Exploratory Data Analysis: North Carolina State Patrol Traffic Stops

The Exploratory examination of North Carolina State Patrol traffic stop data is presented in this report. The Stanford Open Policing Project provided the dataset, which includes comprehensive details about attributes. To guarantee accuracy, data preparation included cleansing, format conversion, and removing unnecessary information. To investigate trends and connections, the research makes use of frequency tables, cross-tabulations, visualizations (bar charts, scatter plots, and jitter plots), and statistical tests. Significant differences by race, age, and sex are revealed using a logistic regression model that forecasts the probability of searches based on demographic variables. The results show trends in the results of traffic stops and possible differences in search behavior by demography.

Variables of Interest

The following are the main variables of importance in this analysis:

1. Date: The day the traffic stop took place.
2. Location: The precise spot where the halt was made.
4. Subject Age: The individual who was stopped's age.
5. Subject Race: The race of the person who was stopped.
6. Subject Sex: The person's gender ceased.
8. Arrest Made: Shows whether an arrest was made (FALSE or TRUE).
9. Found Contraband: Indicates whether or not contraband was found during the search (TRUE/FALSE).
10. Search Conducted: Indicates whether a search took place (TRUE/FALSE).

Data Cleaning

Steps Followed:

1. Select Relevant Columns: Kept only necessary columns for analysis.
2. Convert Date Column: Change "date" to Date type.
3. Transform Categorical Variables: Convert to factor type for "location", "county_name", "subject_race", "subject_sex", "type", and "reason_for_stop".
4. Convert Age to Numeric: Ensure "subject_age" is numeric.
5. Convert Binary Columns to Logical: Change binary columns to logical type.
6. Add Day of the Week: Create "dayofweek" based on "date".
7. Remove Unwanted Race Categories: Remove rows with "subject_race" set to "other," "unknown," or NA.

Initial Analysis Steps

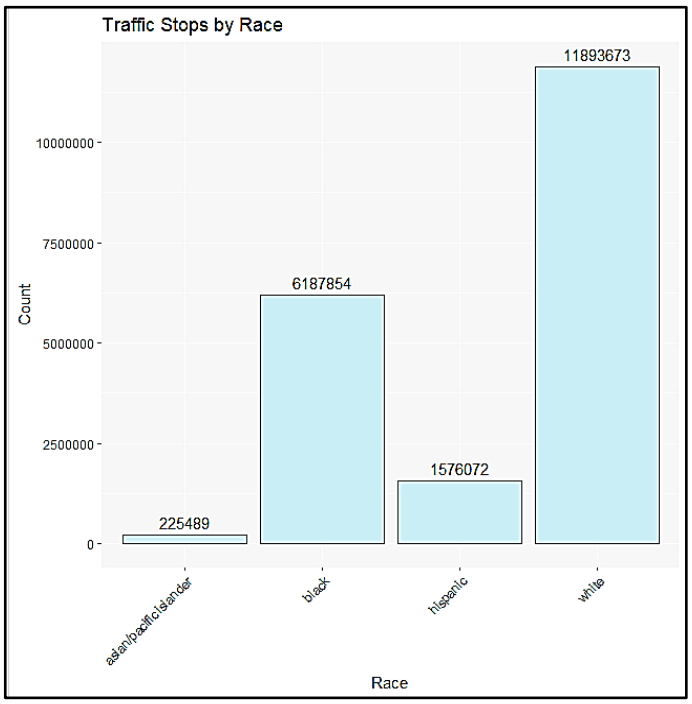
Create Tables of Frequencies: Develop frequency tables to summarize the distribution of each categorical variable in the dataset.

Cross-Tabulate Key Relationships: Create cross-tabulations to examine relationships between key categorical variables, such as "subject_race" and "type" of stop.

Visualize Distributions: Use histograms to display the age distribution ("subject_age") and bar plots to show the distribution of "subject_race."

Create Interactive Plots: Incorporate interactive visualizations to allow users to explore the data in more detail.

