# Predicting Crime Severity Rates in Los Angeles County to Help LAPD Better Allocate Resources

Project Final Report

DSCI 550: Data Science at Scale, Fall 2024

Dr. Seon Ho Kim

By: Tushar Elangovan, Jacob Graber, Yun Cheih Lee (Jack), & Kayla Schwefler

## Problem

With the social and political climate surrounding our world today from the BLM movement during the pandemic and even the election, many Americans have been feeling restless and unsafe. Although we are a couple years behind the peak of its popularity, “defund the police” has been an extremely common sentiment especially surrounding George Floyd. So many taxpayer dollars have been put into helping run this organization, but Los Angeles is still often perceived as a crime-ridden city.

This raised the question: How effective really is the police force that is local to all of us within Los Angeles? With the rising political tensions following the election, we want to address the concerns and effectiveness of LAPD, and maybe even give them credit where it is due. The major issue is that many Los Angeles citizens seem distraught from the police department’s operations, and we would like to either help put those concerns to rest, or at least shed some light on how the LAPD might be able to better allocate its limited resources. In order to accomplish this, we set out to **show that predicting crime severity rates** in different areas of LA County is possible using **deep learning methods**. With more accurate predictions of crime rates by location in LA County, we hope that LAPD can better allocate police presence and funding where it is needed most.

## Dataset

The dataset used in this research consists of close to 1 million crime reports from Los Angeles County, spanning from 2020 to 2024 (Gostinski, C.). It contains 28 columns, including crucial information such as crime types (e.g., robbery, theft, assault, homicide), temporal details (date, time, and location), and victim demographics (age, gender). Additionally, the dataset includes incident-specific data like the involvement of weapons, the premise of the crime, and arrest status. Geographical details such as area names and coordinates further enrich the dataset, allowing for spatial analysis of crime patterns. This comprehensive dataset enables a

multifaceted exploration of crime trends, victim profiles, and the effectiveness of law enforcement efforts across different periods and regions in Los Angeles.

### Handling Error and Outliers:

With such a large dataset, anomalies like null values or incorrect values are impossible to avoid. Upon further inspection, many of the null values appeared in fields like weapon used (simply signifying the lack of a weapon used) or victim details (potentially due to anonymous reports or a victimless crime). We also encountered a few other issues like missing longitude/latitude fields or missing premise descriptions. The dataset also had other matters like negative age values. Given that these were clear errors and impossible occurrences, we treated them as their absolute value instead as it was a likely mistake from human data input error. Overall, because the missing variables were due to the value themselves (and only on <1% of our total dataset), we chose to ignore these missing values as it was unlikely to cause a bias within our inferences.

## Exploratory Data Analysis

Prior to developing our model, we performed a thorough exploratory data analysis (EDA) to uncover essential patterns and relationships within the dataset. Our analysis focused on crime-related trends across different dimensions, including location, time of day, day of the week, and monthly variations. Additionally, we examined crime prevalence across various geographic areas within Los Angeles, observing patterns both broadly and over time. This approach also involved investigating the relationship between crime and demographics, particularly victim age, to better understand the population affected by different types of crimes. By analyzing spatial distribution and temporal patterns, we were able to derive insights that informed the development of our crime index model, ensuring it accounted for key factors influencing crime rates

## Crime Trends

1. Monthly Analysis:
    a. **Most Crime-Active Month**: July with 15% higher crime rates than the average total monthly crime rate.
        i. One of the main factors is that summer months often see increased activity, with warmer weather encouraging more people to be out in public. This increased public presence, especially in densely populated urban areas, creates more opportunities for crimes
    b. July and August are the months with the highest average crime rates.
    c. We can see a general trend of crimes dropping towards the second half of the year.

i. Due to dropping temperatures during fall and winter, the frequency of outdoor social gatherings and street-level crimes decreases. Cooler weather encourages people to stay indoors.

d. **Least Crime-Active Month by**:
   i. Total Crime Rate: **November**
   ii. Average Crime Rate: **February**

