Texas gun violence – exploratory data analysis

Group members: Shuyuan Sun, Cuiting Zhong, Sichun Zuo

1. Introduction

Every day, about 30 people die in gun homicides in the U.S. Another 200 are injured by bullets. These incidents of gun violence are recorded in every corner of the country. Thanks to the Gun Violence Archive, data on gun homicides and nonfatal shootings is available for us.

2. Description of project goals

1) Description

We are investigating Gun Violence Data (over 260k US gun violence incidents from 2013-2018) but focusing only on relevant data of Texas State. We adopt exploratory analysis to interpret the data from different perspectives. We try to find out the trend of this incident and make a forecast of how gun violence can increase or decrease in the future.

2) Importance of the problem

We picked this topic because many of us are international students who first come to the U.S. For many of us, gun violence is rare where they used to live. By investigating the data, we can obtain some insights into Texas gun violence to get familiar with the conditions where we are living. Meanwhile, for all of us who will live here for a long time, this analysis helps to warn us of time periods and areas with high gun violence rate as well as potential criminals.

3. Data overview and data cleaning

We used *the gun-violence-data_01-2013_03-2018 dataset* provided by Kaggle, which was originally from the Gun Violence Archive. This dataset has 239768 observations and 29 variables in total.

Since our objective is to investigate the gun violence in Texas area and there are too many missing values in some columns, we chose the subset where state is Texas as our object and removed columns with very high rates of missing values. Also, there is unstructured data in those text columns (incident_characteristics, participant_gender, etc.), we tried to extract the useful information from them. Hence, we created 24 new variables, including n_victim_killed, n_victim_injured and so on.

4. Exploratory analysis

i. Time related trends

Since the size of sample in 2013 is too small to obtain reasonable analysis, discussion in this section is based on the data from 2014 to 2018Q1.

a) By year (graph 4-i-a)

There are 3133, 3276, 3606, 2875 gun incidents in 2014, 2015, 2016, 2017, respectively. And 676 gun incidents in the first quarter of 2018.

b) By quarter

The average quarterly number of gun violence is 798. From graph 4-i-b, there was a fluctuating increasing trend from 2014Q1(748) to

2017Q1(958) and then it dropped dramatically to 691 in 2017Q2 and kept decreasing during year 2017. However, it began to increase since 2018. In addition, it seems that gun violence happens most frequently during the third quarter in each year, which might be related to the summer holiday factor.

c) By month

On average, there are 266 gun-incidents within a month. From plot graph 4-i-c1, we got the similar conclusion as the quarterly analysis. The number of gun violence increased from 2014 to Jan 2017 and then kept decreasing from Jan 2017 to June 2017. Since July 2017, it reverted to the old pattern as before Jan 2017.

From the bar graph 4-i-c2 (which excludes the data in 2018), in terms of the number of gun violence occurred during 2014-2017, January ranks the highest and June the lowest.

d) By weekday (graph 4-i-d)

We also want to find out if there is some obvious difference between weekend and weekdays. Surprisingly, there are most occurrence in Sunday but the least in Saturday.

e) Forecast

We tried to do forecast using time series model. First, create a time series using the monthly count series.

By ploting the time series and using Augmented Dickey-Fuller Test, we found it is stationary, so we did not need to do any differential or other transformation. Then we tried to do seasonal decompose (graph 4-i-e1), acf plot and pacf plot (graph 4-i-e2) to check its trend and seasonality. Based on the previous results, we chose ARMA (1,0) and ARIMA (1,0,0) x $(1,0,0)_{12}$ as two candidate models. After fitting each model and finishing model diagnosis, both models passed the diagnosis test. Based on the AIC criterion, we decided to use ARIMA (1,0,0) x $(1,0,0)_{12}$ as our final model and used it to do the future one-year forecast. The forecast result is shown in the graph 4-i-e3 and table 4-i-e.

ii. Location related trends

a) Incidents by city

- Total no. of incidents by city

  According to the heatmap (graph 4-ii-a1), incidents are more often in east Texas. To be specific, in graph 4-ii-a2, top 5 cities with most incidents are: Houston, San Antonio, Dallas, Corpus Christi, and Austin. This may be because that there are more crimes in these main cities but may also be that these cities are densely populated. So, we then adjusted the number of incidents by population of each city.

- Incidents per 1k people by city

  Top 20 cities with highest number of incidents per 1k residents (graph 4-ii-a3) are a bit different from top 20 with highest total number of incidents. We only consider cities where total number of violence >10 and average population >1000. According to the tables (table 4-ii-a1), Mansfield, Corpus Christi, Killeen, Longview…etc. appear in both lists. In these cities, one is more likely to encounter gun violence.

- Austin in specific
  In Austin, according to the heatmap (graph 4-ii-a4), more incidents appear in Downtown, around East Riverside, and around Webb Middle School.

b) Victims by city

- Total no. of victims killed and injured by city
  According to graph 4-ii-b1 & b2, 18 cities are in top 20 whether rank by total number of victims killed and injured or by total number of incidents. Though Temple and Humble are not ranked top by number of violence (table 4-ii-b1), there are more people injured or killed.
- No. of victims killed and injured per 1k people by city
  Similarly, in this part we also only consider cities where total number of violence >10 and average population >1000 (graph 4-ii-b3 & b4). Though Converse, San Antonio, La Marque…etc. are not ranked top by number of violence per 1k residents (table 4-ii-b2), there are more people injured or killed per 1k residents.

c) Incident characteristic in Texas total and by city
   For the whole Texas, we find that "Armed robbery" characteristic appeared most with 1496 records and "Officer Involved Incident" characteristic appeared in 1432 records. The third characteristic "TSA Action" with 944 records. Therefore, we can see that suspects in Texas use guns to rob most often. Also, among all gun incidents, officers involved nearly 10% of them. (4-ii-c1)
   For the big cities, we test 15 cities with the highest population. The feature characteristic for Houston is "Armed robbery", whose number is nearly 25% of all armed robbery records. Second city Dallas features "TSA action". San Antonio features "Drive-by". Austin features "TSA action" and Corpus Christi features "Armed robbery". It seems like that armed robbery happens often across Texas and suspects are more likely to carry weapons passing by Dallas and Austin. (4-ii-c2)

iii. Suspect and victim analysis

a) Age distribution

   - Suspects
     The age group with the most suspect is 18, which has 344 suspects, followed by 19, which has 341 suspects. And the number of suspects drops significantly after the age of thirty. Therefore, we conclude that most suspects are young people. (4-iii-a1)
   - Victims
     The victim also has its main age group at 20 with 236 victims. 18~22 age groups all have more than 200 victims. The number of victims decreases mildly from 20s to 80s. The victim also has a

small group of children. From 2~4, each group has about 30 victims. (4-iii-a2)

- Compare
  We find that suspects have on average 100 more people than victims at 20s, while they have almost same number of people after 40s. The Victim stands out on child groups, which have obviously much more people than the suspect at same ages. Therefore, we can tell that some incidents are aiming young children in purpose.

b) Age group percentage in different types of incidents (robbery, drive by, mass-shooting)
We select three featured characteristics to further analyze. To easily compare participants' age difference, we divided ages into three groups. People with age 0 to 11 is regarded as child, 12 to 17 is teenage, and 18+ is adult. And we plot the bars in percentage ways.

- Suspects
  On Average, the teen group is 7% and child group is 0.6% of all suspects. In robbery cases, the percentage of teen is same with the average, while the child percentage is much smaller, only 0.04%. In Drive-by, the teen is the same, while the child increases to 1%. In mass shooting, there is no child and very little teen. (4-iii-b1~3)
- Victims
  In victims, on average teen is about 6% and child is about 3%. In three cases, robbery has 3% teen and 1% child, drive-by has 10% teen and 6% child, and mass shooting has 8% teen and 8% child. (4-iii-b4~6)

c) Gender percentage in different types of incidents (robbery, drive by, mass-shooting)

- Suspects
  On average, female is 9% in suspects. In robbery female is 15%. In drive-by female is 19%. In mass-shooting female is 27%. (4-iii-c1~3)
- Victims
  On average, female is 22% in victims. In robbery female is 5%. In drive-by female is 8%. In mass-shooting female is 5%. (4-iii-c4~6)

5. Conclusions and insights

From this exploratory data analysis, we examined gun violence data in Texas from different perspectives and discovered several interesting trends. For example, from past records, we predict the number of incidents in Texas will increase in fluctuation from 2018 March onwards. Houston, San Antonio, Dallas, Corpus Christi, and Austin have the highest number of incidents, but the number of incidents per 1k residents of these cities are not necessarily high. The age distributions of suspects and victims are both centered around 20s. This analysis helps to warn us of time periods and areas with high gun violence rate as well as potential criminals.

6. References

Gun violence data: https://www.kaggle.com/jameslko/gun-violence-data
City population data:
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

7. Appendices

Original Data:

```
df=pd.read_csv('gun-violence-data_01-2013_03-2018.csv')
df.info()
```

