Project Report - Group 5

Port Authority of New York and NJ Data Analytics Project

BANL 6900-01 - Business Analytics Capstone

1. **Available Data Sources**

Our project analyzes traffic, toll violations, congestion, and forecasting for major Port Authority bridges and tunnels. We are building five goal-specific analytical datasets, each developed from a common foundation of Port Authority data, merged with weather, holiday, and facility mobility data.

Primary Datasets Provided by the Port Authority

1. **Traffic by Facility Datasets**
    - Files: Traffic-Record (original dataset which has 8 facility) we have divided it into 8 different facilities to reduce the loss of rows because csv can handle only 1,048,576 rows.
    - Records: In each facility we have around 80,000k - 100,000k rows per facility.
    - Fields Used: DATE, Facility_Name, TOTAL, EZPASS, CASH, VIOLATION, Autos, Small_T, Large_T, Buses
    - Purpose: Core source for traffic volume, payment type, and vehicle composition.
2. **Weather Dataset**
    - File: Tbl_Weather_clean.csv
    - Records: 5,263
    - Fields Used: DATE, PRCP, SNOW, TMAX, TMIN, AWND
    - Purpose: Adds environmental conditions affecting traffic demand.

- **Total files:** We used 11 CSV files - 8 traffic files (one per facility) and 3 supporting files (Weather, Holidays and Toll Prices).
- **Connection logic:** All datasets are linked through the DATE field.
- **File format:** All files are CSVs with headers, comma-delimited, and contain ~90k-110k rows per facility.

2. **Data Preparation, Integration, and Cleaning**

We started with the raw file **"All Recorded Traffic.txt"** provided by the Port Authority, which contained data for all bridges and tunnels.

- **Using Python we did:**

- All integration and cleaning tasks were done in Python using pandas and NumPy within Jupyter Notebook (Anaconda). Key functions used include pd.read_csv() for loading data and pd.concat() for combining multiple facility datasets.
- Removed duplicates.
- Standardized the DATE field to YYYY-MM-DD and reformatted TIME from numeric values (e.g., 930 to 09:30).
- Grouped data by Facility + Date and summed numeric columns like TOTAL, EZPASS, and VIOLATION.
- Split the large dataset into eight individual csv files one for each facility to manage file size and processing speed.

- **Each facility dataset was analyzed separately, then merged with additional datasets:**
  - Weather dataset (joined on DATE) - adds temperature, precipitation, and wind data.
  - U.S. Holiday dataset (joined on DATE) - identifies holidays and seasonal travel spikes.
  - All merges were done using the DATE column only, since each traffic file already represented a single facility.
  - Missing values were filled using forward-fill or monthly means to maintain data continuity.
  - For loops automated the reading, cleaning, and merging of all eight facility-level CSV files. Each iteration used os.listdir() to locate files and merge() to join datasets consistently on the DATE field.
  - Post-integration validation included checking for missing values using isnull().sum(), duplicates via drop_duplicates(), and date continuity using pd.date_range(). Summary statistics were reviewed with describe() to confirm data integrity.
  - The final output was a clean, daily-level dataset for each facility, ready for forecasting, modeling, and visualization in later goals.

3. **Exploratory Data Analysis (EDA):**
   - EDA was performed in Python (*pandas, NumPy, matplotlib, seaborn*) to explore and validate the Port Authority traffic datasets before modeling.
   - Missing Values: Checked using *isnull().sum()* and handled with forward-fill *(fillna(method='ffill'))* or monthly mean imputation using *groupby('Month').transform(lambda x: x.fillna(x.mean())).*
   - Duplicates: Removed with *drop_duplicates()* during initial cleaning of the raw traffic file.
   - Distinct Values: Verified with *nunique()* to ensure one record per DATE and Facility_Name.
   - Date & Time Formatting: Standardized using pd.to_*datetime() and str.zfill()* for consistency.
   - Outlier Handling: Outliers and extreme spikes in traffic and violation counts were smoothed through daily-level aggregation using groupby*(['DATE', 'Facility_Name']).agg(sum)* before merging with other datasets.

- Exploratory Visualization: Relationships among variables were explored using *seaborn.heatmap()* for correlation analysis and to identify top influencing factors such as weather, holidays, and weekends.

4. External Datasets Used

Yes, we plan to use two external datasets in addition to the company-provided files:

- U.S. Holiday Dataset - Includes fields such as DATE, Holiday, Day_Name, Month, and Year. It is used to capture the impact of national holidays and long weekends on traffic trends.
- Historical Toll Prices Dataset - For the year 2025.

5. Managing Large Datasets

After combining multiple datasets such as traffic, weather, and holidays, a very large number of records were generated. To make the data manageable, we followed a goal-based and facility-level approach using Python.

- The original Port Authority traffic data was split into eight separate facility files (e.g., Bayonne_traffic.csv, GWB_Upper_traffic.csv, Lincoln_traffic.csv etc.), each containing about ~ 90 - 110k rows. This made the data easier to clean, process, and analyze.

- We developed five individual Python scripts, one for each project goal. Each script handles its own joins and cleaning steps according to the goal's requirements and each goal provides its own compact cleaned dataset for modeling and visualization.

    - **Goal 1:** Combined traffic, weather, and holiday data (joined on DATE) for finding top 5 factors.
    - **Goal 2:** No joins, we will use combined traffic data for toll violation modeling which already contains all necessary fields such as total vehicles, E-ZPass, cash payments, and violation counts.
    - **Goal 3:** Merged traffic, weather, and holiday datasets on the DATE field. These joins enabled analysis of how weather, rain, temperature, and holidays influence congestion and traffic peaks.
    - **Goal 4:** Combined all traffic dataset with toll price (joined on DATE) to check shift in traffic pattern.
    - **Goal 5:** No joins will use Combined traffic data and forecasting models (Prophet and SARIMAX) to predict the usage of facilities.