HW 2 Solutions

```
library(tidyverse)

Police.Victims<-read.csv("PoliceKillings.csv", header=TRUE)
```

Question 1)

a)

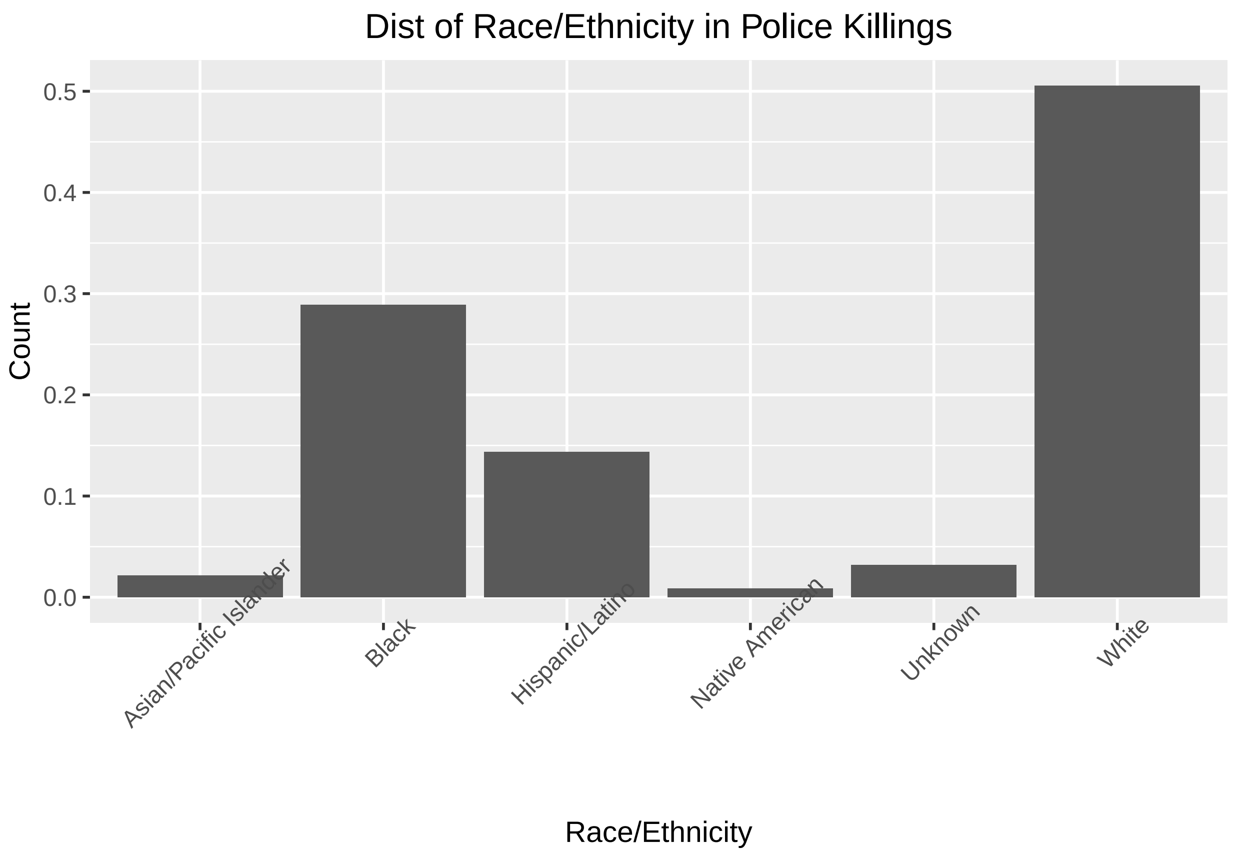
The table of the proportion of victims in each race / ethnicity is shown below.

```
##Table
mytab<-table(Police.Victims$raceethnicity)
round(prop.table(mytab)*100, 2)
```

```
##
## Asian/Pacific Islander                  Black        Hispanic/Latino
##                   2.14                  28.91                  14.35
##        Native American                Unknown                  White
##                   0.86                   3.21                  50.54
```

The bar chart of the proportion of victims in each race / ethnicity is shown below.

```
newData<-Police.Victims%>%
  group_by(raceethnicity)%>%
  summarize(Counts=n())%>%
  mutate(Percent=Counts/nrow(Police.Victims))

##then create a new bar chart, by adding some extra arguments in
##aes and geom_bar
ggplot(newData, aes(x=raceethnicity, y=Percent))+
  geom_bar(stat="identity")+
  theme(axis.text.x = element_text(angle = 45),
        plot.title = element_text(hjust = 0.5))+
  labs(x="Race/Ethnicity", y="Count",
       title="Dist of Race/Ethnicity in Police Killings")
```



Based on the data, about 2% of victims are AAPI, 29% of victims are Black, 14% of victims are Hispanic, less than 1% are Native American, about 51% are white, with the rest being of known race / ethnicity.

