Module 2 Homework

Michael Puchalski

2025-02-02

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
police_df<-read.csv("PoliceKillings.csv", header = TRUE)
election_df<-read.csv("State_pop_election.csv", header =TRUE)
covid_df<-read.csv("stateCovid.csv", header =TRUE)
```

Question 1a

Using the raceethnicity variable, create a table and a bar chart that displays the proportions of victims in each race / ethnic level. Also, use your table and bar chart in conjunction with the US Census Bureau July 1 2024 estimates to explain what your data reveal.

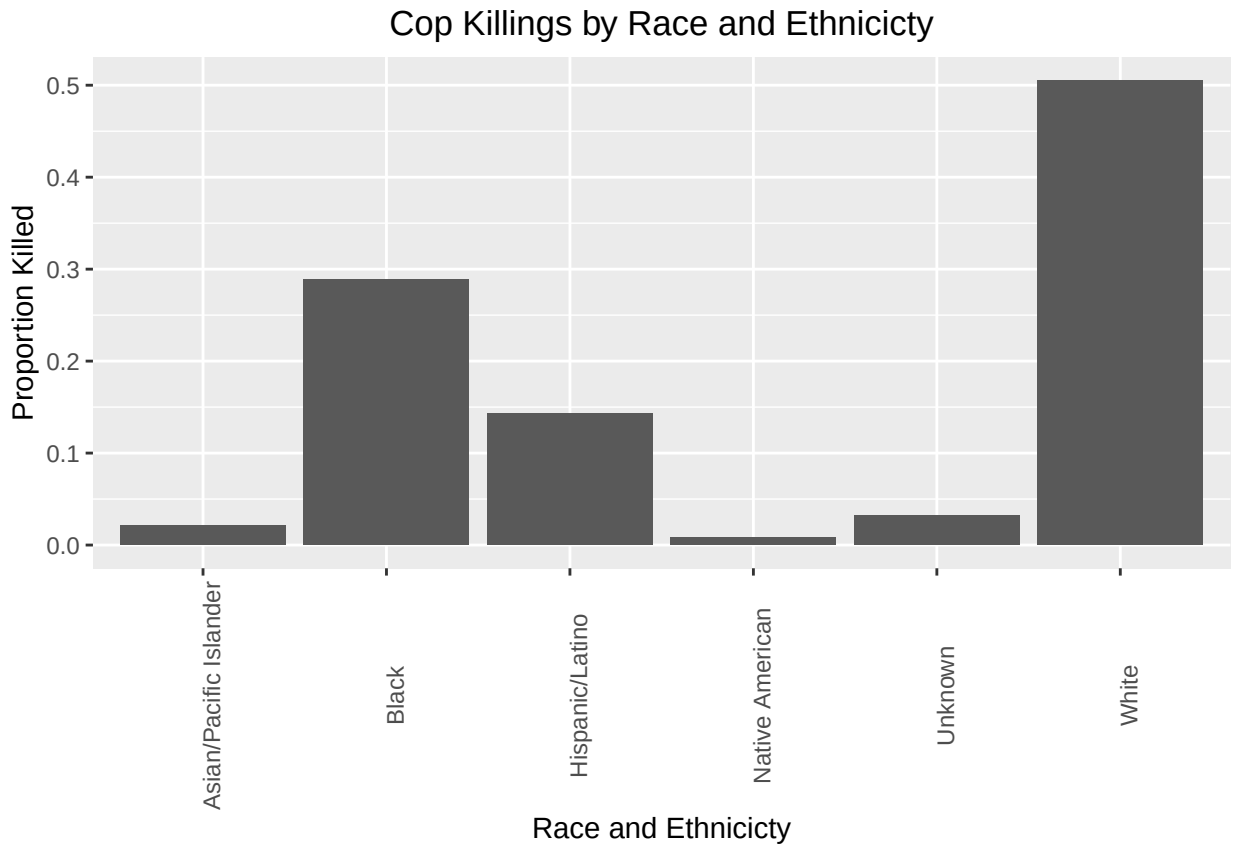
```
#police_df$raceethnicity<-factor(police_df$raceethnicity, levels=c("White", "Black", "Hispanic/Latino",
#table(police_df$raceethnicity)

race_df<-police_df%>%
  mutate(raceethnicity = as.character(raceethnicity))%>%
  group_by(raceethnicity)%>%
  summarize(Counts=n())%>%
  mutate(proportion = Counts/nrow(police_df))
race_df
```

