Project Progress Report 2

Team members:

Giacomo Bizzotto, Rishitha Thatipally, Irfan Shaik, Errolla Vivek, Shraddha Shakya, Sampath Sai Raghav Allaboyina.

1 ) Overview of Data Sources and Structure

Our project leverages an extensive database infrastructure composed of three distinct Microsoft Access databases, each providing critical information for our analysis.

The primary dataset, “TBT\_Traffic\_Database\_PA-1” (1.85 GB), serves as the cornerstone of our study, containing comprehensive traffic-related data essential for forecasting bus passenger flows. This database includes eight different tables, over 40 pre-defined queries, and one table with more than 5 million records, making it the largest and most valuable resource for our project.

The second dataset, “Unit\_571\_Database” (242 MB), complements the primary database with 23 tables providing additional operational and infrastructure data.

Lastly, the “Weather Database-1” (864 KB) offers one table with weather-related information, which will be incorporated into our analysis to assess the impact of environmental conditions on passenger traffic. The integration of these diverse data sources will enable a comprehensive and accurate forecast of bus passenger demand.

2) Data Preparation, Integration, and Cleaning Strategy

To ensure the quality and consistency of our data, we adopted a structured multi-step data preparation process using Python, Excel, and R. These tools enabled us to clean, filter, and transform raw data into a standardized format. The cleaning phase involved identifying and handling missing values, removing duplicates, and correcting inconsistencies in data entries. For missing data, we applied mean/median imputation for numerical attributes, replacing missing values with the mean or median depending on whether the distribution was skewed. Additionally, we used linear interpolation to estimate missing values in time-series data by computing new values based on the trend between known data points. Once the datasets were cleaned, we imported them into SQL Server Management Studio (SSMS) to create a well-structured relational database. First, we defined the table schema, ensuring appropriate data types and constraints were assigned to each column. We then used the BULK INSERT command to efficiently load large CSV files into SSMS. For example, the traffic data was imported using the following command:

BULK INSERT Traffic\_Data

FROM 'C:\Users\giaco\OneDrive\Desktop\UNH\COURSES\Database Management for Business Analytics\Project Group - Port Authority\DATABASES FOR THE PROJECT\CSV\Traffic\_Data.csv'

WITH (

FIRSTROW = 1, -- Start from the first row

FIELDTERMINATOR = ',', -- Define CSV delimiter

ROWTERMINATOR = '\n', -- Line break

TABLOCK );

This method allowed us to efficiently handle large datasets while maintaining data integrity. After importing, we established relationships among different tables by defining primary and foreign keys, enforcing referential integrity, and designing a comprehensive relational schema linking traffic, operational, and weather data. To enhance performance and enable more efficient querying, we created intermediate tables to store pre-processed data, facilitating further transformations and reducing computational overhead when executing complex queries. The combination of these technologies and methodologies ensured the creation of a unified, high-quality dataset ready for analysis, ultimately improving both the accuracy and efficiency of our forecasting models.

3) External Data Utilization Plan

Currently, we do not plan to use any external datasets beyond the ones provided by the company. The data available from the three databases appears to be comprehensive and sufficient to meet the requirements of our analysis. However, should any gaps or additional requirements arise during the project, we will reassess the need for external data sources.

4) Establishing Relationships and Joining Datasets

After performing data synthesis and joins, the resulting dataset is expected to generate a large number of rows. To ensure the data remains manageable and optimized for analysis, we will extensively use the GROUP BY function in SQL Server Management Studio. This approach will allow us to aggregate data into more compact and meaningful categories, such as summarizing passenger counts by time intervals, bus routes, or weather conditions. By grouping the data in this way, we can reduce redundancy, improve query performance, and maintain a dataset that is both easier to handle and more insightful for forecasting purposes.

5) Final Dataset Presentation and Manageability

To support our forecasting goals, we have developed three key datasets by executing SQL queries on our database to extract and transform the necessary information. Each of these datasets serves a specific purpose in achieving our main forecasting objectives. The first dataset projects the estimated number of bus passengers for each terminal and facility from 2025 to 2030, providing long-term insights into demand trends and helping with capacity planning and resource allocation. The second dataset focuses on carrier-specific projections, breaking down passenger distribution across different service providers to analyze trends among bus operators and understand how demand varies by carrier. This dataset is crucial for optimizing service schedules and balancing load distribution. Lastly, the third dataset identifies peak travel times by analyzing historical data, highlighting the busiest days, months, and weeks for each facility. This information is essential for predicting traffic surges and enhancing operational efficiency. Together, these datasets enable a comprehensive forecasting approach, ensuring that our models effectively address future passenger demand, carrier-specific trends, and peak traffic periods.

A table with numbers and letters

AI-generated content may be incorrect.

Figure 1: Database1, to forecast bus passengers from 2025 to 2030 in each facility

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Dataset 2, to project the prediction results by the individual carrier

A screenshot of a graph

AI-generated content may be incorrect.

Figure 3: Dataset 3, to forecast the busiest year, month and week for each facility