IST707: Crime Classification based on Geographical Area

Navya Kiran V B
Madhumitha Saravanan
Sindhuja Maheswaran
Jovita Andrews
Spruha Chandrakant Band

Table of Contents

1. INTRODUCTION

In an era of increasing urbanization and the ever-growing need for information accessibility, the safety and well-being of individuals within specific geographical areas have become a significant concern. Understanding the dynamics of crime within neighborhoods, the composition of local populations, and the safest times for outdoor activities is paramount for both real estate agents, residents, and the general public. With this in mind, we have embarked on a project aimed at developing a robust system for Crime Classification based on Geographical Area.

2. OBJECTIVE

The primary objectives of this project are diverse, aiming to develop a comprehensive and data-driven system that addresses critical concerns related to crime and safety in specific geographical areas. Firstly, the project endeavors to create a sophisticated system that leverages advanced data-driven techniques to categorize and identify various types of crimes based on their occurrence within specific neighborhoods. By doing so, this system will become a valuable resource, shedding light on the intricate crime landscape within these areas, offering insights to law enforcement, researchers, and the general public. Secondly, the project seeks to empower a diverse range of users, including real estate agents, residents, and the wider public, with the ability to assess the safety of different neighborhoods effectively. Through the system, users will gain access to comprehensive crime statistics and trends, enabling them to make informed decisions regarding property investments, residential choices, and daily activities, ultimately enhancing their well-being and peace of mind.

3. ANALYSIS OVERVIEW

3.1 Data Preparation and Cleaning:

- **Data Acquisition:**

  The data for this project was sourced from the Los Angeles Police Department. This dataset is a comprehensive collection of reported crimes in the Los Angeles area, including details such as the type of crime, location, date and time of occurrence, and victim information. The richness of this dataset makes it an ideal source for analyzing crime patterns across different neighborhoods.

- **Cleaning Process:**

  The initial step in the data preparation phase involved cleaning the dataset to ensure its reliability and accuracy for analysis. This process included several key steps:

  Renaming Columns: The original dataset contained column names that were either too cryptic or not intuitive. We renamed these columns for better readability and understanding. For instance, a column named 'VICT_AGE' was renamed to 'Victim Age' for clarity.

  Handling Missing Values: Missing data can lead to inaccurate analysis and biased results. We identified columns with missing values and employed various techniques to handle them. For categorical data, we considered options like filling missing values with the mode (most frequent value) or creating a separate category for missing data. For numerical data like victim age, we used methods such as mean or median imputation.

  Addressing Outliers: Outliers can significantly skew the results of the analysis. We paid particular attention to outliers, especially in the victim age data. We used statistical techniques such as the Interquartile Range (IQR) method to identify and handle outliers. For example, ages that were implausibly high or low (such as a victim age listed as 0 or over 100 years) were examined and addressed appropriately, either by removing these records or adjusting them if it was a data entry error.

3.2 Feature Engineering:

Feature engineering is a critical step in preparing the dataset for meaningful analysis. We added new features to the dataset to enable a more in-depth analysis of the crime patterns:

Crime Severity : We categorized crimes into different levels of severity based on their nature and impact. This categorization was based on criteria such as the level of violence involved, the value of property damage or theft, and the societal impact of the crime. For example, crimes like homicide were classified as 'high severity', while petty theft was classified as 'low severity'.

Victim Demographics: Understanding the demographics of crime victims can provide valuable insights into crime patterns. We created features representing various demographic aspects of the victims, such as age groups, gender, and ethnicity. This allowed us to analyze whether certain demographic groups were more susceptible to specific types of crimes.

3.3 Location-Based Analysis:

Heatmap Creation: We utilized the Folium library to create interactive heatmaps, effectively showcasing areas with higher concentrations of crime. These visualizations provided a clear representation of geographic patterns in crime distribution, aiding in identifying regions with elevated crime rates.

Geospatial Clustering: Applying KMeans clustering techniques, we identified distinct crime hotspots within the city. This analysis revealed significant geographic patterns and trends, highlighting specific areas that are more prone to certain types of criminal activities.

3.4 Victim-Based Analysis

Demographic Breakdown: Our analysis focused on the impact of crimes on different demographics, including age groups, genders (Male, Female, Unknown), and ethnicities (LatinX, White, Black, etc.). This breakdown provided insights into which demographic groups were more frequently targeted by specific crime types.

Visualization Techniques: We employed seaborn and Plotly for generating detailed bar plots and categorical plots. These visualizations illustrated the distribution of crimes across various demographic groups, offering a clearer understanding of the victim profiles.

3.5 Crime-Based Analysis:

Crime Categories: We grouped crimes into distinct categories such as Theft, Violent Crimes, Property Destruction, etc., to enable targeted analysis. This categorization helped in understanding the prevalence and characteristics of different types of criminal activities.

Frequency Analysis: The analysis determined the most common types of crimes and their specific locations. This information is crucial for developing focused crime prevention strategies.

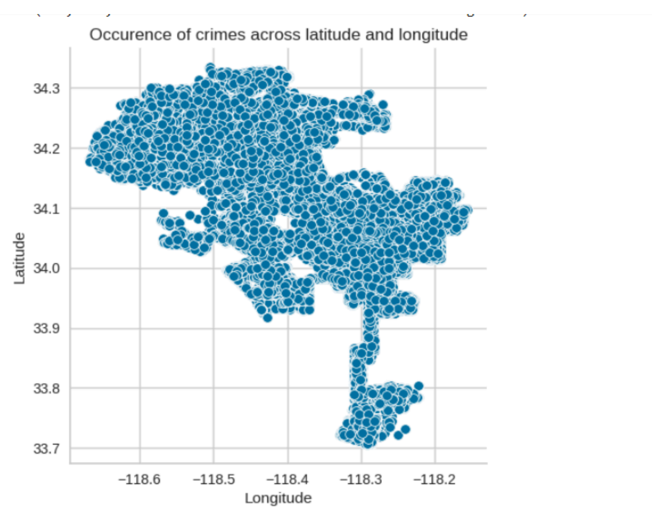
Severity Assessment: We classified crimes into 'Severe' and 'Not Severe' categories based on their respective crime codes. This assessment helped in prioritizing law enforcement resources and attention towards more serious criminal offenses.

3.6 Time Series Analysis

Temporal Trends: Our study extensively examined crime occurrences over different time periods, including years, months, days, and specific time ranges. This temporal analysis was instrumental in identifying patterns and trends in crime rates.
Day and Month Analysis: We identified specific days and months with higher crime rates, noting that Fridays and the month of July were particularly significant. This information can be vital for law enforcement agencies to plan their operations and resources more effectively.

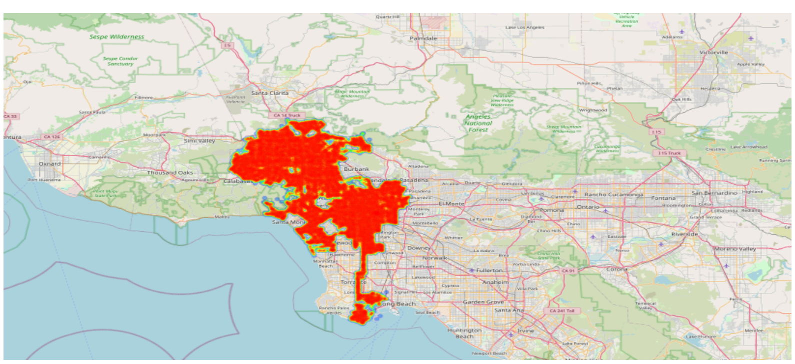
4. EXPLORATORY DATA ANALYSIS

1.Occurrences of crime across Latitude and Longitude



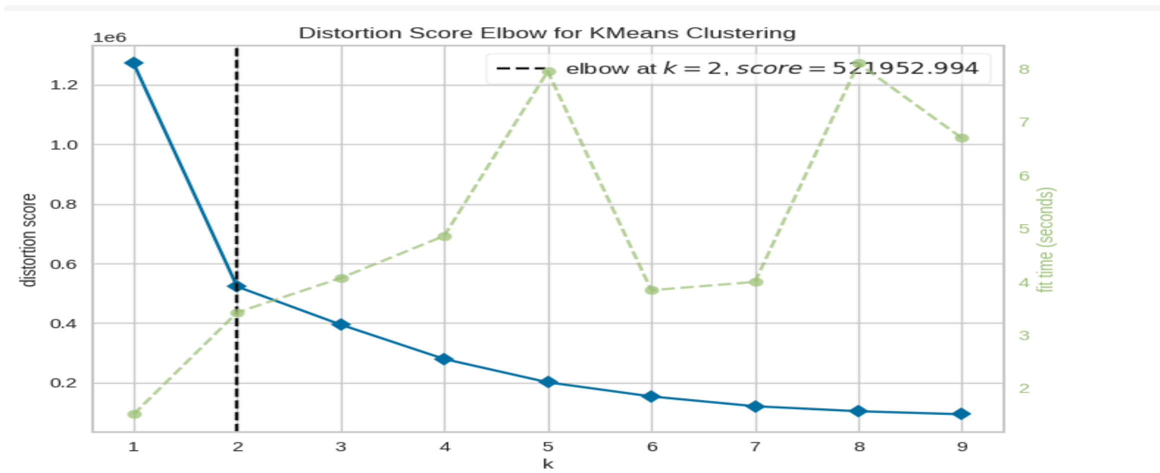
The plot presents a dense clustering of crime incidents based on latitude and longitude, with a high concentration of occurrences in specific areas. The uniform color suggests a singular category or type of crime being represented. The distinct geographic patterns may indicate crime hotspots or areas of particular interest for law enforcement and public safety analysis.

