National College of Ireland

BSc in Technology Management

Data Analytics

2020/2021

Conor Hillyer

x17403374

x17403374@student.ncirl.ie

US Police Shootings

Data Analysis Report

Contents

Executive Summary

The purpose of this report is to give the reader an overall understanding of what my project is based on, what I set out to achieve, what I did achieve, how I did achieve the results of my findings and most importantly how to interpret the data in a meaningful way.

Some of the key points of this report that are most important for the reader to be familiar with are as follows:

- Introduction: To understand what the project is based around, what my aims and objectives were, and understand what technologies I am using to carry out my analysis.
- Data: This is an important section so that anyone reading can understand where I acquired my data and how it is structured.
- Results: A clear and concise way to read and comprehend what was achieved by me completing my analysis on the dataset.
- Conclusion: This contains a summary of what I completed and also describes any advantages/disadvantages and or limitations that I encountered while completing my project.

In terms of recommendations for the use of this report, I suggest that it could be used as a tool to try and show states, cities, and towns of the US how people have been affected by police shootings and how they can be avoided by looking at the results of my analysis as reasons to do so.

1.0 Introduction

1.1. Background

I chose to undertake this project in particular as I believe it will serve a purpose to the citizens of the USA who are affected most by Police shootings. As my data set is built around real people who were killed by US Police over a certain period of time, I am sure that anyone that reads my analysis will understand why it is important to know how the data is broken down so that future shootings and killings can be avoided as much as possible in the future. (Nazir, AN, 2020/1, Kaggle.com)

1.2. Aims

There are a wide range of aims and objectives that I would like to achieve while analysing my data set. The main aim of my project is to break down and analyse my data set to uncover any unusual or abnormal data points from the data set. By uncovering any abnormal or unexpected visualisations we can identify where things need to change to try and avoid any unwarranted deaths produced by police in the US.

1.3. Technology

Throughout the development of this project, I have progressively included more technologies as I progressed further into my investigation. Starting, while preparing for the Mid-Presentation I had use and access to four technologies. These technologies were Excel, SPSS, R in R-Studio and Tableau. After the mid-point presentation, I then moved onto include Python into the analysis as I felt that it would make a meaningful addition when compared to the other technologies that I had already implemented.

Below you can see what each technology was used for along with some challenges that I faced with each one and how I overcame these challenges.

- Excel: Excel is where my data analysis project started. As did not need to learn how to use the software and I already had a good understanding of how it worked I could jump straight in with some exploratory data analysis and filtering of the data. As filtering is very easy to apply to data in excel, I applied that function to the age column so that I could get a better understanding of this column of data. You can see some examples of the filters that were applied in the results section of the report.

❖ Challenge: There were no challenges encountered when using excel as I have had plenty of experience over my work and college experience.

- SPSS: SPSS was a new technology to me in my fourth year of study. After using it in a number of modules during my final year, however, I then felt confident that I could get some good results and outputs from the powerful software. SPSS was mainly used for descriptive data analysis methods as the software creates very easy to read and interpret images based on the data that you feed it. SPSS was also used to build some graphs on the data as it has a very useful graph builder function that can be very useful when you are looking for a very specific graph to be built.

❖ Challenge: One challenge that I did encounter with the use of SPSS is that because SPSS has so many potential uses, I sometimes found it hard to understand what function I needed to use to receive my desired outcome. This was overcome easily however as IBM, who developed SPSS has a great website explaining all of SPSSs functions and operations.

- R: The use of R at the start of my project was limited to filtering and some basic graphs as I did not have the experience that I needed to make the most out of the technology. After achieving some basic filtering and graphing results I moved onto Python as I thought that is all I could achieve with R. After revisiting the technology, however, I did a lot more research on R and then managed to produce some more impressive graphs and outputs.

❖ Challenge: One large issue that I had to overcome was how to properly style the outputs. Producing basic outputs in R is relatively easy but styling them to how you like is what I struggled with when starting with R. With practice and referencing online forums this issue was overcome.

- Tableau: As I had previous experience with Tableau in my third year of study, I thought it would be a good technology to implement into my project. As tableau is very user friendly and requires very little research regarding how to implement it, it made a lot of sense to use this technology as an additional way of visualising the data for the project.

❖ Challenge: The only challenge that I came across while using Tableau in my project was selecting the correct columns to plot onto the sheet. As most of the data in the data set are non-numerical this was a challenge. One way I got around this issue however is that Tableau automatically converts some of the columns into figures by counting the contents of each column. This then allows you to plot the now numerical column against other non-numerical columns.

- Python: The last technology that I decided to implement was Python. As I had no classes that taught me how to use Python, I resorted to Udemy online learning to figure out how to use this technology. After purchasing and completing a course in Python I then felt confident that I could then implement it into my project. Quickly after using Python on my data, I realised how versatile and complex Python can be. The technology was used for descriptive and visualisation methods, but data mining is also a possibility when using Python. For the purposes of my project, however, data mining was not used in Python.

❖ Challenge: Several challenges were encountered while using Python for my project. This was mostly because that I had very little experience with the technology. Getting used to how packages worked and were called upon when compared to R took some getting used to but all of the challenges that I encountered were overcome by implementing trial and error and getting as much experience with the technology as possible.

1.4. Structure

The sections outlined in this project are as follows with brief explanations for why they are included:

- Executive Summary: This section is provided so to give the reader a condensed view of the document and what is included.
- Introduction: Here is where my project is introduced, aims are spoken of and my choice of Technologies are outlined.

- Data: My data set is described and explained in this section so the reader understands where it was sourced and can learn some basic attributes of the data.
- Methodology: In this section, I describe how I went about deciding how to carry out my aims and objectives of the data.
- Analysis: During this part, I go into detail about how I carried out my analysis of my data and why I chose these methods over other techniques
- Results: All of the results from every piece of analysis done on the project can be found here in various formats so that the reader can understand what was achieved from carrying out the analysis.
- Conclusion: A description of the advantages, disadvantages and limitations of the project are included in this section of the report so the reader can be aware of what made the project challenging and difficult.
- Further Development or Research: Here is where I explain what I hope to achieve with any further development and research on this data set.

1.5. Business Use Case

In terms of a business use case for this project. I have come up with several scenarios where this could be useful to both the US public and the US Police forces.

- Informing the Most Affected: One example of how this report could be used to reduce the amount police shootings is to educate and inform the people that are most affected by police shootings. As this group appears to be young males, a solution could be to start running informative talks in schools on how to avoid the kinds of behaviour and activity that statistically ends up with suspects being shot and killed by police. This training could involve giving statistics on gun and knife crimes and how Police are trained to deal with armed suspects. Hopefully, by supplying this group of people with this type of information they can then make more rational decisions on how to avoid being involved with police and hence police shootings.
- Updated Police Training: Even though police training does involve showing police officers how to deescalate situations there is still a large number of shootings occurring that by looking at the data did not need to occur. The types of data that point to that conclusion is by looking at the data of unarmed and non-threatening suspects being shot by police. I believe with more training, that police officers could bring a larger number of suspects under control without using lethal force. With more training on how to deal with suspects with mental illness, suspects who are children and suspects who may not understand how serious the situation may be that these numbers would come down.
- Each State to Receive Personalized Statistics: As the United States is such a large and diverse country it is important to not deploy a national approach to the situation. By giving each state access to detailed statistics like in this report then each state could take on methods that would improve their individual cases. This would give more power to the states to implement plans and solutions that would work for their state but maybe not for another.

2.0 Data

While looking online for suitable data sets for my project I navigated to Kaggle.com to find one as I used this service last year in another module. I liked this service as it allows you to search their online database of data sets for one specific to your needs, whether that be the size of the data set or the topic that the data set is based around. After looking at several sets on Kaggle, including ones involving Formula 1 and traffic accidents in the UK I discovered the data set that I wanted to work with, “US Shootings”. I chose this data set over the others as it was larger so it allowed me to go deeper in the analysis, it felt more important than analysing wins or traffic accidents as hopefully after this analysis is complete anyone can read my report and realise how people are being affected by police shootings every day in the US. It was also a topic that I was not super familiar with as it is not something that I or anyone I know follows regularly as I do with Formula 1, so it was more interesting to me than the other options that I had looked at.

After picking the data set, I had to run some basic statistical techniques to make sure that the data set had enough in it to achieve the kind of high-level analysis that I wanted it to have. By checking how many rows, columns, and the number of missing values for example. This then let me see how well structured the data was for analysis and visualisation of data. Looking at how many variables are available in each column was also a good way to see what type of data I could pull from the data e.g., The “Race” column having six variables and the “Arms_Category” having twelve variables. Having a good number of variables in each column was a good indication to see how many unique filters/visualisations could be made once I started plotting these columns against each other to create visualisations and plots.

I also ran some exploratory data analysis on the data before diving right into visualisations. This involved using SPSS, Python and R to get a better look at the data by running functions such as “Frequency Table”, “Descriptives” and “PPlot” in SPSS, things like “data.shape”, “data.info()”, “data.describe()”, “data.nunique()” and “ProfileReport(data)” in Python and items like “str(Shootings)”, “summary(Shootings)” and “head(Shootings)” in R. The outputs of these commands and functions gave me a much greater understanding on what is contained within the data. The data set itself is contained within a CSV File. The data contained within the file is ready for visualisation out of the box and no cleaning or pre-processing was necessary before I started performing my analysis. You can have a look at the data yourself here: https://www.kaggle.com/ahsen1330/us-police-shootings.

(Nazir, AN, 2020/1, Kaggle.com)

Below you can see some basic attributes of the data:

- Name of File = “shootings”
- Number of columns = 15
- Number of Rows = 4895
- Format = CSV

- Data Structure = CSV, Semi-Structured
- Number of Numerical Columns = 3
- Number of Character Columns = 12
- Size of file = 506K

3.0 Methodology

For the purpose of this project, a data mining methodology was implemented from start to finish of the project. I had two choices for this methodology, one being KDD (Knowledge Discovery Database) and CRISP-DM (Cross Industry Process for Data Mining. As CRSIP-DM is more focused on the business side of data mining as it includes subjects such as business understanding, and this is a research paper I used the KDD Methodology. The overall goal of using a process like KDD is to extract knowledge from data in the form of databases or the case of this project a CSV file. One important thing to note about the KDD Methodology is that it can vary from project to project, you notice when the process is explained that some steps may be not needed or only a small amount of work is needed to fur fill that part of the methodology. To follow and implement the KDD methodology there several steps to follow to ensure that the process is adhered to. These steps are explained in detail below:

- Selection: For this project, this involved me, the writer of the project to go and search for a data set that was suitable for data analysis techniques and also one that would be interesting to run an analysis on. The data that was used in this project was sourced from a well-known website in the data analysis community. Kaggle.com contains thousands of data sets that are available to the public for data analysis and data mining uses such as projects like mine.

- Pre-processing: The process of Pre-processing the data is designed to “clean” the data so that it is more suitable for data analysis and data mining methods. Luckily for me, the data set that I selected for the project was already cleaned before it was uploaded to Kaggle.com. The process of cleaning the data can involve many things but one of the most important is making sure that there are no missing values that could impair the results you produce from the data. My data set has zero missing values, so it was ready for data analysis and data mining from the start of the project.

- Transformation: Like the step of Pre-processing the Transformation section of the KDD methodology is designed to prepare the data to be analysed as we would like. This process largely depends on what type of analysis you would like to do with the data. Again, like the above section, the data was already suitable for data analysis without the need for any transformation of data.

- Data Mining: The data mining section of the KDD is where I performed all of my statistical tests and created as many visualisations as possible to help understand why the data is structured the way it is. This is arguably one of the most important steps in the process and is where the majority of the work occurred. This is also where all of my discussed technologies were implemented.

- Interpretation and Evaluation: In this final section of KDD all of the produced data visualisations and outputs need to be sifted through and organised into what is relevant and what is not. This is an important process as it determines what the outcome of the entire project turns out to be. Based on this review process I decided on what outputs I wanted to include in the report and also which ones we could ignore. The basis for these decisions comes down to a few decisions, these being: Does each output perform as well as we would like it too? Is the report going to have any overlap in outputs (this is likely while using multiple technologies)? Do the outputs that we have selected tell the narrative that the writer is trying to portray? If each visualisation can answer positively to these questions, then it is likely that it will be included in the final report. This was a long process for me to complete as I had to go through and examine all of the outputs that I had produced in four technologies since November of last year. Once I had selected the outputs that I wanted to include in this report I then converted them all to images so that they could be implemented in Word without difficulty. (Rajput, AR, 2019, KDD Process in Data Mining)

4.0 Analysis

My approach to my analysis of my data changed depending on which technology I was using. When I wanted to filter and visualise, I would use R, if I wanted to filter with dates and months, I would use Excel, when I needed statistical analysis, I would use SPSS when I wanted to confirm the results of my created visualisations from other technologies, I would use Tableau and when I wanted to visualise the data in a more advanced way Python would be implemented. This approach was used so that I could make the most out of each of my chosen technologies but also make it as easy as possible for me to get the type of visualisation or statistical figure I wanted to create.

Some of the key attributes that I focused on while analysing the data were “Race”, “Gender”, “Age”, “State”, “Signs of Mental Illness”, “Flee”, “Body Camera” and “Arms Category”. All of these attributes are capable of providing some sort of unique analysis of the data. Some of the most important attributes such as “Age”, “Gender” and “State” were able to provide some very interesting and sometimes shocking data points once analysed.

I chose attributes such as “Age” and “Gender” as some of the main attributes to focus on as these are telling of what types of people are being involved in these incidents. Attributes such as “body camera” and “flee” are interesting for statistical evaluations of the data to see how many times these may have been factors but ultimately, they will not change how often these Police Shooting incidents are taking place.

Once the information was acquired, I could make sense of it and explain what it means to the overall analysis of the data and make a point about whether this particular part of the analysis was worth bringing up in my report or if they seemed to be at normal levels. Deciding which visualisations and code to include in my results section of my report was challenging at times as not all of the visualisations were as useful as others. By looking at what each visualisation or line of code provided me with was the most important thing here.

Now that I have explained some of the rationales of why I selected some attributes over others for analysis is covered, I will now give a detailed description of the type of methods used in each of the technologies that I used to analyse the data for the project.

Excel: While using excel for its ease of use and easy methods for viewing data I primarily used the "Filter" function while trying to access specific parts of the data. This was specifically useful while trying to get a hold of how a certain period of time was a factor in the shootings. This function was also used extensively to pick out specific age groups. This allowed me to see not just the number of people in each age group but also their names and other details associated with their deaths. For details on how different age groups were affected by the shootings please navigate down to the results section.

SPSS: As SPSS is known for its statistical and visualisation power that is what I used it for most. I used the following functions and attributes of SPSS to get the results I wanted: "Frequency Table", "Descriptives", "PPlot", and "Chart Builder".

Tableau: When using Tableau in my analysis, I would use it to confirm the results of visualisations made in other technologies. I would also use its "count" function which helped visualise the categorical columns much easier than in some other technologies.

R: As R is a programming language and not as easy to use like Excel and SPSS there a lot more functions needed to get the outputs that I was looking for. The most important and used functions and packages I used in R were as follows:

- "library(RColorBrewer)": For colour usage.
- "library(waffle)": For use of a specific type of visualisation.
- "library(ggplot2)": The most popular visualisation package on R.
- "Shootings[Shootings$race == "Black"]": How I filtered out specific parts of the data.
- "qplot" = How certain visualisations were created.
- "pie" = How pie charts were created.
- "ggplot" = How most visualisations were made.
- "waffle" = How to visualise lots of data as coloured cubes.
- "geom_density" = Smooth histogram.
- "geom_histogram" = Normal Histogram.
- "facet_wrap" = Multiple panels of a visualisation.
- "geom_violin" = Creates a Violin Plot
- "geom_boxplot" = Creates a boxplot

To view the results of these functions please view the results section of this report.

Python:

Again, as Python is a coding language and not as easy to use or manipulate as SPSS and Excel lots of functions and packages are needed to create the kind of outputs that we are looking to create. A list of the most used and important functions and packages can be seen below alongside a brief explanation.

- “import numpy as np” = For working with arrays.
- “import pandas as pd” = Used for data analysis.
- “import matplotlib.pyplot as plt” = Used for visualisation and graphing.
- “import seaborn as sns” = Used for data visualisation.
- “import plotly.graph_objects as go” = Used for the choropleth maps.
- “import plotly.graph_objects as go” = Used to create violin plots.
- “from pandas_profiling import ProfileReport” = Used to run the pandas report.
- “import missingno as msno” = Used for the visualisation of the missing values.
- “from wordcloud import WordCloud” = Used for the word cloud visualisation.
- “sns.countplot” = Creates a histogram
- “data.groupby” = To select specific parts of the data needed for visualisation.
- “plt.imshow” = Used to plot a word cloud.
- “sns.displot” = Shows a histogram with a line attached.
- “go.Figure” = Using plotly visualisation.
- “sns.boxplot” = Creating a boxplot
- “sns.swarmplot” = Plotting lots of individual data points. (No overlapping)
- “sns.stripplot” = Similar to swarmplot.
- “sns.barplot” = bar graph.
- “sns.kdeplot” = Shows the distribution of the data.
- “go.Choropleth” = Used to create mapping visualisations.

5.0 Results

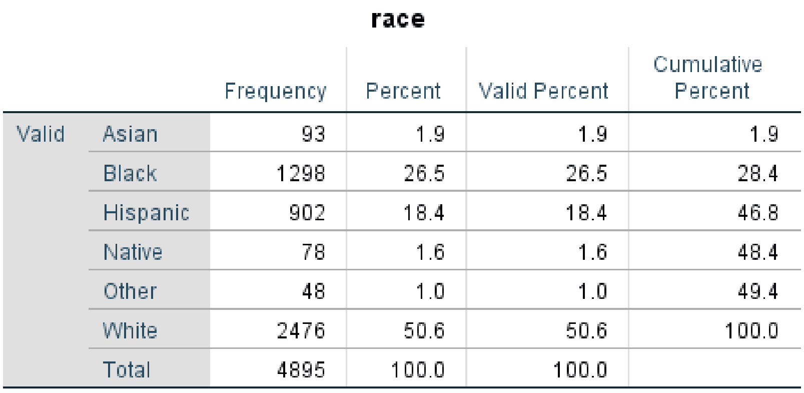
In this section of the report, I will be breaking down each and every output from the four technologies that I have used to analyse the data. The outputs are being displayed in no particular order as there is no perfect way to organise the outputs by what they are displaying. To see how each main category of the data performed please visit the conclusion section where this is provided.

To start, I will be displaying the descriptive statistics and basic statistical information about both the overall data set and its core attributes. These are being shown to give you the reader a brief understanding of how the data is broken down before I start presenting visualisations.

Descriptive Statistics and SPSS Outputs:

Below in "Fig. 1" we can see the descriptive statistics of the "Race" column. This output describes the frequency of each race within the "Race" column, the percentage value that each race takes up of the total deaths. This output was achieved via SPSS. Now that I have explained how this type of output is read, I will now display the others and allow you the reader to interpret them.

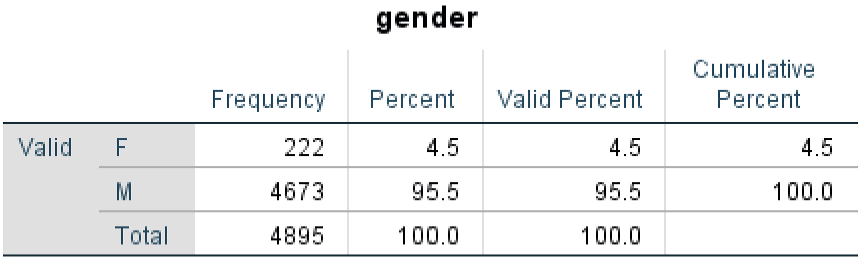
Fig. 1



race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Asian	93	1.9	1.9	1.9
	Black	1298	26.5	26.5	28.4
	Hispanic	902	18.4	18.4	46.8
	Native	78	1.6	1.6	48.4
	Other	48	1.0	1.0	49.4
	White	2476	50.6	50.6	100.0
	Total	4895	100.0	100.0	

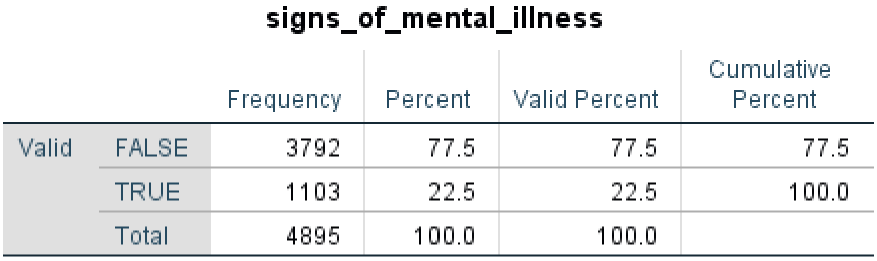
Fig. 2



gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	222	4.5	4.5	4.5
	M	4673	95.5	95.5	100.0
	Total	4895	100.0	100.0	

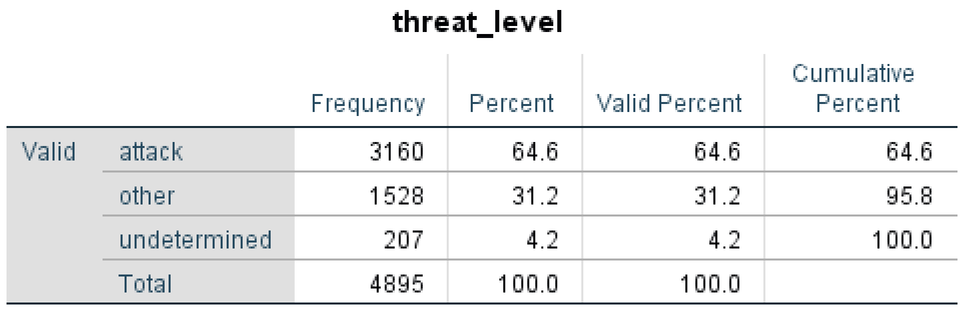
Fig. 3



signs_of_mental_illness

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FALSE	3792	77.5	77.5	77.5
	TRUE	1103	22.5	22.5	100.0
	Total	4895	100.0	100.0	

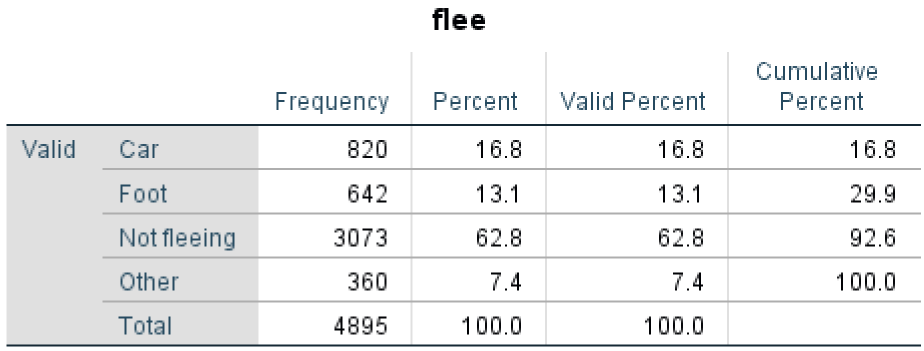
Fig. 4



threat_level

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	attack	3160	64.6	64.6	64.6
	other	1528	31.2	31.2	95.8
	undetermined	207	4.2	4.2	100.0
	Total	4895	100.0	100.0	

