CSE3040 Exploratory Data Analysis

J Component - Project Report

Review III

ANALYSIS OF CRIME AGAINST WOMEN IN INDIA

By

22MIA1061 HRISHIKESH
22MIA1072 PHOOBESH S

M.Tech CSE with Specialization in Business Analytics

Submitted to

Dr.A.Bhuvaneswari,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

May 2024

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

WINTER SEM 23-24

Worklet details

Programme	M.Tech with Specialization in Business Analytics	
Course Name / Code	Exploratory data analysis / CSE 3040	
Slot	F1 Slot	
Faculty Name	Dr.A.Bhuvaneswari	
Digital Assignment		
Team Members Name | Reg. No	HRISHIKESH	22MIA1061
	PHOOBESH S	22MIA1072

Team Members(s) Contributions – Tentatively planned for implementation:

Worklet Tasks	*Contributor's Names*
Dataset Collection	PHOOBESH S
Preprocessing	HRISHIKESH
Architecture/ Model/ Flow diagram	PHOOBESH S
Model building (suitable algorithm)	HRISHIKESH
Results – Tables, Graphs	PHOOBESH S, HRISHIKESH
Technical Report writing	PHOOBESH S
Presentation preparation	HRISHIKESH

ABSTRACT

This abstract presents an analysis of crime against women in India through the lens of exploratory data analysis (EDA) concepts. The dataset under examination encompasses various dimensions of crimes reported against women across different regions and time periods in India. The analysis commences with an introduction to the dataset, followed by a discussion on the significance and urgency of addressing gender-based violence in India. EDA techniques are then applied to gain insights into the patterns, trends, and characteristics of crimes against women. Descriptive statistics provide an overview of the frequency and distribution of different types of crimes, including but not limited to rape, domestic violence, dowry-related violence, and sexual harassment. Temporal analysis explores how the prevalence of crimes against women has evolved over time, uncovering any seasonal variations or long-term trends. Spatial analysis examines the geographical distribution of reported incidents, identifying regions with higher rates of gender-based violence and potential hotspots requiring targeted intervention. Moreover, the analysis delves into demographic factors such as age, education, and socio-economic status to understand their correlation with the incidence of crimes against women. Visualizations such as heatmaps, histograms, and scatter plots aid in elucidating these relationships and highlighting disparities across different demographic groups and regions. By identifying gaps and inefficiencies in the legal framework and law enforcement mechanisms, the analysis seeks to inform policy interventions aimed at enhancing the protection of women's rights and promoting gender equality. In conclusion, this analysis of crime against women in India using EDA concepts provides valuable insights into the multifaceted nature of gender-based violence and its socio-cultural, economic, and institutional determinants. By leveraging data-driven approaches, policymakers, activists, and stakeholders can devise evidence-based strategies to prevent and address crimes against women, fostering a safer and more inclusive society for all.

Table of Contents

INTRODUCTION

Crime against women is a pervasive and deeply concerning issue globally, and India is no exception. Despite legislative reforms and societal awareness campaigns, gender-based violence continues to plague communities across the country, manifesting in various forms such as domestic abuse, sexual assault, dowry-related violence, and trafficking. Understanding the dynamics and patterns of these crimes is crucial for effective prevention and intervention efforts.

This study focuses on the analysis of crime against women in India through the application of exploratory data analysis (EDA) concepts. By examining a comprehensive dataset encompassing reported incidents of gender-based violence across different regions and time periods, this analysis seeks to uncover underlying trends, identify vulnerable demographics, and highlight geographic hotspots requiring targeted intervention.

The urgency of addressing crime against women in India cannot be overstated. Despite significant strides in women's empowerment and legislative measures to protect their rights, the prevalence of gender-based violence remains alarmingly high. According to the National Crime Records Bureau (NCRB), thousands of cases of rape, domestic violence, and other forms of violence against women are reported annually, representing only a fraction of the true extent of the problem due to underreporting and societal stigma.

Through this study, we aim to shed light on the multifaceted nature of gender-based violence in India and its underlying socio-economic, cultural, and institutional determinants. By leveraging data-driven insights, we seek to inform evidence-based policy interventions, strengthen legal frameworks, and enhance support services for survivors of violence. Ultimately, our goal is to contribute to the creation of a safer and more equitable society where women are afforded the fundamental right to live free from fear and discrimination.

PROBLEM BACKGROUND

Crime against women in India presents a complex and multifaceted challenge with deep-rooted socio-cultural, economic, and institutional dimensions. Despite significant progress in various spheres, including legislation and women's empowerment initiatives, the prevalence of gender-based violence persists at alarming levels, posing a significant threat to the safety, well-being, and dignity of women across the country**.**

Furthermore, underreporting of crimes against women remains a significant barrier to accurately assessing the magnitude of the problem and providing adequate support and redressal mechanisms for survivors. Factors such as fear of retaliation, social stigma, and lack of trust in the criminal justice system often deter victims from seeking help or reporting incidents of violence, leading to a gross underestimation of the true prevalence of gender-based violence.

Against this backdrop, there is an urgent need for comprehensive data-driven analysis to understand the dynamics of crime against women in India, identify underlying trends and patterns, and inform evidence-based interventions and policy reforms. By leveraging exploratory data analysis (EDA) techniques, this study aims to fill critical knowledge gaps, raise awareness, and advocate for systemic changes to create safer and more inclusive environments for women, where their rights are respected, protected, and upheld.

LITERATURE REVIEW:

Crime against women in India has garnered significant attention from researchers, policymakers, and activists, reflecting its profound societal implications and the urgency of addressing this pressing issue. A wealth of scholarly literature exists, examining various aspects of gender-based violence and its underlying determinants within the Indian context.

Studies have highlighted the pervasive influence of patriarchal norms and gender inequalities in perpetuating violence against women (Kabeer, 2005; Sen, 2001). These norms dictate women's subordinate roles in society, limiting their autonomy and exacerbating their vulnerability to abuse and exploitation. Furthermore, socio-economic factors such as poverty, lack of education, and limited access to resources have been identified as significant risk factors for experiencing violence (Kishor & Gupta, 2009).

Spatial and temporal analysis of crime data has emerged as a valuable tool for understanding patterns and trends in gender-based violence. Studies have utilized geographic information systems (GIS) to map the spatial distribution of reported incidents, identifying hotspots and areas with elevated risk levels (Kapur & Sharma, 2019). Temporal analysis has revealed seasonal variations and long-term trends in crime rates, shedding light on underlying factors driving fluctuations (Kumar & Pradhan, 2016).

Overall, the literature underscores the need for holistic, multi-sectoral approaches to address crime against women in India. By combining legal reforms, socio-economic empowerment initiatives, and data-driven interventions, stakeholders can work towards creating safer, more equitable societies where women are afforded the rights and protections they deserve.

PROBLEM STATEMENT

Despite concerted efforts to address crime against women in India, the persistence of gender-based violence remains a significant challenge, reflecting deep-rooted societal norms and structural inequalities. The problem is exacerbated by underreporting, inadequate access to justice, and limited support services, perpetuating cycles of violence and impunity. Furthermore, disparities in socio-economic status and geographic location exacerbate the vulnerability of marginalized women. To effectively combat this issue, there is a critical need for comprehensive data-driven analysis to understand the dynamics of gender-based violence, identify underlying trends, and inform evidence-based interventions. This study seeks to address these knowledge gaps by applying exploratory data analysis (EDA) techniques to analyze crime data, ultimately contributing to the development of targeted strategies and policy reforms aimed at promoting women's safety, rights, and well-being in India.

OBJECTIVE

- Conduct comprehensive analysis of crime against women data in India to unveil patterns, trends, and correlations.
- Develop predictive models to accurately forecast trends in gender-based violence incidents.
- Explore the temporal patterns of crime against women to identify peak periods and seasonality.
- Investigate the spatial distribution of incidents to pinpoint geographical hotspots and areas of heightened risk.
- Explore the intersectionality of gender-based violence with other forms of discrimination, including caste, ethnicity, religion, and sexual orientation.
- Identify emerging trends and shifts in patterns of gender-based violence over time.

These objectives aim to provide a holistic understanding of crime against women in India, informing evidence-based interventions and policy reforms to create safer and more inclusive communities for all.

