Midterm EDA: Group 7

2022-10-31

Introduction

From 2015- 2022, in response to a deep lack of reporting within government sources, The Washington Post compiled a database of every fatal police shooting in the United States.We are interested in exploring this data, specifically as it shows the differences between US States.

Setting the Data Up

First we call our packages: dplyr and ggplot2 as well as reading our data:

```
## Loading required package: ggplot2
```

```
##
## 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 dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks plotly::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Then we remove the null values from our dataset

```
## 'data.frame':    6288 obs. of  17 variables:
##  $ id                     : int  3 4 5 8 9 11 13 15 16 17 ...
##  $ name                   : chr  "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffma
##  $ date                   : chr  "10/4/2022" "10/4/2022" "10/3/2022" "10/2/2022" ...
##  $ manner_of_death        : chr  "shot" "shot" "shot and Tasered" "shot" ...
##  $ armed                  : chr  "gun" "gun" "unarmed" "toy weapon" ...
##  $ age                    : int  53 47 23 32 39 18 22 35 34 47 ...
##  $ gender                 : chr  "M" "M" "M" "M" ...
##  $ race                   : chr  "A" "W" "H" "W" ...
##  $ city                   : chr  "Shelton" "Aloha" "Wichita" "San Francisco" ...
##  $ state                  : chr  "WA" "OR" "KS" "CA" ...
##  $ signs_of_mental_illness: logi  TRUE FALSE FALSE TRUE FALSE FALSE ...
##  $ threat_level           : chr  "attack" "attack" "other" "attack" ...
##  $ flee                   : chr  "Not fleeing" "Not fleeing" "Not fleeing" "Not fleeing" ...
##  $ body_camera            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude              : num  -123.1 -122.9 -97.3 -122.4 -104.7 ...
##  $ latitude               : num  47.2 45.5 37.7 37.8 40.4 ...
##  $ is_geocoding_exact     : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...

## [1] 17

## [1] 6288

##    Length     Class      Mode
##      6288 character character

## [1] "character"
```

After Accounting for Null Values: The dataset we are working with has 6574 observations. There is a sample row of the data as well

```
## [1] "Number of observations:"

## [1] 6288

##   id       name       date manner_of_death armed age gender race    city state
## 1  3 Tim Elliot 2022-10-04            shot   gun  53      M    A Shelton    WA
##   signs_of_mental_illness threat_level        flee body_camera longitude
## 1                    TRUE       attack Not fleeing       FALSE      -123
##   latitude is_geocoding_exact month year
## 1     47.2               TRUE    10 2022
```

Basic Stats

He are some basic stats:

Structure:

```
## 'data.frame':    6288 obs. of  19 variables:
##  $ id                     : int  3 4 5 8 9 11 13 15 16 17 ...
##  $ name                   : chr  "Tim Elliot" "Lewis Lee Lembke" "John Paul Quintero" "Matthew Hoffma
```

```
##  $ date                    : Date, format: "2022-10-04" "2022-10-04" ...
##  $ manner_of_death         : chr  "shot" "shot" "shot and Tasered" "shot" ...
##  $ armed                   : chr  "gun" "gun" "unarmed" "toy weapon" ...
##  $ age                     : int  53 47 23 32 39 18 22 35 34 47 ...
##  $ gender                  : chr  "M" "M" "M" "M" ...
##  $ race                    : chr  "A" "W" "H" "W" ...
##  $ city                    : chr  "Shelton" "Aloha" "Wichita" "San Francisco" ...
##  $ state                   : chr  "WA" "OR" "KS" "CA" ...
##  $ signs_of_mental_illness: logi  TRUE FALSE FALSE TRUE FALSE FALSE ...
##  $ threat_level            : chr  "attack" "attack" "other" "attack" ...
##  $ flee                    : chr  "Not fleeing" "Not fleeing" "Not fleeing" "Not fleeing" ...
##  $ body_camera             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ longitude               : num  -123.1 -122.9 -97.3 -122.4 -104.7 ...
##  $ latitude                : num  47.2 45.5 37.7 37.8 40.4 ...
##  $ is_geocoding_exact      : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ month                   : chr  "10" "10" "10" "10" ...
##  $ year                    : chr  "2022" "2022" "2022" "2022" ...
```

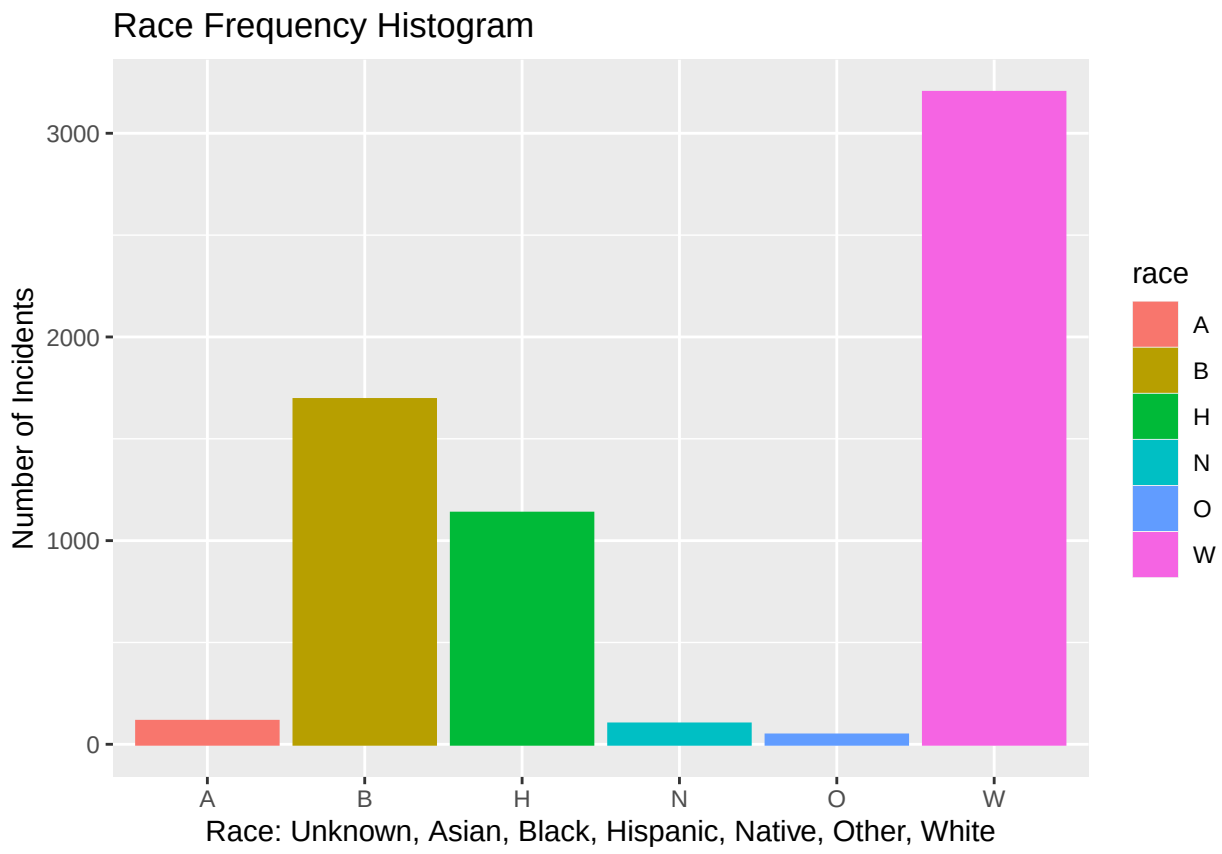
Means and Median for Numeric Variables (Age):