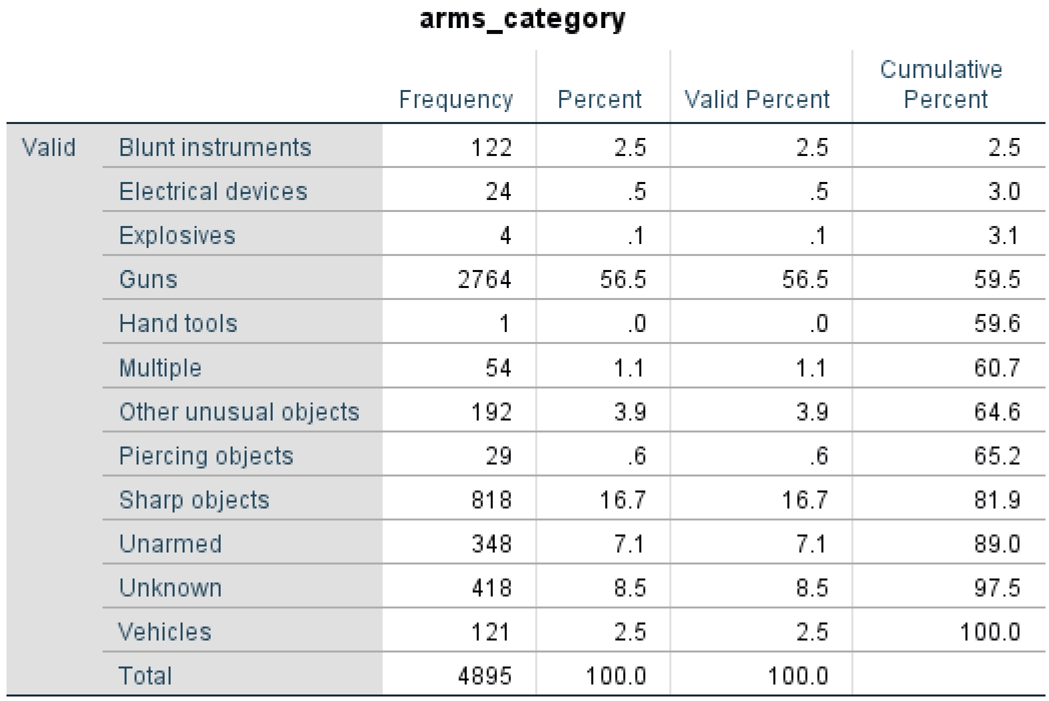
Fig. 5



flee

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Car	820	16.8	16.8	16.8
	Foot	642	13.1	13.1	29.9
	Not fleeing	3073	62.8	62.8	92.6
	Other	360	7.4	7.4	100.0
	Total	4895	100.0	100.0	

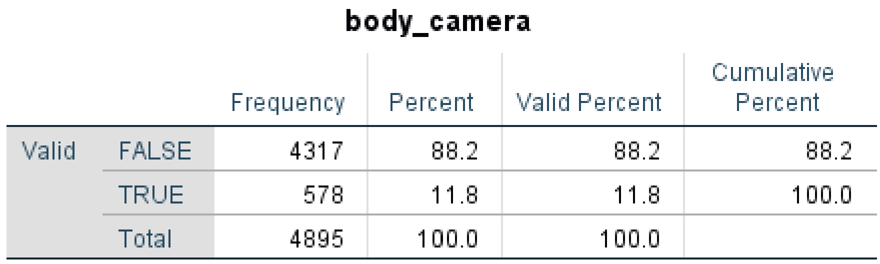
Fig. 6



arms_category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Blunt instruments	122	2.5	2.5	2.5
	Electrical devices	24	.5	.5	3.0
	Explosives	4	.1	.1	3.1
	Guns	2764	56.5	56.5	59.5
	Hand tools	1	.0	.0	59.6
	Multiple	54	1.1	1.1	60.7
	Other unusual objects	192	3.9	3.9	64.6
	Piercing objects	29	.6	.6	65.2
	Sharp objects	818	16.7	16.7	81.9
	Unarmed	348	7.1	7.1	89.0
	Unknown	418	8.5	8.5	97.5
	Vehicles	121	2.5	2.5	100.0
	Total	4895	100.0	100.0	

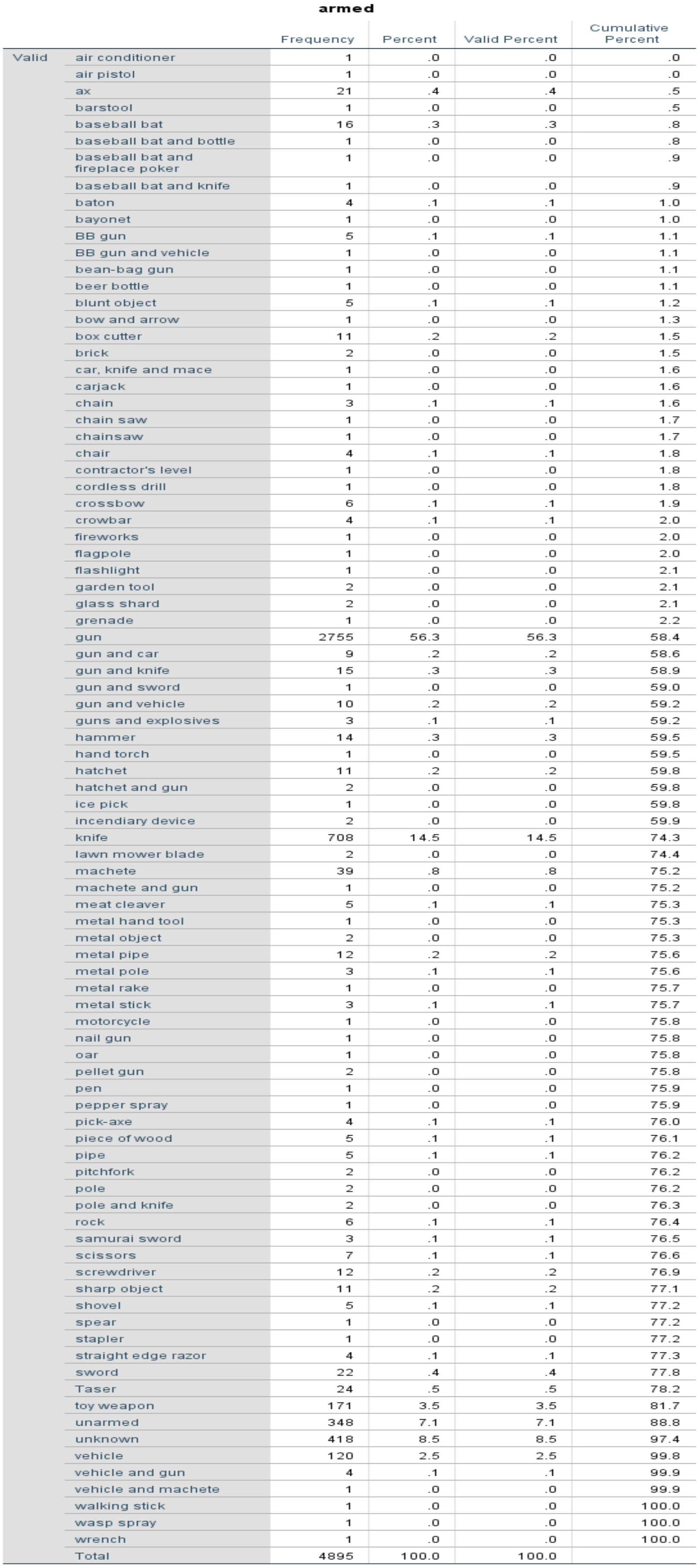
Fig. 7



body_camera

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FALSE	4317	88.2	88.2	88.2
	TRUE	578	11.8	11.8	100.0
	Total	4895	100.0	100.0	

Fig. 8

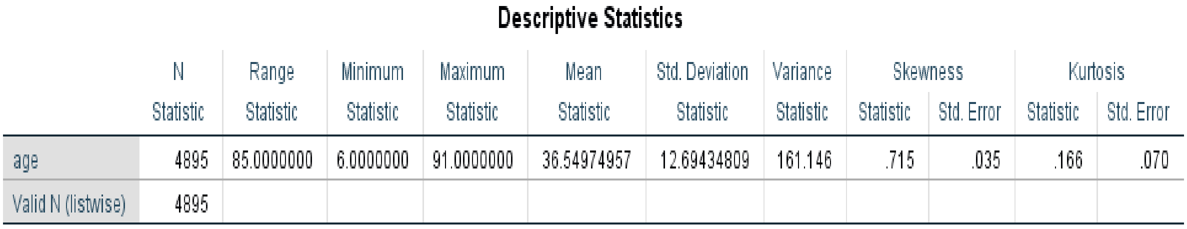


armed

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	air conditioner	1	.0	.0	.0
	air pistol	1	.0	.0	.0
	ax	21	.4	.4	.5
	barstool	1	.0	.0	.5
	baseball bat	16	.3	.3	.8
	baseball bat and bottle	1	.0	.0	.8
	baseball bat and fireplace poker	1	.0	.0	.9
	baseball bat and knife	1	.0	.0	.9
	baton	4	.1	.1	1.0
	bayonet	1	.0	.0	1.0
	BB gun	5	.1	.1	1.1
	BB gun and vehicle	1	.0	.0	1.1
	bean-bag gun	1	.0	.0	1.1
	beer bottle	1	.0	.0	1.1
	blunt object	5	.1	.1	1.2
	bow and arrow	1	.0	.0	1.3
	box cutter	11	.2	.2	1.5
	brick	2	.0	.0	1.5
	car, knife and mace	1	.0	.0	1.6
	carjack	1	.0	.0	1.6
	chain	3	.1	.1	1.6
	chain saw	1	.0	.0	1.7
	chainsaw	1	.0	.0	1.7
	chair	4	.1	.1	1.8
	contractor's level	1	.0	.0	1.8
	cordless drill	1	.0	.0	1.8
	crossbow	6	.1	.1	1.9
	crowbar	4	.1	.1	2.0
	fireworks	1	.0	.0	2.0
	flagpole	1	.0	.0	2.0
	flashlight	1	.0	.0	2.1
	garden tool	2	.0	.0	2.1
	glass shard	2	.0	.0	2.1
	grenade	1	.0	.0	2.2
	gun	2755	56.3	56.3	58.4
	gun and car	9	.2	.2	58.6
	gun and knife	15	.3	.3	58.9
	gun and sword	1	.0	.0	59.0
	gun and vehicle	10	.2	.2	59.2
	guns and explosives	3	.1	.1	59.2
	hammer	14	.3	.3	59.5
	hand torch	1	.0	.0	59.5
	hatchet	11	.2	.2	59.8
	hatchet and gun	2	.0	.0	59.8
	ice pick	1	.0	.0	59.8
	incendiary device	2	.0	.0	59.9
	knife	708	14.5	14.5	74.3
	lawn mower blade	2	.0	.0	74.4
	machete	39	.8	.8	75.2
	machete and gun	1	.0	.0	75.2
	meat cleaver	5	.1	.1	75.3
	metal hand tool	1	.0	.0	75.3
	metal object	2	.0	.0	75.3
	metal pipe	12	.2	.2	75.6
	metal pole	3	.1	.1	75.6
	metal rake	1	.0	.0	75.7
	metal stick	3	.1	.1	75.7
	motorcycle	1	.0	.0	75.8
	nail gun	1	.0	.0	75.8
	oar	1	.0	.0	75.8
	pellet gun	2	.0	.0	75.8
	pen	1	.0	.0	75.9
	pepper spray	1	.0	.0	75.9
	pick-axe	4	.1	.1	76.0
	piece of wood	5	.1	.1	76.1
	pipe	5	.1	.1	76.2
	pitchfork	2	.0	.0	76.2
	pole	2	.0	.0	76.2
	pole and knife	2	.0	.0	76.3
	rock	6	.1	.1	76.4
	samurai sword	3	.1	.1	76.5
	scissors	7	.1	.1	76.6
	screwdriver	12	.2	.2	76.9
	sharp object	11	.2	.2	77.1
	shovel	5	.1	.1	77.2
	spear	1	.0	.0	77.2
	stapler	1	.0	.0	77.2
	straight edge razor	4	.1	.1	77.3
	sword	22	.4	.4	77.8
	Taser	24	.5	.5	78.2
	toy weapon	171	3.5	3.5	81.7
	unarmed	348	7.1	7.1	88.8
	unknown	418	8.5	8.5	97.4
	vehicle	120	2.5	2.5	99.8
	vehicle and gun	4	.1	.1	99.9
	vehicle and machete	1	.0	.0	99.9
	walking stick	1	.0	.0	100.0
	wasp spray	1	.0	.0	100.0
	wrench	1	.0	.0	100.0
	Total	4895	100.0	100.0	

Below in “Fig. 9” we can see a breakdown of how the numerical column of “Age” is structured. We can see a number of important figures here including, the minimum age of a person being shot by police being “6” and the oldest being “91”. The overall average is sitting at “36.5” years of age. This is why descriptive statistics valuable to deploy as they give is a great insight into the data.

Fig. 9

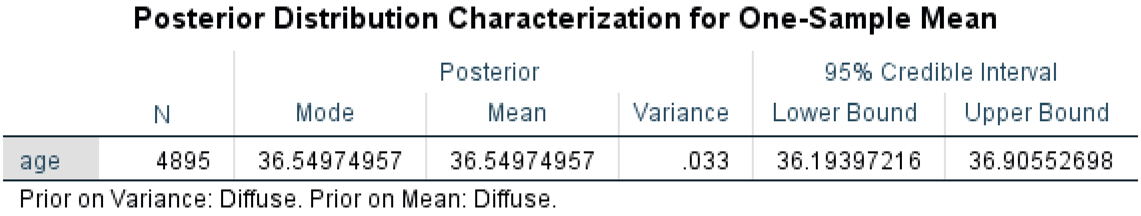


Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	4895	85.0000000	6.0000000	91.0000000	36.54974957	12.69434809	161.146	.715	.035	.166	.070
Valid N (listwise)	4895										

Another form of descriptive statistics that SPSS computed for the “Age” column is a One-Sample Mean test. As we can see below in “Fig. 10” it has given us the mode of 36.54 which is the most common variable in the data set, a variance of 0.33. This low variance is an indication that most of the values within “Age” are a value close to the mean of the data. This makes sense as we know most of the people affected by police shootings are around the age of 36.

Fig. 10



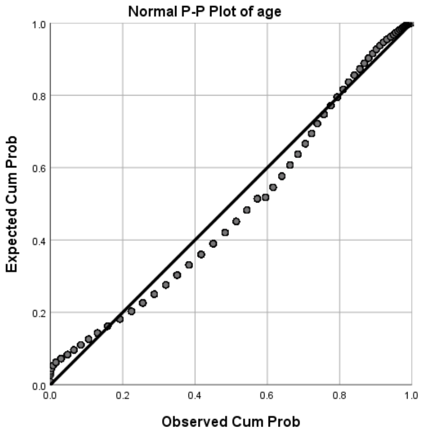
Posterior Distribution Characterization for One-Sample Mean

	N	Posterior			95% Credible Interval	
		Mode	Mean	Variance	Lower Bound	Upper Bound
age	4895	36.54974957	36.54974957	.033	36.19397216	36.90552698

Prior on Variance: Diffuse. Prior on Mean: Diffuse.

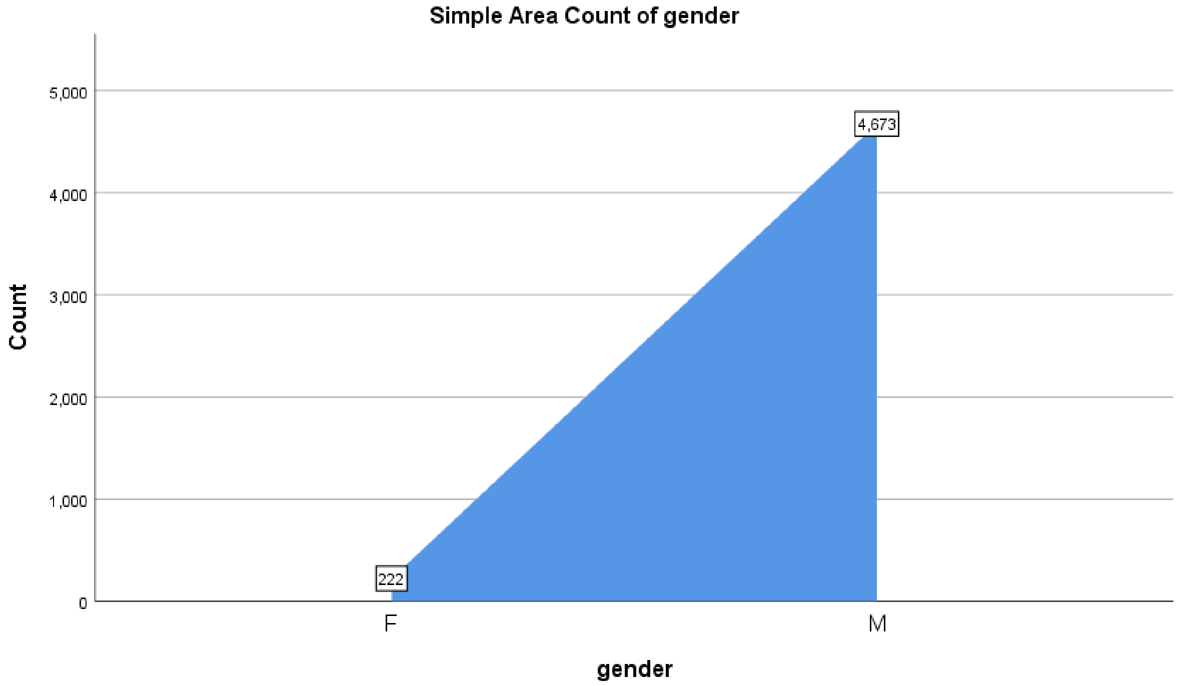
As our data has some numerical values it is also important to check if our data is normally distributed. This means that data close to the mean is common than data far from the mean. As we can see in “Fig. 11” that our data is moving from left to right in almost a straight line which is an indication that our data’s skewness is normally distributed.

Fig. 11



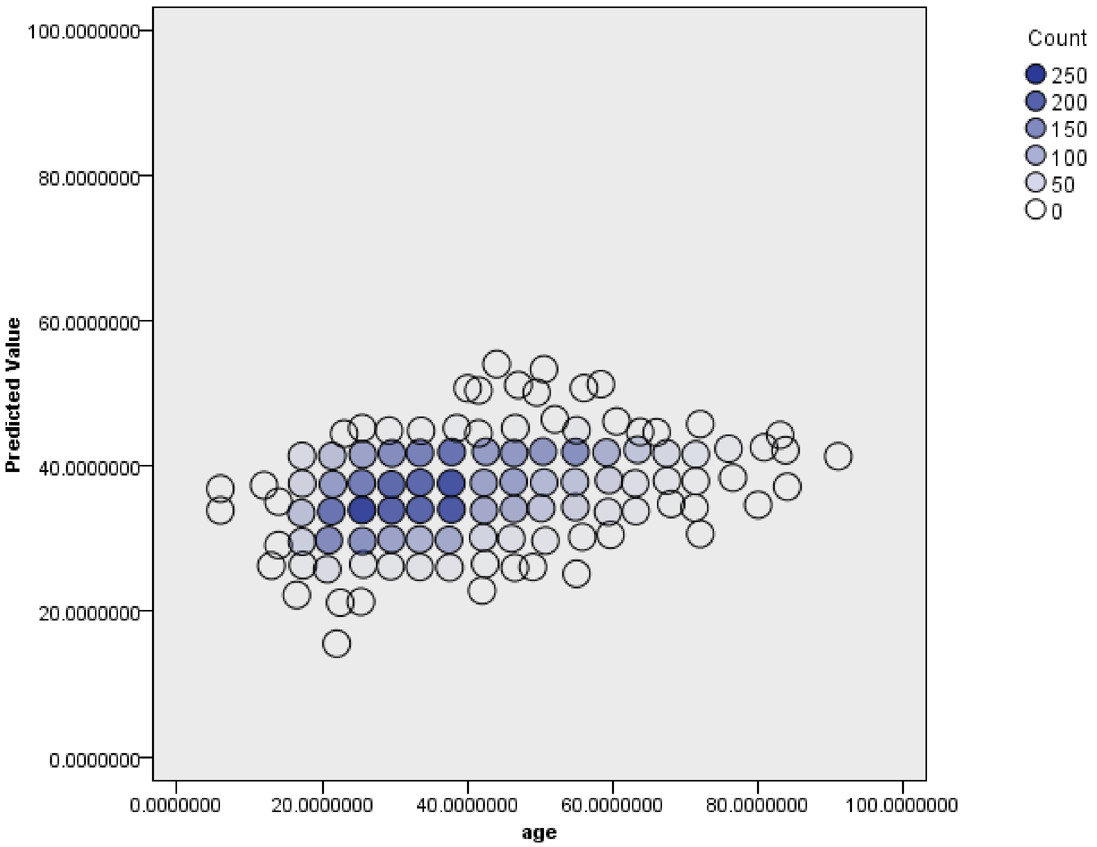
I will now be displaying some other visualisations that were developed from SPSS. Below in "Fig. 12", you can see how both of the represented genders compare regarding to being involved in police shootings. As observed, you can see that men appeared in the data 4,673 times and women appeared just 222 times. That means that men take up 95.45 percent of the shootings in the USA and women take up just 4.54 percent. At this time there is no clear reason as to why this is the case. In terms of how useful this data is, it is vital that men understand how they are being affected and to investigate further why this is the case so that they can try and reduce the overall figures.

Fig. 12



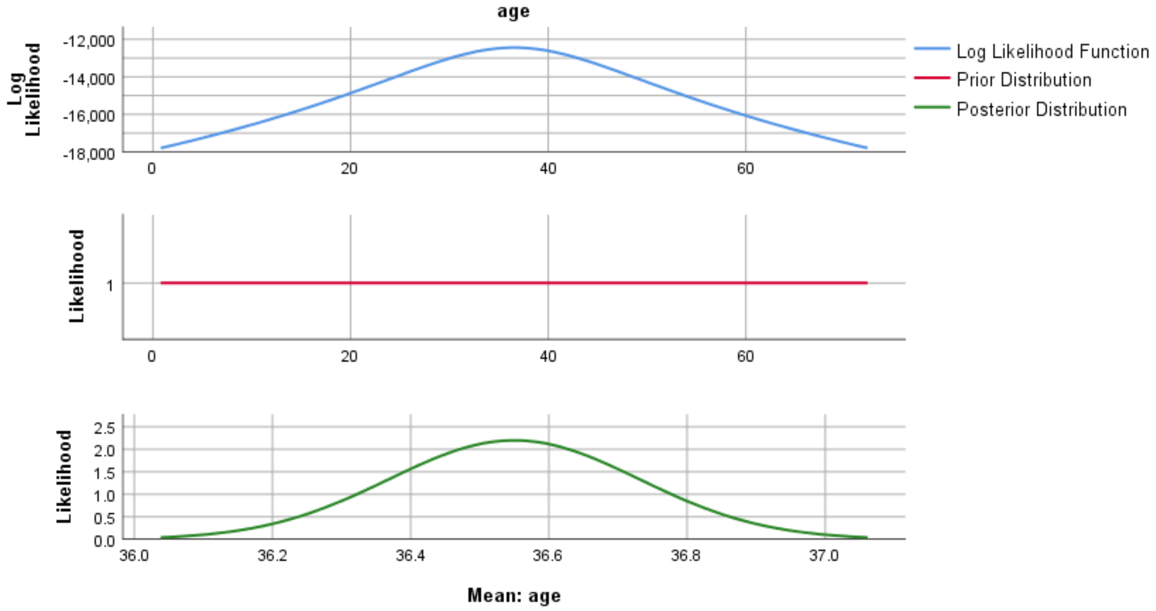
After looking at and discussing “Fig. 12” we understood that men were at much higher risk of being involved in a police shooting. Now looking at “Fig .13” we can see that most people that end up being caught up in a police shooting and end up dying are in the younger age groups. In particular, we can see all the way from around the age of 18 to 40 is at high risk of being involved in a police shooting when compared to other age groups. When we factor this in with that most of these people are male, we can realise that a lot of young men are becoming involved in behaviour that results in them being shot and killed by police.

