Violent Analysis

Code

JianhaoYan

Dec 15, 2018

1 Introduction

Gun violence is a real problem in the USA. So the research on gun violence in America is necessary. The primary purpose of this project is to have an overview of relationships between amounts of gun violence and population. What's more, the dataset we use has many text information. By using this information, we can know how gun violence affects people's health. What's more, we can explore the characteristics of shooters which give government advice to prevent tragedy from happening.

2 Loading the libraries & dataset

We load the dataset and take a look at the dimensions.

The dataset variables are explained as follows:

- Incident Id - Unique ID to each incident
- date - Date of incident occurence
- State - State where incident happened
- City or Country - Country or city where incident took place
- Address - Location of crime
- n_killed - Number of people killed
- n_injured - Number of people injured
- incident_url - URL Describing the incident
- source_url - Same as incident_url ('dataset description doesnt provide much info.Therefore we assume')
- incident_url_fields_missing- Logical indicating whether incident URL is present or not ('dataset description doesnt provide much info.Therefore we assume')
- congressional_district - District number (assume)
- gun_stolen- Status indicating whether the person had his/her gun stolen.
- gun_type - Type of gun
- incident_characteristics - Description of incident.
- latitute- Latitude of the area where crime happened.
- location_description - Place where incident took place.
- longitute - Longitude of the area where crime happened.
- n_guns_involved - Total guns involved in crime
- notes - Comments

- participant_age - Age of people involved
- participant_age_group - Age bracket of the people involved
- participant_gender- Gender of involved people
- participant_name - Name of the involved people
- Participant_relationship - Relationship of the group
- participant_status - status of the people - either arrested,injured or killed in the incident
- participant_type - Either he is a victim or suspect
- sources - Source of the incident information
- state_house_district - state house district Number
- state_senate_district - State senate district number

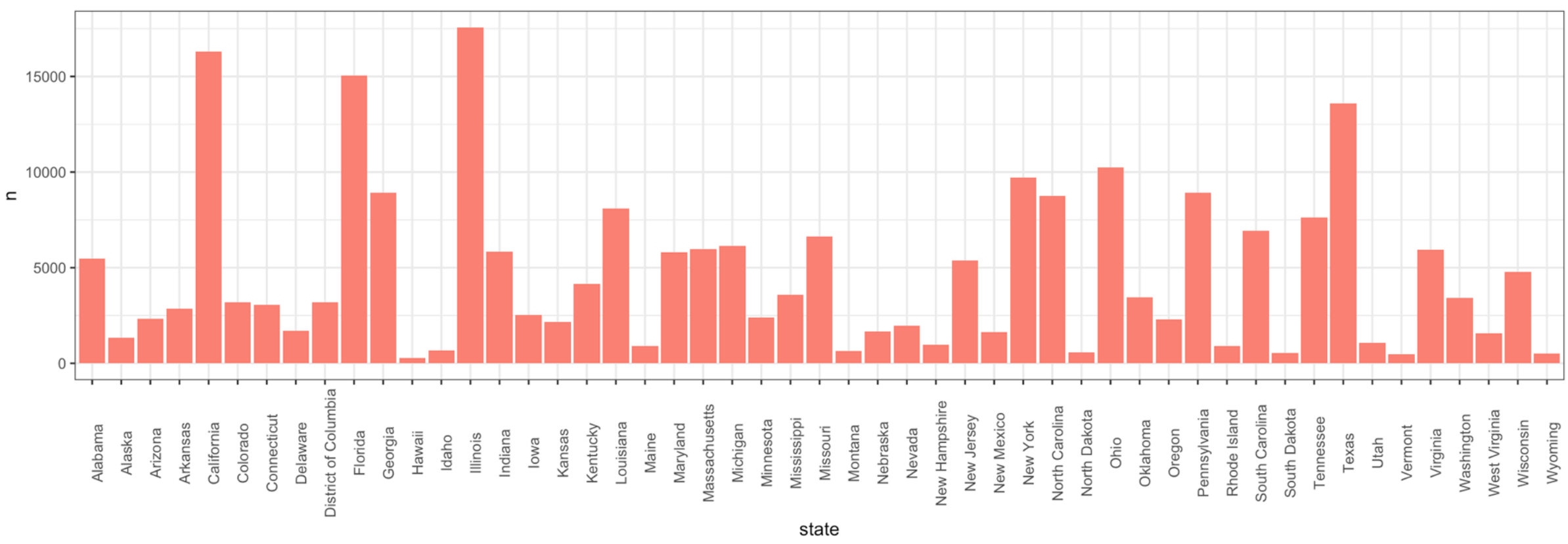
3 Data Cleaning

```
## [1] "2013-01-01" "2013-01-01" "2013-01-01" "2013-01-05" "2013-01-07"
## [6] "2013-01-07"
```

4 Exploratory Data Analysis:

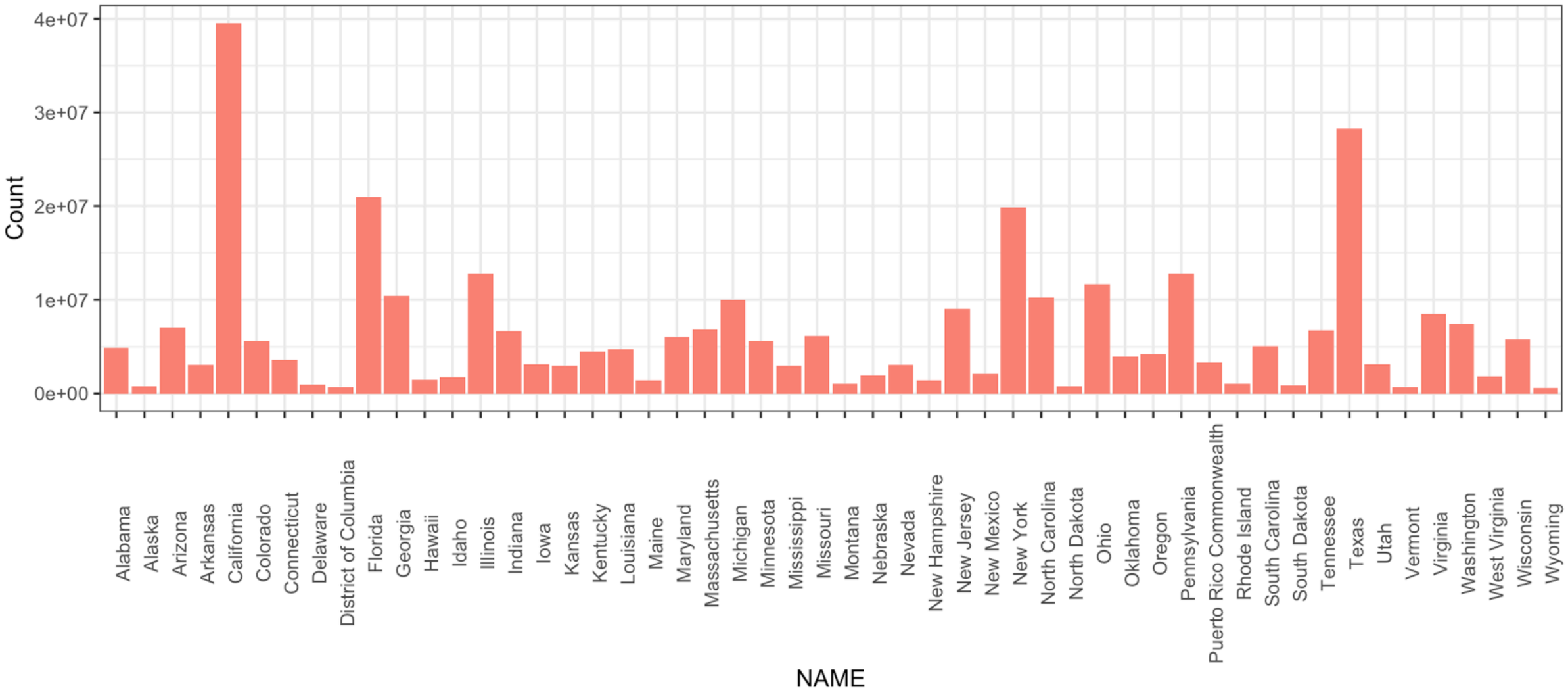
4.1 Distribution of Incidents and gun_violence over state:

We want to get a birdseye view of the number of incidents that have taken place over the state. So this part we maily visulize the relationships between amounts of gun violence and states.



