Gun Violence Analysis

Gun violence is a significant public health concern in the United States, with a lack of easily accessible detailed data. However, a comprehensive dataset containing information on over 260,000 gun violence incidents in the US between January 2013 and March 2018 is now available in CSV format from gunviolencearchive.org.

This data has the potential to enable data scientists and statisticians to study gun violence trends and make informed predictions about future patterns, as well as to understand the underlying factors and trends in gun violence, crucial for developing effective strategies to reduce its impact.

The information provided is based on the following source:

1. The Gun Violence Archive (GVA):
https://www.kaggle.com/datasets/jameslko/gun-violence-data

Dataset Variables

Column	Description
incident_id	The ID number for each incident
date	Date (xxxx-xx-xx)
state	USA State where the incident took place
n_participants	The Number of Participants involved in the incident
n_killed	The Number of people killed in this incident
n_injured	The Number of people injured in this incident
victims	The number of people involved in the incident that are victims
suspects	The number of people involved in the incident that are the suspects
adult	Was an Adult involved in this incident
teen	Was a Teen involved in this incident
child	Was a Child involved in this incident
killed	Qualitative
injured	Were people killed during this incident
arrested	Was someone arrested as a result of this incident
unharmed	Did anyone leave unharmed from this incident

Gun Violence Analysis

The correlation heatmap visually represents the strength and direction of the relationships between variables.

'Suspects' Correlation: The negative correlations of 'suspects' with 'n_killed', 'n_injured', and 'victims' make sense. A suspect would not be expected to be the person killed or injured, and there would be a negative relationship between being a suspect and being a victim. This is reflected in the heatmap by the negative values, indicating an inverse relationship between these variables

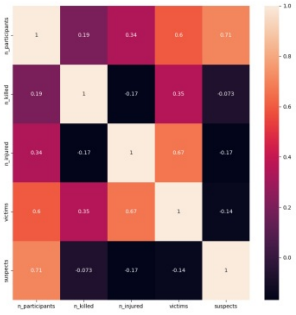
.

'n_killed' and 'n_injured' Correlation: The correlation of -0.16 between 'n_killed' and 'n_injured' makes sense as it reflects that one cannot be both killed and injured at the same time. The negative correlation is indicative of this mutually exclusive relationship, which is captured by the heatmap

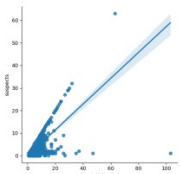
.

'Victims', 'Suspects', and 'n_participants' Correlation: The strong correlation of 'victims' and 'suspects' with 'n_participants' makes sense as it reflects how the number of participants involved in an incident is related to the number of victims and suspects. This correlation is evident in the heatmap, indicating a positive relationship between these variables

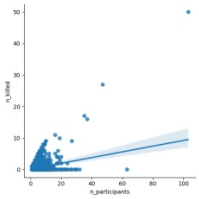
.



Gun Violence Analysis



By employing 'n_participants' as our dependent variable, we conducted an in-depth analysis of correlations and linear trends in relation to 'suspects,' 'victims,' and 'n_killed.'

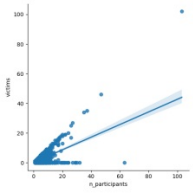


The correlation coefficients provided indicate the strength and direction of the linear relationships between the variables. Here's what each correlation means:

1. The correlation of 0.19 between 'n_participants' and 'n_killed' suggests a weak positive linear relationship. This means that the relationship is not very strong.

2. The correlation of 0.6 between 'n_participants' and 'victims' indicates a moderately strong positive linear relationship. This suggests that as the number of participants increases, the overall number of victims (including both killed and non-killed) also tends to increase.

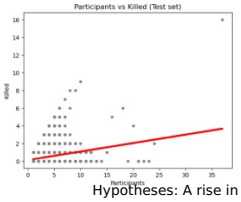
3. The correlation of 0.71 between 'n_participants' and 'suspects' sugg..

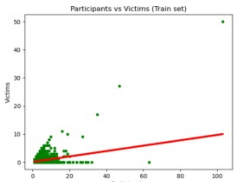


At this stage of our analysis, we can articulate our hypotheses:

1. A rise in the number of participants does not necessarily correspond to a proportional increase in deaths

Gun Violence Analysis





Hypotheses: A rise in the number of participants does not necessarily correspond to a proportional increase in deaths

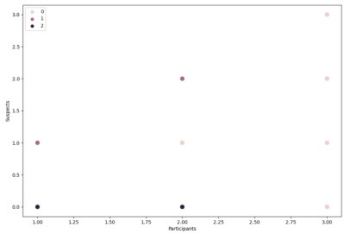
We opted for a simple linear regression given the consideration of a single independent variable. Subsequently, we divided our dataset into a training set and a test set, utilizing the training set for model training.

Statistical metrics, including the slope, were computed to evaluate model accuracy and compare actual data points with predictions. The final step involved assessing accuracy on the testing set.

The linear red lines highlight our slight correlation between the two variables. We can still see many data points that are free from the line, and ultimately, help to support our hypthesis.

Hypothesis: An escalation in the number of participants is expected to coincide with a rise in the number of suspects.

The goal is to categorize data points with higher similarity and distinguish those that exhibit dissimilarity through Cluster Analysis. Employing the Elbow Technique assisted in determining the ideal number of clusters, which, in our scenario, was identified as 3. Subsequently, the model allocated data points to each cluster based on their distances from the respective centroids.

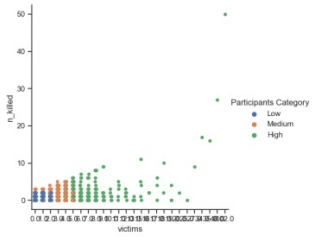


Gun Violence Analysis

In conclusion,

Our analysis, specifically using the Correlation Matrix and dependent variable, did not yield new information. At a glance, we can understand why 'Participants' and 'Victims' would correlate quite easily, correct? The process, albeit necessary, left us wanting more. Ultimately, our hypothesis was simple to both create and prove true.

The Machine Learning Analysis followed a similar trend, as it did not provide us with any new analysis we could not already have pulled from the data prior.



Limitations:

The dataset provided daily accounts of each incident that was logged. This is fantastic to have, but we were unable to use it in our Time-Series Analysis in a meaningful way, as the required codes, and processes, needed totals in a year-over-year rather than daily.

Once outliers were removed, the Machine Learning Scatterplot because difficult to produce. We had several thousand (around 18 thousand to be exact) instances, but most fell within a 3-digit range. Perhaps a more expansive plot could have provide..

Recommendation:

Further analysis, and having a grasp of the required information for our 3 main forms of analysis (Correlation Matrix, Machine Learning..