# SI 671: Analyzing Crime Data Patterns in Los Anngeles for 2022

ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Youheng Fu (youhfu@umich.edu)

## Introduction

- In this changing world, individuals, families, and communities all desire a sense of security in their lives.
- Los Angeles, one of the largest city in the United States, has long been plagued by the large amount and variety of crimes, which can serve as a classic example.
- To reflect the trend of crime rate's changing and make it significant to predict the future situation, I choose the crime data collected in year 2022 to analyze.

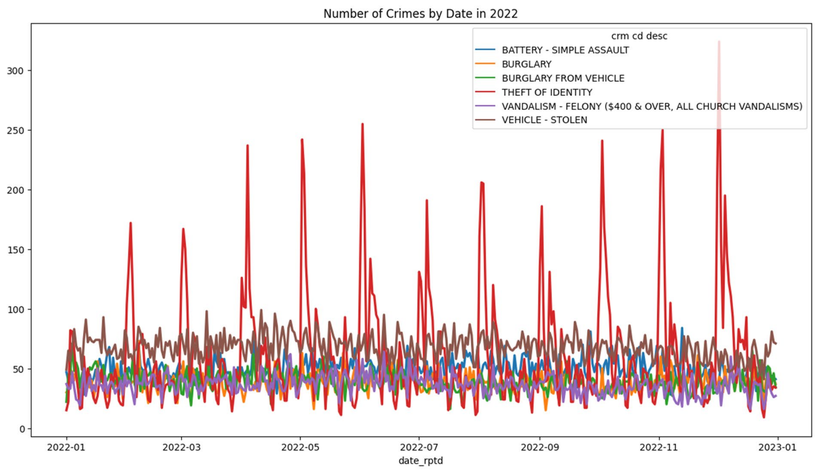
## Model/Problem set-up

Given the data records which contains the time, area, crime type, and other features of the case happened to construct time series and network analysis, I raised the following research questions:

RQ1: What are the seasonal patterns of the crime for different crime types?

RQ2: How can we predict the future crime incidents for different crime types and do they make sense?

RQ3: What are the areas shared with the most similar crime types and have high centralities?



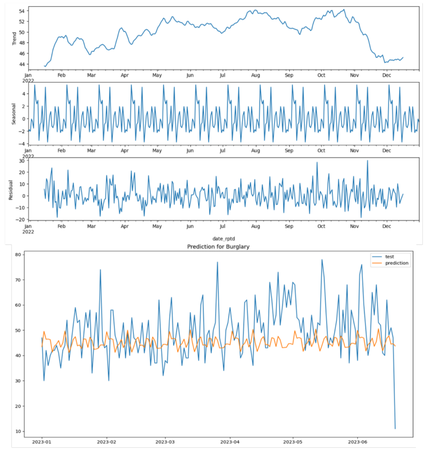

## Methodology

- In time series analysis part, we only utilize top 6 frequent crime types to fit our model. To analyze the seasonal patterns and trends for each crime types, we perform seasonal decomposition with additive model.
- We use seasonal ARIMA model to predict the future performance (we use the first five month data in 2023 to test). We first check the stationarity of the data and then use ACF and PACF to confirm the hyperparameters of our model.
- We construct the graph by setting the nodes be the areas in LA and the weights of edges be the similarities between areas. The similarities are calculated by Jaccard similarity formula with different crime types. We finally use K-means method to cluster the nodes

## Result (time series)

RQ1 : The figure1 indicates that there exist seasonal patterns in the crime data. The crime rate is high at the beginning of the month. LAPD should allocate more resources at that time.

