**Group 17**

**Nimal Kumarr Arunkumar (801317922)**

**Samhitha Mudiam (801329504)**

**Hrushi Pappuri (801060321)**

**Tanuj Darla (801316530)**

**Shravan Naidu Gollu (801318790)**


# Big Data Analytics for Competitive Advantage

## ITCS 6100

### Deliverable 3

**Team Members**

Nimal Kumarr Arunkumar (801317922)

Samhitha Mudiam (801329504)

Hrushi Pappuri (801060321)

Tanuj Darla (801316530)

Shravan Naidu Gollu (801318790)

**Repository and file storage**

GitHub repository :

https://github.com/darlatanuj10/Big_Data_project

Google Drive link :

https://drive.google.com/drive/folders/1oTZcGHwaM4rpUTlRjLQMYEjOGVE5BV_v?usp=share_link

**Communication plan to include project artifact repository**

This communication plan outlines how we will communicate about our project. This communication plan details how and where we will communicate about our project artifacts during our group project.

To distribute project artifacts, we will use the following communication channels:

- WhatsApp: For real-time messaging and file sharing.
- Google Drive is a service for storing and sharing project documents.
- For virtual meetings and screen-sharing sessions, use Zoom.

Repository of Artifacts: Google Drive will serve as our central store for all project artifacts, including:

- Other Documentation
- Code Files
- Datasets

At the following periods, we will communicate about project artifacts:

- Team meetings once a week: We will review project progress, any difficulties or risks, and changes to project artifacts.
- Discussion Hours: We will setup other doubt and meeting sessions when an additional help for a team member is needed.

**Domain**

This domain focuses on analyzing and building predictive models using a gun violence dataset containing information on incidents in the United States from January 2013 to March 2020. Preprocessing steps include data cleaning, feature engineering, data transformation, and feature selection. Various machine learning models can be applied, such as decision trees and random forests, gradient boosting machines, and neural networks, classification methods to achieve good performance metrics. Metrics such as precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic (ROC) curve should be used to measure model performance as they provide more informative results than accuracy alone for imbalanced datasets. By using Python 3.x and necessary libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn, analysts can gain insights into the patterns and factors contributing to gun violence incidents and build models to predict future incidents.

# Gun Violence Data

**Data :**
https://www.kaggle.com/datasets/jameslko/gun-violence-data
https://www.kaggle.com/datasets/konivat/us-gun-violence-archive-2014
https://www.gunviolencearchive.org

## Business Problem or Opportunity

One of the primary business problems that can be addressed with this dataset is to identify the factors associated with gun violence incidents. This includes understanding the demographics of the victims and perpetrators, the types of weapons used, the location and timing of incidents, and other key variables that may be relevant. By analyzing these factors, law enforcement agencies, policymakers, and community organizations can develop targeted interventions to prevent gun violence and improve public safety.

Another business opportunity that can be identified with this dataset is to explore the impact of gun control policies on gun violence. This includes analyzing the effectiveness of existing policies, identifying gaps in policy implementation, and exploring potential policy solutions to reduce gun violence. This information can inform policy development and implementation at the local, state, and national levels.

Moreover, the dataset can be used to identify high-risk areas and populations for gun violence and develop interventions that address the root causes of gun violence, such as poverty, unemployment, and mental health issues. By targeting these factors, stakeholders can reduce the incidence of gun violence and improve the health and safety of communities.

In summary, the gun violence dataset presents a significant business opportunity to better understand the complex issue of gun violence, develop evidence-based interventions, and inform policy decisions. By analyzing this data, stakeholders can work together to reduce the harm caused by gun violence and promote public safety.

## Research Objectives

This project's aim is to examine a comprehensive dataset of gun violence occurrences that took place in the US between January 2013 and March 2020. Finding major patterns and trends that can be used to pinpoint the key causes of these accidents is the goal. A further objective of the project is to develop a machine learning model that can precisely forecast the probability of gun violence occurrences in various situations and environments. Utilizing the proper evaluation metrics, the generated model's performance will be assessed. Finally, the initiative aims to offer data-driven suggestions to legislators, law enforcement groups, and community organizations to implement focused interventions and policies to lower gun violence events.

**Questions**

1) What are the most typical patterns and trends in gun violence instances across the United States from 2013 to 2020?

2) Which aspects (such as place, time, socioeconomic variables, etc.) have the biggest bearing on the frequency of gun violence incidents?

3) Can the discovered factors be used to predict gun violence events using machine learning models like logistic regression, decision trees, random forests, gradient boosting machines, and neural networks?

4) Can gun violence occurrences be accurately predicted using the given parameters using machine learning models such as logistic regression, decision trees, random forests, gradient boosting machines, and neural networks?

5) How does the predictive performance of the various machine learning models compare when evaluated using metrics like precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic (ROC) curve?

6) What practical lessons can be drawn from the data to help policymakers, law enforcement groups, and community organizations create focused strategies for stopping and intervening in gun violence incidents?

7) How did the economic crisis change in the gun violence statistics in the United States ?

**Data understanding**

We have imported a number of Python libraries and modules that are commonly used for data science and machine learning tasks.

- os is a built-in Python module that provides functions for interacting with the operating system.
- numpy is a library for scientific computing with Python. It provides a high-performance multidimensional array object and tools for working with arrays.

- pandas is a library for data analysis and manipulation in Python. It provides high-performance, easy-to-use data structures and data analysis tools.
- matplotlib is a library for creating static, animated, and interactive visualizations in Python.
- seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.
- plotly is a library for creating interactive, web-based visualizations in Python.
- %matplotlib inline is a magic command that tells matplotlib to display plots inline in the Jupyter notebook.

The following modules are imported from the sklearn library, which provides a large collection of machine learning algorithms and tools:

- train_test_split is a function for splitting a dataset into training and testing sets.
- LinearRegression is a linear regression model.
- metrics is a module for evaluating machine learning models.
- mean_squared_error is a function for calculating the mean squared error of a model.
- cross_val_score is a function for evaluating a model using cross-validation.
- Lasso is a linear regression model with L1 regularization.
- Ridge is a linear regression model with L2 regularization.
- DecisionTreeRegressor is a decision tree regressor.
- RandomizedSearchCV is a function for performing a randomized search over a hyperparameter space.

---

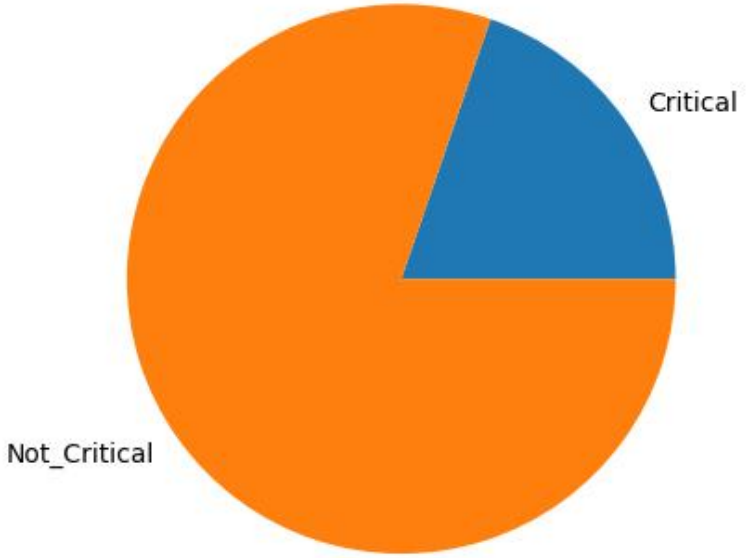
- RandomForestRegressor is a random forest regressor.
- GridSearchCV is a function for performing a grid search over a hyperparameter space.
- DummyRegressor is a dummy regressor.
- StandardScaler is a function for standardizing features.
- Pipeline is a class for creating pipelines of machine learning models.

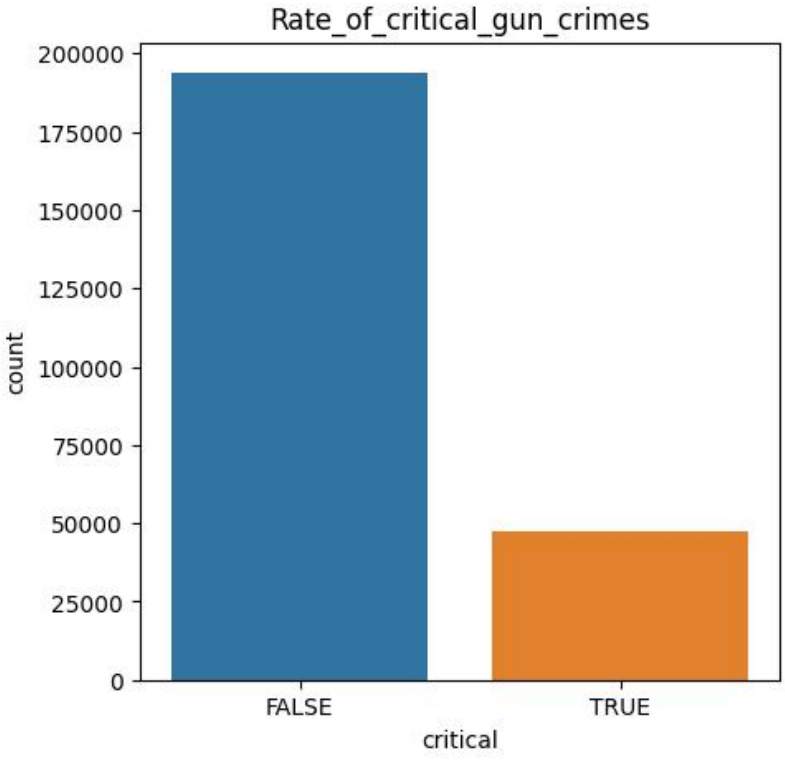
- scale is a function for scaling features.
- MinMaxScaler is a function for min-max scaling features.
- normalize is a function for normalizing features.

