What types of crimes are different victim demographics more likely to experience?

shujun (chloe) Yang

View midterm.html on GitHub

Introduction

Crime is a major concern in urban areas, and understanding the patterns of victimization can help improve public safety. In this project, I am analyzing a large dataset of reported crimes from Los Angeles, which includes details about crime type, victim demographics (such as gender, age, and ethnicity), and other relevant factors. The dataset is sourced from Los Angeles Open Data and contains millions of records covering various types of offenses over several years.

The goal of my analysis is to explore how victim demographics influence the type of crime they experience. Specifically, I aim to answer the following questions:

Do different genders experience different types of crimes? Are certain age groups more vulnerable to specific crimes? How does crime type vary across different ethnic groups? Can we predict the most likely crime a person might experience based on their demographics? To answer these questions, I will conduct statistical analyses, visualize trends using heatmaps and bar charts, and ultimately develop a machine learning model to predict crime risks based on victim attributes. The findings could provide insights into crime prevention efforts and help identify high-risk individuals who may need additional safety measures.

Method

Data Source and Preprocessing

The dataset used in this analysis comes from Los Angeles Open Data, specifically the Crime Data from 2020 to Present provided by the Los Angeles Police Department (LAPD). This dataset includes incidents of crime reported in Los Angeles since 2020, covering various crime types along with victim demographics such as age, gender, and ethnicity. The data is sourced via an API request and contains records transcribed from original crime reports

To ensure data quality, I performed extensive preprocessing on the dataset using R (tidyverse, dplyr, lubridate). First, I converted date columns to Date format and standardized time values to extract the hour of occurrence. Categorical variables such as vict_sex and vict_descent were factorized for easier analysis.

Missing values in categorical columns (crm_cd_desc, status_desc, premis_desc, mocodes, vict_sex, vict_descent) were replaced with "Unknown", while numerical columns (vict_age) were imputed with the median value after filtering out invalid values (e.g., negative ages and zeros). I also filled missing values in premis_cd using the most frequent category.

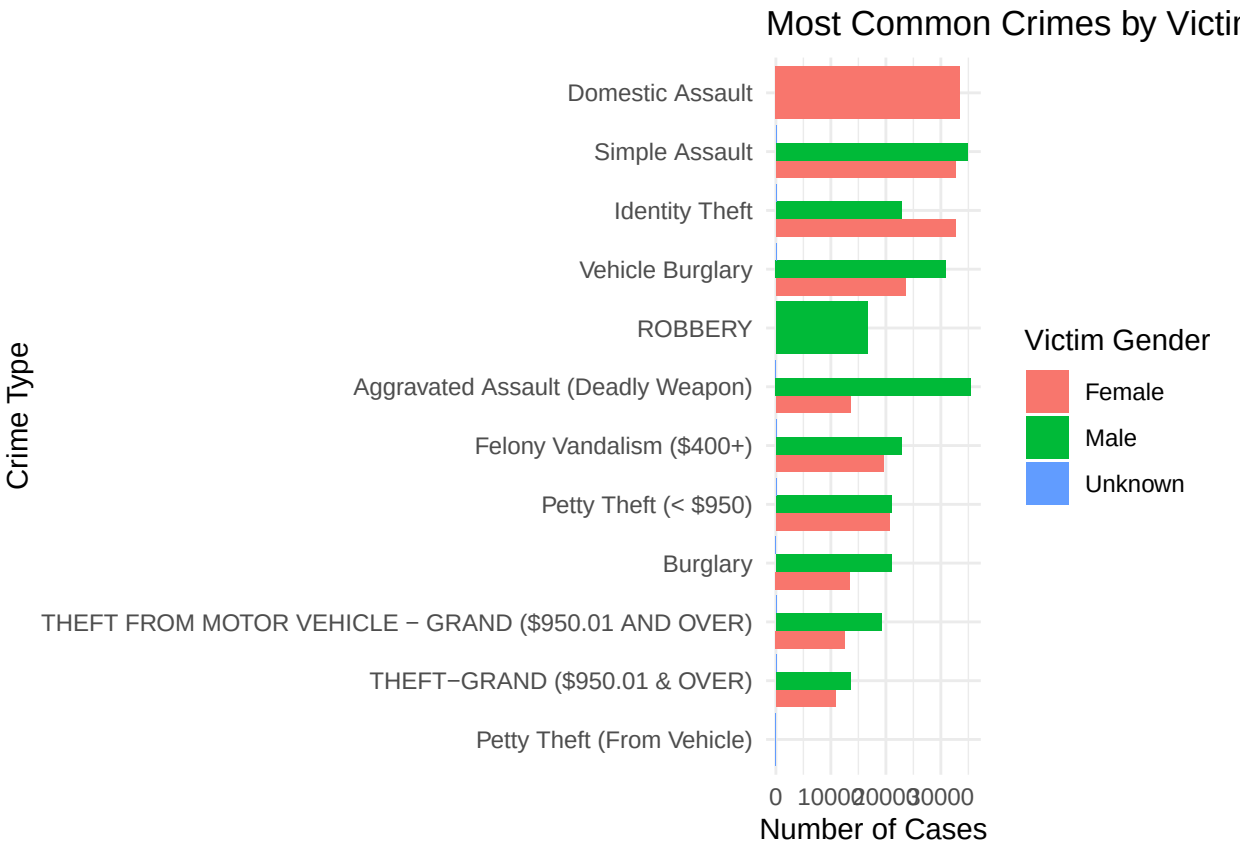
To ensure geographical accuracy, I removed records with invalid latitude and longitude values (e.g., lat=0, lon=0). Additionally, crime descriptions (crm_cd_desc) were standardized and simplified for clarity, and ethnic codes were mapped to full labels (e.g., "B" → "Black", "H" → "Hispanic/Latin/Mexican"). Finally, I dropped high-missing-value columns such as crm_cd_2, weapon_used_cd, and cross_street, and removed records with "Other" or "Unknown" ethnicity to maintain consistency in demographic analysis.

