Data Exploration

Name : SiaoHsuan Jiang | ID : 33029229
Tutors' name : Mohit Gupta, Abhishek Sriramulu | Tutorial number : 1

1. Introduction

Exploration of possible factors related to gun and murder safety issue in United States

Motivation:

I wondered which state is safer to live in, and I feel safety issue including both **gun** and **murder** aspects is worthy of exploring because holding firearms is legal in United States. Especially as an Asian, I've been curious about do racial distributions in all states have something to do with firearm deaths for a long time. In the future, I would like to find an ideal and safer state to live my life, so I decided to take this topic for my project.
I believe these analytics would help me have better insight and understanding about some implicit firearm risk in America.

Q1. Do possible factors such as race, poverty rate and gun ownership have something to do with firearm deaths?

Q2. Whether Murder and Nonnegligent Manslaughter or Firearm Deaths is more likely to occur in more racially complex areas or not. (Use rate data instead because population should be taken into consideration.)

2. Data Wrangling

Description of data source:

A. States with the most (and least) gun violence, which have been reported by USA Today. 50 rows and 7 columns. (https://www.usatoday.com/story/news/nation/2018/02/21/states-most-and-least-gun-violence-see-where-your-state-stacks-up/359395002/)

B. US States by Race 2022 (Total numbers and rates). 52 rows and 8 columns. (https://worldpopulationreview.com/states/states-by-race)

C. Gun violence and ownership ratio in the US. Tabular data from Wikipedia. 50 rows and 9 columns. (https://en.wikipedia.org/wiki/Gun_violence_in_the_United_States_by_state)

Note: All datasets will be used in both Q1 and Q2.

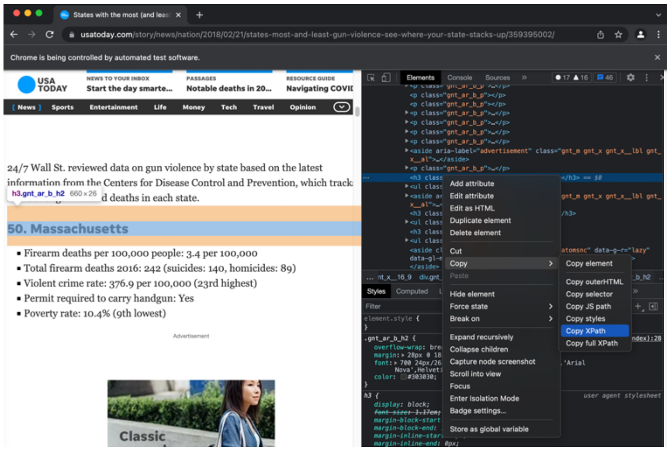
Data Cleaning and Data Transformation:

Dataset A and C:

The source of Dataset A is a textual article and Dataset C is a textual table. Both datasets cannot be downloaded as csv file directly via links provided above, so I need to use some web scraping tools to transform the textual data into the desired format and then export it.

The following graph shows the flow of data transformation. I used selenium tool with XPath method to scrape text information on Python. At the last step, two datasets were written into .csv files as a preparation of further analytics in R and Tableau. For more detail, full pre-processing Python code has been pushed to my personal github.
Github Link:
https://github.com/rock86109/FIT5147_Data_Visualization/tree/main/DEP_Web_Scraping

Get XPath for web scraping



24/7 Wall St. reviewed data on gun violence by state based on the latest information from the Centers for Disease Control and Prevention, which track[...] d deaths in each state.

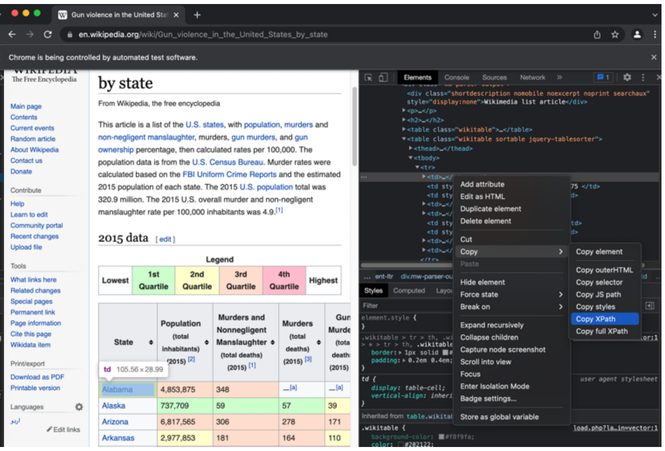
h3.gnt_ar_b_h2 660×26

50. Massachusetts

- Firearm deaths per 100,000 people: 3.4 per 100,000
- Total firearm deaths 2016: 242 (suicides: 140, homicides: 89)
- Violent crime rate: 376.9 per 100,000 (23rd highest)
- Permit required to carry handgun: Yes
- Poverty rate: 10.4% (9th lowest)

```
<p class="gnt_ar_b_p"></p>
<p class="gnt_ar_b_p"></p>
▶<p class="gnt_ar_b_p">…</p>
▶<p class="gnt_ar_b_p">…</p>
▶<p class="gnt_ar_b_p">…</p>
▶<p class="gnt_ar_b_p">…</p>
▶<aside aria-label="advertisement" class="gnt_m gnt_x gnt_x__lbl gnt_
x_al">…</aside>
▶<p class="gnt_ar_b_p">…</p>
<h3 class                     </h3> == $0
```

```
.gnt_ar_b_h2 {
  overflow-wrap: bre
  margin: ▶ 28px 0 18
  font: ▶ 700 24px/26
      Nova',Helveti
  color: #303030;
}
h3 {
  display: block;
  font-size: 1.17em;
  margin-block-start
  margin-block-end:
  margin-inline-star
  margin-inline-end: 0px;
```



by state

From Wikipedia, the free encyclopedia

This article is a list of the U.S. states, with population, murders and non-negligent manslaughter, murders, gun murders, and gun ownership percentage, then calculated rates per 100,000. The population data is from the U.S. Census Bureau. Murder rates were calculated based on the FBI Uniform Crime Reports and the estimated 2015 population of each state. The 2015 U.S. population total was 320.9 million. The 2015 U.S. overall murder and non-negligent manslaughter rate per 100,000 inhabitants was 4.9.[1]

2015 data [edit]

