Crime pattern mining
DSC/CSC 440

Sai Mourya Buchi
Masters in Computer Science
University of Rochester
Email: sbuchi@ur.rochester.edu

Aashrith Maisa
Masters in Data Science
University of Rochester
Email: amaisa@ur.rochester.edu

Abstract—**This study investigates the impact of legislative changes, specifically California's Proposition 47, on crime rates in Los Angeles, as well as the effect of the COVID-19 pandemic on crime patterns. Covering a period from 2010 to 2023, the research employs pattern mining and time series analysis to analyze crime data, aiming to predict future crime trends in light of the recent legislative revisions. The study's objective is to understand how these significant societal and legal changes correlate with crime occurrences, providing insights that could inform policy-making and law enforcement strategies in urban environments.**

1. Introduction

In the dynamic landscape of urban environments, maintaining public safety and managing crime effectively are of paramount importance. The complexity of these challenges is further heightened by legislative changes and global events, such as California's Proposition 47 and the COVID-19 pandemic. These developments offer a unique vantage point to examine the dynamics of urban crime in response to societal shifts. Motivated by this opportunity, our project focuses on understanding the impacts of these changes on crime rates in Los Angeles, California.

The selection of Los Angeles as the study's focus is strategic, given the city's diverse demographic composition and its significance as a microcosm for observing the effects of legislative decisions on crime trends. Spanning from 2010 to 2023, our analysis covers a critical period that includes the enactment of Proposition 47 and the entire duration of the COVID-19 pandemic. [3]

This study transcends academic inquiry, aiming to provide actionable insights for shaping policy decisions and law enforcement strategies. In an era when cities globally are confronting similar issues, comprehending the influence of legislative changes and global crises on crime rates is crucial.

In this research, we have employed advanced data mining techniques [2] not only to analyze but also to predict crime trends. Utilizing frequent pattern mining [1, 4], time series analysis, and clustering techniques, we seek to derive nuanced insights into the nature and future trajectory of crime in Los Angeles.

This project underscores our dedication to contributing to a safer, more informed society. By harnessing the power of data analysis, we endeavor to understand and anticipate the complex dynamics of urban crime, providing a foundation for informed decision-making and policy development.

1.1. Proposition 47: A Legislative Landmark

In 2014, California passed Proposition 47, a law that significantly transformed the state's criminal justice system. This proposition reclassified certain non-violent offenses from felonies to misdemeanors, leading to a considerable decrease in the state's jail and prison populations.

Contrary to the narrative that Proposition 47 increased crime rates, evidence, as highlighted by the Georgetown Journal on Poverty Law & Policy, indicates a different scenario. Post-implementation, statewide violent and property crime rates were lower than in 2010, with some Californian cities even experiencing a decrease or stabilization in crime rates. This suggests that Proposition 47 did not causally affect criminal activity, offering a more complex perspective than initially perceived.

Moreover, the Public Policy Institute of California points out that following Proposition 47, the state's total incarceration rate fell to a 20-year low. While there was a modest rise in property crime, particularly auto thefts, the reduction in incarceration did not lead to an increase in violent crime.

In 2023, the legislative context evolved further with Senate Bill 533, aimed at protecting employees from being required to intervene in criminal situations. This bill, while focused on employee safety rather than criminal activity, adds another layer to the intricate interplay between legislation and crime rates.

1.2. Bill to Repeal Proposition 47: Legislative Response to Retail Theft

In response to rising concerns about retail theft and related crimes, a bill was introduced in the California Assembly to repeal Proposition 47. This proposed legislation aims to lower the felony theft threshold from $950 to $400. This move reflects a significant reevaluation of Proposition 47, particularly regarding its impact on theft and property crimes. The introduction of this bill underscores the ongoing dialogue and reassessment in California's legislative approach to balancing reduced incarceration rates with public safety concerns. Our study takes into account this evolving legislative context, analyzing its implications on urban crime trends in Los Angeles and contributing to the broader understanding of criminal justice policy in California.

2. Data preprocessing

The study utilized two extensive datasets encompassing crime data from Los Angeles over a 13-year period from 2010 to 2023. These datasets are rich in detail, containing various attributes such as the date of occurrence, types of crimes, and victim demographics like gender, age, race, and the geographical area of the crime. The first dataset covers data from 2020 to the present, while the second spans from 2010 to 2019, allowing for a longitudinal analysis of crime trends.