Exploratory Data Analysis (EDA)

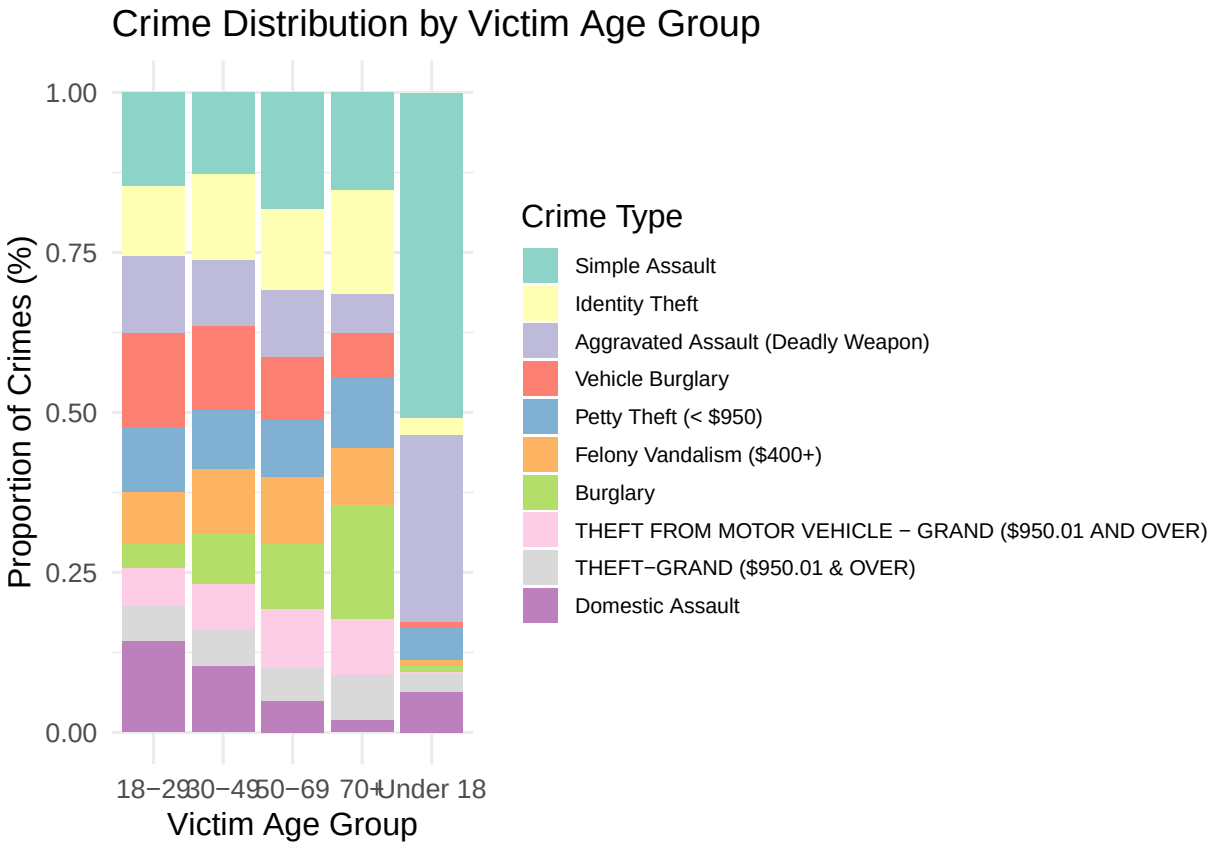
To analyze how victim demographics influence crime type, I generated three key visualizations using ggplot2. The first compares crime types by gender, showing that men are more likely to experience violent crimes (e.g., robbery, aggravated assault), while women are more frequently victims of domestic assault and identity theft. The second visualization explores crime distribution by age group, highlighting that younger victims are more prone to physical crimes, whereas older individuals face a higher risk of financial crimes. Lastly, I examined crime type distribution by ethnicity, revealing that Black and Hispanic victims report more violent crimes, while White and Asian victims experience higher rates of identity theft and property crimes. These patterns confirm that my research question is valid, as clear demographic trends emerge in crime victimization.

