Crime Data Analysis of a Los Angeles for year 2023

Areena Syed, Varshith Reddy Bhimireddy, Dinesh chandra Gaddam, Long Huynh

1. Project Background

Los Angeles is sharing a detailed crime dataset for 2023 to promote transparency by addressing public safety in large cities. This dataset contains a significant amount of information that can be used to study recent crime trends, helping in improving community safety measures.In order to better understand criminal behavior in Los Angeles, the project analyzes this crime data. To better tackle public safety concerns, it's essential to delve into patterns and trends more comprehensively.

Various factors, such as location characteristics, socioeconomic conditions, and temporal variables, can influence crime rates. Nevertheless, this dataset provides an opportunity to uncover the specific attributes and drivers behind crime occurrences in Los Angeles.Using this valuable data, police, policymakers, and community members can learn things to help them make better plans and spend money more wisely to make things safer.

This project is based on the idea that using data is crucial for dealing with complex problems like crime. By using the detailed crime dataset from the City of Los Angeles, we can use this data to make better decisions and make positive changes in our communities. The dataset includes a lot of different information, like what happened in each incident, what kind of crime it was, where it happened, and the status of each case. With this rich dataset, we'll use statistical analysis, explore the data in-depth, and use visualization tools to find meaningful insights.

2. About the Data

The dataset, acquired from Data governance, is a comprehensive collection of Los Angeles crime data, offering a substantial amount of information with over 925,000 observations and 28 different features. It is specifically tailored to cover the occurrences throughout the year 2023, providing a detailed snapshot of crime trends and patterns within the city during that time period.

Its goal is to tackle crime-related concerns and promote community well-being. By thoroughly analyzing this data, we aim to extract insights that can guide the development of strategies to improve public safety and create a safer environment for residents.

2.1 SMART Questions

Some of the SMART questions we initially framed were:

1. How accurately can we predict the likelihood of crime being solved in Los Angeles based on the available data features in 2023?
2. What are the key factors influencing crime rates across various neighborhoods or communities, and how have these factors evolved over the recent years?
3. Can we identify emerging spatial and temporal patterns or hotspots for crime categories to inform proactive and targeted interventions?
   - Predicting crime likelihood (Question1)could be addressed using supervised learning models such as decision trees, random forests, and logistic regression.
   - Identifying the main factors influencing crime(Question2), could involve techniques like feature selection, reducing dimensionality, and using ensemble methods like gradient boosting.
   - Detecting spatial and temporal patterns or hotspots(Question3), might require specialized spatial analysis techniques. However, clustering algorithms like k-means could potentially aid in identifying areas with high crime rates.

Now we will begin our analysis.

3. Initial Data Analysis

Glossary of all the columns in the dataset:

- DR_NO: Division of Records has a specific identification number
- Date Rptd: MM/DD/YYYY
- DATE OCC: MM/DD/YYYY
- TOME OCC: In 24 hour military time
- AREA: The LAPD consists of 21 Community Police Stations, which are known as Geographic Areas within the department.These Geographic Areas are sequentially numbered from 1-21.
- 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.
- Rpt Dist No: A four - digit code that represents a sub-area within a Geographic Area.
- Part 1-2:
- Crm Cd: Indicates the Crime Committed.
- Crm Cd Desc:Defines the crime code provided
- Mocodes: Modus Operandi: Activities associated with the suspect in commission of the crime
- Vict Age: Two character Numeric

- Vict Sex: F- Female M- Male X- Unknown
- Vict Descent: Descent Code: A- Othr Asian B-Black C- Chinese D- Cambodian F- Filipino G- Guamanian H- Hispanic/Latin/Mexican I- American Indian/Alaskan Native J- Japanese K - Korean L- Laotian O- Other P- Pacific Inslander S- Samoan U- Hawaiian V- Vietnamese W- White X- Unknown Z- Asian Indian
- Premis Cd: The type of structure, vehicle, or location where crime took place
- Premis Desc: Defines the Premise Code provided
- Weapon Used Cd: The type of weapon used in the crime
- Weapon Desc: Defines the Weapon Used Code provided
- Status: Defines the status code provided
- Status Desc: Indicates the crime committed. Crime code 1 is the primary and most serious one.
- Crm Cd 1: May Contain a code for an additional crime, less serious than code 1
- Crm Cd 2: May Contain a code for an additional crime, less serious than code 1
- Crm Cd 3: May Contain a code for an additional crime, less serious than code 1
- Crm Cd 4: May Contain a code for an additional crime, less serious than code 1
- LOCATION: Street address of Crime incident
- Cross Street: Cross street of rounded Address
- LAT: Latitude
- LON: Longitude

Source: data.gov

This dataset shows records of crimes that happened in Los Angeles since 2020. The information comes from written crime reports, but since they were typed on paper, there might be mistakes in the data. Some locations are marked as (0°, 0°) when the data is missing. Addresses are only given to the nearest hundred blocks to keep people's privacy. The data is as accurate as what's in the database.

3.1 Understanding the Data

The data is first imported from a CSV file and stored as a Pandas DataFrame. Through initial exploration, we delve into key aspects of the dataset. This includes observing the first few rows to get a glimpse of the data's structure, checking the data types and identifying any missing values. Summary statistics are computed to gain a high-level understanding of the numerical features within the dataset. Additionally, we assess the uniqueness of the data to determine the diversity and distinctiveness of its entries.

To further understand the dataset, various visualization techniques are employed. Histograms, count plots, and box plots are utilized to visually represent the distributions, trends, and relationships present in the data. These visualizations provide insights into critical aspects such as the distribution of victim ages, the frequency of crimes across different areas, common crime descriptions, and the temporal trends in crime occurrences over time. By comprehensively exploring these facets of the dataset, we gain valuable insights that serve as a foundation for subsequent analysis and decision-making processes.

3.2 Pre-processing

The following pre-processing steps were performed:

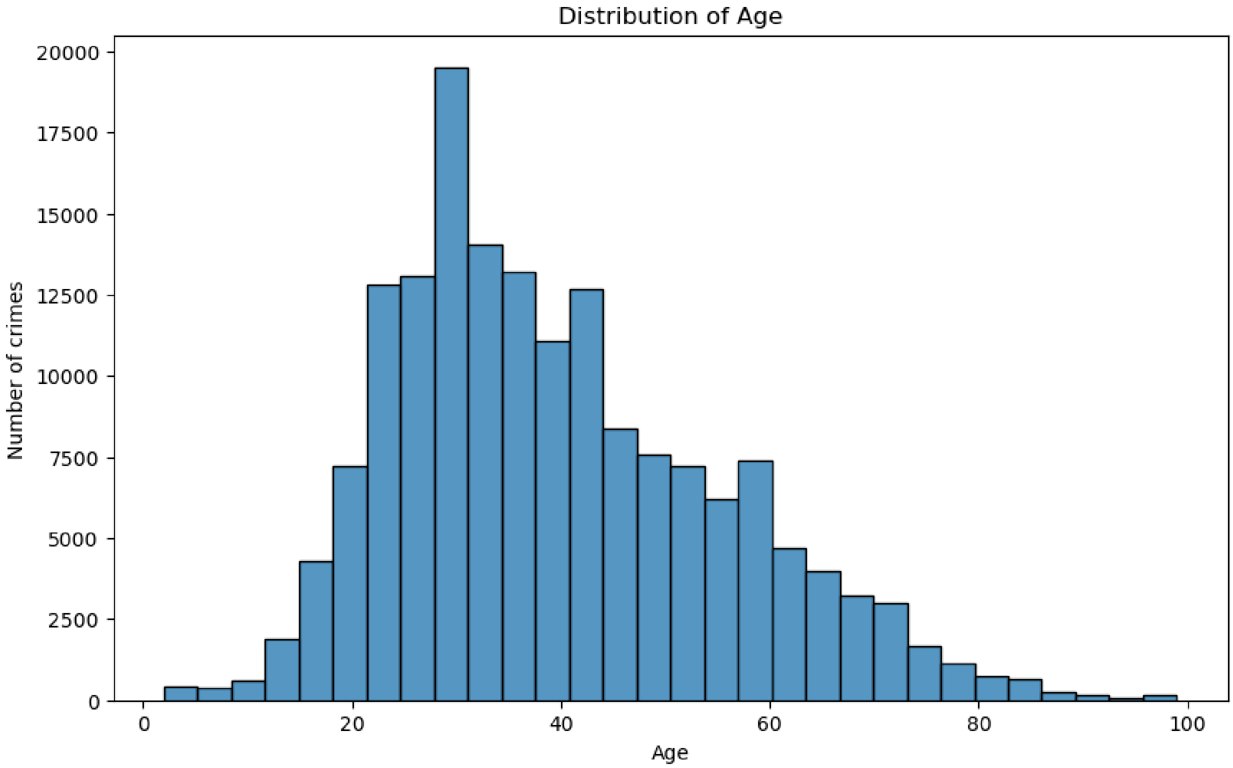
- Column Removal:
  - Unnecessary columns are identified and removed from the dataset to simplify its structure and reduce unnecessary complexity. This helps in focusing on relevant variables for analysis and modeling.
- Filtering Invalid Data:
  - Data integrity is crucial for meaningful analysis. Any invalid or irrelevant data, such as negative age values or other outliers, is filtered out. This ensures that only valid and meaningful data points are retained for analysis, thereby maintaining the integrity of the dataset.
- Handling Missing Values:
  - Missing values can significantly impact the analysis results. These missing values are addressed by either dropping rows containing missing values or filling them with appropriate placeholders like 'Not Reported' or 'Unknown', depending on the context. This step ensures that the dataset is complete and ready for analysis.
- Data Type Conversion:
  - Data types are adjusted to their appropriate formats to facilitate analysis. For example, date columns are converted to datetime format to enable temporal analysis.This helps keep the data organized and makes it easier to do different types of analysis.
- Subset Selection:
  - Sometimes, it's necessary to focus on specific time periods or relevant subsets of data for analysis.In these situations, we take out only the part of the data that we need. For instance, in the given scenario, data for the year 2023 is extracted for further analysis. Subset selection allows for a more targeted analysis of specific trends or patterns within the dataset.

These preprocessing steps collectively ensure that the dataset is cleaned, consistent, and optimized for subsequent exploratory analysis and modeling tasks, thereby enhancing the reliability and validity of the analytical findings.

3.3 Exploratory Data Analysis

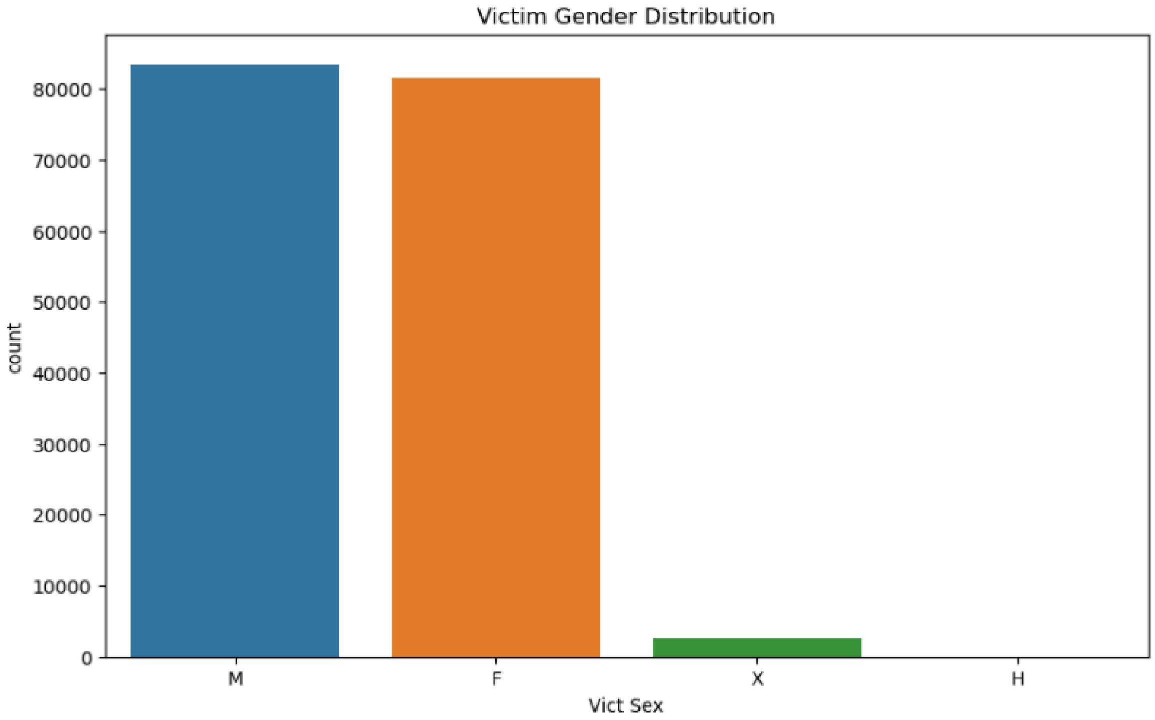
We’re going through the EDA process because it gives us a clear understanding of the crime scene in Los Angeles for 2023. It’s like understanding the big picture before we decide what to do next. By noticing major patterns, we can provide recommendations to help make the city safe.

We started our analysis by looking at the ages of crime victims, which is critical for understanding which age groups are most affected by crime.



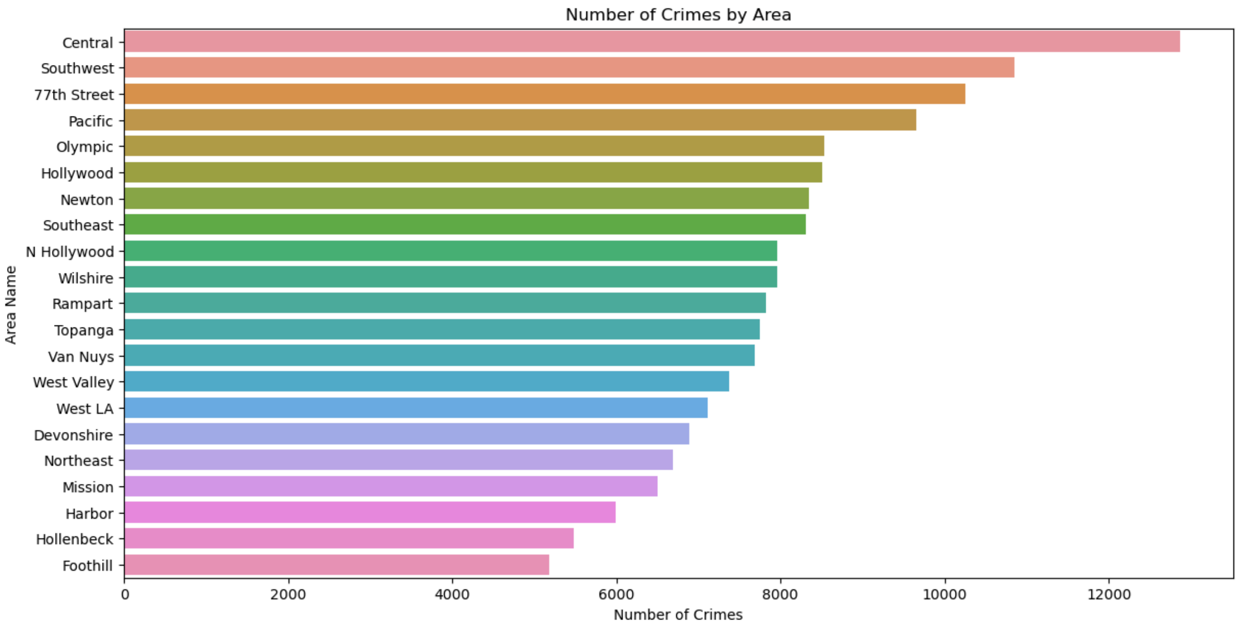
From the above plot, we observed that distribution was skewed towards younger adults, with the highest number of crimes occurring in an age group that looks to be in the mid-20s to early 30s. The frequency of crimes involving victims decreases as the age increases past this peak.

