Stat 6021: HW 2

Tom Lever

09/08/22

1. For this question, we will work on the dataset `PoliceKillings.csv`. This dataset was the basis for this article on police killings in 2015. You may read more about the data and the variable descriptions here.

```
library(dplyr)
police_killings <- read.csv(file = 'PoliceKillings.csv') %>% select(-X)
head(police_killings, n = 1)
```

```
##                 name age gender raceethnicity    month day year streetaddress
## 1 A'donte Washington  16   Male         Black February  23 2015  Clearview Ln
##        city state latitude longitude state_fp county_fp tract_ce     geo_id
## 1 Millbrook    AL 32.52958 -86.36283        1        51    30902 1051030902
##   county_id            namelsad        lawenforcementagency   cause armed  pop
## 1      1051 Census Tract 309.02 Millbrook Police Department Gunshot    No 3779
##   share_white share_black share_hispanic p_income h_income county_income
## 1        60.5        30.5            5.6    28375    51367         54766
##   comp_income county_bucket nat_bucket  pov      urate   college
## 1   0.9379359             3          3 14.1 0.09768638 0.1685095
```

```
number_of_killings <- nrow(police_killings)
number_of_killings
```

```
## [1] 467
```

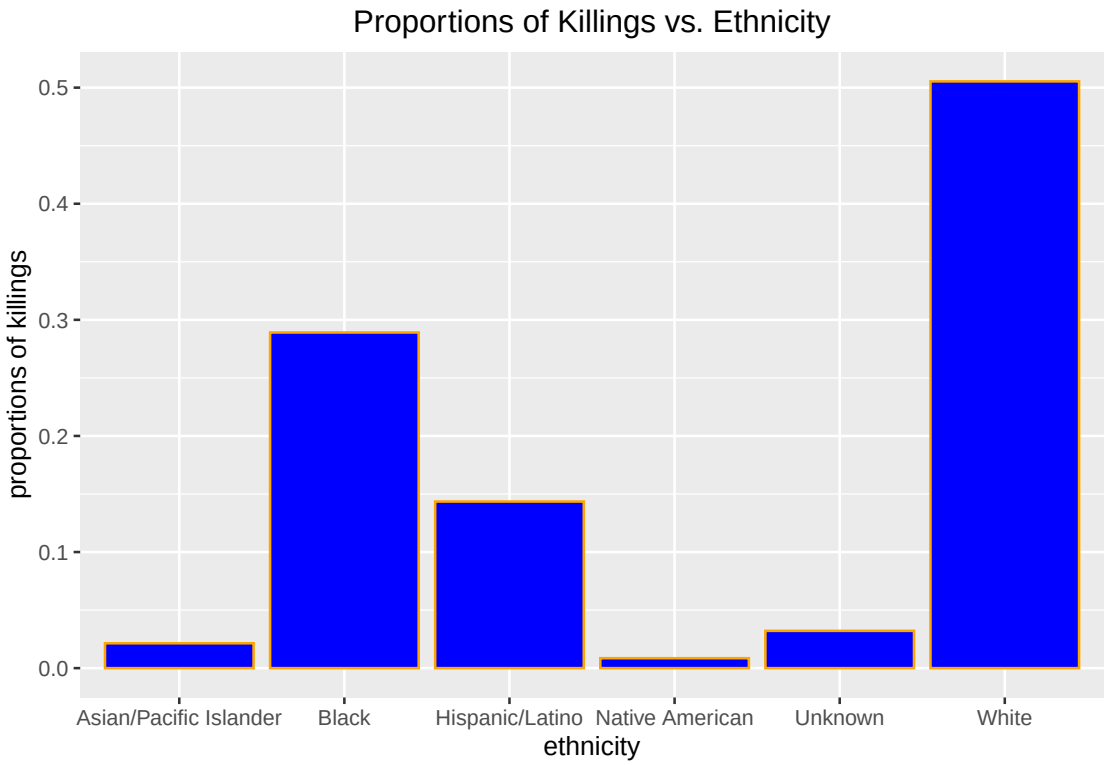
   (a) Using the `raceethnicity` variable, create a table and a bar chart that displays the proportions of victims in each race / ethnic level. Also, use your table and bar chart with the US Census Bureau July 1, 2021 estimates to explain what your data reveal.

```
round(prop.table(table(police_killings$raceethnicity)), 3)
```

```
##
## Asian/Pacific Islander                 Black        Hispanic/Latino
##                  0.021                 0.289                   0.143
##        Native American               Unknown                   White
##                  0.009                 0.032                   0.505
```

```
library(ggplot2)
ethnicity <-
    police_killings %>%
        select(raceethnicity) %>%
        rename(ethnicity = raceethnicity)
proportion <-
    ethnicity %>%
        group_by(ethnicity) %>%
        summarize(numbers_of_killings = n()) %>%
        mutate(
            proportions_of_killings = numbers_of_killings / number_of_killings
```

```
        )
ggplot(data = proportion, aes(x = ethnicity, y = proportions_of_killings)) +
    geom_bar(stat = "identity", fill = "blue", color = "orange") +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    ) +
    labs(
        x = "ethnicity",
        y = "proportions of killings",
        title = "Proportions of Killings vs. Ethnicity"
    )
```



