IDS[C] FINAL REPORT

Prepared by

Zaman Shahriaz

TABLE OF CONTENTS

PROJECT OVERVIEW:

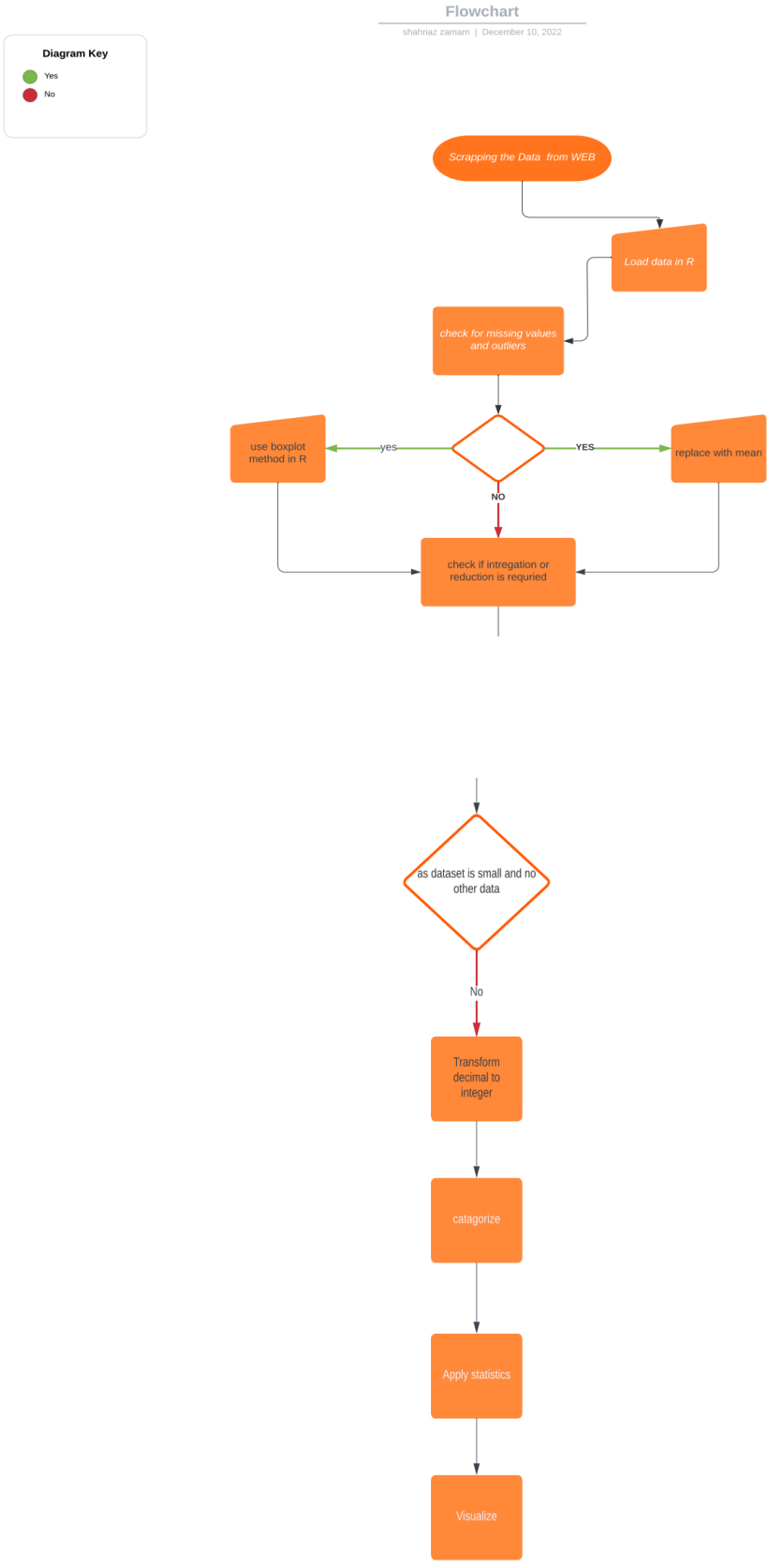
A dataset needed to be scraped from a website using **R-Studio** after scrapping pre-processing and visualizing is applied. In which 50 rows and 8 columns made up the dataset. The dataset concerned information on the number of assault and weapons used in violence in each of the 50 US states concluded by the **FBI**. It also includes the proportion of people who reside in urban areas. The first column listed the names of the states, the second listed the number of aggravated assault, the third to seventh column has listed the number of firearms, knives, other weapon, personal weapons and the eighth[Last] column listed the proportion of population

First the data was scrapped from the web then loaded into the working directory in the

R-Studio.

The dataset needed to be loaded manually in excel file format and again loaded in R-studio in order to preprocess the data that was being contained in the dataset.

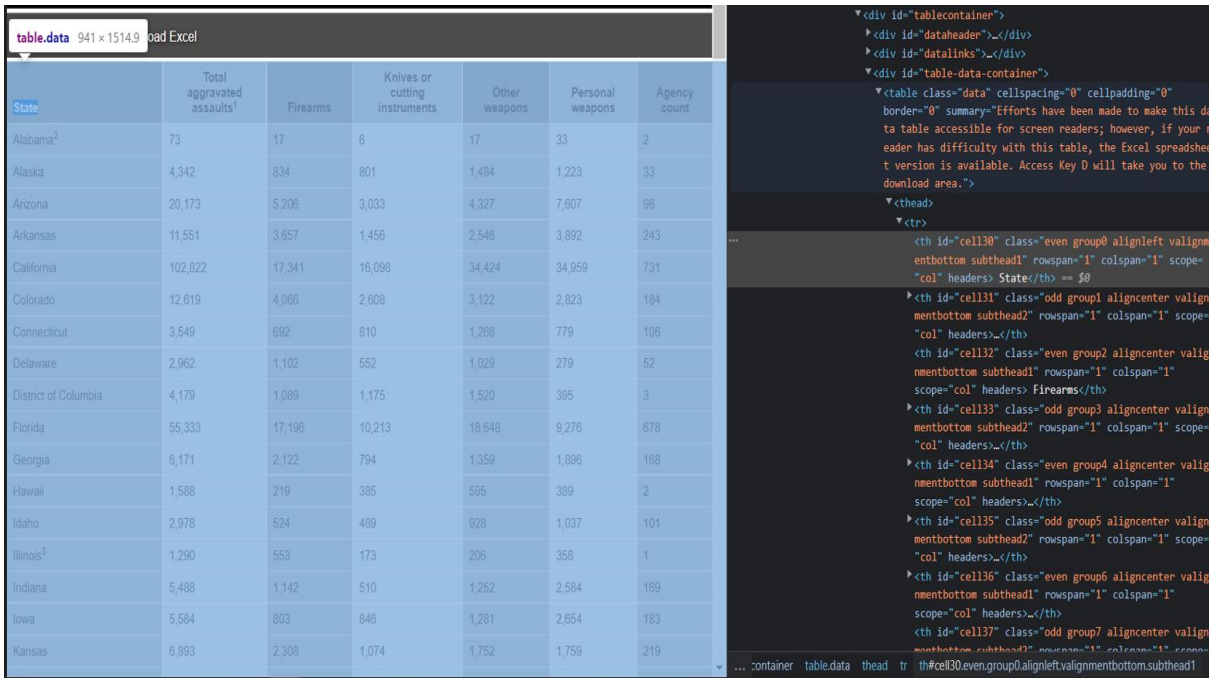
SOLUTION DESIGN:



DATA COLLECTION:

State	Total aggravated assaults[1]	Firearms	Knives or cutting instruments	Other weapons	Personal weapons	Agency count	Population
Alabama[2]	73	17	6	17	33	2	85,670
Alaska	4,342	834	801	1,484	1,223	33	727,792
Arizona	20,173	5,206	3,033	4,327	7,607	96	6,308,226
Arkansas	11,551	3,657	1,456	2,546	3,892	243	2,501,782
California	102,822	17,341	16,098	34,424	34,959	731	38,767,853
Colorado	12,619	4,066	2,608	3,122	2,823	184	4,910,658
Connecticut	3,549	692	810	1,268	779	106	3,492,933
Delaware	2,962	1,102	552	1,029	279	52	970,535
District of Columbia	4,179	1,089	1,175	1,520	395	3	705,749
Florida	55,333	17,196	10,213	18,648	9,276	678	21,424,937
Georgia	6,171	2,122	794	1,359	1,896	168	3,026,506
Hawaii	1,588	219	385	595	389	2	1,142,377
Idaho	2,978	524	489	928	1,037	101	1,751,712
Illinois[3]	1,290	553	173	206	358	1	145,719
Indiana	5,488	1,142	510	1,252	2,584	169	3,004,084
Iowa	5,584	803	846	1,281	2,654	183	2,584,085
Kansas	6,893	2,308	1,074	1,752	1,759	219	1,821,596
Kentucky	5,722	2,160	660	2,323	579	398	4,456,082
Louisiana	16,346	5,699	2,461	4,564	3,622	150	3,988,385
Maine	824	72	147	253	352	135	1,344,212
Maryland	15,727	2,840	3,490	5,898	3,499	151	6,043,312
Massachusetts	16,375	1,666	3,562	7,502	3,645	346	6,726,719
Michigan	29,325	8,993	5,580	8,934	5,818	612	9,492,862
Minnesota	7,338	1,820	1,516	1,954	2,048	368	5,433,467

