Project Report

# PREDICTING CRIME TYPES IN LOS ANGELES USING MACHINE LEARNING ALGORITHMS

**Group Members:**

1. Anshuman Pradhan (qb8291)
2. Jai Bhala(uw9410)
3. Yash Dodia (ww8176)

# Index

# INTRODUCTION

Predicting crimes is a very complex task that is influenced by many factors some of which include factors based on geographical location or temporal conditions. In addition to using machine learning algorithms, we will specifically focus on incorporating the area and time information into our crime-type prediction models. By considering the area where crimes occur and the times at which they happen, we seek to capture spatial and temporal patterns that may be indicative of different crime types.

The dataset we will be using comprises of detailed information on the date of occurrence, time of occurrence, area of occurrence, and various other attributes related to crimes reported in Los Angeles. By analyzing this data, we can uncover patterns and correlations between the crime types and their corresponding areas and time frames. This analysis will allow us to train machine learning algorithms to recognize and predict crime types based on these key features.

By utilizing the area and time attributes in our predictive models, we aim to better the accuracy and specificity of crime-type predictions. This information can provide valuable insights to law enforcement agencies by enabling them to allocate resources effectively, implement targeted crime prevention strategies, and deploy personnel in high-risk areas during specific time periods. Ultimately, our goal is to empower law enforcement agencies with reliable crime-type predictions based on area and time, enabling them to proactively combat criminal activities and enhance public safety in Los Angeles.

# Dataset

For this project, we will utilize the "Crime_Data_from_2020_to_Present.csv" dataset, which is publicly available on the Los Angeles Open Data platform (https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8). This dataset has been sourced directly from the Los Angeles Police Department and contains information on various crimes committed within the city. It consists of 739687 rows and 28 columns, providing detailed attributes such as the date of occurrence, time of occurrence, area of occurrence, premise where the crime took place, crime type, latitude, longitude, the district where it was reported, and victim details including age, sex, and descent. Columns in this Dataset:

| Sequence | Variable Name | Description | Possible Values |
|---|---|---|---|
| 1 | DR_NO | Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5 digits | |
| 2 | Date Rptd | Date Reported | Dates ranging from 2020 to July 2023 |
| 3 | DATE OCC | Date of Occurrence | Dates ranging from 2020 to July 2023 |
| 4 | TIME OCC | Time of Occurrence (in 24 hour military time). | 24 hr time |
| 5 | AREA | The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21. | Area Code ranging from 1-21 |
| 6 | AREA NAME | 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. | 77th Street, Central, Devonshire, Foothill, Harbor, Hollenbeck, Hollywood etc. |
| 7 | Rpt Dist No | A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. | |

| | | | |
|---|---|---|---|
| 8 | Part 1-2 | | |
| 9 | Crm Cd | A unique code for the crime committed. | 113,121,122 etc. |
| 10 | Crm Cd Desc | Defines the Crime Code provided. | ARSON, ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER, ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT etc. |
| 11 | Mocodes | Modus Operandi Codes | 100,101,102 etc. |
| 12 | Vict Age | Victim Age | |
| 13 | Vict Sex | Victim Sex | F - Female M - Male X - Unknown |
| 14 | Vict Descent | Victim Descent. | A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino etc. |
| 15 | Premis Cd | Code for the premise where the crime took place. | 101,102,103, 14,105 etc. |
| 16 | Premis Desc | Defines the Premise Code provided. | ALLEY, BANK, BUS STOP, BEACH, etc. |
| 17 | Weapon Used Cd | Code for the type of weapon used in the crime. | 101,102,103, 14,105 etc. |
| 18 | Weapon Desc | Definition for the weapon code | AIR PISTOL/REVOLVER/RIFLE/BB GUN, ANTIQUE FIREARM, etc. |
| 19 | Status | Status of the case. (IC is the default) | AA, AO, CC, IC, JA, JO |
| 20 | Status Desc | Defines the Status Code provided. | Adult Arrest, Adult Other, Invest Cont, Juv Arrest, Juv Other, UNK |
| 21 | Crm Cd 1 | Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. | |
| 22 | Crm Cd 2 | May contain a code for an additional crime, less serious than Crime Code 1. | |
| 23 | Crm Cd 3 | May contain a code for an additional crime, less serious than Crime Code 1. | |

| | | | |
|---|---|---|---|
| 24 | Crm Cd 4 | May contain a code for an additional crime, less serious than Crime Code 1. | |
| 25 | LOCATION | Street address of crime incident rounded to the nearest hundred block to maintain anonymity. | |
| 26 | Cross Street | Cross Street of rounded Address | |
| 27 | LAT | Latitude | |
| 28 | LON | Longtitude | |

# PROJECT OBJECTIVES

The primary objective of this project is to develop accurate crime prediction models using machine learning algorithms. To achieve this, we have divided our analysis into three classifications: multiclass prediction, binary class prediction (violent and non-violent crimes), and prediction of specific crime types within the category of violent crimes.

- **Multiclass Prediction:**
  In the multiclass prediction classification, we aim to predict all the crime types present in the dataset. By training machine learning algorithms on historical data, we seek to develop a model capable of classifying crimes into their respective types accurately. This approach provides a broader understanding of crime patterns thus aiding in strategic decision-making for law enforcement agencies.
- **Bi-class Prediction (Violent and Non-violent Crimes):**
  In the bi-class prediction classification, our focus shifts to differentiating between violent and non-violent crimes. Using the machine learning algorithms specifically for this binary classification task, we aim to develop a model that can effectively distinguish between these two major categories.
- **Prediction for Specific Crime Types within the Violent Crime Category:**
  Within the category of violent crimes, we aim to further refine our prediction models to identify specific crime types. By focusing solely on crimes categorized as violent, we develop specialized models that differentiate between various violent offenses such as assault, robbery, or battery.

# Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. It involves transforming raw data into a clean, consistent, and suitable format for analysis and model training. By performing various preprocessing techniques, we can improve the quality of the data, address missing values, handle outliers, and transform features to make them more informative for machine learning algorithms. In the context of predicting crime types in Los Angeles, data preprocessing plays a vital role in ensuring accurate and reliable results.

| DR_NO | Date Rptd | DATE OCC | TIME OCC | AREA | AREA NAME | Rpt Dist No | Part 1-2 | Crm Cd | Crm Cd Desc | ... | Status | Status Desc | Crm Cd 1 | Crm Cd 2 | Crm Cd 3 | Crm Cd 4 | LOCATION | Cross Street | LAT | LON |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10304468 | 01/08/2020 12:00:00 AM | 01/08/2020 12:00:00 AM | 2230 | 3 | Southwest | 377 | 2 | 624 | BATTERY - SIMPLE ASSAULT | ... | AO | Adult Other | 624.0 | NaN | NaN | NaN | 1100 W 39TH PL | NaN | 34.0141 | -118.2978 |
| 190101086 | 01/02/2020 12:00:00 AM | 01/01/2020 12:00:00 AM | 330 | 1 | Central | 163 | 2 | 624 | BATTERY - SIMPLE ASSAULT | ... | IC | Invest Cont | 624.0 | NaN | NaN | NaN | 700 S HILL ST | NaN | 34.0459 | -118.2545 |
| 200110444 | 04/14/2020 12:00:00 AM | 02/13/2020 12:00:00 AM | 1200 | 1 | Central | 155 | 2 | 845 | SEX OFFENDER REGISTRANT OUT OF COMPLIANCE | ... | AA | Adult Arrest | 845.0 | NaN | NaN | NaN | 200 E 6TH ST | NaN | 34.0448 | -118.2474 |
| 191501505 | 01/01/2020 12:00:00 AM | 01/01/2020 12:00:00 AM | 1730 | 15 | N Hollywood | 1543 | 2 | 745 | VANDALISM - MISDEAMEANOR ($399 OR UNDER) | ... | IC | Invest Cont | 745.0 | 998.0 | NaN | NaN | 5400 CORTEEN PL | NaN | 34.1685 | -118.4019 |
| 191921269 | 01/01/2020 12:00:00 AM | 01/01/2020 12:00:00 AM | 415 | 19 | Mission | 1998 | 2 | 740 | VANDALISM - FELONY ($400 & OVER, ALL CHURCH VA... | ... | IC | Invest Cont | 740.0 | NaN | NaN | NaN | 14400 TITUS ST | NaN | 34.2198 | -118.4468 |

Fig. 1- First 5 entries of the data frame

The data used for crime prediction in Los Angeles is sourced from the Los Angeles Police Department and contains information such as the date and time of occurrence, area, crime codes, victim demographics, and geographic coordinates. However, raw data often contains inconsistencies, missing values, and irrelevant features, which can hinder the effectiveness of machine learning models. Therefore, data preprocessing becomes essential to mitigate these challenges and extract meaningful insights from the dataset.

