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:

```
##
## 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
```

Then we remove the null values from our dataset

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] 6574

##   id       name      date manner_of_death armed age gender race    city state
## 1  3 Tim Elliot 10/4/2022            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
## 1     47.2               TRUE
```

Basic Stats

He are some basic stats:

Structure:

```
## 'data.frame':    6574 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 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1229] 128 770 810 820 933 941 966 991 1338 1353 ...
##   ..- attr(*, "names")= chr [1:1229] "128" "770" "810" "820" ...
```

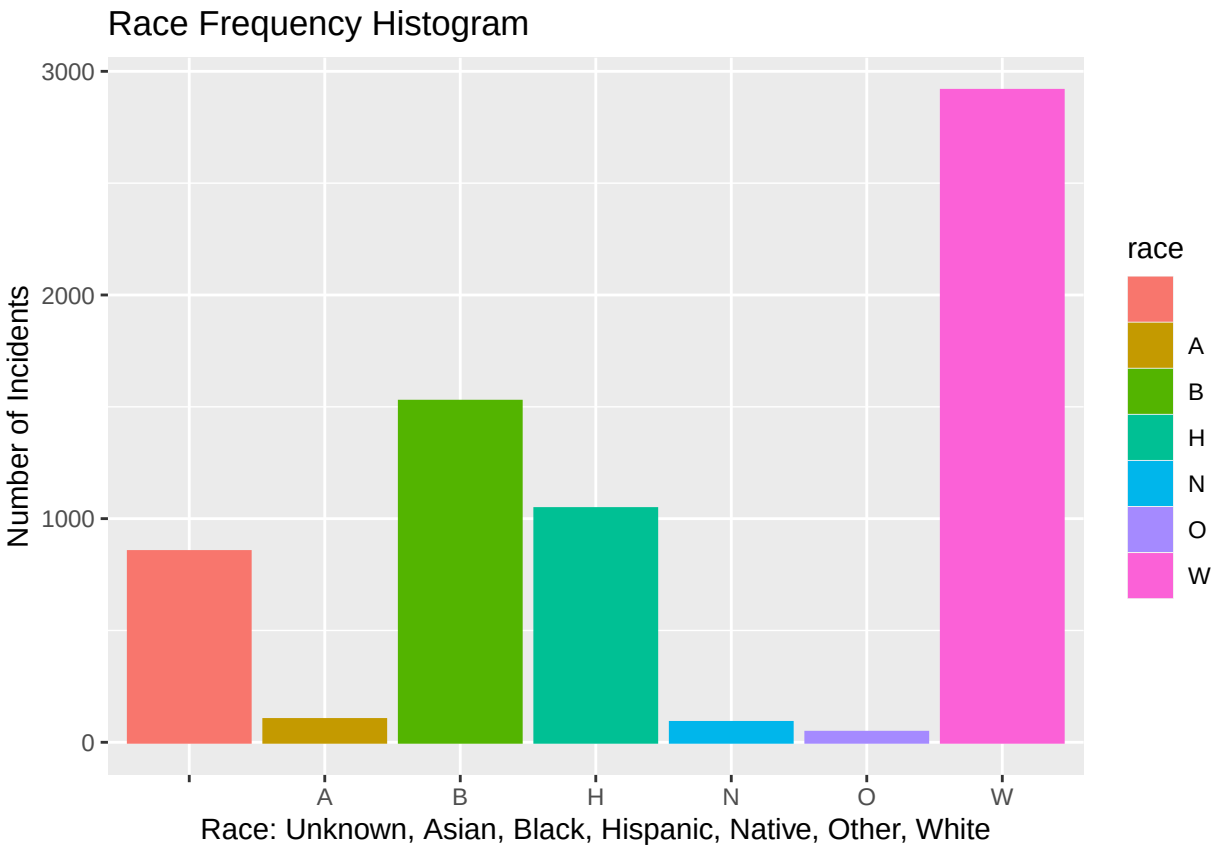
Means and Median for Numeric Variables (Age):

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

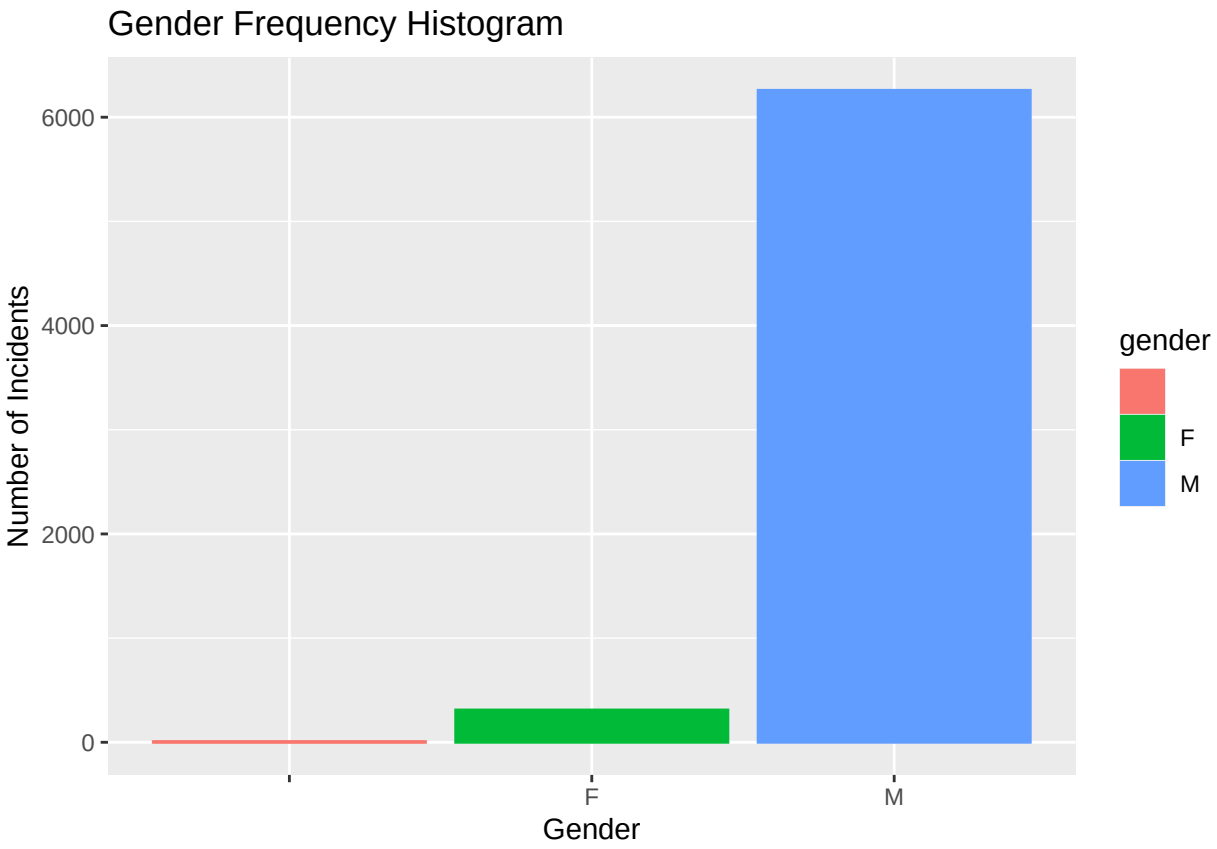
```
## [1] 35
```

#Frequency Graphs for Categorical Variables:

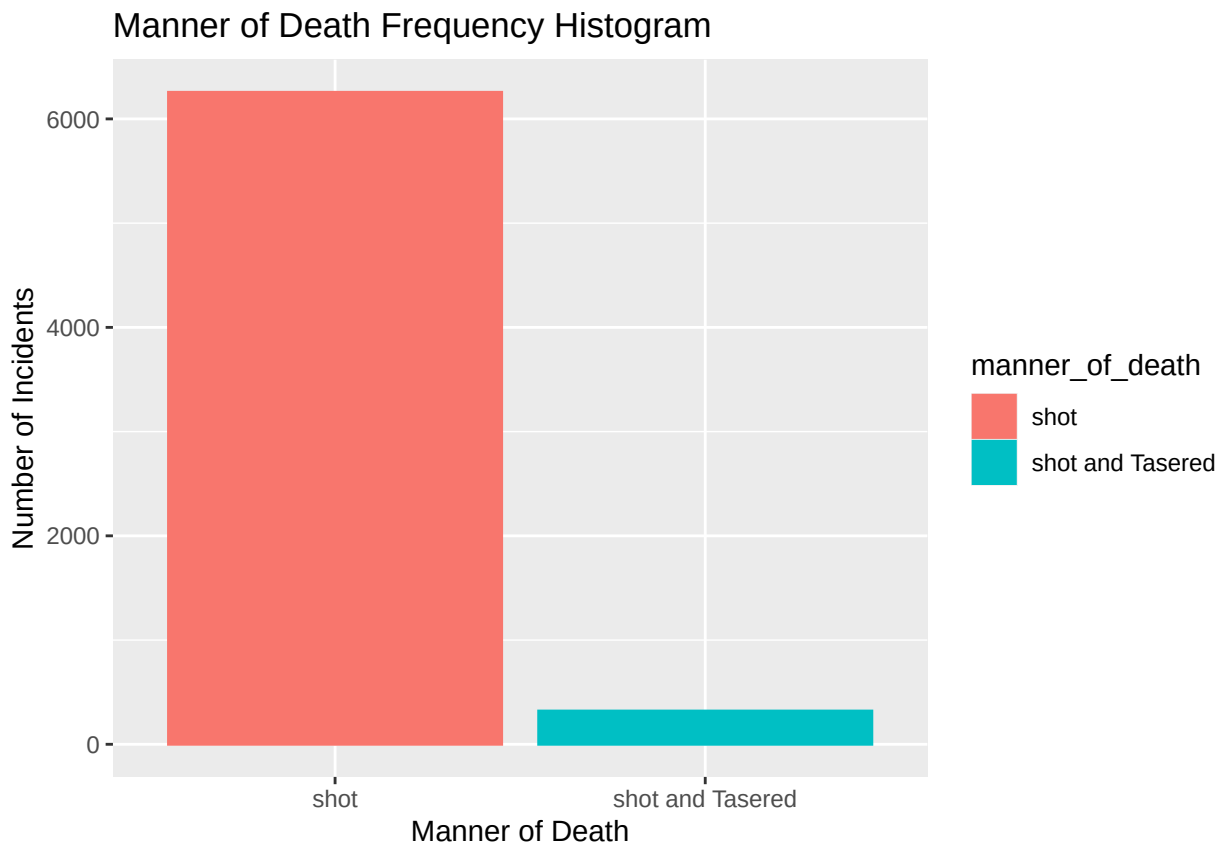
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



