DAS732: Data Visualization Assignment 3

Subham Agarwala
IMT2022110
IIIT Bangalore
Bangalore, India
Subham.Agarwal@iiitb.ac.in

Sarvesh Kumar
IMT2022521
IIIT Bangalore
Bangalore, India
SarveshKumar.A@iiitb.ac.in

Ayush Gupta
IMT2022546
IIIT Bangalore
Bangalore, India
Ayush.Gupta546@iiitb.ac.in

I. Introduction

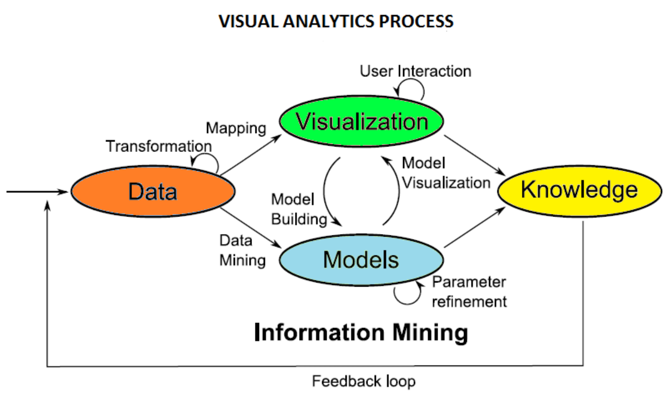


Fig. 1. Visual Analytics Workflow Process outline

Visual analytics is a multidisciplinary field that combines data analysis and visualization to facilitate human decision-making and understanding of complex datasets. At its core, a visual analytics workflow involves the seamless integration of data preprocessing, visualization, and interaction, allowing users to derive insights and iteratively refine their analysis. The workflow often incorporates data transformations, feedback loops, and computational models to create a dynamic and adaptive process for exploring data.

In this assignment, we explore the principles of the visual analytics workflow and demonstrate its application in solving real-world problems. The assignment emphasizes the importance of interactivity and iterative refinement by incorporating feedback loops that guide the analysis. Data transformation techniques are applied to preprocess and optimize raw datasets, ensuring they are suitable for downstream visualizations and modeling.

Additionally, the assignment includes the integration of machine learning models into the workflow, showcasing how predictive and prescriptive analytics can augment human interpretation. By embedding machine learning components, we leverage computational power to uncover patterns, anomalies, and trends that might otherwise go unnoticed.

This assignment can be considered a continuation to Assignment 1 (report attached in appendix) the results of which have been referred to and used in this assignment. To recall, our dataset for Assignment 1 was a crime dataset for the city of LA. The same dataset has been used along with supplementary dataset as and when necessary in the different workflows.

The sections in the assignment report have been divided into the different workflows curated by us. The workflows are:

- **Workflow 1:** Correlation of arrest data with crime data and criminal demographics.
- **Workflow 2**: Cluster Analysis and Visualization of Crime Patterns in Los Angeles
- **Workflow 3**: Time Series Analysis and Forecasting of Crime Trends in Los Angeles
- **Workflow 4**: Looking for correlation between demographics of victims and which area they belong to.
- **Workflow 5**: Analyzing affect of Pandemic on Crimes against various Demographics

II. Workflows

A. ***Workflow 1:*** *Correlation of arrest data with crime data and criminal demographics.*

In Assignment 1, we explored the crime data of Los Angeles with a focus on the temporal and spatial distribution of incidents. This initial analysis served as the foundational step for this workflow, providing crucial insights into the patterns of crime across time and space. Before diving into the deeper analysis for this section, here is some notable information we managed to gather from our visualizations in Assignment 1.

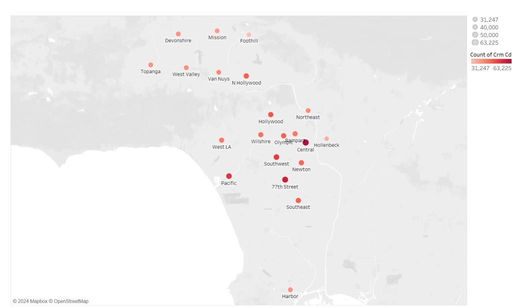


Fig. 2. Crime count in LA (2020-2024)

Fig. 2 depicted the spatial concentration of crimes, highlighting regions with the highest frequency of incidents. Certain regions consistently had higher crime counts, marking them as potential hotspots for targeted interventions.

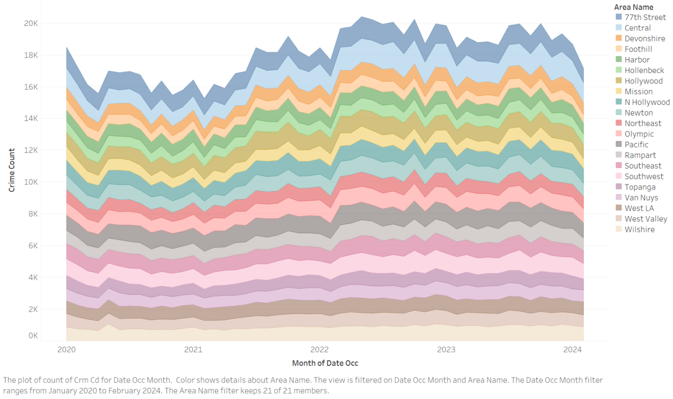


Fig. 3. Crime rate (2020-2024) across different areas of LA

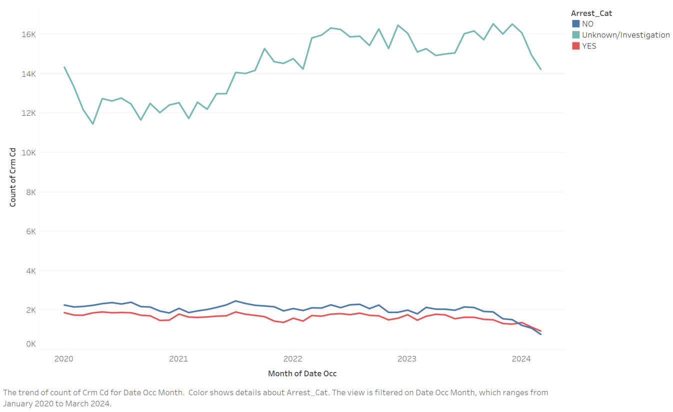


Fig. 4. Criminal arrest data extracted from status column.

Fig. 3 highlighted changes in crime rates across various regions over time. Among all areas, the Central Region showed the most significant fluctuations in crime rates, suggesting unique dynamics or challenges in this area.

The other visualizations provided insights into the demographics or crime victims, including gender, age, race etc.

While the victim data was rich, there was limited information available on the perpetrators of these crimes, making it challenging to draw comprehensive conclusions about criminal behavior. The dataset primarily focused on victims, with minimal data on criminals. The only indirect information available was derived from the Status column, which included labels like Juvenile Arrest or Adult Arrest, allowing us to infer the age category of the perpetrators.

Understanding crime patterns requires not just victim data but also insights into the criminals. This can help identify specific groups or demographics involved in criminal activities, uncover motivations, and analyze how different crimes are distributed across the city.

The objective of this section is to enhance the insights derived from Assignment 1 by incorporating additional datasets into the workflow. Specifically, we have introduced the arrest data of Los Angeles, which provides valuable information about the perpetrators of crimes. This dataset includes attributes such as the gender, race, age, and reason for arrest of the criminal.

By adding this new dataset we can broaden our understanding of criminal profiles, identify common offender groups and analyze crime patterns.

Since the data for arrest was available only from 2010 to 2019, we have extended the time period of the crime dataset too and will be working with the same for this section.

Run 2:

Considering the Assignment 1 as the first iteration in the workflow we use the knowledge gained from the previous visualizations and find a gap in knowledge about the criminals, thus using the additional dataset we first visualize the count of logged arrests in the city of LA on a geographical map along with the logged crimes in the same regions.

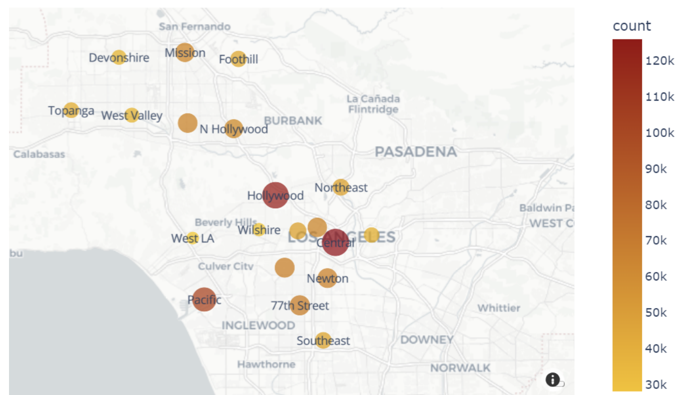


Fig. 5. Arrest count in different regions of LA (2010-2019)

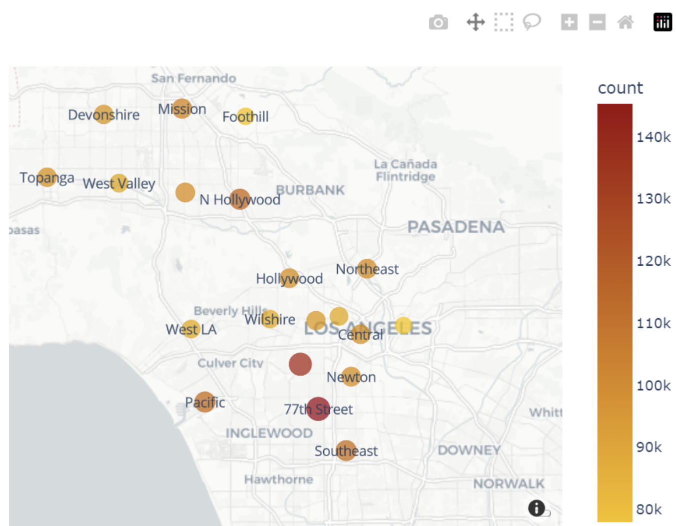


Fig. 6. Crime count in different regions of LA (2010-2019)

Inference 1: Previously, we analyzed crime data from 2020 to 2024, where the Central region of Los Angeles emerged as the largest crime hotspot, showing a sharp increase in crime rates over these 4-5 years. In contrast, when we now examine the data from 2010 to 2019, the Central region was not the leading hotspot as seen in Fig. 6, indicating the rapid rise in crime that occurred more recently. However, the major crime hotspots during both periods remained relatively consistent, including regions like Central, Southwest, 77th Street, and Pacific.

Inference 2: Although the Central region was not the largest crime hotspot among the major areas during the 2010-2019 period, it still recorded the highest number of arrests, followed by Hollywood and the Pacific region as seen in Fig. 5. This suggests that while the frequency of crimes in Central may not have been the highest, it could indicate a higher level of law enforcement activity or more proactive policing in that area. Additionally, the high number of arrests in Hollywood and Pacific might reflect the concentrated efforts of law enforcement in areas known for high foot traffic or more visible crime, such as tourist hotspots and commercial districts. These trends could also point to targeted policing strategies or the types of crimes being committed in each region, such as those with higher arrest rates for certain offenses.

This provides scope to verify whether the reasons for arrests in these regions were primarily the ones for which there is targeted policing.

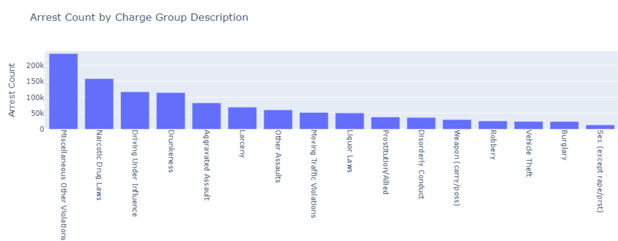


Fig. 7. Reasons for arrest that are most common in LA.

Run 3:

Fig. 7 shows that apart from the miscellaneous causes, violation of Narcotic Drug Laws, Driving under influence, Drunkenness, Aggravated Assault, Larceny, Liquor Laws etc are the most common reasons for arrest in the city of LA. This can be due to more the frequency of these crimes being more or due to the law department being more vigilant about these crimes. Assuming the latter to be true we will plot arrest data for the major arrest hotspots to check if the arrests in those regions were majorly for the reasons mentioned above.

We can see in Fig. 8 that violation of Narcotic Drug Laws (ignoring miscellaneous charges) is one of the major reasons among others like drunkenness, driving under influence, Violtion of liquor laws etc.

Inference: These are also the most common reasons of arrest in the whole of LA which tells us that regions like Central, Hollywood etc have the higher number of arrests because they exhibit crimes that are more prone to attract police attention. The lifestyle in these areas plays a pivotal role in this trend. Central Los Angeles, known for its dense

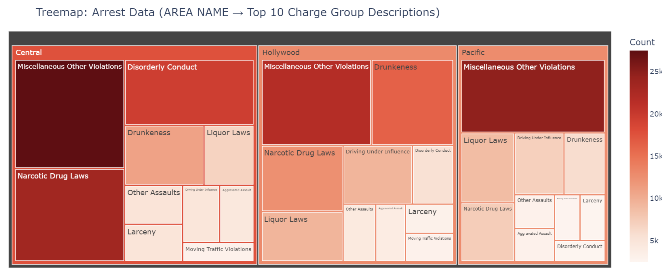


Fig. 8. Reason of arrest most common in major hotspots.

population and a combination of residential, commercial, and entertainment spaces, sees higher rates of crimes like drug violations and public intoxication. The area's bustling atmosphere, coupled with a diverse population, often results in greater visibility of such offenses. Similarly, Hollywood, a global entertainment hub, attracts a steady flow of tourists and individuals involved in nightlife, which can lead to an increased presence of alcohol-related offenses and drug use. The visibility and frequency of these behaviors make them more likely to catch the attention of law enforcement.

These cities' active nightlife and social environments create opportunities for offenses like drunkenness, driving under the influence, and narcotic violations, leading to a higher number of arrests.

Another interesting observation we can make is in the rate of increase or decrease in arrests along with crime across the years. For this we have made a superimposed line graph showing the trend of crimes and arrests (Fig. 9).

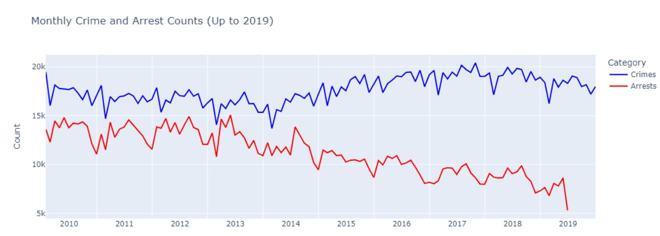


Fig. 9. Crime and Arrest trend across the years in LA.

We see that with a consistently non-decreasing crime rate, the rate of arrests in LA seem to decrease over the years, even for the years we see an increase in the crime rate.

Inference: This could mean that police are focusing on different approaches, like targeting more serious crimes or using other methods instead of making arrests. It might also suggest that there are not enough police resources to handle the growing number of crimes, or that certain offenses are no longer being arrested for as much due to changes in laws or policies. It could also indicate degrading law enforcement in the city. Essentially, the decrease in arrests, even as crime rises, points to shifts in how law enforcement is responding to crime.

Run 4:

In Assignment 1 we had made use of the ”Status” column in the crime dataset which had values like ”Juvenile Arrest” and ”Adult Arrest” to estimate arrest count in the time period (2020-2024) which can be seen in Fig. 4. Now that we have arrest data we can validate the assumption by checking if the numbers add up. Since we have arrest data only until 2019 we have used the Prophet model from Facebook, trained it on available data and used it to forecast the arrest rate for the years 2020-2024. We plotted the forecasted data alongside the actual data in Fig. 10 to make observations.

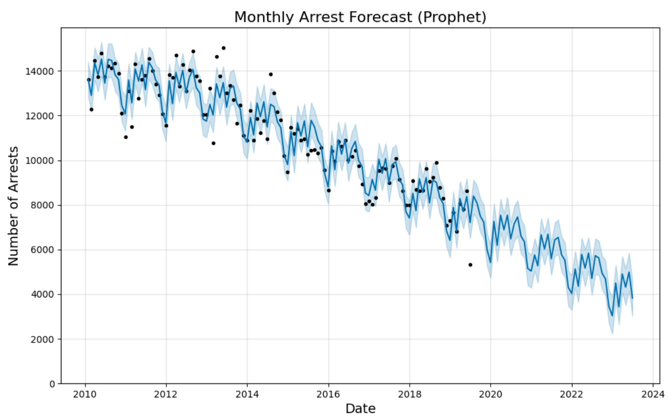


Fig. 10. Monthly forecast of arrest data in LA.

We can see that the forecasted data is more or less acceptable for the time period we trained it on, and predicts the data from 2020 to 2024 based on the decreasing trend seen earlier. We see that the arrest count starts at around 6000-7000 in 2020 and decreases to about 4000 by 2024 which makes an average of about 5000 monthly arrests. According to Fig. 4, we see that our assumption gave a count of about 2000 monthly arrests, but the status of about 14000 cases was unknown.

This discrepancy suggests that the reliance on the ”Status” column alone might have led to an underestimation of arrests. By using the Prophet model to account for the overall trend, we can see that our initial estimate likely overlooked a substantial portion of arrests that were either not categorized or mislabeled.

Run 5:

Having examined the major regions with higher arrest counts and the charges associated with those arrests, it is equally important to analyze the demographics of the individuals involved in these crimes. Understanding the characteristics of the people behind the offenses can help uncover patterns and provide insights into whether certain groups are disproportionately responsible for specific crimes or categories of offenses. This analysis could reveal valuable trends that might inform targeted interventions or policy decisions.

To better understand the demographics of individuals involved in crimes, we created a Sankey diagram (Fig. 11) that illustrates the flow of data from Gender to Age Groups to Descent. This visualization helps us analyze the relationships and proportions across these categories.

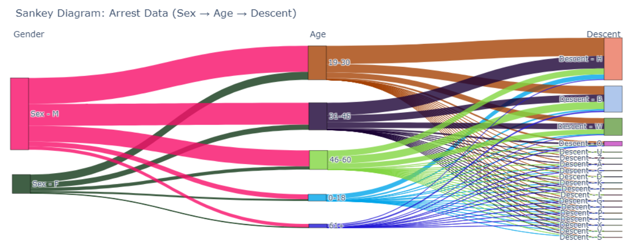


Fig. 11. Sankey diagram showing flow of arrest data (Gender,Age Groups,Descent)

Inference:

The analysis reveals that the majority of individuals arrested are male, with significantly fewer arrests of females. Among the age groups, the largest share of offenders falls within the ranges of 19-30 years, followed by 31-45 years and 46-60 years, indicating that crimes are predominantly committed by young to middle-aged adults. However, there is also a noticeable presence of juvenile offenders, emphasizing the need to address criminal behavior among minors.

