Unveiling Crime Patterns and Predicting Crime Incidents in LA: Integrated Statistical Analysis

Type	Detail	
Group Number	9	
Group Member	**Name**	**ID**
	DAI Li (Group Leader)	1155199260
	LUO Shilong	1155205224
	HUANG Weiyu	1155208079
	FENG Jinhan	1155206031
	JIANG Xijie	1155202866
	YUE Kun	1155208084

I. Introduction

Crime, which is a chronic problem in society, has large numbers, complex causes, and far-reaching effects. As a result, there is a wide range of research aimed at identifying patterns in criminal behavior to prevent or mitigate it. In the United States, the crime rate has remained high for long, which has a great impact on the American and even the world society. Meanwhile, the U.S. government also categorizes and publishes relevant social data, including statistics on crime in various regions.

Based on this background, we divided the "criminal behavior pattern and prediction" into four sub-problems, hoping that through the representative (LA City) data, we may draw conclusions for these problems by doing systematical analysis, and to contribute to the study of society.

II. Data Collection and Preprocessing

Data Source: The data we use is from the official platform of the US government(https://data.gov/), including 258,498 datasets from various aspects in the US. The detailed dataset is the Crime_Data_from_2020_to_Present.csv of Los Angeles, containing about 800 thousand crimes happening in LA from 2020 to Oct 23/2023 (the data is keeping updated), and detailed information of each crime. To conduct the analysis, the main attribute we used is as follows:

Attributes	Meanings
Date	Date occurred and reported, we choose the former
Area	The zone where crime takes place
Victim_Info	Attribute includes age, sex, descent of victims
Weapon_Code	Weapon type of the crime
Time_Occured	Specific time of crimes
Reporting District	A set of numbers of each district
Crime_Code	The type of crime
LAT & LON	Detailed coordinate of crimes

Preprocessing Procedure: This dataset is adequate and detailed, and the data structure is also clear to do the analysis directly. Thus, the preprocessing process is simple to conduct: Firstly, for the various attributes, we change the abstract title into clear statements with a higher readability as is listed above.

Besides, by describing and selecting the extreme value (such as latitude and longitude has extreme value 0, victim age has extreme value) from some of the columns, we discard the invalid values. Because each problem analyzes different objects, special data preprocessing is also taken in each problem. After preprocessing, we dive into 4 questions as following.

III. Data Analysis

Q1: Major Components of Crimes and Attributes of Victims

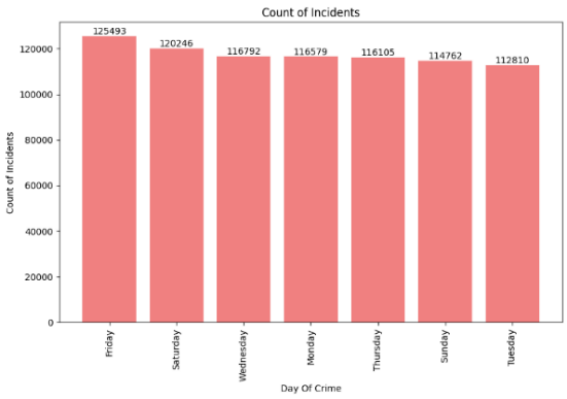
In this question, we will mainly conduct descriptive analysis from the dataset. The main methods include visualization based on modules in python.

For all the characteristics we choose, we have visualized:

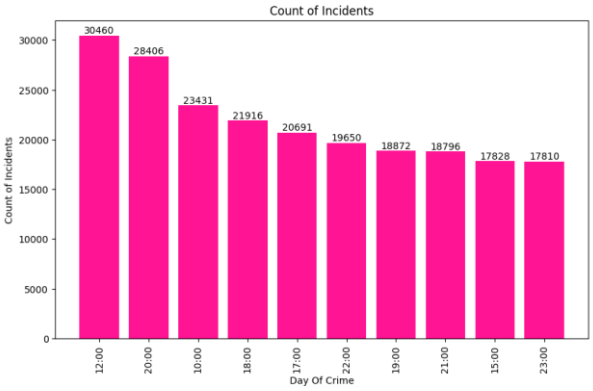
- Time of crimes, including the month, date, weekday, datetime of time happening.
- Victim information, including sex, age, descent, type of crime related to victim sex and age. We also include crime premises and the analysis of weapon types.
- Places of crimes, including distribution of crimes in different districts, mapping of the crime spots, which includes heatmap of LA city, report district map and LA district map of crime distribution.

