Morgan Lind

Github Link:

Final Project Report

Introduction

Los Angeles, one of the most populous and diverse cities in the United States, has a population of approximately 3.8 million residents as of 2023. California is home to nearly 39 million individuals (<https://datacommons.org/place/geoId/0644000utm_medium=explore&mprop=count&popt=Person&hl=en>). This diversity shapes a wide range of social and economic dynamics, including public safety and crime patterns.

According to recent reports from the Mayor of Los Angeles in 2024, present a decrease in crime citywide, with a 14 percent decline in homicides and a 19 percent reduction in shooting victims compared to previous years (<https://mayor.lacity.gov/news/lapd-releases-2024-end-year-crime-statistics-city-los-angeles.>). These improvements highlight the city's ongoing efforts in crime reduction and community safety. However, the decline in crime may not be evenly distributed across all communities and does not include all categories of crime, prompting questions about how demographic characteristics such as ethnicity, age distribution, and gender relate to crime patterns in specific communities.

This project is motivated by the goal of understanding why certain communities are more vulnerable to crime than others. The analysis will examine multiple factors contributing to these disparities. Gaining this information can support the creation of new public policies, the equitable allocation of resources, and the development of strategies to improve safety across all communities.

In this project, I will use two datasets to analyze the relationship between demographics and crime in Los Angeles during 2024 and 2025. One dataset contains crime reports, while the other provides demographic information at the community level. The analysis will focus on identifying high-crime communities and examining whether certain demographic groups are disproportionately affected. This study aims to provide a clearer understanding of the social dynamics influencing public safety across Los Angeles.

Data

This project uses two main data sources. The first dataset, available at <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>, contains crime reports from 2020 to the present. For this analysis, only data from 2024 and 2025 were used. Additional details about this dataset, including its data dictionary, can be found here <https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data>. The second dataset consists of demographic information scraped from <https://www.laalmanac.com/population/po24la.php>, which provides population and ethnicity breakdowns by community. To integrate this information with the crime dataset, the demographic data was grouped and matched to corresponding AREA\_NAME values. Two additional sources were used to help match neighborhoods to their respective divisions: <https://lacounty.gov/government/about-la-county/maps-and-geography/> and <https://www.rubyhome.com/communities/los-angeles-county/>.

*Review Crime Data*

I obtained data from Data.gov in CSV format. Using Excel, I filtered the data to only contained crime records that occurred in 2024 and 2025. Subsequently, I removed columns that were irrelevant to the objectives of this analysis. The data set had numerous missing values, particularly information regarding the victim (VICTIM\_AGE, VICTIM\_SEX, VICTIM\_DESCENT), to fix this, I imputed these values with ‘X’ to indicate unknown. For the remaining missing values, I imputed them with ‘Unknown,’ or I matched similar crime codes or descriptions elsewhere to fill in the appropriate values. After resolving all missing data, the DATE\_OCC column was converted to a datetime format (MM/DD/YYYY). The updated, cleaned, and structured dataset was saved to the csv file cleaned\_crime\_data.csv.

*Review Los Angeles Community Demographic Populations*

The website containing demographic population data for communities in Los Angeles presented all information on a single page, with each table including both raw population counts and population proportions. However, the website’s structure was difficult to scrape due to the irregular formatting like missing data in columns and or page breaks between tables. To successfully retrieve all information, I scraped neighborhood row, skipped over the total population column. Then I continued scraping for each demographic group, only including every other row with raw population numbers. Since the scraped data was already in a table format, no cleaning was required, so I saved the CSV file total\_la\_population.csv.

*Combining Crimes and Population*

Since both datasets referenced geographic location in Los Angeles, I merged them based on those identifiers. However, the crime data set only included broader police district areas (AREA\_NAMES), while the population data set provided a detailed breakdown by individual neighborhoods. To find a solution, I researched to determine which neighborhoods fall under each police department’s geographic range. This was a very time-consuming step due to my unfamiliarity of Los Angeles’ geographic layout. After matching each neighborhood to its corresponding police area, I renamed the Neighborhood column in population data set to AREA\_NAME to align with the crime data set. This enabled a successful merge of the datasets using a left join, allowing for all records in the crime data to be kept and population observations were only included if their identifier matched.

