Exploratory Data Analysis on

CRIME AGAINST WOMEN

by

Group 4

Zeel Ghori
ID: 202201287
Course: BTech(ICT)

Aryankumar Panchasara
ID: 202201056
Course: BTech(ICT)

Nischay Agrawal
ID: 202411031
Course: MTech(ICT)

Course Code: IT 462
Semester: Autumn 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

December 2, 2024

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of my project titled “Crime Against Women” Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am extremely fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

Furthermore, I would like to extend my appreciation to the entire team at DAIICT for fostering an environment of collaboration and innovation. The resources and facilities provided have been crucial in conducting comprehensive research and analysis.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future endeavors.

Once again, thank you for your unwavering guidance and belief in my abilities. Your mentorship has been invaluable, and I am truly grateful for the opportunity to work with you.

Sincerely,
[Zeel Ghori, 202201287]
[Aryankumar Panchasara, 202201056]
[Nischay Agrawal, 202411031]

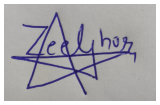
DECLARATION

We, 202201287, 202201056, 202411031 hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

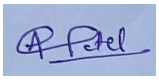
We acknowledge that the data used in this project is obtained from the data.gov.in site. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

We hereby sign the declaration statement and confirm the submission of this report on 3rd December, 2024.



Zeel Ghori
ID: 202201287
Course: BTech(ICT)



Aryankumar Panchasara
ID: 202201056
Course: BTech(ICT)

Nischay Agrawal
ID: 202411031
Course: MTech(ICT)

CERTIFICATE

This is to certify that Group 4 comprising Zeel Ghori, Aryankumar Panchasara, and Nischay Agrawal has successfully completed an exploratory data analysis (EDA) project on the Crime against Women, which was obtained from data.gov.in.

The EDA project presented by Group 4 is their original work and has been completed under the guidance of the course instructor, Prof. Gopinath Panda, who has provided support and guidance throughout the project. The project is based on a thorough analysis of the Crime against Women dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the Crime against Women, which demonstrates the analytical skills and knowledge of the students of Group 4 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

December 2, 2024

Contents

List of Figures 5

1 Introduction 1
- 1.1 Your Project idea . . . 1
- 1.2 Data Collection . . . 1
- 1.3 Dataset Description . . . 1
- 1.4 Packages required . . . 2

2 Data Cleaning 4
- 2.1 Missing data analysis . . . 4
  - 2.1.1 Missing Data Heatmap . . . 4
- 2.2 Imputation . . . 5
- 2.3 Data Cleaning Steps Performed . . . 5
  - 2.3.1 Ensured Years are Numeric . . . 5
  - 2.3.2 Checked for Duplicates . . . 5
  - 2.3.3 Corrected Inconsistent Formatting in 'STATE/UT' Column . . . 5

3 Visualization 6
- 3.1 Univariate analysis . . . 6
  - 3.1.1 Histograms, Violin Plots, and Boxplots . . . 6
  - 3.1.2 Barplot for Crime Count by Crime Head . . . 8
  - 3.1.3 Histograms for Each Crime Head . . . 8
  - 3.1.4 Crime Distribution Across States for Each Year . . . 10
  - 3.1.5 Violin Plots for Each Crime Head . . . 12
- 3.2 Bivariate Analysis . . . 13
  - 3.2.1 Iteration Through Crime Heads and Plot Creation . . . 13
  - 3.2.2 Heatmap for Crime Count by State/UT and Year . . . 14
  - 3.2.3 Pairplot of Year Columns . . . 15
  - 3.2.4 Correlation Matrix for Each Year . . . 15
  - 3.2.5 Crimes from 2001 to 2012 by State/UT . . . 16
- 3.3 Multivariate Analysis . . . 17
  - 3.3.1 Total Crime Count per State Over the Years . . . 17
  - 3.3.2 Total Crime Count for Each Crime Head Over the Years . . . 18
  - 3.3.3 Stacked Crime Count by Crime Head Over the Years . . . 18
  - 3.3.4 Aggregate Data by Year and Crime Head . . . 19
  - 3.3.5 Aggregate Data by State/UT and Crime Head . . . 21

