STAT 6021: Homework Set 2

Question 1

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

Prework

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ------------------------------------------------------------------------ tidyve
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts --------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/hw2")
data = read.csv("PoliceKillings.csv", header = TRUE)
head(data)
```

```
##                  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
## 1          3 14.1 0.09768638 0.16850951 NA
## 2          1 28.8 0.06572379 0.11140236 NA
## 3          3 14.6 0.16629314 0.14731227 NA
## 4          3 11.7 0.12482727 0.05013293 NA
## 5          4  1.9 0.06354983 0.40395421 NA
## 6          1   58 0.07365145 0.10295519 NA
```

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.

*1 Create a table that displays the proportions of victimis in each race/ethnic level.

```
prop.table(table(data$raceethnicity))
```

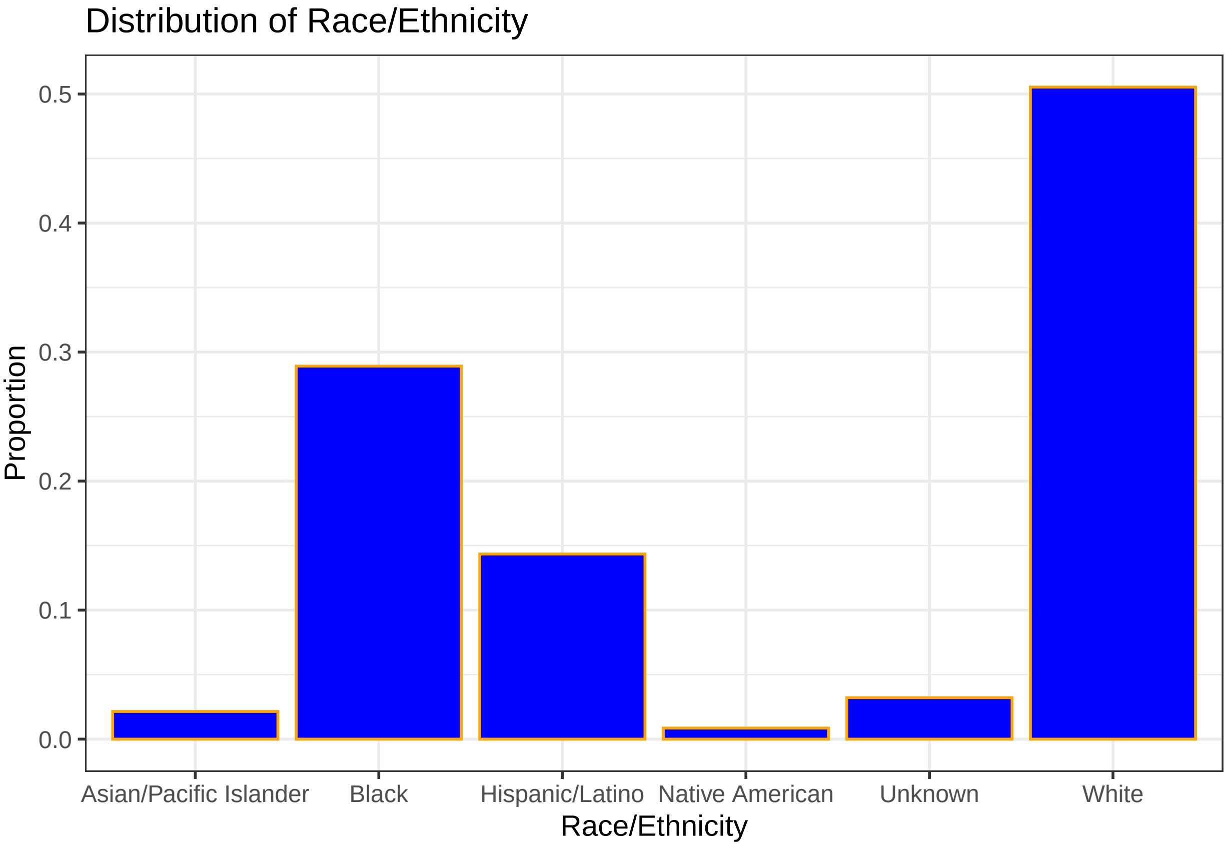
```
## 
## Asian/Pacific Islander                  Black        Hispanic/Latino 
##             0.02141328             0.28907923             0.14346895 
##        Native American                Unknown                  White 
##              0.00856531              0.03211991              0.50535332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