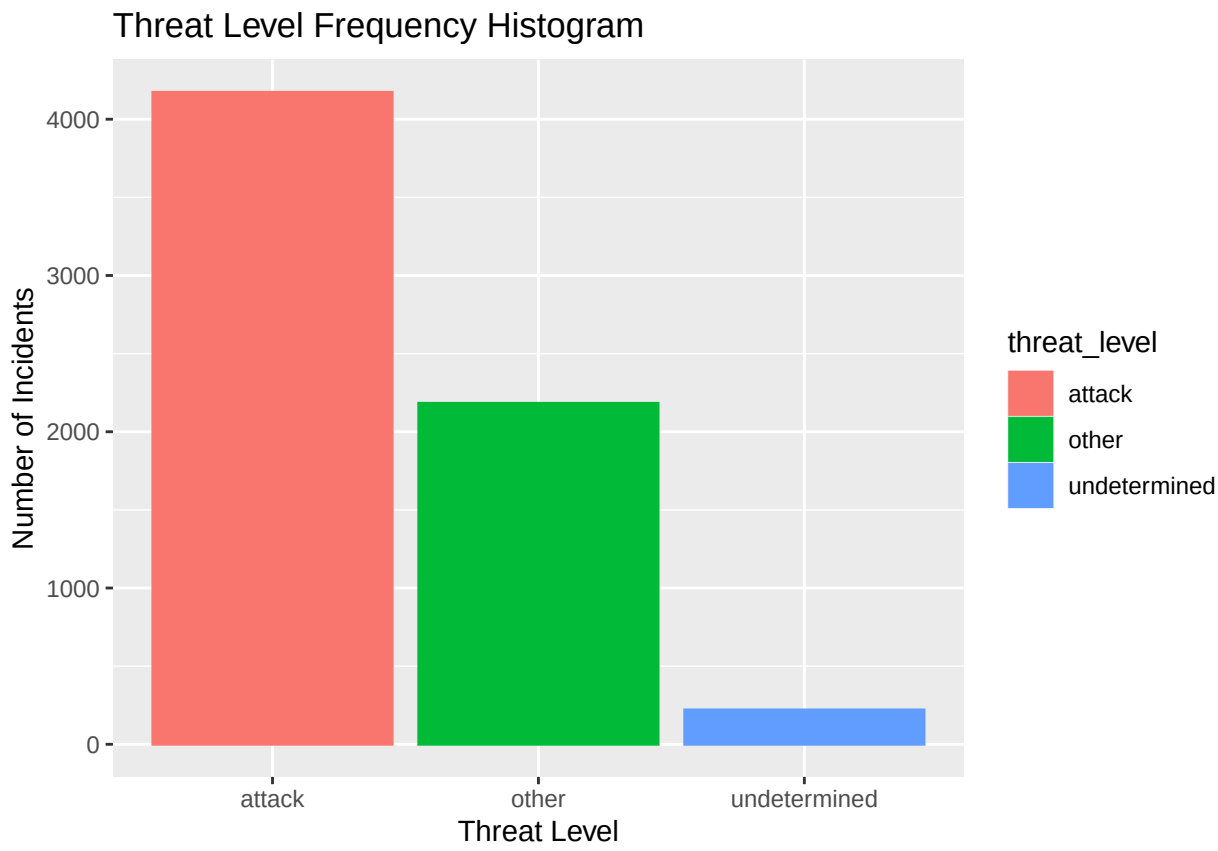

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



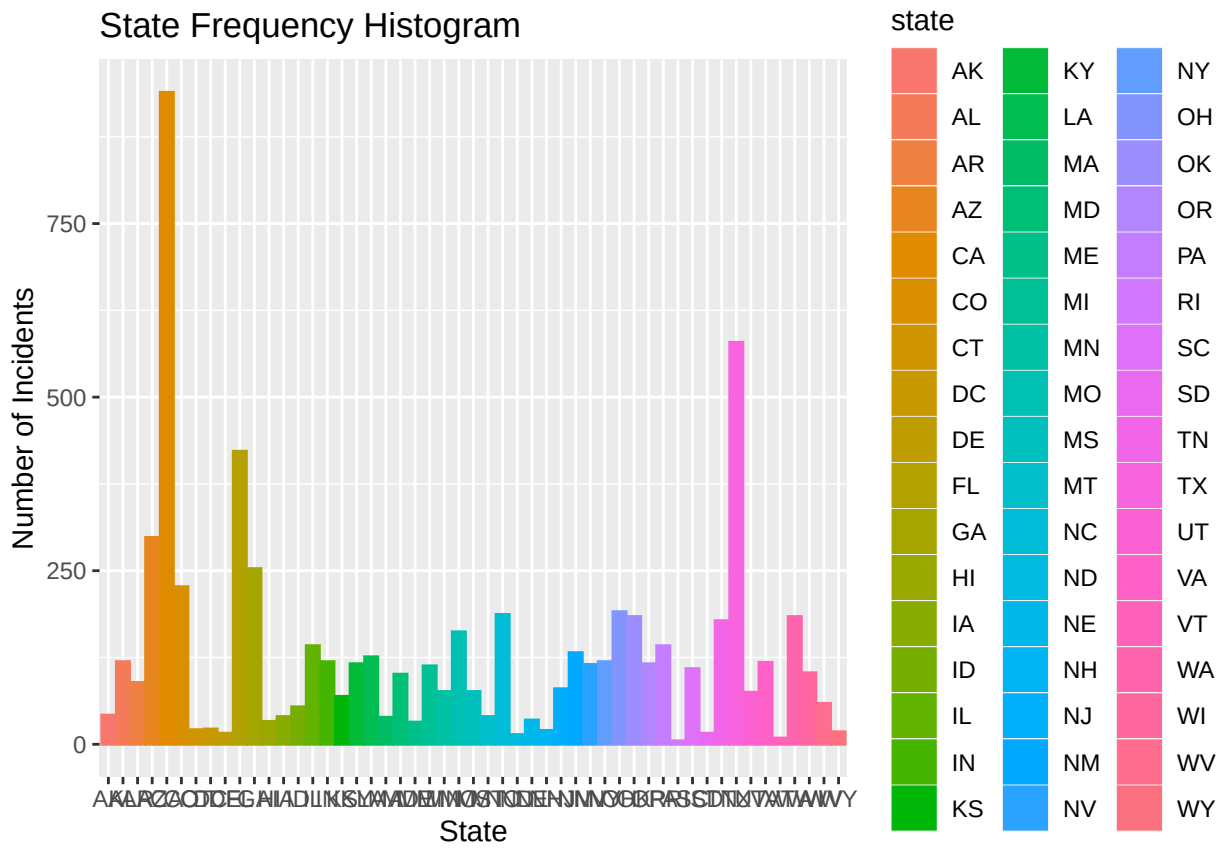

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




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




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



###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] 1810
```