Next, we plotted the relationship between Vict Sex and count:



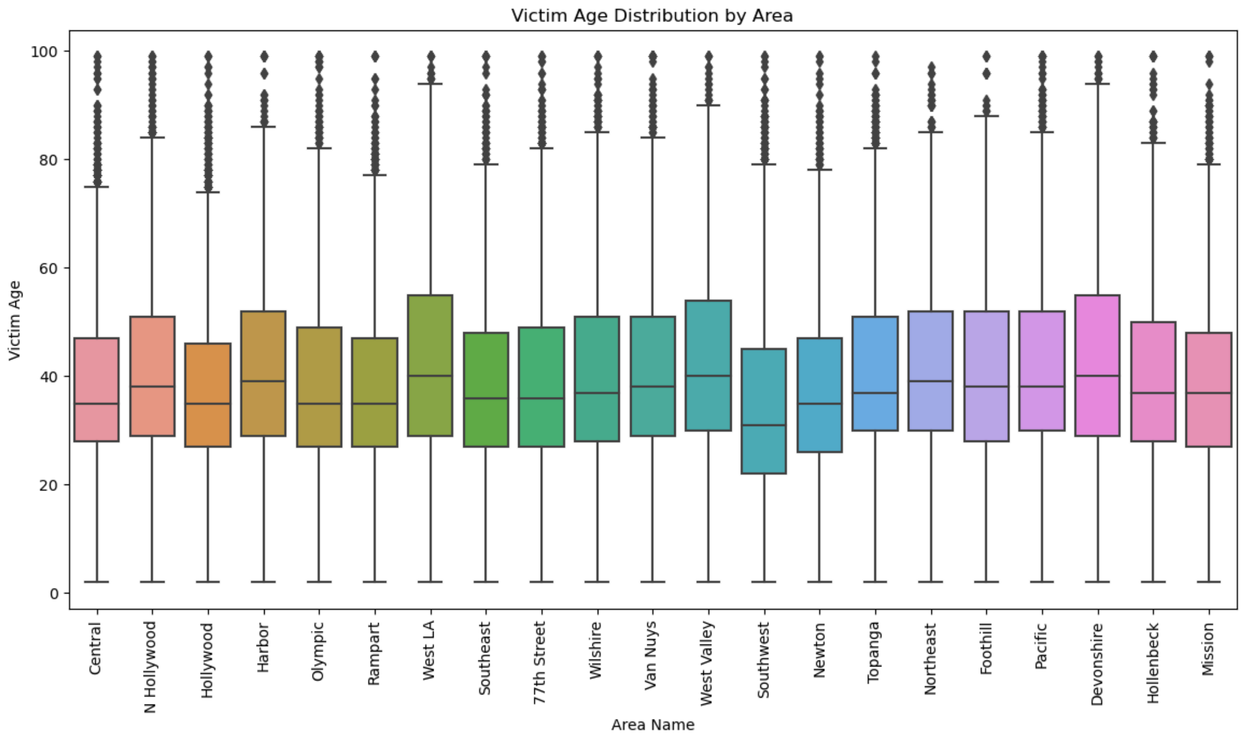
Looking at this plot, we see the distribution of crime victims by gender. Males and females are affected by crime at similar rates.

Next, we plotted a bar chart of crime reported in various areas of Los Angeles during the year 2023.



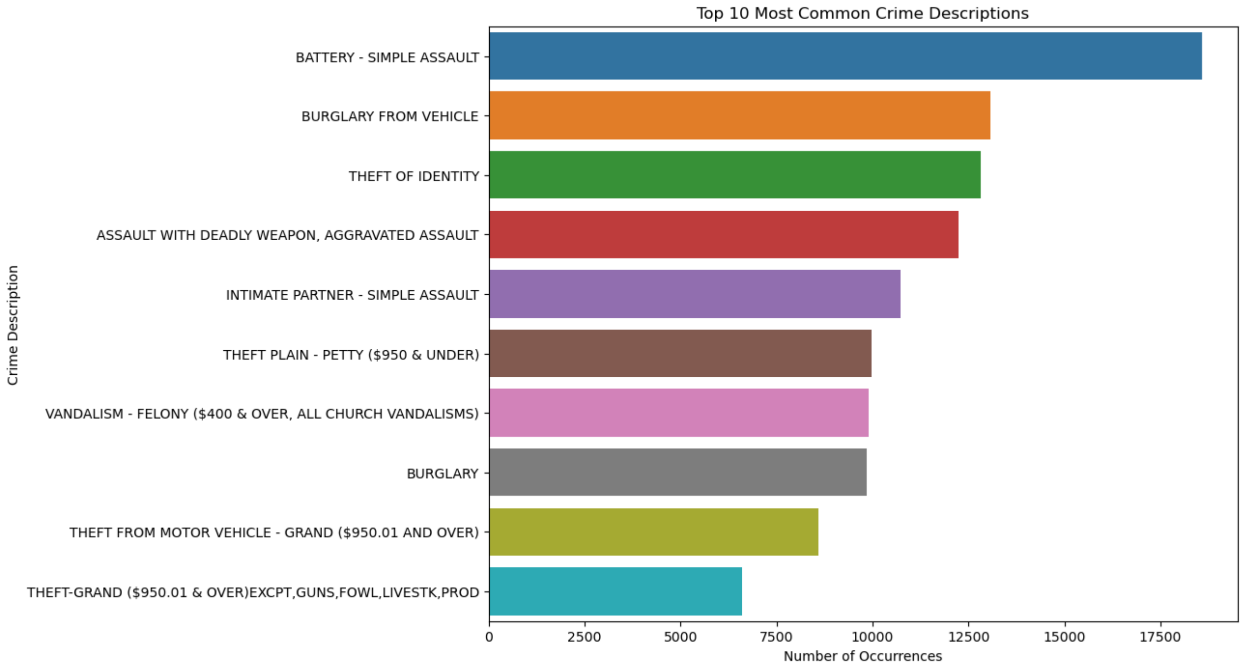
From the graph, we can clearly show which areas are more affected by crime, with the Central region recording the highest number of crimes, suggesting a significant demand for law enforcement resources there. In contrast, areas like Foothill, Hollenbeck, and Harbor report low incidents pointing them as lower-risk areas.

Next, we examine the age distribution of crime victims across various neighborhoods in Los Angeles. The box plot below provides a visual summary of these distributions, highlighting the median age of victims and the variability in ages within each area.



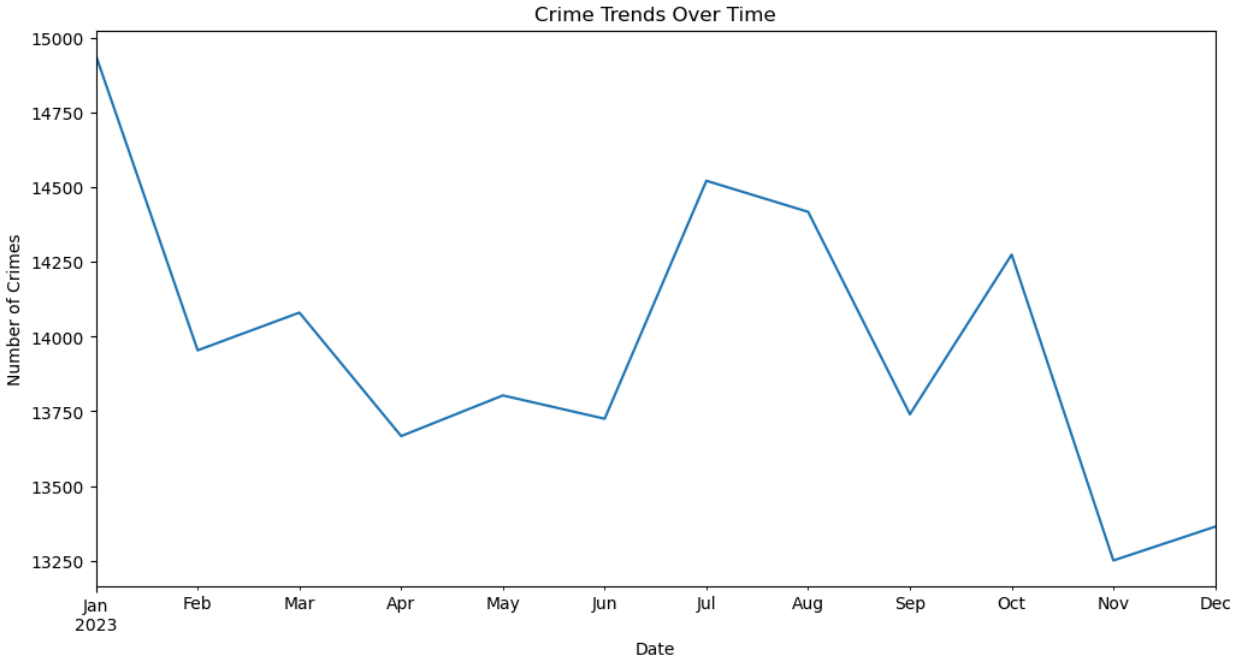
The box plot reveals that the ages of crime victims change across different parts of Los Angeles, but the most common victim age stays about the same. Places like central and Hollywood, in particular, have a broader range of victim ages, suggesting they have a mix of factors that influence different age groups.

We now explore the common crimes in Los Angeles to understand what types of illegal activities happen most often. The bar chart presents an overview of the top ten most commonly reported crimes, providing an understanding of the most prevalent issues faced by the community.



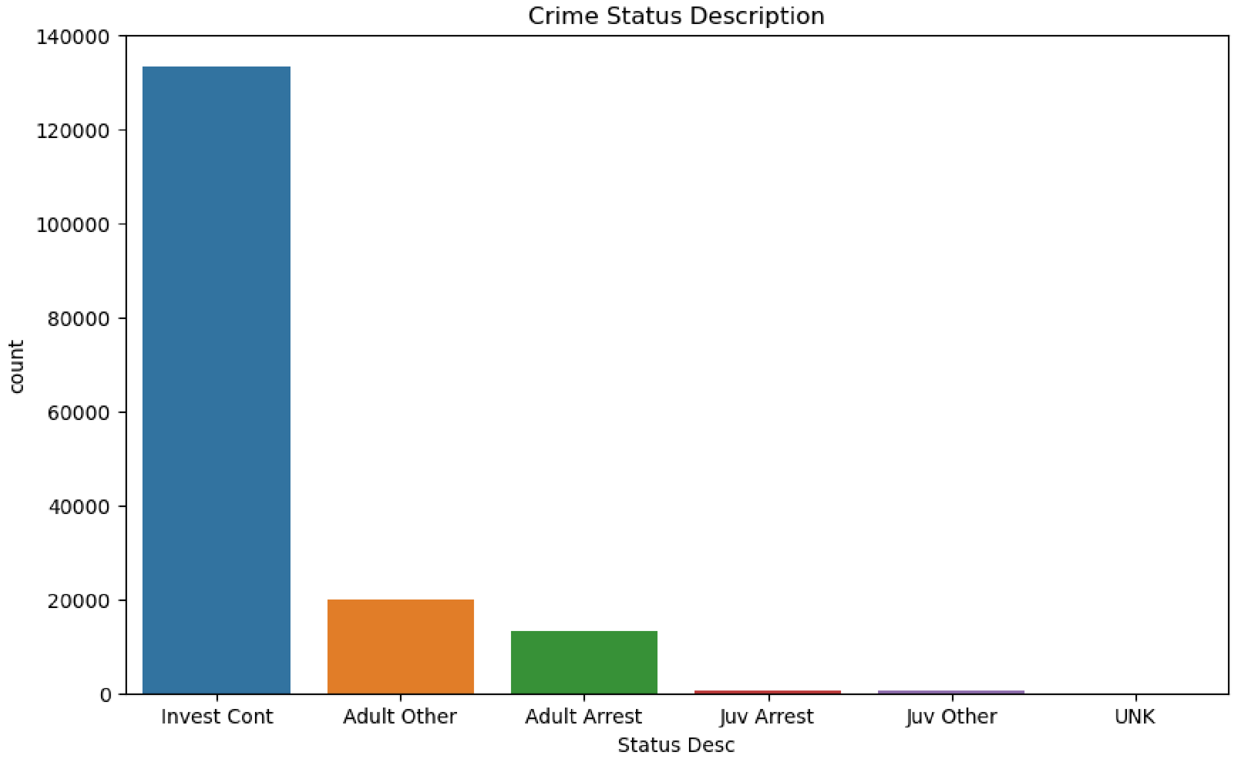
This chart shows that ‘Battery - Simple Assault’ is the crime most often reported, with ‘Burglary from Vehicle’ and ‘Theft of Identity’ also being highly common.

We’ll look at the monthly changes in crime rates across Los Angeles for the year 2023 with a line chart next.



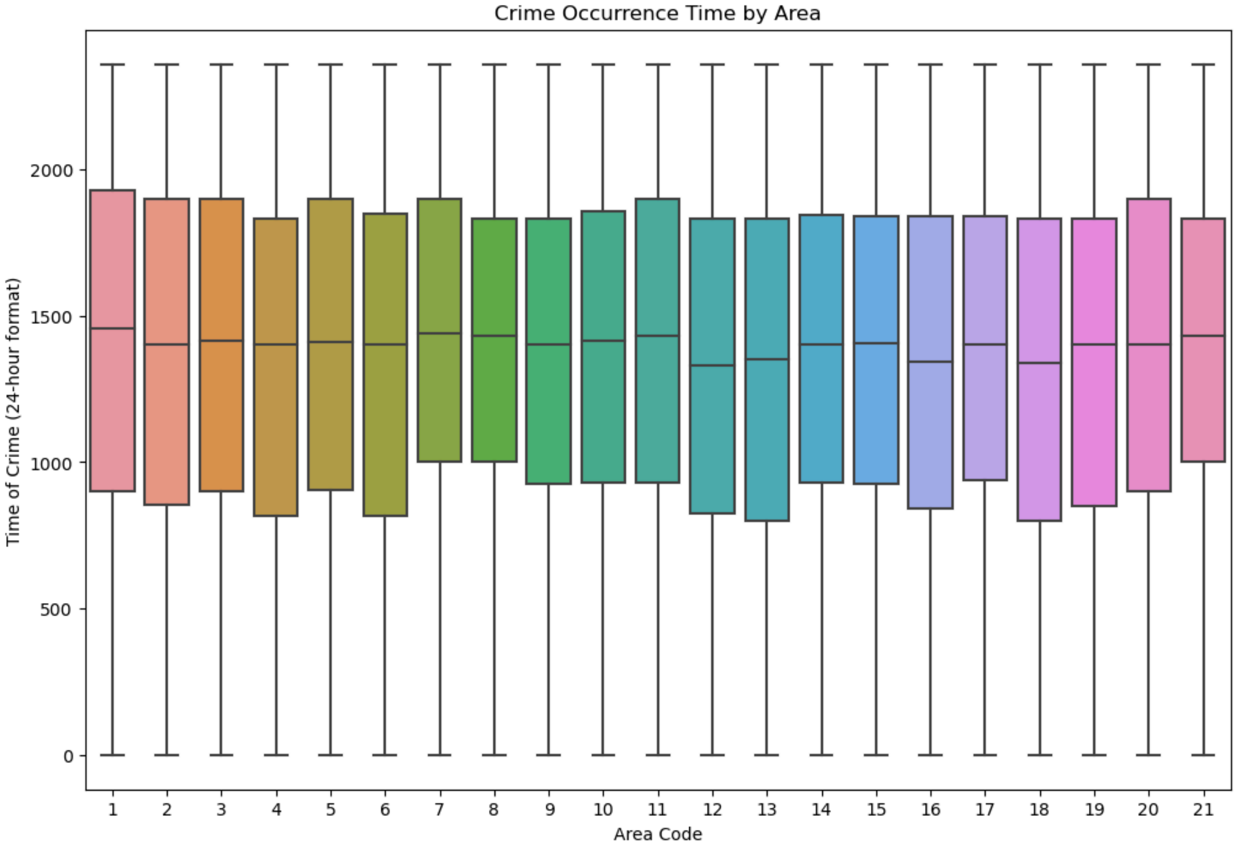
This line graph illustrates the fluctuations of crime throughout the year. There is an increase in crimes in July and October, while February has the lowest. These fluctuations might be linked to seasonal events, local policies, or other factors that could be influencing crime rates.