We find that the states - Illinos,California,Texas have higher incidents of gun shooting. And in our thought, these states have high population. So is there any relationship existing between population and gun violence? Let’s do some explorations.

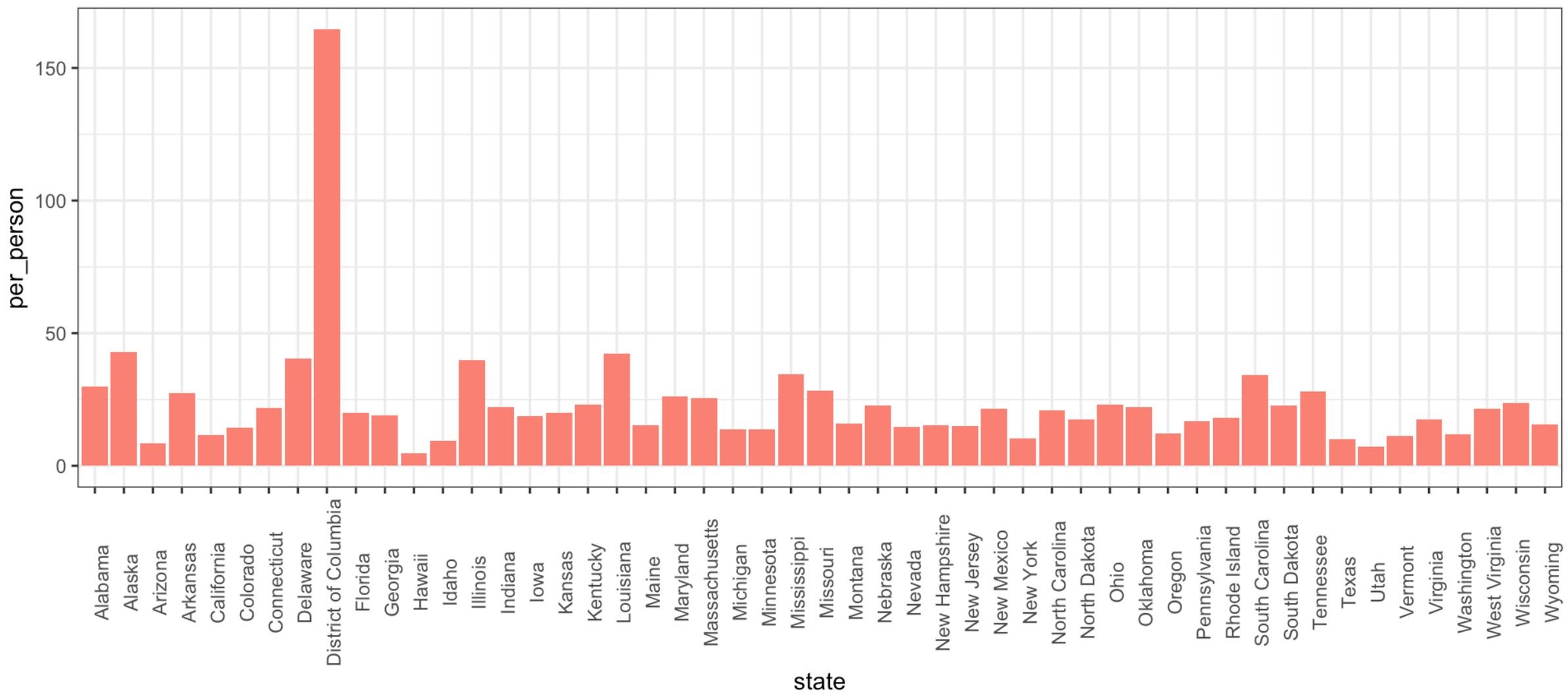
4.2 Population



It is obvious that California, Illinos and Texas have higher populations than other states. So more population can cause more gun violences.

4.3 Census and gun_violenc.

Let’s see gun violence density in each state.

