- https://www.macmillanihe.com/studentstudyskills/page/choosing-appropriate-research-methodologies/ (used this page to help me choose what type of

Analysis of the American Gun Crimes.

American gun laws appear to be out of control. Is this true, or are their gun laws the easy blame for high crime rates? Are the states with stricter approaches to gun laws achieving lower crime rates? Are crime rates influencing the states happiness rates?

Charlene Moore
X15412048

Declaration Cover Sheet for Project Submission

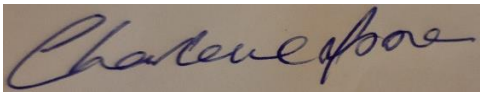
SECTION 1 *Student to complete*

Name: Charlene Moore
Student ID: x15412048
Supervisor: Dr. Eugene O'Loughlin

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature:  **Date:** 6th May 2019

CHARLENE MOORE

National College of Ireland

1 Executive Summary

It is said that Gun Control saves lives (Lopez, 2017). It has also been said that it makes no significant difference to death rates. Which theory is right? This document is a detailed analysis on gun violence in the United States of America. The purpose of this report is to analyse a topic of choice for a final year dissertation, at the National College of Ireland. The motivation behind this document is to gain insights into America's gun violence, and surrounding factors that potentially have impact.

Objectives. To analyse America's gun violence, determining links, trends and patterns within past data on gun violence sourced from Kaggle (Kaggle, 2019), where it was downloaded originally from Gun Violence Archive (GVA) (Gun Violence Archive, 2019). The main objective is to try understanding America's gun laws and gain insights their gun violence rates, determine where gun violence is highest, what time of the year has the highest gun violence rate and what firearms are used most to conduct the crimes. Gun violence could be related to many factors such as drugs, gangs, suicides, mass shootings etc., This document proposes a model that allows law enforcement map out the areas where crime rates are high, along with the types of firearms used in the past in these areas to contribute towards the battle of lowering these rates. With this tool, law enforcements in the areas can be proactive rather than reactive.

Methods. Using GVA's data along with data obtained from the US Census, the selection and exploration method was used to determine links, trends and patterns. In conjunction, a literature review was carried out to examine existing knowledge on gun violence, along with what has been done and how this document can be an addition to the existing body of knowledge.

Results. An examination in trends in the states gun violence rates and the different types of crime committed, revealed that there are trends between alcohol consumption, happiness rankings of a state and the gun laws of a state and their gun violence rates.

Conclusions. There is certainly a link between gun violence laws and their gun violence rates. This requires evaluating to improve rates of gun violence. Identification of the violence categories such as gang, drugs and suicide along with time series forecasting would be a very useful tool to law enforcements. In conjunction with gun sales records and the gun types common in the areas that can be determined based on past trend, law enforcement can be proactive rather than reactive in reducing gun violence rates.

This document is user friendly, providing detailed descriptions on technical terms and visualizations for those with no technical background.

2 Introduction

2.1 Background

Concealed carry and open carry are the two ways that firearms can be carried in America, one being visible and the other hidden (Carry, G. T., 2017). According to (Siegel et al. 2017), laws have been put in place to control the rate of sale, purchases, possession and storage of a firearm in act to prevent gun violence. Most gun laws enacted in America happen at state level across the country, not in the halls of congress and these laws are independent of Federal firearm laws (Carry, G. T., 2017). With the high variation in gun laws, policies and protections on issues such as permits, carry laws, sales and self-defence laws differ between states (Carry, G. T., 2017). State line crossing can be drastically different, one state allowing open carry, the other not. A great example outlined by (Siegel et al. 2017) is California's "discretion when deciding who lawfully can carry a gun, while neighbouring state Arizona allows it's residents to carry a loaded hidden gun without a permit, allows non-residents to get a carry permit through the mail, and does not have expanded background check laws to cover all gun sales.". Such high variation in state laws may be a contributing factor to the high gun violence rates.

An evaluation on all 50 states gun laws by (Siegel et al. 2017) revealed that some states make it too easy for criminals and other people considered dangerous to access guns. Some of the states with weak gun laws consist of Arizona, Mississippi, Idaho, Florida, Wyoming, South Dakota and Alaska. States with stronger gun laws are California, New Jersey, Connecticut, Maryland, Massachusetts, New York and Hawaii. *Figure 1* demonstrates the 50 states ranked in terms of gun laws on an annual gun law score card produced by (Giffords law center, 2018). The scorecard represents state gun laws for today, 2018. It will be interesting to see if there is a trend between the gun laws and the gun violence rates in my analysis like it has shown in the below scorecard.

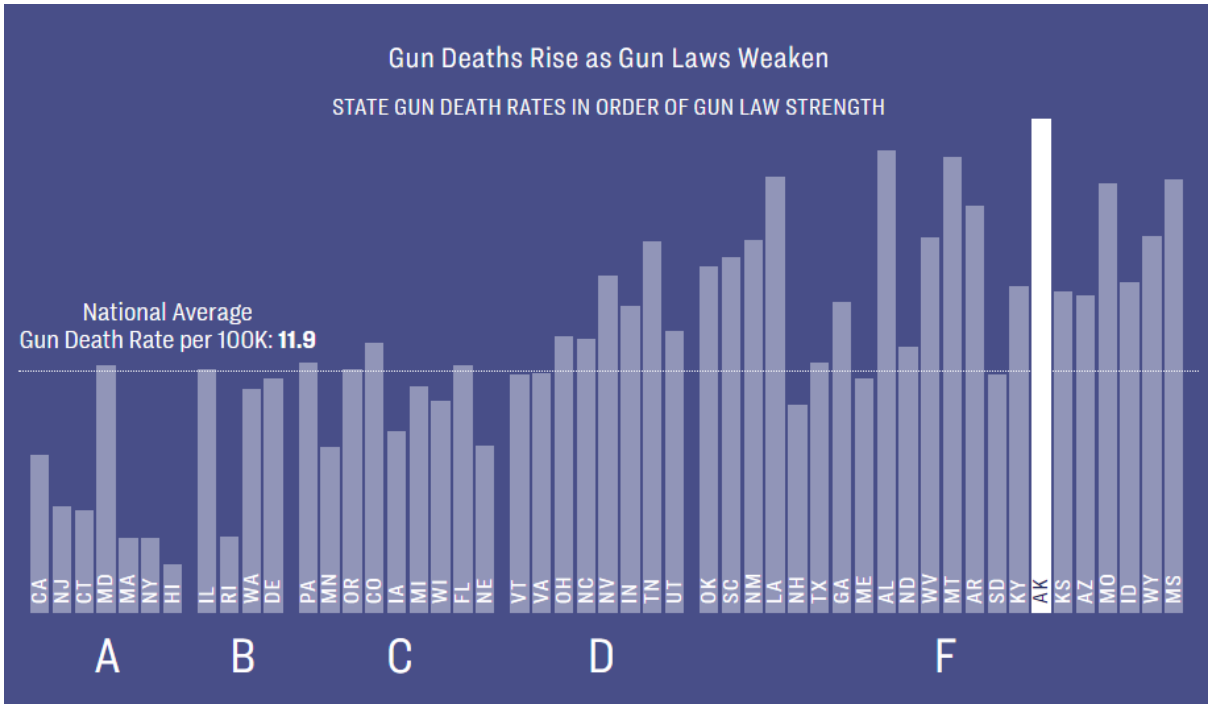


Figure 1: Annual Gun Law Scorecard by (Giffords law center,2018).

The scorecard represents the ranking of gun laws, as well as demonstrating each states gun death rate per 100k people. California ranks 1st in terms of strength of gun laws, followed by New Jersey and Connecticut with considerably low gun death rates. Mississippi and Wyoming rank last, meaning they have considerably weak gun laws according to the scorecard, but high gun deaths. Alaska's gun law is one of the many states with weak gun laws, resulting in the highest gun death rate according to this scorecard.

The problem being addressed in this document is gun violence in America. (Lopez, 2017) state that compared to the developed world, America's gun laws are unique in a sense that they stand alone in the developed country with their gun laws. Lopez highlights the fact that America has more guns than any other country in the world as well as the highest number of gun deaths than any other developed nation.

Both (Giffords law center, 2018) and (Carry, G. T., 2017). highlight the fact that gun laws vary within each state. As shown in the scorecard, California, Minnesota, Maryland, New Jersey and New York are considered strong gun laws. (Guns To Carry, 2018) supports this statement, by stating that these states are the few states that do not have a provision to protect the right to own and bear firearms.

Illustrated in *figure 2* are the stats provided by (Carry, G. T., 2017). on gun laws in states. Private sale NICS checks are required in 18 states. These states are; California, Colorado, Connecticut, Delaware, District of Columbia, Hawaii, Illinois, Iowa, Massachusetts, Michigan, Maryland, New Jersey, Nebraska, North Carolina, Oregon, Pennsylvania, Rhode Island, Washington. Not all states require gun registries. However, to collect data on gun sales. The few states and one district that do require gun registry are California, Hawaii, Maryland, New York and the District of Columbia (Carry, G. T., 2017). Below in *figure 2,* eight states have gun registry. The 8 represents the 5 states that officially require gun registry and the 3 states; Michigan, New Jersey and Washington that collect data on gun sales (Carry, G. T., 2017).

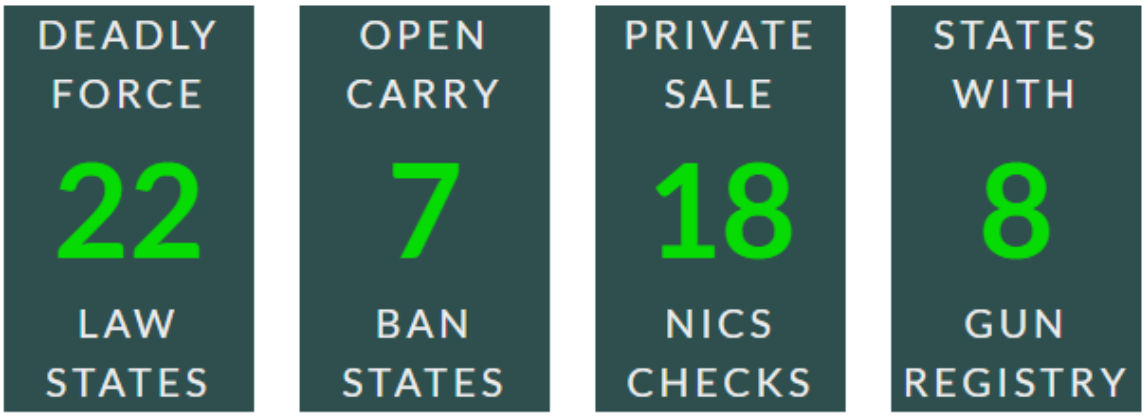


Figure 2: Gun law statistics in each state (Carry, G. T., 2017).

There are 22 states with deadly force laws, 7 states that ban open carry, 18 states that have background checks required to purchase a firearm and 8 states that either collect data on gun sales or require firearm registry.

Gun crimes is a tough issue to study. The topic itself on gun crimes and what factors impact it is controversial itself. Some researchers declare that gun restrictions don't impact gun death rates at all, some reckon the restrictions on gun laws save lives (Lopez 2017). I will delve deeper into these views later in the document, they can be very persuasive.

States happiness have been ranked by Gallup since 2008. The states are ranked in terms of well-being, by factors such as Career, Social, Financial, Community and physical well-being (Inc, Gallup., 2018). Ranking first, seven years in a row; Hawaii has been considered the happiest or lowest well-being state in the united states of America. Ranking among the top ten states, Hawaii has made the cut for 11 executive years, along with Colorado. West Virginia ranked in the lowest well-being for 10 executive years (Inc, Gallup., 2018). It would be interesting to determine if there are trend within these rankings and the states gun violence and gun laws. We will explore these possibilities to feed curiosity.

According to surveys carried out by (RJ Reinhart, 2018) in October of this year, "six in 10 American's support stricter Gun Laws", It is a drop from the survey statistics for March which was 67%. However, it reflects the highest percentage to favour tougher firearms laws in two or more decades. (RJ Reinhart, 2018) Above is one of the questions that were asked in Gallup's survey last October, after the Mass shooting in Vegas. Statistics on whether Americans want the gun laws stricter, to stay the same or weakened. Stricter laws are a higher vote at 61% to the 30% who voted for them to remain the same and the 8% who want the laws to be weakened.

Americans' Attitudes on Gun Control, by Key Demographic Group

In general, do you feel that the laws covering the sale of firearms should be made more strict, less strict or kept as they are now?

	More strict	Kept as they are	Less strict
	%	%	%
All Americans	61	30	8
Gender			
Men	51	38	9
Women	70	22	6
Political affiliation			
Republican	31	55	13
Independent	61	28	7
Democrat	87	10	3
Gun ownership			
Gun owner	38	48	11
Gun nonowner	73	20	6

GALLUP, OCT. 1-10, 2018

Figure 3: Sample Question asked to American's by (RJ Reinhart, 2018).

The results reveal that more women than men protest gun laws to be stricter. More men than women think they should remain the same. And a smaller portion of men and women voted for the gun laws to be weakened, with 9% men and 6% women.

The *inspiration* behind choosing Gun Violence as topic for this document is due to keen interest in the America's gun laws after the Las Vegas shooting, October 1st, 2017. Visiting family with some friends, on the Las Vegas Strip when the tragic event occurred. Luckily, I was not at the concert due to the tickets being sold out. However, I was staying the next resort up from the Mandalay Bay, on the balcony at the time to witness many victims hysterically exiting the direction of the shooting. We were issued a warning to not leave our hotel rooms, due to the suspicion that there was more than one shooter. That day I went from not being aware of the impact guns can have, to being shocked and scared of their capabilities. That day forward I decided I was going to learn more about American gun laws and what defines them. This project presented the perfect opportunity to do this.

The Las Vegas mass shooting, when a single man let fire on hundreds of people at a route-91 country festival, killing 58 people and injuring over 800 others, last October 1st. A tragedy that that touched the heart of people all over the world, raised many questions on gun laws in America. (Economist, The, 2017) The weapon used to carry out this mass shooting, by a gunman called Stephen Paddocks, was a "legal but controversial accessory onto his semi-automatic rifles to enable them to fire hundreds of rounds per minute. Officials say that these devices – known as bump-stocks" were found along with 23 guns inside the gun man's room, on the 32nd floor of the Mandalay Bay Resort and Casino (BBC News, 2017). The fact the guns were legal, considering the fact they were built upon with a "controversial accessory", raises the question, would stricter approaches to gun laws have any effect on lowering the rate of mass shootings like this happening?

The key colour being used throughout this document is orange in the visual representations. The reason for using this colour is systematic as it is represents of seriousness. According to Kandinsky, colour is a powerful tool that can be used to appeal to its target audience. Orange is a justified colour to use in this document as gun violence is a serious matter that deserves serious attention and awareness (Treehouse Island, Inc., 2019).

2.2 Aims

Aim 1: The first aim is to find datasets on gun violence and states population. Addition to this is to gather data or source data on states happiness rankings.

Aim 2: After selecting the data, reading in the data to RStudio and cleansing the data is next. This consists of discarding any columns that are irrelevant to the analysis. Removal of discrepancies in the data such as outliers with given proof of valid reason for removal is done at this stage and throughout the analysis if required.

Aim 3: The third aim is to research and gain understanding as to why it is each state can have such different gun laws and if there is a trend in gun laws and gun violence rates within states.

Aim 4: Exploration of the data for patterns within crime from 2014 to 2017 is a big aim in this document. Patterns such as the day, month and year crime is at peak, the states and cities that have highest crime rate and to delve further into each manner of gun deaths and injuries on which of the three contrasting types of gun mortalities is; accident, suicide, or homicide. Determining if there is a trend in a state's happiness rankings and their gun death rates would be interesting. Although in data analytics we must be sure to bear in mind correlation is not causation.

Aim 5: The fifth aim is to introduce machine learning algorithms into the analysis on gun violence to create a model that will help law enforcement work towards reducing gun violence rates. Such models include Forecasting models such as Arima and Regression analysis such as Logistic Regression and Random Forest.

Aim 6: The final aim in this analysis is to construct and provide a user-friendly document that can insight readers with and without technical backgrounds on America's gun violence and the potential surrounding factors that may be linked.

2.3 Technologies

R: Founded by Ihaka and Gentleman in 1993, R is a functional programming language for statistical computation and graphics. R has integrated help systems which is fantastic for support ("R: What Is R?" n.d.). R is used throughout this analysis.

RStudio: RStudio is an Integrated Development Environment (IDE) for statistical computing and graphics. R statistical language is used in conjunction with RStudio scripts for this analysis.

Microsoft Excel: Built to create grids, text, numbers and formulas specifying calculations, Microsoft Excel is a spreadsheet program, built to supports files in the CSV format. Files which are CSV format can be imported to and exported from programs like Excel. My dataset on Gun Violence from Kaggle is csv format. Excel is a great tool for creating quick and simple charts and for conducting statistical tests on data.

Tableau: Tableau is a visual business intelligence tool used for creating data visualizations, publishing data sources as well as workbooks to Tableau Server. Once findings are found, tableau will be used to display my findings.

Canva: A website that provides templates for posters. Used to create a poster promoting the main points in this analysis.

GitHub: An open source version control system, used to store all work throughout the project.

2.4 Structure

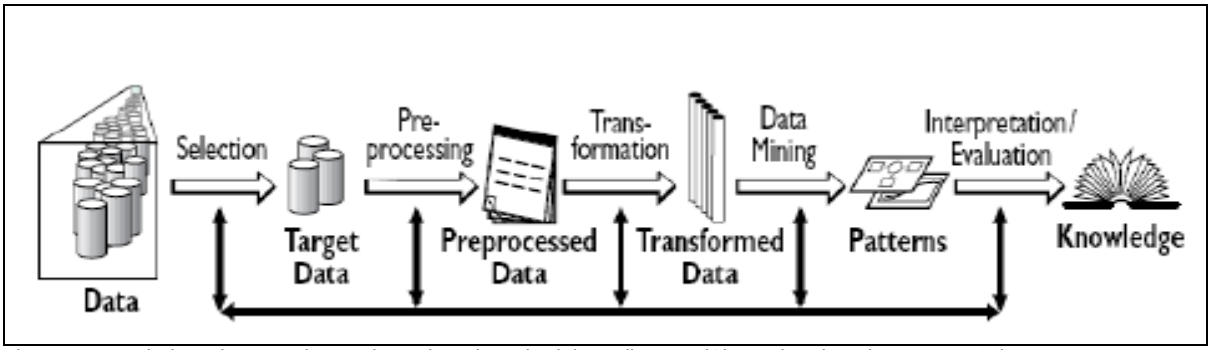


Figure 4: Knowledge Discovery in Database (KDD) Methodology diagram (Shawndra.pbworks.com., 1996).

Throughout this document the data analytics methodology, Knowledge discovery in Databases (KDD) was applied. This process is for extracting useful knowledge from Volumes of Data. (Shawndra.pbworks.com., 1996)

KDD focuses on the overall process of knowledge discovery from data, including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall human-machine interaction can be modelled and supported. (Shawndra.pbworks.com., 1996)

KDD has a series of important stages to complete in order to achieve objectives.