In the initial phase of data preprocessing, the focus was on refining the datasets by removing attributes that were not essential for the analysis. Attributes such as area code, crime code, and status were deemed extraneous and thus excluded. This step was crucial in ensuring that the analysis would be concentrated on the most impactful variables. Additionally, the time of crime reporting was processed and standardized across the datasets to maintain uniformity and accuracy in the temporal analysis.

The next step involved addressing missing and anomalous data within key columns. The columns representing victim's age, sex, and race were scrutinized for missing values and inconsistencies. Entries with implausible values, such as age listed as zero or marked as 'X', or race indicated as 'X', were removed. This cleansing process was vital to uphold the integrity and reliability of the dataset.

Recognizing the extensive variety of crime types in the dataset, a strategic categorization was undertaken. The diverse crime types were consolidated into nine broad categories: violent crimes, sexual offenses, white-collar crimes, property crimes, petty thefts, grand thefts, among others. This categorization was informed by the existing penal codes and supplemented with extensive research, including consulting Google for clarity on certain crime types. This step was instrumental in simplifying the complexity of the crime data, thereby enabling more coherent pattern analysis.

Another significant aspect of the preprocessing was the clustering of continuous age data. Victim age, initially a continuous variable, was segmented into defined categories or buckets. This categorization allowed for a more nuanced analysis of crime trends across different age groups.

The study also acknowledged the diversity in victim races. Given the numerous unique racial identifiers, particularly among minority groups, these were grouped into a singular 'others' category. This consolidation resulted in four main racial categories, streamlining the analysis and ensuring that minority group data were appropriately represented and analyzed.

Geographical categorization of the crime locations was also undertaken. The myriad of crime locations were grouped into seven broader geographical areas. This was achieved through the use of geographical references and mapping tools sourced from Google, ensuring an accurate and meaningful categorization of crime locations.

Upon the completion of these preprocessing steps, the datasets were primed for the application of sophisticated data mining techniques. Frequent pattern mining algorithms were deployed to unearth recurring patterns and associations in crime occurrence. Time series analysis was utilized to explore and interpret trends over the examined period. This approach was particularly significant in projecting future crime patterns, especially in the context of the legislative changes and societal upheavals such as the COVID-19 pandemic.

3. Methodologies

3.1. Frequent Pattern Mining Analysis

In the initial stage of our frequent pattern mining analysis, the preprocessed data was converted into two distinct lists of lists, each representing a different time frame, to effectively apply the FP-Growth algorithm. These lists encapsulated key aspects of each crime occurrence, including victim sex, race, age group, area category, and the specific crime. Initially, we experimented with a minimum support threshold of 10% for the FP-Growth algorithm. However, this threshold did not yield a significant number of relevant patterns. Through iterative adjustment, we discovered that a minimum support of 2% was optimal, uncovering a plethora of meaningful itemsets in both the pre- and post-COVID datasets.

Upon identifying these frequent itemsets, we focused on the top 10 from each dataset for a comparative analysis. We utilized the FOCUS [5] metric to quantify the dissimilarity between these sets of itemsets, where a higher FOCUS value indicated greater divergence. This analysis was pivotal in highlighting the changes in crime patterns before and after significant societal and legislative events.

Further, we delved into association rule mining from these itemsets, particularly prioritizing rules with a confidence level above 55%. This approach was instrumental in revealing the relationships between victim demographics and various crime types. To extend our analysis, we also considered the nine broader crime categories defined in our preprocessing. This step was replicated for both the pre- and post-COVID datasets, allowing us to draw comparisons and understand the shifts in crime patterns across different time periods.

A key observation was the noticeable shift in crime rates between 2014 and 2015, coinciding with the enactment of Proposition 47. We meticulously analyzed the frequent itemsets from periods before and during this legislative change, providing deeper insights into the evolution of crime patterns in response to Proposition 47.

3.2. Time Series Analysis Using LSTM

Our approach to time series analysis involved the development and utilization of an LSTM model, specifically structured with two LSTM layers, two dropout layers, and a final dense layer. This model composition was strategically chosen to adeptly capture and analyze the temporal patterns inherent in the crime data.