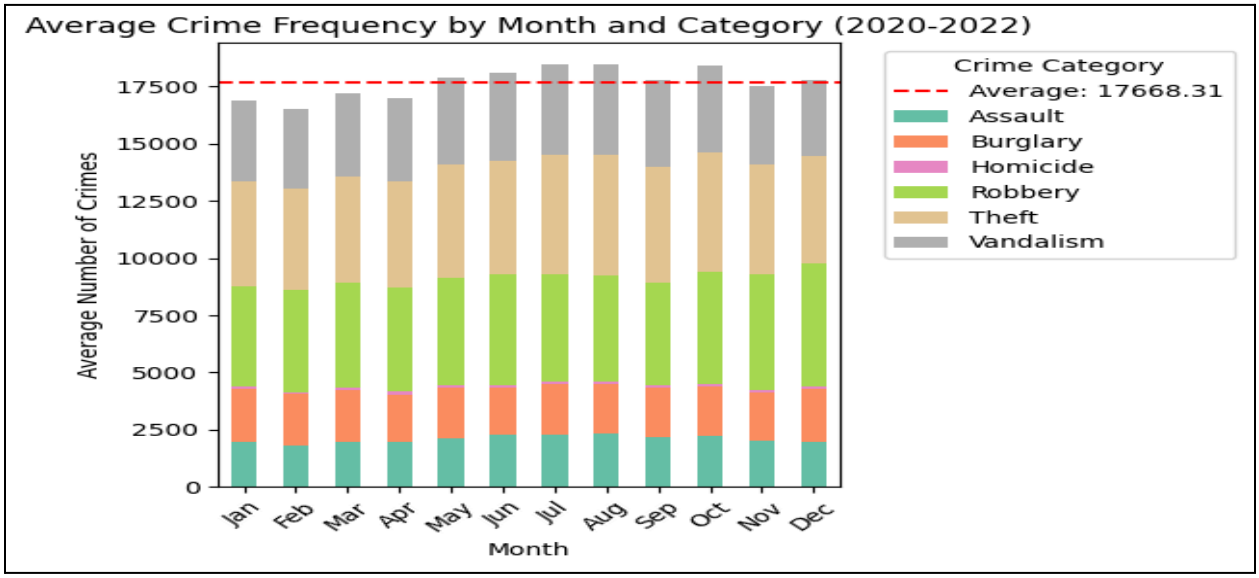


**Fig 1**: Average Crime by Month and Crime Category

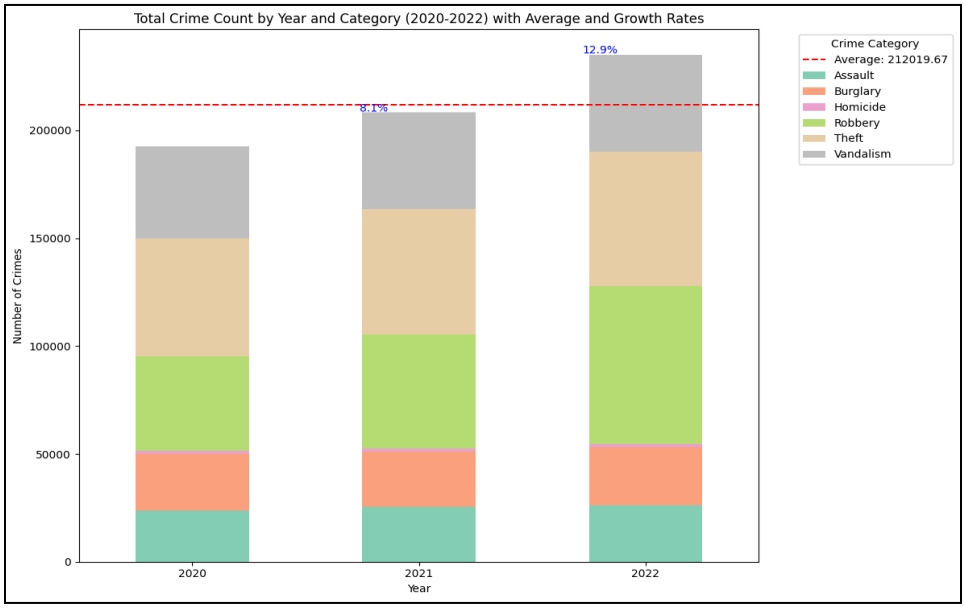


**Fig 2**: Total Crime by Year and Crime Category

2. Yearly Trends:
   a. **Peak Year**: 2022 saw the highest crime rate, with 13% more crimes than average.

b. **Lowest Year**: 2020, possibly due to lockdown effects.

