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 sourced from Kaggle.

Originally 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_county	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 (victims and 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=",")
```

Table 2: Gun Violence Dataset Summary

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	incident_id [integer]	Mean (sd) : 559334.3 (293128.7) min < med < max: 92114 < 543587 < 1083472 IQR (CV) : 508683 (0.5)	239677 distinct values	0 (0%)
2	date [factor]	1. 2013-01-01 2. 2013-01-05 3. 2013-01-07 4. 2013-01-19 [1721 others]	3 (0.0%) 1 (0.0%) 2 (0.0%) 1 (0.0%) 239670 (100.0%)	0 (0%)
3	state [factor]	1. Alabama 2. Alaska 3. Arizona 4. Arkansas [47 others]	5471 (2.3%) 1349 (0.6%) 2328 (1.0%) 2842 (1.2%) 227687 (95.0%)	0 (0%)
4	city_or_county [factor]	1. Abbeville 2. Abbotsford 3. Abbott 4. Abbott Township [12894 others]	37 (0.0%) 3 (0.0%) 1 (0.0%) 1 (0.0%) 239635 (100.0%)	0 (0%)
5	address [factor]	1. 100 block of Kohler Cour 2. 100 block of South Sumne 3. 1000 block of 32nd St 4. 10100 block of South Par [198033 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 223176 (100.0%)	16497 (6.88%)
6	n_killed [integer]	Mean (sd) : 0.3 (0.5) min < med < max: 0 < 0 < 50 IQR (CV) : 0 (2.1)	16 distinct values	0 (0%)
7	n_injured [integer]	Mean (sd) : 0.5 (0.7) min < med < max: 0 < 0 < 53 IQR (CV) : 1 (1.5)	23 distinct values	0 (0%)
8	incident_url [factor]	1. http://www.gunviolencearc/ 2. http://www.gunviolencearc/ 3. http://www.gunviolencearc/ 4. http://www.gunviolencearc/ [239673 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 239673 (100.0%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
9	source_url [factor]	1. /%20—8newsnow.com 2. /%20%20http%3A//www.wdsu 3. /%20http%3A//blog.al.com/ 4. /%20http%3A//ktla.com/201 [213985 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 239205 (100.0%)	468 (0.2%)
10	incident_url_fields_missing [factor]	1. False	239677 (100.0%)	0 (0%)
11	congressional_district [integer]	Mean (sd) : 8 (8.5) min < med < max: 0 < 5 < 53 IQR (CV) : 8 (1.1)	54 distinct values	11944 (4.98%)
12	gun_stolen [factor]	1. 0::Not-stolen 2. 0::Not-stolen||1::Not-sto 3. 0::Not-stolen||1::Not-sto 4. 0::Not-stolen||1::Not-sto [345 others]	1352 (1.0%) 45 (0.0%) 10 (0.0%) 7 (0.0%) 138765 (99.0%)	99498 (41.51%)
13	gun_type [factor]	1. 0::10mm 2. 0::10mm||1::22 LR 3. 0::10mm||1::223 Rem [AR-1 4. 0::10mm||1::45 Auto||2::4 [2498 others]	32 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 140191 (100.0%)	99451 (41.49%)
14	incident_characteristics [factor]	1. Accidental Shooting 2. Accidental Shooting||Acci 3. Accidental Shooting||Acci 4. Accidental Shooting||Acci [18122 others]	1 (0.0%) 20 (0.0%) 8 (0.0%) 1 (0.0%) 239321 (100.0%)	326 (0.14%)
15	latitude [numeric]	Mean (sd) : 37.5 (5.1) min < med < max: 19.1 < 38.6 < 71.3 IQR (CV) : 7.5 (0.1)	101240 distinct values	7923 (3.31%)
16	location_description [factor]	1. 'Taste' Dessert Bar 2. "Anderson Island" 3. "Canadian Shores" 4. "Central West End" [27591 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 42085 (100.0%)	197588 (82.44%)
17	longitude [numeric]	Mean (sd) : -89.3 (14.4) min < med < max: -171.4 < -86.2 < 97.4 IQR (CV) : 14.1 (-0.2)	112347 distinct values	7923 (3.31%)