In terms of descent, individuals of Hispanic/Latin/Mexican, Black, and White backgrounds constitute the largest proportions of law offenders. For men, the 19-30 age group accounts for the highest number of criminals, followed by 31-45 and 46-60 age groups. A similar trend is observed among women, although the overall number of female offenders is smaller.

When looking across all age groups, offenders from the Hispanic/Latin/Mexican descent are the most prevalent, followed by those of Black descent and then White descent. This trend holds true regardless of gender or age, suggesting that these groups are disproportionately represented in arrest records.

To delve deeper into the relationship between gender and other demographic categories, we utilized two UpSet plots, focusing separately on the Male and Female categories.

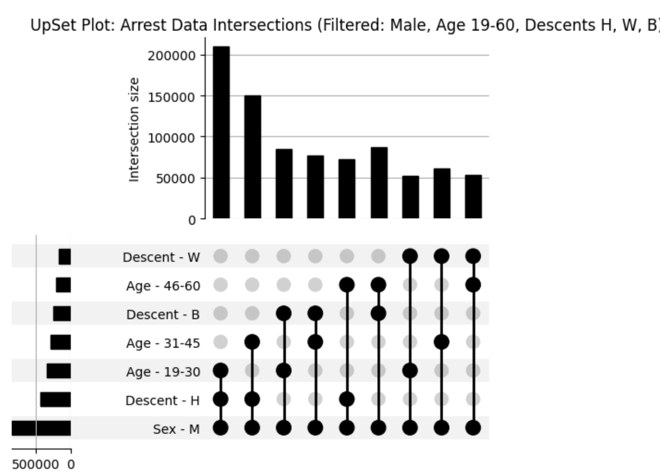


Fig. 12. UpSet plot: Male

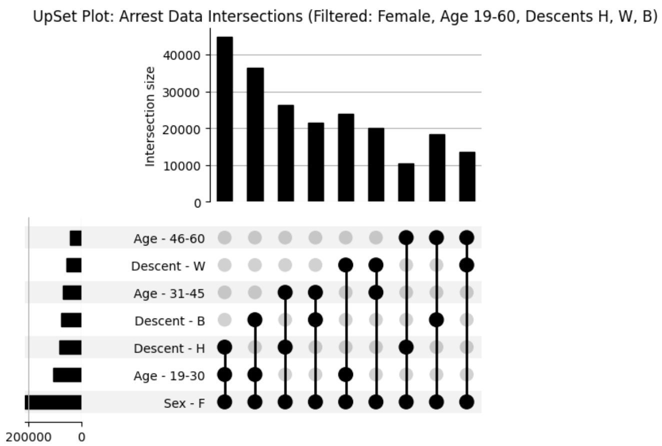


Fig. 13. UpSet plot: Female

For the male category, the intersection with the highest number of arrests involves males aged 19-30 from a Hispanic/Latin/Mexican background. This indicates that young Hispanic/Latin/Mexican males are disproportionately represented in arrest records.

Among Black males, the age group with the highest number of arrests is the 41-60 age group, highlighting a distinct trend in arrests for this demographic compared to others. For White males, the age group most represented in arrests is the 31-45 age group, showing a variation in the age distribution of offenders based on descent.

For the female category, the intersection with the highest number of arrests mirrors that of males, comprising females aged 19-30 from a Hispanic/Latin/Mexican background. This suggests that this demographic group is consistently prominent in arrest records, regardless of gender.

However, unlike males, a noticeable spike is observed among White women in the 19-30 age group, indicating a unique trend for this demographic. This could point to specific factors or circumstances influencing the arrest patterns of young White women compared to other groups.

For other intersections, the trends among females are largely similar to those observed in males, albeit with generally lower overall counts. These patterns highlight both shared and distinct factors affecting arrest demographics across genders and highlight the importance of considering gender-specific dynamics in crime analysis and prevention strategies.

The information gathered so far leaves scope for further analysis to be done on the reason for arrest for different demographic categories. We have made 3 treemps for the same, one each for Age group, Gender and Descent.

Starting with the treemap corresponding to age groups as seen in Fig. 14, we observe distinct patterns in the reasons for arrests across various age demographics:

- For individuals in the 19-30 and 31-45 age groups, the most common reason for arrest is the violation of Narcotic Drug Laws, followed by driving under the influence, aggravated assault, and drunkenness. Larceny and moving traffic violations also feature prominently in these groups, indicating a mix of substance-related offenses and property crimes.
- In the 46-60 and 61+ age groups, the most frequent reason for arrest shifts to drunkenness, followed closely by violation of Narcotic Drug Laws and driving under the influence. This highlights a trend of substance-related offenses persisting into older age groups but with a decreased emphasis on other forms of crime.
- The 0-18 age group shows a distinctly different pattern compared to the adult groups. The most common reason for arrest is pre-delinquency, followed by larceny and non-criminal detention. This suggests that younger individuals are more often involved in crimes related to early-stage criminal behavior, such as theft and being detained for minor offenses. Other notable offenses include assault, robbery, weapon possession or carrying, narcotic drug law violations, burglary, and vehicle theft. These findings indicate that while younger offenders may engage in a range of property and violent crimes, many of the arrests could be for less serious offenses like theft or minor disturbances, with some also involved in more serious crimes like robbery and weapon possession.

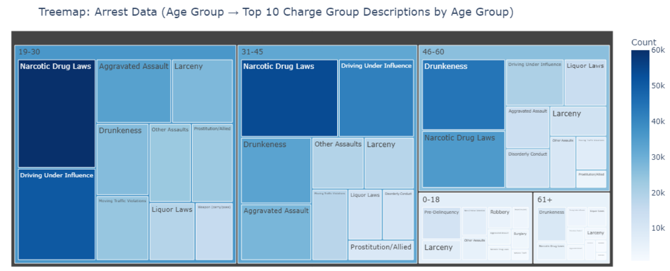


Fig. 14. Treemap showing proportion of arrest charges among different age groups

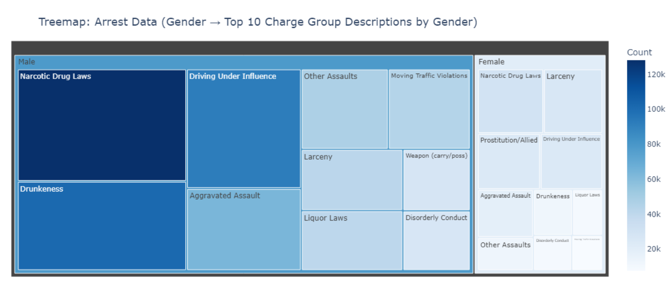


Fig. 15. Treemap showing proportion of arrest charges among male and female categories.

Moving on the treemap that corresponds to gender, labeled as Fig. 15.

Among male offenders, the largest proportion of arrests is related to violations of Narcotic Drug Laws, highlighting the significant role of substance abuse in male criminal activity. This is followed by arrests for drunkenness and driving under the influence, indicating a strong association between alcohol and drug-related offenses. Aggravated assault ranks next, showing that violent crime is also prevalent, albeit less frequent than substance-related crimes. Other notable offenses include moving traffic violations, larceny, liquor law violations, and weapon-related crimes. These trends suggest that males are more likely to engage in both substance abuse-related offenses and violent crimes, with property crimes and traffic violations also being common.

For female offenders, narcotic drug law violations are again the leading cause of arrests, similar to males, highlighting the role of substance abuse among women. However, unlike the male trend, larceny emerges as the second most common offense, suggesting that theft-related crimes are more prominent among females. Prostitution is also notably high among female offenders, reflecting a significant societal issue. Driving under the influence, assault, drunkenness, and liquor law violations follow as less frequent causes for arrests. These patterns suggest that while substance-related offenses are common for both genders, females tend to be more involved in property crimes like larceny and engage more frequently in prostitution compared to males.

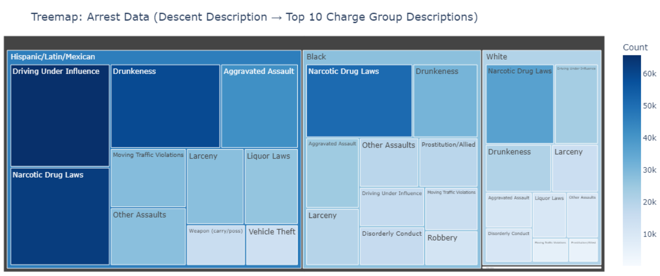


Fig. 16. Treemap showing proportion of arrest charges among people of different Descent

Lastly we have the treemap that corresponds to Descent (considering only Hispanic/Latin/Mexican, Black and White), shown in Fig. 16.

- For the Hispanic/Latin/Mexican descent group, the primary reasons for arrest are Driving Under the Influence (DUI) and Violation of Narcotic Drug Laws. These charges are significantly more common than others. Following these, offenses such as Drunkenness and Assault are also noteworthy, contributing substantially to arrest counts. Additionally, crimes like Moving Traffic Violations and Larceny are prevalent but not as dominant as the top charges. The trend highlights a significant proportion of arrests linked to substance-related offenses, particularly alcohol and drugs, with traffic-related offenses and property crimes being notable secondary concerns.
- Among Black individuals, the most common charge is Violation of Narcotic Drug Laws, which is closely followed by Drunkenness. These two categories make up a large portion of arrests. Assault and Larceny also contribute to the arrest statistics, though to a lesser extent. Other offenses like Prostitution, Driving Under Influence, and Robbery are observed but have a lower occurrence compared to the more frequent narcotics and alcohol-related offenses. The arrest pattern for Black individuals reveals a heavy association with substance use, particularly drugs and alcohol, with violent offenses such as assault and property crimes like larceny being common but secondary. Prostitution is a more more prominent reason among these individuals compared to people from other descents.
- For individuals of White descent, the leading charge is also Violation of Narcotic Drug Laws, followed by Driving Under the Influence and Drunkenness, reflecting a similar trend to the other groups. Larceny and Assault are next in frequency, suggesting a combination of substance-related offenses and property crimes. Liquor Laws Violations also appear among arrests, though less prominently than the other charge types. Prostitution is a charge seen in this group but at a lower rate compared to other offenses. The overall pattern among White individuals shows that narcotics and alcohol-related offenses are the most common reasons for arrest, with property crimes and alcohol-related legal violations being secondary, and a relatively low occurrence of prostitution charges.

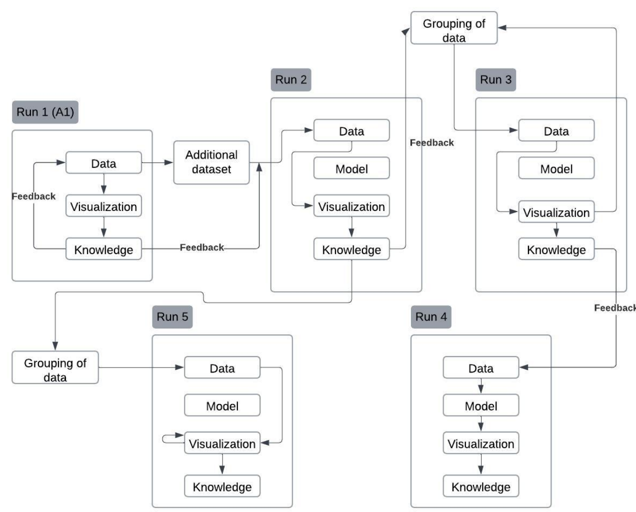


Fig. 17. Workflow diagram

Conclusion: The analysis of crime and arrest data in Los Angeles (2010-2024) highlights significant trends and disparities. Central LA stands out with the highest arrest rates, indicating focused law enforcement efforts in this area. However, rising crime rates across certain categories suggest ongoing challenges. Young males, particularly of Hispanic/Latin/Mexican descent, are disproportionately affected, with narcotics, DUIs, and assaults being common offenses. Women face unique arrest patterns, and juvenile arrests often involve pre-delinquent behavior. Substance-related offenses and regional disparities underline the need for targeted community programs, policy interventions, and resource optimization to address the city's evolving safety concerns.

That concludes the analysis for this section. The workflow diagram for this section can be seen in Fig. 17.

*B. **Workflow 2:** Cluster Analysis and Visualization of Crime Patterns in Los Angeles.*

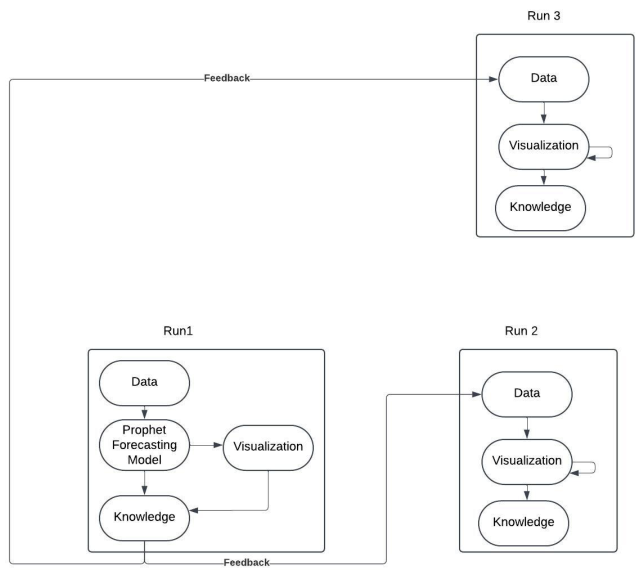


Fig. 18. Workflow Diagram

1) Introduction: This section delves into the analysis of crime patterns in Los Angeles between 2020 and 2023, employing clustering techniques and data visualization methods. The primary focus is to identify meaningful patterns and trends within the crime data by segmenting it into distinct clusters. These clusters are then analyzed to gain insights into crime distribution, victim demographics, and geographic trends. You can see the workflow diagram for this section in Fig. 18

2) Methodology: The analysis follows a structured workflow consisting of the following steps:

- Data Selection and Preprocessing
- Dimensionality Reduction using Principal Component Analysis (PCA)
- Determination of Optimal Number of Clusters
- Geospatial Visualization of Clusters
- Network Analysis through Node-Link Diagrams
- Focused Analysis on Selected Clusters

3) Data Selection and Preprocessing: The initial dataset comprises crime records in Los Angeles from 2020 to 2023. To facilitate meaningful analysis, a subset of relevant columns was chosen, including:

- Date of Occurrence of Crime (DATE OCC)
- Time of Occurrence of Crime (TIME OCC)
- Crime Description (Crm Cd Desc)
- Victim Age (Vict Age)
- Victim Sex (Vict Sex)
- Victim Descent (Vict Descent)
- Weapon Description (Weapon Desc)
- Area Name (AREA NAME)
- Premises Description (Premis Desc)
- Status of the Case Description (Investigation Continued, Juvenile Arrest, Adult Arrest etc) (Status Desc)
- Latitude (LAT)
- Longitude (LON)

Steps in Preprocessing:

- **Rationale for Column Selection:**
  These columns were selected as they encompass vital information regarding the timing and nature of crimes, victim demographics, weapons involved, and location-based details. Additionally, latitude and longitude provide geospatial data critical for clustering and spatial analysis.
- **Data Cleaning:**
  Missing or null values were removed to maintain data quality and integrity, ensuring the dataset is suitable for analysis without introducing bias.
- **Data Transformation:**
  Categorical variables, such as crime descriptions and victim characteristics, were converted into numerical format using one-hot encoding.

By following these preprocessing steps, the dataset was prepared to effectively support the subsequent stages of analysis and visualization.

4) Dimensionality Reduction using Principal Component Analysis (PCA): Due to the high dimensionality (600 features) resulting from one-hot encoding, PCA was employed to reduce the feature space while retaining significant variance:

- **Process:** PCA reduced the dimensions from 603 to 154 components.
- **Benefits:**
  - Computational Efficiency: Reduced computational load for clustering algorithms.
  - Noise Reduction: Mitigated the impact of redundant or less informative features.
  - Visualization: Facilitated visualization in lower-dimensional space.

Run 1:

5) Clustering and Determination of Optimal Clusters: To identify patterns within the crime data, clustering was performed:

- **Elbow Method:**
  - **Approach:** K-means clustering was applied with cluster numbers ranging from 1 to 50 as seen in Fig. 19.
  - **Results:**
    Determining the optimal number of clusters for analysis posed significant challenges. Despite experimenting with various values of k (up to 50 clusters), a clear elbow point was not observed in the elbow method plot. Furthermore, the computational demands of processing larger cluster counts were substantial.
    Given these constraints, $k = 21$ was chosen as the number of clusters, aligning with the 21 geographic

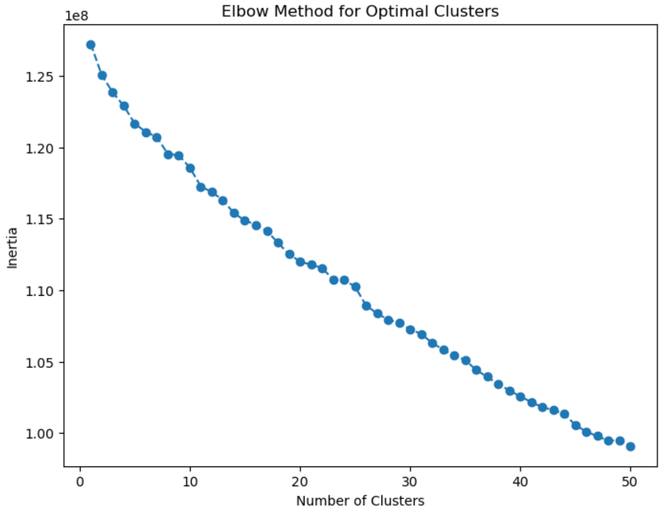


Fig. 19. Elbow Method

areas in Los Angeles. This decision is a reasonable approximation that allows us to explore the dataset meaningfully. However, it is important to note that the clustering model considered multiple features beyond location, leading to groupings influenced by other variables.

The model was fit to the dataset and every data point was assigned to a cluster from cluster 0 to cluster 20. You can see the size of the clusters in Fig. 20.