```
## [1] 36.7
```

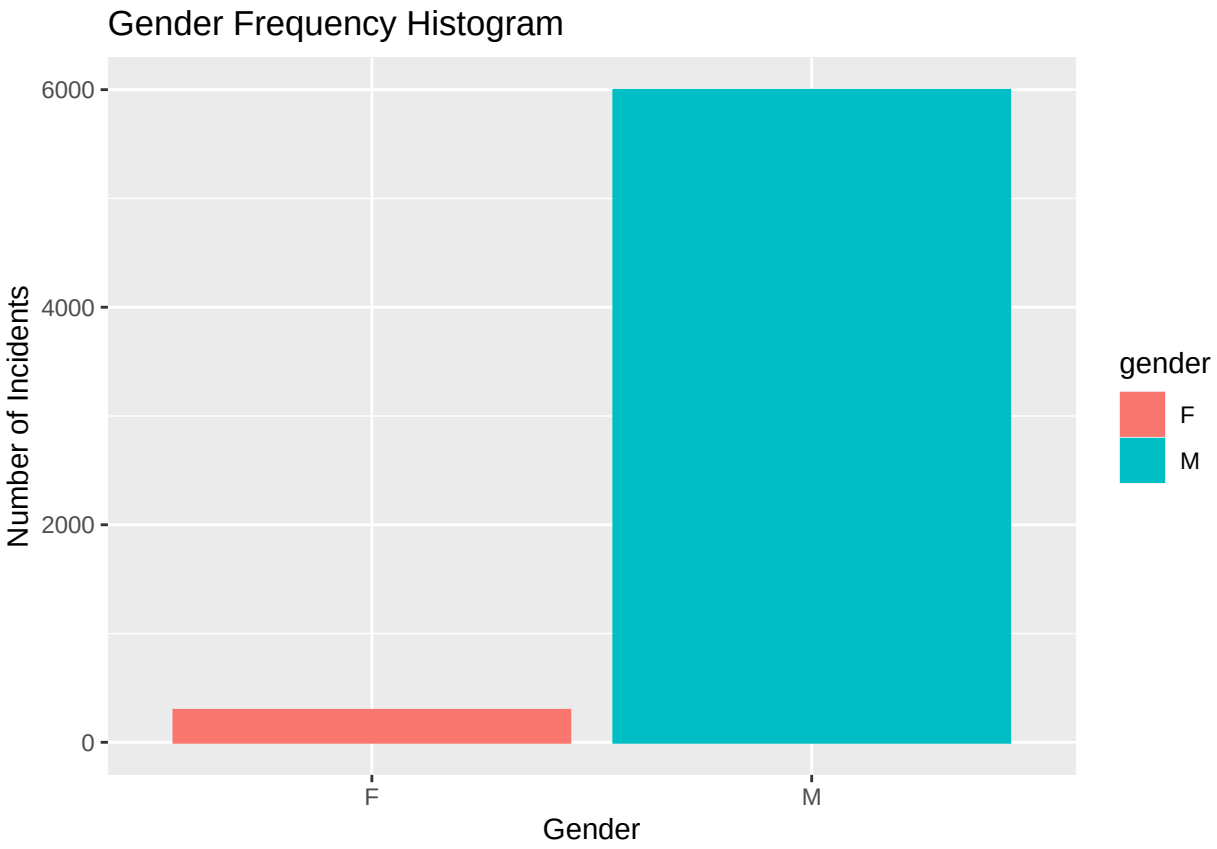
```
## [1] 34
```

#Frequency Graphs for Categorical Variables:

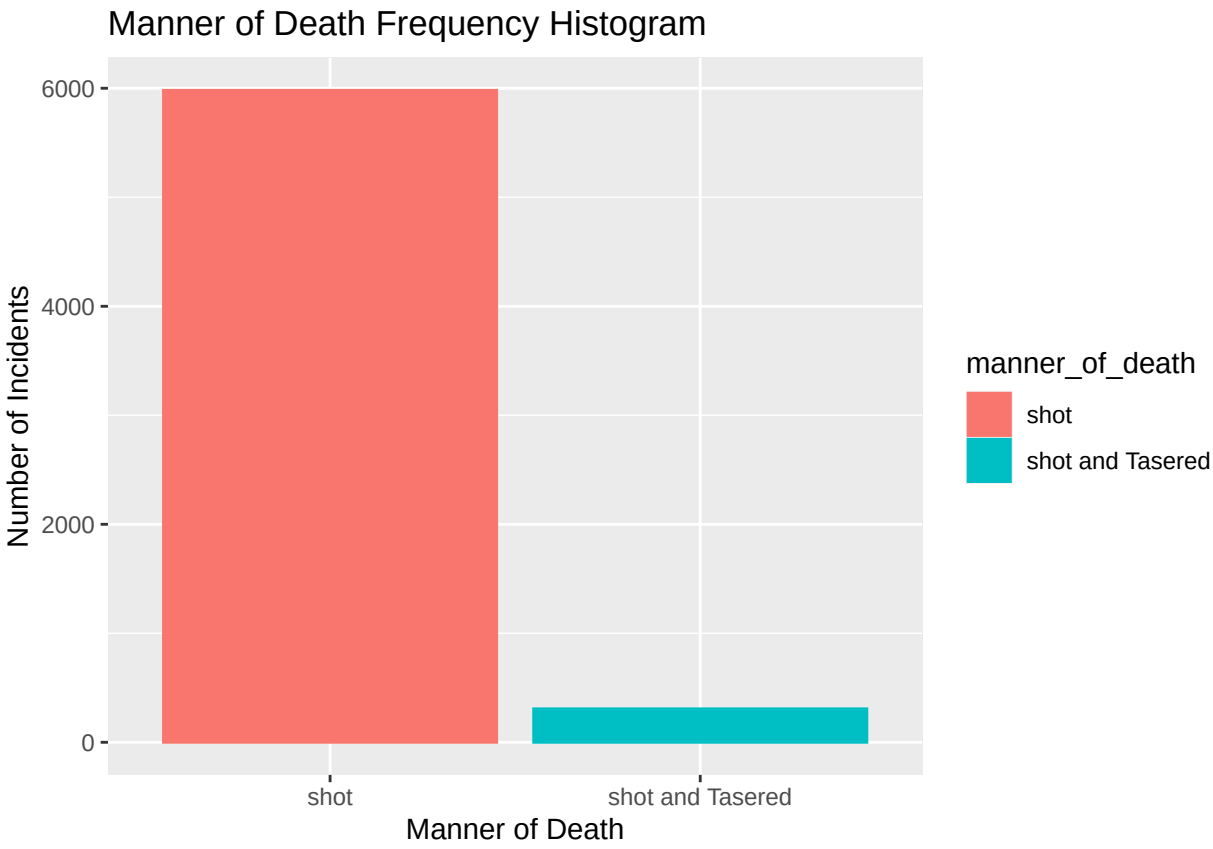
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

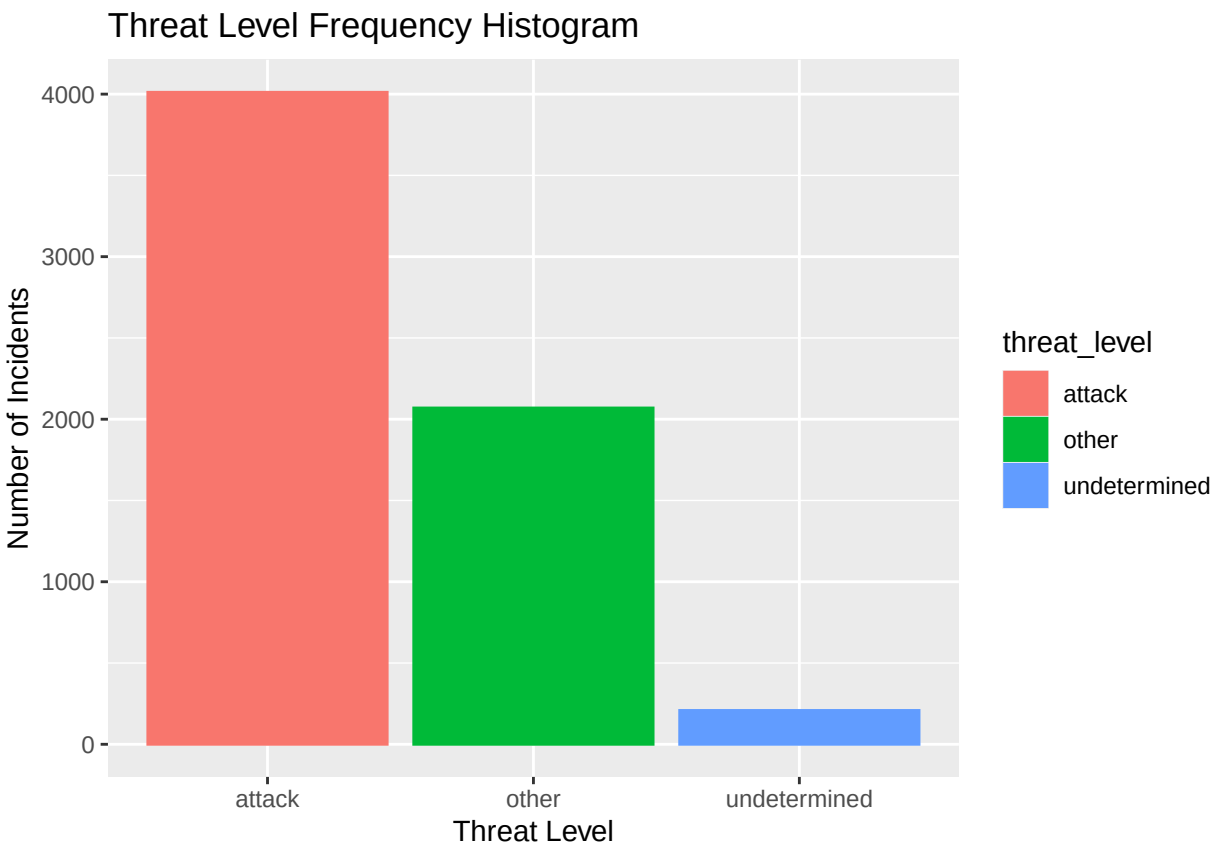