18	n_guns_involved [integer]	Mean (sd) : 1.4 (4.7) min < med < max: 1 < 1 < 400 IQR (CV) : 0 (3.4)	106 distinct values	99451 (41.49%)
19	notes [factor]	1. ' When asked what was goi 2. 'heard shots, felt pain' 3. 'heard shots, felt pain.' 4. 'Heartless Felon' threate [136648 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 158656 (100.0%)	81017 (33.8%)
20	participant_age [factor]	1. 0::0 2. 0::0||1::1||2::28||3::24 3. 0::0||1::18 4. 0::0||1::18||2::20 [18947 others]	12 (0.0%) 1 (0.0%) 2 (0.0%) 1 (0.0%) 147363 (100.0%)	92298 (38.51%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
21	participant_age_group [factor]	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 [894 others]	94671 (47.9%) 49273 (24.9%) 13893 (7.0%) 1 (0.0%) 39720 (20.1%)	42119 (17.57%)
22	participant_gender [factor]	1. 0::Female 2. 0::Female||1::Female 3. 0::Female||1::Female||2:: 4. 0::Female||1::Female||2:: [869 others]	7791 (3.8%) 918 (0.5%) 102 (0.1%) 13 (0.0%) 194491 (95.7%)	36362 (15.17%)
23	participant_name [factor]	1. 0::"Bear"||1::Cardell Mon 2. 0::"Bo" 3. 0::"Chicago" 4. 0::"Chucky" [113484 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 117420 (100.0%)	122253 (51.01%)
24	participant_relationship [factor]	1. 0::Aquaintance 2. 0::Aquaintance||1::Aquain 3. 0::Aquaintance||1::Aquain 4. 0::Armed Robbery [280 others]	63 (0.4%) 8 (0.1%) 1 (0.0%) 370 (2.3%) 15332 (97.2%)	223903 (93.42%)
25	participant_status [factor]	1. 0::Arrested 2. 0::Arrested||1::Arrested 3. 0::Arrested||1::Arrested| 4. 0::Arrested||1::Arrested| [2146 others]	2739 (1.3%) 487 (0.2%) 175 (0.1%) 73 (0.0%) 208577 (98.4%)	27626 (11.53%)
26	participant_type [factor]	1. 0::Subject-Suspect 2. 0::Subject-Suspect||1::Su 3. 0::Subject-Suspect||1::Su 4. 0::Subject-Suspect||1::Su [255 others]	44914 (20.9%) 8922 (4.2%) 3040 (1.4%) 1267 (0.6%) 156671 (72.9%)	24863 (10.37%)
27	sources [factor]	1. http: //1160wccs.com/blair/ 2. http: //1160wccs.com/polic/ 3. http: //1160wccs.com/polic/ 4. http: //1350kman.com/fort-/ [217276 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 239064 (100.0%)	609 (0.25%)
28	state_house_district [integer]	Mean (sd) : 55.4 (42) min < med < max: 1 < 47 < 901 IQR (CV) : 63 (0.8)	275 distinct values	38772 (16.18%)
29	state_senate_district [integer]	Mean (sd) : 20.5 (14.2) min < med < max: 1 < 19 < 94 IQR (CV) : 21 (0.7)	68 distinct values	32335 (13.49%)

Initial observation of the data shows that there is a number of features which do not present any analytical value (Table: 2). 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 favor of the *participant_age_group*. The age group is more suitable for categorization and has much less missing data (16% vs 39%).

The remaining features could be grouped 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 Table: 2). This is not quite acceptable for the analytics, thus the participant 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 returns a named vector object. For example a *participent_type* could be structured as follows:

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

Unfortunately *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 also 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 participant-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 omitted).