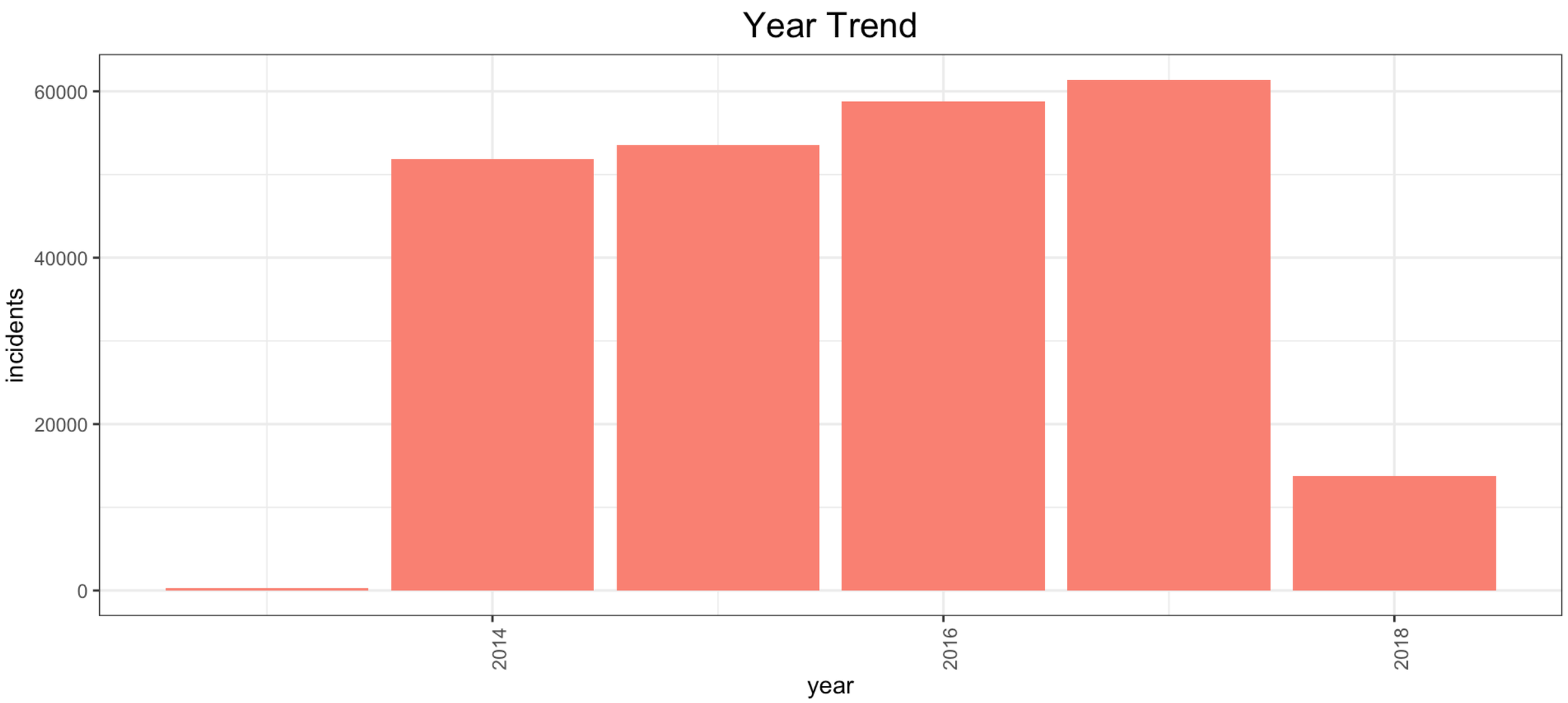


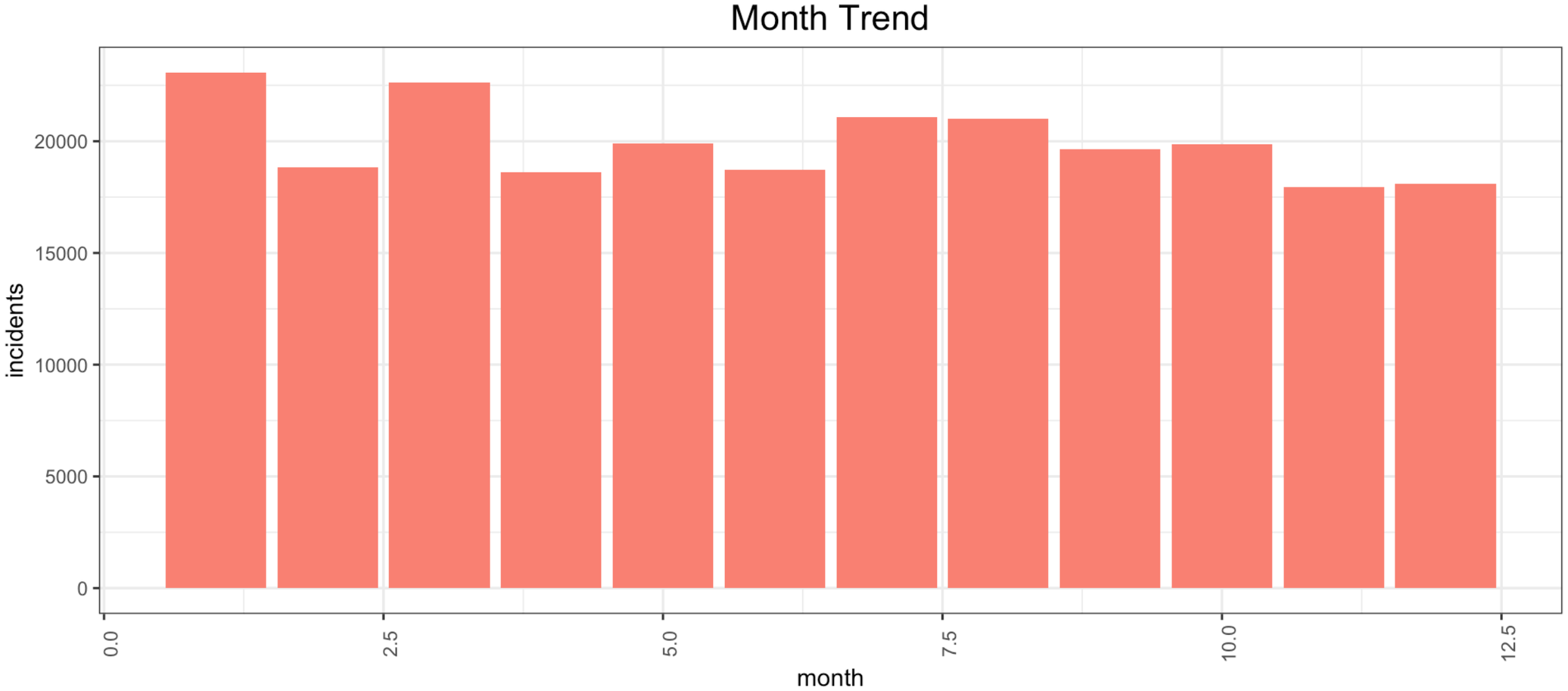
In this graph, we can find distribution of gun violence density in each state differs from distribution of amounts pf gun violence in each state. So we can not say that increase of population can cause gun violence more frequently.

4.4 Trend of incidents and gun_violence over time.

Lets understand whether there were any patterns in the incidents over the year,month， quarter and day

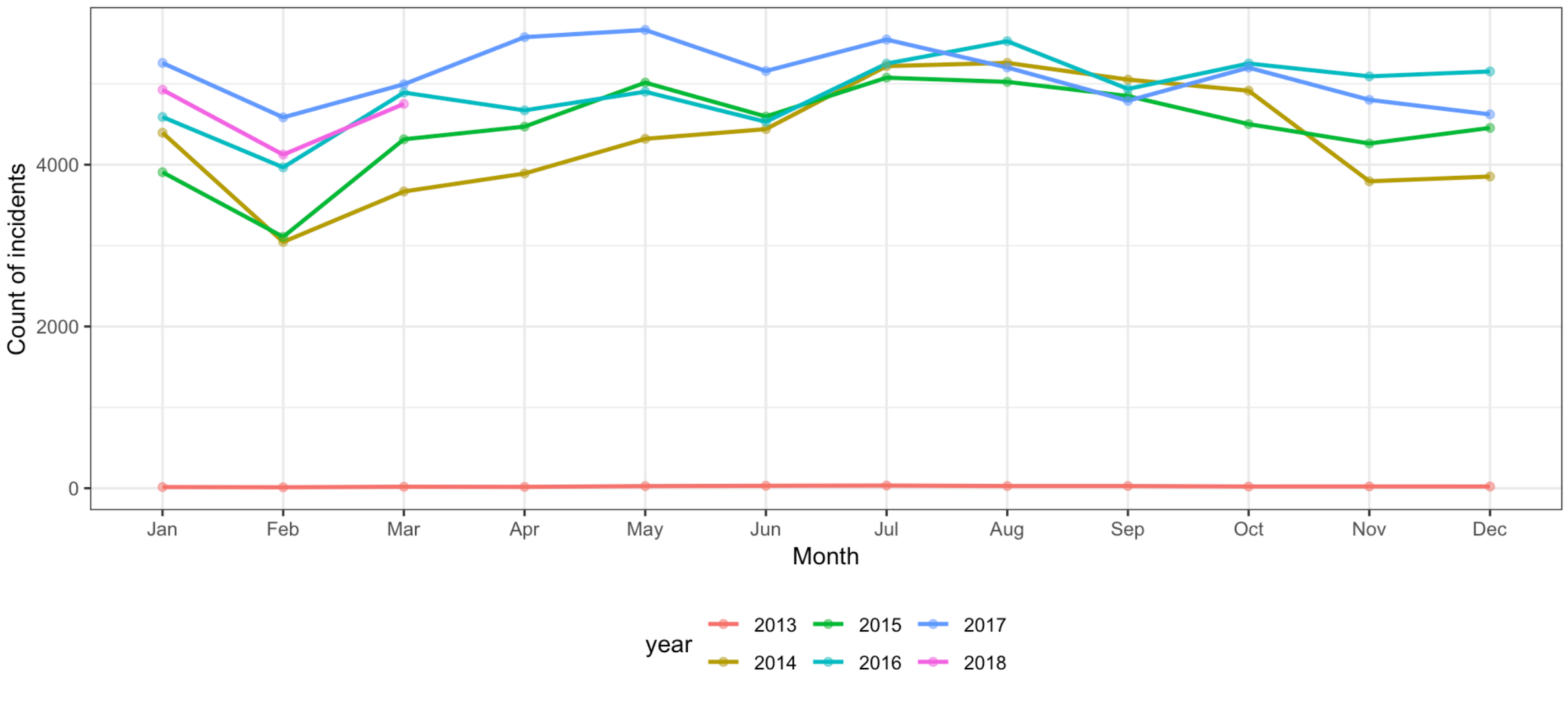
Firstly, let’s look at trends of incidents over year.





