MSc Data Science & Analytics Thesis

Predictive analysis on United States of America Gun violence Dataset involving Socioeconomic parameters

Author

Pradeep Gurunathan

Student Number-19251698

Supervisor

Dr. Peter Mooney

A thesis submitted in fulfilment of the requirements for the degree of MSc in Data Science and Analytics 2019-2020

in the

Department of Computer Science
Maynooth University

August 02,2020

Declaration

With the permission of our supervisor (Dr. Peter Mooney) and Head of Department (Dr. Joseph Timoney) I Pradeep Gurunathan (Student number- 19251698) worked as part of the two-person team with Karthick Pandi (Student number- 19250898). There will be some overlap in the data extraction and processing part of this work. However, both myself and Karthick Pandi worked independently to write our thesis report. Dr. Mooney has read the drafts of our thesis reports to ensure their authenticity.

Signed: Pradeep Gurunathan

Date- 02-08-2020

Abstract

MSc in Data Science and Analytics

Predictive analysis on United States of America Gun violence Dataset involving Socioeconomic parameters

By – Pradeep Gurunathan

Gun Violence is a very serious issue faced by many countries around the world. It is very common to hear every day in the news media about a shooting incident in the USA, even though the United States has the 28th highest rate of deaths related to gun violence in the world in 2017. The US. has seen a rise in gun-related violence and it has the highest rate of murder by firearm among the developed nations. In this report, a detailed analysis is performed on the United states Gun violence data collected from 2013 to 2018 in order to gain a better understanding of how dangerous the gun culture is within the United States. Detailed analysis, coupled with information on age group, gender, state, location, etc. are used to understand more about US gun violence. Extra parameters such as population, per-capita and un-employment rate in each U.S. state are used to predict which state is most dangerous and safest from gun violence. This inclusion of these socioeconomic data will give a clearer picture of the seriousness of the gun violence problem. Apart from this prediction this report also attempts to predict which month of the year and day of the week is dangerous or safer for citizens. Finally, a novel risk score is generated by giving proper weightings to attributes in order to predict which month of the year, day of the week and which U.S. State is most safe and dangerous from gun related violence.

Contents

1.Introduction

A brief introduction of the History and current scenario of United States Gun violence and detailed outline on United States Population, Unemployment and Per-Capita income. This chapter also includes Motivation of the report and objectives that was achieved in this report.

1.1 United States Gun Violence

Prelude

According to an article by renowned American Activist, writer and historian *Roxanne Dunbar-Ortiz* history of Gun culture and its origin in the wars against indigenous people. Article tells us about secret history about Second Amendment of United states constitution in 1791. [1] The United States, she writes, "was founded on conquered land, with capital in the form of slaves, hence the term chattel slavery. This was exceptional in the world and has remained exceptional. The capitalist firearms industry was among the first successful modern corporations. Gun proliferation and gun violence today are among its legacies.". She argues that "understanding the second amendment is key to understand the gun culture in United States". After independence from Britain the newly formed United States established the individual rights for having firearms so that independent militias and settlers to do dirty work of invading indigenous nations or "Taking land by force" from the indigenous communities. She has also disproved a myth like "right to bear arms" has to do with protecting right to hunt, she says that second amendment has nothing to with hunting. Gun industries has peddled its products to promote the Gun culture by supported ideologies like "Gun Rights".

Current Situation

Gun culture has a deeper history roots back to American continental conquest, slavery, second amendment which gave right for arms and later as the leading military power in the world. U.S. is the world's leading buyer and exporter of military weapons, there is a huge power of gun lobby to keep the weapons industry to be highly profitable in United States and keeping the ideology "Gun rights" possession of it is still alive.

There are around 393 million in circulation in the United States that is approximately 120.5 guns for every 100 people [2]. Around 58% of adult experience the gun related violence and around 3 million children witness the gun violence each year according to CDC [3]. The people they died from accident shooting are three times to have firearms in their home than in the controlled environment. Most of the death related to children are due to playing with preloaded guns in absence of their parents. Suicide rates are higher with the higher rates of gun ownership. Reason of Suicide are related to poverty, urbanization, unemployment, mental illness and alcohol or drug abuse. Assault style weapons are responsible for minority of the gun incidents in U.S., but it is the weapon of choice for the Mass shooting which can cause chaos and high fatalities. There is a high risk in the homicide eight-fold for women physically abusive relationship and this domestic violence are more likely to end up deadly with gun at home.

In the current situation even though there is strong voice for "Prohibition of Guns" and some states have stricter laws for sale of firearms and possession of firearms. There is alarming increase in firearm homicide, firearm suicide and public mass shooting in United states, according to 2017 United states ranked 28th for the gun related violence and it is highest among the developed countries. Un-employment, poverty, mental illness, suicide rate, Gun availability are the main driving forces for increase in Gun culture in United states.

1.2 United State Population, Un-Employment and Per-Capita Income.

United State Population

United states are well known for its diverse country with different ethnicities. It has high number of immigrants throughout the country's history, with many people coming for a better life and opportunities. With this search for opportunities it known as "American Dream", it has also become "melting pot" of different culture, nationalities and traditions. United states have third most population with 327.2 million in 2018 and its population is only expected to increase. It is estimated to see a steady growth over the next 40 years and by 2060 the population is estimated to be 400 million people [4].

Many people are moving towards big cities for better life, few such state is California, Illinois, Texas and Florida. Its every come to hear Gun violence or a mass shooting in this state's. Multiple factors such as large population, high number of immigrants each year, also with urbanization, high competition for jobs, Un-employment, poverty and on top of its easy access to firearms will only increase the Gun related violence.

United State Un-Employment and Per-Capita Income.

In a recent study by Daniel Kim, an associate professor of health science at Northeastern provides new insight into how socio-economic circumstances could be driving gun violence in the U.S. On average around 100 people are killed with guns every day. Between 2001 to 2013, gun violence incidents took more lives than the combined number of US people killed in war, AIDS, drug abuse and terrorism. This epidemic proportions are mainly due the roots in socio-economic causes, since there is high number of un-employment due to urbanization, less wages and high cost of living this has direct relation to the Gun violence in USA. After the several massing shooting incidents studies and investigation shows that many of it is due to the economic downtrend seen by the individual who caused the incidents.

1.3 Motivation and Objectives

Motivation

Gun violence has reached alarming proportion as much as 100 people have died every day. Total victim count is more than the total lives lost due to wars, AIDS, drug abuse and terrorism put together. There has been a strong support for gun control in recent days, with 393 million gun in circulation and 120 gun for every 100 people it very difficult to see any reduction anytime soon, it is every important to understand the how Un-employment rate and per-capita income has influence on an individual or a family life, so it is important to understand these factors and find if there is any direct impact and facts related to the high Gun violence incident in certain U.S. States.

Objectives of this Report

Three main objectives of this reports are:

- To Clean and integrate the data from different sources
- To perform machine learning techniques on the data,
- To generate some statistical analysis and risk score ranking on Gun violence dataset involving socio-economic data.

This report focusses on the relationship between the Gun Violence and Socio-Economic data, socio-economic data includes Un-employment and Per-capita income for each U.S. States. In this report there will be a 3D Geo-Visualization of the U.S. Gun violence Data from 2013 to 2018 using Deck-GL and Google Map API, an heatmap is generated to visualize which part of the U.S. Geographical region is highly or less affected by Gun violence and along with heatmap an hexagonal 3D projectile is generated to see the intensity of the Gun violence in certain areas.

First, we will be focusing on the main dataset which includes the gun violence data from 2013 to 2018(*Source – Kaggle*), first and for most important part of this report is data cleaning, data modification that helps with our data analysis and prediction. After the data cleaning and data pre-processing some of the preliminary analysis on the main dataset has been done to answer questions such as which Age, which Gender and is there any correlation between the incidents and victims. Next most important part of the report is use of machine learning methods to answer which day of week, which month of year, which State and which City or county is most safe or dangerous from Gun related violence. Even a risk score ranking system is generated by giving proper weightage to the parameter to rank month, day and States to answer some of the questions.

Just with the Gun violence data we cannot clearly say which State is more safe or dangerous, because it is very common to hear a mass shooting incident in California, Florida, Illinois and Texas but the most dangerous are not these states. To answer this question which state is more dangerous we are considering another dataset which includes Population, Un-Employment and Per-capita income. Finally, a risk score ranking system generated including this extra parameter to get more accurate prediction of which U.S. State is Dangerous and how the socio-economic factor plays a curial role in Gun culture in United States of America.

2. Tools Used for Analysis

2.1 Why Python?

Three main factors why python was used for the Data Analysis: -

- It is easy and flexible.
- It is accepted widely in industries and most popular language of Data enthusiast in the industries.
- It has very vast variety of python libraries for data science.

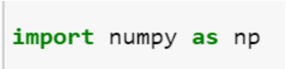
Advantages of Python over R?

[4] We preferred Python over R because unlike Python, R is not a general-purpose programming language. It focusses exclusively on statistical computing and data analysis. But Python acts as a general-purpose programming language since its syntax rules enable developers to build applications with concise and readable code base. Hence many programmers preferred Python over R. Secondly, Python is bundles with several useful packages that we needed for the data analysis like NumPy, Pandas, Seaborn, Matplotlib etc. The main aspect of preferring python over R is speed of the compilation. Several studies suggest that Python is faster than several widely used programming languages. Without investing urge time and effort, the beginners start explores a way to learn a robust programming language for data analysis. Python enables programmers to express concepts without any additional codes with its simple syntax rules. On the other hand, the steep learning curve of R requires beginners to put extra time and efforts. It seems very difficult to learn R if the person doesn't have any prior programming knowledge.

We are all informed of using pip, easy install and virtualenv if gets involved in Python world for too long time and even after using all these tools for installation still we haven't able to reach our requirements. The most important problem with these libraries is that these libraries focus entirely around Python, ignoring other non-python libraries dependencies such as HDF5, MKL, LLVM etc. which doesn't contains setup.py file in their source code and also do not install files into Python's site packages directory. Here comes Conda a packaging tool and installer that aims to do more than what pip, easy install or virtualenv doing currently. Conda successfully handles the library dependencies which is available outside the Python packages and handles the Python packages by themselves. Another important feature of Conda is it also creates a virtual environment like how virualenv creating currently.

2.2 Python Libraries Used

NumPy:

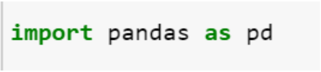


```
import numpy as np
```

[5] Numerical Python simply called as NumPy. A core library for scientific calculations which consist of strong n-dimensional array object. NumPy is written in C and executes rapidly accordingly. By correlation, Python is a powerful language that is deciphered by the CPython translator, changed over to bytecode, and executed. It is greatly used in computations while working with linear algebra, Random number capability etc. For generic data this NumPy library is used as an effective multi – dimensional container. There are many libraries that use NumPy, though a few are usually bundled

with it: SciPy, MatPlotLib, pandas, sympy and nose. NumPy and SciPy are two sides of a coin. Historically, NumPy was formed from two packages, so it contains not just the ndarray type and array manipulation functions but the numeric functions, as well. This array is in the form of rows and columns. We are using this NumPy array instead of using list in python in order to decrease the memory use, improve the speed and for better convenience.

Pandas:

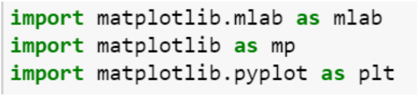


```
import pandas as pd
```

Another important toolkit which is greatly used for data analysis while we are trying to do with Python is Pandas. It has n number of usages ranging from parsing multiple file formats to NumPy matrix array which gets converted from the entire data table. That's why this Pandas is considered as a trusted ally in both Machine Learning as well as Data Analytics. Pandas incorporated with some quick, flexible and expressive data structures intended to make working with "relational" or "labeled" data both easy and intuitive. It's a high-level level building block for performing practical, real world data analysis with python. Here are some of the things why we prefer to use pandas:

1. In real world dataset we might encounter lot of missing data (usually represented as NaN) in the dataset in order to fix this and to proceed further with analyzing the dataset we need to handle this missing values, Pandas does this with good performance for both floating point as well as non-floating point data.
2. The second thing is Size Mutability which means columns can be inserted and gets deleted from the data frame and higher dimensional objects.
3. Data Alignment which gets done automatically as well as explicitly, an important feature of using pandas. Here objects are getting aligned explicitly to a set of labels.

Matplotlib:



```
import matplotlib.mlab as mlab
import matplotlib as mp
import matplotlib.pyplot as plt
```

A complete library for creating static, animated, interactive visualization which performs an essential role in Data Analysis with Python is Matplotlib. These Matplotlib are used for plotting two-dimensional plotting of the arrays. It's a multi-platform data visualization library build on top of NumPy arrays and intended to work with the wider SciPy stack. In other words, we can say it as numerical mathematics extension of NumPy. Pyplot is a Matplotlib module which provides a MATLAB like interface. In general, Matplotlib is created to be as functional as MATLAB, with the ability to use python, and it has the benefit of being free and even as an open source. The main advantage of using Matplotlib is it lets us to visualize huge amount of data in an effortlessly digestible manner. This one encloses variety of plots such as lines chart, bar chart, scatter plot, histogram etc. Publication quality figures in a range of hardcopy formats are produced with the help of this Matplotlib visualization library. These Matplotlib can be used in many areas such as Python Scripts, the Python and IPython Shell, Web application server, as well as various graphical user interface toolkits.

Seaborn:

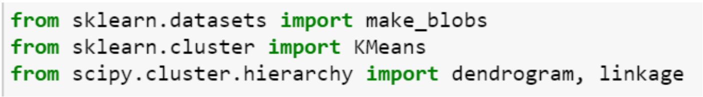
```
import seaborn as sns
```

Seaborn is a library for creating scientific plots in Python. It's based on head of matplotlib and firmly incorporated with pandas info structures. Now is a section of the practicality that seaborn poses:

1. Dataset-arranged API for looking at connections among various factors
2. Help for using all out factors to show opinions or total measurements
3. Choices for picturing univariate or bivariate conveyances as well as for looking at them among subsets of information
4. Programmed assessment and plotting of direct relapse models for several types subordinate factors
5. Helpful viewpoints on the general structure of complicated datasets level deliberations for establishing multi-plot networks that let you efficiently fabricate complicated representations
6. Brief command over matplotlib figure styling along with a few inherent subjects

Seaborn aims to make representation a focal piece of exploring and obtaining information. Its dataset-situated plotting capacities work on data frames and exhibits including entire datasets and inside play out the important semantic planning and measurable conglomeration to create informative plots.

Scikit-Learn:



```
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
from scipy.cluster.hierarchy import dendrogram, linkage
```

Scikit-learn is one of the open machine learning libraries for Python. It has several algorithms like kmeans, support vector machine, Linkage clustering, random forests, and k-neighbors, and numerical and scientific libraries like NumPy and SciPy is also support by this.

It provides various supervised and unsupervised learning algorithms with interface in python. It is built up on the SciPy (Scientific Python) this must be installed before Scikit-Learn. As a extension SciPy is Scikit-Learn which includes all the machine learning algorithms.

Several popular groups of models offered by Scikit-learn involve:

- Feature selection: used for finding significant attributes from which to create supervised models.
- Ensemble methods: used for merging the predictions of multiple supervised models.
- Manifold Learning: Used For summarizing and depicting complex multi-dimensional data.
- Supervised Models: a huge array not limited to generalized linear models, discriminate analysis, naive bayes, lazy methods, neural networks, support vector machines and decision trees.
- Cross Validation: used for estimating the performance of supervised models on unobserved data.
- Datasets: used for test datasets and for generating datasets with properties for investigating model performance.
- Dimensionality Reduction: used for decreasing the number of attributes in data for summarization, visualization and feature selection such as Principal component analysis.
- Parameter Tuning: for getting the most out of supervised models.
- Clustering: used for grouping unlabeled data such as Hierarchical and KMeans.
- Feature extraction: used for important attributes in image and text data.

2.3 Anaconda

Anaconda is a free and open source python distribution for scientific computation that helps for package management and deployment. This involves data science package for linux, windows and macOS. Anaconda has over 250 packages which installs automatically and there are 7500 additional open source packages that can be install by using PyPI, conda package and virtual environment manager. Conda and pip has many differences including how package dependency is managed. Pip while installing packages it will automatically installs all the dependent packages without checking for any other conflicts from the previously installed packages. Whereas, Conda first checks the current environment and other previously installed packages for any conflicts and shows the warnings if any conflicts are there.

Anaconda navigator

Anaconda navigator is desktop application which includes all the anaconda distribution that helps the user to launch application and conda package management, channels and environment without the user typing a single line of commands on command-line. Searching for package on anaconda cloud or on the local repository, to run them the packages and to install those packages the Anaconda navigator is very useful. It is available for Linux, windows and MacOS.

Many of the below application are available in navigator:

1. Jupyter notebook.
2. JupyterLab.
3. Spyder.
4. Glue.
5. Orange.
6. VS code.
7. R Studio.

Note: -

- **For our Data analysis and prediction, we have used Jupyter notebook.**
- **For our 3D data visualization coding we have used Visual studio code.**

2.4 Anaconda Applications Used

Jupyter

The Jupyter Notebook is an open-source web application that permits us to make and distribute documents that contain live code, visualizations, equations and narrative text. Uses include data cleaning, data transformation, data preprocessing, numerical calculation, statistical modeling, machine learning, data visualization, and far more. Jupyter supports more than 40 programming languages such as Python, R, Scala and Julia. It can be share with others via dropbox, GitHub, Email and Jupyter Notebook viewer. An interactive output can be generated with the code to HTML, LaTeX, custom MIME types, images and videos. It can be integrated to big data tools such as Apache Spark.

Visual Studio Code

Visual Studio code is open source code editor by Microsoft for Windows, MacOS and Linux. It has many features such as syntax highlighting, code debugging, code completion, code refactoring, snippets and embedded GitHub. Users can change keyboard shortcuts, themes, preferences and can even install extensions that add extra functionality. Most of the common programming languages such as JavaScript, Typescript, JSON, CSS and HTML, debugging support for NodeJS.

3. Data Science Methodologies

Self-evaluation of the Data and learning from the earlier experience in order to provide results without specific programming methods. It's the main purpose to make computers learn automatically without human assist. Many of the below machine learning and statistical algorithms are used to answer the objectives of our report.

3.1 Clustering

[6] Cluster is group of similar objects. The technique which groups similar objects is called clustering. Objects in the same group are more like one another than the objects in different groups.

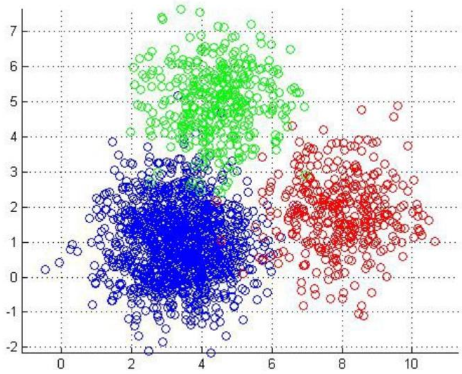


Fig 1. Clustering

Most popular clustering algorithms used by data scientist and analysts.

