title: “Midterm EDA: Group 7” author: “Alex Khater, Rajeev Koneru, and Cora Martin” date: “2022-11-05” output: pdf_document: default word_document: default html_document: default

Introduction

In response to a severe lack of reporting within government sources, The Washington Post compiled a database of every fatal police shooting in the United States from 2015-2022. We are interested in exploring this data, specifically as it relates to differences between U.S. states and regions.

This exploratory data analysis is divided into four main parts: first, we organize the data; second, we perform some basic statistical analyses; third, we reshape the data for state- and region-based comparative analyses; fourth, we ask a SMART research question about our data and attempt to answer this question.

Part 1: Setting Up the Data

First we call our packages. Then we read the data set that comes from a csv file called FPS22.csv.

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

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

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

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x plotly::filter() masks dplyr::filter(), stats::filter()
## x dplyr::lag()     masks stats::lag()
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

After accounting for null values, the data set we are working with has 6,574 observations. Below we have provided a single sample observation:

Name	Date	Manner of Death	Armed	Age	Gender	Race	City
Tim Elliot	10/04/2022	Shot	Gun	53	M	A	Shelton

State	Signs of Mental Illness	Threat Level	Flee	Body Camera	Longitude	Latitude	Is Geocoding Exact?
WA	1	TRUE	Not fleeing	FALSE	-123	47.2	TRUE

The total number of observations:

```
## [1] 6574
```

Part 2: Basic Statistics

We provide some basic statistics about 2015-2022 fatal police shootings in the United States, using information from the Washington Post data set.

Mean age of victims of police violence:

```
## [1] 37.2
```

Median age of victims of police violence:

```
## [1] 35
```

Figure 1 Frequency graph for the age of victims of police violence:

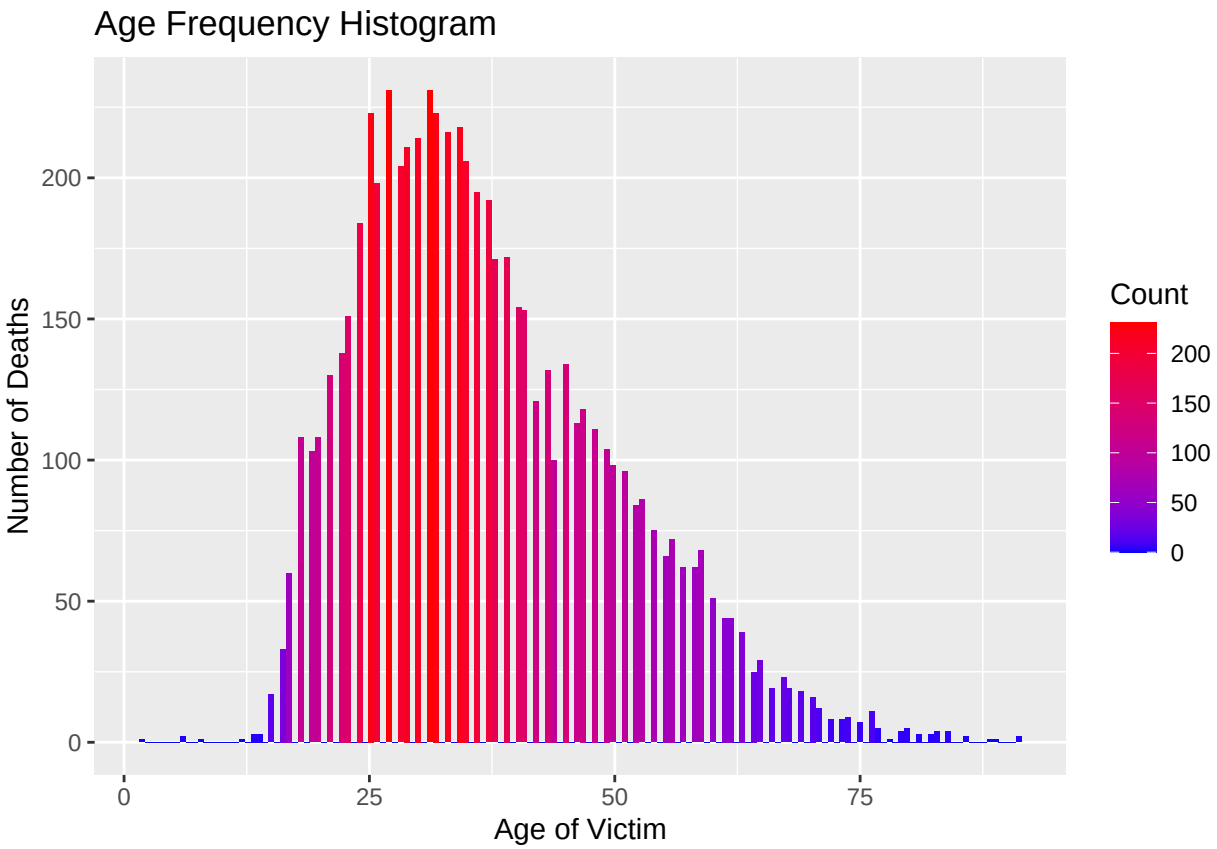


Figure 2 Frequency graph for the race of victims of police violence:

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

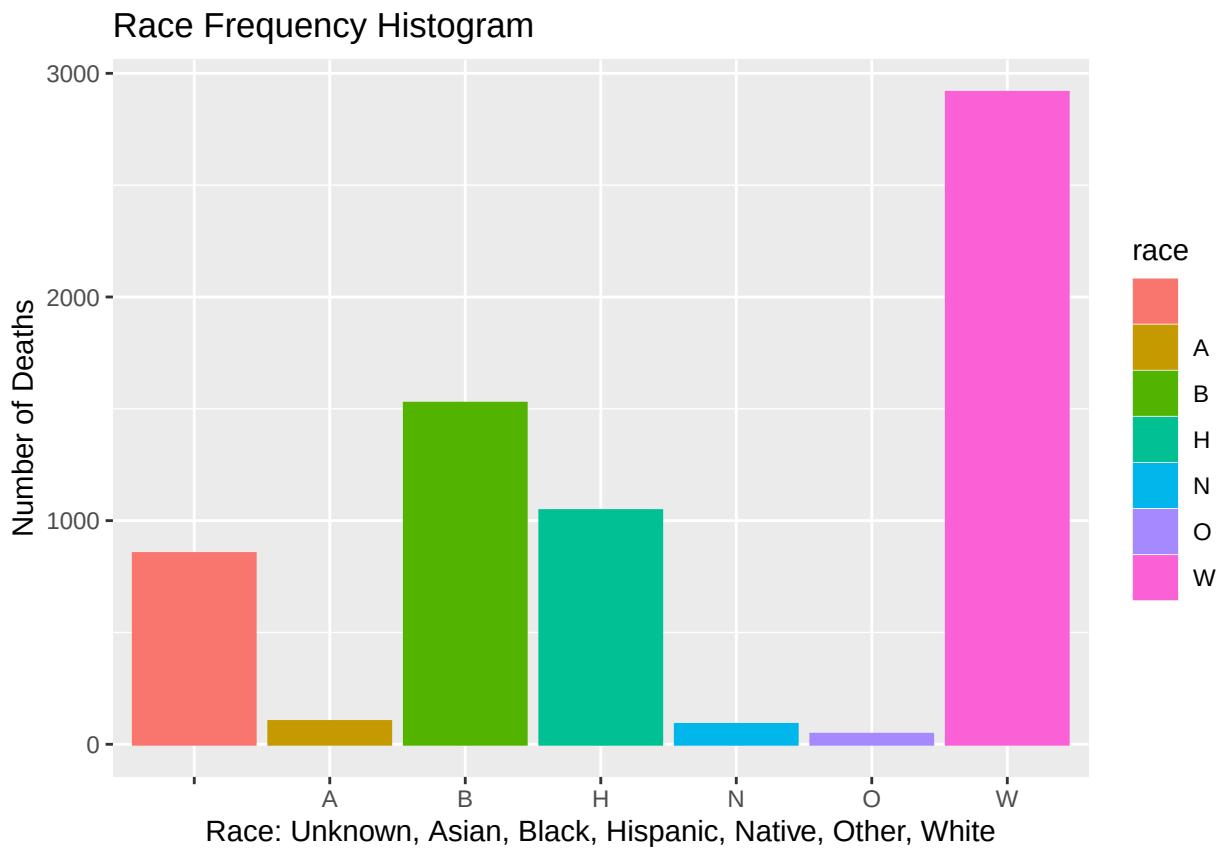


Figure 3 Frequency graph for the gender of victims of police violence:

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

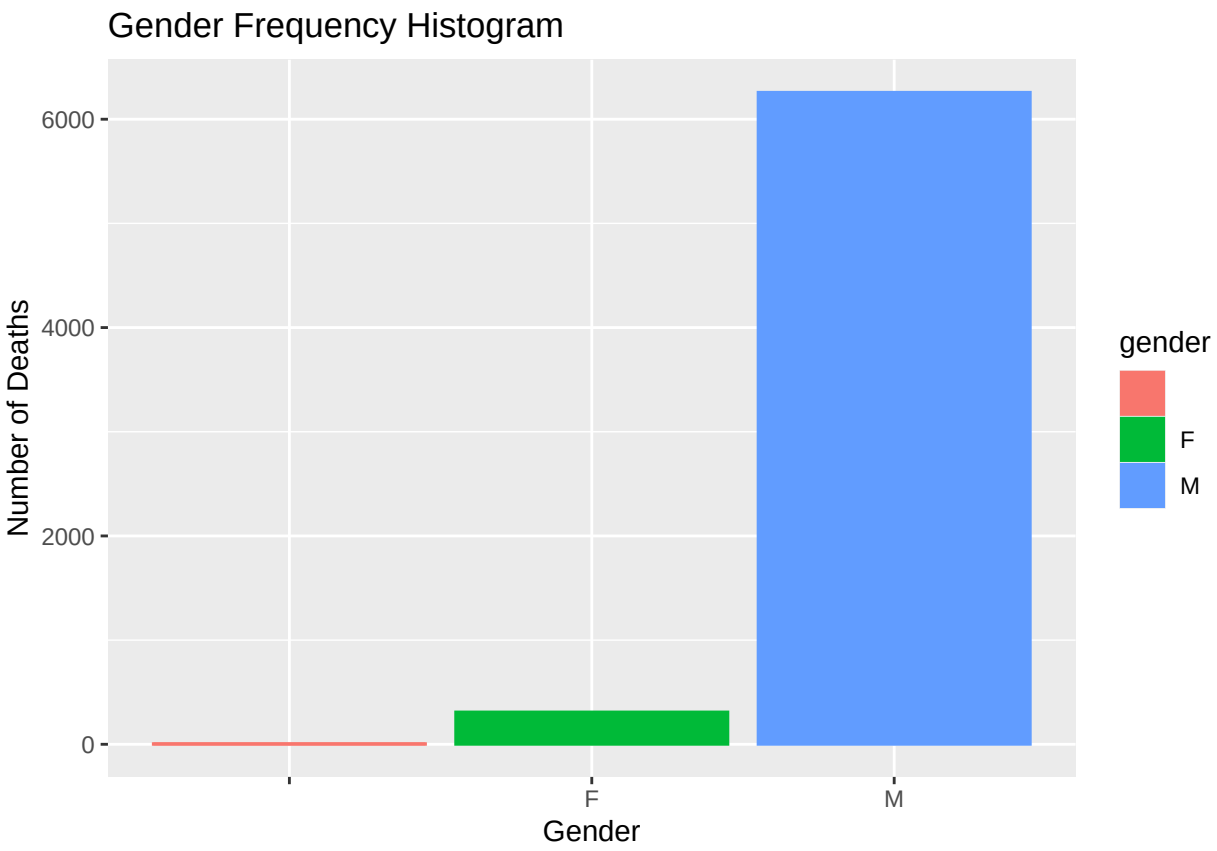


Figure 4 Frequency graph for the manner of death of victims of police violence:

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

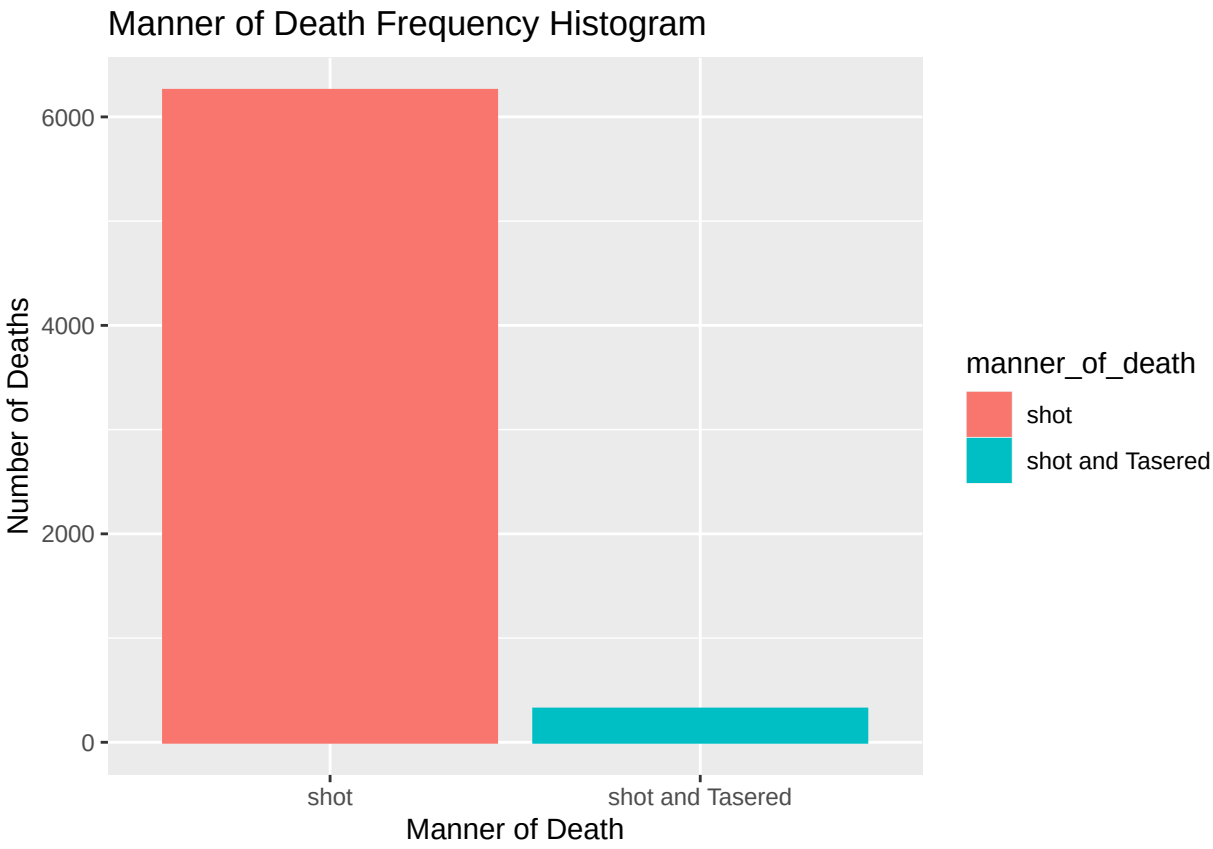


Figure 5 Frequency graph for the threat level of victims of police violence:

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

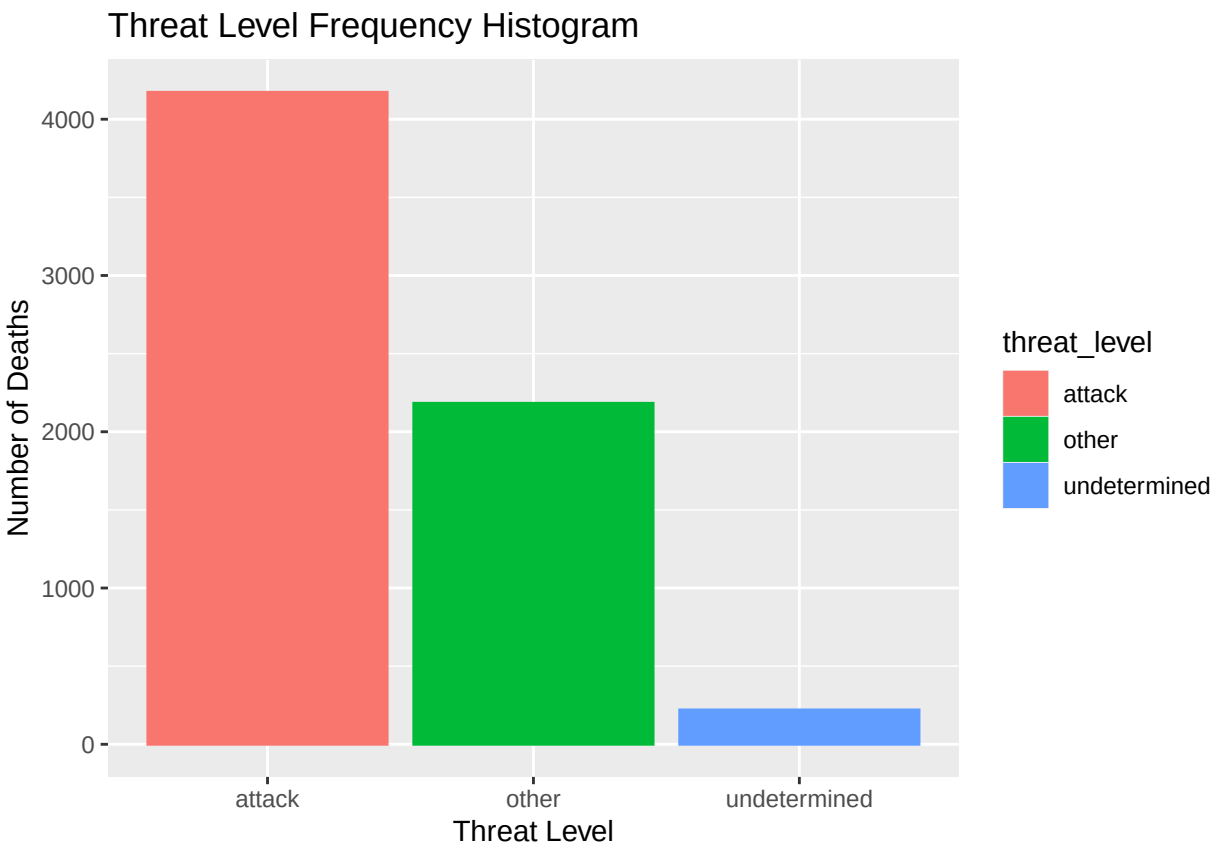


Figure 6 Hover over the map below to see the breakdown of fatal police shootings, divided by the race of the victim. We looked at the total number of deaths in each state by race and following are some of the insights:

1) We see that the state with the highest level of victims of police violence is California with a total of 885 victims, followed by Texas with a total of 553 and then Florida with 427.

2) These results are consistent with the populations of these states, with the highest being California, then Texas, and then Florida.

3) We also observe that the highest number of deaths is for Hispanic people in California, whereas in Texas and Florida there are more fatal shootings of White people.

Figure 7 Now we look at the age of the suspect shot, as well as their race. We made the following observations:

1) We see from the boxplot below that the median age for Black people that have been killed by police is 29 years.

2) White people have a relatively higher median age of 35 years whereas Asian people have the highest median age of around 38 years.

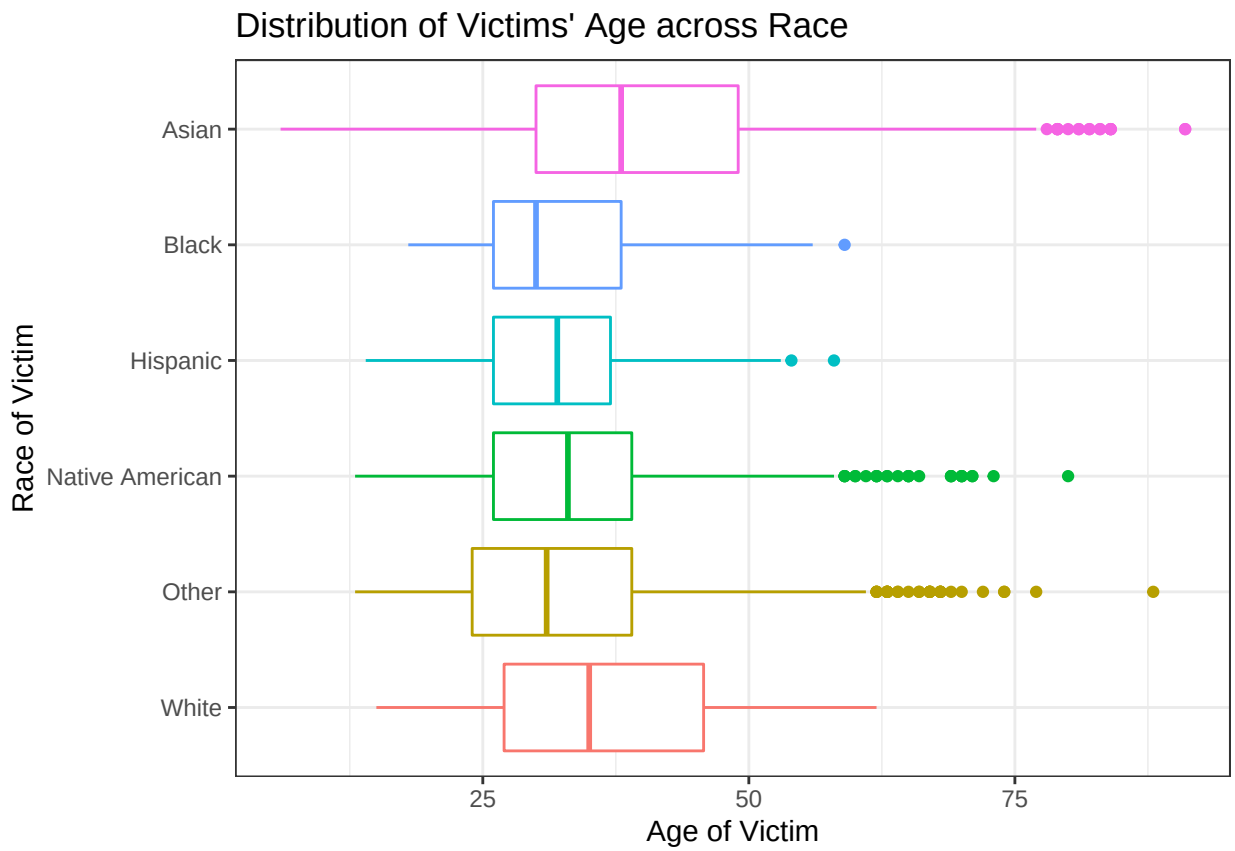


Figure 8 If we look at the age of each victim against the status of their mental health, we can make the following observation: signs of mental illness appear more frequently within the 30s age range while death by police for people age 50 and above are more common for people showing signs of mental illness.

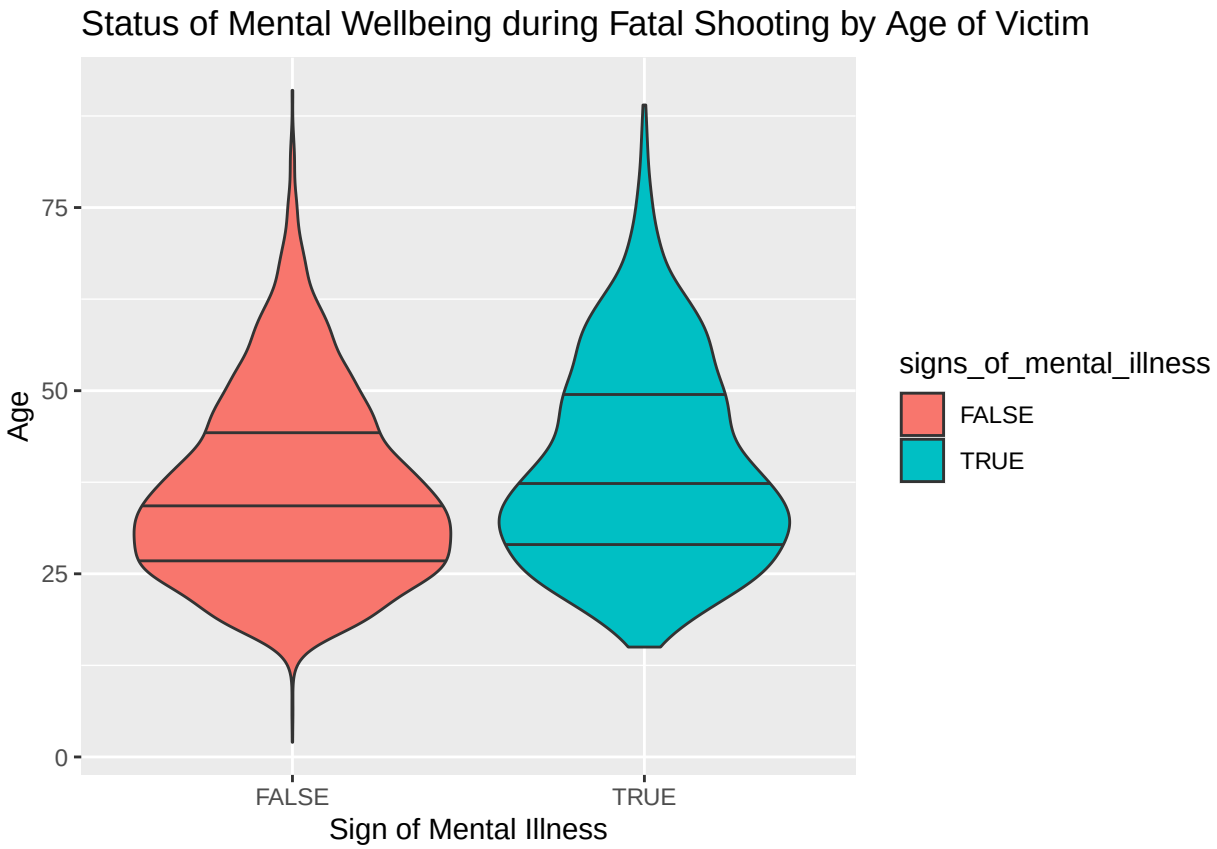


Figure 9 We also looked at the death by race and gender, coming up with the following insight: individuals across all races that were shot and killed by police were more often men.

