SCHOOL OF HUMAN SCIENCES & TECHNOLOGY

GUN VIOLENCE ANALYSIS IN U.S.A

REPORT BY MBD S2
TING-LUN, FAN

DATASET INTRODUCTION

Gun violence in U.S.A analysis

Background

"Gun Violence Data" is a data set that contains detailed information about each gun violence incidents in United States, including region, gun type, age...etc. This dataset is from Gun Violence Archive (GVA) This dataset has 29 unique columns with 240k rows.

Column Explanation

- incident_id
- date (date of crime)
- state (state of crime)
- 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_missingincident_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)
- incident_characteristics (ex: Shot - Dead, Shot - Wounded/Injured...etc)
- latitude (Location of the incident)
- location_description
- longitude (Location of the incident)
- n_guns_involved (Number of guns involved in incident)
- notes (Additional information of the crime)
- participant_age
- participant_age_group
- participant_gender
- participant_name
- participant_relationship (Relationship of participant to other participant(s))
- participant_status (Extent of harm done to the participant)
- participant_type
- sources (Participants source)
- state_house_district (Voting house district)
- state_senate_district (Territorial district from which a senator to a state legislature is elected.)

Goal of Analysis

Understand the gun violence incident geographically

According to report, 30,000 of women, men and children are killed every year in America by guns. Gun violence incident is a serious and tragic issue in United States. If we can find out the potential pattern or regularity about these incidents such as the frequency of gun violence incident in different region, we can definitely increase our awareness or safety level, or even launch new regulation based on the research.

Thus, the ultimate goal for this analysis is to "Find out the frequency and level of gun violence incident in different region geographically. " In other words, we use geographic type column as main criteria to understand which place are more dangerous. To achieve this, we have created the following step.

1. Select useful and relevant columns for our analysis
2. Create evaluate indicator and analyze different region, then focus on the most severe place.
3. Find out the average severity level of each incident in different region.
4. Create a table with all relevant column analyzed based on geographic criteria.

The reason we choose geographic column as our main criteria is because each state in U.S.A has totally different regulation, culture and situation, so the severity of gun violence will also be different, this can be a good indicator.

Through these steps, we expected to understand the different level of each states and the severity of each incident, then create a completed table that contain all helpful information.

Gun Violence in U.S.A Analysis

Step 1: PySpark environment setup

Step 2: Data source and spark data abstraction (dataframe) setup

Extract gun_violence_data.csv as GunDF

Step 3: Dataset metadata analysis

A. Display schema and size of the DataFrame

Check the type of each column and total rows

B. Get one or multiple random samples from the data set

Step 4: Data cleaning

A. Check missing value percentage in every column

Make sure all data is analyzable and accurate, if there's any missing value we have to solve it.

B. Drop the columns which missing value are more than 15% and irrelevant to our analysis

If there are more than 15% of data missing in one column, inserting other value such as mean or median might create bias, so we delete it. Also, we drop the column which are not irrelevant to our analysis topic, which is not related to region. Thus, we delete those columns and keep the relevant columns for further analysis.

C. Check missing value for the rest of the columns

Check those columns, we can see that there are missing values in column state, city_or_county, n_killed and n_injured. For different columns we have different solution.

D. Deal with missing values -- 'state', 'city_or_county' and 'date'

State and city_or_county are two main analytical criteria, the geographical attribute is better to remove rather than fill it. Also we do the same with date, so we delete those rows with missing value directly.

E. Deal with missing values -- 'n_killed' and 'n_injured'

The missing value of column n_killed and and n_injured are all smaller than 5 %. Thus, for n_killed and n_injured, we fill mode into all missing value to lower the effect of bias. The mode value is 0.

Step 5: Column basic profiling

A. Switch datatype of "n_killed" and "n_injured" from str to int

We can see that both column n_injured and n_killed are string datatype, for further aggregation calculation we have to switch them into integer. In this step we switch it.