The training phase of the LSTM model utilized data spanning from 2010 to 2014. This period was crucial for establishing a baseline of crime trends before the implementation of Proposition 47. The subsequent validation phase utilized data from 2015 to 2016, providing a critical comparison to the post-legislation crime scenario.

Looking towards the future, we retrained the LSTM model on more recent data, from 2020 to 2023, to predict crime trends for 2024 to 2025. This predictive analysis was particularly focused on assessing the potential impact of recent legislative revisions on future crime patterns.

Throughout the training process, the model underwent 150 epochs with a batch size of 32. This extensive training was designed to ensure the model's efficacy in learning complex temporal relationships and enhancing its predictive accuracy for future crime trends in Los Angeles.

4. Insights

4.1. Insights from Frequent Pattern Mining

Our in-depth analysis commenced with a visualization phase, where we used pie charts to depict the distribution of various attributes within our dataset. This visual exploration helped us in grasping the demographic and criminal diversity of our dataset. Complementing these pie charts, a heatmap provided a spatial perspective, showcasing the geographic spread of crime occurrences across different areas.

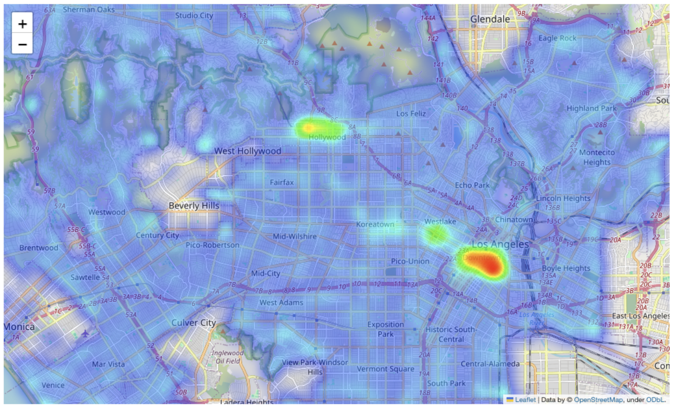


Figure 1. Heat map on crimes

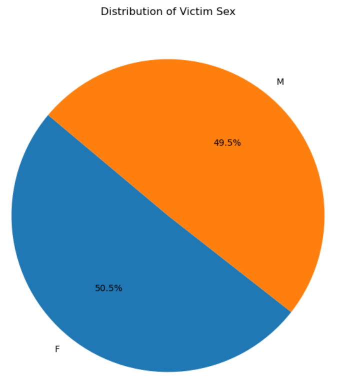


Figure 2. Distribution of Victim Age in Pre-COVID data

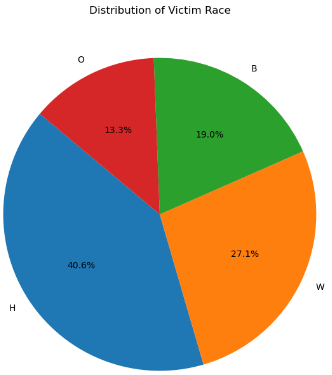


Figure 3. Distribution of Victim Race in Pre-COVID data

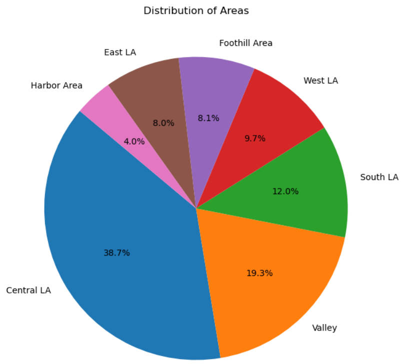


Figure 4. Distribution of Area of crime occurred in Pre-COVID data

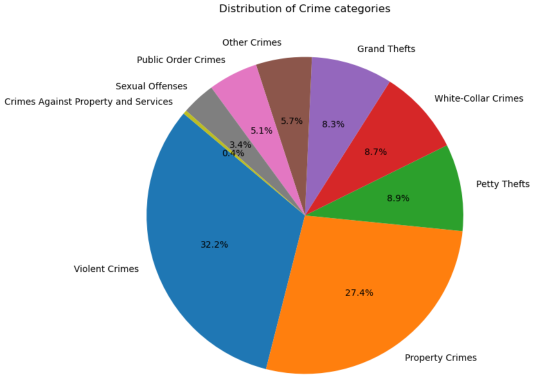
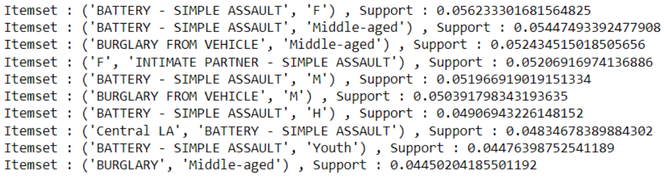


Figure 5. Distribution of Crime Categories in Pre-COVID data

A time series plot of the entire dataset, spanning from 2010 to 2023, offered a temporal view of the crime trends. These visualizations were instrumental in setting expectations for the patterns that our analysis might reveal.

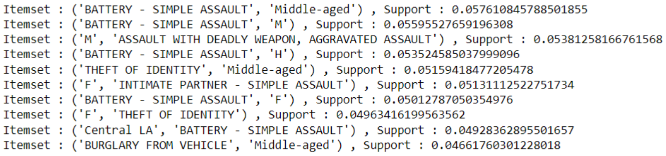
4.1.1. Analysis of Crime Patterns: Pre-Covid and Post-Covid.

