MSc Data Science & Analytics Thesis

Predictive Analysis on American Gun Violence Dataset involving Gun Laws, Suicide and Education Rate

Author

Karthick Pandi

Student Number-19250898

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 09,2020

Declaration

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

Signed: Karthick Pandi

Date- 09-08-2020

Abstract

MSc in Data Science and Analytics

Predictive Analysis on American Gun Violence Dataset involving Gun Laws, Suicide and Education Rate

By – Karthick Pandi

"Gun Violence is not something that appears just in a bad neighborhood or in another part of the world. It appears right here, right outside the door" – Stephen Young.

Gun violence is one of the serious threats which is faced by citizens in many of the countries today. Even though many countries are affected by gun violence, the United States of America is probably the most affected country in the world. According the statistics report released by the BBC, the rate of murder or manslaughter by firearm is the maximum in the developed world. A count of 11,000 deaths were recorded which involves firearms, according to the survey in 2017. This data analysis and prediction project mainly concentrates on gun violence which occurred in the United States of America from the year 2013 – 2018. Within the report we analyze demographic variables such as age, gender, date of the incident happened, gun laws in the US, literacy rates, and suicide rates from each of the states in the US. With the help of the predictors like date of incident we are trying to predict future gun violence events. From this analysis we obtained some interesting findings related to the most dangerous times of year in terms of gun violence incidents and the predicted most dangerous state and city in the future. These predictions are performed using machine learning concepts. We also attempt to predict the most vulnerable and safest (day of the week, month of the year and states) in the US with a help of predictive scoring scheme by giving a sensible weighting to the parameters. We show that this result matches with the machine learning predictions. This project also considers whether factors like gun laws, literacy rate, population and suicide rates have any association with the gun violence incidents happening in the US. Once, after analyzing all these we attempt to grade each of the US states with the help of scoring scheme and generate some interesting results.

Contents

1. Introduction

1.1 United States Gun Violence

A famous quote **"We lose eight children and teenagers to gun violence every day. If a mysterious virus suddenly started killing eight of our children every day, America would mobilize teams of doctors and public health officials. We would move heaven and earth until we found a way to protect our children. But not with gun violence "**– Elizabeth Warren [1]. Every year United States is encountering tens of thousands of deaths and injuries. About a count of 73,505 nonfatal firearm injuries i,e) approximately 23.2 injuries per 100,000 people was recorded in the year 2013.In the same year 33,636 deaths has been recorded due to "injury by firearms" which is 10.6 deaths per 100,000 people approximately. The above death is inclusive of 21,175 suicides, 11,208 homicides, 505 in terms of accidental or negligent discharge of a firearm, and 281 deaths due to firearms use with "undetermined intent" [2]. In 2017, gun deaths reached their highest level since 1968.The main cause for these incidents is because of the spreading gun owing culture. According to the report by BBC, United states is the top most civilian country where gun culture is found common among peoples .Despite the fact that it is not as easy to know accurately how many guns civilians own around the world, but approximately the US is far away from other countries with a count of 390 million more.

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 rotation in the United States that is approximately 120.5 guns for every 100 people. Handguns, Rifles , Shotguns are the commonly found types among the people in United States and these three types occupies the top place for the weapons used by the assailant against the victims and this is acknowledged by the report submitted by BBC [5].

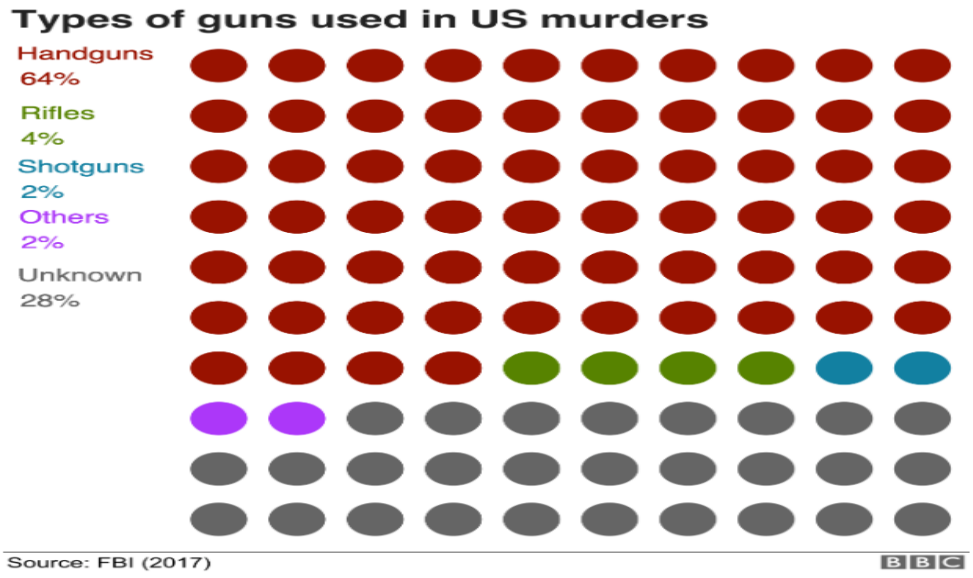


Fig 1: Types of guns used in US murders

According to the social activist, there are some reasons behind this spreading gun culture. But the most and important factors to consider is guns laws among the states. It might be a surprise for a people who came from the countries which does not allow peoples to own their guns legally.

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 its is highest among the developed countries. Un-employment, poverty, mental illness, suicide rate, Gun availability are the other driving forces for increase in Gun culture in United states.

1.2 United States Population.

The US is a country of 50 states encompassing a vast swath of North America with Alaska in the northwest and Hawaii extending the nation's presence into the Pacific Ocean. According to the United States Census Bureau the population of US is 329,989,689 till Jun 21, 2020. It is the third most country with huge population. Immigrants plays an important role in US population since United states is known for its better lifestyle and employment opportunities. As stated by United States Census Bureau one international person is getting migrated into US for every 47 seconds. 400 million is the estimated population and this will be achieved in the next 40 years i,e) within 2060 according to the national report of US. California, Illinois, Texas and Florida are the few vulnerable cities in US so many people are moving to some other cities for better and peaceful life.

Multiple factors such as large population, high number of immigrants each year, also with urbanization, high competition for jobs, Un-employment, poverty, literature rate and on top of it easy access to firearms are considered to be the important factors for these gun violence [4].

1.3 United States Gun Laws

In United States of America, it is common among individuals to have a weapon like gun. Every adult who is residing in US can own or carry one. The above mentioned one is considered as a basic right of the people in United States. That's because when the country was founded the right to buy and carry a gun was written into the constitution, which is a list of basic rules that a country is based on. Only exemptions to not to own the guns is the person with mental health illness and the persons who convicted as criminals, or if they are not a US citizen. As I mentioned earlier United States is composed of 50 different areas called states. Each state can adapt some laws to suit their own residents' views. Since from 1791, these rules are in place, so purchasing and possessing a gun is something that American people have been allowed to do for a very long time. However, regulations on gun rights do vary between various US states and there are unique principles to prevent people from having guns in certain spots like in - or near - schools. In some states, the regulations are less strict than in others. For example, in the state of Nevada, people do not have to tell anyone that they own a gun. Different rules are available in practice for which different types of guns people can own and - again - these vary from state to state [3].

Over the last 60 years, people opinion about changing their guns laws which they follow currently is seems to be changed dramatically. According to polling by Gallup, people has shifted their gears over time against this gun culture. Many of the people is not satisfied with the current gun laws and they need strict legislation.

From the US people survey, it is evident that the people voice against gun and to regulate the United states gun laws is seems to be in increasing trends. Some of the states in the countries already taken some moves to prohibit or strictly regulate the possession of assault weapons. Laws vary by state but California, for example, has banned around 75 types and models of assault weapon [5] .

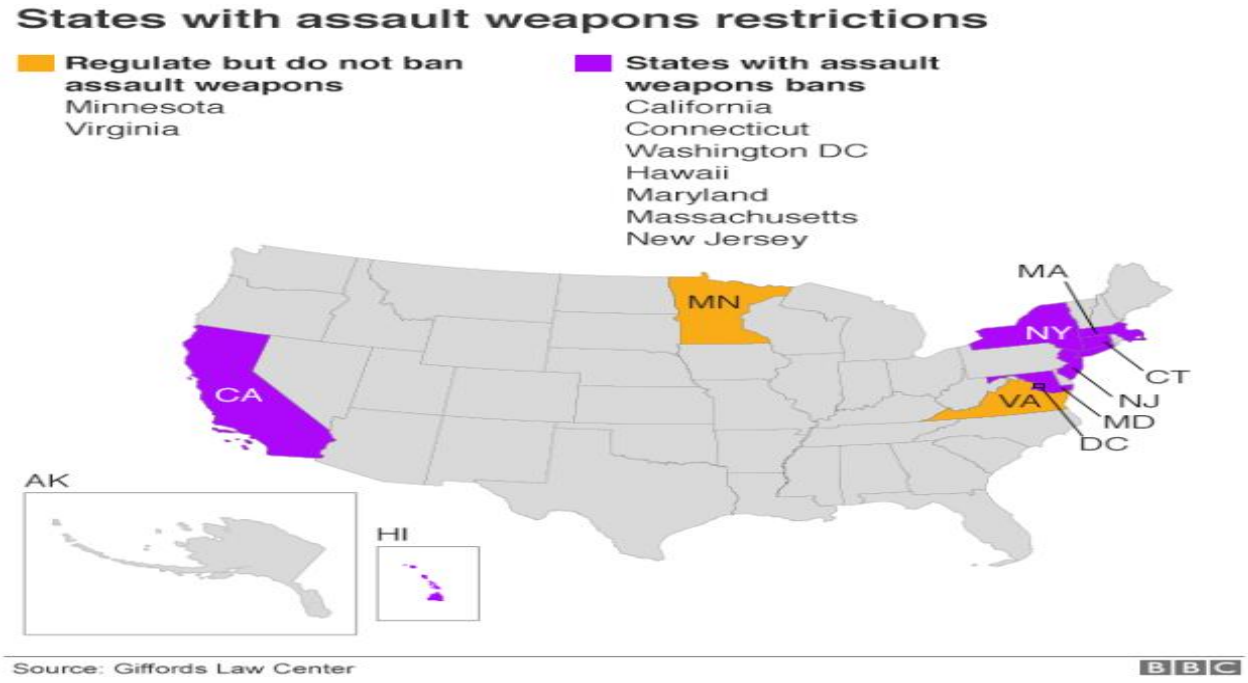


Fig 2: States with assault weapons restrictions

Across the world some of these controls are widely supported by the people including the controls like restricting the sales of guns to people who are mentally ill and who convicted as criminals or who is on watch lists.

1.4 United States Literature Rate

A person is functionally illiterate if he or she is incapable to understand basic sources of written info. These basic sources of information include warning labels and driving directions. Central Intelligence Agency (CIA) of US publishes “The Word Factbook”. "There are no universal definitions and standards of literacy" and the derived statistics are based on some common definition. "the ability to read and write at a specified age." "Detailing the standards that individual countries use to assess the ability to read and write is beyond the scope of the Factbook. Information on literacy, while not a perfect measure of educational results, is probably the most easily available and valid for international comparisons.". According to the world factbook 99% of the people in United states has basic knowledge in reading and writing and is ranked 28th of the 214 nations included worldwide. By, definition, these literacy percentage refers to people who aged 15 or older with basic reading and writing skills.

Some 52% of all Americans i, e) 49% globally have basic or below reading skills. Most of them can be able to do things like sign the forms, comparing the ticket prices for two different events and look up shows in a TV guide. These people cannot be able to find places by using maps and compare viewpoints in two editorials. Around 4% of Americans are non-erudite which implies they have below level1 literacy. They cannot read well enough to perform activities of daily in a modern society. Most cannot identify which candidate earned the fewest votes from a simple table identifying three

candidates and the number of votes they received.14% of people in America has below basic literacy levels. On the other hand, 34% of people has basic literacy level. The people with basic literacy level can be able to count the number of countries in which the generic drug market accounts for 10% or more of drug sales from two paragraphs and a chart of generic drug use in 15 countries. But most cannot identify the link leading to the organization's phone number from a website with several links, including "contact us" and "FAQ." 36% of the American people where the global literature rate on this category is 39% have immediate literacy skills followed by 12% of peoples has proficient literacy skills[12].

1.5 United States Suicide Rates

Suicide is the major health problem in United States of America. It ranked high among all other wealthy nations in the world.48,344 suicide has recorded in the year 2018, which is around 6k more from the year 2014 according to the CDC's National Center for Health Statistics (NCHS). There is an increase of 24% in annual suicide rate between 1999 to 2014 which is approximately 10.5 to 13 suicides among 100,000 people, the highest rate recorded in 28 years. Due to some circumstances these number of cases is getting under reported. In 2016, the CDC released data showing that the cases has increased in number when compared to last 30 years and in later 2018, it released further data that these continues to be increasing in each and every states except Nevada since 1999. "Death of Despair" refers to surging death rates from suicides, drug overdoses, and alcoholism are the three important factors for a consecutive three-year decline of life expectancy in United States. This constitutes shows a drop in life expectancy for the first time after three years in the U.S. since the years 1915–1918.

Suicide was the 7th leading cause for males and 14th leading cause for female with respect to the survey conducted in the year 2015. Additionally, it was the 2nd leading cause of death for people who aged between 15 to 34 in United States Of America and for the people who aged between 10 to 14 it was the third leading cause of death. From 1999 to 2010, the suicide rate among Americans is nearly 30 percent for the people who aged between 35 to 64. The largest spike is found between the people who ages between 60 to 64 is nearly 60%, then men in their fifties, with rates rising nearly 50 percent. In 2008, it was observed that U.S. suicide rates, particularly among middle-aged white women, had increased, although the causes were unclear.[20] As of 2018, about 1.7 percent of all deaths were suicides [10][11].

1.6 Motivation

Devoid Of spotting the numbers, it is easy to dismiss the brutality of our current gun violence epidemic. Our goal is to notify, educate, creating awareness and to provoke changes throughout the country. If these numbers anger and upset you as much as they do us, join our fight and act against gun violence. Everyday around 313 people are shot in US with a breakdown of 103 people are shot and killed, 210 survive with gun injuries, 95 are intentionally shot by someone else, 63 died from gun suicide,10 survived an attempted gun suicide, 1 is killed unintentionally, 90 are shot unintentionally, 1 is killed by legal intervention, 4 are shot by legal intervention, 1 died but the intent was unknown, 12 are shot but the intent was unknown. Total victim count is more than the total lives lost due to wars, AIDS, drug abuse and terrorism put together. So, there has been a drastic increase in people voice against guns control and it is very important to understand the Gun Laws, Literature Rate and Suicide Rate and how it has a direct impact on Gun Violence and facts related to the high Gun violence incident in certain U.S. States.