3.3.6 Crime Trends by State/UT 22

4 Feature Engineering **23**
4.1 Feature extraction 23
4.1.1 Table of Crime Count by State/UT and Crime Head 23
4.1.2 Total Crimes, Growth Rate (CAGR), and Average Yearly Crimes 24
4.2 Feature selection 25
4.2.1 Total Crimes by State/UT Plot 25
4.2.2 Plot of Total Crime Count by Year 26
4.2.3 Crime Distribution by Category 27
4.2.4 Year-wise Percentage Contribution of Each Crime Head 27
4.2.5 CRIME HEAD Distribution for Each Year 28
4.2.6 Top Crime Head for Each State 28
4.2.7 Number of Outliers 29
4.2.8 Capping the Outliers 29
4.2.9 Outlier Treatment and Visualization 29
4.2.10 Data Normalization 30

5 Model fitting **31**
5.1 Regression 31
5.1.1 Linear Extrapolation 31
5.1.2 Polynomial Extrapolation 41

6 Conclusion & future scope **49**
6.1 Findings/observations 49
6.2 Challenges 49
6.3 Future plan 50

List of Figures

2.1 Heatmap of Missing Data 5

3.1 Histogram for the year 2001 6
3.2 Histogram for the year 2002 6
3.3 Histogram for the year 2003 7
3.4 Histogram for the year 2004 7
3.5 Histogram for the year 2005 7
3.6 Histogram for the year 2006 7
3.7 Histogram for the year 2007 7
3.8 Histogram for the year 2008 7
3.9 Histogram for the year 2009 7
3.10 Histogram for the year 2010 7
3.11 Histogram for the year 2011 8
3.12 Histogram for the year 2012 8
3.13 Barplot for Crime Count by Crime Head 8
3.14 Histogram for RAPE 9
3.15 Histogram for KIDNAPPING and ABDUCTION 9
3.16 Histogram for DOWRY DEATH 9
3.17 Histogram for ASSAULT ON WOMEN 9
3.18 Histogram for INSULT TO THE MODESTY OF WOMEN 9
3.19 Histogram for CRUELTY BY HUSBAND OR RELATIVES 9
3.20 Histogram for IMMORAL TRAFFIC (PREVENTION) ACT 10
3.21 Histogram for INDECENT REPRESENTATION OF WOMEN (PREVENTION) ACT 10
3.22 Crime Distribution for 2001 10
3.23 Crime Distribution for 2002 10
3.24 Crime Distribution for 2003 10
3.25 Crime Distribution for 2004 10
3.26 Crime Distribution for 2005 11
3.27 Crime Distribution for 2006 11
3.28 Crime Distribution for 2007 11
3.29 Crime Distribution for 2008 11
3.30 Crime Distribution for 2009 11
3.31 Crime Distribution for 2010 11
3.32 Crime Distribution for 2011 12
3.33 Crime Distribution for 2012 12
3.34 Violin Plot for RAPE 12

3.35 Violin Plot for KIDNAPPING . . . 12
3.36 Violin Plot for DOWRY DEATH . . . 12
3.37 Violin Plot for ASSAULT ON WOMEN . . . 12
3.38 Violin Plot for INSULT TO MODESTY . . . 13
3.39 Violin Plot for CRUELTY BY HUSBAND . . . 13
3.40 Violin Plot for IMMORAL TRAFFIC . . . 13
3.41 Violin Plot for INDECENT REPRESENTATION . . . 13
3.42 Trends for Each Crime Head Over the Years . . . 14
3.43 Heatmap of Crime Count by State/UT and Year . . . 14
3.44 Pairplot of Year Columns . . . 15
3.45 Correlation Matrix for Each Year . . . 16
3.46 Crimes from 2001 to 2012 by State/UT . . . 17
3.47 Total Crime Count per State Over the Years . . . 18
3.48 Total Crime Count for Each Crime Head Over the Years . . . 18
3.49 Stacked Crime Count by Crime Head Over the Years . . . 19
3.50 Pie Chart for Crime Distribution by Year . . . 20
3.51 Pie Chart for Crime Distribution by State/UT . . . 21
3.52 Crime Trends by State/UT . . . 22

