Analyzing Crimes Committed Against Women in San Diego County

Alisa Crowe, Rita Herfi, and Samantha Keppler

Big Data Analytics, San Diego State University

BDA 594: Big Data Science and Analytics Platforms

Prof. Ming-Hsiang Tsou

Dec 16, 2024

Abstract

This data analysis project explores the patterns and trends of reported crimes against women in San Diego County, leveraging data from the San Diego Association of Governments (SANDAG) website. The study aims to identify the most frequently committed types of crimes, analyze their temporal and spatial distributions, and assess the influence of socio-economic factors on crime rates. Through comprehensive data cleaning, visualization, predictive modeling, and analysis, the project highlights key findings, including the roles of race, age, city, and type of incident.

***Keywords*: Reported Crimes, Victim Demographics, Domestic Violence, Data Analysis, Predictive Modeling**

1. Introduction

The issue of crimes against women in San Diego County is a significant concern, affecting the safety, well-being, and quality of life of thousands of women in the region. Despite various efforts to address this problem, the prevalence of such crimes remains alarmingly high. There is a critical need for a detailed and comprehensive analysis of these crimes to understand their patterns, underlying causes, and the effectiveness of current preventative measures.

Overall crime rates in the United States have been dropping in recent years, especially since the institution of stay-at-home mandates during the COVID-19 pandemic (Jackman, 2020). However, evidence suggests that these trends vary dramatically based on area of interest, indicating that nationwide or global trends may not be applicable to San Diego County specifically (Boman & Gallupe, 2020). According to the City of San Diego Official Website, the city and the surrounding area, is currently, “one of the most ethnically and culturally diverse places in the nation”, due to the proximity to the Mexican border among other socioeconomic factors. For the above reasons, it seems necessary to conduct studies on crime trends in San Diego County specifically,

to understand how crime rates affect this melting pot of a metropolitan area.

Several research studies have already investigated crime trends in San Diego in relation to various factors. For example, a 2023 analysis done by Yuyang Han, a scholar at the Scripps Institution of Oceanography at UC San Diego, discovered higher crime rates in the southern coastal region of San Diego, which is also more prone to heat waves. However, another study done by Caetano et al. (2021) found that the rate of violent crimes in San Diego County was significantly lower than the overall state rate in 2017. Furthermore, most existing literature dates back to the early 2000s, before modern data analytics techniques were developed. Overall, there seems to be a lack of consensus throughout studies in the area, and a lack of recent research projects. Beyond that, there is little to no research specifically focused on female victimization.

Why is it important to investigate crimes against women specifically? Multiple studies have suggested that the recent decrease in crime in the US can be attributed to a shift in the nature of crimes. While less severe crimes committed by groups have become less common, more violent crimes committed by individuals have actually increased. Consistent with this, since the start of the pandemic, the US has seen a disturbing increase in domestic violence related incidents (Amber et al., 2020). This is a strong reason to look into crimes against women, as women are more likely to be the victims in domestic violence cases. Although sources disagree on the exact rate of domestic violence against women versus men in the US, they all agree that women experience domestic violence at a higher rate than men do.

Several questions remain: How are crimes against women in San Diego changing over time? Are there any factors that can be used to predict crimes against women, and develop strategies for preventing it?

1.1. Project Objectives

The purpose of this analysis is to comprehensively investigate the incidence and characteristics of crimes against women in San Diego County. By employing advanced data analytics techniques, this study seeks to provide actionable insights that can inform policy decisions, enhance law enforcement strategies,

and support the development of more effective intervention programs. The insights gained from this analysis will be crucial in developing targeted approaches to reduce the occurrence of these crimes and enhance the safety and well-being of women in the region. By addressing this critical issue through meticulous analysis and evidence-based recommendations, the aim is to contribute to the ongoing efforts to combat crimes against women and create a community where all individuals can thrive without fear of violence or discrimination.

2. Data Collection and Cleaning

2.1. Data Collection

For this project, a public dataset was analyzed from the SANDAG Open Data Portal that was collated by the Automated Regional Justice Information System (ARJIS). The dataset, consisting of self-reported crimes from victims to law enforcement agencies in San Diego County, is called California Incident Based Reporting System (CIBRS) Group A Public Crime Data. It had 511K rows, each containing one unique offense, and 21 different fields. The dataset does not only report the highest charge assigned to each incident; each individual incident can have up to 10 offenses assigned to it if multiple offenses were committed. This analysis includes data from January of 2021 to September of 2024 (the time at which the data was pulled).

It is necessary to mention that the SANDAG's portal also had a different dataset of similar nature called CIBRS Group B Public Crime Data. SANDAG differentiates between the two datasets as follows: "Group A offenses include more serious crime categories, such as Rape, Robbery, and Homicide and Group B arrests include less serious crime categories such as Loitering, Driving Under the Influence (DUI), and Liquor Law Violations." Group A was the dataset chosen, because these crimes were more likely to have victims (crimes against Persons) whereas Group B offenses were more likely to be crimes against Property or Society.

Data collected from the US Census Bureau data hub were also used to weigh results based on the observable demographic qualities of San Diego County. Age and Sex data in the area was gathered (Table ID: S0101) as well as Hispanic or Latino Origin by Race (Table ID: B03002).