```
[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 new 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. Children/ teens make majority

participant_gender could also be parsed and replaced with the coded categorical features as described 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 - female dominated group

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"    "killed,"   "unharmed"
[7] "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* (Table: 2) *gun_type* and *gun_stolen* have similar to the participant-related features encoding (*[idx1::value1||idx2::value2]*). Thus they also could be parsed and substituted 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 another new 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 acquired legally. We are going to create a new categorical feature - *gun_origin* which would maintain the following 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* attribute. 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 knowledge 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.

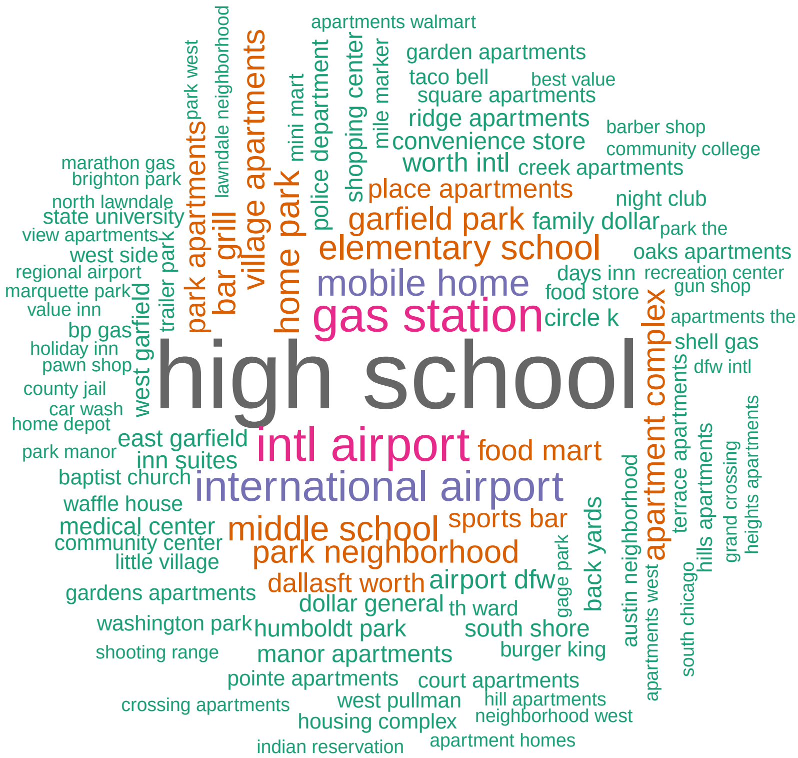

Figure 3: Most Common Bi-gram Terms in Location Decription

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 following 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.

```
print(xtable(tmp[1:50,1:2],
      caption = "\\tt Most Frequently Used Incident Characteristics  Bi-grams"),
      row.names = FALSE,  include.rownames = FALSE)
```

word	freq
shot wounded	93926.00
wounded injured	93313.00
dead murder	53409.00
murder accidental	53409.00
shot dead	53272.00
accidental suicide	52967.00
shots fired	45895.00
nonshooting incident	44761.00
officer involved	38229.00
injured shot	37426.00
fired no	35750.00
no injuries	35552.00
found commission	30863.00
guns found	30863.00
commission crimes	30720.00
possession guns	30646.00
involved incident	23860.00
subject suspect	21886.00
suspect perpetrator	21881.00
home invasion	21244.00
injury death	20540.00
incident officer	20166.00
death evidence	19723.00
evidence dgu	19723.00
robbery injury	19723.00
armed robbery	19502.00
dgu found	19383.00
carry lost	19017.00
flourishing open	19017.00
open carry	19017.00
lost found	18975.00
brandishing flourishing	18519.00
confiscation raid	17991.00
le confiscation	17991.00
atf le	17966.00
raid arrest	17959.00
felon prohibited	17165.00
gun felon	17165.00
prohibited person	17158.00
suicide shot	17153.00
possession gun	17137.00
drug involvement	16976.00
defensive use	16824.00
found shot	16689.00
involved shooting	16658.00
accidental shooting	15641.00
arrest possession	14942.00
car street	13655.00
street car	13655.00
car car	13630.00

Information the *incident_characteristics* 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 - armed robbery
- 4 - suicide
- 5 - raid/ arrest/ warrant
- 6 - domestic violence
- 7 - gun brandishing, flourishing, open demonstration _ We are also be adding a feature that indicates if drugs or alcohol was involved: *is_drug_alcohol*

Date Feature In addition to *date* of incident attribute we add *month* and *day_of_week* to identify any seasonal patterns.

Data Preparation

Prior to generating new features as discussed in the previous paragraph we would need to impute missing latitude and longitude data. To do so we employ OpenCage forward geocoding API. Unlike Google this company offers a free tier. To save time we imputed the missing geo-coordinates and saved the result in the file. The code below is submitted for demonstration purpose only.

```
imputeCoordinates()
```

Now we are going to remove the features identified as redundant

```
data = subset(data, select = c(-incident_id, -incident_url, -source_url,
  -state_house_district, -state_senate_district,-sources, -incident_url_fields_missing,
  -congressional_district, -address, -participant_age, -participant_name,
  -participant_relationship, -notes))
```

Lastly we are going to loop through the entire data fame imputing missing data and adding new features. Again this is a lengthy process that takes about 1.5 hours to finish. The code is also quite long. Thus in order not to clutter the report we submit the code just in the script to illustrates the process, but will not output it into the report.

After we added new feature it is time to remove the columns that are no longer relevant and save the result into a file to be used for unsupervised learning

```
data = subset(data, select = c(-participant_age_group, -participant_type,
      -participant_gender, -participant_status, -location_description,
      -incident_characteristics, -gun_stolen,-gun_type))
```

Resulting Dataset

After a rather lengthy process, we finaly have reached the stage when our data set is eady to be used for exploration by clustering algorithms. This is the summary of the resulting data.