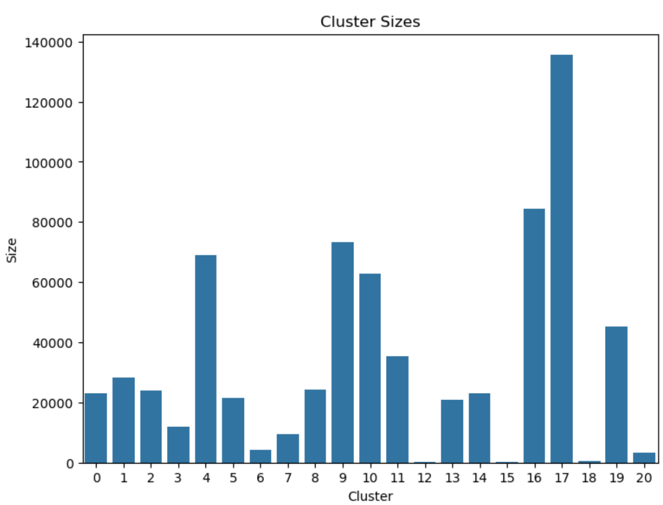


Fig. 20. Sizes of Clusters

6) Geospatial Visualization of Clusters:

- **Method:**
  - Mapped the latitude (`LAT`) and longitude (`LON`) coordinates.
  - Color-coded data points according to their cluster assignments.
- **Observations:**
  - **Cluster 12:** Notably concentrated at the southern tip of Los Angeles, appearing predominantly green on the map as seen in Fig. 21.
  - **Spatial Patterns:** While clusters did not strictly follow geographical locations, certain areas exhibited distinct clustering, indicating the possibility of localized crime patterns.

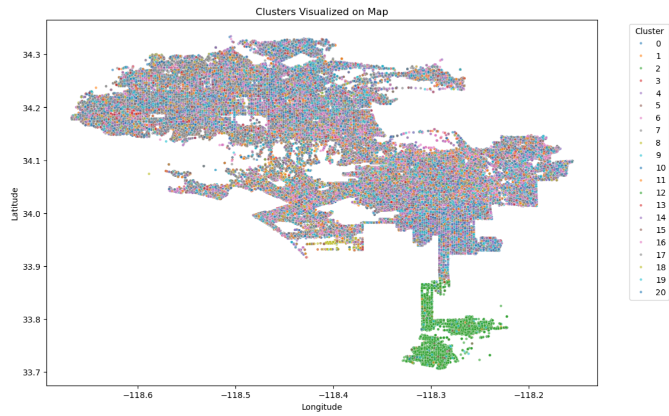


Fig. 21. Geospatial Map Visualization of Clusters

Run 2:

7) Cluster Analysis through Node-Link Diagrams, Geospatial Maps, Parallel Coordinate Plots, Parallel Sets and Upset Plots: To explore relationships within and between clusters:

- **Node-Link Diagrams:**
  - **All Clusters:** Initial node-link diagrams for all clusters resulted in overly complex visuals due to data density, therefore decided to focus on only few clusters.
  - **Selected Clusters:** Focused on clusters **12**, **15**, **18**.
    - **Cluster 12:** Chosen for its distinct geospatial concentration.
    - **Clusters 15:** Contain a manageable number of data points (hundreds), enabling effective visualization.
  - Analyzing clusters with larger sizes proved less insightful as they often lacked granularity and could have been further subdivided. Therefore, for this report, we focus on clusters 12 and 15, as they appear to have been grouped appropriately and provide meaningful insights. Importantly, these clusters are unlikely to change significantly with an increase in k, ensuring the robustness of the analysis.
- **Method:**
  - I used the NetworkX library to create an undirected graph G, where each node represents a crime incident, and edges represent proximity between incidents.
  - Applied k-nearest neighbors (k=4) within clusters to construct network connections. This provides enough connections to reveal meaningful patterns without overcomplicating the graph. Used the

NearestNeighbors class from scikit-learn to find the nearest neighbors for each node.

– Rendered the graph on an interactive OpenStreetMap, where nodes and edges are color-coded by crime type. The visualization allows users to zoom in and out, pan across the map, and explore specific areas in detail, offering a dynamic and user-friendly way to analyze the spatial relationships and clustering of crimes across the city. This interactivity enhances the ability to draw insights by providing a detailed and customizable view of the data.

Cluster 12:

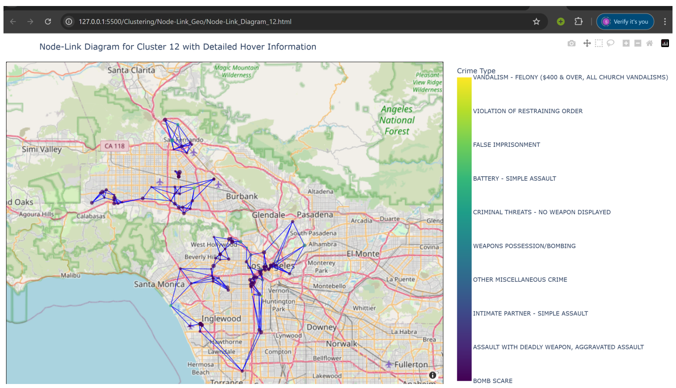


Fig. 22. Node link diagram of Cluster 12 on a Geospatial Map

- **Crime Type Distribution:**
  – Most of the crime types within Cluster 12 are categorized as "Bomb Scare," indicating a potential focus on public safety and heightened security concerns in this cluster. This prevalence might suggest the need for enhanced surveillance or rapid response protocols in areas identified in this cluster.
- **Sub Clusters Within Cluster:**
  – From the node-link diagram seen in Fig. 22, it is evident that there are 3 sub-clusters. Further analysis of crime patterns is needed to investigate these sub-clusters.
- **Proximity to Airports and Airstrips:**
  – A notable observation is the presence of some incidents near airports or airstrips. This proximity suggests a potential linkage to transportation-related security concerns, such as bomb threats targeting critical infrastructure. Such areas might require specialized safety measures and increased vigilance to prevent disruptions or security breaches.

To analyze this cluster further, we will use the following plots:

- **PCP** helps in understanding complex relationships across multiple variables.
- **Parallel Sets** provide a cleaner, flow-based visualization for analyzing proportions and dominant trends.
- **UpSet Plot** is perfect for identifying specific overlaps and quantifying combinations of variables.

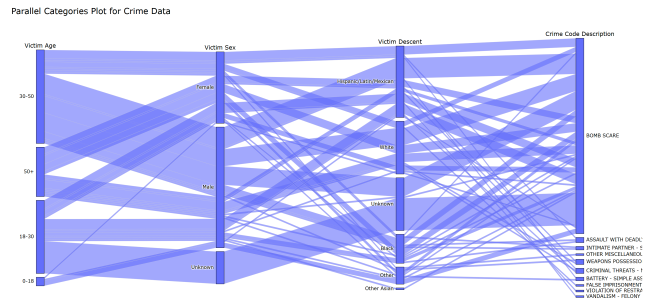


Fig. 23. PCP for Cluster 12

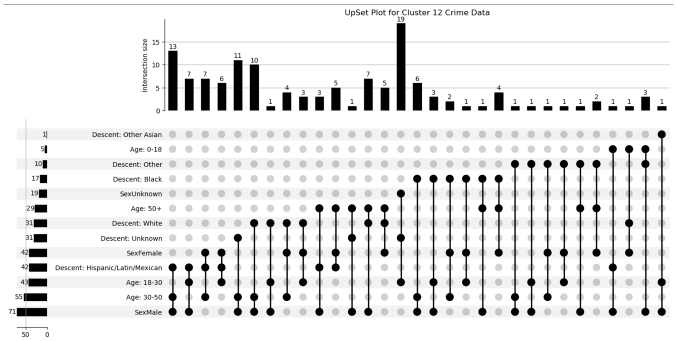


Fig. 24. Upset Plot for Cluster 12

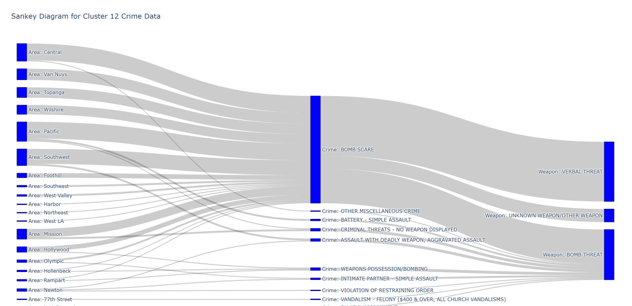


Fig. 25. Sankey Diagram for Cluster 12

Observations and Insights for Cluster 12:

- **Victim Demographics:**
  – Based on the UpSet Plot shown in Fig. 24 and Parallel Coordinates Plot (PCP) shown in Fig. 23, it is evident that male victims are predominantly affected in this cluster, accounting for 71 out of 132 data points. The most common victim age group is 30-50 years, and the Hispanic/Latin/Mexican descent emerges as the most affected victim descent category.
- **Weapon Usage:**
  – The Sankey Diagram Fig. 25 highlights that the primary weapons associated with crimes in this cluster are Verbal Threats and Bomb Threats, indicating a reliance on psychological intimidation or threats rather than physical harm.

- **Intersection Analysis:**
  - The UpSet Plot uniquely provides the exact number of intersections between various attributes, which is particularly useful for deeper analysis. For example, it enables the precise identification of overlapping subsets, such as the intersection of specific victim demographics, crime types, and weapon descriptions, facilitating further investigation of the cluster.

Run 3:
Cluster 15:

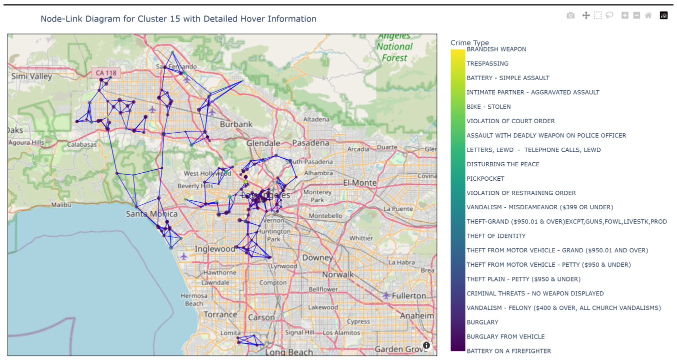


Fig. 26. Node-Link Diagram for Cluster 15

- The node-link diagram for Cluster 15 as shown in Fig. 26 reveals a wide geographic spread of incidents across urban and suburban areas, including Central Los Angeles, Burbank, Santa Monica, and Inglewood, indicating that crimes in this cluster are dispersed rather than concentrated in a single hotspot.
- The most common crime observed is Battery on a Firefighter, with Central LA emerging as a significant hotspot, reflecting the socio-economic dynamics of the region.
- Urban centers like Central Los Angeles are likely to experience higher crime density due to population, economic activity, and transit hubs, leading to diverse crimes. Suburban areas like San Fernando and Burbank may exhibit patterns of property crimes, including burglary and theft from motor vehicles, emphasizing the need for neighborhood surveillance. Coastal areas such as Santa Monica are prone to opportunistic crimes like pickpocketing and petty theft, particularly in tourist-heavy zones.
- The parallel categories plot as shown in Fig. 27 provides an insightful visualization of victim demographics and crime descriptions. The analysis highlights the following key observations:
  - **Victim Age Group:** The majority of affected victims belong to the age group of 30-50 years, indicating that middle-aged individuals are the most vulnerable.
  - **Victim Sex:** The crimes disproportionately impact male victims, showcasing a significant gender disparity in the affected population.
  - **Victim Descent:** Among the descent categories, individuals identified as White are the most frequently affected demographic group.

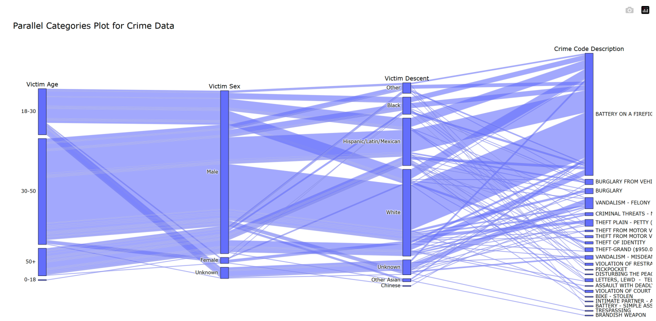


Fig. 27. Parallel Sets Plot for Cluster 15

- This data suggests that crimes predominantly target middle-aged White males.

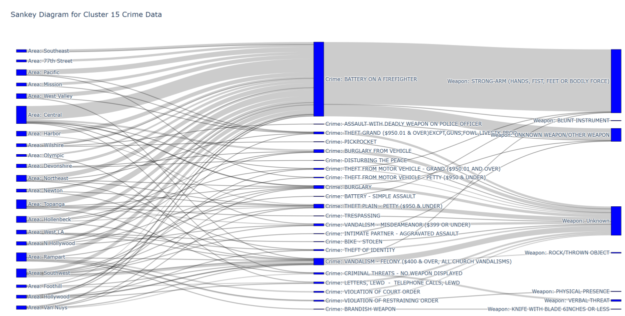


Fig. 28. Sankey Diagram for Cluster 15

- The Sankey Diagram as shown in Fig. 28 highlights that the primary weapons associated with crimes in this cluster are Strong Arm (Physical force).

8) Conclusion: The analysis of Los Angeles crime data (2020-2023) reveals key trends across clusters, highlighting geographic hotspots, demographic vulnerabilities, and crime-specific patterns. Cluster 12, covered cases which, emphasizes the need for enhanced security against non-physical threats like bomb scares, while Cluster 15 highlights widespread urban crimes, often involving physical force. Middle-aged males, particularly from Hispanic and White descents, are disproportionately affected. These insights underscore the importance of localized strategies, targeted interventions, and tailored victim support to improve public safety across diverse areas of the city.

C. ***Workflow 3:*** *Time Series Analysis and Forecasting of Crime Trends in Los Angeles.*

1) Introduction: This report aims to analyze and forecast crime trends in Los Angeles by leveraging historical crime data spanning from 2010 to 2023. Utilizing two comprehensive datasets, we employed time series forecasting techniques to predict future crime occurrences up to the year 2027. The primary objective is to assess the accuracy of the forecasting model so that we can rely on the model's prediction on weekly

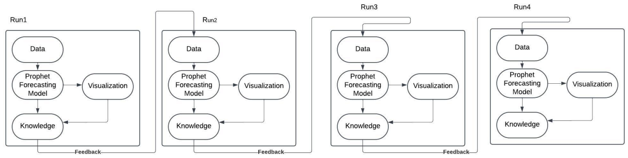


Fig. 29. Workflow Diagram

crimes till the year 2027. You can see the workflow diagram in Fig. 29

Datasets Used:

- **Dataset 1**: Crimes in LA from **2020-2024**
- **Dataset 2**: Crimes in LA from **2010-2019**

By integrating data from these two periods, we strived to enhance the model's predictive capabilities and validate its performance against actual crime data. The analysis involves data preprocessing, outlier detection and removal using the Interquartile Range (IQR) method, forecasting with the Prophet model, and comparing forecasted values with real observations.

2) Methodology: The analysis follows a structured workflow consisting of the following steps:

1) Initial Plotting of Weekly Crime Data and Forecasting
2) Model Validation Using Historical Data
3) Outlier Detection and Removal
4) Refinement of the Forecasting Model
5) Comparison with Actual Values from Dataset 1
6) Final Forecasting Incorporating Both Datasets

Run 1:

3) Initial Plotting of Weekly Crime Data and Forecasting: In Assignment 1 (A1), we began by plotting the weekly crime data from Dataset 1 (2020-2024) as seen in Fig. 13 of Assignment-1. The goal was to visualize the existing trends in crime occurrences. Building upon this, we used the Prophet model to forecast crime rates up to the year **2027**.

- Prophet is a forecasting tool developed by Facebook, suitable for time series data with trends.
- **Data Limitation**: We did not include data points from 2024 onwards in the initial forecasting mentioned reasoning in A1.

Observation: The initial forecast seen in Fig.30 provided a baseline understanding of expected crime trends but lacked validation due to the absence of actual future data.

Run 2:

4) Model Validation Using Historical Data: To evaluate the accuracy of our forecasting model, we needed actual data to compare against the predicted values. Since future data up to 2027 were unavailable, we:

- Utilized **Dataset 2** (2010-2019) to train the model.
- Forecasted weekly crime rates for the years **2020-2023**.
- Compared these forecasts with actual data from **Dataset 1**.

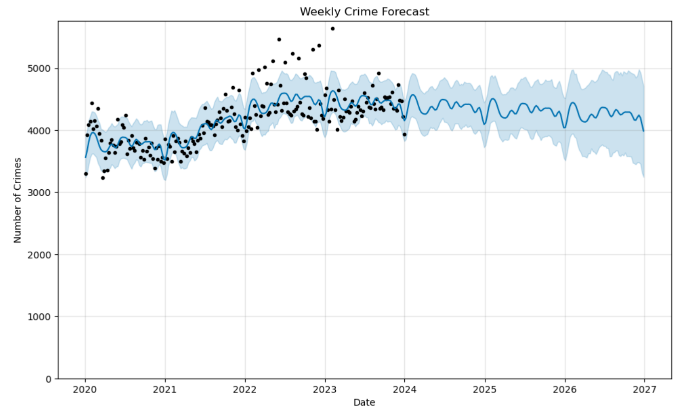


Fig. 30. Weekly Crime Forecast by training on data points from 2020-2023 and forecasting till 2027

Purpose: This approach allowed us to assess how well the model predicts known values, providing insights into its accuracy and reliability.

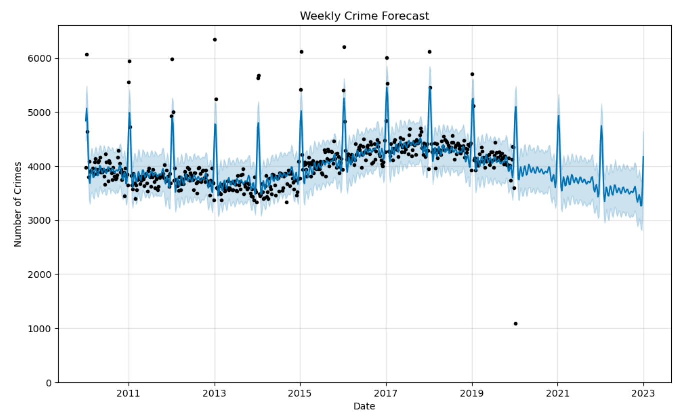


Fig. 31. Weekly Crime Forecast by training on data points including outliers from 2010-2019 and forecasting till 2023

Outlier Detection and Removal: Upon analyzing the forecast shown in Fig, 31, we noticed clear outliers:

- **Yearly Spikes**: There were spikes at the beginning of each year throughout the historical data.
- **Impact on Forecasting**: These spikes carried over into the forecasted years, which was undesirable as they didn't represent typical crime patterns.

Outlier Removal Using Interquartile Range (IQR) Method:

- **IQR Method**: A statistical technique used to identify and remove outliers based on the distribution of the data.
- **Process**:
  - Calculated the first quartile (Q1) and third quartile (Q3) of the weekly crime data.
  - Computed the IQR: **IQR = Q3 - Q1**.
  - Defined the lower and upper bounds:
    * **Lower Bound**: Q1 - 1.5 * IQR
    * **Upper Bound**: Q3 + 1.5 * IQR

– Removed data points outside these bounds.