1. Hierarchical clustering.
2. K-Means clustering.
3. Mean-Shift clustering.
4. Density Based Spatial clustering of Application with Noise (DBSCAN)
5. Expectation-Maximization (EM) clustering using Gaussian Mixture Models (GMM)

3.2 Hierarchical Clustering Algorithm

HCA or Hierarchical Clustering Algorithm is an Unsupervised clustering algorithm which creates cluster that mostly order from top to bottom. Hierarchical Clustering Algorithm is divided into following:

1. Agglomerative Hierarchical Clustering Algorithm
2. Divisive Hierarchical Clustering Algorithm

Agglomerative Hierarchical Clustering Algorithm

Agglomerative Hierarchical Clustering Algorithm is the very widely used type of hierarchical clustering used for grouping objects into clusters based on their similarity. It's otherwise called as Agglomerative Nesting (AGNES). It's a "bottom-up" methodology each observation starts in its individual cluster, and then pairs of clusters are combined as one moves up the hierarchy if the objects are similar.

Steps Involved?

1. Each data object point starts as single-point cluster- forms N clusters
2. Two closest object points and cluster them to one cluster- forms N-1 clusters
3. Take the two closest clusters and make them one cluster- forms N-2 clusters.
4. Step-3 is repeated until we are left with only one cluster.

To calculate the distance between the objects which in turn help to deciding the rules for clustering. These methods are called as linkage methods. Some of these methods are:

1. Single Linkage method: Distance between two clusters is shortest distance between two points in each cluster. Outlier are mostly the one which will get merged at the end.
2. Complete Linkage method: Distance between two clusters is longest distance between two points in each **cluster**. Outlier are mostly the one which will get merged at the first.
3. Centroid Linkage method: Before merging two cluster calculate distance between the centroid of cluster 1 and centroid of cluster 2.
4. Average Linkage method: Distance between two clusters is average distance between two points in each cluster.

There is no hard and fast method to be chosen one, it depends on the individual. Different methods lead to different clusters.

Dendrogram

Dendrogram is tree diagram showing hierarchical cluster relationships among different sets of data.

Note:

1. Distance among data objects represents differences.
2. Height of the blocks signifies the distance amongst clusters.

There can be row or column graph for dendrogram. Dendrogram can be circular or fluid shape.

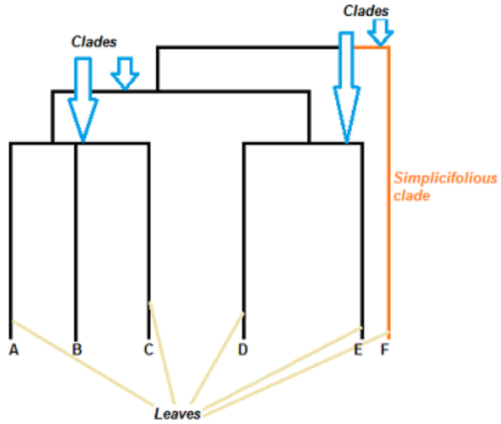


Fig 2: Dendrogram

Clade can ideally have a limitless amount of leaves. Though, the more leaves you have, the tougher the graph will be to decipher.

Divisive Hierarchical Clustering Algorithm

Divisive Hierarchical Clustering Algorithm is a top-down approach clustering where cluster all the objects into one cluster and then divide the cluster to similar two clusters. This continuous till each cluster contains one objects. This approach is opposite to Agglomerative Hierarchical Clustering

Algorithm. Studies say that Divisive Hierarchical Clustering Algorithm produces better accurate hierarchies than Agglomerative Hierarchical Clustering Algorithm. For both the clustering algorithm user must provide the cluster number.

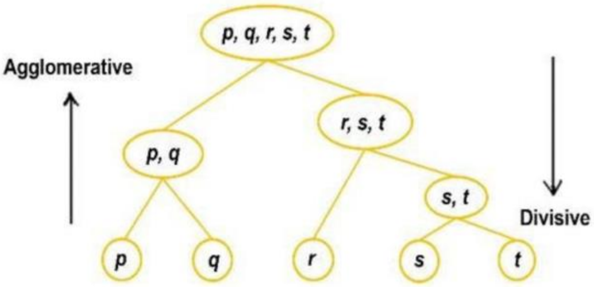


Fig 3: Hierarchical Clustering Algorithm

3.3 K-means Clustering

[7] Within each cluster the points in it should be similar. Hence, our main aim is to minimize the distance between the points in each cluster. K-Means algorithm is to reduce the sum of distances between the points and their corresponding cluster centroid.

There are 5 important steps for k-means to create clusters for these objects.

1. Choosing the number of clusters k- Deciding the number of clusters k in k-means is very important.
2. Random section of k from the data as centroid.
3. Assigning/cluster all the points to the closest cluster centroid.
4. Recalculate the centroids of newly formed clusters.
5. Step 3 and 4 are repeated.

3 steps Criteria to stop K-means clustering

1. If the centroids are not changing for the newly formed cluster.
2. Points in the cluster doesn't change.
3. Maximum number of iterations has been reached.

Challenges faces with K-means clustering algorithm.

One of the frequent challenges we encounter while working with K-Means is that the size of clusters is not the same.

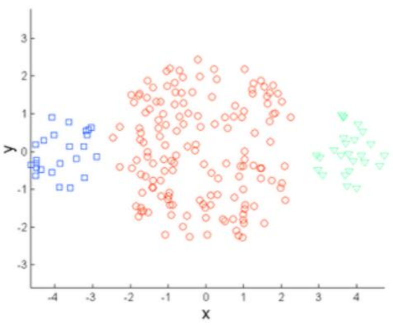


Fig 4: Original Points

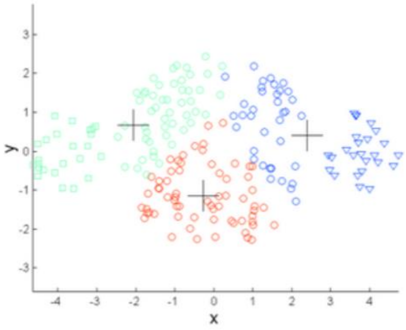


Fig 5: K-means(k=3)

leftmost and the rightmost clusters are of smaller size compared to the central cluster. Now, we apply k-means with k=3 clustering on these points, the outcomes will be somewhat like *Fig 5.*

Second if the densities of the points are different.

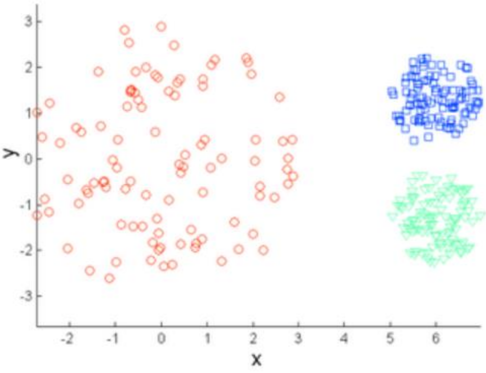


Fig 6: Original Points

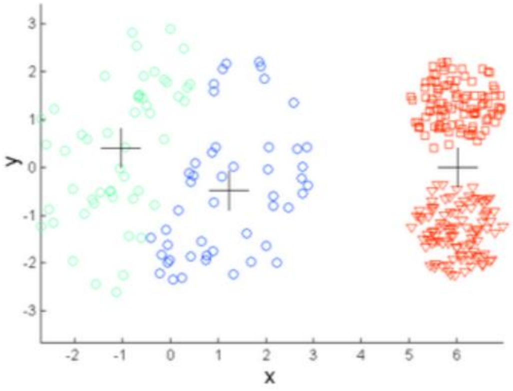


Fig 7: K-means(k=3)

We can see from *Fig 7* points which are compact are in single cluster. Whereas the point which are loosely spread was assigned to single cluster, but now assigned to different clusters.

One solution for this is to have high number of clusters. So instead of using k=3 use a bigger k number like k=10.

3.4 Confidence Intervals

The confidence interval (CI) could be a variety of values that is expected to incorporate a population value with a specific degree of confidence. It's frequently stated in % whereby a population means lie down among an upper and lower interval.

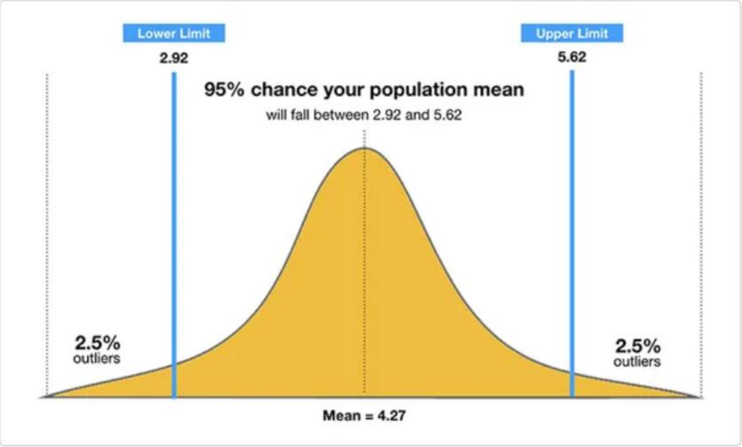


Fig 8: Confidence Interval

The 95% confidence interval may be a variety of values that you just may be 95% sure includes true mean of the population. As sample size increases, the range of interval values will narrow, indicating that you merely know that mean with far more precision compared with a smaller sample size.

4.Data Collection and Overview

One of the most important process before doing the data cleaning and data analysis is understanding the dataset which was collected. Understanding each of the attributes in the dataset is essential to come up with research questions for our analysis. We have made use of two dataset in our analysis.

4.1 Data Collection

What is Data Collection?

Data collection is described as the "method of collecting and measuring information on variables of interest, in a recognized methodical approach that permits one to answer several questions, stated research queries, test hypotheses, and evaluate outcomes."[4]

Why Data Collection?

Significance for Data collection meant for our analysis are as follows:

1. It facilitates to answer few of our basic questions such which age group and which gender is highly affected by Gun violence.
2. For our Geo-Visualization utilizing the longitude and latitude attributes in our dataset.
3. For predicting which month of year and day of week is extremely dangerous and safe from Gun violence.
4. Using number of people killed, people injured, number of fatal incidents and Gun involved attributes to generate Risk score in order to rank the U.S. States its risk for Gun violence. This risk rank is for our initial analysis only based on United States Gun Violence Dataset that was collected from Kaggle.
5. Apart from work with United States Gun Violence Dataset to do initial analysis and preliminary Risk score ranking, we next include United States Population, Per-Capita and Un-Employment Dataset to determine if there is any relationship between extra included attributes and the Gun violence dataset attributes.
6. Finally, with all the collected data attributes we will rank the U.S. state based the danger level for Gun Violence.

Data Collection Source

Source for our Data analysis and Geo-Visualization are follows:

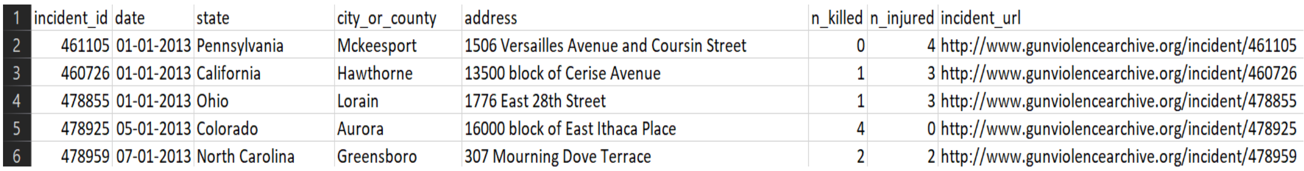
1. Source of the United States Gun Violence Dataset: https://www.kaggle.com/jameslko/gun-violence-data
2. Source of the Dataset for 2018 Population Attribute: https://www.worldatlas.com/articles/us-states-by-population.html
3. Source of the Dataset for Per-Capita Income and Un-Employment attributes: https://www.statista.com/

4.2 United States Gun Violence Dataset

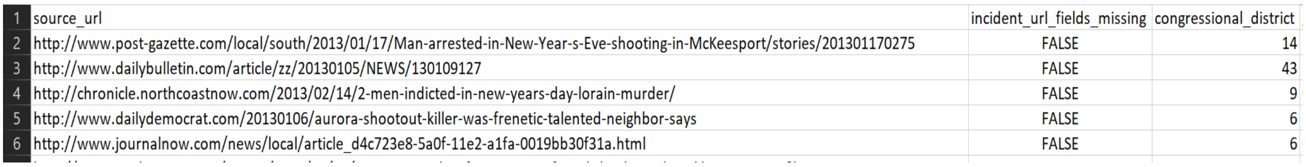
In this analysis we are incorporating a dataset which has over 260k gun violence incidents which happened from 2013 to 2018, with comprehensive information about each incident. With the help of this dataset we are aiming to make informed predictions about forthcoming trends which plays a crucial role in crime reduction in the nearby future. This dataset holds the attributes such as:

1. Incident_id – Unique identity number for each crime.
2. Date – Date of the crime occurred
3. State – State in which the crime occurred
4. City_or_County – City or County in the specified state where the crime occurred
5. Address – Address of the crime happened place.
6. n_killed – number of persons killed because of this crime
7. n_injured - number of persons injured because of this crime
8. source_url – Reference to the reporting source
9. incident_url - URL regarding the incident
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 - Latitude coordinate of the incident or crime occurred
16. location_description – Description of the location of incident or crime
17. longitude - Longitude coordinate of the incident or crime occurred
18. n_guns_involved - Number of guns involved in incident occurred
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 participants at the time crime
22. participant_gender - Gender of participant(s) in the incidents
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 involved in the crime
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.

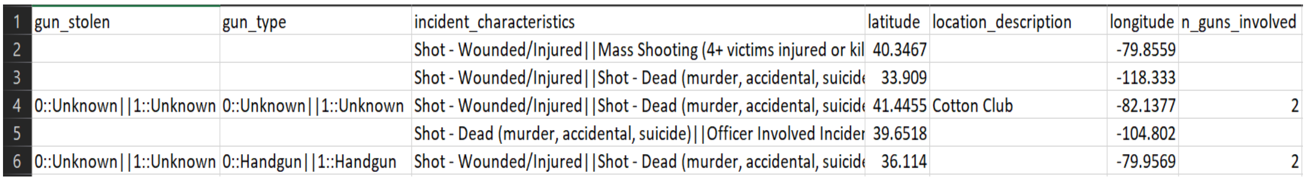
Dataset overview



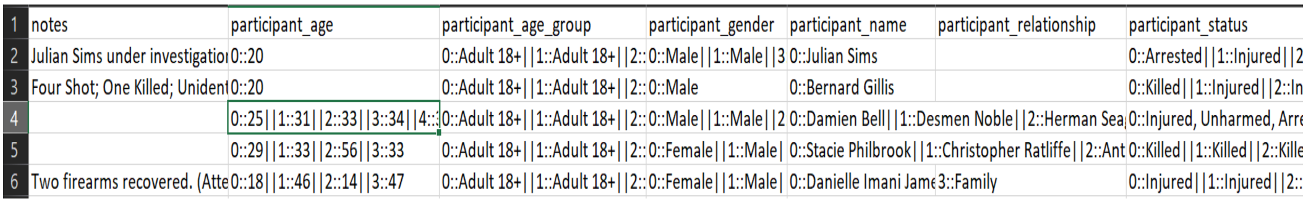
1	incident_id	date	state	city_or_county	address	n_killed	n_injured	incident_url
2	461105	01-01-2013	Pennsylvania	Mckeesport	1506 Versailles Avenue and Coursin Street	0	4	http://www.gunviolencearchive.org/incident/461105
3	460726	01-01-2013	California	Hawthorne	13500 block of Cerise Avenue	1	3	http://www.gunviolencearchive.org/incident/460726
4	478855	01-01-2013	Ohio	Lorain	1776 East 28th Street	1	3	http://www.gunviolencearchive.org/incident/478855
5	478925	05-01-2013	Colorado	Aurora	16000 block of East Ithaca Place	4	0	http://www.gunviolencearchive.org/incident/478925
6	478959	07-01-2013	North Carolina	Greensboro	307 Mourning Dove Terrace	2	2	http://www.gunviolencearchive.org/incident/478959



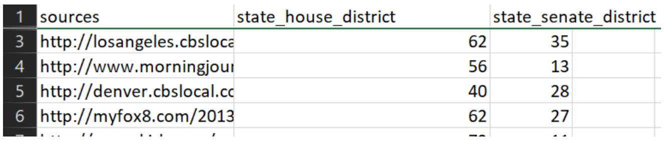
1	source_url	incident_url_fields_missing	congressional_district
2	http://www.post-gazette.com/local/south/2013/01/17/Man-arrested-in-New-Year-s-Eve-shooting-in-McKeesport/stories/201301170275	FALSE	14
3	http://www.dailybulletin.com/article/zz/20130105/NEWS/130109127	FALSE	43
4	http://chronicle.northcoastnow.com/2013/02/14/2-men-indicted-in-new-years-day-lorain-murder/	FALSE	9
5	http://www.dailydemocrat.com/20130106/aurora-shootout-killer-was-frenetic-talented-neighbor-says	FALSE	6
6	http://www.journalnow.com/news/local/article_d4c723e8-5a0f-11e2-a1fa-0019bb30f31a.html	FALSE	6



1	gun_stolen	gun_type	incident_characteristics	latitude	location_description	longitude	n_guns_involved
2			Shot - Wounded/Injured||Mass Shooting (4+ victims injured or kil	40.3467		-79.8559	
3			Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide	33.909		-118.333	
4	0::Unknown||1::Unknown	0::Unknown||1::Unknown	Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide	41.4455	Cotton Club	-82.1377	2
5			Shot - Dead (murder, accidental, suicide)||Officer Involved Incider	39.6518		-104.802	
6	0::Unknown||1::Unknown	0::Handgun||1::Handgun	Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide	36.114		-79.9569	2