According to "Population Estimates, July 1, 2021, (V2021)", 75.8 percent of people in all states and counties, and for cities and towns with a population of 5,000 or more, around July 1, 2021, were white. Let's assume that this subpopulation is representative of the population in the entire United States at any moment in 2015. Considering "FiveThirtyEight Police Killings Dataset" and "Proportions of Killings vs. Ethnicity", for a sample of 467 people killed by police between 01/01/2015 and 06/01/2015, only 50.5 percent were white. Let's assume that this sample is simple random. The number of white people in the sample is greater than 5; the number of non-white people in the sample is greater than 5. Let's test at a significance level 0.05 the claim and null hypothesis that the proportion of white people among people killed by police between 01/01/2015 and 06/01/2015 is greater than or equal to 75.8 percent.

```
library(TomLeversRPackage)
testResult <-
    testNullHypothesisInvolvingProportion(
        0.758,
        0.505,
```

```
        number_of_killings,
        ">=",
        0.05
    )
testResult
```

```
## Since probability 1.27808027844324e-37
## is less than significance level 0.05,
## we reject the null hypothesis.
## We have sufficient evidence to support the alternate hypothesis.
```

Since the above probability is far less than the significance level, we reject the claim and null hypothesis. We have sufficient evidence to conclude that the proportion of white people among people killed by police between 01/01/2015 and 06/01/2015 is less than 75.8 percent. There is a statistically significant difference between the proportion of the population in the entire United States at any moment in 2015 that were white and the proportion of white people among people killed by police between 01/01/2015 and 06/01/2015.

(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). Add this new variable to your data frame.

```
age.num <- as.numeric(
    police_killings %>%
        select(age) %>%
        mutate(age = replace(age, age == "Unknown", "-1")) %>%
        pull(age)
)
age.num[1:24]
```

```
##  [1] 16 27 26 25 29 29 22 35 44 31 76 40 -1 31 23 39 25 54 24 57 21 42 21 36
```

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

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

```
police_killings <- police_killings %>% bind_cols(data.frame(age.num))
head(police_killings, n = 1)
```

```
##                 name age gender raceethnicity    month day year streetaddress
## 1 A'donte Washington  16   Male         Black February  23 2015  Clearview Ln
##        city state latitude longitude state_fp county_fp tract_ce     geo_id
## 1 Millbrook    AL 32.52958 -86.36283        1        51    30902 1051030902
##   county_id            namelsad        lawenforcementagency   cause armed  pop
## 1      1051 Census Tract 309.02 Millbrook Police Department Gunshot    No 3779
##   share_white share_black share_hispanic p_income h_income county_income
## 1        60.5        30.5            5.6    28375    51367         54766
##   comp_income county_bucket nat_bucket  pov      urate   college age.num
## 1   0.9379359             3          3 14.1 0.09768638 0.1685095      16
```

