A Final Project Report on

US Mass Shootings Data: Exploratory Analytics

By: Team Member Name: **Sneha Godbole**

Email id: **snegodbo@umail.iu.edu**

**Table of Contents**

[**Abstract** 3](#_30j0zll)

[**1.**](#_1fob9te) **Introduction** 4

[**Motivation** 4](#_3znysh7)

[**Existing Work** 4](#_2et92p0)

[**Contribution** 4](#_tyjcwt)

[**2.**](#_3dy6vkm) **Data & Methods** 5

[**Data Introduction: US Mass Shootings (50 Years)** 5](#_1t3h5sf)

[**Dataset** 5](#_4d34og8)

[**Data statistics** 5](#_2s8eyo1)

[**Data Limitations** 5](#_17dp8vu)

[**3.**](#_3rdcrjn) **Implementation Approaches & Techniques** 6

[**Data Preprocessing** 6](#_26in1rg)

[**Methods used for Exploratory Analytics / Insights** 6](#_lnxbz9)

[**4.**](#_35nkun2) **Results** 8

[**Novelty** 8](#_1ksv4uv)

[**Insights & Discussions** 9](#_44sinio)

[**Code Base Details** 13](#_3j2qqm3)

[**5.**](#_1y810tw) **Conclusion & Future Work** 13

[**6.**](#_2xcytpi) **REFERENCES** 13

# **Abstract**

The phenomenon of mass shootings has emerged in US over the past 50 years. A high proportion of rampage shootings have occurred in the United States, and secondarily, in European nations with otherwise low firearm homicide rates; yet, paradoxically, shooting massacres are not prominent in the Latin American nations with the highest firearm homicide rates in the world. A review of the scientific literature from 2010 to early 2014 reveals that, at the individual level, mental health effects include psychological distress and clinically significant elevations in posttraumatic stress, depression, and anxiety symptoms in relation to the degree of physical exposure and social proximity to the shooting incident [[3]](#_1ci93xb). Given the historical dataset of the 50 years, I am proposing an in-depth exploratory analysis of the mass shootings in US over this period to find correlations with several important factors (example: age, sex, race, calendar date, etc.), analysis of the on the location, impact, killings, motives of the incident. I am also looking forward to finding an unknown pattern based on the dataset visualization which can be used as an informative analysis for future reference in the same work area.

# **Introduction**

## **Motivation**

The US has witnessed 398 mass shootings in last 50 years that resulted in 1996 deaths and 2488 injured [[2]](#_1ci93xb). The latest and the worst mass shooting of October 2, 2017 killed 58 and injured 515 so far. As per some researchers in this area, the number of people injured in this attack is more than the number of people injured in all mass shootings of 2015 and 2016 combined! When I saw this news on the TV the other day, it disturbed me so much that I started thinking about why would somebody with no criminal background do that? As of today, police are still investigating the motive of the recent Vegas mass shootings, and have not published even a clue on why the shooter did that! We discussed this with at workplace for days after this and we started guessing mental unstability, personal life reasons, depression, psycho-path mindset or an influential incident by a terror organization, etc. could cause a normal person to get into such mindset over the period of years. A more detailed analysis on the same topic out of curiosity informed me the mass shootings have been a big challenge / problem the American government is facing and this has been happening in the US from 50+ years.

This is unfortunately the truth of Modern American history. I could not stop myself in digging in more on this topic which led me to 50+ years mass shooting dataset. As this is one of the biggest challenge in modern American history which would impact the next generation Americans heavily in near future, it gave me motivation to choose this topic for the exploratory analysis project, to find interesting insights and patterns about the historical shootings happened in the US.

## **Existing Work**

I found many research papers and literature reviews on the mass shootings area, some of them are present at [Research Papers Link](https://www.ncbi.nlm.nih.gov/pubmed/?term=mass+shootings) . Many of the papers do analysis of mental illness, American firearms policies, the effects of shooting rampages on children and adolescents, the psychological effects for wounded victims and the emergency healthcare personnel who care for them, the disaster behavioral health considerations for preparedness and response, and the media "framing" of mass shooting incidents in relation to the portrayal of mental health themes [[3]](#_1ci93xb).

