Final Project: Milestone (Project GitHub Repo)

Kadin Wilkins (kwilkins@berkeley.edu), Matilda Orona (matilda_orona@berkeley.edu), Vikram Magal (vikram_magal@berkeley.edu), Anushka Vazirani (avazirani@berkeley.edu)

Motivation

We aim to predict the number of violent crimes in a given zip code and month using past crime trends and zip code population and socioeconomic data. Understanding these patterns is crucial as it enables law enforcement agencies to better allocate resources and respond more effectively and safely. Accurate predictions could guide decisions about where to personnel, ensuring officers are adequately prepared, and facilitating measured de-escalation tactics where possible.

Additionally, uncovering these trends can support community policing efforts by distinguishing between areas prone to violent crime and those primarily experiencing non-violent offenses. Ultimately, our research seeks to enhance public safety by providing actionable insights that help law enforcement agencies strategically prioritize their resources.

Data Overview

We use a dataset published by the Los Angeles Police Department (LAPD) containing all reported crime incidents for the city of Los Angeles from January 1, 2020 to May 29, 2025, containing 1,004,991 rows and 28 columns (LAPD Open Data). Each incident includes detailed information such as time, date, victim demographics, latitude and longitude, and area name. The areas are distinguished by the 21 Los Angeles Community Police Stations districts. Each record also specifies the type of crime, whether a weapon was used, and, if so, the type of weapon.

American Census Survey 5-year data estimates allow us to incorporate population statistics and socioeconomic information gathered from the US Census survey (ACS Data). Crime rates are known to be related to population levels and socioeconomic conditions like median income, income inequality, unemployment rates, and poverty rates, all of which can be pulled and calculated from ACS data estimates. A challenge with the ACS data is that it is not available at the LAPD district boundary level, so we have decided to pull it at the zip code level and conduct our modeling at the zip code level to leverage this additional data source. We will use the most recent ACS 5-year estimates from 2023. While our data spans multiple years, population and socioeconomic characteristics tend to change slowly over time, so the 2023 data serves as an appropriate approximation for these variables over our time period of crime incidents.

Zip Code Tabulation Area (ZCTA) boundary data, sourced from the US Census Bureau's TIGER shapefiles (TIGER Data), are leveraged in this analysis to connect the LAPD crime data to the ACS data at the zip code level. The TIGER shapefiles provide each ZCTA's five-digit zip code and its boundary polygon and coordinates, allowing us to employ spatial joins to identify the zip code of a crime by its recorded latitude and longitude. Since the most recent year of ACS 5-year data is for 2023, we also pulled the 2023 TIGER ZCTA boundaries so the ACS data matches the geographic boundaries.

Data Cleaning, Processing, and Feature Engineering

To leverage ACS data at the zip code level, we merged the LAPD crime incidents with the TIGER ZCTA boundary data by their recorded geographic coordinates. Records with missing or invalid latitude and longitude (*n = 2,240*) or that could not be mapped to a ZIP code (*n = 406)* were dropped. We also limited the data to incidents prior to January 2024 because starting then, LAPD transitioned to a new crime reporting system to comply with FBI-mandated reporting standards. This resulted in a sharp drop in reported incidents in this dataset that does not reflect the complete number of incidents occurring in LA (see appendix). As it would be a challenge to model based on incomplete data, we dropped all incidents beginning January 1, 2024 and after *(n = 127,620)* left a total of *874,725* incidents across *149* zip codes in LA from January 2020 to December 2023.

Violent crimes were identified using a two-step classification process based on federal crime classification methodologies. In this study, violent incidents are those with a populated *weapon_code* or those whose *crime_desc* matched violent offenses per UCR Reporting and COMPSTAT definitions.