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

```
age <- police_killings %>% select(age.num) %>% rename(age = age.num)
ggplot(age, aes(x = age)) +
    geom_density() +
    labs(title = "Probability Density of Killings vs. Age") +
    theme(
        plot.title = element_text(hjust = 0.5),
```

```
        axis.text.x = element_text(angle = 0)
    )
```


```
age <- age %>% pull(age)
shapiro.test(age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  age
## W = 0.96986, p-value = 3.263e-08
```

```
library(moments)
skewness(age)
```

```
## [1] 0.5842849
```

```
library(TomLeversRPackage)
calculateMode(age)
```

```
## [1] 29
```

```
median(age)
```

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

```
mean(age)
```

```
## [1] 37.03854
```

```
quantile(age, 0.25)
```

```
## 25%
```

```
##  27
```

```
quantile(age, 0.75)
```

```
## 75%
##  45
```

```
sd(age)
```

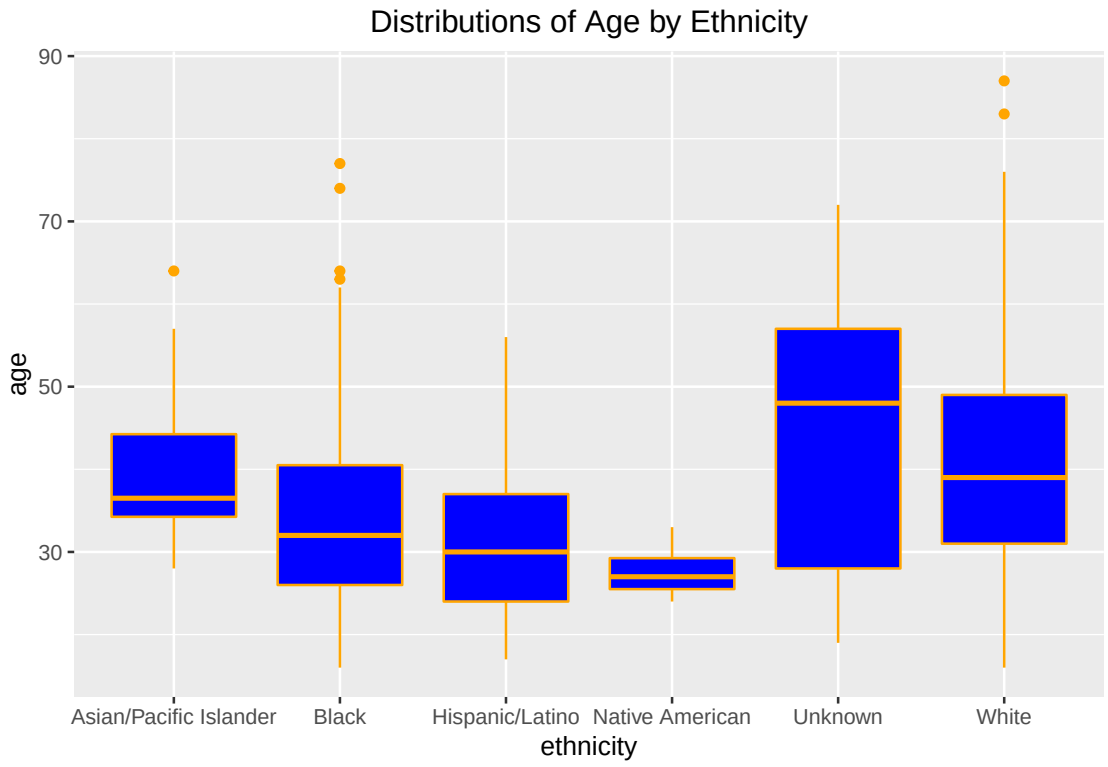
```
## [1] 13.41651
```

By the Shapiro-Wilk Test for Normality with null hypothesis that "Probability Density of Killings vs. Age" is normal, since the probability 3.263e-08 is less than a significance level of 0.05, we reject the null hypothesis and conclude that the distribution is not normal. The distribution is skewed to the right. The area under the distribution is 1. An area under the distribution represents a probability that a person killed has an age within the relevant age range.

