# PROJECT REPORT

## Enterprise Cloud Computing and Big Data

## (BUDT737)

## Gun Incident Analysis (2013 - 2018)

Arjun Rao Kaveti

Neelapu Venkata Sai Dhanushree

Sreelakshmi Suresh

# ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

| | **Typed Name** | **Signature** |
|---|---|---|
| 1 | Arjun Rao Kaveti | Arjun Rao Kaveti |
| 2 | Neelapu Venkata Sai Dhanushree | Neelapu Venkata Sai Dhanushree |
| 3 | Sreelakshmi Suresh | Sreelakshmi Suresh |

# EXECUTIVE SUMMARY

## Introduction

This report presents our comprehensive analysis of gun incident data spanning the years 2013 to 2018, employing advanced PySpark tools. Our investigation primarily centered on understanding incident severity and frequency variations across states and over time. Through meticulous examination, we identified dynamic trends in incidence rates, shedding light on the nuanced nature of gun-related incidents. Additionally, our analysis delved into incident characteristics, including the types of firearms involved and the demographics of affected individuals. This deeper exploration revealed potential risk factors and patterns contributing to gun violence, offering valuable insights for policymakers and stakeholders. By leveraging evidence-based strategies informed by our findings, we aim to address the multifaceted challenges posed by gun violence and promote community safety and well-being. Ultimately, our analysis serves as a crucial step towards fostering informed decision-making and implementing targeted interventions aimed at reducing the incidence of gun-related incidents and mitigating their impact on individuals and society.

# DATA DESCRIPTION

**Data source-** Kaggle

**Data link-** https://www.kaggle.com/code/erikbruin/gun-violence-in-the-us-eda-and-rshiny-app/report

**Data Size-** 29 Variables and 239,678 rows in our data.

**A Small Sample of Observations:-**

| Sr No | Variable Name | Data Type | Example |
|---|---|---|---|
| 1 | incident_id | dbl | 461105, 460726, 478855,478925, 4789… |
| 2 | date | chr | "2013-01-01", "2013-01-01", |
| 3 | state | chr | "Pennsylvania", "California", "Ohio"… |
| 4 | city_or_county | chr | "Mckeesport", "Hawthorne", "Lorain",… |
| 5 | address | chr | "1506 Versailles Avenue and Coursin … |
| 6 | n_killed | dbl | 0, 1, 1, 4, 2, 4, 5, 0, 0, 1, 1, 1, … |
| 7 | n_injured | dbl | 4, 3, 3, 0, 2, 0, 0, 5, 4, 6, 3, 3, … |
| 8 | incident_url | chr | "http://www.gunviolencearchive.org/i… |
| 9 | source_url | chr | "http://www.post-gazette.com/local/s… |
| 10 | incident_url_fields_missing | lgl | FALSE, FALSE, FALSE, FALSE, FALSE, F… |
| 11 | congressional_district | int | 14, 43, 9, 6, 6, 1, 1, 2, 9, 7, 3, 1… |
| 12 | gun_stolen | chr | NA, NA, "0::Unknown\|\|1::Unknown", NA… |
| 13 | gun_type | chr | NA, NA, "0::Unknown\|\|1::Unknown", NA… |
| 14 | incident_characteristics | chr | "Shot - Wounded/Injured\|\|Mass Shooti… |
| 15 | latitude | dbl | 40.3467, 33.9090, 41.4455, 39.6518, … |
| 16 | location_description | chr | NA, NA, "Cotton Club", NA, NA, "Fair… |
| 17 | longitude | dbl | -79.8559, -118.3330, -82.1377, -104.… |
| 18 | n_guns_involved | dbl | NA, NA, 2, NA, 2, NA, 2, NA, NA, NA,… |

| | | | |
|---|---|---|---|
| 19 | notes | chr | "Julian Sims under investigation: Fo… |
| 20 | participant_age | chr | "0::20", "0::20","0::25\|\|1::31\|\|2::… |
| 21 | participant_age_group | chr | "0::Adult 18+\|\|1::Adult 18+\|\|2::Adult… |
| 22 | participant_gender | chr | "0::Male\|\|1::Male\|\|3::Male\|\|4::Female… |
| 23 | participant_name | chr | "0::Julian Sims", "0::Bernard Gillis… |
| 24 | participant_relationship | chr | NA, NA, NA, NA, "3::Family", NA, "5:… |
| 25 | participant_status | chr | "0::Arrested\|\|1::Injured\|\|2::Injured… |
| 26 | participant_type | chr | "0::Victim\|\|1::Victim\|\|2::Victim\|\|3:… |
| 27 | sources | chr | "http://pittsburgh.cbslocal.com/2013… |
| 28 | state_house_district | int | NA, 62, 56, 40, 62, 72, 10, 93, 11, … |
| 29 | state_senate_district | int | NA, 35, 13, 28, 27, 11, 14, 5, 7, 44… |