Diving into the heart of our study, we applied frequent pattern mining to both pre-covid and post-covid datasets. This bifurcation aimed to capture the shifts in crime patterns due to the pandemic's impact. We meticulously extracted the top 10 frequent patterns for each dataset, analyzing them for both specific crimes and broader crime categories. This approach afforded us a granular understanding of the crime landscape, revealing changes in crime dynamics over these distinct periods.



```
Itemset : ('BATTERY - SIMPLE ASSAULT', 'F') , Support : 0.056233301681564825
Itemset : ('BATTERY - SIMPLE ASSAULT', 'Middle-aged') , Support : 0.05447493392477908
Itemset : ('BURGLARY FROM VEHICLE', 'Middle-aged') , Support : 0.052434515018505656
Itemset : ('F', 'INTIMATE PARTNER - SIMPLE ASSAULT') , Support : 0.05206916974136886
Itemset : ('BATTERY - SIMPLE ASSAULT', 'M') , Support : 0.051966919019151334
Itemset : ('BURGLARY FROM VEHICLE', 'M') , Support : 0.050391798343193635
Itemset : ('BATTERY - SIMPLE ASSAULT', 'H') , Support : 0.04906943226148152
Itemset : ('Central LA', 'BATTERY - SIMPLE ASSAULT') , Support : 0.048346783898843O2
Itemset : ('BATTERY - SIMPLE ASSAULT', 'Youth') , Support : 0.04476398752541189
Itemset : ('BURGLARY', 'Middle-aged') , Support : 0.04450204185501192
```

Figure 6. Top 10 frequent pattern for specific crime in Pre-COVID data



```
Itemset : ('BATTERY - SIMPLE ASSAULT', 'Middle-aged') , Support : 0.057610845788501855
Itemset : ('BATTERY - SIMPLE ASSAULT', 'M') , Support : 0.05595527659196308
Itemset : ('M', 'ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT') , Support : 0.05381258166761568
Itemset : ('BATTERY - SIMPLE ASSAULT', 'H') , Support : 0.053524585037999096
Itemset : ('THEFT OF IDENTITY', 'Middle-aged') , Support : 0.05159418477205478
Itemset : ('F', 'INTIMATE PARTNER - SIMPLE ASSAULT') , Support : 0.05131112522751734
Itemset : ('BATTERY - SIMPLE ASSAULT', 'F') , Support : 0.05012787050354976
Itemset : ('F', 'THEFT OF IDENTITY') , Support : 0.04963416199563562
Itemset : ('Central LA', 'BATTERY - SIMPLE ASSAULT') , Support : 0.04928362895501657
Itemset : ('BURGLARY FROM VEHICLE', 'Middle-aged') , Support : 0.04661760301228018
```