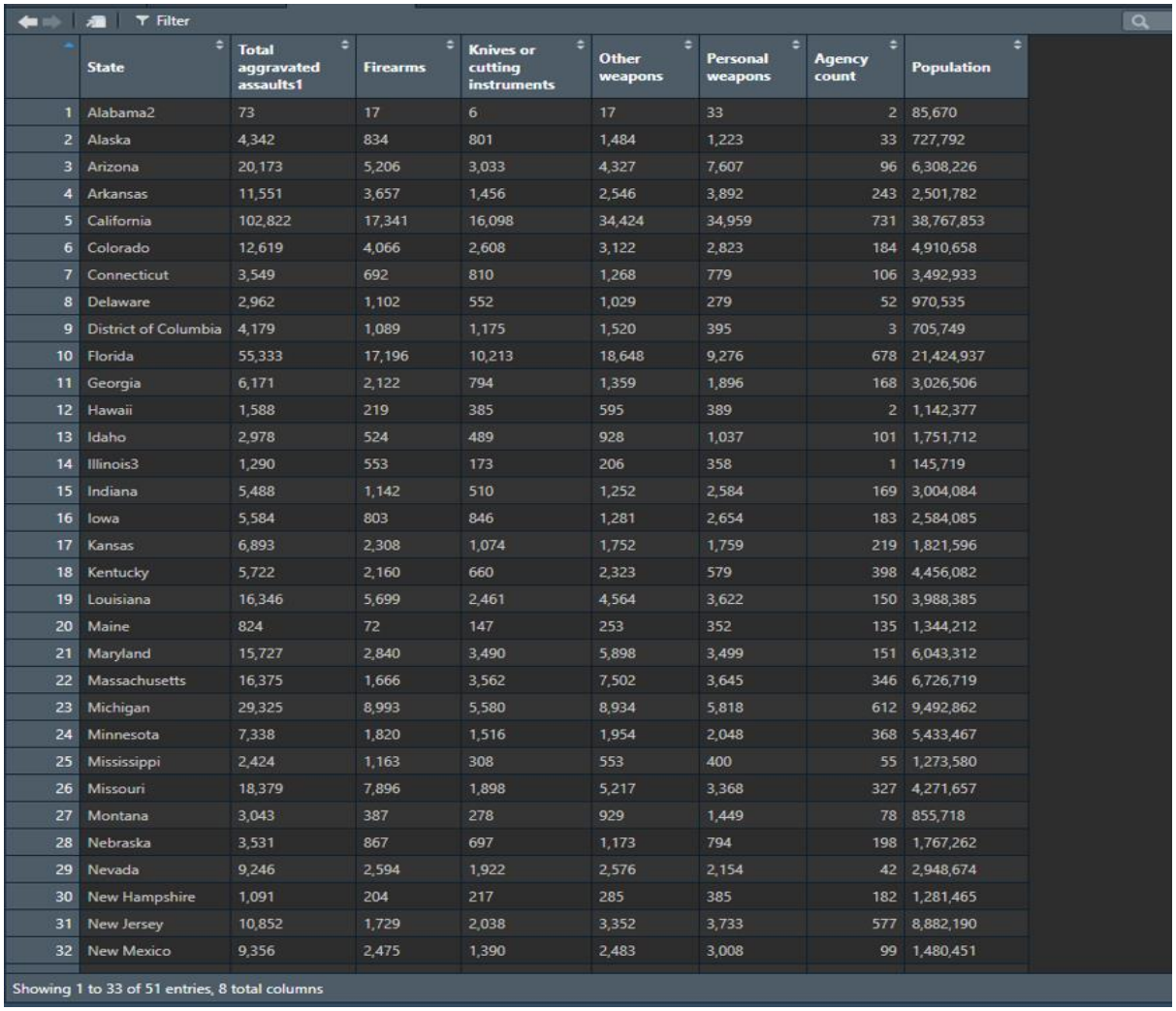
Above image illustrates the raw data in the website in tabular pattern where states are in and attack weapons are in columns.



In the highlighted part the data table is selected form the web page through inspect element and copied into the RStudio panel

```
library(rvest)
library(dplyr)
library(xml2)

col_link = "https://www.patriotsoftware.com/blog/accounting/average-cost-living-by-state/"
col_page = read_html(col_link)

col_table = col_page %>% html_nodes("table.has-fixed-layout") %>%
  html_table() %>% . [[1]]

install.packages("writexl")

View(col_table)
write.csv(col_table, file = "Table_2.csv")

```

By installing library **rvest**, **dplyr** and **xml2** and putting the website URL and table link through pipeline method tables [dataset] is loaded in RStudio



Filter

	State	Total aggravated assaults1	Firearms	Knives or cutting instruments	Other weapons	Personal weapons	Agency count	Population
1	Alabama2	73	17	6	17	33	2	85,670
2	Alaska	4,342	834	801	1,484	1,223	33	727,792
3	Arizona	20,173	5,206	3,033	4,327	7,607	96	6,308,226
4	Arkansas	11,551	3,657	1,456	2,546	3,892	243	2,501,782
5	California	102,822	17,341	16,098	34,424	34,959	731	38,767,853
6	Colorado	12,619	4,066	2,608	3,122	2,823	184	4,910,658
7	Connecticut	3,549	692	810	1,268	779	106	3,492,933
8	Delaware	2,962	1,102	552	1,029	279	52	970,535
9	District of Columbia	4,179	1,089	1,175	1,520	395	3	705,749
10	Florida	55,333	17,196	10,213	18,648	9,276	678	21,424,937
11	Georgia	6,171	2,122	794	1,359	1,896	168	3,026,506
12	Hawaii	1,588	219	385	595	389	2	1,142,377
13	Idaho	2,978	524	489	928	1,037	101	1,751,712
14	Illinois3	1,290	553	173	206	358	1	145,719
15	Indiana	5,488	1,142	510	1,252	2,584	169	3,004,084
16	Iowa	5,584	803	846	1,281	2,654	183	2,584,085
17	Kansas	6,893	2,308	1,074	1,752	1,759	219	1,821,596
18	Kentucky	5,722	2,160	660	2,323	579	398	4,456,082
19	Louisiana	16,346	5,699	2,461	4,564	3,622	150	3,988,385
20	Maine	824	72	147	253	352	135	1,344,212
21	Maryland	15,727	2,840	3,490	5,898	3,499	151	6,043,312
22	Massachusetts	16,375	1,666	3,562	7,502	3,645	346	6,726,719
23	Michigan	29,325	8,993	5,580	8,934	5,818	612	9,492,862
24	Minnesota	7,338	1,820	1,516	1,954	2,048	368	5,433,467
25	Mississippi	2,424	1,163	308	553	400	55	1,273,580
26	Missouri	18,379	7,896	1,898	5,217	3,368	327	4,271,657
27	Montana	3,043	387	278	929	1,449	78	855,718
28	Nebraska	3,531	867	697	1,173	794	198	1,767,262
29	Nevada	9,246	2,594	1,922	2,576	2,154	42	2,948,674
30	New Hampshire	1,091	204	217	285	385	182	1,281,465
31	New Jersey	10,852	1,729	2,038	3,352	3,733	577	8,882,190
32	New Mexico	9,356	2,475	1,390	2,483	3,008	99	1,480,451

Showing 1 to 33 of 51 entries, 8 total columns

	State	Total aggravated assaults1	Firearms	Knives or cutting instruments	Other weapons	Personal weapons	Agency count	Population
21	Maryland	15,727	2,840	3,490	5,898	3,499	151	6,043,312
22	Massachusetts	16,375	1,666	3,562	7,502	3,645	346	6,726,719
23	Michigan	29,325	8,993	5,580	8,934	5,818	612	9,492,862
24	Minnesota	7,338	1,820	1,516	1,954	2,048	368	5,433,467
25	Mississippi	2,424	1,163	308	553	400	55	1,273,580
26	Missouri	18,379	7,896	1,898	5,217	3,368	327	4,271,657
27	Montana	3,043	387	278	929	1,449	78	855,718
28	Nebraska	3,531	867	697	1,173	794	198	1,767,262
29	Nevada	9,246	2,594	1,922	2,576	2,154	42	2,948,674
30	New Hampshire	1,091	204	217	285	385	182	1,281,465
31	New Jersey	10,852	1,729	2,038	3,352	3,733	577	8,882,190
32	New Mexico	9,356	2,475	1,390	2,483	3,008	99	1,480,451
33	New York	43,967	4,535	11,626	12,825	14,981	497	18,382,202
34	North Carolina	21,191	10,813	2,892	3,915	3,571	253	7,678,759
35	North Dakota	1,243	28	141	517	557	103	759,718
36	Ohio	16,947	6,636	3,247	4,931	2,133	414	9,215,222
37	Oklahoma	12,122	3,342	1,956	3,627	3,197	406	3,909,106
38	Oregon	6,933	1,031	1,244	2,382	2,276	151	3,680,010
39	Pennsylvania	5,959	962	599	812	3,586	438	3,297,344
40	Rhode Island	1,408	228	403	444	333	48	1,059,361
41	South Carolina	18,207	8,598	2,675	4,145	2,789	260	4,660,637
42	South Dakota	2,005	277	463	471	794	107	816,216
43	Tennessee	30,687	13,807	5,270	9,587	2,023	456	6,643,476
44	Texas	73,258	25,970	14,156	20,714	12,418	823	27,151,685
45	Utah	4,281	692	812	1,321	1,456	109	2,883,845
46	Vermont	861	131	137	148	445	79	599,538
47	Virginia	10,823	3,535	1,601	3,132	2,555	400	8,403,438
48	Washington	12,990	2,667	2,047	4,311	3,965	219	7,114,406
49	West Virginia	1,996	474	239	453	830	103	953,873
50	Wisconsin	11,351	2,814	1,058	3,090	4,389	409	5,649,131
51	Wyoming	739	102	101	247	289	55	512,713

And finally by installing the package **writexl** it was possible to covert the file into **CSV** format and imported back to RStudio for further pre-processing and visualization.

