Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

Business Problem

Gun violence is certainly not an issue likely to go away anytime soon. More and more spree shooters in particular are causes for additional concerns and Parkland and Sandy Hook still hang in the slight rearview, and the city of Chicago has had over 1,000 homicides this year (on track to set a record from 1994 of 1,141). Each of these has different motives and actions involved with shooting incidents. Understanding these circumstances, and grouping them based on historical events may provide insight into the factors that play into them, which may in turn help us prevent such circumstances from coming together in the first place.

Background/History

Firearms are weapons. From their onset that is what they've been designed to be used for regardless of the person or object on the receiving end. While their use in hunting and wars provides an, arguably, positive effect for the person using them, crimes and accidental shootings involving them are distinctly negative (except in cases of self-defense and other unique scenarios). In recent years, it has continued to be a topic of large debate on whether gun control is effective or should be implemented in order to prevent these negative aspects. As this is a complex ethical topic, the conflict may never be resolved there.

Revolving around that debate, and the emphasis it has cast on gun violence over the past years, we can swiftly remember instances such as Sandy Hook, Parkland, and the Las Vegas shootings. Additionally, in Illinois it has been consistently reported that the city of Chicago is on pace to set a new record for murders in 2021, previously set in 1994 as seen in Figure 1. It is clear that gun violence, and the need to understand why they occur, is still just as important as ever. With that being said, our goal is to identify clusters of violent gun incidents from historical data and compile these groups to identify the factors that go into them, such that we can prevent these 'perfect storm' situations from coming together and resulting in additional fatalities.

Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

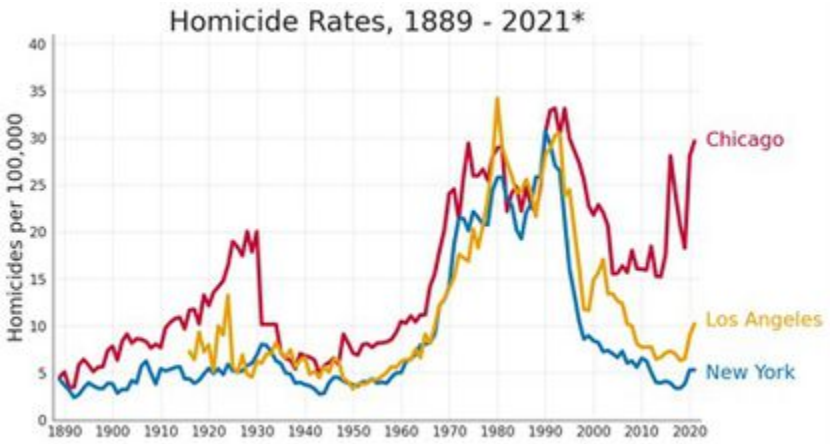

Figure 1: Gun Violence in Chicago over Time

Data Explanation

The primary dataset we will use will be provided from the Kaggle user James Ko, who generated it from the gunviolencearchive.org website, which can be found below:

https://www.kaggle.com/jameslko/gun-violence-data

The dataset is made up of over 239,000 entries and 29 columns of feature data information of varying data types. According to the author, the data may be described as: "...CSV file [that] contains data for all recorded gun violence incidents in the US between January 2013 and March 2018, inclusive…". The originating data columns were as follows: incident_id, date, state, city_or_county, address, n_killed, n_injured, incident_url, source_url, incident_url_fields_missing, congressional_district, gun_stolen, gun_type, incident_characteristics, latitude, location_description, longitude, n_guns_involved, notes, participant_age, participant_age_group, participant_gender, participant_name, participant_relationship, participant_status, participant_type, sources, state_house_district, and state_senate_district.

In order to perform some feature reduction, we first analyzed each column to determine if the columns were set as numerical data or categorical. For categorical columns we aimed to one-hot-encode as many as reasonable for inclusion in our clustering algorithm. Those we could not or deemed irrelevant to our clustering were dropped. As such, once the drops were completed we had only: state, n_killed, n_injured, gun_stolen, gun_type, latitude, longitude, n_guns_involved, participant_age, participant_gender, and participant_type left.

Of the remaining columns, state was then one-hot-encoded. Gun_stolen was listed as a number of inputs for each gun as either unknown, stolen, or not stolen. We switched this column to a boolean value if any guns were found to be stolen. We then parsed out

gun_type and one-hot-encoded that column, and cross-referenced gender and age columns against participant type to identify the ages and genders of the suspects involved in the case. Once the cross-reference was complete, we also one-hot-encoded the gender of suspects.

Wrapping up our data prep, we filled in missing values according to domain knowledge (i.e. since guns were involved in all these incidents, missing values under n_guns was presumed to be 1).

Methods

Once data preparation was complete, we began to implement the dbscan machine learning algorithm. The first step in doing so was determining the minimum points and epsilon value ideal for describing our clusters. Based on research we found (Mullin), we identified that the best minimum points value would be two times the number of dimensions found in our dataset.

