Exploratory Data Analysis: California State Patrol Traffic Stops

Using information from the Stanford Open Policing Project, this study offers a preliminary exploratory analysis of California State Patrol traffic stops. The dataset contains comprehensive data on traffic stops, including the demographics of the people stopped, the causes of the stops, some visualizations that would show relations of age with outcome and race, hypotheses testing, correlation regression analysis.

Key Variables

The primary variables analysed include:

1. subject_race: The halted person's race.
2. subject_sex: The person's gender who stopped.
3. reason_for_stop: The justification given for the stop.
4. search_conducted: Provides TRUE or FALSE if a search was carried out.
5. contraband_found: Provides TRUE or FALSE if contraband was discovered.
6. consequence: The consequence of the stop, such as an arrest, warning, or citation.
7. date: The day that the stop took place.

Data Cleaning

The following data cleaning steps were performed:

1. Only the above-mentioned primary variables were chosen.
2. For easier manipulation, the "date" column was changed to a Date type.
3. "reason_for_stop" was transformed into a categorical component.
4. To categorize stops by day (Sun–Sat), a new variable called "dayofweek" was added.
5. Rows with "subject_race" listed as "other," "unknown," or absent (NA) were eliminated.

Initial Analysis Steps

1. Frequency tables for the primary category variables were created.
2. To examine the connections between variables, cross-tabulations were carried out.
3. To illustrate variable distributions, bar graphs and histograms were made.
4. Interactive charts were created for further in-depth understanding.

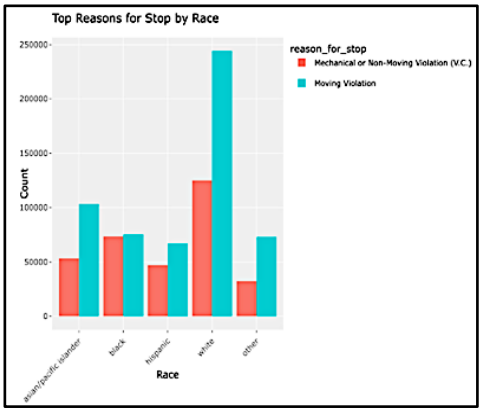
Racial Distribution of Stops

The frequency table of stops by race reveals:

White	Black	Hispanic	Asian	Other
372,318	152,196	116,014	157,684	106,858

‘White’ individuals account for most stops, followed by **‘Asian’** and **‘Black’** individuals.

Bar chart: Reasons for Stop by Race



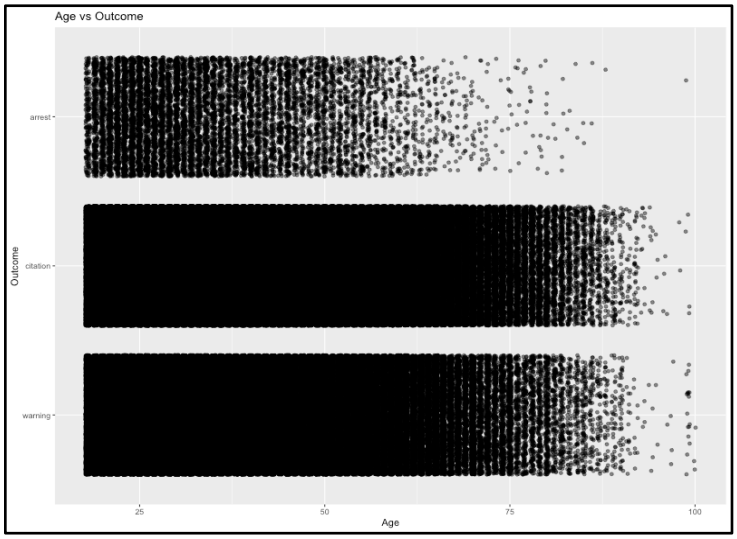
- The bar chart on the left reveals that **Non-Moving** and **Moving Violations** are the top 2 reasons for stop.

- Also, **White** individuals are highest stopped while average for other races is almost the same indicating fair traffic rules implementation by California State Police Department and is significantly higher than other races. This indicates that population of **Whites** is highest.

Scatter Plot: Age vs. Outcome

The association between subject_age (x-axis) and outcome (y-axis), where outcomes are categorized as warning, citation, or arrest, is depicted in the scatter plot. Jittering is used to minimize overlap, and each point corresponds to a traffic stop.