2.2. Data Cleaning Procedures

In the data cleaning process, there were several measures taken to ensure the dataset was suitable for analysis and predictive model training. First, since it was irrelevant to the research topic, the field called Stolen Vehicles, was removed. Next, it was necessary to filter out certain attributes that were irrelevant. In the Crimes Against field, rows that fell under the category of Property and Society were deleted, leaving only Crimes Against People. Then, in the Victim Sex field, rows with male victims were removed, leaving only female victims. The fields Crimes Against and Victim Sex were then deleted, as they were now homogeneous in nature.

2.2.1. Victim Race Field

To clean the Victim Race field, several attributes were grouped together to make larger racial groups. These racial groups were meant to closely match the subtypes of race that are defined by the U.S. Census Bureau. The following groupings were made:

Figure 1

New Aggregate Groupings in Comparison to Original Dataset Categories

New Aggregate Group	Original Categories
American Indian	American Indian
Asian	Asian Indian, Cambodian, Chinese, Filipino, Indian, Japanese, Korean, Laotian, Other Asian, Vietnamese
Black	East African, Black
Hispanic	Hispanic
Other	Other
Pacific Islander	Guamanian, Hawaiian, Pacific Islander, Samoan
Unknown	N/A
White	Middle Eastern, White

Note: Currently, the federal government officially categorizes people with Middle Eastern and North African descent as White (Wang, 2022) so the Census Bureau does the same. This is set to change in the future, so future analyses may consider creating a separate group for those of Middle Eastern and North African descent.

2.2.2. Victim Age Field

For the Victim Age column, the 7,623 rows where age was under 18 were dropped since the analysis is focused on crimes committed against adult women. The remaining null values were filled with the median value for each race group. The median was chosen over

the mean because the age data is right-skewed, making the median a better representation of the typical age seen in the dataset. Additionally, the Incident Date column was converted to a datetime data type for ease of analyzing time-related metrics.

2.2.3. Offense Description Field

Upon seeing the distribution of offense descriptions, all categories with less than 1% frequency were combined into an aggregate column *Other*. These columns were as follows: Sexual Assault with an Object, Murder, Nonnegligent Manslaughter, Negligent Manslaughter, Statutory Rape, Forcible Sodomy, Human Trafficking/Involuntary Servitude, and Incest.

2.2.4. City Field

The City column initially contained 2,402 null values, as well as 3,804 rows that contained the invalid value of *Sheriff*. The main goal for cleaning the City column was to accurately assign one city name to each incident, thereby improving the data's completeness and reliability for later geographical analysis. This process involved removing all values which were not valid city names within San Diego County and leveraging other aspects of the data to fill in missing fields. The Zip Code column, which only had 202 null values, was used to fill in missing data from the City column, reducing the number of null values from over 6,000 to 175. For other missing values of City, the Agency, Beat, and BCS (Beat Codes System) Area columns were used to infer the correct city names. For example, if a record showed no City name but had an Agency of *Chula Vista*, it was inferred that the city the incident occurred in was Chula Vista. Similar strategies were used for the Police Beat and Sheriff BCS Area data. Instances where City data was available but Zip Code was null were handled by filling the latter column with the most frequent zip code used for the specific city in the dataset. Lastly, unincorporated areas were mapped to their corresponding city names to reduce data redundancy; for example, all instances of *El Cajon Uninc* were combined with *El Cajon*. There were 41 rows of data where the City field could not be inferred; these rows were dropped from the dataset to ensure consistency.

2.2.5. HHSA Region Field

Once the City column was completed, a new column was created titled HHSA Region to provide a simpler framework for categorizing geographic data. The San Diego County's Health and Human Services Agency's 2024 assessment categorized the cities into six regions: Central, East, North Central, North Coastal, North Inland, and South. Grouping the data into these broader regions allows for easier interpretation and higher accuracy, as regional classifications are more reliable than the estimates in many of the City and Zip Code fields. Lastly, the 25 rows where City, Zip Code, and HHSA Region were null and could not be inferred were dropped from the dataset.

2.3. Data Limitations

There are a few notable limitations due to the nature of the dataset. Firstly, this dataset only spans three years of recent history in San Diego County, so long-term trends may be difficult to extrapolate from this analysis. Additionally, it is important to note that many incidents of crime, especially domestic violence related incidents, go unreported. This research only contains analysis on crimes that were documented by the various law enforcement agencies in San Diego County. Also noted, some issues with the victim categorization that was done by SANDAG, had some outdated grouping used. This included grouping Middle Eastern individuals in with White individuals, and defining victims by sex rather than by gender.

Despite the thorough data cleaning efforts, there were still several biases that remained in the dataset. Many city names and zip codes were estimated due to incomplete data, introducing some level of uncertainty in the geographic analysis. Values that could not be estimated with confidence were omitted from the dataset. Next, a significant number of ages (7,623) were filled in using the median values grouped by race. This method results in a much higher concentration of ages around the median values, specifically 30-35 years old, and may not accurately reflect the true distribution of ages. These limitations show the challenges of working with incomplete and imperfect data. While the data cleaning process significantly improved the dataset's quality, these inherent limitations should be acknowledged and considered when interpreting the results of any

analysis or predictive modeling conducted using this data.

3. Methodology

Data cleaning and preparation were conducted using Jupyter Notebook, chosen for its interactive environment to execute Python scripts and visualize results. The primary tool for data manipulation was the Pandas library, which allowed data cleaning procedures by filtering, merging, and mapping. Additionally, several Python dictionaries were constructed to map values such as zip codes, unincorporated areas, and sheriff-designated BCS Areas to corresponding city names. Throughout the data cleaning process, intermediate datasets were saved at various stages for traceability to ensure that certain data cleaning decisions could be revisited as needed.

