[redacted]

DISSERTATION

Analysis on the Influence of Mental Illness and Racial Discrimination on Fatal Police Shootings

By

BASMAH ZAHID

[redacted]

A report submitted in partial fulfillment of the requirements of Asia Pacific University of Technology and Innovation for the degree of

BSc (Hons) in Computer Science with specialism in Data Analytics

Supervised by [redacted]

2nd Marker: [redacted]

July-2021

[redacted]

Acknowledgement

Finally, I am at the end of my degree where I have completed my FYP and I have nothing but gratitude and respect for my supervisor, [redacted] She has greatly guided me and offered me help whenever I ran into a problem. She encouraged and helped me stay on a proper schedule to ensure I can finish my work to the best of my abilities. I am extremely appreciative of her supervision and cannot thank her enough.

I am also thankful towards my second marker [redacted] for evaluating my work.

I would like to thank our FYP coordinator, [redacted] for his continuous support in replying to my emails and concerns whenever I had queries regarding any aspect of my project.

I am extremely grateful to my lecturers who taught us in these past three years, I learned a lot from them and will be forever thankful for their patience and helpfulness through my time here.

I am thankful for my friends who stayed up countless nights with me working on our assignments together and supporting each other through tough times.

I would like to thank my family; my parents, and my siblings for supporting me whenever I had tough times and listening to me whenever I had troubles regarding work or life.

Table of Contents

Chapter 1: Introduction to The Study

1.1 Project Background

Various stipulations have been used to describe Fatal Police Shootings. According to the Austin Police Department (2015), they may be referred to as Officer-Involved Shootings which has been defined as a situation where the officer has fired their firearm deliberately or by accident at the victim which has caused the victims death or used a less-fatal weapon or item and struck the victim which contributed to the victim's demise. These actions taken by the police have been classified as police brutality which is the use of unnecessary and illegal force against civilians. Numerous forms come under the term such as "assault and battery", torture, manslaughter, provocation (false arrest), bullying or intimidation and verbal abuse, among other forms of abuse. There have multiple studies that have unfailingly shown that a police officer will most probably shoot a black unarmed man than a white unarmed man. (Hemenway et al., 2018)

Hemenway et al. (2018) state that police officers in the USA contribute to murders of more than 1000 citizens every year, this statement is further backed up by the FBI and Vox (Lopez,2018). It is observed that US fatal police shooting rates are significantly higher than other first world countries; they kill 4 times higher than the rate in Canada, 22 times higher than the rate in Australia, 40 times the rate in Germany and 125 times greater than the rate in England/Wales.

There may be several factors that may contribute to fatal police shootings such as mental illness, racial discrimination, firearm availability, threat level etc. This research is focusing on two of those factors, mental illness, and racial discrimination.

Every year 19% and 4.1% of U.S adults experience some manner of mental illness or a severe mental illness, respectively. Unless treated, most adults usually go on to function normally throughout their lives however ones who have a more severe case end up in certain situations that are out of their control, in reference to this study, police shootings. (Parekh, 2018)

Mental illness refers to several types of conditions which may affect how you function in daily life such as anxiety, depression, schizophrenia and more. Worldwide, many people suffer from certain mental health concerns however it becomes an issue when brings unwanted stress in a person's life. (Mayo Clinic, 2019) The American Psychological Association (2020) released a report

regarding the mental health crisis in 2020. They reported that due to the ongoing coronavirus pandemic, 67% of the adults experienced elevated stress over the past year. While comparing their report in 2019, the respondents for 2020 stated they have various sources of stress, out of those, 62% of adults consider mass shootings a significant source of their stress as compared to the 71% in 2019, while 59% consider police violence against minorities a source of stress as compared to the 36% of adults in 2016. People of colour reported discrimination as a factor for their stress as well; 48% black Americans, 43% Hispanics, 42% native Americas, 41% Asians and 25% whites.

"Racism has been defined as an ideology of racial superiority followed by discriminatory and prejudicial behaviour in three domains: individual, institutional, and cultural." (Pieterse et al., 2012) Racism is the belief of one social group that they are superior in all terms of life while racial discrimination encompasses any intentional or unintentional action that influences singling out people due to race and enforcing burdens on them rather than on others or withdrawing access to benefits accessible to other members of the community. (Thomlinson, 2020) Americans that have been subjected to police violence come from all races, ethnic backgrounds, ages, classes, and sexes.

Data analysis refers to the tools and techniques that are utilized to enhance productivity and enable us to extract and classify data to detect trends and patterns to validate certain key requirements. (Varga, 2018) This research provides an exploratory data and predictive analytics approach on the influence of mental illness and racial discrimination on fatal police shootings using fatal police shooting dataset. Data analysis will greatly aid us in gaining a deeper understanding of the factors that can influence these shootings and whether or not mental illness and race of a person are critical in these situations, the visualization aspect will allow the presentation of these facts and figures to be communicated more effectively.

1.2 Problem Statement

The police force was initially seen as heroes to the people of the United States but now they are seen as criminals and violators of human rights. In 2020, the high-profile murders of George Floyd, Breonna Taylor and Laquan McDonald caused extreme backlash towards the law enforcement however as expected, the officers in question simply were "suspended" or put on a "leave of absence". The majority of these officers are given the green light to use lethal force in the name of "self-defence" or "preventing a crime" irrespective if the threat is towards the civilian

themselves or the officer. The Guardian reported two of the deadliest police departments in the US that underwent investigation in early 2017, The Bakersfield police department and the Kern County Sheriff's office. Both were reported to have the highest number of kills than any other province in America in 2015 and have been accused of corruption and violence. (Laughland and Swaine,2016)

The unjust killing of victims in these fatal police shootings in the United States should be considered as a humanitarian crisis and public health emergency. According to various research conducted; a Latino man is more at risk of being killed than a White man by law enforcement officers, a Black man has 1 in 1000 chance of being killed in their life by the law enforcement meanwhile the risk of being killed by cops rises for both men and women of every race and ethnicity in their early 20's up to the age of 35. (Edwards, Lee and Esposito, 2019) Other studies conducted stated that 23% of victims of fatal police shootings exhibited signs of mental illness. (McGroarty, 2019) There were surveys conducted with the public and police officers as respondents; 67% of police officers stated that they believed the death of black individuals was an isolated event while 31% believed there were other reasons that may have caused the death. 60% of the public believed the opposite and thought it was a part of a bigger issue while only 39% stated it was an isolated event. (Bacon, 2020)

It has been stated time and time again that police officers within the United States are far more likely to kill in an encounter with a victim than any other country, The American Journal of Public Health states this is probably due to weak gun laws (Kivisto et al., 2017); An additional factor can be the anti-black racism within the majorly white police department that further cause these deaths. In recent times, body cams were brought in so there could be solid proof whether the officers reacted in self-defence or killed in cold blood, and it showed that more percentage reacted because of the latter. It is also highly likely that fatal police shootings may directly influence the cause of mass school shootings, as most school shooters are predominantly white and are reacting due to their own racist beliefs. These shootings further fuel racist hate crimes in the US that claim the lives of hundreds to thousands of people per year. This is where this research can be used through its exploratory and predictive analysis of fatal police shootings to see how a person's race or possible signs of mental illness may play a part and how other factors may contribute to these killings. Additionally, it may be used to see whether other factors correlate to the person race or

mental illness at the time of shootings, hence with the use of this project, can be used to enact strict gun laws, reduce officer-related fatalities, and prevent hate crimes in the future.

1.3 Rationale

In 2015, The Washington Post as well as The Guardian noted the rising number of death due to fatal police shootings and created a database which contains all the information regarding these situations. If the average person were to go through every line in a dataset and keep track of what the previous information was, they would not be able to reach a proper conclusion hence to achieve the goals for this project, conducting an exploratory data analysis is appropriate.

Ayer and Ramchand (2021) analysed whether mental illness can be a critical factor in gun violence and found that about 20% of people with mental health crisis have drug-related issues, this may also lead to mass violence committed by a mentally ill person which may cause law enforcement officials to intervene and be forced to shoot.

Ever since various official new sources started collecting data about the total number of deaths per year, with 1000 people killed within 2 years in 2017 and the rise of the black lives matter movement, the justice department under Barack Obama had launched an investigation.

Harte and Mclaughlin (2017) stated that due to this, over 14 police municipalities were under the federal consent decree which obliged them to reform the way their officers used force, profiling, recruitment, training and overseeing,

The Washington Post even suggested that due to their database, it was very likely that police departments may consider restructuring a part of their rule such as making it a strict requirement for a police officer to file a report whenever they do need to point a gun at a victim but do not end up firing at them. (Jennings, 2017)

Exploratory data analysis has the ability to show data trends and illustrations such as histograms, heatmaps, bar charts that can help to recognize and assess the relationship between racism and mental illness with fatal police shootings and offer insights into the police brutality that people of colour endure. It can further be utilized as an aid in reducing these shootings and address the loopholes in certain legislation laws, such as pardoning of police officers instead of giving out proper punishments as well as addressing the need for proper psychiatric care for mentally ill

individuals while data mining models such as regression and decision trees may be used to predict the probability of an individual being shot due to their race or mental illness as well as any other factors. The EDA can further be used in other countries where this issue is also prevalent and be utilized to see the trends and relationships to ensure proper and corrective measures can take place.

1.4 Potential Benefits

Tangible and Intangible are two types of potential benefits for this research. Tangible benefits address the advantages that can be evaluated, such as items and outcomes that can be calculated. Intangible benefits, on the other hand, address the advantages that are indirect and cannot be wholly subjective measures. (Thurimella and Padmaja, 2014)

1.4.1 Tangible Benefits

- Through this project, improvement in processes as the visualization can be used to enact firearm availability laws.
- By conducting this analysis, there can be improved productivity concerning the collection of data to reduce the use of force by police towards civilians (Eng and Wenig, 2020) as well as holding law enforcement officers accountable for their actions which may allow victims to receive the justice they deserve.
- Implementation of training of officers to respond to situations more appropriately.

1.4.2 Intangible Benefits

- Improving the lives of people of colour and mentally ill individuals in America
- Initiate the process of enabling the mentally ill to receive psychiatric care in America.

1.5 Target Users

The target users will be as listed below:

1. Seniors' members of the law enforcement who oversee making decisions regarding officer-involved shootings.
2. The legislative bodies to allow for passing laws regarding firearm availability.
3. The citizens of the United States of America and the people worldwide.
4. Officials who are responsible for providing a proper means of psychiatric care for the mentally ill.

The analysis that is being conducted has been used in similar research areas such as an exploratory data mining analysis which identified patients with depression who are at high risk with suicidal tendencies. This research was then used to decrease the probability of this event occurring by taking certain measures such as identifying patients at an early stage. (Ilgen et al., 2009) Insights were gained from the research to make a well-informed and evidence-based decision. By conducting this analysis on the fatal police shooting dataset, it will aid in improving the lives of innocent civilians who are at risk of being killed at the hand of law enforcement.

1.6 Scopes and Objectives

1.6.1 Aim

Provide insights on how mental illness and racial discrimination impact fatal shootings in the United States by utilizing Exploratory Data Analysis to provide awareness of police brutality. This data can be used to address how police behaviour is influential for other hate crimes all over the country. Moreover, relevant authorities can then push for stricter legislation laws and punishments for these crimes by using the analytical data as support for these actions.

1.6.2 Objectives

1. To investigate the key factors to analyse the influence of mental illness and racial discrimination on fatal police shootings.
2. To analyse the influence of mental illness and racial discrimination on fatal police shootings using explanatory data analysis techniques.
3. To evaluate the performance of analysis results using evaluation metrics such as R-Squared (R^2), Mean Squared Error (M.S.E), Mean Absolute Error (M.A.E) or Misclassification Rate (MISC).
4. To develop a dashboard to visualize the analysis results

1.6.3 Deliverables

Analysing fatal police shootings will allow the discussion of how racial discrimination exists within law enforcement and how it directly affects the lives of innocent people. Furthermore, The EDA can be used by authorities to try and enforce stricter gun laws in the United States, educate the public and raise awareness against the stigma of mental illness and the need for psychiatric care. EDA can be used to make decisions such promote a stricter punishment for law enforcement

instead of simple suspensions. Furthermore, it is hoped to help provide a visualization of the unfairness against the people of colour in America. Deliverables for this project are listed below.

1. Importing dataset into SAS Enterprise Miner for data exploration to distinguish relations among variables.
2. Conduct data exploration to find the relationship between certain variables.
3. Conduct data pre-processing steps such as handling missing or redundant data, removing outliers using SAS Studio and SAS Enterprise Miner.
4. Analyse data set using exploratory data analysis including predictive modelling techniques such as regression and decision trees.
5. Create a dashboard using Power BI to show results of the data analysis using bar charts, ribbon charts, pie charts, decomposition trees, clustered graphs and more.

1.6.4 Nature of Challenge

The challenges that this research may face are firstly data cleaning and data pre-processing. This step is crucial to the project's success, these two phases need to be carried out meticulously to ensure the data is free of inaccuracies and redundancy, as well as make sure there is no loss of important data. Pre-processing is a very important step before and after it is imported to SAS Enterprise Miner and for analysis in Power BI. While being imported to Power BI, it is assumed by the program that the dataset is clean and in case this is untrue, then there is a possibility of unreliable results.

Moreover, certain pre-processing should be carried out before importing into SAS Enterprise Miner as the tool is unable to carry out certain steps on its own. (Sarma, 2017)

Furthermore, another possible challenge can be identifying relationships between the variables in the dataset as well as producing the test data from the actual dataset. The researcher will need to determine how much data needs to adequately allocated to the training, validation, and test datasets.

There are also some known issues that users usually face with Power BI such as processing of a certain number of rows, it is noted that it runs into problems when the rows are over 20000, fortunately for our project, the rows are at a maximum of 6,000 and may reduce after the pre-

processing steps are completed. (Brook, 2020) Later, the dashboard's must be interactive and have no issues occurring when testing out relationships between different variables.

1.7 Overview of the Report

This research project will discuss the context of the problem through journals, articles, and materials relevant to the scope. Chapter 1 consists of the project background and problem statement which talk about mental illness and racial discrimination as well as lists certain studies to refer to the problems listed. The aims, objectives and deliverables will also be discussed as well as the rationale, nature of the problem and potential benefits of the system. This chapter will give the reader an overview of the research. Chapter 2 is the literature review and will delve into the issue in-depth starting with the fatal shootings and police shootings as well as the influence it has had worldwide as well as in the research domain, America, then racial discrimination and mental illness will be discussed as a factor on fatal police shootings. These are discussed by using articles, journals and books published by reliable and academic sources to give an overview as well help the reader understand the problem, without this, the project outcome and deliverables will be unsatisfactory; similar system and research projects will be discussed as well as compared in a table to conclude this chapter. Chapter 3 is the technical research chapter, and it discusses the IDE and tools that this research will be employing as well as provide the advantages and disadvantages for each. Chapter 4 consists of a comparison between data mining methodologies as well as justification on the selected methodology followed by a discussion of each phase within them. CRISP-DM methodology has been chosen as the most suitable methodology to ensure the success of this research project. Chapter 5 will be the data analysis process of constructing and optimizing the selected models based on the target variables and will consist of the data exploration, data pre-processing, visualization, and model building process. Chapter 6 will follow up from the previous chapter and carry out the evaluation of the model based on certain metrics and select the best model for the target variable before proceeding with a discussion based on the best model and relating to relevant or similar studies as to this research project as well as provide suggestions based on the results. Chapter 7 is the final chapter which will contain our conclusion regarding our overall research, the final reflection for the duration of this project, the gaps and limitations, the references used for the project and the appendix which has all the relevant documents including the project proposal form (PPF), project specification form (PSF), project log sheets, fast track, disclaimer form, (list remaining forms) With this, the research project has been completed in its entirety.

1.8 Project Plan

No.	Tasks/Meetings	Starting Date	Ending Date	Days	Status
1	FYP Semester 1 - Meeting 1	3-Dec		0	DONE
2	FYP Semester 1 - Meeting 2	18-Dec		0	DONE
3	FYP Semester 1 - Meeting 3	5-Feb		0	DONE
4	FYP Semester 1 - Meeting 4	26-Feb		0	DONE
5	FYP Semester 2 - Meeting 5	11-May		0	DONE
6	FYP Semester 2 - Meeting 6	17-Jun		0	DONE
7	FYP Semester 2 - Meeting 7	8-Jul		0	DONE
-	Project Proposal Form (PPF)	23-Nov	30-Nov	7	DONE
-	Project Specification Form (PSF)	18-Dec	28-Dec	10	DONE
CHAPTER 1: INTRODUCTION TO THE STUDY		12-Jan	27-Jan	15	DONE
1.1	Project Background	12-Jan	14-Jan	2	DONE
1.2	Problem Statement	14-Jan	16-Jan	2	DONE
1.3	Rationale	16-Jan	17-Jan	1	DONE
1.4	Potential Benefits	17-Jan	20-Jan	3	DONE
1.5	Target Users	20-Jan	21-Jan	1	DONE
1.6	Scopes and Objectives	21-Jan	25-Jan	4	DONE
1.6.1	Aim	21-Jan	22-Jan	1	DONE
1.6.2	Objectives	22-Jan	23-Jan	1	DONE
1.6.3	Deliverables	23-Jan	24-Jan	1	DONE
1.6.4	Nature of Challenge	24-Jan	25-Jan	1	DONE
1.7	Overview of the Report	25-Jan	26-Jan	1	DONE
1.8	Project Plan	26-Jan	27-Jan	1	DONE
CHAPTER 2: LITERATURE REVIEW		27-Jan	16-Feb	20	DONE
2.1	Introduction to Literature Review	27-Jan	28-Jan	1	DONE
2.2	Domain Research	28-Jan	11-Feb	14	DONE
2.3	Exploratory Data Analysis and Visualizations	11-Feb	12-Feb	1	DONE
2.4	Predictive Analytics	12-Feb	13-Feb	1	DONE
2.5	Evaluation Techniques	13-Feb	14-Feb	1	DONE
2.6	Similar Systems	14-Feb	15-Feb	1	DONE
2.7	Summary	15-Feb	16-Feb	1	DONE
CHAPTER 3: TECHNICAL RESEARCH		16-Feb	21-Feb	5	DONE
3.1	IDE (Integrated Development Environment)	16-Feb	17-Feb	1	DONE
3.2	Tools Chosen	17-Feb	18-Feb	1	DONE
3.3	Operating System	18-Feb	19-Feb	1	DONE
3.4	Hardware and Software Requirements	19-Feb	20-Feb	1	DONE
3.5	Summary	20-Feb	21-Feb	1	DONE
CHAPTER 4: METHODOLOGY		21-Feb	1-Mar	8	DONE
4.1	Introduction	21-Feb	21-Feb	0.5	DONE
4.2	Methodologies Comparison	21-Feb	22-Feb	0.5	DONE
4.3	CRISP-DM – Selected Methodology	22-Feb	23-Feb	1	DONE

4.4	Business Understanding	23-Feb	24-Feb	1	DONE
4.5	Data Understanding	24-Feb	25-Feb	1	DONE
4.6	Data Preparation	25-Feb	26-Feb	1	DONE
4.7	Data Modelling	26-Feb	27-Feb	1	DONE
4.8	Evaluation	27-Feb	28-Feb	1	DONE
4.9	Deployment	28-Feb	28-Feb	0.5	DONE
4.10	Summary	28-Feb	1-Mar	0.5	DONE
FYP SEMESTER 1 SUBMISSION – 3rd March 2021					
CHAPTER 5: DATA ANALYSIS		20-May	1-Jul	42	DONE
5.1	Introduction	20-May	21-May	1	DONE
5.2	Initial Data Exploration	21-May	31-May	10	DONE
5.3	Data Cleaning	31-May	15-Jun	15	DONE
5.4	Data Visualization	15-Jun	18-Jun	3	DONE
5.5	Modelling/Data-driven product	18-Jun	28-Jun	10	DONE
5.6	Reports and Dashboard	28-Jun	30-Jun	2	DONE
5.7	Summary	30-Jun	1-Jul	1	DONE
CHAPTER 6: RESULTS AND DISCUSSION		1-Jul	5-Jul	4	DONE
6.1	Introduction	1-Jul	2-Jul	1	DONE
6.2	Model Evaluation	2-Jul	4-Jul	2	DONE
6.3	Discussion	4-Jul	5-Jul	1	DONE
CHAPTER 7: CONCLUSION AND REFLECTION		5-Jul	7-Jul	2	DONE
7.1	Conclusion	5-Jul	5-Jul	0.5	DONE
7.2	Reflection	5-Jul	6-Jul	0.5	DONE
7.3	Challenges	6-Jul	6-Jul	0.5	DONE
7.4	Limitations	6-Jul	7-Jul	0.5	DONE
7.5	Future Improvements	7-Jul	7-Jul	0.5	DONE
CHAPTER 8: REFERENCES AND APPENDICES		18-Dec	12-Jul	159	DONE
8.1	References	18-Dec	12-Jul	159	DONE
8.2	Appendices	18-Dec	12-Jul	159	DONE
8.3	CHECKING OF FYP DOCUMENT	18-Dec	12-Jul	159	DONE

Chapter 2: Literature Review

2.1 Introduction to Literature Review

This project employs exploratory data analysis to investigate the impact of mental illness and racial discrimination on fatal police shootings in America. To ensure the project is successful, it is important to investigate related topics to have an insight into the problem area in different aspects. One of the factors; Racism unfortunately still exists to this day and claims the lives of millions of people worldwide due to mass school shootings, hate crimes or police arrests. Moreover, people also get denied opportunities, education, medical treatment among other things that should not be withheld due to a person's skin colour.

This literature review will dive into the connection between fatal police shootings and police brutality, how mental illness and racial discrimination can be an influence on the former, the use of prediction and exploratory data analysis and its significance on the study; all of this research will help gain insights on the issues and help conduct the analysis.

Finally, research on similar analysis/systems that have been carried out to help gain valuable information to enhance our project.

2.2 Domain Research

2.2.1 Fatal Police Shootings

A police officer works within law enforcement and takes the oath to protect and serve the country and its citizens, prevent crimes and arrest criminals. However, this is not the case for the cops in the US. It has been observed through various sources that fatal police shootings have only increased in the past few years even with the added usage of body cameras and media attention. (Belli, 2020) The number of fatal shootings by police of innocent black people in the United States is more than three times higher than those of white people as found by the Journal of Epidemiology & Community Health with men of colour facing a lifelong threat of being murdered by the police. Black, Indigenous and People of Colour aka BIPOC are the victims most often so much so that the facts and figures have been defined as a "public health emergency". (Lett et al., 2020)

Nagin (2020) states that having access to weapons causes a higher instance of fatal police shootings. However, within the same article another author Cook presented that only 31% of US

citizens possess firearms and are law-abiding citizens who own the gun for legal purposes, though since firearms are fairly easy to access through illegal means in America, it may bring about more repeated encounters with law enforcement.

There are also instances of fatal police shootings in other countries such as Canada. Between Jan 1 and Nov 30, 55 people were killed by law enforcement officers. When talking about race, 48% were indigenous and 19% were black people. Families of the victims stated that some of the people suffered from mental illness or a drug habit and sent the police for wellness checks; 9 shootings out of the 55 people were wellness checks, with each of them being deadly and 4 of the shootings involving people of colour. (Malone et al., 2020)

There have been multiple types of research carried out concerning fatal police shootings and what factors may influence these events with various data analysis techniques. One example is a study carried out by Kivisto et al. (2017) aimed to find the link between firearm regulation and fatal police shootings and found that judicial restriction on firearms helped lower the fatal police shootings instances.

2.2.2 Police Brutality

Police brutality is defined as the use of extreme and unnecessary force on citizens such as shootings, harsh beatings, intimidation, verbal and psychological abuse, physical coercion etc. (Taylor, 2018) Based on international law, the police are only allowed to resort to lethal force as the last option in the case where there is serious imminent danger to themselves or to protect others from harm however this is not the case as in the US – George Floyd and Breonna Taylor, amongst other black people who were murdered by police while they were unarmed. (Amnesty, 2020)

According to Taylor (2018), despite all the media coverage of police brutality, most of them failed to mention that police brutality against BIPOC dates back even before the 1960s. The earliest mention of police brutality by the media is the murder of James Powell in 1965. Although a select few have mentioned police violence before the 1960s, they fail to mention the people efforts to stop the violence.

Perhaps the most prominent case of police brutality which also caused the Black Lives Matter movement to become internationally recognized were the murders of Michael Brown and Eric Garner. In August 2014, Michael Brown was shot down by a white officer after the officer told

him and his friend to use the sidewalk. After a bit of a scuffle, the officer shot Brown and the body remained on the street for four hours until it was taken away. Michael Brown was black, 18 and unarmed. (APNews,2019)

Police brutality in other countries such as Japan comes in the form of torture; Japanese police officers were alleged to have tortured suspects in jail to get them to confess, after being investigated it was concluded that they do indeed participate in physical and psychological torture; lack of privacy, having to ask for permission for each and everything such as lying down or getting up, constantly having their sleep interrupted, it seemed that they rarely used force but they threatened suspects instead. They also denied them access to lawyers which legally is the right of a person held in jail. (Geller and Toch, 1996)

2.2.3 Mental Illness as a Factor on Shootings

When reporting the fatal police shootings, the race and age of the victim are usually at the forefront while overlooking the fact that the person has either suffered or was suffering from mental illness. (Frankham, 2018)

According to Treatment Advocacy Centre, 7.9 million Americans deal with a mental illness that before would have been treatable through medications however over the past 50 years, the system that once would have given them the psychiatric treatment has been slowly demolished. It has been noted by various sources that a person is 16 times more likely than civilians to be at the risk of being killed by a police officer if they have an untreated mental illness. Furthermore, in 2015 people with a mental illness made up an uneven number of the individuals killed when approached by law enforcement. (Fuller et al.,2015)

International Bipolar Foundation states that in 2015, the NY Police Department handled over 12,000 mental health calls every month however it raised questions as to why EMT's were not handling this instead? Mental illness is surrounded by stigma such as being seen as violent and dangerous individuals so much so that people around them end up calling the police which may escalate to fatal encounters as most police officials are not trained to deal with a mental health crisis, and in actuality, mentally ill people are usually the ones who are the victims of an offence, not the one committing it. (Scout, 2015) Jim Fisher, a former FBI investigator and blogger

analyzed “lethal police shooting” in 2011 and found that over 25% of people who were killed in an encounter with the law enforcement were mentally ill. (Fuller et al.,2015)

In September 2020, Ricardo Muñoz was killed by a police officer after his family called the emergency line to help him get psychiatric care. (Sholtis, 2020) It is apparent that mental illness is not the only factor that is considered in these encounters however the link of mental illness to police violence should be addressed.

2.2.4 Racial Discrimination as a Factor on Shootings

3,453 individuals, including African Americans, Latinos, Asian Americans, Native Americans, whites, and LGBTQ adults, as well as men and women, were surveyed by researchers. Participants were asked if they thought they had ever personally encountered kinds of discrimination, both systemic and individual. The research showed that 45% of African Americans were unable to rent or buy apartments due to racial discrimination; 18% of Asian-Americans faced discrimination when interacting with law enforcement while Indian-Americans reported more unfair police treatment and stops than Chinese-Americans. Out of fear of being treated poorly,1 in 5 Latinos averted medical treatment. (RWJF, 2017)

In 2015, The Guardian reported that in that year alone out of 1,134 deaths, black males between the ages of 15 and 34 made up 15% of total death even though they made up roughly 2% of the total US Inhabitants. (Swaine et al., 2015)

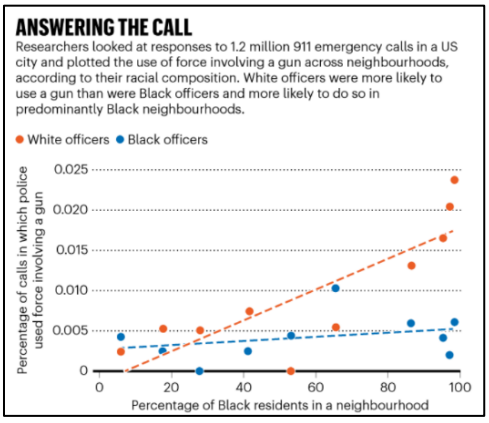


Figure 1 - Answering The Call (Hoekstra and Sloan, 2020)

According to research carried out by economist, Mark Hoekstra, he concluded, derived on data from more than two million 911 calls in two US cities, that white officers who were called to black

communities unloaded their arms 5 times as frequently as black officers sent to the same neighbourhoods for similar calls. (Hoekstra and Sloan, 2020)

"Even if overt racial bias, inherent racial bias, and discrimination by all police officers against Black individuals were eliminated, the racial disparity in fatal police shootings would remain". (Siegel, 2020, p1079) Siegel (2020) further goes on to say that the theory is that institutional racism in the violent use of force is indeed a result of structural racism, particularly urban segregation, resulting in law enforcement officers perceiving and regulating not just black people but black neighbourhoods in profoundly distinct ways. Muslim Americans also raised questions regarding police brutality, including abuse and racial discrimination, after Sep 11, 2001, attacks had occurred because several police officers had conducted illegal undercover operations to infiltrate mosques to find "terrorists" but were unsuccessful.

Even though other minorities also face discrimination, African Americans are the majority of the victims. While assessing certain factors, it was understood that the violence against African Americans came from the antiblack racism within the majorly white police branches. This is not the only factor that plays a part; certain police departments call for loyalty and brotherhood with a "show of force" approach to anybody who dares challenges an officer's power. Furthermore, the attitude and values of the antiblack infused groups are adopted by novice policemen who want to get promoted, feel accepted and successful as well as have an added ego boost. (Moore, 2020)

2.3 Exploratory Data Analysis and Visualizations

Originally developed in the late 1970s by US mathematician John Tukey, exploratory data analysis (EDA) is applied by data scientists to evaluate and examine data sets and summarize their key traits occasionally applying methods of data visualization. It assists to select the best method to exploit data sources to obtain the answers you need which in turn makes it easier for data scientists to find trends, spot anomalies, test a theory or verify assumptions. It is usually used to see what information can be understood from modelling or hypothesis testing and provides greater insight into variables in the dataset and the connection between them. It is significant in ensuring that findings that are produced are legitimate and precise and can be employed for any business outcomes.

Certain statistical techniques are performed such as K-Means clustering which is unsupervised learning where data points are allocated to every k-group. It typically employed in market segmentation, pattern recognition etc. Another analysis method that can be employed is the multivariate visualizations There are four types of EDA's: univariate graphical and non-graphical and multivariate graphical and non-graphical. Some multivariate graphics are scatter plots, bubble charts and heat maps, among others (IBM, 2020) Exploratory Data Analysis can be carried out through various tools and programs such as Tableau, Python, R, SAS Enterprise Miner, Power BI, Weka and more. Data exploring and preparation are important steps when exploratory data analysis is carried out. SAS Studio and SAS Enterprise Miner will be used for the two phases mentioned above to under what the data looks like; what variables are missing or irregular and have a complete understanding of the dataset. This will be discussed more in-depth in the methodology and data analysis chapters.

For this project, the enterprise application, Power BI will be utilized to deploy and visualize the relevant outcomes. The findings can be displayed through, Histograms – to see a specific variable distributes, Scatterplots – check the relationship between two variables, Feature correlation or heat maps – to show the link between numerous variables. Some of the variables that will be analysed are age, race, location, presence of mental illness. (Jabbari and Kienle, 2019). A few examples of the types of graphs that can be used in the project as well as an example of the application of the dashboard in Tableau is discussed in the next section.

2.3.1 EDA in Officer-Involved Shooting

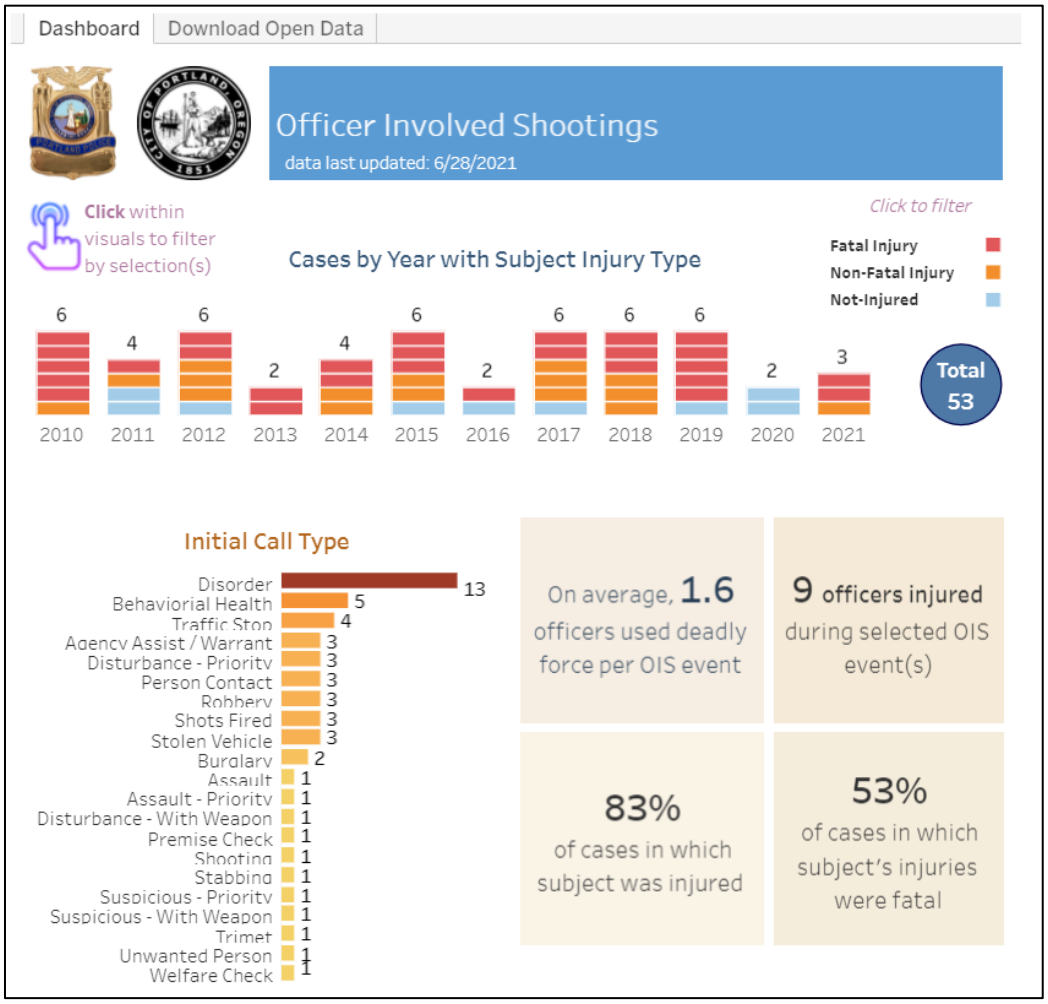


Figure 2 - Dashboard in Tableau (Portland Police Bureau, 2018)

An example of creating a dashboard for exploratory data analysis can be seen by the Portland Police Bureau who created an interactive dashboard to show deadly use of force by the police.

The graphs display the types of calls the officers received about each case as well as described the average number of officers who were involved in using extreme force. By creating such a dashboard, provides easy to understand and comprehend information for the public and can be used to show how important it is to bring change to these situations. (Portland Police Bureau, 2018)

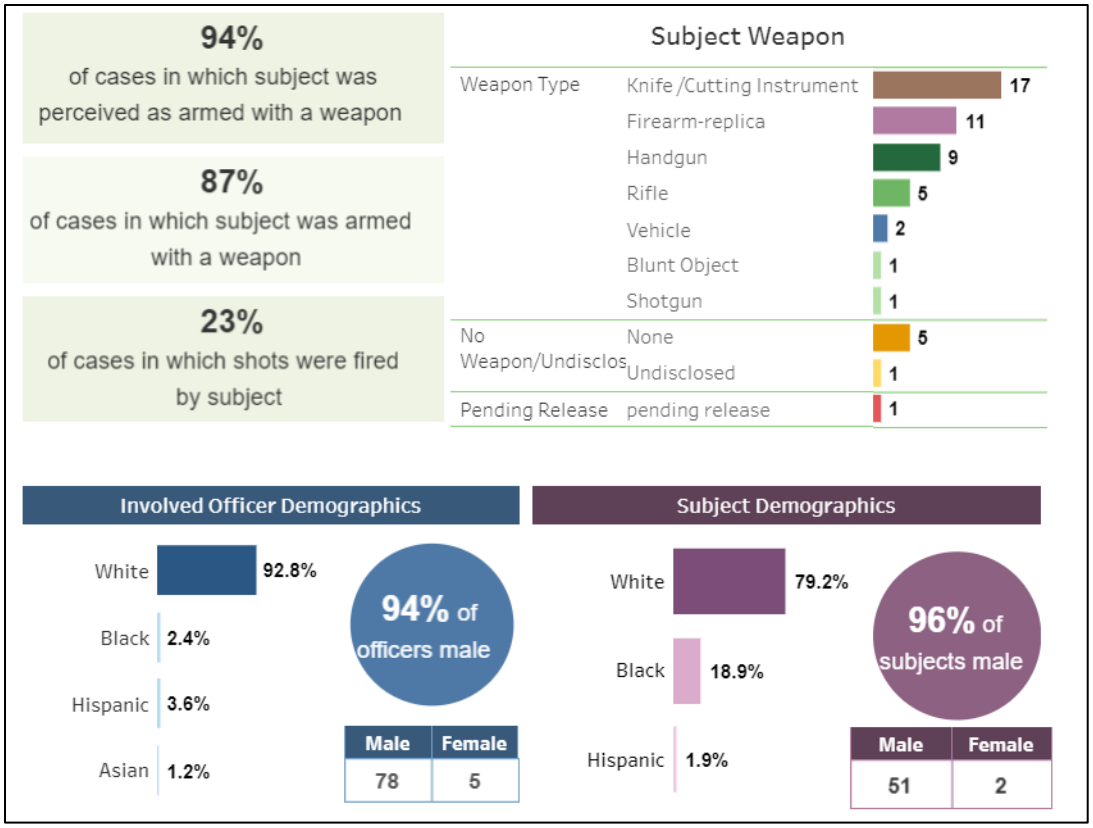


Figure 3 - Continuation of Dashboard

The dashboard also displays the types of weapons the victim may have been holding at the time of these events. Two bar charts for the officer-involved demographic as well as the subject demographic have been displayed. This has helped the law enforcement identify what the race, gender or age of the subject has been and how some of these factors affect these situations.

2.3.2 Visualization Techniques in PowerBI

Visualization in PowerBI can help provide actionable insight that can help change business functions. A dashboard in PowerBI can showcase a variety of charts such as bar plots, decomposition trees, histograms, pie charts etc. These can further be used on the Fatal Police Shooting dataset to showcase the cleaned dataset in various ways such as providing insights on the average age of the victims, the most common region and more. The examples of a few of the charts that may be used are shown below.

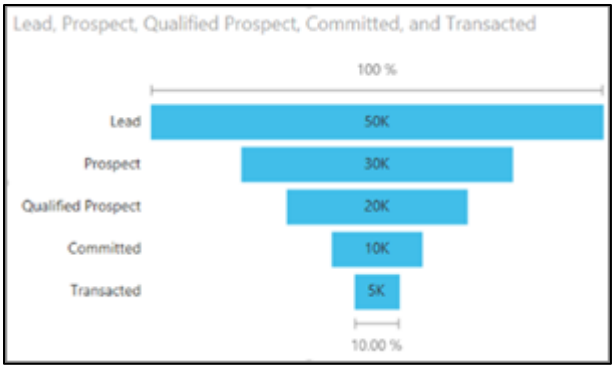


Figure 4 - Funnel Chart (Microsoft, 2019)

A funnel chart aids in the visualization of a linear process with concurrently linked phases. Evey phase of the funnel symbolizes a proportion of the total. In several cases, a funnel chart is transformed like a funnel, with the very first stage becoming the biggest and each following step is smaller than the last. A pear-shaped funnel can also help identify a problem in the process. However, the first phase, the "intake" stage, is usually the largest. (Microsoft, 2019)

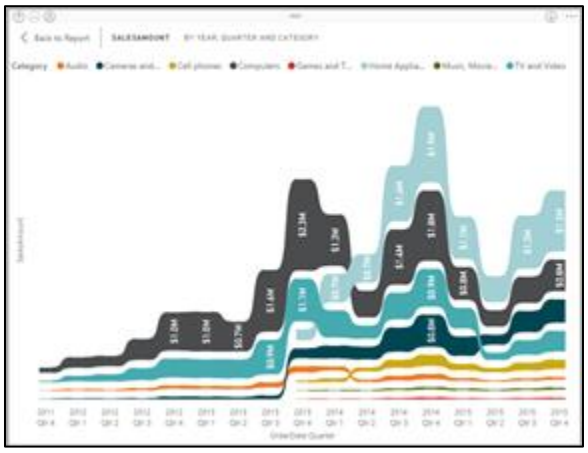


Figure 5 - Ribbon Chart (Microsoft, 2020)

Users can use ribbon charts to visually represent data and quickly determine which data category has the largest amount. Ribbon charts are good for displaying rank change because the highest range (value) is always showcased on top for each time frame. (Microsoft, 2020) For example, for the data analysis, this chart may be used to show the total number of victims who possessed arms at the time of the shooting and which weapon was the most occurring amongst the victims. This information can help initiate regulations of certain laws such as gun laws in the country to reduce easy access to weapons for common civilians.

2.4 Prediction Modeling

The process of detecting meaningful relationships between variables using pattern matching tools, statistics, algorithms, artificial intelligence, and data mining is referred to as predictive analytics. It encompasses the use of business intelligence, data mining and statistical methods to understand the likeliness of a certain event is happening based on the current/historical data. The most common predictive models include regression, decision trees and neural networks. (Patil et al., 2016) Under this sub-section, a discussion regarding the datatypes of the datasets, types of predictive models that can be built for the analysis of this project along with how they function in SAS Enterprise Miner as well explanation about certain algorithms or information that can be seen in the output will also be discussed. The implementation of these models in this project can help gain an understanding of the main factors that may influence these shootings and how do race and mental illness contribute to them. Is it the victim's age or region they were in that may have affected these situations?

One of the most useful prediction data mining techniques is known as classification which allows predicting of information from confirmed historical data. Classification may also be defined as a supervised learning method as the classes are already established before the data is examined. KNN, Decision Trees, Regression, Clustering etc. are all algorithms and techniques which come under the label of classification. (Bhardwaj and Pal, 2011)

2.4.1 Data Types of Variables

Each attribute has a data type and role in a dataset. Data types can be nominal, binary, ordinal, ratio, interval or unary. Nominal describes a scale consisting of categorical variables such as gender, city, the county with no order of importance between them. These can be both on the categorical scale or in a dichotomous scale, which means that there are only two categories.

Ordinal is similar to Nominal however there is rank for the variables such as top 10 flavours of ice cream or best-to-worst places to travel, these all have a rank. Interval shows an order between the numerical variables such as age, test scores, etc. The ratio is similar to Interval however it has a true zero value unlike Interval and there are points on its scale that make sense as ratios, taking the example of age, there can be someone who is 0 years old and if someone is 2 years old, they are 2 times younger than someone who is 4 years old. Interval and Ratio can be collapsed into one category of continuous variables. (Brown, 2011) Binary variables are those that are either true or

false, yes, or no, occurred or did not occur but may also be applied to those attributes that have more than two categories and need to be changed into a binary variable for carrying out analysis. Unary variables are those which have the same value in the whole column such as '1' or 'TRUE'.

2.4.2 Logistic Regression

Regression analysis is a series of statistical techniques applied to approximate the correlations among one or more independent variables and a dependent variable. It can be used to establish the depth of the connection between variables and model the relationship between them in the future. (Trinidad, 2020) Two types of regression will be discussed below.

Logistic regression is very similar to the linear regression modelling process however unlike linear regression, the dependent target variable is not continuous, yet is discrete or a categorical variable. This is very useful when predicting outcomes such as whether a person can have a heart attack or not, will a person purchase an item or not and more. A logistic regression model can be used to model target variables with more than two levels (Parr-Rud, 2014); a link between a set of independent factors and a nominal dependent variable is modelled using nominal logistic regression, often referred called multinomial logistic regression. A nominal variable is a variable that may have a minimum of three categories with no specific order. Logistic regression can be used to address various classification issues such as binomial, multinomial ordinal and cardinal classification. It enables viewing of important insights which help in the research moreover, it is easier to train and apply in comparison to other models. And works well with a dataset that is linearly *independent* which means that when a straight line is drawn, it separates two classes of data from the other. However, logistic regression is not able to forecast a continuous outcome and becomes erroneous for smaller sample sizes. (Thanda, 2020)

Cheney (2020) carried out an analysis to understand deadly police encounters through various types of machine-learning models, one of which was logistic regression to deal with the initial class imbalance in their exploratory data analysis. Other models they used were KNN, SVC, Naïve Bayes, and Random Forest Classifiers. Frankham (2017) used multinomial logistic regression to analyse the fatal shooting of a person with the variable of race and the way the encounter was initiated.

2.4.3 Linear Regression

The result of any analysis can be displayed with a linear model while linear regression can be used to ensure that this model is deployed. Linear regression is defined as a technique in which the input variables (x) and the single output variable (y) have a linear relationship. It is a supervised machine learning algorithm where unlike logistic regression; the expected output is continuous with a constant slope. Depending on the number of input variables, it can either be referred to as simple linear regression or as multiple linear regression. (Brownlee, 2016)

Linear regression function as such that the intercept is used to centre the prediction and the left-over parameter estimates define the trend strength among the dependent and independent variables. The model causes all the predicted values to appear in one direction. The intercept and the parameter estimates are selected to decrease the 'squared error' between the predicted and observed target values which are also known as a least-squares estimation. These predicted estimates can be considered as a linear estimate of the average value of a target trained on observed input values. (Georges et al., 2010) Linear regression is simple and easy to implement but can be susceptible to noise, overfitting, and outliers which can be optimized using techniques such as regularization and gradient descent. (Waseem, 2019)

Ryan (2019) applied multivariate regression to evaluate the dependent variable which is shooting rate versus the independent variables such as health factors (i.e., mental illness). Clark et al. (2020) used a linear probability model to analyse the lower bound of black people who would not have been shot had they been white people. They carried out the analysis with the logit model as well as the difference between both analyses for logit and linear model were negligible.