For each point, we may display 1-2 representative images (Pic 3.1,3.2), others will be put into the appendices.

1. Time Analysis of Crimes



(Pic3.1 Count of Incidents [By day of the week])



(Pic3.2 Count of Incidents [By date])

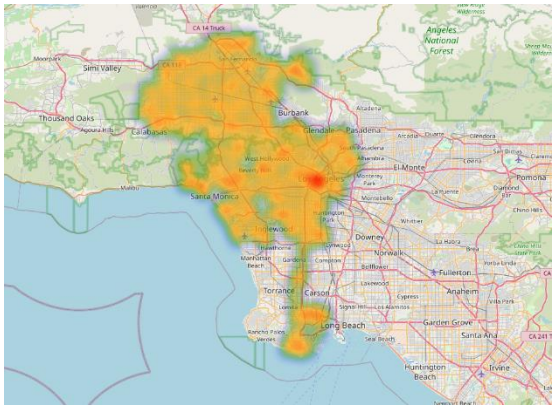
We can draw the conclusion of many phenomena that mostly matches our guess: Friday is the most frequent happen day and the least Tuesday is slightly lower; Among all the months, Dec and Nov have apparently low count of incidents than the others which is near the same. Complete analysis of each graph is included in the code file. However, some of the statistics seem to be surprising as it is displayed in the image that the crime at 12p.m. is significantly higher than other times, which may be due to the problems of the data source of LAPD.

2. Victim and Crime Information Analysis

According to all the output, we can draw the following conclusions: the Male victim number is slightly higher than female. Since most victims fallen on age 25-30, we investigate the crime types of this age group. It can be discovered that the main part is simple assault, which is reasonable for their impulsive personality. The most common crime is Vehicle Stolen happening on the streets, other crimes related with vehicles are also common, which are interesting facts compatible with the famous video game GTA. For the descent of the victim, Asians like Chinese and Japanese are the least common victims, and Hispanic, Latin and Mexican are the most common race of victims, which is most likely because there is a significant difference in the number of people of different races in LA.

3. Place Analysis and Mapping of Crimes

By using folium module to map the points on the open-source OpenStreetMap, we can see the spot location of the crimes directly. Also, on the district map, we can see the number of crimes of a district by moving pointer on that district. The main spot for crime is in downtown LA, where the main population of LA lies. Also, some spots of West Hollywood and Southern part near airport also have a higher rate of crimes.



(Pic3.3 Hotspot1[LA])

(Pic3.4 Hotspot2 [LA])

Q2: Predict the number of future crimes based on past data

This study was analyzed using the ARIMA model, which is constructed through the attributes of the cases themselves and time, and is used to predict the number of cases that will occur at a certain time in the future.

1. Data pre-processing

a. Smoothness Processing

By observing the time series plot of this data, it is considered that the data has an upward growth trend, so it is initially judged that the series is not a smooth time series. The unit root test is performed on the data, and the test result shows that the P value is less than 0.05, which is considered that the original data is non-smooth and needs further differential processing.

