**Project Report - Group 4**

**Port Authority of New York and NJ Data Analytics Project**

**BANL 6430-08- Database Mgmt. Business Analytics**

1. **Business Goals of the Project**

The Port Authority of NY & NJ is responsible for managing transportation infrastructure, including bus terminals, bridges, tunnels, and staging facilities. As transportation demand shifts due to post-COVID recovery, the agency needs data-driven insights to optimize its operations.

This project aims to analyze, predict, and compare passenger traffic trends to support better decision-making for temporary staging facilities, congestion management, and infrastructure planning.

**Additionally, we aim to:**

Identify key factors influencing passenger volume.

1.Predict busiest times for staging facilities (weekly, monthly, yearly).

2.Forecast passenger volume per carrier to optimize scheduling.

3.Compare current trends (2024) to pre-COVID trends (2019).

1. **Tools We Plan to Use** - Python, SQL and Power BI

Purpose of tools in the Project

**SQL** - Used for extracting, cleaning, and preprocessing passenger, traffic, and weather data. It enables efficient filtering, aggregation, and structuring of data from relational databases before advanced analytics.

**Python** – Chosen for its robust data analysis and machine learning capabilities. It will be used for forecasting, regression analysis, and clustering while also processing SQL-extracted data before modeling.

**Power BI** - Utilized for creating interactive dashboards to visualize forecasted trends, peak travel times, and comparisons between 2019 and 2024. It allows stakeholders to explore insights dynamically.

1. **Algorithms & Models we will be using**

|  |  |  |
| --- | --- | --- |
| **Model / Algorithm** | **Python Libraries** | **Purpose** |
| **ARIMA & Prophet** | statsmodels, fbprophet | Predict passenger volume from 2025 to 2030 to aid staging facility planning. |
| **Multiple Linear Regression** | scikit-learn, pandas, numpy | Identify key factors (weather, traffic, holidays) that impact passenger volume. |
| **K – Means Clustering** | scikit-learn, pandas | Categorize travel patterns into high, medium, and low congestion periods. |
| **Decision Trees** | scikit-learn | Determine whether passenger trends increased, decreased, or remained stable compared to 2019. |

1. **Rationale for Choosing These Tools and Models**

The selected tools and models ensure **efficient data processing, accurate forecasting, and actionable insights** to optimize transportation operations for the Port Authority of NY & NJ.

SQL is used for **data extraction, cleaning, and preprocessing**, efficiently handling passenger, traffic, and weather data. Python, with its powerful **machine learning and statistical libraries**, enables forecasting, regression, and clustering for in-depth trend analysis.

Power BI provides **interactive visualizations**, making insights accessible to stakeholders.

For forecasting passenger volume (2025-2030), **ARIMA and Prophet** are chosen. ARIMA captures **long-term trends and seasonality**, while Prophet accounts for **external events and holidays**, making predictions more precise. **Multiple Linear Regression** identifies key factors like **weather, traffic, and public holidays**, aiding in schedule optimization.

To analyze peak travel times, **K-Means Clustering** segments data into **high, medium, and low-traffic periods**, supporting congestion management. **Decision Trees** classify conditions that trigger passenger surges, further refining scheduling and resource allocation.

For comparing **2019 vs. 2024 passenger trends**, **regression and classification models** detect **post-COVID shifts in travel behavior**, ensuring informed long-term infrastructure planning.

This strategic combination of tools and models ensures a **data-driven, scalable, and interpretable approach**, helping the Port Authority enhance efficiency, manage congestion, and optimize staging facility usage.