In data preprocessing the following steps were followed:

## Removing Irrelevant Columns

```
# Removing irrelevant/not meaningful attributes from description of columns data. there are some columns don't needed so we drop them
col = ['Date Rptd','DR_NO','AREA', 'Mocodes', 'Premis Desc', 'Weapon Desc', 'Status Desc','Crm Cd 1',
       'Crm Cd 2', 'Crm Cd 3', 'Crm Cd 4', 'Cross Street','Weapon Used Cd']
df = df.drop(col,axis=1)
```

Fig 2- Removing columns not needed in analysis.

In the initial dataset, certain columns do not provide meaningful information for predicting crime types or are irrelevant to the analysis. These columns, such as "Date Rptd," "DR_NO," "AREA," "Mocodes," "Premis Desc," "Weapon Desc," "Status Desc,"

"Crm Cd 1," "Crm Cd 2," "Crm Cd 3," "Crm Cd 4," and "Cross Street," are dropped from the data frame. This step ensures that the remaining columns are relevant and contribute to the prediction task, reducing noise and computational overhead.

## Handling Missing Values:

```
# Handling null values
df.isna().sum()/df.shape[0]
# filling victim sex empty rows with "X" unknown symbol
df["Vict Sex"].fillna("X",inplace=True)

# filling victim descent empty rows with "X" unknown symbol
df["Vict Descent"].fillna("X",inplace=True)

#dropping any empty rows existing after this
df.dropna(inplace=True)
```

Fig. 3 – Filling missing/null values

Missing values are a common issue in real-world datasets, and they can impact the performance of machine-learning models. To address missing values in the dataset, we perform the following steps:

Calculate the number of missing values in each column using the "isna().sum()" function. This provides an overview of the missing data in the dataset. For categorical columns such as "Vict Sex" and "Vict Descent," missing values are filled with the symbol "X" to represent an unknown category.

Any rows with remaining missing values after filling categorical columns are dropped from the dataframe. This step ensures that the dataset used for training and evaluation is complete and does not introduce bias due to missing information.

## Data Type Handling:

To ensure proper handling and utilization of the data, several data type conversions and extractions are performed. These steps involve transforming certain columns to different data types and extracting relevant information from existing columns. The data type handling steps applied to the dataset are as follows:

Converting the date column to datetime:
The "DATE OCC" column, which contains the date of occurrence of each crime, is converted from a string format to a datetime format using the "pd.to_datetime()" function. This conversion enables efficient manipulation and analysis of temporal data.

Extracting year, month, day, and day of the week:
To capture the temporal aspects of the data, the year, month, day, and day of the week information is extracted from the "DATE OCC" column. Separate columns, namely "year," "month," "day," and "day_of_week," are created to store these extracted values. This allows for more granular analysis based on different time components.

Converting time to string and padding with zeros:
The "TIME OCC" column, representing the time of occurrence of each crime, is converted to a string data type. This conversion facilitates further data manipulation and consistency in handling time-related information. Additionally, a padding operation is performed to ensure that all time values have a consistent length of four digits.

Extracting hour and minute:
To gain more detailed insights into the time aspect, the "TIME OCC" column is split into separate columns, "hour" and "minute," by extracting the corresponding portions from the time string. This separation enables the analysis of crime patterns at different times of the day.

```
# Converting the date column to datetime
df['DATE OCC'] = pd.to_datetime(df['DATE OCC'], format='%m/%d/%Y %I:%M:%S %p')

# Extracting year, month, day into separate columns
df['year'] = df['DATE OCC'].dt.year
df['month'] = df['DATE OCC'].dt.month
df['day'] = df['DATE OCC'].dt.day
df['day_of_week'] = df['DATE OCC'].dt.day_name()

# Converting time to string
df['TIME OCC'] = df['TIME OCC'].astype('str')

# converting the column with 4 digits
df['TIME OCC'] = df['TIME OCC'].apply(lambda x: x.zfill(4))

# extracting hour and minute
df['hour'] = df['TIME OCC'].str[:2]
df['minute'] = df['TIME OCC'].str[2:]
```

Fig. 4- Data Type Handling

By performing these data type handling steps, we ensure that the necessary temporal information is properly extracted and represented in a format suitable for analysis and model training. These transformations enhance the ability to uncover patterns and relationships between crime types, areas, and specific time periods in Los Angeles.

## Mapping Crime Types:

The original dataset contains a wide range of crime types, which can make the prediction task complex and challenging. To simplify the problem, we map the crime types to a reduced set of categories. This step involves defining a function called "Crime_type()" that takes the original crime type as input and maps it to one of the selected crime categories.

```
# due to the huge number of Labels we should map them to reduce its numbers
def Crime_type(crime):

    types = ["STOLEN",'BATTERY','THEFT', 'BURGLARY', 'VANDALISM', 'ASSAULT', 'CRIMINAL THREATS',
             'TRESPASSING','VIOLATION',"CRIMINAL HOMICIDE", 'CHILD ABUSE', 'RAPE','ROBBERY']

    types_name = ["MOTOR VEHICLE THEFT",'BATTERY','THEFT', 'BURGLARY', 'VANDALISM', 'ASSAULT', 'CRIMINAL THREATS',
              'TRESPASSING','VIOLATION',"CRIMINAL HOMICIDE", 'CHILD ABUSE', 'RAPE','ROBBERY' ]

    for name,t in zip(types_name,types) :
        if t in crime:
            return name

    return "OTHERS"
```

```
# apply function on each row
df["Crime"] = df["Crm Cd Desc"].apply(Crime_type)
```

```
df["Crime"].value_counts()
```

```
THEFT                  199071
ASSAULT                 97413
BURGLARY                94335
MOTOR VEHICLE THEFT     91767
VANDALISM               65798
BATTERY                 64073
OTHERS                  52670
ROBBERY                 29164
CRIMINAL THREATS        15410
VIOLATION               15241
TRESPASSING             10299
RAPE                     3177
CRIMINAL HOMICIDE        1259
Name: Crime, dtype: int64
```

Fig. 5- Crime mapping

The function "Crime_type()" compares the original crime type with a set of pre-defined crime types, such as "STOLEN," "BATTERY," "THEFT," "BURGLARY," "VANDALISM," "ASSAULT," "CRIMINAL THREATS," "TRESPASSING," "VIOLATION," "CRIMINAL HOMICIDE," "CHILD ABUSE," "RAPE," and "ROBBERY."

If the original crime type matches any of these categories, it is assigned to the corresponding category.

If the original crime type does not match any of the pre-defined categories, it is categorized as "OTHERS."

This mapping process reduces the number of crime categories and simplifies the prediction task, making it more manageable for machine learning algorithms.

# Data Transformation:

The "Crime Type" column is added to the dataframe, providing a simplified representation of the original crime types.

```
# Converting the following into binary clasification by dividing the crime type into violent and non violent

def categorize_crime(crime_type):
    if crime_type in ['ASSAULT','BATTERY', 'ROBBERY', 'RAPE', 'CRIMINAL HOMICIDE']:
        return 'violent'
    else:
        return 'non-violent'

df['Crime_Type'] = df['Crime'].apply(categorize_crime)

print(df['Crime_Type'].value_counts())
```

```
non-violent    450971
violent        190650
Name: Crime_Type, dtype: int64
```

Fig. 6- Adding new column “crime_type”

This data transformation step simplifies the prediction task by reducing the number of classes and allows us to focus on a smaller, more manageable set of crime categories.

By performing these data preprocessing steps, we ensure that the dataset used for predicting crime types in Los Angeles is cleaned, transformed, and suitable for training machine learning models. The removal of irrelevant columns reduces noise and computational overhead, handling missing values ensures completeness and reliability, mapping crime types simplifies the prediction task, and data transformation provides a reduced set of crime categories. These steps pave the way for effective model development and accurate predictions.

# Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a very important step in the data analysis process as it gives the user an idea about the dataset before applying machine learning algorithms. In our dataset, EDA helped us understand patterns, look for any trends and discover how the data is distributed. Below is a brief overview of the key aspects of EDA for this project:

## Distribution of Crimes:

From the dataset we can get a comprehensive overview of the various crimes in Los Angeles. Amongst the multiple crime categories, “Theft” stood out as the most recorded crime. “Assault” was second most recorded crime type, but it was almost 53% less than “theft”. Then the various crime types in order were “Motor Vehicle Theft” followed by “Assault”, “Vandalism”, “Battery”, and “Robbery”. also contribute significantly to the overall crime landscape. On the other hand, categories such as criminal homicide and rape have relatively lower reported cases. The following distribution gives an idea about the most occurring crimes thus giving the stakeholders an idea on where to allocate resources and implement specialized measures to prevent crimes.