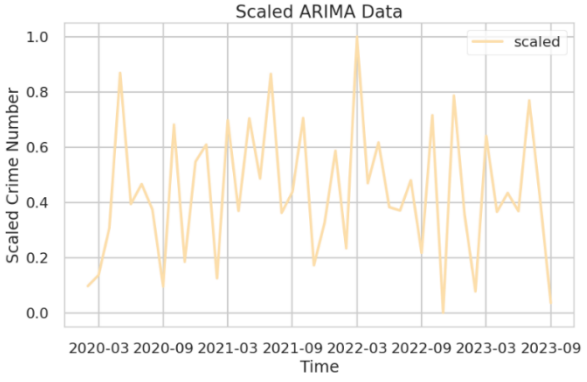
The series is differentiated to make it a smooth series. After the first-order differencing, it was observed that there was no obvious linear trend in the time series plot, and it was considered that the time series had smoothness and model selection could be carried out.

b. Normalization

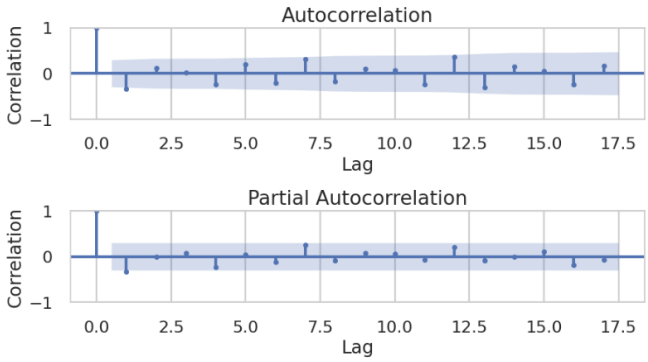
Because the case attributes have various values with different proportions, it is necessary to standardize the time series after performing the smoothness test and differencing to facilitate the output of the final results.

2. Model setup

Observing the autocorrelation plot, the ACF value first gradually decreases and then increases, in the first half of the interval range (0~6), it is considered that there is a slight tail dragging situation, initially selected ARIMA model is (0,1,4). PACF value is smooth, autocorrelation has an obvious 2nd order truncated tail, in summary, set up the whole model as ARIMA (0,1,4)(0,1,2)[1].



(Pic3.5 Scaled ARIMA Data)



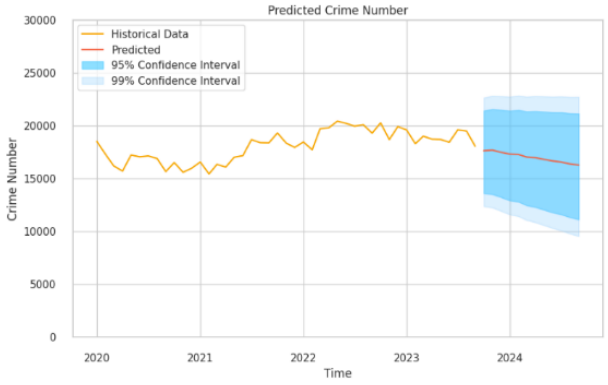
(Pic3.6 Autocorrelation & Partial Autocorrelation)

3. Fitting model

Different ARIMA models were selected for fitting, using great likelihood estimation, adjusting the model parameters to get different AIC values and taking the model with the smallest AIC value. Finally, the model ARIMA(0,1,4)(0,1,2)[1] with AIC value of 219.8 is selected and this result is also in line with the previous observation.

4. Making predictions

Based on the completed model that has been constructed, the prediction is made by substituting the previous months and the number of crimes data and setting the confidence interval, and the following prediction result is obtained. The prediction assumes that the number of crimes will decrease smoothly and gradually in the coming time, and there exists a situation in which the number first rises slightly and then falls back gradually.



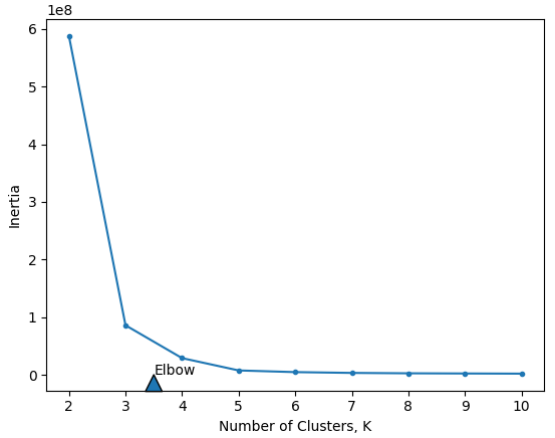
(Pic3.7 Predicted Crime Number)

