

CS109 Final Project:

Analyzing Data from the Stanford Open Policing Project to Study Traffic Stop Patterns

Tony Wang
CS109 Winter 2024

GitHub Repository link: <https://github.com/tonywangs/open-police-project>

Introduction

This project delves into the analysis of Traffic and Pedestrian Stops by Police in Rhode Island, using data from the Stanford Open Policing Project (<https://openpolicing.stanford.edu/>). The objective of this project is to understand the dynamics of traffic stops, focusing on the influence of driver demographics on police actions, the occurrence and outcomes of police searches, and the temporal trends in police traffic stops.

Methodology

The analysis involved several statistical and machine learning techniques to explore the data comprehensively, including:

- **Descriptive Statistics**, which provided an overview of driver ages and the frequency of various violations.
- **Violation Analyses**, which categorized the violations leading to traffic stops, revealing the most common reasons for stops.
- **Search and Outcome Analyses**, which examined the proportion of stops that led to searches, including an analysis of search types and outcomes, and also explored the distribution of stop outcomes (e.g., citations, warnings, arrests).
- **Demographic Analyses**, which investigated the relationship between driver demographics and police actions.
- **Temporal Analyses**, which looked at changes over time in the number of stops or searches.
- **Machine Learning Analyses**, which predicted outcomes of stops based on driver information, time, and various other factors.

Key Findings

Descriptive Statistics and Violation Analysis

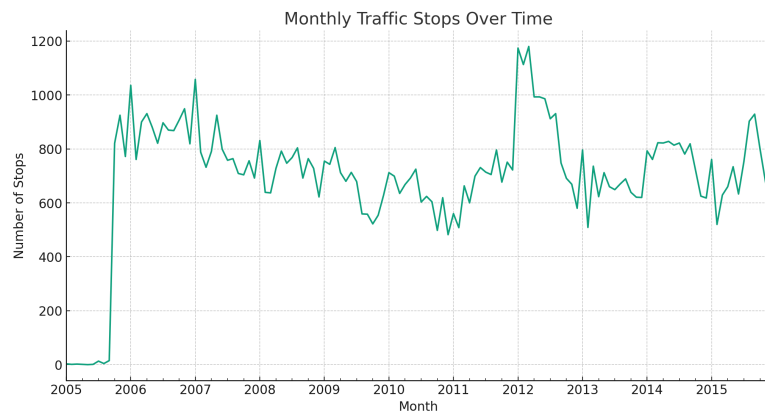
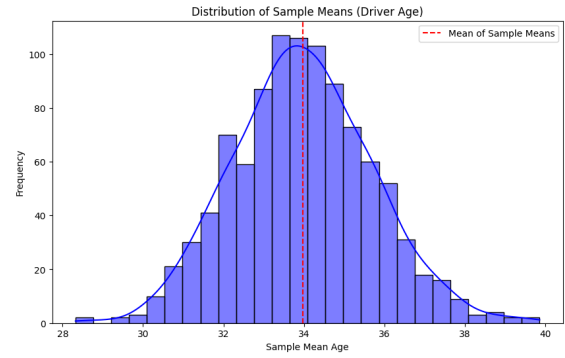
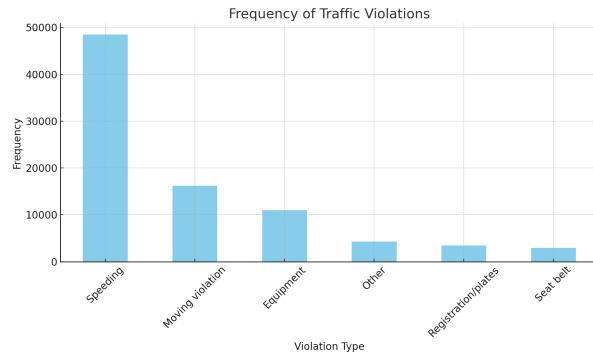
The average age of drivers stopped was approximately 34 years old, with the majority of stops involving speeding violations, significantly outnumbering other violations such as seat belt non-compliance, vehicle equipment issues, or registration/permit violations. The bar chart visualizing these findings starkly highlighted the prevalence of speeding, suggesting a strong enforcement focus on this particular traffic safety issue. This focus on speeding might reflect broader road safety strategies aimed at reducing accidents and fatalities associated with high-speed driving.