```
## # A tibble: 6 x 3
##   raceethnicity          Counts proportion
##   <chr>                   <int>      <dbl>
## 1 Asian/Pacific Islander     10    0.0214
## 2 Black                     135    0.289
## 3 Hispanic/Latino            67    0.143
## 4 Native American             4    0.00857
## 5 Unknown                    15    0.0321
## 6 White                     236    0.505
```

```
# ggplot(police_df, aes(x = raceethnicity))+
  # geom_bar() +
  # theme(axis.text.x = element_text(angle = 90),
    # plot.title = element_text(hjust=0.5))+
  # labs(x="Race and Ethnicicty", y = "Number Killed", title = "Cop Killings by Race and Ethnicicty")

ggplot(race_df, aes(x = raceethnicity, y = proportion))+
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90),
    plot.title = element_text(hjust=0.5))+
  labs(x="Race and Ethnicicty", y = "Proportion Killed", title = "Cop Killings by Race and Ethnicicty")
```



Question 1b

Convert the variable age, the age of the victim, to be numeric, and call this new variable age.num. Use the is.numeric() function to confirm that the newly created variable is numeric (and output the result), and add this new variable to your data frame.

```
police_df_clean<-police_df%>%
  drop_na(age)%>%
  mutate(age.num = as.numeric(age))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age.num = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
is.numeric(police_df_clean$age.num)
```

