Homework2

Dima Mikhaylov

9/11/2021

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Part 1: questions based on `PoliceKillings.csv`

```
PoliceKillings <- read.csv('PoliceKillings.csv', header=TRUE)
```

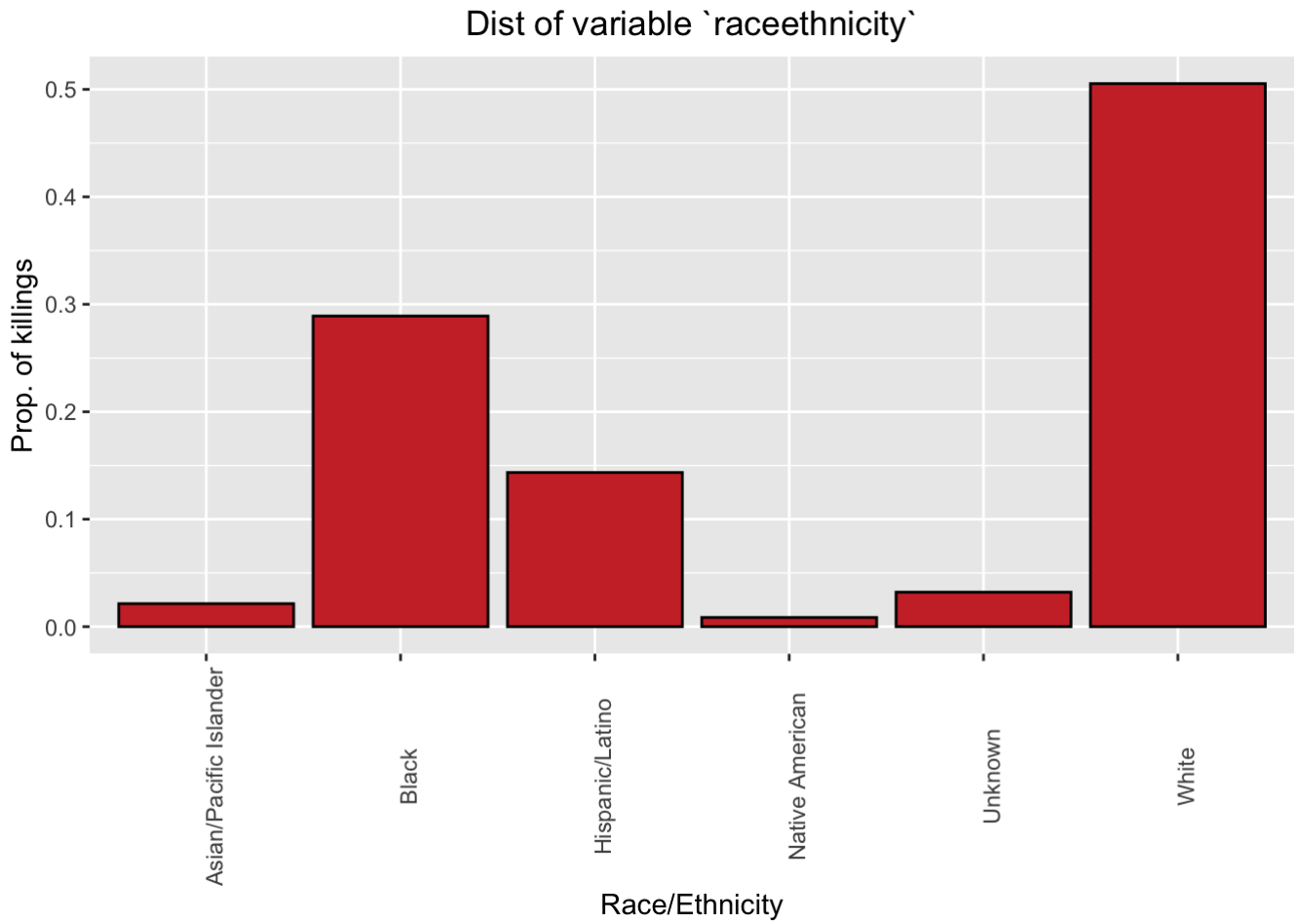
a. Using the `raceethnicity` variable, create a table and a bar chart that displays the proportions of victims in each race / ethnic level. Also, use your table and bar chart in conjunction with the US Census Bureau July 1 2019 estimates to explain what your data reveal.

```
round(prop.table(table(PoliceKillings$raceethnicity), ) *100, 2)
```