Initially, basic visualizations were created using R Studio to begin the analysis. After extracting trends from the data, higher-level visualizations, including maps, were developed using Tableau.

When using location for Tableau some areas were not incorporated so they had to be manually entered in order to be identified. Then, two word clouds were plotted with the cities of where the incidents occurred. These had to have underscores because R would read two worded cities incorrectly as separate entities, but going through similar functions from the Web Exercise 2 in the BDA 594 class, they were generated correctly.

The predictive modeling pipeline was developed using the Python library Scikit-learn for machine learning tasks like model training and evaluation. The process involved data preprocessing, feature extraction, and model training. Key features were experimented with in order to find the most optimal set of input features for each model to maximize predictive performance. A web application was then created using Flask to serve as an API interface, processing user inputs and communicating with the trained models to provide real-time predictions. To make data input more intuitive for the user, an HTML file detailing the specifications of the interactive form was created and embedded into the Google Sites page. The complete system was deployed using Heroku, which allowed us to create a

user-friendly application that runs predictions in real-time.

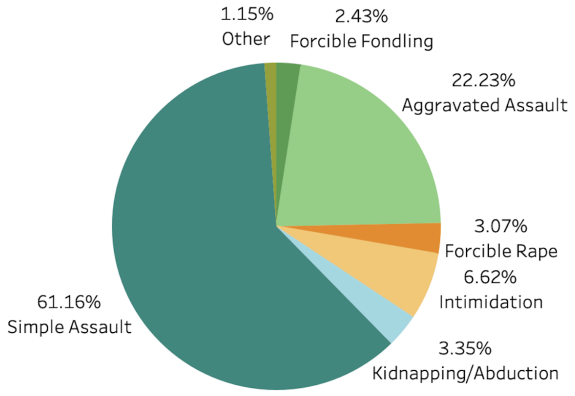
4. Data Analysis and Results

4.1. Offense Description

To explore what types of crimes were being committed against women in San Diego County, the CIBRS Offense Description field of the dataset was mapped out.

Figure 2

Distribution of Offense Description



Over 80% of crimes committed were related to assault, with the majority being attributed to Simple Assault (61.16%) yet a still significant proportion being attributed to the Aggravated Assault (22.25%). Simple assault is defined as, "an act that causes someone to fear physical harm or injury" and can range from threatening to hurt someone to actually "slapping or shoving" someone. On the other hand, aggravated assault is a more severe crime, and, "typically involves the use of a weapon and/or severe bodily injury of another person" (*Understanding the Differences Between Simple and Aggravated Assault: Know Your Rights* 2023). Examples include use of a deadly weapon to attack a victim, or attacking an elderly person. Obviously, it was concerning to see that about one fifth of the crimes committed against women in San Diego fell under the more violent classification of aggravated assault.

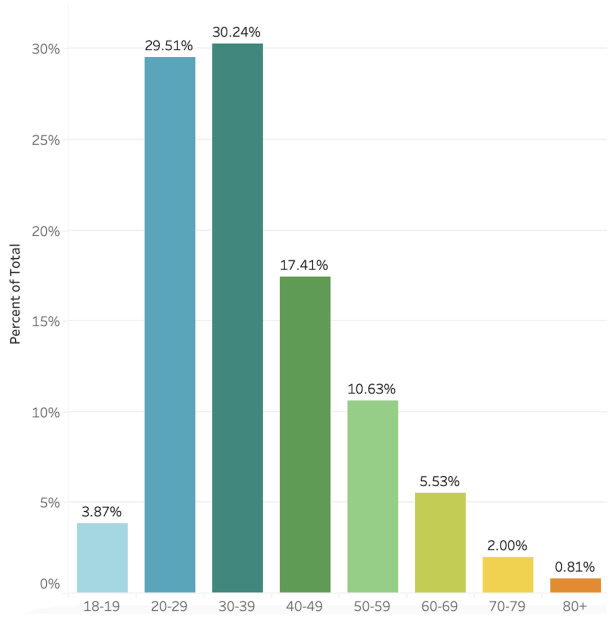
Meanwhile, the remaining crimes of Intimidation (6.62%), Kidnapping/Abduction (3.35%), Forcible Rape (3.07%), Forcible Fondling (2.43%), and Other (1.15%) were reported at significantly lower rates.

4.2. Victim Age

To analyze the distribution of Victim Age, distinct age groups were created in the dataset. The figure below shows the results.

Figure 3

Distribution of Victim Age Groups



Note: Because this project is only investigating crimes committed against adult women, the first age group bin has an interval of two years (18-19) while the rest have intervals of nine years (ex. 20-29)

The distribution was positively skewed, indicating that the majority of victims were younger in age.

Next, these results were weighted to the Census Bureau's age group distribution of females in San Diego County, to determine if any age groups were being disproportionately victimized. To do this, data was referenced from the Census Bureau's Age and Sex table. First, new Per Capita Counts were calculated relative to the female population size using the formula below:

Figure 4

Formula for Calculating Per Capita Rates

$$\textit{Per Capita Count for Group X} = \frac{\textit{Number of Victims in Group X}}{\textit{Total Population of Group X}}$$

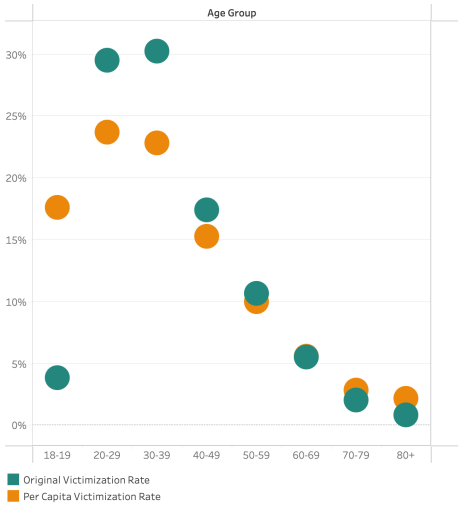
$$\textit{Per Capita Rate for Group X} = \frac{\textit{Per Capita Count for Group X}}{\textit{New Population Total}}$$

Note: The hypothetical New Population Total was calculated by summing all the Per Capita Counts.