Next, we aggregated the number of violent incidents to the zip code-month level. Missing zip-month combinations between January 2020 and December 2023 were added and filled with 0s. We then calculated 1-, 2-, and 3-month violent crime lagging count fields and dropped the earliest months for which a full 3-month history was not available, leaving *6,705* rows of monthly violent crime totals for *149* zip codes from April 2020 to December 2023. This cleaned dataset contains *284,117* violent crimes. We also engineered temporal features, including a categorical season variable based on month and sine/cosine transformations of the month number to represent cyclical trends.

Population and socioeconomic indicators from the ACS data were merged by zip code. Missing ACS socioeconomic values (encoded as -666666 or -666667) were converted to NULL for now. Population is populated across all zip codes, so no records were dropped.

We will conduct further analysis to determine a strategy for imputation if we decide to include socioeconomic features in our models, but a preliminary correlation analysis (see appendix) indicates they may not be useful.

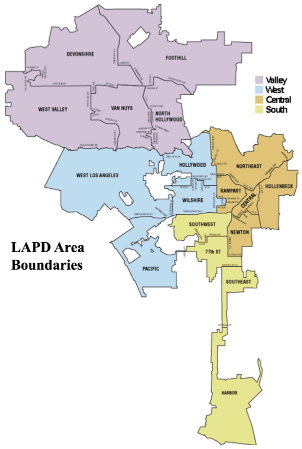
The final modeling dataset contains *6,705* zip code-month observations. Sequential splitting based on the month was employed to create our final modeling datasets.

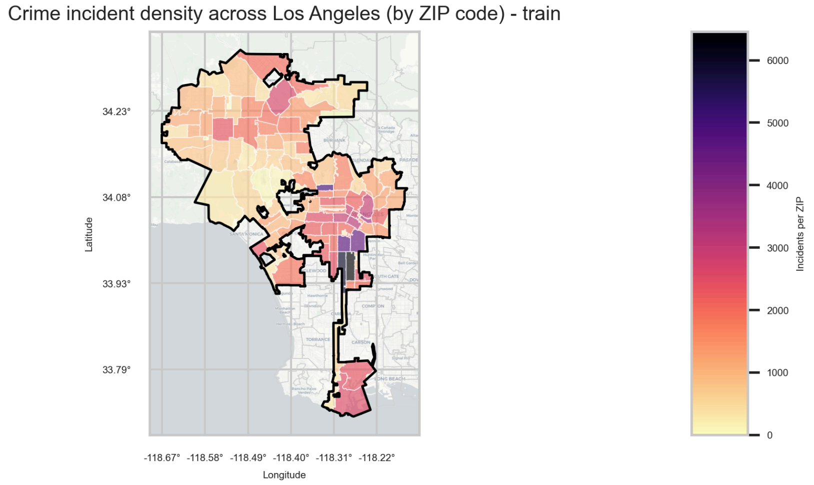
- Training (first 60% of dates): April 2020 - June 2022 (4,023 rows)
- Validation (next 20% of dates): July 2022 - March 2023 (1,341 rows)
- Test (last 20% of dates): April 2023 - December 2023 (1,341 rows)

Data exploration was conducted on the original scale for all variables, but prior to modeling, the target monthly violent crime count and lagging and population features will be standardized and the categorical season feature will be one-hot encoded.

