[BoldFont=latinmodern-math.otf]

Statistical Insights on Open Policing Data

Reshmii Bondili, Srujana Chintala, Murat Al, Sakshi Hegde

2024-12-10

Introduction

The study of policing practices through statistical analysis provides crucial insights into patterns of law enforcement, social disparities, and systemic biases within society. Open policing data, which is made publicly available by the Stanford Open Policing Project — a unique partnership between the Stanford Computational Journalism Lab and the Stanford Computational Policy Lab, serves as a rich resource for exploring these issues. By leveraging this data, we can identify trends and correlations related to factors such as race, gender, location, and the outcomes of police interactions, including citations, warnings, and arrests. This project focuses on analyzing open policing data to uncover statistical insights into how policing practices may differ based on race and sex and to assess the presence of potential biases in law enforcement practices across different cities. Through the application of descriptive and inferential statistical methods, this analysis aims to identify significant patterns in police activity and generate evidence that can inform policy decisions aimed at promoting fairness and equity in policing.

Project Overview

The primary goal of this project is to explore potential biases related to race and sex. Specifically, we aim to examine whether biases exist based on race (subject_race) and sex (subject_sex), and how these biases may influence the likelihood of receiving citations, warnings, or arrests, as well as the time period in which these events occur. To achieve this, we will begin with data preprocessing, which includes handling missing values. Our analysis will focus on identifying key features related to race and sex, using descriptive statistics to gain insights. Additionally, inferential statistics will be applied to draw conclusions and make inferences from the data, first by analyzing each city individually, and then performing a cross-city comparison.

Finding and Cleaning Data

The datasets from cities like San Diego, Charlotte, and Nashville vary in size, and they predominantly consist of categorical variables, with age being the only numerical feature. These specific features were selected because they are common across all three city datasets, and each dataset provides a sufficiently large sample size for analysis. However, there were a few challenges that we encountered and overcame, including dealing with missing values (NaN), duplicates, and inconsistencies where the outcomes of features were similar but labeled differently. In our analysis of individual cities, we focused on features with fewer missing values. For the cross-city analysis, we concentrated on key and comparable features, such as time period, citations and warnings issued, subject sex (subject_sex), subject race (subject_race), subject age (subject_age), and arrests made (arrest_made). For more information on the dataset, please refer to the following link: Visit the Open Policing website for the data files and more information.

Descriptive Analysis

Descriptive statistics focuses on the summarization and representation of the principal characteristics of a dataset. It utilizes numerical measures, graphical formats, and tables to offer a clear perspective on the data. It aims to visualize and articulate the data collected. When examining traffic open policing data, descriptive statistics can be instrumental in answering questions regarding the distribution of traffic violations among different racial and gender demographics. This examination of traffic open policing data addresses several pertinent questions related to demographic distributions. It seeks to understand the distribution of traffic violations among diverse racial and ethnic groups. Additionally, it analyzes how these violations are distributed by gender within each racial category. The research further identifies the predominant types of traffic violations for each racial and gender demographic. Moreover, it investigates the timing and locations of traffic violations, focusing on when (time of day, day of the week) and where (specific areas) these incidents are most prevalent among different racial and gender groups. The average age of drivers involved in these violations is also considered across various demographics. Finally, a comparative analysis is performed to assess the differences in traffic violation rates and the outcomes of traffic stops, including warnings, citations, and arrests, across different racial and gender groups.

```
#Adding libraries:
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

1. Absolute and Relative Frequency Table for Arrests Across the 3 Cities

```r
suppressWarnings(suppressPackageStartupMessages(library(dplyr)))
suppressWarnings(suppressPackageStartupMessages(library(ggplot2)))

#loading the dataset in csv format
data <- read.csv("cities.csv")

data_filtered <- data %>%filter(!is.na(arrest_made))

arrest_summary <- data_filtered %>%
group_by(cityname) %>%
summarise(
  Number_of_Arrests = sum(arrest_made == TRUE),
  Total_Interactions = n()
) %>%
mutate(
  Relative_Frequency = (Number_of_Arrests / Total_Interactions) * 100
)%>%
rename(
  "City" = cityname,
  "Number of Arrests" = Number_of_Arrests,
  "Relative Frequency (%)" = Relative_Frequency
)

print(arrest_summary)
```

```
## # A tibble: 3 x 4
##   City      `Number of Arrests` Total_Interactions `Relative Frequency (%)`
##   <chr>                   <int>              <int>                     <dbl>
## 1 Nashville               11013             723753                      1.52
## 2 SD                       3054             225750                      1.35
## 3 charlotte                4997             246219                      2.03
```

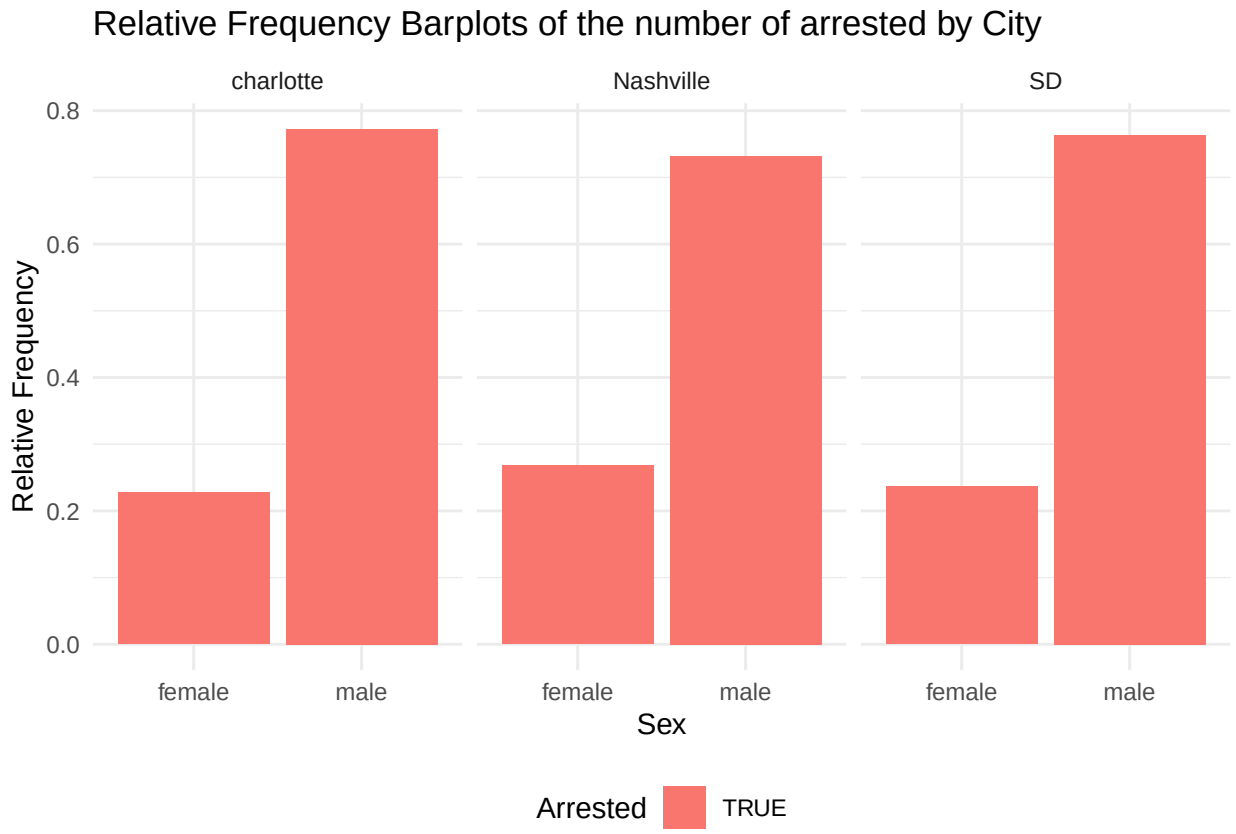
The analysis shows that Nashville recorded the highest number of arrests (11,013) but had a lower relative arrest frequency (1.52%) compared to Charlotte, which had fewer arrests (4,997) but the highest relative frequency (2.03%). San Diego had the fewest arrests (3,054) and the lowest relative frequency (1.35%). These differences highlight variations in arrest practices or interaction contexts across the cities.

2. Relative Frequency Barplots of the number of Males and Females arrested by City

```r
#Sample data
data_bar <- data.frame(
  City = data$cityname[data$arrest_made == "TRUE"],
  Sex = data$subject_sex[data$arrest_made == "TRUE"],
  Arrested = data$arrest_made[data$arrest_made == "TRUE"]
)

data_bar <- na.omit(data_bar)
```

```r
# Calculate relative frequencies
rel_freq <- prop.table(table(data_bar), margin = 1)
rel_freq_df <- as.data.frame(rel_freq)
colnames(rel_freq_df) <- c("City", "Sex", "Arrested", "RelativeFrequency")
```

```r
# Create the barplot
ggplot(rel_freq_df, aes(x = Sex, y = RelativeFrequency, fill = Arrested)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ City) +
  labs(title = "Relative Frequency Barplots of the number of arrested by City", x = "Sex", y = "Relativ
  theme_minimal() +
  theme(legend.position = "bottom")
```



From the graph, it's evident that the rate of male arrests is similar in all three cities. This suggests that there's a common pattern of male involvement in incidents that result in arrests across different areas. Factors like law enforcement approaches, social behaviors, or how people are distributed might be contributing to this consistency.

The almost uniform height of the female bars shows that the arrest rates for women are largely consistent across the cities. Any small differences might be linked to local factors or policies that affect women, but overall, the rates remain pretty steady. This could also suggest that the social conditions or enforcement practices for female incidents are quite similar.

The charts reveal that males consistently have higher bars than females, indicating they face more arrests. This could lead to some interesting questions about the gender-related factors or behaviors that might be influencing these arrest rates.

Summaries of center (mean, median, mode, trimmed mean) and dispersion (IQR, variance, coefficient of variation, skewness)

3. Age Distribution across the three different cities and the measures of centre and dispersion plotted on Histogram

```r
#Load required libraries

#Convert 'subject_age' to numeric
df_clean <- data
df_clean$subject_age <- as.numeric(df_clean$subject_age)

#Remove rows with missing or NA values in 'subject_age' column
df_clean <- df_clean %>% filter(!is.na(subject_age))

#Function to calculate mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

#List of unique cities in the dataset
city_names <- unique(df_clean$cityname)

#Prepare a data frame with statistics for each city
summary_stats <- df_clean %>%
  group_by(cityname) %>%
  dplyr::summarize(
    mean_age = mean(subject_age, na.rm = TRUE),
    median_age = median(subject_age, na.rm = TRUE),
    mode_age = get_mode(subject_age),
    trimmed_mean_age = mean(subject_age, trim = 0.1, na.rm = TRUE),
    IQR_value = IQR(subject_age, na.rm = TRUE),
    variance_value = var(subject_age, na.rm = TRUE),
    cv_value = (sd(subject_age, na.rm = TRUE) / mean(subject_age, na.rm = TRUE)) * 100,
    skewness_value = skew(subject_age, na.rm = TRUE),
    .groups = "drop"  # Avoid grouped data from being carried forward
  )

#Print the statistics for each city separately in the console
for (city in city_names) {
  city_stats <- summary_stats %>% filter(cityname == city)

  cat("\n--- Statistics for", city, "---\n")
  cat("Mean: ", round(city_stats$mean_age, 2), "\n")
  cat("Median: ", round(city_stats$median_age, 2), "\n")
  cat("Mode: ", city_stats$mode_age, "\n")
  cat("Trimmed Mean: ", round(city_stats$trimmed_mean_age, 2), "\n")
  cat("\n")
}
```

