# Final Project: Predicting Trends in California Border Crossings Over Time

Leslie Garcia

INF 6480: Statistics and Data Analysis

Professor Smith

Wayne State University

December 13, 2025

## Introduction

The purpose of this analysis is to explore patterns and seasonal trends in border crossings at California ports of entry, using U.S. Customs and Border Protection's Border Crossing Entry Data, which I downloaded from Kaggle. This dataset contains official U.S. Border Crossing Entry Data collected by the Bureau of Transportation Statistics. The dataset includes 400,000 records of traffic activity across various land ports in the United States, over 120. This includes multiple modes of transportation. The records span from 1996 to 2025. Each row provides the port name, state, border type, date, crossing measure (e.g., trucks, buses, pedestrians, trains, or containers), and the total number of recorded crossings. I selected this dataset because it covers recent years and is well-suited to time-series analysis, forecasting, transportation studies, and visualization projects. My interest in using this data was to examine U.S.–Mexico border crossing data for California ports of entry to determine whether holiday periods exhibit predictable increases in travel volumes. Using descriptive analyses, hypothesis testing, and a multivariate linear regression model, the study evaluated seasonal patterns and key predictors of border activity.

Results showed no statistically significant difference in crossing volumes between holiday and non-holiday months, despite small average increases during July, November, and December. I usually cross the border during federal holidays in those months, which appears consistent with the trend. In contrast, transportation mode emerged as a major driver of crossing volume, with pedestrian and personal-vehicle traffic substantially higher than rail or commercial freight traffic. The regression model, including year, month, holiday status, mode of travel, and standardized geographic variables, explained approximately 68% of the variance in log-

transformed crossing counts, demonstrating strong predictive ability. However, holiday status was not a significant predictor.

Findings suggest that while California border crossings follow long-term growth trends and strong mode-specific patterns, holiday travel does not reliably produce surges large enough to influence overall monthly totals. Future work should consider weekly or daily data, incorporate policy and economic indicators, and evaluate non-linear forecasting models. This led me to examine whether border-crossing trends increase each year, since, overall, regardless of the holiday month, the numbers remain consistent.

Border crossings between Mexico and California represent a major channel of economic activity, tourism, and labor mobility. Understanding patterns in crossing volumes is essential for traffic management, staffing, and forecasting demand at ports of entry. While popular narratives suggest that travel increases during holiday seasons, empirical evidence specific to border movement is limited.

The primary objective of this analysis is to determine whether the holiday months (July, November, and December) are associated with higher border-crossing counts and whether these trends can support predictive modeling. The study focuses on California because it contains several of the nation's busiest land ports, including San Ysidro, Otay Mesa, and Calexico.

The following research questions guide the analysis:

1. Do California border crossings increase significantly during holiday periods?
2. How do crossing volumes vary by transportation mode?
3. What variables best predict crossing counts over time?

4. Can linear regression adequately model and predict border activity?

This report presents the data preparation process, descriptive findings, hypothesis testing, and interpretation of a predictive model built from the dataset.

## Method

Data Source: Data were obtained from the U.S. Customs and Border Protection Border Crossing Entry dataset, covering January 1996 through June 2025. The analysis focused solely on entries through California ports (N = 22,015 observations). Each observation represents a port–month–measure combination (e.g., number of pedestrian crossings at San Ysidro in May 2010).

## Data Cleaning and Variable Construction

Date strings were converted into Date objects using the lubridate package. Derived variables included:

- Year, Month, and Month_Name
- HolidayMonth (July, November, December)
- MajorTravelHoliday (March–May, July, November–December)
- Log-Transformed Values (log1p(Value))
- Standardized latitude/longitude for modeling
- Factor conversion of categorical variables

Observations with missing dates or crossing values were removed.

## Analytical Approach

The analysis proceeded in four stages:

1. Descriptive statistics describing central tendencies and distributions
2. Data visualization, including histograms, time series, and holiday comparisons
3. Hypothesis testing, including:
    - Independent samples t-test (holiday vs non-holiday)
    - One-way ANOVA (differences by mode)
    - Chi-square test of independence (mode × holiday status)
    - Correlation analysis (latitude vs volume)
4. Linear regression modeling, predicting log-transformed monthly crossing counts using:
    - Holiday status
    - Transportation mode
    - Year and month
    - Standardized location variables

A 70/30 train–test split evaluated model performance.

