**EXPLORATORY DATA ANALYSIS**

**CSE3040**

**WINTER SEMESTER 2024-25**

**J Component REPORT**

A close-up of a logo

Description automatically generated

**Team Name:** Group Y

**Team Members:**

DHINESH KUMAR M – 23MIA1102

**Guided By:**

Dr ASNATH VICTY PHAMILA Y

**An Exploratory Analysis of Crimes Against Women in India**

**Acknowledgement**

Anyone is welcome to use the codes and content provided in this repository for their own learning, research, or projects. Please feel free to clone the repository to your local machine and adapt the materials as needed. However, kindly refrain from directly forking or modifying this repository on GitHub. For access, updates, or collaboration, please use the following link: <https://github.com/Dhinesh-Fedor/EDA_Project.git>.

Thank you for respecting these guidelines and contributing to an open, collaborative learning environment.

**ABSTRACT**

In today’s world, ensuring safety and equality for women is essential for building inclusive and just societies. This project examines the patterns and drivers of crimes against women in India, aligned with **Sustainable Development Goal 5 (SDG 5): Gender Equality**, which calls for the elimination of all forms of violence against women and girls. Utilizing a detailed dataset covering various crimes—such as rape, domestic violence, assault, and dowry deaths—across Indian states, the analysis explores the interrelations among these crimes and identifies the factors most influencing their occurrence. The methodology includes **data cleaning, outlier detection, univariate and multivariate exploration, linear regression, feature selection, and K-Means clustering** to uncover meaningful patterns and groupings. Additionally, cluster labelling was performed to assign descriptive categories to each state based on their crime profiles, enabling clearer regional comparisons and targeted analysis. Moving beyond surface-level statistics, this project highlights regions facing the greatest challenges, identifies the most significant contributors to violence, and provides actionable insights to inform targeted interventions and promote gender equality.

**OBJECTIVES**

The primary goal of this project is to identify key patterns, trends, and anomalies in crimes against women across India, using data-driven analysis to support policy decisions and interventions. The project employs a range of advanced analytical techniques to reveal significant insights into the prevalence of crimes such as rape, domestic violence, assault, dowry deaths, and more, ultimately aiming to reduce violence against women, in line with Sustainable Development Goal 5 (SDG 5): Gender Equality.

**1. Outlier Detection and Removal:** The first objective is to clean the dataset by detecting and removing outliers. This ensures that any extreme data points, which may distort analysis or mask underlying trends, are addressed. By doing so, the analysis identifies states or crime categories that deviate from national norms, spotlighting regions that may require special attention or intervention.

**2. Univariate and Multivariate Analysis:** Through both univariate and multivariate analysis, the project investigates the relationships between different types of crimes, including rape, domestic violence, assault, and dowry deaths. The use of statistical methods and visualizations helps to uncover patterns in the data and understand how various crime types relate to one another, providing a deeper understanding of the underlying factors influencing crime rates.

**3. Feature Selection and Regression Analysis:** Feature selection techniques are applied to identify the most significant variables influencing crimes against women. Linear regression is employed to quantify how different factors—such as assault, dowry deaths, and domestic violence—affect the occurrence of rape cases. This step ensures that the subsequent analyses are based on the most impactful and relevant variables.

**4. Clustering Analysis:** K-Means clustering is used to group Indian states based on crime patterns and prevalence. By categorizing states into clusters, the project identifies regional trends and similarities in crimes against women. This clustering helps to understand how different regions face unique challenges, enabling the identification of areas in need of focused, region-specific interventions. Additionally, cluster labelling was performed to assign descriptive categories to each state based on their crime profiles.

**5. Data-Driven Insights for Policy Intervention:** Ultimately, the project seeks to make sense of the complex dynamics of crimes against women in a way that is both technical and socially meaningful. By providing actionable insights into which states and crime types require attention, the analysis supports informed decision-making and policy formulation aimed at addressing gender inequality, improving women's safety, and reducing violence against women across India.

**METHODOLOGY**

**1. Dataset Collection:** The project began by gathering a comprehensive dataset on crimes against women across various states in India. The data was sourced from Kaggle platform, ensuring it included multiple categories of crimes—such as rape, dowry deaths, domestic violence, and assault—along with key contextual information such as the year and state. This provided a solid foundation for exploring the dynamics of gender-based violence in India.

**2. Data Cleaning & Preprocessing:** After collecting the dataset, the next step involved cleaning and preprocessing the data. This process included removing unnecessary columns and renaming the remaining columns to more meaningful and understandable labels. Additionally, missing values were checked and handled appropriately, ensuring that the data was clean and consistent for the subsequent steps. This step was crucial to ensure that the analysis would yield reliable and accurate results.

**3. Exploratory Data Analysis (EDA):** With the data cleaned, exploratory data analysis (EDA) was conducted. Initially, univariate analysis was performed to examine the distributions of individual crime categories, using visualizations such as histograms and box plots. Bivariate analysis followed, where relationships between different crime categories were explored using scatterplots and correlation matrices. This step revealed patterns and potential interrelationships between crime types, providing insights into how they may be interconnected.

**4. Outlier Detection & Removal:** To ensure the integrity of the analysis, outlier detection method—Interquartile Range (IQR) was employed. This technique helped identify extreme data points that could distort the results. Upon detection, each outlier was carefully evaluated to determine whether it was an error or a valid anomaly. Depending on the assessment, the outliers were either removed or adjusted to maintain the robustness of the dataset.

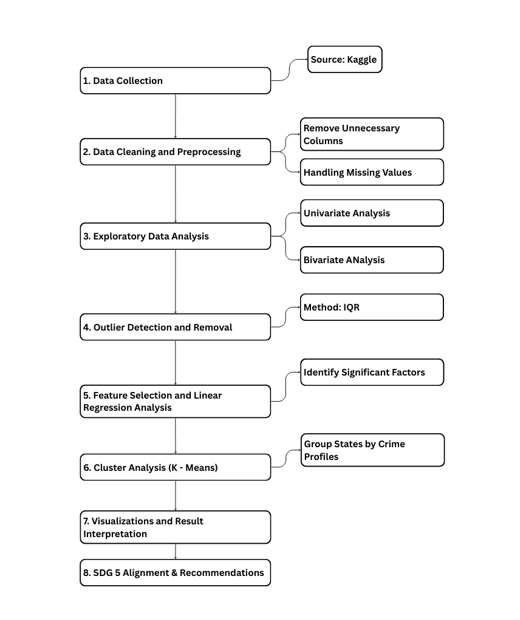
**5. Feature Selection & Linear Regression Analysis:** Feature selection was performed to identify the most significant factors influencing crimes against women. This process helped isolate key variables such as domestic violence and dowry deaths. Linear regression analysis was then conducted to quantify the relationships between these selected factors and the occurrence of rape cases. This step provided insights into which variables were the primary drivers of gender-based violence in India, ensuring that subsequent analyses were based on the most impactful variables.