Reasoning: Outliers can skew the model's understanding of underlying patterns. Removing them helps in improving the model's accuracy by focusing on typical data behavior.

Run 3:

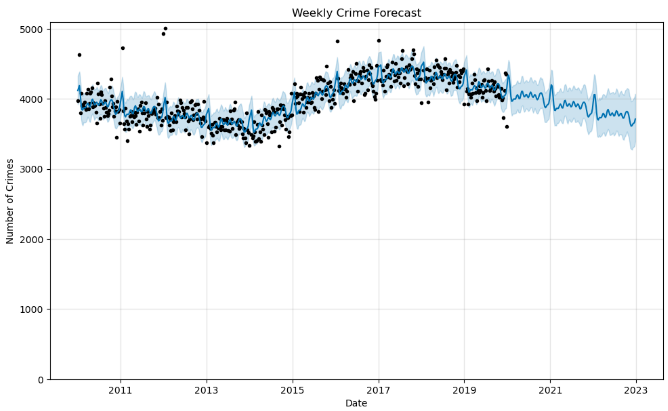


Fig. 32. Weekly Crime Forecast by training on data points excluding outliers from 2010-2019 and forecasting till 2023

5) Refinement of the Forecasting Model: With outliers removed, we retrained the Prophet model. Results of forecasting as seen in Fig. 32

- **Improved Forecast**:
  – The previously observed spikes were no longer present in the forecasted data.
  – The confidence intervals (represented by outer blue lines) provided a clearer indication of the model's certainty.
- **Benefits**:
  – A more accurate representation of expected crime trends.
  – Reduction in the influence of anomalous data points.

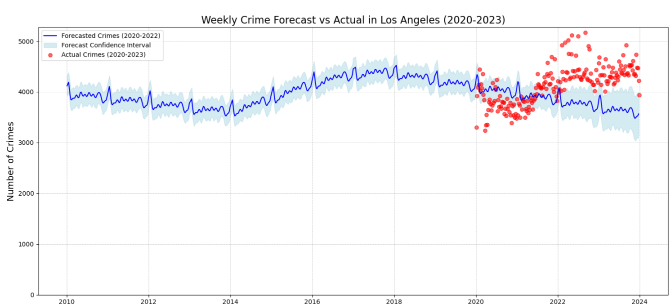


Fig. 33. Weekly Crime Forecast trained on data points excluding outliers from 2010-2019 and forecasting till 2023 vs Actual Data from 2020-2023

6) Comparison with Actual Values from Dataset 1: To further assess the model's performance:

- **Visualization**:
  – In Fig. 33 plotted the actual weekly crime data from **2020-2023** (Dataset 1) using a scatter plot for better clarity.
  – Superimposed the scatter plot onto the forecasted values from the model.

Findings:

- **Accuracy**: The model's predictions closely followed the actual data trends in many instances.
- **Scope for Improvement**: Some deviations indicated potential areas for refining the model further.

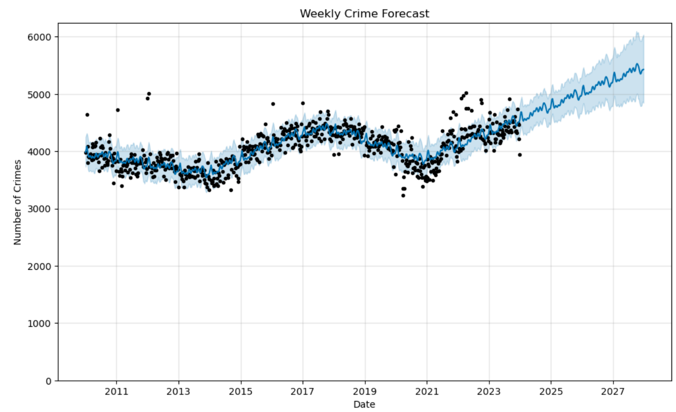


Fig. 34. Weekly Crime Forecast by training on data points from 2010-2023 and forecasting till 2027

7) Final Forecasting Incorporating Both Datasets: **Run 4:** Finally, we combined data from both datasets:

- **Combined Data**: Included crime data from **2010-2019** and **2020-2023**.
- **Forecasting**:
  – Used the extended dataset to forecast crime trends up to **2027**, you can see the results in Fig. 34.

Advantages of Combining Datasets:

- **Enhanced Model Training**: More data provides a better foundation for the model to learn patterns.
- **Improved Accuracy**: Incorporating a broader time span captures more variations and trends.

8) Conclusion: This analysis showcased the application of time series forecasting to predict crime trends in Los Angeles up to the year 2027. By utilizing datasets spanning over a decade, we were able to:

- **Assess Model Performance**: Through validation against actual data from 2020-2023, we gauged the accuracy of the forecasts.
- **Improve Forecast Accuracy**: Outlier removal using the IQR method led to more reliable predictions.
- **Identify Improvement Areas**: Comparing forecasts with real observations highlighted opportunities for model enhancement.

D. *Workflow 4: Looking for correlation between demographics of victims and which area they belong to.*

1) Workflow Analysis: This topic was explored briefly in A1. In Fig. 34 and Fig. 42 of A1 we see pie charts for Victim sex and Descent respectively put on a map. We see some geographical patterns in each visualization independently. More inferences may be drawn when both these features are combined alongside other demographics like victim age. This combined data was not available. Hence, unsupervised clustering the areas using the cumulative demographics from each area as parameters. Further analysis is done on these clusters. A visual representation of the workflow can be seen below. All code for this section can be found in Workflow_4_5.ipynb

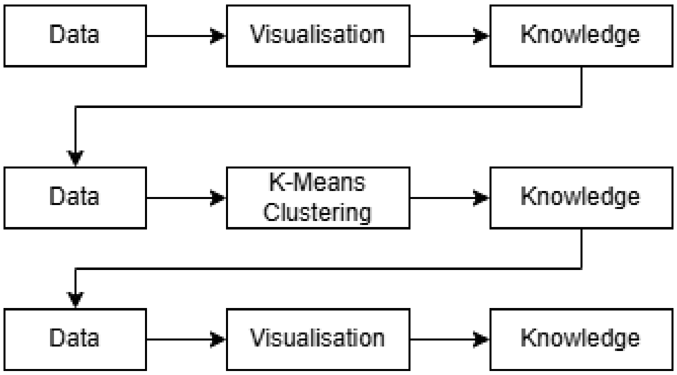


Fig. 35. Flowchart for Workflow 4

2) Clustering: As mentioned in the workflow, first, clustering was done. K-means clustering algorithm was used and the elbow method was used to find optimal number of clusters. The process can be seen in the code for this section. The elbow is present at 5 clusters. Hence, we cluster areas into 5 parts.

After clustering, we assign each cluster a color and areas are marked with their cluster's color on an interactive map that can be seen in map.html.

An interactive visualisation was made and can be seen in map.html.

In Fig. 36, we see 5 different clusters(0 to 4). These clusters were decided purely on the basis of demographics of victims. Yet, most clusters are concentrated in the same geographical area as well. We see the Pink(Cluster 4) are concentrated in the Middle West part of LA. While Green(Cluster 2) are present only in the Middle South part of LA.

3) Analysis of clusters: Let us look at the different clusters and the Areas in them through Fig. 37. We see a Tree map that shows the sizes of the the different clusters.

Next, We can now try to observe properties of these clusters and what makes them different. Seeing the values does not show any obvious differences. We can use pie charts. We do this in Fig. 38.

We see some variances such as Cluster 3 having a majority of Hispanic/Latin/Mexican Victims while cluster 2 has a huge number of Black victims. Hence, from Fig. 36 we can see a high number of crimes against people from Black descent are committed in the green areas or Middle South LA. This could be due to the population density of African Americans in that area. More inferences can be made seeing the pie charts for Victim Sex. This can be seen in Fig. 39.

We see Cluster 2 having a high number of Female Victims. While Cluster 1 has a big number of Male victims. This could

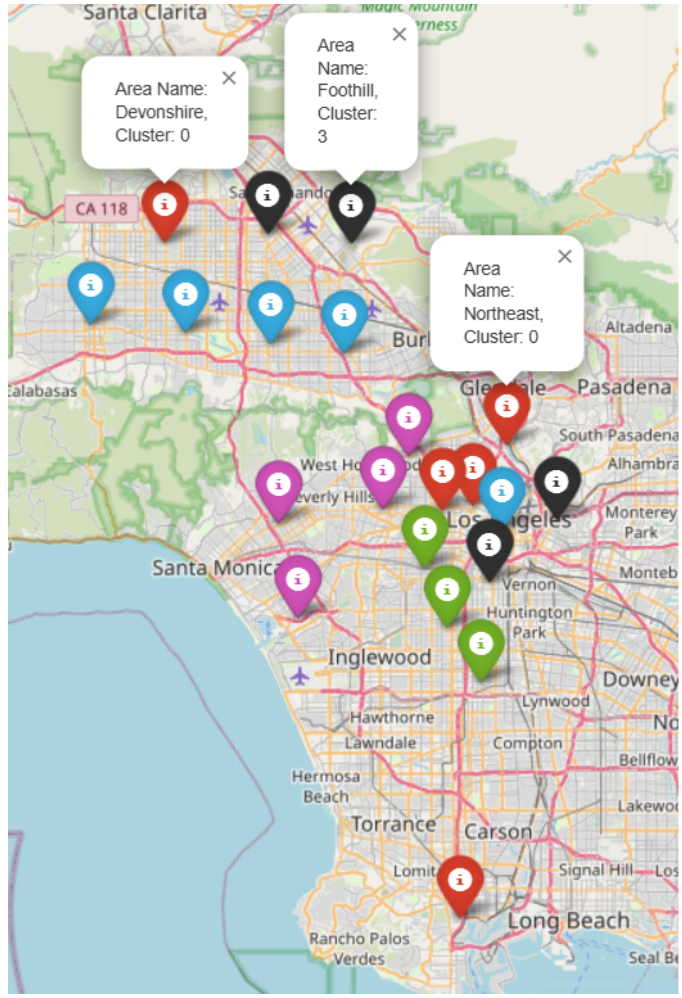


Fig. 36. Screenshot from map.html

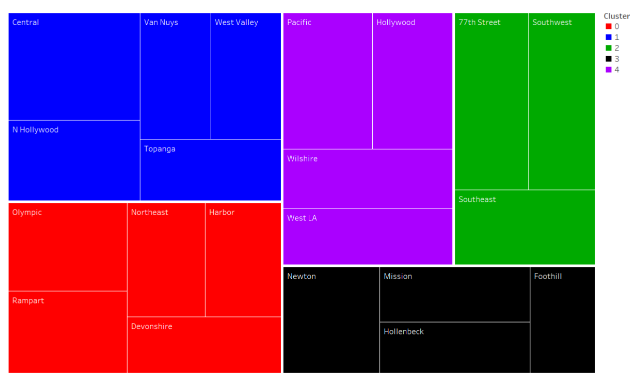


Fig. 37. A view of different clusters and what areas contribute to them

point to Cluster 2 being specifically unsafe for Women. From Fig. 36 we can infer that the South Middle part of LA is comparatively unsafe for women while the Blue regions(Cluster 1) in the north are relatively safer.

With these pie charts, it is still difficult to see the relation between Sex and Descent. This can be seen using a Parallel Categories Plot. It can be seen in parallel_categories.html.

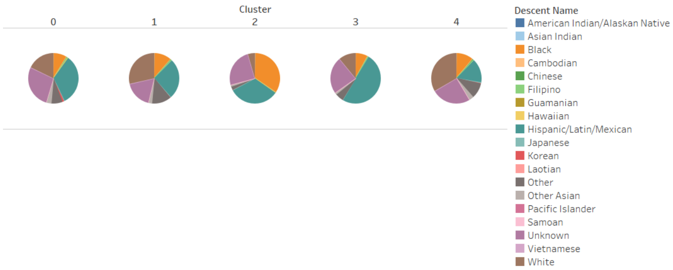


Fig. 38. Pie charts of different clusters and their Victim descents

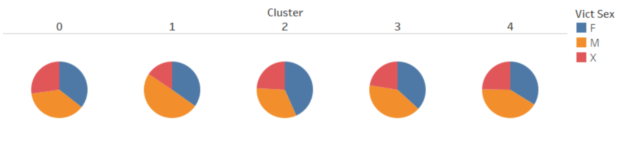


Fig. 39. Pie charts of different clusters and their Victim descents

Since it is interactive, many inferences can be drawn from it. We'll talk about a few of them here.

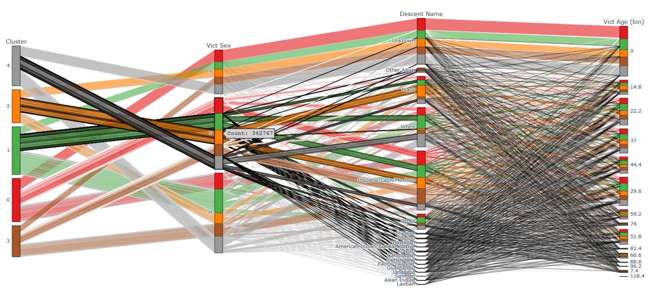


Fig. 40. PCP with lines highlighted for Female victims

In Fig. 40 we can see that a majority of female victims in cluster 2 were also of Black descent. Hence the area is more unsafe for black women. This cannot be inferred from the pie charts. An example can be seen in Fig.41

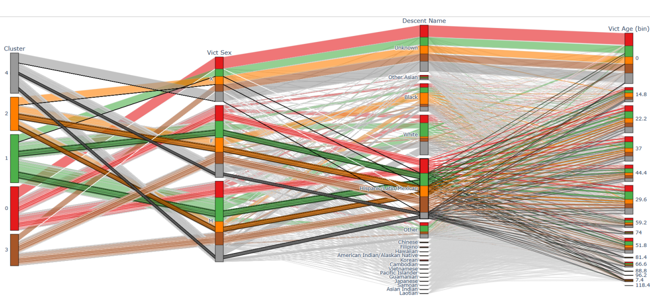


Fig. 41. PCP with lines highlighted for Hispanic/Latin/Mexican victims

We see while Cluster 3 has high rates of both Hispanic/Latin/Mexican descent and Men, the amount of Hispanic/Latin/Mexican men and women is roughly same. Hence, A PCP is extremely useful in this scenario.

4) Conclusion: The clustering analysis revealed key geographic and demographic patterns in crime. Areas were grouped into five clusters, showing regional cohesion and distinct victim profiles. Cluster 2 (Middle South LA) had high proportions of Black victims and women, highlighting specific safety concerns, while Cluster 3 (Middle West LA) was dominated by Hispanic/Latin/Mexican victims. The parallel categories plot uncovered nuanced interactions, such as the vulnerability of Black women in Cluster 2. This approach demonstrates the value of combining clustering and advanced visualizations to better understand crime demographics and inform targeted interventions.

E. *Workflow 5: Analyzing affect of Pandemic on Crimes against various Demographics*

This short workflow tries to establish if there was any difference in how the pandemic affected number of victims belonging to each Demographic. All code for this section can also be found in Workflow_4_5.ipynb.

1) Workflow Analysis: In A1, we attributed a lot of changes in the number of victims over time to the pandemic. We also saw a difference in how different demographics were affected. To confirm our observations, we can use previous data to forecast number of victims form 2020-2024. If there is a difference, it can most likely be attributed to the pandemic. For this, we need previous data. Hence, we append our initial dataset with new crime data from 2010-2019. We can then use it to forecast data for the future.

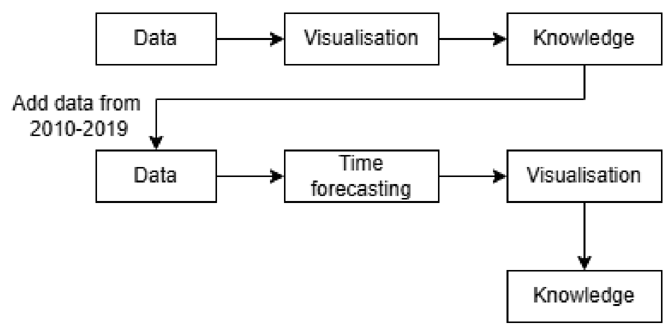


Fig. 42. Flowchart for Workflow 5

2) Forecasting: Let us first perform the analysis on Descent. We'll only look at the 3 major Descents. "Black", "White" and "Hispanic/Latin/Mexican".

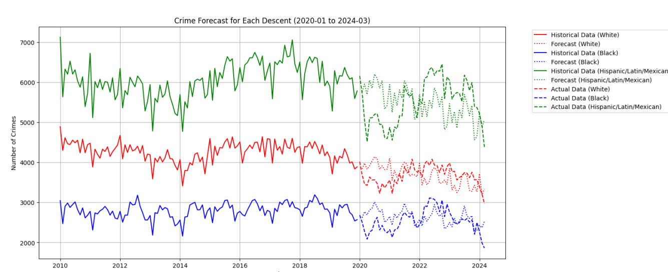


Fig. 43. Forecast vs Actual values for Victim Descent

In Fig. 43, we see the data from 2010 to 2020 then the dotted line shows the forecast while the dashed line shows the actual value. We see that all descents had a dip from predicted values in the beginning of 2020. However, "Hispanic/Latin/Mexican" has a way bigger dip than the other two. They also have a much bigger rise compared to the forecasted prediction when the pandemic ends. This shows they were disproportionately affected. There could be lower victims for various reasons, one being that due to the pandemic, various crimes were not reported in the community. It could also be that the crimes committed against that community took place in public places with a large number of people and the pandemic did not allow those conditions to take place.

Let us now look at how different Genders were affected.

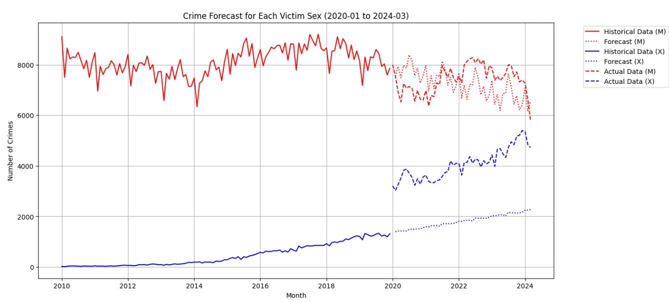


Fig. 44. Forecast vs Actual values for Victim Sex(Male)

Since, Male and Female have similar values, putting them in the same graph led to a lot of overlap and visual clutter. Hence they can be seen in separate graphs in Fig. 44 and Fig. 45.