Following this merge, I observed that the resulting dataset contained over one million rows, with duplicated DR\_NO incident report records. This required data cleaning, starting with dropping all duplicate records. Then converting DATE\_OCC to a datetime format and extracting the year and month from that column. Finally, I dropped columns that were unnecessary for the analysis. The merged dataset was cleaned and formatted to begin the analysis portion.

*Table 1 Data Dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Location | Type | Description |
| DR\_NO | Crime Rates | Integer | Division of Records Number: Official file number made up of a 2-digit year, area ID, and five digits |
| DATE\_OCC | Crime Rates | Date | Date on which crime occurred. |
| AREA\_NAME | Crime Rates | Object | Geographic Areas or Patrol Divisions given a name designation that references a landmark or the surrounding community that it is responsible for. |
| CRM\_CD\_DESC | Crime Rates | Object | Defines of the crime committed. |
| VICT\_AGE | Crime Rates | Float | The age of the victim of a crime. |
| VICT\_SEX | Crime Rates | Object | F - Female M - Male X - Unknown |
| VICT\_DESCENT | Crime Rates | Object | Code: A - Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native Japanese K - Korean L – Laotian O – Other P – Pacific Islander S – Samoan U – Hawaiian V – Vietnamese W – White X – Unknown Z – Asian Indian |
| PREMIS\_DESC | Crime Rates | Object | Defines type of structure, vehicle, or location where the crime took place. |
| WEAPON\_DESC | Crime Rates | Object | Defines type of weapon used in crime |
| LOCATION | Crime Rates | Object | Street address of crime incident rounded to the nearest hundred block. |
| Year | Crime Rates | Integer | Year of when the crime was committed. |
| Month | Crime Rates | Integer | Month of when the crime was committed. |
| American Indian and Alaska Native (Proportion) | Population | Float | Specific demographic group |
| Asian (Proportion) | Population | Float | Specific demographic group |
| African American (Proportion) | Population | Float | Specific demographic group |
| Native Hawaiian and Other Pacific Isander (Proportion) | Population | Float | Specific demographic group |
| child\_victim | Crime Rates | Integer | Indicates if victim of crime is younger than 18 years old. 1 – Child 0 – Not Child |

Analysis

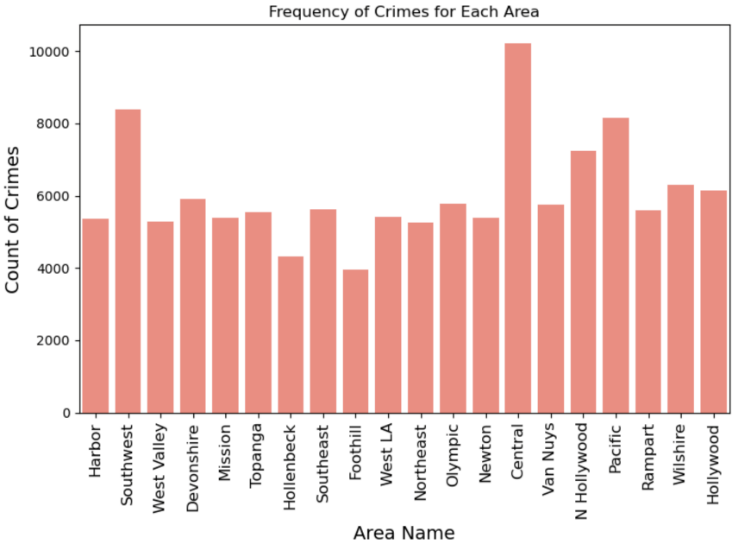
*Sociodemographic Patterns in High-Crime Areas*