2.4.4 Regression Model Representation in SAS Enterprise Miner

Linear and Logistic Regression can be modelled in SAS Enterprise Miner using the regression node in the model category of SAS's data mining procedure. The default settings are kept as logistic regression with the logit function. the type of model can be selected as stepwise, backward, forward or none.

Stepwise regression allows the input variables to enter the model, the model can then remove them at any step, in case they do not meet the significance value anymore. Forward regression is a model with increasing complexity, the inputs values are entered as long as they meet the p-value's

significance levels and start with the most significant while Backward regressions start with all the inputs in the model and remove the least significant inputs one by one until it reaches a point where a rule is kept or there are no more inputs left. (Choueiry, 2021)

In SAS Enterprise Miner, the link function under the regression nodes property can be used to indicate which link function can be utilized. The linear predictor is where the link function relates the response means. The link function in the regression node in SAS EM helps the user select the appropriate function for the regression modelling that will happen. Under linear function the formula for the identity link function is:

$$y^* = \beta' x + U$$

Figure 6 - Link Function Formula (Sarma, 2017)

Y = latent variable of a record
β = Vector Coefficient
x = explanatory variable such as victims age.
U = is the random variable in the equation

U may generate different functions due to the different assumptions that may be made due to how U is distributed. (Sarma, 2017)

The equation generated may look as shown below

$$\Pr(y=1|x) = \Pr(y^* > 0|x) = \Pr(\beta' x + U > 0) = \Pr(U > -\beta' x)$$

Figure 7 - Probability of Response

Y is the target variable here such as race or signs of mental illness.

$$\Pr(y=1|x) = \Pr(U > -\beta' x) = 1 - F(-\beta' x)$$

Figure 8 - Simplified Equation for Probability of Response

F is the *CDF* or the cumulative distribution function for the random variable U (Sarma, 2017)

In Logistic regression – logit, cloglog and probit link functions may be used.

1. **Logit** specifies the default and indicates "the inverse of the cumulative logistic distribution." → g(M) = log (M/1-M)

In this case, the logit is selected the CDF is seen as below.

$$F(-\beta' x)=\frac{1}{1+e^{\beta' x}} \quad \text{and} \quad 1-F(-\beta' x)=1-\frac{1}{1+e^{\beta' x}}=\frac{1}{1+e^{-\beta' x}}$$

Figure 9 - Logistic Regression Equation

$$\Pr(y=1\mid x)=\frac{1}{1+e^{-\beta' x}}=\frac{e^{\beta' x}}{1+e^{\beta' x}}$$

Figure 10 - Probability of Response

$$\log\left(\frac{\Pr(y=1\mid x)}{1-\Pr(y=1\mid x)}\right)=\beta' x.$$

Figure 11 – Link Function

β'x is the linear prediction as it takes the inputs linear combination.
Y is the target variable and x will be known as the explanatory variable
1/1+e $^{-\beta' x}$ is the equation of regression which comes from the cross multiplication in the equation and generates the final link function as shown in figure 11.

2. **Cloglog or complementary log-log function** "specifies the inverse of the cumulative extreme-value function." → g(M) = log (-log (1-M))

$$\Pr(y=1\mid x)=1-\exp(-\exp(\beta' x))$$

Figure 12 – Probability of response

β'x and *y* are the same as the above equations.

3. **Probit** specifies "the inverse of the cumulative standard normal distribution." →g(M) = 1/ ΘM (SAS, 2021)

In the case, probit is used as the link function property, then the U value that was mentioned above will have a normal distribution where its mean will be 0 and its standard deviation will come out as 1. The probability of response equation for it will be as below

$$\Pr(y=1\mid x)=1-F(-\beta' x)=F(\beta' x)$$

$$\text{where } F(\beta' x)=\int_{-\infty}^{\beta' x}\frac{1}{\sqrt{2\pi}}e^{-u^2}du$$

Figure 13 - Probit probability of response

The regression node here uses the probnorm function to determine the probability of the equation hence it is computed *y = 1* as

$$\Pr(y=1\mid x)=probnorm(\beta' x)$$

Figure 14 - probnorm

Where *probnorm* gives *F(β'x).* (Sarma, 2017)

After the regression node is run, we can see the below information in the output window.

Fit Statistics

The fit statistics table is shown as an output when the results of the model are displayed, it shows the measure of the created model's goodness of fit with various evaluation metrics such as RMSE, MAE, ASE and more. Some of these discussed in the evaluation metrics subsection.

Maximum Likelihood Estimates

Logistic regression can use various algorithms for example maximum likelihood estimation. According to Brooks-Bartlett (2018), It can be defined as "Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that was observed." Although, the MLE works best when the dataset is large and not small as it would not produce the best output with small data. The analysis of maximum likelihood estimates display the estimates (regression coefficients), standard errors, the Wald Chi-square value, and their p

values. The estimates inform us of the most important inputs that are related to the output and whether the link between them is a positive or negative relationship.

Type 3 Analysis of Effects

In SAS Enterprise Miner's regression output the Type 3 analysis of effects displays the hypothesis tests for selected inputs along with their Wald Chi-Square values and p-value., A value nearer to 0 in the Pr > ChiSq column implies a significant input while a value closer to 1 suggests an extraneous input.

Odds Ratio Estimate

An odds ratio (OR) is a measure of the relationship between an exposure and a result. The OR predicts the number that an outcome will occur given a specific exposure, as opposed to the probability that the outcome will occur in the absence of that exposure.

The regression coefficient (b1) in logistic regression is the estimated increase in the log odds of the result per unit increase in value of the exposure. It can also be said that the odds ratio associated with a one-unit increase in exposure is the exponential function of the regression coefficient (eb1). The odds ratio in logistic regression denotes the effect of a predictor X on the likelihood that one event will happen. (Grace-Martin, 2012) For the logistic regression models, it may help in finding out which factors are important alongside the targets. (Szumilas, 2010)

Wald Chi-Square

Establishes whether or not one of the regression coefficients of the predictors is not equal to 0 within the model.

2.4.5 Decision Trees

Decision trees are used in data analysis and machine learning for predictive, descriptive analytics as well as classification and regression (Hillier, 2020) and are non-parametric models which means that they are adaptable models that do not expand the number of parameters as more characteristics are added. (Chauhan, 2020) The purpose of this technique is to develop a model that predicts the value of a target variable, and the decision tree solves the problem by using the tree representation,

where the leaf node corresponds to a class label and characteristics are displayed on the internal node of the tree. (Sharma, 2021)

Because they are both simple to learn and effective, decision trees are often the tool of choice for predictive modelling. A decision tree's main purpose is to divide a large amount of data into smaller sections. Prediction is divided into two stages. The tree is generated, evaluated, and optimized using an existing set of data at the first stage of the model's development. The model is then used to forecast an uncertain outcome in the second phase. (Aunalytics, 2015)

Decision trees can use various types of algorithms to help decide the splitting of the nodes. As every sub-node is created, the subsequent nodes become homogeneous in conjunction with the target variables. The algorithms depend on the type of the target variable selected and are listed below.

ID3/C4.5 – the latter is the successor to the former. ID3 utilizes entropy and info gain to specify the split rules and only functions with categorical targets and predictors while C4.5 functions well with categorical and continuous variables, it can handle missing information and aids the user to state the cost of error, it also uses a pruning function.

- Entropy can be known as the estimate of the unpredictability of the data being processed. The greater the entropy, the more difficult it is to draw a valuable conclusion from the data.

$$H(s) = -probablity\ of\ \log_2(p+) - -probability\ of\ \log_2(p-)$$

Figure 15 formula for entropy where (p+) is %of positive class and (p-) is % of the negative class. (Sharma, 2021)

- Information Gain allows users to evaluate at which feature the split should be occurring at every step of the building of the tree. At every node, the value of information measures how much data a feature is providing us for the class and hence the split with the largest info gain will be the first split and so on until the child nodes become pure or the value of info gain reaches 0. (Sharma, 2021)

$$Gain(S,A) = H(s)\sum \frac{/Sv/}{/s/} H(Sv)$$

Figure 16 information gain formula (Sharma, 2021)

CART – (Classification and Regression Tree) is known as a non-parametric algorithm that understands knowledge and justifies decision tree models. It performs similar to the CHAID algorithm however instead of the chi-square test it uses the GINI index. CART also function with both categorical or continuous target variables which are for classification and regression trees, respectively. CART has a benefit over CHAID as it can create a much more precise model than CHAID. (Informit, 2014) The limitation for CART is its proprietary algorithm.

- Gini Index is a cost function that is utilized to analyse dataset splits. It is computed by deducting one from the total of each class's squared probability. (Chauhan, 2020) This can also be known as the Gini coefficient that is between 0 and 1 with them representing perfect equal and perfect inequal, respectively. In case the index value is high, the data is scattered. (Choudhury, 2019) When the Gini is 0, it means it is a pure node. As the targets for this project are binary, the criterion for this is set under the Nominal target criterion tab in SAS Enterprise Miner. The value of p denotes the square of the number of the class by all the cases within a node. Gini Index may also be known as Gini Coefficient or Gini Ratio and can be used to evaluate prediction models in SAS EM.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Figure 17 (Chauhan, 2020)

CHAID – (Chi-square automatic interaction detection or Chi-Square criterion in Enterprise Miner) This algorithm functions only with target variables that are categorical. The algorithm calculates a chi-square test among the target variable and all available predictors before partitioning the sample with the best predictor. This is repeated with every segment until no large splits are left. A limitation for CHAID is the restriction on categorical data. (Informit, 2014) The chi-square technique is explained below in the evaluation technique.

Mashinchi (2020) applied decision trees for their classification problem to predict if the victim was mentally ill, they split their dataset into 80% training and 20% testing. They also used random forest, SVC, etc and cross-validated to improve the overfitting of certain models.

In SAS Enterprise Miner, the decision tree can be modelled by taking the decision tree node under the model tab. In the decision tree, there are node train properties where we can place our splitting rules depending on the type of target criterion. Entropy or Gini Index can be selected to carry out a reduction in the Entropy or Gini index measure along with the type of assessment method we want to utilize to carry out the process, this includes Misclassification Rate, Lift and Average Square Error or Default. The Assessment method also allows for the selection of the largest tree or the maximal tree, 'N' which allows us to specify at what leaf we want the model to stop and lastly the Assessment method itself which chooses the most optimal and pruned tree.

2.4.6 Random Forest Decision Tree

The random forest algorithm takes many of the decision trees and combines them into one. It is employed for supervised learning however is very strong and commonly used. It uses more than one decision due to taking multiple decision trees and arranges the decision depending on other decisions and comes up with the final decision cantered on the majority. It has a smoother prediction, due to making various predictions instead of finding the best one. (Vadapalli, 2020) It can be used in both classification and regression problems. Moreover, it does not have many overfitting problems as if the number of trees is more than enough, then the model will not be overfitted. However, it does have issues with time as having to use so many trees can cause it to become slow and problematic for real-time forecasts. If the prediction is important, then using a random forest would be an issue but if the time is more important, then other faster methods would be used instead. (Donges, 2019)

Streeter (2019) used various machine learning approaches to predict race and to figure out which variables distinguished between black and white people. They used random forest with the classification outcome as race and used the accuracy measure known as the "out-of-bag" error rate. This was the average proportion of wrong estimates made by the decision tree from the sample observation.

2.5 Evaluation Techniques

This section discusses the various evaluation metrics that may be used to evaluate the models in terms of accuracy in prediction in the data analysis chapter.

2.5.1 R-Squared(R^2)/Adjusted R-Squared

The R-Squared value determines how much variability is provided by the dependent variable that can be justified by the prediction model. Moreover, it measures the depth of the link between the target and the model. The regression error can be computed using the value of the input variables. To note, the value of R2 falls between 0 and 1 and the higher the value, the greater the variability.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Figure 18 - R-Squared Formula

The ratio of the residual error (RSS) over the total error (TSS) informs the user about the remaining error within the model. The lower the value of the R-Squared, the better the fit of the model between the prediction and actual value. (Wu, 2020)

2.5.2 Mean Absolute Error (M.A.E)

This is similar to M.S.E but instead, take the sum of the absolute value of error. In comparison to M.S.E, M.A.E does not punish large errors while M.S.E does. (Wu, 2020)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

Figure 19 – Mean Absolute Error (Wu, 2020)

1/n is the division by the total number of data points while Ŷ is the actual output value and Y is the predicted output value. (Pascual, 2018)

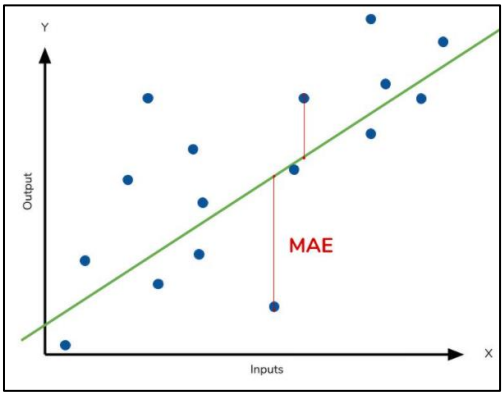


Figure 20 - MAE line (Pascual, 2018)

This is the graph that is displayed when the MAE is predicted with the line representing the prediction and the dots representing the actual data points of the data. Although MAE is the easiest interpretable evaluation metric, it does not consider the overfitting and underfitting of a model and hence may not be the best evaluation metric to use for the models if there are many outliers. (Pascual, 2018)

2.5.3 Mean Square Error (M.S.E)

Mean Square Error is an absolute measure of the goodness of fit. The M.S.E is computed as the sum of the square of prediction error, which is the actual output minus the predicted output, then divided by the number of data points. It gives an absolute value on how much the expected outcomes vary from the actual figure. Most of the understandings from a single outcome cannot be inferred, but it provides a real number to compare with other model results and assist in selecting the best regression model. R.M.S.E is the root of M.S.E and is often used more often than the latter due to easier interpretation. (Wu, 2020) M.A.E and M.S.E cannot be used in comparison because the latter is always squared while the former is not hence will have the bigger value of the two. Moreover, the presence of any outliers in the data may cause the M.S.E to have a higher value as M.A.E does not take any outliers into account in the calculation. (Pascual, 2018)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Figure 21 – Mean Square Error (Wu, 2020)

(Ŷ – Y) provides the user with the square of the difference between the actual and predicted value.

2.5.4 Root Mean Squared Error (R.M.S.E)

This provides the user with information about how close or far the model is from providing the correct answer and shows the mean of the prediction error within the same scale. It is calculated as the standard deviation of the MSE and can be seen in the formula below. The closer the value is to 0, the better the model is. (Borah et al., 2019)

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum \epsilon_i^2}$$

Figure 22 - the ϵi represents the sum of the squared mean from MSE. (Borah et al., 2019)

2.5.5 Confusion Matrix

An N x N (N is the number of target classes) matrix used in the evaluation of the performance of a classification model. It compares the actual target values to the machine learning model's predictions. This provides us with a comprehensive picture of how well our classification model is working and the types of errors it makes.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Figure 23 - rows represent predicted while columns represent actual values (Choudhury, 2019)

1. TP – this is true positive where the predicted is equal to the actual (positive) value and the model has predicted a positive value.
2. TN – this is a true negative where the predicted matches the actual (negative) value and the model has predicted a negative value.
3. FP – this is false positive where the predicted value was incorrectly predicted since the actual value was negative however the model estimated a positive value, and this is further referred to as a Type 1 Error.
4. FN – this is false negative where the predicted value was incorrectly predicted since the actual value was positive however the model estimated a negative value, and this is further referred to as a Type 2 Error.

A confusion matrix is very helpful in evaluating the performance of the models by measuring Recall, Precision, F1 Score and AUC-ROC curve. Recall informs us how many positive classes were predicted accurately and needs to have the highest value possible.

$$Recall = \frac{TP}{TP + FN}$$

Figure 24 - (Narkhede, 2018)

Precision informs us how many of the positive classes that we have predicted are positive.

$$Precision = \frac{TP}{TP + FP}$$

Figure 25 - (Narkhede, 2018)

F1-Score enables us to have a comparison between two models which have high precision and low recall or vice versa, this is where F1-Score aids in measuring both scores at the same time.

$$F\text{ - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

Figure 26 - (Narkhede, 2018)

AUC-ROC is one of the most commonly used evaluation metrics for classification models and allows for measurements at various threshold settings. Curve uses (TPR)True Positive rate (Recall) and False positive rate (FPR) to plot the graph. AUC is the area under ROC and can be used to measure the classification models quality. (Choudhury, 2019)

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Figure 27 (Choudhury, 2019)

2.5.6 Chi-Square Test of Independence

Chi-square criterion is used to construct the CHAID in SAS Enterprise Miner, the formula for chi-square is shown below. The greater the chi-square value, the greater the difference between the parent and child node. (Chauhan, 2020) The link between two categorical or nominal variables is found through the chi-square test of independence which is a statistical hypothesis test. (JMP, 2021)

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where:

χ^2 = Chi Square obtained
$\sum$ = the sum of
O = observed score
E = expected score

Figure 28 - formula for chi-square tests (Chauhan, 2020)

2.5.7 Misclassification Rate (M.I.S.C)

Misclassification rate is when a number of the predictions made are misclassified and not accurate. This can, in turn, be used to find the accuracy of a prediction model by

$$1 - the\ value\ of\ the\ Misclassification\ rate\ x100$$

Figure 29 - Accuracy

Misclassification Rate is calculated by the sum of the false positive and false negatives with the total sum of instances. The lower the value, the less misclassified the model and the more accurate it is. The formula for this is shown below. (Borah et al., 2019) The meaning of false positives, false negatives, true positives, true negatives, precision and accuracy have been discussed under the confusion matrix subheading.

$$\text{Misclassification rate} = \frac{\text{False positives+false negatives}}{\text{Total instances}}$$

Figure 30 - formula for misclassification rate (Borah et al., 2019)

2.5.8 Average Squared Error (A.S.E)

This is the mean value of the square of the variation between both the predicted and actual outcomes. In the regression model, it is known as the validation error assessment measure which allows the selection of the model with the lowest error rate while for decision trees, it chooses the tree with the lowest or smallest average square error.

y_{it} is the actual outcome from the validation dataset while $\bar{y}_t$ is the predicted outcome. i^{th} is the number of the record while t^{th} is the number of the node. (Sarma, 2017)

$$\sum_{t=1}^{T} \frac{n_t}{n} \left[\frac{1}{n_t} \sum_{i=1}^{n_t} (y_{it} - \bar{y}_t)^2 \right].$$

Figure 31 - ASE equation (Sarma, 2017)

2.5.9 Cumulative Lift

Instances in the training and validation data are listed in descending order of forecasted target value. A subset of the ranked data is chosen. This small portion, or decile, correlates to the chart's horizontal plane. Cumulative lift is defined as the ratio of (the proportion of cases with the primary outcome in the selected fraction) to (the proportion of cases with the primary outcome overall). The vertical axis corresponds to cumulative lift. Relatively high lift values indicate that the model seems to be doing a good job of distinguishing between main and secondary cases. (Georges et al., 2010)

2.6 Similar Systems

In this section, there will be a discussion regarding similar research/systems. A brief overview of each system is given and then all the systems are compiled in a table that lists their titles, authors, year published, data mining and evaluation techniques used results and limitations.

1. Mesic et al (2018) used Poisson regression to evaluate the effect of structural racism on Black-White inequality among states in fatal police shootings (FPS) concerning victims who were not reported to be armed from January 1, 2013, to June 30, 2017. They created an index; state racism which has five elements: residential segregation, gaps in incarceration rates, educational attainment, economic indicators, and employment status.
2. Ross (2015) used multi-dimensional Bayesian modelling to examine the degree of racial profiling in the shooting of American citizens by law enforcement officers in recent years utilizing the U.S Police Shooting Database (USPSD).
3. Ryan (2019) used SPSS v.25 to carry out cross tab analyses to establish the bivariate relationships among variables to find out which element indicated the greatest association as well as bivariate regression analysis to find out which factors like mental illness contribute to threat level.
4. Schulberg (2020) carried out exploratory data analysis using R to analyze U.S Police caused fatalities where 20 years' worth of police killings were collected from various datasets and made into one comprehensive dataset.
5. Schwartz and Jahn (2020) utilized Poisson Model and MePoisson command in Stata to estimate annual incident rates and incident rate ratio for police killing in Metropolitan Statistical Areas (MSAs). Sensitivity analysis was used to compare models they created using

mepoisson and lastly used statistical software GeoDa to exhibit spatial autocorrelation including K-nearest neighbour for classifications.

6. Ruess (2019) used the fatal force database and utilized Content Analysis to gather more information from news articles. Exploratory content analysis and statistical analysis techniques were carried to investigate whether situational factors were different with black and white suspects.

7. Wang and Fan (2021) used Weka ML software and carried out various prediction models to predict shooting rate based on state location and predict a victim's race in a fatal police shooting.

Comparison Table of Similar Research

No.	Related Work, Description	Author/s	Dataset	Data Mining Methodology	Evaluation Method	Result/Outcome	Limitations
1	The Relationship Between Structural Racism and Black-White Disparities in Fatal Police Shootings at the State Level.	Aldina Mesic, Lydia Franklin, Alev Cansever, Fiona Potter, Anika Sharma, Anita Knopov, Michael Siegel, 2018.	1.U.S. Police Shootings Database, 2.FatalEncounters.org, 3.KilledbyPolice.net.	Statistical Analysis with Stata Poisson Regression	Regression Coefficient	In every 10-point rise of state racism index, the black-white difference of officer-involved firings for people not known to be armed increased by 24%. Nationally, blacks were shot 3.1 times more than police and unarmed blacks were shot 4.5 times more than white people.	1. Police shooting rates are unstable due to a small number of cases. 2. Analysis limited because of starting time of compilation was from 2013. 3. Disparities amongst another ethnicity not investigated. 4. Investigation of city-level and urban area-specific differences not carried out. 5. Allegedly armed victims were not considered to be unarmed in analysis. (Mesic et al., 2018)
2	A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011–2014.	Cody. T Ross, 2015.	U.S. Police-Shooting Database (USPSD)	1.Multi-level Bayesian Modelling using R.	Regression coefficients, Correlation matrix, Cholesky factor, L	1. The findings indicate a major bias in shootings with unarmed black citizens being 3.49 times more likely to be shot as opposed to unarmed white citizens (unarmed and shot)	1. More detailed analysis is needed as there are differences in certain categories that should be taken into consideration. 2. Higher quality covariate data is required. (Ross, 2015)
3	Lethal Use of Force: Insights into Mental Illness.	Robert A. Ryan, 2019.	WAPO Fatal force database	1. Cross Tab Regression Analysis	Adjusted R-Squared	1.Individuals with mental illness posed threat but were less likely to flee or attack the police	1. Comprehensive data set but did not provide all types of fatal police shootings.

				2.Bivariate Regression Analysis 3.Multi-variate Regression Analysis 4.Descriptive Statistics.			2. Some of the incidences were coded by individuals who were not at the scene for example for threat level – it did not mention that suspects were charging at officers and refused to cooperate and ended up being categorized as other. (Ryan,2019)
4	U.S. Police-caused Fatalities.	Justin Schulberg, 2020.	5 different datasets of individuals killed by police from 2000-2020: demographic information taken from 2014 census bureau data	Exploratory Data Analysis in R	N/A	1. 2015 showed the biggest rise in police-involved shootings, possibly due to the Ferguson protests 2. Throughout the years' black people were being killed at higher rates than any other race	1. Imperfect data collection 2. Bias in data collection 3. Under-reporting of data there may be bias in releasing reports regarding police brutality. (Schulberg, 2020)
5	Mapping fatal police violence across U.S. metropolitan areas: Overall rates and racial/ethnic inequities, 2013-2017	Gabriel L. Schwartz, Jaquelyn L. Jahn, 2020.	FatalEncounters.org (taken from online media reports and public records)	1. Multilevel Poisson Models	mepoisson method, Sensitivity Analysis Spatial Correlation Regression	1. Southwestern Metropolitan areas had the highest rate of police-related fatality and lowest in Midwest and Northeast but with black-white inequities, the result was reversed. 2. The result excluded death from possible accidents 3.Across all MSAs, black people were 3.23 more likely to be killed than whites and Latinx were 1.05 more likely.	1. Possible misclassification of ethnicity as it is not self-reported so POC may be classified as white when reported by media or officials. 2. Cause of death may also be misclassified that can bias the result. (Schwartz and Jahn, 2020)

6	Situational Context of Police Use of Deadly Force: a Comparison of Black and White Subjects of Fatal Police Shootings.	Shana Lynn Meaney Ruess, 2019	1. WAPO Fatal force database 2. News articles through content analysis	1. Exploratory Content Analysis in Atlas ti. 2. Statistical Analysis	t-Test and chi-square tests of independence	1. Black people are killed disproportionately to white people. 2. Black men over 45 who were killed were experiencing mental illness issues. 3. White people who were killed were actively committing a crime and also were suffering from mental health issues. 4. There were certain differences in age, mental illness and threat levels between the black and white people.	1. Inadequate and partial dataset 2. Some instances are not similar such as the use of deadly force with subject alive with cases of non-lethal force 3. Inaccurate reporting of data intentionally by journalists causing inaccurate results. (Ruess,2019)
7	US Fatal Police Shooting Analysis and Prediction	Yuan Wang and Yangxin Fan, 2021	Washington Post Fatal Police Shooting Dataset KilledByPolice Dataset	Pattern Mining, Correlated variables analysis Prediction	Confusion Matrix	The discrepancy in new media reporting and instigating hostility in police departments Shootings depend on various variables	They were not able to find more influence of factors which would have given more conclusive results No evidence of racial bias (Wang and Fan, 2021)

2.7 Summary

In conclusion, this chapter was used to research the domain area for fatal police shootings, police brutality, mental illness, and racial discrimination as a factor in fatal shootings as well as finally discussing some similar research projects carried out.

It provided a clear understanding of the problem at hand and how carrying out prediction models and exploratory data analysis will be used to complete this project successfully. By researching similar systems, it showed various ways that other researchers implemented their analysis by using data mining techniques such as regression, multi-variate Bayesian analysis etc.

According to the literature review conducted, it is important to analyse how mental illness and racial discrimination play a part in officer-involved shootings. This analysis can then be utilized to stress the need for psychiatric care for the mentally ill and proper legislation laws to avoid illegal firearm availability that influences both fatal police shootings as well as other crimes as well implementation of other laws such as the application of body cameras in every encounter with civilians.

Chapter 3: Technical Research

In this chapter, the tools and techniques used to conduct the project will be discussed along with their aforementioned features, benefits, and disadvantages.

3.1 IDE (Integrated Development Environment)

An IDE, or Integrated Development Environment, allows coders to merge the different aspects of programming into one convenient location. IDEs improve programmer efficiency by integrating popular software development tasks such as editing source code, creating executables, and debugging into a single application. Without an IDE, a developer must choose, implement, incorporate, and handle all of these tools individually. The integrated toolset is intended to streamline the development of software and can detect and reduce coding errors and spelling mistakes. Various types of IDEs suit a developer's needs, they range from cloud-based to mobile, language-specific, or web-based. For example, HTML, JavaScript, Microsoft Visual Studio Code – which is a web-based IDE. NetBeans, Eclipse, IntelliJ IDEA are more of the popular IDEs used by developers as they allow a multi-language-based IDE. (Gillis and Silverthorne, 2017) For this project, SAS Studio, a web-based IDE will be utilized.

3.1.1 SAS Studio

SAS Studio is a web-based development environment for SAS that can be accessed via a web browser. Usually, professionals' program in SAS using software on their PC desktop or SAS server. SAS Studio is unique in that it is a program that allows developers to write and execute SAS code directly from their internet browser. SAS Studio enables users to access their metadata, libraries, and current programs, as well as build new ones. To produce SAS code, users can also use SAS Studio's preconfigured tasks (SAS, 2016)

SAS Studio launches on an internet browser every time users run it. Once users input their SAS programs, the code is routed to the SAS server connected with the SAS Studio session, and the outcomes are displayed on the user's browser. The SAS server could be a personal computer, a server on the network, or a cloud-hosted server. (Delwiche and Slaughter, 2018)

Figure 32 - (SAS, 2016)

What makes SAS Studio unique is that it runs in a browser and has easy to access and understand tools that can greatly benefit even a novice programmer. Within SAS Studio, the programmer can convert the task into a SAS Program to easily add or edit the code that has been created by the program.

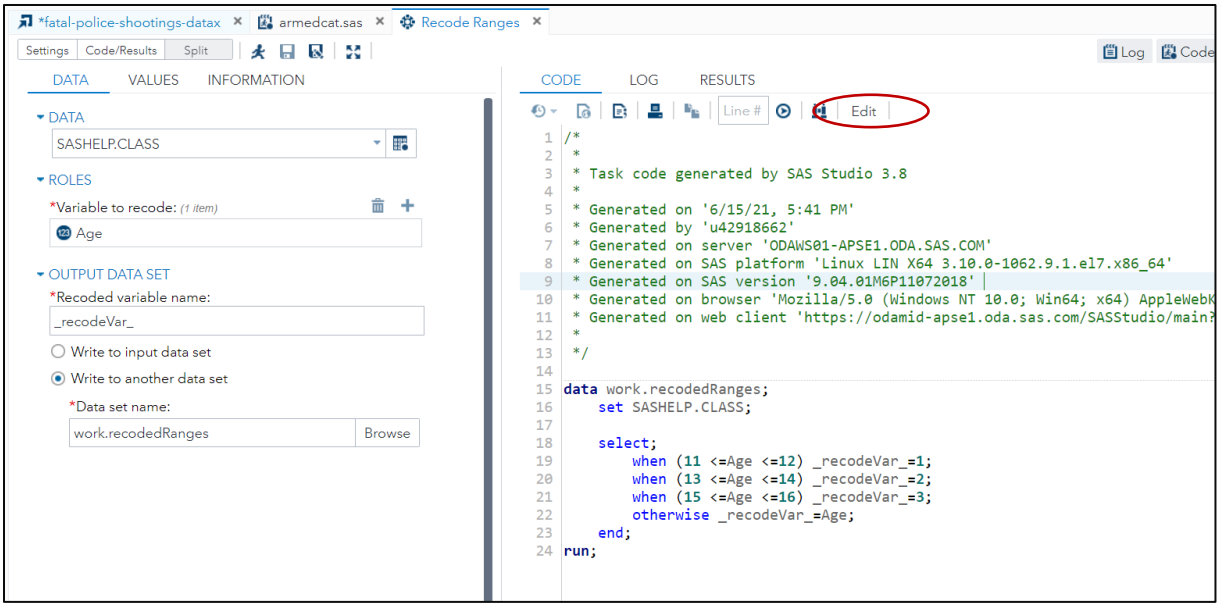


Figure 33 - Screenshot from SAS Studio showing 'Recode Range' task

In the screenshot above it can be seen that when a user enters a value they would like to recode into a different variable, the SAS code is shown on the right. The user may click on the code and edit it to their liking and run the output to produce the output.

SAS Studio has various benefits such as being an easy to learn programming tool and not requiring the need for installing software onto the user's desktop however using it on an online browser may bring about certain issues such as having a temporary loss of internet connection can prevent the user from using the online software or lose any progress they may have made before. For this project, it may be used to aid pre-processing and cleaning of the dataset.

3.2 Tools Chosen

3.2.1 Power BI Desktop

Figure 34 – Power BI Desktop Logo (Microsoft, 2020)

Power BI is the umbrella term for a collection of cloud-based apps and resources that allow businesses to obtain, organise, and analyse information from a multitude of sources using a user-friendly interface. It may refer to the Power BI Desktop (the windows desktop version), Power BI Service (Software as a Service, SaaS) or Power BI apps that are available on Windows, iOS, and Android devices. Originally, Power BI started as project 'Gemini' by Amir Netz. It allowed SSAS or SQL Server Analysis service to be accessible as an in-memory engine. In 2009, it was renamed Power Pivot and was distributed as an Excel add-in by Microsoft. By 2015, it had gained some traction and was renamed Power BI and was described as a SaaS solution that will provide data to our devices like Acumatica.

Power BI primarily gathers and processes information, transforming it into actionable insights through the use of aesthetically pleasing and simple-to-understand graphs and charts. This enables users to create and share clear and actionable snapshots of what is going on in their company. Power BI allows users to see not only what has occurred in the past and what is occurring now,

but also what may occur in the future. Machine learning capabilities are built into Power BI, allowing it to detect trends in information and then using those patterns to make intelligent assumptions and run "what if" scenarios. These predictions can be used to create forecasts and plan for future needs and other performance factors.

Power BI comes with many advantages, firstly companies can import a large amount of information that possibly may not have been handled by other applications. In-built ML features allow detecting trends much easier and making more informed assumptions. Since Power BI is cloud-based, the user is presented with modern intelligence resources and frequently updated powerful algorithms. Moreover, by creating dashboards, the companies get a much better sense of their data as well as access it whenever necessary. The application can work with other platforms such as Office 365, Google Analytics, Hadoop and more. Finally, Power BI guarantee data security by offering full control on internal and external accessibility. (Wright, 2019) As mentioned before, Power BI has several versions, but only Power BI Desktop and Power BI mobile are free while the others are paid for services.

	Application	**Description**
1.	*Power BI Desktop*	Free, intended from SME's. (Small-medium enterprises)
2.	*Power BI Service*	1. Power BI Premium – licenses by scale, meant for large companies. 2. Power BI Pro – paid/license to get access to certain feature and share reports.
3.	*Power BI Mobile*	Application that can be downloaded on to mobile devices.
4.	*Power BI Embedded*	Allows to be embedded into independent software vendors application and is known as a white label version of Power BI, through this they do not need to build their own analytic elements.
5.	*Power BI Report Server*	This is version which companies who would like to keep their information and dashboard on their own servers.

Figure 35 - Comparison of PowerBI (Wright, 2019)

3.2.2 SAS Enterprise Miner

Figure 36 - (SAS Enterprise Miner, 2016)

SAS Enterprise Miner est. 1966 at North Carolina State University, is a tool for creating precise predictive and descriptive models on large quantities of data from various sources within an organization. SAS is the leading independent provider within the business intelligence industry and the market leader in business analytics software and services. SAS has been serving customers since 1976 and now helps customers at over 70,000 locations boost efficiency and generate value by producing actionable insights faster using creative solutions.

Business Analysts use SAS EM to model their data with various features and functions, moreover, SAS EM can be used to implement business applications such as identifying fraud, reducing risk, resource requirements, campaigns and more. SAS enables us to create models and explore data through versatile data preparation and management capabilities as well as make actionable insight and decisions due to its predictive analytics and data mining capabilities which can further be incorporated into business processes. The main components for SAS predictive analytics and data mining are exploratory data analysis to carry out data exploration, modelling and deployment. SAS EM will be used to carry out data exploration, pre-processing and modelling.

SAS EM can help carry out various tasks due to its endless features such as clustering, market basket analysis, linear and logistic regression – these models will be utilized for the data modelling phase to find the relationship between variables and predict the possibility of fatal police shootings due to person race and ethnicity or mental illness. (PredictiveAnalyticsToday, 2016)

3.3 Operating System

The operating system used for this project will be Windows 10 – Home edition.

3.4 Hardware and Software Requirements

These are the minimum requirements needed for us to carry out the project successfully.

1. Processor – Intel(R) Core (TM) i7-8565U CPU @ 1.80GHz, 1992 MHz, 4 Core(s), 8 Logical Processor(s)
2. Memory – 16 GB RAM
3. Keyboard and Mouse
4. Wi-Fi (Wireless Fidelity)/LAN cable
5. Windows 10 Operating System
6. Google Chrome/Microsoft Edge

3.5 Summary

To conclude this chapter, we have discussed in detail the technical research aspect of this project which consisted of chosen integrated development environments, operating systems, database management systems and tools.

Application features and benefits of the required IDE, SAS Studio. SAS EM will be used for the data exploration, preprocessing and modelling. Next, a visualization tool, Power BI that will be used in the deployment phase to display actionable insights and show data trends was examined. There was further discussion on the operating system that will be used for the project, Windows 10 as well as the minimum hardware and software requirements.

Chapter 4: Methodology

4.1 Introduction

The methodology is vital in researching and analysing our data. It allows us to gather accurate and important results that can help us gain valuable insight into our data. In a data analytics project, there are steps such as data cleaning, modelling, preparing and evaluation and that may take a lot of time, hence it is important to have a suitable methodology.

The detection and extraction of trends and information from broad or challenging data sets are data mining. This involves a wide range of activities, involving grouping or clustering, dependency discovery, and detection of anomalous instances within the data. With greater ease of access to larger volumes of data, a larger emphasis has been put on how to use that data efficiently and to implement tools and processes that methodically deliver new insight into the links between data wherever possible. (Mellor et al., 2018) Data mining necessitates a standard method that can convert business problems into data mining activities, proposing effective data transformations and techniques for data mining, and presenting means for determining the feasibility of the findings and reporting the experience. (Writh and Hipp, 2000)

4.2 Methodologies Comparison

There can be various types of system development methodologies the user can apply to a data analytics project such as KDD or CRISP-DM. A comparison between the two will be shown below with a justification for the chosen methodology at the end.

4.2.1 KDD

KDD is known as knowledge discovery in databases and is used to understand certain types of information from a dataset or collection of data. It may also be defined as the extensive process of discovering information in data, with a focus on the elevated application of certain data mining techniques. The main goal of using KDD is to obtain information from large databases.

It contains five or seven steps. (Qaiser and Shafique, 2014) According to Azevedo and Santos (2008), the five main stages are known as the selection process which consists of finding a dataset or creating your own; preprocessing of the data to make it clean and useable for the transformation phase, the data mining phase where a specific data mining tasks are set with the suitable algorithms

as well as their implementation which will aid in finding patterns from the data and lastly the evaluation phase where the patterns found in the previous phases are interpreted. When the cycle is finished and all the steps are accomplished, the expert is given the evaluation information, which shows whether the knowledge was acquired. Otherwise, the cycle begins again with new goals, until the objective is met. (Pyvovar, 2019)

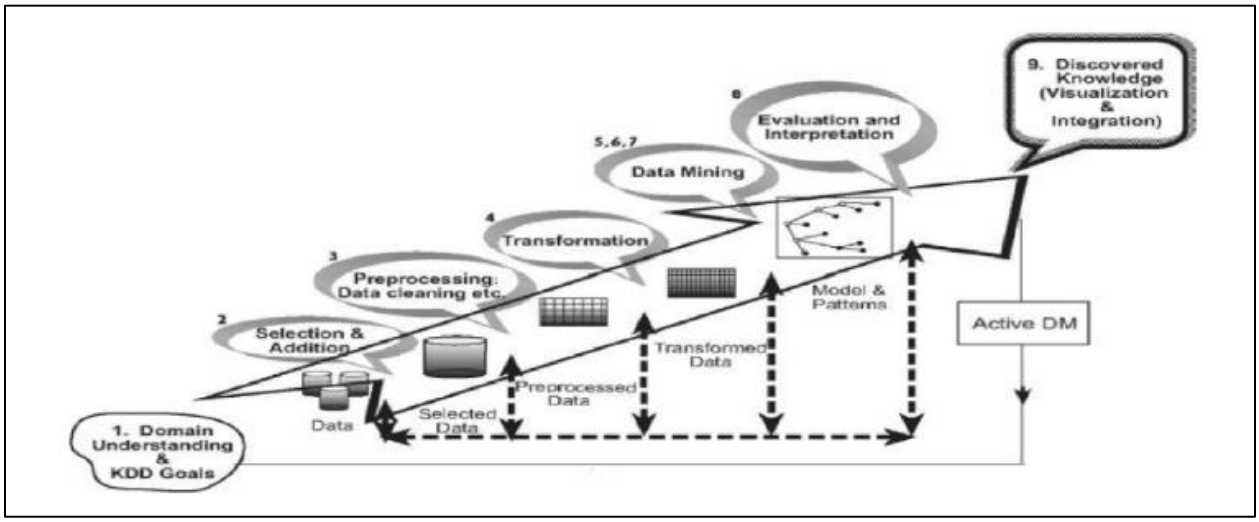


Figure 37 - KDD (Qaiser and Shafique, 2014)

In KDD, it may be necessary to go back and correct a prior step. (Dåderman and Rosander, 2018) It can figure out issues within the current process due to anomaly identification such as finding privacy concerns for a certain project and can use this information to improve the process in the future however this causes the whole process to take more time as every time there is a new knowledge gained or a problem uncovered, it is applied back to strengthen the process's subsequent repetition, not the current one. (Data Science Project Management, 2021)

Furtherly, KDD is prone to security concerns when data is collected such as hacking and privacy concerns from the user when gathering as much information from the user without violating their privacy.

4.2.2 CRISP-DM

CRISP-DM, established in 1999, is known as Cross Industry Standard Process for Data Mining. It is a well-known methodology that is applied to various projects within the data science and analytics industry.

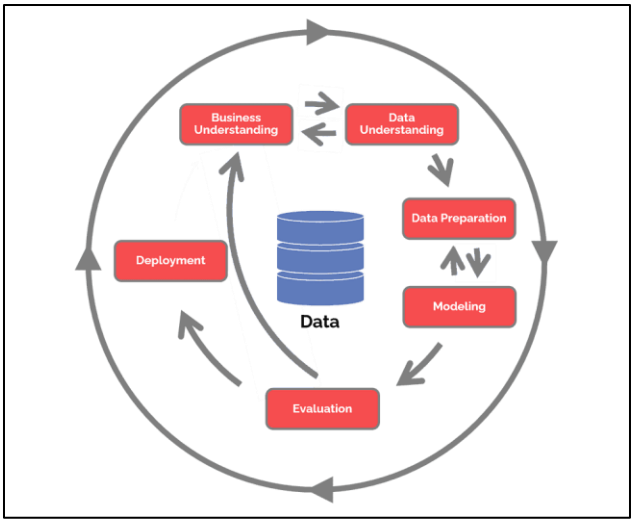


Figure 38 - CRISP-DM (Data Science Project Management, 2021)

A few projects use a combination of CRISP-DM along with other types of methodologies to complete their projects and produce optimal outcomes. (Data Science Project Management, 2021) CRISP-DM consists of six phases. Business understanding: here the objectives, goals and project plans are decided. Data understanding is where the required information is collected, described, explored, and verified before the data preparation phase under which data is cleaned and formatted accordingly. The modelling phase consists of applying certain modelling techniques like classification while evaluation is where the models are evaluated and as well ensure the mode taken to create the model is certain to achieve the required objectives. Lastly, the deployment phase where the model is shown, however, this is not the end of the phase and simply can be used to present the data and knowledge in a clean and easy to understand manner. (Azevedo and Santos, 2008) CRISP-DM shows the most common relationship between each of the phases which allows us to go back to certain phases and correct our mistakes before moving on to the next phase, however it does have a limitation that there is the link between certain phases. For example, if the user has found an issue with the current data in the modelling aspect, they can revert to data preparation to choose different target variables without restarting the cycle, however, cannot revert to data understanding. (Pyvovar, 2019)

CRISP-DM's works to prevent this whole issue altogether as in its business understanding phase, it aims to ensure that a data analyst or scientist is informed on their objectives and business needs before moving on to solving the problem. (Data Science Project Management, 2021) In a project,

the result of the prior phases makes way for which phase or task must be carried out next. (Writh and Hipp, 2000)

A con for CRISP-DM is that it is not a project management approach since it assumes that its user is a single individual or a small team, ignoring the need for teamwork collaboration in bigger projects, moreover, it is very document-heavy due as each task in the phase needs documentation and can cause the process to slow down.

A table is shown below to show how each phase for the methodologies can correspond to each other, they are both considered as iterative processes.

KDD	**CRISP-DM**
Pre-KDD	Business Understanding
Selection	Data Understanding
Pre-processing	
Transformation	Data Preparation
Data Mining	Modelling
Interpretation/Evaluation	Evaluation
	Deployment

Table 1 - Comparison

KDD is the oldest framework as it was established in 1996 while CRISP-DM was established in 1999. CRISP-DM is more practical as compared to KDD, moreover, even though both are iterative processes, CRISP-DM is more iterative and complete due to its iteration flow. KDD is outdated compared to CRISP-DM however some argue that CRISP-DM is also not modern enough for big data projects either. (Kumar, 2020)

Since CRISP-DM is a cross-industry standard, it applies to any type of data science project or work regardless of the field of research while KDD is only applicable to data mining projects. CRISP-DM further fosters strategies and makes it possible for projects to be replicated using it while KDD uses the gained information to apply it to the next cycle hence current cycle cannot be implemented.

KDD is also more on the expensive side due to the storage of the collected data. Costs may also incur due to compilation and maintenance before the process may even start. (Data Science Project Management, 2021) As compared to KDD, CRISP-DM is noted as a cost-saving method since it

incorporates a variety of processes that automate common data mining operations, and the procedures are well-known in the industry. (Chatterjee, 2020) KDD has either 5 or 7 steps while CRISP-DM is set on 6 steps. CRISP-DM consists of both a Business understanding and Deployment phase while KDD does not consider those phases and instead has a pre and post KDD phase.

4.2.3 Summary of Comparison

The user has researched both the methodologies and given suitable explanations regarding each of them along with certain advantages and disadvantages. After considering each point and analyzing which methodology will be suitable for this project, the user has decided to choose CRISP-DM and will provide the justification regarding it below.

4.3 CRISP-DM (Cross Industry Standard Process for Data Mining)

A vital part of data science management is to make sure the data is of the highest quality hence to ensure the success of this project, CRISP-DM is most suitable. It also fits all the requirements set out by the user and has many pros over KDD as well as any of its cons be overcome.

CRISP-DM is more adaptable and flexible therefore can give the user benefits that they would achieve with agile principles. Moreover, for this project, the user needs to be able to correct any problems in their previous phases and this can be done much more efficiently when using CRISP-DM. It is an easy methodology to follow and gives proper guidance, especially for a novice data scientist.

Furtherly, it has a “right-start” and “right-end” and matches perfectly with the objectives set by the users, and although it is considered to be document-heavy, this project is a research project and requires it to be fruitful. The teamwork aspect can also be ignored as the project is carried out by one user.