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

Table 4: `Engineered Gun Violence Dataset Summary`

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
1	date [factor]	1. 2013-01-01 2. 2013-01-05 3. 2013-01-07 4. 2013-01-19 [1721 others]	3 (0.0%) 1 (0.0%) 2 (0.0%) 1 (0.0%) 239670 (100.0%)	0 (0%)
2	state [factor]	1. Alabama 2. Alaska 3. Arizona 4. Arkansas [47 others]	5471 (2.3%) 1349 (0.6%) 2328 (1.0%) 2842 (1.2%) 227687 (95.0%)	0 (0%)
3	city_or_county [factor]	1. Abbeville 2. Abbotsford 3. Abbott 4. Abbott Township [12894 others]	37 (0.0%) 3 (0.0%) 1 (0.0%) 1 (0.0%) 239635 (100.0%)	0 (0%)
4	n_killed [integer]	Mean (sd) : 0.3 (0.5) min < med < max: 0 < 0 < 50 IQR (CV) : 0 (2.1)	16 distinct values	0 (0%)
5	n_injured [integer]	Mean (sd) : 0.5 (0.7) min < med < max: 0 < 0 < 53 IQR (CV) : 1 (1.5)	23 distinct values	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
6	latitude [numeric]	Mean (sd) : 37.5 (5.2) min < med < max: -39 < 38.6 < 71.3 IQR (CV) : 7.5 (0.1)	107051 distinct values	0 (0%)
7	longitude [numeric]	Mean (sd) : -89.2 (15) min < med < max: -171.4 < -86.2 < 176.2 IQR (CV) : 14.1 (-0.2)	118198 distinct values	0 (0%)
8	n_guns_involved [integer]	Mean (sd) : 0.8 (3.6) min < med < max: 0 < 1 < 400 IQR (CV) : 1 (4.5)	107 distinct values	0 (0%)
9	month [integer]	Mean (sd) : 6.4 (3.4) min < med < max: 1 < 6 < 12 IQR (CV) : 6 (0.5)	12 distinct values	0 (0%)
10	day_of_week [integer]	Mean (sd) : 4.1 (2) min < med < max: 1 < 4 < 7 IQR (CV) : 4 (0.5)	7 distinct values	0 (0%)
11	victim_gender [integer]	Mean (sd) : 0.7 (0.8) min < med < max: 0 < 1 < 4 IQR (CV) : 1 (1.1)	5 distinct values	0 (0%)
12	suspect_gender [integer]	Mean (sd) : 0.6 (0.7) min < med < max: 0 < 1 < 4 IQR (CV) : 1 (1.1)	5 distinct values	0 (0%)
13	victim_age_group [integer]	Mean (sd) : 0.6 (0.7) min < med < max: 0 < 1 < 4 IQR (CV) : 1 (1.1)	5 distinct values	0 (0%)
14	suspect_age_group [integer]	Mean (sd) : 0.6 (0.7) min < med < max: 0 < 1 < 4 IQR (CV) : 1 (1.1)	5 distinct values	0 (0%)
15	n_victim_killed [integer]	Mean (sd) : 0.2 (0.5) min < med < max: 0 < 0 < 49 IQR (CV) : 0 (2.2)	15 distinct values	0 (0%)
16	n_victim_injured [integer]	Mean (sd) : 0.5 (0.7) min < med < max: 0 < 0 < 53 IQR (CV) : 1 (1.6)	23 distinct values	0 (0%)
17	n_victims [integer]	Mean (sd) : 0.8 (0.8) min < med < max: 0 < 1 < 102 IQR (CV) : 1 (1.1)	26 distinct values	0 (0%)
18	n_suspects [integer]	Mean (sd) : 0.8 (1) min < med < max: 0 < 1 < 63 IQR (CV) : 1 (1.2)	33 distinct values	0 (0%)
19	n_arrested [integer]	Mean (sd) : 0.4 (0.8) min < med < max: 0 < 0 < 63 IQR (CV) : 1 (2)	31 distinct values	0 (0%)
20	gun_type_involved [integer]	Mean (sd) : 0.3 (0.8) min < med < max: 0 < 0 < 4 IQR (CV) : 0 (3)	5 distinct values	0 (0%)
21	gun_origin [integer]	Min : 0 Mean : 0 Max : 1	0 : 230802 (96.3%) 1 : 8875 (3.7%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Missing
22	place_type [integer]	Mean (sd) : 0.8 (1.7) min < med < max: 0 < 0 < 5 IQR (CV) : 0 (2)	6 distinct values	0 (0%)
23	incident_type [integer]	Mean (sd) : 1.3 (2.2) min < med < max: 0 < 0 < 7 IQR (CV) : 2 (1.7)	6 distinct values	0 (0%)
24	is_drug_alcohol [integer]	Min : 0 Mean : 0.1 Max : 1	0 : 210310 (87.8%) 1 : 29367 (12.2%)	0 (0%)

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