```
## [1] TRUE
```

```r
head(police_df_clean, n=6)
```

```
##                 name age gender   raceethnicity    month day year
## 1 A'donte Washington  16   Male           Black February  23 2015
## 2     Aaron Rutledge  27   Male           White    April   2 2015
## 3        Aaron Siler  26   Male           White    March  14 2015
## 4       Aaron Valdez  25   Male Hispanic/Latino    March  11 2015
## 5       Adam Jovicic  29   Male           White    March  19 2015
## 6      Adam Reinhart  29   Male           White    March   7 2015
##            streetaddress         city state latitude  longitude state_fp
## 1           Clearview Ln    Millbrook    AL 32.52958  -86.36283        1
## 2 300 block Iris Park Dr    Pineville    LA 31.32174  -92.43486       22
## 3   22nd Ave and 56th St      Kenosha    WI 42.58356  -87.83571       55
## 4      3000 Seminole Ave   South Gate    CA 33.93930 -118.21946        6
## 5         364 Hiwood Ave Munroe Falls    OH 41.14857  -81.42988       39
## 6    18th St and Palm Ln      Phoenix    AZ 33.46938 -112.04332        4
##   county_fp tract_ce      geo_id county_id             namelsad
## 1        51    30902  1051030902      1051  Census Tract 309.02
## 2        79    11700 22079011700     22079     Census Tract 117
## 3        59     1200 55059001200     55059      Census Tract 12
## 4        37   535607  6037535607      6037 Census Tract 5356.07
## 5       153   530800 39153530800     39153    Census Tract 5308
## 6        13   111602  4013111602      4013 Census Tract 1116.02
##              lawenforcementagency   cause   armed  pop share_white share_black
## 1     Millbrook Police Department Gunshot      No 3779        60.5        30.5
## 2 Rapides Parish Sheriff's Office Gunshot      No 2769        53.8        36.2
## 3       Kenosha Police Department Gunshot      No 4079        73.8         7.7
## 4    South Gate Police Department Gunshot Firearm 4343         1.2         0.6
## 5          Kent Police Department Gunshot      No 6809        92.5         1.4
## 6       Phoenix Police Department Gunshot      No 4682           7         7.7
##   share_hispanic p_income h_income county_income comp_income county_bucket
## 1            5.6    28375    51367         54766   0.9379359             3
## 2            0.5    14678    27972         40930   0.6834107             2
## 3           16.8    25286    45365         54930   0.8258693             2
## 4           98.8    17194    48295         55909   0.8638144             3
## 5            1.7    33954    68785         49669   1.3848678             5
## 6             79    15523    20833         53596   0.3887044             1
##   nat_bucket  pov      urate    college  X age.num
## 1          3 14.1 0.09768638 0.16850951 NA      16
## 2          1 28.8 0.06572379 0.11140236 NA      27
## 3          3 14.6 0.16629314 0.14731227 NA      26
## 4          3 11.7 0.12482727 0.05013293 NA      25
## 5          4  1.9 0.06354983 0.40395421 NA      29
## 6          1   58 0.07365145 0.10295519 NA      29
```

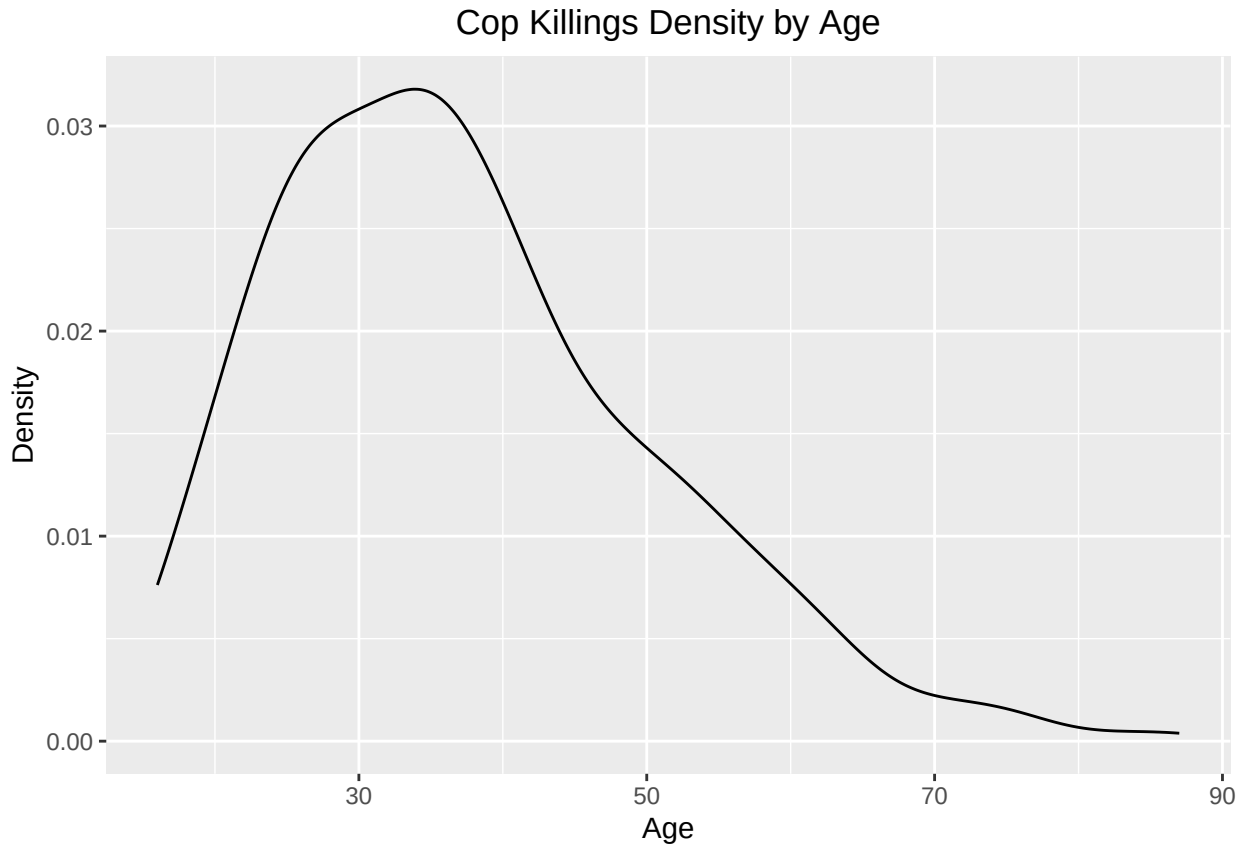
Question 1c

Create a density plot of the variable age.num. Comment on this density plot