	A	B	C	D	E	F	G	H	I	J
1		State	Total aggravated assaults1	Firearms	Knives or cutting instruments	Other weapons	Personal weapons	Agency count	Population	
2	1	Alabama2	73	17	6	17	33	2	85,670	
3	2	Alaska	4,342	834	801	1,484	1,223	33	727,792	
4	3	Arizona	20,173	5,206	3,033	4,327	7,607	96	6,308,226	
5	4	Arkansas	11,551	3,657	1,456	2,546	3,892	243	2,501,782	
6	5	California	102,822	17,341	16,098	34,424	34,959	731	38,767,853	
7	6	Colorado	12,619	4,066	2,608	3,122	2,823	184	4,910,658	
8	7	Connecticı	3,549	692	810	1,268	779	106	3,492,933	
9	8	Delaware	2,962	1,102	552	1,029	279	52	970,535	
10	9	District of	4,179	1,089	1,175	1,520	395	3	705,749	
11	10	Florida	55,333	17,196	10,213	18,648	9,276	678	21,424,937	
12	11	Georgia	6,171	2,122	794	1,359	1,896	168	3,026,506	
13	12	Hawaii	1,588	219	385	595	389	2	1,142,377	
14	13	Idaho	2,978	524	489	928	1,037	101	1,751,712	
15	14	Illinois3	1,290	553	173	206	358	1	145,719	
16	15	Indiana	5,488	1,142	510	1,252	2,584	169	3,004,084	
17	16	Iowa	5,584	803	846	1,281	2,654	183	2,584,085	
18	17	Kansas	6,893	2,308	1,074	1,752	1,759	219	1,821,596	
19	18	Kentucky	5,722	2,160	660	2,323	579	398	4,456,082	
20	19	Louisiana	16,346	5,699	2,461	4,564	3,622	150	3,988,385	
21	20	Maine	824	72	147	253	352	135	1,344,212	
22	21	Maryland	15,727	2,840	3,490	5,898	3,499	151	6,043,312	
23	22	Massachus	16,375	1,666	3,562	7,502	3,645	346	6,726,719	
24	23	Michigan	29,325	8,993	5,580	8,934	5,818	612	9,492,862	
25	24	Minnesota	7,338	1,820	1,516	1,954	2,048	368	5,433,467	
26	25	Mississippi	2,424	1,163	308	553	400	55	1,273,580	
27	26	Missouri	18,379	7,896	1,898	5,217	3,368	327	4,271,657	
28	27	Montana	3,043	387	278	929	1,449	78	855,718	
29	28	Nebraska	3,531	867	697	1,173	794	198	1,767,262	
30	29	Nevada	9,246	2,594	1,922	2,576	2,154	42	2,948,674	
31	30	New Hamp	1,091	204	217	285	385	182	1,281,465	
32	31	New Jerse	10,852	1,729	2,038	3,352	3,733	577	8,882,190	
33	32	New Mexi	9,356	2,475	1,390	2,483	3,008	99	1,480,451	

DATA PRE PROCESSING STEPS:

Data Cleaning:

The data set had some noisy value or corrupted data which was manipulated, and organized

Handling Missing Data:

Missing data in a database can arise for a number of reasons, including the absence of a value for that column or the failure to capture the data at the time of data collection, but the dataset did not had any missing value as codes ran through showed valid results.

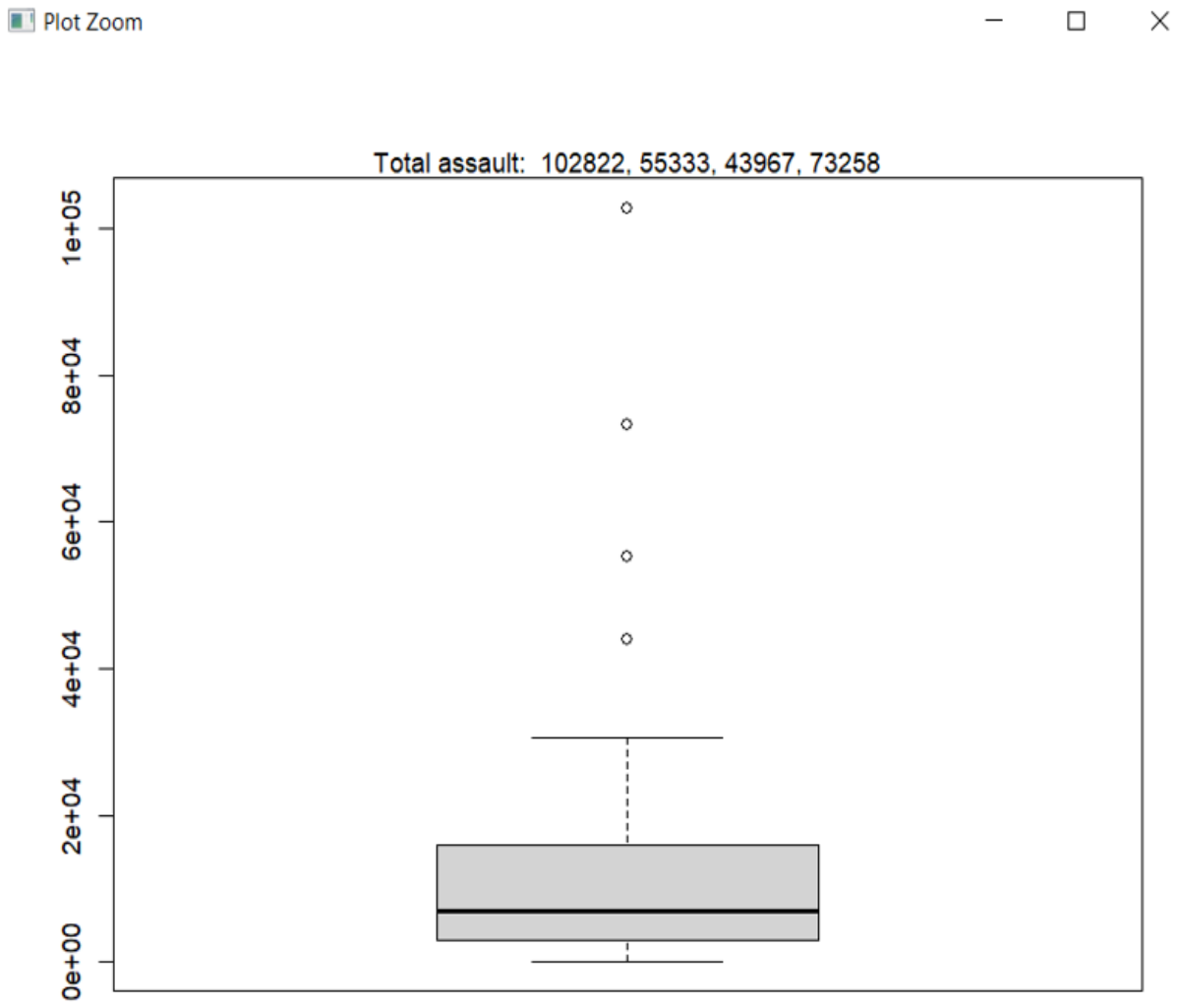
```
sum(is.na(Table_17$`Total aggravated assaults1`))
sum(is.na(Table_17$Firearms))
sum(is.na(Table_17$`Other weapons`))
sum(is.na(Table_17$`Personal weapons`))
sum(is.na(Table_17$`Agency count`))
sum(is.na(Table_17$Population))
```

```
> sum(is.na(Table_17$`Other weapons`))
[1] 0
> sum(is.na(Table_17$`Total aggravated assaults1`))
[1] 0
> sum(is.na(Table_17$`Personal weapons`))
[1] 0
> sum(is.na(Table_17$Population))
[1] 0
> sum(is.na(Table_17$`Agency count`))
[1] 0
```

Smoothing Noisy Data: Here, by using **Boxplot** and **mtext** function it is seen that some columns have outliers. Mean function was used to get rid of the outliers.