1.7 Objective of this Report

This Data Analysis and Visualization report mainly focus on Gun shooting incidents happened all through the United States and the impacts of Population of the United States, Gun Laws, Literature Rate and Suicide Rate in each U.S. States. The geo graphical visualization of the incident occurred place for the year 2013 to 2018 is done with the help of Deck.gl library and Google Map API.

Apart from visualization the focus of this analysis includes the dataset which holds information about US gun shooting incidents for the year 2013 to 2018 (*Source – Kaggle*). Data cleaning and pre-processing is the first and foremost things to be done before stating a data analysis. Once after completing the prerequisites like data cleaning, pre-processing and standardize the data, some preliminary analysis is done on the dataset to answer some basic questions. In this data analysis we are trying to predict the future and it done with the help of machine learning approaches like **K – Means Clustering.** Even risk scoring analysis is computed for finding dangerous month, and day by providing proper weightage to the parameters and these results are getting compared with the results obtained from the prediction using Machine Learning.

Upon further research about these incidents there might be some hidden reason which directs these gun shooting incidents. To answer these questions, I am considering some factors like gun laws, Population of the United States, Literature, and suicide rate of the state and checking the correlation of these predictor variables with the gun shooting incidents. Finally, a risk score ranking system generated including these extra parameters to get more accurate prediction of which U.S. State is Dangerous and how the above factors plays a curial role in Gun culture in the United States of America.

1.8 Structure of this Report

In this report, at the first part we have seen some history, statistical facts and reasons that drives this spreading gun culture in the United States of America. The, next section of this report holds the basic details about the tools and technologies we opt for doing the research on Gun violence dataset. Followed by the predictive analysis, we had done for this American gun violence dataset using some machine learning concepts. Dataset overview and Geo visualization are the next two sections which implies the overview about the dataset like the number of rows and columns it holds, the names of the columns and in Geo visualization it holds the details and steps that we used to geo- locate the gun shoot incident happened places in the United States using Deck. gl library.

The first step to do before starting our research is preparing the data and how this has been done is enclosed in the section five of this thesis. The sixth chapter of this thesis is the one which holds the research findings about the US gun violence dataset and its interesting observations. Finally, the conclusion of the thesis and the materials which we used as a reference is mentioned in the last two chapters of this thesis.

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?

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 conscious 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 have not able to reach our specific requirements. The main problem with these libraries is that these libraries focus entirely around Python, neglecting 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
```

Numerical Python simply called as NumPy. A core library for scientific calculations which consist of powerful 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 to reduce the memory usage, increase the speed and also 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 is why this Pandas is considered as a trusted ally in both Machine Learning as well as Data Analytics. Pandas incorporated with some quick, adaptable and flexible, with expressive data structures intended to make works with "relational" or "labeled" data both easy and intuitive. In python, it is considered as a high-level building block for doing practical, real world data analysis. 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 wide-ranging library for establishing static, animated, interactive visualization which plays an important 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 designed to work with the broader SciPy stack. In other words, we can say it as numerical mathematics extension of NumPy. A module named Pyplot which provides an interface like MATLAB which is also a module from Matplotlib. In general, Matplotlib is designed to be as usable as MATLAB, with the ability to use python, and it has the advantage of being free and even as a open source. The main advantage of using Matplotlib is it allows us to visual huge amount of data in a easily digestible manner. It encloses variety of plots such as lines chart, bar chart, scatter plot, histogram etc. Publication excellence figures in a variety 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, and various graphical user interface toolkits.

Seaborn:

```
import seaborn as sns
```

Factual designs are getting done in Python with the help of library called Seaborn. It is built on top of matplotlib which resembles the panda's information structures. Here is a portion of the usefulness that seaborn offers:

1. Connections between various features are looked with the help of dataset organized API.
2. Matplotlib API is comparatively low level. Doing advanced statistical visualization is feasible, but often involves a lot of boilerplate code. In order effortless this we opt for Seaborn.
3. Seaborn offers an API on top of Matplotlib that poses massive brands for plotting and coloring the graphs.
4. Programmed estimation and plotting of direct relapse models for various types subordinate factors
5. Helpful perspectives onto the general structure of complex datasets level deliberations for organizing multi-plot networks that let you effectively fabricate complex representations
6. Seaborn also provides high level commands to create a range of plot categories useful for statistical evaluation, and even for some statistical model fitting.

Seaborn intends to make representation a focal piece of investigating and getting information. Its dataset-situated plotting capacities work on data frames and exhibits containing entire datasets and inside play out the important semantic planning and measurable conglomeration to create enlightening plots.

Scikit-Learn:

Scikit-learn is one of the machine learning libraries available in Python for free of cost. It has various

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

algorithms like kmeans, support vector machine, Linkage clustering, random forests, and k-neighbours, and NumPy and SciPy are some of the numerical and scientific libraries which is get supported by this. 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.

Scikit-learn has some popular groups of models which include:

- Feature selection: for finding significant attributes from which to create supervised models.
- Ensemble methods: for combining the predictions of multiple supervised models.
- Manifold Learning: It is used for briefing and depicting complicated multi-dimensional data.
- Supervised Models: a vast array not limited to generalized linear models, discriminate analysis, naive bayes, lazy methods, neural networks, support vector machines and decision trees.
- Cross Validation: valuing the performance of supervised models on unseen data is done.
- Datasets: It is for testing the dataset and to generate the dataset with some specific properties for investigating model behavior.

- Dimensionality Reduction: for lowering the count 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: for grouping unlabeled data such as Hierarchical and K-Means.
- Feature extraction: attributes in image and text data are gets defined.

2.3 Anaconda

Anaconda is a free of charge and open source python distribution for scientific computation that helps for package management and deployment. This comprises 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 also 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:

Jupyter notebook, JupyterLab, Spyder, Glue, Orange, VS code, 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 Jupyter Notebook

The Jupyter Notebook is an open-source web application that permits us to make and share 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 much more.

Jupyter adopts over 40 programming languages such as Python, R, Scala, and Julia. It can be share with others via drop box, 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.

2.5 Visual Studio Code

Finally, the IDE which we preferred to develop this 3D Geo graphical visualizations using Deck.gl is Visual Studio 2013.Visual Studio code is open source code editor by Microsoft for Windows, MacOS and Linux. 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, JavaScript function timing and embedded with GitHub. So, these above features makes this IDE an odd one while compared with other IDE available in the markets. Users can change keyboard shortcuts, themes, preferences and can even install extensions that add extra functionality. Software developers preferred to use this IDE for developing their codes using the common programming languages such as JavaScript, Typescript, JSON, CSS and HTML.

3.Data Science Methodologies

3.1 Clustering

Clustering is the method where we can be able to identify the similar group of data in a dataset. Clustering is one among the most popular technique avails in data science. Each group which has entities are more like other objects in the same group and have more dissimilarities with the other objects in some other groups.

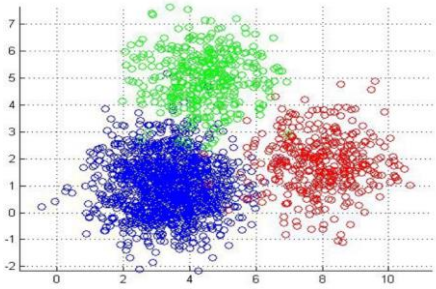


Fig 3: Clustering

Types of Clustering

In general clusters is broadly classifies into two different categories. They are:

1. Hard Clustering
2. Soft Clustering

Hard Clustering

Hard clustering is nothing but, each data point is either belongs to a cluster completely or not.

Soft Clustering

On the other hand, soft clustering is nothing but as an alternative of putting each data point into an individual cluster, a probability of those data points in those clusters is assigned.

3.2 Types of Clustering Models

There are more than 100 types of clustering algorithm available. Every methodology follows some different set of rules for defining the “similarities” among the data points. Out of these very few algorithms are used or preferred by the data analyst, let’s look them in detail:

- Connectivity Models
- Centroid Models
- Distribution Models
- Density Models

Connectivity Models

Connectivity models follows two approaches. The first one is that it starts classifying all data points in a separate cluster and then aggregating them with the decreasing distance. On the other hand, second approach follows by assigning all data points into a single cluster and then partitioned with the increasing distance. It is to remember that the choice of distance is subjective. The advantage of this

method is these models is very easy to interpret but it lacks scalability of handling large dataset. Hierarchical clustering and its variants are the most popular clustering algorithms which falls under this category.

Centroid Models

In this model the relationship is obtained with the familiarity of the data points to the centroid of the clusters and it works on the principle of iterative clustering algorithms. K – means clustering algorithm is a widespread algorithm technique which drops into this classification. In these methods the number of clusters needed at the end must be stated beforehand, which makes it vital to obtain the preceding information of the dataset. These models run recursively to locate the local optima.

Distribution Models

This model is based on the principle of what probability is the data points in the cluster belongs to the same distribution i,e) the distributions like Gaussian, Normal etc. The main drawback of this model is it habitually suffer from overfitting. Expectation-Maximization Algorithm is the one which falls under this model which uses multi variate normal distributions.

Density Models

This model works with the principle of searching the data space for areas of varied density of data points in the data space. Various density regions are gets isolated and data points are assigned into those regions in the same cluster. Popular examples of density models is DBSCAN and OPTICS.

3.3 Types of Clustering Algorithms

Now will dive into the two extremely widespread clustering algorithms in detail.

K – Means Clustering

K – Means works on the principle of iterative clustering algorithm principle. In each iteration its aims to find the local maxima. This algorithm follows the below five steps.

1) Desired number of clusters must be specified.
2) Initializing the centroid is done with the help of shuffling the dataset and then selecting the K data points randomly for the centroids without replacement.
3) Centroid computation is done iteratively until no further improvements needed i,e) assignment data points to clusters is not changing.
    - Sum of squared distance between centroid and all data points is getting computed.
    - Assigning each data point to the closest centroid
    - Centroids for the clusters is computed by taking the average of all data points that belongs to each cluster.

Hierarchical Clustering Algorithm

As the name suggests it is an algorithm that assembles hierarchy of clusters. This algorithm begins with all the data points allotted to a cluster of their own. Then the next step is like the two nearest cluster is merged as same cluster. In the end the algorithms get terminate when there is only one cluster left. The result of hierarchical clustering is shown with the help of dendrogram. Interpretation of dendrogram is done as follows:

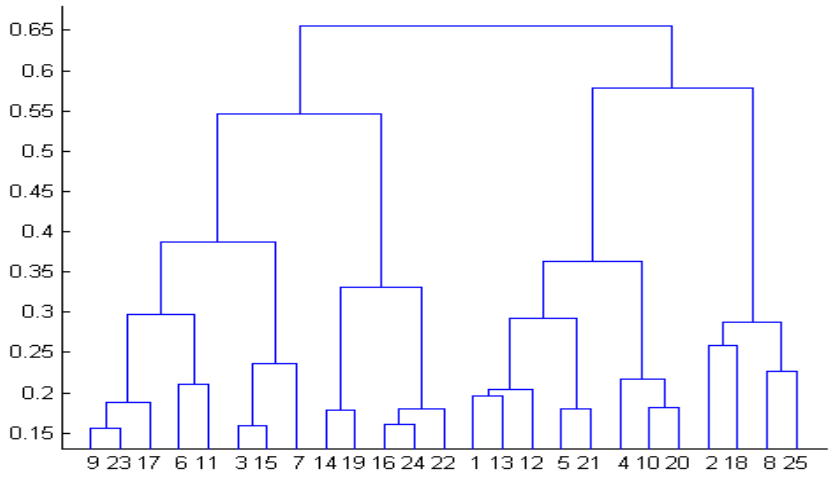


Fig 4: Hierarchical Clustering

Here we can see that at first it gets started with 25 data points, where each assigned to separate clusters. Then the two closest cluster is gets merged and acts as a single cluster, it will continue until we have single cluster at the top. The distance between two clusters is represented by the height of the dendrogram at which the two clusters gets merged. By observing the dendrogram we can decide the number of clusters that can best depict the different groups. The best choice of choosing the number of clusters is the number of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster. In the above example, the best choice for number of clusters is calculated as 4 since the distance AB is the maximum vertical distance covered by the horizontal line in this dendrogram.

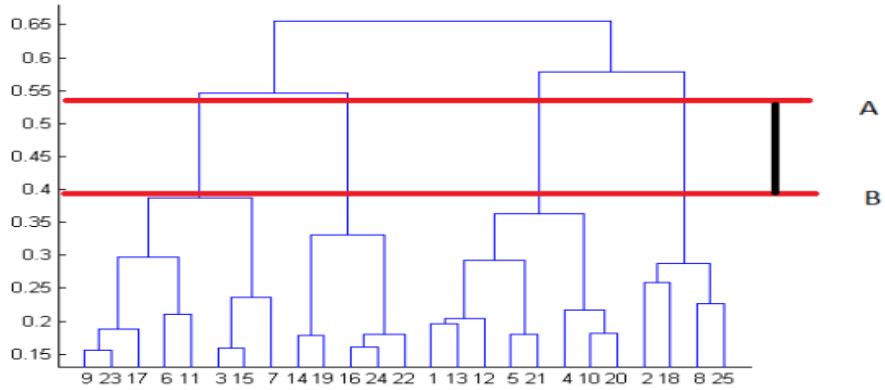


Fig 5: Dendrogram

Hierarchical clustering follows two different approaches namely:

1) Bottom Up Approach
2) Top Down Approach

The above example is work on the principle of bottom up approach. It is also possible to solve this with the help of top down approach beginning with all data points allocated into the similar cluster and performing splits recursively till each data points is allocated a separate cluster.

The merging of two clusters is decided based on the closeness of the clusters. There are several metrics for determining the closeness of the two clusters. They are:

- Euclidean distance: $||a-b||_2 = \sqrt{(\Sigma(a_i-b_i))}$
- Squared Euclidean distance: $||a-b||_2^2 = \Sigma((a_i-b_i)^2)$
- Manhattan distance: $||a-b||_1 = \Sigma|a_i-b_i|$
- Maximum distance:$||a-b||_\infty = max_i\,|a_i-b_i|$
- Mahalanobis distance: $\sqrt{((a-b)^T\, S^{-1}\, (-b))}$ {where, s : covariance matrix}

3.4 Hierarchical Vs K - Means Clustering

- Handling big dataset is not possible with hierarchical but K means can manage large dataset
- $O(n^2)$ is the time complexity for k-means which is linear in nature whereas, hierarchical has its time complexity as quadratic in nature.
- In the hyper spherical space like circle in 2D, sphere in 3D k-means works well.
- Parameterizing the value K in k-means, which is the number of clusters we want to create. Creating k centroids is the initial step while using this type of algorithm. It then iterates between assign step and then update step, whereas the hierarchical clustering produces the clusters incrementally, by using dendrogram.
- With the help of dendrogram which is produced in hierarchical clustering understanding the data is very easy when compared with K-means clustering.
- The algorithm can never undo any previous steps. So, for example, the algorithm clusters 2 points, and later we see that the connection was not a good one, the program cannot undo that step.

3.5 Confidence Intervals

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

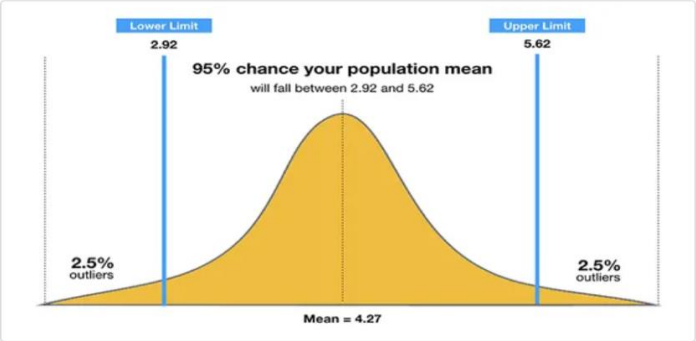


Fig 6: Confidence Interval

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

4.Data Collection and Overview

Understanding the data in the dataset is the most important process before starting to analyze the dataset. Each of the attributes in the dataset must be understood clearly to come up with some research questions for our analysis. In this data analysis we have used two different dataset one with United states gun violence incidents for the year 2013 – 2018 and the other one with United states gun laws, suicide, and literature rate.

4.1 Data Collection:

What is Data Collection?

It is nothing but the process of gathering information on targeted variables from an establishes system with the help of this one can able to respond related questions and evaluate the outcomes. It is an important element in research in all fields of study including humanities and business.

Why Data Collection?

- Data Collection routes to answer our basic questions such as which state is more dangerous in United States.
- The attributes like longitude and latitude is very useful in geo locating our incident locations.
- For predicting which month of the year and day of the week is more dangerous.
- To find which age category and gender category involved more in this type of crimes
- Is gun laws plays an important role in gun violence?
- Does suicide and literature rate are highly correlated with Gun Violence.

4.2 United States Gun Violence Dataset

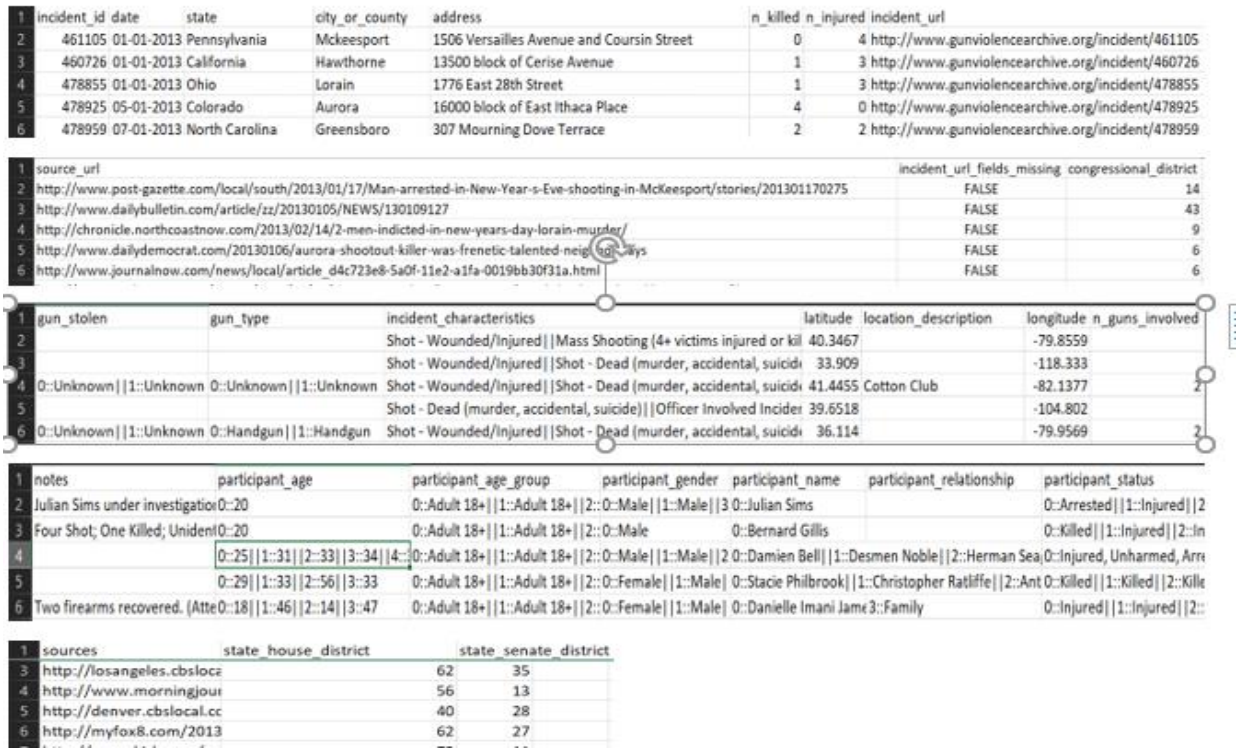
In this analysis we are incorporating a dataset which has over 260k gun violence incidents which happened from 2013 to 2018, with comprehensive information about each incident. With the help of this dataset we are aiming to make informed predictions about forthcoming trends which plays a crucial role 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.

Note:
Source of the Dataset is Kaggle. https://www.kaggle.com/jameslko/gun-violence-data

Dataset Overview:



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

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

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

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

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

Fig 7: Overview of US gun violence dataset

4.3 United States Population, Gun Laws, Suicide Rate and Literature Rate Dataset

Population of the United states, Literature and suicide rate of the states in the US and finally the gun laws with respect to specific states are considered for the further analysis such as whether these factors has much influence on gun shooting incidents which is happening all through the country. The above-mentioned factors are considered for risk score calculation by providing appropriate weightage to the parameters. Once after computing these risk score it gets compared with the previous result i,e) with the Dataset 1 (US gun shooting dataset from kaggle). By observing these results, we can conclude that which U.S. State is Dangerous and how the factors like Population, Suicide Rate, Literature Rate, and Guns laws 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: Names of the states in US
2. Pop: 2018 Population of 51 US States
3. Permit: Whether the people in the states are allowed to use the guns legally
4. High School Graduate or Higher: Literature rate who passed high schools
5. Bachelor's Degree: Literature rate who passed bachelors
6. Advanced Degree: Literature rate who passed masters
7. Average: Average literature rate
8. Suicide Rate: Average suicide rate for the years 2013 to 2018

Note:

- Source of the Dataset for 2018 Population Attribute:https://www.worldatlas.com/articles/us-states-by-population.html
- Source of the Dataset for Suicide rate: https://www.cdc.gov/nchs/pressroom/sosmap/suicide-mortality/suicide.htm
- Source of the Dataset for Gun Permit: https://en.wikipedia.org/wiki/Gun_laws_in_the_United_States_by_state
- Source of the Dataset for Literature: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment

Dataset Overview:

	A	B	C	D	E	F	G	H
1	State	Permit	High school graduate or higher	Bachelors degree	Advanced Degree	Average	Sucide Death Rate	Population
2	Alabama	No	85.3	24.5	9.1	39.63333333	15.64	4908621
3	Alaska	No	92.4	29	10.4	43.93333333	25.28	734002
4	Arizona	No	86.5	28.4	10.7	41.86666667	18.26	7378494
5	Arkansas	No	85.6	22	7.9	38.5	18.74	3038999
6	California	Yes	82.5	32.6	12.2	42.43333333	10.54	39937489

Fig 8: Overview of US Population, suicide, literature and gun laws dataset

5.Geo Visualization

5.1 Introduction

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 to help out with the people who wanted to do big visualization on the web. It is 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. 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:

```
41  //HeatMap layer
    Complexity is 3 Everything is cool!
42  const heatmap = () => new HeatmapLayer({
43      id: 'heat',
44      data: sourceData,
45      getPosition: d => [d.longitude, d.latitude],
46      getWeight: d => d.n_killed + (d.n_injured * 0.5),
47      radiusPixels: 60,
48  });
49
```

Fig 11: Heatmap Generation code

Below code helps us to generate 3D hexagonal layer:

```
50  // Hexagon layer
    Complexity is 3 Everything is cool!
51  const hexagon = () => new HexagonLayer({
52      id: 'hex',
53      data: sourceData,
54      getPosition: d => [d.longitude, d.latitude],
55      getElevationWeight: d => (d.n_killed * 2) + d.n_injured * 1,
56      elevationScale: 1000,
57      extruded: true,
58      radius: 10000,
59      opacity: 0.6,
60      coverage: .8,
61      lowerPercentile: 50
62  });
```

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 Geo-Visualization Result

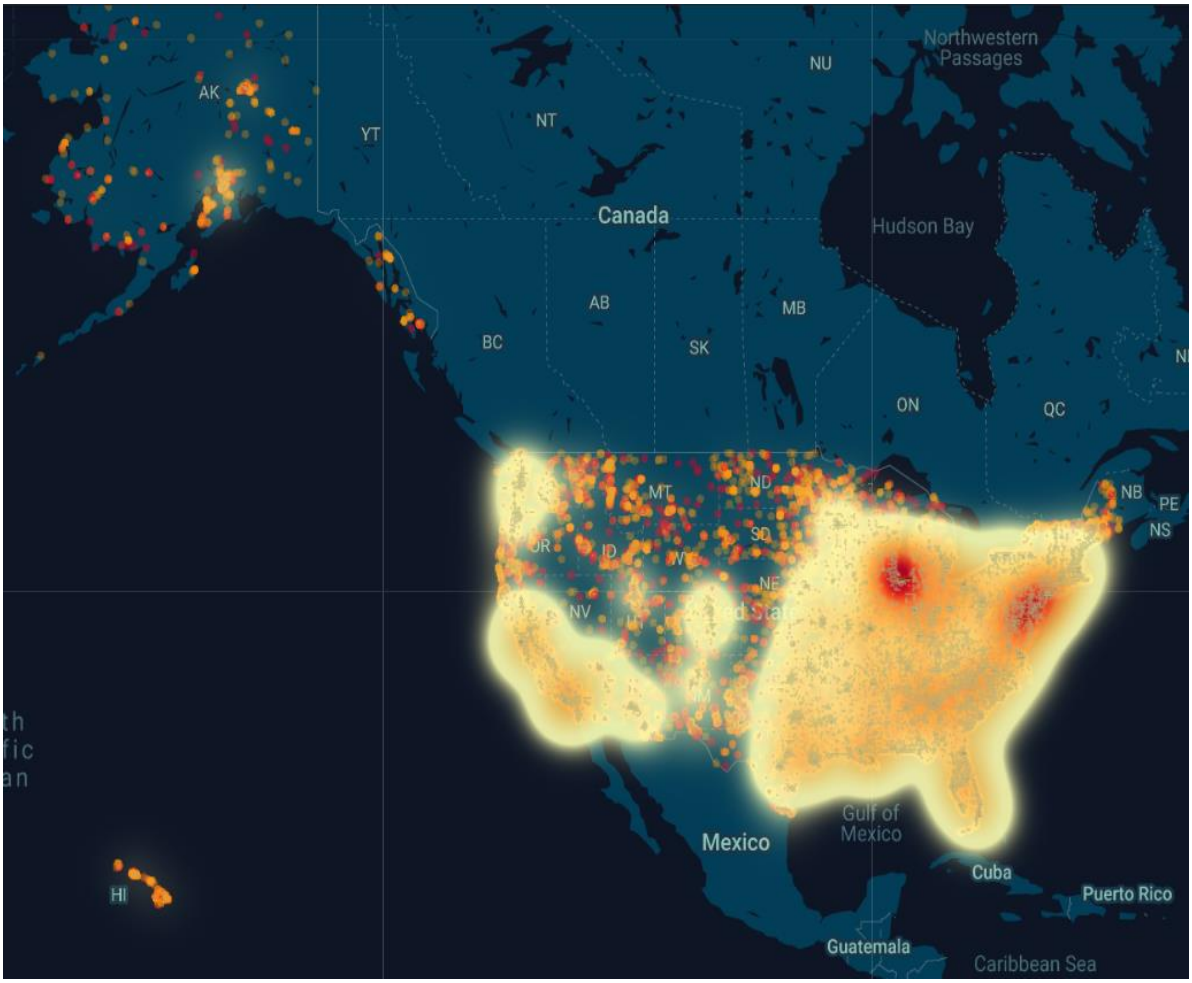


Fig 14: Heat Map of the United States Gun Violence from 2013 to 2018

The above geo-graphical map represents the United States gun violence incidents for the year 2013 to 2018.It is observed from the graph that:

- North-East and South-East part of the United states has high density gun violence incidents.
- 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 collection, Data cleaning and Data modification are the steps involved in Data Preparation of the dataset. Dataset which was collected from different source are stored as a CSV file format. These datasets have to be imported to perform data cleaning, to found any mismatch in the data, to remove null values and to perform some data modification which is useful for our data analysis.

6.1 Data Importing

Pandas a library in Python is very much useful in importing dataset which needed for our analysis.

Reading our two datasets for our data analysis.