<table>
<tr><th>1</th><th>notes</th><th>participant_age</th><th>participant_age_group</th><th>participant_gender</th><th>participant_name</th><th>participant_relationship</th><th>participant_status</th></tr>
<tr><td>2</td><td>Julian Sims under investigatio</td><td>0::20</td><td>0::Adult 18+||1::Adult 18+||2::</td><td>0::Male||1::Male||3</td><td>0::Julian Sims</td><td></td><td>0::Arrested||1::Injured||2</td></tr>
<tr><td>3</td><td>Four Shot; One Killed; Unident</td><td>0::20</td><td>0::Adult 18+||1::Adult 18+||2::</td><td>0::Male</td><td>0::Bernard Gillis</td><td></td><td>0::Killed||1::Injured||2::In</td></tr>
<tr><td>4</td><td></td><td>0::25||1::31||2::33||3::34||4::</td><td>0::Adult 18+||1::Adult 18+||2::</td><td>0::Male||1::Male||2</td><td>0::Damien Bell||1::Desmen Noble||2::Herman Sea</td><td></td><td>0::Injured, Unharmed, Arre</td></tr>
<tr><td>5</td><td></td><td>0::29||1::33||2::56||3::33</td><td>0::Adult 18+||1::Adult 18+||2::</td><td>0::Female||1::Male|</td><td>0::Stacie Philbrook||1::Christopher Ratliffe||2::Ant</td><td></td><td>0::Killed||1::Killed||2::Kille</td></tr>
<tr><td>6</td><td>Two firearms recovered. (Atte</td><td>0::18||1::46||2::14||3::47</td><td>0::Adult 18+||1::Adult 18+||2::</td><td>0::Female||1::Male|</td><td>0::Danielle Imani Jame</td><td>3::Family</td><td>0::Injured||1::Injured||2::</td></tr>
</table>



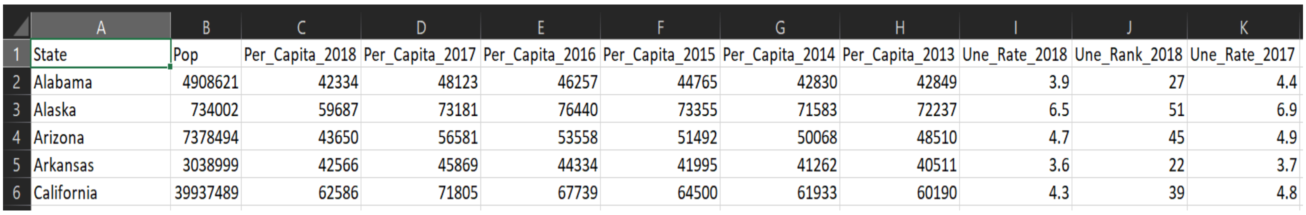
1	sources	state_house_district	state_senate_district
3	http://losangeles.cbsloca	62	35
4	http://www.morningjour	56	13
5	http://denver.cbslocal.cc	40	28
6	http://myfox8.com/2013	62	27

4.3 United States Population, Per-Capita and Un-Employment Dataset

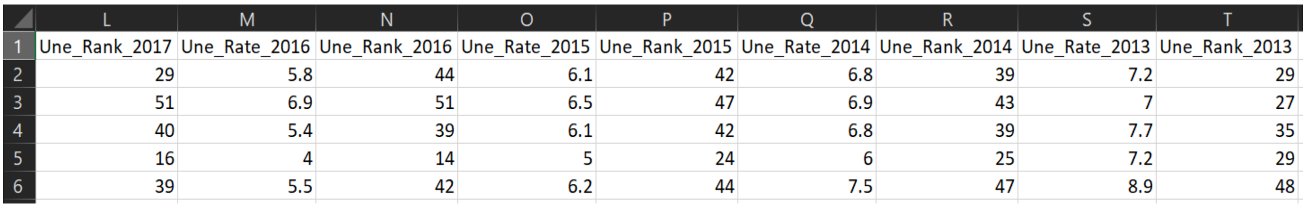
United States Population, Per-Capita and Un-Employment Dataset is the Second Dataset that we will consider for further analysis such Risk score generation to understand the Socio-Economic status of an individual factor has relationship with Gun violence data (Dataset 1). Risk score ranking system generated including this extra parameter to get more accurate prediction of which U.S. State is Dangerous and how the socio-economic factor plays a curial role in Gun culture in United States of America. Dataset consists data of 51 United states. This dataset holds the attributes such as:

1. State: United States 51 states name
2. Pop: 2018 Population of 51 US States
3. Per_Capita_2018: Per-capita income of 51 US states in 2018
4. Per_Capita_2017: Per-capita income of 51 US states in 2017
5. Per_Capita_2016: Per-capita income of 51 US states in 2016
6. Per_Capita_2015: Per-capita income of 51 US states in 2015
7. Per_Capita_2014: Per-capita income of 51 US states in 2014
8. Per_Capita_2013: Per-capita income of 51 US states in 2013
9. Une_Rate_2018: Un-Employment rate of 51 US states in 2018
10. Une_Rank_2018: Un-Employment Rank of 51 US states in 2018 based on the Un-employment rate
11. Une_Rate_2017: Un-Employment rate of 51 US states in 2017
12. Une_Rank_2017: Un-Employment Rank of 51 US states in 2017 based on the Un-employment rate
13. Une_Rate_2016: Un-Employment rate of 51 US states in 2016
14. Une_Rank_2016: Un-Employment Rank of 51 US states in 2016 based on the Un-employment rate
15. Une_Rate_2015: Un-Employment rate of 51 US states in 2015
16. Une_Rank_2015: Un-Employment Rank of 51 US states in 2015 based on the Un-employment rate
17. Une_Rate_2014: Un-Employment rate of 51 US states in 2014
18. Une_Rank_2014: Un-Employment Rank of 51 US states in 2014 based on the Un-employment rate
19. Une_Rate_2013: Un-Employment rate of 51 US states in 2013
20. Une_Rank_2013: Un-Employment Rank of 51 US states in 2013 based on the Un-employment rate

Dataset Overview



	A	B	C	D	E	F	G	H	I	J	K
1	State	Pop	Per_Capita_2018	Per_Capita_2017	Per_Capita_2016	Per_Capita_2015	Per_Capita_2014	Per_Capita_2013	Une_Rate_2018	Une_Rank_2018	Une_Rate_2017
2	Alabama	4908621	42334	48123	46257	44765	42830	42849	3.9	27	4.4
3	Alaska	734002	59687	73181	76440	73355	71583	72237	6.5	51	6.9
4	Arizona	7378494	43650	56581	53558	51492	50068	48510	4.7	45	4.9
5	Arkansas	3038999	42566	45869	44334	41995	41262	40511	3.6	22	3.7
6	California	39937489	62586	71805	67739	64500	61933	60190	4.3	39	4.8



	L	M	N	O	P	Q	R	S	T
1	Une_Rank_2017	Une_Rate_2016	Une_Rank_2016	Une_Rate_2015	Une_Rank_2015	Une_Rate_2014	Une_Rank_2014	Une_Rate_2013	Une_Rank_2013
2	29	5.8	44	6.1	42	6.8	39	7.2	29
3	51	6.9	51	6.5	47	6.9	43	7	27
4	40	5.4	39	6.1	42	6.8	39	7.7	35
5	16	4	14	5	24	6	25	7.2	29
6	39	5.5	42	6.2	44	7.5	47	8.9	48

5. Geovisualization

5.1 Introduction

In order to make easier the high-performing Geo based visualization action we choose for DeckGl library. It was built by the group of software people from Uber Visualization Team. In 2016, they launched the version 1 of deck.gl in order to help with the people who wanted to do big visualization on the web. It's an open source software and the main purpose of this to invite other peoples to use and build on this framework.Deck.gl version4 integrates the supports for advanced geospatial exploration along with new non-spatial visualization capabilities. In addition to that it came up with many demo's and examples which invited lot of software developers and passionate visualization engineers to use this framework much easier and enabling quicker and more seamless development of Web-GL powered visualizations. This is the main objective of the version5 which we opt to develop our Geo based visualization for Gun Laws dataset US.

The focus of deck.gl version 5 is it overcomes some challenges. They are:

1. A Pure JavaScript API
2. Framework agnosticism
3. Scripting support
4. Ease Of Use

Pure JavaScript API:

Before Version 5 deck.gl is unique for the developers who are much more comfortable with using React Framework. In order to make available this library to all users they are making it possible to use deck.gl without react. They applied a concept called "One API" philosophy, meaning that there are some minor differences how these API's is getting initialized, almost all classes and properties have the same name and semantics across versions.

Framework agnosticism:

Though version4 has lot of dependencies on React framework, version 5 engineered out all the React dependencies from deck.gl. Now deck.gl officially supports being used without any specific JavaScript UI framework and now used as a base for building integrations with other UI frameworks

Scripting Support:

Automatic Mapbox base map integration, an additional ingredient in React- independent JavaScript API was developed and published as a script version of deck.gl.

Ease of Use:

This has been done with added features like it automates the highlighting, resizing of the visualization images, automatic loading of layered data, automatic component positioning, automatic controls and with more declarative API's.

Data usually an array of **JSON** objects is mapped into a stack of visual layers using deck.gl. These layers can be any of the forms like icons, polygons, texts and these created layers can be viewed with the help of views such as map, first-person, orthographic.

An important feature which makes this odd from other visualization libraries is it can handle some challenges like:

1) Performance rendering and updating larger dataset
2) Event handling with more interaction such as picking, highlighting and filtering
3) Cartographic projections and integration with major basemap providers
4) A catalog of proven, well-tested layers

5.2 Why Deck.GL?

Deck.gl is a library that resolves this problem by running expensive computations on the GPU with webGL. This means we can run real time 3D visualizations on datasets with millions of geographical points. It is better when compared to geo visualizations which is available in python. For our dataset it has around 260K records which slows down the performance if we proceed with python geographical packages. Deck.gl greatly increases the performance of the system as well as it shows the 3-dimensional view of the geo graphical visualization.

IDE Used

IDE which we preferred to develop this 3D Geo graphical visualizations using Deck.gl is Visual Studio 2013.The reason behind this is as we developed this visualizations with the help of JavaScript an scripting language, Visual Studio is the best and preferred IDE to play with JavaScript as it has lot of features like good code editing with navigation, syntax highlighting, code folding, debugging and JavaScript function timing. So, these above features make this IDE an odd one while compared with other IDE available in the markets.

5.3 Steps Involved

As mentioned above now Deck.gl acts as a framework independent which means it works on any type of frameworks. Our 3D visualizations images run on a framework called vanilla JavaScript with Webpack. The first and foremost things to start developing this 3D visualization images are to obtain an API key from Google Cloud for Google Maps JS.

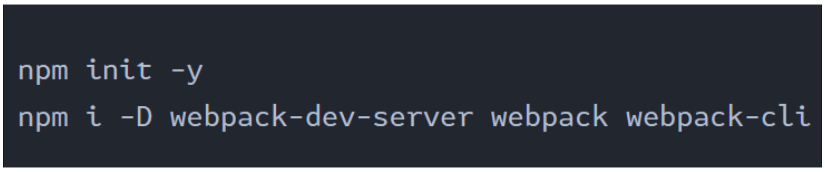
We have made use of the below attributes from the United States Gun Violence Dataset:

1. City_or_County
2. Longitude
3. Latitude
4. n_killed
5. State
6. Incident_id

Next major step is setting up of Google maps API Key from below URL.
https://developers.google.com/maps/documentation/javascript/get-api-key

After setting up the API key we integrate with our framework.

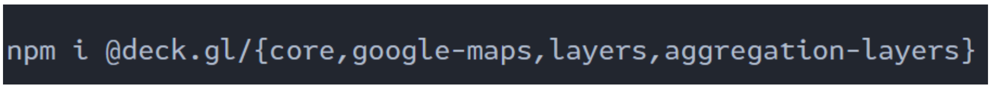
Add necessary package and create webpack project:



```
npm init -y
npm i -D webpack-dev-server webpack webpack-cli
```

Fig 9: Create Webpack project

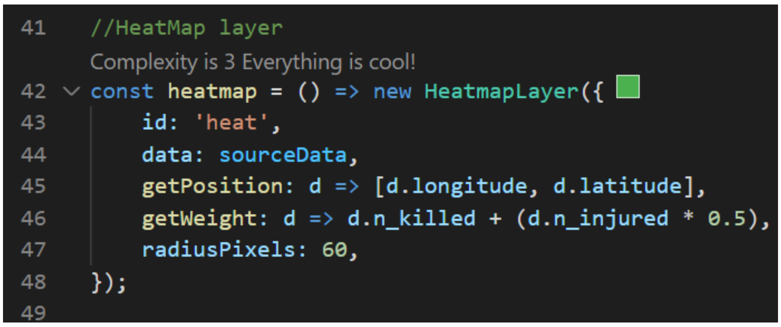
Install Deck.GL package for our project



```
npm i @deck.gl/{core,google-maps,layers,aggregation-layers}
```

Fig 10: Add Deck.GL package

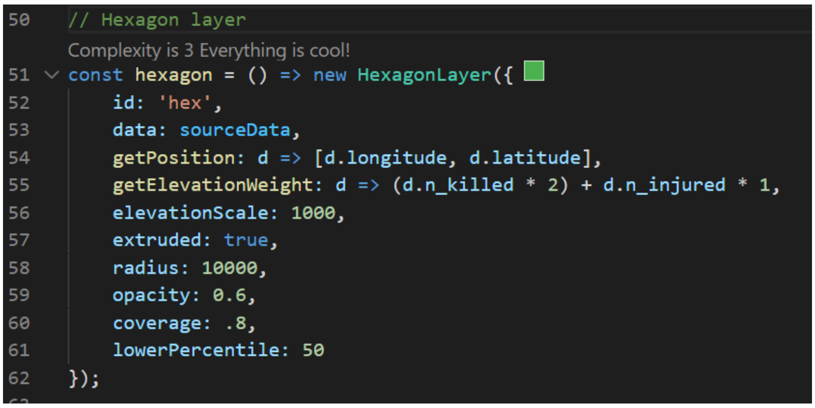
Below code help us to generate Heatmaps for our Dataset:



```
//HeatMap layer
Complexity is 3 Everything is cool!
const heatmap = () => new HeatmapLayer({
    id: 'heat',
    data: sourceData,
    getPosition: d => [d.longitude, d.latitude],
    getWeight: d => d.n_killed + (d.n_injured * 0.5),
    radiusPixels: 60,
});
```

Fig 11: Heatmap Generation code

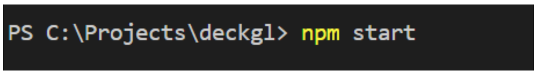
Below code helps us to generate 3D hexagonal layer:



```
// Hexagon layer
Complexity is 3 Everything is cool!
const hexagon = () => new HexagonLayer({
    id: 'hex',
    data: sourceData,
    getPosition: d => [d.longitude, d.latitude],
    getElevationWeight: d => (d.n_killed * 2) + d.n_injured * 1,
    elevationScale: 1000,
    extruded: true,
    radius: 10000,
    opacity: 0.6,
    coverage: .8,
    lowerPercentile: 50
});
```

Fig 12: Hexagonal Projection code

After completion of our coding we will use below line of command to run our application in localhost server.



```
PS C:\Projects\deckgl> npm start
```

Fig 13: To start the Application

Once after running our 3D geo visualization starts running in localhost server.

5.4 Geovisualization Result

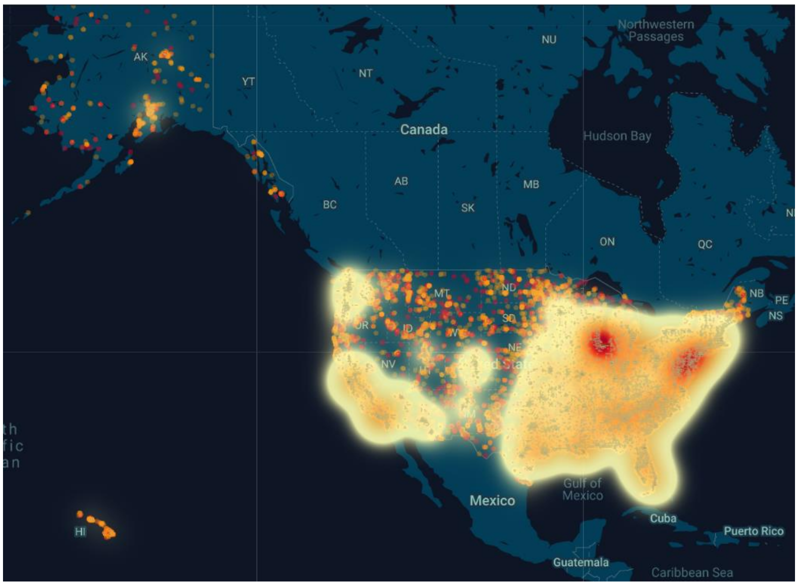


Fig 14: Heat Map of the United States Gun Violence from 2013 to 2018

Above Heat Map shown is for the United states Gun violence incidents from 2013 to 2018. Following observation are made from the above heat map:

- There is a high density of the Gun violence incident towards the North-East and South-East of United states.
- Alaska, Hawaii and North-west part has a smaller number of incidents.
- In South-West that is California and in North-East that is Illinois and New York region has high incident based on the Heat-Map

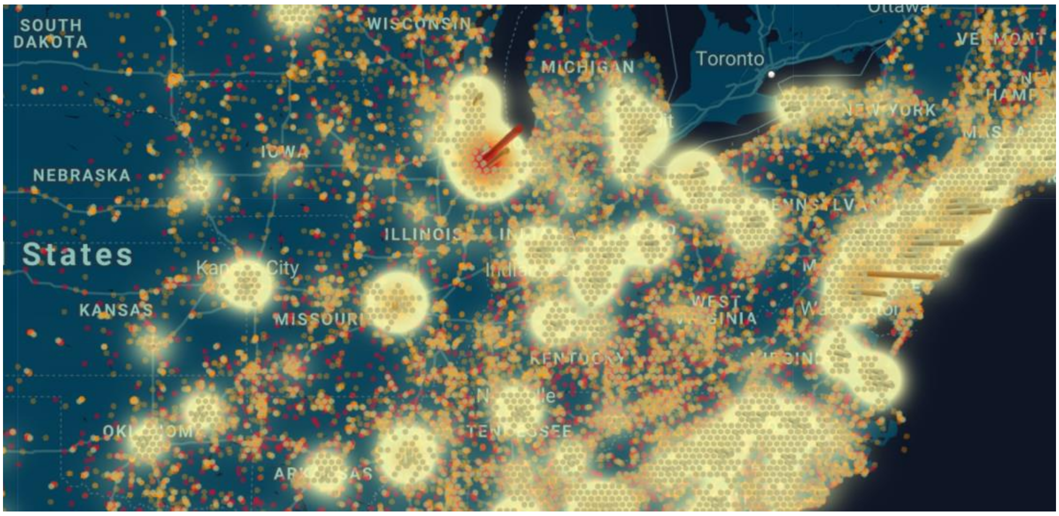


Fig 15: Hexagon projection Layer Map of the United States Gun Violence from 2013 to 2018

Hexagon projection helps us to understand intensity of the gun violence in **10kms radius**. *Fig 13* shows Chicago region in Illinois state has the highest number of incident than any other regions in the United States.

