Dig Into Crimes in Los Angeles

GitHub Repository: https://github.com/yiranelmo/Crime_Data

Ximiao Li
ximiaoli@umich.edu

Kewen Li
kewenl@umich.edu

Yiran Liu
elmokkkk@umich.edu

1 Introduction

Grand Theft Auto V is one of the most well-known games after 2000. The main reason is that it vividly depicts life based on Los Angeles. In Los Angeles, like many other major metropolitan areas, a large number and types of crimes are happening every day. Exploring and understanding the patterns, causes, and temporal development of crime is not only essential for law enforcement agencies but also for policymakers, sociologists, and researchers aiming to create safer communities.

Our work aims to shed light on the nuances of criminal activities, their spatial distribution, temporal variations, and sociodemographic influences. Using crime data from 2020 through October 2023, we performed regional clustering on victim distribution and a time-series analysis of the crime volume, which provided us valuable insights into effective law enforcement, community engagement, and the well-being of its residents.

2 Related Work

Crime analysis is a fundamental component of contemporary policing strategies, focusing on proactive problem-solving and informed decision-making, as noted by Hinkle et al., 2020[1]. Crime analysts employ various tools such as calculations, databases, and computer systems to interpret data. They identify criminal trends and patterns, target high-risk areas and individuals, and optimize the allocation of police resources, ultimately working toward reducing overall criminal activity (Fyfe et al., 2017[2]).

Different techniques are applied to obtain a deeper understanding of crimes and provide a solid foundation for society's future improvement. For instance, crime maps play a pivotal role in identifying hotspots, indicating areas with a disproportionate number of crimes (Boba Santos and Taylor, 2014[3]). This analytical technique heavily relies on Geographic Information Systems (GIS) technology that requires precise information and data to accurately depict crime clusters (Kumar and Chandrasekar, 2011[4]).

Among the police departments that implemented preventive measures grounded in analytical products and data, a number of major cities in the United States achieved substantial reductions in crime rates within specific areas (Perry, 2013[5]).

3 Dataset

3.1 Crime Data

For this project, we will use "Crime Data from 2020 to Present" on Kaggle. This data is originally a publicly available dataset from the Los Angeles Open Data Portal that reflects crime incidents in the city of Los Angeles since 2020. Since the data were transcribed from paper versions of the original crime reports, there may be some errors in the data.

The data contains 28 columns and 815883 rows detailing the time, location (latitude and longitude), and type of crime, as well as the victims' basic information, and weapon of the perpetrator. This data contains many variables on location, which in combination with the time variable can provide a lot of valuable reference information. While this data is essential for understanding trends and patterns in criminal activity, it can provide many valuable references for informed decision-making on crime prevention and intervention strategies.

The following table 1 shows the names of some of the important columns in the dataset along with a description.

Column Name	**Description**
DR NO	Record Number
Date Rptd	Incident Reported Date MM/DD/YYYY
DATE OCC	Incident Occurred Date MM/DD/YYYY
TIME OCC	Incident Occurred Time (24 hour military time format)
AREA NAME	21 Geographic Areas
Rpt Dist No	A Four-digit Code Represents Sub-area
Crm Cd	Type of Crime Committed (Code)
Crm Cd Desc	Type of Crime Committed (Description)
Vict Age	Victim's Age
Vict Sex	Victim's Sex
Vict Descent	Victim's Ethnicity
Premis Desc	Where the crime occurred
Weapon Used Cd	Weapon Used Code
Weapon Desc	Weapon Used Description
Status Desc	Status Of Incident
Cross Street	Name of Cross Street
LAT	Latitude
LON	Longitude

Table 1: Columns Description

3.2 Geographical Data

We also used LAPD Division geographic data from the City of Los Angeles Hub for exploratory data analysis to better present the results. The data contains geographic location information for 21 areas under the jurisdiction of Los Angeles. Visualization data geographically can be very intuitive to display the distribution of the number of crimes in the Los Angeles area across the 21 districts under the jurisdiction of the city of Los Angeles and capture useful and specific information on different areas.

3.3 Data Preprocessing

Since this data has many literal descriptive variables and will not be used in subsequent analyses, we removed them first. Initial checks were then performed on each column individually against the dataset description file to determine if there was a difference in the description. We have organized the data and filled in the missing values. The detailed procedure code has been uploaded to GitHub. Finally, we stored the organized data as pickle files to save disk space and further use.