The proposed analysis question I wanted to answer was ‘What is the predominant demographic group in the three Los Angeles communities with the highest recorded crime rates?’ For this analysis, I used crime data from the LA Crime dataset, focusing on the number of crime incidents (DR\_NO), and demographic data from the LA Population dataset. Given my limited familiarity with Los Angeles communities, I was unsure what to expect from the analysis and visualizations. However, based on background research, I hypothesized that the three high-crime communities will have varied demographic compositions, with no distinct group consistently dominating, and that demographic factors may contribute differently to crime rates in each area. To gain an initial understanding of crime distribution across Los Angeles communities, I examined the frequency of reported incidents to discover if there were any outliers with low or high crime rates. I created a histogram with communities on the x-axis and the count of reported crimes (DR\_NO) was along the y-axis. The results revealed that Central, Southwest, and Pacific were the communities with the highest crime frequencies, with incident records of 10,217, 8,379, and 8,165 incidents. The mean number of reported crimes in the top three high-crime communities was 8,920 incidents, with a standard deviation of 921. This highlights that crime is not evenly distributed, with few areas experiencing higher crime rates. After viewing the total crime counts, I wanted to look at the percentages to see how evenly crime was spread across the top three communities.

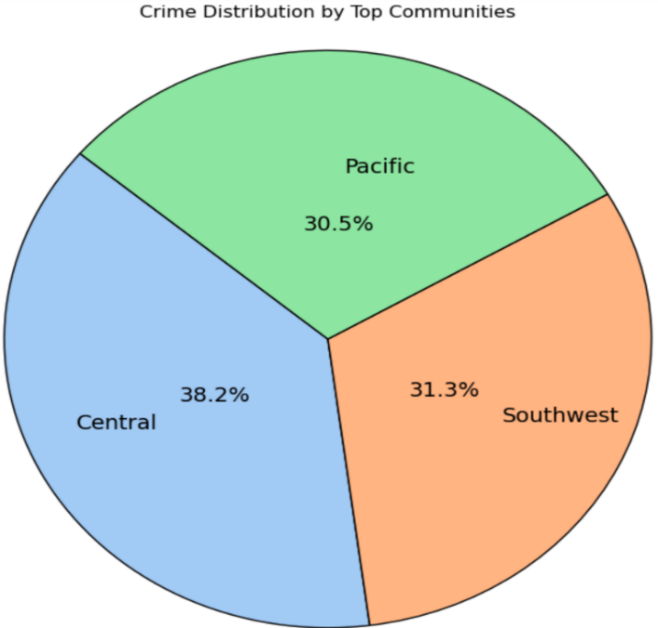
After identifying the communities with the highest crime rates, I wanted to examine how their demographics compared I used the FacetGrid bar chart to visualize the demographic distribution in each community, allowing for easy comparison between the three areas. By reshaping the demographic data using the melt function, I ensured that each demographic group was represented as a proportion across all areas, making it easier to compare relative distributions.

The results showed notable differences in demographic compositions across these communities. Southwest shows a significantly higher proportion of Hispanic or Latino residents, while Central has a more even distribution, with White and Hispanic or Latino populations being the most prominent. In contrast, Pacific is predominantly White. These variations in community demographics provide social context for interpreting patterns in crime data.