Q3: Areas and Times of Highest Crime Frequency and Crime code classification

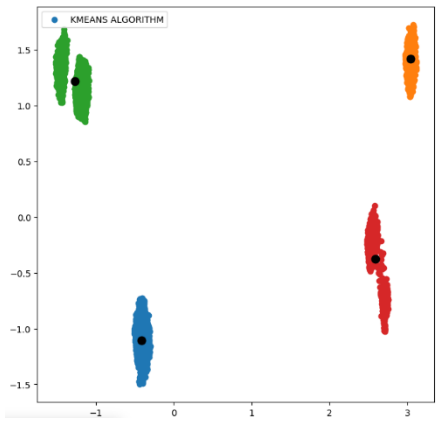
For this question, we aim at telling from the data of the time and location which have the highest rate of crime.

1. K-means with PCA

Firstly, we preprocessed data, including removing null values, using Zscore and Winsorize. Secondly, we decide 2 principal components for our analysis, which is more direct for viewing and matches the 2 dimensions: time and location of our prediction purpose. Thirdly, we use the elbow plot and Silhouette Score to determine the best cluster number for K-means (we now set it as 4). By doing K-means of the above setting, we gained the following results:



(Pic3.8 Predicted Crime Number)



(Pic3.9 Predicted Crime Number)

This output image interprets 4 main spots (centroids) of the 4 divided clusters. Roughly speaking, large inter-cluster distances and small intra-cluster distances are desirable clustering results. Specifically, we can classify crime data into 4 different clusters, and we could see the locations and time of crime where police have to do more patrolling from the centroids of each cluster.

Overall, result of K-means clustering analysis is that the number of clusters is 4 and the silhouette score is 0.9136435, which is a nice result.

2. Logistic Regression

After K-means clustering, we try to use Logistic regression for the prediction of crime code categories. In the initial attempt, our test set classification accuracy was difficult to improve at around 70, we guessed that it was because of the excessive variety of Crime codes in the dataset that made it difficult for the model to make classification predictions, so we selected the top 5 crime codes with the highest number of crime occurrences on the basis of the Q1 data visualization, and classified them for the study. The model we fit can get a nice result for the Accuracy score at 89.0%.

Furthermore, we try to use Regularization and Grid Search method to further improve accuracy score, but the improvement isn't apparent. The model accuracy of the test set is only improved by 0.5% to 89.5%. We speculate that since regularization is a way to optimize the model by reducing the model complexity, but we did not use complex model and many features, so it is difficult for regularization to bring significant improvement.

Q4: Predict crime position with specific crime attributes

In this study, the KNN method is used to categorize the victim attributes under different types of cases, which is used to predict the types of cases to which victims with certain attributes may belong.

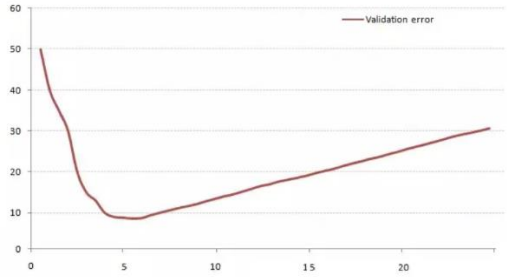
1. Data preprocessing

a. Normalization

Because different cases have different characteristics, and different characteristics of the value of the scale is different, so we need to normalize the sample so that different characteristics have different weights, and ultimately can be equal to the important. We use this formula to normalize: $M_j = \max_{i=1,\dots,m} x_{ij} - \min_{i=1,\dots,m} x_{ij}$

2. Model setup

Cross-validation was used in this study and the results are shown in Fig(Pic 3.10). The final result is that the K value is most effective when it is chosen as 5.