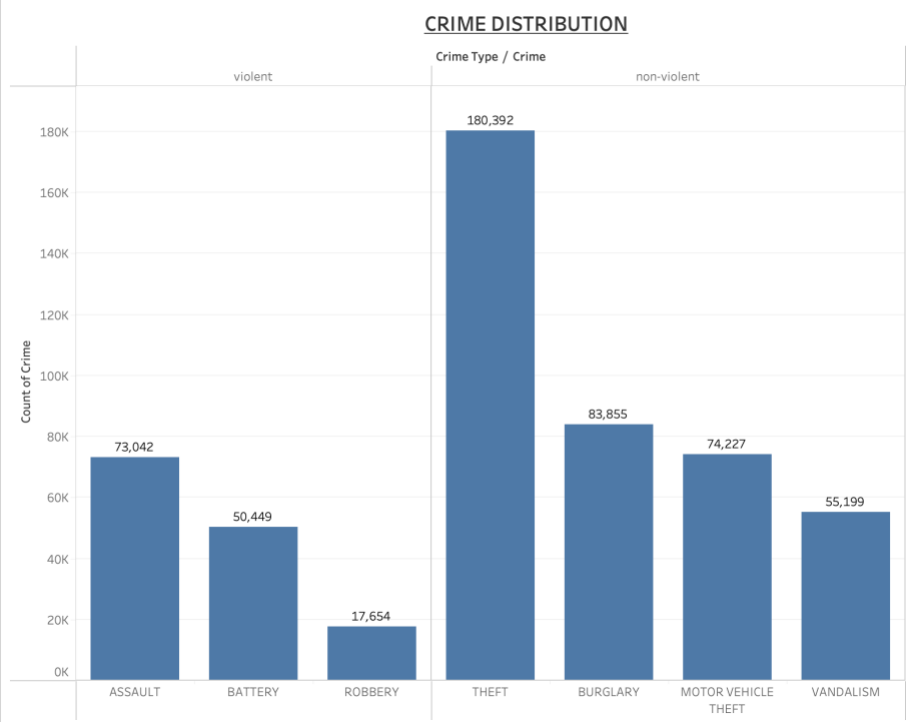


Fig. 7- Crime Distribution

Now we adopt a second point of view, where the crimes are categorized as either “Violent” or “Non-Violent”. Most of the reported incidents fall under non-violent category which comprises of

“Theft”, “Burglary”, “Motor Vehicle Theft”, “Vandalism”, and others. The non-violent crimes comprise a huge chunk of the dataset. Meanwhile, the “Violent” category consisting of “Assault”, “Battery”, “Robbery’, and other violent offenses, represents a substantial number of reported incidents as well.

For the various law enforcement agencies to create successful tactics and allocate resources to protect public safety, they must have a thorough understanding of the balance between violent and nonviolent crimes.

## Geographical Analysis:

Having an understanding about the how the crimes are distributed across the region. This in turn helps in effective crime prevention and an efficient resource allocation.
In this analysis, we explore various aspects of geographic patterns in crime within Los Angeles.

## Heat Map:

We have used head map to visualize the intensity of the crimes recorded across different regions in Los Angeles. The Central region of the city has a higher concentration of crimes in comparison to the North and South regions. We can infer this concentration as areas with higher crime rates, often referred to as crime hotspots. Hotspot identification in turn enables the law enforcement agencies to address the fundamental causes of crime in those specific locations, strategically allocate resources and carry out focused interventions.

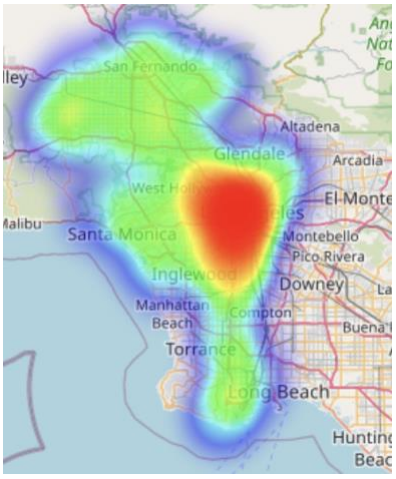


Fig.8- Heat Map- 1

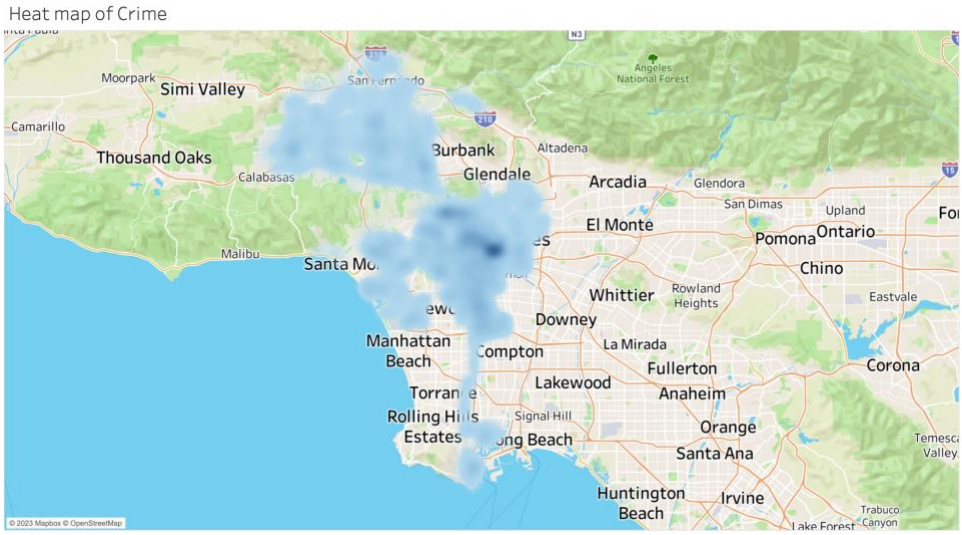


Fig.9 Heat Map-2

## Map with Crime Type:

Here the various crime types are overlaid on a map, by doing so we seek to uncover any spatial patterns or any correlation between the crime types and any specific region. However, we can infer from the plot below that there are no discernible trends in crime types based on geographic areas. The crime types overlap across different regions making it difficult to comprehend any clear spatial relationship between the crime types and specific regions. This finding suggests that there may be some foreign factors other than the geographic location (such as socio-economic factors, individual circumstances etc.) that have a stronger influence on the occurrence of these crimes.

