# Kyle Kingston

DSC630

Professor Andrew Hua

Murder stats and caught prediction

# Introduction

Crime in the us has been a hot button issue at both a national, and local level. Overall murder rates appear to have been dropping until recent years where we have seen a spike.

A graph with a line going up

Description automatically generated

I would like to see if we can accurately predict what crime is being committed based on various aspects of the victim. I am wanting to find the answers to the following questions:

1. What characteristics of a victim seem to lead to a crime not being solved.
2. How does the distribution look between the states
3. Do larger cities have more murders per 100k population
4. What factors seem to have the largest correlation to being murdered.

I’m open to other questions as the data becomes more familiar and we can explore other areas within the data.

# Data Information

Dataset link: <https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset/data>

This data set was gathered by the Us Government, Crime department in collaboration with local law enforcement agencies. This data was gathered and uploaded to Kaggle.com at the link above. The dataset consists of the following fields and descriptions:

* Record ID: Unique identifier for the crime
* Agency Code: Code for Agent who recorded the crime
* Agency Name: Name of agency
* Agency Type: Type of agency
* City: City where crime was committed
* State: State where crime was committed
* Year: Year that crime was committed
* Month: Month that crime was committed
* Incident: Number of incidents
* Crime Type: [Murder or Manslaughter, Manslaughter by Negligence]
* Crime Solved: If the crime was solved
* Victim Sex: Sex of the Victim
* Victim Age: Age of the Victim
* Victim Race: Race of the Victim
* Victim Ethnicity: Ethnicity of the Victim
* Perpetrator Sex: Sex of the Perpetrator
* Perpetrator Age: Age of the Perpetrator
* Perpetrator Race: Race of the Perpetrator
* Perpetrator Ethnicity: Ethnicity of the Perpetrator
* Relationship: Relationship between the victim and Perpetrator
* Weapon: Weapon used in the crime
* Victim Count: Number of Victims
* Perpetrator Count: Number of Perpetrators
* Record Source: Source of the crime record

# Ethical Implications

This project can potentially run into some ethical issues. While the data is public record there is still a level of privacy that is being violated by bringing these individual matters into the forefront. This evaluation will reveal some factors that make you more prone to being murdered. That can cause some people to feel uncomfortable being a potential target

# Project step outline

Project steps:

1. Import data and save as data frame
2. Preform EDA
3. Visualize variables to better represent their relationship with the dataset
4. Remove variables that will cause overtraining [any perpetrator variables]
5. Divide dataset and train model
6. Evaluate model performance
7. Hyperparameter tune if necessary

For this project I plan to use a logical regression model with a target variable of the solving of the murder. This decision was based off the target variable being a binary one, either the crime was solved or not solved. This type of target variable leans strongly toward a logistic regression model. if my logistic regression model does not meet expectations, then I will consider pivoting to a knn or neural network model.

# Model Evaluation

Evaluation metrics:

1. Accuracy: This will help to evaluate when our predictions are correct and when they are not based on our training and test data sets. We will be taking the mean of the cross validation scores of our model.
2. Roc Curve: This will display a graph of our true positive vs false positive rate.
3. Confusion Matrix: This is a way to compare the predictions the model makes vs the correct variable. This helps to visualize where the model is getting confused when creating its predictions
4. Classification Report: This report will provide a wide array of values that will help to evaluate the model.
   1. Precision: provides a proportion of positive identifications. If the model produces no false positives, it will have a score of 1.0
   2. Recall: similar to precision but the score is reflective based on false negatives instead of false positives
   3. F1 score: A combination of precision and recall. If both precision and recall are perfect and no false positives or false negatives are present, then the f1 would be 1.0
   4. Accuracy: the accuracy of the predictions in the model
   5. Macro avg: average of precision, recall and f1
   6. Weighted avg: average of precision, recall and f1 weighted by how many samples of each classification

# Contingency plan Overview

Dataset link: <https://www.kaggle.com/datasets/ikynahidwin/depression-professional-dataset>

If my original project fails for some reason, I will use a separate data set to predict the outcome of depression. Essentially, this dataset has characteristics of people and records if they are diagnosed depressed or not. I would also plan to use a logical regression on this data set. The questions I would like to answer would be:

* 1. What factors have the largest effect on depression diagnosis
  2. What demographic groups seem to have a disproportionate amount of depressed population?

The condition of utilizing this contingency is as follows:

1. The dataset appears to be lacking quality
   1. If the data has a large number of missing values to the point where it would not provide a large enough sample to properly train a model
2. The model performs poorly even after tuning methods are implemented

The potential risks of using this contingency is the timeline of the project potentially being compromised. If the contingency is implemented during the early stages of the project this will likely not be a factor but if toward the end of the project that the contingency is used, then that may compromise the quality of the analysis if the project is rushed. I may perform eda on this dataset during the coming weeks in hopes of alleviating the concerns of time if this plan is implemented in the 11th hour.

# Milestone 3

# Will I be able to answer the questions I want to answer with the data I have?

1. What characteristics of a victim seem to lead to a crime not being solved.
   1. This question can be answered within this dataset, we have various characteristics and can identify the most prevalent features of the victims based on histograms and bar plots of the data.
2. How does the distribution look between the states
   1. Yes this can be done with the dataset. We have recorded murders and the location of those murders. Once again, a bar chart will be extremely helpful in this identification and visualization.
3. Do larger cities have more murders per 100k population
   1. For this question we will need to pull population statistics in order to determine the population of the cities in question. This should be relatively easy to find and incorporate, likely from the census bureau
4. What factors seem to have the largest correlation to being murdered.
   1. With solved being the target variable of our Logistical regression model we can utilize that data to find what feature had the greatest impact on the model during training to answer this question.

# What visualizations are especially useful for explaining my data?

I find the Histograms and bar plots to be essential in understanding the data. Having an Idea of how the data is diswtributed helps to gather an understanding of the dataset as a whole. For instance, the histogram for victim count shows a large amount of 0 count offences, this raises the question of how that could be possible or what does that represent in the data. After digging further into this I have not found a conclusive answer

# Do I need to adjust the data and/or driving questions?

We should be able to answer all the questions with the data on hand with the exception of the population data that we will need to pull in at a future step. Overall the questions seem to have a straight forward path for solution. Some will need to be gathered through some additional analysis.

# Do I need to adjust my model/evaluation choices?

My model performs rather accurately, and I feel will yield a reasonable result for this determination. I utilized a pipeline when applying the model so it should be easy to apply additional models in order to find what yields best results. Right now, our accuracy is in the acceptable range being 72% accurate. Though I have hopes that integrating an additional model or hyperparameter tuning this model will increase the accuracy.

# Are my original expectations still reasonable?

I feel the original expectations are still reasonable. The analysis seems to be coming along smoothly, and I don’t foresee any insurmountable problems. If somehow issues arise, I can lean back on the backup dataset but at this point I feel like that is very unlikely to be necessary.

# Sources:

Agrawal, A. (2023, October 2). *US crime dataset*. Kaggle. https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset

Jatmiko, N. A. (2024, November 20). *Depression professional dataset*. Kaggle. https://www.kaggle.com/datasets/ikynahidwin/depression-professional-dataset

Korhonen, V. (2024, November 12). *USA: Reported murder and Nonnegligent Manslaughter Rate 2023*. Statista. https://www.statista.com/statistics/191223/reported-murder -and-nonnegligent-manslaughter-rate-in-the-us-since-1990/