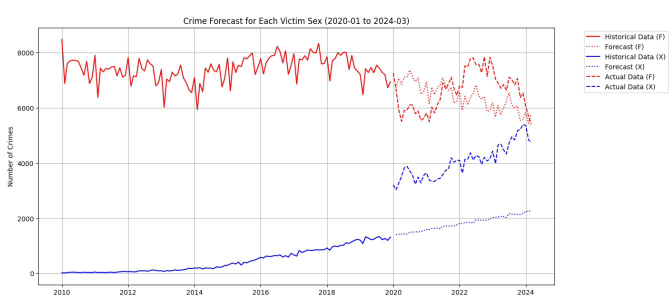


Fig. 45. Forecast vs Actual values for Victim Sex(Female)

We see that Males do not see that much of a change, in fact, the forecast is quite accurate, which implies male victims were not largely affected by the pandemic. Females, however, see quite a big change compared to the forecast, with a significant dip in the early stages of the pandemic followed by a sharp rise as restrictions eased. This suggests that the pandemic had a disproportionate impact on crimes involving female victims, potentially due to shifts in crime patterns, reporting behaviors, or societal dynamics during this period. These variations underscore the need for targeted analysis and interventions to address the unique vulnerabilities of different demographic groups during crises.

We also see a huge rise in the number of cases with unknown gender during and after the pandemic period. This could indicate disruptions in data collection or reporting mechanisms, where essential victim details, such as gender, were omitted or improperly recorded. It may also reflect an increase in anonymous reporting or crimes involving unidentified victims.

3) Conclusion: The forecasting analysis highlights the significant impact of the pandemic on victim demographics. By comparing predicted and actual values for 2020–2024, we observed notable differences, particularly among Hispanic/Latin/Mexican victims, who experienced a pronounced dip followed by a sharp rise in numbers. This suggests a disproportionate effect on this community, likely due to pandemic-induced changes in crime occurrence and reporting.

Similarly, analysis by gender revealed that female victims were more affected than males, with substantial deviations from forecasts during and after the pandemic. The sharp rise in cases with unknown gender further emphasizes challenges in data accuracy and reporting during this period. These findings underscore the importance of robust data collection and targeted interventions to address demographic disparities in crime impact during crises.

CONTRIBUTIONS

- **IMT2022110 Subham Agarwala:** All visualizations and analysis for workflow 1. Inferences were made on data and visualizations related to arrests, and the relation of arrests with the demographics and the geography was highlighted, and forecasted arrest data for a later time period to check the validity of an assumption made in A1. Crime and arrest trends were analyzed.
- **IMT2022521 Sarvesh Kumar:** Workflow 2 focused on cluster analysis and visualization of crime patterns in Los Angeles (2020-2023). Clustering techniques, including PCA and k-means, were used to identify patterns , with geospatial visualizations and network analyses providing insights into localized crime trends. For instance, Cluster 12 highlighted bomb threats , while Cluster 15 revealed widespread urban crimes like battery on firefighters, emphasizing victim demographics such as middle-aged males of Hispanic or White descent. Workflow 3 employed time series forecasting using the Prophet model to analyze weekly crime trends (2010-2023) and predict trends up to 2027. By integrating historical and recent datasets, outlier removal using the IQR method improved forecast accuracy, validated against real data.
- **IMT2022546 Ayush Gupta:** All visualizations and analysis for workflow 4 and 5. Clustering was done on areas based on demographics to strengthen inferences made in A1 and to make new ones. These clusters were analysed using pie charts and PCP. In workflow 5, forecasting was done to see affect of pandemic on number of victims from each demographic to verify assumptions made in A1. Analysis was done on forecasts and the actual value.

REFERENCES

Dataset for A1: https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000

Arrest data: https://www.kaggle.com/datasets/cityofLA/los-angeles-crime-arrest-data?select=arrest-data-from-2010-to-present.csv

Extended A1 dataset: https://www.kaggle.com/datasets/sumaiaparveenshupti/los-angeles-crime-data-20102020

APPENDIX

DAS732: Data Visualization Assignment 1

Subham Agarwala
IMT2022110
IIIT Bangalore
Bangalore, India
Subham.Agarwal@iiitb.ac.in

Sarvesh Kumar
IMT2022521
IIIT Bangalore
Bangalore, India
SarveshKumar.A@iiitb.ac.in

Ayush Gupta
IMT2022546
IIIT Bangalore
Bangalore, India
Ayush.Gupta546@iiitb.ac.in

I. Introduction

This report is a compilation of our work on the "USA Big City Crime Data"[1] dataset, which includes data visualizations and inferences to answer the question *"What are the major crime trends in big cities, and how do factors such as location, type of crime, and demographic details influence these trends?"*. We have divided our analysis into three tasks which are:

- **T1:** Analysis of crime trends based on location
- **T2:** Analysis of crime trends based on the type of crime and frequency
- **T3:** Analysis of crime trends based on victim demographics

II. Preprocessing of Dataset

Before performing any visualizations or analysis following are the steps we took to preprocess the dataset:

- We dropped the dataset for Chicago as it did not provide any more information than just the location of the crime and the crime type.
- Removed duplicate entries
- Removed rows with negative victim age
- Grouped null values for victim sex and victim descent column into a single category "X" to represent unknown values.
- Removed "DR_NO", "Location", "CrossStreet" and "Mocodes" due to lack of useful information.
- Removed rows with Longitude value as "0" as it does not belong to LA.
- Removed "CrmCd1" as it was the same as "CrmCd".
- Computed the following new fields to ease and better the visualizations:
  - "AvgLong" and "AvgLat: grouped latitudes and longitudes by area bu computing their avg."
  - "Hour of the Day": extracted the hour from the "Time Occ" column
  - "Arrest_Cat": specifies whether the criminal was arrested or not. Extracted from "status" column.
  - "Modified_Part1-2": converted values to string.
  - "Modified_Weapon_Desc": replaces null values in the "Weapon Desc" column.
  - "Time_Diff_Occ_Repo": The difference in number of days between "Date Occ" and "Date Rptd" columns
  - "Criminal_Age_Category": classifies arrested criminals into adult and juvenile. Extracted from "Status" column.
  - "Descent Name": expands the labels given in "Vict Descent" column to their actual meaning.

III. Tasks

A. ***Task 1:*** *Analysis of crime trends based on location*

The goal of this task is to analyze the distribution of crimes across different geographic areas in the dataset to identify which areas experience higher crime rates and potential crime hotspots. This analysis aims to provide insights into how crime is concentrated across different parts of the city and to help law enforcement or policymakers focus resources on the areas most affected.

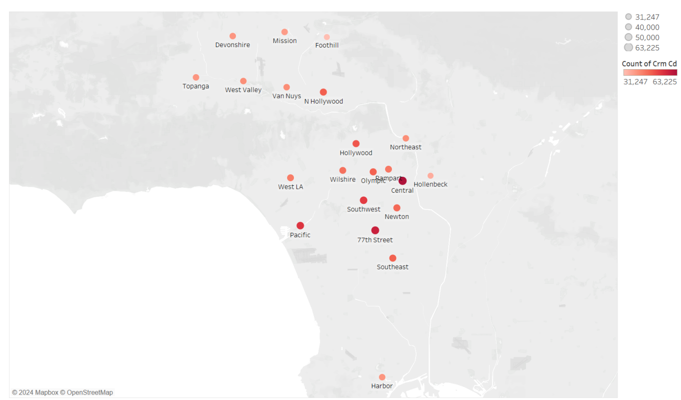


Fig. 1. Geospatial map that shows the crime hotspots in LA.

First we plot the data on a geospatial map to identify crime hotspots in LA (Fig. 1).From the visualization we can see that areas like Central, 77th Street, Pacific and Southwest have the highest crime count among the rest of the areas in LA with Central having the highest crime count. Foothill and Hollenbeck seem to have the lowest crime count.

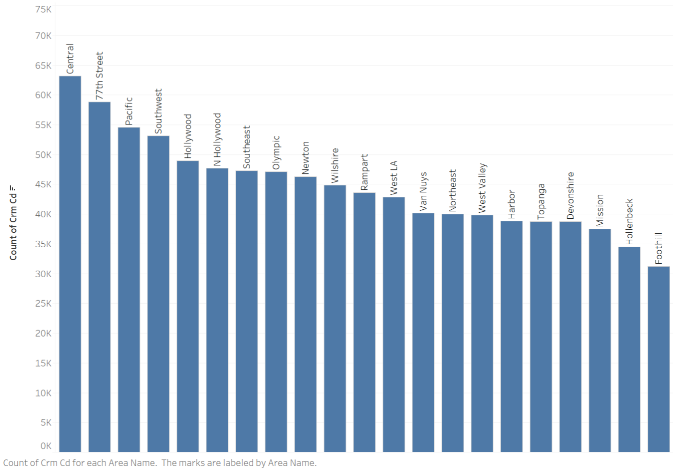


Fig. 2. Bar graph of crime count in the respective areas.

Fig. 2 gives us a better understanding of the crime count in the different areas of LA.

Now that we have identified that the Central area of LA is the biggest hotspot for crimes in the city, we will look deeper into the specifics of the crimes taking place there. We will now explore the premises in the Central region of LA where major volume of crimes take place.

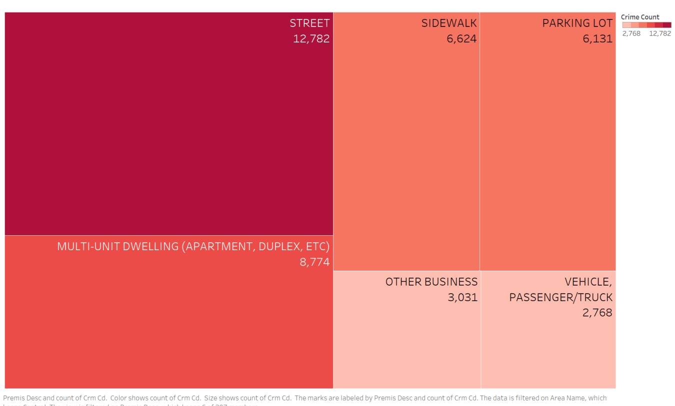


Fig. 3. Premises in Central region with maximum crime counts.

Fig. 3 shows the most common premises for crime occurrences in Central LA. As depicted, the majority of crimes take place on the street, which suggests that public spaces, where individuals are more exposed, are particularly vulnerable to criminal activity. Apartments/Duplexes rank as the second most frequent crime scene, indicating that residential areas are also significant targets for crime. This may highlight issues related to property crimes, domestic incidents, or security vulnerabilities in housing complexes.

To observe the hours of the day in which maximum crimes take place in the above premises in Central LA we have visualized the data in the form of a heat map (Fig. 4).

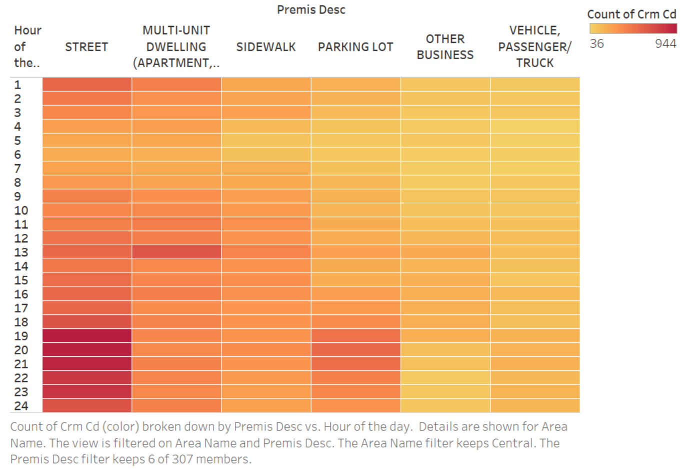


Fig. 4. Heat map showing density of crimes taking place at different hours of the day in Central LA.

We see that the maximum number of crimes in Central LA in all premises take place during the later hours of the day. This pattern is consistent across most premises, indicating that evening and nighttime are the most vulnerable periods for criminal activity in Central LA. This could be due to reduced visibility, fewer people in public spaces, or a higher likelihood of individuals being outside during these hours.

For the "Other Business" category, we observe a distinct peak in crime incidents between the 13th and 19th hour (1 PM to 7 PM). This could suggest that certain types of businesses, such as retail stores, restaurants, or entertainment venues, may experience more crime during business hours, possibly linked to theft, robbery, or other opportunistic crimes.

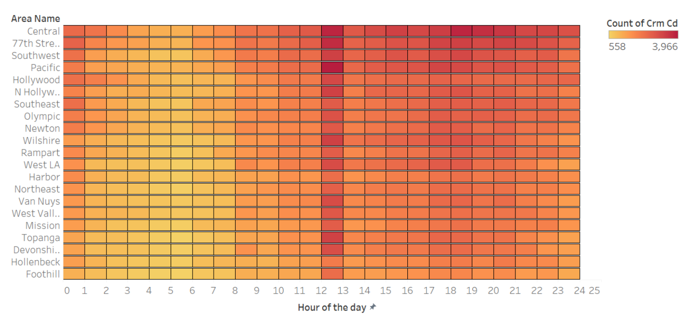


Fig. 5. Heat map illustrating density of crimes taking place at different hours of the day in LA.

Fig. 5 displays the distribution of crime incidents across different hours of the day throughout Los Angeles. The data reveals a similar pattern to that observed in Fig. 4 for Central LA, where the highest crime activity occurs during the evening and nighttime hours. However, we see an unusual spike in crime count during the first hour (12 AM) and the 12th hour (12 PM). This anomaly is present across all areas of LA and likely stems from 00:00 and 12:00 being used as default placeholders for unknown or missing time values in the dataset, skewing the crime count at these times.

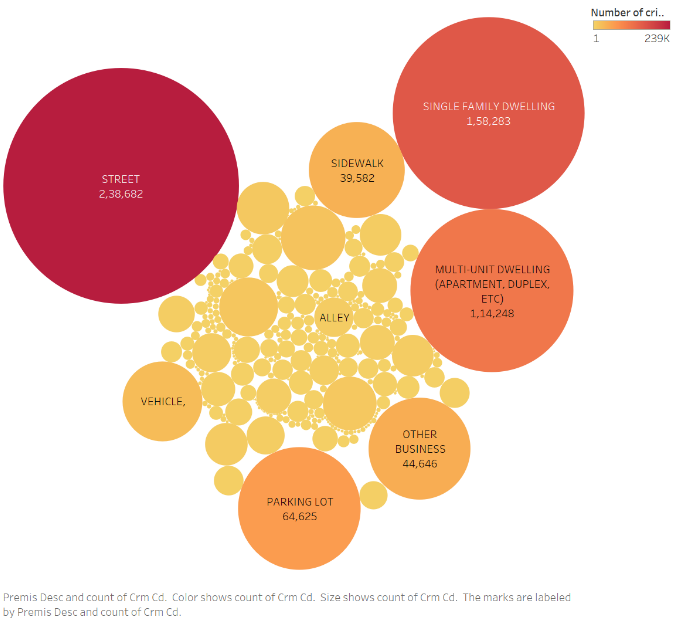


Fig. 6. Crime count by premises across all of LA

When analyzing crime count by premises across all of Los Angeles in Fig. 6, we observe a similar pattern to that seen in Central LA, with the street remaining the most frequent location for criminal incidents. However, in contrast to Central LA, single-family dwellings emerge as the second most frequent premises for crime in the broader LA region. This shift highlights the vulnerability of residential areas across the city.

The prevalence of crime in single-family dwellings suggests that crimes such as burglary, theft, and domestic violence may be more common in residential neighborhoods than in apartment complexes or multi-family housing units, which were more prominent in Central LA. Streets and public spaces continue to be hotspots, indicating that crimes in public or semi-public areas (e.g., parking lots, sidewalks) are still a significant concern.

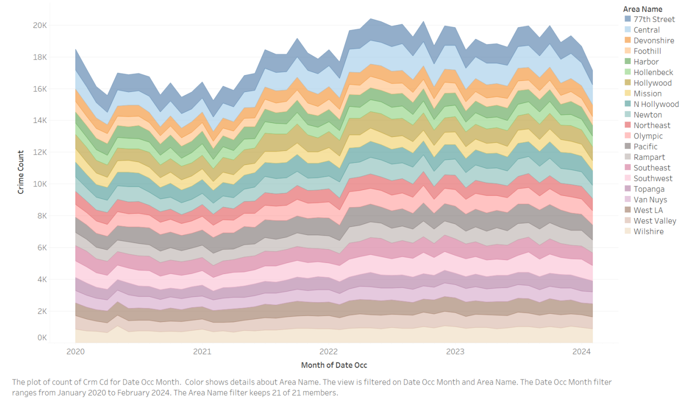


Fig. 7. Area chart showing crime count by area in LA over time

Fig. 7 is an area chart illustrating the trend in crime counts across different regions of Los Angeles over the years. The chart provides a visual representation of how the number of crimes has fluctuated or remained stable in various areas.

The most striking observation from the chart is the significant increase in crime count in Central LA. This region shows a notable and steady rise in criminal incidents, shown more clearly in Fig. 8, setting it apart from other areas in Los Angeles. In contrast, other regions either show a slight increase in crime over time or maintain a relatively constant rate. These areas appear to be less affected by the broader trends influencing crime in Central LA.

Fig. 8 effectively highlights the dominant contribution of Central LA to the overall crime increase in the city, suggesting that Central LA may be experiencing unique social or economic factors that are driving this upward trend.

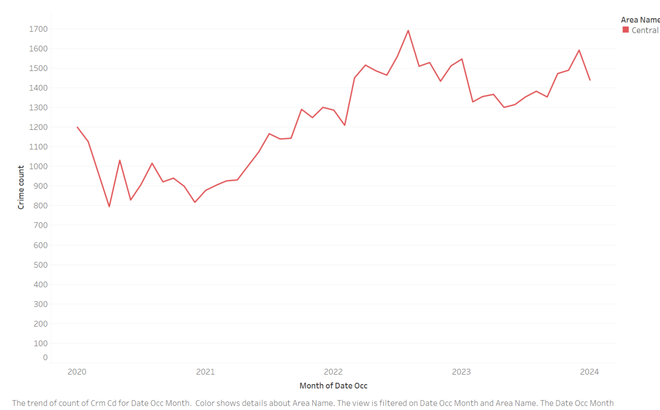


Fig. 8. Line chart showing the increase in crime in Central LA over the years

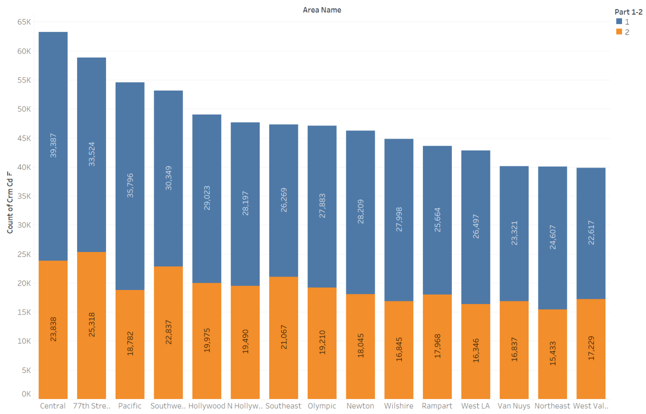


Fig. 9. Distribution of part 1 and part 2 crimes across top areas in LA with respect to crime count.