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

```
raceethnicity_prop <- PoliceKillings %>%
  group_by(raceethnicity)%>%
  summarise(Counts=n())%>%
  mutate(Percent=Counts/nrow(PoliceKillings))

ggplot(raceethnicity_prop, aes(x=raceethnicity, y=Percent))+
  geom_bar(fill="brown3", color='black', stat="identity")+
  theme(axis.text.x=element_text(angle=90),
        plot.title=element_text(hjust=0.5))+
  labs(x="Race/Ethnicity", y="Prop. of killings", title="Dist of variable `raceethnicity
`")
```



Observation: it seems noteworthy that Black people, although account for only 13% of the US population, have a relatively high proportion of almost 30%. In contrast, White people account for 76% of the population but only 50% proportion of the killings.

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.

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

```
## [1] FALSE
```

```
PoliceKillings$age.num <- as.numeric(PoliceKillings$age)
```

```
## Warning: NAs introduced by coercion
```

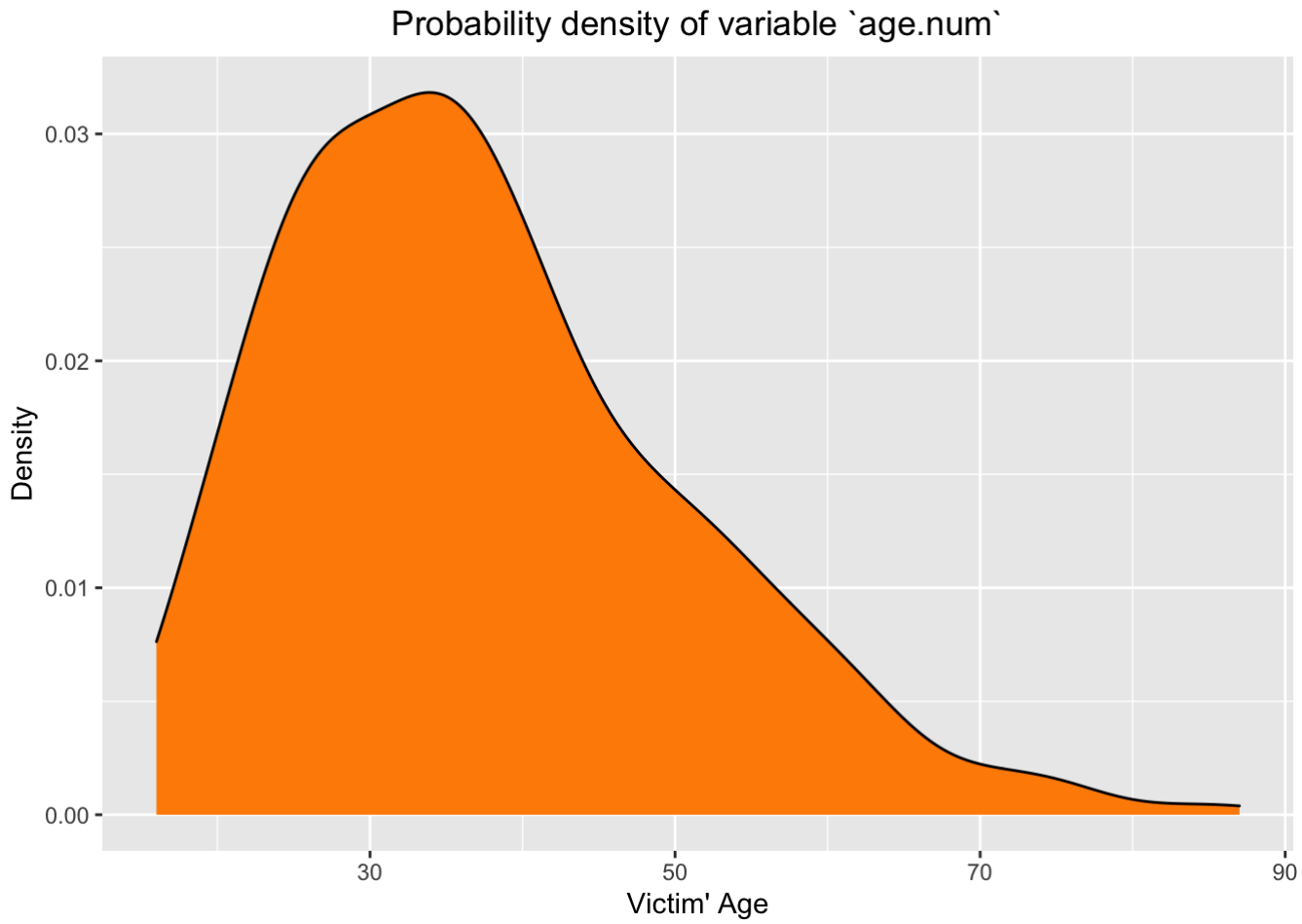
```
is.numeric(PoliceKillings$age.num)
```

