Capstone 2 Final Report

- The problem I want to solve for my second capstone project is to determine what factors related to location and age are leading to gun related incidents where injuries or death are occurring and based on that where laws and potential new training programs based on that information should be targeted at.
- My client is the entirety of the United States government, but this is something that affects people beyond that. With my analysis there will be a way of determining what locations are in the biggest need of targeting as far as testing stricter gun laws or better training programs. In addition we will be able to get a better idea of what the age should be for acquiring a gun.
- The data I'm using will be a list of gun incidents over the past few years in the United States. I have obtained the dataset from Kaggle and will continue to clean and adjust it to form a data frame that is optimal to work with for this problem
- The steps I will take to solve the problem as of right now are as follows:
    - Acquire and clean the data
    - Make various scatter plots to get ideas on trends
    - Run significance tests in order to confirm or deny suspicions
    - Create predictive models
- My deliverables for this project will be the code for the project, a write up on the project and a slide presentation.
- Who is my client?
    - My Client is the United States government.

    - My client will care because this information can potentially be used to reduce gun related deaths.
- What data am I using?
    - The data I am using is as follows:
        - Dates of gun related incidents
        - List of major cities where gun related incidents happened
        - Amount of deaths caused in related incidents
        - Age of shooters

Data Wrangling

- The original dataset was a very large csv file that included every recorded gun incident in the United States over the past 5 years. For this project there was far too much information so a fair bit of cleaning had to be done.
- The following steps were taken in order to clean the data.
    - First I determined what I wanted the data set to show.
        - I wanted to be able to look at gun incidents that lead to at least one death in larger cities.
    - The second step was to remove all the columns that I wasn't planning on using.
    - The third step was to drop all of the rows that contain null values.

- The fourth step was to get rid of any incident listed that didn’t result in at least one death.
- The fifth step was really tricky. For the fifth step my goal was to remove all rows representing incidents that happened in cities with less than 500,000 people. For this I separated this into substeps. I didn’t have the information about the cities within the data set so a fair bit of manually removing things was done.
    - The first sub step was to determine what states didn’t have any cities with 500,000 people or more and remove rows representing incidents that happened in those states.
    - Next I went in and manually removed the cities that were left with 500,000 people or less.
    - The final sub step was to make a bar graph comparing the amount of gun related deaths in each city and then see if I missed any small cities and then to remove those cities.

As I noted in the steps I removed any rows with missing values.

