Analyzing Crime Trends and Predicting Hot Spots in Los Angeles

Dominique Akinyemi

2024-11-03

Project Overview

This project analyzes recent crime trends in Los Angeles to uncover meaningful patterns and insights that inform the general public, local communities, and government stakeholders.

By exploring the types of crimes reported, victim demographics, and geographic distributions, this analysis seeks to empower individuals and organizations to make data-driven decisions that enhance safety and awareness. Visualizations and trend analyses will provide accessible, actionable insights to help citizens understand their surroundings and guide policy and resource planning at a local level.

Data Set Access Link: LA Crime Data from 2020 to Present

Tools

- **MySQL Workbench and Command Line Interface**: Used for writing and executing SQL queries to manage and analyze data.
- **R Markdown**: Used for creating project documentation.
- **Jupyter Notebook (Python)**: Used for data extraction, including `pdfplumber` to process pdf.
- **Tableau**: Used to create visualizations and dashboards for exploring and presenting insights.

Project Workflow

Throughout this project, SQL was used extensively to clean, transform, and analyze the crime data. Key techniques included filtering missing or irrelevant values (e.g., 'Unknown'), manipulating and pivoting tables to create tidy data sets, and performing aggregation (`GROUP BY`) to identify trends across demographics, geography, and time. Advanced queries integrated multiple fields to uncover actionable insights. These results were exported to Tableau for visualization and further analysis. Some, more in-depth queries are stored in this documentation, while other queries are summarized to streamline documentation and increase readability.

Table of Contents

I. Introduction to Data Source

'Crime Data from 2020 to Present' owned by LAPD Open Data

- LA's Crime data set contains the records for reported crime incidents in the city of Los Angeles from 2020 to the present (updated bimonthly). At the time of this project, it was most recently updated on October 30, 2024.
  - Note: On March 7, 2024, LAPD adopted a new records management system for crime reporting which has temporarily resulted in incomplete records, as only incidents reported on the old system are included in the published data. *For this reason, yearly analysis will not include 2024.*
- This data set, provided and maintained by the Los Angeles Police Department, includes a wide range of information about each reported crime, including unique report number, date and time of occurrence and reporting, victim and location information, crime codes, and more.

Original Attributes & Descriptions: Noted according to data source info.

Column	Description
dr_no	Unique identifier for police report made up of 2 digit year, 2 digit area ID, and 5 digits (9 digits total)
date_rptd	Date crime was reported, from 2020-2024
date_occ	Date crime occurred, from 2020-2024
time_occ	24 hour military time (ex: 2200)
area	Geographical areas within police department, numbered 1-21
area_name	Name designation that references landmark or surrounding community
rpt_dist_no	Four digit code that represents sub-area within area
part_1_2	Indicates part 1 or 2 offense
crm_cd	Indicates crime committed (same as crm_cd_1)
crm_cd_desc	Defines crime code provided
mocodes	Modus Operandi (activities associated with suspect in commission of crime)
vict_age	Two character numeric age
vict_sex	Victim sex (F: Female, M: Male, X: Unknown)
vict_descent	Victim descent code
premis_cd	Type of structure, vehicle, or location where the crime took place
premis_desc	Defines the premise code provided
weapons_used_cd	Type of weapon used in the crime
weapons_desc	Defines the weapon used code provided
status	Case status code (IC is default)
status_desc	Defines the status code provided
crm_cd_1	Indicates the crime committed; crime code 1 is primary and most serious
crm_cd_2	May contain code for an additional crime, less serious than 1
crm_cd_3	May contain code for an additional crime, less serious than 2
crm_cd_4	May contain code for an additional crime, less serious than 3
location	Street address of crime incident, rounded to nearest hundred block
cross_street	Cross street of rounded address
lat	Latitude, decimals standardized to include 6 decimal points
lon	Longitude, decimals standardized to include 6 decimal points

II. Data Import and Preparation

Steps

1. Verified Local Infile Setting

- Set `local_infile` using to enable `LOAD DATA LOCAL INFILE` for data set import.

2. Created Database

- Used `CREATE DATABASE` to create `crime_db`.

3. Created Table and Column Names

- Used `CREATE TABLE` to prepare for import of CSV file and add column names according to data source info.
- Initially using VARCHAR type (without size limits) for all columns to ensure successful import of more complex data (for example, numerical codes with trailing/leading zeros and military time formatted as string). Data types will be converted to proper data type during cleaning steps.
- Set dr_no as primary key.

4. Loaded Data Set From CSV

- Used CLI to import files for speed and efficiency.
- Data was imported using `LOAD DATA LOCAL INFILE` into `crime_data` table.

5. Verified Successful Import

- Confirmed data was imported by selecting all columns from `crime_data` table and using `DESCRIBE` to inspect columns. Total rows on import: 990,293.

6. Created Backup Data Set

- Copied raw data set as raw_crime_backup before further manipulation.
  - (Restore Process: Use `DROP TABLE` to delete the existing table and `CREATE TABLE` to restore from the backup)

III. Data Cleaning

Cleaning Summary

Category	Affected Columns	Issue Identified	Action Taken	Rationale
Null Values & Blanks	All columns	Missing or blank values in columns, represented as nulls or white spaces.	Replaced nulls/blanks with standardized placeholders like `'0000'`, `'Unknown'`, or `'X'`.	Ensures completeness for analysis and compatibility with SQL queries.
Duplicate Rows	All columns	2,876 duplicate rows found with identical values across all columns.	Removed duplicates using a `ROW_NUMBER` window function.	Prevents duplicate counting in analysis and ensures data integrity.
Data Type Standardization	All columns	Columns were imported as strings to ensure compatibility and needed to be converted to correct data type for analysis.	Converted dates to `DATE`, times to `TIME`, lat/lon to `DECIMAL`, and ages to `INTEGER`. Set size limits for string columns.	Ensures proper data types for accurate calculations and compatibility with SQL and Tableau.

Category	Affected Columns	Issue Identified	Action Taken	Rationale
Outliers	`vict_age`, `lat`, `lon`	Invalid values present (e.g., victim ages > 120, out-of-range geographic coordinates).	Removed rows with invalid outliers (e.g., victim age > 120 or latitude/longitude outside valid LA ranges).	Prevents skewed analysis and maintains realistic data.
Column Splitting	`mocodes`	Combined multiple M.O. codes in a single column, separated by spaces.	Split into 10 separate columns (`mo_1`–`mo_10`) using `SUBSTRING_INDEX`.	Enables detailed analysis of individual M.O. codes.
Code Mapping and Lookup	`vict_sex`, `vict_descent`, `crime_code`	Codes lacked descriptive values (e.g., `'M'` for Male or `'A'` for Asian). Crime codes separate from descriptions.	Mapped codes to descriptive values using lookup tables for `vict_sex`, `vict_descent`, and crime codes.	Enhances interpretability and analytical value of categorical data. Also reduces redundant columns.
Inconsistent Formatting	`crm_cd_desc`, `premis_desc`, `weapon_desc`, etc.	Inconsistent capitalization, extra spaces, and redundant descriptions.	Standardized capitalization, removed white spaces, and addressed duplicates descriptions (e.g., "RETIRED (DUPLICATE)").	Improves readability and consistency in textual data.
Unnecessary Columns	`crime_code_2`–`crime_code_4`, `premis_cd`, `weapon_code`, etc.	Columns were redundant or lacked sufficient documentation for analysis.	Dropped unnecessary columns with limited analytical value.	Simplifies data set and focuses on relevant attributes.
Outdated/ Invalid Entries	`premis_desc`, `crm_cd_desc`	Rows marked as "RETIRED (DUPLICATE)" or invalid premise codes.	Removed rows with invalid premise descriptions or outdated codes.	Ensures only valid and current data is included in analysis.
String Cleaning	`location`, `cross_street`, `crm_cd_desc`, etc.	Presence of leading/trailing spaces and inconsistent formatting.	Applied `TRIM` to remove white spaces and standardized text case across multiple columns.	Maintains consistency and improves usability for textual and geographic data.

Detailed Cleaning Steps and Code

1. Validated Data

Verified column values using SQL queries, checking for format and value consistency.

- `dr_no`: Removed 5 invalid rows not matching the 9-digit format.
- `vict_age`: Excluded 131 records where age exceeded 120.
- `vict_sex` and `vict_descent`: Mapped invalid or blank values to standardized placeholders ('X').

- Other columns: Confirmed values adhered to the documented ranges with no further action required.

2. Counted Distinct Variables

Identified unique values for categorical fields to ensure consistency. Compared to data source info for accuracy.

- `area`: 21 distinct areas & area names
- `crm_cd`: 140 distinct crime codes & descriptions
- `vict_sex`: 3 distinct victim sexes
- `vict_descent`: 19 distinct victim descents
- `weapons_used_cd`: 80 distinct weapon codes & descriptions
- `status`: 6 distinct status codes & descriptions
- `location`: 66,364 distinct locations
- `cross_street`: 10,354 distinct cross streets
- `premis_cd`: 315 distinct premise codes & 307 distinct descriptions
  - These query results indicate there are premise codes with the same descriptions. These repeated descriptions were investigated and resolved:

```sql
WITH desc_code_counts AS (
    SELECT
        premis_desc,
        COUNT(DISTINCT premis_cd) as unique_codes
    FROM crime_data
    WHERE premis_desc IS NOT NULL
    GROUP BY premis_desc
    HAVING COUNT(DISTINCT premis_cd) > 1
)

SELECT
    cd.premis_desc,
    GROUP_CONCAT(DISTINCT cd.premis_cd ORDER BY cd.premis_cd) as premise_codes,
    dcc.unique_codes as number_of_codes,
    COUNT(*) as total_records
FROM crime_data cd
INNER JOIN desc_code_counts dcc
    ON cd.premis_desc = dcc.premis_desc
GROUP BY
    cd.premis_desc,
    dcc.unique_codes
ORDER BY
    dcc.unique_codes DESC,
    cd.premis_desc;
```

    - Query results show "RETIRED (DUPLICATE) DO NOT USE THIS CODE" accounts for two of the repeated codes and white space fields for the other 8. These errors were removed from the data.

3. Handled Null Values

Replaced nulls and blanks with standardized placeholders for each column.

- Dates (`date_rptd`, `date_occ`): '0000-00-00'
- Strings (e.g., `crm_cd_desc`): 'Unknown'
- Numeric columns: Zero (`0` or `0000`).

4. Removed White Spaces

Used `UPDATE` and `TRIM` to remove leading/trailing white spaces in string columns: `area_name`, `crm_cd_desc`, `premis_desc`, `weapon_desc`, `status_desc`, `location`, `cross_street`.

5. Adjusted Data Types

Converted column data types and added size parameters; confirmed with `DESCRIBE`.

- Dates: Converted to `DATE`.
- Times: Initially kept as strings, later converted to `TIME` with proper formatting.
- Numeric fields (e.g., `vict_age`, `lat`, `lon`): Converted to `INTEGER` or `DECIMAL` as required.
- Strings (& Primary Key): Applied size limits (e.g., `VARCHAR`).

6. Converted Text to Lowercase

Standardized the capitalization in several textual columns of the `crime_data` table.

- Ensures that the first letter of each field (`crm_cd_desc`, `premis_desc`, `weapon_desc`, `location`, and `cross_street`) is capitalized, while the rest of the letters are converted to lowercase.
- Used `CONCAT()` and `SUBSTRING()` functions to adjust the case of each string field.

7. Removed Duplicates

Eliminated 2,876 duplicate rows using a `ROW_NUMBER()` window function.

```
WITH RankedData AS (
    SELECT
        *,
        ROW_NUMBER() OVER (
            PARTITION BY 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, weapons_used_cd,
                         weapon_desc, status, status_desc, crm_cd_1, crm_cd_2,
                         crm_cd_3, crm_cd_4, location, cross_street, lat, lon
            ORDER BY dr_no -- Ordered by primary key to keep the first instance
        ) AS row_num
    FROM crime_data
)
DELETE FROM crime_data -- Deleted rows where row_num is greater than 1 (duplicates)
WHERE dr_no IN (
    SELECT dr_no FROM RankedData WHERE row_num > 1
);
```

8. Standardized Column Names

Renamed columns for consistency and clarity, including replacing abbreviations.

```sql
ALTER TABLE crime_data
    RENAME COLUMN dr_no TO report_no,
    RENAME COLUMN date_rptd TO date_reported,
    RENAME COLUMN date_occ TO date_occurred,
    RENAME COLUMN time_occ TO time_occurred,
    RENAME COLUMN area TO area_code,
    RENAME COLUMN area_name TO area_name,
    RENAME COLUMN rpt_dist_no TO report_district_no,
    RENAME COLUMN part_1_2 TO part_no,
    RENAME COLUMN crm_cd TO crime_code,
    RENAME COLUMN crm_cd_desc TO crime_description,
    RENAME COLUMN mocodes TO mo_codes,
    RENAME COLUMN vict_age TO victim_age,
    RENAME COLUMN vict_sex TO victim_sex,
    RENAME COLUMN vict_descent TO victim_descent,
    RENAME COLUMN premis_cd TO premise_code,
    RENAME COLUMN premis_desc TO premise_description,
    RENAME COLUMN weapons_used_cd TO weapon_code,
    RENAME COLUMN weapon_desc TO weapon_description,
    RENAME COLUMN status TO status_code,
    RENAME COLUMN status_desc TO status_description,
    RENAME COLUMN crm_cd_1 TO crime_code_1,
    RENAME COLUMN crm_cd_2 TO crime_code_2,
    RENAME COLUMN crm_cd_3 TO crime_code_3,
    RENAME COLUMN crm_cd_4 TO crime_code_4,
    RENAME COLUMN location TO location,
    RENAME COLUMN cross_street TO cross_street,
    RENAME COLUMN lat TO latitude,
    RENAME COLUMN lon TO longitude;
```

9. Separated M.O. Codes

Split `mo_codes` into separate columns (`mo_code_1` to `mo_code_10`) using `SUBSTRING_INDEX`.