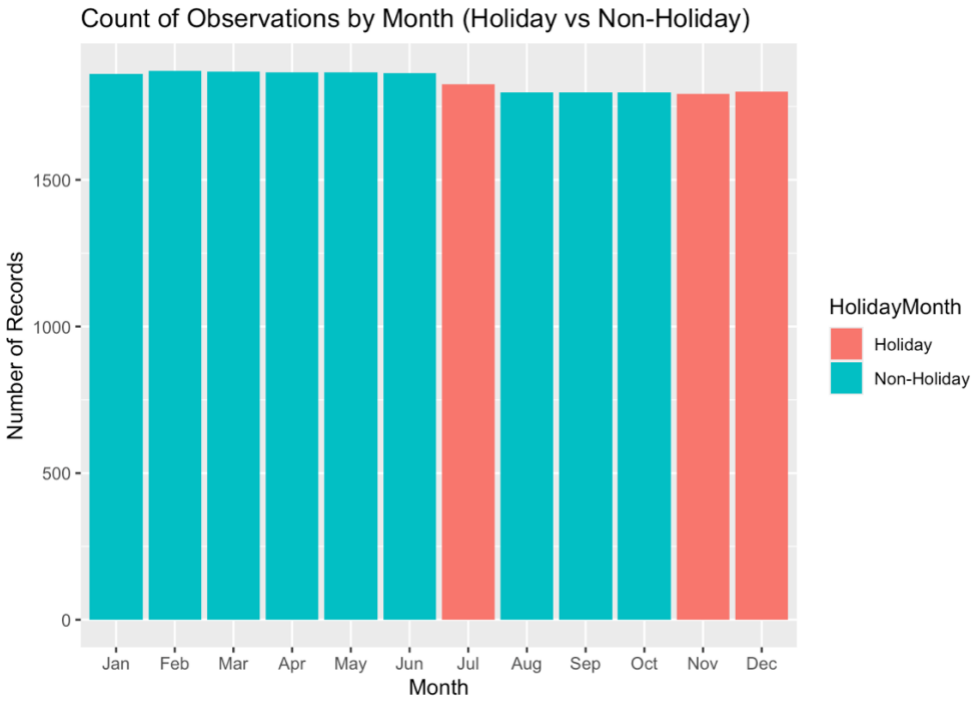
## Results

Crossing counts were highly skewed (mean = 140,649; median = 1,063), reflecting a small number of extremely high-volume observations. Holiday months showed slightly higher means, but median values were inconsistent, suggesting that any holiday effect may be non-systematic. Transportation modes were unevenly represented: personal vehicles and pedestrian

crossings dominated, while rail traffic comprised far fewer observations. A long-term upward trend was visible in time-series plots, despite temporary downturns linked to global and policy events (e.g., 2008 recession, pandemic-related restrictions).