Important Takeaways from the Scatter Plot:

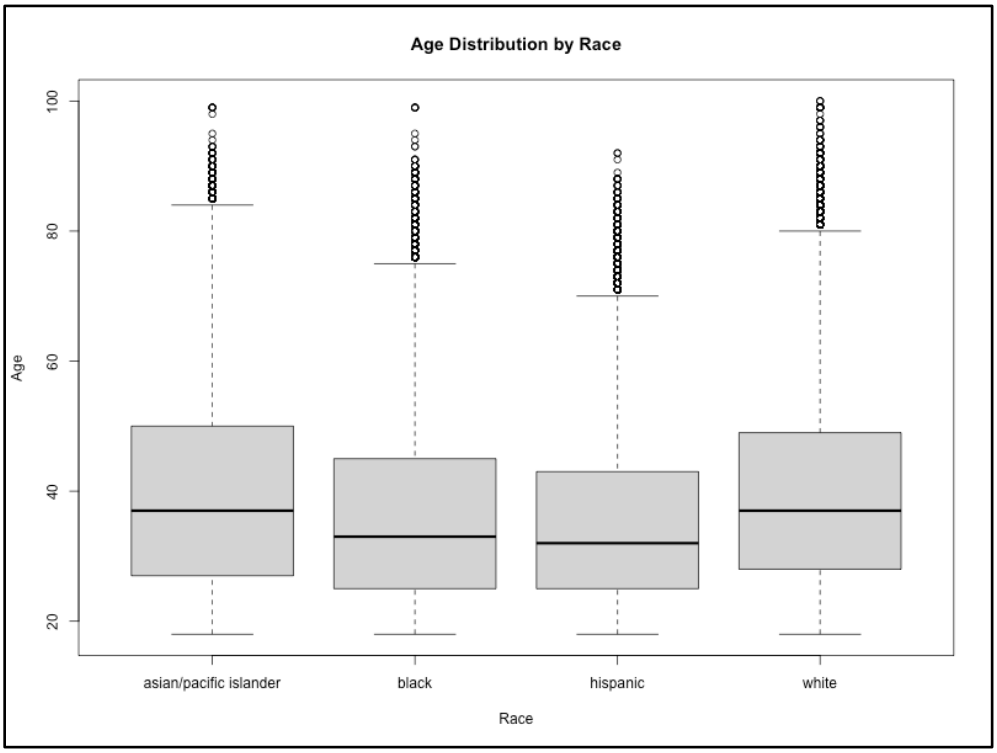


1. **Distribution of Ages by Outcome:**
- Although people of all ages are involved in traffic stops that result in arrests, citations, and warnings, most of them are between the ages of 20 and 50.
- People between the ages of 20 and 60 see a greater concentration of stops, particularly in the areas of **citations** and **warnings.**

2. **The Arrest Category Has Fewer Points Than Citations and Warnings:**
- The **"arrest"** category has significantly fewer points than the **"citation"** and **"warning"** categories, suggesting that arrests are comparatively rare occurrences.
- Arrests seem to be concentrated among younger people, but "**citations"** and "**warnings"** are more evenly dispersed across age groups.

3. **Arrest Outliers Among Older Age Groups:**
- There are notable outliers where arrests occurred among older people (ages 55 to 95), indicating deviations from the general pattern that arrests are more common among younger people.
- These anomalies imply that arrests for older age groups can still happen, but infrequently.

Boxplot: Age Distribution by Race



1. **Age Distribution Across Racial Groups**:
   - For the majority of racial groupings, the median age is between 35 and 40 years old.
   - The **Hispanic** group shows slightly **lower median age** comparatively.
   - The **Asian/Pacific** group displays a **wider interquartile range (IQR)**, suggesting greater age variability within this group.
2. **Outliers in Age**:
   - Across all racial groups, there are several **outliers** where individuals aged **70 and above** were stopped.
   - These outliers are especially prominent among the **White** racial group, though they are observed in all categories.
3. **Comparing Racial Groups**:
   - While age distributions are generally similar across racial groups, there are subtle differences worth further exploration, such as why some groups exhibit more **age variability** or a greater number of **outliers** than others.

Hypotheses Testing:

Hypothesis Testing is a statistical method that uses sample data to draw conclusions about a population. Here, in California State Patrol Traffic Dataset, following two hypotheses are performed:

1. One Sample T-test for mean age:

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 = 196, df = 842741, p-value <0.0000000000000002
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 38 38
sample estimates:
mean of x
       38
```

The p-value comes out to be very less than significance level of 0.5 and hence therefore we **reject the null hypothesis**.
There is evidence to suggest that mean age of individuals stopped is significantly different from 35 years of age.

2. Likelihood of being searched depending on Black Race and White Race:

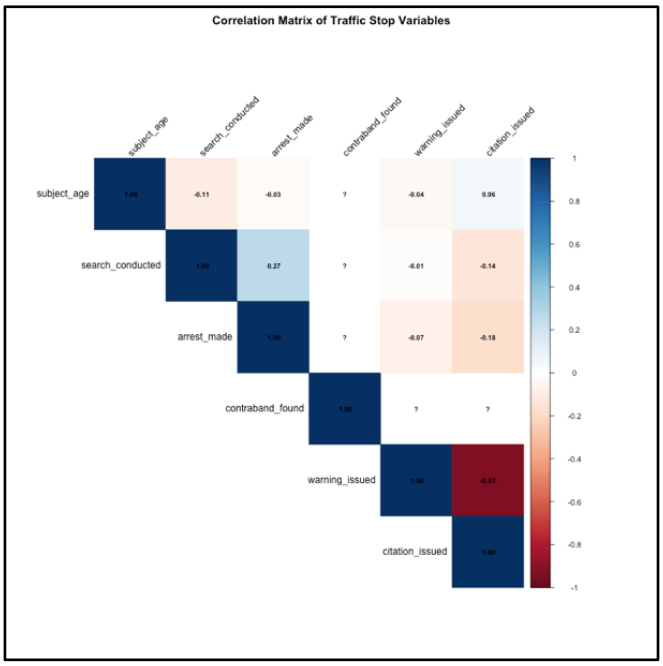
Null Hypothesis (H_0): There is no association between race and the likelihood of being searched during a traffic stop.
Alternative Hypothesis (H_1): There is an association between race and the likelihood of being searched during a traffic stop.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  observed
X-squared = 24314, df = 1, p-value <0.0000000000000002
```

