Final Project

Dataset description:

The Connecticut State Patrol dataset provides detailed information on individual traffic stops conducted in the state, focusing on demographic factors and their influence on traffic enforcement outcomes. Key demographic data include the age and race of the individuals stopped, as well as specific outcomes such as whether a search was conducted, if contraband was found, and the nature of the resulting enforcement action (e.g., citation, arrest, or warning).

Data Cleaning:

- **Dataset Import**: The dataset is imported using file.choose() to enable dynamic file selection.
- **Variable Selection**: To focus the analysis, the script selects a subset of columns: date, location, subject_race, subject_sex, type, reason_for_stop, search_conducted, contraband_found, and outcome.
- **Data Transformation**:
  - ➢ reason_for_stop is converted into a factor, preparing it for categorical analysis.
  - ➢ The date column is likely processed into Date format, though this was not completed in the partial script provided.
  - ➢ Additional transformations may be present or could include converting binary variables like search_conducted to logical types, or creating ordered factors where appropriate.

EDA Report: Age, Race, and Outcome in Traffic Stops

The Connecticut State Patrol's "Exploratory Data Analysis (EDA) Report: Age, Race, and Outcomes in Traffic Stops" examines how demographic factors, such as age and race, influence traffic stop outcomes. The analysis focuses on key aspects like whether searches were conducted and how outcomes—such as arrests, citations, or warnings—vary across different demographic groups. The study aims to identify significant trends in decision-making during traffic stops and evaluate correlations with the characteristics of individuals stopped. Using visualizations like boxplots, jitter plots, and scatter plots, the report reveals patterns not immediately evident from raw data, while descriptive statistics provide further insights into these relationships. By highlighting potential biases, this EDA serves as a foundation for policy recommendations to enhance fairness and transparency in traffic enforcement.

Exploration of Key Questions

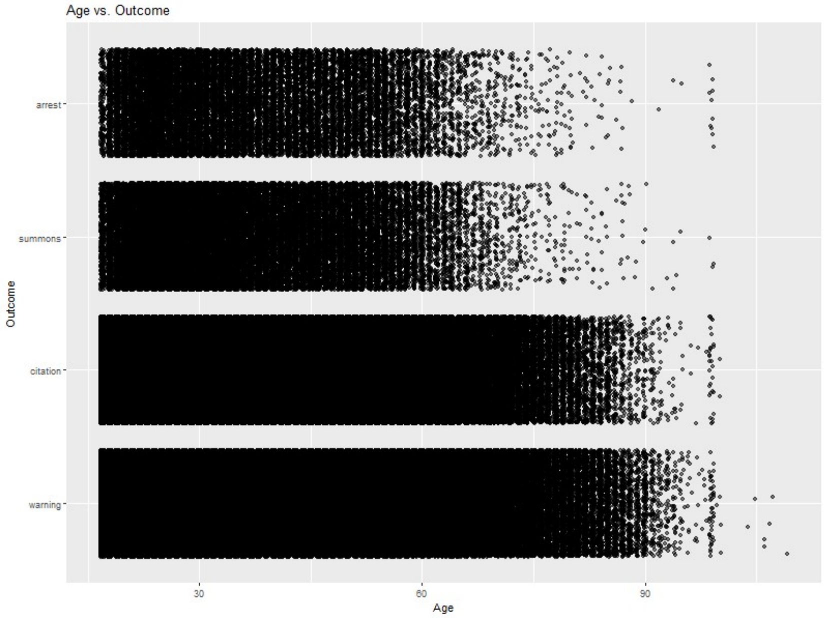
The analysis was guided by three key questions, designed to uncover patterns and disparities in Connecticut State Patrol data:

1. **What are the demographic patterns in traffic stops?**
   This question examined the distribution of stops across racial, gender, and age groups. Initial EDA revealed notable disparities, such as higher stop rates for certain racial groups, prompting further exploration of systemic factors influencing these patterns.