Following are the three EDA graphes:

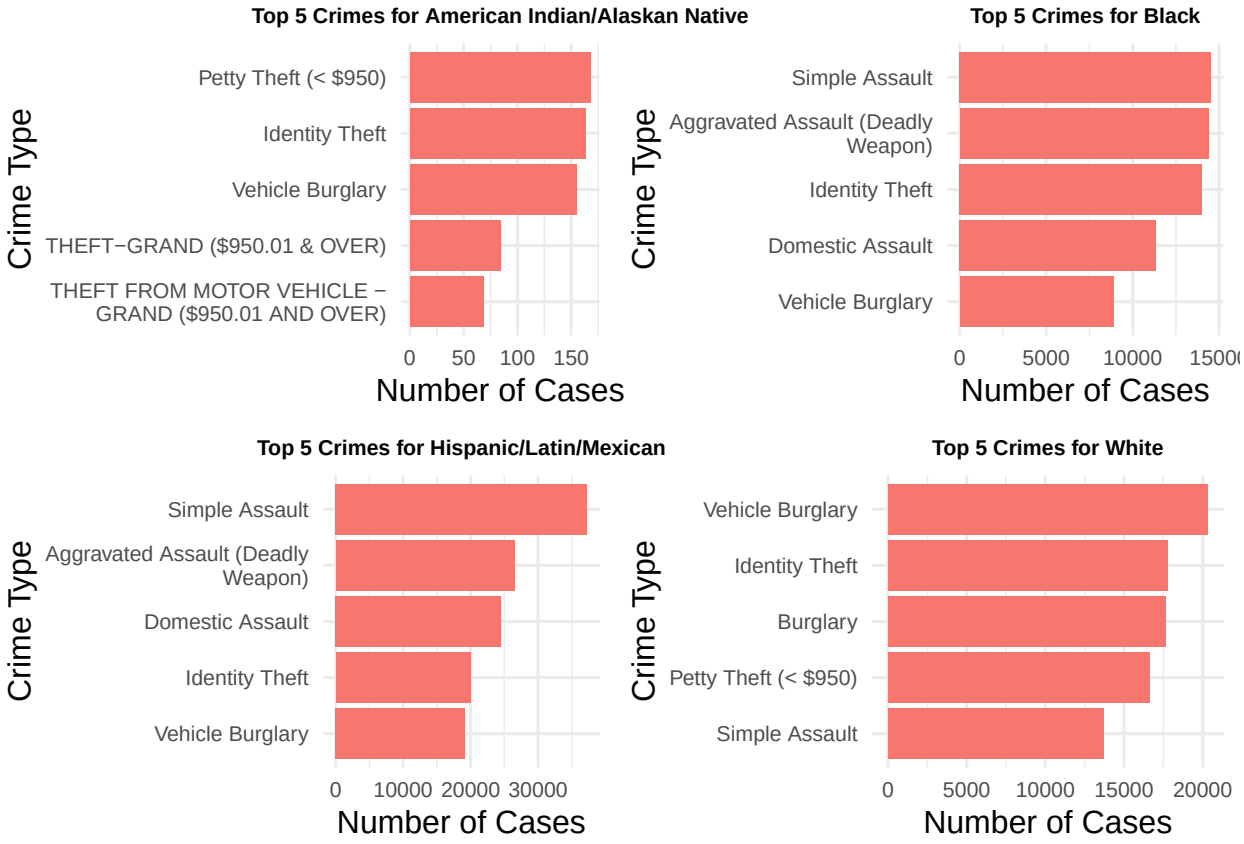
1. Victim Gender vs. Crime Type



2. Victim Age vs. Crime Type



3.Victim Ethnicity vs. Crime Type



After completing the EDA, I confirmed that my hypothesis was reasonable. Next, I conducted chi-square tests for each of the three key questions to statistically verify that gender, age, and ethnicity significantly influence the types of crimes victims are most likely to experience.

Chi-Square Test and Analysis

1. Do men and women experience different types of crimes?

Method

To answer this, crimes were grouped into seven types:

- Assault (e.g., domestic assault, aggravated assault)
- Fraud (e.g., identity theft, credit card fraud)
- Property Damage (e.g., vandalism, arson)
- Public Order (e.g., stalking, resisting arrest)
- Sexual Crimes (e.g., rape, lewd conduct)
- Theft (e.g., burglary, vehicle theft)
- Violent Crimes (e.g., robbery, homicide)

A table was created to compare victim gender across crime types. A Chi-Square Test was used to check if gender and crime type are related.

Analysis

Crime Type Distribution by Victim Gender

Victim Gender	Crime Type	Count
Female	Assault	105858
Male	Assault	97038
Unknown	Assault	107
Female	Fraud	34365
Male	Fraud	25750
Unknown	Fraud	119
Female	Property_Damage	30807
Male	Property_Damage	34781
Unknown	Property_Damage	79
Female	Public_Order	30267
Male	Public_Order	16945
Unknown	Public_Order	34
Female	Sexual_Crimes	8970
Male	Sexual_Crimes	1336
Unknown	Sexual_Crimes	5
Female	Theft	103346
Male	Theft	137514
Unknown	Theft	487
Female	Violent_Crimes	13804
Male	Violent_Crimes	29661
Unknown	Violent_Crimes	23

