Milestone 4: Traffic Violation
STAT 479
Group 4
Code: https://github.com/hparkailey/479_group4/blob/main/m4.Rmd

1. Introduction

Driving is the most popular mode of transportation that Americans use to commute between home and work [1]. However, traffic accidents can result in road crash fatalities and disabilities, which is being recognized as a major public health concern [2]. We believe that minimizing traffic violations would prevent severe accidents and create safer local communities. Therefore, we aim to investigate if there's any patterns in traffic violations based on population demographics of racial groups, age groups, and gender by studying a Maryland County traffic violations dataset [3]. By studying demographic relationships with traffic violations, we hope that policymakers will be able to come up with a plan tailored to the specific social group where traffic violations and accidents will be minimized in the future. We used the "Traffic and Drugs Related Violations" dataset [3], which includes 12 variables, such as driver's age, gender, and race, relating to 65,000 traffic records in Maryland.

2. Shiny App

Our app allows users to interactively visualize traffic violations and draw interesting insights. There are three tabs: description of our app, static plots for overview, and interactive plots for exploration. The description page explains our app with a word cloud of keywords. In the static plots page, there is a pie chart showing different violations followed by a bar chart comparing the percentage of races of Maryland and races in our dataset. Below, there is a density plot showing the relationship between drivers' race and age and a ridgeline plot showing the relationship between drivers' gender and age for violations. In our third page, we have user interactions with a slider bar for age, a checkbox for gender, and select input for race. Below, there is a scatter plot where users can brush over the plot to reveal further data within an interval of dates. We defined reactive functions that create data frames to represent the brushed points. We used renderPlot() functions to render plots that were drawn based on user inputs and selections. First, a line plot shows the frequency of violations at different hours of the day followed by two violin plots. On the left, a violin plot displays the distribution of arrests across ages and on the right, a violin plot displays the distribution of drug-related stops across ages. Overall, our app enables users to study traffic violation patterns based on demographics and dates.

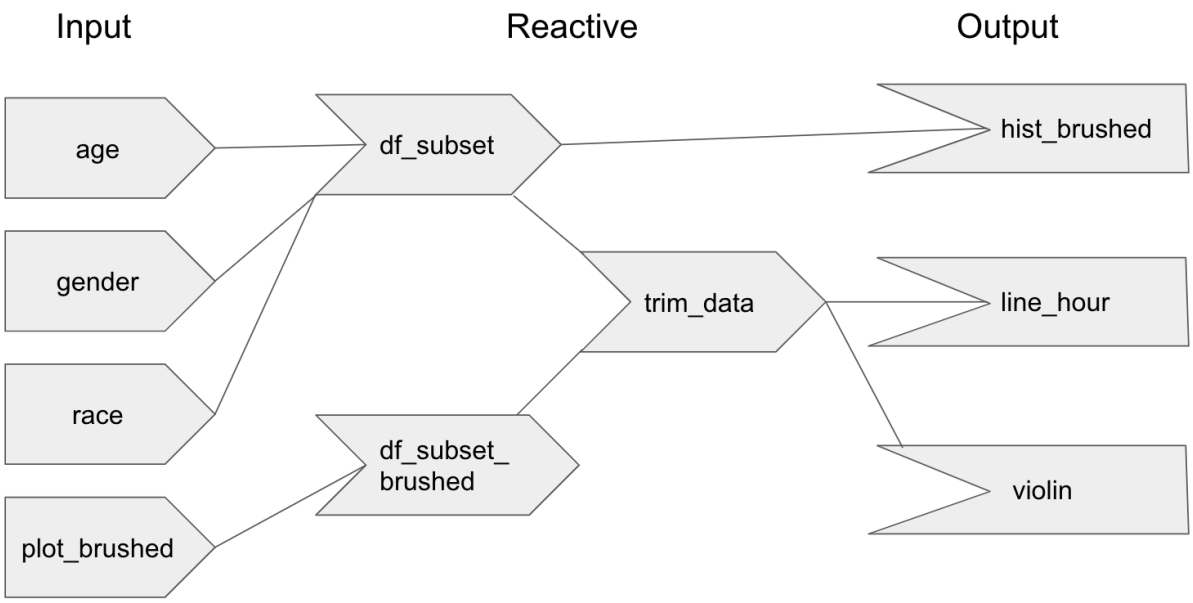

Figure 1. The reactive graph for the interactive plots for exploration.

3. Interesting Findings

For the density plot colored by gender which shows the relationship between driver age and traffic violations, we see that age distribution for each violation type somewhat differs based on drivers' gender, but not to a significant extent. As for the density plot that's colored by race which represents the relationship between the driver's age and traffic violation types, we found some subtle differences in traffic violation patterns across ages based on different races. For equipment violations, roughly the majority of White and Hispanic drivers in our dataset committed equipment violations at the age of 23. Most Asian drivers tend to commit equipment violations at a higher age of around 30. For violations related to speeding and registration/plates, age distributions are skewed to the right for all races, which indicates that drivers of younger age (around mid-20s) committed more of these violations in our dataset.

In general, all kinds of violation types decrease from midnight to 5:00 AM shown by the line plot. Violations jump dramatically, especially for speeding violations, from 5:00 AM to 10:00 AM, which can be due to an increase in traffic due to the morning traffic rush. After the morning rush, all violation types decrease again until it slightly increases again at around noon, and then decrease gradually till later in the evening. After 8:00 PM, we can notice an increase in all traffic violations again. Thus, the line plot shows that hours at which violation counts are high are correlated with hours at which the traffic is the busiest.

Both violin plots show that the median age of drivers involved in traffic violations and getting arrested and drivers who got stopped due to drugs is mid-20s to 30s. There were no

noticeable differences in the median age based on gender. Although traffic violation counts for each violation type are roughly similar for all racial groups, there are differences in the shape of the distribution based on races for both violin plots. Overall, shapes of both violin plots are almost identical to shapes of distribution for white drivers, since the majority of the population within our dataset is classified as white. Compared to the shape of distribution across all races, shapes of distribution for black drivers seem to be concentrated on the left side, indicating that, within our dataset, younger black drivers are arrested and stopped more due to drugs compared to older black drivers. As for Asian drivers, we can tell that most Asian drivers who got arrested are of younger age compared to other races, although there aren't that many Asian driver records within our dataset. Hispanic drivers have roughly similar shapes of distribution as the shape distribution of the overall population.

4. Conclusion

Using our app, we can investigate the driver demographic of our dataset as well as the relationship between each traffic violation pattern and driver demographic information. Overall, there weren't any substantial differences in the median age of a driver who got involved in traffic violations based on gender nor race, with all of them around mid-20 to 30s. The stop hour line graph revealed that the traffic violation counts coincide with the typical rush hour when traffic is the busiest. Violin plots show that there are some differences in the shape of the distribution that can vary for certain races, but we will need more data to verify this information. For future research, we hope to gather data from other counties as well, since the dataset we've used for this project was only collected in Maryland. By gathering more data from other counties, we hope to have more diverse driver demographics representative of the entire population.

Works Cited:

[1] https://www.statista.com/chart/18208/means-of-transportation-used-by-us-commuters/
[2] https://www.asirt.org/safe-travel/road-safety-facts/
[3] https://www.kaggle.com/shubamsumbria/traffic-violations-dataset/discussion/304376