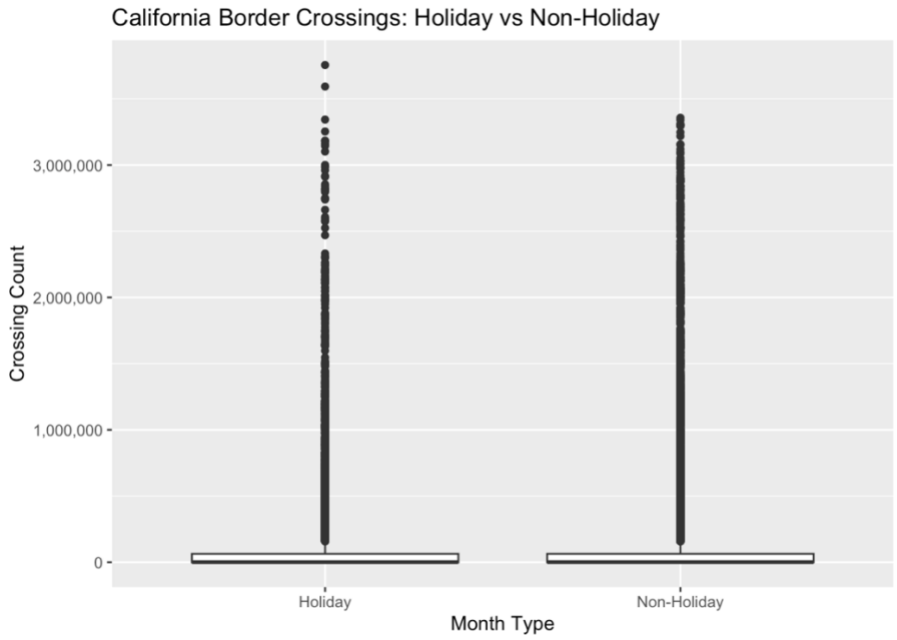
## Descriptive Statistics and Data Visualization

Figure 1. Count of Observations by Month (Holiday vs Non-Holiday)



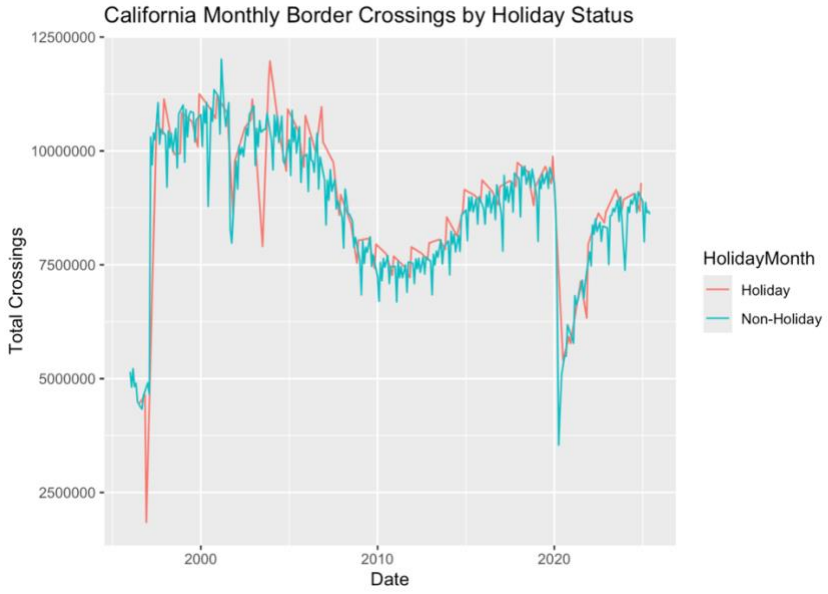

The distribution of observations across months is uniform, confirming that data availability does not bias results. The holiday months (July, November, and December) are correctly labeled. This figure does not reflect crossing volume, but it validates the dataset structure.