4.1 Total Crimes by State/UT . . . 26
4.2 Total Crime Count by Year . . . 26
4.3 Crime Distribution by Category . . . 27
4.4 Year-wise Percentage Contribution of Each Crime Head . . . 28
4.5 CRIME HEAD Distribution for Each Year . . . 28
4.6 Top Crime Head for Each State/UT . . . 29
4.7 Boxplot Before Outlier Treatment . . . 29
4.8 Boxplot After Outlier Treatment . . . 29
4.9 Effect of Normalization on Data (Before v/s After) . . . 30

5.1 Extrapolated Crime Trends (2013-2017) . . . 32
5.2 Extrapolated Crime Trends (2013-2017) (Normalized) . . . 33
5.3 State-wise Predicted Crime Rates (2013-2017) . . . 34
5.4 State-wise Predicted Crime Rates (2013-2017) (Normalized) . . . 34
5.5 Extrapolated Crime Head Trends (2013-2017) . . . 35
5.6 RMSE for Each Crime Head . . . 36
5.7 R-squared score for Each Crime Head . . . 36
5.8 Residuals of Linear Regression Extrapolation for Each Crime Head . . . 37
5.9 Assault on women with intent to outrage her modesty . . . 37
5.10 Cruelty by husband or relatives . . . 38
5.11 Dowry death . . . 38
5.12 Immoral Traffic (prevention) act . . . 39
5.13 Indecent representation of women (prevention) act. . . . 39
5.14 Insult to the Modesty of women . . . 40
5.15 Kidnapping and Abduction . . . 40
5.16 Rape . . . 41
5.17 Polynomial Extrapolation of Crime Trends (2013-2017) . . . 42
5.18 R-Squared for Polynomial Regression (2001-2012) . . . 42

5.19 Assault on women with intent to outrage her modesty (Polynomial Regression) 43
5.20 Cruelty by husband or relatives (Polynomial Regression) 43
5.21 Dowry death (Polynomial Regression) 44
5.22 Immoral Traffic (prevention) act (Polynomial Regression) 44
5.23 Indecent representation of women (prevention) act. (Polynomial Regression) 45
5.24 Insult to the Modesty of women (Polynomial Regression) 45
5.25 Kidnapping and Abduction (Polynomial Regression) 46
5.26 Rape (Polynomial Regression) 46
5.27 Comparison of Linear and Polynomial Fits 47

List of Tables

2.1 Missing Values Analysis 4
4.1 Crime Count by State/UT and Crime Head 24

Abstract

This project explores crime trends in India from 2001 to 2012 using comprehensive crime statistics across various states and Union Territories. Through rigorous data preprocessing, including cleaning and imputation, the dataset was prepared for analysis. Visualization techniques, such as univariate and multivariate analysis, were employed to uncover patterns, anomalies, and relationships in the data. The findings from this analysis provide actionable insights for policymakers, law enforcement, and researchers. These insights can guide resource allocation, policy development, and public awareness efforts to address crime effectively and strategically. This report emphasizes the potential of data-driven approaches in tackling societal challenges and fostering informed decision-making.

Chapter 1. Introduction

1.1 Your Project idea

The project titled "Crime Against Women" focuses on analyzing patterns and trends in crimes committed against women in India from 2001 to 2014. The project aims to uncover critical insights into the prevalence and dynamics of such crimes. By examining this data, the project seeks to identify patterns across regions, study year-on-year changes, and understand correlations between different types of offenses. This initiative is designed to provide a data-driven perspective on how societal, cultural, and regional factors contribute to these crimes, ultimately helping policymakers and stakeholders to implement targeted interventions and preventive measures for ensuring the safety and empowerment of women.

1.2 Data Collection

The data for this project was sourced from the official government portal data.gov.in, ensuring its authenticity and reliability. The process began with identifying relevant datasets related to crimes against women by searching the portal using appropriate keywords. After shortlisting the datasets, we carefully reviewed the metadata and descriptions provided to ensure they matched the scope and objectives of our study. Once the most suitable dataset was identified, it was downloaded in its original format for analysis.[1]

1.3 Dataset Description

The dataset used for this project includes data on crimes committed against women across different states and Union Territories (UTs) in India from 2001 to 2014. This dataset was sourced from reliable government databases and provides comprehensive statistics on various crime categories. The data covers the following states and UTs:

- Andhra Pradesh
- Arunachal Pradesh
- Assam
- Bihar
- Chhattisgarh
- Goa
- Gujarat
- Haryana
- Himachal Pradesh
- Jammu & Kashmir
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra

- Manipur
- Meghalaya
- Mizoram
- Nagaland
- Odisha
- Punjab
- Rajasthan
- Sikkim
- Tamil Nadu
- Tripura
- Uttar Pradesh
- Uttarakhand
- West Bengal
- Andaman & Nicobar Islands
- Chandigarh
- Dadra & Nagar Haveli
- Daman & Diu
- Delhi UT
- Lakshadweep
- Puducherry

The dataset comprises information on various types of crimes against women, including but not limited to:

- Indecent Representation of Women (Prevention) Act
- Immoral Traffic (Prevention) Act
- Cruelty by Husband or Relatives
- Insult to the Modesty of Women
- Assault on Women with Intent to Outrage Her Modesty
- Dowry Death
- Kidnapping & Abduction
- Rape

1.4 Packages required

In this project, several Python packages were utilized to perform data analysis, visualization, and modeling. These packages provided the necessary tools for data preprocessing, exploration, and building predictive models. Below is a description of the key packages used:

1. **Pandas**:
   A powerful data manipulation and analysis library that is essential for handling tabular data. Pandas was used to load, clean, and organize the dataset, as well as to perform various data operations, such as filtering, grouping, and aggregating data.

2. **NumPy**:
   A fundamental package for numerical computation in Python. NumPy provided efficient data structures, such as arrays, and support for mathematical operations, enabling faster computation and data manipulation.

3. **Missingno**:
   A specialized library for visualizing missing data. Missingno was used to identify and understand patterns in missing values within the dataset, allowing for informed decisions during the data cleaning and imputation processes.

4. **Matplotlib**:
   A widely-used plotting library that provided the tools for creating static, animated, and interactive visualizations. Matplotlib was used for generating various charts and plots to represent the distribution and relationships in the data.

5. **Seaborn**:
   A data visualization library built on top of Matplotlib that provides an interface for creating attractive and informative statistical graphics. Seaborn was employed to create complex visualizations such as heatmaps, bar plots, and violin plots, which helped to reveal trends and insights in the data.

6. **Scikit-Learn (sklearn)**:
   A comprehensive library for machine learning that provides tools for data preprocessing, feature selection, and building predictive models. Scikit-Learn was used for model fitting, including training and evaluating regression models to predict trends and relationships in the dataset.

7. **SciPy**:
   A library used for scientific and technical computing. SciPy was utilized for advanced mathematical operations, such as statistical tests and optimizations, aiding in hypothesis testing and analytical tasks.

These packages were essential for transforming raw data into meaningful insights and ensuring that the data analysis was both accurate and efficient. By leveraging their combined functionalities, we were able to conduct a comprehensive exploratory data analysis, build models, and visualize trends and patterns in the dataset.

Chapter 2. Data Cleaning

Data cleaning is a crucial step in the data analysis pipeline, ensuring the dataset is consistent, accurate, and ready for analysis. The following steps were taken to clean and preprocess the dataset before further analysis:

2.1 Missing data analysis

To begin the data cleaning process, we first examined the dataset for any missing values. A detailed analysis was conducted to identify the presence of null values in the dataset. The following table summarizes the check for missing values:

	Missing Values	% of Total Values
STATE/UT	0	0.0
CRIME HEAD	0	0.0
2001	0	0.0
2002	0	0.0
2003	0	0.0
2004	0	0.0
2005	0	0.0
2006	0	0.0
2007	0	0.0
2008	0	0.0
2009	0	0.0
2010	0	0.0
2011	0	0.0
2012	0	0.0

Table 2.1: Missing Values Analysis

2.1.1 Missing Data Heatmap

A heatmap was plotted to visualize the presence of missing data across the dataset. This visualization confirmed that there were no missing values, as expected. Including this figure would further validate the quality of the dataset and make it easier to present the findings.