```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



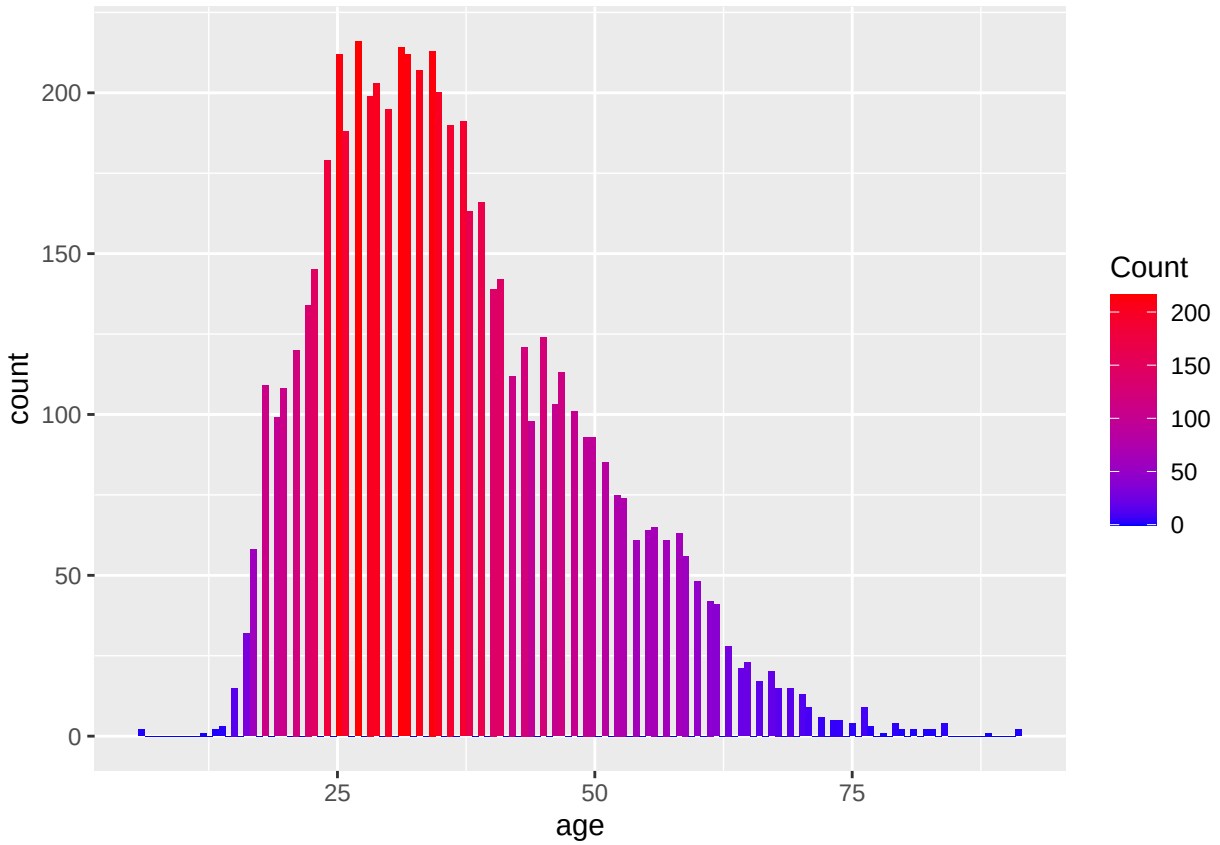

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```


Manner of Death Frequency Histogram
6000
4000
2000
0
Number of Incidents
shot
shot and Tasered
Manner of Death
manner_of_death
shot
shot and Tasered



##AGE Distribution

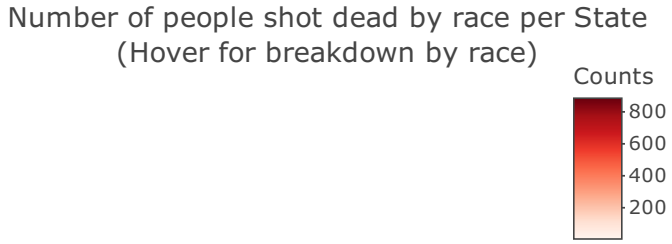
```
## Warning: Removed 125 rows containing non-finite values (stat_bin).
```



##Geospatial Analysis Interesting Finding 1 : California has the highest police shootings, and highest suspects shot in California are Hispanic and not White/Black. We looked at the total deaths in each state by race and following are some of the insights:

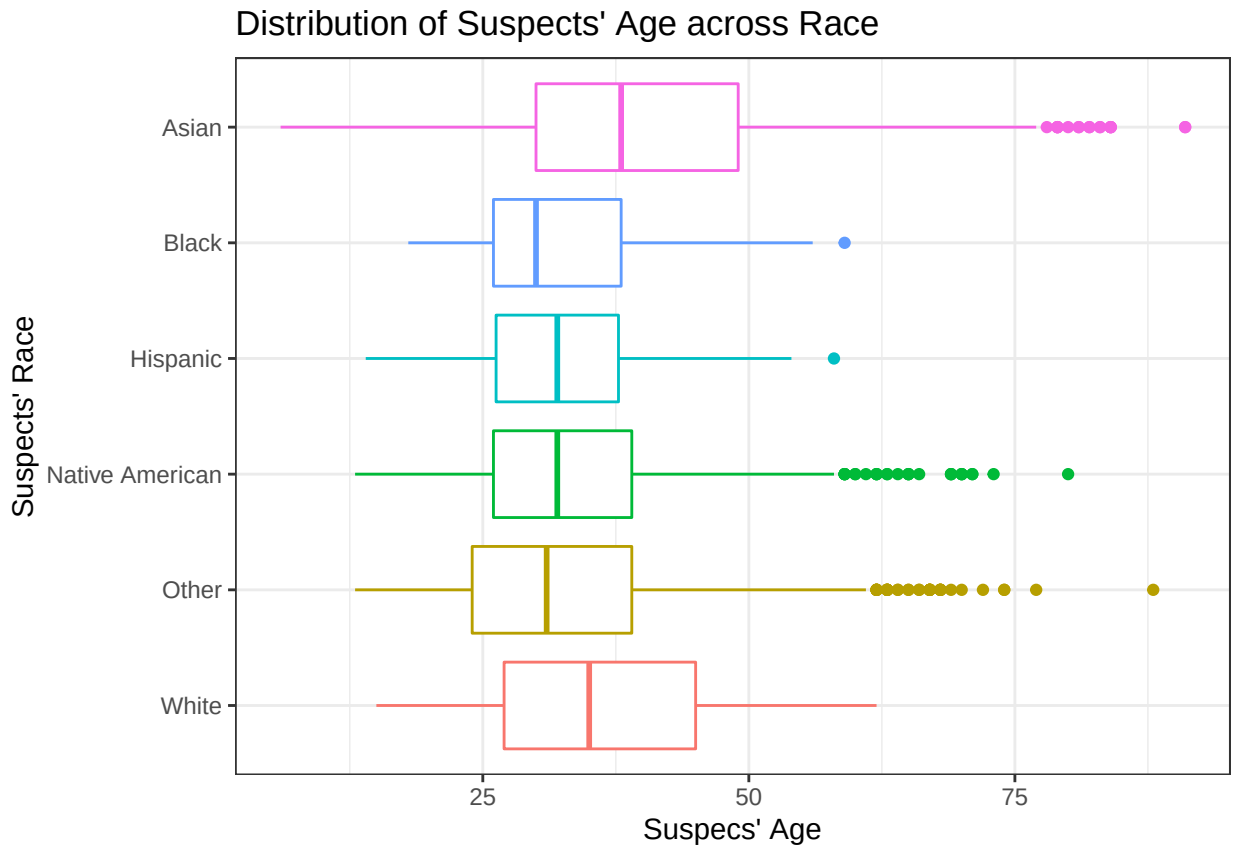
1)We see that police has shot the most people in California - a total of 885, followed by Texas with a total of 553 and then Florida with 427 deaths. 2)These results are consistent with the relative population of these states. Highest being California, then Texas and Florida . 3)We also observe that the highest number of deaths is for Hispanic in California, whereas in Texas and Florida there are more deaths amongst White.

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



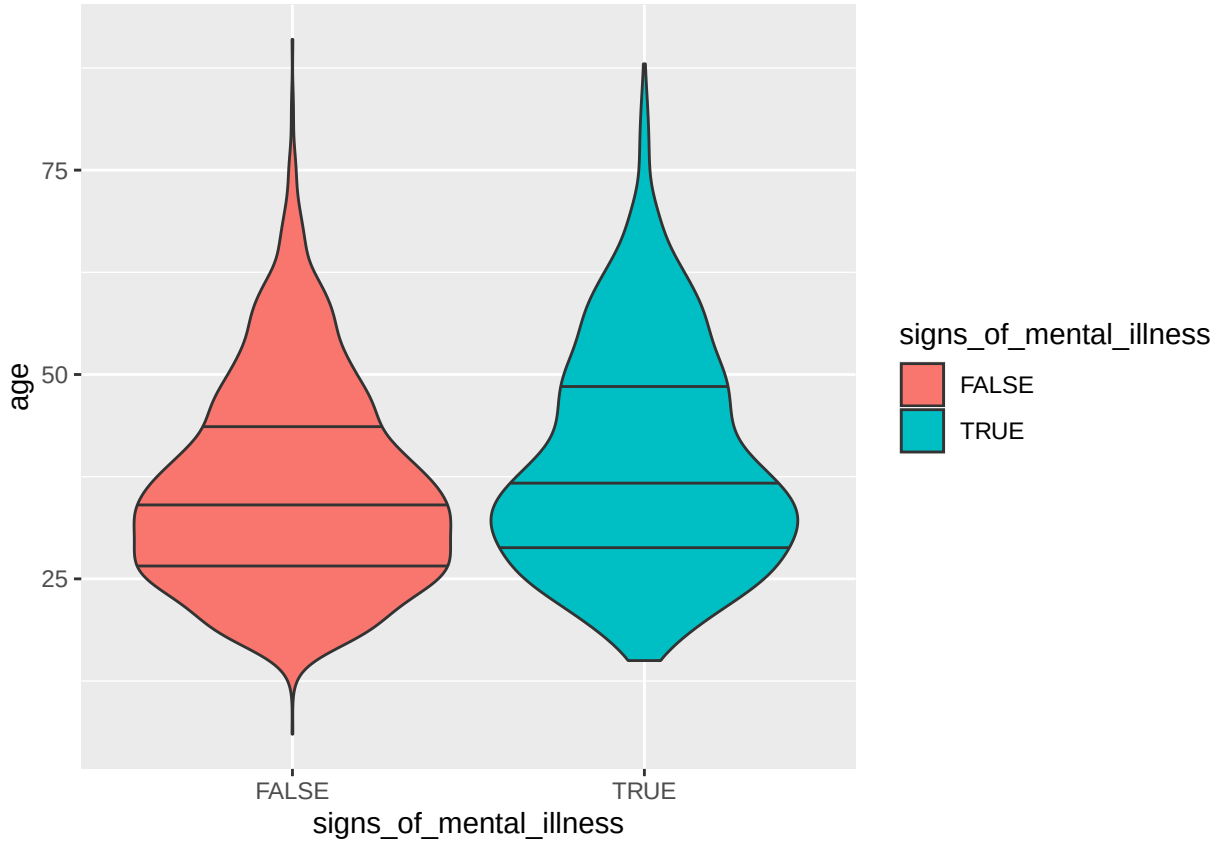
##Race/Age/Gender Analysis Interesting Finding 2 - Black people shot were relatively younger compared to other race. We are looking at the age of the suspect shot vs their race. The observations are as follows:

1)We see from the boxplot below, that the median age for Black that have been shot is 29 years. 2)White have relatively higher median age of 35 years whereas Asian have the highest median age of around 38 years. 3)signs of mental illness appear more frequently within 30s while the distribution of ages above 50 are more larger for people showing signs of mental illness. ##



###age against signs of mentall illness

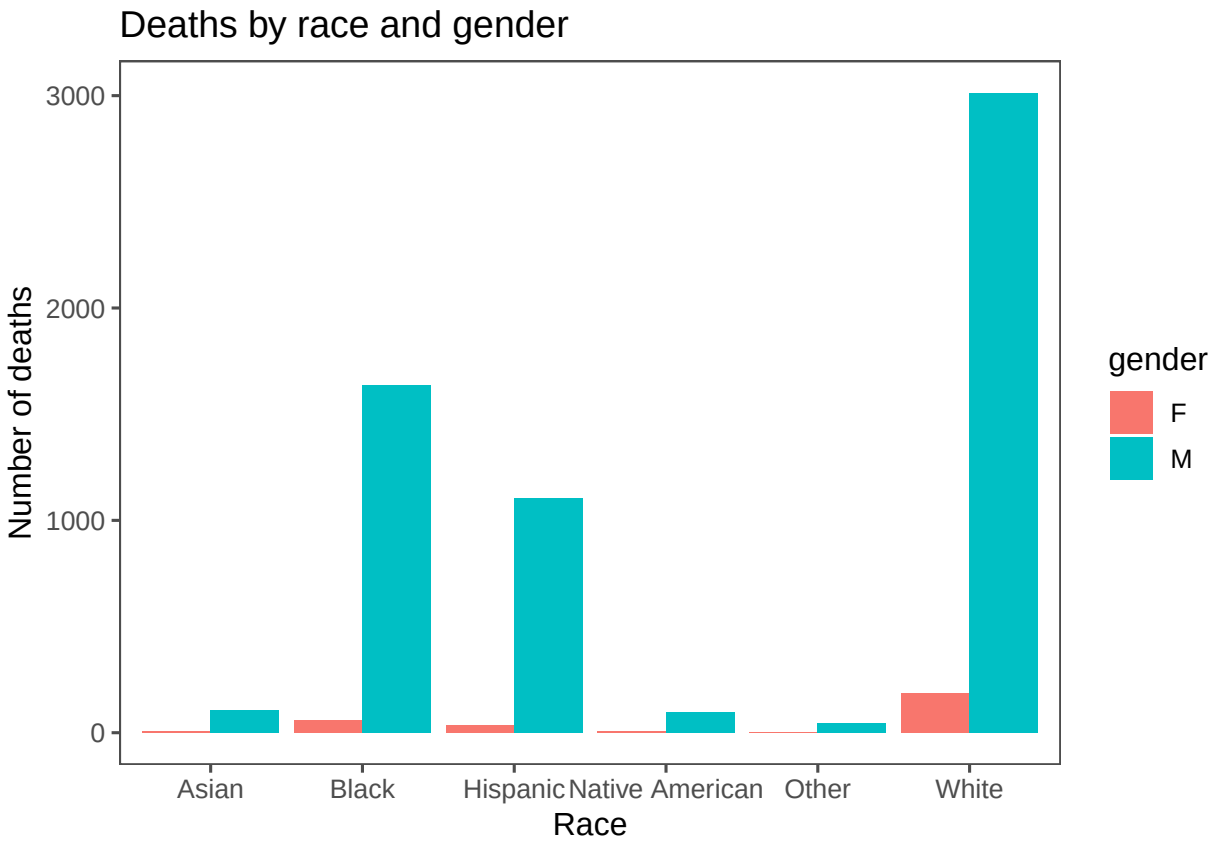
```
## Warning: Removed 125 rows containing non-finite values (stat_ydensity).
```



##Interesting Finding 3 - Hardly any female death shootings has been observed We looked at the deaths by race and gender and following are some of the insights:

1)Maximum number of suspects shot were males and there were very few females. 2)Maximum number of suspects shot were White, however this does not necessarily mean that higher proportion of white popultion is shot. These are absolute numbers and they are high as white have a significantly large population compared to other race.

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



Suspect's Condition Interesting Finding 4 - Higher % of unarmed Black suspects were shot than any other race We looked at the distribution of deaths by Race and top 5 armed categories. Following are some key observations:

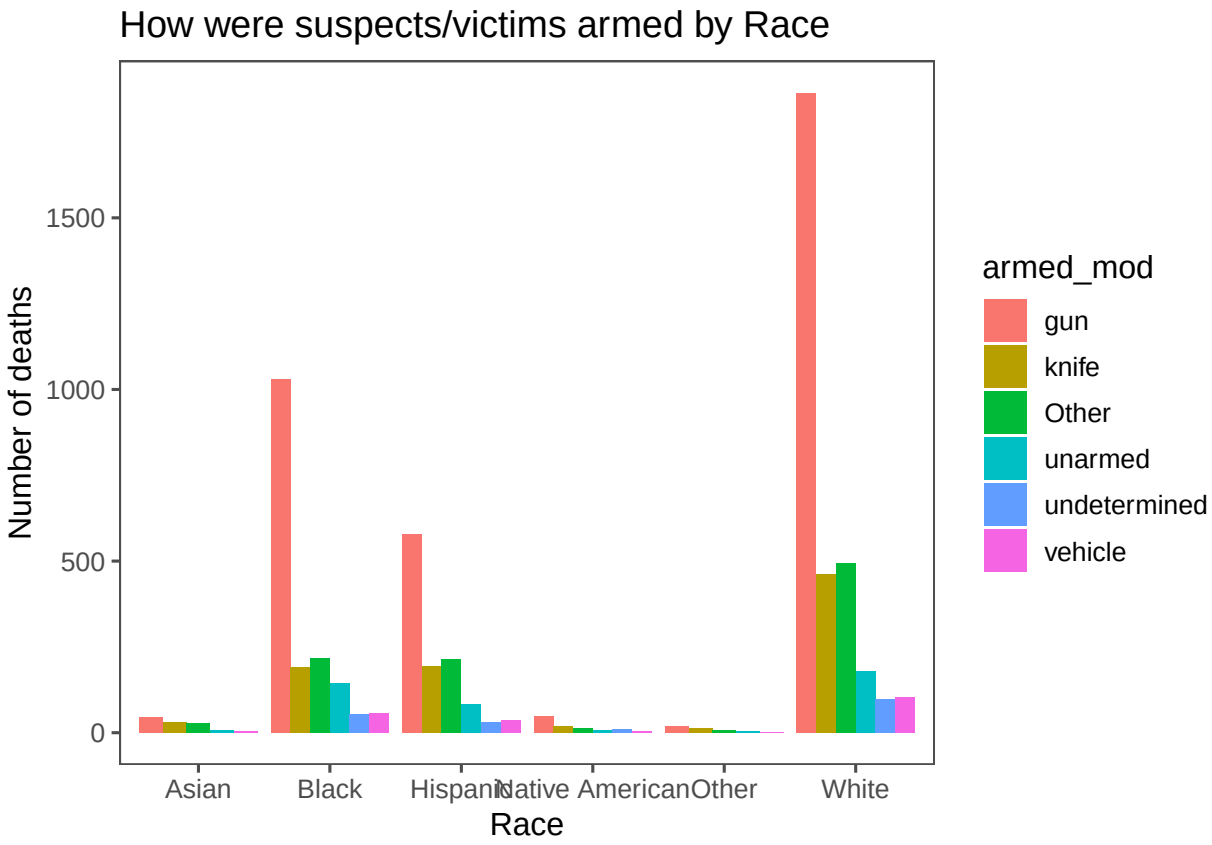
1)Around ~9% of the Black suspects were unarmed whereas only ~6% of the White suspects were unarmed, Guns are the most popular weapon across all the races except for Asians (Asian suspects have a higher proportion of Knives)