This dataset aims to represent a record of more than 260k gun violence incidents, with detailed information about each incident, available in CSV format. We hope this data will make it easier for data scientists and statisticians to study gun violence and make informed predictions about future trends. The CSV file contains data for all recorded gun violence incidents in the US between January 2013 and December 2021, inclusive. Next up, we take a look at our columns in the dataframe and perform various analysis based on the time of the incident. For that reason we convert our date feature into a datetime object. We checkup for the various columns present in the dataset and look up for missing values in the data. We then drop off the columns consisting of null values since this are not needed and will help in making our analysis crisp and accurate. We then change up the column names for easy understanding of our data. For our next step, we come up with some intermediate analysis and visualization of our dataset for better understanding.

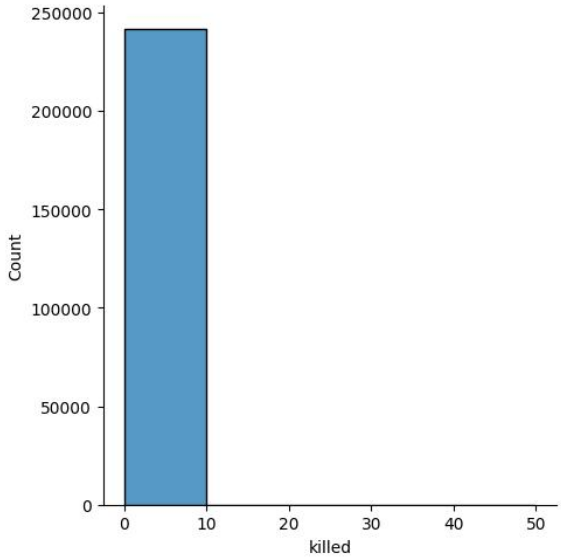
From our dataset, we check for the toatal number of cases which are critical and the number of cases which are not critical.

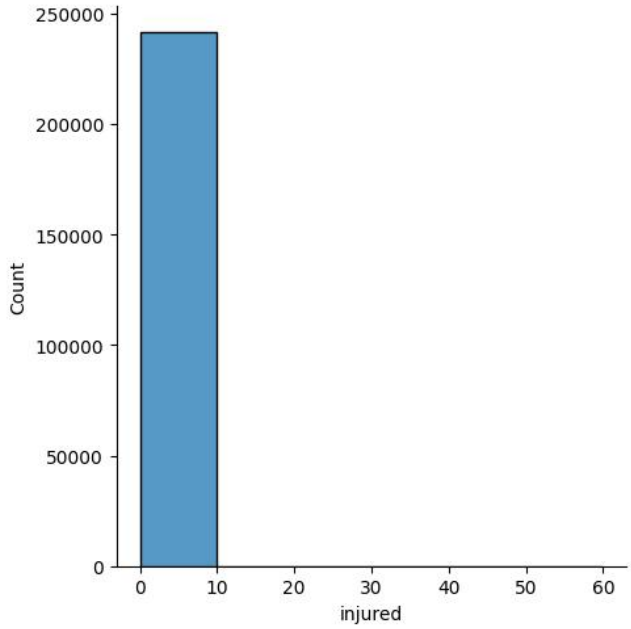
Number of Critical cases : 47607
Number of Non-critical cases : 193790


Critical
Not_Critical


Rate_of_critical_gun_crimes
200000
175000
150000
125000
100000
75000
50000
25000
0
count
FALSE
TRUE
critical

Then we come up with graphs to display the total killed and total injured counts.




**Data preparation**

The goal of this machine learning algorithm implementation is to create a supervised machine learning algorithm model using linear regression that can predict the number of people killed/injured as a result of gun violence based on a number of feature variables.

- Drop columns that are irrelevant to the project .
- Remove columns/rows with excessive amount of missing data .
- Choose age-group column over age column for better prediction .
- Create pseudo-dummy columns that counts the number of genders/age groups in a single row .
- Include only the top 15 cities with most incidents for better prediction .
- Add new date columns (year, month, weekday) .
- Remove incident characteristics due to data leakage .
- Create new numerical columns for Categorical columns (ex.mapped_cities) .
- Nearly 200,000 rows of data were dropped as a result .

- Using the Chi-Square Test for Independence and previously creating numerical columns for the Categorical data, features were selected for the machine learning models.
- Two types of feature sets were created: one for predicting the number of people killed (killed) and the other for predicting the number of people injured (injured).
- Both feature sets included the same features shown on the right, except the set for predicting the number of people killed included n_injured as a feature and vice versa.

- From performing inferential statistics and checking correlation between variables, we know which features we want to select to predict the number of people killed (n_killed) and the number of people injured (n_injured).

- In the inferential statistics notebook, we've also turned the age-group and participant-gender columns into its own separate categorical columns. For age-group, we had 3 columns for child (people aged 0-11), teen (people aged 12-17) and adult (people aged 18+) and each one counted the number of times a person of that age group was part of the incident. For participant-gender, we had 3 columns for male, female and unknown (cases where the gender was unknown).

- We also used the pandas function pd.Categorical() to transform the column 'city_or_county' into a Categorical column where each value is represented by numbers. Ex. Chicago is no longer a string called 'Chicago' and is instead represented by the number 1.

- Through these steps, we've ensured that the feature variables will be numerical while retaining its original information.

**Data analysis**

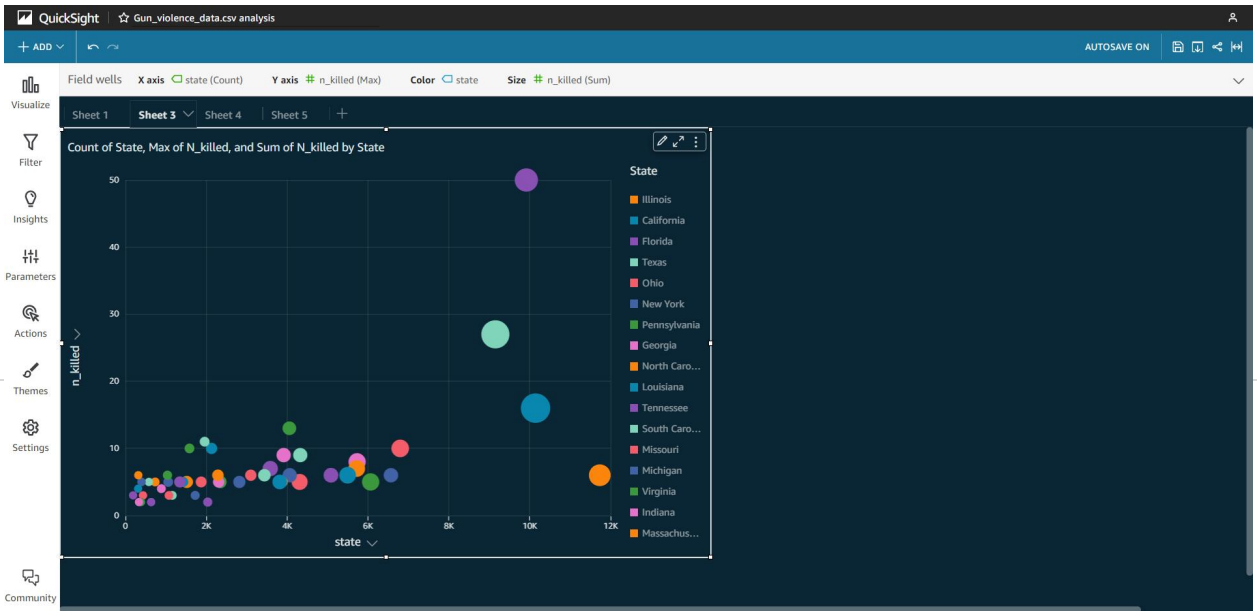
We have created a dashboard with the most of our analysis based on the day, month, the critical rate of the crime, the killed and injured rate and so on using an onlone visulaization tool Amazon Quicksight for our purpose.

