Predicting Crime and Safety: Los Angeles Crime Data Analysis and Visualization

Al-Amin Hossain
Department of Computer Science
University of South Dakota
Vermillion, SD, USA
alamin.hossain@coyotes.usd.edu

Abstract— **This study explores the patterns of crime in Los Angeles(LA), using the dataset from the Los Angeles Police Department (LAPD) spanning from 2020 to the present. The primary focus was to analyze the criminal activities and understand if LA is safe to visit for new tourists. The project provides a multifaceted view of urban criminal activity by focusing on various aspects of crime, such as geographical hotspots, crime count monthly and yearly trends, demographic, age, and gender impacts, and the nature of offenses. A significant part of the study involved identifying areas with the highest crime rates. The analysis revealed a few places as the most crime-prone areas. It also discovered the months with the highest crime incidence and the days and times when crimes peaked. The study also indicates the victim's group according to age, color, and sex. This insight is crucial for developing targeted social interventions and preventive measures. Moreover, the study observed the nature of crime, premises, and weapon types. A key component of the study was the application of machine learning models to predict crime counts in upcoming days. Tested and evaluated multiple models to find the perfect fit. This project has an immense scope. This could be helpful for general people to know about LA's crime pattern. Law enforcement authorities could be aware of the hot spots and prepare with necessary precautionary measures. Overall, the study not only paints the crime patterns it also describes the potential uses of the data analysis and predictive modeling for a safe society.**

Keywords— crime, count, location, Los Angeles, predicted, safety

I. Introduction

Crime and criminal activities are common in big cities. This is becoming one of the biggest problems in society. Nowadays, it is increasing in many cities around the world at an alarming rate. Anyone could be the victim of any form of criminal offense if someone is not aware of the area and surroundings. Crime in an area is often linked to different things like race, age, gender, infrastructure, timing, day, and months. It is easy for a person to be aware of these mentioned parameters if he or she has been living in that area for a long time. But a tourist or traveler can't know about everything without being there. So, it will be helpful and sometimes lifesaving for them to be aware of the crime-prone areas, most of the occurrence time, crime type, and related information before the trip begins.

Los Angeles, often referred to as the "City of Angels", is a sprawling metropolis on the West Coast of the United States. It is renowned for its cultural diversity, entertainment industry, and iconic landmarks. However, beneath the glamorous facade, the city grapples with complex challenges. According to Wikipedia, the population of Los Angeles is 3.8 million [2] which is the third highest for a city in North America, second in the United States, and the first in California. The distinctive combination of various communities, socioeconomic variations, and geographical characteristics in Los Angeles creates an environment conducive to criminal activities.

This research project focuses on studying crime in Los Angeles. I formulated a few questions about the safety and crime in LA. I gathered crime data from 2020 to the present from the Los Angeles Police Department (LAPD). I then explored and analyzed the collected data to find the answers to the questions which led me to identify crime-prone areas, the most common crimes, peak times and days for crime, types of crime, locations, and the impact of age and sex on crime rates. Additionally, I looked at trends in crime over the past few years. Using the same data, I attempted to predict the average daily crime count using various machine-learning models. I compared the results of each model using evaluation metrics to determine the most suitable one for this dataset.

II. Related Work

Learned and used the technique of data visualization for outlier detection from this project. Ismail Sefa (2020) has nicely represented the process.[6]

Adopted data cleaning technique, box plotting for outlier detection and visualization technique from this project by Sidney Kung [8]

The approach towards finding the answers has been shown in this project. Ambarish(2017) has explained how to find out the expected results from this large dataset. [5]

The application of machine learning models in crime prediction has shown promising results. Patel and Singh (2020) successfully used regression models to forecast crime rates, a methodology I adopted in our study. [3]

The effectiveness of different predictive models, as discussed by Lee and Kang (2021), informed our selection of the Autoregressive model. [4]

III. Methodology

This project tries to inform people about crime and safety in Los Angeles by analyzing the reported incidents. At first, I carefully prepared a few questions that would answer

a. which areas of LA have the highest crime?

b. which months are most crime-prone, which day of the week, and which time of the day?

c. is there any connection between crime count, age, and sex of the victims?

d. how much crime has fluctuated in the last 3 years, which type of crime mostly occurs, which premises the crime occurs most, and what kind of weapons are mostly used for crime?

e. what are the top 3 impacted demographics in LA?

To find the answer I followed the steps described in figure 1. This is a high-level diagram of the steps that I have followed to find the answers. The steps will be further described below.