Fig. 13



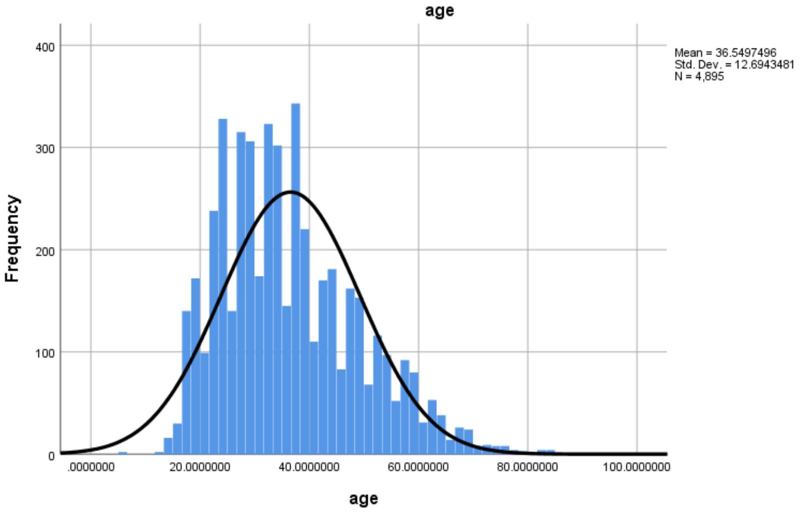
If we take a look at “Fig. 14” below we can see some Likelihood estimates performed by SPSS. If we pay attention to the blue and green lines, we can see that the likelihood of something being involved in a police shooting is its highest around the 35-year-old sage category and the mean age of the people involved sitting around 36.

Fig. 14



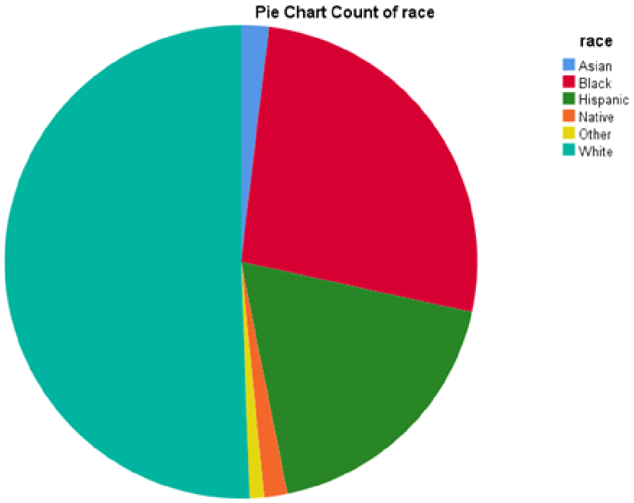
In “Fig. 15” we have the data points of “age” plotted onto a histogram; we can also see the curve of standard deviation plotted on top of the traditional histogram.

Fig. 15



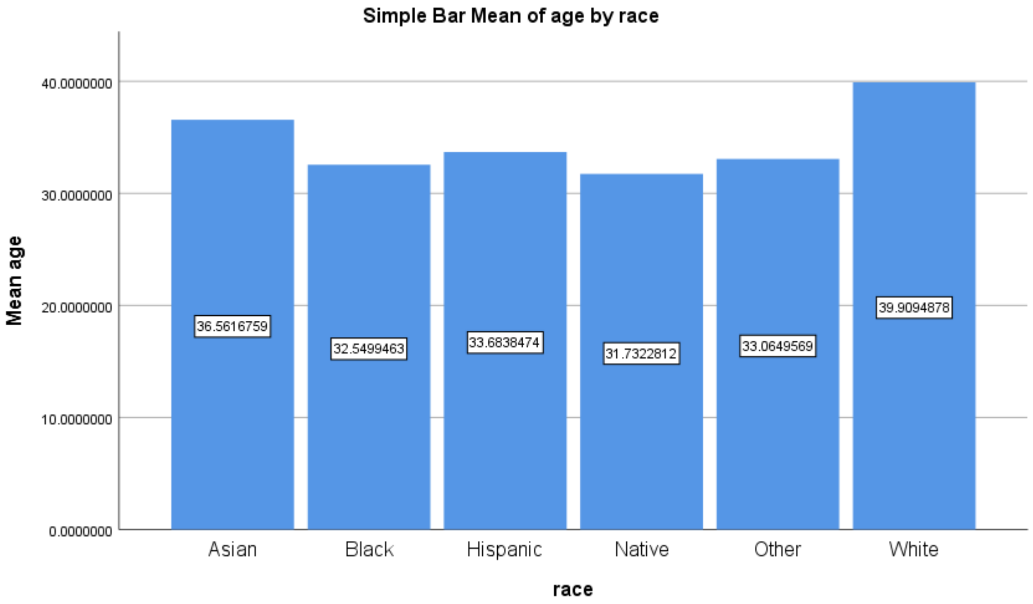
We are now going to have a look at how SPSS interpreted “Race” within the data set. Below in “Fig. 16” we can see a very simple pie chart that displays how each of the 6 represented races make up the 4895 shooting deaths. Now that we can see that the “White” race makes up around 50% of total deaths but this does not tell the whole story as we need to know how the “White” race takes up the general population. I go into further detail in other “Race” related results further into this report.

Fig. 16



Each “Race” within the data set is affected by police shootings in a number of different ways. If we look at how the age groups of “Race” is different between them in “Fig. 17” we can see quite a large difference. The “White” race has an average age of 39.9 when they were killed by police whereas the “Black”, “Hispanic” and “Native” races have a much lower average of between 31 to 33. The tells that the “Black”, “Hispanic” and “Native” races are more affected by police shootings at a younger age for at this time an unknown reason.

Fig. 17



R:

I will now be exploring the outputs of my R script. I will start with the explanation of some filters that I have ran and then progress onto the visualisations.

1: While looking through the data in excel I noticed a weapon type labelled "Toy Weapon". Logically I presumed that children use "Toy Weapons", so I then had a dig around the values associated with these types of weapon. Using the following lines of code, I could receive the youngest people in the data set that were shot and killed for having and displaying a "Toy Weapon" which the police misidentified as a real weapon.

- Shootings[Shootings$armed =="toy weapon" & Shootings$age == "13",]# Youngest Person killed for having a "Toy Weapon"
- Shootings[Shootings$armed =="toy weapon" & Shootings$age == "14",]# Next Youngest Person killed for having a "Toy Weapon"
- Shootings[Shootings$armed =="toy weapon" & Shootings$age == "15",]# Next Youngest Person killed for having a "Toy Weapon"
- Shootings[Shootings$armed =="toy weapon" & Shootings$age == "16",]# 3 Children killed aged 16 for "Toy Weapon"
- Shootings[Shootings$armed =="toy weapon" & Shootings$age == "17",]# 4 Children killed aged 17 for "Toy Weapon"

As you can see above from the results that 10 children were killed over 5 years for being "armed" with a "Toy Weapon". These types of shootings were a lot more common than I had anticipated, it is obvious from this result that either more training for police with children and toy guns is needed or some more rules for parents on what a safe and suitable toy is for a young child.

2: Another attribute that I wanted to explore in R was the "unarmed" option in the "arms category" column. By filtering the below queries I could find how many "Unarmed" people were shot and killed instead of being apprehended like they would have been in Ireland. I also selected that Body Camera footage of the killing was unavailable. I selected the largest five states in the US to best show the results of this filter:

- Shootings[Shootings$body_camera =="FALSE" & Shootings$arms_category == "Unarmed" & Shootings$state == "CA",] # 43 unarmed people killed where body cameras were unavailable, in California
- Shootings[Shootings$body_camera =="FALSE" & Shootings$arms_category == "Unarmed" & Shootings$state == "TX",] # 33 unarmed people killed where body cameras were unavailable, in Texas
- Shootings[Shootings$body_camera =="FALSE" & Shootings$arms_category == "Unarmed" & Shootings$state == "FL",] # 23 unarmed people killed where body cameras were unavailable, in Florida
- Shootings[Shootings$body_camera =="FALSE" & Shootings$arms_category == "Unarmed" & Shootings$state == "NY",] # 4 unarmed people killed where body cameras were unavailable, in New York
- Shootings[Shootings$body_camera =="FALSE" & Shootings$arms_category == "Unarmed" & Shootings$state == "IL",] # 4 unarmed people killed where body cameras were unavailable, in Illinois

As we can see from the results that a total of 107 people who were "Unarmed" were shot and killed over 5 years in the US with no body camera footage available after the fact.

3: In this filter, I am showing the number of people who showed “Signs of Mental Illness” with “Race” in the state of “CA”(Largest population) to see what the figures are. I did this analysis across 5 of the biggest states but have chosen California for this filter.

- Shootings[Shootings$race == "Black" & Shootings$signs_of_mental_illness =="TRUE" & Shootings$state == "CA",] #California, 22 Deaths of Black People with Mental Illness
- Shootings[Shootings$race == "White" & Shootings$signs_of_mental_illness =="TRUE" & Shootings$state == "CA",] #California, 43 Deaths of White People with Mental Illness
- Shootings[Shootings$race == "Asian" & Shootings$signs_of_mental_illness =="TRUE" & Shootings$state == "CA",] #California, 6 Deaths of Asian People with Mental Illness
- Shootings[Shootings$race == "Hispanic" & Shootings$signs_of_mental_illness =="TRUE" & Shootings$state == "CA",] #California, 55 Deaths of Hispanic People with Mental Illness
- Shootings[Shootings$race == "Native" & Shootings$signs_of_mental_illness =="TRUE" & Shootings$state == "CA",] #California, 2 Deaths of Native People with Mental Illness

As you can see from filter output that “White”, “Hispanic” and “Black” people were the most affected demographic in the state of California, but the population of California is 71% “White”. This means that people of the “Hispanic” or “Black” races are at a much higher risk of being involved in a fatal police shooting when being mentally ill when compared to the population percentage of the state of California.

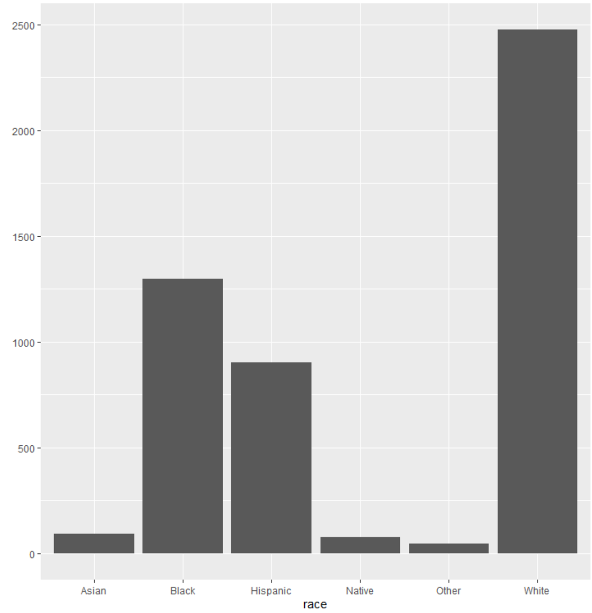
4: In this next filter I will be showing how many times “Sharp Objects” or “Piercing Objects” was used in each of the five largest states.

- Shootings[Shootings$arms_category =="Sharp objects" & Shootings$state =="CA",] 165 + 4 169
- Shootings[Shootings$arms_category =="Sharp objects" & Shootings$state =="TX",] 60 + 6 = 66
- Shootings[Shootings$arms_category =="Sharp objects" & Shootings$state =="FL",] 58 + 1 = 59
- Shootings[Shootings$arms_category =="Sharp objects" & Shootings$state =="NY",] 24 + 1 = 25
- Shootings[Shootings$arms_category =="Sharp objects" & Shootings$state =="IL",] 11 + 0 = 11
- Shootings[Shootings$arms_category =="Piercing objects" & Shootings$state =="CA",] 4
- Shootings[Shootings$arms_category =="Piercing objects" & Shootings$state =="TX",] 6
- Shootings[Shootings$arms_category =="Piercing objects" & Shootings$state =="FL",] 1
- Shootings[Shootings$arms_category =="Piercing objects" & Shootings$state =="NY",] 1
- Shootings[Shootings$arms_category =="Piercing objects" & Shootings$state =="IL",] 0

By looking at the returned figures we can see that 169 people were shot and killed for using a “Sharp Object” or “Piercing Object” in the state of California over the five years. Even though these kinds of figures seem low when compared to “Guns” it is still an important category to keep an eye on as Knives are much easier to procure than “Guns”.

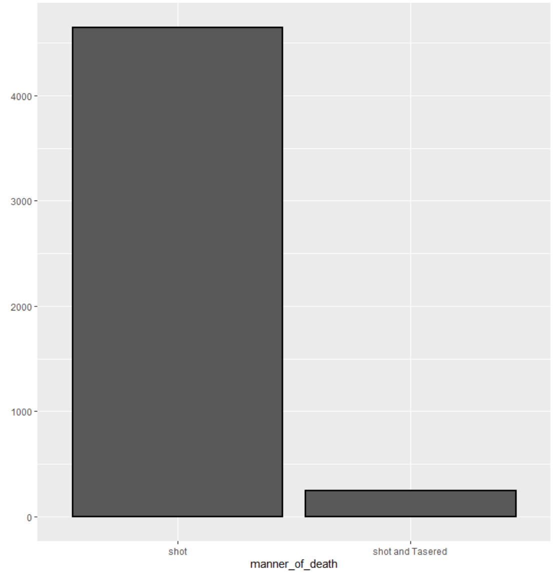
5: Starting with a basic but functional histogram we can see the six “Races” plotted against each other. As we can see the “White” race as before from our SPSS output is a large percentage of the data. This is purely because the “White” race takes up a very large percentage of the USA populations and is not accurate in terms of which “Race” is proportionately more affected. Looking at the graph, however, the “White” race is the leader for most people affected by police shootings when analysed from purely an analytical point of view.

Fig. 18



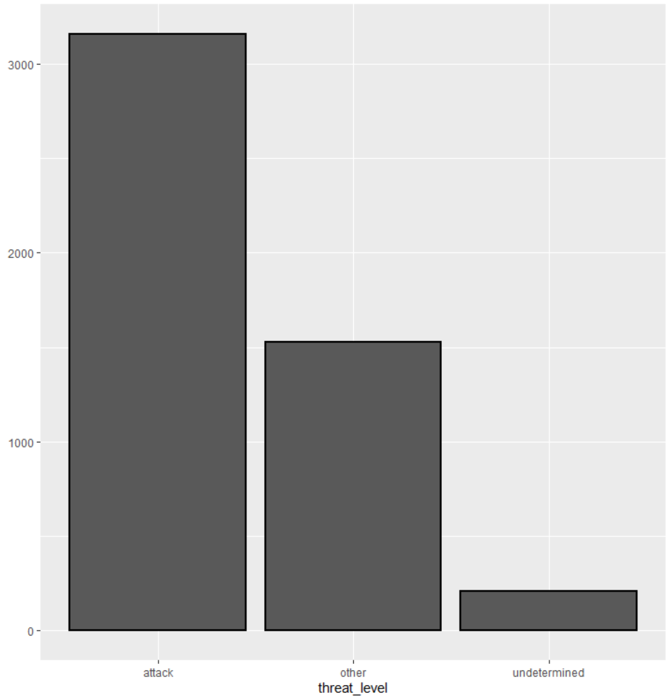
6: Now looking at the factor of how the people of the data set were killed. As we can see by looking at “Fig. 19” we can see that the majority of people were simply shot by the police officer. A small portion of the people within the data set was “shot and tasered”. This is likely due to the fact that officers try escalating to live fire in increments. This does not seem to be the case in the majority of the cases, however.

Fig. 19



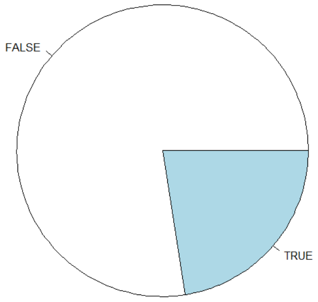
7: In “Fig. 20” we have how the people that were shot and killed were behaving before they were shot. A large portion of the people involved were “attacking” the police officer, but another large portion of the persons shot were labelled as “other” or “undetermined”. These very vague and meaningless titles simply do not provide enough detail into how the person was behaving before they were shot and killed.

Fig. 20



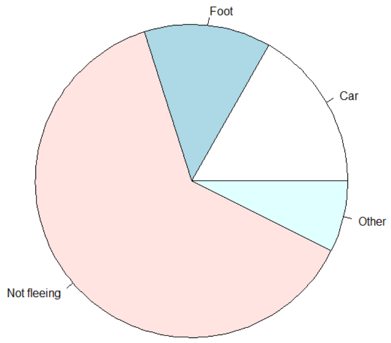
8: In “Fig. 21” I have displayed a pie plot that displays how Mental Illness has been a factor in the total number of deaths. As we can see from the pie plot that most incidents did not have this as a factor for the shooting that occurred. There does appear to a fair number of incidents in the data set that does have this a factor in the shooting taking place. As mental illness becomes talked about more and more in our societies, we will hopefully see police officers receive training on how to communicate properly with someone that appears to be struggling with their mental health.

Fig. 21



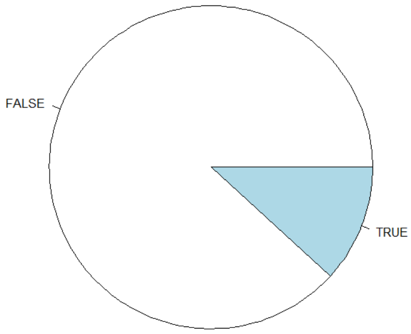
9: In “Fig. 22” we are showing how the persons that were shot and killed were fleeing or not and if they were by what medium. By looking at the pie chart we can see that most people were not fleeing so this did not factor into the shooting. We can see that a good portion of people did flee both on foot, in a car or other forms of transport. This could be a factor in the person being shot and killed as they tried to evade the police which could put others at risk.

Fig. 22



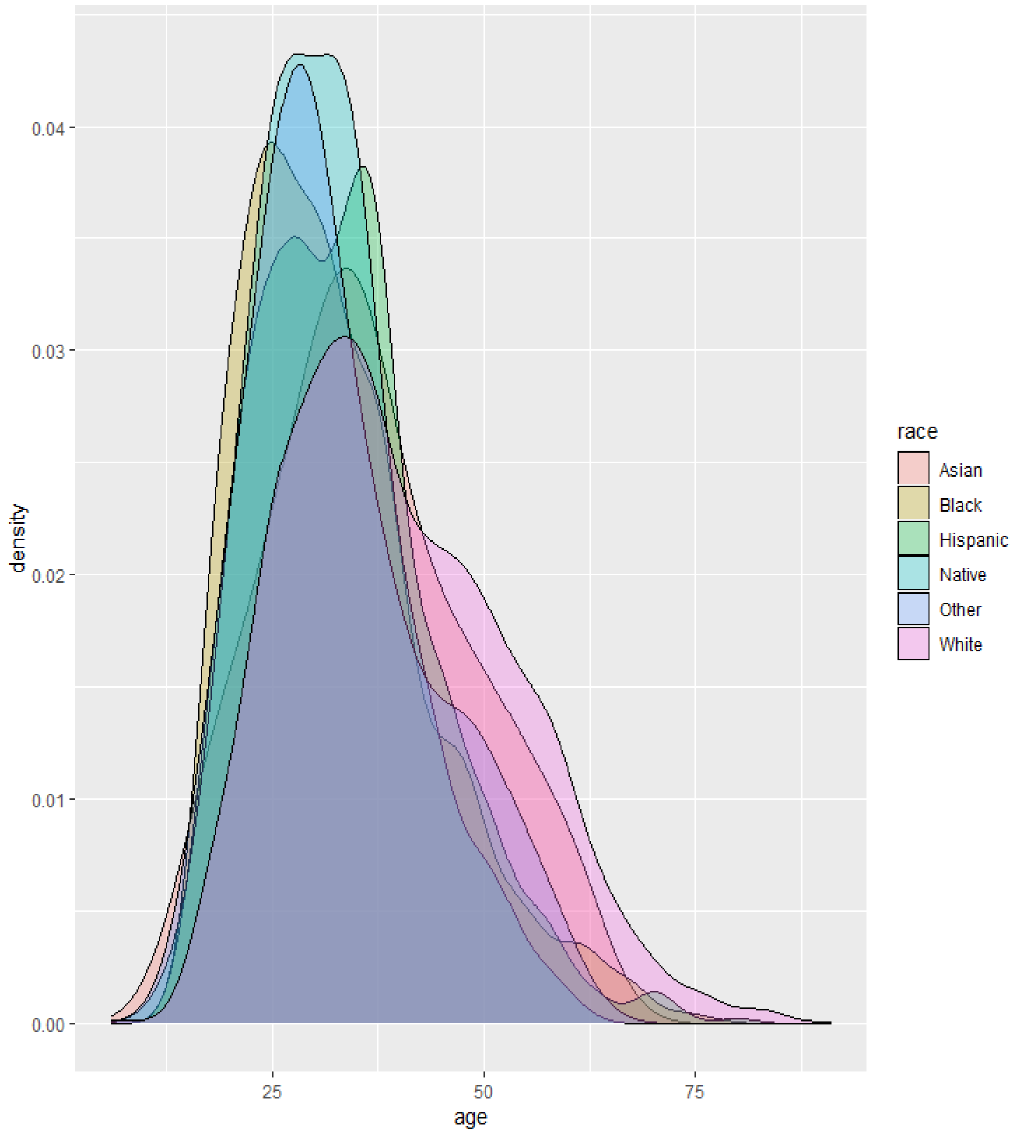
10: Below in “Fig. 23” we are showing if body cameras were worn when the person was shot and killed. This sort of information is vital to the investigation as if no video is available then we have to rely on the police officer’s version of events if no other witnesses were on the scene. Hopefully, as time goes on that this technology is implemented in more and more police departments.

Fig. 23



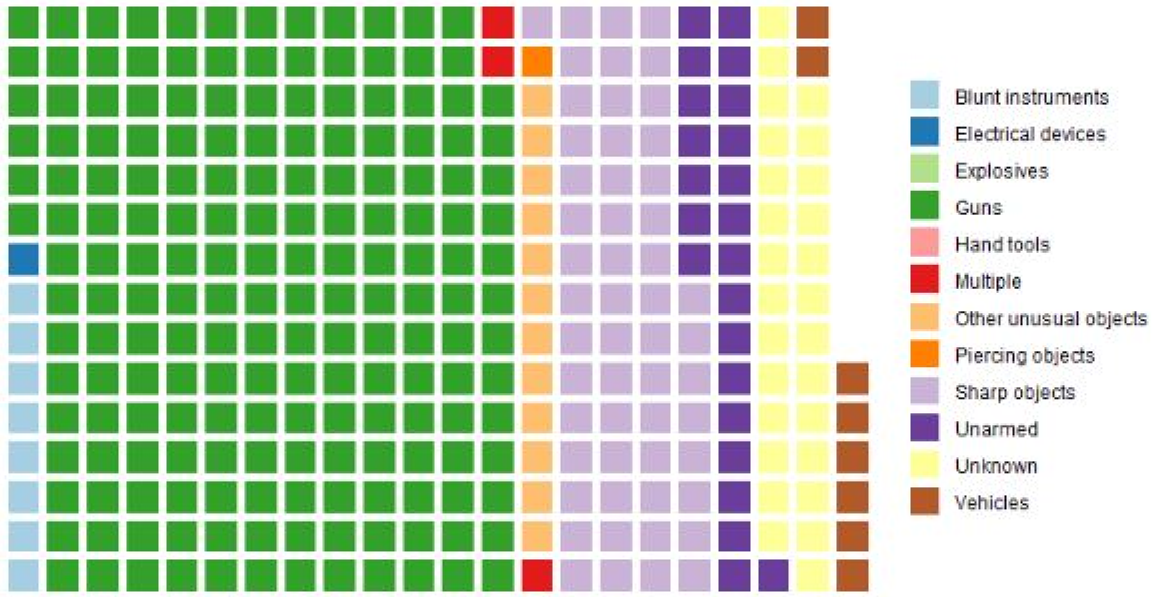
11: Below in “Fig. 24” we can see how each of the races that are included in the data set stack up against each other when compared on age. When looking at this type of plot we need to pay attention to how steep the curve is on both sides of the peak. By looking at each side and the peak of each “Race” we can see how where each “Race” ends up on the X-axis which represents age group. For example, we can see that the “Black” race reaches its peak very quickly and early on in the X-axis which indicates that the “Black” race has more people affected at a younger age than the other races in this plot.