```r
ggplot(police_df_clean, aes(x=age.num))+
  geom_density()+
```

```
  theme(axis.text.x = element_text(angle = 0),
    plot.title = element_text(hjust=0.5))+
  labs(x="Age", y = "Density", title = "Cop Killings Density by Age")
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_density()`).
```



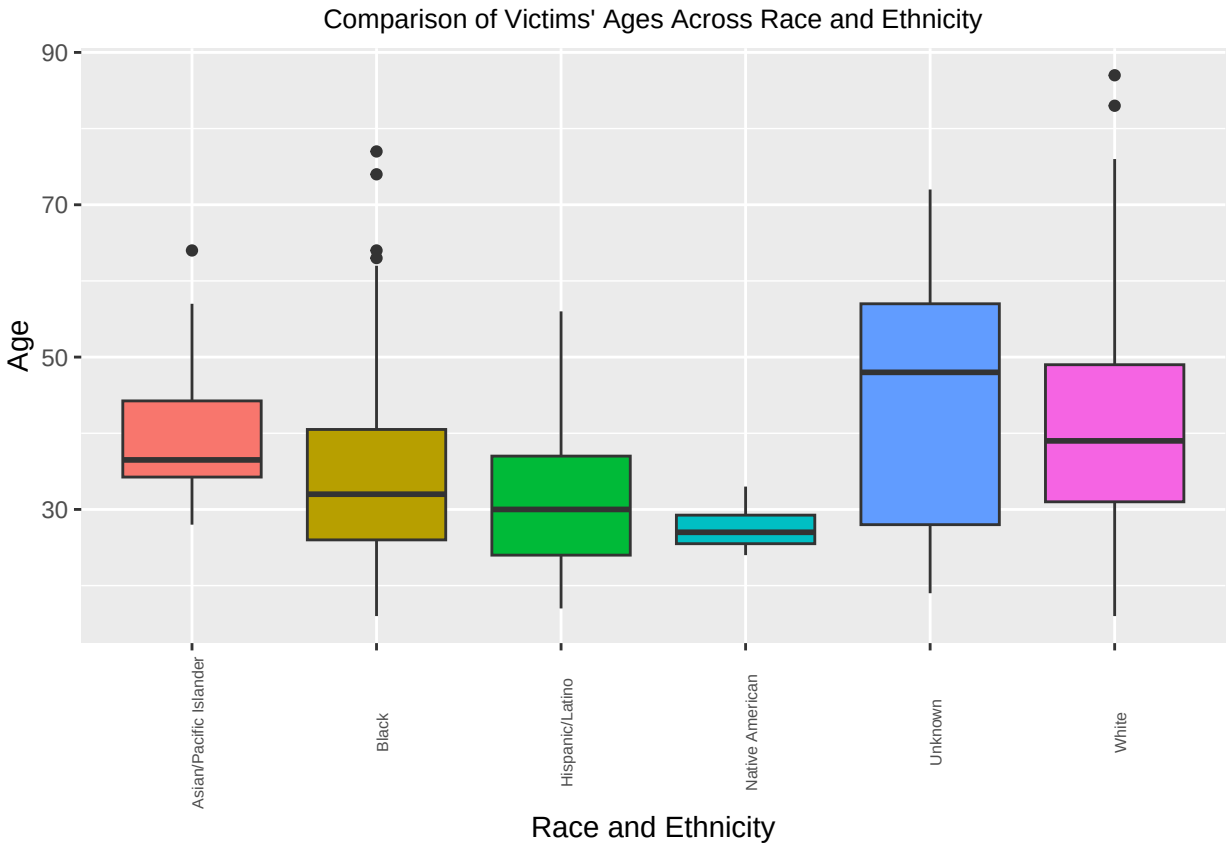
We can see looking at the density plot that there is a bell curve type scale where you see a higher concentration of cop killings happen towards what would be considered middle ages with it decreasing as the groupings get older.

Question 1d

Create a visualization to compare the ages of victims across the different race / ethnicity levels. Comment on the visualization

