Capstone Project - Gun Violence in the US

Joshua Kim

Springboard - Data Science Career Track

Table of Contents

Introduction - The Problem & The Goal

- The Problem:
    - One of the biggest issues that America has faced in the past few decades has been the rise of gun violence in civilian life and it has become particularly prevalent in the past decade.
    - According to the U.S. Centers for Disease Control and Prevention, 33,636 Americans were killed in 2013 and that figure rose to 38,658 deaths in 2016.[1] While the official number for the total number of deaths in 2017 has yet to be released, the organization estimates it to surpass 2016 based on end-of-the-year figures.
- The Goal:
    - To analyze and explore data on gun violence in the US over the last few years.
    - To create a Machine Learning model that can predict the number of people killed and injured based on features about the shooting incidents.

Clients - Who Cares?

- Government Agencies (ex. U.S. Centers for Disease Control and Prevention)
  - Identify which factors can help predict the number of casualties in shooting incidents.
  - Determine which models are most efficient for analyzing gun violence data.
- Pro - Gun Control Organizations
  - Inform the public where the most dangerous cities are in terms of gun violence.
  - Understand which factors are most important in predicting casualties even if they don’t have the data.
- You!
  - Learn about the demographics of the most susceptible people to gun violence.
  - Understand the characteristics of the incident participants.