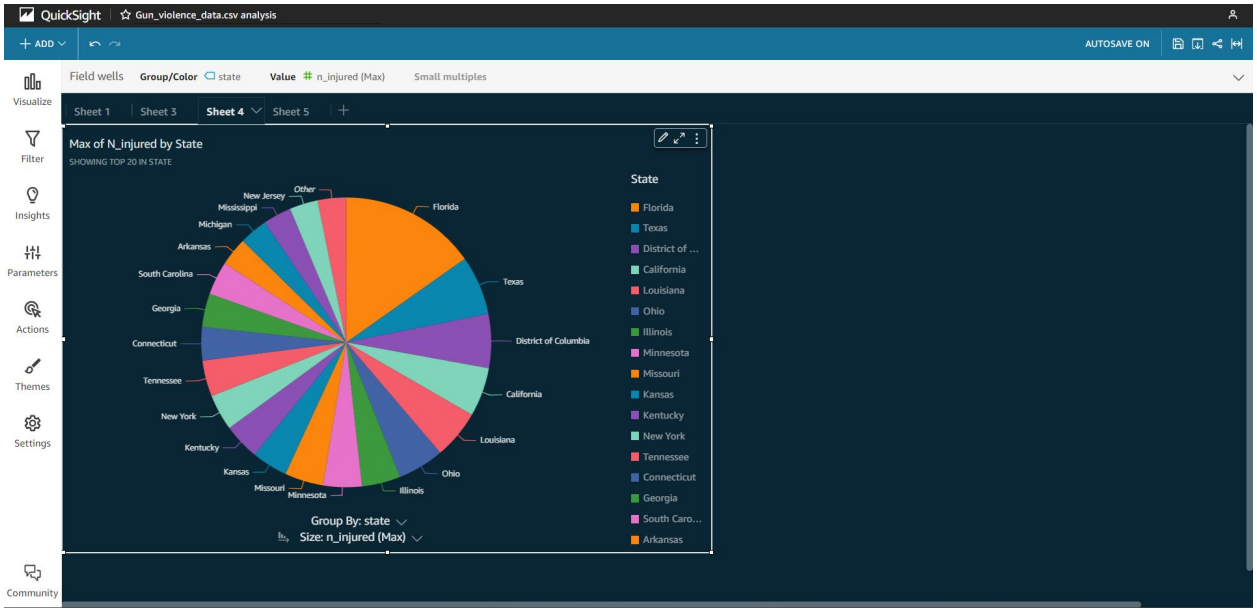
Amazon Quicksight is a business analytics service provided by AWS.

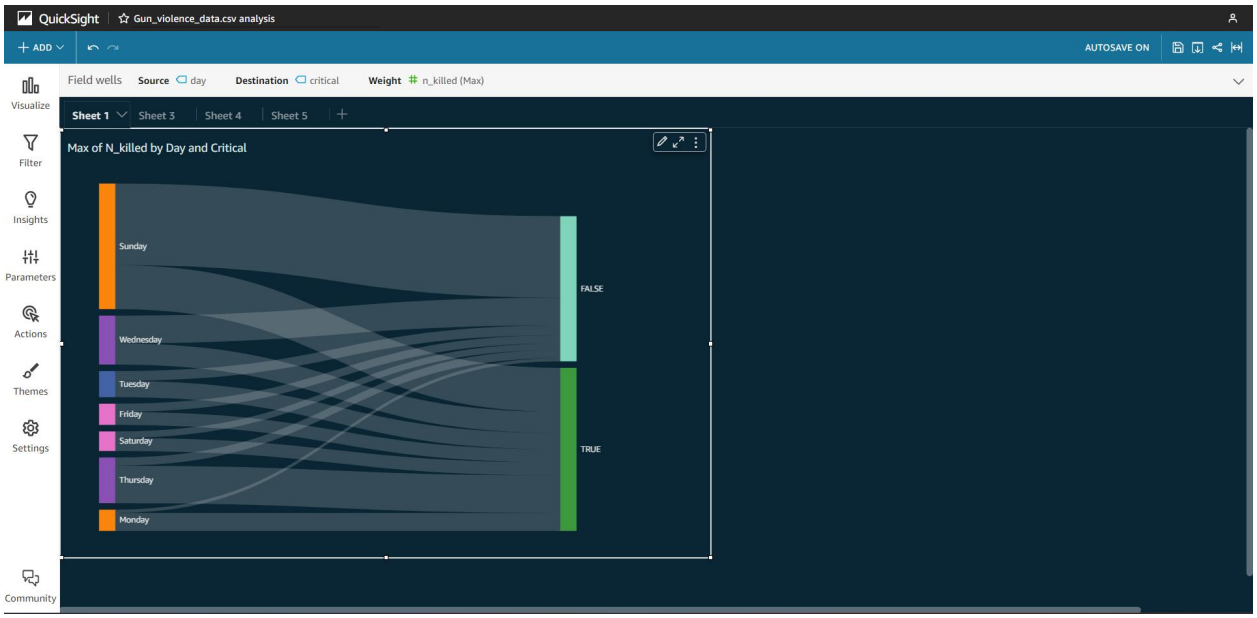
Quicksight provides easy to use tools to build visualizations, perform ad-hoc analysis, get business insights from the data and share the results with others.

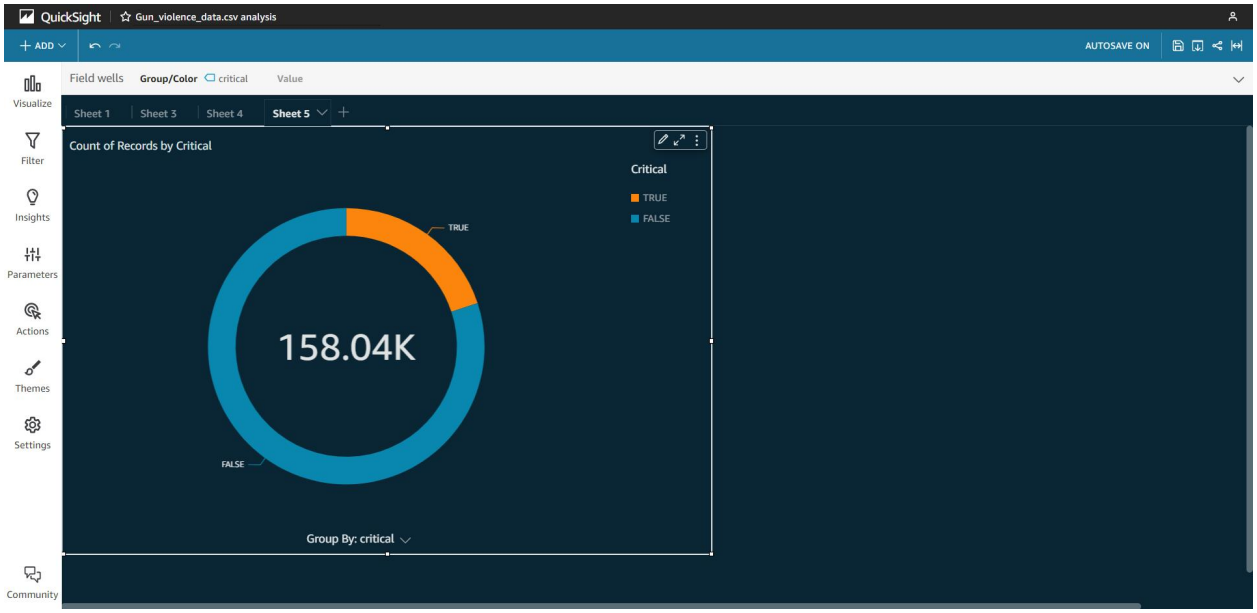
Quicksight supports the following:

- A wide range of data sources ( Amazon Athena, Apache Spark, etc).
- Different ways to filter data
- Different visualization methods and charts (bar charts, Pie charts, scatter maps, etc)
- Pattern and anomaly detection
- Interactive dashboards

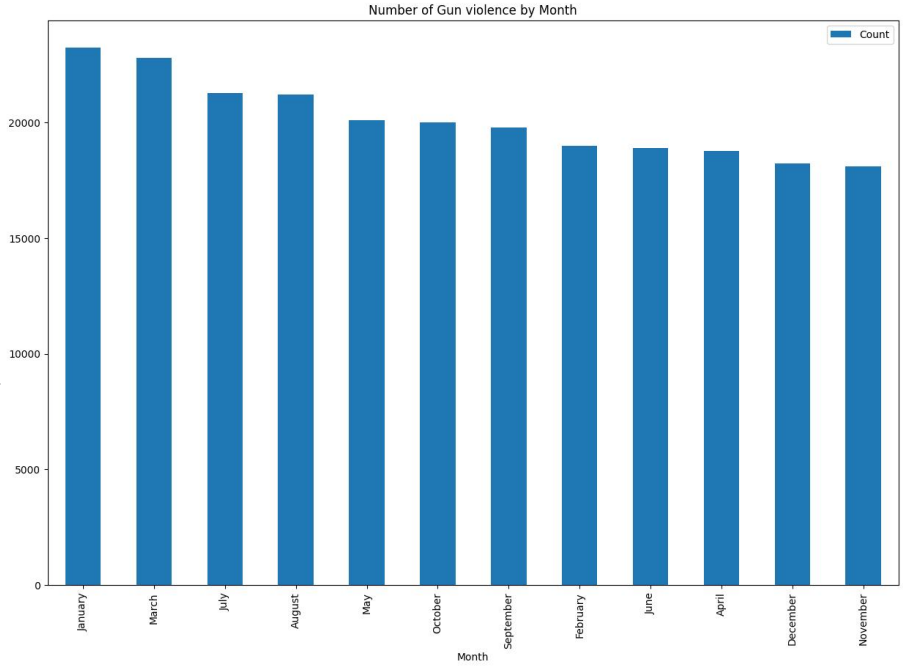



QuickSight
Gun_violence_data.csv analysis
ADD
AUTOSAVE ON
Field wells
Group/Color state
Value n_injured (Max)
Small multiples
Visualize
Filter
Insights
Parameters
Actions
Themes
Settings
Community
Sheet 1
Sheet 3
Sheet 4
Sheet 5
Max of N_injured by State
SHOWING TOP 20 IN STATE
Florida
Texas
District of Columbia
California
Louisiana
Ohio
Illinois
Minnesota
Missouri
Kansas
Kentucky
New York
Tennessee
Connecticut
Georgia
South Carolina
Arkansas
Michigan
Mississippi
New Jersey
Other
State
Florida
Texas
District of ...
California
Louisiana
Ohio
Illinois
Minnesota
Missouri
Kansas
Kentucky
New York
Tennessee
Connecticut
Georgia
South Caro...
Arkansas
Group By: state
Size: n_injured (Max)