DATA SET & TOOLS USED

The dataset for analysis and prediction was obtained from www.kaggle.com. The dataset files with data related to cases and arrests of crimes against women in the past years. This dataset encompasses details such as area name, year, group name, sub-group name, and various case details such as the number of cases acquitted or discharged, cases charge-sheeted, cases compounded or withdrawn and cases convicted.

KEY FEATURES:

- **Area Name**: The dataset likely includes information about different regions or areas where these crimes occurred. This could be at the city, district, or state level.
- **Year:** The year of the reported data. It's essential to understand the temporal context and any trends over time.
- **Group and Sub-Group Names:** These categories help classify the types of crimes. For example:
- **Group:** Crimes related to violence, harassment, or exploitation.
- **Sub-Group:** Specific types of crimes within each group (e.g., domestic violence, sexual assault, human trafficking).
- **Case Details:** Number of Cases Acquitted or Discharged: Indicates the cases where the accused were not held responsible.
- **Number of Cases Charge-Sheeted:** Refers to cases where charges were formally filed against the accused.
- **Number of Cases Compounded or Withdrawn:** Instances where victims or their families decided not to pursue legal action.
- **Number of Cases Convicted:** Indicates successful legal outcomes where the accused were found guilty

ALGORITHMS

LINEAR REGRESSION

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc. This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.

RANDOM FOREST

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative

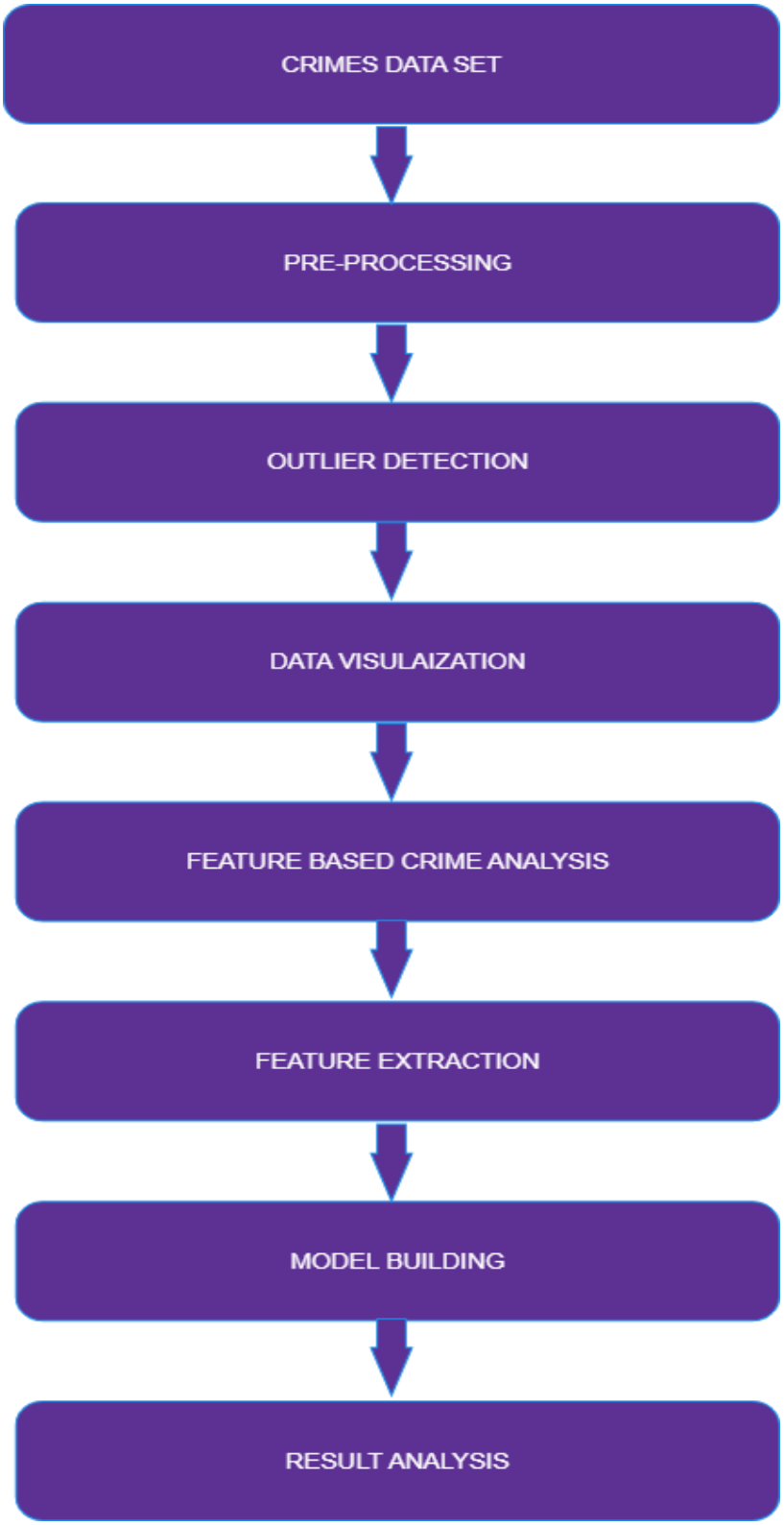
decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.

ONE HOT ENCODER

One-hot encoding is a technique used in machine learning and data preprocessing, particularly in the context of categorical variables. It's a way of representing categorical data so that it can be used in machine learning algorithms, which typically require numerical input.
Here's how it works:

1. **Categorical Variables:** Categorical variables are variables that can take on a limited, fixed number of values. For example, "color" could be a categorical variable with values like "red," "blue," and "green."
2. **Integer Encoding:** Before applying one-hot encoding, categorical variables are often converted into integer values. Each unique category is assigned an integer identifier. For example:
3. **One-Hot Encoding:** After integer encoding, each integer value is represented as a binary vector that is all zero values except for the index of the integer, which is marked with a 1. This is why it's called "one-hot" encoding.

SYSTEM ARCHITECTURE



IN THIS ARCHITECTURE:

1. **Dataset:** The Crime dataset serves as the initial data source, containing records of crimes against women.
2. **Preprocessing**: Data preprocessing involves cleaning, transforming, integrating, and normalizing the dataset to prepare it for further analysis.
3. **Feature Extraction:** Relevant features are extracted from the dataset to capture important aspects of customer behavior and preferences.
4. **Data Visualization:** Visualizations are generated to explore the relationships between features, identify patterns, and gain insights into the underlying structure of the data.
5. **Feature Scaling:** Feature scaling is applied to standardize the range of features in the dataset, ensuring all features contribute equally to the analysis.
6. **Visualization:** Visualizations are generated to visualize crime patterns and facilitate interpretation.
7. **Result Analysis:** The extensive analysis that has been done provides the overall brief of crimes against women over the years in India.

MODULE DESCRIPTION

The project begins by meticulously curating datasets tailored for comprehensive crime analysis in India. Through careful filtering and integration of external sources, a robust dataset is assembled to delve into the intricacies of gender-based violence and crime patterns across different regions and timeframes.

Data preprocessing emerges as a pivotal step, ensuring the dataset's integrity and readiness for analysis. Addressing challenges like missing data and inconsistencies, this phase lays the groundwork for accurate insights into crime dynamics.

Cleaning the data involves meticulous error rectification and integrity maintenance. Techniques such as imputation and outlier detection are employed to handle noisy data, ensuring that the dataset is well-structured and conducive to meaningful analysis.

Visualization techniques, facilitated by tools like Matplotlib, offer intuitive representations of crime patterns and trends in India. Through visually compelling graphs and charts, the analysis unveils insights into the spatial and temporal dynamics of gender-based violence, aiding in the identification of hotspots and trends.

Feature selection plays a crucial role in identifying the key determinants of gender-based violence. Through advanced analysis techniques, redundant or irrelevant features are pruned, enhancing the predictive accuracy of the models and providing valuable insights into the factors driving crime rates.

This project serves as a comprehensive exploration of crime analysis methodologies in the Indian context, offering actionable insights to inform policy interventions and support services. Additionally, it provides a practical learning experience in data analysis techniques and Python programming, empowering stakeholders with evidence-based strategies to address gender-based violence effectively.