Crimes in the dataset are categorised into part 1 and part 2 crimes. Fig. 9 presents the distribution of Part 1 and Part 2 crimes for the top 15 areas in Los Angeles based on total crime count. The visualization highlights a clear trend: in all of these high-crime areas, Part 1 crimes are more prevalent than Part 2 crimes.

Part 1 crimes typically include serious offenses such as homicide, robbery, assault, and burglary, while Part 2 crimes encompass less severe offenses, like vandalism, public intoxication, and disorderly conduct. The fact that Part 1 crimes consistently outnumber Part 2 crimes across all these areas points to the gravity of criminal activity in these regions.

The areas with the highest crime count show particularly large disparities between Part 1 and Part 2 crimes, suggesting that these regions may be hotspots for violent or serious criminal offenses.

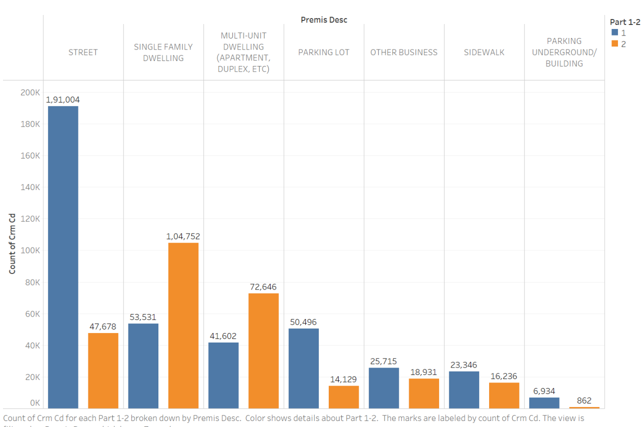


Fig. 10. Distribution of part 1 and part 2 crimes in top premises in LA with respect to crime count

Now that we have seen the distribution of Part 1 and Part 2 crimes across different areas in LA, let's examine their distribution across the top premises in LA.

In Fig. 10, we observe that streets, parking lots, businesses, and sidewalks have a significantly higher number of Part 1 crimes compared to Part 2 crimes. This aligns with the nature of these locations, where offenses like robbery, assault, and burglary (classified under Part 1 crimes) are more likely to occur.

However, we see a contrasting trend when it comes to single-family dwellings and multi-unit dwellings. Residential areas show a higher concentration of less severe crimes (Part 2), possibly reflecting neighborhood disputes, minor offenses, or violations that are less violent in nature.

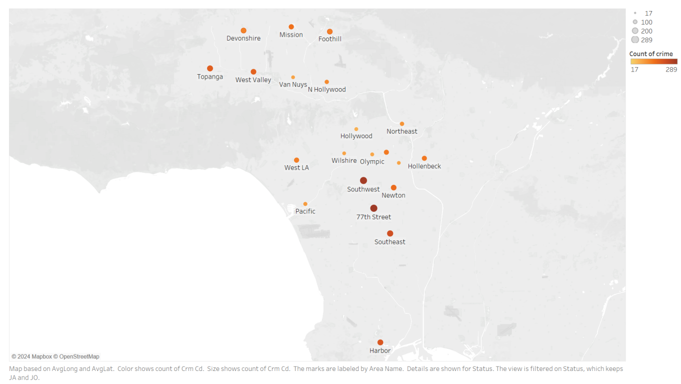


Fig. 11. Geographical plot showing number of juvenile criminals in the different areas of LA.

We also utilized the "Status" column of the dataset, which provides information about actions taken against offenders, to gain insight into the number of juvenile criminals in different areas of Los Angeles. By filtering the data for statuses related to juvenile actions, we were able to estimate the number of juvenile offenders in each area, as depicted in Fig. 11.

Despite having the highest overall crime count, the Central LA area shows a very low number of juvenile criminals. This may indicate that the crimes committed in Central LA are more likely to involve adult offenders or that youth-related criminal activity is minimal in this area. The low juvenile crime rate in Central LA, despite its high overall crime rate, suggests that crime prevention measures for juveniles might already be effective or that adult-related crimes are the primary concern in this region. On the other hand, areas such as 77th Street and Southeast LA report the highest numbers of juvenile offenders. These regions appear to have a greater concentration of youth involvement in criminal activities compared to other areas.

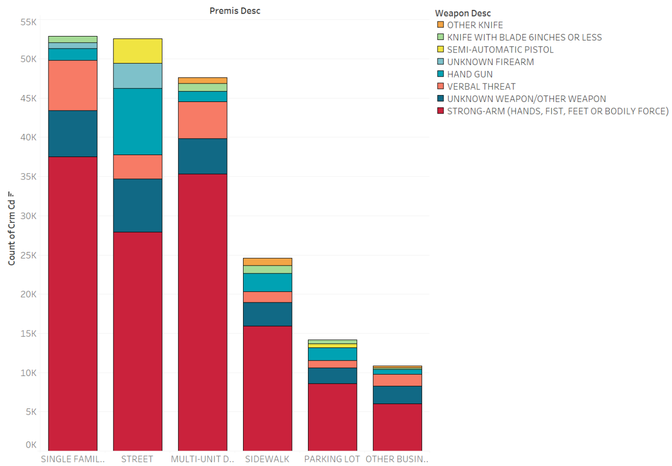


Fig. 12. Top weapon used in the premises with maximum number of crimes.

Our final crime analysis focuses on the top weapons used in the premises with the highest crime counts across Los Angeles. As shown in Fig. 12, we observe distinct patterns in weapon usage based on the type of location.

Strong-arm force (which includes hands, fists, feet, or bodily force) is the most commonly used method to commit crimes across all premises, indicating a high prevalence of physical assault in both public and residential areas. In residential areas like single-family dwellings and multi-unit dwellings, verbal threats rank as the second most common method used, followed by the use of handguns and unknown firearms. This suggests that while physical force dominates, verbal intimidation and firearms also play a role in residential crimes. On streets, handguns are used more frequently than verbal threats, indicating a higher tendency for gun-related crimes in public areas. This is followed by the use of unknown firearms and semi-automatic pistols, showing that gun violence is a major concern in these locations. In sidewalks and parking lots, aside from strong-arm tactics, knives with blades 6 inches or less are also commonly used, adding a layer of concern for weapon-related incidents in these public spaces.

B. ***Task 2:*** *Analysis of crime trends based on the type of crime and frequency*

The goal of this task is to analyze the distribution of crimes by type and frequency within Los Angeles. This analysis aims to identify which types of crimes are most prevalent and how their frequency varies across different time periods. By examining these patterns, we seek to gain insights into the predominant crime issues affecting the city and understand trends that may inform targeted interventions.

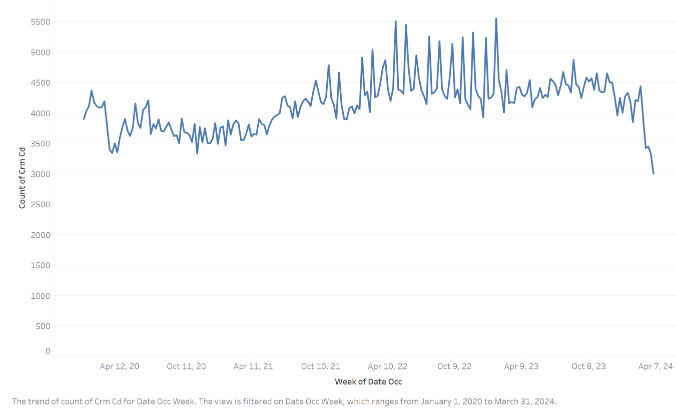


Fig. 13. Weekly Crime Trends in Los Angeles

The time series graph [Fig. 13] of weekly crime data in Los Angeles reveals a notable spike in criminal activity during 2022 and 2023. This significant rise may be linked to various factors such as economic changes, shifts in policing, or other influences. The data underscores the need to investigate the causes behind this surge and to develop targeted strategies to address the increase in crime. Notably, the graph shows a dip in crime rates after 2023, which could be due to incomplete data for 2024.

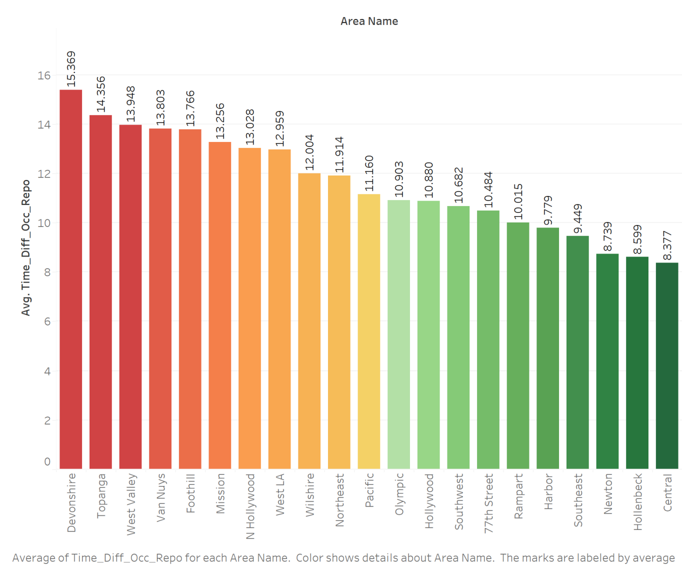


Fig. 14. Average Crime Reporting Delays Across LA Regions

[Fig. 14] shows that crime reporting times vary significantly across different regions of Los Angeles. Central, Hollenbeck, and Newton report crimes the quickest, with an average delay of only 8 days. In contrast, Devonshire and Topanga exhibit the longest delays, averaging over two weeks. These differences may indicate variations in community responsiveness, police efficiency, or local infrastructure. The disparity in reporting times raises important questions about the underlying causes and potential interventions to improve crime reporting efficiency in slower regions.

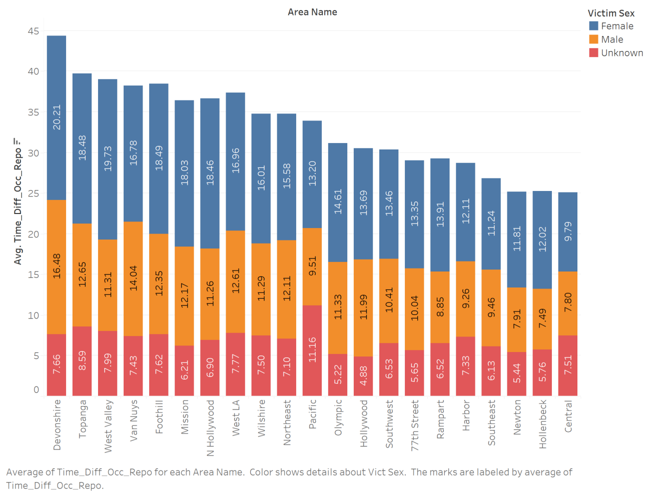


Fig. 15. Gender Disparity in Crime Reporting Across LA Regions

[Fig. 15] highlights the average number of days to report a crime across LA regions, segmented by gender. On average, crimes with male victims are reported 4-5 days earlier than those involving female victims. This trend is consistent across regions, with Hollenbeck, and Newton having the shortest delays, but still showing a notable gender gap in reporting times.

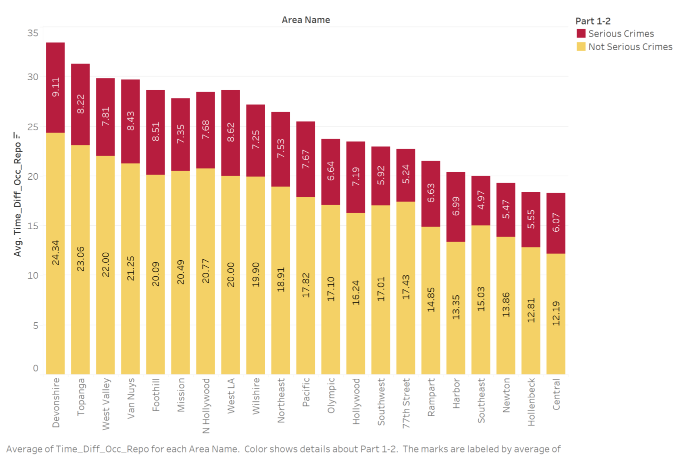


Fig. 16. Reporting Time Delays by Crime Seriousness across LA Regions

[Fig. 16] shows the average number of days to report a crime in LA, categorized by crime seriousness. It reveals that serious crimes are reported over a week faster than less serious ones, demonstrating a prompt response to urgent cases. This insight can be used to evaluate the effectiveness of crime reporting and response mechanisms across different regions.

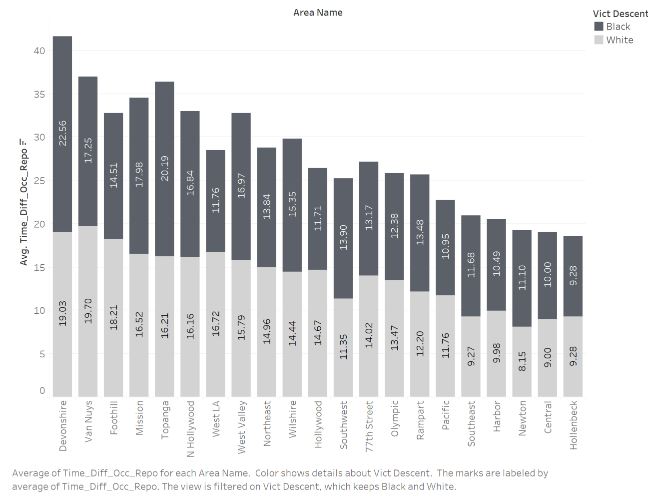


Fig. 17. Racial Equality in Crime Reporting

[Fig. 17] shows the average number of days to report a crime in LA by victim's skin color. The data indicates minimal delays between reporting times for white and Black victims, suggesting negligible racial discrimination in reporting processes

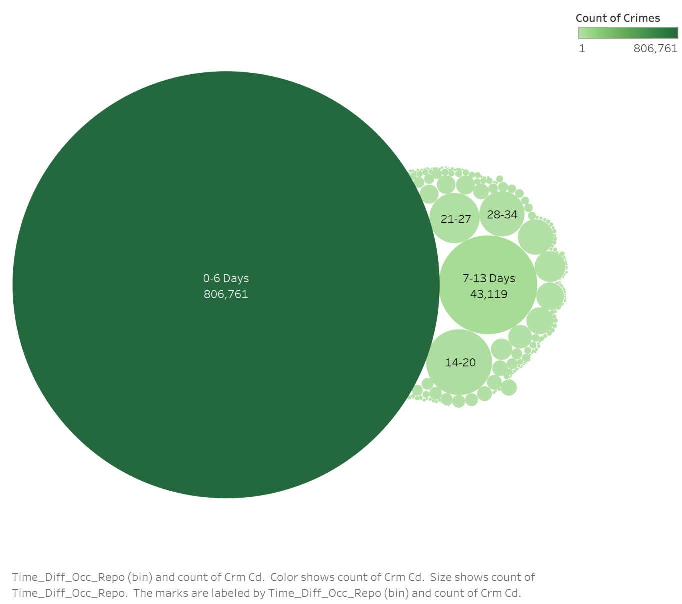


Fig. 18. Crime Reporting Timeliness in LA

[Fig. 18] show that most crimes in LA are reported within the first week, indicating a generally prompt reporting process. There are minimal delays between reporting times for different skin colors, suggesting negligible racial discrimination, and serious crimes are reported significantly faster, reflecting a prioritized response to urgent cases. Although most crimes are reported within the first week, some experience significant delays, which skews the average reporting times and affects the perceived efficiency across regions.

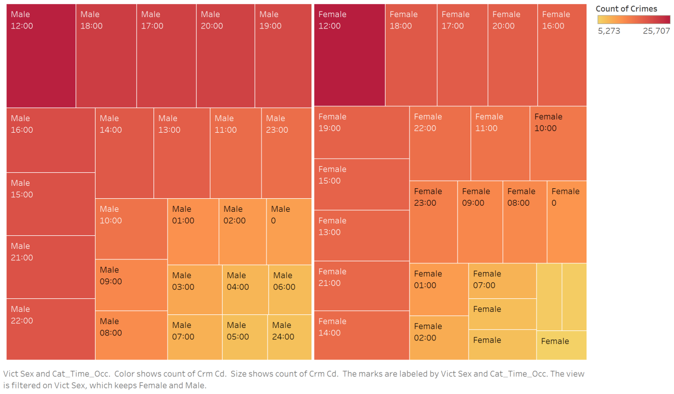


Fig. 19. Crime Distribution by Hour and Gender

[Fig. 19] presented as a tree map, shows the distribution of crimes by hour of the day, segmented by gender. The data reveals a notable spike in crime reports at 12:00, which may indicate potential data inaccuracies or a default time entry. Overall, the count of crimes is nearly equal between men and women, with higher incidents occurring during late hours and fewer during earlier times of the day.

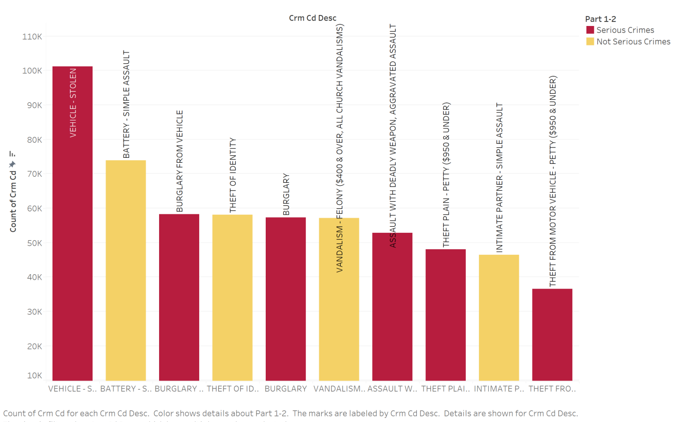


Fig. 20. Top 10 Most Common Crimes in LA

[Fig. 20] displays the top 10 most common crimes in LA, categorized by crime seriousness. Vehicle theft emerges as the most frequent and serious crime. This visualization is useful for identifying prevalent crime types and their severity, aiding in resource allocation and targeted interventions to address the most critical issues.