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

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

```
ggplot(PoliceKillings, aes(x=age.num))+
  geom_density(fill="darkorange", color='black')+
  theme(plot.title=element_text(hjust=0.5))+
  labs(x="Victim' Age", y="Density", title="Probability density of variable `age.num`")
```

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

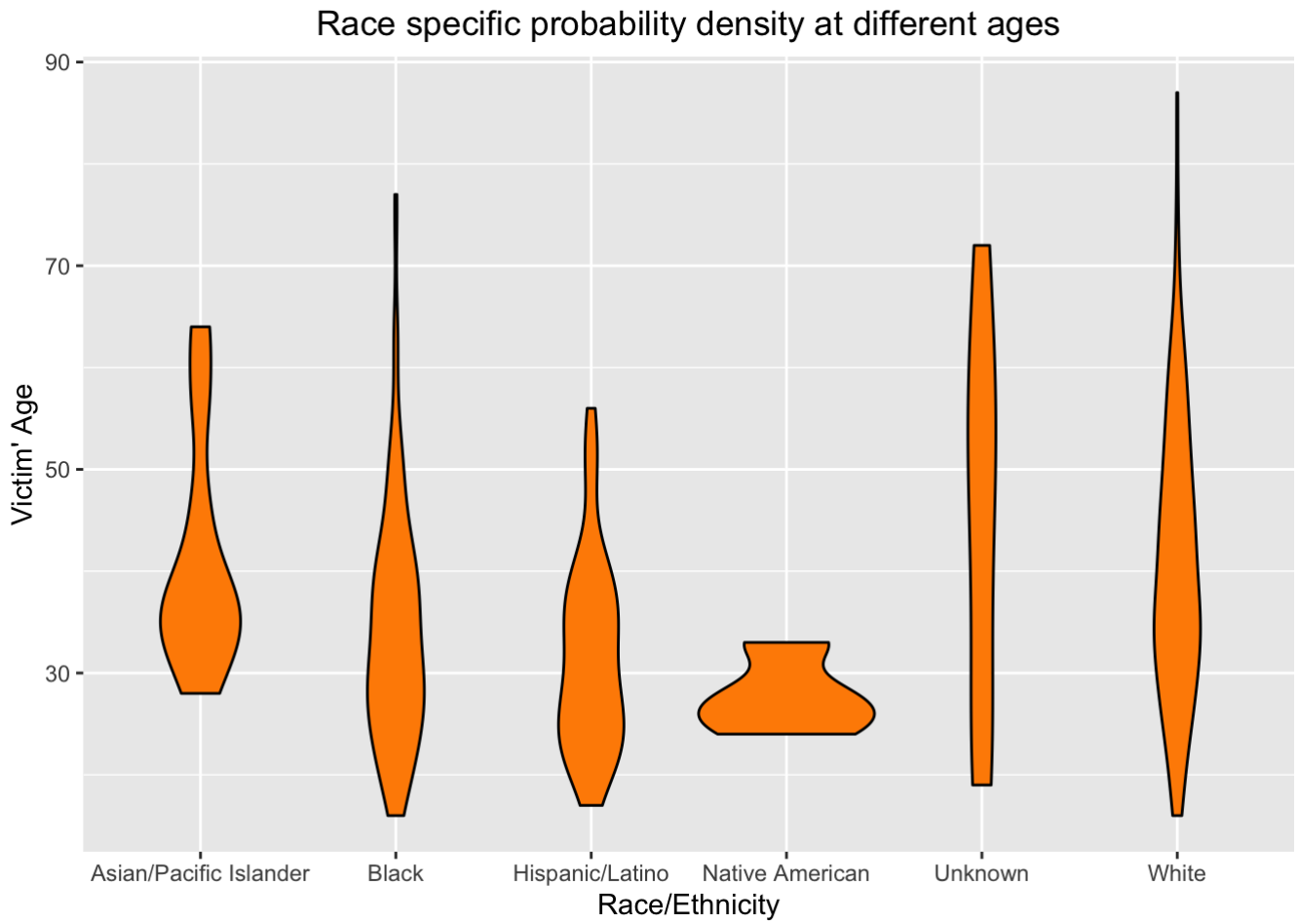


Observation: it seems that the density constantly increases for younger victims, picks at around 35 years and declines till the age of 65 years. The density function is almost flat around zero after 70 years.