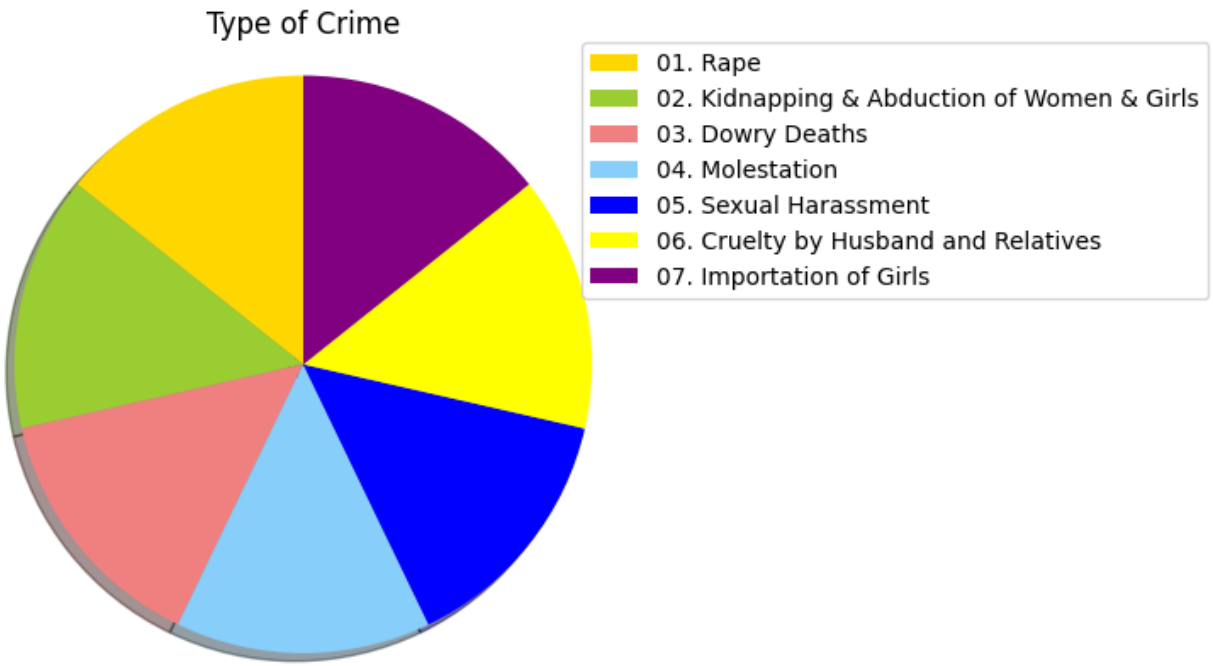
RESULT ANALYSIS

CORRELATION MATRIX OF CRIME DATA

0.626288	0.894387	0.886829	0.979532	1.000000
0.479304	0.773011	0.790971	0.898131	0.908630
0.003758	0.028246	0.031851	0.029464	0.031259
0.200507	0.040446	0.040308	0.088885	0.071187
0.564572	0.995146	0.995713	0.887791	0.925765
0.511221	0.840507	0.852710	0.819567	0.838302
0.341075	0.562538	0.578592	0.730773	0.742083
0.517174	0.830166	0.835874	0.878659	0.881364
0.508417	0.854536	0.860175	0.883146	0.896084
0.183977	0.365005	0.380630	0.588188	0.550037
0.853060	0.660652	0.643107	0.728930	0.715242
0.825602	0.650743	0.636559	0.719263	0.697667
0.420858	0.901250	0.912612	0.730547	0.763909
0.325041	0.216511	0.215407	0.402348	0.336929
0.462551	0.782535	0.797537	0.844944	0.846065

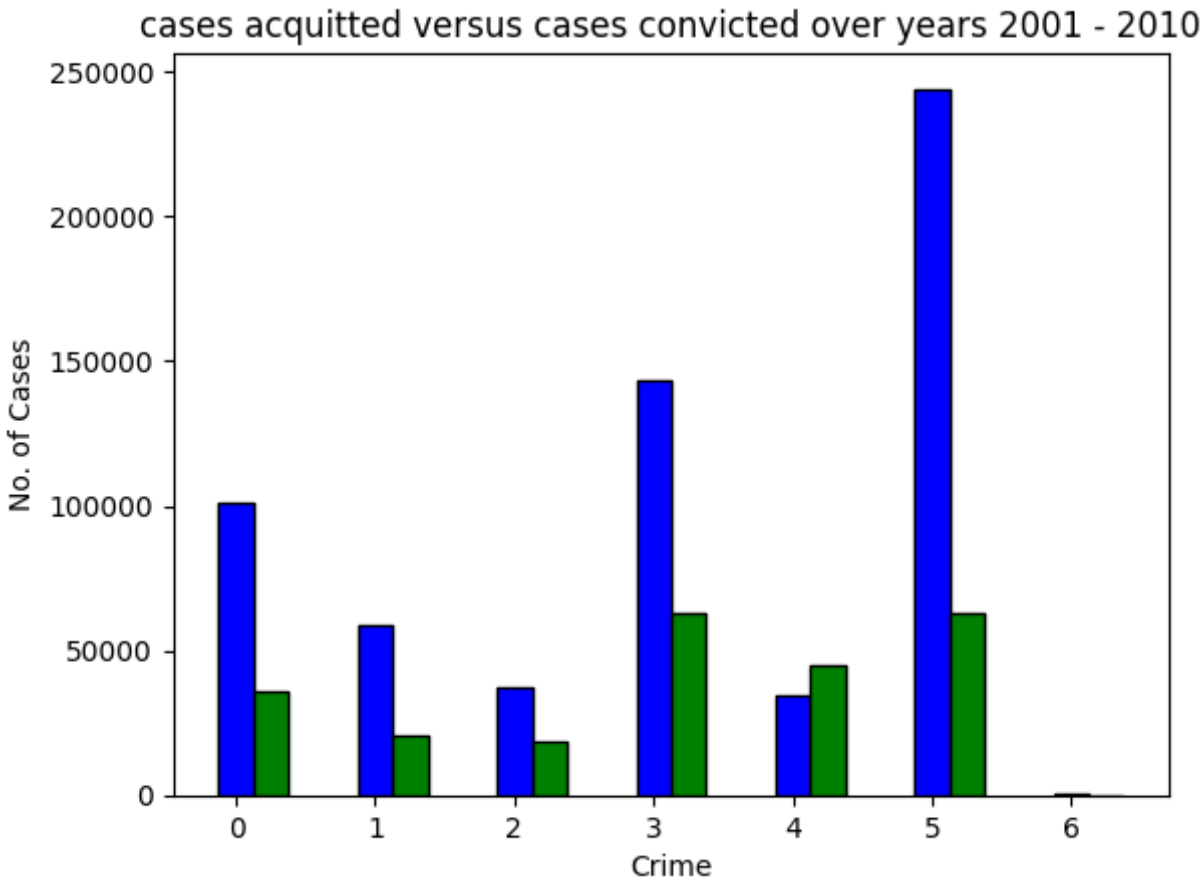
The heatmap displays correlations between variables, with cooler colors indicating negative correlations and warmer colors indicating positive correlations.