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

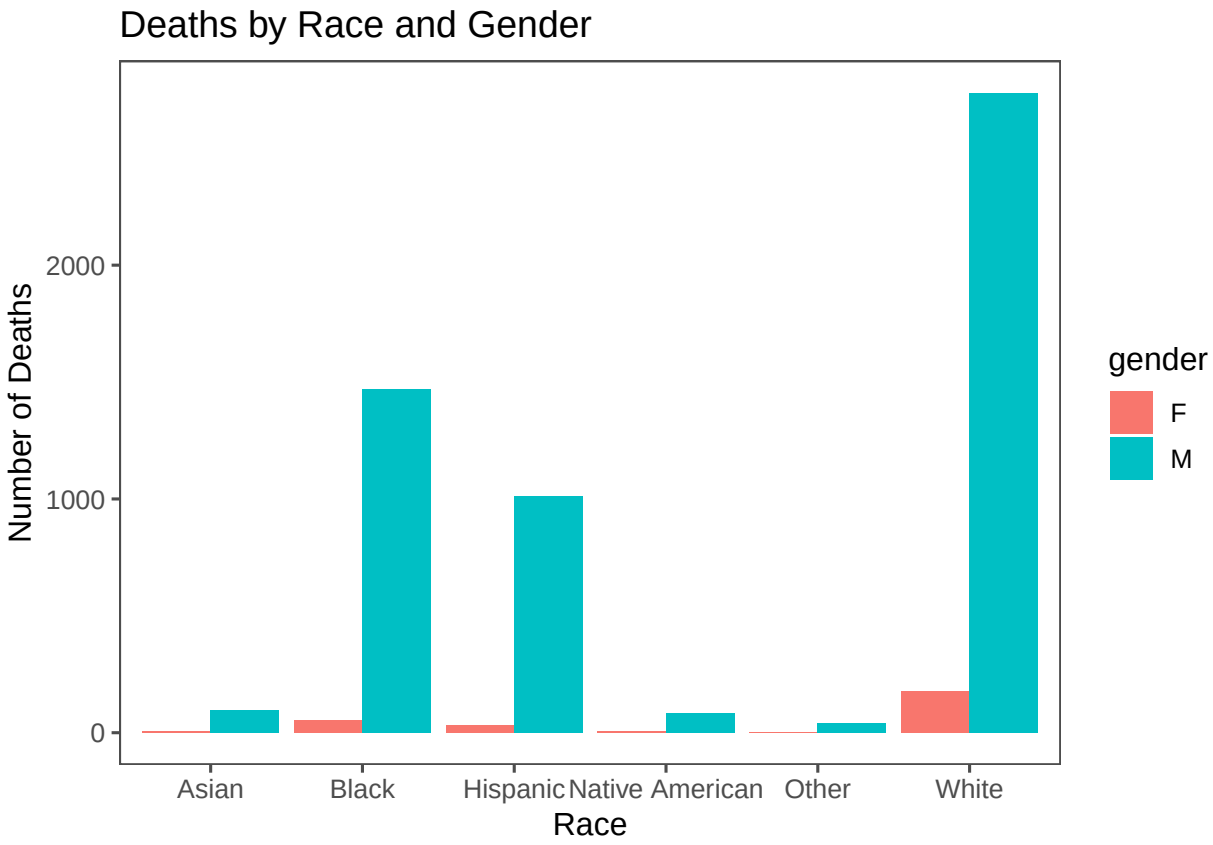


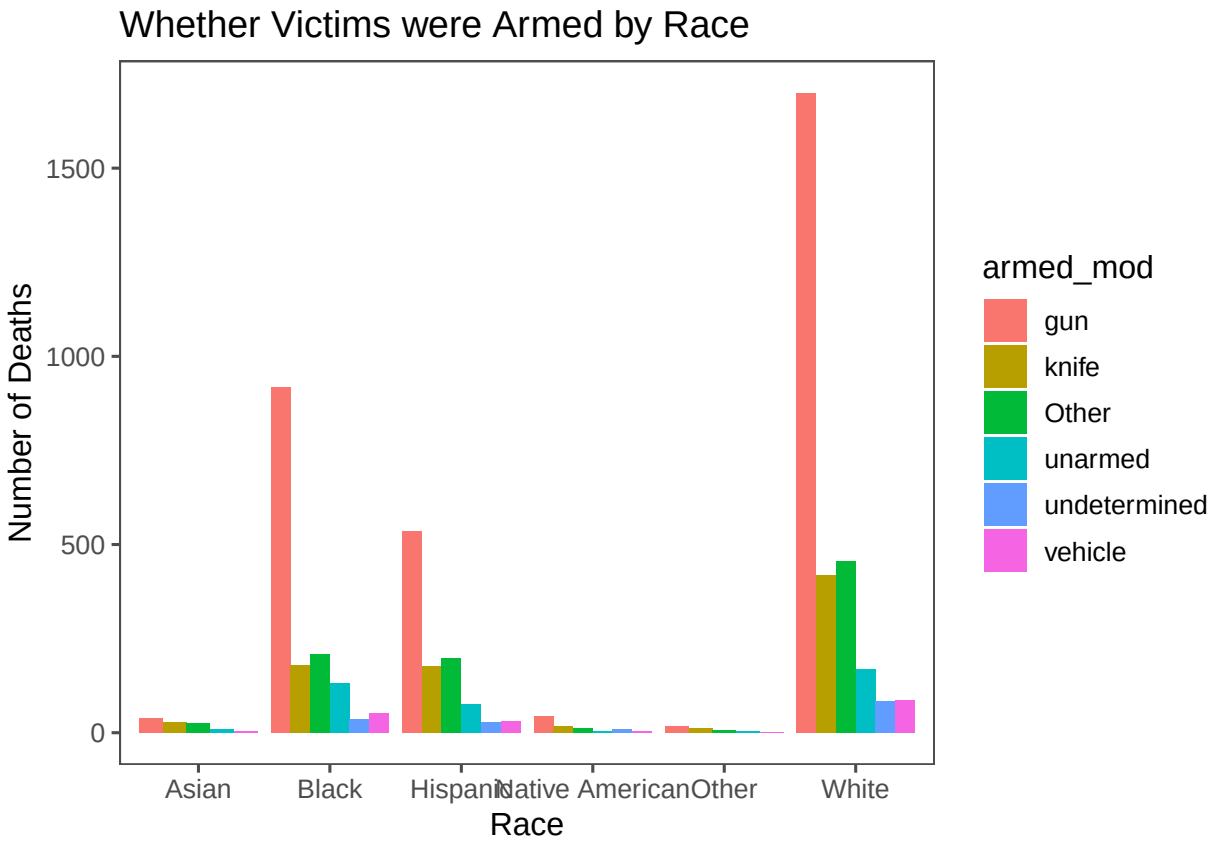
Figure 10 We then looked at the distribution of deaths by race and the top 5 armed categories. We discovered that around 9% of the Black victims were unarmed whereas only approximately 6% of the White victims were unarmed. Guns were the most used weapon across all races except for Asian individuals. Asian victims were more often wielding knives.

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

Distribution of Deaths by Armed Category and Race:

```
##   race  gun knife Other unarmed undetermined vehicle
## 1    A 38.2  27.4  23.5    7.84         0.00    2.94
## 2    B 60.1  11.7  13.6    8.66         2.43    3.41
## 3    H 51.2  16.9  19.0    7.37         2.58    2.87
## 4    N 50.6  18.0  13.5    5.62         8.99    3.37
## 5    O 40.0  28.9  15.6   11.11         0.00    4.44
## 6    W 58.3  14.4  15.6    5.80         2.88    2.98
```

Figure 11 The following graph illustrates the deaths per year by race from 2015-2022:



We looked at the distribution of deaths by suspects' race and whether they were trying to flee or not. The following are some of our most interesting observations:

1) Only 53% of Black victims shot were not fleeing whereas 71% of the Asian victims who were shot were not trying to flee.

2) The car is the most popular method of fleeing among White victims whereas for Black victims, the most popular method of fleeing was by foot.