Legend

Lowest	1st Quartile	2nd Quartile	3rd Quartile	4th Quartile	Highest

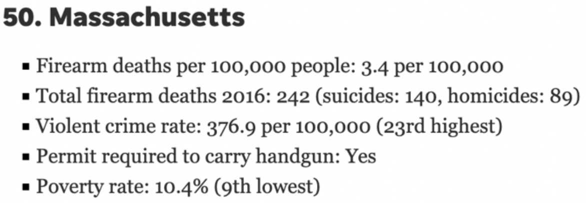
td 105.56×28.99

State	Population (total inhabitants) (2015) [2]	Murders and Nonnegligent Manslaughter (total deaths) (2015) [1]	Murders (total deaths) (2015) [3]	Gun Murde (tota death (2015)
Alabama	4,853,875	348	—[a]	—[a]
Alaska	737,709	59	57	39
Arizona	6,817,565	306	278	171
Arkansas	2,977,853	181	164	110

```
<div class="shortdescription nomobile noexcerpt noprint searchaux"
style="display:none">Wikimedia list article</div>
▶<p>…</p>
▶<h2>…</h2>
▶<table class="wikitable">…</table>
▼<table class="wikitable sortable jquery-tablesorter">
  ▶<thead>…</thead>
  ▼<tbody>
    ▼<tr>
      ▶<td>…</
```

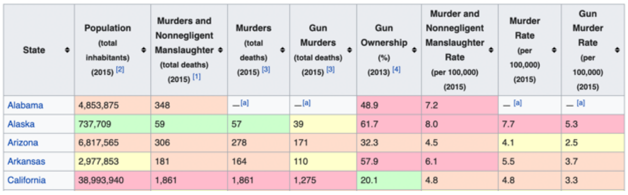
```
element.style {
}
.wikitable > tr > th, .wi
> * > tr > th, .wikitable
  border: ▶ 1px solid
  padding: ▶ 0.2em 0.4em;
}
td {
  display: table-cell;
  vertical-align: inheri
}
Inherited from table.wikital
.wikitable {
  background-color: #f8f9fa;
  color: #202122;
```

Original textual data information



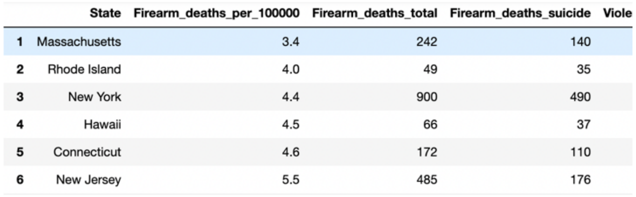
50. Massachusetts

- Firearm deaths per 100,000 people: 3.4 per 100,000
- Total firearm deaths 2016: 242 (suicides: 140, homicides: 89)
- Violent crime rate: 376.9 per 100,000 (23rd highest)
- Permit required to carry handgun: Yes
- Poverty rate: 10.4% (9th lowest)

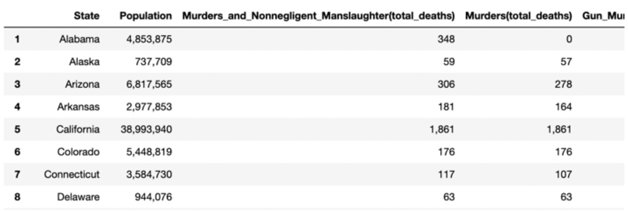


State	Population (total inhabitants) (2015) [2]	Murders and Nonnegligent Manslaughter (total deaths) (2015) [1]	Murders (total deaths) (2015) [3]	Gun Murders (total deaths) (2015) [3]	Gun Ownership (%) (2013) [4]	Murder and Nonnegligent Manslaughter Rate (per 100,000) (2015)	Murder Rate (per 100,000) (2015)	Gun Murder Rate (per 100,000) (2015)
Alabama	4,853,875	348	—[a]	—[a]	48.9	7.2	— [a]	— [a]
Alaska	737,709	59	57	39	61.7	8.0	7.7	5.3
Arizona	6,817,565	306	278	171	32.3	4.5	4.1	2.5
Arkansas	2,977,853	181	164	110	57.9	6.1	5.5	3.7
California	38,993,940	1,861	1,861	1,275	20.1	4.8	4.8	3.3

Data wrangling into desired format and then export the data

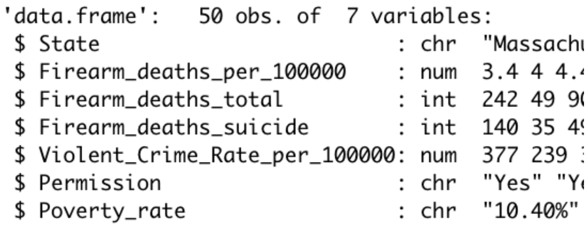


	State	Firearm_deaths_per_100000	Firearm_deaths_total	Firearm_deaths_suicide	Viole
1	Massachusetts	3.4	242	140	
2	Rhode Island	4.0	49	35	
3	New York	4.4	900	490	
4	Hawaii	4.5	66	37	
5	Connecticut	4.6	172	110	
6	New Jersey	5.5	485	176	

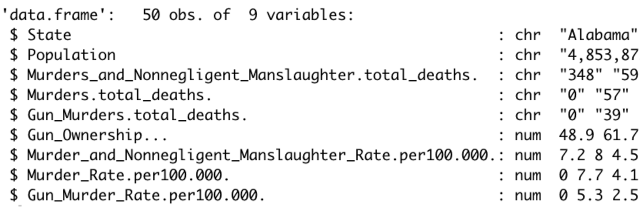


	State	Population	Murders_and_Nonnegligent_Manslaughter(total_deaths)	Murders(total_deaths)	Gun_Mu
1	Alabama	4,853,875	348	0	
2	Alaska	737,709	59	57	
3	Arizona	6,817,565	306	278	
4	Arkansas	2,977,853	181	164	
5	California	38,993,940	1,861	1,861	
6	Colorado	5,448,819	176	176	
7	Connecticut	3,584,730	117	107	
8	Delaware	944,076	63	63	

Brief summary outputs of two exported datasets.



