Crime Analysis of California.

By Rohit Raman

Date: 10/03/2025- 16/06/2025

Type of column in the datasets.

Business Questions:

1. What factors influence whether a crime case in California is closed or remains open?

2. What are the most prevalent crime types in California, and how are they distributed across different times of the day?

3. Which areas in California are the top crime hotspots, and what are their peak crime hours and dominant crime types?

4. What is the frequency of serious versus less serious crimes, and what are their yearly trends, hourly distribution, and hotspot areas?

5. What are the demographic characteristics of crime victims in California, and how do they relate to crime timing?

6. What are the peak hours, days, and seasons for crime occurrence in California?

7. What are the most common weapons used in crimes, and how do they vary by crime type and geographic area?

8. What are the prevalent modus operandi (MO) codes, and how frequently do they occur?

9. What is the closure rate for different crime types, and which crimes are most challenging to resolve?

10. Can distinct crime patterns or groups be identified based on victim age and crime type to support targeted prevention strategies?

11. What is the frequency of crime by hours?

12. What is the frequency of crime by Day of Week?

13. What is the crime frequency by Season?

14. What are the top 10 crime areas?

15. What are the top 10 most frequent crimes and what are the monthly trends of those crimes?

16. What are the top 10 crime areas; what are the peak hour crime in that area and what all type crimes happen over there?

17. What is the demographic of victims?

18. What is the relationship between the victim's age and the time of the crime?

19. What is the frequency of serious versus less serious crimes in California? Provide the yearly trends for both categories, along with their hourly distribution and the areas where these crimes most commonly occur.

20. What is the most common weapon used in the crime, what kind of weapon used in what type of crime, and weapon used in the Geographical area?

21. What is the modus operandi, and how many crimes have occurred under this code?

22. Analyze the rate at which crimes are being closed or resolved.

23. What types of crimes are most prevalent in the city, and how are they distributed across different times of the day?

24. Can we identify distinct patterns or groups of crimes based on victim age and crime type to support targeted crime prevention strategies?

1. **Basic Overview of the datasets.**

```
[ ] df.columns

Index(['DR_NO', 'Date Rptd', 'DATE OCC', 'TIME OCC', 'AREA', 'AREA NAME',
       'Rpt Dist No', 'Part 1-2', 'Crm Cd', 'Crm Cd Desc', 'Mocodes',
       'Vict Age', 'Vict Sex', 'Vict Descent', 'Premis Cd', 'Premis Desc',
       'Weapon Used Cd', 'Weapon Desc', 'Status', 'Status Desc', 'Crm Cd 1',
       'Crm Cd 2', 'Crm Cd 3', 'Crm Cd 4', 'LOCATION', 'Cross Street', 'LAT',
       'LON'],
      dtype='object')
```

Visual 1: Type of column in the dataset

The above visual 1 represents the type of column we have in the datasets. These are the column we have:

- **DR_NO**: Report number (unique identifier for the case)
- **Date Rptd**: Date the crime was reported
- **DATE OCC**: Date the crime occurred
- **TIME OCC**: Time the crime occurred
- **AREA**: Numeric code for the geographical area

- **AREA NAME**: Name of the geographical area
- **Rpt Dist No**: Reporting district number
- **Part 1-2**: Classification of the crime (Part 1 includes serious crimes like homicide, while Part 2 includes less severe offenses)
- **Crm Cd**: Crime code
- **Crm Cd Desc**: Description of the crime
- **Mocodes**: Modus operandi codes (method used to commit the crime)
- **Vict Age**: Age of the victim
- **Vict Sex**: Sex of the victim
- **Vict Descent**: Ethnic or racial background of the victim
- **Premis Cd**: Premises code (type of location where the crime occurred)
- **Premis Desc**: Description of the premises
- **Weapon Used Cd**: Code for the weapon used
- **Weapon Desc**: Description of the weapon used
- **Status**: Status of the case (e.g., open, closed, pending)
- **Status Desc**: Description of the case status
- **Crm Cd 1, 2, 3, 4**: Additional crime codes if multiple offenses occurred
- **LOCATION**: Address or coordinates of the crime scene
- **Cross Street**: Nearest cross street
- **LAT**: Latitude of the crime location
- **LON**: Longitude of the crime location

2. Missing values in the datasets.

```
df.isnull().sum()
```

	0
DR_NO	0
Date Rptd	0
DATE OCC	0
TIME OCC	0
AREA	0
AREA NAME	0
Rpt Dist No	0
Part 1-2	0
Crm Cd	0
Crm Cd Desc	0
Mocodes	151741
Vict Age	0
Vict Sex	144765
Vict Descent	144777
Premis Cd	16
Premis Desc	588
Weapon Used Cd	677885
Weapon Desc	677885
Status	1
Status Desc	0
Crm Cd 1	11
Crm Cd 2	935996
Crm Cd 3	1002835
Crm Cd 4	1005085
LOCATION	0
Cross Street	850909
LAT	0
LON	0

dtype: int64

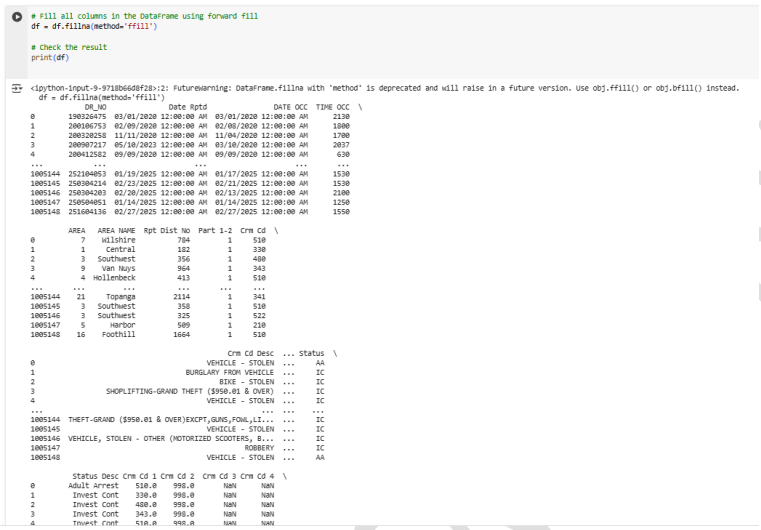
Visual 2: Checking the missing value

The dataset provided has a mix of complete and incomplete fields, with several columns containing a significant number of null (missing) values. Key columns such as DR_NO, Date Rptd, DATE OCC, TIME OCC, AREA, AREA NAME, Rpt Dist No, Part 1-2, Crm Cd, Crm Cd Desc, Vict Age, LOCATION, LAT, and LON are fully populated, which is advantageous for core crime analysis. However, there are substantial gaps in several other important fields. For instance, Vict Sex and Vict Descent are missing in over 144,000 records, likely due to privacy concerns or incomplete victim reporting. The Mocodes field, which may indicate modus operandi codes, is also missing in over 150,000 entries, suggesting it may not be recorded for all crime types. Fields related to weapons—Weapon Used Cd and Weapon Desc—are missing in over 677,000 rows, implying that weapons were either not used or not reported in a majority of cases. Additionally, the Crm Cd 2, Crm Cd 3, and Crm Cd 4 fields, which allow for recording multiple crime codes per incident, are largely empty, likely because most incidents are classified under a single primary code. The Cross Street field also has a very high number of missing values, limiting spatial granularity. Despite some missing data in Premis Cd and Premis Desc, these are relatively minor and manageable. Overall, while the dataset is strong

in foundational crime reporting data, it would benefit from careful handling of missing values, particularly for demographic and weapon-related analysis.

3. Filling the missing Values:

In visual 3: I used fillna(method='ffill') because the missing values in my dataset aren't actually unknown—they just weren't repeated in the current rows since the previous records already had that information. In many cases, data is structured so that repeated details are only listed once and left blank afterward to avoid redundancy. By applying forward fill, I make sure that these missing values are filled with the last valid entry, which helps keep my dataset complete and consistent without losing important information.



```
# Fill all columns in the DataFrame using forward fill
df = df.fillna(method='ffill')

# Check the result
print(df)
```

```
<ipython-input-9-9710b66d8f28>:2: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj.bfill() instead.
  df = df.fillna(method='ffill')
           DR_NO            Date Rptd                 DATE OCC  TIME OCC  \
0       190326475  03/01/2020 12:00:00 AM  03/01/2020 12:00:00 AM      2130
1       200106753  02/09/2020 12:00:00 AM  02/08/2020 12:00:00 AM      1800
2       200320258  11/11/2020 12:00:00 AM  11/04/2020 12:00:00 AM      1700
3       200907217  05/10/2023 12:00:00 AM  03/10/2020 12:00:00 AM      2037
4       200412582  09/09/2020 12:00:00 AM  09/09/2020 12:00:00 AM       630
...           ...                     ...                     ...       ...
1005144  252104053  01/19/2025 12:00:00 AM  01/17/2025 12:00:00 AM      1530
1005145  250304214  02/23/2025 12:00:00 AM  02/21/2025 12:00:00 AM      1530
1005146  250304203  02/20/2025 12:00:00 AM  02/13/2025 12:00:00 AM      2100
1005147  250504051  01/14/2025 12:00:00 AM  01/14/2025 12:00:00 AM      1250
1005148  251604136  02/27/2025 12:00:00 AM  02/27/2025 12:00:00 AM      1550

         AREA   AREA NAME  Rpt Dist No  Part 1-2  Crm Cd  \
0          7    Wilshire          784         1     510
1          1     Central          182         1     330
2          3   Southwest          356         1     480
3          9     Van Nuys          964         1     343
4          4  Hollenbeck          413         1     510
...      ...         ...          ...       ...     ...
1005144   21     Topanga         2114         1     341
1005145    3   Southwest          358         1     510
1005146    3   Southwest          325         1     522
1005147    5     Harbor          509         1     210
1005148   16    Foothill         1664         1     510

                                    Crm Cd Desc  ... Status  \
0                                VEHICLE - STOLEN  ...     AA
1                              BURGLARY FROM VEHICLE  ...     IC
2                                   BIKE - STOLEN  ...     IC
3                  SHOPLIFTING-GRAND THEFT ($950.01 & OVER)  ...     IC
4                                VEHICLE - STOLEN  ...     IC
...                                           ...  ...    ...
1005144  THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LI...  ...     IC
1005145                          VEHICLE - STOLEN  ...     IC
1005146  VEHICLE, STOLEN - OTHER (MOTORIZED SCOOTERS, B...  ...     IC
1005147                                  ROBBERY  ...     IC
1005148                          VEHICLE - STOLEN  ...     AA

        Status Desc  Crm Cd 1  Crm Cd 2  Crm Cd 3  Crm Cd 4  \
0       Adult Arrest     510.0     998.0       NaN       NaN
1        Invest Cont     330.0     998.0       NaN       NaN
2        Invest Cont     480.0     998.0       NaN       NaN
3        Invest Cont     343.0     998.0       NaN       NaN
4        Invest Cont     510.0     998.0       NaN       NaN
```

Visual 3: Filling the missing values.

```
df.isnull().sum()
```

	0
DR_NO	0
Date Rptd	0
DATE OCC	0
TIME OCC	0
AREA	0
AREA NAME	0
Rpt Dist No	0
Part 1-2	0
Crm Cd	0
Crm Cd Desc	0
Mocodes	1
Vict Age	0
Vict Sex	0
Vict Descent	0
Premis Cd	0
Premis Desc	0
Weapon Used Cd	44
Weapon Desc	44
Status	0
Status Desc	0
Crm Cd 1	0
Crm Cd 2	0
Crm Cd 3	246
Crm Cd 4	246
LOCATION	0
Cross Street	6
LAT	0
LON	0

```
# Drop rows with any null (NaN) values in the DataFrame
df = df.dropna()

df.isnull().sum()
```

	0
DR_NO	0
Date Rptd	0
DATE OCC	0
TIME OCC	0
AREA	0
AREA NAME	0
Rpt Dist No	0
Part 1-2	0
Crm Cd	0
Crm Cd Desc	0
Mocodes	0
Vict Age	0
Vict Sex	0
Vict Descent	0
Premis Cd	0
Premis Desc	0
Weapon Used Cd	0
Weapon Desc	0
Status	0
Status Desc	0
Crm Cd 1	0
Crm Cd 2	0
Crm Cd 3	0
Crm Cd 4	0
LOCATION	0
Cross Street	0
LAT	0

Visual 4: Checking missin value after ffill and again droping the remaing missing values.

After filling the missing values using forward fill, I noticed that some rows still contained missing data across multiple columns, which didn't make sense. Since these rows were incomplete and potentially unreliable, I decided to drop them from the dataset.

Data type of columns:

```
df.info()
<class 'pandas.core.frame.DataFrame'>
Index: 1004903 entries, 246 to 1005148
Data columns (total 28 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   DR_NO           1004903 non-null  int64
 1   Date Rptd       1004903 non-null  object
 2   DATE OCC        1004903 non-null  object
 3   TIME OCC        1004903 non-null  int64
 4   AREA            1004903 non-null  int64
 5   AREA NAME       1004903 non-null  object
 6   Rpt Dist No     1004903 non-null  int64
 7   Part 1-2        1004903 non-null  int64
 8   Crm Cd          1004903 non-null  int64
 9   Crm Cd Desc     1004903 non-null  object
 10  Mocodes         1004903 non-null  object
 11  Vict Age        1004903 non-null  int64
 12  Vict Sex        1004903 non-null  object
 13  Vict Descent    1004903 non-null  object
 14  Premis Cd       1004903 non-null  float64
 15  Premis Desc     1004903 non-null  object
 16  Weapon Used Cd  1004903 non-null  float64
 17  Weapon Desc     1004903 non-null  object
 18  Status          1004903 non-null  object
 19  Status Desc     1004903 non-null  object
 20  Crm Cd 1        1004903 non-null  float64
 21  Crm Cd 2        1004903 non-null  float64
 22  Crm Cd 3        1004903 non-null  float64
 23  Crm Cd 4        1004903 non-null  float64
 24  LOCATION        1004903 non-null  object
 25  Cross Street    1004903 non-null  object
 26  LAT             1004903 non-null  float64
 27  LON             1004903 non-null  float64
dtypes: float64(8), int64(7), object(13)
memory usage: 222.3+ MB
```

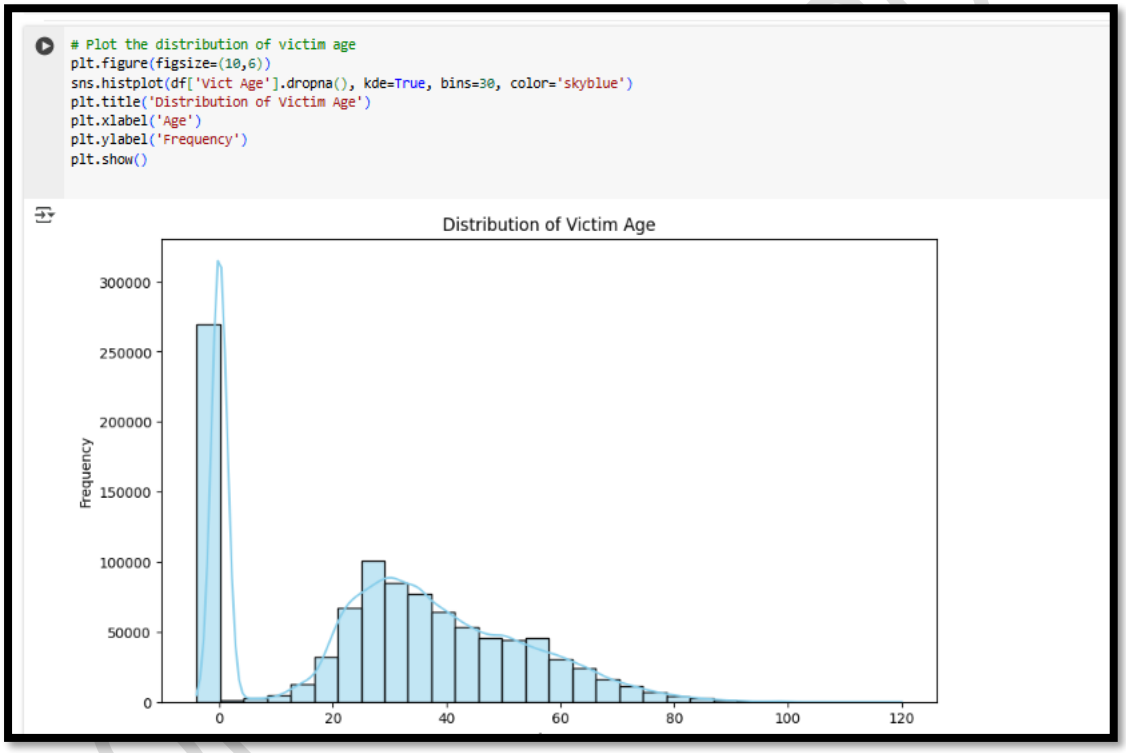
Visual 5: Data type of the column.