Moreover, according to various surveys and polls, CRISP-DM was chosen to be the most popular approach for carrying out data science projects or works. KDnuggets Polls carried out polls every 2-3 years from 2002-2007 and asked the respondents about the main method they are using to carry out data mining. They further increased the capacity of this poll in 2014 by adding analytics, data science and data mining. The total number of respondents lay between 150-200 and showed CRISP-DM to remain the most popular methodology throughout the years (Piatetsky, 2014)

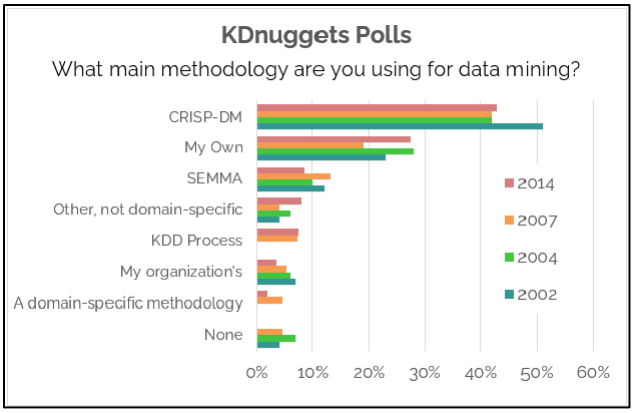


Figure 39 KDnuggets poll (Data Science Project Management, 2021)

Other polls conducted by datascience-pm.com within recent years that are similar to the KDnuggets polls as well as analysing the google searches shows that CRISP-DM is popular among other methodologies even in the current time. Therefore, after considering all the factors and research, the user believes CRISP-DM is the optimal methodology for this project.

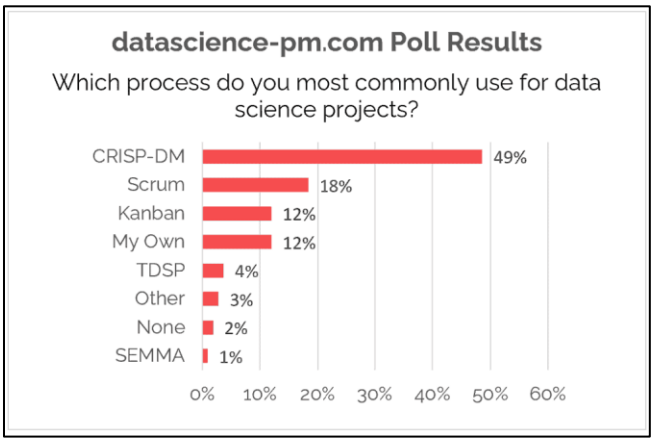


Figure 40 (Data Science Project Management, 2021)

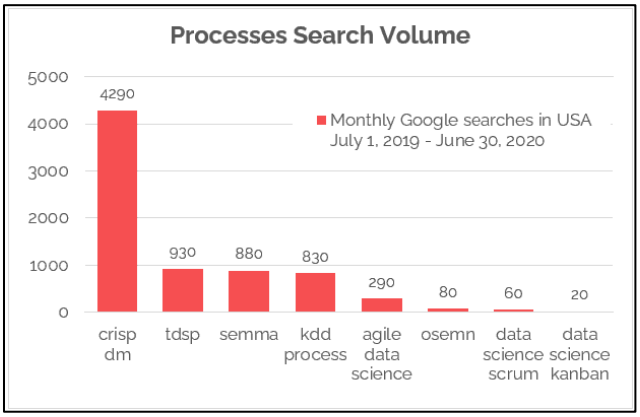


Figure 41 (Data Science Project Management, 2021)

4.4 Business Understanding

This is the initial stage of CRISP-DM and as the name indicates, it is to recognize what is needed to be achieved from the project. It focuses on identifying the aims and requirements of the project after which the knowledge earned is transformed into a "data mining problem definition" and a project plan is formed to attain the objectives. For this project, it is important to find out how this data can be used to show how mental illness and racial discrimination influences fatal police shootings. The primary goal is to show an in-depth analysis of how this greatly affects people of colour and mentally ill individuals.

1. Assessing situation –In this, the resources that will be required are determined such as the datasets, relevant IDEs, articles journals and more. The projects requirements and goals need to be understood to carry out the project successfully as well as determine if any risks may be involved.
2. Determining data mining goals – This defines how the project accomplishment will be seen from data mining perception. The primary goal for this phase is to do an in-depth analysis of the reasons for this project, how it will affect the target users and more. For example, the success of this project in producing actionable insights can be used to advocate for psychiatric care for the mentally ill in America. The data mining goals for this project are found under the objectives and deliverable section.
3. Developing the project plan – The user must keep the data mining objectives in mind as they move forward with this strategy. The measures that will take place in the later stages, including all of the CRISP-DM phases and their implementation strategies and methods, will need to be outlined here. (Smart Vision Europe, 2020)

4.5 Data Understanding

This phase allows the researcher to recognize what can be accomplished with the data; the data quality is assessed in terms of being useful and complete and is an important step before the Data Preparation phase. This is a critical part of the project as it demonstrates how feasible and reliable the outcome will be. This phase detects any noteworthy subsets that may be obscured, evaluates the data to find patterns for actionable insights and so on. (Smart Vision Europe, 2020)

4.5.1 Collection of the initial data

This step starts with collecting our initial data, which includes data loading that is important for this phase. Here the data sets are obtained along with how they were acquired as well as any issues that have been encountered. By recording this, it will be able to be implemented into future projects or projects within the same scope. The data requirements are outlined, and the availability of the data is verified, for example, knowing what data types will be required to achieve the goals as well as confirming that the information is correct and viable.

4.5.2 Describing the Data

The information that has been collected needs to be defined such as the surface/gross properties of the data. The quantity of data such as the number of fields or variables in the table, the format of the data and the data types all need to be identified. After this, it can be determined whether the collected data meets the data mining goals and project requirements.

The dataset used is published by the Washington Post (WAPO) which will be used for analysing the influence of mental illness and racial discrimination on fatal police shootings, it is a data set that originated in 2015 and continues to be updated whenever there is a recent shooting that may have occurred. The dataset consists of names, ages, the race of the victims as well as if the victim were suffering from a mental illness among other fields. The file contains around 6000 records with 17 variables. A data dictionary will be provided in a separate file along with it.

4.5.3 Exploration of Data

This part of the phase looks at problems related to data mining using reporting and data visualization techniques. It includes links between small sets or a certain number of attributes, aggregations results and simple statistical analysis. The data mining goals could be directly answered through these analyses. They may also contribute to or enhance the data definition and quality reports, and feed into the conversion and other data preparation steps required for further research. Through the use of the tools assigned for this project, the data could be explored in a multitude of ways such as using various nodes within SAS Enterprise Miner. File Import node allows the exploration of data in its very raw form, through the use of the explore tab on the node's editor, the user can see the distribution of the variables before any steps are taken. Nodes like StatExplore and Multiplot can allow the user to see the statistical summary and plots of the variables against the target. This will further be touched upon in chapter 5.

4.6 Data Preparation

Data Preparation is known to be the most important step as well as the step that takes the most time in CRISP-DM. This phase involves ETLS that changes a portion of information into something valuable by using certain methods and algorithms. During this process, any missing values, outliers, or irregularities in the raw dataset can be detected.

SAS Enterprise Miner and SAS Studio can be utilized for this phase of CRISP-DM. The dataset will be imported into EM for analysis using exploratory data analysis techniques to look at relationships, trends, and anomalies to gain an understanding of the dataset. This can help select the relevant rows, columns and attributes required for the project using the FileImport node. Some examples of graphs that can be used to carry out this analysis are boxplots and histograms that can be utilized to show the minimum, maximum, mean, median, quartiles and range of the data which will help in identifying any missing values or irregular data as well as scatterplots to show the relationship between variables and to detect outliers. The data can be partitioned using the Data Partition node while the Variable Selection node can be used to select the most important variables for modelling against the targets, this can be done before and after the transformation of data depending on how the process is carried out.

After exploring the data thoroughly, the process can move onto data cleaning, this is the process of removing or replacing missing values, formatting the data, and selecting the relevant variables. Data can be cleaned using impute node in EM – this allows replacement of the missing values using the mean/most occurring value or median etc. Drop node can be used to 'drop' or delete certain columns if they are not useful to the evaluation. Moreover, the dataset should be continuously cleaned to ensure that the dataset is optimal for the project. SAS Studio may also be employed to aid in certain aspects of data preparation which may be easier to carry out in the online IDE instead of SAS EM.

The last step of this phase is data formatting, if there is a need for the variable's format to be changed, it can be done through nodes like Transform Variables, Interactive Binning, etc.

4.7 Data Modelling

This is an extremely important core step that will be accountable for the outcomes that should be satisfactory to the project aims. It takes the least amount of time from the project if previous phases

have been carried out efficiently and are sufficient. Due to any reasons such as result is not satisfactory or can be additionally improved then data preparation phase can be revisited and enhance the available information. (Rodrigues, 2020) This phase has four sub-phases that will be explained below.

1. Select Modelling Technique – this sub-phase is where the selected modelling techniques are discussed in detail and how they will be carried out in the data analysis section of the project.
2. Generate Test Designs – description of the plan to be used for training, testing, and evaluating data set such as partition the data into three datasets: training, validation, and testing using the Data Partition node.
3. Build Model – this includes setting certain parameters and creating real models using whatever modelling tool selected.
4. Assess Model - Many models typically compete against one another, and the data analyst needs to analyse the results of the model based on domain knowledge, pre-defined requirements for performance, and test design.

When the data pre-processing is complete, the data modelling phase can be initiated. In the selected dataset there are multiple instances of categorical data such as age, gender, and race, these can be changed to a numerical value by making design variables (aka dummy variables) to ensure data is prepared for modelling.

Test data can come from a sub-set of the original data set, and this can be prepared through the data partition node within the SAS enterprise miner. Test Data can be utilized in a confirmatory manner, for example, to ensure that a particular set of inputs to a function gets the desired result. (Kumar, 2017) However, the evaluation can also be considered according to the validation dataset. More on this will be discussed in the next chapter.

The model can be constructed within the SAS Enterprise Miner and age, race, gender, manner of death, signs of mental illness, etc. can be some of the variables that may be used as independent and dependent variables. To assess the model, a report can be generated regarding the model's analysis.

4.8 Evaluation

The second last step is the evaluation phase where all the results are verified and validated. In case the results are unsatisfactory, the previous phases can be revisited and the mistakes which caused such results can be amended. (Rodrigues, 2020) It is extremely important to evaluate as it helps understand the performance of the model as well as help in presenting it efficiently. There are various evaluation methods to assess the model's efficiency and performance, some of them have been discussed in the literature review such as:

1. Mean Absolute Error (M.A.E)
2. Mean Squared Error (M.S.E)
3. Root Mean Squared Error (R.M.S.E)
4. R-Squared (R^2)
5. Misclassification Rate

Wu (2020) says that M.A.E, M.S.E and R-Squared are the three best metrics to evaluate the regression model while Chi-square and Confusion matrix is best for decision trees.

4.9 Deployment

Model production is normally not the end of the project. Even if the model intends to increase awareness of the data, it will be important to organize and present the knowledge acquired in a way that the user will use it. The deployment phase could be as straightforward as producing a document or as complicated as implementing a consistent data mining process, based on the requirements. It is critical to know ahead of time what steps must be taken to use the models that have been created. (Writh and Hipp, 2000)

For this project, after all the phases are complete, the last phase, deployment will remain. The cleaned dataset will be exported from SAS Enterprise Miner and imported to Power BI for the creation of an interactive dashboard. This will be carried out by selecting variables that will produce the intended outcome for this project. Certain pre-processing steps will need to be taken care of when producing the dashboard in PowerBI. There can be 3-4 dashboards for this project such as one for a general overview of the data, one for showing certain important variables against the target variables, and one dashboard each with regards to both targets interactively showing their best model. The initial two dashboards can consist of charts such as boxplots, funnel

diagrams, pie charts, heat maps. The last two charts could be a line chart or decomposition tree based on the best model selection.

4.10 Summary

For this chapter, a description regarding each phase has been given as well as what the outcome may be after the individual phases are complete. Business Understanding gave an overview of what the objective was for this project overall, for example, investigating the influence of mental illness and racial discrimination on fatal police shootings using predictive analytics techniques and EDA. For the Data Understanding phase, the importance of understanding the key attributes of the data was discussed. The process of carrying out data exploration and data cleaning using SAS Enterprise Miner and SAS Studio was recognized in the data preparation phase, so the data is optimal for the modelling and evaluation phase. In the data modelling phase, the technique that will be the most suitable for this project which will be regression and decision trees were touched upon. In the Evaluation phase, the top evaluation metrics for the final assessment of the models were mentioned such as MSE, RMSE, MISC and MAE. Finally, for the Deployment phase, it was discussed what steps will be taken to import the data from SAS Enterprise Miner to Power BI for visualization which will ensure the success of this project.

Chapter 5: Data Analysis

5.1 Introduction

Data Analysis is important when analysing a large amount of information to discover hidden patterns, trends, relationships, and correlations to help produce actionable insights. It allows the prediction and visualization of the types of crimes that may occur due to certain variables. The geographical location, possible timings, and reasons are some of the things that can be discovered through the application of big data analysis. This data may further be used to help enact laws and regulations or place pressure on law enforcement to ensure these situations may not occur again. (Vadakkanmarveettil, 2020) For this research paper, data analytics is applied in both a descriptive and predictive manner which is both exploring the data and coming up with questions and whether certain attributes will cause a situation to occur, respectively. The relationship between independent and dependent variables will be analysed through predictive modelling with regression and decision tree using target variables of *Signs Of Mental Illness* and *Race*.

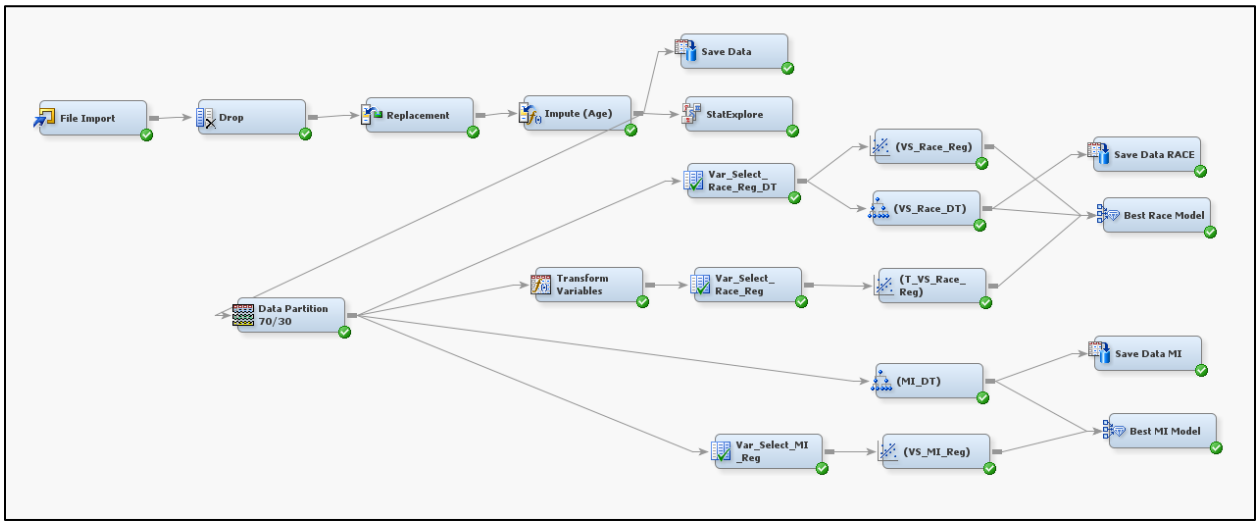


Figure 42 - Workspace showcasing prediction model in SAS EM

5.2 Initial Data Exploration

The data set has been downloaded from the official GitHub of the Washington Post's database (WashingtonPost, 2015) for 'fatal police shootings' (FPS) and a column has been merged from the 'Mapping Police Violence'(MPV) database to help condense the locations. The dataset is in a Microsoft Excel CSV file and consists of over 6,000 records with 18 attributes. A data dictionary is created showing the original dataset below.

Some of the data cleanings will be taking place in SAS Studio before being imported into SAS Enterprise Miner, this will be discussed under the Data Cleaning subheading. The dataset is first loaded into SAS Enterprise Miner in CSV format to get an initial understanding of the overall data and define their roles and levels. Any data inconsistencies can be discovered here and cleaned in the following stages.

5.2.1 Data Dictionary

No.	Attribute	Role	Level	Description	Values
1	id	Input	Interval	unique id to distinguish each victim	
2	name	Input	Nominal	Name of the victims	
3	date	Time_ID	Interval	Date the police shot the victim	YYYY-MM-DD
4	manner_of_death	Input	Nominal	How was the victim killed?	shot shot and tasered
5	armed	Input	Nominal	Did the victim possess any weapons when they were shot	Baseball bats, screwdriver, hammer, etc.
6	age	Input	Interval	Age of the victims	
7	gender	Input	Nominal	Gender of the victims	F/M
8	race	Target	Nominal	Race of the victims	A/B/H/O/N
9	city	Input	Nominal	City victim was in when they were shot.	Location of shooting may consist of county names
10	state	Input	Nominal	two-letter post code acronym	Two-letter post code abbreviation
11	signs_of_mental_illness	Target	Binary	According to news reports, the victim was experiencing mental illness or had a history of it, or had expressed a suicidal tendency	True/False
12	threat_level	Input	Nominal	Was the victim a threat to themselves and anyone around them	attack other undetermined
13	flee	Input	Nominal	Did the victim try to flee and with what methods	car foot not fleeing other

14	body_camera	Input	Binary	If the police officer was wearing a body camera or not during the killing	True/False
15	longitude	Input	Interval	Coordinates where the shooting took place	
16	latitude	Input	Interval		
17	is_geocoding_exact	Input	Interval	Whether or not the coordinates are exact or not.	True/False

Table 2 - WaPo Fatal Police Shooting Dataset

1	County	Input	Nominal	The County where the police officers' departments were located and the victim was shot	Ex: Name of the County such as Los Angeles County

Table 3 - Mapping Police Violence 'county' attribute

5.2.2 File Import Node – Data Exploration

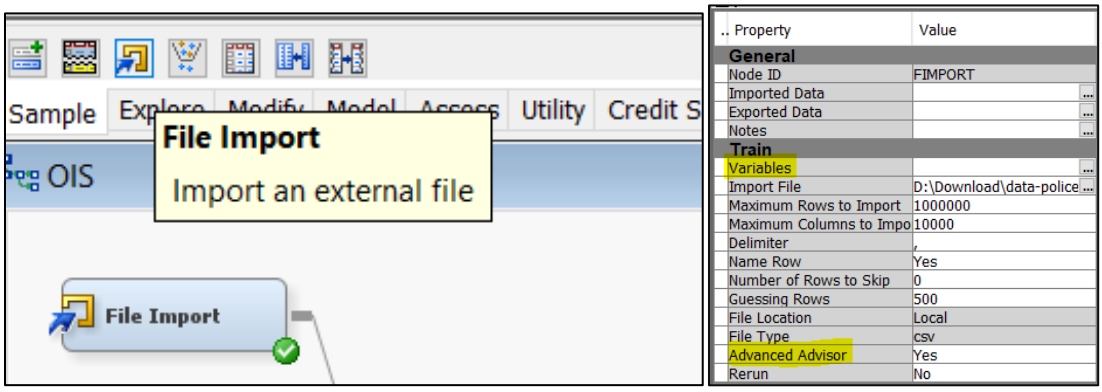


Figure 43 & 44 - FileImport Node and Property Window

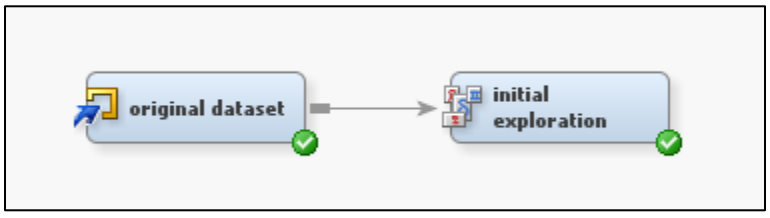


Figure 45 - Initial data exploration of the original dataset

A new diagram is created under the new project and the file import node is dragged onto the workspace from the Sample Tab in the tool bar. A file import node can be used to import the fatal police shooting (FPS) spreadsheet to an accurate format for Enterprise Miner which will allow it

to be recognized as the data source to be used in the data mining process. (SAS, 2021) It further allows the selection of the metadata information for the table such as selecting the role and level as shown under the "Variable" property in the file import node property. The advanced advisor was turned to 'Yes' to see what levels and roles the program would automatically assign the variables.

Roles refers to the model role of a variable such as input which is independent and is used in the prediction of the target, target refers to variables where the current data is known to the user however for new data it is unknown and is the dependent variable of the dataset. Time ID is used to identify time while ID is the unique indicator for every observation in a dataset. Rejected is used for the removal of irrelevant variables in the dataset. Level refers to the measurement of a variable such as Nominal, Ordinal, Binary, Interval and Unary. (SAS, 2021) The meaning of each level has been discussed in detail in the literature review (see Section 2.4.1)

Name	Role	Level	Comment
age	**Input**	**Interval**	
armed	**Rejected**	**Nominal**	Exceeds maximum number of levels cutoff
body_camera	**Input**	**Binary**	
city	**Rejected**	**Nominal**	Exceeds maximum number of levels cutoff
date	**Time ID**	**Interval**	
flee	**Input**	**Nominal**	
gender	**Input**	**Binary**	
id	**ID**	**Interval**	
is_geocoding_c	**Input**	**Binary**	
latitude	**Input**	**Interval**	
longitude	**Input**	**Interval**	
manner_of_de	**Input**	**Binary**	
name	**Rejected**	**Nominal**	Exceeds maximum number of levels cutoff
race	**Target**	**Nominal**	
signs_of_ment	**Target**	**Binary**	
state	**Rejected**	**Nominal**	Exceeds maximum number of levels cutoff
threat_level	**Input**	**Nominal**	

Figure 46 - FileImport Node Variable editor

The node was then run, and it can be seen that name, armed, city and state were all rejected due to exceeding the 'maximum number of levels cut-off'. If these variables are kept, they would bring up issues while trying to carry out the data analysis process.

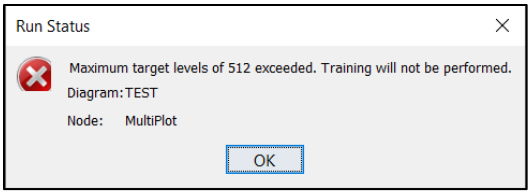


Figure 47 - Error Message if variables are not rejected by FileImport Node

For the remaining variables, everything was assigned appropriately; RACE and SIGNS_OF_MENTAL_ILLNESS were turned into target variables as those were the dependent variables meanwhile the remaining were kept as input, time_id or id as those would be the independent variables. The target variables were explored and shown with their frequency below.

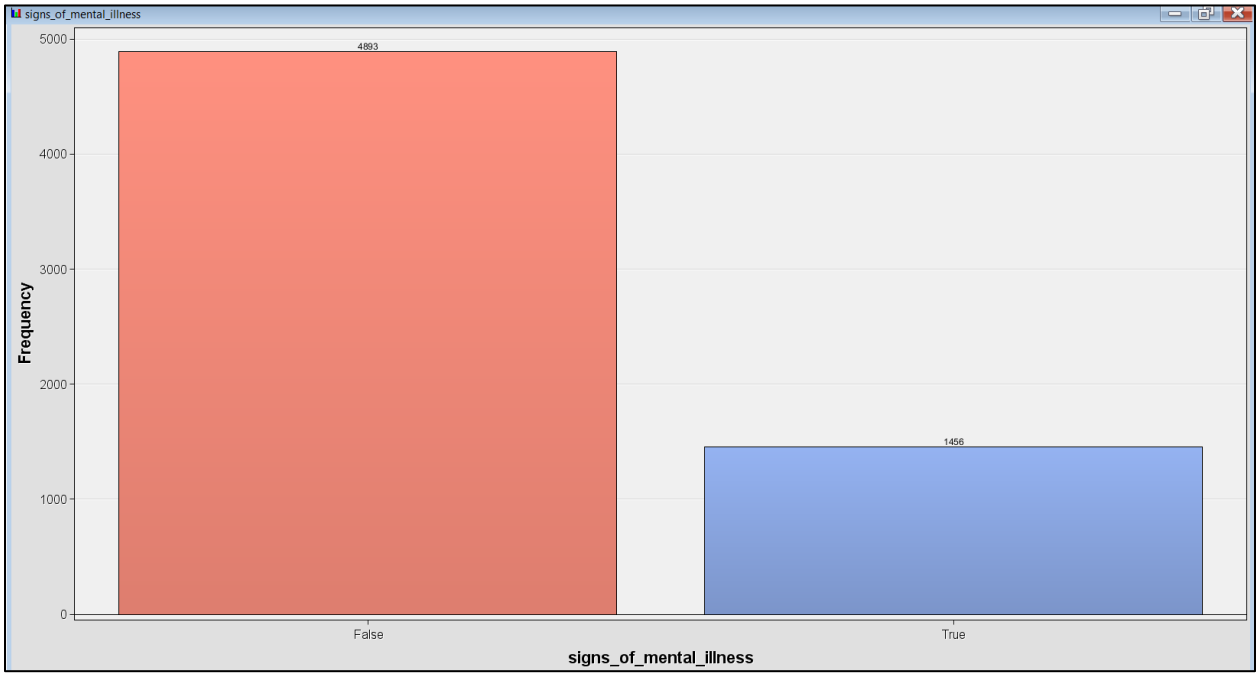


Figure 48 – Bar-chart for Target Variable: SIGNS_OF_MENTAL_ILLNESS

It can be seen that out of 6349 victims, 4893 had no signs of mental illness while 1456 were known to be suffering from a mental illness. According to the Mapping Police Violence dataset, when a victim was suffering from mental illness, it also includes victims who had a known history of drug and alcohol abuse. There were also victims for who it was not known whether they were suffering from mental illness, or it was unclear at the time of the shooting, (Sinyangwe et al., 2021) this is however not included in the WaPo dataset as the SIGNS_OF_MENTAL_ILLNESS is categorized under true or false, therefore it was decided to keep it as the binary target variable.

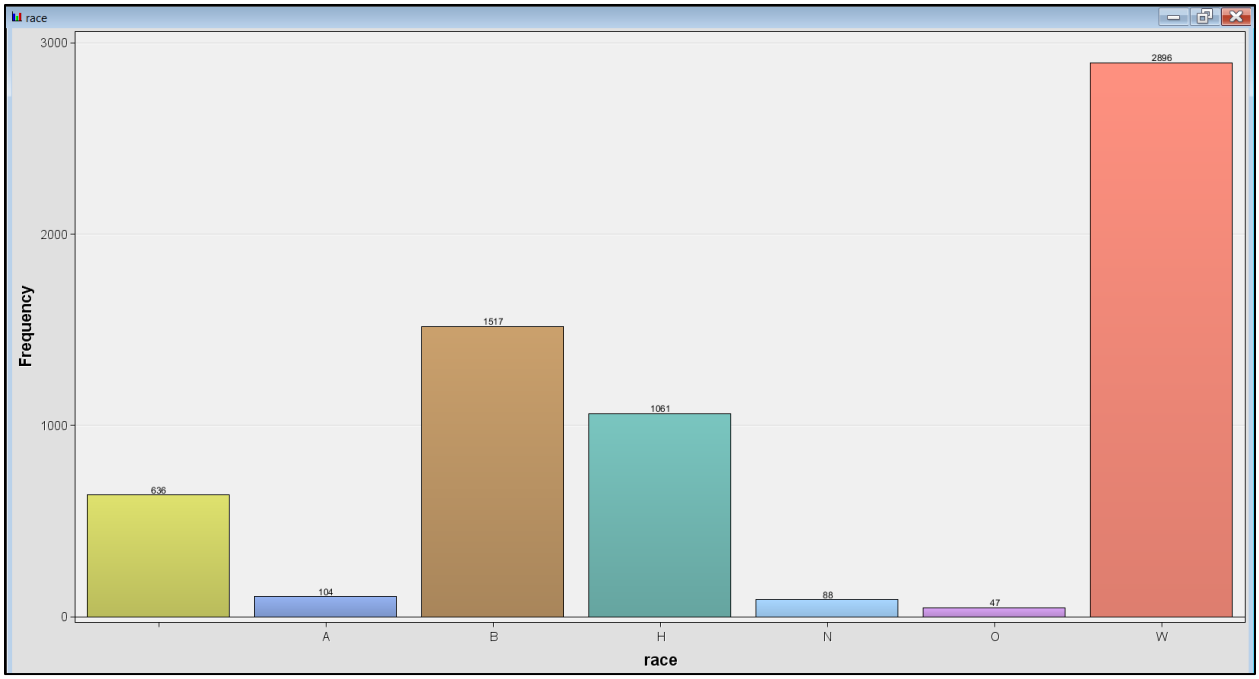


Figure 49 – Bar-chart for Target Variable: RACE

For the target variable RACE, missing values can be seen which stand at 636 victims, while 'Asian', 'Native' and 'Other' victims were at 104, 88 and 47, respectively. 'Black', 'Hispanic', and 'White victims' were the top 3 races that had the most victims in this dataset with 1517, 1061 and 2896, respectively.

5.2.3 StatExplore Node – Statistical Information and Variable Relationships

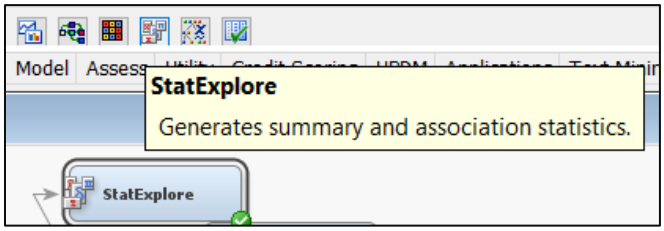


Figure 50 – StatExplore Node

There are two target variables, RACE, and SIGNS_OF_MENTAL_ILLNESS. After file import, specific nodes from the Explore tab in SAS Enterprise Miner can be used, such as StatExplore and Multiplot to carry out data exploration. DMDB node may also be used however it only provides the statistical information therefore using StatExplore which shows more than just statistics is much more efficient for this project.

In the property window for StatExplore Node, under the sub-section of chi-square statistics, The 'Chi-square' is set to yes here to carry out the calculation and display it for every variable. It shows the relationship between the target variables and categorical input variables. For example, the age of the victim is an interval variable and if there was a need for chi-square statistic to be measured, then the 'interval variables' is set to yes. The number of bins can be specified; however, the default setting can be kept as 5.

In the first few lines, the variable summary can be seen which shows the roles, measurement level and the number of variables that correlate with it. There is one ID variable, 5 binary variables including one target variable, 3 interval variables, 3 nominal variables including a target variable and 4 rejected variables.

The variable level summary shows the ID which accounts for each observation in the dataset with a total number of unique records being 6349 and the two target variables. It is noted for RACE the frequency count is 7, there is one category under RACE for missing values which will be removed in the data cleaning process.

Under the class variable summary, it shows all the categorical nominal variables and displays their role, the number of levels within the variables, the missing values and mode, the most occurring variable. For FLEE, GENDER and RACE, there are 414, 1 and 636 missing values, respectively, meanwhile, for interval variable summary, it shows all interval variables, AGE, LATITUDE, and LONGITUDE. As interval variables are numerical, it calculates their mean, standard deviation, median, skewness, kurtosis, minimum, maximum, missing and non-missing values. Mean is the average of all the values of a variable, the standard deviation is the measure of how spread out a data is, the lower it is, the closer to the mean and the higher it is, the further away from the mean. Skewness is when the distribution has deviated from the actual normal distribution that is usually symmetrical on each side. (Trinidad, 2020), while Kurtosis determines if the values are heavy-tailed or light-tailed in comparison to a normal distribution.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	BINARY	4
INPUT	INTERVAL	3
INPUT	NOMINAL	2
REJECTED	NOMINAL	4
TARGET	BINARY	1
TARGET	NOMINAL	1

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
id	ID	6349
race	TARGET	7
signs_of_mental_illness	TARGET	2

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	body_camera	INPUT	2	0	FALSE	86.69	TRUE	13.31
TRAIN	flee	INPUT	5	414	Not fleeing	60.67	Car	16.19
TRAIN	gender	INPUT	3	1	M	95.59	F	4.39
TRAIN	is_geocoding_exact	INPUT	2	0	TRUE	99.87	FALS	0.13
TRAIN	manner_of_death	INPUT	2	0	shot	94.90	shot and Tasered	5.10
TRAIN	threat_level	INPUT	3	0	attack	64.51	other	32.46
TRAIN	race	TARGET	7	636	W	45.61	B	23.89
TRAIN	signs_of_mental_illness	TARGET	2	0	FALSE	77.07	TRUE	22.93

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	37.08719	12.99832	6067	282	6	35	91	0.72334	0.183642
latitude	INPUT	36.65859	5.37914	6042	307	19.498	36.102	71.301	0.608864	2.386085
longitude	INPUT	-97.179	16.6106	6042	307	-158.137	-94.357	-67.867	-0.57598	-0.26012

Figure 51 - Result Output for StatExplore

5.2.3.1 Chi-Square and Variable Worth Plot

The chi-square plot shows the relationship between the input and target variables, it displays the strength between the two. The larger the chi-square value, the stronger the relationship between that input variable and the target and the variable's predictive value. For seeing the relationship between the input and target, some of the rejected variables such as STATE, ARMED and CITY were used while IS_GEOCODING_EXACT, NAME and ID were excluded. Through this, certain aspects of the data cleaning process can also be justified.

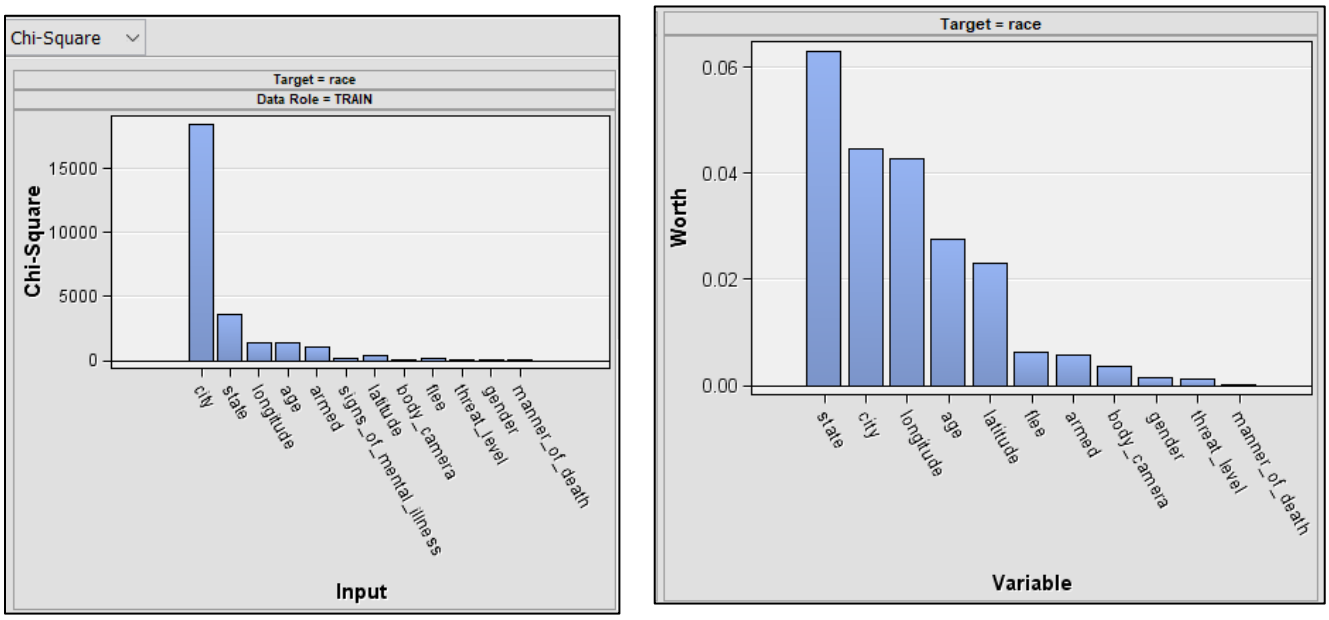


Figure 52 & Figure 53 – Chi Square and Variable Worth Plot

Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=race

Input	Chi-Square	Df	Prob
city	18497.5648	16500	<.0001
state	3597.9843	300	<.0001
longitude	1438.5614	30	<.0001
age	1390.9327	30	<.0001
armed	1016.9376	594	<.0001
signs_of_mental_illness	137.9709	6	<.0001
latitude	407.2723	30	<.0001
body_camera	68.8002	6	<.0001
flee	173.4463	24	<.0001
threat_level	47.6198	12	<.0001
gender	35.9459	12	0.0003
manner_of_death	7.0611	6	0.3152

Figure 54 - Chi-Square Statistics

The chi-square plot shows CITY having the highest chi-square value compared to any other variable, 18497.5 followed by STATE at 3597.9, LONGITUDE at 1438.5 and AGE at 1390.9. The

Df column represents the degree of freedom which tells us how many numbers within our chi-square are independent. DF is calculated by subtracting the total categories or levels of a variable by 1.

$$DF\ (Degree\ of\ Freedom) = DF - 1$$

Figure 55 - DF formula

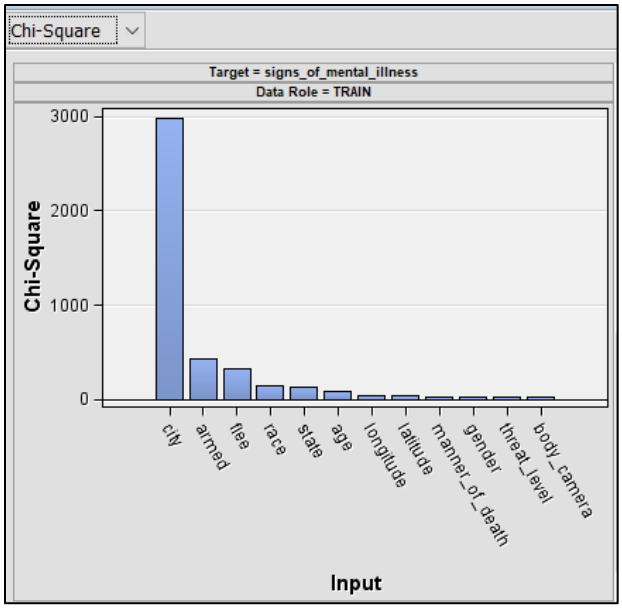
For CITY, STATE and ARMED, the DF is too high. This may be due to the missing values or different types of categories that correspond to each race. CITY and STATE can be pre-processed under one column which is REGION that can condense the location based on their geographical locations on the US Map. As a result, LONGITUDE and LATITUDE may also be dropped.

The variable worth window can be seen next which displays the "worth" of each independent input variable against the target variable. Variable worth is computed by the corresponding p-value to the chi-square test statistic. The formula for this is shown below. The variable worth chart classifies the input variables corresponding to their computed worth.

$$P(x^2 >= \text{calculated chi-Square statistic}) = p$$
$$\text{Worth of the input is } -2\log(p)$$

Figure 56 - Variable Worth Formula (Sarma, 2017)

According to the variable worth plot for target RACE, the STATE is the variable with the highest worth, 0.0631, followed by CITY = 0.044, LONGITUDE = 0.043, AGE = 0.027 and so on. MANNER_OF_DEATH has the lowest variable worth with 0.0001319



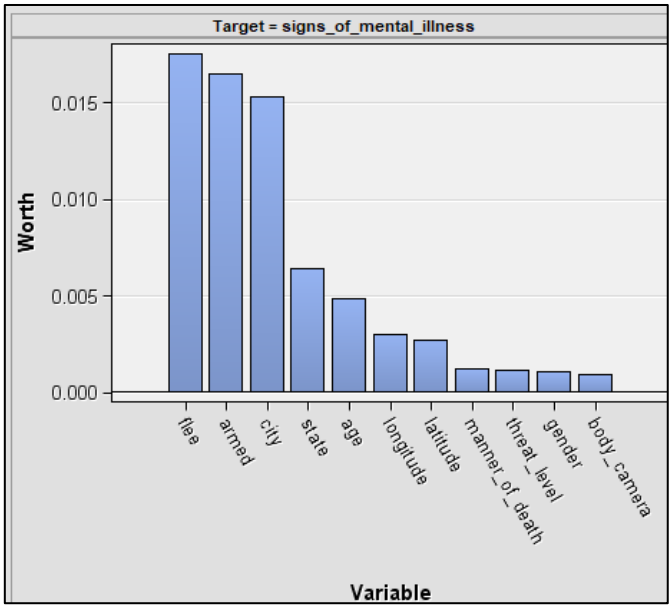


Figure 57 & Figure 58 – Chi Square and Variable Worth Plot

Data Role=TRAIN Target=signs_of_mental_illness

Input	Chi-Square	Df	Prob
city	2977.0574	2750	0.0014
armed	430.0537	99	<.0001
flee	315.8319	4	<.0001
race	137.9709	6	<.0001
state	128.1889	50	<.0001
age	76.5077	5	<.0001
longitude	39.0221	5	<.0001
latitude	32.4982	5	<.0001
manner_of_death	22.1556	1	<.0001
gender	21.2642	2	<.0001
threat_level	21.1740	2	<.0001
body_camera	17.2213	1	<.0001

Figure 59 - Chi-Square Statistics

For SIGNS_OF_MENTAL_ILLNESS as a target variable, CITY again has the highest chi-square at 2977.05 followed by ARMED and FLEE at 430 and 315.8, respectively. The variable worth plot however displays that FLEE has a higher variable worth than the CITY variable.

The DF for CITY here is at 2750 which is lower than the DF for CITY with the target variable, RACE. This is due to the reason that RACE has 7 levels while SIGNS_OF_MENTAL_ILLNESS has only 2 levels. This is another thing that needs to be pre-processed to help bring consistency to the chi-square statistics. SAS Enterprise Miner has difficulties with multinomial analysis which will be touched upon in the sections below.

The variable worth plots show FLEE with the highest variable worth of 0.0175, followed by ARMED = 0.0165, CITY = 0.0153, and so on. BODY_CAMERA has the lowest variable worth, 0.00095 against the target, SIGNS_OF_MENTAL_ILLNESS.

5.3 Data Cleaning

This considers the result from the StatExplore node and carries out certain pre-processing steps such as transforming data, removing redundant data such as rows with many empty values, checking for outliers and more. Some pre-processing will take place in SAS Studio such as recoding certain categorical variables to interval variables, while some will take place after being imported to SAS Enterprise Miner such as taking care or replacing missing values.

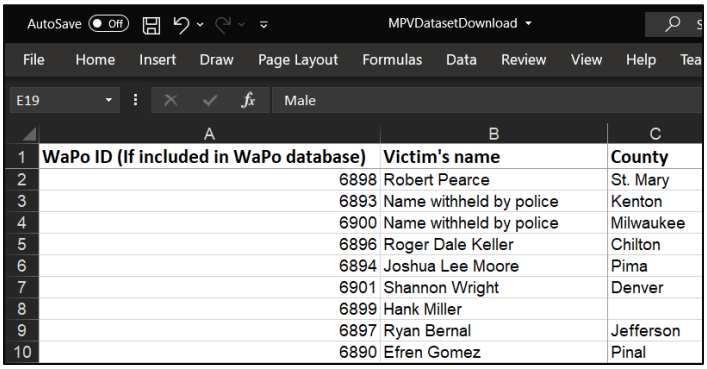
5.3.1 Data Cleaning – Before Importing to SAS Enterprise Miner

5.3.1.1 Data Cleaning in Excel

To start with making the dataset optimal to carry out data analysis, it was first decided to merge the county column from the MPV dataset with the FPS dataset. This was done within Excel itself using the VLOOKUP function as below.

=VLOOKUP(lookup value, range containing the lookup value, the column number in the range containing the return value, Approximate match (TRUE) or Exact match (FALSE)). (Microsoft, 2019)

The lookup value would be the ID number as it was present in both datasets, the range would be the columns in the MPV dataset which consisted of WaPo ID, Victim's NAME and COUNTY, column number would be '3' as that is the column where the data for COUNTY is located and lastly an 'Exact Match' would be required hence FALSE would be selected.



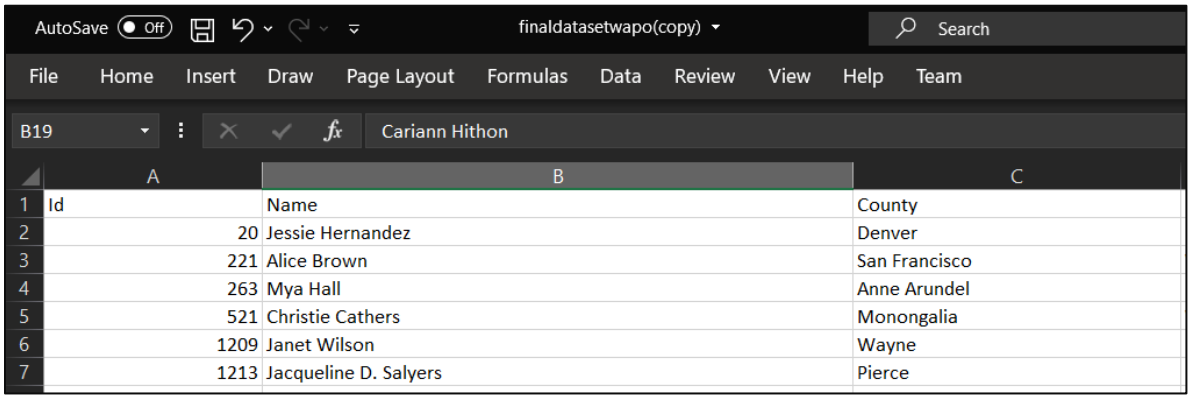
	A	B	C
1	WaPo ID (If included in WaPo database)	Victim's name	County
2	6898	Robert Pearce	St. Mary
3	6893	Name withheld by police	Kenton
4	6900	Name withheld by police	Milwaukee
5	6896	Roger Dale Keller	Chilton
6	6894	Joshua Lee Moore	Pima
7	6901	Shannon Wright	Denver
8	6899	Hank Miller	
9	6897	Ryan Bernal	Jefferson
10	6890	Efren Gomez	Pinal

Figure 60 - WaPo ID in MPV dataset

```
=VLOOKUP(A32,'[MPVDatasetDownload.xlsx]2013-2020 Police Killings'!$A$2:$C$9205,3,FALSE)
```

Figure 61 - VLOOKUP function

After this function is performed, the COUNTY column has been added to the WaPo FPS dataset.