4 Exploratory Data Analysis

We perform exploratory data analysis to get an overview of the dataset. By dividing the features into four parts: crime, victims, criminals, and seasonal decomposition, it is easier for us to capture the patterns and trends. Detailed Exploratory Data Analysis code will be available on GitHub.

4.1 Crimes

As shown in Figure 1, even though the number of crimes in Los Angeles is numerous, they are not spread evenly. The majority of crimes occurred around central areas or downtown nearby. The further the area is from the central, the fewer crime cases occurred.

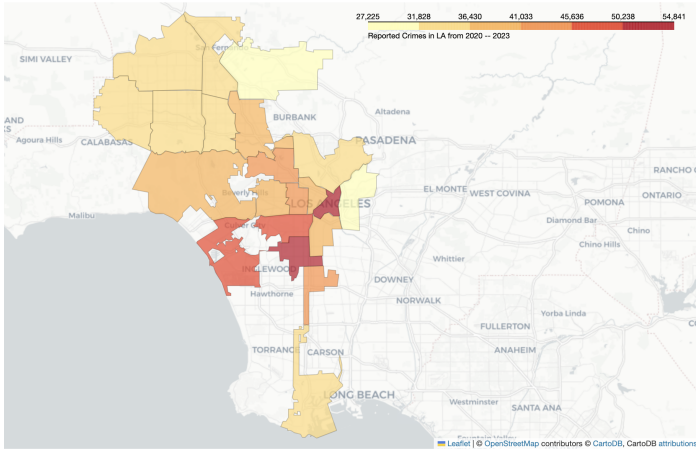

Figure 1: Number of Crimes in Different Areas of Los Angeles

4.2 Victims

As Shown in Figure 2, among all of the recorded victims, the number of female victims is an average of 5,000 more than male ones in most areas.

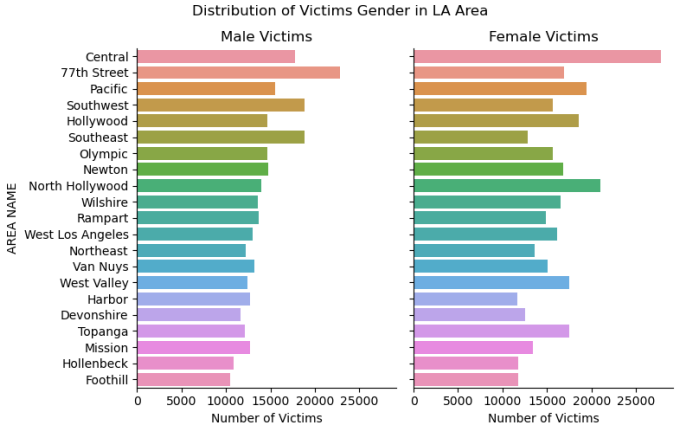

Figure 2: Distribution of Victims Gender in LA Area

According to Figure 3, the black race is more likely to suffer more than other races from crimes. Besides, different areas exhibit distinct patterns. Some areas such as 77th Street have the most Hispanic victims whereas little Hispanic victims in other areas.

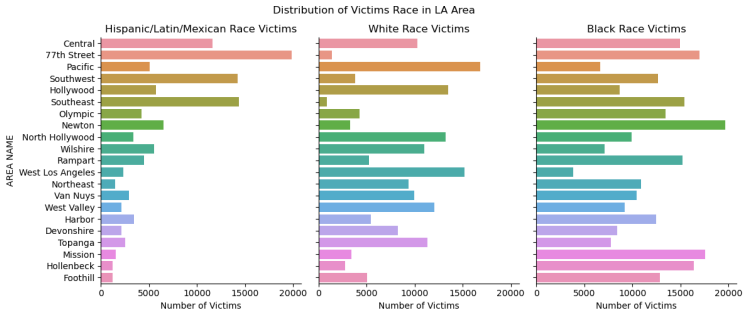

Figure 3: Distribution of Victims Race in LA Area

In Figure 4, the histograms show victims' age regardless of race as well as incident occurred time. The victims are generally kids and adults. And the most frequently time frame of the crime could be in the afterboon and at night.