Figure 2. Boxplot of Holiday vs Non-Holiday Crossings



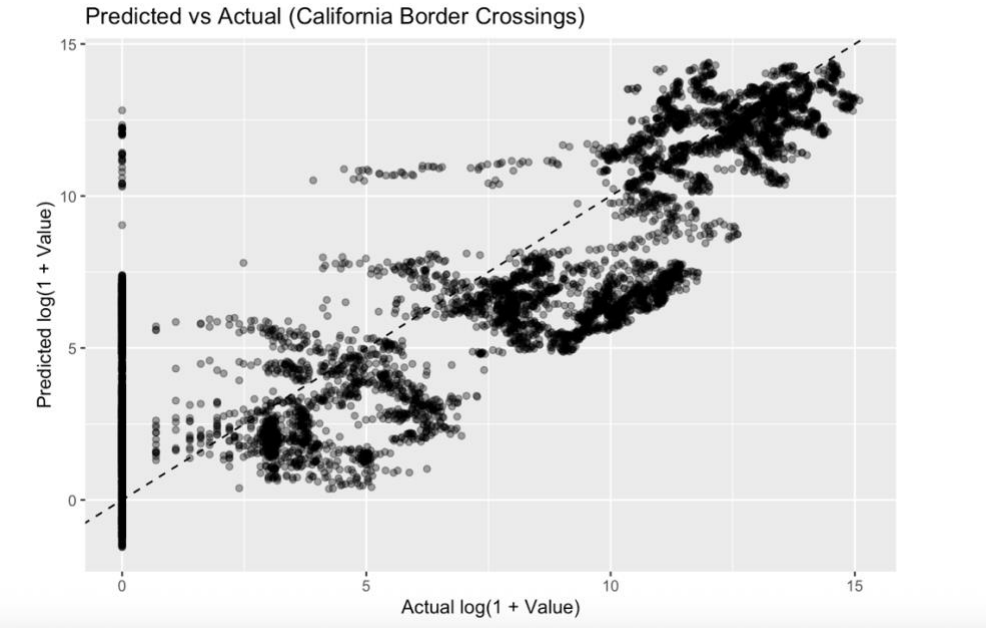

Holiday and non-holiday months have nearly identical medians and overlapping interquartile ranges. This visual evidence supports the conclusion that holidays do not substantially influence monthly crossing volume.

Figure 3. Holiday vs Non-Holiday Trends Over Time




Holiday and non-holiday lines nearly overlap, with no clear or consistent holiday spikes. Occasional peaks happen, but they vary from year to year.

Figure 4. Predicted vs. Actual Values




The regression model's predictions align closely with actual log-crossing values, demonstrating solid predictive performance. Points follow the 45-degree line, with some spread indicating unexplained variation.

## Hypothesis Testing

### Holiday Effect (T-Test)

The t-test comparing log-transformed mean crossing counts showed:

- $t(9186) = -0.66$, $p = .51$

Interpretation: There is no statistically significant difference between holiday and non-holiday months. Holidays do not reliably increase border traffic at a monthly resolution.

### Differences by Transportation Mode (ANOVA)

$F(11, 22003) \approx 3605$, $p < .001$

Interpretation: Transportation mode accounts for substantial variation. Pedestrian and personal-vehicle crossings are substantially higher than those for rail or freight, suggesting that port operational demands differ markedly by mode.

**Mode × Holiday Status (Chi-Square)**

$\chi^2(11) \approx 0.05$, $p = 1.00$

Interpretation: Holiday months do not change the distribution of transportation modes. Travelers do not shift modes (e.g., from cars to walking) based on holidays.

**Latitude Correlation**

$r = -.18$

Interpretation: More southern ports tend to have higher volumes, with San Ysidro and Calexico accounting for large portions of overall traffic.

**Linear Regression Model**

The final regression model included holiday status, transportation mode, year, month, and standardized location measures.

**Model Fit**

- Train $R^2 = .686$
- Test $R^2 = .680$
- RMSE ≈ 2.87 (log scale)

Interpretation: The model explains roughly two-thirds of the variance in crossing volume, indicating strong predictive value for a linear model.

### Key Predictors

- Year was strongly positive, reflecting long-term growth in crossings.
- Latitude and longitude were significant, with southern ports exhibiting higher counts.
- Transportation mode had the largest effects, with pedestrian and passenger vehicle crossings substantially exceeding commercial traffic.
- HolidayMonth was not significant ($p = .13$), supporting earlier findings.

## Findings

This analysis sought to determine whether California border crossings exhibit predictable increases during holiday periods. Evidence from descriptive analyses, hypothesis testing, and regression modeling suggests that holiday months do not reliably affect total crossing volumes. Instead, the strongest and most consistent predictors were:

- Transportation mode
- Geographic location
- Year (long-term upward trend)

The lack of a holiday effect might suggest that using more specific holiday dates could better capture short-term surges. Economic and policy factors tend to have a greater impact on border activity than seasonal travel. Additionally, port infrastructure and operational constraints may limit the visibility of spikes during holiday periods. The model's strong predictive performance demonstrates that border-crossing volumes can be reasonably well forecast using linear methods, although nonlinear models may capture additional variance.