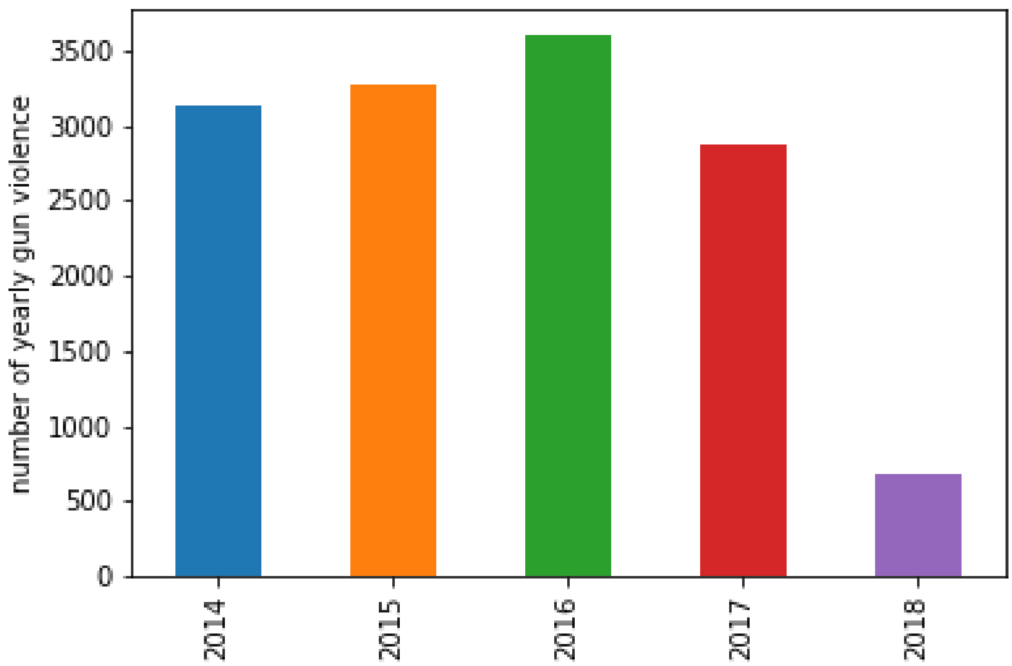
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 29 columns):
incident_id                    239677 non-null int64
date                           239677 non-null object
state                          239677 non-null object
city_or_county                 239677 non-null object
address                        223180 non-null object
n_killed                       239677 non-null int64
n_injured                      239677 non-null int64
incident_url                   239677 non-null object
source_url                     239209 non-null object
incident_url_fields_missing    239677 non-null bool
congressional_district         227733 non-null float64
gun_stolen                     140179 non-null object
gun_type                       140226 non-null object
incident_characteristics       239351 non-null object
latitude                       231754 non-null float64
location_description           42089 non-null object
longitude                      231754 non-null float64
n_guns_involved                140226 non-null float64
notes                          158660 non-null object
participant_age                147379 non-null object
participant_age_group          197558 non-null object
participant_gender             203315 non-null object
participant_name               117424 non-null object
participant_relationship       15774 non-null object