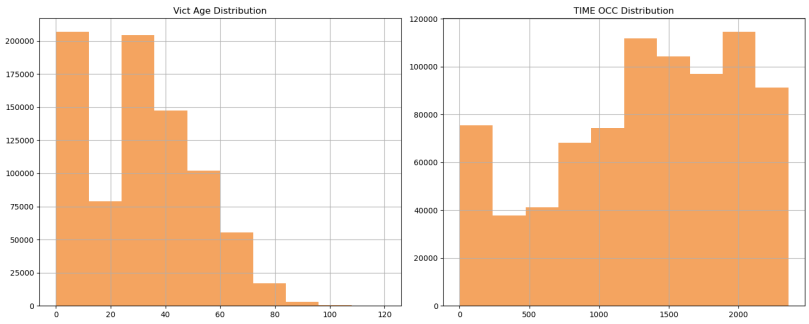

Figure 4: Histogram of Victim Age and Time Occurred

4.3 Criminals

In Figure 5, Strong-Arm, which refers to hands, fist, feet or bodily force, are used overwhelmingly by the criminals compared to other weapons. Interestingly, lots of unknown weapons are appeared in the records.

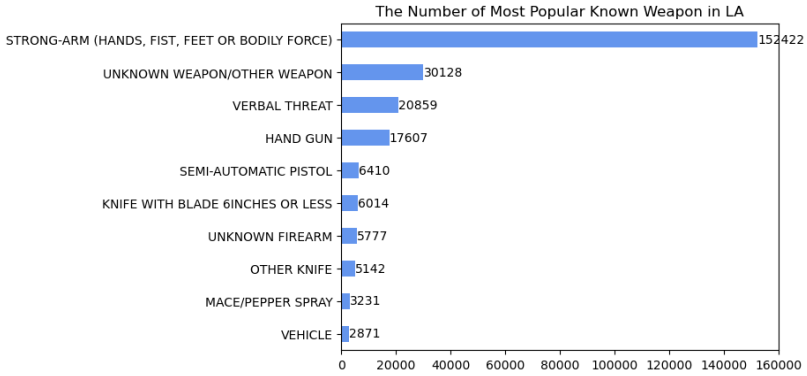

Figure 5: Number of Different Known Weapon Was Used

4.4 Seasonal decomposition

We decomposed the number of crimes over time using seasonal decomposition to explore potential hidden seasonal patterns. The results are shown in Figure 6:

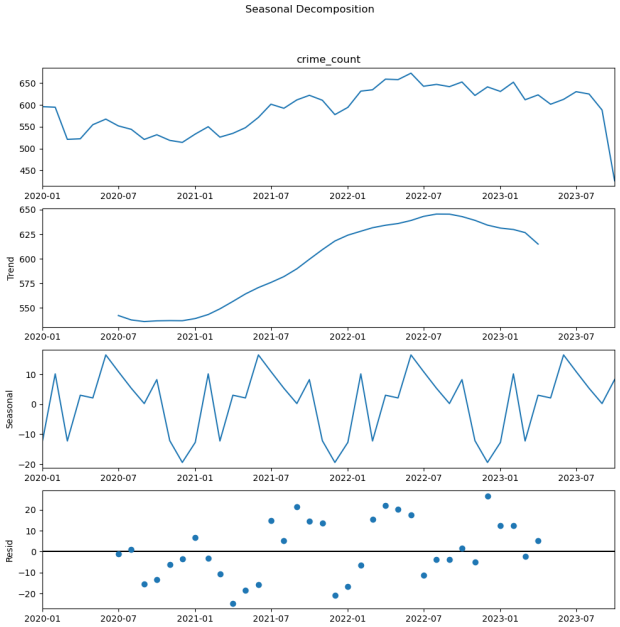

Figure 6: Seasonal Decomposition

We can conclude the findings for each panel:

1. **Top Panel - Observed Data**: This panel shows the original data, which is the count of crimes spanning from January 2020 to October 2023. The crime count fluctuates over time but does not show a clear upward or downward trend within this timeframe.

2. **Second Panel - Trend**: This panel extracts the underlying trend from the data, smoothing out the seasonal fluctuations and irregular components. The line presents a curvilinear shape in general. The trend line starts relatively flat and begins to rise around early 2021, reaching a peak around early to mid-2022. After the peak, it shows a slight decline, suggesting that the overall crime counts was increasing for a period before starting to decrease.

3. **Third Panel - Seasonal**: This panel shows the seasonal component, which captures the regular pattern that repeats over time. In this case, there is a clear seasonal pattern that repeats annually. Peaks often appears in June to August, indicating summer is the specific time of the year that a substantial number of crimes occurred.