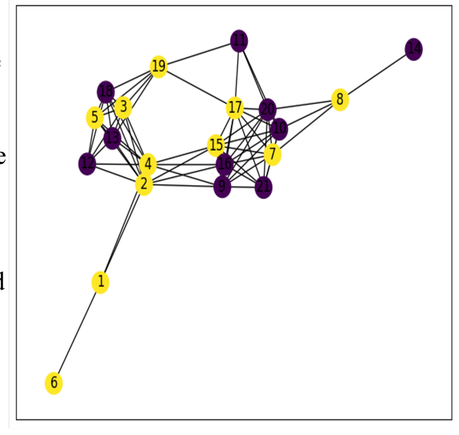
RQ2: The results show ARIMA model gives an acceptable prediction result except for data type "theft of identity". RMSE and MAE are relatively small for other models.




| | RMSE | MAE |
|---|---|---|
| BATTERY - SIMPLE ASSAULT | 3.33 | 8.45 |
| BURGLARY | 2.99 | 7.21 |
| BURGLARY FROM VEHICLE | 3.25 | 8.84 |
| THEFT OF IDENTITY | 7.13 | 33.19 |
| VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) | 2.83 | 6.42 |
| VEHICLE - STOLEN | 3.16 | 8.22 |

## Results (network)

RQ3 : Analyzing this similarity network, we find that the area 2 and 4 have the highest page rank, highest betweenness centrality and highest degree centrality. The nodes in the graph are clustered by centrality. The result indicates that area 2 and 4 have highly overlapped crime types comparing to other areas. The LAPD should allocate more experienced polices in these areas.




## Limitation and Future work

- The model ARIMA is still not so accurate for some data types, we may train a model utilizing deep learning like LSTM in the future analysis.
- We only use data in 2022. If more data are collected, the model can gain more information and improve the result.
- Some other features in the original dataset are not used.

## Significant References

- Data can be retrieved from the url:
- https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8.

COLLEGE OF ENGINEERING
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

# Analyzing Crime Data Patterns in Los Angeles for 2022

Youheng Fu (youhfu@umich.edu)

December 10, 2023

## 1 Abstract

Nowadays, people focus more on self-security as the crime rate still keeps at a level. This project centers on Los Angeles, a city grappling with a diverse array of crimes, and aims to scrutinize seasonal crime patterns, construct a predictive model for crime occurrences, and compare crime types across different districts. It utilizes a time series to find potential seasonal patterns in the crime data and uses a seasonal ARIMA model to predict the result. Additionally, a network analysis is applied to demonstrate the similarity of crime types. The research finally finds that there truly exist seasonal patterns in those crime types and the prediction model is relatively reasonable. Besides, the areas 2 and 4 have the highest overlapped amount of crime types. The LAPD can refer to this project to allocate police resources more efficiently. In the future study, we can use more advanced models like LTSM to improve the results. The code can be retrieved from the url.[1].

## 2 Introduction

Social security and stability have always been the focus of the people's attention. In this changing world, individuals, families, and communities all desire a sense of security in their lives. Crime rate, crime type, and other attributes related to crime are essential factors affecting social security and stability. Hence, analyzing them becomes a significant topic for understanding crime trends and developing strategies for crime prevention. Los Angeles, one of the largest cities in the United States, has long been plagued by the amount and large variety of crimes. As the crime rate trend is decreasing with fluctuation, the type and the location distribution of crime are increasing, which makes the situation more complex. In the project, the main goal is to analyze the seasonal patterns of crime cases and try to construct a prediction model for crime amount. Additionally, we will also try to compare the different crime types in different districts of LA. I hope the outcomes can help to deter criminality before it occurs and allocate police resources in different areas more effectively.

Given the background information and motivations to analyze the crime data patterns, the research questions of this project are shown below:

- What are the seasonal patterns of crime for different crime types?
- How can we predict the future crime incidents for different crime types and do they make sense?
- What are the areas shared with the most similar crime types and have high centralities?

## 3 Work Review and Data Source

Some researchers tried many different ways to predict and prevent crimes. Shah and Bhagat (2021) [1] achieved this by machine learning and computer version methods to recognize crime patterns. Kshatri et al. (2021) [2] claimed a classification work with the help of stacking machine learning models such as random tree algorithm, neural network, support vector machine, etc. Adrams (2020) [3] discussed more on the relationship between COVID-19 and crime rates. However, those researches do not focus on the pattern and changes of crime data in a specific region, which is what this project wants to figure out.