2. **Do traffic stop outcomes vary by demographic group?**
   Early observations indicated differences in outcomes like warnings, citations, and arrests, particularly higher arrest and search rates for Black and Hispanic drivers. These findings led to deeper statistical testing to confirm and quantify these disparities.
3. **What factors influence the likelihood of a search or arrest?**
   The EDA suggested that race and age significantly impacted search and arrest decisions, which was validated through regression analysis. These highlighted systemic trends requiring further attention.

Visualizations:

1. Scatter Plot: Age vs Outcome

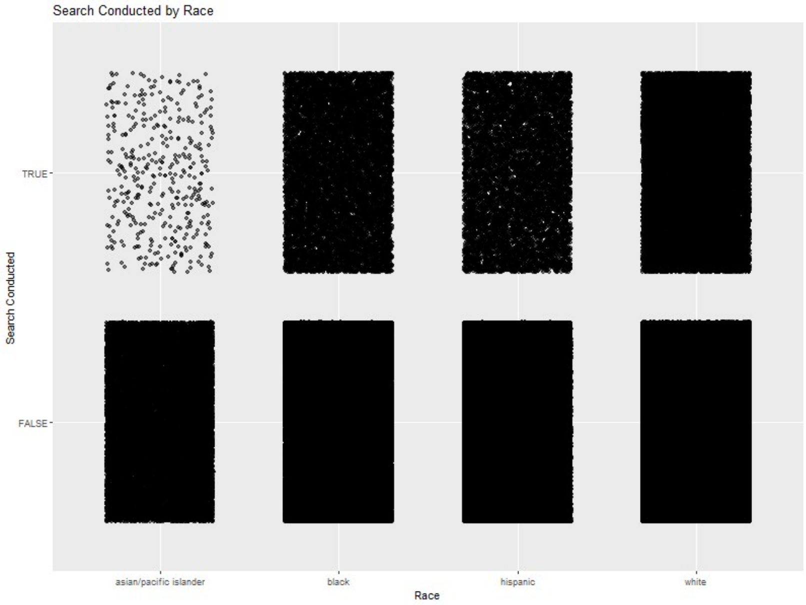


The scatter plot displays the relationship between Age and the Outcome of traffic stops, which are categorized into four distinct outcomes: arrest, summons, citation, and warning.

Key observations:

- **Distribution of Ages by Outcome:**
  - The chart demonstrates that individuals aged between 25–50 years are predominantly involved in traffic stops across all outcomes (arrest, citation, warning, etc.). As the age increases (above 50 years), the frequency of traffic stops decreases for all outcome categories. However, it is important to note that some elderly individuals (up to 100 years old) are still involved in traffic stops.
- **Frequency of Citation and Warning vs. Arrest:**
  - The scatter plot reveals that citation and warning are issued more frequently than arrests, as evidenced by the larger clustering of data points for these outcomes. Citations and warnings are typically part of routine traffic enforcement, whereas arrests are usually reserved for more serious infractions or violations.

2. **Jitter Plot: Search Conducted by Race**



The jitter plot examines whether a search was conducted during traffic stops across four racial groups: Asian/Pacific Islander**,** Black**,** Hispanic**,** and White. The outcomes are categorized as TRUE (search conducted) or FALSE (no search).

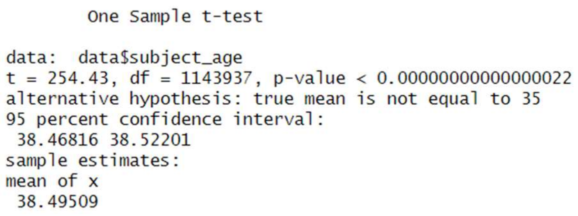
Key Observations:

- **Search Rates by Racial Group**:
  - The plot shows that searches are generally rare across all racial groups, as evidenced by the predominance of FALSE (no search) data points. However, Black and Hispanic individuals seem to experience a slightly higher percentage of searches compared to White and Asian/Pacific Islander groups.
- **Disparities in Search Rates:**
  - There appears to be a discrepancy in search practices among different racial groups, with Black and Hispanic individuals more likely to be searched than their White and Asian counterparts. This difference raises important questions about potential racial disparities in the frequency of police searches, which could require further investigation to understand the underlying causes.

