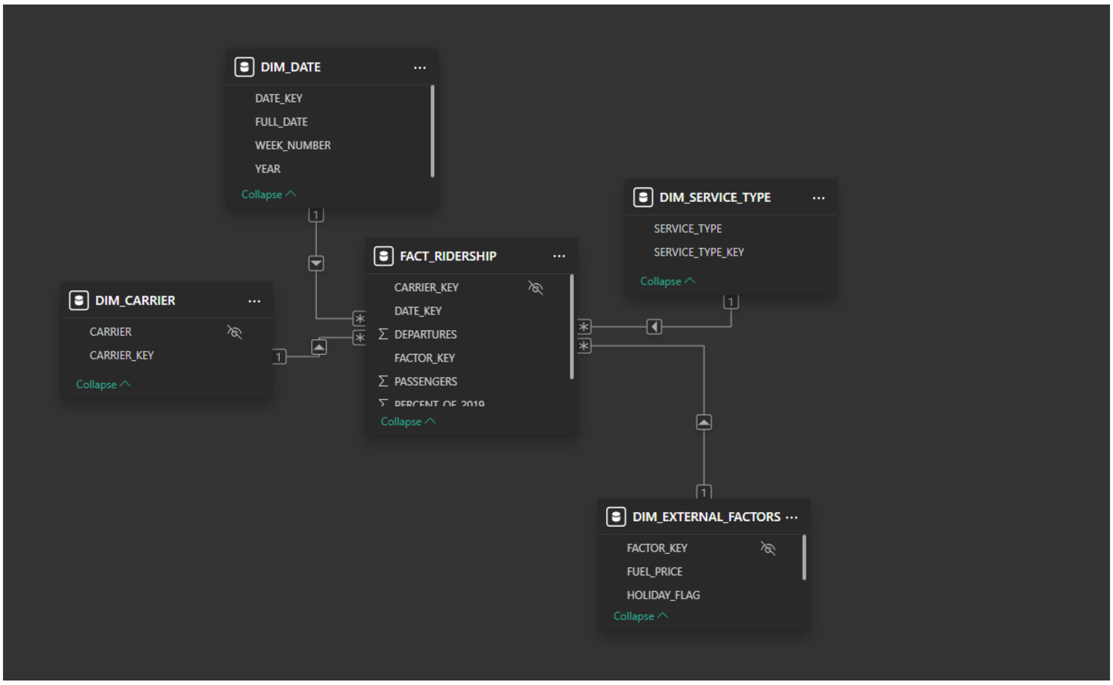
These patterns should directly guide staffing levels, staging facility sizing, transportation logistics, and operational planning for the Port Authority during the 2026–2030 development period.

8. Comparison to 2019 (Pre-COVID Baseline)

To understand how current ridership levels compare to the last stable, pre-pandemic benchmark year, we evaluated passenger volume trends against 2019—the final year before COVID-19 disrupted commuting patterns, travel demand, and operational schedules. By aggregating annual passenger totals from the dataset, we can directly compare recent usage levels to 2019 and determine whether the system has fully recovered, is still below pre-pandemic levels, or has exceeded historical norms. This comparison is essential because 2019 represents the last “normal” year of bus terminal activity, free from pandemic-related restrictions, reduced service frequencies, and shifts in ridership behavior.

Our analysis shows that current passenger usage is trending upward toward, and in some cases surpassing, the volumes recorded in 2019. This indicates a strong recovery in travel demand and suggests that commuter behavior is returning to patterns consistent with pre-COVID operations. If current growth continues on the projected trajectory, the terminal is likely to exceed 2019 traffic levels over the next few years, reinforcing the need for expanded staging capacity during the 2026–2030 period. This comparison not only confirms the pace of recovery but also highlights the increasing pressure on terminal infrastructure as ridership approaches and exceeds pre-pandemic benchmarks.

