



# Capstone Project - Gun Violence in the US

Joshua Kim

Springboard - Data Science Career Track

# Table of Contents



1. Cover Page
2. Table of Contents
3. Introduction
4. Clients
5. Dataset
6. Data Dictionary
7. Data Wrangling
8. Number of Casualties
9. Casualties by City
10. Casualties by Date
11. Casualties by Gender
12. Casualties by Incident Characteristics
13. Feature Selection
14. Baseline/Linear Regression
15. Decision Tree
16. Random Forest
17. Random Forest Model Visualization
18. Best Models
19. Conclusion
20. Appendix

# 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.<sup>1</sup> 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.

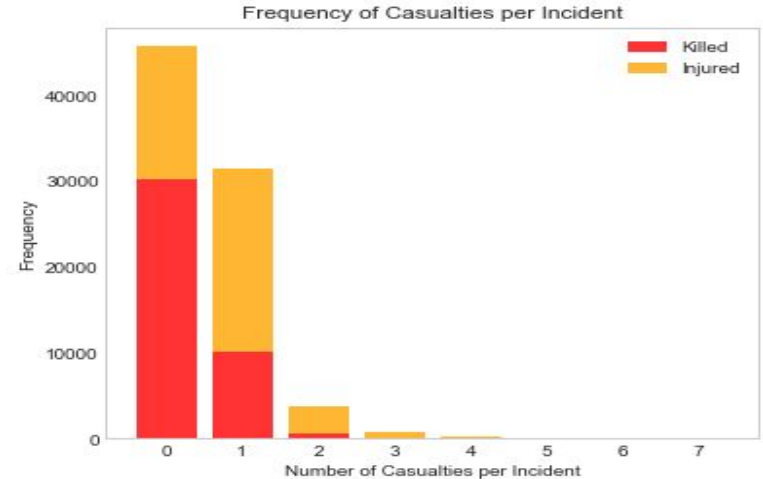
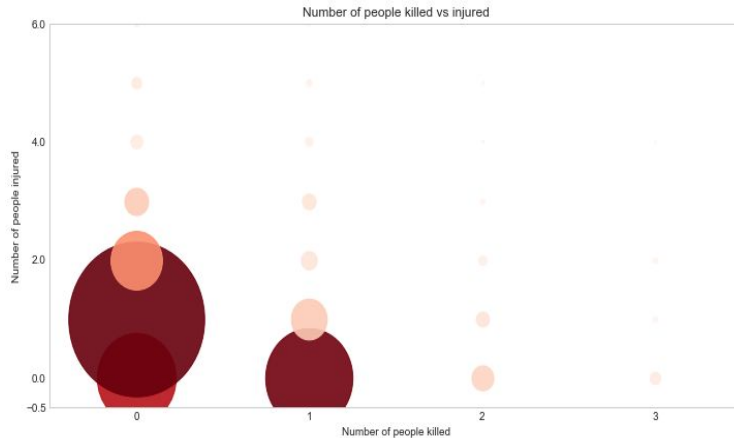
# Data Wrangling

- ★ 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)
- ★ Create 2 columns for the first 2 incident characteristics from each row
- ★ 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 28 columns):
date                35727 non-null datetime64[ns]
state               35727 non-null object
city_or_county      35727 non-null category
n_killed            35727 non-null int64
n_injured           35727 non-null int64
congressional_district 35727 non-null float64
gun_type            35727 non-null object
incident_characteristics 35727 non-null object
participant_age_group 35727 non-null object
participant_gender   35727 non-null object
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_total 35727 non-null int64
participant_gender_male 35727 non-null int64
participant_gender_female 35727 non-null int64
participant_gender_unknown 35727 non-null int64
primary_characteristic 35727 non-null category
secondary_characteristic 15240 non-null category
mapped_cities       35727 non-null int8
mapped_primary      35727 non-null int8
mapped_secondary     35727 non-null int8
dtypes: category(3), datetime64[ns](1), float64(3), int64(13), int8(3), object(5)
memory usage: 7.7+ MB
```

# Number of Casualties (Deaths and Injuries)

- ★ Most shooting incidents resulted in no deaths 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.