**6. Clustering Analysis (K-Means):** K-Means clustering was applied to group Indian states based on their crime profiles. Relevant features, such as rape cases, dowry deaths, and domestic violence, were selected for this analysis. The Elbow Method was used to determine the optimal number of clusters, which revealed states with similar crime patterns. Clustering helped identify geographical regions facing shared challenges, providing deeper insights into regional variations and enabling the design of targeted interventions. Additionally, cluster labelling was performed to assign descriptive categories to each state based on their crime profiles.

**7. Visualization & Result Interpretation:** The findings from the analysis were presented through a series of visualizations, including charts, graphs, and plots. These visual representations helped convey the results clearly and concisely, making it easier to interpret the data. The visualizations highlighted key patterns, trends, and areas of concern, offering stakeholders a straightforward way to understand the most significant findings.

**8. SDG 5 Alignment & Recommendations:** The results were aligned with Sustainable Development Goal 5 (SDG 5), which focuses on gender equality. The analysis assessed how the identified patterns related to the goal of reducing violence against women and promoting gender equality. Based on the insights obtained, targeted recommendations for policy interventions were developed, focusing on high-risk states and providing actionable steps for improving safety and justice for women across India.

**ARCHITECTURE**



**1. Collecting the Data:** The process began with sourcing data from Kaggle, ensuring access to comprehensive and reliable information on crimes against women across Indian states. This provided a solid foundation for all subsequent analysis.

**2. Cleaning and Preparing the Data:** Next, the dataset was cleaned and pre-processed to ensure quality and consistency. This involved removing unnecessary columns, addressing missing values, and standardizing the data. These steps were essential to eliminate errors and prepare the dataset for meaningful analysis.

**3. Exploring the Data (EDA):** Exploratory Data Analysis was conducted to gain initial insights. Univariate analysis helped in understanding the distribution of individual crime types, while bivariate analysis revealed relationships between different features. Visualizations such as charts and plots were used to uncover trends and patterns.

**4. Finding Outliers:** Outlier detection was performed using the Interquartile Range (IQR) method. This step identified data points that deviated significantly from the norm, allowing for their removal or further examination to ensure robust and unbiased results.

**5. Identifying Key Factors:** Feature selection and linear regression analysis were applied to pinpoint the most significant variables influencing crimes against women. This helped focus the analysis on the factors that matter most for understanding and addressing the issue.

**6. Grouping States by Crime Profiles:** Cluster analysis using K-Means was employed to group states based on their crime profiles. This clustering revealed patterns and similarities among states, aiding in the identification of regions with shared challenges and characteristics. Additionally, cluster labelling was performed to assign descriptive categories to each state based on their crime profiles.

**7. Visualizing and Interpreting Results:** The findings were presented through clear visualizations and concise interpretations. These visuals made it easier to communicate key insights and support data-driven conclusions.

**8. Aligning with SDG 5 and Recommendations:** Finally, the results were linked to Sustainable Development Goal 5, which focuses on achieving gender equality and eliminating violence against women. Based on the analysis, actionable recommendations were proposed to support policy-making and targeted interventions toward safer, more equitable societies.

**DATASET OVERVIEW**

**Link:** <https://www.kaggle.com/datasets/balajivaraprasad/crimes-against-women-in-india-2001-2021>.

The dataset provides a comprehensive, multi-year record of crimes against women across Indian states and union territories, enabling in-depth analysis of crime patterns, trends, and risk factors at both state and national levels.

**Key Features of the Dataset**

**1. Extensive Temporal and Geographic Coverage:** The dataset spans two decades (2001–2021) and includes all major Indian states and union territories, offering a longitudinal and pan-India perspective on crimes against women.

**2. Diverse Crime Categories:**  
It records annual counts for several major crime types:

* + *Rape (Rape)*
  + *Kidnapping & Abduction (K&A)*
  + *Dowry Deaths (DD)*
  + *Assault on Women (AoW)*
  + *Assault on Modesty (AoM)*
  + *Domestic Violence (DV)*
  + *Witchcraft-related Crimes (WT)*  
    This granularity allows for detailed analysis of specific crime types and their evolution over time.

**3. State/UT and Yearly Breakdown:** Data is organized by state/UT and year, enabling cross-sectional and time-series comparisons to identify regional disparities and temporal trends.

**4. Comparability and Policy Relevance:** The structure of the dataset supports direct comparison of crime rates and patterns across states and years, facilitating benchmarking and identification of high-risk regions or periods.

**5. Alignment with SDG 5:** The dataset directly supports analysis related to Sustainable Development Goal 5 (SDG 5): Gender Equality, particularly the target of eliminating all forms of violence against women and girls. By providing granular, reliable data on various forms of violence, it enables evidence-based assessment of progress and policy needs.

**Why SDG 5 Was Chosen for This Dataset**

**1. Direct Link to Gender-Based Violence:** The dataset’s focus on crimes against women—such as rape, dowry deaths, and domestic violence—aligns precisely with SDG 5’s mandate to eliminate violence against women and girls in all spheres of life.

**2. Insight into Risk Factors and Trends:** By tracking multiple crime types over time and across regions, the dataset allows for the identification of persistent risk factors, emerging threats, and the effectiveness of interventions—key for SDG 5 monitoring and action.

**3. Regional and Temporal Comparisons:** The ability to compare states and years helps highlight where progress is being made and where challenges remain, supporting targeted, data-driven interventions and resource allocation.

**4. Foundation for Evidence-Based Recommendations:** This dataset enables rigorous analysis to inform policies and programs aimed at reducing violence against women, supporting India’s and global efforts to achieve gender equality and ensure women’s safety.

**RESULTS**

**Outlier Detection and Capping**

To ensure data quality and minimize the influence of extreme values, outlier detection and capping were performed on all numerical features in the dataset using the Interquartile Range (IQR) method.