When I ran df.info() on my dataset, it gave me a summary of all the columns, their data types, and how many non-null values each one has. For example, some columns like DR_NO, DATE OCC, and TIME OCC are of type int64, meaning they store whole numbers. Columns like LAT and LON are float64 because they contain decimal values representing geographic coordinates. The columns such as AREA NAME, Crm Cd Desc, Status Desc, and LOCATION are of type object, which usually means they contain text or string data. Some date-related fields like Date Rptd or DATE OCC might also be stored as object if they haven't been converted to datetime format yet. By checking this info, I got a clear idea of the structure of my data and which columns might need cleaning or type conversion.

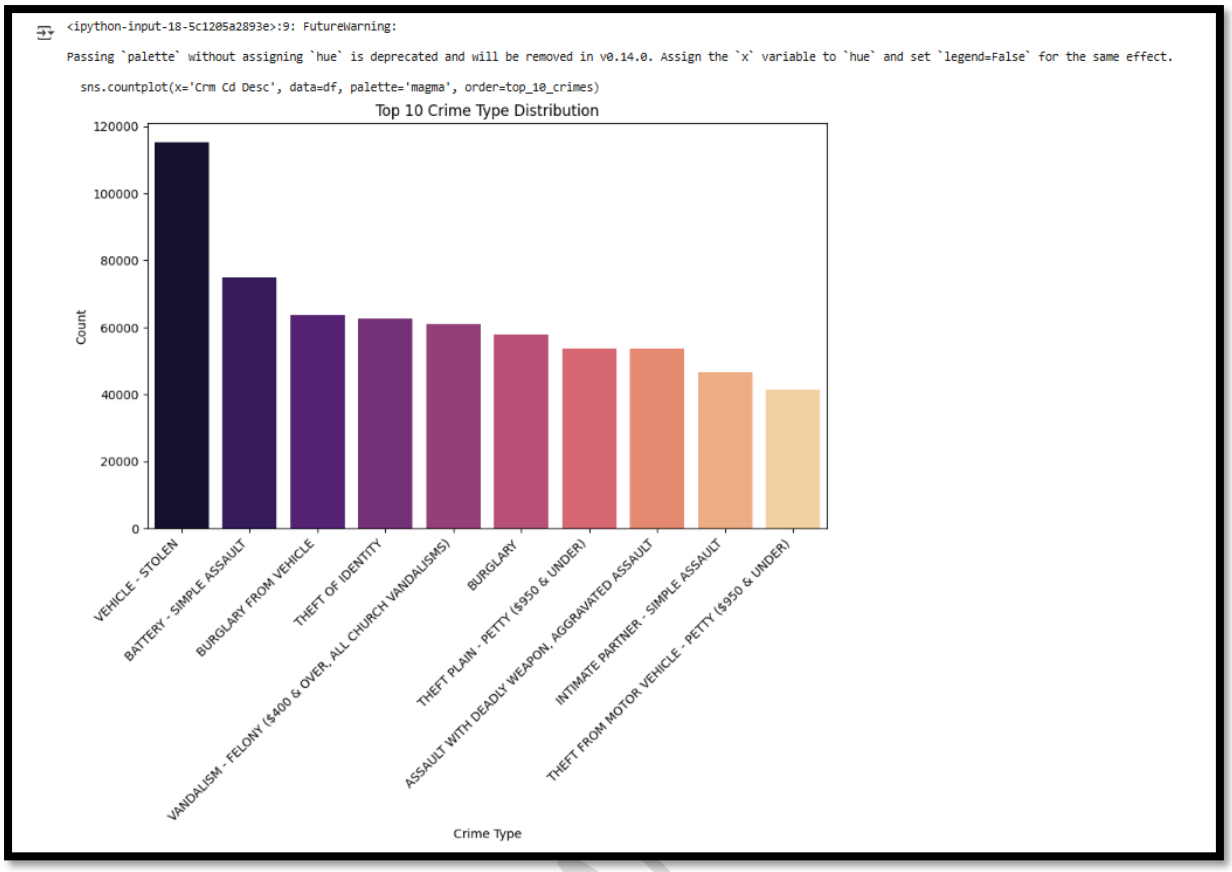
Data distribution:

Age Distribution of Victim

In California the victim who is subjected to crime range from 20-60 years.



Visual 6: Distribution of Victim Age.

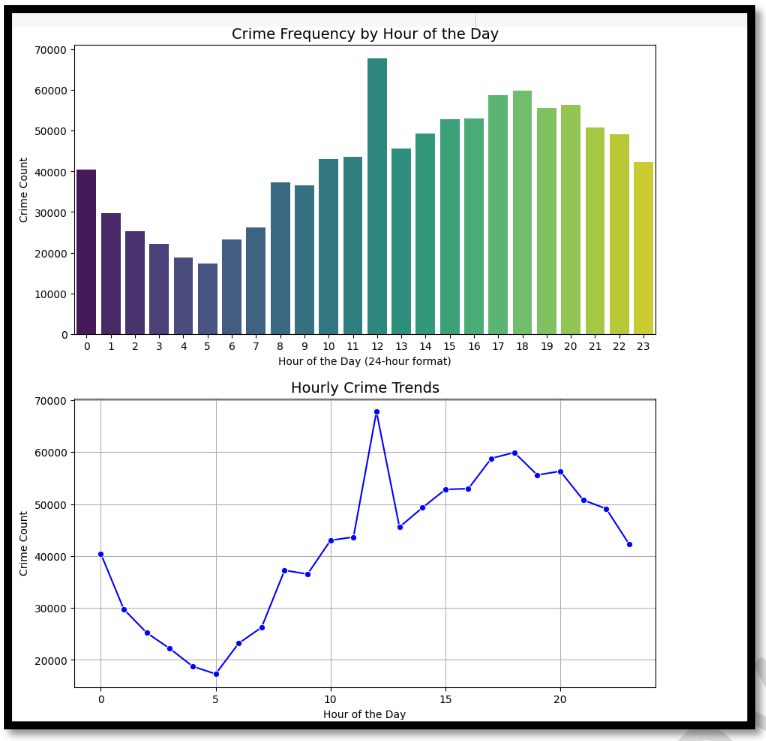


Visual 7: Crime of California city.

"Vehicle theft is the most common crime occurring in California, with a total of 110,000 reported cases. Second is simple assault, with around 700,000 reports. Third is burglary from vehicles, with approximately 600,000 cases. Fourth comes identity theft, vandalism, and burglary, each with around 580,000 reported incidents."

Temporal Analysis: Identifying Peak Crime Hours, Days, and Seasonal Trends:

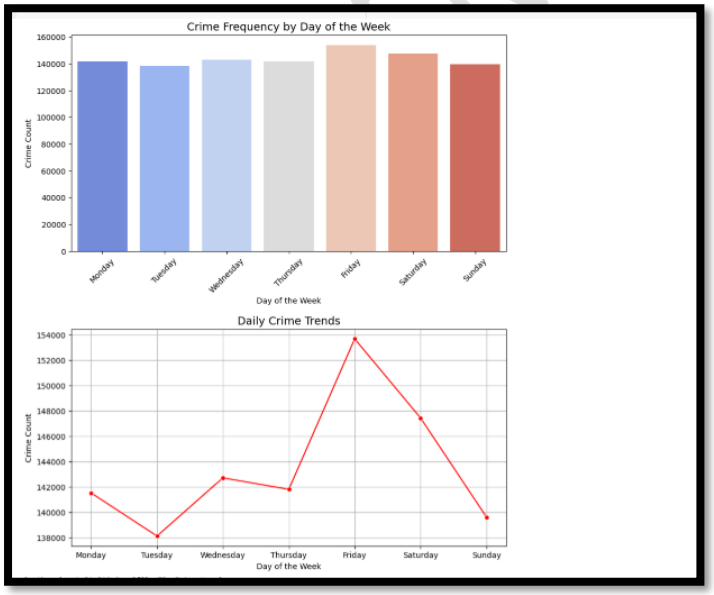
1. **What is the frequency of crime by hours???**



Visual 8 Peak crime Hours.

In Visual 8, I analyzed the peak crime hours. The bar and line charts show that the highest number of crimes occur between 10 AM and 3 PM, with crime activity beginning as early as 5 AM.

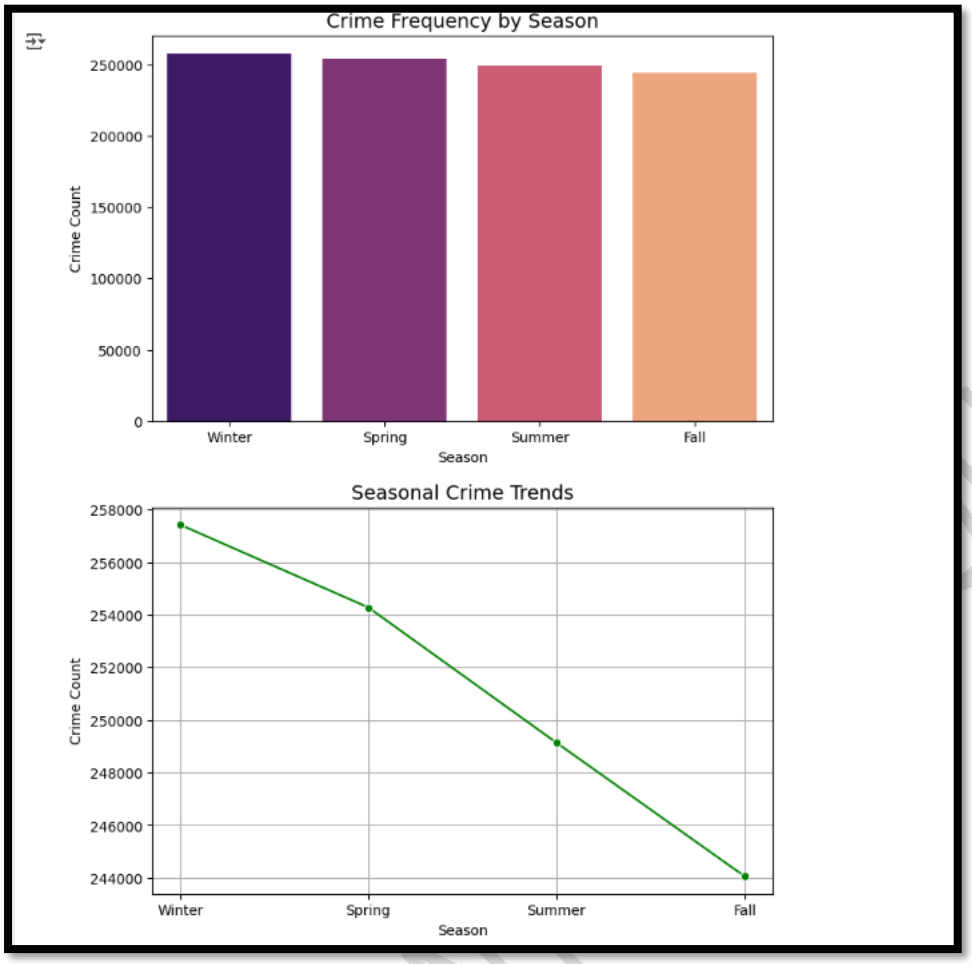
2. **What is the frequency of crime by Day of Week??**



Visual9: Crime Frequency by day of the week

From Visual 12, Thursday, Friday, and Saturday are the most prominent days when the highest number of crimes occur, with values of 142,000, 154,000, and 148,000 respectively.