```
##
## --- Statistics for SD ---
## Mean:  36.95
## Median:  34
## Mode:  25
## Trimmed Mean:  35.64
##
##
```

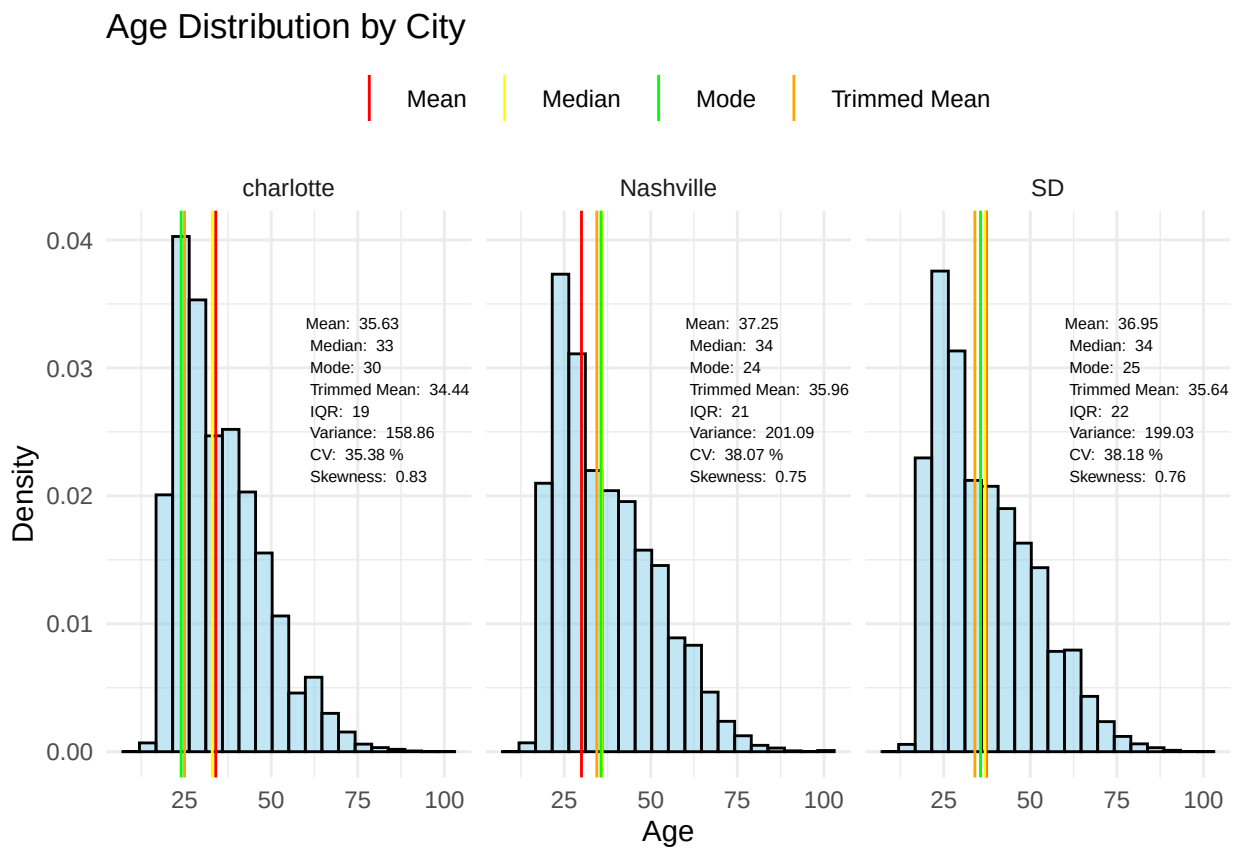
```
## --- Statistics for charlotte ---
## Mean:  35.63
## Median:  33
## Mode:  30
## Trimmed Mean:  34.44
##
##
## --- Statistics for Nashville ---
## Mean:  37.25
## Median:  34
## Mode:  24
## Trimmed Mean:  35.96
```

```r
#Prepare the data frame for adding lines (mean, median, mode, trimmed mean) to the legend
line_data <- data.frame(
  cityname = rep(city_names, each = 4),  # Repeat each city for 4 statistics
  line_type = rep(c("Mean", "Median", "Mode", "Trimmed Mean"), times = length(city_names)),
  xintercept = c(
    summary_stats$mean_age,
    summary_stats$median_age,
    summary_stats$mode_age,
    summary_stats$trimmed_mean_age
  ),
  color = c("red", "yellow", "green", "orange")
)

#Function to add text labels for statistics
add_labels <- function(city) {
  stats <- summary_stats %>% filter(cityname == city)
  paste(
    "Mean: ", round(stats$mean_age, 2), "\n",
    "Median: ", round(stats$median_age, 2), "\n",
    "Mode: ", stats$mode_age, "\n",
    "Trimmed Mean: ", round(stats$trimmed_mean_age, 2), "\n",
    "IQR: ", round(stats$IQR_value, 2), "\n",
    "Variance: ", round(stats$variance_value, 2), "\n",
    "CV: ", round(stats$cv_value, 2), "%\n",
    "Skewness: ", round(stats$skewness_value, 2)
  )
}

#Plot histograms for each city with relative frequencies, annotations, and lines for mean, median, mode
ggplot(df_clean, aes(x = subject_age)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "skyblue", color = "black", alpha = 0.5) +
  facet_wrap(~ cityname) +  # Separate plots by city
  theme_minimal() +
  scale_y_continuous(labels = scales::comma) +  # Prevent scientific notation on y-axis
  labs(title = "Age Distribution by City", x = "Age", y = "Density") +

#Add vertical lines for Mean, Median, Mode, Trimmed Mean
  geom_vline(data = line_data, aes(xintercept = xintercept, color = line_type), linetype = "solid", siz

#Add labels for the statistics
  geom_text(data = summary_stats, aes(x = 60, y = 0.0275, label = mapply(add_labels, cityname)),
```

```
            inherit.aes = FALSE, size = 2, hjust = 0) +

#Customizing the legend
  scale_color_manual(values = c("Mean" = "red", "Median" = "yellow", "Mode" = "green", "Trimmed Mean" =
  theme(legend.title = element_blank(),  # Remove legend title
        legend.position = "top")  # Position the legend at the top
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



SD and Nashville have a younger age concentration (mode around 24-25), but the mean age is slightly higher in Nashville (37.25) than SD (36.95), suggesting that there is a significant population of older individuals in Nashville as well. Charlotte has a mode of 30, indicating that this city has a more middle-aged demographic compared to SD and Nashville. Its mean age is also the lowest (35.63), suggesting a younger population in general.

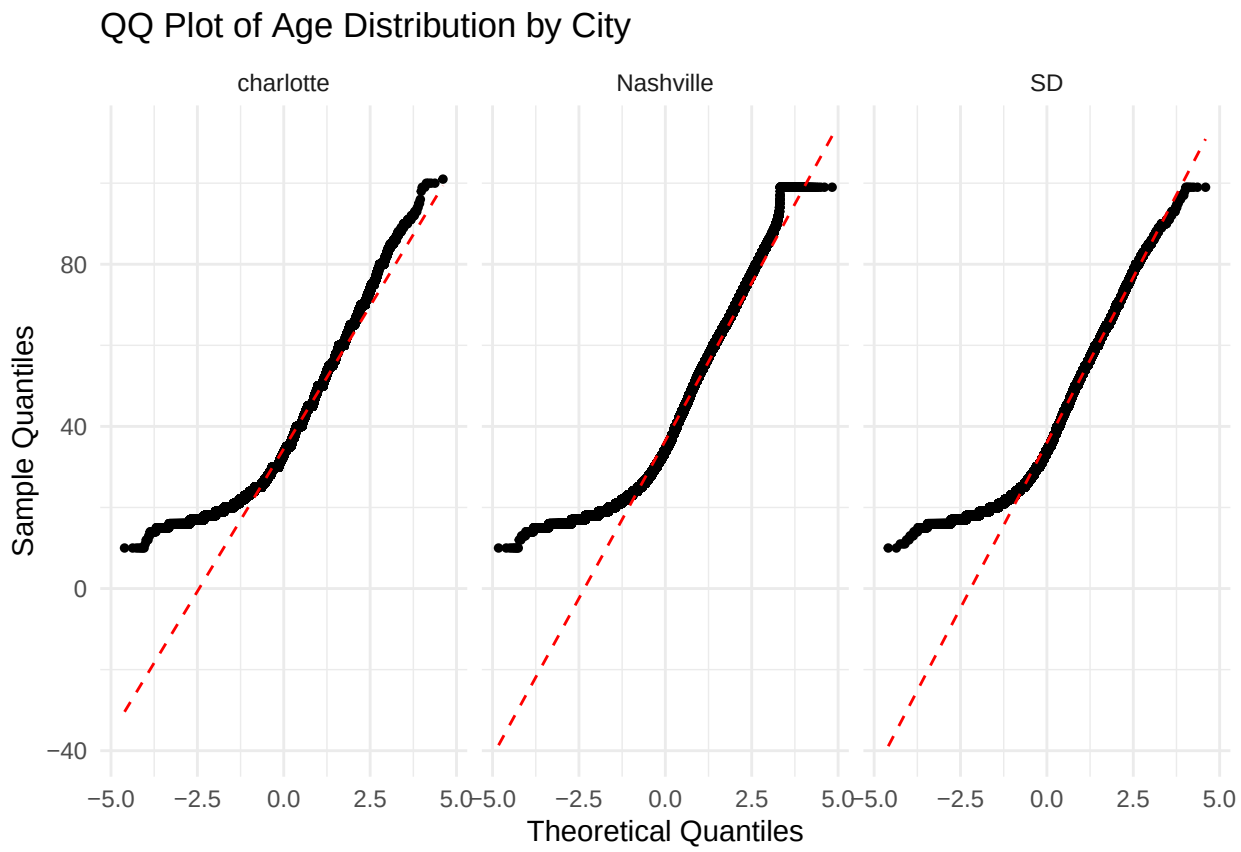
Each of the three cities has a positive skew, which means their age distributions are a bit right-leaning, with a longer tail for older ages. This tells us that even though younger people are more numerous, there's still a solid group of older individuals. The coefficient of variation is quite similar, ranging from 35.38% in Charlotte to 38.18% in SD, showing comparable age variability. SD and Nashville have a higher variance, around 200, suggesting a broader age distribution, while Charlotte's variance is lower at about 158, indicating a more concentrated age range.

4. Quantile Plots:

```
# QQ plot for each city
ggplot(data, aes(sample = subject_age)) +
  geom_qq(color = "black", size = 1) +  # QQ plot points
  geom_qq_line(color = "red", linetype = "dashed") +  # QQ line for normal distribution comparison
  facet_wrap(~ cityname) +  # Facet by city
  labs(title = "QQ Plot of Age Distribution by City", x = "Theoretical Quantiles", y = "Sample Quantiles
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 6400 rows containing non-finite outside the scale range
## (`stat_qq()`).
```

```
## Warning: Removed 6400 rows containing non-finite outside the scale range
## (`stat_qq_line()`).
```



QQ plots provide insight into the normality of the age distribution in each city. As observed, the lower tail of the distribution deviates from the normal line, which suggests a right-skewed distribution. In this case,

the peak of the distribution is shifted towards the lower values (on the left side), with more data points concentrated in the lower half of the age range. This deviation indicates that the data has a tail on the right, typical of a right-skewed distribution, where the majority of the observations are clustered at the lower end, and there are a few higher-value outliers.

5. Box Plot For Age vs Race Across Cities

```r
#library(ggplot2)
#library(dplyr)

cleaned_data <- data %>%
  filter(
    !is.na(subject_age) &
    !is.na(subject_race) &
    !is.na(cityname) &
    subject_race != "unknown"
  )
ggplot(cleaned_data, aes(x = subject_race, y = subject_age, fill = cityname)) +
  geom_boxplot() +
  labs(
    title = "Age Distribution by Race Across Cities",
    x = "Race",
    y = "Age",
    fill = "City"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```

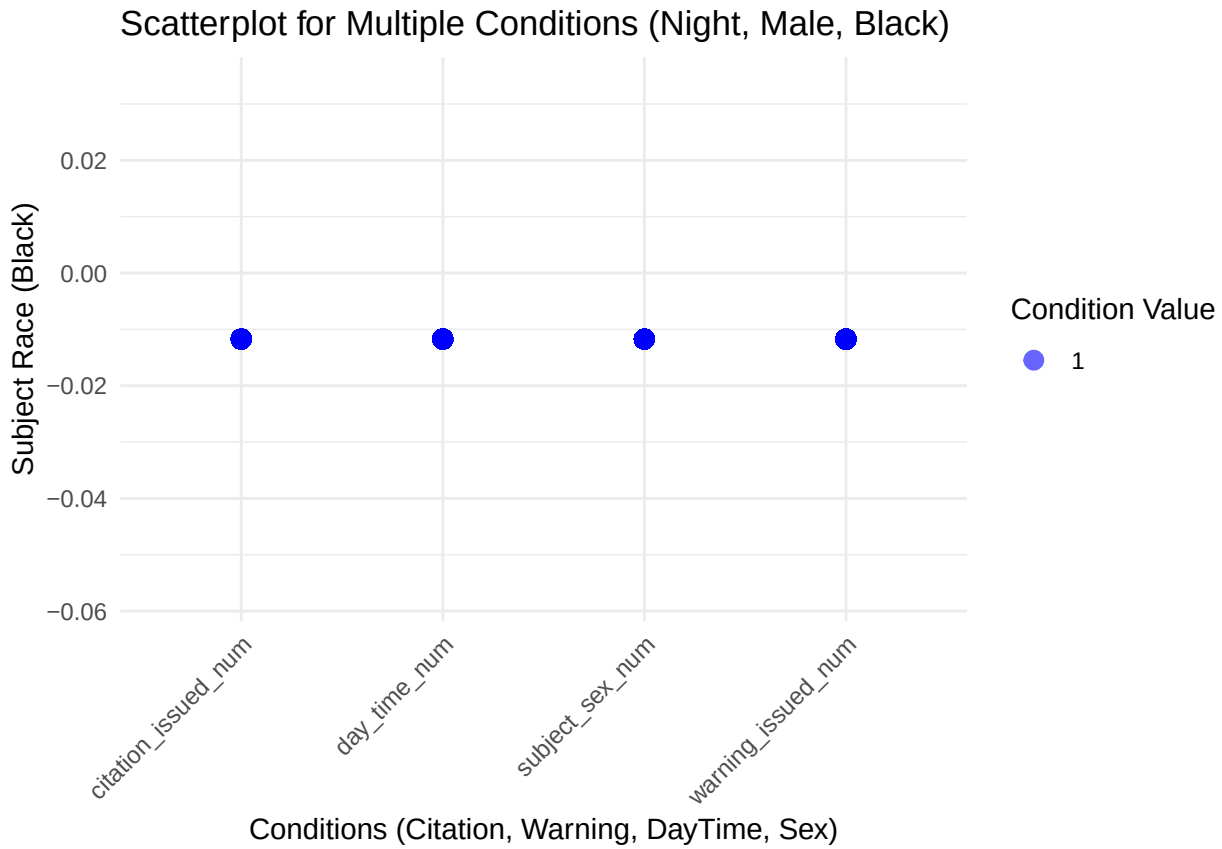
For most racial groups, the median age of police interactions is relatively consistent, hovering around the mid-30s. However, there is greater variability in the age distribution for certain groups, such as "White" and "Black" individuals, as indicated by their wider interquartile ranges and a greater number of outliers, particularly in Nashville and San Diego."Other" and "Hispanic" groups show more compact distributions, suggesting fewer extremes in the ages of individuals involved in police interactions. These differences may reflect varying population demographics or differing patterns of police engagement across racial groups in the cities.

6. Scatterplot(s), including correlation

We want to observe the correlation between multiple variable for instance day_period = Night, subject_race = 'black' and subject_sex = 'male' across all the cities since they were prominent features in our descriptive analysis which gave us insight on biases among black male race in night time.

