**Project Proposal**

The Pursuit: Unraveling Homicides

**INDEX**

**Introduction**……………………………………………………………………………….. 3

1. Project overview …………………………………………………………………….3
2. Context of our data ………………………………………………………………….3

**Question of interest** ………………………………………………………………………..4

**Data Processing and analysis** ……………………………………………………………..5

1. Data set descriptions …………………………………………………………… ….5
2. Data processing tasks ………………………………………………………………6
3. Data analysis ………………………………………………………………………..6

**Expected findings**…………………………………………………………………………..9

**Project Timeline** …………………………………………………………………………..11

# **Introduction**

***Project overview***

In order to better understand the context, trends, and underlying causes of homicide in a particular location or across numerous regions, our study intends to provide a thorough analysis of homicide report data. This analysis includes the investigation of a number of homicide-related issues, including victim characteristics, crime scenes, suspects' intentions, and the influence of societal, economic, and environmental factors.

***Context for your data***

Homicide is a serious social problem with ramifications across society. It affects not only the criminal justice system but also critically affects social and public health. The focus of our study is to look deeply into homicide report data, which normally contains details about each homicide case, including the date, time, location, victim's characteristics, suspect information, cause of death, and, in some circumstances, the crime's motivation.

We will try to answer a variety of questions in hopes of understanding if it is possible to prevent or solve crime in a more efficient manner. Through temporal/spatial analysis, we can learn more about the data given and successfully answer our questions of interest. Our dataset has range from 1980 to 2014, giving us almost 35 years of homicide data to work with. We hope to be able to analyze hot spots, predict trends, and other such data analysis/processing tasks.

***Why is this interesting? Why should we care about it?***

Public safety: It's critical for law enforcement and public safety organizations to comprehend homicide trends and patterns. It aids in the development of crime prevention plans and more efficient resource allocation.

Impact on social life: Homicides have a profound effect on communities, not only harming the victims and their families but also instilling a sense of fear and insecurity. We can spot high-risk areas and populations through data analysis and take preventative action.

Policy Development: To develop and improve policies pertaining to gun control, mental health care, education, and other areas that may help reduce homicides, policymakers can make use of the findings from this analysis.

Resource Allocation: Based on data-driven insights, resource allocation for victim support, neighborhood activities, and law enforcement is more efficient.

Academic Research: The information helps researchers better understand the underlying factors that lead to homicides in the fields of criminology, sociology, and public health.

Future Homicide Prevention: We can endeavor to put preventive measures into place to lower the total number of killings by identifying common elements and patterns in homicide instances.

In conclusion, our project's analysis of homicide report data is crucial because it can help us better understand the causes of homicides and guide the development of interventions and policies that would lessen this senseless loss of life and its related social and financial costs.

## ***Questions of Interest***

Questions we would like to answer in the context we described using our data:

* What are the primary demographic traits of homicide perpetrators and victims, and how have they changed over time? What was the most affected segment of the population?
* What are the most frequent causes/weapons used in homicides, and how do they alter depending on geography and demographics? What are the regions or cities that have the most homicides and which causes are the most prevalent in each state?
* Are there times when the number of homicides significantly rises or falls, and if so, what may have caused these changes—such as modifications to policy, changes in the economy, or adjustments to law enforcement tactics? Would any economic suffering or recession period contribute a significant effective relation to the frequency of homicide? Does the frequency vary based on season?
* Correlation between perpetrator age and gender vs. victim age. Which generation would be the most affected by homicides? Is there any generation gap between the victim and the perpetrator? Can we decrease this crime rate? (Analyzing ages and age groups)
* Does the relationship between the victim and the perpetrator matter? Is it more common that the two sides know each other or don’t?
* For the crimes that were solved, which agency types were the most effective? Which states’ agencies were the best at solving the crimes? Which were the worst?
* What kind of weapons did certain age groups prefer to use? What kind of weapons did the different genders given use?
* What makes some areas more likely to have homicides, and can we use predictive analysis to find out which neighborhoods or city blocks are most at risk?