DISTRIBUTION OF CRIME TYPES



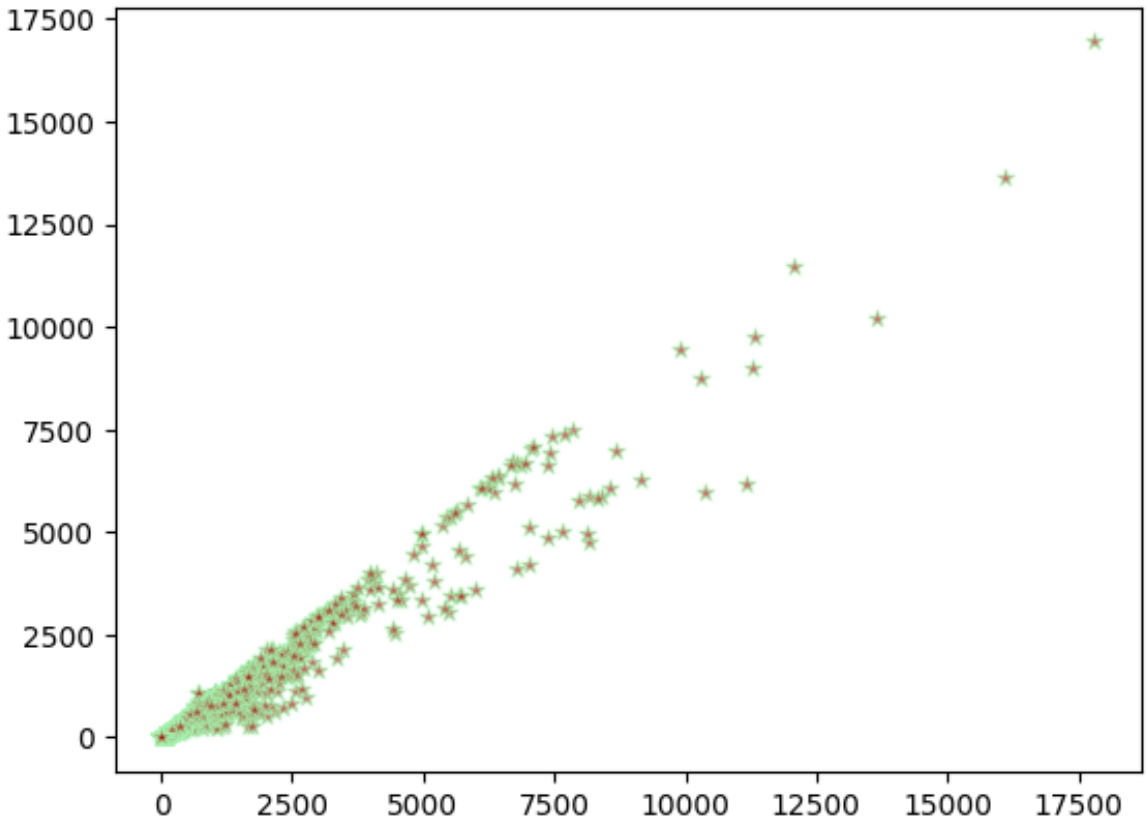
The pie chart visualizes the distribution of different crime types, with colors indicating each type. "Type of Crime" chart illustrates the relative frequencies distinctly.

COMPARISON OF CASES ACQUITTED AND CONVICTED (2001-2010)



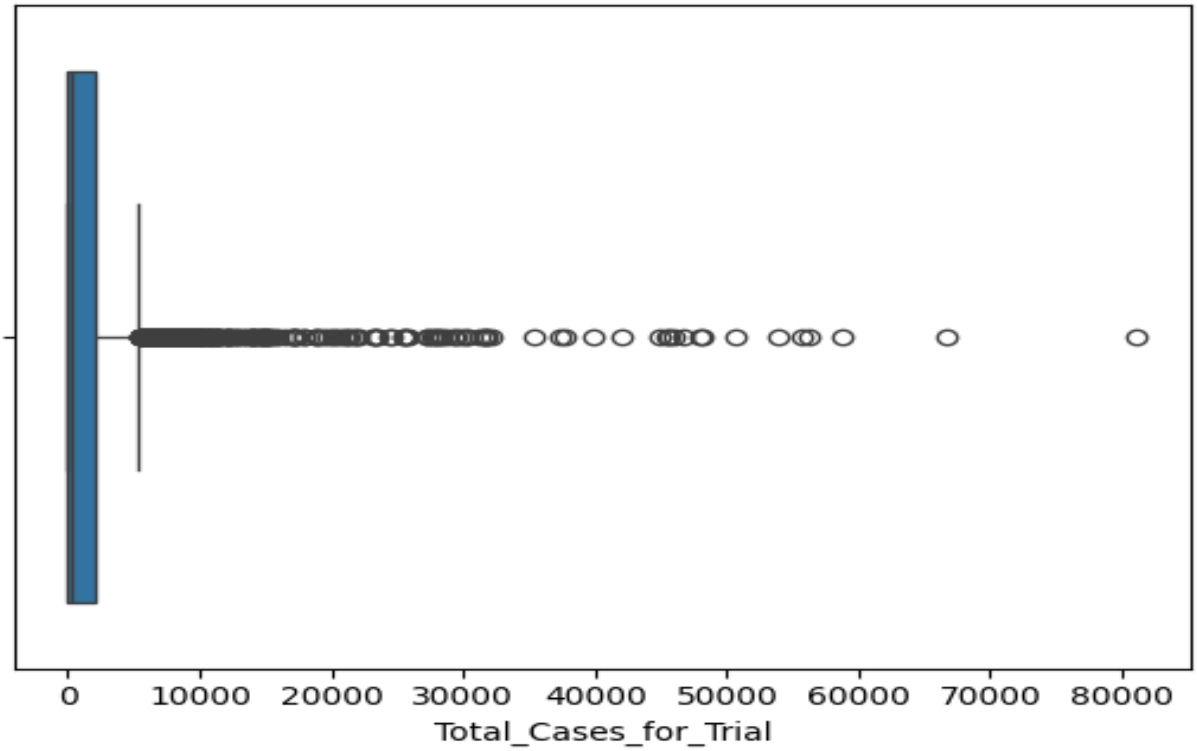
The bar graph displays the comparison between cases acquitted and cases convicted for different crimes from 2001to 2010, illustrating their respective frequencies.

RELATIONSHIP BETWEEN CASES REPORTED AND CASES SENT FOR TRIAL



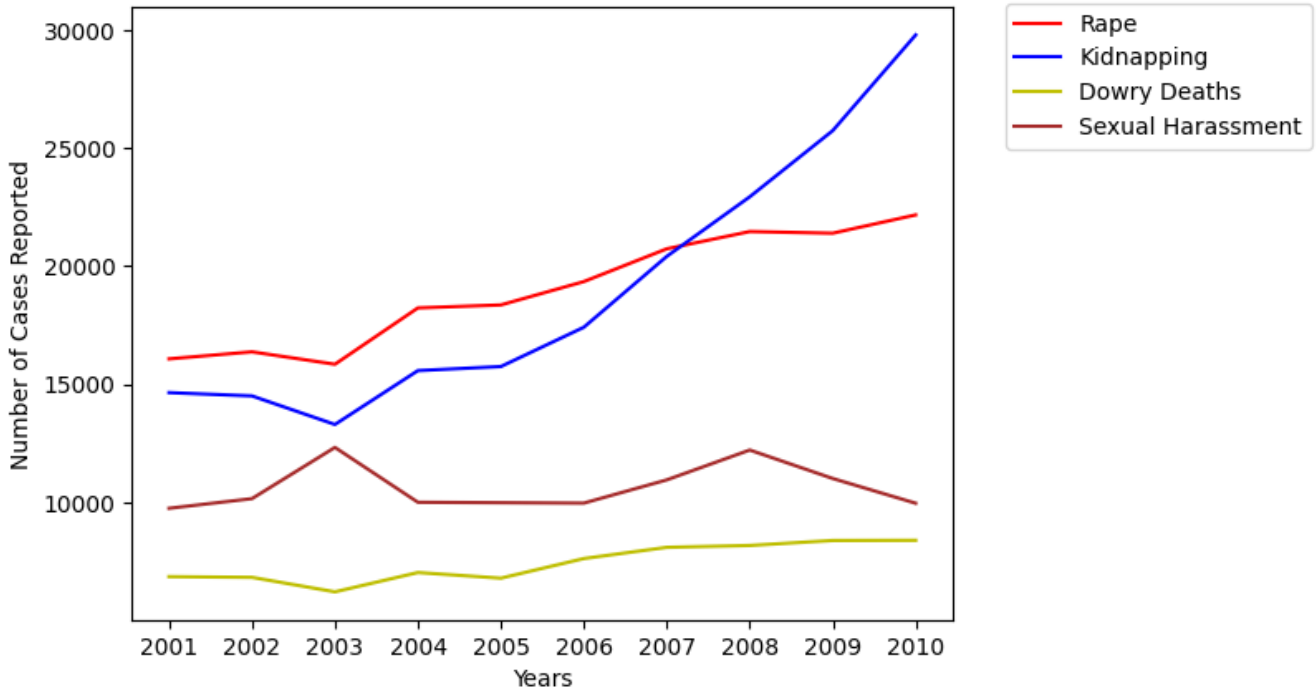
The scatter plot showcases the relationship between cases reported and cases sent for trial, with varying markers and colors indicating different data points, emphasizing their distribution and correlation.