```
ggplot(police_df_clean, aes(x=raceethnicity, y=age.num, fill=raceethnicity))+
  geom_boxplot()+
  theme(axis.text.x = element_text(size =6, angle = 90),
        plot.title = element_text(size = 10, hjust = 0.5),
        legend.position = "none")+
  labs(x = "Race and Ethnicity", y = "Age", title = "Comparison of Victims' Ages Across Race and Ethnic
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



From this visualization you can show that the overall spread of age density changes across racer and ethnicity with Native American Deaths being significantly more concentrated towards the lower age groups compared to other ethnicities while white victims are the largest spread with them having the great aged outliers. Meanwhile, outside of outliers of Black and Hispanic/Latino groups, they are highly concentrated between about 25 and 30 years of age.

Question 1e

Create a visualization to compare the different causes of death (variable cause) across the different race / ethnicity levels. Comment on this visualization, specif- ically on whether the cause of death appears to be independent of the victim's race / ethnicity.

```
mytab2<-table(police_df_clean$raceethnicity, police_df_clean$cause)
print("Table 1. Cause Counts")
```

```
## [1] "Table 1. Cause Counts"
```

```
mytab2
```

```
##
##                          Death in custody Gunshot Struck by vehicle Taser
##   Asian/Pacific Islander                0       7                 1     2
##   Black                                 8     110                 3    13
##   Hispanic/Latino                       1      64                 0     1
##   Native American                       1       2                 0     1
##   Unknown                               0      14                 0     1
##   White                                 4     214                 8     9
##
```

```
##                           Unknown
##   Asian/Pacific Islander       0
##   Black                        1
##   Hispanic/Latino              1
##   Native American              0
##   Unknown                      0
##   White                        1
```

```
print("Table. 2 Percentage of Cause per Race and Ethnicity")
```

```
## [1] "Table. 2 Percentage of Cause per Race and Ethnicity"
```

```
round(prop.table(mytab2,1)*100, 2)
```

```
##
##                           Death in custody Gunshot Struck by vehicle Taser
##   Asian/Pacific Islander             0.00   70.00             10.00 20.00
##   Black                              5.93   81.48              2.22  9.63
##   Hispanic/Latino                    1.49   95.52              0.00  1.49
##   Native American                   25.00   50.00              0.00 25.00
##   Unknown                            0.00   93.33              0.00  6.67
##   White                              1.69   90.68              3.39  3.81
##
##                           Unknown
##   Asian/Pacific Islander    0.00
##   Black                     0.74
##   Hispanic/Latino           1.49
##   Native American           0.00
##   Unknown                   0.00
##   White                     0.42
```

```
print("Table 3. Percentage of Cause across Race and Ethnicity")
```

```
## [1] "Table 3. Percentage of Cause across Race and Ethnicity"
```

```
round(prop.table(mytab2,2)*100, 2)
```

```
##
##                           Death in custody Gunshot Struck by vehicle Taser
##   Asian/Pacific Islander             0.00    1.70              8.33  7.41
##   Black                             57.14   26.76             25.00 48.15
##   Hispanic/Latino                    7.14   15.57              0.00  3.70
##   Native American                    7.14    0.49              0.00  3.70
##   Unknown                            0.00    3.41              0.00  3.70
##   White                             28.57   52.07             66.67 33.33
##
##                           Unknown
##   Asian/Pacific Islander    0.00
##   Black                    33.33
##   Hispanic/Latino          33.33
##   Native American           0.00
##   Unknown                   0.00
##   White                    33.33
```

In this table figure, you can see both the total number of death by cause and race/ethnicity in table 1. In tables 2 and 3, you can see the percentage of the total deaths each cause has across the groups followed by the percentage of each cause each enthnic group makes up respectively.