**Process Overview**

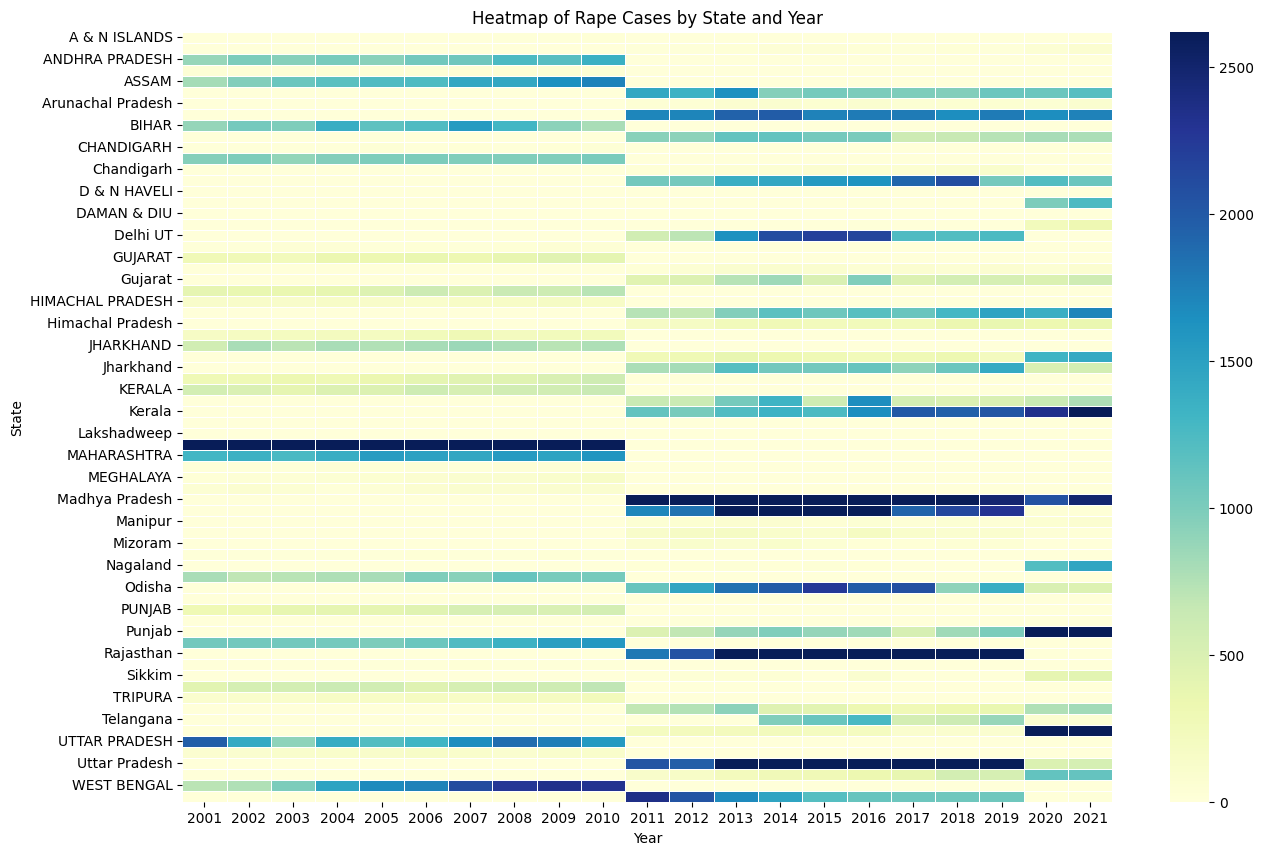
* For each numerical column, the distribution and boxplot were visualized to assess the presence and extent of outliers.
* The IQR was calculated, and lower and upper bounds were established at 1.5 times the IQR below the first quartile (Q1) and above the third quartile (Q3), respectively.
* Data points falling outside these bounds were identified as outliers, with the total count of outliers reported for each feature.
* Outliers were then capped: values below the lower bound were set to the lower bound, and values above the upper bound were set to the upper bound, preserving the majority of the data’s structure while reducing the impact of extreme values.

**Visualization**

* For each numerical feature, four key plots were generated:
  1. Histogram with KDE: Displayed the original distribution and highlighted the presence of outliers.
  2. Boxplot: Provided a visual summary of the spread and outliers for each feature.
  3. Scatter Plot (Original): Showed the raw data points, with outliers visible as points far from the main cluster.
  4. Scatter Plot (Cleaned): Illustrated the effect of capping, with extreme values brought within the defined bounds.

**Univariate and Bivariate Analysis**

**Heatmap of Rape Cases by State and Year**

****

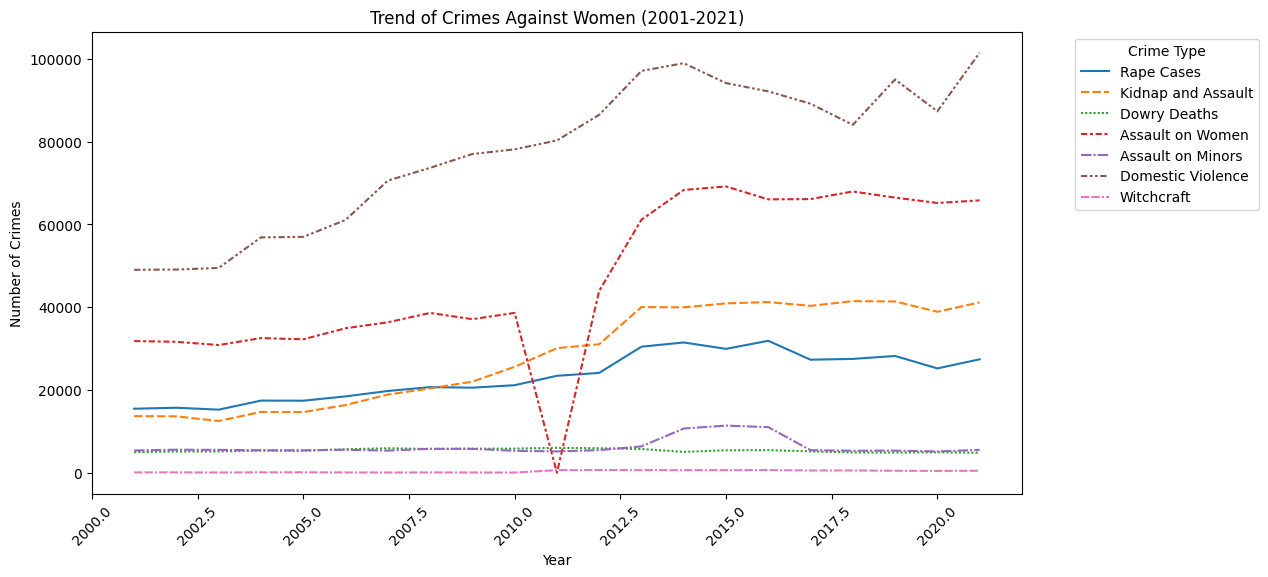
**Visualization Details**

* **X-axis:** Year (2001–2021)
* **Y-axis:** State/Union Territory
* **Colour Intensity:** Number of reported rape cases (darker shades indicate higher counts; lighter shades indicate lower counts)
* **Colour Bar:** Provides a scale for interpreting the number of cases, with the darkest blue representing the highest values (over 2,500 cases)

**Inference**

* The heatmap reveals significant variation in the number of reported rape cases across both states and years.
* States such as Maharashtra, Uttar Pradesh, Madhya Pradesh, Rajasthan, and West Bengal consistently show darker bands, indicating a higher number of reported cases throughout the observed period.
* There is a visible increase in reported cases in many states after 2012, which may correspond to heightened awareness, changes in reporting practices, or actual increases in incidence following high-profile national events and legal reforms.
* Some states and union territories, such as Lakshadweep, Sikkim, and the northeastern states, report consistently lower numbers, as shown by lighter shades.
* The temporal and regional patterns highlighted by this heatmap can guide targeted interventions, resource allocation, and policy focus on regions with persistently high or rising incidence of rape cases.