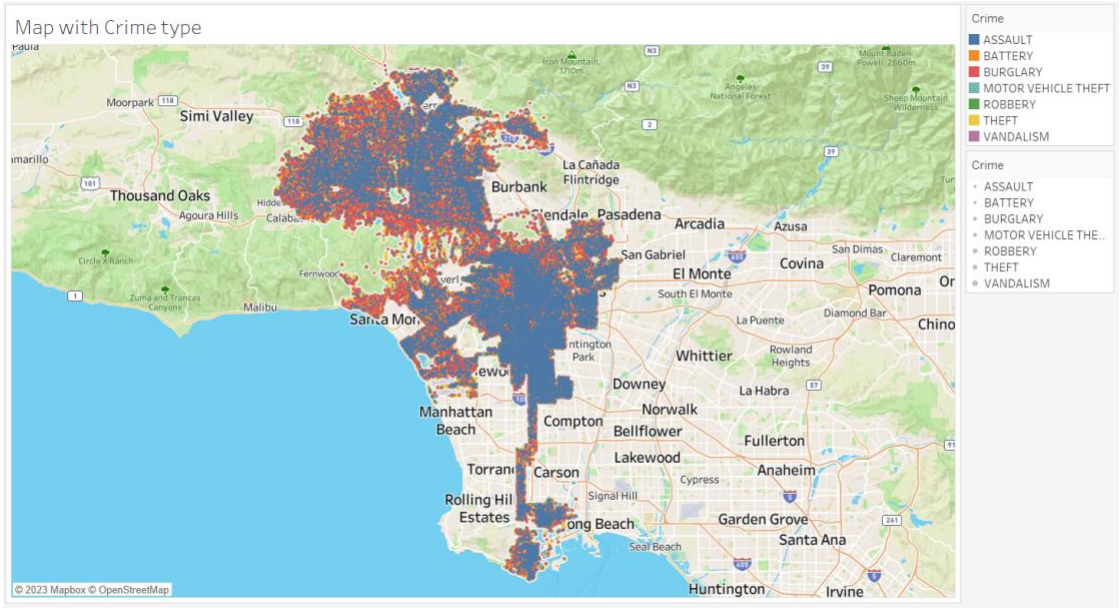


Fig. 10 Map with individual crime types

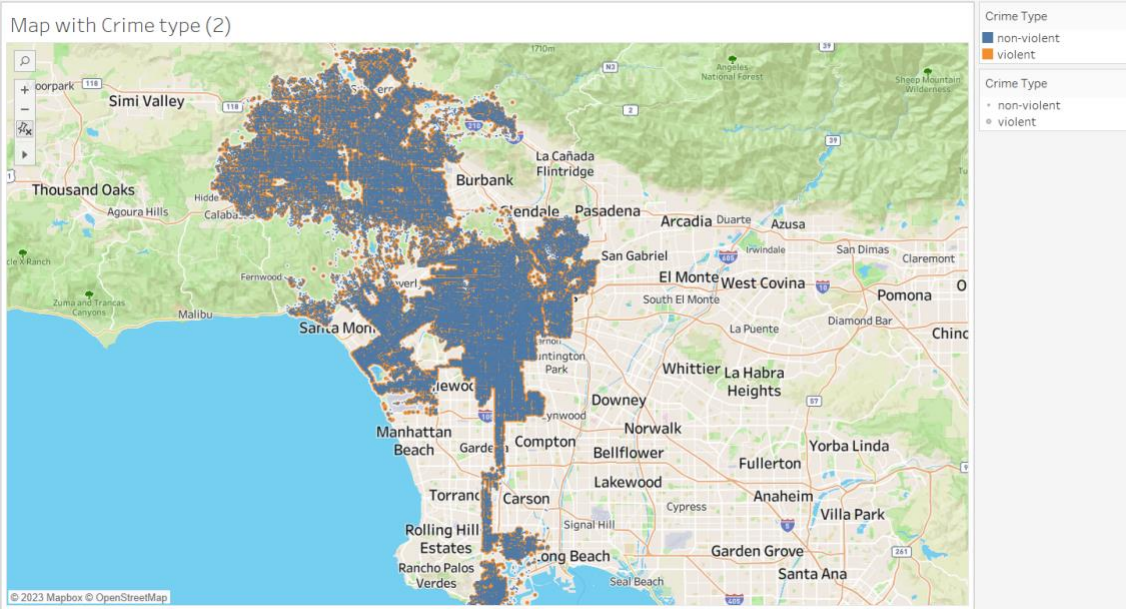


Fig. 11- Map where the crime types are either Violent or Nonviolent

## Crime Over Region:

In this scenario we have considered how the number of crimes recorded has been distributed across different regions of Los Angeles. The plot below highlights that region like Central, Pacific, 77th Street, Southwest, and N Hollywood exhibit some of the highest number of reported crimes. This implies that these regions require more surveillance and more monitoring so that the count of occurrences can be reduced. On the other hand, regions like Harbor, Mission, Hollenbeck, and Foothill have comparatively lower crime rates. By being aware of these variations in crime rates throughout different regions, law enforcement organizations can concentrate their efforts and resources in areas that require greater attention while still taking into account relatively safer localities.

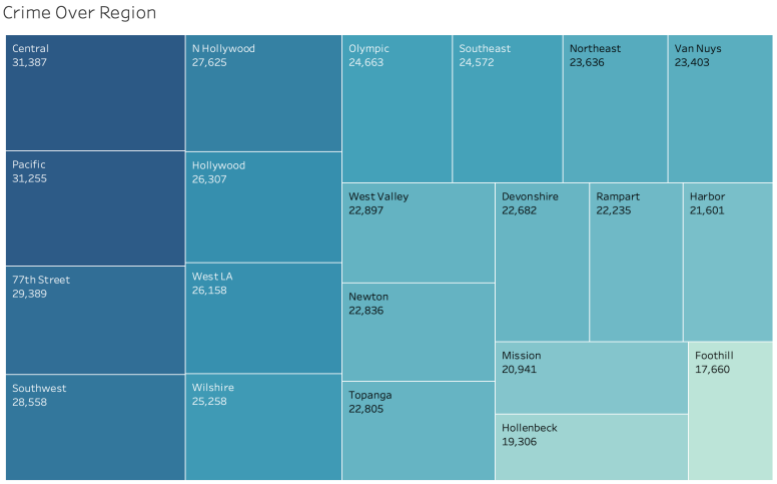

Fig. 12- Recorded crimes over different regions

Overall, the spatial analysis of crime trends in Los Angeles helps to create evidence-based plans for resource allocation, targeted interventions, and crime prevention. Law enforcement organizations can attempt to make communities safer and improve public safety by taking into account the spatial distribution of crime and its link with different causes.

## Crime Over Premise:

The plot below reveals that nearly 70% of the reported crimes take places across 12 similar premises which have categorized into premise codes. This is a strong indicator that a small number of particular premises contribute to majority of criminal incidents. By identifying establishments or dwellings that fall under these 12 premises, law enforcement can greatly decrease the number of criminal incidents.

Some measure to reduce the crime count may include increasing police presence, implementing surveillance systems, enhancing community engagement, and implementing preventive measures tailored to the specific premises associated with high crime rates.

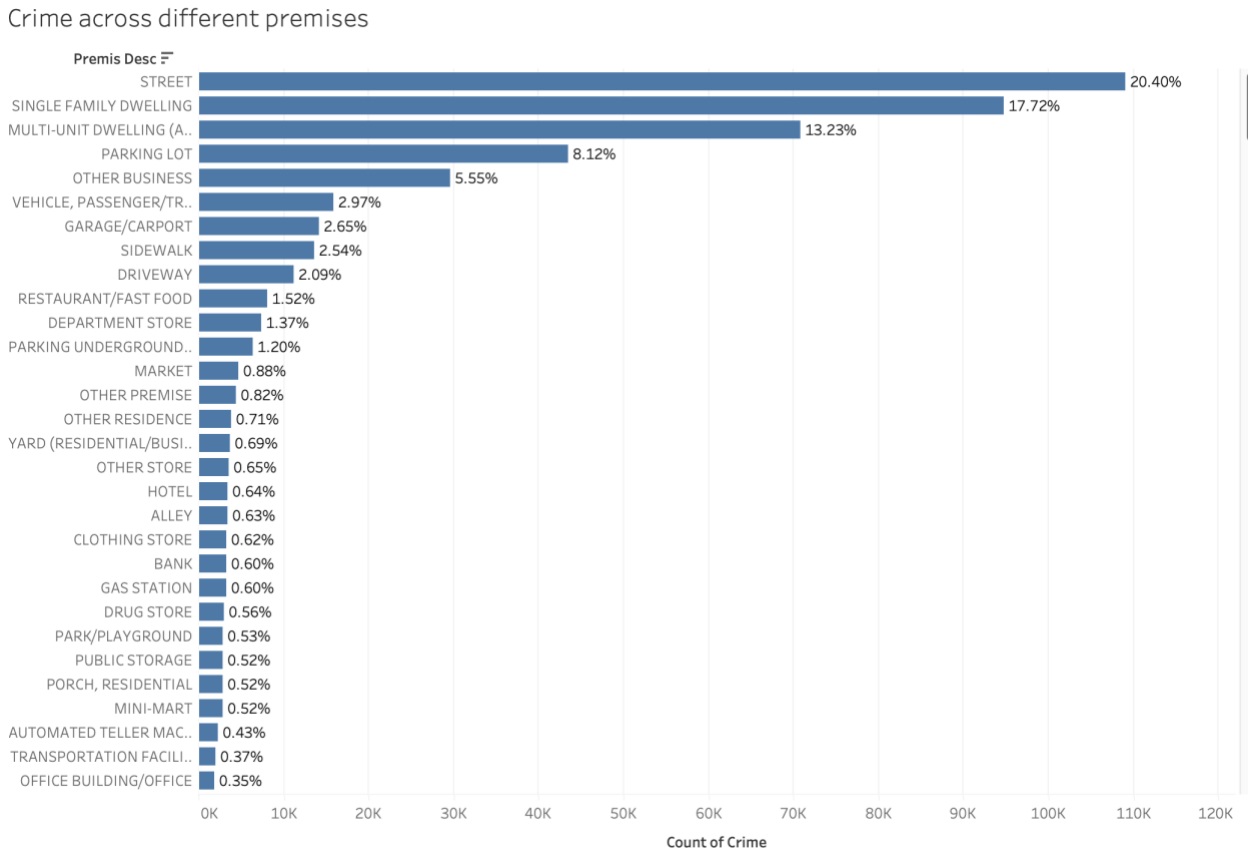


Fig. 13 Crime count across different premises

## Temporal Patterns:

Temporal analysis of crime patterns can be used to gather insights into the variations in criminal activities based on differing time-related factors. We examine the time patterns of crime in Los Angeles in this analysis based on two important factors: crime over an hour and crime over weekdays.