```
library(ggplot2)
library(tidyr)

# Filter the dataset based on the condition subject_race == "black"
df_filtered <- subset(data, subject_race == "black" &
                            citation_issued == TRUE &
                            warning_issued == TRUE &
                            day_period == "Night" &
                            subject_sex == "male")

# Convert the conditions to numeric (for better plotting)
df_filtered$citation_issued_num <- as.numeric(df_filtered$citation_issued)
df_filtered$warning_issued_num <- as.numeric(df_filtered$warning_issued)
df_filtered$day_time_num <- as.numeric(df_filtered$day_period == "Night")
```

```r
df_filtered$subject_sex_num <- as.numeric(df_filtered$subject_sex == "male")

# Reshape the data into a long format for easier plotting
df_long <- df_filtered %>%
  pivot_longer(cols = c(citation_issued_num, warning_issued_num, day_time_num, subject_sex_num),
               names_to = "Condition",
               values_to = "Value")

# Create the scatter plot with ggplot
ggplot(df_long, aes(x = Condition, y = jitter(0), color = as.factor(Value))) +
  geom_point(size = 3, alpha = 0.6) +
  scale_color_manual(values = c("0" = "red", "1" = "blue")) +  # Color based on TRUE/FALSE
  labs(
    title = "Scatterplot for Multiple Conditions (Night, Male, Black)",
    x = "Conditions (Citation, Warning, DayTime, Sex)",
    y = "Subject Race (Black)",
    color = "Condition Value"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for clarity
```



Through the observation, it is evident that since we are dealing with categorical variables and only one numerical variable is present in our dataset which is the age we believe scatter-plot is not a good indicator to get the inference of the correlation among these features so that's why we used the box plot to analyse the same.

Histogram(s) with the appropriate number of bins and vertical lines corresponding to the mean and median

Quantile plot(s);

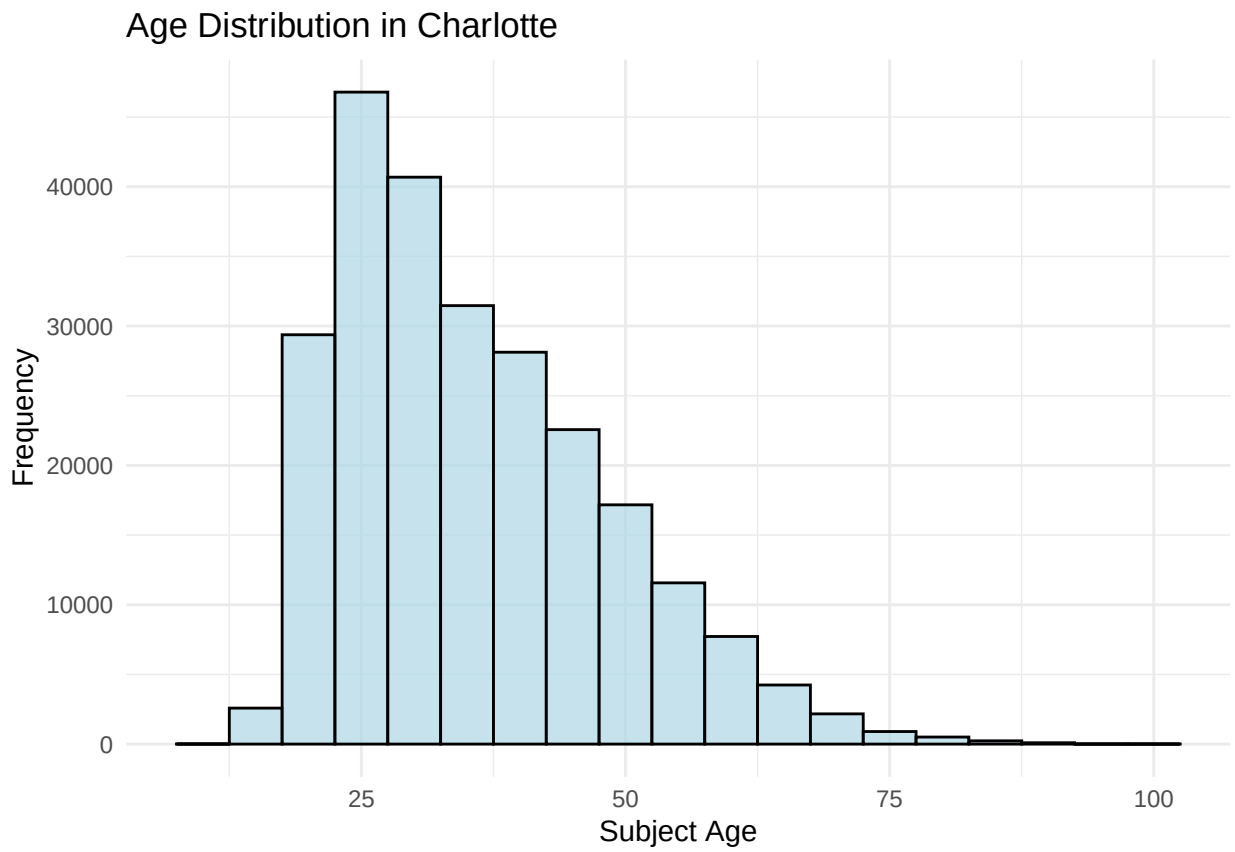
Box-Cox transformation(s) (only if certain data looks non-normal)

7. The box-cox transformation of age distribution across the cities

We are using box- cox transformation to observe that is there any imbalance across the age distribution and if the same is observed then we can transformed the orginal age distribution to a normalised one.

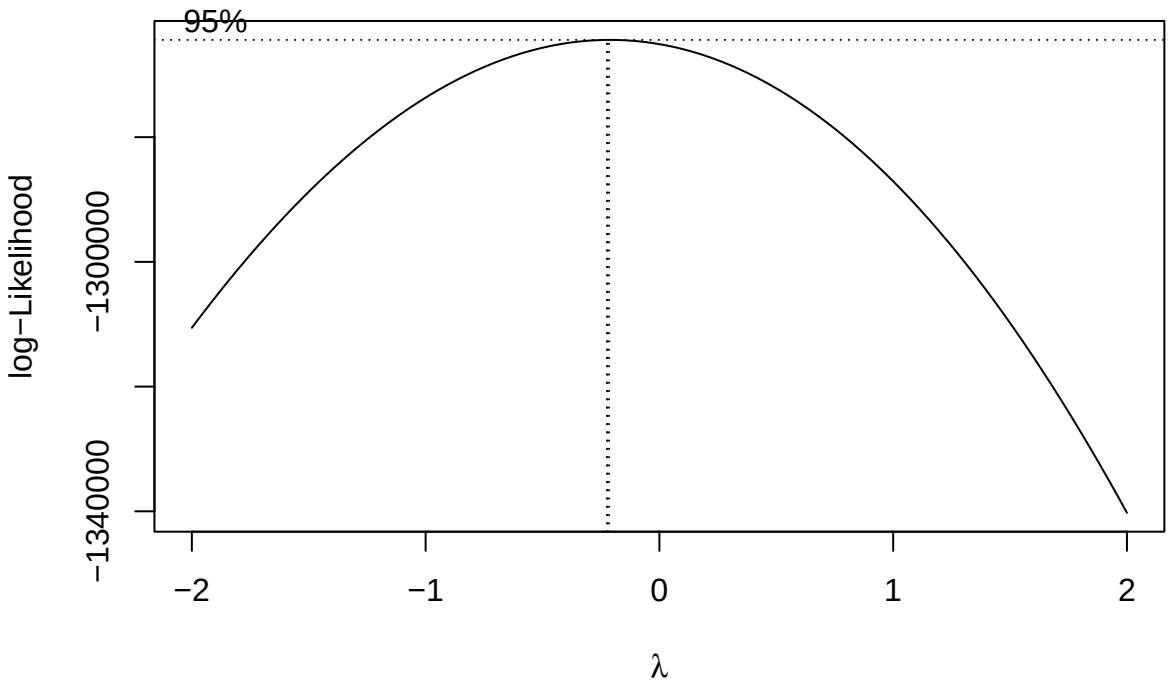
Box cox transformation for Charlotte city

```
# Load necessary libraries
# library(ggplot2)
library(MASS)  # For Box-Cox transformation
# library(dplyr)

# Filter the data for a particular city (e.g., "Charlotte")
charlotte_data <- data %>% filter(cityname == "charlotte", !is.na(subject_age))

# Plot the histogram for 'subject_age' in Charlotte
ggplot(charlotte_data, aes(x = subject_age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Age Distribution in Charlotte",
       x = "Subject Age", y = "Frequency") +
  theme_minimal()
```



Now applying the box cox transformation

```r
# Apply Box-Cox transformation
# Ensure that the subject_age has no zero or negative values
# If necessary, add a constant to make the values positive
charlotte_data$subject_age_transformed <- charlotte_data$subject_age

# Box-Cox transformation (choose the best lambda automatically)
boxcox_result <- boxcox(subject_age_transformed ~ 1, data = charlotte_data, lambda = seq(-2, 2, by = 0.
```



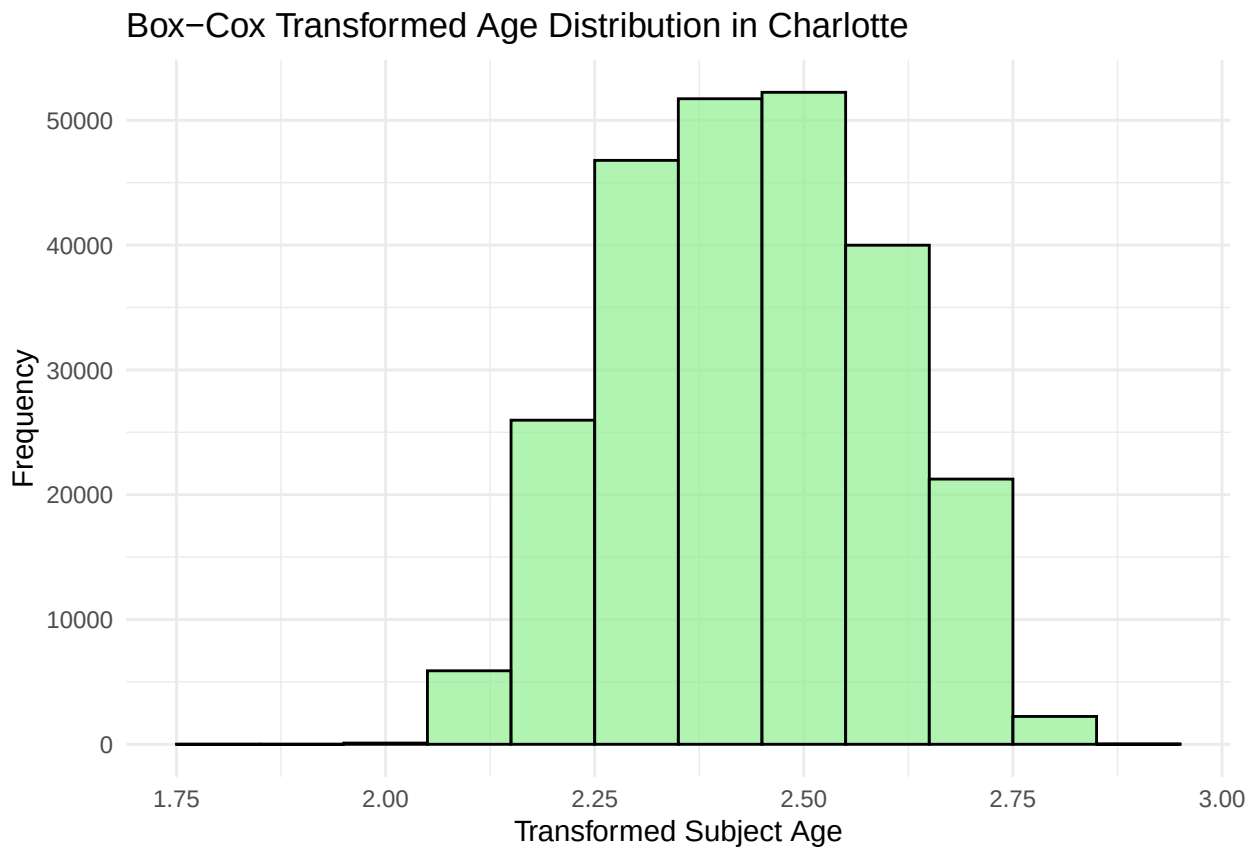