Dataset

- ★ The data was provided by James Ko: https://www.kaggle.com/jameslko/gun-violence-data
- ★ Data was web-scraped from: http://www.gunviolencearchive.org/
- ★ Contains gun-violence data in the US from 2013 - 2018.
- ★ Holds 239, 677 rows of data and 29 columns.
- ★ Missing a lot of data from 2013 and only contains incidents up until March 2018.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 29 columns):
incident_id                    239677 non-null int64
date                           239677 non-null datetime64[ns]
state                          239677 non-null object
city_or_county                 239677 non-null object
address                        223180 non-null object
n_killed                       239677 non-null int64
n_injured                      239677 non-null int64
incident_url                   239677 non-null object
source_url                     239209 non-null object
incident_url_fields_missing    239677 non-null bool
congressional_district         227733 non-null float64
gun_stolen                     140179 non-null object
gun_type                       140226 non-null object
incident_characteristics       239351 non-null object
latitude                       231754 non-null float64
location_description           42089 non-null object
longitude                      231754 non-null float64
n_guns_involved                140226 non-null float64
notes                          158660 non-null object
participant_age                147379 non-null object
participant_age_group          197558 non-null object
participant_gender             203315 non-null object
participant_name               117424 non-null object
participant_relationship       15774 non-null object
participant_status             212051 non-null object
participant_type               214814 non-null object
sources                        239068 non-null object
state_house_district           200905 non-null float64
state_senate_district          207342 non-null float64
dtypes: bool(1), datetime64[ns](1), float64(6), int64(3), object(18)
memory usage: 51.4+ MB
```

Data Dictionary

1. incident_id: ID of the crime report
2. date: Date of crime
3. state: State of crime
4. city_or_county: City/ County of crime
5. address: Address of the location of the crime
6. n_killed: Number of people killed
7. n_injured: Number of people injured
8. incident_url: URL regarding the incident
9. source_url: Reference to the reporting source
10. incident_url_fields_missing: TRUE if the incident_url is present, FALSE otherwise
11. congressional_district: Congressional district id
12. gun_stolen: Status of guns involved in the crime (i.e. Unknown, Stolen, etc...)
13. gun_type: Typification of guns used in the crime
14. incident_characteristics: Characteristics of the incidence
15. latitude: Location of the incident
16. location_description: Location description
17. longitude: Location of the incident
18. n_guns_involved: Number of guns involved in incident
19. notes: Additional information of the crime
20. participant_age: Age of participant(s) at the time of crime
21. participant_age_group: Age group of participant(s) at the time crime
22. participant_gender: Gender of participant(s)
23. participant_name: Name of participant(s) involved in crime
24. participant_relationship: Relationship of participant to other participant(s)
25. participant_status: Extent of harm done to the participant
26. participant_type: Type of participant
27. sources: Participants source
28. state_house_district: Voting house district
29. state_senate_district: Territorial district from which a senator to a state legislature is elected.

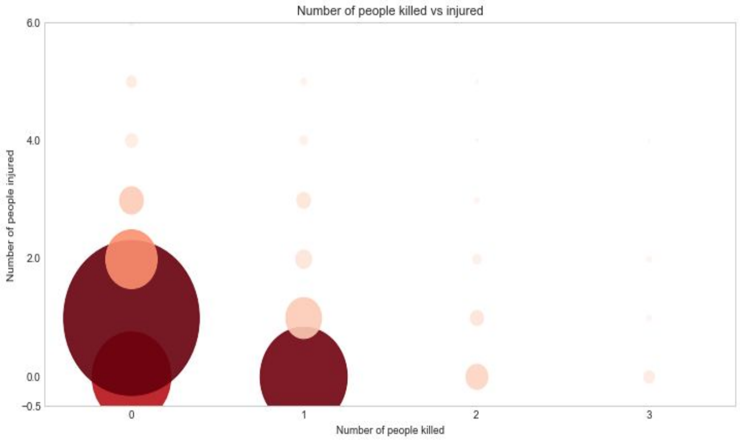
Data Wrangling

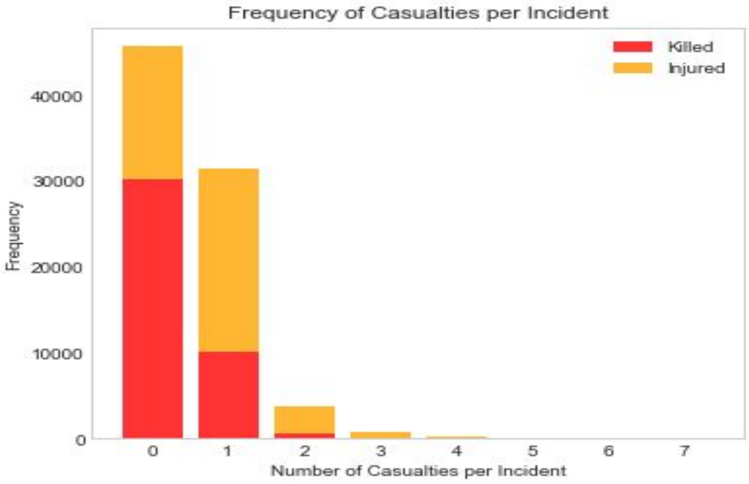
★ Drop columns that are irrelevant to the project
★ Remove columns/rows with excessive amount of missing data
★ Choose age-group column over age column for better prediction
★ Create pseudo-dummy columns that counts the number of genders/age-groups in a single row
★ Include only the top 15 cities with 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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35727 entries, 7 to 239675
Data columns (total 15 columns):
n_killed                    35727 non-null int64
n_injured                   35727 non-null int64
congressional_district      35727 non-null float64
state_house_district        35727 non-null float64
state_senate_district       35727 non-null float64
agegroup_child              35727 non-null int64
agegroup_teen               35727 non-null int64
agegroup_adult              35727 non-null int64
year                        35727 non-null int64
month                       35727 non-null int64
monthday                    35727 non-null int64
weekday                     35727 non-null int64
participant_gender_male     35727 non-null int64
participant_gender_female   35727 non-null int64
mapped_cities               35727 non-null int8
dtypes: float64(3), int64(11), int8(1)
memory usage: 5.4 MB
```

Number of Casualties (Deaths and Injuries)

★ <u>Most shooting incidents resulted in no deaths</u> but there were still nearly 10,000 incidents where 1 person died.
★ There were more incidents that resulted in an injury than incidents without an injury.



