Gun Violence in the US. Application of Unsupervised Learning Methods for Trend Exploration

by Sumaira Afzal, Viraja Ketkar, Murlidhar Loka, Vadim Spirkov

Abstract To do

Background

Objective

The objective of this research is to ...

Data Analysis

The data set used for this research contains 260k of gun violence incidents in the US between January 2013 and March 2018. The data has been soursed from Kaggle.

Originallly the data set was uploaded to Kaggle from Gun Violence Archive (GVA) Web site gunviolencearchive.org. This is a not for profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States. GVA will collect and check for accuracy, comprehensive information about gun-related violence in the U.S. and then post and disseminate it online.

Data Dictionary

Column Name	Column Description
incident_id	Incident ID
date	Date of crime
state	State
city_or_countyCity	City/county of crime
address	Address of the location of the crime
n_killed	Number of people killed
n_injured	Number of people injured
incident_url	URL regarding the incident
source_url	Reference to the reporting source
incident_url_fields_missing	TRUE if the incident_url is present, FALSE otherwise
congressional_district	Congressional district id
gun_stolen	Status of guns involved in the crime (i.e. Unknown, Stolen, etc. . .)
gun_type	Typification of guns used in the crime
incident_characteristics	Characteristics of the incidence
latitude	Location of the incident
location_description	Description of the location
longitude	Location of the incident
n_guns_involved	Number of guns involved in incident
notes	Additional information of the crime
participant_age	Age of participant(s) at the time of crime (victicms nad suspects)
participant_age_group	Age group of participant(s) at the time crime
participant_gender	Gender of participant(s)
participant_name	Name of participant(s) involved in crime
participant_relationship	Relationship of participant to other participant(s)

Column Name	Column Description
participant_status	Extent of harm done to the participant
participant_type	Type of participant (victim or suspect)
sources	Participants source
state_house_district	Voting house district
state_senate_district	Territorial district from which a senator to a state legislature is elected.

Data Exploration

Firstly we are going to load and examine content and statistics of the data set

```
#data = read.csv("../data/gun-violence-data_01-2013_03-2018.csv", header = T,
#                na.strings = c("NA","","#NA"),sep=",")

data = read.csv("../data/gun-violence-sample.csv", header = T,
                na.strings = c("NA","","#NA"),sep=",")

 print(dfSummary(data, valid.col = F, max.distinct.values = 4),
       caption = "\\tt Gun Violence Dataset Summary")
```