*2 Create a bar chart that displays the proportions of victims in each race/ethnic level.

```
byRace = data %>%
  group_by(raceethnicity) %>%
  summarize(Counts=n()) %>%
  mutate(Proportion = Counts/nrow(data))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(byRace, aes(x=raceethnicity, y=Proportion)) +
  geom_bar(stat="identity", fill="blue", color="orange") +
  labs(x="Race/Ethnicity", y="Proportion", title="Distribution of Race/Ethnicity") +
  theme_bw()
```



Looking at the graphical representation/table above approximatly we can see the trend of census data to see if this graph/data follows the same trend. The following is how the data compares to the census data.

```
prop.table(table(data$raceethnicity))
```

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

```
race=c("White", "Black", "Hispanic/Latino","Unknown","Asian/Pacific Islander","Native American")
props = c(0.50535332,0.28907923,0.14346895,0.03211991,0.02141328,0.00856531)
census = c(60.1, 13.4, 1.8, NA, 5.9+0.2, 1.3)/100
```

```
difference = props - census
data.frame(race, props, census,difference)
```

```
##                    race      props census  difference
## 1                 White 0.50535332  0.601 -0.09564668
## 2                 Black 0.28907923  0.134  0.15507923
## 3        Hispanic/Latino 0.14346895  0.018  0.12546895
## 4               Unknown 0.03211991     NA          NA
## 5 Asian/Pacific Islander 0.02141328  0.061 -0.03958672
## 6        Native American 0.00856531  0.013 -0.00443469
```

The census bureau does not include information about unknown race, therefore in regards to all other races, the trend for police killings and race distribution of the population seems to be similar with White and Black race/ethnicities being the highest proportion for both datasets. However, it was interesting to see that in the difference between the observed population and police killing distribution, Black and Hisplanics had an opposite trend. Based on data it might seem that the Black and Hispanic/Latino community should have a lower proportion of police killings in respect to the relative population of the ethnicity, but their police killing distribution was a bit higher. This might be something that could be investigated further.

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.

```
data$age.num = as.numeric(data$age)
```