Data Selection: The data for this document was selected from multiple datasets on Kaggle, US Census Bureau and Gallup. This data plays a vital role in achieving the objectives outlined in this document. Data scraped and provided by my supervisor Dr. Eugene O'Loughlin is also included in this analysis. The data is Happiness Ranking of states in America, for the years 2012 and 2013. An overall analysis of the data was conducted, followed by a closer look into selected states individually.

Pre-processing: The data was pre-processed in R, cleansed and transformed when required, to be consistent and suitable. This included detecting, correcting or removing corrupt or inaccurate data records from the datasets. This stage is very important in preparation for algorithms and models to run smoothly. Mice package was explored at this stage, and later decided against due to R struggling to handle the large amount of data in the gun violence dataset. Subsets and cleansing throughout the analysis were applied alternatively.

Transformation: RStudio was used to read the data in to its environment from the csv files to build a transformed dataset. Tibble was applied at this stage for a nicer printing method, which is useful when working with datasets that are large like gun violence.

Data mining: During this phase data mining techniques were applied using RStudio. Data mining was used to translate the problems and questions in my project into effective results via statistical and visual outputs. Data mining is the process of extracting patterns (models) from data. Patterns provide us with the tools to make predictions (Shawndra.pbworks.com., 1996).

Interpretation/ Evaluation: For the evaluation stage, the results of the previous stages are interpreted and visualised through Tableau and R using visualization packages ggplot2 and plotly. In completion of this stage, the answers to the questions asked at the beginning of the analysis should be answered. If the answers are not found, the stages are repeated until the objectives of the analysis are met. Different methods of evaluation were applied throughout, such as statistical testing and Cross Validation.

2.5 Literature Review

2.6 Acronyms, Definition and Abbreviations

Missing Completely at Random (MCAR) refers to data that has no relationship between the missingness of the data and any values, observed or missing. In other words, the value of that missing data does not impact the rest of the data values.

Missing Not at Random (MNAR) data is when missing values on a variable are related to the values of that variable itself, even after controlling the other variables (Eekhout, Iris., 2019). For example, a variable for gender contains the value 1 and the rest are missing values. This is a case of data being MNAR. 1 would represent a gender, let's say male, and then the missing values would be female. Therefore, removing all the rows of data with 0, would leave you with a dataset recording only male values.

To assist in the reader's understanding throughout this document included is a list of acronyms, definitions and abbreviations.

KDD stands for Knowledge Discovery in Databases. Knowledge discovery in Databases is the process of discovering knowledge using several techniques in fundamental stages that include data selection, pre-processing of the data, Transformation of the data, data mining and Interpretation and Evaluation of the results.

Gun Violence Archive (GVA) is a non-profit corporation that provides free online public access to information gun-related violence in the United States (Gun Violence Archive, 2019).

Kaggle is a platform for researchers and statisticians to complete competitions which are based on Machine Learning, Data Science, Deep learning or AI related (Kaggle, 2019).

US Census Bureau's mission is to serve as the nation's leading provider of quality data about its people and economy (Bureau, US Census., 2019).

Machine Learning is an application of artificial intelligence (AI) that provide systems the ability to automatically learn and improve from experience without being explicitly programmed (Expert System, 2017).

Multivariate Imputation via Chained Equation (MICE) is a popular package used for creating multiple imputations as opposed to single imputations such as the mean. It is useful for many reasons, one being that it takes care of uncertainty in missing values (Analytics Vidhya, 2016).

Statistical testing consists of a Null Hypothesis. A Null Hypothesis (HO) proposes that no significant difference exists in given observations. Alternative is what we are testing for, to reject the HO. The Alternative Hypothesis (HI) states that there is a significant difference in the given observations. If a difference if found, the HO is rejected. If there is no significant difference found, we fail to reject HO, accepting the Alternative Hypothesis (HI). For example:

> *Null Hypothesis:* Given two sample means are equal, there is no significant difference.

Alternative Hypothesis: Given two sample means are not equal, there is a significant difference.

Cross Validation is a method to validate the stability of your models results. This can be done with the error estimation for the model after training. The disadvantage of this technique is that the validation is valid for the data being trained. It is difficult to determine how well the model will work with unseen data (Gupta, Prashant., 2016).

Google Scholar is a platform that allows users to search for articles, theses, books, literature reviews and abstracts. This is a very useful tool when conducting research throughout projects. It is here and *the National college of Ireland's library* search engine where the foundations of this document were researched.

Time Series Analysis refers to the use of statistical methods to analyse time series data and extract meaningful statistics and characteristics about the data. This document discusses an autoregressive integrated moving average (*ARIMA)* model at a later stage.

ARIMA is a model that is used to better understand the data or to make forecasts. Arima basis it's predictions of future values of a time series using a linear combination of its past values and a series of errors (Smarten, 2018).

Dropbox is a workspace used to store files on its servers and when changes are made it automatically updates the files (Kapocsi, 2018)

2.7 Challenges

2.7.1 Limitations on RStudio:

Challenge: R has limitations with handling large datasets. This is since all computation is carried out in the main memory of the computer.

Solution: I have decided to read my files in to r individually to avoid RStudio from alternatively crashing on me.

2.7.2 Computation efficiency:

Challenge: R packages such as Boruta and Mice require a lot of time and ram torun efficiently on large datasets.

Solution: To avoid long waits for code to run to progress with my analysis, borrowed a second laptop to use to conduct research on and progress with this document while the packages ran to optimise time.

2.7.3 Compatibility:

Challenge: Ensuring data is compatible was a challenge. Datasets with different timelines.

Solution: To overcome this challenge and ensure this analysis was within the same timeline of 2014-2017, resulted in excluding some datasets and searching for alternatives.

2.7.4 Project Restrictions

Data: Gun Violence contains 260k records of data on gun violence in America. There aren't many datasets out there that contain this many detailed records. This dataset is large and easily accessible through Kaggle. Happiness Ranks for all states across America isn't as easily accessible as one would think. Accessing data for the required years 2014 until 2018 was not straight forward, resulting in a gap in this analysis that has been added to future work.

Time: The data timeline is from 2013 to 2018. 2013 and 2018 were not as applicable in the analysis as the other years due to 2013 not containing enough gathered data on crime. And 2018, only having record of the first yearly quarter.

Software: No special software is required at this time. But if pursued in being used as a tool for detecting crime; costs and additional software may be needed.

2.7.5 Missing values

Challenge: Some machine learning algorithms don't work well with missing values, resulting in poor performing models. Figure ?, shows percentages of missing data in each variable of the gun violence dataset. Over 5% is considered a lot.

incident_id	date	state
0.0000000000	0.0000000000	0.0000000000
city_or_county	address	n_killed
0.0000000000	6.8829846711	0.0000000000
n_injured	incident_url	source_url
0.0000000000	0.0000000000	0.1952619765
incident_url_fields_missing	congressional_district	gun_stolen
0.0004172264	4.9837698913	41.5136140989
gun_type	incident_characteristics	latitude
41.4940044560	0.1364330477	3.3056851275
location_description	longitude	n_guns_involved
82.4389389097	3.3056851275	41.4935872295
notes	participant_age	participant_age_group
33.8024349335	38.5095836915	17.5735778837
participant_gender	participant_name	participant_relationship
15.1716052370	51.0076018658	93.4186700490
participant_status	participant_type	sources
11.5267150093	10.3739183404	0.2545081317
state_house_district	state_senate_district	
16.1771209706	13.4914343411	

Figure 5: Missing data percentages.

Using R, the percentages were calculated for the missing data in each variable. As you can see participant_relationship is missing 93% of its values. Anything above 5% is considered not good. This is a given reason to remove this variable from the analysis. Location_description too is missing a large amount of data at 82%. The variables gun_type, notes, participant_name, n_guns_involved and gun_stolen all have high percentages of missing data values. Due to our data being so large and a record of 51 states including District of Columbia, we will not be removing any variables as it is not evident yet where exactly the missing values lye and if they are missing completely at random (MCAR) or (MNAR).

Solution: Mice package was tested to attempt to impute missing values with multiple imputations as opposed to single imputation like the mean value. Due to the data set being so large and the data not missing at random (MAR). MICE was not the most practical solution. Instead, following KDD; data was cleansed, and missing values were handled throughout the analysis using the appropriate methods. With gun violence containing 260k records of data in most cases missing values were not a issue. Missing values tend to be big issues with small samples of data.

Challenge: Understanding my dataset wasn't the easiest. The `participant_age` and `participant_status` variables along with a few others were hard to cleanse and understand due to there being more than one value per observation. For example, if there were 3 participants, there were three ages. In some cases, only 2, if the age for the third wasn't known.
Solution: I studied the dataset thoroughly, trying to make sense of it. Researching the data source, along with figuring out ways to cleanse data in this way.

Challenge: Gun Control is a tough issue to study. There is so much research to do, before I could even move on to the cleansing and preliminary analysis stage of my project. I spent more time than planned on researching Gun Crimes, State Laws, and articles on past crimes.

Solution: I adjusted my plan to assure the extra time on research didn't impact other stages in the project. I will now take the time after exams, and before the start of semester 2, to touch up on what was affected by this. E.g. My Gantt Chart isn't to the standard I would like it to be.

Challenge: Picturing how my project will end was hard due to some tools and techniques I have not learned yet. For Example, Machine Learning. This is certainly going to be the biggest challenge to conquer when completing this project.

Outcome: I researched Machine Learning Algorithms whilst defining the scope of my project. With hard work and attending my lectures in semester 2 for the Data and Web Mining which covers Machine Learning, I should hopefully be fine in achieving my objectives, creating a Machine Learning algorithm to predict when crime is likely to occur based on trends from my model of crime patterns.

Challenge: The amount of observations for each state differs in size. Of course, this is normal as some states are bigger than others. I found this a challenge to grasp at first. I created charts and did a t Test on data that was selected wrongly. This is

nothing but trial and error. But again, it cost me some delicate time. It was a challenge to ensure the data was “like by like” for my tests to be accurate. Calculating the rate per state population was something I couldn’t grasp at first. But after research and discussing it with my supervisor I managed to achieve the percentage of crime rates by dividing the number of reported crimes by the total population of each state; the result was multiplied by 100,000. For example, in 2013 there were 16,307 shootings reported in California and the population was 38,347,383. This equals a robbery crime rate of 149.6 per 100,000 general population.

Outcome: I was suspicious of my approach to the t Test. As selecting data like I did didn’t make logical sense. I asked my supervisor and it was proved I had made a mistake. My supervisor guided me on how I should select my data for my tests. I also watched a few videos on YouTube to be sure I knew what I was doing.

3 System

3.1 Requirements

The following section consists of the requirements required for the analysis to be completed.

3.1.1 Data

The data requirements to meet objectives are:

The datasets are open source and are applicable in an analysis of this sort. To analyse crime rates per state population data was required. To achieve insights into trends relating states happiness, data was also required on states happiness rankings for the timeline of the analysis. Figure 5, 9 and ? illustrate the structure to the three datasets used. As mentioned above, these datasets were sourced from Kaggle, US Census and Gallup.

```
Observations: 239,678
Variables: 29
$ incident_id                <dbl> 461105,...
$ date                       <chr> "01/01/...
$ state                      <chr> "Pennsy...
$ city_or_county             <chr> "Mckees...
$ address                    <chr> "1506 V...
$ n_killed                   <dbl> 0, 1, 1...
$ n_injured                  <dbl> 4, 3, 3...
$ incident_url               <chr> "http:/...
$ source_url                 <chr> "http:/...
$ incident_url_fields_missing <lgl> FALSE, ...
$ congressional_district     <int> 14, 43,...
$ gun_stolen                 <chr> NA, NA,...
$ gun_type                   <chr> NA, NA,...
$ incident_characteristics   <chr> "Shot -...
$ latitude                   <dbl> 40.3467...
$ location_description       <chr> NA, NA,...
$ longitude                  <dbl> -79.855...
$ n_guns_involved            <dbl> NA, NA,...
$ notes                      <chr> "Julian...
$ participant_age            <chr> "0::20"...
$ participant_age_group      <chr> "0::Adu...
$ participant_gender         <chr> "0::Mal...
$ participant_name           <chr> "0::Jul...
$ participant_relationship   <chr> NA, NA,...
$ participant_status         <chr> "0::Arr...
$ participant_type           <chr> "0::Vic...
$ sources                    <chr> "http:/...
$ state_house_district       <int> NA, 62,...
$ state_senate_district      <int> NA, 35,...
```

Figure 6: Gun Violence data structure.

The Gun Violence dataset consist of over 260k records initially. When reading the data in from R, an argument was set to replace empty strings i.e. (“ ”) with NA’s to help in the process of cleaning and understanding our data rather than having both NA’s and empty strings. Gun Violence contains four data types; numeric (dbl), characters, integers and logical values. Coercion of data types was carried out throughout the analysis when required. For example, date variable was converted to a Date datatype. This will all be illustrated later in the report in the analysis section.

The Gun Violence dataset contains records of data for all 51 states including District of Columbia. Figure 7 demonstrates the count of gun violence recorded for each state, followed by a bar chart visually representing this data for easier interpretation in figure 8 and a graphical representation in figure 8.

Alabama	Alaska	Arizona	Arkansas
5471	1349	2328	2842
California	Colorado	Connecticut	Delaware
16306	3201	3067	1685
District of Columbia	Florida	Georgia	Hawaii
3195	15029	8925	289
Idaho	Illinois	Indiana	Iowa
661	17556	5852	2517
Kansas	Kentucky	Louisiana	Maine
2145	4157	8103	907
Maryland	Massachusetts	Michigan	Minnesota
5798	5981	6136	2408
Mississippi	Missouri	Montana	Nebraska
3599	6631	638	1651
Nevada	New Hampshire	New Jersey	New Mexico
1953	964	5387	1645
New York	North Carolina	North Dakota	Ohio
9712	8739	573	10244
Oklahoma	Oregon	Pennsylvania	Rhode Island
3455	2286	8929	895
South Carolina	South Dakota	Tennessee	Texas
6939	544	7626	13577
Utah	Vermont	Virginia	Washington
1072	472	5949	3434
West Virginia	Wisconsin	Wyoming	
1575	4787	494	

Figure 7: A count of gun violence for each state in the Gun Violence dataset.

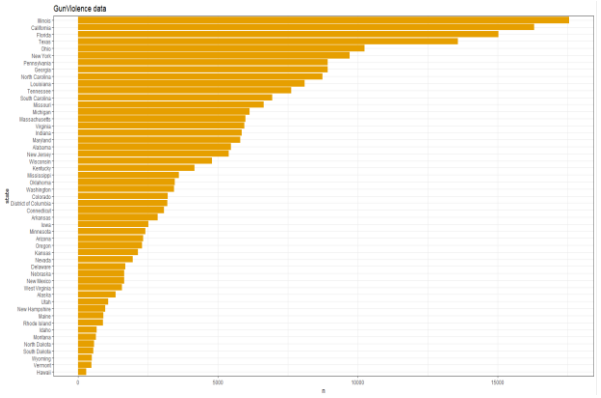


Figure 8: Bar chart of the count of gun violence for each state in the Gun Violence dataset.

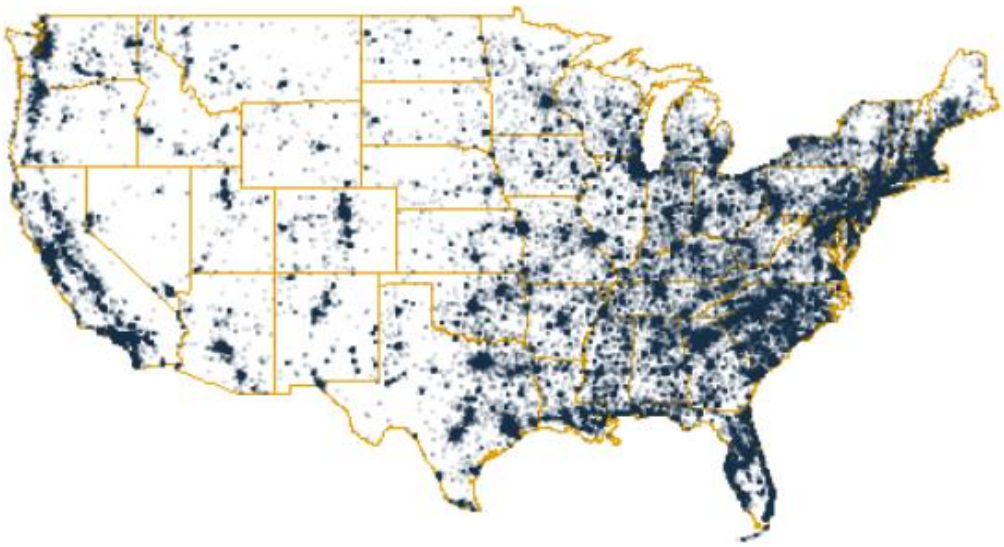


Figure 9: A map of America containing Gun Violence data plotted, illustrating where the crimes happened. The shaded clustered areas represent where there is a higher count of crime recorded for.

It appears California (16,306) has the highest count of gun violence recorded in the dataset over the rest of the states. Hawaii (289) stands for the lowest count of gun violence in the dataset. That's a big difference in the two states. However, this is not a shock, the different population rates of the two states would certainly be a factor as to why this is. A graphical representation of the data (Figure 8) is included for those with little geographical knowledge on where states are to give an idea on where the clusters of crime are. Labels of states will be included in charts below to assist in this. Figure 6, 7 and 8 were plotted using R studio's package ggplot.

```
'data.frame':   51 obs. of  2 variables:
 $ state: Factor w/ 51 levels "Alabama","Alaska",..: 5 44 10 33 39 14 36 11 34 23 ...
 $ X2017: int  39536653 28304596 20984400 19849399 12805537 12802023 11658609 10429379 10273419 9962311 ...
```

Figure 10: State Population data structure.

The State Population dataset consists on two variables, State and the year 2017. Starting of 2017 was the year examined on all 51 states in the data. Later in the analysis there is a deeper look into Alaska's and California's crime rate.

```
'data.frame':   50 obs. of  4 variables:
 $ State: Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ X2012: int  45 31 23 46 18 2 16 26 34 33 ...
 $ X2013: int  47 16 19 45 17 7 31 28 30 27 ...
 $ X2018: int  44 3 15 49 14 6 16 8 20 23 ...
```

Figure 11: Happiness Rankings data Structure.