2.Heatmap



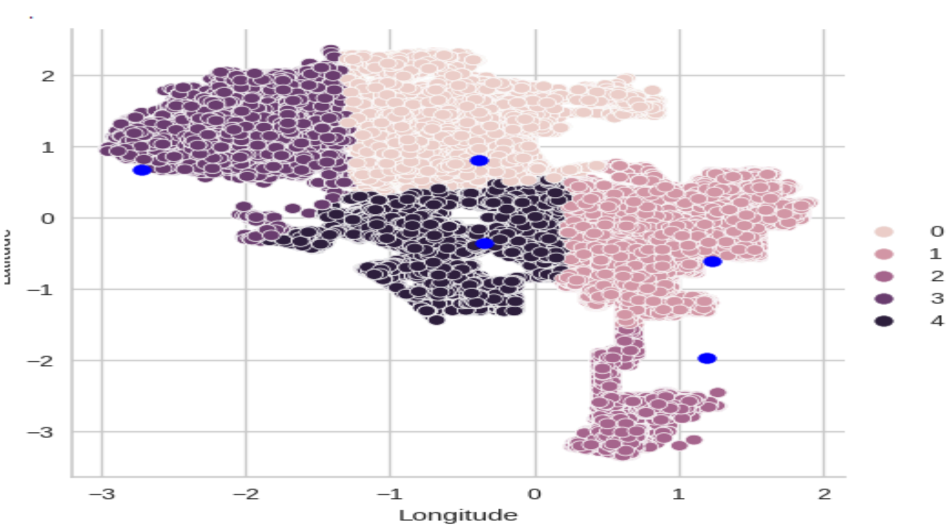
The image depicts a heat map overlay on a geographic map, highlighting areas with high incidences of an unspecified metric, possibly crime or environmental data, concentrated around the Burbank and Santa Monica regions. The intensity of the colors from green to red indicates an increase in frequency or severity of the reported incidents. This visual representation emphasizes areas that may require more focused attention or resources.

3.K-means Clustering:



The graph illustrates the results of an Elbow Method analysis for determining the optimal number of clusters (k) in KMeans clustering. A sharp "elbow" is observed at k=8, suggesting it as the ideal cluster count, after which the distortion score diminishes at a slower rate.

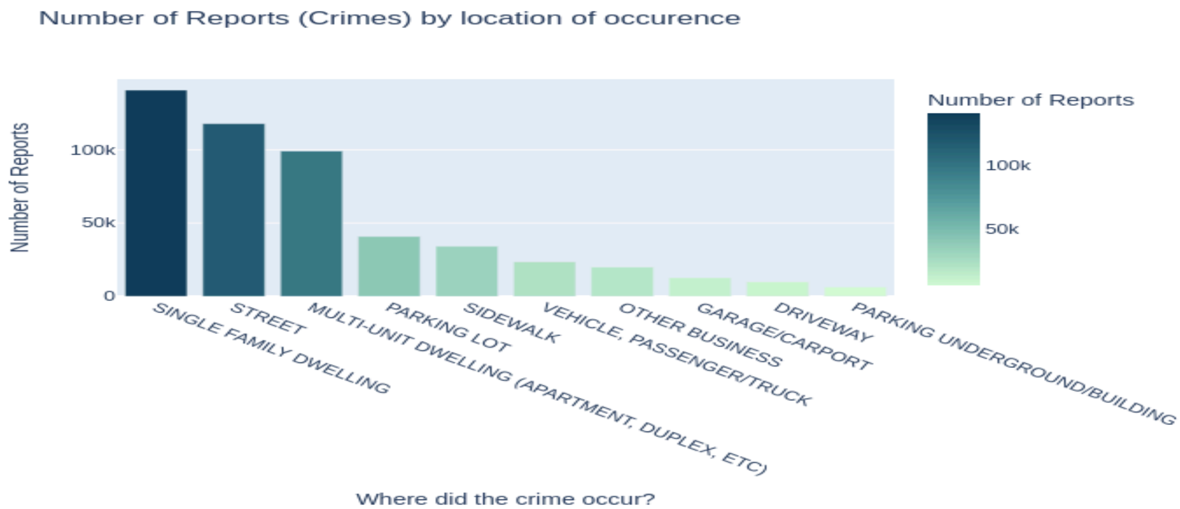
4.Crimes based on Sub regions:



The scatter plot visualizes geographic data points categorized into five groups based on latitude and longitude, displaying distinct clusters that may indicate regions of similar characteristics. A gradient from light to dark shades signifies varying intensities or rankings within the groups.

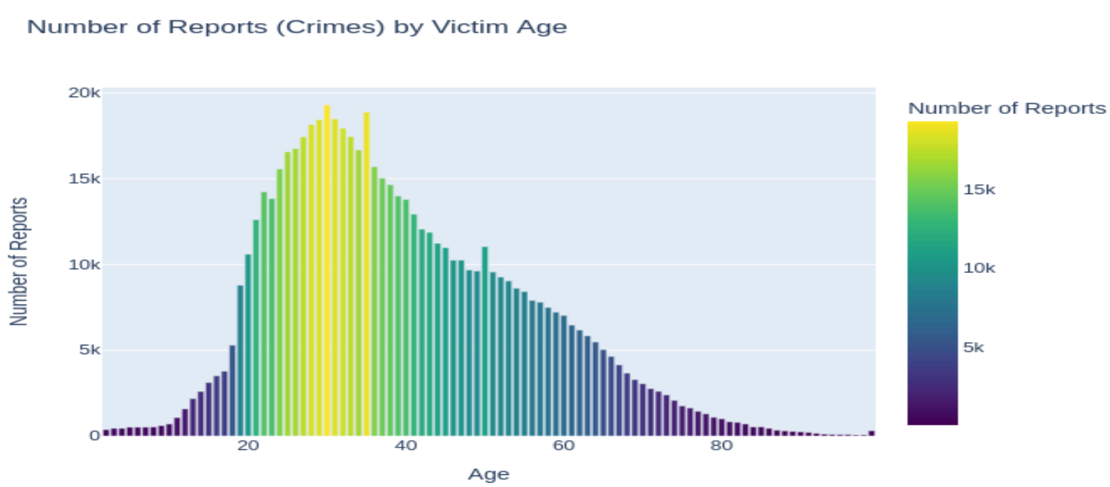
Notably, several blue points stand out, potentially highlighting anomalies or points of special interest.

5. Frequency of crimes reported in various location



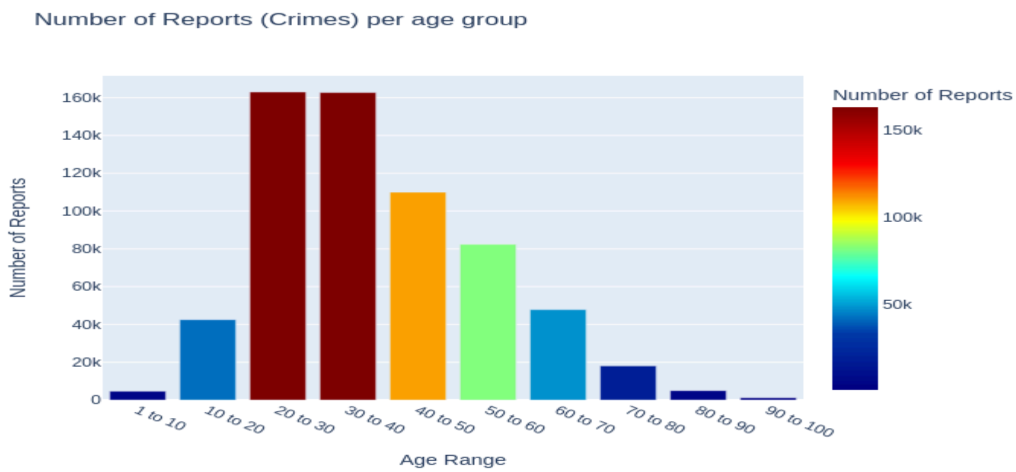
The bar chart displays the number of crime reports categorized by the location of occurrence, with the highest number reported at single-family dwellings, followed by streets. The number decreases significantly in locations such as sidewalks and parking lots, and it is lowest for crimes reported in underground buildings. This visual data suggests that most reported crimes occur in residential areas and on streets, with relatively fewer reports from parking areas and specialized locations like airports and driveways.

