Predictive Risk Modeling for Safety Interventions in Transportation Networks Using Spatial Crime History

Nazmus Sakib Pallab
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
npallab2@illinois.edu

Jiarui Yu
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
jiaruiy9@illinois.edu

Favour Jack
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
frjack2@illinois.edu

Muhammad Fahad Ali
Civil & Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
mali19@illinois.edu

***Abstract*—Integrating personal safety into transportation and pedestrian planning requires systematic use of crime data. Information on crime location, time, and type can be analyzed to identify unsafe streets, intersections, and transit hubs, uncovering vulnerable areas in the urban network. Such insights enable engineers to propose design interventions such as reducing dead-end streets, improving pedestrian connectivity, and strategically relocating public transit drop-off points to enhance safety.**

In this study, raw crime record data will be transformed into actionable hotspot maps and predictive risk models to optimize the allocation of traffic police and patrol routes, ensuring coverage in the areas of highest need. Using advanced machine learning techniques, the study predicts crime types based on factors such as location, time of day, victim profile, and premises description. These results provide Civil Engineers and Urban Planners with evidence-based tools to prioritize infrastructure improvements and safety investments, while also identifying specific locations likely to evolve into future hotspots for proactive deployment of patrols, surveillance, and safety infrastructure.

Together, these predictive and spatial approaches are expected to enhance response efficiency and guide long-term city planning initiatives—from upgrading street lighting and redesigning public spaces to improving transit accessibility and targeting community resources—thereby strengthening the overall resilience and safety of urban infrastructure.

***Index Terms*—Transportation safety, Crime data analysis, Predictive risk modeling, Hotspot mapping, Machine learning, Urban infrastructure planning, Pedestrian safety**

I. Description of Dataset

The dataset used for this project is the ***"Crime Data from 2020 to Present"*** dataset for the City of Los Angeles, which is publicly available on DATA.GOV [1]. It is maintained and released by the Los Angeles Police Department (LAPD) as part of the city's open-data initiative, based on official crime reports filed by law enforcement officers.

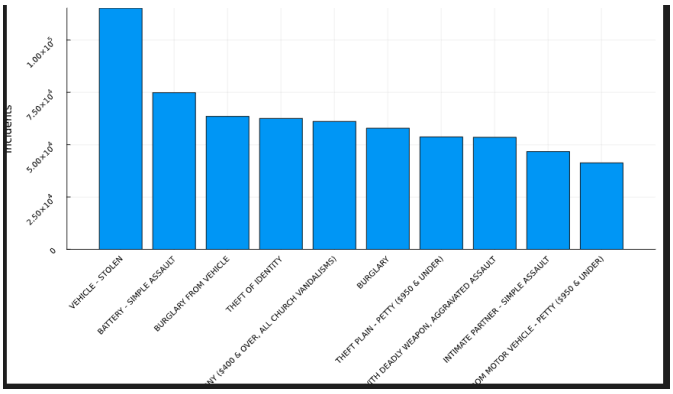
The dataset is provided in CSV format and contains over 1 million rows of crime incidents. The full dataset consists of 28 columns, while our project will focus on the following 12 key attributes:

- `DR_NO (Text)`: Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits.
- `Date Rptd (Floating and Timestamp)`: Date the incident was reported.
- `DATE OCC (Floating and Timestamp)`: Date the incident occurred.
- `TIME OCC (Text)`: Time the incident occurred.
- `AREA (Text)`: Area where the incident occurred.
- `AREA NAME (Text)`: ID of the area where the incident occurred.
- `Rpt Dist No (Text)`: A four-digit code that represents a sub-area within a Geographic Area.
- `Vict Age (Text)`: Age of the victim.
- `Vict Sex (Text)`: Sex of the victim.
- `Vict Descent (Text)`: Descent of the victim.
  - `LAT (Number)`: Latitude coordinate of the incident.
  - `LON (Number)`: Longitude coordinate of the incident.

This dataset provides both spatial (latitude/longitude, area, district) and socio-demographic (victim age, sex, descent) attributes, along with temporal information (date and time of crime occurrence), enabling spatial, temporal, and predictive risk modeling for transportation safety interventions.

II. Exploratory Data Analysis (EDA)

A. Crime Type Distribution



Group and rank by "Crm Cd Desc". Plot 1: bar chart of top 10 crime types. Insight: most frequent vs. least frequent crimes. (Favour Jack)

Check if image loads properly.