Visualizations:
Bar Chart of Traffic Stops by Race

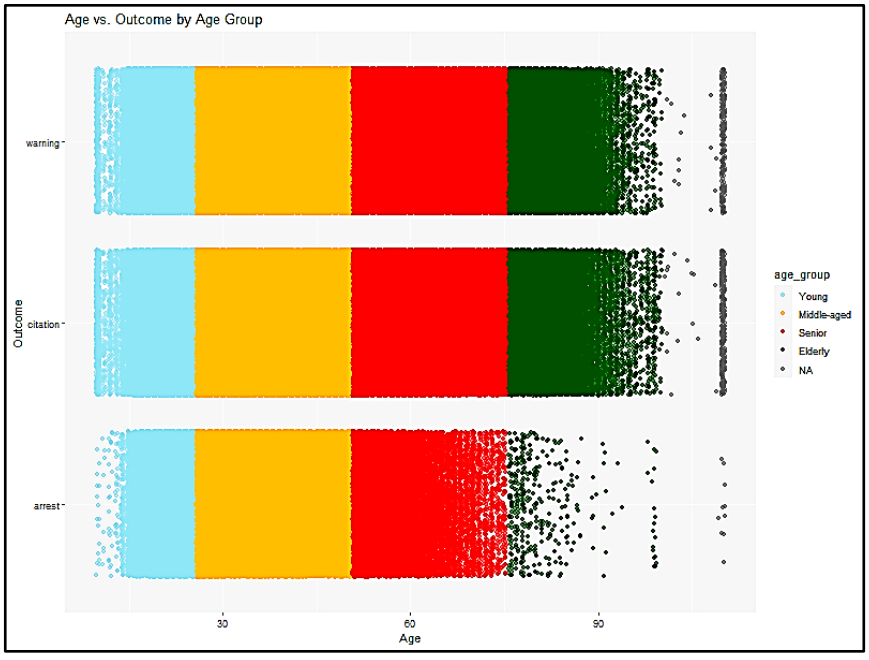


This overview highlights racial patterns in traffic stops. This bar chart displays traffic stops by racial group, showing that white individuals have the highest number of stops, followed by other groups, experienced the highest number of traffic stops, totaling 11,893,673, followed by Black individuals with 6,187,854 stops. Hispanic individuals and Asian/Pacific Islanders had the lowest count.

This suggests a significant variation in traffic stops by race, with white and Black individuals representing the largest shares

.

Scatter Plot: Age vs. Outcome



In the scatter plot, the correlation between age (along the bottom) and different types of outcomes (on the side), with outcomes like arrest, citation, warning, and NA (unknown). Different colors represent age groups: (0, 25, 50, 75, 100); Young (blue), Middle-aged (orange), Senior (red), Elderly (green), and NA (gray), respectively.

1. **Age and Outcomes:**

- Middle-aged and Senior individuals make up a big portion of all outcomes, with Seniors (red) which is clearly evident in arrests.
- Elderly individuals (green) are mostly seen in citation and warning outcomes, meaning they rarely face arrest.