```r
# The lambda value that maximizes the log-likelihood
best_lambda_c <- boxcox_result$x[which.max(boxcox_result$y)]
cat("Optimal Lambda for Box-Cox transformation: ", best_lambda_c, "\n")
```

```
## Optimal Lambda for Box-Cox transformation:  -0.2222222
```

```r
# Apply the transformation with the best lambda
charlotte_data$subject_age_boxcox <- (charlotte_data$subject_age_transformed^best_lambda_c - 1) / best_l
```

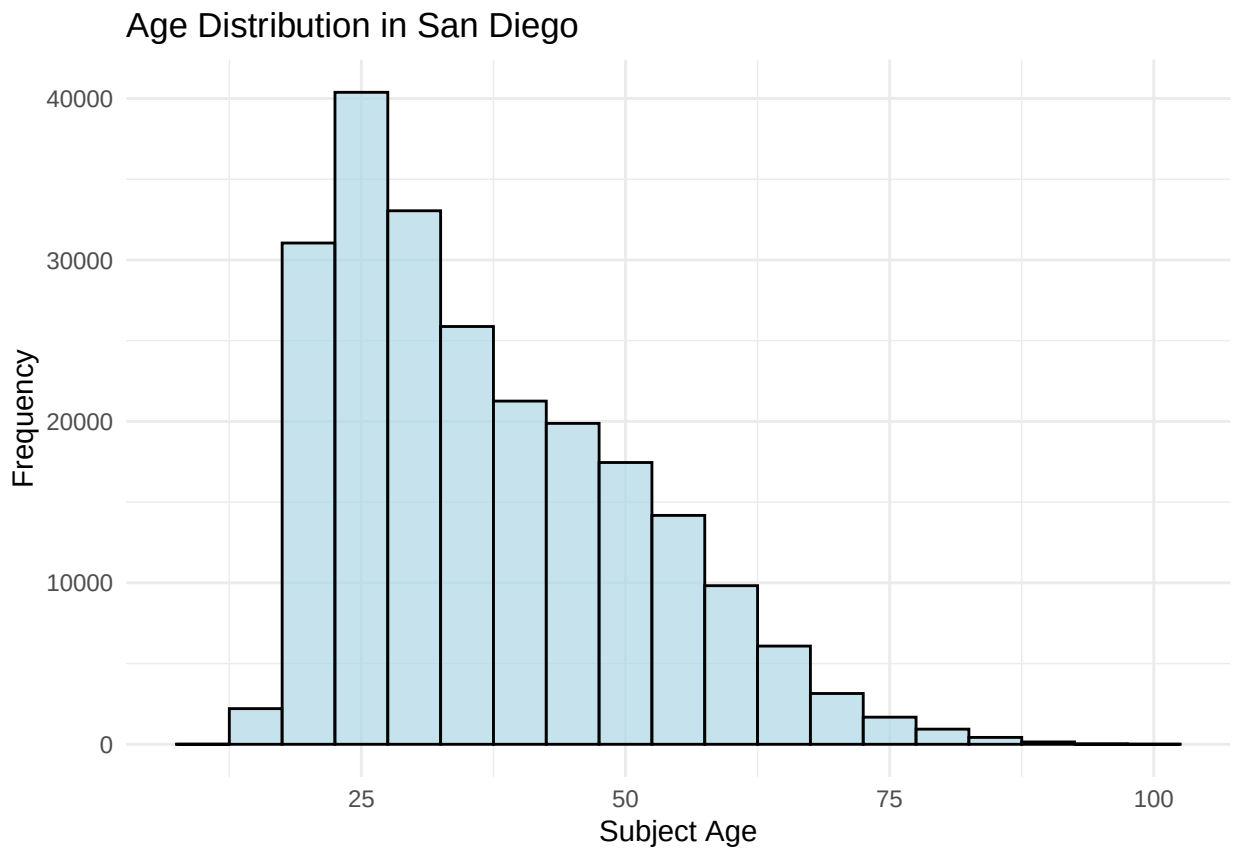
The normalized histogram

```r
# Plot histogram of Box-Cox transformed 'subject_age'
ggplot(charlotte_data, aes(x = subject_age_boxcox)) +
  geom_histogram(binwidth = 0.1, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Box-Cox Transformed Age Distribution in Charlotte",
       x = "Transformed Subject Age", y = "Frequency") +
  theme_minimal()
```

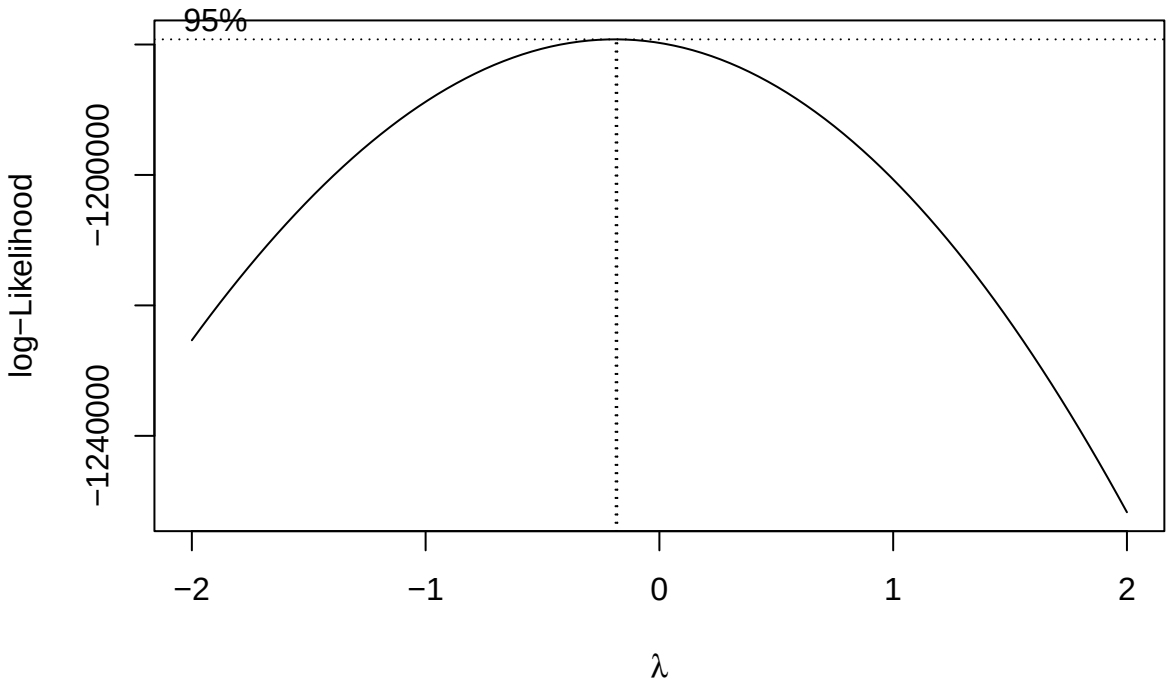


Similarly we are repeating the same procedure for San Diego and Nashville

```
#San Diego
# Filter the data for a particular city
sandiego_data <- data %>% filter(cityname == "SD", !is.na(subject_age))

# Plot the histogram for 'subject_age' in San Diego
ggplot(sandiego_data, aes(x = subject_age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Age Distribution in San Diego",
       x = "Subject Age", y = "Frequency") +
  theme_minimal()
```




```
# Apply Box-Cox transformation
# Ensure that the subject_age has no zero or negative values
# If necessary, add a constant to make the values positive
sandiego_data$subject_age_transformed <- sandiego_data$subject_age

# Box-Cox transformation (choose the best lambda automatically)
boxcox_result <- boxcox(subject_age_transformed ~ 1, data = sandiego_data, lambda = seq(-2, 2, by = 0.1
```

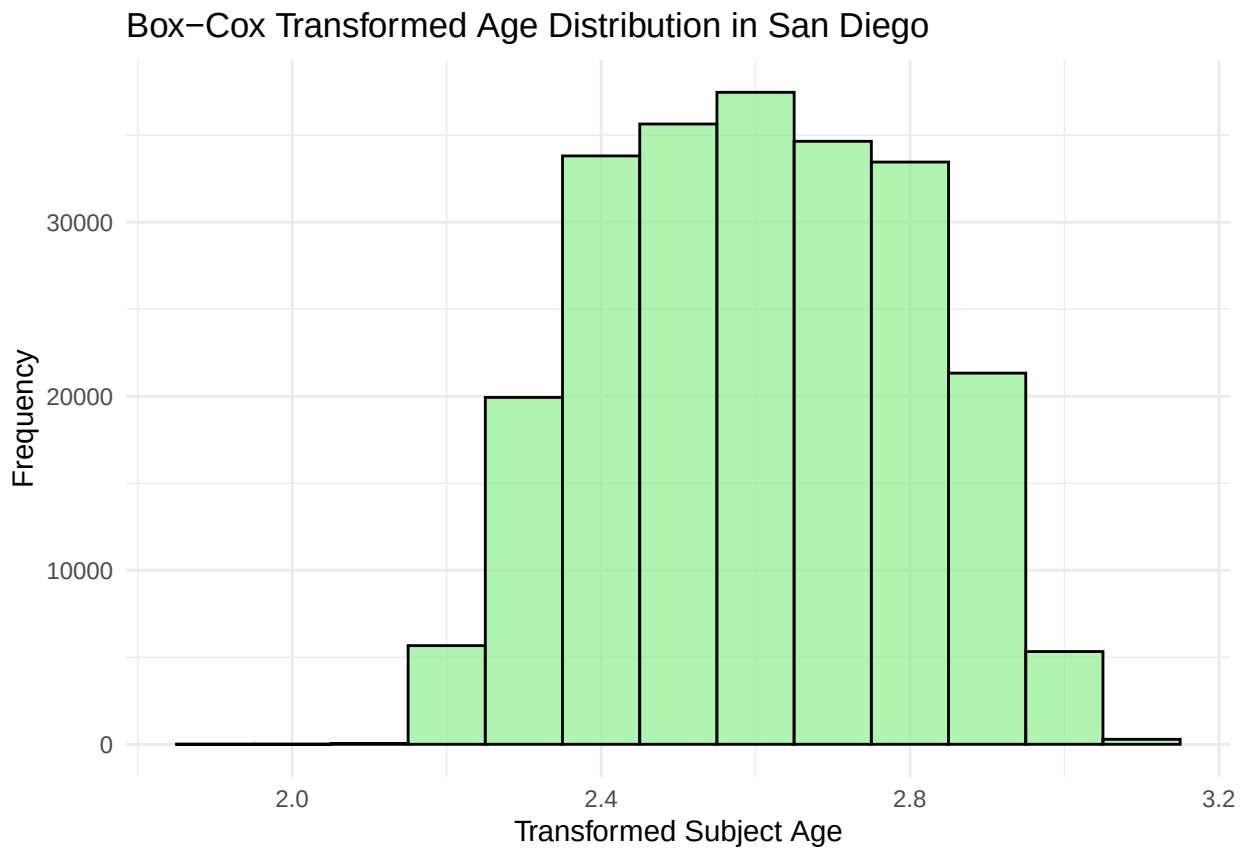



```r
# The lambda value that maximizes the log-likelihood
best_lambda_s <- boxcox_result$x[which.max(boxcox_result$y)]
cat("Optimal Lambda for Box-Cox transformation: ", best_lambda_s, "\n")
```