B. *Temporal Analysis*

(i) Temporal Patterns of Different Crime Categories:

The temporal profile of generalized crime categories reveals that motor vehicle and bicycle theft consistently rank as the most frequent crime types, averaging 30,000–32,000 cases per year. These are followed by simple assault, personal or retail theft, and vandalism, which collectively account for a substantial share of the total crime volume. Aggravated assault and theft from vehicle form the next major tier, reflecting a stable yet diversified pattern of property and personal crimes. While total incident counts remained relatively stable from 2020–2023, a modest uptick was observed in 2022–2023, coinciding with post-pandemic normalization of urban activity. A slight decline in 2024 may partially reflect data latency or reporting lag rather than an actual reduction in crime rates.

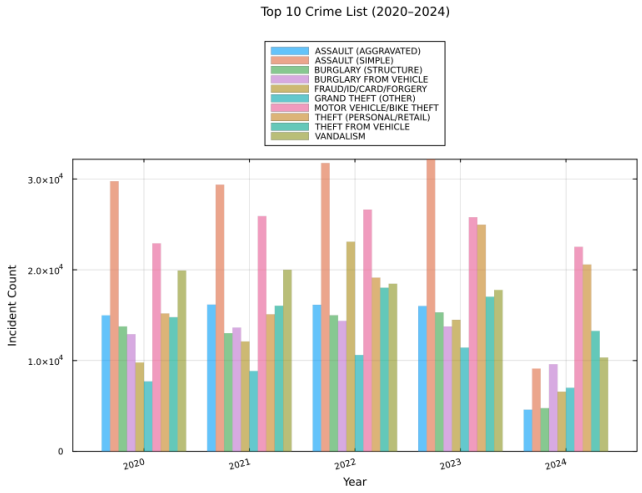


Fig. 1. Top ten generalized crime categories by year (2020–2024).

(ii) Incident Frequency Heatmap:

Incident frequency exhibits a strong diurnal and weekly rhythm, consistent with human activity cycles in an urban environment. As shown in the hourly heatmap, the lowest activity occurs during the early morning hours (03:00–06:00), followed by a sharp increase after 08:00 that persists throughout the day. Evening hours remain active, peaking around typical commuting and social periods, and the highest overall frequencies are observed on Fridays and weekends. This pattern aligns with nightlife, leisure, and mobility trends, highlighting the influence of temporal human behavior on incident dynamics.

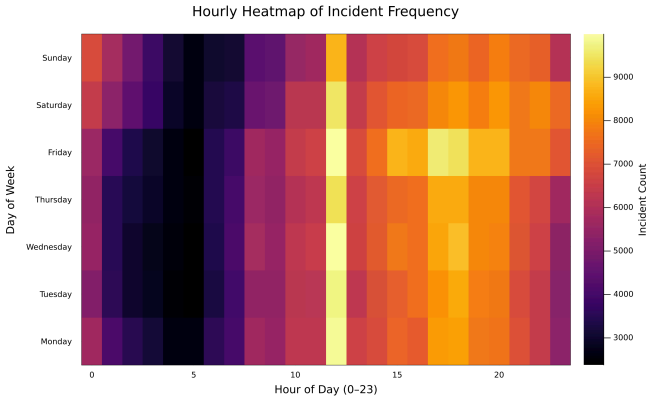


Fig. 2. Hourly heatmap showing diurnal and weekly variations in incident frequency.

(iii) Reporting Delay Distribution:

The analysis of reporting delay is defined as the difference between the date reported and the date of occurrence (Date Rptd − DATE OCC) — reveals a pronounced right-skewed distribution. Most incidents are reported either on the same day or within a few days of occurrence, with a mean delay of 2.96 days and a median delay of 1 day. Approximately 90% of all incidents are reported within five days, indicating generally prompt reporting behavior across most crime categories. The long upper tail in the delay distribution likely reflects crimes with delayed discovery or complex administrative workflows, such as fraud, forgery, or identity-theft–related offenses. This temporal asymmetry underscores the importance of considering both immediate and delayed reporting in operational planning and predictive modeling.