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

```
## [1] "% distribution of deaths by Armed Category in each Race"
```

```
##   race  gun knife Other unarmed undetermined vehicle
## 1    A 39.8  26.6  23.9    7.08         0.00    2.65
## 2    B 60.9  11.2  12.8    8.56         3.13    3.43
## 3    H 50.9  17.2  18.8    7.40         2.64    3.08
## 4    N 48.0  18.0  14.0    7.00        10.00    3.00
## 5    O 41.3  28.3  15.2   10.87         0.00    4.35
## 6    W 58.2  14.5  15.4    5.59         3.03    3.25
```

##Graph : For better visualization, plotting the above results from the table in a stacked bar chart below



##Interesting Finding 5 - Higher proportion of Asians were not fleeing but still shot We looked at the distribution of deaths by suspects' race and whether they were trying to flee or not. Following are some of the interesting observations:

1)Only 53% of the Black suspects shot were not fleeing whereas 71% of the Asian suspects who were shot were not trying to flee 2)Car seems to be the most popular method of fleeing among White suspects whereas for Black suspects (16%), most popular method of fleeing was by foot (19%)

```
## `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`.
```

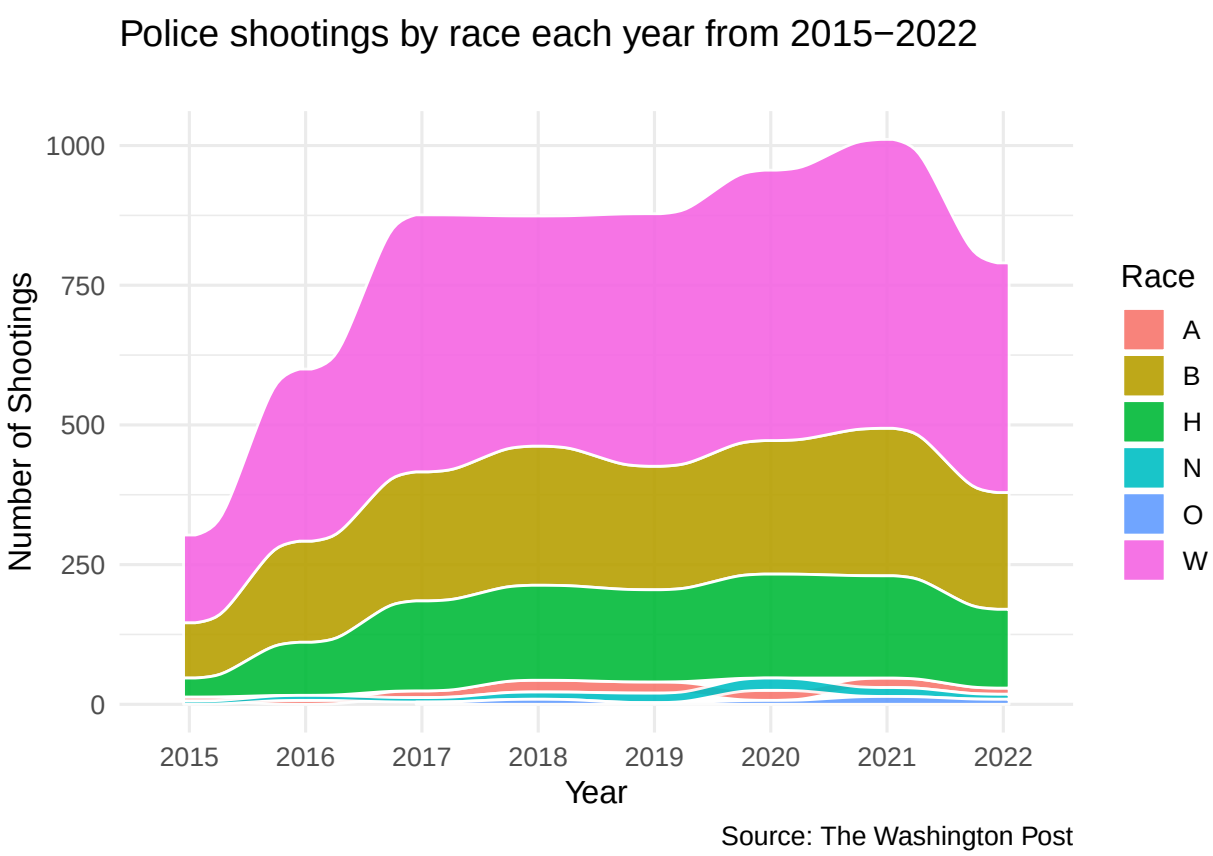
```
## [1] "% distribution of deaths by suspects' status (Fleeing or not fleeing) by Race"
```

```
##   race    V1  Car  Foot Not fleeing Other
## 1    A  7.08 10.6  9.73        71.7  0.88
## 2    B  7.56 15.6 19.43        53.3  4.08
## 3    H  7.49 16.4 14.19        56.8  5.11
## 4    N 14.00 11.0 18.00        53.0  4.00
## 5    O  2.17 19.6 10.87        63.0  4.35
## 6    W  8.50 16.5  9.62        62.0  3.37
```