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

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

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

```
data$age.num
```

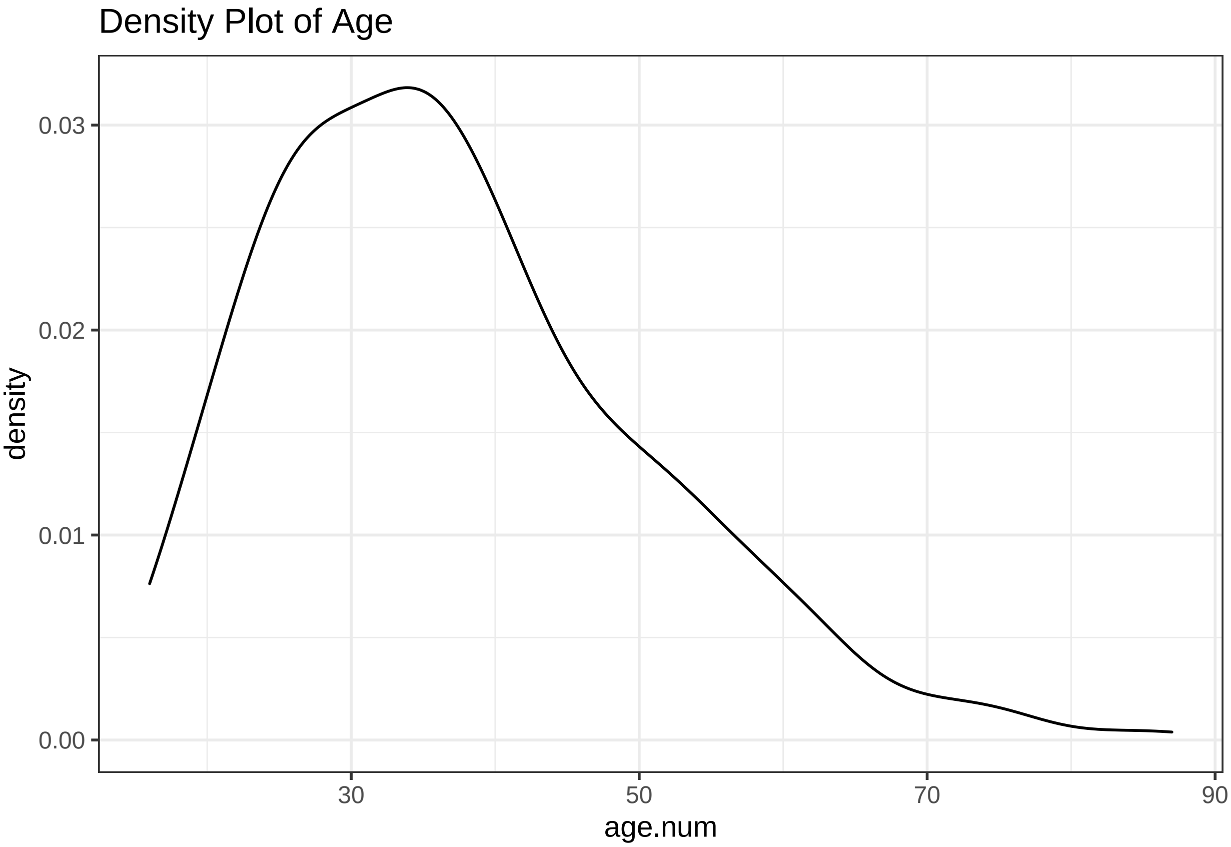
```
##   [1] 16 27 26 25 29 29 22 35 44 31 76 40 NA 31 23 39 25 54 24 57 21 42 21 36 26
##  [26] 49 54 26 48 33 21 41 48 36 41 29 27 45 32 35 36 35 40 18 34 39 21 62 43 44
##  [51] 29 35 50 18 25 31 29 49 23 45 26 35 34 46 29 39 28 51 67 53 25 30 24 35 43
##  [76] 24 29 38 31 36 23 38 53 24 26 28 34 28 40 51 44 25 56 37 58 39 37 35 26 47
## [101] 31 60 32 45 42 17 22 33 18 39 59 33 58 58 47 41 64 45 53 24 48 29 40 34 21
## [126] 26 45 25 17 24 42 29 42 30 29 39 63 49 41 27 30 60 77 19 37 54 29 30 24 21
## [151] 44 40 32 22 43 52 27 34 20 25 24 20 46 42 43 41 59 25 42 64 22 24 63 56 60
## [176] 54 37 22 39 45 57 42 41 19 26 34 69 64 35 40 19 27 37 17 39 74 42 47 43 46
## [201] 44 31 47 41 43 40 32 31 20 20 33 22 41 41 32 16 29 42 29 47 53 18 47 34 36
## [226] 63 36 27 28 33 32 42 31 17 28 24 71 51 28 53 54 45 33 48 34 23 35 33 32 52
## [251] 30 23 35 42 37 56 36 27 30 31 46 51 72 28 63 28 33 24 27 24 28 28 17 46 52
## [276] 39 49 30 51 16 18 22 40 61 52 51 36 36 59 17 18 41 33 25 23 47 58 47 34 28
## [301] 37 87 39 27 35 36 24 26 34 51 49 41 54 36 26 35 22 27 42 32 32 25 26 53 26
## [326] 40 55 29 31 19 57 40 35 35 39 37 36 62 43 32 34 37 37 33 35 40 21 30 23 26
## [351] 39 33 34 37 26 24 25 31 49 59 50 37 28 26 23 32 24 42 34 68 31 83 35 29 50
## [376] 56 43 38 63 27 36 55 36 68 61 46 47 26 37 22 18 39 49 23 47 32 45 51 31 54
## [401] 31 23 29 28 31 24 27 57 39 38 34 39 20 35 36 38 33 57 38 72 37 47 43 37 75
## [426] 21 20 29 37 41 22 23 64 34 49 32 25 39 53 27 36 20 39 19 34 36 34 31 45 34
## [451] NA NA NA 31 28 57 29 50 40 35 53 59 18 28 52 38 48
```