The mode, median, mean, first quartile, and third quartile ages are 29, 35, 37, 27, and 45. The standard deviation of the distribution is 13.417 years. 50 percent of ages lie between 27 and 45.

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

```
age_and_ethnicity <-
    police_killings %>%
        select(raceethnicity, age.num) %>%
        rename(ethnicity = raceethnicity, age = age.num) %>%
        filter(age != -1)
ggplot(age_and_ethnicity, aes(x = ethnicity, y = age)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(title = "Distributions of Age by Ethnicity") +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
sorted_ages <-
    age_and_ethnicity %>%
    filter(ethnicity == "White") %>%
    select(age) %>%
    arrange(desc(age))
head(sorted_ages, n = 3)
```

```
##   age
## 1  87
## 2  83
## 3  76
```

```
min(age_and_ethnicity %>% filter(ethnicity == "Black") %>% pull(age))
```

```
## [1] 16
```

```
min(age_and_ethnicity %>% filter(ethnicity == "White") %>% pull(age))
```

```
## [1] 16
```

```
median(
    age_and_ethnicity %>% filter(ethnicity == "Native American") %>% pull(age)
)
```

```
## [1] 27
```

```
median(age_and_ethnicity %>% filter(ethnicity == "Unknown") %>% pull(age))
```

```
## [1] 48
```

```
median(age_and_ethnicity %>% filter(ethnicity == "White") %>% pull(age))
```

```
## [1] 39
```

```
table(police_killings$raceethnicity)
```

```
##
## Asian/Pacific Islander                   Black        Hispanic/Latino
##                     10                     135                      67
##         Native American                 Unknown                   White
##                      4                      15                     236
```

Including outliers, the oldest person killed was 87 and white. Excluding outliers, the oldest person killed was white and 76. The youngest people were black or white and 16. Including people with unknown ethnicities, all median ages fall between 27 and 48. Excluding people with unknown ethnicities, all median ages fall between 27 and 39, and the median, the third-quartile, and most first-quartile ages for non-white people are less than the median age for white people. Excluding people with unknown ethnicities, interquartile range falls with number of people killed; perhaps all interquartile ranges trend toward the same value. Subsample sizes are small for Native American, Asian / Pacific-Islander, and unknown-ethnicity people.

(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 on the victim's race / ethnicity.

```
ethnicity_and_cause_of_death <-
    police_killings %>%
        select(raceethnicity, cause) %>%
        rename(ethnicity = raceethnicity, cause_of_death = cause)
contingency_table <- table(
    ethnicity_and_cause_of_death$ethnicity,
    ethnicity_and_cause_of_death$cause_of_death
)
contingency_table
```

```
##
##                           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
```

```
chisq.test(contingency_table)
```

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

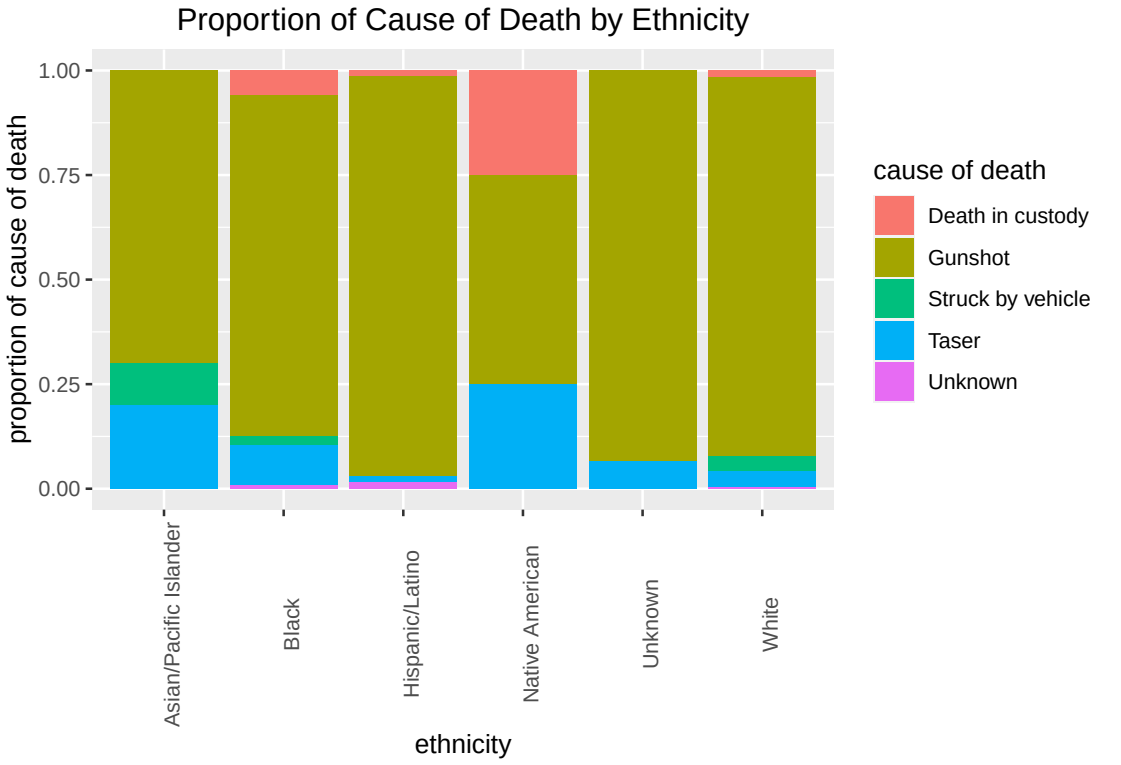
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
```

