Joshua Fraser (jdf254)
Ayman Naji (an443)
Laura Gonzalez (lg458)

November 8, 2020

Project Midterm Report

Description of Dataset:

We have decided to change our dataset from one that strictly encompassses fatal shootings, to one that encompasses fatal and non-fatal shootings. Our dataset **(linked here)** from VICE is sourced from the 50 largest police departments in the US which employ nearly 148,000 officers and serve 54 million Americans. The data was sourced directly from law enforcement agencies and district attorneys, and utilized local media reports in cases where law enforcement agencies did not relay information.

The dataset consists of 4400 police shootings from 2010-2016. It contains descriptive data about the victims such as age, gender and race. In terms of how the shooting transpired, the dataset includes the date and city of the event, nature of event, as well as the race and gender of the responding police officer. Due to the variation in transparency and methods of record-keeping for the law enforcement agencies, the dataset has some missing entries ("NA") and slight inconsistencies that would have enriched the analysis further. For example, in cases where the number of individuals involved in the shooting was unknown, the dataset assumed there to be one victim. Nonetheless, this dataset is highly comprehensive as it includes features not found in other datasets, and is distinguished by the inclusion of non-fatal shootings. This inclusion of non-fatal shootings data can allow us to predict whether a future shooting was fatal or not based on features such as race, gender, age, whether the subject was armed, etc.

Categories of Data:

Date (Int)
NumberOfSubjects (Float64)
Fatal (Char)
SubjectArmed (Char)
SubjectRace (Char)
SubjectGender (Char)
SubjectAge (Int)
NatureOfStop (String)

NumberOfShots (Int)
NumberOfOfficers (Int)
OfficerRace (String)
OfficerGender (String)
Department (String)
FullNarrative (String)
City (String)
Notes (String)

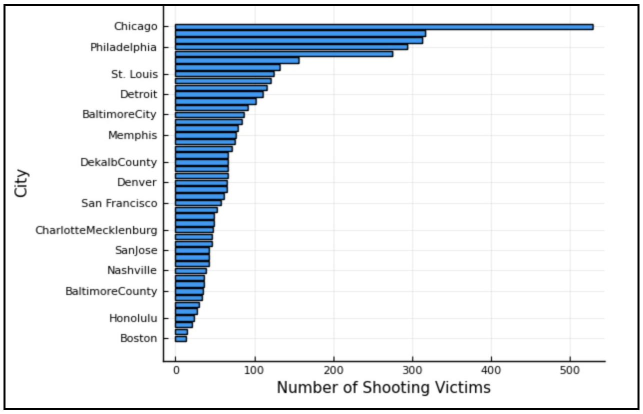
Testing the Effectiveness of our Models:

When testing the effectiveness of our model we will divide our data into a training and test set. We plan on selecting the first 80% as the train data, with 20% held out for validation. We must shuffle the data because our dataset of the shooting and create two random samples. This also ensures the model isn't trained on the training set, and we can make general predictions about future police shootings.

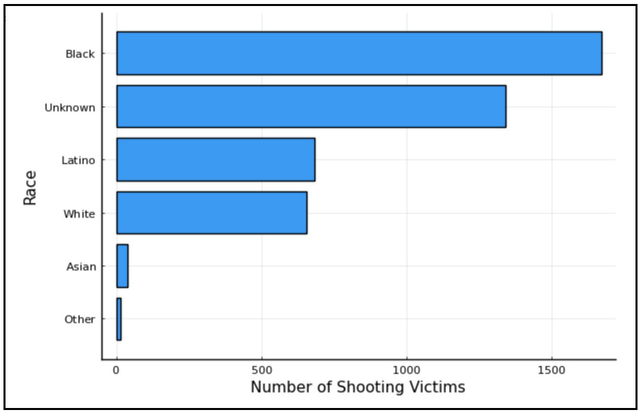
Descriptive Statistics of the Data and Histograms:

In our dataset, Chicago had by far the most police shootings per city with 12%. The majority of police shooting victims were black with 38%. 62% of all police shooting victims were fatal and only 38% of all police shooting victims are armed.