Based on Census data, the proportion of Americans who are white is around 76%, while the proportion of Americans who are Black is around 14%. Compared to the proportion of Americans, a higher proportion of Blacks are victims of police killings, while a lower proportion of whites are victims of police killings.

b)

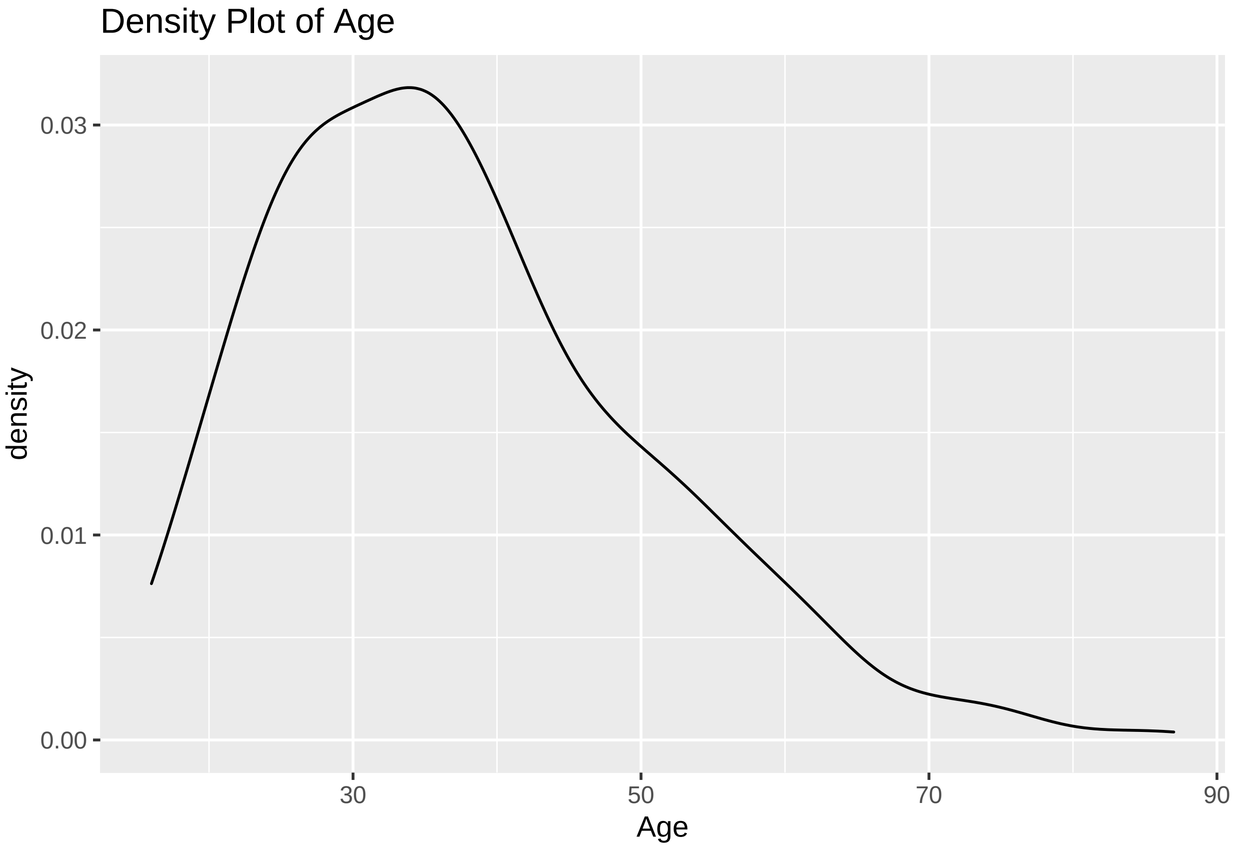
```
Police.Victims<-Police.Victims%>%
  mutate(age.num=as.numeric(as.character(age)))

is.numeric(Police.Victims$age.num)
```