6. Analysis of Crime Incidents by Victim Age



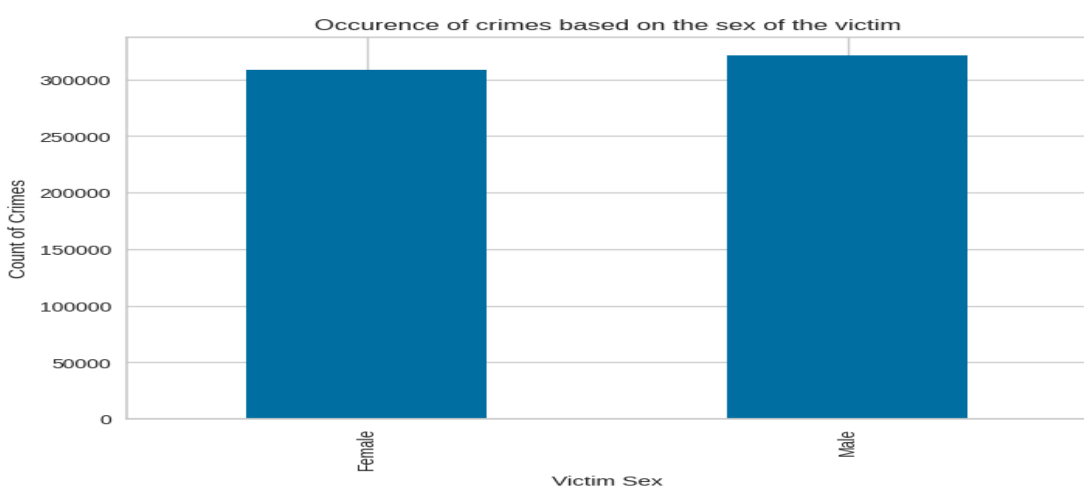
The histogram indicates the distribution of crime reports by victim age, with the highest frequency of reports for victims in their late teens to early thirties. The number of reports declines as age increases, with a significant drop after age 40. The color gradient from purple to yellow to green represents an increase in age groups, visually emphasizing the age range with the highest occurrence of reported crimes.

7. Crime reports distribution by Age groups



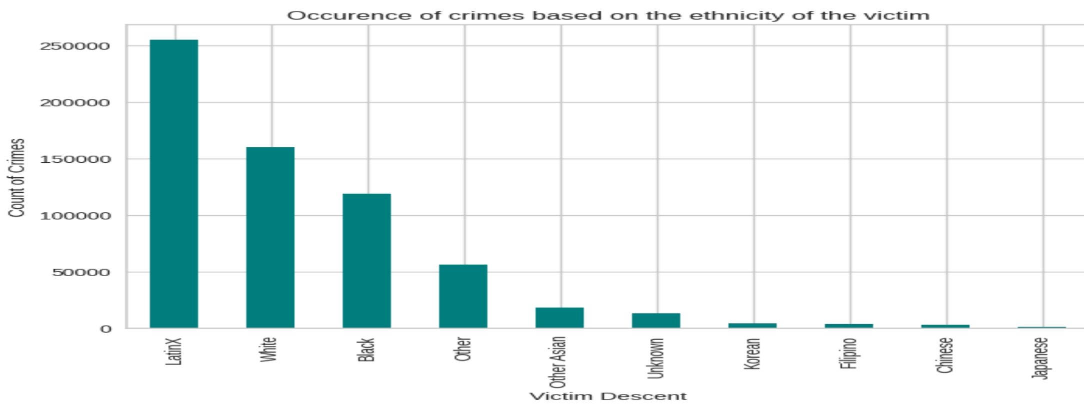
The bar chart shows the number of crime reports segmented by age ranges, with a notably higher number of reports for individuals in their 20s to 40s. The count of reports decreases for younger (below 20) and older age groups (above 60), with the least number of reports for ages 90 to 100. The color gradient corresponds to the age ranges, visually indicating that middle-aged individuals are most frequently reported as crime victims.

8. Comparative analysis of Crime victim count by Gender



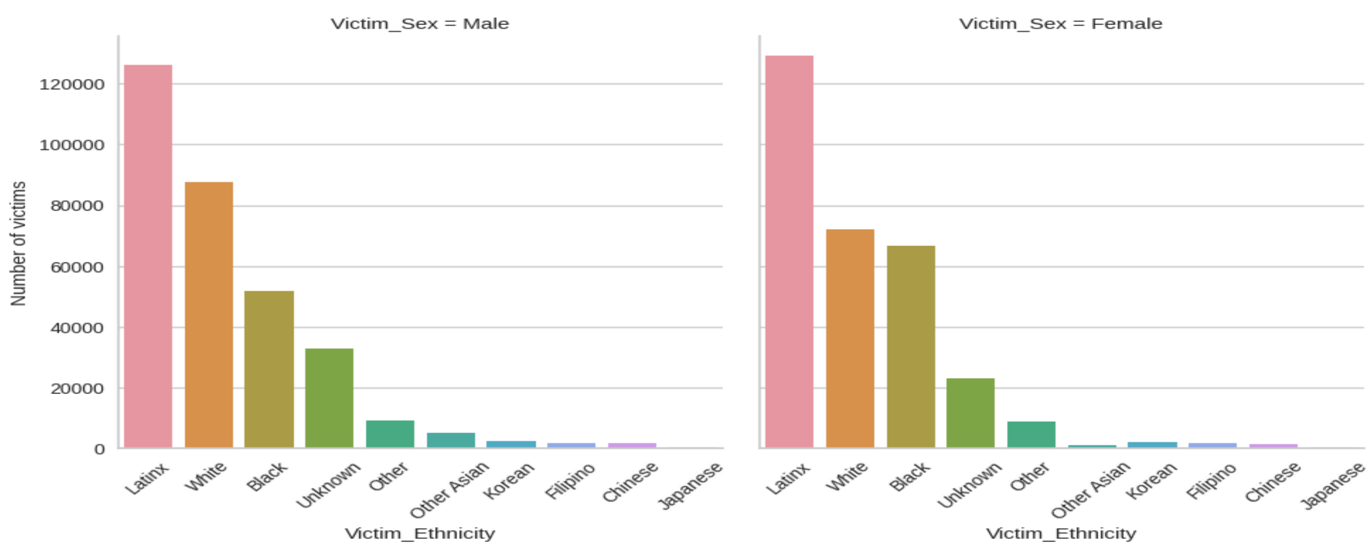
The bar chart compares the occurrence of crimes with respect to the victim's sex, showing a roughly equal distribution between female and male victims. Both bars are close in height, suggesting a marginal difference in crime occurrence by sex. The chart visually represents that the overall count of crimes is high for both categories, exceeding 200,000 incidents for each.

9. Crime statistics by ethnicity of victims



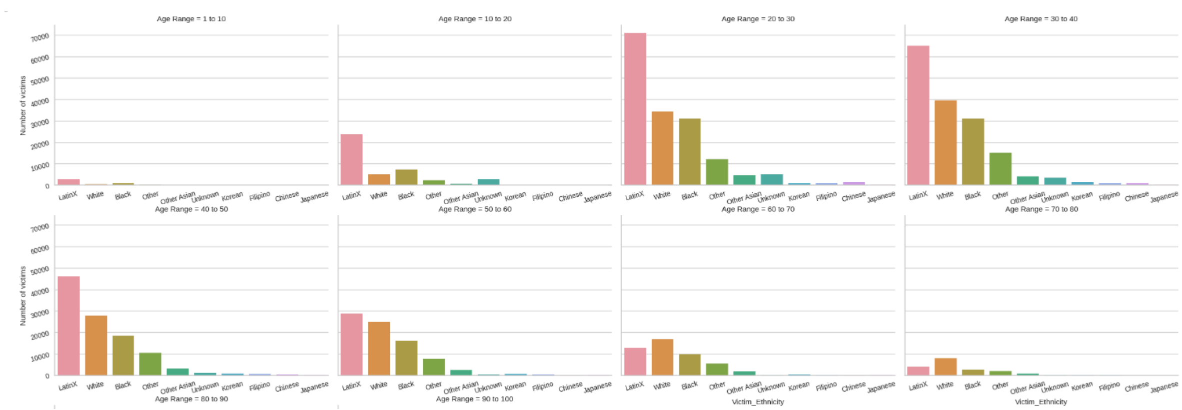
The bar chart details the occurrence of crimes categorized by the ethnicity of the victims, with the highest count for Latinx victims, followed by White and Black. Other ethnic groups, such as Other/Asian, Korean, Filipino, Chinese, and Japanese, have significantly lower reported crime occurrences. This data visualization could be indicative of the demographic distribution within the area studied or of reporting practices.

10. Comparative analysis of Crime victim demographics by Gender and Ethinicity



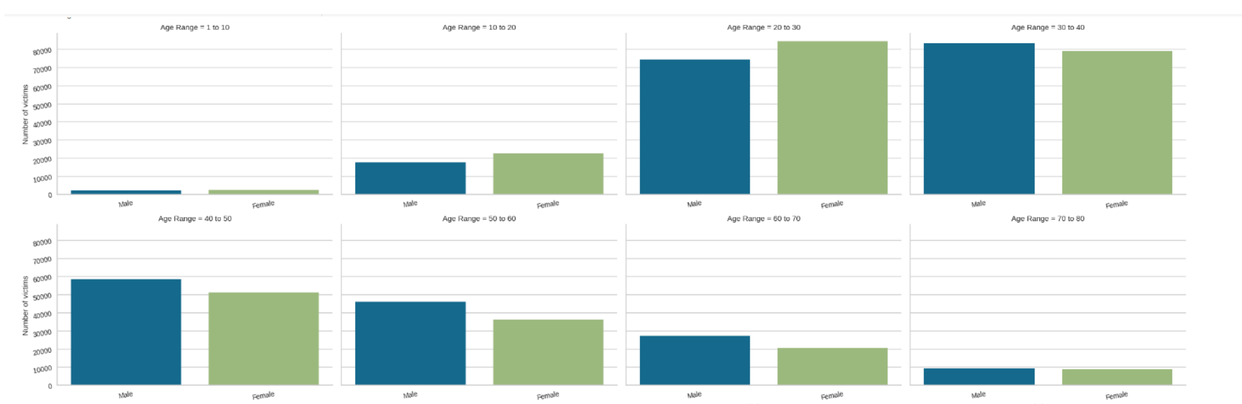
The dual bar charts compare the number of crime victims by ethnicity, separated by sex (male and female). For both males and females, the highest number of victims are reported to be of Latinx ethnicity, followed by White and Black. The charts show a similar distribution pattern across ethnicities for both sexes, with males experiencing a slightly higher number of victimizations in the most represented categories.