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

```
ggplot(PoliceKillings, aes(x=raceethnicity, y=age.num))+
  geom_violin(fill="darkorange", color='black')+
  theme(plot.title=element_text(hjust=0.5))+
  labs(x="Race/Ethnicity", y="Victim' Age", title="Race specific probability density at
 different ages")
```

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



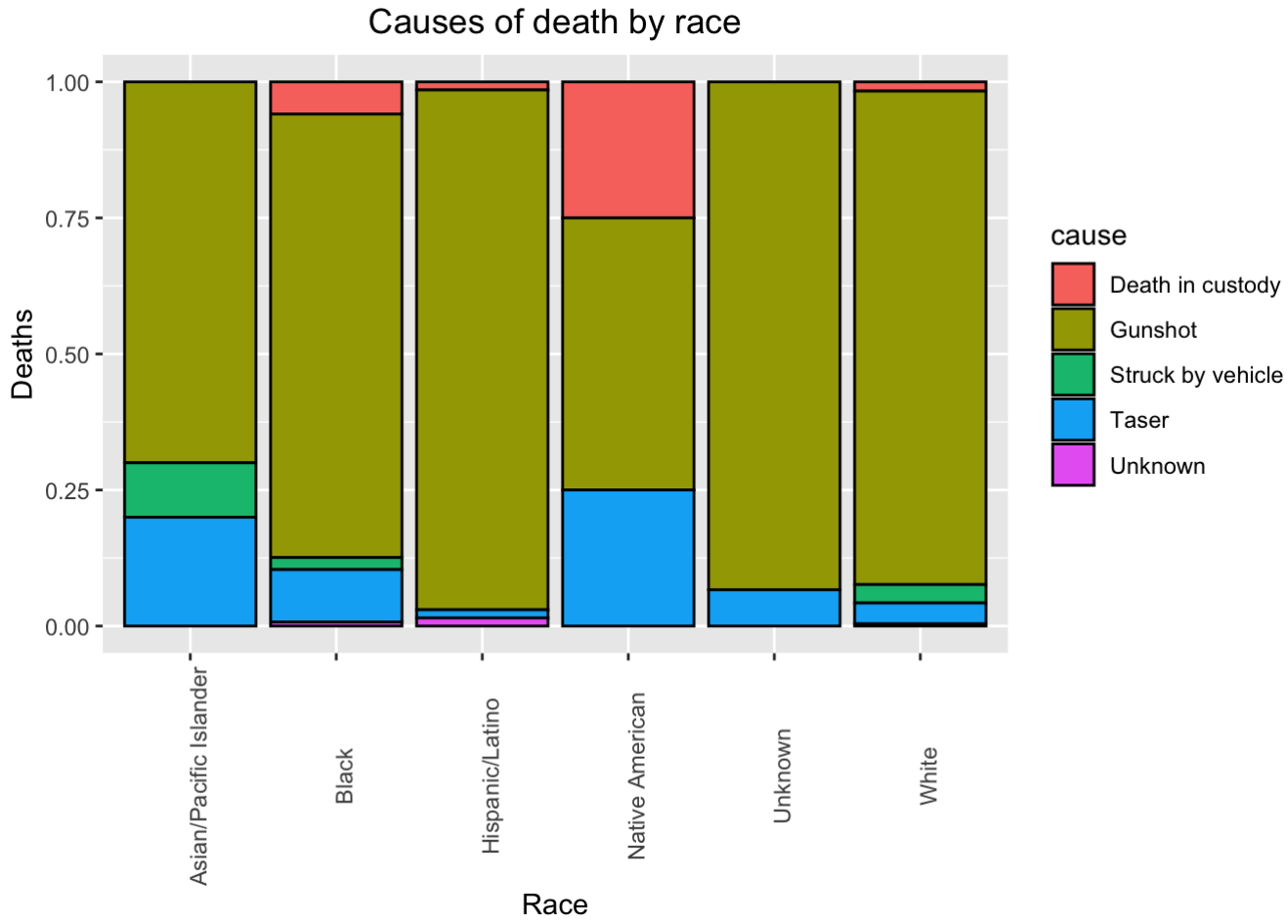
Observation: it seems that Asian/Pasific victims tend to be older. In constract, Black and Native American victims tend to be younger. Also, Native American victims appear to be extremely homogeneous in terms of age, possibly due to a very small number of observations

