uwa6xv_M02_HW

Alanna Hazlett

2024-02-06

Problem 1

Import Police Killings csv

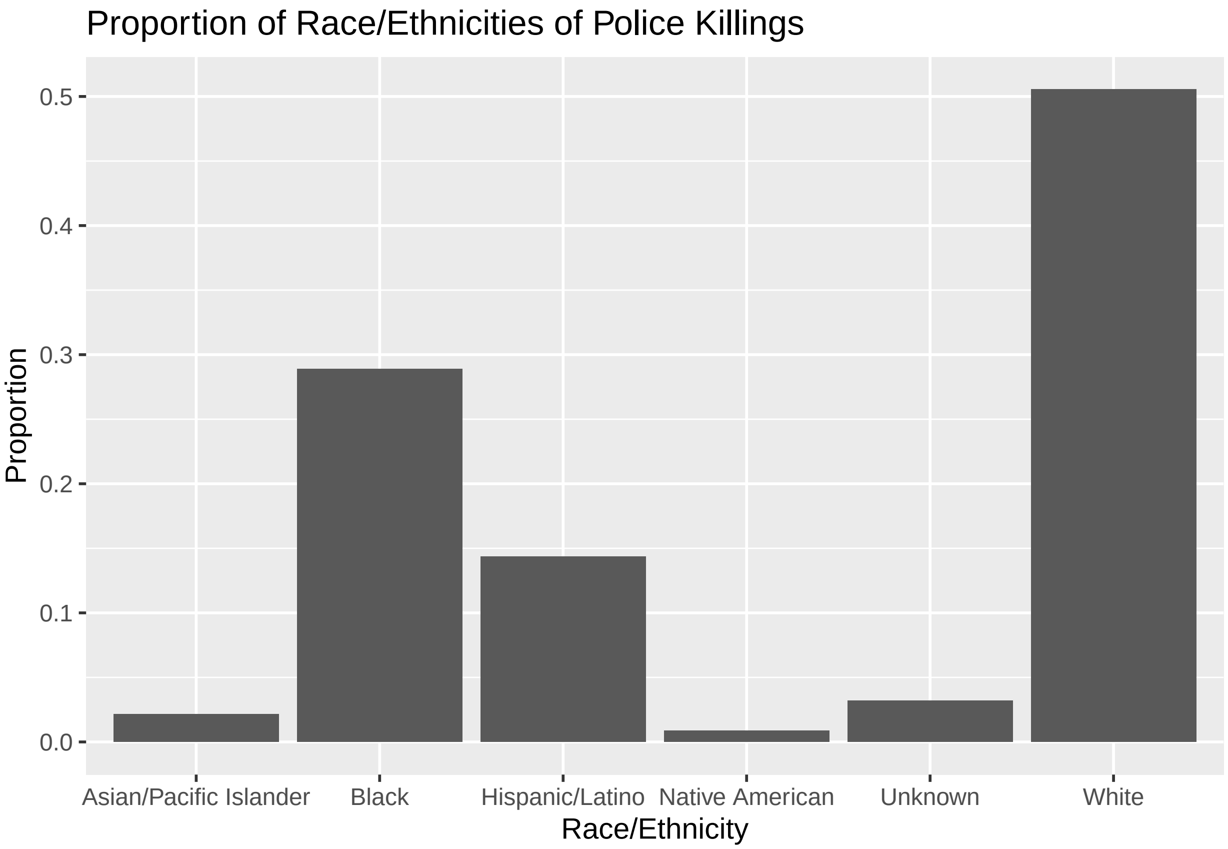
```
PoliceKillings<-read.csv("PoliceKillings.csv")
view(PoliceKillings)
```