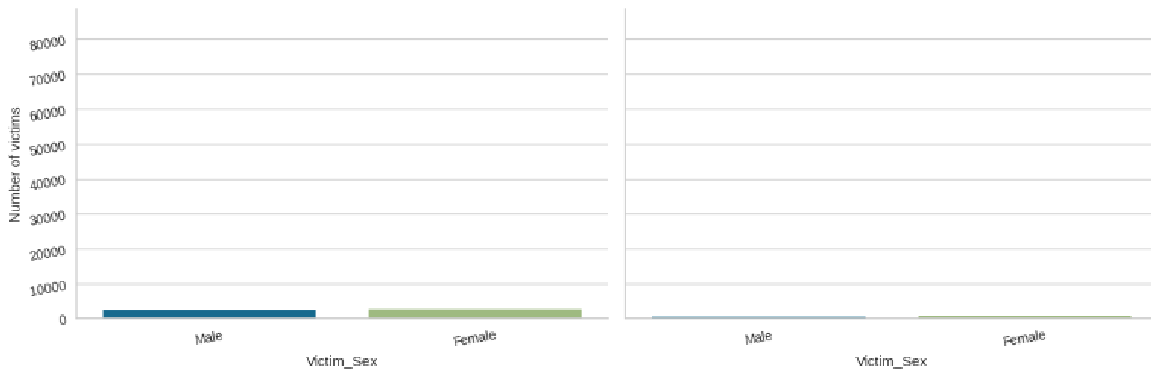
11. Democratic distribution of crime victims by Age and Ethinicity



The series of bar charts depict the distribution of crime victims by ethnicity across different age ranges. Each chart represents a specific age bracket, showing the number of victims in categories such as Latinx, White, Black, and other ethnicities. The charts reveal that certain ethnicities have a higher number of victims within specific age ranges, with notable variations in the distribution patterns as age increases.

12. Age and Gender distribution of Crime Victims



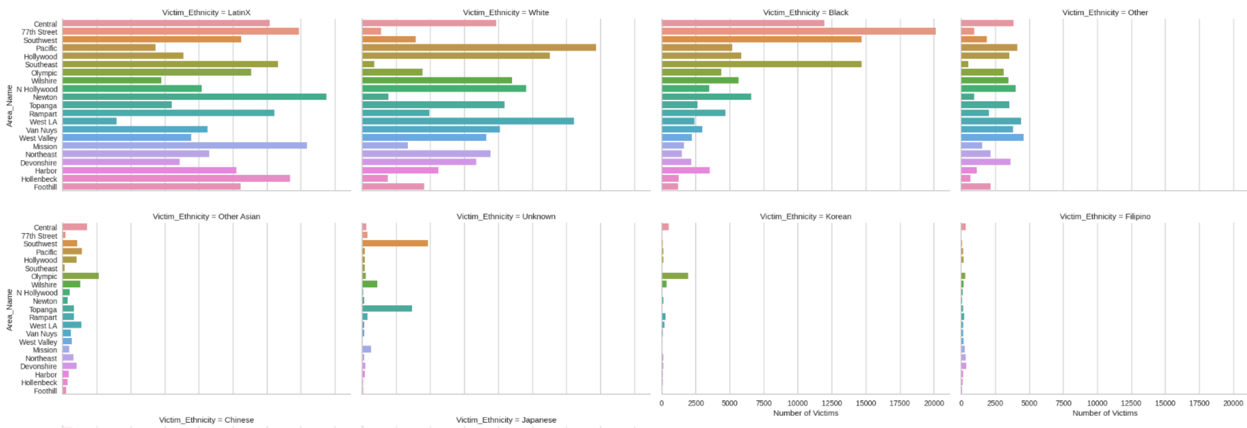


The compilation of bar charts presents crime victimization data segmented by age ranges and further divided by sex. The top set of charts shows the victim count for males and females within specified age brackets, indicating variations in victimization across the lifespan. The bottom chart compares the overall number of crimes for male and female victims, showing a considerable disparity with a higher count for one sex over the other. These visualizations highlight how victimization rates differ by age and sex.

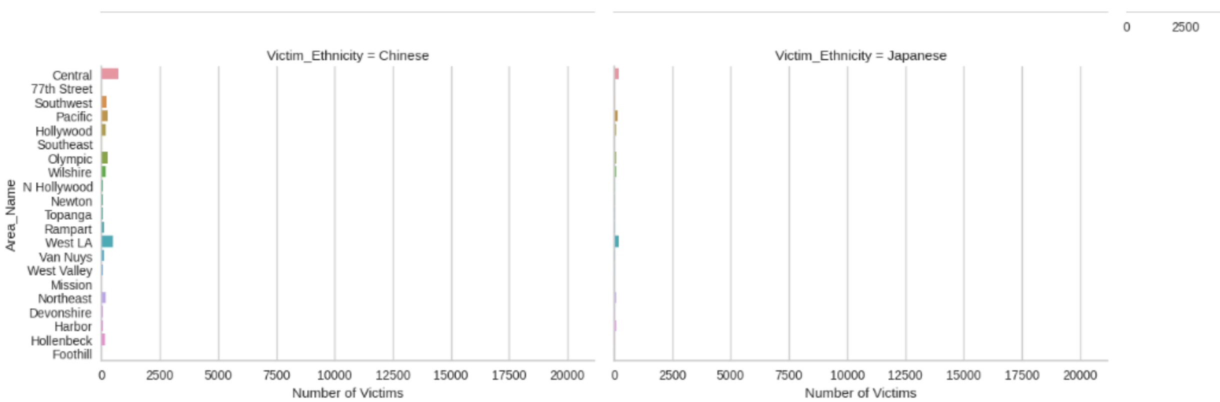
13. Crime Report Frequency by Neighborhood in Los Angeles

The bar chart illustrates the distribution of crime reports by location, with the highest number of reports occurring in 'Single Family/Duplex' and 'Streets/Sidewalks'. The number of reports gradually decreases across various locations such as 'Parking Lots', 'Multi-unit Dwellings', and 'Convenience Stores'. The color gradient suggests a descending order of report frequency from one location type to the next, visually emphasizing the areas with the highest and lowest crime report counts

14. Distribution of Crime Ethnicity

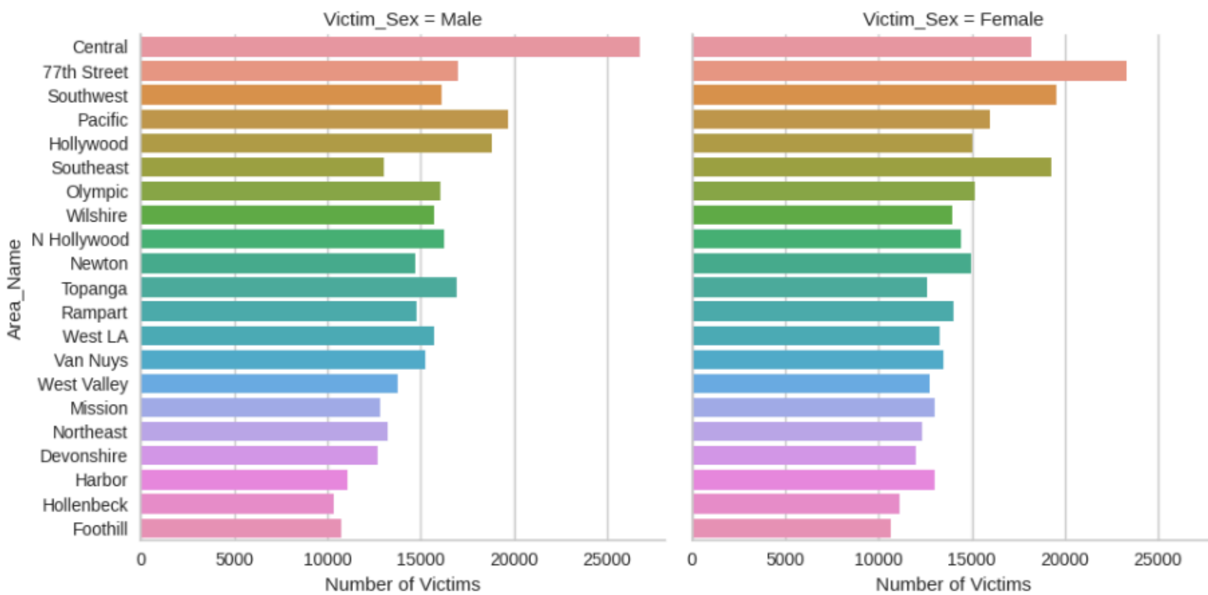


The series of horizontal bar charts display the number of crime victims categorized by ethnicity and the location where the crime occurred. Each chart corresponds to a different ethnic group, including Latinx, White, Black, and several Asian subcategories, among others. The charts show a diverse range of locations from 'Single Family/Duplex' to 'Parking Lots', with varying numbers of victims for each ethnic group, highlighting how victimization patterns differ by location and ethnicity.