- First, identified number of M.O. columns needed by selecting `MAX(LENGTH(mo_codes)`.
- Then, added mo_co columns to table and used the `TRIM()` and `SUBSTRING_INDEX()` functions to handle white space and extract the correct code for each position.

10. Modified Time Format

Reformatted `time_occurred` by adding a colon separator (e.g., `1200` → `12:00`) and converted the column to `TIME` type.

```sql
UPDATE crime_data
SET time_occurred = CONCAT(LPAD(FLOOR(time_occurred / 100), 2, '0'), ':',
LPAD(MOD(time_occurred, 100), 2, '0'))
WHERE LENGTH(time_occurred) = 4;
ALTER TABLE crime_data
MODIFY COLUMN time_occurred TIME;
```

11. Standardized String Formatting

Applied trimming and uniform capitalization across location-related columns (`location`, `cross_street`).

- Removed remaining white spaces located inside strings (not trailing or leading).

12. Rechecked Duplicates and Nulls

After making other changes, including data type adjustments, the data set was rechecked for duplicates and null values. Confirmed no missing fields or duplicate entries.

13. Re-examined Categorical Variables

Calculated statistics for `vict_age` (mean, median, percentiles) and viewed distributions for other demographic fields to verify integrity.

Summary of Variables:

Variable	Unique Var.	Most Common	Count	Least Common	Count
area	21	Central (01)	68,064	Foothill (16)	32,612
crime	140	Vehicle - stolen	111,538	Train wrecking	1
victim sex	3	M (Male)	397,988	X (Unknown)	234,146
victim descent	19	H	293,267	S	56
premise	305	Street	255,221	Tram/streetcar(box wag on rails)*	1
weapon	80	000 (None)	661,005	M1-1 semiautomatic assault rifle	1
status	6	IC (Invest Cont)	788,348	CC (Unknown)	6

Variable	Unique Var.	Min	Max	Average
victim_age	100	0	120	29

14. Dropped and Renamed Columns

Removed redundant or undefined columns.

- MO Codes - Dropped grouped `mocodes` column.
- Area - Deleted `area_code`, renamed `area_name` to `area`.
- Crime - Mapped descriptions to codes.
  - Compared the counts for codes and their descriptions to confirm they matched before mapping.
  - Replaced crime codes:

```sql
-- Created mapping table
CREATE TEMPORARY TABLE crime_code_mapping AS
SELECT DISTINCT crime_code, crime_description, COUNT(*) AS count
FROM crime
GROUP BY crime_code, crime_description
ORDER BY crime_code;
-- Created indexes
CREATE INDEX idx_crime_data_code ON crime(crime_code);
CREATE INDEX idx_crime_code_mapping_code ON crime_code_mapping(crime_code)
```

```sql
-- Updated the main crime_code column with descriptions
UPDATE crime cd
JOIN crime_code_mapping ccm ON cd.crime_code = ccm.crime_code
SET cd.crime_code = ccm.crime_description
```

  - `DROP` crime code columns: The crime codes in columns 2-4 do not match any descriptions. They are not described in any of the data source documentation. As a result, these columns cannot contribute to analysis and will be dropped. Since crime_code_1 is also a duplicate of crime_code, it will be dropped.
  - Then, renamed crime code column and dropped crime description.
- Premise, Weapon, and Status - Dropped `premise_code`, `weapon_code`, `status_code` and renamed columns.
  - Previous cleaning steps confirmed matching codes and descriptions for these columns.

15. Replaced M.O. Codes

Using PDF attached to data source, numerical M.O. codes were manually matched to their description.

- Used python to extract data from MO code pdf using pdfplumber. (See attached notebook).
- Exported to csv file.

Created table in MySQL database and imported codes and descriptions from csv. Then, replaced mo_codes with descriptions.

```
UPDATE crime cd
JOIN mo_codes mc ON cd.mo_code_1 = mc.mo_code
SET cd.mo_code_1 = mc.description;
-- Repeated for mo_code_2 through mo_code_10
```

Replaced 0000 with 'None,' confirmed changes, and renamed columns for clarity.

16. Replaced Sex and Descent

Mapped `vict_sex` and `vict_descent` codes to descriptions using lookup tables.

```
-- victim_sex
CREATE TABLE victim_sex_lookup (
    sex_code CHAR(1),
    sex_description VARCHAR(50)
);
INSERT INTO victim_sex_lookup (sex_code, sex_description) VALUES
('M', 'Male'),
('F', 'Female'),
('X', 'Unknown');
-- victim_descent
CREATE TABLE victim_descent_lookup (
    descent_code CHAR(1),
    descent_description VARCHAR(50)
);
INSERT INTO victim_descent_lookup (descent_code, descent_description) VALUES
('A', 'Other Asian'),
('B', 'Black'),
('C', 'Chinese'),
('D', 'Cambodian'),
('F', 'Filipino'),
('G', 'Guamanian'),
('H', 'Hispanic/Latin/Mexican'),
('I', 'American Indian/Alaskan Native'),
('J', 'Japanese'),
('K', 'Korean'),
```

```
('L', 'Laotian'),
('O', 'Other'),
('P', 'Pacific Islander'),
('S', 'Samoan'),
('U', 'Hawaiian'),
('V', 'Vietnamese'),
('W', 'White'),
('X', 'Unknown'),
('Z', 'Asian Indian');
```

Mapped to crime table.

```
-- Updated victim_sex
UPDATE crime cd
JOIN victim_sex_lookup vsl ON cd.victim_sex = vsl.sex_code
SET cd.victim_sex = vsl.sex_description;
-- Updated victim_descent
UPDATE crime cd
JOIN victim_descent_lookup vdl ON cd.victim_descent = vdl.descent_code
SET cd.victim_descent = vdl.descent_description;
```

17. Created Data Set Backups and Exported Data

Saved cleaned and raw copies backups of the data set for recovery and exploration. Saved the cleaned table as a CSV file, prepared for exploratory analysis and visualization.

IV. Exploratory Data Analysis

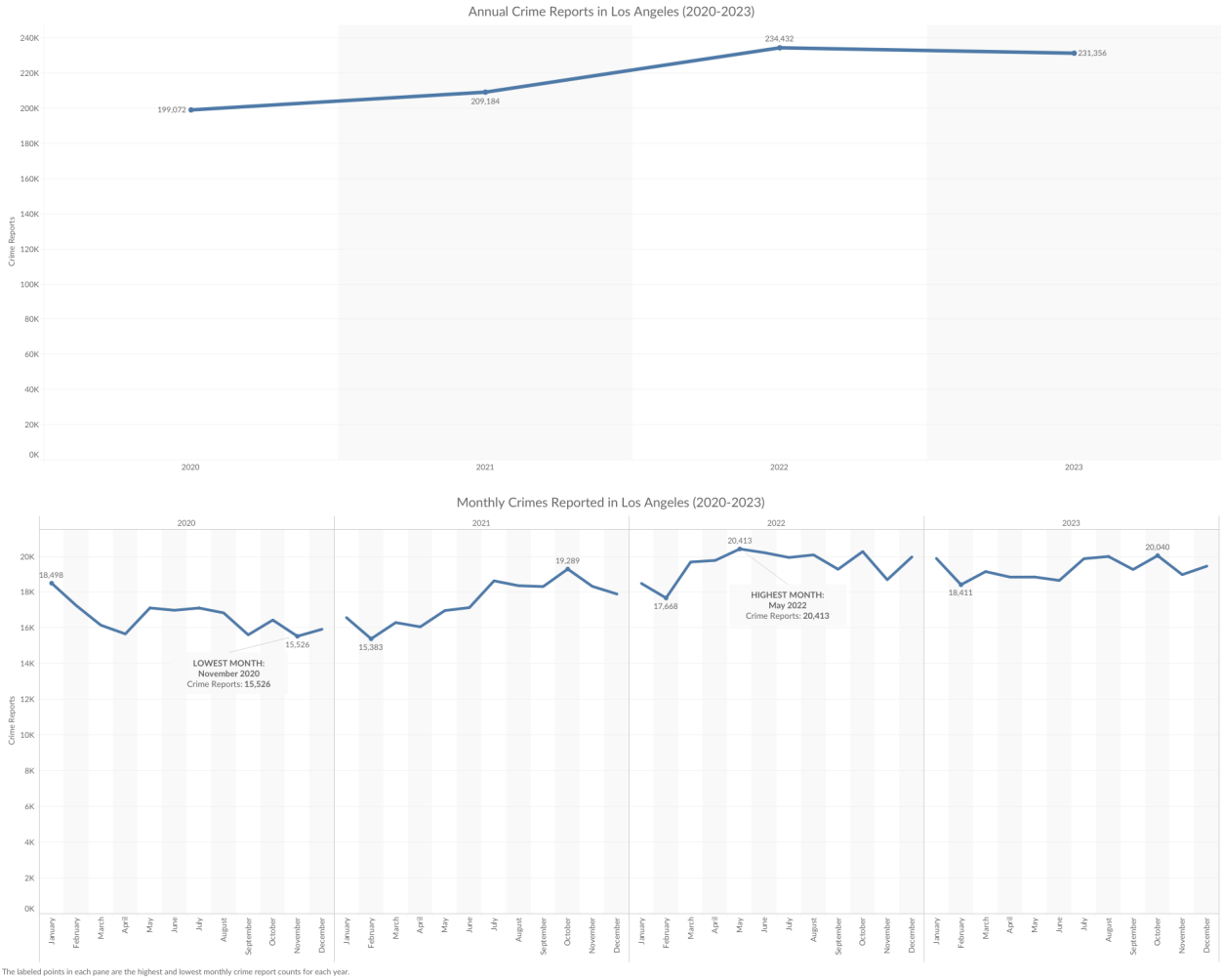
1. Examined Crime Volume and Trends Over Time

Queried annual and monthly crime counts:

```
SELECT YEAR(date_occurred) AS year, COUNT(*) AS yearly_crime_count
FROM crime
GROUP BY year
ORDER BY year;

SELECT YEAR(date_occurred) AS year, MONTH(date_occurred) AS month,
COUNT(*) AS monthly_crime_count
FROM crime
GROUP BY year, month
ORDER BY year, month;
```

Visualized trends with Tableau using line graphs for annual and monthly crime counts.



Insights

Yearly Trends:

- Crime steadily increased from 2020, peaking in 2022. A slight decline in 2023 may indicate stabilization post-COVID disruptions.

Monthly Trends:

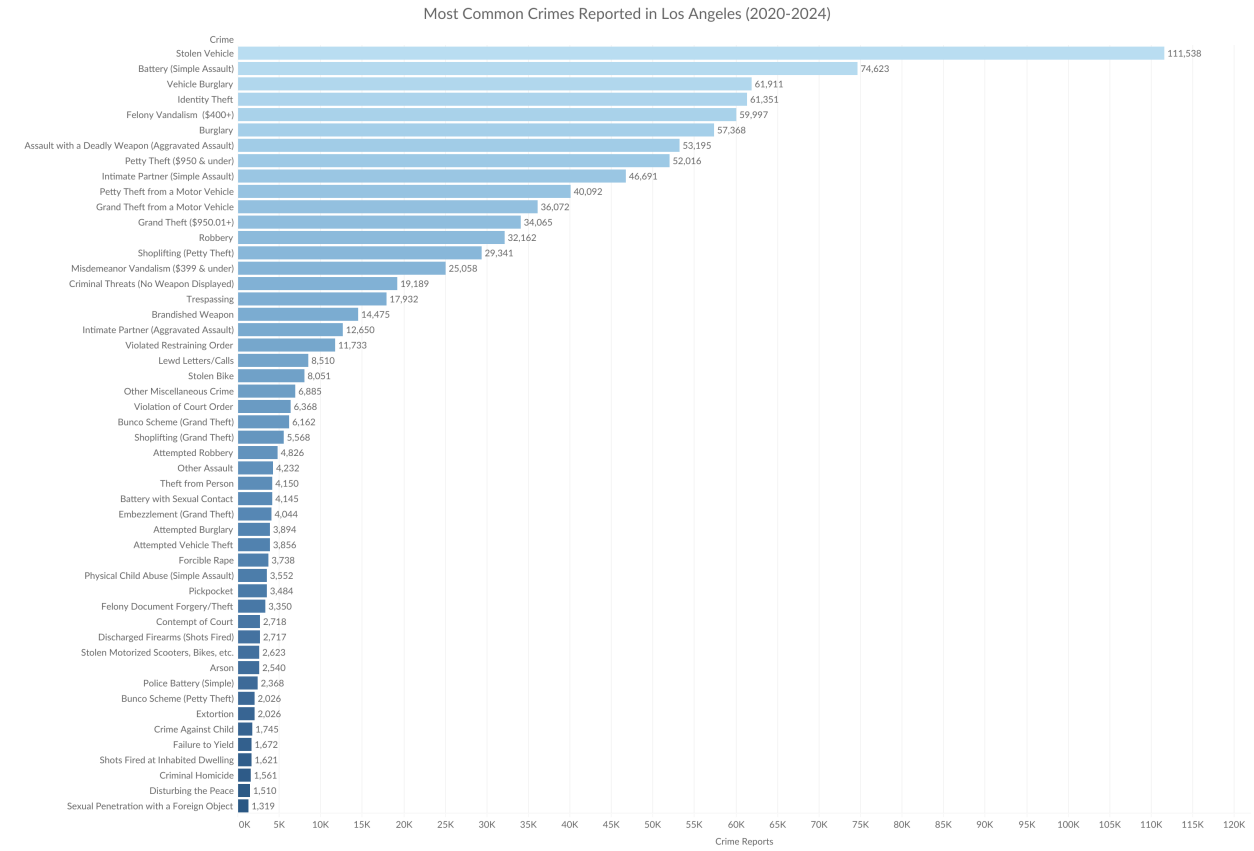
- A decrease in crime in 2020 correlates with pandemic restrictions.
- A sharp rise from mid-2021 onward reflects eased restrictions and potential economic factors.
- Seasonal fluctuations highlight peaks during summer months, likely linked to increased outdoor activity.