(a) 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 2022 estimates to explain what your data reveal.

```
RETable<-table(PoliceKillings$raceethnicity)
REPropTable<-prop.table(RETable)
head(REPropTable)
```

```
##
## Asian/Pacific Islander                  Black        Hispanic/Latino
##             0.02141328             0.28907923             0.14346895
##        Native American                Unknown                  White
##             0.00856531             0.03211991             0.50535332
```

```
PoliceKillings %>%
  group_by(raceethnicity) %>%
  summarize(counts=n()) %>%
  mutate(percent=counts/nrow(PoliceKillings)) %>%
    ggplot(aes(x=raceethnicity,y=percent))+
      geom_bar(stat="identity")+
      labs(x="Race/Ethnicity",y="Proportion",title="Proportion of Race/Ethnicities of Police Killings")
```



* Census states that Whites are 75.5%,Black or African American are 13.6%, American Indian or Alaska Native are 1.3%, Asian are 6.3%, Native Hawaiian or Pacific Islander are 0.3%, and Hispanic are 19.1%.
* For Asian or Pacific Islander there is an increase of 0.84% in our data compared to the census. For people identifying as Black there is an increase of 15.3% compared to the census. For Hispanic/Latino there is a decrease in our data of 4.75% compared to the census. For Native American there is a decrease in our data of 0.44% compared to the Census. For people identifying as White there is about 20% decrease in our data compared to the census. Unknown is not a category for the Census. The White groups has the biggest decrease in our data compared to the Census. The Black groups had the largest increase in our data compared to the Census.

(b)

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.
Received help from Skye on mutate(age.num=as.numeric(as.character(age))).

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

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

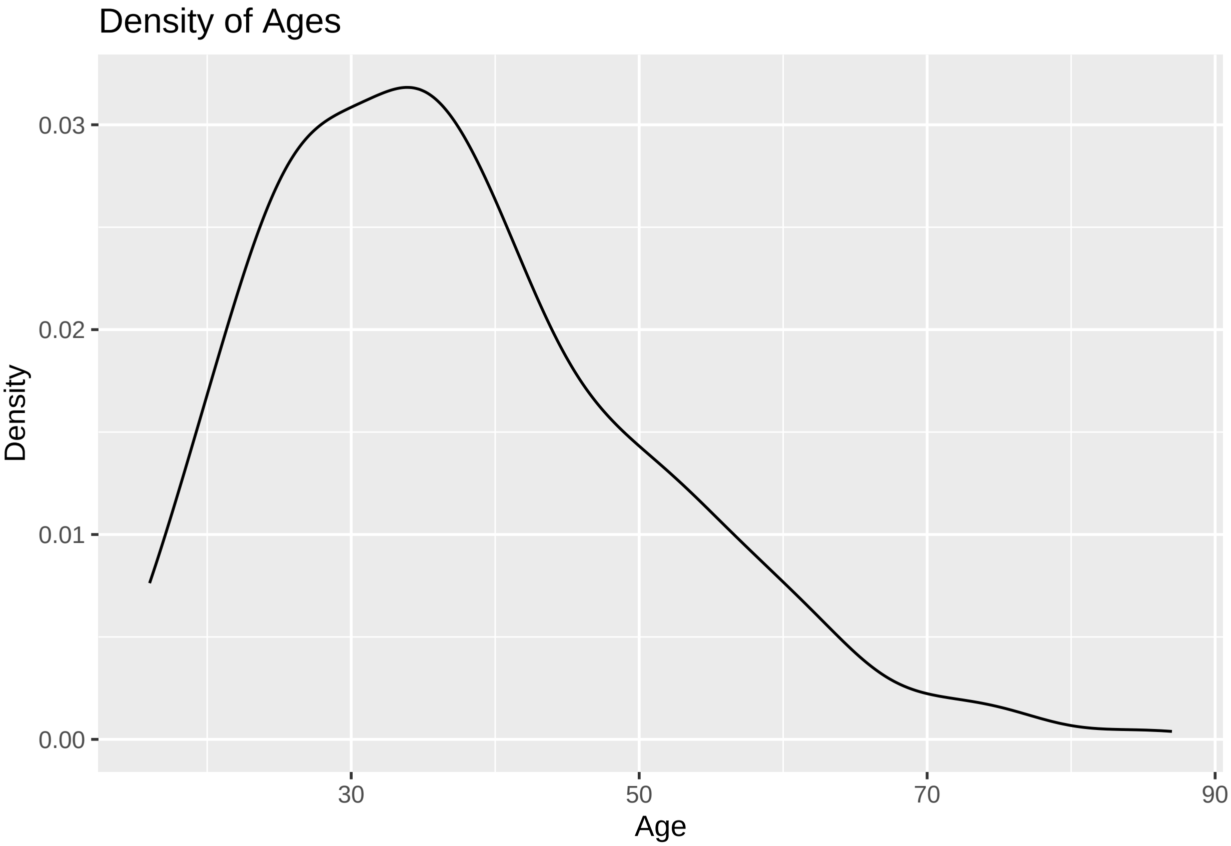
```
is.numeric(PoliceKillings$age.num)
```