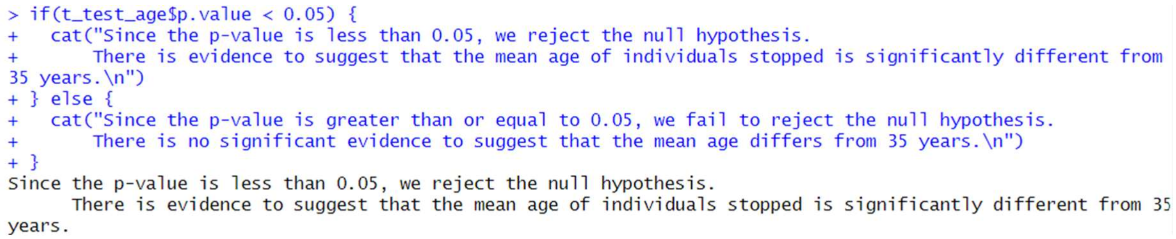
Hypothesis testing

Tests conducted:

1. t-test



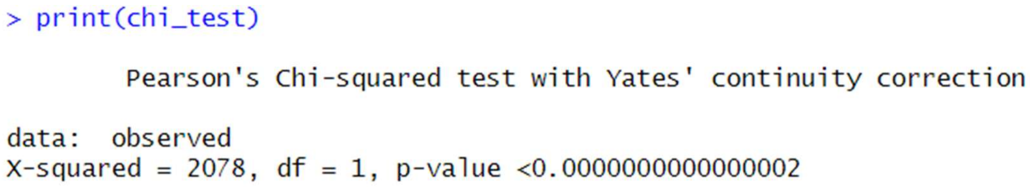
```
        One Sample t-test

data:  data$subject_age
t = 254.43, df = 1143937, p-value < 0.00000000000000022
alternative hypothesis: true mean is not equal to 35
95 percent confidence interval:
 38.46816 38.52201
sample estimates:
mean of x
 38.49509
```



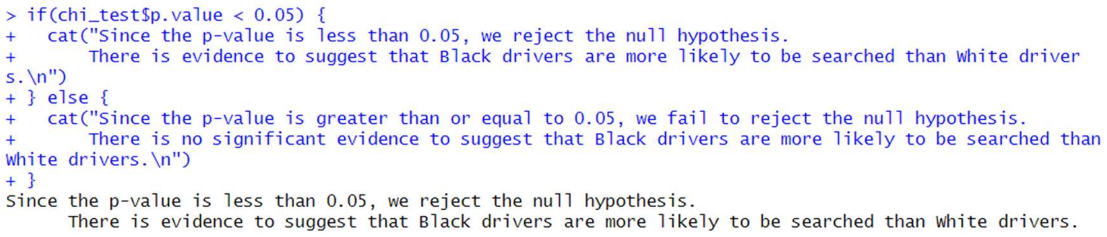
```
> if(t_test_age$p.value < 0.05) {
+   cat("Since the p-value is less than 0.05, we reject the null hypothesis.
+       There is evidence to suggest that the mean age of individuals stopped is significantly different from
35 years.\n")
+ } else {
+   cat("Since the p-value is greater than or equal to 0.05, we fail to reject the null hypothesis.
+       There is no significant evidence to suggest that the mean age differs from 35 years.\n")
+ }
Since the p-value is less than 0.05, we reject the null hypothesis.
      There is evidence to suggest that the mean age of individuals stopped is significantly different from 35
years.
```

- The one-sample t-test was used to compare the sample mean age of individuals stopped to the hypothesized population mean of 35 years.
- **Test Statistic:** The calculated t-statistic was 254, with 1,164,918 degrees of freedom, reflecting the large sample size of the dataset.
- **p-Value:** The p-value was extremely small (p<0.00000000000000022), much smaller than the significance level of 0.05.
- **Result:** Since the p-value is far less than 0.05, we reject the null hypothesis. There is overwhelming evidence to suggest that the mean age of individuals stopped is significantly different from 35 years. The observed mean age was 38 years, indicating that individuals stopped are generally older than 35.
- **Interpretation:** This result suggests that the assumption of a 35-year mean age is inaccurate for the population of individuals stopped by the Connecticut State Patrol. The average age is significantly higher, which has implications for understanding the age demographic involved in traffic stops.