Fig. 24



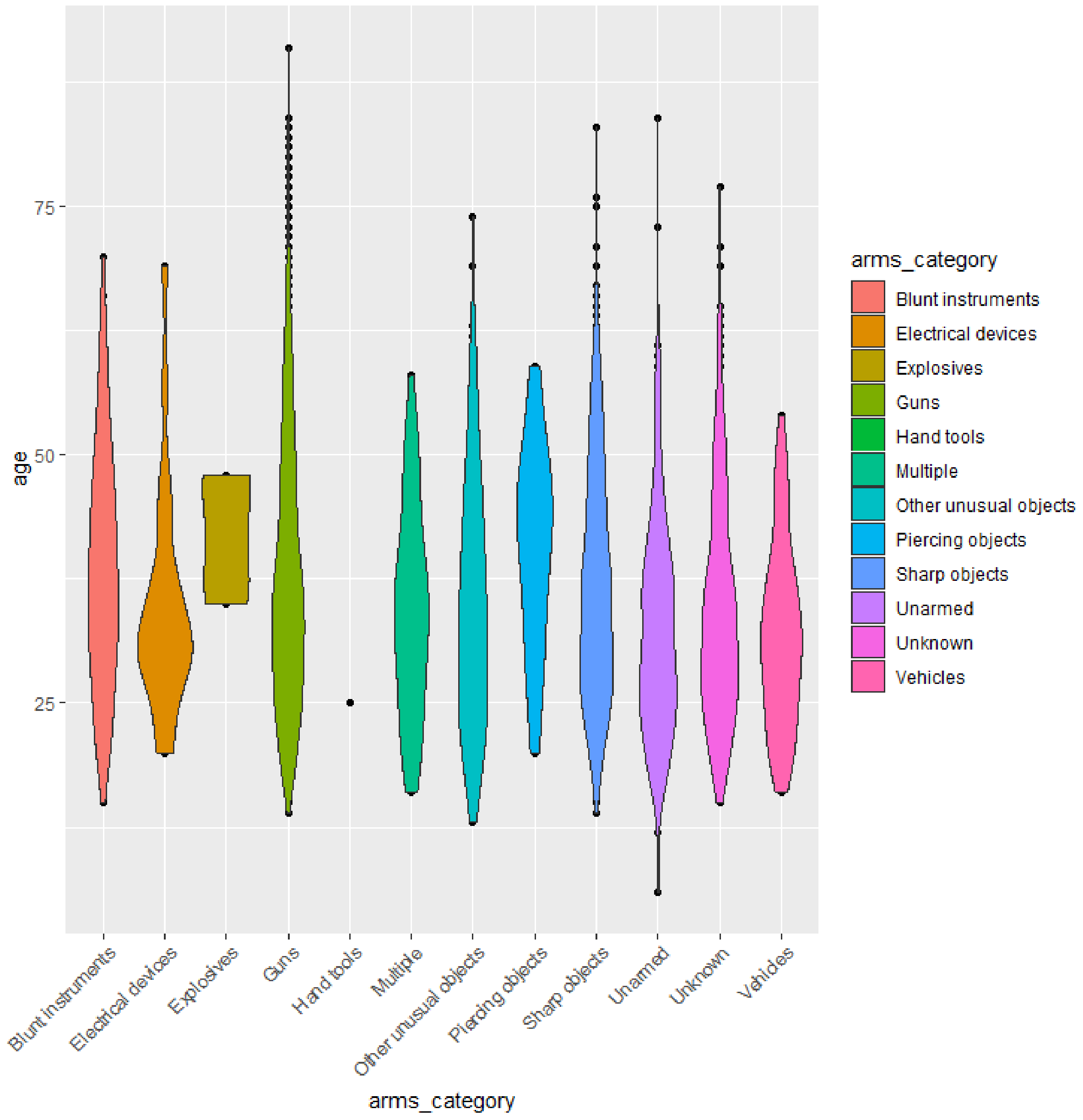
12: Below in “Fig. 25” we can see how different weapon types are represented in the data set. In this plot, each coloured cube is equal to 15 of that weapon type appearing in the data set. As we expected we can see that the green cube which represents “Guns” is appearing more than any other weapon type and almost the same as all other weapon types combined.

Fig. 25



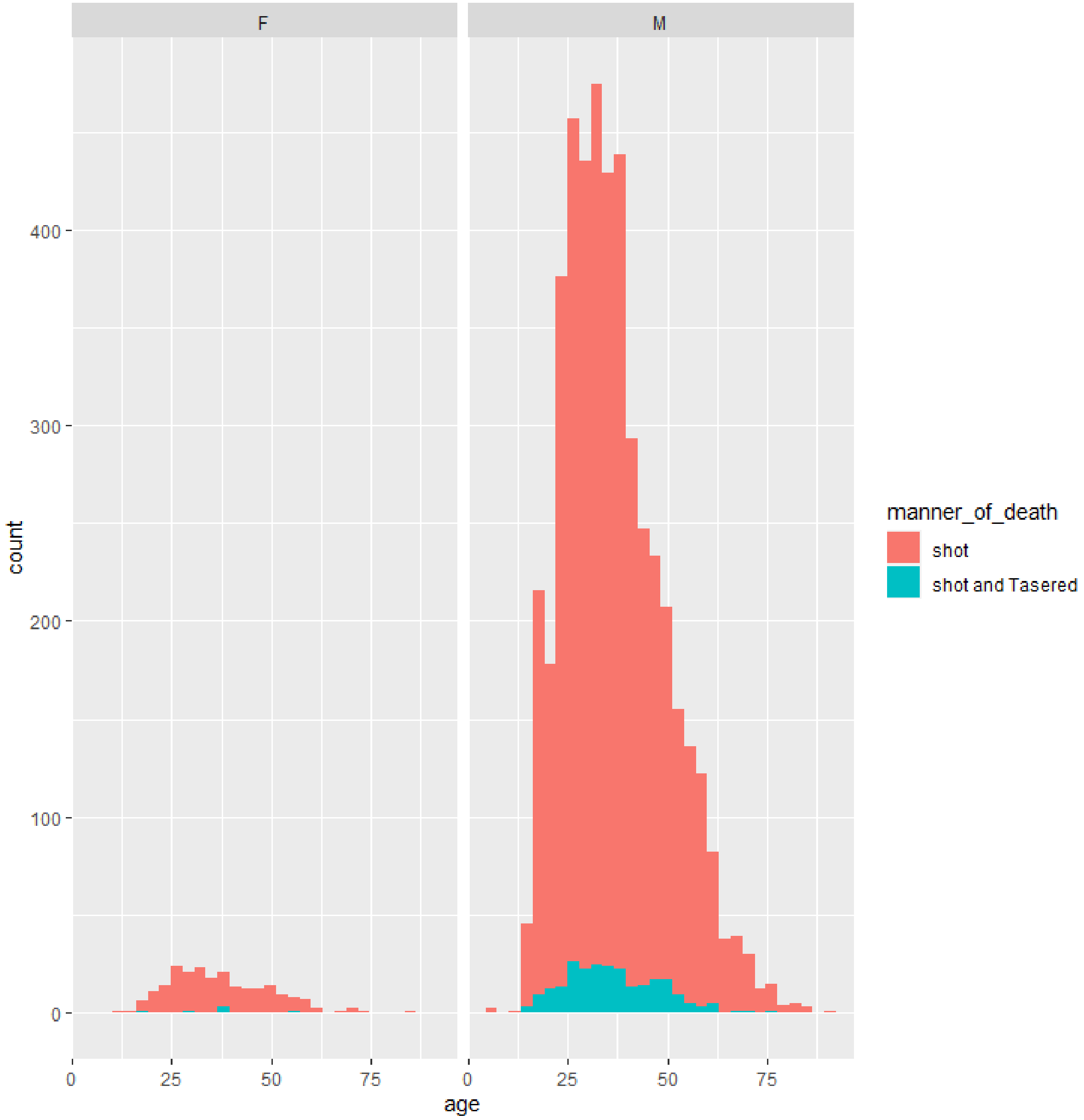
13: In “Fig. 26” we have the arms category potted against the age variable on a violin plot. The results end up showing us how different age groups used different types of weapons categories. For example, that “Guns” was used by a broad range of age groups, but “Explosives” were used almost exclusively by people in their late 30s to late 40s. The thickness of the violin plot at any given place on the plot gives us an idea of how many people at that age used that category of weapon.

Fig. 26



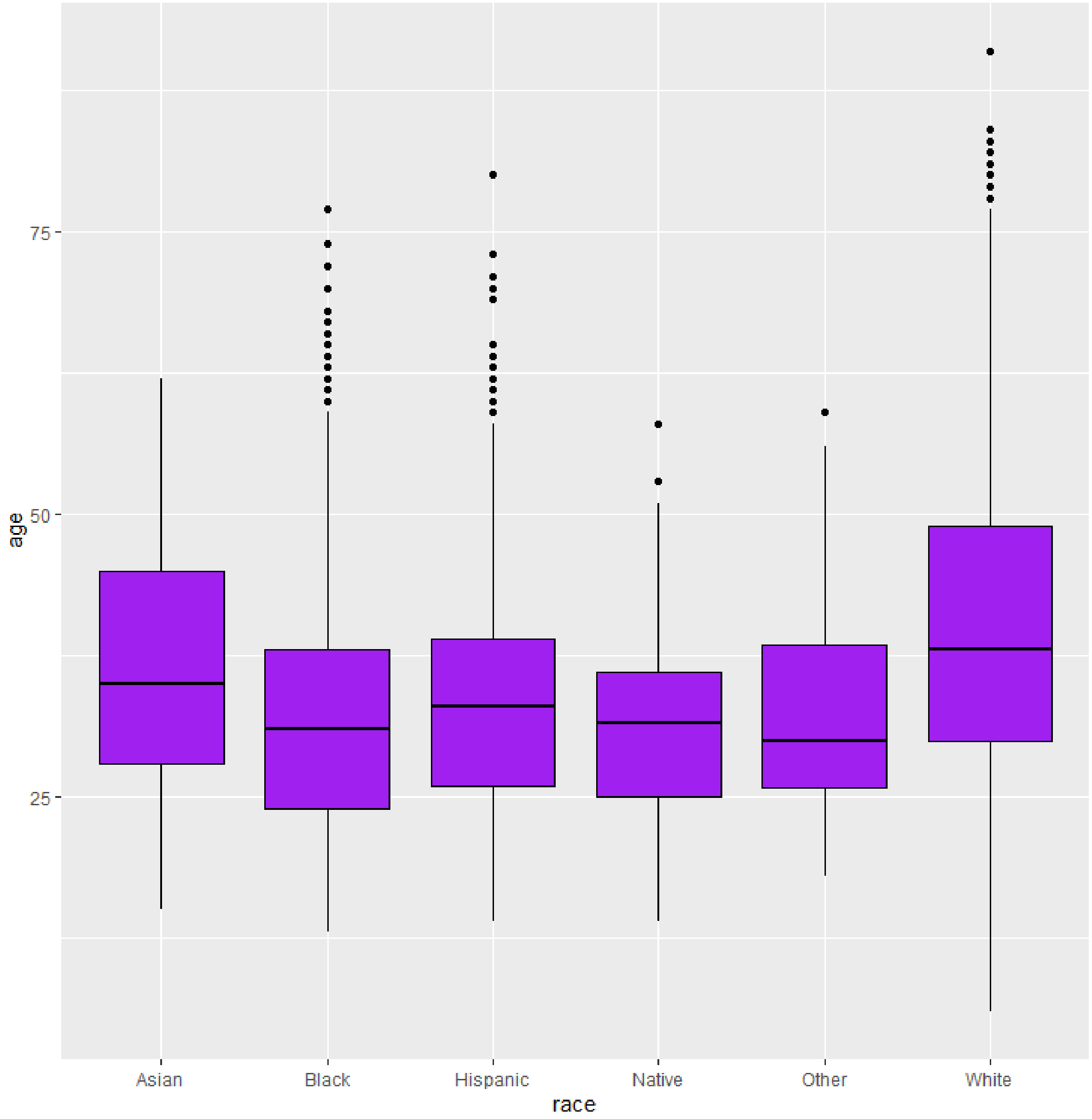
14: In “Fig. 27” I am displaying the difference between the two genders when we consider how the person died. As we can see from the plot that the male gender had far more incidents of being both “shot” and “shot and Tasered” when compared with the female gender.

Fig. 27



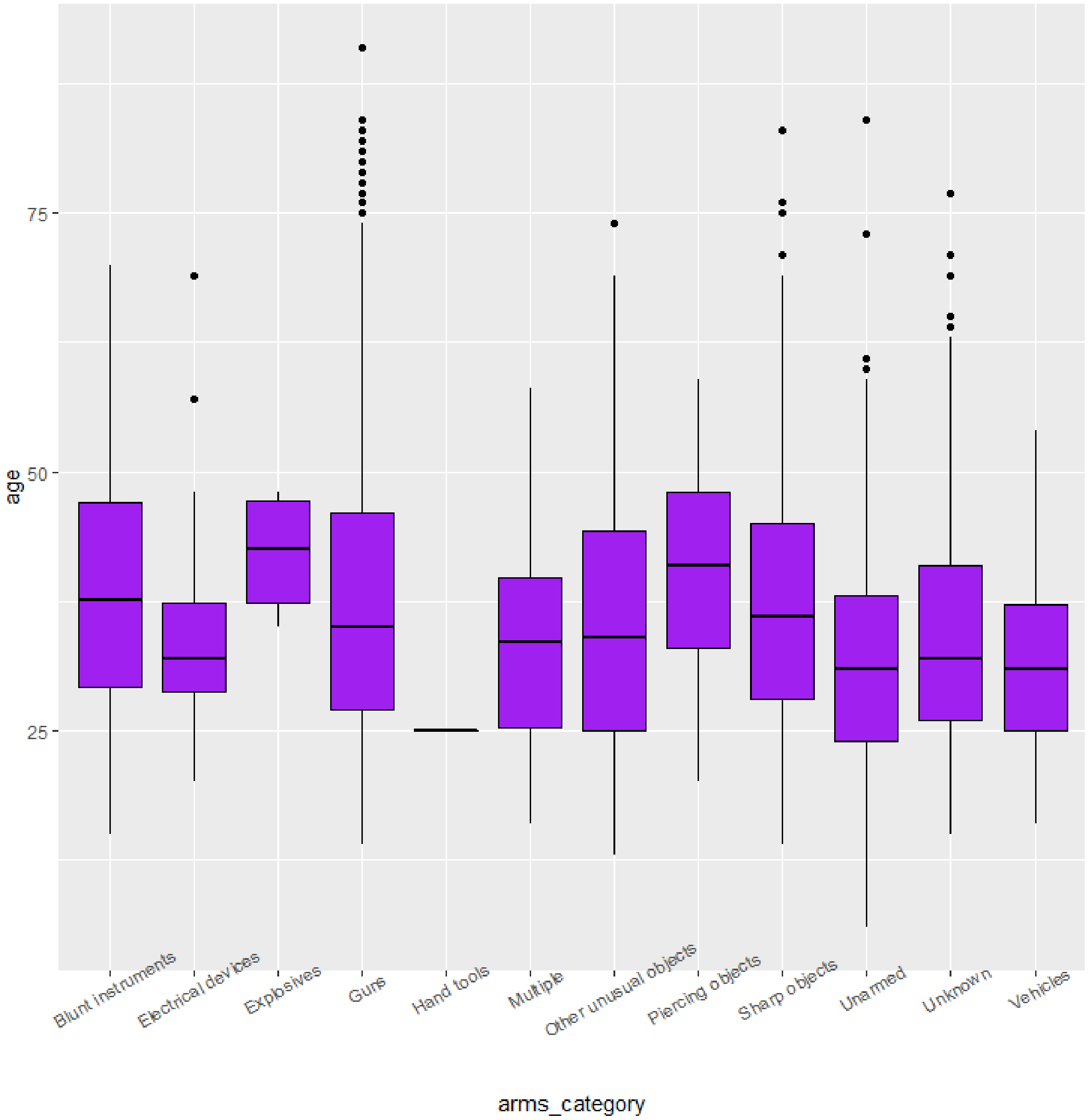
15: In “Fig. 28” we can see a boxplot which is displaying how age and race are factors within this data set. As we can see from the plot that as discussed before that the “White” race is affected more in its older age groups when compared to the “Black”, “Hispanic”, and “Native” races. This plot effectively shows how age is a factor across the different races of people.

Fig. 28



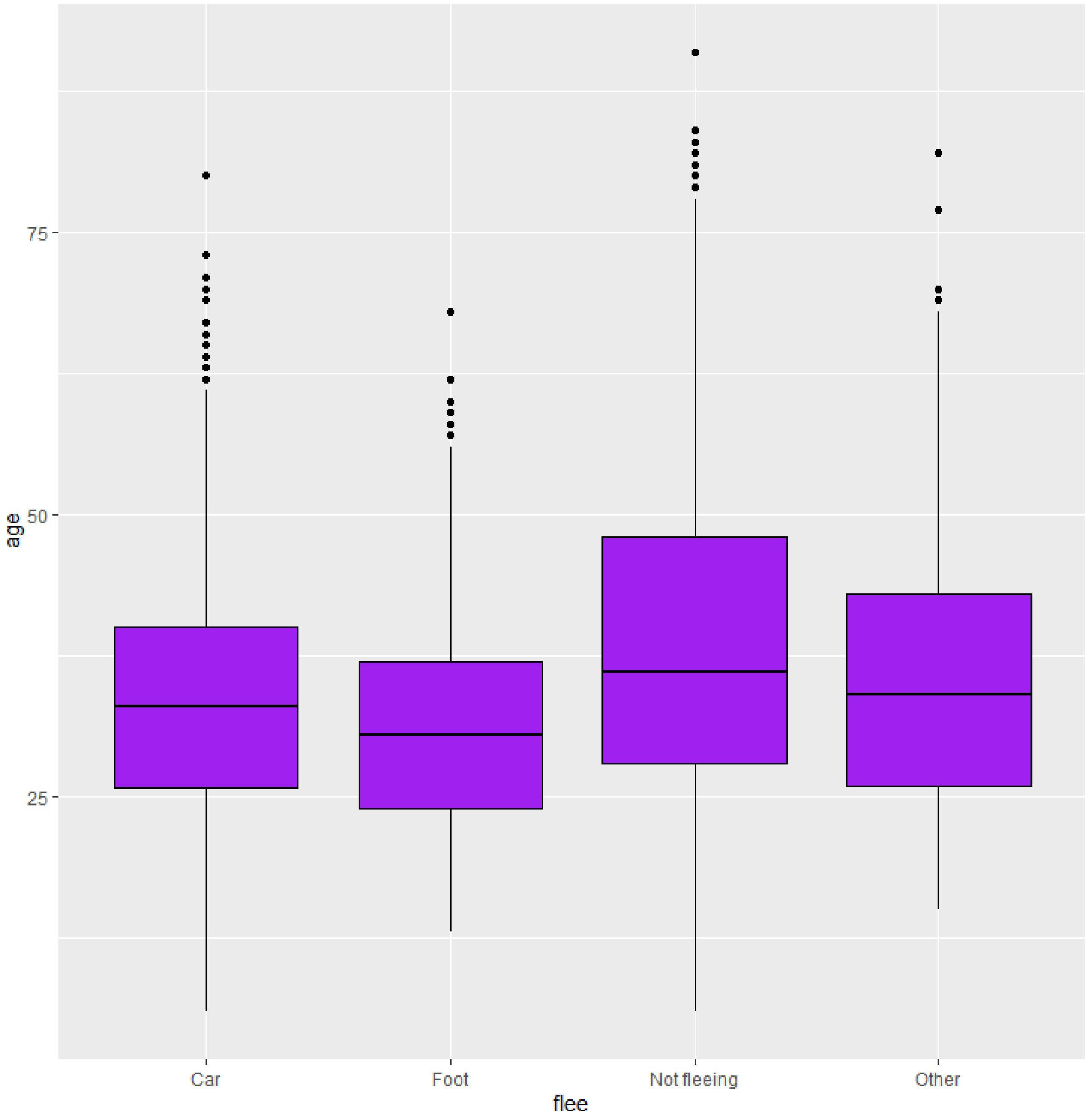
16: Showing a similar style of boxplot as shown above we can see how the different age groups are using different weapon types. This type of plot is also similar to the violin plot that was in a previous R output. Again, we can see how the "Gun" weapon category along with "Sharp Objects" and "Piercing Objects" are used by a wide range of age groups within the data set.

Fig. 29



17: In “Fig. 30” we can see how different age groups used different methods of “Flee” if any at all. As seen before most people did not “Flee” from the police. We can however get a better idea of what kinds of age groups used “Car” or “On Foot” to try and evade the police before ultimately being shot and killed.

Fig. 30



Excel

For the use of my next technology, I used Excel for its easy-to-use filter function to find and display outputs that show some very important statistics for this data set.

1: In this first filter I went through every “Age” group along with “Gender” to identify the worst affected age groups and gender within that bracket.

- Below Twenties: Male = 264, Female = 8, Total = 272
- Twenties: Male = 1307, Female = 63, Total = 1370
- Thirties: Male = 1480, Female = 69, Total = 1549
- Forties: Male = 841, Female = 42, Total = 883
- Fifties: Male = 525, Female = 31, Total =556
- Sixties: Male = 198, Female = 5, Total = 203
- Seventies: Male = 46, Female = 3, Total = 49
- Eighties: Male = 10, Female = 1, Total = 11
- Nineties: Male = 1, Female = 0, Total = 1

As we can see from the above bracket of age groups that the “Thirties” bracket proved to be the worst age group to be involved with a police shooting. This can be down to several factors such as financial stresses, lack of education or gang involvement but regardless of why this age group is the worst for both men and woman is that something needs to be done to try and bring these numbers down.

We can also see that there is a huge discrepancy between Male and Female throughout the entire filter of age groups. This is largely because men are more likely to enter gangs, own weapons, and break high tier laws that could result in police shooting the suspect involved in these crimes.

2: A second more advanced filter that I did in excel is as follows: The person was “Unarmed”, No “Signs of mental illness” was a factor, no “threat level” was determined, they were not “fleeing”, and no “body camera” footage is available.

This returned 15 victims that filled these requirements. I found this number to be quite worrying as it seems from the data that the victims in this filter were not doing anything wrong to warrant them being shot. This does not mean however that this is the case as important information about the situation may be unknown or unavailable.