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

(c)
Create a density plot of the variable age.num. Comment on this density plot.
Received help from Skye on correcting x=age to age.num

```r
ggplot(PoliceKillings,aes(x=age.num))+
  geom_density()+
  labs(x="Age",y="Density",title="Density of Ages")
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_density()`).
```

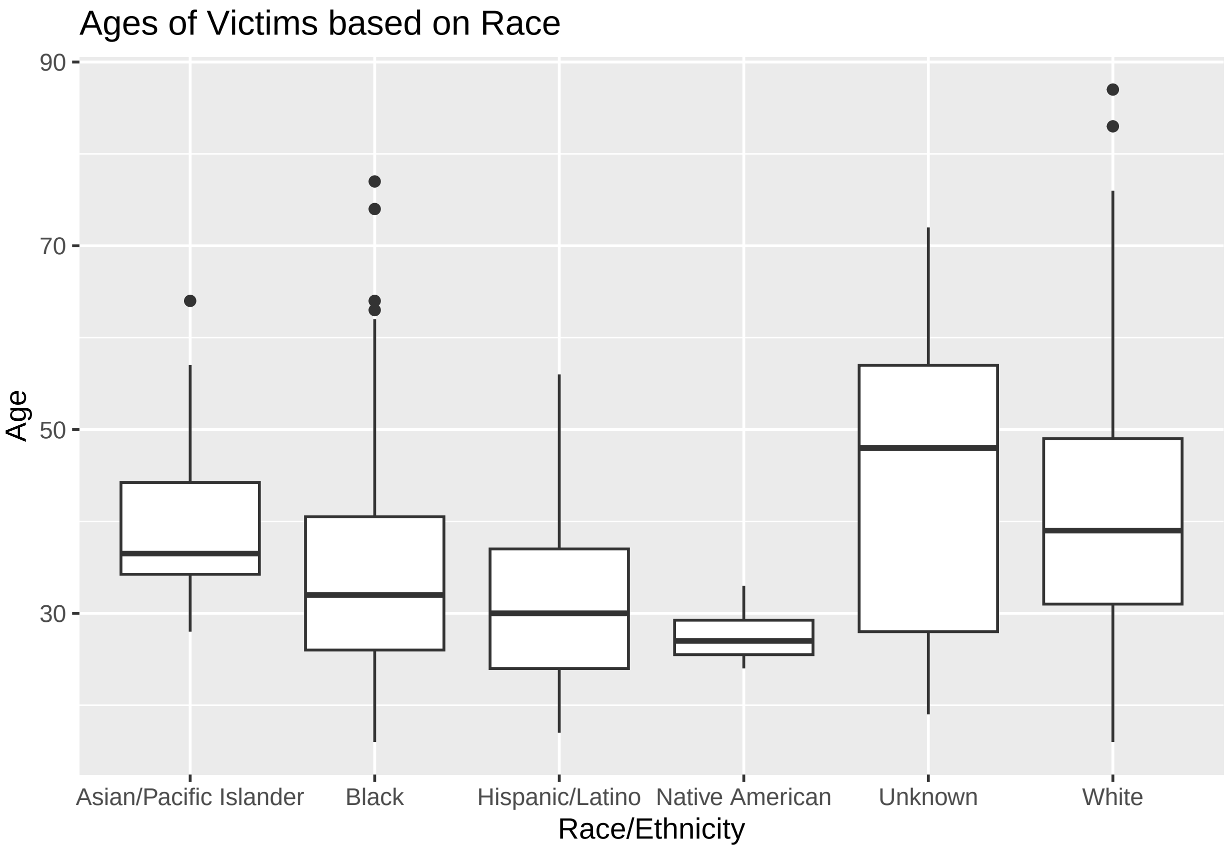


The data ranges from about 18 years old to 86 years old. The mode of the data is around 35 years old. There is a sharp increase from 18 to 35 and then a sharp decrease from 35 to about 45 and then a more gradual taper from 45 to about 86.