## **Contribution**

I did study the paper I used for implementation, along with the other papers referred in the paper to understand what research and insights has been proposed by the authors. I studied the mass shootings dataset and came up with the scope of my project work.

I could successfully use and analyze and verify the research work insights mentioned in the referred paper, the details about this work are present in [implementation approaches and techniques section](#_3whwml4).

I also proposed and designed few **additional insights** to be implemented on the given dataset, the details are present in [results](#_2bn6wsx) section.

# **Data & Methods**

## **Data Introduction: US Mass Shootings (50 Years)**

I found the dataset containing information on last 50 years of US Mass shootings incidences which I would use for this work:

* Geography: United States of America
* Time: 1966-2017
* Unit of analysis: Mass Shooting Attack
* Dataset: The dataset contains detailed information of 398 mass shootings in the United States of America that killed 1996 and injured 2488 people.
* Variables: The dataset contains Serial No, Title, Location, Date, Summary, Fatalities, Injured, Total Victims, Mental Health Issue, Race, Gender, and Latitude and Longitude information.

## **Dataset**

The dataset is available at [Dataset Link](https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years/data). I did verify and used this dataset in the implementation and the dataset looks very good and informative.

## **Data statistics**

The dataset contains mass shootings information data for last 50 years, containing ~323 rows representing unique incidents, each described with 21 metadata columns. The metadata fields contain very important incident information such as Location, Date, Area, Target, Cause, Summary, Suspect information, Victim's information, etc.

## **Data Limitations**

Some of the incidents mentioned in the dataset (meaning, the rows in the dataset) does not have values (example: Longitude, Latitude), and some of the metadata columns (example : Race, Gender) do not have standardized values. Some of the column’s(example: Age) values were missing so couldn't use that column for my analysis and visualization purpose. I did cleanse and standardize these values during implementation.

# **Implementation Approaches & Techniques**

## **Data Preprocessing**

1. As a requirement of doing exploratory analytics by year / month / date level, I standardized all the time stamp values in the given dataset. This was achieved by implementing a custom function and using it on entire dataset.
2. I used step #1 to populate one extra columns namely “year”,”city”,”dow” for each row in a given dataset.
3. Result of steps #1 and #2 was stored in the form of Data Frame and was used for all further analytics / insights exploration
4. I did cleansing and standardization of the **Race** and **Location,Gender** columns, to implement the exploratory analysis for some use cases.

## **Methods used for Exploratory Analytics / Insights**

I have validated few of the insights implementations published at the [Link](https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years). The exploratory analysis implementation led to interesting insights which are described in the results section below in the document.

I have used panda library which is very helpful for calculations like grouping,counting etc.

Following meaningful insights were implemented and visualized on the dataset –

1. **Number of Fatalities vs Injuries per Year**

Analysis and visualization of fatalities vs injuries over all the years preset in the dataset.

**Approach**:

I analyzed the dataset, aggregated #fatalities and #injuries for all the years available in the dataset and plotted this insight using a multi bar chart for all the years on Log Scale. This shows an insight on how the shootings have impacted on deaths and injuries.

1. **Total victims per year**

Analysis and visualization of total victims per year, for all the years preset in the dataset.

**Approach**:

I analyzed the dataset, aggregated #Total victims of all the attacks for all the years present in the dataset and visualized the insight using **scatterplot**.

1. **No of attacks by races**

Analysis and visualization of number of attacks by races, for all the years preset in the dataset.

**Approach**:

I analyzed the dataset and calculated the count of number of attacks by #Races for all the years present in the dataset and visualized the insight using simple **Bar Chart**.

1. **Number of attacks per year**

Analysis and visualization of number of attacks per year, for all the years preset in the dataset.