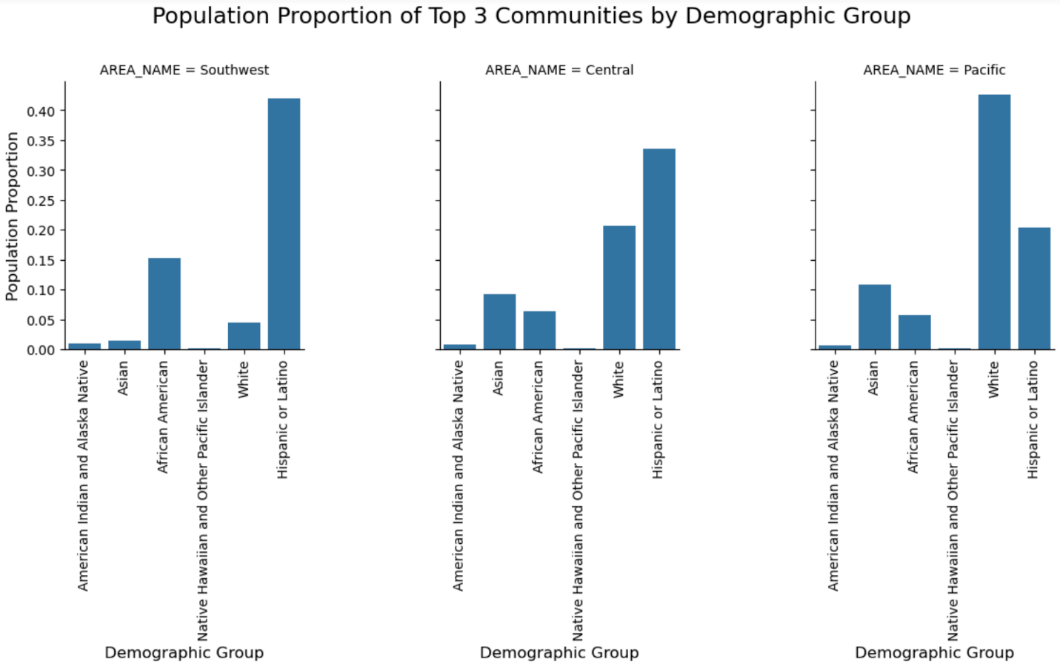
*Figure 1 Distribution of Crime Frequency*



*Figure 2 Percentage Distribution of Top Three Communities*



*Figure 3 Demographic Proportions in High-Crime Communities*



*Crime Rate Fluctuations in Los Angeles Communities During 2024*

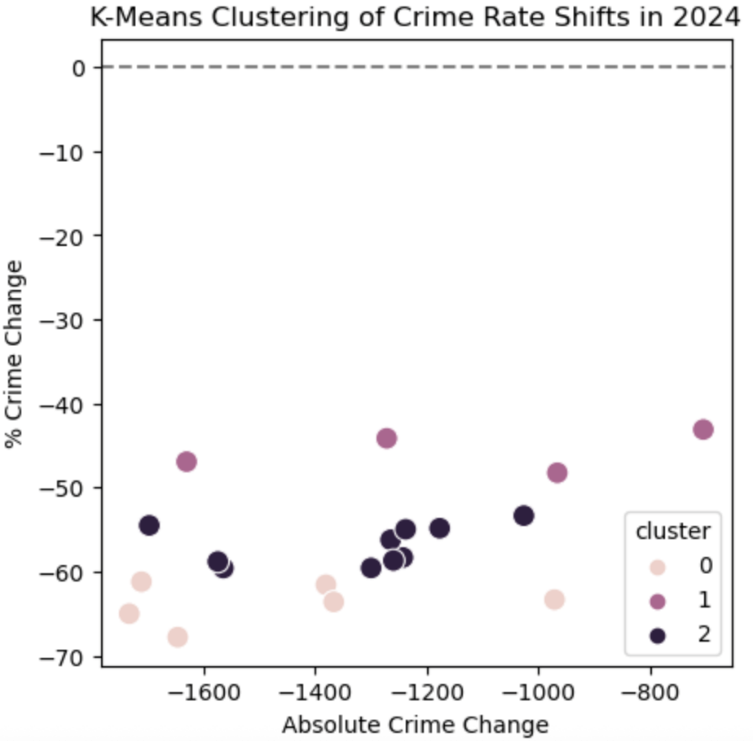
Since the distribution of crime incident reports in each area have been found, the following proposed analysis question was ‘How have crime rates across neighborhoods in Los Angeles changed between January - March 2024 and October - December 2024?’ To explore this question, I hypothesized that crime rates would be consistently low during the winter months at both the beginning and end of the year. This hypothesis assumed that colder and harsher winter conditions would limit outdoor activity, reducing the opportunities for crimes being committed.

I began by filtering the dataset to only include crime incidents from 2024. Then, I divided the data into two periods with January - March representing the beginning of the year and October - December representing the end. For each neighborhood, I calculated the total crime reports in each period to then compute both the absolute and percentage change. This allowed for a comparison of crime trends between the two periods across Los Angeles neighborhoods.