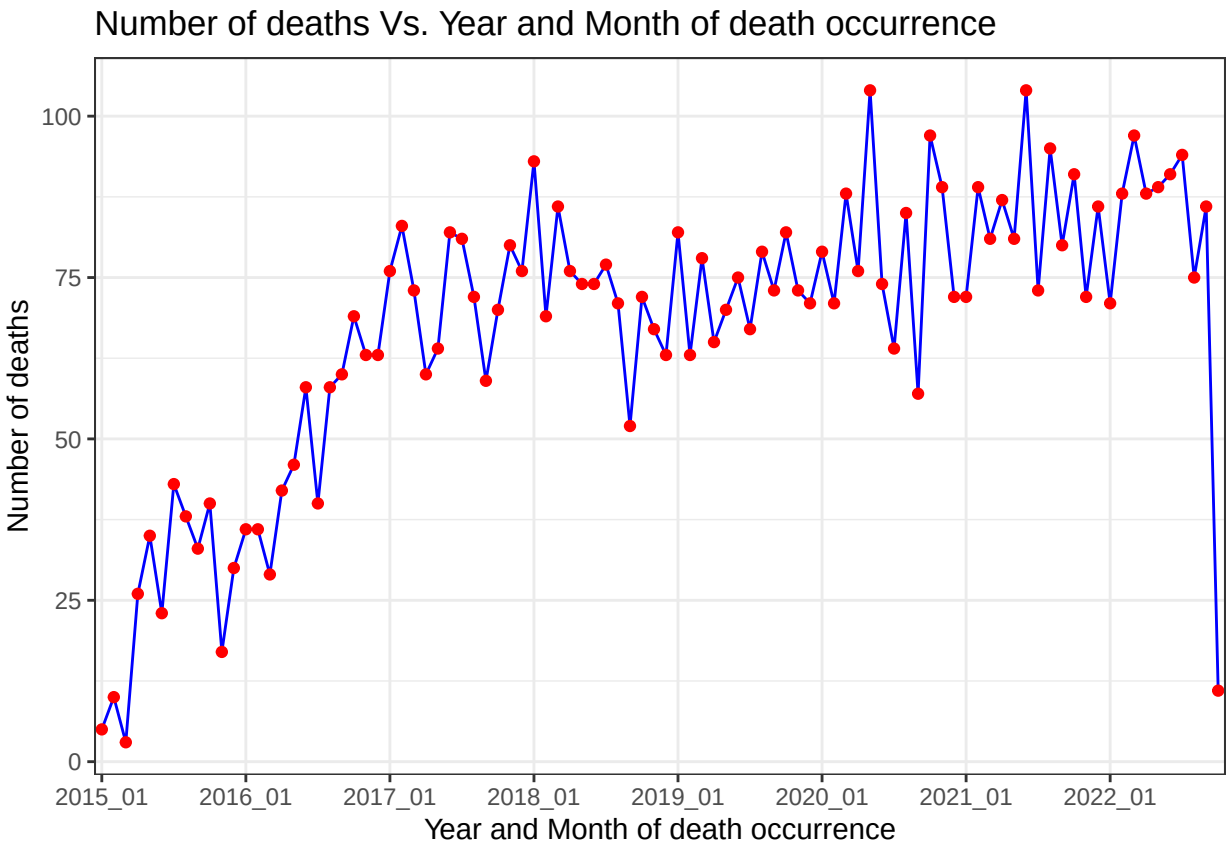
```
## [1] "data.frame"
```

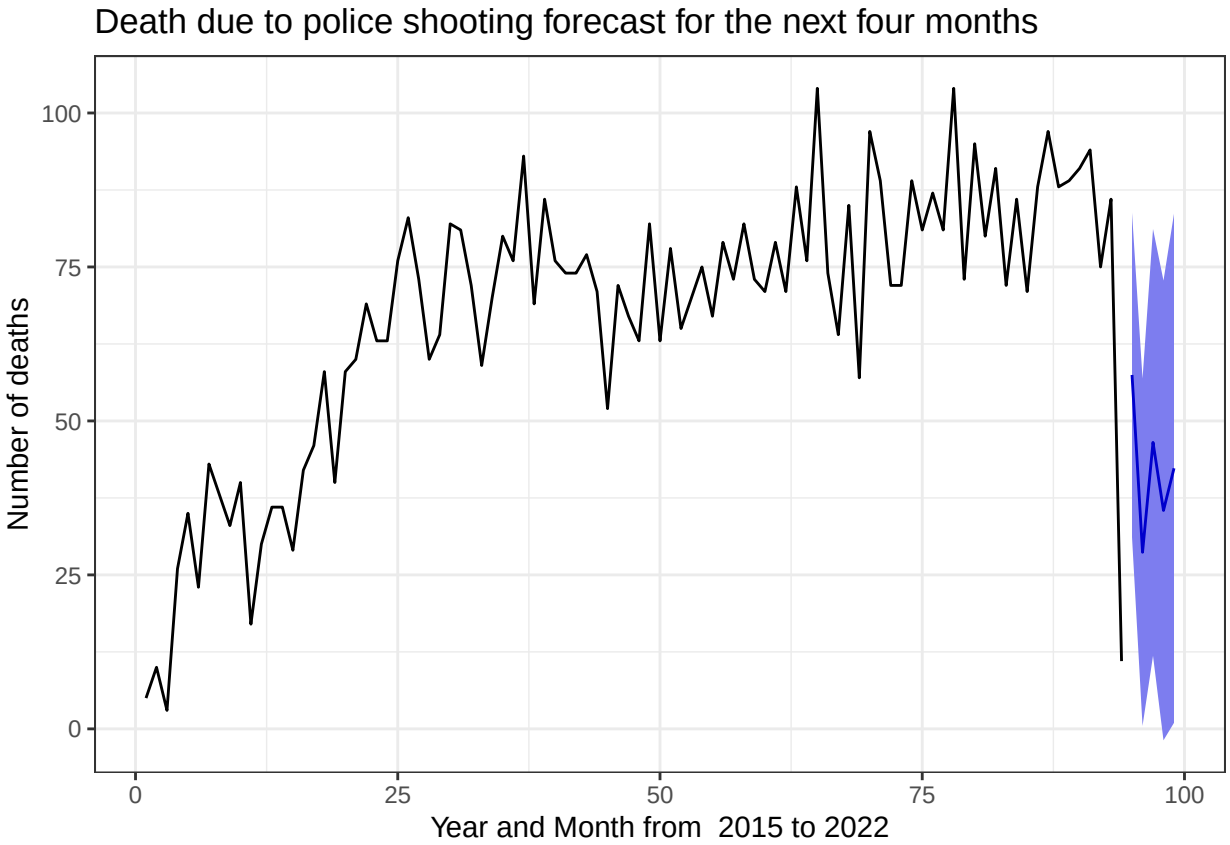
```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

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



##Time Series Analysis Interesting Pattern 6 - Suprisingly there is seasonaility across year or months in police shootings We looked into the monthly trend for 8 years and used ARIMA to forecast the crime for next four months. Since, there is seasonality into the police shootings, even the forecast predicts average shootings for the next four months with a wide confidence interval.


Number of deaths Vs. Year and Month of death occurrence
Number of deaths
100
75
50
25
0
2015_01
2016_01
2017_01
2018_01
2019_01
2020_01
2021_01
2022_01
Year and Month of death occurrence



###Reshaping the Data for State Comparison

We are particularly interested in using this data to view differences between US States and Regions.

The Regions:

NW (North West): CA, WA, OR, NV, ID, UT, MT, CO, WY, AK

SW (South West): NM, AZ, TX, OK, HI

MW (Mid West): IL, WI, IN, MI, MN, MO, IA, KS, ND, SD, NE ,OH

SE(South East): GA, AL, MS, LA, TN, NC, SC, FL, AR, WV, DC, VA

NE (North East): NY, RI, MD, VT, PA, ME, NH, NJ, CT, MA

```
## [1] "Incidents in NW:"
```

```
## [1] 1677
```

```
## [1] "Incidents in SW:"
```

```
## [1] 1162
```

```
## [1] "Incidents in MW:"
```

```
## [1] 1058
```

```
## [1] "Incidents in SE:"
```

```
## [1] 1868
```

```
## [1] "Incidents in NE:"
```

```
## [1] 523
```

We have created two sub datasets by grouping our data by state and by reigion (for graphical purposes). Here is the structure of both:

```
## [1] "By_State:"
```

```
##     state              month               year           regions
##  Length:6288        Length:6288        Length:6288        MW:1058
##  Class :character   Class :character   Class :character   NE: 523
##  Mode  :character   Mode  :character   Mode  :character   NW:1677
##                                                           SE:1868
##                                                           SW:1162
##
##
##      stbcp            gen.p            smi.p            flee.p      att.p
##  Min.   :0.000   Min.   :0.800   Min.   :0.000   Min.   :0   Min.   :0.375
##  1st Qu.:0.099   1st Qu.:0.940   1st Qu.:0.188   1st Qu.:0   1st Qu.:0.588
##  Median :0.134   Median :0.946   Median :0.224   Median :0   Median :0.643
##  Mean   :0.144   Mean   :0.953   Mean   :0.225   Mean   :0   Mean   :0.638
##  3rd Qu.:0.183   3rd Qu.:0.965   3rd Qu.:0.267   3rd Qu.:0   3rd Qu.:0.677
##  Max.   :0.388   Max.   :1.000   Max.   :0.600   Max.   :0   Max.   :1.000
##
##     armed.p          MoD.p          age.avg     Non_White_prop
##  Min.   :0.786   Min.   :0.810   Min.   :32     Min.   :0.000
##  1st Qu.:0.916   1st Qu.:0.936   1st Qu.:35     1st Qu.:0.371
##  Median :0.924   Median :0.948   Median :37     Median :0.501
##  Mean   :0.932   Mean   :0.951   Mean   :37     Mean   :0.491
##  3rd Qu.:0.952   3rd Qu.:0.971   3rd Qu.:38     3rd Qu.:0.589
##  Max.   :1.000   Max.   :1.000   Max.   :44     Max.   :0.909
##                                  NA's   :5597
```

```
## [1] "By Region:"
```

```
##     state              month               year              stbcp
##  Length:6288        Length:6288        Length:6288        Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:0.099
##  Mode  :character   Mode  :character   Mode  :character   Median :0.134
##                                                           Mean   :0.144
##                                                           3rd Qu.:0.183
##                                                           Max.   :0.388
##
##      gen.p            smi.p            flee.p      att.p          armed.p
##  Min.   :0.800   Min.   :0.000   Min.   :0   Min.   :0.375   Min.   :0.786
##  1st Qu.:0.940   1st Qu.:0.188   1st Qu.:0   1st Qu.:0.588   1st Qu.:0.916
##  Median :0.946   Median :0.224   Median :0   Median :0.643   Median :0.924
##  Mean   :0.953   Mean   :0.225   Mean   :0   Mean   :0.638   Mean   :0.932
##  3rd Qu.:0.965   3rd Qu.:0.267   3rd Qu.:0   3rd Qu.:0.677   3rd Qu.:0.952
```

```
##  Max.   :1.000   Max.   :0.600   Max.   :0   Max.   :1.000   Max.   :1.000
##
##      MoD.p          age.avg     Non_White_prop
##  Min.   :0.810   Min.   :32     Min.   :0.000
##  1st Qu.:0.936   1st Qu.:35     1st Qu.:0.371
##  Median :0.948   Median :37     Median :0.501
##  Mean   :0.951   Mean   :37     Mean   :0.491
##  3rd Qu.:0.971   3rd Qu.:38     3rd Qu.:0.589
##  Max.   :1.000   Max.   :44     Max.   :0.909
##                  NA's   :5597
```

As you can see, the groups are identical, besides their grouping.

SMART Question and Answer

Within our dataset of fp1 shootings from 2015 to 2020 in the United States, is there a significant difference between the states?

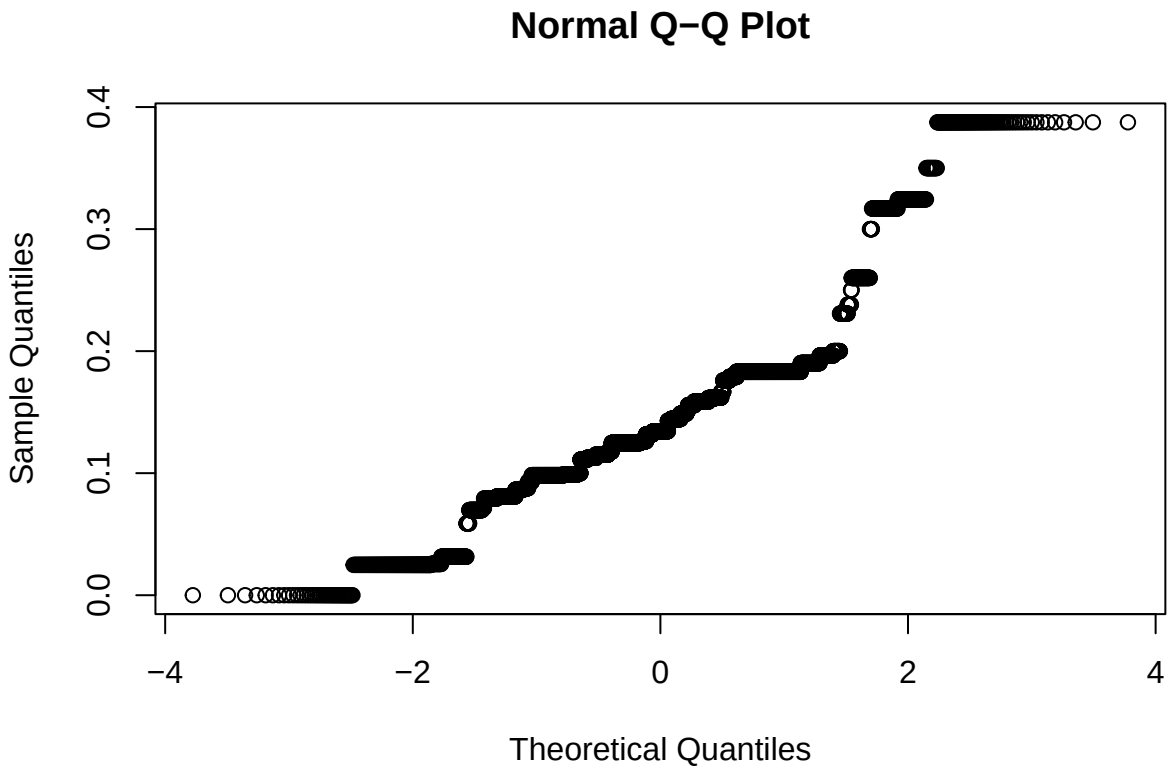
First let's take a look at our data after it has been grouped by state and reorganized into the following variables:

- state: State of Observation
- region: Region of Observation
- stbcp: State Body Camera On Proportion
- genp.p: proportion of male identified shooting victims by state
- smi.p: proportion of shooting victims by state with a documented sign of mental illness
- flee.p: proportion of shooting victims by state that we fleeing
- att.p: proportion of shooting victims by state that we attacking
- armed.p: proportion of shooting victims by state that were not unarmed
- MoD.p: proportion of shooting victims by state who where shot (rather than shot and tased)
- age.avg: average age by state
- Non_White_prop: Proportion of shooting vicitms by state that were not identified as white/caucasian