OUTLIER ANALYSIS OF TOTAL CASES FOR TRIAL



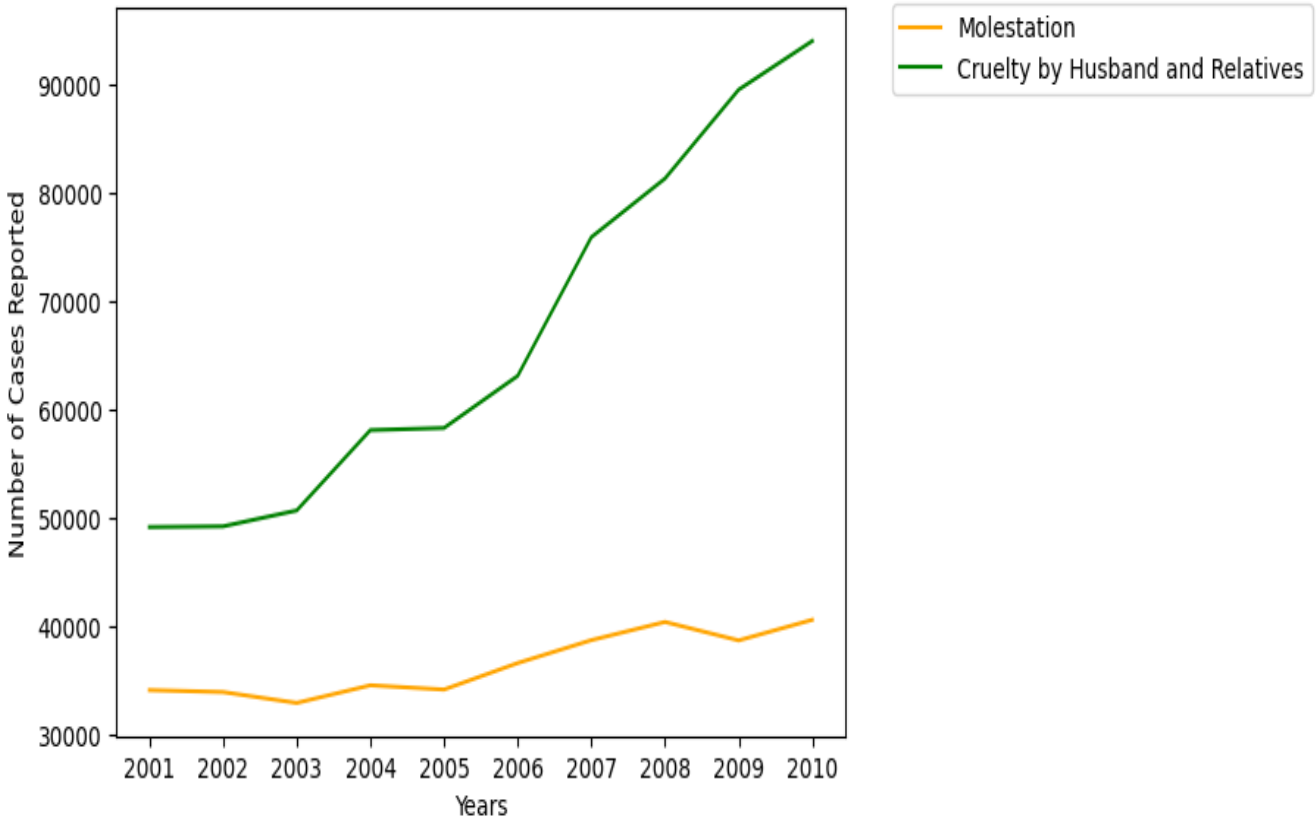
The box plot visualizes the distribution of total cases for trial, highlighting any outliers and providing insights into the spread and central tendency of the data. Since the total cases for trial are outliers, we have to drop those set of rows which are a cumulation of all the cases for trial.

TRENDS IN REPORTED CASES OF SPECIFIC CRIMES OVER 10 YEARS



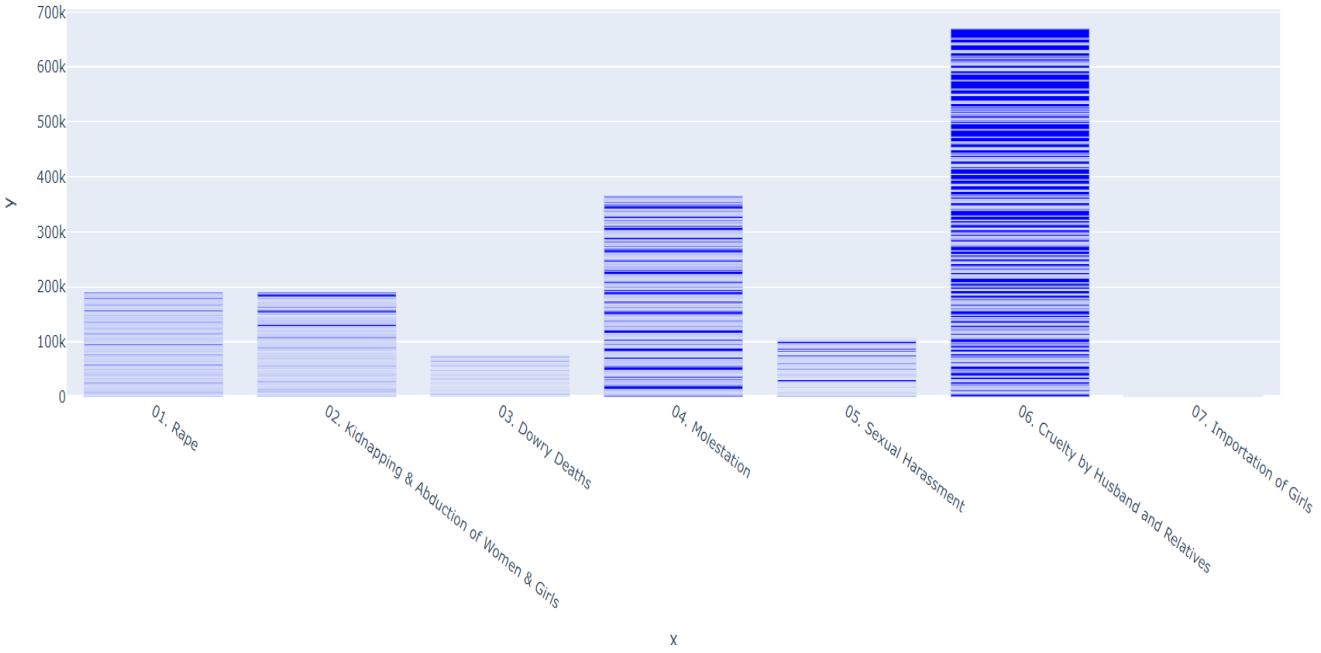
The line graph compares the annual cases reported for different crimes like rape, kidnapping, dowry deaths, sexual harassment, and cruelty, showcasing trends over the years.