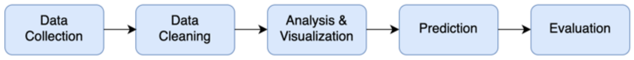


Fig. 1. Methodology steps

A. *Data Collection*

For this project, I have collected the data from the Los Angeles Police Department (LAPD) official website. This dataset is open for all and they keep it updated every alternate week. I have taken the dataset "Crime Data from 2020 to Present" [1], which reflects the crime incidents in Los Angeles dating back to 2020. The downloaded data was last updated on September 11, 2023. It has 798242 rows and 28 columns consisting of the victim's age, sex, descent, weapon used for the crime, location of crime with latitude and longitude, etc. The data provided in this dataset is numeric and string. Each row belongs to each reported incident. Data in each instance belong to different areas of the City of Los Angeles. The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21[1]. The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example, 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles. [1]Some of the information which is utilized in our analysis is as follows:

TABLE I. COLLECTED DATA AND TYPES

Column Name	**Description**
DATE OCC	The timestamp at which the crime occurred. Data Type: Floating Timestamp
TIME OCC	The time at which the crime occurred. Data Type: Text
AREA NAME	Name of the geographic area. Data Type: Text
Crm cd Desc	Indicates the crime committed. Data Type: Text
Vict Age	Indicates the age of the victim. Data Type: Numeric
Vict Sex	Indicates the sex of the victim. Data Type: Text
Vict Descent	Indicates the race of the victim. Data Type: Text
Premis Desc	Indicates the premises where the crime occurred. Data Type: Text
Weapon Desc	Weapon used for the crime. Data Type: Text
LAT	It gives the latitude of the crime. Data Type: Numeric
LON	It gives the longitude of the crime. Data Type: Numeric

B. *Data Cleaning*

The dataset used for this project has a few instances that contain some missing values or null values and probably outliers. To perform data processing, it is required to improve the data quality. There are various techniques available to improve the data quality. I used Python for this project. So, I used the Python Pandas library to manipulate and clean the data.

The data set has 28 columns. First, I dropped the unnecessary columns which are not needed. I dropped 12 columns and kept DR_NO, DATE OCC, TIME OCC, AREA, AREA NAME, Crm Cd, Crm Cd Desc, Vict Age, Vict Sex, Vict Descent, Premis Cd, Premis Desc, Weapon Used Cd, Weapon Desc, LAT and LON. Then renamed those with the proper name in CamelCase format without keeping the spaces in between.

I visualized the dataset using the Python missingo library to check the missing and available data visually. The figure shows dark squares and light squares with the column name on the top. The dark means the data is present and the light square means the data is missing. From the visualization, I found that VictimSex, VictimDescent, WeaponUsedCode, WeaponDesciption column has missing values. Then I checked how many null values were in the dataset for each column via using df.isna().sum(). I found VictimSex has 104653, VictimDescent has 104661, PremiseCode has 9, PremiseDescription has 472, WeaponUsedCode and WeaponDescription has 520347 null values. I have to take care of these values to improve the data quality.

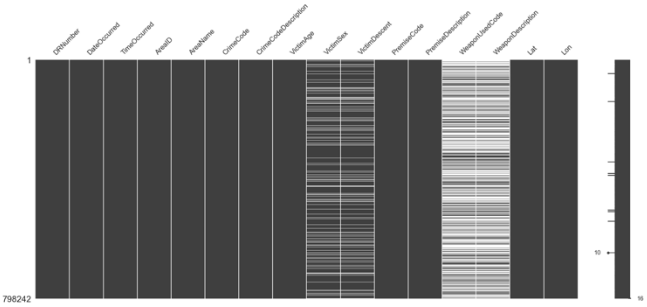
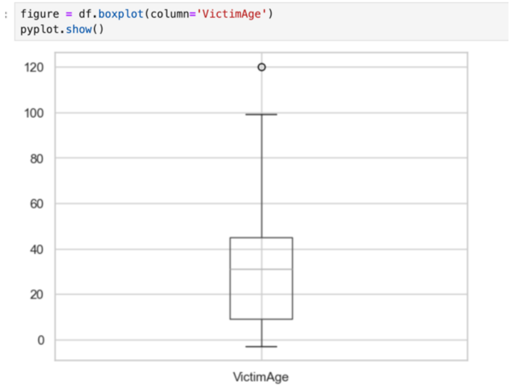


Fig. 2. Missingo Diagram to check the null values

I did not notice any null value for the VictimAge column. Then I visualize the data of VictimAge column using a boxplot to see if there are any outliers present.



```
figure = df.boxplot(column='VictimAge')
pyplot.show()
```

Fig. 3. Checking outliers using boxplot

According to figure 3 we can notice few rows have an age less than zero which is impossible. So, I checked and found that 68 rows have ages below zero. I dropped those rows to avoid invalid values. Rather than those, I also noticed a row that contains age over 100. Although, age over 100 is not impossible but I dropped that too to avoid outliers. After removing 69 rows the remaining row count became 798173.