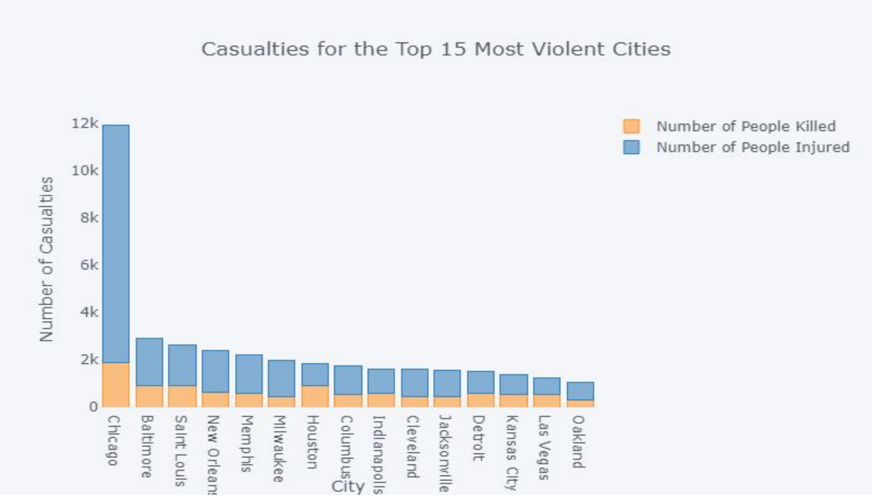


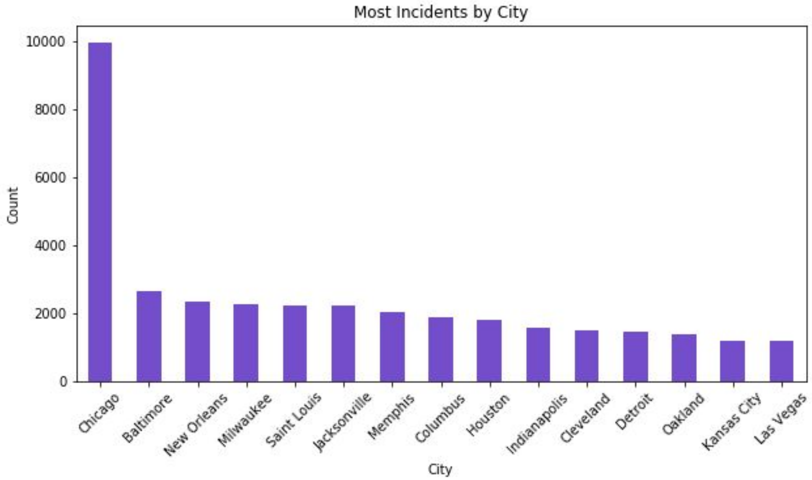
★ The most frequent casualty scenario was when there were <u>0 deaths and 1 injury</u>.
★ Very few incidents that had more than 2 people killed or more than 3 people injured.

Casualties by City

★ Chicago by far had the most number of shooting incidents based on the dataset.

★ Other violent cities included Baltimore, New Orleans and Milwaukee.





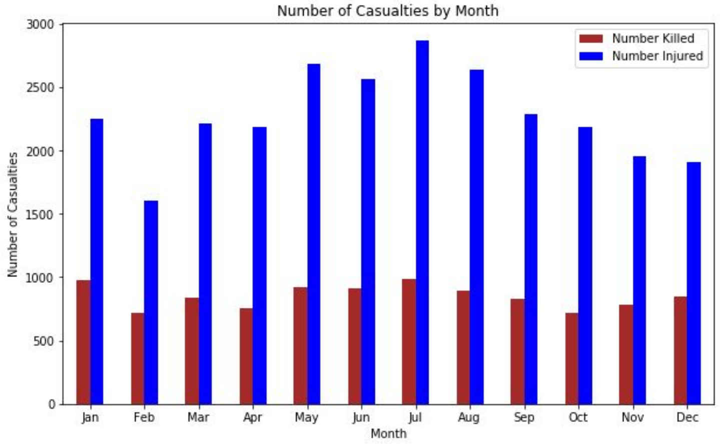
★ Chicago also had the most deaths and injuries.

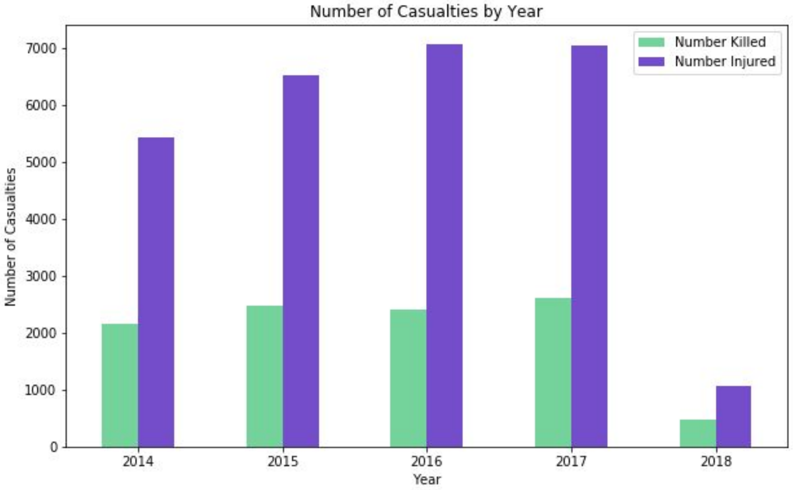
★ Houston had the 2nd most deaths but was only 7th in overall casualties.