participant_status             212051 non-null object
participant_type               214814 non-null object
sources                        239068 non-null object
state_house_district           200905 non-null float64
state_senate_district          207342 non-null float64
dtypes: bool(1), float64(6), int64(3), object(19)
```

Cleaned data set:

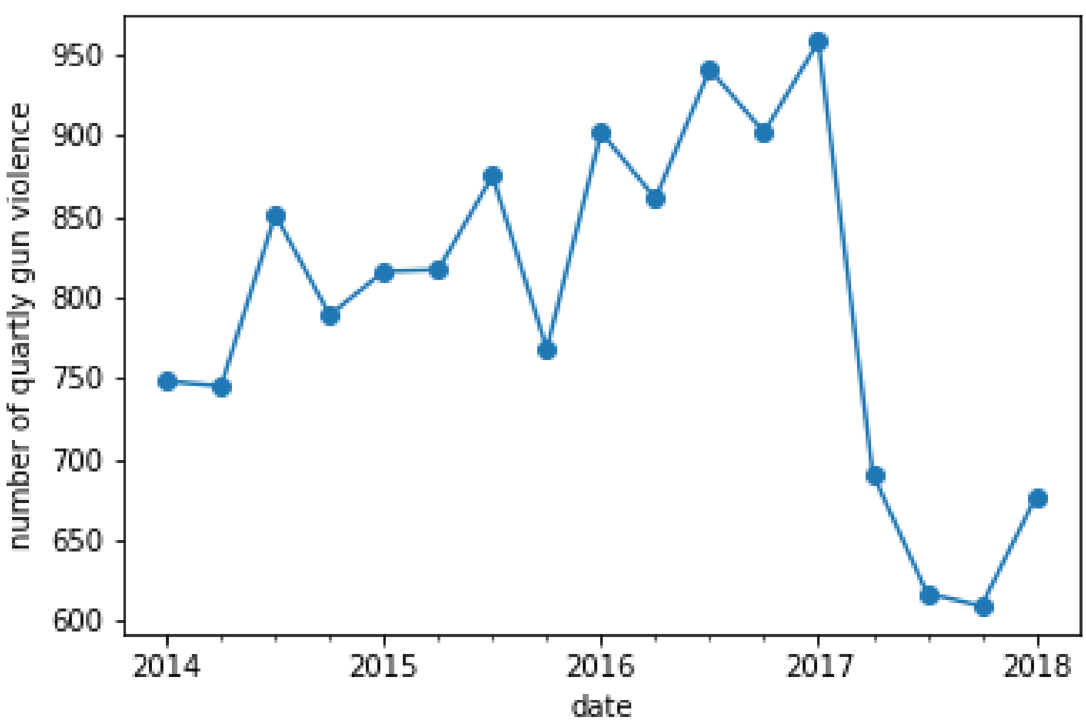
```
df=pd.read_csv('gun_cleaned.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13577 entries, 0 to 13576
Data columns (total 32 columns):
Unnamed: 0              13577 non-null int64
incident_id             13577 non-null int64
date                    13577 non-null object
city_or_county          13577 non-null object
address                 12446 non-null object
latitude                13017 non-null float64
longitude               13017 non-null float64
location_description    3049 non-null object
n_victim                13577 non-null int64
n_suspect               13577 non-null int64
n_victim_male           13577 non-null int64
n_victim_female         13577 non-null int64
n_suspect_male          13577 non-null int64
n_suspect_female        13577 non-null int64
n_victim_Child          13577 non-null int64
n_victim_Teen           13577 non-null int64
n_victim_Adult          13577 non-null int64
n_suspect_Child         13577 non-null int64
n_suspect_Teen          13577 non-null int64
n_suspect_Adult         13577 non-null int64