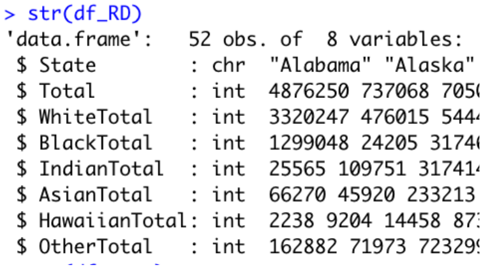
```
'data.frame':	50 obs. of  7 variables:
 $ State                      : chr  "Massachu
 $ Firearm_deaths_per_100000  : num  3.4 4 4.
 $ Firearm_deaths_total       : int  242 49 9
 $ Firearm_deaths_suicide     : int  140 35 4
 $ Violent_Crime_Rate_per_100000: num  377 239
 $ Permission                 : chr  "Yes" "Y
 $ Poverty_rate               : chr  "10.40%"
```



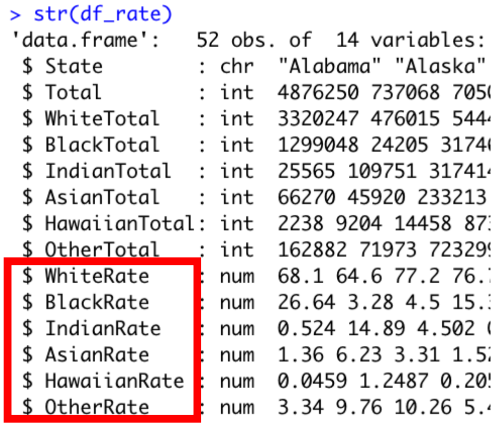
```
'data.frame':	50 obs. of  9 variables:
 $ State                                          : chr  "Alabama"
 $ Population                                     : chr  "4,853,87
 $ Murders_and_Nonnegligent_Manslaughter.total_deaths. : chr  "348" "59
 $ Murders.total_deaths.                          : chr  "0" "57"
 $ Gun_Murders.total_deaths.                      : chr  "0" "39"
 $ Gun_Ownership...                               : num  48.9 61.7
 $ Murder_and_Nonnegligent_Manslaughter_Rate.per100.000.: num  7.2 8 4.5
 $ Murder_Rate.per100.000.                        : num  0 7.7 4.1
 $ Gun_Murder_Rate.per100.000.                    : num  0 5.3 2.5
```

Dataset B

Dataset B can be downloaded directly via the link provided above.
However, it only provides either the total numbers of different races or race rates of different races in each state. Thus, I downloaded the former one and did a transformation. With given numbers in the dataset, I can get the rate data after calculation. Taking advantages of mutate function in dplyr package, I added 6 new columns (in the red frame below) to this dataset.



```
> str(df_RD)
'data.frame':	52 obs. of  8 variables:
 $ State       : chr  "Alabama" "Alaska"
 $ Total       : int  4876250 737068 705
 $ WhiteTotal  : int  3320247 476015 544
 $ BlackTotal  : int  1299048 24205 3174
 $ IndianTotal : int  25565 109751 31741
 $ AsianTotal  : int  66270 45920 233213
 $ HawaiianTotal: int  2238 9204 14458 87
 $ OtherTotal  : int  162882 71973 72329
```

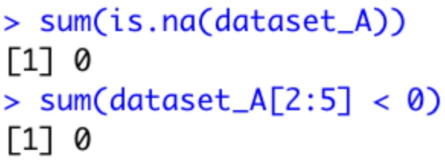


```
> str(df_rate)
'data.frame':	52 obs. of  14 variables:
 $ State       : chr  "Alabama" "Alaska"
 $ Total       : int  4876250 737068 705
 $ WhiteTotal  : int  3320247 476015 544
 $ BlackTotal  : int  1299048 24205 3174
 $ IndianTotal : int  25565 109751 31741
 $ AsianTotal  : int  66270 45920 233213
 $ HawaiianTotal: int  2238 9204 14458 87
 $ OtherTotal  : int  162882 71973 72329
 $ WhiteRate   : num  68.1 64.6 77.2 76.
 $ BlackRate   : num  26.64 3.28 4.5 15.
 $ IndianRate  : num  0.524 14.89 4.502
 $ AsianRate   : num  1.36 6.23 3.31 1.5
 $ HawaiianRate : num  0.0459 1.2487 0.20
 $ OtherRate   : num  3.34 9.76 10.26 5.
```

3. Data Checking

Dataset A

Dataset A is the cleanest dataset among these three. I used sum and is.na functions in R to check whether there is unreasonable number. The result is there are neither missing value in whole data nor negative numbers in numeric columns.

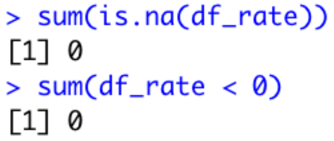


```
> sum(is.na(dataset_A))
[1] 0
> sum(dataset_A[2:5] < 0)
[1] 0
```

Dataset B

First

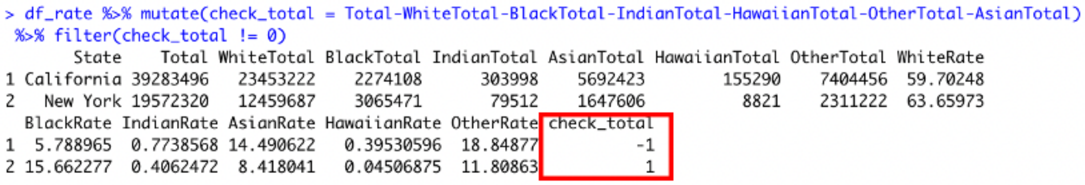
I slightly adjusted the code I used in Dataset A to check missing value and unreasonable values in Dataset B. There is no problem in this part as well.



```
> sum(is.na(df_rate))
[1] 0
> sum(df_rate < 0)
[1] 0
```

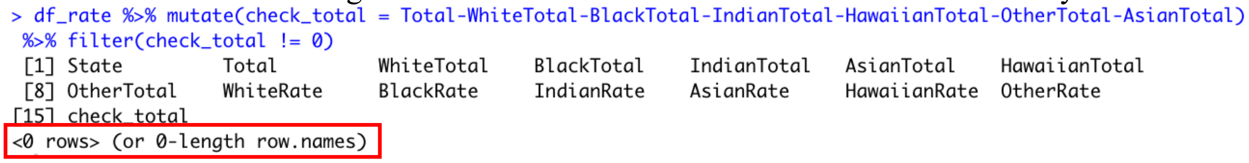
Second

I wondered whether the sum of population of all races would be equal to total population in each state. I used mutate function again to generate a new column called “check_total” which is calculated by subtracting population of all different races from total population, and then I used filter function to find check_total in what states is not equal to zero.
I found **California** and **New York** have unreasonable numbers as the following picture shows.



```
> df_rate %>% mutate(check_total = Total-WhiteTotal-BlackTotal-IndianTotal-HawaiianTotal-OtherTotal-AsianTotal)
 %>% filter(check_total != 0)
       State    Total WhiteTotal BlackTotal IndianTotal AsianTotal HawaiianTotal OtherTotal WhiteRate
1 California 39283496   23453222    2274108      303998    5692423        155290    7404456  59.70248
2   New York 19572320   12459687    3065471       79512    1647606          8821    2311222  63.65973
  BlackRate IndianRate AsianRate HawaiianRate OtherRate check_total
1  5.788965  0.7738568 14.490622   0.39530596  18.84877          -1
2 15.662277  0.4062472  8.418041   0.04506875  11.80863           1
```

The way I dealt with this problem is adjusting the total population numbers in those two states. That is to say, I added 1 to total population in California and minus 1 from total population in New York. Run the code again as a final check. There is no unreasonable value anymore.