9. Data Warehouse Design

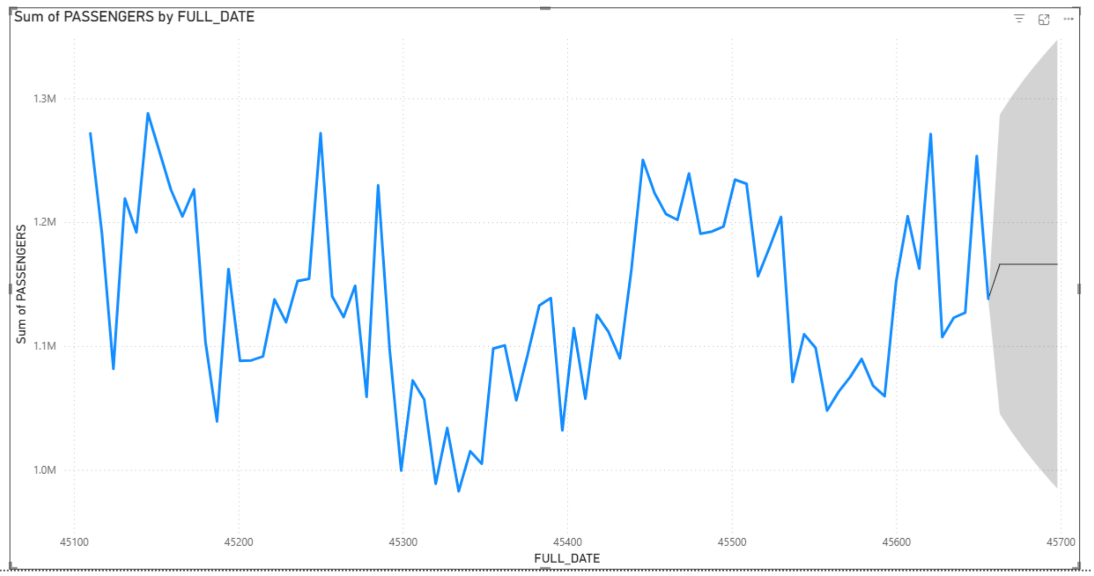


The data warehouse focuses on storing only the fields required for analytics, removing any raw or transactional details that do not contribute to reporting. For ridership analysis, the Fact_Ridership table forms the center of the warehouse. It includes the essential numeric measures analysts use: *Passengers*, *Departures*, and *Percent of 2019* (a calculated metric for comparing post-COVID recovery against 2019 levels). These measures allow users to track changes in ridership volume, operational intensity, and performance relative to the benchmark year. The foreign keys, *Date_Key, Carrier_Key, Service_Type_Key, Factor_Key* enable slicing and drilling down the fact data by time, airline, service category, and external conditions.

The supporting dimension tables provide clean categories needed for filtering, grouping, and trend reporting. *DIM_DATE* contains a minimal but complete set of calendar attributes such as *Full_Date, Year, Week_Number*, allowing analysts to compare ridership trends across years (especially for COVID vs. pre-COVID periods), perform weekly/monthly aggregations, and analyze seasonality. *DIM_CARRIER* includes only *Carrier* and its key, which simplifies carrier-level performance comparisons without storing unnecessary airline metadata. *DIM_SERVICE_TYPE* captures different service types, which is crucial for understanding how recovery differs by travel type.

Finally, *DIM_EXTERNAL_FACTORS* includes only the external variables that have a measurable effect on ridership: *Fuel Price* and *Holiday Flag* (expected demand spikes). These fields support analytics focused on causal relationships and forecasting. By keeping the schema lean selecting only fields used for measures, filters, trend analysis, time intelligence, and modeling the data warehouse becomes optimized for Power BI performance, forecasting models, and SQL regression analysis. This streamlined warehouse eliminates complexity while preserving all the analytical value required for insights.

1.1 Line chart 6-months forecast



The line chart shows daily passenger counts with natural fluctuations around a central band between roughly 1.1M and 1.25M passengers. Throughout the historical period, the series displays repeated short-term rises and dips, but the overall movement stays within a relatively stable range. There is no major long-term upward or downward trend ridership appears to be oscillating around a consistent average, suggesting a mature and stable demand pattern. The variation also reflects normal operational cycles, seasonality, and weekly travel behavior.