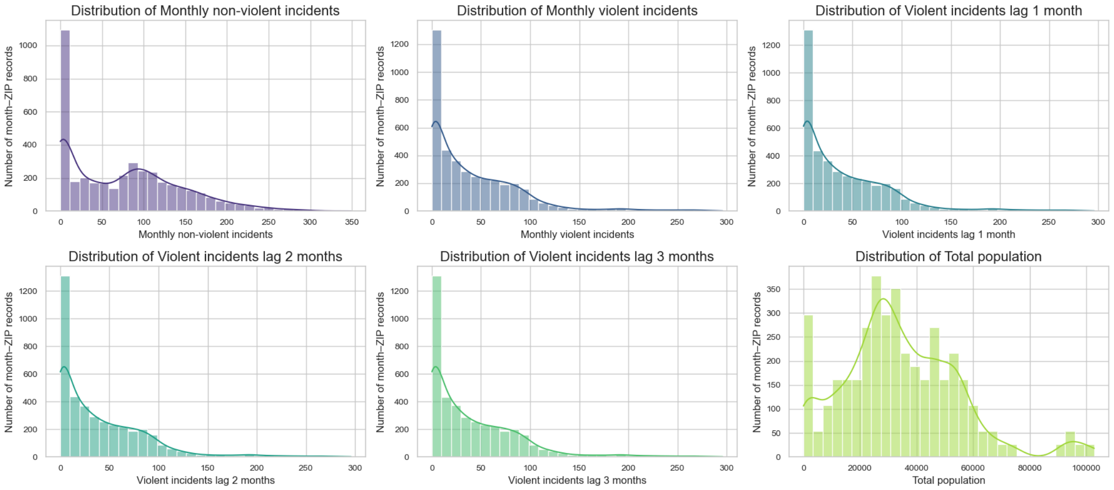
Exploratory Data Analysis

The following data exploration is conducted on our training dataset which contains 166,533 total violent incidents occurring in 149 zip codes in LA between April 1, 2020 and June 30, 2022.

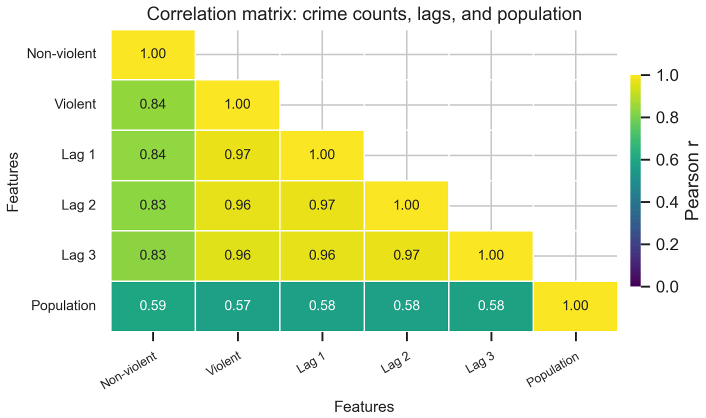




The map above on the right shows total violent crimes by ZIP code, where darker shading denotes a higher number of violent crimes. This allows us to see that certain zip codes have a significantly higher concentration of violent incidents, highlighting clear geographic disparities. When comparing violent crime by zip code to the LAPD area boundary map on the left, we see that there are varying levels of violent crime occurring within each LAPD area. This justifies the need for a prediction model at a smaller, more precise, level than the LAPD districts. If we are able to accurately predict future violent crime rates by zip code, it can help each LAPD district properly allocate their resources to the specific zip codes where violent crime is more likely to occur.



Both monthly violent and non-violent incidents demonstrate pronounced right-skewed distributions, with most ZIP-month pairs reporting relatively few incidents, while a small subset experience consistently high volumes. Part of this skew may reflect coverage artifacts, since some of the zip codes don't fully overlap the LAPD districts, leading to potential underreporting of crimes for those zip codes. We will investigate these zip codes during model error analysis to assess how they affect our model accuracy. The total population histogram presents a broader, multimodal distribution, predominantly centered between 20,000 and 50,000 residents per ZIP code. Notably, we identified five zip codes with a population of zero residents. Further analysis revealed that these zip codes were tied to PO boxes and universities such as UCLA and Pierce College. Despite their lack of residents, these ZIP codes had crime incidents reported and thus were retained in our analysis.



Our correlation analysis shows lagged violent crimes are almost perfectly correlated with the current violent crime count *(r = 0.96-0.97)*, indicating strong temporal trends. Total population by zip code is moderately correlated with our target *(r = 0.57)*, suggesting that composition and structural differences between communities have a significant impact and the population count alone is only a part of the story. It is also important to note that our lagging features are very highly correlated with each other. Including all of these could cause multicollinearity issues with models like linear regression and generally leaves us with redundant features. We elaborate in the next section on how we'll work with different models to handle this properly.