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

c)

The density plot for the age of police victims is shown below.

```
ggplot(Police.Victims,aes(x=age.num))+
  geom_density()+
  labs(x="Age", title="Density Plot of Age")
```

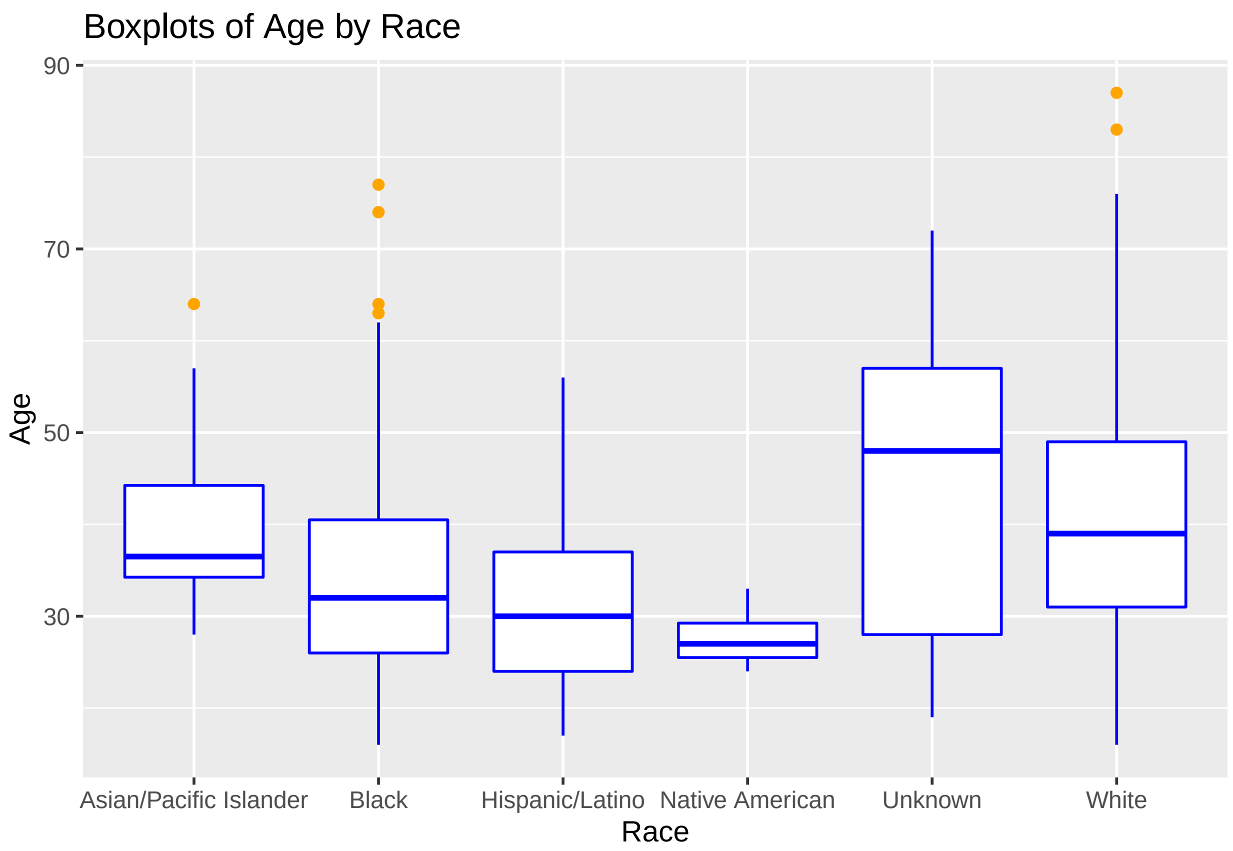


We can see that the distribution of ages is right skewed. Most of the victims are around their early 30s.