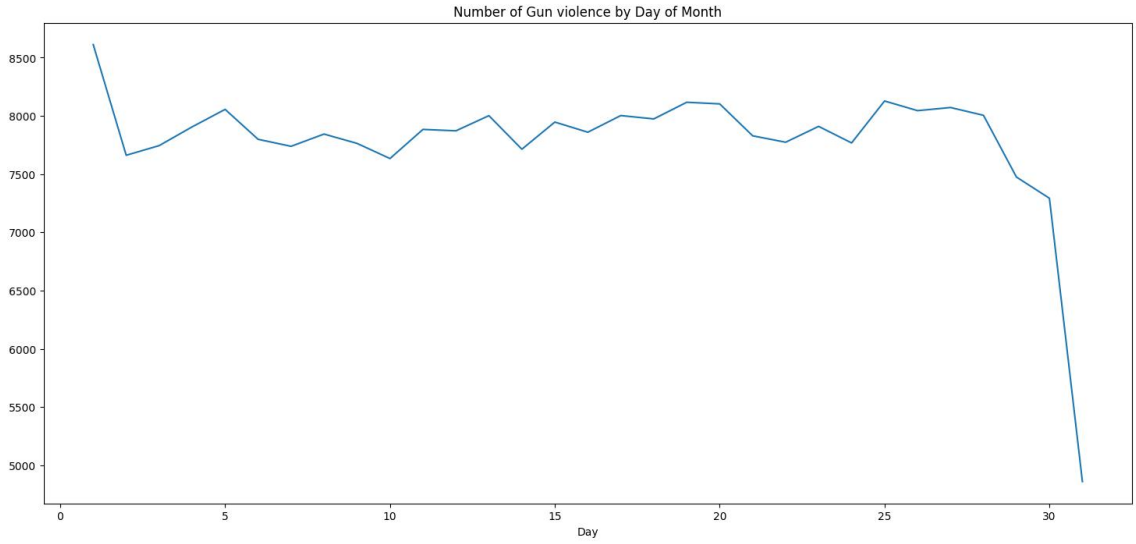

QuickSight
Gun_violence_data.csv analysis
ADD
AUTOSAVE ON
Field wells
Source day
Destination critical
Weight n_killed (Max)
Visualize
Filter
Insights
Parameters
Actions
Themes
Settings
Community
Sheet 1
Sheet 3
Sheet 4
Sheet 5
Max of N_killed by Day and Critical
Sunday
Wednesday
Tuesday
Friday
Saturday
Thursday
Monday
FALSE
TRUE



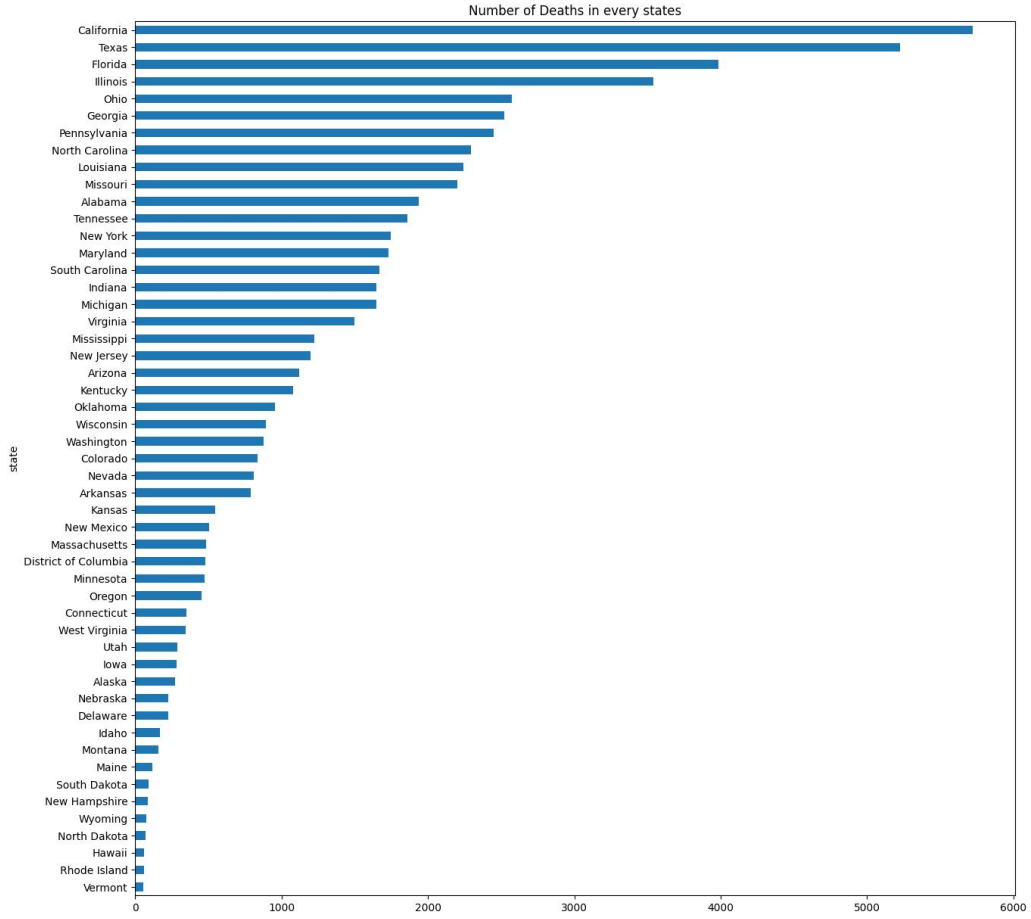

Then for understanding each column and time line for our gun violence determination we present various visulaizations for the dataset.



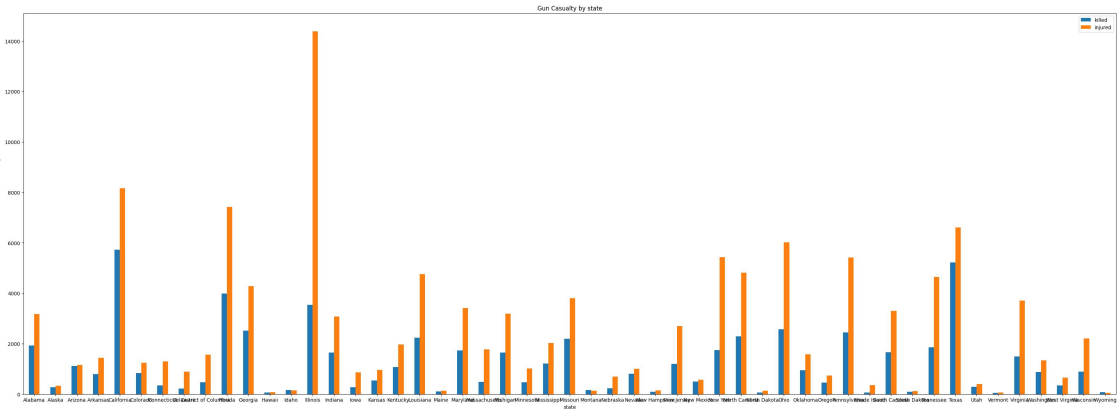
Here we tried to visualize which month was more likely to see gun incidents and to no surprise January, March and July are high on the list due to obvious two big celebrations in New Year and 4th of July.




Based on this figure we can see, the incidents are really common during the start of the month as people are more likely to go out partying, drinking at the start of the month (Fresh Paychecks) compared to the end of the month with the incident almost crashing down.