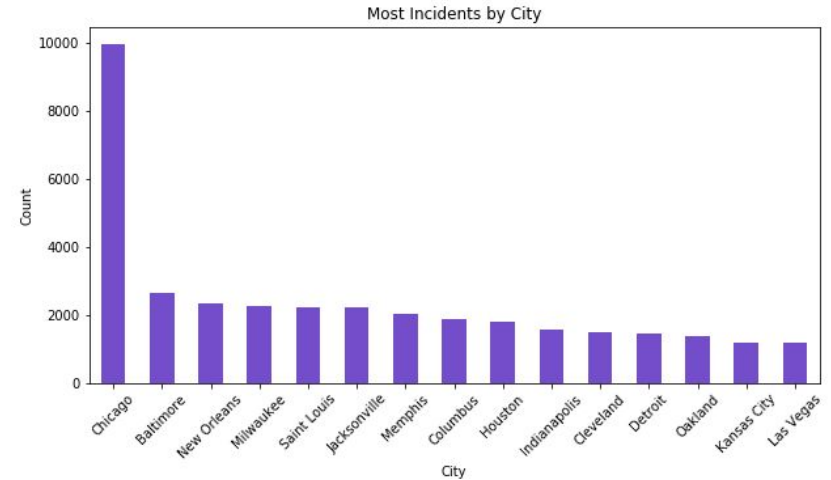
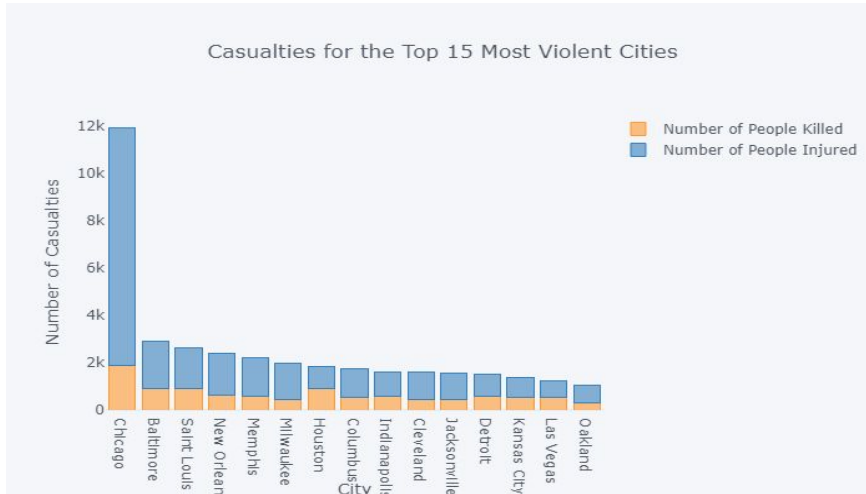


- ★ The most frequent casualty scenario was when there were 0 deaths and 1 injury.
- ★ Very few incidents that had more than 2 people killed or more than 3 people injured.



# Casualties by City

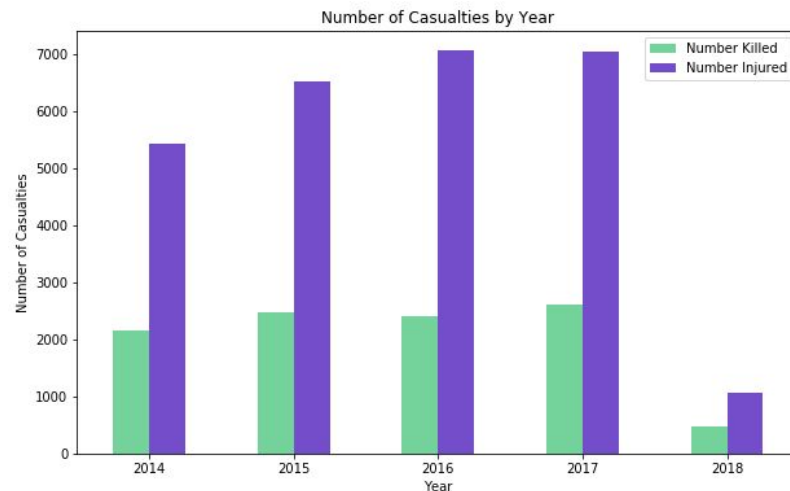
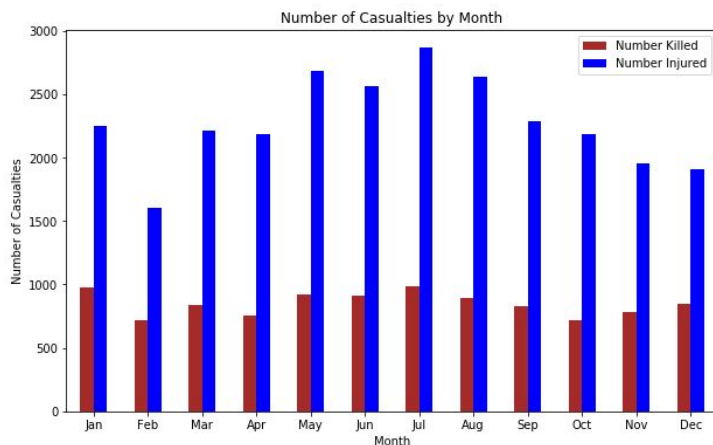
- ★ Chicago by far had the most number of shooting incidents based on the dataset.