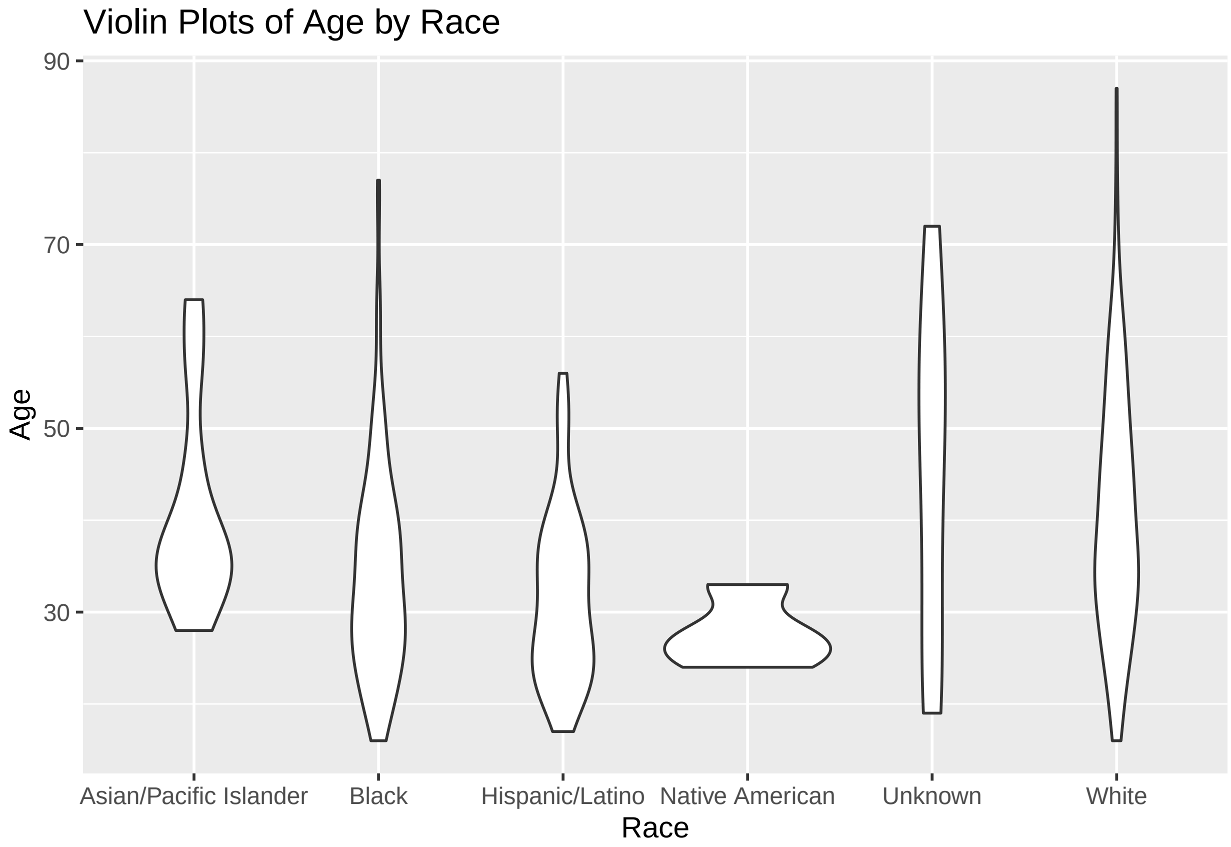
d)

Boxplots and violin plots comparing the distribution of ages for each race / ethnicity are shown below.

```
##boxplot of age and race
ggplot(Police.Victims, aes(x=raceethnicity, y=age.num))+
  geom_boxplot(color="blue", outlier.color = "orange" )+
  labs(x="Race", y="Age", title="Boxplots of Age by Race")
```



```
##violin plot of age and race
ggplot(Police.Victims, aes(x=raceethnicity, y=age.num))+
  geom_violin()+
  labs(x="Race", y="Age", title="Violin Plots of Age by Race")
```