(d)
Create a visualization to compare the ages of victims across the different race / ethnicity levels. Comment on the visualization.
Received help from Skye on utilizing boxplot instead of trying to use bar graph.

```r
ggplot(PoliceKillings,aes(x=raceethnicity,y=age.num))+
  geom_boxplot()+
  labs(x="Race/Ethnicity",y="Age",title="Ages of Victims based on Race")
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_boxplot()`).
```



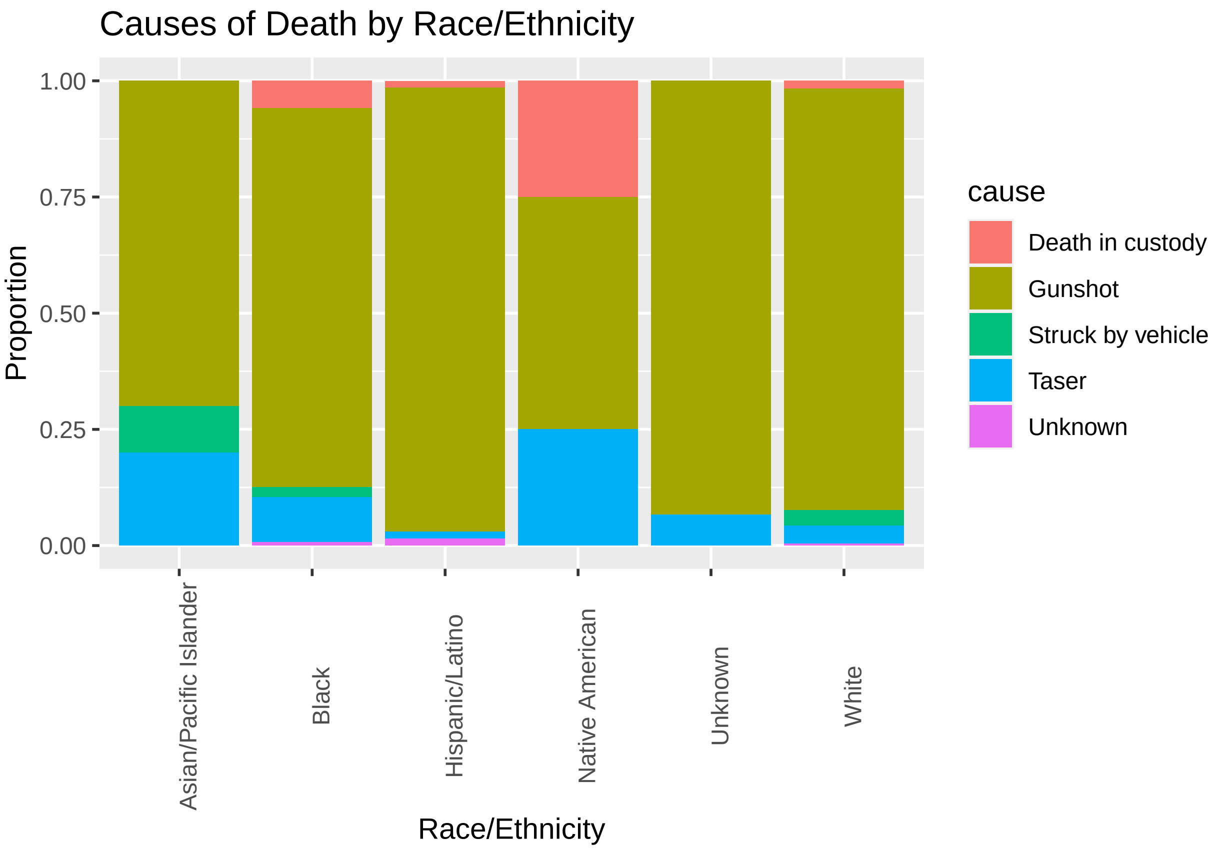
In ascending order, our median age of victims by race are Native American, Hispanic/Latino, Black, Asian/Pacific Islander, White, and Unknown. For our known Race/Ethnicities Native American has the smallest distribution of ages and White has the largest distribution of ages.