```
## # A tibble: 6 x 13
## # Groups:   state [6]
##   state month year  regions  stbcp gen.p smi.p flee.p att.p armed.p MoD.p
##   <chr> <chr> <chr> <fct>    <dbl> <dbl> <dbl>  <dbl> <dbl>   <dbl> <dbl>
## 1 WA    10    2022  NW      0.113  0.960 0.331      0 0.517   0.921 0.940
## 2 OR    10    2022  NW      0.0792 0.980 0.297      0 0.485   0.960 0.950
## 3 KS    10    2022  MW      0.143  0.921 0.206      0 0.714   0.937 0.937
## 4 CA    10    2022  NW      0.183  0.946 0.224      0 0.577   0.916 0.936
## 5 CO    10    2022  NW      0.115  0.963 0.143      0 0.618   0.949 0.982
## 6 OK    10    2022  SW      0.190  0.978 0.217      0 0.685   0.902 0.924
## # ... with 2 more variables: age.avg <dbl>, Non_White_prop <dbl>
```

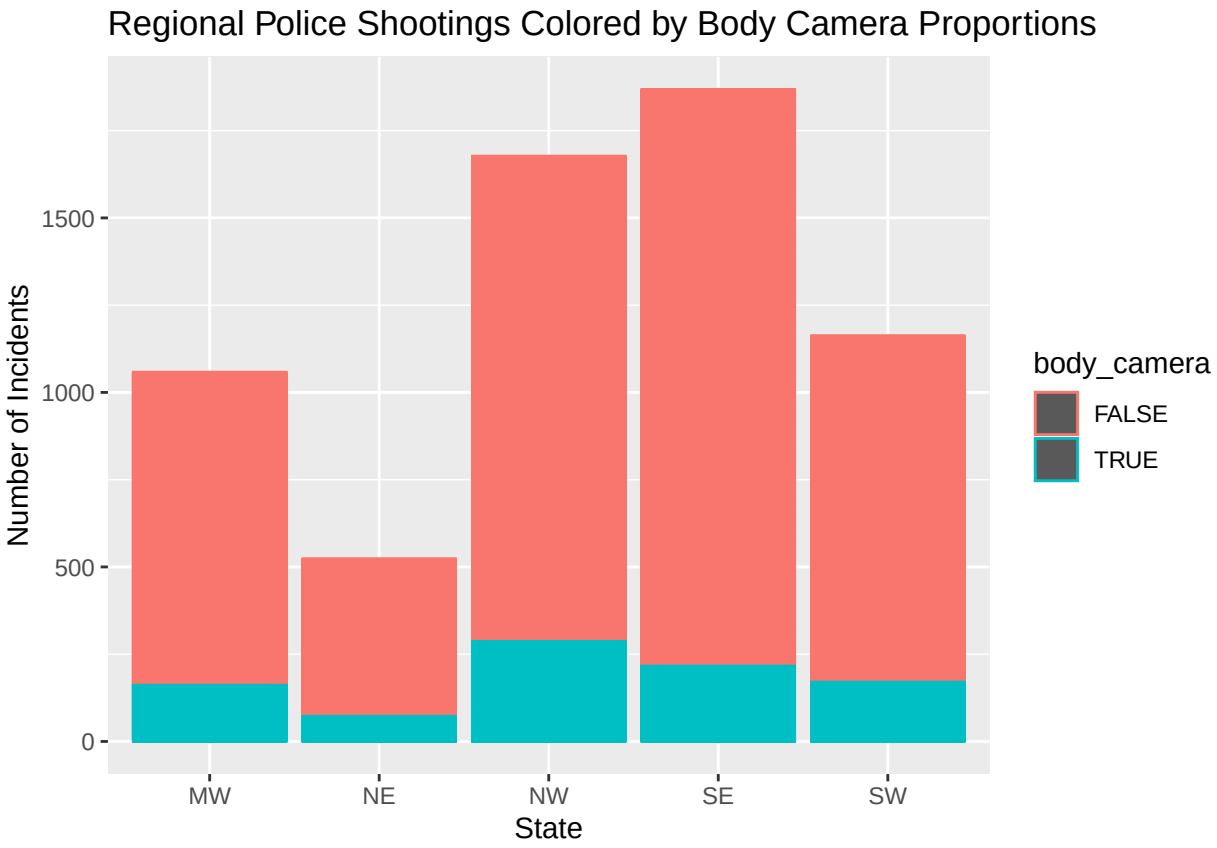
```
##     state              month               year           regions
##  Length:6288        Length:6288        Length:6288        MW:1058
##  Class :character   Class :character   Class :character   NE: 523
##  Mode  :character   Mode  :character   Mode  :character   NW:1677
##                                                           SE:1868
##                                                           SW:1162
##
##
##      stbcp            gen.p            smi.p            flee.p      att.p
##  Min.   :0.000   Min.   :0.800   Min.   :0.000   Min.   :0   Min.   :0.375
##  1st Qu.:0.099   1st Qu.:0.940   1st Qu.:0.188   1st Qu.:0   1st Qu.:0.588
##  Median :0.134   Median :0.946   Median :0.224   Median :0   Median :0.643
##  Mean   :0.144   Mean   :0.953   Mean   :0.225   Mean   :0   Mean   :0.638
##  3rd Qu.:0.183   3rd Qu.:0.965   3rd Qu.:0.267   3rd Qu.:0   3rd Qu.:0.677
##  Max.   :0.388   Max.   :1.000   Max.   :0.600   Max.   :0   Max.   :1.000
##
##     armed.p          MoD.p          age.avg     Non_White_prop
##  Min.   :0.786   Min.   :0.810   Min.   :32     Min.   :0.000
##  1st Qu.:0.916   1st Qu.:0.936   1st Qu.:35     1st Qu.:0.371
##  Median :0.924   Median :0.948   Median :37     Median :0.501
##  Mean   :0.932   Mean   :0.951   Mean   :37     Mean   :0.491
##  3rd Qu.:0.952   3rd Qu.:0.971   3rd Qu.:38     3rd Qu.:0.589
##  Max.   :1.000   Max.   :1.000   Max.   :44     Max.   :0.909
##                                  NA's   :5597
```

We now would like to check our data for normality:



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

Now let us look at the body camera proportions by state:

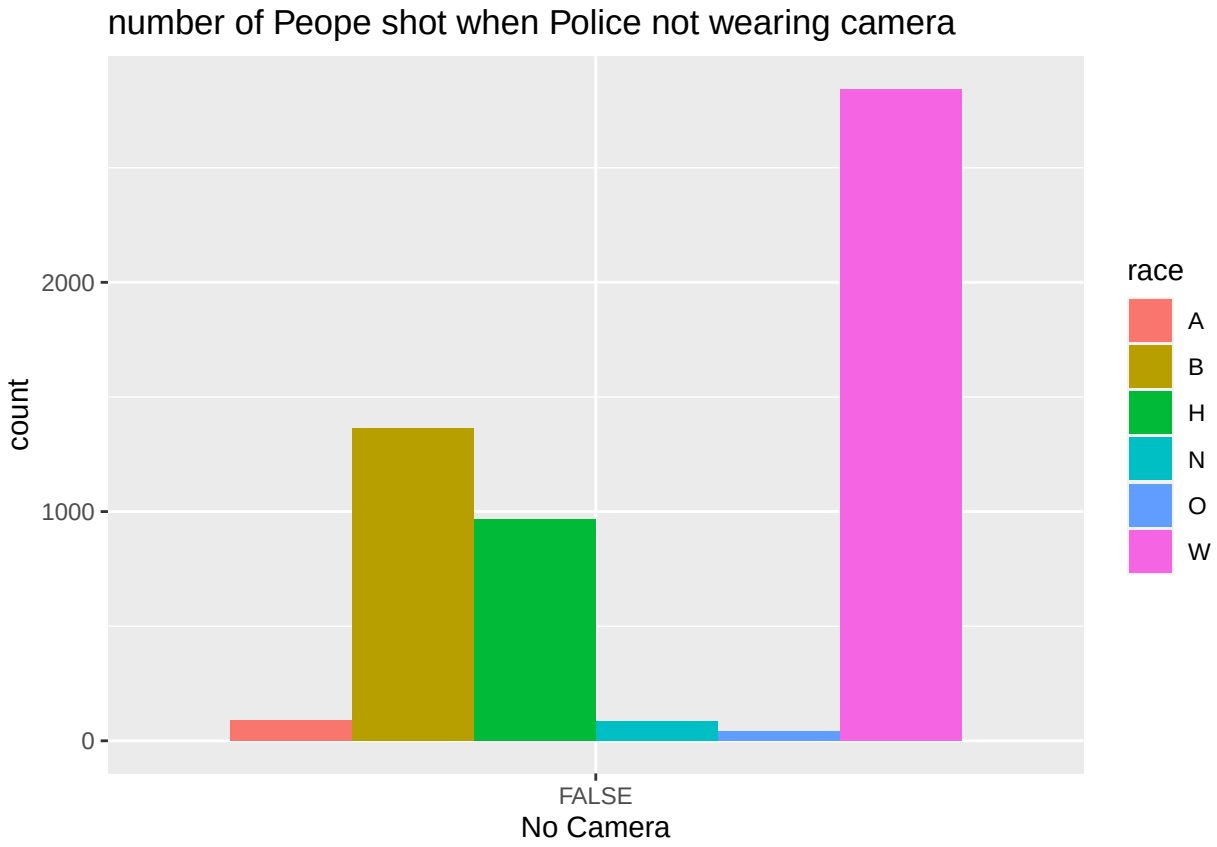


count of body camera = TRUE and count of body camera = FALSE

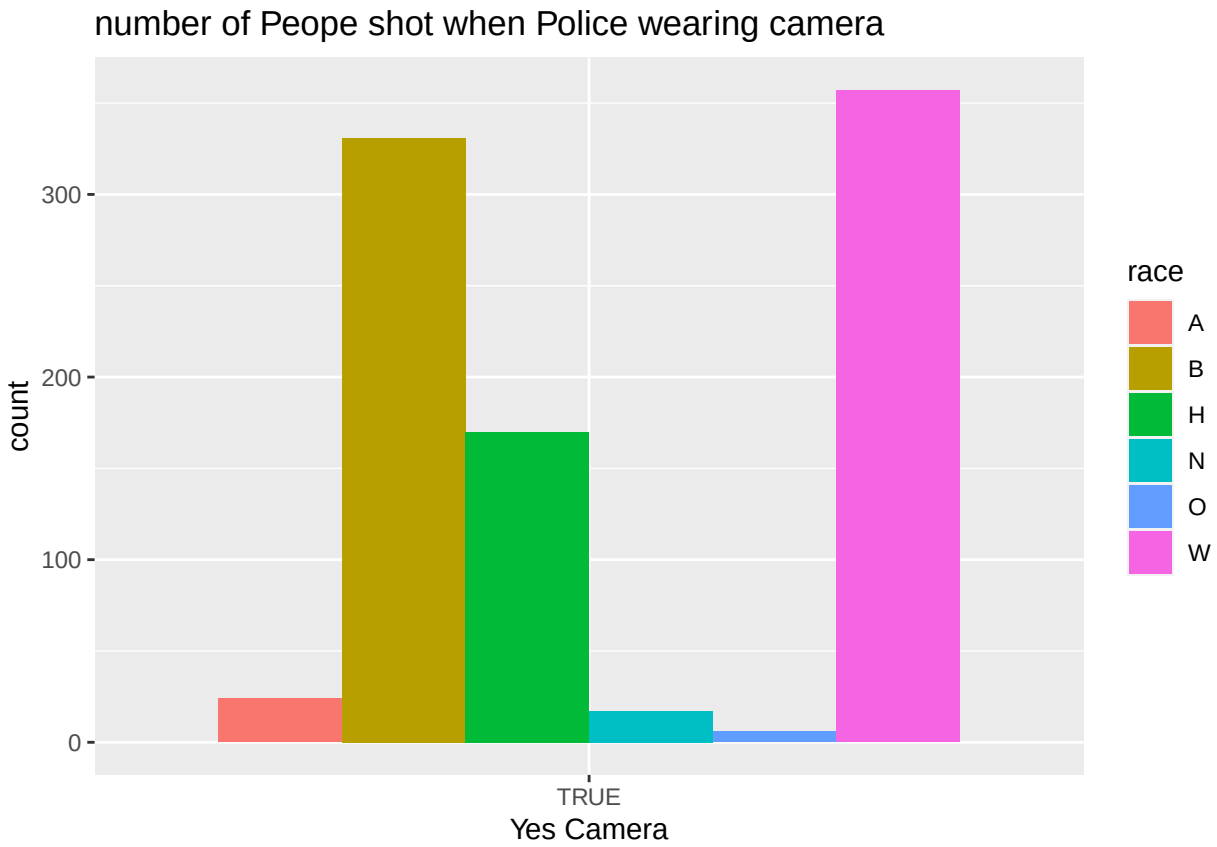
```
##   body_camera   n
## 1        TRUE 905
```