3. **What is the crime Frequency by Season???**



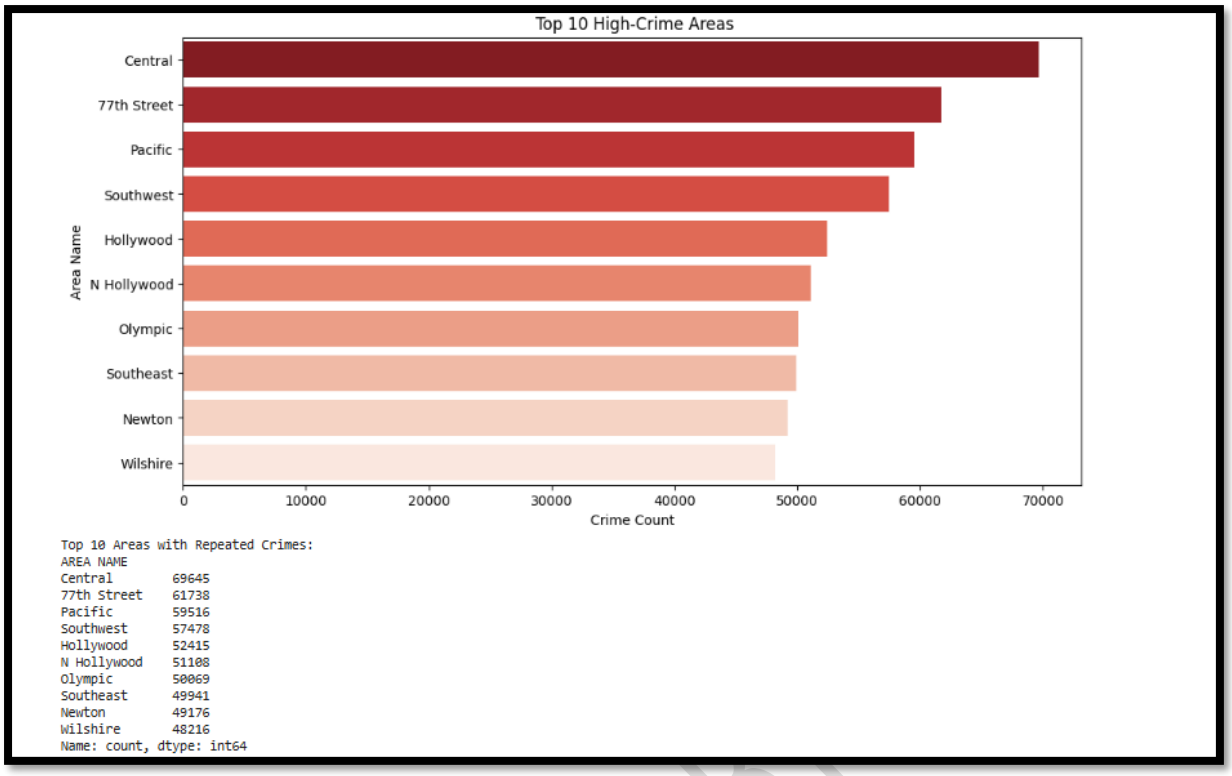
Visual 10: Peak crime Season

From Visual 10, it is evident that the highest number of crimes occur in winter, with a total of 257,500 cases, followed by spring and summer with approximately 254,000 and 249,500 cases, respectively.

Geospatial Crime Hotspot Analysis: Using Area Name, LAT and LON.

Identify crime hotspot based on latitude (LAT), longitude (LON) and Area Name.

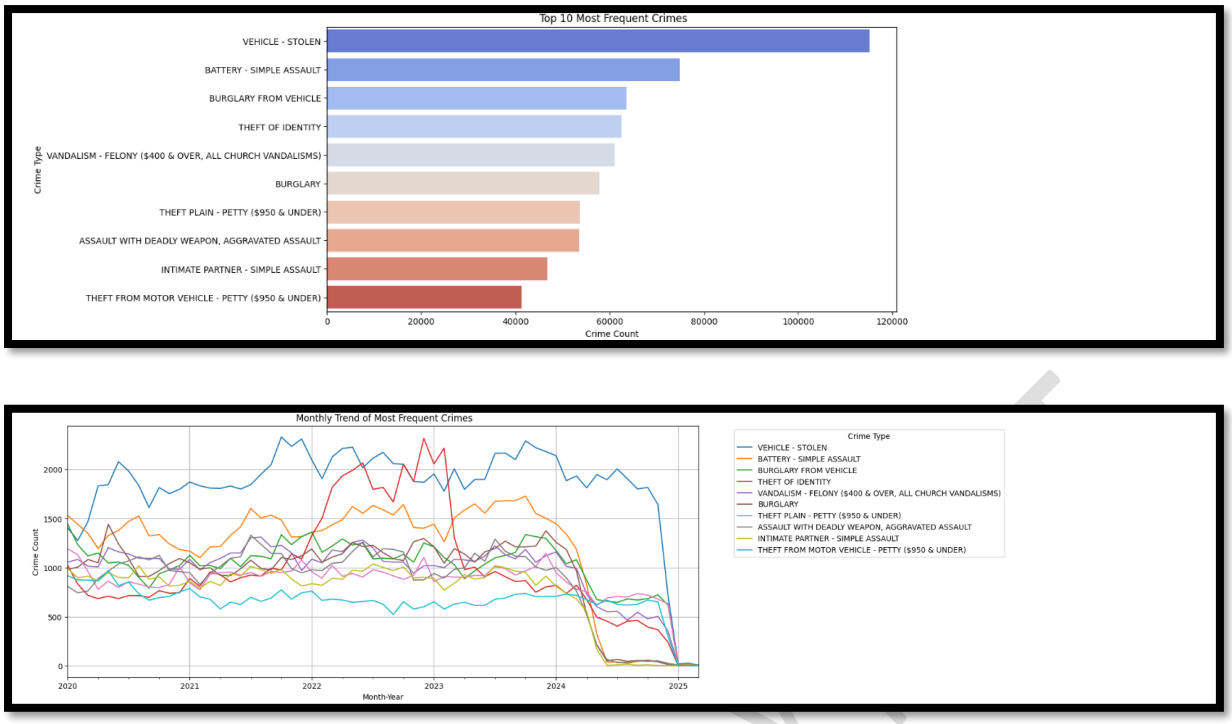
4. **What are the top 10 crime areas??**



Visual 11: Top 10 crime prone area.

From Visual 11: When I reviewed the crime data, **Central** topped the list with **69,645** reported incidents—it's the city's busiest area, so that tracks. **77th Street** followed with **61,738** crimes, known for gang-related issues and violent offenses. I was surprised that **Pacific** had **59,516** cases; despite its beachside vibe, theft and car break-ins are rampant. **Southwest** recorded **57,478** crimes, still struggling with gang presence and community safety. **Hollywood** came in at **52,415**, which made sense—tourists, nightlife, and drug activity keep crime levels high. **North Hollywood** had **51,108**, mostly property crimes and assaults. **Olympic** showed **50,069** cases, a mix of business and residential crimes, especially vehicle-related. **Southeast** reported **49,941** incidents—it's an underserved area where crime often goes hand in hand with poverty. **Newton** had **49,176**, with frequent police activity but stubbornly high crime. Finally, **Wilshire** rounded out the list with **48,216**, where even upscale areas deal with constant issues, especially at night. These numbers really opened my eyes to how localized and persistent crime can be.

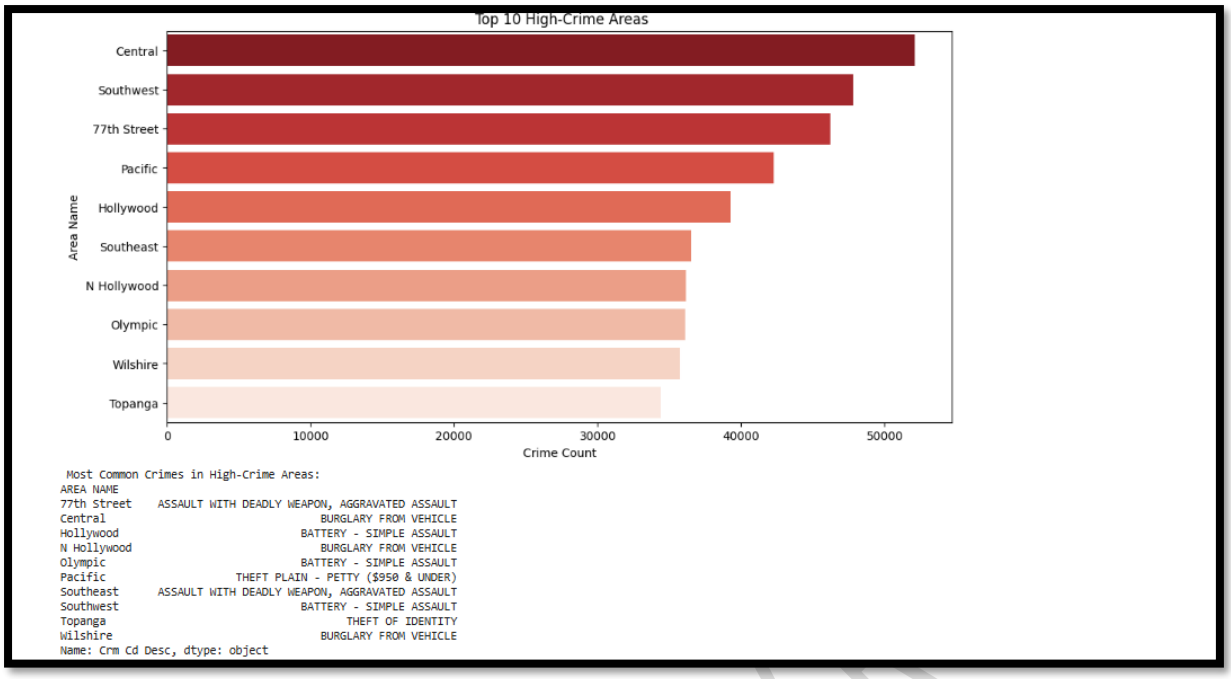
5. **What are the top 10 most frequent crimes and what are the monthly trends of those crimes???**



Visual 12: Top 10 most frequent crime and its yearly variations.

From Visual 12; When I looked at the crime data, I saw that **vehicle theft** is the most common crime with over **115,000** cases, which really surprised me. Next, there are about **74,800** reports of **simple assaults**, where people get hurt without weapons. **Break-ins into vehicles** happen a lot too, with over **63,500** incidents. I also noticed more than **62,500** cases of **identity theft**, which means people's personal information is being stolen frequently. **Felony vandalism**, including damage to churches, showed up over **61,000** times. Regular **burglaries** of homes or buildings are also common, with nearly **58,000** cases. There are many **petty thefts** under $950, around **53,700** reports, and serious **assaults with deadly weapons** numbering over **53,500**. Sadly, **intimate partner assaults** are high too, with over **46,700** cases. Finally, **petty theft from cars** happens frequently, with more than **41,300** reports. Seeing these numbers, it's clear which crimes affect people the most.

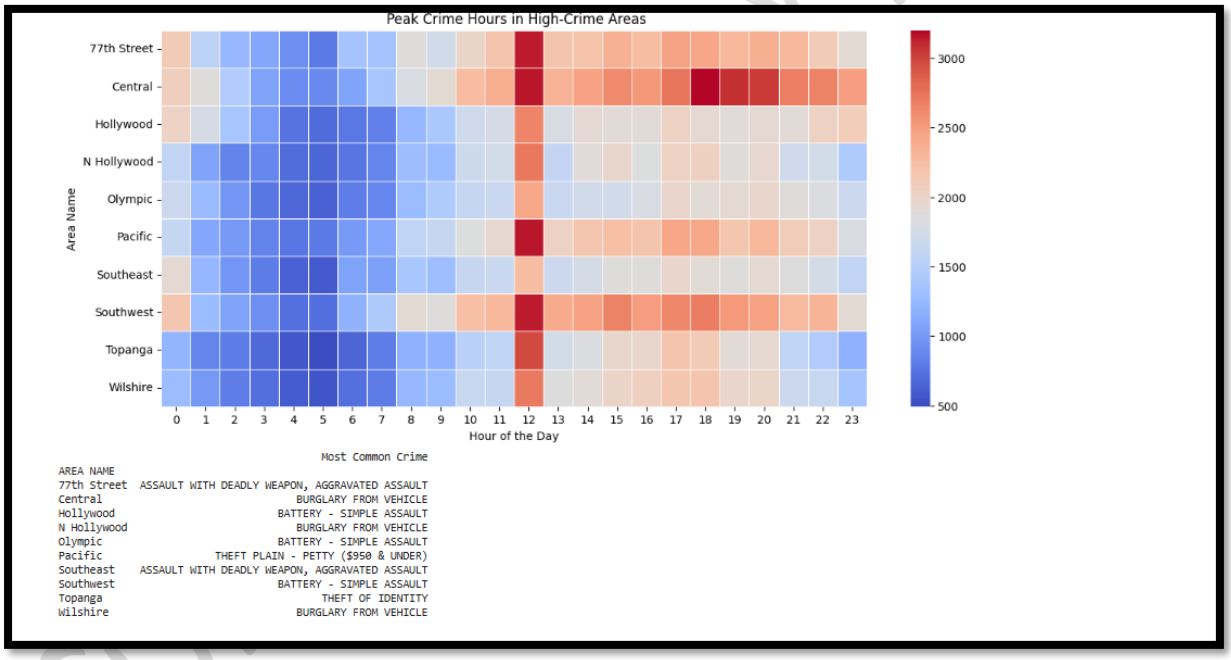
6. **What are the top 10 crime area; what are the peak hour crime in that area and what all type crime happens over there???**



Most Common Crimes in High-Crime Areas:

AREA NAME	
77th Street	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Central	BURGLARY FROM VEHICLE
Hollywood	BATTERY - SIMPLE ASSAULT
N Hollywood	BURGLARY FROM VEHICLE
Olympic	BATTERY - SIMPLE ASSAULT
Pacific	THEFT PLAIN - PETTY ($950 & UNDER)
Southeast	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Southwest	BATTERY - SIMPLE ASSAULT
Topanga	THEFT OF IDENTITY
Wilshire	BURGLARY FROM VEHICLE

Name: Crm Cd Desc, dtype: object



AREA NAME	Most Common Crime
77th Street	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Central	BURGLARY FROM VEHICLE
Hollywood	BATTERY - SIMPLE ASSAULT
N Hollywood	BURGLARY FROM VEHICLE
Olympic	BATTERY - SIMPLE ASSAULT
Pacific	THEFT PLAIN - PETTY ($950 & UNDER)
Southeast	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Southwest	BATTERY - SIMPLE ASSAULT
Topanga	THEFT OF IDENTITY
Wilshire	BURGLARY FROM VEHICLE

Visual 13: Overview of the Top 10 Crime-Prone Areas, Highlighting Their Peak Crime Hours and Most Common Types of Crimes.