## Limitations

1. Monthly data likely masks short-term patterns such as weekend or single-day holiday peaks.
2. External factors (e.g., policy changes, border enforcement, currency fluctuations) were not included.
3. The dataset records entries only, not exits, potentially limiting completeness.
4. Linear regression may not capture nonlinear temporal trends or interactions.

## Conclusion

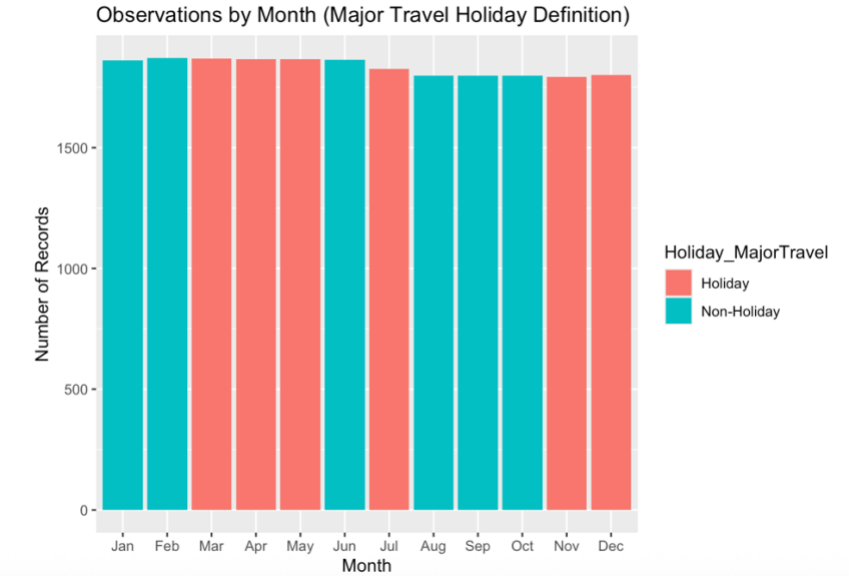
California border crossings display distinct mode-specific and long-term temporal patterns; however, they do not exhibit significant holiday effects at the monthly level. Variables such as transportation mode and port location predominantly influence predictive models, whereas holiday status offers minimal contribution to explanatory power. Future research incorporating weekly or daily data, along with broader contextual factors, may provide more precise insights into holiday travel behavior. Additionally, analyzing specific major holidays or weekends can offer further insights into the anticipated surges.

## References

Naik, U. (2025). *USA_Border_Crossing_Entry_Data*. Kaggle.com. https://www.kaggle.com/datasets/utkarsh1093/usa-border-crossing-entry-data