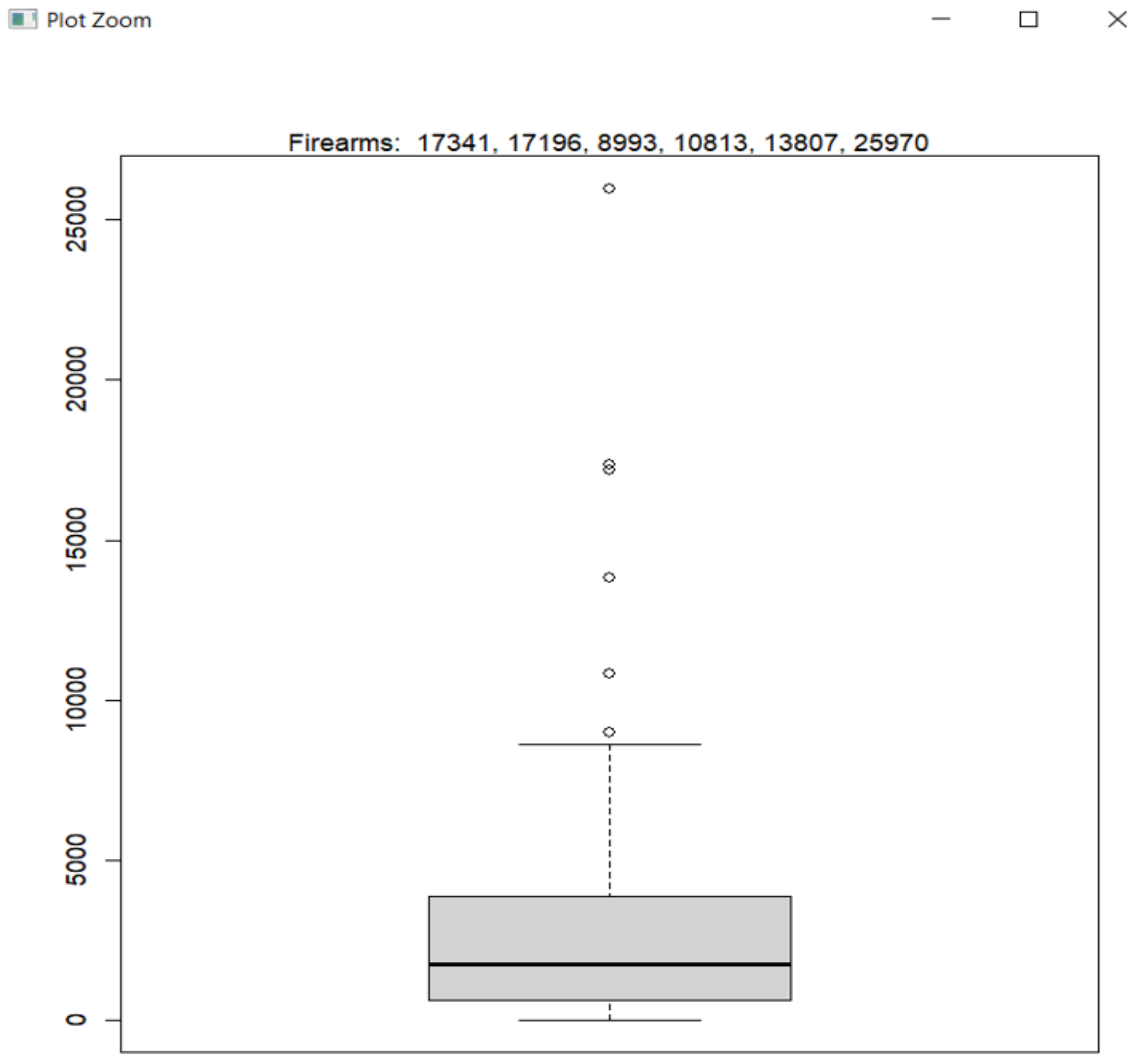
Total Assault

```
boxplot(Table_17$`Total aggravated assaults1`)
out <- boxplot.stats(Table_17$`Total aggravated assaults1`)$out
mtext(paste("Total assault: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 102822] <- mean(Table_17$`Total aggravated assaults1`)
Table_17[Table_17 == 55333] <- mean(Table_17$`Total aggravated assaults1`)
Table_17[Table_17 == 43967] <- mean(Table_17$`Total aggravated assaults1`)
Table_17[Table_17 == 73258] <- mean(Table_17$`Total aggravated assaults1`)
Table_17[Table_17 == 29325] <- mean(Table_17$`Total aggravated assaults1`)
Table_17[Table_17 == 30687] <- mean(Table_17$`Total aggravated assaults1`)
```



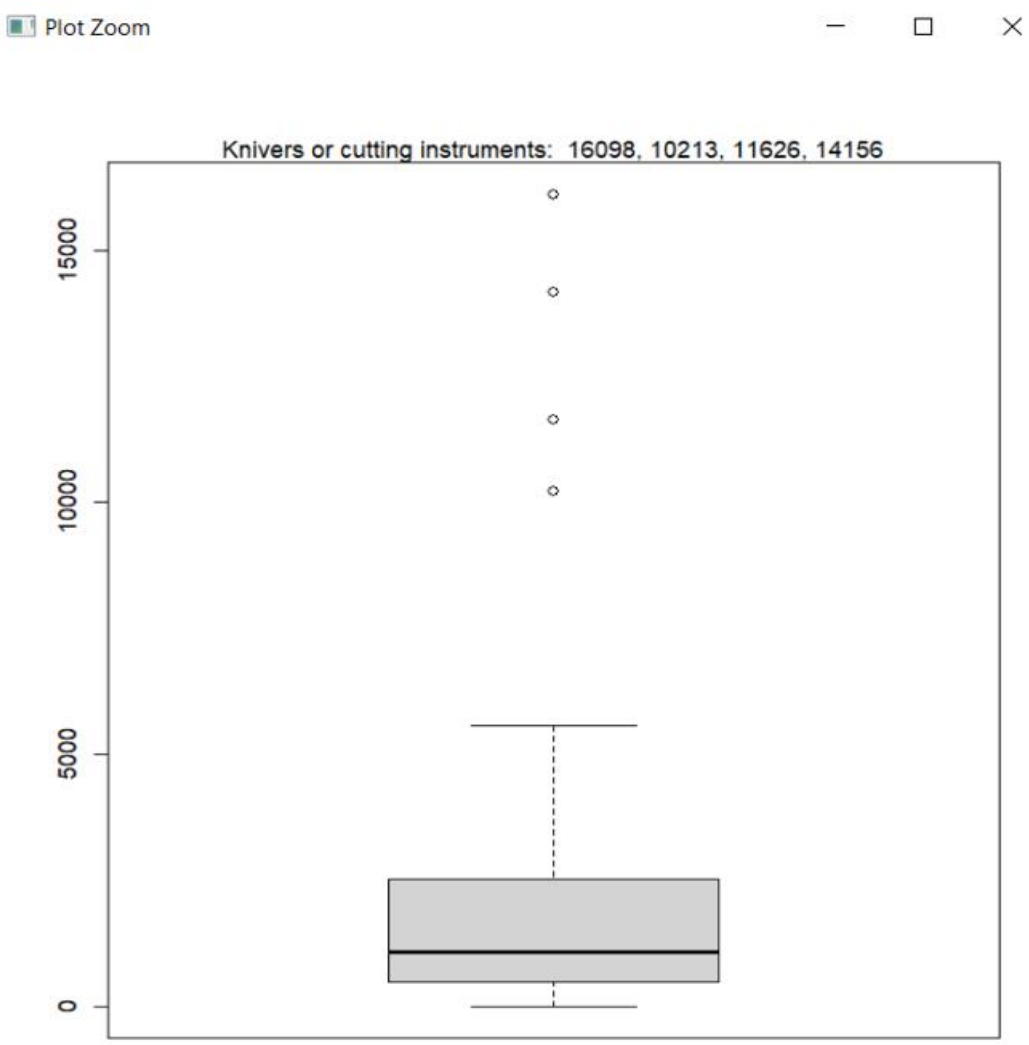
Firearms

```
boxplot(Table_17$Firearms)
out <- boxplot.stats(Table_17$Firearms)$out
mtext(paste("Firearms: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 8993] <- mean(Table_17$Firearms)
Table_17[Table_17 == 10813] <- mean(Table_17$Firearms)
Table_17[Table_17 == 13807] <- mean(Table_17$Firearms)
Table_17[Table_17 == 17196] <- mean(Table_17$Firearms)
Table_17[Table_17 == 17341] <- mean(Table_17$Firearms)
Table_17[Table_17 == 25970] <- mean(Table_17$Firearms)
Table_17[Table_17 == 7896] <- mean(Table_17$Firearms)
Table_17[Table_17 == 8598] <- mean(Table_17$Firearms)
Table_17[Table_17 == 6636] <- mean(Table_17$Firearms)
```



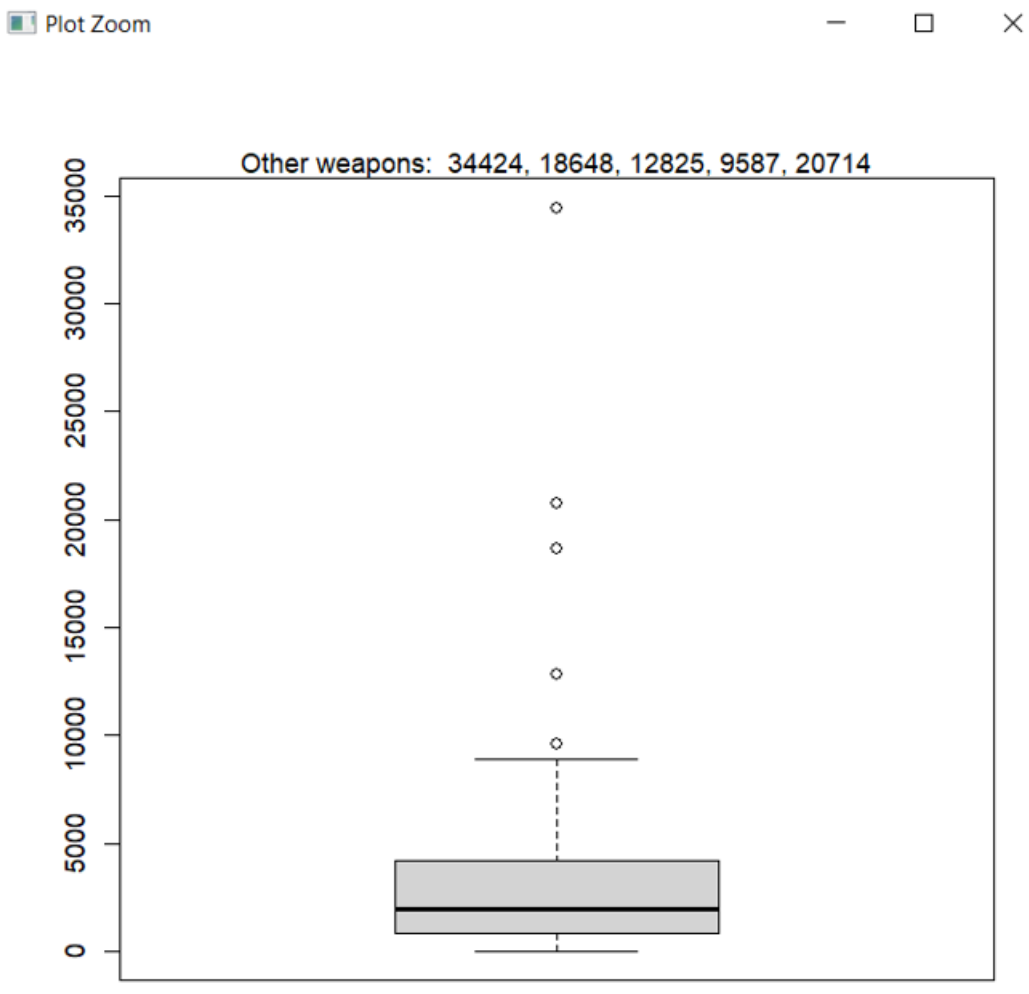
Knives

```
boxplot(Table_17$`Knives or cutting instruments`)
out <- boxplot.stats(Table_17$`Knives or cutting instruments`)$out
mtext(paste("Knivers or cutting instruments: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 16098] <- mean(Table_17$`Knives or cutting instruments`)
Table_17[Table_17 == 10213] <- mean(Table_17$`Knives or cutting instruments`)
Table_17[Table_17 == 11626] <- mean(Table_17$`Knives or cutting instruments`)
Table_17[Table_17 == 14156] <- mean(Table_17$`Knives or cutting instruments`)
Table_17[Table_17 == 5580] <- mean(Table_17$`Knives or cutting instruments`)
Table_17[Table_17 == 5270] <- mean(Table_17$`Knives or cutting instruments`)
```