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if `.name_repair` is
## Using compatibility `.name_repair`.
```

Number of deaths by victims' status (fleeing or not fleeing) by race:

```
##   race    V1  Car  Foot Not fleeing Other
## 1    A  7.84 11.8 10.78        68.6  0.98
## 2    B  7.08 15.5 19.28        54.4  3.74
## 3    H  7.27 16.2 13.78        57.9  4.88
## 4    N 13.48 11.2 17.98        52.8  4.49
## 5    O  2.22 17.8 11.11        64.4  4.44
## 6    W  8.40 15.6  9.95        62.7  3.33
```

```
## [1] "character"
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

Figure 12 This graph shows the victims of police violence by race over time (2015-2022):

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

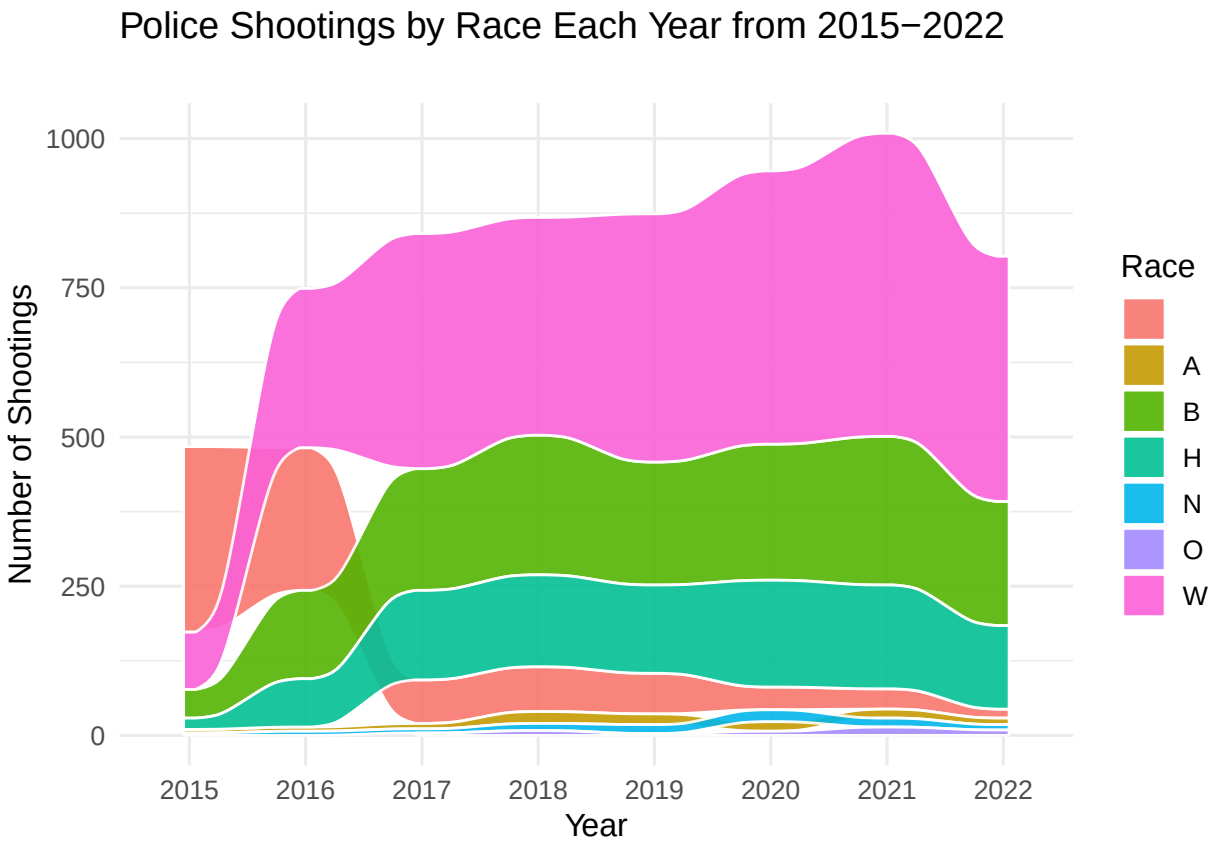


Figure 13 Surprisingly, there is seasonality across years or months in police shootings. We looked into the monthly trend over 8 years and used ARIMA to forecast the likely number of police shootings over the next four months.

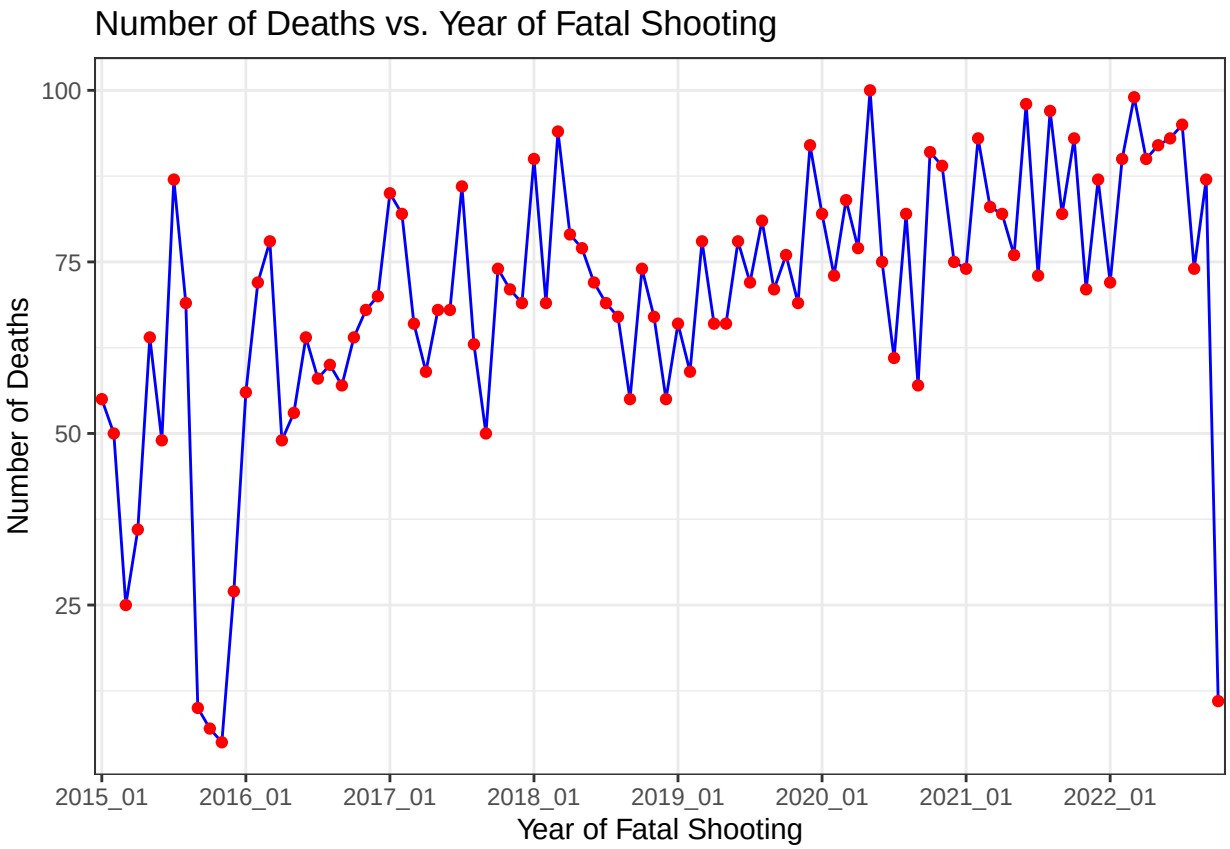
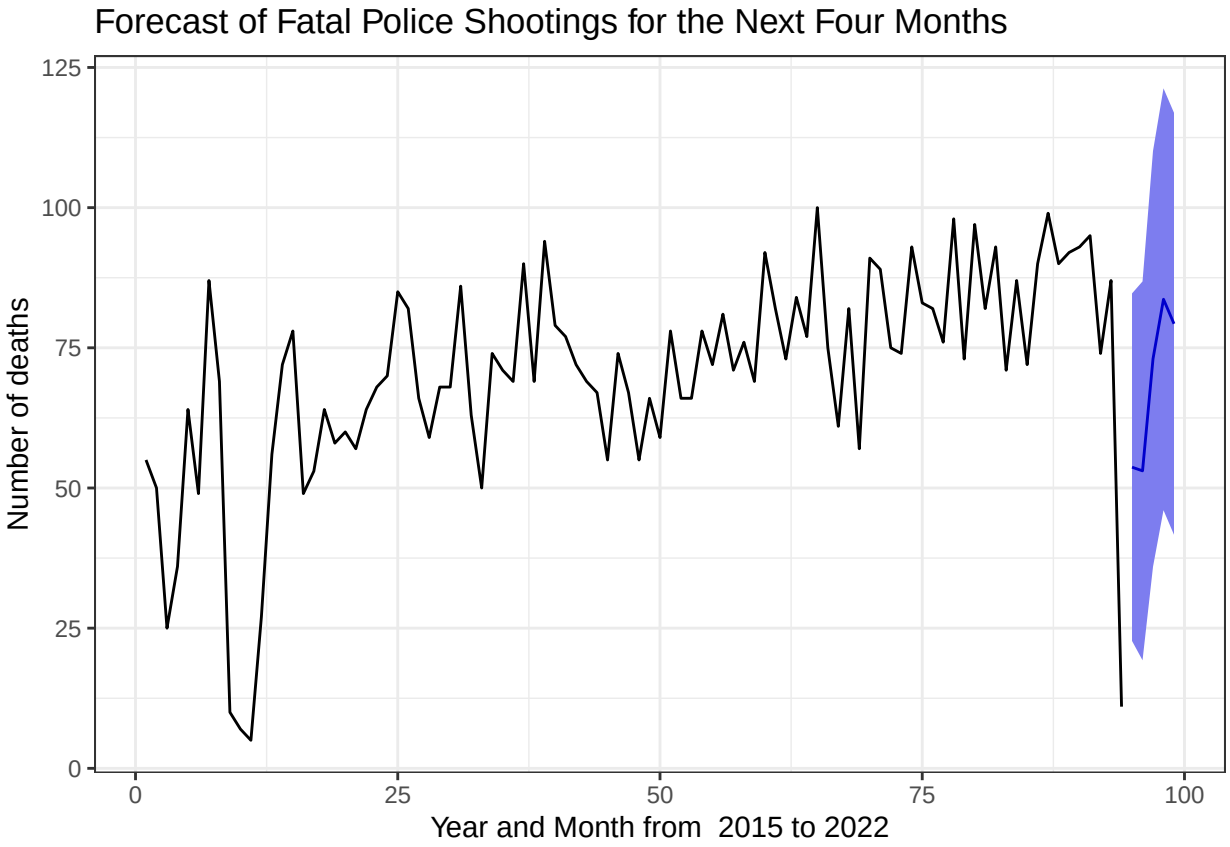