n_victim_Killed         13577 non-null int64
n_victim_Injured        13577 non-null int64
n_victim_Unharmed       13577 non-null int64
n_suspect_Killed        13577 non-null int64
n_suspect_Injured       13577 non-null int64
n_suspect_Unharmed      13577 non-null int64
n_suspect_Arrested      13577 non-null int64
mass_shooting           13577 non-null int64
robbery                 13577 non-null int64
Drive-by                13577 non-null int64
quarter                 13577 non-null int64
dayofweek               13577 non-null int64
dtypes: float64(2), int64(26), object(4)
```

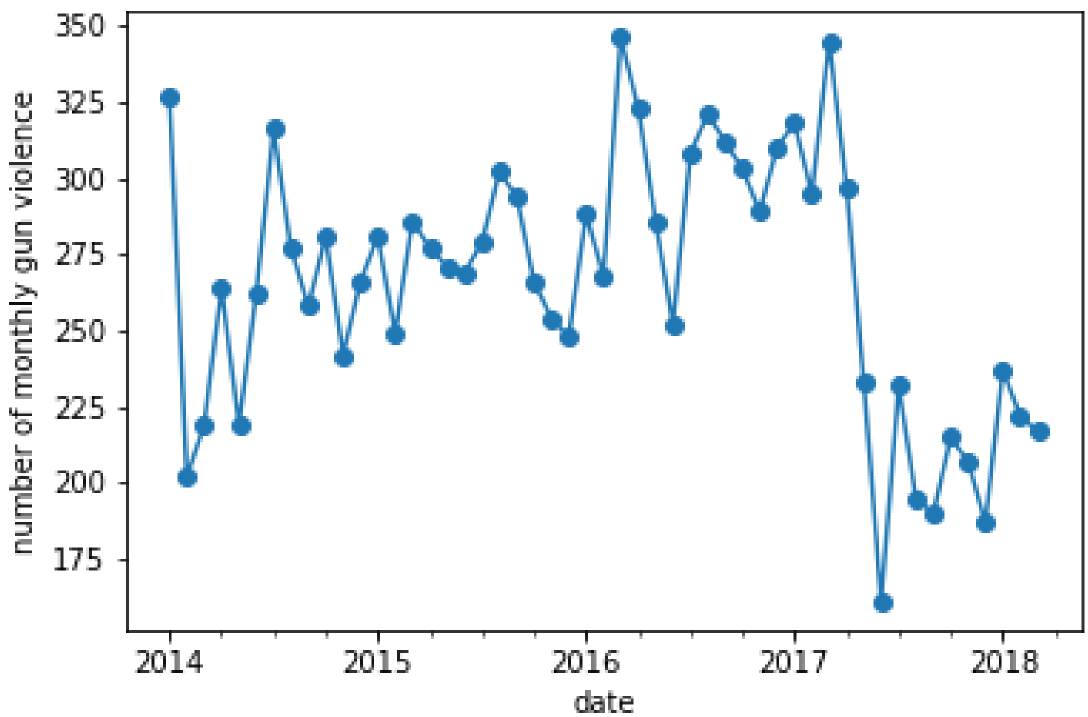
graph 4-i-a : yearly number of gun violence bar plot



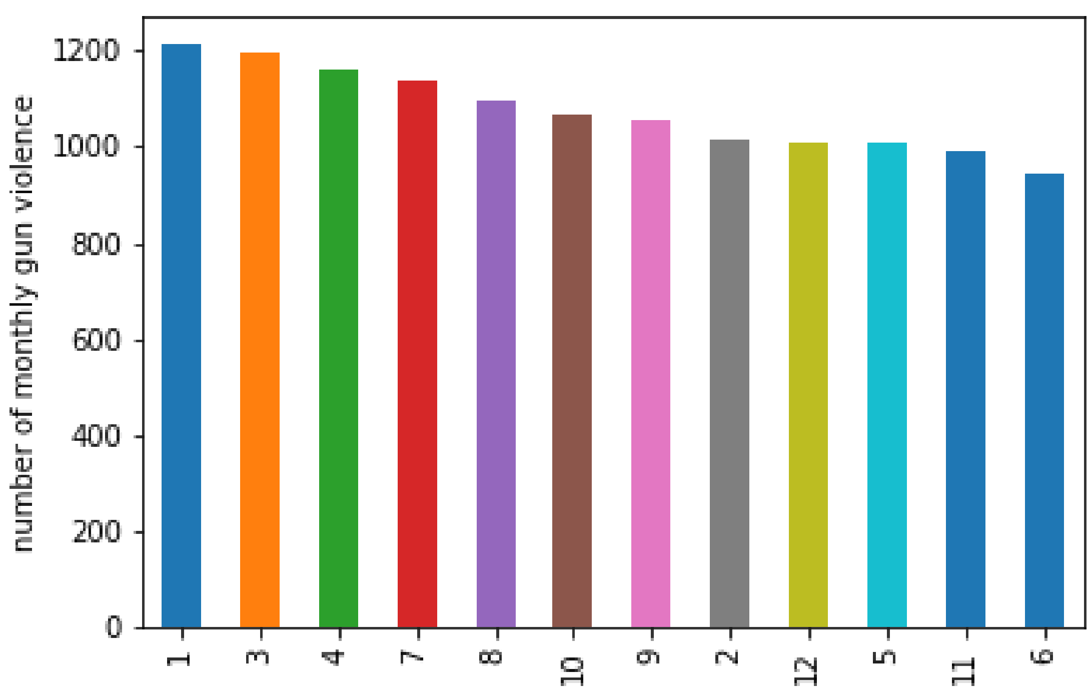
graph 4-i-b: Quarterly number of gun violence



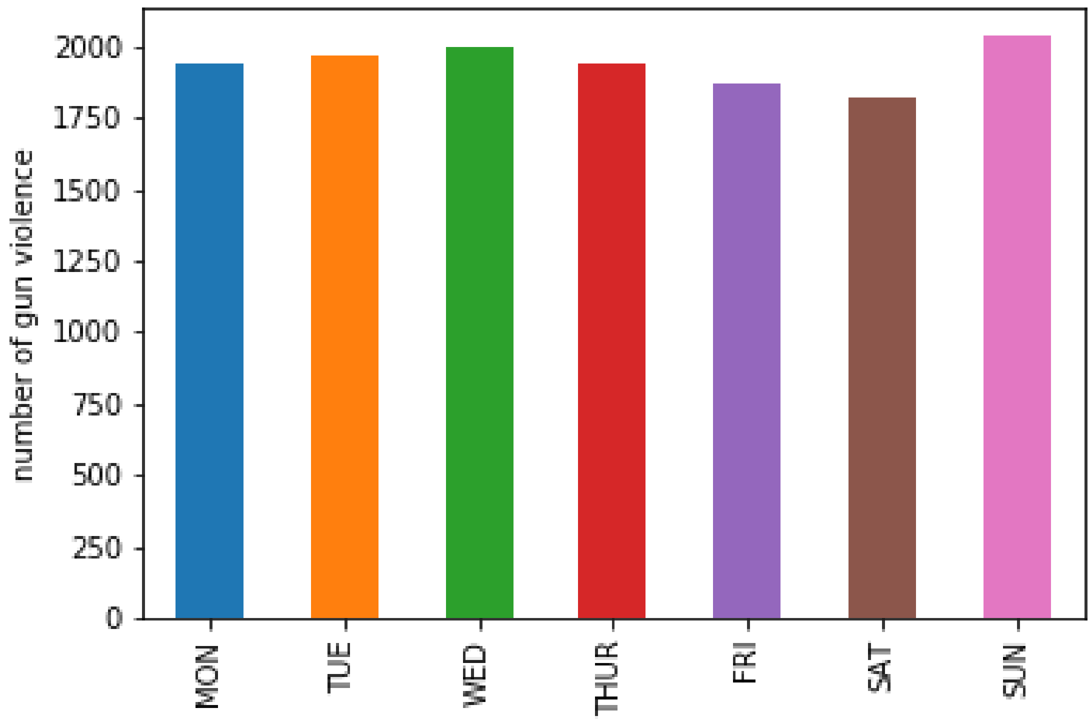
graph 4-i-c1: Monthly number of gun violence

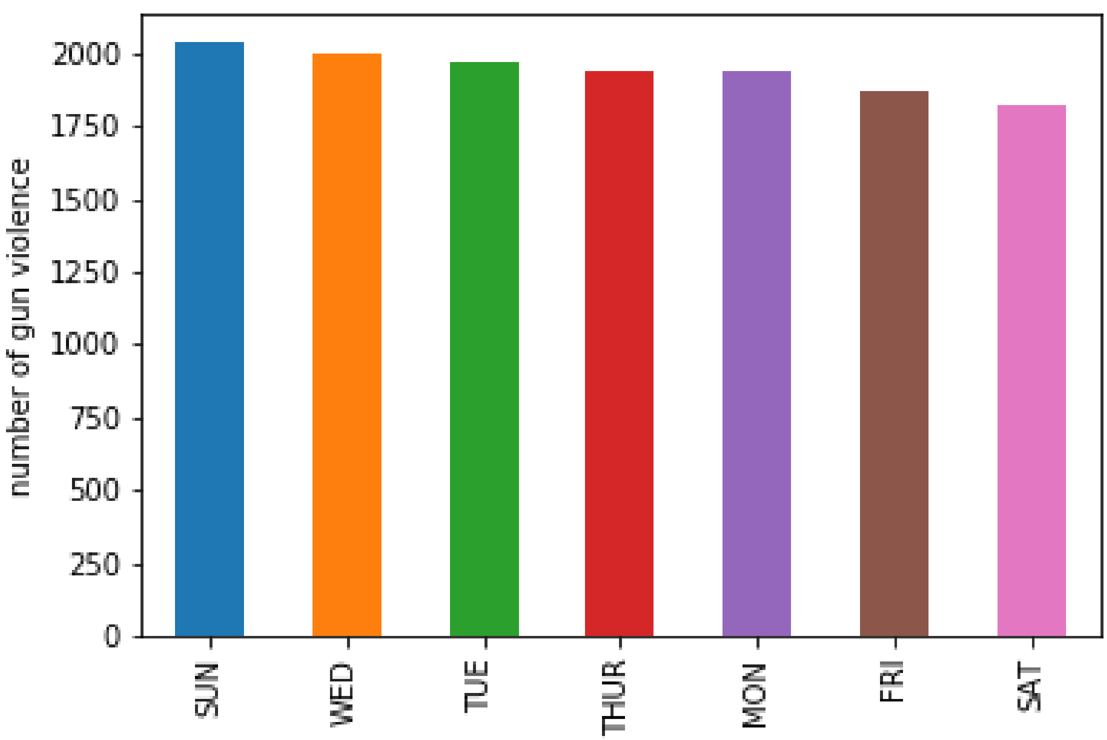


graph 4-i-c2: Monthly number of gun violence bar plot

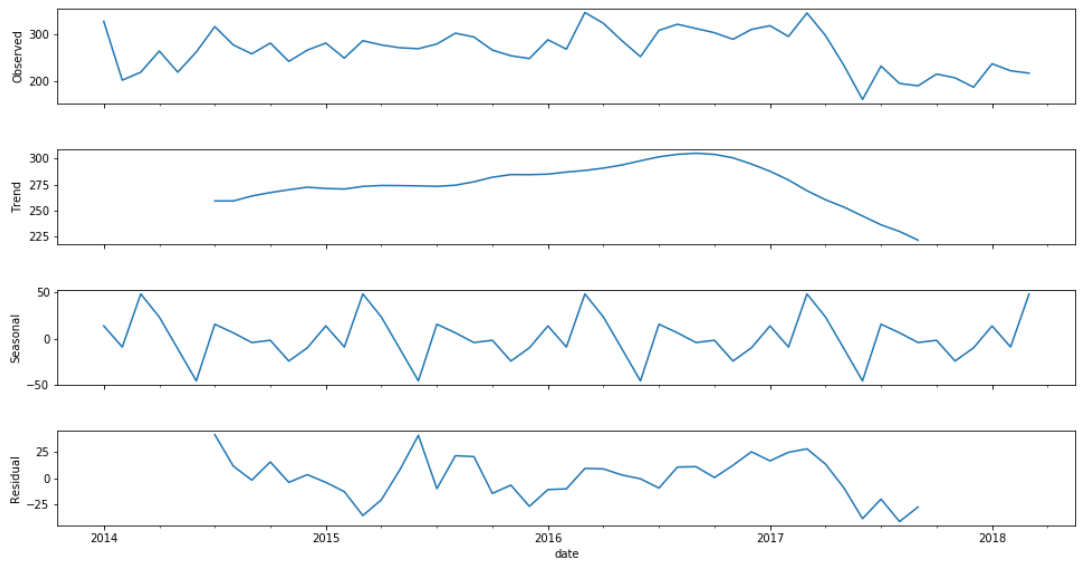


graph 4-i-d: Dayofweek number of gun violence

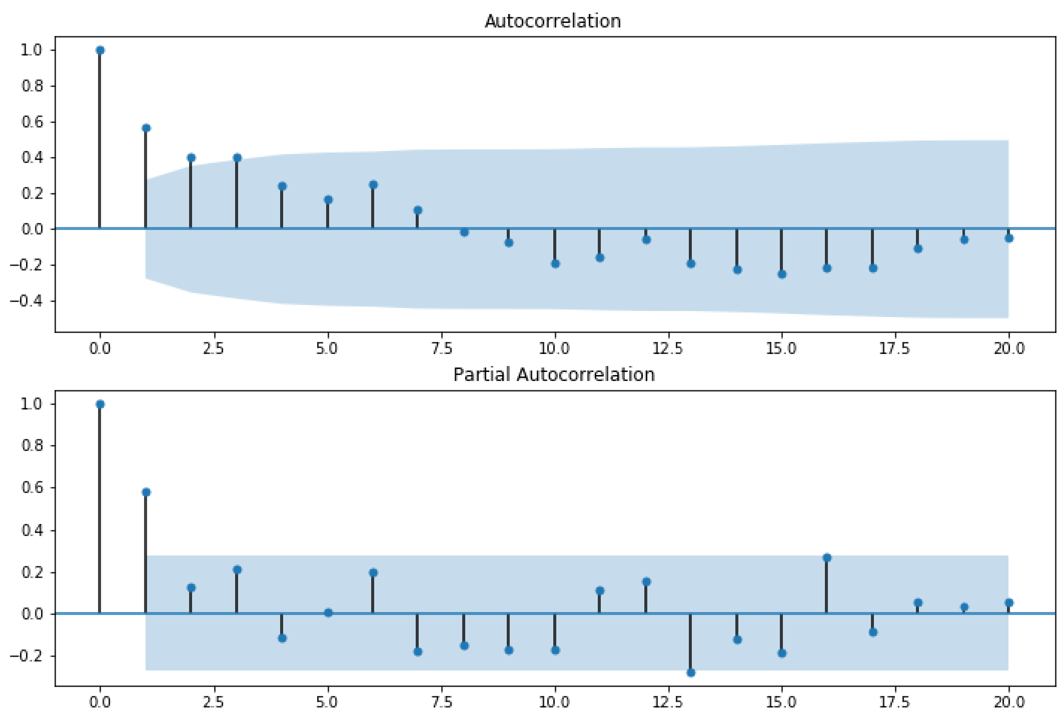




graph 4-i-e1: Decomposing trend and seasonality



graph 4-i-e2: ACF and PACF plot



graph 4-i-e3: One-year forecast

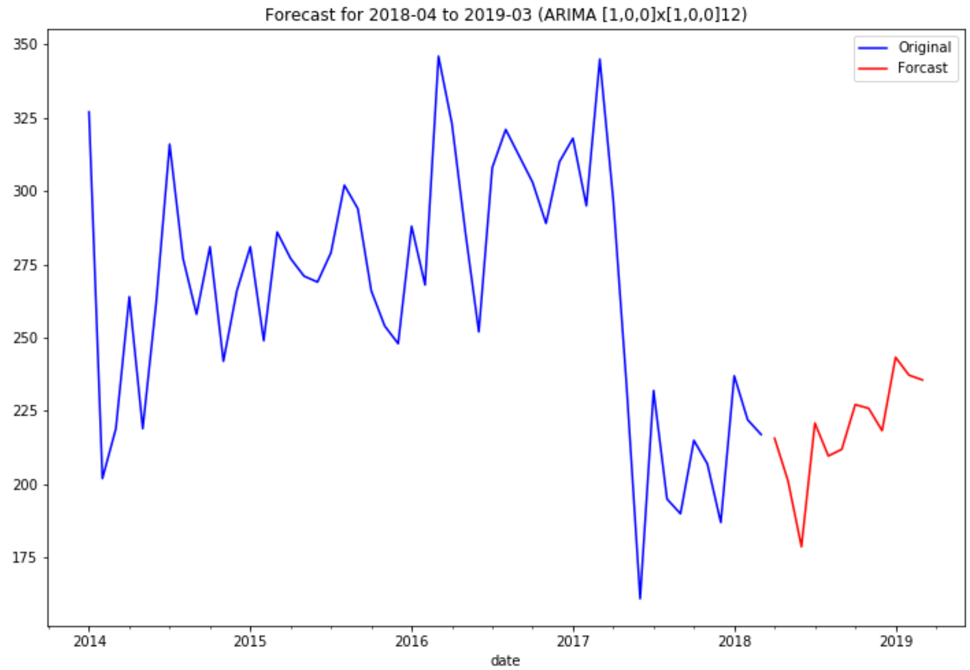
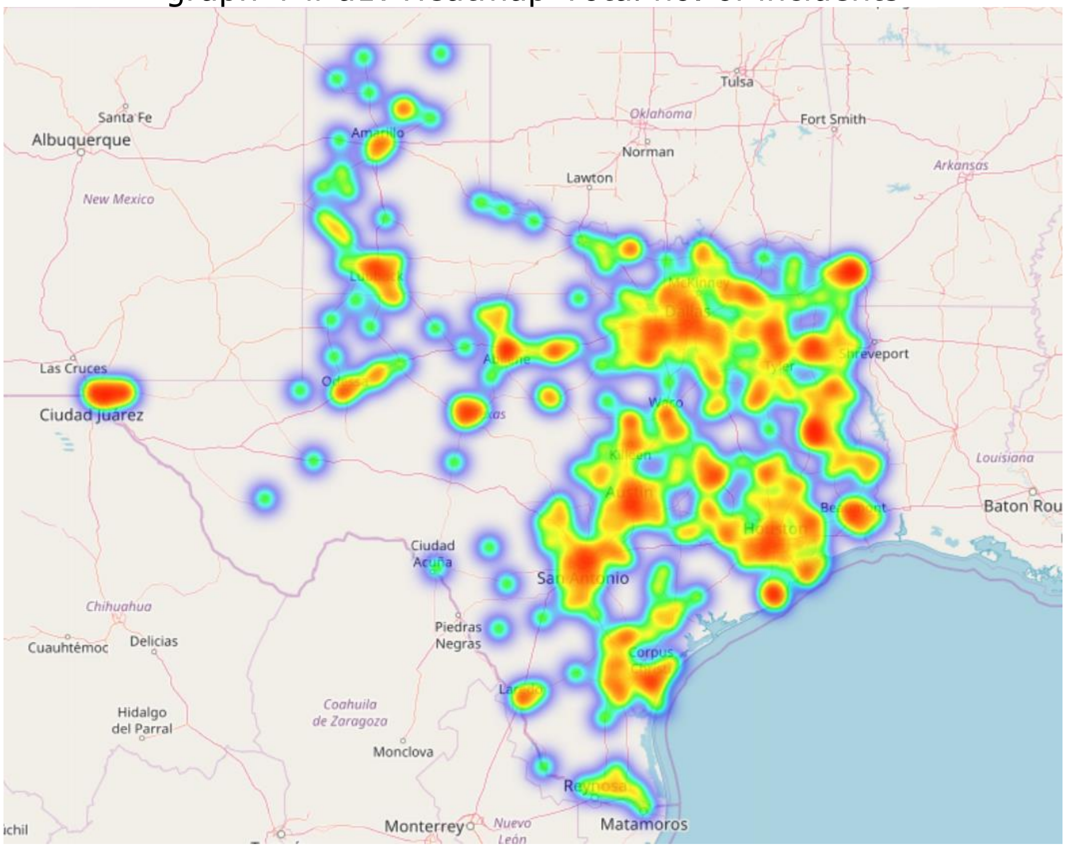


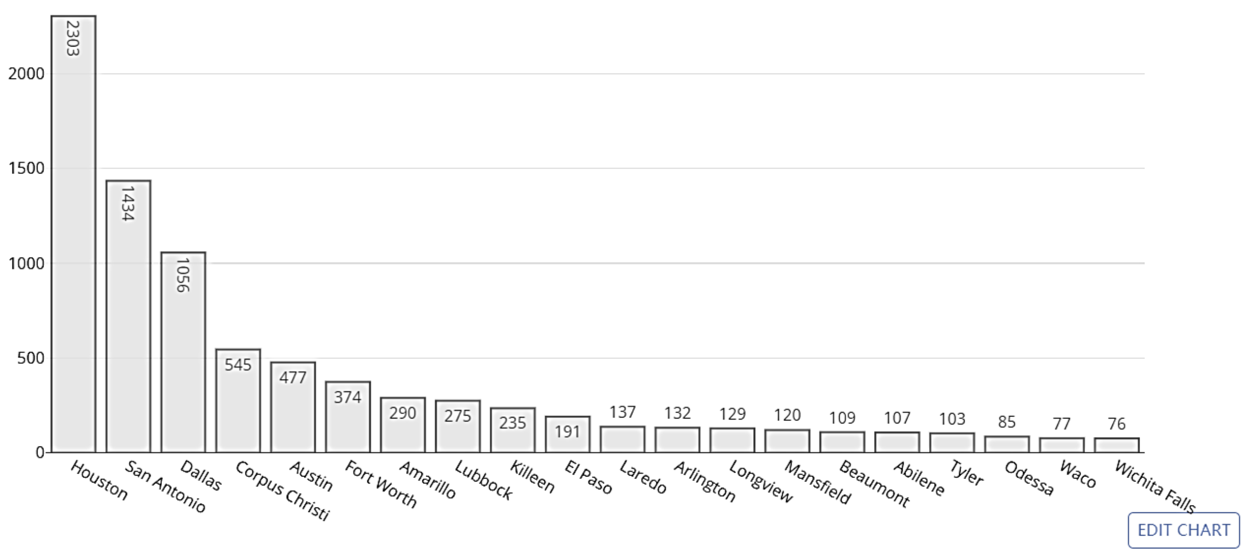
Table 4-i-e: future one-year forecast

<table>