2. Identified Common Crime Types

Ranked crime types by frequency:

```
SELECT crime, COUNT(*) AS count
FROM crime
GROUP BY crime
ORDER BY count DESC;
```

Visualized the top 50 crime types using Tableau bar charts.



Insights

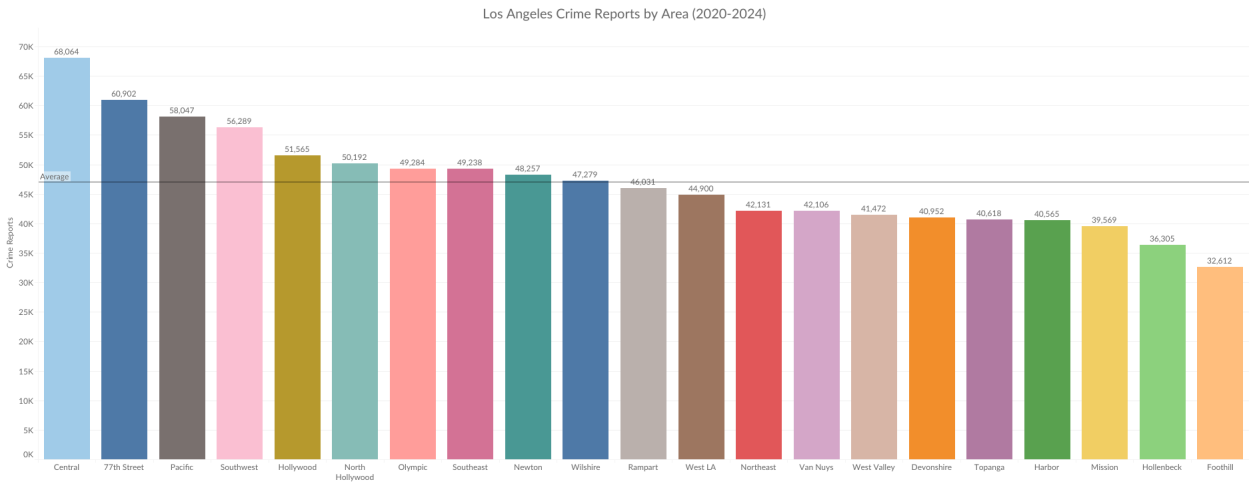
- **Dominant Crimes:** Theft and vehicle-related crimes (e.g., stolen vehicles, burglary) dominate, comprising over 40% of total reports.
- **Violent Crimes:** Assault (battery, aggravated assault) ranks highly, suggesting a need for targeted interventions in specific areas.
- **Long-Tail Crimes:** Rare crimes like train wrecking appear infrequently but still warrant monitoring.

3. Examined Geographic Area Crime Patterns

Queried crime counts by area:

```
SELECT area, COUNT(*) AS crime_count
FROM crime
GROUP BY area
ORDER BY crime_count DESC;
```

Created a bar chart in Tableau for crime counts by area.



Insights

- **Hot spots:** Central LA, 77th Street, and Pacific areas report the most crime, aligning with their population density and socio-economic challenges.
- **Low Crime Areas:** Foothill and Hollenbeck consistently report fewer crimes, possibly due to suburban or less dense settings.

4. Examined Victim Demographics

Analyzed victim sex and descent:

```sql
SELECT victim_sex, COUNT(*) AS victim_count
FROM crime
GROUP BY victim_sex;

SELECT victim_descent, COUNT(*) AS victim_count
FROM crime
GROUP BY victim_descent;
```

Calculated statistics for victim age and categorized age groups:

```sql
WITH ordered_ages AS (
    SELECT victim_age,
    ROW_NUMBER() OVER (ORDER BY victim_age) AS row_num,
    COUNT(*) OVER () AS total_count
    FROM crime
    WHERE victim_age > 0
)

SELECT
    COUNT(victim_age) AS total_records,
    COUNT(DISTINCT victim_age) AS unique_ages,
    MIN(victim_age) AS youngest,
    MAX(victim_age) AS oldest,
    ROUND(AVG(victim_age), 2) AS mean_age,
    ROUND(STDDEV(victim_age), 2) AS standard_deviation,

    (SELECT ROUND(AVG(victim_age), 2)
    FROM ordered_ages
    WHERE row_num IN ((total_count+1)/2, (total_count+2)/2)
```

```
    ) AS median_age,

    SUM(CASE WHEN victim_age < 18 THEN 1 ELSE 0 END) AS minors,
    SUM(CASE WHEN victim_age BETWEEN 18 AND 25 THEN 1 ELSE 0 END) AS young_adults,
    SUM(CASE WHEN victim_age BETWEEN 26 AND 40 THEN 1 ELSE 0 END) AS adults,
    SUM(CASE WHEN victim_age BETWEEN 41 AND 60 THEN 1 ELSE 0 END) AS middle_aged,
    SUM(CASE WHEN victim_age > 60 THEN 1 ELSE 0 END) AS seniors,

    (SELECT ROUND(AVG(victim_age), 2)
    FROM (
      SELECT victim_age
      FROM ordered_ages
      WHERE row_num IN (FLOOR(total_count*0.25), CEIL(total_count*0.25))
      ) AS percentile_25_subquery) AS percentile_25,

      (SELECT ROUND(AVG(victim_age), 2)
      FROM (
        SELECT victim_age
        FROM ordered_ages
        WHERE row_num IN (FLOOR(total_count*0.75), CEIL(total_count*0.75))
        ) AS percentile_75_subquery) AS percentile_75,

        (SELECT victim_age
        FROM crime
        WHERE victim_age > 0
        GROUP BY victim_age
        ORDER BY COUNT(*) DESC
        LIMIT 1) AS mode_age
FROM crime
WHERE victim_age > 0;
```

The max age returned by the query is 120 which is also the upper limit set for the column during cleaning. After investigating distinct values, confirmed that 120 is an error and the max age is actually 99. The error value was removed from SQL table and filtered from visualizations in Tableau.

```
DELETE FROM crime
WHERE victim_age = 120;
```

The query was rerun after filtering errors, which gave the following results:

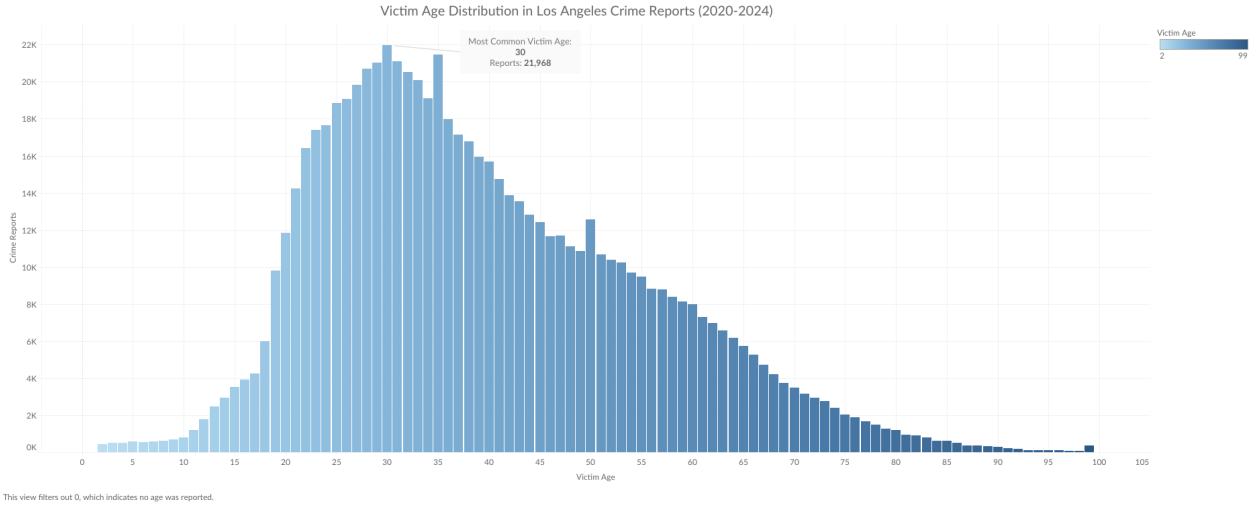
Query	Result
Total Records (excluding nulls)	726,007
Unique Ages	98
Youngest Victim	2
Oldest Victim	99
Mean Age	40
Standard Deviation	16
Median Age	37
Minors (<18)	25,367
Young Adults (18-25)	112,206
Adults (26-40)	288,456
Middle Aged (41-60)	217,981
Seniors (61+)	81,997
25th Percentile	28

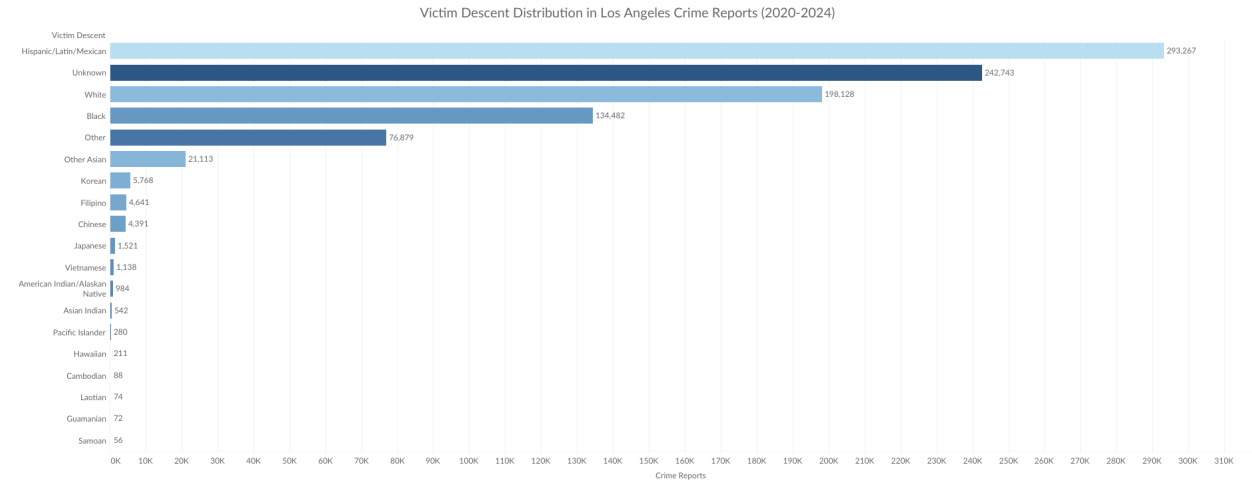
Query	Result
75th Percentile	50
Mode	30

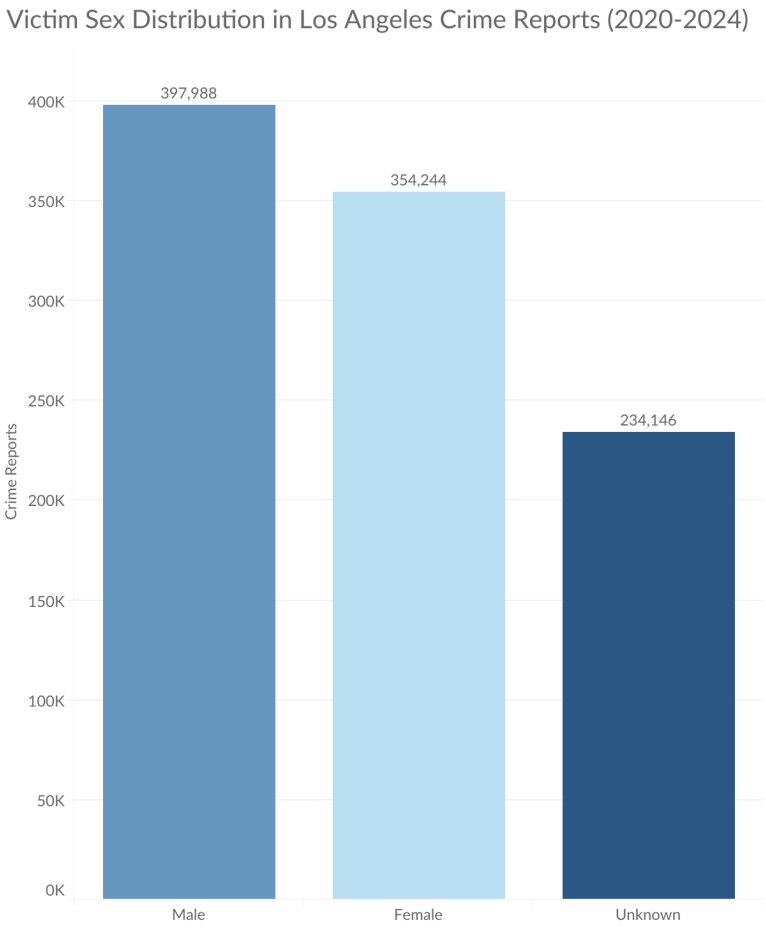
Calculated each age group's percentage of the overall victim population:

Age Group	Calculation	**Percentage**
Minors	25,367/726,007	**3.5%**
Young Adults	112,206/726,007	**15.5%**
Adults	288,456/726,007	**39.7%**
Middle Aged	217,981/726,007	**30%**
Seniors	81,997/726,007	**11.3%**

Finally, I visualized findings using bar charts and histograms.







