Exploratory Data Analysis: Maryland State Patrol Traffic Stops

Joel Thomas Zachariah

College of Professional Studies, Northeastern University

ALY 6010 Probability and Statistics

Professor Terry-Jon Sizemore

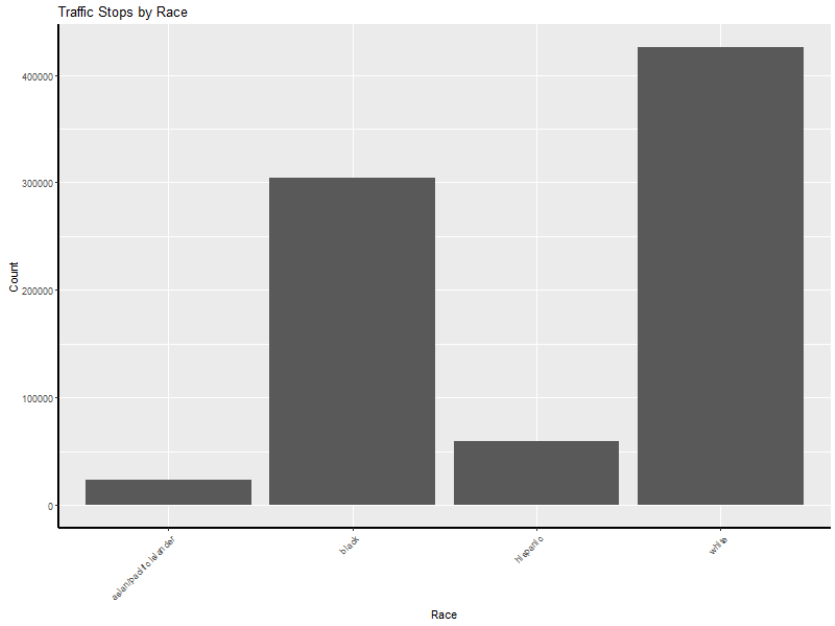
December 09,2024

This report provides an initial exploratory analysis of traffic stops conducted by the Maryland State Patrol. The dataset, sourced from the Stanford Open Policing Project, includes comprehensive details on traffic stops, covering demographics of individuals stopped, reasons for the stops, and their outcomes.

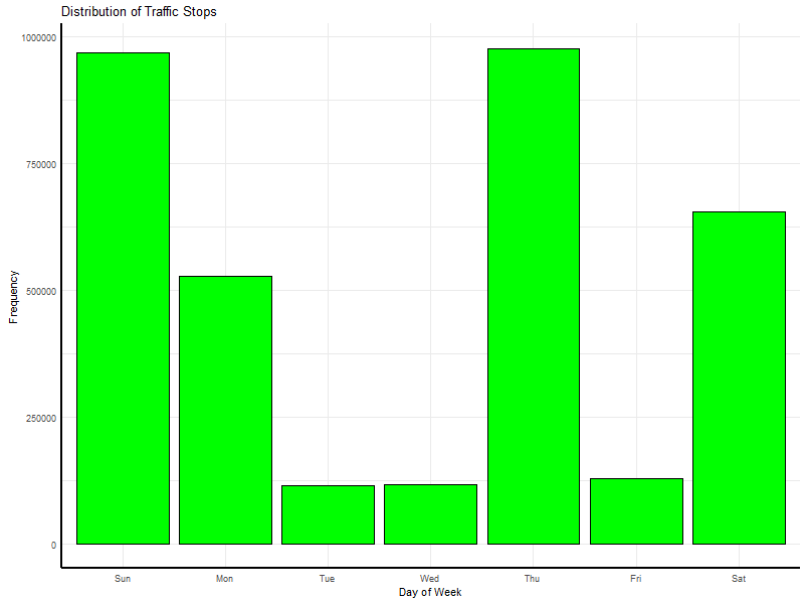
Below are the findings of my analysis

Traffic Stops by Race

The bar plot shows a difference in traffic stop frequency among racial groups, with White individuals having the highest number of stops in this dataset.