After dealing with outliers, I converted the DateOccurred format, removed the time from the value, and saved it in YYYY-MM-DD format. Then, set this column as the index of the data frame. I filled all null values of VictimSex with 'X'. Then I checked the types of VictimSex and found these types: 'M', 'F', 'X', 'H', and '-'. I decided to keep 'M', 'F', and 'X' and convert rest '-' and 'H' into 'X'.

Then I checked the count of VictimDescent for each descent. Replaced '-' with 'X' and all the null values are with 'X`'. Following the same technique replaced all the null values of PremiseCode and WeaponUsedCode with zero. Adding to that filled the null values of PremiseDescription an WeaponDescription with UNKNOWN.

Finally, again checked the whole data frame if there is any null value still exists or not and saved the cleaned data.

C. *Analysis and Visualization*

Data analysis is the process of inspecting, transforming, and modeling data to discover useful information. It involves a variety of techniques and methods to uncover patterns, trends, relationships, and insights within a dataset. The primary objective of data analysis is to extract actionable insights that can inform better understanding, predictions, or decisions. Data visualization is the representation of data in graphical or visual formats, such as charts, graphs, maps, and dashboards, to facilitate understanding and interpretation. Data visualization aims to convey complex information in a clear, intuitive, and accessible manner. It allows individuals to grasp patterns, trends, and insights more easily than through raw data alone.

In this project, I aimed to address the specified questions using the gathered dataset. Initially, I determined the crucial features for answering the questions.

To identify the areas in LA with the highest crime counts, I utilized the AreaName feature. I grouped and visualized the top five areas with the highest crime numbers through a horizontal bar graph. Fig. 4 in the result shows the areas with the highest crime count. Subsequently, I mapped the data to provide a clear visualization of the regions with the highest crime rates. Fig. 5 in the result section shows a map where crime occurs most.

For understanding the temporal aspects of crime, I employed the DataOccurred and TimeOccurred columns to determine the most and least crime-prone months, days of the week, and times of the day. Fig.6, Fig.7, Fig.8, Fig.9 showing the findings in result section. Additionally, I investigated the most common crime during peak times using the CrimeCodeDescription feature in conjunction with TimeOccurred. Fig.10 showing the which crimes occurs in the highest occurs hour.

To explore the impact of victim sex and age on crime counts, I analyzed the VictimSex, VictimAge, and CrimeCodeDescription. Using bar graphs(shows in Fig. 11 result section), I illustrated the gender most affected and employed a combined bar plot to identify the types of crimes associated with each gender. I applied an age filter over 70 on the same plot to pinpoint common crimes among the elderly.(visible in Fig. 12) Furthermore, I created histograms to highlight the most victimized age groups based on gender. The visualization is shown in the Fig. 13.

I used DateOccurred and generated a line plot to find out the trend of crime count over the last three years(Fig. 14). Then I used the CrimeCodeDescription and prepared a horizontal bar graph to find out which kind of crimes occurs most in LA(Fig. 16). Similarly, generated two bar graphs using PremiseDescription and WeaponDescription to find out where the most crime occurs and what kind of weapon is being used to commit the crime (Fig. 15 and Fig. 17). This also indicates if the crime is life-threatening or not. I prepared another graph combining PremiseDescription and CrimeCodeDescription to find out which kind of crimes occurs in the most occurred crime premises.

Lastly, I investigated victim descent, focusing on the top three descents most impacted. By incorporating AreaName, I determined the areas where these descents were most frequently victimized. (Fig. 18 and Fig. 19)

D. *Prediction*

Predicting crime involves anticipating crimes before they occur, necessitating the use of tools to forecast these incidents. Machine Learning algorithms can enhance predictive capabilities. Leveraging a collected dataset and a machine learning model, it becomes possible to anticipate the upcoming crime count.