# **Data Processing and Analysis**

## ***Dataset Description***

The dataset "Homicide Reports, 1980-2014" contains information about homicides that occurred in the United States, as the name implies. As of now, this is the most comprehensive source of homicide data in the United States. This massive dataset includes homicide records from the FBI's Supplementary Homicide Report spanning the years 1976 to 2014, as well as supplementary data collected from over 22,000 unreported homicides made available through the Freedom of Information Act.

This dataset contains critical information such as the age, race, gender, and ethnicity of both victims and perpetrators, as well as information about the victim and perpetrator's relationship and the type of weapon used. Moreover, this also includes the timeline in which the incident occurred and the agency to which the case was assigned to and whether the case was solved or not. The Murder Accountability Project, founded by Thomas Hargrove, collated and made the data available.

The dataset is made up of structured data, mostly in tabular format, and contains a wide range of variables connected to homicide incidents. It includes information about the victims, offenders, and their relationships, as well as specifics regarding the occurrence. The dataset includes factors such as victim age, gender, and race, offender age, gender, and race, weapon used, victim-offender relationship, and incident location.

The data is organized into 24 variables and 638454 observations, with a total of 15322896 data points.

The information in this dataset was gathered from numerous law enforcement agencies and reporting systems across the United States. It contains standardized reporting forms as well as information from law enforcement officers and agencies.

The dataset is provided in .csv format, and users can access it on Kaggle.

Dataset Link: <https://kaggle.com/datasets/murderaccountability/homicide-reports/data>

## ***Data Processing Tasks***

Describe the data processing tasks you will need to do in bullet points. Here are some examples. Please provide a description along with the task.

* Data import
  + Description: We found the data set on Kaggle. It is 1 csv file. We inspected all the rows and columns well before selecting it for the project. We will directly download it from the website and import it into our analysis environment.
* Data cleaning/ Managing missing data
  + Description: Here we will analyze our entire data set and look for missing values, unknown values, and/or values that do not fit the outcomes needed for that particular column. If we find any duplicate records, we will drop them. We will drop any rows or columns that we do not think are useful for our analysis, as well. Depending on the number of missing values, we will evaluate if we will populate based on trends or drop.
* Exploratory Data Analysis
  + Description: Conduct preliminary EDA to understand data distributions, identify outliers, and visualize patterns.
* Indexing, selection, and filtering
  + Description: Creating new relevant dataframes for each question that include only the rows and columns needed to answer that particular query.
* Data Encoding
  + Description: Encode categorical variables into numerical format. For example, we need to encode gender and weapons for respective analysis.
* Data transformation / Merging and reshaping data
  + Description: Deciding how we want to categorize our data (for example, splitting ages into bins).
  + Create derived variables or features that are relevant to your analysis. For example, we need to combine year and month for time series analysis.
  + Aggregate data to the desired level of granularity for our analysis. For example, you might group data by year, month, or location.
* Data Splitting
  + Description: Here, we split our dataset into a training set(80%) and test set(20%) in order to train our model and after our model is trained we can use the test set to predict the crime rates.

## ***Data Analysis***

To answer some of our questions of interest, we have multiple ideas on the types of visualizations we would like to utilize. For visualizations, we are using numerous charts including but not limited to bar charts (specifically stacked bar charts), pie charts to determine the proportions, heat map to determine the patterns, pair plot to determine the relationships, and violin plot(male and female) to determine the outliers in our data, etc. Let’s get a bit more specific on the visualizations we would like to use based on each of our questions of interest.