**Why is our data interesting?**

Understanding the patterns and characteristics of gun violence incidents in the United States is of paramount importance for policymakers, law enforcement agencies, researchers, and advocacy groups. By analyzing such data, stakeholders can identify trends, risk factors, and vulnerable populations, leading to the development of targeted interventions aimed at reducing the incidence of gun-related violence and promoting community safety and well-being.

# RESEARCH QUESTIONS

In investigating gun violence data from 2013 to 2018, our project aims to address several critical questions that bear directly on public safety, law enforcement strategies, and policymaking. These questions, framed in the context of the business area of public safety and policy development, include:

**Geographical Distribution of Gun Violence:** Which states and cities witness the highest incidence of gun violence? Are there identifiable hotspots where gun violence is significantly more prevalent?

**Temporal Trends in Gun Violence:** How has gun violence changed over time? Are there specific years or months when incidents spike, suggesting seasonal trends or responses to legislative changes?

**Severity of Incidents:** What patterns can be observed in the severity of gun violence incidents? How does the severity correlate with geographical location, time of year, or other factors?

**Effectiveness of Legislation**: Can any correlations be drawn between changes in gun violence incidence and the enactment of specific gun control laws at state or federal levels?

**Community and Demographic Factors**: How do community and demographic factors influence the prevalence and severity of gun violence? Are certain communities disproportionately affected?

# METHODOLOGY

**1. Data Collection**

- Utilization of a comprehensive dataset recording gun incidents in the U.S. from 2013 to 2018.
- Why?: A robust dataset ensures that the analysis covers a wide range of incidents across different states and years, allowing for a more thorough understanding of gun violence trends in the United States.

**2. Exploratory Data Analysis (EDA)**

- Statistical summaries, visualization tools (e.g., histograms, box plots, scatter plots), and missing value analysis.
- Why?: EDA is critical for getting acquainted with the dataset's structure, identifying anomalies or outliers, understanding the distribution of key variables, and spotting any potential biases or inconsistencies. This stage sets the direction for more detailed analysis and informs the preprocessing steps.

**3. Preprocessing**

- String splitting with regex for extracting structured information from textual data, and VectorAssembler for combining multiple features into a single vector feature.
- Why?: Preprocessing prepares the raw data for analysis by cleaning and transforming it. Regex is used for its ability to parse complex string patterns, essential for extracting meaningful information from unstructured data fields. VectorAssembler is crucial for machine learning tasks in PySpark, as it aggregates features into a format suitable for model input, facilitating efficient analysis.

**4. Clustering**

- Machine learning algorithms such as K-means for unsupervised classification of incidents.
- Why?: Clustering helps in identifying natural groupings within the data based on similarities across several dimensions, such as incident severity, geographical location, and temporal factors. These insights can reveal underlying patterns and trends in gun violence, which are not immediately apparent, informing targeted interventions.

**5. Graph Analysis**

- Graph frames for modeling relationships between entities (e.g., incidents, states, years) in the dataset.
- Why?: Graph analysis offers a powerful way to visualize and analyze complex relationships and interdependencies within the data. By representing incidents and their attributes as nodes and

edges, one can uncover connectivity patterns, identify clusters, and explore the influence of geographical and temporal factors on gun violence in a more nuanced way.

**6. Streaming Analysis**

- PySpark's streaming capabilities for processing data in real-time or simulating real-time analysis.
- Why?: Streaming analysis mimics the real-world flow of data, providing insights into how gun violence incidents unfold over time. This approach is valuable for understanding dynamic trends, assessing the immediate impact of policy changes, or detecting emerging patterns, thereby offering a timely basis for decision-making.