	A	B	C
1	Id	Name	County
2	20	Jessie Hernandez	Denver
3	221	Alice Brown	San Francisco
4	263	Mya Hall	Anne Arundel
5	521	Christie Cathers	Monongalia
6	1209	Janet Wilson	Wayne
7	1213	Jacqueline D. Salyers	Pierce

Figure 62 - the result of the VLOOKUP function

All rows that consisted of missing values for the RACE column were deleted as it was deemed that they were not necessary to the analysis. There was another variable known as Other, this was still kept as although the race was not known, it was still able to define many rows. In total 636 rows were deleted. The next missing value was for the GENDER column. After filtering, it was shown that only one victim's RACE was defined.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of	threat	flee	body_	longitu	latitud	is_geo	county
2956	Scout Schultz	16/9/2017	shot	knife	21		W	Atlanta	GA	TRUE	other	Not fleeir	FALSE			TRUE	Fulton

Figure 63 - missing value for gender

A quick search on the victim produced various new articles helping deduce that the victim was a 21-year-old male shot in Georgia (Fabian, 2017), hence this was manually changed within Excel as well, with no more missing values for RACE.

Overall, after the initial data exploration, it was noted that ARMED, CITY, STATE and NAME were rejected. This is because there were too many levels for the SAS Enterprise Miner to work with, this can be solved by categorizing the state column into regions within the US and the ARMED column into ARMED CATEGORIES as while going through the dataset, it was seen that many victims may have been holding weapons that can be classified as sharp weapons or were holding blunt weapons. NAME and CITY can be left as is, as NAMES are not required for privacy and ethical reasons meanwhile, for CITY, STATE, LONGITUDE, LATITUDE, the new column REGION can be used for location description instead.

5.3.1.2 Data Cleaning in SAS Studio

The merged file is imported into SAS Studio and run so data pre-processing can begin.

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
6	age	Num	8	BEST.		age
5	armed	Char	32	$32.	$32.	armed
14	body_camera	Num	8	BEST.		body_camera
9	city	Char	30	$30.	$30.	city
18	county	Char	80	$80.	$80.	county
3	date	Num	8	MMDDYY10.		date
13	flee	Char	11	$11.	$11.	flee
7	gender	Char	1	$1.	$1.	gender
1	id	Num	8	BEST.		id
17	is_geocoding_exact	Num	8	BEST.		is_geocoding_exact
16	latitude	Num	8	BEST.		latitude
15	longitude	Num	8	BEST.		longitude
4	manner_of_death	Char	16	$16.	$16.	manner_of_death
2	name	Char	33	$33.	$33.	name
8	race	Char	1	$1.	$1.	race
11	signs_of_mental_illness	Num	8	BEST.		signs_of_mental_illness
10	state	Char	2	$2.	$2.	state
12	threat_level	Char	12	$12.	$12.	threat_level

Figure 64 - Import Result in SAS Studio

The result of running the dataset is shown below with all the variables and their formats being displayed.

Figure 65 - Displaying dataset in SAS Studio

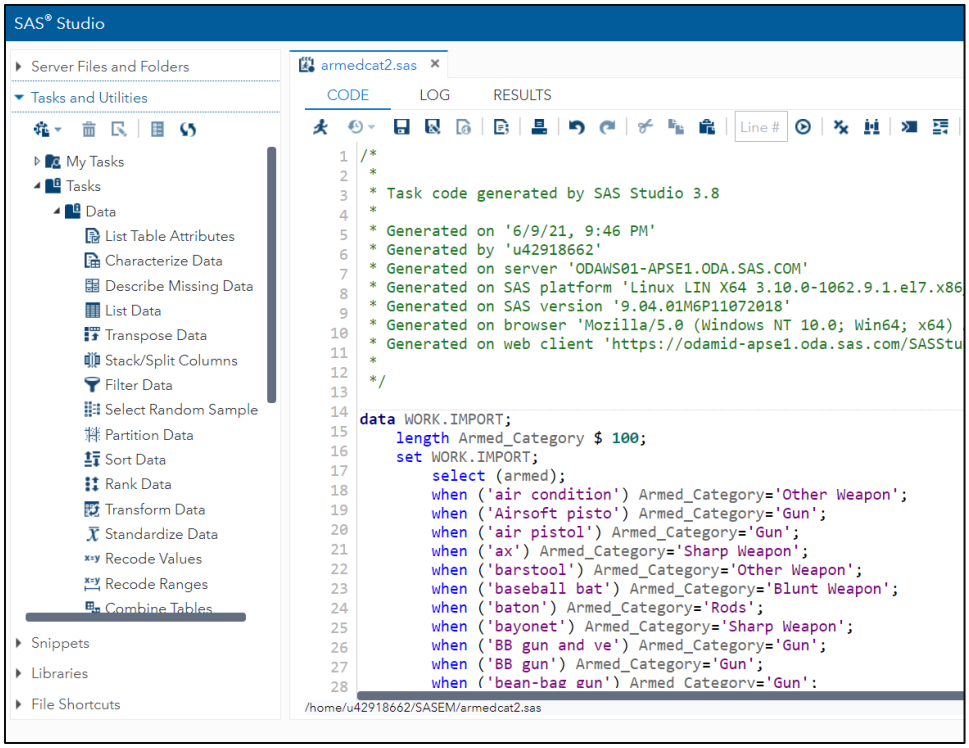


Figure 66 – Tasks in SAS Studio

Firstly, the ARMED column must be recoded with the recode value task that can be found as such Tasks and Utilities>Tasks>Data>Recode Values. This is displayed in the screenshot above.

The task will be used to group certain weapons under one category. After importing, the table below shows how each armed variable was imported as well as the determined category. The categories were determined through own judgement as well as thorough research. Each grouping category will be given a justification below.

Category	Justification	Values
Blunt Weapon	These are weapons that simply require a person's strength to inflict damage such as blunt force.	baseball bat, baseball bat and bottle, baseball bat and fireplace poker, baseball bat and knife, blunt object, brick, hammer, hand torch, metal hand tool, oar, piece of wood, rock, samurai sword, screwdriver, shovel, stapler, wrench
Sharp Weapon	These are weapons that are sharp and can stab or penetrate a victim upon contact.	ax, bayonet, bow and arrow, box cutter, carjack, chainsaw, chain saw, cordless drill, crossbow, glass shard, contractor's level, hatchet, hatchet and gun, ice pick, knife, knife and vehicle, lawn mower blade, machete, machete and gun, meat cleaver, metal rake, pen, pick-axe, pitchfork, railroad spikes, scissors, sharp object, spear, straight-edge razor, sword,

Other Weapon	These are weapons that are different in the category, they can be both sharp and blunt.	air conditioner, barstool, beer bottle, binoculars, bottle, chain, chair, fireworks, flashlight, garden tool, grenade, incendiary device, metal object, microphone, pepper spray, taser, tire iron, wasp spray
Gun	These consist of all types of guns as well as takes any weapon classified as 'gun and..' under this category.	air pistol, airsoft pistol, bb gun, bb gun and vehicle, bean-bag gun, gun, gun and car, gun and knife, gun and machete, gun and sword, gun and vehicle, gun and explosives, pellet gun, nail gun
Vehicle	These consist of machines used as transportation and includes any weapons classified as 'vehicles and..' under this category.	car, knife and mace, motorcycle, vehicle, vehicle and gun, vehicle and machete,
Unarmed	All unarmed victims	unarmed
Undetermined	All victims who may or may have not been carrying a weapon	undetermined, unknown weapon
False Claim	These victims were falsely accused of having a weapon	claimed to be armed
Rods	All weapons that are cylindrical or circular	walking stick, pole and knife, pole, pipe, metal stick, metal pole, metal pipe, flagpole, baton
Toy Weapon	This was already an existing category hence will remain the same.	toy weapon

Table 4 - Category table for new Attribute ARMED_CATEGORY

An **if function** was used to recode the values into a new column 'ARMED_CATEGORY'

```
data WORK.IMPORT;
    length Armed_Category $ 50;
    set WORK.IMPORT;

    if armed='unarmed' then
        Armed_Category='Unarmed';

    if armed='toy weapon' then
        Armed_Category='Toy Weapon';

    if armed='undetermined' or armed='unknown weapon'  then
        Armed_Category='Undetermined';


    if armed='claimed to be armed' then
        Armed_Category='False Claim';

    if armed='air conditioner' or armed='barstool' or armed='beer bottle' or
        armed='binoculars'or armed='bottle' or armed='chain' or armed='chair'or
        armed='fireworks' or armed='flashlight' or armed='garden tool' or
        armed='grenade' or armed='incendiary device' or armed='metal object' or
        armed='microphone'or armed='pepper spray' or armed='Taser' or
        armed='tire iron' or armed='wasp spray' then
            Armed_Category='Other Weapon';

    if armed='Airsoft pistol' or armed='air pistol' or armed='BB gun and vehicle' or
        armed='BB gun' or armed='bean-bag gun' or armed='gun' or
        armed='guns and explosives' or armed='gun and vehicle'or armed='gun and knife' or
        armed='gun and car'or armed='gun and machete' or armed='gun and sword' or
        armed='nail gun' or armed='pellet gun' then
            Armed_Category='Gun';

    if armed='baseball bat' or armed='baseball bat and bottle' or armed='baseball bat and fireplace poker' or
    armed='baseball bat and knife' or armed='blunt object' or armed='brick' or
        armed='crowbar' or armed='hammer' or armed='hand torch' or
        armed='metal hand tool' or armed='oar'or armed='piece of wood' or
        armed='rock'or armed='screwdriver' or armed='shovel' or armed='stapler' or
        armed='wrench' then
            Armed_Category='Blunt Weapon';

    if armed='baton' or armed='flagpole' or armed='metal pipe' or armed='pipe' or
        armed='metal stick' or armed='metal pole' or armed='pole' or
        armed='pole and knife'or armed='walking stick' then
            Armed_Category='Rods';

    if armed='car, knife and mace' or armed='motorcycle'or armed='vehicle'or
        armed='vehicle and gun'or armed='vehicle and machete' then
            Armed_Category='Vehicle';

    if armed='ax' or armed='bayonet' or armed='bow and arrow' or
        armed='box cutter' or armed='carjack' or armed='chainsaw' or
        armed='chain saw' or armed='cordless drill'or armed='crossbow' or
        armed='glass shard'or armed='hatchet' or armed='hatchet and gun' or
        armed='ice pick' or armed='knife' or armed='knife and vehicle' or
        armed='lawn mower blade' or armed='machete' or armed='machete and gun' or
        armed='meat cleaver' or armed='metal rake' or armed='pen' or armed='pick-axe'
        or armed='pitchfork' or armed='railroad spikes' or armed='samurai sword' or
        armed='scissors'or armed='sharp object' or armed='spear' or
        armed='straight edge razor' or armed='sword' then
            Armed_Category='Sharp Weapon';
```

Figure 67 - code for categorizing armed into ARMED_CATEGORY

Next, the date was separated into the day, month, and year for better exploratory and predictive analysis.

```
data IMPORT;
set IMPORT;
day = Day(date);
run;
data IMPORT;
set IMPORT;
month = month(date);
run;
data IMPORT;
set IMPORT;
year = year(date);
run;
```

Figure 68 - code for extracting date

day	month	year
18	3	2020
22	3	2017
23	6	2018
16	2	2015
28	4	2016
26	9	2016

Figure 69 - the result of the day, month, year extraction from date

RACE column was listed as 1 letter to describe each race, this was recoded to specify each race accordingly as shown below.

```
data WORK.IMPORT;
    length Race_Label $ 10;
    set WORK.IMPORT;
    select (race);
        when ('W') Race_Label = 'White';
        when ('B') Race_Label = 'Black';
        when ('O') Race_Label = 'Other';
        when ('N') Race_Label = 'Native';
        when ('H') Race_Label = 'Hispanic';
        when ('A') Race_Label = 'Asian';
        otherwise Race_Label = race;
    end;
run;
```

Figure 70 - result of recoding

	Race_Label
1	Black
2	Black
3	White
4	White
5	White

Figure 71 - recoding race into appropriate labels

As mentioned above, multinomial analysis is unable to be carried out in SAS Enterprise Miner, hence the analysis can be done by recoding RACE into two unordered binary categories of 0 and 1.

According to (CSUSM, 2020), Natives and Asians can come under one category, known as APIDA: “East Asian, South Asian, Southeast Asian, and Pacific Islander populations”. This will also include Black victims. Hispanics may also include White Hispanics hence Whites, Hispanics and Other will be coded together.

```
data WORK.IMPORT;
    length Race_Coded $ 10;
    set WORK.IMPORT;

    select (race);
        when ('W') Race_Coded='0';
        when ('H') Race_Coded='0';
        when ('O') Race_Coded='0';
        when ('B') Race_Coded='1';
        when ('N') Race_Coded='1';
        when ('A') Race_Coded='1';
        otherwise Race_Coded=race;
    end;
run;
```

Figure 72 - recoding race into binary variables

Race_Coded	Race_Label
1	Black
1	Black
0	White
0	White
0	White
0	White
0	White
0	White

Figure 73 - the result of the recoding

W, H and O have been coded as 0 while B, N, A are coded as 1. Usually, Binary variables are known to take 1 or 0 as a ‘response’ or ‘no response’ however they may be used for grouping of categorical variables as well.

```
data WORK.IMPORT;
    length NewGender $ 6;
    set WORK.IMPORT;

    select (gender);
        when ('F') NewGender='Female';
        when ('M') NewGender='Male';
        otherwise NewGender=gender;
    end;
run;
```

NewGender
Male
Male
Male
Male
Male
Male
Male

Figure 74 & Figure 75 – recoding gender and result of the recoding

In the GENDER column, 'F' and 'M' were changed to their respective labels, Female and Male.

```
data WORK.IMPORT;
    length Manner_of_Death $ 20;
    set WORK.IMPORT;

    select (manner_of_death);
        when ('shot') Manner_of_Death='Shot';
        when ('shot and Tasered') Manner_of_Death='Shot and Tasered';
        otherwise manner_of_death=Manner_of_Death;
    end;
run;
```

MannerofDeath
Shot
Shot
Shot and Tasered
Shot
Shot
Shot

Figure 76 - recoding

Figure 77 - result of recoding

```
data WORK.IMPORT;
    length threat_level $ 20;
    set WORK.IMPORT;

    select (threat_level);
        when ('other') threat_level ='Other';
        when ('attack') threat_level ='Attack';
        when ('undetermined') threat_level ='Undetermined';
        otherwise threat_level = threat_level;
    end;
run;
```

	threatlevel
29	Attack
30	Attack
31	Attack
32	Attack
33	Attack
34	Attack
35	Other

Figure 78 - recoding

Figure 79 - result of recoding

Pre-processing may also include minor changes to the data such as capitalizing certain variables, this is done for MANNER_OF_DEATH as well as THREAT_LEVEL. Binary variables such as SIGNS_OF_MENTAL_ILLNESS, BODY_CAMERA and IS_GEOCODING_EXACT automatically changed to binary numerical upon import to SAS Studio in XLSX format.

```
data WORK.IMPORT;
    length state $ 2;
    set WORK.IMPORT;

    select (state);
        when ('WA') state='DC';
        otherwise state=state;
    end;
run;
```

Figure 80 - recoding of state

WA was changed to DC in the STATE column as Washington state is known as DC, not WA.

```
data WORK.IMPORT;
    length Region $ 20;
    set WORK.IMPORT;

    if state='IA' or state='IL' or state='IN' or state='KS' or state='MI' or
        state='MN' or state='MO' or state='ND' or state='NE' or state='OH' or
        state='SD' or state='WI' then
            Region='Midwest';

    if state='CT' or state='MA' or state='ME' or state='NH' or state='NJ' or
        state='NY' or state='PA' or state='RI' or state='VT'
        then Region='Northeast';

    if state='AL' or state='AR' or state='DE' or state='FL' or state='GA' or
        state='KY' or state='LA' or state='MD' or state='MS' or state='NC' or
        state='SC' or state='TN' or state='VA' or state='WV' then
            Region='Southeast';

    if state='AZ' or state='NM' or state='OK' or state='TX' then
            Region='Southwest';

    if state='AK' or state='CA' or state='CO' or state='DC' or state='HI' or
        state='ID' or state='MT' or state='NV' or state='OR' or state='UT' or
        state='WY' then Region='West';
```

Figure 81 - recoding to group state by region

All the states were grouped under 5 different regions, West, Southeast, Northeast, Southwest, and Midwest. They were grouped according to their location on the U.S Map (National Geographic Society, 2012)

SIGNS_OF_MENTAL_ILLNESS and RACE are the target variables, and they are converted to numerical binary variables, both are nominal however for SIGNS_OF_MENTAL_ILLNESS '1' stands for True while '0' stands for False meanwhile for RACE, 1 and 0 are referencing to the group of races under each number.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
id	name	age	Race_	Race_	race	gende	NewGi	Armed	armed	signs_	body_	date	day	month	year	Manni	threat	flee	Regio	state	city	county	manni	threat	longit	latitu	is_gec	ding_exact
3	Tim Ellio	53	1	Asian	A	M	Male	Gun	gun	1	0	1/2/2015	2	1	2015	Shot	Attack	Not fleei	West	DC	Shelton	Mason	shot	attack	-123.122	47.247	1	
4	Lewis Le	47	0	White	W	M	Male	Gun	gun	0	0	1/2/2015	2	1	2015	Shot	Attack	Not fleei	West	OR	Aloha	Washing	shot	attack	-122.892	45.487	1	
5	John Pau	23	0	Hispanic	H	M	Male	Unarmed	unarmed	0	0	1/3/2015	3	1	2015	Shot and	Other	Not fleei	Midwest	KS	Wichita	Sedgwick	shot and	other	-97.281	37.695	1	
8	Matthew	32	0	White	W	M	Male	Toy Weal	toy weap	1	0	1/4/2015	4	1	2015	Shot	Attack	Not fleei	West	CA	San Fran	San Fran	shot	attack	-122.422	37.763	1	
9	Michael I	39	0	Hispanic	H	M	Male	Gun	nail gun	0	0	1/4/2015	4	1	2015	Shot	Attack	Not fleei	West	CO	Evans	Weld	shot	attack	-104.692	40.384	1	
11	Kenneth	18	0	White	W	M	Male	Gun	gun	0	0	1/4/2015	4	1	2015	Shot	Attack	Not fleei	Southwe	OK	Guthrie	Logan	shot	attack	-97.423	35.877	1	
13	Kenneth	22	0	Hispanic	H	M	Male	Gun	gun	0	0	1/5/2015	5	1	2015	Shot	Attack	Car	Southwe	AZ	Chandler	Maricopa	shot	attack	-111.841	33.328	1	
15	Brock Nic	35	0	White	W	M	Male	Gun	gun	0	0	1/6/2015	6	1	2015	Shot	Attack	Not fleei	Midwest	KS	Assaria	Saline	shot	attack	-97.564	38.704	1	
16	Autumn S	34	0	White	W	F	Female	Unarmed	unarmed	0	1	1/6/2015	6	1	2015	Shot	Other	Not fleei	Midwest	IA	Burlingto	Des Moir	shot	other	-91.119	40.809	1	
17	Leslie Sa	47	1	Black	B	M	Male	Toy Weal	toy weap	0	0	1/6/2015	6	1	2015	Shot	Attack	Not fleei	Northeas	PA	Knoxville	Alleghen	shot	attack	-79.991	40.413	1	
19	Patrick W	25	0	White	W	M	Male	Sharp We	knife	0	0	1/6/2015	6	1	2015	Shot and	Attack	Not fleei	West	CA	Stockton	San Joaq	shot and	attack	-121.299	37.93	1	
20	Jessie He	17	0	Hispanic	H	F	Female	Vehicle	vehicle	0	0	01/26/2015	26	1	2015	Shot	Other	Not fleei	West	CO	Denver	Denver	shot	other	-104.909	39.754	1	
21	Ron Snee	31	1	Black	B	M	Male	Gun	gun	0	0	1/7/2015	7	1	2015	Shot	Attack	Not fleei	Southwe	TX	Freeport	Brazoria	shot	attack	-95.369	28.955	1	
22	Hashim H	41	1	Black	B	M	Male	Sharp We	knife	1	0	1/7/2015	7	1	2015	Shot	Other	Not fleei	Midwest	OH	Columbu	Franklin	shot	other	-82.885	39.999	1	
25	Nicholas	30	0	White	W	M	Male	Gun	gun	0	0	1/7/2015	7	1	2015	Shot	Attack	Car	Midwest	IA	Des Moir	Polk	shot	attack	-93.609	41.582	1	

Figure 82 - exporting after pre-processing in SAS Studio

The above is the cleaned dataset after exporting from SAS Studio, certain columns can either be deleted here or dropped in SAS Enterprise Miner, it is chosen to drop them in SAS EM.

No	Variable	Role	Level	Description	Original Values	Pre-processed in SAS Studio
1	Id	Input	Interval	This is a unique id to distinguish each victim	-	-
2	Name	Input	Nominal	Name of the victims	-	-
3	Date	Time_ID	Interval	Date the police shot the victim	DD/MM/ YYYY	-
4	Day	Input	Interval	Day extracted from Date	-	DD
5	Month	Input	Nominal	Month extracted from Date	-	MM
6	Year	Input	Interval	Year extracted from Date	-	YYYY
7	Manner_Of_Death	Input	Nominal	How was the victim killed?	shot shot and Tasered	Shot Shot and Tasered
8	Armed	Input	Nominal	Did the victim possess any weapons when they were shot	-	-
9	Armed_Category	Input	Nominal	-	-	Blunt Weapon Sharp Weapon Other Weapon Undetermined Unarmed False Claim Toy Weapon Gun Vehicle Rods
10	Age	Input	Interval	Age of the victims	6 - 91	-
11	Gender	Input	Nominal	Gender of the victims	F/M	Female/Male
12	Race_Label	Input	Nominal	Race of Victims	A/B/H/O /N	Asian, Black, Hispanic, Other, Native
14	Race	Target	Binary	Race of Victim separated into two categories	1 = Asian, Black, and Native 0 = White, Hispanic , and Other	1/0
15	City	Input	Nominal	City victim was in when they were shot	-	-
16	State	Input	Nominal	The State that the victim was shot in	-	WA changed to DC; the rest

						remain the same.
17	Region	Input	Nominal	States grouped into Regions	-	Southeast, Southwest, Northeast, West, Midwest
18	Signs_Of_Mental_Illness	Target	Binary	Whether the victim displayed signs of mental illness	True/False	1/0
19	Mental_iIlness	Input	Binary	The original mental illness column	TRUE/FALSE	Yes/No
20	Threat_Level	Input	Nominal	Was the victim a threat to themselves and anyone around them	attack/other/undetermined	Attack/Other/Undetermined
21	Flee	Input	Nominal	Did the victim try to flee and with what methods	Car/foot/fleeing/other/blank	-
22	Body_Camera	Input	Binary	If the police officer was wearing a body camera or not during the killing	True/False	1/0
23	Longitude	Input	Interval	Coordinates where the shooting took place	-	-
24	Latitude	Input	Interval		-	-
25	Is_Geocoding_Exact	Input	Interval	Whether or not the coordinates are exact or not.	True/False	1/0

Table 5 – pre-processed dataset ready for further pre-processing in SAS Enterprise Miner

5.3.2 Data Cleaning – After Importing to SAS Enterprise Miner

After the data has been cleaned in SAS Studio, it is imported to SAS Enterprise Miner and the file import node is run again.

Name	Role	Level
ID	ID	Interval
Race_Label	Input	Nominal
Month	Input	Nominal
Age	Input	Interval
Day	Input	Interval
Armed_Category	Input	Nominal
Flee	Input	Nominal
MannerofDeath	Input	Nominal
Region	Input	Nominal
Threat_Level	Input	Nominal
Is_Geocoding_Exact	Input	Binary
Longitude	Input	Interval
Mental_Illness	Input	Binary
Body_Camera	Input	Binary
Gender	Input	Binary
Year	Input	Interval
Latitude	Input	Interval
State	Rejected	Nominal
County	Rejected	Nominal
City	Rejected	Nominal
Armed	Rejected	Nominal
Name	Rejected	Nominal
Race	Target	Binary
Signs_Of_Mental_Illness	Target	Binary
Date	Time ID	Interval
OLD_DATE	Time ID	Interval

Figure 83 - Variables in FileImport Node

It can be seen that there are now two target binary variables which are RACE and SIGNS_OF_MENTAL_ILLNESS. The variables that were rejected before are rejected again after running the import for the same reason as above, the comment can be seen for them under the comment column, each variables format type may also be seen under the Format Type column. Two columns RACE_LABEL and MENTAL_ILLNESS will be used as input nominal and binary variables for modelling with SIGNS_OF_MENTAL_ILLNESS and RACE_LABEL, respectively.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	BINARY	2
INPUT	INTERVAL	6
INPUT	NOMINAL	7
REJECTED	NOMINAL	5
TARGET	BINARY	2
TIMEID	INTERVAL	2

The CONTENTS Procedure

Data Set Name	EMWS3.FIMPORT_DATA	Observations	5704
Member Type	DATA	Variables	25
Engine	V9	Indexes	0
Created	06/16/2021 12:06:09	Observation Length	336
Last Modified	06/16/2021 12:06:09	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 84 - Output of FileImport node showing the total number of variables and records/observations

5.3.2.1 Drop Node – Removing Redundant Columns

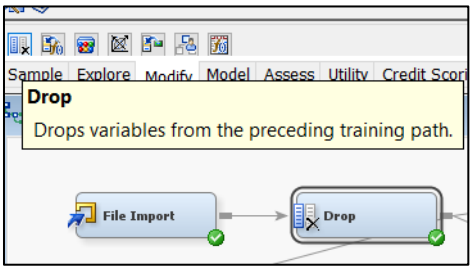
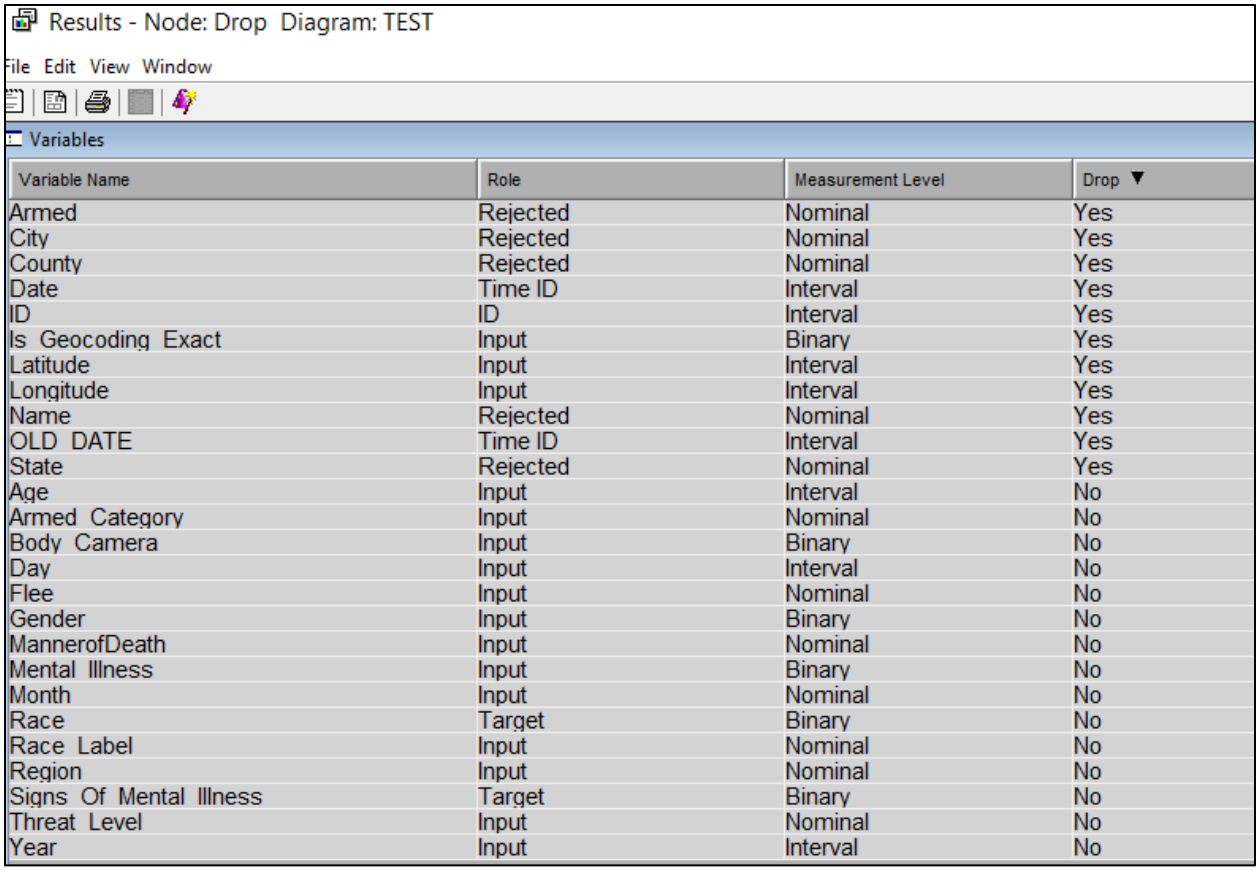


Figure 85 - Drop Node

The drop node is then dragged to the workspace and connected to the file import node. The Drop node is useful for removing variables from the data set or metadata. It can then be used to remove variables from a data set that was formed by the previous node. In this dataset, it will be used to drop any other columns which may not be relevant such as CITY, STATE, NAME, ID, DATE, OLD_DATE, IS_GEOCODING_EXACT, LONGITUDE and LATITUDE as these have either been grouped into categories or will not be providing any results to the analysis.



Variable Name	Role	Measurement Level	Drop
Armed	Rejected	Nominal	Yes
City	Rejected	Nominal	Yes
County	Rejected	Nominal	Yes
Date	Time ID	Interval	Yes
ID	ID	Interval	Yes
Is Geocoding Exact	Input	Binary	Yes
Latitude	Input	Interval	Yes
Longitude	Input	Interval	Yes
Name	Rejected	Nominal	Yes
OLD DATE	Time ID	Interval	Yes
State	Rejected	Nominal	Yes
Age	Input	Interval	No
Armed Category	Input	Nominal	No
Body Camera	Input	Binary	No
Day	Input	Interval	No
Flee	Input	Nominal	No
Gender	Input	Binary	No
MannerofDeath	Input	Nominal	No
Mental Illness	Input	Binary	No
Month	Input	Nominal	No
Race	Target	Binary	No
Race Label	Input	Nominal	No
Region	Input	Nominal	No
Signs Of Mental Illness	Target	Binary	No
Threat Level	Input	Nominal	No
Year	Input	Interval	No

Figure 86 – Variables are shown in Drop Node

The result of the drop column is shown above. Columns that were rejected and were analysed as being irrelevant have been dropped.

5.3.2.2 Replacement Node – Missing Input Categorical Variables

After dropping, the replacement node is dragged from Modify tab and connected to the drop node.

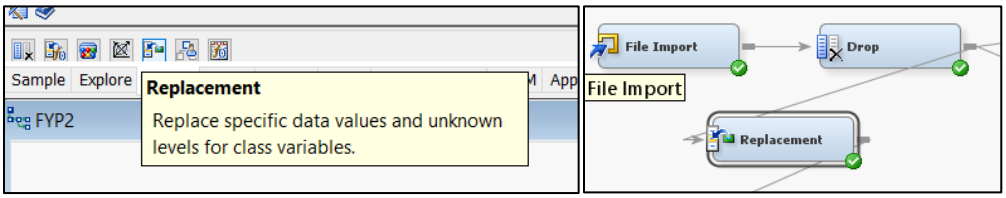


Figure 87 & Figure 88 – replacement node and connected to the previous node

A replacement node is used to prevent loss of data while being able to filter it, unlike the Filter node, which allows the altering of distributions of any variable with the dataset and changes bi-modal or distorted distribution more balanced. Missing class variable values can be replaced with the most predominant value, distribution-based replacement, tree-based imputation, or a constant

while missing interval variables can be replaced using mean, median, mid-range and more. The Class Variable Replacement "editor" under the replacement node's property window is clicked to replace the missing values.

For ARMED_CATEGORY there are 261 missing values, and one category already exists under the name "Undetermined" hence it is decided to categorize them together along with the 'False Claim' as it is only 1 record. Blunt Weapon and Rods have also been grouped under 'Blunt Weapon'.

For FLEE, the missing 331 values and the 'Other' 200 values have been grouped under 'Undetermined'. According to the Washington Post, they have defined FLEE as something that can either happen before or after a police chase, so it seems appropriate to label this as undetermined as in those cases it was not known how the victim reacted.

Replacement Values for Class Variables

Variable	Formatted Value	Type	Character Unformatted Value	Numeric Value	Replacement Value
Armed_Category		C		.	Undetermined
Armed_Category	Blunt Weapon	C	Blunt Weapon	.	Blunt Weapon
Armed_Category	Rods	C	Rods	.	Blunt Weapon
Armed_Category	False Claim	C	False Claim	.	Undetermined
Flee		C		.	Undetermined
Flee	Other	C	Other	.	Undetermined
Mental_Illness	No	C	No	.	No
Mental_Illness	Ye	C	Ye	.	Yes

Figure 89 - Class Variable Replacement Editor

5.3.2.3 Impute Node – Missing Input Numerical Variable

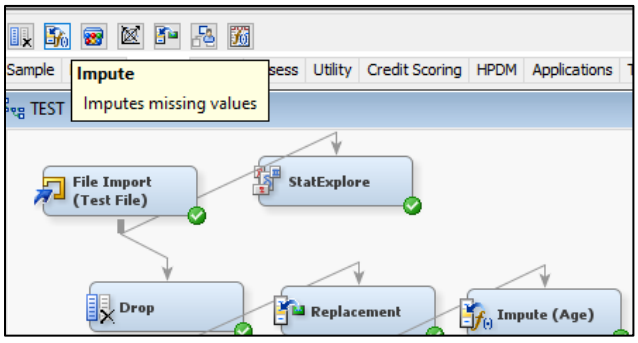


Figure 90 - Impute Node

After the categorical variables have been replaced, the impute node is dragged from the modify tab and connected to the replacement node and 'Variables' under the Train heading in the property window is clicked. It opens up a table showing the variables; according to the StatExplore node, AGE had 104 missing values, so AGE is selected for imputation with the mean method. Various other methods can be used for both interval and categorical variables. Some of them include maximum, minimum, medium, count, tree surrogate and amongst others. (Sarma, 2017)

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP_Age	36.6452	INPUT	INTERVAL		104

Figure 91 - Result of Impute Node

The StatExplore node is connected to the impute node to show the updated chi-squares of all the variables after the pre-processing has been complete. The results are shown below for both targets. After all the pre-processing steps, the new chi-square values and their probabilities can be seen against the target variables.

Data Role=TRAIN Target=Race

Input	Chi-Square	Df	Prob
Race_Label	5704.0000	5	<.0001
IMP_Age	193.2688	4	<.0001
Region	167.4103	4	<.0001
REP_Flee	71.1921	3	<.0001
REP_Mental_Illness	67.0959	1	<.0001
Signs_Of_Mental_Illness	67.0959	1	<.0001
Body_Camera	49.8234	1	<.0001
REP_Armed_Category	27.3814	7	0.0003
Month	12.2017	11	0.3487
Gender	6.4476	1	0.0111
Day	4.8074	4	0.3076
Year	2.9815	4	0.5609
Threat_Level	2.1129	2	0.3477
MannerofDeath	1.1460	1	0.2844

Figure 92 – Chi-Square for RACE

Data Role=TRAIN Target=Signs_Of_Mental_Illness

Input	Chi-Square	Df	Prob
REP_Mental_Illness	5704.0000	1	<.0001
REP_Armed_Category	269.7280	7	<.0001
REP_Flee	269.5163	3	<.0001
Race_Label	134.7714	5	<.0001
Race	67.0959	1	<.0001
IMP_Age	63.0578	4	<.0001
Threat_Level	20.4667	2	<.0001
MannerofDeath	17.7935	1	<.0001
Year	17.4700	4	0.0016
Gender	16.2422	1	<.0001
Region	16.0933	4	0.0029
Month	15.7697	11	0.1499
Body_Camera	14.3538	1	0.0002
Day	10.5247	4	0.0325

Figure 93 - Chi-Square for SIGNS_OF_MENTAL_ILLNESS

Here, it can be seen that RACE_LABEL shows up when the target is RACE. This can be ignored as when prediction modelling will be done for RACE, the RACE_LABEL will be set as rejected. Moreover, SIGNS_OF_MENTAL_ILLNESS and REP_MENTAL_ILLNESS can be seen. These two have the same properties and are both binary, however, REP_MENTAL_ILLNESS is needed in the table for use as an input variable when modelling with RACE. The same justification can be said for when modelling with SIGNS_OF_MENTAL_ILLNESS as a target. The REP_MENTAL_ILLNESS will be rejected then. Moreover, RACE_LABEL will be used as an input in its natural nominal form as it shows that it has a high chi-square value as compared to if it was kept as the binary form.

5.4 Data Visualization

This section of the chapter will be used to carry out some visualization on the preprocessed data. Now that the missing values have been taken care of and all the relevant information has been replaced appropriately, good, and actionable insights can be gained from the below visualizations. A few charts have been visualized concerning chi-square and self-understanding of possible important variables. A brief explanation regarding how the data was exported from SAS EM and imported into PowerBI is given in Section 5.6.

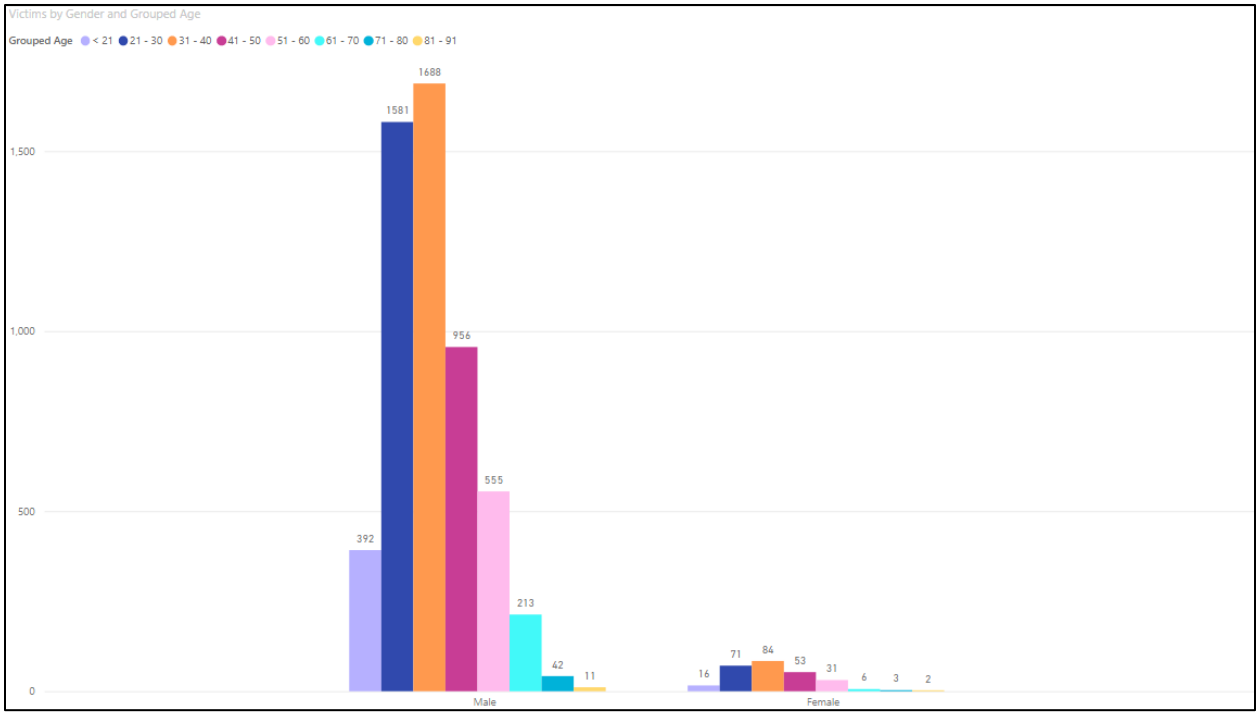


Figure 94 – Cluster Bar Plot of Victims by gender and grouped age

This chart displays the victims by gender and grouped age. It is already understood that the majority of the victims are men, this is because they are more likely to be shot as compared to women. The reasons for this are touched upon both in the literature review as well as the discussion after the evaluation of the prediction models.

Overall, it can be seen that for both genders, age groups, 21-30 and 31-40 are high in number for them, respectively. For male victims, 1668 fall between 31- 40 while for female victims, this age group is only 84. It is still the highest number for females. The age group with the lowest number of victims 61 and above with 81–91-year-old victims only making up 11 and 2 of the victims for male and female victims, respectively. This is further analysed below in terms of race.

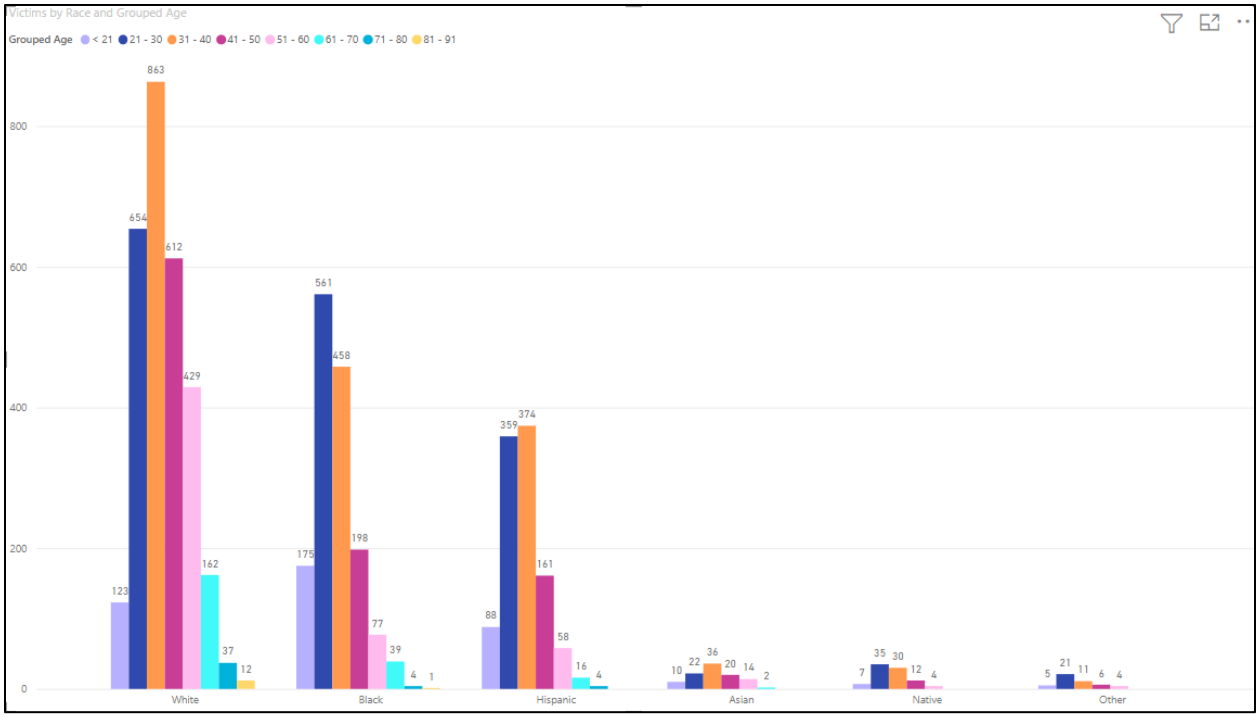


Figure 95 - Clustered Bar Plot of Victims by Race and Age

Victims of FPS were analysed by AGE and GENDER, and now it will be looked at from the perspective of RACE and AGE. AGE was one of the most important variables in terms of chi-square for both mental illness and race.

From this graph, it can be deduced that victims are who fall in the ranges of 21 – 30 and 31 – 40 years old are most in number with White victims taking 863 victims out of the total, following them are 21–30-year-old victims at 654 total. As compared to black victims, they seem to be younger than the White victims with 21 – 30-year-old is being 561 of the total number of black victims. The age group of 31-40 make up over 31% of the total victims. Native and Asian victims are the victims with the lowest number. The oldest victims were few with the highest out of this category being 12 under White, 81–91-year-old victims. For Hispanic victims, the difference between the 21-30 and 31- 40 age group was not much as compared to the Black and White victims. This shows that Hispanics between the ages of 21-40 years old were more likely to be shot.

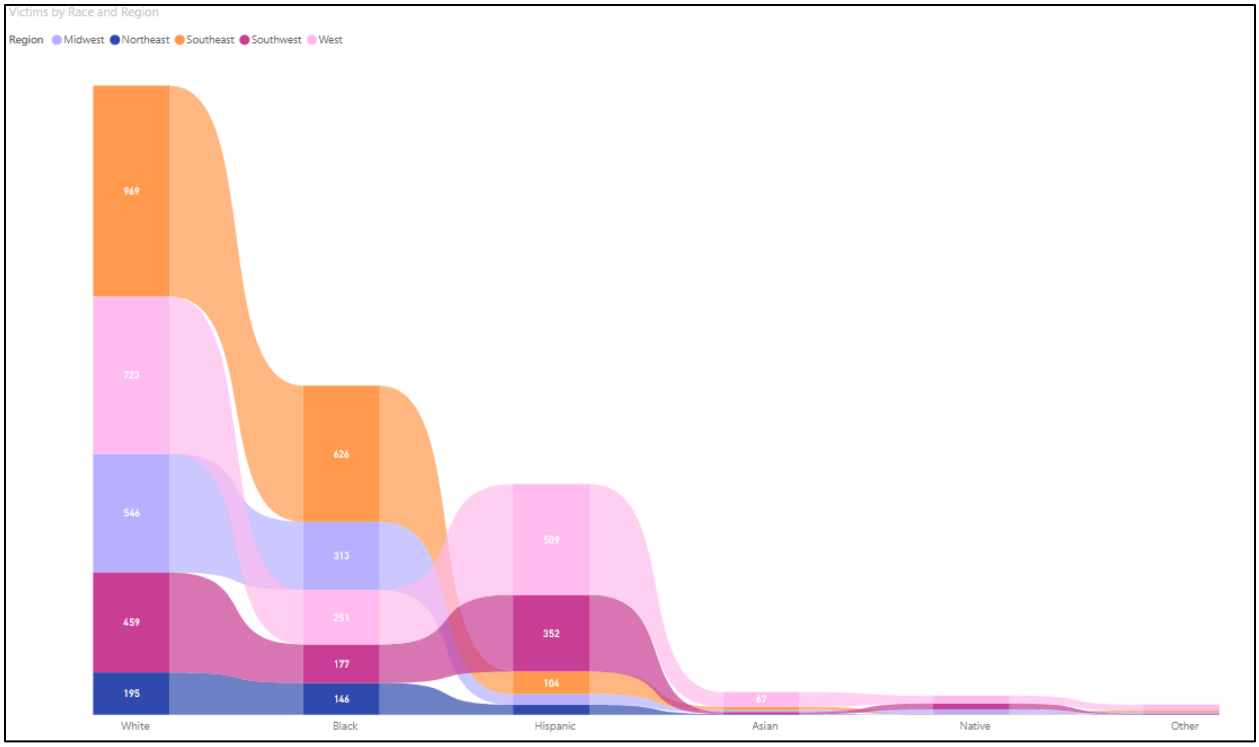


Figure 96 - Ribbon Chart for Victims by Race and Region

This is a ribbon chart showing the victims by race and region. The southeast region can be seen leading in terms of the number of victims with it having the highest number for both black and white victims with 626 and 969 respectively, however, seems to taper off for Hispanic victims and has no victims or few victims from the remaining three races: Asian, Native and Others.