After improving the quality of the dataset by data cleaning the total count of reported incidents was 798173. Calculated the count of crimes for each day. A total of 1350 unique days were found when the crimes were reported. I noticed that I am trying to predict a value which is a continuous value. So, I chose a few regression models and defined the evaluation metrics to evaluate which model is best performing for this dataset. I chose Linear Regression(LR), Autoregressive(AR), Moving Average(MA), and Autoregressive Integrated Moving Average (ARIMA) to predict the count of the crime for each day. I also selected Mean Average Error (MAE), Mean Square Error(MSE), and Root Mean Square Error(RMSE) to compare the models' performance.

First, I selected the independent variable as a unique date (each day) and the dependent variable is crime count. Then shuffled the data and kept 30% as the test set and 70% as the training set. Then, fitted training data in the Linear Regression model. After fitting the data successfully, predict the crime count on future dates which were kept for testing data. Then, we compared the crime count of each day from testing data with the predicted count by the LR model and calculated the MAE, MSE, and RMSE. (Table III)

Again, divided the crime count data into training and testing sets. I kept the 2020-01-01 to 2022-12-31 data as a training set and 2023-01-01 to the rest of the data as a testing set. Then, fitted the data into the Autoregressive model, Moving Average model, and Autoregressive Integrated Moving Average model. Then, I calculated the MAE, MSE, and RMSE values for each model. (Table III)

E. Evaluation

After all the calculations of MAE, MSE, and RMSE I compared the values of the four tested models and decided which was the best-performing model for the dataset.

IV. Result

A. Result of Visualization

a. Which areas in LA have the highest crime rate

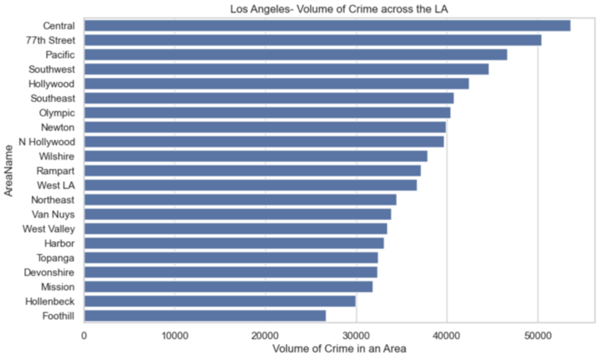


Fig. 4. Areas with highest Crime Count

Here we can see, the top 5 areas where most of the crimes in LA have occurred. Those are Central, 77th Street, Pacific, Southwest, and Hollywood.

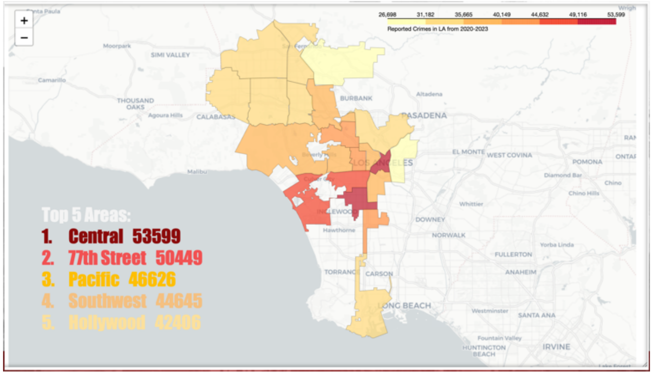


Fig. 5. Map showing the areas with color which have the highest crime count

The dark red color describes the most crime-occurred area, light red is the second highest, dark yellow is the third highest, light yellow is the fourth, and lighter yellow is the fifth highest.

b. Which months are most crime-prone, which day of the week, and which time of the day?

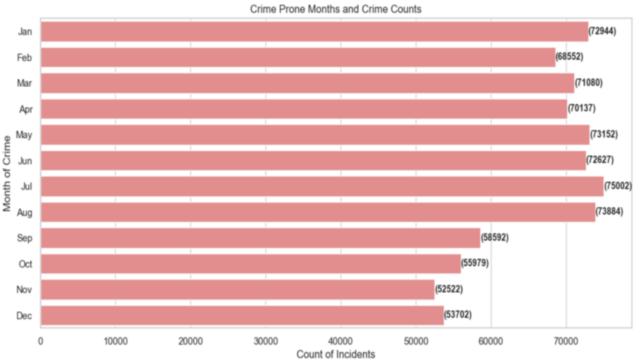


Fig. 6. Crime Count monthly

July and August have the highest crime occurrence count. In contrast, we can observe that November and December have the lowest Crime incidents count.