**Trend Analysis of Crimes Against Women (2001–2021)**

****

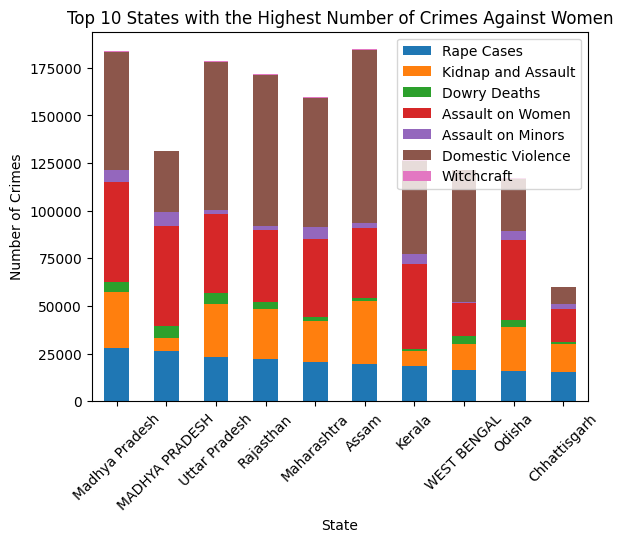
**Visualization Details**

* **X-axis:** Year (2001–2021)
* **Y-axis:** Number of Crimes
* **Lines:** Each line represents a different crime type:
  + Rape Cases
  + Kidnap and Assault
  + Dowry Deaths
  + Assault on Women
  + Assault on Minors
  + Domestic Violence
  + Witchcraft
* **Legend:** Located on the right, mapping line colour and style to each crime type.

**Inference**

* **Domestic Violence** is the most prevalent crime type, showing a steady and significant increase over the two decades, peaking above 100,000 reported cases by 2021.
* **Assault on Women** and **Kidnap and Assault** also demonstrate a rising trend, with a noticeable surge after 2012, possibly reflecting increased reporting and awareness following major national incidents and legal reforms.
* **Rape Cases** have steadily increased, with a marked rise post-2012, aligning with broader national attention and policy changes.
* **Dowry Deaths** and **Witchcraft-related crimes** remain relatively lower in absolute numbers and show less pronounced trends, though dowry deaths are consistently present.
* There is a sharp dip around 2011 across all crime types, likely due to data reporting anomalies or missing data for that year.
* **Assault on Minors** shows a moderate increase, indicating growing awareness or incidence over time.

**Top 10 States with the Highest Number of Crimes Against Women**

****

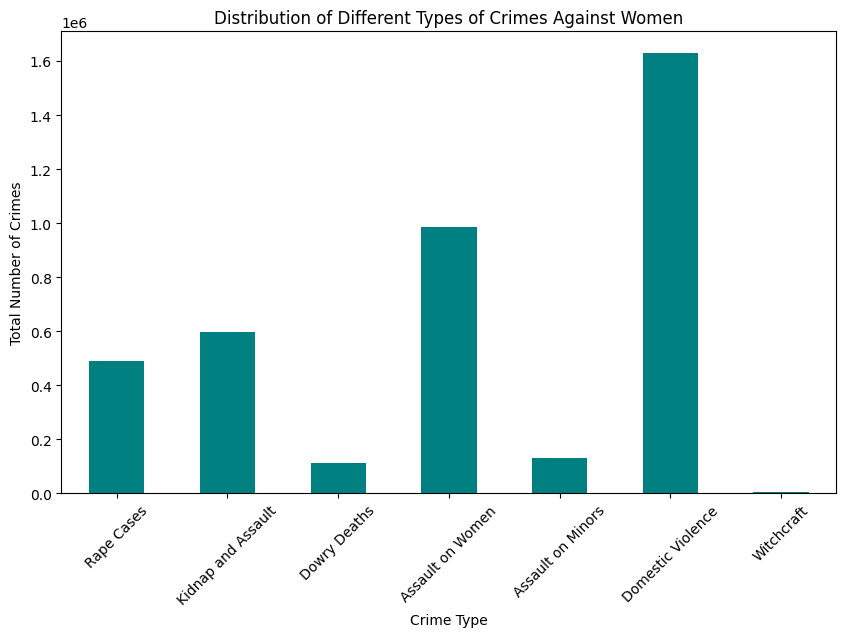
**Visualization Details**

* **X-axis:** State (Top 10 states by total crime count)
* **Y-axis:** Number of Crimes
* **Bars:** Stacked to show the contribution of each crime type within the total for each state:
  + Rape Cases
  + Kidnap and Assault
  + Dowry Deaths
  + Assault on Women
  + Assault on Minors
  + Domestic Violence
  + Witchcraft
* **Legend:** Maps each colour to a specific crime type for clear interpretation.

**Inference**

* **Madhya Pradesh, Maharashtra, and Assam** lead in the total number of crimes against women, each exceeding 175,000 reported cases.
* **Domestic Violence** is the dominant crime type across all top states, forming the largest portion of each bar.
* **Assault on Women** and **Kidnap and Assault** are also significant contributors to the overall crime burden in these states.
* States like **Uttar Pradesh, Rajasthan, and West Bengal** also report high levels of multiple crime types, indicating a broad spectrum of challenges related to women’s safety.
* **Witchcraft-related crimes** are present but constitute a very small fraction of the total, visible as the thin topmost segment in some bars.
* The stacked bar format highlights not only which states have the highest overall crime rates but also the relative prevalence of different crime types within each state, providing insights for targeted intervention and policy focus.