```
## X-squared = 34.679, df = 20, p-value = 0.02188
```

```
round(prop.table(contingency_table, 1) * 100, 1)
```

```
##
##                           Death in custody Gunshot Struck by vehicle Taser
##   Asian/Pacific Islander              0.0    70.0               10.0  20.0
##   Black                               5.9    81.5                2.2   9.6
##   Hispanic/Latino                     1.5    95.5                0.0   1.5
##   Native American                    25.0    50.0                0.0  25.0
##   Unknown                             0.0    93.3                0.0   6.7
##   White                               1.7    90.7                3.4   3.8
##
##                           Unknown
##   Asian/Pacific Islander     0.0
##   Black                      0.7
##   Hispanic/Latino            1.5
##   Native American            0.0
##   Unknown                    0.0
##   White                      0.4
```

```
ggplot(ethnicity_and_cause_of_death, aes(x = ethnicity, fill = cause_of_death)) +
    geom_bar(position = "fill") +
    scale_fill_discrete(name = "cause of death") +
    labs(
        y = "proportion of cause of death",
        title = "Proportion of Cause of Death by Ethnicity"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90)
    )
```




```
table(ethnicity_and_cause_of_death$ethnicity)
```

```
##
## Asian/Pacific Islander                 Black        Hispanic/Latino
##                     10                   135                    67
##        Native American               Unknown                 White
##                      4                    15                   236
```

Given a significance level 0.05, since the above probability 0.022 is less than the significance level, we reject the null hypothesis of the Pearson's Chi-squared test of independence, which states that there is no association between the row variable `ethnicity` and the column variable `cause_of_death`. We have sufficient evidence to conclude that there is an association between the row variable `ethnicity` and the column variable `cause_of_death`. We note that the "Chi-squared approximation may be incorrect".

People of all ethnicities were killed by gunshot and by taser. The majority of people were killed by gunshot, with the majority being vast and decreasing for Latinx, white, unknown-ethnicity, and black people, and smaller for Asian / Pacific-Islander and Native-American people. The second most significant cause of death was tasering, with the proportion decreasing among Native-American, Asian / Pacific-Islander, black, unknown-ethnicity, white, and Latinx people. Asian / Pacific-Islander and unknown-ethnicity people did not die in custody. The third most significant cause of death was custody, with the proportion decreasing among Native American, black, and Latinx and white people. Only white, black, and Asian / Pacific-Islander people were struck by a vehicle. The fourth most significant cause of death was being struck by a vehicle, with the proportion decreasing among Asian / Pacific-Islander, white, and black people. Zero or very small proportions of Latinx and white people were killed by taser, custody, or being struck by a vehicle. Significant proportions of Native American and Asian / Pacific-Islander, black, and unknown-ethnicity people were killed by taser. Significant proportions of Native American and black people died in custody. Significant proportions of Asian / Pacific-Islander, white, and black people were struck by a vehicle. Subsample sizes are small for Native American, Asian /

Pacific-Islander, and unknown-ethnicity people.

(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.