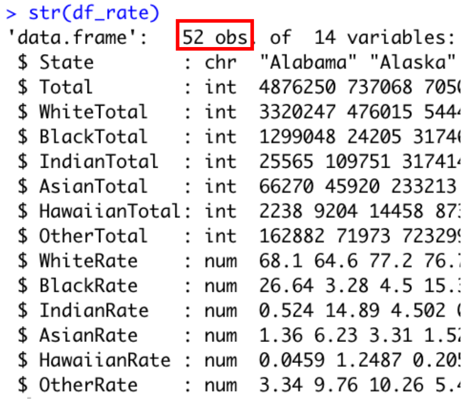


```
> df_rate %>% mutate(check_total = Total-WhiteTotal-BlackTotal-IndianTotal-HawaiianTotal-OtherTotal-AsianTotal)
  %>% filter(check_total != 0)
 [1] State         Total         WhiteTotal    BlackTotal    IndianTotal   AsianTotal    HawaiianTotal
 [8] OtherTotal    WhiteRate     BlackRate     IndianRate    AsianRate     HawaiianRate  OtherRate
[15] check_total
<0 rows> (or 0-length row.names)
```

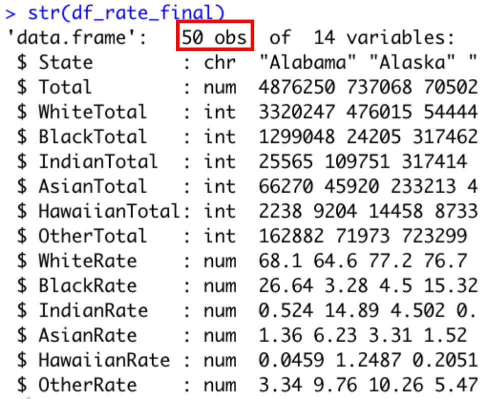
Third

I observed that there are 52 states in Dataset B (see the left screenshot below) by str() function in R. However, there are only 50 states in Dataset A and C, so what I need to do is identifying redundant states and remove them from Dataset B.

I checked whether a list contains items in another list or not by using **%in%** in R, and then realized **Puerto Rico** and **District of Columbia** are redundant states. The right screenshot shows the information about dataset after data checking. (Left one is before checking while right one is after)



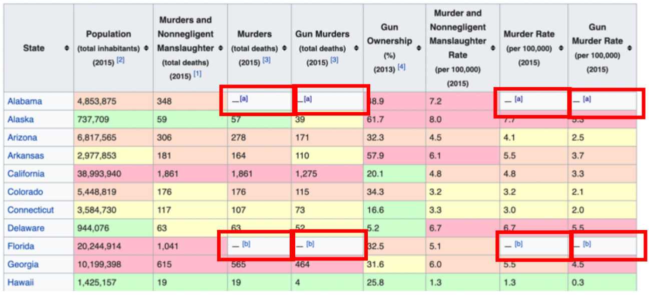
```
> str(df_rate)
'data.frame':	52 obs. of  14 variables:
 $ State        : chr  "Alabama" "Alaska"
 $ Total        : int  4876250 737068 705
 $ WhiteTotal   : int  3320247 476015 544
 $ BlackTotal   : int  1299048 24205 3174
 $ IndianTotal  : int  25565 109751 31741
 $ AsianTotal   : int  66270 45920 233213
 $ HawaiianTotal: int  2238 9204 14458 87
 $ OtherTotal   : int  162882 71973 72329
 $ WhiteRate    : num  68.1 64.6 77.2 76.
 $ BlackRate    : num  26.64 3.28 4.5 15.
 $ IndianRate   : num  0.524 14.89 4.502 
 $ AsianRate    : num  1.36 6.23 3.31 1.5
 $ HawaiianRate : num  0.0459 1.2487 0.20
 $ OtherRate    : num  3.34 9.76 10.26 5.
```



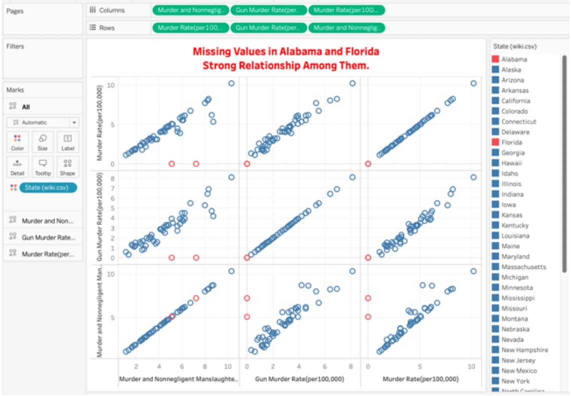
```
> str(df_rate_final)
'data.frame':	50 obs. of  14 variables:
 $ State        : chr  "Alabama" "Alaska" "
 $ Total        : num  4876250 737068 70502
 $ WhiteTotal   : int  3320247 476015 54444
 $ BlackTotal   : int  1299048 24205 317462
 $ IndianTotal  : int  25565 109751 317414
 $ AsianTotal   : int  66270 45920 233213 4
 $ HawaiianTotal: int  2238 9204 14458 8733
 $ OtherTotal   : int  162882 71973 723299
 $ WhiteRate    : num  68.1 64.6 77.2 76.7
 $ BlackRate    : num  26.64 3.28 4.5 15.32
 $ IndianRate   : num  0.524 14.89 4.502 0.
 $ AsianRate    : num  1.36 6.23 3.31 1.52
 $ HawaiianRate : num  0.0459 1.2487 0.2051
 $ OtherRate    : num  3.34 9.76 10.26 5.47
```

Dataset C

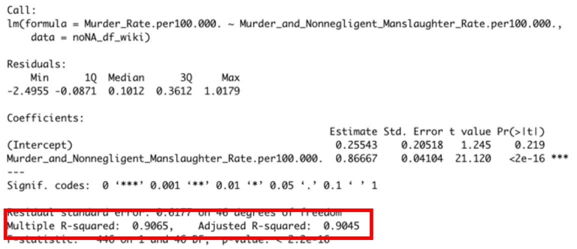
In the original Dataset C, it is observable 8 values are missing in the table (seeing red frames). Although I had replaced them with 0 when scraping, those values would still not make sense. Therefore, in this step, I have to replace these values with more reasonable imputation values.