4. **Bottom Panel - Residual**: The residual component represents the randomness in the data after the trend and seasonal components have been removed. The dots appears to be randomly scattered around the zero line, with no apparent pattern. This indicates the seasonal and trend components have captured most of the systematic information in the data.

Based on these four types of analysis, we discovered significant regional crime patterns and trends over time. Therefore, we decided to split our tasks into two parts: similarity analysis on the regional distribution of victims and time series forecast on crime counts.

5 Methods

5.1 Similarity Analysis on the Regional Distribution of Victims

Los Angeles exhibits a diverse population distribution with varying densities across the city. The downtown core has high population density, while suburban areas offer lower densities. The city's ethnic and cultural diversity is reflected in its neighborhoods, with distinct characteristics and economic disparities.

With such a dynamic and multifaceted population distribution pattern in Los Angeles, studying the geographical distribution of victim characteristics can assist residents in enhancing their awareness of prevention measures, while also enabling law enforcement to adopt more targeted area management practices.

5.1.1 Data Processing

We use the victim's age, sex, and ethnicity information from the cleaned dataset for our analysis. For the victim's age variable (Vict Age), we first calculated the quartiles for all the data and then determined the proportion of victims falling into different quartile age groups within each area. For the victim's sex variable (Vict Sex), we excluded data with unknown gender and calculated the male and female ratio within each area. For the victim's ethnicity variable (Vict Descent), various ethnicities were reclassified into 5 distinct categories: Asian, Black, White, Hispanic, and Others. Then, we derived the proportion of each category within

a different area of Los Angeles. Finally, we combined the above 3 parts together and indexed them by area (AREA NAME), establishing our data for future clustering.

5.1.2 Methodology

We initially attempted pairwise similarity analysis, utilizing both cosine similarity and Pearson correlation coefficients. After plotting the cosine similarity matrix in a heatmap (Figure 7), we discovered that solely exploring pairwise correlations was insufficient to learn about the geographical distribution of different victim groups, so we applied K-means clustering on different areas to investigate victim pattern similarities among regions.

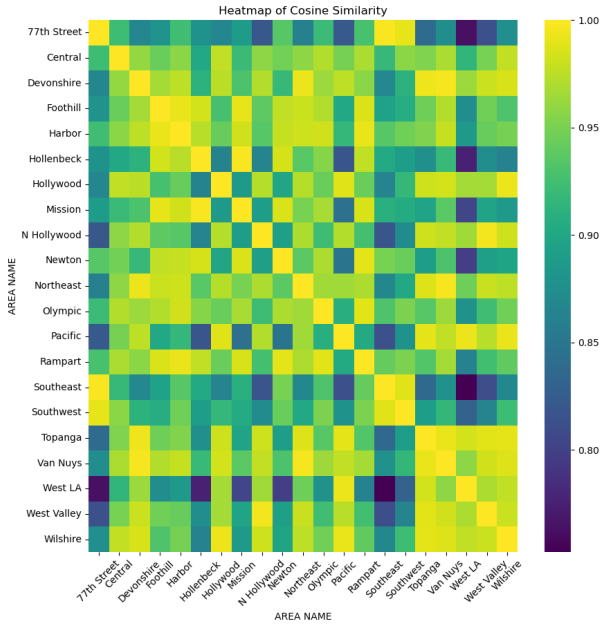

Figure 7: Heatmap of Cosine Similarity

K-means clustering is a widely used method for cluster analysis that aims to partition a set of objects into K clusters in such a way that the sum of the squared distances between the objects and their assigned cluster mean is minimized. We initialized our output number of clusters from 1 to 10 and applied k-means to our data respectively. An elbow occurred in the scree plot when the number of clusters was set to 3, so we chose 3 clusters as our final output.