Casualties by Date

★ Number of casualties has overall been growing each year from 2014 - 2017 for the top 15 cities.

★ There were slightly more injuries in 2016 than 2017, but fewer deaths.

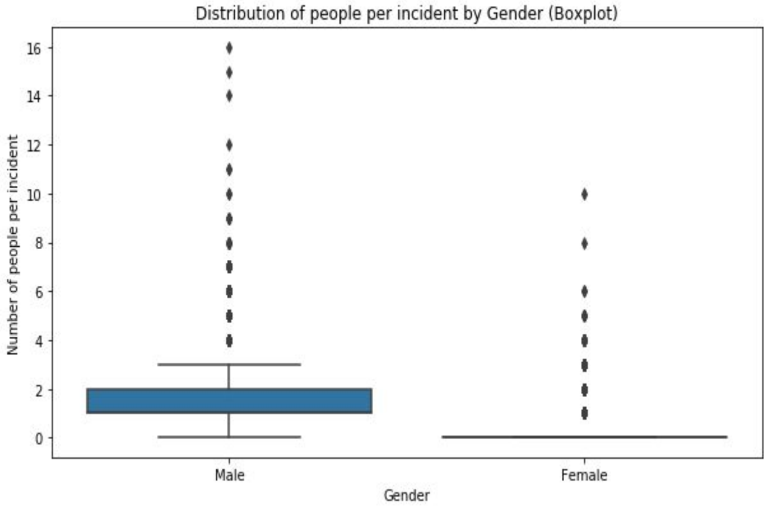


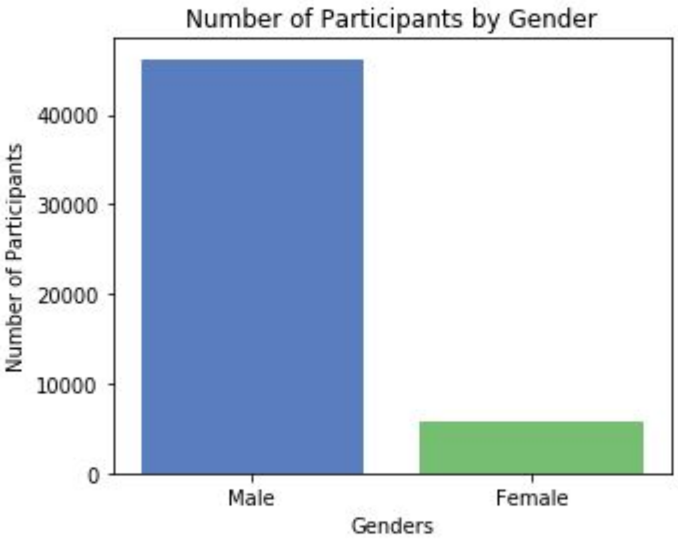


★ Casualties are at their highest in the summer months, with the peak being in July for both deaths and injuries.

Casualties by Gender

★ There were over 40,000 male participants compared to less than 6,000 female participants.





★ Most incidents had at least 1 male involved.

★ There were so few females that it skews the bulk of the distribution to nearly 0.

★ Even having 1 female is considered an outlier.