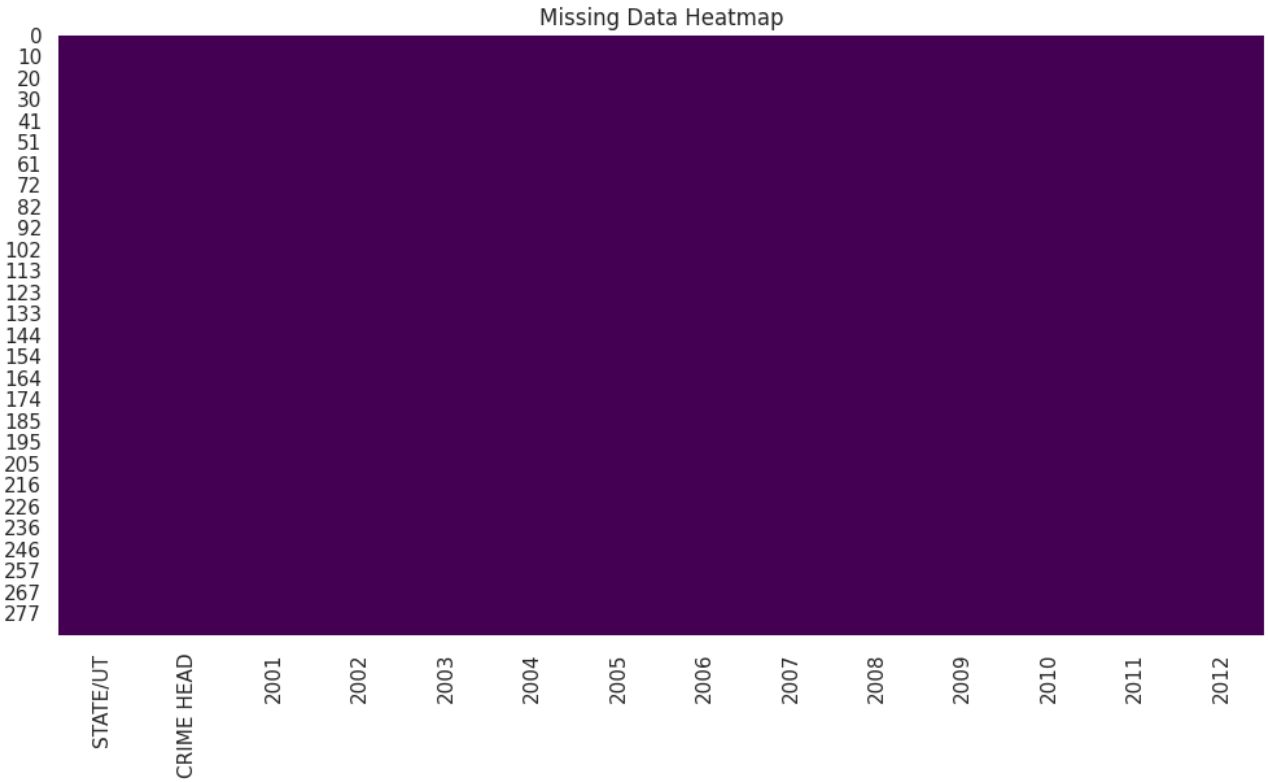


Figure 2.1: Heatmap of Missing Data

2.2 Imputation

The dataset was further examined for any potential missing data that could affect analysis. It was confirmed that there were no missing values, which eliminated the need for imputation.

2.3 Data Cleaning Steps Performed

2.3.1 Ensured Years are Numeric

The data in the ”Year” column was verified and converted to a numeric format to maintain consistency and facilitate analysis. This step was essential to ensure that the year data could be used accurately in subsequent analysis.

2.3.2 Checked for Duplicates

The dataset was inspected for duplicate entries, and it was confirmed that there were no duplicate records. This ensured that our analysis would not be skewed by repeated data points.

2.3.3 Corrected Inconsistent Formatting in ’STATE/UT’ Column

The ’STATE/UT’ column was standardized to ensure consistent naming conventions for all states and union territories. This step was vital to avoid mismatches and errors when analyzing the data by region.

Chapter 3. Visualization

The process of data visualization is fundamental in understanding and interpreting complex datasets. By creating visual representations of data, we can identify patterns, trends, and relationships that may not be immediately apparent in raw data form. This chapter covers various techniques used for analyzing and visualizing crime data across different states and years. Visualizations enable stakeholders to gain insights, make informed decisions, and formulate strategies for addressing issues effectively. The following sections detail univariate, bivariate, and multivariate analyses, using a combination of histograms, bar plots, heatmaps, violin plots, pie charts, and line plots.