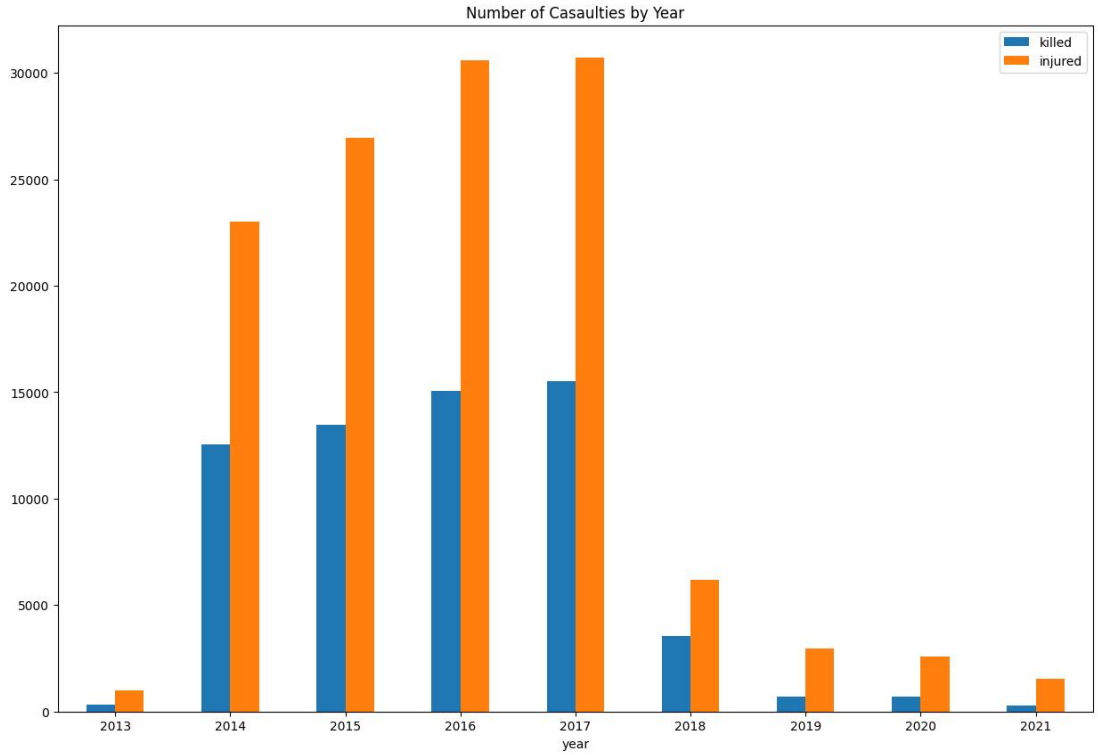



This graph however shows an interesting information as California comes across as the state with the most number of Gun related deaths closely followed by Texas and Florida. Kudos to Vermont, Rhode Island and Hawaii.

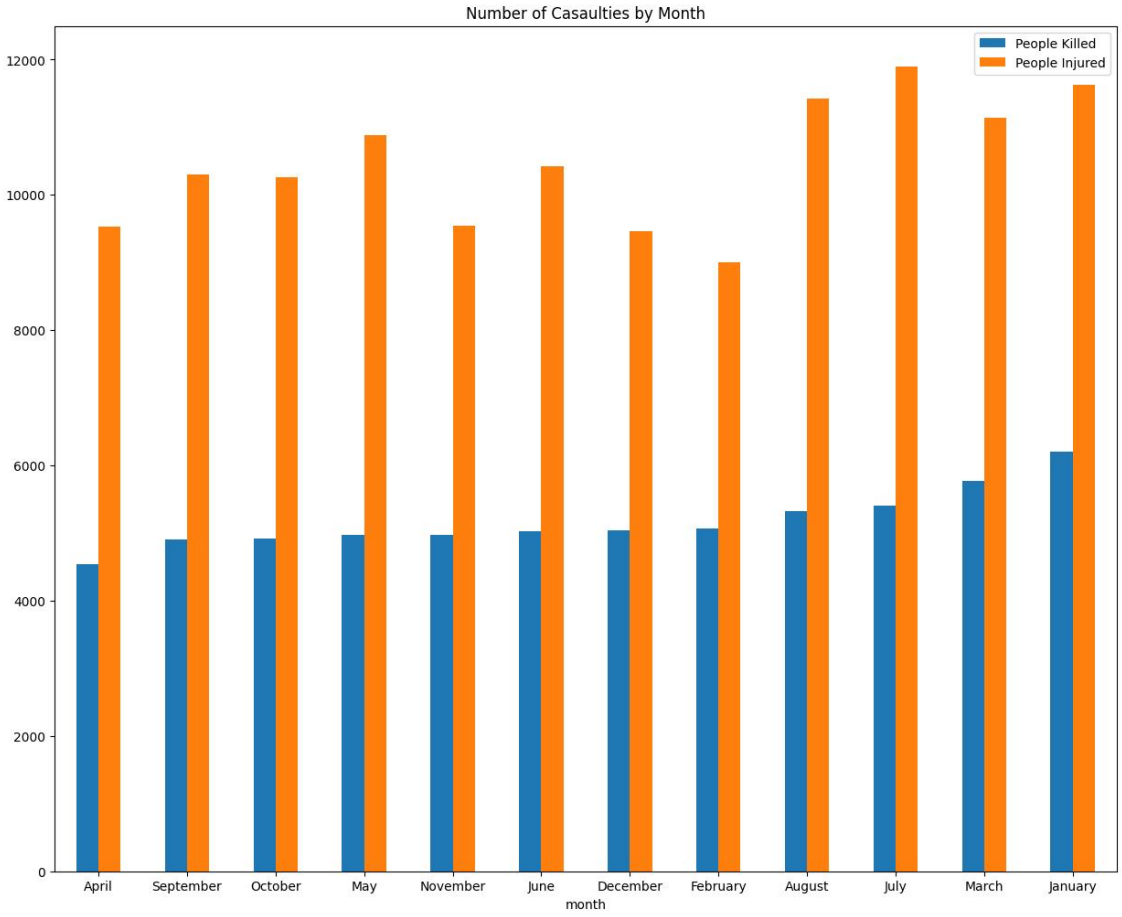



Here we plot the deaths along with the incidents that didn't result in death and it gives us a completely different picture. State of Illionois which had 4th highest number of

Gun deaths, surpasses every single state massively in terms of incidents which didn't result in death.

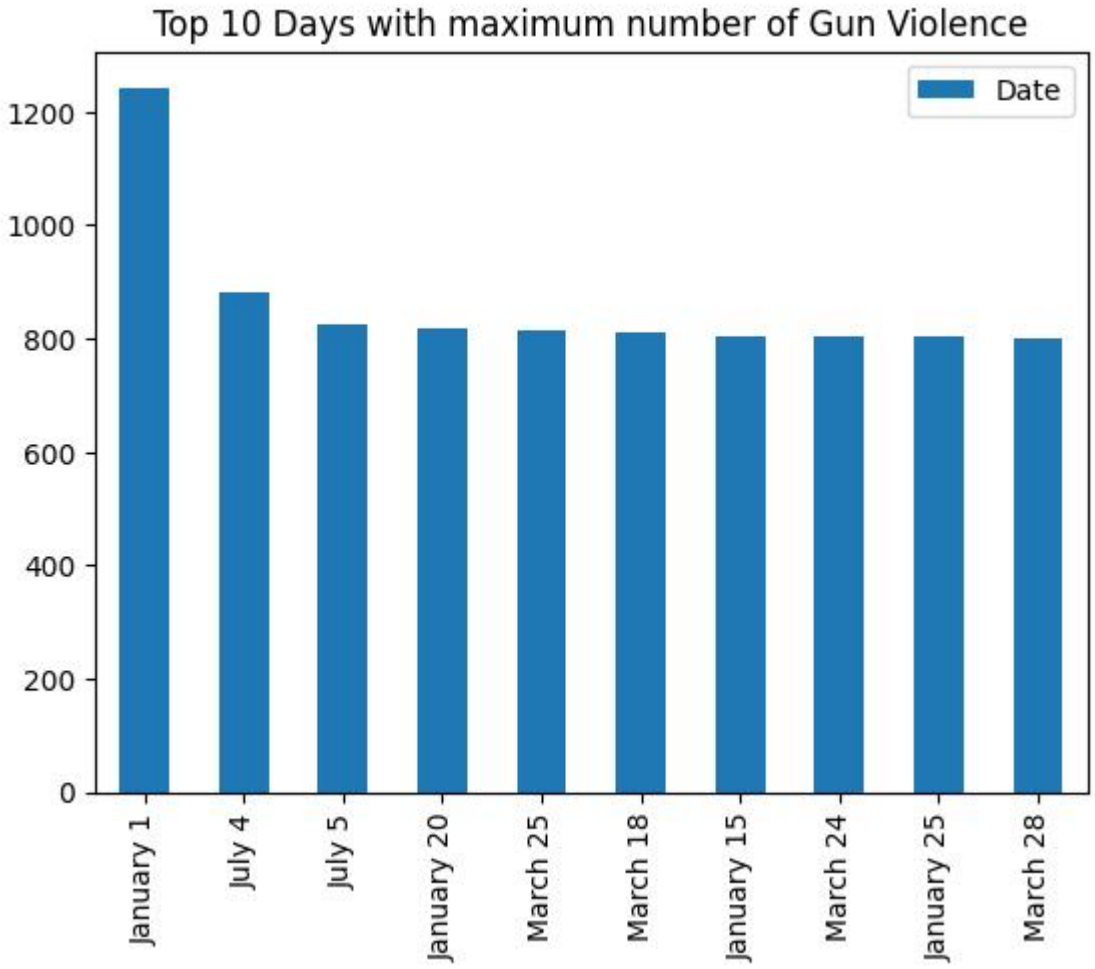



From the graph we have projected the number of casulaties by year and can see the highest casulaities are in 2016 and 2017. We can see that the gun violence rate has gradually decreased after financial crisis in 2018 and covid disease in 2020. This both factors have impacted a lot on the gun violence rates in the USA.