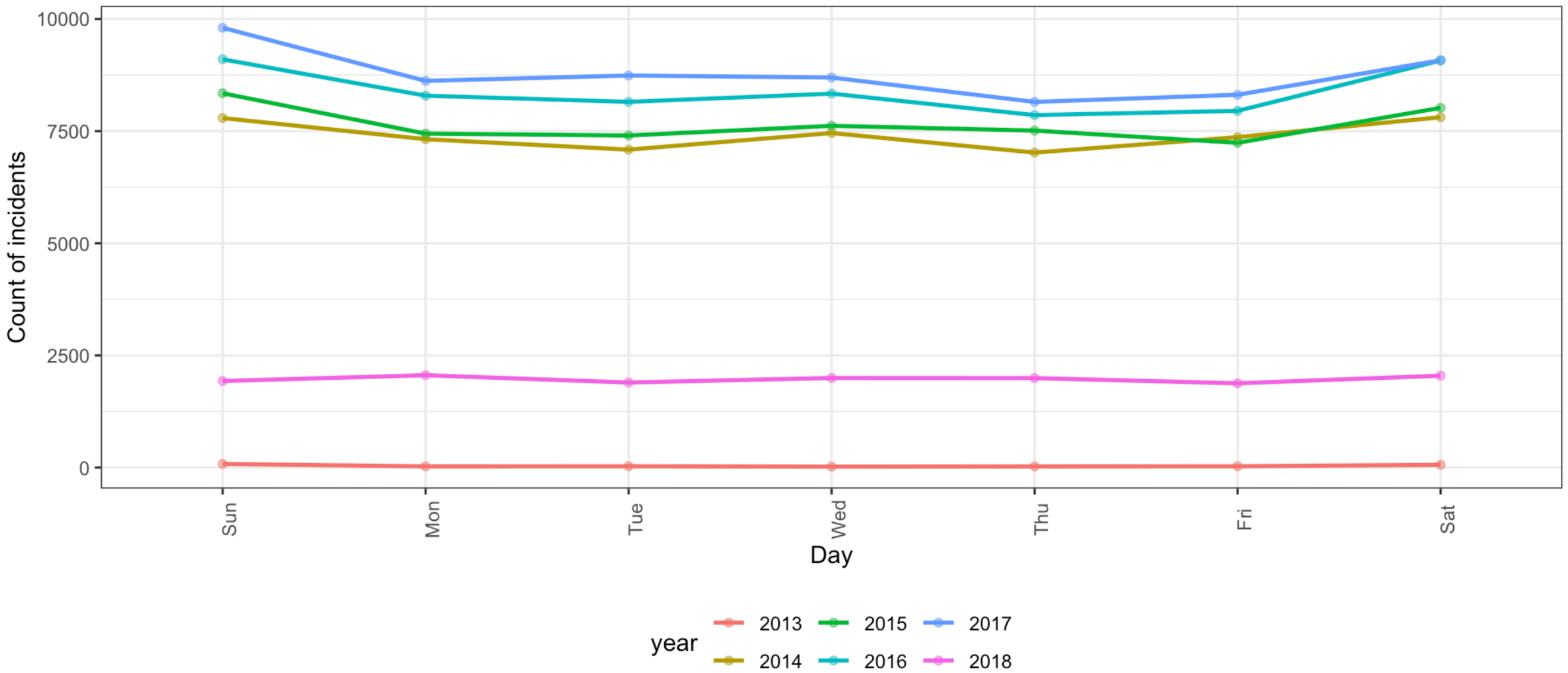
Because the data of 2013 and 2018 are not complete, so the incidents numbers in 2013 and 2018 are very low. However, we can still find the increasing numbers of gun violence with time goes by. What's more, the month does not present some trends. However, we can still find that the incidents happened in Jan and March are more than other months.

So let's move our steps to see what relationships between numbers of incidents and month.



Compared to the bar chart, this line graph can give us a much more clear view of the incidents each month. We can still find that there is no clue showing that some incidents have relationships with the month. But this graph can also tell us the gun violence become much more than previous years.

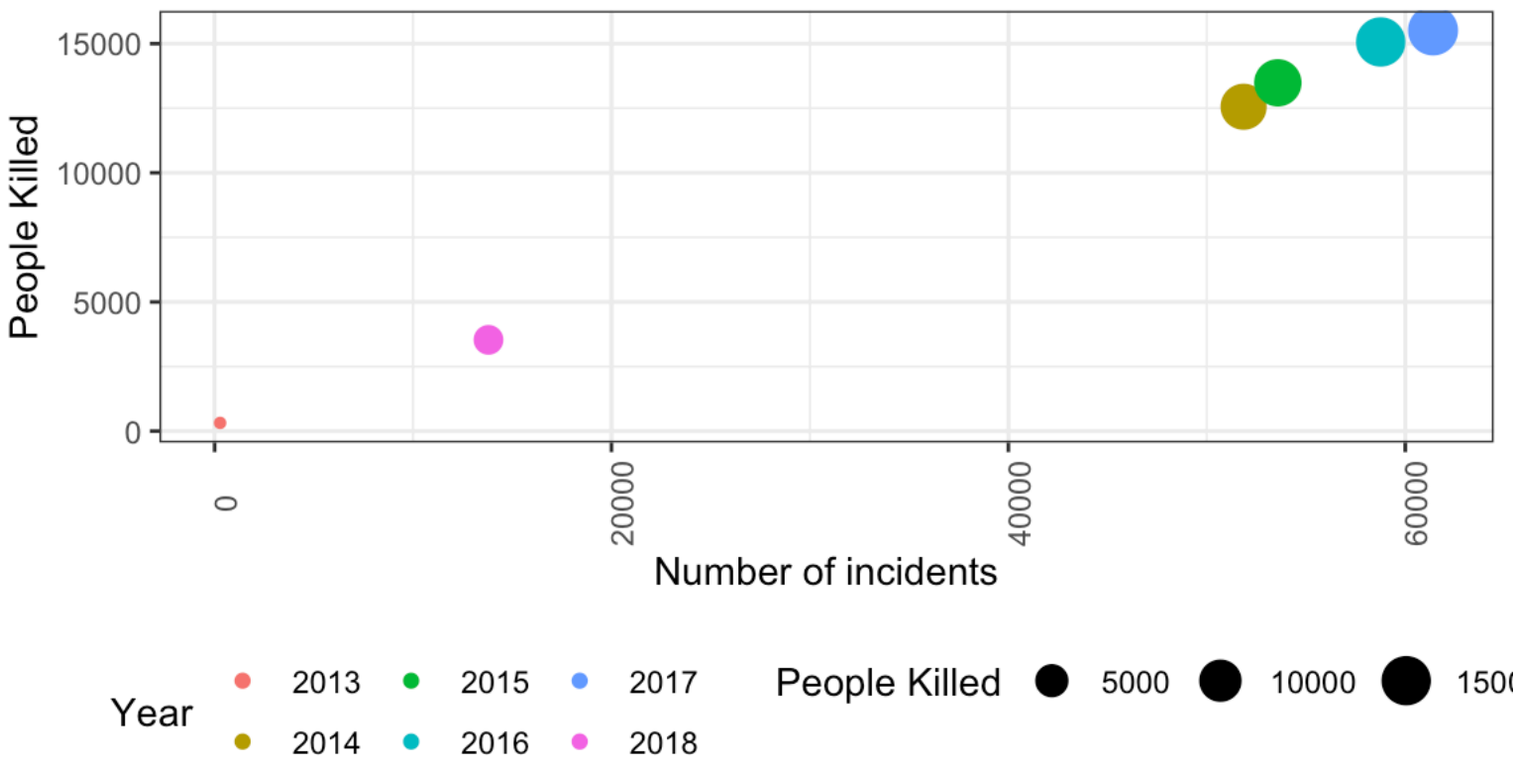
Now, let's look at trends of incidents by day.