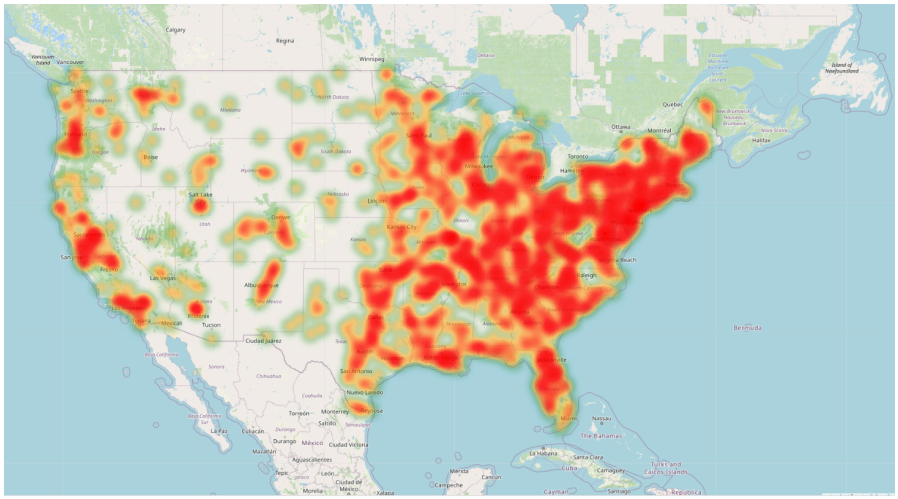
**7. Classification**

- We used a random forest classifier with a multi class classifier to categorize them into low high medium severity and we achieved an accuracy of 77.89%.

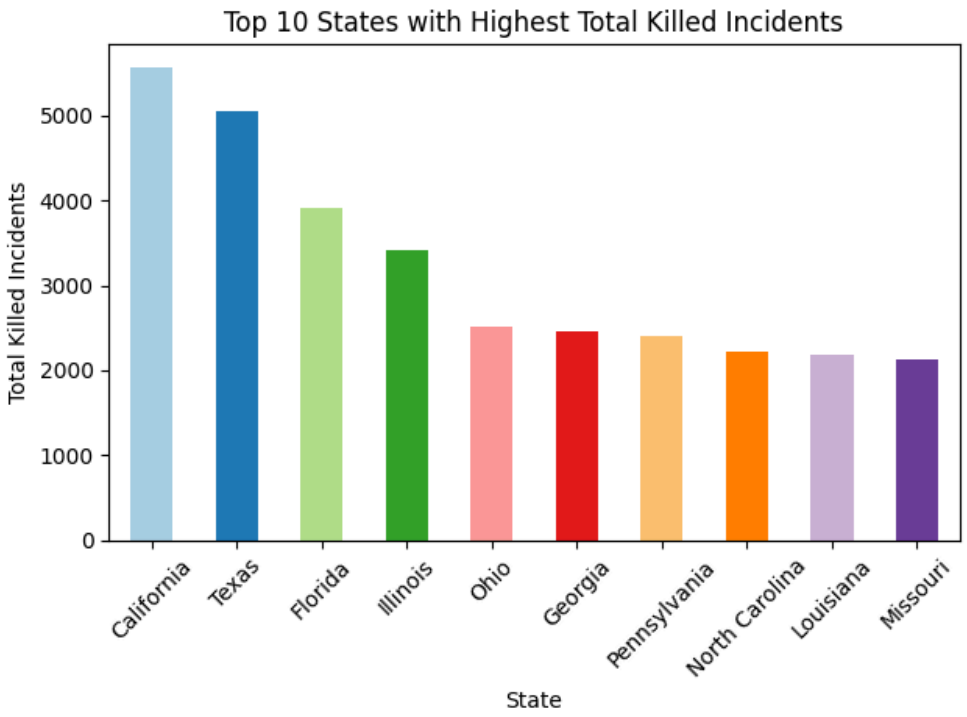
Each technique was chosen for its specific contribution to understanding the complex nature of gun violence across the United States. Together, they provide a comprehensive framework for analyzing the data, uncovering insights, and informing policy and interventions.