**Distribution of Different Types of Crimes Against Women**

****

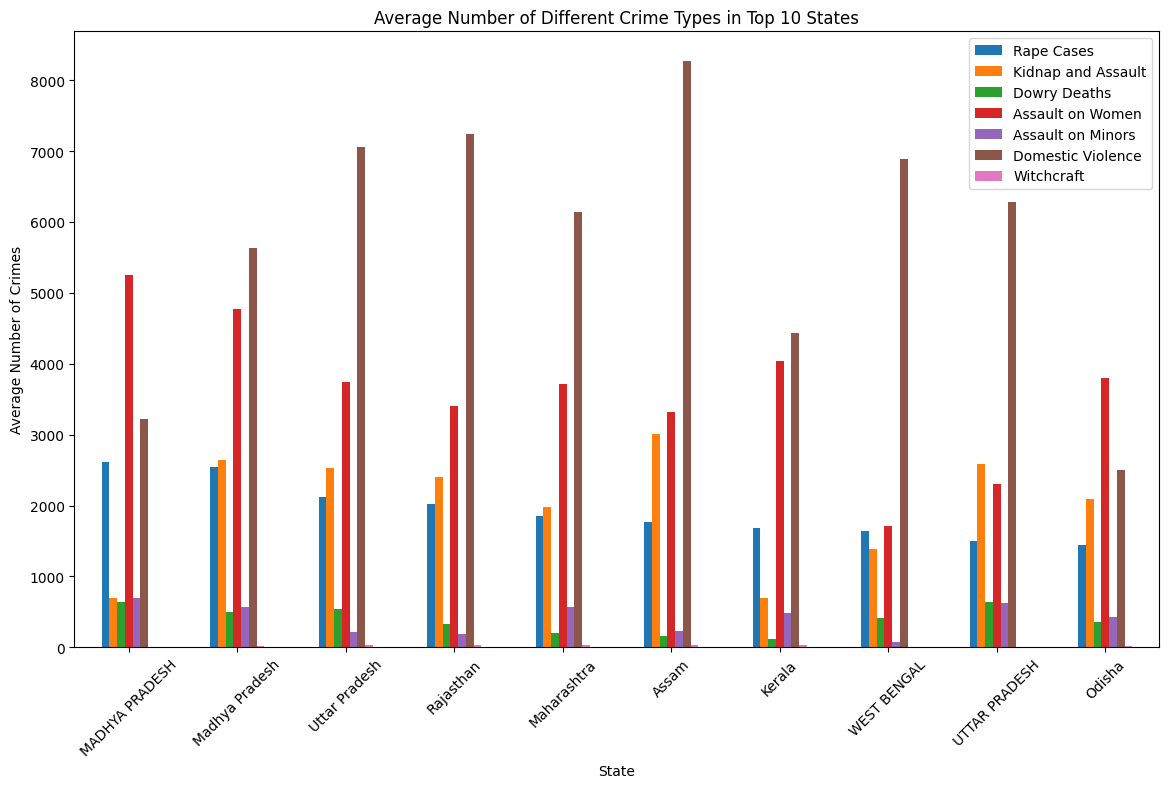
**Visualization Details**

* **X-axis:** Crime Type (Rape Cases, Kidnap and Assault, Dowry Deaths, Assault on Women, Assault on Minors, Domestic Violence, Witchcraft)
* **Y-axis:** Total Number of Crimes (aggregated across all years and states)
* **Bar Heights:** Represent the cumulative count for each crime category

**Inference**

* **Domestic Violence** is by far the most prevalent crime type, with a total count exceeding 1.6 million cases, highlighting it as the most significant challenge among crimes against women.
* **Assault on Women** and **Kidnap and Assault** are also highly reported, with totals close to 1 million and 600,000 cases, respectively.
* **Rape Cases** account for over 500,000 incidents, underscoring the severity and frequency of this crime nationally.
* **Dowry Deaths** and **Assault on Minors** have lower but still substantial totals, each contributing over 100,000 cases to the overall burden.
* **Witchcraft-related crimes** are the least reported, with negligible numbers compared to other categories.
* The bar chart clearly demonstrates that while all these crimes are serious, domestic violence and assault on women are the most widespread, indicating priority areas for intervention and policy action.

**Average Number of Different Crime Types in Top 10 States**

****

**Visualization Details**

* **X-axis:** State (Top 10 states with the highest overall crime counts)
* **Y-axis:** Average Number of Crimes (per year)
* **Bars:** Multiple bars for each state, each representing a different crime type:
  + Rape Cases
  + Kidnap and Assault
  + Dowry Deaths
  + Assault on Women
  + Assault on Minors
  + Domestic Violence
  + Witchcraft
* **Legend:** Located in the upper right, mapping each colour to a specific crime type.

**Inference**

* **Domestic Violence** consistently registers the highest average annual cases across all top 10 states, with Assam exhibiting the highest average (over 8,000 cases per year).
* **Assault on Women** is the second most prevalent crime type in most states, followed by **Kidnap and Assault** and **Rape Cases**.
* **Dowry Deaths**, **Assault on Minors**, and **Witchcraft-related crimes** show much lower averages, with Witchcraft being almost negligible in most states.
* The pattern is consistent across states, with Domestic Violence and Assault on Women dominating the crime landscape, indicating persistent and widespread challenges in these areas.
* This grouped bar chart enables a direct comparison of the average burden of each crime type across the most affected states, highlighting priority areas for targeted interventions and resource allocation.

**Linear Regression Results: Predicting Rape Cases**

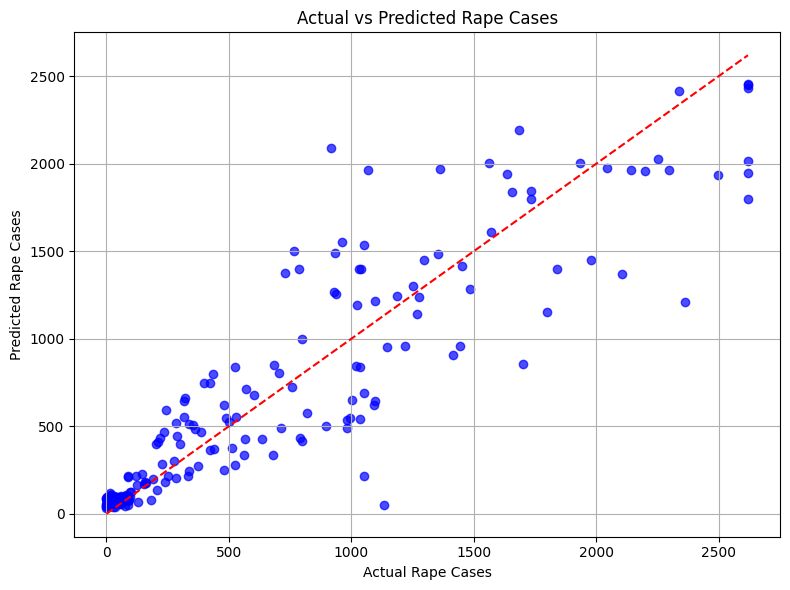
**Model Inputs**

* **Features Used:** Year, Kidnap and Assault, Dowry Deaths, Assault on Women, Assault on Minors, Domestic Violence, Witchcraft
* **Target Variable:** Rape Cases