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

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

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

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

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

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

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

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

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           regions       stbcp           gen.p           smi.p
##  Length:6574        MW:1080   Min.   :0.000   Min.   :0.818   Min.   :0.000
##  Class :character   NE: 568   1st Qu.:0.101   1st Qu.:0.938   1st Qu.:0.200
##  Mode  :character   NW:1810   Median :0.133   Median :0.952   Median :0.219
##                     SE:1890   Mean   :0.144   Mean   :0.952   Mean   :0.223
##                     SW:1226   3rd Qu.:0.183   3rd Qu.:0.966   3rd Qu.:0.265
##                               Max.   :0.409   Max.   :1.000   Max.   :0.556
##      flee.p      att.p          armed.p          MoD.p          age.avg
##  Min.   :0   Min.   :0.350   Min.   :0.778   Min.   :0.810   Min.   :33.1
##  1st Qu.:0   1st Qu.:0.564   1st Qu.:0.918   1st Qu.:0.938   1st Qu.:35.7
##  Median :0   Median :0.644   Median :0.934   Median :0.948   Median :36.9
##  Mean   :0   Mean   :0.635   Mean   :0.937   Mean   :0.951   Mean   :37.2
##  3rd Qu.:0   3rd Qu.:0.679   3rd Qu.:0.958   3rd Qu.:0.969   3rd Qu.:38.6
##  Max.   :0   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :44.4
##  Non_White_prop
##  Min.   :0.250
##  1st Qu.:0.455
##  Median :0.563
##  Mean   :0.557
##  3rd Qu.:0.635
##  Max.   :0.939
```

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

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

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

SMART Question and Answer

Within our dataset of police 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:

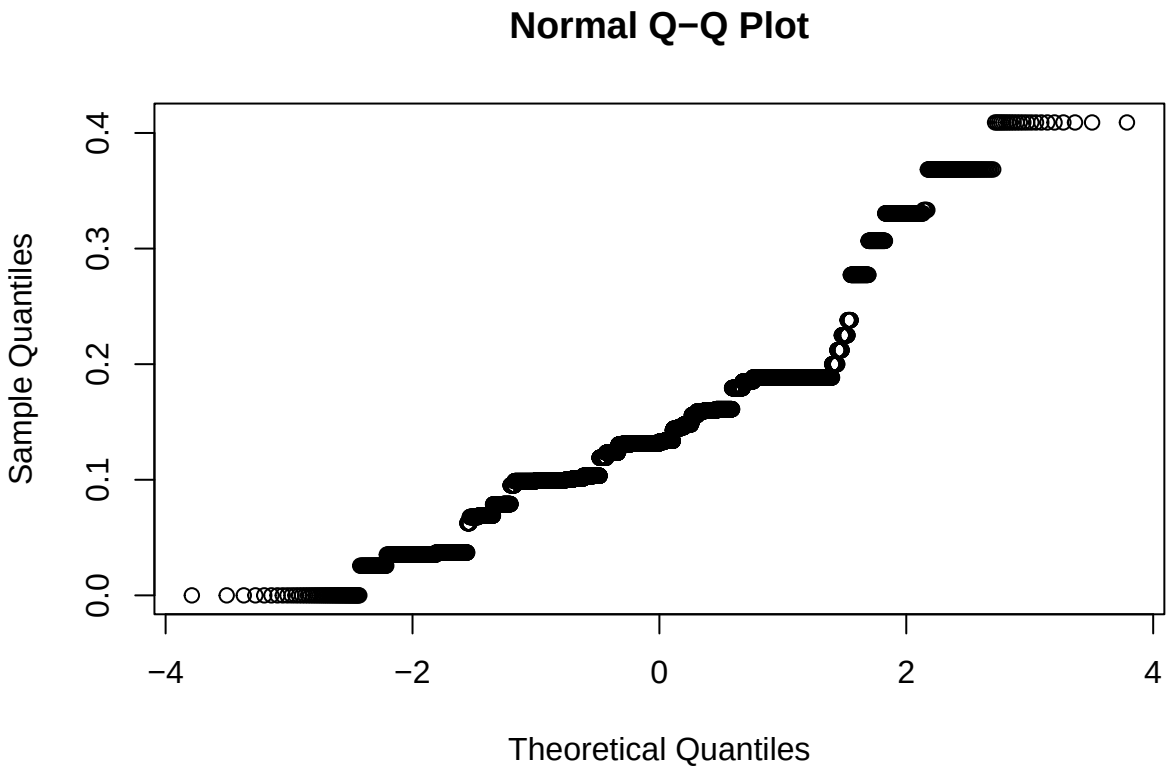
- state: State of Observation
- region: Region of Observation
- stbcp: State Body Camera On Proportion