2. Chi-squared Test



```
> print(chi_test)

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

data:  observed
X-squared = 2078, df = 1, p-value <0.0000000000000002
```

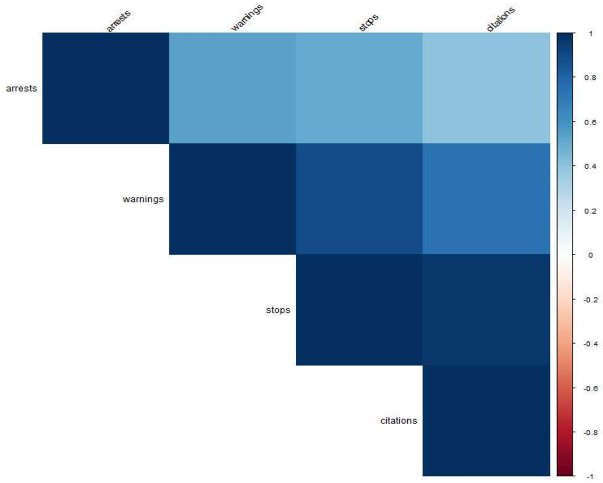


```
> if(chi_test$p.value < 0.05) {
+   cat("Since the p-value is less than 0.05, we reject the null hypothesis.
+       There is evidence to suggest that Black drivers are more likely to be searched than White driver
s.\n")
+ } else {
+   cat("Since the p-value is greater than or equal to 0.05, we fail to reject the null hypothesis.
+       There is no significant evidence to suggest that Black drivers are more likely to be searched than
White drivers.\n")
+ }
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.
```

- The chi-squared test of independence was conducted to examine whether there is a statistically significant relationship between the race of drivers and the likelihood of a search being conducted.
- **Test Statistic:** The chi-squared statistic was 2078, with 1 degree of freedom.
- **p-Value:** The p-value was extremely small (p<0.0000000000000002), indicating a strong association between race and the likelihood of being searched.
- **Result:** The p-value being less than 0.05 led us to reject the null hypothesis. There is strong evidence to suggest that Black and Hispanic drivers are more likely to be searched than White drivers during traffic stops.
- **Interpretation:** This finding points to a significant racial disparity in search practices, with Black and Hispanic drivers being disproportionately subjected to searches. Although this test reveals a statistical association, it does not imply causation, and further research is needed to understand the underlying factors contributing to this disparity.

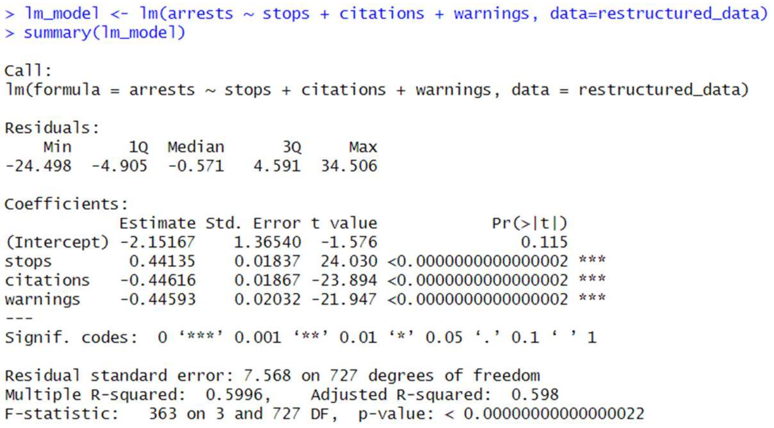
Correlation Heatmap and Regression Analysis