Southeast in total has 1708 victims across all races followed by the West region at 1614 victims and Midwest region at 948 victims. Northeast has the least number of victims for all the races. Hispanic and White victims are prone to get killed in the West region at 509 and 723 respectively as compared to Black victims at 251. Law enforcement should consider this and see exactly what other factors are contributing to these vast number of differences and find ways to reduce the killings.

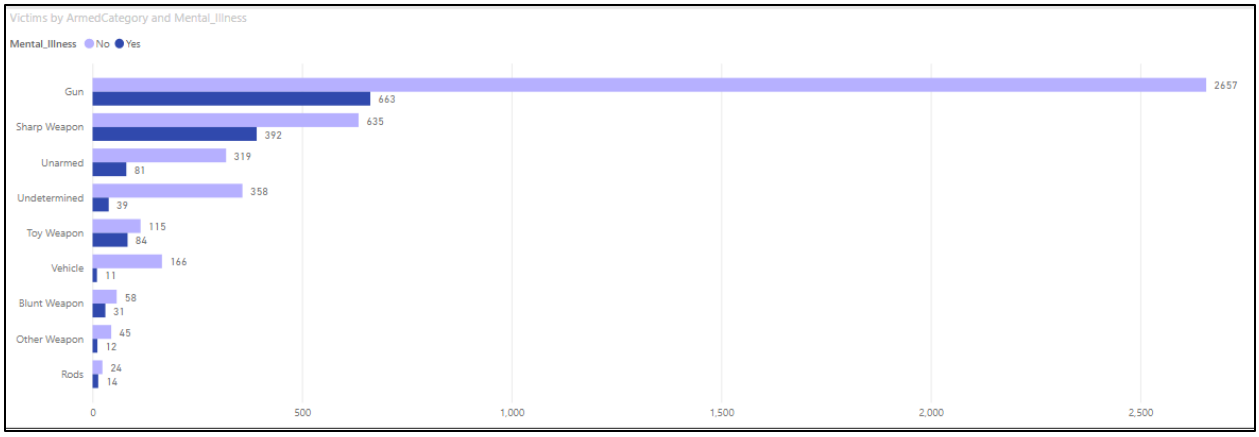


Figure 97 - Victims by Armed Category and Signs of Mental Illness

This chart compares the victims of FPS by ARMED_CATEGORY and MENTAL_ILLNESS. At a glance, it can be seen that most of the time when a victim was armed, there were less likely to have any signs of mental illness. To note; for this research, signs of mental illness do not include whether or not a person was under the influence of any types of drugs or alcohol, in case that data was available, this could be a more productive graph. Most victims resorted to holding a gun with over 2637 victims showing no signs of mental illness and 663 showing signs of mental illness. This can also be attributed to the fact that in the US, there are no laws in place against the possession of guns, so the average American has a type of gun in their possession. (Masters, 2019)

Sharp Weapon's category shows a difference that is less than any other category, for mentally ill victims, 635 possessed a sharp weapon while 392 did not. Unarmed victims did not have any weapons and also made up 7% of the victims who were shot regardless. For victims who were not known to have had possession of a weapon, about 358 of them were killed by the police while showing no signs of mental illness.

These results can be used to see why a police officer shot an unarmed civilian without a reason, police officers are usually trained on how to deescalate a situation, and now there are some states which make them wear a body camera to ensure the truth is shown in fatal police encounters. Law enforcement officials should take note of this data to enforce stricter gun laws that prevent normal civilians from possessing a gun and possibly help reduce these shootings. Other countries have already moved towards enforcing strict gun possession laws, the U.S should also follow suit. (Masters, 2019)

5.5 Modelling/Data-driven product

After the data exploration, pre-processing and visualization was completed, the modelling process was carried out. For this project, it was decided to build decision trees and regression models for both target variables as they were the best choice in terms of prediction models. Both are target variables are binary variables hence these prediction models are binary classification models.

5.5.1 Pre-Modelling Process

This section describes the usage of transform variable and variable selection nodes in the modelling process. For both target variables, both nodes were utilized in a variety of ways before the best method to carry out the modelling process was decided.

5.5.1.1 Data Partition Node – Validation, Training and Testing Dataset

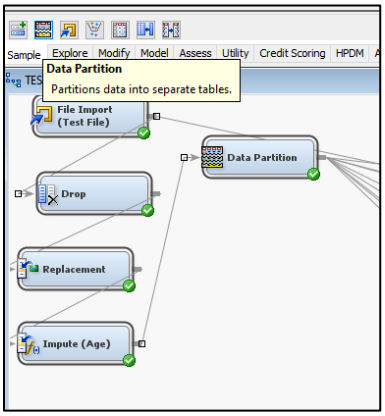


Figure 98 - Data Partition Node

Next, the data partition node is dragged and connected to the impute node to partition the dataset into two or three sub-datasets: Validation, Training and Testing, and connect to the Impute Node. The training dataset is later used in our modelling techniques: Decision Tree and Regression while Validation data will be used for evaluation of these modelling techniques and choose the most optimal one. Fine-tuning is a term used to describe the process of picking the optimal model. The Test data can be utilized to conduct an independent evaluation of the chosen model. This may also be carried out using the Validation dataset. (SAS Institute Inc, 2003)

In the property window for this node under the Train subheading "Partitioning Method"; there are four types of methods: Stratified, Cluster, Simple Random or Default. Default and Stratified are the same and are preferable for binary target variables hence we leave it as 'Default'. Training, Validation and Test dataset are kept as 70, 30, 0 respectively. A research carried out by Dhanabal

et.al (2020) used a data partition of 70,30, 0 hence it will be used as a reference for the models. Mashinchi (2020) carried out an analysis with both mental illness and race as target variables and allotted 80 per cent of their data to the training dataset and 20 per cent of the data to the validation dataset. Moreover, according to Georges et al. (2010), allotting test data a certain percentage can be considered wasteful and is much better utilized in allotting it to the training and validation dataset. Having a higher percentage allotted to training may also help prevent overfitting when building prediction models.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS7.Impt2_TRAIN	5704
TRAIN	EMWS7.Part_TRAIN	3990
VALIDATE	EMWS7.Part_VALIDATE	1714

Figure 99 - Partition Summary

The partition summary displays the dataset that has been divided into subsamples of train and validate data along with its observation, 3990 and 1714, respectively.

5.5.1.2 Transform Variable Node – Modifying Categorical Inputs

Transforming variables allows conducting statistical analysis with limited erroneous values, however, it also changes the units and characteristics of a variable hence it is important to carry out this process when crucially required. (Lee, 2020)

A transform variable node can provide a multitude of ways to transform interval inputs such as using best which takes the best transformation based on best chi-square value while others include, binning, bucket, quantile, maximum normal etc. This may also be referred to as a numeric variable transformation; for some models such as Support Vector Machine, it can be seen as an important step however for models such as Decision Trees, it may not be required as they are not as sensitive to the scale and skewness as other models. (Dei, 2019)

To transform class inputs, there are two ways; dummy indicators and grouping rare levels. The former gives the class variables each a value of 0 or 1, while the latter combines all the important levels into another group known as _OTHER_. A rare level is defined by setting a cut-off value

through the use of the Cut-off Value property. (Sarma, 2017) This type of transformation may also be known as categorical variable transformation or encoding. For most modelling techniques this is very important as they are only able to work with numeric variables. (Dei, 2019)

A transform variable node was dragged from the modify tab and connected to the data partition node for use in individual prediction models of the dependent variables. The default setting was kept as is, and only the class variables with more than 2 levels were transformed using dummy indicators except for RACE_LABEL.

Transformations Statistics

Source	Method	Variable Name	Number of Levels	Formula	Label	Non Missing	Missing
Input	Original	REP_Armed_Category	8		Replacement: Armed_Category		0
Input	Original	REP_Flee	4		Replacement: Flee		0
Input	Original	Region	5				0
Input	Original	Threat_Level	3				0
Output	Computed	TI_REP_Armed_Category1	2	Dummy	REP_Armed_Category:Blunt Weapon		0
Output	Computed	TI_REP_Armed_Category2	2	Dummy	REP_Armed_Category:Gun		0
Output	Computed	TI_REP_Armed_Category3	2	Dummy	REP_Armed_Category:Other Weapon		0
Output	Computed	TI_REP_Armed_Category4	2	Dummy	REP_Armed_Category:Sharp Weapon		0
Output	Computed	TI_REP_Armed_Category5	2	Dummy	REP_Armed_Category:Toy Weapon		0
Output	Computed	TI_REP_Armed_Category6	2	Dummy	REP_Armed_Category:Unarmed		0
Output	Computed	TI_REP_Armed_Category7	2	Dummy	REP_Armed_Category:Undetermined		0
Output	Computed	TI_REP_Armed_Category8	2	Dummy	REP_Armed_Category:Vehicle		0
Output	Computed	TI_REP_Flee1	2	Dummy	REP_Flee:Car		0
Output	Computed	TI_REP_Flee2	2	Dummy	REP_Flee:Foot		0
Output	Computed	TI_REP_Flee3	2	Dummy	REP_Flee:Not fleeing		0
Output	Computed	TI_REP_Flee4	2	Dummy	REP_Flee:Undetermined		0
Output	Computed	TI_Region1	2	Dummy	Region:Midwest		0
Output	Computed	TI_Region2	2	Dummy	Region:Northeast		0
Output	Computed	TI_Region3	2	Dummy	Region:Southeast		0
Output	Computed	TI_Region4	2	Dummy	Region:Southwest		0
Output	Computed	TI_Region5	2	Dummy	Region:West		0
Output	Computed	TI_Threat_Level1	2	Dummy	Threat_Level:Attack		0
Output	Computed	TI_Threat_Level2	2	Dummy	Threat_Level:Other		0
Output	Computed	TI_Threat_Level3	2	Dummy	Threat_Level:Undetermined		0

Figure 100 - Transformed Class Variables

For AGE, log was initially used to see whether there would be any type of advantage towards the modelling process however it produces a similar result hence it was decided that it was best to forego any transformation.

5.5.1.3 Variable Selection Node – Selection of Best Independent Inputs

The variable selection allows the choice of the most determinant variables to the modelling process and removing the irrelevant variables, this can help in producing models with the best fit and provide accurate predictions. The dataset is relatively small compared to the large databases hence it will be easier to carry out variable selection; as the targets are binary categorical variables, both Chi-Square and R-square criteria can be applied.

R-Square and Chi-Square Criterion

Chi-square is only used when there are binary targets and unlike R-Square, it constructs a tree based on the chi-square which can also be referred to as a CHAID and is an important criterion when logistic regression is to be built and identifies all the important input variables however an

R-square criterion helps in reducing loss of information by not binning the continuous variables hence it is suggested to use it alongside the chi-square criterion as variables usually overlap in importance and pass as selected inputs when the variable selection node is run. The R-Square criterion works by implementing a 'forward stepwise least squares regression' which boosts the model R-Square value. Three steps happen with this criterion which is 'computing the squared correlations' where the R2 value of each input is calculated and compared with minimum R-square value, the 'forward stepwise regression' and finally the 'logistic regression for binary targets. (Sarma, 2017)

According to Sarma (2017), after the variables are selected, the rejected variables will not come forth in the predictive modelling process. Moreover, for both target individual variable selection nodes are used as for there is a difference between the importance for variables when they are used as separate targets. The variable selection node is dragged from the explore tab and connected to the data partition node as well as to the transform variable node as needed. The overall setting is kept default except, the selection of the target model, minimum R-square and use group variables are changed to 'R and Chi-square', '0.002' and 'No', respectively.

Train	
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	R and Chi-square
Manual Selector	...
Rejects Unused Input	Yes
Bypass Options	
Variable	None
Role	Input
Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.002
Stop R-Square	5.0E-4
Use AOV16 Variables	No
Use Group Variables	No
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default

Figure 101 - Property Window for Variable Selection Node

There are three variable selection nodes, and each will be discussed below.

Figure 102 - Transformed Variable Selection

This variable selection node selects the best-transformed variables passed from the Transform Variable node. It selects independent variables based on the target variable RACE. This node was followed by a regression node.

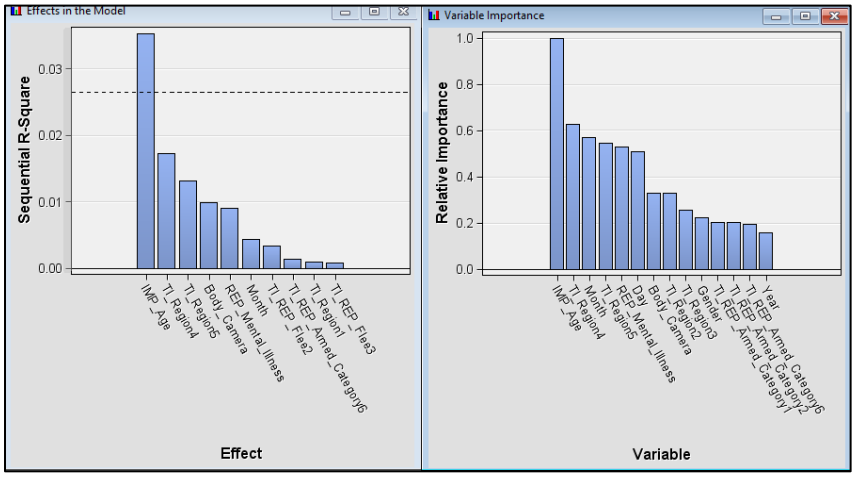


Figure 103

The variables that overlap both in the variable importance plot and the effect plot are the selected variables that pass through the variable selection node.

Variable Selection

Variable Name	Label	Reasons for Rejection ▲
Body_Camera		
IMP_Age	Imputed Age	
Month		
REP_Mental_Illness	Replacement: Mental_Illness	
TI_REP_Armed_Category6	REP_Armed_Category:Unarmed	
TI_Region4	Region:Southwest	
TI_Region5	Region:West	
TI_REP_Flee2	REP_Flee:Foot	Varsel:Small Chi-square value
TI_REP_Flee3	REP_Flee:Not fleeing	Varsel:Small Chi-square value
TI_Region1	Region:Midwest	Varsel:Small Chi-square value
Day		Varsel:Small R-square value
Gender		Varsel:Small R-square value
TI_REP_Armed_Category1	REP_Armed_Category:Blunt Weapon	Varsel:Small R-square value
TI_REP_Armed_Category2	REP_Armed_Category:Gun	Varsel:Small R-square value
TI_Region2	Region:Northeast	Varsel:Small R-square value
TI_Region3	Region:Southeast	Varsel:Small R-square value
Year		Varsel:Small R-square value
MannerofDeath		Varsel:Small R-square value, Small Chi-squar...
TI_REP_Armed_Category3	REP_Armed_Category:Other Weapon	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Armed_Category4	REP_Armed_Category:Sharp Weapon	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Armed_Category5	REP_Armed_Category:Toy Weapon	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Armed_Category7	REP_Armed_Category:Undetermined	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Armed_Category8	REP_Armed_Category:Vehicle	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Flee1	REP_Flee:Car	Varsel:Small R-square value, Small Chi-squar...
TI_REP_Flee4	REP_Flee:Undetermined	Varsel:Small R-square value, Small Chi-squar...
TI_Threat_Level1	Threat_Level:Attack	Varsel:Small R-square value, Small Chi-squar...
TI_Threat_Level2	Threat_Level:Other	Varsel:Small R-square value, Small Chi-squar...
TI_Threat_Level3	Threat_Level:Undetermined	Varsel:Small R-square value, Small Chi-squar...

Figure 104 - Variable Selection

The selected variables are shown below as well as the rejected values along with their reasons for rejections. It can be seen that half the variables were rejected based on both their chi-square and r-square values.

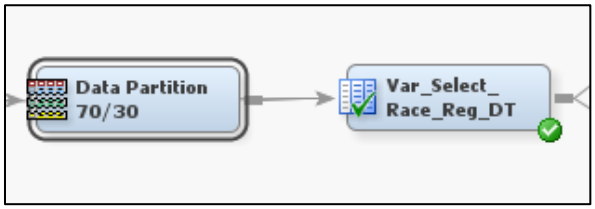


Figure 105 - Pure Variable Selection

This variable selection is also based on the use of RACE as a target variable to select the important features from the dataset, this variable is followed by a regression node as well as a decision tree node.

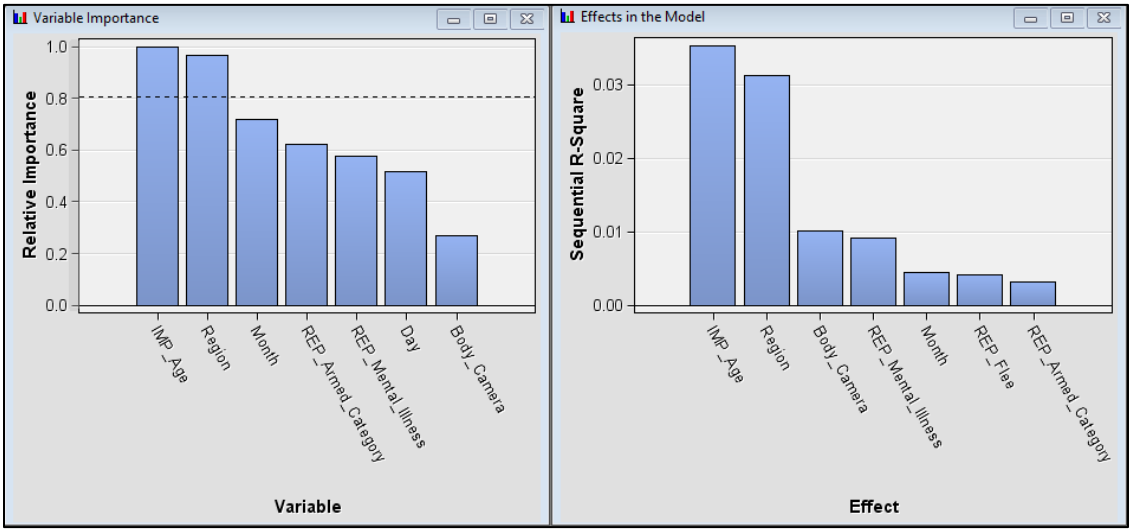


Figure 106 –Variable Importance and Effects Plot for Regular Variable Selection

Variable Selection

Variable Name	Role	Reasons for Rejection ▲
Body_Camera	Input	
IMP_Age	Input	
Month	Input	
REP_Armed_Category	Input	
REP_Mental_Illness	Input	
Region	Input	
REP_Flee	Rejected	Varsel5:Small Chi-square value
Day	Rejected	Varsel5:Small R-square value
Gender	Rejected	Varsel5:Small R-square value, Small Chi-square value
MannerofDeath	Rejected	Varsel5:Small R-square value, Small Chi-square value
Threat_Level	Rejected	Varsel5:Small R-square value, Small Chi-square value
Year	Rejected	Varsel5:Small R-square value, Small Chi-square value

Figure 107 – Results of Variable Selection

For both the transformed variable selection and the regular variable selection for the RACE target, it can be seen that they have selected variables from the same attributes which show the importance of these variables to the target.

Figure 108 - Pure Variable Selection

The final variable selection node takes SIGNS_OF_MENTAL_ILLNESS as a target variable to select the most important variables. No transformed variables were passed through this variable selection node for this target as they did not produce any results in the modelling process. Moreover, it was used only in the regression model for the target variable.

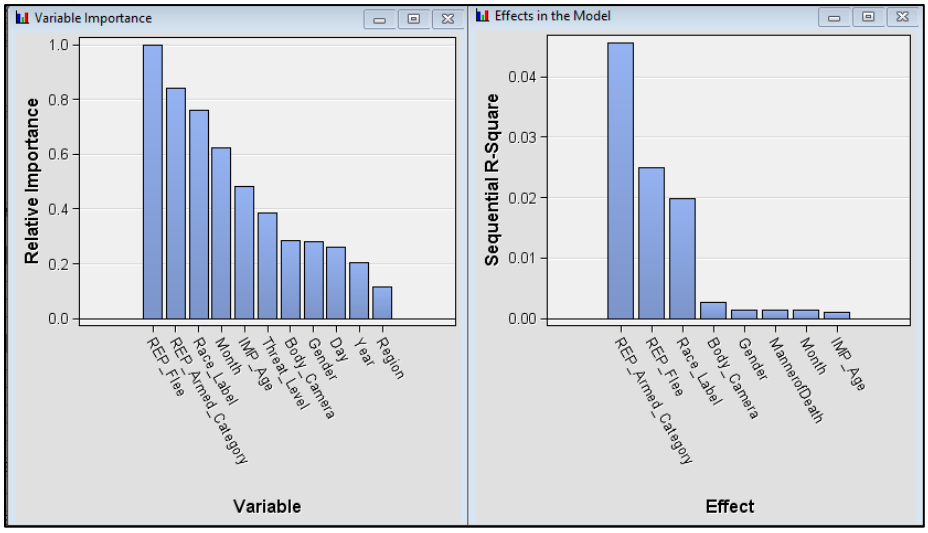


Figure 109 - Variable Importance and Effects Plot for Variable Selection

Variable Selection

Variable Name	Role	Reasons for Rejection ▲
Body_Camera	Input	
Gender	Input	
IMP_Age	Input	
Month	Input	
REP_Armed_Category	Input	
REP_Flee	Input	
Race_Label	Input	
MannerofDeath	Rejected	Varsel2:Small Chi-square value
Day	Rejected	Varsel2:Small R-square value
Region	Rejected	Varsel2:Small R-square value
Threat_Level	Rejected	Varsel2:Small R-square value
Year	Rejected	Varsel2:Small R-square value

Figure 110 - Variable Selection Result

It was tried to pass variables by only using the chi-square criterion however it was unnecessary as the decision tree and regression would not report them in the output. However, when both Chi-square and R-Square criterion was used, it provided the most optimal variables which were suitable for the models that employed variable selection. Multiple methods of applying the variable selection to the modelling process were used and the above settings were decided after continuous trial and error. It can be seen that the variable section is used in the decision tree for the RACE target, this was done as it provided a better model. More on this will be discussed in the modelling subsection.

5.5.2 Influence of Race on Fatal Police Shootings

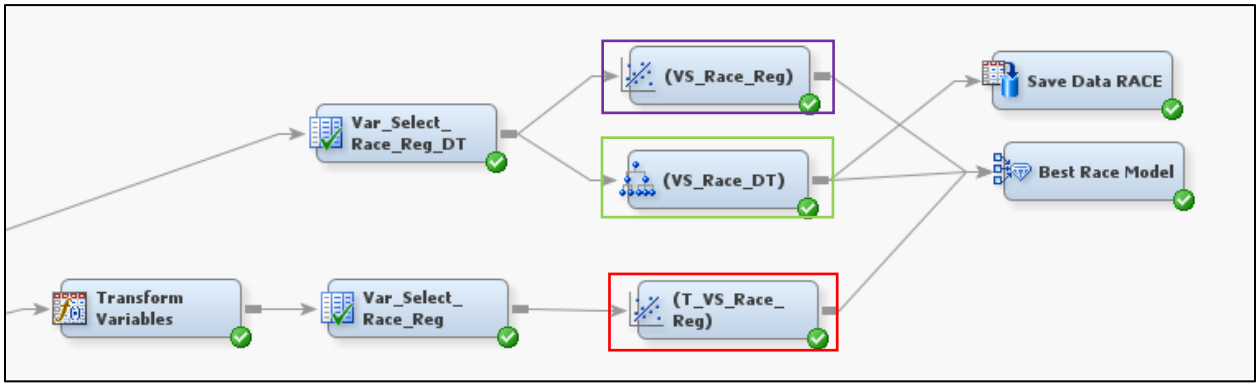


Figure 111 - Prediction Model for the RACE target variable

Three models were made for the RACE target variable to analyse the influence of factors on the race of victims in fatal police shootings. The decision tree and logistic regression model was built, one of the regression models was built using the transform variable node while the other simply used the variable selection node's output to build the model. The variable selection node was also used in the decision tree. Overall, the decision tree was the best model in terms of performance on the validation data.

5.5.2.1 Logistic Regression

Model Settings

The regression type chosen was 'Logistic Regression' as our target variable was binary (Sarma, 2017), and the link function was 'Logit' by default. In the property window for the regression node, under the model selection, the selection model chosen was 'Stepwise' and the selection criterion was kept at 'Validation Misclassification' as for binary categorical variables, it is the

most suitable optimization and criterion method, and lastly the Use Selection Defaults was changed to 'No'. The remaining properties were kept as the default.

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Misclassification
Use Selection Defaults	No
Selection Options	...

Figure 112 & 113 - Property Window

The Entry Significance level, Stay significance level and Maximum Number of Steps were changed and tested with 1.0, 0.5 and 30 respectively as well as other possible values were tried with a different number of steps however it did not optimize the result hence it was decided to keep it at default.

T_VS_Race Regression Model

The *T_VS_Race* Regression model (shown in the red box in Fig. 91) was built by transforming the variables and selecting the best variables according to the chi-square and r-square criteria. The results of the model are shown below.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Race		AIC	Akaike's Informati...	4521.434	.
Race		ASE	Average Squared ...	0.190167	0.188743
Race		AVERR	Average Error Fun...	0.565092	0.560296
Race		DFE	Degrees of Freed...	3984	.
Race		DFM	Model Degrees of ...	6	.
Race		DFT	Total Degrees of ...	3990	.
Race		DIV	Divisor for ASE	7980	3428
Race		ERR	Error Function	4509.434	1920.694
Race		FPE	Final Prediction Er...	0.19074	.
Race		MAX	Maximum Absolut...	0.95911	0.9557
Race		MSE	Mean Square Error	0.190454	0.188743
Race		NOBS	Sum of Frequencies	3990	1714
Race		NW	Number of Estima...	6	.
Race		RASE	Root Average Su...	0.436082	0.434445
Race		RFPE	Root Final Predicti...	0.436738	.
Race		RMSE	Root Mean Squar...	0.43641	0.434445
Race		SBC	Schwarz's Bayesi...	4559.183	.
Race		SSE	Sum of Squared E...	1517.535	647.0106
Race		SUMW	Sum of Case Wei...	7980	3428
Race		MISC	Misclassification ...	0.27594	0.275963

Figure 114 - Fit Statistics for Regression Model (T_VS_Race)

According to the MISC (misclassification rate), the model has a 0.27594 MISC in the train data while the validation data has 0.275963. The ASE (average squared error) is 0.190167 for the train set and 0.188743, if both are rounded to the nearest decimal, then both are 0.19. The RMSE(root mean squared error) value can also be seen to be 0.43641 for the train data while the validation data is 0.434445. The lower the values of all three evaluation metrics, the better. According to the validation data, 27% of the data was misclassified but had an Accuracy of 73%. RMSE is also low and falls on the lower end of the 0 – 1 scale.

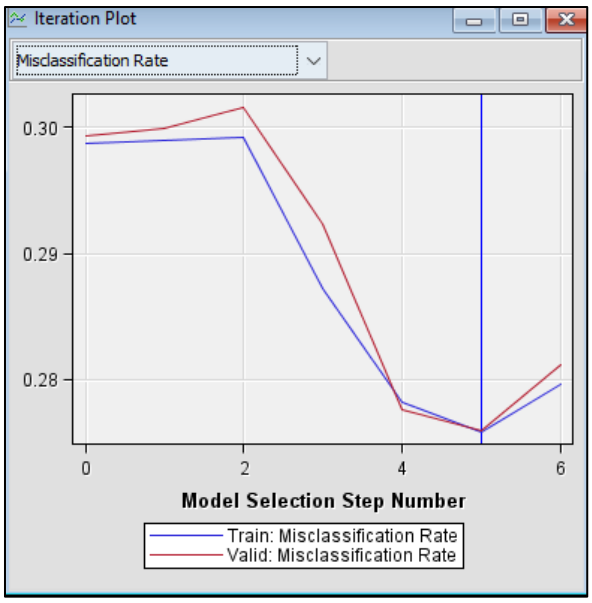


Figure 115 - Misclassification Rate Iteration Plot

The iteration plot for the misclassification rate shows that the smallest value for MISC occurs at step 5 rather than at step 6 where the model ends. By optimizing the model with the appropriate setting, this plot was achieved. This means that at step 5, the predictions will be as accurate as predictions at step 6. Moreover, it is stated that having lower iteration plots is better in terms of performance, partitioning the data and carrying out all the pre-processing steps aids the model in terms of producing an efficient output.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	IMP_Age	1	1	141.1918		<.0001	0.2999
2	TI_Region5	1	2	52.9139		<.0001	0.3016
3	TI_Region4	1	3	68.2312		<.0001	0.2923
4	REP_Mental_Illness	1	4	38.6832		<.0001	0.2777
5	Body_Camera	1	5	40.6966		<.0001	0.2760
6	TI_REP_Armed_Category6	1	6	6.0478		0.0139	0.2812

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 5. It consists of the following effects:

Intercept Body_Camera IMP_Age REP_Mental_Illness TI_Region4 TI_Region5

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
4866.194	4509.434	356.7603	5	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Body_Camera	1	39.9216	<.0001
IMP_Age	1	130.9479	<.0001
REP_Mental_Illness	1	41.8849	<.0001
TI_Region4	1	70.7277	<.0001
TI_Region5	1	91.8737	<.0001

Figure 116 - Summary of Stepwise

The summary of the stepwise selection shows that IMP_Age, T1_Region5, T1_Region4, REP_Mental_Illness, Body_Camera were selected as the effects with optimal complexity. _REP_Armed_Category6 was not included as it a low probability and score. Moreover, it also shows that after step 6, the addition or removal of any input would not change the fit statistic, hence instead a summary was produced. It also shows at what step each variable entered into the model.

The Type 3 analysis of effects displays the parameters in the regression model. As variable selection node has already been used, all the inputs have a Pr > ChiSq values closer to the 0 and have the most effect on the race of the victim.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-0.00592	0.1285	0.00	0.9633		0.994
Body_Camera	0	1	-0.3191	0.0505	39.92	<.0001		0.727
IMP_Age		1	-0.0362	0.00316	130.95	<.0001	-0.2539	0.964
REP_Mental_Illness	No	1	0.3012	0.0465	41.88	<.0001		1.352
TI_Region4	0	1	0.4390	0.0522	70.73	<.0001		1.551
TI_Region5	0	1	0.4197	0.0438	91.87	<.0001		1.522

Figure 117 - Analysis of Maximum Likelihood Estimates

The analysis of maximum likelihood estimates displays the variables where 3 variables, REP_MENTAL_ILLNESS, T1_REGION4 and T1_REGION5 have a positive estimate or regression coefficient as compared to IMP_AGE and BODY_CAMERA. The intercept of the model shows a degree of freedom (df) of 1, this informs the researcher that the model believes

producing one equation for calculation of the odds of both levels is more than enough, but this is not an efficient method as it is better to have two equations for both levels in the target for a much more optimal performance hence why regression can be seen to not be the best model for this prediction. The odds ratio output signifies what factors can contribute to the Race of a person being either 1 or 0 when they are shot. Using this, the importance of that input can be judged.

Odds Ratio Estimates

Effect		Point Estimate
Body_Camera	0 vs 1	0.528
IMP_Age		0.964
REP_Mental_Illness	No vs Yes	1.827
TI_Region4	0 vs 1	2.406
TI_Region5	0 vs 1	2.315

Figure 118 - Odds Ratio Estimates

- For IMP_AGE, the odds ratio estimate equals 0.964. This means that for victims coming under the 1 category in RACE, the odds of victims being '1' in fatal police shootings change by a factor of 0.964, a 1% decrease.

- For BODY_CAMERA, the odds ratio is 0.528, which means that the victims are 0.528 times less likely to be '1' in cases where BODY_CAMERA is 0 than odds in cases where it is 1.

- For REP_MENTAL_ILLNESS, the odds ratio is 1.827 which means in cases where the victim has 'No' case of mental illness, odds of their RACE being 1 are 1.827 higher than in cases of 'Yes' for REP_MENTAL_ILLNESS

- For both TI_REGION4 and TI_REGION5, the odds ratio (0 vs 1) estimate is 2.406 and 2.315, respectively. This means for cases with 0 for each, the odds of RACE of victim being 1 are 2.406 and 2.315 higher than the odds of their RACE being 1 for cases with a value of 1.

VS_Race_Regression Model

Fit Statistics

Target	Target Label	Fit Statistics ▲	Statistics Label	Train	Validation
Race		AIC	Akaike's Informati...	4520.398	.
Race		ASE	Average Squared ...	0.189926	0.189023
Race		AVERR	Average Error Fun...	0.564461	0.560958
Race		DFE	Degrees of Freed...	3982	.
Race		DFM	Model Degrees of ...	8	.
Race		DFT	Total Degrees of ...	3990	.
Race		DIV	Divisor for ASE	7980	3428
Race		ERR	Error Function	4504.398	1922.963
Race		FPE	Final Prediction Er...	0.190689	.
Race		MAX	Maximum Absolut...	0.959883	0.956452
Race		MISC	Misclassification ...	0.274436	0.278296
Race		MSE	Mean Square Error	0.190307	0.189023
Race		NOBS	Sum of Frequencies	3990	1714
Race		NW	Number of Estima...	8	.
Race		RASE	Root Average Su...	0.435805	0.434768
Race		RFPE	Root Final Predicti...	0.436679	.
Race		RMSE	Root Mean Squar...	0.436242	0.434768
Race		SBC	Schwarz's Bayesi...	4570.73	.
Race		SSE	Sum of Squared E...	1515.606	647.9718
Race		SUMW	Sum of Case Wei...	7980	3428

Figure 119 - Fit Statistics for Regression Model (VS_Race_Reg)

This regression model was built using the untransformed variables through a variable selection node (See Section 5.5.1.3). As compared to the first regression model built with the transformed variables, the value of the train data's MISC had improved with 0.274436 however saw a slight increase in the validation data's MISC with 0.278296. As compared to the T_VS_Race model, the difference between the train and validation data's MISC was a little higher in the VS_Race model as compared to the former's model.

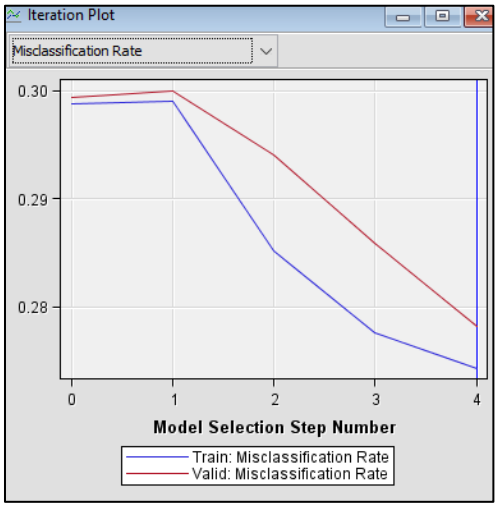


Figure 120 - Misclassification Rate Iteration Plot

The iteration plot for the misclassification rate shows that the smallest value for MISC occurs at step 4 and that is where the model ends. As mentioned above, there is a bit of difference between the train and validation data's MISC however it is not over or underfitting.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	IMP_Age	1	1	141.1918		<.0001	0.2999
2	Region	4	2	128.9300		<.0001	0.2940
3	REP_Mental_Illness	1	3	39.0250		<.0001	0.2859
4	Body_Camera	1	4	41.4965		<.0001	0.2783

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 4. It consists of the following effects:

Intercept Body_Camera IMP_Age REP_Mental_Illness Region

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
4866.194	4504.398	361.7964	7	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Body_Camera	1	40.6912	<.0001
IMP_Age	1	132.7486	<.0001
REP_Mental_Illness	1	42.2807	<.0001
Region	4	135.8104	<.0001

Figure 121 - Summary of Step-Wise Selection and Type 3 Analysis Effects

This regression model reached its optimal state with the variables at step 4 which were IMP_Age, REGION, REP_MENTAL_ILLNESS and BODY_CAMERA, the summary also shows what step each variable had entered the model.

The Global Null Hypothesis's Likelihood Ratio Chi-Square was at 361.7. This improved from the previous models 356.9 which shows that this model produced more significant results. The Type 3 analysis effects show the Wald Chi-Square values of the inputs. All inputs in this model have higher values than the inputs in the first regression model.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	0.5438	0.1253	18.84	<.0001		1.723
Body_Camera	0	1	-0.3223	0.0505	40.69	<.0001		0.724
IMP_Age		1	-0.0366	0.00318	132.75	<.0001	-0.2567	0.964
REP_Mental_Illness	No	1	0.3031	0.0466	42.28	<.0001		1.354
Region	Midwest	1	0.1942	0.0767	6.41	0.0113		1.214
Region	Northeast	1	0.5265	0.1080	23.78	<.0001		1.693
Region	Southeast	1	0.3508	0.0634	30.59	<.0001		1.420
Region	Southwest	1	-0.5550	0.0830	44.74	<.0001		0.574

Figure 122 - Analysis of Maximum Likelihood

This analysis of maximum likelihood estimates displays all the inputs. One of the inputs which are REGION's Midwest has the lowest chi-square amongst all the other inputs but still meets the significance level of <0.05, otherwise, all the inputs that have entered into the model have a significant effect on the performance of the model for predicting race. The intercept generated for the race target variable is only for both levels. It would have been much better to have 2 intercepts for assessment of the levels as mentioned above since it would bring a better interpretation to the model and hence this is one of the reasons why the logistic regression model is not optimal.

Odds Ratio Estimates

Effect		Point Estimate
Body_Camera	0 vs 1	0.525
IMP_Age		0.964
REP_Mental_Illness	No vs Yes	1.834
Region	Midwest vs West	2.036
Region	Northeast vs West	2.838
Region	Southeast vs West	2.381
Region	Southwest vs West	0.962

Figure 123 - Odds Ratio Estimate

- For IMP_Age, the odds ratio estimate equals 0.964. This means that for victims coming under the 1 category in RACE, the odds of victims being '1' in fatal police shootings change by a factor of 0.964, a 1% decrease in IMP_Age.

- For BODY_CAMERA, the odds ratio is 0.525, this means that the odds of the victim being of RACE 1 are 0.525 times lower in cases where BODY_CAMERA is 0 than odds in cases where BODY_CAMERA is 1.

- For REP_MENTAL_ILLNESS, the odds ratio is 1.834 which means in cases where the victim has 'No' case of mental illness, odds of their RACE being 1 are 1.834 higher than in cases of 'Yes' for REP_MENTAL_ILLNESS

- For REGION, the odds ratio estimate is 0.962, which means that for victims who come under 1 category in RACE, the odds of the victim being '1' in fatal police shootings decreases by 1% in Southwest compared to West Region.

- REGION of Midwest (2.036), Northeast (2.838) and Southeast (2.381) against West have a higher chance of predicting race of victim as RACE 1 than 0.

5.5.2.2 Decision Tree

Model Settings

⊟ Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	Gini
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
⊞ Node	
⊞ Split Search	
⊟ Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25

Figure 124 - Decision Tree Property Settings

When a target is a binary variable, its criterion is changed under the 'Nominal Target Criterion' tab. The model was implemented using Prob Chi-Sq and Entropy however the best performing model out of them was shown to be a model with the Gini criterion. The pruning method under the Sub-Tree section was kept as Assessment as it helps the model choose the best tree from the sequence and the Assessment Measure is kept at misclassification. Moreover, as the target is binary, MISC or Misclassification Rate is the best option for assessing our model based on making the best decisions.

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
Race	_NOBS_	Sum of Frequencies	3990	1714
Race	_MISC_	Misclassification Rate	0.267419	0.271295
Race	_MAX_	Maximum Absolute Error	0.806174	0.806174
Race	_SSE_	Sum of Squared Errors	1511.054	661.0647
Race	_ASE_	Average Squared Error	0.189355	0.192843
Race	_RASE_	Root Average Squared Error	0.43515	0.439139
Race	_DIV_	Divisor for ASE	7980	3428
Race	_DFT_	Total Degrees of Freedom	3990	.

Figure 125 - Fit Statistics for Decision Tree (VS_Race_DT)

The fit statistics show the train and validation data's value for MISC with 0.267419 and 0.271295, respectively. The ASE is also relatively good with 0.189355 and 0.1928843, respectively. It has been already mentioned above that Decision Tree was the best performing model out of the three models built with the RACE target variable. It is noted that according to MISC 73% of the data was classified correctly as compared to the 27% misclassification.

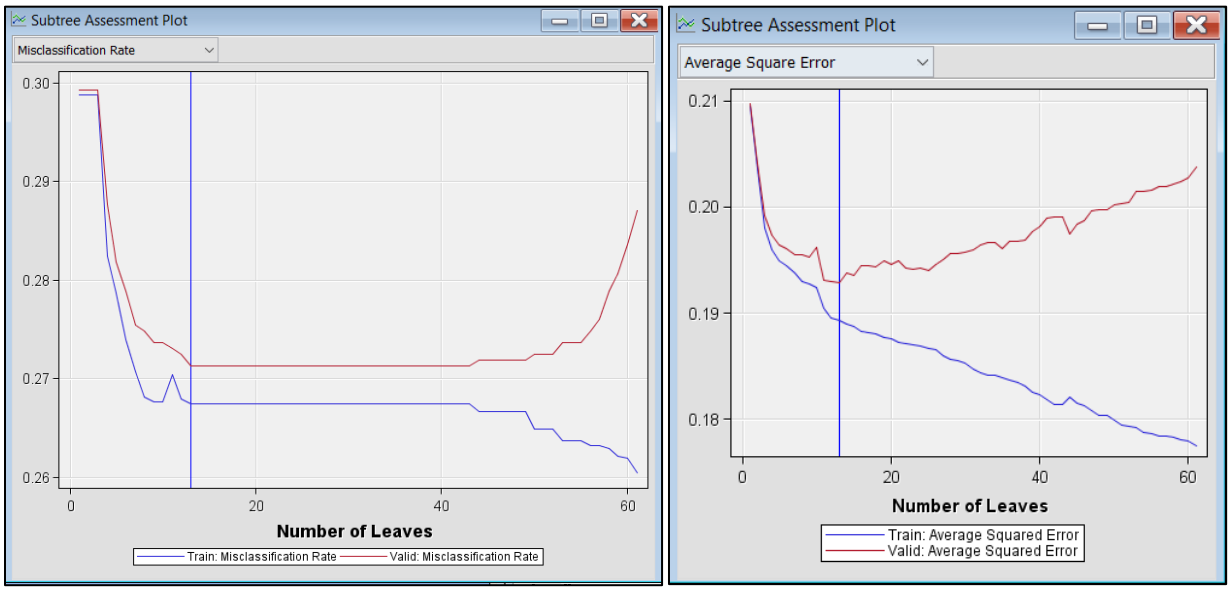


Figure 126 - Misclassification Rate and Average Square Error Sub Tree Assessment Plot

A maximal tree is the one that makes a tree with as many nodes as possible, for this model it is 62 nodes. As pruning and criterion have been selected to model the tree, all 62 nodes do not show up in the final model. It can be seen that the tree stops at the optimal level which is 13 leaves for the misclassification rate iteration plot. Just for comparisons, this can be compared with the average square error assessment plot which shows that the optimal tree for ASE was also at 13 leaves. The

lower the MISC or ASE, the better the performance of the model, however, this does not mean they will produce the same trees. Moreover, after step 13, the tree's accuracy starts to decrease.

Variable Importance

Variable Name	Number of Splitting Rules	Importance ▼	Validation Importance	Ratio of Validation to Training Importance	Label
IMP_Age	3	1.0000	1.0000	1.0000	Imputed A...
Region	2	0.9391	0.8412	0.8957	
REP_Mental_Illness	1	0.4979	0.4792	0.9625	Replacem...
Month	3	0.4011	0.3017	0.7522	
REP_Armed_Category	1	0.3348	0.1030	0.3076	Replacem...
Body_Camera	2	0.3087	0.3637	1.1782	

Figure 127 - Variable Importance

The variable importance chart shows us the important inputs in the decision trees. The value of the important statistic is proportional to the amount of variability in the target represented by the related input in comparison to the input at the top which is IMP_Age. It can be seen that REGION explains a total of 94% of variability explained by the IMP_Age, followed by REP_MENTAL_ILLNESS at 49.7% and so on.