```
colnames(police_killings)
```

```
##  [1] "name"                 "age"                  "gender"
##  [4] "raceethnicity"        "month"                "day"
##  [7] "year"                 "streetaddress"        "city"
## [10] "state"                "latitude"             "longitude"
## [13] "state_fp"             "county_fp"            "tract_ce"
## [16] "geo_id"               "county_id"            "namelsad"
## [19] "lawenforcementagency" "cause"                "armed"
## [22] "pop"                  "share_white"          "share_black"
## [25] "share_hispanic"       "p_income"             "h_income"
## [28] "county_income"        "comp_income"          "county_bucket"
## [31] "nat_bucket"           "pov"                  "urate"
## [34] "college"              "age.num"
```

```
gender_and_cause_of_death <- police_killings %>% select(gender, cause)
contingency_table <- table(
    gender_and_cause_of_death$gender,
    gender_and_cause_of_death$cause
)
contingency_table
```

```
##
##          Death in custody Gunshot Struck by vehicle Taser Unknown
##   Female                0      19                 2     1       0
##   Male                 14     392                10    26       3
```

```
chisq.test(contingency_table)
```

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

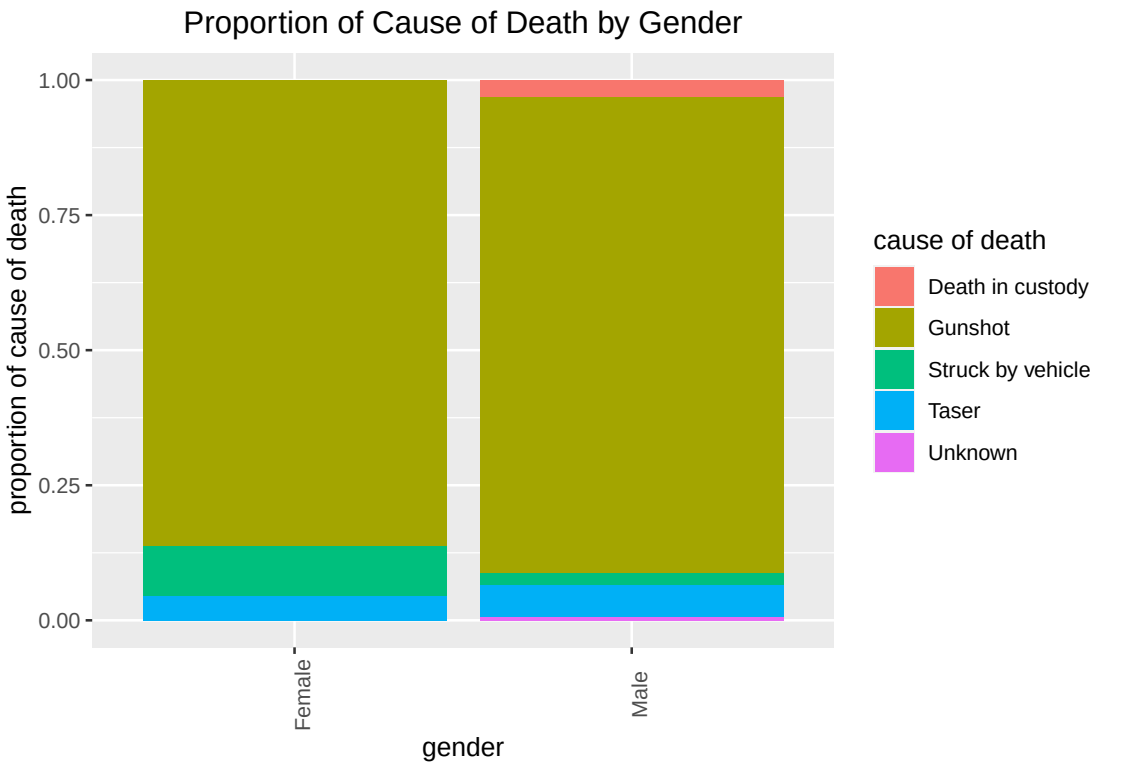
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 4.7296, df = 4, p-value = 0.3162
```

```
round(prop.table(contingency_table, 1) * 100, 1)
```

```
##
##          Death in custody Gunshot Struck by vehicle Taser Unknown
##   Female              0.0    86.4                9.1   4.5     0.0
##   Male                3.1    88.1                2.2   5.8     0.7
```

```
ggplot(gender_and_cause_of_death, aes(x = gender, fill = cause)) +
    geom_bar(position = "fill") +
    scale_fill_discrete(name = "cause of death") +
    labs(
        y = "proportion of cause of death",
        title = "Proportion of Cause of Death by Gender"
    ) +
```

```
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 90)
    )
```




```
table(gender_and_cause_of_death$gender)
```

```
##
## Female   Male
##     22    445
```