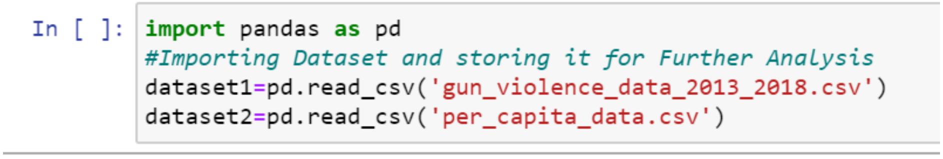
6. Data Preparation for Analysis

Data preparation involves data collection, data cleaning and data modification of the Datasets. Dataset which was collected from different sources are stored as CSV file format. These datasets are imported for cleaning if any incorrect, mismatch or missing data is found, and cleaned data is modified if it useful for our analysis.

6.1 Data Importing

CSV (Comma-separated value) are the most used file format to store and transfer data. Using python reading, writing and manipulate data is a key skill for any data scientist and business analyst. Python Pandas library is used to read our Two datasets and manipulate the data frame using it.

We are loading our two datasets for analysis.



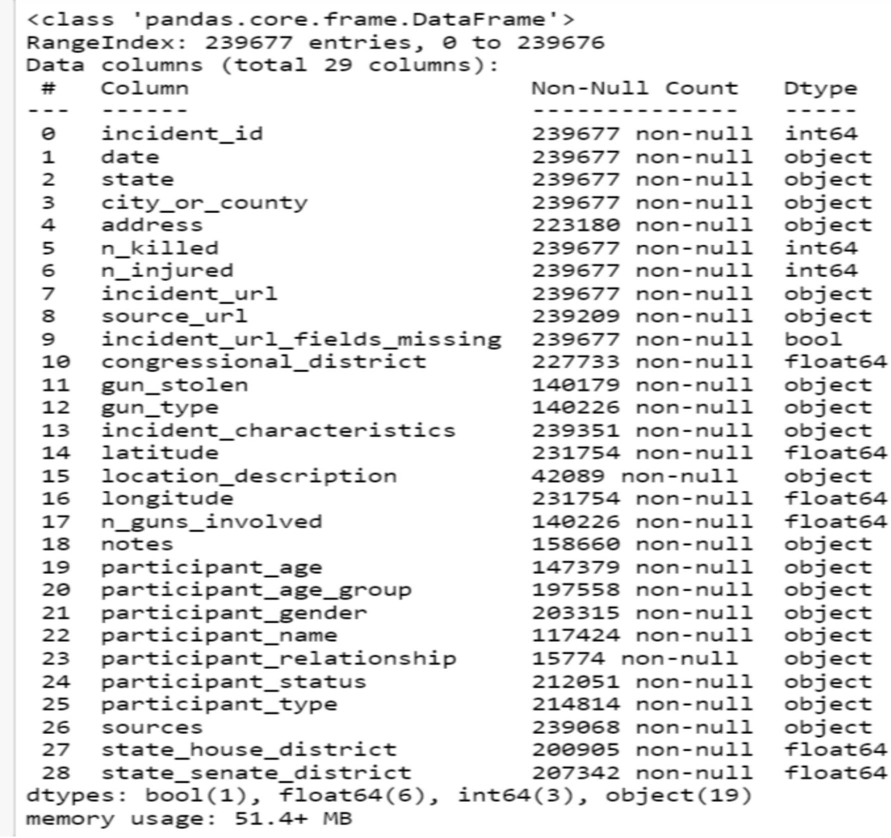
```
In [ ]: import pandas as pd
        #Importing Dataset and storing it for Further Analysis
        dataset1=pd.read_csv('gun_violence_data_2013_2018.csv')
        dataset2=pd.read_csv('per_capita_data.csv')
```

Fig 16: Data importing using Python Pandas library

Our first dataset contains United States Gun Violence from 2013 to 2018 and second dataset has United States population, per-capita and Un-Employment data.

Initial Description of the Datasets:

1. United States Gun Violence Dataset details before data cleaning:

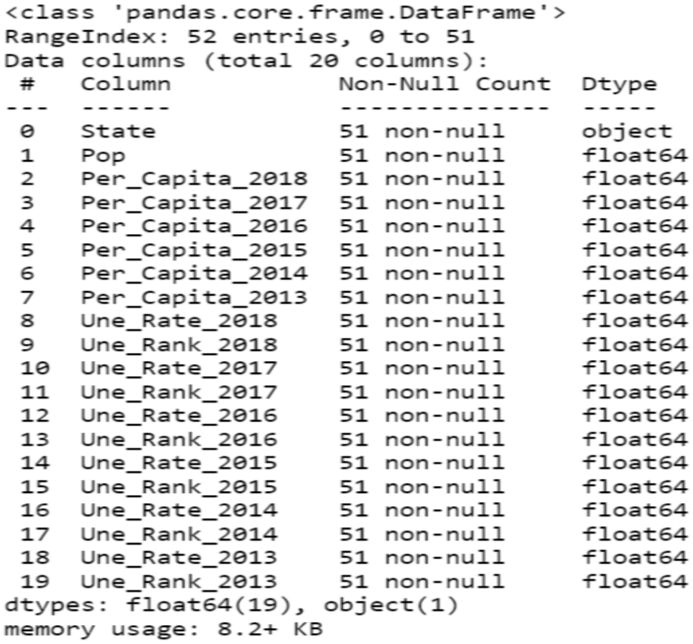


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 29 columns):
 #   Column                         Non-Null Count   Dtype  
---  ------                         --------------   -----  
 0   incident_id                    239677 non-null  int64  
 1   date                           239677 non-null  object 
 2   state                          239677 non-null  object 
 3   city_or_county                 239677 non-null  object 
 4   address                        223180 non-null  object 
 5   n_killed                       239677 non-null  int64  
 6   n_injured                      239677 non-null  int64  
 7   incident_url                   239677 non-null  object 
 8   source_url                     239209 non-null  object 
 9   incident_url_fields_missing    239677 non-null  bool   
 10  congressional_district         227733 non-null  float64
 11  gun_stolen                     140179 non-null  object 
 12  gun_type                       140226 non-null  object 
 13  incident_characteristics       239351 non-null  object 
 14  latitude                       231754 non-null  float64
 15  location_description           42089 non-null   object 
 16  longitude                      231754 non-null  float64
 17  n_guns_involved                140226 non-null  float64
 18  notes                          158660 non-null  object 
 19  participant_age                147379 non-null  object 
 20  participant_age_group          197558 non-null  object 
 21  participant_gender             203315 non-null  object 
 22  participant_name               117424 non-null  object 
 23  participant_relationship       15774 non-null   object 
 24  participant_status             212051 non-null  object 
 25  participant_type               214814 non-null  object 
 26  sources                        239068 non-null  object 
 27  state_house_district           200905 non-null  float64
 28  state_senate_district          207342 non-null  float64
dtypes: bool(1), float64(6), int64(3), object(19)
memory usage: 51.4+ MB
```

Fig 17: United States Gun Violence Dataset Details before Data Cleaning

United States Gun Violence Dataset consist of 239677 entries and around 29 attributes in our datasets. From our initial dataset description, we see that many of the attributes contains Null Values. It evident that we must do data cleaning before doing further data analysis and even some of the attributes are not required for our analysis. Some of the attributes can be used to extract new attributes that we will do in our data pre-processing.

2. United States Population, Per-Capita and Un-Employment Dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   State            51 non-null     object
 1   Pop              51 non-null     float64
 2   Per_Capita_2018  51 non-null     float64
 3   Per_Capita_2017  51 non-null     float64
 4   Per_Capita_2016  51 non-null     float64
 5   Per_Capita_2015  51 non-null     float64
 6   Per_Capita_2014  51 non-null     float64
 7   Per_Capita_2013  51 non-null     float64
 8   Une_Rate_2018    51 non-null     float64
 9   Une_Rank_2018    51 non-null     float64
 10  Une_Rate_2017    51 non-null     float64
 11  Une_Rank_2017    51 non-null     float64
 12  Une_Rate_2016    51 non-null     float64
 13  Une_Rank_2016    51 non-null     float64
 14  Une_Rate_2015    51 non-null     float64
 15  Une_Rank_2015    51 non-null     float64
 16  Une_Rate_2014    51 non-null     float64
 17  Une_Rank_2014    51 non-null     float64
 18  Une_Rate_2013    51 non-null     float64
 19  Une_Rank_2013    51 non-null     float64
dtypes: float64(19), object(1)
memory usage: 8.2+ KB
```

Fig 18: United States Population, Per-Capita and Un-Employment Dataset before Data Cleaning

United States Population, Per-Capita and Un-Employment Dataset consists of 52 entries and 20 attributes. This dataset consists of Population, Per-Capita and Un-Employment details of U.S. States. In this dataset all the entries and attributes for our analysis related to finding relationship between Gun violence and this extra attributes in this dataset.

Step Involved after Data Importing:

1. First, we have to Data exploration or initial analysis of the entries and attributes in our dataset. This step has been done above.
2. Next, we must do the Data cleaning and pre-processing so that the data is good to start with our Data analysis.

6.2 Data Cleaning

What is Data cleaning?

Data that we have need to be cleaned for our analysis, this process of preparing data for data analysis by removing or modifying the incorrect, duplicate, irrelevant and improper data can be called as Data Cleaning. Data that we get mostly be unnecessary or not helpful they may result in an inaccurate result. Data cleaning is not only about removing data, but it is rather than finding a way to maximize the dataset accuracy without necessarily deleting information. This process includes more than removing data, such as correcting spelling, standardizing data and fixing mistakes such as duplicate, missing and empty values.

Why Data cleaning?

Data cleaning is the most important part is any data science related projects or analysis. This process results to the following:

1. More accurate results.
2. Less processing times.
3. Results in less algorithm for analysis.
4. Improves better understanding of models.

6.3 Data cleaning and modification for our Analysis.

First, we must do initial investigation on our Datasets before we do any deleting or modifying on our dataset. It is very important to do an investigation on our dataset so that we come to a concussion which attributes are very important for our analysis. Investigation involves for checking incorrect, missing and duplicate data in dataset.

United States Gun Violence Dataset Cleaning.

After importing Dataset first, we will plot a heatmap to analysis for importance of the attributes.

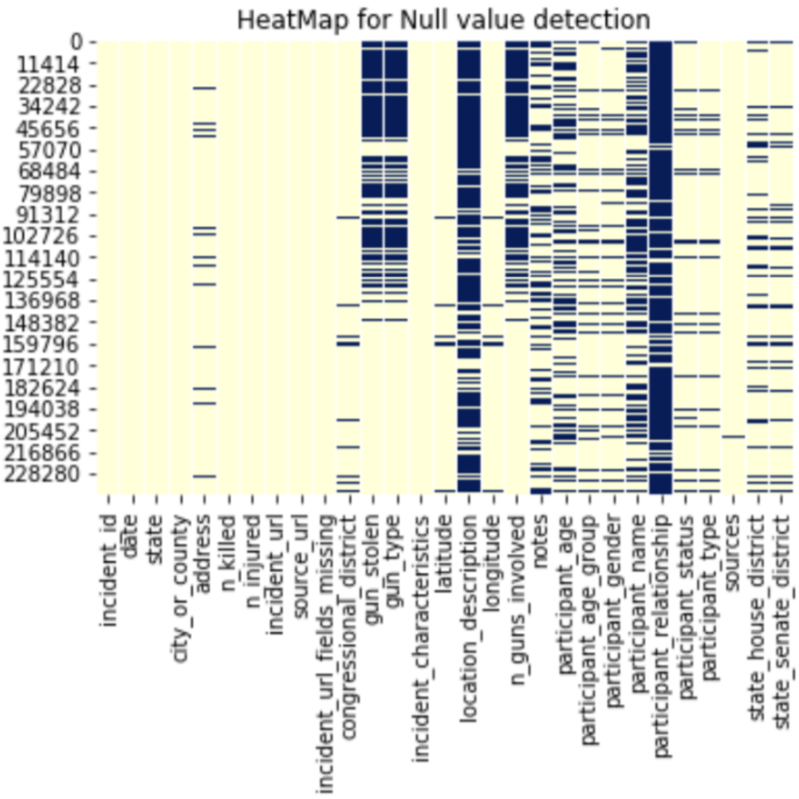


Fig 19: HeatMap to Analyze importance of Data Attributes in our Gun violence dataset.

Following Observation are done from our Gun violence dataset description *Fig 15* and HeatMap *Fig 17*:

1. There are total 239676 entries and 29 attributes in our dataset.
2. Since we know all the description of all our attributes in our dataset from Chapter 4 we can exclude attributes like "incident_id ", "address", "source_url", "incident_url", "

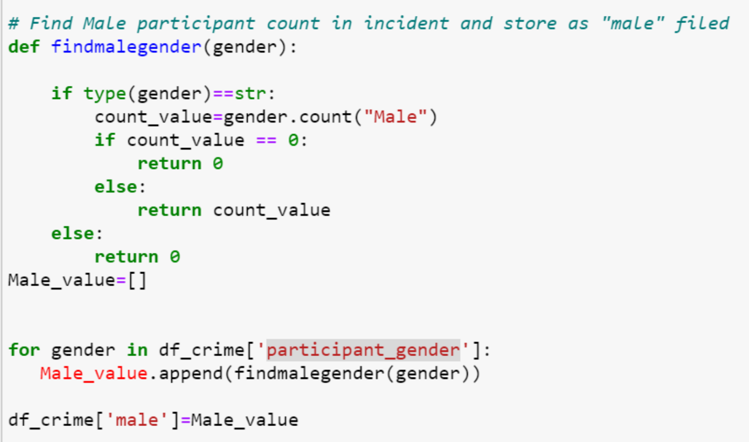
incident_url_fields_missing”, “congressional_district”, ”incident_characteristics”,” participant_status”, ” participant_type”, ”sources”, “notes”, ” state_house_district” and ” state_senate_district”. These attributes may have fewer null values, but we are removing only because these attributes won’t provide necessary importance for our main Analysis objective. Most of these attributes are unique so there won’t be a correct answer for our analysis questions.

3. Data cleaning process involves in deleting attributes if the attributes having more than 40 % Null values in our Dataset. Hence, we can remove “gun_stolen”, “gun_type”, “location_description”, “participant_age_group”, “participant_name” and “participant_relationship”.
4. In our Data analysis we are not making use of “latitude” and “longitude” attributes. We are using these attributes only in our Geo-Visualization.
5. We have checked for duplicate entries in Gun violence dataset.

United States Gun Violence Dataset Modification.

Gun violence Dataset is missing some very explicit attributes such as Number of Male and female involved in incident, Age of each participants, year of the incident, total individuals in each incident, which Month of year and which day of week for each incident. To extract or calculate these attributes we make use of existing attributes in our dataset.

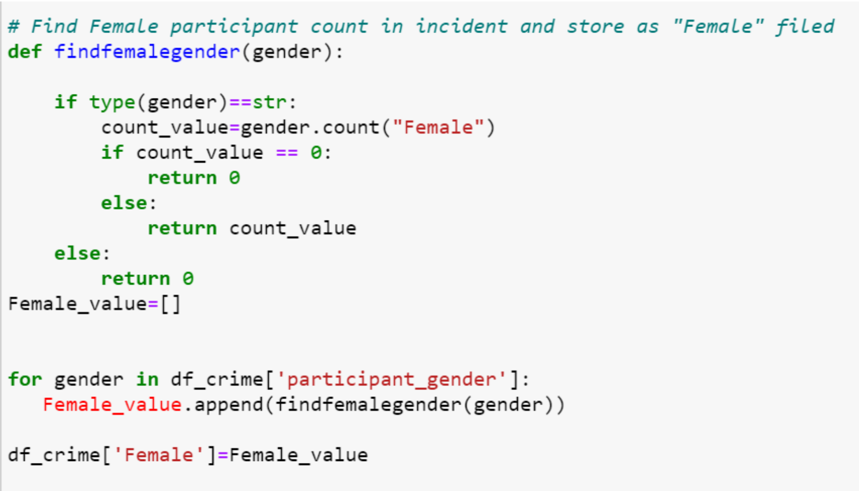
1. To calculating Male participant count from “participant_gender” attribute. We are extracting the “Male” string from this attribute and counting this and storing as a new attribute “male” in United States Gun Violence dataset



```
# Find Male participant count in incident and store as "male" filed
def findmalegender(gender):

    if type(gender)==str:
        count_value=gender.count("Male")
        if count_value == 0:
            return 0
        else:
            return count_value
    else:
        return 0
Male_value=[]


for gender in df_crime['participant_gender']:
   Male_value.append(findmalegender(gender))

df_crime['male']=Male_value
```

Fig 20: Male participant count in incident and store as "male" attribute.

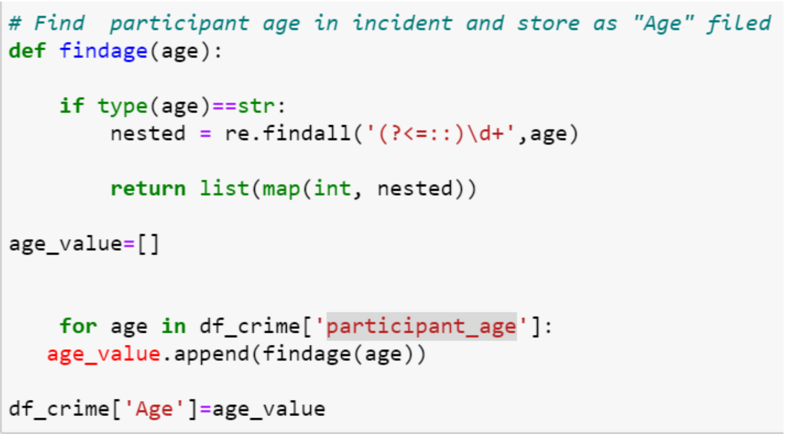
2. To calculating Male participant count from “participant_gender” attribute. We are extracting the “Female” string from this attribute and counting this and storing as a new attribute “female” in United States Gun Violence dataset



```python
# Find Female participant count in incident and store as "Female" filed
def findfemalegender(gender):

    if type(gender)==str:
        count_value=gender.count("Female")
        if count_value == 0:
            return 0
        else:
            return count_value
    else:
        return 0
Female_value=[]


for gender in df_crime['participant_gender']:
   Female_value.append(findfemalegender(gender))

df_crime['Female']=Female_value
```

Fig 21: Female participant count in incident and store as "female" attribute.

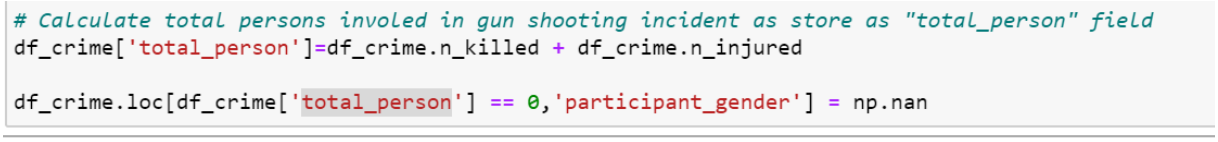
3. Extracting the Participants age Using regular expression from “participant_age” attribute and storing in as new attribute “Age” in United States Gun Violence dataset



```python
# Find  participant age in incident and store as "Age" filed
def findage(age):

    if type(age)==str:
        nested = re.findall('(?<=::)\d+',age)

        return list(map(int, nested))

age_value=[]


    for age in df_crime['participant_age']:
  age_value.append(findage(age))

df_crime['Age']=age_value
```

Fig 22: Participant age in incident and store as "Age" attribute.

