**Project Progress Report**

**Q1️ Explain to the class the available data sources for your project.**

* We are working with three Microsoft Access (.accdb) databases to analyze and forecast bus terminal passenger trends. The following databases **TBT Traffic Database,** **Unit 571\_Database**, and **Weather Database** contains structured data related to historical passenger counts, bus departures, traffic patterns, transit usage, and weather conditions.
* **TBT Traffic Database** consists of 8 datasets, 43 queries, 34 reports, and 31 forms, making it a crucial source for analyzing *traffic flow*, *congestion trends,* and *transit operations*. **Unit 571 Database** focuses on *bus carrier departures, traffic counts across various Port Authority facilities,* and *departure schedules* for both *terminals*. **Weather Database** provides climate-related insights to assess how *weather conditions* impact transit ridership and operational efficiency.
* To effectively utilize these databases, we will standardize *data formats, handle missing values, remove duplicates, and integrate key fields for analysis*. Relationships among the databases will be established using date-based, carrier-based, and facility-based joins to ensure seamless data integration. Additionally, strategies such as filtering relevant years, aggregating data, indexing for performance, and splitting datasets into manageable sections will be applied to optimize processing efficiency. This structured approach will enable us to generate reliable forecasts for passenger volumes and transit operations for *2025-2030*.

**Q2️ How do you plan to prepare, integrate, and clean your data from the available sources?**

* To ensure data accuracy and consistency, we will standardize date fields to the YYYY-MM-DD format and validate numeric fields such as passenger counts and traffic volumes. *Missing data* will be handled by removing columns with excessive gaps, using *linear interpolation* for time-series data, and imputing categorical values based on frequency.
* *Duplicate records* will be *identified and removed*, and filtering will be applied to retain only records from 2015 to 2024, ensuring a strong historical foundation for forecasting 2025-2030 trends. Irrelevant fields that do not contribute to forecasting will be excluded for better performance.
* Additionally, *outlier detection* techniques will be used to identify anomalies, and referential integrity checks will ensure proper relationships among datasets. Necessary transformations will align *data formats*, and *data normalization* will be applied where required for consistency. These measures will create a *clean, structured dataset*, supporting *accurate forecasting and effective decision-making*

*.*

**Q3️ How do you plan to establish relationships among the various datasets? How do you plan to join the datasets?**

Since we have multiple data sources, it is essential to establish key relationships before performing joins. The primary relationships between datasets are categorized based on different factors.

* Date based joins link datasets such as **MBT NJT NYC Trips Passengers**, **MBT Passenger Departures**, **Reconciled Traffic**, and **Tbl Weather** to correlate passenger counts with weather and traffic conditions. Carrier based joins connect **MBT Bus Departures**, **Bus Carrier Departure,** and **MBT NJT NYC Trips Buses**, enabling an analysis of ridership per bus operator. Additionally, facility-based joins establish relationships between **Reconciled Traffic**, **MTA TBTA Hourly Traffic**, and **Traffic Data**, allowing an assessment of congestion's impact on bus passenger numbers.
* To ensure consistency in data integration, the *primary key* for *joins* is the *date*, while secondary keys include the bus operator and transit facility. The join type primarily used is left joins, as this approach retains full records from main datasets, particularly passenger data, ensuring comprehensive analysis without loss of critical information.

**Q4️ Do you plan to use any external datasets apart from those provided by the company?**

* Currently, no external datasets are planned. The datasets provided to us are sufficient for future analysis. However, we may refer to external information available on the internet.

**Q5️ How do you plan to manage the large dataset after joins?**

* Since joining multiple datasets will generate a high number of rows, we will implement *data optimization* techniques using ***SQL, Python, and Excel*** to enhance efficiency and performance. In ***SQL***, we will focus on **filtering data to relevant years**, retaining only records from **2015 to 2024** for analysis. **Aggregation techniques** will be applied using both ***SQL* queries** and ***Python*** to convert **hourly data into daily summaries** and **daily data into monthly aggregates** for forecasting purposes.
* To optimize performance**, indexing strategies** will be implemented in SQL, including **date indexing** for faster *time-series analysis* and **operator indexing** to improve carrier-specific lookups. **Python** will be used for ***data preprocessing, cleaning, and handling large datasets***, while **Excel** will assist in **validating and summarizing key insights** before final analysis. Additionally, to streamline data handling, we will **split the dataset into two separate data frames** a **passenger-based dataset** containing *date, passenger counts,* and *carrier details*, and a ***traffic-based dataset*** including *traffic counts* and *weather data*. These techniques will ensure efficient data management and enhance analytical accuracy.