```
##   body_camera    n
## 1       FALSE 5383
```

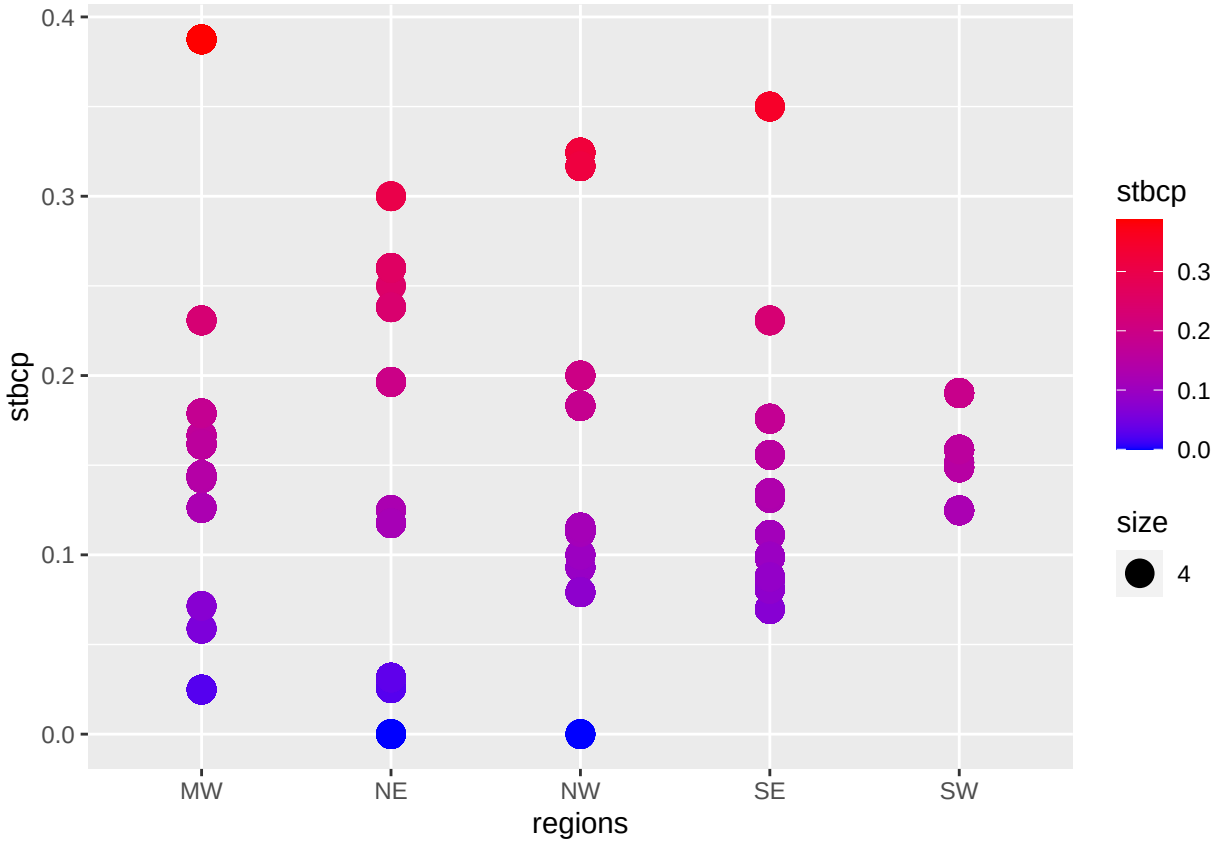
number of Peope shot when Police not wearing camera



##number of Peope shot when Police wearing camera



##Scatter Plot of Body Camera Proportion by Region



And finally, let us check out the mean body camera on proportion off all states:

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

And now let us do a chi-square test to see if there is a siginficant difference between the proportions of each state.

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

Alternative Hypothesis: 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   0.099   0.134   0.144   0.183   0.388
```

```
## Warning in chisq.test(contable): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  contable
## X-squared = 3e+05, df = 2400, p-value <2e-16
```

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

For Further Analysis: We intend to delve into why there are differences and research what factors may explain these differences between states.