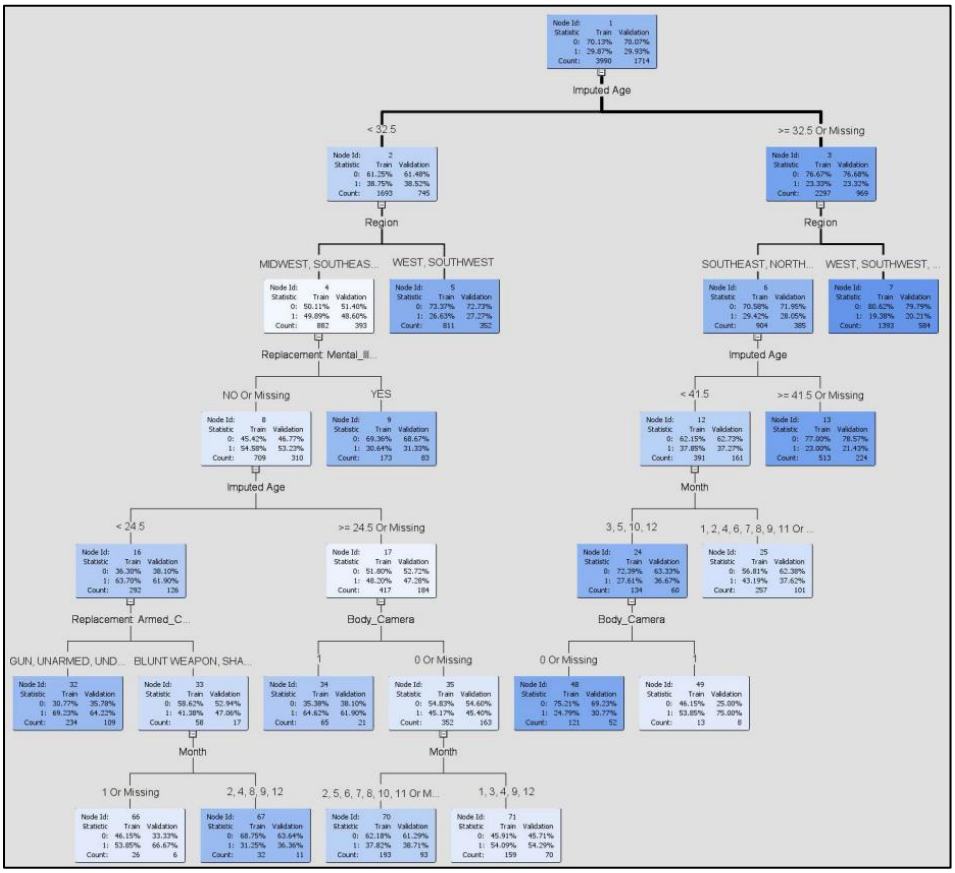


Figure 128 – Generated Tree

This is showing a 13-leave tree optimized under the validation misclassification assessment measure. Each split of the data is based on the log worth to separate subregions of the input space with a high proportion of victims under RACE category 1 and 0. There are 13 node rules however only a few of them will be discussed based on the number of observations as well as their position in the tree as child nodes. To note, the decision tree displays 'missing' as to show how future missing values will show up, this does not mean that they are any missing values in the dataset. (SAS Support, 2014)

```
*------------------------------------------------------------*
 Node = 5
*------------------------------------------------------------*
if Region IS ONE OF: WEST, SOUTHWEST
AND Imputed Age < 32.5
then
 Tree Node Identifier   = 5
 Number of Observations = 811
 Predicted: Race=1 = 0.27
 Predicted: Race=0 = 0.73
```

Figure 129 - Node Rule 5

This node rule has 811 observations and 2 inputs; *REGION* being WEST and SOUTHWEST and *IMPUTED AGE* being less than 32.5 that results in the victims predicted RACE being 0.73 times more likely to be 0 while 0.27 are predicted to be 1.

```
*------------------------------------------------------------*
 Node = 7
*------------------------------------------------------------*
if Region IS ONE OF: WEST, SOUTHWEST, MIDWEST or MISSING
AND Imputed Age >= 32.5 or MISSING
then
 Tree Node Identifier   = 7
 Number of Observations = 1393
 Predicted: Race=1 = 0.19
 Predicted: Race=0 = 0.81
```

Figure 130 - Node Rule 7

This node rule has 1393 observations and 2 inputs; *REGION* being WEST, SOUTHWEST, MIDWEST and *IMPUTED AGE* being greater than 32.5 and those results in the victims predicted RACE to be 0.81 times more likely to be 0 while 0.19 are predicted to be 1.

```
*-------------------------------------------------------------*
 Node = 13
*-------------------------------------------------------------*
if Region IS ONE OF: SOUTHEAST, NORTHEAST
AND Imputed Age >= 41.5 or MISSING
then
 Tree Node Identifier   = 13
 Number of Observations = 513
 Predicted: Race=1 = 0.23
 Predicted: Race=0 = 0.77
```

Figure 131 - Node Rule 13

This node rule has 513 observations and 2 inputs; *REGION* being SOUTHEAST and NORTHEAST and *IMPUTED AGE* being greater or equal to 41.5 that results in the victims predicted RACE to be 0.77 times more likely to be 0 while 0.23 are predicted to be 1.

```
*-------------------------------------------------------------*
 Node = 32
*-------------------------------------------------------------*
if Replacement: Mental_Illness IS ONE OF: NO or MISSING
AND Replacement: Armed_Category IS ONE OF: GUN, UNARMED, UNDETERMINED
AND Region IS ONE OF: MIDWEST, SOUTHEAST, NORTHEAST or MISSING
AND Imputed Age < 24.5
then
 Tree Node Identifier   = 32
 Number of Observations = 234
 Predicted: Race=1 = 0.69
 Predicted: Race=0 = 0.31
```

Figure 132 - Node Rule 32

This node rule has 234 observations and 4 inputs; *REPLACEMENT: MENTAL_ILLNESS* is said to be NO, *REPLACEMENT: ARMED_CATEGORY* is said to be GUN, UNARMED or UNDETERMINED, *REGION* is said to MIDWEST, SOUTHEAST, NORTHEAST or MISSING with an *IMPUTED AGE* being greater than 24.5, this predicts that in this case, the predicted RACE of the victim is 0.69 more likely to be 1 and 0.31 if they are 0.

```
*-------------------------------------------------------------*
 Node = 25
*-------------------------------------------------------------*
if Region IS ONE OF: SOUTHEAST, NORTHEAST
AND Month IS ONE OF: 1, 2, 4, 6, 7, 8, 9, 11 or MISSING
AND Imputed Age < 41.5 AND Imputed Age >= 32.5
then
 Tree Node Identifier   = 25
 Number of Observations = 257
 Predicted: Race=1 = 0.43
 Predicted: Race=0 = 0.57
```

Figure 133 - Node Rule 25

This node rule has 257 observations and 3 inputs; *REGION* with SOUTHEAST and NORTHEAST next is *MONTH* is one of 1,2,4,6,7,8,9 and 11 and *IMPUTED_AGE* are inclusive between 32.5 and 41.5 with the predicted RACE of the victim being 0.57 more likely to 0 and 0.43 if they are 1.

```
*-------------------------------------------------------------*
 Node = 71
*-------------------------------------------------------------*
if Replacement: Mental_Illness IS ONE OF: NO or MISSING
AND Region IS ONE OF: MIDWEST, SOUTHEAST, NORTHEAST or MISSING
AND Month IS ONE OF: 1, 3, 4, 9, 12
AND Imputed Age < 32.5 AND Imputed Age >= 24.5 or MISSING
AND Body_Camera IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 71
 Number of Observations = 159
 Predicted: Race=1 = 0.54
 Predicted: Race=0 = 0.46
```

Figure 134 - Node Rule 71

This node consists of 5 inputs and 159 observations: *REPLACEMENT MENTAL ILLNESS* being NO, *REGION* being MIDWEST, SOUTHEAST, NORTHEAST, *MONTH* being 1,3,4,9 and 12, *IMPUTED AGE* being greater than or equal to 24.5 but less than 32.5 and lastly, *BODY_CAMERA* being 0, this result predicted RACE of the victim to be 0.54 more likely to 1 and 0.46 for 0.

5.5.3 Influence of Mental Illness on Fatal Police Shootings

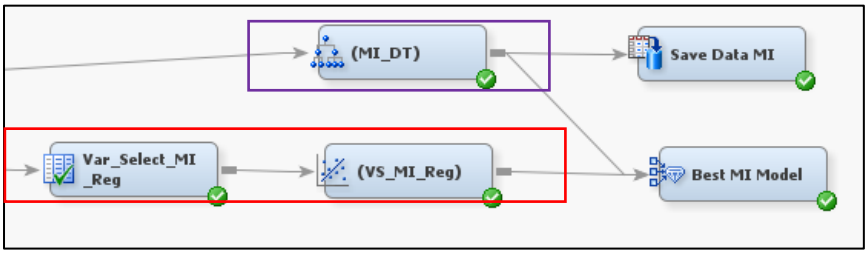


Figure 135 - Prediction Models for SIGNS_OF_MENTAL_ILLNESS

The target variable SIGNS_OF_MENTAL_ILLNESS has two values 1 or 0 and two models; Logistic Regression and Decision Tree will be built for its prediction analysis. After evaluation, a decision tree was selected as the best model.

5.5.3.1 Logistic Regression

Model Settings

As mentioned, for binary categorical variables, logistic regression is the best model to carry out prediction modelling. The settings for the regression model were kept the same as the model for RACE with the 'Stepwise' model and 'Validation Misclassification' selection criterion. Again, no other properties were changed in the model as they did not provide any changes to the output of the model. The model is shown in the diagram above in the red box (Fig. 135) and was built using a variable selection node. The model was first built by transforming the class inputs however it was not optimal in terms of the difference between its validation and train data's scores hence we decided no variable transformation for the regression model will happen.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation ▼
Signs Of Mental ...		DIV	Divisor for ASE	7980	3428
Signs Of Mental ...		SUMW	Sum of Case Wei...	7980	3428
Signs Of Mental ...		NOBS	Sum of Frequencies	3990	1714
Signs Of Mental ...		ERR	Error Function	3932.453	1677.5
Signs Of Mental ...		SSE	Sum of Squared E...	1285.926	549.9181
Signs Of Mental ...		MAX	Maximum Absolut...	0.972447	0.957568
Signs Of Mental ...		AVERR	Average Error Fun...	0.492789	0.489352
Signs Of Mental ...		RASE	Root Average Su...	0.401427	0.400524
Signs Of Mental ...		RMSE	Root Mean Squar...	0.402285	0.400524
Signs Of Mental ...		MISC	Misclassification ...	0.226817	0.231039
Signs Of Mental ...		ASE	Average Squared ...	0.161144	0.16042
Signs Of Mental ...		MSE	Mean Square Error	0.161833	0.16042
Signs Of Mental ...		AIC	Akaike's Informati...	3966.453	.
Signs Of Mental ...		DFE	Degrees of Freed...	3973	.
Signs Of Mental ...		DFM	Model Degrees of ...	17	.
Signs Of Mental ...		DFT	Total Degrees of ...	3990	.
Signs Of Mental ...		FPE	Final Prediction Er...	0.162523	.
Signs Of Mental ...		NW	Number of Estima...	17	.
Signs Of Mental ...		RFPE	Root Final Predicti...	0.403141	.
Signs Of Mental ...		SBC	Schwarz's Bayesi...	4073.409	.

Figure 136 - MISC and ASE for Regression Model

The value of MISC for this regression model is 0.226817 for the train data while the validation data's MISC is 0.231039. The ASE is 0.161144 and 0.16042 for the train and validation data, respectively. The RMSE value may also be looked at showing 0.402285 and 0.400524 as well.

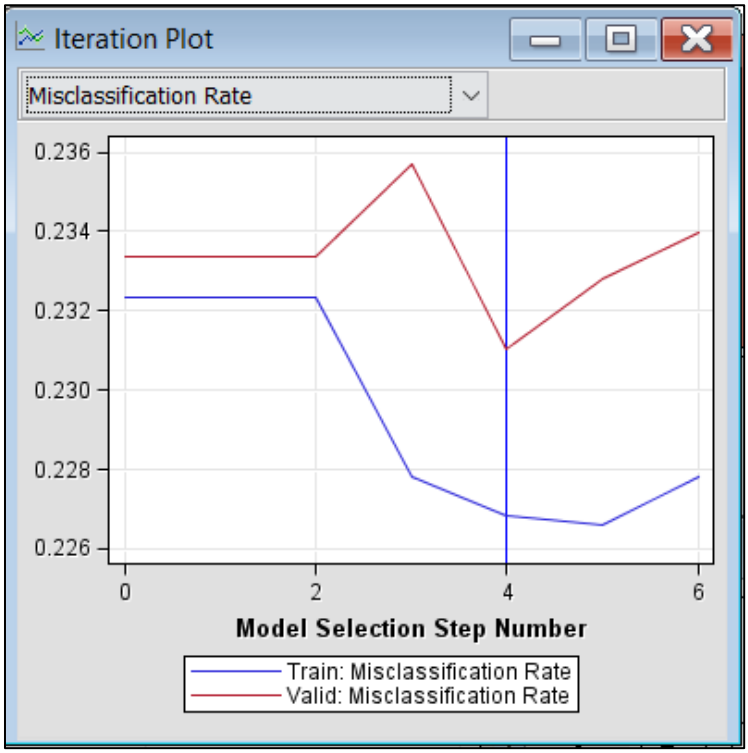


Figure 137 - Misclassification Rate Iteration Plot

The iteration plot for the misclassification rate shows that the smallest value for MISC occurs at step 4 rather than at step 6 where the model ends. By optimizing this model with the appropriate setting, we were able to achieve this plot. This means that at step 4, the predictions will be as accurate as predictions at step 6.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	REP_Flee	3	1	171.0607		<.0001	0.2334
2	REP_Armed_Category	7	2	110.3664		<.0001	0.2334
3	Race_Label	5	3	86.3074		<.0001	0.2357
4	Body_Camera	1	4	11.7025		0.0006	0.2310
5	Gender	1	5	6.0759		0.0137	0.2328
6	IMP_Age	1	6	4.5001		0.0339	0.2340

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 4. It consists of the following effects:

Intercept Body_Camera REP_Armed_Category REP_Flee Race_Label

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
4325.778	3932.453	393.3250	16	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Body_Camera	1	11.6178	0.0007
REP_Armed_Category	7	100.8334	<.0001
REP_Flee	3	92.3389	<.0001
Race_Label	5	88.9654	<.0001

Figure 138 - Summary of Stepwise and Type 3 Analysis of Effects

The summary of the stepwise selection shows that REP_FLEE, REP_ARMED_CATEGORY , RACE_LABEL and BODY_CAMERA were selected as the effects with optimal complexity. Moreover, it also shows that after step 6, the addition or removal of any input would not change the fit statistic, hence instead a summary was produced. The above inputs entered into the model at steps 1, 2,3 and 4, respectively. The Type 3 analysis of effects displays the effects that were entered into the regression model and their Wald Chi-Square as well as the probability. BODY_CAMERA had the lowest probability yet had met the significance level of <0.05, hence was entered into the model with a 0.0007 probability.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-1.7864	0.1438	154.23	<.0001		0.168
Body_Camera	0	1	-0.1896	0.0556	11.62	0.0007		0.827
REP_Armed_Category	Blunt Weapon	1	0.4952	0.2135	5.38	0.0204		1.641
REP_Armed_Category	Gun	1	-0.0567	0.0980	0.33	0.5632		0.945
REP_Armed_Category	Other Weapon	1	-0.3141	0.3568	0.77	0.3787		0.730
REP_Armed_Category	Sharp Weapon	1	0.6756	0.1112	36.94	<.0001		1.965
REP_Armed_Category	Toy Weapon	1	0.8830	0.1724	26.22	<.0001		2.418
REP_Armed_Category	Unarmed	1	-0.0184	0.1626	0.01	0.9099		0.982
REP_Armed_Category	Undetermined	1	-0.6477	0.1998	10.51	0.0012		0.523
REP_Flee	Car	1	-0.3542	0.1146	9.55	0.0020		0.702
REP_Flee	Foot	1	-0.2845	0.1135	6.28	0.0122		0.752
REP_Flee	Not fleeing	1	0.6520	0.0697	87.48	<.0001		1.919
Race_Label	Asian	1	0.1003	0.2399	0.17	0.6759		1.105
Race_Label	Black	1	-0.2165	0.1311	2.73	0.0986		0.805
Race_Label	Hispanic	1	-0.2082	0.1374	2.30	0.1296		0.812
Race_Label	Native	1	-0.1356	0.3013	0.20	0.6528		0.873
Race_Label	Other	1	-0.1150	0.3960	0.08	0.7714		0.891

Figure 139 - Analysis of Maximum Likelihood Estimate

As it can be seen, all the level of the class variables that were entered into the model have their estimate, error, wald chi-square, and Pr> ChiSq values displayed. Although all levels are showing their specific values, it does not mean each of them is important to the model. REP_ARMED_CATEGORY (Blunt Weapon, Sharp Weapon, Toy Weapon), REP_FLEE (Car, Foot, Not Fleeing) are the most important inputs that will be influencing the prediction of mental illness of a victim. Again, like the previous regression models, there is only one intercept generated, it would have been better to have 2 intercepts to assess each level of mental illness for a better and more efficient interpretation of the model.

Odds Ratio Estimates		
Effect		Point Estimate
Body_Camera	0 vs 1	0.684
REP_Armed_Category	Blunt Weapon vs Vehicle	4.536
REP_Armed_Category	Gun vs Vehicle	2.612
REP_Armed_Category	Other Weapon vs Vehicle	2.019
REP_Armed_Category	Sharp Weapon vs Vehicle	5.433
REP_Armed_Category	Toy Weapon vs Vehicle	6.685
REP_Armed_Category	Unarmed vs Vehicle	2.714
REP_Armed_Category	Undetermined vs Vehicle	1.447
REP_Flee	Car vs Undetermined	0.711
REP_Flee	Foot vs Undetermined	0.762
REP_Flee	Not fleeing vs Undetermined	1.945
Race_Label	Asian vs White	0.622
Race_Label	Black vs White	0.453
Race_Label	Hispanic vs White	0.457
Race_Label	Native vs White	0.491
Race_Label	Other vs White	0.501

Figure 140 - Odds Ratio Estimate

The odds ratio explanation has been discussed in the literature review chapter. It is understood that decision trees were the best modelling technique for analysis on mental illness. The below inputs are based on the practicality of the regression model and only a few of them will be discussed regarding their odds ratio.

- For BODY_CAMERA, the odds ratio is 0.684, this means that victims are 0.684 less likely to be mentally ill if there is no BODY_CAMERA on the police officer.
- The Odds ratio for REP_ARMED_CATEGORY (Blunt Weapon Vs Vehicle) is 4.536 which means that the odds of a victim being mentally ill are 4.536 times higher when they had a blunt weapon in their possession than the odds of them being mentally ill when they were in a vehicle. This statement is similar to the Odds ratio for (Sharp Weapon vs Vehicle and Toy Weapon vs Vehicle with an odds ratio of 5.433 and 6.685, respectively.)
- The Odds ratio for REP_FLEE (Not Fleeing vs Undetermined) was as that in cases of not fleeing, the victim was 1.945 times likely to be mentally ill than in undetermined cases.
- The odd ratio for RACE_LABEL (Asian vs White) shows a 0.622 decrease which means Asian victims are less likely to be mentally ill in cases of shooting as compared to White victims.

5.5.3.2 Decision Tree

Model Settings

The model setting for decision tree for SIGNS_OF_MENTAL_ILLNESS was kept the same as for modelling with RACE with target criterion being Gini and assessment method and measure being 'Assessment' and 'Misclassification Rate', correspondingly as it was also a binary categorical variable. Like the previous decision tree for the RACE target, this model was also tried and tested with different criteria like Entropy and Prob Chi-Square but did not produce as good results as Gini.

Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation
Signs Of Mental Illness	NOBS	Sum of Frequencies	3990	1714
Signs Of Mental Illness	MISC	Misclassification Rate	0.221554	0.229288
Signs Of Mental Illness	MAX	Maximum Absolute Error	0.895522	0.895522
Signs Of Mental Illness	SSE	Sum of Squared Errors	1268.083	551.1575
Signs Of Mental Illness	ASE	Average Squared Error	0.158908	0.160781
Signs Of Mental Illness	RASE	Root Average Squared Error	0.398632	0.400975
Signs Of Mental Illness	DIV	Divisor for ASE	7980	3428
Signs Of Mental Illness	DFT	Total Degrees of Freedom	3990	.

Figure 141 - Fit Statistics for SIGNS_OF_MENTAL_ILLNESS

The fit statistics show the train and validation data's value for MISC being 0.221554 and 0.229288, respectively. Although it is not the main assessment metric, ASE is also good with 0.1589908 and 0.160781, respectively. Overall, the decision tree model is approximately 77% accurate.

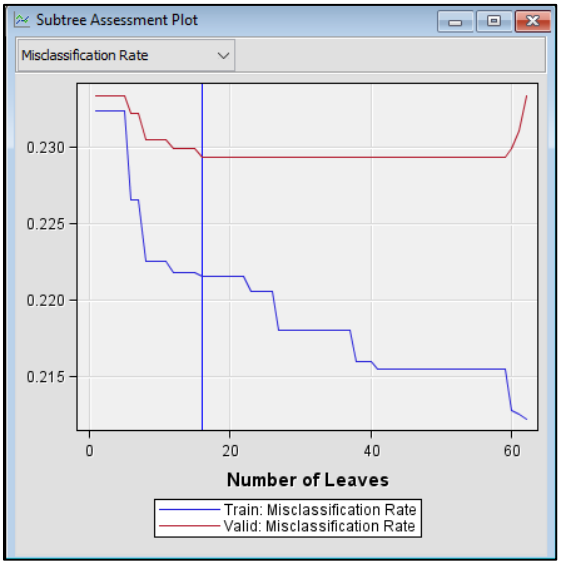


Figure 142 - Subtree Assessment Plot for Misclassification Rate

The maximal tree seems to also end at 62 nodes. If the assessment method and measure was not selected, then the tree would show all 62 nodes. The optimal leaves for this tree can be seen at

node 16 after being pruned, after this point, the model is prone to overfitting and less accurate as the nodes increase.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance ▼
Threat_Level		1	0.2326	0.2678	1.1512
REP_Flee	Replacement: Flee	1	1.0000	1.0000	1.0000
Body_Camera		2	0.2514	0.2380	0.9465
Region		1	0.1798	0.1632	0.9080
REP_Armed_Category	Replacement: Arme...	4	0.8338	0.7319	0.8777
Race_Label		2	0.7219	0.5021	0.6955
Month		3	0.4716	0.1743	0.3696
Day		1	0.2009	0.0000	0.0000
Year		0	0.0000	0.0000	.
IMP_Age	Imputed Age	0	0.0000	0.0000	.
MannerofDeath		0	0.0000	0.0000	.
Gender		0	0.0000	0.0000	.

Figure 143 - Variable Importance

These are the important inputs in the decision tree. The value of the important statistic is proportional to the amount of variability in the target represented by the related input in comparison to the input at the top which is REP_FLEE. REP_ARMED_CATEGORY explains about 83% of variability explained by REP_FLEE, followed by RACE_LABEL at 73% and so on. YEAR, IMP_AGE, MANNEROFDEATH and GENDER have no splits and are not relevant to the model at all.

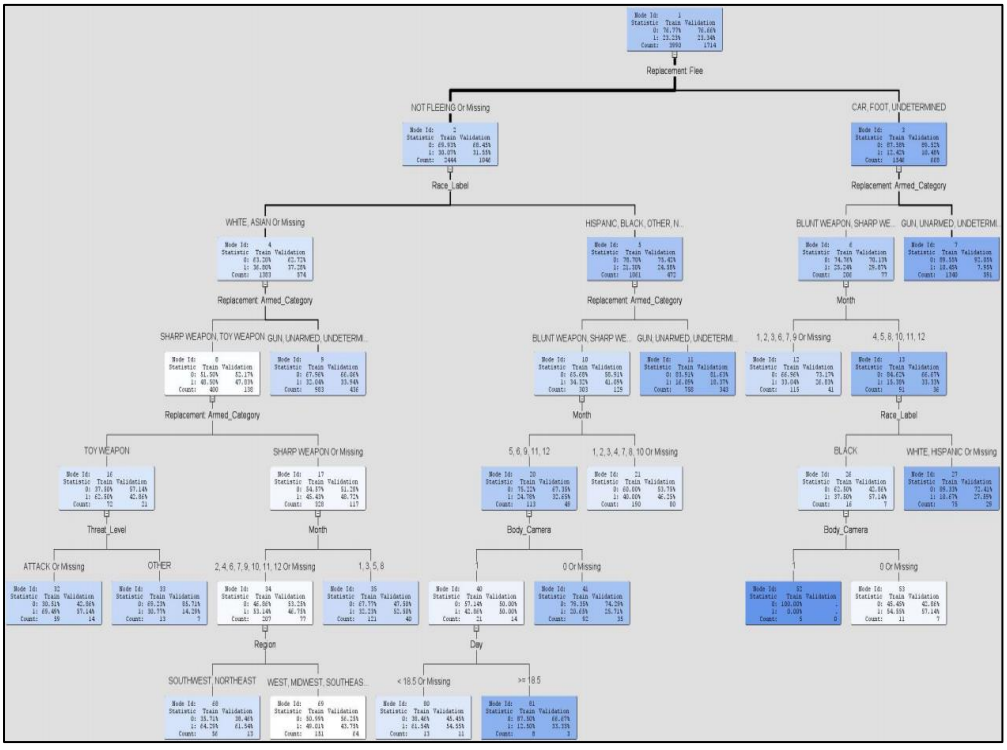


Figure 144 - Generated Tree

This is showing a 16-leaves tree optimized under the validation misclassification assessment measure. Each split is based on the log worth to separate subregions of the input space with a high proportion of victims under SIGNS_OF_MENTAL_ILLNESS.

Although there are a total of 16 leaves, only the top 6 nodes with the highest observations will be discussed. The meaning of why missing appears has been discussed in Section 5.5.2.2.

```
*------------------------------------------------------------*
 Node = 7
*------------------------------------------------------------*
if Replacement: Flee IS ONE OF: CAR, FOOT, UNDETERMINED
AND Replacement: Armed_Category IS ONE OF: GUN, UNARMED, UNDETERMINED, OTHER WEAPON, VEHICLE or MISSING
then
 Tree Node Identifier   = 7
 Number of Observations = 1340
 Predicted: Signs_Of_Mental_Illness=1 = 0.10
 Predicted: Signs_Of_Mental_Illness=0 = 0.90
```

Figure 145 - Node Rule 7

This node rule has 1340 observations and 2 inputs; *REPLACEMENT: FLEE* is CAR, FOOT, and UNDETERMINED and *REPLACEMENT: ARMED_CATEGORY* is GUN, UNARMED, UNDETERMINED, OTHER WEAPON, VEHICLE and if these both occur then the victim is 90% likely to not be suffering from a mental illness and only 10% of victims may have a mental illness.

```
*------------------------------------------------------------*
 Node = 9
*------------------------------------------------------------*
if Replacement: Flee IS ONE OF: NOT FLEEING or MISSING
AND Replacement: Armed_Category IS ONE OF: GUN, UNARMED, UNDETERMINED, BLUNT WEAPON, OTHER WEAPON, VEHICLE or MISSING
AND Race_Label IS ONE OF: WHITE, ASIAN or MISSING
then
 Tree Node Identifier   = 9
 Number of Observations = 983
 Predicted: Signs_Of_Mental_Illness=1 = 0.32
 Predicted: Signs_Of_Mental_Illness=0 = 0.68
```

Figure 146 - Node Rule 9

This node rule has 983 observations and has 3 inputs; if *REPLACEMENT: FLEE* is NOT FLEEING and *REPLACEMENT: ARMED_CATEGORY* is GUN, UNARMED, UNDETERMINED, BLUNT WEAPON, OTHER WEAPON, VEHICLE and *RACE_LABEL* is WHITE or ASIAN, then the predicted probability of the victim not having a mental illness is 0.68 as compared to having signs of mental illness at 0.32.

```
*-------------------------------------------------------------*
 Node = 11
*-------------------------------------------------------------*
if Replacement: Flee IS ONE OF: NOT FLEEING or MISSING
AND Replacement: Armed_Category IS ONE OF: GUN, UNARMED, UNDETERMINED, TOY WEAPON, OTHER WEAPON, VEHICLE or MISSING
AND Race_Label IS ONE OF: HISPANIC, BLACK, OTHER, NATIVE
then
 Tree Node Identifier   = 11
 Number of Observations = 758
 Predicted: Signs_Of_Mental_Illness=1 = 0.16
 Predicted: Signs_Of_Mental_Illness=0 = 0.84
```

Figure 147 - Node Rule 11

This node rule has 758 observations and 3 inputs starting; if *REPLACEMENT: FLEE* is NOT FLEEING and *REPLACEMENT: ARMED_CATEGORY* is GUN, UNARMED, UNDETERMINED, TOY WEAPON, OTHER WEAPON, VEHICLE and *RACE_LABEL* is HISPANIC, BLACK, OTHER or NATIVE, then predicted probability of the victim not having a mental illness is 0.84 as compared to having signs of mental illness at 0.16.

```
*-------------------------------------------------------------*
 Node = 21
*-------------------------------------------------------------*
if Replacement: Flee IS ONE OF: NOT FLEEING or MISSING
AND Replacement: Armed_Category IS ONE OF: BLUNT WEAPON, SHARP WEAPON
AND Race_Label IS ONE OF: HISPANIC, BLACK, OTHER, NATIVE
AND Month IS ONE OF: 1, 2, 3, 4, 7, 8, 10 or MISSING
then
 Tree Node Identifier   = 21
 Number of Observations = 190
 Predicted: Signs_Of_Mental_Illness=1 = 0.40
 Predicted: Signs_Of_Mental_Illness=0 = 0.60
```

Figure 148 - Node Rule 21

This node rule has 190 observations and 4 inputs; if *REPLACEMENT: FLEE* is NOT FLEEING and *REPLACEMENT: ARMED_CATEGORY* is SHARP WEAPON and BLUNT WEAPON and *RACE_LABEL* is HISPANIC, BLACK, OTHER or NATIVE and MONTH is one of 1,2,3,4,7,8 or 10 then the predicted probability of the victim not having a mental illness is 0.60 as compared to having signs of mental illness at 0.40.

```
*-------------------------------------------------------------*
 Node = 69
*-------------------------------------------------------------*
if Replacement: Flee IS ONE OF: NOT FLEEING or MISSING
AND Replacement: Armed_Category IS ONE OF: SHARP WEAPON or MISSING
AND Region IS ONE OF: WEST, MIDWEST, SOUTHEAST or MISSING
AND Race_Label IS ONE OF: WHITE, ASIAN or MISSING
AND Month IS ONE OF: 2, 4, 6, 7, 9, 10, 11, 12 or MISSING
then
 Tree Node Identifier   = 69
 Number of Observations = 151
 Predicted: Signs_Of_Mental_Illness=1 = 0.49
 Predicted: Signs_Of_Mental_Illness=0 = 0.51
```

Figure 149 - Node Rule 69

This node rule has 151 observations and 4 inputs; if *REPLACEMENT: FLEE* is NOT FLEEING and *REPLACEMENT: ARMED_CATEGORY* is SHARP WEAPON and *REGION* is one of WEST, MIDWEST, SOUTHEAST and *RACE_LABEL* is WHITE or ASIAN and *MONTH* is one of 2,4,6,7,9,10,11 or 12 then the predicted probability of the victim not having a mental illness is 0.51 as compared to having signs of mental illness at 0.49.

```
*-------------------------------------------------------------*
 Node = 12
*-------------------------------------------------------------*
if Replacement: Flee IS ONE OF: CAR, FOOT, UNDETERMINED
AND Replacement: Armed_Category IS ONE OF: BLUNT WEAPON, SHARP WEAPON, TOY WEAPON
AND Month IS ONE OF: 1, 2, 3, 6, 7, 9 or MISSING
then
 Tree Node Identifier   = 12
 Number of Observations = 115
 Predicted: Signs_Of_Mental_Illness=1 = 0.33
 Predicted: Signs_Of_Mental_Illness=0 = 0.67
```

Figure 150 - Node Rule 12

This node rule has 115 observations and 3 inputs; if *REPLACEMENT: FLEE* is CAR, FOOT or UNDETERMINED and *REPLACEMENT: ARMED_CATEGORY* is BLUNT WEAPON, SHARP WEAPON, TOY WEAPON and *MONTH* is one of 1,2,3,6,7 or 9 then the predicted probability of the victim not having a mental illness is 0.67 as compared to having signs of mental illness at 0.33.

5.6 Dashboards

In this subsection, descriptive modelling will be carried out with the cleaned dataset in PowerBI Desktop. The data has been exported after the impute node as all the missing values have been removed and taken care of. The data is loaded and transformed in PowerBI. It is important to make

sure that all the values are whole numbers, and all the class inputs are classified as text data. When this aspect is satisfactory, then the visualization process can begin.

Firstly, the data was saved from the Save Data node and the format was kept as CSV. The node was run and the file was exported to SAS Studio.

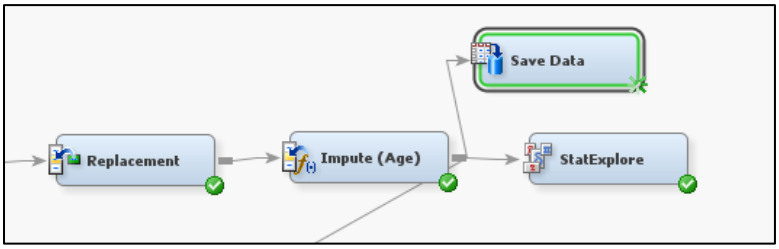


Figure 151 - Save Data Node

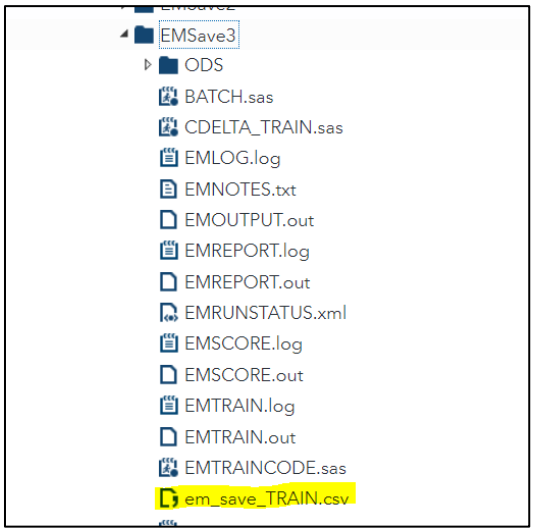


Figure 152 - Exported Dataset

The dataset is shown above in SAS Studio, from where it was then downloaded and imported into PowerBI.

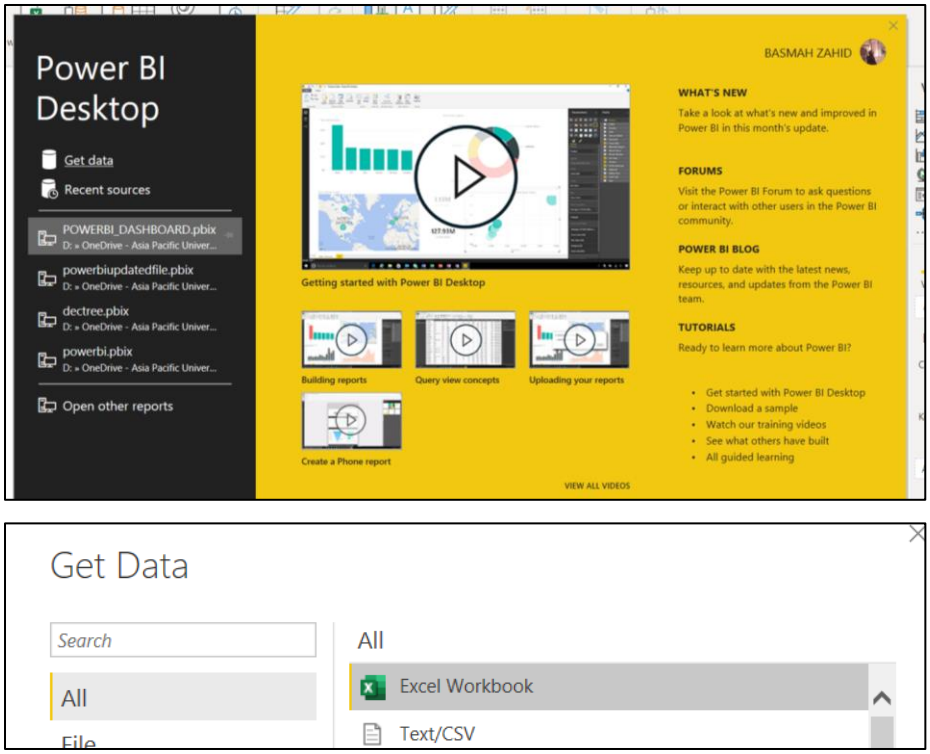


Figure 153 - PowerBI Screen

'Get Data' is selected and it asks for the file. The file can then be chosen from the place it was downloaded to and a display showing the dataset is shown to the user.

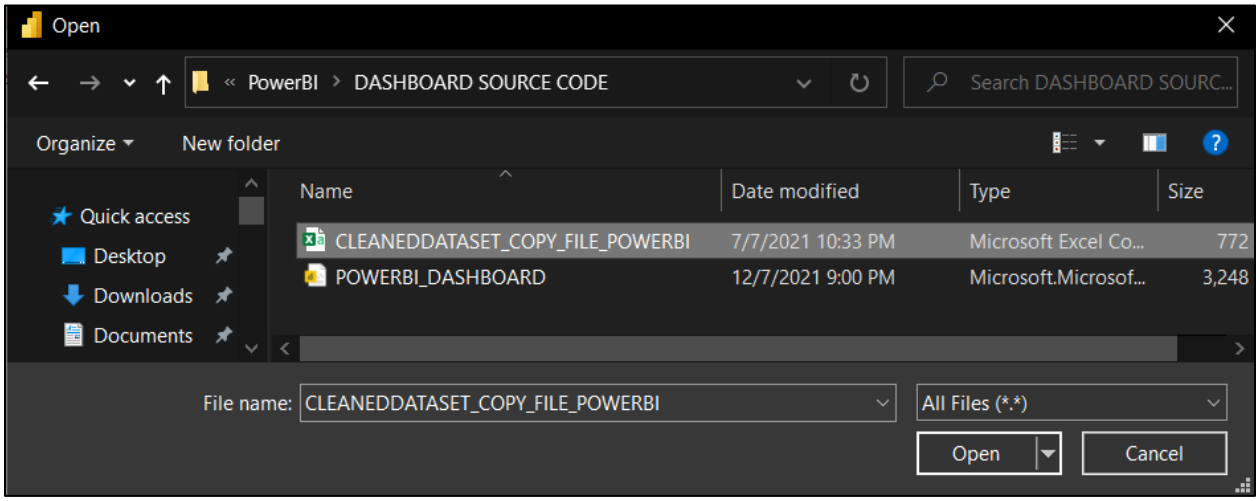
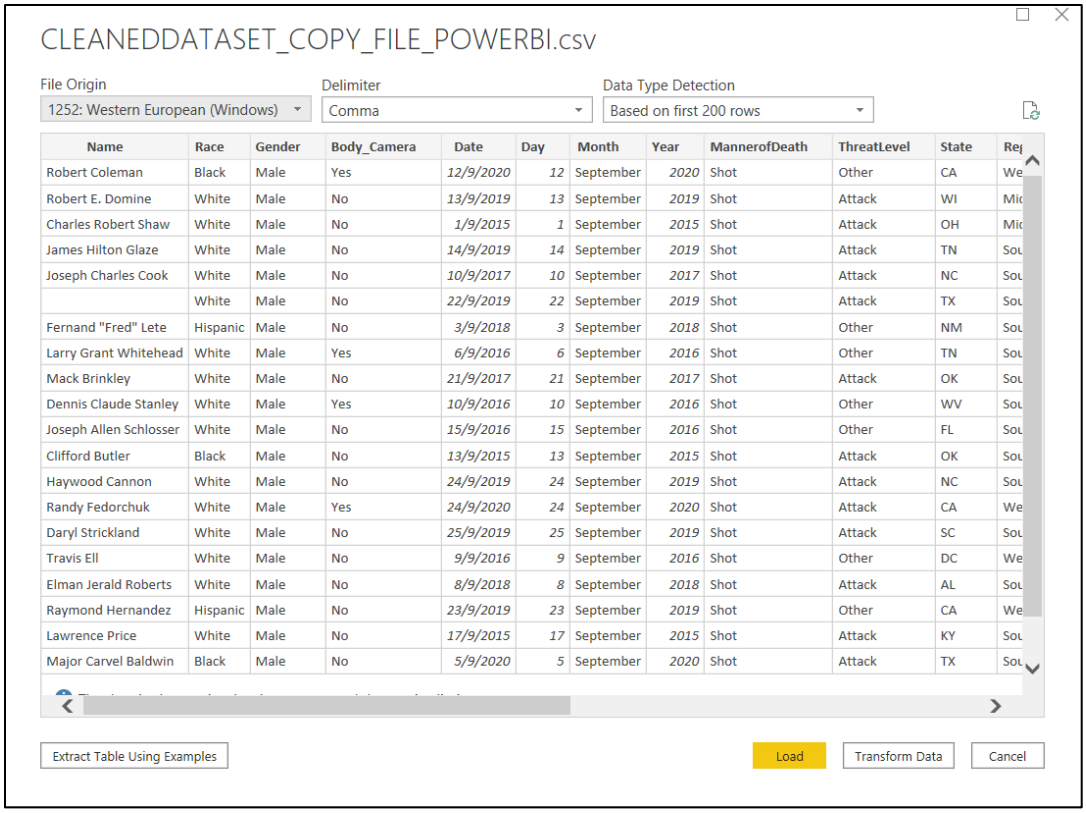


Figure 154 - Selecting File



Name	Race	Gender	Body_Camera	Date	Day	Month	Year	MannerofDeath	ThreatLevel	State	Re
Robert Coleman	Black	Male	Yes	12/9/2020	12	September	2020	Shot	Other	CA	We
Robert E. Domine	White	Male	No	13/9/2019	13	September	2019	Shot	Attack	WI	Mic
Charles Robert Shaw	White	Male	No	1/9/2015	1	September	2015	Shot	Attack	OH	Mic
James Hilton Glaze	White	Male	No	14/9/2019	14	September	2019	Shot	Attack	TN	Sou
Joseph Charles Cook	White	Male	No	10/9/2017	10	September	2017	Shot	Attack	NC	Sou
	White	Male	No	22/9/2019	22	September	2019	Shot	Attack	TX	Sou
Fernand "Fred" Lete	Hispanic	Male	No	3/9/2018	3	September	2018	Shot	Other	NM	Sou
Larry Grant Whitehead	White	Male	Yes	6/9/2016	6	September	2016	Shot	Other	TN	Sou
Mack Brinkley	White	Male	No	21/9/2017	21	September	2017	Shot	Attack	OK	Sou
Dennis Claude Stanley	White	Male	Yes	10/9/2016	10	September	2016	Shot	Other	WV	Sou
Joseph Allen Schlosser	White	Male	No	15/9/2016	15	September	2016	Shot	Other	FL	Sou
Clifford Butler	Black	Male	No	13/9/2015	13	September	2015	Shot	Attack	OK	Sou
Haywood Cannon	White	Male	No	24/9/2019	24	September	2019	Shot	Attack	NC	Sou
Randy Fedorchuk	White	Male	Yes	24/9/2020	24	September	2020	Shot	Attack	CA	We
Daryl Strickland	White	Male	No	25/9/2019	25	September	2019	Shot	Attack	SC	Sou
Travis Ell	White	Male	No	9/9/2016	9	September	2016	Shot	Other	DC	We
Elman Jerald Roberts	White	Male	No	8/9/2018	8	September	2018	Shot	Attack	AL	Sou
Raymond Hernandez	Hispanic	Male	No	23/9/2019	23	September	2019	Shot	Other	CA	We
Lawrence Price	White	Male	No	17/9/2015	17	September	2015	Shot	Attack	KY	Sou
Major Carvel Baldwin	Black	Male	No	5/9/2020	5	September	2020	Shot	Attack	TX	Sou

Figure 155 - Dataset

The user can then select to either load the data or transform some variables. For this project, only Impale needed to be transformed into a whole number, which was done in the PowerQuery and the dataset was loaded and ready to start making the dashboards.

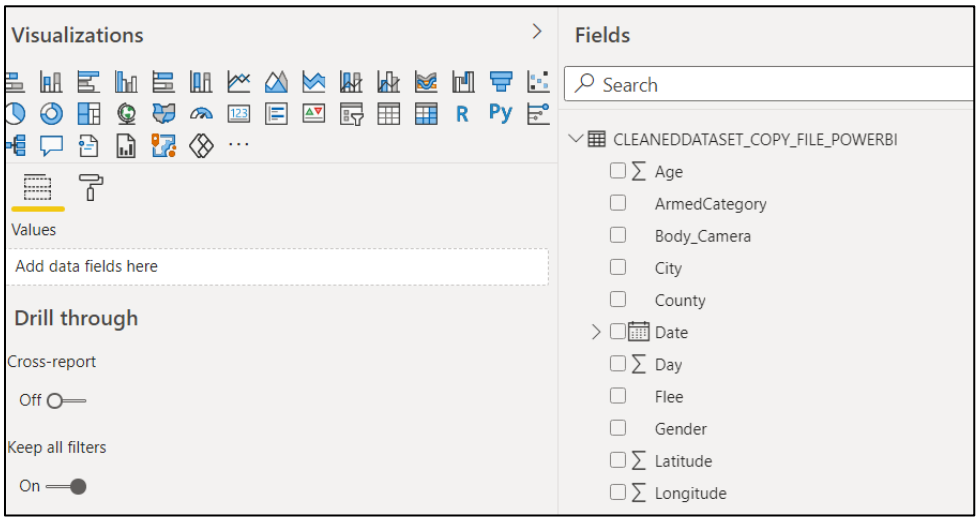


Figure 156 - Loaded Dataset in Power BI

There are four dashboards in total consisting of a general dashboard giving an overview of the dataset followed by a dashboard with some variables plotted against the targets and two dashboards consisting of the best model for each target variable: Interactive decision trees with their top 6 nodes. These dashboards along with the results of the prediction models may be used to enact laws and regulations that can reduce fatal encounters and allow for strict actions to be taken when these encounters do occur. The final discussion regarding this with all the possible suggestions is given in 6.3.