When interpreting this, the table with the most story to tell is Table 3 where it shows that black victims make up a disproportionate amount of deaths while in custody compared to the other groups and almost half of the deaths from tazing. Meanwhile white victims make up over half of all gunshot victims and victims who die by vehicle strike.

When combining this data with the data from table 2, one can see that the majority of all cop killings are a result of gunshots, while black vicitims make up a concerning amount of death cases that could be resulting from police brutality via death in custody and tasing.

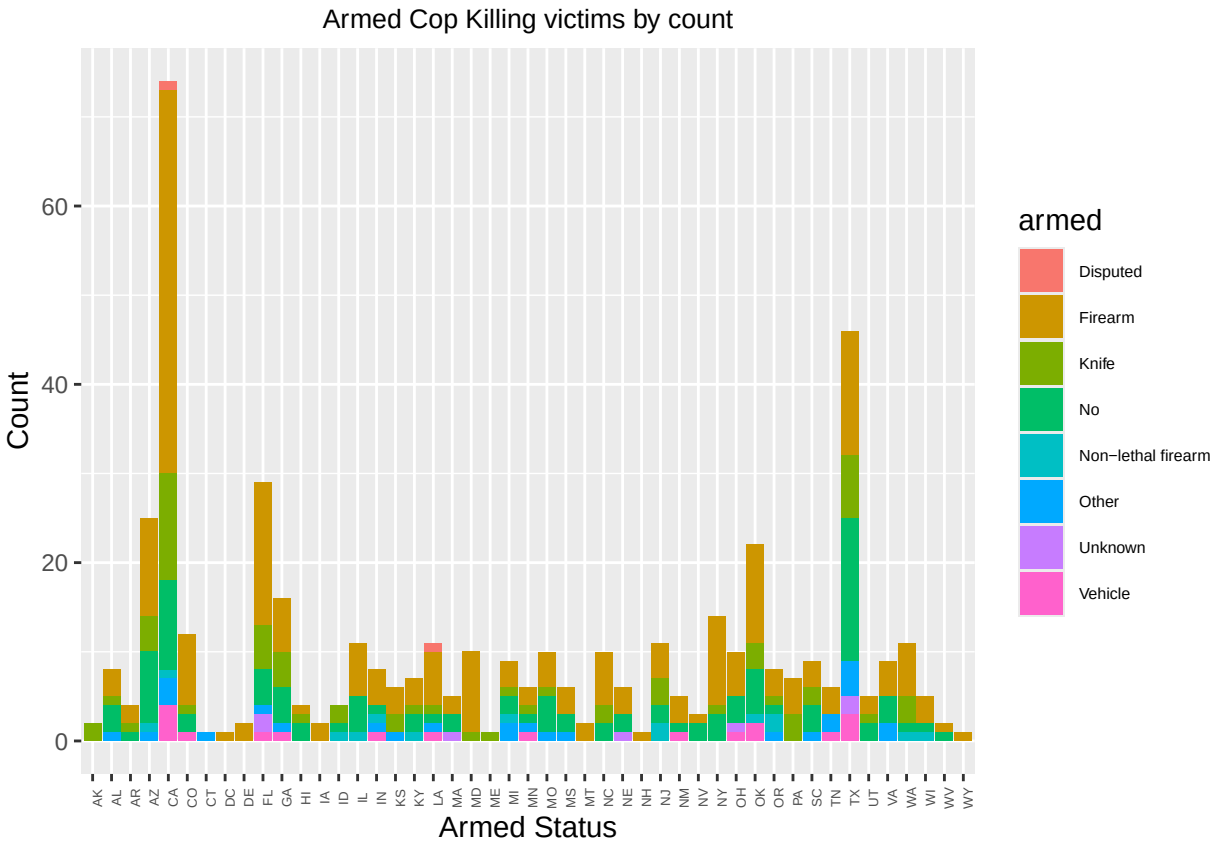
Question 1f

Pick at least two variables from the dataset and create a suitable visualization of the variables. Comment on what the visualization reveals. You may create new variables based on existing variables, and decribe how you created the new variables.

```
armed_df<-police_df_clean%>%
  drop_na(armed)%>%
  drop_na(state)%>%
  group_by(state, armed)
armed_df
```

```
## # A tibble: 467 x 36
## # Groups:   state, armed [154]
##    name   age   gender raceethnicity month   day  year streetaddress city  state
##    <chr>  <chr> <chr>  <chr>         <chr> <int> <int> <chr>         <chr> <chr>
##  1 A'don~ 16    Male   Black         Febr~    23  2015 Clearview Ln  Mill~ AL
##  2 Aaron~ 27    Male   White         April     2  2015 300 block Ir~ Pine~ LA
##  3 Aaron~ 26    Male   White         March    14  2015 22nd Ave and~ Keno~ WI
##  4 Aaron~ 25    Male   Hispanic/Lat~ March    11  2015 3000 Seminol~ Sout~ CA
##  5 Adam ~ 29    Male   White         March    19  2015 