TRENDS IN REPORTED CASES OF MOLESTATION AND CRUELTY BY HUSBAND AND RELATIVES OVER 10 YEARS



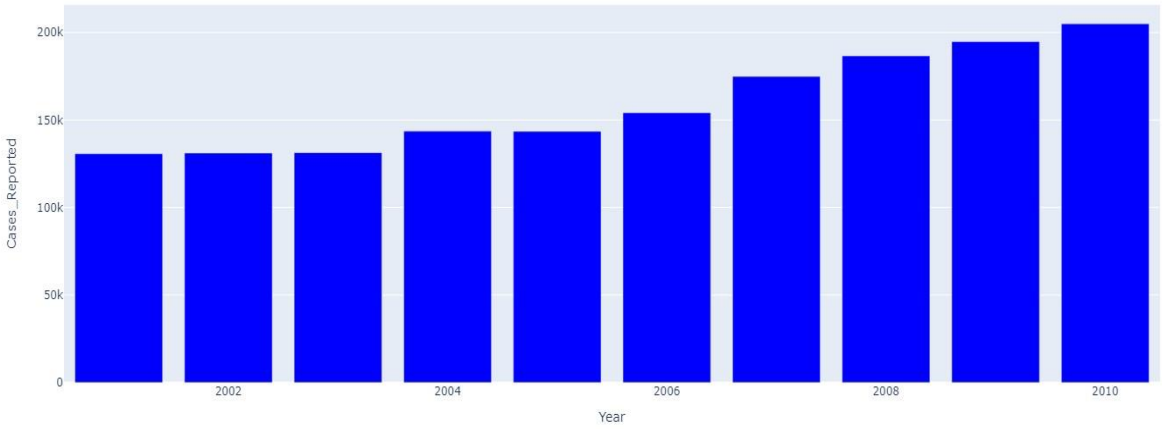
The graph displays trends in reported cases of molestation and cruelty by husband and relatives over ten years, highlighting variations and patterns in their occurrence annually.

CASES REPORTED FOR EACH CRIME SUBGROUP



The Plot bar chart illustrates cases reported for different crime subgroups, with distinct colours indicating variations in frequency, offering a clear visual comparison of their occurrences.

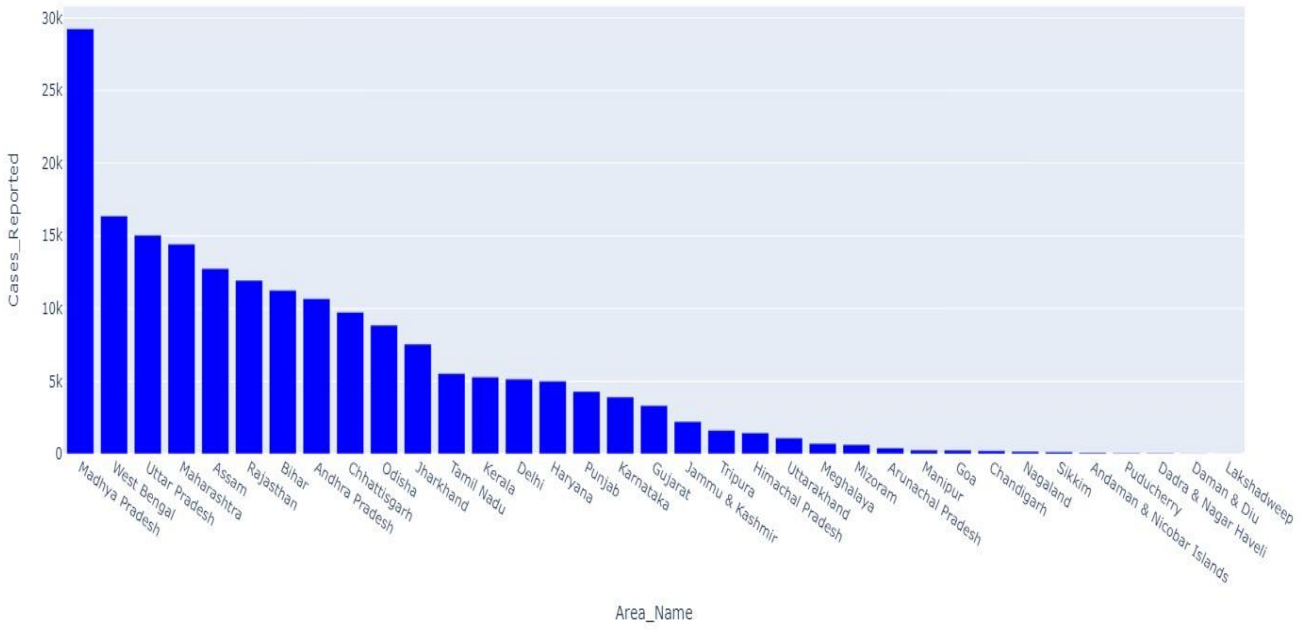
ANNUAL DISTRIBUTION OF REPORTED CASES



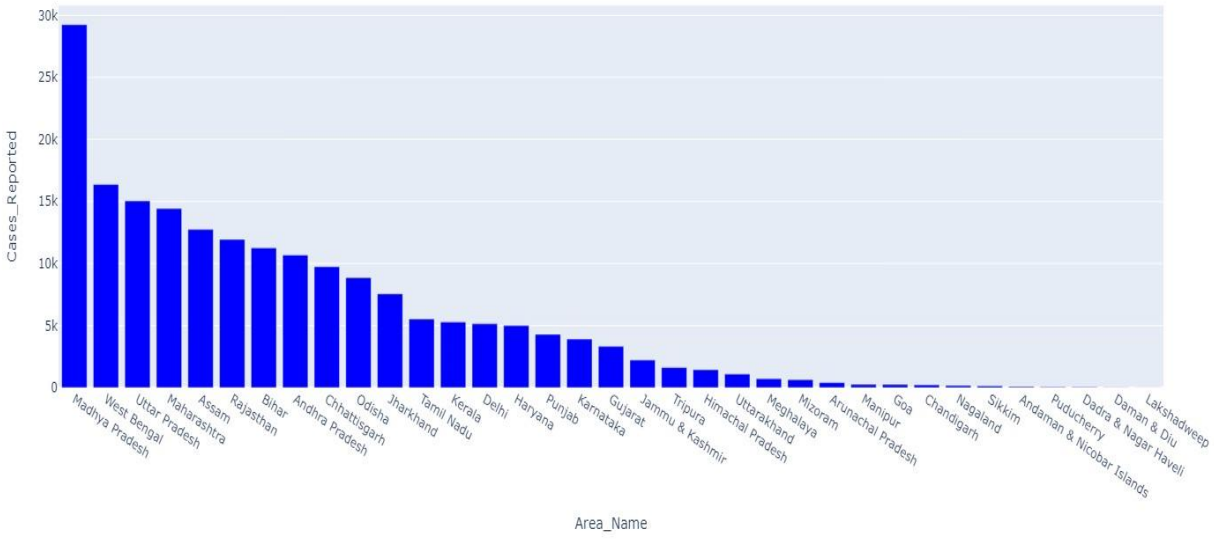
The bar chart visualizes the annual distribution of reported cases, with blue hues representing variations across years, offering insights into trends and fluctuations over time.