5.6.1 Visualization Dashboards

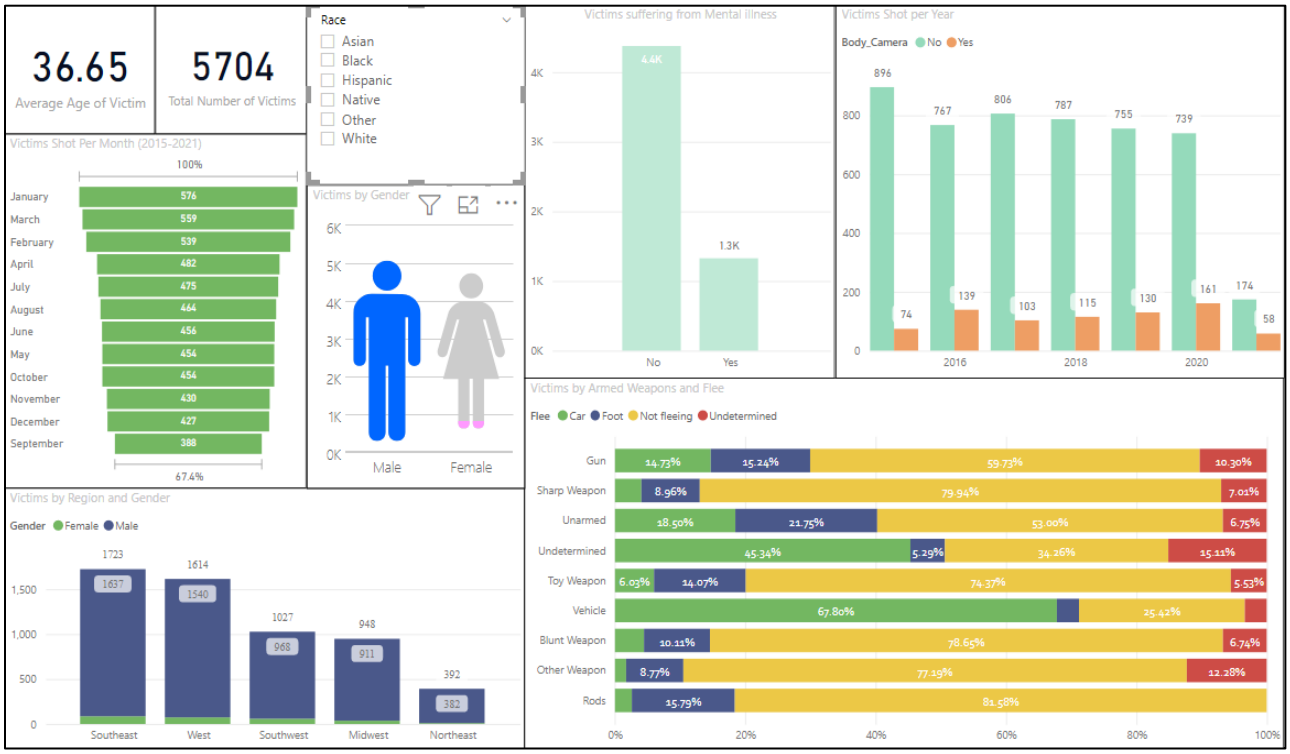


Figure 157 - Funnel Plot of Victims killed by Month

The victims shot per month funnel plot visualization takes the total number of victims that were shot from the years 2015 to 2021 and displays how many victims were killed per month. It can be seen that January has the highest number of deaths at 576 victims out of the total 5704 followed by March at 559, February at 539 and so on. November, December, and September have the least number of victims at 430, 427 and 388, respectively. This can attribute to the fact the start and end of the years come with a lot of festivities and may be a cause of shootings to happen due to the increased probability of crimes. In these situations, the victims may also be civilians who just happened to be in the wrong place at the wrong time

This stacked bar chart shows the victims per region and gender shows that the total number of victims was higher for Male (5,438) than females (266). For region Southeast in terms of Gender made up 28.70% of total Victims. There was the least number of victims killed in the Northeast however the majority of those killed in the Northeast consisted of male victims with 382 out of the total 392 victims. Southeast saw a similar difference between Male and Female victims as well with 1637 male victims and only 87 female victims, this was followed by West where there were 1614 total victims killed, 1027 in Southwest and 948 in Midwest. The police enforcement officials can take the opportunity to see what the reason for the shootings in the Northeast was to have such few victims as compared to the Southeast and try to implement it to reduce the overall shootings. An infographic also displays the difference between male and female victims, with the majority of 5704 victims being men.

Another stacked bar chart showing victims shot by armed weapon and flee displays the 9 types of weapon categories that a victim may have had at the time of their shooting. If it looked at whether the victim was fleeing or not, it can be seen that for all the weapon categories, the majority of the victims did not flee except for the Vehicle category. The majority of the victims had guns as a weapon but did not flee. However, when this is compared with another graph that plots Threat Level against Armed Weapon, it can be seen that the victims were deemed by the officer a threat to their own or the officer's life.

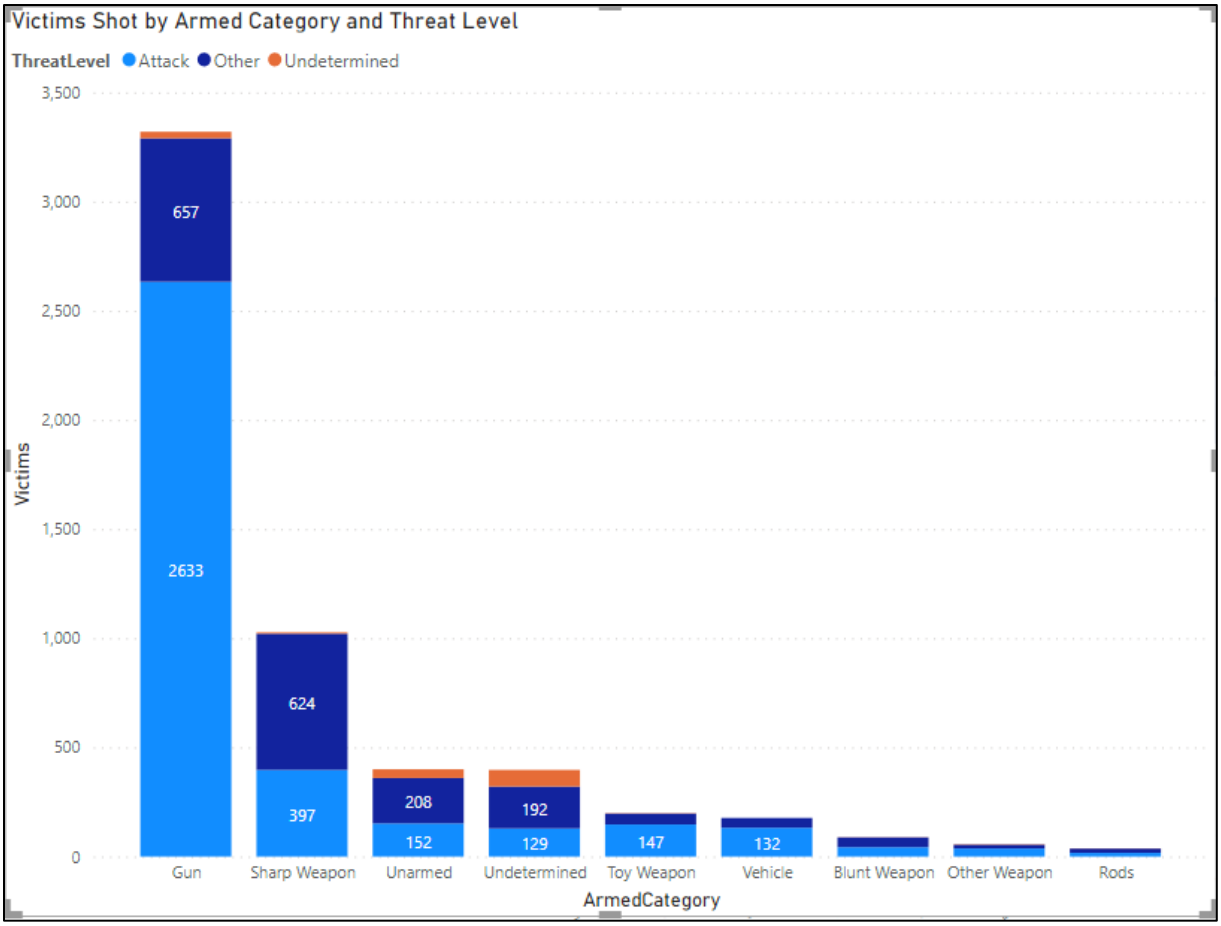


Figure 158 - Victims shot by Armed Category and Threat Level

It can also be seen those victims holding equipped with Sharp Weapon were not fleeing and around 624 out of the total of 1021 were Undetermined or in the Other category for threat level. These types of victims may have been scared and instead of shooting them, future encounters should take note of how to deescalate the situation and allow for the person holding a weapon to understand their life is not a threat. This may help reduce the shootings of innocent and scared people. The average age of victims can be seen as 36.65 and shows how young these victims are, more on this has also been discussed in chapter 6 concerning the best model's results. The total number of victims in this dataset is 5704 and out of those approximately 77% did not suffer from a mental illness at the time of the shootings.

This bar chart shows the number of victims killed per year since the start of the collection of this data. It displays the Victims against the year and categorizes them based on whether the police officer had a body camera on them or not. It can be seen that the average number of victims per

year remain above 700 in cases where the police officer does not have a body camera and remains below 200 in the cases that they do. This visualization can be used as an insight to see that when a police officer can be held accountable for their actions, the number of shootings decrease, hence it should be a law all over America to ensure that police officer in each state is equipped with a body camera. In certain states, this has already been implemented on the federal agents starting June 2021. (Gurman, 2021)

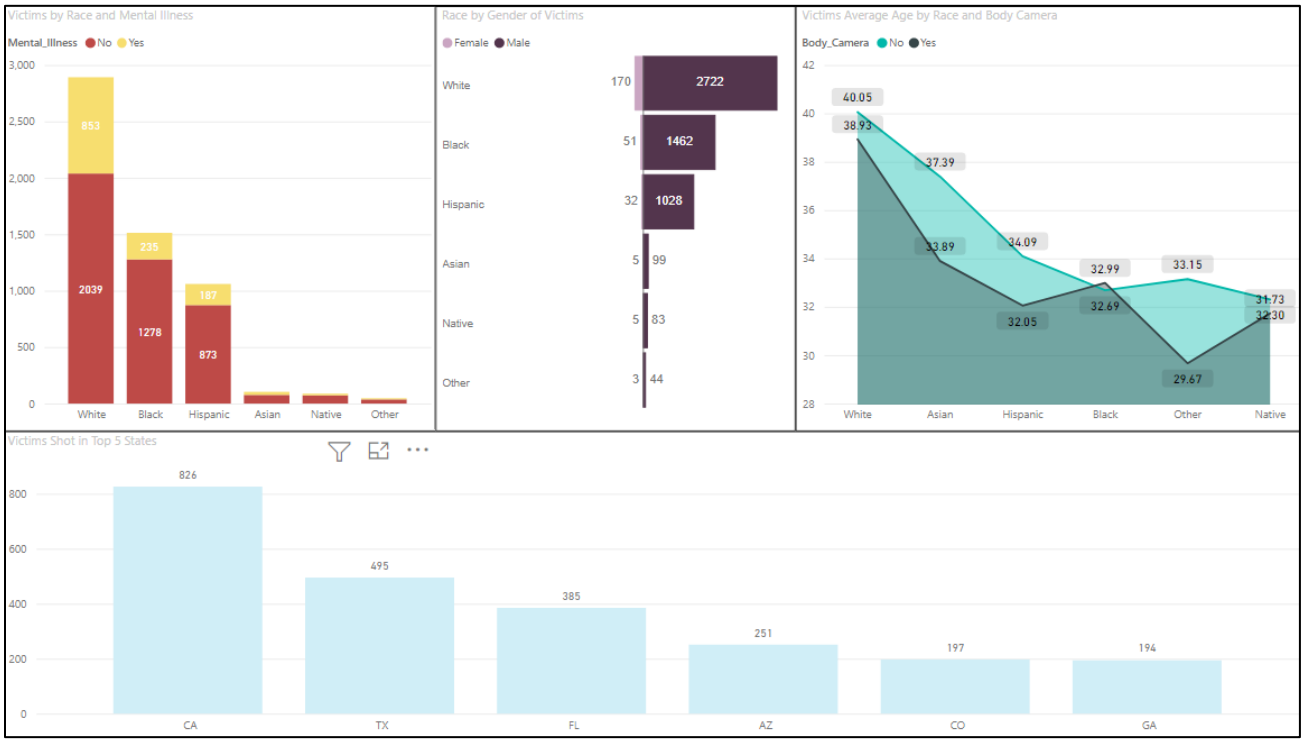


Figure 159 - Dashboard showing critical inputs against both target variables

The above dashboard shows few graphs with the target variables. The researcher has selected the inputs which seemed to be the most important in the prediction models and plotted them against both targets either as multivariate or bivariate plot.

The victims by mental illness and race show how many victims suffered from signs of mental illness and what was their race at the time of the shooting. The second graph plots the gender of the victims according to their race and it can be seen women make up very few of the victims, with Asian and Hispanic women making up a total of 10 female victims between them.

The average age of the victim was 36.65 and a graph has been plotted to show the average age of the victims by race and presence of police officers body cameras. For white people, they have a

higher average age in both cases of body cameras with 40.05 and 38.95 for No and Yes to body cameras, respectively. This shows that it is usually the older white population who are shot.

For Asians, the average age falls, with 37.39 and 33.89 for No and Yes to body camera respectively, this trend continues for Hispanic and Other race with 34.09, 32.05 and 33.15, 29.67, respectively. Oddly enough, it can be seen that for black victims, the average age remains similar in both cases of body cameras with 32.99 and 32.69. Research has been referenced (Chapter 6) with regards to the race of black victims where younger black people are at a high risk of being shot by the police and an example of this is Jenoah Donald who was only 30 years old at the time of his death and was shot by Clark county officers. (Brynelson, 2021) It was reported that he was not armed and initially cooperated, but the officers presumed that he had weapons and had a tussle with him which led to his death.

Another graph displays the victims shot by the top 5 states with California having the highest shootings at 826 deaths in the last 6 years followed by Texas at 495, Florida at 385, Arizona at 251, Colorado at 197 and Georgia at 194. California has a very high population and is a big tourist hot spot and thus these two facts may contribute to this large number. (Statista, 2020)

5.6.2 Dashboard – Decomposition Trees

These decision trees are made according to the top 6 node rules that have been described under the prediction modelling phase for each target variable. It was created in PowerBI using the data from the decision tree node by exporting the data using the save data node.

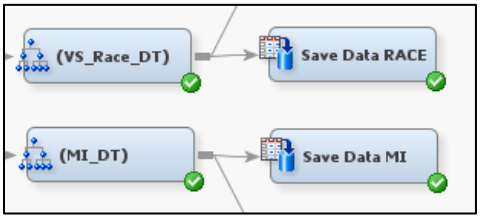


Figure 160 - Save Data Node

The save data node was used to export the data in a CSV format, an excel file. Both the validation and training datasets were downloaded from SAS Studio and combined under one file for each target variable. The file is then imported into PowerBI for model construction. An explanation for this has been given in Section 5.6.1. A decomposition tree visual is selected, and the trees are built

with the predicted targets outcomes average values. The variables for modelling are selected according to their importance in each decision tree and constructed.

5.6.2.1 Dashboard – Predicting Victim's Race

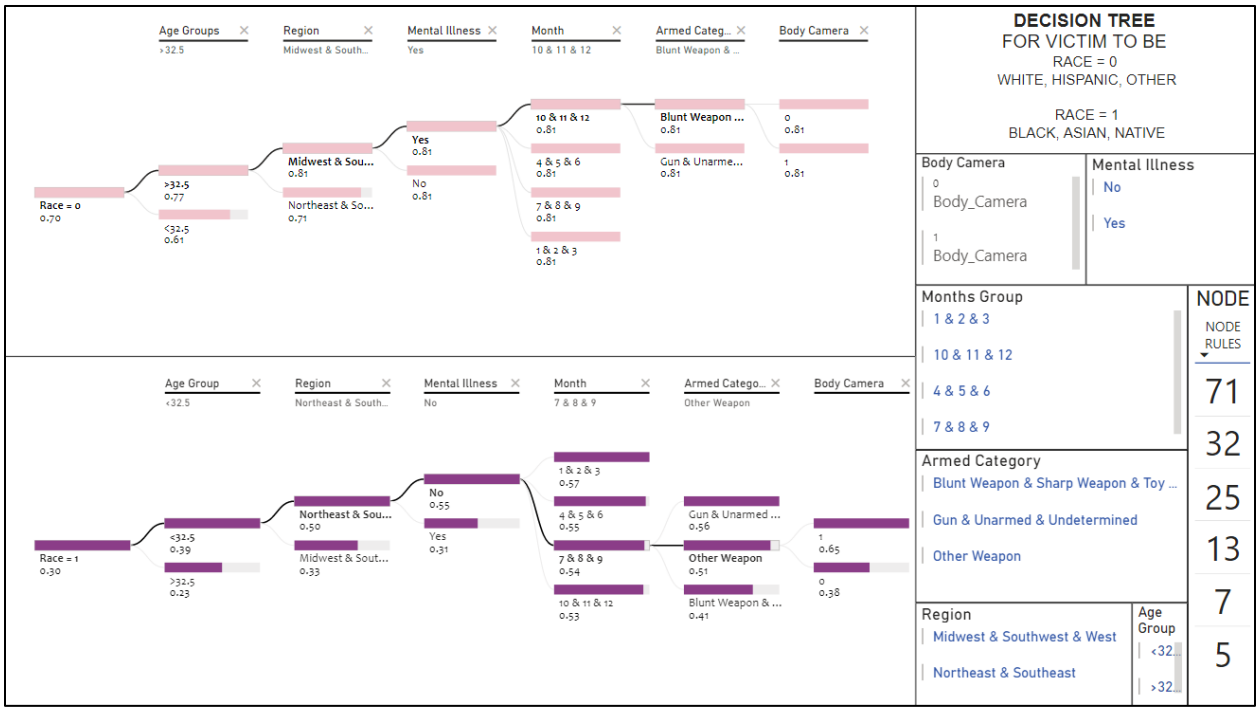


Figure 161 - Decomposition Tree in PowerBI (Race)

The dashboard consists of two decision trees. The tree on the top represents the predicted outcome of 0 while the tree at the bottom represents the predicted outcome of 1 for the race target variable. On the bottom right, there are the top 6 nodes of the decision tree that can be toggled with to see the prediction scores and the relative variables which are boldened when a node value is selected based on their importance. The value of the node that is selected is shown right at the beginning under the "Race =0" or "Race =1". These node rules have already been discussed under the decision tree modelling process. A few labels have been placed to show the inputs that have been used in the decision trees and how they were grouped according to the node rules.

5.6.2.2 Dashboard – Predicting Victim's Signs of Mental Illness

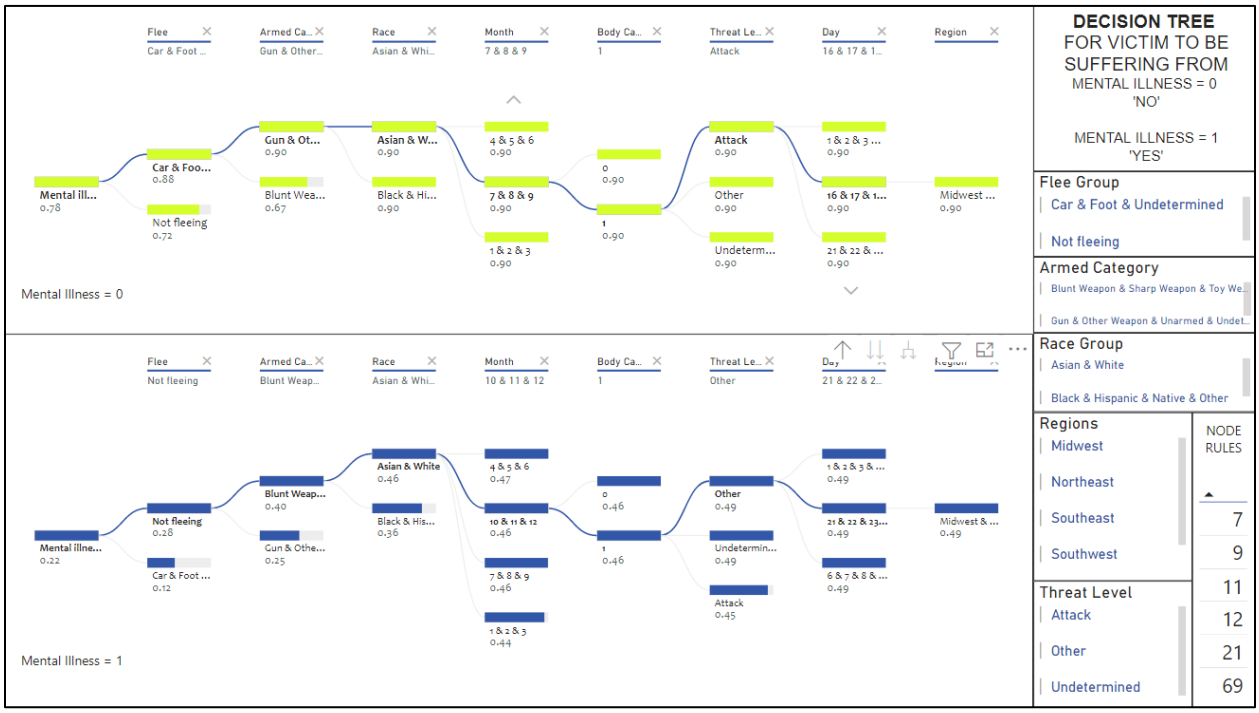


Figure 162 - Decomposition Tree in Power BI (Mental Illness)

The dashboard also consists of two decision trees. The tree on the top represents the predicted outcome of 0 while the tree at the bottom represents the predicted outcome of 1 for the signs of mental illness target variable. On the bottom right, there are the top 6 nodes of the decision tree that can be toggled with to see the prediction scores and the relative variables which are boldened when a node value is selected based on their importance. The value of the node that is selected is shown right at the beginning under the "Mental Illness =0" or "Mental Illness =1". These node rules have already been discussed under the decision tree modelling process. A few labels have been placed to show the inputs that have been used in the decision trees and how they were grouped according to the node rules.

5.7 Summary

Throughout this chapter, the entire data analysis process has been carried out in building an individual prediction model for the target variables: race and mental illness. The initial data exploration was carried out using the File Import and StatExplore node to get an understanding of the missing values and data type. This data was then used in SAS Studio to be pre-processed after which it was imported back to SAS EM to continue the data pre-processing by imputing missing

values such as age and grouping certain class variables into groups. Predictive modelling was then carried out with the aid of data partition, variable selection and transform variable node, after which individual regression and decision trees were modelled for each target variable. At an initial glance, it was understood that for both targets, the decision tree is the best performing model. Both the pre-processed data and the data from the decision tree was used to carry out the Data Visualization in Section 5.4 and the Reports and Dashboards in 5.6. The model evaluation will be done in the next chapter.

CHAPTER 6: Results and Discussion

6.1 Introduction

The purpose of this project was to evaluate how mental illness and race of a victim might contribute to the fatal police shootings in America and whether a person who displays a sign of mental illness may be shot or whether people of colour are more likely to be shot due to their race. This is very important as stated by the objectives of this paper; The Washington Post where the dataset has been selected, has been gathering information on these shootings and updating it daily for insight into these brutal events however, there is more that can be contributed to the data gathering which can further develop more detailed insights and answers. The results we have gathered from the above chapter will be discussed in detail along with some relevant studies below.

6.2 Model Evaluation

6.2.1 Model Comparison for RACE

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable ▲	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors
Y	Tree	Tree	(VS_Race_DT)	Race		0.271295	0.267419	0.189355	0.192843	3990	0.806174	1511.054
	Reg	Reg	(T_VS_Race_Reg)	Race		0.275963	0.27594	0.190167	0.188743	3990	0.95911	1517.535
	Reg4	Reg4	(VS_Race_Reg)	Race		0.278296	0.274436	0.189926	0.189023	3990	0.959883	1515.606

Figure 163 - Fit Statistics

Misclassification Rate is used as the assessment measure to build the models hence it will be used to assess each model based on its validation data. It is noted that the lower the MISC, the better the accuracy. It can be seen that for the Decision Tree (VS_Race_DT), the validation data's MISC is the lowest amongst all three models, as well as in the train data, hence it is selected as the best model amongst the three. The MISC for the decision tree is 0.271 as compared to the Reg's (T_VS_Race_Reg) 0.275 and Reg4's (VS_Race_Reg) 0.278. The decision tree model is 73% accurate as compared to the approximately 72% accuracy of the regression models. According to the decision tree, the most important variables that affected whether a person's race would be 1 or 0 at the time of the shooting were AGE, REGION, MENTAL ILLNESS, MONTH, ARMED CATEGORY, AND BODY CAMERA.

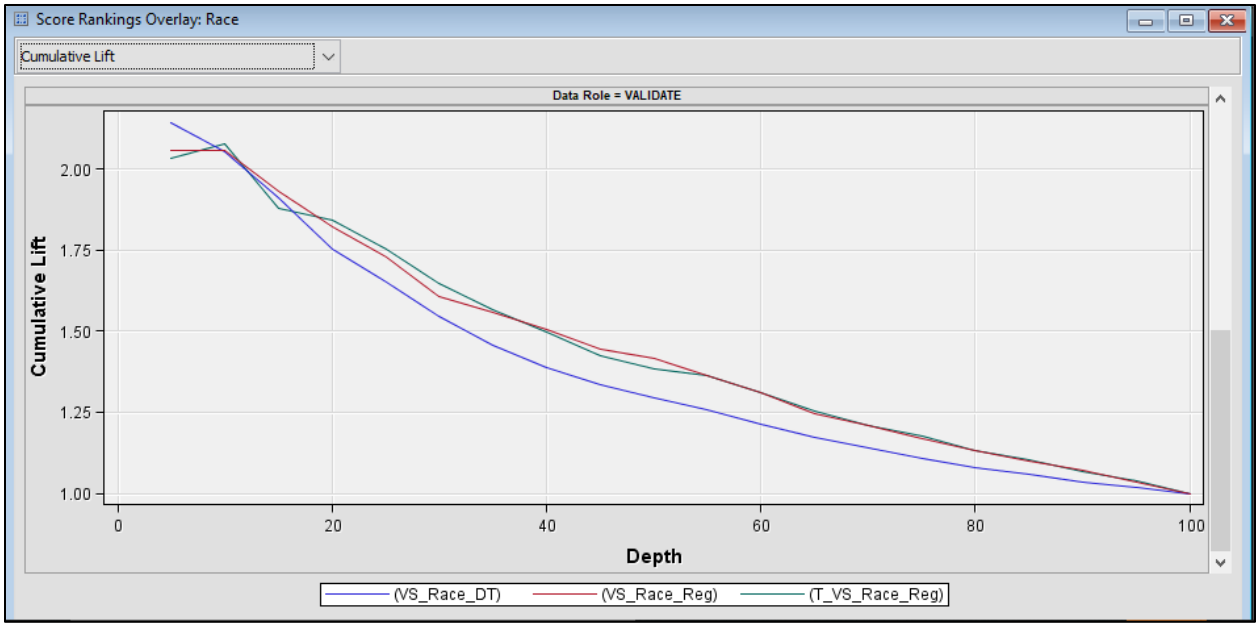


Figure 164 - Cumulative Lift Chart

The cumulative lift can also be used to assess the model and determines which model is the best performing. The blue line represents the decision tree (VS_Race_DT) and has the highest cumulative lift amongst the model in the validation data with 2.146 while the (T_VS_Race_Reg) model had a value of 2.033 and the (VS_Race_Reg) model had a value of 2.059.

For the RACE target, as it was described that the validation misclassification for the first regression model (VS_Race_Reg) was 0.278296 while the second (T_VS_Race_Reg) regression model produced a MISC of 0.275963. Moreover, as the 2nd model was transformed before variable selection, it can be referred to as a more complex model due to having a greater number of steps.

6.2.2 Model Comparison for SIGNS_OF_MENTAL_ILLNESS

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error	Target Label	Train: Sum of Frequencies	Train: Maximum Absolute Error	Train: Sum of Squared Errors
Y	Tree3	Tree3	(MI_DT)	Signs_Of_Mental_Illness	0.229288	0.221554	0.158908	0.160781		3990	0.895522	1268.083
	Reg2	Reg2	(VS_MI_Reg)	Signs_Of_Mental_Illness	0.231039	0.226817	0.161144	0.16042		3990	0.972447	1285.926

Figure 165 - Fit Statistics

Misclassification Rate is used as the assessment measure to build the models hence it will be used to assess each model based on its validation data. For SIGNS_OF_MENTAL_ILLNESS as a target, the decision tree was selected as the best model amongst both models based on the

validation data's misclassification rate. It is 0.229 as compared to the regression models 0.231 The decision tree model was 77.1% accurate while the regression model was 76% accurate. The inputs that very seen to impact whether a victim would be suffering from mental illness at the time of the shooting would be THREAT LEVEL, FLEE, BODY CAMERA, REGION, RACE AND MONTH.

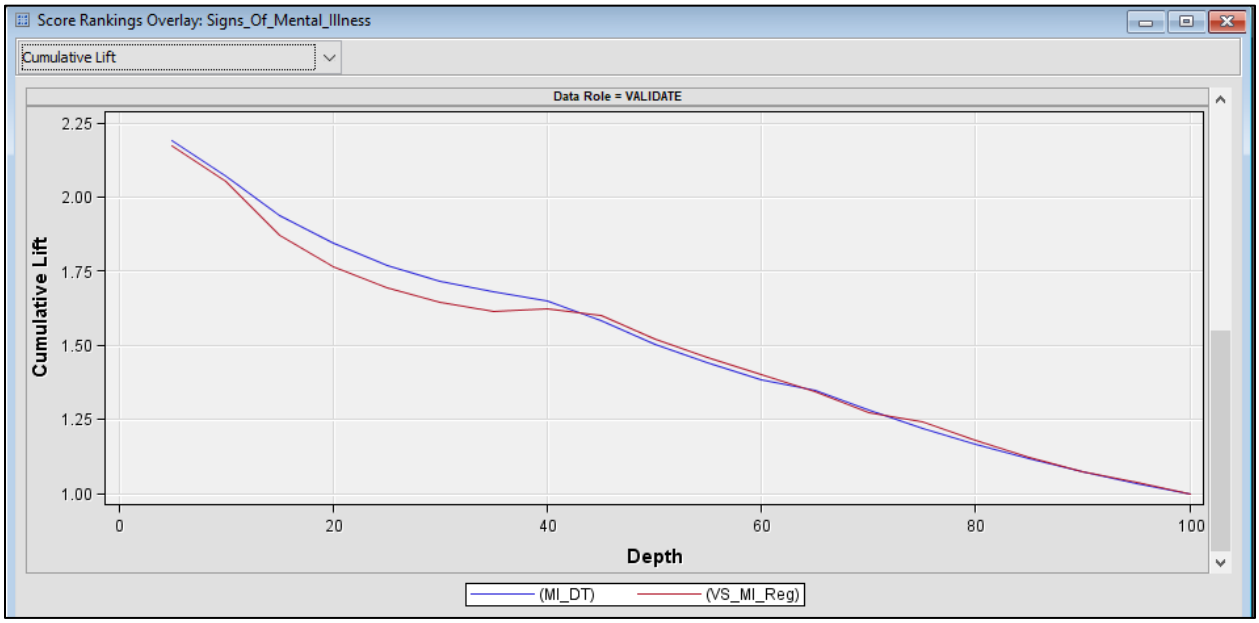


Figure 166 - Cumulative Chart

The chart shows that both the regression model and decision tree are close in being the best model, but the decision tree is just slightly better than the regression model with a CL of 2.189 and 2.171, respectively.

6.3 Discussion

The model evaluation has been discussed in the section before and the result that was obtained will be critically analyzed in this section for both targets: RACE and SIGNS_OF_MENTAL_ILLNESS along with a comparison with similar existing studies as well as suggestions using the produced output of the decision tree and the dashboard will be given.

As discussed throughout chapter 5, both the targets were binary variables hence it was suitable to analyze them from the perspective of the misclassification rate evaluation metric. According to the model evaluation above, it has been determined based on the MISC for both targets, a decision tree is the best optimal model for prediction analysis. The criterion Gini was applied to both the

decision trees. Gini as a criterion was not found in many studies instead other research using different but still, binary target variables used other criteria such as CHAID or Entropy or even Random Forest model which produced the best model for them in terms of their evaluation metrics. One such example is where Dhanabal, Kopasz and Lindsay (2020), built various decision tree models to predict whether a case of officer-involved shooting would be disposed of or not and determined that for their model, CHAID worked best with both evaluation metrics. Streeter (2019) analyzed RACE as an outcome variable and converted it to a binary variable with 1 signifying African Americans and 0 signifying Non-White Hispanics. They used a random forest model to carry out their prediction, out of all of their models, it was the best performing in terms of accuracy. Analysis using entropy and Gini in the decision tree and random forest respectively were also carried out by Yerpude and Gudur (2017).

6.3.1 Victim's Race

As per the decision trees node rules for RACE of the victim, it could be seen that when the victim's age was greater than 32.5 years old and they were within the region of west, southwest, midwest they were more likely to be either White, Hispanic or some Other race. According to a study conducted by Edwards, Lee and Espostio (2019), they state that the risk of being shot by the police is more likely to happen between the ages of 20 and 35 years old for all types of ethnic groups and Haynes (2020) has said that there is a higher chance of having a fatal police encounter in the West and South region than any other region in America. The decision tree had also considered when the body camera of a police officer was not available, and the age was between 25 and 32 then the race of the victim would 54% more likely to fall under the 1 category. Studies have shown that when the body camera is 'On' a police officer, there is a drop in the use of force by police which shows that it is important to implement the rule of body cameras in America to see a drop in these shootings. (Corley, 2021) Just recently, the justice department made it mandatory for federal agents to wear a body camera while carrying out arrest warrants and any activities as such. (Gurman, 2021) This is a step in a good direction, as the countrywide implementation of body cameras can greatly help in situations where officers are guilty of murder but there is no solid evidence to back it up or they are innocent and can use the footage of the body camera to prove it. The dashboard above further proves the minimal fatal encounters when a police officer had a body camera as compared to not having them.

Moreover, the victim's possibility of suffering from mental illness and being armed was also analyzed. When a victim had no signs of mental illness, it showed that the race of the victim was predicted to be 1 by 69% (Black, Asian and Natives). This change is further categorized by the fact that it is a victim younger than the age of 24. Referring back to Haynes (2020), she further reiterates the fact that young black people are at a tremendously high risk of being killed such as a 650% chance of black Chicagoans being killed than white Chicagoans. (Chicago falls under the Midwest region) This finding further suggests that there is some bias when police officers think that a person of colour is armed. This can be because they have been in situations with armed minorities where the latter was a threat. (Nix et al.,2017)

It is important to remove racial intolerance from police officials to ensure that all people are treated fairly when they encounter law enforcement, this is where senior officials in these departments can make it mandatory for police officers to be trained as well as make them understand the implications of killing victims unnecessarily and justifying it through stereotypical bias. Moreover, stricter punishment for officers who killed victims unreasonably should be carried out. Law enforcement should also find ways to bring about a positive outlook on the citizens instead of being seen as 'dangerous' or 'scary' such as suggested by Nix et al. (2017) where a community policing activity with some minorities in specific neighbourhoods will bring positive changes on both sides such as where the officer is also exposed to citizens who are *not* committing any crimes as well as exposing citizens to officers who are seen as friendly and helpful. This can inherently reduce bias in both parties during a police encounter.

6.3.2 Victim's Signs of Mental Illness

The decision tree model for the victims having signs of mental illness display shows that when there is unsureness about the victim's possibility of fleeing or being armed, they are less likely to suffer from any mental illness by 90%. In a study conducted by White (2002), he states that in the case a police officer suspects the victim is armed or planning to run, they will use deadly force in self-defence or prevent the victim from hurting themselves. Due to the news in recent times, any person when encountering a police officer feels extremely nervous and may resort to "defending" themselves from a possible "threat" even if the police officer intends to just check upon them. This sometimes may escalate and cause a fatal encounter to occur.

The model further assesses the race of the victim to be either White or Asian along with the fact they are not fleeing anymore, however are armed with a weapon. It shows an increase in the possibility of the victim suffering from mental illness by 32% at the time of their shooting. In some situations, victims are suffering from mental illness but that is not visible to the officer, such as a 911 call from the family or friend. As they are unable to lay out the full situation over the phone, there are not able to mention specific factors to the officials. (Frankham, 2017) Frankham further states that in case they were a possibility of informing the officers of these factors, situations would not escalate to deadly police encounters. Another factor that is seen here is the region. The region was also seen in the decision tree for the race target variable. A possible way to reduce the volume of killings as a result of 911 calls is to ensure that there is as much information retrieved from the person or family making the call as possible, this can help inform the law enforcement officials who will be attending to that person and can also ensure that only unarmed and trained officers are responding to these crises. This is further restated in a study by Edward, Lee and Esposito (2019).

When the region of the victim falls under Midwest, Southeast or West and their race is either white or Asian coupled with the fact they are not fleeing but are armed, it shows that there is almost an equal possibility of the victim having a mental illness as to not have any signs of mental illness. This suggests that when there is a higher chance of the victim being armed, then there is also an increase in the possibility of them suffering from mental illness, but it is still 51% likely they do not suffer from mental illness when they were shot. Frankham (2018) states that sometimes when people suffering from mental illness are ordered by the police officer to carry out a task, they may be unable to do it due to their illness which causes them to at risk of death. Moreover, she states that they are also less likely to be armed with dangerous weapons. Mentally ill people can easily be calmed down if the police officer is trained to do it well as they do not react well when they are stuck in situations out of their control which can cause them to become violent without meaning to. The above findings also suggest that when the race of the victims is seen to be white or Asian, they are more likely to be coming in contact with the police in these encounters as they are suffering from a mental health crisis. This is further echoed in a study conducted out by Streeter (2019) where they use race as a target outcome and found that factors such as intoxication and mental health crisis, in most cases always point towards a victim who is of white descent. It is important that mental health is taken seriously in America and various psychiatric care centres are established to allow mentally ill people to seek help. If people who are suffering from mental

illness whether as a result of intoxication or drug abuse can find a safe place to seek help and get better, they are less likely to be victims of fatal shootings where they possibly may be unaware that there is a possibility, they could pose a threat to themselves, and the other officials or people involved.

For both targets it is noted that gender does not seem to be an important variable to consider in police use of force in fatal police encounters, however, various studies mention that police officers are more likely to use lethal force on males. (Nix et al., 2017) This type of information may also be seen for the dashboards that have been created above and from the general understanding of the dataset where 90% of the victims are male.

Nix et al., (2017) and Wood et al., (2020) brought attention to the fact that when police officers are trained, they were less likely to use force when they encountered any person. Officers who received procedural training showed more patience and a better attitude when dealing with citizens in an encounter as compared to those who did not have any training. This proves that if training is provided along with having consequences for the actions carried out by law enforcement, then these fatal encounters can decrease and become less likely to occur.

Chapter 7: Conclusion and Reflection

7.1 Conclusion

Due to access to datasets such as the fatal police shooting dataset used in this project, there were various points and relations that were able to be touched upon. The insights that were gained can greatly help change how these encounters happen but there is still a long way to go.

The above suggestions given are very important to enact change in many aspects as it is important to ensure the protection of the civilians of a country without bringing in racial discrimination or stigma against people suffering from mental illness, neither should be their discrimination based on gender as it was seen men were at a very high risk of being shot irrespective of race. Law enforcement officials have a massive responsibility towards their civilians. By using the analytical insights and knowledge gained, it can greatly help bring back the trust the community has lost in the officials as well as ensure that proper punishment has been given. Moreover, through this, occurrences of other crimes influenced by these fatal encounters such as hate crimes and mass school shootings may also see a reduction, such as through the implementation of strict gun laws. Dashboards and the prediction analysis have given an easy to articulate method of presenting the data in facts and figures which helps gain a deeper understanding without losing sight of the matter at hand.

7.2 Reflection

Four main objectives were stated at the beginning of the project that was needed to be accomplished to deem this project a success. Each objective will be touched upon and how it was accomplished and where there could be improvements in the future.

1. To investigate the key factors to analyse the influence of mental illness and racial discrimination on fatal police shootings.

The analysis produced several variables in relation to each target which helped predict how they would have an effect on the prediction of a person's mental illness or race at the time of the shooting.

A drawback was due to the nature of the SAS Enterprise Miner program, the race variable that had to be used as a target had to be changed to a binary categorical variable. According to certain specifications, the races were divided with 1 = 'Black', 'Asian', 'Native' while 0 = 'Hispanic', 'White', 'Other'. So instead of predicting exactly what race the victim is likely to be due to certain factors, the analysis instead had to be done according to grouped races. For each target, the factors overlapped in importance, the table below shows important factors for both targets and are ranked in terms of importance according to the result of the decision tree.

Race	**Signs of Mental Illness**
Age	Flee
Region	Armed Category
Mental Illness	Race
Month	Month
Armed Category	Body Camera
Body Camera	Threat Level
	Day
	Region

For both, it can be seen that Armed Category, Month, Body Camera, Region overlap as the critical factors. Moreover, it is noted that each target also appears in the others analysis showing that there is a possibility of a link between the two.

2. To analyse the influence of mental illness and racial discrimination on fatal police shootings using explanatory data analysis techniques.

An exploratory data analysis was carried out using the pre-processed dataset after importing for SAS Enterprise Miner and various graphs were plotted to see how the relationship between different variables occur as well as with the targets. These were carried out under the visualization subsection in chapter 5 as well as a thorough discussion was carried out. A gap here can be not using the census data for the population of the United States. This is something that can be further

explored and be used to get a deeper insight on the percentage of races to the population of victims killed and may give an even deeper insight. Edwards, Lee and Esposito (2019) carried out an analysis to see the link between the age, race and ethnicity of a person and found that there is a high risk of death for young men who are BIPOC by the police. This was found in this research as well with most victims being men.

3. To evaluate the performance of analysis results using evaluation metrics such as R-Squared (R^2), Mean Squared Error (M.S.E), Mean Absolute Error (M.A.E) or Misclassification Rate (MISC).

As stated, there were two model types developed for each target to assess which would be a better-suited model to aid in the prediction of a person's signs of mental illness or race in possible fatal shootings. For the race target, there were two regression models built with different nodes and one decision tree while for signs of mental illness there were one of each model built. For both targets, the decision tree built using the Gini criterion was deemed to the best model in terms of Misclassification Rate. The Average Square Error was also analysed for each model and although the regression models showed a smaller value, they were prone to overfitting as the validation values were smaller than the train data hence in terms of ASE, decision trees were still deemed to be the better models. For future research, more evaluation metrics such as the confusion metric should be used to select the best prediction model.

An issue the researcher encountered was the lack of information on which type of decision tree would be most suited for this dataset. Other research that has been discussed in chapter 6 used other algorithms such as Entropy or Probability Chi-Square for the decision trees and found them to be better suited for their prediction models. There was uncertainty whether the algorithm used would provide the best result however the researcher carried out the prediction using a decision tree with each criterion and assessment measure and eventually found that Gini worked best with both targets and produce better results in terms of ASE and MISC for all decision trees hence selected Gini as the criterion.

4. To develop a dashboard to visualize the analysis results

The data of the decision tree was saved from SAS EM in excel file format. Both the validation and training data were merged into one file and copied into PowerBI where a decomposition tree was

built using the excel file with each outcome for both targets modelled, along with the node rules with the most observations being displayed, so whenever a specific number of the node is clicked, it shows the outcome of the decision tree and boldens the important variable and displays the prediction value.

Overall, the researcher believes the project has been a success in terms of carrying out the specified objectives as well as ensuring it is carried out ethically by not using the names of the victims. Multiple relationships of every variable were analyzed with the targets. The most critical factors were found and used in prediction models to see how it would be of effect to a person's mental illness or race at the time of a possible fatal police encounter.

7.3 Challenges

There were many challenges that the researcher faced while carrying out this project. Some of them have already been discussed above. Other included the pre-processing of the data. This by far was the most challenging aspect of the whole project as there were multiple crashes, and the researcher was also overwhelmed on how to model with race without compromising on the quality of the variable. This was overcome by changing its data type. This was also the process that took the most time as it needed to be done as accurate as possible to produce the desired outcomes. The next issue was finding articles and papers that correlated with this research however the information was scarce as well as restrictions on the most important papers, but this issue was overcome by using different keywords and using other platforms to access the data.

Lastly, a minor issue to some that seemed difficult for the researcher was the management of time as the time zone difference was great. There was a lot of difficulties managing the meetings as well as ensuring proper progress was ongoing. This truly taught the researcher the meaning of time management and discipline.

7.4 Future Improvements

An important thing to consider for future work is having more data. The fatal police shooting data, although comprehensive, lacks a lot of information that possibly can help determine more, such as marital status, the race of the police, the time of the encounter, just to name a few possible variables. Signs of mental illness should also be broken down into whether a person was intoxicated or angry or any manner of such as that would further help develop the analysis. The

race of the unknown victims should also be researched and corrected instead of being left blank or as 'Other'. The researcher had decided to code the unknown as Other and remove the missing rows as they would skew the results. Other predictive modelling tools and techniques may also further improve the analysis, specifically programs that can handle multinomial values or categorical data without having to modify the data too much. The researcher found evidence of that in programs such as SPSS and R however took on the challenge of developing this project in SAS Enterprise Miner to bring a new perspective and challenge to the study.

Finally, these research results and models may be applied to other countries such as the UK, Pakistan, Kenya as these are some of the countries which have minimal laws concerning the use of lethal force in police encounters and may be used to see what factors are linked and what can be done to improve the situation.