3. Top Crime Categories and Percentage of total Crimes
   **a.** Theft - 27.4%, Robbery - 27.2%, Vandalism - 20.7% - Total Contribution **75.3%**
   b. High levels of **poverty and income inequality** in Los Angeles may drive some individuals to commit theft and robbery as a means of survival or for quick financial gain.
4. Location and Crime Analysis:
   a. Top 3 areas with the highest crime count and their most common crime type:
      i. Central: Burglary, 77th Street : Theft , Pacific: Theft
      ii. **Central and Pacific LA**: These areas are known for having both **high population density** and a significant presence of social issues such as **homelessness, substance abuse, and gang-related violence**.

## Age Demographic

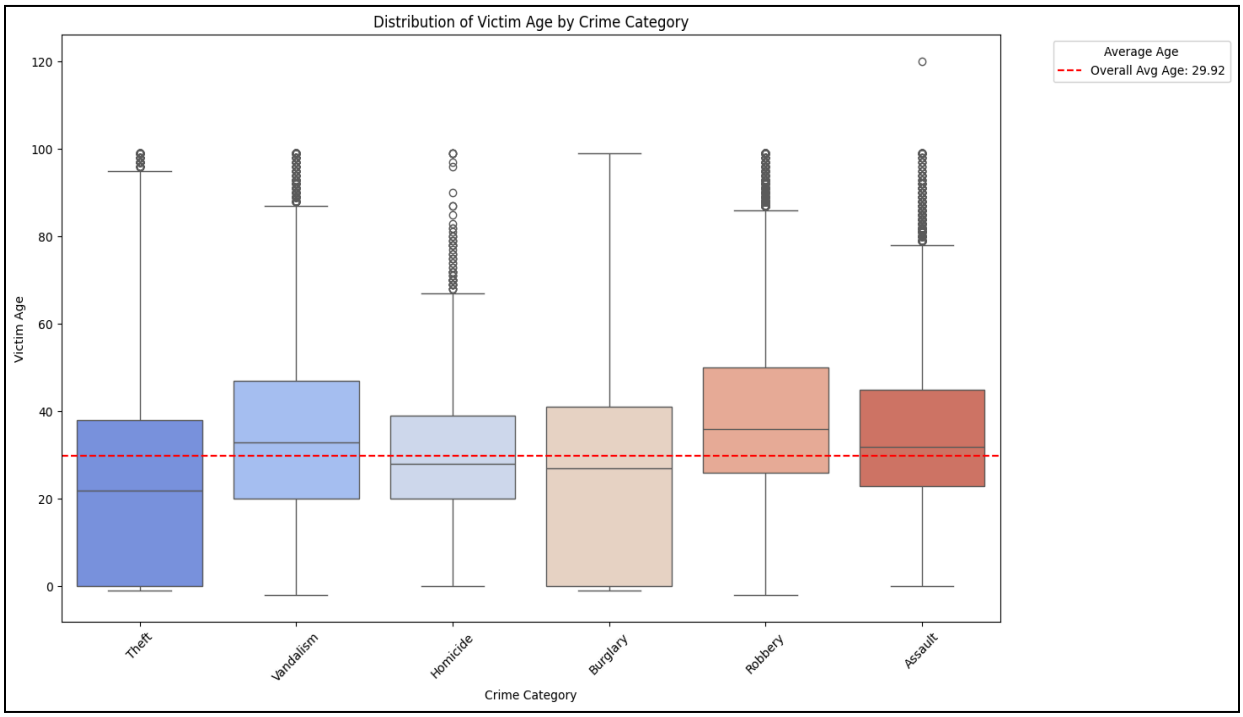


**Fig 3**: Victim Age distribution by Crime Category

1. Approximately **45.63%** of all recorded crimes involve victims within the Adult age group, specifically those between 27 and 35 years old. This could be influenced by several factors such as
   a. **Economic Aspect:** This age group seems to be the most active in terms of them being in the early to mid stages of their careers and navigating financial

instability. These economic stressors can lead to an increased risk of involvement in or victimization by crimes like robbery, especially in urban settings where economic disparity is more pronounced.

b. **Crime Opportunities**: Individuals in the 27-35 age group are more likely to be out in public spaces at times when crimes such as robbery are more likely to occur (e.g., late nights, weekends). Average Victim Age: **30 years**

2. Crimes such as Vandalism, Robbery and Assault seem to have a higher victim average age as compared to Theft, Homicide and Burglary. Potential reasons being:
   a. **Theft** and **burglary** tend to disproportionately affect lower-income and younger populations who may be more prone to opportunistic crimes, such as taking advantage of unattended property or entering empty homes.
   b. **Robbery** can disproportionately affect people who hold higher-value items, often linked to adults with more financial stability and material assets.