364 Hiwood A~ Munr~ OH
##  6 Adam ~ 29    Male   White         March     7  2015 18th St and ~ Phoe~ AZ
##  7 Adria~ 22    Male   Hispanic/Lat~ March    27  2015 4000 Union A~ Bake~ CA
##  8 Adria~ 35    Male   Hispanic/Lat~ March    26  2015 1500 Bayview~ Wilm~ CA
##  9 Alan ~ 44    Male   White         Janu~    28  2015 Pickett Runn~ Suns~ TX
## 10 Alan ~ 31    Male   White         Febr~     7  2015 200 Abbie St~ Wyom~ MI
## # i 457 more rows
## # i 26 more variables: latitude <dbl>, longitude <dbl>, state_fp <int>,
## #   county_fp <int>, tract_ce <int>, geo_id <dbl>, county_id <int>,
## #   namelsad <chr>, lawenforcementagency <chr>, cause <chr>, armed <chr>,
## #   pop <int>, share_white <chr>, share_black <chr>, share_hispanic <chr>,
## #   p_income <chr>, h_income <int>, county_income <int>, comp_income <dbl>,
## #   county_bucket <int>, nat_bucket <int>, pov <chr>, urate <dbl>, ...
```

```
ggplot(armed_df, aes(x=state, fill = armed))+
  geom_bar() +
  theme(axis.text.x = element_text(size =5, angle = 90),
        plot.title = element_text(size = 10, hjust = 0.5),
        legend.text = element_text(size = 6))+
  labs(x = "Armed Status", y = "Count", fill = "armed", title = "Armed Cop Killing victims by count")
```



The armed victims had two high spiking States, one with stricter gun/weapon laws in California, meanwhile Texas with more lenient laws towards weapons has be come the second largest producer of weapon related police killings. This suggests that this type of statistic can be heavily influenced when batched together due to these two states also being the largest populations in the United States.