3.1 Univariate analysis

Univariate analysis involves examining a single variable at a time to understand its distribution and properties. This section focuses on different visualization techniques used to assess the distribution of crime data across various years and categories.

3.1.1 Histograms, Violin Plots, and Boxplots

For each year from 2001 to 2012, histograms, violin plots, and boxplots were plotted to visualize the distribution of crime data. These plots help in understanding the frequency, distribution, and outliers present in the data over the years.

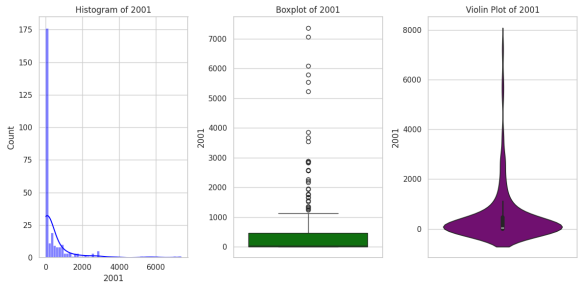


Figure 3.1: Histogram for the year 2001

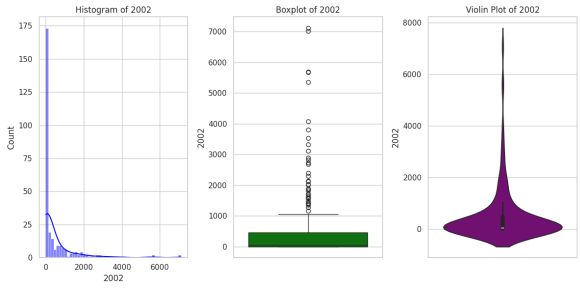


Figure 3.2: Histogram for the year 2002

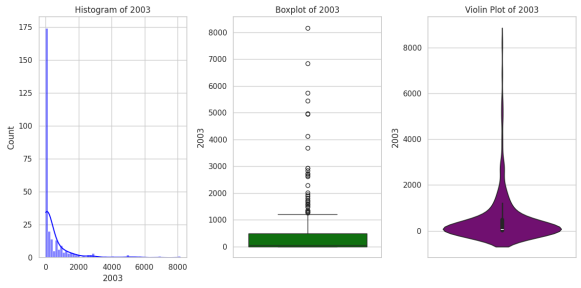


Figure 3.3: Histogram for the year 2003

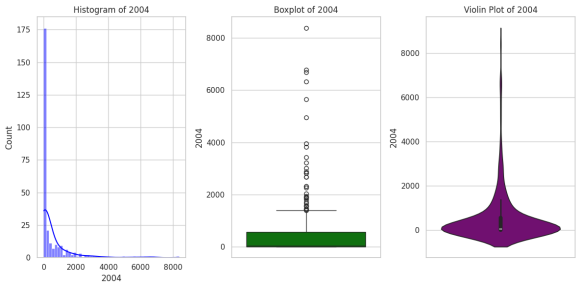