## Spatial Distribution

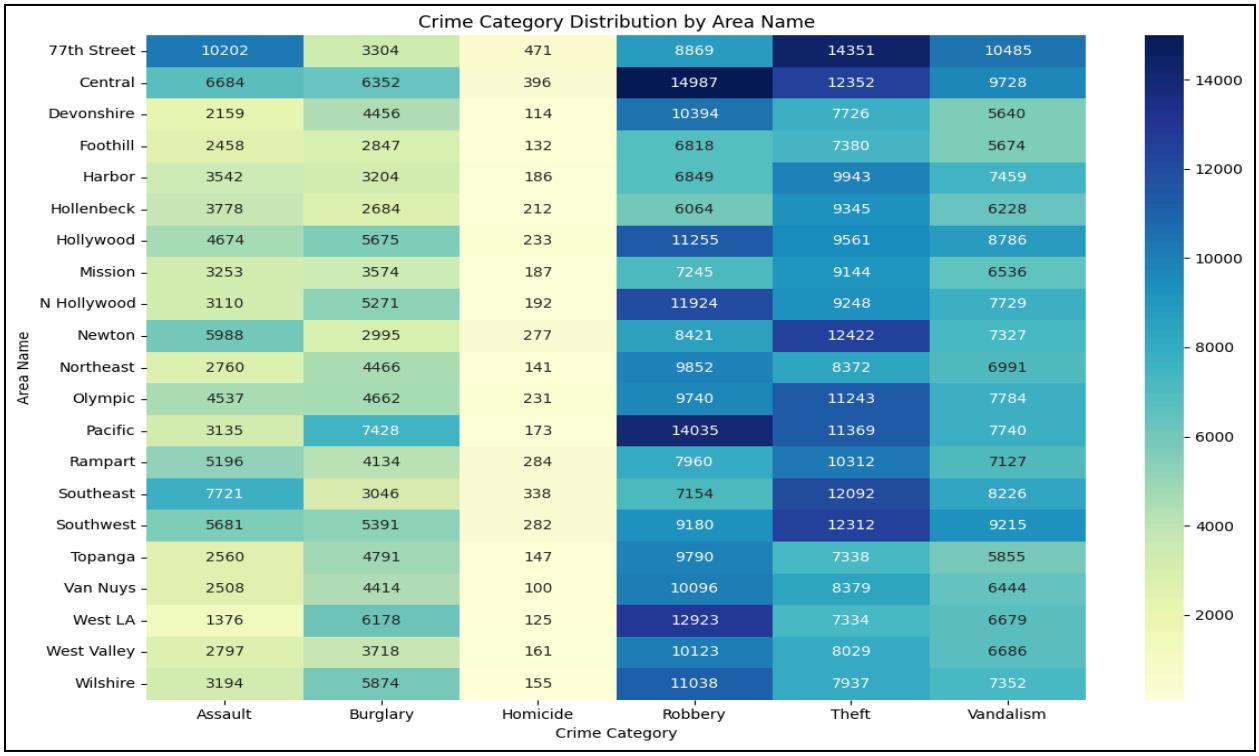


**Fig 4:** Crime Category Distribution by Area Name

1. **Areas with Highest Crime Rates**:
   - **Central LA** represents **7.04%** of the total crime incidents, with a significant 25% annual increase in crime rates from 2020 to 2022.
   - The **77th Street** area represents **6.41%** of the total crime incidents.
   - Top 3 unsafe Areas: **1. Central, 2. 77th Street, 3. Pacific**

2. **Safe Zones**:
    - **Foothill** has the lowest crime reports, at just **3.40%** of the total crime in LA.

### EDA Final Recommendations:

1. In order to address the prevalence of crimes such as Robbery, Theft, and Vandalism, it is recommended that the LAPD focus increased attention on the **Central LA** area to investigate the underlying factors driving these criminal activities. This area also shows a **25% yearly increase in crimes over the past 3 years.**
2. Additionally, the LAPD should explore the reasons behind the notably low crime rates in the **Foothill area**, identifying any positive patterns or approaches that could be adapted to higher-crime areas.
3. Given challenges such as **homelessness and population density**, the LAPD could consider reallocating resources, such as **increasing patrol presence** or enhancing safety measures like **surveillance systems**, to ensure that both residents and tourists feel secure.