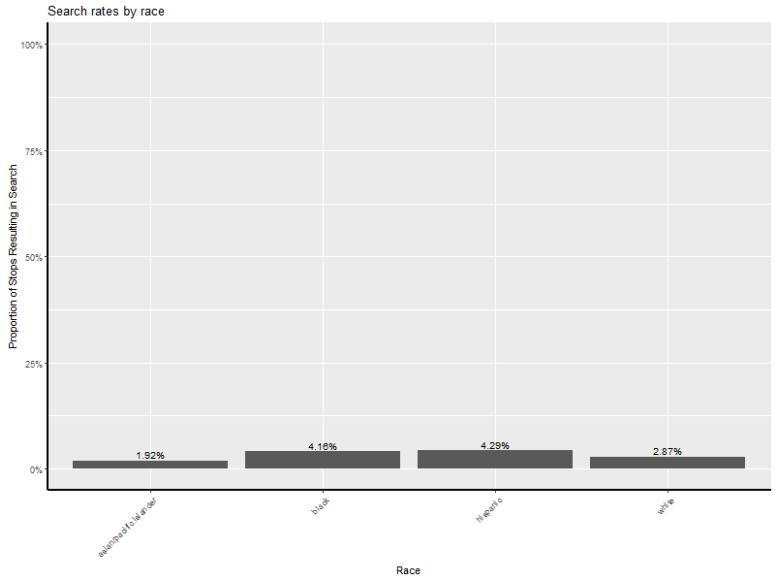


Day of Week Distribution of Stops



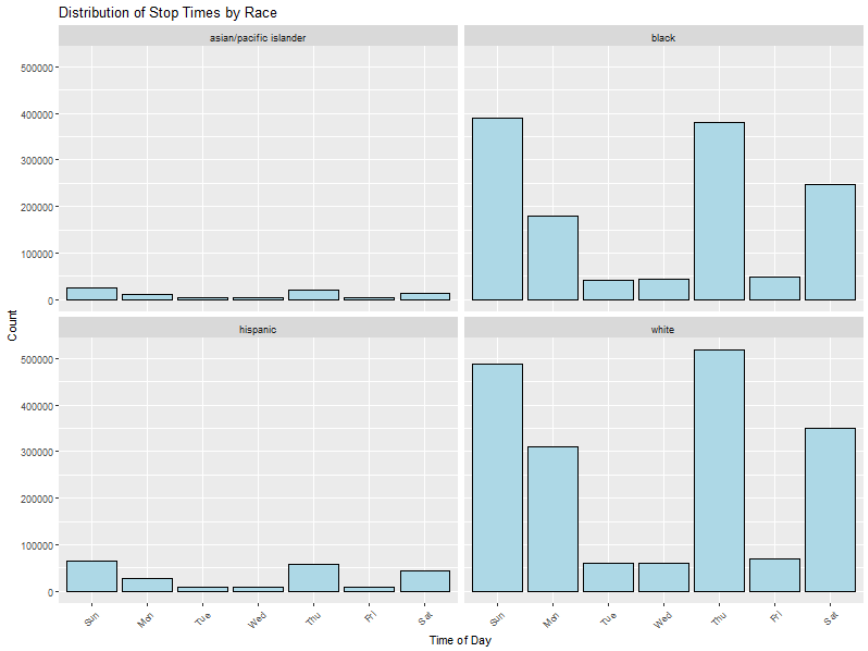
The bar chart shows that distribution of stop time is different for each day of the week. The highest amounts are the same for Sunday and Thursday and the lowest can be seen for Friday.

Search Rates by Race



This bar graph shows the proportion of traffic stops that resulted in a search, broken down by race, white has lower than both Hispanic and Black individuals. Asian/Pacific Islander individuals have the lowest search rate.