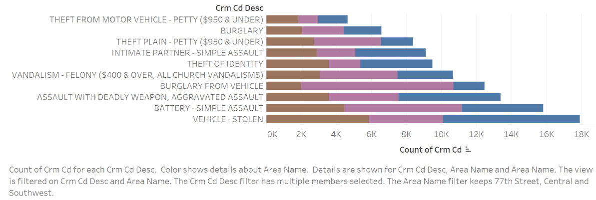


Fig. 21. Crime Trends in Violent LA Areas

[Fig. 21] highlights the most common types of crimes in the violent areas of LA—Central, 77th Street, and Southwest. It shows that Burglary from Vehicle is notably prevalent in Central, while Vehicle Stolen is most common in 77th Street. The data provides insights into the specific crime patterns in these high-violence areas, which can be crucial for targeted policing and crime prevention strategies.

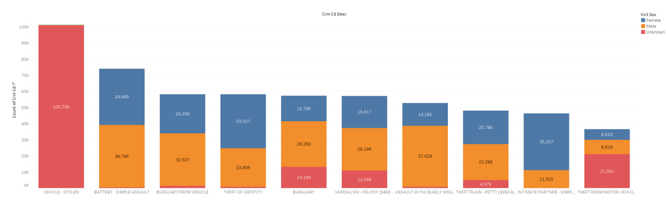


Fig. 22. Gender-Specific Crime Patterns in LA

[Fig. 22] illustrates the top 10 crimes in LA segmented by gender. The data reveals that the victim's sex for Vehicle Stolen is often unknown. Crimes like Intimate Partner Violence and Theft of Identity predominantly affect females, while Assault with Deadly Weapon primarily impacts males.

This visualization is useful for identifying gender-specific crime patterns and understanding which crimes disproportionately affect each gender.

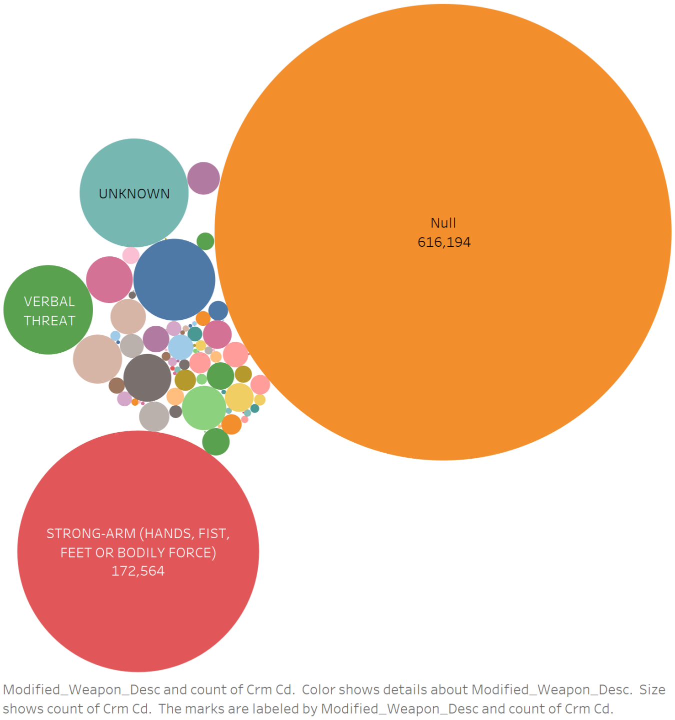


Fig. 23. Weapon Usage in LA Crimes

[Fig. 23] highlights the most common weapons used in crimes across LA. The data shows that most crimes involve no weapon, while in cases where a weapon is used, strong arms, body, or fists are the most frequent.

This visualization is useful for understanding the nature of physical violence in crimes.

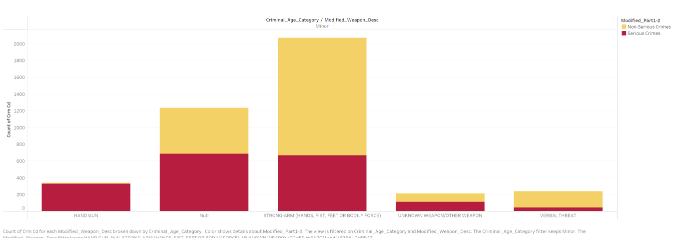


Fig. 24. Weapons Used by Minors in LA Crimes

[Fig. 24] showcases the most common weapons used by minors arrested for crimes in LA, segmented by the seriousness of the offense. The data reveals that minors frequently rely on strong arms and fists, but serious crimes have also been committed using handguns, fists, and even no weapons at all.

This visualization is useful for understanding the role of minors in both non-lethal and serious crimes, aiding in the development of focused intervention strategies for youth crime prevention and rehabilitation.

In addition to minors, I visualised that adults arrested in LA display similar patterns in weapon usage. However, adults commit crimes at a significantly larger volume, approximately 30-35 times more than minors.

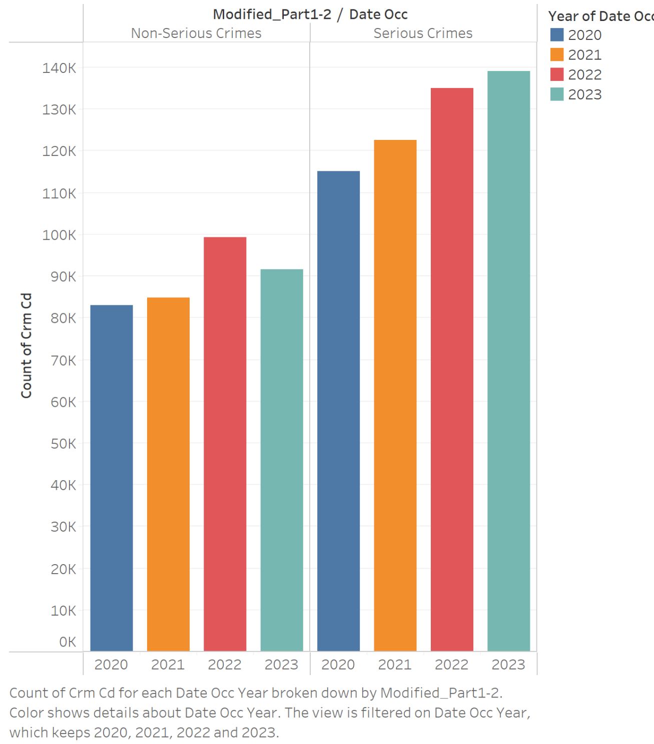


Fig. 25. Yearly Crime Trends in LA (2020-2023) Categorized by Crime Seriousness

[Fig. 25] illustrates the trends in crime seriousness from 2020 to 2023 on a yearly basis. Serious crimes have consistently increased year by year, while non-serious crimes rose until 2022, then saw a decline in 2023. The data for 2024 is excluded due to incomplete months.

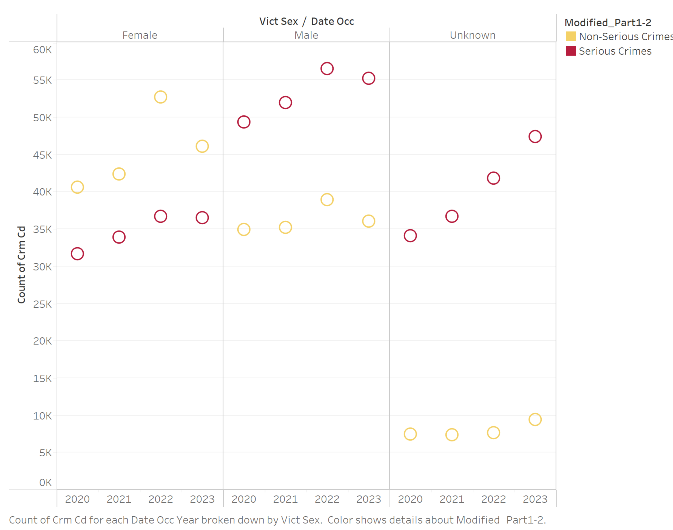


Fig. 26. Gender-Specific Trends in Crime Seriousness (2020-2023)

[Fig. 26] shows the yearly trends in crime seriousness from 2020 to 2023, segmented by victim sex. The data reveals that more male victims are affected by serious crimes than females, while non-serious crimes are more common among female victims. For males, serious crimes outnumber non-serious ones, whereas non-serious crimes on females peaked in 2022.

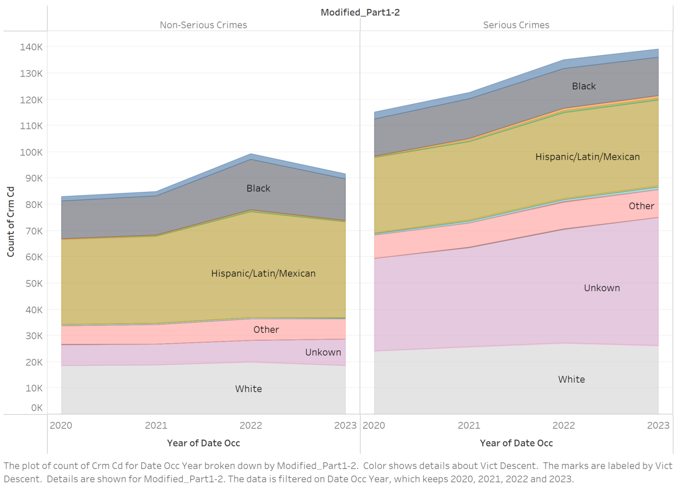


Fig. 27. Yearly Trends in Crime Seriousness by Victim Descent (2020-2023)

[Fig. 27] displays crime seriousness from 2020 to 2023, segmented by victim descent. Hispanic/Latin/Mexican victims are disproportionately affected by both serious and non-serious crimes compared to other descent groups. This pattern persists across the years, highlighting the vulnerability of this community.

The above visualization is useful for recognizing descent-based crime patterns, aiding in the development of targeted crime prevention and victim support initiatives for affected groups.

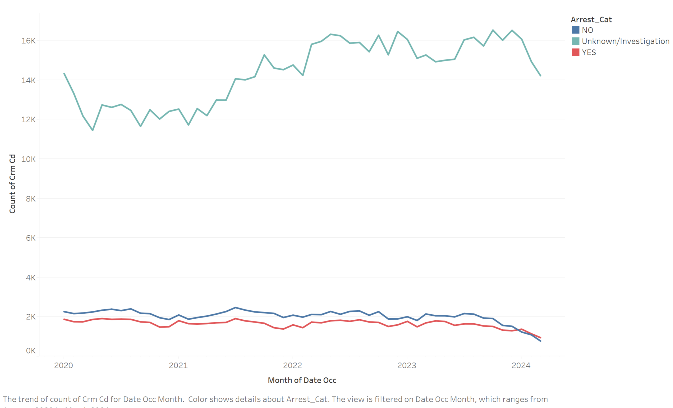


Fig. 28. Monthly Crime and Arrest Statistics

[Fig. 28] tracks the number of crimes and the status of criminals from January 2020 to March 2024 on a monthly basis, highlighting the breakdown of cases where arrests have been made, where investigations are ongoing, and where no action is recorded. The data shows that approximately 2,000 criminals have been arrested monthly, with a similar number still not arrested. A significant proportion of cases remain under investigation or lack clarity on the action taken.

This visualization is crucial for understanding the dynamics of crime resolution and identifying areas where investigative processes may be lagging.

Delaying case resolutions can lead to prolonged risks for public safety, as unresolved cases might continue to pose threats.Efficient case management is essential for maintaining trust in the criminal justice system and ensuring that offenders are held accountable promptly.

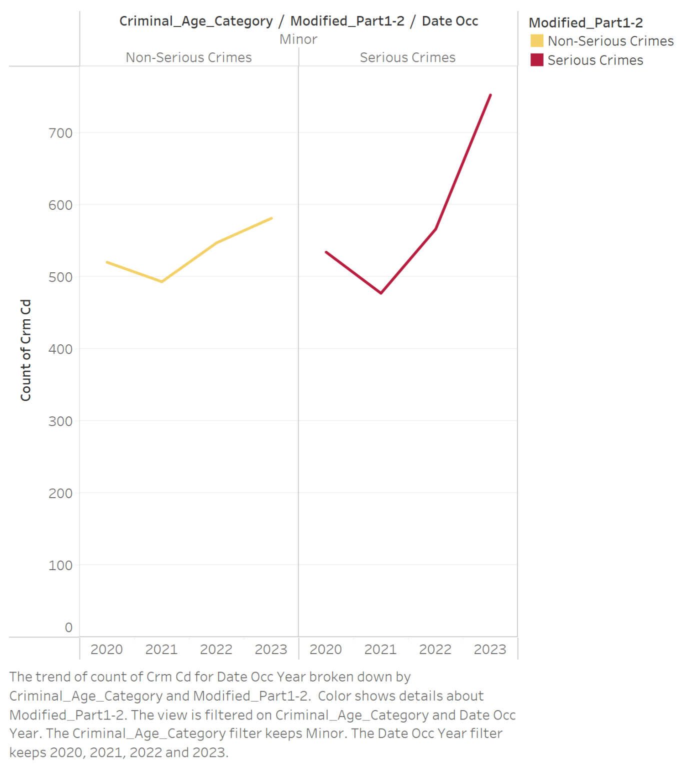


Fig. 29. Trend in Seriousness Crimes Committed by Minors(2020-2023)

[Fig. 29] displays the trend in the seriousness of crimes committed by minors from 2020 to 2023 on a yearly basis. The data indicates that serious crimes committed by minors hit their lowest point in 2021 but have been on the rise since then, with a notably steeper increase compared to non-serious crimes, which also show a rising trend.

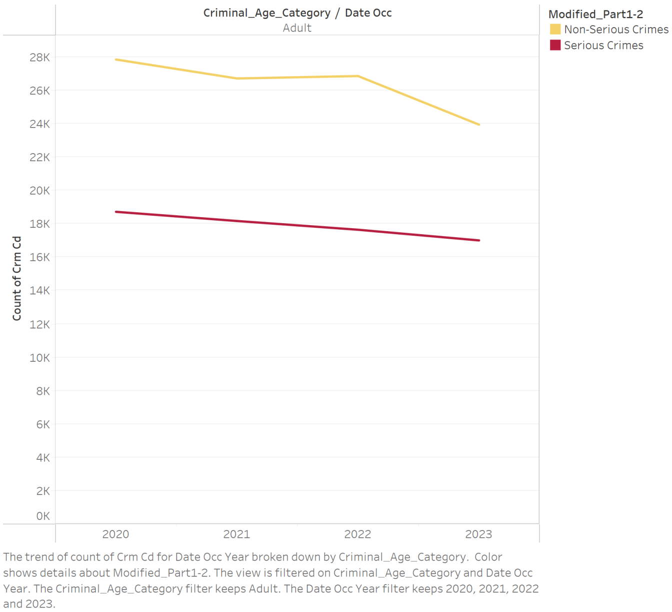


Fig. 30. Yearly Trends in Crime Seriousness Committed by Adults (2020-2023)

[Fig. 30] tracks the trend in the seriousness of crimes committed by adults from 2020 to 2023 on a yearly basis. It reveals that non-serious crimes significantly outnumber serious crimes committed by adults. Additionally, there is a noticeable decline in the number of both serious and non-serious crimes committed by adults over the years.

The series of visualizations and plots presented offer a comprehensive analysis of crime trends in Los Angeles, highlighting key insights into the seriousness of crimes, reporting patterns, weapons used and demographic impacts.

1. Reporting Timeliness: The analysis reveals that crimes are generally reported quickly, with most reports made within the first week. However, some cases experience significant delays, which can skew averages and impact the perception of crime trends.

2. Crime Seriousness and Demographics: Serious crimes are observed to affect males more significantly than females, with non-serious crimes showing a higher prevalence among females. Notably, Hispanic/Latin/Mexican victims are disproportionately impacted by both serious and non-serious crimes.

3. Weapon Usage: Most crimes involve no weapon, but when weapons are used, strong arms, body, or fists are the most common. This pattern is consistent among both minors and adults, although minors tend to use these methods more frequently in serious crimes.

4. Juvenile vs. Adult Crime Trends: The data indicates a troubling increase in serious crimes committed by minors since 2021, contrasted with a decrease in both serious and non-serious crimes committed by adults. This suggests a shift in crime patterns that requires targeted intervention strategies.

5. Crime Resolution and Investigation: There is a significant number of cases with ongoing investigations or unknown action taken against criminals. The consistent number of arrests and unarrested criminals each month underscores the need for more effective case management and resolution strategies.

Overall, these insights are essential for understanding crime dynamics and developing strategies for crime prevention, victim support, and more effective law enforcement. Addressing the issues identified, such as the rise in serious crimes among minors and the need for better data on crime resolution, will be crucial for improving community safety and justice outcomes.

C. *Task 3: Victim Demographics Analysis*

The goal of this task is to analyze the demographics of Los Angeles crime victims from 2000 to May 2024, focusing on age, gender, and descent. By examining these attributes, we aim to uncover trends in victimization across different demographic groups and how they are affected by various crimes. This section presents visualizations and statistics to highlight trends and correlations between demographics and crime types.

*1) **Victim Age** Analysing trends based on age of the victims:* We begin by analysing the correlation between victim age and the crimes committed. The correlation is not fully explored in this section and is further covered while looking at victim sex and descent.

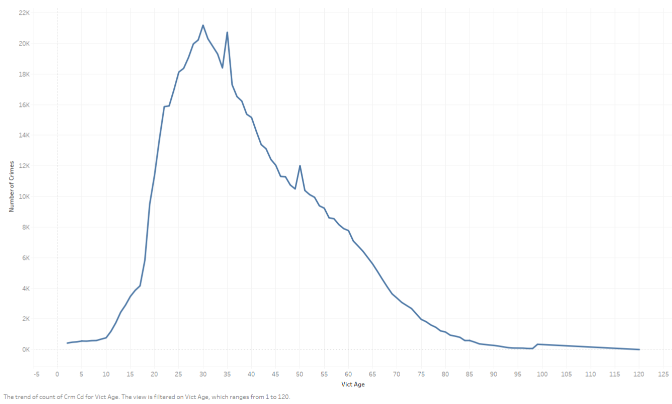
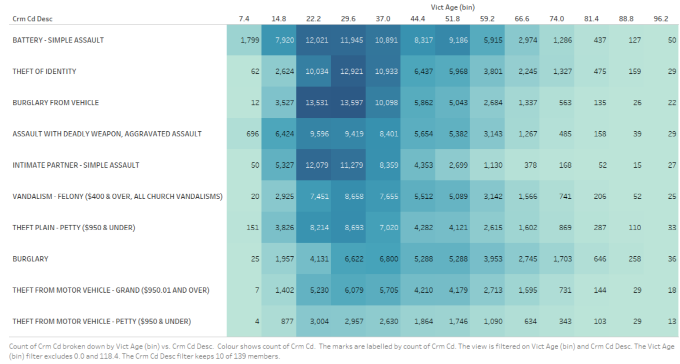


Fig. 31. Crime Victims by Age

Figure 31 is a line chart depicts the number of crimes in relation to the age of victims, revealing trends in victimization across various age groups. The sudden peaks at age 35 and 50 could be due to to inaccurate reporting.