3: This filter shows the number of killings in the top five largest states over each year, the average shootings over the 5 years are also included broken down to the months:

- 2015: CA = 182, FL = 61, IL = 21, NY = 18, TX = 99
- 2016: CA = 127, FL = 55, IL = 26, NY = 17, TX = 77
- 2017: CA = 143, FL = 56, IL = 20, NY = 15, TX = 60
- 2018: CA = 96, FL = 62, IL = 19, NY = 14, TX = 70
- 2019: CA = 111, FL = 57, IL = 10 , NY = 18, TX = 87
- 2020(As of June): CA = 42, FL = 33, IL = 3 , NY = 8, TX = 33

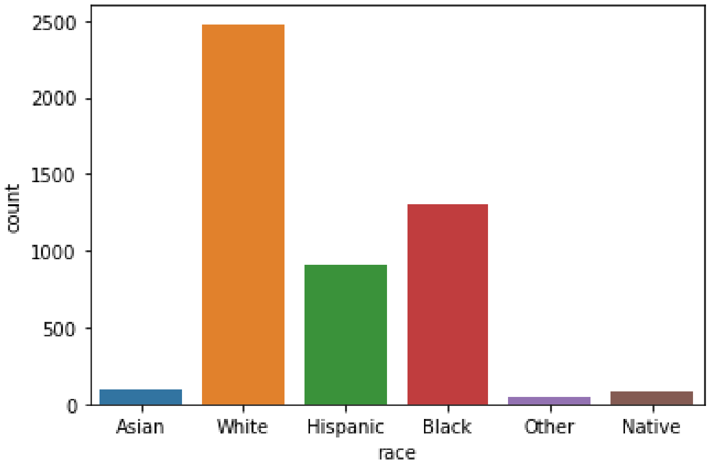
In terms of the most fatal time of year, the average per month is 347.75 with a minimum of 276 in May and a max of 423 in January.

Python

For my fourth technology, I will be displaying all of the relevant outputs produced in Python. I will be starting with some basic histograms and graphs before diving into some more interesting visualisations.

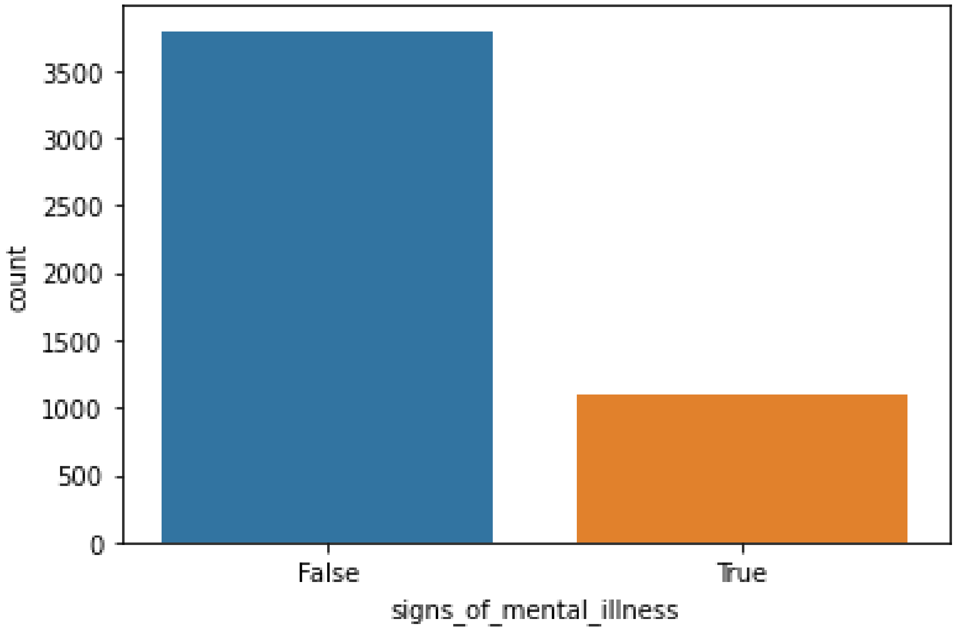
1: Below in "Fig. 31" we can see the different "Races" plotted on a histogram. This is achieved via a count plot. As discussed in previous sections we can see the most affected "Races" of police shootings. The main difference between this histogram and the ones previously discussed is the styling and the addition of figures on the count axis.

Fig. 31



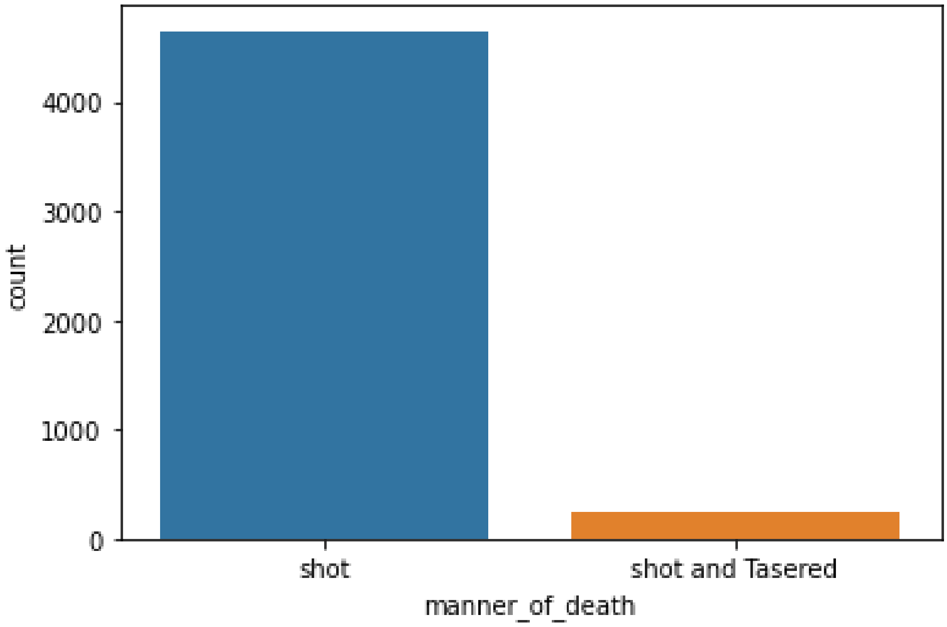
2: In a similar style plot as above, we can see the signs of "mental illness" being displayed as a count plot in "Fig. 32".

Fig. 32



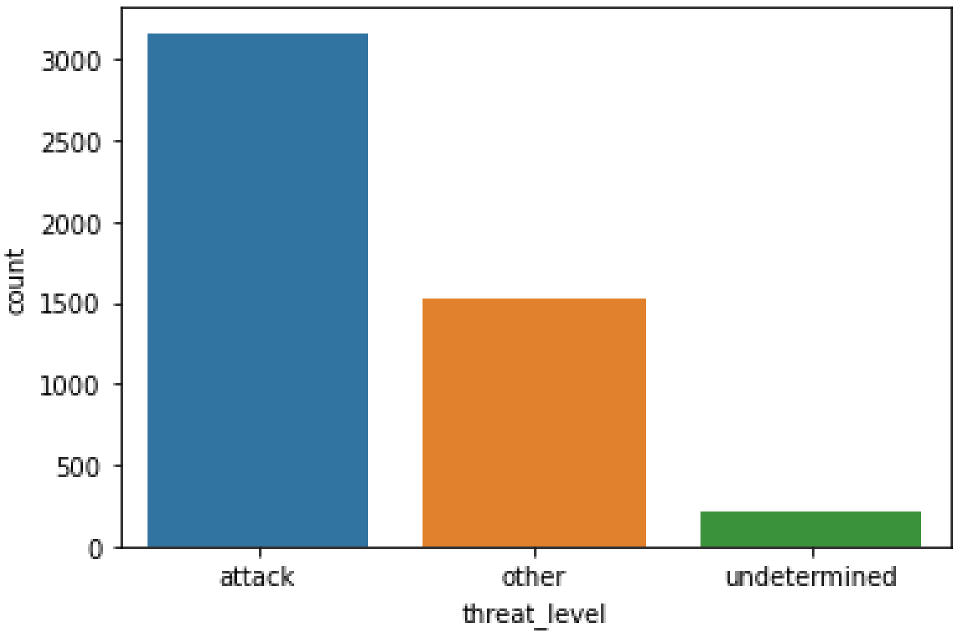
3. We can now see how the “manner of death” variable is displayed on a count plot below in “Fig. 33”.

Fig. 33



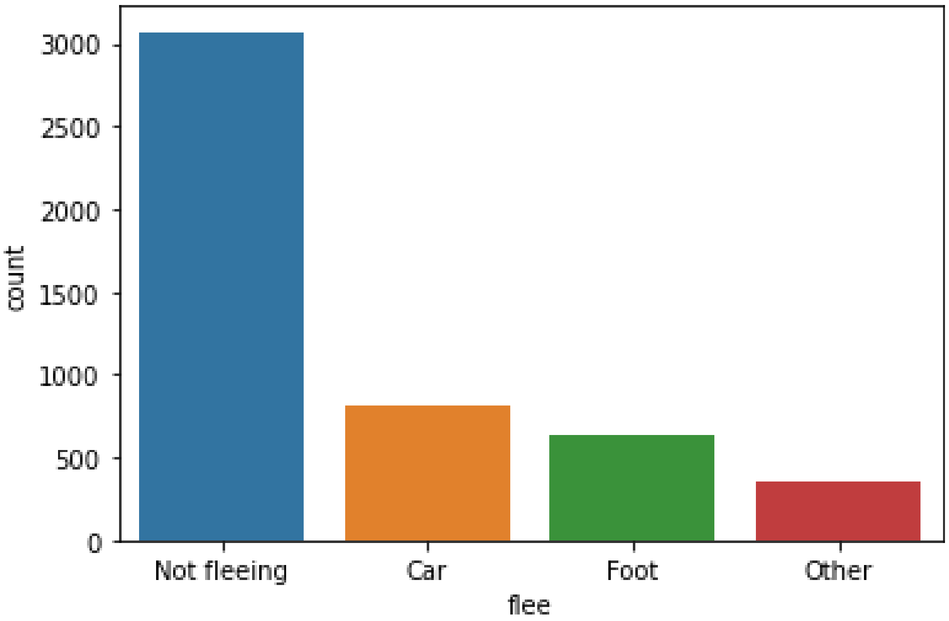
4. In “Fig. 34” we can now see how the “threat level” variable performed on the count plot visualisation.

Fig. 34



5. Below in “Fig. 35” we can see the final count plot being run on the “Flee” variable of the data.

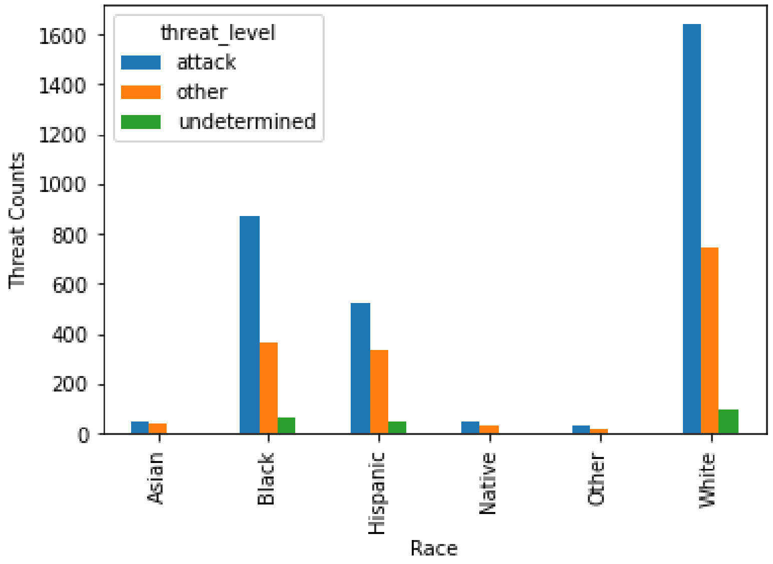
Fig. 35



6. In “Fig. 36” we are observing how each “Race” variable is performing when the “Threat Level” is added as a hue to the data. By observing the visualisation, we can see that the “White” and “Black” races have almost half the amount of “other” threat level when compared to the “attack” attribute.

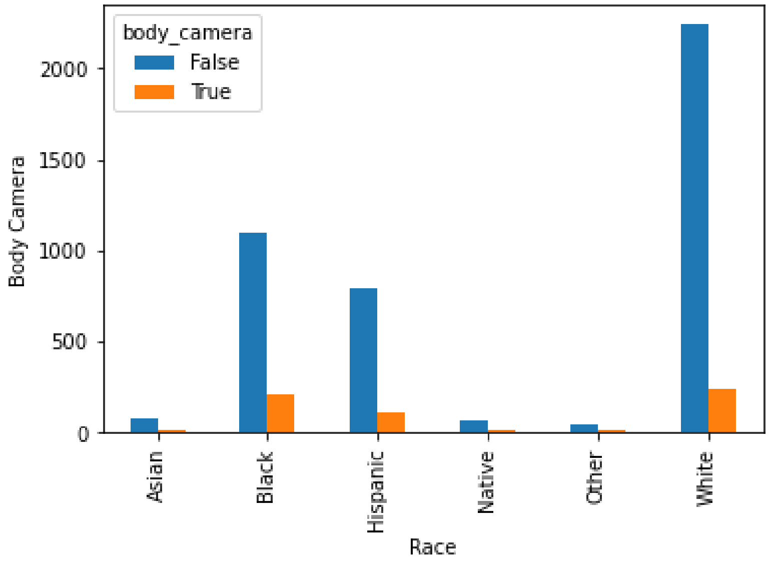
But when we look at the “Hispanic” race we can see that they did not appear as “Attacking” vs “Other” when compared to other races.

Fig. 36



7: Below in “Fig. 37” we have displayed how the use of “body cameras” is used between the different “races” available within the data set. As we can see by looking at the data in the visualisation that both the “Black” and “White” races have similar amounts of body cameras being used when the shooting. This seems to indicate that the “White” race seems to be more affected when it comes to police not using body cameras.

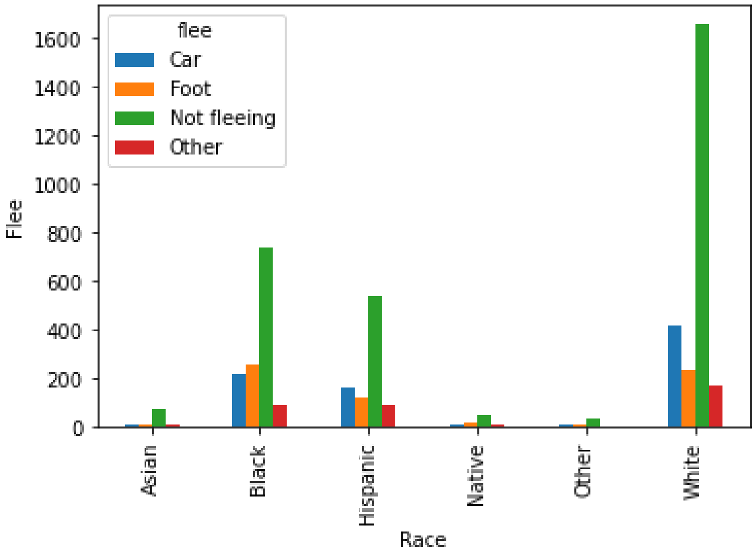
Fig. 37



8: In “Fig. 38” we have the “Race” variable being displayed along with the “Flee” variable to show how different “Races” of people tried to flee the scene. We can notice that the “Black” and

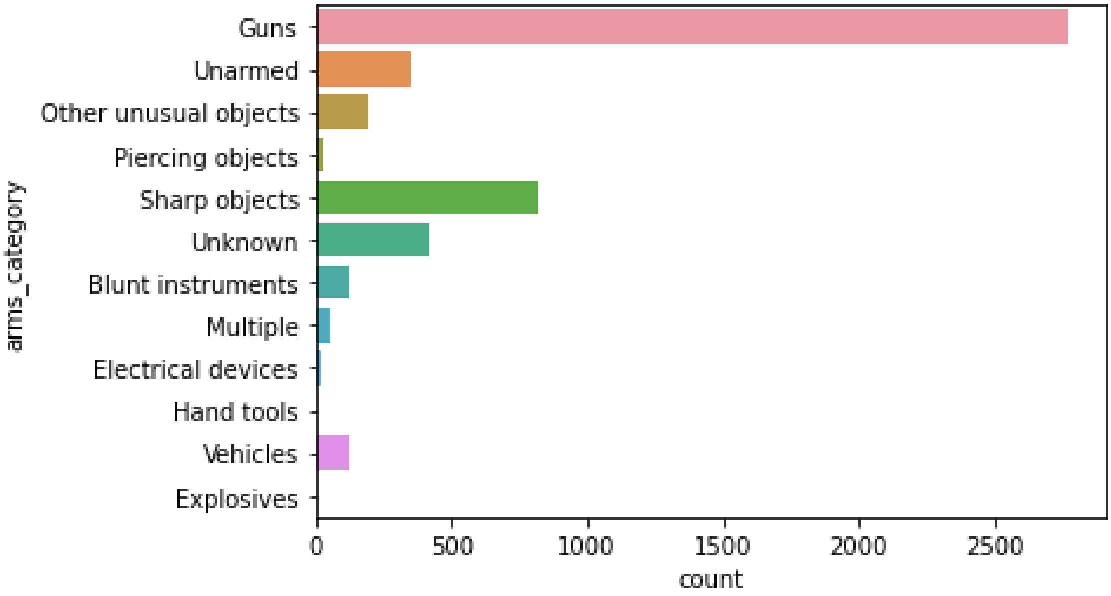
“Hispanic” races seemed to flee more in terms of their proportion to the “White” race. We would expect to see more felling numbers in the “White” race if the figures were even across the board.

Fig. 38



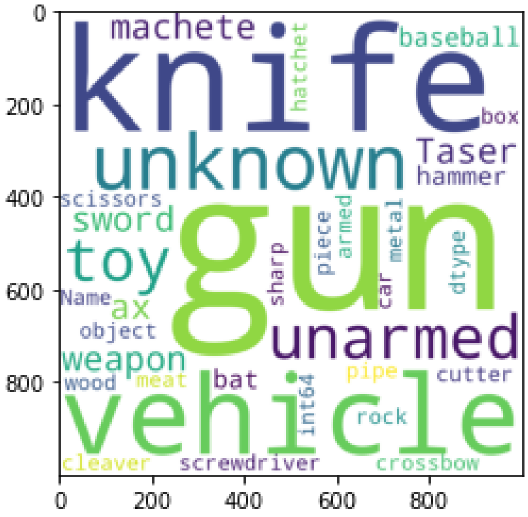
9: Below in “Fig. 39” we can see another breakdown of how the “arms category” is broken down across the data set. Again, seeing that “Guns”, “Unarmed” and “Sharp Objects” being some of the most common categories being used.

Fig. 39



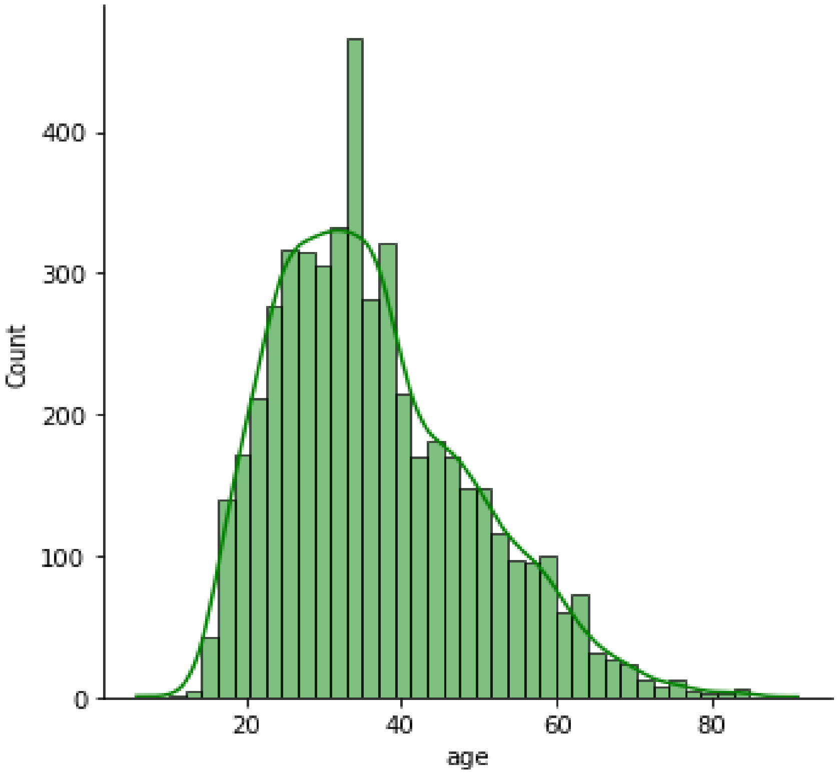
10: In “Fig. 40” we can observe a word cloud that has been built around the “arms category” variable. This shows not just the main attributes of that variable but also some other weapons that we had not noticed in other visualisations.

Fig. 40



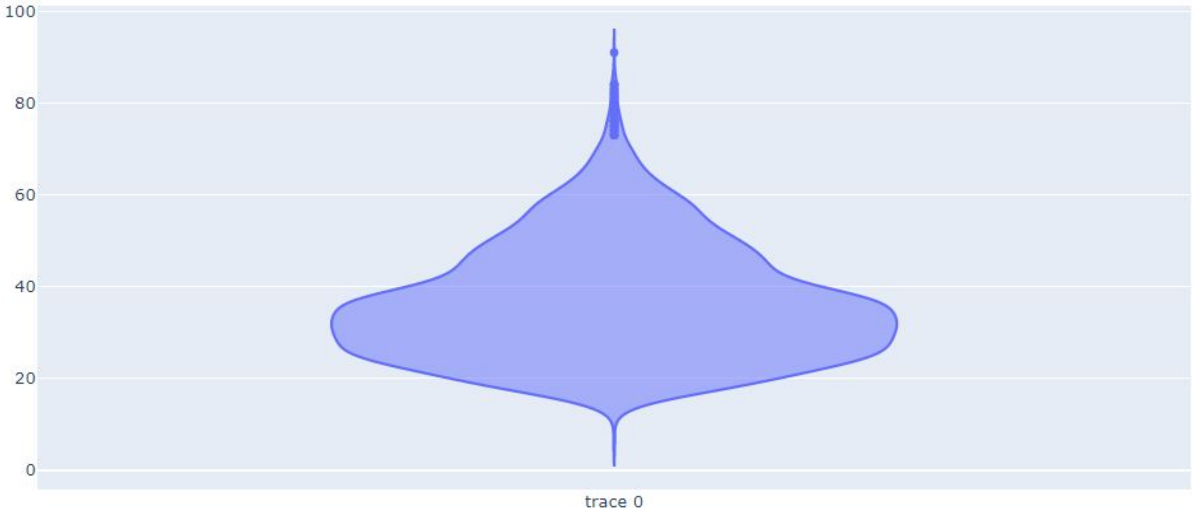
11: In “Fig. 41” we are observing how the age of people affected by police shootings. We also have the distribution of the data displayed in a solid line going from left to right.