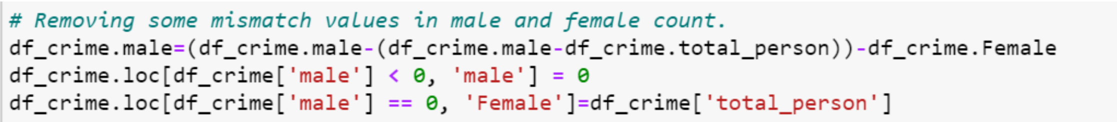
4. Creating new attribute “total_person” which stores the total count of participants who got killed and injured in United States Gun Violence dataset



```python
# Calculate total persons involed in gun shooting incident as store as "total_person" field
df_crime['total_person']=df_crime.n_killed + df_crime.n_injured

df_crime.loc[df_crime['total_person'] == 0,'participant_gender'] = np.nan
```

Fig 23: Calculate total persons involved in gun shooting incident as store as "total_person" field.

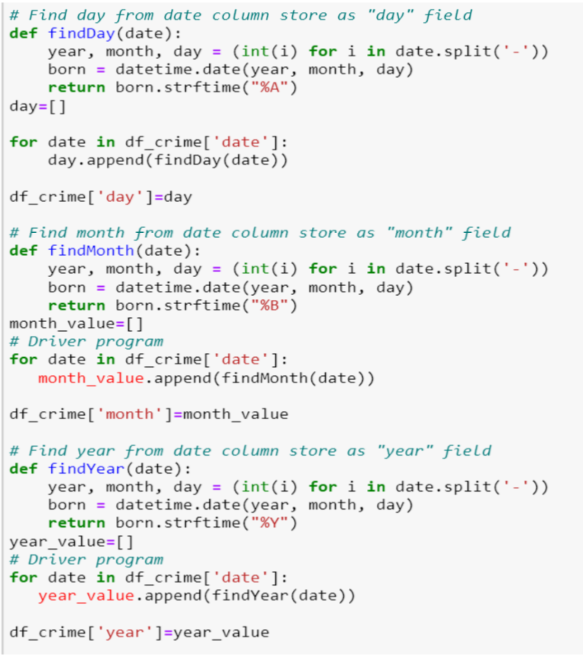
5. Removing or correcting some mismatches in “Male” and “Female” count in United States Gun Violence dataset



```python
# Removing some mismatch values in male and female count.
df_crime.male=(df_crime.male-(df_crime.male-df_crime.total_person))-df_crime.Female
df_crime.loc[df_crime['male'] < 0, 'male'] = 0
df_crime.loc[df_crime['male'] == 0, 'Female']=df_crime['total_person']
```

Fig 24: Removing some mismatch values in male and female count.

6. Using datetime library which day of week, month of year and year are extracted from “date” attribute from United States Gun Violence dataset and stored in “day”, “month” and “year attributes.



```
# Find day from date column store as "day" field
def findDay(date):
    year, month, day = (int(i) for i in date.split('-'))
    born = datetime.date(year, month, day)
    return born.strftime("%A")
day=[]

for date in df_crime['date']:
    day.append(findDay(date))

df_crime['day']=day

# Find month from date column store as "month" field
def findMonth(date):
    year, month, day = (int(i) for i in date.split('-'))
    born = datetime.date(year, month, day)
    return born.strftime("%B")
month_value=[]
# Driver program
for date in df_crime['date']:
   month_value.append(findMonth(date))

df_crime['month']=month_value

# Find year from date column store as "year" field
def findYear(date):
    year, month, day = (int(i) for i in date.split('-'))
    born = datetime.date(year, month, day)
    return born.strftime("%Y")
year_value=[]
# Driver program
for date in df_crime['date']:
   year_value.append(findYear(date))

df_crime['year']=year_value
```

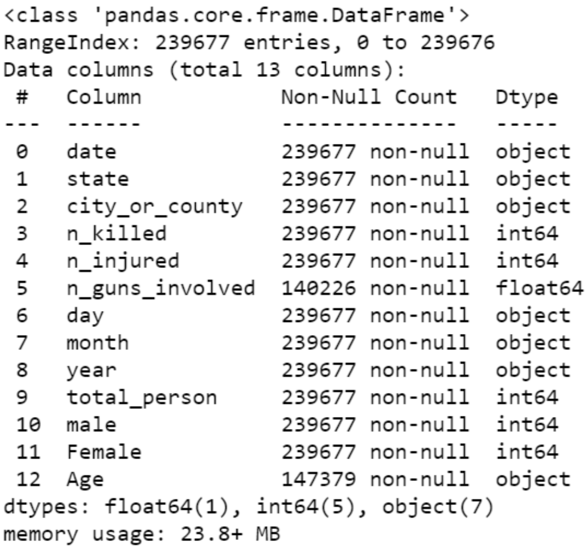
Fig 25: Extracting Day, month and year from "date" attribute.

Note:

- United States Population, Per-Capita, Un-Employment Dataset doesn't need any data cleaning or modification.

After Data cleaning and data modification final dataset that will be used for answering data analytics Questions.