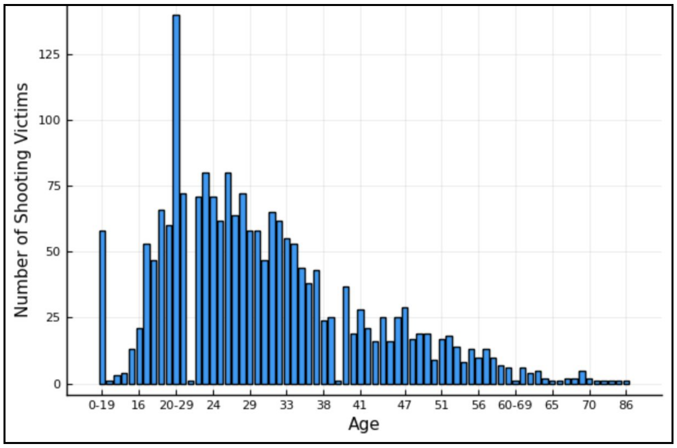
Police Shootings by State:



Race of Shooting Victims:



Age of Shooting Victims:



Features:

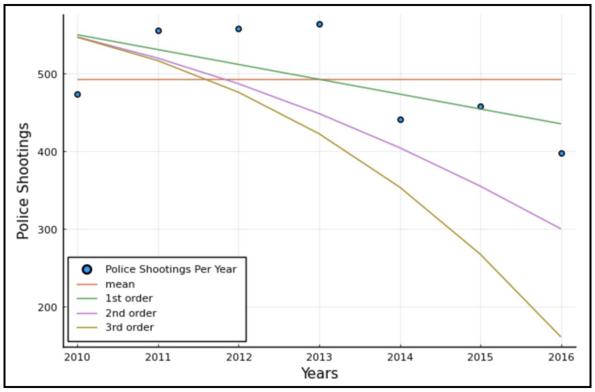
Our dataset consists of 4,400 police shooting victims from 2010-2016. The data has 16 columns that include descriptive data about the victims and information about the when and where the shootings occured.

Corrupted and Missing Data:

Many columns in our dataset had missing or incomplete entries. We know these columns contain missing data because they contain entries with “U” listed, which means the information is unable to be determined and “NA” listed, which means the data is missing. As an example for how we corrected this, we altered the SubjectGender feature to correctly count the genders in cases where there was more than one victim in a singular event (originally were strings with genders separated by semicolons) and change the “UNKNOWN”, “U”, “N/A” all to be of type missing in Julia. This missing information made it difficult to compute summary statistics for our data and thus we had to manipulate the data to find the descriptive statistics. The “NatureOfStop” and “NumberOf Shots” columns contained the most missing information leading us to eliminate the data in our models. Overall, once we deleted the columns that contained mostly missing information, our dataset became much easier to utilize for our preliminary analyses and future developments in our project.

Preliminary Analysis:

For our preliminary analysis, we created a model that shows the cumulative amount of shootings by year and determined the best fit using polynomial regression. Our goal is to investigate if there exists a relationship between the amount of shootings and the span of years and, if so, to explore the nature of the correlation with additional features.



We fit multiple polynomial models using one feature of the data (Year) in order to predict the future trends of police shootings past 2016 (*see plot above*). These models were fit using the least squares for 1st, 2nd, and 3rd order polynomials (as well as the overall mean of Police Shootings throughout 2010-2016). Overall we noticed that there was a gradual decline in shootings after the year 2013, but in the future we’d like to see if there are any other significant features such as race, gender, or the fatality of the shootings that can better predict different suspect scenarios. Furthermore, we will amass more data points by including the months (or breaking up the year into quarters), in order to get closer fits for higher order polynomials.

Further Steps:

The following work needs to be done for our final report:

- The input space needs to be normalized and standardized and data needs to be adjusted to be uniformly categorized (i.e.: age column needs to fix range values)
- We may create new columns to help with our forecasts
- We will test different error metrics
- We will continue to experiment with different models such as multi-layer perceptron and different loss functions
- We will include more features and represent shootings over monthly or yearly quarter increments in order to explore higher ordered polynomial fits