- genp.p: proportion of male identified shooting victims by state
- smi.p: proportion of shooting victims by state with a documented sign of mental illness
- flee.p: proportion of shooting victims by 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 11
## # Groups:   state [6]
##   state regions  stbcp gen.p smi.p flee.p att.p armed.p MoD.p age.avg Non_Whit~1
##   <chr> <fct>    <dbl> <dbl> <dbl>  <dbl> <dbl>   <dbl> <dbl>   <dbl>      <dbl>
## 1 WA    NW      0.103  0.967 0.337      0 0.549   0.935 0.946    36.9      0.576
## 2 OR    NW      0.0690 0.974 0.302      0 0.517   0.957 0.957    39.2      0.328
## 3 KS    MW      0.130  0.913 0.217      0 0.696   0.928 0.942    36.7      0.406
## 4 CA    NW      0.188  0.952 0.219      0 0.564   0.918 0.938    35.5      0.736
## 5 CO    NW      0.123  0.952 0.137      0 0.634   0.952 0.969    35.7      0.507
## 6 OK    SW      0.179  0.978 0.212      0 0.707   0.908 0.924    37.5      0.413
## # ... with abbreviated variable name 1: Non_White_prop
```

```
##     state           regions       stbcp           gen.p           smi.p
##  Length:6574        MW:1080   Min.   :0.000   Min.   :0.818   Min.   :0.000
##  Class :character   NE: 568   1st Qu.:0.101   1st Qu.:0.938   1st Qu.:0.200
##  Mode  :character   NW:1810   Median :0.133   Median :0.952   Median :0.219
##                     SE:1890   Mean   :0.144   Mean   :0.952   Mean   :0.223
##                     SW:1226   3rd Qu.:0.183   3rd Qu.:0.966   3rd Qu.:0.265
##                               Max.   :0.409   Max.   :1.000   Max.   :0.556
##      flee.p      att.p         armed.p         MoD.p         age.avg
##  Min.   :0   Min.   :0.350   Min.   :0.778   Min.   :0.810   Min.   :33.1
##  1st Qu.:0   1st Qu.:0.564   1st Qu.:0.918   1st Qu.:0.938   1st Qu.:35.7
##  Median :0   Median :0.644   Median :0.934   Median :0.948   Median :36.9
##  Mean   :0   Mean   :0.635   Mean   :0.937   Mean   :0.951   Mean   :37.2
##  3rd Qu.:0   3rd Qu.:0.679   3rd Qu.:0.958   3rd Qu.:0.969   3rd Qu.:38.6
##  Max.   :0   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :44.4
##  Non_White_prop
```