Distribution of Stop Time by Race



This plot displays the distribution of police stop times by day of the week and race. The stop patterns vary significantly by race and day of the week, with Black and White individuals facing higher stop counts, particularly on Sundays and midweek days.

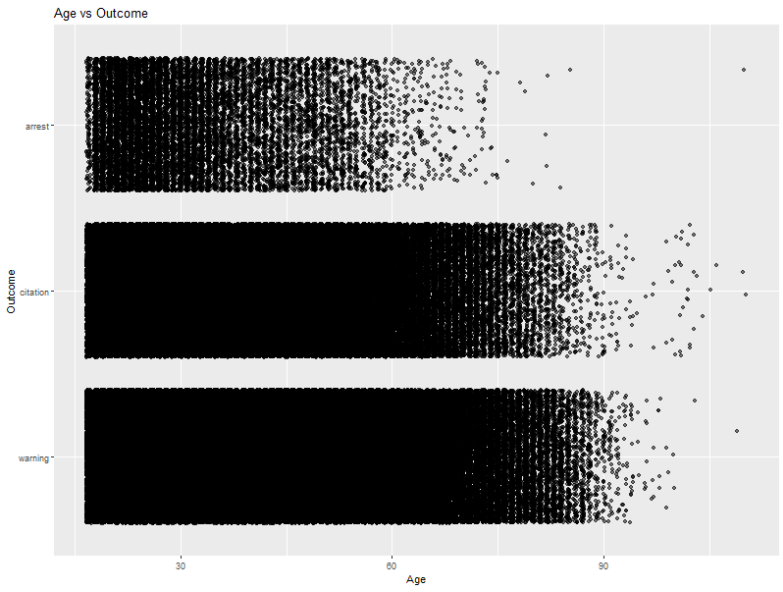
Descriptive Statistics

	X1
vars	1.00000000
n	810131.00000000
mean	37.81714686
sd	14.40657354
median	35.00000000
trimmed	36.62484165
mad	16.30860000
min	17.00000000
max	110.00000000
range	93.00000000
skew	0.64855279
kurtosis	-0.28614533
se	0.01600601

	subject_race	mean_age	sd_age	min_age	max_age	N
1	asian/pacific islander	39.15783	14.28861	17	102	23189
2	black	37.07226	13.41113	17	110	303295
3	hispanic	34.83919	10.79997	17	96	59665
4	white	38.69575	15.42150	17	110	423982

We begin by doing an initial summary for the attribute ‘subject_age’ and Mean age categorization according to each race
From the above analysis we have plotted the below graphs

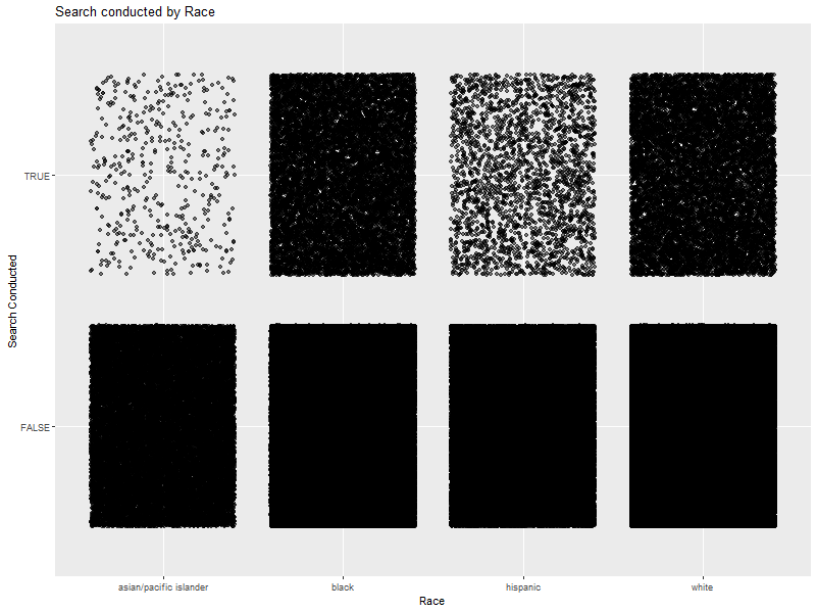
Age vs Outcome



The scatter plot shows the relationship between age and stop outcomes (arrest, citation, warning). Citations and warnings are common across ages 20–60, while arrests are denser among younger