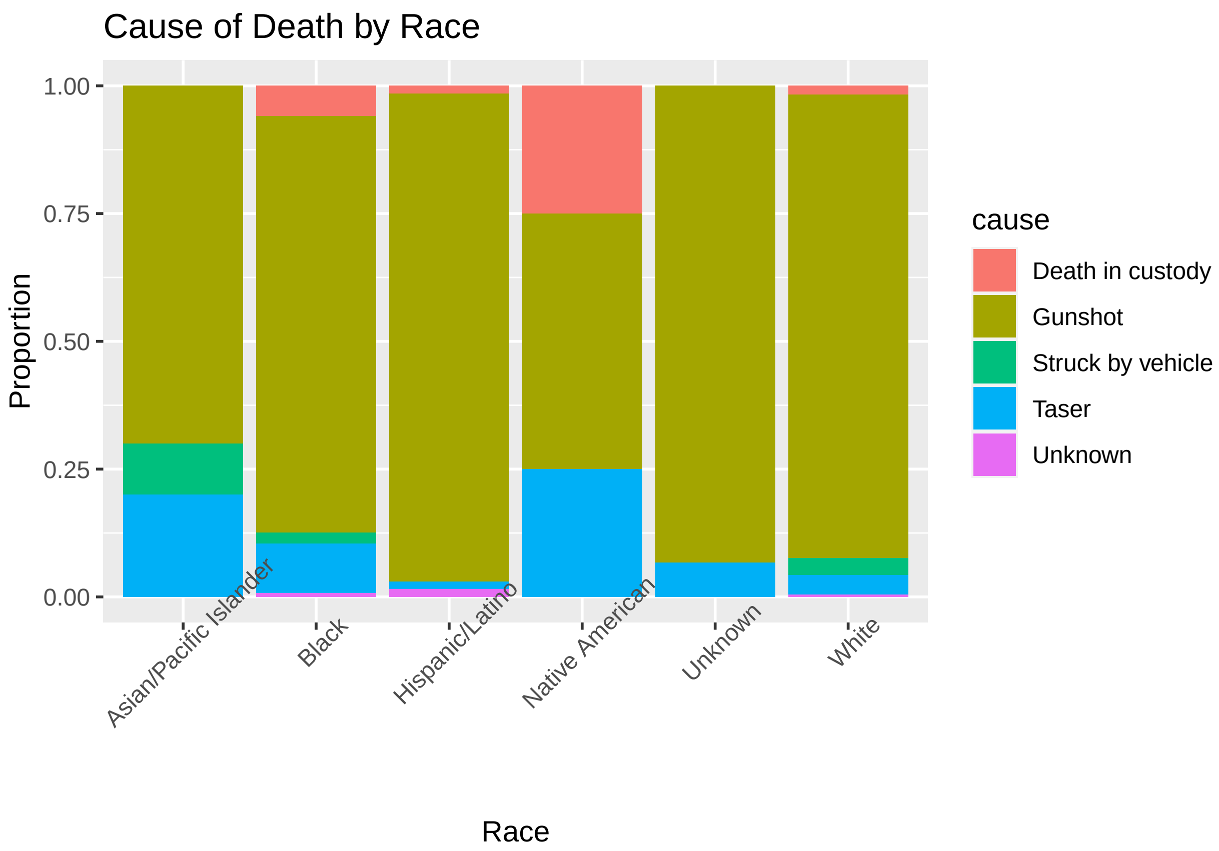


From the boxplots, we can see that based on median ages, Native American victims tend to be the youngest, followed by Hispanic, Black, AAPI, and then white victims. We see the same general trend with the violin plots, although the distributions tend to be right skewed as well. A higher proportion of victims are younger.

e)

A bar chart comparing the cause of deaths for each race / ethnicity is shown below.

```
##cause of death by race
ggplot(Police.Victims, aes(x=raceethnicity, fill=cause))+
  geom_bar(position = "fill")+
  labs(x="Race", y="Proportion",title="Cause of Death by Race")+
  theme(axis.text.x = element_text(angle = 45))
```



Cause of death does not seem to be independent of race. If cause of death and race are independent, we expect the proportions for causes of death to be similar across all races / ethinicities. For example, the proportion of AAPI and Native American victims by gunshot is less than for other races.

f)

Answers vary.