The resulting Per Capita Victimization Rates are displayed below alongside the original rates.

Figure 5

Comparison of Original Victimization Rates and Per Capita Victimization Rates for Age Groups



The Per Capita Rates for the 70-79 and 80+ groups were slightly higher than the Original Rates, indicating that these older groups were disproportionately victimized. Also, the Per Capita Rate for 18-19 was significantly higher than the Original Rate, suggesting that this age group is targeted the most disproportionately.

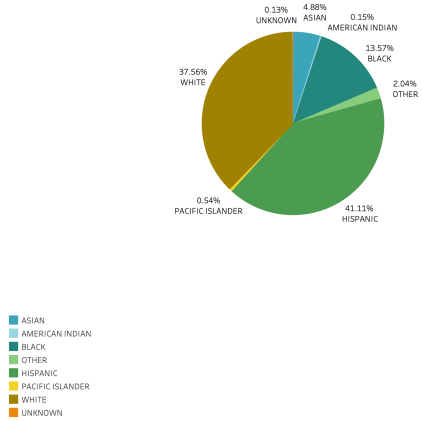
On the other hand, the Per Capita Rates for age groups 40-49 and 50-59 were slightly lower than the Original Rates, indicating that they were slightly underrepresented as victims. And the Per Capita Rates for age groups 20-29 and 30-39 were significantly lower than the Original Rates, suggesting that they were significantly underrepresented as victims.

4.3. Victim Race

Next, an investigation of the race of the victims was performed, according to the prespecified racial groups.

Figure 6

Distribution of Victim Race



Most victims were of White (37.56%) or Hispanic (41.11%) background, matching the fact that these are the largest racial groups in San Diego County according to the US Census Bureau. To investigate whether certain races of women are more frequently targeted, the crime rates were adjusted according to the population rates of each race in the county.

Referencing the Hispanic or Latino Origin by Race table from the Census Bureau data, matching their field categories to the category names created as follows:

Figure 7

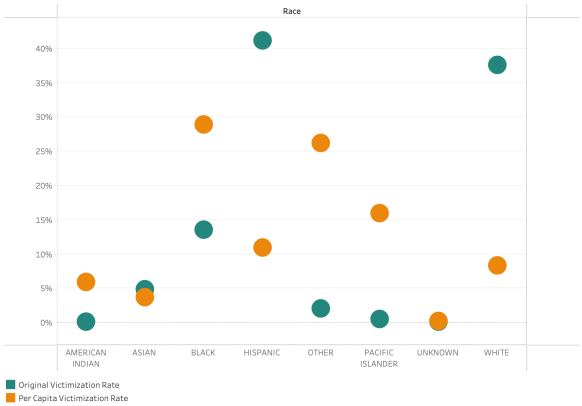
Comparison of Census Bureau's Category Names to SANDAG Category Names

Census Bureau Category	SANDAG Category
American Indian and Alaska Native alone	American Indian
Asian alone	Asian
Black or African American alone	Black
Native Hawaiian and Other Pacific Islander alone	Pacific Islander
Some Other Race alone	Other
Two or More Races	Unknown
White alone	White

For each of these matched groups, their population rates were taken in San Diego and used the same formulas as above to calculate Per Capita Rates. The resulting Per Capita Victimization Rates are displayed below alongside the original rates:

Figure 8

Comparison of Original Victimization Rates and Per Capita Victimization Rates for Victim Race



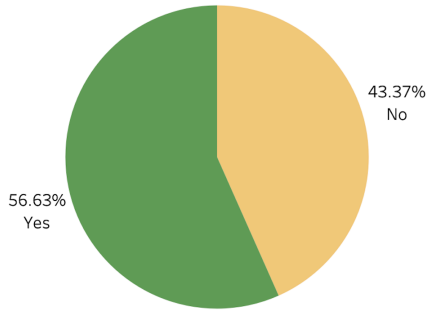
Notably, the Per Capita Rates for American Indian, Black, Other, and Pacific Islander women were significantly higher than their Original Rates, indicating that these communities were disproportionately targeted. On the other hand, the Per Capita Rates for Hispanic and White women were significantly lower than their Original Rates, suggesting that they were underrepresented as victims relative to their population size. Meanwhile, the Per Capita Rates for Asian women and women of Unknown racial origins were close to their Original Rates, indicating that they were represented relatively proportionally.

4.4. Domestic Violence

Then, the distribution of domestic violence related incidents versus non-domestic violence related incidents were explored.

Figure 9

Distribution of Domestic Violence Occurrence



It was discovered, more than half of cases were considered to be incidents of domestic violence (56.63%). This was another concerning finding.

4.5. Incident Date

In an effort to identify temporal crime hotspots, a time analysis on the number of Incidents Over Time was done. As a reminder, the data ranges from January of 2021 to September of 2024.