```
## Optimal Lambda for Box-Cox transformation:  -0.1818182
```

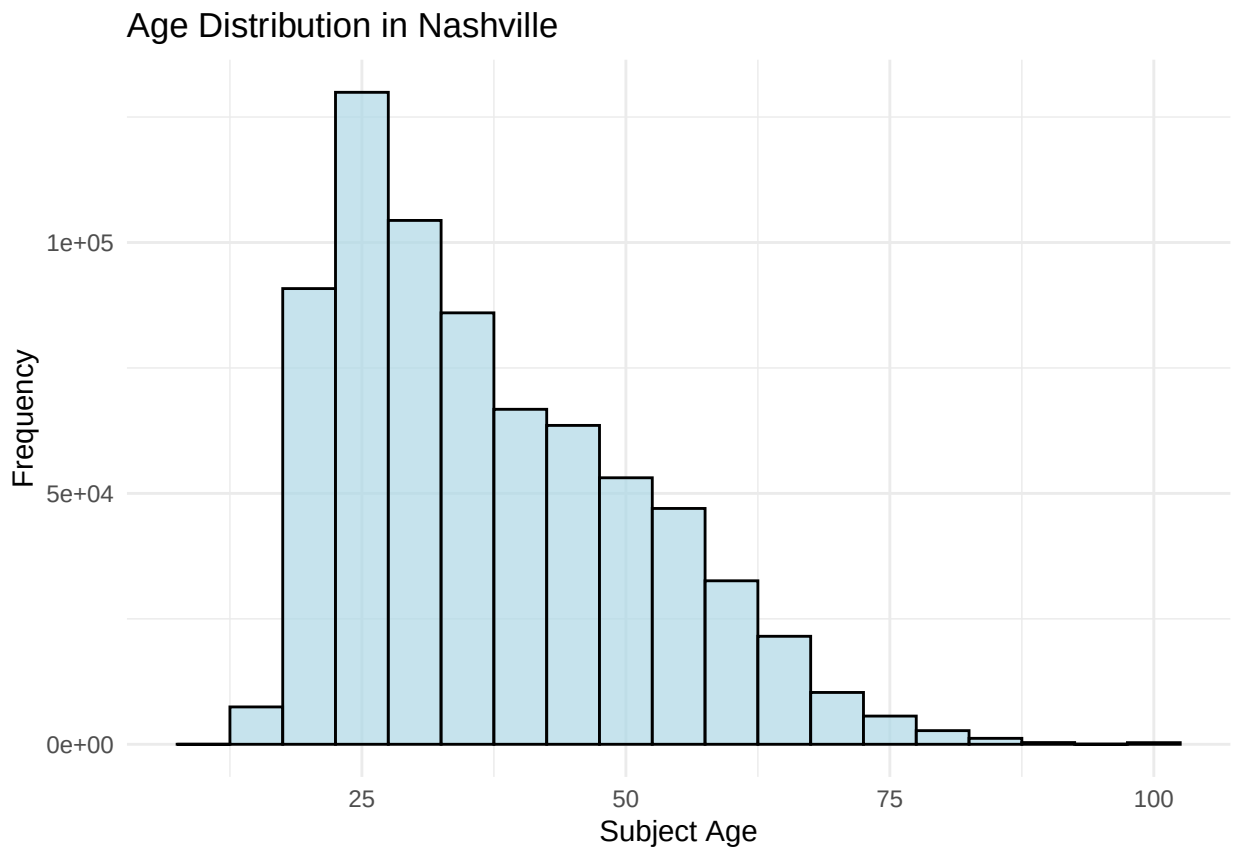
```r
# Apply the transformation with the best lambda
sandiego_data$subject_age_boxcox <- (sandiego_data$subject_age_transformed^best_lambda_s - 1) / best_lar
```

```r
# Plot histogram of Box-Cox transformed 'subject_age'
ggplot(sandiego_data, aes(x = subject_age_boxcox)) +
  geom_histogram(binwidth = 0.1, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Box-Cox Transformed Age Distribution in San Diego",
       x = "Transformed Subject Age", y = "Frequency") +
  theme_minimal()
```

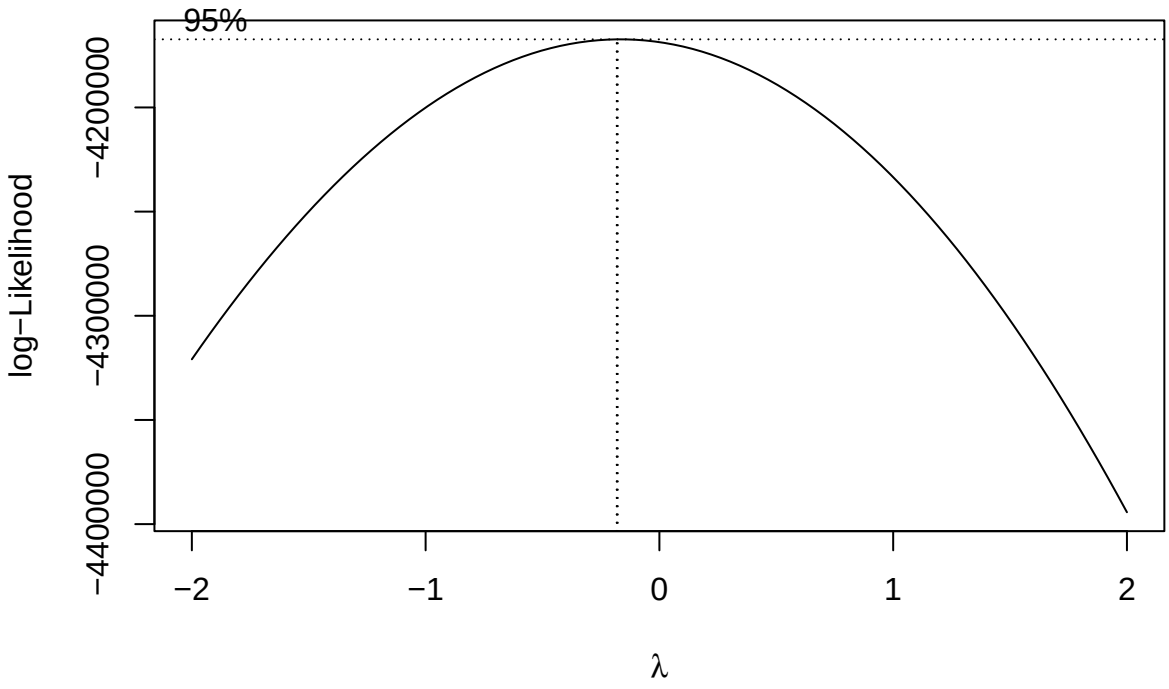


Now box cox transformation of Nashville

```
#Nashville
# Filter the data for a particular city
nashville_data <- data %>% filter(cityname == "Nashville", !is.na(subject_age))

# Plot the histogram for 'subject_age' in Nashville
ggplot(nashville_data, aes(x = subject_age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Age Distribution in Nashville",
       x = "Subject Age", y = "Frequency") +
  theme_minimal()
```




```r
# Apply Box-Cox transformation
# Ensure that the subject_age has no zero or negative values
# If necessary, add a constant to make the values positive
nashville_data$subject_age_transformed <- nashville_data$subject_age

# Box-Cox transformation (choose the best lambda automatically)
boxcox_result <- boxcox(subject_age_transformed ~ 1, data = nashville_data, lambda = seq(-2, 2, by = 0.
```

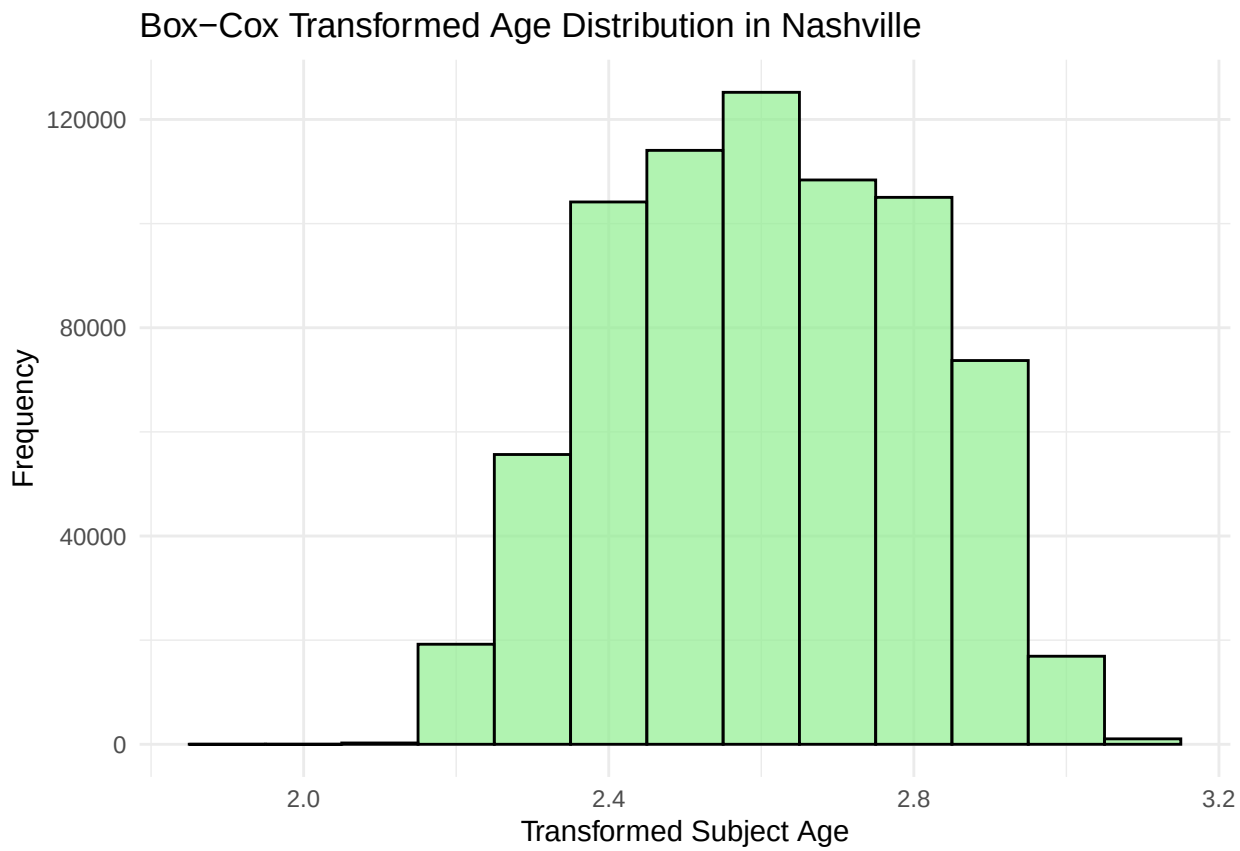



```
# The lambda value that maximizes the log-likelihood
best_lambda_n <- boxcox_result$x[which.max(boxcox_result$y)]
cat("Optimal Lambda for Box-Cox transformation: ", best_lambda_n, "\n")
```

```
## Optimal Lambda for Box-Cox transformation:  -0.1818182
```

```
# Apply the transformation with the best lambda
nashville_data$subject_age_boxcox <- (nashville_data$subject_age_transformed^best_lambda_n - 1) / best_l
```