We will now examine a bar chart that categorizes crimes into various groups based on their status like ongoing investigations, arrests made, or cases that are unresolved.



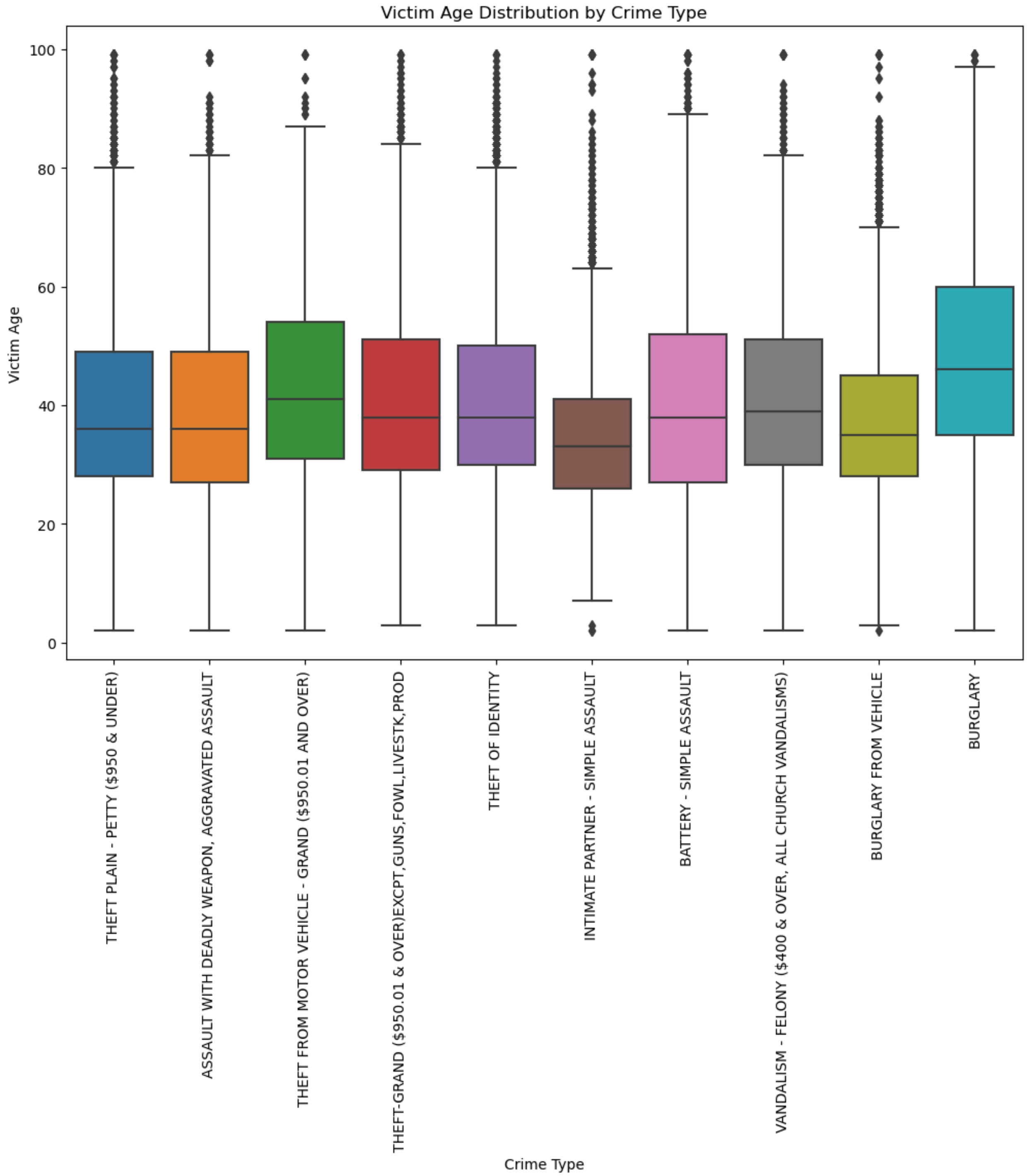
The chart indicates that most crimes are still under investigation, as indicated by the ‘Invest Cont bar’. There are a smaller number of cases that have resulted in an arrest, with adults being arrested more frequently than juveniles. A very small percentage of the cases are marked as ‘UNK’, their outcomes are not clear.

After looking into the types of crimes that happen most, we’re now going to look at when these crimes typically take place. The next box plot will show us what times of day crimes happen in different parts of Los Angeles.



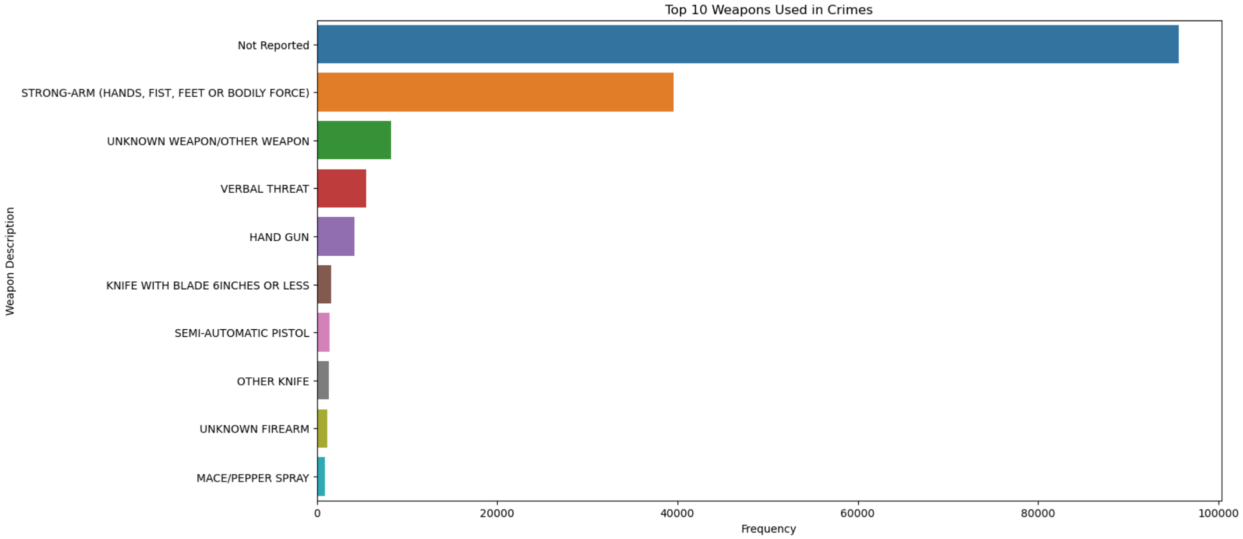
The plot illustrates the timing of crimes across different areas codes with consistency. This consistency in crime times could suggest similar operational factors or social behaviors across the regions.

In order to gain a better understanding of the connection between victim characteristics and the type of crime committed, we will analyze a box plot that displays how victim ages are distributed across different crime categories.



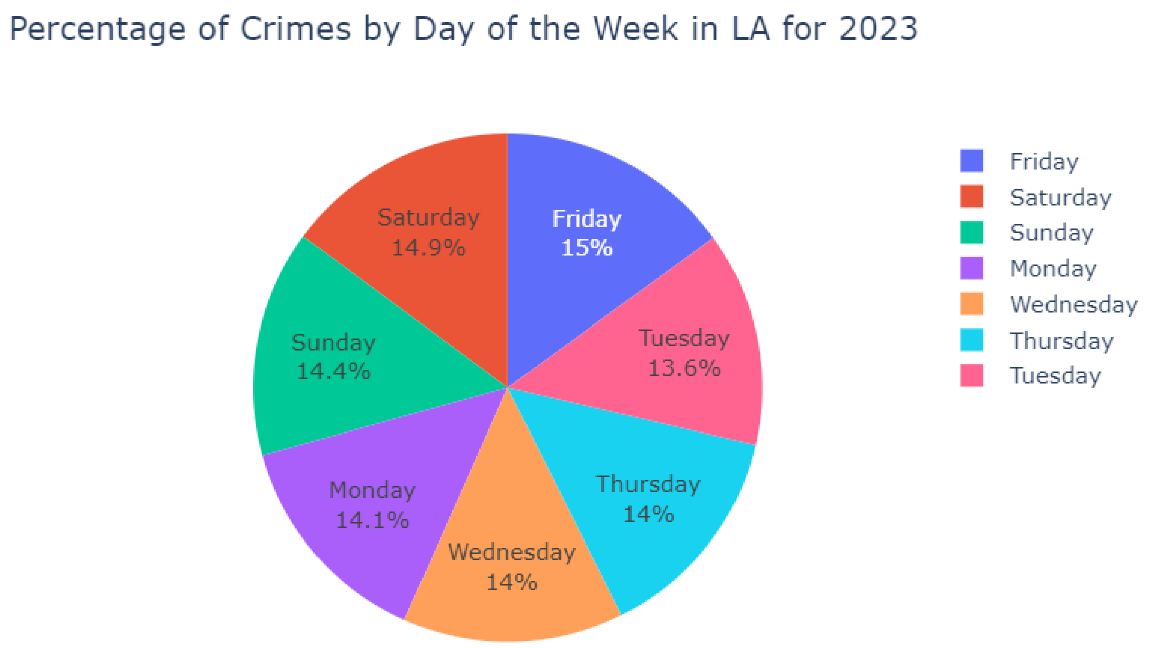
The box plot shows us the variation in victims' ages for different crimes. Some crimes, like theft and battery show a wide range of victim ages, while others like burglary from a vehicle, tend to have victims in a more specific age group. These patterns might help in creating targeted crime prevention programs.

In our continued analysis, we next focus on the tools of crime by examining the types of weapons reported in criminal incidents.



The bar chart shows that 'Not Reported' leads by a large margin, indicating many incidents where the weapon used was not specified. Following this, 'Strong-arm' methods, which include hands, fists, feet, or bodily force, are the most commonly reported, suggesting that many crimes involve personal assault. Less frequently used are guns and knives, though they still represent a significant number of incidents.

Next, we’re looking at when crimes happen during the week. We'll check out a pie chart that shows how many crimes happen on each day in Los Angeles for 2023.



The pie chart shows that crimes are fairly evenly spread throughout the week, but there's a small increase on Fridays. The days of the weekend, Saturday and Sunday, have a bit more crime compared to certain weekdays. This might hint at more crimes happening when people are off work. Understanding these patterns could be really useful for planning police patrols and safety activities.

To wrap it up, this EDA has helped us see the clear patterns in Los Angeles's crime data. Knowing these patterns, we've got what we need to suggest smart ways to make the city safer.

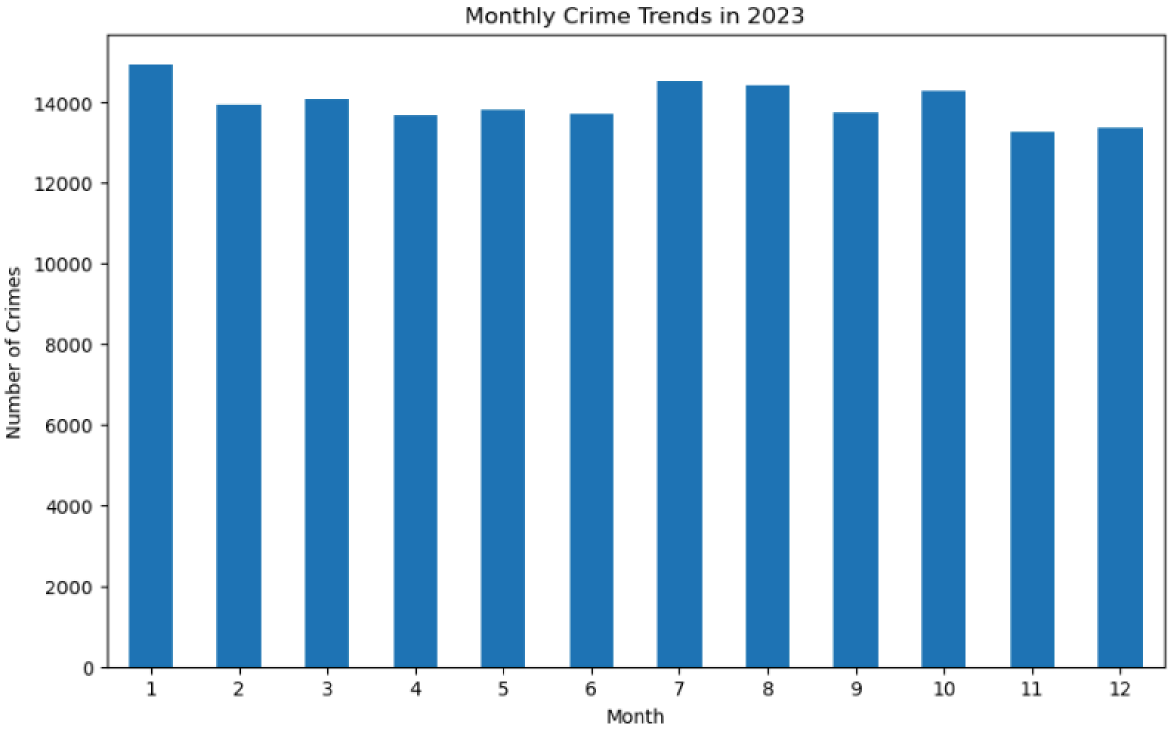
4. Smart questions

Q1: Can we identify emerging spatial and temporal patterns or hotspots for crime categories to inform proactive and targeted interventions?

We conducted an analysis of the data by location and time to detect emerging spatial and temporal patterns or hotpots for different crime categoriesThese hotspots were then studied in terms of crime category, uncovering precise patterns that could be useful for targeted interventions.