Figure 10

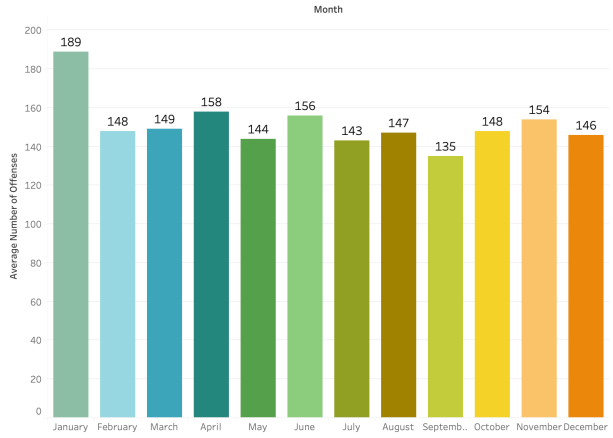
Number of Offenses Over Time

Note: This chart measures the frequency of individual offenses, not individual incidents

The rate of crimes committed against women fluctuated significantly during this time period. Starting at 1224 incidents in January of 2021, the rate experienced several notable peaks: May 2021 (1429 incidents), July 2021 (1423 incidents), and July 2023 (1340 incidents). It also experienced several significant low points: February 2021 (1064), February 2023 (998) and February 2024 (1054). To determine if February experienced significantly lower crime rates, the average offense rates for each month were investigated.

Figure 11

Average Number of Offenses Per Month



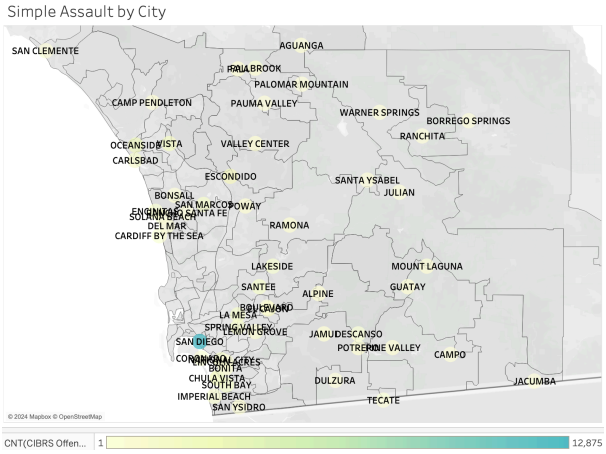
It appears that February did not have a significantly lower average rate of crime in comparison to other months. In fact, February's average rate of about 148 incidents per month was higher than that of May (144), July (143), and September (143). Even more significant is the highest average rating, which is January's 189.

4.6. Geographic Analysis

To understand the distribution of each type of reported incident across the county, pictured below is the geographic visualization for cases that fall under the category of Simple Assault.

Figure 12

Geographic Distribution of Simple Assault Cases



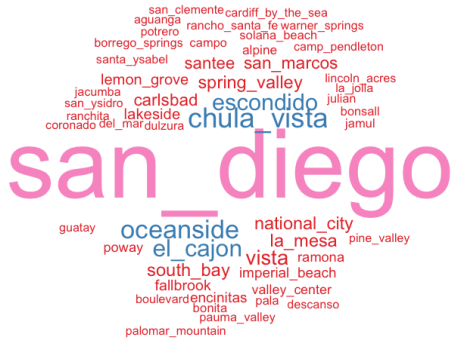
Upon conducting similar analyses with each offense description, it was discovered the same geographic distribution, with most cases being located in the city of San Diego. So, only one of these visualizations was included in this report.

4.6.1 Area Word Clouds

Word clouds are another way to visualize the frequency of certain words or in this case, areas, in a dataset. This method explores total crime rates per area.

Figure 13

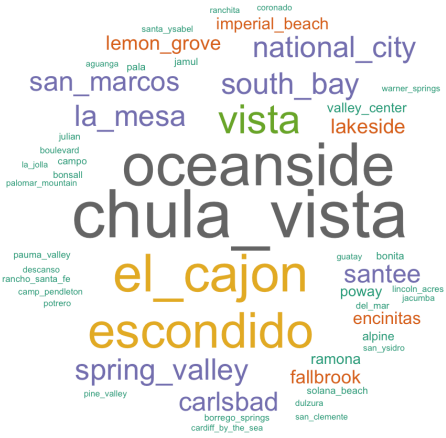
Word Cloud Visualizing the Geographic Distribution of Total Crime Occurrences



Because the city of San Diego took up so much space in this word cloud, "San_Diego" was used as a stopword for the next word cloud so more information could be gleaned on the other cities.

Figure 14

Word Cloud Visualizing Geographic Distribution With Stopword "San_Diego"



Following that, a large representation from some of the major cities in the county, such as Chula Vista and Oceanside showed more prominently. From there, Per Capita Rates were calculated for each area according to the formula in order to determine if any areas were experiencing disproportionately high rates of crimes against women.