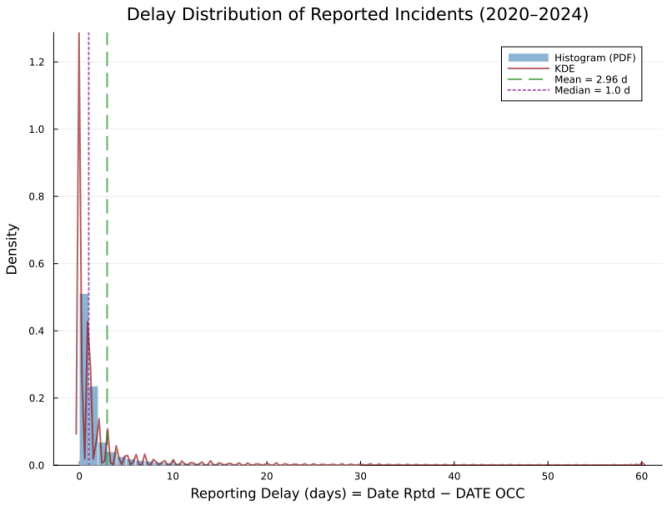


Fig. 3. Distribution of reporting delays (All recorded incidents, 2020–2024).

(iv) Spatial Distribution and Temporal Stability:

Spatially, incident clusters remain highly consistent across the five-year period, concentrating in Downtown, Hollywood, Westlake, and South Los Angeles. These areas exhibit persistent activity regardless of month or year, suggesting enduring socioeconomic and infrastructural factors driving higher incident density. The six representative panels display the months of peak activity for each year between 2020 and 2024. The scatter of points lies almost entirely within the official Los Angeles city boundary, confirming the spatial integrity and proper geocoding of the dataset. Such spatial persistence provides a reliable foundation for hotspot-based predictive modeling and targeted resource deployment.

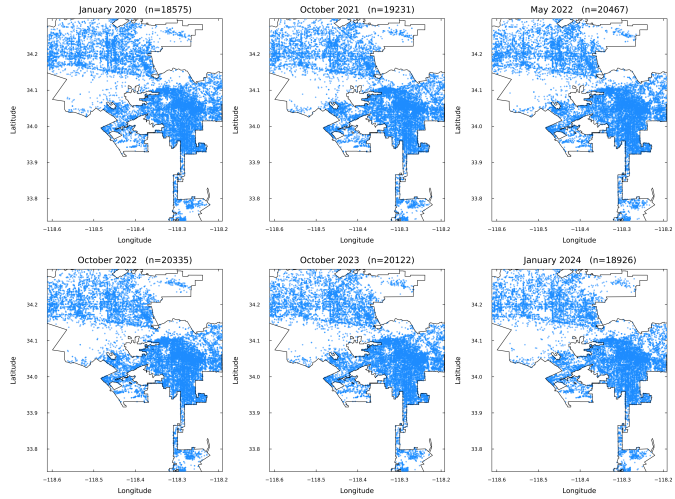


Fig. 4. Annual Peak Months of Incident Distributions (2020–2024)

C. *Spatial Analysis*

Another important aspect of our exploratory data analysis is the spatial distribution of crimes across Los Angeles. By mapping the latitude–longitude of each incident, we visualize hotspots and identify areas with high crime density.

We work with three spatial datasets: the set of crime points $C = \{C_i\}_{\{i=1\}}^{\{N_c\}}$, the set of street-lamp points $L = \{L_l\}_{\{l=1\}}^{\{N_l\}}$, and the city-boundary polygon B. We ensure longitudes and latitudes are numeric, finite, and within plausible ranges so that subsequent geometry remains meaningful.

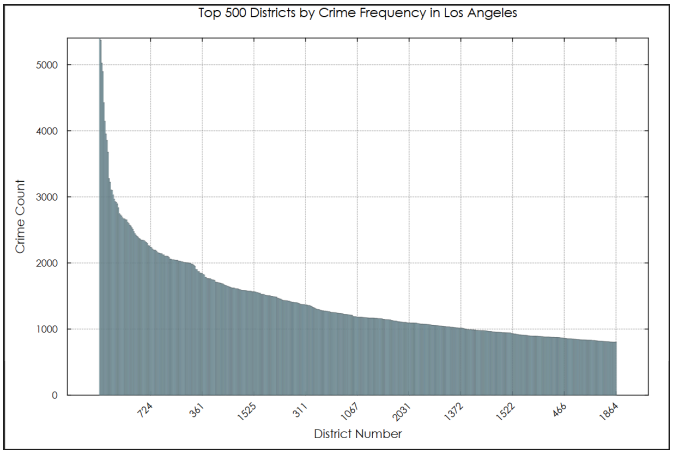


Fig. 5. Top 500 crime count in each designated LAPD district.

