# Fatal Encounters

## Executive Summary

This report consists of the data cleaning, analysis and inference of a dataset obtained from Fatal Encounters. It consists of 31,498 deceased individuals due to encounters with the police in the United States of America between \_\_ and \_\_\_

## Data Preparation

### Data Cleaning

94.44% of the dataset contained null values. Looking more into the variables themselves most of these null values are a result of redundant variables. Hence, the following were dropped for the sake of dimensionality that is covered later in the analysis.

"...34","...33", "Unique ID formula", "URL Temp", "Description Temp", "UID Temporary", "Name Temporary", "URL of image (PLS NO HOTLINKS)", "Race with imputations", "Imputation probability", "Unique identifier (redundant)", "Supporting document link", "Dispositions/Exclusions INTERNAL USE, NOT FOR ANALYSIS", "Unique ID", "Name", "Longitude", "Latitude", "Full Address", "Location of death (county)", "Location of death (zip code)", "Location of death (city)", "Location of injury (address)" , "Agency or agencies involved", "Brief description", "Intended use of force (Developing)"

A graph with blue dots and white text

Description automatically generated

Furthermore, it was also observed that due to the nature of the types of variables almost all are nominal categorical variables with multiple levels. However, some of these variables could be factorized and be put on the ordinal scale and the levels themselves.

That being dealing with the nulls values within these levels were the next hurdle. To be respectful to the amount of effort and time it has taken to gather the data. This analysis did not remove any null values but rather changed the nature of the null values themselves.

### Target Encoding Nominal Variables & Dealing with Null Values

It was observed that some variables contained as high as 268 and low as 4 levels. In order for the unsupervised machine learning model to work without having to compute large dimensions after one-hot encoding, the option to target encode these variables were taken.

Let’s take an example, below is a sampled list of alleged weapons that were on the deceased individual. The way how the types of weapons are either classified as either edged weapons or firearms were most helpful to reduce the levels, but the null values were instead encoded as “Uncertain”, since it is neither the absence nor presence of a weapon itself but the inability to have a clear indication of either. Such was the methodology used to tone down the levels for each of the other nominal variables. The full visualizations of the target encoded variables are present in the appendix.

|  |  |
| --- | --- |
| **Type of Alleged Weapon** | **Count** |
| Null | 14,421 |
| None | 6,125 |
| Firearm/Handgun | 3,585 |
| Firearm/Gun | 2,914 |
| Edged weapon/Knife | 1,419 |
| Firearm/Rifle | 646 |
| Firearm/Shotgun | 472 |
| Uncertain | 468 |
| Firearm/Pellet gun | 260 |
| Firearm/Replica firearm | 155 |
| Edged weapon/Machete | 87 |

Moving on to the only numerical variable in this dataset, age, when observing the variables’ distribution, it is evident that it is slightly skewed to the right.

A graph of a number of age

Description automatically generated

When one thinks about the age of a person one would think it would be a whole number but, in this case, it is best left as ratio data. Why? Because context matters regarding these individuals. When looking at the granularity of the data in the opposing tail ends of the distribution – the youngest, 0.08 years old, had the following description,

“*Timothy Gilles sped away from St. Cloud police and crashing into a vehicle. The crash killed 30-year-old Timothy Casanova of Starbuck. His passenger, 22-year-old Megan Thomsen of Glenwood, was pregnant and gave birth after the crash to a baby that died days after his emergency delivery. Thomsen was also injured. He was sentenced to 432 months in prison.*”

And the oldest, 107 years old, had the following description,

“*Police arrived at a residence in response to reports of a disturbance. 107 y.o. Monroe Isadore allegedly confronted them with a handgun and retreated into a bedroom, firing on them when they attempted to enter. SWAT officers arrived as backup and, after failed negotiations, released gas into the room and broke down the door. Isadore fired on them as they entered, and they returned fire, killing him.”*

But what of the missing values, before removing or imputing these variables let’s look at how the distribution is affected if we were to make these changes.

A group of different colored graphs

Description automatically generated with medium confidence

The closest to the original distribution is noted to be imputing the mean over the median and this was what was chosen going forward into the analysis.

## Unsupervised Machine Learning

### Data Preprocessing

Looking at the distribution of the gender type it is evident that Males make up the most compared to Other, Transgenders, and Unknowns. In this case, these groups were grouped as Others.

A graph with a red bar

Description automatically generated

The main issue that this project will face when cluster analysis is conducted is the number of samples present in the dataset. For the analysis to effectively work the option of dimensionality reduction is a must to the one-hot encoding that will be done on some of these variables. There is also another option, some variables that we could convert from a nominal scale to an ordinal scale. The answer isn’t exactly yes, but observing some of the variables namely, highest level of force, armed/unarmed, fleeing or not feeling, and foreknowledge of mental illness it could be done.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Variable Name** | | | | **Ordinal Scale Level** |
| **Group Name** | **Highest Level of Force** | **Armed or Unarmed** | **Fleeing or Not Fleeing** | **Foreknowledge of Mental Illness** |
| Medical Emergency | Unarmed | No | No | 0 |
| Undetermined |
| Other |
| Less than lethal force | Uncertain | Uncertain | Uncertain | 1 |
| Gunshot | Armed | Yes | Yes | 2 |
| Vehicle |
| Tasered |

The rest of the variables were one hot encoded due to cluster analysis being limited to numerical data points in the vector space.

### Multicollinearity

With all the variables being numerical (nominal and ordinal) the next step is to check for trends in the data, such as multicollinearity which could affect the results of the cluster analysis. One way to mitigate this is to use dimension reduction.

A close-up of a grid

Description automatically generated

### Dimension Reduction – Principal Component Analysis (PCA)

The purpose of using PCA before the cluster analysis is for the sake of performing the analysis on lower dimensional data is more efficient, to eliminate the multicollinearity between the variables, to find latent variables, and to remove the noise.

A graph with numbers and a line

Description automatically generated

A graph with blue lines and words

Description automatically generated

### K-means Clustering

### Partitioning Around Medoids (PAM)