```
#Importing Dataset and storing it for Further Analysis
dataset1=pd.read_csv('gun_violence_data_2013_2018.csv')
dataset3=pd.read_csv('US_GunPermit_Edu_Suc.csv')
df_crime=dataset1
df_Gun_Edu=dataset3
```

Our first dataset contains the records about the United States gun violence incidents for the year 2013 to 2018 and in the second dataset it is with the US gun laws, Population, Suicide and Literature Rate.

Initial Description of the Datasets:

1. United States Gun Violence Dataset for the year 2013 to 2018. These below pictures represent the details of the first dataset 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 16: United States Gun Violence Dataset Details before Data Cleaning

This dataset holds 239677 entries with 29 columns. Incident_id, data, state are some of the columns that doesn't have any null values whereas participant_gender, participant_age, participant_name are some of the columns with null values. Data cleaning has to performed to remove those null values before dive into our data analysis process. Some columns have to be dropped since it doesn't holds any useful information related to our analysis.

2. United States Gun Laws, Population, Suicide and Literature Rate

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 8 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   State                           51 non-null     object
 1   Permit                          51 non-null     object
 2   High school graduate or higher  51 non-null     float64
 3   Bachelors degree                51 non-null     float64
 4   Advanced Degree                 51 non-null     float64
 5   Average                         51 non-null     float64
 6   Sucide Death Rate               51 non-null     float64
 7   Population                      51 non-null     int64
dtypes: float64(5), int64(1), object(2)
memory usage: 3.3+ KB
```

Fig 17: United States Gun Laws, Suicide, literature and population Dataset Details before Data Cleaning

This dataset holds 51 entries with 8 columns and none of the columns has any missing values. This dataset holds the information about the external factors of these US gun violence incidents and how these factors gets correlated with the Gun Violence.

Steps involved after Data Importing:

1. Data exploration or initial data analysis is the first and fore most step that has to be done after data importing.
2. Data cleaning and Data preprocessing must be done before our data analysis.

6.2 Data Cleaning

The process of correcting or detecting an error in a record set, table, or database and refers to identifying the data with incomplete information or inaccurate information and the replacing those data or even deleting those data. Data wrangling tools or with the help of batch processing these data cleaning can be performed interactively.

- Removing the null values:

```
#Investigating Dataset 1 for Null values for Data Cleaning using heatmap
sns.heatmap(df_crime.isnull(), cbar=False,cmap="YlGnBu");
plt.title('HeatMap for Null value detection')
plt.savefig("data_cleaning_1.png")
```

There are some null values available in our dataset. Columns like location_description, participant_relationship has more null values.

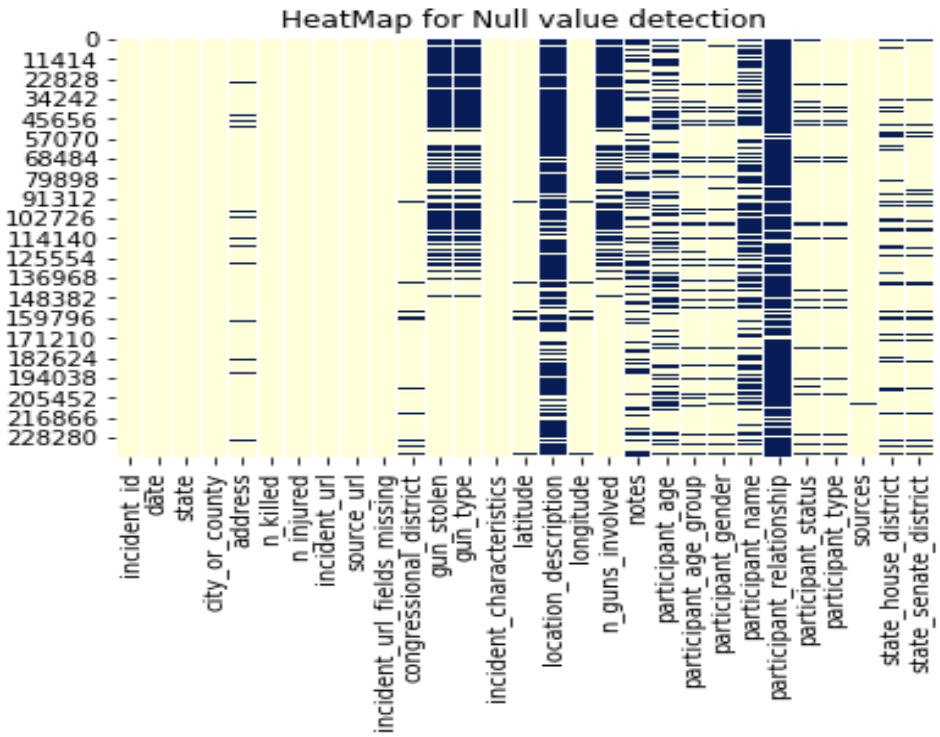


Fig 18: Heat map for detecting null values

- Dropping Unwanted columns which is not used for further analysis:

```
# Dropping unnessary field from dataset that we will not be using for analysis purpose.
df_crime.drop(labels=['incident_id','incident_url','latitude','longitude','gun_type','source_url','incident_url_fields_missing',
'location_description','congressional_district','notes','sources','state_house_district','participant_status',
'state_senate_district','participant_type','participant_relationship','participant_name','participant_gender',
'participant_age_group','participant_age','incident_characteristics','address','gun_stolen'],axis=1,inplace=True)
```

These are the columns in our gun shooting dataset which is not useful for further analysis. So, these columns have to be dropped from the dataset and it is done with the help of drop command.

6.3 Data Modification

Data Modification occurs when a stored value is changed to different forms. In addition to that extracting some part values from the original values and computing the values from other values are also comes under data modification.

- Extracting the Day, Month, Year from the Date field and stored in the separate column

```
# 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
```

- Computing the number of males and number of females involved in the incident from the column participant_gender. We are extracting the string “Male” and “Female” from the column participant_gender, counting this occurrence of string and storing this is a new column with the names male and female respectively in US 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
```

```
# 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
```

- Computing the age of the participants in the gun shooting incident from the column participant_age and stored in the separate columns called age.

```
# 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
```

- Total_Person is the new attribute which holds the total count of participants who got killed and injured in the US gun violence dataset.

```
# 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
```

- Removing or correcting some mismatches in “Male” and “Female” count in the United States gun shooting incidents datasets for the year 2013- 2018.

```
# 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']
```

Note:

- Dataset 2 United States Gun laws, Population, Suicide and Literature Rate does not need any data cleaning to be performed.

7. Data Exploration and Analysis

1) To find any relationship or correlation between the fields in our main dataset (Dataset 1)

This is to check how the attributes in our US gun violence dataset is correlated with each other.

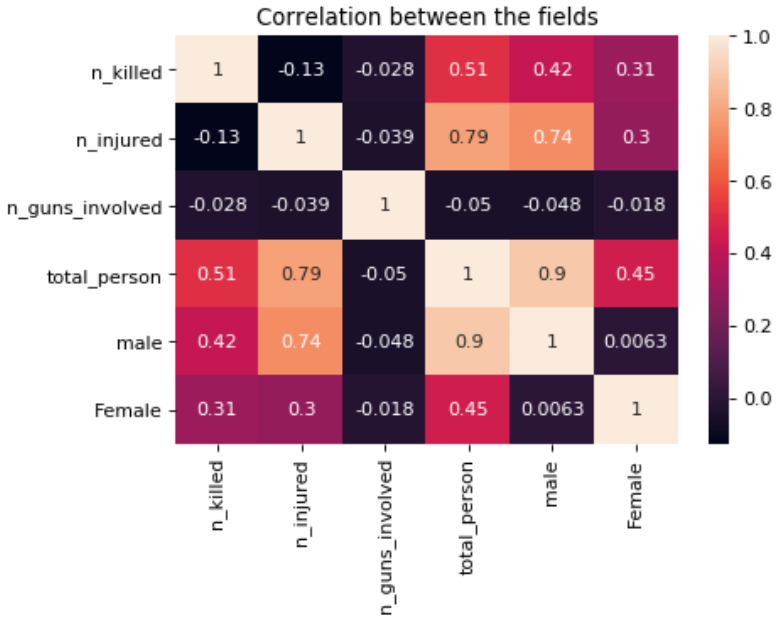


Fig 19: Correlation Matrix

Observations:

- From this correlation matrix we can conclude that Male is highly correlated with the total_person when compared to female. i,e) from the correlation values we can observe that males involvement in the incident is twice when compared to the female involvement.
- In terms of killed and injured the number of persons get affected in males is higher when compared to females.

2) What age group is most affected in USA Gun Shooting Data between 2013-2018

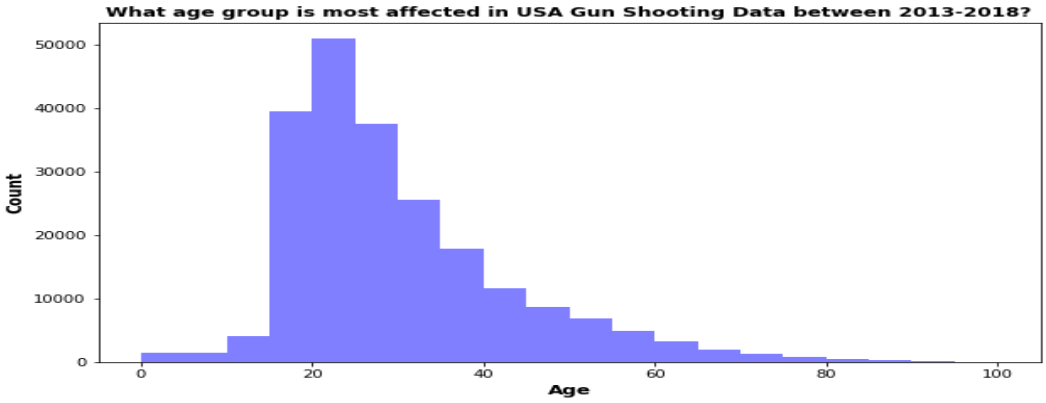


Fig 20: Histogram for age group

To find the age group which is most affected because of this United States Gun Violence for the year 2013 to 2018, we need to extract the age column from our dataset 1.

```
In [17]: round(df_Age['Age'].mean())
Out[17]: 29

In [18]: round(df_Age['Age'].median())
Out[18]: 26

In [19]: round(df_Age['Age'].mode())
Out[19]: 0    19.0
         dtype: float64
```

Observations:

- It is evident from the histogram that the people with the age between 25 – 30 is affected more with this gun shooting incidents. More young adults are getting affected because of this spreading gun culture in United States.
- Ages 29,26 and 19 are the mean, median, mode ages for these gun shooting incidents participants.
- Highest number of people are aged 19 who involved in this crime incidents.

3) Which USA State is the Dangerous based on the Gun shooting Data from 2013-2018?

Total_Peron is the columns which hold the count of **numbers of persons gets injured + number of persons gets killed**. These columns have to be extract for answering this question

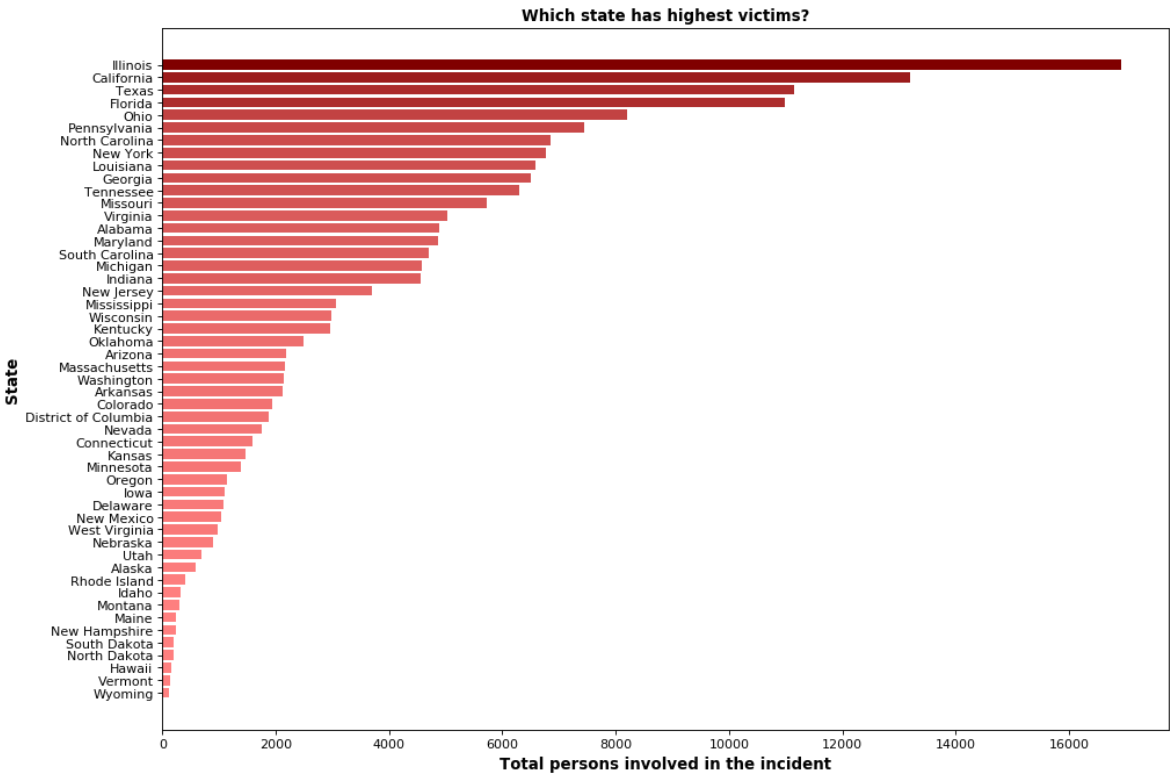


Fig 21: Dangerous state based on gun data (2013 – 2018)

Observations:

- It is conclusive from the graph that **Illinois is the most dangerous state in the United States**. On the other hand, **Wyoming is the safest state in the United States of America**.

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

Male, Female, Year, and the total count of male and female for each year has to use for this analysis.

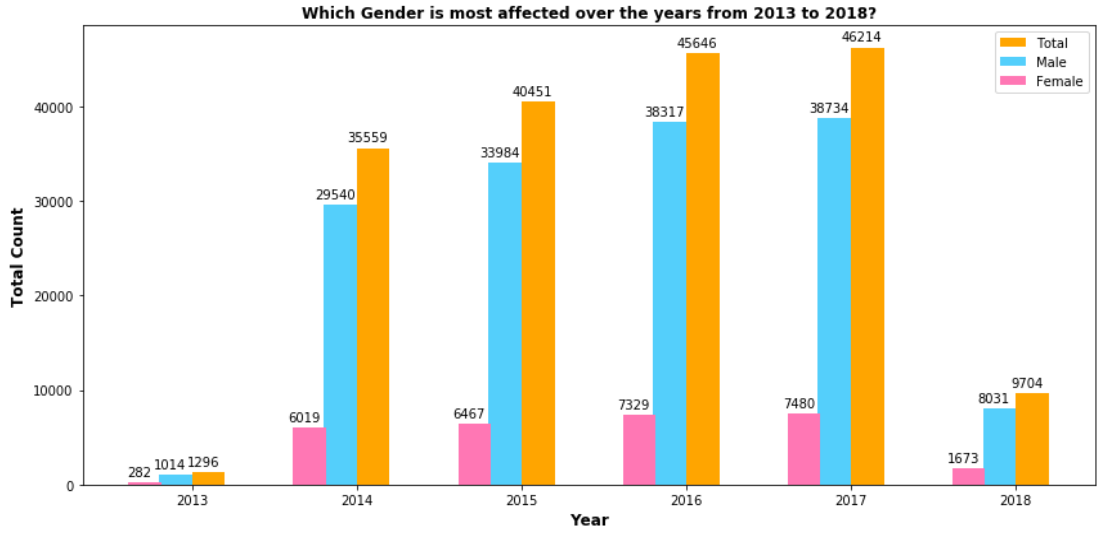


Fig 22: Gender based incident count (2013 – 2018)

Observations:

- It is observed from the graph that in each year males got affected in high number when compared to females. In addition to that we can be able to visualize that every year the crime rate is increasing steadily.
- The year 2018 has a smaller number of incidents since we don't have any enough records to process for that particular year. So, the year 2017 has the highest number of crimes and the year 2013 has lowest number of crimes.

Prediction based on K - Means clustering

5) Which USA states are highly dangerous from Gun violence using K means predictive analysis?

The Victim count and the Incident count gets grouped based on State and Year and the outcoming results are used in this analysis.

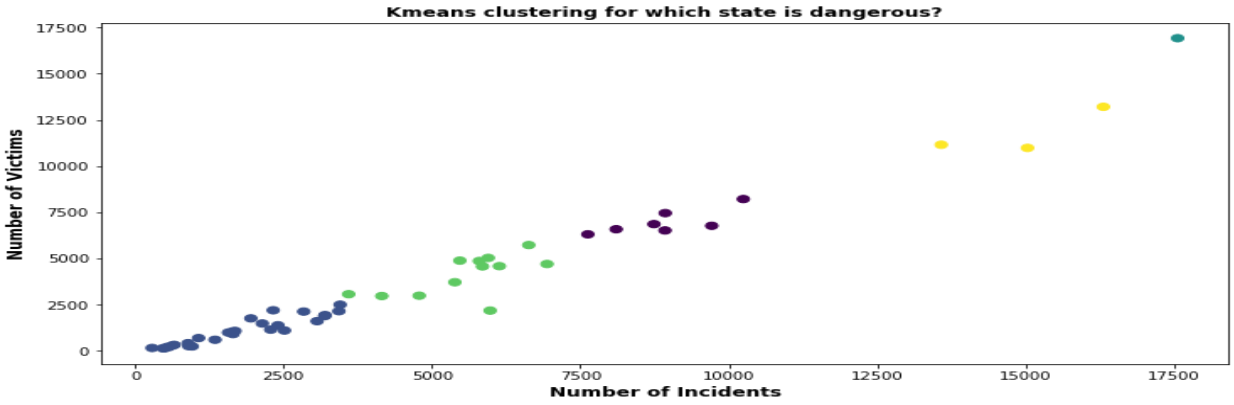


Fig 23: K means predictive clustering for states in US

Observations:

- **Illinois** is the state which is more dangerous in US according to the predictive results and it appears at the top right corner in the figure 23.