* What are the primary demographic traits of homicide perpetrators and victims, and how have they changed over time? What was the most affected segment of the population?
  + For this particular question, we could potentially use a stacked bar chart. This way, viewers can understand the demographics of perpetrators versus victims and the evolution of these data points over time. It would be represented using a double bar for each year, with one bar being the perpetrators and the other being the victims which is then stacked by demographics within each bar. This would provide the user with an easy method of viewing all information needed in order to tackle a question like this. We can use this type of chart to analyze different demographic traits, such as: male vs. female by type of crime, number of perpetrators vs. number of victims by year, etc.
* What are the most frequent causes of homicides, and how do they alter depending on geography and demographics? What are the states or regions that have the most homicides and which causes are the most prevalent in each state?
  + Here we can utilize a heat map. We can prepare two separate maps with the first relating to geography (states) and the second relating to demographics. This way, we can map the most frequent causes and see heat color darken based on the number of homicides within that cause.
* Are there times when the number of homicides significantly rises or falls, and if so, what may have caused these changes—such as modifications to policy, changes in the economy, or adjustments to law enforcement tactics? Would any economic suffering or recession period contribute a significant effective relation to the frequency of homicide? Does the frequency vary based on season?
  + We can use a line chart here with the particular events marked on the chart in order for the viewer to easily be able to view if certain dips or increases were caused by certain factors. If we want to analyze based on the season, we could even use a 3D bubble chart, so that each bubble represents its own season and the size of the bubble represents the number of cases during that season over the years.
* Does the relationship between the victim and the perpetrator matter? Is it more common that the two sides know each other or don’t?
  + A pie chart would be the best visualization here so that we can view the percentage of the perpetrator and the victim knowing each other versus not knowing each other.
* For the crimes that were solved, which agency types were the most effective? Which states’ agencies were the best at solving the crimes? Which were the worst?
  + Using a sunburst chart, we can display the amount of solved and unsolved cases here because the sunburst chart is used to visualize hierarchically.
* Correlation between perpetrator age and gender vs. victim age and gender. Which generation would be the most affected by homicides? Is there any generation gap between the victim and the perpetrator? Can we decrease this crime rate? (Analyzing ages and age groups)
  + We can analyze this by creating additional columns in order to view the correlations or we can use a scatter plot to have each gender represented by a different colour and then plotting based on ages.
* What kind of weapons did certain age groups prefer to use? What kind of weapons did the different genders given use?
  + Based on age groups or genders, we can plot the frequency of certain weapons used using either an area chart or a cluster chart. Cluster charts can reveal patterns for us to use to create proper inferences on weapon use in homicides and why certain weapons can be more frequently based on age or gender.

***Predictive Analysis***

* Model selection for predictive analysis
* Description: To predict the crime rates an appropriate model should be selected.
* The model which can Be chosen is either the simple linear regression model or Random forest or gradient boosting.
* Linear regression for a simple baseline model and random forest for more complex operations that can potentially capture complex patterns.
* Model evaluation
  + Description: Here, the model is evaluated with the help of various metrics like the Mean Absolute Error, Mean squared error or Root Mean Squared Error.
  + This will give us a clear, numerical indication of how much the predicted values differ from the actual value.
* Prediction
  + Description: Using the trained model, we will predict the crime rates in different places.

**Why predictive analysis?**

In our journey of predictive analysis, we embark on a mission to unveil the dominant types of

criminal cases that haunt our communities. Our ultimate goal is to shed light on where these incidents tend to manifest most frequently. This endeavor is driven by a noble purpose - to assist various agencies and government bodies in making strategic decisions about resource allocation. By understanding where crimes are most prevalent, we aim to help them deploy their efforts effectively, ensuring that our neighborhoods become safer havens for all. Our story is one of data-driven exploration, a narrative of dedication to reducing crime rates and improving the lives of our citizens.

# 

# 

# **Expected Findings**

Expected Findings:

* Demographic Trends:

- Victim-Perpetrator Demographics: Assuming that the majority of perpetrators are male (63%) and handguns are the most common weapon used (50%), it might be inferred that young males might be a high-risk demographic for both victims and perpetrators. Further analysis can explore age groups within this category.

- Trends Over Time: Analyzing trends over time, especially related to gender and weapon use, can reveal if there are any changes in these patterns. For example, observing a decrease in male perpetrators or a shift in weapon preferences might indicate changing societal norms or law enforcement efforts.