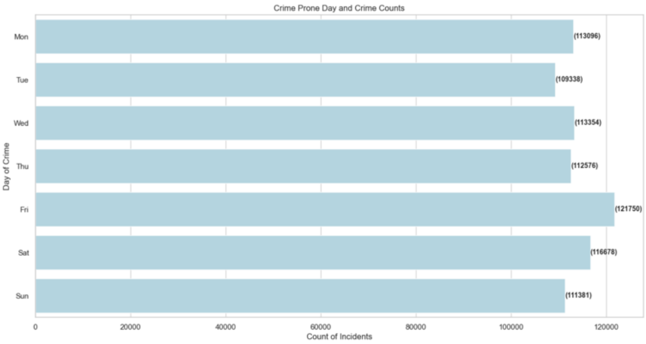


Fig. 7. Crime Count per Day

The figure shows that Friday has the highest and Tuesday has the lowest crime count.

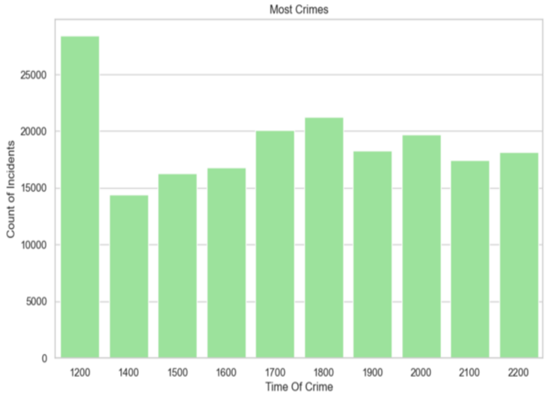


Fig. 8. Crime Count hourly

The plot shows the hours of the day when most crimes occur. We can see a very strange thing! Crime occurs mostly at 1200 hours. Why would someone commit a crime so much in the middle of the day, rather than at night? So, we find out when the least crime occurs and what type of crime occurs most at noon hour.

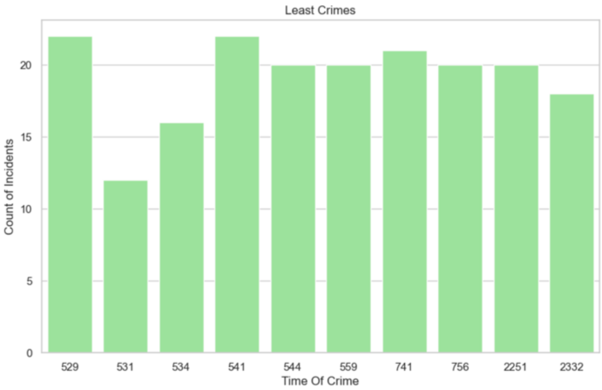


Fig. 9. Crime Count hourly

The crime occurrence count is least at 700 hours. Is it because people have woken up and getting ready for the crime? This is interesting.

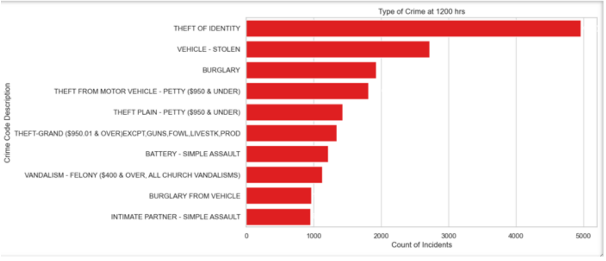


Fig. 10. Mostly occurred crime type at 12:00

The figure shows that Theft of Identity is the most common crime type that happened during noon time.

c. Is there any connection between crime count, age, and sex of the victims?

TABLE II. VICTIM SEX AND COUNT

Victim Sex	Count
Male (M)	330086
Female (F)	294367
Unknonw (X)	173720

From the table II, we can say that Males are the most victims according to this collected dataset. There is a probability that Females could exceed the count of males if we could figure out the unknown portion. The possibility of vice versa is also there.

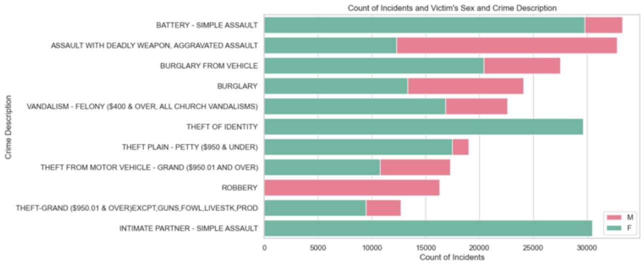


Fig. 11. Incidents count according to victim's sex and crime type

We can see that Females are the only victims of Intimate Partner-Simple Assault and Theft of Identity. On the other hand, Robbery has only male victims.