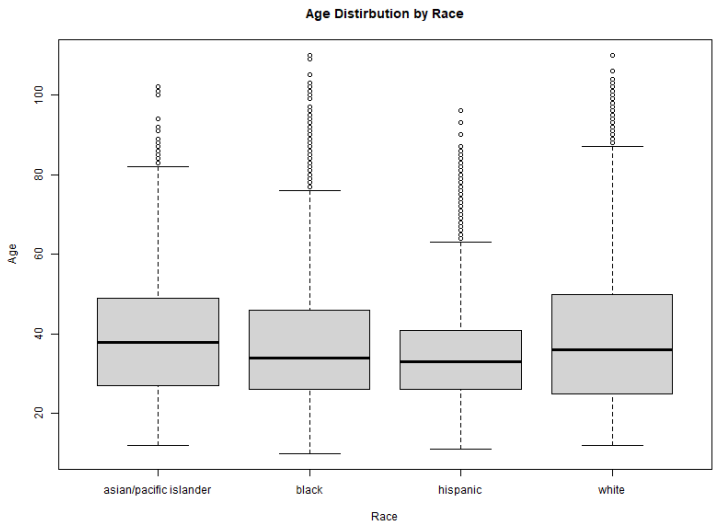
individuals, particularly ages 20–40. Outcomes decrease significantly with age, with few incidents above 70. Sparse outliers exist for individuals over 90.

Search Conducted by Race:



The plot highlights racial disparities in police searches during stops. Black and Hispanic individuals are searched more frequently, as shown by higher densities in the "TRUE" row, while Asian/Pacific Islander and White individuals have lower search rates. Most stops across all races do not involve a search ("FALSE" row), but the differences in search likelihood suggest unequal practices.

Age Distribution by Race



The median age of individuals stopped is 30–40 across all racial groups, with slight variations. Stops decrease with age, but outliers above 60 exist for all groups, and White individuals show a broader age range with higher maximum outliers. Hispanic individuals have a slightly lower median age compared to Asian/Pacific Islander and White individuals. Rare outliers above age

100 may reflect data errors or exceptional cases. Overall, age distributions are similar, centered on young to middle-aged adults.

Hypothesis Test

Two hypothesis called the t-test and chi-test are being performed to conduct analysis about the age and the search conducted on races. For hypothesis we will use the confidence level of 95% as default.

T-test

- **Null Hypothesis (H_0):** The mean age of individuals stopped is equal to 35 years.
- **Alternative Hypothesis (H_1):** The mean age of individuals stopped is not equal to 35 years.

```
        One Sample t-test

data:  data$subject_age
t = 171.75, df = 812531, p-value < 0.00000000000000022
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 37.71936 37.78215
sample estimates:
mean of x
 37.75075
```

We have performed the analysis, and the exact mean age is provided as '37.75075'. We use this to compare the p-value with the 95 per confidence level (0.05).
The p-value is less than 0.05 and hence we reject the null hypothesis

Chi-Test

Null Hypothesis (H_0): There is no association between race(asian/pacific islander vs. hispanic) and the likelihood of being searched during a traffic stop.
Alternative Hypothesis (H_1): There is an association between race (asian/pacific islander vs.hispanic) and the likelihood of being searched during a traffic stop.