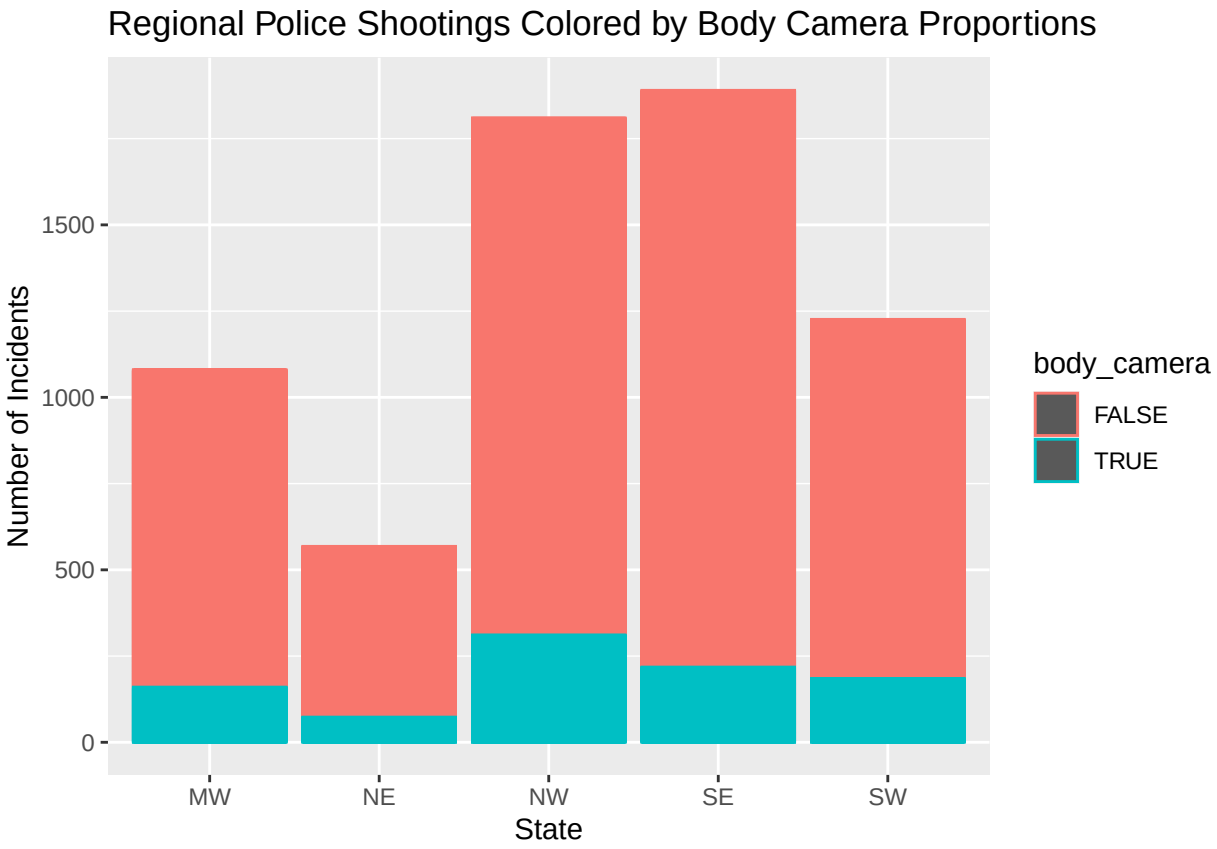
```
##  Min.   :0.250
##  1st Qu.:0.455
##  Median :0.563
##  Mean   :0.557
##  3rd Qu.:0.635
##  Max.   :0.939
```