Question 2a

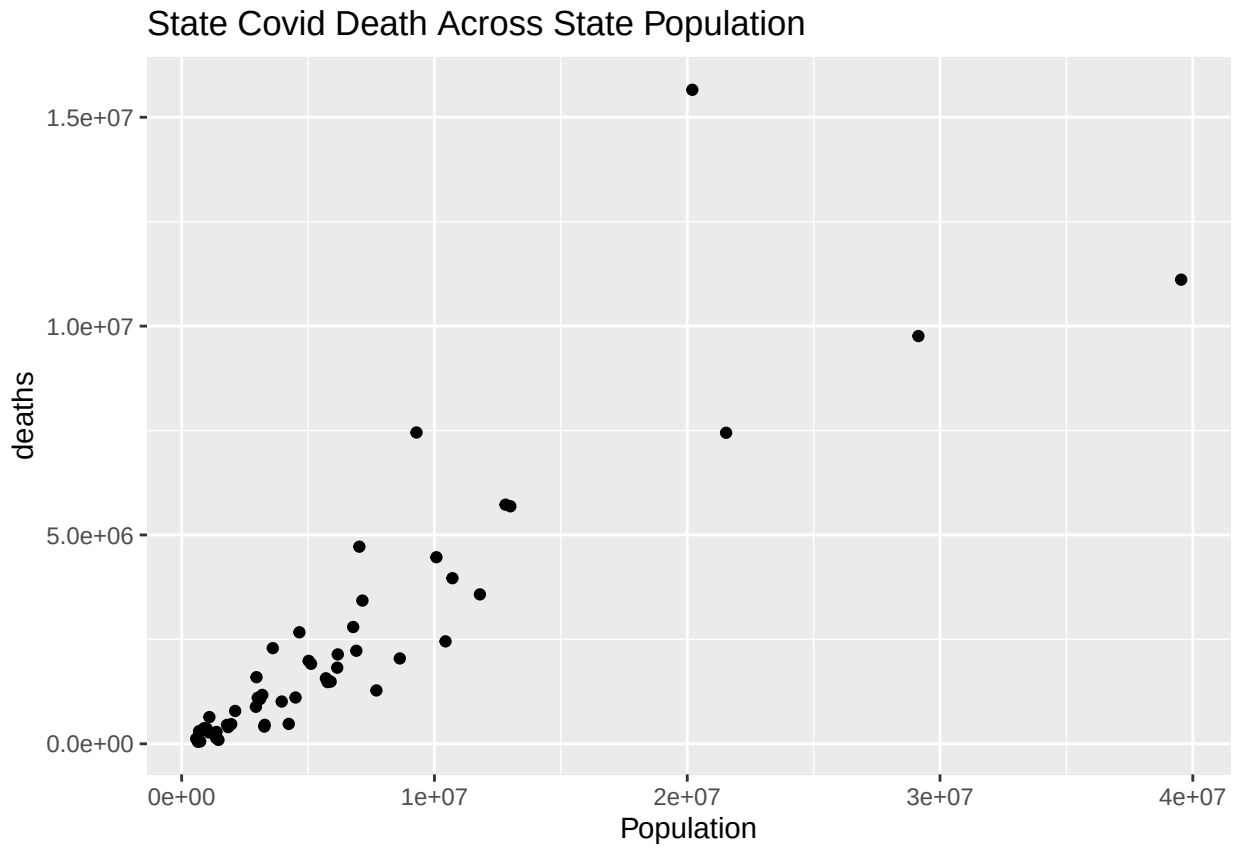
```
election_df <- election_df %>%
  rename("state" = "State")

merge_df <- covid_df%>%
  inner_join(election_df, by = c("state"))
head(merge_df, n=6 )
```

```
##        state     cases   deaths state.rate Population Election
## 1    Alabama 108217598  1986203       1.84    5024279    Trump
## 2     Alaska  12176769    57974       0.48     733391    Trump
## 3    Arizona 170946356  3429632       2.01    7151502    Biden
## 4   Arkansas  67032054  1101842       1.64    3011524    Trump
## 5 California 714568374 11112159       1.56   39538223    Biden
## 6   Colorado  93331929  1477994       1.58    5773714    Biden
```

Question 2b

```
ggplot(merge_df, aes(x = Population, y = deaths))+
  geom_point()+
  labs(title = "State Covid Death Across State Population")
```



Looking at this visualization image for State covid deaths, the size of the population of a state absolutely influenced its covid death count. This is why in a point chart, just as you would expect from an epidemiological background, the Population and the Death count are almost linear in relation.