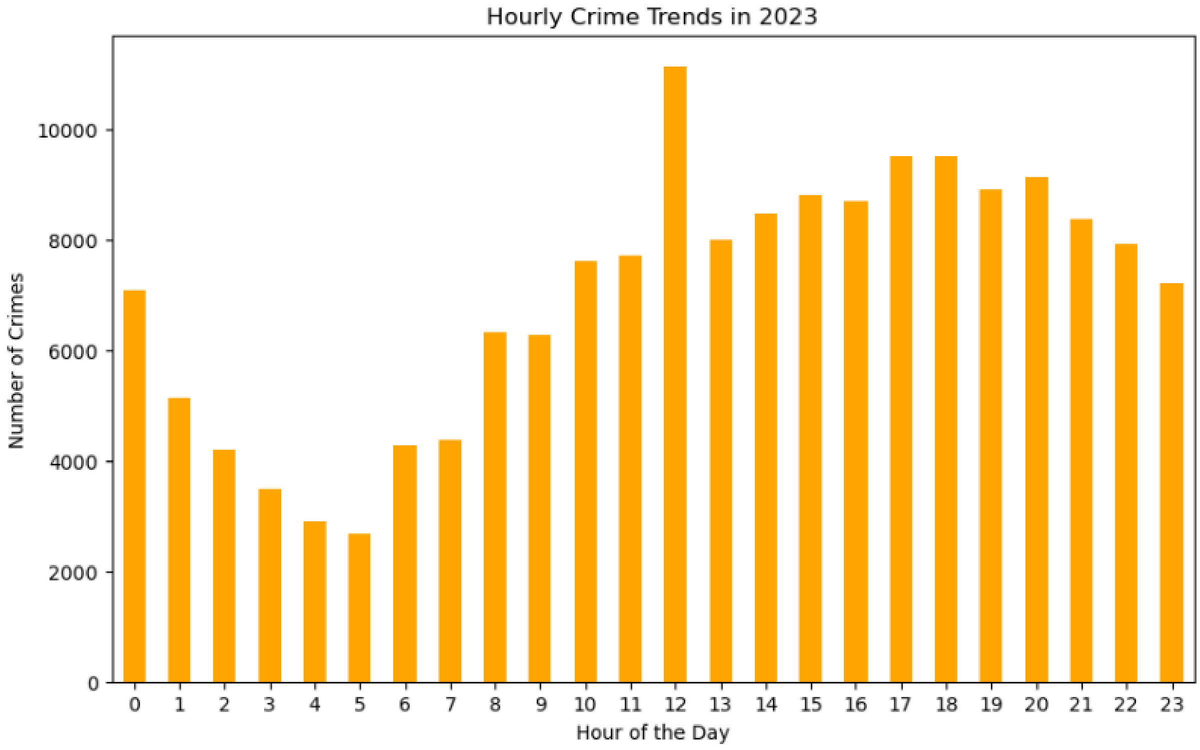
Temporal analysis:

Our analysis began with examining the month-to-month crime frequency in Los Angeles for 2023, aiming to spot any monthly patterns or seasonal effects on crime rates.



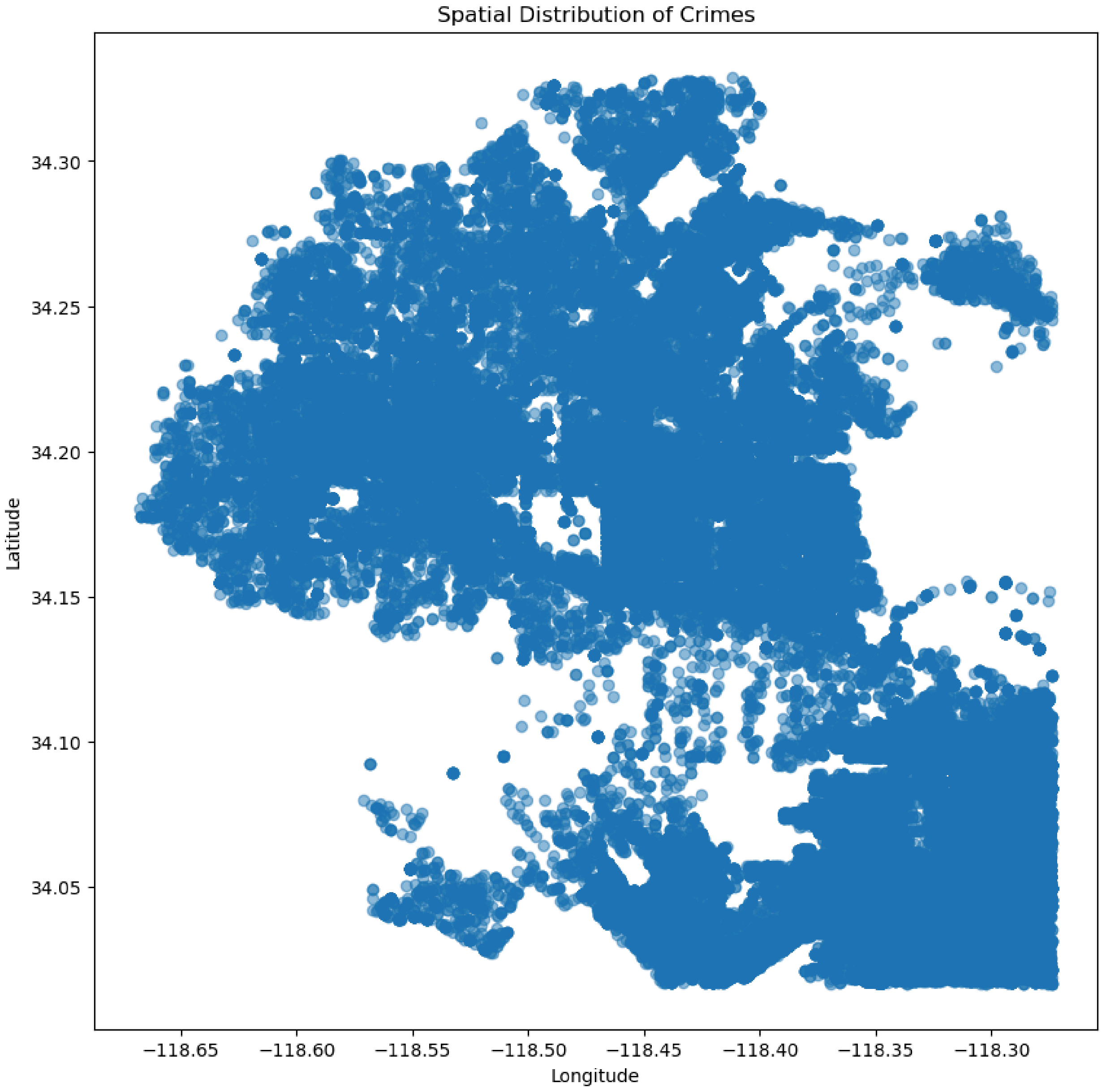
The charts showing each month in 2023 presents a consistent level of crime across all months in Los Angeles. This indicates that police or law enforcement must increase patrols and inform the public about safer times to be out.

Next, we explored the distribution of crime reports by the hour to identify crimes that are more frequent during certain times of the day or night.



Looking at the hourly crime data showed a clear rise in crime around midday. This information is really important because it can help figure out when police patrols should be most active and when people should be more alert.

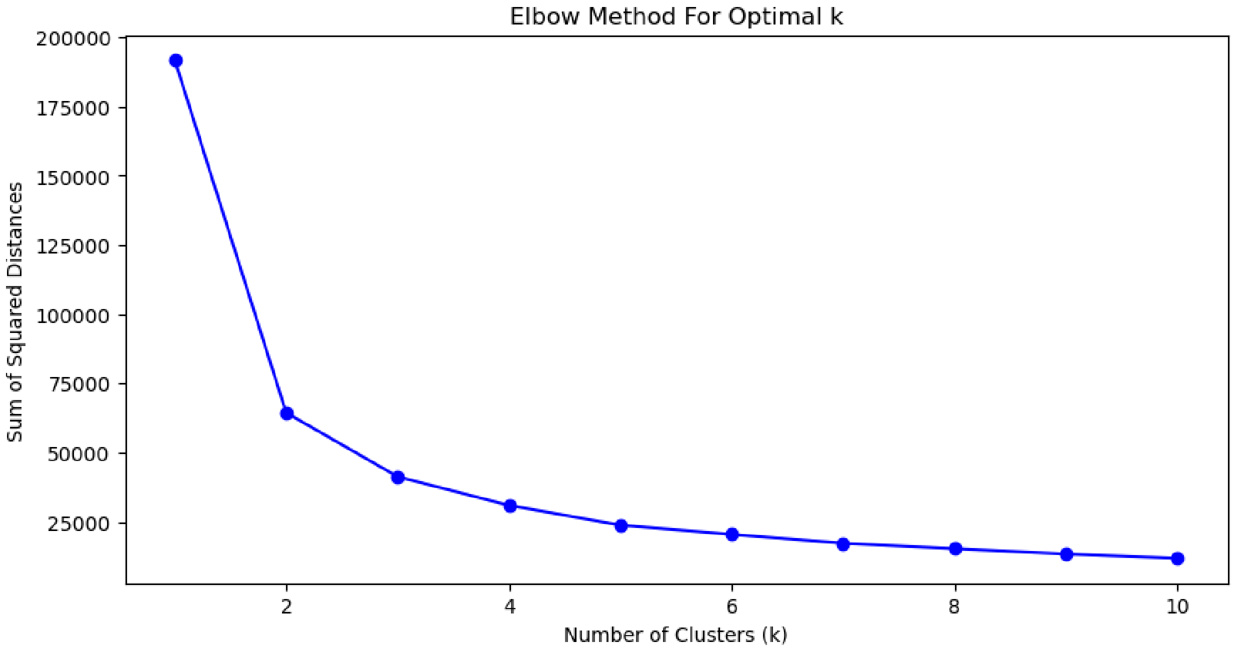
Spatial Analysis:



The scatter plot gives us a clear picture of where crimes are happening in Los Angeles. We see some spots with a lot of dots close together, showing us the hotspots where crime is more common.

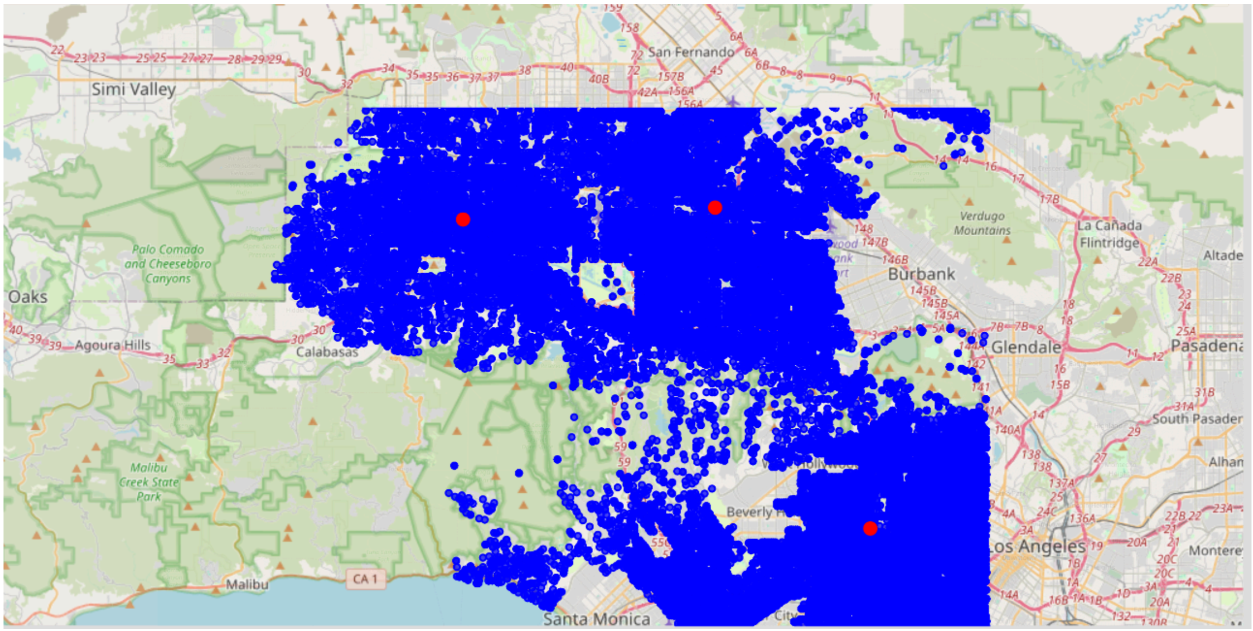
Clustering of Hotspots:

We used a method called K-Means clustering to find out where the most crime happens in Los Angeles. The Elbow method was used to determine the best number of crime hotspots to look at.



From the Elbow graph, the number of clusters against the sum of squared distances, shows a clear leveling at k = 3.

Consider k= 3 and applying K-Mean clustering(taking latitude and longitude)



From the map illustrate clusters of crime in Los Angeles, blue dots indicate that individual crimes and red dots indicate the central points of the highest crime activity.

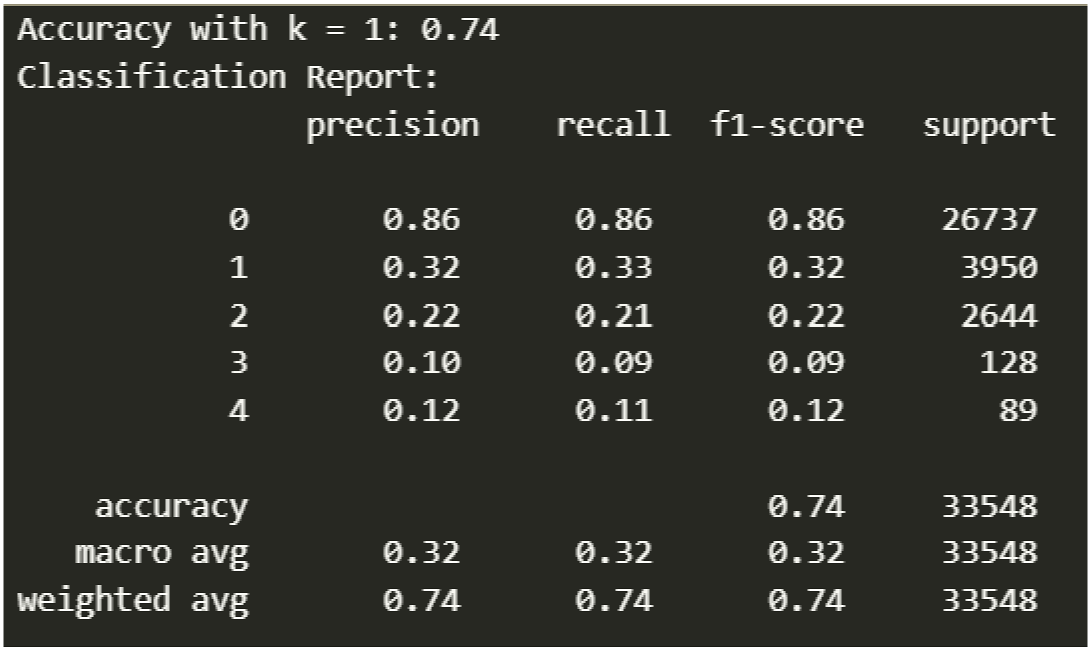
The identifications of these hotspots through clustering is helpful to law enforcement and city planners. It allows increased police patrols or community involvement activities in areas that are most affected by crime,by doing this we can stop crimes before it happens . This targeted approach is more efficient and has a higher impact on public safety.

Q2: How accurately can we predict the likelihood of crime being solved in Los Angeles based on the available data features in 2023?