4.6.2. Crime Rate Standardization Score

To visualize these weighted rates on a map, a standardized metric was created to represent the difference between an area's Original Rate and Per Capita rate. This metric, termed the Crime Rate Standardization Score (CRSS), was calculated using z-scores for both the Original and Per Capita rates:

Figure 15

Formula for Calculating Crime Rate Standardization Scores (CRSS)

$$z_{\text{original}} = \frac{\text{rate}_{\text{original}} - \mu_{\text{original}}}{\sigma_{\text{original}}} \quad z_{\text{per capita}} = \frac{\text{rate}_{\text{per capita}} - \mu_{\text{per capita}}}{\sigma_{\text{per capita}}}$$

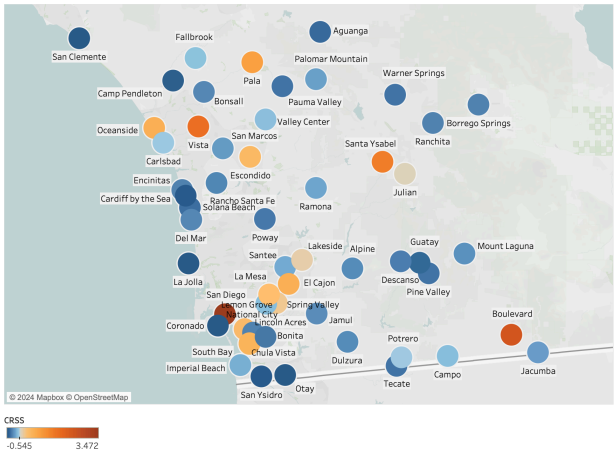
$$\text{Crime Rate Standardization Score} = \frac{z_{\text{original}} + z_{\text{per capita}}}{2}$$

Negative CRSS values indicate that a city is underrepresented in the data. Positive values indicate that a city is overrepresented in

the data. The further the value is from zero, the stronger the misrepresentation is.

Figure 16

Geographic Distribution of Crime Rate Standardization Scores



The city of San Diego still stands out as having high rates in relation to its population. Meanwhile, most areas were slightly underrepresented in the data, with slightly negative CRSS values. However, the areas that were overrepresented were overwhelmingly so, with very positive CRSS values. This is why the lowest CRSS value is -0.545 while the highest value is 3.472.

These areas with positive scores, or overrepresentation in the dataset, were San Diego (3.472), Boulevard (2.351), Vista (1.690), Santa Ysabel (1.409), Pala (0.9030), Oceanside (0.6078), El Cajon (0.5891), Chula Vista (0.5124), South Bay (0.4564), Escondido (0.4013), National City (0.3928), La Mesa (0.3136), Spring Valley (0.1072), Lakeside (0.1072), and Julian (0.0453).

These positive values may suggest that women in these areas are disproportionately targeted, however it is important to note that these rates are only for *reported* crimes. This dataset does not have information about unreported crimes, and it is not known how many crimes go unreported.

4.7. Factors Over Time

Analyses were conducted for each of the factors explored (Offense Description, Victim Age, Victim Sex, Domestic Violence, Incident Date, Geographic Analysis) to identify trends over the three-year period of the dataset. While no significant trends were found for individual factors, an overall decrease in total crime rates between 2023 and 2024 was observed, which was reflected across all factors.

5. Predictive Modeling

The predictive models for Incident Prediction and Domestic Violence Risk were developed using Scikit-learn's Random Forest framework. For the predictive models, several key datetime attributes, including Hour, Day of Week, and Month, were derived from the Incident Time field and added as separate columns. This section will provide a comprehensive overview of the purpose, evaluation metrics, and future goals for both models.

5.1. Incident Prediction Model

The first predictive model that was developed inputs incident details such as victim demographic information, location of the incident, and time to predict the most likely incident type for a given situation. The purpose of this model is to raise awareness about the types of incidents that women are most likely to face, enabling them to take proactive measures for their safety. The feature columns to be defined were Victim Age, Overall Race, City, Hour, Day of Week, Day of Month, and Month. For the geographic input, City (with 55 unique values) was selected over the HHSA Region (5 unique values) or Zip Code (112 unique values) columns because it provided greater generalizability than the specificity of Zip Code allowed for, but still offered a more localized level of detail than the HHSA Region column. The data was split into testing and training sets, with 90% of the data used for training and the remaining 10% used for testing and model evaluation. A label encoder was used to map the categorical columns like Overall Race and City to numeric values to be interpreted by the model. The Random Forest used 100 estimators (Decision Trees) to make its predictions, and it was found that increasing this number did not have any significant effect on performance. The classification report for the finalized model was generated with Scikit-learn's classification report method.

Figure 17

Classification Report for Incident Type Prediction Model

Class	precision	recall	f1-score
Aggravated Assault	0.22	0.04	0.07
Forcible Fondling	0.33	0.03	0.06
Forcible Rape	0.14	0.02	0.03
Intimidation	0.38	0.04	0.08
Kidnapping/Abduction	0.01	0.01	0.01
Other	0.18	0.04	0.06
Simple Assault	0.63	0.94	0.76
accuracy	0.59	0.59	0.59
macro avg	0.27	0.16	0.15
weighted avg	0.48	0.59	0.49

The classification report highlights significant challenges in predicting incident types due to the imbalanced nature of the dataset. The majority classes, particularly Simple Assault and Aggravated Assault, take up approximately 83% of all cases in the dataset. This leads to higher precision, recall, and F1-scores for these classes, notably Simple Assault, compared to others. In contrast, minority classes like "Kidnapping/Abduction" and "Forcible Rape" have poor predictive performance, with low precision, recall, and F1-scores. Each of the minority classes—except for the case of Kidnapping/Abduction—exhibits significantly higher precision than recall. For example, the Intimidation class has a recall of 0.04 but a precision of 0.38, which is more than nine times its recall. This indicates that the model has few false positives for minority classes; when it does predict a minority class, it is often correct. However, the low recall rates suggest that the model rarely predicts these minority classes, likely due to the imbalance in the training data. As a result, predictions are heavily biased toward the majority class, favoring it at the expense of accurately identifying minority incidents.