Figure Fig. 5 ranks Los Angeles (city) LAPD reporting districts by total reported incidents (2020–present). The x-axis lists districts in descending order (leftmost = highest), and the y-axis shows incident counts. The distribution is heavy-tailed: a few districts concentrate many incidents, followed by a long tail of moderate activity.

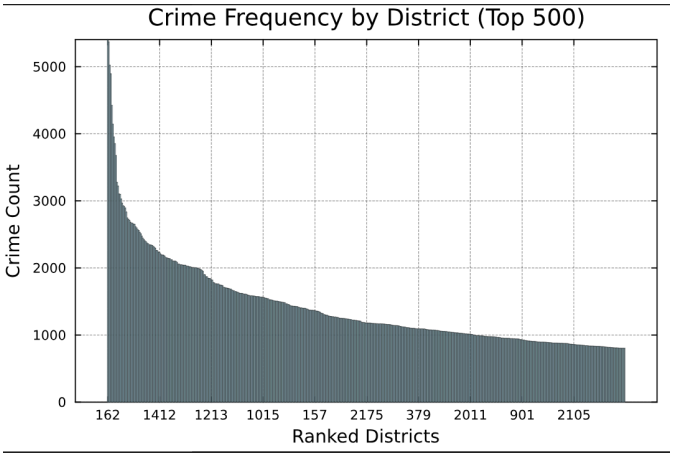


Fig. 6. LA top-500 crime counts (ranked).

Figure Fig. 6 shows the top 500 LAPD reporting districts sorted by total incidents, with the x-axis as rank (left = highest). The y-axis is the incident count. The curve is heavy-tailed: a few districts account for many incidents, followed by a long taper of moderate counts across the remaining ranked districts.

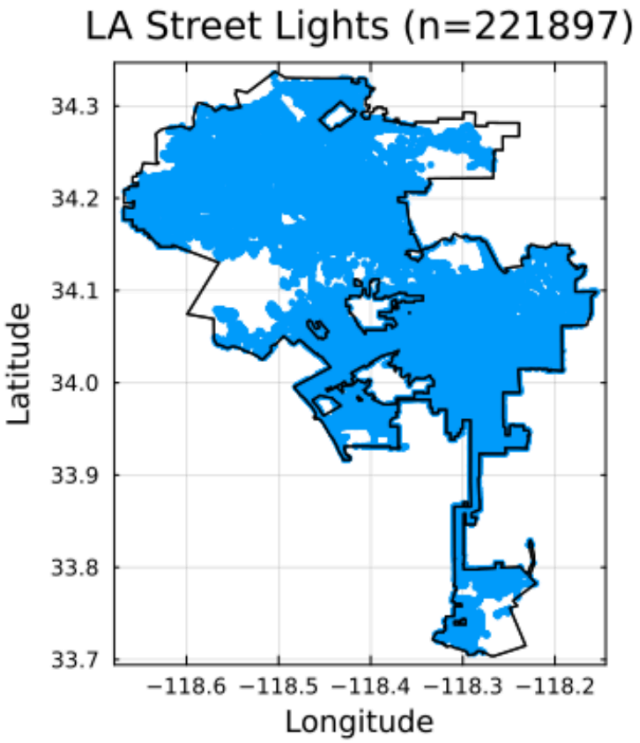


Fig. 7. LA street lights within the city boundary (n=221,897).

This map Fig. 7 plots the reported street-light point locations for Los Angeles in geographic coordinates (longitude/ latitude). The high point density makes many neighborhoods appear as solid filled regions, revealing broad coverage across the city and sparser coverage near edges and open spaces. Axes are in degrees to match the source data.

a) *Finding relationship between crime locations and street lights:*

We project all coordinates into a single metric CRS so distances are measured in meters:

$((x_i, y_i) = \Phi(\lambda_i, \varphi_i),\ (u_l, v_l) = \Phi(\lambda'_l, \varphi'_l))$

where:

- (λ_i, φ_i) and (λ'_l, φ'_l) are geographic (lon/lat, degrees) for crime C_i and lamp L_l,
- Φ denotes the projection to a local metric CRS,
- (x_i, y_i) and (u_l, v_l) are projected (planar) coordinates in meters.

A local projected CRS is used because Euclidean lengths in degrees are not physically meaningful, and a single metric CRS avoids unit mismatches.

We restrict the analysis to the jurisdiction by clipping points to the city polygon B, keeping only crimes and lamps whose projected coordinates fall inside B. This removes out-of-area points and reduces boundary artifacts that would otherwise inflate nearest-distance values.