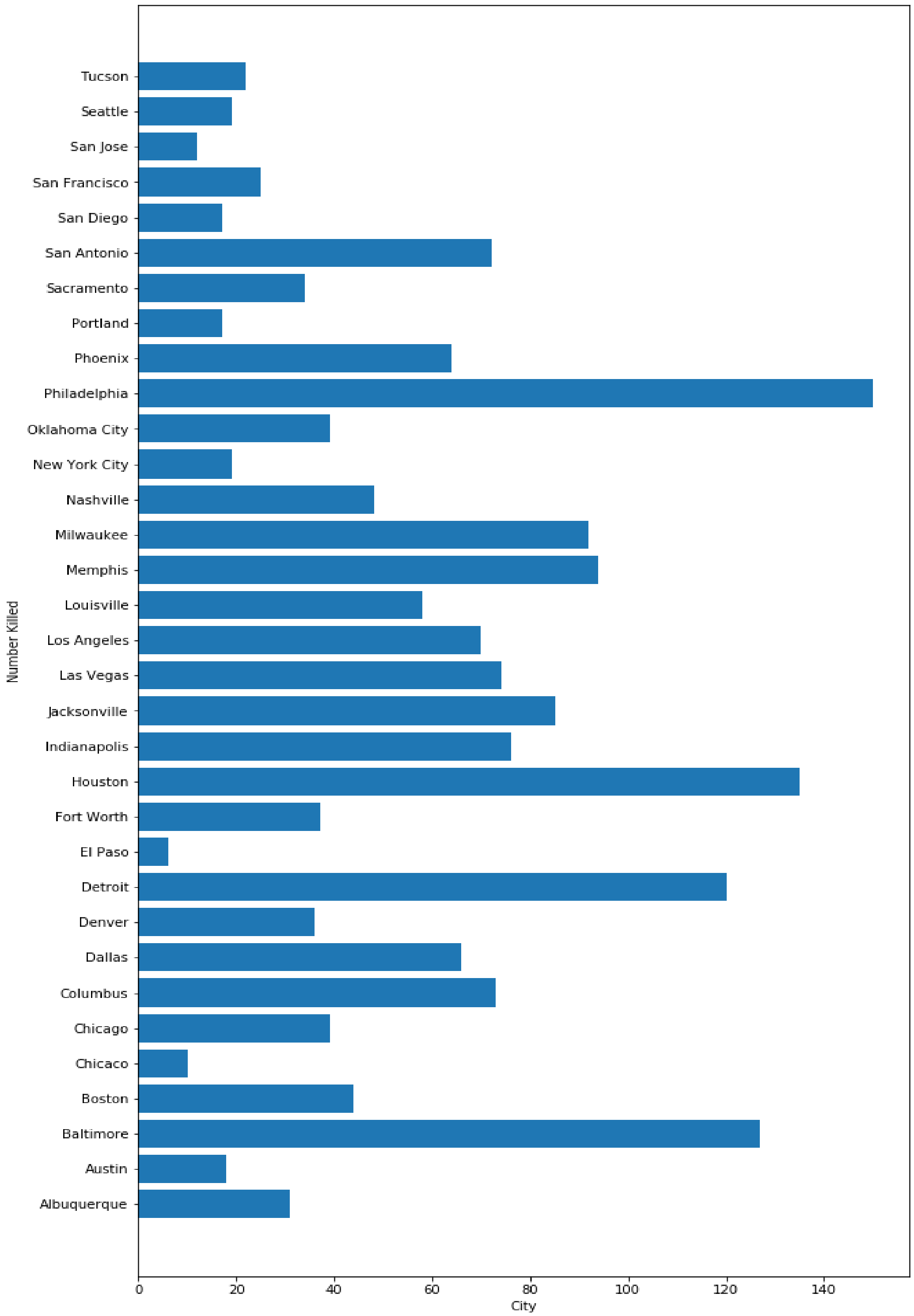
For this there were a few outliers which I removed depending on how much of an outlier they were. Mass shootings were removed from the data set. There weren’t any outliers on the lower end so that side wasn’t a problem.

