Analysing Demographic and Circumstantial Factors in Fatal Police Shootings

The issues:

The fatal-police-shootings-data raises the following questions:

1. What are the demographic characteristics of individuals involved in fatal police shootings (e.g., age, gender, race) and are there disparities in the demographics of those shot by the police?
2. Are there geographic patterns or hotspots for fatal police shootings?
3. Do certain areas or regions have higher rates of police shootings?
4. What were the circumstances leading to each shooting (e.g., armed, unarmed, mental illness) and were there any common risk factors associated with these incidents?
5. What are the common characteristics of the victims, including their race and ages?
6. Are there any trends or patterns in victim demographics?

Geospatial analysis of the data can highlight geographic patterns and hotspots where fatal police shootings are more frequent. This information can be crucial for understanding the distribution of such incidents and potentially identifying areas in need of intervention or reform.

Also, analysing the data may reveal significant demographic disparities in fatal police shootings. These may include disparities in age, gender, and race, shedding light on potential areas for concern regarding equity and justice.

Findings:

Geographic Disparities: The geospatial analysis of the Washington Post dataset on police shootings uncovered significant regional disparities in these incidents. Notably, a concentration of incidents was observed along the west and east coasts, while central areas of the country had fewer incidents. This geographic variation prompts questions about the underlying factors contributing to this pattern, with population density emerging as a plausible explanation.

Age Distribution: Analysing age distributions within the dataset revealed intriguing patterns. The age distribution for White individuals displayed a left-skewed pattern in contrast to Black and Hispanic age distributions. T-tests confirmed that these mean age differences were statistically significant. This trend remained consistent when comparing the ages of Hispanic individuals with those of White individuals.

Statistically Significant Mean Age Differences: To delve deeper into these disparities, an analysis of variance was conducted for all possible race combinations present in the data. Additionally, Monte-Carlo simulations employed to validate mean age differences highlighted a striking pattern of an average age difference of approximately 7 years between Black and

White individuals involved in police shootings. This outcome was echoed when applying Monte-Carlo simulations to age differences between White and Hispanic individuals.

These findings not only shed light on these disparities but also emphasize their significance for future research and discussions surrounding this complex issue.

Discussions:

Upon receiving the Washington Post dataset on police shootings in Excel format, we embarked on a comprehensive exploration. Our first step involved geospatial mapping, although not all data points had geographic coordinates. Despite this limitation, this analysis unveiled regional disparities, with higher concentrations of incidents on the west and east coasts, sparking questions about the underlying reasons for this pattern, potentially linked to population density.

Next, we delved into the age distributions of the victims, considering all races and specific groups such as Black, White, and Hispanic individuals. This required careful data extraction and analysis of descriptive statistics. Notably, we observed left-skewed age distributions for White individuals, prompting further investigation to ascertain whether these differences were statistically significant.

To this end, we conducted T-tests to explore potential underlying factors for age differences. This analysis was extended to evaluate age disparities across various racial groups through an analysis of variance. Monte-Carlo simulations were used to assess the differences in means for White vs. Black ages and White vs. Hispanic ages, yielding insights into the approximately 7-year age difference between Black and White victims.

As a final validation, we presented age distribution plots for all races, reinforcing the finding that the average age for White individuals is higher than for other racial groups. This comprehensive exploration provides critical insights into regional disparities in fatal police shootings and age differences among racial groups.

Appendix A: Method

We obtained the Washington Post dataset on police shootings in an Excel format, comprising various data columns such as age, gender, race, geographic coordinates (latitude and longitude), and more. Upon a closer examination of the data, we discovered intriguing avenues for exploration.

Our initial step involved geospatial mapping of the data. Although the dataset included 8002 unique data points, the 'latitude' and 'longitude' columns had information for only 7162 points. Nonetheless, this geospatial analysis yielded valuable insights. We observed regional disparities in fatal police shootings, with a clear concentration of data points along the west and east coasts, while central areas of the country exhibited fewer incidents. This leads to the question of what factors contribute to this disparity, with one plausible explanation being the higher population density in coastal regions. This approach provided a geographical perspective on fatal police shootings, which is crucial for further investigations.