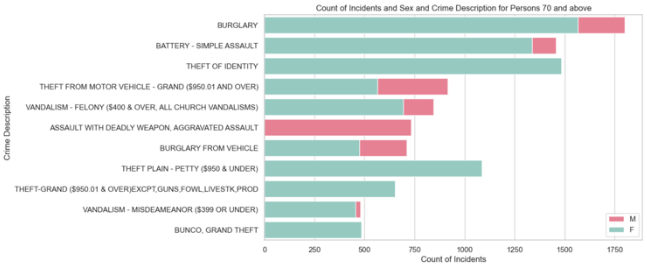


Fig. 12. Incidents count according to victim's sex and crime type over 70 years

We observe that Burglary is the most occurred crime committed by persons aged 70 and above.

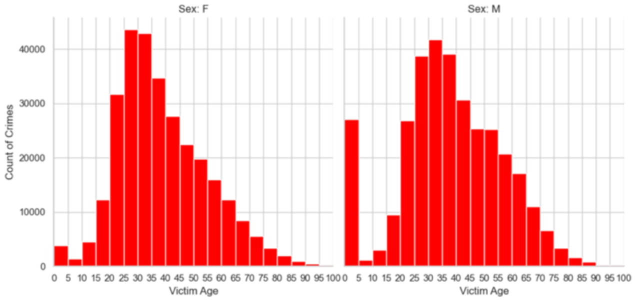


Fig. 13. Victims group by age and gender

According to the figure, most victims are around the age of 25 to 30. It proves that Females are victimized more than males in the age group 25 to 30.

d. how much crime has fluctuated in the last 3 years, which type of crime mostly occurs, which premises the crime occurs most, and what kind of weapons are mostly used for crime?

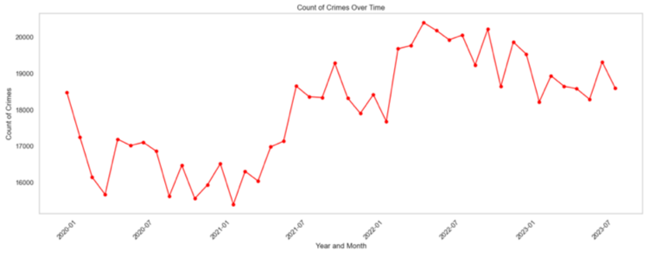


Fig. 14. Line graph showing the crime count trend

From the line graph, we can see that the crime count dropped in the months of 2020. Probably, it happened because people were inside due to covid situation. Later, crime started to increase in 2021. We can see that 2022 has the highest crime count. Till September 2023 the crime count is still lower than 2022 and there is a low chance to increase as the most crime-prone months July and August have already passed.

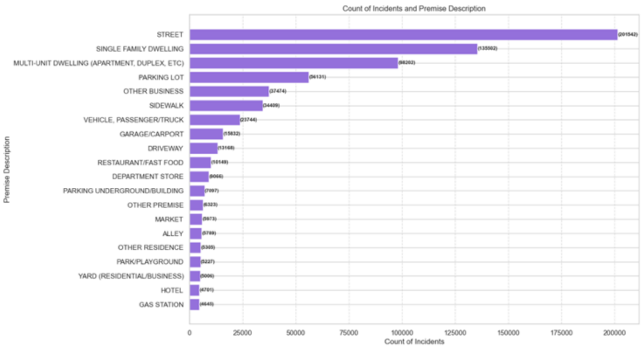


Fig. 15. The highest crime occurrd premise

According to figure Street(201542), Single Family Dwellings (135502), Multi-Unit Dwellings (98202), and Parking lots(56131) are the top four places for Crimes.

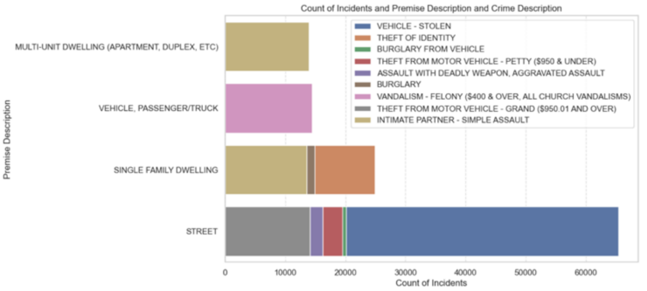


Fig. 16. Showing the highest occurred premise with the type of crimes

We can see the Stolen Vehicle, Theft from Motor Vehicle and Burglary from Vehicle are most common Crime Types on Street(which is the most crime occurred place)