* Causes and Motivations:

- Frequent Causes: Considering that handguns are commonly used (50%), it might be inferred that impulsive crimes or crimes of passion are prevalent motives. Additionally, the high percentage of municipal police involvement (77%) suggests that some homicides might be related to law enforcement activities or confrontations.

- Geographical and Demographic Variances: Since California and Texas have higher percentages of homicides (16% and 10%, respectively), exploring the causes specific to these states might reveal localized issues. Urban areas within these states might exhibit different motives compared to rural regions.

* Spatial and Temporal Patterns:

- Spatial Hotspots: Given that municipal police are involved in 77% of cases, spatial analysis can focus on areas with strained police-community relations. Understanding the reasons behind these hotspots, especially in states like California and Texas, is crucial for targeted law enforcement efforts.

- Temporal Trends: Since 99% of cases are categorized as murder or manslaughter, studying temporal patterns might help in understanding the timing of impulsive crimes or premeditated attacks, aiding law enforcement in strategic planning.

* Weapon Use and Availability:

- Firearm Usage: With handguns being the most common weapon (50%), analyzing the availability and sources of these firearms can provide insights into gun control policies. Additionally, studying the use of alternative weapons can reveal specific scenarios where impulsive crimes might occur.

- Suffocation: The suffocation of homicide, there should be a relationship between victim and perpetrator based on the age and gender. Suffocation means using no weapon but hands. Using two hands around the victim's neck. putting pressure on the victim's neck using a body part like a foot, arm, knee, or anything else. Assuming elders would have the difficulty to become a perpetrator in the case of suffocation.

* External Factors and Policy Impact:

- Economic Factors: Since economic factors can influence crime rates, exploring the relationship between economic fluctuations and homicide rates is essential. Policies addressing socioeconomic triggers might be particularly effective in reducing homicides in vulnerable communities. For instance, perpetrators would use other weapons instead of guns due to the economic impacts of the recession because guns may be too expensive compared to other weapons.

- Policy Changes:Studying the impact of gun control laws and law enforcement strategies, especially in states like California and Texas, can provide valuable insights into the effectiveness of such policies.

* Victim-Offender Relationships:

- Intimate Partner Violence: Given that 22% of victims are female, focusing on intimate partner violence prevention programs and legal measures is crucial. Understanding patterns related to domestic violence can inform targeted interventions and support systems for victims.

- Stranger vs. Non-Stranger Crimes: With 77% of perpetrators being male, distinguishing between stranger and non-stranger crimes involving male perpetrators can guide preventive measures, especially for vulnerable populations.

- Children vs. Spouse: Based on immediate family members, the ratio of son and daughter should be much lower than the ratio of spouses. Thus, The ratio of son and daughter should have the lowest percentage as a victim.

* Seasonal and Event-Driven Trends:

-Seasonal Variations:Certain crimes might exhibit seasonal patterns. For example, outdoor-related crimes could increase during warmer months. Identifying such trends can aid law enforcement in seasonal planning.

-Event-Related Spikes:Events like holidays or public gatherings might influence crime rates. Recognizing these spikes and their causes can help in proactive law enforcement measures during such periods.

* Community Impact and Fear Perception:

- Community Response: High homicide rates can instill fear and negatively impact communities. Understanding community responses and their subsequent impact on social cohesion and safety perceptions is valuable for community-oriented interventions.

- Fear Perception:Analyzing public surveys or sentiment data related to fear of crime can provide qualitative insights. Perceptions of safety and fear can affect community well-being and might reveal areas needing targeted social support or law enforcement engagement.

* Correlation between officers of each title and the case solved ratio in each state:

-There should be a positive correlation to each officer’s title, and the municipal police should have the highest correlation with case case-solved ratio compared to other titles. The primary municipal police duty is to investigate local crimes effectively.

* Perpetrator Age: middle-aged would create a gap in the frequency of homicide.

-For instance, the case number of the perpetrator who is over age 40 is less than the case number of the perpetrator who is less than age 40.

* Agency and success rate : Agencies with low success rates can enhance project outcomes by increasing budgets for improved resources and technology, enabling comprehensive planning. Additionally, expanding and upskilling the workforce fosters collaboration, innovation, and efficient task management, leading to enhanced productivity and overall project success.