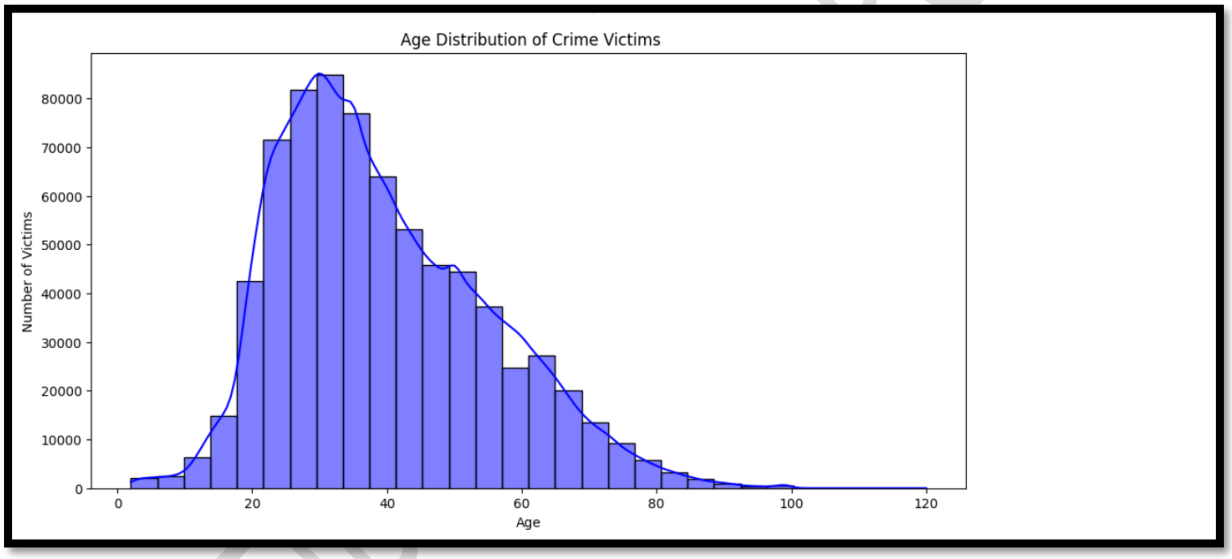
From the data and the bar graph from visual 13, it's clear that Central, Southwest, 77th Street, Pacific, and Hollywood are the top areas where the highest number of crimes occur. Each area also has its most frequent type of crime. For example, **77th Street** is dominated by **assault with a deadly weapon and aggravated assault**, making violent crime a serious concern there. In **Central**, the most common crime is **burglary from vehicles**, showing how vulnerable cars are in that area. **Hollywood** sees mostly **simple battery assaults**, meaning fights and minor physical attacks happen often. Similarly, **Southwest** and **Olympic** areas also

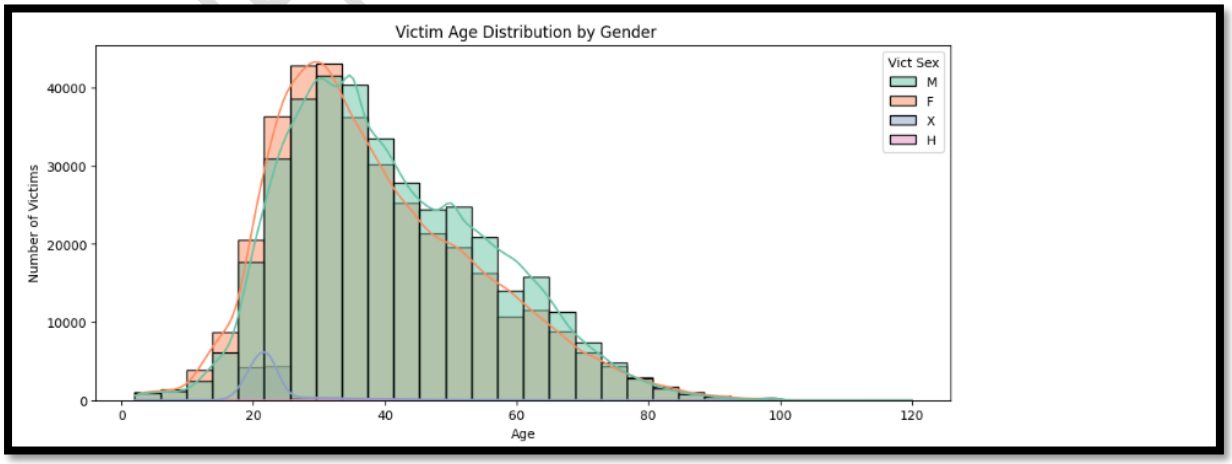
report a lot of **simple battery assaults**. In **North Hollywood** and **Wilshire**, **burglary from vehicles** is the leading crime, while **Pacific** experiences mainly **petty thefts under $950**. **Southeast** faces high levels of **assault with deadly weapons**, and **Topanga** struggles most with **identity theft**.

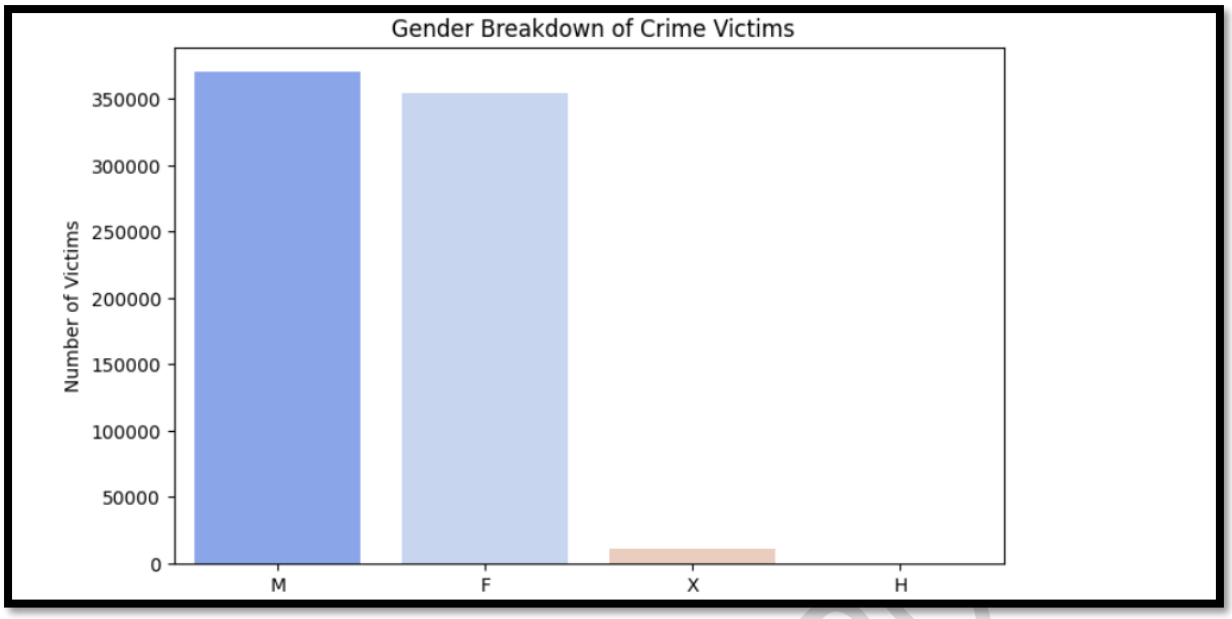
Looking at the times when these crimes peak, **Central** experiences the most crime around **12 PM**, then again between **6 PM and 8 PM**. In **77th Street**, the highest crime rate is at **12 PM**, but there's also a noticeable spread of criminal activity from **1 PM through 10 PM**, indicating a long window of risk. For **Hollywood**, **12 PM** stands out as the most common hour for crimes. This timing information helps to understand not just where crime happens, but when people need to be most cautious.

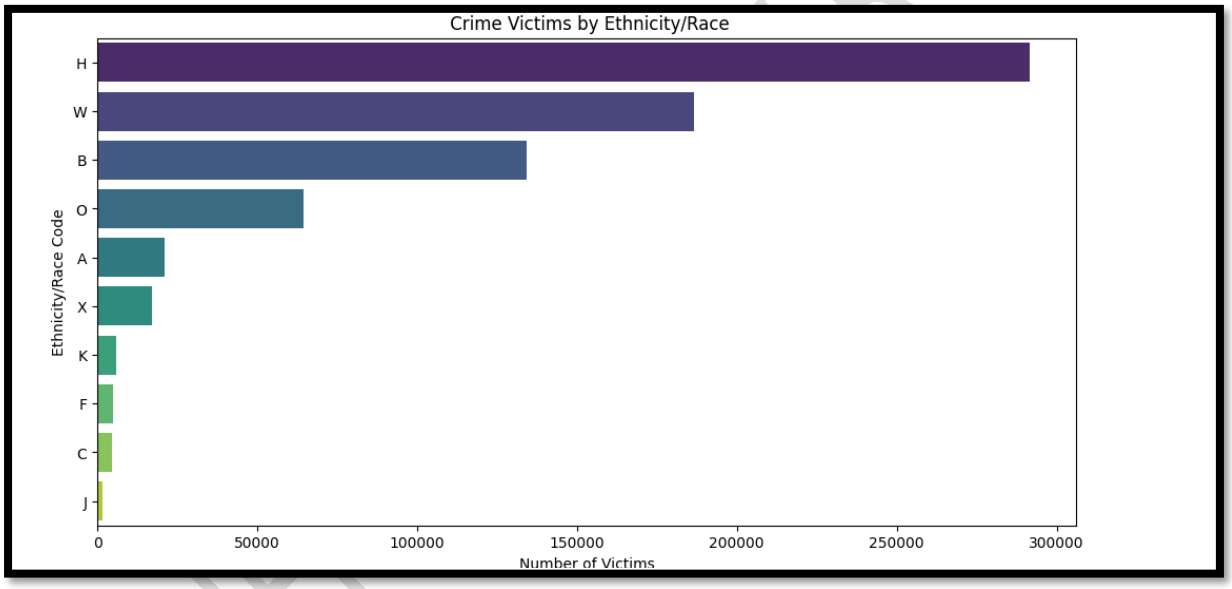
Victim Analysis:

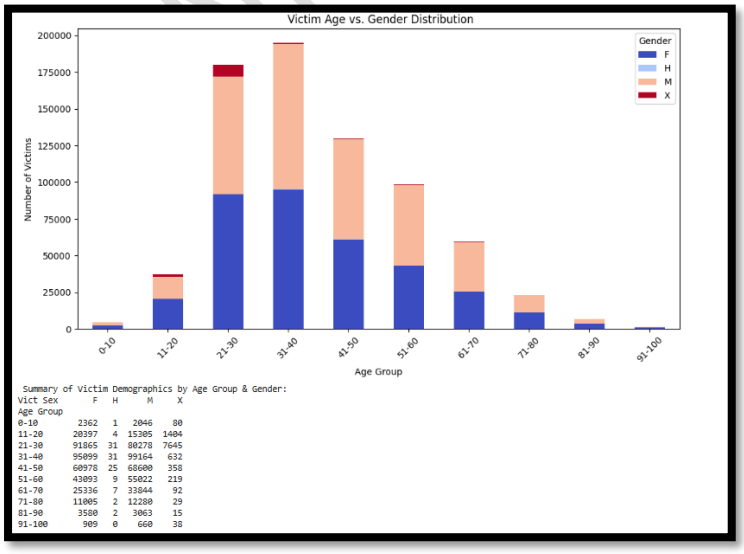
7. **What is the demographic of victims?**











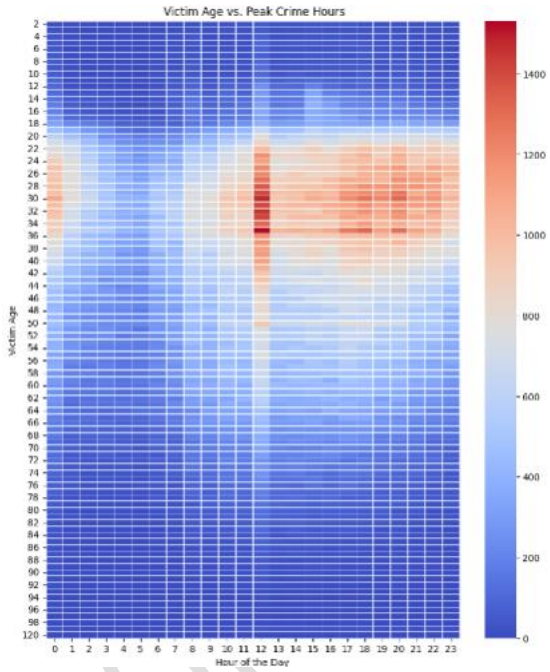
Summary of Victim Demographics by Age Group & Gender:

Vict Sex	F	H	M	X
Age Group				
0-10	2362	1	2046	80
11-20	20397	4	15305	1404
21-30	91865	31	80278	7645
31-40	95099	31	99164	632
41-50	60978	25	68600	358
51-60	43093	9	55022	219
61-70	25336	7	33844	92
71-80	11005	2	12280	29
81-90	3580	2	3063	15
91-100	909	0	660	38

Visual 14: Visualisation on Age Distribution, Gender and ethnicity of victims.

From Visual 14, we analysed the victim's age, gender, and ethnicity. The data shows that victim's range in age from 0 to 100, with the highest number of victims in the 21-30 and 31-40 age groups, which have 91,865 and 95,099 victims, respectively. In terms of gender, the distribution is relatively balanced, with 350,000 male victims and 340,000 female victims, although males have slightly more victims in most age groups. Regarding ethnicity, the most vulnerable groups are H, W, and B, with victims totalling approximately 290,000, 190,000, and 130,000, respectively. Looking at the breakdown by age group, we observe that the number of victims starts at 2,362 in the 0-10 range and peaks at 95,099 in the 31-40 range, before gradually declining in older age groups. Specifically, the 11-20 group has 20,397 victims, the 21-30 group has 91,865, and the 31-40 group has 95,099 victims. Older age groups show significantly lower numbers, with the 81-90 and 91-100 groups having 3,580 and 909 victims, respectively.

8. **What is the relationship between the victim's age and the time of the crime?**



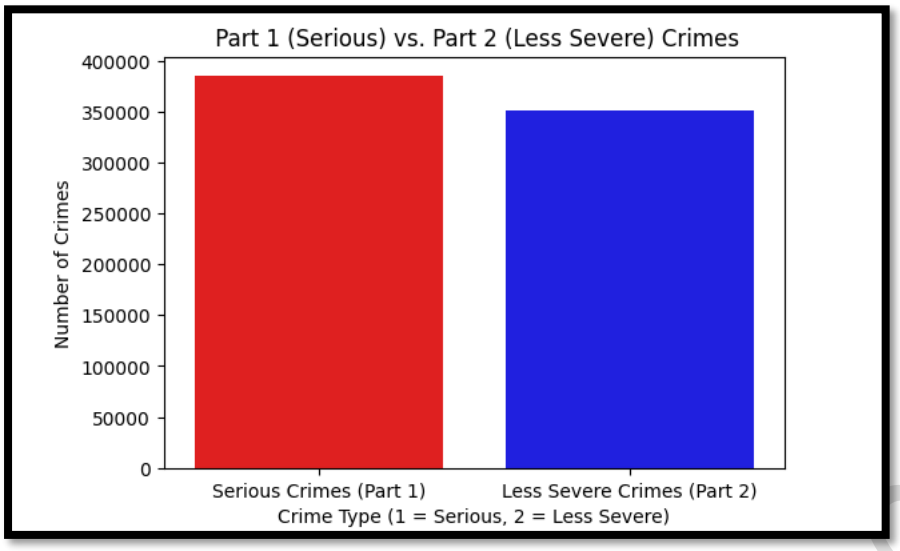
Visual 14: Heat map showing the relationship between victim age and peak crime hours.