The areas are aggregated by victim patterns as follows:

- Cluster 1: Foothill, Harbor, Hollenbeck, Mission, Newton, Olympic, Rampart
- Cluster 2: Central, Devonshire, Hollywood, N Hollywood, Northeast, Pacific, Topanga, Van Nuys, West LA, West Valley, Wilshire
- Cluster 3: 77th Street, Southeast, Southwest

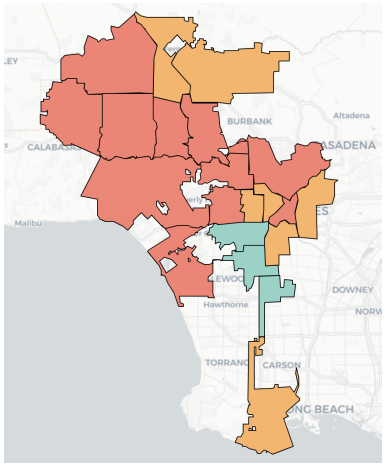

Figure 8: Clustering by Vict's Age, Gender and Race

We denoted clusters 1, 2, and 3 in Figure 8 using orange, red and green colors respectively. Moreover, we can infer the victim features from the cluster centroids. It indicates that victims within cluster 1 are mainly black and Hispanic females, aged between 8 to 31 years old. Hispanic and white males aged 45 to 120 are more likely to be victimized in areas of cluster 2. For cluster 3, Hispanic people aged 8 to 31 are the main target group of victims.

5.2 Time Series Forecasting on Crime Volumn

Forecasting the crime counts that would possibly occur over time in the future would also be a core issue in crime analysis. This enables law enforcement agencies and policymakers to anticipate and respond to changes in crime volume by allocating resources effectively. Furthermore, time series forecasting aids in evaluating the impact of past interventions and assessing the efficiency of crime reduction programs, enabling data-driven decision-making for a safer community.

5.2.1 The Differencing at Different Orders

Before modeling the time series data, we first look at different orders for differencing to identify the better-performing order and then let it be compared with the original data in the baseline modeling stage. The result for different orders is shown in Table 2.

Order	P-value
Period 1	0.004051
Period 2	0.002753
Period 3	0.968683
Period 4	0.845303

Table 2: Different Orders for Differencing

The p-value of the Dickey-Fuller test is minimized when period = 2, so period = 2 is used as the differential time series.

5.2.2 Autocorrelation Function and Partial Autocorrelation Function

After that, we looked at the Autocorrelation Function and Partial Autocorrelation Function for the raw data to determine how to select a model. By observation, we see that both images in Figure 9 exhibit tail-off, which means that they decay to zero asymptotically. This suggests that using the ARIMA model seems to give better performance.

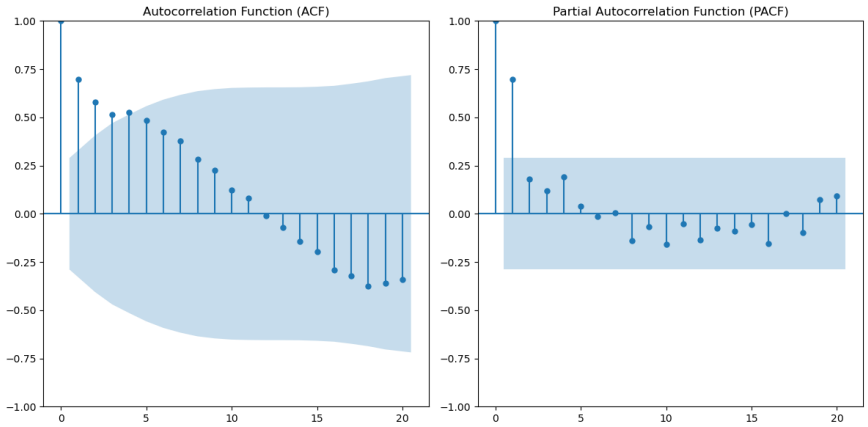

Figure 9: ACF & PACF Plots

5.2.3 Model Evaluation

Our team chose the Autoregressive model and Moving-Average model as the baseline model and looked at their performance on the original time data and the 2nd-order data. Based on the autocorrelation function and autocorrelation function it was concluded that using the ARIMA model seems to give better performance. To verify the conjecture, set the parameter 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 ARIMA model. The results of the models are shown in Table 3.

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 3: Models Evaluation

By comparison, we can see that the ARIMA model has the lowest AIC score after iteration when p = 1 and q = 4 and it also has the lowest RMSE compared to the baseline model. Therefore, we will choose the ARIMA model as the final model for the time series forecasting task to predict the trend in the number of crimes in the Los Angeles area over time.