C

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

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

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



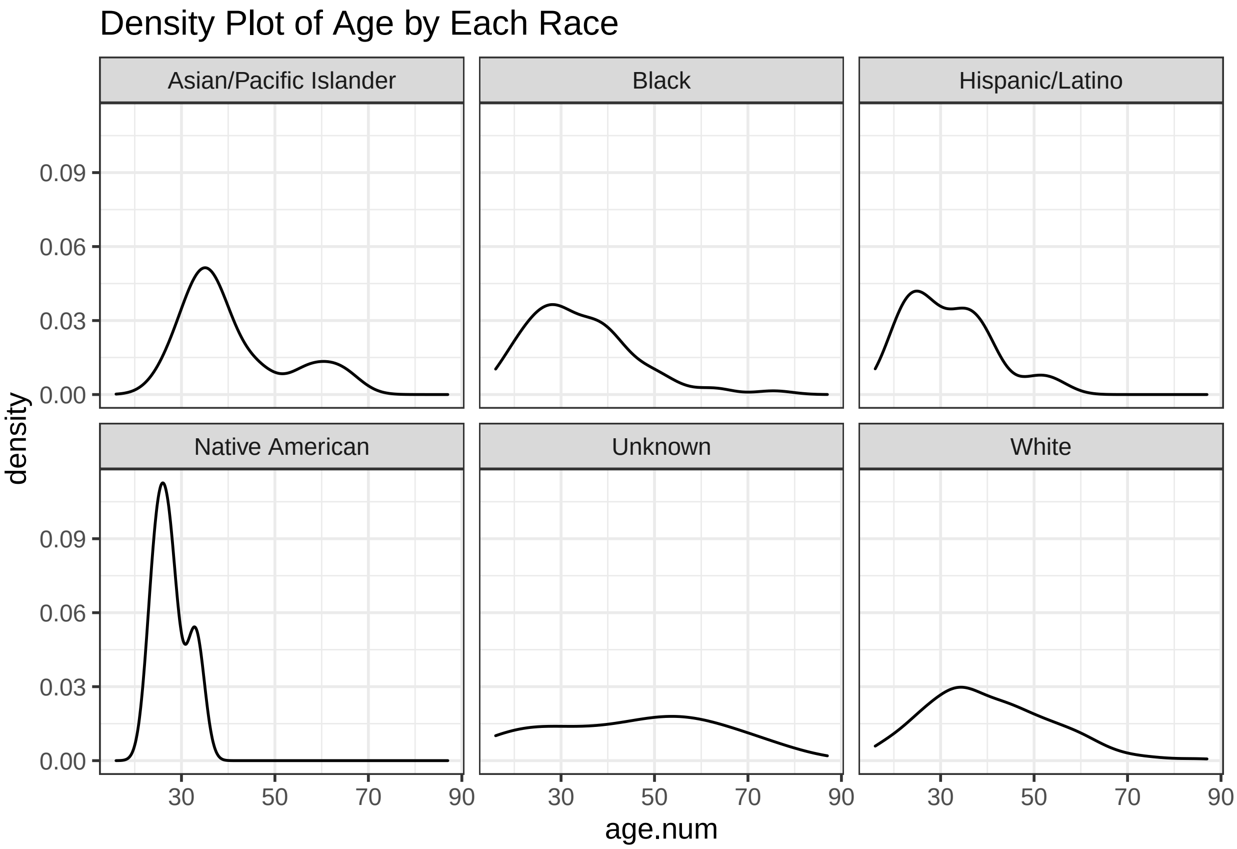
Based on this density plot, it seems that the majority of the data for police killings occured between the range of the age 20-50 years of age. This is interesting since 21 is when you are a full adult with the introduction of alcohol and all responsibilites and as age increases it could be that people commite less crime that could lead to police killings.