**Performance Metrics**

* **Mean Absolute Error (MAE):** 185.17
* **R² Score:** 0.83

**Visualization 1: Actual vs Predicted Rape Cases**

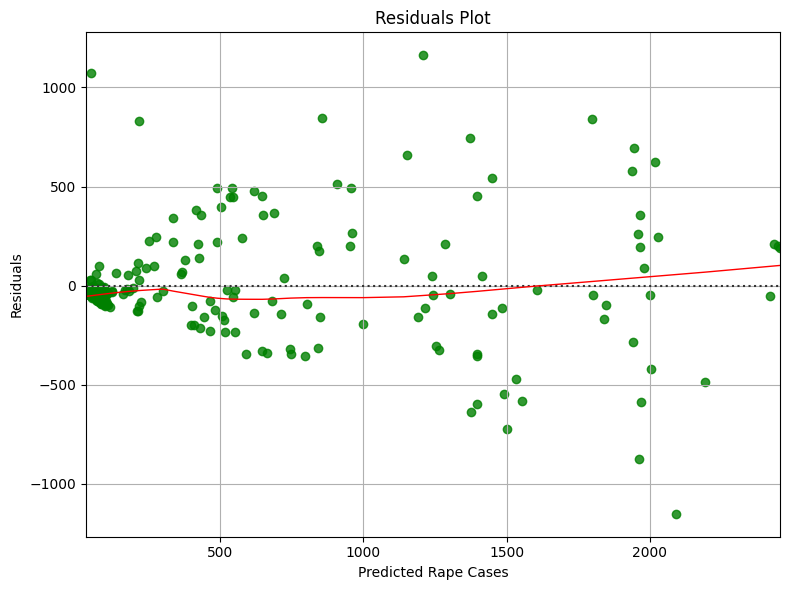
****

* **X-axis:** Actual Rape Cases (test set)
* **Y-axis:** Predicted Rape Cases
* **Blue Dots:** Each represents a test data point.
* **Red Dashed Line:** Represents perfect prediction (where predicted equals actual).

**Inference**

* Most points cluster close to the red dashed line, indicating strong predictive performance.
* The model successfully captures the general trend in rape case numbers, with some variance at higher values.
* The high R² score (0.83) suggests that the model explains a substantial portion of the variance in rape case counts based on the selected features.

**Visualization 2: Residuals Plot**

****

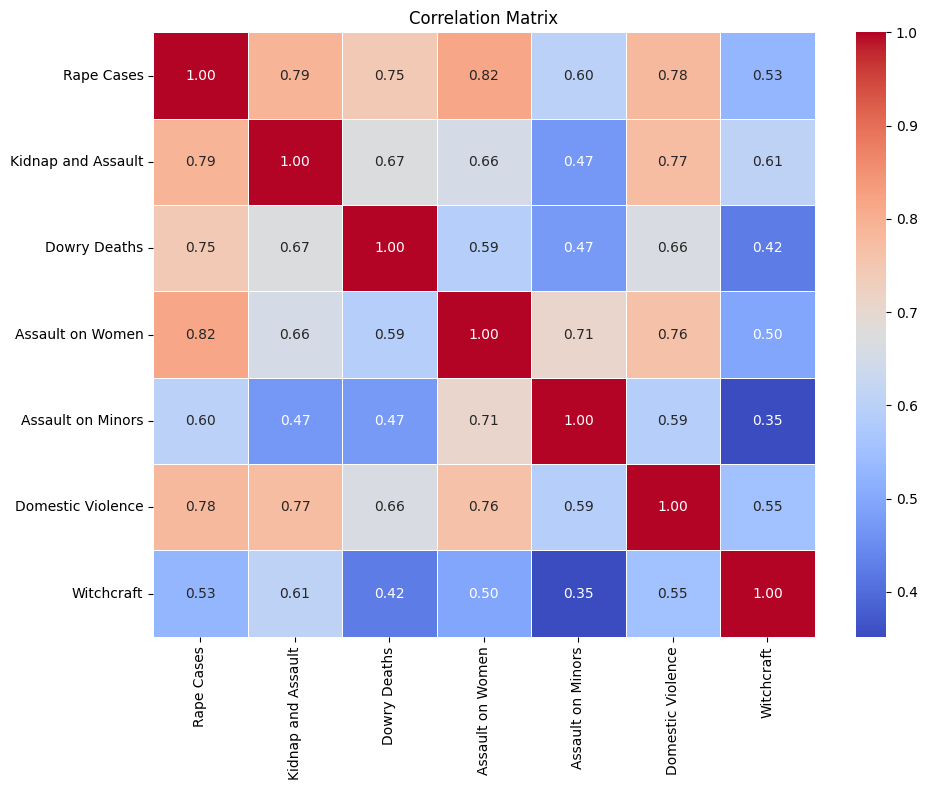
* **X-axis:** Predicted Rape Cases
* **Y-axis:** Residuals (difference between actual and predicted values)
* **Green Dots:** Each represents a residual for a test data point.
* **Red Lowess Line:** Shows the trend in residuals.

**Inference**

* Residuals are mostly cantered around zero, indicating that the model does not systematically over- or under-predict.
* There is some spread and a few larger residuals at higher predicted values, suggesting occasional under- or overestimation for extreme cases.
* The absence of a strong pattern in the residuals supports the validity of the linear regression model.

**Feature Selection**

To enhance the accuracy and relevance of clustering and further analysis, a feature selection process was conducted. This involved evaluating each crime category based on its statistical characteristics—such as variance—and its correlation with other features. The aim was to retain variables that hold analytical value while excluding those with limited contribution.



**1. Rape Cases**

* **Standard Deviation:** 977.02 (High variance)
* **Highly Correlated With:** Assault on Women (0.80), Kidnap and Assault (0.70)
* **Decision:** **Retained**
* **Rationale:** Despite correlation with other crimes, rape is a critical and socially significant crime category with substantial variability across states and years. It provides strong discriminatory power in clustering.

**2. Kidnap and Assault**

* **Standard Deviation:** 1993.54 (Very high variance)
* **Highly Correlated With:** Rape (0.70), Domestic Violence (0.68)
* **Decision:** **Retained**
* **Rationale:** Although correlated with other features, it captures a distinct dimension of crime. Its high variance makes it influential in cluster formation.

**3. Dowry Deaths**

* **Standard Deviation:** 424.92 (Moderate variance)
* **Highly Correlated With:** Kidnap and Assault (0.69)
* **Decision:** **Retained**
* **Rationale:** Dowry-related crimes remain a pressing social issue in India. Despite moderate variance, its relevance to SDG 5 justifies its inclusion.