Insights

- **Sex:** Males comprise a majority of victims, but a significant portion remains unclassified.
- **Descent:** Hispanic victims are the most reported, but high counts of "Unknown" descent reduce interpretability.
- **Age:** The average victim age is 40, with a majority aged 26-40. Victims below 18 represent only 3.5% of cases.
  - The age distribution is slightly right skewed, indicating there are more reports with younger victims than older ones.

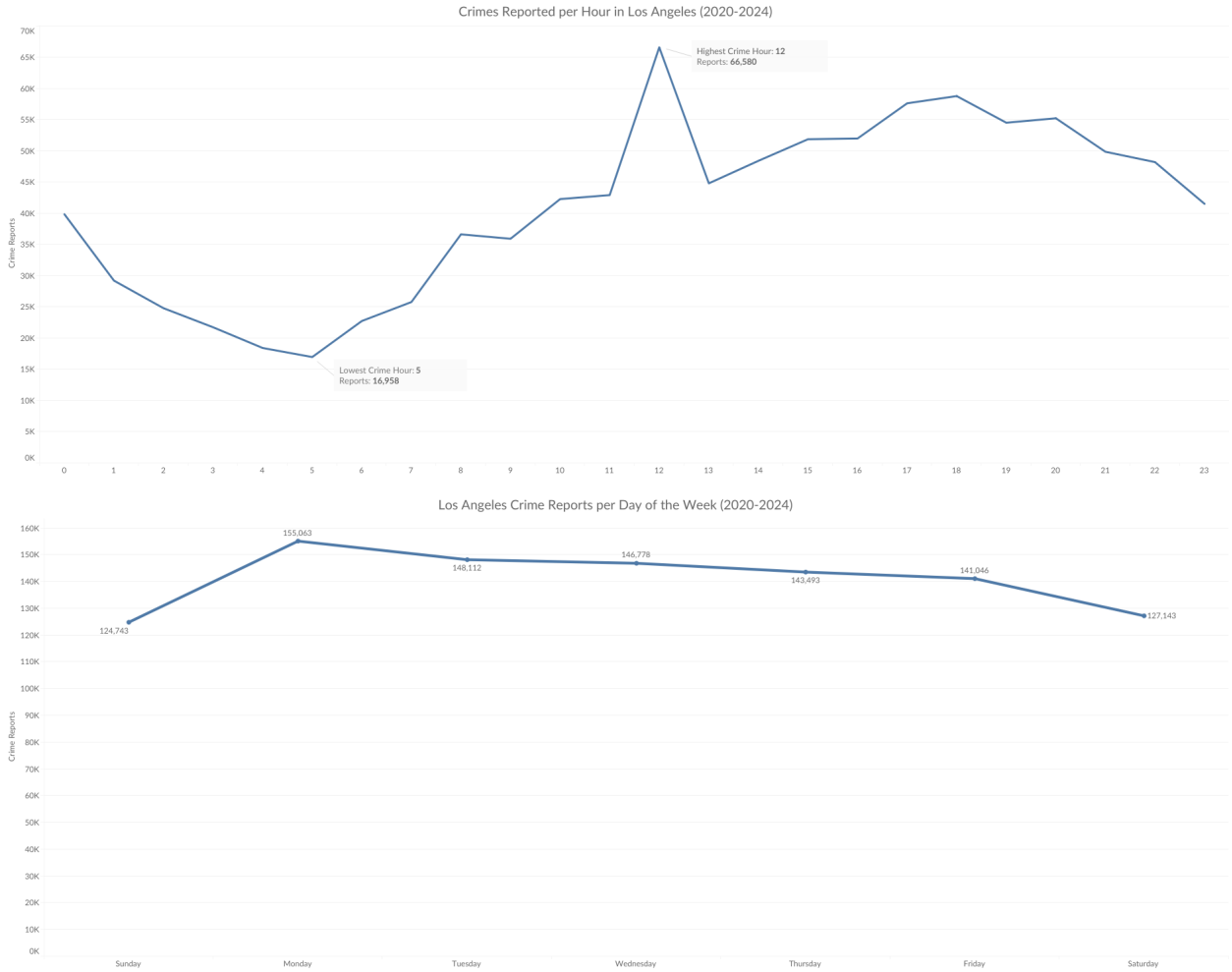
5. Examined Patterns Over Time and Day

Queried hourly and daily crime patterns:

```
SELECT HOUR(time_occurred) AS hour, COUNT(*) AS crime_count
FROM crime
GROUP BY hour
ORDER BY hour;

SELECT DAYOFWEEK(date_occurred) AS day_of_week, COUNT(*) AS crime_count
FROM crime
GROUP BY day_of_week
ORDER BY day_of_week;
```

Visualized results using Tableau line graphs for hour/day trends.



Insights

Hourly Patterns

- Crime peaks at midday and early evening, with a significant drop from 1-5 AM.

Daily Patterns

- Crime is relatively steady during the week (Tuesday-Friday), with lower reports on the weekend and a weekly high on Monday, likely due to weekend incident being reported after delays.

6. Examined Crime Severity (Part I or II)

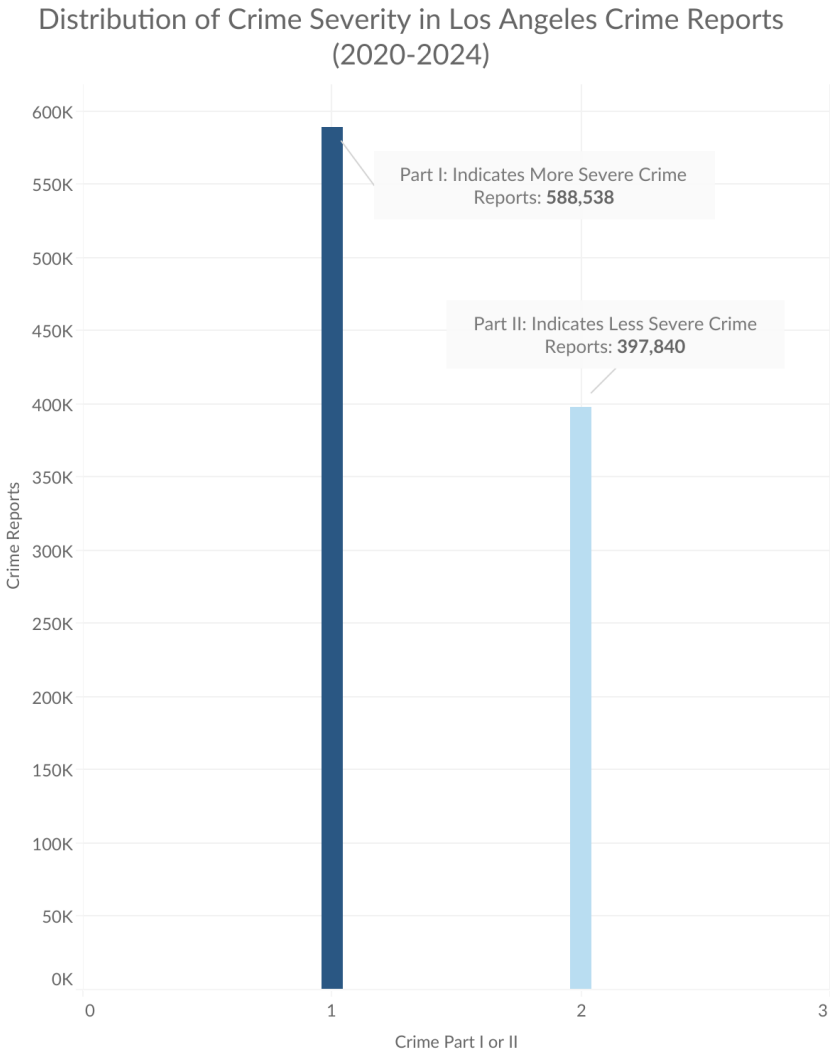
Notes on Part: According to the UCR handbook provided with the data source,

- Part I offenses include Criminal Homicide, Rape, Robbery, Aggravated Assault, Burglary, Larceny Theft, Motor Vehicle Theft, Arson, and Human Trafficking (Commercial Sex Acts and Involuntary Servitude). *These crimes are considered most severe/significant and rank higher if multiple crimes are committed.*
- Part II offenses are all other classifications.

Analyzed distribution of Part I (severe) vs. Part II (less severe) offenses:

```sql
SELECT part_no, COUNT(*) AS crime_count
FROM crime
GROUP BY part_no;
```

Visualized findings with a bar chart in Tableau.



Insights

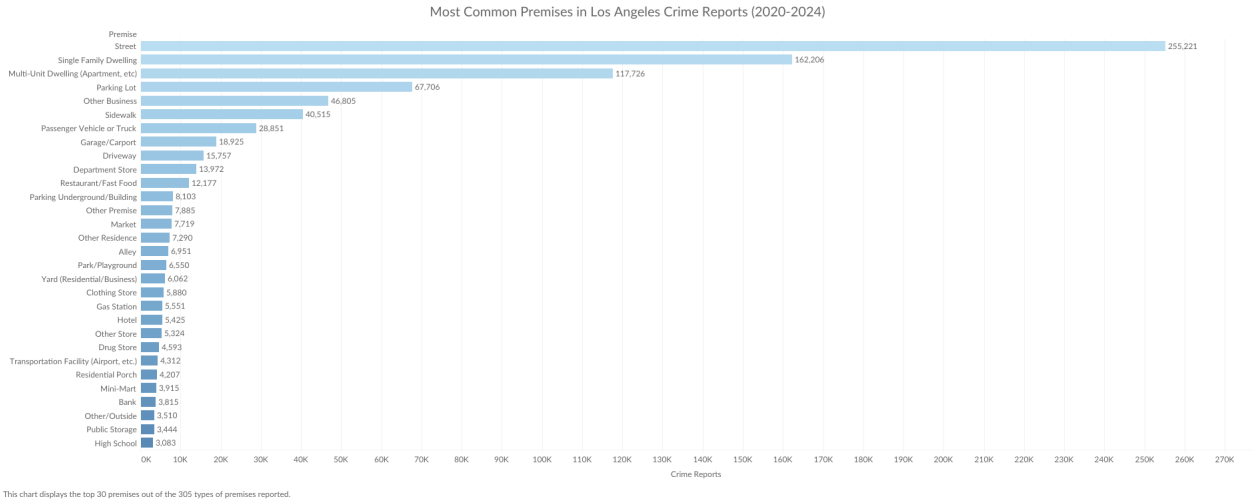
- Severe offenses like assault, robbery, and vehicle theft comprise the majority of reported cases.

7. Identified Common Premises

Queried premise descriptions and counts:

```sql
SELECT premise, COUNT(*) AS premise_count
FROM crime
GROUP BY premise
ORDER BY premise_count DESC;
```

Visualized common premises using bar charts.



This chart displays the top 30 premises out of the 305 types of premises reported.

Insights

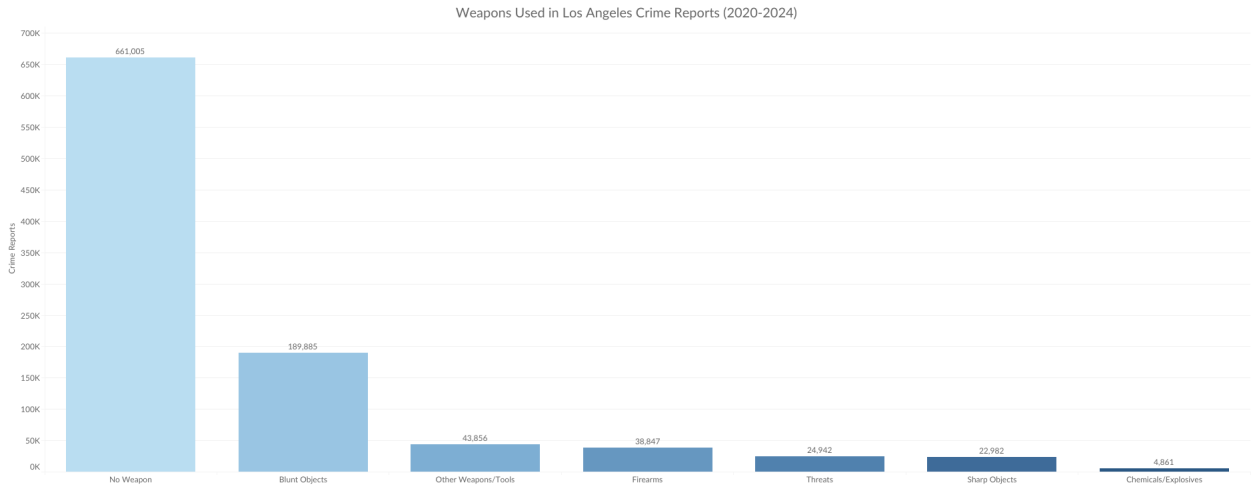
- Streets, private residences, and parking lots are the most frequent crime locations, correlating with high numbers of vehicle-related offenses.