Chi-Square Test Results for Gender and Crime Type

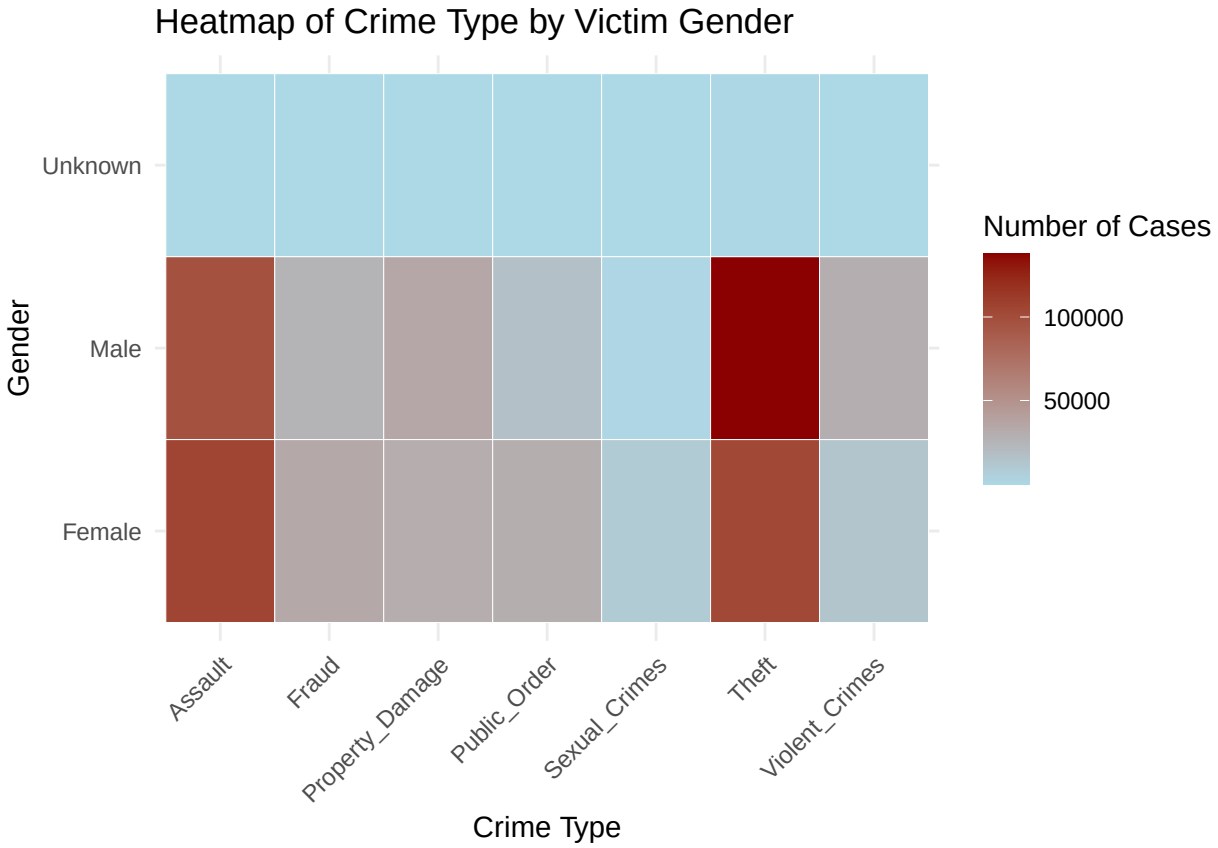
Chi-Square	DF	P-Value
21,814.26	12	$<$ 2.22e-16

The p-value is very small ($p < 0.001$), confirming a strong connection between gender and crime type.

Women are more often victims of assault, fraud, and sexual crimes. This aligns with EDA findings, where domestic assault and identity theft were more common for female victims.

Men are more often victims of theft, violent crimes, and property damage. This matches EDA results, which showed more burglary, robbery, and vandalism cases involving male victims.

Public order crimes are more frequent among female victims, supporting EDA observations on stalking and harassment. This also aligns with the trend we observed in the heatmap.



2. Do different age groups experience different types of crimes?

Method

To answer this, crimes were grouped into seven types as before:

- Assault (e.g., domestic assault, aggravated assault)
- Fraud (e.g., identity theft, credit card fraud)
- Property Damage (e.g., vandalism, arson)
- Public Order (e.g., stalking, resisting arrest)
- Sexual Crimes (e.g., rape, lewd conduct)
- Theft (e.g., burglary, vehicle theft)
- Violent Crimes (e.g., robbery, homicide)

A table was created to compare crime types across different age groups. A Chi-Square Test was used to check if age group and crime type are related.

Analysis

The p-value is very small ($p < 0.001$), confirming a strong connection between age group and crime type.

Individuals aged 30-49 experience the highest number of crimes overall, especially theft and assault. This aligns with EDA findings, where this age group had the most reported incidents.

Younger individuals (under 18) have a higher proportion of sexual crimes, which matches EDA observations.