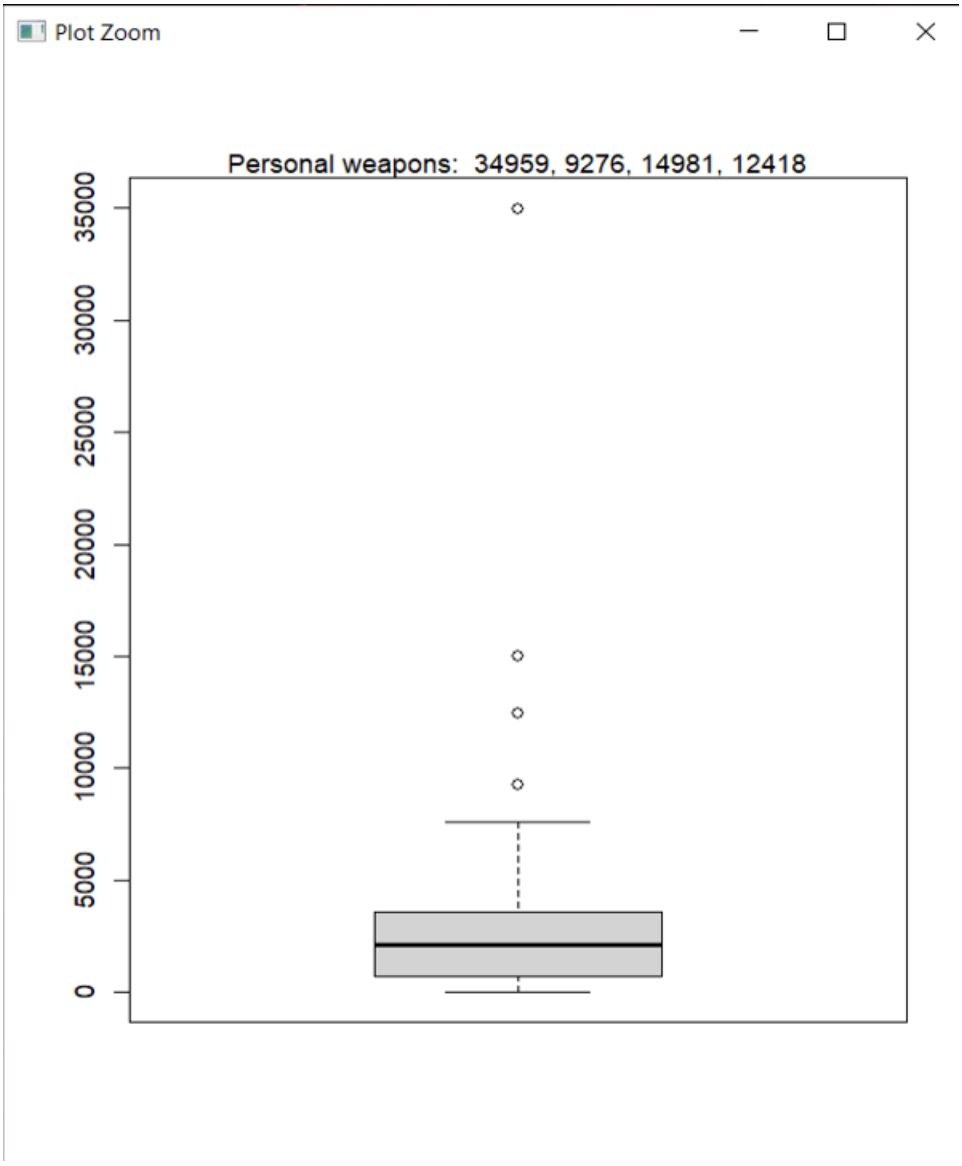
Other Weapons

```
boxplot(Table_17$`Other weapons`)
out <- boxplot.stats(Table_17$`Other weapons`)$out
mtext(paste("Other weapons: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 34424] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 18648] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 12824] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 9587] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 20714] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 7502] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 8934] <- mean(Table_17$`Other weapons`)
Table_17[Table_17 == 12825] <- mean(Table_17$`Other weapons`)
```



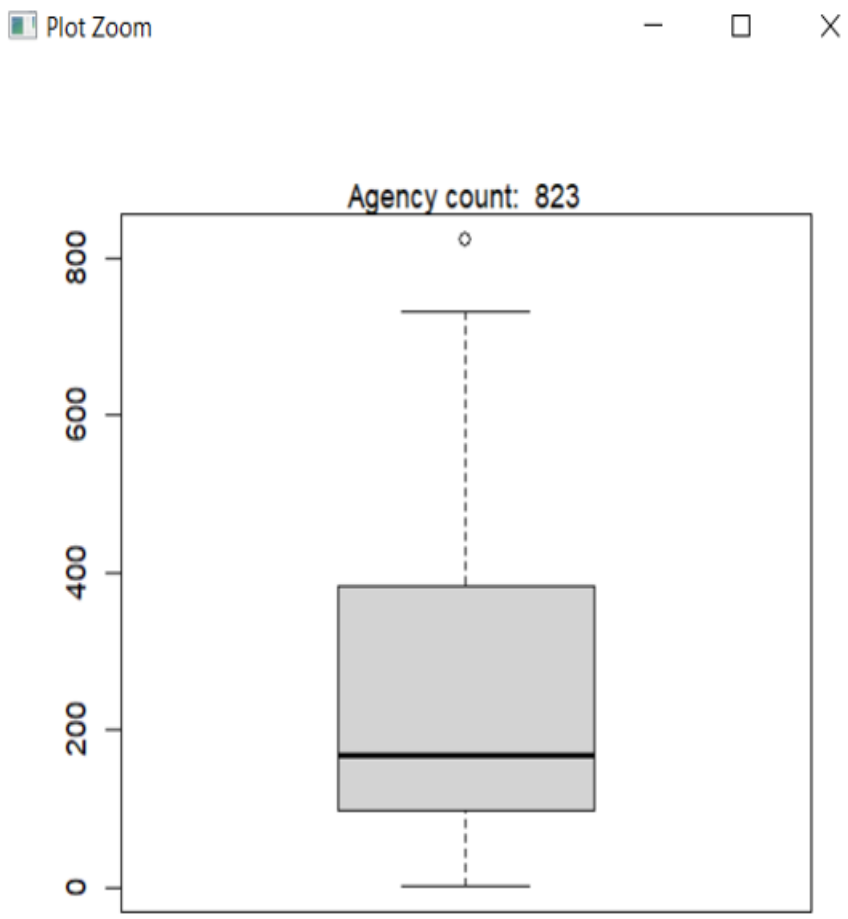
Personal Weapon

```
boxplot(Table_17$`Personal weapons`)
out <- boxplot.stats(Table_17$`Personal weapons`)$out
mtext(paste("Personal weapons: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 34959] <- mean(Table_17$`Personal weapons`)
Table_17[Table_17 == 9276] <- mean(Table_17$`Personal weapons`)
Table_17[Table_17 == 14981] <- mean(Table_17$`Personal weapons`)
Table_17[Table_17 == 12418] <- mean(Table_17$`Personal weapons`)
Table_17[Table_17 == 7607] <- mean(Table_17$`Personal weapons`)
```