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

```
round(prop.table(table(PoliceKillings$cause))*100, 2)
```

```
##
##  Death in custody           Gunshot Struck by vehicle             Taser
##              3.00              88.01              2.57              5.78
##           Unknown
##              0.64
```

```
ggplot(PoliceKillings, aes(x=raceethnicity, fill=cause))+
  geom_bar(color='black', position="fill")+
  theme(axis.text.x=element_text(angle=90),
        plot.title=element_text(hjust=0.5))+
  labs(x="Race", y="Deaths", title="Causes of death by race")
```



Observation: from this plot above it does not appear that the cause of death is dependent of the victim's race / ethnicity. Another stronger conclusion is probably that all groups suffer from gunshot injuries greatly. Also, surprisingly, there are a number of deaths related to tasers, presumably a non-lethal weapon.

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 describe how you created the new variables.

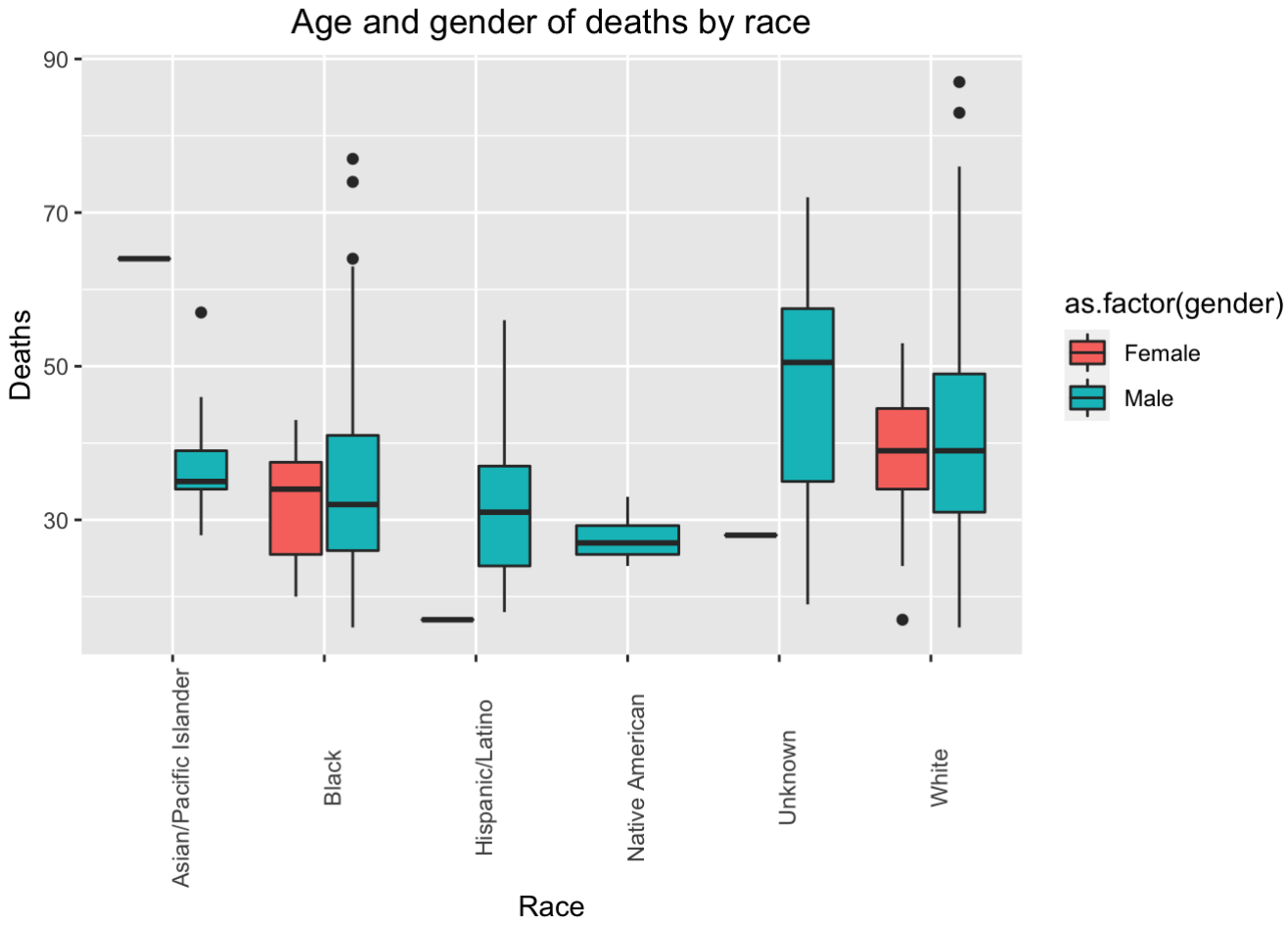
```
round(prop.table(table(PoliceKillings$raceethnicity, PoliceKillings$gender), ) *100, 2)
```

```
##
##                          Female  Male
##   Asian/Pacific Islander   0.21  1.93
##   Black                    1.50 27.41
##   Hispanic/Latino          0.21 14.13
##   Native American          0.00  0.86
##   Unknown                  0.43  2.78
##   White                    2.36 48.18
```

```
ggplot(PoliceKillings, aes(x=raceethnicity, y=age.num, fill=as.factor(gender)))+
  geom_boxplot(fcolor='black')+
  theme(axis.text.x=element_text(angle=90),
        plot.title=element_text(hjust=0.5))+
  labs(x="Race", y="Deaths", title="Age and gender of deaths by race")
```

```
## Warning: Ignoring unknown parameters: fcolour
```