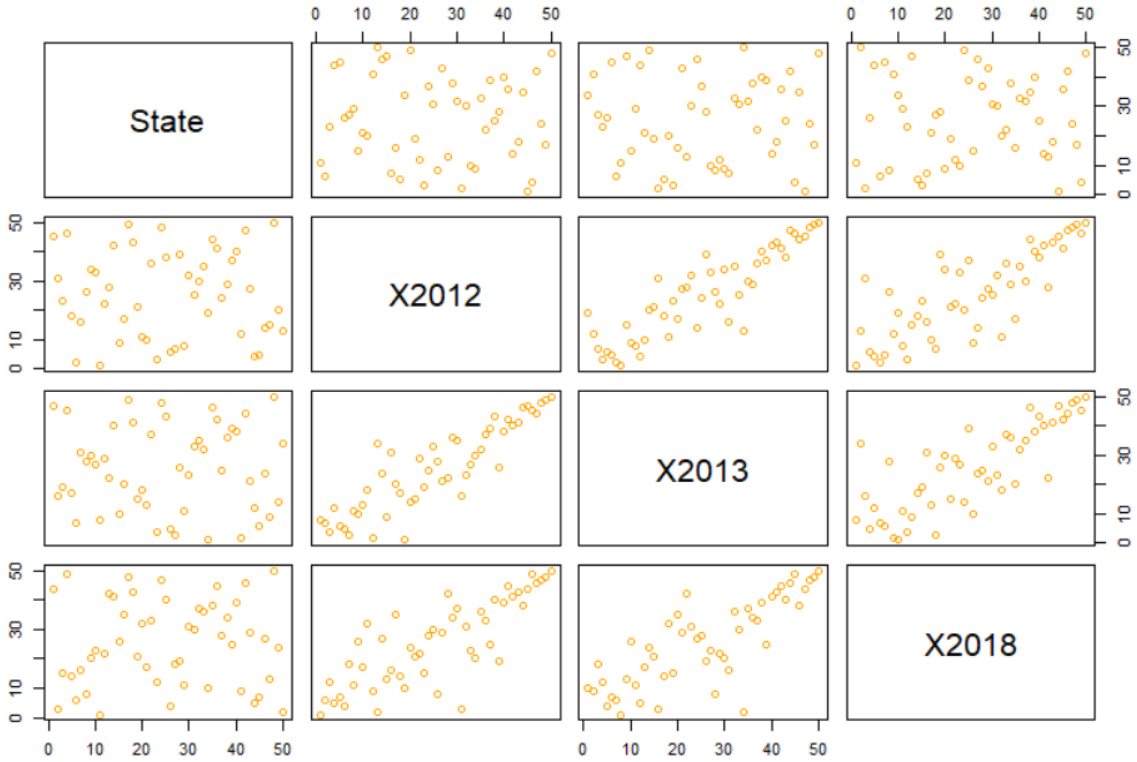


Figure 12: Correlation plot of Happiness Dataset.


Happiness Rankings 2012
2012 Ranking
50
1
15
6
19
3
21
24
22
12
20
36
30
13
7
9
29
39
4
2
28
42
44
18
17
38
49
14
35
23
25
41
46
40
48
45
33
27
43
31
34
Powered by Bing
© GeoNames, MSFT, Navteq


Happiness Rankings 2013
2013 Ranking
50
1
9
5
1
4
15
25
29
2
14
37
35
34
3
10
36
26
12
22
40
46
17
7
20
43
49
24
19
33
42
45
32
38
48
47
27
21
41
16
30
Powered by Bing
© GeoNames, MSFT, Navteq


Happiness Rankings 2018
2018 Ranking
50
1
13
4
10
12
21
28
22
9
24
33
37
2
18
26
34
19
5
42
41
38
14
6
35
40
27
48
36
15
30
45
49
39
47
44
23
29
43
3
20
Powered by Bing
© GeoNames, MSFT, Navteq

3.1.2 Functional

The functional requirements within this project were achieved by adhering to the process of the knowledge discovery in Databases (KDD) methodology. The 5 fundamental steps were followed throughout the project to ensure structure and attention to detail throughout the analysis.

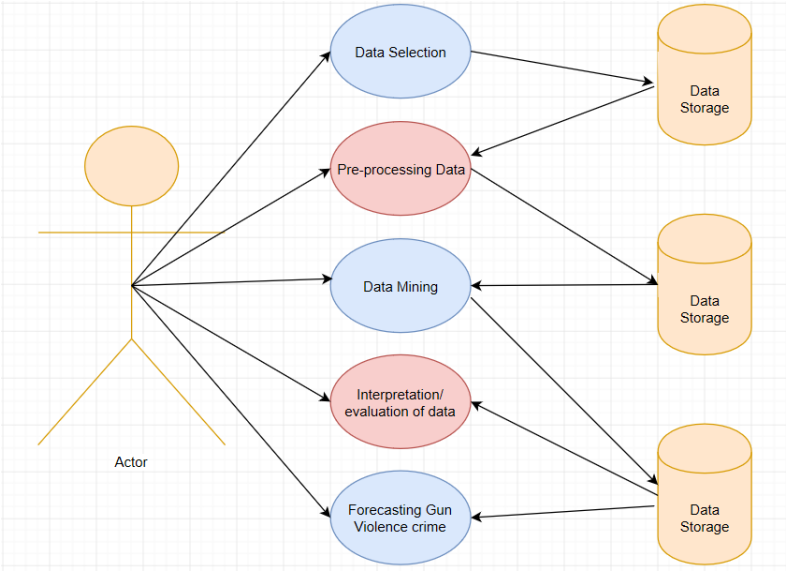


Figure 12: Use case diagram illustrating all functional requirements.

Priority 1	Priority 1 is considered vital to the completion of the project.
Priority 2	Priority 2 is required, but alternatives are viable.
Priority 3	Priority 3 is not considered vital to the analysis but would be a bonus.

Figure 13: Priority table. Measuring the level of priority, a task is.

Requirement 1: Data Selection.

Data selection is priority 1 as it is vital in the completion of the analysis and this document. Analysis on data can't be conducted without data, nor can machine learning algorithms be applied.

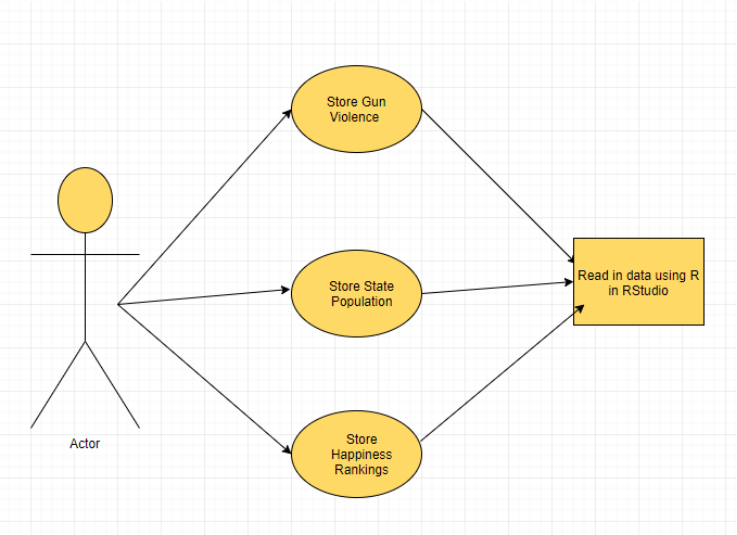


Figure 14: Use case of the storage of the three datasets. The datasets are read in to RStudio's working environment using R.

Requirement 2: Pre-processing of the data.

Pre-processing of the data is priority 2. It is an important requirement in the analysis. However, it is not vital to do it in one way.

Requirement 3: Data Mining.

Data mining is priority 1. It is this fundamental requirement that is the core of this document's objectives. Links, trends and patterns are determined using a combination of data mining tools and techniques.

Requirement 4: Interpretation/ Evaluation of results.

Evaluation is a key stage in any project. It is this requirement that you make sense of your workings and reflect on what has been done. Interpretation/ Evaluation requirement is ranked priority 1.

Requirement 5: Forecasting Gun Violence Crime.

As much as making forecasts at this time would be a great achievement in this project. It is not going to define the project. This document presents future workings. Any work that is not in this document will be pursued at a later stage to dwell on this report in aim to present a model that can forecast gun violence based on past trends.

Requirement 6: Data Visualization

Description & Priority:

Data Visualization will be at the end stages of this project. The data will be displayed on a Tableau Dashboard. Data Visualization is ranked priority 1 as it is through visuals that insights are found, and this document is based around analysing numerous visual representations of data. It was a priority 2 at first as there are many methods to present visuals. However due to a aim being that this document is user friendly and the visuals being majority of the analysis it is now ranked priority 1.

Use Case:

Scope:

This use case is a representation of the interaction the user can have with the Tableau Dashboard to visualise the results of this analysis.

Use Case Diagram:

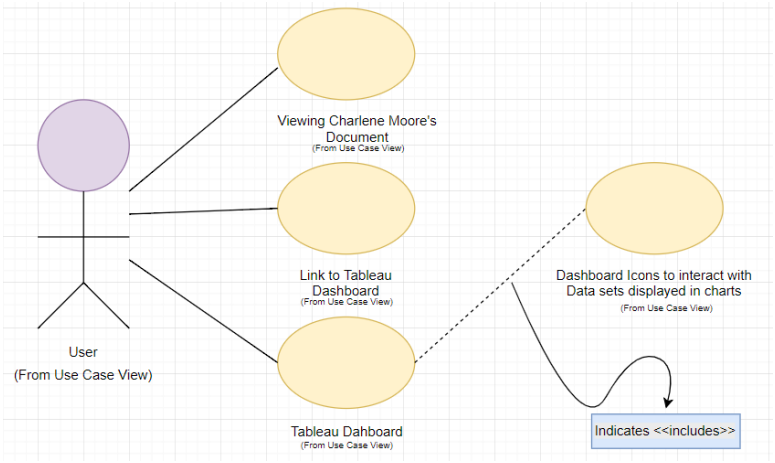


Figure 15: Tableau Dashboard Use Case

Flow Description:

Pre-condition: Analysis of the data is complete and accessibility to Tableau Dashboard is provided.

Activation: The use case starts when the link in the final report of this project is used to access the dashboard on Tableau.

Main Flow:

The user accesses the link from the final document for this project.

The user opens the dashboard via the link. Opening Tableau.

The user filters the dashboard visuals to study the results of the analysis conducted in this project.

Exceptional Flow:

The link fails to direct the user to the dashboard.

The user tries another web server to access the link.

The link works and directs the user to the dashboard.

Termination: The dashboard is exited.

Post-condition: Tableau software stores the data visualisations securely.

GUI:

Below is a mock-up of the Dashboard on Tableau created using Balsamiq Mock-up software. This provides an insight as to what it will look like. The filters will allow users to interact with the data to visualise the results of my analysis on gun crimes.

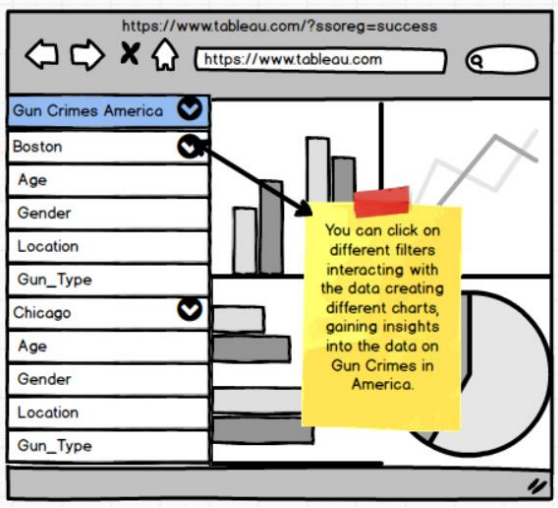


Figure 16: Balsamic mock-up of Tableau Dashboard.

Ggplot and plotly were used throughout the analysis to visualise the results. At the ned of the analysis Tableau is used to present an interactive dashboard for the users to interact with the data.

3.1.3 Non-Functional Requirements

Requirement 1: Performance

Performance of the models is considered priority 2 in this analysis. If models error rates are high, we can only evaluate and interpret as to why this is and adjust the model or try another one. Performance as an overall, in terms of performing the analysis is priority 1. It is important that the KDD methodology is followed to enhance performance throughout the analysis and ensure vital insights are not missed or excluded in this report.

Requirement 2: Security

Security is important in data analytics. With the new GDPR regulations, storage and use of data is more delicate than ever. My data is sourced from open source publicly accessible websites. The data included in this analysis does not contain delicate data that in the hands of others could be a danger. However, there is no harm ensuring all data is securely stored and managed. Dropbox and GitHub were used throughout the cycle of this project to store any information or data relating this analysis.

Requirement 3: Integrity

The datasets selected should contain all relevant information required for an accurate analysis.

Requirement 4: Compatibility

The datasets selected should be compatible with each other, containing data within the same timeline of the analysis. Data for 2000 on crime compared to population data for 1950 is not going to provide accurate results.

3.2 Design and Architecture

Figure 17 is a visual representation of the systems architecture made up of many components. The selecting and storing of the data are key components. The API component is with the ggmap () to plot the coordinates in the gun Violence dataset on a geographical map of America (see figure 18). The data is analysed using data mining tools and techniques, followed by visual representations of the results. The results are then interpreted and evaluated. The system will have the ability to be represented visually and Tableau allows the users to interact with the visuals. A method used for evaluation of models and statistics is statistical tests such as T tests like Anova and Kruskal Wallis tests. For model evaluation cross validation is often used. The system will have the ability to perform these evaluation techniques.

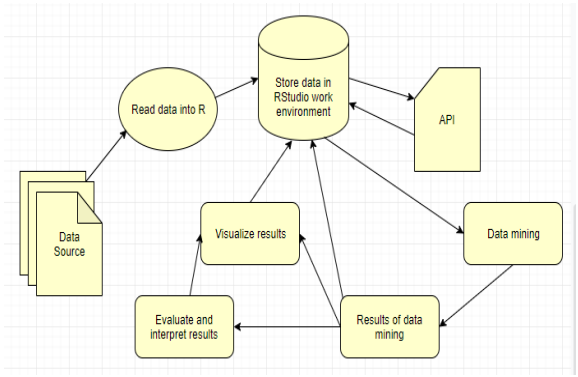


Figure 17: System Architecture that demonstrates the different components of the analysis.

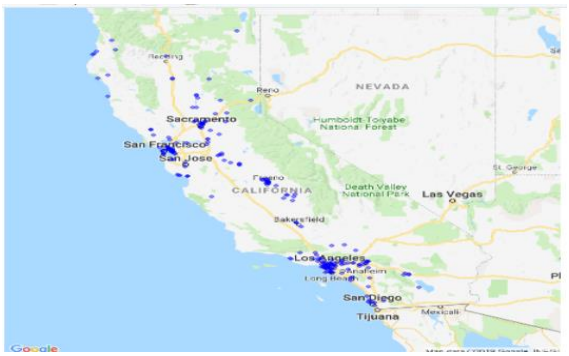


Figure 18: Ggmap plot of all the gun violence data using a Google Api.

3.3 Implementation

A detailed description will be included in this section of what was carried out throughout the project. Attached is a link to Tableau dashboard of visualisations of the analysis and an attachment of the code from RStudio which generated the graphs and charts in this document is included also.

3.3.1 Low Level System Architecture

A low-level system architecture is included to demonstrate the workflow throughout the projects lifecycle. The KDD methodology provided structure and ensured a level of completeness throughout the project. The steps included consist of data selection, ore-processing of the data, transformation, data mining and interpretation and evaluation of the results. By adhering to this methodology, it provided structure in order to meet key objectives.

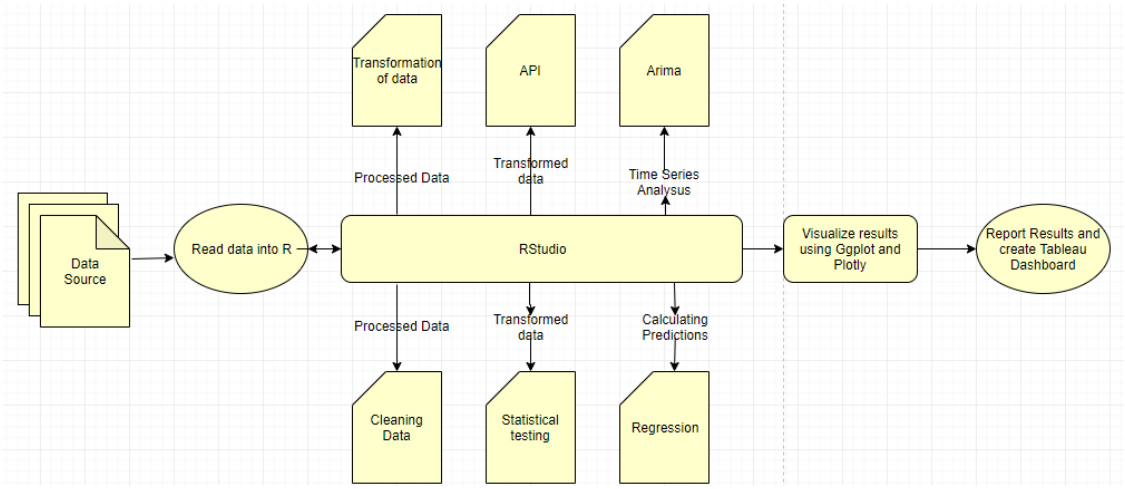


Figure 19: Low-Level System Architecture.

3.3.2 Important R functions used

- str()
- set.seed()
- glimpse()
- summary()
- data.frame()
- as_tibble()
- as.numeric()
- as.date()
- cor()
- table()
- read.csv()
- colnames()
- plot()
- Acf()
- gsub()
- map_data
- ggplot()
- dplyr()
- aes()
- geom_jitter()

3.3.3 Important R packages used

moments () is used for descriptive statistical tests such as testing for skewness and Kurtosis of data. These are important to determine if data is normal distributed.

ggplot2(), a popular package in the R community, offers a powerful graphics language for creating elegant and complex plots (Kabacoff, R. I., 2017).

as_tibble () is a new S3 generic with more effective methods for matrices and data frames (RDocumentation, 2019).

stringr package provides a comprehensive set covering almost anything you can imagine. Stringr focusses on the most important and commonly used string manipulation functions (RDocumentation, 2019).

plyr package is a very useful tool to make simple splits in data, transform that data and easily put that data back together. If there is a big problem needed to be tackled, the plyr tool is the perfect solution to break that problem down in order to tackle it and then combine it all together after.

Ggmap package allows spatial data and models be visualized on top of static maps from various online sources such as Google Maps.

Leaflet package is used for interactive maps.

Tidyr package is used to make it easier to pull apart columns that represents multiple variables.

Lubridate package for statistical computing works with date-times and time-spans. It makes working with dates easy and fun (RDocumentation, 2019).

Forecast package provides methods and tools for analysing univariate forecasts (RDocumentation, 2019).

3.3.4 Initial Analysis

Gun Violence dataset proved suitable for analysis as it contains descriptive accounts of gun violence crimes across America. The dataset required pre-processing. To do this when reading in the data and argument was set to replace empty strings with NA's to then determine where the data's potential anomalies were. The dataset consists of a date variable containing the day, month and year of the crime. This required splitting to allow analysis on each of the individually. The lubridate package gave us the tools to do this simply. The variables containing data on participants in the gun violence is complicated to interpret and transform into insights without pre-processing. To overcome this complication the gsub function was used to split each value in a single variable column to their own column.

Due to the motivation behind this topic being the Las Vegas shooting, it was disappointing to discover the dataset did not contain the tragic event due to reason explained below in figure 20 by James Ko, Gun Violence dataset creator. The Las Vegas shooting happened within the timeline of this analysis, therefor it has been added to the dataset as it is not only a big part of why this topic was chose, but also due to the analysis being accurate justifies adding it. Although with the injury and death count being so high for this incident, it may skew the results of a model for 2017. This requires monitoring.

> " I actually had to specifically remove the Las Vegas shooting along with one other incident; see here for why. Basically, it was causing problems because my program expected an HTML webpage, but that incident was the only one in the entire database that got its own special PDF. If this is a problem for you, I'd be open to a GitHub PR that manually updates the dataset with details from that shooting; then I'll make sure the Kaggle dataset gets updated as well."
>
> - (James Ko, 2018)

Figure 20: Gun Violence dataset creator James Ko response to a discussion on Kaggle as to why the tragedy was not included in the dataset.