B. Get summary profile

Get summary profile and work separately with same attribute, including:

"Summary of columns **date**"

"Distinct values amount in columns **date**"

"Summary of columns **state**, **city_or_county**"

"Distinct values amount in columns **state**, **city_or_county**"

"Summary of columns **n_killed**, **n_injured**"

"Distinct values amount in columns **n_killed**, **n_injured**"

Step 6: Analysis 1_State

A. Calculate each state with their total incident separately

In this part, we want to know the amount of gun violence incident in each state, from the table A, we can see that the states are align based on the amount of incident. The reason we are doing this is because our goal is to analyze the gun violence geographically, with state, we can clearly see which state is more dangerous potentially.

A.

state	totalincident
Illinois	17556
California	16306
Florida	15029
Texas	13577
Ohio	10244
New York	9712
Pennsylvania	8929
Georgia	8925
North Carolina	8739
Louisiana	8103
Tennessee	7626
South Carolina	6939
Missouri	6631
Michigan	6136
Massachusetts	5981
Virginia	5949
Indiana	5852
Maryland	5798
Alabama	5471
New Jersey	5387

B.

state	totalincident	GunViolenceSeverityLevel
Illinois	17556	Black Alert
California	16306	Black Alert
Florida	15029	Black Alert
Texas	13577	Black Alert
Ohio	10244	Black Alert
New York	9712	Red Alert
Pennsylvania	8929	Red Alert
Georgia	8925	Red Alert
North Carolina	8739	Red Alert
Louisiana	8103	Red Alert
Tennessee	7626	Yellow Alert
South Carolina	6939	Yellow Alert
Missouri	6631	Yellow Alert
Michigan	6136	Yellow Alert
Massachusetts	5981	Blue Alert
Virginia	5949	Blue Alert
Indiana	5852	Blue Alert
Maryland	5798	Blue Alert
Alabama	5471	Blue Alert
New Jersey	5387	Blue Alert

B. Create alert level for different amount of gun violence incident

By creating a column for indicating the different level of gun violence severity, from table B we can see black indicates the most dangerous state. This allow us to clearly understand the different dangerous level of each state, even if the table is not order by amount, this also help us cluster into different group and easier for other deeper analysis.

C. Select the most serious level- Black alert for the following analysis

In our analysis, we aim to understand the dangerous state and its city inside, so we filter the states with black alert as our main analysis group. Table C shows the filter result for our further analysis.

state	totalincident	GunViolenceSeverityLevel
Illinois	17556	Black Alert
California	16306	Black Alert
Florida	15029	Black Alert
Texas	13577	Black Alert
Ohio	10244	Black Alert

Step 7: Analysis 2_Black alert state analysis - Illinois as an example

A. Get deeper in states with black alert - city or county - Illinois as an example

We have filtered the most severe states in U.S.A, however the range of each state is too big to clearly understand where is actually dangerous. Thus, we start to get deeper into its city and county. We use the same method, ranking city or county with the amount of its gun violence incident, and filter the top ones. Table D shows the top dangerous place in Illinois state, we can clearly see that the specific dangerous place in Illinois - If you want to travel Illinois, be careful if you are in Chicago!

D.

state	city_or_county	totalincident
Illinois	Chicago	10814
Illinois	Peoria	920
Illinois	Rockford	842
Illinois	Chicago (Englewood)	542
Illinois	Springfield	303

E.

state	city_or_county	averagekilled
inois	Manchester	6.0
inois	Dwight	3.0
inois	Geneseo	2.5
inois	Danforth	2.0
inois	Bay View Gardens	2.0
inois	Braidwood	2.0
inois	Warsaw	2.0
inois	Columbia	2.0
inois	Benton (West City)	2.0
inois	Morton Grove	2.0
inois	Avon	2.0
inois	Oakwood	2.0
inois	Hazel Crest	1.67
inois	Cherry Valley	1.5

B. Average of people get killed in city or county - Illinois as an example

Another indicator is number of killed in each incident, its helpful if we can also understand the severity for each gun violence incident. For example, maybe in some city, there is not too many incidents, however, they have very fierce fight each time, with a lot of people get killed or injured. By using aggregate function, we calculate the average amount of people get killed in each incident. This allow us to better understand the average severity level of each city or county. Then we still need to take care of these specific area. With this analysis, we can offer a more complete and unbiased analysis.