United States Gun Violence Dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 13 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   date             239677 non-null  object 
 1   state            239677 non-null  object 
 2   city_or_county   239677 non-null  object 
 3   n_killed         239677 non-null  int64  
 4   n_injured        239677 non-null  int64  
 5   n_guns_involved  140226 non-null  float64
 6   day              239677 non-null  object 
 7   month            239677 non-null  object 
 8   year             239677 non-null  object 
 9   total_person     239677 non-null  int64  
 10  male             239677 non-null  int64  
 11  Female           239677 non-null  int64  
 12  Age              147379 non-null  object 
dtypes: float64(1), int64(5), object(7)
memory usage: 23.8+ MB
```

Fig 26: United States Gun Violence Dataset

7. Data Exploration and Analysis

7.1 Initial Data Exploration

Is there Relationship between Attributes?

This is to find out whether there is any correlation between the data attributes(fields) of the United States Gun Violence Dataset.

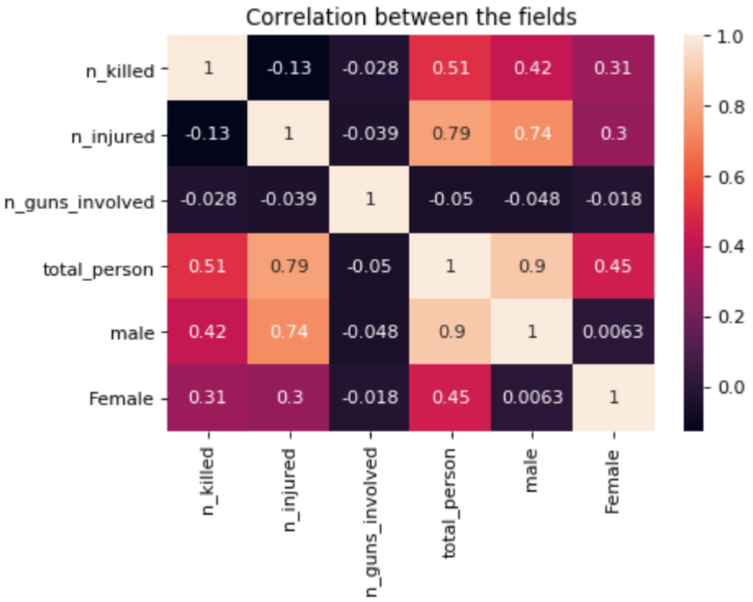


Fig 27: Relationship between the Data attributes.

- There is a high chance of people getting injured than getting killed.
- Male population involved gun violence incidents are more than Female population.
- Proportion of male getting killed and injured are more than female.
- There is no correlation between the Guns involved and total people affected by violence.

What Age Group is most affected in United States Gun violence between 2013 to 2018?

To find which age group is most affected by Gun related violence.

- Gun violence Participants Age Histogram are right skewed, with average Age of 29.
- Median Age of the participants is 26.
- Highest of 11535 participants are aged around 19.
- 46446 people are less than 20 years of age and 40133 people are more than 40 years of age.
- More number are people are age between 20 to 40 with 131795 count.

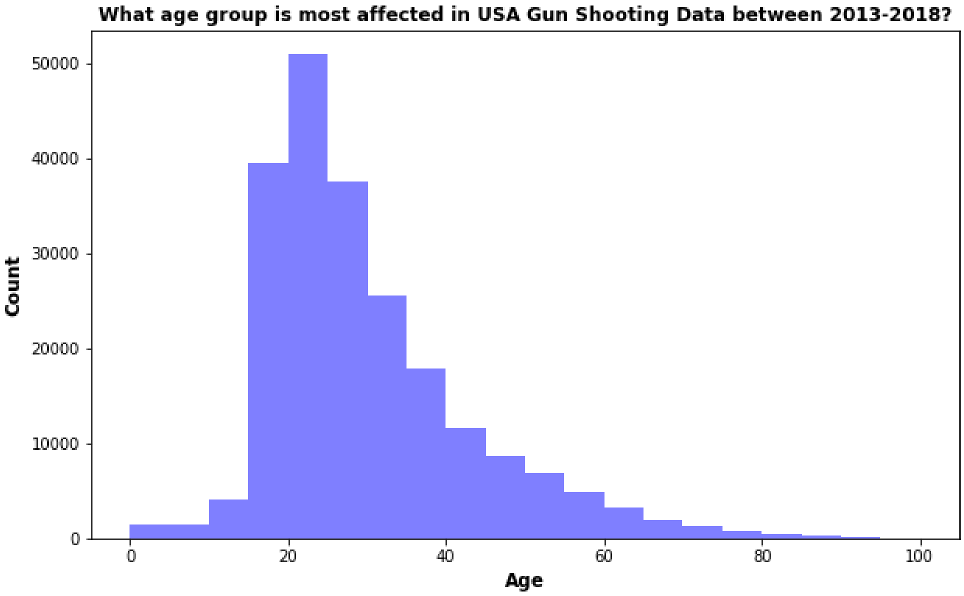


Fig 28: Histogram for participants Age

Which Gender is most affected over the years from 2013 to 2018?

This question helps to answer which gender is most affected in the U.S. due to gun related violence from 2013 to 2018.

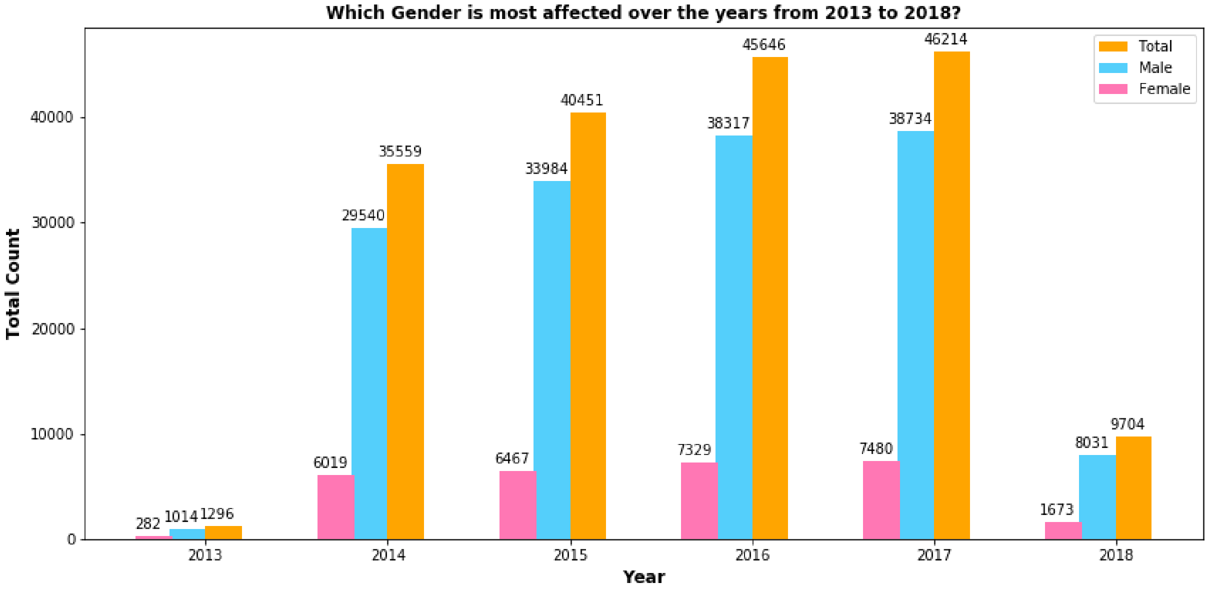


Fig 29: Male and Female population affected by Gun Violence from 2013 to 2018

- It is every evident that Men are more affected than women.
- 2017 has the highest number of victims with 46214 count.

Which U.S. State is more Gun Violence Victims over the years from 2013 to 2018?

Based on the Total victims U.S. States are ranked to seriousness of the Gun violence in each State.

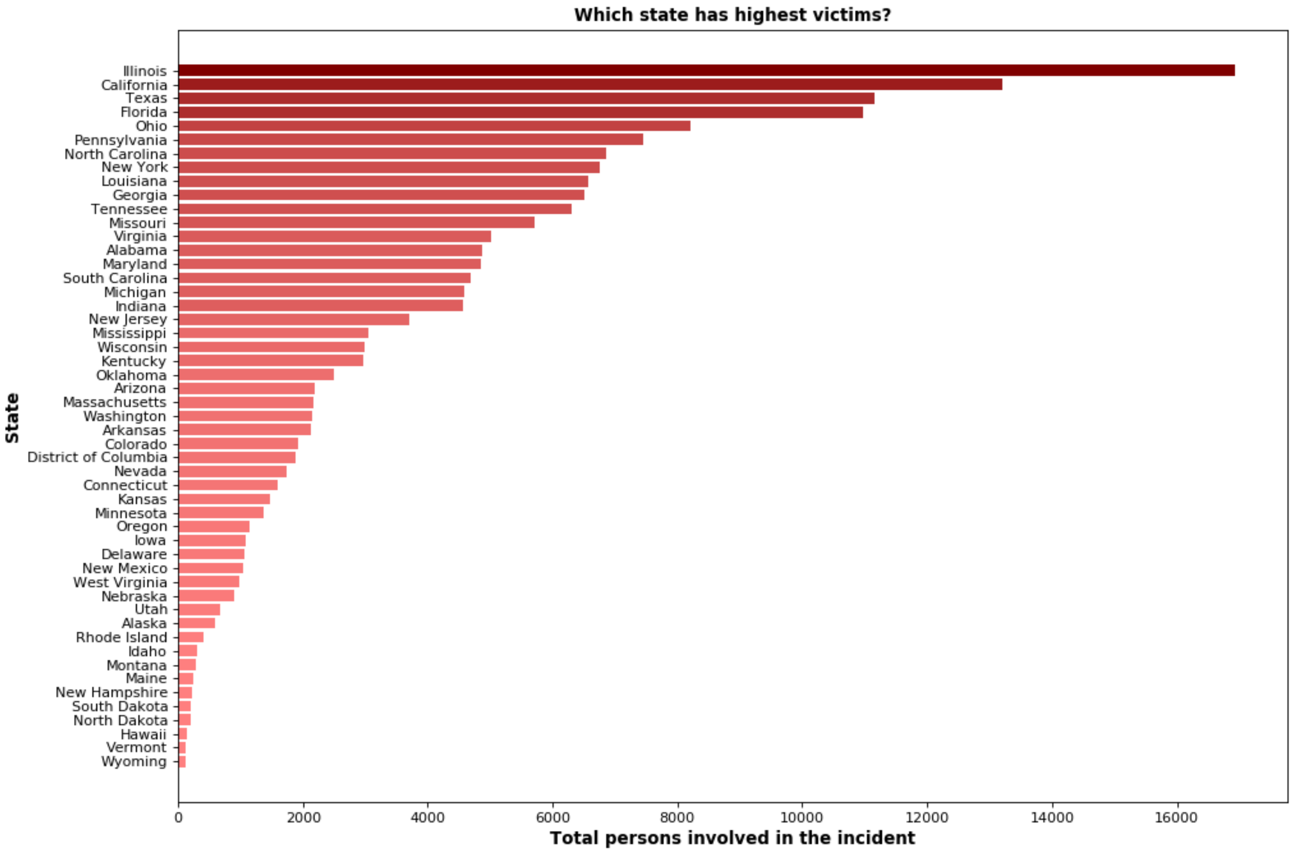


Fig 30: U.S. States Ranking on Total Victim Count.

- Ranking is based on the total victims not based on any other attributes.
- Illinois ranked 1 with the highest victim count 16923 and Wyoming ranked 51 has least number of victims 125.

Which Top 3 U.S. state has Highest and Lowest Kills, Injured and Incidents?

An interesting findings are found from our Gun violence data from 2013 to 2018:

- **Top 3 U.S. States with highest Fatalities (Killed).**
  1. California – 5562
  2. Texas – 5046
  3. Florida – 3909
- **Top 3 U.S. States with lowest Fatalities (Killed).**
  1. Vermont – 57
  2. Rhode Island – 63
  3. Hawaii – 63
- **Top 3 U.S States with highest non-fatal victims (Injured).**
  1. Illinois – 13514
  2. California – 7644
  3. Florida – 7072
- **Top 3 U.S States with lowest non-fatal victims (Injured).**
  1. Wyoming – 52
  2. Vermont – 73

3. Hawaii – 85

- **Top 3 U.S. States with highest Incidents.**
  1. Illinois – 17556
  2. California – 16306
  3. Florida – 15029
- **Top 3 U.S. States with lowest Incidents.**
  1. Wyoming – 125
  2. Vermont – 130
  3. Hawaii – 148

7.2 K-means Cluster Analysis

Which USA states are highly dangerous and safe from Gun violence?

Based on K-means clustering 51 U.S. states are divided into 5 cluster involving incidents and total victims to find out which State has high risk of Gun violence.

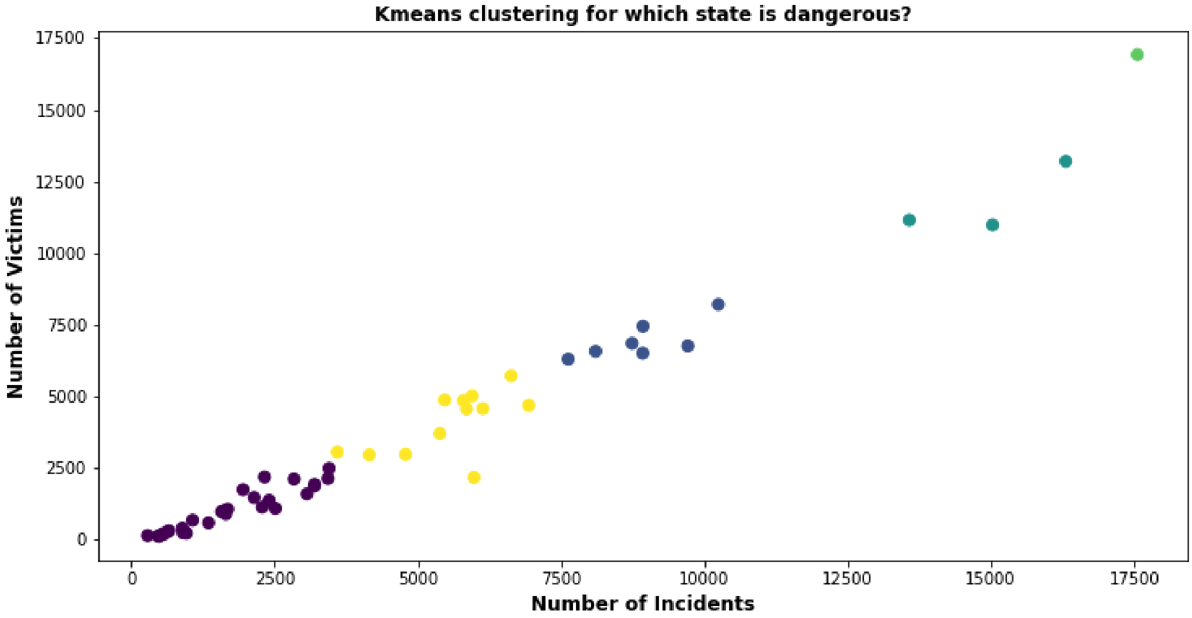


Fig 31: K-means Clustering for U.S. States.

- Cluster 0 consist of 28 states which are the safe states with less incidents and victims.
- Cluster 1 and 4 consists of 7 and 12 states respectively which are moderately safe.
- Cluster 2 consists of 3 U.S. States which are dangerous.
- Cluster 3 has 1 U.S States which is the most dangerous compare to other 50 states.

Which USA Cities are Dangerous and safe from Gun violence?

Based on K-means clustering U.S. Cities are divided into 3 cluster involving incidents and total victims.

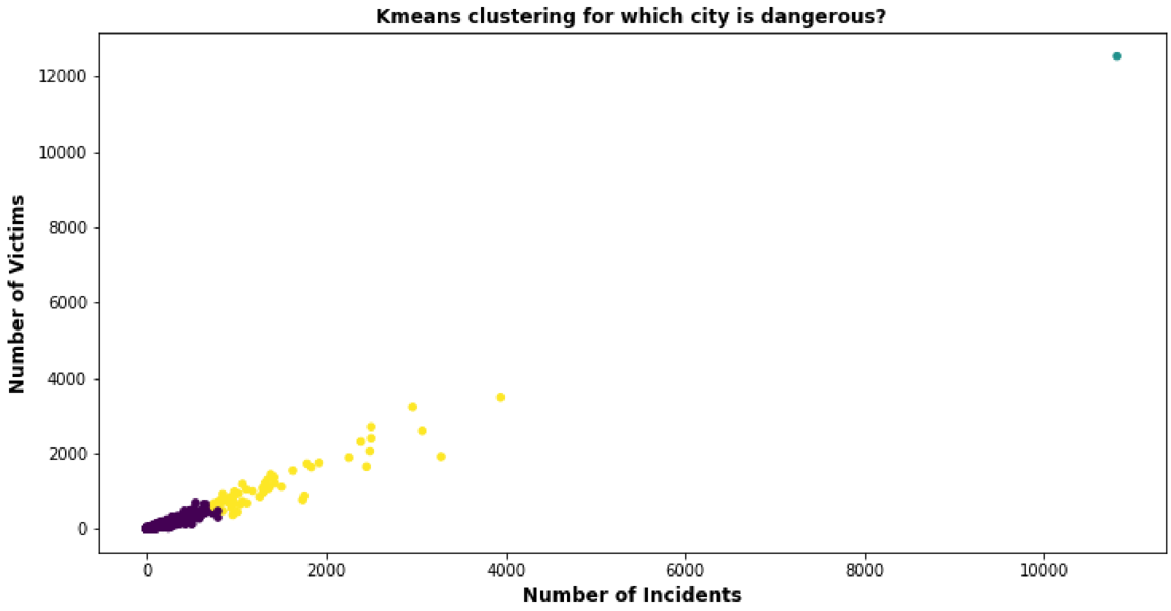


Fig 32: K-means Clustering for U.S. Cities.

- Cluster 0 consist of 12842 cities which are the safe with less incidents and victims.
- Cluster 2 consists of 55 cities which are moderately safe.
- Cluster 1 consists of Chicago city which has very high incident and victims.

Chicago, Baltimore and Philadelphia cities have highest number of victims. Chicago, Baltimore and Washington cities have highest number of Gun violence incidents.

7.3 Hierarchical Cluster Analysis.

Which Day of week has high chances of Gun violence?

One of the objectives of our Report is to find which Day of week is most dangerous from Gun related violence.

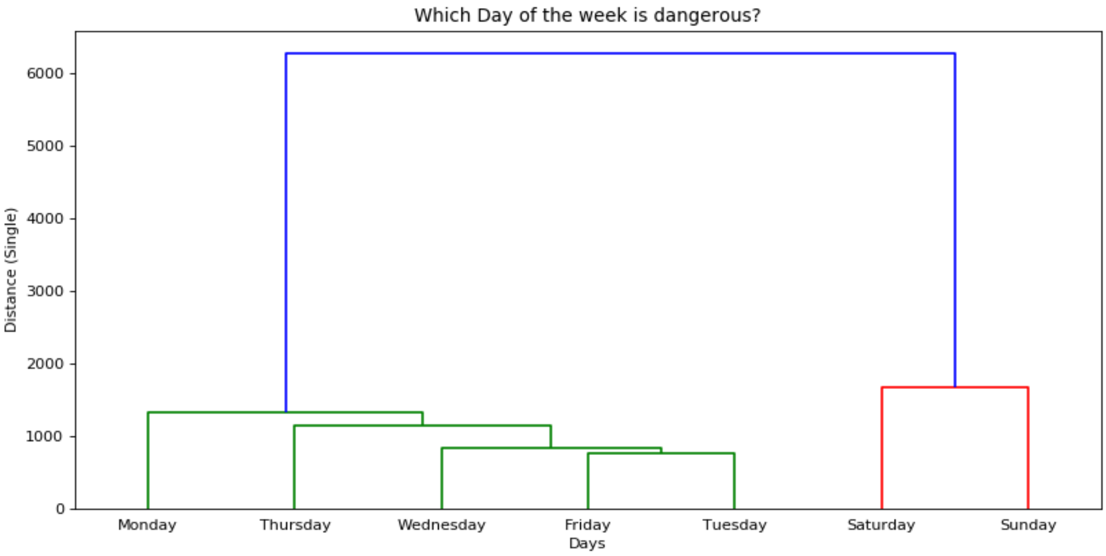


Fig 33: Linkage Clustering for Dangerous Day of week.

Dendrogram shows that weekends are dangerous when compared to weekdays. We can’t answer whether Sunday or Saturday is most dangerous among the two and similarly with Safe day whether Friday or Tuesday. To answer this Risk score is generated for each day in *Chapter 8.1*.

Which Month of year has high chances of Gun violence?

One of the objectives of our Report is to find which m of week is most dangerous from Gun related violence.

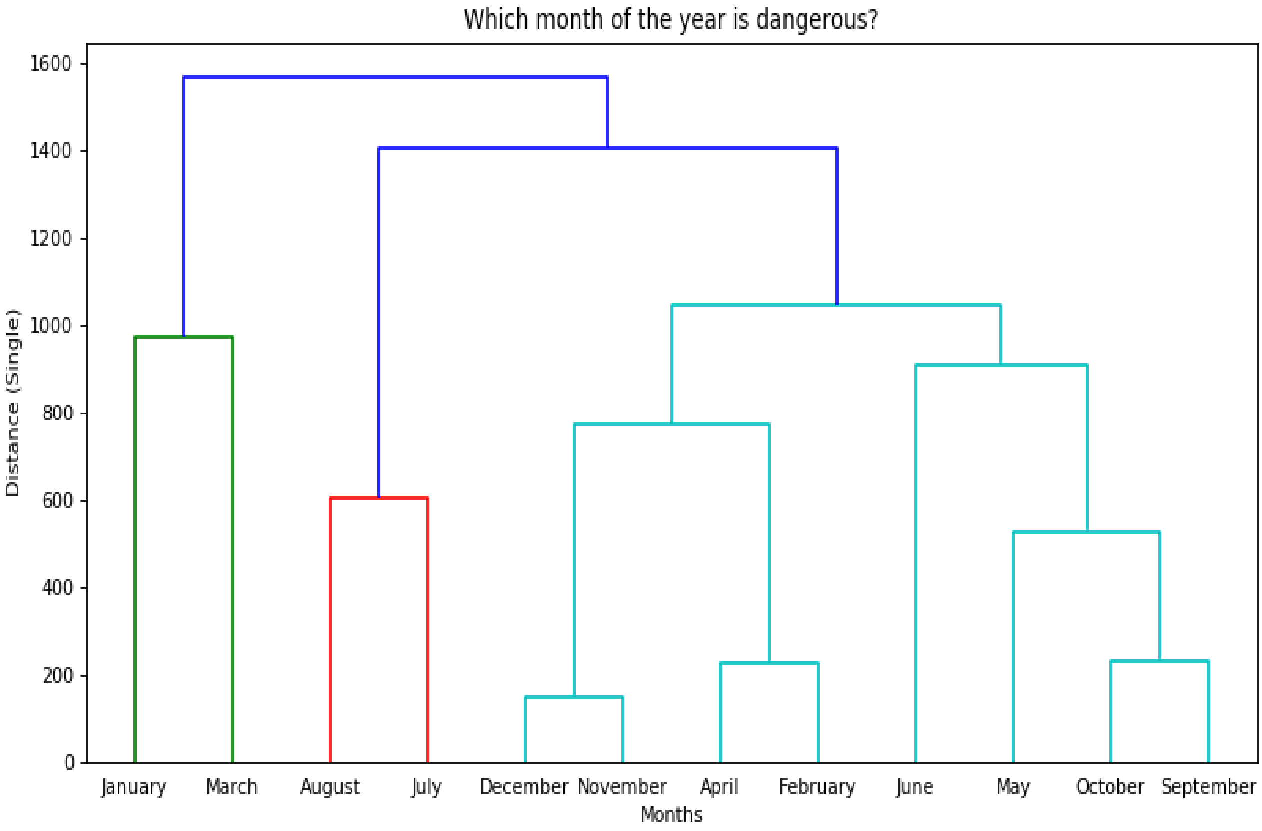


Fig 33: Linkage Clustering for Dangerous Month of year.

December and November are two most secure months among the 12 months. January and March months are two dangerous months amongst 12 months. Risk score that is generated in Chapter 8.2 to answer which among two months are safe and dangerous.

7.4 United States Un-Employment, Per-Capita and Gun Violence Death Rate Data Analysis.

Which U.S. States has Un-Employment Rate more than National Average Rate?

This question helps us to understand the importance of this attribute on the Gun violence.

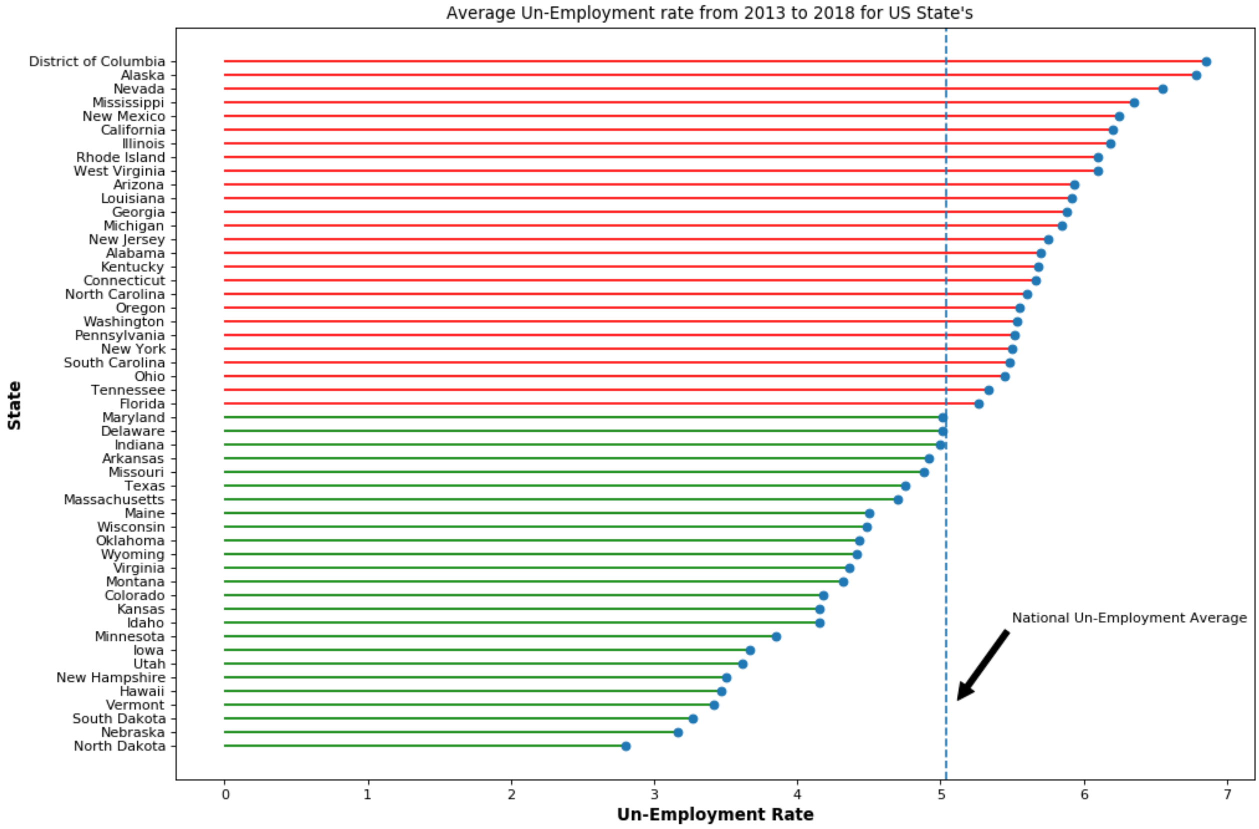


Fig 34: Average Un-Employment Rate from 2013 to 2018.

Average National Un-Employment from 2013 to 2018 is about 5.03. Around 26 U.S. States are more than National Average with District of Columbia and Alaska being Top States. There are 25 States with less than National Average and North Dakota having least Un-employment Rate. U.S. Un-Employment Rate will be used to generate the Risk Score. Finding out whether the Un-Employment rate has an influence on the Gun Violence in U.S. states is very important.

Which U.S. States has Per-Capita Income Rate more than National Average Rate?

U.S States average per-capita income from 2013 to 2018 are used to find importance of this data in our analysis.

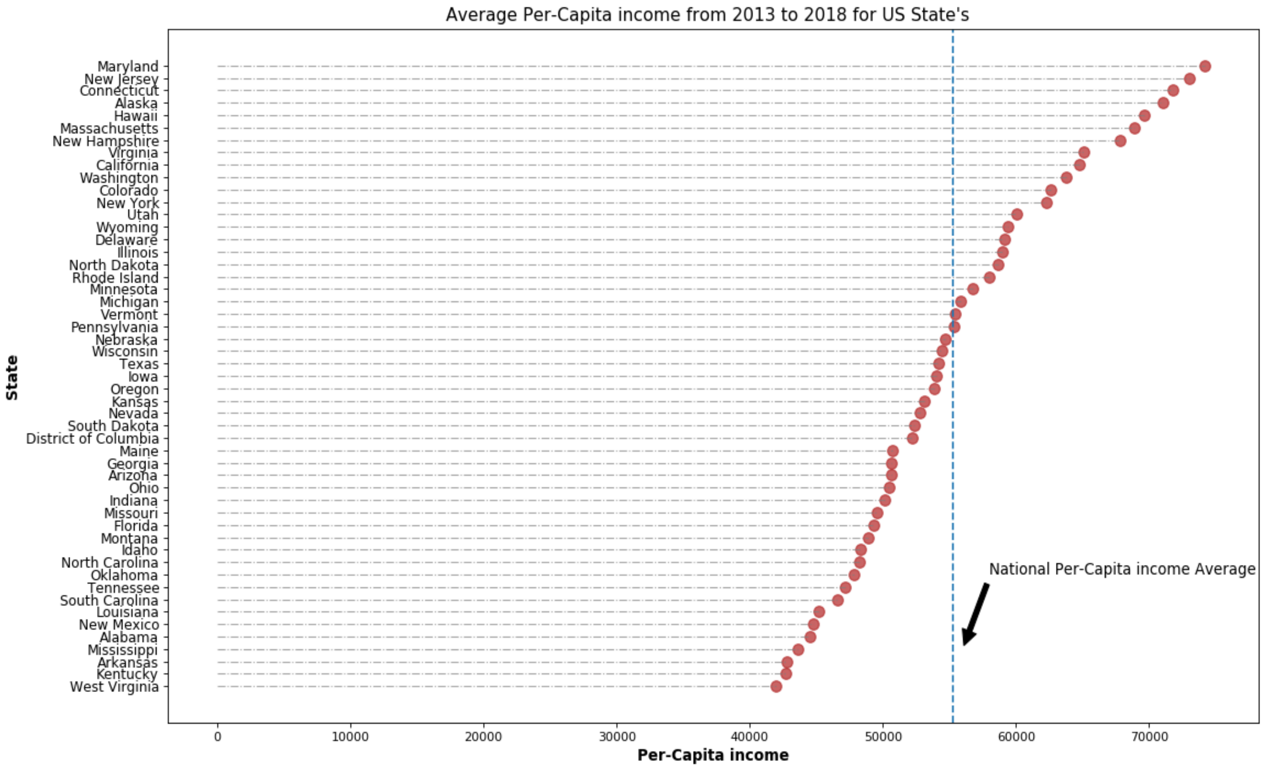


Fig 35: Average Per-Capita Income Rate from 2013 to 2018.

National Average per-capita income from 2013 to 2018 is around 55265.91. Maryland, New Jersey and Connecticut are the top 3 states with highest per-capita income. West Virginia, Kentucky and Arkansas bottom 3 states less than National average.

Which state is safer based on death rate for every 100,000 population?

Involving the population attribute gives us better understanding to find out which U.S. state is safe and dangerous. There can be a greater number of Gun violence incidents and fatalities in Illinois, Florida, California and Texas States, it is common to here about mass shooting incidents in these States. Only involving the population and calculating Death rate for every 100,000 population will answer the question.

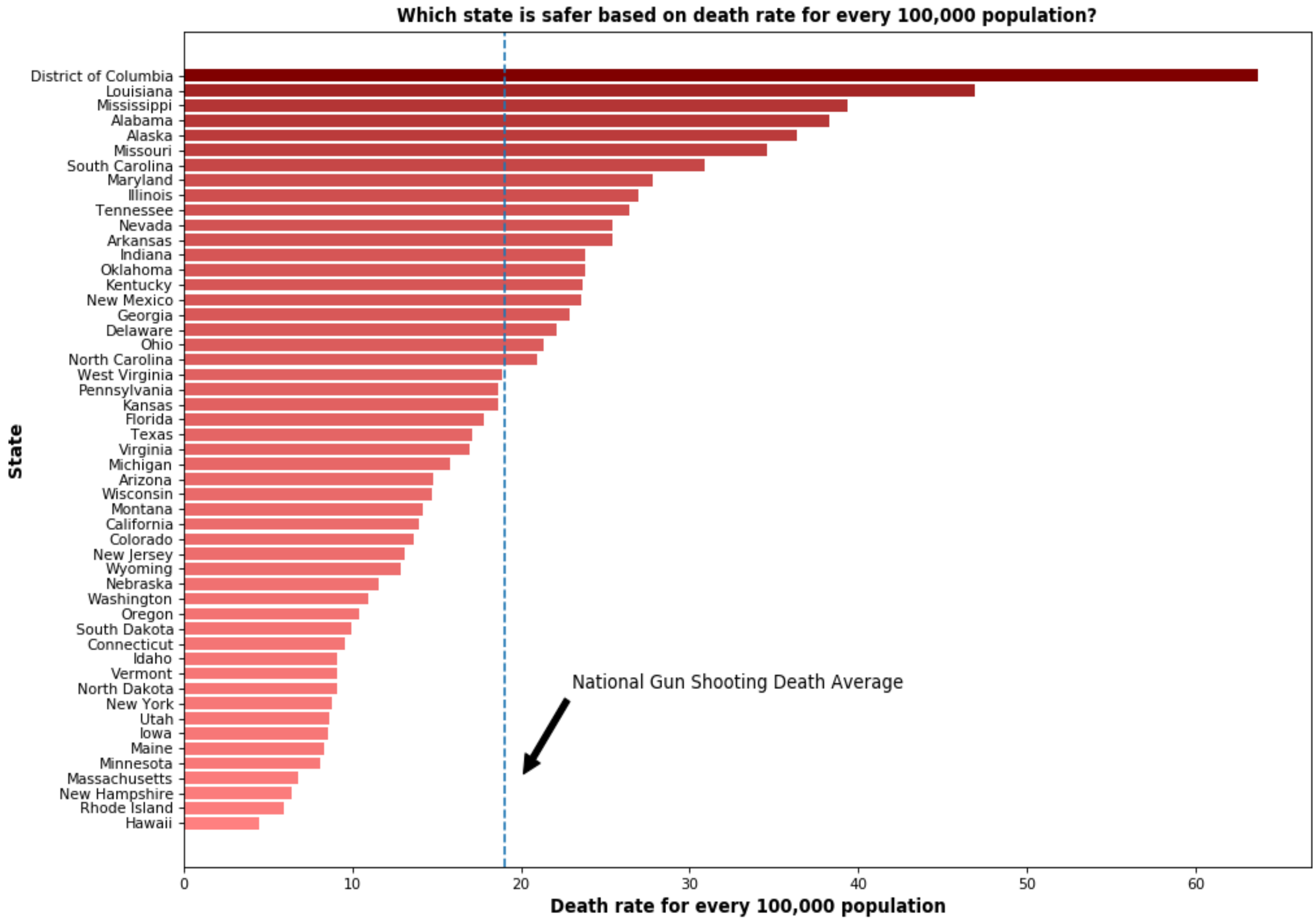


Fig 36: Death Rate for every 100,000 population.

- National Gun violence Death rate is 19.03.
- District of Columbia has 63.6892 as highest death rate from 2013 to 2018.
- Hawaii has least death rate for every 100,000 population with 4.45.

Note:

It is evident from Un-employment rate, Per-Capita income and Death rate analysis

- **If U.S. State has higher Un-employment rate than national average, then they are having less Death rate than National average.**
- **If U.S. State has higher per-capita income than national per-capita income average then death rate is less than national death rate.**