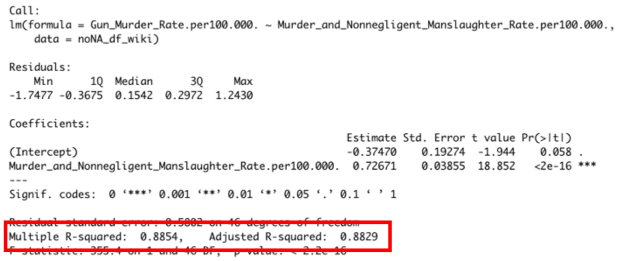
State	Population (total inhabitants) (2015) [2]	Murders and Nonnegligent Manslaughter (total deaths) (2015) [1]	Murders (total deaths) (2015) [3]	Gun Murders (total deaths) (2015) [3]	Gun Ownership (%) (2013) [4]	Murder and Nonnegligent Manslaughter Rate (per 100,000) (2015)	Murder Rate (per 100,000) (2015)	Gun Murder Rate (per 100,000) (2015)
Alabama	4,853,875	348	—[a]	—[a]	48.9	7.2	— [a]	— [a]
Alaska	737,709	59	57	39	61.7	8.0	7.7	5.3
Arizona	6,817,565	306	278	171	32.3	4.5	4.1	2.5
Arkansas	2,977,853	181	164	110	57.9	6.1	5.5	3.7
California	38,993,940	1,861	1,861	1,275	20.1	4.8	4.8	3.3
Colorado	5,448,819	176	176	115	34.3	3.2	3.2	2.1
Connecticut	3,584,730	117	107	73	16.6	3.3	3.0	2.0
Delaware	944,076	63	63	52	5.2	6.7	6.7	5.5
Florida	20,244,914	1,041	— [b]	— [b]	32.5	5.1	— [b]	— [b]
Georgia	10,199,398	615	565	464	31.6	6.0	5.5	4.5
Hawaii	1,425,157	19	19	4	25.8	1.3	1.3	0.3



It seems to be a relationship among the last three columns (right Tableau graph above). That is, **Murder and Nonnegligent Manslaughter Rate**, **Murder Rate**, and **Gun Murder Rate**. Hence, I removed Alabama and Florida rows to create a temporary dataset for regression fitting and further analysis. The results have been printed and visualized as follow.

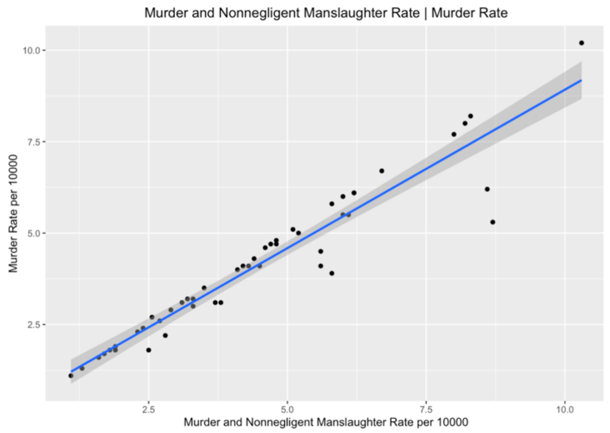


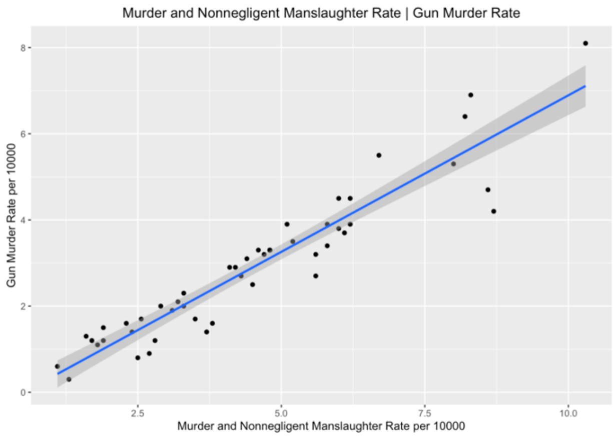
```
Call:
lm(formula = Murder_Rate.per100.000. ~ Murder_and_Nonnegligent_Manslaughter_Rate.per100.000.,
    data = noNA_df_wiki)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4955 -0.0871  0.1012  0.3612  1.0179 

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                            0.25543    0.20518   1.245    0.219    
Murder_and_Nonnegligent_Manslaughter_Rate.per100.000.  0.86667    0.04104  21.120   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6177 on 46 degrees of freedom
Multiple R-squared:  0.9065,	Adjusted R-squared:  0.9045 
F-statistic:   446 on 1 and 46 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = Gun_Murder_Rate.per100.000. ~ Murder_and_Nonnegligent_Manslaughter_Rate.per100.000.,
    data = noNA_df_wiki)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7477 -0.3675  0.1542  0.2972  1.2430 

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                           -0.37470    0.19274  -1.944    0.058 .  
Murder_and_Nonnegligent_Manslaughter_Rate.per100.000.  0.72671    0.03855  18.852   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5802 on 46 degrees of freedom
Multiple R-squared:  0.8854,	Adjusted R-squared:  0.8829 
F-statistic: 355.4 on 1 and 46 DF,  p-value: < 2.2e-16
```

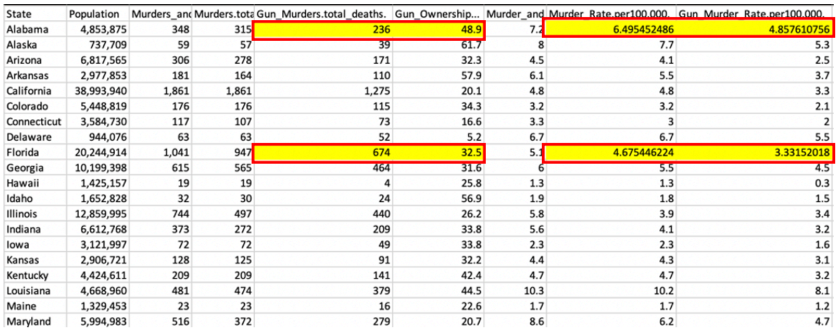
I got two sets of intercept and slope. One is **Murder and Nonnegligent Manslaughter Rate** and **Murder Rate**, and the other is **Murder and Nonnegligent Manslaughter Rate** and **Gun Murder Rate**. The scatter plots with regression lines are shown below. Moreover, both sets are in strong relationships because R squares are over 0.85. As a sequence, I believe without hesitation applying linear regression imputation to this case is suitable.