D

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

```
ggplot(data, aes(x=age.num)) +
  geom_density() +
  labs(title="Density Plot of Age by Each Race") +
  theme_bw() +
  facet_wrap(~raceethnicity)
```

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

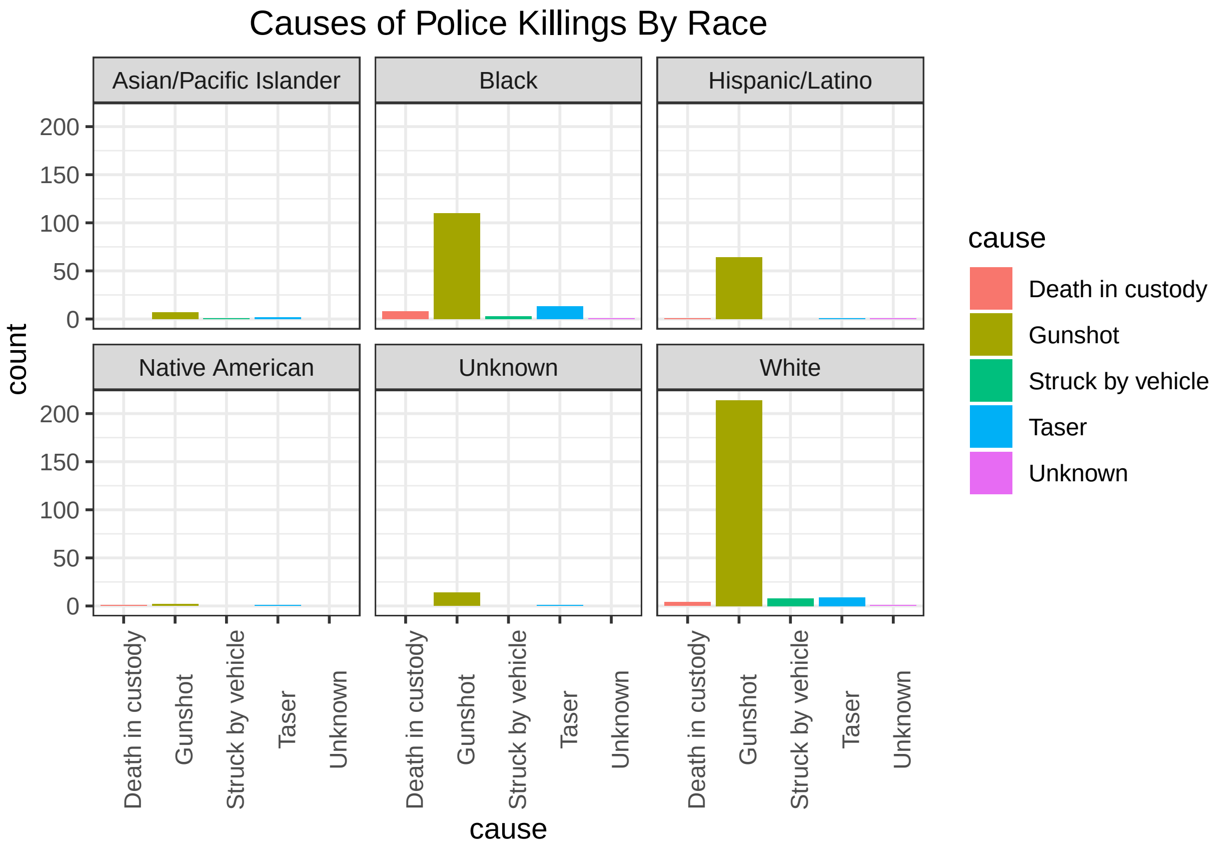


With the exception of unknown race in teh data, we can see that all races follow a simlar pattern of the most police killings being taken place between the age of ~20 - ~50. The Native American Race/Ethnicity has a slighltly higher proportion happening in the lower 20's and 30's when compared to the other race/ethnicity groups, but the general trend seem to be similar.

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.

```r
ggplot(data, aes(x=cause, fill=cause)) +
  geom_bar() +
  labs(title="Causes of Police Killings By Race") +
  facet_wrap(~raceethnicity) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90), plot.title = element_text(hjust = 0.5))
```