(e)
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.

```
deathstable<-table(PoliceKillings$raceethnicity,PoliceKillings$cause)
round(prop.table(deathstable,2),2)
```

```
##
##                         Death in custody Gunshot Struck by vehicle Taser
##   Asian/Pacific Islander            0.00    0.02              0.08  0.07
##   Black                             0.57    0.27              0.25  0.48
##   Hispanic/Latino                   0.07    0.16              0.00  0.04
##   Native American                   0.07    0.00              0.00  0.04
##   Unknown                           0.00    0.03              0.00  0.04
##   White                             0.29    0.52              0.67  0.33
##
##                         Unknown
##   Asian/Pacific Islander    0.00
##   Black                     0.33
##   Hispanic/Latino           0.33
##   Native American           0.00
##   Unknown                   0.00
##   White                     0.33
```

```r
ggplot(PoliceKillings,aes(x=raceethnicity,fill=cause))+
  geom_bar(position="fill")+
  labs(x="Race/Ethnicity",y="Proportion",title="Causes of Death by Race/Ethnicity")+
  theme(axis.text.x=element_text(angle = 90))
```



The causes of death are relatively similar among the groups Black, Hispanic/Latino, Unknown, and White. These all show a majority of casuses of deaths to be by gunshot, with some causes by taser and some death in custody. The Hispanic/Latino group, White group, and Black group do show unknown cause of death. The White group and Black group do show some deaths by stuck by vehicle.

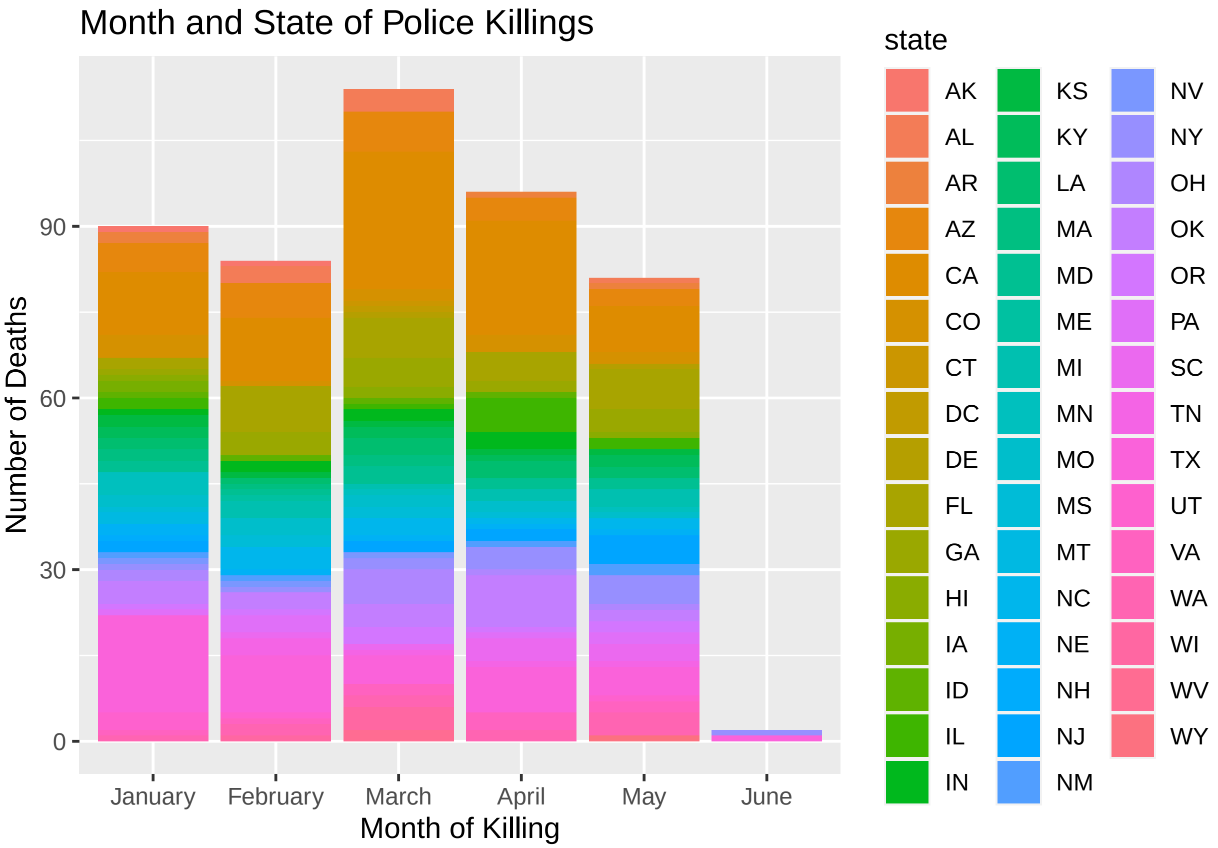
There are more significant differences in the Native American group and the Asian/Pacific Islander group. They have a smaller proportion of deaths by gunshot. The Asian/Pacific Islander group has a larger taser and struck by vehicle proportions. The Native American group has larger taser and death in custody proportions.