## Crime over an Hour:

Upon examination of the distribution of crime over a 24-hour period, we observe that there is a clear trend in the total number of crimes reported throughout the day. From the plot we can see that the total number of crimes generally increases from 1:00 AM to 12:00 PM, peaking at 12:00 PM (noon), and then gradually decreases in the afternoon and evening hours. From this temporal pattern it is safe to conclude that the highest concentration of criminal activities occurs during the daytime hours, particularly around midday. By using these temporal trends law enforcement agencies can efficiently allocate resources and implement targeted measures during the times of the day when crime rates are highest.

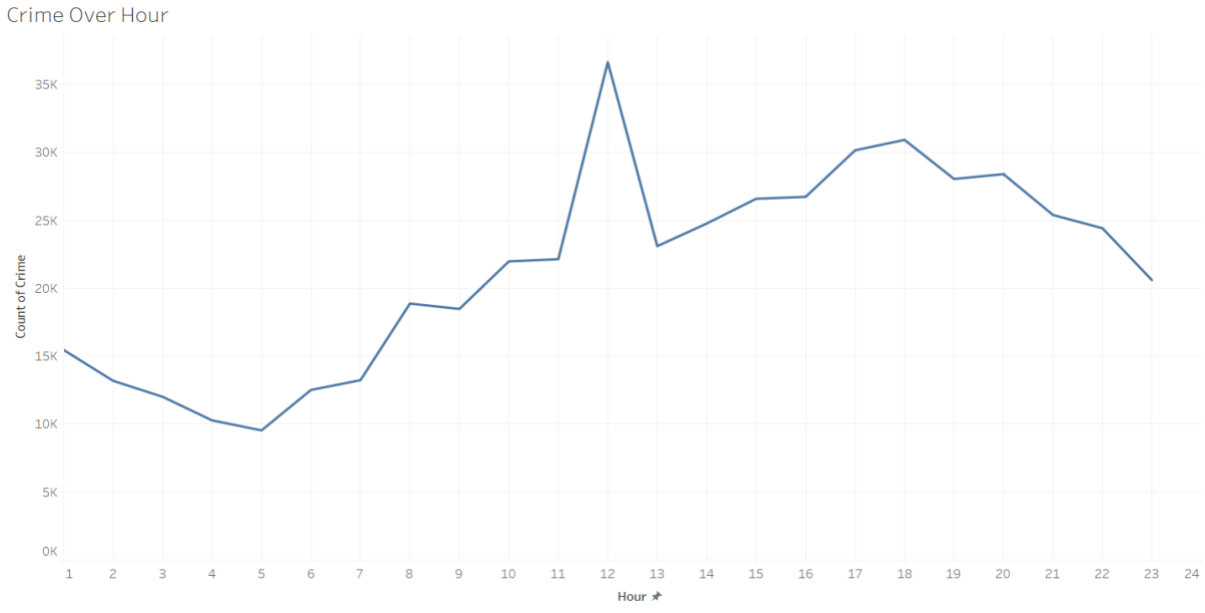


Fig. 14 Crime count across different hours

## Crime over Days of week:

By Analyzing the distribution of crime over weekdays we get to know about any variations in criminal activities based on different days of the week. The plot below shows that there is relatively little difference in the total count of crimes reported across weekdays. However, when compared to other days of week Friday and Saturday tend to have the highest number of reported crimes, with approximately 82,000 and 77,000 crimes, respectively. On the other hand, Sunday records the lowest count of reported crimes, with 73,508 crimes. These findings suggest that weekends, particularly Friday and Saturday, may experience higher levels of criminal activities, while Sundays tend to have slightly lower crime rates. Understanding these temporal patterns can help law enforcement agencies allocate resources and enhance patrol efforts during peak crime days and times.

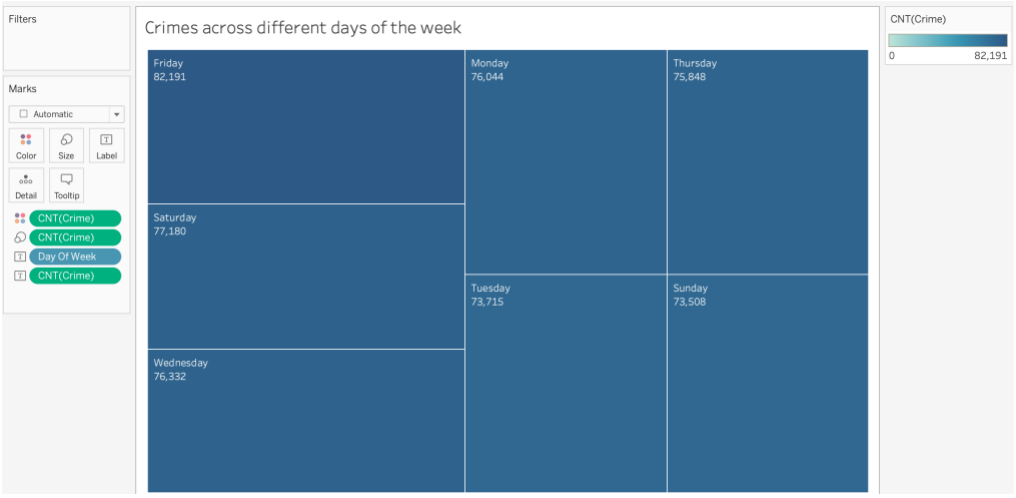


Fig. 15- Crime over days of week

The analysis of crime data in Los Angeles provides valuable insights for law enforcement agencies. By understanding the distribution of crime types, geographic patterns, high-frequency premise codes, and temporal trends, agencies can allocate resources effectively, target interventions, and optimize patrol schedules. These findings enable proactive strategies for crime prevention, resource allocation, and enhanced public safety in Los Angeles.

## Total Crime by Crime Type and Area

The plot depicts the total count of different crime types across various regions. It provides valuable insights into the crime distribution and highlights the prevalence of specific crime types in different areas. The analysis reveals that theft is the most common crime across all regions, indicating a higher frequency of theft incidents compared to other crime types.

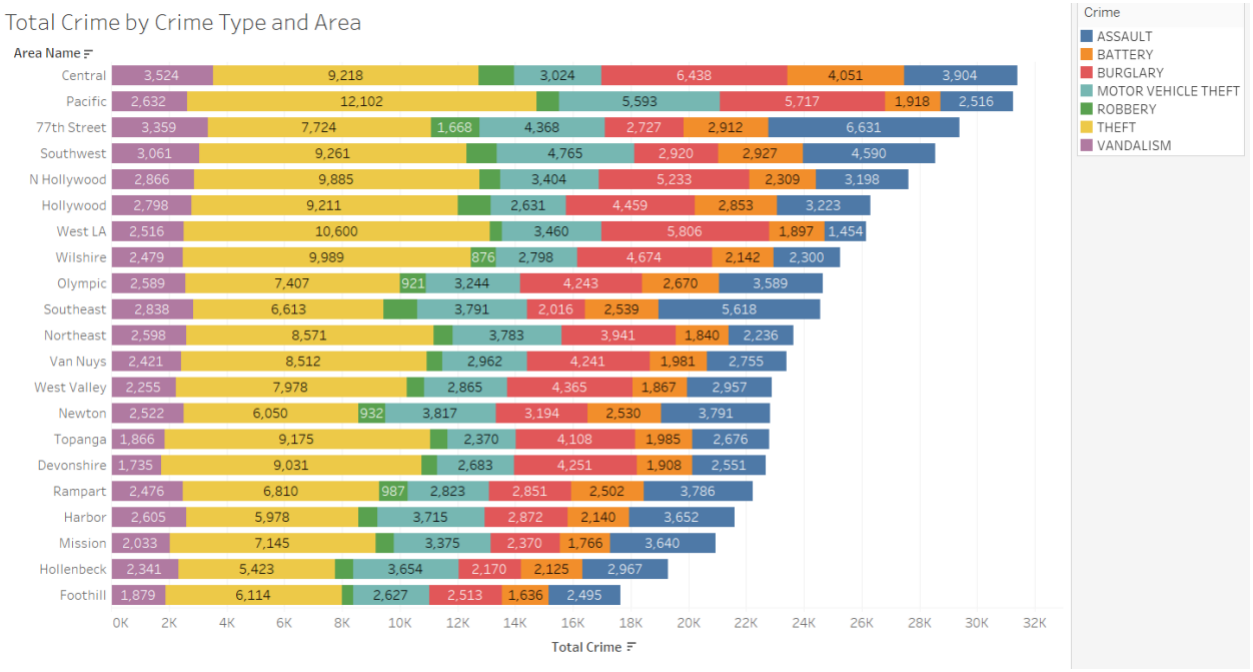


Fig 16. - Total Crime by Crime Type and Area

Burglary is the second most prevalent crime in the dataset, occurring with a relatively lower frequency compared to theft. This suggests that burglaries are less frequent but still significant in terms of occurrence across the regions. The lower count of burglaries compared to theft may imply a higher level of security measures taken by individuals or businesses to protect their properties.