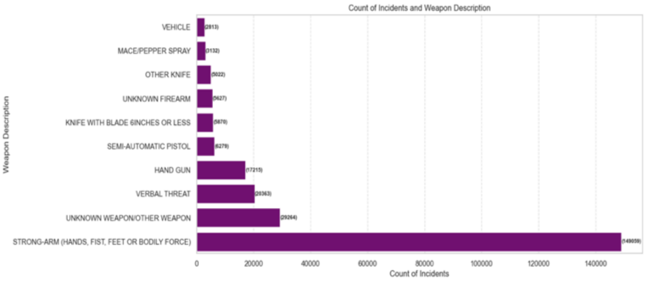


Fig. 17. Type of weapons used for most of the crimes

We can observe that No weapon has been used for most of the crimes. This also means that most of the crimes are not directly life-threatening.

e. What are the top 3 impacted demographics in LA?

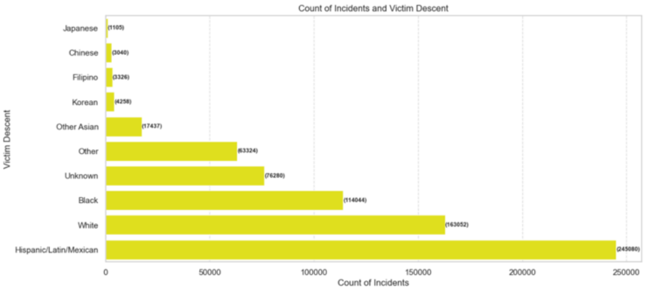


Fig. 18. Higest impacted descent

For all the crimes reported from January 2020 to September 2023, Latin(245080) is the highest victim of crimes, White (163052) is the second highest and Black(114044) is the third. So, it seems 31% are of Latin descent, 20% are white and 14% are black.

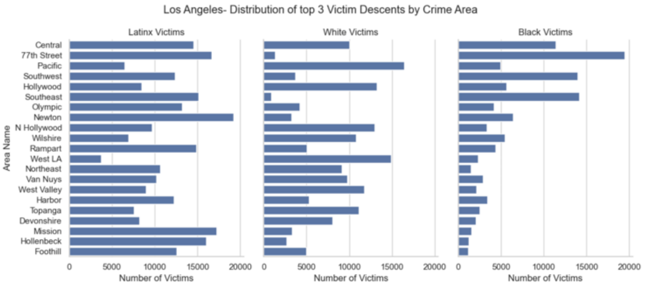


Fig. 19. The areas where the top three descent are impacted

Looking at the top 3 vulnerable demographics in LA, Black, and Latinx communities have suffered in the highest crime-occurring areas of 77th Street and Southeast. This could also be that fewer white people are living in these two areas. White victims are highest Pacific area, West LA, and Hollywood, this could be because these areas have higher white populations. Similarly, areas such as Mission, Newton, Rampart, Foothill, and Hollenbeck have a higher number of Latinx victims than the other descents.

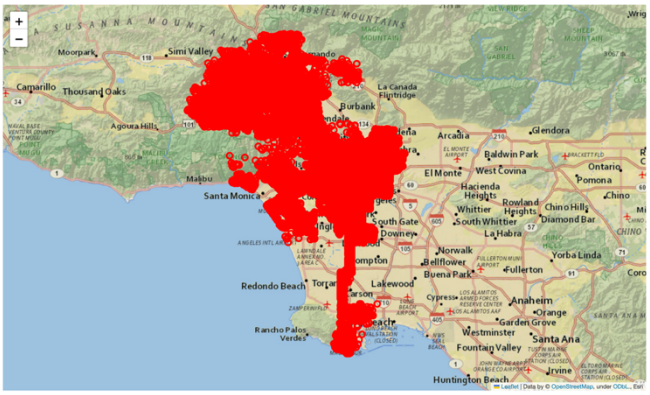


Fig. 20. Map showing the exact location of crime

The map shows the exact locations where the crime occurred.

B. *Result of Prediction*

As I described earlier, I fitted the data in Linear Regression, Autoregressive, Moving Average, and Autoregressive Integrated Moving Average. Calculated the MAE, MSE, and RMSE for each model. Table III shows the result of each model's MAE, MSE, and RMSE.