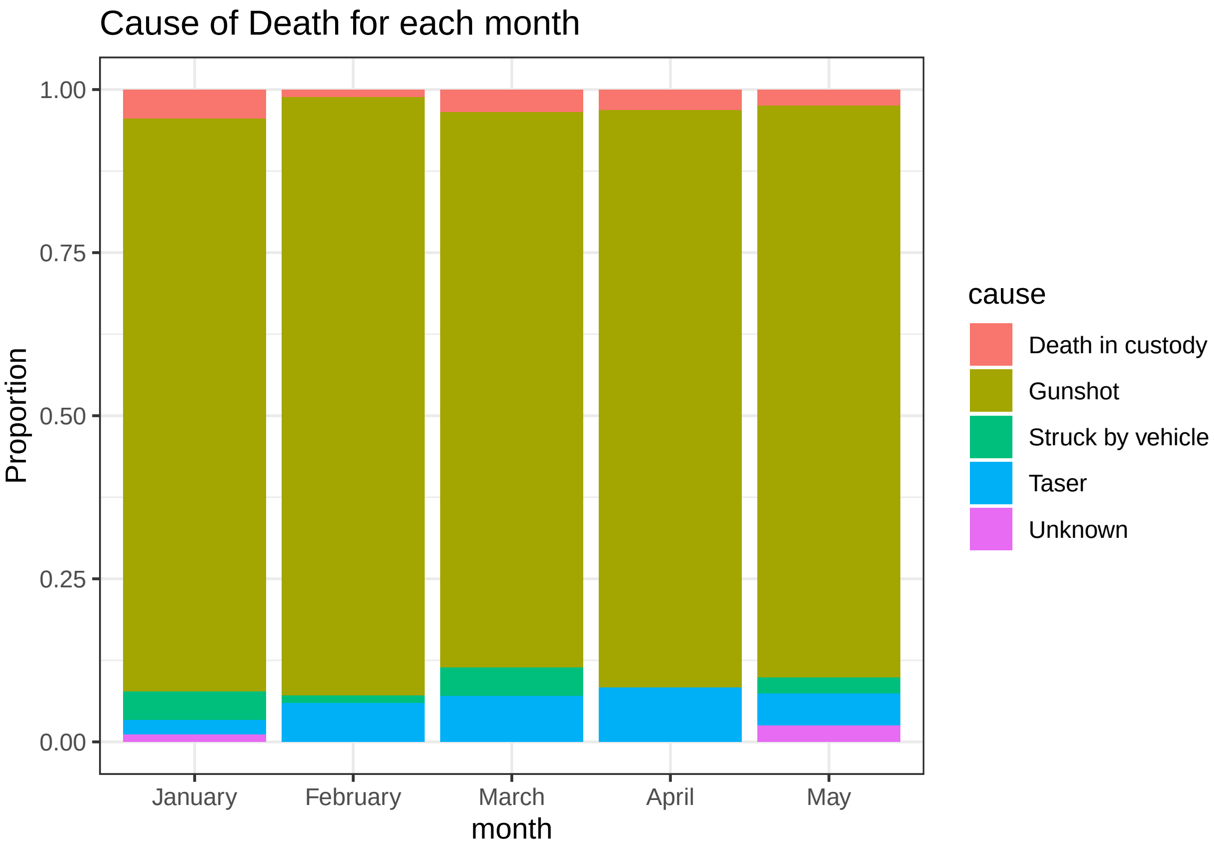
Looking at the casues of death for each race, it was shown for White, Black, and Hispanic/Latino communities Gunshot was the biggest cause of death while for the other communities all causes were similar. All other causes except gunshot seems to be independent of race/ethnicity except for gunshot.

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.