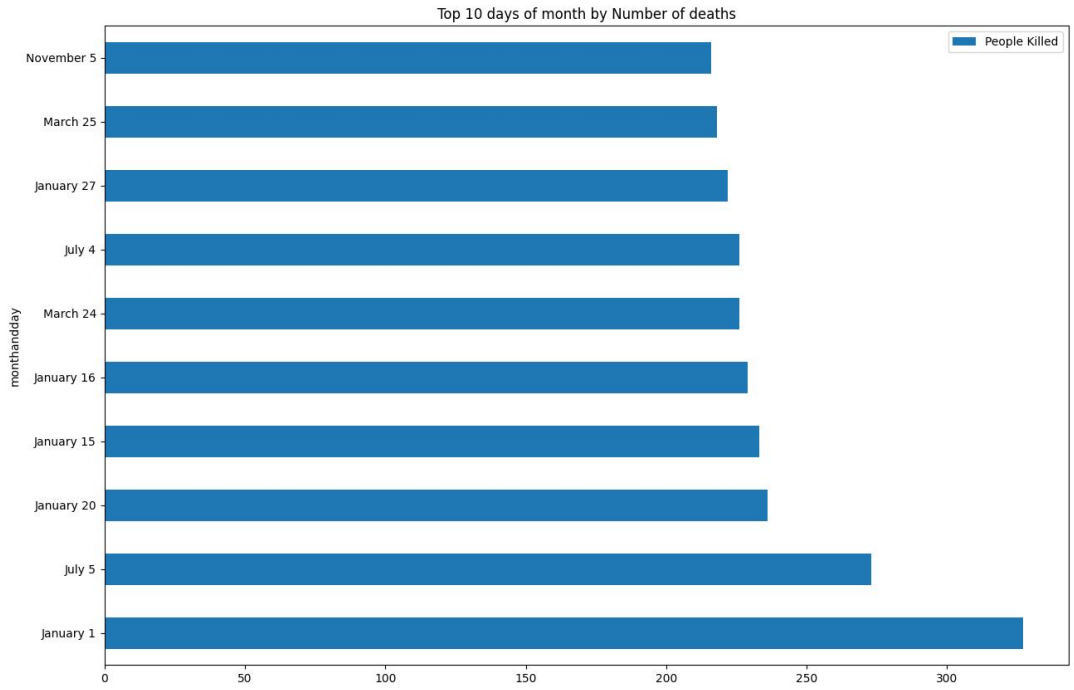

Number of Casaulties by Month
People Killed
People Injured
12000
10000
8000
6000
4000
2000
0
April
September
October
May
November
June
December
February
August
July
March
January
month


Number of deaths by day of the month
day
5
1
23
26
19
25
11
20
28
10
12
27
7
4
14
17
18
9
15
8
24
16
6
22
13
3
2
30
21
29
31
0
500
1000
1500
2000
People Killed

This is up for interpretation as you can see more deaths occur on the 5th of a month but the constant thing has been that the end of the month are fairly quiet in terms of gun violence.




It comes as no surprise that two of the most severe days are 1st of January and 4th of July. The reason could be a large number of public events and people being intoxicated and prone to being violent early.




## Machine Learning

- Let's begin by using the most standard regression classifier: Linear Regression. It's very simple since it does not require any hyperparameters.
- First we will create a baseline using the DummyRegressor function and set the strategy to 'mean'. This will give us a way to check how accurate the predictions are.
- Our baseline r-squared for killed is -0.0001 and our baseline RMSE is 0.5067.
- Our baseline r-squared for injured is -0.00002 and our baseline RMSE is 0.7654.
- Although the r-squared seems bad, it's important to remember it only serves as a baseline against which we can compare the results of our models.
- By comparing the results of the linear regression to the n_killed baseline, we can see that the r-squared of 0.3193 and RMSE of 0.4123 are better than the baseline.
- Similarly for n_injured, the r-squared of 0.4207 and RMSE of 0.6005 are better than the baseline.

- Let's plot the predicted versus the actual response values so that we can see how well the linear regression model performed from a visual standpoint.

- There are a few takeaways from the predicted vs actual plots. First, the model predicted quite a few negative values for n_killed and n_injured, which is practically impossible. Secondly, the model did not predict any values greater than 2 for n_killed nor did it predict any values greater than 10 for n_injured. It also used floating point numbers instead of only integer values. This shows the limitations of the Linear Regression model.

- The model also set most of the coefficients close to 0. In fact, the only variables that had a coefficient absolute value greater than 0.2 were n_injured when n_killed was the response variable and n_killed when n_injured was the response variable. In both cases, they had negative correlations which implies they have an inverse relationship. On average, if we assume that there is a fixed number of casualties per incident, it makes sense that with more injuries, there will be fewer deaths and vice versa.

- Next, let's use the Lasso Regularization and Ridge Regularization on the data to help compensate for the possibility of overfitting. For these classifiers, we will be specifying the hyperparameter alpha, which multiples the L1 term. In order to decide which alpha to use, we will be using GridSearchCV to iterate over a list of values for alpha. This will return the best alpha value and its "score".

- For killed, the R-squared score for Lasso (0.3123 vs 0.3193) became worse compared to the Linear Regression model without any regularization. However, the RMSE (0.4112 vs 0.4123) did slightly improve. Overall, it's still better than the baseline.

- For injured, the R-squared score slightly improved for Lasso (0.4279 vs 0.4207) as well as the RMSE score (0.5809 vs 0.6005).

- We will next use the Ridge regularization to see if there are any improvements on the model.

- For killed, all 3 regression models (Linear without regularization, Linear with Lasso, Linear with Ridge) had relatively similar RMSE scores but their R-Squared scores were noticeably different. Linear without any regularization had the best r-squared score at 0.3193 and the best RMSE at 0.4123.

- When looking at the predicted vs actual plots, all 3 models looked similar in their distribution of values. They all had a majority of their values fit between -2 and 2. The feature coefficients plots of all 3 methods were practically the same with only minute differences in the coefficient values.

- For injured, both regularization methods improved the Linear Regression model. Ridge regularization had the best R-Squared score (0.4437) while Lasso regularization had the best RMSE (0.5809).
- The feature coefficients plots of the Linear Regression without regularization and Linear Regression with Lasso regularization were very similar. However, the plot of Ridge regularization was different and almost identical to the Ridge regularization plot for killed.

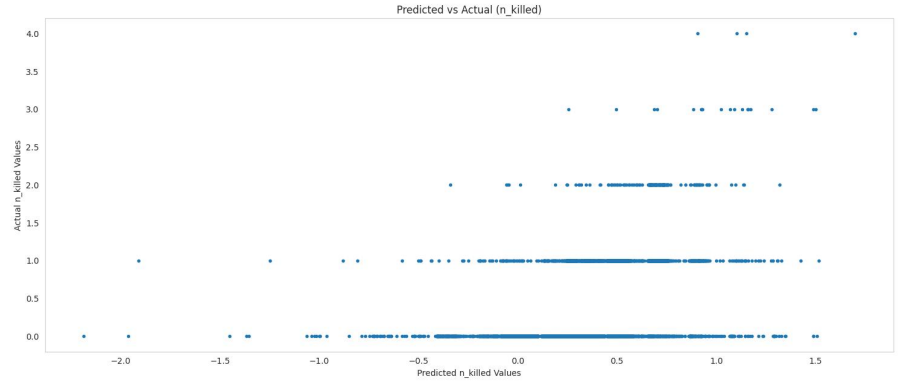
| Model for n_killed | R-Squared | RMSE |
|---|---|---|
| Linear Regression | 0.3193 | 0.4123 |
| Linear Regression w/ Lasso | 0.3123 | 0.4112 |
| Linear Regression w/Ridge | 0.2973 | 0.4144 |
| Base-line | -0.0001 | 0.5067 |

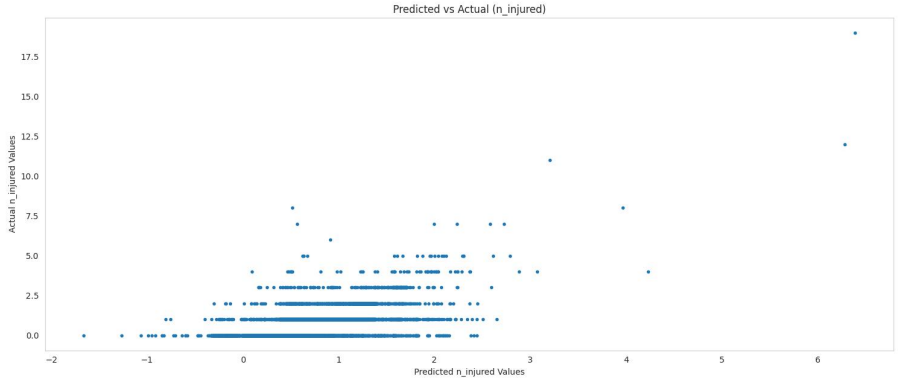
| Model for n_injured | R-Squared | RMSE |
|---|---|---|
| Linear Regression w/Ridge | 0.4437 | 0.5993 |
| Linear Regression w/ Lasso | 0.4279 | 0.5809 |
| Linear Regression | 0.4207 | 0.6005 |
| Base-line | 0.0000 | 0.7654 |