- ★ Other violent cities included Baltimore, New Orleans and Milwaukee.



- ★ Chicago also had the most deaths and injuries.
- ★ Houston had the 2nd most deaths but was only 7th in overall casualties.

# Casualties by Date

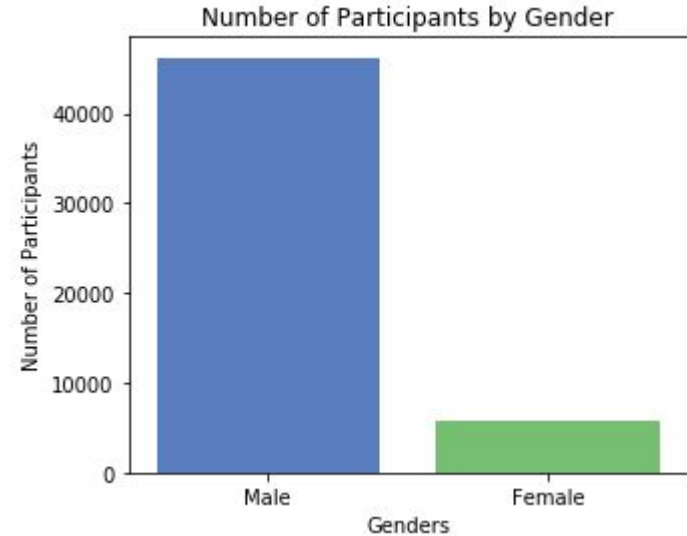
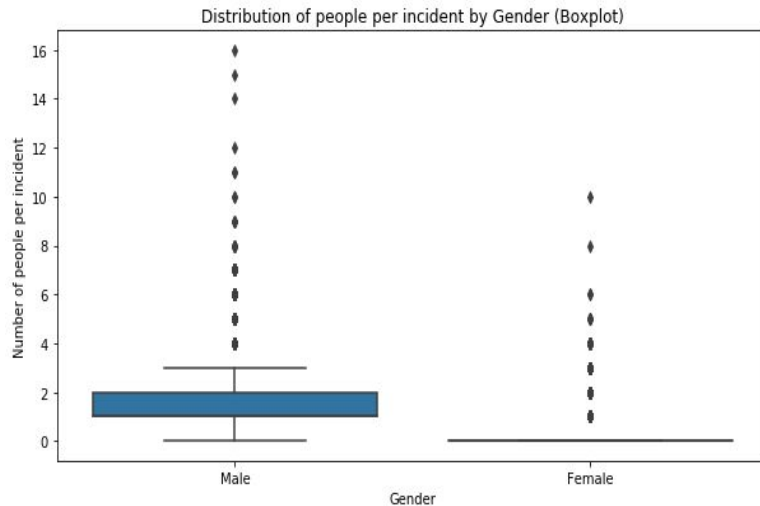
- ★ Number of casualties has overall been growing each year from 2014 - 2017 for the top 15 cities.
- ★ There were slightly more injuries in 2016 than 2017, but fewer deaths.



