**Port Authority of New York & New Jersey Data Analytics Project - Report 2**

**Group 2**

**Introduction**

The Port Authority of New York and New Jersey (PANYNJ) manages critical transportation infrastructure, including tunnels and bridges. Our project focuses on analyzing traffic patterns, toll violations, and external factors affecting congestion. This report outlines our approach to data preparation, integration, and analysis.

**Data Sources Available**

We have access to four key datasets provided by the Port Authority:

1. **Traffic Data:** Contains timestamp, location, vehicle count, and average speed, helping analyze traffic trends.
2. **Toll Violation Data:** Includes timestamps, vehicle type, toll amount, and violation status, used to assess unpaid tolls.
3. **Weather Data:** Features date, temperature, precipitation, and wind speed, essential for analyzing weather impact.
4. **Event Data:** Lists event type, location, and estimated attendance, helping identify congestion causes.

Additionally, we may use external datasets such as **historical traffic records from the NYC Department of Transportation** and **detailed weather data from OpenWeatherMap** for improved accuracy.

**Data Preparation, Integration, and Cleaning**

To ensure data quality, we will first handle missing values by inputting numerical values with averages and replacing unknown categorical values. We will remove duplicates and standardize all date formats to MM-DD-YYYY. Outliers will be identified using statistical methods like **Interquartile Range (IQR) and Z-score filtering** to prevent distorted analysis.

For integration, we will establish relationships using **timestamps and locations as primary keys**, while vehicle type and event type will serve as foreign keys. We plan to **join datasets using SQL queries**, such as:

* **INNER JOIN:** Merging traffic and weather data based on timestamps and locations.
* **LEFT JOIN:** Connecting toll violation data while preserving all traffic records.
* **RIGHT JOIN:** Incorporating event data to ensure all events are accounted for.

**Managing Large Data Volumes**

Since merging multiple datasets can generate a high volume of rows, we will manage this by **sampling**, summarizing traffic records by hour or day, and indexing key fields for faster queries. These techniques will ensure the dataset remains manageable without losing essential insights.

**Final Dataset for Analysis**

The final cleaned dataset will include fields such as timestamp, location, traffic volume, speed, toll violations, weather conditions, and event type. This structured dataset will be optimized for analysis and visualization.

To maintain data quality, we will conduct final checks to confirm **no missing values**, ensure proper standardization, and verify logical consistency across records.

**Tools and Technologies Used**

For data processing, we will use **SQL for queries, Python (Pandas, NumPy, Scikit-learn) for cleaning and statistical analysis, Tableau/Power BI for visualization, and Azure AutoML for predictive modeling**.

**Conclusion**

Our approach ensures a **structured, high-quality dataset** ready for insightful analysis. We have a total of 50,000 rows and 11 columns. By integrating multiple data sources and using advanced analytical tools, we aimed to uncover patterns that will support better traffic management and policy decisions.