**4. Assault on Women**

* **Standard Deviation:** 2463.96 (Very high variance)
* **Highly Correlated With:** Rape (0.80)
* **Decision:** **Retained**
* **Rationale:** Represents a broad spectrum of gender-based violence. The high standard deviation indicates significant variation, supporting its role in pattern recognition.

**5. Assault on Minors**

* **Standard Deviation:** 806.02 (Moderate variance)
* **Correlation:** Lower correlation with other features
* **Decision:** **Retained**
* **Rationale:** While less correlated, crimes involving minors are critical for social insight and may highlight unique patterns within clusters.

**6. Domestic Violence**

* **Standard Deviation:** 4042.00 (Very high variance)
* **Highly Correlated With:** Rape, Kidnap and Assault, Assault on Women (0.68 – 0.76)
* **Decision:** **Retained**
* **Rationale:** As one of the most prevalent and variable crimes, it is essential for understanding regional patterns and underlying sociocultural factors.

**7. Witchcraft**

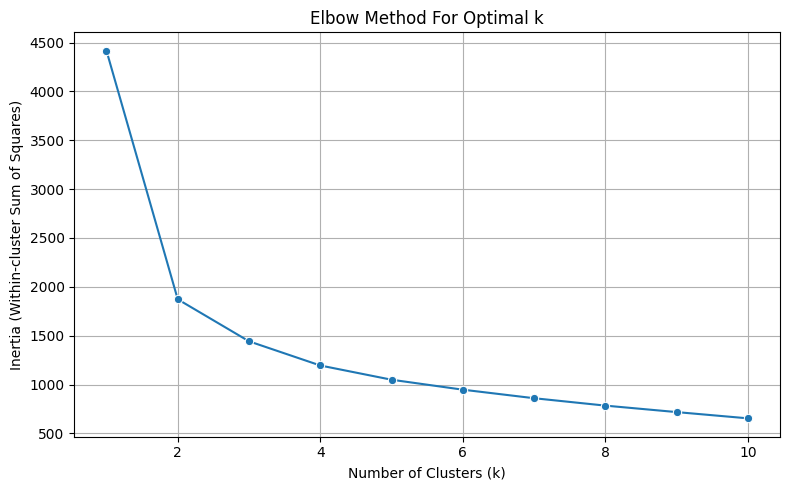
* **Standard Deviation:** 79.99 (Low variance)
* **Correlation:** Weak correlation with other features
* **Decision:** **Dropped**
* **Rationale:** The low variance and weak correlations indicate limited influence on clustering outcomes. It was excluded to streamline the feature set.

**Clustering (K-Means)**

To uncover hidden patterns in crimes against women across Indian regions and support targeted policy interventions aligned with SDG 5 (Gender Equality), a **KMeans clustering** approach was employed. The clustering focused on multiple forms of violence, aiming to segment regions based on crime intensity and type. This helps in identifying **critical zones**, **emerging hotspots**, and **relatively safer areas** that can guide data-driven resource allocation and awareness initiatives.

**Optimal Cluster Determination**

The Elbow Method was used to determine the ideal number of clusters by examining how the Within-Cluster Sum of Squares (WCSS) decreased with increasing *k*. A noticeable elbow was observed at k = 3, suggesting that the dataset naturally groups into three distinct categories.

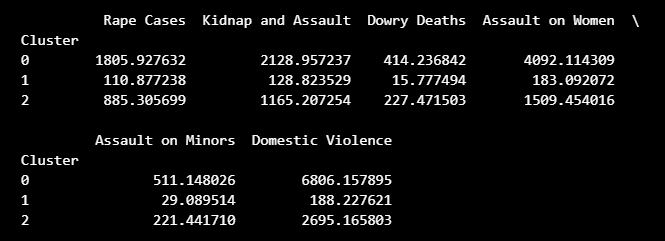


**Cluster Profiles**

Clustering was performed on six critical crime categories:

* Rape Cases
* Kidnap and Assault
* Dowry Deaths
* Assault on Women
* Assault on Minors
* Domestic Violence

**The clusters formed revealed meaningful socio-criminal groupings**

****

**Cluster 0 – High-Crime Hotspots:** These regions report alarmingly high levels of all crimes analysed. They demand urgent multi-pronged intervention, such as better policing, fast-track courts, support systems for survivors, and stricter enforcement of laws.

**Cluster 2 – Moderate-Crime Zones**: This group includes regions where crime rates are moderate but non-negligible, pointing to a need for preventive strategies, community engagement, and early intervention to prevent escalation.

**Cluster 1 – Low-Incidence Zones**: These areas reflect significantly lower crime rates, potentially indicating safer environments or underreporting. These findings could help benchmark effective practices in these regions, or call for further investigation into reporting mechanisms.

**Visual Exploration**

A pair plot visualization was created to show how clusters differ across combinations of crime types. This multidimensional perspective helps visualize overlaps, density, and separability between different regional profiles.



This clustering approach allows policymakers and gender rights organizations to prioritize action based on data-driven insights. By identifying regional patterns of violence, this analysis supports the larger goal of promoting gender equality and ensuring safety for women and girls, directly contributing to the achievement of SDG 5.

**Cluster Label Assignment and Export**

The code assigned descriptive labels to each cluster generated from the clustering analysis, mapping numeric cluster identifiers to human-readable categories:

* **0 → High Crime**
* **1 → Low Crime**
* **2 → Moderate Crime**

A new column, **'Crime Cluster Label'**, was added to the Data Frame to reflect these categories. The resulting Data Frame, containing the year, state, numeric cluster, and descriptive cluster label, was exported as a CSV file (states\_with\_clusters.csv).

****

**SDG 5 ALIGNMENT AND RECOMMENDATIONS**

**SDG 5** aims to “achieve gender equality and empower all women and girls,” with a strong emphasis on eliminating all forms of violence against women and girls in public and private spheres. The data-driven analysis of crimes against women across Indian states directly supports the targets and indicators outlined under SDG 5, particularly:

* **Target 5.2:** Eliminate all forms of violence against all women and girls, including trafficking, sexual and other types of exploitation.
* **Target 5.3:** Eliminate harmful practices such as child, early and forced marriage, and dowry deaths.
* **Target 5.1:** End all forms of discrimination against women and girls everywhere.

**Key Insights from Data Analysis**

**1. Crime Patterns and Regional Disparities**

* States with persistently high rates of rape, dowry deaths, and domestic violence (e.g., Uttar Pradesh, Madhya Pradesh, Rajasthan, West Bengal) highlight regions where women face heightened risks and systemic barriers to safety and justice.
* States with lower reported crime rates may either reflect effective interventions or, in some cases, underreporting due to social stigma or lack of access to justice.