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

# Casualties by Gender

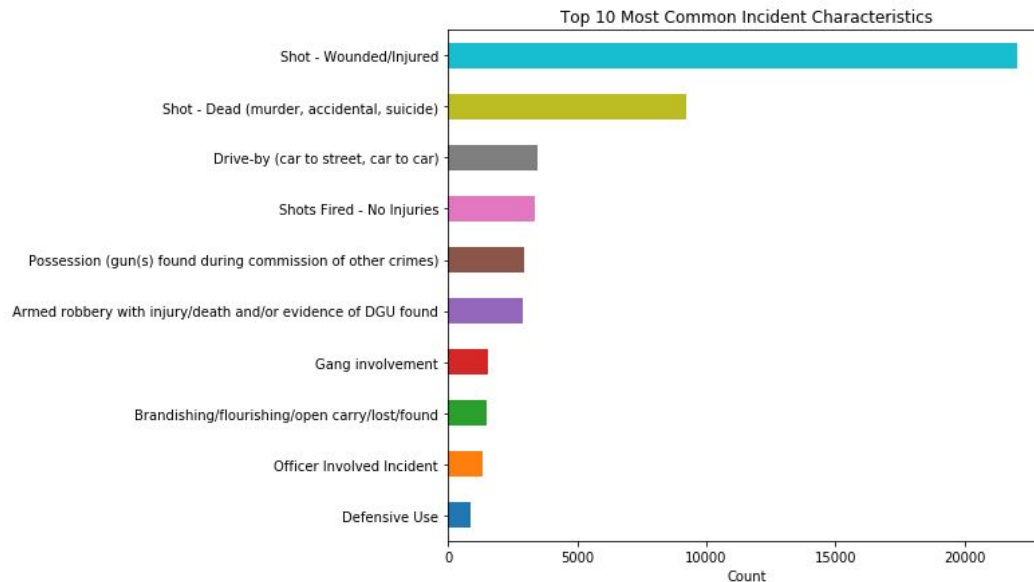
- ★ There were over 40,000 male participants compared to less than 6,000 female participants.



- ★ 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.

# Casualties by Incident Characteristics

- ★ The two most common incident characteristics were 'Shot - Wounded' and 'Shot - Dead'.
- ★ 'Shots Fired - No Injuries' came up less frequently than 'Drive-by', indicating there were more incidents with casualties than none.



# 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 16 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_adult            35727 non-null int64
agegroup_teen             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
participant_gender_unknown 35727 non-null int64
mapped_primary            35727 non-null int8
mapped_cities             35727 non-null int8
mapped_secondary          35727 non-null int8
dtypes: float64(3), int64(10), int8(3)
memory usage: 5.2 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
- ★ Linear Regression without Regularization performed the best out of the 3 but did not have good results compared to the ensemble methods.

<u>Baseline</u>	n_killed	n_injured
RMSE	0.5067	0.7654
R-Squared	-0.0001	0

<u>Linear Regression</u>	n_killed	n_injured
RMSE	0.3924	0.5719
R-Squared	0.3835	0.4747

# 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 to no impact on the response variables)
  - Outliers (incidents with more than 3 people injured or more than 2 people killed).
- ★ This resulted in a better RMSE and R-Squared for both response variables.

