Dig Into Crimes in Los Angeles

Yiran Liu, Ximiao Li, Kewen Li

Rackham Graduate School, University of Michigan

Background

Los Angeles, often referred to as the 'City of Angels,' is one of the most diverse cities in the United States. With a rich cultural tapestry and dynamic urban landscape, Los Angeles has faced unique challenges related to crime and public safety. Understanding and analyzing crime patterns in the city is essential for effective law enforcement, community engagement, and the well-being of its residents.

Data

In the project, we'll use **'Crime Data from 2020 to Present'** from Kaggle. This data is a publicly available dataset from the Los Angeles Open Data Portal that reflects crime incidents in the city of LA from 2020 to October 2023. The data is detailed with **crime case time**, and **location**, as well as **the victims' basic information**, and **the weapon of the perpetrator**. We did some data cleaning and output the final data as Python pickle files to support data analysis and time series forecast afterward. Detailed code will be available on **GitHub**.

Exploratory Data Analysis

After data preprocessing, we perform exploratory data analysis to get an overview of the dataset. Detailed Exploratory Data Analysis code will be available on **GitHub**.

- **Distribution of the crimes in different areas**: The geographic plot1 shows the distribution of caseloads in the 21 districts under the jurisdiction of LA.
- **Number of victims in different regions by sex/Ethnicity**: The plot displays victim differences between regions, suggesting potential regional patterns.
- **Seasonal decomposition of crime counts**: The plot reveals a curvilinear trend and a yearly seasonal pattern with non-constant residuals over time.

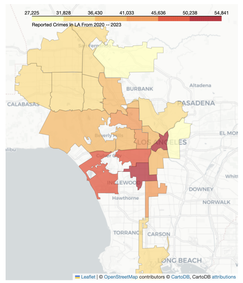

Figure 1. Distribution of the crimes in different areas

We discovered significant regional crime patterns and trends over time of after EDA. Therefore, we decided to split our tasks into two parts.

Similarity Analysis on the regional Distribution of Victims

Los Angeles exhibits a diverse population distribution with varying densities across the city. Studying the geographical distribution of victim characteristics can assist residents in enhancing their awareness of prevention measures and enabling law enforcement to adopt more targeted area management.

- **Generate data**. Victim's age, sex and ethnicity information from preprocessed data are used in this section. Specifically, we calculate the proportion each feature of different categories within each area:
  - Sex: Male/Female
  - Age: 0-8/8-31/31-45/45-120 (Quartile)
  - Ethnicity: Asian/Black/White/Hispanic/Others
- **Pairwise similarity analysis**. Using cosine similarity and Pearson correlation coefficients.
- **K-means clustering**. 3 clusters are output in our final model.

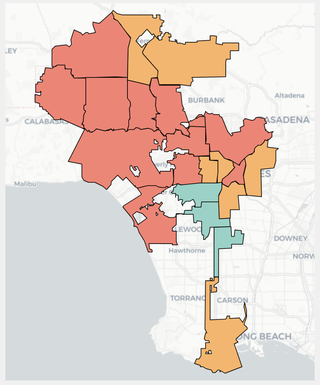

Figure 2. Clustering by Vict's Age, Gender and Ethnicity

Victim Persona

- **Red Region**: Male, Hispanic/White, Age 45 - 120
- **Orange Region**: Female, Black/Hispanic, Age 8 - 31
- **Light Blue Region**: Male $\gtrsim$ Female, Hispanic, Age 8 - 31

Time Series Analysis

In this section, we will conduct a **Time Series Forecast** by building a model to predict future trends in the number of crimes in the Los Angeles area to help inform the LAPD's future police scheduling.

We chose the **Autoregressive** and **Moving-Average** models as baseline models to check their performance on original and differenced data, respectively. Then, we will also evaluate by examining the ACF and PACF to determine whether to use the AR model, the MA model or the **ARMA** model.

1. **The differencing at different orders**. When differencing at 2nd order, we got the lowest P-value on the Dickey-Fuller test. So, the 2nd order is the differenced data to build the model.
2. **Autocorrelation Function Plot & Partial Autocorrelation Function Plot**.
3. **Model Evaluation**. Evaluation of the models is done through RMSE scores and AIC scores to determine the final model to perform the prediction task.

These will be the main steps to make subsequent predictions. After model evaluation, we chose the best-performing model for time series forecasting to simulate trends in the number of crimes in Los Angeles.

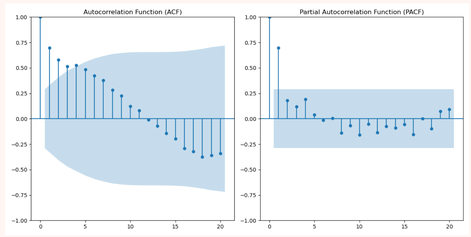

Figure 3. ACF & PACF Plots

By observation, we see that both images exhibit **tail-off**, which means that **they decay to zero asymptotically**. This suggests that using the ARMA model seems to give better performance. To test this conjecture, we set **d to 2** because of the lower P-value in differencing at 2nd order and then **iterated with p and q values** set in the range of 1-10 to find the best combination to ensure the best performance of the ARMA model. The results of the models are shown below.

Model	RMSE	AIC
AR(1) on original data	59.91	330.878
MA(1) on original data	60.10	360.238
AR(1) on 2nd order data	70.01	337.746
MA(1) on 2nd order data	70.63	322.864
ARIMA(1, 2, 4)	**55.12**	**316.602**

Table 1. Models Evaluation

By comparison, we can find that the **ARMA model with p = 1 and q = 4** has the **lowest AIC scores after iterations**, and it also has the **lowest RMSE** when compared to the baseline model. Therefore, we will choose **the ARMA model as the final model for the time series forecasting task**.

Time Series Forecast

The detailed modeling code is hosted on **GitHub**.

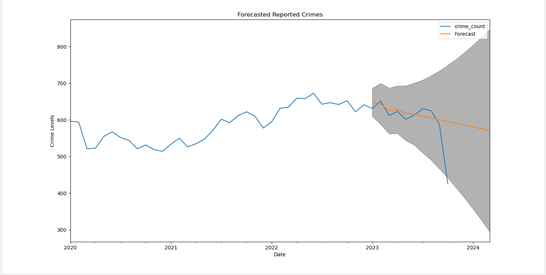

Figure 4. Time-Series Forecasting Results

The blue line in the image shows the actual crime data over time, the orange line shows the results predicted by the ARMA model. The plots show that the total number of crimes in Los Angeles has been steadily trending downward month by month over time.

Summary

Based on the results of our analysis and projections, we can provide some suggestions for the LAPD to consider:

- Redistribute police resources to maintain stability and security in the LA area, with more police presence in high-crime areas such as Central, 77th Street, and Pacific.
- Society and government should pay more attention to the development of the Black and Latino communities as they are the main victims in violent areas.
- If the government can provide more historical data, the prediction will be more accurate and effective, and it will also be a better reference for policymaking.