TABLE III. MAE, MSE, RMSE VALUES OF ML MODELS

MODEL NAME	MAE	MSE	RMSE
LINEAR REGRESSION (LR)	57.02	8364.99	91.46
AUTOREGRESSIVE (AR)	24.02	1790.86	42.32
MOVING AVERAGE (MA)	33.64	1912.75	43.74

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)	38.62	3151.03	56.13

C. *Result of Evaluation*

All the evaluation metrics are the calculation of Error. So, the less will indicate the better model. From the figure, we can see the MAE, MSE, and RMSE values of the Autoregressive model are 24.02, 1790.86, and 42.32 respectively, which are the lowest compared to the rest of the models. We can see that the Autoregressive model will be the best fit for the collected dataset to predict the count of crime for each day.

V. Recommendation

Based on the visualization results and predictions, it is evident that Central, 77th Street and Pacific are the top three areas where crime occurred and was reported, based on LAPD incident reports from January 2022 to September 2023. The majority of street-related crimes are associated with vehicles, emphasizing the importance of vigilance regarding personal belongings when in these areas. Additionally, individuals in the age group of 25-30, both male and female, should exercise increased caution when visiting LA.

A recommended precautionary measure involves the Los Angeles Police Department deploying more units near locations with significant Latin and Black communities. Moreover, the LAPD should enhance readiness during July and August, identified as the months with the highest incidence of crime.

It is noteworthy that a significant proportion of crimes in LA occurs without the use of firearms, suggesting a lower direct threat to life. In summary, LA can be considered safe for visits by taking precautionary measures and being aware of the surrounding conditions.

VI. Future work

The future work aims to further elevate the predictive capabilities of our analytics system by incorporating advanced models, specifically focusing on the implementation of the Prophet Model. Additionally, the study will explore the impact of different features on model performance and conduct ongoing evaluations with incoming data.

VII. Conclusion

This study embarked on a journey to unravel the intricate patterns of crime in Los Angeles. My exploration began with a detailed examination of crime data from the LAPD which spans from 2020 to the present.

The project identified specific areas in LA - Central, 77th Street, Pacific, Southwest, and Hollywood - as hotbeds of criminal activity. The analysis uncovered that the summer months, particularly July and August, along with Fridays, were peak times for criminal activities. This temporal pattern provides a roadmap for strategic policing and community vigilance.

The study highlighted that 25-30 groups are more impacted. It is also noticed that the crime rate dropped in 2020. The probable reason could be COVID-19 locked down. Then observed a significant increase in 2021 and 2022. The study also identified that Latin and Black are mostly impacted in common areas like 77th Street and Southeast.

The project also predicted the crime count for each day using machine learning models. Moving forward, the integration of advanced predictive models and continuous data analysis will be essential. This ongoing effort will not only refine our understanding of urban crime but also enhance the effectiveness of law enforcement strategies.

This study presents a comprehensive picture of the crime landscape in Los Angeles. By harnessing the power of data analysis and predictive modeling, we can take a significant step towards a safer and more informed society. This research is not just a reflection of the current state but a beacon guiding future urban safety initiatives.

References

[1] "Crime Data from 2020 to Present | Los Angeles - Open Data Portal." 2023. December 6, 2023. https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data.

[2] (Wikipedia contributors 2023b)

[3] Patel, R., & Singh, A. K. (2020). Forecasting Crime Rates Using Regression Models. *Journal of Quantitative Criminology*, 36(4), 675-692.

[4] Lee, J., & Kang, M. (2021). Comparative Analysis of Machine Learning Models in Crime Prediction. *Security and Crime Science Review*, 7(3), 45-59.

[5] Ambarish. 2017. "EDA LACrimes -Maps & TimeSeriesForecasts & XGBoost." September 26, 2017. https://www.kaggle.com/code/ambarish/eda-lacrimes-maps-timeseriesforecasts-xgboost.

[6] Ismailsefa. 2020. "Crimes Data Analysis and Visualization (EDA)." Kaggle. March 25, 2020. https://www.kaggle.com/code/ismailsefa/crimes-data-analysis-and-visualization-eda/notebook.

[7] Alex Arnold. n.d. "GitHub - Alexarnold630/Dallas_Crime_Analysis_and_Prediction: Machine Learning Engine Predicting Dallas Crime Incident Status with Visualization Analysis Using Tableau for Crime Time and Locations." GitHub. https://github.com/alexarnold630/Dallas_Crime_Analysis_and_Prediction.

[8] Sidneykung. n.d. "GitHub - Sidneykung/LA_crime_forecasting: Time Series Modeling Project to Forecast LA Reported Crime Rates Based on 10 Years Worth of Recent Data." GitHub. https://github.com/sidneykung/LA_crime_forecasting.