Feature Selection

★ 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 (n_killed) and the other for predicting the number of people injured (n_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.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35727 entries, 7 to 239675
Data columns (total 13 columns):
congressional_district      35727 non-null float64
state_house_district        35727 non-null float64
state_senate_district       35727 non-null float64
agegroup_child              35727 non-null int64
agegroup_teen               35727 non-null int64
agegroup_adult              35727 non-null int64
year                        35727 non-null int64
month                       35727 non-null int64
monthday                    35727 non-null int64
weekday                     35727 non-null int64
participant_gender_male     35727 non-null int64
participant_gender_female   35727 non-null int64
mapped_cities               35727 non-null int8
dtypes: float64(3), int64(9), int8(1)
memory usage: 4.8 MB
```

Baseline/Linear Regression

- ★ A simple baseline was created using the DummyRegressor regressor to compare with the results of the models.
- ★ 2 primary metrics were used:
  - RMSE (Root Mean Squared Error)
  - R-Squared
- ★ 3 different Linear Regression Models were used:
  - Linear Regression without Regularization
  - Lasso Regularization
  - Ridge Regularization
- ★ Application: The number of people injured had the highest coefficient value for predicting the number of people killed, and vice versa.
  - For every person injured in a shooting incident, there was a *decrease* of 0.3 deaths. In other words, there was a negative correlation between the 2 variables.

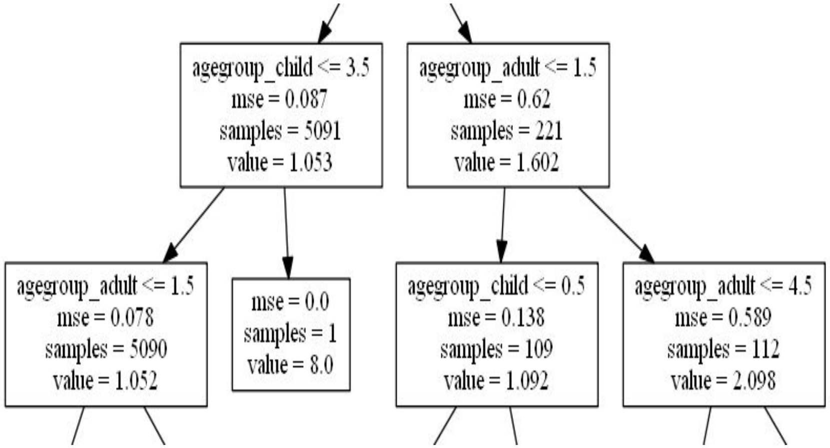
Baseline	n_killed	n_injured
RMSE	0.5067	0.7654
R-Squared	-0.0001	0
Linear Regression	n_killed	n_injured
RMSE	0.4123	0.6005
R-Squared	0.3193	0.4207

Decision Tree

- ★ Models were run using RandomSearchCV to obtain the optimal hyper-parameters.
- ★ Using Decision Tree model had better results than using Linear Regression due to its nature of asking sequential questions and ability to handle large datasets well.
- ★ The model was further optimized by removing "noisy" features (little impact on the response variables) and outliers (ex. Incidents with more than 2 people killed).

- ★ Application: The decision tree splits at each node based on a characteristic. Ex. One branch made its decisions based on how many adults or children were part of a shooting incident.