Methods and Experiments

Given that the lagged features are so highly correlated with a month's violent crime count, our baseline model will employ a basic linear regression using only the 1-month lag feature (# of violent crimes in the previous month) to predict the number of violent crimes for a month. To improve on our baseline, we will run various experiments with tree-based models and feature combinations, including the ACS variables. Tree-based models are generally better at handling highly correlated features, so we can include them all in these experiments. With a decision tree, we can visualize relationships within our data, before moving to RF and XGBoost. A random forest can be helpful for us, due to relative resistance to outliers, which is especially helpful for us given the variability in violent crime rates across zip codes. The feature importance scores from these models will also give us insight into the most important predictors of violent crime rates. We will also evaluate if XGBoost offers additional improvements because it also provides feature importance scores and has high predictive accuracy on large datasets like ours.

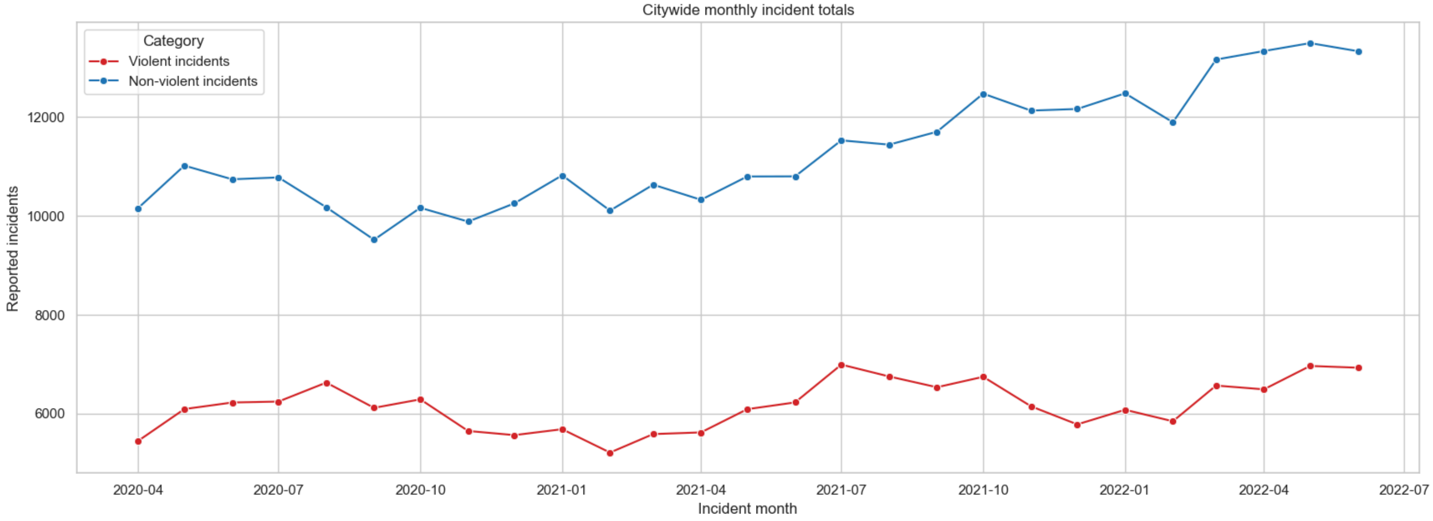
For model tuning we will use the validation dataset. With decision trees, we can tune hyperparameters such as maximum tree depth and minimum number of samples per split. Random forest also allows us to tune the maximum depth, as well as number of features. With XGBoost, we can tune similar metrics as well as the learning rate. Again, given that we have a lagging feature that is so highly correlated with our target variable, we may see high model performance without much model tuning, but will run these experiments to see what improvements they can offer.