```
In [31]: # States count with corresponding cluster number
         df_state.groupby('cluster',as_index=False).count()['state']

Out[31]: 0     7
         1    28
         2     1
         3    12
         4     3
         Name: state, dtype: int64
```

```
In [34]: # cluster 2
         df_state.query('cluster==2')['state']

Out[34]: 13    Illinois
         Name: state, dtype: object
```

```
In [36]: # cluster 4
         df_state.query('cluster==4')['state']

Out[36]: 4     California
         9        Florida
         43         Texas
         Name: state, dtype: object
```

Observations:

- It is evident that our predictive results get match with the analysis results. In both we found that Illinois is the most dangerous state in the United States.

- California, Texas, Florida occupies the next three places in the most dangerous states in US.

6) Which USA Cities are Dangerous from Gun violence?

The Victim count and the Incident count must be grouped based on City and Year and the outcoming results are used in this analysis.

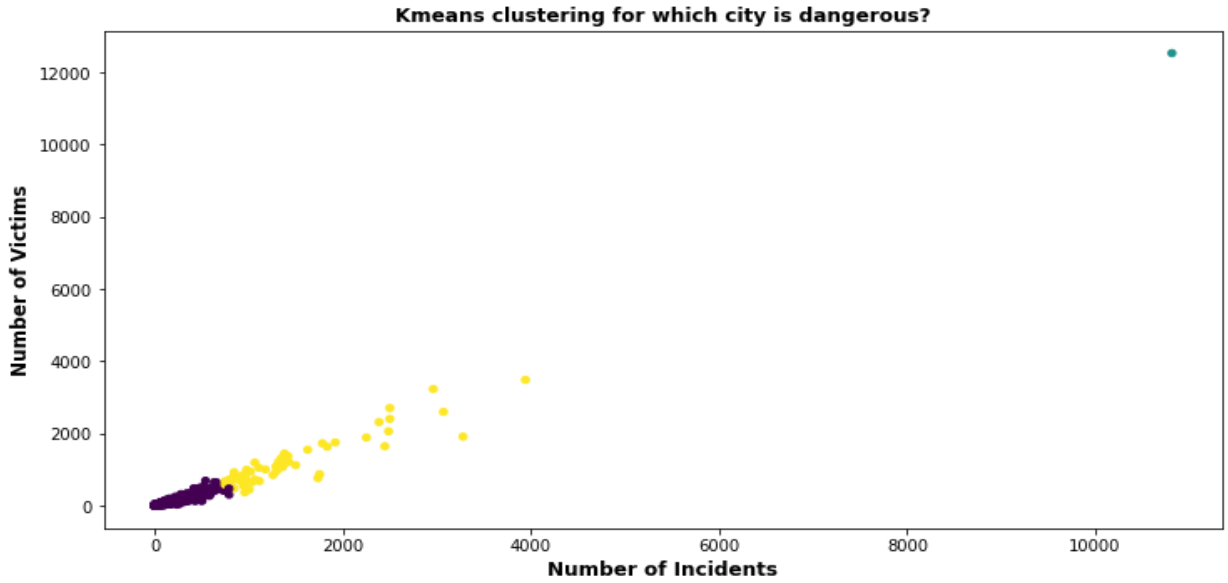


Fig 24: K means predictive clustering for cities in US

Observations:

- **Chicago** is the city which is more dangerous in US according to the predictive results and it appears at the top right corner in the figure 24.

```
In [40]: df_city.groupby('cluster').count()['city_or_county']
Out[40]: cluster
         0    12842
         1        1
         2       55
         Name: city_or_county, dtype: int64