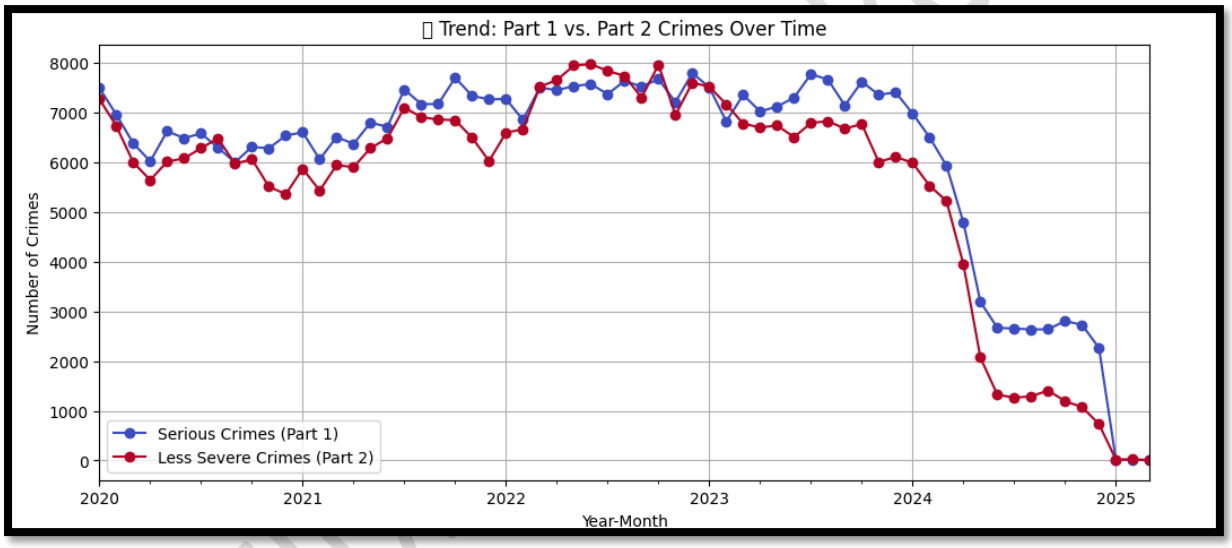
We observe that the highest number of crimes occur among individuals aged 24 to 36, primarily between 11 AM and 10 PM, which are the peak hours for criminal activity.

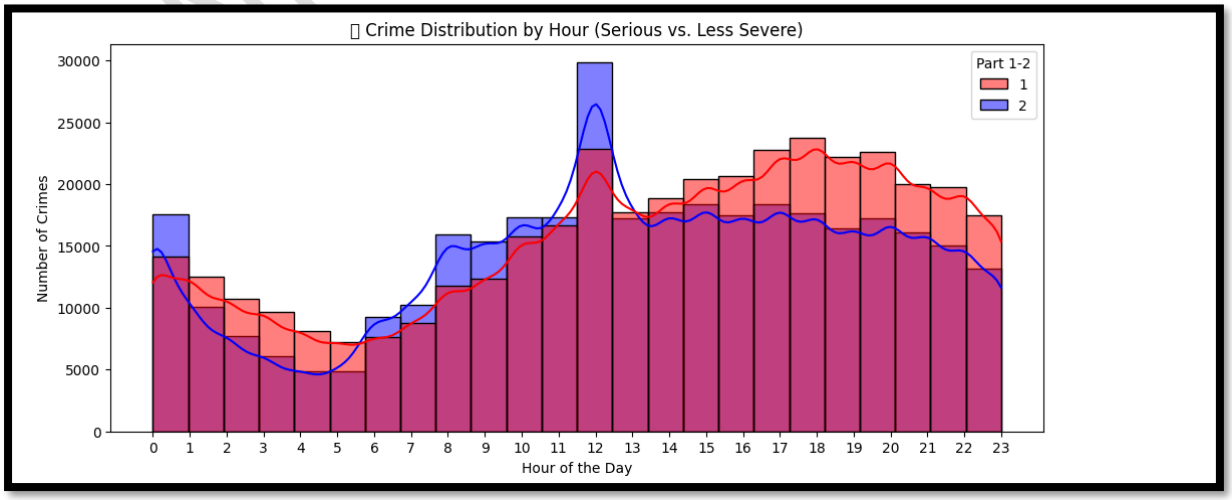
Crime Severity & Weapons Analysis

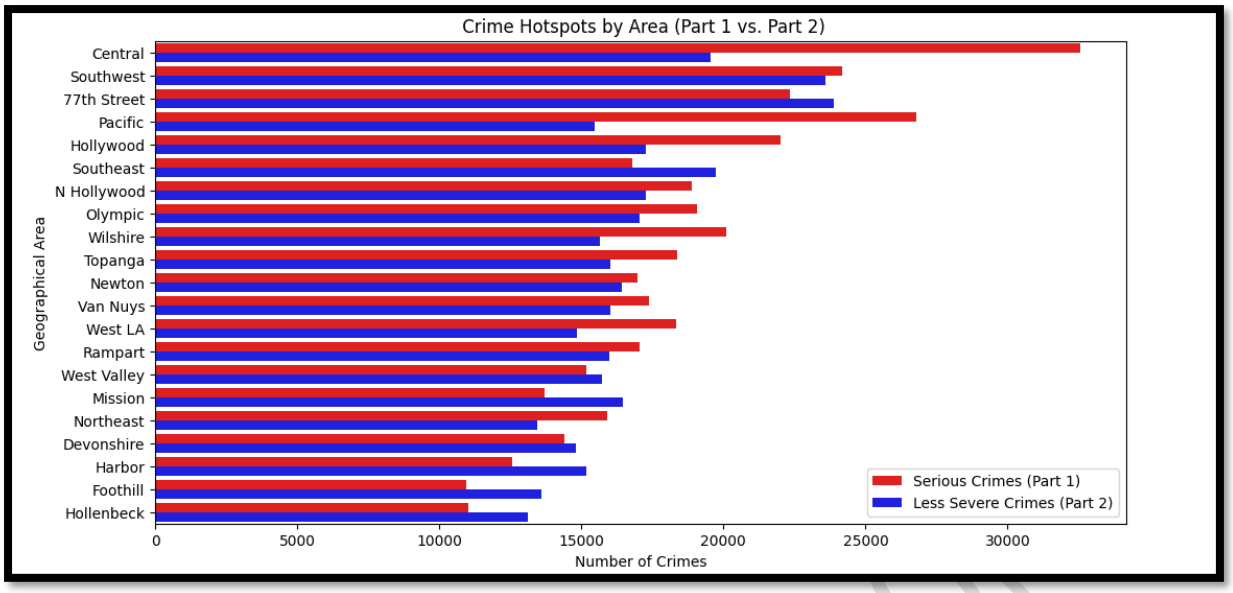
9.What is the frequency of serious versus less serious crimes in California? Provide the yearly trends for both categories, along with their hourly distribution and the areas where

these crimes most commonly occur.









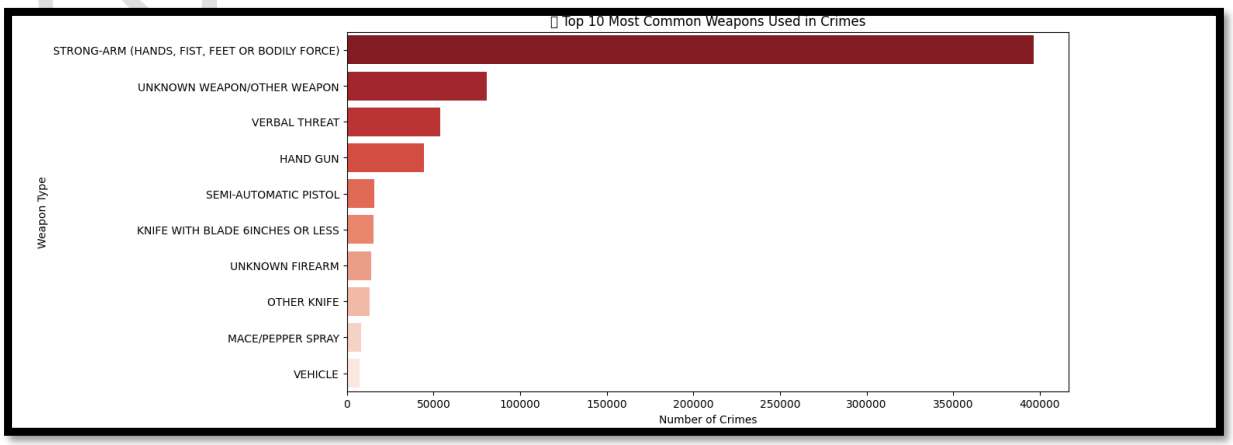
Visual 16: Crime Severity and it trends by years, day area and month.

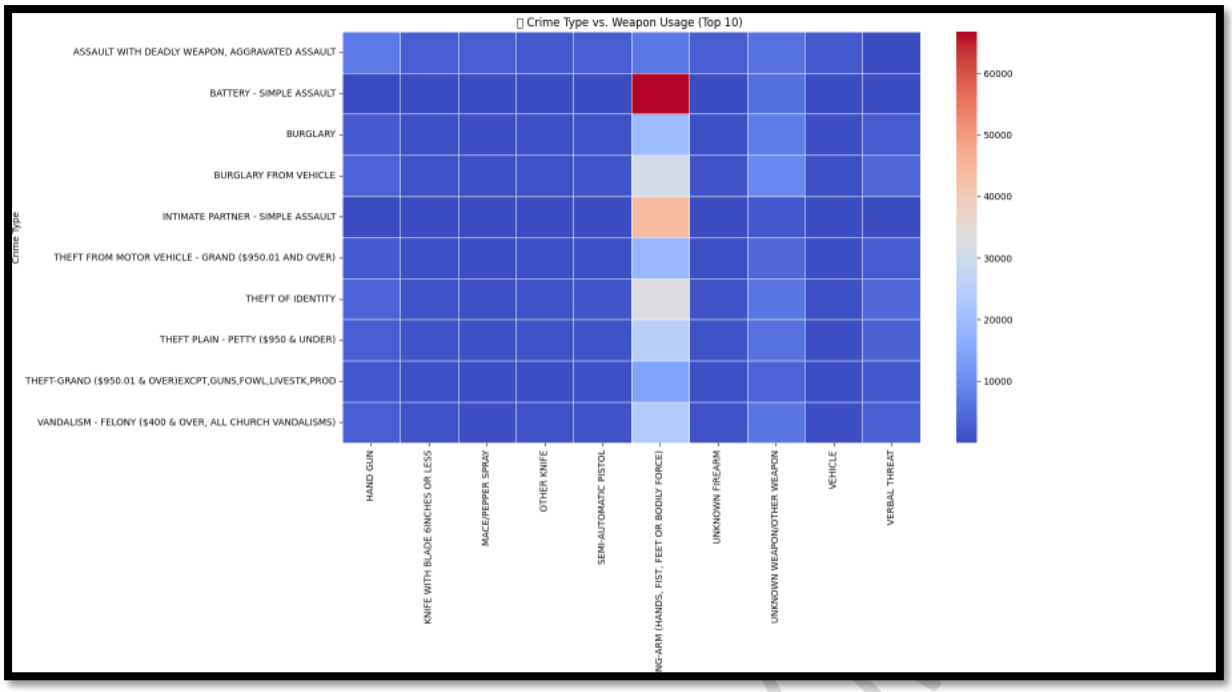
From Visual 16, crimes have been categorized based on their severity for analysis. The first bar plot indicates that serious crimes occur more frequently than less severe ones, with total counts of approximately 390,000 and 350,000, respectively. However, the yearly trend line graph shows that both severe and less severe crimes were consistently high between 2021 and 2024, peaking around 2023–2024, followed by a sharp decline starting in 2024.

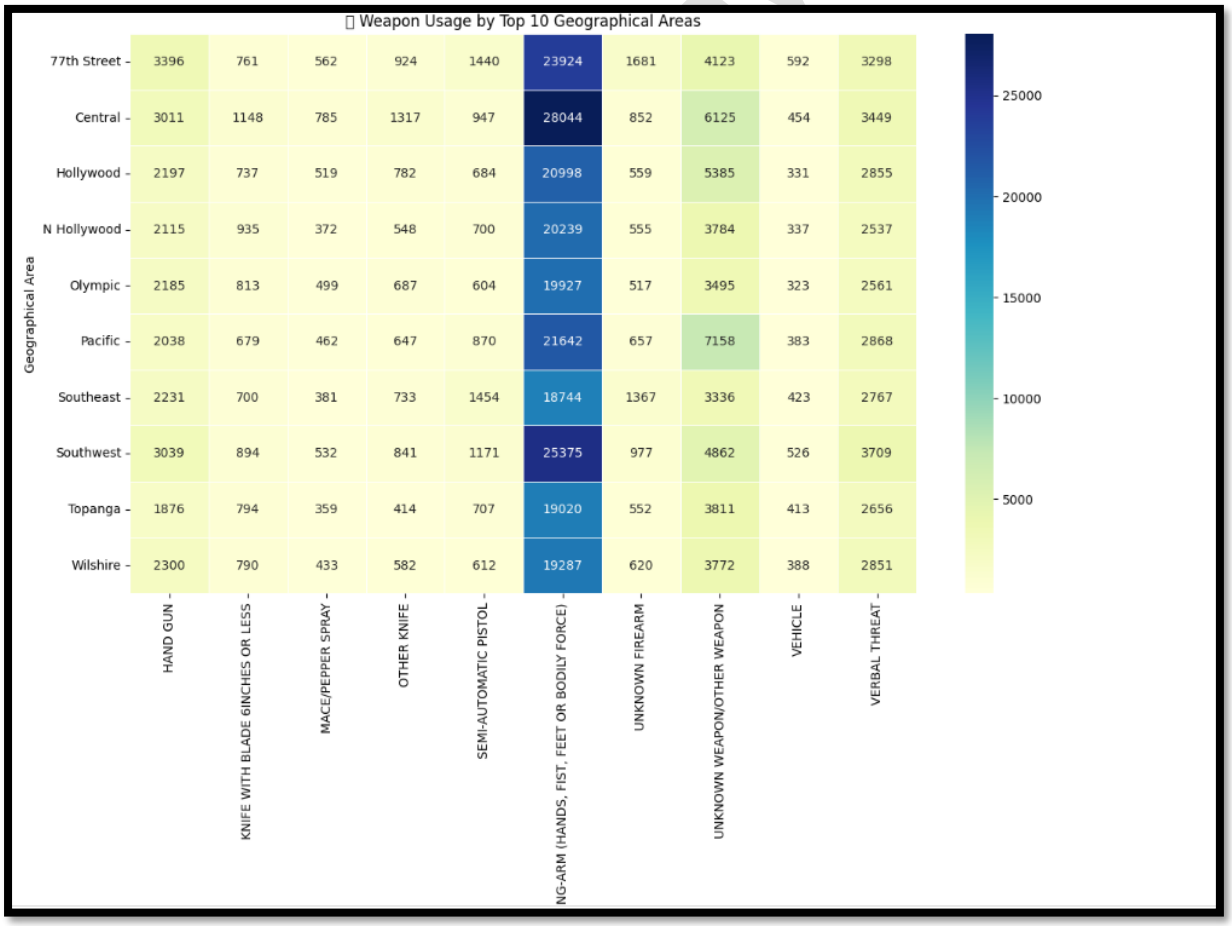
The hourly crime distribution reveals that 12 PM is the peak time for both crime types. Additionally, severe crimes tend to occur more frequently between 5 PM and 10 PM. Lastly, the grouped bar chart highlights key crime hotspot areas—Central, Southwest, 77th Street, Pacific, and Hollywood—as the top locations for both severe and less severe crimes.