[1] https://github.com/youhfu/SI671-final-project

The dataset used in this project is from the official data recording website [4]. The dataset contains all crime records from January 2020 to May 2023. In the analysis part of the project, to make the information more similar and related to the current situation, we only use the data in 2022. Additionally, the dataset contains many features, including the crime reported date, the area where that crime happened, the gender of victims, etc. We will then only use crime types and dates in the time series analysis part and use areas and crime types in the network analysis part.

# 4 Methodology

## 4.1 Time Series Analysis

To figure out the patterns for different crime types, we need to determine which crime types to analyze. Here, we choose the crime types whose amounts are over 40,000 to be our research objects, since we can gain more information from them. After that, the histogram of crime types and the time series of them in 2022 are shown as follows:

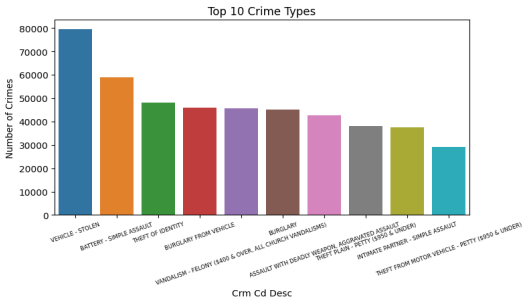



Figure 1: Top 10 amount crime types

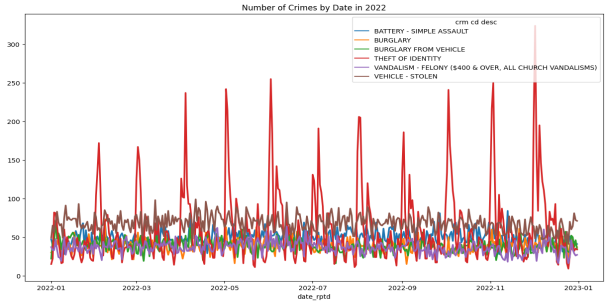



Figure 2: Time series of 6 crime types

To find the seasonal patterns, we use seasonal decomposition with the additive model here. The reason that we choose the additive model instead of the multiplicable model is that the time series graph gives us an intuition that the amplitude of seasonal fluctuation does not change a lot. We set the parameter "period" to 30 since we want to figure out the changing pattern of crime cases within a month.

For the prediction part, we first need to check the stationarity of our time series. We choose to utilize an augmented-dickey fuller (ad-fuller) test to show the stationarity. If we reject the null hypothesis, which means the p-value is less than 0.1, the time series is stationary. The results of p-values for different crime types are shown in the below table:

| Crime Type | p-value |
|---|---|
| Battery - simple assault | $2.89 \times 10^{-13}$ |
| Burglary | 0.0051 |
| Burglary from vehicle | 0.001 |
| Theft of identity | $1.74 \times 10^{-10}$ |
| Vandalism - felony ($400 & over, all church vandalisms) | $8.96 \times 10^{-4}$ |
| Vehicle - stolen | 0.151 |

Table 1: p-value of different crime types in ad-fuller test

According to table 1, we can find that the first five crime types have stationary time series, but the last one does not. Then we calculate the rolling mean of vehicles - stolen and test the p-value using the ad-fuller test again. The p-value is 0.09 in the end and hence it is stationary in this case. We then tune the hyperparameters of our prediction model with this result.

We option to use the ARIMA (autoregressive integrated moving average) model for prediction. Before fitting the model, we need to choose the hyperparameters of the ARIMA model. There are three parameters in the model: p, d, and q. The first parameter p means the order of autoregression, which signifies the past time steps to consider for prediction. The second parameter d represents the

order of integration, which determines the order of differencing applied. The last parameter q means the order of moving average, which signifies the number of lagged forecast errors in the prediction equation. Since the time series are all stationary, we choose parameter d to be zero. Here We will draw ACF (autocorrelation function) and PACF (partial-autocorrelation function) graphs to find the parameters p and q. If there exists a significant spike at lag p in the PACF graph and a significant spike at lag q in the ACF graph, we can choose them as model parameters. If not, we fit the model starting from both p and q equal to one.