In order to evaluate model performance and select a final model, we will compare mean squared error across the training, validation, and test sets. We aim to find a highly generalizable model without needing an overly complex model, so if we find that more complex models do not offer much lower error rates in comparison to our baseline, we will not favor those.

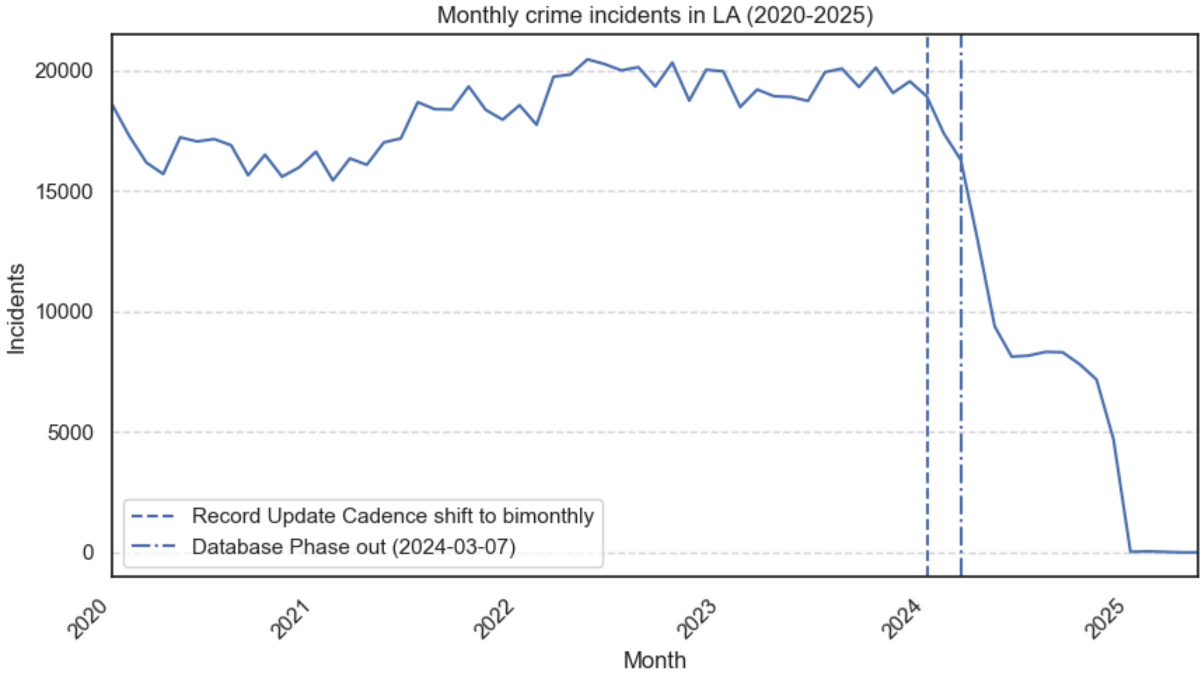
Team Contributions

- **Vikram**: Jupyter notebooks (01 EDA.ipynb). Violent crime classification description in report. I had limited availability this week due to travel but remained in contact with my team and will take on more work for the next deliverable.
- **Matilda**: Conducted EDA, data cleaning (outlier treatment), and visualizations for the LAPD and the final training dataset (crime_data_eda_MO.ipynb, eda_matilda.ipynb). Contributed to the motivation, data description and EDA sections.
- **Kadin**: Conducted analysis on dataset features, and how to best implement them for our baseline model. Created a baseline as we explored how to best utilize our data (EDA-KW.ipynb), including variable transformations and visuals. Assisted in documentation.
- **Anushka**: Jupyter notebooks for data cleaning, preprocessing, and EDA (download_ACS.ipynb, preprocessing_AV.ipynb, eda_AV.ipynb). Wrote data overview and data processing section of report, helped refine other sections.