3.3.4.1 Data Exploration

3.3.4.1.1 States crime

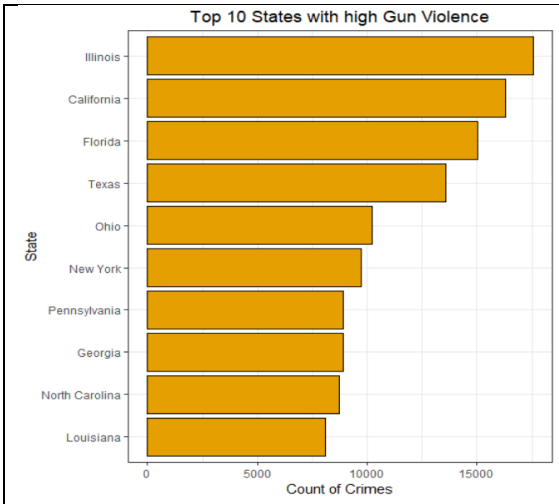


Figure 20: Top 10 states with High Gun Violence

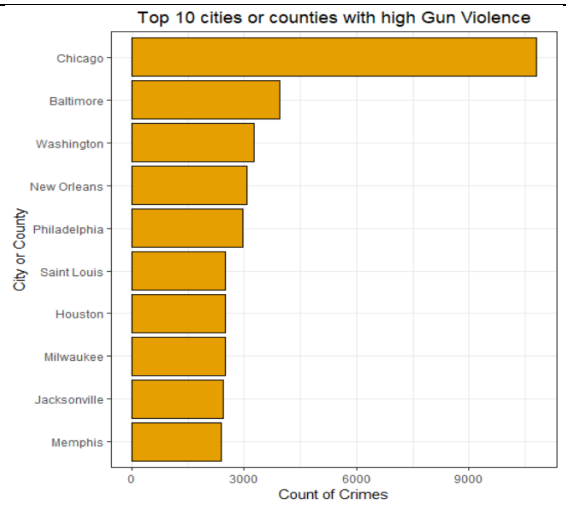


Figure 21: Top 10 cities or countries with high gun Violence

Figures 20 and 21 are bar charts plotted based on the gun violence dataset alone. According to gun Violence data set, Illinois (17556) and California have the highest gun violence rate out of the 51 states including District of Columbia. Hawaii and Vermont appear to have the lowest. The above charts are based on a count of incidents from the gun Violence data. State population surely impacts the rate of crime in a state. Thus, State Population data will be included in this analysis for a more accurate output. But first, we will continue exploring our data set to gain a deep understanding of it before introducing more data.

3.3.4.1.2 Gun Violence Year

To determine the time frame of the data the summary function was used to determine the min and max values in the date column. See figure 22.

```
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"2013-01-01" "2015-03-07" "2016-04-05" "2016-03-13" "2017-04-03" "2018-03-31"
        NA's
         "2"
```

Figure 22: Summary of gun violence date variable.

The variable date, containing the timestamp of when the crime occurred is in the format of Year/month/day and is classed as a Character. To use the lubridate package and to delve deep into the timeline of this analysis, the variable date was coerced to date datatype and reformatted to day/month/year.

The timeline of this analysis is based between 2013 and 2018, as this is the timeline gun Violence contains data on. Let's delve deeper into this timeline.

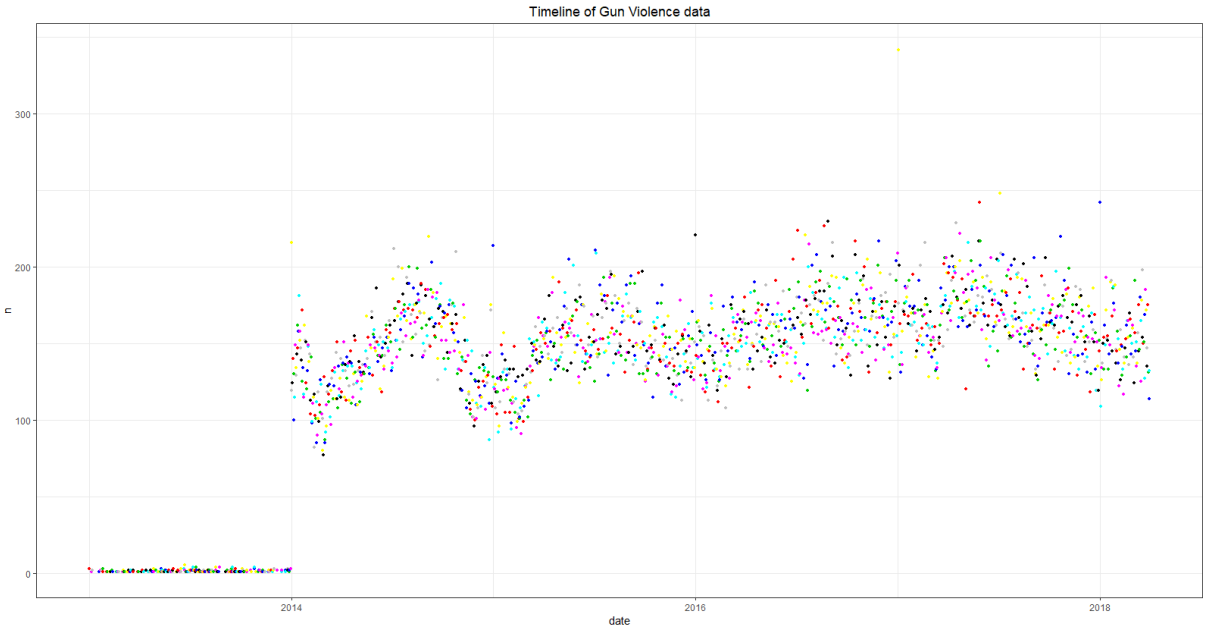


Figure 23: Gun Violence data plotted by date.

It appears from the scatter plot in figure 23 that before 2014 there seems to be very little data. 2018 also seems to be missing data. Let's look at this closer to really understand the timeline of the analysis. See figure 24 showing a boxplot of the year of the crimes, extracted using lubridate.

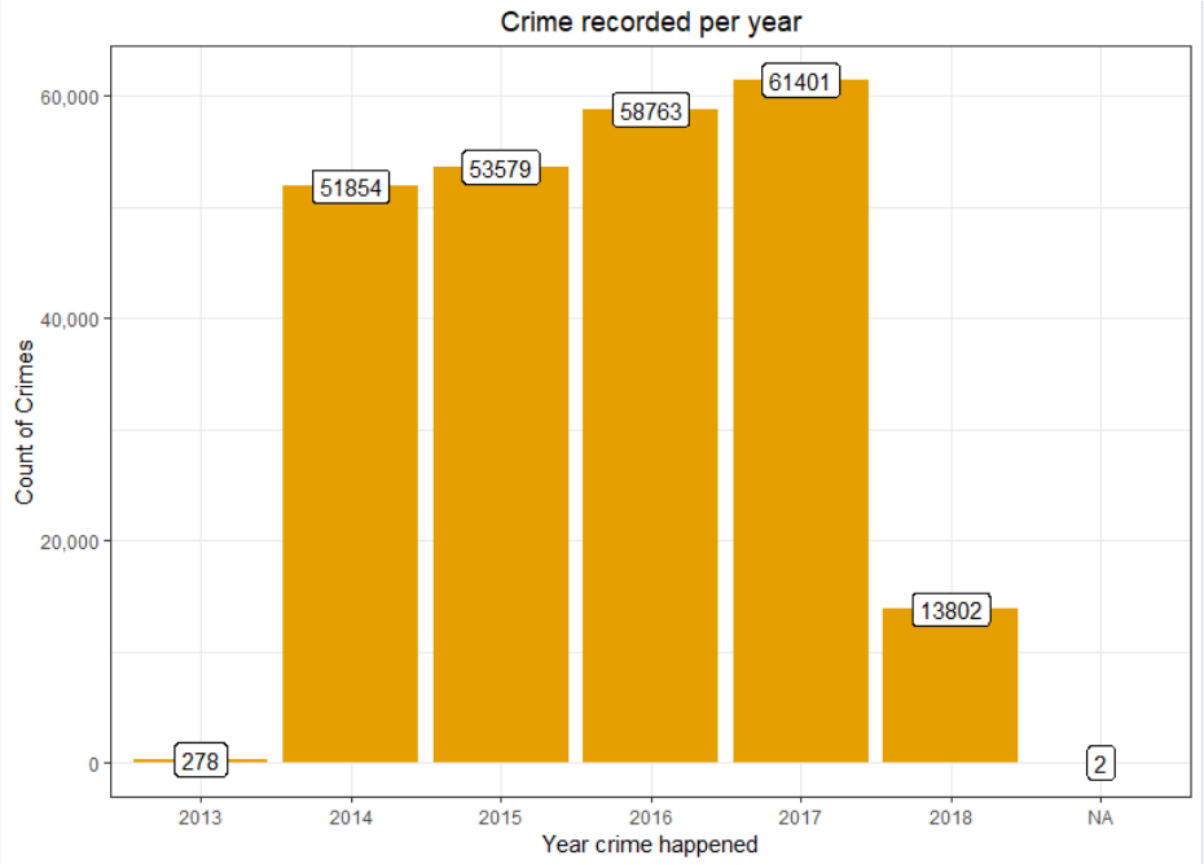


Figure 24: Gun Violence data plotted by date.

According to figure 24 There is very little gun Violence recorded for 2013. It would be naive to think that the crime rate was this low that year. But more likely that it was 2013 that the gun violence archive started gathering data on gun violence and perhaps their methods of gathering data were limited, resulting in little data gathered.

2017 has the highest count of gun violence recorded in the gun violence dataset. This may be due to the crime rate being at peak that year, or due to their methods of gathering data improving. The positive increase in gun violence from 2013 to 2017 suggests that it is possible that accessibility to data has gotten better, allowing more records of gun violence to be gathered. This justifies adding in population data and basing the analysis per capita as opposed to just the gun violence data to ensure accurate results.

2018 data seems to be too low to believe that it is the rate of gun violence for that year. This would require more attention to try understanding as to why the data count is so low for this year also.

A deeper look into the years is illustrated in figure 24. Lubridate allows you to easily divide year data into yearly quarters. This clearly tells us that only the first yearly quarter of 2018 was recorded. And with looking at the publication of the data set gun violence this makes sense and confirms that.

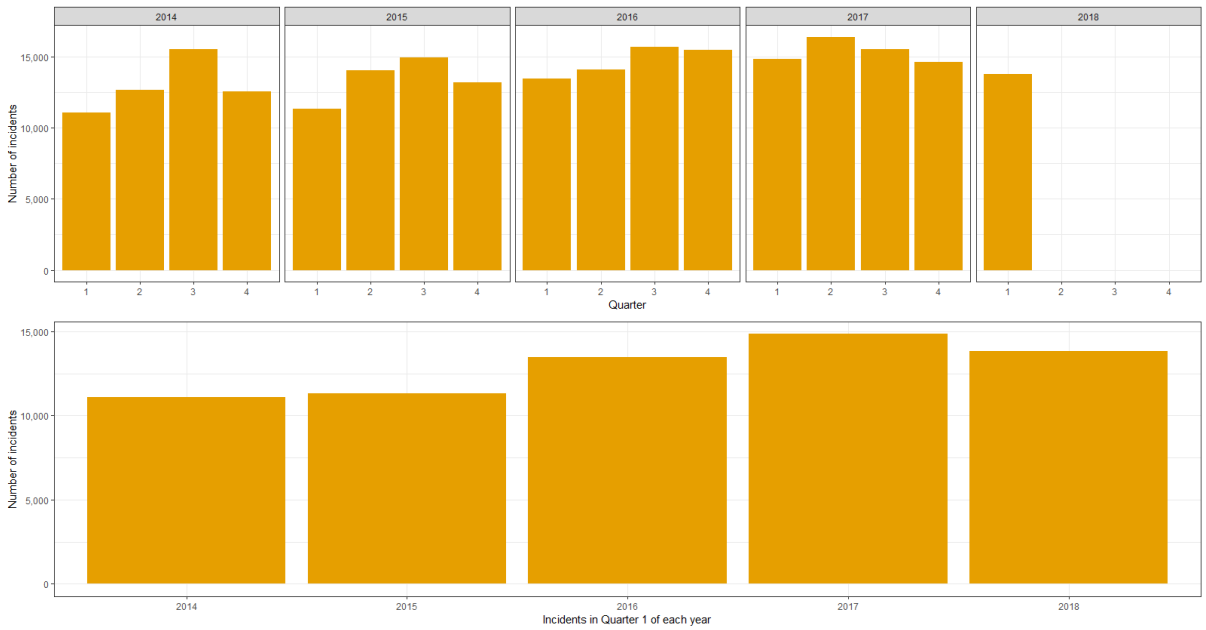


Figure 24 Timeline of the data explored by yearly quarter.

The first quarter of each year, January to March appears to decrease in January which is good as 2016 and 2017 gradually increased. It is hopeful examining the first quarter of 2018, that the remainder of the year will decrease too.

The third quarter of the year 2014 peaks. This is the months of July, August and September; holiday season. It is at a peak the following two years also. It then drops in the month of July 2017.

According to (World Tourism Organization, 2018), America's overnight visitor's (*tourists*) have increase over the past few years. A total of (69,995) tourists visiting in 2013, increasing to (75,022) in 2014, (77,774) in 2015, (76,407) in 2016 and (76,941) in 2017. With this steady increase, it could be related to the graduate increase in gun violence. However, we do not have the proof yet to state this. Therefore, this will certainly be included in future workings as it does not fit in the scope of this project unfortunately.

Independence day is a celebration that marks the holiday of Declaration of Independence of the United States on July 4, 1776 (Williams, S., Chaplain, C., 2018). This being a big celebration across the country may be contribution to the peak in the third quarter of the years 2014, 2015 and 2016. This will certainly be worth investigating. So, let's do this!!

First, we will investigate each month to see if it is even July that causes the peak.

3.3.4.1.3 Gun Violence Month

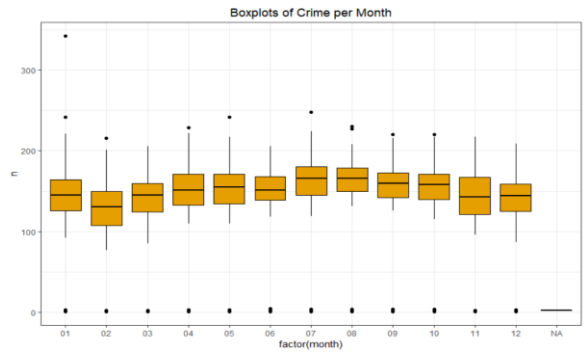


Figure 26: Boxplot of Incidents in Gun Violence dataset per month.

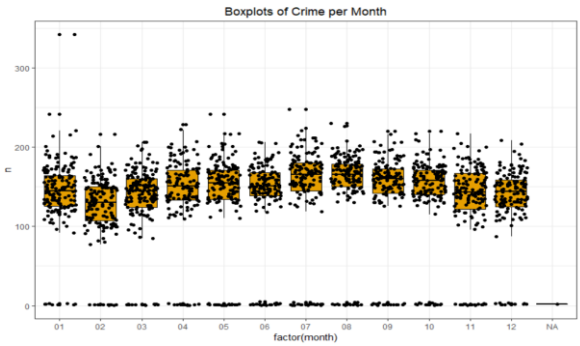


Figure 27: Boxplot displaying datapoints of Incidents in Gun Violence dataset per month.

July is month seven. It is evident by the boxplots that the month of July mean June violence over the years of this analysis (2014-2017), is higher than the remaining months.

An Anova test was conducted to determine if there was a significant difference between the month of July and the other months. The results as shown in Figure 28, determine there is a significant difference in the number of gun Violence in the month July, that the other months of the year.

Ho: There is no significant difference in the mean of the data samples on months.

HI: There is a significant difference in at least two of the means of data samples on months.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
JanSample	1	523973	523973	0.243	0.6248
FebSample	1	5084147	5084147	2.359	0.1328
MarSample	1	116940	116940	0.054	0.8170
AprSample	1	5840440	5840440	2.710	0.1079
MaySample	1	5794	5794	0.003	0.9589
JunSample	1	6323145	6323145	2.934	0.0949
AugSample	1	968953	968953	0.450	0.5066
SepSample	1	2959422	2959422	1.373	0.2485
OctSample	1	1354694	1354694	0.629	0.4328
NovSample	1	2059941	2059941	0.956	0.3344
DecSample	1	40567	40567	0.019	0.8916
Residuals	38	81885522	2154882		

Figure 27: Anova output which was conducted on samples of each month to determine if there is a significant difference between July and the rest of the months.

A Two Anova test was conducted to determine if a significant difference lies between the mean of the crimes accounted for the month of July the timeline of this analysis, in comparison to the mean of the other months. The results based on a 95% significance level, determine that there is a significant difference between July's gun Violence count, and the remainder months. February, April, June and September are relatively high p values. This suggests there is a very significant difference between those months and the month of July. May's p value is low, lower than 0.05, which suggests there is not a significant difference between May and July in the amount of gun violence recorded in the gun Violence dataset. It will be interesting to see if when the population is introduced into the analysis if this changes.

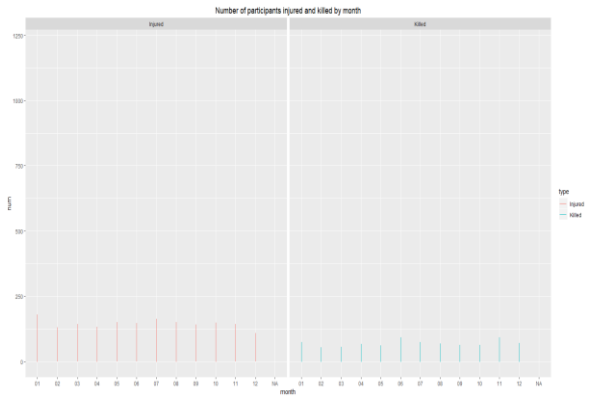


Figure 29: Graph displaying the count of injuries and count of deaths side by side for analysis.

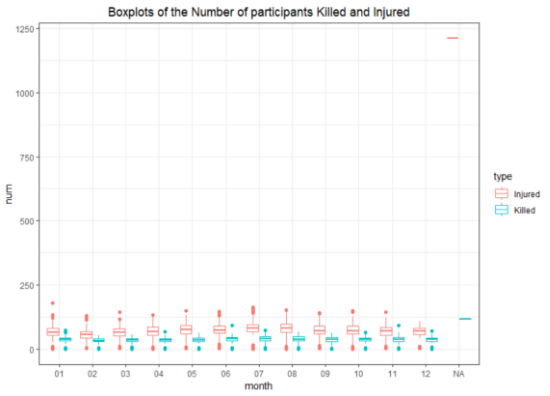


Figure 30: Boxplots displaying the count of injuries and count of deaths side by side for analysis.

Figures 29 and 30 display the counts of both injuries and deaths in the gun violence dataset, categories by the months of the year over the timeline of the analysis 2013-2018.

Both graphs aren't the best representations of this as they are hard to read due to the large scale the graph was set at. See figure 31 for a closer look at the boxplots from figure 30.