Future improvements could focus on addressing this imbalance through techniques such as oversampling the minority classes (e.g., Synthetic Minority Oversampling Technique or SMOTE), undersampling the majority classes, or applying class-weighting strategies in the model. Additionally, collecting more data for underrepresented classes could enhance predictive performance for minority categories. These efforts would create a more robust and equitable model capable of handling diverse incident types. Alternatively, future work could involve developing a binary classifier to predict whether an incident is "Simple Assault" or not,

as this approach would result in a more balanced distribution and potentially improve overall model performance.

5.2. Domestic Violence Prediction Model

The Domestic Violence Risk Prediction model was designed to classify incidents as either domestic violence (DV) or non-domestic violence. Its purpose was to provide a tool for identifying high-risk DV cases, enabling early intervention and resource allocation to support victims. Unlike the Incident Prediction model, which predicted multiple incident types, this model focused on a binary classification task. The same features—Victim Age, Overall Race, City, Hour, Day of Week, and Month—were used. The model was designed to minimize false negatives, as misclassifying domestic violence cases could have serious consequences. To prioritize caution, a lower threshold of 40% was implemented for predictions. This means that if the model was at least 40% confident that a case involved domestic violence, it was classified as such, ensuring a higher likelihood of identifying potential DV cases even at the risk of increased false positives.

Figure 18

Classification Report for Domestic Violence Risk Model

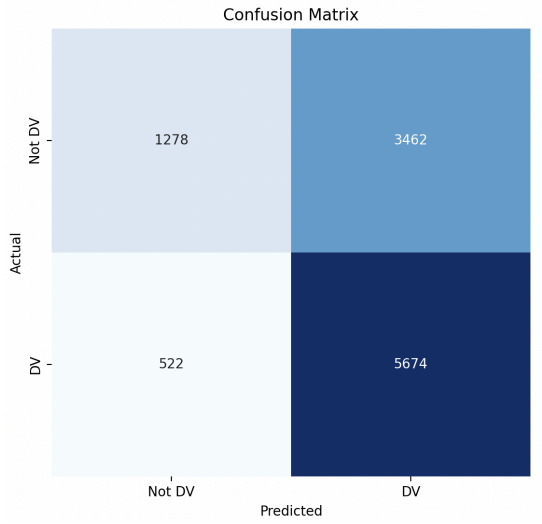
Class	precision	recall	f1-score
DV	0.62	0.92	0.74
Not DV	0.71	0.27	0.39
accuracy	0.64	0.64	0.64
macro avg	0.67	0.59	0.57
weighted avg	0.66	0.64	0.59

This classification report reveals a notable imbalance in performance metrics between the DV and Not DV classes. The model achieved a recall of 0.92 for the DV class, meaning it was able to successfully identify the majority of true DV cases. The precision of 0.62, indicates a higher rate of false positives. Conversely, for the Not DV class, precision is higher at 0.71, but recall drops significantly to 0.27, meaning many cases that truly were not domestic violence-related were misclassified as DV. This trade-off reflects the model's prioritization of minimizing false negatives for

DV cases, which is an important consideration given the potential risks of under-identifying these incidents. Compared to the Incident Prediction model, which struggled more significantly with minority classes, the DV model demonstrates stronger recall for its minority class (Not DV) largely because the training data was not as skewed, with approximately 57% of incidents being classified as Domestic Violence.

Figure 19

Confusion Matrix of Test Data for Domestic Violence Risk Model



The confusion matrix shows that the model correctly identified 5,674 true DV cases while misclassifying 522 DV cases as Not DV (false negatives). This highlights its strong recall for DV incidents and aligns with the goal to minimize false negatives. However, the model struggled with the Not DV class, correctly identifying only 1,278 true Not DV cases while misclassifying 3,462 Not DV cases as DV (false positives). This imbalance indicates a trade-off: the model favors classifying cases as DV to ensure potential incidents are not missed, at the expense of a high false positive rate for Not DV cases. This suggests the need for further refinements to improve precision for the Not DV class without sacrificing the high recall for DV cases. Adjusting thresholds, incorporating additional features, or fine-tuning class weighting could help achieve a better balance. In practical terms, this confusion matrix demonstrates that the model is highly effective at flagging DV incidents, but may require additional post-prediction validation steps to handle false positives effectively.

5.3. Implementation and Deployment

Both predictive models were integrated into user-friendly applications using Flask-based APIs. These APIs enabled real-time interaction with the models, allowing users to input relevant features and receive predictions and probabilities

immediately. To provide an intuitive interface, HTML files were created for each model, allowing users to enter features such as victim age, race, city, and time-related details through dropdown menus and forms. These HTML interfaces were embedded into a Google Sites page. This approach ensured that the models were not only functional but also practical and convenient for end-users, requiring no specialized software or technical expertise to interact with the predictions. More details on how the models were trained and implemented can be found in the Github repository here: https://github.com/alisa-crowe/bda594-project

6. Discussion

The analysis of the SANDAG CIBRS Group A Public Crime Data victim demographic information indicated that these communities of women may be disproportionately targeted as victims of crime in San Diego County:

- Women who are 18-19 years in age
- Women who are above the age of 60
- Women of color, namely women of American Indian, Black, Other, and Pacific Islander heritage
- Women living in the following areas: Boulevard, Chula Vista, El Cajon, Escondido, Julian, Lakeside, La Mesa, National City, Oceanside, Pala, San Diego, Santa Ysabel, South Bay, Spring Valley, and Vista.