With this being said, we identified that if we count one-hot-encoded columns as 1 dimension, we'd have a total of 10 total dimensions, so our best minimum points value would thus be 20.

Once this value was determined, we could then discern the best value for epsilon by performing k-nearest neighbors analysis to identify the elbow point of the distances from the center of the clusters. Plotted out as seen in Figure 2, we found the elbow to be ε=2. From here we applied our DBScan method with our minimum points and epsilon value accordingly.

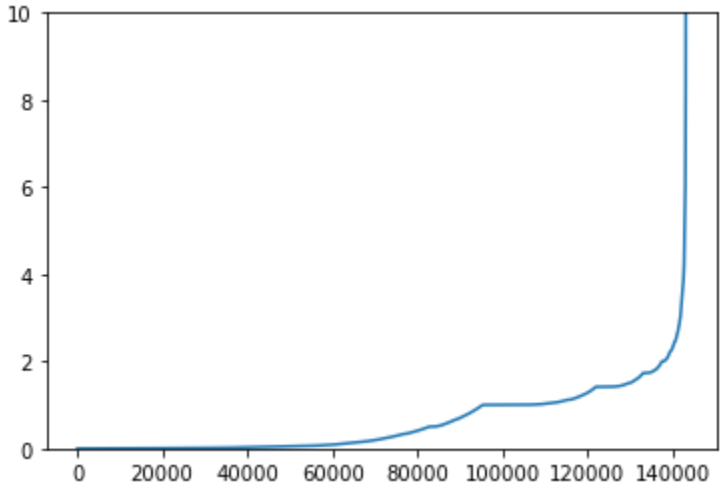

Figure 2: Points Sorted by Distance to the 20th Nearest Neighbor

Analysis

Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

The resulting outputs of our clustering algorithm was that we identified that from the roughly 143,000 incidents of gun violence provided during the 5 year span from 2013 to 2018, these could be clustered into 474 unique clusters of incident types. We identified our silhouette score, which ranges from -1 (representing too many or too few clusters) to +1 (representing that the clustering configuration is appropriate) based on how similar each incident is compared to other incidents within the cluster as opposed to those incidents found in other clusters, to be a respectable 0.39. From here we also visualized the cluster/noise spectrum via matplotlib found in Figure 3 below:

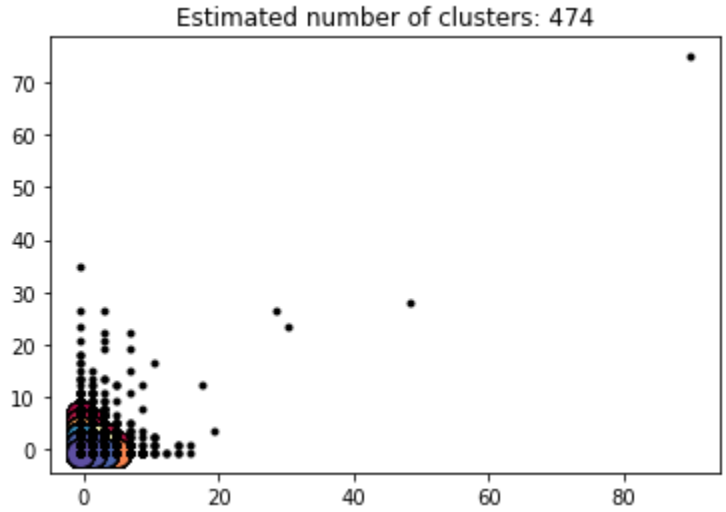

Figure 3: Clustering Distributions

Conclusion

Re-examining the analysis found and how it relates to the business problem, we have uniquely identified 474 distinct clusters out of the multitude of gun-related incidents over the past 5 years. These groupings can be deciphered as situationally implicit identifiers for individuals and environments that can be at risk. Through our groupings, further predictive work could be done on additional data to identify clusterings a target may belong to and discern risk factors that can then be minimized to prevent future gun violence incidents.

Assumptions

One of the largest assumptions from our data is that leveraging historical data clustered here, one can predict possible incidents in the future. Additionally, several assumptions were made in the data prep stage surrounding missing values and aggregation of data including:

- If multiple guns were used in incident, any stolen guns should characterize incident as including stolen gun

- If multiple suspects were involved in an incident, an adequate measure of analyzing age was determined as finding the mean of all suspect ages
- If multiple suspects were involved in an incident, an adequate categorization of the column would be identified as “both”
- If no number of guns was listed, 1 was assumed as all had to have at least one gun involved to be included in dataset
- Latitude and longitude were set to 0 if not provided

Limitations

Some limitations of the analysis include the lack of recent data coming after 2018, lack of analysis of incident characteristics included in the dataset as one-hot-encoding of these values would grossly expand the number of columns and thus the complexity and processing time of the clustering, similar lack of analysis on zip code/other geographically specific data for same reason. Additional clustering techniques could also have been used for comparison of cluster size.

Challenges

Challenges faced during the analysis of the dataset were largely around getting the data into an interpretable and encodable format. Most were expressed with numerical delimiters relating to either a victim or suspect.