- **Based on this analysis we will calculate Risk score ranking by providing high weightage to Death Rate, Per-capita income and Un-Employment rate.**

7.5 Confidence Interval Analysis.

This analysis is based on the 95% confidence interval to find the range for the Death Rate, Un-Employment and Per-Capita income.

- Confidence Interval for Un-Employment Rate based on the U.S. National Gun violence Death rate for every 100,000 people.

States having more than National Death Rate?	95% Confidence Interval	
	Minimum Un-Employment rate	Maximum Un-Employment rate
Yes (If States are more than National Death Rate)	5.27	6.08
No (If States are less than National Death Rate)	4.19	5.13

- Confidence Interval for Per-Capita Income based on the U.S. National Gun violence Death rate for every 100,000 people.

States having more than National Death Rate?	95% Confidence Interval	
	Minimum Per-Capita Income	Maximum Per-Capita Income
Yes (If States are more than National Death Rate)	46467.1	56681.9
No (If States are less than National Death Rate)	54391	61519.9

- Confidence Interval for Gun Violence Death Rate for every 100,000 people based on the U.S. Per-Capita Income.

States having more than National Per-Capita Income?	95% Confidence Interval	
	Minimum Death Rate	Maximum Death Rate
Yes (If States are more than National Per-Capita Income)	9.58	19.22
No (If States are less than National Per-Capita Income)	16.94	29.24

- Confidence Interval for Gun Violence Death Rate for every 100,000 people based on the U.S. Un-Employment Rate.

States having more than National Un-Employment Rate?	95% Confidence Interval	
	Minimum Death Rate	Maximum Death Rate
Yes (If States are more than National Un-Employment Rate)	16.88	30.93
No (If States are less than National Un-Employment Rate)	10.86	18.83

8. Risk Score Generation

Risk Score are generated based on providing proper weightage to attributes so that a proper answer is derived for our questions.

8.1 Risk Score on Day of Week

Which day of the week is most unsafe for Gun Violence?

Objective of this question is to find out which Day of week is more prone to Gun Violence. This answer the questions raised in our dendrogram cluster analysis.

To answer this question Risk score is generated by giving proper weightage to attributes. We are giving below weightage for the attributes.

- 15% weightage to "Number of non-fatal incidents".
- 30% weightage to "number of Fatal incidents".
- 30% weightage to "Number of people got killed".
- 15% weightage to "Number of people got injured".
- 10% weightage to "Number of Guns involved".
- Total weightage accounts 100% to all the attributes.

This weightage is given based on the impact of each attributes and relationship with the Gun violence. Higher weightage is given to fatal incident and less to non-fatal incidents.

Note: Based on the studies from initial analysis from Chapter 7, weightage is given based on the importance of the attribute and impact on our analysis.

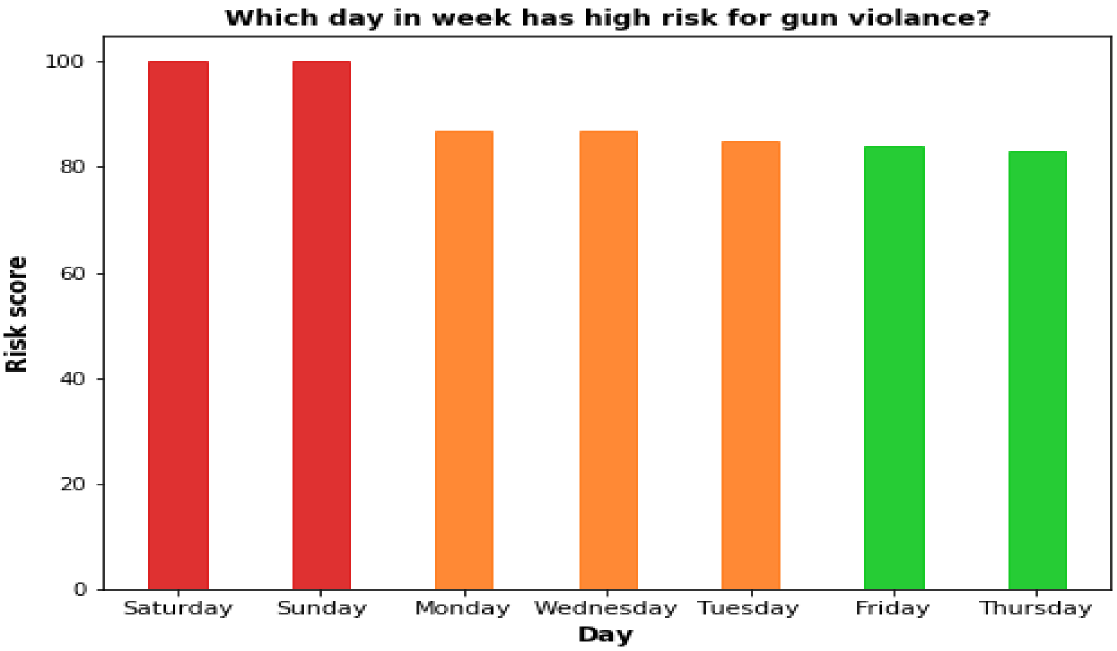


Fig 37: Risk Score for Day of Week.

- Weekends have high chance of Gun violence. Saturday is the most dangerous than Sunday.
- Thursday seems to be safer than compared to other days.

- Monday, Wednesday and Tuesday have moderate risk of Gun violence.

8.2 Risk Score on Month of Year

Which Month of the Year is most unsafe for Gun Violence?

Objective of this question is to find out which month of year is more prone to Gun Violence. This answer the questions raised in our dendrogram cluster analysis.

To answer this question Risk score is generated by giving proper weightage to attributes. We are giving below weightage for the attributes.

- 15% weightage to “Number of non-fatal incidents”.
- 30% weightage to “number of Fatal incidents”.
- 30% weightage to “Number of people got killed”.
- 15% weightage to “Number of people got injured”.
- 10% weightage to “Number of Guns involved”.
- Total weightage accounts 100% to all the attributes.

This weightage is given based on the impact of each attributes and relationship with the Gun violence. Higher weightage is given to fatal incident and less to non-fatal incidents.

Note: Based on the studies from initial analysis from Chapter 7, weightage is given based on the importance of the attribute and impact on our analysis.

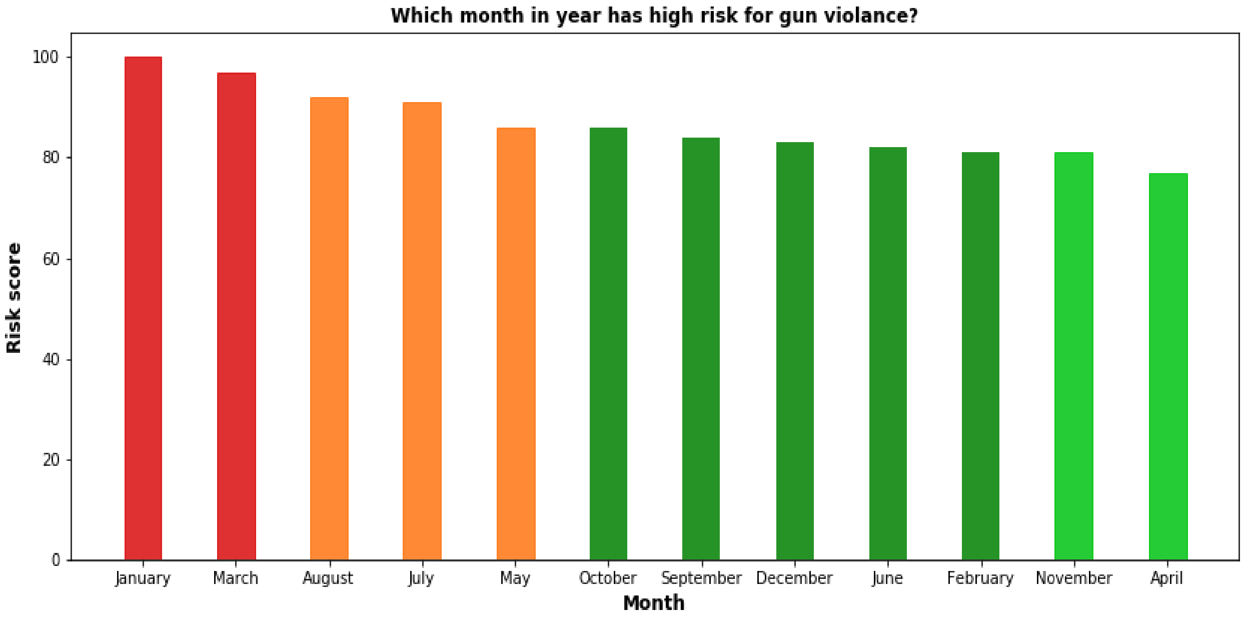


Fig 38: Risk Score for Month of Year.

- January has high chance of Gun violence compared to other months.
- April is the safest month compared to other months.

8.3 Risk Score Ranking on U.S. States without Involving Socio-economic Attributes.

Which U.S. State is safer from Gun violence?

To answer this question Risk score is generated by giving proper weightage to attributes. We are giving below weightage for the attributes in Gun Violence Dataset.

Note: We are not involving any socio-economic attributes.

- 15% weightage to “Number of non-fatal incidents”.
- 30% weightage to “number of Fatal incidents”.
- 30% weightage to “Number of people got killed”.
- 15% weightage to “Number of people got injured”.
- 10% weightage to “Number of Guns involved”.
- Total weightage accounts 100% to all the attributes.

This weightage is given based on the impact of each attributes and relationship with the Gun violence. Higher weightage is given to fatal incident and less to non-fatal incidents.

Note: Based on the studies from initial analysis from Chapter 7, weightage is given based on the importance of the attribute and impact on our analysis.

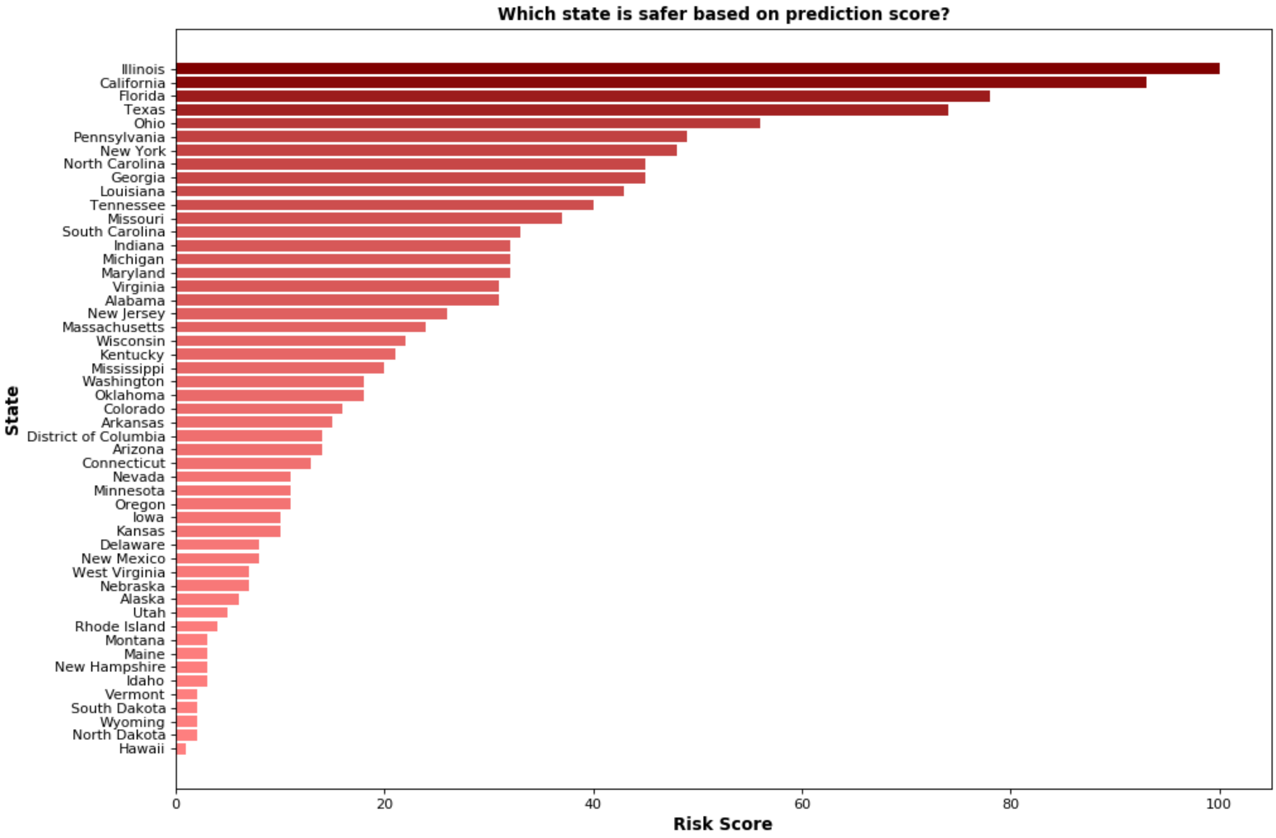


Fig 39: Risk Score Rank for U.S. State without involving Socio-Economic Attributes

- Based on the Risk score generated the U.S. States are ranked.
- There is a high chance of Gun Violence in Illinois and California.

- Hawaii ranked 51 and it has lesser chance of Gun related violence.

Since we are involving other attributes like “Number of Fatal incidents”, “Number of non-fatal incidents” and “Guns involved” along with total victims in incidents there is a change in ranking from the U.S. State rank based on only total victims as in *Fig 30.*

From *Fig 39:*

- 26 State moved up in the ranking to have high chance of the Gun violence. This is due to Risk score ranking by giving weightage to the attributes.
- Massachusetts ranks 20 and seems to be much more dangerous than previously by just taking victims into account.
- Around 17 States moved down the ranking and seems to be much safer than previous ranking.
- Arizona seems too much safer with ranking dropping from 24 to 29.

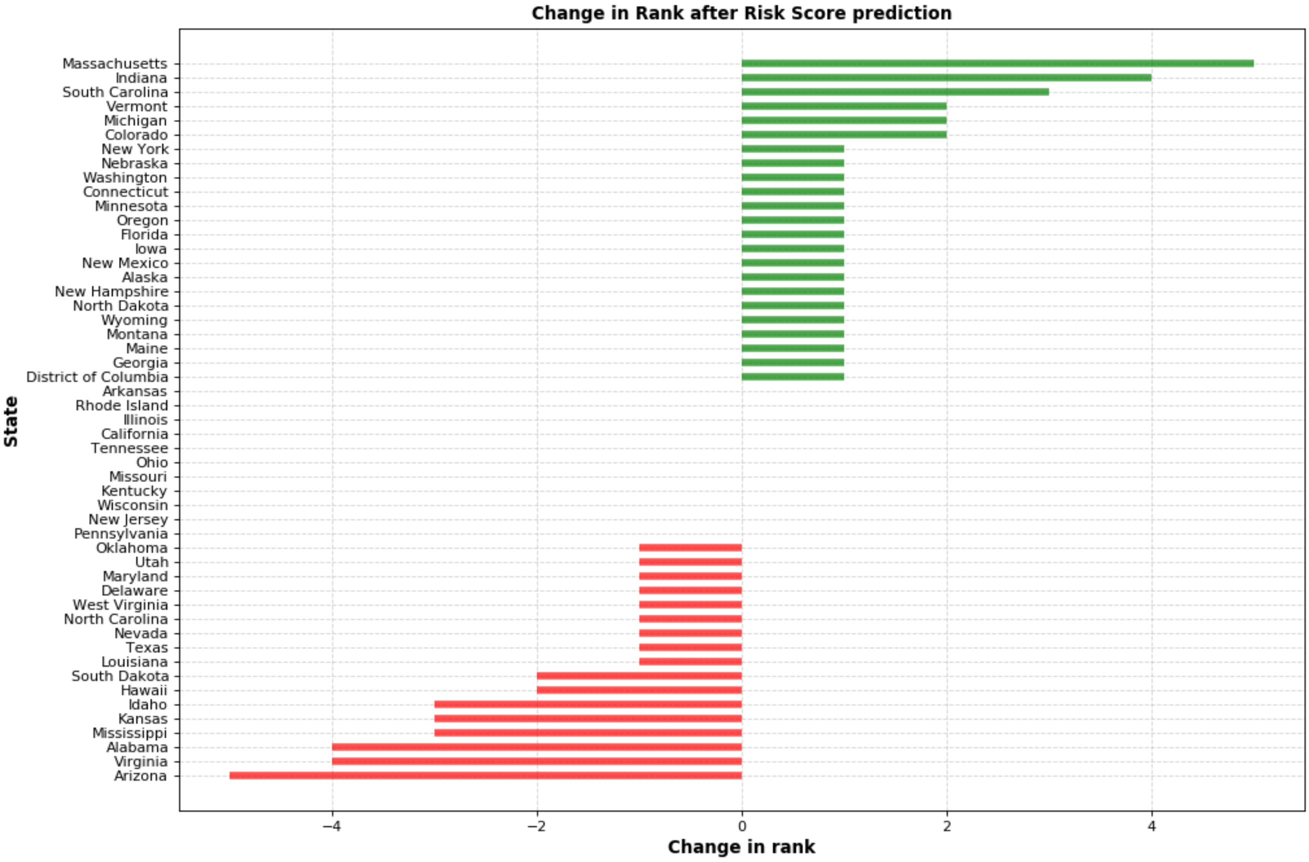


Fig 40: Change in Rank from “*Fig 30*” Ranking.

This Risk score ranking, and rank change tells about the importance of each attribute since we are giving high weightage to fatal incidents and number of people getting killed in each U.S. States

If we just take in consideration of only number of people killed or incidents, we can’t tell how safe or dangerous is the U.S. states are from Gun related violence.

Next, we are considering socio-economic attributes to give a better picture on the U.S States Gun Violence Safety.

8.4 Risk Score Ranking on U.S. States Involving Socio-economic Attributes.

Which U.S. State is safer from Gun violence?

To answer this question Risk score is generated by giving proper weightage to attributes. We are giving below weightage for the attributes in Gun Violence Dataset and Socio-economic Attributes.

Note: We are involving any socio-economic attributes.

- 5% weightage to “Number of non-fatal incidents”.
- 10% weightage to “number of Fatal incidents”.
- 25% weightage to “Death Rate for every 100,000 population”.
- 5% weightage to “Number of people got injured”.
- 5% weightage to “Number of Guns involved”.
- 25% weightage to “Per-capita Income”
- 25% weightage to “Un-Employment Rate”
- Total weightage accounts 100% to all the attributes.

This weightage is given based on the impact of each attributes and relationship with the Gun violence. Higher weightage is given to death rate, Un-employment, Per-Capita Income and less to non-fatal incidents and Guns involved.

Note: Based on the studies from initial analysis from Chapter 7, weightage is given based on the importance of the attribute and impact on our analysis.

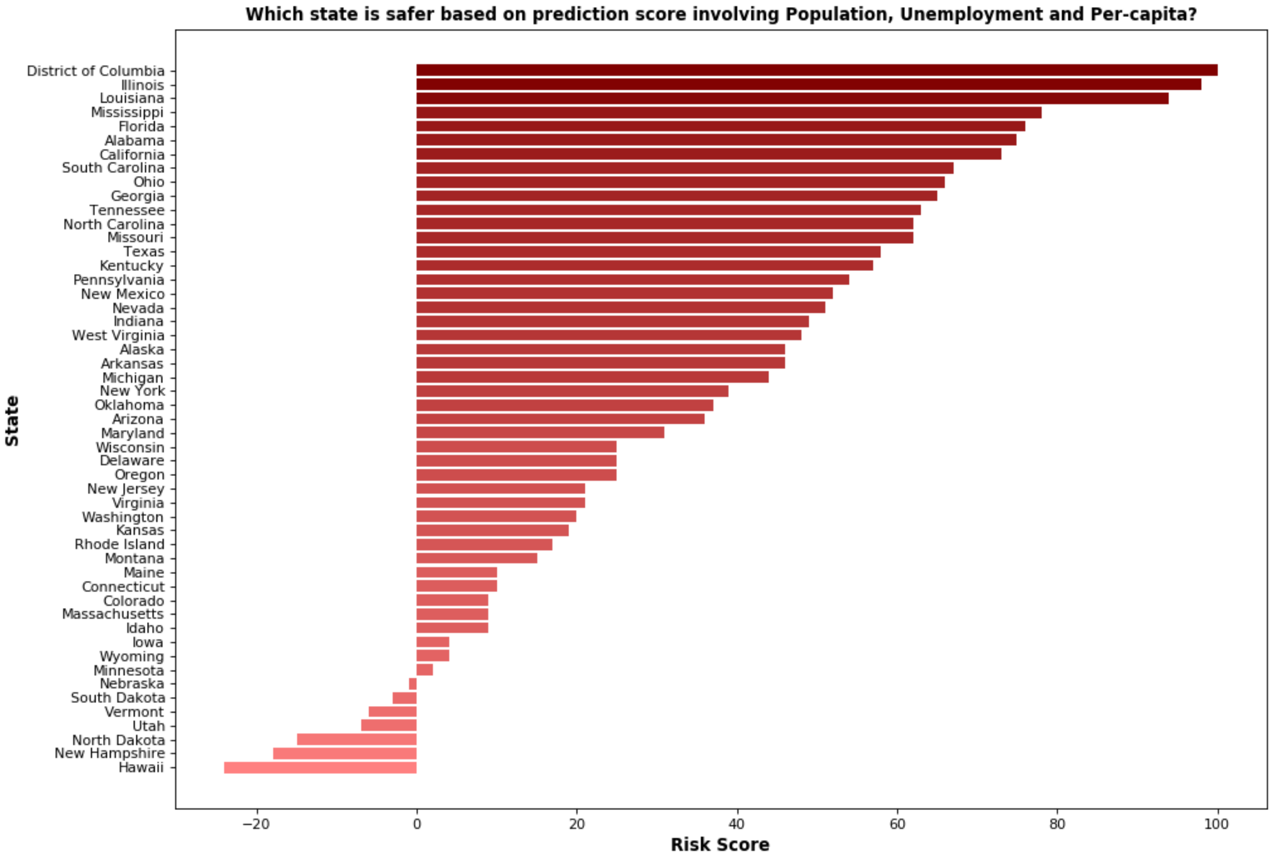


Fig 41: Risk Score Rank for U.S. State involving Socio-Economic Attributes.

Based on the study conducted on the school massing shooting by Adam Pah, professor at Northwestern University it had been found that there's a correlation between Un-Employment and Gun Shooting. The correlation among economic uncertainty and gun violence in schools adds a disastrous dimension to what's already known about the influence of unemployment and financial tension on families. Economic uncertainty increases feelings of frustration, rage, anxiety, depression, further as a scarcity of safety. Now, this research reveals that economic security, with lost confidence and reduced prospects for employment, may connect on to gun violence.