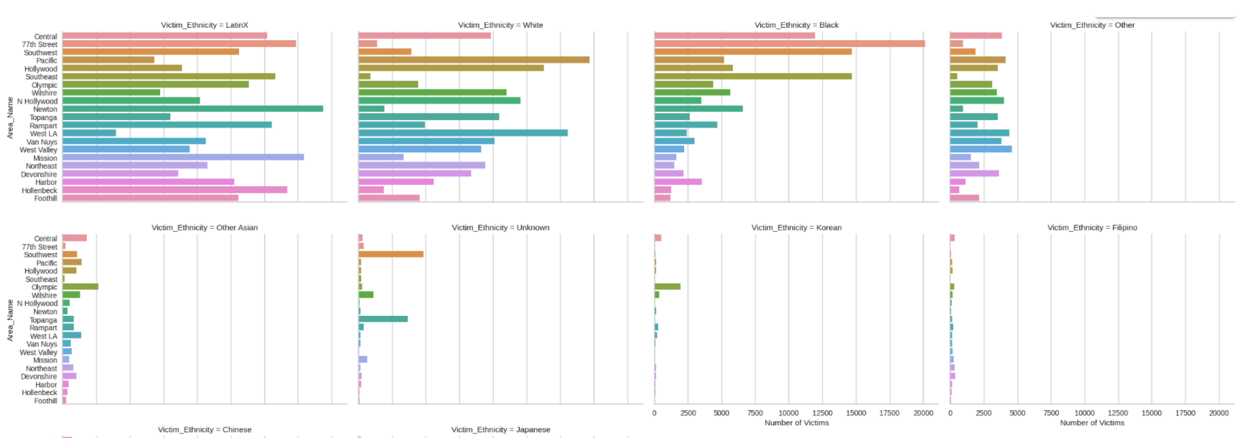
The two horizontal bar charts display crime victimization data for individuals of Chinese and Japanese ethnicity across various police districts. The distribution of victims is uneven, with certain areas showing a higher incidence of crimes for these ethnic groups. Both charts illustrate disparities in the number of victims, suggesting geographic patterns of crime occurrences within these communities.

15. Gender-Specific Crime Victim Statistics Across Los Angeles Police Districts

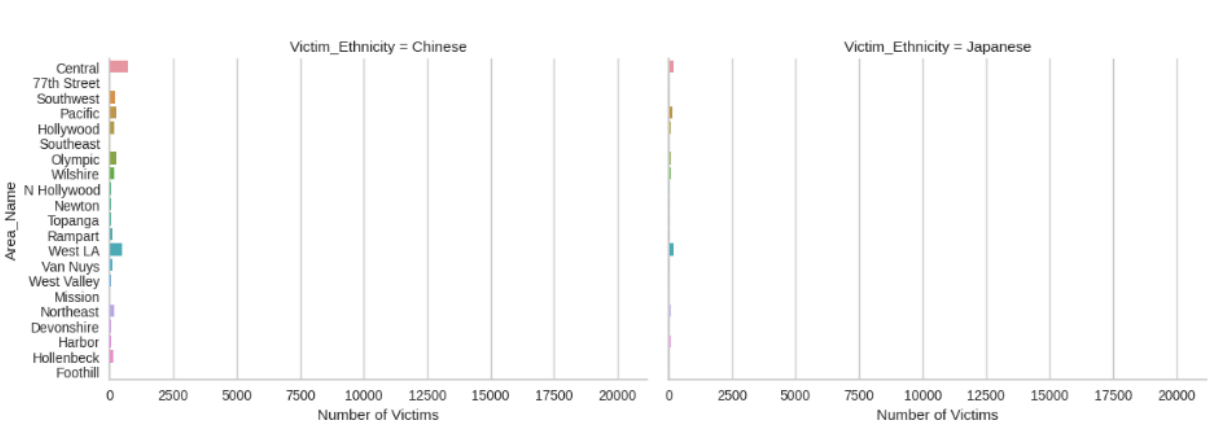


The dual bar charts compare the number of crime victims by sex across various police districts. The chart on the left represents male victims and the one on the right represents female victims, with each bar corresponding to a different district. Both charts show a similar pattern, with the Central district having the highest number of victims for both sexes, followed by the 77th Street and Southwest districts, indicating areas with higher crime incidents.

16. Ethnicity-Based Crime Victim Statistics Across Los Angeles Police Districts

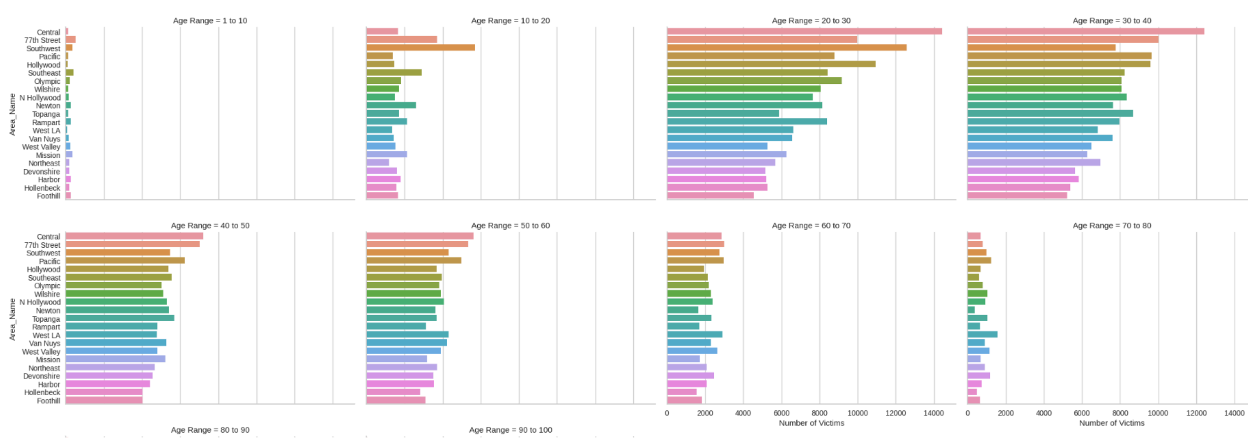


The set of horizontal bar charts displays the number of crime victims in various police districts, segmented by ethnicity. Each chart is dedicated to a different ethnic group—Latinx, White, Black, Other, Other Asian, Unknown, Korean, Filipino, Chinese, and Japanese—showing the count of victims per district. These charts illustrate the distribution of crime across ethnicities and locations, revealing which districts have higher incidences of crime for specific ethnic groups.

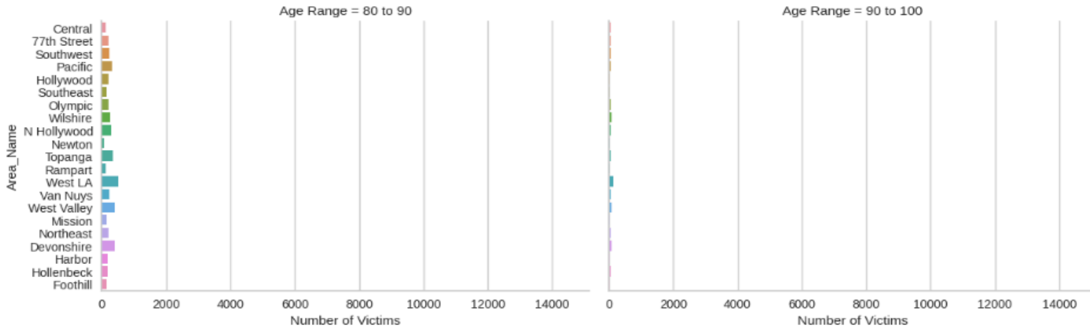


The two horizontal bar charts display the number of crime victims by police district for two ethnic groups: Chinese and Japanese. The "Central" district shows the highest number of victims for both ethnicities. The charts indicate that certain districts have higher crime rates affecting these specific ethnic groups, while other districts report significantly fewer incidents.

17. Crime Victim Statistics by Age Range Across Los Angeles Police Districts

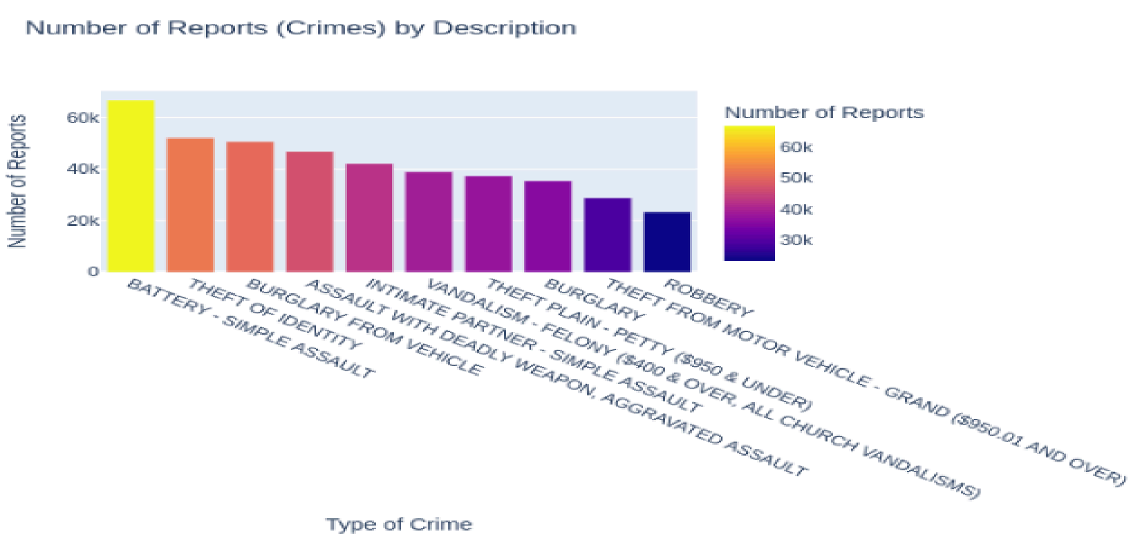


The collection of bar charts represents crime victimization data across different age ranges in various police districts. Each chart is categorized by a specific age bracket, from 0-10 to over 60 years old. The charts show the number of victims in each district within these age groups, highlighting the distribution of crime across both age and location, with some age groups experiencing higher rates of victimization in certain districts.