Manual extractions were used in order to separate these and accordingly designate column values in order to be further one-hot-encoded for categorical analysis through the clustering algorithm. The manual extraction process is cpu-intensive and any error messages along the way required re-running this piece which took additional time. Reworking said sections would be key to making code more efficient.

Future Uses / Additional Applications

Future use of these cluster outputs would revolve around predictive applications to discern clusters a possible incident would fit in based on suspect, state, lat, long, and other available information may fit in and accordingly attempt to prevent such circumstances as those found within the cluster from matching up to at-risk situations. Additionally, linking up information found here with that of additional indicators that can be found for these incidents that can further define the clusters would be ideal to expand the understanding of each cluster’s features and further fleshing them out.

Recommendations

Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

Leveraging this model, particularly if paired with predictive insights, can assist in determining scenarios that could evolve into gun violence. Determining these situations and actively intercepting them or minimizing exposures of certain individuals/communities to risk factors found in-cluster are keys to lowering gun violence going forward and saving lives of victims and in many cases the suspect as well.

Implementation Plan

In order to implement this model according to the recommendations, the steps needed to be taken include:

- Identify at-risk individuals, or at-risk situations
- Leveraging modeling technique to identify cluster of individual/situation to determine similar in-cluster incidents
- Minimize risk factors that uniquely identify the group

Ethical Assessment

Characteristics found in common for these groupings need to be maintained as general predictions or summarizations, not sureties of potential misdoings. General data purity and avoidance of data manipulation needs to be evaluated. Data collection only was maintained through 2018, so lack of direct recency, which may be augmented with additional data later in study if found.

Rows not containing suspect information were dropped from the study in the data cleaning phase. These could provide further cluster distinction that provides differing insights. Inclusion of gender data could be misleading, particularly if said features were found to be largely deciding factors for any clustering, and could, if necessary, be dropped from the study entirely if found to be overly indicative.

Questions:

1. What predictive analytics algorithm would you leverage in order to apply these modelings to future incidents?
2. How did you discern your silhouette measurement to be significant?
3. Which features were most important in determining the clusterings chosen?
4. How would this modeling technique compare to that of a k-means clustering model?
5. Keying off that, what made you make the initial decision to use DBScan in the first place?
6. The elbow graph you showed in Figure 2 does not show precisely where epsilon value is best. How would changing that value affect the number of clusters created?

Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

7. As opposed to finding the mean average of the ages, would the median ages be a better assumption to use?
8. Are there any ways to prevent ethical challenges regarding incidents revolving around specific communities?
9. Besides more data, are there any other ways you can think of to improve the silhouette score of your clustering algorithm?
10. Is there a way you could have included the date column into your clustering to provide it as a method of clustering?

Rickord, Jake JJ
Applied Data Science
Project 1 Milestone 2

Appendix

1. C. (2015, June 16). *GitHub - choffstein/dbscan: Python implementation of "Density Based Spatial Clustering of Applications with Noise."* GitHub. https://github.com/choffstein/dbscan
2. Chauhan, N. S. (2020, April). *DBSCAN Clustering Algorithm in Machine Learning*. KDnuggets. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html
3. GeeksforGeeks. (2019, June 6). *Implementing DBSCAN algorithm using Sklearn*. https://www.geeksforgeeks.org/implementing-dbscan-algorithm-using-sklearn/
4. *Gun Violence Archive*. (2013). Gun Violence Archive. https://www.gunviolencearchive.org/
5. *How to determine epsilon and MinPts parameters of DBSCAN clustering*. (2021, October 8). Amir Masoud Sefidian. http://sefidian.com/2020/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/
6. Lutins, E. (2020, December 4). *DBSCAN: What is it? When to Use it? How to use it - Evan Lutins*. Medium. https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
7. Mullin, T. (2021, December 15). *DBSCAN Parameter Estimation Using Python - Tara Mullin*. Medium. https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd
8. Prado, K. S. D. (2019, June 3). *How DBSCAN works and why should we use it? - Towards Data Science*. Medium. https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
9. Shreiber, A. (2021, December 14). *A Practical Guide to DBSCAN Method - Towards Data Science*. Medium. https://towardsdatascience.com/a-practical-guide-to-dbscan-method-d4ec5ab2bc99
10. *So You Have Some Clusters, Now What?* (2017, November 9). Square Corner Blog. https://developer.squareup.com/blog/so-you-have-some-clusters-now-what/
11. Victor, D., & Taylor, D. B. (2021, December 2). *Mass Shootings in the United States in 2021*. The New York Times. https://www.nytimes.com/article/mass-shootings-2021.html

12. *What is Clustering?* | *Clustering in Machine Learning* |. (2020, February 10). Google Developers. https://developers.google.com/machine-learning/clustering/overview

13. (2012, December 19). *Mass Shooting Psychology: Spree Killers Have Consistent Profile, Research Shows*. HuffPost. https://www.huffpost.com/entry/mass-shooting-psychology-spree-killers_n_2331236