We define planar Euclidean distance between a crime and a lamp:

$$d(C_i, L_l) = \sqrt{(x_i - u_l)^2 + (y_i - v_l)^2} \qquad (1)$$

For each crime, we keep its nearest-lamp distance: $d_i = m \underset{\{1lelleN_l\}}{\in} d(C_i, L_l), q \quad i = 1, ..., n$, where n is the number of crimes inside B. This yields a single, interpretable proximity value per incident. A spatial index keeps these queries fast.

From $\{d_i\}_{\{i=1\}}^n$ we build the empirical cumulative distribution function (ECDF): $F(r) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{d_i \le r\}, \quad r \ge 0$, so the coverage at an operational radius R is simply: $\text{Coverage}(R) = F(R)$.

We report robust summaries of proximity (median) and tail behavior (p90, p99) along with the maximum distance. Quantiles are preferred over the mean because the distribution is typically right-skewed and outlier-sensitive.

To guard against faulty records, we flag implausible distances above a conservative cap and, if needed, compute robust z-scores using the median and MAD to identify unusual d_i values.

Finally, to see how coverage accumulates with distance, we partition r into analyst-chosen bands (e.g., 0–50–100–250 m, …) and tabulate the share of crimes whose nearest-lamp distance falls into each band. This "coverage by band" view highlights where most gains occur (very small radii) and where diminishing returns set in as the radius grows.

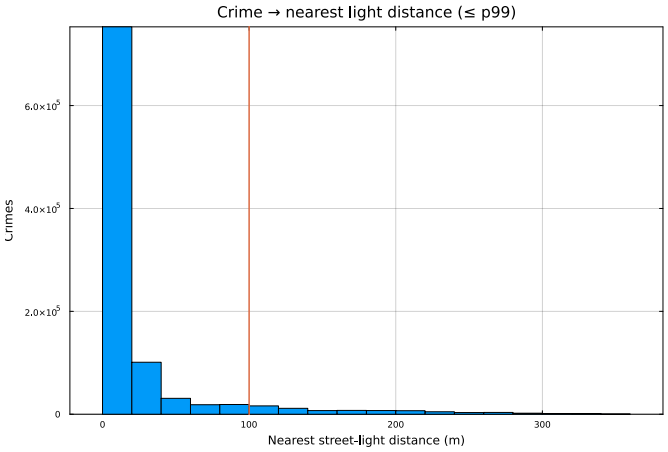


Fig. 8. Crime → nearest light distance (≤ p99), LA.

Figure Fig. 8 shows a histogram of distances (meters) from each crime point to its nearest street light, limited to the 99th percentile to avoid outliers. Most crimes fall very close to a light (left-heavy bar mass), and the frequency declines rapidly with distance. The vertical orange line marks the chosen radius R used later as a working cutoff for proximity.

TABLE I
LA DISTANCE-TO-LIGHT SUMMARY.

Metric	Value
N (valid)	1,004,996
min (m)	0.0852
mean (m)	25,702.8
std (m)	5.43e3
p25 (m)	10.8821
median (m)	14.2868
p75 (m)	20.8421
p90 (m)	81.4454
p95 (m)	152.942
p99 (m)	342.602
max (m)	1.15167e7
≤R (count)	922,403
≤R (%)	91.7822

Table I summarizes the distance from each reported crime in Los Angeles to its nearest street light. It shows total valid records, distribution percentiles (p25–p99), the maximum, and how many / what share fall within the working radius **R = 100 m**. Most crimes are close to a light, with a long right tail driven by a small set of far-out points. More such comparison will be carried out based upon availability of the data.

d. Victim and Incident Attributes

4) Correlations and Relationships

5) Implications for Police Station Planning

Summarize the evidence:

Which areas have high and persistent crime density?

Which times need more coverage (e.g., night hours)?

Support with map + table of "Top 5 areas by crime density and trend".

D. Demographic Anlysis

Besides analyzing crime types and its patterns over time and space, exmining the demographic characteristics of crime victims might provide some useful sights. We analyzed age, sex, and descent compositions of victims. Overall, the victim population is 40.19% male, 35.68% female, and 24.13% unknown or missing. The age distribution of victims is shown in Fig. 9. It shows that the age group of 30-34 has the highest number of victims, followed by the age group of 25-29. The descent distribution of victims is shown in Fig. 10. It shows that the major victim descent groups are Hispanic/Latin/ Mexican, White, and Black, with the percentages of 34.45%, 23.41%, and 15.79%, respectively.