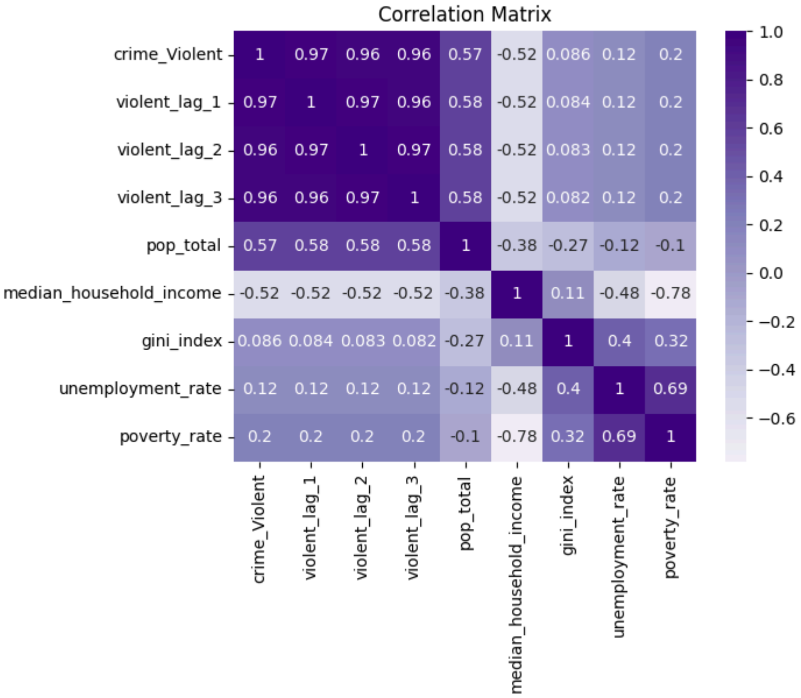
Appendix



Citywide crime incidents total for training set. Shows seasonal trends with higher crime rates in summers, lower in winter



Monthly crime incidents in raw LAPD data download, showing sharp decline after reporting standards changed.



	crime_Violent	violent_lag_1	violent_lag_2	violent_lag_3	pop_total	median_household_income	gini_index	unemployment_rate	poverty_rate
crime_Violent	1	0.97	0.96	0.96	0.57	-0.52	0.086	0.12	0.2
violent_lag_1	0.97	1	0.97	0.96	0.58	-0.52	0.084	0.12	0.2
violent_lag_2	0.96	0.97	1	0.97	0.58	-0.52	0.083	0.12	0.2
violent_lag_3	0.96	0.96	0.97	1	0.58	-0.52	0.082	0.12	0.2
pop_total	0.57	0.58	0.58	0.58	1	-0.38	-0.27	-0.12	-0.1
median_household_income	-0.52	-0.52	-0.52	-0.52	-0.38	1	0.11	-0.48	-0.78
gini_index	0.086	0.084	0.083	0.082	-0.27	0.11	1	0.4	0.32
unemployment_rate	0.12	0.12	0.12	0.12	-0.12	-0.48	0.4	1	0.69
poverty_rate	0.2	0.2	0.2	0.2	-0.1	-0.78	0.32	0.69	1

Correlation matrix included socioeconomic fields from ACS data. Preliminary correlation analysis on ACS socioeconomic fields.

Related work

- Predicting the Probability of Crime Related Danger in Los Angeles
  This paper attempts to utilize multiple machine learning algorithms to accurately estimate the danger of crime by area within Los Angeles.
- Prevention Is better than cure: Predicting violent crime in US counties using machine-learning methods
  This project evaluated multiple ml models including Linear Regression, KNN, and Gradient Boosting(XGB) on data from over 2100 US counties.
- Using Machine Learning Algorithms to Analyze Crime Data
  This research focuses on predicting violent crime specifically in the state of Mississippi.