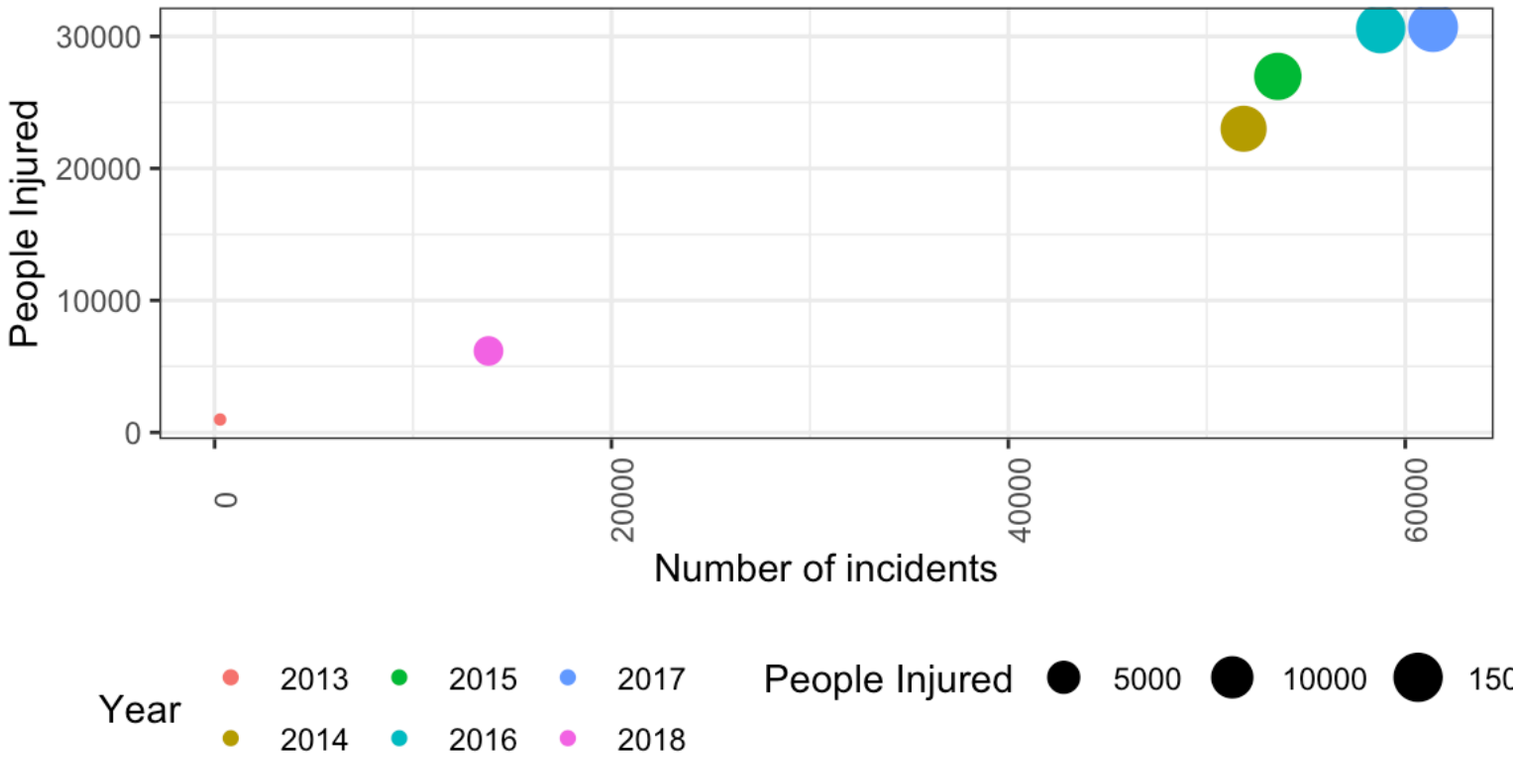


We can find it seems like that incidents are easier to be happened on weekend, especially on Sunday.

4.5 Time Series for Number of people killed and injured:

Let us now visualise the number of people killed or injured in the incident as a function of time.





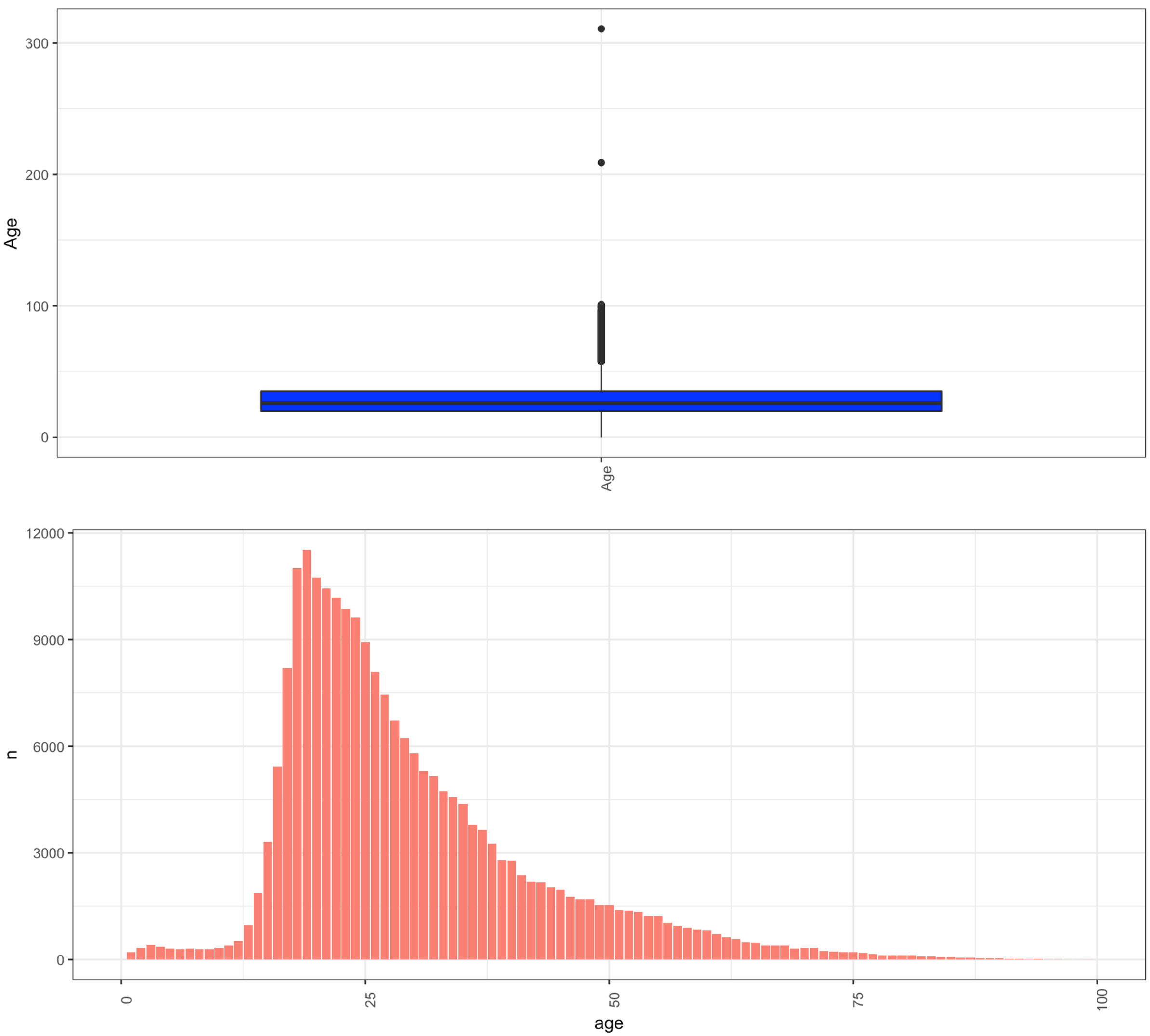
We find that there is a perfect cause and effect relationship. As the number of incidents rises, the number of people injured or killed has raised for the year. This is nothing strange and is very normal. But this also proves that our data is correct.

4.6 Age distribution of preparators

```
## [1] 0::20                                  0::20
## [3] 0::25||1::31||2::33||3::34||4::33 0::29||1::33||2::56||3::33
## [5] 0::18||1::46||2::14||3::47        0::23||1::23||2::33||3::55
## 18952 Levels:  0::0 0::0||1::1||2::28||3::24 ... 9::28
```

We can find the virsion of age information in this data is werid. So we need to preprocess it.