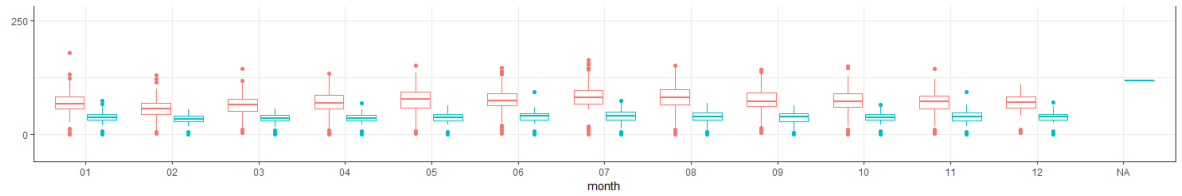


Figure 31: Zoom in on boxplot of the month's crime took place.

With a clearer view of the boxplot in figure 30, there appears to be outliers in all months. These are data points outside the whiskers of the boxplot which represent a standard deviation from the mean of data points. The potential outliers may represent mass shootings. i.e. The Las Vegas shooting in 2017.

It is evident in the boxplots that there is a higher average of injuries than deaths in the dataset. This is expected. The red boxplots represent injuries and the blue, deaths.

Ggplot is a fantastic package for plotting graphs and interpreting data. However instead of cropping graphs it would be more beneficial for this analysis if the graphs were interactive allowing up to zoom in. Plotly will be used in conjunction with ggplot to achieve this. See figure 32.

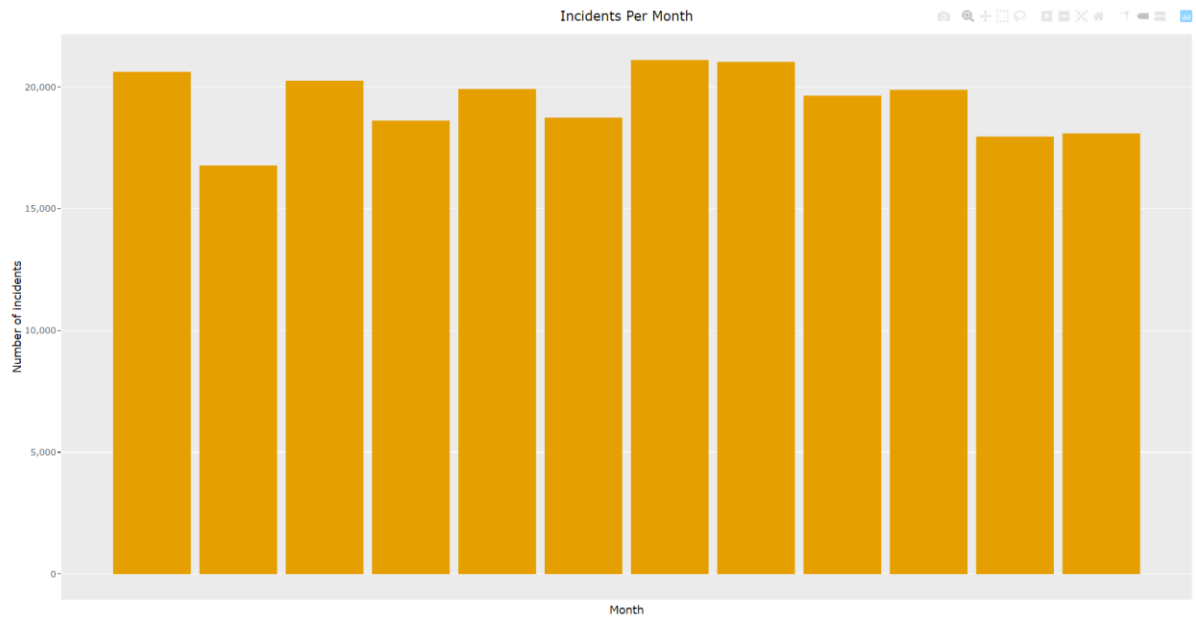


Figure 32: Boxplot created using plotly of our gun violence data per month.

January, July and August are highest in the bar plot above (figure 32). July is the highest at a total of 21,109 records of gun violence recorded. Followed by August (21,026) and January (20,620). February has the lowest count, at (16,773). We know that Independence day is in July, but what causes the increase in the other months? Let's delve deeper and look at the dates a bit closer. See figure 33.

date	Total number of incidents
1 1	1115
7 4	876
7 5	820
7 30	788
10 25	742
7 17	740
7 19	740
8 13	734
7 25	730
8 1	730

Figure 33: Focus in on the days of the years that have the highest count of gun violence crime.

What do we know, the 4th of July is the second highest day of the year for gun violence according to our data? Above it, the first of January, New Year's day. This makes sense. The early morning after the count down into the new year, when alcohol consumption is involved, it's not surprising. The day after Independence Day follows, with a total of (820) gun violence's recorded in the gun violence dataset. Just like New Years, this would be the early hours of the morning after the fireworks and celebrations for Independence Day, a high level of consumption of alcohol across the country no doubt.

According to (Branas, C., n.d), in 2008 a total of 46 laws in 31 states restricted the interaction of alcohol and firearms. Stating that over one-third of firearm injuries descendants had acutely consumed alcohol prior to their death makes this not come as a surprise that these laws were enacted. Most states prohibit firearm holders to even access a place where alcohol is served or consumed in attempt to prevent gun violence.

There is no specific cause of the high gun violence count for the 30th of July that can be identified at this moment. The 25th of October too does not have a significant holiday that could explain the high crime rate. It is clothes to Hallowe'en, but not such. Perhaps it's due to breaks in schools or colleges and again, high alcohol consumption on breaks like this. This is only speculation, which gives reason to investigate more as future work.

3.3.4.1.4 Gun Violence Day of the week

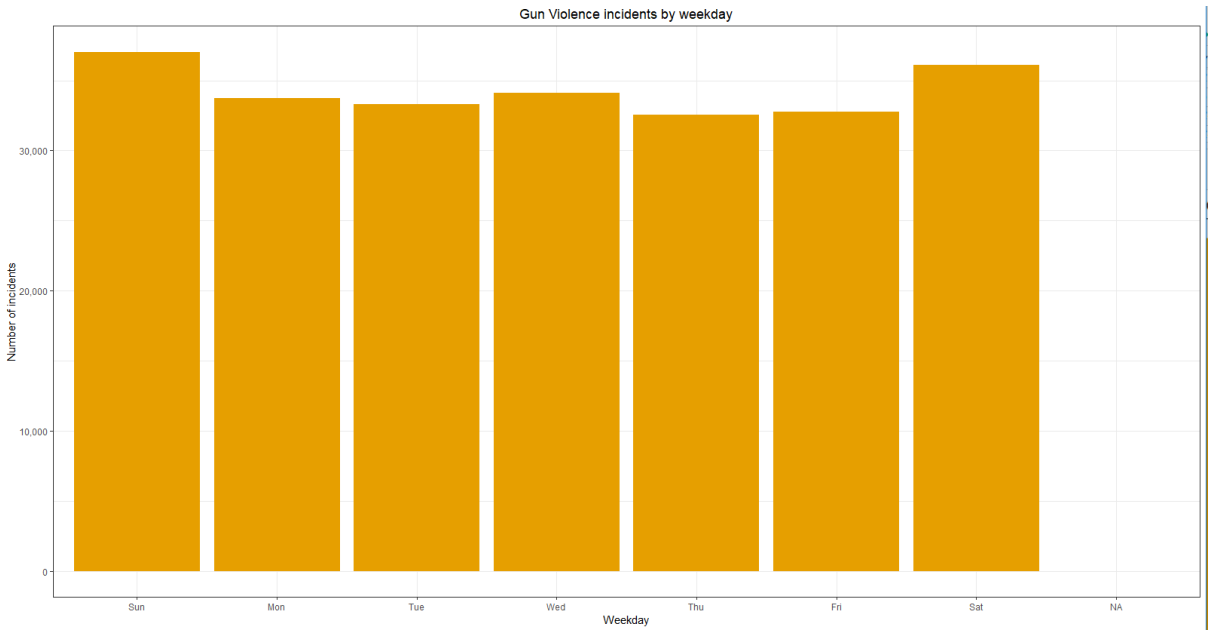


Figure 34: Gun Violence recorded by weekday.

A lot of comparisons have been made to alcohol and gun violence already. But it is certainly a massive factor in the rates of gun violence crime. Figure 34, a bar plot of gun violence by day of the week again gives reason to think that alcohol has a big factor in gun violence rates. Sunday has the highest count of incidents recorded, followed by Saturday. Both days that have high consumption of alcohol rates. Sunday being the highest peak is understandable due to Saturday being the most common day for people to consume alcohol and the early hours of Sunday morning is usually when parties occur, and things end up getting out of hand. This again is only speculating and alcohol consumption and its relationship with gun violence is an entire project worth investigating. Wednesday's peak, higher than Friday is surprising.

3.3.4.1.5 State Population

As mentioned before, Illinois (17,556) and California (16,306) have the highest count of incidences in the gun violence dataset. Hawaii ranks last, with as little as (289) gun violence's recorded in our dataset. It would be very naïve for us to think that they have the highest and lowest crime rate as population rate compared to smaller and bigger states is massive, meaning there is far more people included in the statistics of it. Therefore, including the states populations in this analysis was certainly justified.

State populations are tracked throughout the year. Therefore, when seeking data on states population it wasn't too straight forward. Analysis like this have used many different datasets on states population. Some for the same timeline of this. After research, I decided to base this analysis on the US Census Bureau data who are a principal agency in the U.S. Federal Statistical System who are responsible for providing data about America's economy and people.

First, we will look at 2017, as it is the year that we have the highest count of gun violence data recorded for.

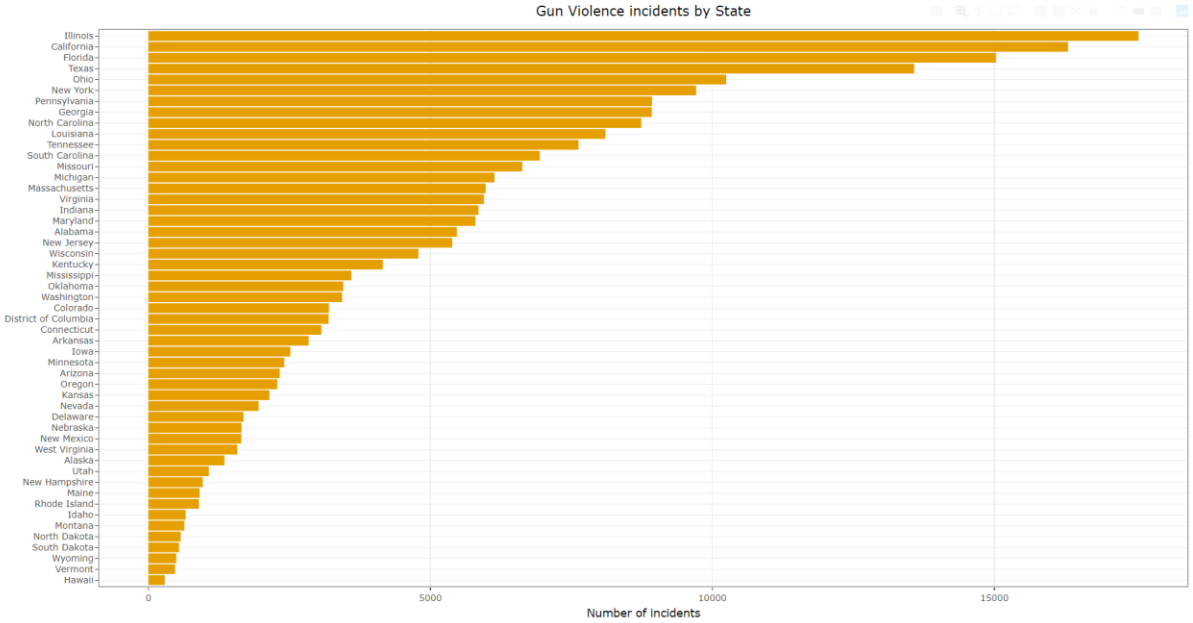


Figure 35: Bar plot illustrating the highest to lowest count of gun violence's in our gun violence dataset.

Figure 35 shows in descending order which states have the highest and lowest count of gun violence in our dataset. It will be interesting to see if there is a significant change when the population dataset is included. Let's investigate.

state	stateIncidents	X2017	Per100000
Alabama	5471	4874747	112
Alaska	1349	739795	182
Arizona	2328	7016270	33
Arkansas	2842	3004279	95
California	16306	39536653	41
Colorado	3201	5607154	57
Connecticut	3067	3588184	85
Delaware	1685	961939	175
District of Columbia	3195	693972	460
Florida	15029	20984400	72

Figure 36: Top 10 states the highest gun violence rate per 100,000.

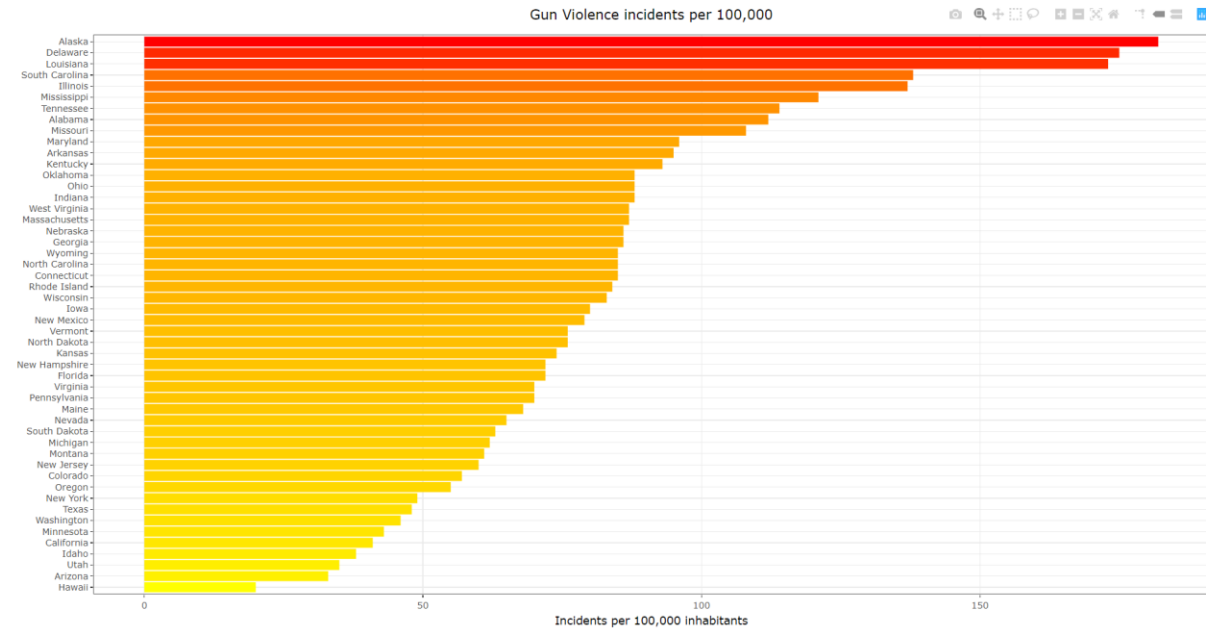


Figure 37: States ranked in order of danger with the use of statistical semantic representations. Red representing danger, yellow safe.

state	stateIncidents	X2017	Per100000
<fct>	<int>	<int>	<dbl>
Alaska	1349	739795	182
California	16306	39536653	41
Florida	15029	20984400	72
Illinois	17556	12802023	137

Figure 38: Closer look at states gun violence rate by state.

A proportion test was conducted to determine if there is a significant difference in the size difference in crime rate between Alaska and the remainder states.

p

```
        2-sample test for equality of proportions with continuity correction

data:  c(incidents_alaska, incidents_restUS) out of c(n_alaska, n_restUS)
X-squared = 1191.4, df = 1, p-value < 0.00000000000000022
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0009921687 0.0011880484
sample estimates:
      prop 1       prop 2
0.0018234781 0.0007333696
```

Figure 37: Output for 2-sample test for equality of proportions with continuity correction.

The pvalue is very small, therefore HO is rejected, accept Alternative Hypothesis that the Test is two sided.

We can conclude that the incidents rate in Alaska is significantly higher than the rest of the US.

	state	sumVic	sumInj	sumDeath	PercDeath	sumIncidents	vicPerInc
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>
1	Alabama	4878	2998	1880	0.39	5471	0.89
2	Alaska	592	325	267	0.45	1349	0.44
3	Arizona	2190	1096	1094	0.5	2328	0.94
4	Arkansas	2120	1347	773	0.36	2842	0.75
5	California	13206	7644	5562	0.42	16306	0.81
6	Colorado	1929	1133	796	0.41	3201	0.6

Figure 38: Closer look at the stats for sum of the incidents, the amount of victims and how many deaths

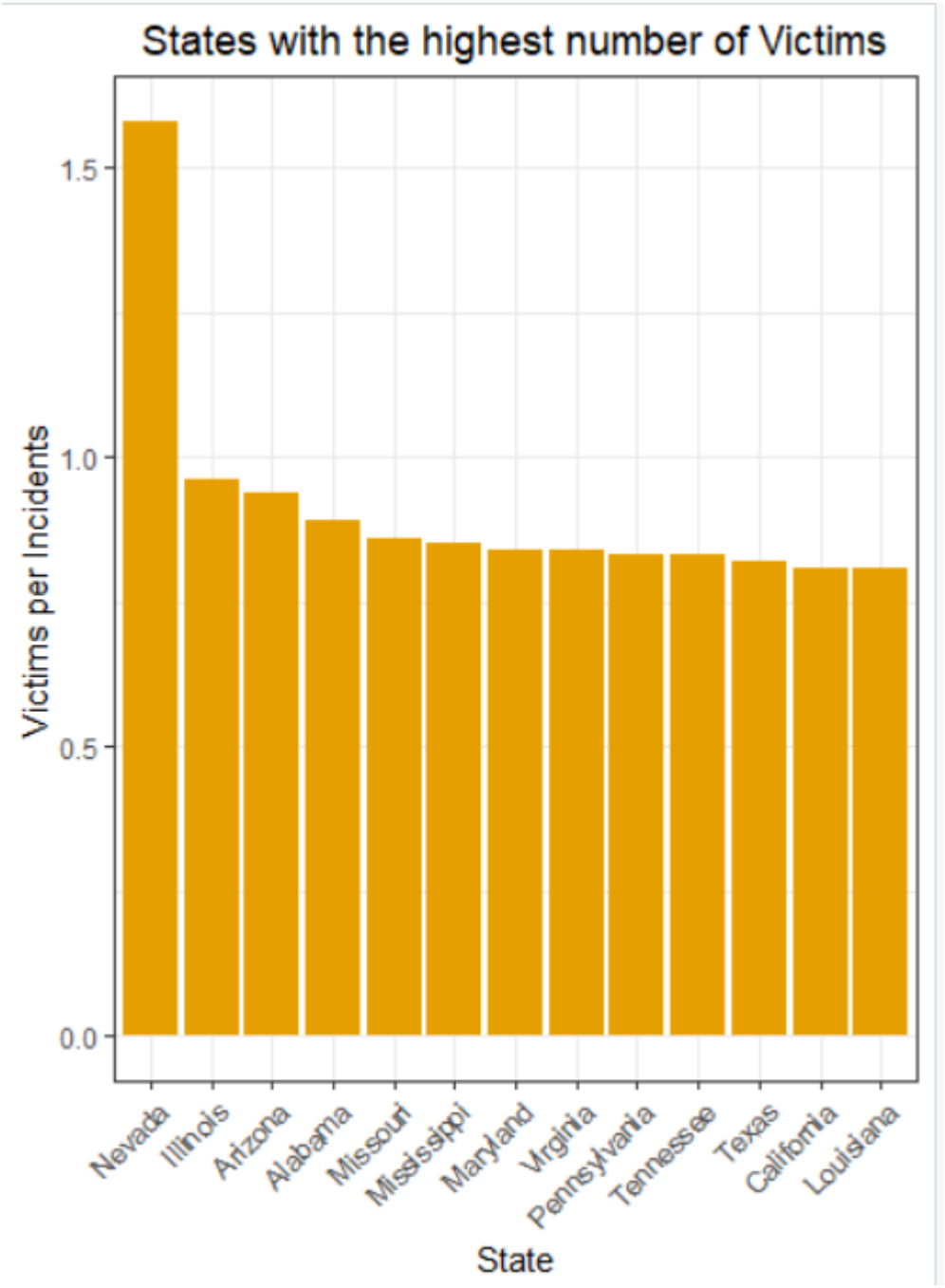


Figure 40: States with the highest count of victims.