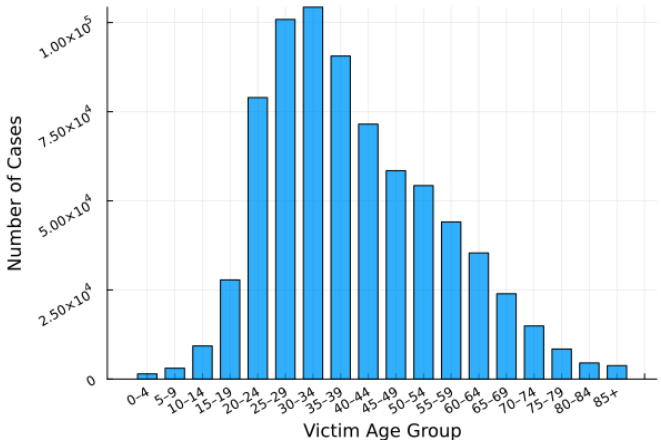


Fig. 9. Age Distribution of Crime Victims

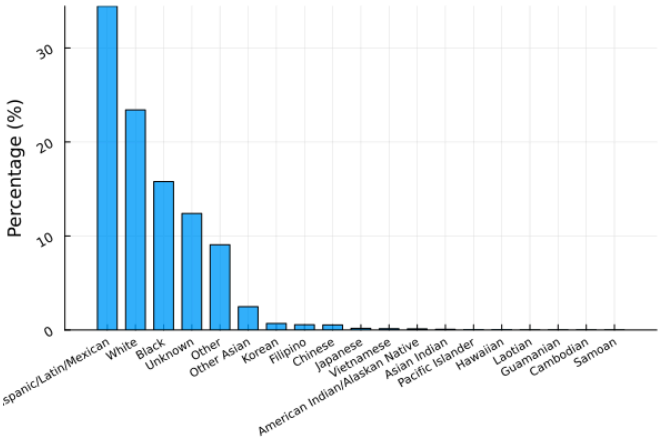


Fig. 10. Descent Distribution of Crime Victims

To see the correlation between age and descent, we created a heatmap shown in Fig. 11. The pattern of age among all descent groups is similar, with the age group of 25-34 having the highest number of victims across all descent groups.

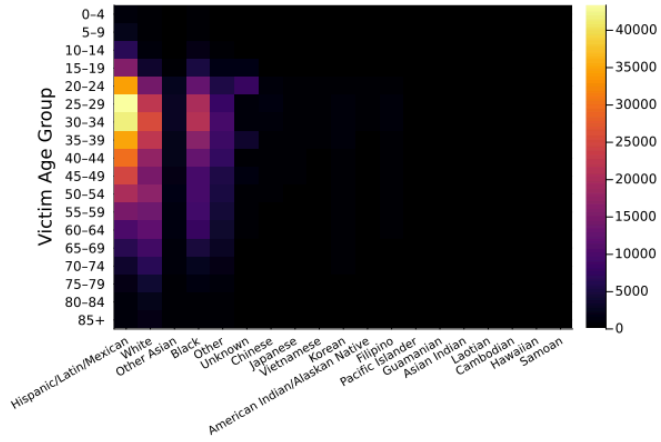


Fig. 11. Heatmap of Victim Age vs. Descent

E. *Summary Statistics of Dataset*

Metric	**Value**
Total records (2020–2024)	≈ 1,004,991 incidents
Average monthly incidents	≈ 20,000–23,000
Distinct crime types	≈ 140
Generalized crime types	37 aggregated categories
Earliest date of Dataset	2020-01-01
Latest date of Dataset	2024-03-31
Mean Hotspot for Crime (lat / lon)	34.05 / −118.32 (≈ Downtown LA)
Mean reporting delay after incident	2.96 days (mean), 1 day (median)
Victim Sex Composition	40.19% male, 35.68% female
Largest Victim Age Group	30–34 years
Top Victim Descent	Hispanic/Latin/Mexican

III. Predictive Modeling

The plan for the modeling will be:

1) Dig into correlations between all features and introduce some new variables if necessary.
2) Come up with a suitable formulation (i.e. decide the structure of the model equation to calculate the number of crimes.)
3) Apply the machine learning scheme we learned in class to train the model and validate it.

References

[1] Los Angeles Police Department / LAPD OpenData, “Crime Data from 2020 to Present.” data.lacity.org / Data.gov, 2025.