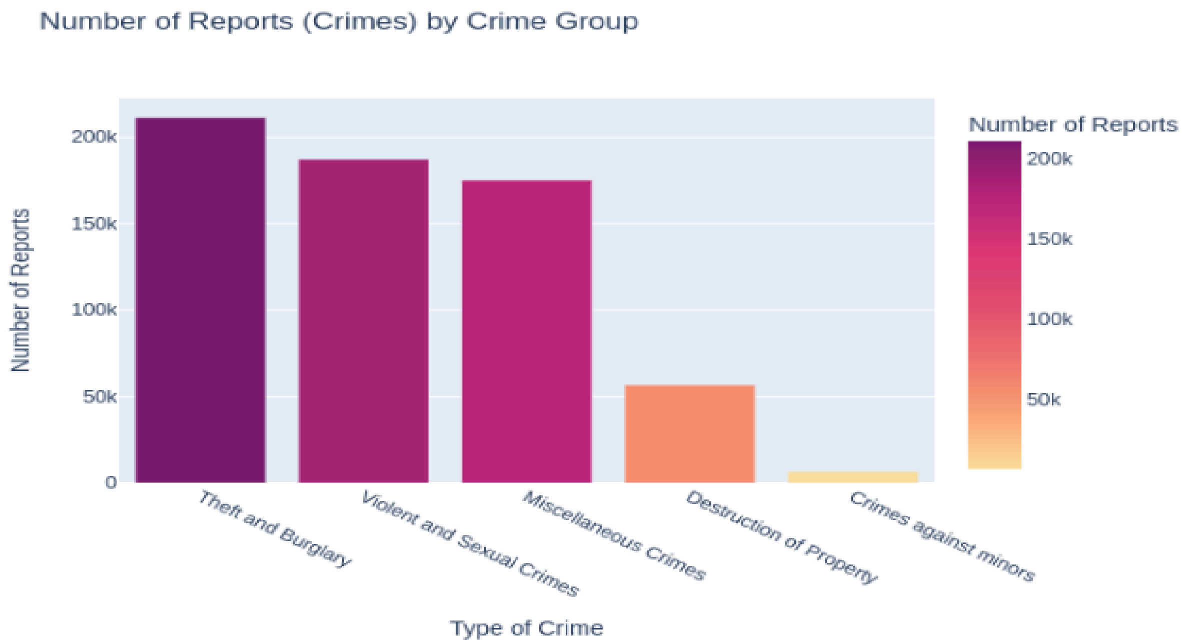
The two horizontal bar charts display the number of crime victims within the age ranges of 80 to 90 and 90 to 100 across various police districts. The charts show a significantly lower number of victims in these higher age groups compared to younger demographics. The distribution across the districts is uneven, with some districts reporting a slightly higher number of victims even within these older age ranges.

18. Frequency of Crime Reports by Specific Offense Description



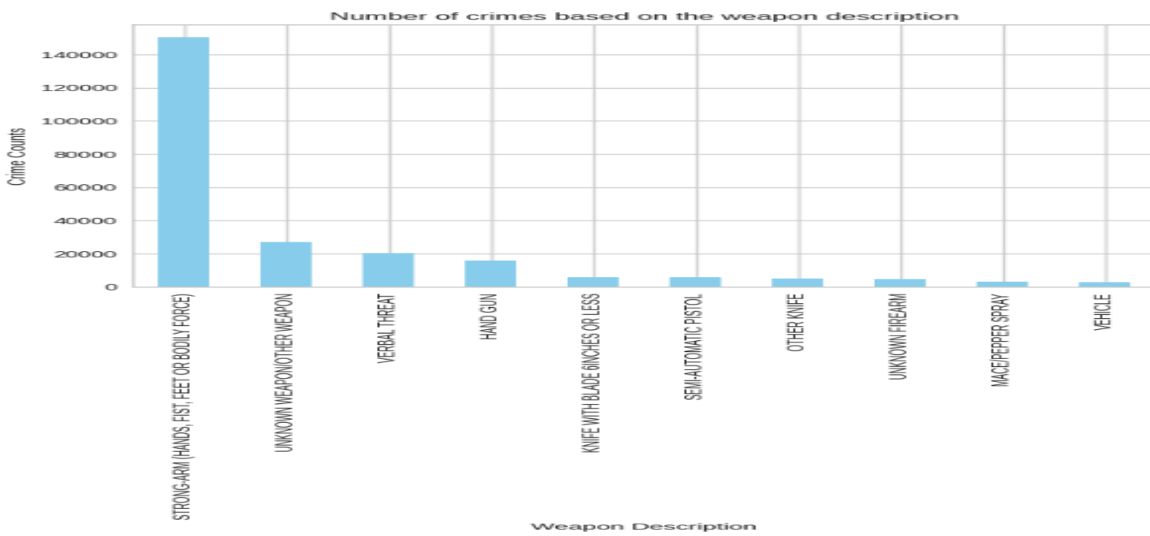
The bar chart visualizes the number of crime reports categorized by the type of crime, with battery-simple assault having the highest incidence, followed by theft from vehicle and burglary. The number of reports for each type of crime decreases from left to right, with aggravated assault being the least frequent. The color gradient from yellow to dark blue represents a descending order of frequency, visually emphasizing the most to least common types of reported crimes.

19. Reported Crime Incidents Categorized by Type of Offense



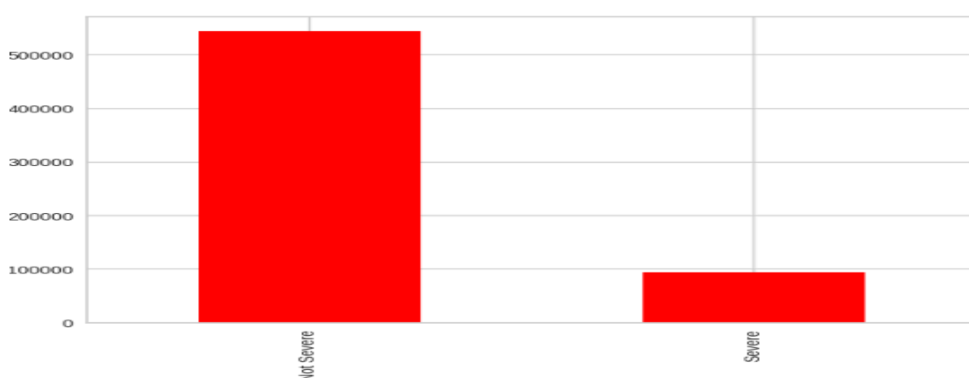
The bar chart categorizes crime reports into five groups, with 'Theft and Burglary' and 'Violent and Sexual Crimes' being the most reported, indicating high occurrences of these types of offenses. 'Miscellaneous Crimes' and 'Destruction of Property' follow with significantly fewer reports, and 'Crimes against Minors' are the least reported group. The color-coding differentiates the crime groups, visually emphasizing the relative frequency of each category.

20. Crime Counts Categorized by Weapon Used



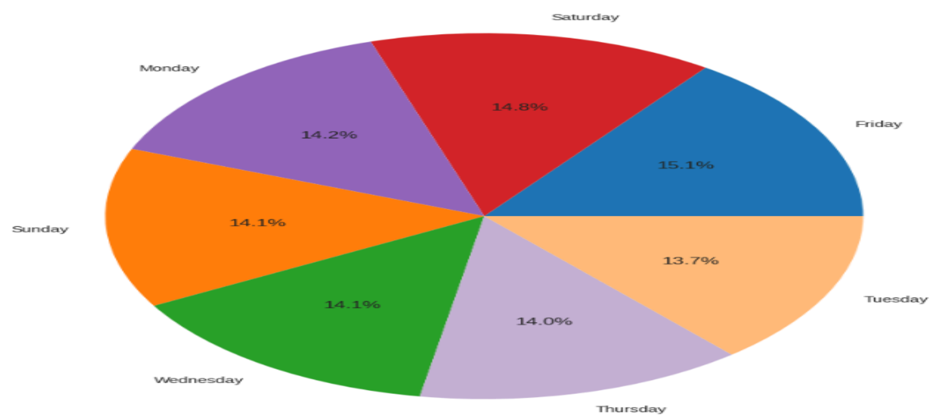
The bar chart illustrates the number of crimes associated with various weapon descriptions, with 'Strong-arm' (no weapon) being the most common, followed by much lower counts for crimes involving knives, firearms, and other weapons. The sharp drop-off in frequency from 'Strong-arm' to other categories indicates that a significant portion of reported crimes do not involve weapons. The chart serves to compare the prevalence of different weapon types used in criminal activities.