```
# Plot histogram of Box-Cox transformed 'subject_age'
ggplot(nashville_data, aes(x = subject_age_boxcox)) +
  geom_histogram(binwidth = 0.1, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Box-Cox Transformed Age Distribution in Nashville",
       x = "Transformed Subject Age", y = "Frequency") +
  theme_minimal()
```



So we observed that before transformation the age distribution across all cities is right skewed and we normalised them by applying the box-cox transformation and the optimal lambda for box transformation in the case of Charlotte, San Diego and Nashville are -0.2222222, -0.1818182 and -0.1818182. So we analysed that lambda values are the same in the case of San Diego and Nashville, which gives us the insight that the distribution of age is similar.

Inferential Statistics

Inferential statistics offers robust methodologies for the examination of traffic policing data, enabling a deeper understanding of patterns, correlations, and disparities among diverse groups. This analytical approach is essential for guiding policy formulation and promoting equitable and effective law enforcement practices. We were able to address inquiries such as whether certain racial groups exhibit a higher propensity for specific categories of traffic violations compared to others, the existence of a significant correlation between race and the probability of receiving a traffic citation, and the extent of variability in traffic violations across different racial groups, as well as how this variability may fluctuate depending on the nature of the offense. The statistical techniques employed include Confidence Intervals, Regression Analysis, ANOVA, Inferences on Variances, Inference on Means, and Proportion tests.

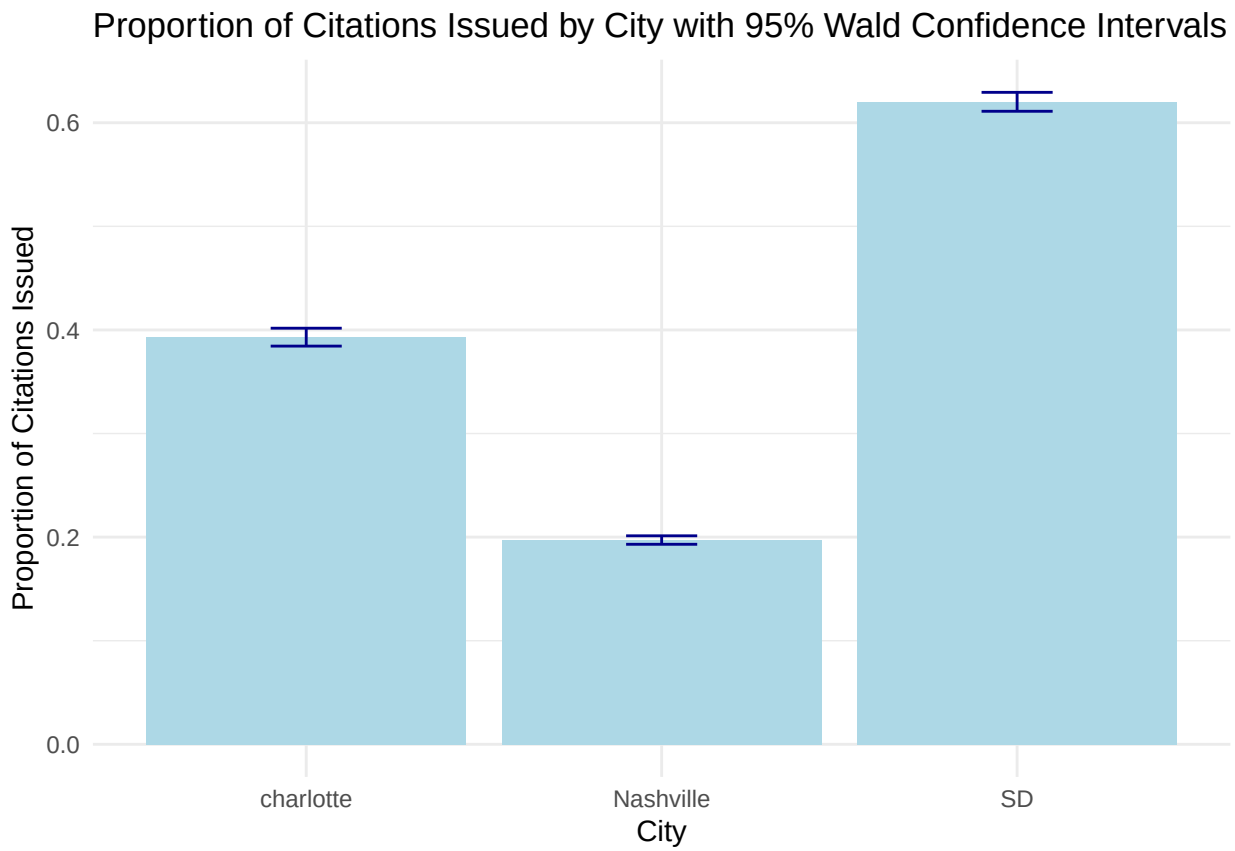
1.**Wald Confidence Interval for Three Proportions: Citations Issued Across the Cities***

The Question we are trying to answer is **Does the proportion of citations issued differ significantly across the three cities at a 95% confidence level?**

```r
library(dplyr)
#library(ggplot2)
options(dplyr.width = Inf)  # Ensures all columns are displayed

data <- read.csv("cities.csv")

# Set seed for reproducibility
set.seed(123)

# Take a random sample (e.g., 5% of the data)
data_sampled <- data %>%
  sample_frac(0.05) %>%
  filter(
    !is.na(citation_issued)
  )


citation_summary_sampled <- data_sampled %>%
  group_by(cityname) %>%
  summarise(
    proportion_citations = mean(citation_issued == TRUE, na.rm = TRUE),
    standard_error = sqrt((proportion_citations * (1 - proportion_citations)) / n()),
    z_value = qnorm(0.975),
    ci_lower_wald = proportion_citations - z_value * standard_error,
    ci_upper_wald = proportion_citations + z_value * standard_error
  )

# Print the summary table
print(citation_summary_sampled)
```

```
## # A tibble: 3 x 6
##   cityname  proportion_citations standard_error z_value ci_lower_wald
##   <chr>                    <dbl>          <dbl>   <dbl>         <dbl>
## 1 Nashville                0.197        0.00209    1.96         0.193
## 2 SD                       0.620        0.00467    1.96         0.611
## 3 charlotte                0.393        0.00441    1.96         0.384
##   ci_upper_wald
##           <dbl>
## 1         0.201
## 2         0.629
## 3         0.402
```

```r
ggplot(citation_summary_sampled, aes(x = cityname, y = proportion_citations)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  geom_errorbar(aes(ymin = ci_lower_wald, ymax = ci_upper_wald), width = 0.2, color = "darkblue") +
  labs(
    title = "Proportion of Citations Issued by City with 95% Wald Confidence Intervals",
    x = "City",
    y = "Proportion of Citations Issued"
  ) +
  theme_minimal()
```



Confidence Interval for Nashville: [0.1931, 0.2013]
Confidence Interval for San Diego: [0.6111, 0.6294]
Confidence Interval for Charlotte: [0.3844, 0.4017]

The confidence intervals for the proportions of arrests in Nashville, San Diego, and Charlotte do not overlap, indicating the proportion of citations issued differ significantly across the three cities at a 95% confidence level. San Diego has the highest citations proportion (61.11%–62.94%), followed by Charlotte (38.44%–40.17%), and then Nashville (19.31%–20.13%). The lack of overlap suggests statistically significant differences in citations given across these cities.

2.**Confidence Interval for Regression Coefficients: Sex vs Contraband Found**

The question we are trying to answer is **Does the subject's sex (male vs. female) significantly influence the likelihood of contraband being found in Charlotte at a 95% confidence level?***

```
charlotte_data <- read.csv("charlotte_data.csv")
charlotte_data_cleaned <- charlotte_data %>%
 filter(!is.na(contraband_found), !is.na(subject_sex))

# Fit logistic regression model
model <- glm(contraband_found ~ subject_sex, data = charlotte_data_cleaned, family = binomial)

summary(model)
```

```
##
## Call:
```

```
## glm(formula = contraband_found ~ subject_sex, family = binomial,
##     data = charlotte_data_cleaned)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.11477    0.02053 -54.310  < 2e-16 ***
## subject_sexmale  0.05994    0.02206   2.718  0.00657 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 105816  on 92881  degrees of freedom
## Residual deviance: 105808  on 92880  degrees of freedom
## AIC: 105812
##
## Number of Fisher Scoring iterations: 4
```

```
# Confidence intervals for regression coefficients
conf_intervals <- confint(model)
```

```
## Waiting for profiling to be done...
```

```
print(conf_intervals)
```

```
##                      2.5 %     97.5 %
## (Intercept)     -1.15513549 -1.0746721
## subject_sexmale  0.01682963  0.1032901
```

The analysis shows that **yes**, subject sex significantly influences the likelihood of contraband being found in Charlotte at a 95% confidence level. Specifically, males are slightly more likely than females to have contraband found during police stops, with an odds ratio of approximately `1.062`. This odds ratio represents a 6.2% increase in odds of contraband being found for males compared to females. This value is derived from exponentiating the coefficient for `subject_sexmale`:

$$e^{0.05994} = 1.062$$

The 95% confidence interval for this effect is `[1.017, 1.109]`, derived by exponentiating the confidence interval for the coefficient `[0.0168, 0.1033]`, indicating statistical significance because the interval does not include `1`.

The p-value (`0.00657`) for the `subject_sexmale` coefficient is less than `0.05` (alpha), confirming this significance at the 95% confidence level. For females (the reference category), the baseline probability of contraband being found is approximately `24.7%`, calculated using the intercept (`-1.11477`) in the logistic regression model as:

$$P(\text{contraband} \mid \text{female}) = \frac{e^{-1.11477}}{1 + e^{-1.11477}}$$

Being male slightly increases this likelihood to approximately `25.8%`, calculated by adding the `subject_sexmale` coefficient (`0.05994`) to the intercept and using the same formula.

---

3. Problem statement - Are there significant disparities in the number of traffic-related arrests by police among various racial or ethnic groups, when accounting for factors such as the type of offense, location, and time of day, at a 95% confidence interval?

Step 1 – *Check the conditions required for the validity of the test*

The two variables City and the ArrestMade are categorical and it approximately follows a ^{2} distribution with $(r-1)(c-1)$ degrees of freedom. Since all the values are greater than 5, so we can apply Test of Independence with ^{2} test

Step 2 – *Define the parameter of interest*

We're categorizing places like SD, Charlotte, and Nashville, figuring out if the arrest is true or false, and the race we're considering for the test is black.

Step 3 – *State the the desired Significance level*

The Significance level is mentioned in the question which is 0.05. $= 0.05$

Step 4 – *State the Null Hypothesis*

Null Hypothesis states that the arrest rate of the black race is independent of the location

H_0 : *Arrest rate of blackrace is independent of the location*

Step 5 – *State the Alternative Hypothesis*

Alternative Hypothesis states that there is an association between the arrest rate of the black race and the location

H_1 : *Arrest rate of blackrace is associated with the location*

Step 6 – *Determine the test and calculate the test statistic*

```
filtered_black <- subset(data, data$subject_race == 'black')
```

```
#Create the contingency table
contingency_table_one <- table((filtered_black$cityname), (filtered_black$arrest_made))

contingency_table_one
```

```
##
##              FALSE   TRUE
##   charlotte 126060   2923
##   Nashville 262394   6055
##   SD         24432    463
```

```
chisq.test(contingency_table_one, correct=T)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table_one
## X-squared = 17.127, df = 2, p-value = 0.0001909
```

Step 7 − *Calculate the p − value or the critical value*

Since this is a two-sided Hypothesis test

$$P - value\ is\ 0.0001909$$

Step 8 − *Makereject/fail to reject decision*

So we have, $p = 0.0001909$ and $= 0.05$, and $p <$, hence we reject H_0

Step 9 − *State your conclusion in the context of the problem*

$p <$, hence we can conclude that the arrest rate of black race is associated with the location.

---

4. Problem statement - In analyzing the age distribution of individuals across three cities, does gender influence the variability of ages among drivers? Specifically, at a significance level of $= 0.05$, do males in these cities demonstrate a more consistent (less variable) age profile compared to females, or is the age variability between genders comparable across these locations?

Step 1 − *Check the conditions required for the validity of the test*

Since we are comparing two different populations, age of male individuals vs age of female individuals, the type of distribution demonstrated is F distribution

Step 2 − *Define the parameter of interest*

$\frac{2}{1}$ is the variance of the age of male individuals and $\frac{2}{2}$ is the variance of the age of female individuals

Step 3 − *State the the desired Significance level*

The Significance level is mentioned in the question which is 0.05. $= 0.05$

Step 4 − *State the Null Hypothesis*

Null Hypothesis states that the variance of the age of male individuals is equal to the variance of the age of female individuals

$$H_0 : \frac{2}{1} = \frac{2}{2}$$

Step 5 − *State the Alternative Hypothesis*

Alternative Hypothesis states that the variance of the age of male individuals is not equal to the variance of the age of female individuals

$$H_1 : \frac{22}{12}$$

Step 6 – *Determine the test and calculate the test statistic*

F test statistic is the test statistic

```
#Remove rows with NA values in the 'subject_sex' column
var_pop <- data[!is.na(data$subject_sex), ]

#Remove rows with NA values in the 'subject_age' column
var_pop <- data[!is.na(data$subject_age), ]

female_pop <- var_pop$subject_age[var_pop$subject_sex == 'female' |    var_pop$subject_sex == 'Female'

male_pop <- var_pop$subject_age[var_pop$subject_sex == 'male' |    var_pop$subject_sex == 'Male' | var_
```

```
var.test(male_pop, female_pop, ratio = 1, alternative = c("two.sided"),
 conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  male_pop and female_pop
## F = 1.0588, num df = 713650, denom df = 473027, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.054569 1.063104
## sample estimates:
## ratio of variances
##           1.058832
```

Step 7 – *Calculate the* p – *value or the critical value*

This is a two-sided Hypothesis test

$$P - value\ is\ < 2.2e - 16$$

Step 8 – *Makereject*/*fail to reject decision*

So we have, $p < 2.2e - 16$ and $= 0.05$, and $p <$, hence we reject H_0

Step 9 – *State your conclusion in the context of the problem*

$p <$, hence we can conclude that the variance of the ages of female population is not equal to the variance of the ages of the male population.

The F-statistic indicates a slight difference in the variances between male and female populations, with the variance for males being about 1.06 times larger than that for females. The p-value is extremely low, which

provides strong evidence against the null hypothesis, leading us to reject the claim that the variances are equal. Furthermore, the confidence interval for the ratio of variances is entirely above 1, further suggesting that the variance of the male population is significantly higher than that of the female population.

---

5. Problem statement - We want to obtain the inference on mean age of black population based on variables like citation issued, warning issued and arrest across three cities to assess whether race along with the mentioned factors influences the variability of mean ages among cities. Specifically, from descriptive analysis we found that a particular mean of age especially among black race have more counts than any other race across these cities. So we want to verify is there really any difference in mean age of black or not?

Since in this case we will be using Welch test but in order to comply with the welch test condition we will be performing the variance test of age of black population based on the same factors across the cities to ensure that there is an unequal variance of age.

$$Step\ 1 - Check\ the\ conditions\ required\ for\ the\ validity\ of\ the\ test$$

Since we are comparing two different cities, age of black population in city1 (e.g. Nashville) vs age of black population in city2 (e.g. San Diego), the type of distribution demonstrated is F distribution

$$Step\ 2 - Define\ the\ parameter\ of\ interest$$

$\frac{2}{1}$ is the variance of the age of black population based on the mentioned factors in city1 and $\frac{2}{2}$ is the variance of the age of black population based on the mentioned factors in city2

$$Step\ 3 - State\ the\ the\ desired\ Significance\ level$$

The Significance level is mentioned in the question which is 0.05. = 0.05

$$Step\ 4 - State\ the\ Null\ Hypothesis$$

Null Hypothesis states that the variance of the age of black population based on factors in one of the city is equal to the variance of the age of black population based on factors in another city,

$$H_0 : \frac{2}{1} = \frac{2}{2}$$

$$Step\ 5 - State\ the\ Alternative\ Hypothesis$$

Alternative Hypothesis states that the variance of the age of black population based on factors in one of the city is not equal to the variance of the age of black population based on factors in another city

$$H_1 : \frac{22}{12}$$

$$Step\ 6 - Determine\ the\ test\ and\ calculate\ the\ test\ statistic$$

F test statistic is the test statistic

$$Step\ 7 - Calculate\ the\ p - value\ or\ the\ critical\ value$$

This is a two-sided Hypothesis test