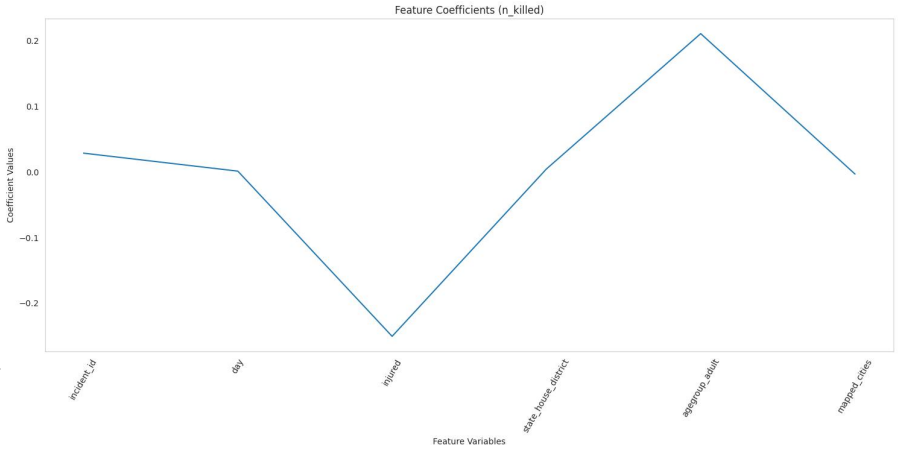
**Linear regression**

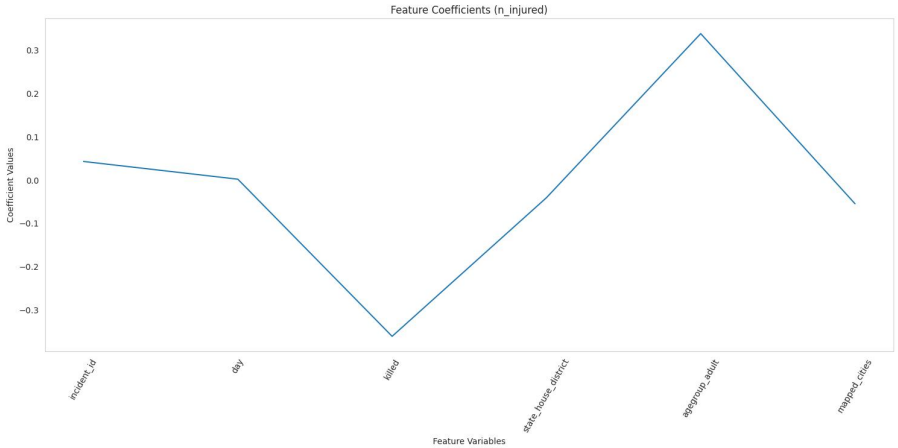
Baseline R-squared (n_killed): -0.0001
Baseline Root Mean Squared Error (n_killed): 0.5067
Baseline Average 10-Fold CV Score (n_killed): -0.0053 Baseline R-squared (n_injured): -0.00002
Baseline Root Mean Squared Error (n_injured): 0.7654
Baseline Average 10-Fold CV Score (n_injured): -0.0126

R-squared (n_killed): 0.2900
Root Mean Squared Error (n_killed): 0.4211
Average 10-Fold CV Score (n_killed): 0.2663
R-squared (n_injured): 0.3450
Root Mean Squared Error (n_injured): 0.6386
Average 10-Fold CV Score (n_injured): 0.3196


Predicted vs Actual (n_killed)
Actual n_killed Values
Predicted n_killed Values


Predicted vs Actual (n_injured)
Actual n_injured Values
Predicted n_injured Values

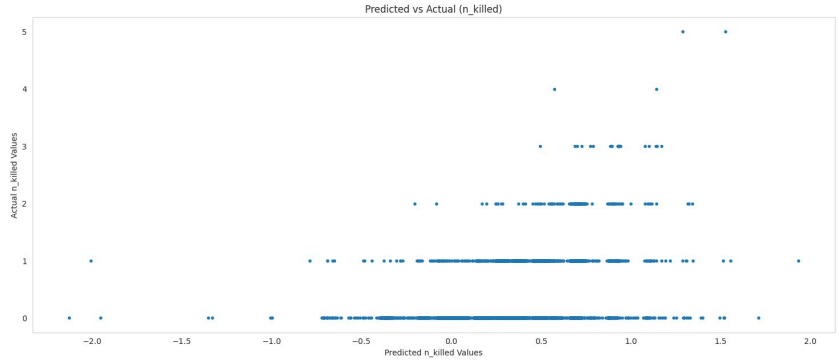

Feature Coefficients (n_killed)
Coefficient Values
incident_id
day
injured
state_house_district
agegroup_adult
mapped_cities
Feature Variables

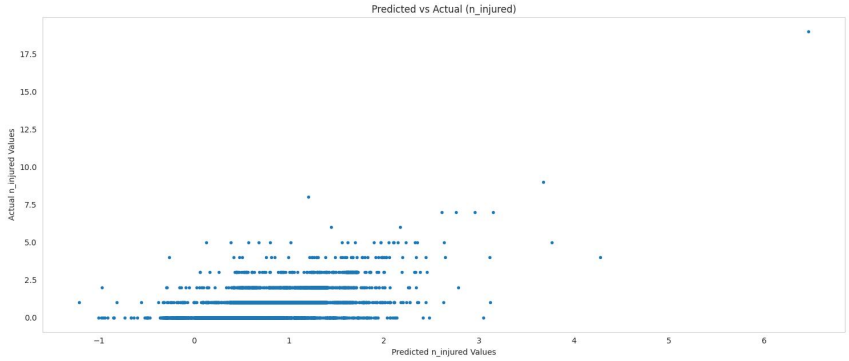

Feature Coefficients (n_injured)
Coefficient Values
incident_id
day
killed
state_house_district
agegroup_adult
mapped_cities
Feature Variables

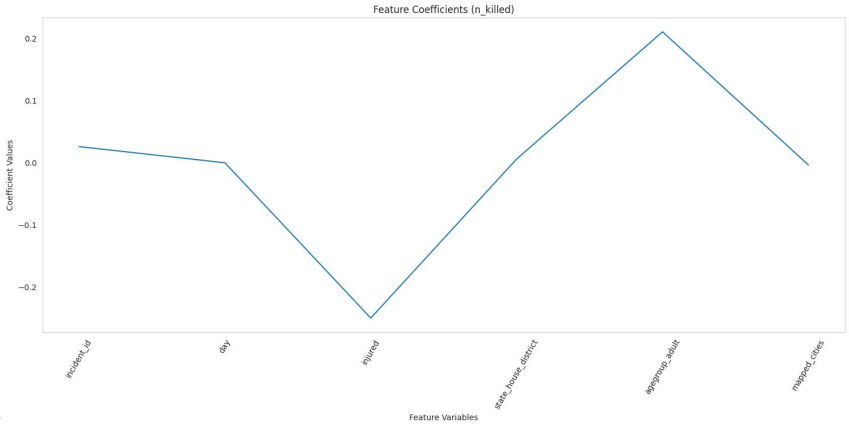
**Linear Regression w/ Lasso**

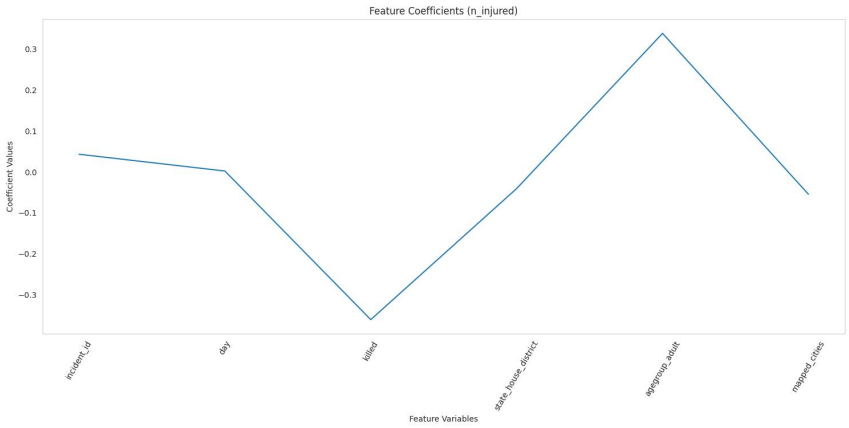
The best score is (n_killed):0.2860
The best parameters are (n_killed): {'lasso__alpha': 0.001}
The best score is (n_injured):0.3201
The best parameters are (n_injured): {'lasso__alpha': 0.001}

R-squared (n_killed): 0.2883
Root Mean Squared Error (n_killed): 0.4183
Average 10-Fold CV Score (n_killed): 0.2665
R-squared (n_injured): 0.3401
Root Mean Squared Error (n_injured): 0.6239
Average 10-Fold CV Score (n_injured): 0.3196