5.2.4 Forecasting Result

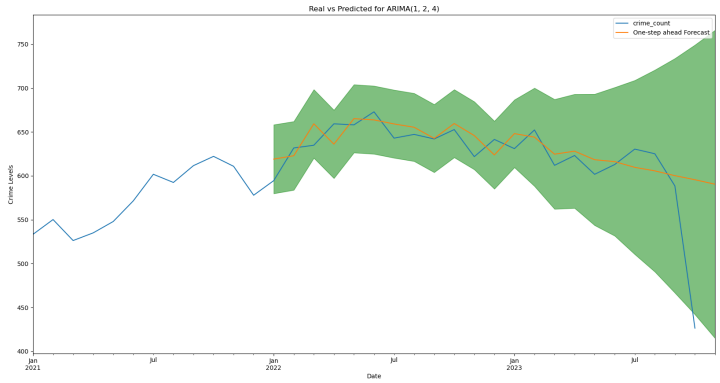

Figure 10: Real vs Predicted for ARIMA(1, 2, 4)

From Figure 10, It can be noticed that the predicted values(the Orange Line) are very close to the observed values(the blue line) and also correlate with the trend and seasonality of the time series.

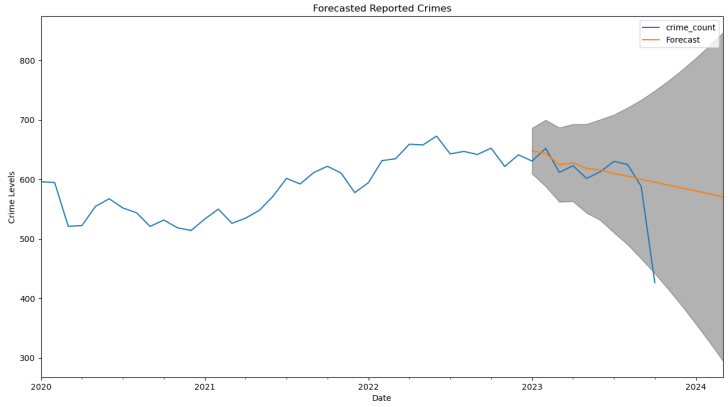

Figure 11: Forecasted Reported Crimes

From the above prediction result plot 11, as the step increases, the predicted values(the Orange Line) gradually deviate from the observed values(the blue line) and the confidence intervals in the gray area also increase.

6 Conclusions

Based on the results of our analysis and projections, we gathered valuable insights about criminal activities, their spatial distribution and temporal variations. Therefore, we wish to share the information with LAPD to make Los Angeles communities better.

- 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.
- The government may allocate more police resources to deter and manage the rise in criminal activity, and citizens need to heighten their vigilance during summer. The police department could disseminate this message and offer practical safety tips to help the public safeguard themselves and their property more effectively during this high-risk period.

7 Limitations

- We only collected data from 2020 to the present, which contains a three-year-long period. However, including more data points or a longer time frame is important for us to make a more accurate time series forecasting.
- The data features are limited, we have a limited understanding of the criminals. A few factors are important to measure why citizens choose to commit crimes, such as their income, education level, and condition of mental health. These features can provide more insights into our research question
- As we discussed earlier, we have both categorical and numerical data, a more appropriate or direct way is to use the kprototypes algorithm, which is an updated version of the K-means algorithm that can be applied to categorical and numerical data. However, the method is inapplicable due to its low efficiency and high computational cost.
- As pointed out by LAPD's official website, the data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data

8 Future Work

- We will collect previous or new data points to enlarge our data set. Meanwhile, researching about criminals and their purpose and adding those to the data is in the plan.
- We will spend appropriate time on the cloud platform to train original kprototype models and speed up by using higher computational machine.
- We might look for a better algorithm to improve our time series forecasting.

References

[1] Joshua C Hinkle, David Weisburd, Cody W Telep, and Kevin Petersen. Problem-oriented policing for reducing crime and disorder: An updated systematic review and meta-analysis. *Campbell systematic reviews*, 16(2):e1089, 2020.

[2] Helene Oppen Gundhus, Kira Vrist Rønn, and Nick Fyfe. *Moral issues in intelligence-led policing*. Routledge, 2017.

[3] Rachel Boba Santos and Bruce Taylor. The integration of crime analysis into police patrol work: Results from a national survey of law enforcement agencies. *Policing: an international journal of police strategies & management*, 37(3):501–520, 2014.

[4] M Vijaya Kumar and C Chandrasekar. Gis technologies in crime analysis and crime mapping. *International Journal of Soft Computing and Engineering*, 1(5):115–121, 2011.

[5] Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.