21. Incidence of Crimes by Severity Level



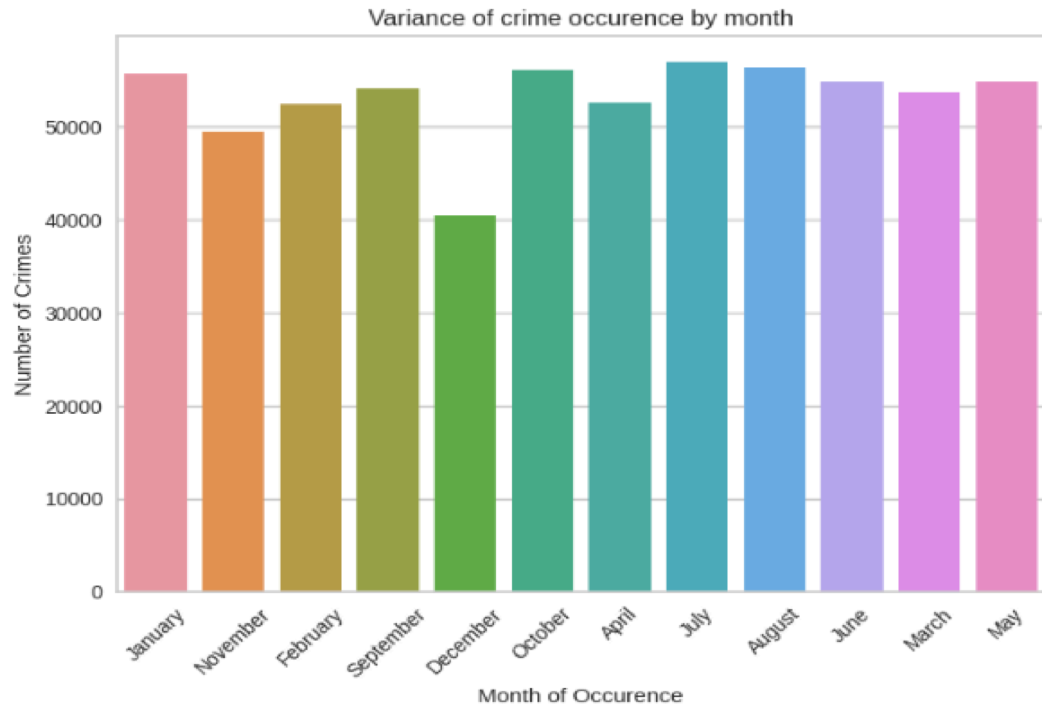
The bar chart contrasts two categories: 'Not Severe' and 'Severe', with 'Not Severe' incidents occurring at a significantly higher frequency than 'Severe' ones. The stark difference in the height of the bars indicates a much larger number of 'Not Severe' cases reported. This visualization is likely meant to emphasize the prevalence of less severe incidents over severe ones within a dataset.

22. Proportion of Weekly Crime Incidents Highlighting Peak Days



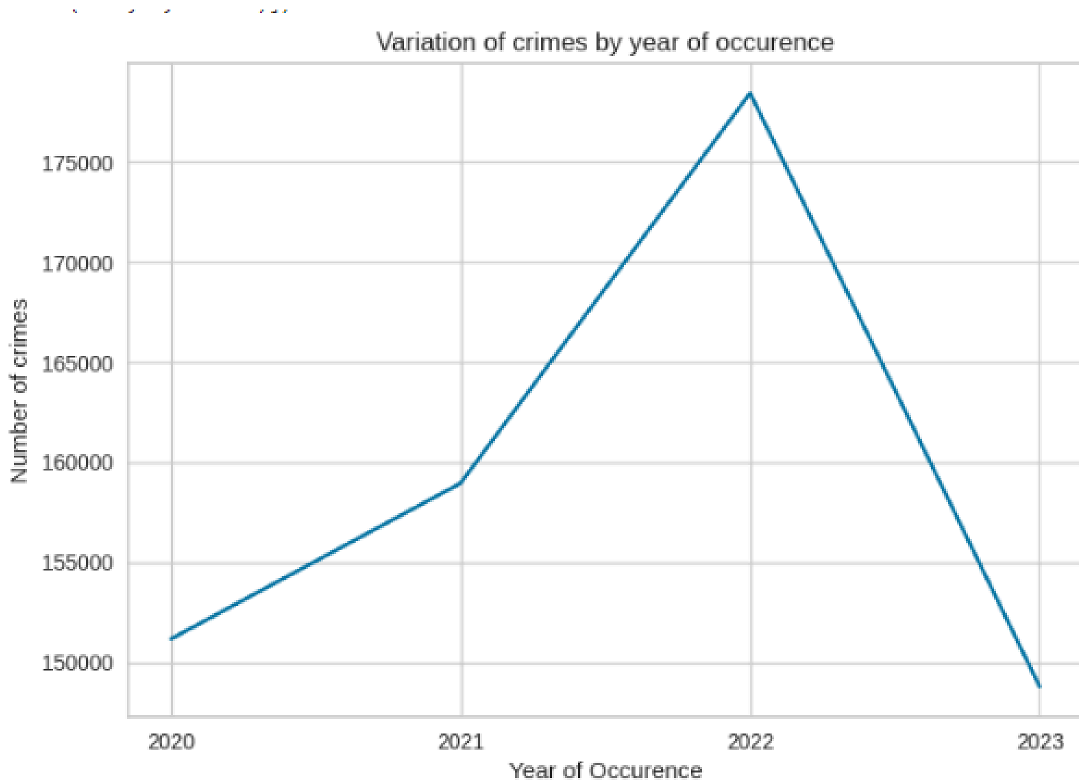
The pie chart displays the distribution of events (possibly crimes) by days of the week, with the highest percentage occurring on Friday (15.1%) and the lowest on Thursday (13.7%). The relatively even distribution suggests that the number of events does not vary drastically by day. The chart visually communicates that while there are slight variations, events are fairly consistent throughout the week.

23. Monthly Crime Occurrence Variation Throughout the Year



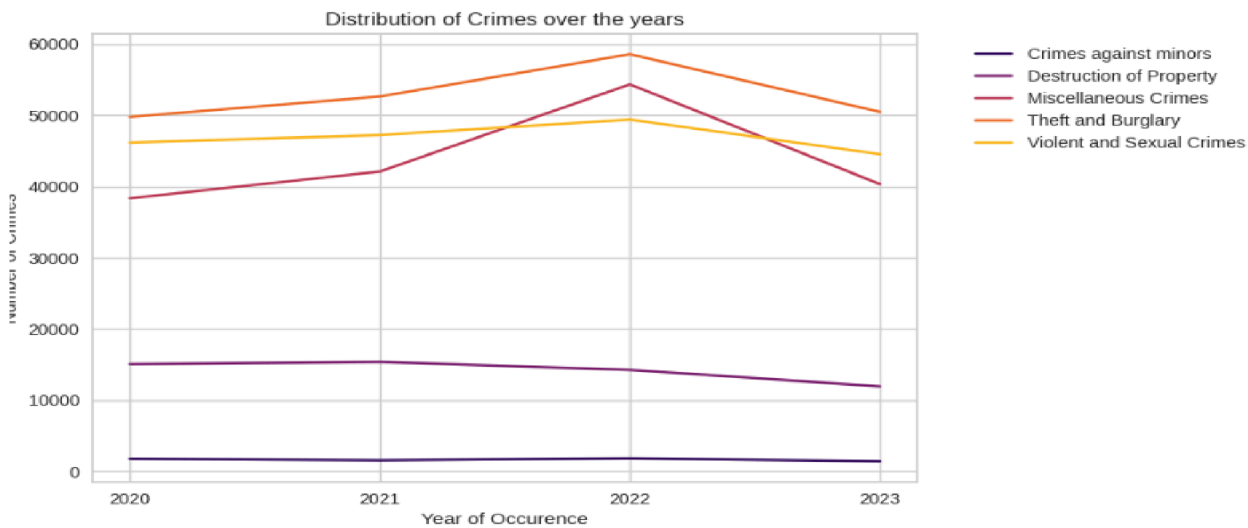
The bar chart illustrates the variance in the number of crimes reported each month, showing that January has the highest number, while September has the lowest. The chart displays a general decrease in crime from the beginning of the year towards the end, with a slight increase again in October. The color gradient does not seem to follow a particular pattern, offering a visual differentiation between the months.