Among Black, Hispanic/Latino, Unknown, and White groups the causes of death are relatively the same and I would say independent of their race/ethnicity. Among the Asian/Pacific Islander and Native American groups the causes of death vary more significantly and I would say the causes of deaths are not independent of their race/ethnicity.

(f)

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.
Received help from Kimberly Gonzales on addition of order before factor.

```r
PoliceKillings2 <- PoliceKillings[order(PoliceKillings$month), ]
PoliceKillings2$month<-factor(PoliceKillings2$month, levels=c("January","February", "March","April","Ma
ggplot(PoliceKillings2,aes(x=month,fill=state))+
  geom_bar(position="stack",stat="count")+
  labs(x="Month of Killing",y="Number of Deaths",title="Month and State of Police Killings")
```



What I wanted to know: Is there a state that clearly has a larger number of police killings than other states? The distribution of police killings among states seem to be fairly even. There are not any states that seem to have significantly more or fewer police killings than the other states.
Is there a month where police killings seem higher than others or lower than others?
There seems to be about the same number of police killings in the months of January, February, March, and April. There is a drastically lower number of police killings in June. There is a significantly higher number of police killings in March.

Problem 2

Import stateCovid.csv. If extra column, remove.

```
stateCovid<-read.csv("stateCovid.csv")
stateCovid<-stateCovid[,-c(1)]
```

(a)

Merge these two datasets, stateCovid.csv and State_pop_election.csv. Use the head() function to display the first 6 rows after merging these two datasets.

```
election<-read.csv("State_pop_election.csv")
# remove non-matching rows: Guam, Virgin Islands, Northern Marina Islands
stateCovid2<-stateCovid[-c(12,37,50),]
electionCovid<-data.frame(election,stateCovid2)
```