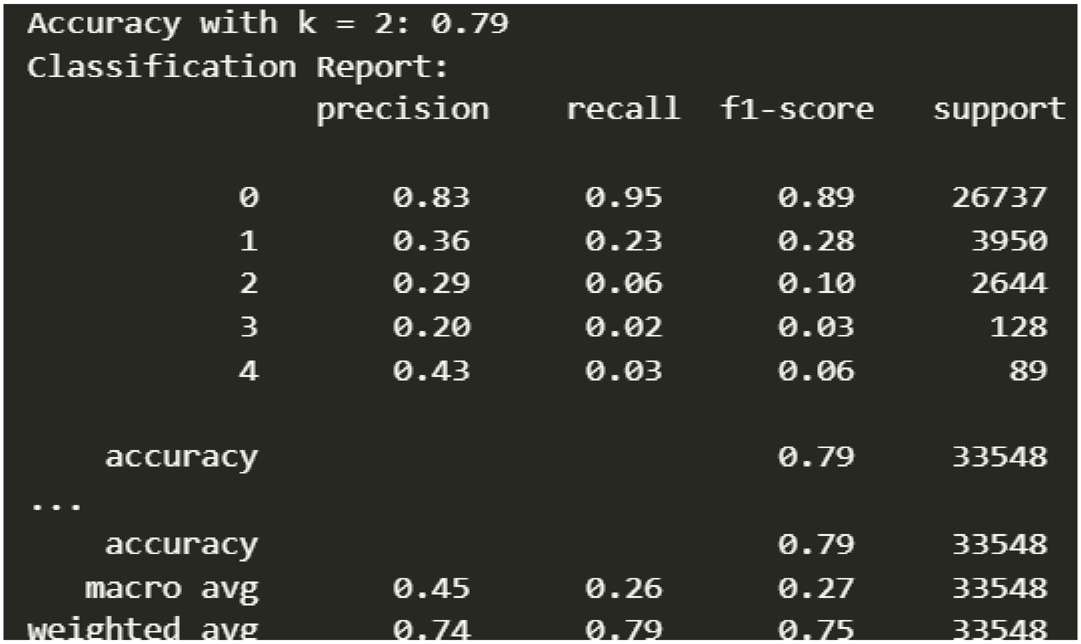
Predicting the likelihood of crime being solved in a city like Los Angeles involves a complex interplay of various factors, including but not limited to the type of crime, location, time of day, socioeconomic conditions, law enforcement resources, and technological advancements.The features that affect chances of crime being solved are Area, Vict age, crime type(crm cd), Weapon Used, Month.The Dependent variable is “Status” (tells if crime is solved). I have used the following algorithms to answer the question.

KNN ALGORITHM

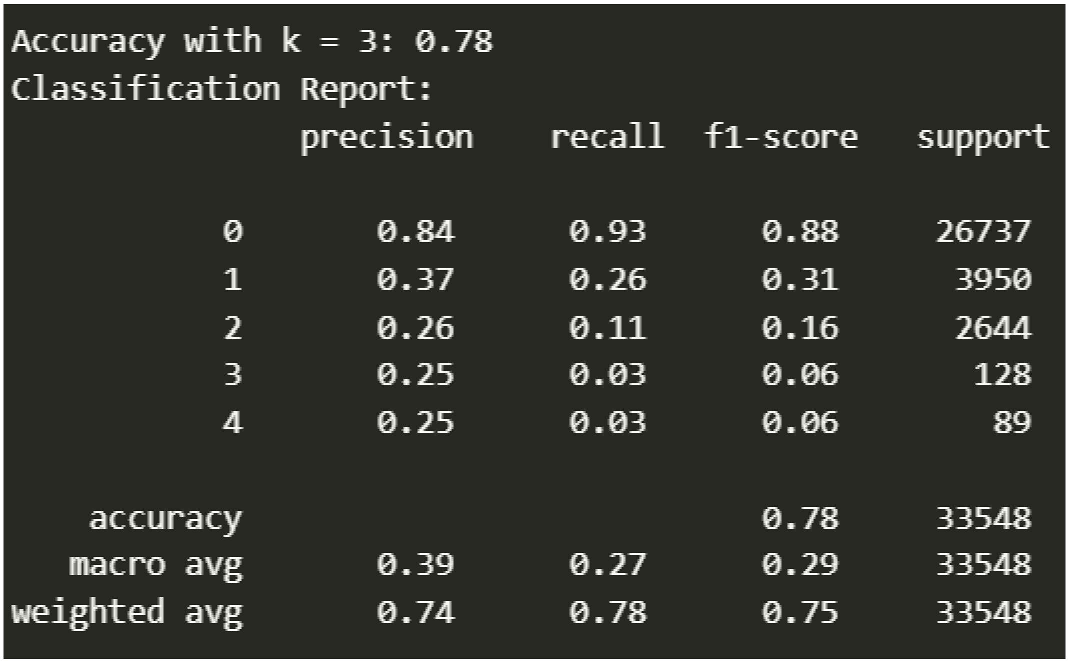
The k-nearest neighbor’s method, generally known as KNN or k-NN, is a non-parametric, supervised learning classifier that utilizes proximity to classify or predict the grouping of a single data point.



```
Accuracy with k = 1: 0.74
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.86      0.86     26737
           1       0.32      0.33      0.32      3950
           2       0.22      0.21      0.22      2644
           3       0.10      0.09      0.09       128
           4       0.12      0.11      0.12        89

    accuracy                           0.74     33548
   macro avg       0.32      0.32      0.32     33548
weighted avg       0.74      0.74      0.74     33548
```

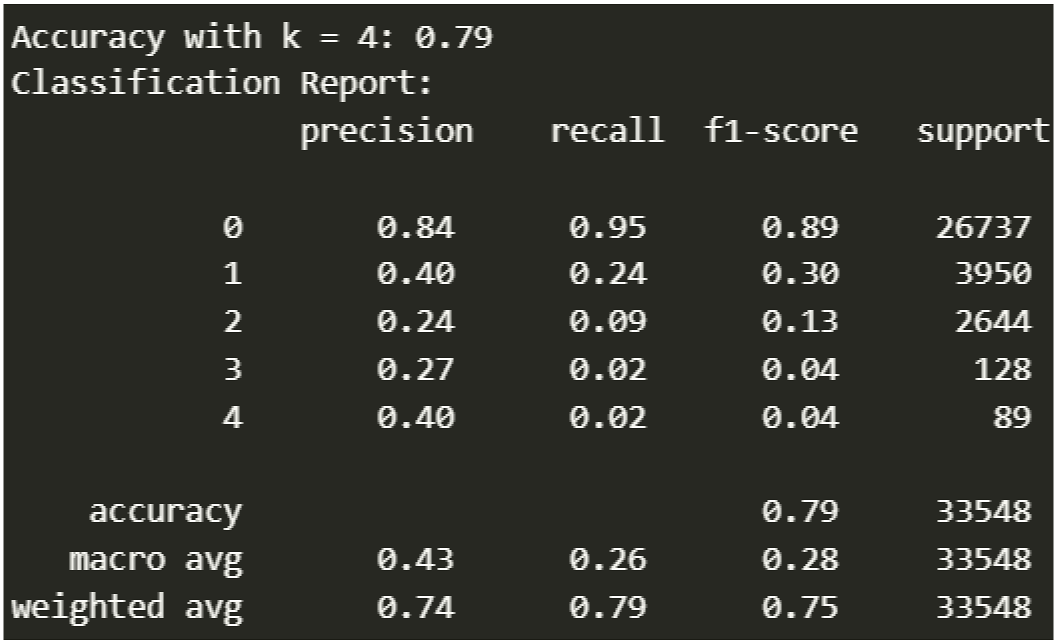


```
Accuracy with k = 2: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.95      0.89     26737
           1       0.36      0.23      0.28      3950
           2       0.29      0.06      0.10      2644
           3       0.20      0.02      0.03       128
           4       0.43      0.03      0.06        89

    accuracy                           0.79     33548
...
    accuracy                           0.79     33548
   macro avg       0.45      0.26      0.27     33548
weighted avg       0.74      0.79      0.75     33548
```

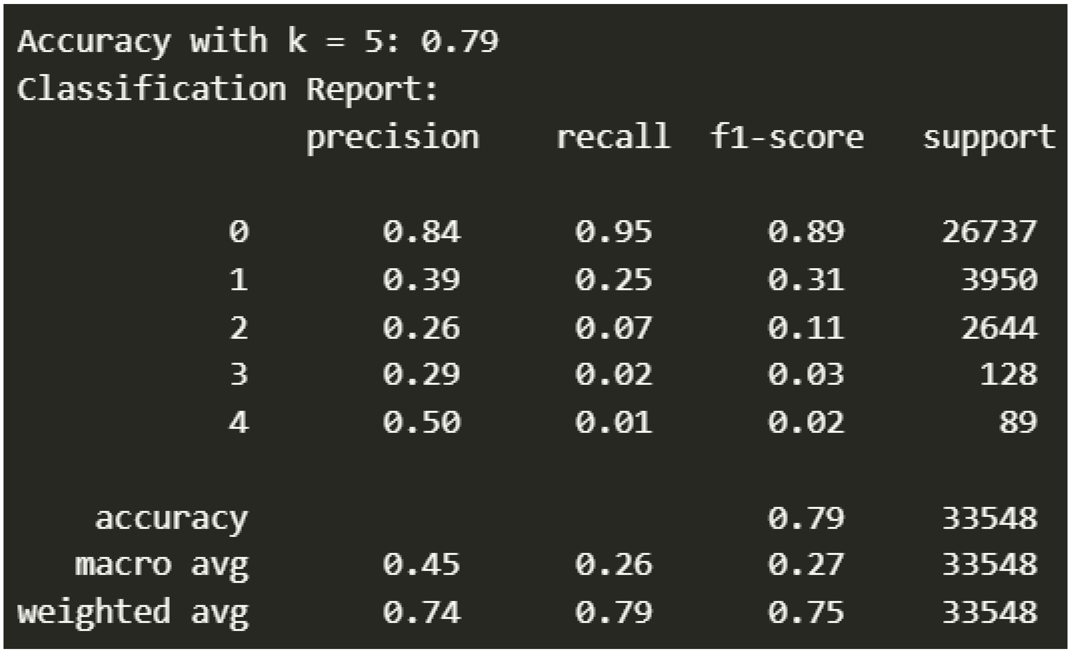


```
Accuracy with k = 3: 0.78
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.93      0.88     26737
           1       0.37      0.26      0.31      3950
           2       0.26      0.11      0.16      2644
           3       0.25      0.03      0.06       128
           4       0.25      0.03      0.06        89

    accuracy                           0.78     33548
   macro avg       0.39      0.27      0.29     33548
weighted avg       0.74      0.78      0.75     33548
```

I have applied KNN with different values of k (number of neighbors) and found that k=2 gave the highest accuracy of 0.79. KNN is simple to implement and works well for datasets with clear decision boundaries but can be sensitive to irrelevant or noisy features.



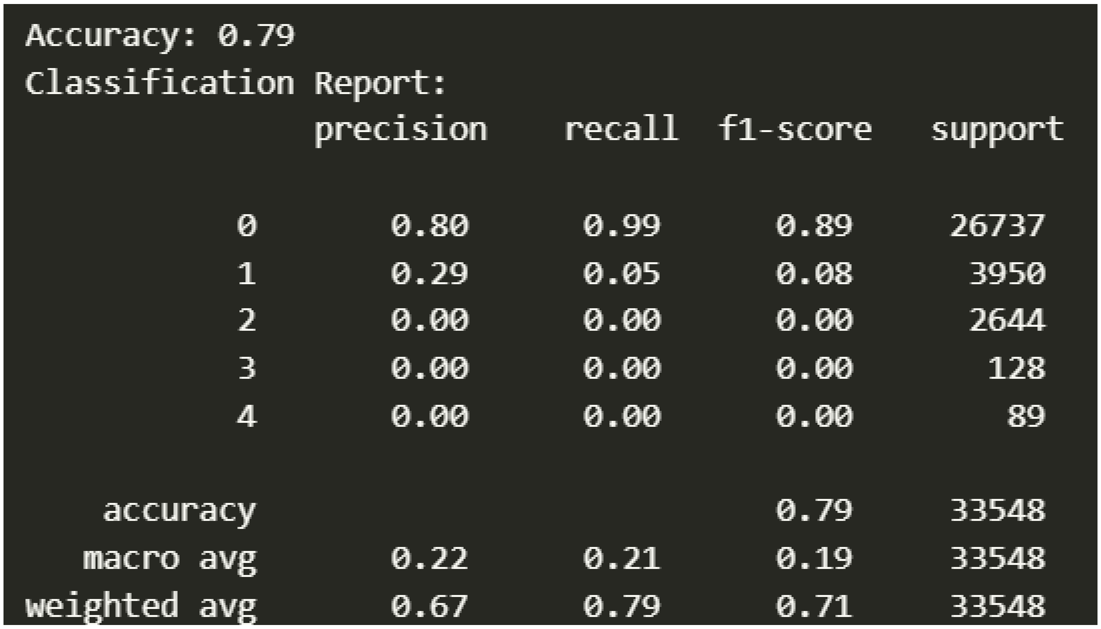
```
Accuracy with k = 4: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.95      0.89     26737
           1       0.40      0.24      0.30      3950
           2       0.24      0.09      0.13      2644
           3       0.27      0.02      0.04       128
           4       0.40      0.02      0.04        89

    accuracy                           0.79     33548
   macro avg       0.43      0.26      0.28     33548
weighted avg       0.74      0.79      0.75     33548
```



```
Accuracy with k = 5: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.95      0.89     26737
           1       0.39      0.25      0.31      3950
           2       0.26      0.07      0.11      2644
           3       0.29      0.02      0.03       128
           4       0.50      0.01      0.02        89

    accuracy                           0.79     33548
   macro avg       0.45      0.26      0.27     33548
weighted avg       0.74      0.79      0.75     33548
```

Logistic Regression using the Sklearn library

Logistic regression is a binary classification algorithm that estimates the probability of a binary outcome based on one or more independent variables. In the analysis, I used logistic regression to predict the status of reported crimes based on features such as area, victim age, crime code, premises code, weapon used code, and month. After preprocessing and scaling the data, logistic regression model achieved an accuracy of 0.79, indicating that it correctly predicted the crime status nearly 80% of the time. Logistic regression is efficient for binary classification tasks and provides interpretable coefficients for understanding feature importance.

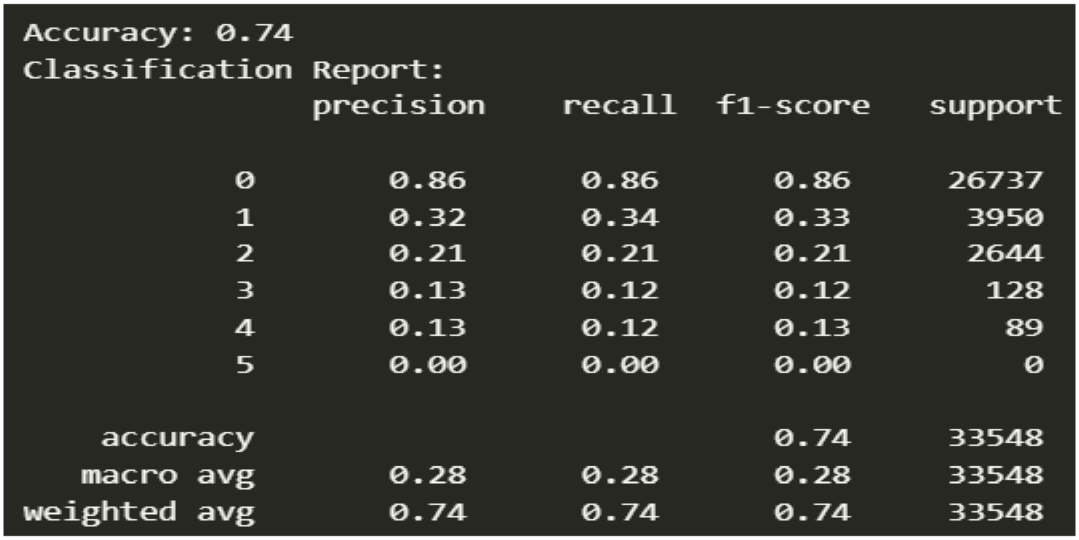


```
Accuracy: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.99      0.89     26737
           1       0.29      0.05      0.08      3950
           2       0.00      0.00      0.00      2644
           3       0.00      0.00      0.00       128
           4       0.00      0.00      0.00        89

    accuracy                           0.79     33548
   macro avg       0.22      0.21      0.19     33548
weighted avg       0.67      0.79      0.71     33548
```

Decision Tree

Decision trees are versatile and can handle both classification and regression tasks. They create a tree-like structure of decisions based on feature values.