## Crime Index Scoring

We chose to build our own index of crime severity because the Cambridge Crime Harm Index data had very broad ranges for harm values (Administrator, 2020). Murders would be in the 5000's while a simple misdemeanor was in the 10's. This put an extreme weight towards more serious crimes which would heavily skew our dataset, and hindered the training of our model.

To normalize the dataset, the crimes were compared and rated with severity scores from **0 to 100**. This gave a much more typical range and **helped reduce bias and skew** while allowing us to add other factors into consideration. Besides the base severity score, our normalization also considers weapons used as well as victim age.

The victim's age being greater than 0 and less than 18 (meaning the victim was a child) or of age larger than 65 (meaning the victim was an elder) gave it a **+10 on our index** due to the more **vulnerable population**. The usage of a **weapon also further increased index score by either 5**, **10, or 15** depending on the type of weapons used to commit the crime– this was considered because deadlier weapons were far more likely to have both physical and psychological impacts on the victims. The inspiration was taken from the legal court's bail system where bail costs are increased in accordance with the usage of a weapon (Federal Bureau of Investigation, 2011).

By generating our own index, we gave ourselves an additional layer of flexibility to transform our analysis while reducing bias when performing our analysis or training our model. This final crime severity index scoring provided the last piece that we needed to proceed with our analysis.

## Experiment

In order to help better allocate police attention and LAPD resources in general, our team attempted to create a time-series forecasting model to predict crime severity scores in each of the 21 areas in our dataset on any given day. To accomplish this goal, we first grouped all the crime index scores for a given day and area, turning our dataset into a 1280 row by 21 column dataframe, where each cell represents the total crime severity score for its respective area and day. Then, this dataset was reformatted into input sequences of 30 days of crime scores for all 21 regions, and output sequences of the 31st day of scores for all 21 regions. Finally, these sequences and their corresponding outputs were split into 80% training and 20% testing data, respectively.

After researching the most optimal neural network model architectures to accomplish our goal, our team decided that a convolutional neural network (CNN) would be the best suited to our task. This is because the CNN model, unlike an RNN or LSTM model, performs much better at maintaining the 'shape' of input data, which in our case is a 30 row by 21 column input set. After trying several different model architectures and going through extensive hyperparameter tuning, we ended up with two models that can somewhat accurately forecast crime severity scores for each of the 21 regions. The first of these, called 'Weekday Model', incorporates a 22nd column in the input sequence to capture the day of the week in order to better capture cyclical temporal dependencies. The second model, called 'Scaled Model', uses a MinMaxScaler to better normalize the data before training. Both models use the same model architecture, which consists of Conv1D, BatchNormalization, Dropout, GlobalAveragePooling1D, and Dense layers.