8. Identified Commonly Used Weapons

Queried weapon types and their counts:

```sql
SELECT weapon, COUNT(*) AS weapon_count
FROM crime
GROUP BY weapon
ORDER BY weapon_count DESC;
```

Grouped weapon types in Tableau for clearer visualization.



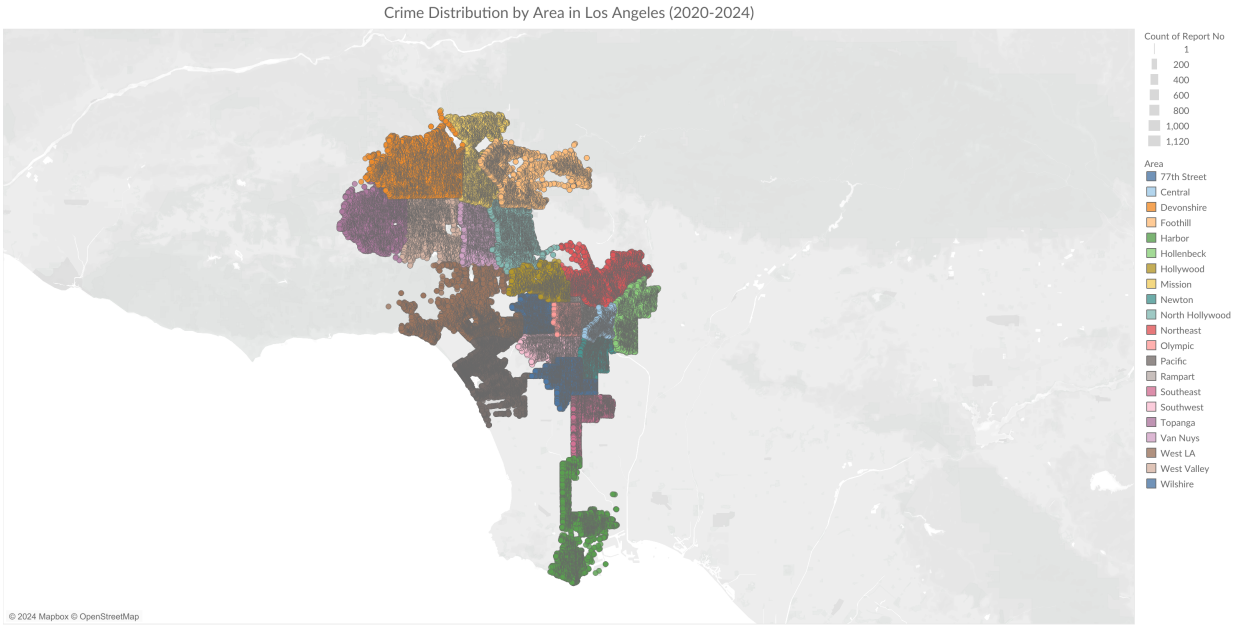
Insights

- A vast majority of crimes committed involve no weapon at all, including no threats or physical force. This coincides with the data on crime types, which shows that reports in recent years have involved significantly more theft than violent crimes. When violent crimes occur, they more commonly involve physical assault rather assault with a weapon.

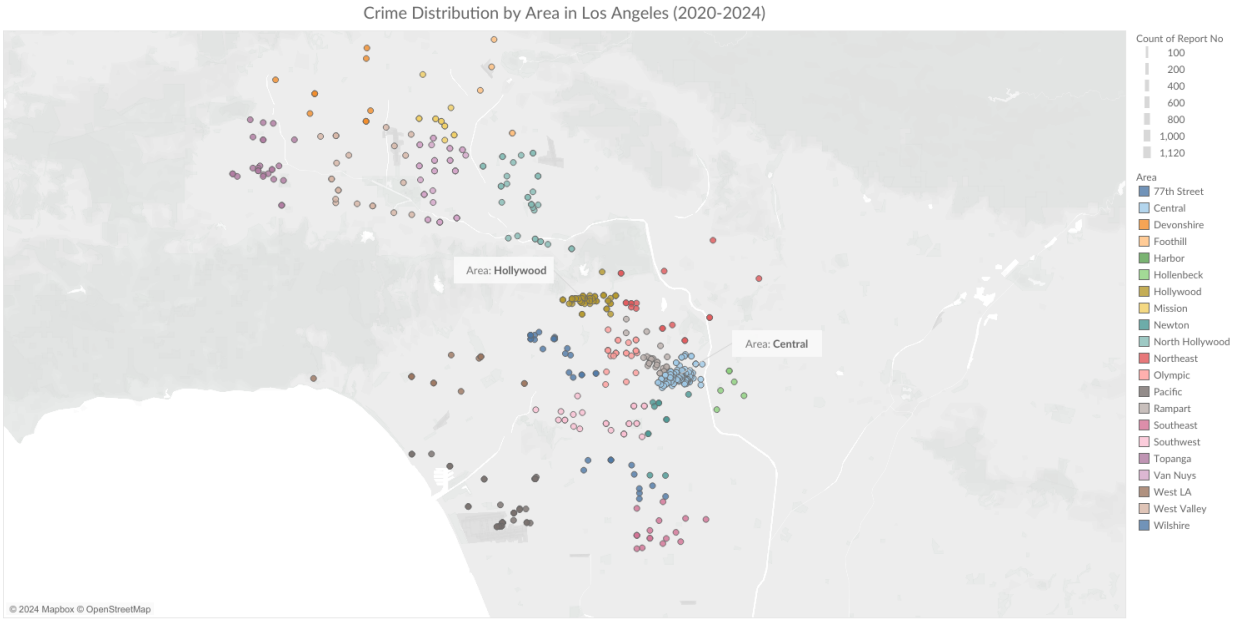
9. Visualized Crime Reports by Plotting Coordinates

Geographical Map

- Used latitude and longitude coordinates to create a spatial map in Tableau, using color to show areas and mark size to show number of reports.
- Excluded visible error points where `area_name` did not align with the coordinates, using Tableau's "exclude" feature.
- Added **Crime Type** and **Victim Sex** to the tooltip for better interactivity and contextual insights.



- Filtered counts of reports to show only locations with 100-1,120 (max) incidents per point. This significantly reduces the number of points on map, highlighting clusters of points where a high number of incidents have occurred.

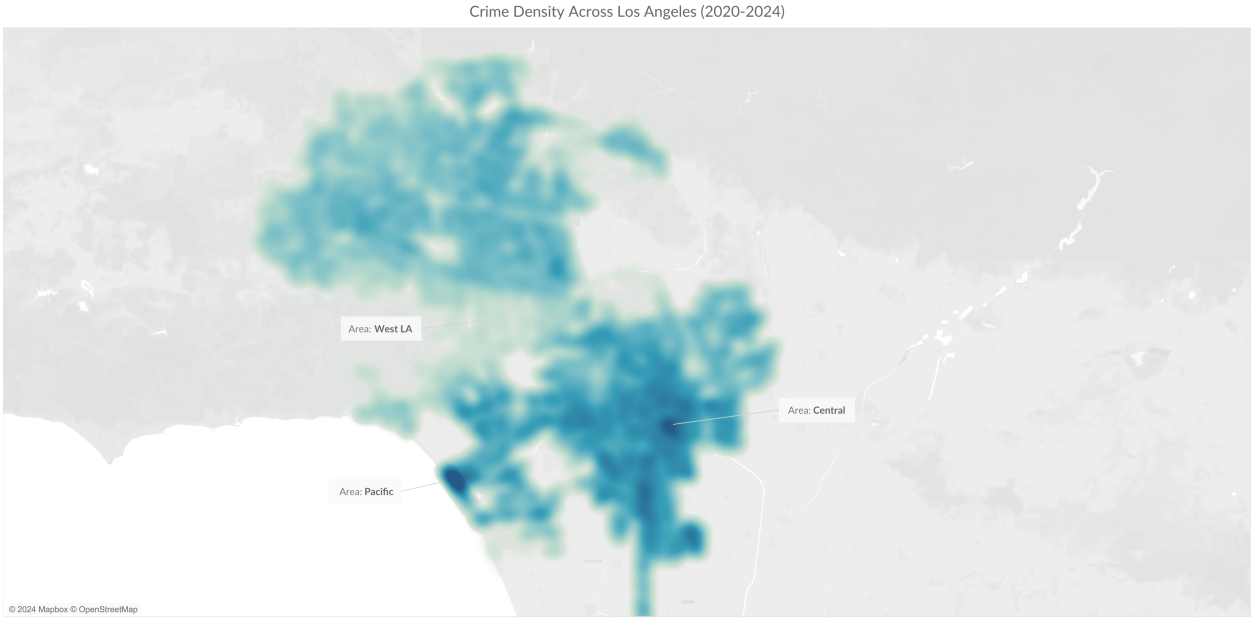


This map visualizes reported crimes across Los Angeles, with points color-coded by area and sized according to the number of reports per location. Hovering over a point reveals additional details, including the type of crime and the victim's sex, allowing for an interactive exploration of localized crime patterns.

- **Key Observations**:
  - Crime is heavily concentrated in downtown, inner areas of LA.
  - Central and Hollywood are two areas with significantly more points with over 100 incidents than other areas of LA. Following these, Rampart, Olympic, and Wilshire (all centrally located) have the next highest concentrations of crime.

Heat Map

- Created a density-based heat map showing the intensity of crime occurrences across Los Angeles.



This heat map illustrates the density of crime incidents across Los Angeles, transitioning from teal (low density) to blue (high density). Areas of higher density indicate crime hotspots, offering insights into regions with significant crime activity.

- **Key Observations**
  - Central LA stands out again as the most significant hot spot, along with other downtown areas,

followed by coastal areas such as Pacific.

– Peripheral regions like West LA, West Valley, and Harbor show minimal crime density.

V. Advanced Insights

1. Examined Temporal Trends in Dominant Crime Types

Used `GROUP BY` and `COUNT` functions to aggregate crime occurrences by time periods.

Year and Month:

```
SELECT YEAR(date_occurred) AS year, MONTH(date_occurred) AS month,
       crime, COUNT(*) AS crime_Count
FROM crime
GROUP BY year, month, crime
ORDER BY year, month;
```

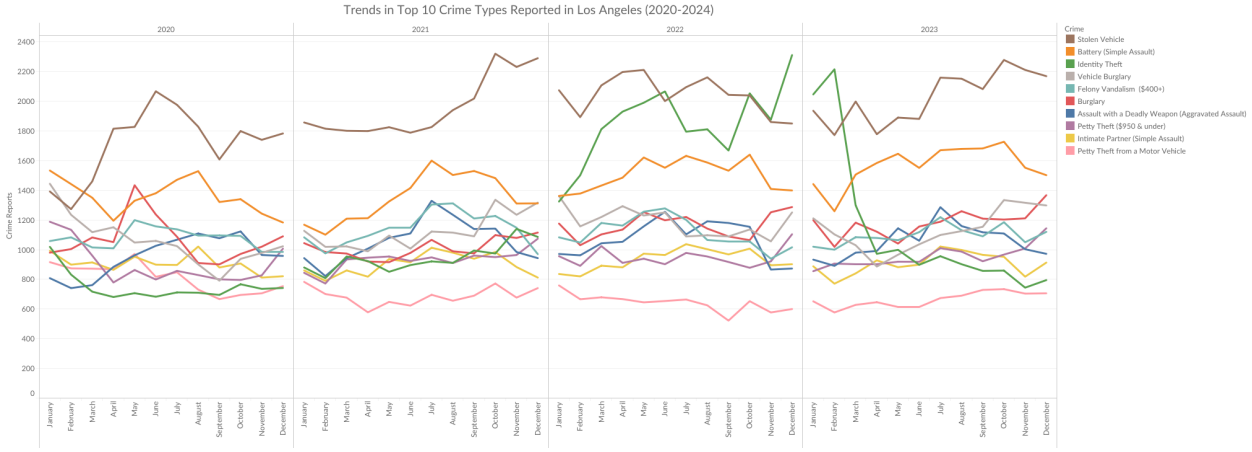
Day of the Week:

```
SELECT DAYNAME(date_occurred) AS DayOfWeek,
       crime, COUNT(*) AS crime_count
FROM crime
GROUP BY DayOfWeek, crime
ORDER BY FIELD(DayOfWeek, 'Monday', 'Tuesday', 'Wednesday', 'Thursday',
               'Friday', 'Saturday', 'Sunday');
```

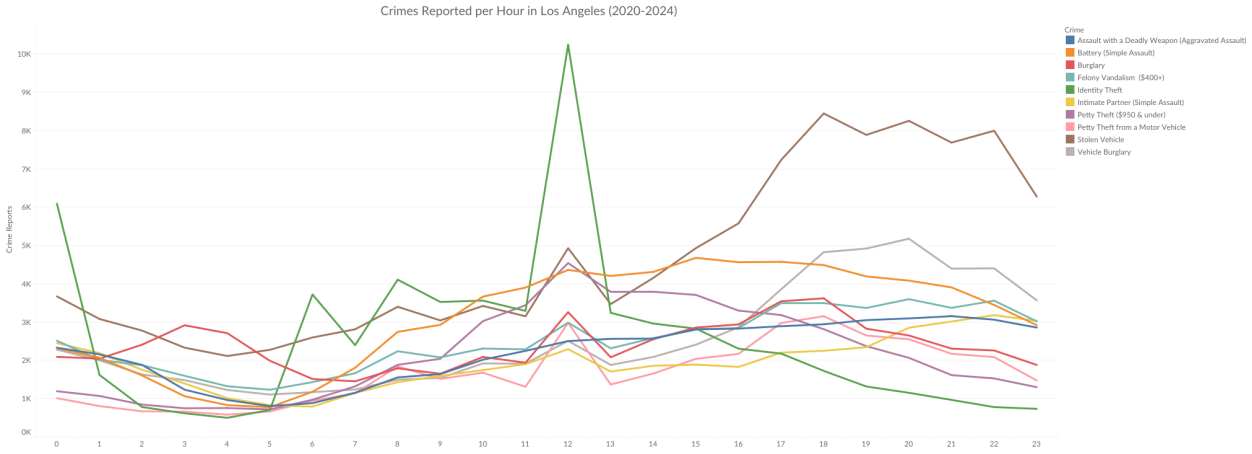
Hour:

```
SELECT HOUR(time_occurred) AS hour, crime, COUNT(*) AS crime_count
FROM crime
GROUP BY hour, crime
ORDER BY hour;
```

Created line charts showing trends in the top 10 crime types reported.