On the other hand, robbery is observed to have the lowest count among the three crime types. This indicates that robberies occur less frequently in comparison to theft and burglary incidents. The lower occurrence of robberies may be attributed to various factors, including the presence of effective law enforcement measures, public awareness, and the implementation of security protocols.

By analyzing the plot, it becomes evident that theft is a prevalent crime across all regions, followed by burglaries, while robberies have the least occurrence. This information can be useful for law enforcement agencies and policymakers to allocate resources and devise crime prevention strategies accordingly. By focusing on addressing theft-related crimes, authorities can make a significant impact on reducing overall crime rates and enhancing public safety. Additionally,

understanding the patterns of burglaries and identifying areas with higher incidences can aid in implementing targeted preventive measures to minimize these crimes.

## Crime over top 10 Premise code with type

The plot illustrates the distribution of crime types across the top 10 premise codes with the highest crime rates. It offers valuable insights into the specific premise codes associated with the highest number of criminal incidents and the corresponding patterns in the types of crimes committed within each code.

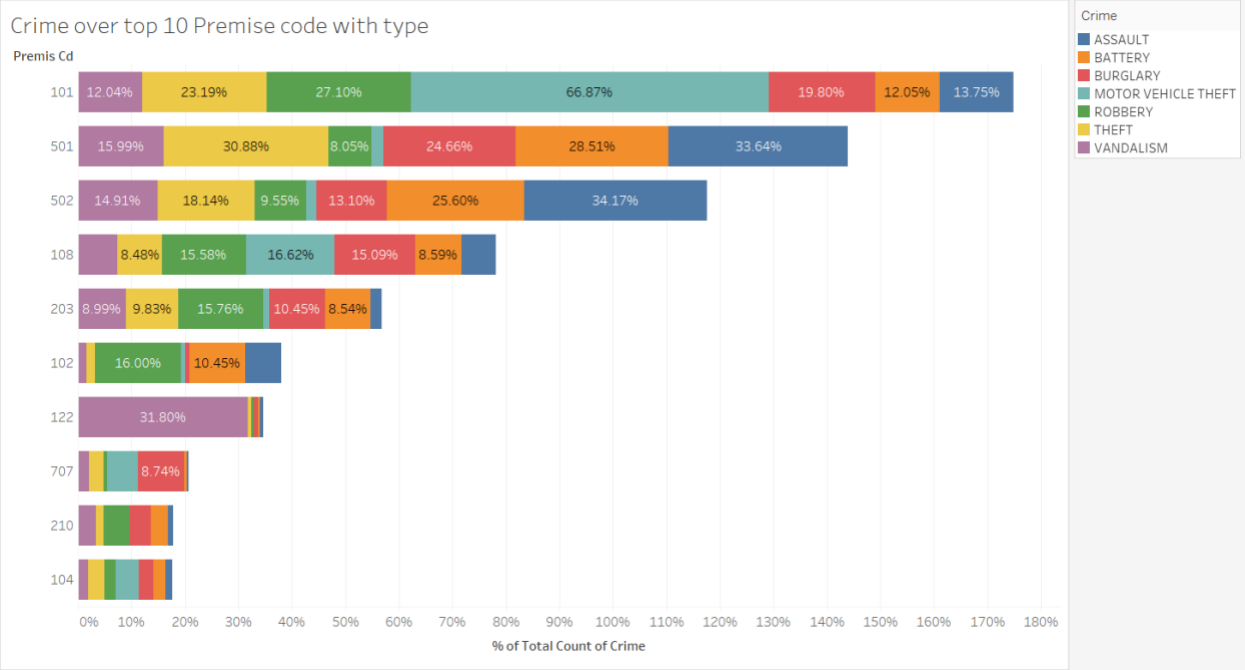


Fig. 17 – Crime over premise code with type

The analysis of the plot reveals that premise code 101 has the highest number of crimes, and approximately 70% of these cases are attributed to motor vehicle theft. This indicates a significant prevalence of motor vehicle thefts within the context of premise code 101. The high proportion of motor vehicle theft suggests that this particular premise code may be associated with locations or circumstances that make vehicles more susceptible to theft, such as parking lots or areas with inadequate security measures. Understanding this pattern can help law enforcement agencies allocate resources and implement targeted strategies to prevent motor vehicle thefts in these specific areas.

Code 122, on the other hand, is associated with around 32% of vandalism cases. This indicates that a substantial portion of crimes under premise code 122 involves acts of vandalism. The higher percentage of vandalism within this premise code suggests that it may be linked to specific locations or situations where acts of vandalism are more likely to occur, such as public parks,

abandoned buildings, or areas with limited surveillance. Recognizing this pattern enables authorities to prioritize surveillance efforts, increase patrols, or implement preventative measures to mitigate vandalism incidents and protect public and private property.

# Modelling and Results

**Feature Selection:**
It refers to the process of selecting features from the dataset that are more influential on predicting the crime. This procedure is important as it reduces complexity and increases the performance of the models.

```
features = ['AREA NAME','Rpt Dist No','Premis Cd','Block','Street','LAT','LON','hour','day_of_week','Crime','Crime_Type']

df = df[features]
df.head()
```

| | AREA NAME | Rpt Dist No | Premis Cd | Block | Street | LAT | LON | hour | day_of_week | Crime | Crime_Type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Southwest | 377 | 501 | 1100 | W 39TH PL | 34.0141 | -118.2978 | 22 | Wednesday | BATTERY | violent |
| 1 | Central | 163 | 102 | 700 | S HILL ST | 34.0459 | -118.2545 | 03 | Wednesday | BATTERY | violent |
| 3 | N Hollywood | 1543 | 502 | 5400 | CORTEEN PL | 34.1685 | -118.4019 | 17 | Wednesday | VANDALISM | non-violent |
| 4 | Mission | 1998 | 409 | 14400 | TITUS ST | 34.2198 | -118.4468 | 04 | Wednesday | VANDALISM | non-violent |
| 6 | Central | 161 | 404 | 700 | S FIGUEROA ST | 34.0483 | -118.2631 | 13 | Thursday | THEFT | non-violent |

Fig. 18- First 5 rows of the features selected for creating the models.

The goal of the feature selection method was to choose relevant features that would aid in the classification of different types of crimes. The chosen features primarily included geographical and temporal data, which are crucial factors in understanding and predicting crimes.

**Geographical Data:**
Geographic features included attributes like “AREA NAME”, “Rpt dist No.”, “Premise Code”, “Block”, “Street”, “Latitude”, “Longitude” were taken in as part of analysis.

**Time Data:**
Temporal features comprised of the day of the week, and hour of the day. These features capture temporal patterns and variations in criminal activities. Understanding the temporal aspects of crime is essential for resource allocation and implementing targeted interventions during specific times when crime rates are highest.

**Sampling:**
Sampling is conducted when the population is too large to include in analysis. It is an important step in the modeling process to address data imbalance and reduce time complexity. Initially, a random sample was selected from the dataset to ensure a manageable size for training the models. This random sample represents a subset of the original data, preserving its overall characteristics.

```
# sample size
sample_size = 50000  # Adjust the sample size as desired

# Randomly sample the data
sampled_df = df.sample(n=sample_size, random_state=42)
sampled_df.reset_index(drop=True, inplace=True)

# shape of sampled dataframe
sampled_df.shape
```

```
(50000, 11)
```

```
X = sampled_df.drop(columns=['Crime','Crime_Type'])
y = sampled_df['Crime']
```

```
y.value_counts()
```

```
THEFT                 16949
BURGLARY               7806
MOTOR VEHICLE THEFT    6815
ASSAULT                6795
VANDALISM              5296
BATTERY                4676
ROBBERY                1663
Name: Crime, dtype: int64
```

Fig. 19- Sampling data

**Imbalanced Data:**

In our existing dataset the target variable is imbalanced. To tackle that problem under sampling was performed. Undersampling involves reducing the number of instances from the majority class to match the number of instances in the minority class(es). By balancing out the class distribution, the models can then be implemented learn from an equal representation of different crime types. As a result, there will be there won't be any sampling bias and the prediction accuracy can be improved. Since undersampling reduces the overall number of instances to be processed by the models it results in faster model training and inference hence, allowing for more efficient implementation in real-world scenarios.