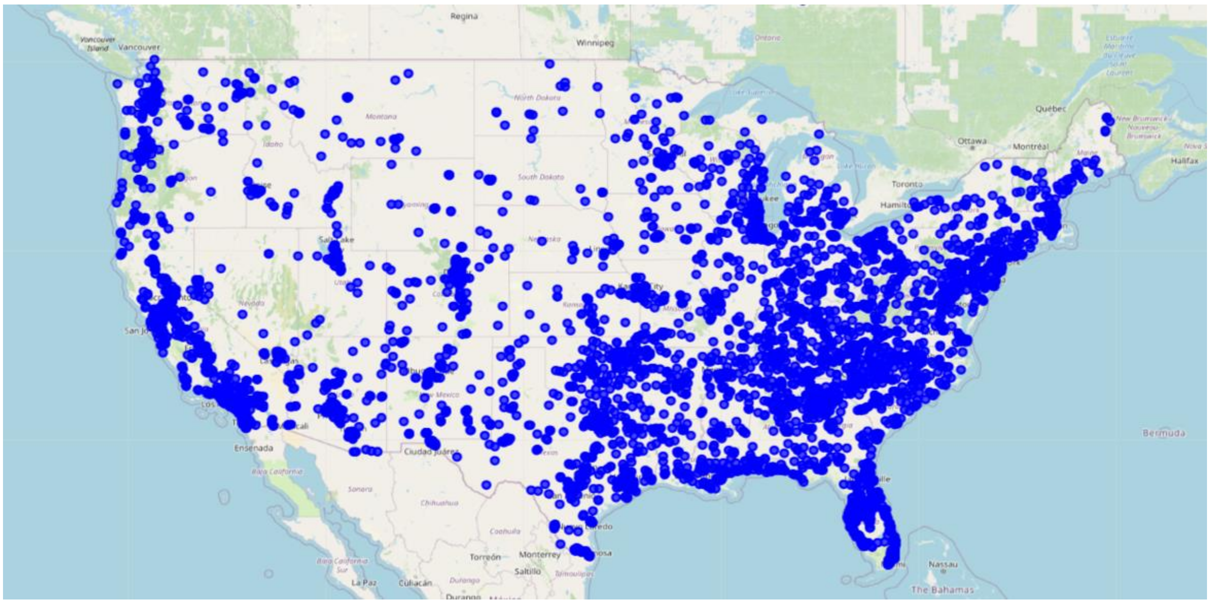
Our next approach involved plotting the age distributions. Initially, we created age distribution graphs for all races, followed by separate ones for Black, White, and Hispanic individuals. To do this, we extracted ages from cells with age entries and analysed the descriptive statistics of the age distribution. Notably, we observed that the age distribution for White individuals appeared more left-skewed compared to Black and Hispanic age distributions. To assess whether these differences in means were occurring by chance, we conducted T-tests. The same pattern held when comparing Hispanic ages with those of White individuals.

Further investigation was conducted to identify statistically significant groups with similar characteristics. An analysis of variance was performed on all possible combinations of races within the data. To validate mean age differences, Monte-Carlo simulations were run on White and Black ages, as well as White and Hispanic ages. Frequency distributions of the observed mean differences revealed an average of approximately 7 years of age difference between Black and White individuals shot. Similar outcomes were obtained when applying Monte-Carlo simulations to White and Hispanic ages.

Finally, to reinforce the observation of mean age differences, we plotted the age distribution for all races, reinforcing the finding that the average age for White individuals is higher than that of all other races.