**Alignment Insight:** Identifying these regional disparities enables targeted policy interventions and resource allocation, supporting SDG 5’s mandate to eliminate violence and discrimination at all levels.

**2. Outlier Detection Reveals Hidden Vulnerabilities**

* Outlier analysis (using IQR) uncovers states and years with exceptionally high or low crime rates, flagging periods or locations that require urgent investigation or intervention.
* For example, sudden spikes in dowry deaths or domestic violence can signal breakdowns in social protection, legal enforcement, or public awareness.

**Alignment Insight:** Spotting these anomalies helps direct SDG 5 efforts to areas with acute vulnerabilities, ensuring no region is left behind in the pursuit of gender equality.

**3. Feature Selection and Multivariate Analysis Inform Policy**

* Linear regression and feature selection highlight key drivers of violence, such as the correlation between high rates of domestic violence and increased incidents of rape or dowry deaths.
* These relationships provide empirical evidence for prioritizing interventions that address root causes, such as patriarchal norms, lack of education, or economic dependency.

**Alignment Insight:** Evidence-based identification of risk factors supports the design of holistic, multi-sectoral programs—recognized as effective in addressing violence against women and girls.

**4. Clustering Supports Targeted Interventions**

* K-Means clustering groups states with similar crime profiles, revealing clusters with shared challenges (e.g., high rates of both domestic violence and dowry deaths).
* This approach allows for the development of region-specific strategies and sharing of best practices between states facing similar issues.

**Alignment Insight:** Cluster-based insights enable more nuanced, context-sensitive policy responses, in line with SDG 5’s focus on inclusivity and local adaptation.

**5. Supporting Evidence-Based Policy and Resource Allocation**

* The combination of EDA, outlier detection, and clustering provides a robust empirical foundation for monitoring progress, evaluating interventions, and justifying the allocation of resources.
* This approach enhances transparency and accountability in SDG 5 implementation.

**Alignment Insight:** Data-driven monitoring and evaluation are essential for scaling up successful interventions and ensuring sustained progress towards eliminating violence against women.

**Recommendations and Area of Focus**

| **Area of Focus** | **Recommendations** |
| --- | --- |
| High crime rate states | Implement intensive awareness campaigns,  strengthen legal enforcement, and expand  support services. |
| Regions with spikes/outliers | Conduct targeted investigations and rapid-  response interventions to address  underlying causes. |
| States with high domestic violence & dowry deaths | Develop integrated programs addressing  social norms, economic empowerment,  and legal literacy. |
| Underreporting/  low crime states | Improve reporting mechanisms, community  outreach, and victim support to ensure  accurate data. |
| Multi-sectoral coordination | Foster collaboration between police,  judiciary, health, education, and  community organizations. |
| Monitoring and evaluation | Establish regular data audits and  transparent reporting to track SDG 5  progress at state and district levels. |

***“By linking real-world crime data with SDG 5, this project provides actionable, data-driven insights that empower governments, NGOs, and global agencies to design effective policies, allocate resources efficiently, and accelerate progress toward gender equality and the elimination of violence against women and girls.”***

**FUTURE SCOPE**

Building on the insights gained from this analysis of crimes against women in India, several future directions can further enhance the depth and impact of the research:

**1. Deep-Dive into Subcategories of Crime**  
Future studies can focus on more granular subcategories within each crime type (such as age group, urban vs. rural incidence, or relationship to perpetrator) to uncover nuanced patterns and risk factors that may be masked in aggregate data.

**2. Policy Impact Assessment**  
By integrating timelines of major legal reforms, awareness campaigns, or government initiatives, future analyses can evaluate the effectiveness of specific policy interventions in reducing crime rates or shifting reporting patterns.

**3. Sentiment and Media Analysis**  
Incorporating data from news reports, social media, and public sentiment surveys can provide valuable context on societal attitudes, underreporting, and the influence of media coverage on crime trends and reporting behaviour.

**4. Predictive Hotspot Mapping**  
Utilizing advanced spatial analytics and machine learning, future work can develop predictive models to identify potential future hotspots for crimes against women, enabling pre-emptive action from authorities and support organizations.

**5. Cross-State Comparative Studies**  
Comparing the effectiveness of different state-level policies, law enforcement practices, and community programs can highlight best practices and inform the design of more effective interventions in high-risk regions.

**6. Victim Support and Rehabilitation Analysis**  
Expanding the scope to include data on victim support services, rehabilitation outcomes, and survivor reintegration can provide a more holistic view of the response ecosystem and highlight gaps in support structures.

**7. Collaboration with Stakeholders**  
Engaging with NGOs, law enforcement, healthcare providers, and community leaders in future research can facilitate richer data collection, ground-truthing of findings, and the co-creation of actionable solutions.

**8. Integration of Real-Time Reporting Tools**  
Exploring the use of mobile applications or online platforms for real-time crime reporting and resource mapping can improve data timeliness and empower communities to participate in prevention and response efforts.

**CONCLUSION**

This project systematically explored the landscape of crimes against women in India, guided by the framework of Sustainable Development Goal 5 (SDG 5): Gender Equality. Utilizing a comprehensive dataset spanning two decades and multiple crime categories—including rape, domestic violence, assault, dowry deaths, and more—the analysis combined rigorous data cleaning, outlier detection, univariate and multivariate exploration, feature selection, linear regression, and K-Means clustering with descriptive cluster labelling.

The results revealed persistent regional disparities and rising trends in crimes against women, with certain states consistently reporting higher incidences and distinct crime profiles. Feature selection and regression analysis identified key factors most strongly associated with the prevalence of specific crimes, while clustering and cluster labelling provided a clear framework for comparing states and targeting interventions.

By transforming raw data into actionable insights, the project not only highlights areas of urgent concern but also demonstrates the value of data-driven approaches for policy-making and resource allocation. The methodology and findings support evidence-based strategies to address violence against women, while the open dataset and code offer a foundation for further research and innovation.

Looking ahead, future work can extend this analysis through time-series forecasting, integration of socio-economic variables, geospatial mapping, real-time dashboards, and community-driven reporting tools. These enhancements will further empower stakeholders to monitor progress, design targeted interventions, and accelerate India’s journey toward achieving gender equality and ending violence against women.

**REFERENCES**

1. Varaprasad, B. (2023). *Crimes Against Women in India (2001–2021)*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/balajivaraprasad/crimes-against-women-in-india-2001-2021>
2. United Nations Department of Economic and Social Affairs. (n.d.). *Goal 5: Achieve gender equality and empower all women and girls*. Retrieved from <https://sdgs.un.org/goals/goal5>.
3. United Nations. (n.d.). *Gender equality and women's empowerment – Goal 5*. Retrieved from <https://www.un.org/sustainabledevelopment/gender-equality/>