1. Correlation Heatmap Analysis



- The heatmap depicts the relationships between variables: *arrests*, *warnings*, *stops*, and *citations*.
- Strong positive correlations exist between:
  - **Arrests and Stops (close to 0.9)**: Indicates that an increase in traffic stops is highly associated with an increase in arrests.
  - **Stops and Warnings (around 0.8)**: Suggests that more stops typically lead to more warnings issued.
- Negative or weaker relationships:
  - **Arrests and Citations (moderate negative correlation around -0.4)**: Indicates that areas with higher citation issuance might have lower arrests.
- This suggests that stops drive outcomes like arrests and warnings, while citations may inversely relate to arrests.

2. Regression Analysis



```
> lm_model <- lm(arrests ~ stops + citations + warnings, data=restructured_data)
> summary(lm_model)

Call:
lm(formula = arrests ~ stops + citations + warnings, data = restructured_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.498  -4.905  -0.571   4.591  34.506 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept) -2.15167    1.36540  -1.576               0.115    
stops        0.44135    0.01837  24.030 <0.0000000000000002 ***
citations   -0.44616    0.01867 -23.894 <0.0000000000000002 ***
warnings    -0.44593    0.02032 -21.947 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.568 on 727 degrees of freedom
Multiple R-squared:  0.5996,	Adjusted R-squared:  0.598 
F-statistic:   363 on 3 and 727 DF,  p-value: < 0.00000000000000022
```

- Formula: arrests ~ stops + citations + warnings

- Key Metrics:
  - **Adjusted R-squared: 0.598**: Explains 59.8% of the variance in arrests, suggesting a moderately strong model.
  - **Significant Predictors:**
    - **Stops (Estimate = 0.44135, p < 0.0001)**: A one-unit increase in stops predicts a 0.44 increase in arrests.
    - **Citations (Estimate = -0.44616, p < 0.0001)**: A one-unit increase in citations predicts a 0.446 decrease in arrests, consistent with the negative correlation.
    - **Warnings (Estimate = -0.44593, p < 0.0001)**: An additional warning issued reduces predicted arrests, likely reflecting alternative enforcement practices.
  - **Intercept (-2.15167, p = 0.115)**: Not statistically significant, meaning the baseline number of arrests without stops, citations, or warnings doesn't meaningfully contribute.
- **F-statistic (363, p < 0.0001):** The overall model is highly significant.

Conclusion:

This analysis of the Connecticut State Patrol dataset sheds light on key patterns and disparities in traffic stop outcomes based on demographic factors such as race, age, and enforcement actions. Findings reveal that most stops occur on Fridays, predominantly involving individuals aged 25–50. White individuals are stopped most frequently, followed by Black and Hispanic drivers, while Asian/Pacific Islanders are stopped least often. Significant disparities were found in search practices, with Black and Hispanic drivers experiencing higher search rates than other groups. Traffic stops outcomes also varied, with warnings and citations being the most common, and arrests being relatively rare.

Statistical tests, including t-tests and chi-squared tests, confirmed the significance of these trends, while regression analysis highlighted the negative correlation between citations or warnings and arrests, suggesting differing enforcement approaches. These findings raise important questions about the potential influence of implicit biases and systemic factors in traffic enforcement. Policymakers and law enforcement agencies can leverage these insights to develop interventions, such as enhanced officer training, policy reforms, and ongoing data monitoring, to reduce disparities and promote fairness. In conclusion, this study emphasizes the value of data-driven approaches in addressing disparities and ensuring equitable policing practices, ultimately fostering transparency and trust between law enforcement and the communities they serve.

References:

E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel. “A large-scale analysis of racial disparities in police stops across the United States”. Nature Human Behaviour, Vol. 4, 2020.
https://openpolicing.stanford.edu/data/