## Results

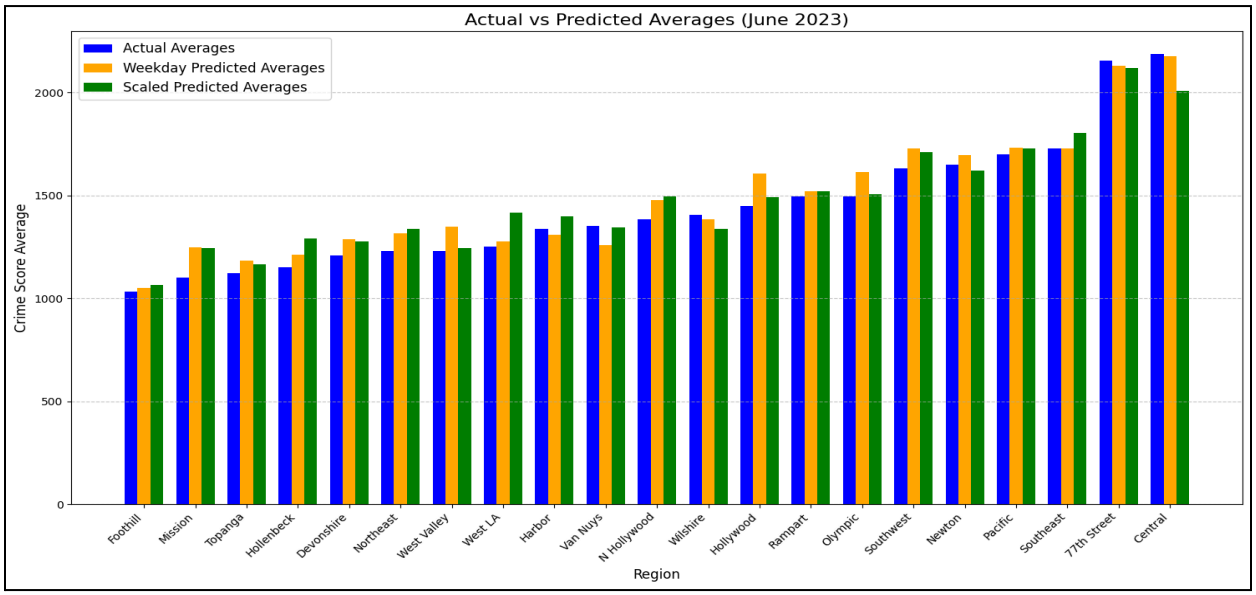


**Fig 5:** Actual vs. Predicted Average Crime Severity Scores by Region (June 2023)

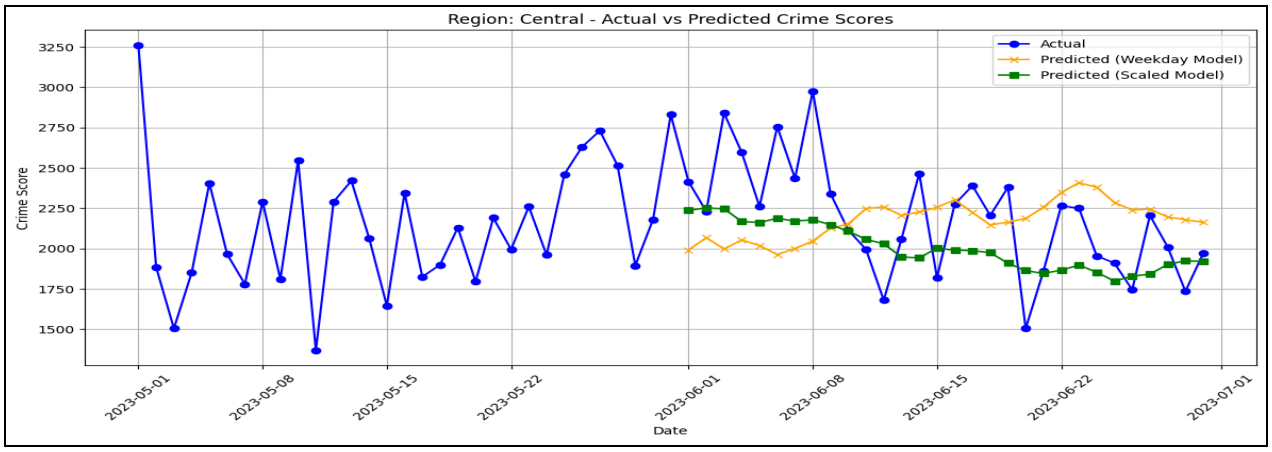


**Fig 6:** Actual vs. Predicted Crime Severity Scores (Central Region, June 2023)

As we can see here, the two models struggle slightly at capturing day-to-day fluctuations in crime severity scores, but ultimately are relatively effective at capturing overall trends for a given area over a 30 day period. It can also be seen that the 'Scaled Model' slightly outperforms the 'Weekday Model.'

## Conclusions and Observations

In conclusion, we believe that the crime severity rate prediction model could significantly assist the LAPD in better allocating its resources. By predicting the location and expected crime severity rates, the LAPD would be able to plan more effectively, selecting the most appropriate measures not only for prevention but also for deploying response units. While the current model may not fully capture day-to-day fluctuations in crime rates, we are confident that with the inclusion of more current data and improvements in methods for capturing spatiotemporal dependencies, this deep learning model could produce more valuable crime forecasts.

## References


1. Administrator. (2020, September 29). The Cambridge Crime Harm Index (CCHI). Institute of Criminology. https://www.crim.cam.ac.uk/research/thecambridgecrimeharmindex
2. Federal Bureau of Investigation. (2011). Crime measures. In Crime in the United States 2011. U.S. Department of Justice. Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/crime-measures
3. City of Los Angeles. (2023). Operating budget: Police department. Open Budget. Retrieved from https://openbudget.lacity.org/#!/year/2023/operating/0/department_name/Police/0/program_name?vis=lineChart
4. Gostinski, C. (n.d.). Crime data analysis [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/candacegostinski/crime-data-analysis