Figure 14 The forecast predicts average shootings for the next four months with a wide confidence interval.



Part 3: Reshaping the Data for State and Regional Comparative Analysis

After pursuing the above exploratory analysis, we decided to do some comparative analyses between states and regions to create a specific, measureable, achievable, relevant, and time-oriented research question to pursue for the remainder of the project.

To do this, wee began by dividing the data into regions for easier visualization and comparative analysis. The regions divide each US state as follows:

Northwest (NW)	Southwest (SW)	Midwest (MW)	Southeast (SE)	Northeast (NE)
California	New Mexico	Illinois	Georgia	New York
Washington	Arizona	Wisconsin	Alabama	Rhode Island
Oregon	Texas	Indiana	Mississippi	Maryland
Nevada	Oklahoma	Michigan	Louisiana	Vermont
Idaho	Hawaii	Minnesota	Tennessee	Pennsylvania
Utah	-	Missouri	North Carolina	Maine
Montana	-	Iowa	South Carolina	New Hampshire
Colorado	-	Kansas	Florida	New Jersey
Wyoming	-	North Dakota	Arkansas	Connecticut
Arkansas	-	South Dakota	West Virginia	Massachusetts
Arkansas	-	Nebraska	DC	-
-	-	Ohio	Virginia	-

Fatal shootings in the Northwest United States:

```
## [1] 1810
```

Fatal shootings in the Southwest United States:

```
## [1] 1226
```

Fatal shootings in the Midwest United States:

```
## [1] 1080
```

Fatal shootings in the Southeast United States:

```
## [1] 1890
```

Fatal shootings in the Northeast United States:

```
## [1] 568
```

We then created two sub-data sets by grouping the data by state and by region for visualization purposes. The contents of both groups are identical, besides their grouping.

Part 4: SMART Question and Answer

Within our data set of 6,574 observations of police shootings from 2015 to 2022 in the United States, is there a correlation between the U.S. state of observation and whether a body camera was turned on during the shooting?

First let's take a look at our data after it has been grouped by state and reorganized into the following variables:

Variable	Meaning
state	State of observation
region	Region of observation
stbcp	Body camera on proportion by state
genp.p	Proportion of male victims by state
smi.p	Proportion of victims by state with signs of mental illness
flee.p	Proportion of victims by state the were fleeing
att.p	Proportion of victims by state that were attacking
armed.p	Proportion of victims by state that were armed
MoD.p	Proportion of victims by state that were shot
age.avg	Average age by state
Non_White_Prop	Proportion of non-White victims by state

The state data subgroup can be summarized as follows:

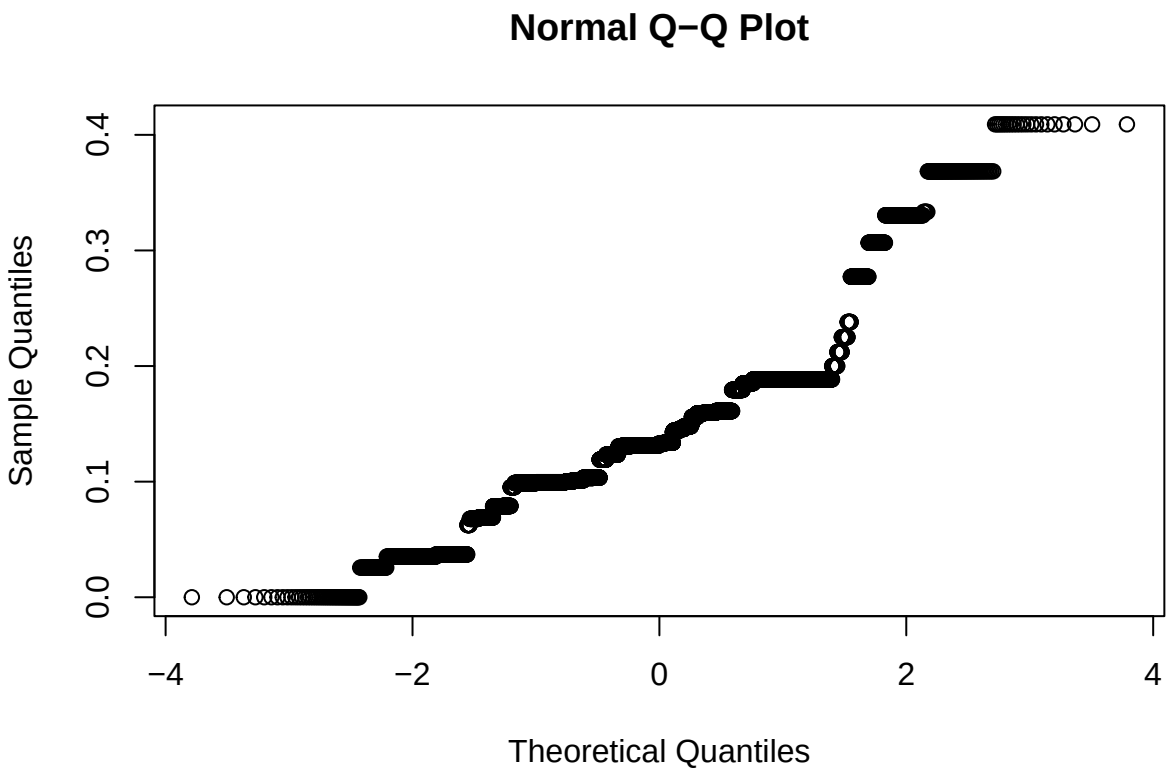
```
##     state              month              year           regions
##  Length:6574        Length:6574        Length:6574        MW:1080
##  Class :character   Class :character   Class :character   NE: 568
##  Mode  :character   Mode  :character   Mode  :character   NW:1810
```

```
##                                                       SE:1890
##                                                       SW:1226
##
##      stbcp           gen.p           smi.p           flee.p      att.p
##  Min.   :0.000   Min.   :0.818   Min.   :0.000   Min.   :0   Min.   :0.350
##  1st Qu.:0.101   1st Qu.:0.938   1st Qu.:0.200   1st Qu.:0   1st Qu.:0.564
##  Median :0.133   Median :0.952   Median :0.219   Median :0   Median :0.644
##  Mean   :0.144   Mean   :0.952   Mean   :0.223   Mean   :0   Mean   :0.635
##  3rd Qu.:0.183   3rd Qu.:0.966   3rd Qu.:0.265   3rd Qu.:0   3rd Qu.:0.679
##  Max.   :0.409   Max.   :1.000   Max.   :0.556   Max.   :0   Max.   :1.000
##     armed.p          MoD.p          age.avg     Non_White_prop
##  Min.   :0.778   Min.   :0.810   Min.   :33.1   Min.   :0.250
##  1st Qu.:0.918   1st Qu.:0.938   1st Qu.:35.7   1st Qu.:0.455
##  Median :0.934   Median :0.948   Median :36.9   Median :0.563
##  Mean   :0.937   Mean   :0.951   Mean   :37.2   Mean   :0.557
##  3rd Qu.:0.958   3rd Qu.:0.969   3rd Qu.:38.6   3rd Qu.:0.635
##  Max.   :1.000   Max.   :1.000   Max.   :44.4   Max.   :0.939
```