Figure 7. Top 10 frequent pattern for specific crime in Post-COVID data

4.1.2. Measuring Dissimilarity with FOCUS Metric.

We applied the FOCUS metric to quantify the dissimilarity between the pre-covid and post-covid datasets. For specific crimes, the dissimilarity score stood at 0.521, indicating a significant divergence in crime patterns following the onset of COVID-19. When this analysis was extended to broader crime categories, the dissimilarity decreased to 0.495. This drop was anticipated, as the aggregation into broader categories tends to dilute specific crime variations.

$$dis(A,B) = \frac{\sum_{X \in A \cup B} |supp_D(X) - supp_E(X)|}{\sum_{X \in A} supp_D(X) + \sum_{X \in B} supp_E(X)}$$

4.1.3. Association Rules: Unveiling Hidden Relationships.

Our frequent pattern mining also led to the discovery of fascinating association rules. In the pre-covid data, one notable rule suggested that if a simple assault by an intimate partner occurred in Central LA, there was a 79.2% likelihood the victim was female. Another rule indicated a 61.1% confidence that victims of identity theft who were male were also middle-aged. In the context of violent crimes in East LA, there was an 80.43% confidence that the victim was Hispanic. Additionally, for property crimes involving Hispanic male victims, there was a 62.3% chance the victim was middle-aged.

Post-covid data revealed its own set of intriguing rules. For instance, in cases of aggravated assault in Central LA, the likelihood of the victim being male was 73.2%. Another interesting finding was that for grand thefts involving male victims, there was a 61.5% probability of the victim being middle-aged.

4.1.4. Proposition 47: Analyzing Legislative Impact.

The enactment of Proposition 47 in 2014, which reclassified certain crimes as misdemeanors, prompted us to analyze the dissimilarity in crime patterns around this legislative change. The FOCUS metric revealed a dissimilarity of 0.74 for specific crimes and 0.72 for crime categories between the periods before and immediately after 2014, underscoring the significant impact of the proposition. The presence of patterns involving petty thefts in the top 10 itemsets between 2014 and 2015 was particularly telling, reflecting the proposition's influence on crime trends.

4.2. Time Series Analysis and Prediction

Our study employed a comprehensive time series analysis and LSTM (Long Short-Term Memory) model training to understand and predict crime trends in Los Angeles, particularly in light of legislative changes like Proposition 47.

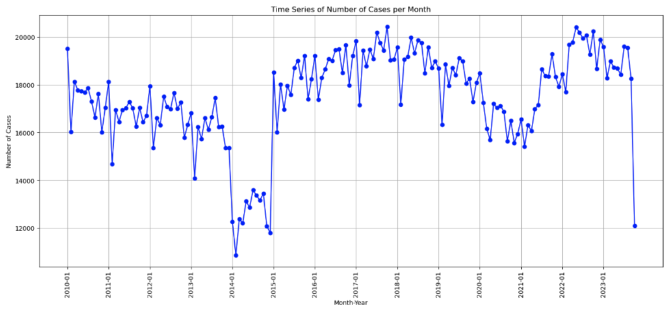


Figure 8. Crime rate from 2010 to 2023

Initially, we focused on crime data from 2010 to 2014. This period was crucial as it predated the enactment of Proposition 47, allowing us to establish a baseline for crime patterns before this significant legislative change. Our time series analysis aimed to identify underlying trends and seasonal patterns in various crime categories, with a special focus on petty thefts and grand thefts.

The LSTM model, known for its effectiveness in handling time series data due to its ability to capture temporal dependencies, was then trained on this dataset. LSTM's recurrent neural network architecture is particularly suited for predicting time-dependent data, making it an ideal choice for our analysis.

4.2.1. Validation and Insights (2015-2016). Post-training, the LSTM model was validated using data from 2015 to 2016, the period immediately following the implementation of Proposition 47. This proposition reclassified certain non-violent offenses from felonies to misdemeanors, impacting crime rates and incarceration numbers. The validation process was crucial to test the model's accuracy and reliability.

The model demonstrated a high degree of accuracy, as evidenced by a robust Root Mean Square Error (RMSE) metric. Our analysis during this phase revealed a significant decline in petty thefts and grand thefts, aligning with the expected outcomes of Proposition 47's reclassification of certain crimes.