Weapon Analysis:

10: What is the most common weapon used in the crime, what kind of weapon used in what type of crime, and weapon used in the Geographical area??







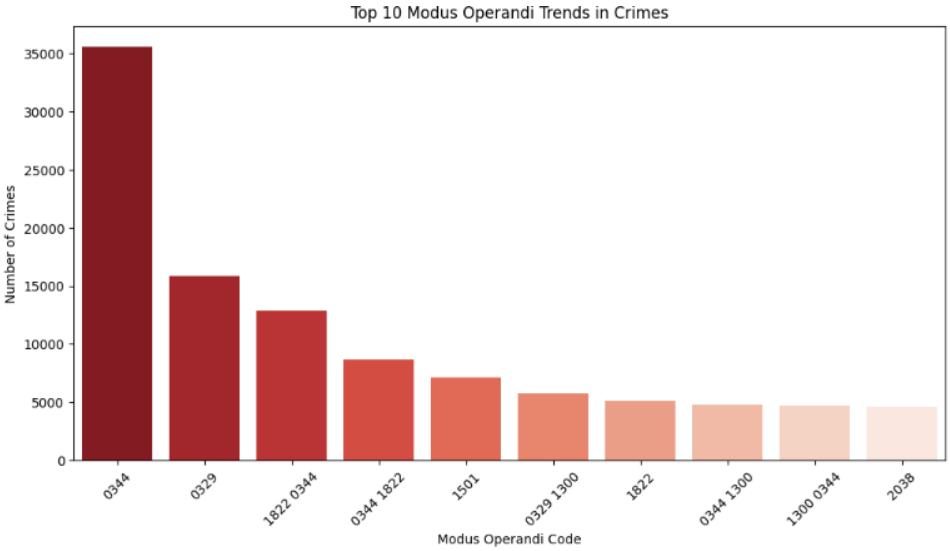
Visual 17: Weapon Analysis –

This visual 17 illustrates the types of weapons used across various crimes and highlights the specific areas where each weapon type is most commonly involved.

The bar chart reveals that the most commonly used weapon types are strong-arm force, unknown weapons, verbal threats, and handguns. The heatmap further highlights that in cases of simple assault, the crime is predominantly committed using strong-arm force alone. The final bar heatmap indicates that strong-arm force is predominantly used across all types of crimes, with the highest frequency observed in areas such as Central, Southwest, and 77th Street.

Criminal Modus Operandi Trends: Identify crime patterns and MO code.

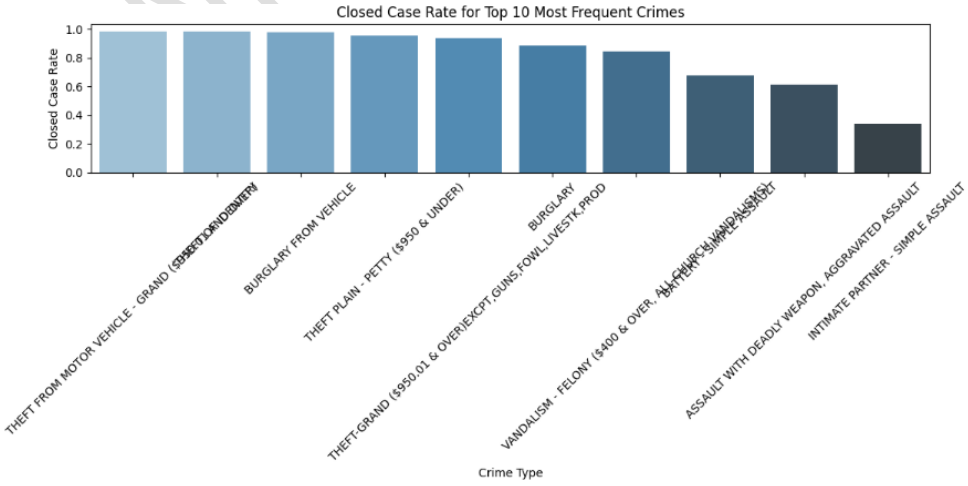
11. What is the modus operandi, and how many crimes have occurred under this code?



Visual 18: Bar graph to show criminal modus operandi trends.

Visual 18 shows that the most prevalent crime code is 0344, with a total of 35,000 reported incidents. This is followed by crime code 0329, and then 18220344, which has a total of 13,000 reported crimes.

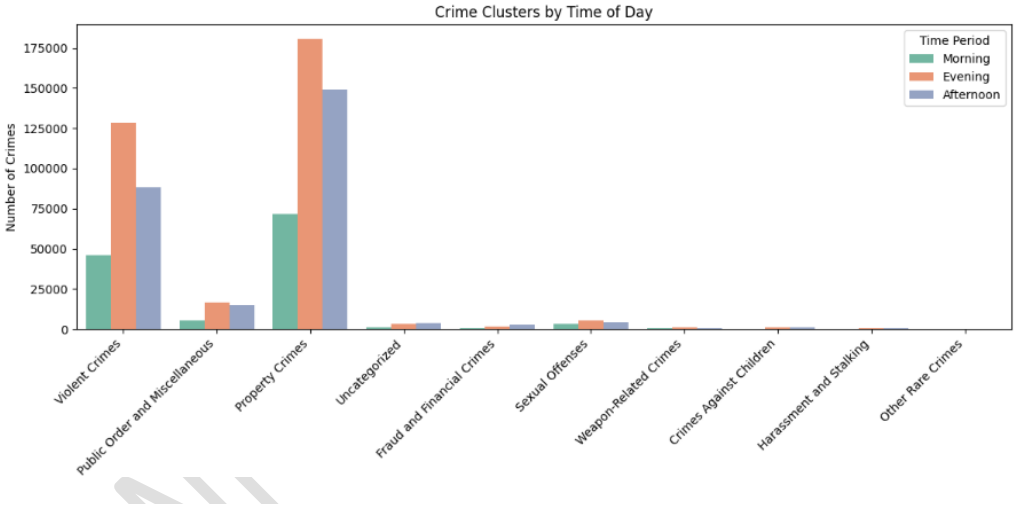
11. Analyse the rate at which crimes are being closed or resolved.

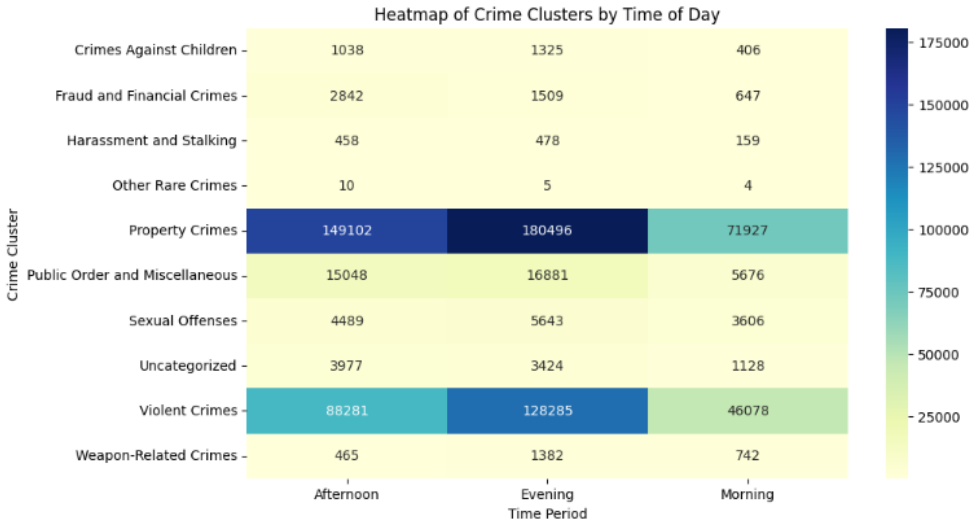


Visual 19: Crime closure rate.

From visual 19; to assess the resolution efficiency across different crime types, I first categorized raw status codes such as 'AA', 'IC', 'JA', and others into two groups: **Open** and **Closed**. After confirming all statuses were properly mapped, I identified the **top 10 most frequently reported crimes** and grouped them by both crime type and status. For each crime type, I calculated the **Closed Rate** as the proportion of closed cases to the total number of reported cases (Open + Closed). The results revealed that crimes such as **"Theft from Motor Vehicle - Grand ($950.01 and Over)"**, **"Theft of Identity"**, and **"Burglary from Vehicle"** had exceptionally high closure rates, all above **98%**. Similarly, **petty theft** and **grand theft** also showed strong closure performance, with rates above **93%**. In contrast, crimes involving **assault**, particularly **"Intimate Partner - Simple Assault"**, had significantly lower closed rates — as low as **34%** — suggesting potential complexities or investigative challenges associated with these cases. This analysis highlights both the strengths and gaps in case resolution across different crime categories.

12. What types of crimes are most prevalent in the city, and how are they distributed across different times of the day?

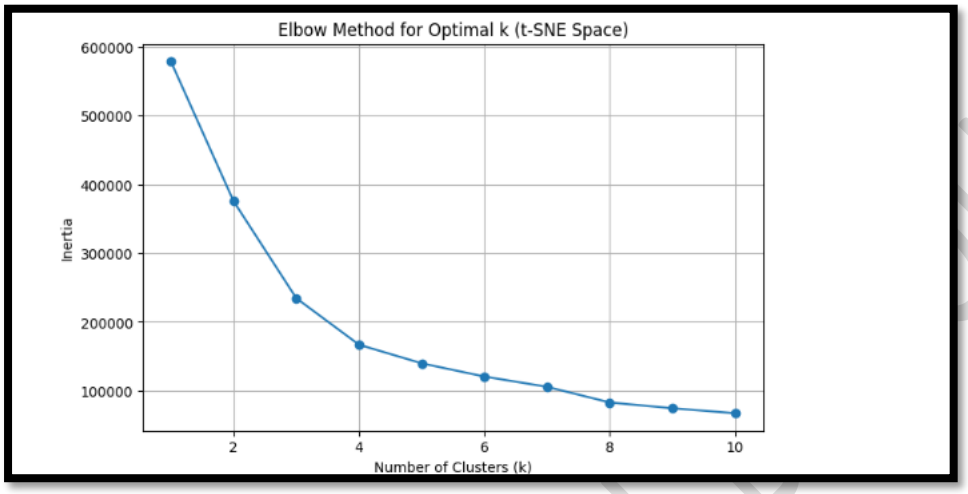


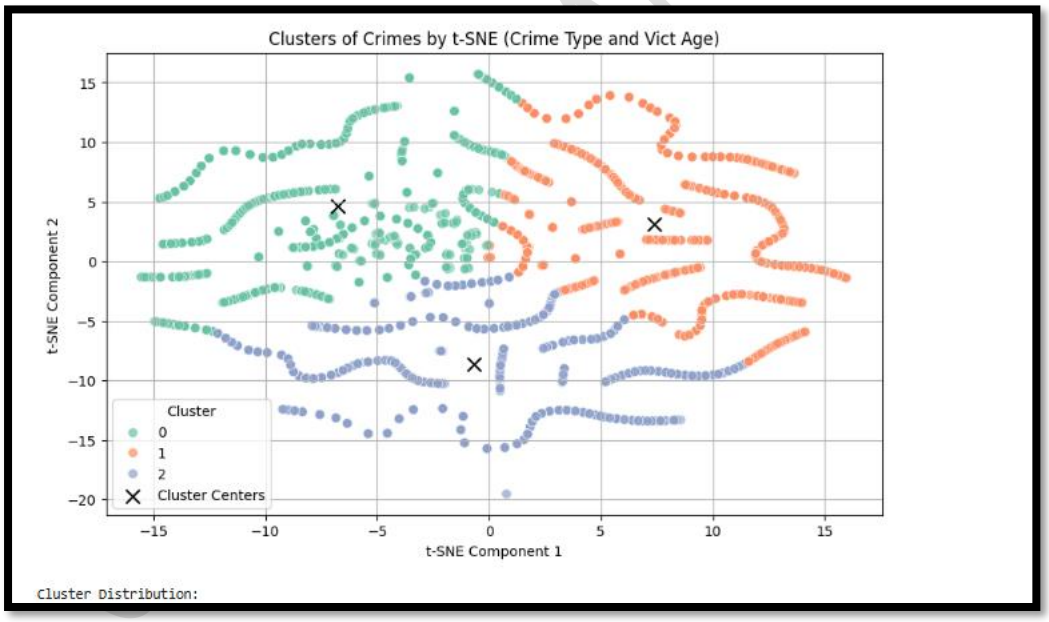


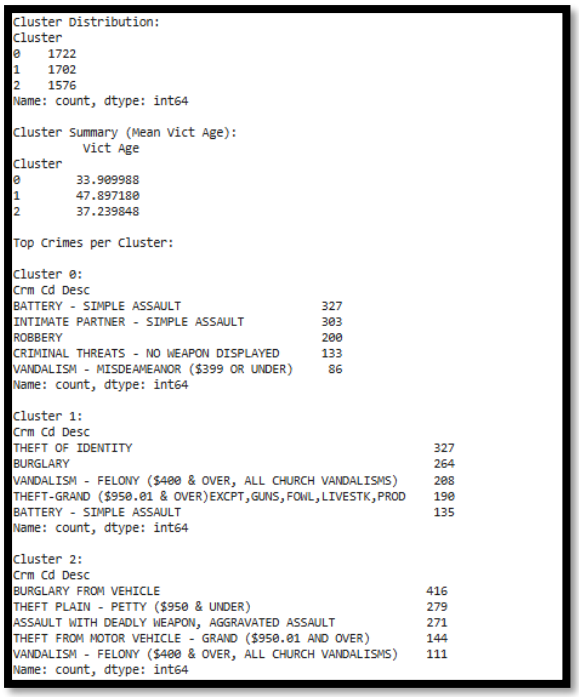
Visual 20: Crime Types Occurring at Different Times of Day

From Visual 20, violent crimes occur throughout the day, with the highest number of cases in the evening, followed by the afternoon. The same pattern is observed for property crimes, which peak in the evening and then in the afternoon.