Crime Type Distribution by Victim Age Group

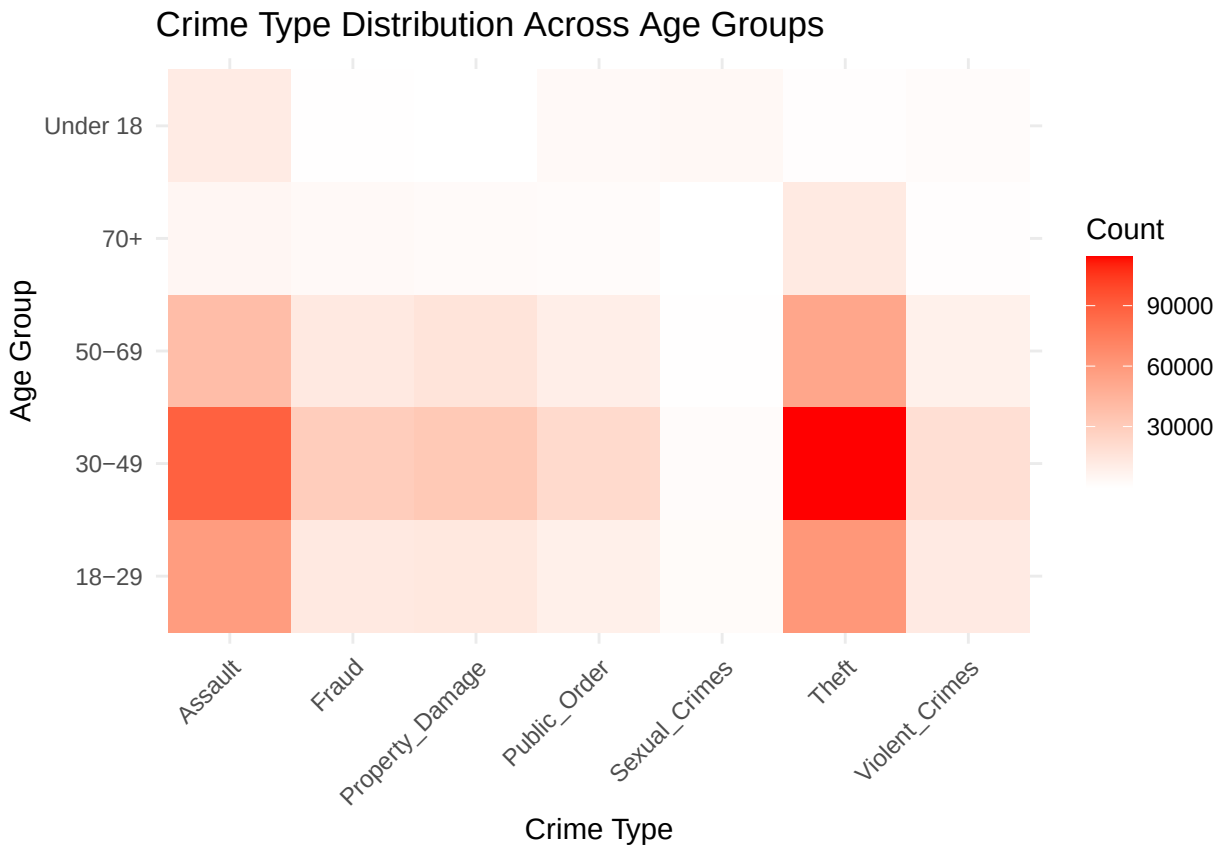
Age Group	Crime Type	Count
18-29	Assault	58083
30-49	Assault	88667
50-69	Assault	39036
70+	Assault	5190
Under 18	Assault	12027
18-29	Fraud	13333
30-49	Fraud	29606
50-69	Fraud	13263
70+	Fraud	3595
Under 18	Fraud	437
18-29	Property_Damage	13974
30-49	Property_Damage	32404
50-69	Property_Damage	16157
70+	Property_Damage	2888
Under 18	Property_Damage	244
18-29	Public_Order	9214
30-49	Public_Order	22321
50-69	Public_Order	9959
70+	Public_Order	2194
Under 18	Public_Order	3558
18-29	Sexual_Crimes	2674
30-49	Sexual_Crimes	2444
50-69	Sexual_Crimes	651
70+	Sexual_Crimes	65
Under 18	Sexual_Crimes	4477
18-29	Theft	60929
30-49	Theft	114021
50-69	Theft	52542
70+	Theft	12729
Under 18	Theft	1126
18-29	Violent_Crimes	12480
30-49	Violent_Crimes	19023
50-69	Violent_Crimes	8616
70+	Violent_Crimes	1096
Under 18	Violent_Crimes	2273

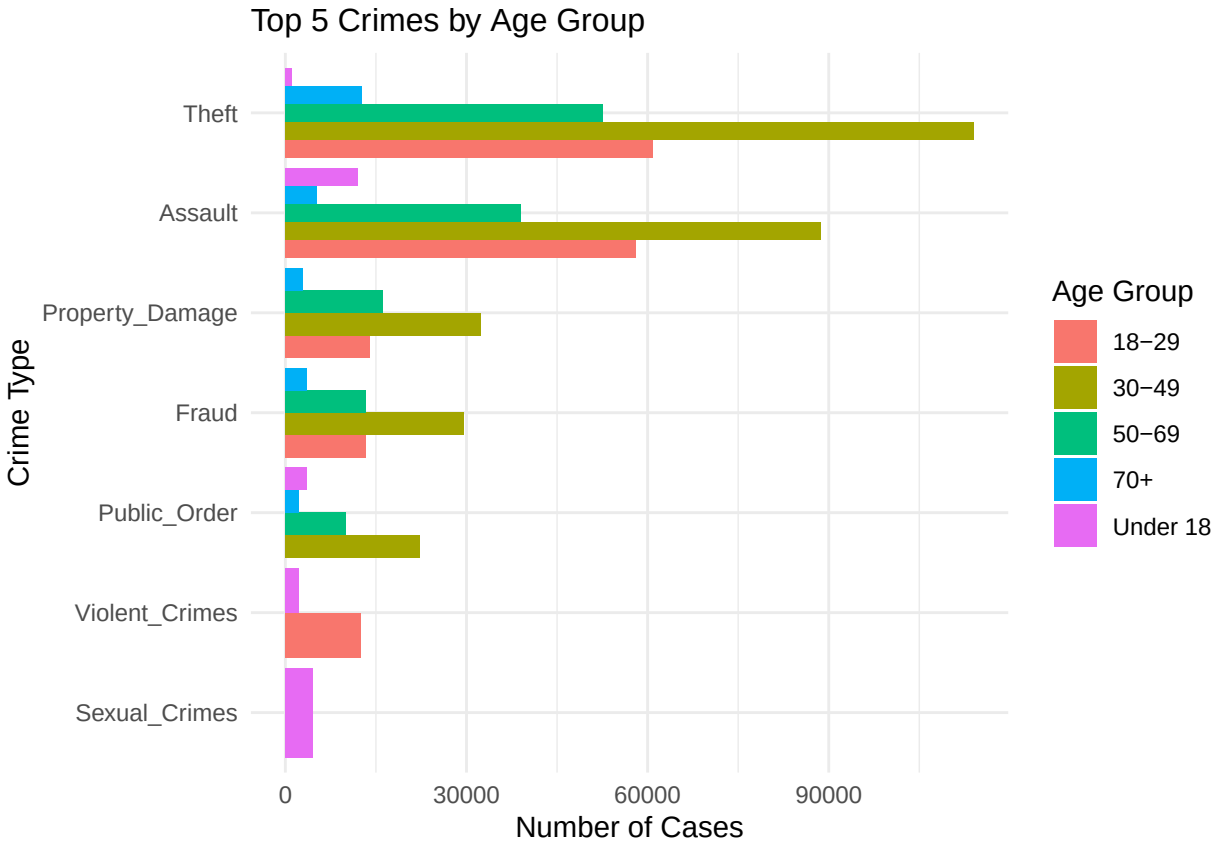
Older individuals (70+) experience fewer crimes overall, but fraud and property damage are relatively more common among them.