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



Observation, when splitting by gender, first females (from the table above) constitute an absolute minority of the vistims, and second, their age distribution largely mimics males from the same race / ethnicity categories.

Part 2: questions based on `stateCovid.csv` and `State_pop_election.csv`

The dataset should contain 4 columns: • the name of the state (55 “states”, the 50 states, plus DC, Puerto Rico, Guam, Northern Mariana Islands, and the Virgin Islands) • the number of cases • the number of deaths • the death rate, defined as the number of deaths divided by the number of cases You may realize that when you exported the data file as a .csv file, an extra column was added to the dataframe. Remove this column.

```
Covid <- read.csv('stateCovid.csv', header=TRUE) #, row.names=1)
dim(Covid)
```

```
## [1] 55  4
```

```
head(Covid)
```

```
##        state   cases deaths state_rate
## 1    Alabama  545028  11188       2.05
## 2     Alaska   69826    352       0.50
## 3    Arizona  882691  17653       2.00
## 4   Arkansas  341889   5842       1.71
## 5 California 3793055  63345       1.67
## 6   Colorado  547961   6746       1.23
```

a. There is a dataset on Collab, called `State_pop_election.csv` . The data contain the population of the states from the 2020 census (50 states plus DC and Puerto Rico), as well as whether the state voted for Biden or Trump in the 2020 presidential elections. 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', header=TRUE)