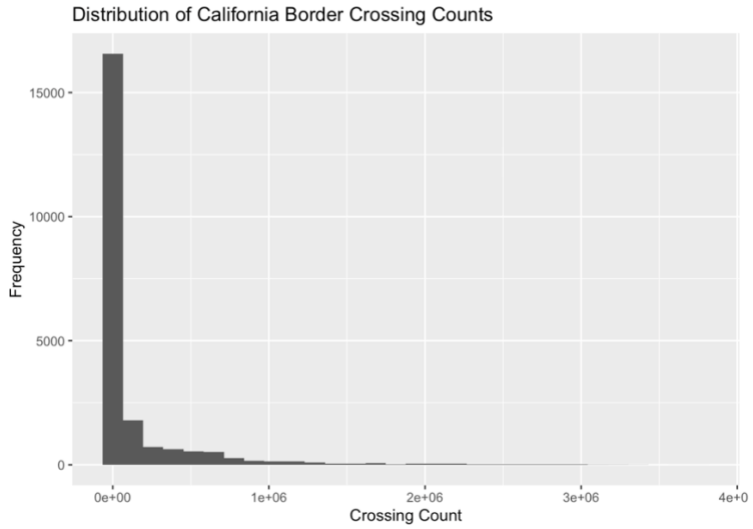
# Appendix

Table 1. Observations by Month under Expanded Holiday Definition



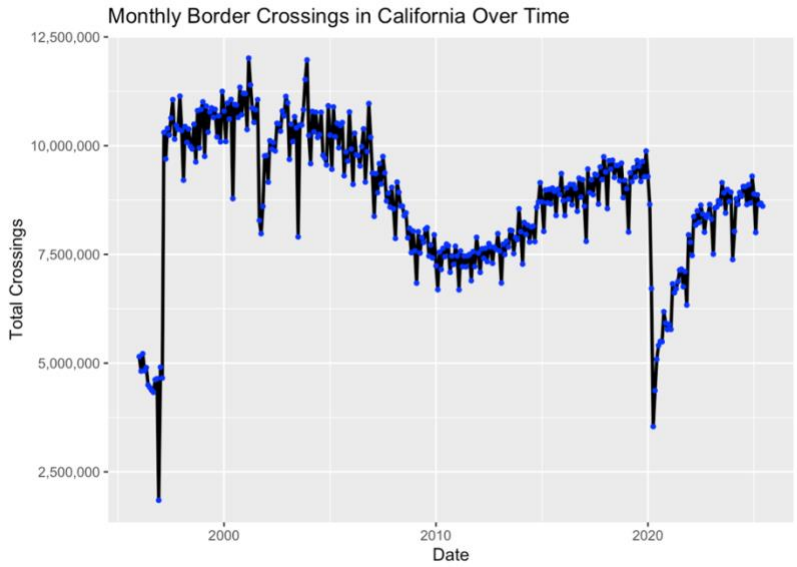

After using an expanded holiday classification, the distribution stays even. This shows that any differences seen in crossing volume are caused by real activity patterns—not variations in record frequency.

Table 2. Histogram of Crossing Counts



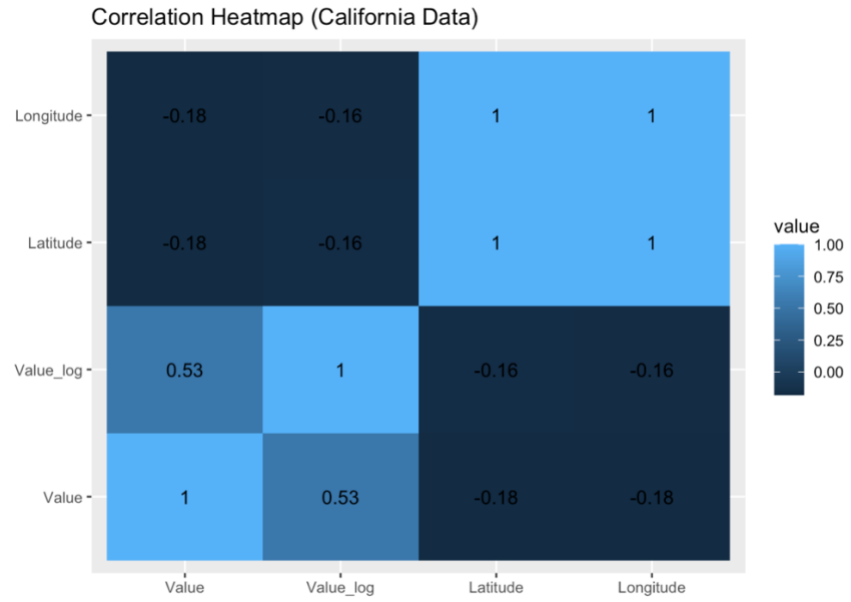

Crossings are extremely right-skewed, with most observations under 50,000 and a small number exceeding 3 million. This validated the log transformation used in all inferential tests and regression modeling.

Table 3. Monthly Crossings Over Time




The time series plot illustrates long-term growth, with recession dips around 2008 and a sharp decline in 2020 due to the pandemic. These macroeconomic and policy changes outweigh any seasonal effects observed at the month level.

Table 4. Correlation Heatmap




Value and Value_log are strongly correlated; Value has mild negative correlations with latitude Value and Value_log are strongly correlated; Value shows mild negative correlations with latitude and longitude, indicating higher volumes at southern ports. The absence of strong correlations suggests no multicollinearity issues. multicollinearity issues.