<u>Decision Tree</u>	n_killed	n_injured
RMSE	0.1518	0.3940
R-Squared	0.9085	0.7462

<u>Optimized Decision Tree</u>	n_killed	n_injured
RMSE	0.1258	0.3143
R-Squared	0.9325	0.7776

# Random Forest

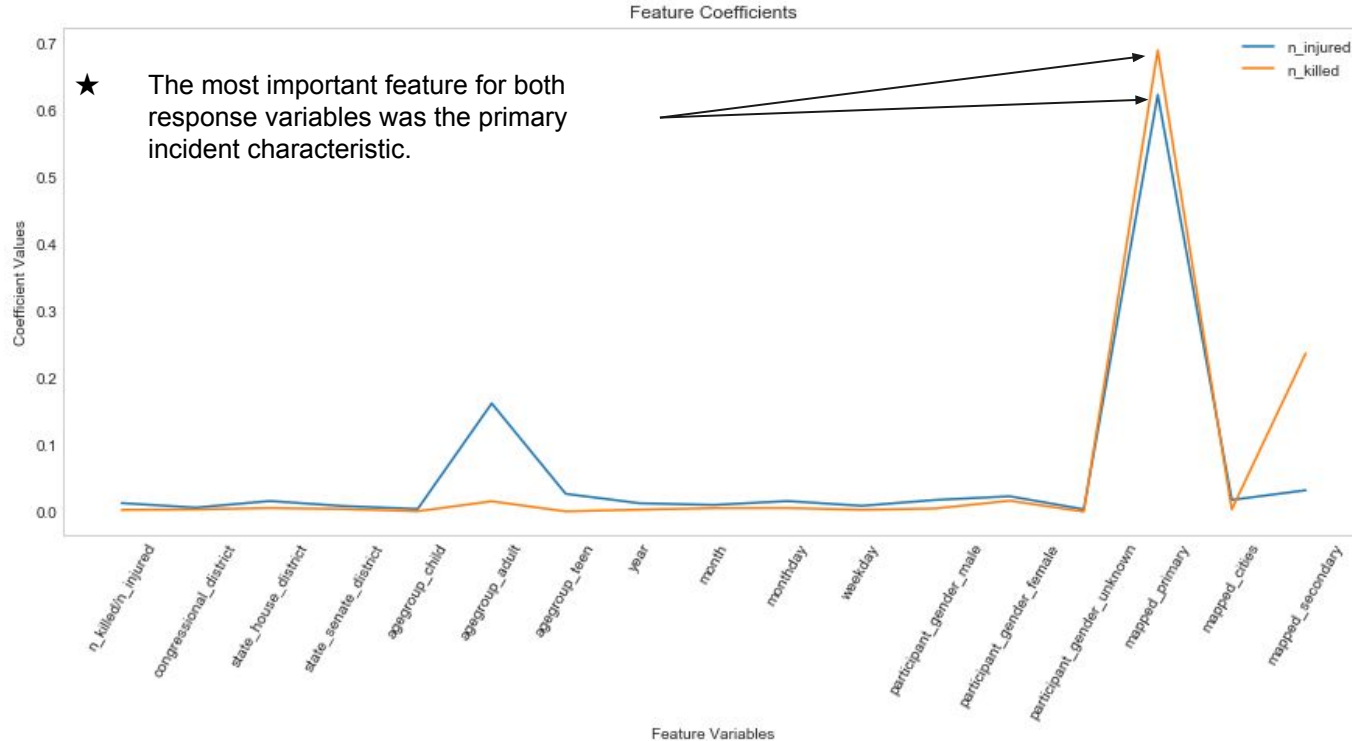
- ★ Random forest uses a collection of decision trees and results are aggregated into a final result.
- ★ The model was optimized the same way as the decision tree (removing same features/outliers).
- ★ The most important features for predicting the number of people killed were the primary characteristics, secondary characteristics, the number of female participants and number of adults.
- ★ There were several more important features for predicting the number of people injured such as the year and the city.

<u>Random Forest</u>	n_killed	n_injured
RMSE	0.1566	0.3589
R-Squared	0.9023	0.7822

<u>Optimized Random Forest</u>	n_killed	n_injured
RMSE	0.1205	0.2919
R-Squared	0.9368	0.8149



# 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
<u>Random Forest with Optimization</u>	0.1205	0.9368	0.2919	0.8149
<u>Decision Tree with Optimization</u>	0.1258	0.9325	0.3143	0.7776
<u>Random Forest</u>	0.1566	0.9023	0.3589	0.7822

# Conclusion



- ★ Main factors in predicting the number of people killed or injured:
  - Incident characteristics (ex. 'Shot - Dead', 'Shot - Wounded', 'Drive-by', 'Armed Robbery')
  - Number of adults involved (ages 18 and over)
  - Number of females involved
- ★ Recommendations:
  - Use incident characteristics to identify patterns regarding the shooting incidents.
  - Females are involved significantly less in shooting incidents than males but their presence correlates to higher injury and death counts.
- ★ 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 gun types that were used by shooters.
  - Analyze other big cities such as New York and Los Angeles.

# Appendix



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