	Vict Age (bin)												
Crm Cd Desc	7.4	14.8	22.2	29.6	37.0	44.4	51.8	59.2	66.6	74.0	81.4	88.8	96.2
BATTERY - SIMPLE ASSAULT	1,799	7,920	12,021	11,945	10,891	8,317	9,186	5,915	2,974	1,286	437	127	50
THEFT OF IDENTITY	62	2,624	10,034	12,921	10,933	6,437	5,968	3,801	2,245	1,327	475	159	29
BURGLARY FROM VEHICLE	12	3,527	13,531	13,597	10,098	5,862	5,043	2,684	1,337	563	135	26	22
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	696	6,424	9,596	9,419	8,401	5,654	5,382	3,143	1,267	485	158	39	29
INTIMATE PARTNER - SIMPLE ASSAULT	50	5,327	12,079	11,279	8,359	4,353	2,699	1,130	378	168	52	15	27
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	20	2,925	7,451	8,658	7,655	5,512	5,089	3,142	1,566	741	206	52	25
THEFT PLAIN - PETTY ($950 & UNDER)	151	3,826	8,214	8,693	7,020	4,282	4,121	2,615	1,602	869	287	110	33
BURGLARY	25	1,957	4,131	6,622	6,800	5,288	5,288	3,953	2,745	1,703	646	258	36
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)	7	1,402	5,230	6,079	5,705	4,210	4,179	2,713	1,595	731	144	29	18
THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	4	877	3,004	2,957	2,630	1,864	1,746	1,090	634	343	103	29	13

Count of Crm Cd broken down by Vict Age (bin) vs. Crm Cd Desc. Colour shows count of Crm Cd. The marks are labelled by count of Crm Cd. The view is filtered on Vict Age (bin) and Crm Cd Desc. The Vict Age (bin) filter excludes 0.0 and 118.4. The Crm Cd Desc filter keeps 10 of 139 members.

Fig. 32. Heatmap of Age and Crime Types

Figure 32 is a heatmap which allows us to understand the relation between age and victim for 10 of the most frequent crimes. We can see for certain crimes such as "Intimate Partner - Sexual Assault", there is a highly concentrated victim age while for some such as "Burglary" it is spread out.

*2) **Victim Sex** Analysing trends based on sex of the victims:* Now that we have looked at how age influences various aspects of the crime, lets look at gender. As a general legend for all plots, "M" label stands for male, "F" stands for female and "X" is unknown.

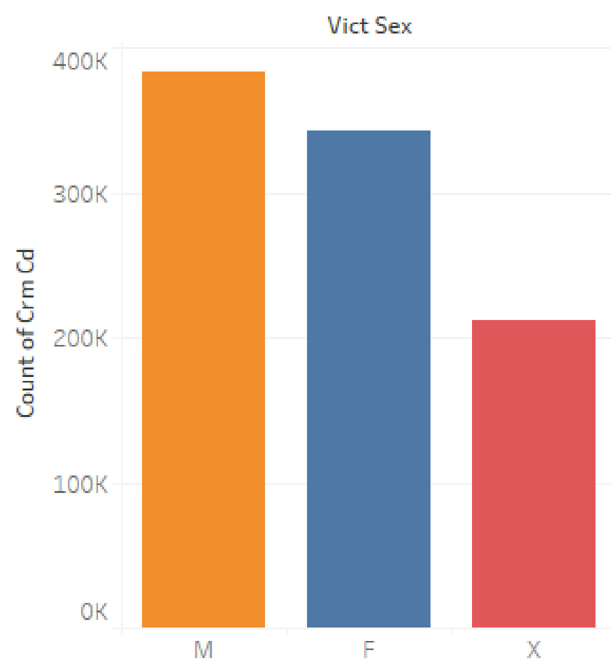


Fig. 33. Number of Crimes by Gender

In Fig. 33, the bar chart displays the number of crimes committed, categorized by the gender of the victims. Each bar represents the total number of crimes associated with male, female, and unknown gender victims. We can see there were more crimes committed against Males than Females. Also we can see that there is a significant number unknown values which will need to be excluded for further visualisations.

Let us now look at the ratios of male to female crime in various areas. In many of the areas the ratio is almost 1:1 these points are excluded and we'll look at points with a measurable difference in ratio.

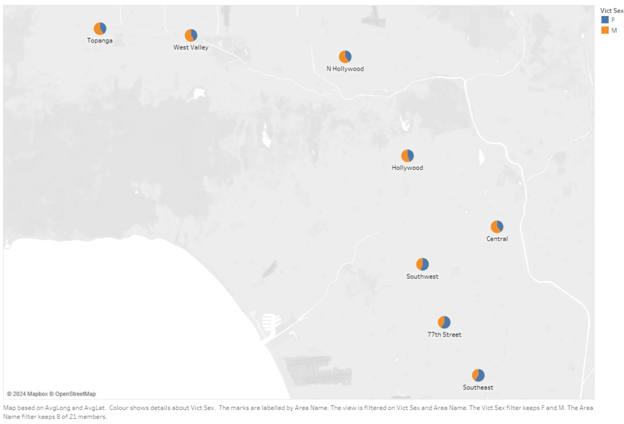


Fig. 34. Map with pie charts showing ratio of male and female victims

We see that there areas leaning towards both sides however, sides with more crimes against females are more significant as the overall dataset has more males. There can be many reasons for this imbalance between areas. This will depend on the factors that affect Male and Female crimes. We will now visualise other fields to find such factors.

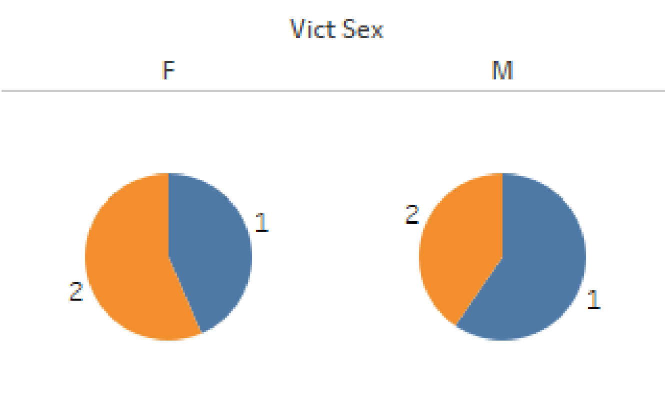


Fig. 35. Part of Crime by Gender

Fig. 35 is a pie chart displaying how many crimes of part 1 and part 2 are committed against Males and Females. Part 1 crimes refer to more serious and violent crimes. Part 2 crimes are less serious crimes as explained earlier. We see that males face a majority of serious crimes while its the opposite for females. If we look at the Figure 9, we see that the areas with more female victims had more part 2 crimes.

Now we can look at what part 1 and part 2 crimes are faced by males and females.

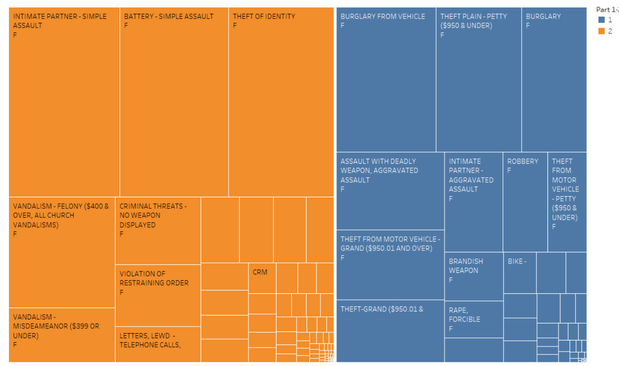


Fig. 36. Crimes Faced by Females in Part 1 and Part 2

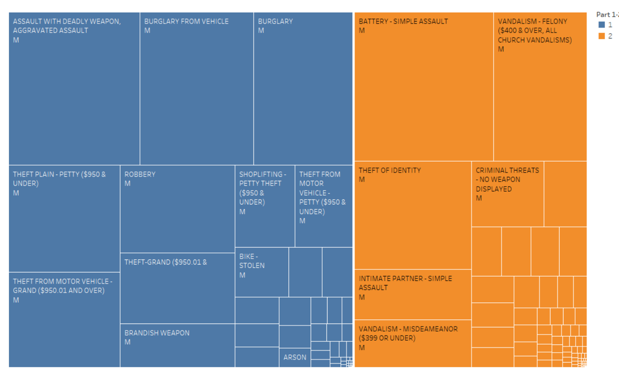


Fig. 37. Crimes Faced by Males in Part 1 and Part 2

We get some very interesting details from these two charts. We can see that females face more "Simple Assaults" Which refers to just threatening instead of a physical altercation. Males face more violent crimes such as Aggravated Assault.

We also see crimes that have a huge frequency in females but are barely present in the other. Such as various "Intimate Partner" crimes and "Rape - forcible".

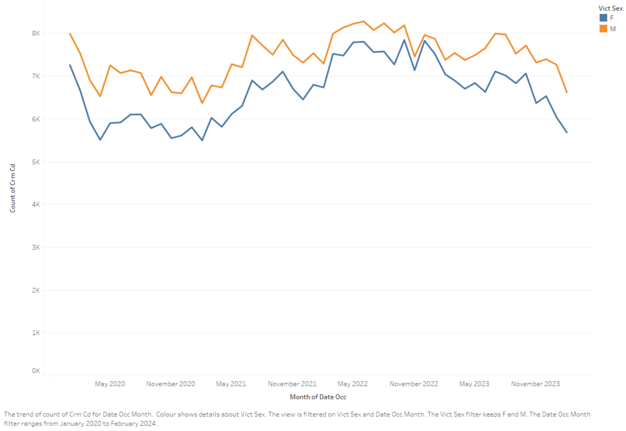


Fig. 38. Gender of Victims over Time

We can now analyse the trend of crime against both genders from 2020 to 2024. We see a dip in both during 2020. This can be reasoned by the pandemic. We then see an rise in both in 2022 although the rise is more in females. We can try to see why this unequal rise happens.

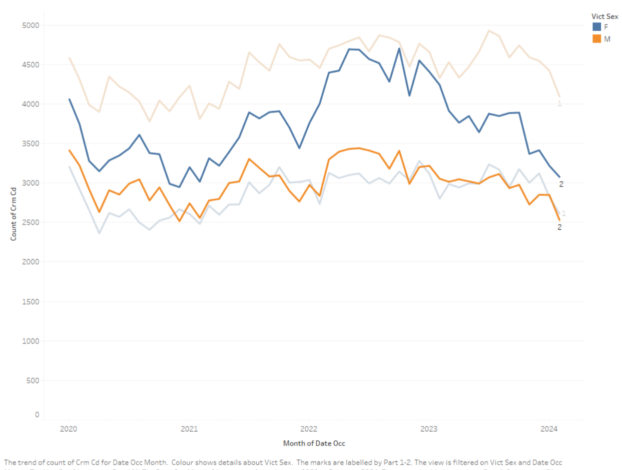


Fig. 39. Gender of Victims by Part over Time

In figure 39, we see that in the time when the spike happened, part 1 crimes against females remained stable but there was a huge spike in number of part 2 crimes. This spike is observed at a smaller level in males too. This could be the reason for the bigger rise in females after the dip.

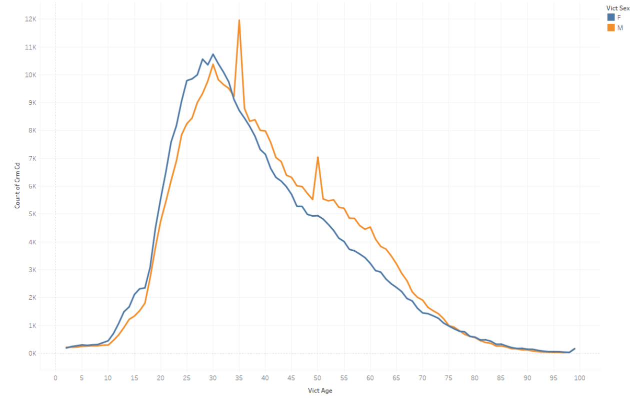


Fig. 40. Age of Victims by Gender

In figure 40 we see the relation between age and gender of victims. We see that there's actually more female victims than male up to the age of 33.

We also see spikes in males at the ages of 35 and 50 which corresponds to the spikes in the combined age graph. We see that there are so many more male victims over the age of 33, they become greater in number overall.

*3) **Victim Descent** Trends based on descent of the victims:* In this section we look at the cultural origin of the victims. This data could be helpful in seeing if certain communities are being targeted and what communities need to be protected better. We first look at the communities we have in the dataset. The word "Community" has been used to refer to the values of Descent.

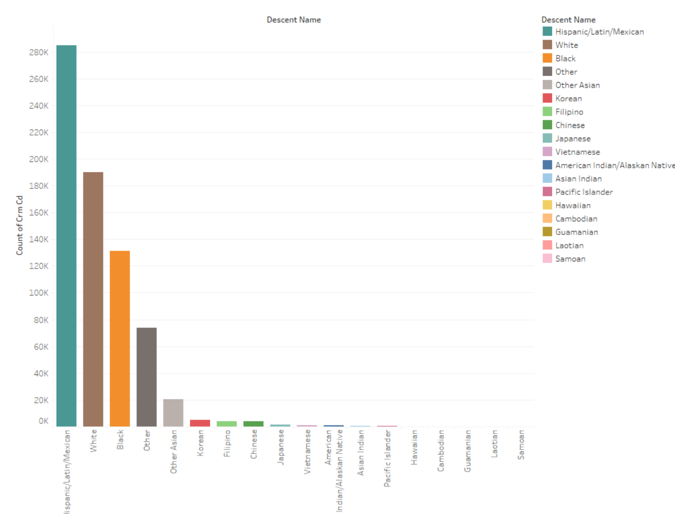


Fig. 41. Communities Present in the Dataset

In figure 41, we see the various communities present. There's 3 very dominant values here.

"Hispanic/Latin/Mexican", "White" and "Black". Further visualisation can be done on these 3.

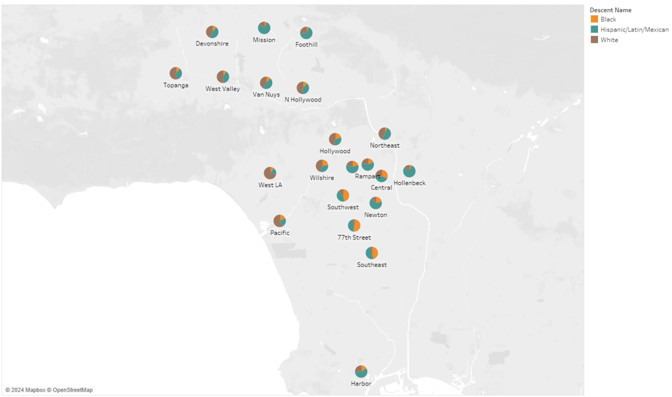


Fig. 42. Descent Based Analysis of Each Area

Figure 42 contains pie charts showing the split of the main 3 communities for each area. This could be due to different levels of population of each community in different areas. There are extremely different pie charts for different areas, for example "Southeast" and "Pacific". This shows the geographical and cultural variation of victims.

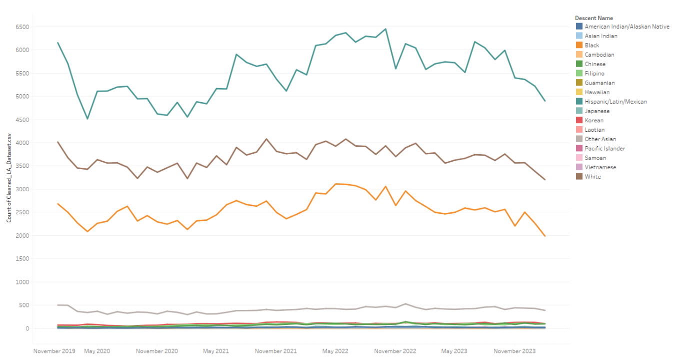


Fig. 43. Descent Based Analysis Over Time

In figure 43, we can see the number of victims by descent from 2020 to 2024. We see that there is a slump in "Hispanic/Latin/Mexican" victims that matches the slump in the overall data. However, the quantity of rest of the values were not affected or barely affected over that period. We also see a dip towards the end in all communities.

Fig. 44. Descent Based Analysis of Part

Figure 44 shows the pie charts for part 1 and 2 for various Descent values. We see that a lot of the communities have a heavy majority of part 1 i.e serious crimes. There are only 3 communities, "Black", "Hispanic/Latin/Mexican", "Laotian", that have a part 2 majority.

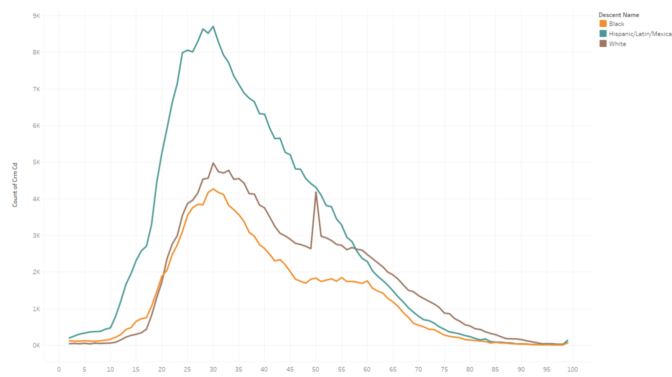


Fig. 45. Age based analysis of Descent

In figure 45, we can see the age distribution for the main 3 communities. We see that "Hispanic/Latin/Mexican" has a peak early then falls off. The other 2 communities are more spread out. We see a sudden peak at 50 for "White". This is seen in the overall age chart. However there is no peak in these communities at 30. Thus that peak must be cause by unknown or "Other" Descent Values.

IV. Author's contributions

- **Subham Agarwala:** visualization and analysis for Task 1
- **Sarvesh Kumar:** visualization and analysis for Task 2
- **Ayush Gupta:** visualization and analysis for Task 3
  Preprocessing of data was a cumulative contribution by all team members.

V. References

References

[1] Kaggle Crime Data. *Los Angeles and Chicago Crime Data from 2000*. Available at: https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000?resource=download&select=Chicage+Crime+Data.csv.