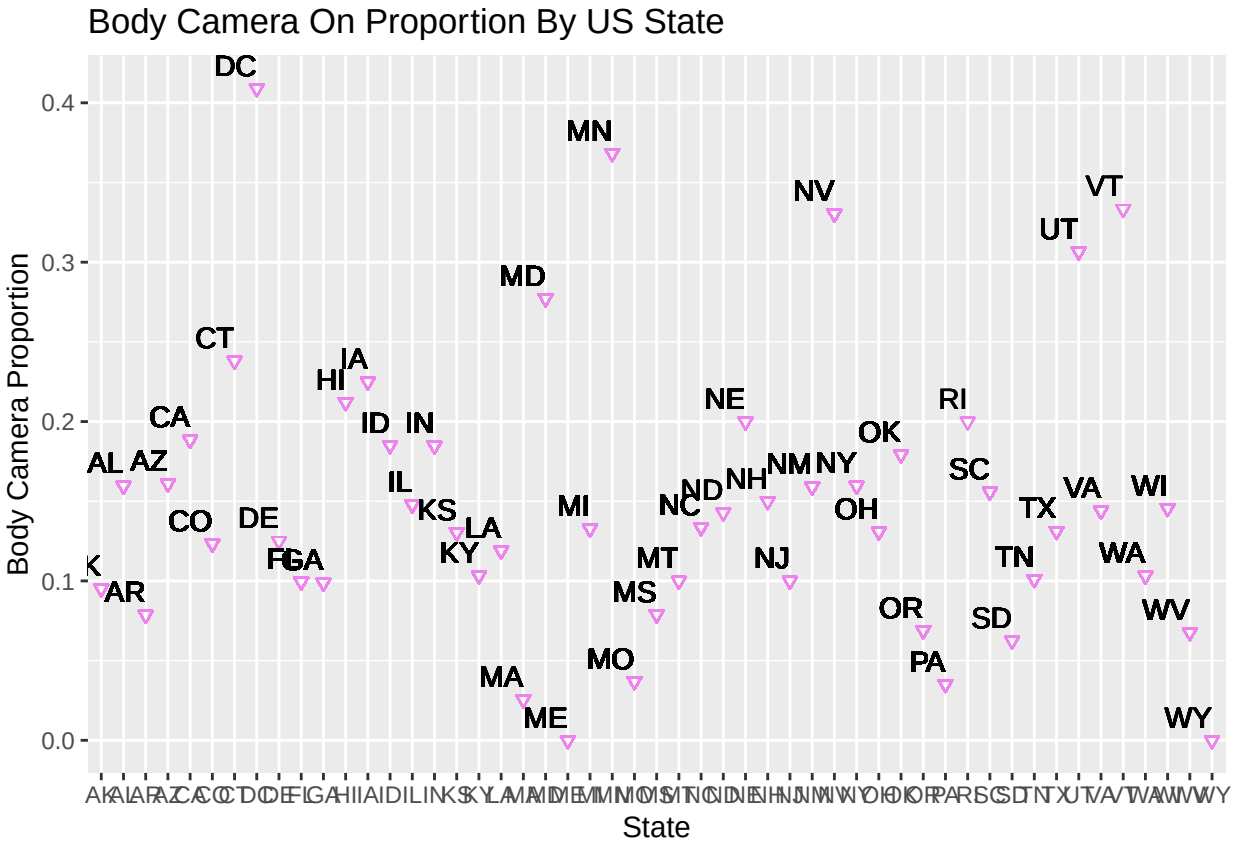
We now would like to check our data for normality:



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:


Regional Police Shootings Colored by Body Camera Proportions
Number of Incidents
1500
1000
500
0
MW
NE
NW
SE
SW
State
body_camera
FALSE
TRUE



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   0.101   0.133   0.144   0.183   0.409
```

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

```
##
##  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 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.