# **Project Timeline**

Each group member should feel responsible for aspects of the project. In the table below, assign group members to take the lead on tasks and assign tentative due dates. Feel free to modify the table to suit your needs. Note: Other group members can/should participate in tasks.

| **Task** | **Task Lead(s)** | **Due Date** |
| --- | --- | --- |
| ~~Data collection and Inspection~~ | ~~All~~ | ~~10/13/2023 (8PM)~~ |
| ~~Identifying the problem~~ | ~~Krithika~~ | ~~10/13/2023 (8PM)~~ |
| ~~Define scope~~ | ~~Krithika~~ | ~~10/13/2023 (8PM)~~ |
| ~~Decide how to visualize~~ | ~~Lahari~~ | ~~10/13/2023 (8PM)~~ |
| ~~Project proposal submission~~ | ~~Lahari~~ | ~~10/13/2023 (8PM)~~ |
| ~~Performing analysis to access the quality of the data (EDA)~~ | ~~Krithika~~ | ~~10/13/2023 (8PM)~~ |
| ~~Deciding whether or not predictive analysis is feasible for our dataset~~ | ~~Siddharth~~ | ~~10/13/2023 (8PM)~~ |
| ~~Deciding method of data storage / reorganizing data to be in a readable format~~ | ~~Niharika~~ | ~~10/18/2023~~ |
| ~~Managing missing data / data cleaning~~ | ~~Lahari~~ | ~~10/22/2023~~ |
| ~~Exploratory data analysis~~ | ~~Matthew~~ | ~~10/22/2023~~ |
| ~~Indexing/ Selection/ Filtering~~ | ~~Abhishek~~ | ~~10/31/2023~~ |
| ~~Data transformation / Merging and reshaping data~~ | ~~Niharika~~ | ~~11/7/2023~~ |
| ~~Stacked bar chart~~ | ~~Lahari~~ | ~~11/7/2023~~ |
| Geospatial analysis | Niharika | 11/7/2023 |
| ~~Time series analysis~~ | ~~Krithika~~ | ~~11/7/2023~~ |
| Demographic analysis | Priya | 11/7/2023 |
| Select the model (predictive analysis) | Siddharth | 11/8/2023 |
| ~~Finding observations~~ | ~~Abhishek~~ | ~~11/10/2023 (8PM)~~ |
| ~~Creating inferences based on observations~~ | ~~Matthew~~ | ~~11/10/2023 (8PM)~~ |
| ~~Markdowns (prior to update)~~ | ~~Biruk~~ | ~~11/10/2023 (8PM)~~ |
| ~~Comments (prior to update)~~ | ~~Biruk~~ | ~~11/10/2023 (8PM)~~ |
| ~~Code review for optimal code~~ | ~~Niharika~~ | ~~11/10/2023 (8PM)~~ |
| ~~Scrum style meetings~~ | ~~Niharika~~ | ~~11/10/2023 (8PM)~~ |
| ~~Project update submission~~ | ~~Niharika~~ | ~~11/10/2023 (8PM)~~ |
| Sunburst chart | Priya | 12/2/2023 |
| Scatter plot | Biruk | 12/2/2023 |
| Cluster chart | Siddharth | 12/2/2023 |
| Splitting the data set (predictive analysis) | Siddharth | 11/15/2023 |
| Training the model (predictive analysis) | Abhishek | 11/25/2023 |
| Come up with predictions (predictive analysis) | Lahari | 12/2/2023 |
| Markdowns (prior to report/showcase) | Lahari | 12/4/2023 (8PM) |
| Comments (prior to report/showcase) | Priya | 12/4/2023 (8PM) |
| Project report submission | Matthew | 12/4/2023 (8PM) |
| Project showcase submission | Matthew | 12/4/2023 (8PM) |
| Group peer eval submission | All | 12/12/2023 (8PM) |

UPLOAD DATASETS, HTML, NOTEBOOK