Data Frame Summary

data
Dimensions: 5000 x 29
Duplicates: 0

Table 2: `Gun Violence Dataset Summary`

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	incident_id [integer]	Mean (sd) : 559918.8 (293493.9) min < med < max: 92119 < 549418.5 < 1083435 IQR (CV) : 515148.2 (0.5)	5000 distinct values	0 (0%)
2	date [factor]	1. 2013-04-14 2. 2014-01-01 3. 2014-01-02 4. 2014-01-03 [1475 others]	1 (0.0%) 4 (0.1%) 3 (0.1%) 4 (0.1%) 4988 (99.8%)	0 (0%)
3	state [factor]	1. Alabama 2. Alaska 3. Arizona 4. Arkansas [47 others]	95 (1.9%) 31 (0.6%) 50 (1.0%) 62 (1.2%) 4762 (95.2%)	0 (0%)
4	city_or_county [factor]	1. Abbeville 2. Aberdeen 3. Abilene 4. Abingdon [1621 others]	1 (0.0%) 2 (0.0%) 3 (0.1%) 1 (0.0%) 4993 (99.9%)	0 (0%)
5	address [factor]	1. 200 West Henry Street 2. 2100 block of East 12th 3. 2100 block of Pauger Str 4. 5404 S.E. 14th St [4571 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 4631 (99.9%)	365 (7.3%)
6	n_killed [integer]	Mean (sd) : 0.2 (0.5) min < med < max: 0 < 0 < 4 IQR (CV) : 0 (2)	5 distinct values	0 (0%)
7	n_injured [integer]	Mean (sd) : 0.5 (0.7) min < med < max: 0 < 0 < 9 IQR (CV) : 1 (1.5)	10 distinct values	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
8	incident_url [factor]	1. http://www.gunviolencearc/ 2. http://www.gunviolencearc/ 3. http://www.gunviolencearc/ 4. http://www.gunviolencearc/ [4996 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 4996 (99.9%)	0 (0%)
9	source_url [factor]	1. /%20http%3A//ktla.com/201 2. /%20http%3A//www.nola.com 3. /%20http%3A//www.pressand 4. /blog.tsa.gov/2014/04/tsa [4863 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 4990 (99.9%)	6 (0.12%)
10	incident_url_fields_missing [factor]	1. False	5000 (100.0%)	0 (0%)
11	congressional_district [integer]	Mean (sd) : 8 (8.6) min < med < max: 0 < 5 < 53 IQR (CV) : 8.2 (1.1)	53 distinct values	268 (5.36%)
12	gun_stolen [factor]	1. 0::Not-stolen 2. 0::Not-stolen||1::Stolen 3. 0::Stolen 4. 0::Stolen||1::Stolen [53 others]	30 (1.0%) 2 (0.1%) 89 (3.0%) 20 (0.7%) 2802 (95.2%)	2057 (41.14%)
13	gun_type [factor]	1. 0::10mm 2. 0::12 gauge 3. 0::12 gauge||1::12 gauge| 4. 0::20 gauge [167 others]	2 (0.1%) 3 (0.1%) 1 (0.0%) 3 (0.1%) 2935 (99.7%)	2056 (41.12%)
14	incident_characteristics [factor]	1. Accidental Shooting||Acci 2. Accidental Shooting||Acci 3. Accidental Shooting||Acci 4. Accidental Shooting||Acci [1179 others]	2 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 4984 (99.9%)	11 (0.22%)
15	latitude [numeric]	Mean (sd) : 37.6 (5.3) min < med < max: 19.5 < 38.7 < 71.3 IQR (CV) : 7.5 (0.1)	4602 distinct values	191 (3.82%)
16	location_description [factor]	1. "The Grove" business dist 2. (Blacklick) 3. (Brownsville) 4. (Burnside) [788 others]	1 (0.1%) 1 (0.1%) 1 (0.1%) 1 (0.1%) 862 (99.5%)	4134 (82.68%)
17	longitude [numeric]	Mean (sd) : -89.7 (14.8) min < med < max: -159.4 < -86.5 < -68.1 IQR (CV) : 14.6 (-0.2)	4590 distinct values	191 (3.82%)
18	n_guns_involved [integer]	Mean (sd) : 1.6 (8.4) min < med < max: 1 < 1 < 400 IQR (CV) : 0 (5.2)	32 distinct values	2056 (41.12%)