Appendix B: Results

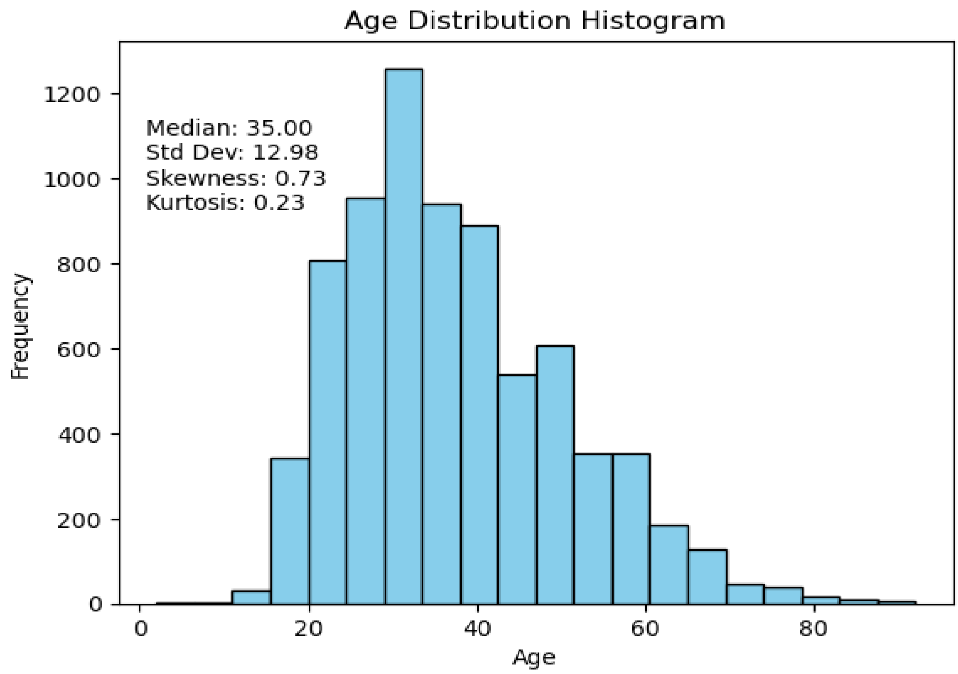
<u>Geospatial plot of the data using Folium:</u>



We noticed variations in fatal police shootings across regions, marked by a distinct clustering of data points on the west and east coasts, coupled with a lower incidence in central areas of the country.

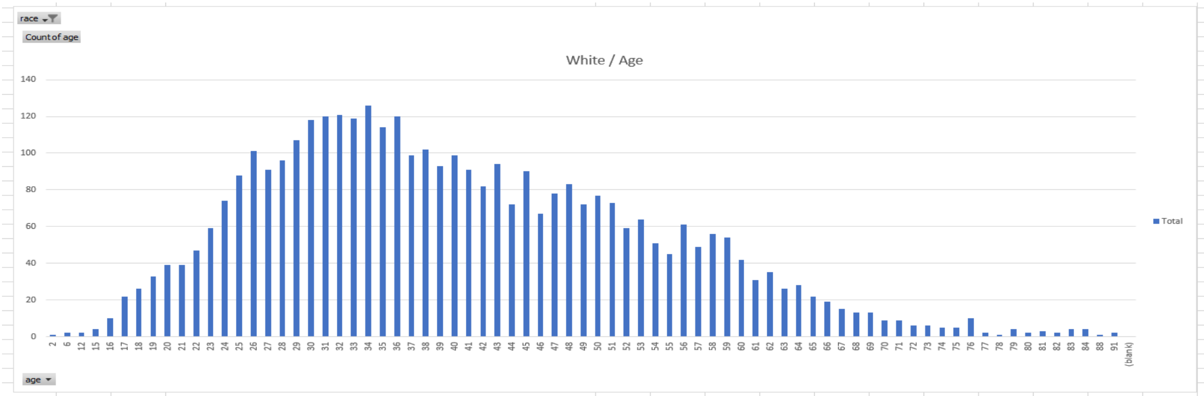
Distribution for ages of all races of people that were shot:

```
count    7499.000000
mean       37.209228
std        12.979490
min         2.000000
25%        27.000000
50%        35.000000
75%        45.000000
max        92.000000
Name: age, dtype: float64
```