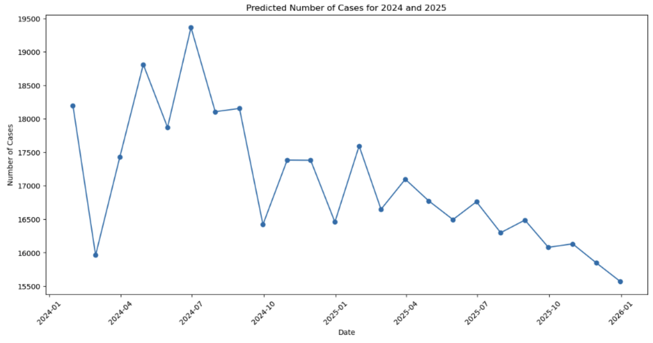


Figure 9. Predicted crime rate

4.2.2. Predictive Analysis for Future Trends (2024-2025). Leveraging the insights and patterns uncovered, we extended our analysis to predict crime trends for the years 2024 to 2025. This prediction was particularly challenging and significant as it aimed to gauge the potential impact of the revised Proposition 47. For this predictive analysis, the LSTM model was retrained on more recent data from 2020 to 2023.

Our predictive model indicated an initial spike in crime rates at the beginning of 2024, which gradually decreased over the months. This trend aligns with the anticipated effects of the revised Proposition 47, suggesting that while there might be an initial adjustment period, the overall crime rate could decrease as the year progresses.

5. Conclusion

Our research embarked on a quest to decode the intricate impacts of legislative shifts, particularly Proposition 47, and the COVID-19 pandemic on crime patterns in Los Angeles. The insights garnered from our in-depth frequent pattern mining and advanced time series analysis have illuminated critical correlations and trends, resonating well with our initial hypotheses.

The implementation of Proposition 47, a landmark legislative change, was found to significantly influence crime dynamics. Our analysis revealed an evident transformation in the types of crimes post-legislation, showcasing a shift towards lesser felonies, as anticipated from such legal reforms. This underlines the potent influence of legislative decisions on urban crime patterns.

Concurrently, the onset of the COVID-19 pandemic introduced a distinct paradigm shift in criminal activities. Our findings indicated not just an escalation in overall crime rates but also a transition in crime nature, with a surge in offenses like identity theft and aggravated assault. This shift is reflective of the changing societal landscape during the pandemic.

The predictive analysis using our LSTM model, particularly projecting into 2024-2025, provided a forward-looking perspective on how crime trends might evolve in response to ongoing and future legislative and societal changes. The model's predictions, based on the nuanced understanding of past and present trends, suggest a landscape of urban crime that continues to evolve in complexity.

In summarizing, our study reaffirms the significant impact of legislative changes and global crises on crime trends. The research contributes valuable insights into the dynamic interplay between societal forces and criminal activities, offering crucial implications for policy-making and law enforcement. It underscores the necessity for adaptive, informed strategies in crime management and prevention, attuned to the evolving tapestry of societal norms and legislative frameworks.

References

[1] Kumar, Dr K Ramesh. (2014). A complete survey on application of frequent pattern mining and associa:on rule mining on crime paIern mining. Interna:onal journal of Advances in Computer Science and Technology. 3. 264-275.

[2] S. V. Nath, Crime PaIern Detec:on Using Data Mining, 2006 IEEE/WIC/ACM Interna:onal Conference on Web Intelligence and Intelligent Agent Technology Workshops, Hong Kong, China, 2006, pp. 41-44, doi: 10.1109/WI-IATW.2006.55.

[3] Liu L, Chang J, Long D, Liu H. Analyzing the Impact of COVID-19 Lockdowns on Violent Crime. Int J Environ Res Public Health. 2022 Nov 23;19(23):15525. doi: 10.3390/ijerph192315525. PMID: 36497600; PMCID: PMC9739108.

[4] Ganti, V., Gehrke, J., Ramakrishnan, R.: A framework for measuring changes in data characteristics. In: ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), pp. 126–137. ACM Press, New York (1999)

[5] Ntoutsi, Irene & Theodoridis, Yannis. (2008). Comparing Datasets Using Frequent Itemsets: Dependency on the Mining Parameters. 5138. 212-225. 10.1007/978-3-540-87881-0_20.