Trends throughout the day for the top 10 crime types:



Insights

- **Dominant Crime Types Across Years:**
  - Stolen Vehicle consistently has the highest number of reported cases across all years, indicating a persistent issue with vehicle theft in Los Angeles.
  - Battery follows as a major crime, reflecting recurring interpersonal violence.
- **Significant Increases/Decreases:**
  - Identity Theft shows a dramatic spike in early 2023 compared to previous years, suggesting a sudden surge in this crime type, possibly linked to economic factors.
  - Felony Vandalism and Vehicle Burglary exhibit steady or slight increases over time.
  - Burglary and Intimate Partner Assault remain relatively stable over the years, indicating no significant intervention or worsening in these areas.
- **Seasonal or Monthly Trends:**
  - Crimes such as Stolen Vehicle and Battery peak during warmer months (April to August) in most years. This could be tied to increased social activity or reduced school/work obligations during summer.
  - In contrast, crimes like Petty Theft from a Motor Vehicle do not exhibit a strong seasonal pattern, suggesting they occur consistently throughout the year.
- **Effects of External Events (e.g., COVID-19):**
  - Early 2020 shows relatively low crime levels across all types, likely due to the onset of the pandemic and associated lock downs. Post-pandemic recovery in 2021 shows a gradual uptick in most crime types, with Stolen Vehicle and Battery leading the rebound.
- **Extreme Shifts in Crime Types:**
  - Identity Theft has risen significantly over the years, becoming a more prominent crime type by 2023–2024, whereas Petty Theft has slightly declined in volume, indicating potential shifts in criminal behavior or law enforcement focus.
- **Long-Term Trends:**
  - Most crime types show an overall upward trajectory from 2020 to 2024, indicating a general rise in reported crimes. However, some crimes like Intimate Partner Assault remain relatively flat, which may suggest stable but unaddressed rates of occurrence.
- **Hourly Trends:**

- Vehicle theft and burglary peak in the evening (6 PM-11 PM), likely due to low light, reduced vigilance, and fewer witnesses.
- Assault incidents peak in the evening as well (6 PM–10 PM), correlating with active social hours.
- Identity theft shows an unusual peak at 12 PM, which may indicate time being estimated during reporting.

Potential Actionable Insights for Stakeholders

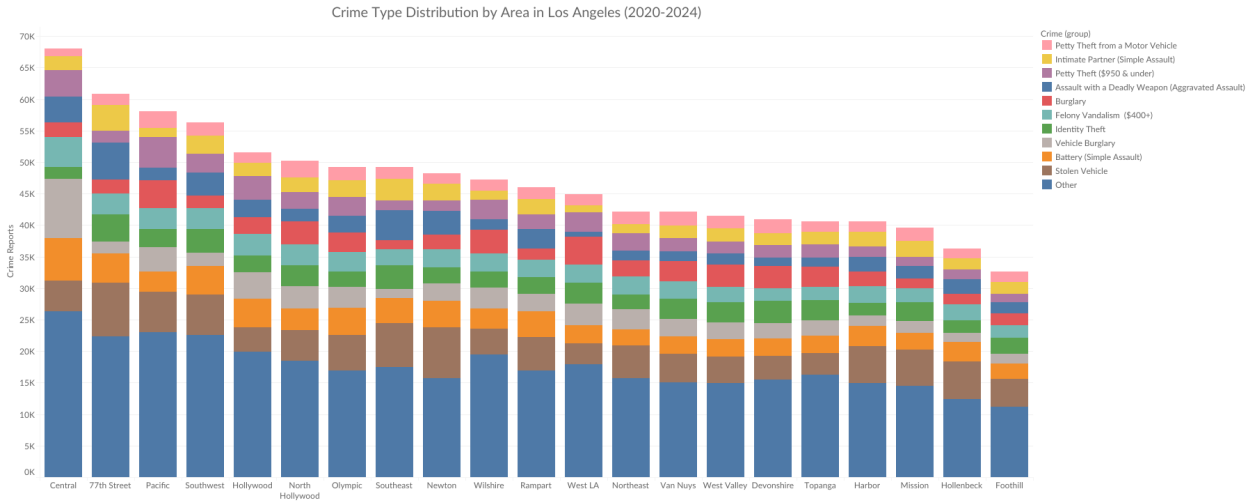
- **Prioritization for Law Enforcement:**
  - Resources should focus on curbing Stolen Vehicle crimes through preventive measures like vehicle tracking technology or community patrols.
  - Address the sharp increase in Identity Theft through public awareness campaigns and digital fraud detection measures.
- **Community Outreach and Prevention:**
  - Summer months could benefit from targeted campaigns or initiatives to reduce vehicle theft during peak times.
  - Neighborhood-specific programs can focus on preventing Felony Vandalism and Burglary, especially in areas showing consistent trends.

2. Examined Crime Types by Area

Grouped data area and crime type:

```sql
SELECT area, crime, COUNT(*) AS crime_count
FROM crime
GROUP BY area, crime
ORDER BY crime_count DESC;
```

Created a stacked bar chart, showing the counts for the top 10 crime types and grouping others, for each area in LA.

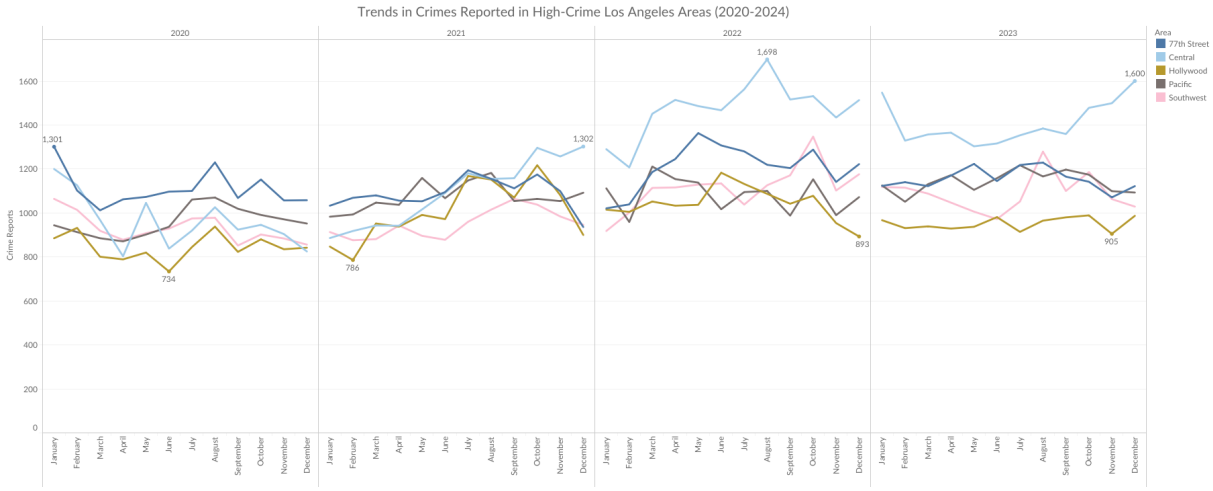


Insights

- **Dominant Crime Type Across Areas:**
  - Stolen Vehicle appears, again, to be the dominant crime type in nearly all areas, occupying a substantial portion of the total reports. This confirms a widespread issue with auto theft across the city.

- **Variation in Crime Composition:**
  - Some areas, such as Hollywood and North Hollywood, have relatively higher proportions of Identity Theft and Petty Theft, possibly due to tourist activity or dense commercial zones.
  - Southwest and Southeast have a more significant share of Battery and Assault with a Deadly Weapon, which may indicate specific safety challenges in these areas.

Created a line chart to examine trends in crime in the 5 areas with the most crime.



Insights

- While Central was a higher-crime area in 2022-2023, it trended similarly to other areas. This could indicate that the post-pandemic crime rise was concentrated in Central LA and surrounding downtown areas.

3. Correlation Analysis and Multivariate Temporal Analysis

Victim Demographics Analyzed correlations between victim demographics and crime types:

```
SELECT victim_age, victim_sex, crime, COUNT(*) AS crime_count
FROM crime
GROUP BY victim_age, victim_sex, crime;
```

Most Common Combinations

No.	Victim Age	Victim Sex	Crime	Crime Count
1	Unknown	Unknown	Vehicle Stolen	111,065
2	Unknown	Unknown	Petty Theft Motor Vehicle	23,874
3	Unknown	Unknown	Petty Theft Shoplifting	15,661
4	Unknown	Unknown	Burglary	12,248
5	Unknown	Unknown	Felony Vandalism	10,455
6	Unknown	Unknown	Trespassing	7,273
7	Unknown	Male	Burglary	5,236
8	Unknown	Unknown	Grand Theft	5,066
9	Unknown	Unknown	Robbery	4,971
10	Unknown	Unknown	Petty Theft	4,851

Excluding Unknown Ages

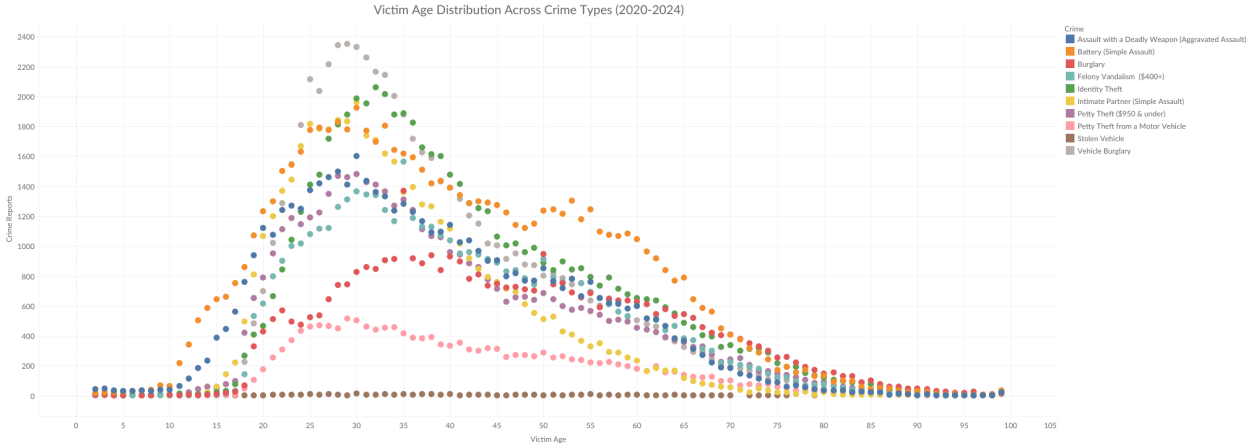
No.	Victim Age	Victim Sex	Crime	Crime Count
1	30	Female	Intimate Partner Assault	1,516
2	25	Female	Intimate Partner Assault	1,446
3	26	Female	Intimate Partner Assault	1,391
4	29	Female	Intimate Partner Assault	1,391
5	27	Female	Intimate Partner Assault	1,379
6	24	Female	Intimate Partner Assault	1,327
7	30	Male	Burglary from Vehicle	1,323
8	31	Female	Intimate Partner Assault	1,310
9	30	Female	Identity Theft	1,273
10	29	Male	Burglary from Vehicle	1,268

Insights

- **Prevalence of Missing Age Data:**
  - A significant portion of crime reports lack victim age information, with "Unknown" victims dominating across all major crime types. These include high-frequency crimes like vehicle theft (111,065 cases), petty theft from motor vehicles (23,874 cases), and shoplifting (15,661 cases), highlighting gaps in reporting or data collection.
- **Patterns in Intimate Partner Violence:**
  - Among reports with valid age and sex data, **females aged 24–31** are overwhelmingly represented in intimate partner - simple assault cases, indicating a high risk for this demographic.
- **Age-Specific Trends in Burglary from Vehicles:**
  - Males in their late 20s and early 30s (e.g., ages 29–31) are frequently victims of burglary from vehicles, with consistent counts (~1,300 cases per age group).
- **Identity Theft and Female Victims:**
  - Females around age 30 are prominently targeted for identity theft (1,273 cases), suggesting a demographic focus for such fraud-related crimes.

Then, visualized these relationships between variables using Tableau.