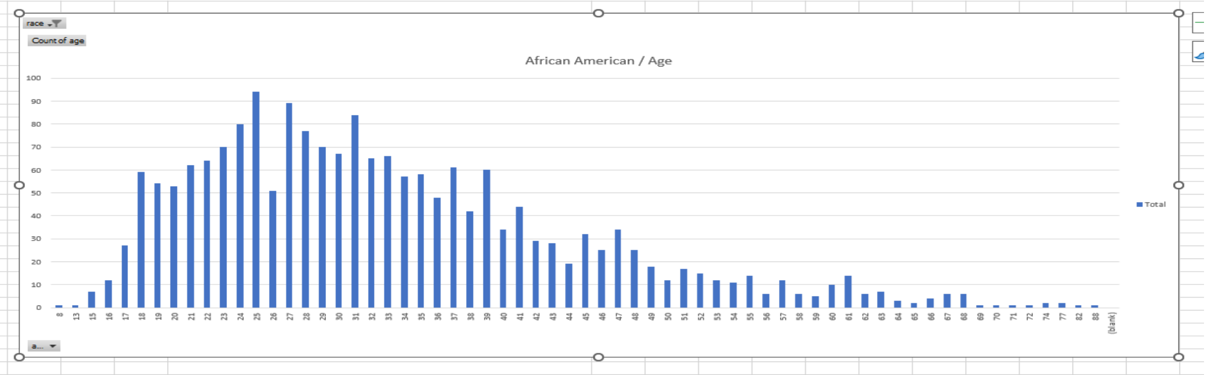


We observed a slight variation between the mean and median, hinting at a subtle right-leaning skew in the age distribution, with a skewness value of around 0.73. It's important to highlight that the kurtosis is near 3, indicating the absence of a distinct peak around the mean, and the distribution's tails are not substantially skewed. If we were to graph a normal distribution with the same mean and standard deviation, it would closely resemble the observed age distribution.

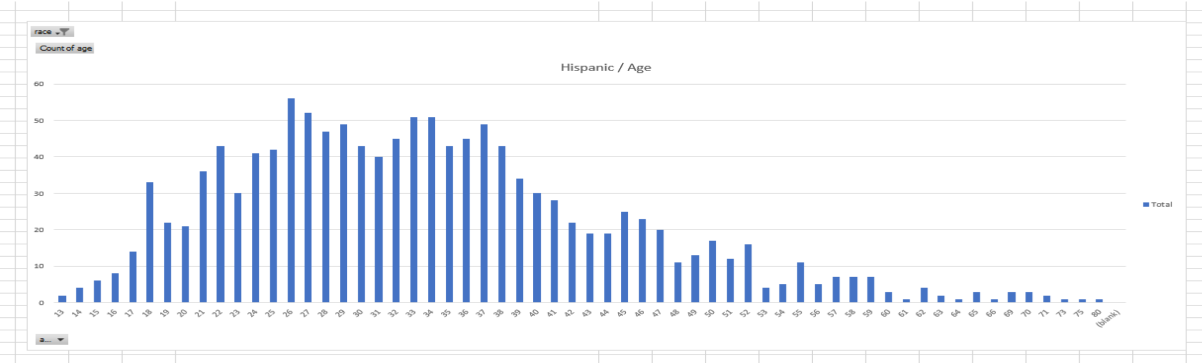
Distribution for ages of white people that were shot:



Distribution of Black People that were shot



Distribution of Hispanic People that were shot



As we can see from this data, the graphs for White ages is more left skewed than that for Black ages or that for Hispanic. From a measure of their descriptive statistics, we can see that difference clearly.