After confirming the choice of parameters, we employ the data in 2022 as the training dataset and data in 2023 as the original dataset as the testing set. We then examine the prediction results by RMSE (root mean square error) and MAE (mean absolute error). Finally, we apply a Granger causality to find some interesting relationships between pairwise crime types.

### 4.2 Network Analysis

In this part, we try to figure out the similarity of different areas in LA. We calculate the Jaccard similarity by using the top 10 frequent crime types in each area. Then we create the graph by employing each area be the nodes and each similarity be the weight of the edges'. To make the graph not so dense, we only choose the edges which have a weight larger than 0.5. Then the graph can show the overlap of crime types in the areas more significantly.

After creating the network, we calculate the page rank, close centrality, betweenness centrality, and degree centrality of each node in the graph. We then apply the K-means method to cluster the nodes into two clusters by those centrality features.

## 5 Result

### 5.1 Time Series Analysis

The trend, seasonal patterns, residual of the first feature "battery - simple assault" is shown as follow:

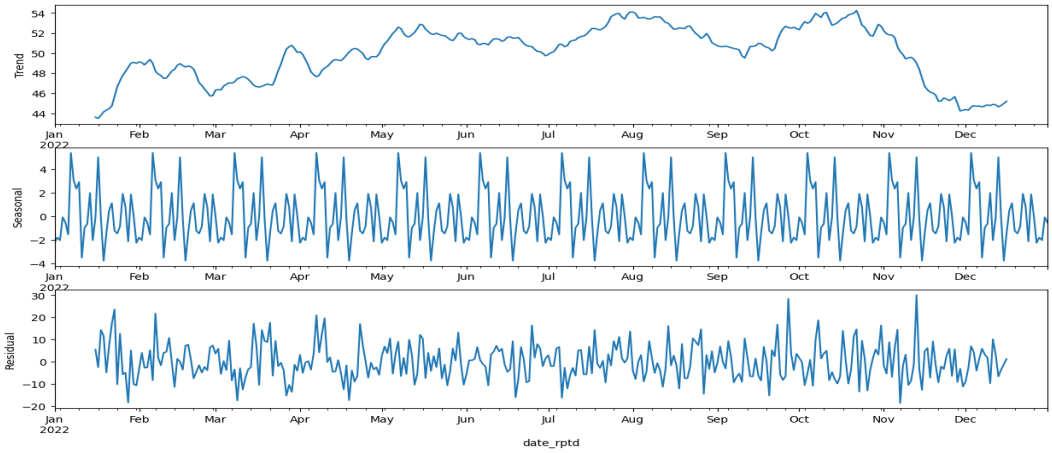



Figure 3: Seasonal decomposition of battery - simple assault

The seasonal patterns of other crime types are shown as below:

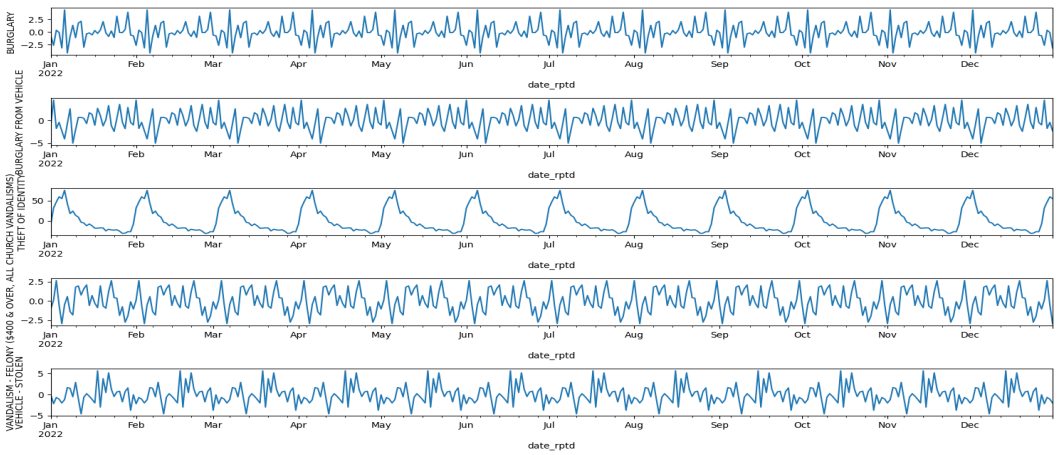



Figure 4: Seasonal components of other 5 crime types

The figure of ACF for each crime type is shown as follow:

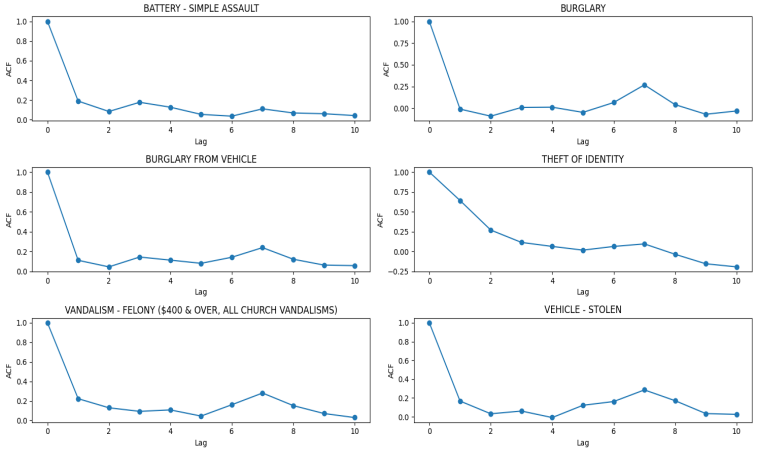



Figure 5: ACF for each crime type

The figure of PACF for each crime type is shown as below:

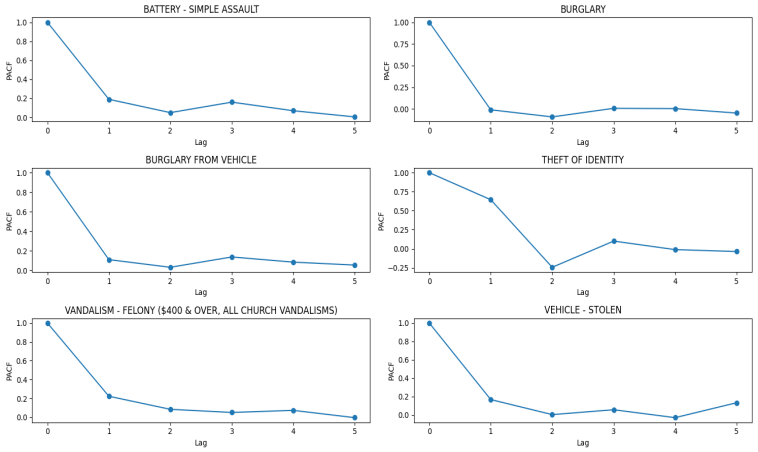



Figure 6: PACF for each crime type

According to the result of the ACF and PACF graph, we start to tune our model by setting

parameter (p,d,q)=(1,0,1), and increasing the parameters to examine different metric results. The final result shows that the model performs best when (p,d,q)=(1,0,1). We also tried to apply the original ARIMA model and seasonal ARIMA model, but the ARIMA model only gives a straight-line prediction (see figure) and seasonal ARIMA performs better. The results of best prediction metrics (RMSE and MAE) are shown in the below table:

| Crime type | RMSE | MAE |
|---|---|---|
| Battery - simple assault | 3.33 | 8.45 |
| Burglary | 2.99 | 7.21 |
| Burglary from vehicle | 3.25 | 8.84 |
| Theft of identity | 7.13 | 33.19 |
| Vandalism - felony ($400 & over, all church vandalisms) | 2.83 | 6.42 |
| Vehicle - stolen | 3.16 | 8.22 |