REGIONAL DISTRIBUTION OF REPORTED CASES FOR EACH CRIME SUBGROUP

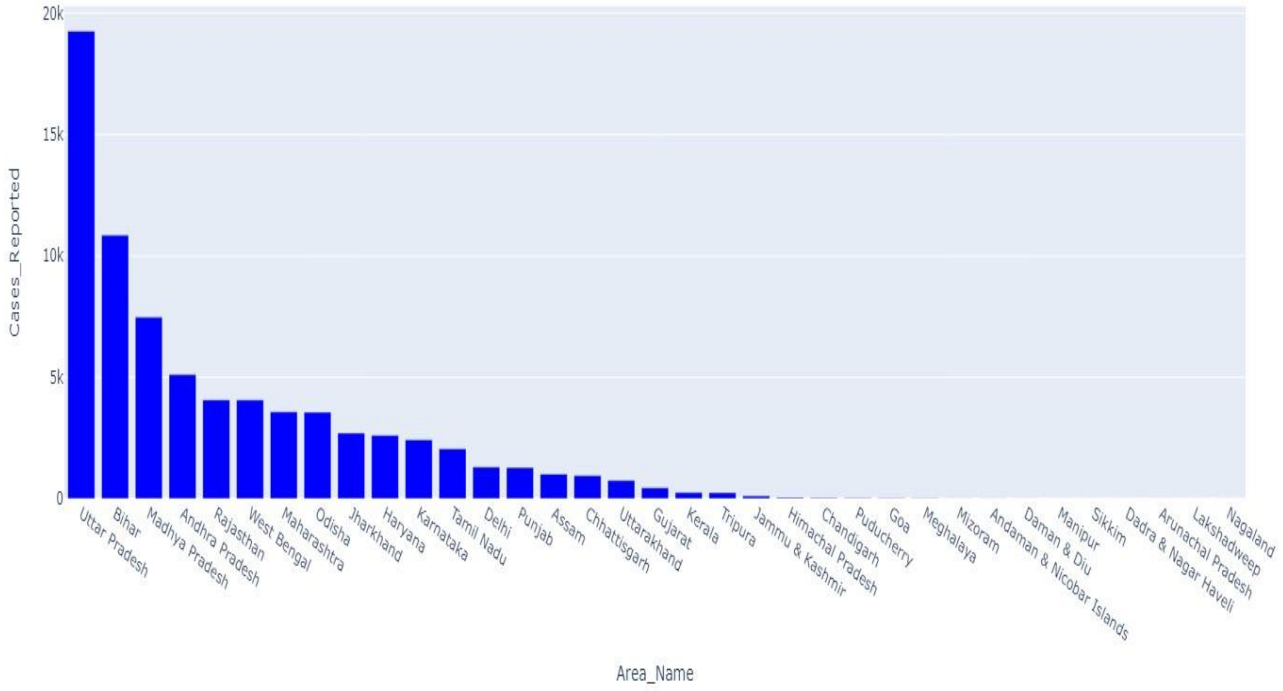
01. RAPE



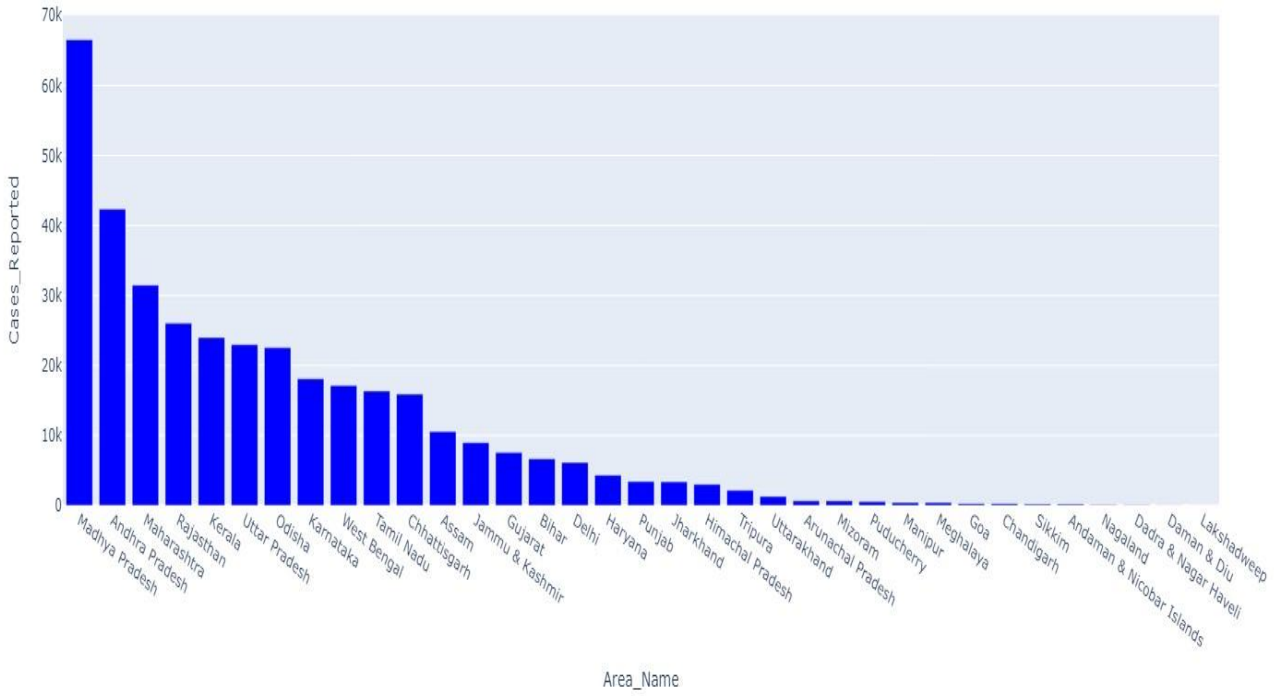
02. KIDNAPPING & ABDUCTION OF WOMEN & GIRLS