		Ages	
	Black	Hispanic	White
Mean	33.0129	33.7486	40.206
Median	31	33	38
Std Dev	11.4415	10.8634	13.1479

.

So, we further checked by doing T tests to check if the differences in the means is occurring by chance or if there is some underlying reason.

```
> comp1

        Welch Two Sample t-test

data:  bage and wage
t = -20.071, df = 3976.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.900401 -6.494292
sample estimates:
mean of x mean of y
 32.92812  40.12546
```

Therefore, we can tell by the T value that this is significant, even the P value indicates its highly unlikely for such a difference to occur by such chance.

The same is true for comparing Hispanic ages and those of White people.

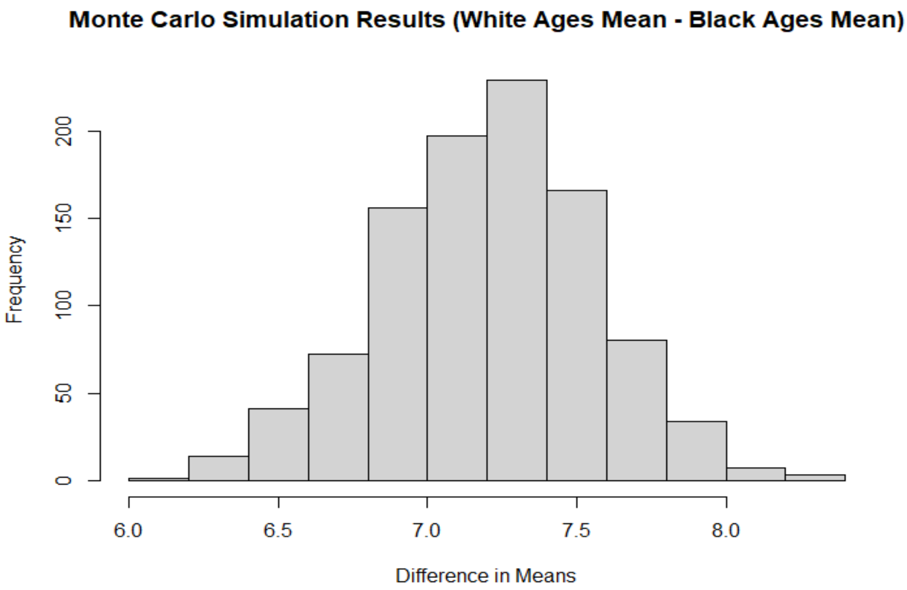
```
> comp2

        Welch Two Sample t-test

data:  hage and wage
t = -16.588, df = 2402.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.307128 -5.762139
sample estimates:
mean of x mean of y
 33.59083  40.12546
```

If we do a further analysis to check if we are missing out on an any statistically significant groups showing similar characteristics, we do an analysis of variance on all possible combinations of races possible in our data.

```
R 4.3.1 · ~/
Residuals   6344 943622     149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1652 observations deleted due to missingness
> TukeyHSD(a1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = age ~ race, data = fatal_police_shootings_data)

$race
          diff         lwr        upr     p adj
B-A -3.0318841  -6.2521133  0.1883451 0.0785797
H-A -2.3691711  -5.6456080  0.9072658 0.3079860
N-A -3.3095146  -7.9359263  1.3168972 0.3201318
O-A -2.4863158 -11.0468312  6.0741996 0.9624301
W-A  4.1654624   0.9965842  7.3343406 0.0024871
H-B  0.6627130  -0.6663834  1.9918094 0.7140416
N-B -0.2776305  -3.8039805  3.2487194 0.9999224
O-B  0.5455683  -7.4740254  8.5651620 0.9999624
W-B  7.1973465   6.1613697  8.2333232 0.0000000
N-H -0.9403435  -4.5180951  2.6374081 0.9756437
O-H -0.1171447  -8.1594731  7.9251836 1.0000000
W-H  6.5346335   5.3352960  7.7339709 0.0000000
O-N  0.8231988  -7.8571061  9.5035036 0.9998057
W-N  7.4749770   3.9954573 10.9544966 0.0000000
W-O  6.6517782  -1.3473340 14.6508904 0.1668247

> |
```

Given from the P values we can see that our pairs of White - Black and White-Hispanic show the most statistically significant difference so we are on the right track.