Table 2: RMSE and MAE of final model

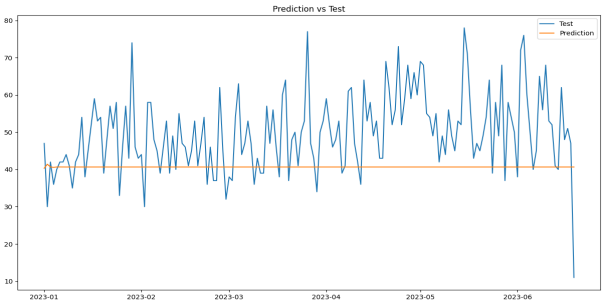



Figure 7: Prediction result of original ARIMA model (Battery)

The figure of prediction result of seasonal ARIMA model is shown as follow:

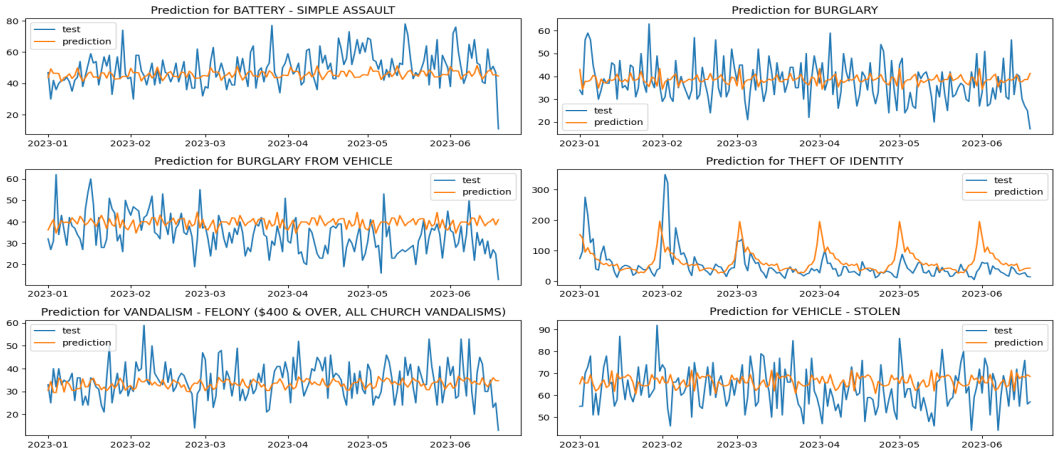



Figure 8: Prediction result of seasonal model

For Granger causality, we only show the pairs with p-value less than 0.05, which means they do not have Granger causality:

| crime_x | crime_y | p-value |
|---|---|---|
| Battery - simple assault | Burglary | 0.045 |
| Burglary | Vandalism - felony | 0.018 |
| Burglary from vehicle | Theft of identity | 0.033 |
| Vandalism - felony | Battery - simple assault | 0.00042 |
| Vehicle - stolen | Battery - simple assault | 0.017 |
| Vehicle - stolen | Vandalism - felony | 0.00052 |

Table 3: p-value¡0.05 for Granger causality

### 5.2 Network Analysis

For network analysis part, the final graph with cluster labels is shown below:

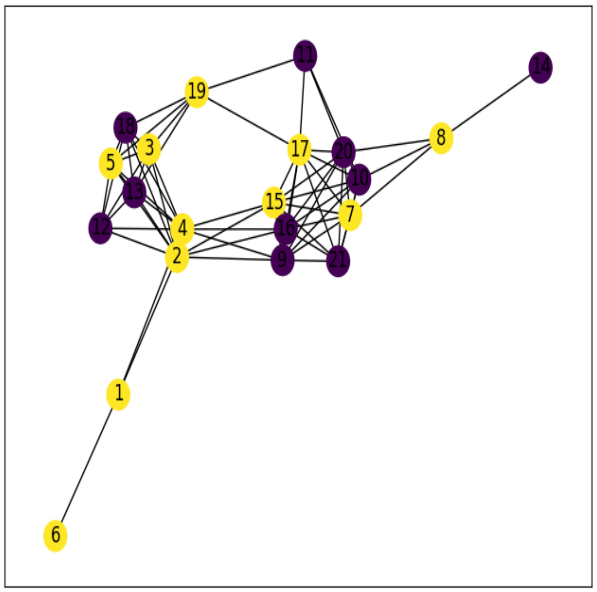



Figure 9: Network with clustering

The top 5 nodes with high page rank are shown as follow:

| node no. | page_rank | close_centrality | between_centrality | degree_centrality |
|---|---|---|---|---|
| 4 | 0.065903 | 0.588235 | 0.196626 | 0.50 |
| 2 | 0.065903 | 0.588235 | 0.196626 | 0.50 |
| 9 | 0.060023 | 0.606061 | 0.099100 | 0.45 |
| 20 | 0.060429 | 0.512821 | 0.072837 | 0.45 |
| 10 | 0.060429 | 0.512821 | 0.072837 | 0.45 |

Table 4: Top 5 nodes with high page rank

## 6 Analysis and Conclusion

### 6.1 RQ1

The results (figure 3 and figure 4) of patterns of different crime types indicate that there truly exist some seasonal patterns in the time series for the period (=30) we chose. We can also observe that for most crime types, the crime rate reaches a peak at the beginning of the month and gets smaller at the end of the month. The trend of battery - simple assault suggests that the crime rate is relatively high in fall and summer and becomes smaller in winter. The LAPD can allocate more resources in these times and needs to strengthen their precautions.

### 6.2 RQ2

For the prediction part, the result generated by the ARIMA model (figure 7) demonstrates that it cannot perform well for this problem since it gives us a straight line. The reason can be that the ARIMA model cannot reflect the seasonal property of the time series, which we have found in RQ1. The seasonal ARIMA model then performs relatively better, since the RMSE and MAE values are all relatively small except for crime-type theft of identity. The MAE value is relatively high compared to other crime types. From the prediction result (figure 8), the test data reveals that the fluctuation of crime patterns changes after February 2023. Hence, we may use other prediction models in the future to improve the result. In conclusion, the result provides insight and can become a guide for LAPD to allocate policies in the future.

The Granger causality reveals another interesting result to us. We can find that only 6 pairs of crime types have no Granger causality, which means that crime type x cannot imply crime type y. Although Granger causality cannot be regarded as true causality, it still suggests that there possibly exists causality between different crime types. The LAPD can try to focus and make great efforts on some crime types in the future, and the rate of other crime types may decrease accordingly.

### 6.3 RQ3

In the network analysis part, the results imply that nodes 4 and 2 have the highest page rank, betweenness centrality, and degree centrality, which means these two areas have the most amount of overlapped crime types compared to other areas. Hence the LAPD should assign police with more experience who can deal with cases in many different types in those areas.

The K-means clusters here do not give us very clear results, which means the graph is not highly clustered.

## 7 Future Work and Limitations

There are still some limitations to this project. First, the seasonal ARIMA model is still not very accurate for some crime types. In the future, we can use more advanced models like deep learning model LSTM to make the prediction.

Second, we only use the crime data in 2022 to analyze. In further research, we can utilize more data to let our model collect more data from history and improve the result.

Finally, the original dataset is very large and we only employ a few features in it. We can utilize more features like gender in the future to explore more interesting results.

# References


[1] Bhagat N. Shah M. Crime forecasting: a machine learning Shah, N., computer vision approach to crime prediction, and 9 (2021). https://doi.org/10.1186/s42492-021-00075-z prevention. Vis. Comput. Ind. Biomed. Art 4.

[2] Sapna Singh Kshatri, Deepak Singh, Bhavana Narain, Surbhi Bhatia, Mohammad Tabrez Quasim, and G. R. Sinha. An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. *IEEE Access*, 9:67488–67500, 2021.

[3] David S. Abrams. Covid and crime: An early empirical look. *Journal of Public Economics*, 194:104344, 2021.

[4] Data Link: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to Present/2nrs-mtv8/about$_data$.