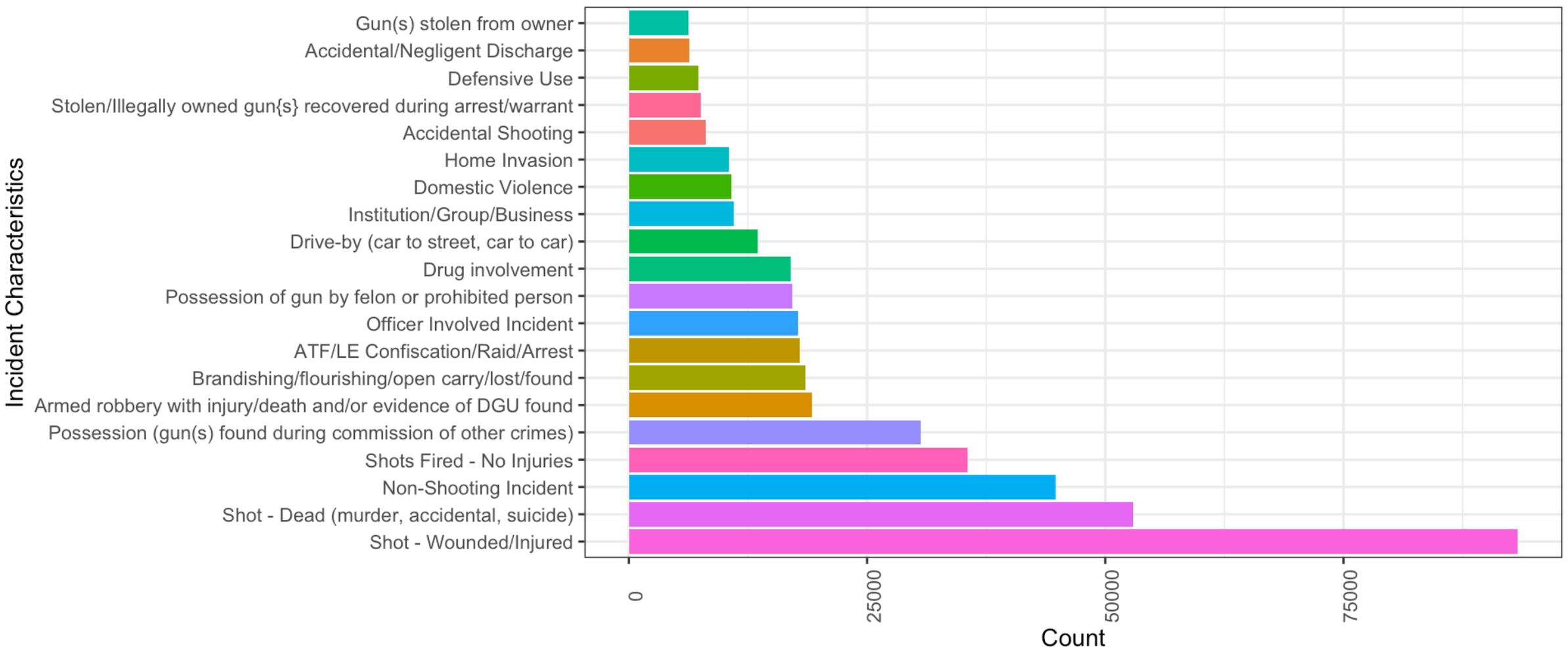
```
## [1] 20 20 25 31 33 34
```



We see a large number of outliers. What' more, ages of bad guys gather around 25 years old. So young people should calm down and reflect themselves.

4.7 Incident Characteristics:

Lets visualise the characteristics of each incident.

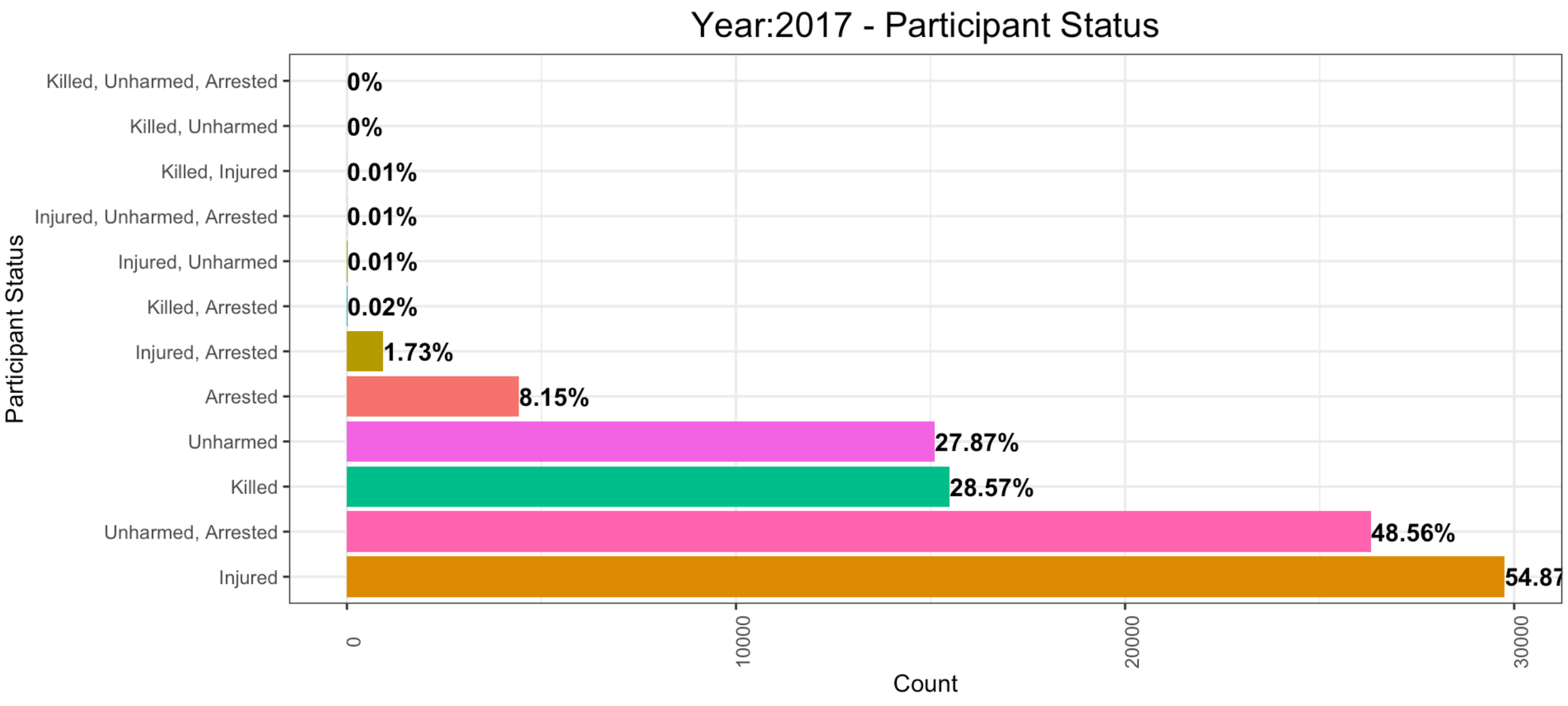


This is the text mining part in my project. We can find the discriptions of gun violences in data set have typical characteristics. There are various reasons for gun violence.

4.8 Participant status:

It is important to know what happened after the incident happened.The dataset provides information about the type of participant(either victim or suspect) and what the person's status was after the incident.Lets use these features to visualise and understand the scenario.

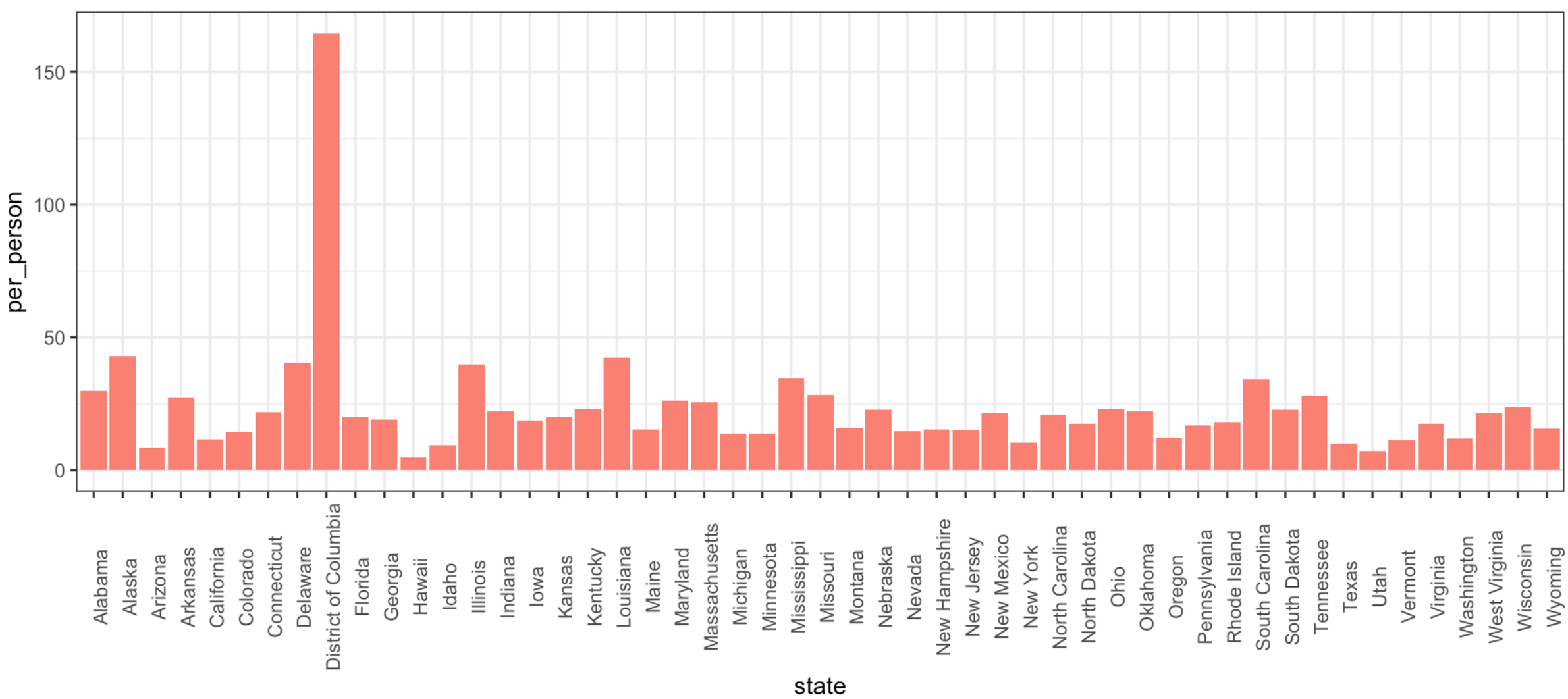
```
## [1] "Killed"              "Unharmed, Arrested" "Unharmed, Arrested"
## [4] "Unharmed, Arrested" "Injured"            "Unharmed, Arrested"
```



As infered from previous graphs,we find that nearly in half of the instances people were injured followed closely by arrest.28 % of the people were killed where as nearly same amount were unharmed.