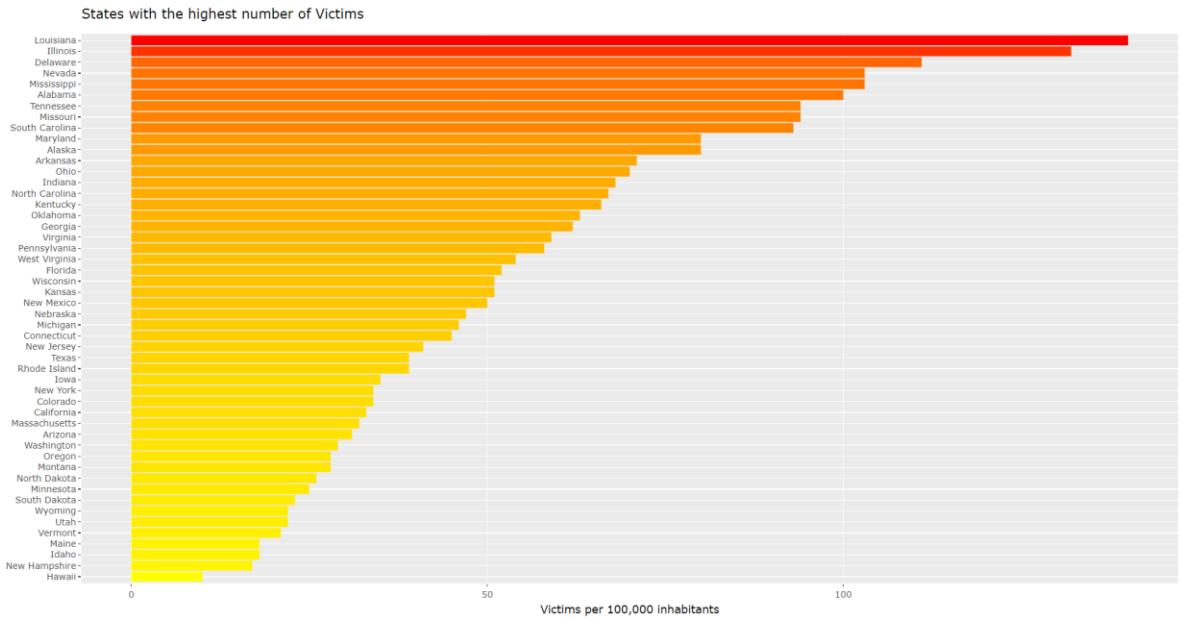


Figure 41: States with the highest count of victims plotted. Red represents danger, gradient to yellow which is considered safest.

	city	state	cityIncidents
	<fct>	<fct>	<int>
1	Bronx	New York	845
2	Brooklyn	New York	1405
3	New York (Manhattan)	New York	360
4	Queens	New York	130
5	Staten Island	New York	411

Figure 42: Incidents per city.

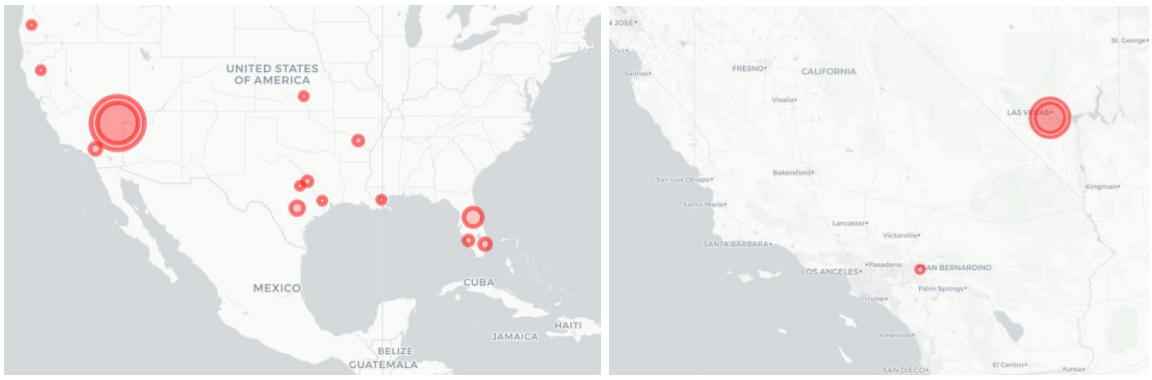


Figure 43: Map plotted using leaflet. Bigger circles represent highest crime. Graph to the right is a zoom in on Las Vegas. If you hover over it in the link of code provided, it shows the coordinates of the mass shooting in the Mandalay Bay.

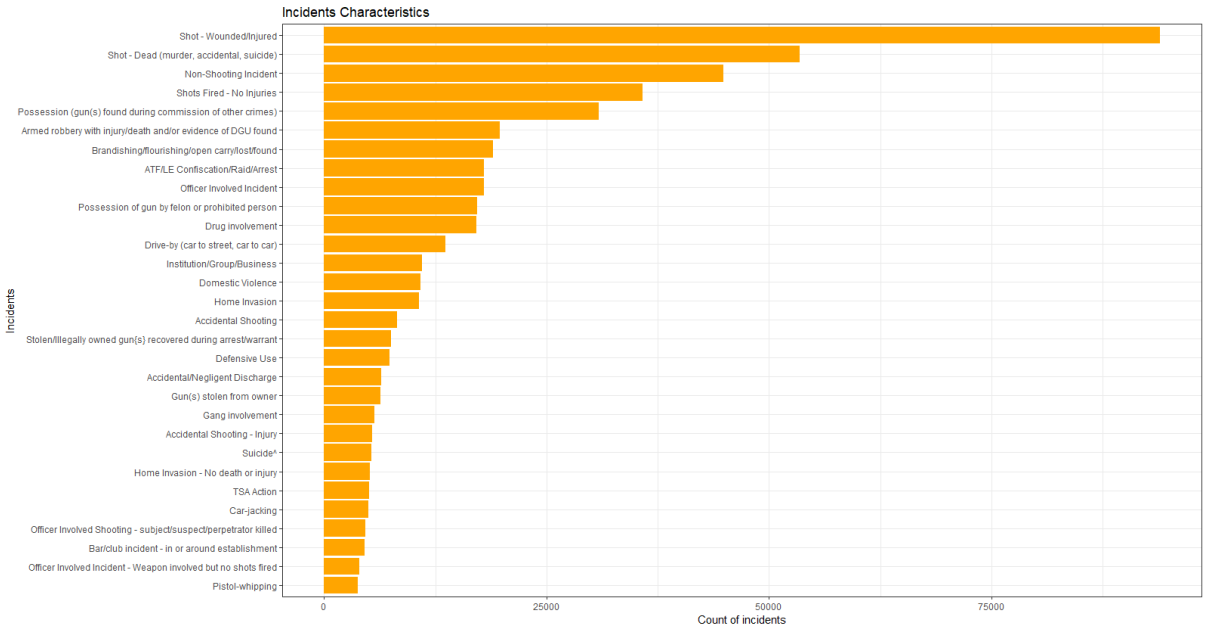


Figure 44: Characteristics of the incident

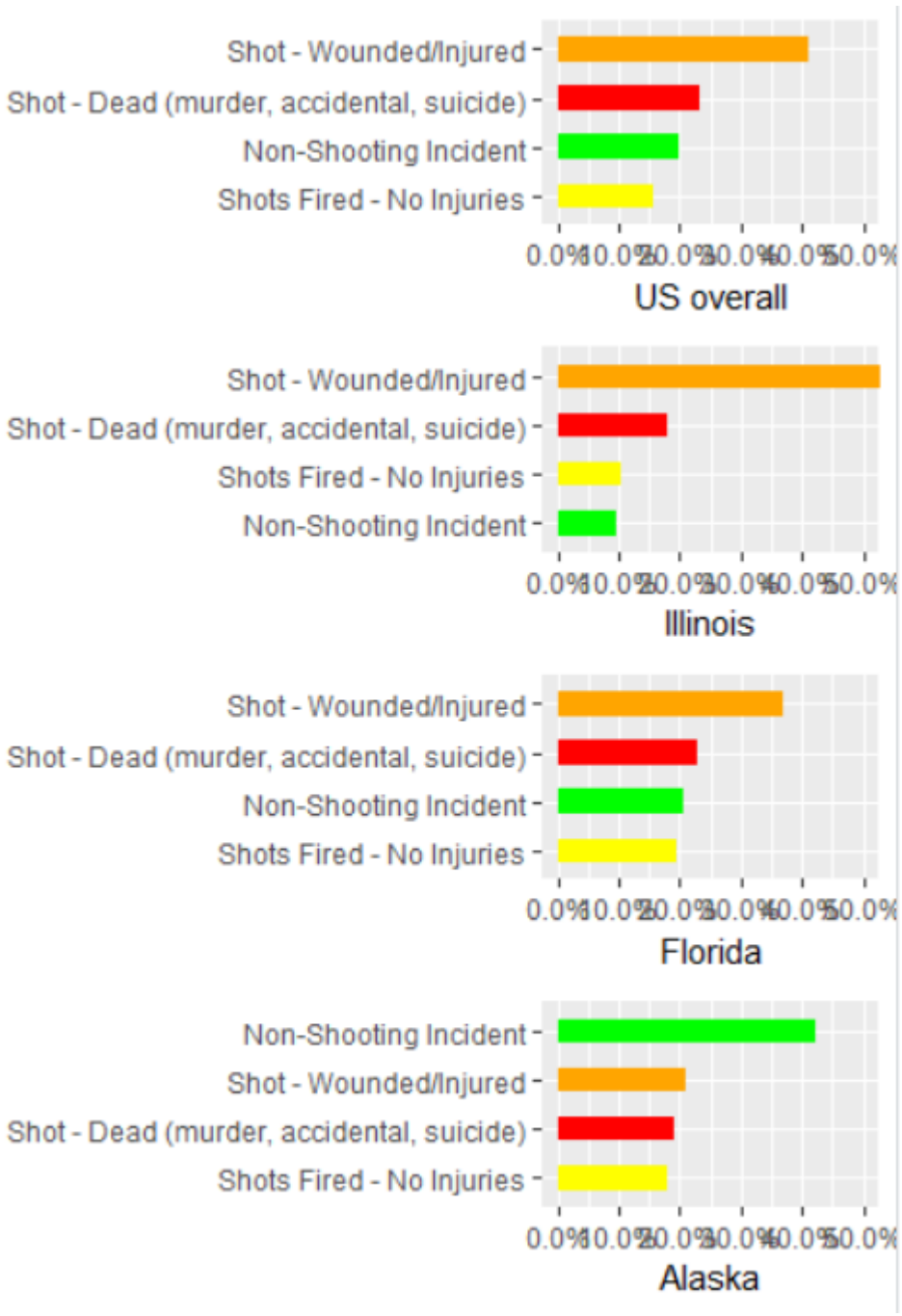


Figure 45: Insight into Florida, Alaska and Illinois compared to the US overall in regard to incident where people were shot dead, where no injuries occurred and where participants were injured or wounded.

:

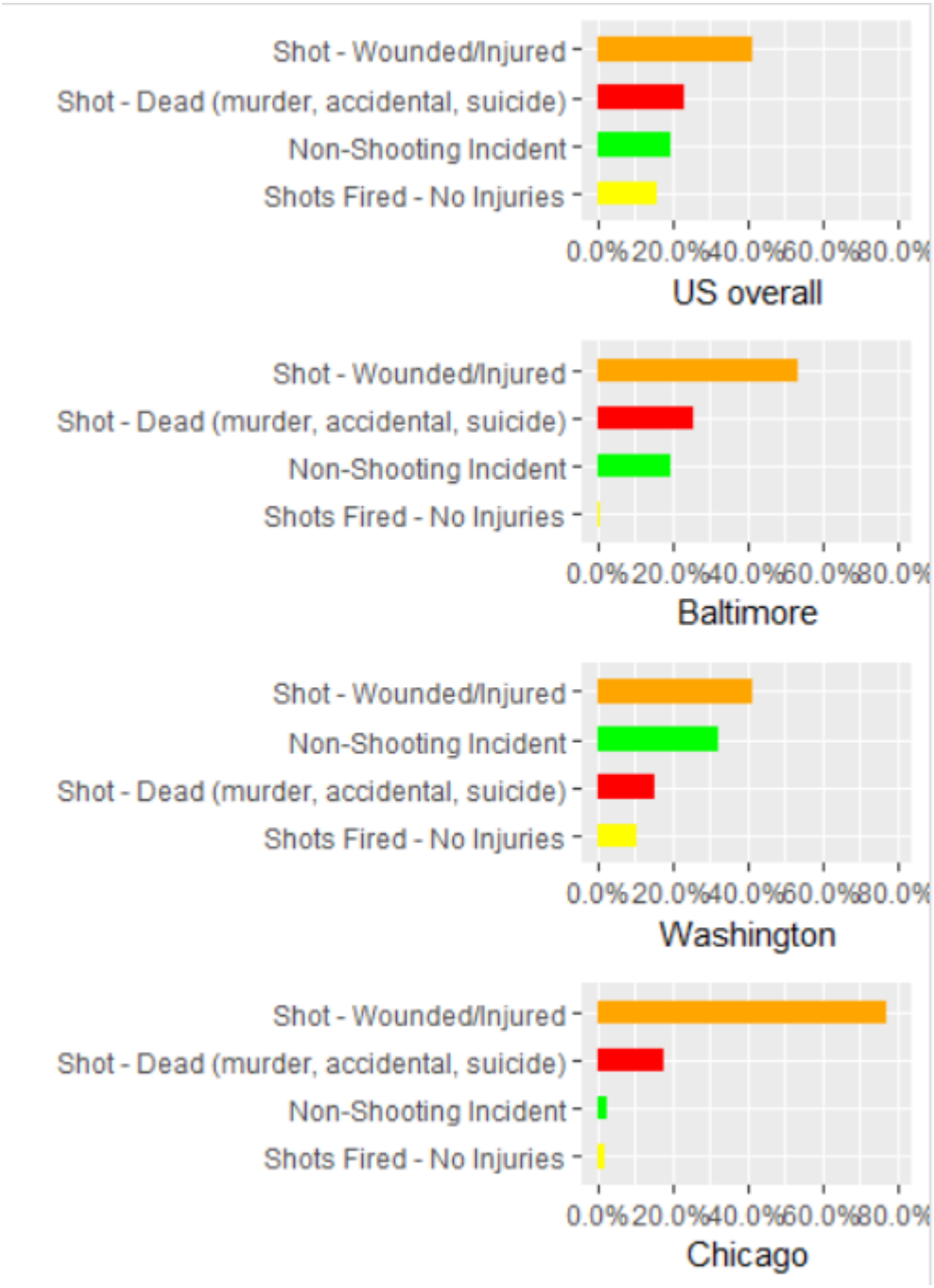


Figure 46: Incidents by city.

	incident_characteristics	n
	<fct>	<int>
1	Accidental Shooting - Death	1729
2	Attempted Murder/Suicide (one variable unsuccessful)	469
3	Mass Murder (4+ deceased victims excluding the subject/suspect/perpetrator , one location)	98
4	Murder/Suicide	2480
5	Shot - Dead (murder, accidental, suicide)	53409
6	Suicide^	5344

Figure 47: Incidents Characteristics where people died.

Figure 50: Gang involvement in gun incidents illustrated in California.

Figure 51: Terrorism in America.

The bigger circles represent the mass shooting that took place in Florida and Las Vegas.

3.3.4.1.6 State Happiness

State	X2012.Ranking	X2013.Ranking	X2018.Ranking
Hawaii	1	8	1
Colorado	2	7	6
Minnesota	3	4	12
Utah	4	12	5
Vermont	5	6	7
Montana	6	5	4
Nebraska	7	3	18
New Hampshire	8	11	11
Iowa	9	10	26
Massachusetts	10	13	17

Figure 52: Hawaii ranks umber one for happiness, followed by Colorado. Hawaii's crime rate is also the lowest. Would this suggest that there is a relationship between the two?

	State	X2012.Ranking	X2013.Ranking	X2018.Ranking
41	Oklahoma	41	42	45
42	Indiana	42	40	41
43	Louisiana	43	41	43
44	Ohio	44	46	38
45	Alabama	45	47	44
46	Arkansas	46	45	49
47	Tennessee	47	44	46
48	Mississippi	48	48	47
49	Kentucky	49	49	48
50	West Virginia	50	50	50

Figure 53: Bottom 10 states. West Virginia ranks 50, suggesting that the well being of the state isn't as well as the others. Would this be related to gun violence?

3.3.4.2 Time Series Analysis

3.3.4.2.1 Arima

Arima is an Auto-Regressive (use of historical data) Integrated Moving Average. A model that forecasts future predictions based on historical data trends.

Arima uses lags, these are current and historical variables, at a value of -1. Lag 1 is -1, lag 2 is -2 and so on. This continues to lag(N). Arima has three parameters which are d for the amount the integrated moving average differentiates. A simple arima assumes non-seasonality and that the data is stationary. This must be tested before an Arima model could be fitted.