Fig. 41



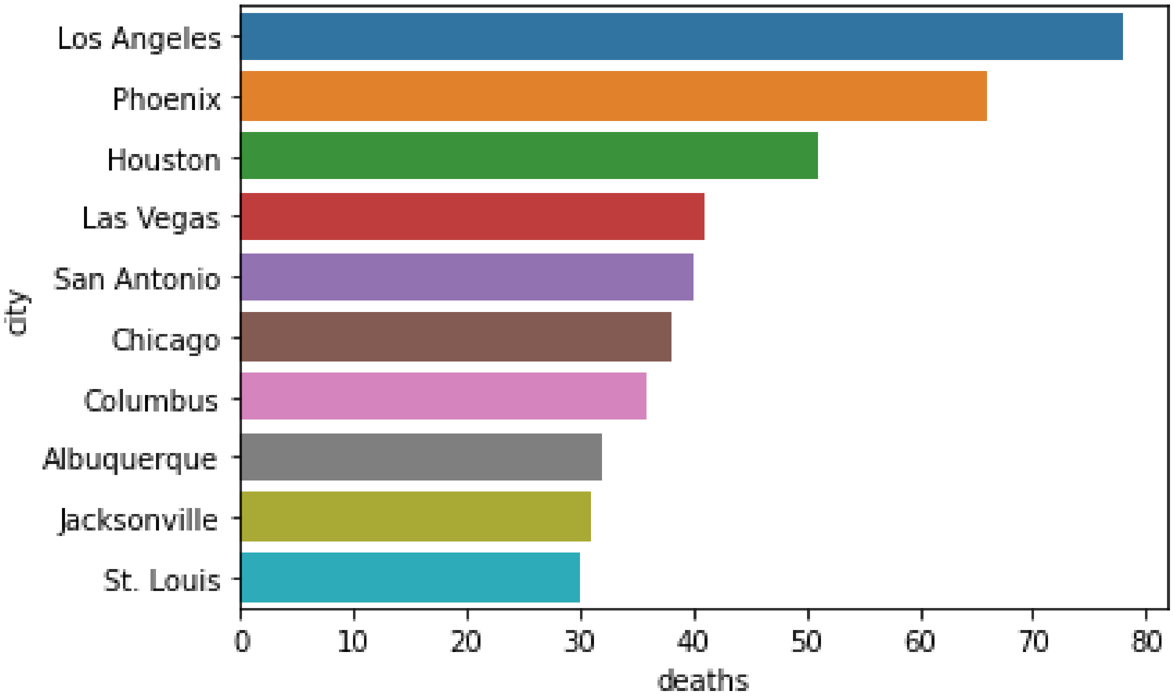
12: Below in “Fig. 42” we can see a different style of graph to visualise how age is divided across the data set. As we observed and discussed before, we can see that the ages of 20 to late ’30s are the most affected age group of the data.

Fig. 42



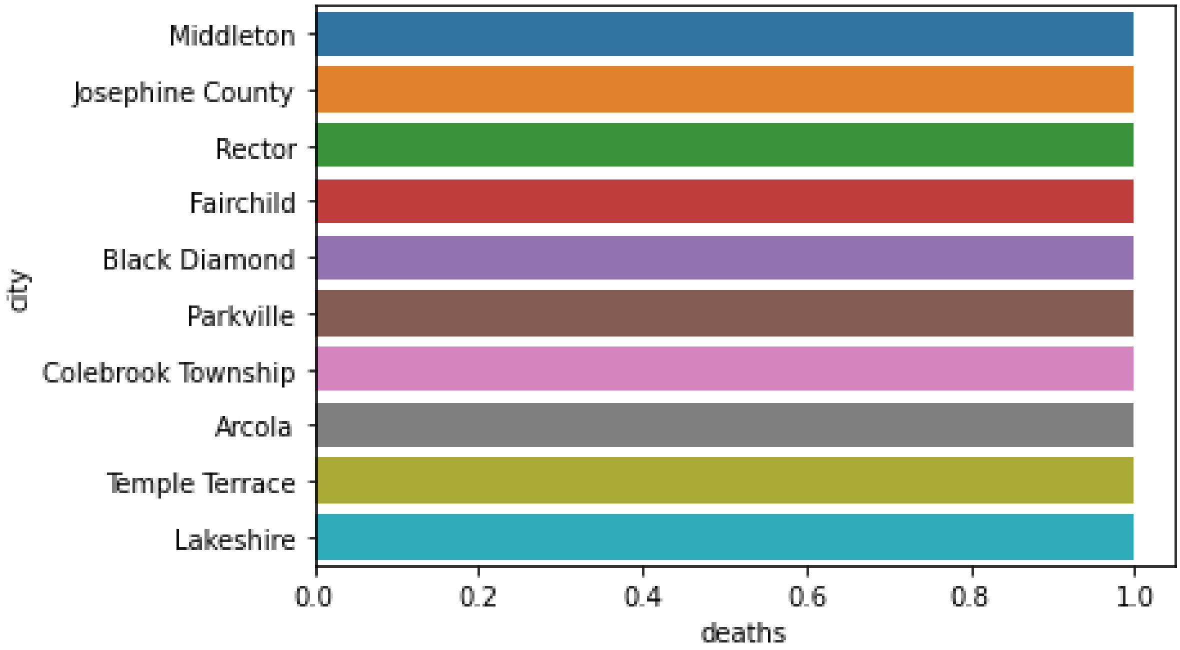
13: In “Fig. 43” we are now displaying the top ten worst cities in the USA for the number of fatal shootings by police. We can see Los Angeles topping out here with quite a lead over other cities such as Phoenix and Houston.

Fig. 43



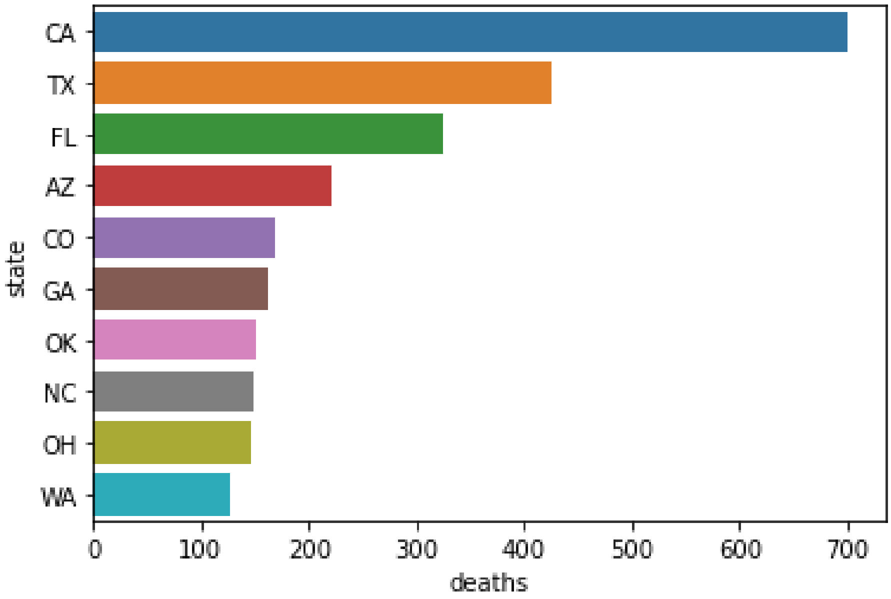
14: In “Fig. 44” we are now visualising some cities in the USA that have some of the least shootings by police. All are under one shooting over the last 5 years on average.

Fig. 44



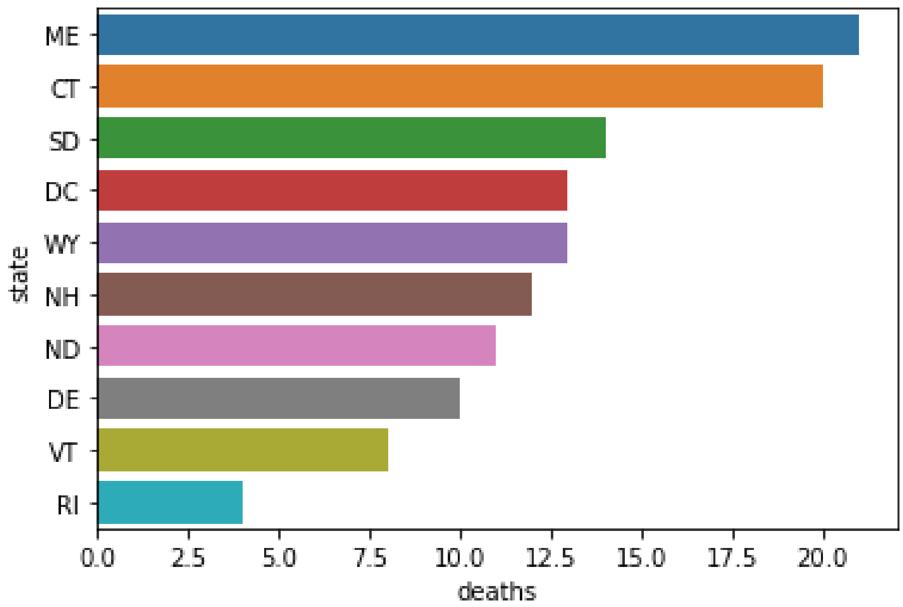
15: In “Fig. 45” we are visualising the top ten worst states in the USA for police shootings. We can see that California takes the top place here followed by Texas and Florida.

Fig. 45



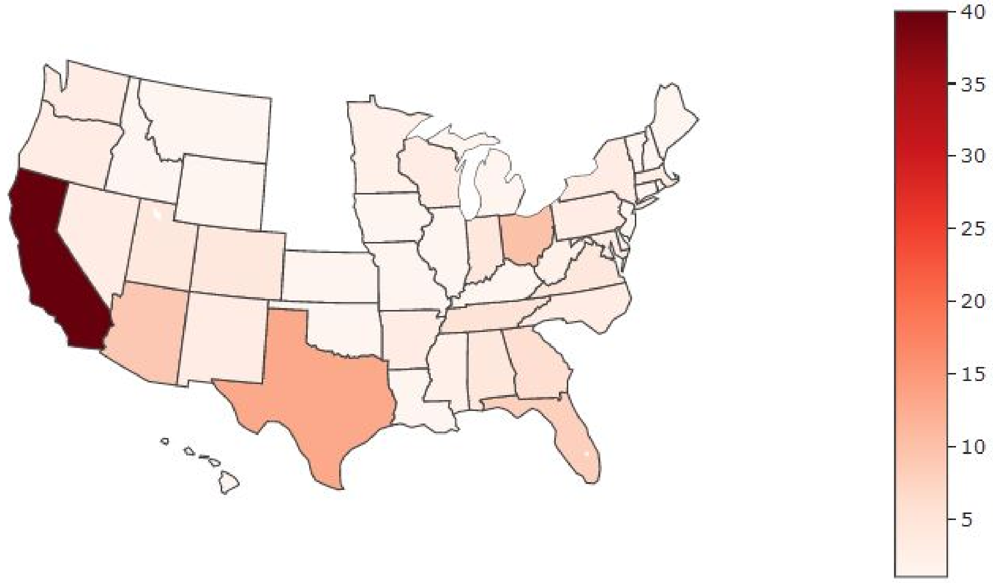
16: Now we are visualising the top 10 states with the least amount police shootings. We can see that Rhode Island(RI) is the best state as it only had an average of around 4 shootings over the last 5 years.

Fig. 46



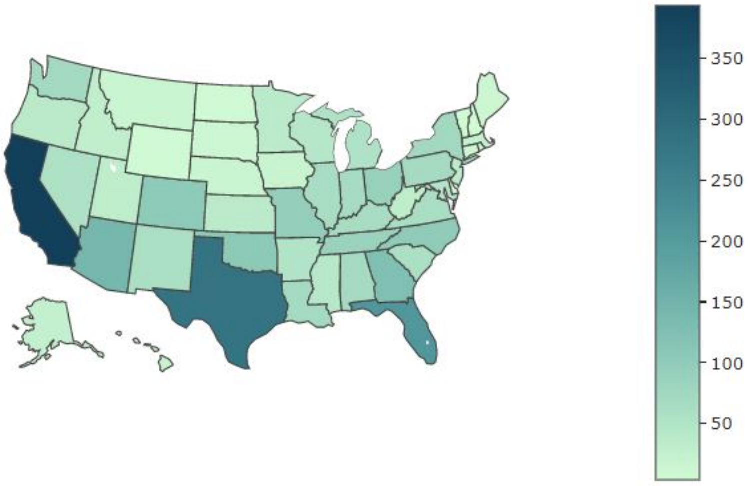
17: In "Fig. 47" I am visualising how "Toy Weapons" have been a factor in the shooting of an individual. I am displaying this visualisation on a Choropleth US State Heatmap. As per the heat map, we can see that California, Texas, Florida, and a handful of other states have had run-ins with "Toy Weapon" situations that ended up with the person holding it being shot and killed. This is clearly a nationwide issue as only a handful of states had no occurrences of this issue.

Fig. 47



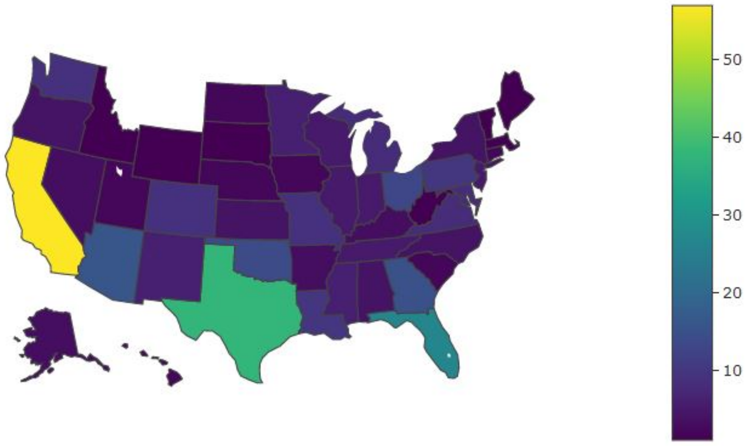
18: Below "Fig. 48" we have another heat map of the USA; this time it is visualising the states with the greatest number of times a police officer had encountered "Attacking" behaviour. As expected, we see this being high in the states with high populations but also as an issue in smaller states. This overall aggression towards the police forces is something that seems to be a regular occurrence in the people of the US.

Fig. 48



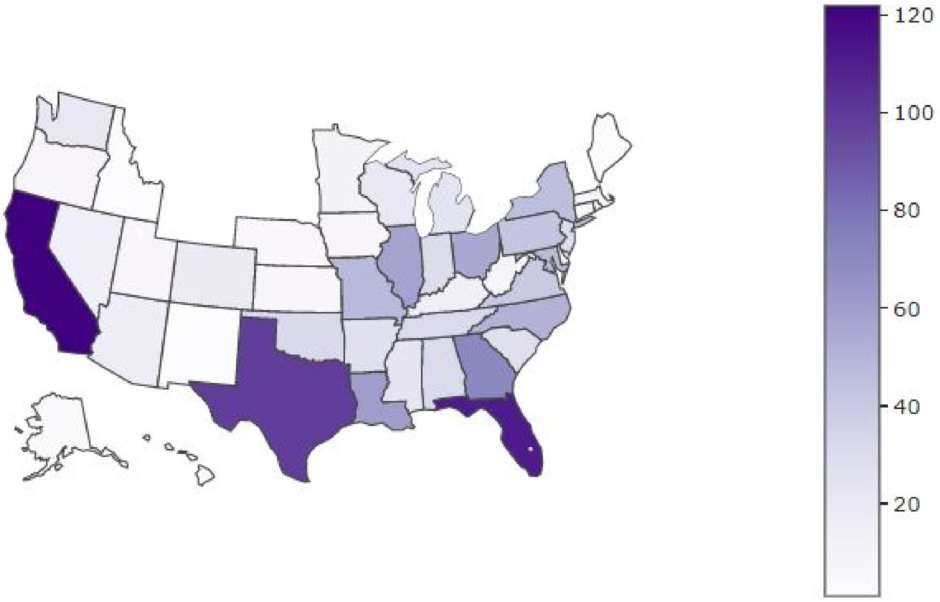
19: In the below heat map "Fig. 49" of the USA we have visualised the number of people that were killed even when being unarmed. The reason for the shooting may have been due to other reasons but I found this hard to believe that the person could not have been subdued without them being shot. This is definitely a statistic that all states need to work to bring down.

Fig. 49



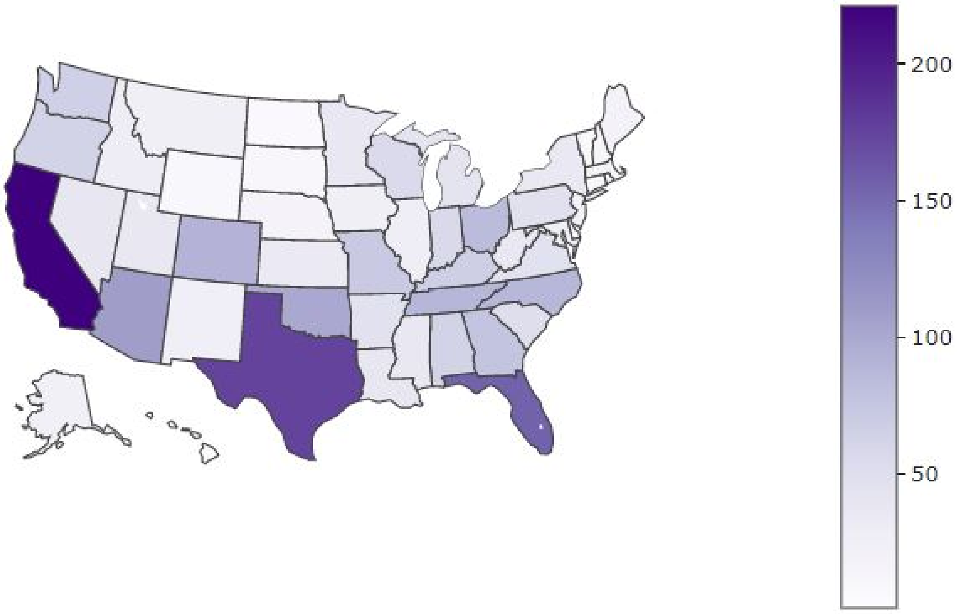
20: In “Fig. 50” of the USA we are now visualising how the “Black” race has been affected by police shootings across the entire country. We can see that the “Black” race is affected across most states of the USA and affected most in the states of California, Texas, and Florida. They also affected more than I had realised in the smaller eastern states.

Fig. 50



21: In “Fig. 51” of the USA we are now visualising how the “White” race has been affected by police shootings across the entire country. We can see that the “White” race is affected almost everywhere in the USA. There does not seem to be any comparisons to be made between states like I could for the “Black” race.

Fig. 51

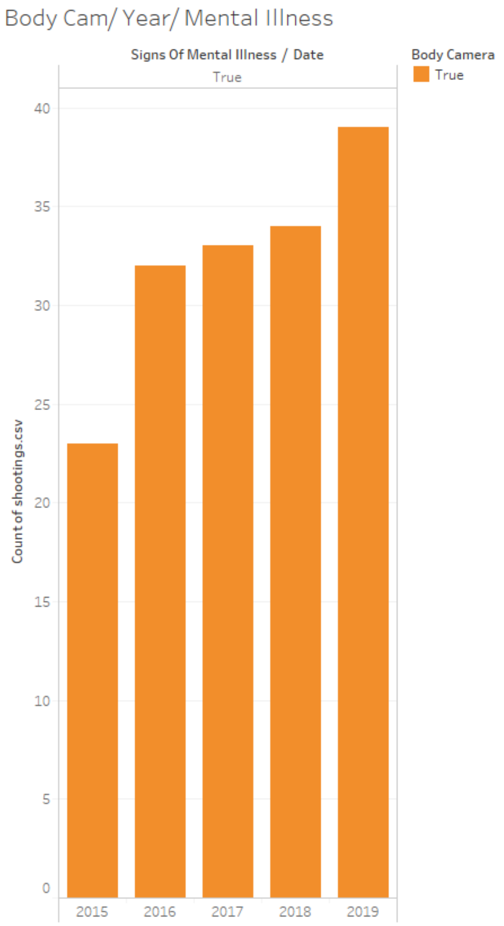


Tableau

For my last technology that I implemented into this report, I decided to use Tableau as it has easy to use graph build/design methods. I used this technology to confirm the results of my other technologies and also run some other visualisations that you can see below.

1: Below in “Fig. 52” we have the “Mental Illness” and “Body Camera” as “True being displayed over the five years with full amounts of data available(Hence 2020 has been removed). We can see that the cases of Mental Illness being a factor in a police shooting has increased quite dramatically over the four years being displayed.

Fig. 52



Count of shootings.csv for each Date Year broken down by Signs Of Mental Illness. Colour shows details about Body Camera. The view is filtered on Date Year, Signs Of Mental Illness and Inclusions (Body Camera,Signs Of Mental Illness,YEAR(Date)). The Date Year filter excludes 2020. The Signs Of Mental Illness filter keeps True. The Inclusions (Body Camera,Signs Of Mental Illness,YEAR(Date)) filter keeps 5 members.

2: Below in “Fig. 53” I am displaying the cases of “Gender” that had “Mental Illness” as a factor. As expected, we have a large difference between the male and female numbers. But if we look at how the genders perform when represented as a percentage of how many people per Gender appeared to have a mental illness the results are very different.

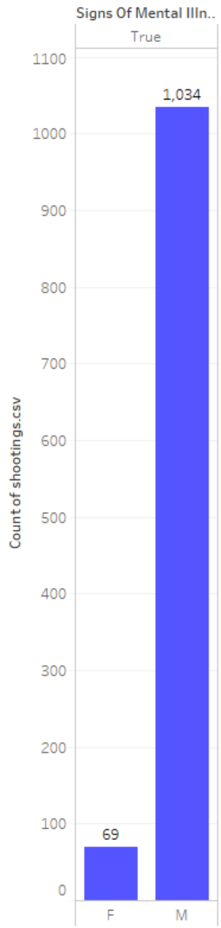
1034 Men in the data set appeared to have a mental illness out of 4673 who were shot. That is 22.12 percent of men within this data set appeared to have a mental illness.

69 Women in the data set appeared to have a mental illness out of 222 who were shot. That is 31.08 percent of women within this data set appeared to have a mental illness.

Overall, that is around a 9% difference between the two genders that were represented in the data set. This points out that women seem to have a higher chance of “Mental Illness” being a factor in a police shooting.

Fig. 53

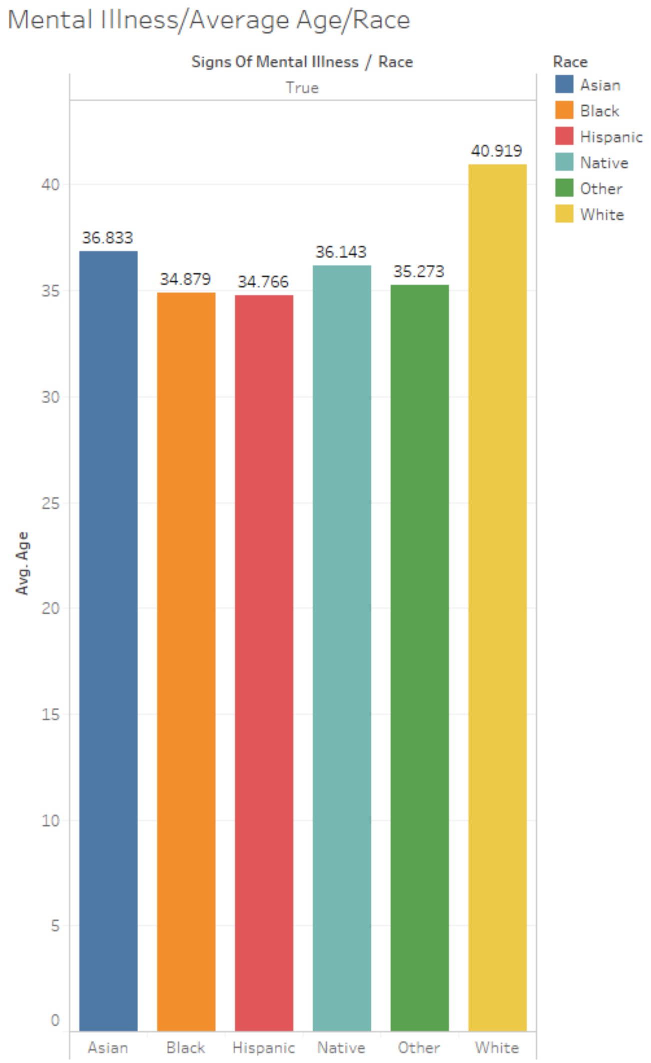
Mental Illness as "T" with Gender



Count of shootings.csv for each Gender broken down by Signs Of Mental Illness. The view is filtered on Signs Of Mental Illness, which excludes False.