Search and Outcome Analysis

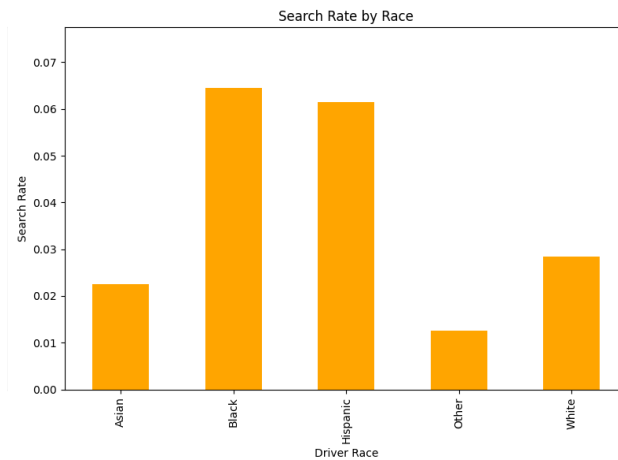
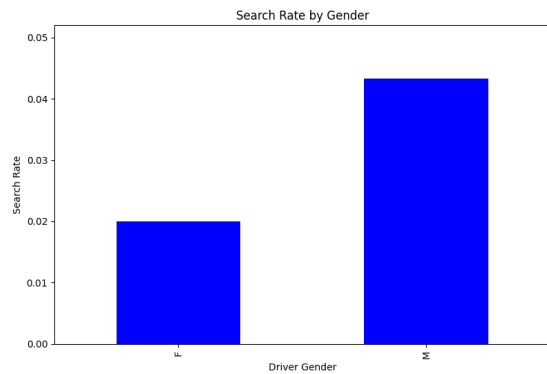
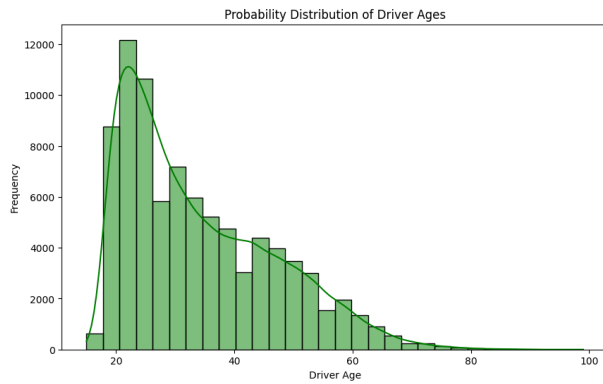
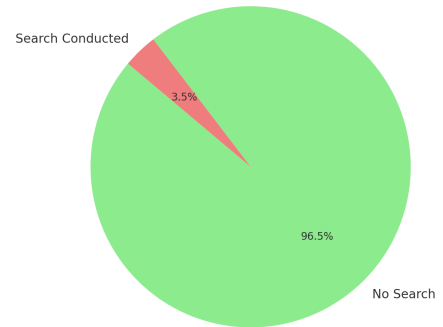
Only a small fraction (approximately 3.5%) of traffic stops actually led to police searches. A pie chart illustrating the proportion of stops leading to searches highlighted this rare occurrence, with searches more likely when violations were involved. The low rate of searches could also indicate adherence to legal standards requiring reasonable suspicion or probable cause. A significant proportion of stops resulted in citations, with fewer instances leading to arrests or searches related to drugs. This indicates that in Rhode Island there seems to be a primary focus on enforcing traffic laws over drug-related offenses.

Demographic Analysis

Multinomial Logistic Regression analysis revealed significant disparities in search rates, particularly concerning male drivers and drivers of Black or Hispanic race. These drivers were found to be more likely to be subjected to searches compared to their counterparts. This disparity raises critical questions about the factors influencing search decisions and highlights potential areas of concern regarding bias in traffic stop searches.



Proportion of Stops Leading to a Search



Temporal Analysis

The Time Series Analysis conducted on the monthly traffic stops within Rhode Island revealed interesting patterns of fluctuation without demonstrating a consistent long-term trend. The observed fluctuations suggest variability in traffic stop frequencies, which might be influenced by a myriad of factors such as seasonal variations in driver behavior, changes in traffic law enforcement priorities, or external events impacting road usage. For example, increased traffic stops during holiday seasons might reflect a strategic emphasis on road safety due to higher traffic volumes. Conversely, a decline in stops during certain periods could correlate with external factors such as severe weather conditions limiting road travel. These cyclical patterns, marked by minor increases and decreases, suggest that while traffic enforcement activity varies, it remains relatively stable over time, balancing out any significant deviations.

Random Variable (RV) Analyses

I also decided to simulate a binomial distribution scenario by considering the event of a search being conducted during a traffic stop as a 'success'. We can assume the probability of a search being conducted is the overall rate of searches in the dataset. Thus, I simulated 1000 traffic stops and visualized the distribution of the number of searches conducted. If I had access to the exact timestamps of each traffic stop, I would also try using the Poisson distribution, since it is useful for modeling the number of occurrences of events (like traffic stops) over a continuous time interval, assuming the events happen independently and at a constant rate.

Machine Learning Analyses

The Random Forest Classifier, used to predict whether a search would be conducted, achieved an accuracy of approximately 96%. However, the model's precision, recall, and F1-score for predicting searches were low, suggesting there were challenges in accurately forecasting search decisions. This discrepancy suggests that while the model is adept at recognizing the majority class (no search conducted), it struggles with the minority class (search conducted), highlighting a common challenge in machine learning known as class imbalance. This outcome prompts further examination of model parameters, feature selection, and possibly the exploration of techniques designed to better handle imbalanced datasets, such as SMOTE (Synthetic Minority Over-sampling Technique) or targeted cost functions to improve model performance for minority class predictions.

Discussion

The comprehensive analysis of traffic and pedestrian stops in Rhode Island reveals several critical insights into policing practices and their implications. The visualizations generated throughout this analysis played a crucial role in elucidating the patterns and trends within the data, guiding my conclusions and recommendations. While traffic law enforcement (e.g., citations for speeding) constitutes the bulk of police stops, the demographic analysis and machine learning models point to disparities in search practices. Specifically, male drivers and drivers of certain races (Black, Hispanic) are disproportionately searched, raising questions about fairness and bias. The temporal analysis suggests that while the overall number of stops fluctuates, there is no clear upward or downward trend.

Conclusions

This research has illuminated the complex dynamics of traffic stops in Rhode Island, combining statistical analyses with machine learning models to uncover patterns in enforcement practices. While the majority of stops adhere to traffic law enforcement, disparities in search rates based on demographics highlight areas for further investigation and potential reform. Future research should continue to monitor these trends and explore the effectiveness of interventions aimed at reducing bias and ensuring fair policing practices.

Citation

As authorized in the assignment instructions, ChatGPT was very helpful in the completion of the assignment, helping with method ideation, code writing, paper drafting, etc.

Data source

<https://www.kaggle.com/datasets/faressayah/stanford-open-policing-project/data>