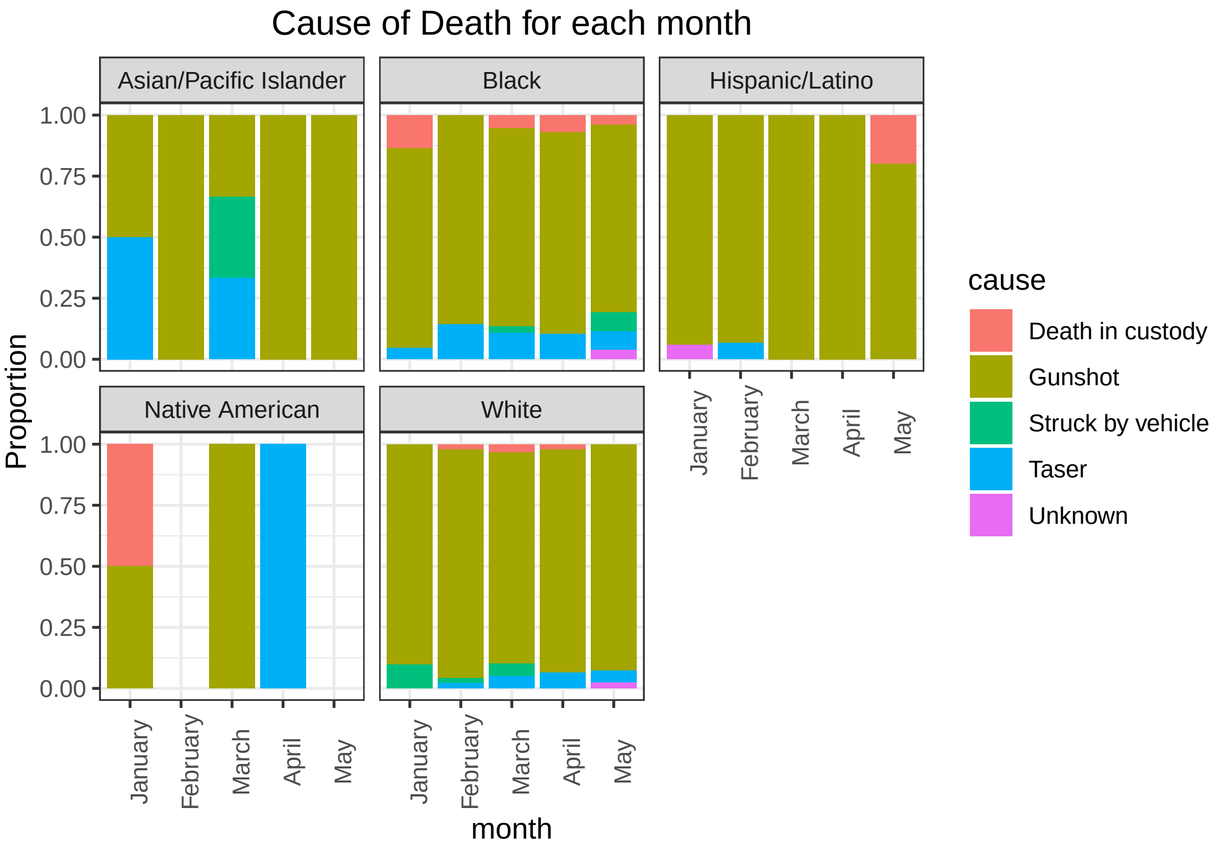
I thought it will be interesting to see whether the month (time of year) had a play in the cause of death for police killings.

```
data$month <- factor(data$month, levels=c("January", "February", "March", "April", "May"))
data %>%
  filter(!is.na(month)) %>%
  ggplot(aes(x=month, fill=cause)) +
  geom_bar(position = "fill") +
  labs(title="Cause of Death for each month", y="Proportion") +
  theme_bw()
```



From the visualization above there wasn't that much difference of causes of death for each month. The Gunshot category was the dominating factor. However, we saw that for the white, black, hispanic/latino communities gunshot was thee dominating factor. Therefore, we could also look at this plot by race/ethnicity.

```
data %>%
  filter(!is.na(month)) %>%
  filter(raceethnicity != "Unknown") %>%
  ggplot(aes(x=month, fill=cause)) +
  geom_bar(position = "fill") +
  facet_wrap(~raceethnicity) +
  labs(title="Cause of Death for each month", y="Proportion") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90), plot.title = element_text(hjust = 0.5))
```



This visualization is much more interesting but based on how similar the causes were for the race/ethnicity groups that are not White, Black or hispanic/latino it is still unclear to make a decisive hypothesis.

Question 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 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.

Prework

```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/hw2")
stateCovid = read.csv("stateCovid.csv", header = TRUE)
dim(stateCovid)
```