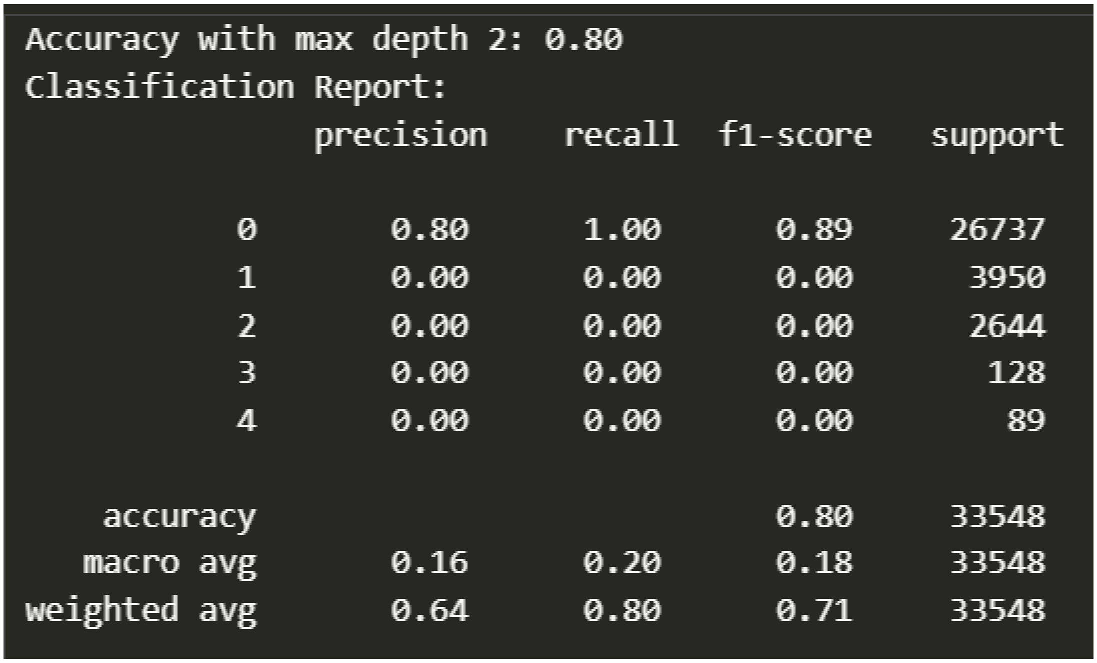
Before tuning:



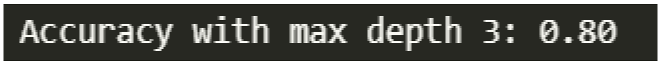
```
Accuracy: 0.74
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.86      0.86     26737
           1       0.32      0.34      0.33      3950
           2       0.21      0.21      0.21      2644
           3       0.13      0.12      0.12       128
           4       0.13      0.12      0.13        89
           5       0.00      0.00      0.00         0

    accuracy                           0.74     33548
   macro avg       0.28      0.28      0.28     33548
weighted avg       0.74      0.74      0.74     33548
```

The decision tree classifier was trained on the same features as logistic regression and achieved an accuracy of 0.74 initially. To improve the decision tree's performance, we performed hyperparameter tuning by varying the maximum depth of the tree.

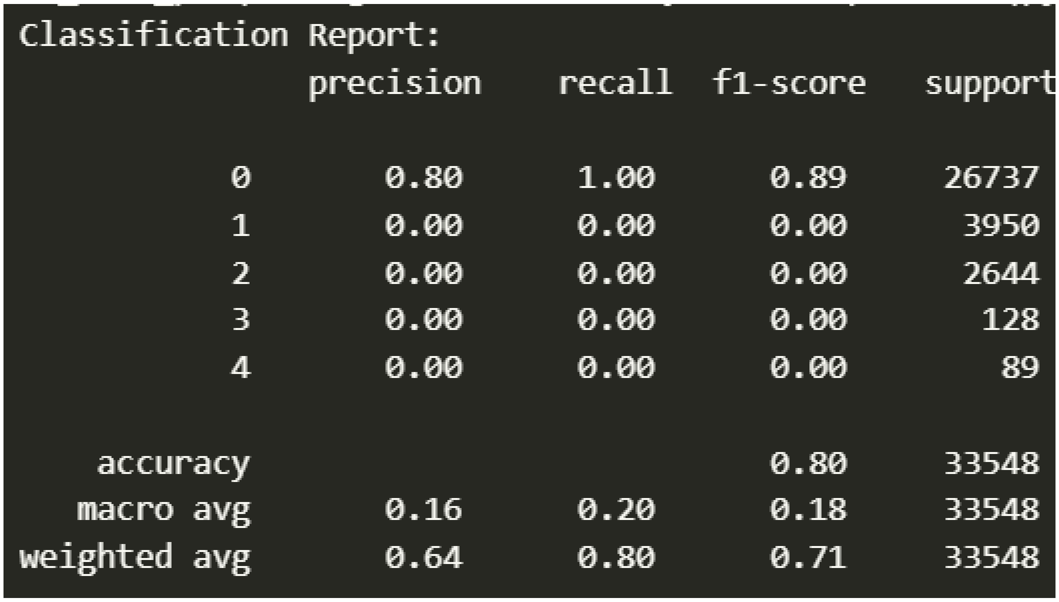
After tuning:



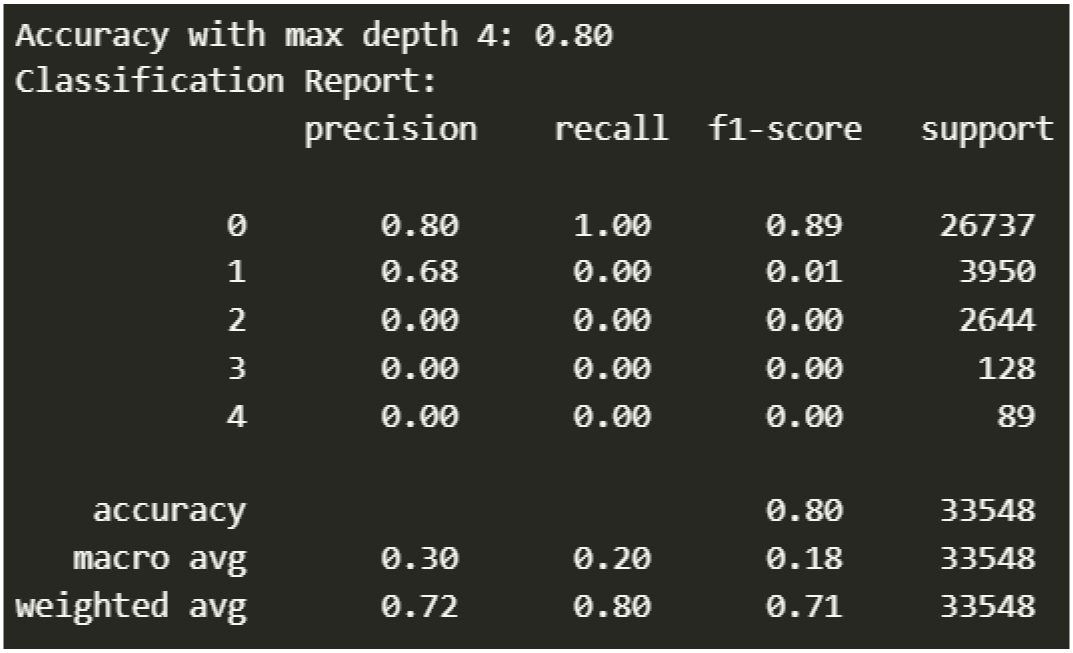
```
Accuracy with max depth 2: 0.80
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     26737
           1       0.00      0.00      0.00      3950
           2       0.00      0.00      0.00      2644
           3       0.00      0.00      0.00       128
           4       0.00      0.00      0.00        89

    accuracy                           0.80     33548
   macro avg       0.16      0.20      0.18     33548
weighted avg       0.64      0.80      0.71     33548
```



```
Accuracy with max depth 3: 0.80
```

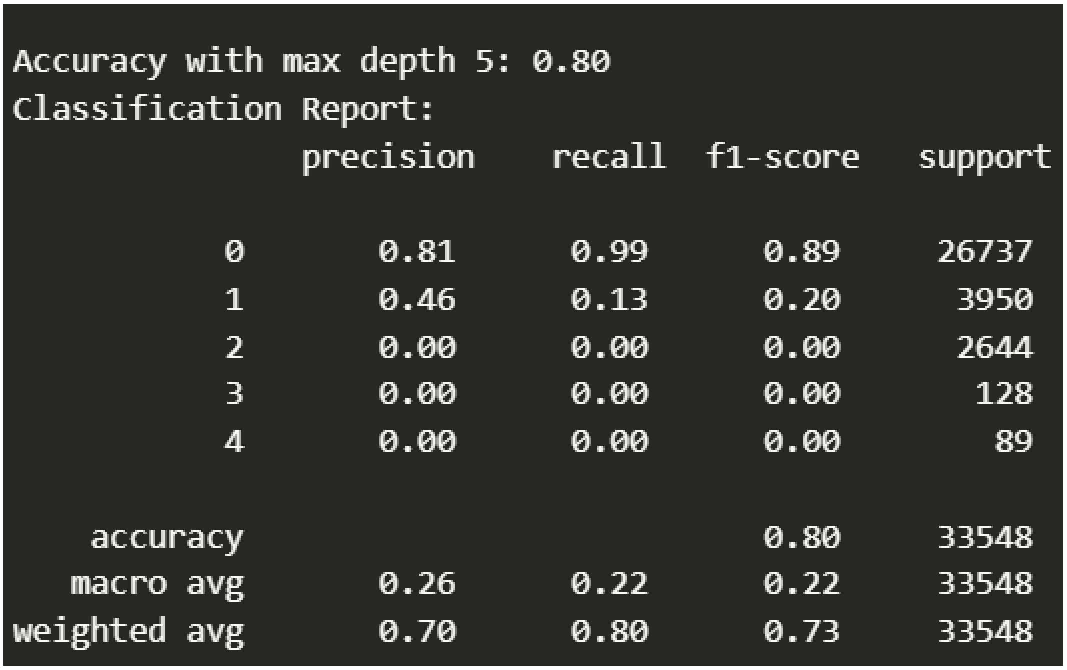


```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     26737
           1       0.00      0.00      0.00      3950
           2       0.00      0.00      0.00      2644
           3       0.00      0.00      0.00       128
           4       0.00      0.00      0.00        89

    accuracy                           0.80     33548
   macro avg       0.16      0.20      0.18     33548
weighted avg       0.64      0.80      0.71     33548
```



```
Accuracy with max depth 4: 0.80
Classification Report:
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     26737
           1       0.68      0.00      0.01      3950
           2       0.00      0.00      0.00      2644
           3       0.00      0.00      0.00       128
           4       0.00      0.00      0.00        89

    accuracy                           0.80     33548
   macro avg       0.30      0.20      0.18     33548
weighted avg       0.72      0.80      0.71     33548
```

With a change in depth, the accuracy improved to 0.80. Decision trees are easy to interpret and can capture complex interactions between features but may be prone to overfitting without proper tuning.



```
Accuracy with max depth 5: 0.80
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.99      0.89     26737
           1       0.46      0.13      0.20      3950
           2       0.00      0.00      0.00      2644
           3       0.00      0.00      0.00       128
           4       0.00      0.00      0.00        89

    accuracy                           0.80     33548
   macro avg       0.26      0.22      0.22     33548
weighted avg       0.70      0.80      0.73     33548
```

Accuracy with max depth 3: 0.80

Is the Model OverFitting while tuning:

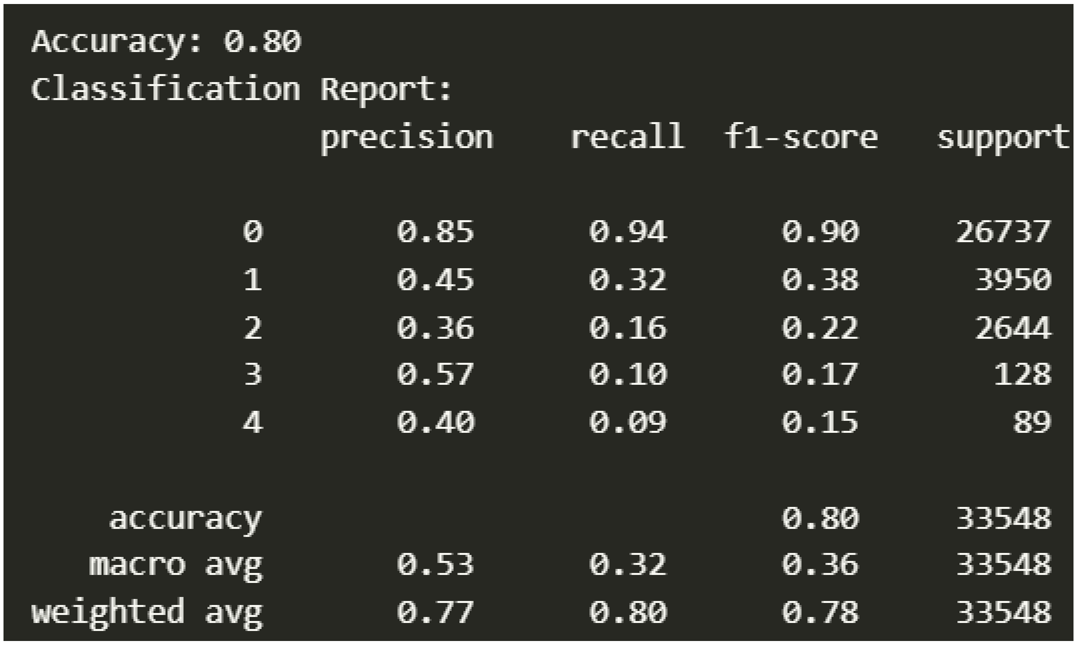
Based on the classification report, the model does not appear to be an overfit model. The accuracy score with max depth 2,3,4,5 is 0.80, which is reasonably good but not exceptionally high, indicating that the model is not overfitting the training data.The precision, recall, and F1-score values vary across different classes. Class 0 has a high precision (0.81) and recall (0.99), while other classes like 1, 2, and 3 have lower scores. The macro average precision (0.26) and recall (0.22) are relatively low, suggesting that the model's performance is not outstanding across all classes. The weighted average precision (0.70) and recall (0.80) are higher than the macro averages, indicating that the model performs better on the more prevalent classes.

Overfitting typically occurs when a model performs exceptionally well on the training data but fails to generalize to unseen data. In these cases, the metrics suggest that the model is not overfitting, but it may be struggling to capture the patterns in certain classes, possibly due to class imbalance or other factors.

In summary, the model indicates a moderately performing model that is not overfitting, but there may be room for improvement in terms of overall performance and handling of minority classes.

Random Forest

During training, random forests (also known as random choice forests) generate a huge number of decision trees to use as an ensemble learning approach for classification, regression, and other problems. The output of a random forest is the class selected by the vast majority of trees, which is useful for solving classification issues. When a regression task is given, the average prediction of the individual trees is given back. Decision trees may overfit their training data, although random decision forests mitigate this problem. Random forests are more effective than decision trees in most cases. So, for the random forest, we just used the default parameters, and the accuracy of the model is 80%. Moreover, the classification report for the random forest is shown below.