**Approach**:

I have created new column in the DataFrame called “year” from original “Date” column so that I can easily count the number of attacks per year values. I haveand visualized the insight using simple **Line Chart**.

1. **Causes of attack**

Analysis and visualization of different causes behind the attack, for all the years preset in the dataset.

**Approach**:

I have calculated causes of attack by grouping the the attacks by “Cause”. I haveand visualized the insight using simple **Bar Chart**.

1. **Overall No. of Attacks per Weekdays**

Analysis and visualization of attacks happened over weekends vs weekdays, for all the years preset in the dataset.

**Approach**:

From the “Date” column from dataset, I have calculated the number of attacks over the weekend by using pandas DateTime property “DayofWeek”.

For example:

weekday = date.dt.dayofweek / 5 == 1.

I haveand visualized the insight using **Pie Chart**.

1. **Shooting Fatalities/Injuries by Latitude/Longitude in United States**

I have used latitude/Longitude data from dataset, and calculated each bubbles radius depending upon number of fatalities. When you hover over on each bubble you can see Number of Fatalities/Injured people and the location where this attack took place.

1. **Top 10 Cities of mass shooting**

I have calculated city name from location column and grouped them in descending order. Then took only first 10 records to show top 10 cities.

# **Results**

## **Novelty**

I I have validated the insights from research paper and proposed, designed and implemented **following additional use cases** on the same datasets as follows –

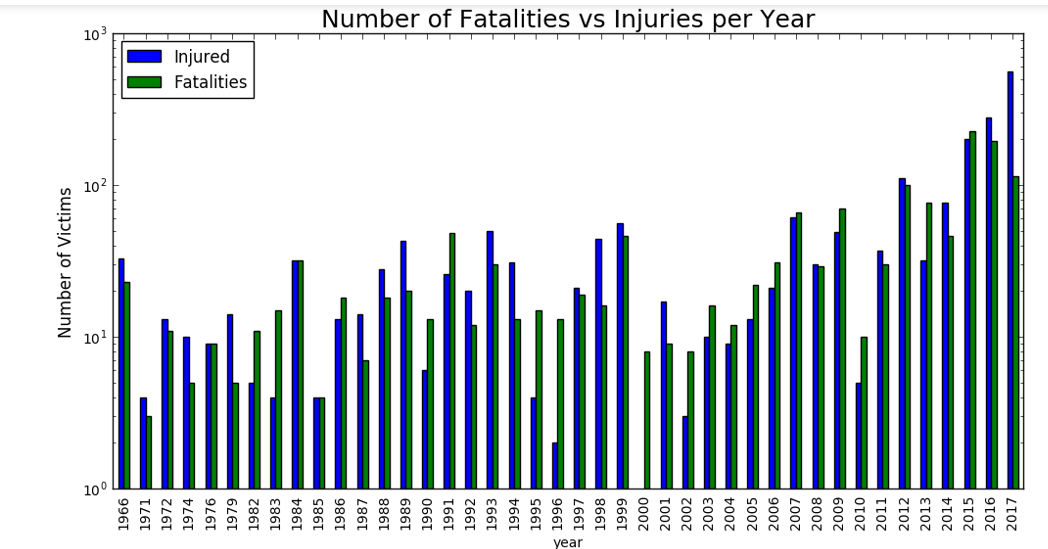
1. Show number of fatalities/Injuries using bubble map.
2. Causes of the attack
3. Number of attacks on weekend vs weekday on pie chart.

I found **interesting insights** and patterns on all these **additional use cases** and thedetails are present in section below.