- Murder Rate = intercept1 + Murder and Nonnegligent Manslaughter Rate * slope1
- Gun Murder Rate = intercept2 + Murder and Nonnegligent Manslaughter Rate * slope2
- Murders = Murder Rate * Population
- Gun Murders = Gun Murder Rate * Population

Finally, I can get all desired values by following the formulas above. Imputed dataset is shown as follow.

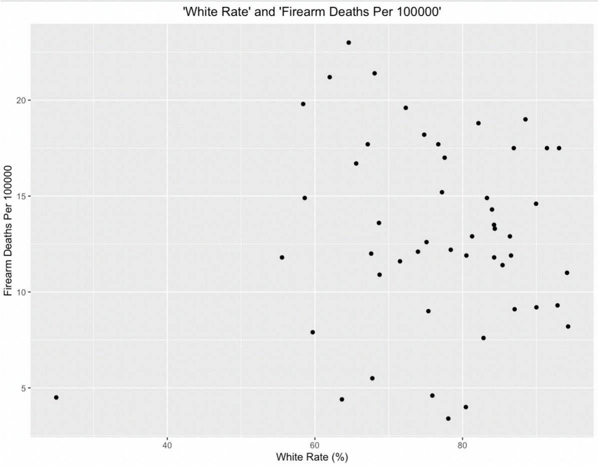


State	Population	Murders_anc	Murders.tota	Gun_Murders.total_deaths.	Gun_Ownership...	Murder_and	Murder_Rate.per100.000.	Gun_Murder_Rate.per100.000.
Alabama	4,853,875	348	315	236	48.9	7.2	6.495452486	4.857610756
Alaska	737,709	59	57	39	61.7	8	7.7	5.3
Arizona	6,817,565	306	278	171	32.3	4.5	4.1	2.5
Arkansas	2,977,853	181	164	110	57.9	6.1	5.5	3.7
California	38,993,940	1,861	1,861	1,275	20.1	4.8	4.8	3.3
Colorado	5,448,819	176	176	115	34.3	3.2	3.2	2.1
Connecticut	3,584,730	117	107	73	16.6	3.3	3	2
Delaware	944,076	63	63	52	5.2	6.7	6.7	5.5
Florida	20,244,914	1,041	947	674	32.5	5.1	4.675446224	3.33152018
Georgia	10,199,398	615	565	464	31.6	6	5.5	4.5
Hawaii	1,425,157	19	19	4	25.8	1.3	1.3	0.3
Idaho	1,652,828	32	30	24	56.9	1.9	1.8	1.5
Illinois	12,859,995	744	497	440	26.2	5.8	3.9	3.4
Indiana	6,612,768	373	272	209	33.8	5.6	4.1	3.2
Iowa	3,121,997	72	72	49	33.8	2.3	2.3	1.6
Kansas	2,906,721	128	125	91	32.2	4.4	4.3	3.1
Kentucky	4,424,611	209	209	141	42.4	4.7	4.7	3.2
Louisiana	4,668,960	481	474	379	44.5	10.3	10.2	8.1
Maine	1,329,453	23	23	16	22.6	1.7	1.7	1.2
Maryland	5,994,983	516	372	279	20.7	8.6	6.2	4.7

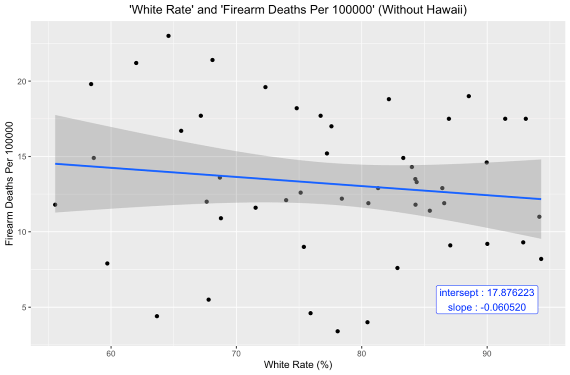
4. Data Exploration

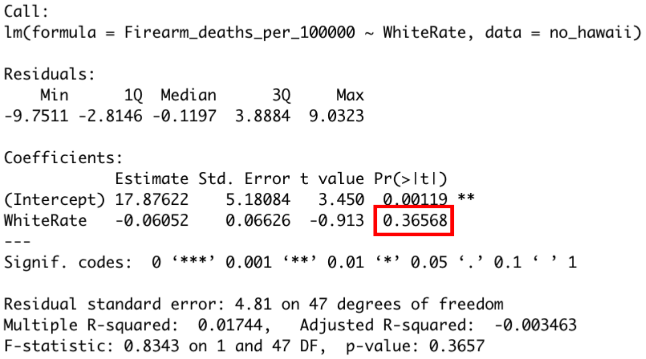
Q1.
How do possible factors such as race, poverty rate and gun ownership have impacts on firearm deaths?

First
I took a quick look on race and **Firearm Deaths Rate** by utilizing scatter plot. In Q1, I just used **White Rate**, which occupied the most proportion in most states, to see the relationship with **Firearm Deaths Rate**, and I will have a closer analysis on each race in Q2.
It's observable that there is a State, Hawaii, on the bottom left is significantly having a much lower white rate than the other states. Therefore, I believe it's reasonable to exclude it from the following model fitting because it doesn't belong to the same group as the other states.



After removing Hawaii, I fitted a regression model to check a relationship between two variables.