Age and Crime Type

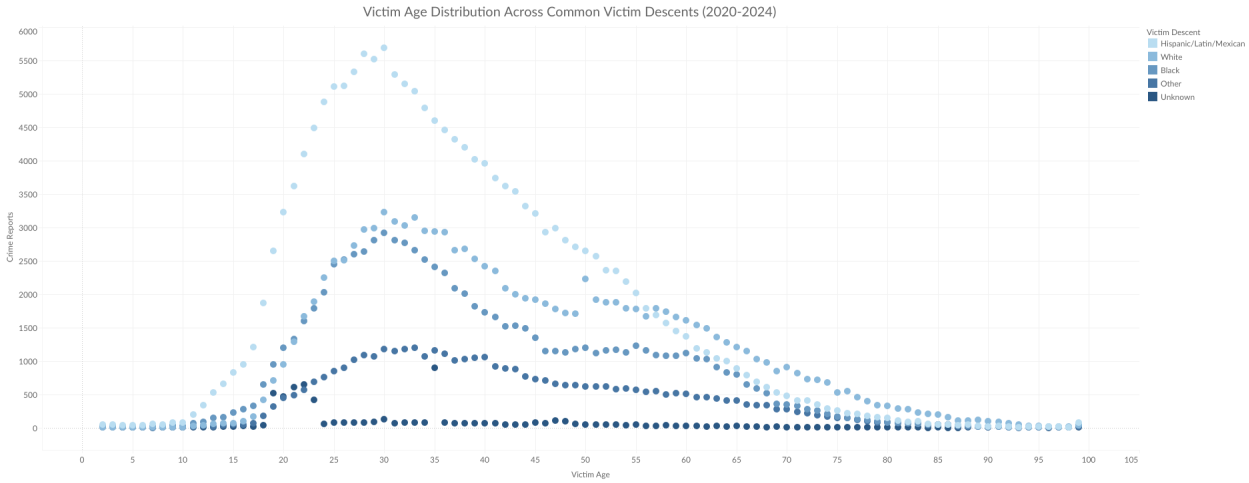


Insights

- The age distributions for prevalent crime types are somewhat similar, with some variation.

- Petty theft from a motor vehicle has slightly more younger victims than older ones. Intimate partner assault also disproportionately affects younger victims (women, as shown in SQL analysis) more than older ones.

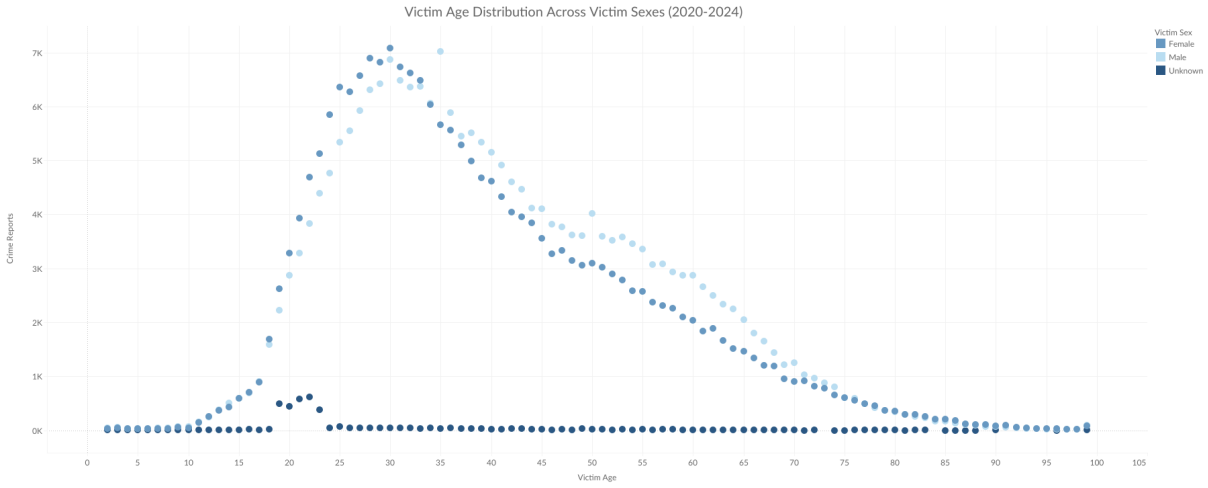
Age and Descent



Insights

- Hispanic victims are most prevalent in the 26–40 age group, reflecting Los Angeles's demographic distribution.
- Black and White victims show similar distributions, but Black victims are slightly over-represented in younger age groups (18–25) compared to White victims.
- Victims with "Unknown" descent are evenly distributed, highlighting data collection gaps that limit further demographic analysis.

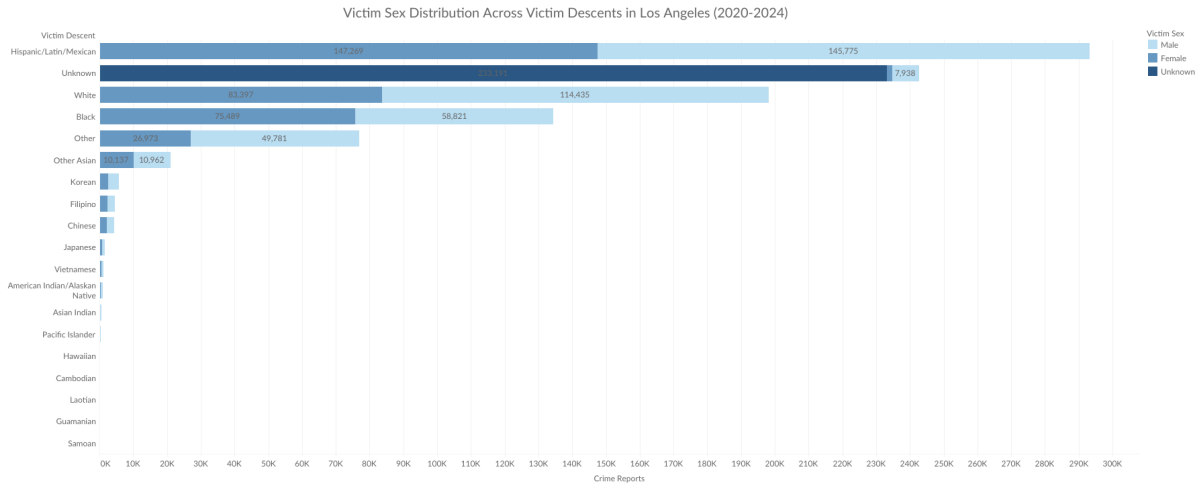
Age and Sex



Insights

- Male and female victims show similar age distributions, except males aged 45–70 are disproportionately represented in crime reports.

Descent and Sex

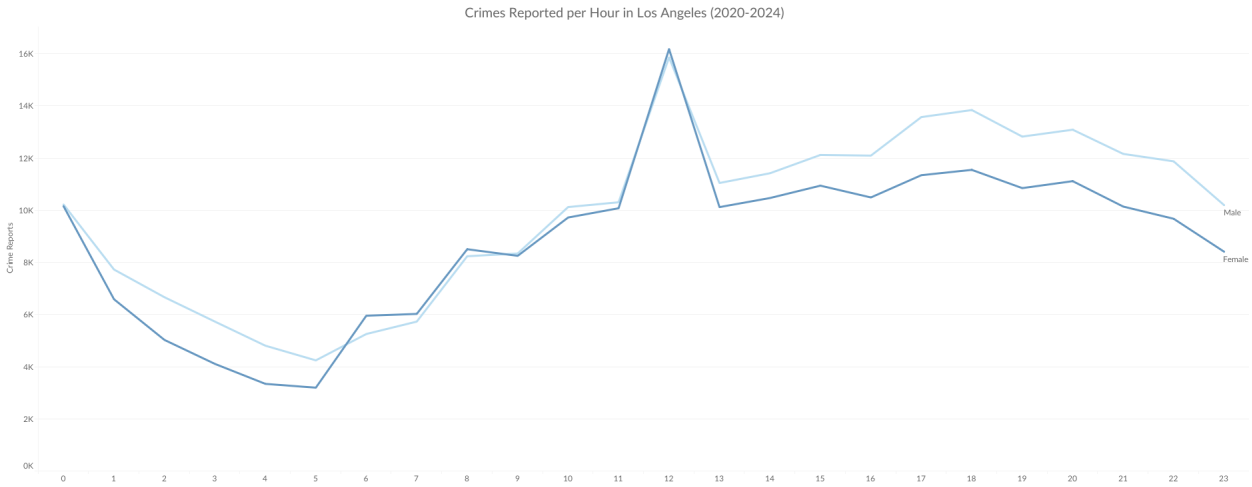


Insights

- Black female victims make up a larger portion of Black victims than males. In contrast, the distributions for other victim descents are relatively even or male-dominated. This indicates particular vulnerability of Black women and high rates of victimization.

Created line charts showing hourly, yearly, and monthly trends in the following variables:

Hourly Trends by Victim Sex



Insights

- Crimes reported by male and female victims remains similar throughout midday (6 AM-1PM, though males have more crime reports during early morning and late evening hours than females.

Weapon and Premise Studied weapon usage by premise type:

```
SELECT premise, weapon, COUNT(*) AS crime_count
FROM crime
GROUP BY premise, weapon;
```

Most Common Weapon/Premise Combinations

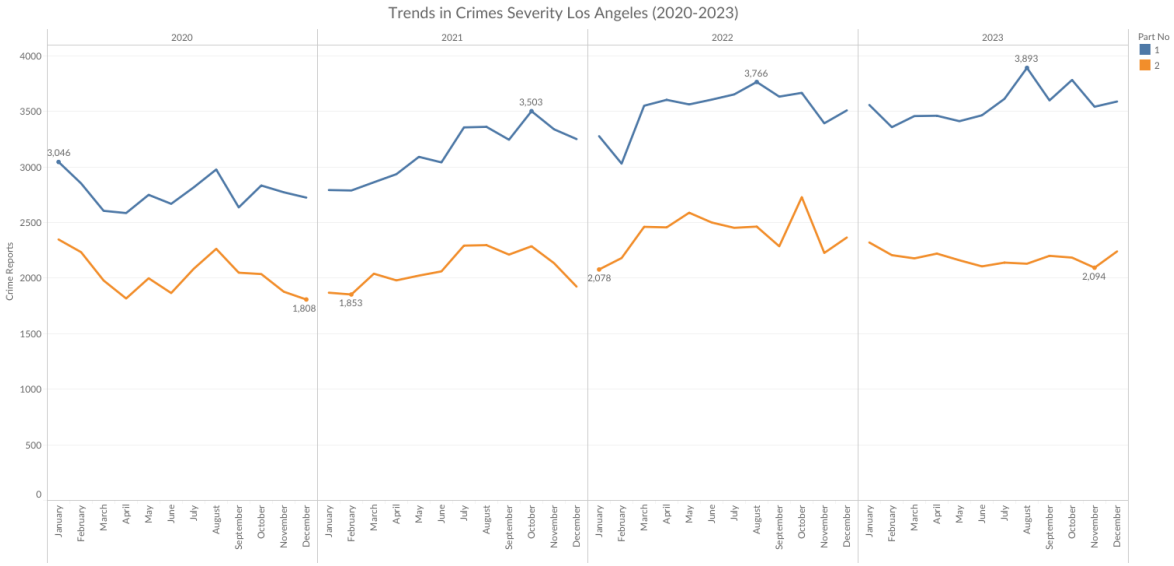
No.	Premise	Weapon	Count
1	Street	None	186,430
2	Single Family Dwelling	None	102,183
3	Multi-Unit Dwelling	None	63,634
4	Parking Lot	None	49,560
5	Single Family Dwelling	Strong Arm	37,839
6	Multi-Unit Dwelling	Strong Arm	35,576
7	Other Business	None	28,289
8	Street	Strong Arm	24,659
9	Vehicle	None	24,659
10	Garage/Carport	None	17,405

Weapon and Premise Insights

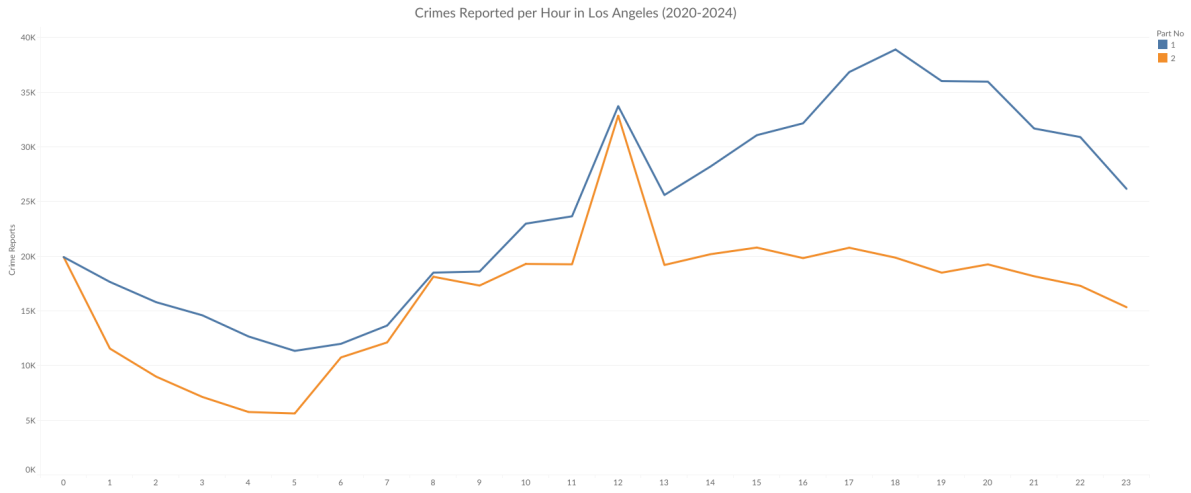
- Most crimes occur without weapons, with streets (186,430 cases) and single-family dwellings (102,183 cases) being the predominant crime locations. *This suggests that many incidents are property-related or non-violent.*
- Strong-arm (hands, fist, feet, or bodily force) incidents are prevalent in single-family dwellings (37,839 cases) and multi-unit dwellings (35,576 cases), indicating that physical confrontations often occur in residential settings.
- Handgun-related incidents (8,611 cases) are most frequently reported on streets, suggesting a higher risk of armed violence in public spaces, likely associated with robberies or disputes
- Premises like department stores (12,724 cases), parking lots (49,560 cases), and restaurants/fast food establishments (7,564 cases) predominantly report crimes without weapons, likely reflecting shoplifting, petty theft, or property damage.
- Premises such as vehicles (24,659 cases) and garages/carports (17,405 cases) primarily report crimes without weapons, aligning with trends in auto-related thefts.