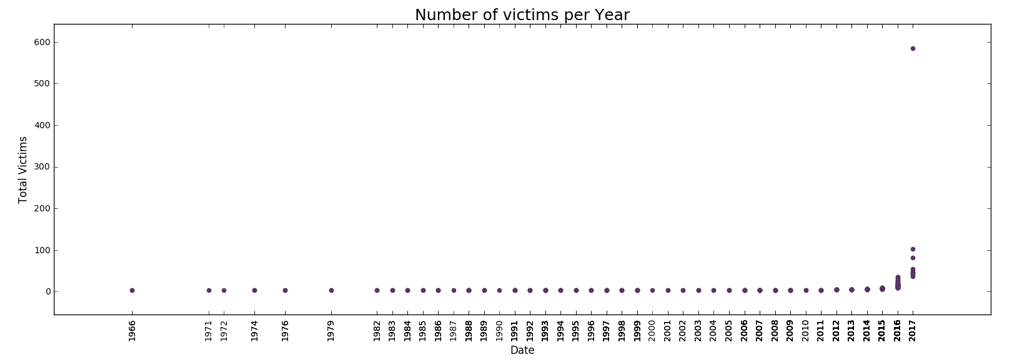
## **Insights & Discussions**

In this section, we will go over all the exploratory analytics / insights I have found after implementation –

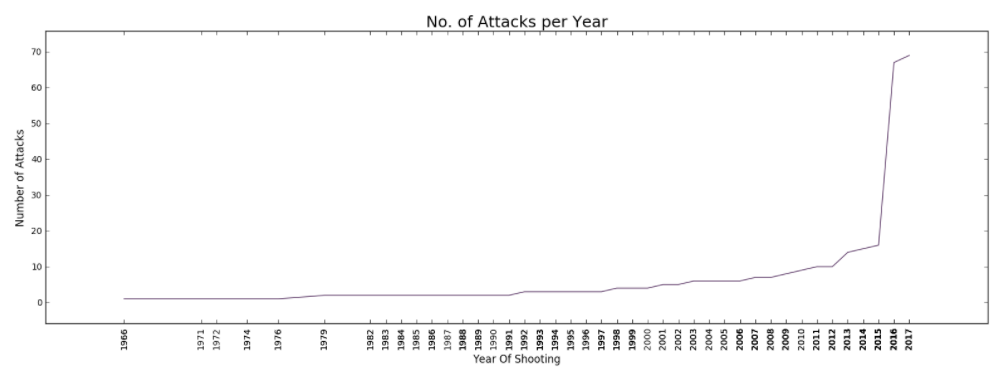
1. In 2017, highest number of people injured from the attack’s (558). In 2015, highest number of people died in the attack (226).



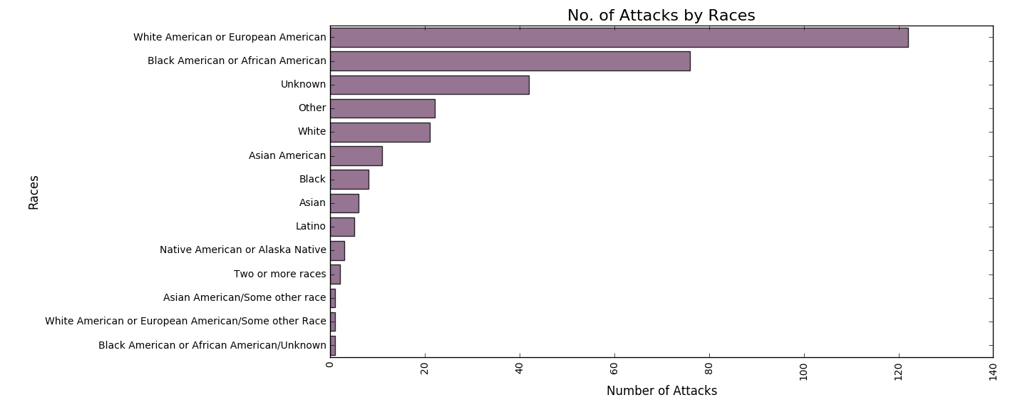
1. The scatterplot below shows how the rate of victims per year increased over the years. This rate has increased drastically from 2016.