3: Below in “Fig. 54” we have the average age of each “Race” when “Mental Illness” was a factor in a fatal police shooting. Again, we can see that the “White” race has by far the highest overall average age for being involved in a fatal police shooting. This also points out that the “Black”, “Hispanic”, and “Oher” race categories are being affected at much younger age groups than members of the “White” category.

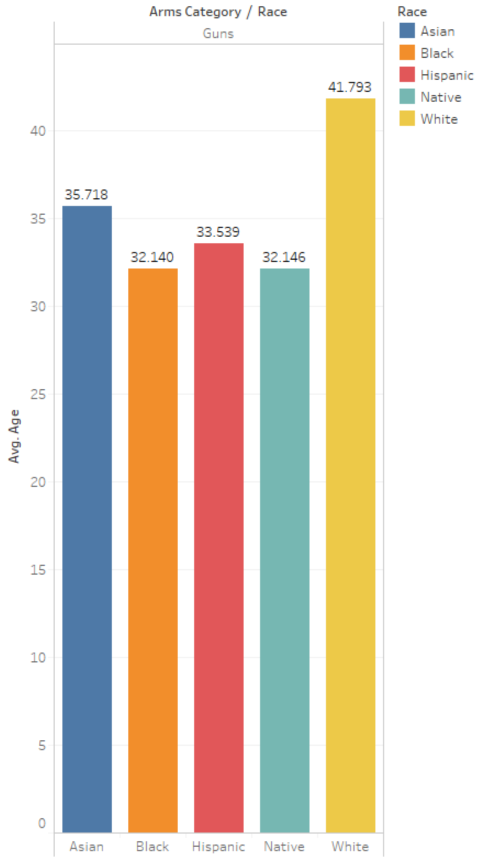
Fig. 54



Average of Age for each Race broken down by Signs Of Mental Illness. Colour shows details about Race. The view is filtered on Signs Of Mental Illness, which excludes False.

4: Below in “Fig. 55” we have each of the “Race” categories being compared to the average age that they were killed at when “Guns” was the weapons category involved in the incident. This shows that the “Black” and “Native” category of “Race” are using “Guns” that ultimately end up with them being shot and killed by police a lot more than the “White” and “Asian” race categories. This is something of concern as it indicates that the “Black” and “Native” race categories are getting involved with weapons at a younger age than other race categories.

Fig. 55

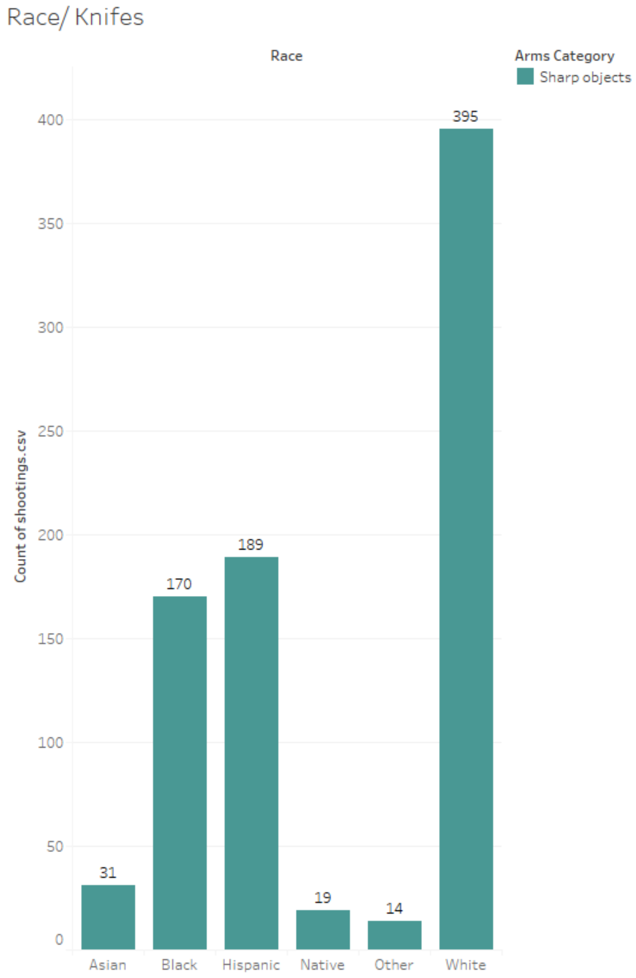


Race / Average Age / Guns Only

Average of Age for each Race broken down by Arms Category. Colour shows details about Race. The view is filtered on Race and Arms Category. The Race filter excludes Other. The Arms Category filter keeps Guns.

5: In “Fig. 56” we have visualised how “Sharp Objects” such as knives, and swords have been used in incidents involving fatal police shootings across the “Race” variable in the data set. Even though that the “White” race category has the largest overall number, the “Black” and “Hispanic” race categories have much worse numbers when we consider that the US population is 75% “White”. Even though that the “Black” and “Hispanic” race categories make up a much smaller percentage of the population they still represent a large amount of the overall “Sharp Object” Crime that is occurring.

Fig. 56

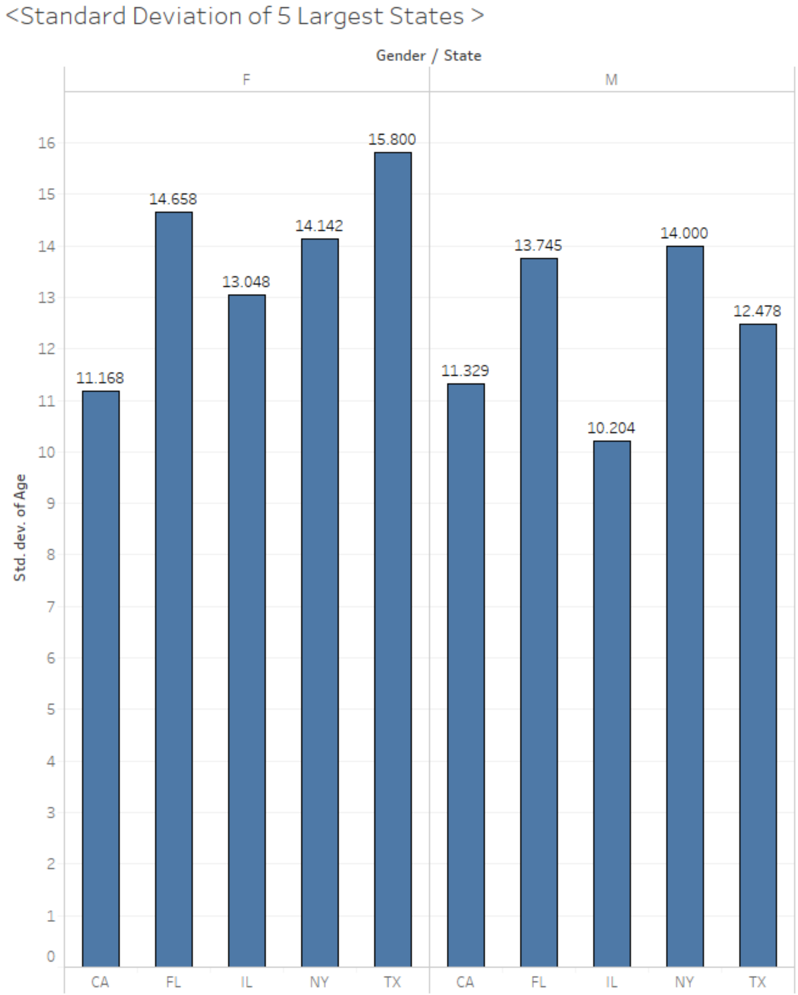


Race/ Knifes

Count of shootings.csv for each Race. Colour shows details about Arms Category. The view is filtered on Arms Category and Exclusions (Arms Category,Race). The Arms Category filter keeps Piercing objects and Sharp objects. The Exclusions (Arms Category,Race) filter keeps 52 members.

6: In “Fig. 57” we have the five largest states in the US being plotted against the “Gender” and “Age” categories in the data set when Standard Deviation is considered. The lower the standard deviation means that the figure for that state is close to the average number. The higher the number goes the further away from the average the state becomes. We can see that California (CA) and Illinois (IL) have the closest standard deviation for police shootings when age is the variable being considered. This means that these states have very few outliers, and all data points are relatively close to the normal average of age. When we look at Texas(TX) for Females and New York(NY) for males we can see a much higher standard deviation which indicates that not all of the data points are close to the average.

Fig. 57



Standard deviation of Age for each State broken down by Gender. Details are shown for State. The view is filtered on State, which keeps CA, FL, IL, NY and TX.

(N/A, N/A, 2021, QuickFacts California), (N/A, N/A, 2019, List of U.S. States with Codes and Abbreviations), (N/A,N/A, 2021, QuickFacts United States)

6.0 Conclusion

In this section of the report, I will be analysing the most important findings from the results section of the report. I will be breaking this section into the variables of the data set so that it can be read and interpreted easily.

Date: In terms of how time was a factor across the incidents occurring I will now discuss the results of some filtering that was done in excel to identify some interesting findings. Across the five largest states we can see the number of fatal shootings that occurred in each year

- 2015: CA = 182, FL = 61, IL = 21, NY = 18, TX = 99
- 2016: CA = 127, FL = 55, IL = 26, NY = 17, TX = 77
- 2017: CA = 143, FL = 56, IL = 20, NY = 15, TX = 60
- 2018: CA = 96, FL = 62, IL = 19, NY = 14, TX = 70
- 2019: CA = 111, FL = 57, IL = 10 , NY = 18, TX = 87
- 2020(As of June): CA = 42, FL = 33, IL = 3 , NY = 8, TX = 33

Below we can now observe the average number of deaths for each of the five largest states in the US.

- Average for California excluding 2020: 131.8.
- Average for Florida excluding 2020: 58.2.
- Average for Illinois excluding 2020: 19.2.
- Average for New York excluding 2020: 16.4.
- Average for Texas excluding 2020: 78.6.

Manner of Death: The "Manner of Death" column in the data set could not add much value to my overall investigation, as all of the people contained in the data are victims of police shootings. Regardless I will provide some of the basic statistical figures that I have available. This category had two variables, one being "Shot" and the other being "Shot and Tasered". This does provide some insight into what happened at the incident, but it does not change the result. A total of 248 people were tasered before being shot by police, the other 4,647 people were just shot.

The one thing that I took away from these results was why were not more of the 4,647 suspects tasered before being shot. I thought a non-lethal method of apprehending someone was preferred but the officer's safety often overrules that method of thinking.

Armed Category/Armed: By looking at the visualisations created in the above section of the report I can identify how different weapon types were used across the other variables. Starting with the most used weapons category, "Guns" appeared in the data set 2,764 times. This is made up 56.5 percent of the "Arms Category". "Sharp Objects" and "Piercing Objects" also made up over 17 percent of the "Arms Category". If the US government and its states could focus on just "Gun" and Knife" crime, they could drastically reduce the number of times a police officer needs to shoot and kill one of its citizens.

One concerning variable within this category that I noticed early on was the variable of "Toy Weapon". This variable indicates that a portion of the people shot and killed by police were done so for holding or using a "Toy Weapon". When I looked further into this weapon type I came across the following results:

- The youngest person killed for having a “Toy Weapon” was 13.
- The next youngest person killed for having a “Toy Weapon” was 14.
- The next youngest person killed for having a “Toy Weapon” was 15.
- A total of 3 children were killed at age 16 for using a “Toy Weapon”.
- And finally, 4 17-year-olds were shot and killed for using a “Toy Weapon”.

By looking at a heatmap of all of the “States” that had reported “Toy Weapons” the states of California had an alarming number even when compared to other large states such as Texas and Florida. This is something the state of California needs to investigate further. I find the results of the “Toy Weapon” category quite difficult to stomach as I think most people would agree that there needs to be better processes put in place to protect children from armed police officers.

Another variable in the “Arms Category” that I found particularly interesting was the “Unarmed” section. This includes 348 people or 7.1 percent of the “Arms Category” and it identifies people who were shot and killed by police even though that they were unarmed. By digging even further into the data and using some other variables I discovered that a total of 15 suspects had been shot and killed with the following variables being considered: The person was unarmed, there were no signs of a mental illness, there was no threat level assigned, the person was not fleeing the scene and there was also no body camera footage available. To myself and I presume anyone else who will read this report can agree that the shooting of those 15 suspects seemed unnecessary based on the information available from the data set.

Age: Regarding “Age” as a factor in fatal police shootings, this is a hugely important topic for this data set. Starting with some basic information on this variable, the youngest age of the affected suspects was age 6 and the oldest of age 91. That leaves quite a large “Age” range of 85. The overall average for the “Age” variable was 36.5. The overall variance of the “Age” variable was just 0.033. This means that most of the figures associated with “Age” were close to the average of 36.5. People who were adversely young or old would affect this variance figure. By observing a number of the visualisations in the results section of this report we can acknowledge that most people affected by police shootings are in the age groups of the early ’20s to late ’30s. As we also know that most people affected by police shootings are “Men”, we can now assume that most of these men are within the age groups of their 20’s to late ’30s. This is concerning as the trend of younger people being involved with police is remaining at a high level. If we now take a look at how each “Race” compared when “Age” is skewed against it, we can see the following results.

- Asian Average Age: 36.5
- Black Average Age: 32.5
- Hispanic Average Age: 33.6
- Native Average Age: 31.2
- Other Average Age: 33.0
- White Average Age: 39.9

We can see from the results that the “Black”, “Hispanic”, “Native” and “Other” race categories have considerably lower average age rates when compared to the other race variables.

The “Arms Category” variable was also a surprising one to visualise when compared with age as we saw specific age groups lean more to specific types of weapons. For example, we saw “Guns” being used by a wide range of age groups, but we only saw “Explosives” being used by the 35’s to 50’s age group.

I have also created a breakdown of the “Age” variable along with gender so that the totals can be displayed. See below:

- Below Twenties: Male = 264, Female = 8, Total = 272
- Twenties: Male = 1307, Female = 63, Total = 1370
- Thirties: Male = 1480, Female = 69, Total = 1549
- Forties: Male = 841, Female = 42, Total = 883
- Fifties: Male = 525, Female = 31, Total =556
- Sixties: Male = 198, Female = 5, Total = 203
- Seventies: Male = 46, Female = 3, Total = 49
- Eighties: Male = 10, Female = 1, Total = 11
- Nineties: Male = 1, Female = 0, Total = 1

Gender: When starting with this project I did not anticipate that there would be such a large disparity between the two genders that were represented within the data set. The male representation within this data set made up 95.5 percent of all of the fatal police shooting incidents. Women made up just 4.5 percent of these incidents.

Where “Gender” appeared to be one of the biggest points of concern was when I considered how “Mental Illness” was a factor across the two genders. When I looked into this, I saw that 31% of women who were involved with fatal police shootings had a “Mental Illness” as a factor when they were killed. This figure is just 22% for men so there is a clear difference between the two genders when this is accounted for. The reasoning for this disparity is unknown by analysing the data available but it is a finding that needs to be investigated further so that we can recognise why it is occurring. As “Gender” is a factor in some other findings you will see more “Gender” related findings in other sections of the conclusion.

While looking at the Standard Deviation of “Age” against the “Gender” variable and also the five largest “States” I identified a noticeable difference between how high the standard deviation is with women compared to men. The men saw a much lower standard deviation figure when compared to women. This indicates that the women in the data set were not as close to the average age of women affected by a police shooting. This also points out how much younger the male gender is affected by police shootings.

Race: Race was one of the most important factors that I needed to analyse in this project. As of writing this report, the news is filled with race-related police shootings and also the backlash that follows on from that. I will start with some of the basic statistics of the race-related data. By looking at our SPSS descriptive output we can see the following:

- “Asian” = Appeared 93 times within the data and makes up 1.9 percent of the “Race” data column.
- “Black” = Appeared 1298 times within the data and makes up 26.5 percent of the “Race” data column.
- “Hispanic = Appeared 902 times within the data and makes up 18.4 percent of the “Race” data column.
- “Native” = Appeared 78 times within the data and makes up 1.6 percent of the “Race” data column.
- “Other” = Appeared 48 times within the data and makes up 1.0 percent of the “Race” data column.

- “White” = Appeared 2476 times within the data and makes up 50.6 percent of the “Race” data column.

Now by looking at this data we can recognise which “Races” are most affected by fatal police shootings. Both the “White” and “Black” races are ones of concern within the data due to their high figures. One very important piece of information that we need to recognise is that the USA is 76.3 percent “White” and only 13.4 percent “Black” when we have a look at the latest census information available from https://www.census.gov/. Now that we know how the population is divided based on its race, we can identify that the “Black” race is being disproportionately affected by fatal police shootings as the “Black” race makes up 26.5% of the police shootings over the last five years while only representing 13.4 percent of the population. If we compare that to the “White” race we see that they represent 76.3 percent of the population but only made up 50.6 percent of the data within the data set. When looking at the “Asian” race we can see from the census information that they represent 5.9% of the population but only took up 1.9 percent of the fatal police shooting data meaning that the “Asian” race is affected in the opposite way to how the “Black” race is. Now looking at the “Hispanic” data we can see that the “Hispanic” race makes up 18.5 percent of the US population and appeared in the data set at 18.4 percent of the time. This means that out of all of the largely represented “Races” that the “Hispanic” race is the only one in the data set that appears to perform as we expect based on the population percentage and how they appeared in the data set. Now that we know that the “Black” race is disproportionately affected by police shootings the states and the government need to recognise this and implement measures to try and reduce these figures.

I will now discuss some of the other results produced by the visualisations within my chosen technologies. Firstly, we will see how the “mental illness” and “Race” variables perform in the state of California. California was chosen as it is the largest state within the US.

- The “Black” race appeared 22 times.
- The “White” race appeared 43 times.
- The “Asian” race appeared 6 times.
- The “Hispanic” race appeared 55 times.
- The “Native ” race appeared 2 times.

That is a total of 128 persons that were shot and killed in California over the last five years that had a “Mental Illness” variable attached to them when they were killed. Firstly, we need to look at the demographics of California to recognise what a reasonable number of appearances for each “Race” will be in this filter. Based on the https://www.census.gov/quickfacts/CA website we can see that in California the “Race” demographic is divided into the following:

- White: Makes up 71.9 percent of the California population.
- Black: Makes up 6.5 percent of the California population.
- Asian: Makes up 15.5 percent of the California population.
- Hispanic: Makes up 39.4 percent of the California population.
- Native: Makes up 0.5 percent of the California population.

Now by seeing how much each “Race” represents out of the total of 128 persons we can see if these percentages are similar to the population demographic or if some “Races” are disproportionately affected:

- White: Makes up 33.59 percent of the total persons shot and killed by police in California who also had a “Mental Illness” attribute considered.
- Black: Makes up 17.18 percent of the total persons shot and killed by police in California who also had a “Mental Illness” attribute considered.
- Asian: Makes up 6.68 percent of the total persons shot and killed by police in California who also had a “Mental Illness” attribute considered.
- Hispanic: Makes up 42.96 percent of the total persons shot and killed by police in California who also had a “Mental Illness” attribute considered.
- Native: Makes up 1.5 percent of the total persons shot and killed by police in California who also had a “Mental Illness” attribute considered.

Again, we can see that the “White” race is not appearing as much as we expect within the data set. That means that the “White” race is not as much affected by police shootings then we would expect based on the high “White” population percentage. With a 71.9 percentage population rate, they are only taking up 33.59 of the total police shootings when “Mental Illness” is a factor. Now when we look at the “Black” race and see how they have been affected, we can see that even though the “Black” race makes up just 6.5 percent of the California population that they are appearing 17.18 percent of the time in the filter that has been applied. This is quite a large disparity of what we would be expecting if the playing field were level between the “Races”. I will allow you the reader to interpret the results of the other “Races”.

Another concerning value that was gathered from a histogram plot shows the average “Age” of each “Race” but only concerning when a “Gun” was in use at the crime. When we look at the results, we can see that the “Black” race has an average “Gun” related “Age” of 32.14 when they were shot and killed by the police, which is the youngest average age out of the “Races”. When the “White” race is considered however we can see that their average is 41.79, which is the highest average. This displays again how much more affected younger “Black” people are being affected by police shootings.

If we now look at how “Sharp Objects” were a factor across the different “Races” available in the data, we can see that there is another large disparity amongst how the data is distributed. Looking at the histogram that was created in Tableau we can see that the “White” race had the most “Sharp Object” related to police shootings. However, if we look at the “Black” and “Hispanic” races we can see that their levels of use of this type of weapon are much higher than the “White” race when we consider how the percentage of race in terms of the country’s population is considered.

When looking at the visualisations that show how the different “Races” perform with other variables such as “Threat Level”, “Body Camera” and “Flee” we can see no disproportionate difference between each of the represented “Races”.

If I look at the results of the heat map which displays how each of the “Races” is distributed across the country, we can see the results which I would expect. We see dark areas in the larger states and light levels as the smaller states are observed.

You can also find how “Race” is a factor within other variables in the data by looking at the “Age” and “Signs of Mental Illness” sections of this results section.

City: By looking at the information provided by the created visualisations we can now say that the most affected Cities when it comes to fatal police shootings is Los Angeles, Phoenix, Houston, Las Vegas, and San Antonio. These five states had a range of forty to eighty deaths over the five years available. In terms of some of the safest cities, we can now confirm that the cities of Middleton, Josephine County, Rector, Fairchild, and Black Diamond have some of the lowest police shooting incidents in the US over the last 5 years. All five of those states had only one fatal police shooting over the five years of data.

State: In terms of how different states appeared within the data set I can now confirm that the states of California(CA), Texas(TX), Florida(FL), Arizona(AZ) and Colorado(CO) had the most fatal police shootings over the last five years. All of those states had 200 or above police shootings over the five years available. The states with the least amount of fatal police shootings were Rhode Island(RI), Vermont(VT), Delaware(DE), North Dakota(ND) and New Hampshire(NH). All of these states had less than 13 deaths over the five years that the data runs for.

Signs of Mental Illness: As mental health and illness is still a relatively taboo topic in organisations such as Police forces and Governments it was important that I had a good look at how mental health was a factor in police shootings. As I discussed, I have already gone over how women are almost 9% more likely to have a mental illness when they are shot and killed by police.

The "Mental Illness" disparity continues across some other variables in the dataset. When looking at how "Race" was a factor in "Mental Illness" I noticed that the "Black", "Hispanic" and "Other" race categories had an average "Age" of around 34 when they were shot and killed but the "White" race had an average of 40 when "Mental Illness" was a factor to them being shot and killed. This disparity between the different "Races" needs to be investigated further.

When looking at how "Mental Illness" has been increasing or decreasing over the years available we also see a concerning trend. Year on year we see increments of increase to "Mental Illness" being a factor in why a police shooting occurred. This sort of ever-increasing variable is something that needs to be looked into further as if it is not understood sooner rather than later then the police force of the US will be dealing with suspects suffering from "Mental Illness" much more often than they are now. This may result in the police forces needing additional training on how to deal with these kinds of suspects as safe as possible. (N/A, N/A, 2019, List of U.S. States with Codes and Abbreviations), (N/A, N/A, 2021, QuickFacts California), (N/A,N/A, 2021, QuickFacts United States)

7.0 Further Development or Research

In terms of future development or research of my project. There are a number of things that I would have liked to implement or use in the project but ultimately could not because of a wide range of factors. I will now discuss a handful of topics that could have improved my data analysis. These topics may become available to me in the future so this is definitely a topic I would like to revisit in the future.