Hence, keeping in this in mind we are involving Un-Employment and Per-Capita Income attribute to predict how the U.S. States are safer from Gun related violence.

- District of Columbia has highest death rate for every 100,000 population and Highest Un-Employment hence, it is Ranked 1st in our Risk Score Ranking involving Un-employment, Per-Capita Income.
- Hawaii stands at 51st rank as the death rate and un-employment rate are very low. On top of it per-capita income more than the national average.
- States with higher death rate than national average has moved up, along with that if State has lower un-employment rate than national average that U.S, State moves further up the Ranking.
- If State having higher per-capita income only than the State moves down the ranking to be much safer.

This influences on the Ranking of U.S. States for the safety of Gun related violence.

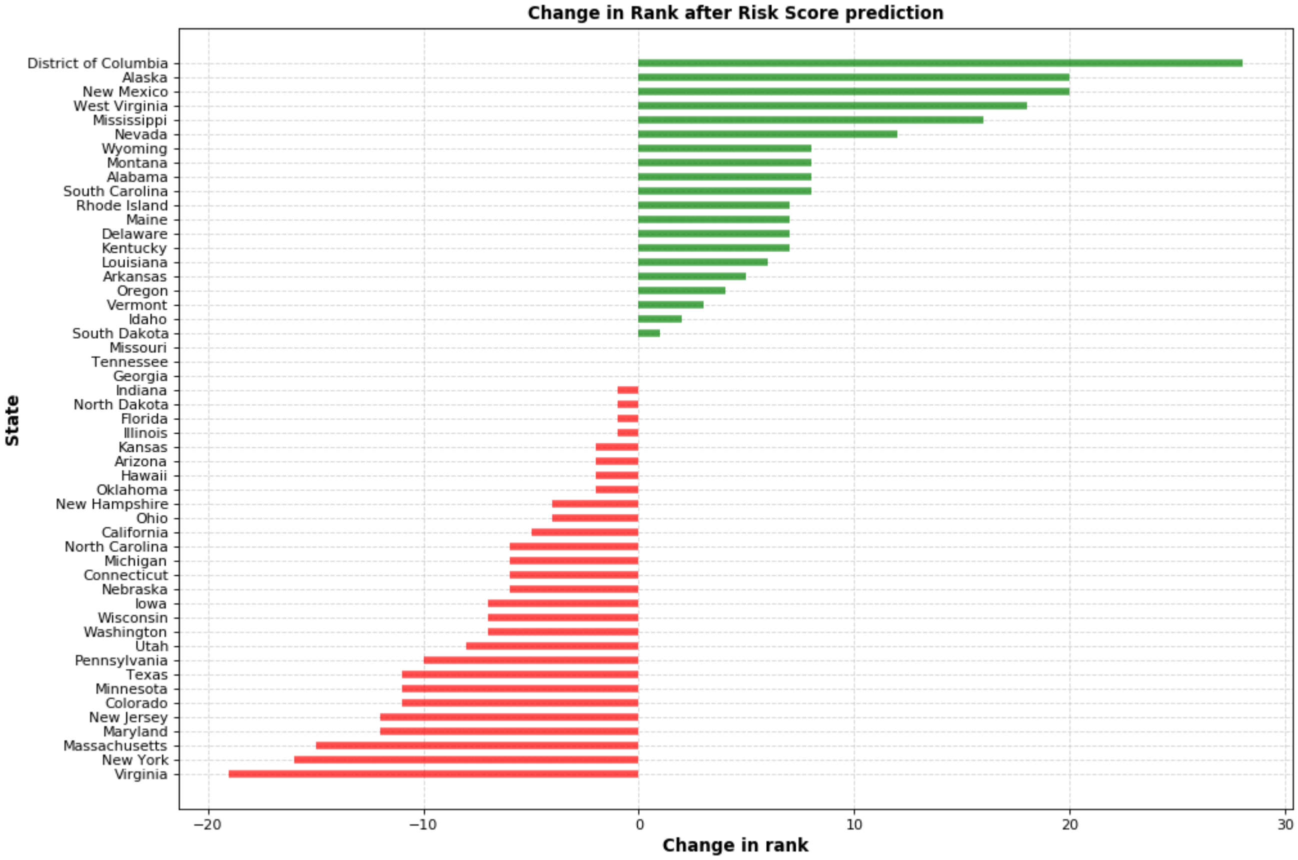


Fig 42: Change in Rank from "*Fig 30*" Ranking.

- To a greater extend District of Columbia have moved up by 28 ranks to come on top of the ranking list and be the most unsafe U.S. State from Gun related violence.
- Virginia has moved down to 32nd rank and it is much more safer U.S. State than earlier ranking based on the victims.
- Georgia, Tennessee and Missouri are the only three U.S. States which didn’t see any change in their ranking even after considering Death rate, Un-Employment and Per-Capita Income.

9. Conclusion and Future Work.

9.1 Conclusion

Objective of our data analysis report has been achieved from our data collection, data cleaning, data science methodologies and Risk score ranking, We have answered many basic questions like which Age group, which Gender is most affected and finding out if there is any correlation between incidents, victim and Guns involved from our data analysis. Majority of our victims are aged 19 and average age of victims is 29. There are around 131795 victims aged between 20 to 40 which is high when compared to other age group. There is a high chance for the victim to be a Male when a person sufferer from Gun violence. With a higher level of probability to get injured than getting killed if anyone gets involved in the incident. Apart from the basic questions we are focusing on which City, State, Day of Week and Month of year is safe from Gun related violence. Based on the data science method K-means clustering U.S. cities are divided into 3 clusters. Chicago has the highest number of Gun shooting incidents and victims, and it is way higher than the second most dangerous city that is Baltimore. Dendrogram shows that weekdays are safer than weekends and from the risk score ranking it is clear than Saturday is most dangerous and Thursday being safest day of the week. This difference may be due to more people spend their time outside places like restaurant, pubs, clubs, malls, shopping center, and many more. Shooting at workplace and educational institutes may account for majority of the incidents in weekdays. Both dendrogram and Risk score ranking on the gun violence data January and March are two most dangerous month for Gun related incidents, this may be due to high number of holiday's like New year on January 1st, January 15 Martin Luther King, Jr. Day, Epiphany on January 6, Orthodox day on January 7 and holiday season. Even March has many holidays like Palm Sunday, St. Patrick's Day, Good Friday and Easter. There are 13 holidays in January and more than 20 holidays in March some of which varies in each U.S. States may be reasons for these high Gun related violence. Summer months May, June, July and August can be equally dangerous, June is much safer compared to other summer months. April is much safer out of 12 months.

Main focus of our analysis is to predict which U.S. State is more vulnerable to Gun related incidents. Based on the initial ranking based on total victim count in each U.S. State, Illinois, California and Texas are Top 3 States with high victims and Wyoming, Vermont and Hawaii are the bottom 3 States which are having less victim count. Next Risk score ranking involving only attributes related to Gun violence dataset and with proper weightage given to all the attributes, Illinois, California and Florida are Top 3 and Hawaii, North Dakota and Wyoming are bottom 3 State for Gun safety. In this initial ranking based on total victims and risk score ranking involving attributes only from Gun violence dataset without involving basic parameter like population of U.S States to gauge the severity of these violence in each State for every 100,000 people which gives the Death rate and without involving the population parameter the results based only Gun violence dataset is inappropriate. After including population attribute and finding the death rate for every 100,000-population district of Columbia has the highest death rate with 63.68 from 2013 to 2018 and Hawaii seems to be safest with 4.45 death rate. Many studies based on the massing shooting has found correlation between Un-Employment and Per-Capita Income, with our risk score ranking involving this socio-economic attributes District of Columbia is most dangerous which moved up by 28 ranks since is has Highest Un-employment rate and low per-capita income. Hawaii is still the safest even after considering the socio-economic attributes. Confidence interval gives us a clear correlation in our data how the high Un-Employment rate may lead to higher Gun violence and with high per-capita income there is less chance of gun related incidents. People who are employed and have high per-capita income will increases sense of security, confidence and safety which leads to a better mental health. Gun violence can be reduced with a stricter Gun laws to an extent but improving the economic condition of an individual or a family is very important than need for a Stricter Gun laws.

9.2 Future Work

In this report we have only focused on United States Gun violence dataset with 260k entries from 2013 to 2018 and in addition to this we included population, Un-Employment and Per-capita Income to answer our analysis research queries.

We can do further analysis by including additional Gun violence data before 2013 and after 2018 to find out interesting and new findings to answer and accurately predict how serious is Gun violence. Apart from this we can do analysis with additional attributes like Gun permit law, suicide rate and educational details. Analysis on the type of location involved in the violence whether pubs, clubs, schools, malls, etc. which one is most dangerous and type of gun incident whether it is suicide, homicide, mass shooting, etc. Overall seriousness of Gun violence around the world can be analyzed in our future work.

10. Reference

[1]. The roots of gun violence in the United States, Brian Ward, “https://isreview.org/issue/109/roots-gun-violence-united-states”

[2]. Gun Violence: Facts and Statistics, Children’s Hospital of Philadelphia, “https://injury.research.chop.edu/violence-prevention-initiative/types-violence-involving-youth/gun-violence/gun-violence-facts-and#.Xx7dDp5KhnJ”

[3]. Gun Violence in America, Everytown research, https://everytownresearch.org/wp-content/uploads/2018/07/Gun-Violence-in-America-CDC-Update-022020C.pdf

[4]. Python vs R for Data science, Medium, “https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197”

[5]. Python Libraries, GeeksforGeeks, “https://www.geeksforgeeks.org/python-programming-language/”

[6]. Nagesh Singh Chauhan, kdnuggets, “https://www.kdnuggets.com/2019/09/hierarchical-clustering.html”

[7]. In depth: K-means Clustering, jakevdp, “https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html”

[8]. United States Gun violence Dataset, Kaggle, “https://www.kaggle.com/jameslko/gun-violence-data”

[9]. United States population, Worldatlas, “https://www.worldatlas.com/articles/us-states-by-population.html”

[10]. Per-capita income and Un-employment rate, Statista, “https://www.statista.com/”

[11]. Google API, Google Maps API. “https://developers.google.com/maps/documentation/javascript/overview”

11. Appendix

11.1 GitHub Repository

Geovisualization GitHub repository

Geovisualization on U.S. Gun violence Dataset from 2013 to 2018 using DeckGL

https://github.com/Pradeepgurunathan/Gun-Violence-Geovisualization

Predictive Analysis GitHub repository

Predictive analysis GitHub repository for the United states Gun violence data collected from 2013 to 2018.

https://github.com/Pradeepgurunathan/Predictive-analysis-on-United-States-of-America-Gun-violence-Dataset