We can find califonia has the most people in America.

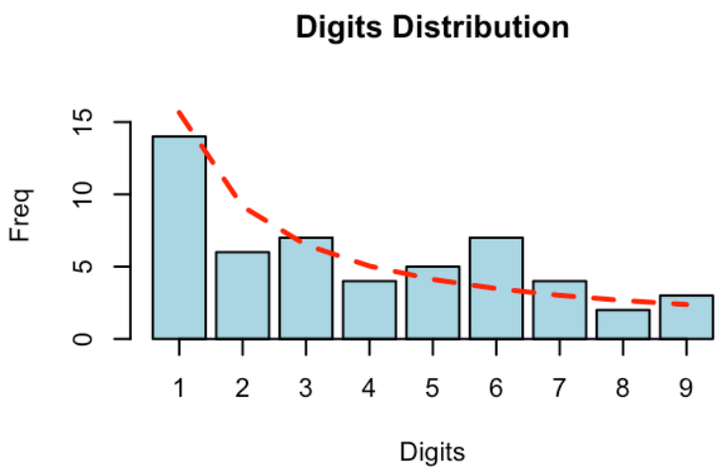
4.9 Census and gun_violence

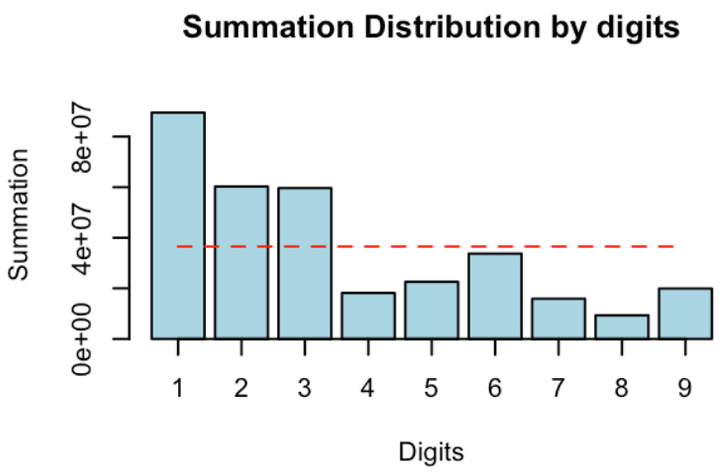


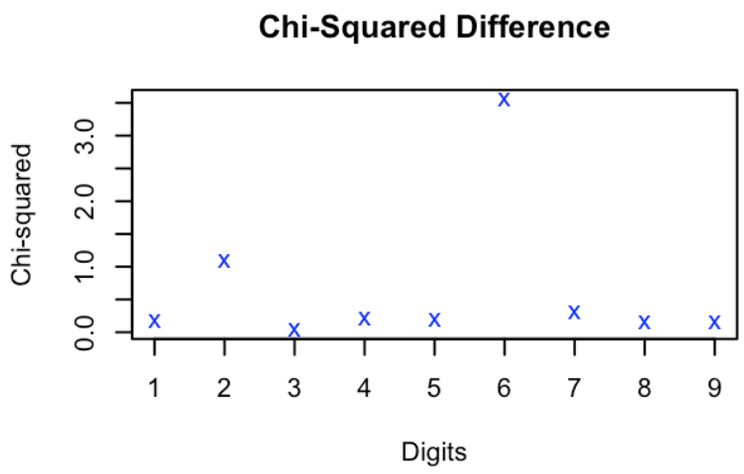
It seems like that density of gun incidents in Columbia is the highest around America.

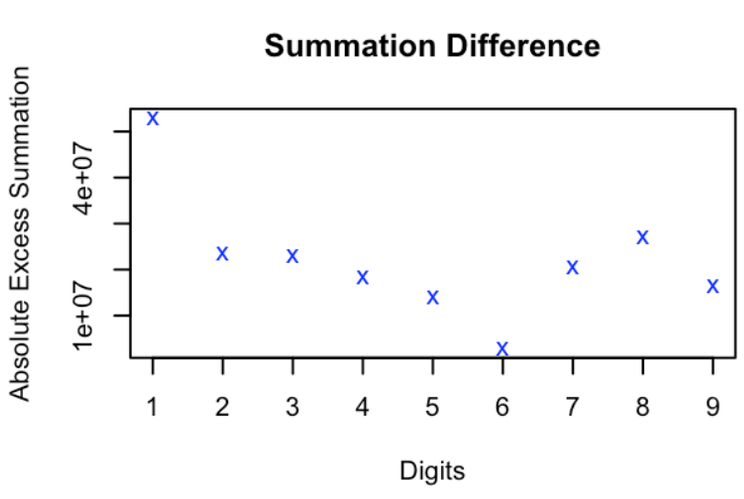
4.10 Benford.Analysis Package

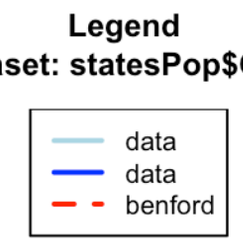
We use benford to find out potential fraud cases in this census data.












```
## [1] 0.02781862
```

number	duplicates
4874747	1
739795	1
7016270	1
3004279	1
39536653	1
5607154	1

3588184	1
961939	1
693972	1
20984400	1

It’s acceptable that at most three counties have the same population. However, our data is not consistent with Benforld law perfectly.

The ‘suspicious’ observations according to Benford’s Law:

state	Count
District of Columbia	693972
Florida	20984400
Indiana	6666818
Kansas	2913123
Maryland	6052177
Massachusetts	6859819
Mississippi	2984100
Missouri	6113532
Nevada	2998039
New Mexico	2088070
Tennessee	6715984
Texas	28304596
Vermont	623657

The first digits ordered by the mains discrepancies from Benford’s Law:

digits	absolute.diff
6	3.5187669
2	3.1567455
1	1.6535598
4	1.0393207
7	0.9844188
5	0.8825752
8	0.6599312
9	0.6206105
3	0.5031857

```
##
##  Pearson's Chi-squared test
##
## data:  statesPop$Count
## X-squared = 5.9091, df = 8, p-value = 0.6574
```

The p-value is 0.654 so that we cannot reject null hypothesis, which means that the distances between data points and benford points are not significantly different.

4.11 BenfordTests Package

```
##
##  JP-Square Correlation Statistic Test for Benford Distribution
##
## data:  statesPop$Count
## J_stat_squ = 0.18996, p-value = 0.3009
```

Joenssen's JP-square Test for Benford's Law: The result signifys that the square correlation between signifd(statesPop$Count,2) and pbenf(2) is not zero.

Code

```
##
##  Euclidean Distance Test for Benford Distribution
##
## data:  statesPop$Count
## d_star = 0.74657, p-value = 0.6826
```

"edist.benftest" takes any numerical vector reduces the sample to the specified number of signif- icant digits and performs a goodness-of-fit test based on the Euclidean distance between the first digits' distribution and Benford's distribution to assert if the data conforms to Benford's law.

The p-value is greater than 0.05 so that we can not reject the null hypothesis. Therefore, the goodness-of-fit test based on the Euclidean distance between the first digits' distribution and Benford's distribution shows the data does conform to Benford's law very well.

4.12 Benford analysis conclusion

Even though all the tests and plots we've done signify that our data follows well the Benford Law, we can't arbitrarily say that there are not frauds in these census observations.

5 Conclusion

From the above analysis, we can find the population has positive relationships with gun violence. What's more, the outcome of gun violence is serious because almost every gun violence accompanies people injured and dead. Besides, gun violence can cause severe mental health issue which is hard to cure. There are various reasons for gun violence. But it is evident that people in 25~30 years old are more likely to do some bad things. So government should take some actions.