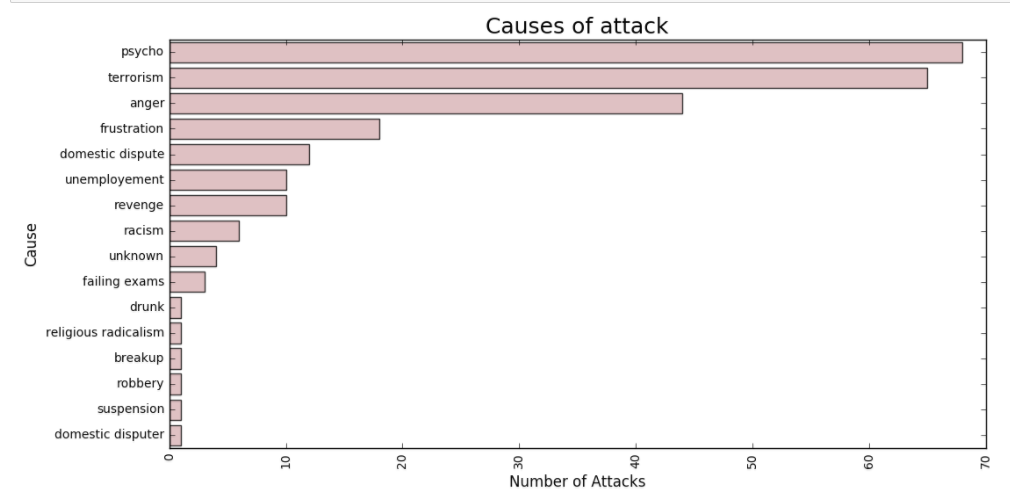


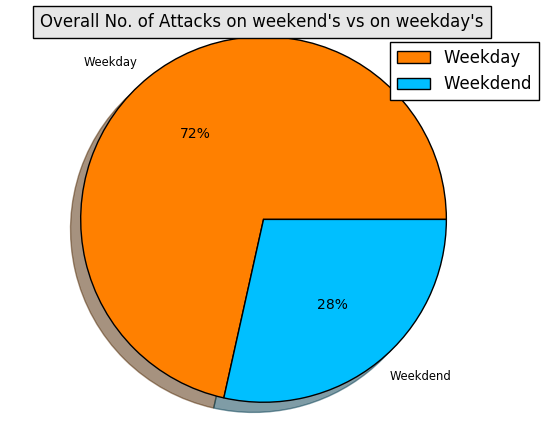
1. The line chart below clearly represents the rate of number of attacks per year has spiked n 2016. The rate has increased 4 times in 2016 than 2015.



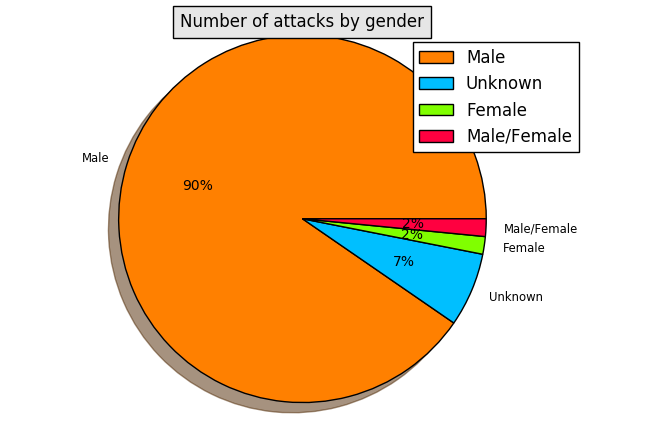
1. From the bar chart below, we can see that White American or European American Race is way high than others. If we combine White with it, it will become twice higher than Black American or African American Race.