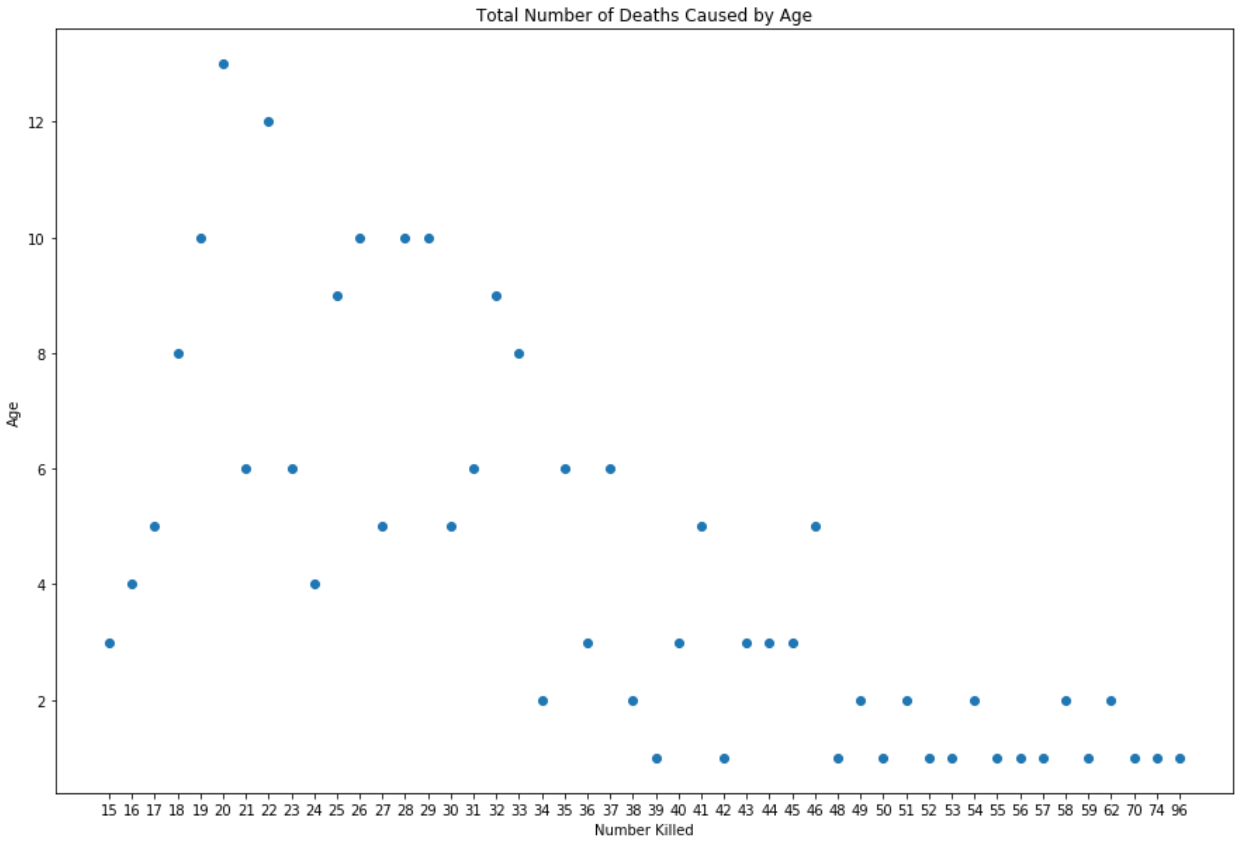
Exploratory Data Analysis

For my exploratory data analysis I made a couple graphs so far. Both are bar graphs comparing the amount of deaths caused by gun violence. The first graph shows a comparison of gun related deaths in major US cities. Then the second graphs shows a comparison of gun related deaths by age. My goal with these graphs is to be able to target certain locations and age groups for an analysis of gun control laws while also determining in which locations are they most needed.

- As far as the problem goes at the time of the second milestone it remains the same as before.
- As far as my target client goes at the time of the second milestone it remains the same as before.
- As far as the data I'm using goes at the time of the second milestone it remains the same as before.
- Now for the steps I've taken between the first and second milestone:
    - I've cleaned the data even more for use with the following steps:
        - I've renamed cities that were essentially the same place to one name each.
        - The steps I had previously taken to manually remove the cities has been turned into merely two lines of code at this time.
        - I've better separated the suspects and victims in order to get information about the shooters specifically.
    - I've visualized the data even further:

- First I created an updated bar graph comparing the total number of deaths attributed to gun violence in each of the large population United States cities. This was done using the following steps:
    - First I summed the number of gun related deaths in each city to set up the data.
    - Next I visualized the data
- Next I created a scatter plot comparing the total number of deaths attributed to gun violence in each age that caused those crimes. This was done using the following steps:
    - First I summed the number of gun related deaths caused by each age to set up the data.
    - Next I visualized the data
- Below are images of the visualizations:

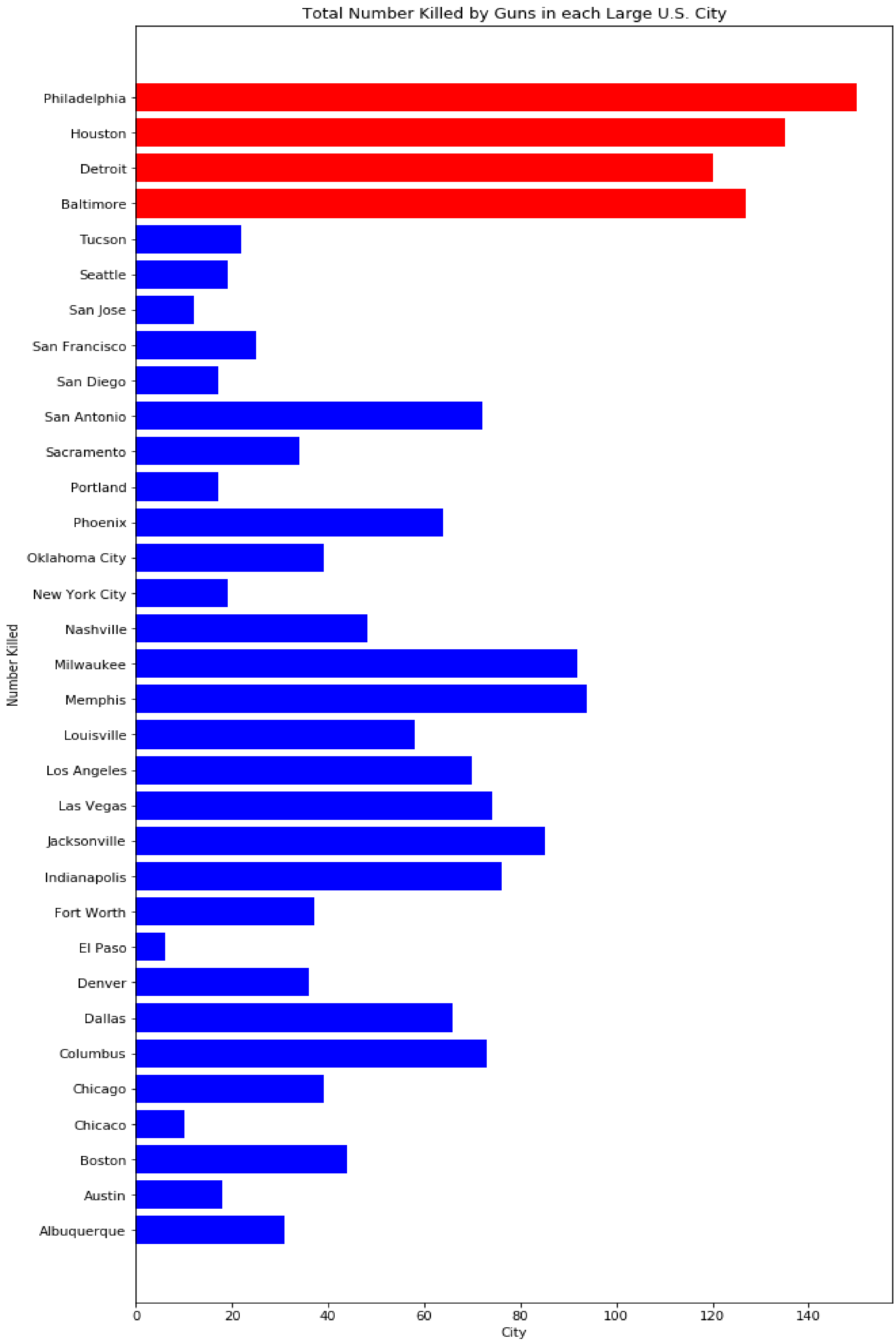

Tucson
Seattle
San Jose
San Francisco
San Diego
San Antonio
Sacramento
Portland
Phoenix
Philadelphia
Oklahoma City
New York City
Nashville
Milwaukee
Memphis
Louisville
Los Angeles
Las Vegas
Jacksonville
Indianapolis
Houston
Fort Worth
El Paso
Detroit
Denver
Dallas
Columbus
Chicago
Chicaco
Boston
Baltimore
Austin
Albuquerque
Number Killed
0
20
40
60
80
100
120
140
City