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

```
head(stateCovid)
```

```
##        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.

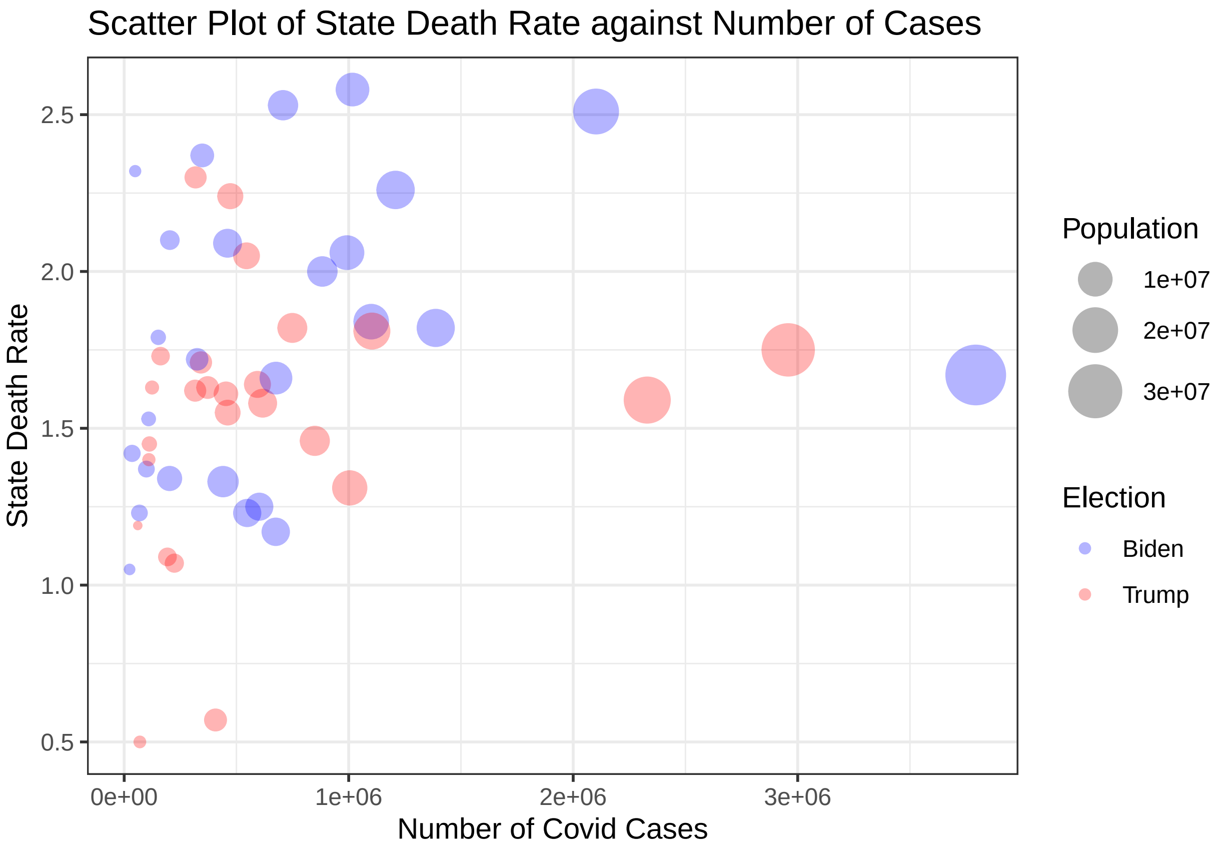
```
stateElection = read.csv("State_pop_election.csv", header=TRUE)
stateMerged = merge(stateCovid, stateElection, by.x="state", by.y="State")
head(stateMerged)
```

```
##        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
```

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.

```
stateMerged %>%
  filter(!is.na(Election)) %>%
  ggplot(aes(x=cases, y=state.rate, size=Population, color=Election)) +
  scale_color_manual(values=c("blue", "red")) +
  geom_point(alpha=0.3) +
  scale_size(range = c(1,10)) +
  labs(title="Scatter Plot of State Death Rate against Number of Cases",
       y="State Death Rate",
       x="Number of Covid Cases") +
  theme_bw()
```



This was a very interesting visual to see considering our nations covid status. Interstingly enough in my personal point of view we can see that the higher death rates voted for Biden. The states with the lowest state death rates tend to have voted fro Trump. However, in the middle range for state death rate (1.0-2.0) we could see that they are quite evenly mixed together so it might be hard to make a hard definitive statement. Also the trend the scatter plot is hard to decipher it doesn't seem like the more number of cases the higher the death rate given that there were many states with a lower number of covid cases and high death rates. And of these lower number of covid cases and high deathrates there wasn't a definitive winner for the trend of whether that state voted from Trump or Biden. The variables selected were selected based on the thought that maybe there was a trend between death ratess and presidential election result, but much more analysis and testing must be made to make a decision since the visual is not a clear visual in the trend.