We created a contingency table for the observed frequencies to find the search conducted between the races "asian/pacific islander, Hispanic".

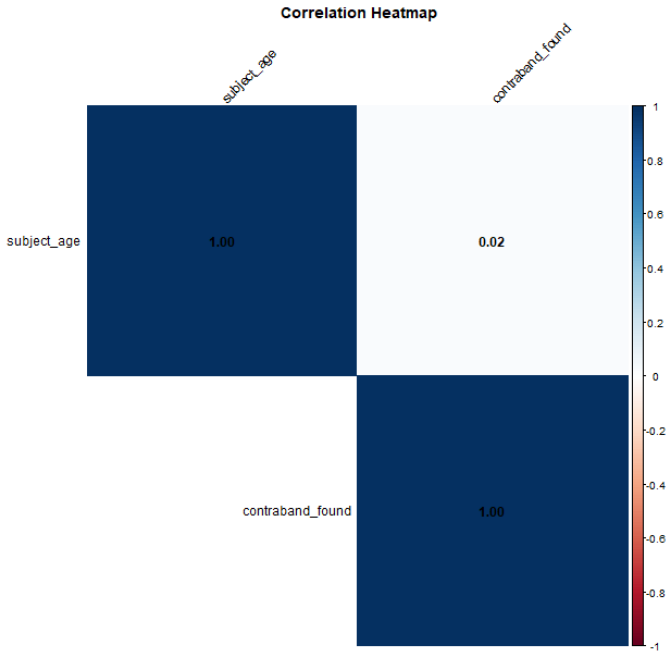
```
                        Search_Conducted
Race                       Yes   No
  asian/pacific islander  81621 1291
  hispanic               212977 8704
> |
```

Maximum number of searches were conducted for the Hispanic race.
Result of the chi-test is as below

- **Chi-Square Statistic (X^2):** 1066.6 **, Degrees of Freedom (df):** 1
- **p-value:** 0.00000000000000022

Using the data we provide the result that the p-value is less than 0.05 and hence we reject the null hypothesis

Correlation and Heatmap



A heatmap of correlations among subject_age, search_conducted, and contraband_found was produced, but search_conducted had no variance, leaving correlations for subject_age and contraband_found only.

- The correlation coefficient (0.0238) shows an extremely weak positive relationship, indicating negligible change in contraband likelihood with age.
- As expected, contraband_found is perfectly correlated with itself (correlation = 1).

Logistical Model

A logistic model showed a negligible positive effect of subject_age (coefficient = 0.00038) on the odds of finding contraband, despite statistical significance. Search_conducted was excluded due to singularity, as contraband_found depends directly on it. The minimal deviance reduction (339.33 to 339.14) indicates weak explanatory power for the predictors and a weak relationship between subject_age and contraband_found.

Conclusion

- **Search Rates by Race**: Black and Hispanic individuals experience searches more frequently during traffic stops compared to other racial groups.
- **Stop Outcomes by Race**: The outcomes of stops, such as arrests, citations, or warnings, vary significantly by race, with arrests being more common among younger individuals.

- **Day-of-Week Variation**: Traffic stop distributions differ across the days of the week, showing notable variations in stop frequency on certain days.
- **Age and Citation Trends**: Individuals aged 20–40 are more likely to receive citations or warnings, while arrests are less frequent but occur more among younger individuals.
- **Age Distribution and Search Disparities**: Boxplots show similar age distributions across most racial groups, with outliers present in each group. However, jitter plots reveal disparities in search rates, with Asian/Pacific Islander individuals appearing to face searches more often.
- **Correlation Analysis:** between subject_age and contraband_found indicates an extremely weak positive relationship. Age has little to no effect on the likelihood of contraband being found.
- **Logistic Model Analysis**: The subject_age coefficient suggests a negligible impact on the odds of finding contraband, though statistically significant and variable search_conducted was excluded due to singularity, as it is directly tied to the outcome (contraband_found).