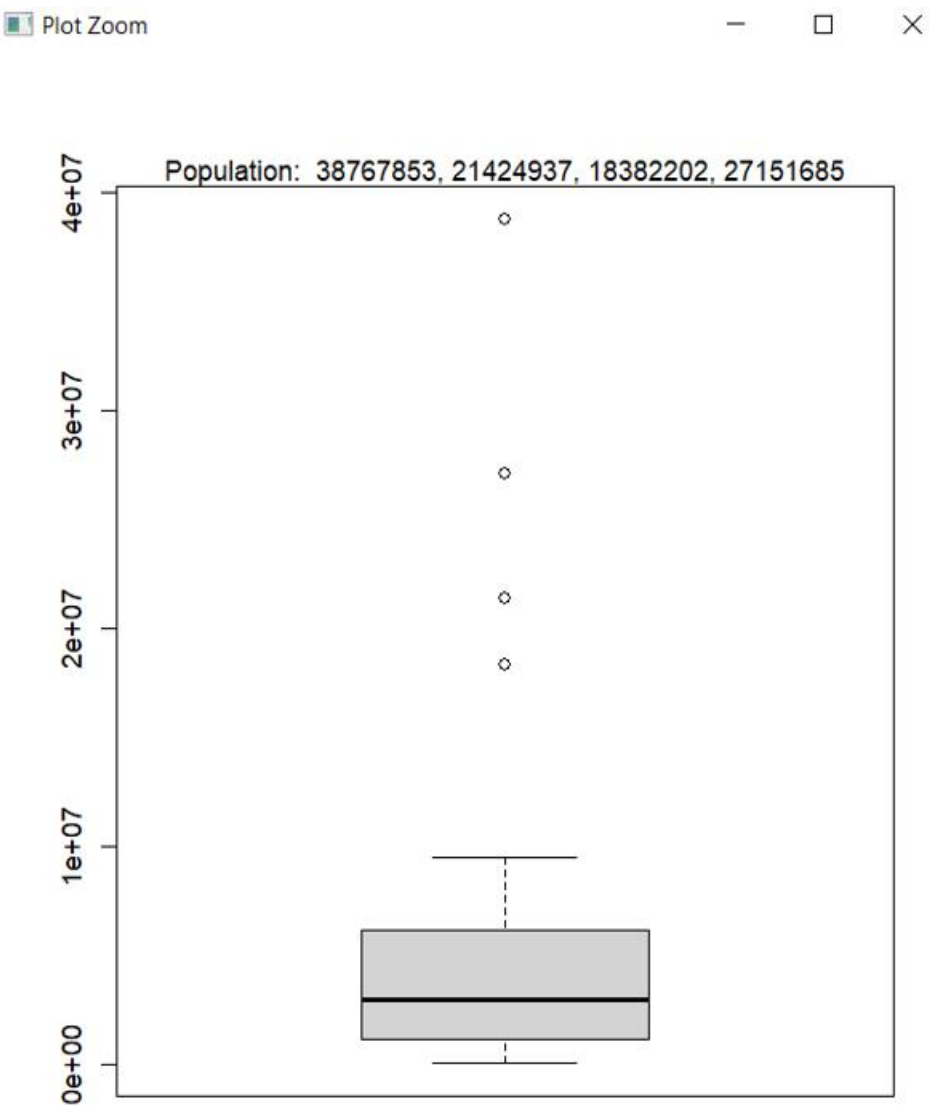
Agency

```
boxplot(Table_17$`Agency count`)
out <- boxplot.stats(Table_17$`Agency count`)$out
mtext(paste("Agency count: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 823] <- mean(Table_17$`Agency count`)
```



Population

```
boxplot(Table_17$Population)
out <- boxplot.stats(Table_17$Population)$out
mtext(paste("Population: ", paste(out, collapse = ", ")))
Table_17[Table_17 == 38767853] <- mean(Table_17$Population)
Table_17[Table_17 == 21424937] <- mean(Table_17$Population)
Table_17[Table_17 == 18382202] <- mean(Table_17$Population)
Table_17[Table_17 == 27151685] <- mean(Table_17$Population)
```



Data Wrangling: Since cleaning step is complete there is no need of data wrangling.

Data Integration: There is no other data set to integrate into the data so the step is skipped.

Data Transformation: . Finally values were transformed into integer using **celling** decimal numbers for assault is irrational.

```
Table_17$`Total aggravated assaults1`=lapply(ceiling(Table_17$`Total aggravated assaults1`),as.integer)
```

Data Reduction: This step was omitted since the amount of data was so minimal.

Data Discretization:

Dividing the data into groups based on population size: According to the size of the population, the data was categorized. Small (50%), medium (60%), large (>70%), and extra-large (>70%) were the categories used to group the data. [dplyr] library and mutate function was used to manipulate the data.

```
library(dplyr)
Table_17<-Table_17 %>% mutate(Type =
                              case_when(Population  <  86670 ~    "Small",
                                        Population  <  866718 ~   "Medium",
                                        Population  <  9910658 ~  "Large",
                                        Population  >= 15000000 ~ "Extra Large")
)
View(Table_17)
```

```
> library(dplyr)
> Table_17<-Table_17 %>% mutate(Type =
+                               case_when(Population  <  86670 ~    "Small",
+                                         Population  <  866718 ~   "Medium",
+                                         Population  <  9910658 ~  "Large",
+                                         Population  >= 15000000 ~ "Extra Large")
+ )
> View(Table_17)
>
```

Descriptive Statistics:

Mean, Median, Variance, Standard Deviation, Inter Quartile Range of multiple columns in the dataset were illustrated through RStudio.

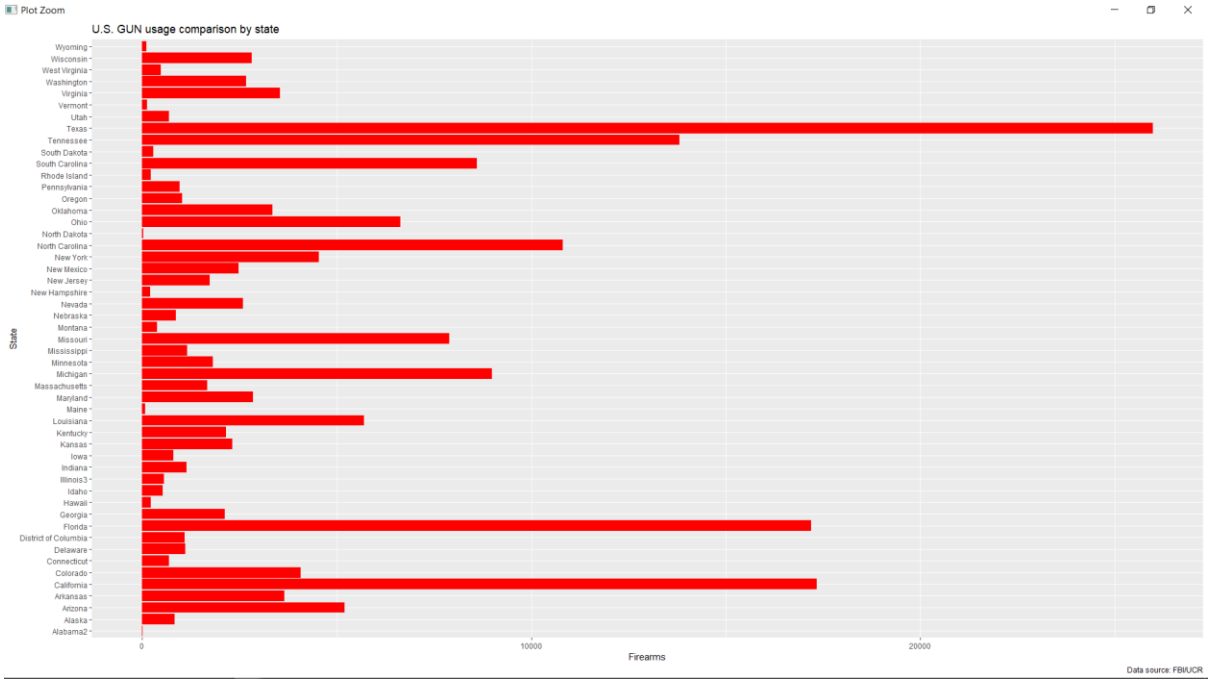
```
#mean
mean(Table_17$Population)
mean(Table_17$Firearms)

#median
median(Table_17$`Agency count`)

#Range
max(Table_17$Population)-min(Table_17$Population)

#variance
var(Table_17$Population)
var(Table_17$`Agency count`)

#standard deviation
sd(Table_17$Population)
sd(Table_17$`Agency count`)

#Quantile
quantile(Table_17$Population)
quantile(Table_17$`Agency count`)

#InterQuartile
IQR(Table_17$Population)
IQR(Table_17$`Agency count`)
```