(Pic3.10 Cross-validation)

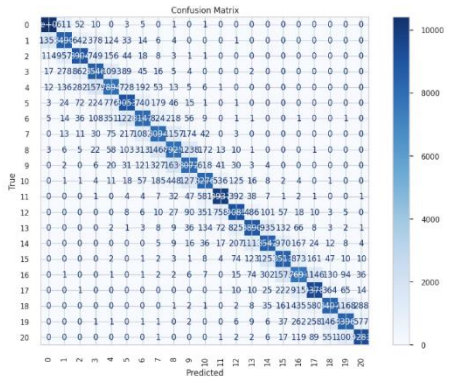
3. Prediction Evaluation

a. Confusion Matrix

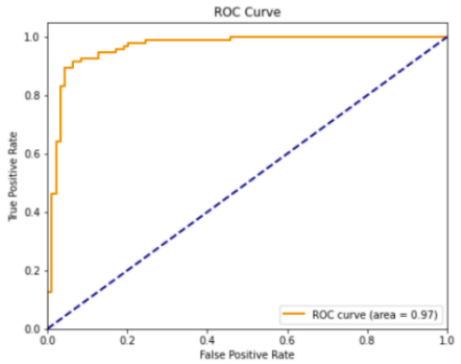
This study uses the confusion matrix to carry out the detection of the prediction model, and the final prediction graphs obtained are as follows, and the prediction effect can be known to be better according to the color of the legend.

b. AUC

The AUC was also used for this study and the final result value was about 0.97, which is close to 1, indicating that the model predicts the type of case it belongs to more perfectly for a larger number of threshold scenarios.



(Pic3.11 Confusion Matrix)



(Pic3.12 ROC/AUC)

IV. Discussion of the results.

By doing the above analysis on each 4 questions, we have achieved these results:

1. On the descriptive analysis, we summarize the features of crimes on dimensions as time, victims' information and the relationship between victim and crime type. We also show the frequency of crimes based on locations. Through these results, we can take preventive measures to mitigate the crime accordingly.

2. By constructing the ARIMA model based on the data, we make the prediction through the past data, which displays the trend of the number of crimes. This decreasing trend implies a good situation for future security work.

3. We use K-means with PCA to get the hotspot of areas and times for crimes. Through figuring out the principal components and clustering, we got the hotspot for crimes and an ideal silhouette score of 0.91364. We also use Logistic regression for the prediction of crime code categories. After optimizing the model, we got our best model with 0.893 accuracy score.

4. We use KNN to categorize the victim attributes under different types of cases, the model reaches a good result based on the prediction evaluation outputs of Confusion Matrix and AUC.

V. Conclusion & Future work

1. Conclusion

In this study, analyzing crime data in Los Angeles, CA from 2020 to 2023, we found that Friday is the most frequent happen day and Tuesday the least. Dec and Nov have apparently low count of incidents than the others. We also get the daytime features and the location distribution.

By using the ARIMA model, we analyze the time series data to predict a potential decline in crime rates. We also collect the hotspot data to demonstrate the crime concentrations by K-means and PCA algorithm. Finally, we use KNN method to categorize the areas of crime via attributes of crime and victims. A confusion matrix is used to explain the good outcomes of KNN.

2. Future Work

There is future work for the content and methodology of this study.

a. ARIMA Model: Add auxiliary method & data macro dimension.

ARIMA method for time series is useful but easily influenced by recent data, choosing other black-box methods such as machine learning could help to improve the accuracy.

The dimensions of research data are limited, add more macro data might help explain the causes of crime better.

b. Unbalanced samples: Use Undersampling & Oversampling

As the attributes of a certain type of crime are more unbalanced (the majority of victims in a certain type of case are female), it is necessary to do undersampling for an extremely large amount of data and oversampling for an extremely smaller amount of data

Appendices

I. Picture & Figure

Reference File Folder: Pic.&Fig.

II.Raw Data

Reference File: raw_data.csv

III.Programming Code

Reference File: programming_code.py