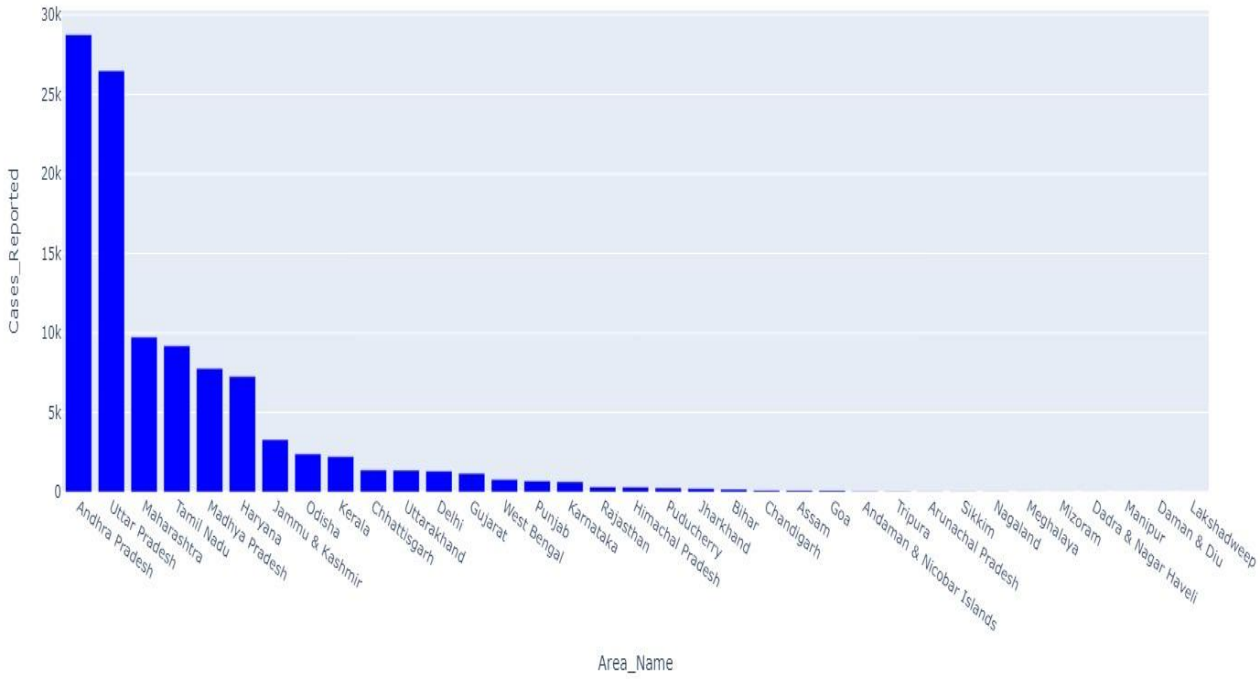
03. DOWRY DEATHS



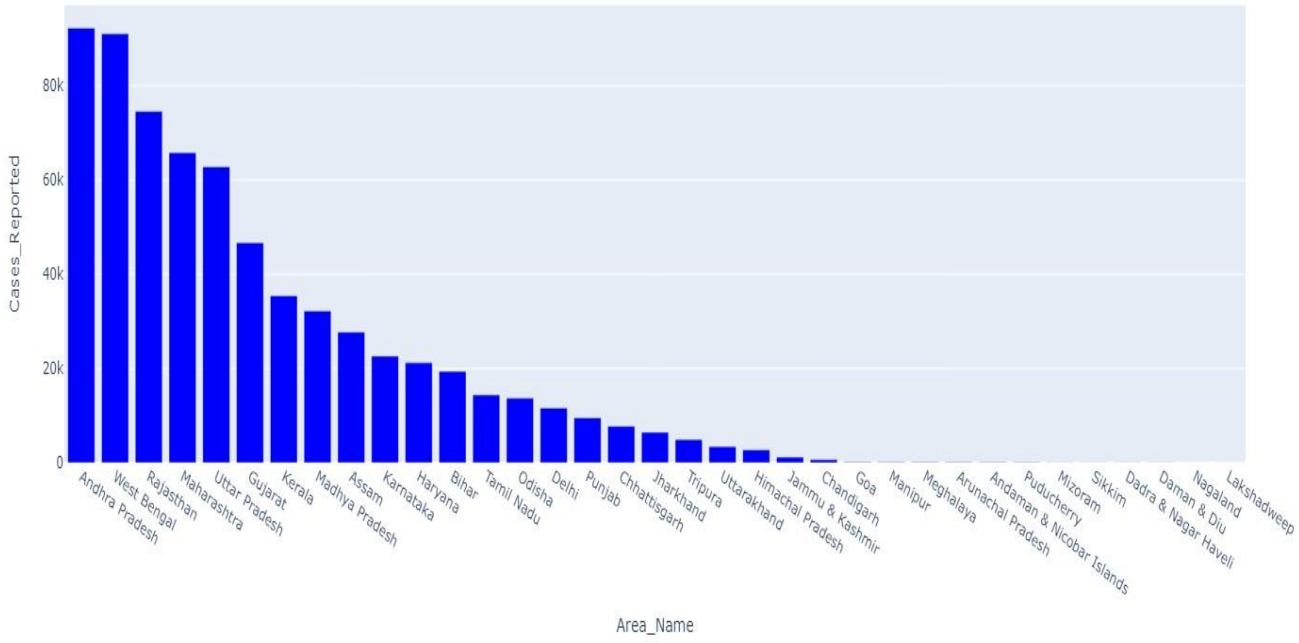
04. MOLESTATION



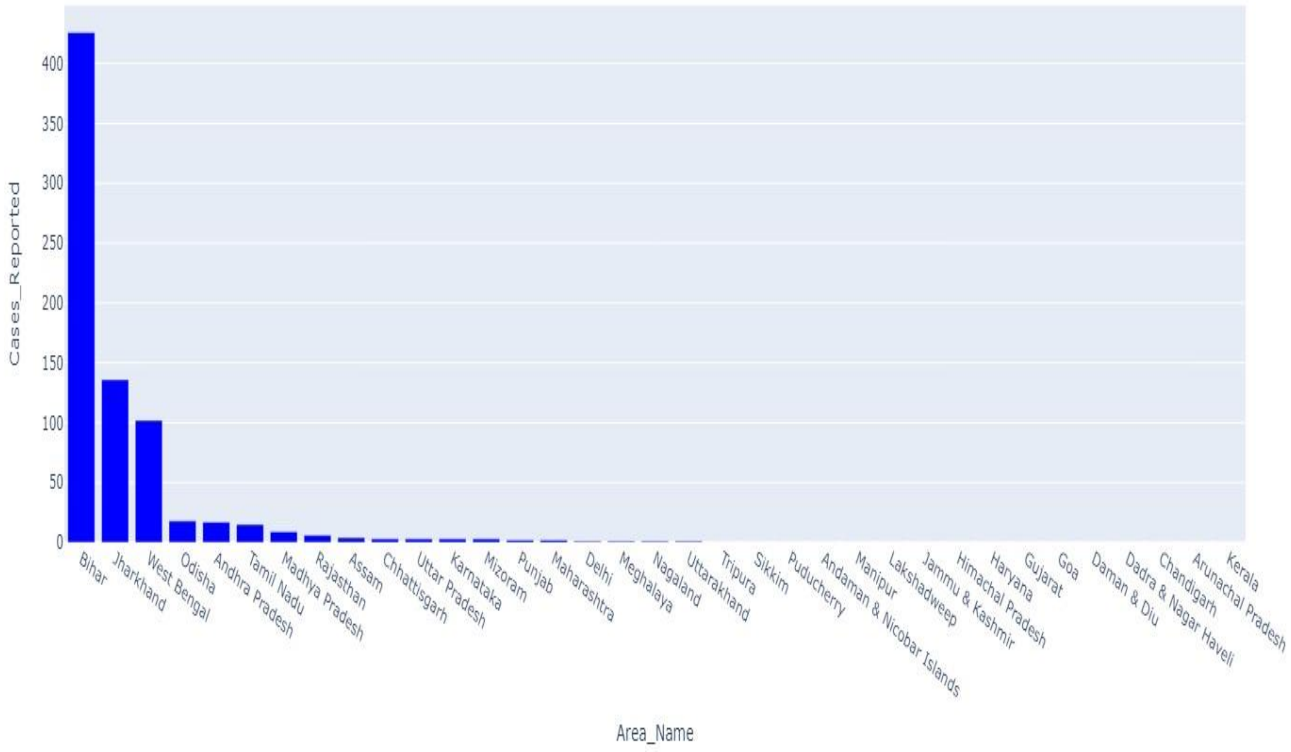
05. SEXUAL HARASSMENT



06. CRUELTY BY HUSBAND AND RELATIVES



07. IMPORTATION OF GIRLS



These above seven graphs represent different crime subgroups, aggregates reported cases by area, and displays them in bar charts. Each chart represents the distribution of reported cases across areas for a specific crime subgroup, allowing comparison of regional prevalence. The blue bars indicate varying levels of reported cases, providing insights into geographical patterns of crime incidence for each subgroup.

CONCLUSION & FUTURE ENHANCEMENTS

CONCLUSION:

The analysis of crime against women in India using exploratory data analysis (EDA) techniques has provided valuable insights into the complex dynamics of gender-based violence. Through meticulous data collection, preprocessing, and analysis, significant patterns and trends have been uncovered, shedding light on the spatial, temporal, and socio-economic factors influencing crime rates. Visualization techniques have facilitated a deeper understanding of the data, allowing stakeholders to identify hotspots, trends, and disparities across different regions and demographics. Furthermore, feature selection has highlighted key determinants of gender-based violence, informing evidence-based interventions and policy reforms.

FUTURE ENHANCEMENTS:

- **Incorporating more diverse datasets:** Expand the analysis to include a wider range of datasets, such as socio-economic indicators, cultural data, and victim/survivor demographics, to provide a more comprehensive understanding of gender-based violence.
- **Advanced predictive modeling:** Explore advanced machine learning algorithms and predictive modeling techniques to improve the accuracy of forecasts and identify early warning signs of gender-based violence hotspots.
- **Longitudinal analysis:** Conduct longitudinal analysis to track changes in crime rates over time, enabling the identification of emerging trends and the evaluation of the effectiveness of interventions.

- **Incorporating qualitative data:** Integrate qualitative data through surveys, interviews, or case studies to capture nuanced aspects of gender-based violence experiences and perceptions, enriching the analysis with qualitative insights.
- **Stakeholder engagement:** Collaborate with stakeholders, including government agencies, NGOs, and community organizations, to ensure the analysis addresses their needs and priorities and facilitates the co-creation of effective interventions and support services.
- **Continuous monitoring and evaluation:** Establish a framework for continuous monitoring and evaluation to track the implementation and impact of interventions over time, enabling iterative improvements and adaptive responses to evolving challenges.

Individual Contribution of the Team

Worklet Tasks	*Contributor's Names*
Dataset Collection	PHOOBESH S
Preprocessing	HRISHIKESH
Architecture Diagram	PHOOBESH S
Model building (suitable algorithm)	HRISHIKESH
Results – Tables, Graphs	PHOOBESH S, HRISHIKESH
Technical Report writing	PHOOBESH S
Presentation preparation	HRISHIKESH

GITHUB LINK OF THE PROJECT

https://github.com/PHOOBESH/PHOOBESH/blob/main/EDA%20-%20ANALYSIS%20OF%20CRIME%20AGAINST%20WOMEN%20IN%20INDIA.ipynb

REFERENCES

https://link.springer.com/chapter/10.1007/978-3-030-78750-9_13
https://www.ijitee.org/wp-content/uploads/papers/v8i6s4/F12970486S419.pdf
https://courses.ischool.berkeley.edu/i247/s14/reports/CrimesAgainstWomen_Rathi_Heiser_Bhosle.pdf
https://vc.bridgew.edu/jiws/vol22/iss5/1/
https://www.jetir.org/papers/JETIR2106104.pdf
https://www.kaggle.com/code/sugandhkhobragade/eda-of-murders-in-india