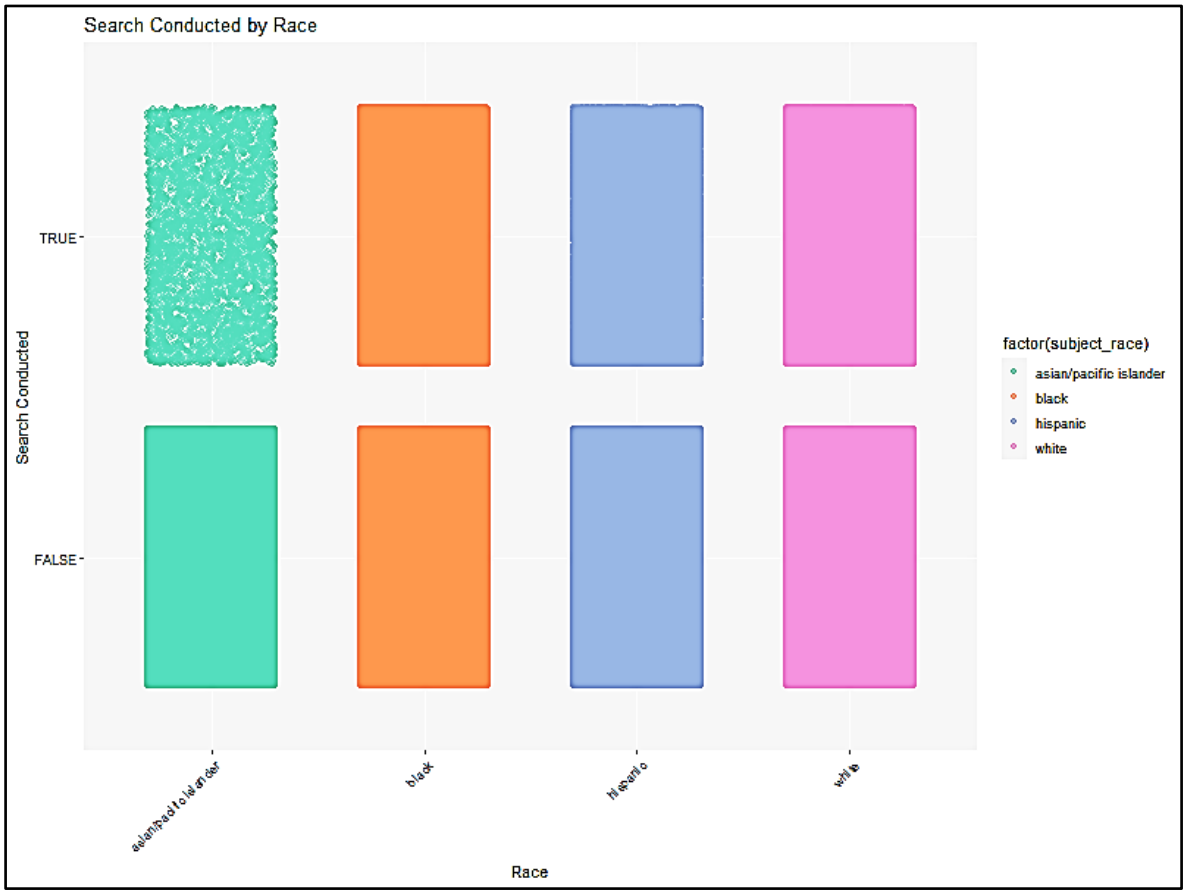
2. **Outcome Patterns by Age Group:**

- Arrests are more common for Seniors, while citations and warnings are spread across Middle-aged and Senior groups.
- Young people (blue) are present in each outcome, but not as much as other age groups.

3. **Older Individuals and Outcomes:**

- Elderly people mostly receive citations and warnings, showing they are less likely to be arrested.
- The NA category, has a lot of young and middle-aged individuals.

Jitter Plot: Search Conducted by Race



The jitter plot above displays whether or not a search was performed on various racial groupings during traffic stops. "True" represents stops where a search occurred, while "False" indicates no search. Jittering spreads the points slightly, minimizing overlap to make individual instances more visible.

Understanding the Jitter Plot:

1. **Search Volumes by Race Group:**
   - As evidenced by the lower number of points in the "TRUE" category, searches are less common across all racial groupings.
   - Searches are comparatively more scattered among Asian/Pacific Islander people than other ethnicities.
2. **Disparities in Searches:**
   - While searches are generally infrequent, there appear to be some differences between racial groups. Asian/Pacific Islander and Black individuals seem to experience searches at a higher rate compared to Hispanic, White, and Other racial groups.
   - This pattern suggests there may be disparities in search rates that could need further investigation.

Hypothesis Testing

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 = 21.766, df = 20284198, p-value < 0.00000000000000022
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 35.05973 35.07155
sample estimates:
mean of x
 35.06564
```

1) **t-statistic**: The number, **21.766**, shows the standard errors that separate the sample mean from the predicted mean.
2) **Degrees of Freedom (df)**: **20,284,198** the sample size -1
3) **p-value**: The p-value is a very modest value of **< 0.00000000000000022**. It indicates that there is a very statistically significant divergence between the sample mean and the predicted mean.
4) **95 percent confidence interval**: It is calculated that the real mean falls between **35.05973** and **35.07155**.
5) **Sample Mean (mean of x)**: The sample mean of **35.06564** is marginally different from the value of 35 that was hypothesized.

We **reject the null hypothesis** since the **p-value is significantly less than 0.05**.
There is enough information to conclude that the population's actual **mean age is not 35.**
Because a two-sided t-test considers both greater than and less than when determining if the sample mean differs substantially from the predicted mean, we used "two-side" while constructing the code.

Hypothesis:

The null hypothesis (H_0) states no correlation between the chance of being searched during a traffic stop and race (Black vs. white). This means Black drivers are not more likely than White drivers to be searched.

The alternative hypothesis (H_1) states that the chance of being searched during a traffic stop is correlated with race (Black vs. White). This means Black drivers are more likely than White drivers to be searched.

To run the hypothesis test we would require the “subject_race” and “subject_conducted” cross-tabulation.