13. Can we identify distinct patterns or groups of crimes based on victim age and crime type to support targeted crime prevention strategies?







```
Cluster Distribution:
Cluster
0    1722
1    1702
2    1576
Name: count, dtype: int64

Cluster Summary (Mean Vict Age):
          Vict Age
Cluster
0        33.909988
1        47.897180
2        37.239848

Top Crimes per Cluster:

Cluster 0:
Crm Cd Desc
BATTERY - SIMPLE ASSAULT                    327
INTIMATE PARTNER - SIMPLE ASSAULT           303
ROBBERY                                     200
CRIMINAL THREATS - NO WEAPON DISPLAYED      133
VANDALISM - MISDEAMEANOR ($399 OR UNDER)     86
Name: count, dtype: int64

Cluster 1:
Crm Cd Desc
THEFT OF IDENTITY                                       327
BURGLARY                                                264
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)     208
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD    190
BATTERY - SIMPLE ASSAULT                                135
Name: count, dtype: int64

Cluster 2:
Crm Cd Desc
BURGLARY FROM VEHICLE                                   416
THEFT PLAIN - PETTY ($950 & UNDER)                      279
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT          271
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)     144
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)    111
Name: count, dtype: int64
```

Visual 21: t-sne Cluster with it statistics.

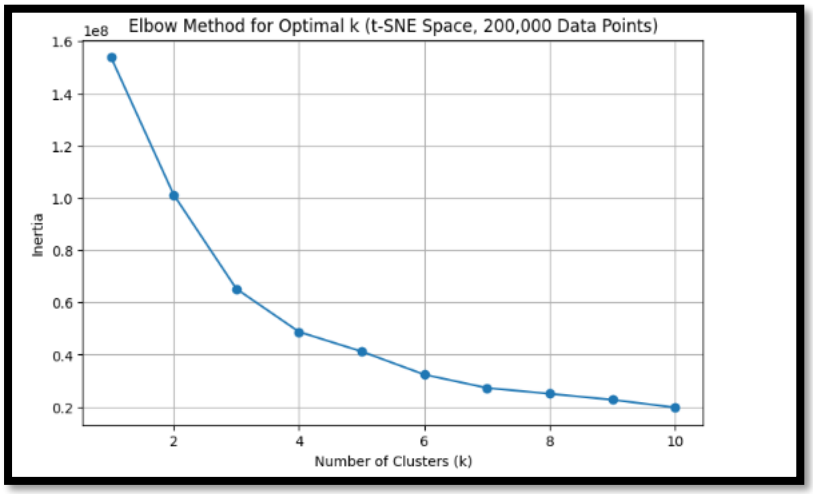
t-SNE (t-distributed Stochastic Neighbor Embedding) was used in this analysis as a dimensionality reduction technique to visualize high-dimensional data in a 2D space. The dataset contained multiple binary indicators for crime types (generated through one-hot encoding) along with victim age, resulting in a complex feature space. t-SNE effectively reduced this dimensionality while preserving local patterns and similarities, allowing for easier visualization and identification of meaningful clusters within the data.

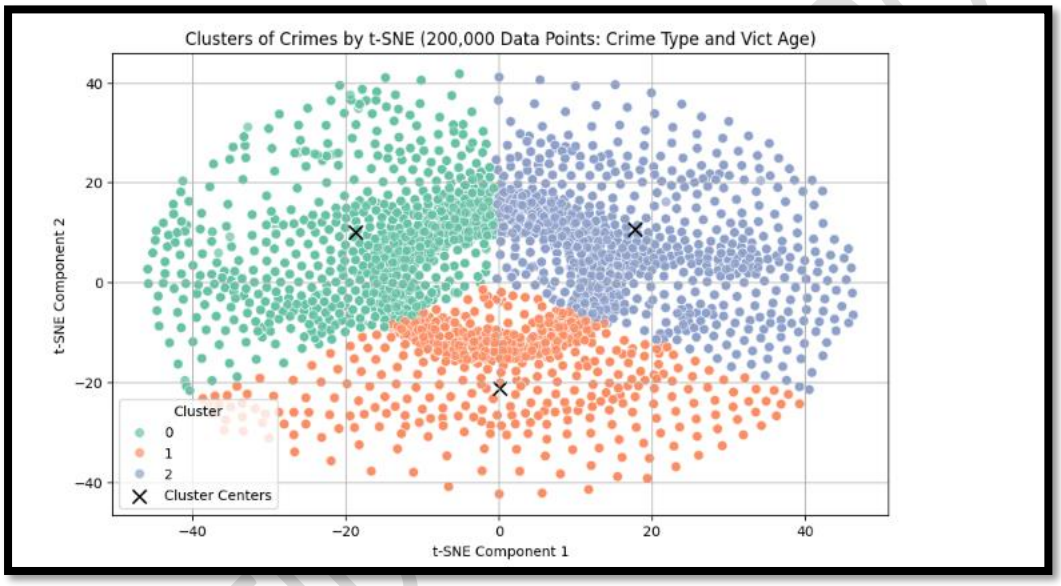
To manage computational efficiency and enhance the performance of t-SNE, the dataset was sampled down to 5,000 records. If the original dataset contained fewer than 5,000 entries, all data points were included. The features used for this analysis consisted of victim age and encoded crime type indicators.

Once dimensionality was reduced using t-SNE (with two components), KMeans clustering was applied to group similar observations in the transformed 2D space. The elbow method was employed to determine the optimal number of clusters, which was found to be three.

The resulting clusters were distributed as follows: Cluster 0 included 1,722 data points, Cluster 1 had 1,702, and Cluster 2 consisted of 1,576 data points. The average victim age within each cluster was approximately 33.9 years for Cluster 0, 47.9 years for Cluster 1, and 37.2 years for Cluster 2. Each cluster also exhibited distinct dominant crime types: Cluster 0 was primarily associated with assault and robbery-related crimes, Cluster 1 included cases of identity theft, burglary, felony vandalism, and grand theft, while Cluster 2 showed a concentration of burglary from vehicles, petty theft, and aggravated assault. This clustering helps uncover patterns in the data, offering insights that could support more targeted crime prevention strategies and effective resource allocation.

T-SNE Cluster on large Data points






```
Cluster Distribution:
Cluster
2    68148
1    67477
0    64375
Name: count, dtype: int64

Cluster Summary (Mean Vict Age):
           Vict Age
Cluster
0        29.838276
1        37.703158
2        50.437709

Top Crimes per Cluster:

Cluster 0:
Crm Cd Desc
INTIMATE PARTNER - SIMPLE ASSAULT                11952
ROBBERY                                           6715
BATTERY - SIMPLE ASSAULT                          5746
CRIMINAL THREATS - NO WEAPON DISPLAYED            4260
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT    4058
Name: count, dtype: int64

Cluster 1:
Crm Cd Desc
BURGLARY FROM VEHICLE                                        16706
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT               10041
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)       9484
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)           6972
BATTERY - SIMPLE ASSAULT                                      6345
Name: count, dtype: int64

Cluster 2:
Crm Cd Desc
THEFT OF IDENTITY                          12951
BURGLARY                                   10590
THEFT PLAIN - PETTY ($950 & UNDER)          8288
BATTERY - SIMPLE ASSAULT                    7968
VANDALISM - MISDEAMEANOR ($399 OR UNDER)    4293
Name: count, dtype: int64
```

Visual20: T-sne Clustering for big data points.

After performing initial clustering on a small sample for quick insights, we advanced to a more robust analysis by scaling up to **200,000 data points**. This allowed for **greater accuracy and reliability** in identifying crime patterns.

Using **t-SNE** for dimensionality reduction and **KMeans** for clustering, we identified **three distinct clusters** based on crime type and victim age.

- **Cluster 0**: Dominated by **younger victims (avg. age ~29.8)**, with high occurrences of **intimate partner assaults and robberies**.
- **Cluster 1**: Victims in their **mid-thirties (avg. age ~37.7)**, with a mix of **vehicle-related crimes** and **aggravated assaults**.
- **Cluster 2**: Characterized by **older victims (avg. age ~50.4)**, mostly affected by **identity theft, burglary, and petty theft**.

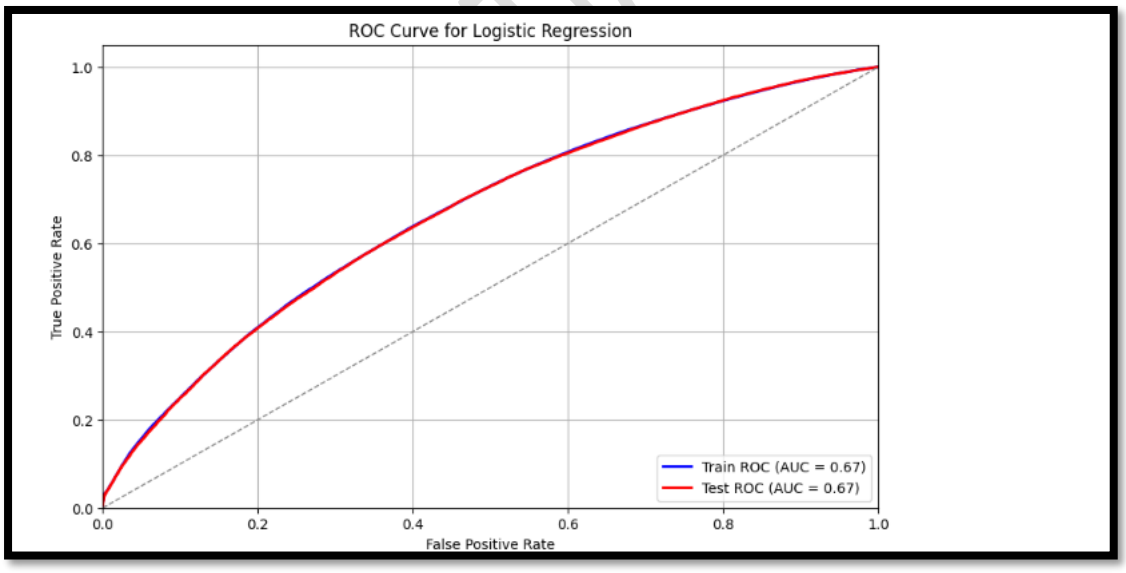
This expanded clustering reveals **clear, age-based crime patterns**, supporting **targeted prevention strategies** and smarter **resource allocation**. By working with a larger dataset, the insights are not only more nuanced but also more actionable

Logistic Regression Model:

```
Optimization terminated successfully.
         Current function value: 0.502427
         Iterations: 271
         Function evaluations: 272
         Gradient evaluations: 272

Statsmodels Logistic Regression Summary:
                           Logit Regression Results
==============================================================================
Dep. Variable:          Status_Binary   No. Observations:               588404
Model:                          Logit   Df Residuals:                   588351
Method:                           MLE   Df Model:                           52
Date:                Thu, 27 Mar 2025   Pseudo R-squ.:                0.05915
Time:                        06:48:30   Log-Likelihood:           -2.9563e+05
converged:                       True   LL-Null:                  -3.1422e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                  3.7218      2.917      1.276      0.202      -1.995       9.438
Hour                  -0.0332      0.003    -10.270      0.000      -0.040      -0.027
Vict Age               0.1090      0.003     32.432      0.000       0.102       0.116
Part 1-2              -0.7050      0.007   -106.143      0.000      -0.718      -0.692
Month                  0.0293      0.003      9.090      0.000       0.023       0.036
AREA_2                -0.5183      0.020    -25.912      0.000      -0.557      -0.479
AREA_3                -0.1388      0.019     -7.186      0.000      -0.177      -0.101
AREA_4                -0.3907      0.022    -17.707      0.000      -0.434      -0.347
AREA_5                -0.7142      0.020    -34.901      0.000      -0.754      -0.674
AREA_6                -0.0821      0.021     -3.941      0.000      -0.123      -0.041
AREA_7                -0.1365      0.021     -6.418      0.000      -0.178      -0.095
AREA_8                -0.1902      0.022     -8.548      0.000      -0.234      -0.147
AREA_9                -0.6490      0.020    -32.455      0.000      -0.688      -0.610
AREA_10               -0.8181      0.020    -40.700      0.000      -0.858      -0.779
AREA_11               -0.4932      0.021    -23.211      0.000      -0.535      -0.452
AREA_12               -0.4386      0.018    -23.874      0.000      -0.475      -0.403
AREA_13                0.0291      0.021      1.373      0.170      -0.012       0.071
AREA_14                0.0574      0.021      2.673      0.008       0.015       0.099
AREA_15               -0.7028      0.020    -36.009      0.000      -0.741      -0.665
AREA_16               -0.6829      0.021    -32.032      0.000      -0.725      -0.641
AREA_17               -0.5146      0.021    -24.353      0.000      -0.556      -0.473
AREA_18                0.1230      0.021      5.933      0.000       0.082       0.164
AREA_19               -0.6903      0.020    -34.327      0.000      -0.730      -0.651
AREA_20               -0.5133      0.020    -25.749      0.000      -0.552      -0.474
AREA_21               -0.7642      0.020    -38.098      0.000      -0.803      -0.725
Vict Sex_H             0.1781      0.357      0.499      0.618      -0.521       0.877
Vict Sex_M             0.2405      0.007     36.266      0.000       0.228       0.254
Vict Sex_X            -0.0846      0.043     -1.985      0.047      -0.168      -0.001
Vict Descent_A        -0.8732      2.917     -0.299      0.765      -6.590       4.843
Vict Descent_B        -1.2844      2.917     -0.440      0.660      -7.001       4.432
Vict Descent_C         1.6776      2.920      0.575      0.566      -4.045       7.400
Vict Descent_D         0.3715      2.960      0.126      0.900      -5.430       6.173
Vict Descent_F         1.0316      2.918      0.354      0.724      -4.687       6.751
Vict Descent_G        -0.3859      2.945     -0.131      0.896      -6.159       5.387
```