The analysis of the nature of offenses also suggests that these types of offenses are being reported at high rates:

- Assault-related offenses, simple or aggravated
- Domestic-violence related offenses
- Offenses occurring during the month of January

The predictive modeling involved two primary tasks: incident type prediction and domestic violence risk prediction. These models have demonstrated the potential to identify patterns and predict future risks by analyzing historical data. The feature importance charts reveal critical insights about what factors are most influential in predicting outcomes. In both models, Victim Age is the most important predictor, highlighting the vulnerability of certain age groups that are consistent with the findings of the analyses. Temporal factors such

as Hour and Month are also critical, suggesting patterns in when crimes and domestic violence incidents are likely to occur. While it was found that the city in which the incident occurred as well as the victim's race were not largely influential in the models' outcomes, this does not diminish their significance as factors that could provide valuable context and inform targeted interventions. With further development, these models have potential to support more targeted interventions and resource allocation.

To interact with the data visualizations or predictive models, please visit the comprehensive project website using this URL: https://sites.google.com/sdsu.edu/bda-594-final-project-star/home

7. Conclusion

The findings from this research could have significant policy implications. For example, they could inform local law enforcement strategies and community support programs aimed at preventing crimes against women. Comparing the trends observed in San Diego county with those of similar urban areas could highlight unique or common factors, and provide broader applications to different areas.

This study could also pave the way for future research. Investigating other factors such as economic conditions, social services availability, and perpetrator demographics could further understand why crimes against women occur, and help develop more effective prevention strategies. Additionally, investigating these same factors and how they affect crimes against *men* in San Diego County, could provide a good basis for comparison and discovery of which factors affect women uniquely. On top of this, it would be nice to conduct similar analyses on a dataset that defines victims by gender, not by sex, because studies show that trans women, who are likely considered 'Male' in this dataset, are multiple times more likely to be victims of violent crimes (Flores, et al., 2017-2018). Finally, analyses like this would benefit from understanding how many crimes and what kinds of crimes tend to go unreported, as well as what victims report crimes at lower rates. It may be interesting to explore relationships between police agencies and people in border counties, and how this may affect rates of crime reporting

in San Diego County. This knowledge could be applied to determine what percentage of the above findings can be attributed to actual observable crime trends.

In conclusion, addressing crimes against women in San Diego County requires a focused and nuanced approach, given the unique demographic and socioeconomic factors at play. While overall crime rates in the United States have been declining, the specific trends in San Diego highlight the need for localized studies to understand the dynamics of female victimization. By investigating the changing patterns of crimes against women and identifying predictive factors, more effective prevention strategies and support systems can be developed. This comprehensive analysis is crucial for enhancing the safety and well-being of women in this diverse metropolitan area.

References

Amber, P., Alina, P., Megan, O., Kelly, T., Niyati, S., Sabine, O.-P., & Nicole, van G. (2020, April). *(PDF) pandemics and Violence Against Women and Children*. UN.org. https://www.researchgate.net/publication/344164631_Pandemics_and_Violence_Against_Women_and_Children

Boman, J. H., & Gallupe, O. (2020, July 8). *Has covid-19 changed crime? crime rates in the United States during the pandemic - american journal of criminal justice*. SpringerLink. https://link.springer.com/article/10.1007/s12103-020-09551-3

Caetano, R., Vaeth, P. A. C., Gruenewald, P. J., Ponicki, W. R., Kaplan, Z., & Annechino, R. (2021, February 25). *Proximity to the U.S./Mexico border, alcohol outlet density and population-based sociodemographic correlates of spatially aggregated violent crimes in California*. Science Direct. https://www.sciencedirect.com/science/article/abs/pii/S1047279721000302?via%3Dihub

Han, Y. (2023, September 15). *Relationship between extreme heat and violent crime in San Diego County: Analysis and recommendations for crime prevention and climate change mitigation*. eScholarship, University of California. https://escholarship.org/uc/item/9xk653gx

Flores, A. R., Meyer, I. H., Langton, L., & Herman, J. L. (2021). Gender Identity Disparities in Criminal Victimization: National Crime Victimization Survey, 2017-2018. *American journal of public health*, *111*(4), 726–729. https://doi.org/10.2105/AJPH.2020.306099

Jackman, T. (2020, May 19). *Amid pandemic, crime dropped in many U.S. cities, but not all - The Washington Post*. The Washington Post. https://www.washingtonpost.com/crime-law/2020/05/19/amid-pandemic-crime-dropped-many-us-cities-not-all/

Population. City of San Diego Official Website. (n.d.). https://www.sandiego.gov/economic-development/sandiego/population#:~:text=Because%20of%20its%20proximity%20to,diverse%20places%20in%20the%20nation

San Diego CIBRS Group A Data Report. SANDAG Open Data Portal. (n.d.). https://opendata.sandag.org/stories/s/San-Diego-CIBRS-Group-A-Data-Report/75u9-txik/#:~:text=Group%20A%20offenses%20include%20more,)%2C%20and%20Liquor%20Law%20Violation

Understanding the Differences Between Simple and Aggravated Assault: Know Your Rights. Home. (2023, April 11). https://www.gregglawdallas.com/understanding-the-differences-between-simple-and-aggravated-assault-know-your-rights

Wang, H. L. (2022, February 17). *The U.S. Census sees middle eastern and North African people as white. many don't*. NPR. https://www.npr.org/2022/02/17/1079181478/us-census-middle-eastern-white-north-african-men