- A limited number of numerical factors: Within the data set that I chose for this project there are only three numerical columns to utilise. Where this becomes an issue is when we look further into these columns and realise that the "ID" column even though it is numerical is virtually worthless to the analysis as it is only used to number the number of attributes on each row. This could not be used for any statistical test or data mining technique. The other two columns being "Date" and "Age" can be used for statistical tests and data analysis, but we are limited with what types of tests we can use and also what we can compare these numerical columns against. If in the future another data set is compiled that has more numerical values included, then this is something that I would be greatly interested in analysing. For the purpose of the report, however, I made the most out of what I had available at the time.

- There only being data over the last five years: As most people would be aware, shootings carried by US police has been a problem in the USA far longer than 2015. Being able to analyse a data set that spanned a much greater distance of time would be beneficial to this report, but this unfortunately was unavailable to me at the time. I failed to find any data sets that had as much detail as the one I had selected that spanned say over several decades. It appears that the more you go back in time the less detail there is in each data set and hence lessens the quality of an analysis report. That is why I chose a data set with lots of detail, zero missing values and spans a shorter period of time. If I manage to find a data set that can fur fill all of the spoken of requirements, then that would be something I would be interested in pursuing.

- Introducing more Technologies: The inclusion of even more technologies would also be a very good addition to my analysis in the future. Each technology has its own advantages and disadvantages so the more technologies that get implemented increases the chances of you finding something that a previous technology may not have highlighted. It is also very difficult to be proficient at lots of technologies so using as many technologies as best you can be a sure advantage in the future development of this project.

8.0 References

Reference 1: Nazir, AN, 2020/1, US Police Shootings, viewed October 2020, < https://www.kaggle.com/ahsen1330/us-police-shootings

Reference 2: N/A, N/A, 2019, List of U.S. States with Codes and Abbreviations, Viewed November 2020, and April 2021, < https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=53971

Reference 3: N/A, N/A, 2021, QuickFacts California, Viewed November 2020 and April 2021, < https://www.census.gov/quickfacts/CA

Reference 4: N/A,N/A, 2021, QuickFacts United States, Viewed November 2020, and April 2021, < https://www.census.gov/quickfacts/fact/table/US/PST045219

Reference 5: Rajput, AR, 2019, KDD Process in Data Mining, Viewed December 2020, < https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Coding Reference 1: Portilla, JP, 2019, Python for Data Science and Machine Learning Bootcamp, Viewed and Studied from December 2020, < https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/

Coding Reference 2: N/A, N/A, 2021, Choropleth Maps in Python, Viewed in April 2021, < https://plotly.com/python/choropleth-maps/

Coding Reference 3: N/A, N/A, 2018, Wordcloud, Viewed April 2021, < https://www.python-graph-gallery.com/wordcloud/

Coding Reference 4: N/A, N/A, 2021, Pandas_Profiling, Viewed in April 2021, < https://pandas-profiling.github.io/pandas-profiling/docs/master/index.html

Coding Reference 5: Dutta, SD, 2019, Visualize missing values (NaN) values using Missingno Library, Viewed April 2021, < https://www.geeksforgeeks.org/python-visualize-missing-values-nan-values-using-missingno-library/

Coding Reference 6: Kumra, AK, 2020, R – Waffle Chart, Viewed in April 2021, < https://www.geeksforgeeks.org/r-waffle-chart/

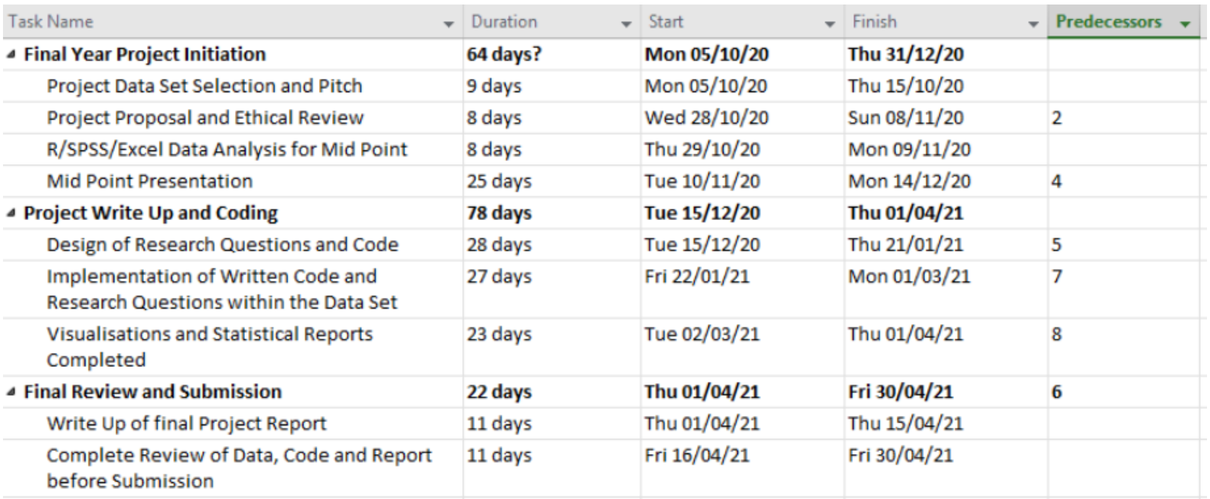
Coding Reference 7: N/A, N/A, 2021, R Color Brewer's Palettes, Viewed in April 2021, < https://www.r-graph-gallery.com/38-rcolorbrewers-palettes.html

Coding Reference 8: Eremenko, KE, 2021, R Programming A-Z™: R For Data Science With Real Exercises, Viewed in September 2020, < https://www.udemy.com/course/r-programming/

9.0 Appendices

9.1. Project Plan

Please see below a copy of my project plan. An older version of this plan is available in the Project proposal section of the report.



Task Name	Duration	Start	Finish	Predecessors
Final Year Project Initiation	**64 days?**	**Mon 05/10/20**	**Thu 31/12/20**	
Project Data Set Selection and Pitch	9 days	Mon 05/10/20	Thu 15/10/20	
Project Proposal and Ethical Review	8 days	Wed 28/10/20	Sun 08/11/20	2
R/SPSS/Excel Data Analysis for Mid Point	8 days	Thu 29/10/20	Mon 09/11/20	
Mid Point Presentation	25 days	Tue 10/11/20	Mon 14/12/20	4
Project Write Up and Coding	**78 days**	**Tue 15/12/20**	**Thu 01/04/21**	
Design of Research Questions and Code	28 days	Tue 15/12/20	Thu 21/01/21	5
Implementation of Written Code and Research Questions within the Data Set	27 days	Fri 22/01/21	Mon 01/03/21	7
Visualisations and Statistical Reports Completed	23 days	Tue 02/03/21	Thu 01/04/21	8
Final Review and Submission	**22 days**	**Thu 01/04/21**	**Fri 30/04/21**	**6**
Write Up of final Project Report	11 days	Thu 01/04/21	Thu 15/04/21	
Complete Review of Data, Code and Report before Submission	11 days	Fri 16/04/21	Fri 30/04/21	

9.2. Reflective Journals

9.2.1. October Journal

October 2020

Unfortunately, the start of this module/year for me personally did not get off to a good start. Off the bat, I have to do a specialisation that I do not want to do. During the summer NCI allowed the students of BSHTM to apply for our specialisations. A number of us picked the Business Analysis stream but now have to do the Data Analytics stream. As a student who personally dislikes coding, this was an unsettling realisation to get to grips with.

My feelings towards this specialisation got worse since starting the Project class. I do not agree with the logic that we have to simultaneously study new subjects and then produce a high standard year 4 project on those subjects that we have just studied. Students in the Computing class seem to have a significant advantage over BSHTM students as they have been doing classes that are relevant to their final project throughout previous years of their course.

Personal thoughts aside about I feel about how this is being managed I have successfully picked a suitable dataset that I will be using for my Project. The dataset that I have chosen is based around US Fatal Police Shootings over the last number of years. I believe that once I learn how to do suitable data analysis of the data set that it could reveal previously unseen statistics that could prove useful to police forces and the people of the US.

In terms of classes that are relevant to the Final Year Project for the Data Analytics students: Programming for Big Data and Business Data Analysis progress is slow for now. In Programming for Big Data, we are learning to use R. As all of the other classes such as the computing students have experience with R or Python the lecturer is not teaching the program from the start. This has left me and the others behind looking for other ways to teach ourselves how to program in R. The class structure revolves around the lecturer waiting for students to ask questions on labs that have been uploaded to Moodle. These labs are pretty advanced for even the full-on programming students, so I am finding these difficult. I have resorted to online courses on Udemy to try and get up to speed. The Data Analysis class is a level playing field and I have no issues with so far keeping up with the lecturer.

As of 29/10/2020, I have also gotten access to my Supervisor. I am planning to use this resource as best I can to achieve the highest grade possible.

Reflections

In terms of reflections on the project so far. My focus going forward for the project is to complete my Project Proposal for next week which may be difficult as I am unable to talk about my project on any technical level as I do not have that knowledge yet.

A focus on Programming in R and getting familiar with Data Analysis is a priority over the coming weeks so to improve in those areas that are vital for the success of my Project.

9.2.2. November Journal

November 2020

This was a busy month for all of the modules that I currently have within my specialisation. Every module had assignments or projects due within this period so getting work done on the project was difficult to achieve.

Projects in my programming for big data class in particular were heavy when it came to workload. As I was busy with other modules, I did not achieve any meaningful progress on the project. However, as I was using R for one of my projects I got to learn how to query and develop visualisations using R on a data set. Now that I have newfound knowledge on how to pull interesting findings from my data set, I will now be implementing these new methods of coding into my analysis.

Reflections

My focus for the coming month is to focus on the Midpoint Presentation. As my workload has lessened for December before it comes back up again in early January, I can now focus on making some progress with my project. A focus on learning as many technologies that can make my analysis as accurate as possible is going to be my approach to the midpoint presentation. I am aware that this monthly report is short but as I explained at the start, I did not have the time to provide any meaningful progress to the project during this period. My December journal will have much more information about how I have done my analysis for my Midpoint presentation and how I went about putting it all together.

9.2.3. December Journal

December 2020

In general, this month was a focus and final push to get ready for the Mid-Point Presentation. This mainly involved preparing lots of visualisations and information that I have managed to pull from my dataset which I could then use in my Presentation.

Work I did on the Mid-Point Presentation involved: preparing my visualisations, making sense of the information, learning how to use R effectively, learning how to use SPSS and Excel to my advantage and writing up reports and filming my presentation. Some challenges I faced while doing my point presentation were mainly based around learning new technologies that I previously have had no experience with such as SPSS and R. These challenges will continue into the next semester where I will be implementing more technologies to try and bring more out of the data set.

Reflections

In terms of reflections and lessons learned for the month of December. One big thing that I am thankful for doing is starting all of my assignments and projects as early as I could. This allowed me to have sufficient time to give each assignment and project enough time that they needed. For something that I wish I had done differently in December is that I would have preferred to have a better understanding of the technologies that I have implemented into my project.

My overall focus for tor the final runup of my course is to focus on learning the technologies that I need to get what I want out of my dataset. As I only have another two modules I will much more time to apply to the end of this project.

9.2.4. January Journal

January 2021

As for January, I was focusing on my end of semester assignments so I could not give the project the time I gave it in previous months. As we go into February the focus will be getting up and running with new technologies that were not included in the mid-point presentation. These new technologies include Tableau and Python. Tableau will be used to visualise my data set in ways that I previously had not done before. Python will also be used to visualise the data set but before I can start with Python, I will begin learning how to code in this new language.

As of the second semester of my final year, I have also started to study “Advanced Business Data Analysis” and “Data and Web Mining”. As I progress in these new modules, I also hope to include any new skills that can be obtained in these classes into my project so to improve my analysis of my data set.

Reflections

As my focus was on my end of semester assignments for the month of January, I have no notable reflections to add to this month’s reflective journal.

9.2.5. February Journal

February 2021

The month of February was mostly used to get up to speed with a new technology that I have implemented into my project. This new technology is called Tableau and is a very powerful piece of software that is used to visualise large amounts of data. Luckily for me, the college has a license for this software, so it was easy for me to get it up and running on my machine.

The output for Tableau is sheets that contain the visualisation that you have created. I have created and designed several sheets with different visualisations in each that provide insights into my dataset. Each of these visualisations will be included in my final report at the end of the year. These visualisations will add another layer of understanding to the issues that I am trying to make sense of within my dataset. As my dataset only has a limited amount of numerical data, I am mainly trying to visualise how different character columns interact and then decide what the significance of this is after seeing the results. This process involves a lot of time-consuming trial and error as getting a visualisation is easy but getting one that is relevant is proving to be difficult.

In February I have also gotten started and more familiar with Python and its capabilities. In the next monthly report, I will go into detail about what I have achieved within Python as of now my progress has been focused on getting familiar with the new technology.

Reflections

In terms of reflections for February. I am very happy with the work that I have completed in Tableau as it has provided me with a large number of high-quality visualisations that will prove very valuable when writing up my final project report at the end of the year. With regards to Python, it is now time to shift my time from Tableau to Python so that I can add another technology to my project.

9.2.6. March Journal

March 2021

As discussed in last month's report I will be describing what exactly I have learned while getting up to speed with Python and how I am implementing it into my final year project.

As the modules, I am taking this semester do not cover how to use Python as an analytical tool I had to find other resources to help me get up to speed. I have used Udemy previously to learn how to use R for other modules and this project, so I looked there for a course in Python.

Once I had found a suitable course, I got started a few days later. The course in question has both videos and practical sessions for learning how Python works. The course is taught through 165 lectures in total in the form of videos. It also has detailed code notebooks for each lecture in case I needed to reference a piece of code in the lecture.

The course covered a wide variety of elements for Python. I will not include everything here, but some examples include NumPy, Pandas, Seaborn, Matplotlib, Plotly, Sckit-Learn, Data Visualisation, Linear Regression and Machine Learning.

In terms of what I have implemented so far into my project, I have used the following parts of Python on my dataset: NumPy, Pandas, Seaborn and Matplotlib. These are used mostly for data analysis, plotting and visualisations of my data.

As soon as I get more comfortable with these basic parts of Python, I will then start to implement more advanced methods of analysis. I would like to include methods of data mining and machine learning if I can manage to implement those technologies onto my data. In the report for April, I will have hopefully more information on the more advanced methods of Python.

Once I have got everything I can from Python I will be in really good shape in terms of items to include in my report, subsequently making the report writing easier and more interesting.

Reflections

Not much to add this much in terms of reflections. I need to focus on getting as much as possible from Python. April will be a busy month for me in terms of assignments so this could slow down progress for the project, but this is to be expected.

9.2.7. April Journal

April 2021

As I discussed in last month's Reflective Journal, I will be going through what I have achieved in Python over the last couple of weeks. I will also go over and explain what I have been up to over the last month.

Before I could continue with finishing up the final year project, I needed to complete assignments and Terminal Based Assessments of my other modules. Due to the technical difficulties that impaired the college a few weeks back these submissions were delayed along with the final year project submission date. Once these were completed, however, I could then completely focus on my project.

Some core tasks that I have been working on during the month of April were the following:

- Cleaning up all code across all of the technologies: This process involved going through all of the code that I had written throughout this project and deleting any unnecessary code, adding sufficient notes to each line of code, and tidying up each script so that it was easy to understand and reproduce.
- Planning and researching better methods for writing my final report: This process was important as it would cement how I would be formatting the final report in terms of how I would display my outcomes and explain their results. By having a plan like this I could make sure that the report serves its purpose and is easy to interpret.
- Deciding which outputs from my technologies should be included in the report as not all are suitable or worthwhile putting in the report: I needed to go through all of the outputs and scripts that I had written and decide which ones were most powerful and should be included. Once I had decided the outputs that would be included, I placed them in a folder so I could then copy them into the report when I needed them.
- Python: Previously in my work with Python I had only used it for basic descriptive statistics and linear regression models, but I now had to use it for visualisation methods. By using the information, I learned in the Udemy course I had completed and various online sources I managed to get some very interesting visualisations applied to my data set.

Reflections

In terms of Reflections for this Month, I do not have much to touch on. All in all, the hard work is done and all that's left is to is polish off all of my code and finish writing up the report.

9.3. Project Proposal

9.3.1. Objectives

The objective of me choosing the Data Set "US Police Shootings" is to apply an in-depth analysis of the data to expose any trends or potential anomalies within the data that may not be visible at first glance of the dataset. It will also become useful to any third parties who read my eventual report on the data and decide to act on the findings.

Some potential points may include the following:

- Understand if any race, in particular, is proportionately more at risk than others. This piece of analysis may turn out to be the most important finding as we are still seeing alarming numbers of Black Americans being killed each year by police for an apparent trivial reason.
- The location as to where these shootings are occurring is also vital as realising which State and City have the worst statistics would allow the authorities in those states and cities to implement new solutions to try and reduce their numbers.
- Identify the age and gender of the people who are most likely to be involved in a police shooting. By educating this age group on this issue they may be able to avoid situations that lead to police shootings better.
- Identify the reasoning for the shooting. See if weapons were used and if none were then to see what caused the shooting.
- What percentage of Police officers who were involved in a fatal shooting were using a body camera? Some states require body cameras by law to be used, so seeing how many of the shootings in those states did not have body cameras available could be an indicator to foul play of some description.
- Ideally, once the analysis is complete the findings can be used to introduce new solutions to help prevent more shootings in the future. The findings could be useful to local police, state police, the US Government, the general population of the US and most importantly the most vulnerable people to be involved in a police shooting.
- Compare the fatality rate of someone to police had "Signs_of_Mental_Illness" versus someone who seemed to be mentally healthy.
- Look to see how many of the killings had guns involved and whether that state where the shooting occurred had strict background checks for gun ownership.

9.3.2. Background

While searching for potential data sets online the one that caught my eye was “US Police Shootings”. The reason this data set stood out to me among the others I had looked into was that this data could serve a real purpose if analysed and distributed correctly. The most recent US police shooting that I can remember was the shooting of Breonna Taylor in March 2020 so the data set of “US Police Shootings” seemed like it could give off some interesting analytical findings.

Over the last decade or so I have witnessed multiple shootings and acts of violence in the US involving guns. Other data sets I researched and was interested in such as ones based on Formula 1 and space travel were interesting but ultimately would not serve much of a purpose and would not have the same meaning behind them.

9.3.3. Technical Approach

In terms of research into my chosen data set type. I have looked at data from previous years but from looking at them I think my data set may uncover some more interesting statistical information and or anomalies. I believe this to be the case as my data set spans 5 years from 2015 – 2020. This means that we have fresh data to look at, especially over the last few years to look at and potentially spot any new information that has occurred over the last five years. Lots of other data sets of this type also usually only look at 2 or 3 years max for their statistical findings. I am hoping that by looking at a longer period I will be able to show some more accurate and compelling data.

In terms of my approach to how I will extract the data I talk about above; I will be relying heavily on the information I am learning this year studying R and data analysis. This will help me pick out the best parts of the data set to analyse and report on.

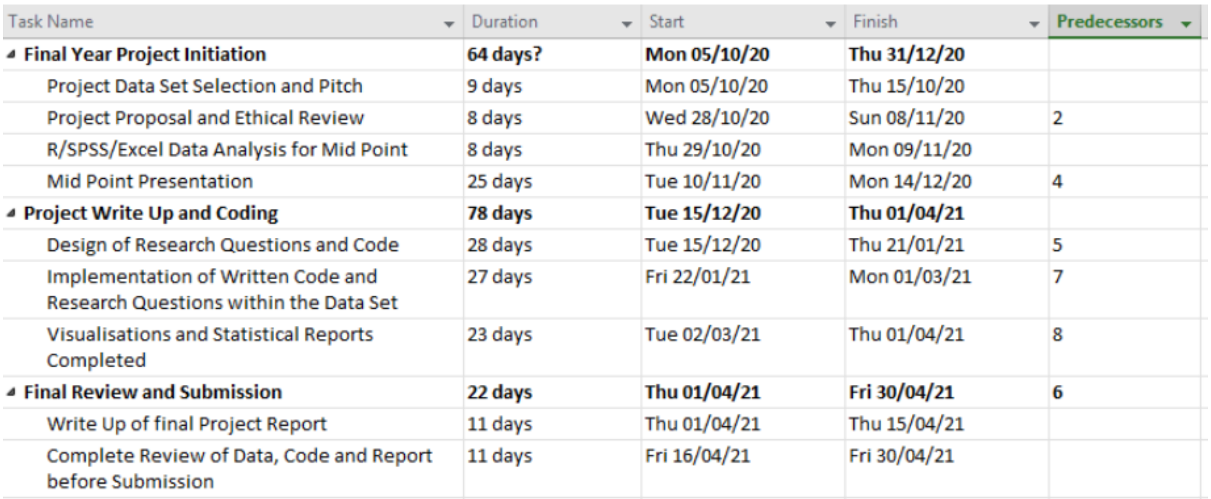
I will be running plots, graphs, tables, and statistical analysis methods where best suited to achieve the results I want. This will be achieved through programs I have already mentioned such as R, Excel, and SPSS.

9.3.4. Special Resources Required

In terms of special resources, I currently am unaware of any such resources being necessary.

9.3.5. Project Plan

Please see below the first draft of the project plan.



Task Name	Duration	Start	Finish	Predecessors
Final Year Project Initiation	**64 days?**	**Mon 05/10/20**	**Thu 31/12/20**	
Project Data Set Selection and Pitch	9 days	Mon 05/10/20	Thu 15/10/20	
Project Proposal and Ethical Review	8 days	Wed 28/10/20	Sun 08/11/20	2
R/SPSS/Excel Data Analysis for Mid Point	8 days	Thu 29/10/20	Mon 09/11/20	
Mid Point Presentation	25 days	Tue 10/11/20	Mon 14/12/20	4
Project Write Up and Coding	**78 days**	**Tue 15/12/20**	**Thu 01/04/21**	
Design of Research Questions and Code	28 days	Tue 15/12/20	Thu 21/01/21	5
Implementation of Written Code and Research Questions within the Data Set	27 days	Fri 22/01/21	Mon 01/03/21	7
Visualisations and Statistical Reports Completed	23 days	Tue 02/03/21	Thu 01/04/21	8
Final Review and Submission	**22 days**	**Thu 01/04/21**	**Fri 30/04/21**	**6**
Write Up of final Project Report	11 days	Thu 01/04/21	Thu 15/04/21	
Complete Review of Data, Code and Report before Submission	11 days	Fri 16/04/21	Fri 30/04/21	

I have chosen to leave the timeline out of this current proposal as the image is very large and would be difficult to read in word.

9.3.6. Technical Details

As I am a Data Analytics student, I will be doing most of my analytical findings through several programmes and languages that will help me squeeze as much analytical data from the dataset as possible. The applications and coding languages that I intend to use are as follows but not limited to the following:

Excel: As most people know Excel is an extremely powerful and capable piece of software that provides the base for my data set. It allows me to view it in a manageable way without getting too lost in the details. It will also allow me to run basic statistical and analytical formulas through the dataset.

SPSS: This is a Statistical software package used for interactive or batched statistical analysis. This is a piece of software that I have zero experience with and have never used since this semester. I am currently learning how to use it in my Data Analysis class so I will be learning how best to use this software to my advantage so to achieve the most interesting results from my data set.

R: Coding in R is used for statistical computing and graphics. It is widely used by statisticians and data miners for developing statistical data analysis. I am new to using this software, so I am having to learn as I go. I will be trying to get the most out of my dataset using R as best I can. R should be considered as one of the main ways I will be retrieving data from my data set as my project goes on.

9.3.7. Evaluation

In terms of testing my data set, I will not have to do any sort of unit testing or program testing as I am in the Data Analytics stream. Instead, I believe I will be confirming that the statistical information that I retrieve from the data set is accurate and is of use to my overall project aims.

I will be able to confirm that code works by seeing if the data that I wish to be shown via graphs and plots etc are being displayed correctly.