Public order crimes are most frequent in the 30-49 age group, supporting EDA findings on offenses like stalking and resisting arrest. The overall trend suggests that crime type varies significantly by age group.

Chi-Square Test Results for Age Group and Crime Type

Chi-Square	DF	P-Value
69,869.05	24	$< 2.22e\text{-}16$





3. Do different ethnic groups experience different types of crimes? Method

To answer this, crimes were grouped into seven types as before:

- Assault (e.g., domestic assault, aggravated assault)
- Fraud (e.g., identity theft, credit card fraud)
- Property Damage (e.g., vandalism, arson)
- Public Order (e.g., stalking, resisting arrest)
- Sexual Crimes (e.g., rape, lewd conduct)
- Theft (e.g., burglary, vehicle theft)
- Violent Crimes (e.g., robbery, homicide)

A table was created to compare victim ethnicity across crime types. A Chi-Square Test was used to check if ethnicity and crime type are related.

Analysis

The p-value is very small ($p < 0.001$), confirming a strong connection between ethnicity and crime type.

Hispanic/Latin/Mexican and Black victims experience the highest number of reported crimes, particularly assault and theft. White victims also show a high occurrence of theft, aligning with previous EDA findings.

Asian and Pacific Islander groups, including Chinese, Japanese, and Filipino victims, have relatively fewer recorded crimes. Fraud and public order offenses are more prominent among these groups compared to other crime types.

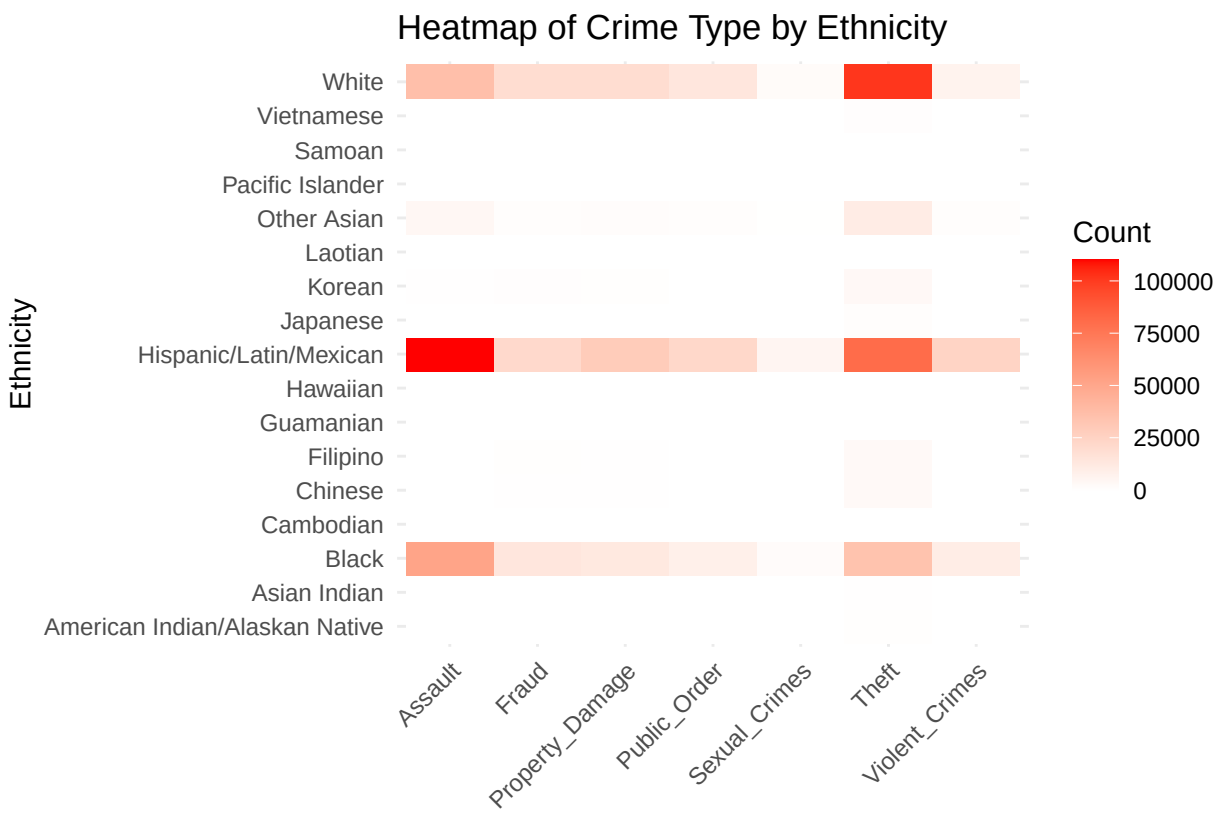
The heatmap reveals that theft and assault are the most reported crimes across multiple ethnic groups, reinforcing the trends observed in the table. Ethnic minorities generally report fewer violent crimes than Hispanic and Black victims.