24. Annual Trend of Crime Rates from 2020 to 2023



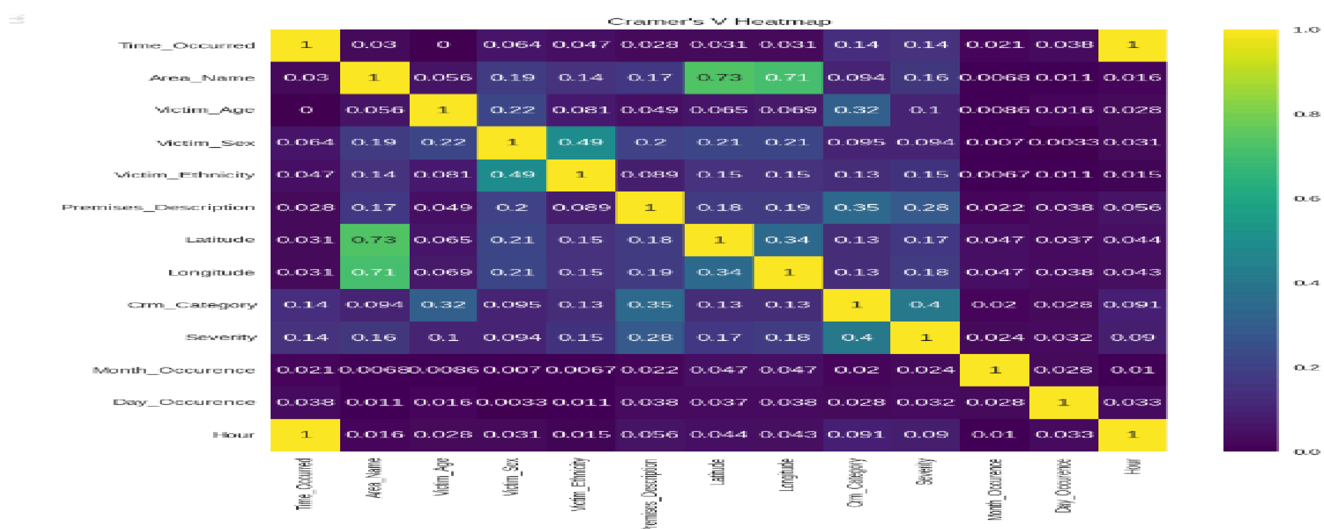
The line graph depicts the variation in the number of crimes from 2020 to 2023, showing a sharp increase in 2022, followed by a significant decrease in 2023. The peak in 2022 suggests a notable year-over-year rise in crime occurrences before dropping to a count lower than the starting point in 2020. This visualization highlights a volatile period with a significant fluctuation in crime rates over the observed years.

25. Yearly Distribution of Different Crime Categories from 2020 to 2023



The line graph shows the distribution of different types of crimes from 2020 to 2023. 'Theft and Burglary' and 'Violent and Sexual Crimes' are the most prevalent, peaking in 2022 before declining in 2023. The graph illustrates a general trend of increasing crime rates across most categories from 2020 to 2022, followed by a substantial decrease in 2023, with 'Crimes against minors' remaining relatively low over the years.

26. Correlation Heatmap of Crime Data Attributes



The heatmap represents Cramér's V statistic values, which measure the strength of association between categorical variables in a dataset related to crimes. High values (closer to 1, indicated in yellow) suggest a stronger association, such as between 'Area_Name' and 'Latitude' or 'Longitude'. Most variables show low to moderate association (indicated by cooler colors), with 'Hour' and 'Time_Occurred' showing a strong association, which is intuitive since they are time-related. This heatmap is useful for identifying which variables have the strongest relationships within crime data.

5. METHODS

Random Forest classifier before SMOTE

	precision	recall	f1-score	support
0	0.78	0.90	0.84	755
1	0.59	0.35	0.44	293
accuracy			0.75	1048
macro avg	0.68	0.63	0.64	1048
weighted avg	0.73	0.75	0.73	1048

XGBoost classifier before SMOTE

```
Training Accuracy:  0.966844044207941 Testing Accuracy:  0.7280534351145038
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.85      0.82       755
           1       0.52      0.41      0.46       293

    accuracy                           0.73      1048
   macro avg       0.65      0.63      0.64      1048
weighted avg       0.71      0.73      0.72      1048
```

Logistic Regression before SMOTE

```
Training Accuracy:  0.7142857142857143 Testing Accuracy:  0.7204198473282443
Classification Report:
              precision    recall  f1-score   support

           0       0.72      1.00      0.84       755
           1       0.00      0.00      0.00       293

    accuracy                           0.72      1048
   macro avg       0.36      0.50      0.42      1048
weighted avg       0.52      0.72      0.60      1048
```

Random Forest classifier after SMOTE

	precision	recall	f1-score	support
0	0.82	0.72	0.77	755
1	0.45	0.59	0.51	293
accuracy			0.68	1048
macro avg	0.63	0.66	0.64	1048
weighted avg	0.72	0.68	0.69	1048

XGBoost classifier after SMOTE

Training Accuracy: 0.9418338108882521 Testing Accuracy: 0.6784351145038168

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.70	0.76	755
1	0.45	0.62	0.52	293
accuracy			0.68	1048
macro avg	0.64	0.66	0.64	1048
weighted avg	0.72	0.68	0.69	1048

Logistic Regression after SMOTE

Training Accuracy: 0.5527220630372492 Testing Accuracy: 0.5524809160305344

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.55	0.64	755
1	0.33	0.56	0.41	293
accuracy			0.55	1048
macro avg	0.54	0.56	0.53	1048
weighted avg	0.64	0.55	0.58	1048

```
Epoch 1/20
14341/14341 [==============================] - 27s 2ms/step - loss: 1.2175 - accuracy: 0.4368 - val_loss: 1.1817 - val_accuracy: 0.4717
Epoch 2/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1751 - accuracy: 0.4740 - val_loss: 1.1687 - val_accuracy: 0.4759
Epoch 3/20
14341/14341 [==============================] - 24s 2ms/step - loss: 1.1626 - accuracy: 0.4811 - val_loss: 1.1539 - val_accuracy: 0.4865
Epoch 4/20
14341/14341 [==============================] - 27s 2ms/step - loss: 1.1535 - accuracy: 0.4854 - val_loss: 1.1460 - val_accuracy: 0.4911
Epoch 5/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1486 - accuracy: 0.4874 - val_loss: 1.1431 - val_accuracy: 0.4917
Epoch 6/20
14341/14341 [==============================] - 26s 2ms/step - loss: 1.1451 - accuracy: 0.4900 - val_loss: 1.1403 - val_accuracy: 0.4946
Epoch 7/20
14341/14341 [==============================] - 26s 2ms/step - loss: 1.1416 - accuracy: 0.4924 - val_loss: 1.1369 - val_accuracy: 0.4954
Epoch 8/20
14341/14341 [==============================] - 27s 2ms/step - loss: 1.1390 - accuracy: 0.4945 - val_loss: 1.1346 - val_accuracy: 0.4991
Epoch 9/20
14341/14341 [==============================] - 26s 2ms/step - loss: 1.1372 - accuracy: 0.4957 - val_loss: 1.1358 - val_accuracy: 0.4990
Epoch 10/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1359 - accuracy: 0.4967 - val_loss: 1.1317 - val_accuracy: 0.4998
Epoch 11/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1349 - accuracy: 0.4971 - val_loss: 1.1330 - val_accuracy: 0.4992
Epoch 12/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1339 - accuracy: 0.4982 - val_loss: 1.1322 - val_accuracy: 0.5007
Epoch 13/20
14341/14341 [==============================] - 26s 2ms/step - loss: 1.1330 - accuracy: 0.4989 - val_loss: 1.1290 - val_accuracy: 0.5014
Epoch 14/20
14341/14341 [==============================] - 24s 2ms/step - loss: 1.1322 - accuracy: 0.4989 - val_loss: 1.1311 - val_accuracy: 0.4990
Epoch 15/20
14341/14341 [==============================] - 27s 2ms/step - loss: 1.1313 - accuracy: 0.4999 - val_loss: 1.1309 - val_accuracy: 0.5002
Epoch 16/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1302 - accuracy: 0.5007 - val_loss: 1.1273 - val_accuracy: 0.5034
Epoch 17/20
14341/14341 [==============================] - 27s 2ms/step - loss: 1.1293 - accuracy: 0.5015 - val_loss: 1.1264 - val_accuracy: 0.5032
Epoch 18/20
14341/14341 [==============================] - 24s 2ms/step - loss: 1.1276 - accuracy: 0.5028 - val_loss: 1.1276 - val_accuracy: 0.5055
Epoch 19/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1264 - accuracy: 0.5037 - val_loss: 1.1274 - val_accuracy: 0.5036
Epoch 20/20
14341/14341 [==============================] - 25s 2ms/step - loss: 1.1250 - accuracy: 0.5052 - val_loss: 1.1227 - val_accuracy: 0.5075
3984/3984 [==============================] - 6s 1ms/step - loss: 1.1252 - accuracy: 0.5062
Test accuracy: 50.62%
```

In the analysis of Los Angeles crime data, we used the XGBoost classifier and RandomForest classifier, logistic regression,and Recurrent Neural Network was trained and evaluated for predicting crime occurrences. Before applying SMOTE (Synthetic Minority Over-sampling Technique), the RandomForestClassifier showed an accuracy of 75% with a higher recall for the majority class (0.90) but lower for the minority class (0.35). The XGBoost Classifier pre-SMOTE showed slightly lower accuracy (72.8%) with similar trends in recall. Post-SMOTE intended to balance the class distribution, the RandomForest and XGBoost models showed an overall accuracy of 68%. Logistic Regression, maintaining a consistent accuracy of 72%.
A SimpleRNN-based neural network model is trained for a classification task. The target variable is encoded and one-hot encoded, while the predictor features are scaled using StandardScaler. After training for 20 epochs, the model achieves a test accuracy of approximately 50.62%.

6. CONCLUSION AND FUTURE SCOPE

The development of a location-based crime identification system designed to address the specific needs and inquiries of real estate agents and community members has the potential to significantly improve safety awareness and decision-making. It empowers individuals to make informed choices about their living arrangements and outdoor activities, ultimately contributing to safer and more informed communities. To ensure its effectiveness, collaboration with reliable data sources, the implementation of advanced machine learning techniques, a user-friendly interface, stringent privacy protection, and accessibility considerations are vital. This system has the potential to make a positive and lasting impact on the lives of its users, enhancing their overall quality of life.