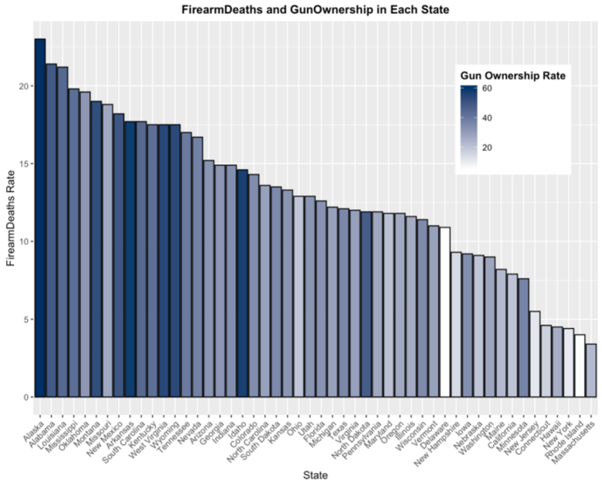
```
Call:
lm(formula = Firearm_deaths_per_100000 ~ WhiteRate, data = no_hawaii)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.7511 -2.8146 -0.1197  3.8884  9.0323 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 17.87622    5.18084   3.450  0.00119 **
WhiteRate   -0.06052    0.06626  -0.913  0.36568   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.81 on 47 degrees of freedom
Multiple R-squared:  0.01744,	Adjusted R-squared:  -0.003463 
F-statistic: 0.8343 on 1 and 47 DF,  p-value: 0.3657
```

Looking at the output above, since the p-value is greater than 0.05, we can't reject the null hypothesis (Ho: beta1 = 0). There is no evidence to say that there is a relationship between **White Rate** and **Firearm Deaths Rate.**

Second

To analyze the impacts of **Gun Ownership** on **Firearm Deaths Rate**, instead of using regression model again, I chose bar graph this time.

I sorted the order of bar graphs by **Firearm Deaths Rate**, and then fill the bars with **Gun Ownership Rate** which is continuous. We can clearly observe that bars on the left are having relatively darker colors which means states with the higher **Firearm Deaths** are having relatively higher **Gun Ownership Rates**.



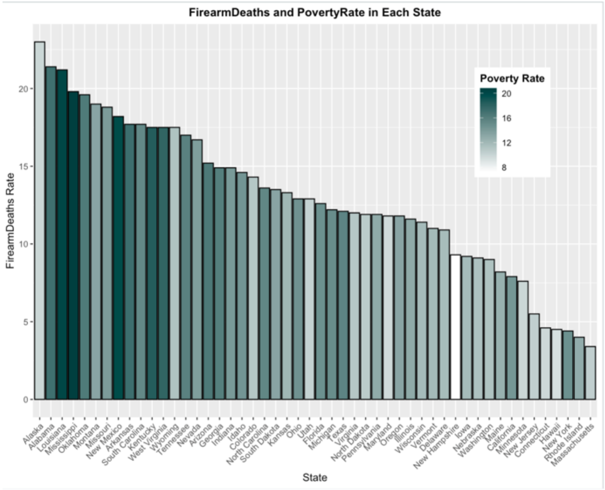
Third

I believe **Poverty Rate** is another factor that affects the **Firearm Deaths Rate**. I took the same way with the previous one and replaced the **Gun Ownership** with **Poverty Rate**.

Apparently, bars on the left are having relatively darker colors. The higher the **Poverty Rate** is, **Firearm Deaths** are relatively more likely to happen in that state.

Note:

Poverty Rate represent the percentage of citizens who are living their life under a specific poverty standard in that state.



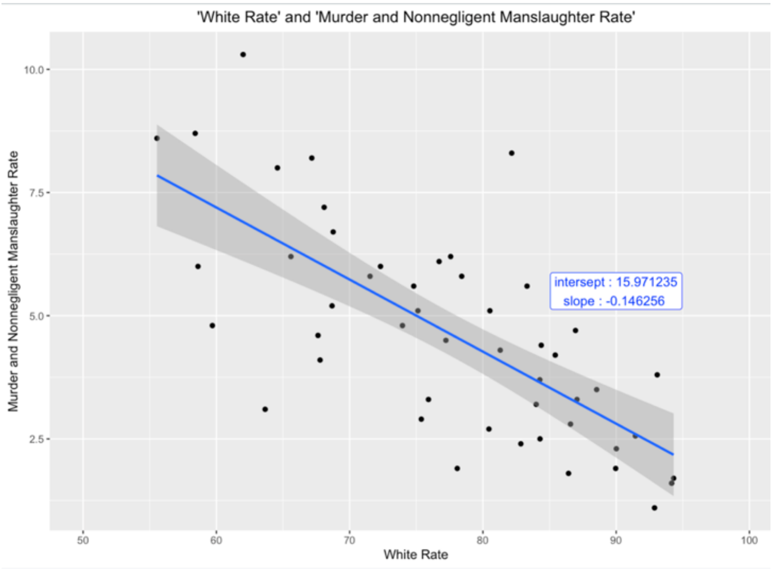
Q2.

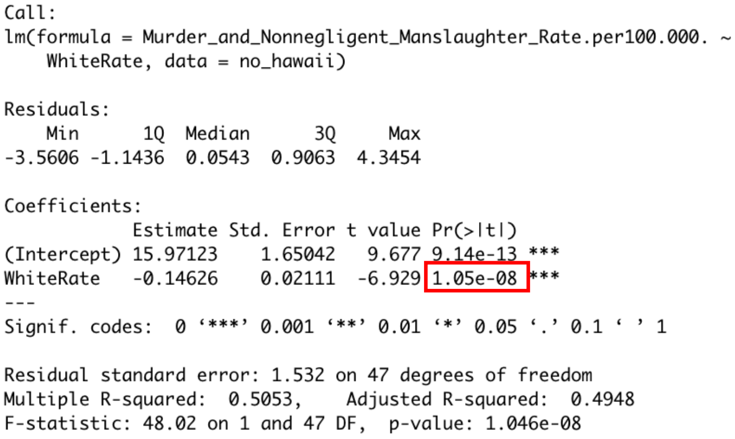
Whether Murder and Nonnegligent Manslaughter or Firearm Deaths is more likely to occur in more racially complex areas or not.

I took not only **Firearm Deaths Rate** but also **Murder and Nonnegligent Manslaughter Rate** aspect to discuss safety in this question.

First

The graph indicates the linear relationship between **White Rate** and **Murder and Nonnegligent Manslaughter Rate**. Since the p-value is less than 0.05, we **reject** the null hypothesis (Ho: beta1 = 0). We can conclude that there is a statistically significant relationship between two variables in the linear regression model.