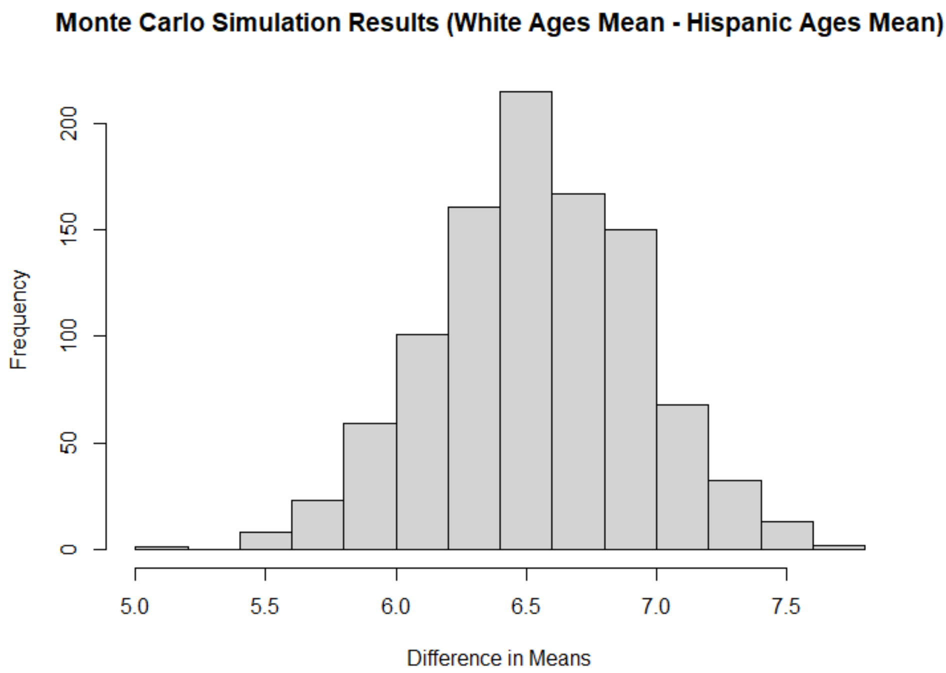
Running monte carlo simulations on the data for white ages and black ages, and for white and hispanic ages, we plot the frequency distribution in the differences of means observed over the various simulations. This takes the shape of roughly normal distributions.



The graph for black ages clearly shows that we are close to an average of roughly 7 years age difference for black and white people shot. The data also shows that the means difference 95% of the time is 7.1 years.

```
> # Display the p-value
> cat("Observed Mean Difference:", observed_mean_diff, "\n")
Observed Mean Difference: 7.197346
> cat("Proportion of Simulated Differences >= Observed Difference:", p_value, "\n")
Proportion of Simulated Differences >= Observed Difference: 0.522
> summary(observed_mean_diff)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  7.197   7.197   7.197   7.197   7.197   7.197 
>
```

We conduct similar experiments using the Monte-Carlo simulation for the White and Hispanic ages and we find similar results.