In [41]: # cluster 1
         #Most dangerous city or county in USA
         df_city.query('cluster==1')
```

Out[41]:

	city_or_county	total_2013	total_2014	total_2015	total_2016	total_2017	total_2018	incident_2013	incident_2014	incident_2015	incident_2016
2019	Chicago	77.0	2307.0	2771.0	3553.0	3304.0	519.0	15.0	2042.0	2369.0	3075.0

Observations:

- From the above prediction, **Chicago** is the most dangerous city in the US.

Prediction based on linkage clustering

7) Which Day of week has high chances of Gun violence?

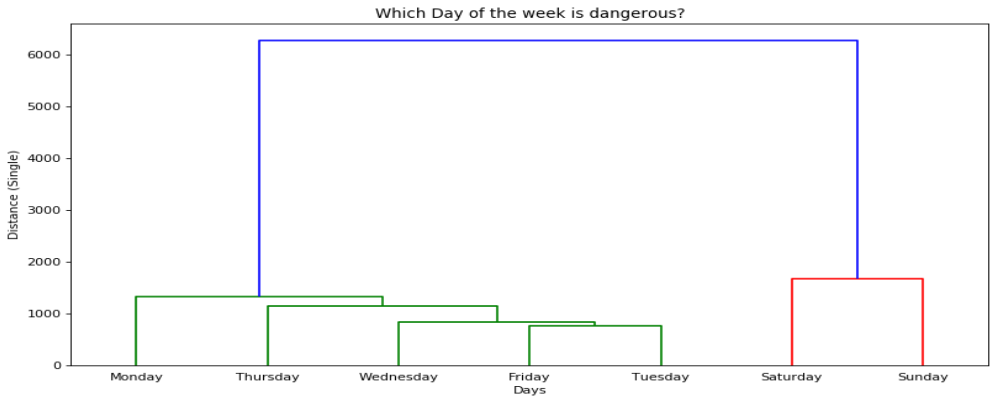


Fig 25: Linkage clustering to find dangerous day of the week in US

Observations:

- **Saturday and Sunday** are the most dangerous days of the week. On these two days more number of crimes occurred in US.
- **Friday and Tuesday** are the day with less crimes. So, these are the safest days of the week.

8) Which Month of the year has high chances of Gun violence?

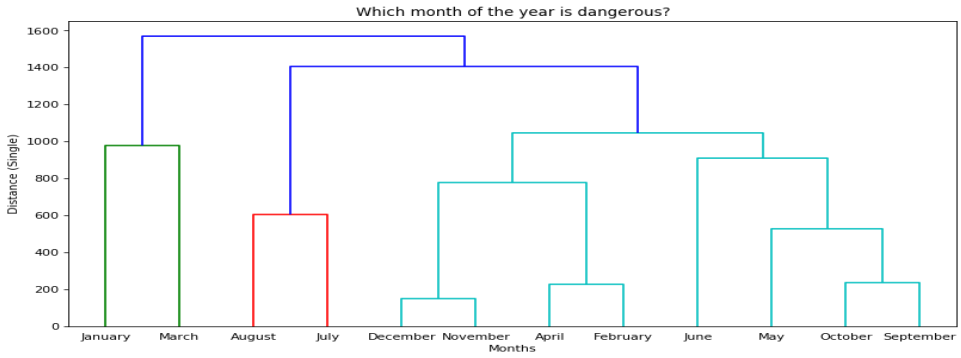


Fig 26: Linkage clustering to find dangerous month of the year in US

Observations:

- **January and March** are the most dangerous months of the year. On these two months a greater number of crimes happened in US.
- **December and November** are the months with less crimes. So, these are the safest months of the year.

9) Which Top 3 US state has Highest and Lowest Kills, Injured, Guns Involved, Incidents and Total Victims?

- **Illinois, California, Texas** are the top three states with highest victim rates.
- **Illinois, California, Florida** are the top three states with highest incidents rates.
- **Hawaii, Vermont, Wyoming** are the top three states with least incidents rates.
- **Wyoming, Vermont, Hawaii** are the top three states with least victim rates.
- **California, Illinois, Florida** are the top three states with high guns involved in US.

- **Hawaii, Vermont, Wyoming** are the top three states with least guns involved in US.
- **California, Texas, Florida** are the top three states with highest kills.
- **Illinois, California, Florida** are the top three states with highest injury rate.
- **Vermont, Rhode Island, Hawaii** are the top three states with least kills.
- **Wyoming, Vermont, Hawaii** are the top three states with least injury rates.

10) Risk Score Calculation involving number of fatalities, incidents and Guns involved:

The risk score calculation is used for predicting (the most dangerous state and city, the safest state and city, the dangerous and safest month of the year , and the dangerous and safest day of the week). This risk score calculation involves number of fatalities, incidents and guns involved.

- The weightage includes **0.15 for harmless incidents and number of persons gets injured, 0.30 for fatal incident and number of persons killed and 0.10 for the gun** involvement.
- 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.

Our risk score calculation is exactly gets matches with the predictive data. In both results we are resulting that **Illinois, California, Florida** are the top three states with lot of crime activities. So, these are the top three dangerous state in US. But in top three safest state **Vermont is replaced with South Dakota**.

Note: Based on the studies from initial analysis from Chapter 6, weightage is given based on the importance of the attribute and impact on our analysis.

10.1) Which of the two days is more dangerous and safest day of the week?

We Predicted that **(Saturday and Sunday)** are the **most dangerous days** and **(Friday and Thursday)** are **the safest days** in the week. But we can't be precise about which is best out of these two days. So, we are performing risk score evaluation.

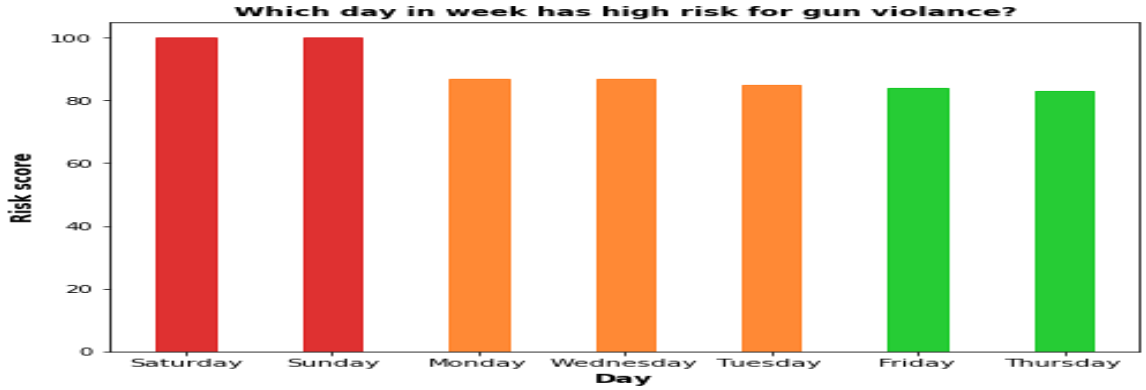


Fig 27: Dangerous day of the week

Observations:

- With the help of this risk score analysis we can conclude that **Saturday is more vulnerable when compared to Sunday** after providing proper weightage to all parameters.

- With the help of this risk score analysis we can conclude that **Thursday is safer when compared to Friday** after providing proper weightage to all parameters.

10.2) Which of the two months is more dangerous and safest month of the year?

We Predicted that **(January and March)** are the **most dangerous month** and **(April and November)** are **the safest month** of the year. But we can't precise about which is best out of these two months. So, we are performing risk score evaluation.

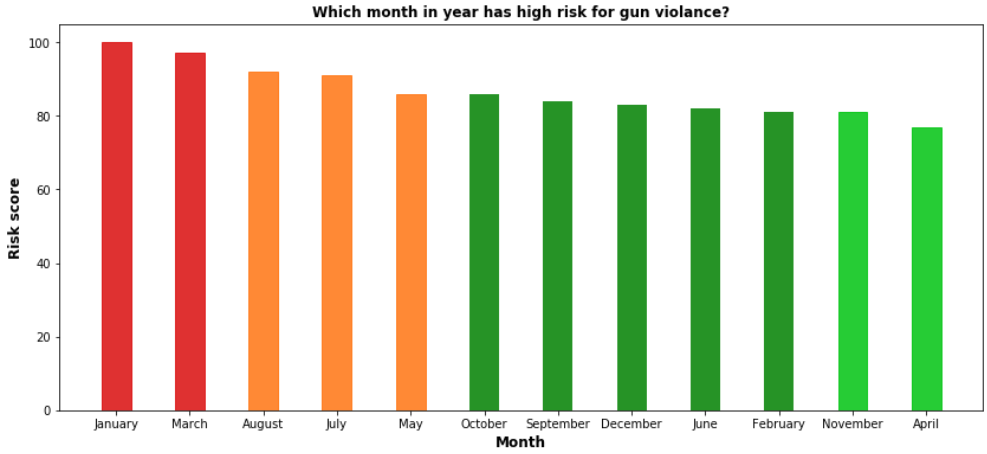


Fig 28: Dangerous month of the year

- **January** is the month with a greater number of gun incidents records. So, it is more dangerous than the month March.
- **April is the safest month of the year in Unites States**. These months has very less number of crime incidents when compared to November and ranked at the bottom of the most dangerous state in the US.

10.3) Which USA State has high chance of Gun violence based on Risk score generated?

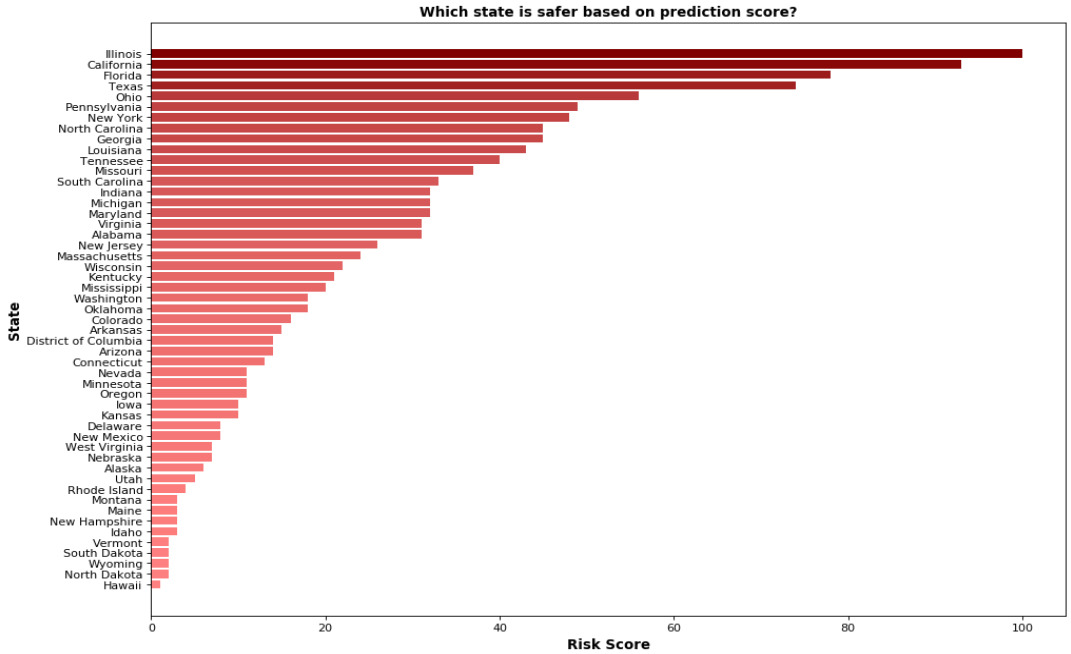


Fig 29: State rank based on risk score prediction

Observations:

- From the above picture**, Illinois, California, Florida** are the top three states with high number of crime incidents, and it is the **most dangerous states** in the US.
- **Hawaii, North Dakota, Wyoming** are the three states with a smaller number of crime incidents and it is the **most safest states** in the United States.

10.4) How Does the safety ranking of USA state affect based on Risk score generated?

Observations:

- **Massachusetts** is the state which got affected more in this risk score generation. This state moved **five places on top** from the old rank in the most dangerous state list.
- **Arizona** is the state which moved **five places to bottom** from the old rank in the most dangerous state list.

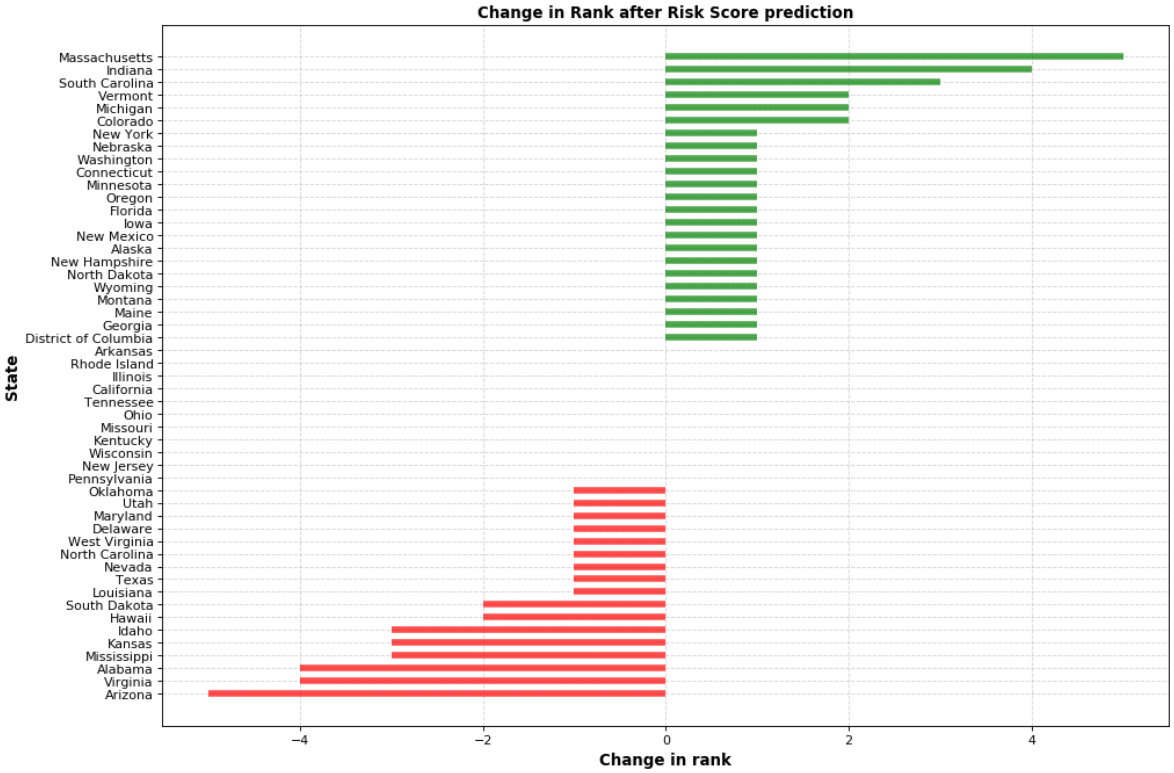


Fig 30: Change in rank based on risk score prediction

As of now we have evaluated and analyzed the dangerous and safest states in US based on Gun incidents records happened. But there are some reasons avails behind the screen which leads to these types of incidents. There are many of them, but we are taking few of those factors and analyzing the influence of those factors on the dangerous and safest state in United States. Factors includes Population, Gun Laws, Literature Rate, Suicide Rate.

11. Risk score calculation involving Population, Gun Permit law, Higher education rate and Suicide rate.

The risk score calculation is used for predicting (the most dangerous state and city, the safest state and city, the dangerous and safest month of the year, and the dangerous and safest day of the week). This risk score calculation involves number of fatalities, incidents guns involved and some external parameters like US gun laws, population, suicide, and literature rate.

11.1) Risk score calculation involving Population and Gun law permit

The weightage includes **0.05 for both harmless incidents and number of persons gets injured, 0.10 for both fatal incident and gun involvement and 0.35 for both death rate and gun permit.** Since death rate is high in the states where the people can use their guns legally. So, here death rate and gun permit yields more weightage.