Optimized Decision Tree	n_killed	n_injured
RMSE	0.3304	0.4898
R-Squared	0.5348	0.4666



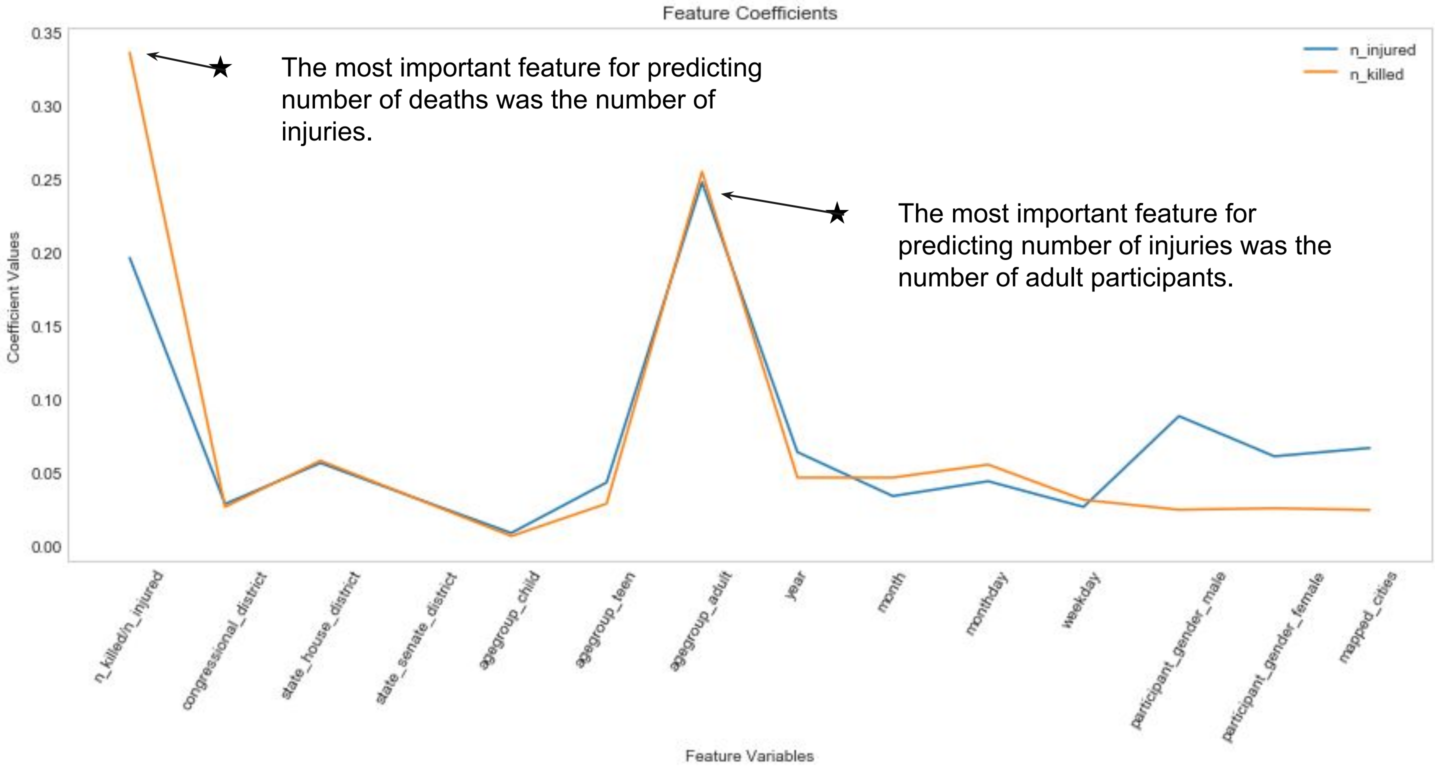
Random Forest

- ★ Random forest uses a collection of decision trees so that results are aggregated into a final result.
- ★ This is done to reduce the risk of overfitting the model by averaging the results of multiple decision trees.

- ★ Application: The most important features for predicting the number of people killed were the number of people injured, number of adults and state house district.
- ★ For predicting the number of people injured, it was the number of adults, the number of people killed, and the number of males.

Random Forest	n_killed	n_injured
RMSE	0.3401	0.5379
R-Squared	0.5398	0.5110
Optimized Random Forest	n_killed	n_injured
RMSE	0.3137	0.4680
R-Squared	0.5718	0.5245

Random Forest Model Visualization



Best Models

- ★ The 3 best performing models in this project were:
  - Random Forest with optimization
  - Decision Tree with optimization
  - Random Forest without optimization
- ★ Random Forest with optimization came out on top for predicting both the number of people killed and number of people injured.

	n_killed		n_injured	
	RMSE	R-Squared	RMSE	R-Squared
Random Forest with Optimization	0.3137	0.5718	0.4680	0.5245
Decision Tree with Optimization	0.3304	0.5348	0.4898	0.4666
Random Forest	0.3401	0.5398	0.5379	0.5110

Conclusion

- ★ Main factors in predicting the number of people killed or injured:
  - Number of adults involved (ages 18 and over).
  - Number of injuries (for predicting number of deaths) and vice versa.
  - Number of males involved.
- ★ Recommendations:
  - Number of injuries can be a useful predictor of the number of deaths (and vice versa) due to the negative correlation between the 2 variables.
  - Check the number of adults; incidents with adults are more likely to result in casualties than with children or teens.
- ★ Next Steps:
  - Use classification models (ex. Logistic Regression) to classify incidents based on thresholds (ex. Incidents with 2+ deaths vs incidents with 0 or 1 deaths)
  - Find other datasets that contain more information regarding the shooting incidents (ex. Gun type, ethnicity).
  - Analyze other big cities such as New York and Los Angeles.

Appendix

1. https://www.thetrace.org/rounds/gun-deaths-increase-2017/