<table>
<tr><th>No</th><th>Variable</th><th>Stats / Values</th><th>Freqs (% of Valid)</th><th>Missing</th></tr>
<tr><td>19</td><td>notes
[factor]</td><td>1. 'heard shots, felt pain.'
2. Man attempted to shoot i
3. Mike's Food Store
4. "...shot himself in the t
[3127 others]</td><td>1 (0.0%)
1 (0.0%)
1 (0.0%)
1 (0.0%)
3296 (99.9%)</td><td>1700
(34%)</td></tr>
<tr><td>20</td><td>participant_age
[factor]</td><td>1. 0::0||1::19
2. 0::1
3. 0::1||1::19||2::20
4. 0::1||1::22
[1045 others]</td><td>1 (0.0%)
1 (0.0%)
1 (0.0%)
1 (0.0%)
3094 (99.9%)</td><td>1902
(38.04%)</td></tr>
<tr><td>21</td><td>participant_age_group
[factor]</td><td>1. 0::Adult 18+
2. 0::Adult 18+||1::Adult 18
3. 0::Adult 18+||1::Adult 18
4. 0::Adult 18+||1::Adult 18
[104 others]</td><td>1962 (47.6%)
1036 (25.1%)
264 (6.4%)
99 (2.4%)
765 (18.5%)</td><td>874
(17.48%)</td></tr>
<tr><td>22</td><td>participant_gender
[factor]</td><td>1. 0::Female
2. 0::Female||1::Female
3. 0::Female||1::Female||2::
4. 0::Female||1::Female||2::
[108 others]</td><td>151 (3.6%)
13 (0.3%)
2 (0.0%)
2 (0.0%)
4055 (96.0%)</td><td>777
(15.54%)</td></tr>
<tr><td>23</td><td>participant_name
[factor]</td><td>1. 0::A.J. Hagner||1::Wayne
2. 0::Aaron Parkinson
3. 0::Aaron Roberts||2::Gary
4. 0::Aaron T Vincent||1::Gr
[2449 others]</td><td>1 (0.0%)
1 (0.0%)
1 (0.0%)
1 (0.0%)
2456 (99.8%)</td><td>2540
(50.8%)</td></tr>
<tr><td>24</td><td>participant_relationship
[factor]</td><td>1. 0::Aquaintance
2. 0::Armed Robbery
3. 0::Armed
Robbery||1::Arme
4. 0::Armed
Robbery||1::Arme
[44 others]</td><td>1 (0.3%)
11 (3.5%)
3 (0.9%)
1 (0.3%)
301 (95.0%)</td><td>4683
(93.66%)</td></tr>
<tr><td>25</td><td>participant_status
[factor]</td><td>1. 0::Arrested
2. 0::Arrested||1::Arrested
3. 0::Arrested||1::Arrested|
4. 0::Arrested||1::Arrested|
[261 others]</td><td>59 (1.3%)
11 (0.2%)
5 (0.1%)
1 (0.0%)
4344 (98.3%)</td><td>580
(11.6%)</td></tr>
<tr><td>26</td><td>participant_type
[factor]</td><td>1. 0::Subject-Suspect
2. 0::Subject-Suspect||1::Su
3. 0::Subject-Suspect||1::Su
4. 0::Subject-Suspect||1::Su
[67 others]</td><td>913 (20.4%)
190 (4.2%)
68 (1.5%)
18 (0.4%)
3282 (73.4%)</td><td>529
(10.58%)</td></tr>
<tr><td>27</td><td>sources
[factor]</td><td>1. http://13wham.com/news/lo/ 2. http://13wham.com/news/to/ 3. http://44news.wevv.com/gu/ 4. http://44news.wevv.com/te/ [4860 others]</td><td>1 (0.0%)
1 (0.0%)
1 (0.0%)
1 (0.0%)
4977 (99.9%)</td><td>19
(0.38%)</td></tr>
<tr><td>28</td><td>state_house_district
[integer]</td><td>Mean (sd) : 55.1 (40.5)
min < med < max:
1 < 46 < 209
IQR (CV) : 64 (0.7)</td><td>183 distinct values</td><td>853
(17.06%)</td></tr>
<tr><td>29</td><td>state_senate_district
[integer]</td><td>Mean (sd) : 20.3 (14.4)
min < med < max:
1 < 18 < 67
IQR (CV) : 22 (0.7)</td><td>65 distinct values</td><td>717
(14.34%)</td></tr>
</table>

Initial observation of the data shows that there is a number of features which do not present any analytical value (Figure: **??**). They are:

- *incident_id*
- *incident_url*
- *source_url*
- *state_house_district*
- *state_senate_district*
- *congressional_district*
- *sources*
- *incident_url_fields_missing*

We also going to drop *participant_age* feature in favour of the *participant_age_group*. The age group is more suatable for categarization and has much less missing data (16% vs 39%).

The reamining features could be groupd as follows.

Participant Features This group describes suspects and victims found on the crime scene. The content of the features of this group is structured as follows: *[idx1::value1||idx2::value2]* (see Figure **??**). This is not quite acceptable for the analytics, thus the particiapnt related features would have to be parsed to extract valuable information about the crime.