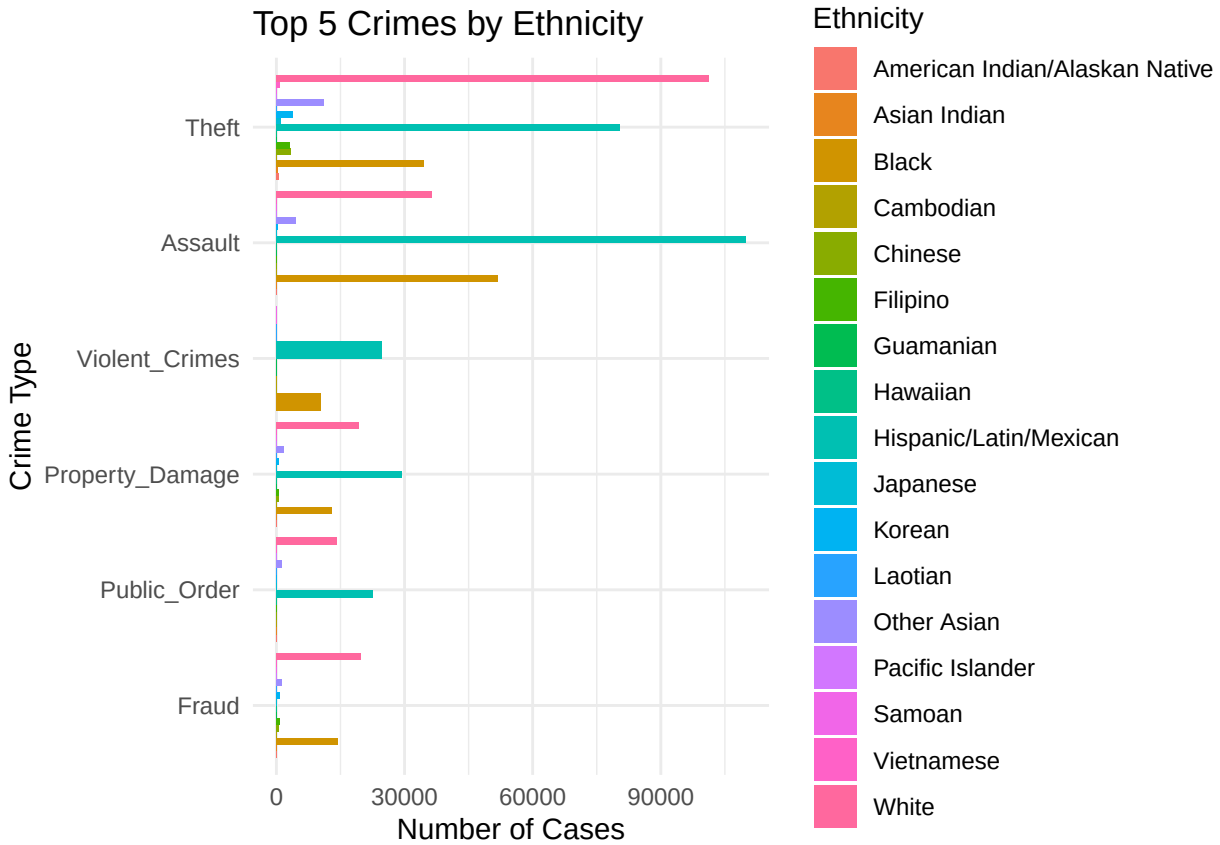
Table 1: Observed Counts: Crime Type by Ethnicity

	Assault	Fraud	Property_Damage	Public_Order	Sexual_Crimes	Theft	Violent_Crimes
American Indian/Alaskan Native	53	169	82	34	3	645	20
Asian Indian	6	75	48	14	0	429	1
Black	51798	14495	12903	8754	2025	34601	10447
Cambodian	1	14	11	1	0	62	2
Chinese	19	545	490	66	4	3456	7
Filipino	102	727	600	101	9	3232	18
Guamanian	17	7	8	2	2	33	4
Hawaiian	9	33	54	4	1	117	1
Hispanic/Latin/Mexican	109843	21759	29356	22645	5653	80400	24792
Japanese	16	188	175	19	1	1166	4
Korean	290	875	679	147	14	3843	81
Laotian	0	20	20	1	0	34	1
Other Asian	4449	1378	1790	1219	236	11028	1137
Pacific Islander	9	59	25	9	3	176	3
Samoan	3	20	4	2	0	26	3
Vietnamese	11	149	105	18	3	888	2
White	36377	19721	19317	14210	2357	101211	6965