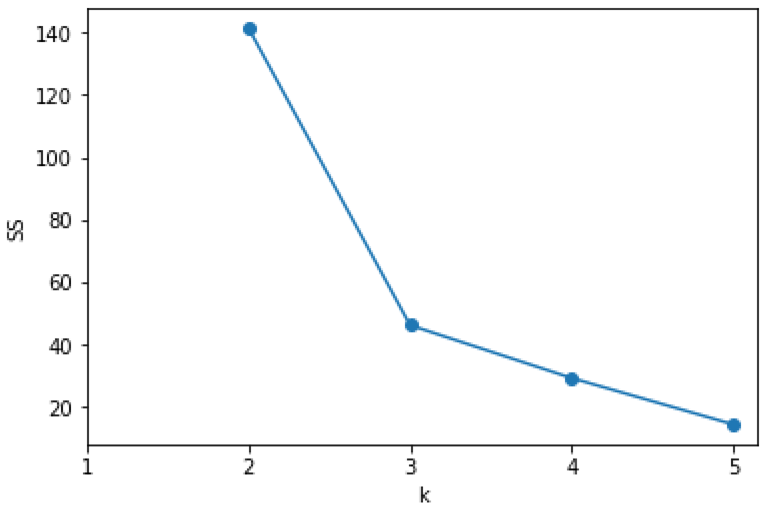


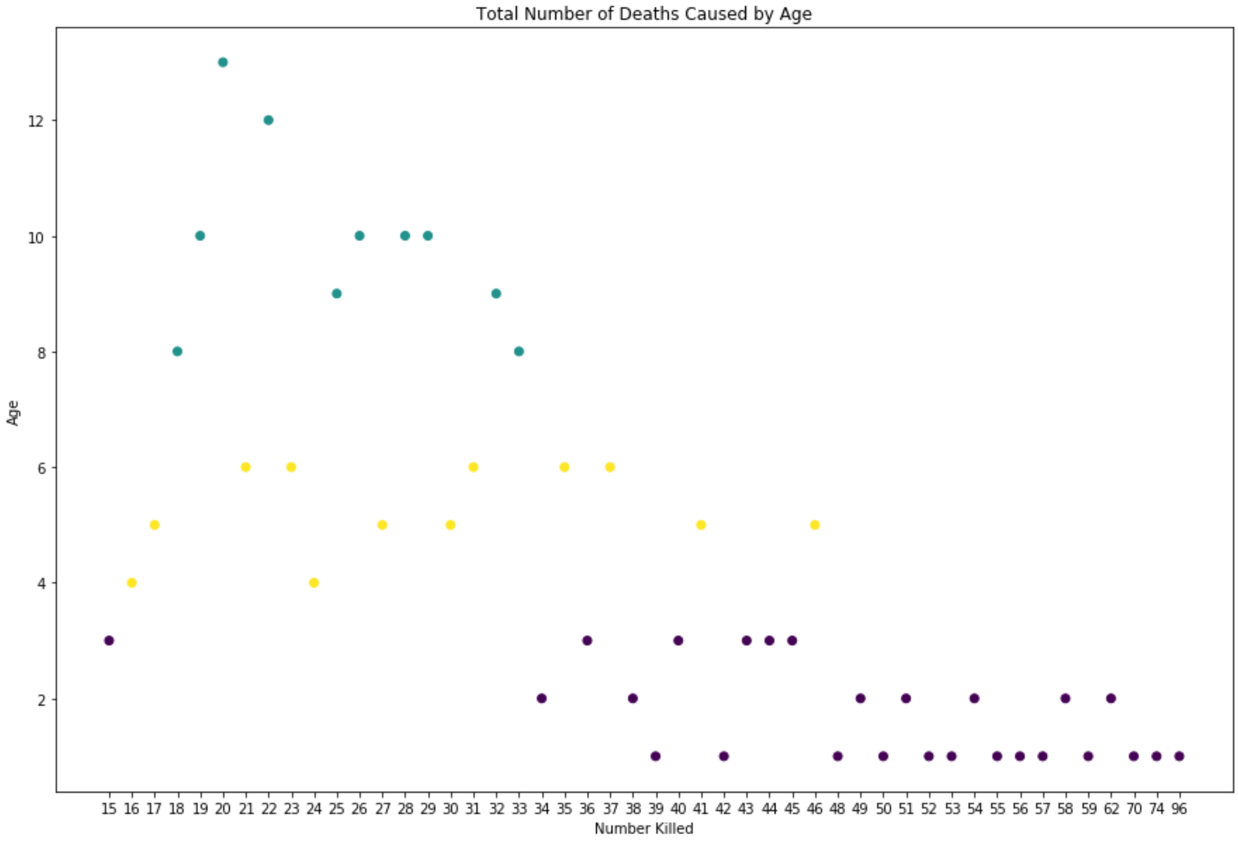
- The final steps I took as far as this project goes were continuing the visualizations.
- The first thing I did as far as final steps for visualizations went was update the bar graph to highlight the 4 cities that had 100 or gun related deaths be highlighted in red.
- The next thing I did was separate the previously created scatter plot into clusters:
  - For this part, first I use the elbow test to determine that the correct number of clusters was 3.