It is feasible. *utils.R* script contains *parseFeature* function, which parses *[idx1::value1||idx2::value2]* structure and returnes a named vector object. For example a *participent_type* could be structured as follows:

0	1	2	3
Victim	Victim	Subject-Suspect	Subject-Suspect

Unforunatley *participant_relationship* feature missing **93%** of values. It is not possible to impute the missing data thus we will drop it. For obvious reasons we are aslo going to get rid of *participant_name*. The rest of the participant-related features will be parsed and replaced wit the new categorical attributes. In order to do so we have to understand what possible values each particpant-related feature can have. for this we will employ text mining technique.

We begin with *participant_type* feature

```
participantType = data %>%  mutate(text = trimws(gsub('\\|\\||:|\\|'," ",participant_type,
        fixed = F))) %>% filter(text != "0" ) %>% select(text)

pCorups = VCorpus(VectorSource(participantType))
pCorups  = tm_map(pCorups , removeNumbers)
pTermMatrix = tm::TermDocumentMatrix(pCorups)
# count frequent words
print(tm::findFreqTerms(pTermMatrix, 10))
```

```
[1] "subject-suspect" "victim"
```

As we can see the *participan type* may have two values *vitim* and *subject-suspect*. If the *participant type* is missing we will consider it as **unknown**. Thus we will be employing *participant type* feature as a basis to impute all other participant stats.

Let's find the possible values of *parctipant_age_group* feature (the coded is ommitted).

```
[1] "adult" "child" "teen"
```

Further examination of the feature data shows that there the age group values are:

- Adult 18+
- Teen 12-17
- Child 0-11

Thus using *participant_age_group* feature data we will create two ne ones: *vicitm_age_group* and *suspect_age_group*. These new categorical features will be coded as follows:

- 0 - no info
- 1 - all adults
- 2 - children/ teens
- 3 - adults and children/ teens . Adults make majority
- 4 - adults and children/ teens. Chlidren/ teens make majority

participant_gender could also be parsed and replaced with the coded categorical features as descrived below.

```
participantGender = data %>%  mutate(text = trimws(gsub('\\|\\||:|\\|'," ",participant_gender,
        fixed = F))) %>% filter(text != "0" ) %>% select(text)
pCorups = VCorpus(VectorSource(participantGender))
pCorups  = tm_map(pCorups , removeNumbers)
pTermMatrix = tm::TermDocumentMatrix(pCorups)
print(tm::findFreqTerms(pTermMatrix, 10))
```

```
[1] "female" "male"
```

As a result we will be adding two new features:

victim_gender - gender of the victims *suspect_gender* - gender of the suspects

Gender Codes

- 0 - no info
- 1 - male

- 2 - female
- 3 - male dominated group
- 4 - femail dominated goup

The last feature of the group is *participant_status*. It maintains the outcome of the incident. Let's review the content of the attribute.

```
[1] "arrested"  "injured"   "injured,"  "killed"    "unharmed"  "unharmed,"
```

Based on our findings we will be creating three new numerical features:

- n_victim_killed - number of victims killed
- n_victim_injured - number of victims injured
- n_arrested - number of suspects arrested

Gun Related Features There are three attributes that describe gun types: *gun_stolen*, *gun_type* and *n_guns_invoved* (Figure: **??**) *gun_type* and *gun_stolen* have similar to the participant-related faetures encoding (*[idx1::value1||idx2::value2]*). Thus they also could be parsed and substituded with the categorical features. We begin with the gun type.

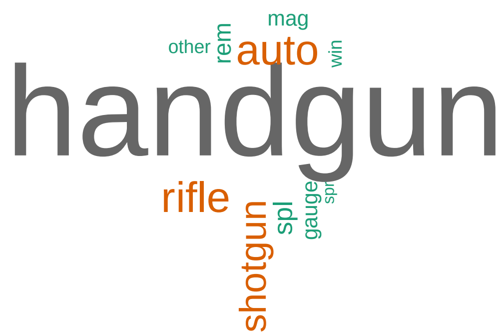

Figure 1: The Most Frequently Used Gun Types

Employing simple text mining techniques we can see that **handgun**, **rifle**, **shotgun** and **auto** make the majority. Thus we will add ane feature *gun_type_involved* to categorize the gun types as follows:

- 0 - unknown
- 1 - handgun
- 2 - shotgun/ rifle
- 3 - automatic
- 4 - mix/other

gun_stolen attribute tells if the gun was stolen or aquired legally. We are going to create a new categorical feature - *gun_origin* which would maintain the folloing data:

- 0 - unknown
- 1 - all stolen
- 2 - all acquired legally
- 3 - mix of stolen and legal guns

Location Related Features To analyze geography of the crimes we will be employing *state*, *city_or_county*, *latitude* and *longitude* atrribute. since we have the coordinates the *address* feature does not present much value for unsupervised learning. We will be using it though to impute missing latitude and longitude values. This activity will be covered in greater details in **Missing Data** paragraph.