The region data subgroup can be summarized as follows:

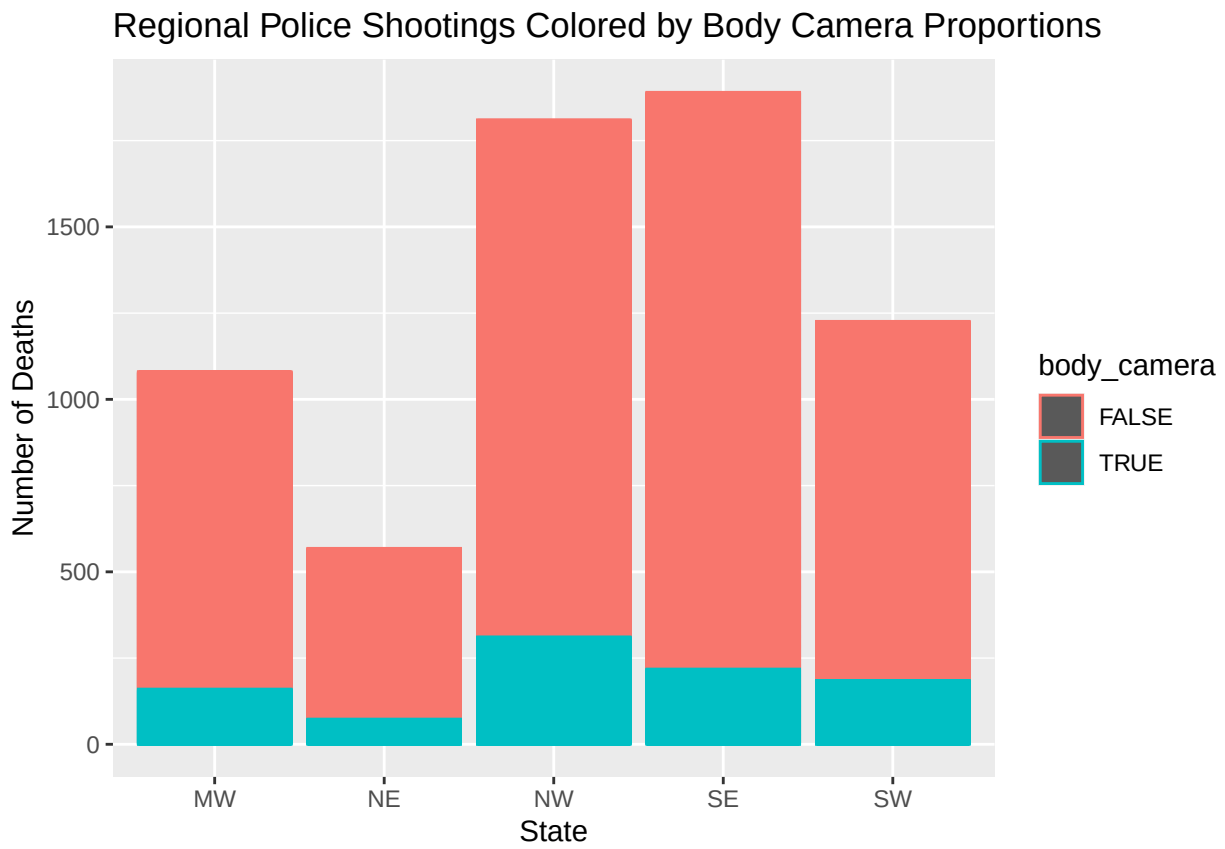
```
##     state              month               year              stbcp
##  Length:6574        Length:6574        Length:6574        Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:0.101
##  Mode  :character   Mode  :character   Mode  :character   Median :0.133
##                                                           Mean   :0.144
##                                                           3rd Qu.:0.183
##                                                           Max.   :0.409
##      gen.p           smi.p           flee.p      att.p          armed.p
##  Min.   :0.818   Min.   :0.000   Min.   :0   Min.   :0.350   Min.   :0.778
##  1st Qu.:0.938   1st Qu.:0.200   1st Qu.:0   1st Qu.:0.564   1st Qu.:0.918
##  Median :0.952   Median :0.219   Median :0   Median :0.644   Median :0.934
##  Mean   :0.952   Mean   :0.223   Mean   :0   Mean   :0.635   Mean   :0.937
##  3rd Qu.:0.966   3rd Qu.:0.265   3rd Qu.:0   3rd Qu.:0.679   3rd Qu.:0.958
##  Max.   :1.000   Max.   :0.556   Max.   :0   Max.   :1.000   Max.   :1.000
##      MoD.p          age.avg     Non_White_prop
##  Min.   :0.810   Min.   :33.1   Min.   :0.250
##  1st Qu.:0.938   1st Qu.:35.7   1st Qu.:0.455
##  Median :0.948   Median :36.9   Median :0.563
##  Mean   :0.951   Mean   :37.2   Mean   :0.557
##  3rd Qu.:0.969   3rd Qu.:38.6   3rd Qu.:0.635
##  Max.   :1.000   Max.   :44.4   Max.   :0.939
```

Figure 15 We will now check our data for normality:



Because the plot is relatively linear, we can conclude this data is close enough to normality for our purpose.

Figure 16 Now let us look at the body camera proportions by state. In the below bar graph, TRUE signifies a police body camera that was on, while FALSE indicates the body camera was off:



Number of fatal shootings where the body camera was on:

```
##   body_camera   n
## 1        TRUE 947
```

Number of fatal shootings where the body camera was off:

```
##   body_camera    n
## 1       FALSE 5627
```

Figure 17 The below graph illustrates the number of victims shot and killed by race when a body camera was off:

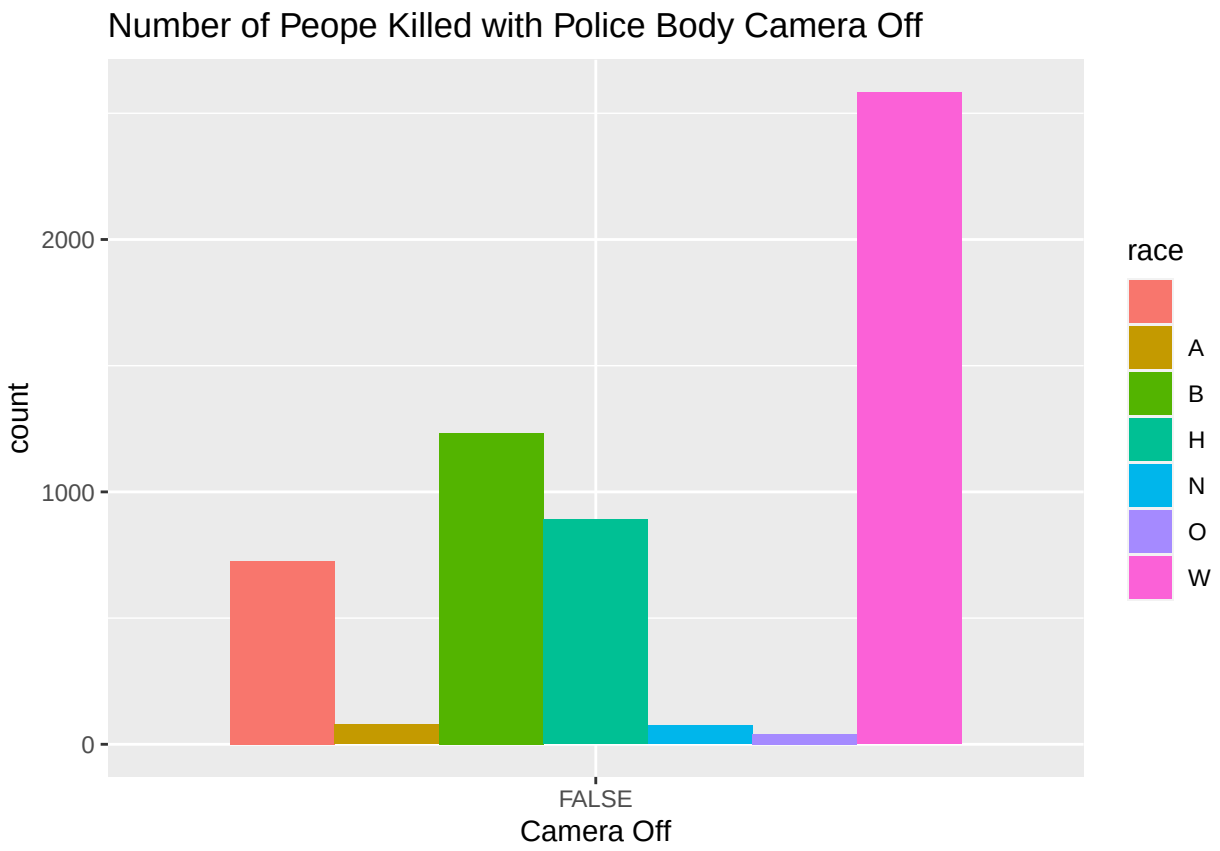


Figure 18 The below graph illustrates the number of victims shot and killed by race when a body camera was on:

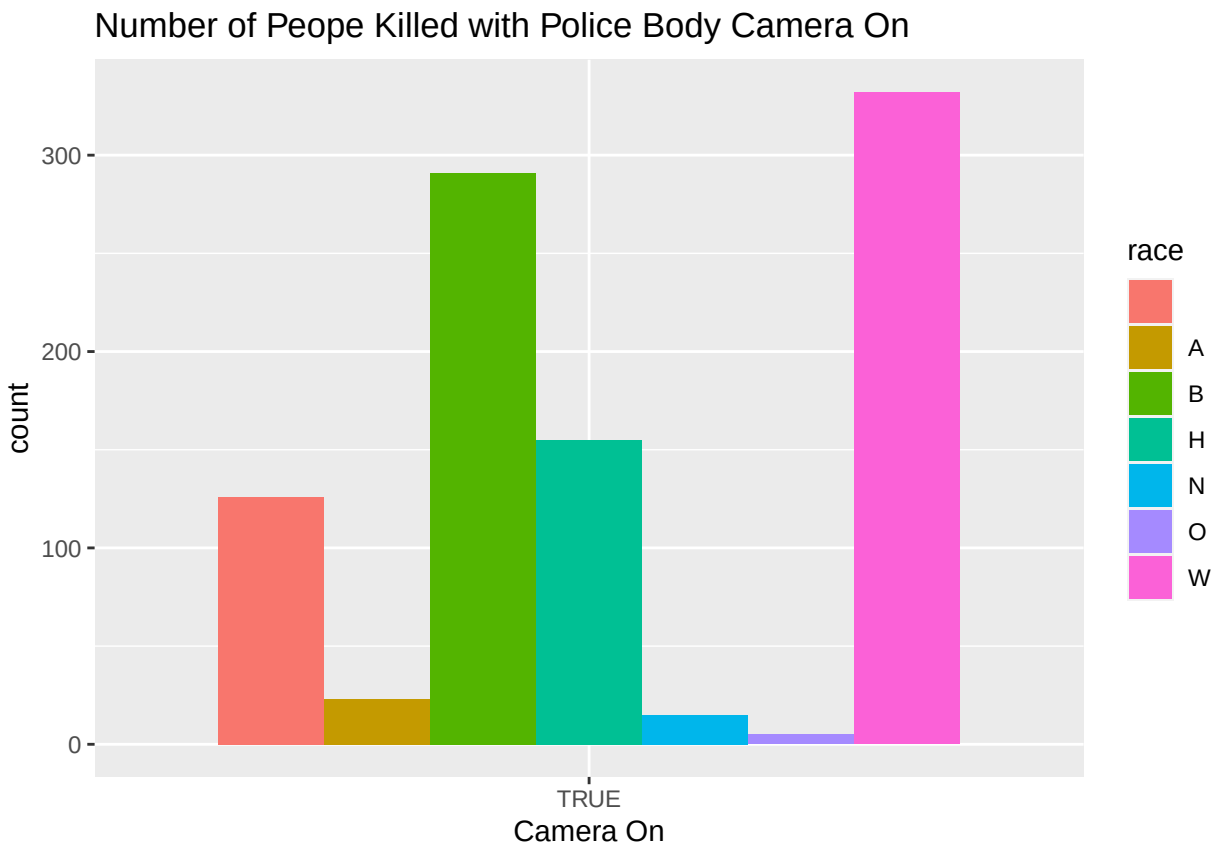
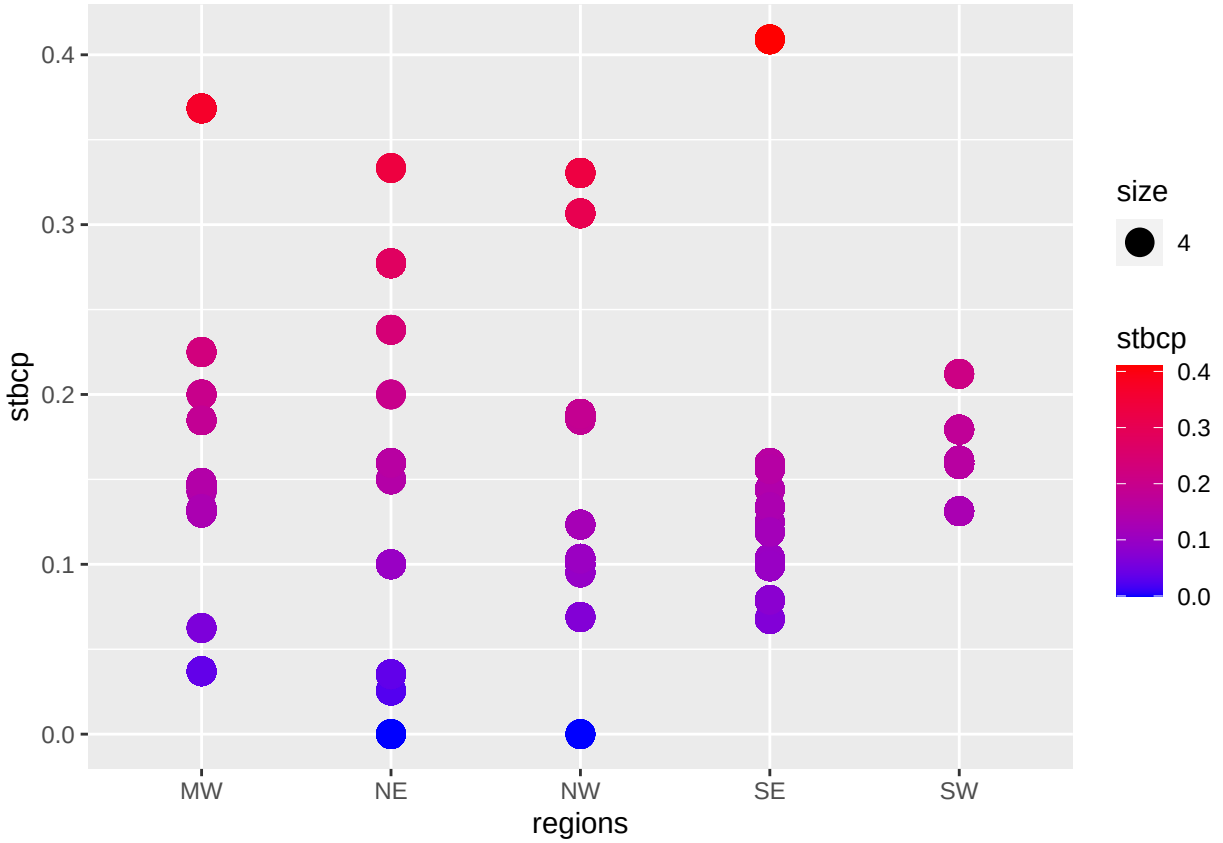


Figure 19 This scatter plot shows the proportion of fatal shootings when cameras were on by state (the variable stbcp). Each point on the graph depicts a state's proportion of shootings where the police body camera was turned on during the incident). We can see that there is very little variation in Southwest, and many differences among states in the Midwest.



Finally, let us check out the mean body camera on proportion for all states:

```
## [1] 0.144
```

And the stbcp median body camera on proportion for all states:

```
## [1] 0.133
```

We will now perform a chi-square test to see if there is a significant difference between the proportions of each state.

H_0: There is no significant differences between US States in the proportion of body cameras being turned on during police shootings

H_A: There is a significant difference between US State in the proportion of body cameras being turned on during police shootings

Significance Level: $\alpha = 0.05$"

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.101   0.133   0.144   0.183   0.409
```

```
##
##  Pearson's Chi-squared test
##
## data:  contable
## X-squared = 3e+05, df = 2300, p-value <2e-16
```

With a p-value of 2e-16, we easily pass our significance level of alpha=0.05 and have shown that there exists significant differences between different states' proportions of body camera usage during fatal police shootings.

This exploratory data analysis has shows that there is significant difference in the level body camera usage in police shootings between states and regions in the United States. We intend to delve into the reasons why there are differences and research what factors may explain these differences between states. This will require understanding state laws and policies regarding the use of police body cameras. We must also understand the police force consequences for turning off body cameras during police activity in different states.

Studying the use of body cameras in police work is an important topic of study for data-driven policy research in the United States. We hope to be able to apply this correlation between the U.S. state of observation and whether the body camera was on or off during the shooting to state policy on body cameras during police work.