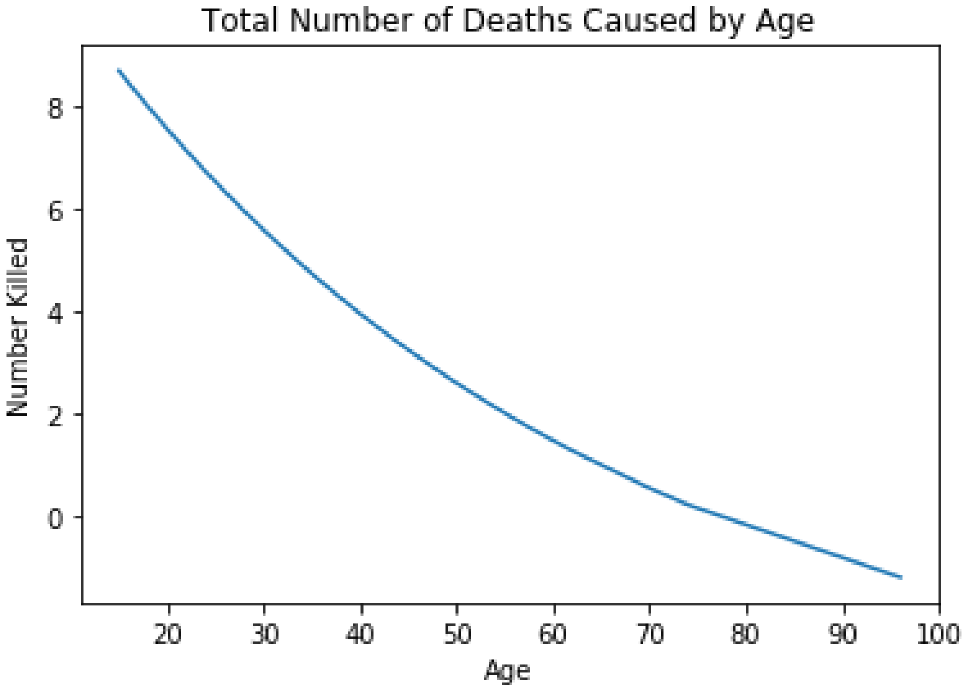
- Next I applied some new code to the scatter plot into 3 different clusters.

- Finally I created a regression to illustrate an exponential curve which I feel best estimates the relationship between age of the criminal and the total number killed by that age.
- All of these visualizations are shown below:


Total Number Killed by Guns in each Large U.S. City
Number Killed
Philadelphia
Houston
Detroit
Baltimore
Tucson
Seattle
San Jose
San Francisco
San Diego
San Antonio
Sacramento
Portland
Phoenix
Oklahoma City
New York City
Nashville
Milwaukee
Memphis
Louisville
Los Angeles
Las Vegas
Jacksonville
Indianapolis
Fort Worth
El Paso
Denver
Dallas
Columbus
Chicago
Chicaco
Boston
Austin
Albuquerque
0
20
40
60
80
100
120
140
City


SS
140
120
100
80
60
40
20
1
2
3
4
5
k


Total Number of Deaths Caused by Age
Age
12
10
8
6
4
2
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 48 49 50 51 52 53 54 55 56 57 58 59 62 70 74 96
Number Killed



So, now for the recommendation. Based on this data I would recommend that the United States start to implement stricter gun laws in the cities of Houston, Baltimore, Detroit, and Philadelphia. In addition based on the exponential curve and the clustering of the scatter plot. I would immediately look into increasing the age requirement for buying a gun to 21 based on the extremes as far gun related deaths attributed to an age that are seen in the ages of 19 and 20.