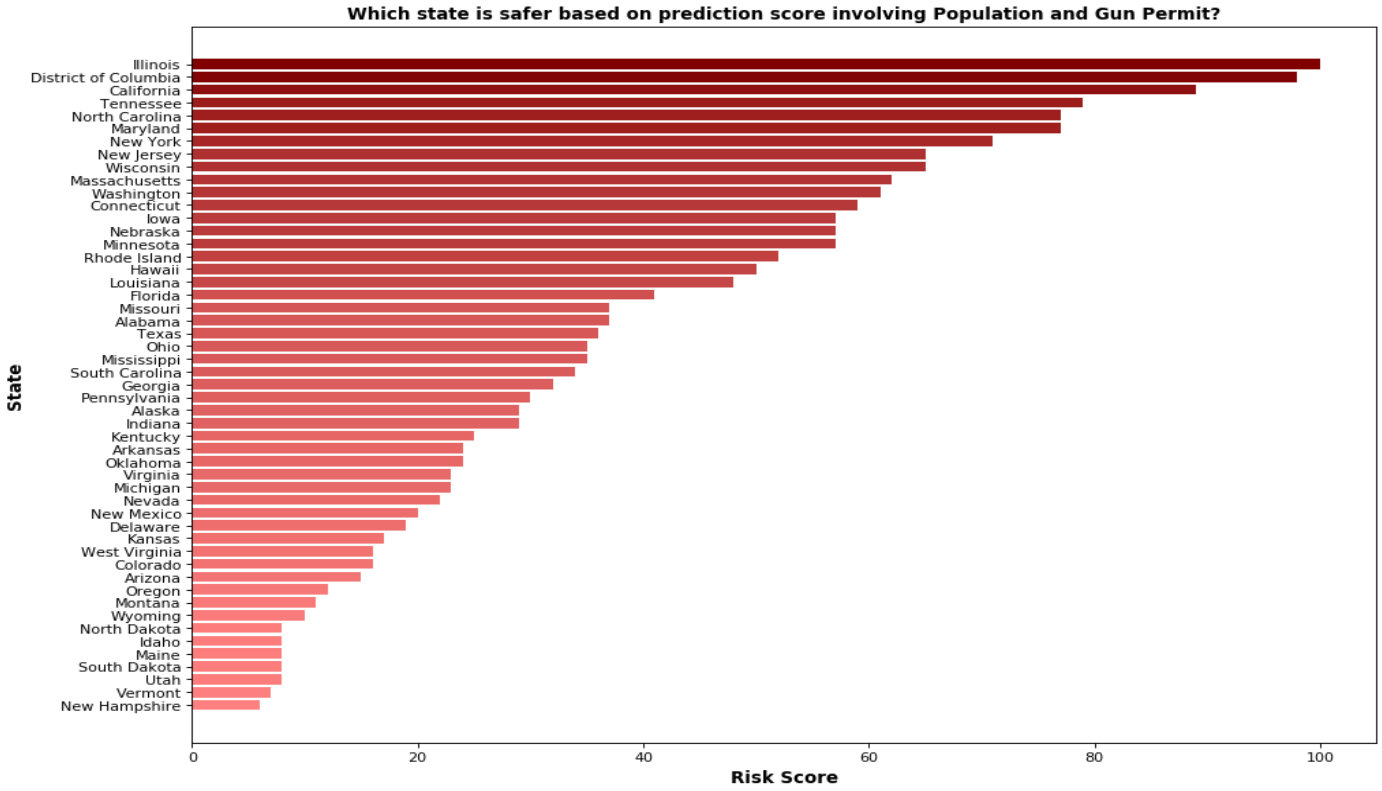


Fig 31: Prediction score involving Population and Gun law

Observation:

- **Illinois, District of Columbia and California** are the three states which ranked top in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Gun Law.**
- **New Hampshire, Vermont, Utah** are the three states which ranked bottom in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Gun Law.**

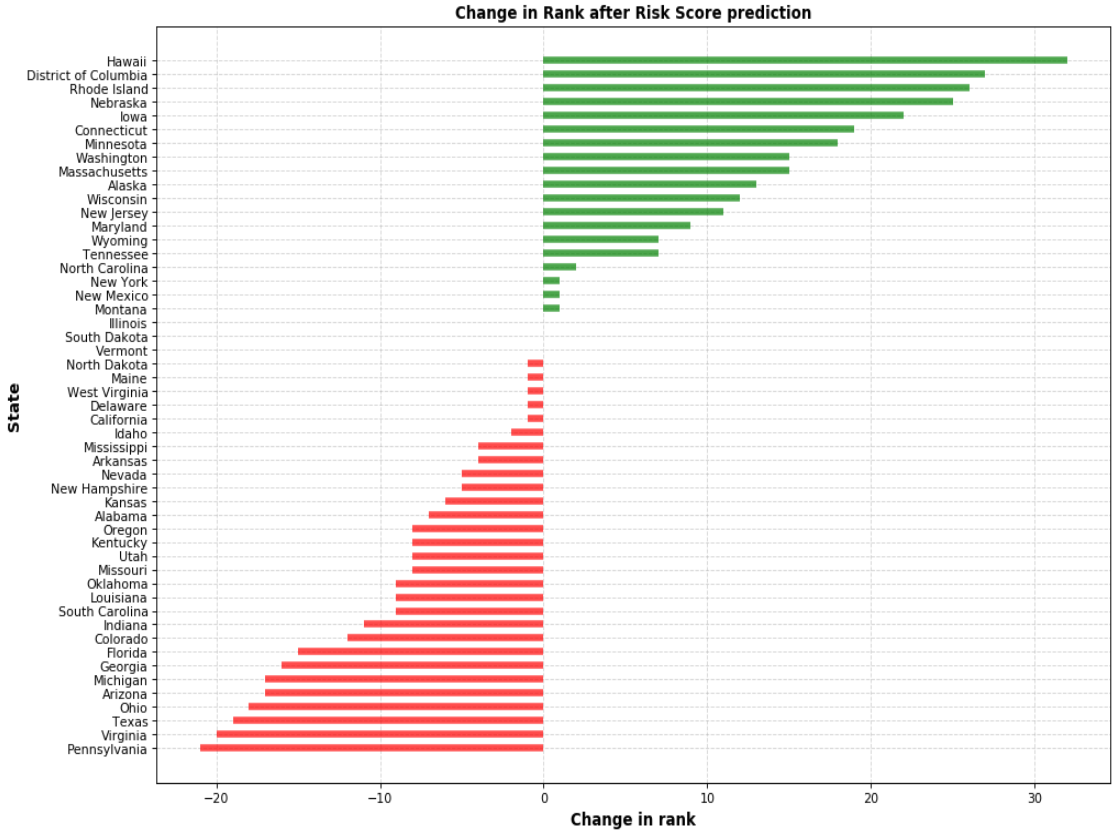


Fig 32: Change in rank involving Population and Gun law

Observations:

- **Hawaii, District of Columbia, and Rhode Island** are the top three states which moves forward in the ranking after the risk score prediction.
- **Pennsylvania, Virginia and Texas** are the least three states decrease its rank after the risk score prediction.

11.2) Risk score calculation involving Population and Higher Education Rate

The weightage **includes 0.05 for both harmless incidents and number of persons gets injured, 0.10 for both fatal incident and gun involvement and 0.35 for both death rate and education.** Since death rate is high in the states where the people literature rate is low. To check the influence of education rate and death rate in this gun violence culture we are providing high weightage to death and education rate.

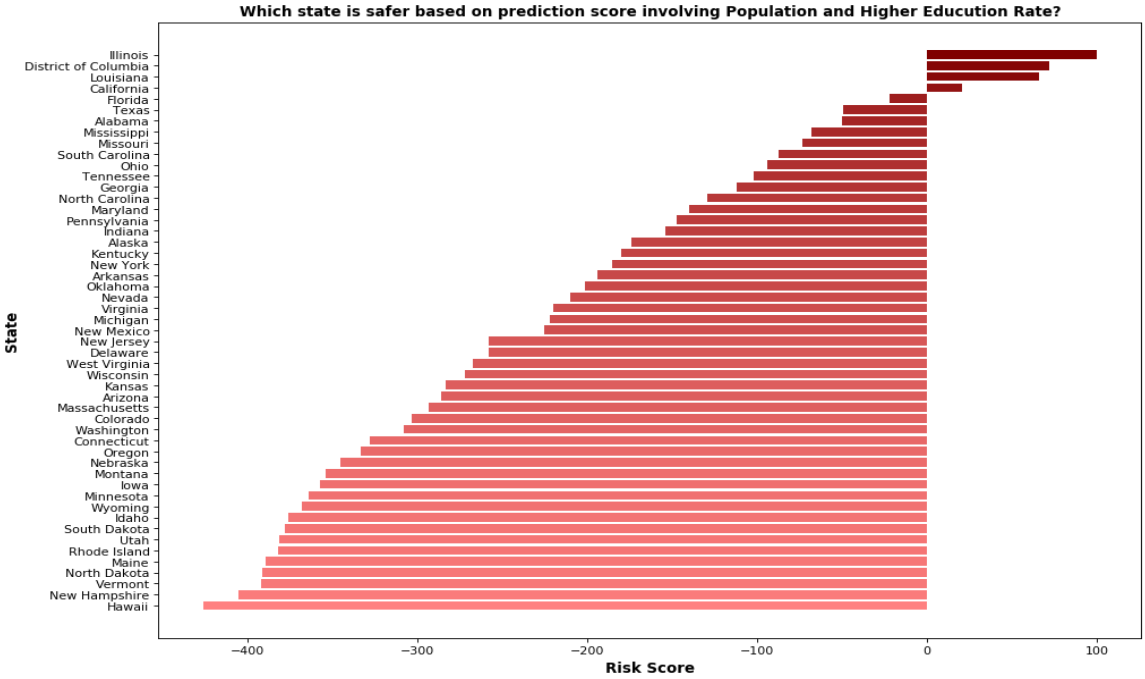


Fig 33: Prediction score involving Population and Higher Education

Observations:

- **Illinois, District of Columbia and Louisiana** are the three states which ranked top in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Higher Education.**
- **Hawaii, Hew Hampshire, Vermont** are the three states which ranked bottom in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Higher Education.**

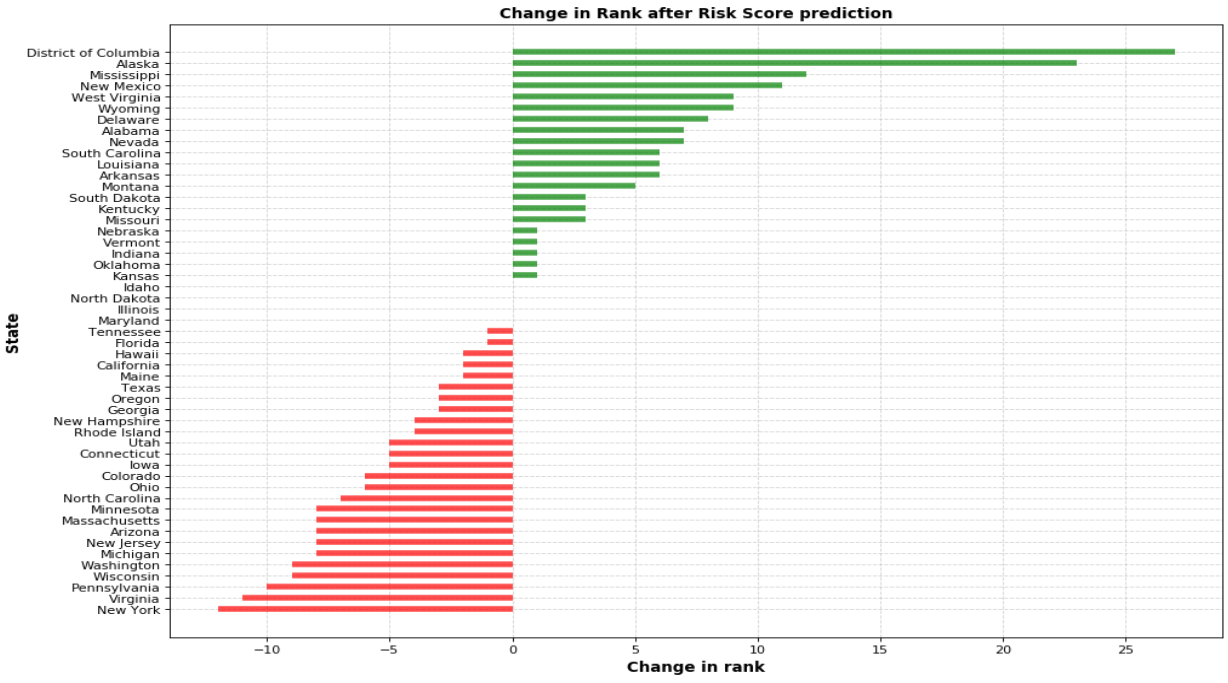


Fig 34: Change in rank involving Population and Higher Education

Observations:

- **District of Columbia, Alaska and Mississippi** are the top three states which moves forward in the ranking after the risk score prediction.
- **New York, Virginia and Pennsylvania** are the least three states decrease its rank after the risk score prediction.

11.3) Risk score calculation involving Population and Suicide Rate

The weightage **includes 0.05 for both harmless incidents and number of persons gets injured, 0.10 for both fatal incident and gun involvement and 0.35 for both death rate and suicide rate.** Since death rate is high in the states where the suicide rate is high. To check the proportionality between the death rate and suicide rate in this gun violence culture we are providing high weightage to death and suicide rate.

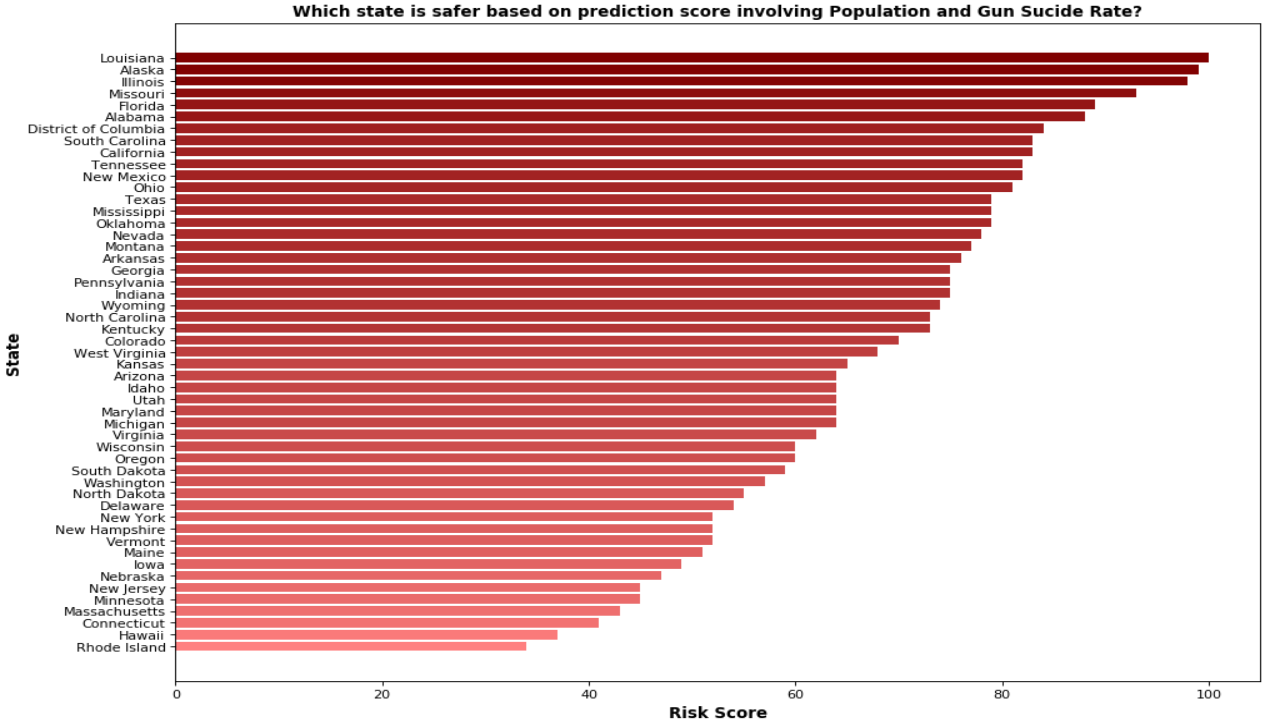


Fig 35: Prediction score involving Population and Suicide Rate

Observations:

- **Louisiana, Alaska, and Illinois** are the three states which ranked top in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Suicide Rate.**
- **Connecticut, Hawaii, and Rhode Island** are the three states which ranked bottom in the most dangerous state list after allocating proper weightage to all parameters with inclusive of factors **Population and Suicide Rate.**

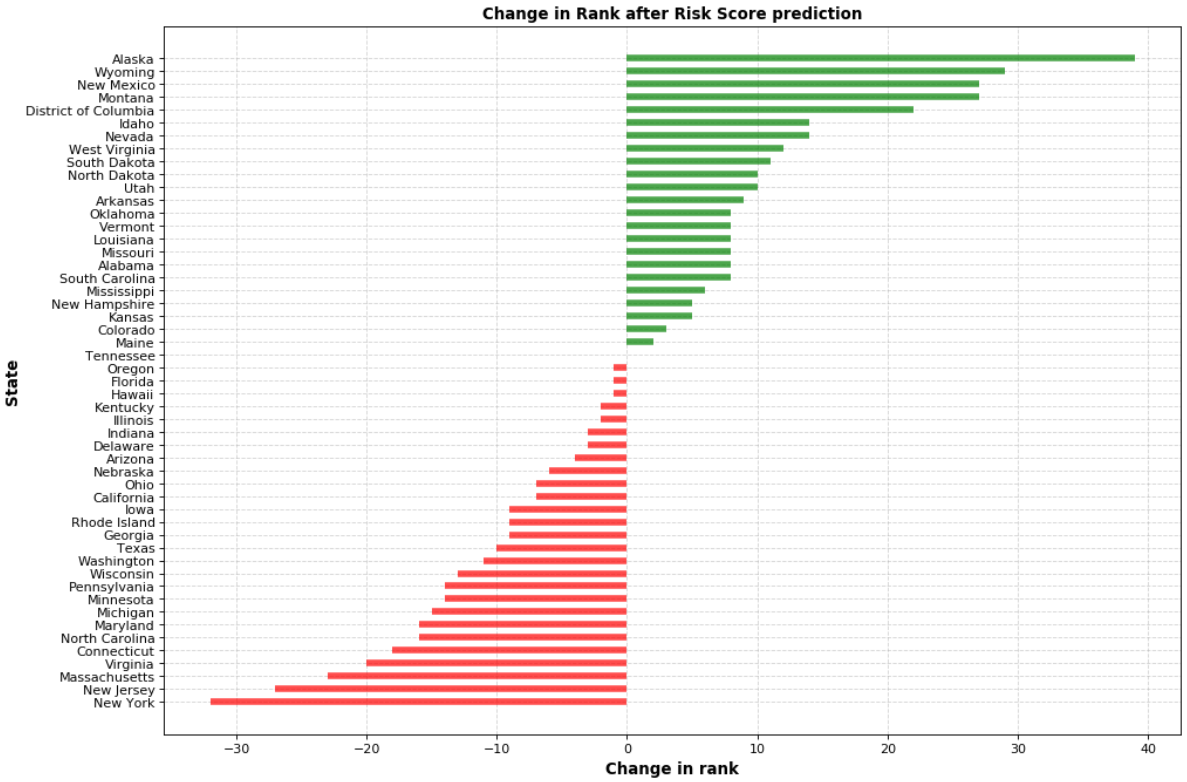


Fig 36: Change in rank involving Population and Suicide Rate

Observations:

- **Alaska, Wyoming, New Mexico** are the top three states which moves forward in the ranking after the risk score prediction.
- **Massachusetts, New Jersey, and New York** are the least three states decrease its rank after the risk score prediction.

11.4) Risk score calculation involving Population, Gun Permit Law, Higher Education Rate, and Suicide Rate

The weightage includes **0.05 for (harmless incidents, fatal, gun involvement, and number of persons gets injured) , 0.20 for (death rate, permit, higher education and suicide rate).** Here, we are including all the four parameters with equal weightage to check the influence in gun violence.

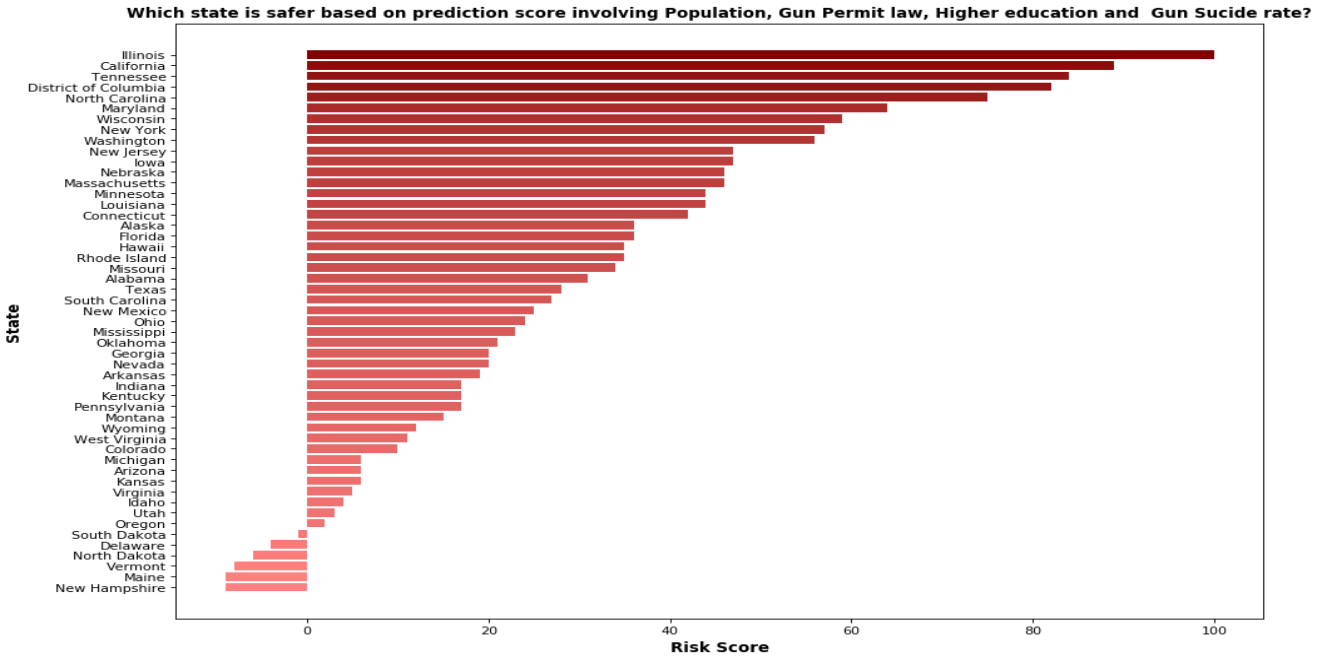


Fig 37: Prediction score involving Population, Gun Law, Higher Edu and Suicide Rate

- **Illinois, California, Tennessee** are the three states which ranked top in the most dangerous state list with inclusive of all factors.
- **Vermont, Maine and New Hampshire** are the three states which ranked least in the most dangerous state list with inclusive of all factors.

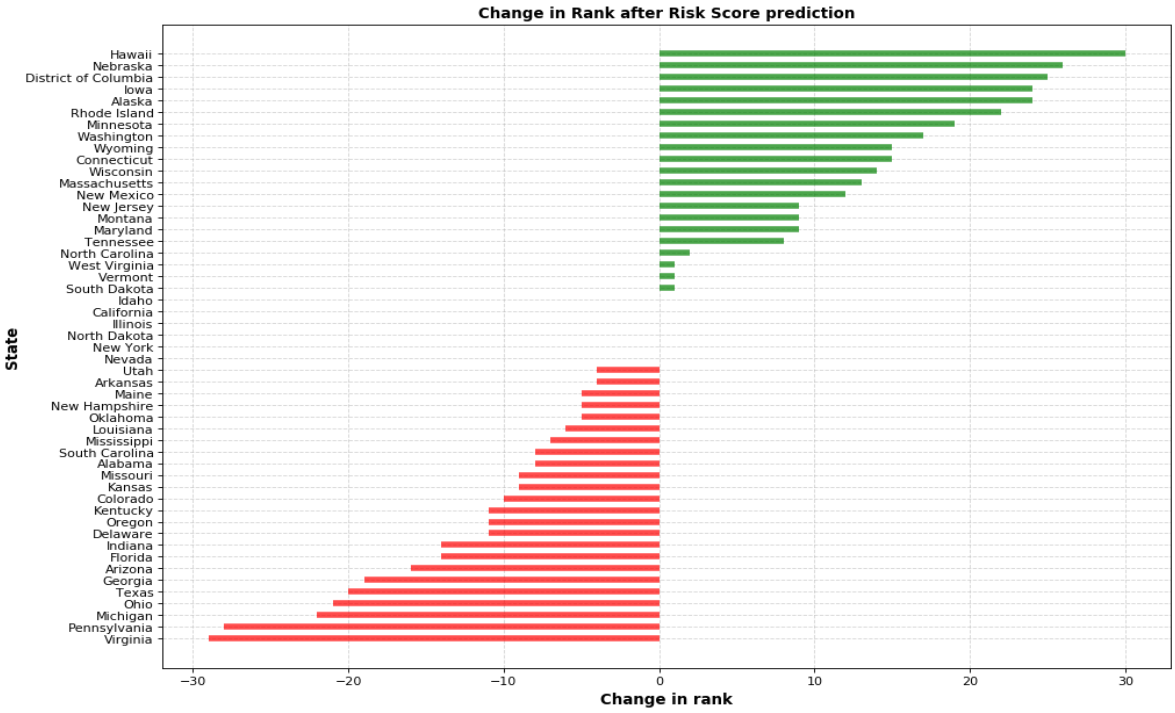


Fig 38: Change in rank involving Population, Gun law, Higher Edu and Suicide Rate

Observations:

- **Alaska, Nebraska, and District of Columbia** are the top three states which moves forward in the ranking after the risk score prediction.
- **Virginia, Pennsylvania, and Michigan** are the least three states decrease its rank after the risk score prediction.

12) Confidence Interval for Suicide rate with the people who has gun permit?

It is conclusive that the 95% CI is (9.17,13.86) after taking 20 samples iteratively for1000 times using bootstrap principle. The overall mean value of Suicide rate for the people with gun permit is 10.691.

Interval range for Suicide rate	95% Confidence Interval	
	Minimum Suicide rate	Maximum Suicide rate
Yes (gun allowed state)	9.17	13.86

13) Confidence Interval for Suicide rate with the people who does not has gun permit?

It is conclusive that the 95% CI is (9.17,13.86) after taking 20 samples iteratively for1000 times using bootstrap principle. The overall mean value of Suicide rate for the people with gun permit is 16.371.

Interval range for Suicide rate	95% Confidence Interval	
	Minimum Suicide rate	Maximum Suicide rate
No (gun not allowed state)	15.755	19.83

From the research 12 and 13, it is conclusive that the people with the states allowed to use their guns legally has low suicide rate. On the other hand, states with no gun permit has high suicide rate.

14) Confidence Interval for literature rate with the people who has gun permit?

It is conclusive that the 95% CI is (86.94,91.01) after taking 20 samples iteratively for1000 times using bootstrap principle. The overall mean value of literature rate for the people with gun permit is 88.85.

Interval range for Suicide rate	95% Confidence Interval	
	Minimum Suicide rate	Maximum Suicide rate
Yes (gun allowed state)	86.94	91.01

15) Confidence Interval for literature rate with the people who doesn't has gun permit?

It is conclusive that the 95% CI is (87.16,90.35) after taking 20 samples iteratively for1000 times using bootstrap principle. The overall mean value of literature rate for the people with gun permit is 89.93.

Interval range for Suicide rate	95% Confidence Interval	
	Minimum Suicide rate	Maximum Suicide rate
No (gun not allowed state)	87.16	90.35

From the research 14 and 15, it is conclusive that the people in the states with high literature rate are not allowing to use the guns legally, henceforth the suicide rate is also low in those states as per the previous observation. It is evident that literature plays an vital role in this gun violence. On contradiction, people from the state with low literature rate are using the guns legally.

8. Conclusion and Future Work

8.1 Summary of the Thesis