Predicted vs Actual (n_killed)
Actual n_killed Values
Predicted n_killed Values


Predicted vs Actual (n_injured)
Actual n_injured Values
Predicted n_injured Values

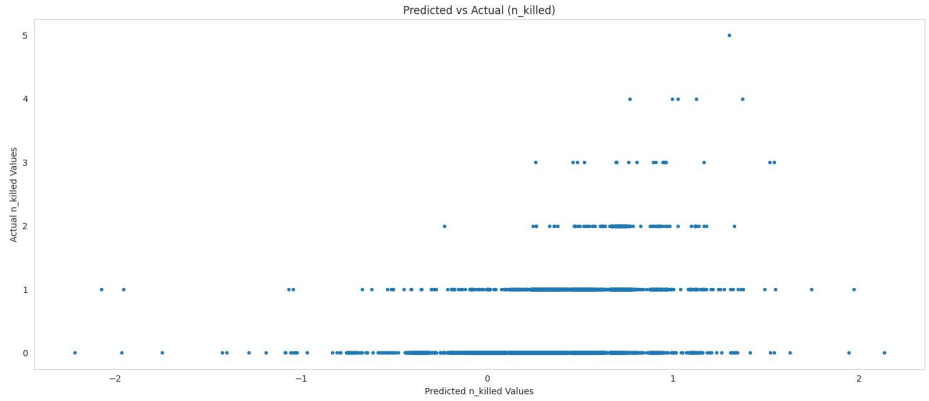

Feature Coefficients (n_killed)
Coefficient Values
incident_id
day
injured
state_house_district
agegroup_adult
mapped_cities
Feature Variables

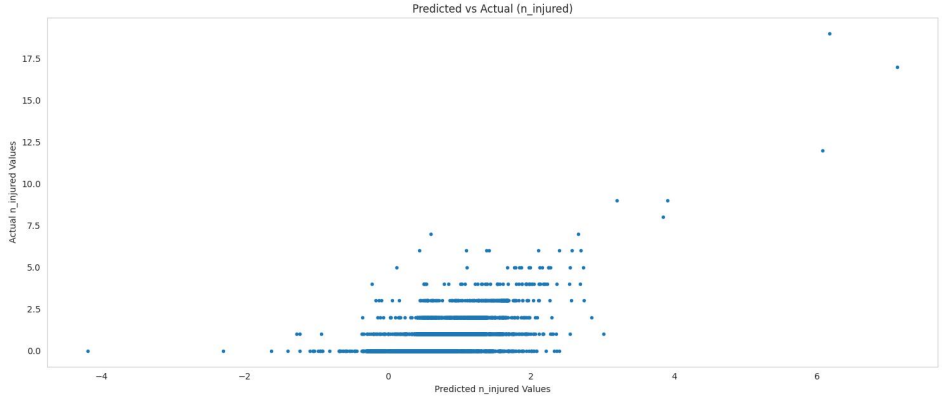

Feature Coefficients (n_injured)
Coefficient Values
incident_id
day
killed
state_house_district
agegroup_adult
mapped_cities
Feature Variables

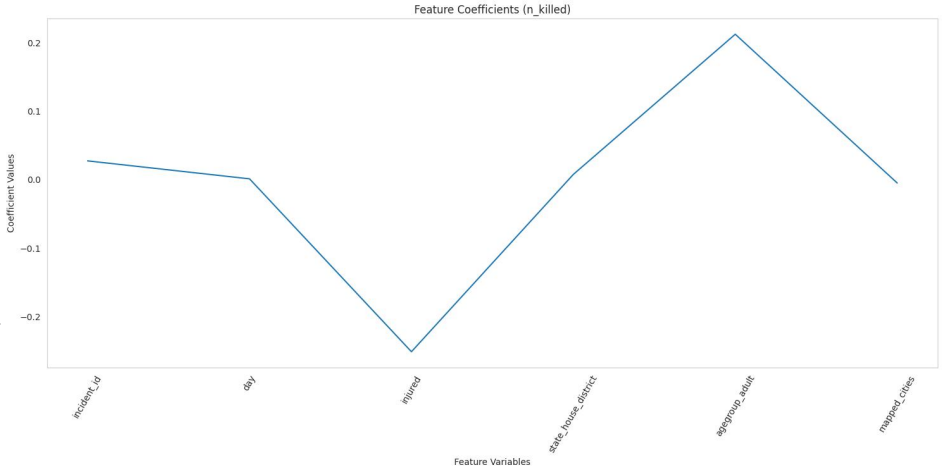
**Linear regression w/ Ridge**

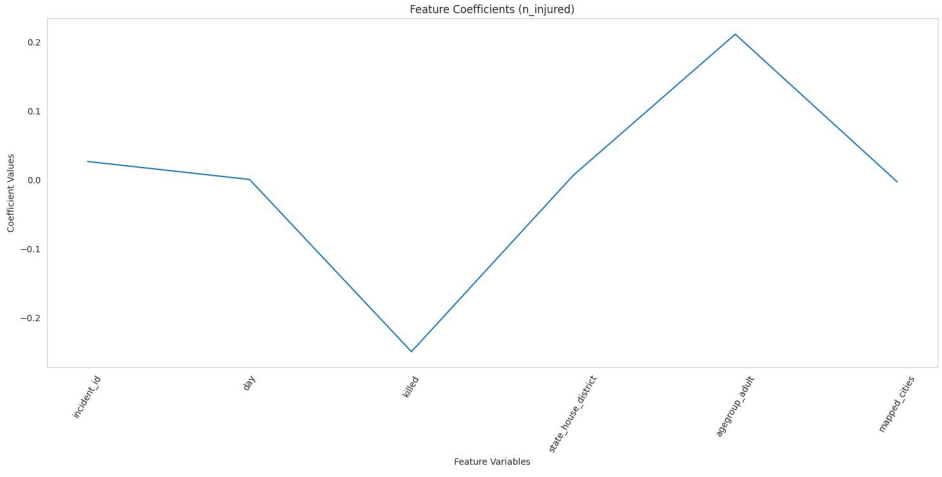
The best score is (n_killed):0.2922
The best parameters are (n_killed): {'ridge__alpha': 10}
The best score is (n_injured):0.3078
The best parameters are (n_injured): {'ridge__alpha': 0.001}

squared (n_killed): 0.2734
Root Mean Squared Error (n_killed): 0.4214
Average 10-Fold CV Score (n_killed): 0.2664
squared (n_injured): 0.3594
Root Mean Squared Error (n_injured): 0.6431
Average 10-Fold CV Score (n_injured): 0.3196


Predicted vs Actual (n_killed)
Actual n_killed Values
Predicted n_killed Values


Predicted vs Actual (n_injured)
Actual n_injured Values
Predicted n_injured Values


Feature Coefficients (n_killed)
Coefficient Values
incident_id
day
injured
state_house_district
agegroup_adult
mapped_cities
Feature Variables


Feature Coefficients (n_injured)
Coefficient Values
incident_id
day
killed
state_house_district
agegroup_adult
mapped_cities
Feature Variables

**Conclusion**

1. What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

**Ans** : The data is highly imbalanced, with a large number of null examples (no gun violence) and a small number of positive examples (gun violence). This made it difficult to train a machine learning model. The data had many unwanted rows which can skew the results of machine learning models.

2. Did you create any new additional features / variables?

**Ans** : We have created new data rows with the help of date. We have created month, day and year for providing useful analysis inputs for our data analysis part in answering the questions.

3. What was the process you used for evaluation? What was the best result?

**Ans** : We have used linear regression algorithm to come up with our analysis. We have used lasso and ridge regularization to find the best fit for our model and came to see that linear regression using lasso regularization has provided us with the best result.

4. What were the problems you faced? How did you solve them?

**Ans** : The most problem which we faced while performing our analysis came in with the large number of data available. The data was rigid and we have drop and look for features which would help in our analysis. The data different datatypes for different columns which made up the challenge while encoding the data and had to do best assistance for each indivdual row. The model also needed high computation power due to large dataset and took more time in analysis the required answers.

5. What future work would you like to do?

**Ans : Expand the dataset to include more features:** The current datasets on gun violence in the United States are limited in their scope. They typically only include information about the location, time, and type of gun violence incident. Expanding the dataset to include more features, such as the demographics of the victims and perpetrators, the socioeconomic status of the area, and the presence of gun control laws, could help to better understand the causes of gun violence and develop more effective interventions.

**Develop interventions based on the findings of the analysis:** Once the data has been analyzed, it can be used to develop interventions that are aimed at reducing gun violence. These interventions could include things like gun control laws, mental health services, and social programs.

**Reference**


- https://www.aafp.org/about/policies/all/gun-violence.html
- https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019
- Jaffe, S. (2018). Gun violence research in the USA: the CDC's impasse. The Lancet (British Edition), 391(10139), 2487–2488. https://doi.org/10.1016/S0140-6736(18)31426-0
- https://www.jec.senate.gov/public/_cache/files/69fcc319-b3c9-46ff-b5a6-8666576075fe/the-economic-toll-of-gun-violence-final.pdf
- https://www.w3schools.com/python/matplotlib_pyplot.asp