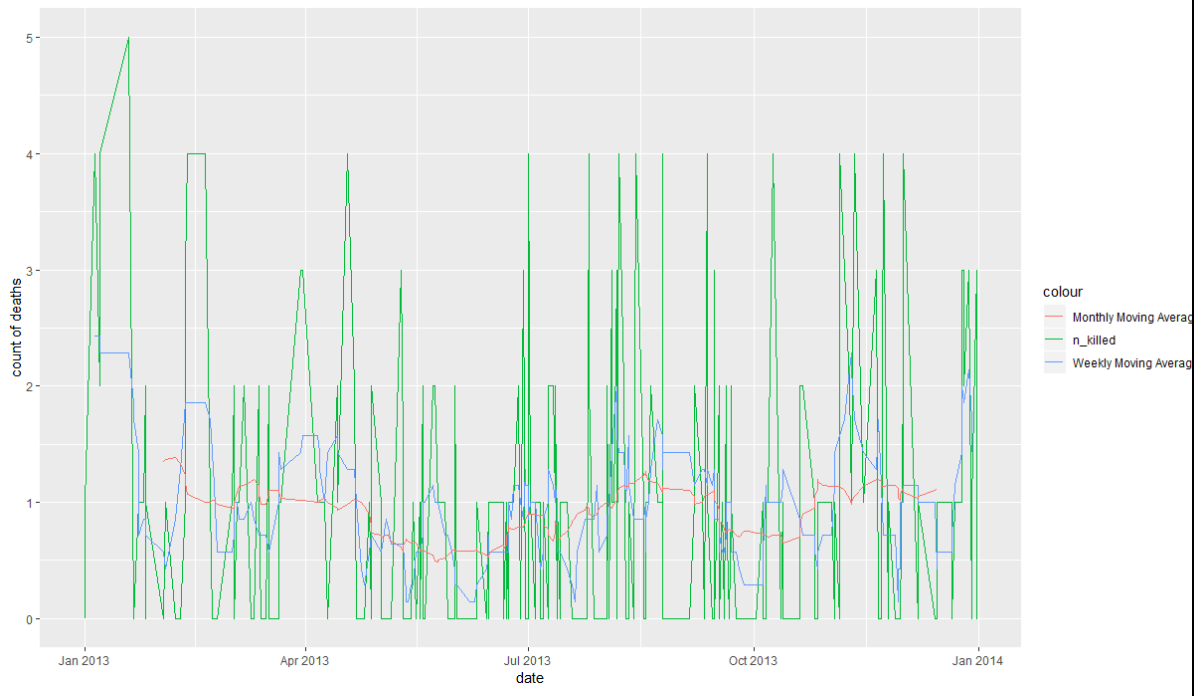


Figure 54: Weighted Moving Average of 2013.

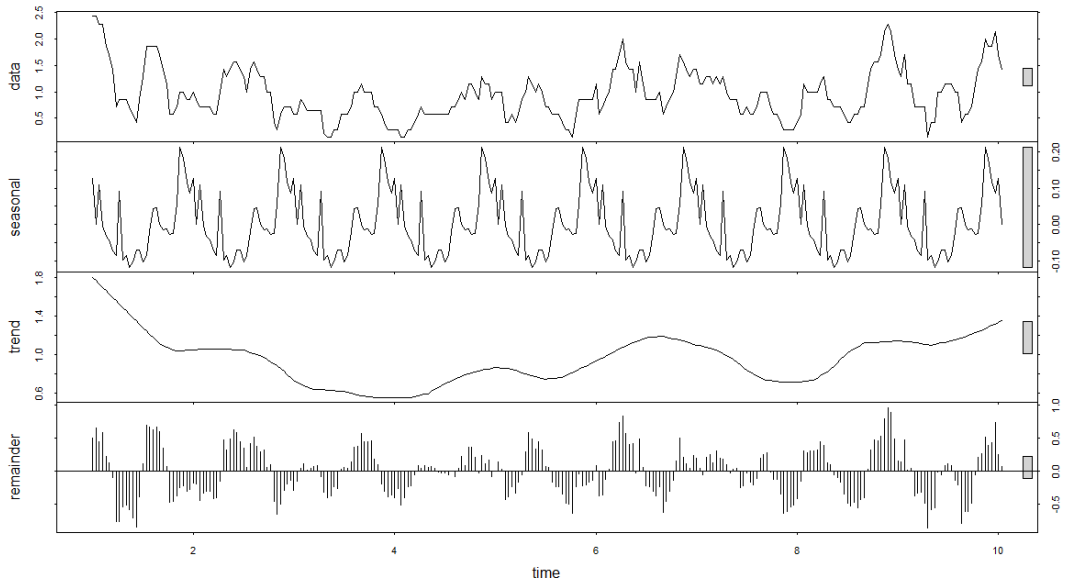


Figure 55: Decomposed Time Series

```
        Augmented Dickey-Fuller Test

data:  nkilledd_ma
Dickey-Fuller = -6.3476, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Figure 56: Results of AugmentedDickey-Fuller Test

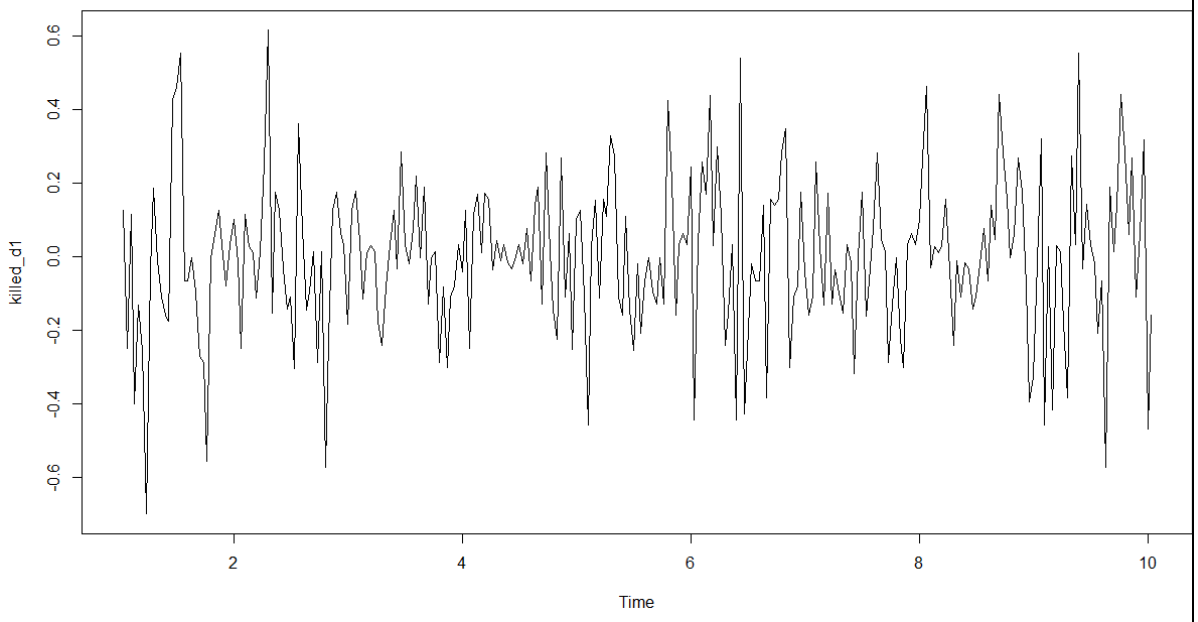


Figure 57: Differentiated over time

```
        Augmented Dickey-Fuller Test

data:  killed_d1
Dickey-Fuller = -10.718, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Figure 58: Results of Augmented Dickey-Fuller Test.

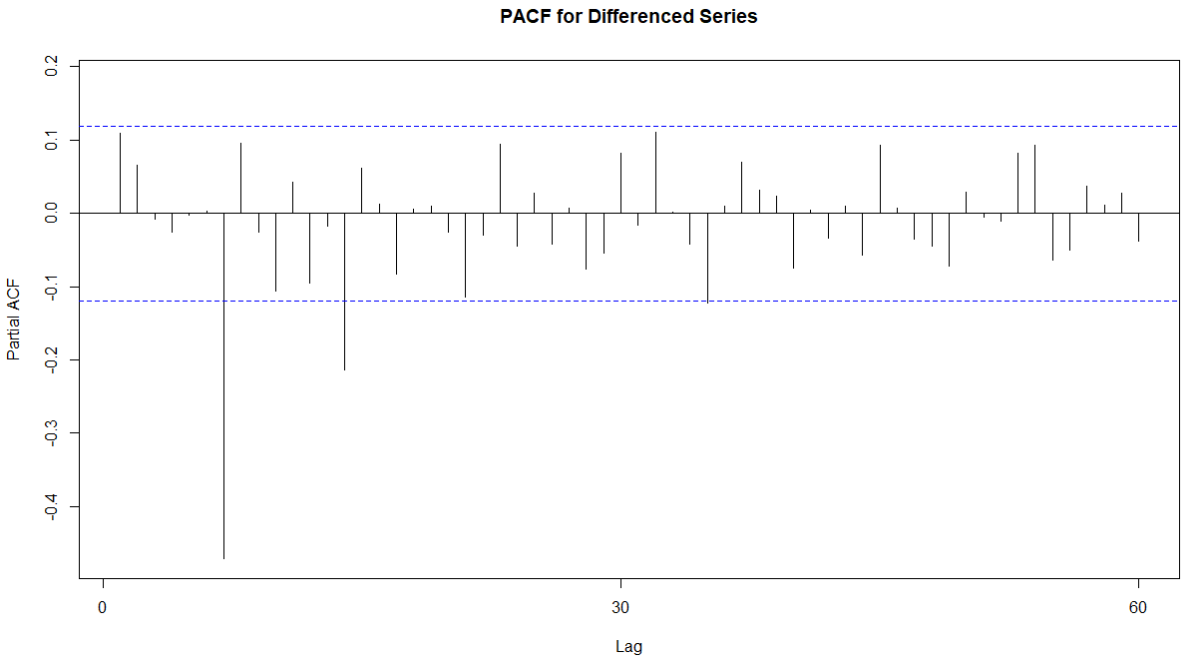


Figure 59: PACF for Difference Series Plot

```
Series: seasonal_nkilled
ARIMA(2,0,1) with non-zero mean

Coefficients:
          ar1      ar2      ma1    mean
      1.6827  -0.7450  -0.6369  0.9555
s.e.  0.0905   0.0811   0.1047  0.0704

sigma^2 estimated as 0.03994:  log likelihood=53.11
AIC=-96.21   AICc=-95.99   BIC=-78.18
```

Figure 60: Arima output.

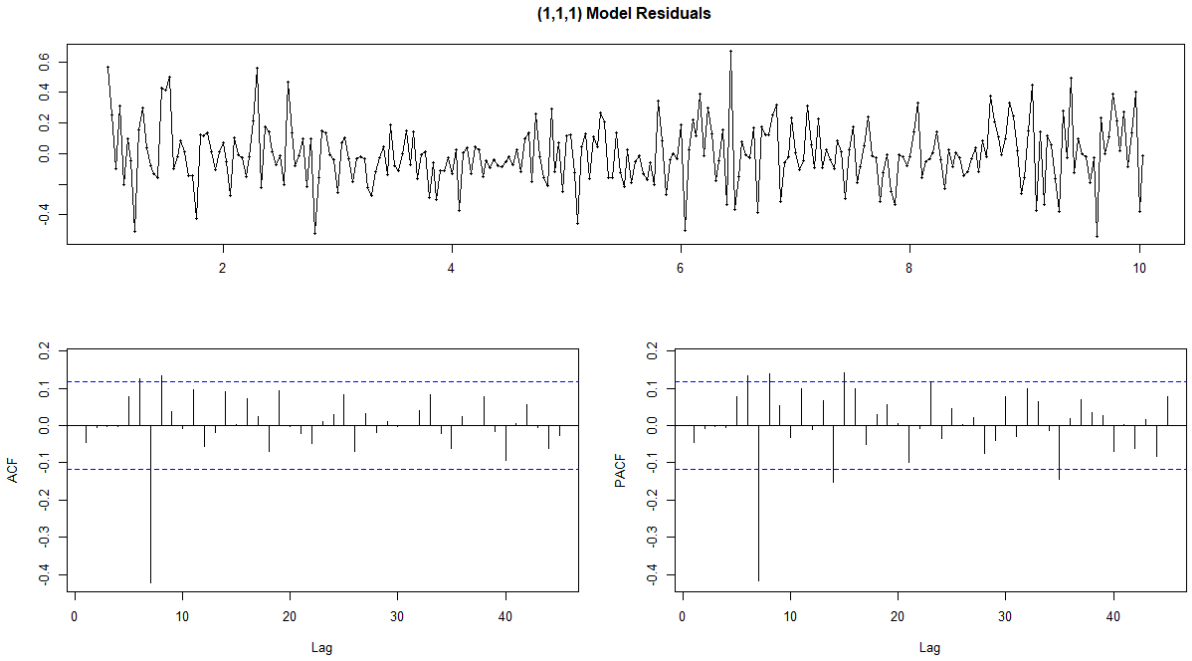


Figure 61: Model Residuals for Arima.

```
Call:
arima(x = seasonal_nkilled, order = c(1, 1, 7))

Coefficients:
          ar1      ma1     ma2      ma3     ma4      ma5      ma6      ma7
      0.2279  -0.0607  0.0541  -0.0446  0.0287  -0.0096  -0.0029  -0.9649
s.e.  0.0621   0.0355  0.0352   0.0401  0.0403   0.0369   0.0361   0.0383

sigma^2 estimated as 0.0235:  log likelihood = 112.94,  aic = -207.89
```

Figure 62: Call output

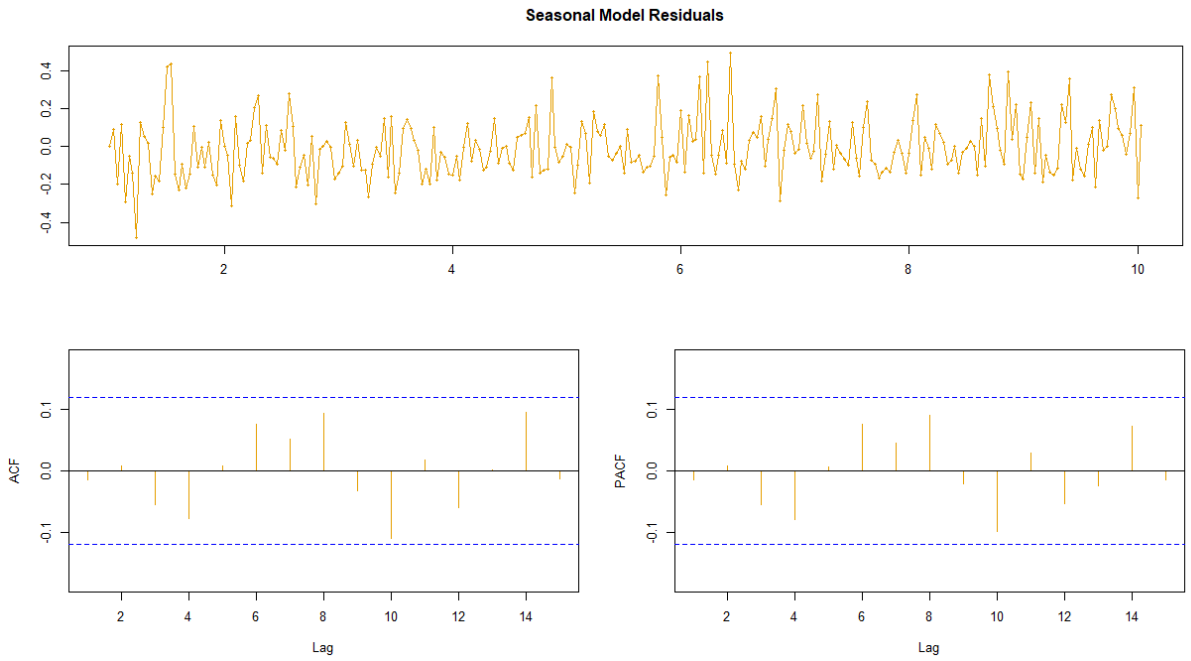


Figure 63: Lags

```
Call:
arima(x = seasonal_nkilled, order = c(1, 1, 7))

Coefficients:
          ar1      ma1     ma2      ma3     ma4      ma5      ma6      ma7
      0.2279  -0.0607  0.0541  -0.0446  0.0287  -0.0096  -0.0029  -0.9649
s.e.  0.0621   0.0355  0.0352   0.0401  0.0403   0.0369   0.0361   0.0383

sigma^2 estimated as 0.0235:  log likelihood = 112.94,  aic = -207.89
```

Figure 64: Arima

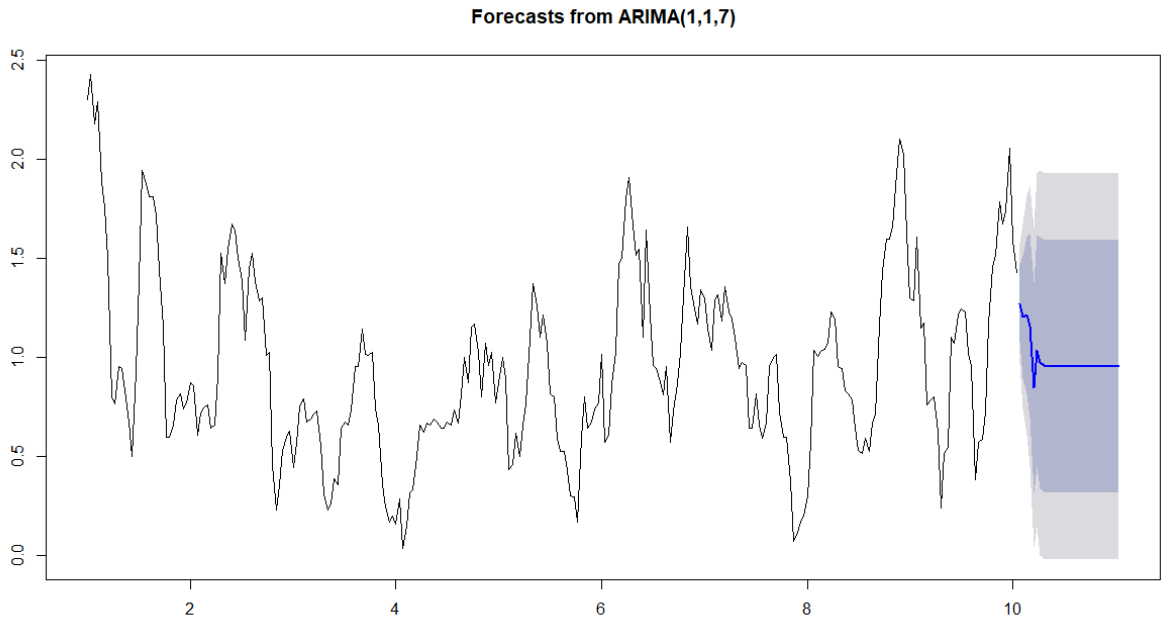


Figure 65: Forecast from Arima, not very good.