Given a significance level 0.05, since the above probability 0.032 is less than the significance level, we reject the null hypothesis of the Pearson's Chi-squared test of independence, which states that there is no association between the row variable `gender` and the column variable `cause`. We have sufficient evidence to conclude that there is an association between the row variable `gender` and the column variable `cause`. We note that the "Chi-squared approximation may be incorrect".

The proportion of people killed by gunshot is approximately equal across gender. Only male people died in custody. A significantly higher proportion of female people were struck by a vehicle. A significantly higher proportion of male people were killed by taser.

2. For this question, use the `.csv` data file that you created at the end of the previous homework set, `stateCovid.csv`. The dataset should contain 4 columns:

   - the name of the political entity (i.e., one of the 50 states, 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.

```
state.level <-
    read.csv(file = "../../Module_1--Data_Wrangling/Homework/stateCovid.csv") %>%
    select(-X) %>%
    rename(State = state, Death_Rate = state.rate)
head(state.level, n = 3)
```

```
##            State  Cases Deaths Death_Rate
## 1         Alaska  69826    352       0.50
## 2           Utah 406895   2308       0.57
## 3 Virgin Islands   3512     28       0.80
```

(a) There is a dataset on Collab in `State_pop_election.csv`. The dataset contains 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 from `stateCovid.csv` and `State_pop_election.csv`. Use the `head()` function to display the first 6 rows after merging these two datasets.

```
states_populations_and_votes <- read.csv(file = "State_pop_election.csv")
head(states_populations_and_votes, n = 3)
```