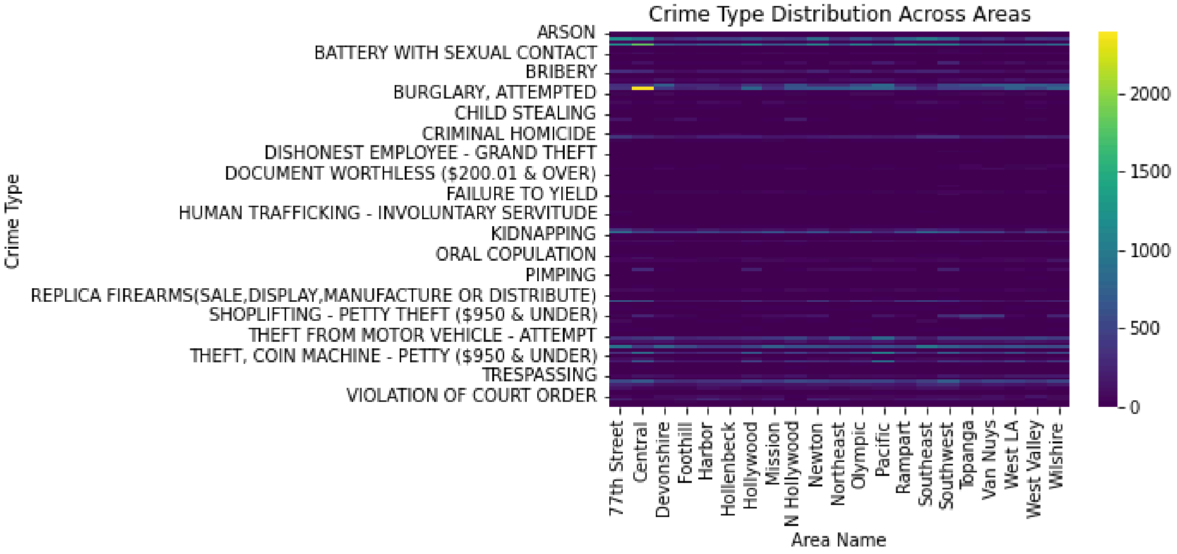


```
Accuracy: 0.80
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.94      0.90     26737
           1       0.45      0.32      0.38      3950
           2       0.36      0.16      0.22      2644
           3       0.57      0.10      0.17       128
           4       0.40      0.09      0.15        89

    accuracy                           0.80     33548
   macro avg       0.53      0.32      0.36     33548
weighted avg       0.77      0.80      0.78     33548
```

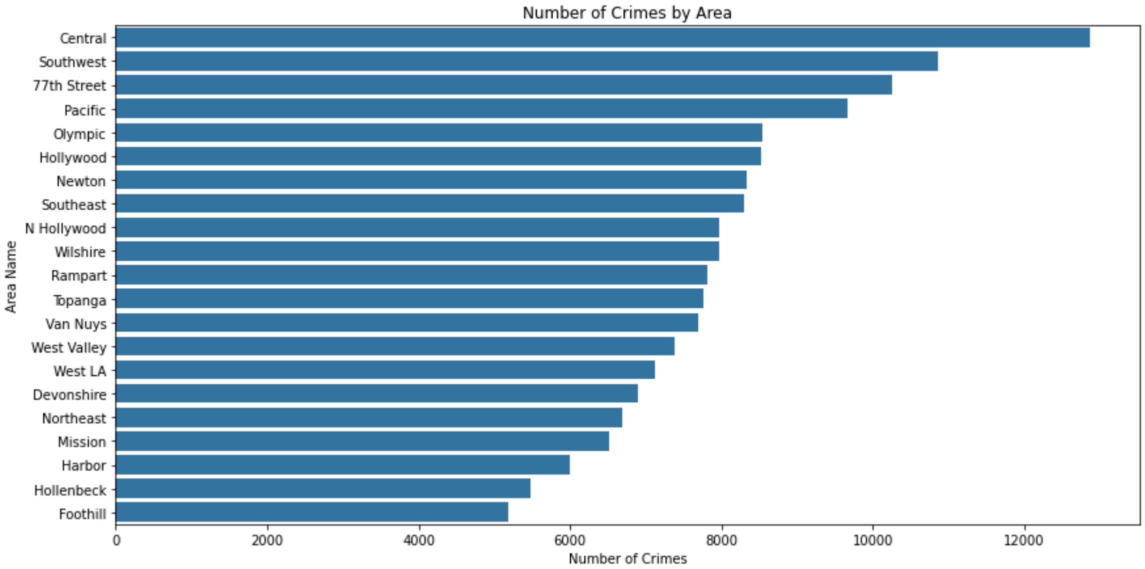
observation: Random forests reduce overfitting compared to single decision trees and can handle large datasets with high dimensionality.

Q3: What are the key factors influencing crime across the neighborhood?

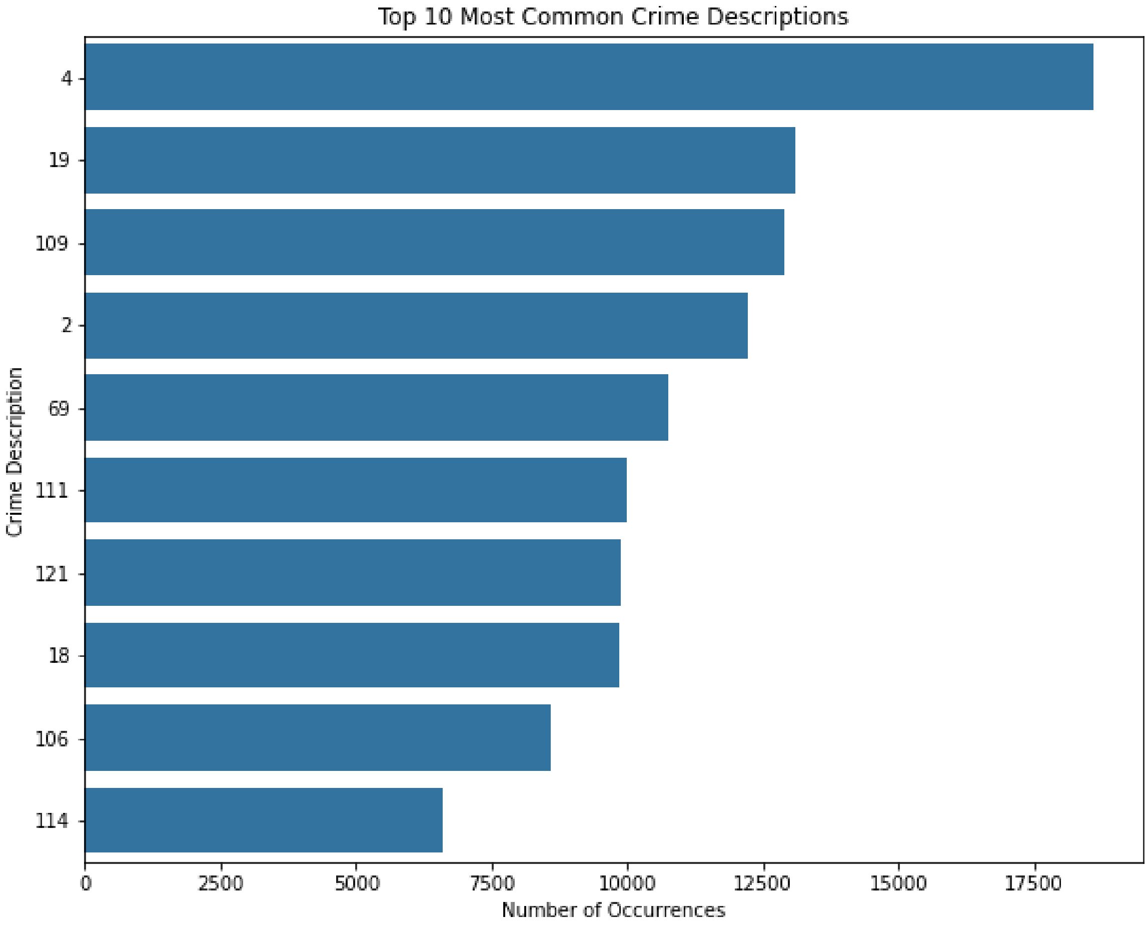
To start, we take look back at a few of our EDA visualizations to find the potential variables that we could use in our statistical test



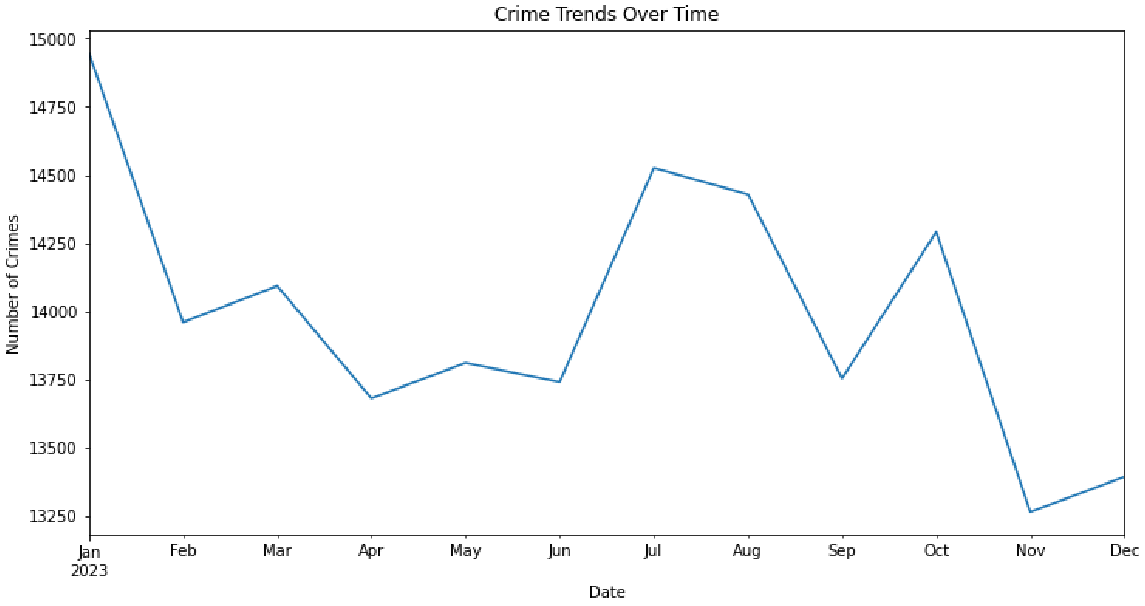
Crime Type Distribution: The heatmap shows how different crime types are distributed across various areas. This can help in identifying if certain crimes are more concentrated in specific regions.



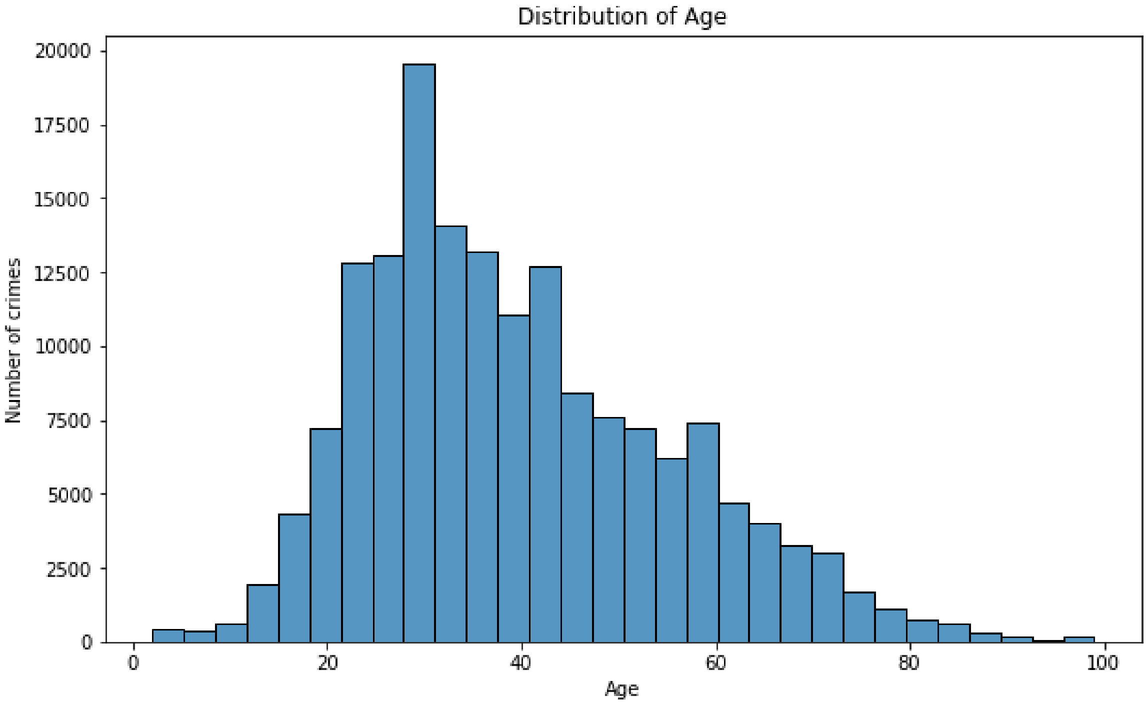
Crime Distribution by Area: The bar chart shows the frequency of crimes in different areas. This visualization clearly identifies which areas have higher crime rates and may require further investigation into what makes them more prone to crime.



Top Crime Types: The list of top crime types gives an idea of the most common crimes, which is useful for prioritizing resources and prevention efforts.



Crime Trends Over Time: The line chart plots the number of crimes per month, giving insights into how crime rates fluctuate throughout the year. It can be helpful to examine these trends in relation to specific events or changes in policy.



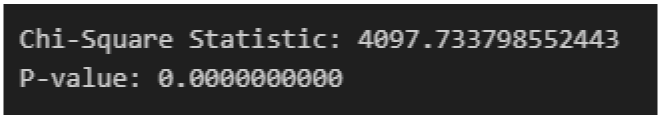
Distribution of Victim Age: The histogram provides an overview of the age distribution of crime victims. This could be cross-referenced with the type of crimes to find age-related trends.

Statistical Tests

Throught statistical test, we can:

- Confirm Relationships: Validate whether the relationships we observe in the EDA are statistically significant.
- Feature Selection: Identify which features have a significant relationship with our target variable and should be included in our models.

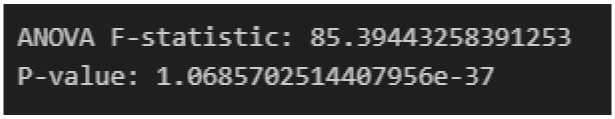
1. Chi-Square Test for Independence: To test if there's a significant relationship between two categorical variables (e.g., AREA NAME and Status).



```
Chi-Square Statistic: 4097.733798552443
P-value: 0.0000000000
```

Chi-Square Statistic: 4097.73 suggests a very strong relationship. P-value: Essentially 0 (even after attempting to show more decimal places), indicating that the relationship between the area and the status of the crime is statistically significant.

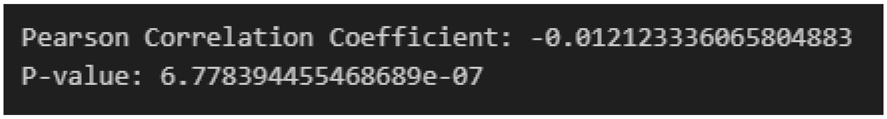
2. ANOVA Test: To compare the means of a continuous variable across multiple categories (e.g., Vict Age across different AREA NAME).



```
ANOVA F-statistic: 85.39443258391253
P-value: 1.0685702514407956e-37
```

F-statistic: 85.39 is relatively high, indicating a strong between-group variance compared to within-group variance. P-value: Approximately 1.07*10^−37, which is virtually zero, shows that there are statistically significant differences in victim age among the different areas.