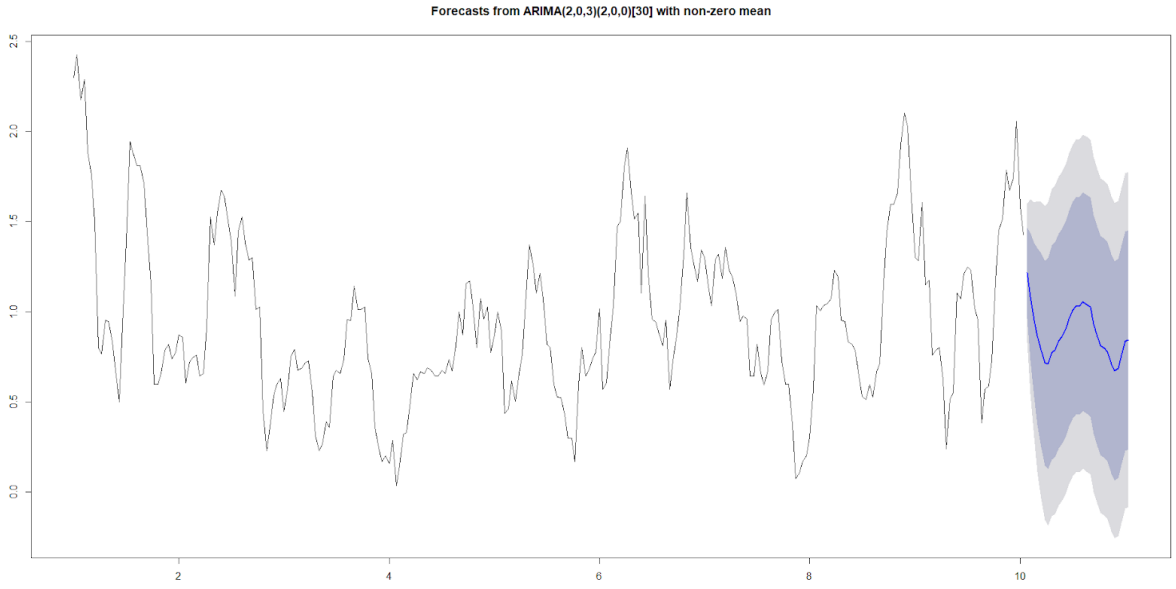


Figure 66: Arima model forecast with non-zero means.

```
Series: seasonal_nkilled
ARIMA(2,0,3)(2,0,0)[30] with non-zero mean

Coefficients:
          ar1      ar2      ma1     ma2     ma3     sar1
      1.6619  -0.7525  -0.6801  0.0745  0.1118  -0.0226
s.e.  0.0942   0.0823   0.1053  0.0687  0.0864   0.0714
         sar2    mean
      -0.2074  0.9355
s.e.   0.0738  0.0560

sigma^2 estimated as 0.03872:  log likelihood=57.96
AIC=-97.92   AICc=-97.24   BIC=-65.47
```

Figure 67: Arima model forecast

After adjusting the seasonality and lags, the Arima model performed well for 2013, however our analysis was stirred, excluding 2013. This would justify creating an Arima model and testing it on the other years within the timeline of he project.

3.3.4.3 Word Cloud

```
> head(gunViolence$incident_characteristics)
[1] "Shot - Wounded/Injured||Mass Shooting (4+ victims injured or killed excluding the subject/suspect/perpetrator, one location)||Possession
 (gun(s) found during commission of other crimes)||Possession of gun by felon or prohibited person"
[2] "Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide)||Mass Shooting (4+ victims injured or killed excluding the subject/susp
ect/perpetrator, one location)||Gang involvement"
[3] "Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide)||Shots Fired - No Injuries||Bar/club incident - in or around establishm
ent"
[4] "Shot - Dead (murder, accidental, suicide)||Officer Involved Incident||Officer Involved Shooting - subject/suspect/perpetrator killed||Dru
g involvement||Kidnapping/abductions/hostage||Under the influence of alcohol or drugs (only applies to the subject/suspect/perpetrator )"
[5] "Shot - Wounded/Injured||Shot - Dead (murder, accidental, suicide)||Suicide^||Murder/Suicide||Attempted Murder/Suicide (one variable unsuc
cessful)||Domestic Violence"
[6] "Shot - Dead (murder, accidental, suicide)||Home Invasion||Home Invasion - Resident killed||Mass Shooting (4+ victims injured or killed ex
cluding the subject/suspect/perpetrator, one location)||Armed robbery with injury/death and/or evidence of DGU found"
>
```

Figure 68: Head shot of the top few values for incident characteristics.

With some values being extremely long and containing a lot of information a word cloud was created in figure x to visually see the top 50 words in the incident characteristics variable.

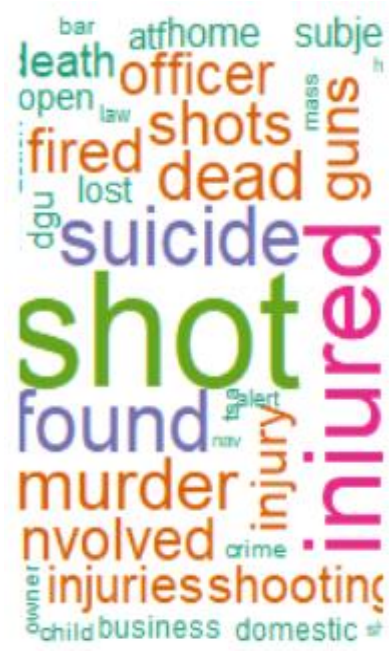


Figure 69: Incident characteristics word cloud.

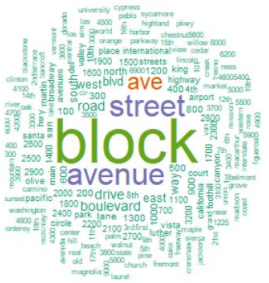


Figure 70: Location description word cloud for California.

Figure 71: Word cloud for location description

	word	freq
block	block	7108
avenue	avenue	3098
street	street	3070
ave	ave	1976
road	road	812
drive	drive	775
boulevard	boulevard	737
way	way	628
blvd	blvd	601
east	east	550

	word	freq
park	park	265
district	district	126
school	school	86
apartments	apartments	72
county	county	70
center	center	59
police	police	59
department	department	58
inn	inn	57
high	high	51

Figure on right showing how frequent top 10 frequent words come up in location description for California state. Figure on left showing top 10 frequent words in California's addresses.

"theater"	"lake"	"park"	"college"	"santa"	"area"	"recreation"	"market"
"airport"	"angeles"	"lax"	"los"	"marina"	"san"	"rosa"	"store"
"walgreens"	"oakland"	"beach"	"inn"	"aid"	"pharmacy"	"rite"	"motel"
"center"	"community"	"office"	"cal"	"arts"	"bay"	"east"	"highland"
"hospital"	"apartment"	"cents"	"county"	"fresno"	"range"	"sunset"	"lounge"
"golden"	"hotel"	"dennys"	"restaurant"	"river"	"south"	"city"	"long"
"super"	"crossroads"	"trailer"	"club"	"gun"	"oak"	"tree"	"bar"
"fish"	"grill"	"house"	"apts"	"village"	"burger"	"storage"	"creek"
"resort"	"church"	"regional"	"desert"	"palm"	"shell"	"firearms"	"king"
"gas"	"station"	"valero"	"mission"	"health"	"lot"	"parking"	"diego"
"training"	"nightclub"	"ranch"	"elementary"	"school"	"smf"	"sfo"	"travel"
"bur"	"liquor"	"target"	"big"	"shopping"	"sna"	"manor"	"lincoln"
"united"	"indian"	"hills"	"high"	"pier"	"country"	"home"	"mobile"
"apartments"	"box"	"jack"	"laguna"	"green"	"department"	"police"	"bank"
"union"	"mall"	"casa"	"ont"	"valley"	"oaks"	"twin"	"street"
"pacific"	"sports"	"metro"	"lucky"	"row"	"district"	"mart"	"mini"
"north"	"bell"	"taco"	"subway"	"hill"	"bart"	"gardens"	"sacramento"
"bowl"	"food"	"place"	"cambridge"	"courthouse"	"mountain"	"intl"	"hyde"
"neighborhood"	"potrero"	"gate"	"complex"	"view"	"national"	"mcdonalds"	"sierra"
"meadows"	"attorneys"	"west"	"villa"	"burgers"	"california"	"university"	"garden"
"hookah"	"library"	"rancho"	"square"	"suites"	"chevron"	"cafe"	"del"
"sol"	"martin"	"state"	"dollar"	"general"	"walmart"	"arco"	"medical"
"shop"	"ampm"	"mexican"	"lodge"	"town"	"reservation"	"bear"	"casino"
"jewelry"	"hollywood"	"grove"	"supermarket"	"estates"	"star"	"solano"	"7eleven"
"sea"	"beverly"	"america"	"sheriffs"	"starbucks"	"middle"	"shooting"	"ontario"
"kings"	"valencia"	"fastrip"	"francisco"	"international"	"stop"	"kfc"	"circle"
"vista"	"depot"	"academy"	"fashion"	"jose"	"hall"	"family"	"silver"
"bayview"	"heights"	"heritage"	"memorial"	"blue"	"canyon"	"boyle"	"red"
"safeway"	"plaza"	"collective"	"american"	"grocery"	"costco"	"best"	"civic"
"central"	"florence"	"jail"	"chinatown"	"charter"	"ocean"	"tenderloin"	"mesa"
"line"	"factory"	"oceanview"	"castle"	"vermont"	"quality"	"compton"	"fairgrounds"
"gateway"	"forest"	"berkeley"	"bridge"	"campus"	"huntington"	"royal"	"less"
"ridge"	"midcity"	"arden"	"block"	"terrace"	"sheriff"	"bakersfield"	"oildale"
"seaport"	"lakeview"	"smoke"	"humboldt"	"eureka"	"attorney"	"richmond"	

Figure 72: Words that came up at least four times in the dataset regarding gun violence.

Looking at these words alone give you insight into where gun violence occurs. Of course, it's very broad, but never the less, still intriguing.

	word	freq
0unknown	0unknown	98093
0handgun	0handgun	16628
auto	auto	5868
09mm	09mm	5446
1unknown	1unknown	4891
1handgun	1handgun	3207
0shotgun	0shotgun	2640
022	022	2611
0rifle	0rifle	2267
040	040	2244

Figure 73: Top 10 most frequent gun types in gun violence dataset.

3.3.4.4 Machine Learning

Machine Learning is a vital part of crime detection and prevention. It is used broadly across the world and is only getting bigger and at a higher demand. Machine learning is defined as the giving a computer the ability to learn without being explicitly programmed. There are two major parts, supervised and unsupervised learning.

Machine learning algorithms were not considered priority one as outlined in the early stages of this report. This is not due to them not being important, they are. But merely due to the fact an analysis on a topic like gun violence provided a journey of gaining insights that required more time than initially anticipated when starting the project. Linear Regression and Random Forest will be finished included in the code of this document for further analysis.

3.3.4.5 Tableau

A link to an interactive dashboard is attached with the code of this analysis for further exploration of the data.

3.3.5 Testing

A mixture of statistical tests were carried out such as the Shapiro Wilk test, which is a test for normality, An Augmented Dickey-Fuller Test, and a two sample proportionate z test.

As mentioned above, for an machine learning algorithms used in this analysis testing will be conducted to determine the performance accuracy of the model on future unseen data.

3.3.6 Evaluation and Recommendations

3.3.6.1 Key Analysis

Concluding from our analysis alcohol has a big contribution to gun violence, however this certainly requires a deeper investigation. With the different state laws for gun violence it would be beneficial for the law enforcement to invest in a tool proposed in this document, a tool that forecasts the location a crime is likely to happen based on past trends in that area. The model will take account of the types of guns use in that area over time to allow the law enforcement to be proactive instead of reactive. Gaining insight into the category of an incident, whether it is suicide, gang or drug related would also be useful to include in the model to also inform the law enforcement of the areas of the high rates for these categories, to encourage proactive action on lowering the rates and putting support in place for the areas that have high suicide rates relating gun violence, for the high drug related areas to perhaps open more drug rehabilitation centres and the same with gang related activity. All this information is very valuable to present a tool that can really make an impact of the rate of gun violence. Gun Violence is not the solution, and those in dark places considering suicide and gangs battling it out using lethal weapons like guns, it's the problem we are trying to face and tackle. If gun laws were more stricter and the same across states, it would lower the risk of states with stronger gun laws, having neighbours with friendlier approaches to gun laws crossing over and it effecting their states.

3.3.6.2 Recommendations

A deeper look into *America's tourism* would certainly not go to waste. It is during the holiday season that gun violence is at peak. Correlation is not causation, but research into the relationship between the two would be interesting. It would be of interest to determine how many tourists are charged for committing a crime when visiting the country. And more specifically, how many of those charges are gun related.

July 30th and October 25th have significantly high counts of gun violence recorded for. With them not being public holidays, or a significant date in history, a deeper look into why this might be worth be worthwhile.

3.3.6.3 Key Findings

The peaks of gun violence relate to special occasions such as New Years day, Independence Day, Holiday season/ tourist season which tend to relate to alcohol consumption. It would be near impossible to stop this completely, and with most states already prohibiting the handling of firearm around alcohol it would appear there isn't much more to be done. But no, there is always more to be done. Whether security checks in areas attracting alcohol is increased, or if the purchase of alcohol requires a brief search for firearms on person, something needs to be done. This itself is a very controversial topic and requires special attention in the further.

3.3.6.4 Project Changes

If I were to change this project, I would narrow in more on one state to gain a deeper insight into the trends within. This analysis has provided the ground works to do this, therefore this would be my next step. Throughout the project many changes were made. Discoveries led to unexpected paths, and research and data exploration exceeded the time limit set for these areas resulting in machine learning algorithms results not to be included in this report. The links to these will be included with the code. Arima was carried out on 2013, which was not then included in the rest of the analysis, which would be worth while investing time in after to conduct on the rest of the timeline of the analysis.

4 Conclusions

Gun Violence is such a massive topic. The more I analysed the data the more and more I began to realise the number of factors that are related to it have impact and require such attention. Alcohol was mentioned a lot throughout the analysis. One might think, well duh of course it impacts gun violence. But the level it impacts it, with the times gun violence is at peak, the day of he weak; its mid blowing. The discovery of this justifies future investigation into each states gun laws regarding alcohol along with the statistics on how many crimes relating guns are in fact alcohol related. This document briefly touches on this, referencing to (Branas, C., n.d) statement that in 2008 a total of 46 laws in 31 states restricted the interaction of alcohol and firearms. Overall a better understanding into the structure of America's gun laws and the times and places gun violence is at peak, factoring the state's population, the category of the incident, whether it was gang, suicide and drug related were discovered. These discoveries along with the fact that Hawaii has the lowest gun violence crime and has been ranked the happiest state in America numerous times and is considered one of the states with strong gun laws, suggest that there is in fact a relationship between gun laws and the happiness of a state. The factors that contribute to the happiness are based on well being which is not directly relating gun violence. This leads to curiosity on whether violence was included in the happiness rankings as to what the results would be. Another one to investigate in the future. This analysis has provided the groundings of a very powerful and useful too and the hunger to delve deeper into gun violence in America and even perhaps other parts of the world.

5 Further Development or Research

Bibliography

Lopez, G. (2017). The research is clear: gun control saves lives. Washington: Vox Media. Available at: https://www.vox.com/policy-and-politics/2017/10/4/16418754/gun-control-washington-post [Accessed 2 May 2019].

Kaggle, (2019). 'Datasets', Kaggle. Available at: https://www.kaggle.com/datasets [Accessed 4 May 2019].

Gun Violence Archive, (2019). 'Gun Violence Archive', Available at: https://www.gunviolencearchive.org/ [Accessed 3rd May 2019].

Bureau, US Census., (2019) "About the Bureau.", Available at: https://www.census.gov/about-us [Accessed 4 May 2019].

Giffords law center, (2018) 'Giffords Law Center's Annual Gun Law Scorecard.', Available at: https://lawcenter.giffords.org/scorecard. [Accessed May 3, 2019].

BBC News. (2017, October 4). Las Vegas shooting: Gun used `bump-stock` device to shoot faster. BBC News, p. 1. Available at: https://www.bbc.com/news/world-us-canada-41482010 [Accessed 3 May 2019].

Inc, Gallup., (2018), "Hawaii Tops U.S. in Wellbeing for Record 7th Time." Gallup, Available at: https://news.gallup.com/poll/247034/hawaii-tops-wellbeing-record-7th-time.aspx [Accessed 3 May 2019].

Economist, The., (2017, October 5). 'The Las Vegas shooting has reinvigorated calls for gun control', The Economist Newspaper Limited, Available at: https://www.economist.com/united-states/2017/10/05/the-las-vegas-shooting-has-reinvigorated-calls-for-gun-control [Accessed 29 September 2018].

Carry, G. T. (2017). 'Gun Laws By State', Guns To Carry, Available at: https://www.gunstocarry.com/gun-laws-state/ [Accessed 12 October 2018].

Kabacoff, R. I. (2017). 'Graphics with ggplot2. Quick-R', Available at: https://www.statmethods.net/advgraphs/ggplot2.html [Accessed 12 October 2018].

RJ Reinhart, (2018). 'Six in 10 Americans Support Stricter Gun Laws.', Available at: https://news.gallup.com/poll/243797/six-americans-support-stricter-gun-laws.aspx [Accessed 12 October 2018].
Siegel, Michael, Molly Pahn, Ziming Xuan, Craig S. Ross, Sandro Galea, Bindu Kalesan, Eric Fleegler, and Kristin A. Goss. (2017). "Firearm-Related Laws in All 50 US States, 1991-2016." American Journal of Public Health 107 (7): 1122–29. https://doi.org/10.2105/AJPH.2017.303701. [Accessed 12 October 2018].
Shawndra.pbworks.com. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communication Of The ACM. Available at: http://shawndra.pbworks.com/f/The%20KDD%20process%20for%20extracting%20useful%20knowledge%20from%20volumes%20of%20data.pdf [Accessed 12 December 2018].
RDocumentation, (2019), 'as_tibble', RDocumentation. Available at: https://www.rdocumentation.org/packages/tibble/versions/1.4.2/topics/as_tibble [Accessed 4 May 2019].
Gupta, Prashant., (2016). 'Cross-Validation in Machine Learning', Analytics Vidhya. Available at: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f [Accessed 3 May 2019].
Smarten (2018). 'What is ARIMA Forecasting and How Can it Be Used for Enterprise Analysis?', Smarten Advance Data Discovery. Available at: https://www.smarten.com/blog/what-is-arima-forecasting-and-how-can-it-be-used-for-enterprise-analysis/ [Accessed 4 May 2019].
Expert System (2017). 'What is Machine Learning?', Expert Systems, Machine Learning definitions. Available at: https://www.expertsystem.com/machine-learning-definition/ [Accessed 12 December 2018].
Eekhout, Iris (2019).'Incomplete data can contain valuable information', Don't Miss Out! Available at: https://www.iriseekhout.com/ [Accessed 4 May 2019].
Kapocsi, C., (2018). 'A Beginner's Guide on How to Use Dropbox', Cloudwards. Available at: https://www.cloudwards.net/how-to-use-dropbox/#one [Accessed 5 May 2019].
Treehouse Island, Inc., (2019).'How Colour Communicates Meaning', treehouse. Available at: https://blog.teamtreehouse.com/how-colour-communicates-meaning [Accessed 5 May 2019].
World Tourism Organization, (2018). 'United States: Country-specific: Arrivals of non-resident tourists at national borders, by country of residence 2013 - 2017 (09.2018)', UNWTO World Tourism Organization. Available at: https://www.e-unwto.org/doi/abs/10.5555/unwtotfb0840011220132017201809 [Accessed 5 May 2019].
Williams, S., Chaplain, C., (2018).'What is American Independence Day? And why is it on July 4th?', *Evening Standard*. Available at: https://www.standard.co.uk/news/world/what-is-american-independence-day-2018-and-why-is-it-celebrated-on-4th-july-a3877171.html [Accessed 4 May 2019].
Branas, C., (n.d).'Alcohol & Firearms: Research, Gaps in Knowledge, and possible interventions', Charles Branas, PhD. Available at: whats a system architecture in data analytics project [Accessed 4 May 2019].

6 Appendix

Table 40.


ANALYSIS OF GUN VIOLENCE IN AMERICA
National College of Ireland
TECHNOLOGY MANAGEMENT, DATA ANALYTICS
Charlene Moore
DATA
INSIGHTS
ACTIONS
BIG DATA
R Studio
Analytics
Insight
Gun Violence incidents per 100,000
GUN VIOLENCE IS NOT THE SOLUTION
IT'S THE PROBLEM