<tr><td rowspan="2">Date</td><td colspan="9">2018</td><td colspan="3">2019</td></tr>
<tr><td>Apr</td><td>May</td><td>June</td><td>July</td><td>Aug</td><td>Sep</td><td>Oct</td><td>Nov</td><td>Dec</td><td>Jan</td><td>Feb</td><td>Mar</td></tr>
<tr><td>#of gun violence</td><td>216</td><td>201</td><td>179</td><td>221</td><td>210</td><td>212</td><td>227</td><td>226</td><td>218</td><td>243</td><td>237</td><td>236</td></tr>
</table>

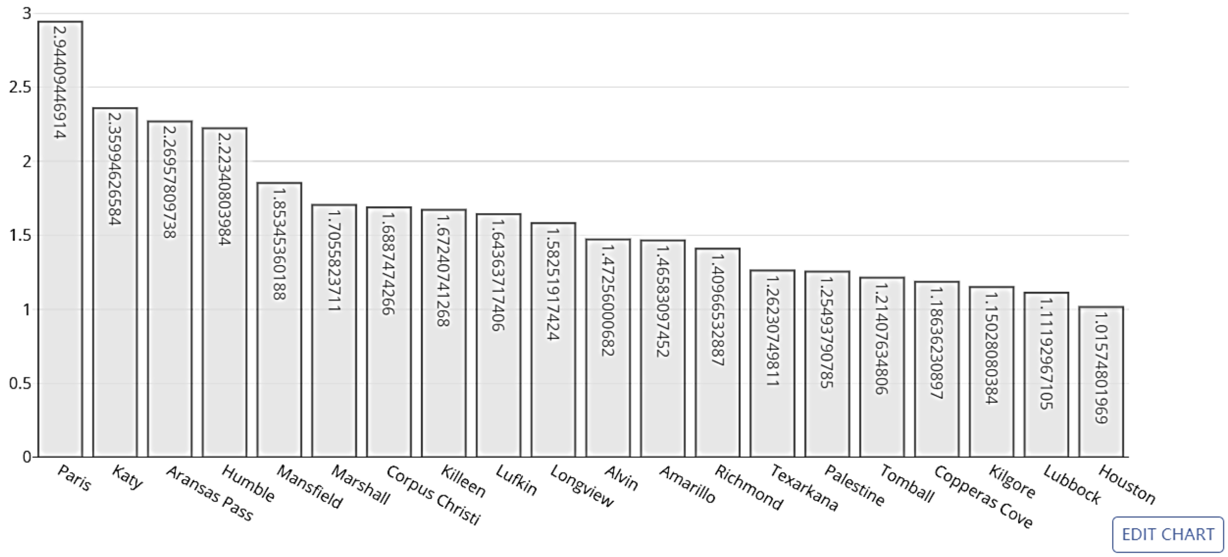
graph 4-ii-a1: Heatmap-Total no. of incidents



graph 4-ii-a2: Total no. of incidents by city-TOP20



graph 4-ii-a3: Incidents per 1k people by city-TOP20



graph 4-ii-a4: Heatmap-Total no. of incidents in Austin

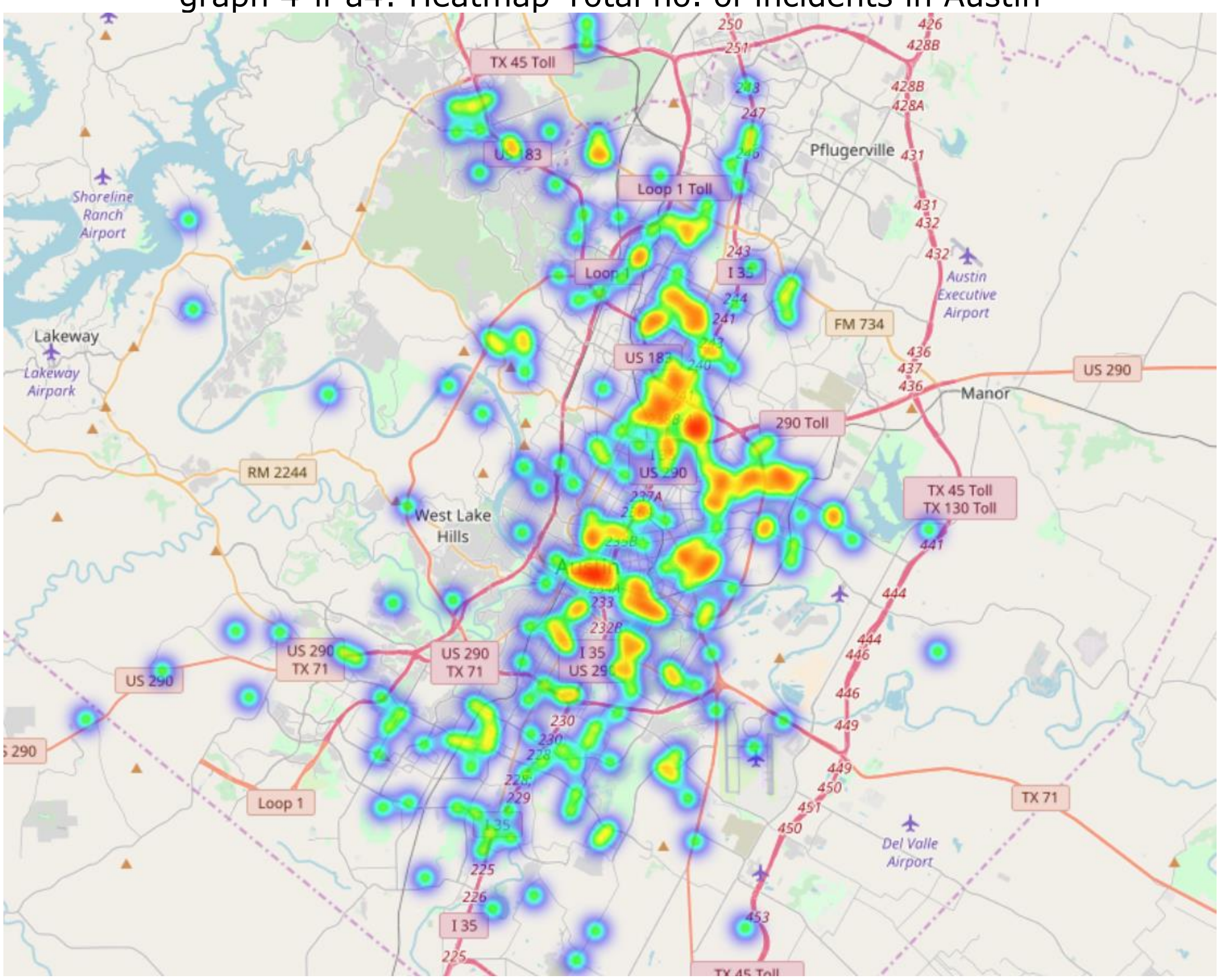
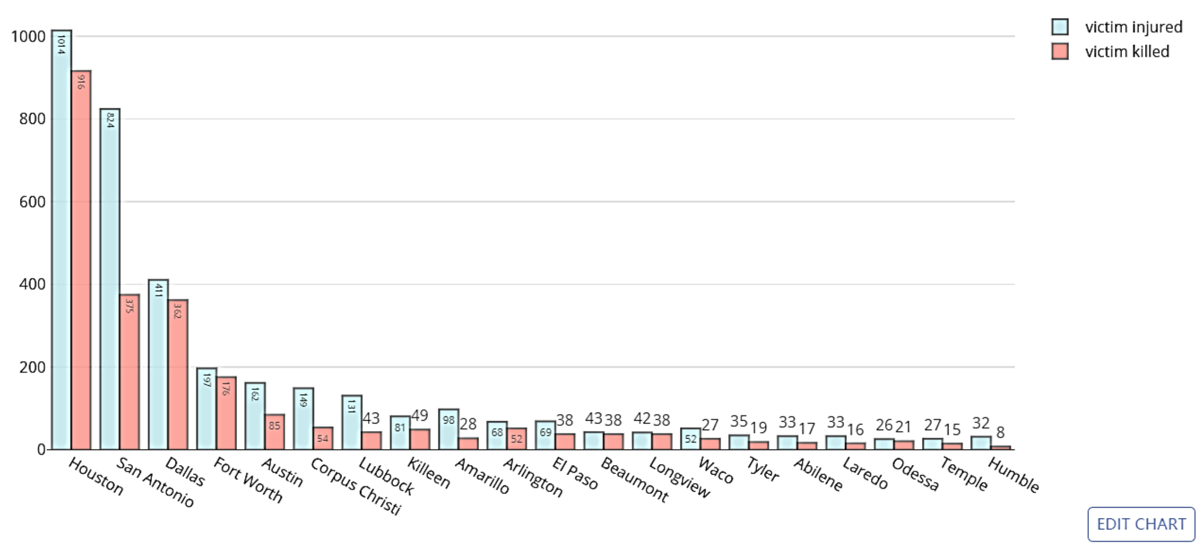


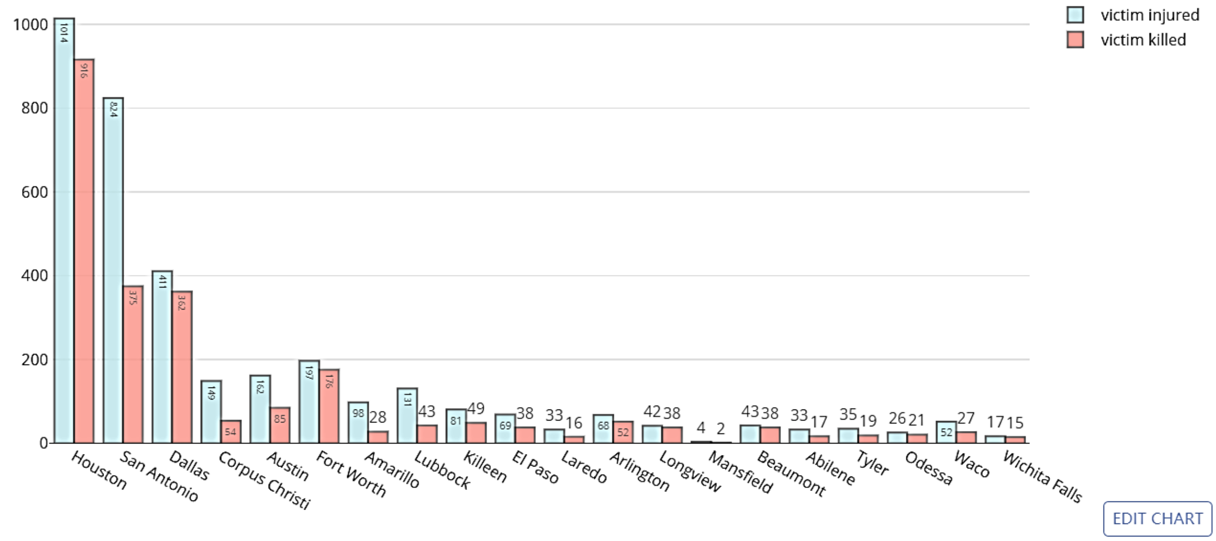
table 4-ii-a1: Cities in both TOP20

	PLACE	per1k_vio_x	Total_vio_x	AVGpop_x
0	Mansfield	1.853454	120	64744.0
1	Corpus Christi	1.688747	545	322724.4
2	Killeen	1.672407	235	140516.0
3	Longview	1.582519	129	81515.6
4	Amarillo	1.465831	290	197840.0
5	Lubbock	1.111930	275	247317.8
6	Houston	1.015748	2303	2267294.6

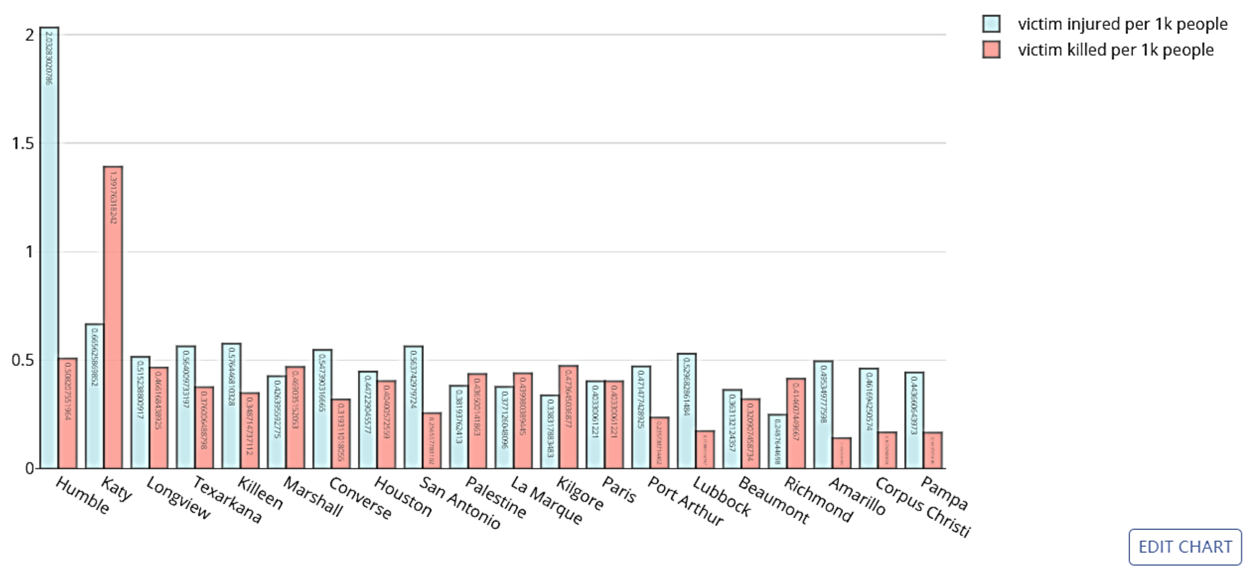
graph 4-ii-b1: Total no. of victims killed and injured by city (rank by total loss)



graph 4-ii-b2: Total no. of victims killed and injured by city (rank by total violence)



graph 4-ii-b3: No. of victims killed and injured per 1k people by city (rank by total loss per 1k people)