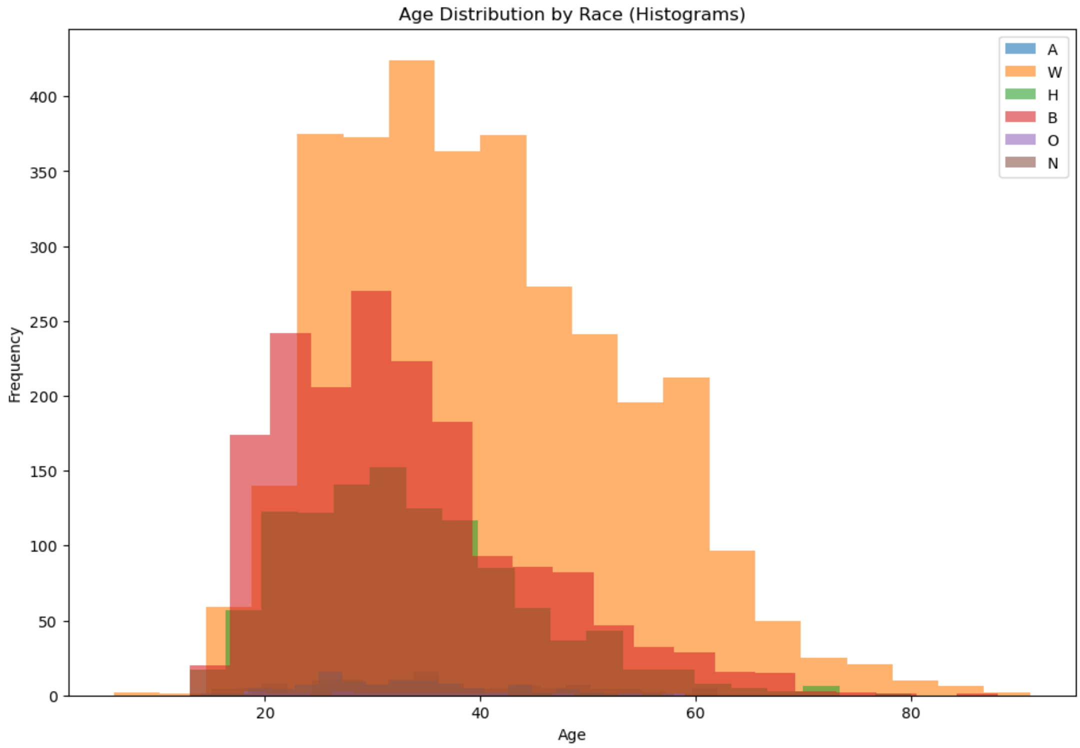


Here the graphs show an average of around a 6.5 year difference in the means over a large number of simulations from the data.

It also shows that for 95% of the time, the data shows a difference in mean ages of 6.5 years for Hispanic people.

```
> summary(observed_mean_diff)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.535   6.535   6.535   6.535   6.535   6.535 
> 
```

Age Distribution by Race (All Races):



Descriptive Statistics:

race	count	mean	std	min	25%	50%	75%	max
A	125.0	35.960000	11.592127	15.0	27.0	35.0	45.0	62.0
B	1725.0	32.928116	11.388649	13.0	24.0	31.0	39.0	88.0
H	1134.0	33.590829	10.743505	13.0	26.0	32.0	39.0	80.0
N	103.0	32.650485	8.994234	14.0	26.5	32.0	37.5	58.0
O	19.0	33.473684	11.796273	18.0	25.5	31.0	40.5	59.0
W	3244.0	40.125462	13.162144	6.0	30.0	38.0	49.0	91.0

To further confirm the mean difference in ages we plotted the age distribution of all races which displays that the average ages for White people is higher than all the other races.