Random sampling and undersampling techniques were combined, and the resulting refined dataset addressed the data imbalance while being a representative of the original smaple. This sampling strategy enables the models to learn from a balanced dataset, enhancing their ability to accurately predict crime types and improving overall model performance.

```
from imblearn.under_sampling import RandomUnderSampler

# Create an instance of RandomUnderSampler
rus = RandomUnderSampler(random_state=42)

# Resample the data using RandomUnderSampler
X_resampled, y_resampled = rus.fit_resample(X, y)
```

```
y_resampled.value_counts()
```

```
ASSAULT                1663
BATTERY                1663
BURGLARY               1663
MOTOR VEHICLE THEFT    1663
ROBBERY                1663
THEFT                  1663
VANDALISM              1663
Name: Crime, dtype: int64
```

Fig. 20- Random Undersampling the data

**Label Encoding:**

The dependent feature “Crime Type” was encoded using label encoding. This transformation assigns unique numerical labels to each crime type, enabling the models to work with categorical data.

```
from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder
label_encoder = LabelEncoder()

df['Crime'] = label_encoder.fit_transform(df['Crime'])
```

```
df.Crime.value_counts()
```

```
0    1663
1    1663
2    1663
3    1663
4    1663
5    1663
6    1663
Name: Crime, dtype: int64
```

Fig. 21- Label Encoding

**One-Hot Encoding:**

Multiple categorical independent variables were transformed using one-hot encoding. This process converts categorical variables into binary vectors, allowing the models to understand and interpret the categorical information during training. The variable such as “AREA NAME”, “Rpt dist No.”, “Premise Code”, “Block”, “Street”, “day of the week”, and “hour of the day” were one hot encoded.

```
# One-hot encoding all columns except for LAT, LON, and Crime
df_encoded = pd.get_dummies(df[['AREA NAME','Rpt Dist No','Block','Street','Premis Cd','hour','day_of_week']],drop_first=True)

# Concatenate the one-hot encoded columns with the original dataframe
new_df = pd.concat([df[['LAT', 'LON', 'Crime']], df_encoded], axis=1)

new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11641 entries, 0 to 11640
Columns: 4380 entries, LAT to day_of_week_Wednesday
dtypes: float64(2), int64(1), uint8(4377)
memory usage: 48.9 MB
```

```
new_df.head()
```

| | LAT | LON | Crime | AREA NAME_Central | AREA NAME_Devonshire | AREA NAME_Foothill | AREA NAME_Harbor | AREA NAME_Hollenbeck | AREA NAME_Hollywood | AREA NAME_Mission | ... | hour_20 | hour_21 | hour_22 | hour_23 | day_of |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 34.2197 | -118.4536 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | |
| 1 | 33.9256 | -118.2674 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| 2 | 34.2027 | -118.3745 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | |
| 3 | 34.0491 | -118.4614 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 1 | |
| 4 | 34.0276 | -118.3351 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 1 | 0 | 0 | 0 | |

5 rows × 4380 columns

Fig. 22- One hot encoding

**Train and Test Split:**

In this step the data is partitioned into training and testing sets. While training set is used to train the classification models, the testing set is utilized to examine their performance and how effectively they generalize to new, untested data.

```
#Split dataset to Training Set & Test Set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=101)

#define a function for train ML model
def train_model(clf, X_train, y_train):
    clf.fit(X_train, y_train)

#define a function for predict ML model
def predict_data(clf, features, target):
    y_pred = clf.predict(features)

    acc = accuracy_score(target,y_pred)
    f1 = f1_score(target, y_pred, average='micro')
    con_matrix = confusion_matrix(target,y_pred)
    class_report = classification_report(target,y_pred)
    return f1, acc, con_matrix, class_report

#define a function for create ML model
def model(clf, X_train, y_train, X_test, y_test):
    train_model(clf, X_train, y_train)

    f1, acc, con_matrix, class_report = predict_data(clf, X_test, y_test)
    print("Test Data:")
    print("-" * 20)
    print("F1 Score:{}".format(f1))
    print("Accuracy:{}".format(acc))
    print("Confusion Matrix")
    print(con_matrix)
    print("Classfication Report")
    print(class_report)
    print("-" * 20)
```

Fig. 23- Splitting intro training and testing sets

# Building Models

Multiple models were built for prediction. The models used are listed below:

- Logistic Regression: Logistic regression can be used to predict the probability(s) of a binomial or multiple mutually exclusive outcomes.

- Decision Tree Classifier: It is a classification model that can be used in both multinomial and binomial modelling to predict the category to which the output will belong to.

- K Nearest Neighbors Classifier: It can be used in both binomial and multinomial modelling. In binomial models, it can be used to get the outcome by finding the nearest neighbors for a data point and finding the majority class among those neighbors. In multinomial models, it follows the same approach as binomial but instead of 2 there will be greater number of neighbors.

- Random Forest Classifier: It is used to predict outcomes by building a large number of decision tress (binary or multiple) and combining the prediction of all trees.

- Gaussian Naïve Bayes: It follows the assumption the input features all follow a gaussian distribution. In binomial modelling it assumes that the features are independent, and the output is predicted by estimating the probability of each class in given features. In multinomial the same methodology is followed but input classes are multiple instead of binary.

- Ada Boost Classifier: It is a method used in both multinomial and binomial modelling which combines multiple weak classifiers to create a strong classifier.

- Multi Layer Perceptron Classifier: This method predicts binary/multiple outcomes by training a network with interconnected nodes, learning patterns through backpropagation, with the final prediction being based on the output of the last layer.

- Quadratic Discriminant Analysis: This classification algorithm that assumes multiple covariance matrices for each class and fits quadratic decision boundaries to separate those classes. It estimates the parameters for each individual class and that parameter is then used to calculate the likeliness of a new data point belonging to those classes. The class with the greatest likelihood designated the predicted class.

- Gradient Boosting Classifier: This method also combines multiple weak classifiers (usually decision trees) to create a strong classifier.

- XGBoost(Extreme Gradient) Classifier: This method sequentially trains an ensemble of weak learners, then updates the model's prediction based on errors made by the previous learners and finally aggregates the predictions of weak learners to make the final prediction.

- CatBoost Classifier: It's working is similar to XGBoost with some of their difference being in their handling of categorical variables, training speed etc.

- Light Gradient Boosting MClassifier: Its working methodology is same as XGBoost but it comes with an increased efficiency and speed

## THREE DIFFERENT CASES

The modelling was done in 3 multiple cases. They are:

- Multinomial Modelling- The dependent variables are "ASSAULT","BATTERY", "BURGLARY", "MOTOR VEHICLE THEFT" , "ROBBERY", "THEFT", "VANDALISM".
- Binomial Modelling- The dependent variables are either "Violent" or "Non-Violent."
- Only Violent Crimes (Multinomial)- The dependent variables are "ASSAULT", "BATTERY", "ROBBERY".

# RESULTS

## Multinomial

| Models | Accuracy |
|---|---|
| CatBoost Classifier | 35.8% |
| XGBoost Classifier | 35.3% |
| Gradient Boosting Classifier | 35.1% |
| AdaBoost Classifier | 34.6% |
| LightGBM Classifier | 33.9% |
| Logistic Regression | 33.7% |
| Random Forest Classifier | 32.3% |
| Decision Tree | 28.2% |
| Multilayer Perceptron (MLP) Classifier | 25.8% |
| K Nearest Neighbors (KNN) | 20.6% |
| Gaussian Naive Bayes | 18.5% |
| Quadratic Discriminant Analysis | 18.2% |
| Baseline Model | 17.2% |

Table 1 Results of Multinomial models

In the multiclass prediction classification, the CatBoost Classifier, XGBoost Classifier, and Gradient Boosting Classifier emerged as the top performers, achieving high accuracy levels of 35.8%, 35.3%, and 35.1% respectively. These models demonstrated a better performance in accurately classifying different crime types when compared to other models.

**Binomial**

| Models | Accuracy |
|---|---|
| CatBoost Classifier | 70.4% |
| Random Forest Classifier | 70.2% |
| Multilayer Perceptron (MLP) Classifier | 70.0% |
| XGBoost Classifier | 69.6% |
| LightGBM Classifier | 69.5% |
| Logistic Regression | 67.5% |
| AdaBoost Classifier | 66.8% |
| Gradient Boosting Classifier | 66.7% |
| Decision Tree | 65.5% |
| Gaussian Naive Bayes | 62.6% |
| K Nearest Neighbors (KNN) | 61.7% |
| Quadratic Discriminant Analysis (QDA) | 60.6% |
| Baseline Model | 50% |

Table 2 Results of Binomial models

For the binary classification of violent and non-violent crimes, the CatBoost Classifier showed the highest performance with an accuracy of 70.37%. The XGBoost Classifier, LightGBM Classifier, Random Forest Classifier, and Multilayer Perceptron (MLP) Classifier also performed well, with accuracy levels ranging from 69.46% to 70.17%. These models proved reliable in accurately distinguishing between violent and non-violent crimes.