3. Correlation Test: To measure the strength of the association between two continuous variables (e.g., TIME OCC and Vict Age).



```
Pearson Correlation Coefficient: -0.012123336065804883
P-value: 6.778394455468689e-07
```

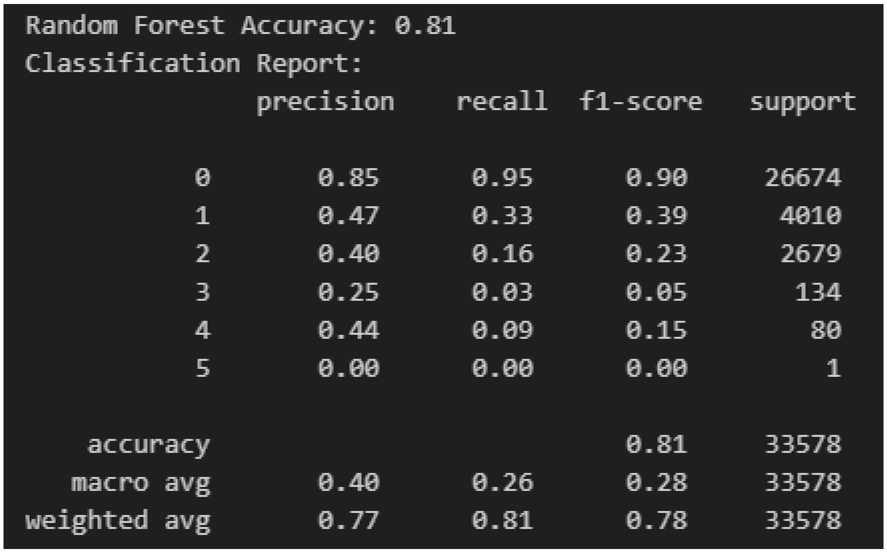
Coefficient: -0.012 suggests a very small negative correlation between the time of the occurrence and the victim's age, which is probably not practically significant.
P-value: Approximately 6.78*10^−7, indicates that the small correlation is statistically significant, but given the size of the dataset, small correlations can become statistically significant even when they might not be meaningful in a practical sense
Based on these results, we have evidence to consider these variables as potentially important features in our predictive models. The next step would be to use these insights to build models

that could understand the factors that contribute to crime in different areas.

We have chosen Random Forest, Gradient Boost, XGBoost, and CatBoost for our predictive model. These chosen models offer a balanced approach, combining performance, flexibility, and interpretability. They are suitable for exploring complex relationships and handling the challenges presented by our dataset. Our selected features are time occurrence, area, crime code, victim age, and weapon. Our target feature is the status

features = ['TIME OCC', 'AREA', 'Crm Cd', 'Vict Age', 'Weapon Used Cd']
target = 'Status'

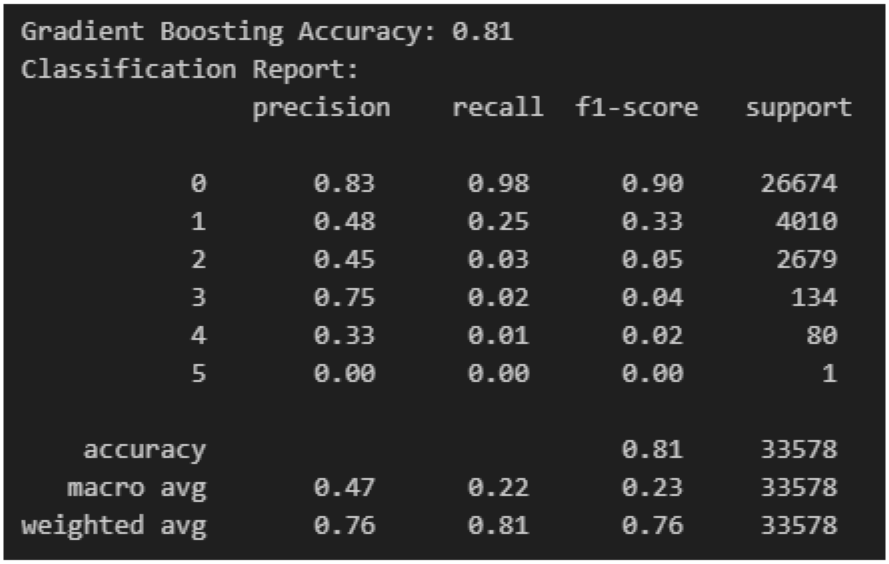
Random Forest



```
Random Forest Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.95      0.90     26674
           1       0.47      0.33      0.39      4010
           2       0.40      0.16      0.23      2679
           3       0.25      0.03      0.05       134
           4       0.44      0.09      0.15        80
           5       0.00      0.00      0.00         1

    accuracy                           0.81     33578
   macro avg       0.40      0.26      0.28     33578
weighted avg       0.77      0.81      0.78     33578
```

Overall Accuracy: 0.81 indicates that the model is reasonably effective at classifying the most prevalent class. Class 0 (Most Common) has high precision (0.85) and recall (0.95) demonstrate that Random Forest is very effective at identifying the most common crime types, with a good balance between identifying positive cases and minimizing false positives. Minority Classes (1-5) significantly drop in precision and recall, particularly for classes with fewer instances (2-5). This suggests difficulty in predicting less common crimes accurately, likely due to insufficient training data for these classes or features that do not distinguish well between classes.

Gradient Boost

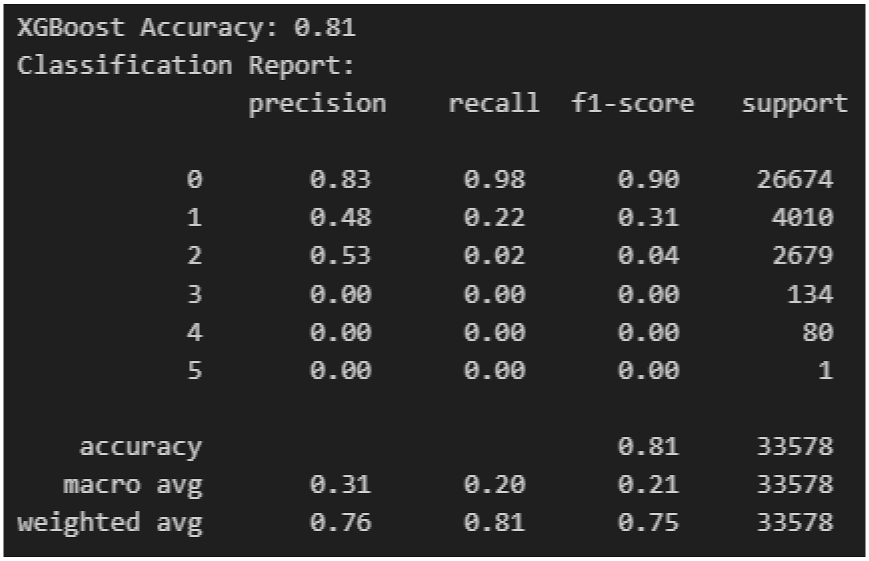


```
Gradient Boosting Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.98      0.90     26674
           1       0.48      0.25      0.33      4010
           2       0.45      0.03      0.05      2679
           3       0.75      0.02      0.04       134
           4       0.33      0.01      0.02        80
           5       0.00      0.00      0.00         1

    accuracy                           0.81     33578
   macro avg       0.47      0.22      0.23     33578
weighted avg       0.76      0.81      0.76     33578
```

Overall Accuracy: Matches Random Forest at 0.81, showing similar effectiveness on a broad level.

Class 0 slightly has lower precision (0.83) but higher recall (0.98) than Random Forest, suggesting Gradient Boosting might be more inclined to classify ambiguous cases as Class 0, potentially increasing false positives. Minority Classes show a similar struggle with slightly better recall for class 3 but still very low overall effectiveness in these categories.

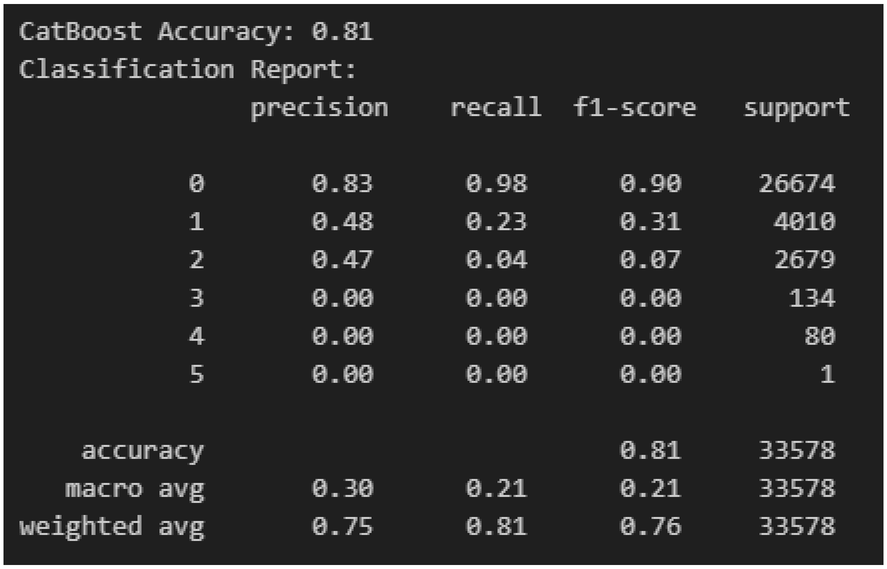
XGBoost



```
XGBoost Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.98      0.90     26674
           1       0.48      0.22      0.31      4010
           2       0.53      0.02      0.04      2679
           3       0.00      0.00      0.00       134
           4       0.00      0.00      0.00        80
           5       0.00      0.00      0.00         1

    accuracy                           0.81     33578
   macro avg       0.31      0.20      0.21     33578
weighted avg       0.76      0.81      0.75     33578
```

Overall Accuracy: Also 0.81, consistent across models. Class 0 is similar to Gradient Boosting in precision and recall, indicating robustness in predicting common crime types. Minority Classes: Marginal improvement in recall for class 2 compared to Gradient Boosting and Random Forest, but overall still low. This slight improvement might be due to better handling of class imbalance or learning subtle patterns not captured by other models. Classes 3, 4, and 5 have no successful predictions (0% recall), which may require further data balancing or feature engineering to improve.

CatBoost



```
CatBoost Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.98      0.90     26674
           1       0.48      0.23      0.31      4010
           2       0.47      0.04      0.07      2679
           3       0.00      0.00      0.00       134
           4       0.00      0.00      0.00        80
           5       0.00      0.00      0.00         1

    accuracy                           0.81     33578
   macro avg       0.30      0.21      0.21     33578
weighted avg       0.75      0.81      0.76     33578
```

Overall Accuracy: Again at 0.81, indicating all models perform comparably in general crime prediction. Class 0 has a high recall (0.98) similar to Gradient Boosting, suggesting CatBoost is also potentially over-predicting this class at the expense of others. Minority Classes show no improvement over the other models in handling the least frequent classes, with effectiveness very low or nonexistent.

Key Findings:
The visualizations revealed significant variations in crime rates across different areas. Areas with high-population density like Central and Hollywood showed higher crime instances. Chi-square tests confirmed a significant association between the area and crime type, suggesting that certain areas are predisposed to specific types of crime. Time-based data analysis highlighted trends such as increased crime rates during specific months or times of day. Statistical analysis showed significant differences in victim age across different crime types, indicating demographic targeting based on age in certain crimes.

From these findings, we conclude the following factors influence crimes in the LA neighborhood:

1. Geographic Area: Different neighborhoods exhibit distinct crime patterns.

2. Time of Crime: Specific times of day and certain months or seasons see higher crime rates.
3. Type of Crime: Variations in crime types across areas indicate different local conditions.
4. Victim Demographics: The age of victims varies with different types of crimes, suggesting targeting or vulnerability of specific groups.

5. Summary and Conclusions

The key results that we identified from performing exploratory data analysis are:

- Publishing articles on Weekends may result in higher shares.

- Publishing articles on topics such as social media and technology may result in higher shares.

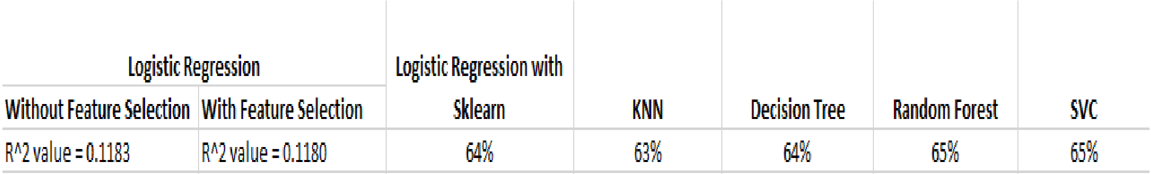
- A good article is one written with a good balance of personal opinions and factual information. This results in an increased number of shares.

- Remaining neutral towards content is important. Although, in some fringe examples, having extremely negative or positive opinions results in a higher share but it's a hit-or-miss situation.

- Number of Words in the title to be between 10 – 18, not too long; not too short to catch the attention of the reader.

- Number of Words in the content should be in the range of 0-2000. Making the article short and catchy results in higher shares.

For Modeling, the summary of the performance of all the Models used can be seen below:



Logistic Regression		Logistic Regression with Sklearn	KNN	Decision Tree	Random Forest	SVC
Without Feature Selection	With Feature Selection					
R^2 value = 0.1183	R^2 value = 0.1180	64%	63%	64%	65%	65%

- Random Forest and SVC gave the best results out of all the models in terms of accuracy.

- The AUC for Random Forest and SVC was also the highest – 0.71, 21% greater than random chance.

Conclusion

The detailed analysis of the 2023 crime data from Los Angeles provided by our team has yielded insightful results, enhancing our understanding of the crime dynamics within the city. Through diligent exploration, preprocessing, and application of advanced statistical methods and machine learning models, we have successfully identified significant patterns and factors influencing crime occurrences.

Our findings clearly show that crime in Los Angeles is highly influenced by geographic and temporal factors. High crime rates in areas like Central Los Angeles call for targeted policing and community initiatives. Furthermore, the temporal analysis highlights specific times of the year and days that require heightened vigilance and resource allocation.

The analysis of victim demographics such as age and sex underscores the need for community support programs tailored to protect vulnerable groups more effectively.

By identifying the most prevalent crimes, our study helps law enforcement agencies to tailor their strategies and training programs to address these common threats more effectively.

Predictive Modeling Efficacy: The deployment of various predictive models has demonstrated that while it is feasible to predict crime outcomes with a reasonable degree of accuracy (approximately 81%), challenges remain in improving the prediction accuracy for less frequent crime types.

In summary, this project not only sheds light on the current state of crime in Los Angeles but also sets the groundwork for future efforts to enhance community safety through data-driven decision-making. Our collaborative efforts have demonstrated the power of data in understanding and combating crime, and we are optimistic that the continued use of these analytical techniques will lead to a safer urban environment for all residents of Los Angeles.

6. References

1. K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.
2. UCI Online News Popularity Data Set, Machine Learning Repository of University of California at Irvine (2015). Retrieved November 6th, 2022, from www.kaggle.com/datasets/thehapyone/uci-online-news-popularity-data-set