CHAPTER 8: References and Appendices

8.1 References

1. American Psychological Association. (2020). *Stress in America 2020: A National Mental Health Crisis.* [online] Available at: <https://www.apa.org/news/press/releases/stress/2020/sia-mental-health-crisis.pdf> [Accessed 27 February 2021].
2. Amnesty (2020) What Is Police Brutality? . [online] Available at: <https://www.amnesty.org/en/what-we-do/police-brutality/> [Accessed 24 December 2020].
3. AP NEWS. (2019). *Timeline of events in shooting of Michael Brown in Ferguson.* [online] Available at: <https://apnews.com/article/9aa32033692547699a3b61da8fd1fc62> [Accessed 25 February 2021].
4. Aunalytics. (2015). *Decision Trees: An Overview - Aunalytics*. [online] Aunalytics. Available at: https://www.aunalytics.com/decision-trees-an-overview/#:~:text=Decision%20trees%20tend%20to%20be,of%20data%20into%20smaller%20segments.&text=A%20regression%20tree%20is%20used%20to%20predict%20continuous%20quantitative%20data. [Accessed 19 May 2021].
5. Austin Police Department. (2015). Officer - Involved Shootings: 2000 - 2014. [online] Austin: Austin Police Department, p.2. Available at: <https://web.archive.org/web/20161228075326/http://www.austintexas.gov/sites/default/files/files/Police/OIS_report_2014.pdf> [Accessed 24 December 2020].
6. Ayer, L. and Ramchand, R. (2021). *Is Mental Illness a Risk Factor for Gun Violence?* [online] Rand.org. Available at: https://www.rand.org/research/gun-policy/analysis/essays/mental-illness-risk-factor-for-gun-violence.html [Accessed 9 July 2021].
7. Azevedo, A. and Santos, M. (2008). KDD, semma and CRISP-DM: A parallel overview. In: IADIS European Conference Data Mining 2008. [online] Amsterdam: IADIS, pp.182-185. Available at: <https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview> [Accessed 16 May 2021].
8. Bacon, P. (2020). How The Police See Issues Of Race And Policing. [online] FiveThirtyEight. Available at: <https://fivethirtyeight.com/features/how-the-police-see-issues-of-race-and-policing/> [Accessed 27 February 2021].
9. Belli, B.(2020). *Racial Disparity In Police Shootings Unchanged Over 5 Years* . [online] YaleNews. Available at: <https://news.yale.edu/2020/10/27/racial-disparity-police-shootings-unchangedover-5-years> [Accessed 25 December 2020].

10. Bhandari, A., 2020. Difference Between R-Squared and Adjusted R-Squared. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/07/difference-between-r-squared-and-adjusted-r-squared/> [Accessed 28 February 2021].
11. Bhardwaj, B. and Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. IJCSIS) International Journal of Computer Science and Information Security, [online] 9(4). Available at: https://arxiv.org/ftp/arxiv/papers/1201/1201.3418.pdf [Accessed 10 July 2021].
12. Borah, S., Ashour, A.S., Babo, R., Dey, N. and Panigrahi, R. (2019). Social Network Analytics Computational Research Methods and Techniques. ed. [online] www.sciencedirect.com, Academic Press, pp.1–19. Available at: https://www.sciencedirect.com/science/article/pii/B9780128154588000013 [Accessed 10 July 2021].
13. Brook, J. (2020). *Known Issues and Limitations with PowerBI.* [online] Supermetrics Support Forum. Available at: <https://support.supermetrics.com/support/solutions/articles/19000092165-known-issues-and-limitations-with-power-bi> [Accessed 26 February 2021].
14. Brooks-Bartlett, J. (2018). *Probability concepts explained: Maximum likelihood estimation*. [online] towardsdatascience. Available at: https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1 [Accessed 16 June 2021].
15. Brown, J.D. (2011). Likert items and scales of measurement? SHIKEN: JALT Testing & Evaluation SIG Newsletter, [online] pp.1-5. Available at: <https://hosted.jalt.org/test/PDF/Brown34.pdf> [Accessed 4 June 2021].
16. Brownlee, J. (2016). *Linear Regression for Machine Learning.* [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/linear-regression-for-machine-learning/ [Accessed 19 May 2021].
17. Brynelson, T. (2021). Jenoah Donald, 30-year-old Black man shot by Clark County deputies, has died. [online] opb. Available at: https://www.opb.org/article/2021/02/12/jenoah-donald-clark-county-sheriffs-office-shooting-death/ [Accessed 11 July 2021].
18. Chatterjee, T.K. (2020). *Why using CRISP-DM will make you a better Data Scientist?* [online] GreatLearning Blog: Free Resources what Matters to shape your Career! Available at: https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/ [Accessed 18 May 2021].
19. Chauhan, N.S. (2020). *Decision Tree Algorithm, Explained - KDnuggets*. [online] KDnuggets. Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html [Accessed 16 June 2021].

20. Cheney, A.J. (2020). *Understanding Deadly Police Encounters with Data Science*. [online] Medium. Available at: https://towardsdatascience.com/understanding-deadly-police-encounters-with-data-science-3cf1192d9778 [Accessed 17 May 2021].
21. Choudhury, A. (2019). *10 Model Evaluation Techniques Every ML Enthusiast Must Know*. [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/10-model-evaluation-techniques-every-machine-learning-enthusiast-must-know/ [Accessed 20 May 2021].
22. Choueiry, G. (2021). *Understand Forward and Backward Stepwise Regression – Quantifying Health.* [online] Quantifyinghealth.com. Available at: https://quantifyinghealth.com/stepwise-selection/ [Accessed 10 July 2021].
23. Clark, T., Cohen, E., Glynn, A., Owens, M., Gunderson, A. and Schiff, K. (2020). Are Police Racially Biased in the Decision to Shoot?. [online] Atlanta: Emory University, pp.1-50. Available at: <https://static1.squarespace.com/static/58d3d264893fc0bdd12db507/t/5ed6859f3c31fe420713f58b/1591117221040/Racial_Bias_in_Shootings.pdf> [Accessed 17 May 2021].
24. Corley, C. (2021). *Study: Body-Worn Camera Research Shows Drop In Police Use Of Force*. [online] NPR.org. Available at: https://www.npr.org/2021/04/26/982391187/study-body-worn-camera-research-shows-drop-in-police-use-of-force [Accessed 8 July 2021].
25. CSUSM. (2020). *Defining Diaspora: Asian, Pacific Islander, and Desi Identities | Cross-Cultural Center | CSUSM*. [online] Available at: https://www.csusm.edu/ccc/programs/diaspora.html [Accessed 12 June 2021].
26. Dåderman, A. and Rosander, S. (2018). *Evaluating Frameworks for Implementing Machine Learning in Signal Processing A Comparative Study of CRISP-DM, SEMMA and KDD.* [online] Diva-Portal. Available at: http://www.diva-portal.org/smash/get/diva2:1250897/FULLTEXT01.pdf [Accessed 17 May 2021].
27. Data Science Project Management. (2021). *CRISP-DM - Data Science Project Management*. [online] Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 10 July 2021].
28. Data Science Project Management. (2021). *KDD and Data Mining - Data Science Project Management*. [online] Available at: https://www.datascience-pm.com/kdd-and-data-mining/ [Accessed 10 July 2021].
29. Dei, M. (2019). Catalog of Variable Transformations To Make Your Model Work Better. [online] Medium. Available at: https://towardsdatascience.com/catalog-of-variable-transformations-to-make-your-model-works-better-7b506bf80b97 [Accessed 5 July 2021].
30. Delwiche, L. and Slaughter, S. (2018). *SAS Studio: A New Way to Program in SAS, continued*. [online] SAS Institute Inc, pp.1-20. Available at: <https://www.lexjansen.com/wuss/2016/30_Final_Paper_PDF.pdf> [Accessed 15 June 2021].

31. Dhanabal, A., Kopasz, M. and Lindsay, A. (2021). What Happens After Police Shootings? *SAS GLOBAL FORUM 2020*. [online] Oklahoma: SAS Institute, p.2. Available at: <https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/5200-2020.pdf> [Accessed 23 June 2021].
32. Donges, N. (2019). *A complete guide to the random forest algorithm*. [online] Built In. Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 18 May 2021].
33. Edwards, F., Lee, H. and Esposito, M. (2019). Risk of being killed by police use of force in the United States by age, race–ethnicity, and sex. *Proceedings of the National Academy of Sciences*, [online] 116(34), pp.16793–16798. Available at: https://www.pnas.org/content/116/34/16793#sec-3 [Accessed 25 December 2021].
34. Eng, C. and Wenig, B. (2020). *Analysis of Police Fatal Shootings in the U.S*. [online] Databricks. Available at: <https://databricks.com/blog/2020/11/16/fatal-force-exploring-police-shootings-with-sql-analytics.html> [Accessed 27 February 2021].
35. Fabian, L. (2017). *GBI investigating after Georgia Tech police shoot and kill student with knife*. [online] Macon Telegraph. Available at: https://www.macon.com/news/local/crime/article173895406.html [Accessed 12 June 2021].
36. Frankham, E. (2017). Ethnoracial and Mental Health Disparities in How Fatal Police-Public Encounters are Initiated. Ph.D. University of Wisconsin–Madison.
37. Frankham, E. (2018). Mental Illness Affects Police Fatal Shootings. Contexts, [online] 17(2), pp.70-72. Available at: <https://journals.sagepub.com/doi/pdf/10.1177/1536504218776970> [Accessed 25 February 2021].
38. Fuller, D., Lamb, H., Biasotti, M. and Snook, J. (2015). Overlooked in the Undercounted: THE ROLE OF MENTAL ILLNESS IN FATAL LAW ENFORCEMENT ENCOUNTERS. [online] Office of Research and Public Affairs, pp.1-14. Available at: <https://www.treatmentadvocacycenter.org/storage/documents/overlooked-in-the-undercounted.pdf> [Accessed 25 February 2021].
39. Geller, W. and Toch, H. (1996). *Police Violence: Understanding and Controlling Police Abuse of Force*. 1st ed. New Haven, CT: Yale University Press, pp.273-276.
40. Georges, J., Thompson, J., Wells, C., Bohannon,T., Hardin,M., Kelly, D., Lucas,B., and Walsh, S. (2010). *Applied Analytics Using SAS Enterprise Miner Course Notes.* Cary: SAS Institute Inc.
41. Gillis, A.S. and Silverthorne, V. (2017). *integrated development environment (IDE)*. [online] SearchSoftwareQuality. Available at: https://searchsoftwarequality.techtarget.com/definition/integrated-development-environment [Accessed 2 March. 2021].

42. Grace-Martin, K. (2012). *Why use Odds Ratios in Logistic Regression - The Analysis Factor*. [online] The Analysis Factor. Available at: https://www.theanalysisfactor.com/why-use-odds-ratios/ [Accessed 6 July 2021].
43. Gurman, S. (2021). U.S. Mandates Body Cameras for Federal Law-Enforcement Officers. [online] WSJ. Available at: https://www.wsj.com/articles/u-s-mandates-body-cameras-for-federal-law-enforcement-officers-11623110510 [Accessed 9 July 2021].
44. Harte, J. and Mclaughlin, T. (2017). *Emboldened by Trump, some police unions seek to overhaul Obama's reforms*. [online] U.S. Available at: https://www.reuters.com/article/us-usa-police-trump-insight-idUSKBN15E106 [Accessed 9 July 2021].
45. Haynes, D. (2020). Study: Black Americans 3 times more likely to be killed by police. [online] UPI. Available at: https://www.upi.com/Top_News/US/2020/06/24/Study-Black-Americans-3-times-more-likely-to-be-killed-by-police/6121592949925/ [Accessed 8 July 2021].
46. Hemenway, D., Azrael, D., Conner, A. and Miller, M. (2018). Variation in Rates of Fatal Police Shootings across US States: the Role of Firearm Availability. *J Urban Health*, [online] 96(1). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6391295/> [Accessed 27 February 2021].
47. Hillier, W. (2020). What Is a Decision Tree and How Is It Used? [online] CareerFoundry. Available at: https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/ [Accessed 19 May 2021].
48. Hoekstra, M. and Sloan, C. (2020). *Does Race Matter for Police Use of Force? Evidence from 911 Calls*. [online] Available at: https://www.nber.org/papers/w26774 [Accessed 26 February 2021].
49. IBM. (2020). What is Exploratory Data Analysis?.[online] Available at: <https://www.ibm.com/cloud/learn/exploratory-data-analysis> [Accessed 25 December 2020].
50. Ilgen, M.A., Downing, K., Zivin, K., Hoggatt, K.J., Kim, H.M., Ganoczy, D., Austin, K.L., McCarthy, J.F., Patel, J.M. and Valenstein, M. (2009). Exploratory Data Mining Analysis Identifying Subgroups of Patients With Depression Who Are at High Risk for Suicide. *The Journal of Clinical Psychiatry*, [online] 70(11), pp.1495–1500. Available at: https://pubmed.ncbi.nlm.nih.gov/20031094/ [Accessed 9 July 2021].
51. Informit. (2014). *Decision Tree Learning | Predictive Analytics Techniques | InformIT*. [online] Available at: https://www.informit.com/articles/article.aspx?p=2248639&seqNum=11 [Accessed 16 May 2021].
52. Jabbari, G. and Kienle,F. (2019). *Exploratory Data Analysis – An Important Step in Data Science - CAMELOT Blog. [online] CAMELOT Blog.* Available at: <https://blog.camelot-

group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/> [Accessed 1 March 2021].

53. Jennings, J.T. (2017). *Want to reduce fatal police shootings? This policy makes a big difference.* [online] Washington Post. Available at: https://webcache.googleusercontent.com/search?q=cache:KQh-t_yAk34J:https://www.washingtonpost.com/news/monkey-cage/wp/2017/03/14/want-to-reduce-fatal-police-shootings-this-policy-makes-a-big-difference/+&cd=1&hl=en&ct=clnk&gl=sa [Accessed 9 July 2021].
54. JMP. (2021). *Chi-Square Test of Independence*. [online] Jmp.com. Available at: https://www.jmp.com/en_us/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html [Accessed 20 May 2021].
55. Kivisto, A., Ray, B. and Phalen, P. (2017). Firearm Legislation and Fatal Police Shootings in the United States. American Journal of Public Health, [online] 107(7), pp.1068-1075. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463213/> [Accessed 25 December 2020]
56. Kumar, B. (2020). *Comparison of KDD, SEMMA, CRISP-DM, and the Crucial Steps of Business Understanding - Bharani Kumar*. *YouTube*. Available at: https://www.youtube.com/watch?v=RbxdkTixxLo&ab_channel=BharaniKumar [Accessed 18 May 2021].
57. Kumar, P. (2017). *Importance of Test Data*. [online] Prashant Kumar. Available at: https://qanalysisblog.wordpress.com/2017/06/24/importance-of-test-data/#:~:text=Test%20data%20is%20the%20Input%20feed%20for%20Testing%20the%20Application.&text=Test%20Data%20helps%20the%20developers,function%20produces%20some%20expected%20result. [Accessed 20 May 2021].
58. Laughland, O. and Swaine, J.(2016). *Two 'Deadliest' Police Departments In US To Be Investigated In California*. [online] the Guardian. Available at: <https://www.theguardian.com/usnews/2016/dec/22/california-police-kern-county-investigation-kamala-harris> [Accessed 25 December 2020].
59. Lee, D. (2020). Data transformation: a focus on the interpretation. Korean Journal of Anesthesiology, [online] 73(6), pp.503-506. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7714623/> [Accessed 5 July 2021].
60. Lett, E., Asabor, E.N., Corbin, T. and Boatright, D. (2020). Racial inequity in fatal US police shootings, 2015–2020. *Journal of Epidemiology and Community Health*, [online] 75(4), pp.394–397. Available at: https://jech.bmj.com/content/75/4/394.citation-tools [Accessed 25 December 2020].

61. Lopez, G. (2018). *American Police Shoot and Kill Far More People Than Their Peers In Other Countries.* [online] Vox. Available at: <https://www.vox.com/identities/2016/8/13/17938170/uspolice-shootings-gun-violence-homicides> [Accessed 24 December 2020].
62. Malone, K., Omstead, M. and Casey, L.(2020). *Police shootings in 2020: The effect on officers and those they are sworn to protect | CBC News*. [online] CBC. Available at: <https://www.cbc.ca/news/canada/manitoba/police-shootings-2020-yer-review-1.5849788> [Accessed 26 February 2021]
63. Mashinchi, N. (2020). *An Examination of Fatal Force by Police in the US - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/an-examination-of-fatal-force-by-police-in-the-us-db897d97085c [Accessed 19 May 2021].
64. Masters, J. (2019). U.S. Gun Policy: Global Comparisons. [online] Council on Foreign Relations. Available at: https://www.cfr.org/backgrounder/us-gun-policy-global-comparisons [Accessed 11 July 2021].
65. Mayo Clinic. (2019). *Mental Illness - Symptoms And Causes* . [online] Available at: < https://www.mayoclinic.org/diseases-conditions/mental-illness/symptoms-causes/syc-20374968 > [Accessed 24 December 2020].
66. McGroarty, M. (2019). *Mental Health Crises Significant Factor in Police Shootings*. [online] Chicago Policy Review. Available at: <https://chicagopolicyreview.org/2019/01/11/mental-health-crises-significant-factor-in-police-shootings/> [Accessed 27 February 2021].
67. Mellor, J.C, Stone, M.A. and Keane, J. (2018). 'Application of Data Mining to "Big Data" Acquired in Audiology: Principles and Potential.', *Trends in Hearing*, 22, pp. 1-10. Available at: <https://journals.sagepub.com/doi/full/10.1177/2331216518776817> [Accessed 28 February 2021].
68. Mesic, A., Franklin, L., Cansever, A., Potter, F., Sharma, A., Knopov, A. and Siegel, M. (2018) . The Relationship Between Structural Racism and Black-White Disparities in Fatal Police Shootings at the State Level. Journal of the National Medical Association, [online] 110(2), pp.106-116. Available at: <https://www.sciencedirect.com/science/article/pii/S0027968417303206> [Accessed 26 February 2021].
69. Microsoft (2019). *VLOOKUP function*. [online] Microsoft.com. Available at: https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1 [Accessed 12 June 2021].
70. Microsoft (2019). Use ribbon charts in Power BI - Power BI. [online] Microsoft.com. Available at: https://docs.microsoft.com/en-us/power-bi/visuals/desktop-ribbon-charts [Accessed 11 July 2021].

71. Microsoft (2020). Funnel charts - Power BI. [online] Microsoft.com. Available at: https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-funnel-charts [Accessed 11 July 2021].
72. Moore, L. (2020.) *Police Brutality In The United States | Definition, History, Causes, & Examples* . [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/topic/PoliceBrutality-in-the-United-States-2064580> [Accessed 24 December 2020].
73. Nagin, D. (2020). Firearm Availability and Fatal Police Shootings. *The ANNALS of the American Academy of Political and Social Science,* [online] 687(1), pp.49-57. Available at: <https://journals.sagepub.com/doi/pdf/10.1177/0002716219896259> [Accessed 26 February 2021].
74. Narkhede, S. (2018). *Understanding Confusion Matrix - Towards Data Science.* [online] Medium. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 [Accessed 20 May 2021].
75. National Geographic Society (2012). *United States Regions*. [online] National Geographic Society. Available at: https://www.nationalgeographic.org/maps/united-states-regions/ [Accessed 13 June 2021].
76. Nix, J., Campbell, B.A., Byers, E.H. and Alpert, G.P. (2017). A Bird's Eye View of Civilians killed by Police in 2015. *American Society of Criminology*, [online] 16(1), pp. 309-331. Available at: < https://psychology.usu.edu/research/factotum/files/A%20Birds%20Eye%20View%20of%20Civilians%20Killed%20by%20Police%20in%202015.pdf> [Accessed 9 July 2021].
77. Pascual, C. (2018). Tutorial: Understanding Regression Error Metrics in Python. [online] Dataquest. Available at: https://www.dataquest.io/blog/understanding-regression-error-metrics/ [Accessed 11 July 2021].
78. Parekh, R. (2018). *What Is Mental Illness?* . [online] Psychiatry.org. Available at: <https://www.psychiatry.org/patients-families/what-is-mentalillness#:~:text=Mental%20illnesses%20are%20health%20conditions,Mental%20illness%20is%20common.> [Accessed 24 December 2020].
79. Parr-Rud, O. (2014) Business Analytics Using SAS® Enterprise Guide® and SAS® Enterprise Miner®: A Beginner's Guide. Cary: SAS Institute Inc.
80. Patil, Y., Kumari, S. and Jeble,S. (2016). 'Role of big data and predictive analytics', International Journal of Automation and Logistics, 2(4) pp. 307-331
81. Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects - KDnuggets*. [online] KDnuggets. Available at: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html [Accessed 10 July 2021].

82. Pieterse, A.L., Todd, N.R., Neville, H.A. and Carter, R.T. (2012). Perceived racism and mental health among Black American adults: A meta-analytic review. *Journal of Counseling Psychology,* 59(1), pp.1–9.
83. Portland Police Bureau. (2018). Dashboard Walkthrough | Officer Involved Shootings | The City of Portland, Oregon. [online] Portlandoregon.gov. Available at: https://www.portlandoregon.gov/police/article/684064 [Accessed 11 July 2021].
84. PredictiveAnalyticsToday. (2016). *SAS Enterprise Miner.* [online] Available at: <https://www.predictiveanalyticstoday.com/sas-enterprise-miner/> [Accessed 28 February 2021].
85. Pyvovar, N. (2019). *Data Science project management methodologies - DataDrivenInvestor*. [online] Medium. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb [Accessed 16 May 2021].
86. Qaiser, H. and Shafique, U., 2014. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research,* [online] 12(1), pp.217-222. Available at: <https://www.researchgate.net/project/A-Comparative-Study-of-Data-Mining-Process-Models-KDD-CRISP-DM-and-SEMMA> [Accessed 16 May 2021].
87. Rodrigues, I. (2020). CRISP-DM methodology leader in data mining and big data. [online] Medium. Available at: <https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781> [Accessed 28 February 2021].
88. Ross, C. (2015). A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011–2014. [online] 10(11). Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141854&utm_source=Iterable&utm_medium=email&utm_campaign=newsletter-20200605> [Accessed 27 February 2021].
89. Ruess, S. (2019). Situational Context of Police Use of Deadly Force: a Comparison of Black and White Subjects of Fatal Police Shootings. [online] Portland: Portland State University, pp.4,20-28,43,69-76. Available at: <https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=6211&context=open_access_etds> [Accessed 1 March 2021].
90. RWJF. (2017). *Discrimination in America: Experiences and Views.* [online] Available at: <https://www.rwjf.org/en/library/research/2017/10/discrimination-in-america--experiences-and-views.html> [Accessed 26 February 2021].
91. Ryan, A.R. (2019). Lethal Use of Force: Insights into Mental Illness. [online] New York: John Jay College of Criminal Justice, pp.1-5,36-60. Available at: <https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1101&context=jj_etds> [Accessed 1 March 2021].

92. Sarma, K.S. (2017) *Predictive Modelling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition.* 3rd edn. Cary: SAS Institute Inc.
93. SAS Institute Inc. (2003) *Data Mining Using SAS Enterprise Miner: A Case Study Approach, Second Edition.* Cary: SAS Institute Inc.
94. SAS Support. (2014). *47186 - The Decision Tree node displays "Missing Values Only" even when data contains no missing values*. [online] Available at: https://support.sas.com/kb/47/186.html [Accessed 8 July 2021].
95. SAS. (2016). *SAS Studio Help Center*. [online] Sas.com. Available at: https://support.sas.com/software/products/sas-studio/faq/SASStudio_whatis.htm [Accessed 15 Jun. 2021].
96. SAS. (2021). *SAS Help Center.* [online] Sas.com. Available at: https://documentation.sas.com/doc/en/emref/14.3/n1jqzz8cssr9m2n1ktx2iyv87q56.htm#p01xxszbcpburun1tl4xo7g7ggqn [Accessed 19 May 2021].
97. SAS. (2021). *SAS Help Center*. [online] Sas.com. Available at: https://documentation.sas.com/doc/en/emref/14.3/p1rk96oj5sk2tyn1esay58oha0o3.htm#:~:text=You%20use%20the%20File%20Import,the%20SAS%20Enterprise%20Miner%20toolbar. [Accessed 2 June 2021].
98. SAS . (2021). SAS Help Center. [online] Sas.com. Available at: https://documentation.sas.com/doc/en/emref/14.3/p0ldp4l9cnob3gn1dytmimkunaa6.htm#p035bdkivtqrwpn1oywdrtwibgiu [Accessed 18 June 2021].
99. Scout, C. (2015). How Mental Illness Affects Police Shooting Fatalities - International Bipolar Foundation. [online] International Bipolar Foundation. Available at: <https://ibpf.org/how-mental-illness-affects-police-shooting-fatalities/> [Accessed 25 February 2021].
100. Schulberg, J.(2020). *Exploratory Data Analysis of U.S. Police-caused Fatalities*. [online] Medium. Available at: <https://towardsdatascience.com/exploratory-data-analysis-of-u-s-police-caused-fatalities-fce47a2b7198> [Accessed 1 March 2021].
101. Schwartz, G. and Jahn, J. (2020). Mapping fatal police violence across U.S. metropolitan areas: Overall rates and racial/ethnic inequities, 2013-2017. [online] 15(6). Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229686> [Accessed 27 February 2021].
102. Siegel, M. (2020). *RACIAL DISPARITIES IN FATAL POLICE SHOOTINGS: AN EMPIRICAL ANALYSIS INFORMED BY CRITICAL RACE THEORY*. [online] Boston: Boston University School of Public Health, pp.1069-1092. Available at: <https://www.bu.edu/bulawreview/files/2020/05/10-SIEGEL.pdf> [Accessed 26 February 2021].

103. Sinyangwe, S., McKesson, D. and Elzie, J. (2021). Mapping Police Violence - Planning Team. [online] Mapping Police Violence. Available at: <https://mappingpoliceviolence.org/planning-team> [Accessed 18 June 2021].
104. Sharma, A. (2021). *Decision Tree Algorithm for Classification: Machine Learning 101*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/#:~:text=The%20goal%20of%20this%20algorithm,internal%20node%20of%20the%20tree. [Accessed 19 May 2021].
105. Sholtis, B. (2020). *Family Mourns Man with Mental Illness Killed by Police and Calls for Change.* [online] khn.org. Available at: https://khn.org/news/police-killing-mental-illness-emergency-crisis-care-pennsylvania/ [Accessed 25 February 2020]
106. Smart Vision Europe. (2020). *Crisp DM methodology - Smart Vision Europe*. [online] Available at: <https://www.sv-europe.com/crisp-dm-methodology/> [Accessed 25 February 2021].
107. Statista. (2020). State Population: United States 2020 | Statista. [online] Available at: https://www.statista.com/statistics/183497/population-in-the-federal-states-of-the-us/ [Accessed 11 July 2021].
108. Streeter, S. (2019). Lethal Force in Black and White: Assessing Racial Disparities in the Circumstances of Police Killings. The Journal of Politics, 81(3), pp.1124–1132.
109. Swaine, J., Laughland, O., Lartey, J. and McCarthy, C. (2015). *Young Black Men Killed By US Police At Highest Rate In Year Of 1,134 Deaths.* [online] the Guardian. Available at: <https://www.theguardian.com/us-news/2015/dec/31/the-counted-police-killings-2015-youngblack-men> [Accessed 25 December 2020].
110. Szumilas, M. (2010). Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, [online] 19(3), pp.227–9. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/ [Accessed 11 July 2021].
111. Taylor, C. (2018). Fight the Power: African Americans and the Long History of Police Brutality in New York City. 1st ed. New York: NYU Press, pp.1-2.
112. Thanda, A. (2020). *What is Logistic Regression? A Beginner's Guide [2021]*. [online] CareerFoundry. Available at: https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/ [Accessed 7 February 2021].
113. Thomlinson, R. (2020). *Understanding racism and racial discrimination: Recognizing & responding to the problem in Canada - Rubin Thomlinson*. [online] rubinthomlinson.com. Available at:

https://rubinthomlinson.com/understanding-racism-and-racial-discrimination-recognizing-responding-to-the-problem-in-canada/ [Accessed 10 July 2021].

114. Thurimella, A. and Padmaja, T. (2014). Chapter 2 - Economic Models and Value-Based Approaches for Product Line Architectures. [online] pp.11-36. Available at: <https://www.sciencedirect.com/science/article/pii/B9780124104648000027> [Accessed 27 February 2021].
115. Trinidad, C. (2020). Regression Analysis. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/ [Accessed 18 May 2021].
116. Trinidad, C. (2020). *Skewness*. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/knowledge/other/skewness/ [Accessed 11 June 2021].
117. Vadakkanmarveettil, J. (2020). *Big Data Analytics To Beat Crime | Jigsaw Academy*. [online] Jigsaw Academy. Available at: https://www.jigsawacademy.com/big-data-analytics-to-beat-crime/ [Accessed 11 June 2021].
118. Vadapalli, P. (2020). *Random Forest Vs Decision Tree: Difference Between Random Forest and Decision Tree | upGrad blog*. [online] upGrad blog. Available at: https://www.upgrad.com/blog/random-forest-vs-decision-tree/#:~:text=A%20decision%20tree%20combines%20some,forest%20model%20needs%20rigorous%20training. [Accessed 20 May 2021].
119. Varga, S. (2018). *Exploring a Data Set with Simple Statistics in Power BI – Data Inspirations*. [online] Blog.datainspirations.com. Available at: <http://blog.datainspirations.com/2018/03/18/exploring-a-data-set-with-simple-statistics-in-power-bi/> [Accessed 1 March 2021].
120. Wang, Y. and Fan, Y. (2021). US Fatal Police Shooting Analysis and Prediction. [online]. Available at: https://arxiv.org/pdf/2106.15298.pdf [Accessed 10 July 2021].
121. Waseem, M. (2019). *Linear Regression for Machine Learning | Intro to ML Algorithms | Edureka*. [online] Edureka. Available at: https://www.edureka.co/blog/linear-regression-for-machine-learning/ [Accessed 16 May 2021].
122. WashingtonPost (2015). washingtonpost/data-police-shootings. [online] GitHub. Available at: https://github.com/washingtonpost/data-police-shootings [Accessed 18 December 2020].
123. White, M.D. (2002). Identifying situational predictors of police shootings using multivariate analysis. *Policing: An International Journal of Police Strategies & Management*, 25(4), pp.726–751.
124. Wood, G., Tyler, T.R. and Papachristos, A.V. (2020). Procedural justice training reduces police use of force and complaints against officers. *PNAS*, [online] 117(18), pp. 9815-921. Available at: < https://www.pnas.org/content/pnas/117/18/9815.full.pdf> [Accessed 12 July 2021].

125. Wright, N. (2019*). Everything you ever wanted to know about Microsoft Power BI*. [online] Nigel Frank. Available at: <https://www.nigelfrank.com/blog/everything-you-ever-wanted-to-know-about-microsoft-power-bi/> [Accessed 28 February 2021].

126. Writh, R. and Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.* [ebook] Ulm, pp.1-11. Available at: <http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf> [Accessed 28 February 2021].

127. Wu, S. (2020). *What are the best metrics to evaluate your regression model?* [online] Medium. Available at: <https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b> [Accessed 28 February 2021].

128. Yerpude, P. and Gudur, V. (2017). Predictive Modelling of Crime Dataset Using Data Mining. *International Journal of Data Mining & Knowledge Management Process*, [online] 7(4), pp.43–58. Available at: <https://aircconline.com/ijdkp/V7N4/7417ijdkp04.pdf> [Accessed 8 July 2021].

8.2 Appendices

8.2.1 FYP Poster

ANALYSIS ON THE INFLUENCE OF MENTAL ILLNESS AND RACIAL DISCRIMINATION ON FATAL POLICE SHOOTINGS

Basmah Zahid
[redacted]
B.Sc (Hons) in Computer Science with Specialism in Data Analytics - [redacted]
Supervised by: [redacted]
Second Marker [redacted]

Introduction

The world sees daily news reports of innocent civilians shot by the law enforcement and questions why it is happening? People at home wonder if they will be next, persons of color in America are worried about their children and family in case they encounter these situations with officers. But is it really something racially motivated on behalf of the law enforcement?

By using fatal police shooting data with exploratory data analysis and predictive modelling, these questions can be answered, this research project will use various tools and techniques for the building of dashboard and models to examine what factors are present at the time of these deadly encounters and whether racial discrimination or mental illness of a victim correlate with them.

Objective

1. To investigate the key factors to analyze the influence of mental illness and racial discrimination on fatal police shootings.
2. To analyze the influence of mental illness and racial discrimination on fatal police shooting using explanatory data analysis techniques.
3. To evaluate the performance of analysis results using evaluation metrics such as R-Squared (R2), Mean Squared Error (M.S.E), Mean Absolute Error (M.A.E) or Misclassification Rate (MISC).
4. To develop a dashboard to visualize the analysis results.

Methodology and Techniques

CRISP-DM

Exploratory Data Analysis

Predictive Modelling

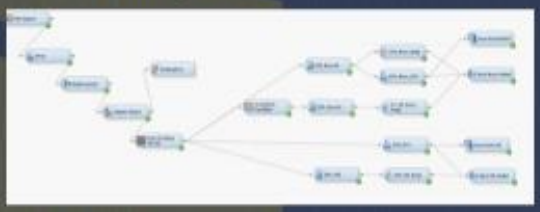
Models

Decision Trees

Logistic Regression

Implementation

- Data Preprocessing, Modelling and Evaluation with SAS Studio and Enterprise Miner
- Prediction Model Diagram in SAS EM



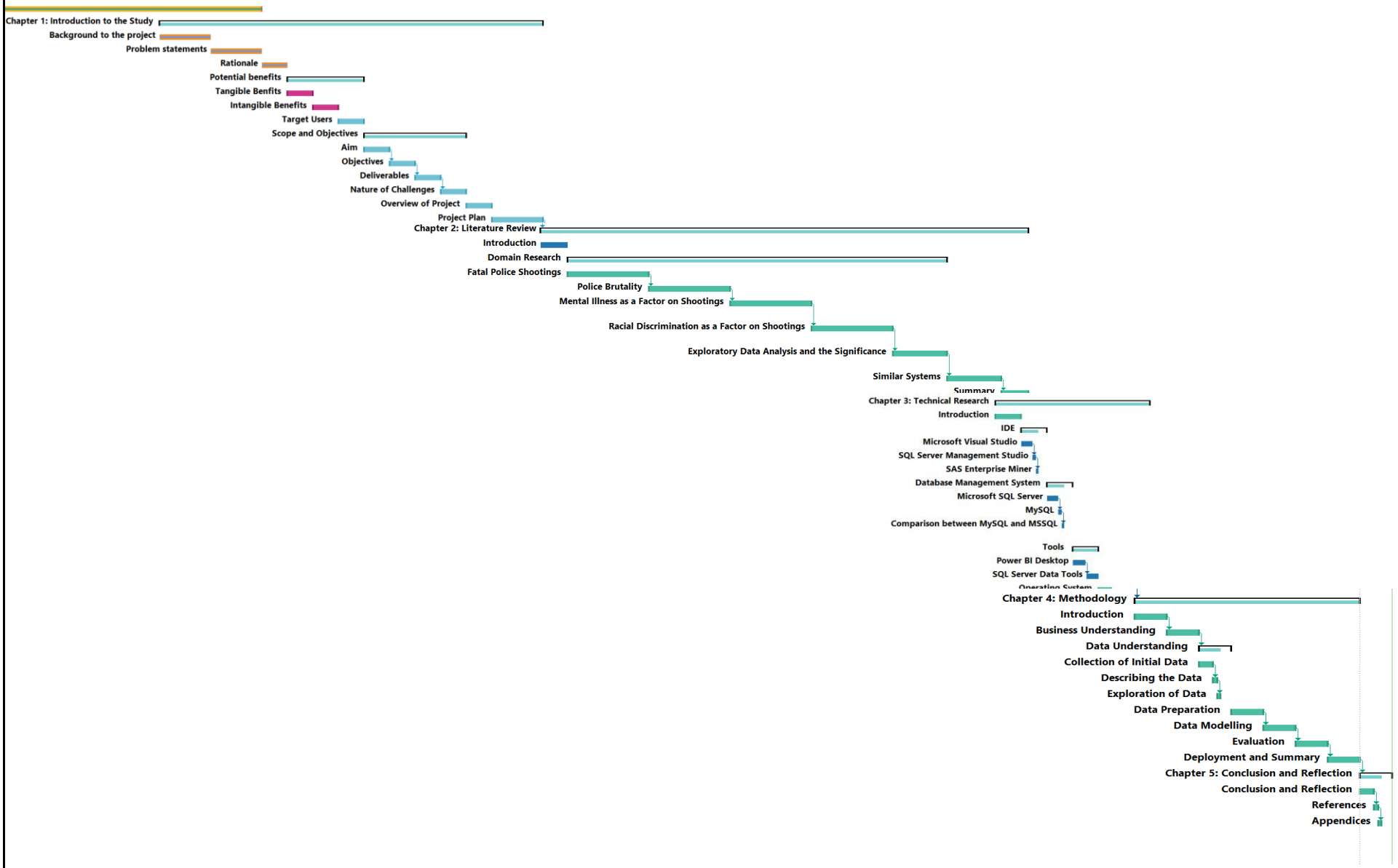
- Dashboard Visualization in PowerBI

Conclusion

Evaluation metrics misclassification rate and average squared error were used to assess the two models from which decision tree was found to be the best model to carry out prediction for both targets: signs of mental illness and race. The results presented that the likelihood of a person suffering from mental illness was predicted to be low but not impossible while in terms of race, white victims were most predicted to be shot in case of fatal encounters however in certain events, other races such as black, Asian and native victims were more likely to be shot. The variables that correlate most with race were age, region, signs of mental illness, month, armed category and body camera while for mental illness; race, threat level, flee, body camera, region, armed category, month and day were most critical.

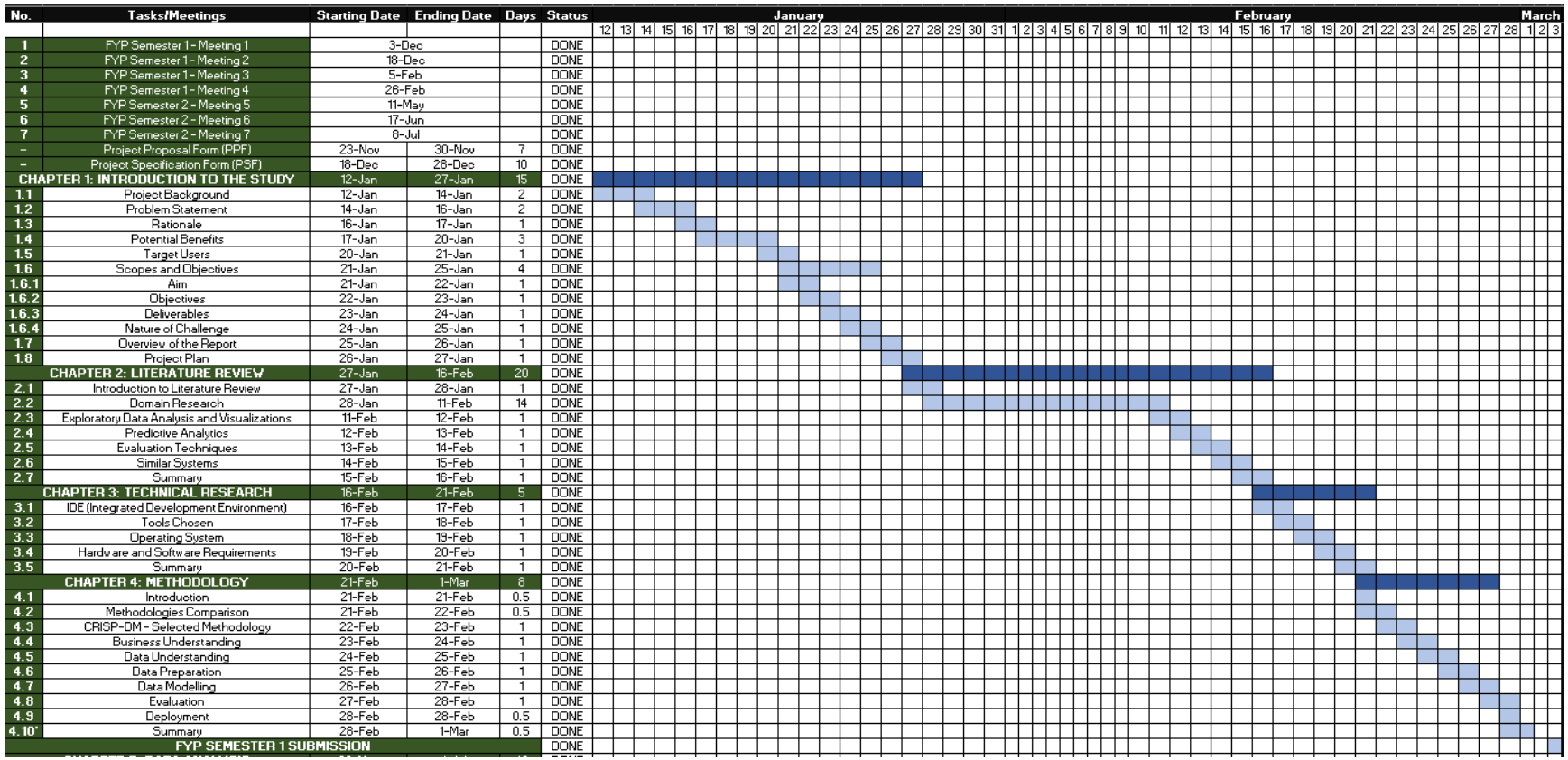
8.2.2 Gantt Chart

FYP Semester 1 – this is the initial Gantt chart made in the semester. All the new changes have been reflected in the next Gantt chart.



Updated Gantt Chart – FYP Semester 2

This is the whole Gantt chart for both semesters and due to page limitations, it needed to be divided into two screenshots.



No.	Tasks/Meetings	Starting Date	Ending Date	Days	Status
1	FYP Semester 1 - Meeting 1	3-Dec			DONE
2	FYP Semester 1 - Meeting 2	18-Dec			DONE
3	FYP Semester 1 - Meeting 3	5-Feb			DONE
4	FYP Semester 1 - Meeting 4	26-Feb			DONE
5	FYP Semester 2 - Meeting 5	11-May			DONE
6	FYP Semester 2 - Meeting 6	17-Jun			DONE
7	FYP Semester 2 - Meeting 7	8-Jul			DONE
-	Project Proposal Form (PPF)	23-Nov	30-Nov	7	DONE
-	Project Specification Form (PSF)	18-Dec	28-Dec	10	DONE
CHAPTER 1: INTRODUCTION TO THE STUDY		12-Jan	27-Jan	15	DONE
1.1	Project Background	12-Jan	14-Jan	2	DONE
1.2	Problem Statement	14-Jan	16-Jan	2	DONE
1.3	Rationale	16-Jan	17-Jan	1	DONE
1.4	Potential Benefits	17-Jan	20-Jan	3	DONE
1.5	Target Users	20-Jan	21-Jan	1	DONE
1.6	Scopes and Objectives	21-Jan	25-Jan	4	DONE
1.6.1	Aim	21-Jan	22-Jan	1	DONE
1.6.2	Objectives	22-Jan	23-Jan	1	DONE
1.6.3	Deliverables	23-Jan	24-Jan	1	DONE
1.6.4	Nature of Challenge	24-Jan	25-Jan	1	DONE
1.7	Overview of the Report	25-Jan	26-Jan	1	DONE
1.8	Project Plan	26-Jan	27-Jan	1	DONE
CHAPTER 2: LITERATURE REVIEW		27-Jan	16-Feb	20	DONE
2.1	Introduction to Literature Review	27-Jan	28-Jan	1	DONE
2.2	Domain Research	28-Jan	11-Feb	14	DONE
2.3	Exploratory Data Analysis and Visualizations	11-Feb	12-Feb	1	DONE
2.4	Predictive Analytics	12-Feb	13-Feb	1	DONE
2.5	Evaluation Techniques	13-Feb	14-Feb	1	DONE
2.6	Similar Systems	14-Feb	15-Feb	1	DONE
2.7	Summary	15-Feb	16-Feb	1	DONE
CHAPTER 3: TECHNICAL RESEARCH		16-Feb	21-Feb	5	DONE
3.1	IDE (Integrated Development Environment)	16-Feb	17-Feb	1	DONE
3.2	Tools Chosen	17-Feb	18-Feb	1	DONE
3.3	Operating System	18-Feb	19-Feb	1	DONE
3.4	Hardware and Software Requirements	19-Feb	20-Feb	1	DONE
3.5	Summary	20-Feb	21-Feb	1	DONE
CHAPTER 4: METHODOLOGY		21-Feb	1-Mar	8	DONE
4.1	Introduction	21-Feb	21-Feb	0.5	DONE
4.2	Methodologies Comparison	21-Feb	22-Feb	0.5	DONE
4.3	CRISP-DM - Selected Methodology	22-Feb	23-Feb	1	DONE
4.4	Business Understanding	23-Feb	24-Feb	1	DONE
4.5	Data Understanding	24-Feb	25-Feb	1	DONE
4.6	Data Preparation	25-Feb	26-Feb	1	DONE
4.7	Data Modelling	26-Feb	27-Feb	1	DONE
4.8	Evaluation	27-Feb	28-Feb	1	DONE
4.9	Deployment	28-Feb	28-Feb	0.5	DONE
4.10'	Summary	28-Feb	1-Mar	0.5	DONE
FYP SEMESTER 1 SUBMISSION					DONE

No.	Tasks/Meetings	Starting Date	Ending Date	Days	Status
	CHAPTER 5: DATA ANALYSIS	20-May	1-Jul	42	DONE
5.1	Introduction	20-May	21-May	1	DONE
5.2	Initial Data Exploration	21-May	31-May	10	DONE
5.3	Data Cleaning	31-May	15-Jun	15	DONE
5.4	Data Visualization	15-Jun	18-Jun	3	DONE
5.5	Modelling/Data-driven product	18-Jun	28-Jun	10	DONE
5.6	Reports and Dashboard	28-Jun	30-Jun	2	DONE
5.7	Summary	30-Jun	1-Jul	1	DONE
	CHAPTER 6: RESULTS AND DISCUSSION	1-Jul	5-Jul	4	DONE
6.1	Introduction	1-Jul	2-Jul	1	DONE
6.2	Model Evaluation	2-Jul	4-Jul	2	DONE
6.3	Discussion	4-Jul	5-Jul	1	DONE
	CHAPTER 7: CONCLUSION AND REFLECTION	5-Jul	7-Jul	2	DONE
7.1	Conclusion	5-Jul	5-Jul	0.5	DONE
7.2	Reflection	5-Jul	6-Jul	0.5	DONE
7.3	Challenges	6-Jul	6-Jul	0.5	DONE
7.4	Limitations	6-Jul	7-Jul	0.5	DONE
7.5	Future Improvements	7-Jul	7-Jul	0.5	DONE
	CHAPTER 8: REFERENCES AND APPENDICES	18-Dec	12-Jul	159	DONE
8.1	References	18-Dec	12-Jul	159	DONE
8.2	Appendices	18-Dec	12-Jul	159	DONE
8.3	CHECKING OF FYP DOCUMENT	18-Dec	12-Jul	159	DONE