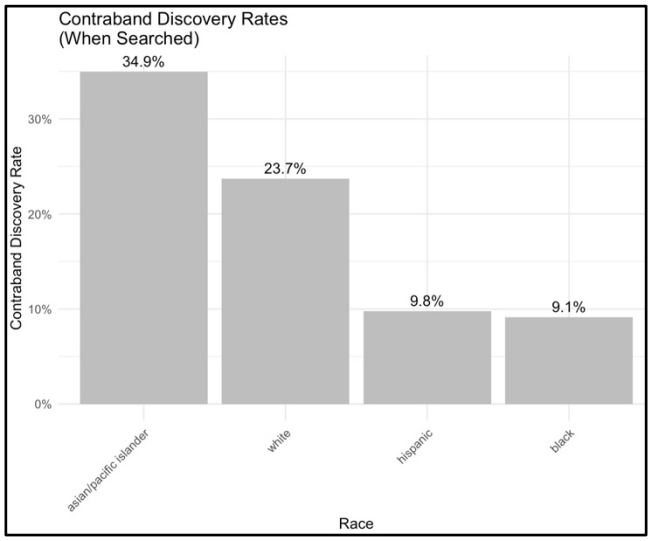
Since the p-value is less than 0.05, we **reject the null hypothesis**.
There is evidence to suggest that Black Drivers are more likely to be searched than White Drivers.

Correlation Heatmap:



- There is strong relationship between **'warning_issued'** and **'citation_issued'**.
- Therefore, when warning is issued, citations are rarely given and hence are vice-versa.
- **'subject_age'** indicate weak linear relationship with other variables.

Contraband Discovery Rates:



- **Asian/Pacific Islander** has highest contraband discovery rate of 34.9%.
- **Hispanic** and **Black** groups have relatively lower rates and is approximately have same rates.

Contraband Discovery Model:

```
Call:
glm(formula = contraband_found ~ subject_age + subject_race +
    subject_sex, family = binomial, data = subset(traffic_data,
    search_conducted == 1))

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)          -0.97574    0.05980   -16.3 <0.0000000000000002 ***
subject_age           0.01969    0.00107    18.4 <0.0000000000000002 ***
subject_raceblack    -1.63818    0.04712   -34.8 <0.0000000000000002 ***
subject_racehispanic -1.50086    0.05225   -28.7 <0.0000000000000002 ***
subject_raceother    -0.72757    0.05884   -12.4 <0.0000000000000002 ***
subject_racewhite    -0.58612    0.04669   -12.6 <0.0000000000000002 ***
subject_sexmale      -0.40440    0.03172   -12.8 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42106  on 50598  degrees of freedom
Residual deviance: 39396  on 50592  degrees of freedom
AIC: 39410

Number of Fisher Scoring iterations: 5
```

Predicting contraband found during a search based on age, race, and sex.

- The likelihood of having contraband is marginally higher for older people (+0.01969).
- Black, Hispanic, Other, and White individuals are less likely compared to the reference race.
- Males are less likely than females (-0.40440) in this category.
- Significance: $p < 0.001$ indicates that every variable is extremely significant.

Conclusion:

White people make up the largest racial group involved in traffic stops, with moving and non-moving offenses accounting for the majority of stops. While people up to 60 years old are mostly given warnings or citations, those between the ages of 20 and 50 are stopped the most often, especially for arrests, citations, and warnings. The median age for all racial groups is usually between 35 and 40 years old, with the Hispanic group having a somewhat lower median age and the Asian/Pacific group showing more fluctuation. A p-value less than 0.05 indicates that the mean age of those halted is substantially different from 35 years, according to statistical analysis.

There are racial differences in the likelihood of being searched during a stop; a statistically significant p-value indicates that Black drivers are more likely than White drivers to be searched. The Asian/Pacific Islander group has the highest rate of contraband discovery (34.9%), whereas the Black and Hispanic groups had relatively lower rates. There are minor linear connections between subject age and other characteristics, with contraband likelihood being somewhat higher for older people but lower for men and non-reference races. The likelihood of finding contraband varies significantly based on age, race, and sex.