# RESULTS AND KEY FINDINGS

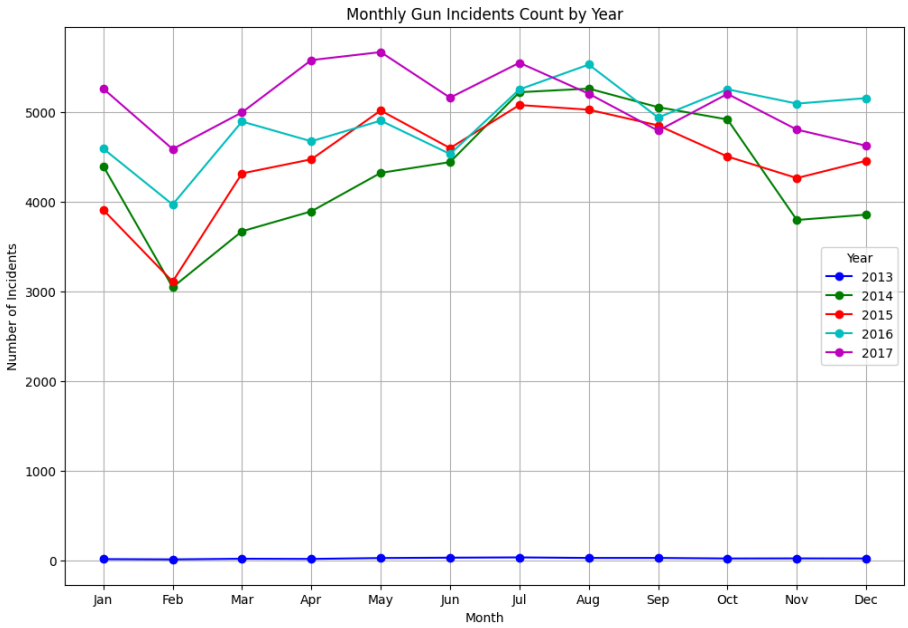
**Incident Severity and Trends:** Clustering identified specific regions as high-risk zones. Notably, certain states consistently exhibited higher violence levels, validating the approach for targeted intervention.




**Geographical Insights:** California emerged as a critical focus with more number of deaths across all years., with graph analysis underscoring its status as a hotspot for incidents yielding maximum fatalities. This aligns with historical data and underscores the need for specific policy attention.

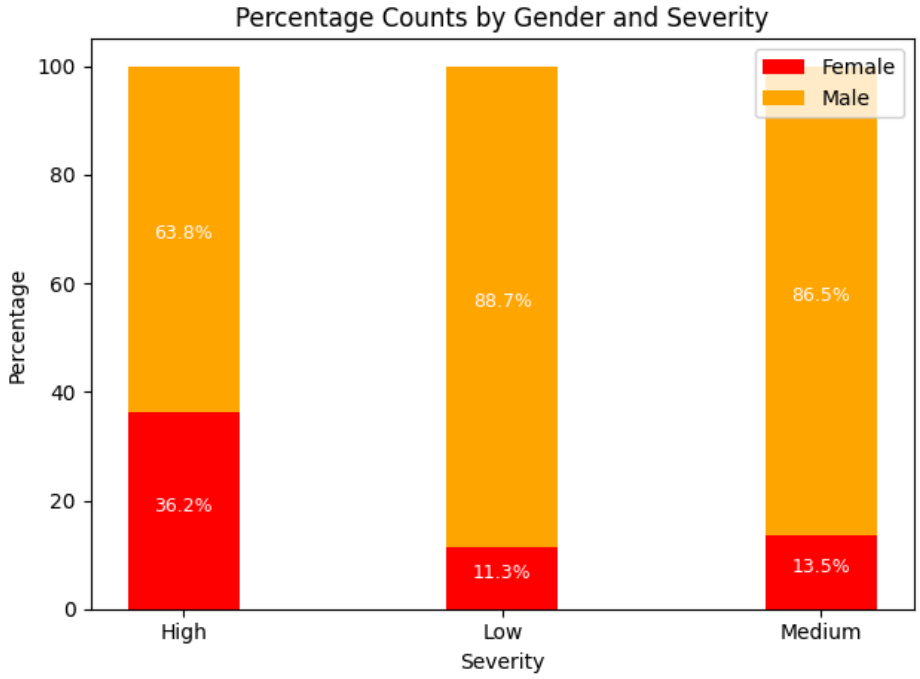



**Temporal Patterns:** Analysis highlighted evolving trends in gun incidents, including an increase in specific incident types over the years. This real-time analysis capability offers avenues for timely law enforcement and policy responses.




**Male Predominance in Incidents:** The data shows a higher percentage of male involvement across all severity levels of gun incidents. This predominance could suggest that males are more frequently involved

in situations leading to gun-related incidents, whether as victims or perpetrators. This finding aligns with broader criminological research indicating higher male involvement in violent crimes.




Accuracy for the random forest  Classification:-

```
[39] evaluator = MulticlassClassificationEvaluator(labelCol="newseverity", predictionCol="prediction", metricName="accuracy")
     accuracy = evaluator.evaluate(forestresults)
     print("Test set accuracy = " + str(accuracy))
```

```
Test set accuracy = 0.7789143041237113
```

# CONCLUSION

Our comprehensive analysis of gun incident data has provided valuable insights into the nature and distribution of gun violence across the United States. Through innovative data processing and analytical techniques, we have uncovered patterns that can inform effective strategies to reduce violence and its impact on communities. Our team's work demonstrates the power of PySpark tools in deriving meaningful insights from complex datasets, offering a roadmap for data-driven decision-making in public safety and policy development.