Temporal Trends in Crime Severity Created line charts showing hourly, yearly, and monthly trends the number of Part I and II crimes.

Yearly and Monthly Trends in Part No



Hourly Trends in Part No



Part Insights

- Part I (more severe) crimes occur more than Part II throughout the entire day, though they rise significantly compared to Part II crimes in the afternoon and evening from 1 PM-11PM. In the morning hours from 6 AM-12 PM, this gap is much less significant.
- Yearly and monthly trends in crime severity are similar, though when increases in the amount of crime occurs, Part I crimes increase more steeply than Part II.

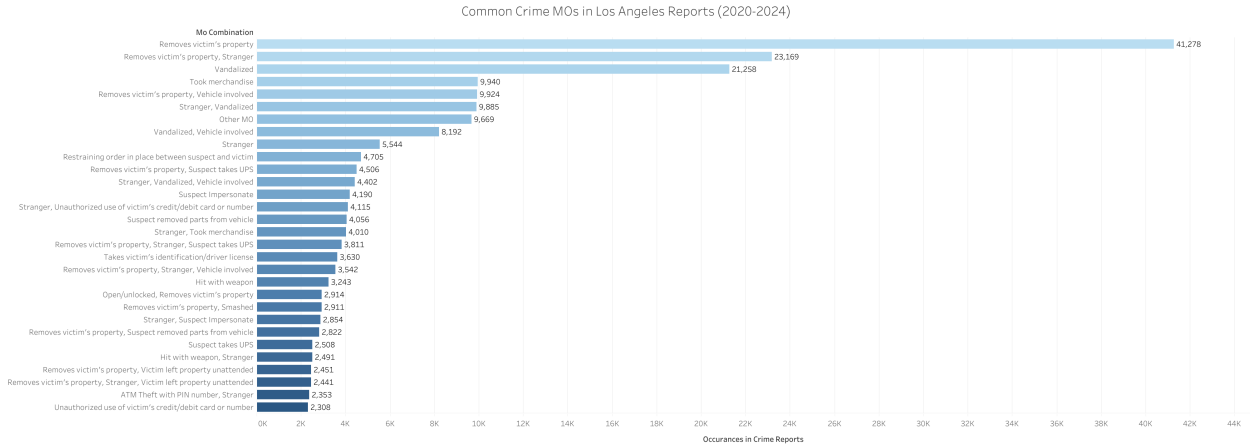
4. Examined Modus Operandi (M.O.)

Since the M.O. Codes were split into different columns in order to be mapped to relevant descriptions, they needed to be combined again for analysis.

To analyze the M.O. data, the `mo_1` to `mo_10` columns were unpivoted into a single column using `UNION ALL`. Duplicate entries within each crime report were removed, and the M.O. codes were normalized into alphabetical order using `GROUP_CONCAT` with sorting. This ensured that combinations such as 'Threaten to kill, Stranger' and 'Stranger, Threaten to kill' were treated as identical.

The consolidated data was grouped and counted to identify the most common M.O. combinations, which revealed property theft and strong-arm tactics as dominant patterns.

Exported the results for visualization in tableau, creating a bar chart showing the most common M.O.s.



M.O. Insights

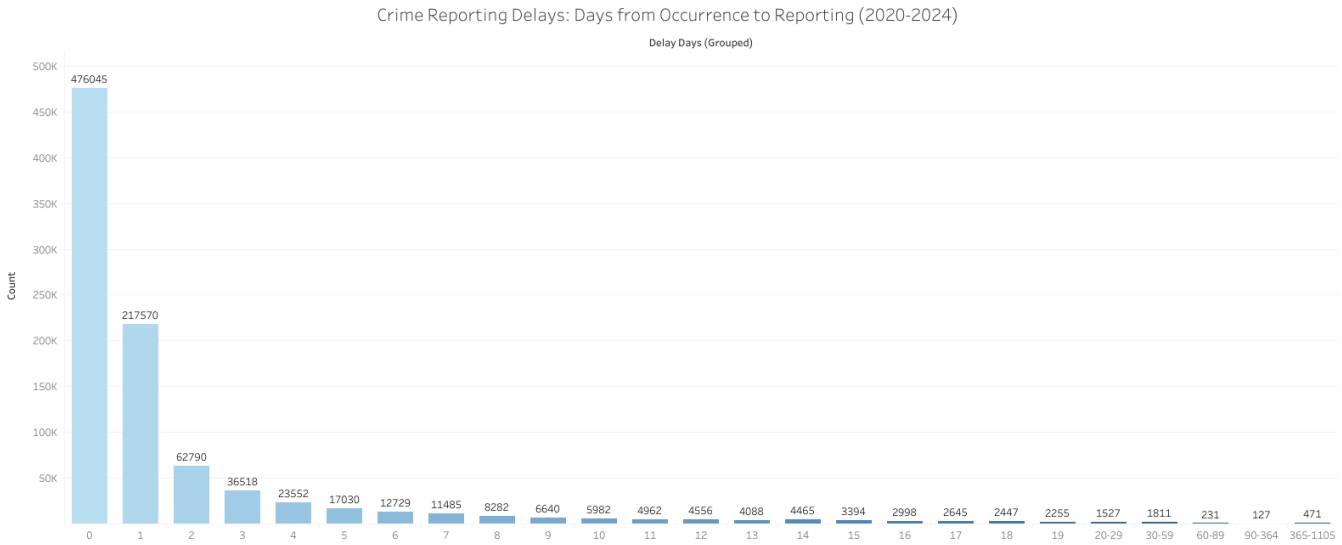
- While there are 1,000 unique combinations M.O.s, three dominate the others: 1. Removes victim's property (41,278), 2. Removes victim's property and is a stranger (23,169), 3. Vandalized (21, 258)
- Following these top 3, there is a steep drop off with the next M.O., "took merchandise" or shoplifting (9,940).
- Most of the M.O. combinations involve the taking of property (by far), then other types of theft and vandalization, usually involving vehicles.

5. Analyzed Reporting Delays

Investigated reporting delays using a `DATEDIFF` query. Exported results for import to Tableau.

```
SELECT DATEDIFF(date_reported, date_occurred) AS delay_days, COUNT(*) AS count
FROM crime_data
GROUP BY delay_days
ORDER BY delay_days;
```

Created a histogram to view the distribution of reporting delays.



This chart illustrates the distribution of delays (in days) between crime occurrence and reporting. The distribution is highly right-skewed, with the majority of reports filed within a few days. For clarity, delays exceeding 20 days have been grouped into bins as these represent a relatively insignificant portion of the data.

Reporting Delay Insights

- **Majority of Reports Filed Quickly:**
  - A significant proportion of crime reports (476,045 cases) were filed on the same day the crime occurred.
  - A steep drop follows, with 217,570 reports filed one day after the crime.
- **Short Delays Dominate:**
  - Reporting delays of up to three days account for a substantial portion of the data, highlighting that most victims or witnesses report crimes promptly.
  - Beyond three days, the count of reports diminishes progressively.
- **Long-Tail Distribution:**
  - The histogram shows a highly right-skewed distribution with a long tail, indicating that some crimes are reported weeks, months, or even up to a year later.

  - Delays beyond 20 days are grouped, but they collectively represent a small fraction of total reports.

- **Implications of Long Reporting Delays:**
  - Delays over 30 days are rare, with even fewer cases reported beyond 60 days.
  - These long delays could represent challenges such as hesitation to report, lack of immediate awareness of the crime, or systemic issues in reporting mechanisms.

- **Opportunities for Improvement:**
  - The high volume of same-day reporting indicates that many reporting systems work efficiently, especially for urgent cases.
  - Efforts to address longer delays could focus on community outreach, education, and reducing barriers to timely reporting.

VI. Key Findings

1. Temporal Crime Trends:

- Crime peaked in 2022, driven primarily by increases in vehicle theft and battery. This spike correlates with pandemic recovery periods, possibly reflecting heightened economic and social activity.
- Seasonal trends show higher crime rates during summer months, particularly in July and August, with significant peaks in the evening hours (6 PM–10 PM). Part I crimes, such as theft and assault, dominate these high-crime periods.

2. Geographical Area Hot Spots:

- Central LA and 77th Street consistently rank as the most crime-heavy areas, reporting high levels of vehicle theft, battery, and robbery.
- Residential crime patterns emerge in Pacific and Southwest, where burglary and vandalism are more prevalent.
- Lower-crime areas, such as Hollenbeck and Foothill, show a balanced distribution of minor crimes like petty theft and vandalism.

3. Victim Demographics:

- Adults aged 26–40 account for the majority of crime reports, reflecting their vulnerability to theft, assault, and robbery.
- Hispanic victims dominate across all age groups, aligning with Los Angeles's population demographics, while Black victims are slightly over-represented in younger age categories (18–25).
- Black women make up a higher portion of Black victim than males, while White male victims dominate over female ones. Women aged 20-30 (of all descents) are the most common victims of intimate partner assault.
- Unknown entries in victim descent (30%) and sex (20%) limit the reliability of demographic analyses, emphasizing the need for improved data collection.

4. Crime Severity:

- Part I crimes (violent and severe offenses) far outnumber Part II crimes (less severe). Vehicle theft is the most reported crime across all years, followed by battery and burglary.
- Part II crimes, such as identity theft and fraud, are growing steadily, suggesting a shift in criminal behavior toward non-violent but impactful offenses.

5. Reporting Delays:

- Most crimes are reported on the same day or within three days, indicating an efficient response for time-sensitive incidents.

  Delays exceeding 20 days are rare but disproportionately affect non-violent crimes like fraud and identity theft, where discovery often happens weeks or months after the incident.

VII. Dashboard Planning

1. Main Dashboard Panels:

- Crime Trends Over Time: Line charts for yearly/monthly trends, filterable by crime type and area.
- Victim Demographics: Bar charts showing age, sex, and descent distributions.
- Geographic Distribution: Heat maps and point maps to highlight high-crime areas like Central LA.

2. Supporting Panels:

- Modus Operandi: Bar charts showing top M.O. combinations by area or crime type.
- Crime Severity: Comparative charts for Part I and Part II crimes.

3. Interactivity:

- Filters for time, location, crime type, and demographics.
- Tool tips with details on victim information and weapon use.

4. Custom Features:

- Drill-downs for specific areas or crime types.
- Animations showing crime trends over time.

VIII. Final Recommendations

1. Crime Prevention and Awareness Initiatives

- **Vehicle Theft Prevention:** Increase public awareness about vehicle security, especially in hot spots like Central LA and 77th Street. Campaigns could include distributing steering wheel locks or promoting GPS trackers.
- **Community Engagement in High-Crime Areas:** Host neighborhood meetings in Pacific and Southwest to educate residents about burglary prevention, emphasizing home security measures and reporting suspicious activity.

2. Targeted Law Enforcement Strategies

- **Resource Allocation:** Deploy additional patrols in high-crime areas during peak hours (6 PM–10 PM) and high-crime months (July–August).
- **Focus on Violent and Property Crimes:** Allocate resources to areas with frequent strong-arm crimes, such as streets and sidewalks, to deter assaults and robberies.

3. Improving Data Collection

- Enhance reporting mechanisms to address gaps in victim demographic data. Mandatory fields for sex and descent, alongside public awareness campaigns, could reduce the high proportion of unknown entries.
- Adopt clear protocols for capturing M.O. data to standardize crime analysis and provide actionable insights.

4. Digital and Economic Crime Mitigation

- **Combat Identity Theft and Fraud:** Collaborate with financial institutions to identify patterns in fraud targeting specific age groups (e.g., women aged 30–40). Launch online safety campaigns tailored for seniors and working professionals.
- **Public Reporting Awareness:** Encourage timely reporting of non-violent crimes like fraud through simplified online platforms, reducing reporting delays and improving the accuracy of analyses.

5. Public-Focused Resources

- **Interactive Dashboards:** Develop user-friendly dashboards for public access, providing real-time updates on crime trends by area, crime type, and time of day.
- **Safety Tips:** Include localized safety tips and resources, such as contact information for local police stations and crime prevention workshops.

IX. Limitations and Future Directions

1. Limitations

- **Limited 2024 Data**:
  Due to the LAPD's transition to a new FBI-mandated reporting system and technical issues, 2024 data is incomplete and less frequently updated, affecting the accuracy and reliability of analyses for this year.
- **Incomplete Victim Demographics**:
  A significant portion of victim demographic data, including age, sex, and descent, is marked as "Unknown," limiting the depth and accuracy of analyses involving victim characteristics.

2. Future Enhancements

- Incorporate external socio-economic and census data for contextual analysis.
- Develop predictive models to forecast crime hot spots.
- Analyze trends in repeat offenses and victim patterns for enhanced prevention strategies.

X. References

1. Crime Data from 2020 to Present - Los Angeles City Open Data (Access data set, UCR handbook, COMPSTAT and M.O. Codes)
2. LA Crime Dashboard (Tableau)