Descriptive Features *notes*, *location_description* and *incident_characteristics* are free-text features that might provide additional insights about the crime scene. We are going to take a close look at each feature and decide if we could utilize it.

Lets' begin with the *notes*

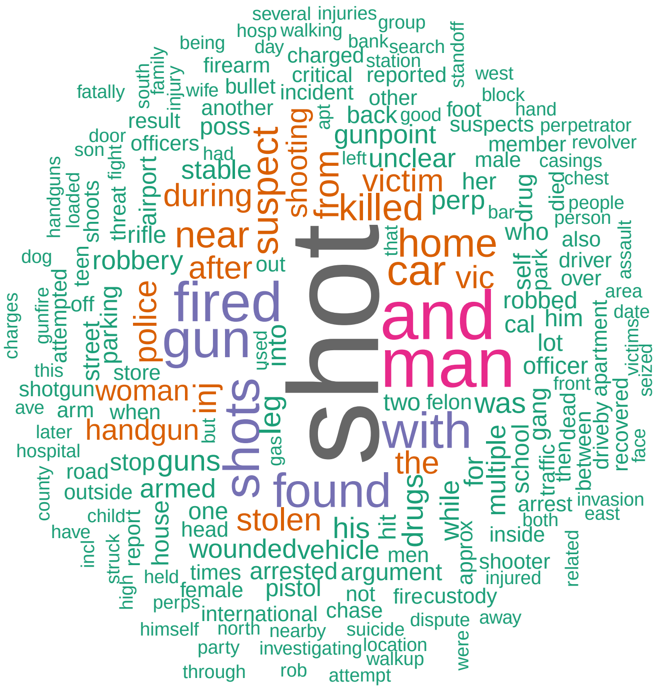

Figure 2: Most Common Words in Notes

Unfortunately *notes* feature does not provide more knowlege to what the others features already supply. Thus it will be dropped.

location_description on the other hand, could be useful to classify location type. Unfortunately 82% of the data is missing. Nonetheless this feature appears to be too important to ignore. Let's see how much we can salvage.