graph 4-ii-b4: No. of victims killed and injured per 1k people by city (rank by total violence per 1k people)

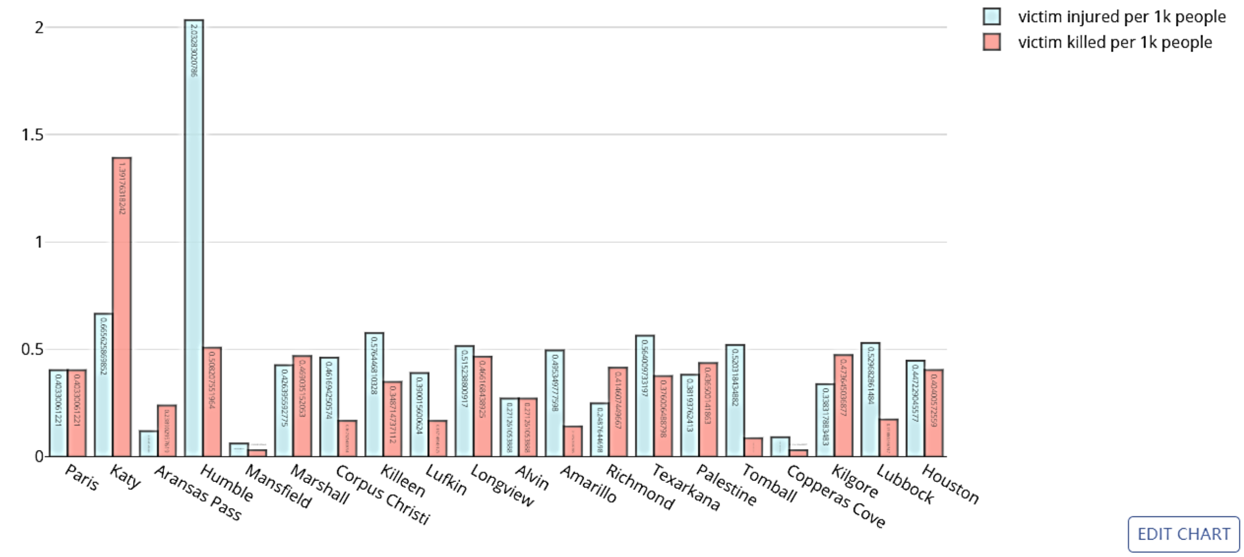


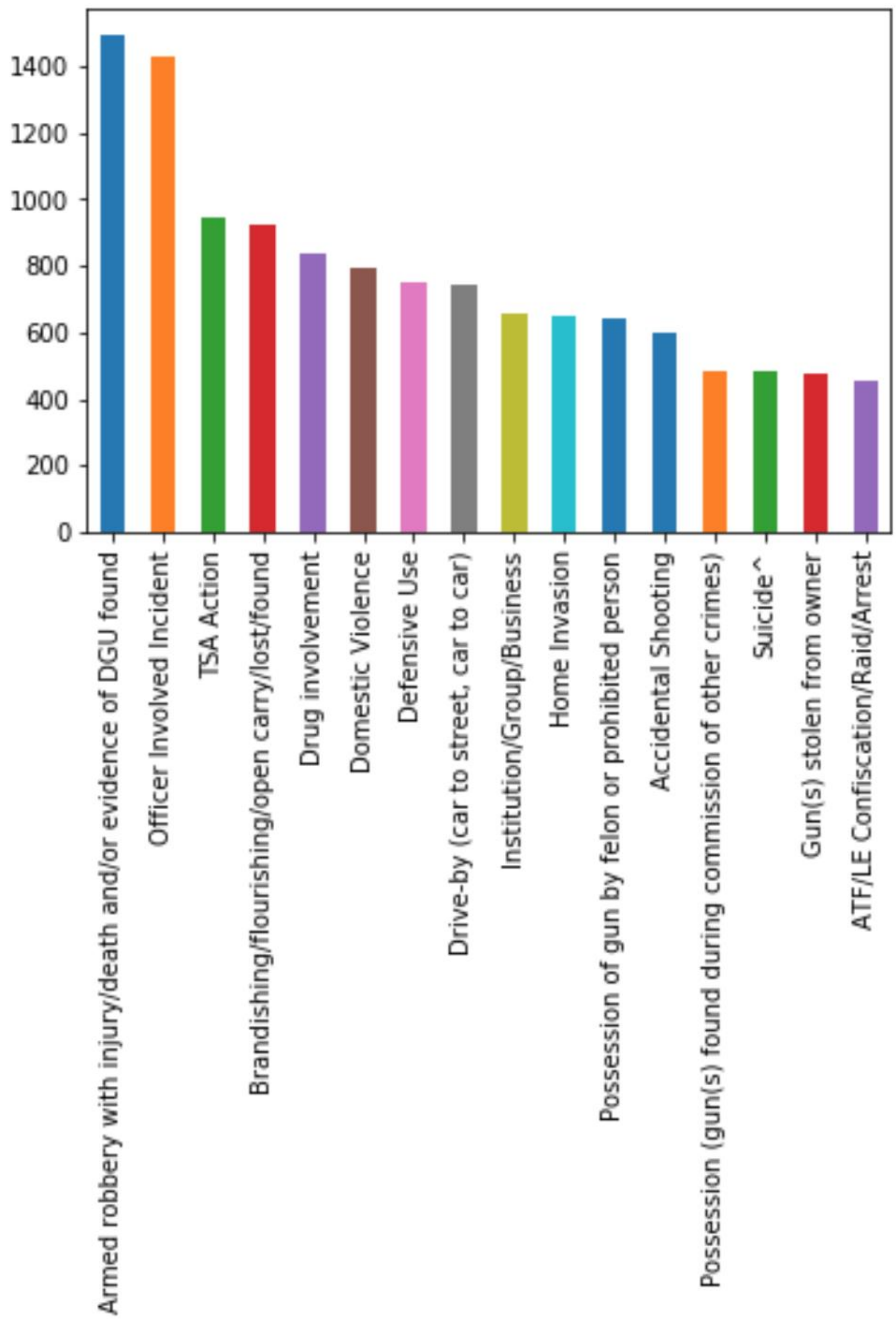
table 4-ii-b1: Cities with more people killed or injured while fewer number of incidents

	PLACE	n_injured	n_killed	total_loss	Total_vio
446	Temple	27	15	42	47
255	Humble	32	8	40	35

table 4-ii-b2: Cities with more people killed or injured per 1k residents while fewer number of incidents per 1k residents

	PLACE	per1k_inj	per1k_kill	total_loss_1k	per1k_vio
157	Converse	0.547390	0.319311	0.866701	0.775470
402	San Antonio	0.563743	0.256558	0.820301	0.981077
289	La Marque	0.377126	0.439980	0.817106	0.879961
375	Port Arthur	0.471477	0.235739	0.707216	0.779751
33	Beaumont	0.363132	0.320907	0.684040	0.920498
360	Pampa	0.443661	0.166373	0.610033	0.720949

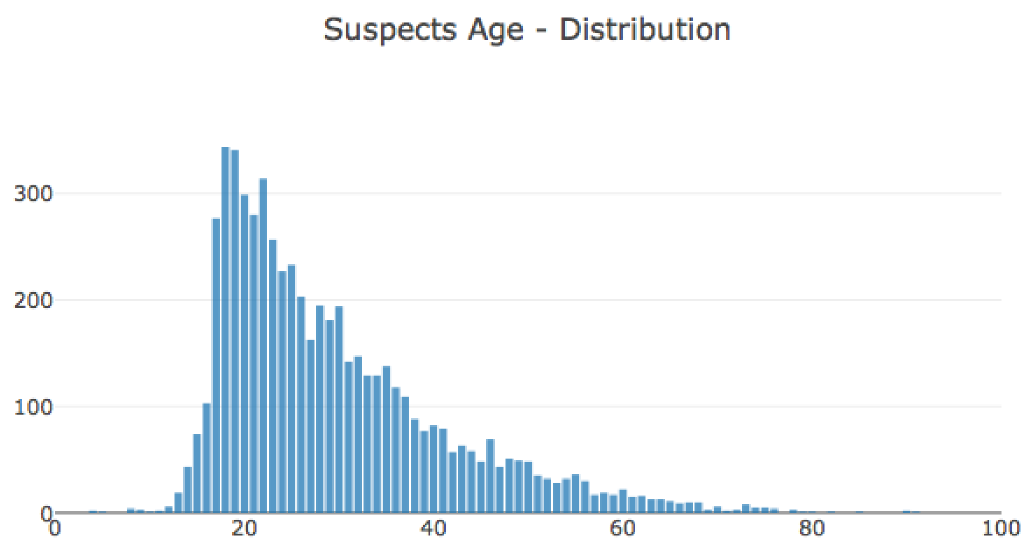
(4-ii-c1)


0
200
400
600
800
1000
1200
1400
Armed robbery with injury/death and/or evidence of DGU found
Officer Involved Incident
TSA Action
Brandishing/flourishing/open carry/lost/found
Drug involvement
Domestic Violence
Defensive Use
Drive-by (car to street, car to car)
Institution/Group/Business
Home Invasion
Possession of gun by felon or prohibited person
Accidental Shooting
Possession (gun(s) found during commission of other crimes)
Suicide^
Gun(s) stolen from owner