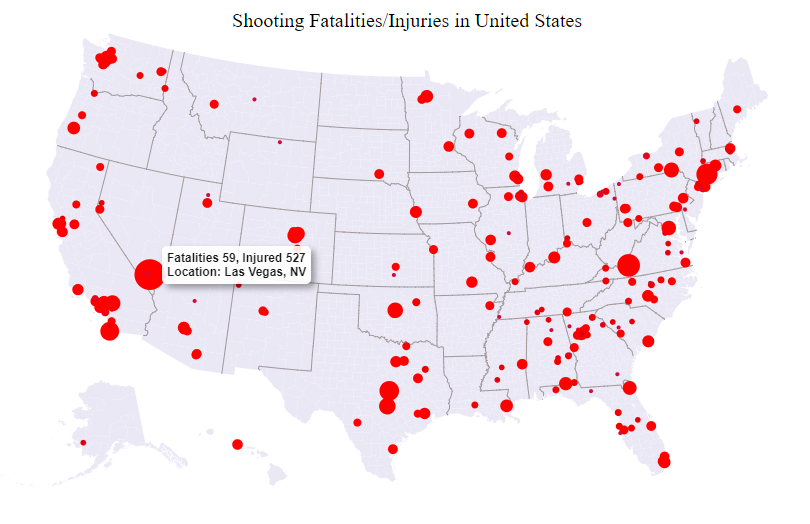
1. Following insight seems very interesting to me because from this chart you can easily realize that the most attacks are caused by mental health reasons. If we add psyco, anger, frustration then it's almost double than terrorism. So from here we learn that if person finds solution on their mental health problems then there would be less attacks.
2. Following pie chart shows that number of shooting 2.5 times high on weekday as compared to the weekends.

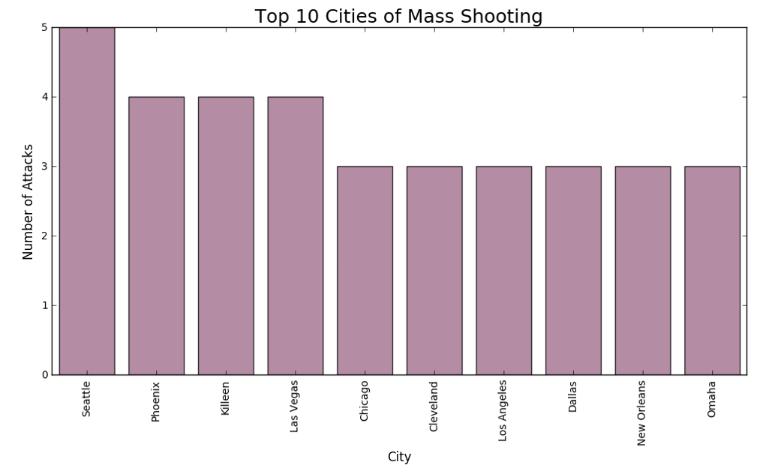


1. A pie chart below shows 90% shooters are male.



1. Following map shows number of Fatalities/Injuries happened at different locations in US. Each bubble represents the scale of Fatality. We can observe the largest mass shooting in terms of fatalities and injuries is Vegas, in 2017.



1. A chart below shows top 10 cities where most of the attacks happened. From this chart it shows that Seattle city had more number of attacks.

## **Code Base Details**

The code base for this implementation is available [here](https://github.com/snehagodbole/US-Mass-Shootings-Data-Exploratory-Analytics.git)

# **Conclusion & Future Work**

* I could successfully implemented, and verified most of the research work, insights and literature survey published in the research [paper](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318286/) and [dataset](https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years/data) provided.

* The implementation led to interesting exploratory analysis and I got many meaningful insights from the US mass shootings data. I found the overall end to end experience very exciting and interesting.
* Future work item could be to extend these explorations and techniques to other similar datasets for other countries / areas and publish the results, correlate them wherever applicable.

# **REFERENCES**

[1]. Mental Illness, Mass Shootings, and the Politics of American Firearms: [Link](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318286/)

[2]. <https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years>

[3]. <https://www.ncbi.nlm.nih.gov/pubmed/25085235>