```r
#variance test for age of black and white across all cities
# Load necessary libraries
library(dplyr)

# Filter the data for black individuals who have citation issued, warning issued, or arrest made
df_filtered_black <- data %>%
  filter(subject_race == "black",
         citation_issued == TRUE | warning_issued == TRUE | arrest_made == TRUE)

# Get unique city names
city_names <- unique(df_filtered_black$cityname)

# Create an empty list to store variance test results
variance_test_results <- list()

# Perform variance test (F-test) for subject_age between each pair of cities
for (i in 1:(length(city_names) - 1)) {
  for (j in (i + 1):length(city_names)) {

    # Filter data for the two cities
    city1_data <- df_filtered_black %>% filter(cityname == city_names[i])
    city2_data <- df_filtered_black %>% filter(cityname == city_names[j])

    # Perform variance test between the two cities for subject_age
    var_test_result <- var.test(city1_data$subject_age, city2_data$subject_age)

    # Store the result in the list
    variance_test_results[[paste(city_names[i], city_names[j], sep = "_vs_")]] <- list(
      city1 = city_names[i],
      city2 = city_names[j],
      p_value = var_test_result$p.value,
      statistic = var_test_result$statistic,
      conf_int = var_test_result$conf.int
    )
  }
}

# View the variance test results
variance_test_results
```

```
## $SD_vs_charlotte
## $SD_vs_charlotte$city1
## [1] "SD"
##
## $SD_vs_charlotte$city2
## [1] "charlotte"
##
## $SD_vs_charlotte$p_value
## [1] 0
##
## $SD_vs_charlotte$statistic
##        F
## 1.226088
```

```
##
## $SD_vs_charlotte$conf_int
## [1] 1.201610 1.251247
## attr(,"conf.level")
## [1] 0.95
##
##
## $SD_vs_Nashville
## $SD_vs_Nashville$city1
## [1] "SD"
##
## $SD_vs_Nashville$city2
## [1] "Nashville"
##
## $SD_vs_Nashville$p_value
## [1] 8.740003e-29
##
## $SD_vs_Nashville$statistic
##         F
## 0.8939787
##
## $SD_vs_Nashville$conf_int
## [1] 0.8768577 0.9115818
## attr(,"conf.level")
## [1] 0.95
##
##
## $charlotte_vs_Nashville
## $charlotte_vs_Nashville$city1
## [1] "charlotte"
##
## $charlotte_vs_Nashville$city2
## [1] "Nashville"
##
## $charlotte_vs_Nashville$p_value
## [1] 0
##
## $charlotte_vs_Nashville$statistic
##         F
## 0.7291309
##
## $charlotte_vs_Nashville$conf_int
## [1] 0.7222563 0.7360830
## attr(,"conf.level")
## [1] 0.95
```

$$P - value\ is\ < 2.2e - 16$$

Step 8 − *Makereject*/*fail to reject decision*

So we have, $p < 2.2e - 16$ and $= 0.05$, and $p <$, hence we reject H_0

Step 9 − *State your conclusion in the context of the problem*

$p <$, hence we can conclude that the variance of the age of black population based on factors in one of the city is not equal to the variance of the age of black population based on factors in another city

Now performing the Welch since we found variance of age of black population across the cities to be unequal.

Step 1 − *Check the conditions required for the validity of the test*

Since we are comparing two different cities, mean age of black population in city1 (e.g. Nashville vs mean age of black population in city2 (e.g. San Diego), the type of distribution demonstrated is by t distribution.

Step 2 − *Define the parameter of interest*

$\frac{2}{1}$ is the mean age of black population based on the mentioned factors in city1 and $\frac{2}{2}$ is the mean age of black population based on the mentioned factors in city2

Step 3 − *State the the desired Significance level*

The Significance level is mentioned in the question which is 0.05. = 0.05

Step 4 − *State the Null Hypothesis*

Null Hypothesis states that the mean age of black population based on factors in one of the city is equal to the mean age of black population based on factors in another city.

$$H_0 : \frac{2}{1} = \frac{2}{1}$$

Step 5 − *State the Alternative Hypothesis*

Alternative Hypothesis states that the mean age of black population based on factors in one of the city is not equal to the mean age of black population based on factors in another city.

$$H_1 : \frac{22}{12}$$

Step 6 − *Determine the test and calculate the test statistic*

t test statistic is the test statistic

Step 7 − *Calculate the p − value or the critical value*

This is a two-sided Hypothesis test

```
# Load necessary libraries
library(dplyr)

# Filter for black individuals who were issued citations, warnings, or arrests
df_filtered_black <- data %>%
  filter(subject_race == "black" &
         (citation_issued == TRUE | warning_issued == TRUE | arrest_made == TRUE))

# Function to perform Welch's t-test comparing mean age of black individuals across different cities
```

```r
perform_welch_test_cities <- function(data, age_col, city_col) {

  # Get unique city names
  city_names <- unique(data[[city_col]])

  # Create an empty list to store results
  t_test_results <- list()

  # Loop through all pairs of cities and perform Welch's t-test
  for (i in 1:(length(city_names) - 1)) {
    for (j in (i + 1):length(city_names)) {

      city1 <- city_names[i]
      city2 <- city_names[j]

      # Filter data for the two cities
      city_data <- data %>%
        filter(!!sym(city_col) %in% c(city1, city2))

      # Perform Welch's t-test comparing the mean age of black individuals between the two cities
      t_test_result <- t.test(
        subject_age ~ cityname,  # Use direct column names
        data = city_data,
        var.equal = FALSE
      )

      # Store the result
      t_test_results[[paste(city1, city2, sep = "_vs_")]] <- list(
        city1 = city1,
        city2 = city2,
        p_value = t_test_result$p.value,
        statistic = t_test_result$statistic,
        conf_int = t_test_result$conf.int,
        estimate = t_test_result$estimate
      )
    }
  }

  # Return the list of results
  return(t_test_results)
}

# Run the function to perform Welch's t-test comparing the mean ages of black individuals who were issu
welch_results <- perform_welch_test_cities(df_filtered_black, "subject_age", "cityname")

# Print the results
welch_results
```

```
## $SD_vs_charlotte
## $SD_vs_charlotte$city1
## [1] "SD"
##
## $SD_vs_charlotte$city2
```

```
## [1] "charlotte"
##
## $SD_vs_charlotte$p_value
## [1] 6.350808e-17
##
## $SD_vs_charlotte$statistic
##         t
## -8.363681
##
## $SD_vs_charlotte$conf_int
## [1] -0.9783913 -0.6068796
## attr(,"conf.level")
## [1] 0.95
##
## $SD_vs_charlotte$estimate
## mean in group charlotte        mean in group SD
##                34.80984                35.60248
##
##
## $SD_vs_Nashville
## $SD_vs_Nashville$city1
## [1] "SD"
##
## $SD_vs_Nashville$city2
## [1] "Nashville"
##
## $SD_vs_Nashville$p_value
## [1] 6.95137e-16
##
## $SD_vs_Nashville$statistic
##        t
## 8.076451
##
## $SD_vs_Nashville$conf_int
## [1] 0.5665077 0.9295927
## attr(,"conf.level")
## [1] 0.95
##
## $SD_vs_Nashville$estimate
## mean in group Nashville        mean in group SD
##                36.35053                35.60248
##
##
## $charlotte_vs_Nashville
## $charlotte_vs_Nashville$city1
## [1] "charlotte"
##
## $charlotte_vs_Nashville$city2
## [1] "Nashville"
##
## $charlotte_vs_Nashville$p_value
## [1] 1.7209e-280
##
## $charlotte_vs_Nashville$statistic
```

```
##         t
## -35.82809
##
## $charlotte_vs_Nashville$conf_int
## [1] -1.624969 -1.456403
## attr(,"conf.level")
## [1] 0.95
##
## $charlotte_vs_Nashville$estimate
## mean in group charlotte mean in group Nashville
##                34.80984                36.35053
```

$$P - value\ is\ < 2.2e - 16$$

$$Step\ 8 - Makereject/fail\ to\ reject\ decision$$

So we have, $p < 2.2e - 16$ and $= 0.05$, and $p <$, hence we reject H_0

$$Step\ 9 - State\ your\ conclusion\ in\ the\ context\ of\ the\ problem$$

$p <$, hence we can conclude that the mean age of black population based on factors in one of the city is not equal to the mean age of black population based on factors in another city.

---

6. Problem Statement - Now, we want to obtain the inference on the mean age of black population and white population based on variables like citation issued, warning issued and arrest across three cities to assess whether races along with the mentioned factors influences the variability of mean ages among cities. Specifically, from descriptive analysis we found that a particular mean of age especially among black race have more counts than any the white race across these cities. So we want to verify is there really any difference in mean age of black and white?

Since in this case we will be using Welch test but in order to comply with the welch test condition we will be again performing the variance test of age of black and white population based on the same factors across the cities to ensure that there is an unequal variance of age among the two races.

$$Step\ 1 - Check\ the\ conditions\ required\ for\ the\ validity\ of\ the\ test$$

Since we are comparing two different cities, age of black population across the cities vs age of white population across all the cities (e.g. San Diego), the type of distribution demonstrated is F distribution

$$Step\ 2 - Define\ the\ parameter\ of\ interest$$

$\frac{2}{1}$ is the variance of the age of black population based on the mentioned factors in cities and $\frac{2}{2}$ is the variance of the age of white population based on the mentioned factors in cities.

$$Step\ 3 - State\ the\ the\ desired\ Significance\ level$$

The Significance level is mentioned in the question which is 0.05. $= 0.05$

$$Step\ 4 - State\ the\ Null\ Hypothesis$$

Null Hypothesis states that the variance of the age of black population based on factors in cities is equal to the variance of the age of white population based on factors in cities.

$$H_0 : {}^{2}_{1} = {}^{2}_{2}$$

Step 5 – State the Alternative Hypothesis

Alternative Hypothesis states that the variance of the age of black population based on factors in cities is not equal to the variance of the age of white population based on factors in cities.

$$H_1 : {}^{22}_{12}$$

Step 6 – Determine the test and calculate the test statistic

F test statistic is the test statistic

Step 7 – Calculate the p – value or the critical value

This is a two-sided Hypothesis test

```r
#variance test for age of black and white
# Load necessary libraries
library(dplyr)

# Filter the data for black and white individuals who have citation issued, warning issued, or arrest m
df_filtered <- data %>%
  filter(subject_race %in% c("black", "white"),
         citation_issued == TRUE | warning_issued == TRUE | arrest_made == TRUE)

# Get unique city names
city_names <- unique(df_filtered$cityname)

# Create an empty list to store variance test results
variance_test_results <- list()

# Perform variance test (F-test) for subject_age between black and white individuals across cities
for (city in city_names) {

  # Filter data for the current city
  city_data <- df_filtered %>%
    filter(cityname == city)

  # Perform variance test between black and white individuals for subject_age in the current city
  var_test_result <- var.test(subject_age ~ subject_race, data = city_data)

  # Store the result in the list
  variance_test_results[[city]] <- list(
    city = city,
    p_value = var_test_result$p.value,
    statistic = var_test_result$statistic,
    conf_int = var_test_result$conf.int
  )
```

```
}

# View the variance test results
variance_test_results
```

```
## $SD
## $SD$city
## [1] "SD"
##
## $SD$p_value
## [1] 6.026472e-125
##
## $SD$statistic
##         F
## 0.7712833
##
## $SD$conf_int
## [1] 0.7554225 0.7875818
## attr(,"conf.level")
## [1] 0.95
##
##
## $charlotte
## $charlotte$city
## [1] "charlotte"
##
## $charlotte$p_value
## [1] 0
##
## $charlotte$statistic
##         F
## 0.7377987
##
## $charlotte$conf_int
## [1] 0.7286990 0.7470001
## attr(,"conf.level")
## [1] 0.95
##
##
## $Nashville
## $Nashville$city
## [1] "Nashville"
##
## $Nashville$p_value
## [1] 4.847367e-124
##
## $Nashville$statistic
##         F
## 0.9193588
##
## $Nashville$conf_int
## [1] 0.9130052 0.9257609
## attr(,"conf.level")
```