ATF/LE Confiscation/Raid/Arrest

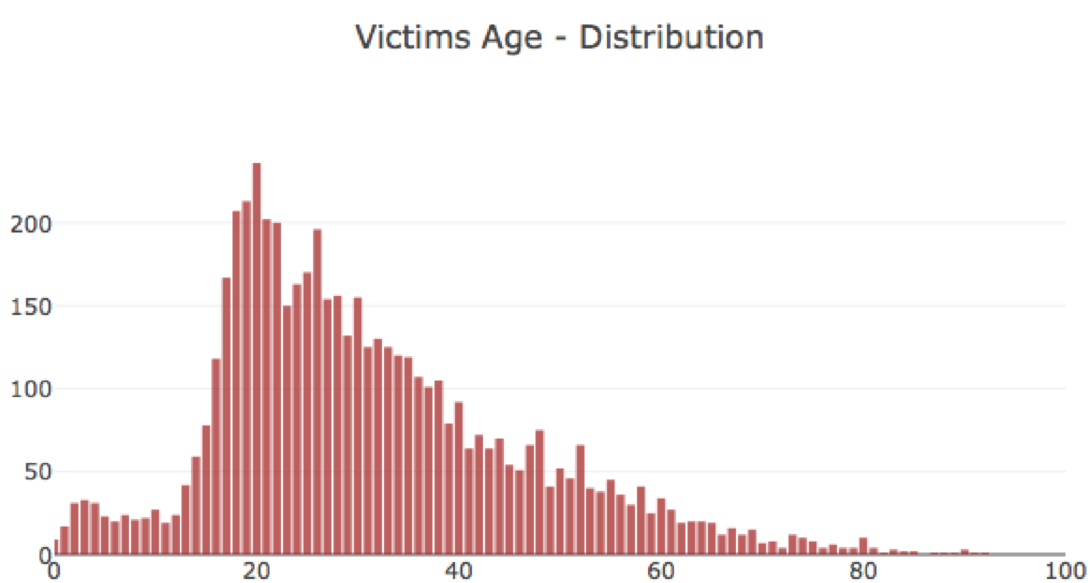
(4-ii-c2)

	Most common incident	Num incident
city_or_county		
Houston	Armed robbery with injury/death and/or evidenc...	394
Dallas	TSA Action	266
San Antonio	Drive-by (car to street, car to car)	182
Austin	TSA Action	108
Corpus Christi	Armed robbery with injury/death and/or evidenc...	68
Fort Worth	Armed robbery with injury/death and/or evidenc...	53
Lubbock	Armed robbery with injury/death and/or evidenc...	42
Irving	TSA Action	40
Amarillo	Armed robbery with injury/death and/or evidenc...	34
El Paso	Officer Involved Incident	31
Arlington	Officer Involved Incident	25
Laredo	Brandishing/flourishing/open carry/lost/found	20
Grand Prairie	Domestic Violence	8
Garland	Armed robbery with injury/death and/or evidenc...	6
Plano	Armed robbery with injury/death and/or evidenc...	5

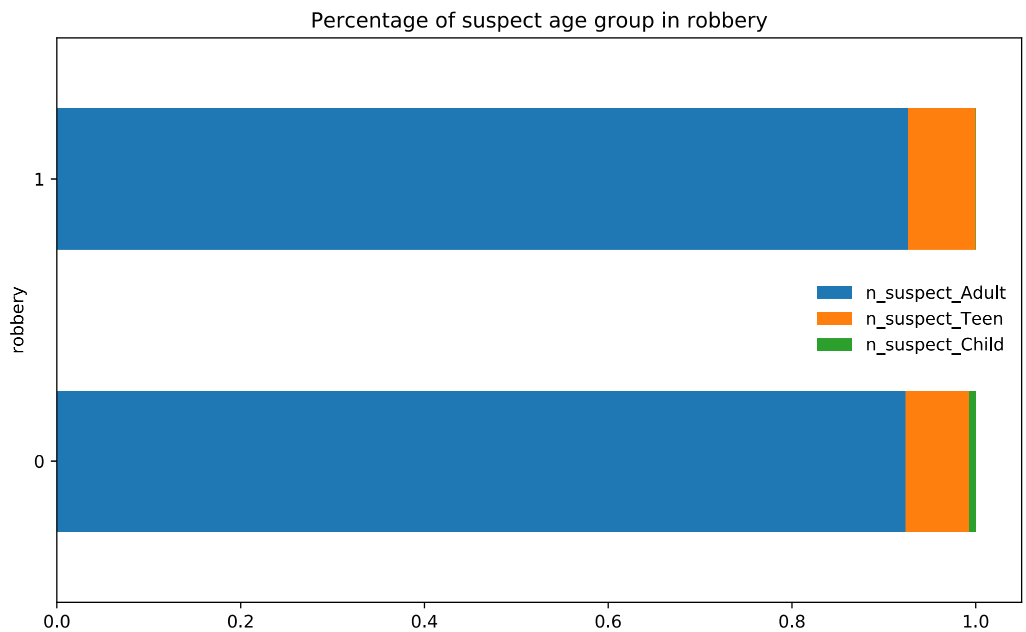
(4-iii-a1)

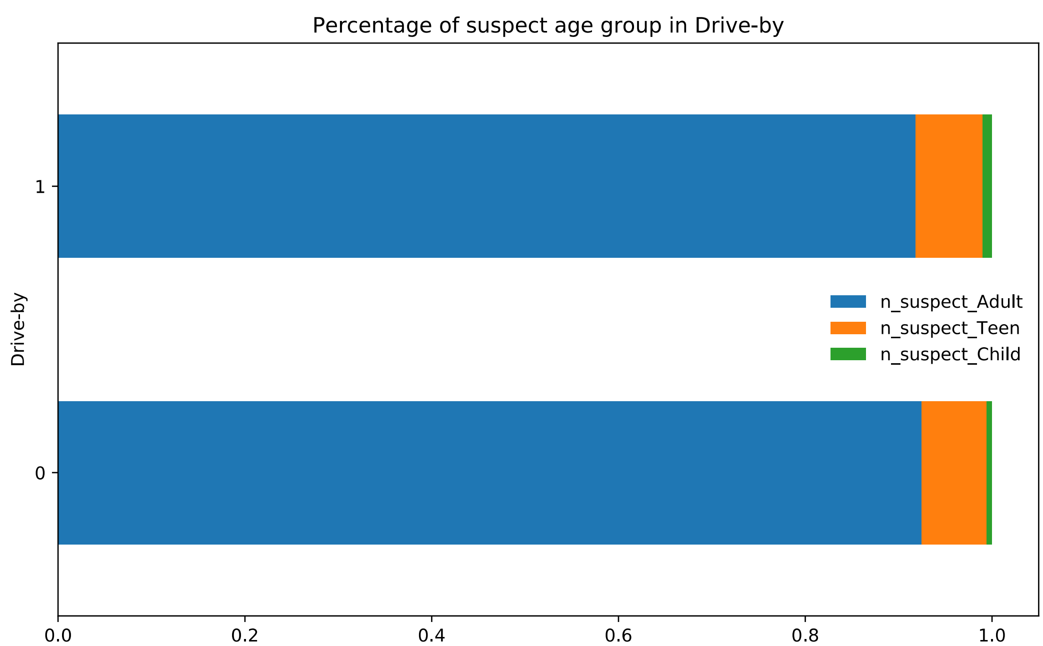


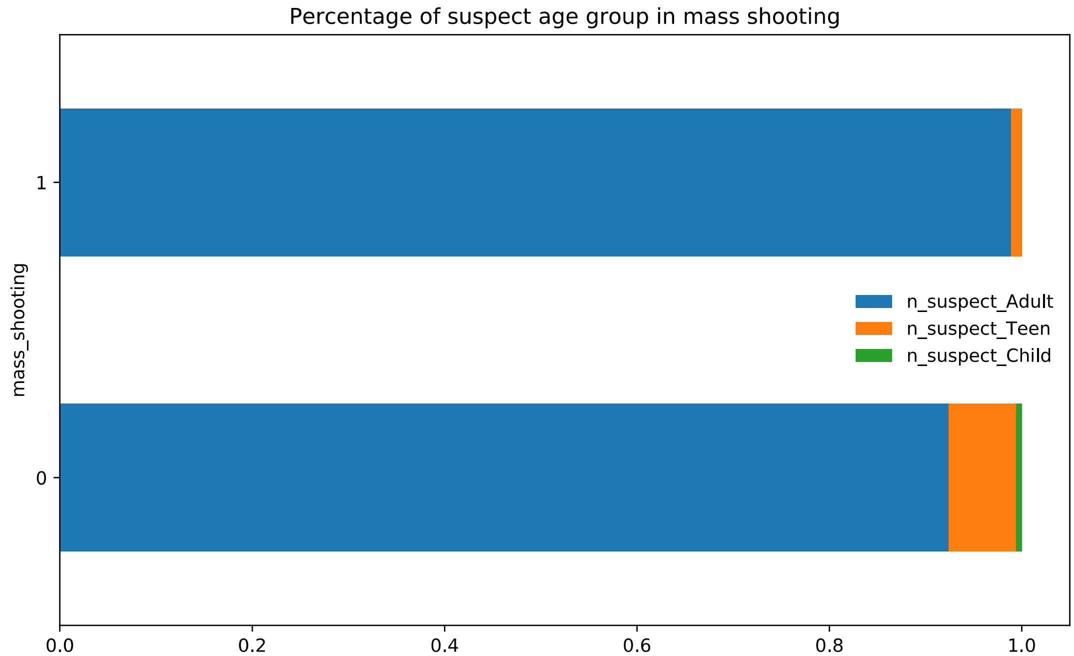
(4-iii-a2)



(4-iii-b1~3)

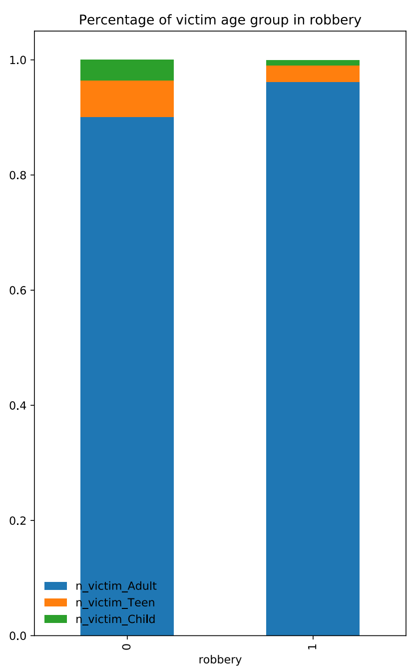

Percentage of suspect age group in robbery
robbery
1
0
n_suspect_Adult
n_suspect_Teen
n_suspect_Child
0.0
0.2
0.4
0.6
0.8
1.0


Percentage of suspect age group in Drive-by
Drive-by
1
0
n_suspect_Adult
n_suspect_Teen
n_suspect_Child
0.0
0.2
0.4
0.6
0.8
1.0

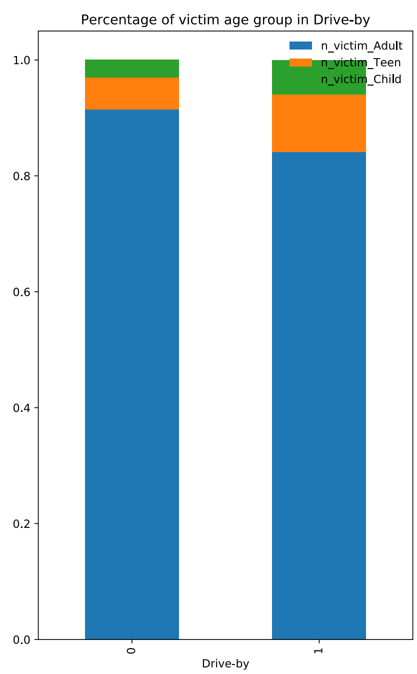


Percentage of suspect age group in mass shooting
mass_shooting