AREA_10	-0.8181	0.020	-40.700	0.000	-0.858	-0.779
AREA_11	-0.4932	0.021	-23.211	0.000	-0.535	-0.452
AREA_12	-0.4386	0.018	-23.874	0.000	-0.475	-0.403
AREA_13	0.0291	0.021	1.373	0.170	-0.012	0.071
AREA_14	0.0574	0.021	2.673	0.008	0.015	0.099
AREA_15	-0.7028	0.020	-36.009	0.000	-0.741	-0.665
AREA_16	-0.6829	0.021	-32.032	0.000	-0.725	-0.641
AREA_17	-0.5146	0.021	-24.353	0.000	-0.556	-0.473
AREA_18	0.1230	0.021	5.933	0.000	0.082	0.164
AREA_19	-0.6903	0.020	-34.327	0.000	-0.730	-0.651
AREA_20	-0.5133	0.020	-25.749	0.000	-0.552	-0.474
AREA_21	-0.7642	0.020	-38.098	0.000	-0.803	-0.725
Vict Sex_H	0.1781	0.357	0.499	0.618	-0.521	0.877
Vict Sex_M	0.2405	0.007	36.266	0.000	0.228	0.254
Vict Sex_X	-0.0846	0.043	-1.985	0.047	-0.168	-0.001
Vict Descent_A	-0.8732	2.917	-0.299	0.765	-6.590	4.843
Vict Descent_B	-1.2844	2.917	-0.440	0.660	-7.001	4.432
Vict Descent_C	1.6776	2.920	0.575	0.566	-4.045	7.400
Vict Descent_D	0.3715	2.960	0.126	0.900	-5.430	6.173
Vict Descent_F	1.0316	2.918	0.354	0.724	-4.687	6.751
Vict Descent_G	-0.3859	2.945	-0.131	0.896	-6.159	5.387
Vict Descent_H	-1.2579	2.917	-0.431	0.666	-6.974	4.458
Vict Descent_I	0.4109	2.921	0.141	0.888	-5.314	6.136
Vict Descent_J	1.4065	2.924	0.481	0.630	-4.324	7.137
Vict Descent_K	0.4078	2.917	0.140	0.889	-5.310	6.126
Vict Descent_L	0.3964	2.966	0.134	0.894	-5.417	6.210
Vict Descent_O	-0.8698	2.917	-0.298	0.766	-6.586	4.847
Vict Descent_P	0.4769	2.931	0.163	0.871	-5.268	6.222
Vict Descent_S	0.1782	2.977	0.060	0.952	-5.657	6.013
Vict Descent_U	0.4965	2.941	0.169	0.866	-5.267	6.260
Vict Descent_V	1.7372	2.928	0.593	0.553	-4.002	7.477
Vict Descent_W	-0.7788	2.917	-0.267	0.789	-6.495	4.938
Vict Descent_X	-0.7330	2.917	-0.251	0.802	-6.450	4.984
Vict Descent_Z	1.3000	2.934	0.443	0.658	-4.450	7.050
DayOfWeek_Monday	-0.0606	0.012	-5.036	0.000	-0.084	-0.037
DayOfWeek_Saturday	-0.1059	0.012	-8.928	0.000	-0.129	-0.083
DayOfWeek_Sunday	-0.2138	0.012	-18.081	0.000	-0.237	-0.191
DayOfWeek_Thursday	-0.0220	0.012	-1.813	0.070	-0.046	0.002
DayOfWeek_Tuesday	-0.0375	0.012	-3.085	0.002	-0.061	-0.014
DayOfWeek_Wednesday	-0.0367	0.012	-3.043	0.002	-0.060	-0.013

```
==============================================================================

Training Accuracy: 0.6050026852298761
Training Confusion Matrix:
 [[ 85456  47341]
 [185077 270530]]

Testing Accuracy: 0.604662071215891
Testing Confusion Matrix:
 [[21386 11818]
 [46337 67561]]
```



Visual 21: Logistic regression model with statistics.

A logistic regression model trained on a dataset of 588,404 samples with 52 features, including continuous variables like Hour, Vict Age, and Month, and binary-encoded categorical features such as AREA_2 to AREA_21, Vict Sex, Vict Descent, and DayOfWeek, all

free of missing values. It predicts a binary target (Status_Binary) and achieved convergence after 271 iterations using Maximum Likelihood Estimation, with a Pseudo R-squared of 0.05915 and a Log-Likelihood of -2.9563e+05, indicating modest explanatory power. The model demonstrates balanced performance, with a training accuracy of 60.50% and a testing accuracy of 60.47%, showing no signs of overfitting. Its training confusion matrix ([85,456 TN, 47,341 FP], [185,077 FN, 270,530 TP]) yields a precision of 0.851, recall of 0.594, and F1 score of 0.699, while the testing confusion matrix ([21,386 TN, 11,818 FP], [46,337 FN, 67,561 TP]) maintains similar metrics (precision: 0.851, recall: 0.593, F1: 0.698). Significant predictors include Part 1-2 (coef: -0.7050), Vict Age (coef: 0.1090), and AREA_10 (coef: -0.8181), though some Vict Descent features are less impactful (p>0.05).This model stands out for its robust generalization and balanced handling of both classes, making it a reliable choice for applications where minority class performance matters.

Explanation of Regression model:

Influential Variables and Justification

The following variables significantly influence the response variable Status_Binary (0 = Open, 1 = Closed) in Model 2's logistic regression, based on $p < 0.05$, indicating statistical significance. Their impact is driven by coefficient magnitude, z-scores, and crime-related context:

- **Part 1-2 (coef: -0.7050, z: -106.143, p < 0.001)**: Part 2 crimes (less serious) are less likely to be closed than Part 1 crimes (serious, e.g., murder), as serious crimes receive more investigative resources, boosting closure rates.
- **Vict Age (coef: 0.1090, z: 32.432, p < 0.001)**: Older victims increase closure likelihood, possibly due to involvement in specific crime types or clearer testimony aiding investigations.
- **Vict Sex_M (coef: 0.2405, z: 36.266, p < 0.001)**: Male victims are more likely to have closed cases than the reference (likely female), reflecting differences in crime types or reporting patterns.
- **Vict Sex_X (coef: -0.0846, z: -1.985, p = 0.047)**: The "X" sex category (unknown/non-binary) slightly reduces closure likelihood, possibly due to incomplete data complicating investigations.
- **AREA_2, AREA_3, AREA_4, AREA_5, AREA_6, AREA_7, AREA_8, AREA_9, AREA_10, AREA_11, AREA_12, AREA_15, AREA_16, AREA_19, AREA_20, AREA_21 (coefs: -0.8181 to -0.0821, z: -40.700 to -3.941, p < 0.001)**: These areas decrease closure likelihood compared to the reference (AREA_1), with AREA_10 (-0.8181) and AREA_21 (-0.7642) showing the strongest effects. Geographic variations likely stem from differences in crime density, resources, or socioeconomic factors.

- **AREA_14 (coef: 0.0574, z: 2.673, p = 0.008), AREA_18 (coef: 0.1230, z: 5.933, p < 0.001)**: These areas increase closure likelihood, suggesting better policing or simpler cases in these locations.
- **Hour (coef: -0.0332, z: -10.270, p < 0.001)**: Later hours reduce closure likelihood, likely due to fewer witnesses or investigative challenges at night.
- **Month (coef: 0.0293, z: 9.090, p < 0.001)**: Later months slightly increase closure likelihood, possibly reflecting seasonal investigative patterns.
- **DayOfWeek_Monday, DayOfWeek_Saturday, DayOfWeek_Sunday, DayOfWeek_Tuesday, DayOfWeek_Wednesday (coefs: -0.2138 to -0.0367, z: -18.081 to -3.043, p ≤ 0.002)**: Cases on these days are less likely to close than on Friday (reference), with Sunday showing the strongest effect, likely due to limited resources on weekends or early weekdays.

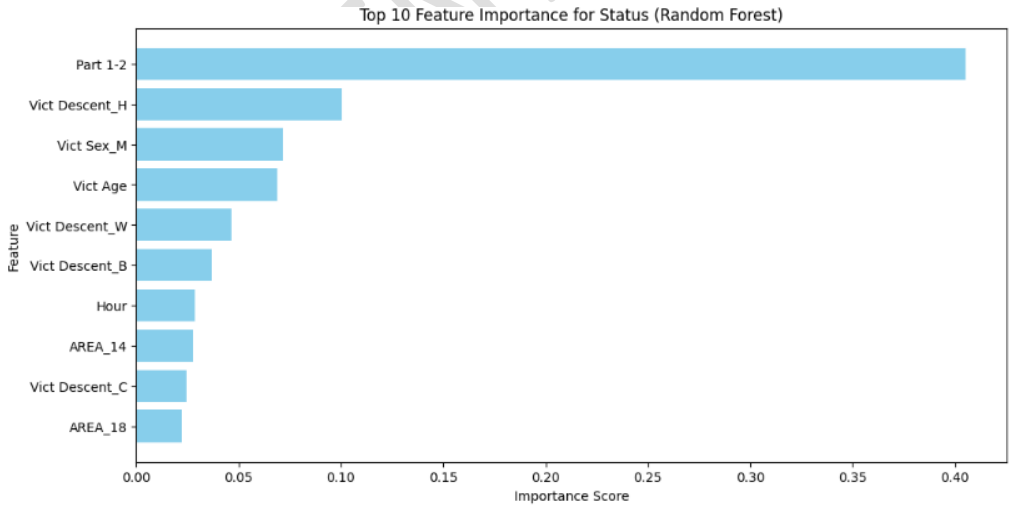
Summary of Regression model:

The response variable, Status_Binary, indicates whether a case is **"Open" (0) or "Closed" (1)**. The explanatory variables include Hour, Vict Age, Month, Part 1-2, and one-hot encoded AREA, Vict Sex, Vict Descent, and DayOfWeek. **Significant influencers (p < 0.05) include Part 1-2, Vict Age, Vict Sex_M, Vict Sex_X, most AREA variables (especially AREA_10, AREA_5, AREA_15, AREA_19, AREA_21, AREA_18), Hour, Month, and most DayOfWeek variables (especially Sunday, Saturday, Monday**). These variables drive case closure predictions due to their **statistical significance, large coefficients**, and relevance to crime investigation dynamics. Non-significant variables (Vict Descent_A to Z, Vict Sex_H, AREA_13, DayOfWeek_Thursday) have minimal impact due to weak associations or sparse data. The model's balanced performance (training accuracy: 60.50%, testing accuracy: 60.47%, F1 ≈ 0.698) reflects the effectiveness of these influential variables in capturing patterns in case resolution.

Random Forest:

14. "What factors influence whether a crime case is closed or remains open?"

```
Training Accuracy: 0.6176895466380242
Training Confusion Matrix:
 [[ 85103  47694]
 [177259 278348]]

Testing Accuracy: 0.6147027232804448
Testing Confusion Matrix:
 [[21147 12057]
 [44621 69277]]

Feature Importance (All Variables):
                  Feature  Importance
2              Part 1-2    0.405195
33       Vict Descent_H    0.100377
25           Vict Sex_M    0.071542
1              Vict Age    0.069110
43       Vict Descent_W    0.046590
28       Vict Descent_B    0.036983
0                  Hour    0.028820
16              AREA_14    0.027529
29       Vict Descent_C    0.024521
20              AREA_18    0.022085
31       Vict Descent_F    0.017240
36       Vict Descent_K    0.014934
3                 Month    0.013366
14              AREA_12    0.012516
15              AREA_13    0.012343
12              AREA_10    0.008356
8                AREA_6    0.007699
21              AREA_19    0.007678
7                AREA_5    0.007173
5                AREA_3    0.006987
10               AREA_8    0.006508
23              AREA_21    0.005627
38       Vict Descent_O    0.005511
9                AREA_7    0.004981
17              AREA_15    0.004207
48     DayOfWeek_Sunday    0.004032
18              AREA_16    0.003544
35       Vict Descent_J    0.003153
44       Vict Descent_X    0.002315
27       Vict Descent_A    0.001630
42       Vict Descent_V    0.001593
49   DayOfWeek_Thursday    0.001519
4                AREA_2    0.001510
11               AREA_9    0.001479
47   DayOfWeek_Saturday    0.001442
13              AREA_11    0.001438
46     DayOfWeek_Monday    0.001394
22              AREA_20    0.001365
50    DayOfWeek_Tuesday    0.001319
51  DayOfWeek_Wednesday    0.001294
26           Vict Sex_X    0.001191
19              AREA_17    0.000722
34       Vict Descent_I    0.000468
6                AREA_4    0.000258
```



Visual22: Random forest model.

As part of my machine learning project, I developed a Random Forest classification model to predict whether a reported crime case in Los Angeles would be closed or remain open. I started by cleaning the dataset and transforming the target variable Status into a binary

format: "Closed" as 1 and "Open" as 0. My predictor variables included factors such as the hour of occurrence, victim age, sex, descent, crime seriousness (Part 1-2), and location-based attributes like police area and day of the week. After handling missing values and encoding categorical variables using one-hot encoding, I split the data into training and test sets. I then applied feature scaling to the numeric columns and trained a RandomForestClassifier using 100 estimators with a maximum depth of 10 and class balancing enabled. The model achieved a training accuracy of **61.77%** and a testing accuracy of **61.47%**, which indicates the model was stable but had moderate predictive power. The testing confusion matrix showed **44,621 false negatives**, meaning the model often failed to correctly identify cases that were actually closed. Feature importance analysis revealed that Part 1-2 was the most dominant variable, contributing **40.5%** to the model's decisions, followed by victim descent (H, W, B, C) and victim sex, which together added over **25%** more. Surprisingly, temporal features like month and hour had relatively low influence, and many categorical features such as Vict Descent_L, Vict Descent_Z, and DayOfWeek_Tuesday contributed almost nothing. This project provided valuable insight into class imbalance, feature engineering, and model interpretability.

Summary:

Crime Analysis Summary for California

The California crime dataset, analyzed through logistic regression and random forest models, provides insights into crime patterns and case resolution dynamics. The dataset includes 25 columns with varied data types: DR_NO, TIME OCC, AREA, Rpt Dist No, Part 1-2, Crm Cd, Premis Cd, Weapon Used Cd, and Crm Cd 1-4 are integers (int64); LAT and LON are floats (float64); and Date Rptd, DATE OCC, AREA NAME, Crm Cd Desc, Mocodes, Vict Sex, Vict Descent, Premis Desc, Weapon Desc, Status, Status Desc, LOCATION, and Cross Street are strings (object). Key findings reveal vehicle theft (115,000 cases), simple assault (74,800), and burglary from vehicles (63,500) as the most prevalent crimes, with high closure rates for theft-related crimes (>93%) but lower for assaults (e.g., 34% for intimate partner assault). Central, 77th Street, and Pacific are crime hotspots, with peak crime hours between 10 AM and 3 PM, and Thursday to Saturday showing the highest frequencies. Winter sees the most crimes (257,500 cases). Victim demographics show a balanced gender distribution (350,000 males, 340,000 females), with ages 21–40 most affected, and ethnic groups H, W, and B being the most victimized.

The logistic regression model (Model 2) predicts case closure (Status_Binary: 0 = Open, 1 = Closed) with a training accuracy of 60.50% and testing accuracy of 60.47% (F1 ≈ 0.698), indicating balanced performance. Influential variables include Part 1-2 (serious crimes more likely to close), Vict Age (older victims increase closure likelihood), Vict Sex_M (male victims linked to higher closure), and most AREA variables (e.g., AREA_10, AREA_21 decrease closure; AREA_18 increases it). Temporal features like Hour (later hours reduce closure) and

DayOfWeek_Sunday (lower closure rates) also matter. The random forest model, with 61.77% training and 61.47% testing accuracy, confirms Part 1-2 as the dominant predictor (40.5% importance), followed by victim descent and sex, while temporal features have less impact. Missing values, particularly in Vict Sex, Vict Descent, and Weapon Used Cd, were handled with forward fill and dropping incomplete rows, ensuring data integrity. Clustering via t-SNE revealed age-based crime patterns, with younger victims linked to assaults and older ones to theft, supporting targeted prevention strategies.