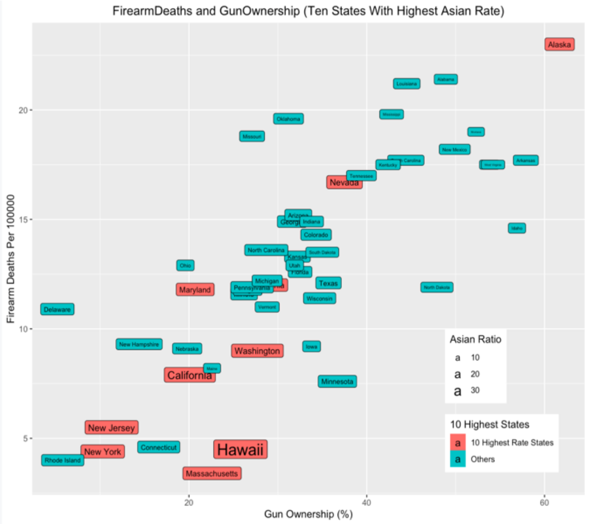


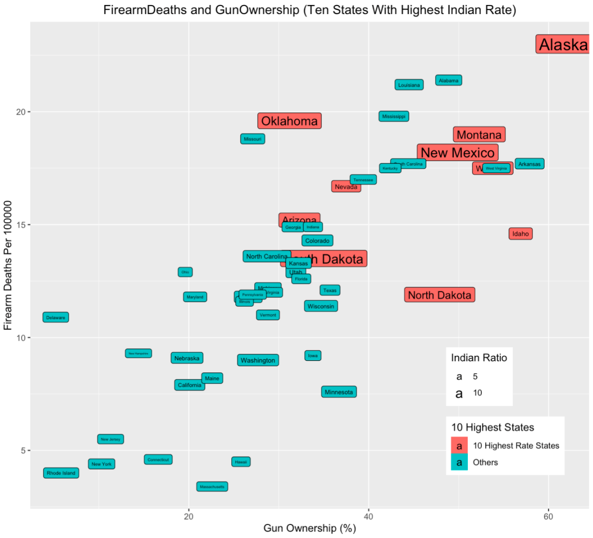
```
Call:
lm(formula = Murder_and_Nonnegligent_Manslaughter_Rate.per100.000. ~
    WhiteRate, data = no_hawaii)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5606 -1.1436  0.0543  0.9063  4.3454

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.97123    1.65042   9.677 9.14e-13 ***
WhiteRate   -0.14626    0.02111  -6.929 1.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.532 on 47 degrees of freedom
Multiple R-squared:  0.5053,    Adjusted R-squared:  0.4948
F-statistic: 48.02 on 1 and 47 DF,  p-value: 1.046e-08
```

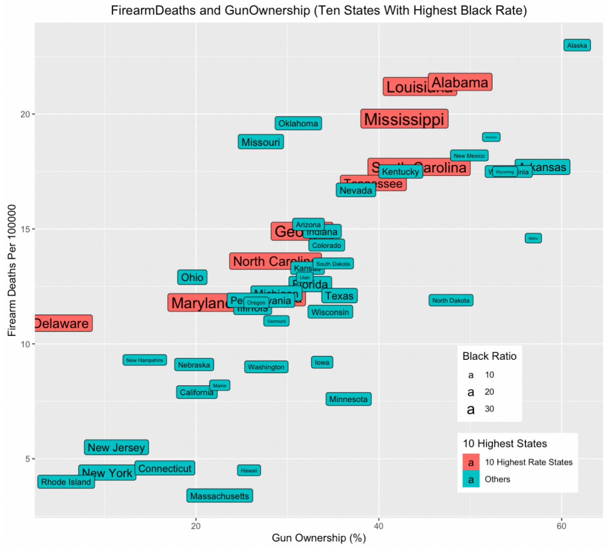
Second

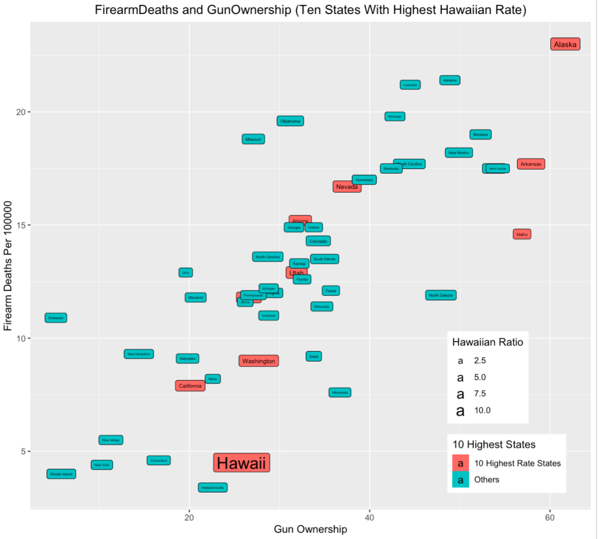
Taking a closer analysis on each race in each state, I found some interesting things in visualizations.

Red labels represent ten states with the highest racial rate out of 50 states, and text size of each label is determined by the corresponding racial rate in each state.









By observing the distribution of top ten states, I found that states with relatively higher **Asian Rate** (top left graph) are having lower **Gun Ownership** and **Firearm Deaths** rates, while states with relatively higher **Indian Rate** (top right graph) are having higher **Gun Ownership** and **Firearm Deaths** rates overall. However, I can't observe the similar trend in **Black Rate** (bottom left graph) and **Hawaiian Rate** (bottom right graph).

Important:

It is noteworthy that although the graph shows that states with higher **Indians Rate** are having comparably higher **Gun Ownership** and **Firearm Deaths Rates**, it doesn't mean Indians caused these two rates to be high. It probably just a coincidence because the data only shows the results instead of showing cause and effect.

5. Conclusion

I applied two different ways to find the relationship among different factors and target variables. In Q1, what surprised me is that I didn't find relationships between **White Rate** and **Firearm Deaths Rate** through the analytics which I assumed completely opposite. On the contrary, the results that **Gun Ownership** and **Poverty Rate** are having something to do with **Firearm Deaths Rate** are exactly what I expected.

In Q2, I discussed different races separately, and found some interesting results in **White**, **Asian** and **Indian Rate.** There is statistical evidence showing the relationship between **White Rate** and **Murder and Nonnegligent Manslaughter Rate**. Additionally, visualized graphs indicate that states with higher **Indian Rates** are more dangerous in **Firearm Deaths** while states with higher **Asian Rates** are less dangerous overall.

6. Reflection

In this project, I not only learned how to practically put what we learned in class into practice but also learned that not every kind of races would be associated with **Firearm Deaths**. Like I assumed **White** and **Black Rate** would have the strongest relationship with it but apparently not. I should have reduced the subjective judgment when doing data exploration so that I could have analyzed the data more neutrally and saved more time on finding useless information.

7. Bibliography

Imputation (statistics) - Wikipedia. (2022). Retrieved 11 April 2022, from https://en.wikipedia.org/wiki/Imputation_(statistics)#Regression