To evaluate whether the difference in crime counts followed a normal distribution, I applied the Shapiro-Wilk test. The test yielded a statistic of 0.9437 and a p-value of 0.2812. Since the p value exceeds the threshold of 0.05, we fail to reject the null hypothesis, confirming that the differences in crime rates were normally distributed. The lack of statistical significance implies that the changes in crime rates were not substantial enough to support a strong seasonal pattern across neighborhoods.

To further investigate the changes in crime rates across neighborhoods, I applied K means based on percent change. The algorithm grouped neighborhoods into three distinct groups representing increase, decrease, or slight change in crime, assigning each based-on similarity in percent changes. The clustering approach identified trends in crime rate changes across Los Angeles neighborhoods in 2024. To visualize this, I created a scatter plot with the absolute crime change on the x-axis and percent change on the y-axis. Each point represents a neighborhood and is colored by its assigned cluster, and a horizontal line at zero percent to highlight neighborhoods with no notable change. Importantly, no neighborhood experienced an increase in crime throughout the duration of 2024. All changes showed declines, with percentage decreases ranging from 40 to 70. This indicates an overall decrease in reported crime from January to March and October to December. Therefore, the clusters now represent a significant decrease, moderate decrease, or minimal decrease effectively illustrating the broad and consistent decline across all neighborhoods.

*Figure 5 Crime Trends Scatter Plot*



*Child Victimization Patterns Across Los Angeles Communities*

The final analysis addressed the question, ‘Are there specific communities where crimes involving child victims are more prevalent, and what are the annual trends?’ Based on earlier findings about demographic composition and overall crime rates, the analysis focused on the impact crime has on minors. I hypothesized that crime counts to be higher at the beginning of the year, based on the observed annual declines in neighborhood crime rates. Additionally, I expected the neighborhoods with the highest child victims counts to correspond with neighborhoods that experience the most crimes. However, I did not have a strong hypothesis about the descent of child victims due to the demographic diversity in Los Angeles.

To begin, missing victim age data was imputed using the median value for consistency. I filtered the dataset to isolate crimes reported involving minors and grouped it by area and month. For the times series analysis, I extracted the year and month from the crime reported dates and split the data into 2024 and 2025 to examine trends separately. This allowed me to create line charts with monthly trends for each year. From the results in the 2024 graph, monthly crime rates against children were notably high between February and May. After that period, there was a significant decline, with less than 10 child victimization reports per neighborhood in each month. In 2025, the rates continued at similar level, indicating a steady decrease in crimes involving minors.

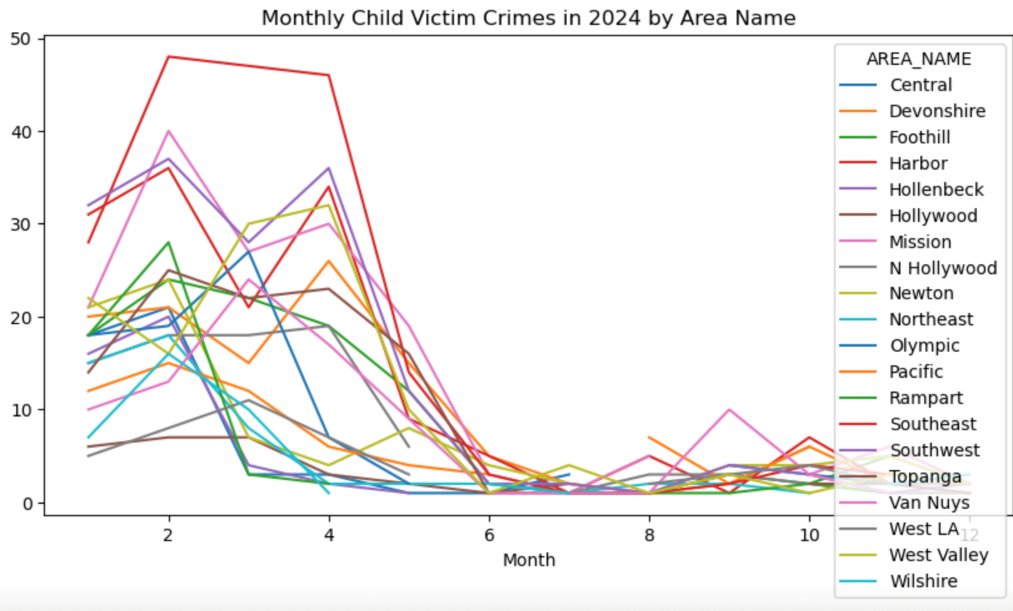
To further explore the impact of time of year and neighborhood on crimes committed against children. I created a binary variable for victim age, where 1 represents victims under 18 and 0 represents victims over 18, and one-hot encoded both AREA\_NAME and Month to use in a Random Forest Classifier. This model was trained with an 80/20 split and a balanced class weight to adjust the model to focus on the minority class, child victims. The Random Forest model achieved an accuracy score of 0.6283. The classification report demonstrated robust performance in identifying adult victims, with an F1 score of 0.77, but struggled to correctly predict child victims, with a significantly lower F1 score of 0.06. While recall for child victims was relatively high at 0.73, precision was only 0.03, indicating many predictions for child victims were incorrect. Despite using class balance, the results suggest that this model had difficulty predicting child victims due to the imbalance between crimes against adults and children.

