Project Overview:

The project analyzes US-Canada and US-Mexico border crossing data to understand traffic patterns, trends, and anomalies. It uses data from the US Department of Transportation and involves data cleaning, analysis, visualization, and basic predictive modeling.

What's Done:

Data Acquisition: The project retrieves border crossing data from a public API and loads it into a Pandas DataFrame. The data is then stored in a MongoDB database for easier access and analysis.
Data Cleaning and Preparation: The data is cleaned by handling missing values, converting data types, and removing duplicates. It's also prepared for analysis by creating new features like month, weekday, and year.
Exploratory Data Analysis: The code performs various analyses, including:
Time-series analysis to identify seasonal patterns and monthly trends.
Geographic analysis to locate high-traffic border ports.
Measure analysis to understand the distribution of different crossing types (e.g., personal vehicles, trucks).
Anomaly detection to identify unusual traffic patterns.
Cross-border comparisons to see how traffic differs between states and border types.
Visualization: The analysis is visualized using various charts, including line charts, bar charts, pie charts, scatter plots, and box plots.
Predictive Modeling: A Random Forest Regressor is used to predict border crossing volumes based on features like date, location, and crossing type.
Tools and Technologies Used:

Python: The primary programming language used for data analysis and visualization.
Pandas: A library for data manipulation and analysis.
Requests: A library for making API requests.
MongoDB: A NoSQL database used to store the data.
PyMongo: A library for interacting with MongoDB.
Matplotlib and Seaborn: Libraries for data visualization.
Scikit-learn: A library for machine learning and data preprocessing.
Key Results from Plots and Tables:

Seasonal Patterns: Border crossings show clear seasonal variations, with peaks during summer months.
Busiest Ports: Certain ports consistently have higher traffic volumes compared to others.
Crossing Types: Personal vehicle crossings are generally the most frequent.
Anomalies: Unusual spikes or drops in traffic can be identified using anomaly detection techniques.
Border Differences: US-Mexico borders tend to have higher traffic compared to US-Canada borders.
Predictive Accuracy: The Random Forest model achieves a reasonable level of accuracy in predicting future traffic volumes.
Golden Tables:

The analysis produces several key tables:

geo_agg: Summarizes traffic volumes at different port locations.
measure_contribution: Shows the relative contribution of different crossing types.
cross_border_agg: Provides a cross-border comparison of traffic volumes.
predictive_agg: Gives average and threshold values for different measures, useful for forecasting.