**Violent only (Multinomial)**

| Models | Accuracy |
|---|---|
| Random Forest Classifier | 59.55% |
| CatBoost Classifier | 59.48% |
| LightGBM Classifier | 59.30% |
| XGBoost Classifier | 59.22% |
| Gradient Boosting Classifier | 58.39% |
| AdaBoost Classifier | 58.03% |
| Logistic Regression | 58.00% |
| Multilayer Perceptron (MLP) Classifier | 57.30% |
| Decision Tree | 53.35% |
| K Nearest Neighbors (KNN) | 52.27% |
| Baseline Model | 33.33% |
| Gaussian Naive Bayes | 21.42% |
| Quadratic Discriminant Analysis (QDA) | 15.78% |

Table 2 Results of Violent only models

In the classification of specific crime types within the violent crime category, the top-performing models were the XGBoost Classifier, CatBoost Classifier, LightGBM Classifier, AdaBoost Classifier, and Random Forest Classifier. These models achieved accuracies ranging from 58.03% to 59.55%. Their performance indicated a reasonable level of accuracy in classifying violent crimes, with potential for further refinement and real-world application.

Overall, the developed crime prediction models demonstrated effectiveness in accurately classifying crime types and distinguishing between violent and non-violent crimes. These models have the potential to aid law enforcement agencies in strategic decision-making and targeted interventions for crime prevention.

# Related Works

1. GitHub - priyanka21sk/LA-Crime-Data-Analysis-using-Tableau –
   This analysis uses LAPD crime data from 2012 to 2015, focusing on crime patterns, bureau-wise crime rates, and juvenile crimes. It employs Tableau for interactive data visualization, providing valuable insights for law enforcement, policymakers, and researchers working on crime analysis and prevention in Los Angeles.

2. 2014 Chicago Crime Data Analysis | PPT-
   This analysis delves into crime patterns across the city of Chicago, focusing on the ten highest and lowest recorded crime types. The study also explores the distribution of crimes based on time of day, months, and locations across different streets. The analysis aims to provide valuable insights into crime trends in Chicago, which can be utilized by law enforcement agencies and policymakers to develop targeted crime prevention strategies.

3. GitHub - srinu844/Group_4: LA crime data analysis-
   This analysis specifically investigates the potential relationship between crimes in Los Angeles and weather conditions, including seasons such as summer, spring, and winter. Additionally, the study identifies the crime types with the highest occurrence rates. By examining these connections, the analysis contributes to a better understanding of how external factors like weather might influence crime patterns in LA.

4. Los Angeles Crime Data Analysis Using Pandas | by Fabio Rodrigues | Analytics Vidhya | Medium-
   This analysis comprehensively explores various aspects of crime in Los Angeles, such as the demographics of victims (age, sex, descent), the most frequently occurring crimes, areas with higher crime occurrences, and commonly used weapons in criminal activities. The analysis, carried out using Pandas, provides valuable insights into the characteristics and distribution of crimes in LA, which can inform law enforcement strategies and resource allocation.

5. Machine Learning in Los Angeles Crime Analysis 2010-2019-
   This analysis leverages LAPD crime data and Census data to identify patterns and visualize crime trends on the map of Los Angeles. By employing machine learning techniques, the study aims to discover hidden correlations and predictive patterns that can assist law enforcement agencies in understanding crime dynamics and implementing proactive measures.

6. Crime Data Analysis and Prediction for city of Los Angeles-
   This work focuses on data analysis and visualization to identify crime hotspots within the city of Los Angeles. By examining crime data, the analysis aims to pinpoint specific geographic areas where criminal activities are concentrated. This information can be crucial for law enforcement to strategically allocate resources and enhance their crime prevention efforts in high-risk regions.

# CONCLUSION

In this project, our primary objective was to develop accurate crime prediction models using various machine-learning algorithms. Our primary focus was on three classifications: multiclass prediction, bi-class prediction (violent and non-violent crimes), and prediction of specific crime types within the category of violent crimes. By analyzing historical crime data and then training various machine learning models on it we sought to provide law enforcement agencies with insights and effective tools to better understand crime patterns, severity and types thus resulting in strategic decision-making and better-targeted interventions for crime prevention.

In the multiclass prediction classification modelling was aimed to predict the different crime types in the dataset. After evaluation of multiple models, the CatBoost Classifier, XGBoost Classifier, and Gradient Boosting Classifier stood out as top performers while demonstrating high accuracy levels of 35.8%, 35.3%, and 35.4% respectively, showcasing their ability to accurately classify instances of different crime types. This prediction provides an idea as to how the crimes take place and any patterns so that stakeholders can effectively allocate resources.

In the bi-class prediction classification, our primary focus shifted to differentiating between violent and non-violent crimes. Multiple models showcased impressive performance in accurately distinguishing between these two major categories. The CatBoost Classifier emerged as the top performer with an accuracy of 70.37%, closely followed by the XGBoost Classifier (69.60%) and the LightGBM Classifier (69.46%). Additionally, the Random Forest Classifier (70.17%) and the Multilayer Perceptron (MLP) Classifier (69.80%) also exhibited reliant accuracy.
Then we shifted our focus solely to Violent crimes, we sought to develop specialized models capable of differentiating between offenses such as “Assault”, “Robbery”, or “Battery”. Among the evaluated models, the XGBoost Classifier, CatBoost Classifier, LightGBM Classifier, AdaBoost Classifier, and Random Forest Classifier emerged as the top performers. These models achieved accuracies ranging from 58.03% to 59.55%, indicating their potential to accurately classifying specific types of violent crimes

As a concluding note, our project aimed to develop crime prediction models to support decision-makers in law enforcement agencies by uncovering crime patterns and facilitating better decision-making. While our project encountered certain shortcomings, it also showcased notable strengths and provided valuable insights for further exploration in this field.

The CatBoost Classifier and XGBoost Classifier, two classification models utilized in our project, demonstrated promising accuracy levels, making them reliable options for crime classification tasks. These models showed potential in assisting law enforcement agencies with crime prevention strategies, resource allocation, and response mechanisms, which can contribute to creating safer communities.

However, it is important to acknowledge the limitations of our project. We encountered challenges in gathering comprehensive and diverse data, which can impact the models' performance and there is scope of improvement, Further research on this topic consisting of alternative approaches will be helpful in utilizing the data and gaining outcomes with increased accuracy scores.

Despite these limitations, our project's strengths lie in providing a foundation for future research and development. We have made progress in understanding crime prediction methodologies and have established the potential of the employed classification models. This groundwork can guide future endeavors to refine the models, incorporate additional features, and explore alternative approaches.

Furthermore, exploring real-time data integration and adapting to the evolving nature of crime can increase the relevance and applicability of the prediction models. By acknowledging both the strengths and shortcomings of our project, we pave the way for continuous improvement and contribute to the advancement of crime prediction methodologies. Through iterative refinement and exploration, we can empower decision-makers in law enforcement agencies to make more informed decisions, ultimately working towards safer communities.

# References


1. Purushu, P (2017) "LA Crime Data Analysis Using Tableau", GitHub repository, https://github.com/priyanka21sk/LA-Crime-Data-Analysis-using-Tableau

2. Li, Y (2015) “2014 Chicago Crime Data Analysis”, Slideshare, https://www.slideshare.net/Yawenli/2014-chicago

3. Vaidya, A, Liou, M, Singh, S, and Hanumansetty, S (2020) “Data Analysis of Los Angeles Crimes Arrest Data and Weather (API)”, GitHub repository, https://github.com/srinu844/Group_4

4. Rodrigues, F (2021) “Los Angeles Crime Data Analysis Using Pandas”, Medium Publication, https://medium.com/analytics-vidhya/los-angeles-crime-data-analysis-using-pandas-a68780d80a83

5. Chen, Y (2021) “Machine Learning in Los Angeles Crime Analysis 2010-2019”, ArcGIS StoryMaps, https://storymaps.arcgis.com/stories/e1a9166fe2b649af8d8757476ade0b6a

6. Parekh, H (2021) “Crime Data Analysis and Prediction for city of Los Angeles”, Slideshare, https://www.slideshare.net/HetaParekh1/crime-data-analysis-and-prediction-for-city-of-los-angeles