```
> # Cross-tabulation of subject race and search conducted
> table(data$subject_race, data$search_conducted)

                          FALSE    TRUE
  asian/pacific islander  221669    3820
  black                  5909944  277910
  hispanic               1504205   71867
  white                 11629274  264399
  other                   157569    3330
  unknown                 238952    3688
```

We then create a contingency table for the observed frequencies because it organizes the observed frequencies into a clear format for comparison

`

	Yes	No
Black	277910	5909944
White	264399	11629274

By, manually calculating the proportion of which individuals were searched, Black individuals were searched at a higher rate of **4.49%** than White individuals at **2.22%**.

Chi-Square test:

The chi-square test is used to evaluate whether there is a significant association between **race** and whether a search was conducted.

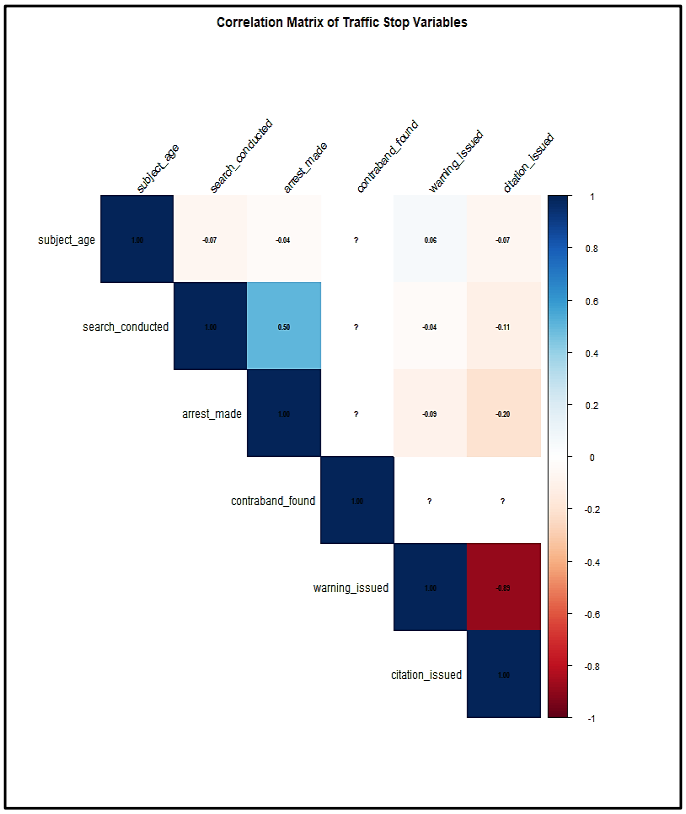
```
        Pearson's Chi-squared test with Yates' continuity correction