6 Acknowledgment

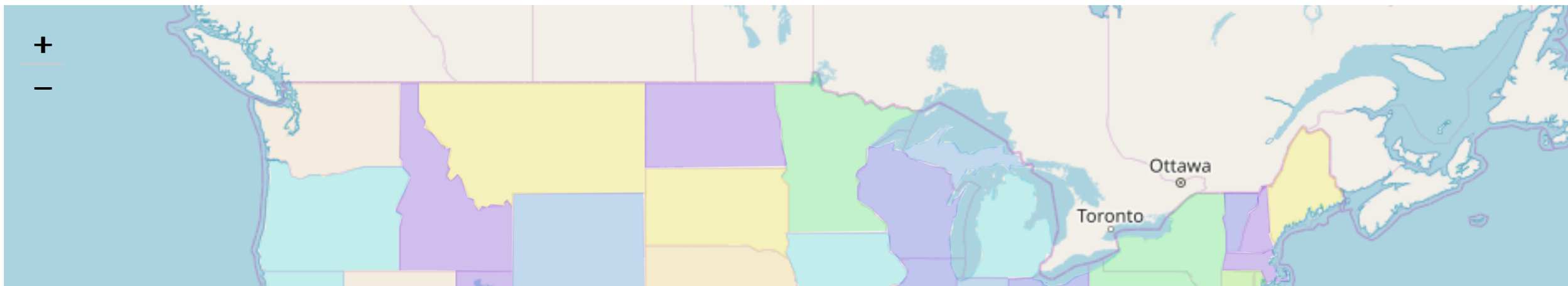
Firstly, thank Professor Haviland for letures in this semester. Your kind words helped me in study. Thank you for the extra help you gave me during your office hours on the few concepts I struggled with. I really appreciate your talent and dedication to your avocation and your students. I really appreciate my TA Brian. You really help me solve many problems in my study.

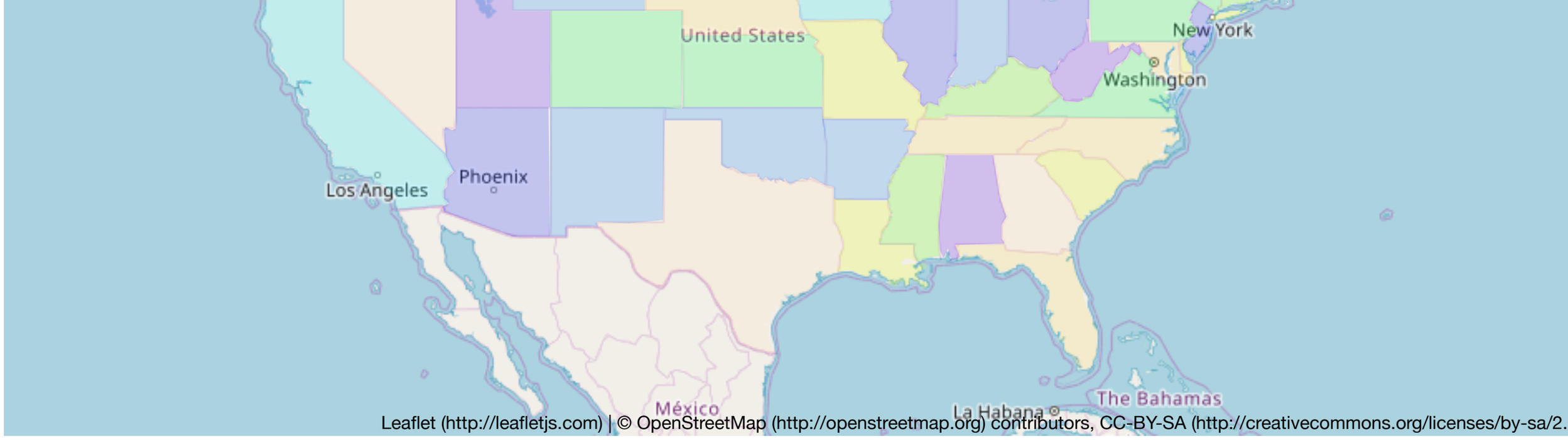
Besides,I would also like to show my gratitude to so many outstanding data scientists for sharing their pearls of wisdom. Their kindness really change world a lot.

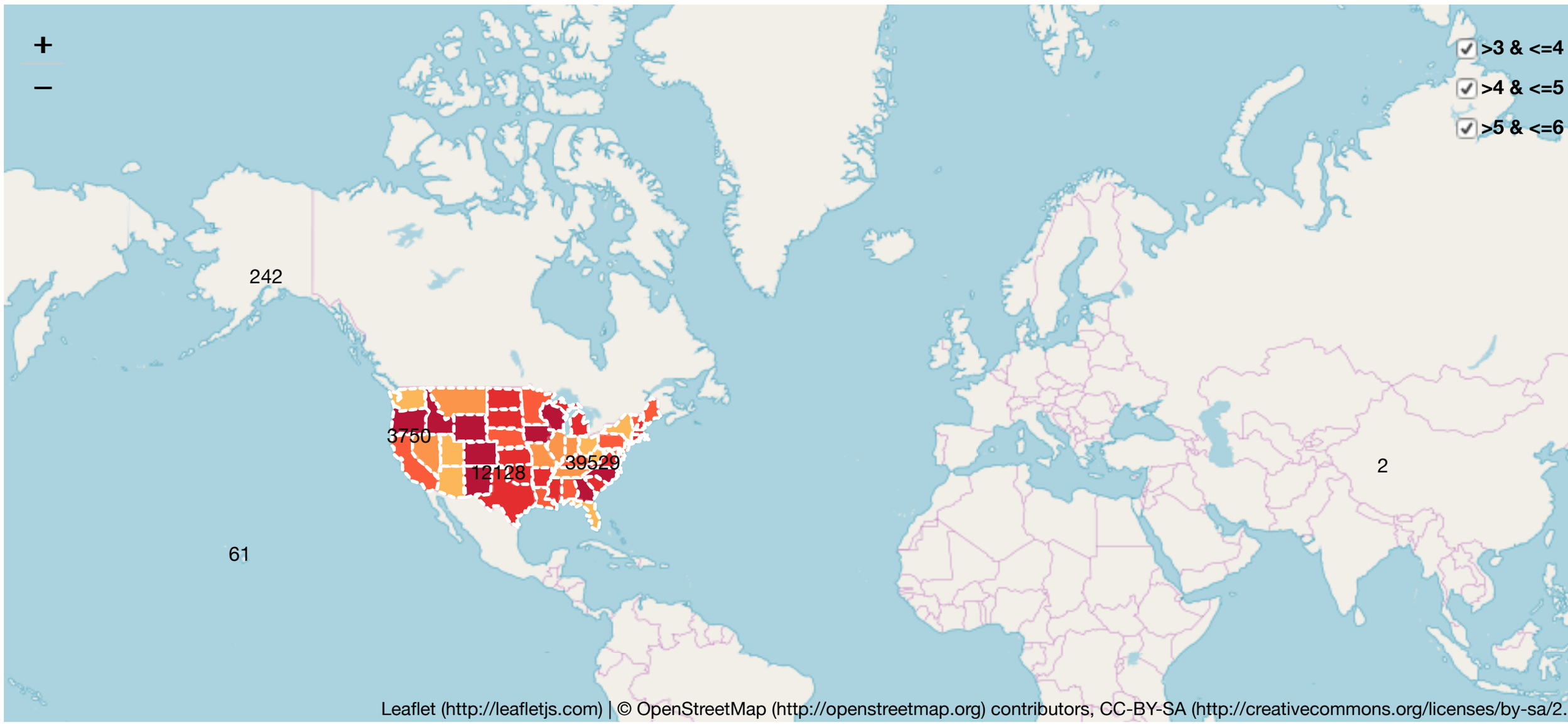
At last, I should restate that data used in this project comes from kaggle provided by HomeCredit Company. I will only use this data for nonprofit research such as final project.

6.1 map

Code




United States
New York
Washington
Phoenix
Los Angeles
México
La Habana
The Bahamas
Leaflet (http://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.


Code
+
−
>3 & <=4
>4 & <=5
>5 & <=6
242
3750
12128
39529
61
2
Leaflet (http://leafletjs.com) | © OpenStreetMap (http://openstreetmap.org) contributors, CC-BY-SA (http://creativecommons.org/licenses/by-sa/2.