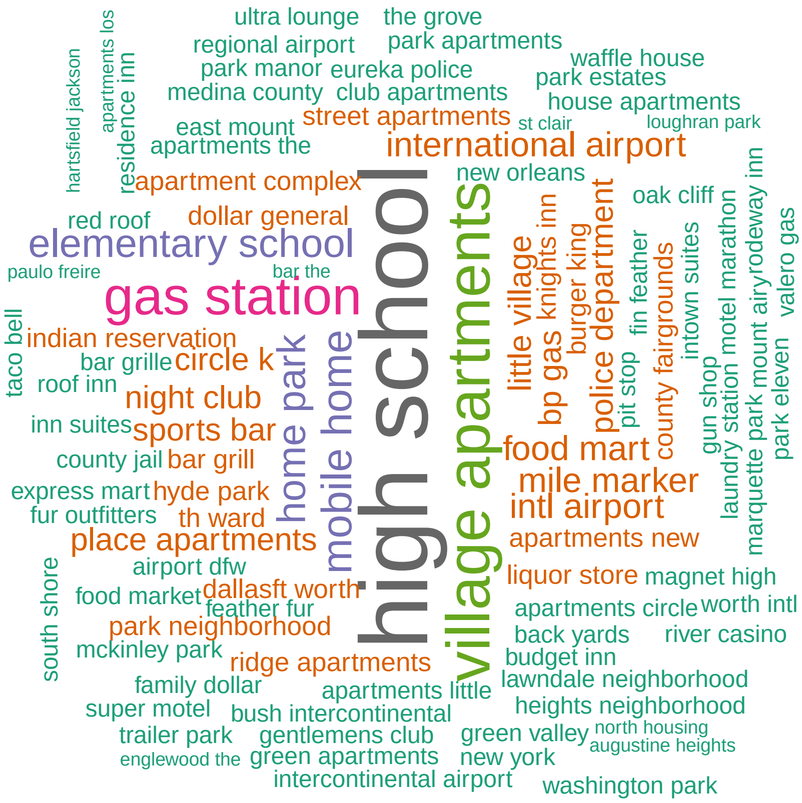

Figure 3: Most Common Bi-gram Terms in Location Decription

As plot 3 shows we could utilize bi-gram terms if the location description is available. Thus let's add new categorical feature - *place_type*, which would have the fllowing values:

- 0 - unknown
- 1 - school/ university/ college
- 2 - community center/ shopping center/ hospital/ church
- 3 - home invasion
- 4 - street/drive by
- 5 - other public places

The last feature in the group is *incident_characteristics*. The feature is almost 100% populated. As with the previous two we are going to find out what info it maintains.

```
                                    word freq
shot wounded                shot wounded 1946
wounded injured          wounded injured 1938
dead murder                  dead murder 1111
murder accidental      murder accidental 1111
shot dead                      shot dead 1111
accidental suicide    accidental suicide 1103
nonshooting incident nonshooting incident  962
shots fired                  shots fired  930
injured shot                injured shot  770
officer involved        officer involved  753
fired no                        fired no  716
no injuries                  no injuries  714
found commission        found commission  639
guns found                    guns found  639
commission crimes      commission crimes  638
possession guns          possession guns  635
involved incident      involved incident  487
subject suspect          subject suspect  423
```

```
suspect perpetrator         suspect perpetrator  423
injury death                       injury death  413
incident officer               incident officer  408
carry lost                           carry lost  401
flourishing open               flourishing open  401
lost found                           lost found  401
open carry                           open carry  401
brandishing flourishing brandishing flourishing  398
death evidence                   death evidence  397
evidence dgu                       evidence dgu  397
robbery injury                   robbery injury  397
dgu found                             dgu found  391
home invasion                     home invasion  391
armed robbery                     armed robbery  390
defensive use                     defensive use  363
atf le                                   atf le  361
confiscation raid             confiscation raid  361
le confiscation                 le confiscation  361
raid arrest                         raid arrest  361
suicide shot                       suicide shot  359
felon prohibited               felon prohibited  354
gun felon                             gun felon  354
prohibited person             prohibited person  354
possession gun                   possession gun  353
drug involvement               drug involvement  345
found shot                           found shot  334
accidental shooting         accidental shooting  320
involved shooting             involved shooting  316
arrest possession             arrest possession  302
car car                                 car car  269
car street                           car street  269
driveby car                         driveby car  269
```

Information the feature provides proved to be useful. It can support two features: *place_type*, which was introduced above and *inceident_type*. The *incident_type* is going to be a categorical attribute with the following codes:

- 0 - unknown
- 1 - accidental
- 2 - defensive use
- 3 - armded robbery
- 4 - suicide
- 5 - raid/ arrest/ warrant
- 6 - domestic violence
- 7 - gun brandishing, flourishing, open demonstration

Missing Data

Takeaways from Data Exploration Excersize

Data Preparation

Data Imputing

Modeling and Evalutation

Feature Selection

Data Upsampling

Partitioning Clustering Approach

Hierarchical Clustering Approach

Density-based Clustering Methods

Clustering Method Evaluation

Model Deployment

Conclusion

Note from the Authors

This file was generated using *The R Journal* style article template, additional information on how to prepare articles for submission is here - Instructions for Authors. The article itself is an executable R Markdown file that could be downloaded from Github with all the necessary artifacts.

Sumaira Afzal
York University School of Continuing Studies

Viraja Ketkar
York University School of Continuing Studies

Murlidhar Loka
York University School of Continuing Studies

Vadim Spirkov
York University School of Continuing Studies