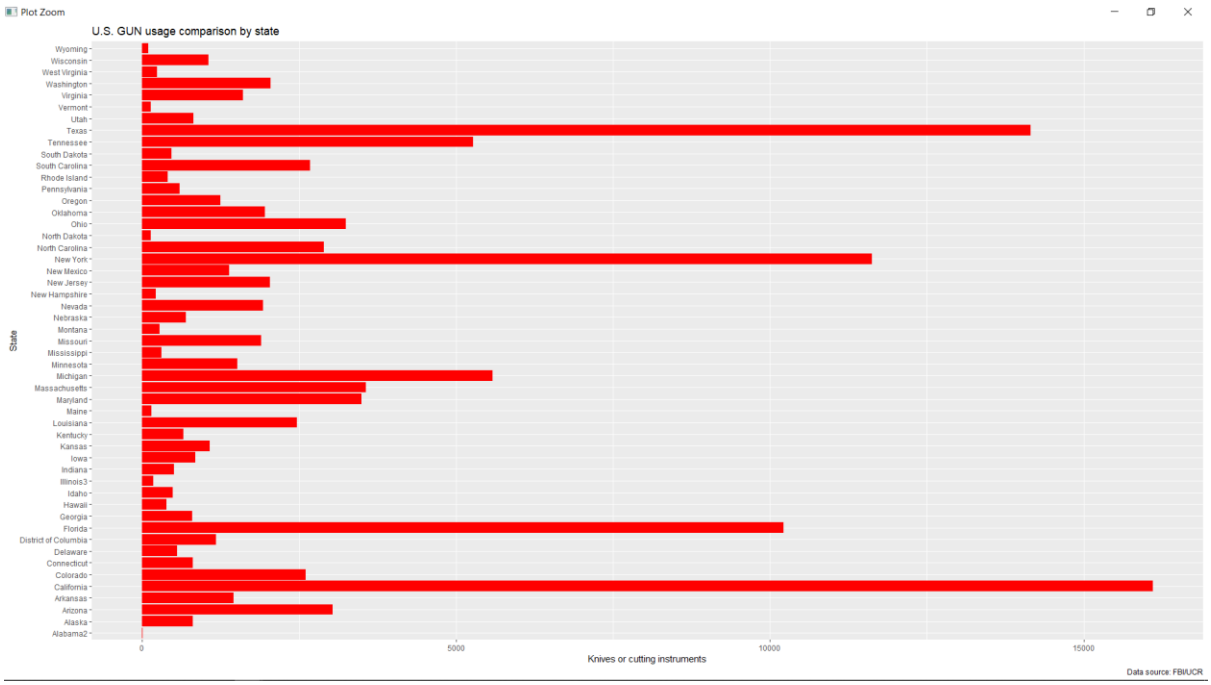
```
> mean(Table_17$Population)
[1] 3515533
> median(Table_17$`Agency count`)
[1] 169
> max(Table_17$Population)-min(Table_17$Population)
[1] 9407192
> var(Table_17$Population)
[1] 6.753872e+12
> var(Table_17$`Agency count`)
[1] 34120.82
> sd(Table_17$Population)
[1] 2598821
> sd(Table_17$`Agency count`)
[1] 184.7182
> quantile(Table_17$`Agency count`)
   0%   25%   50%   75%  100%
  1.0  97.5 169.0 357.0 731.0
> IQR(Table_17$Firearms)
[1] 2204.5
> IQR(Table_17$`Agency count`)
[1] 259.5
> mean(Table_17$Firearms)
[1] NA
```

Visualization: Library **ggplot2** and **mosaic-Data** were used to visualize the following graph between all 8 column present in the table Firearms and Knives graphs are illustrated via red whilst personal and other weapon are illustrated via Orange and so on.

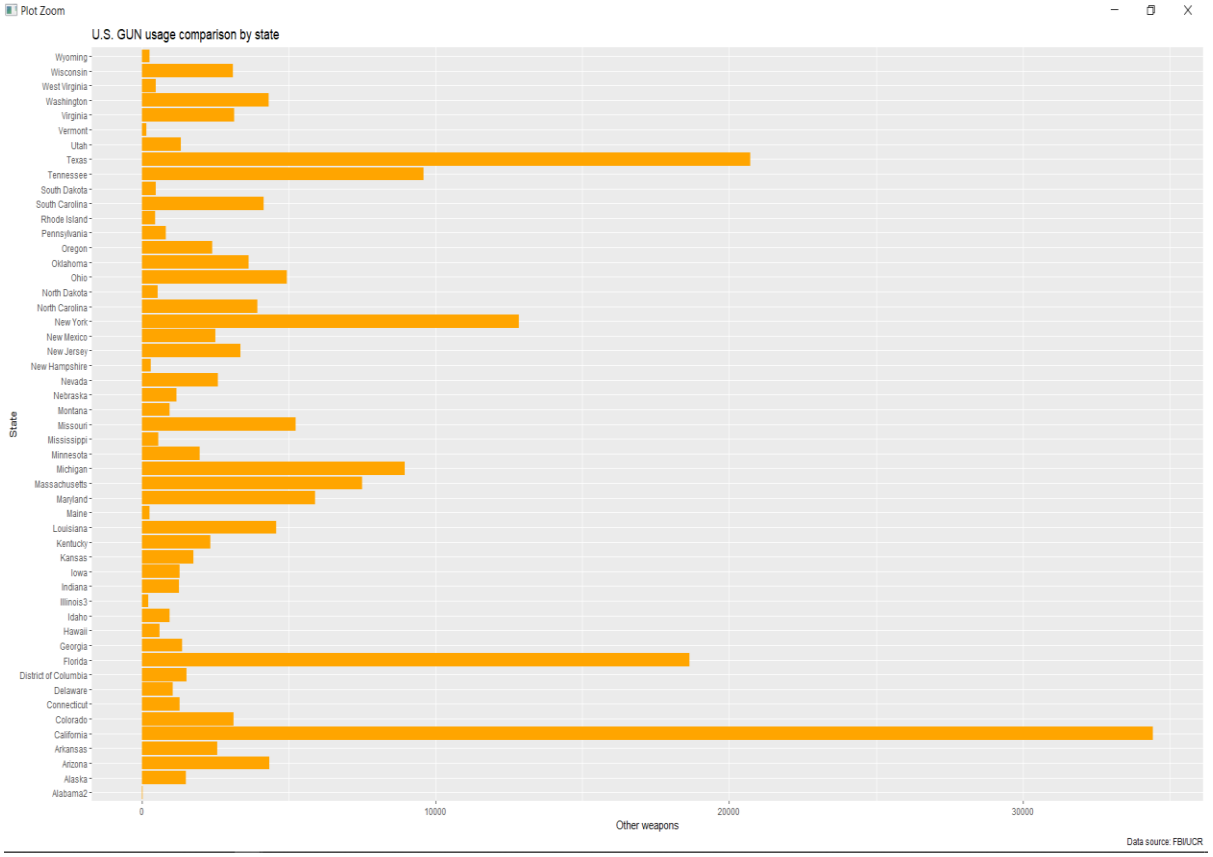
```
library(ggplot2)
library(ggplot2)
library(mosaicData)

ggplot(Table_17)+geom_bar(aes(x = State, y = Firearms),stat="identity" ,fill = "green")+
  labs(title = "U.S. GUN usage comparison by state",caption = "Data source: FBI/UCR") + coord_flip()
```