Table 2: Chi-Square Test Results for Ethnicity and Crime Type

	Chi_Square_Statistic	Degrees_of_Freedom	P_Value
X-squared	63339.88	96	0.00e+00


Heatmap of Crime Type by Ethnicity
White
Vietnamese
Samoan
Pacific Islander
Other Asian
Laotian
Korean
Japanese
Hispanic/Latin/Mexican
Hawaiian
Guamanian
Filipino
Chinese
Cambodian
Black
Asian Indian
American Indian/Alaskan Native
Ethnicity
Assault
Fraud
Property_Damage
Public_Order
Sexual_Crimes
Theft
Violent_Crimes
Crime Type
Count
100000
75000
50000
25000
0


Top 5 Crimes by Ethnicity
Theft
Assault
Violent_Crimes
Property_Damage
Public_Order
Fraud
Crime Type
0
30000
60000
90000
Number of Cases
Ethnicity
American Indian/Alaskan Native
Asian Indian
Black
Cambodian
Chinese
Filipino
Guamanian
Hawaiian
Hispanic/Latin/Mexican
Japanese
Korean
Laotian
Other Asian
Pacific Islander
Samoan
Vietnamese
White

Prediction Modeling

To predict the most likely type of crime a victim might experience, I built and compared several classification models using victim demographics and contextual features such as age, gender, ethnicity, time of occurrence, area, and type of location. The target variable was the categorized crime type. The dataset was split into training (70%), validation (15%), and testing (15%) subsets.

I began with a multinomial logistic regression model using hard-label prediction as a baseline. While it achieved a reasonable Top-1 accuracy of 46% on the test set, its performance across individual crime categories—especially those with fewer samples—was poor.

To address class imbalance, I implemented a weighted random forest model. Although the Top-1 accuracy dropped to 33%, the class-wise metrics improved notably, especially for underrepresented crime types such as fraud and sexual crimes. This tradeoff highlighted the model's improved fairness in minority class detection, which is critical for real-world risk warnings.

Building on this, I transitioned from hard prediction to soft prediction using the probability outputs of the weighted random forest. This allowed for Top-N accuracy evaluation. I also trained an XGBoost model for comparison. Both models showed similar performance in Top-N accuracy, with Top-3 reaching over 85% and Top-5 nearing 97%. Given these results, I propose using a Top-3 prediction threshold for generating alerts in a crime risk prediction system. This approach balances predictive accuracy with actionable insight, allowing authorities or individuals to be warned about the most likely threats based on demographic profiles.

Table 3: Table 1. Test Accuracy: Multinomial Logistic Regression Model

Crime Category	Accuracy (%)
Overall (Validation)	46.53
Overall (Test)	46.89
Class: Assault	66.24
Class: Fraud	11.46
Class: Property_Damage	32.96
Class: Public_Order	0.10
Class: Sexual_Crimes	0.06
Class: Theft	62.29
Class: Violent_Crimes	0.00

Table 4: Table 2. Test Accuracy: Weighted Random Forest Model

Crime Category	Accuracy (%)
Overall (Validation)	33.54
Overall (Test)	33.65
Class: Assault	13.86
Class: Fraud	67.34
Class: Property_Damage	36.45
Class: Public_Order	27.96
Class: Sexual_Crimes	53.02
Class: Theft	37.98
Class: Violent_Crimes	53.49

Table 5: Table 3. Top-N Accuracy (Validation & Test Sets) using Weighted Random Forest

Top-N	Validation Accuracy (%)	Test Accuracy (%)
Top-1	49.11	49.50
Top-2	73.28	73.62
Top-3	85.61	85.84
Top-4	92.93	93.02
Top-5	97.31	97.26

Table 6: Table 4. Top-N Accuracy (Validation & Test Sets) using XGBoost Classifier

Top-N	Validation Accuracy (%)	Test Accuracy (%)
Top-1	49.01	49.07
Top-2	73.27	73.58
Top-3	85.78	85.92
Top-4	93.17	93.12
Top-5	97.38	97.34

Conclusion and Summary

This study explores the relationship between victim demographics and crime types in Los Angeles using a large-scale dataset. Through exploratory data analysis and statistical testing, we found clear patterns: men are more likely to experience violent crimes and theft, while women are more frequently victims of fraud, sexual offenses, and domestic assault. Age also plays a significant role—young adults face higher risk of physical and sexual violence, whereas older individuals are more likely to experience fraud and property crimes. Ethnic disparities are also evident: Hispanic/Latin and Black victims are disproportionately represented in assault and theft categories, while White and Asian victims more often report fraud and property crimes.

To build a predictive system, we trained and compared several classification models. While the multinomial logistic regression achieved decent overall accuracy, it failed to perform well on less frequent crime types. A weighted random forest model improved class-wise sensitivity, particularly for rare classes. To maximize practical use, we adopted soft prediction strategies and evaluated Top-N accuracy. Both weighted Random Forest and XGBoost showed strong performance, with Top-3 accuracy exceeding 85%.

Rather than relying on strict Top-1 predictions that may miss important risks, we propose a Top-3 alert strategy. This threshold balances precision and breadth—ensuring that likely threats are captured without overwhelming users with too many warnings. Such a strategy is practical for real-world early warning systems that support public safety and proactive interventions.

This predictive framework has potential applications in public safety, community outreach, and urban policy design. However, several limitations remain. The model does not account for dynamic factors such as recent crime spikes, neighborhood policing, or personal behavior. Additionally, prediction accuracy is constrained by data imbalance and the quality of original police reports. Future work could incorporate spatial-temporal modeling, neural networks, or external socioeconomic data to improve performance and interpretability.