data:  observed
X-squared = 71976, df = 1, p-value < 0.00000000000000022
```

Since the **p-value is far less than 0.05**, we **reject the null hypothesis**.

This indicates that there is **a significant association between race and whether a search was conducted**. There is evidence to suggest that Black drivers are more likely to be searched than White drivers.

Correlation Matrix of Traffic Stop Variables:



***Strong Positive Correlation**:

- **Search Conducted & Arrest Made (0.50)**: Searches are moderately associated with arrests.

***Strong Negative Correlation**:

- **Warning Issued & Citation Issued (-0.89)**: When warnings are issued, citations are rarely issued (and vice versa).

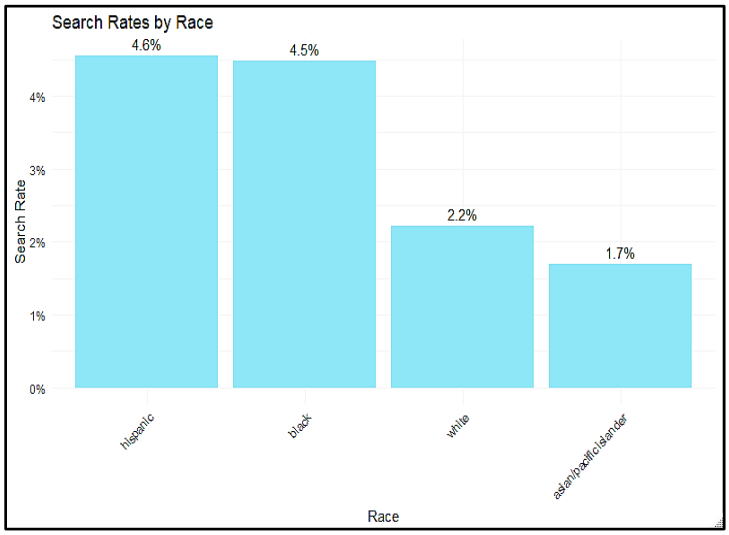
***Weak Relationships**:

- **Subject Age** has weak correlations with all other variables, meaning age has minimal impact on these outcomes.
- **Search Conducted** slightly decreases the likelihood of warnings (-0.04) or citations (-0.11).

***Missing Data**:

- **Contraband Found** correlations are missing with most variables, suggesting unavailable data for those relationships.

This bar chart displays search rates by race during traffic stops:



1. **Highest Search Rates**:
   - Hispanic individuals: **4.6%**
   - Black individuals: **4.5%**
2. **Lower Search Rates**:
   - White individuals: **2.2%**
   - Asian/Pacific Islander individuals: **1.7%**

Key Insight:

Hispanic and Black individuals have notably higher search rates compared to White and Asian/Pacific Islander individuals, suggesting potential disparities in search practices based on race.

Search Probability Model:

Logistic Regression Result

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

Coefficients:
                      Estimate Std. Error  z value             Pr(>|z|)
(Intercept)         -3.7503964  0.0170120 -220.456 < 0.0000000000000002 ***
subject_age         -0.0348915  0.0001182 -295.089 < 0.0000000000000002 ***
subject_raceblack    1.0593159  0.0165075   64.172 < 0.0000000000000002 ***
subject_racehispanic 0.8256799  0.0168375   49.038 < 0.0000000000000002 ***
subject_raceother    0.2204537  0.0240389    9.171 < 0.0000000000000002 ***
subject_raceunknown -0.1911082  0.0233657   -8.179 0.000000000000000286 ***
subject_racewhite    0.3274977  0.0165093   19.837 < 0.0000000000000002 ***
subject_sexmale      1.0645332  0.0034448  309.023 < 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: 5572941  on 20274393  degrees of freedom
Residual deviance: 5277413  on 20274386  degrees of freedom
AIC: 5277429

Number of Fisher Scoring iterations: 7
```

Goal: To predict if a search is conducted based on age, race, and sex.

Findings:

- Age: Older individuals are less likely to be searched (-0.0349).
- Race: Black (+1.06), Hispanic (+0.83), and White (+0.33) individuals are more likely to be searched; "Unknown" race is less likely (-0.19).
- Sex: Males are more likely to be searched (+1.06).

Significance: All predictors are highly significant ($p < 0.001$).

Model Fit: Reduced deviance (5572941 → 5277413) and AIC = 5277429.

From this, we can say that searches are more likely for males, Black, Hispanic, and White individuals, while older individuals and those with "unknown" race are less likely to be searched.

All variables significantly impact the likelihood of a search being conducted.

Conclusion

- **Search Rates by Race**:
  Black (4.5%) and Hispanic (4.6%) individuals experience higher search rates during traffic stops compared to White (2.2%) and Asian/Pacific Islander (1.7%) individuals.
- **Search and Arrest Link**:
  Traffic stops involving searches are more likely to lead to arrests, while warnings and citations are rarely issued together.
- **Demographic Influence on Searches**:
  Males, Black, and Hispanic individuals are searched more frequently, while older individuals and those with "unknown" race are searched less often.

Additionally, the correlation matrix highlights a moderate relationship between searches and arrests, suggesting that searches frequently result in legal repercussions. These results are corroborated by hypothesis testing, since the chi-square test shows a statistically significant correlation between search probability and race. A minor but statistically significant variance in the mean age of those stopped is revealed by the one-sample t-test, indicating that age may potentially influence the results of traffic stops.

Age, sex, and race all have a substantial impact on the likelihood that a search will be undertaken, according to the predictive search probability model. Older people and those with unclear racial identities are less likely to be searched than men and people who identify as Black or Hispanic.