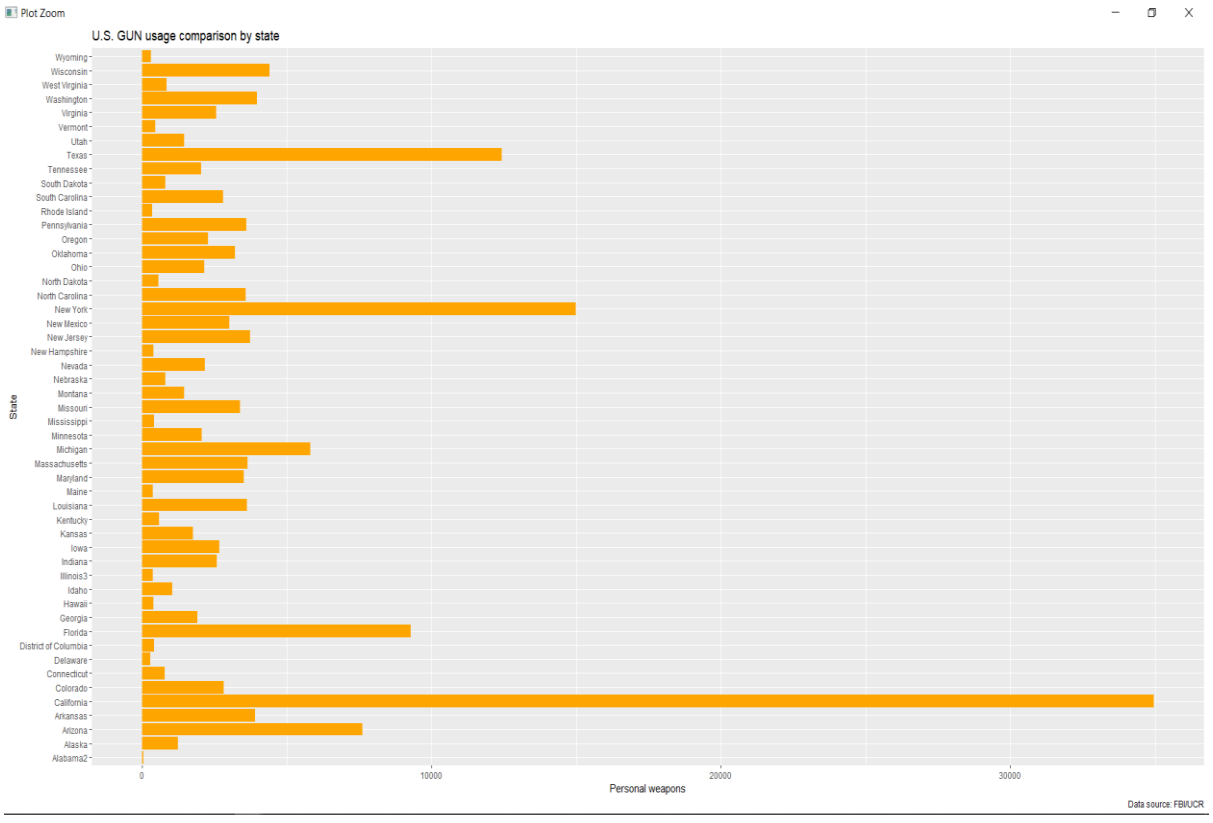
State verses firearm verses knives



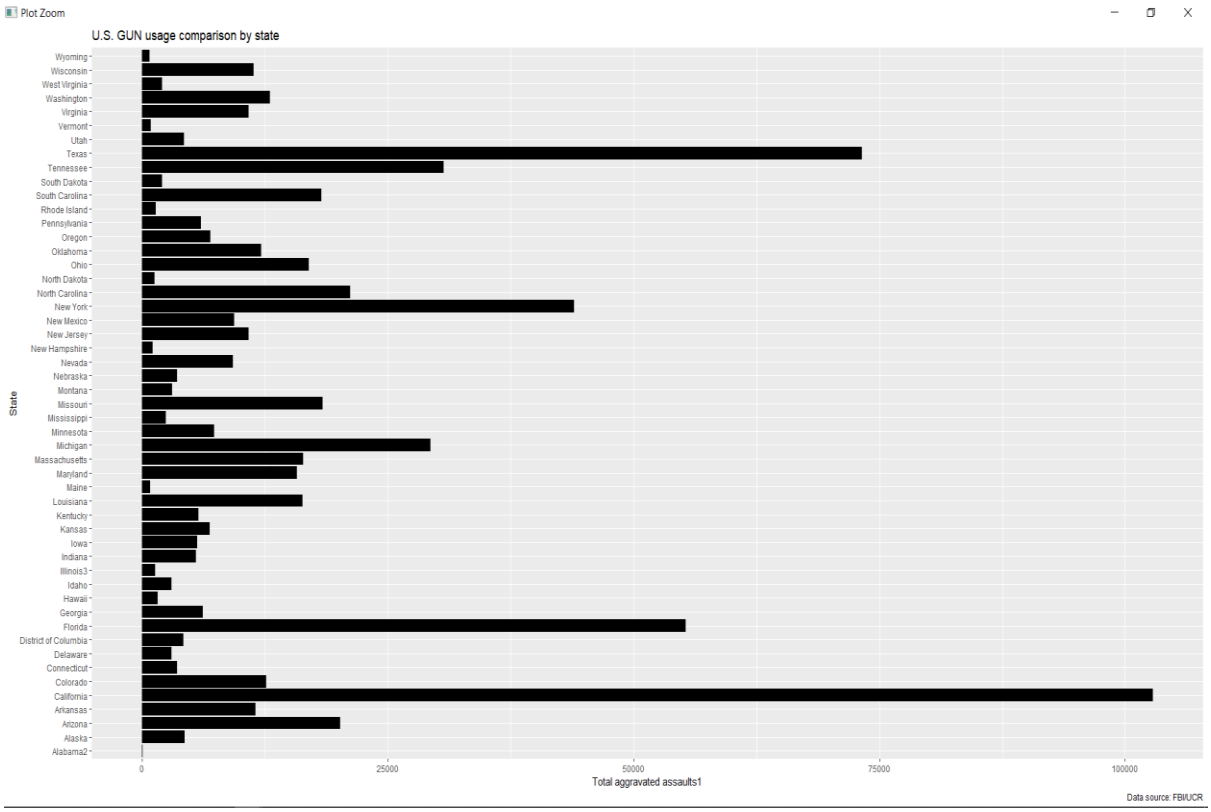


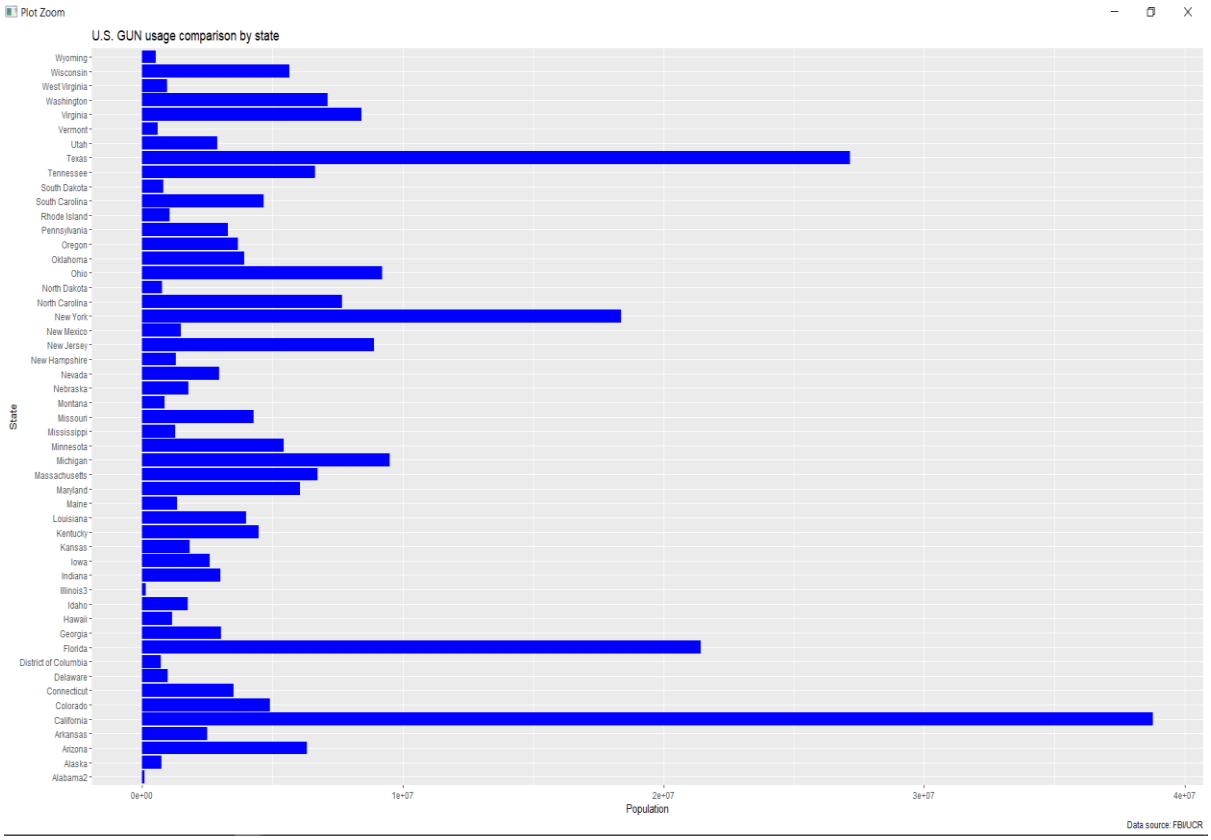
State verses other weapon and personal weapon



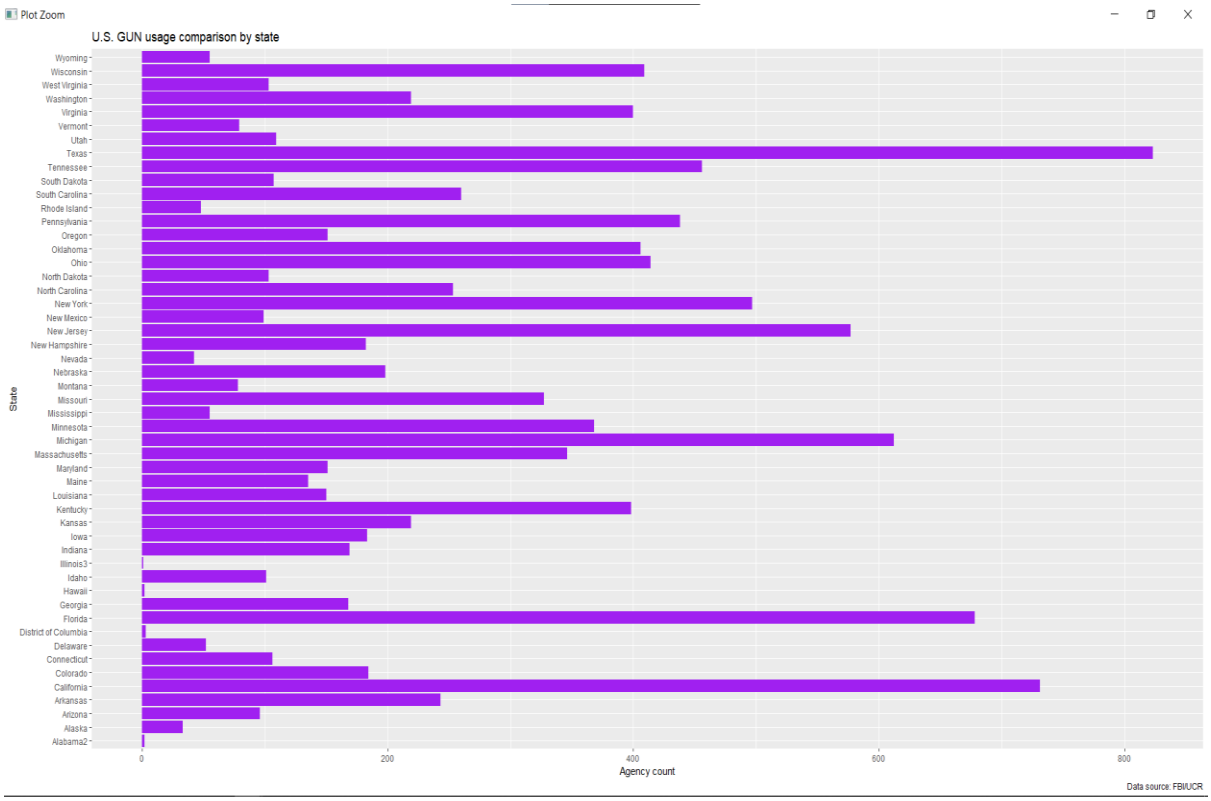


Total Assault comparison in different states with various Population





State Agency count



PROPOSED AMENDMENT:

The statistical analysis language R is quite potent. This project provided a great opportunity to become familiar with R's different features. One of the most crucial pillars of data science is the pre-processing of data since accurate analysis is extremely difficult to perform without accurate data.

CONCLUSION:

Data pre-processing was the topic of this study, which is a crucial component of data science. The provided data was thoroughly processed using a variety of R language tools. The stored cleaned dataset has 50 rows and 8 columns. The dataset is now properly organized and visualized in order to make everyone understand the focal points of the collected data with minimal effort.