```
## [1] 0.95
```

$$P - value\ is\ < 2.2e - 16$$

$$Step\ 8 - Makereject/fail\ to\ reject\ decision$$

So we have, $p < 2.2e - 16$ and $= 0.05$, and $p <$, hence we reject H_0

$$Step\ 9 - State\ your\ conclusion\ in\ the\ context\ of\ the\ problem$$

$p <$, hence we can conclude that the variance of the age of black population based on factors in cities is not equal to the variance of the age of white population based on factors in cities.

Now performing Welch test for mean age between black and white since the above condition is satisfied

$$Step\ 1 - Check\ the\ conditions\ required\ for\ the\ validity\ of\ the\ test$$

Since we are comparing two races, mean age of black population in cities (e.g. Nashville vs mean age of white population in cities, the type of distribution demonstrated is by t distribution.

$$Step\ 2 - Define\ the\ parameter\ of\ interest$$

$\frac{2}{1}$ is the mean age of black population based on the mentioned factors in cities and $\frac{2}{2}$ is the mean age of white population based on the mentioned factors in cities

$$Step\ 3 - State\ the\ the\ desired\ Significance\ level$$

The Significance level is mentioned in the question which is 0.05. $= 0.05$

$$Step\ 4 - State\ the\ Null\ Hypothesis$$

Null Hypothesis states that the mean age of black population based on factors in cities is equal to the mean age of white population based on factors in cities.

$$H_0 : \frac{2}{1} = \frac{2}{1}$$

$$Step\ 5 - State\ the\ Alternative\ Hypothesis$$

Alternative Hypothesis states that the mean age of black population based on factors in cities is not equal to the mean age of white population based on factors in cities.

$$H_1 : \frac{22}{12}$$

$$Step\ 6 - Determine\ the\ test\ and\ calculate\ the\ test\ statistic$$

t test statistic is the test statistic

$$Step\ 7 - Calculate\ the\ p - value\ or\ the\ critical\ value$$

This is a two-sided Hypothesis test

```r
# Load necessary libraries
library(dplyr)

# Filter for black and white individuals who were issued citations, warnings, or arrests
df_filtered_black_white <- data %>%
  filter(subject_race %in% c("black", "white") &
         (citation_issued == TRUE | warning_issued == TRUE | arrest_made == TRUE))

# Function to perform Welch's t-test comparing mean age of black and white individuals within each city
perform_welch_test_cities <- function(data, age_col, city_col) {

  # Get unique city names
  city_names <- unique(data[[city_col]])

  # Create an empty list to store results
  t_test_results <- list()

  # Loop through all cities and perform Welch's t-test comparing mean age between black and white indiv
  for (city in city_names) {

    # Filter data for the current city
    city_data <- data %>%
      filter(!!sym(city_col) == city)

    # Perform Welch's t-test comparing the mean age between black and white individuals within the city
    t_test_result <- t.test(
      subject_age ~ subject_race,  # Comparing mean age between black and white individuals
      data = city_data,
      var.equal = FALSE
    )

    # Store the result
    t_test_results[[city]] <- list(
      city = city,
      p_value = t_test_result$p.value,
      statistic = t_test_result$statistic,
      conf_int = t_test_result$conf.int,
      estimate = t_test_result$estimate
    )
  }

  # Return the list of results
  return(t_test_results)
}

# Run the function to perform Welch's t-test comparing the mean ages of black and white individuals who
welch_results <- perform_welch_test_cities(df_filtered_black_white, "subject_age", "cityname")

# Print the results
print(welch_results)
```

```
## $SD
## $SD$city
```

```
## [1] "SD"
##
## $SD$p_value
## [1] 1.825383e-192
##
## $SD$statistic
##         t
## -29.76706
##
## $SD$conf_int
## [1] -3.231358 -2.832106
## attr(,"conf.level")
## [1] 0.95
##
## $SD$estimate
## mean in group black mean in group white
##            35.60248            38.63421
##
##
## $charlotte
## $charlotte$city
## [1] "charlotte"
##
## $charlotte$p_value
## [1] 0
##
## $charlotte$statistic
##         t
## -47.98474
##
## $charlotte$conf_int
## [1] -2.926779 -2.697068
## attr(,"conf.level")
## [1] 0.95
##
## $charlotte$estimate
## mean in group black mean in group white
##            34.80984            37.62176
##
##
## $Nashville
## $Nashville$city
## [1] "Nashville"
##
## $Nashville$p_value
## [1] 0
##
## $Nashville$statistic
##         t
## -53.95122
##
## $Nashville$conf_int
## [1] -1.982235 -1.843260
## attr(,"conf.level")
```

```
## [1] 0.95
##
## $Nashville$estimate
## mean in group black mean in group white
##            36.35053            38.26327
```

So we have, $p < 2.2e-16$ and $= 0.05$, and $p <$, hence we reject H_0

Step 9 – State your conclusion in the context of the problem

$p <$, hence we can conclude that the mean age of black population based on factors in cities is not equal to the mean age of white population based on factors in cities.

---

7. Applying confidence interval

Problem statement - We want to know whether the proportion of black individuals across the cities is a true proportion?

So we are using proportion test i.e. (prop.test()) for the same.

```
# Filter data for 'black' individuals
df_black <- data %>%
  filter(subject_race == "black")

# Get unique city names
city_names <- unique(df_black$cityname)

# Create an empty list to store results
confidence_intervals <- list()

# Loop through each city and calculate the proportion of black individuals and its confidence interval
for (city in city_names) {

  # Filter data for the current city
  city_data <- df_black %>%
    filter(cityname == city)

  # Calculate the number of black individuals in the city
  num_black <- nrow(city_data)

  # Calculate the total number of individuals in the city (including other races)
  total_in_city <- nrow(data %>% filter(cityname == city))

  # Perform a proportion test to get the confidence interval for the proportion of black individuals
  prop_test_result <- prop.test(num_black, total_in_city, conf.level = 0.95)

  # Store the results (city name, proportion, and confidence interval)
  confidence_intervals[[city]] <- list(
    city = city,
    proportion_black = num_black / total_in_city,
    lower_bound = prop_test_result$conf.int[1],
```

```
    upper_bound = prop_test_result$conf.int[2]
  )
}

# View the results
confidence_intervals
```

```
## $SD
## $SD$city
## [1] "SD"
##
## $SD$proportion_black
## [1] 0.1093573
##
## $SD$lower_bound
## [1] 0.1080968
##
## $SD$upper_bound
## [1] 0.1106306
##
##
## $charlotte
## $charlotte$city
## [1] "charlotte"
##
## $charlotte$proportion_black
## [1] 0.5238548
##
## $charlotte$lower_bound
## [1] 0.5218797
##
## $charlotte$upper_bound
## [1] 0.5258291
##
##
## $Nashville
## $Nashville$city
## [1] "Nashville"
##
## $Nashville$proportion_black
## [1] 0.3709092
##
## $Nashville$lower_bound
## [1] 0.3697964
##
## $Nashville$upper_bound
## [1] 0.3720235
```

We observe a tight confidence interval between the proportions, it means that there is less uncertainty about the estimate of the proportion, and the data you have provides strong evidence for the value of the proportion.

8. Applying Mean of Age comparison between cities using 1-Way Anova:

QUESTION: Is the mean age of subjects in each city similar?

H0=Mean of ages is equal for all 3 cities. H1=There is at least one pair of cities which have different means of age. Significance level: 0.05 Test This is typical candidate for 1-way ANOVA

```
# Group by 'cityname' and calculate mean of 'subject_age' and most frequent 'day_period'
result <- data %>%
  group_by(cityname) %>%
  summarise(
    mean_age = mean(subject_age, na.rm = TRUE),
   )

# View the result
print(result)
```

```
## # A tibble: 3 x 2
##   cityname  mean_age
##   <chr>        <dbl>
## 1 Nashville     37.2
## 2 SD            37.0
## 3 charlotte     35.6
```

Anova 1-way:

```
anova_cities=aov(subject_age~cityname,data=data)
summary(anova_cities)
```

```
##                  Df    Sum Sq Mean Sq F value Pr(>F)
## cityname          2    484498  242249    1262 <2e-16 ***
## Residuals   1197481 229933095     192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 6400 observations deleted due to missingness
```

Conclusion: As the p value(<2e-16) is much less than the alpha(0.05) we REJECT H0 and conclude that there is at least one pair which have different age means.**

Post-Hoc Tukey for 'Age Mean' pairs:

For each pair we say H0= The mean difference of the cities paires is 0. H1= The mean difference is not 0.

```
TukeyHSD(anova_cities)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = subject_age ~ cityname, data = data)
##
## $cityname
##                           diff        lwr        upr p adj
## Nashville-charlotte  1.6198062  1.5440357  1.6955768     0
## SD-charlotte         1.3246071  1.2301755  1.4190387     0
## SD-Nashville        -0.2951991 -0.3732451 -0.2171531     0
```

The Tukey test also shows that all pairs are siginificantly different then each other since each pair comparison resulted in *p_adj=0(very small) that is way less than alpha. Also you can see the lower and upper bounds above where none of them has 0 in, that also shows the there is no equality.*

9. A confidence of interval analysis:

The CI for age in SD to be considered as equal to Charlotte by 95% interval:

```
SD_mean = mean(data %>% filter(cityname == 'San Diego') %>% pull(subject_age),na.rm = TRUE)
SD_sd=sd(data %>% filter(cityname == 'San Diego') %>% pull(subject_age),na.rm = TRUE)
margin=qnorm(0.975)*SD_sd
sd_selected = data %>% filter(cityname == 'San Diego')

SD_min=SD_mean-margin/sqrt(nrow(sd_selected))

SD_max=SD_mean+margin/sqrt(nrow(sd_selected))

print(paste0("San Diego Mean=",SD_mean))
```

```
## [1] "San Diego Mean=NaN"
```

```
print(paste0("San Diego Max=",SD_max))
```

```
## [1] "San Diego Max=NaN"
```

```
print(paste0("San Diego Min=",SD_min))
```

```
## [1] "San Diego Min=NaN"
```

#The *95% confidence interval for age* in the city of SD:

$$\left[36.397, 37.011\right]$$

** Using this interval above, look at the means of other cities below and see that none of the means fall into the interval of San Diego. This is just a sample work to show the the difference **using a CI approach.**

```
# Group by 'cityname' and calculate mean of 'subject_age' and most frequent 'day_period'
result <- data %>%
  group_by(cityname) %>%
  summarise(
    mean_age = mean(subject_age, na.rm = TRUE),
   )

# View the result
print(result)
```

```
## # A tibble: 3 x 2
##   cityname  mean_age
##   <chr>        <dbl>
## 1 Nashville     37.2
## 2 SD            37.0
## 3 charlotte     35.6
```

10. LOGISTIC REGRESSION:

Using the data features, we aimed to predict the Arrest Results, which is a binary (True-False):

```
# Train the logistic regression model on the training set
log_reg_model <- glm(arrest_made ~ cityname + subject_race + day_period,
                     family = binomial, data = data)
```

```
# Predict probabilities on the dataset (using the same data for predictions here)
predictions_prob <- predict(log_reg_model, newdata = data, type = "response")

# Convert probabilities to class labels (threshold = 0.3)
predicted_labels <- ifelse(predictions_prob > 0.3, TRUE, FALSE)
```

```
# Confusion matrix
confusion_matrix <- table(predicted_labels, data$arrest_made)
print(confusion_matrix)
```

```
##
## predicted_labels   FALSE    TRUE
##            FALSE 1174164   19027
```

As you can see due to the very low arrest rate, *that is a very significant imbalance towards FALSE* (19027/1174164=1.6%) the model predicted all as False even at lower threshold set above. (0.3 instead of 0.5)**.

Notes and Logistics

Statistics turned out to be super helpful for me because it equipped me with the skills to interpret data and draw conclusions about the world around me. It gave me a solid way to think critically about uncertainty and variability. By getting a grip on things like probability, distributions, and hypothesis testing, I felt more confident and accurate when tackling problems. The professor has a knack for simplifying complex ideas, making even the toughest topics feel approachable. He often shared real examples that were easy to relate to, which helped me see how abstract concepts applied in the real world. His clear teaching style, great resources, and readiness to assist made the subject not just manageable bly fun.

I learned a lot more than I expected in this class. I thought I knew enough about statistics but compared to what I know now, I feel more comfortable and knowledgeable about statistics, probability, and different Hypothesis Tests. I have also experienced using R which is a great plus for all of us. Working with a team and dealing with various challenges in finding, cleaning, and understanding data was also a great experience for me.