Question 2)

```
##remove column one
state.level<-read.csv("stateCovid.csv", header=TRUE)
state.level<-state.level[,-1]
```

a)

```
election<-read.csv("State_pop_election.csv", header=TRUE)

##merge data frames
state.data<-state.level %>%
  inner_join(election, by="State")

head(state.data)
```

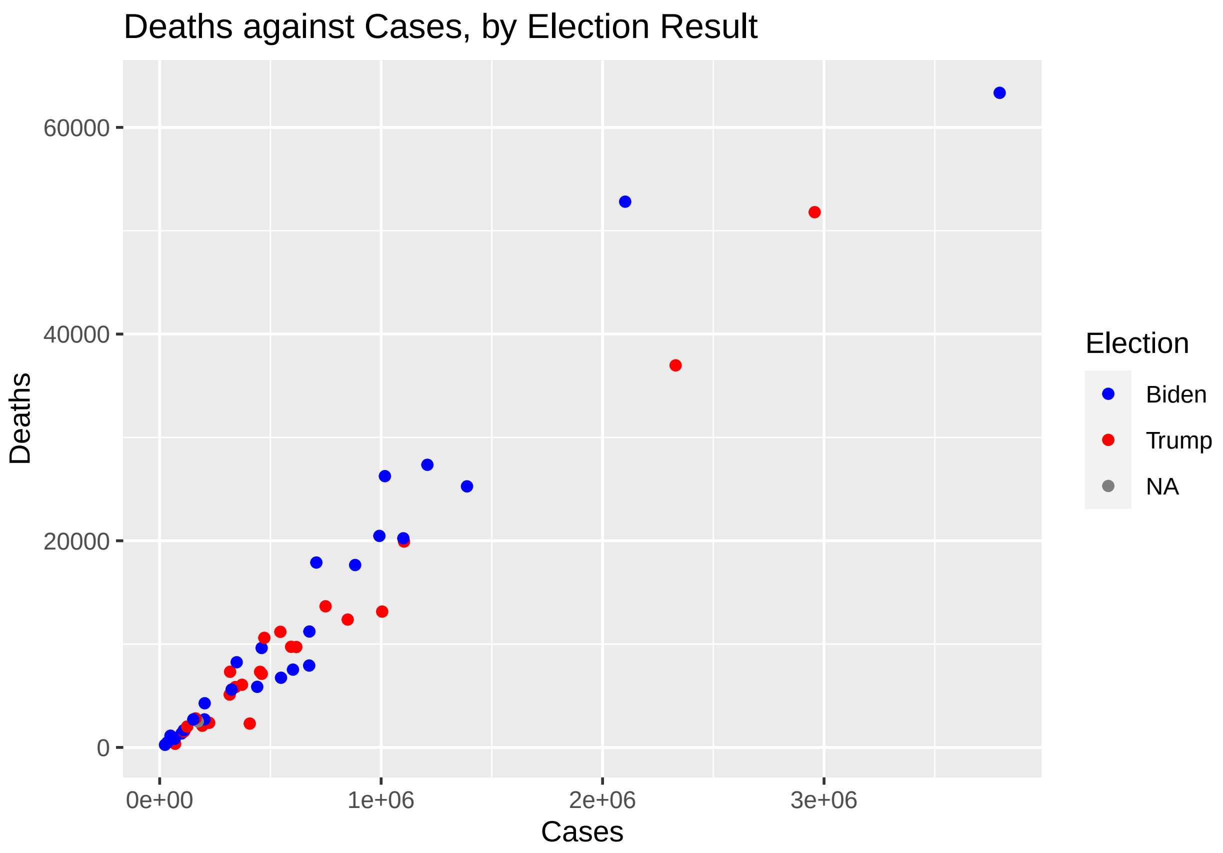
```
##       State  Cases Deaths state.rate Population Election
## 1    Alaska  69826    352       0.50     733391    Trump
## 2      Utah 406895   2308       0.57    3271616    Trump
## 3   Vermont  24240    255       1.05     643077    Biden
## 4  Nebraska 223517   2385       1.07    1961504    Trump
## 5     Idaho 192704   2103       1.09    1839106    Trump
## 6 Wisconsin 675152   7923       1.17    5893718    Biden
```

Note: I had saved the original file (at the end of HW1) sorted by death rate. If you sorted it differently that is fine. I didn't specify a specific sorting to save the file.

b)

Answers vary. A common visualization is deaths against cases, separate colors by vote for President.

```
##scatterplot
ggplot(state.data, aes(x=Cases,y=Deaths, color=Election))+
  geom_point()+
  labs(x="Cases", y="Deaths", title="Deaths against Cases, by Election Result")+
  scale_color_manual(values=c("blue", "red"))
```


Deaths against Cases, by Election Result
Deaths
60000
40000
20000
0
0e+00
1e+06
2e+06
3e+06
Cases
Election
Biden
Trump
NA