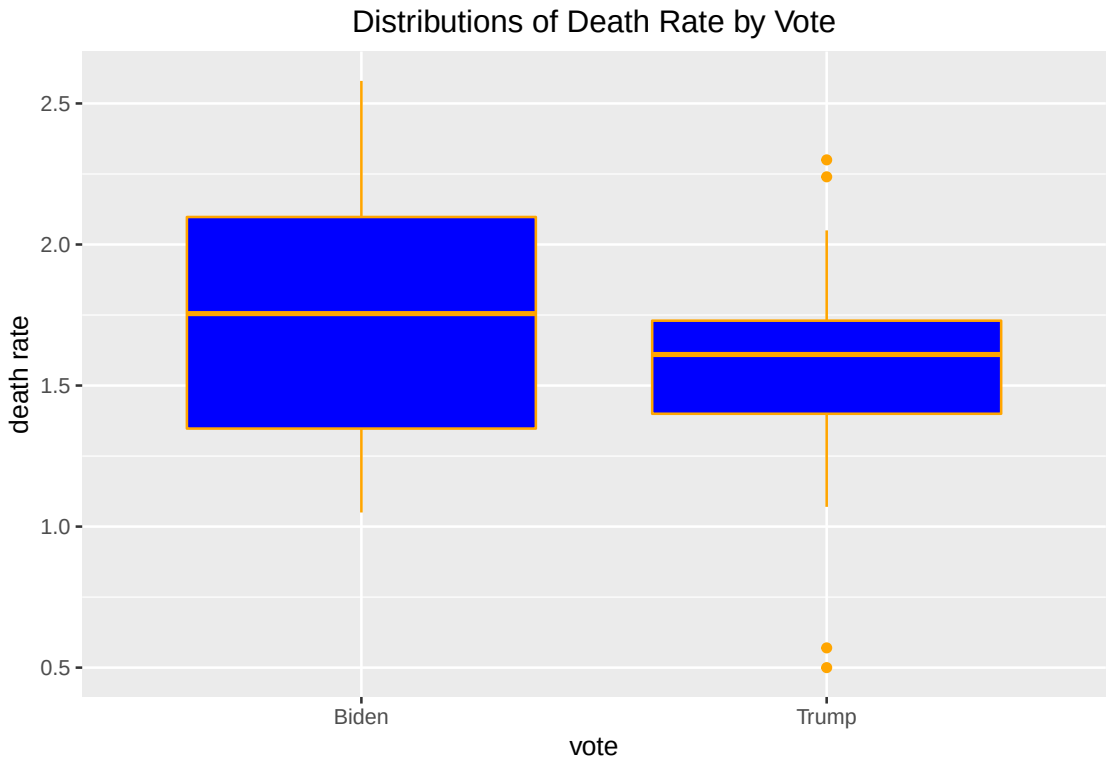
```
##     State Population Election
## 1 Alabama    5024279    Trump
## 2  Alaska     733391    Trump
## 3 Arizona    7151502    Biden
```

```
states_cases_deaths_death_rates_populations_and_votes <-
    state.level %>% full_join(states_populations_and_votes, by = "State")
head(states_cases_deaths_death_rates_populations_and_votes, n = 6)
```

```
##            State  Cases Deaths Death_Rate Population Election
## 1         Alaska  69826    352       0.50     733391    Trump
## 2           Utah 406895   2308       0.57    3271616    Trump
## 3 Virgin Islands   3512     28       0.80         NA     <NA>
## 4        Vermont  24240    255       1.05     643077    Biden
## 5       Nebraska 223517   2385       1.07    1961504    Trump
## 6          Idaho 192704   2103       1.09    1839106    Trump
```

(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 describe how you created the new variables.

```
state_vote_and_death_rate <-
    states_cases_deaths_death_rates_populations_and_votes %>%
        select(State, Election, Death_Rate) %>%
        filter(!is.na(Election)) %>%
        rename(state = State, vote = Election, death_rate = Death_Rate)
ggplot(state_vote_and_death_rate, aes(x = vote, y = death_rate)) +
    geom_boxplot(fill = "Blue", color = "Orange") +
    labs(
        y = "death rate",
        title = "Distributions of Death Rate by Vote"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```




```
death_rate <- state_vote_and_death_rate %>% select(death_rate)
head(
    state_vote_and_death_rate %>%
        arrange(desc(death_rate)) %>%
        filter(vote == "Biden"),
    n = 3
)
```

```
##           state  vote death_rate
## 1    New Jersey Biden       2.58
## 2 Massachusetts Biden       2.53
## 3      New York Biden       2.51
```

```
head(
    state_vote_and_death_rate %>%
        arrange(desc(death_rate)) %>%
        filter(vote == "Trump"),
    n = 3
)
```

```
##         state  vote death_rate
## 1 Mississippi Trump       2.30
## 2   Louisiana Trump       2.24
## 3     Alabama Trump       2.05
```

The maximum death rate occurs for New Jersey, whose electors voted for Joe Biden in 2020. Including outliers, the second and third largest death rates occur for Mississippi and Louisiana, whose electors voted for Trump. Excluding outliers, the second largest death rate occurs for Massachusetts, whose electors voted for Biden. Including outliers, the minimum death rate occurs for Alaska, whose electors voted for Trump. Excluding outliers, the minimum death rate occurs for Vermont, whose electors voted for Biden. The median and third quartile death rates and

interquartile range among states whose electors voted for Biden is greater than the corresponding death rates and interquartile range among states whose electors voted for Trump. The first quartile death rate among states whose electors voted for Biden is less than the corresponding first quartile death rate among states whose electors voted for Trump.