Appendix C: code

Code for geospatial plot using Folium:

```
import pandas as pd
import folium
from folium.plugins import MarkerCluster

# Specify the file path
```

```
file_path = r'C:\Users\Tiyasa\Desktop\Courses_Sem1\MTH 522\fatal-police-shootings-data.xls'

# Load the Excel file into a DataFrame
df = pd.read_excel(file_path)
df.head(5)

df.dropna(subset=['latitude', 'longitude'], inplace=True)

# Create a Folium map centered around the United States
m = folium.Map(location=[37.0902, -95.7129], zoom_start=4)  # Coordinates for the United States

# Add dots for each data point
for index, row in df.iterrows():
    folium.CircleMarker(
        location=[row['latitude'], row['longitude']],
        radius=5,  # Adjust the radius as needed
        color='blue',  # Customize the color of the dots
        fill=True,
        fill_color='blue',
        fill_opacity=0.7,
    ).add_to(m)

# Display the map
m.save('map3.html')  # Save the map as an HTML file
```

Code for age distribution of Black and White people:

```
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats

# Specify the file path
file_path = r'C:\Users\Tiyasa\Desktop\Courses_Sem1\MTH 522\fatal-police-shootings-data.xls'

# Load the Excel file into a DataFrame
df = pd.read_excel(file_path)

# Filter data for Black and White people
black_data = df[df['race'] == 'B'].dropna(subset=['age'])
white_data = df[df['race'] == 'W'].dropna(subset=['age'])

# Plot histograms for Black and White people's ages
plt.figure(figsize=(12, 6))  # Adjust figure size as needed

plt.subplot(1, 2, 1)  # Create the left subplot
plt.hist(black_data['age'], bins=20, color='skyblue', edgecolor='black')
plt.xlabel('Age')
plt.ylabel('Frequency')
```

```
plt.title('Age Distribution for Black People')

plt.subplot(1, 2, 2)  # Create the right subplot
plt.hist(white_data['age'], bins=20, color='lightcoral', edgecolor='black')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution for White People')

# Calculate descriptive statistics for Black people's ages
statistics_black = black_data['age'].describe()
median_black = black_data['age'].median()
std_dev_black = black_data['age'].std()
skewness_black = stats.skew(black_data['age'])
kurtosis_black = stats.kurtosis(black_data['age'], fisher=False)

# Calculate descriptive statistics for White people's ages
statistics_white = white_data['age'].describe()
median_white = white_data['age'].median()
std_dev_white = white_data['age'].std()
skewness_white = stats.skew(white_data['age'])
kurtosis_white = stats.kurtosis(white_data['age'], fisher=False)

# Adjust layout and show the plots
plt.tight_layout()
plt.show()

# Print the statistics
print("Statistics for Black People's Ages:")
print(statistics_black)
print(f"Median: {median_black:.2f}")
print(f"Standard Deviation: {std_dev_black:.2f}")
print(f"Skewness: {skewness_black:.2f}")
print(f"Kurtosis: {kurtosis_black:.2f}")
print("\nStatistics for White People's Ages:")
print(statistics_white)
print(f"Median: {median_white:.2f}")
print(f"Standard Deviation: {std_dev_white:.2f}")
print(f"Skewness: {skewness_white:.2f}")
print(f"Kurtosis: {kurtosis_white:.2f}")

# Calculate the mean age for Black and White people
mean_age_black = black_data['age'].mean()
mean_age_white = white_data['age'].mean()

# Calculate the mean age difference
mean_age_difference =  mean_age_white - mean_age_black

# Print the mean ages and mean age difference
print(f"\nMean Age for White People: {mean_age_white:.2f}")
print(f"Mean Age for Black People: {mean_age_black:.2f}")
print(f"Mean Age Difference (White - Black): {mean_age_difference:.2f}")
```

Contribution:

Tiyasa Saha: Worked on the Issues, Findings, Discussion, Methods, Code and Results sections. Also self-plotted the graphs to analyse the data using the various methods discussed in the report.

Kanishka Patre: Worked on the Issues, Findings, Methods, Code and Result sections. Plotted graphs and used various regression analysis models and tests to analyse the data.

Srikanth Koncherry: Worked on identifying issues, writing code for and the analysis models and producing the graphs for them

Gautam Marathe: Worked on initial analyses, cleaning data, analysing, and looking for different models to fit the data on, testing various fits for their errors, testing non-linear models on the data to describe trends between predictors.