n_suspect_Adult
n_suspect_Teen
n_suspect_Child

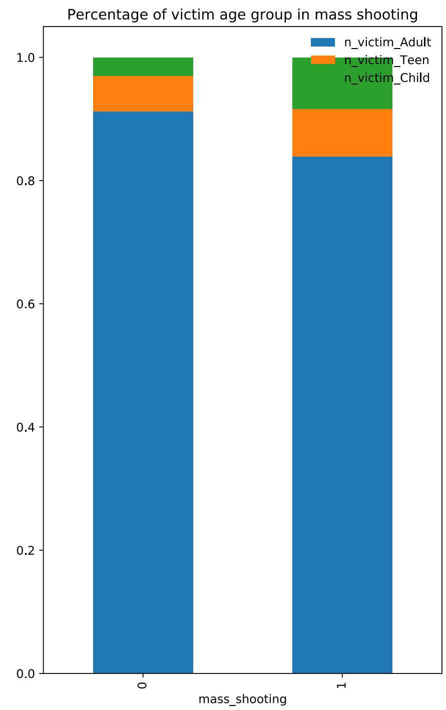
(4-iii-b4~6)



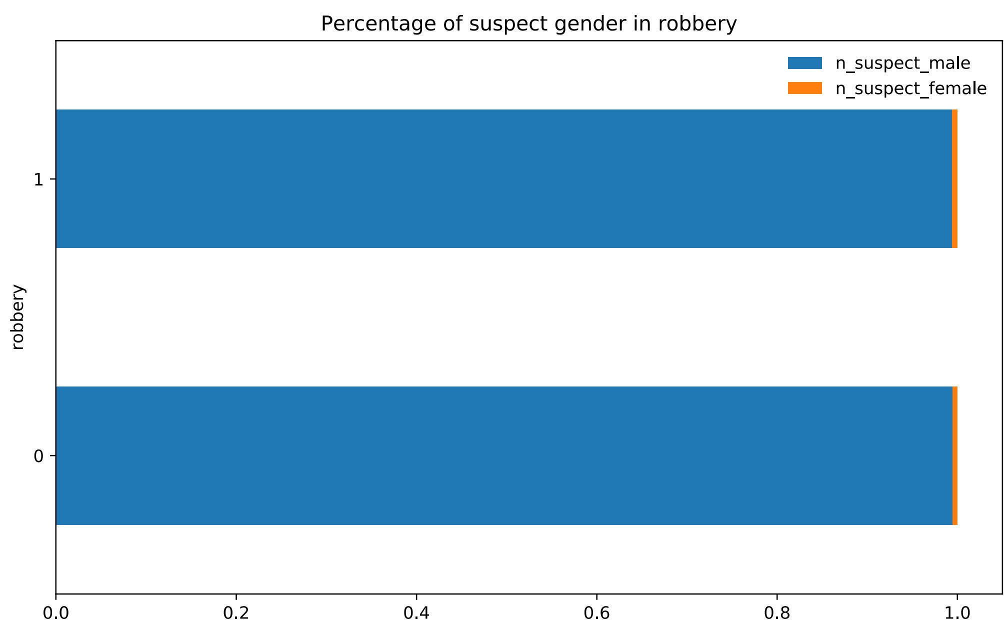
Percentage of victim age group in robbery
n_victim_Adult
n_victim_Teen
n_victim_Child
robbery

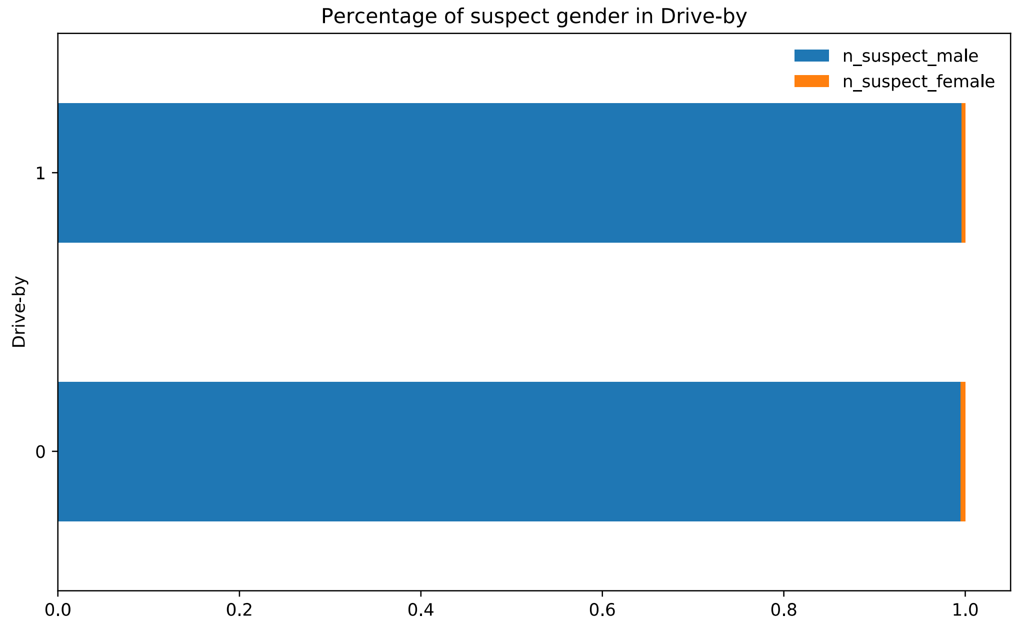


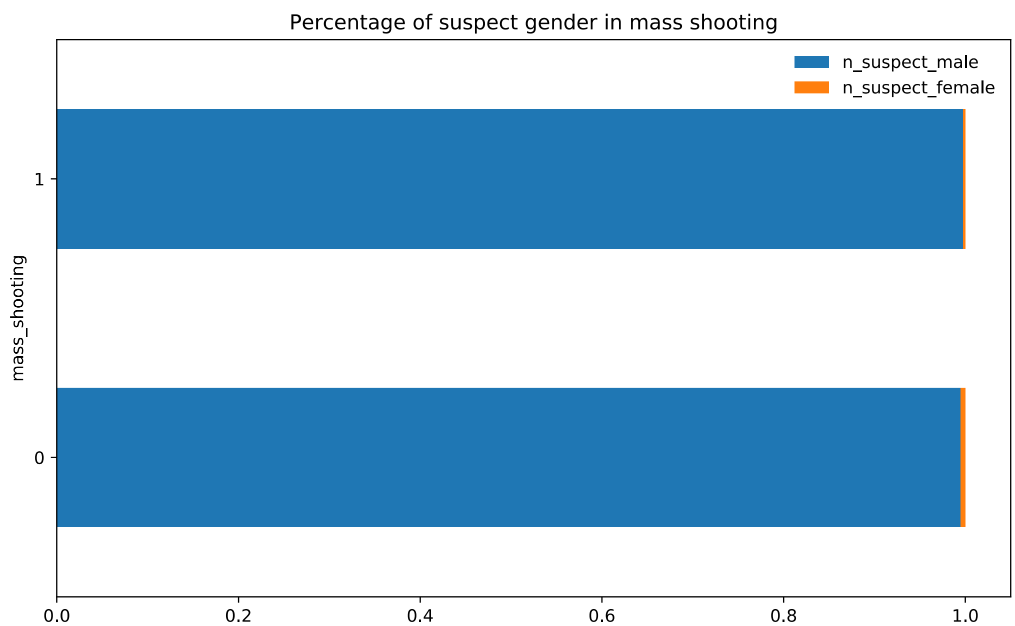
Percentage of victim age group in Drive-by
n_victim_Adult
n_victim_Teen
n_victim_Child
Drive-by


Percentage of victim age group in mass shooting
n_victim_Adult
n_victim_Teen
n_victim_Child
1.0
0.8
0.6
0.4
0.2
0.0
0
1
mass_shooting

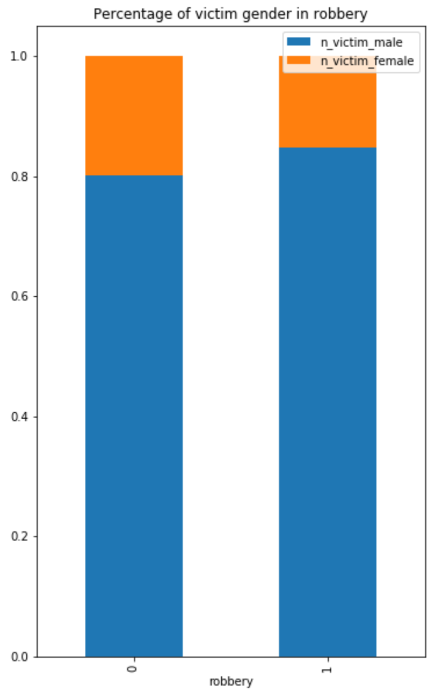
(4-iii-c1~3)

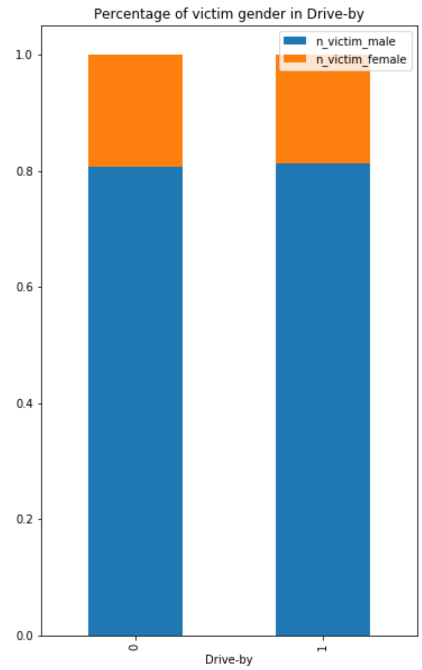

Percentage of suspect gender in robbery
n_suspect_male
n_suspect_female
1
0
robbery
0.0
0.2
0.4
0.6
0.8
1.0

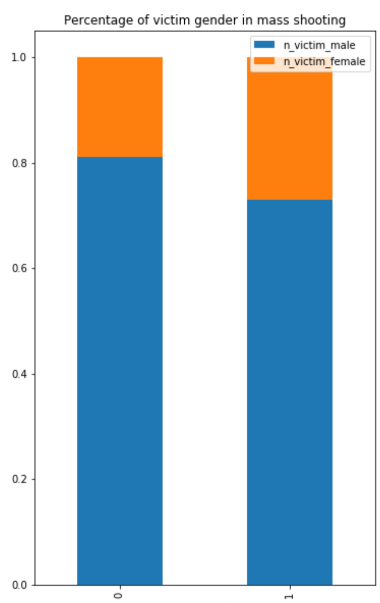

Percentage of suspect gender in Drive-by
n_suspect_male
n_suspect_female
Drive-by
1
0
0.0
0.2
0.4
0.6
0.8
1.0


Percentage of suspect gender in mass shooting
n_suspect_male
n_suspect_female
mass_shooting
1
0
0.0
0.2
0.4
0.6
0.8
1.0

(4-iii-c4~6)


Percentage of victim gender in robbery
n_victim_male
n_victim_female
1.0
0.8
0.6
0.4
0.2
0.0
0
1
robbery


Percentage of victim gender in Drive-by
n_victim_male
n_victim_female
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Drive-by


Percentage of victim gender in mass shooting
n_victim_male
n_victim_female
1.0
0.8
0.6
0.4
0.2
0.0
0
1
mass_shooting