To visualize the distribution of child victims by location and descent, I created three graphs. The first vertical bar chart presents the number of child victims across neighborhoods. The results displayed the communities with the highest count, Southeast reported 200 cases, Mission 168, and Southwest 160. These results contradicted by original hypothesis since these neighborhoods did not correspond to the ones with the highest overall crime rates. Southwest was the only neighborhood that appeared in both overall crime and child victimization analyses. Therefore, indicating that crimes reported against minors are prevalent in specific communities.

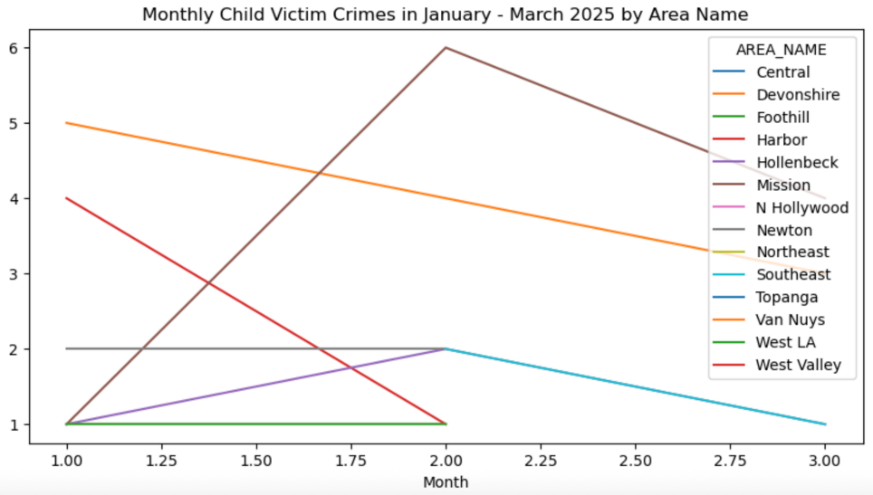
The second bar chart illustrates the number of child victims by descent. It reveals that Hispanic/Latin/Mexican are the most affected, with 1,262 reported cases, followed by African American children with 326. These results highlight a significant disparity in victimization among different demographic groups.

To combine both geographic and demographic results, I created a FacetGrid. This visualization shows which descents are most impacted within the neighborhoods with the most child victimization reports. Based on previous results, I expected moderate diversity, but the results revealed that children of Hispanic/Latin/Mexican are severely impacted across all three areas, African American children were next most affected in the Southwest, and Asian children were more affected in Southeast. Collectively, these visualizations reveal distinct patterns in child victimization across Los Angeles, emphasizing the increased vulnerability in certain communities and ethnic groups.