Merged <- merge( Covid, Election, by.x="state", by.y = "State")
head(Merged)
```

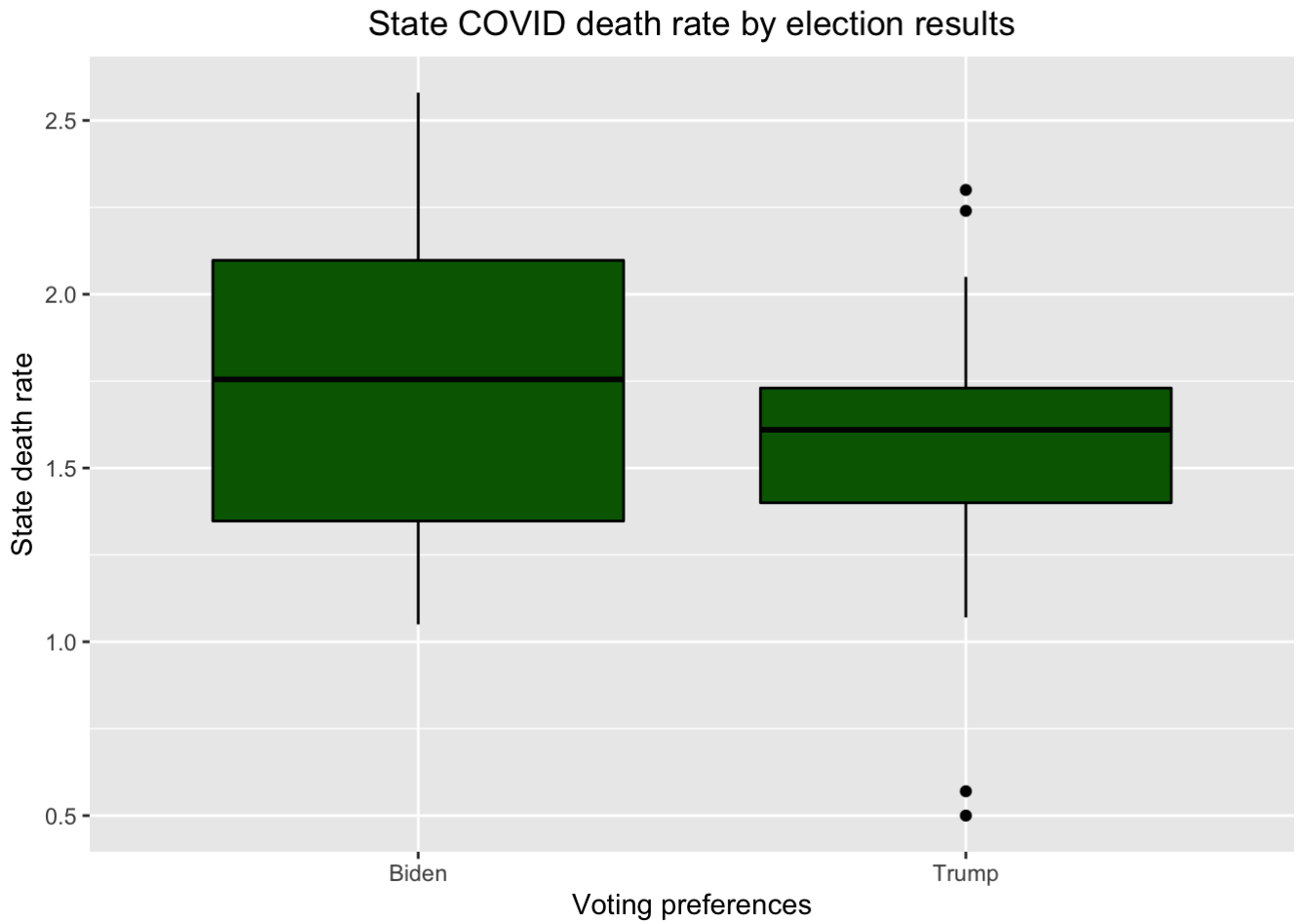
```
##        state   cases deaths state_rate Population Election
## 1    Alabama  545028  11188       2.05    5024279    Trump
## 2     Alaska   69826    352       0.50     733391    Trump
## 3    Arizona  882691  17653       2.00    7151502    Biden
## 4   Arkansas  341889   5842       1.71    3011524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

```
dim(Merged)
```

```
## [1] 52  6
```

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.

```
ggplot(Merged[!is.na(Merged$Election),], aes(x=Election, y=state_rate))+
  geom_boxplot(fill="darkgreen", color='black')+
  theme(plot.title=element_text(hjust=0.5))+
  labs(x="Voting preferences", y="State death rate", title="State COVID death rate by el
ection results")
```



Observation: as shown on the boxplot above, distribution of state death rate does vary by aggregate voting preferences, as measured by which candidate, Biden or Trump, prevailed in the last presidential elections. Noteworthy, that the death rate is somewhat higher and much more variant around the mean in sthe tates supporting Biden. Trump supporting states had lower average and much smaller variance, with some large outliers.