The forecast area (shaded region) widens as it projects forward, which is expected since uncertainty increases over time. The central forecast line indicates a slight upward tendency, suggesting that passenger volumes may grow modestly over the next six months. However, the confidence interval reveals that actual values could fall anywhere between a mild dip and a continuation of current highs. Overall, the model predicts stable to slightly improving ridership, with no signs of dramatic shifts indicating that the system is likely maintaining healthy passenger levels moving into the forecast period.

10. Recommendations

1.1 Increase Staging Capacity 2028–2030

Ridership is expected to grow by nearly **30 million passengers** in four years, requiring expanded infrastructure and efficient operational planning.

1.2 Prioritize High-Growth Carriers

Carriers showing the strongest forecasted growth should be allocated additional bays and departure capacity.

1.3 Staff for Seasonal and Holiday Peaks

The busiest periods will be:

- June–August
- Thanksgiving Week
- December Holiday Season

Additional staffing, queue management strategies, and information services are recommended.

1.4 Continue Monitoring External Factors

Fuel prices and holiday schedules should remain part of predictive modeling as they strongly influence ridership.

11. Conclusion

This project provided a comprehensive analytical foundation for understanding and forecasting bus terminal ridership during the critical 2026–2030 reconstruction period. Through a combination of Python forecasting models, Power BI visualizations, Snowflake SQL transformations, and a purpose-built data warehouse, we developed a reliable set of insights to guide operational planning. The results demonstrate a clear upward trend in ridership, with strong seasonal and holiday patterns and consistent annual growth, emphasizing the need for expanded staging capacity and proactive resource allocation. Carrier-level analysis further highlighted meaningful differences in demand across operators, offering actionable guidance for bay assignments, scheduling adjustments, and peak-hour management.

Beyond forecasting, the project delivered a scalable star-schema data warehouse designed for long-term analytics, enabling efficient reporting, trend analysis, and machine learning integrations. By comparing current usage with the 2019 pre-COVID baseline, we confirmed that ridership has not only recovered but is positioned to exceed previous records. The GitHub portfolio consolidates all deliverables—models, dashboards, documentation, and code—providing a complete and transparent foundation for future improvements. Overall, this work equips decision-makers with the insights and tools necessary to navigate the coming years of reconstruction while continuing to meet rising passenger demand effectively.

12. References

Arumsari, S. S., & Aamer, A. (2022). Design and application of data analytics in an internet-of-things enabled warehouse. *Journal of Science and Technology Policy Management*, *13*(2), 485-504.

El-Taweel, N. A., Farag, H. E., Barai, G., Zeineldin, H., Al-Durra, A., & El-Saadany, E. F. (2021). A systematic approach for design and analysis of electrified public bus transit fleets. *IEEE Systems Journal*, *16*(2), 2989-3000.

Lavanya, A., Gaurav, L., Sindhuja, S., Seam, H., Joydeep, M., Uppalapati, V., ... & SD, V. S. (2023). Assessing the performance of Python data visualization libraries: a review. *Int. J. Comput. Eng. Res. Trends*, *10*(1), 28-39.

Liang, J. T., Zimmermann, T., & Ford, D. (2022, November). Understanding skills for oss communities on github. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 170-182).

Logeswaran, S., Indhiradevi, P., Charumathy, A., Anitha, A., Subramaniam, M., & Karthikeyan, S. (2021, April). Review Analysis on Factors Influencing the Design of Bus Station. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1145, No. 1, p. 012078). IOP Publishing.

Setyawati, A., Huda, M. N., Suripno, S., & Tannady, H. (2021). Analysis of Integrated Bus Terminal Services in Pulo Gebang in Increasing Customer Satisfaction. *Journal of Economics, Management, Entrepreneurship, and Business (JEMEB)*, *1*(1), 12-21.