In this thesis, we have analyzed and visualized the gun shooting incidents which happened across the United States for the year 2013 – 2018. In the first chapter of this thesis, we gave a brief introduction to the problem that we are trying to address, followed by what are the objectives that we have in mind to achieve by the end of the project and finally the overall solutions that we came up with. In the next chapter we gave a short note about the libraries and the prerequisites that we have installed and used to analyze this gun US shooting dataset. Followed by there is a brief introduction about the machine learning concepts we have implemented for the prediction analysis. Clustering is one the machine learning predictive analysis we opted for predicting the future with the help of US gun shooting records which avails in our database. Linkage clustering is another method which is also used for prediction. In fourth chapter, of this thesis there would be a short intro about the dataset which we have used for this thesis and the details about each column in the dataset. Geo Visualization of this US gun shooting records for the year 2013 – 2018 is done with help of three dimensional geo visualization library called Deck.Gl and the installation, tools and libraries need for this visualization is explained briefly in the fifth chapter of this thesis. At last, the finding, and the prediction that we computed from this dataset is explained with some research questions.

8.2 Overall Evaluation

The main objective of this thesis is to find the most dangerous and safest states in the United States of America in terms of gun shooting incidents or from spreading gun culture. Apart from this, we come up with the statistics like which age category who actively gets participated in these crime incidents, and the gender-based analysis shows which gender involved more in this gun shooting incident. In addition to that we came up with the most dangerous and safest month of the year and day of the week. Before starting our data analysis, we had successfully done with the data cleaning an important pre requestees for data analysis. Because data cleaning is the first and fore most step that must be completed before starting our data modification. Data modification is the step which is the backbone of creating customized variables which is very much useful to answer our research questions. Customized columns like Day, Month and Year holds the value which gets extracted from the column Date. Finally, the geo locations of these crime incidents are visualized with the help of **Deck.gl** a visualization library, where we can get the 3D picture of the crime location.

After analyzing the US gun shooting dataset, it is conclusive that Illinois is the most dangerous state where lot of gun shooting incidents happened when compared to other states in US, whereas Hawaii is the safest state in US. This analysis is purely based on gun shooting incidents and it does not include any other factors. Once, after involving the factors like Population, Literature rate, Gun Laws and Suicide rate we come up with slightly different conclusions. Even though Illinois is the most dangerous state after adding the factor population and guns laws, the proportion of crimes with respect to population is high in District of Columbia and New Hampshire is the safest state. In the states like Illinois and District of Columbia, American government permits people to use their guns legally. This might be a reason in the increasing number of crimes.

“Education is the most powerful weapon which you can use to change the world.” With this quote we can observe that education is the main factor for controlling the crimes in any countries. So, adding Higher Education rate as a factor Illinois and District of Columbia maintains its position at the top with the label dangerous state. On the other hand, Hawaii is the safest state. In US guns are widely preferred for committing suicides so adding suicide rate as a factor Louisiana is the state with worst suicide percentage and it places at the top of the list in most dangerous state of the US and Rhode Island is the safest state in terms of suicide rate. By incorporating all the four factors it is evident that Illinois is the state with the worst tag name as dangerous state in United States and New Hampshire as the safest state. But in terms of death proportion District of Columbia is the most dangerous state in US.

In addition to that we conclude that January and March are most dangerous months of the year. In these months US encounters some important state holidays like English New year, Orthodox Christmas Day, Epiphany and Martin Luther King Jr. Day etc. On these days’ peoples use to spend lot of their times on pups and restaurants which might be a major cause for these types of gun shooting incidents. Saturday and Sunday are the most dangerous day of the week where most of the peoples spend their times in pups, clubs, and party halls on weekend.

8.3 Future Work

In this analysis we mainly focused on US gun shooting incidents for the year 2013 to 2018.With these 260 K records we came up with some interesting and mysterious findings. But it would be very useful if we got some records beyond the year 2013 so that we would find some answers for the whodunit questions. Adding that some more factors like Un Employment, Per Capita Income also might be an external factor for these types of crimes. So, we try to include these factors in our analysis in the near future by keeping this thesis as the base.

9.References

[1] US gun violence quotes: https://www.goodreads.com/quotes/tag/gun-violence

[2] US Gun Violence Statistics: https://en.wikipedia.org/wiki/Gun_violence_in_the_United_States

[3] .American Gun Laws: https://en.wikipedia.org/wiki/Gun_laws_in_the_United_States_by_state and https://www.bbc.co.uk/newsround/41483003.

[4] . American Population : https://www.census.gov/popclock/

[5] . American gun culture, price of the gun, public opinion about the gun laws, weapons restrictions in each state: https://www.bbc.com/news/world-us-canada-41488081

[6].Machine learning concepts like K-means and linkage clustering: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

[7] . Libraries in Python: https://www.geeksforgeeks.org/python

[8] . Google API creation : https://developers.google.com/maps/documentation/javascript/overview

[9]. Geo visualization using Deck gl: https://fireship.io/lessons/deckgl-google-maps-tutorial/

[10]. Statistics about American suicide rate: https://www.cdc.gov/nchs/pressroom/sosmap/suicide-mortality/suicide.htm

[11]. Statistics about American suicide rate:

https://www.americashealthrankings.org/explore/annual/measure/Suicide/state/DC?edition-year=2018

[12]. Statistics about American literature rate:

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment

10.Appendix

10.1 Github Repository

Predictive analysis on American Gun Violence dataset involving Gun Laws, Suicide and Education rate for the year 2013 – 2018.

https://github.com/Karthickjessy/American-Gun-Violence-Dataset-Analysis-and-Geo-Visualization.git