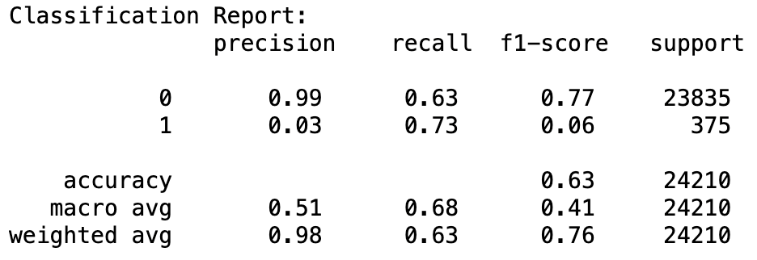
*Figure 6 Child Victim Crime Trends by Month in 2024*



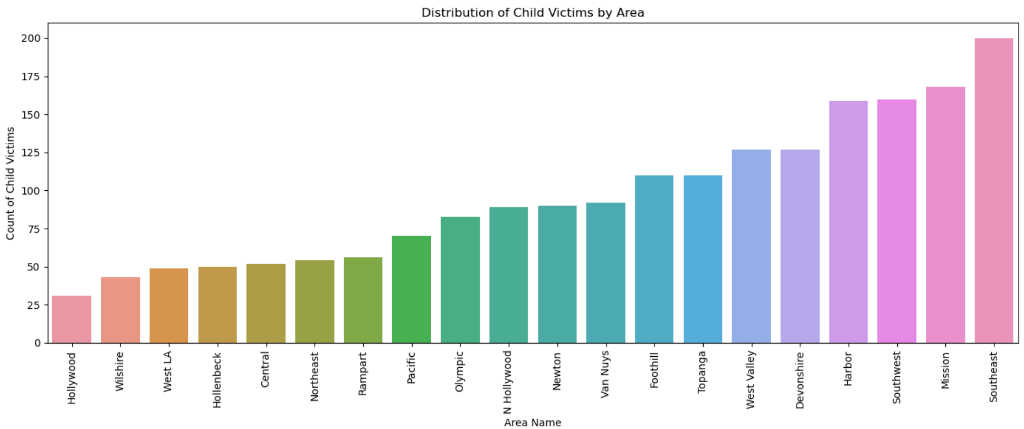
*Figure 7 Child Victim Crime Trends by Month in 2024*