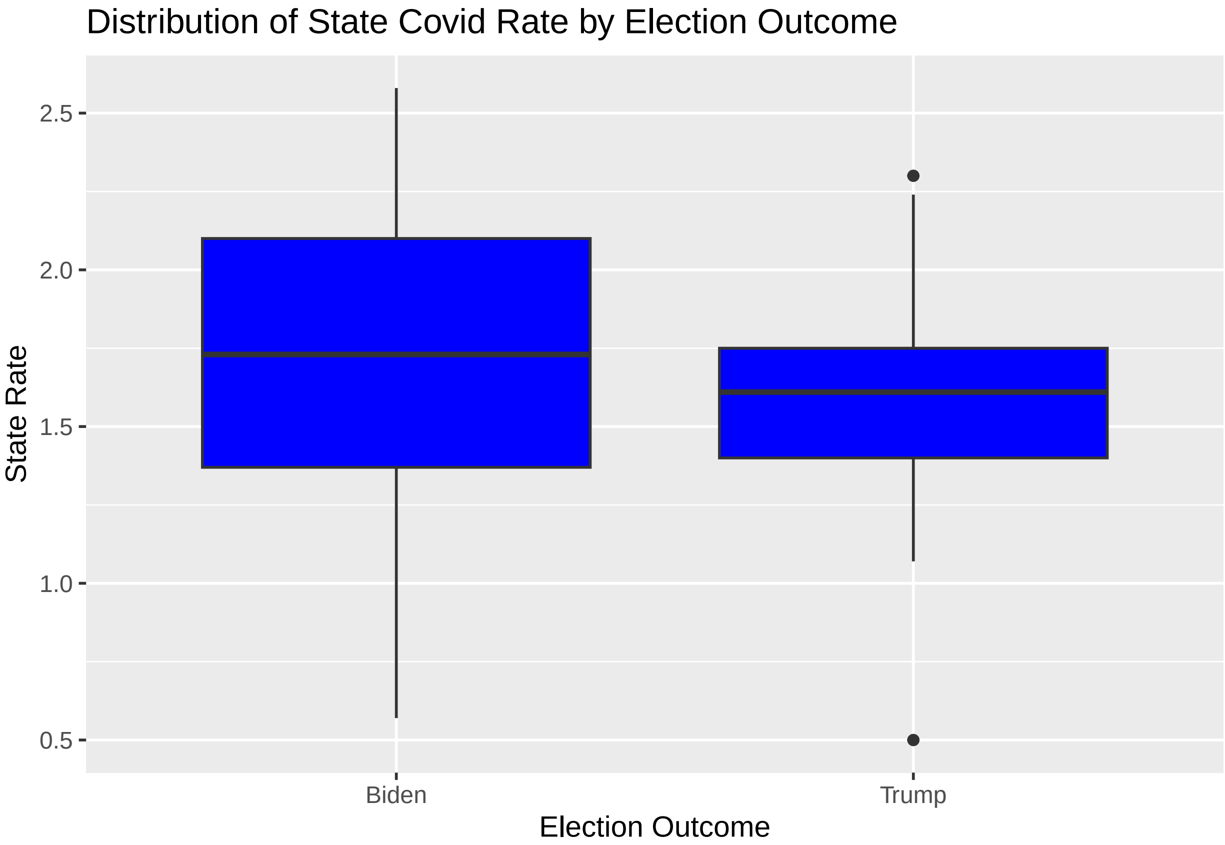
```
electionCovid<-electionCovid[,-c(4)]
head(electionCovid)
```

```
##        State Population Election   cases deaths state.rate
## 1    Alabama    5024279    Trump  545028  11188       2.05
## 2     Alaska     733391    Trump   69826    352       0.50
## 3    Arizona    7151502    Biden  882691  17653       2.00
## 4   Arkansas    3011524    Trump  341889   5842       1.71
## 5 California   39538223    Biden 3793055  63345       1.67
## 6   Colorado    5773714    Biden  547961   6746       1.23
```

(b)

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.

```
# I want to see the proportion of covid deaths (state rate) compared to election outcome (Trump or Bide
electionCovid_nomiss<-na.omit(electionCovid)
ggplot(electionCovid_nomiss,aes(x=Election, y=state.rate))+
  geom_boxplot(fill="Blue")+
  labs(x="Election Outcome", y="State Rate", title="Distribution of State Covid Rate by Election Outcom
```



For the states that had elected Biden in 2020 there is a wider distribution of state rates of Covid, particularly from the 25th to 75th percentile. State rates is defined as the percent of covid deaths per covid cases for each state. The median of the states that elected Biden is slightly higher than the median of the states that elected Trump.