C. Average of people get injured in city or county - Illinois as an example

The same indicator as n_killed, we use the content of these two columns to better analyze each region.(Table F)

F.

state	city_or_county	averageinjured
Illinois	Centreville	5.0
Illinois	Hanna City	3.0
Illinois	Lake Bluff	3.0
Illinois	Calumet Park	2.0
Illinois	Mchenry (county)	2.0

D. Merge table together - Illinois as an example

Now we have all table with different analysis, the last step is to merge it together to create a completed table with all the columns, we use join function to merge table E and table F and become table G, then merge table G with table D to get our final table H. This table shows our ultimate analysis goal - to understand which regions are more dangerous geographically.

G.

state	city_or_county	averageinjured	averagekilled
Illinois	Addison	0.5	1.0
Illinois	Braidwood	0.0	2.0
Illinois	Chicago (Englewood)	1.01	0.26
Illinois	Lockport	0.27	0.27
Illinois	Dekalb	0.11	0.22
Illinois	Kendall (county)	0.0	0.0
Illinois	Libertyville	0.0	0.5
Illinois	O'fallon	0.0	0.0
Illinois	Homer Glen	0.5	1.0
Illinois	Melrose Park	0.43	0.71
Illinois	New Boston	0.0	0.0
Illinois	Quincy	0.29	0.08
Illinois	Crest Hill	0.0	0.0
Illinois	Lynn Center	0.0	0.0
Illinois	Marengo	0.67	0.0
Illinois	Saline (county)	0.5	0.0
Illinois	Spring Valley	0.0	0.0
Illinois	Wood River	0.14	0.0
Illinois	Carlock	0.0	1.0
Illinois	Grayville	1.0	0.0

H.

state	city_or_county	totalincident	averageinjured	averagekilled	GunViolenceSeverityLevel
Illinois	Chicago	10814	0.96	0.19	Black Alert
Illinois	Peoria	920	0.32	0.04	Green Alert
Illinois	Rockford	842	0.47	0.09	Green Alert
Illinois	Chicago (Englewood)	542	1.01	0.26	Green Alert
Illinois	Springfield	303	0.37	0.13	Green Alert
Illinois	Champaign	213	0.38	0.12	Green Alert
Illinois	Joliet	198	0.46	0.23	Green Alert
Illinois	Aurora	191	0.45	0.15	Green Alert
Illinois	Kankakee	186	0.34	0.1	Green Alert
Illinois	Chicago (Roseland)	159	0.96	0.26	Green Alert
Illinois	Decatur	140	0.54	0.16	Green Alert
Illinois	East Saint Louis	120	0.37	0.59	Green Alert
Illinois	Danville	102	0.57	0.25	Green Alert
Illinois	Urbana	95	0.26	0.07	Green Alert
Illinois	Evanston	84	0.33	0.1	Green Alert
Illinois	Carbondale	81	0.38	0.07	Green Alert
Illinois	Belleville	69	0.28	0.12	Green Alert
Illinois	Chicago (Chicago ...	67	1.01	0.24	Green Alert
Illinois	Alton	66	0.38	0.14	Green Alert
Illinois	Quincy	66	0.29	0.08	Green Alert

Conclusion

Our goal is to understand the dangerous level of gun violence in different region of U.S.A, and we choose region as our main analysis criteria, the analysis clearly shows the following observation.

- Different severity level of each state
- Most dangerous state, city and county with highest frequency of incident.
- Severity level can be main focus criteria, however the severity level of each incident is also important, don't relax!

This analysis is useful for...

- Traveler : Get fully prepared before travelling!
- Government/ police office : To understand with area is more dangerous, then they can either primarily focus on those area or announce alert to travelers and residents.

On the other hand, we can also use our table to understand the safest place, it not listed but the logic and method are all the same!