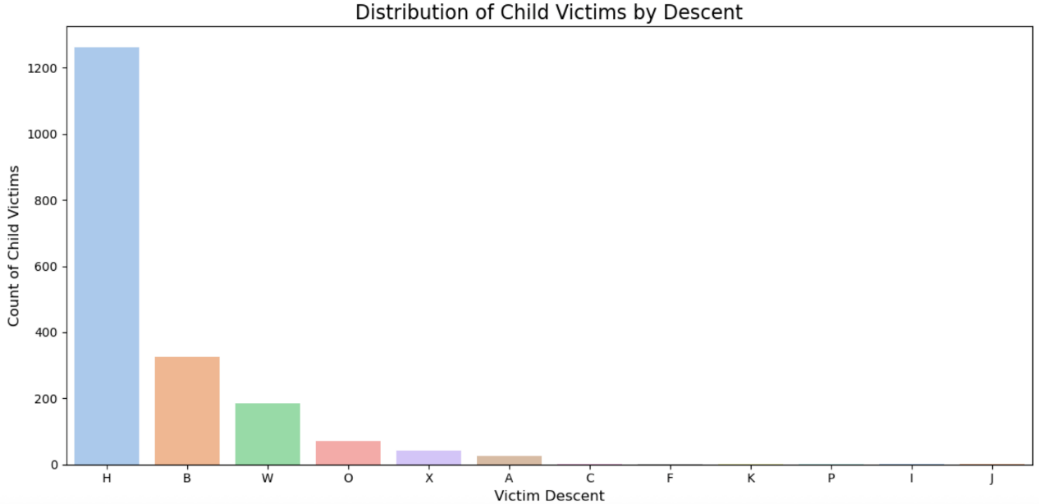
*Figure 8 Classification Report*



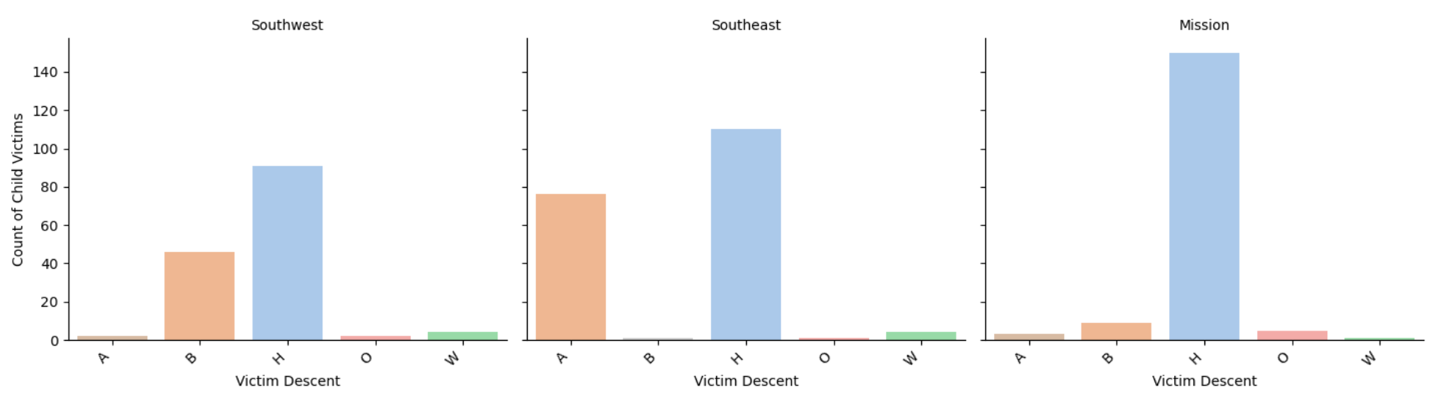
*Figure 9 Area Distribution of Child Victimization*



*Figure 10 Distribution of Child Victim Descent*



*Figure 11 Child Victim Descent in Top Communities*



Conclusion

This analysis explored how crime patterns in Los Angeles vary across communities, demographic groups, and time, using data from 2024 and 2025. Across the proposed questions, the results emphasized that crime is unevenly distributed, with certain communities and demographic groups are disproportionately impacted. Areas with the highest crime rates, which include Central, Southwest, and Pacific, each had distinct demographic compositions, indicating that there is not a single dominant group residing in these neighborhoods. Citywide crime rates in 2024 experienced a gradual decline, which contradicted the hypothesis of seasonal changes. The analysis of child victimization revealed that the neighborhoods with the highest crimes rates differed from those most affected by crimes involving minors. Importantly, Hispanic/Latin/Mexican children were significantly overrepresented among victims, revealing a clear disparity. Overall, these finding demonstrate the complexity and frequency of crime rates across Los Angeles, emphasizing the need to address the disparities and gain a better understanding of the underlying factors contributing to these crimes.

Through the project, I encountered limitations that affected the extent and comprehensiveness of the analysis. An initial challenge was addressing missing values in critical fields, impacting the dataset’s quality. Additionally, geographic identifiers in the datasets did not align, which made it difficult to accurately match police areas and neighborhood population data. Another issue was infrequent updating of public crime records, leading to inconsistent and unrealistic comparisons between 2024 and 2025. Finally, the machine learning model performed poorly when predicting child victimization due to the limited available features and imbalance in the dataset.

If conducting a future analysis, there multiple improvements that can be made to enhance the complexity and accuracy. One change that would benefit is obtaining more granular and consistent geographic identifiers, improved data on victim demographics, and additional datasets that include more features such as socioeconomic indicators. Furthermore, using more advanced machine learning methods that take class imbalance into consideration could improve understanding as to why certain demographic groups are disproportionally affected. Altogether, continued analysis on crime patterns can assist in the advancement of public safety and help those more vulnerable to crime.