Figure 3.4: Histogram for the year 2004

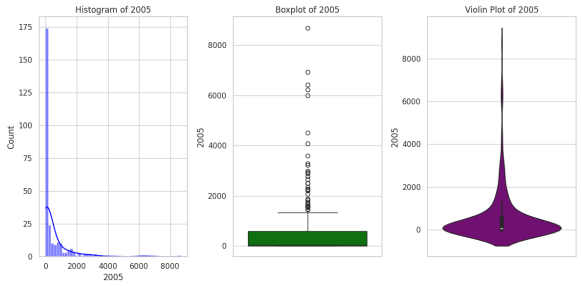


Figure 3.5: Histogram for the year 2005

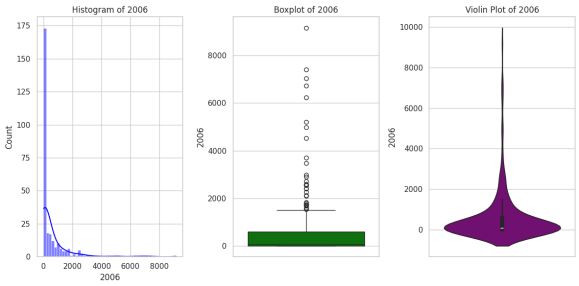


Figure 3.6: Histogram for the year 2006

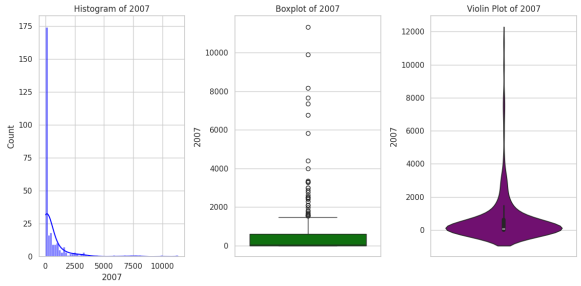


Figure 3.7: Histogram for the year 2007

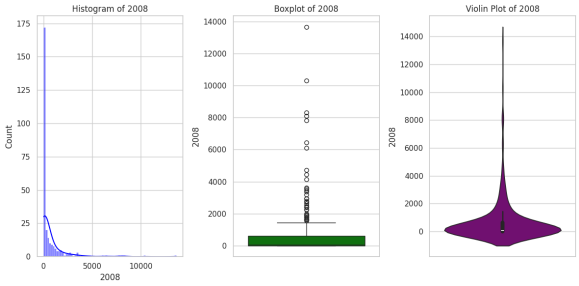


Figure 3.8: Histogram for the year 2008

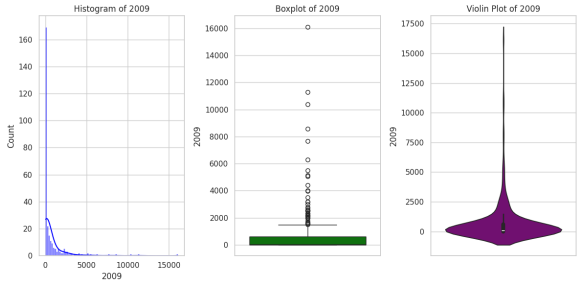


Figure 3.9: Histogram for the year 2009

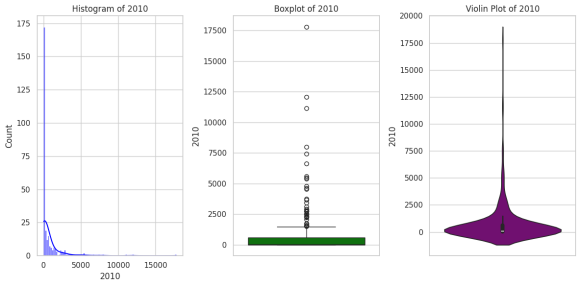


Figure 3.10: Histogram for the year 2010

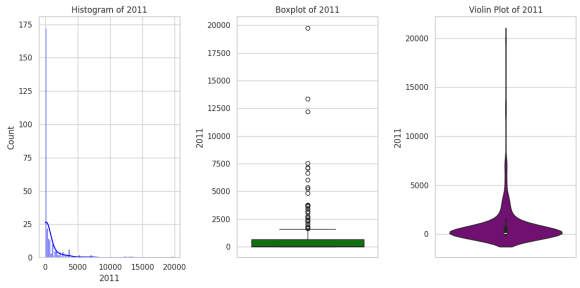


Figure 3.11: Histogram for the year 2011

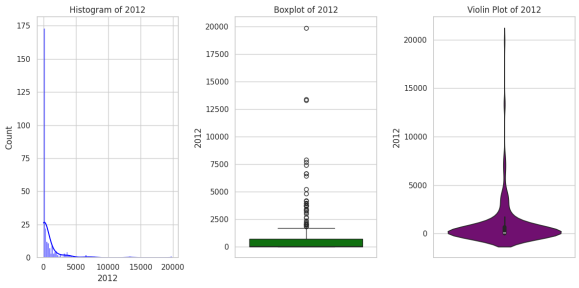


Figure 3.12: Histogram for the year 2012

3.1.2 Barplot for Crime Count by Crime Head

A barplot was created to show the distribution of crimes by different crime heads.

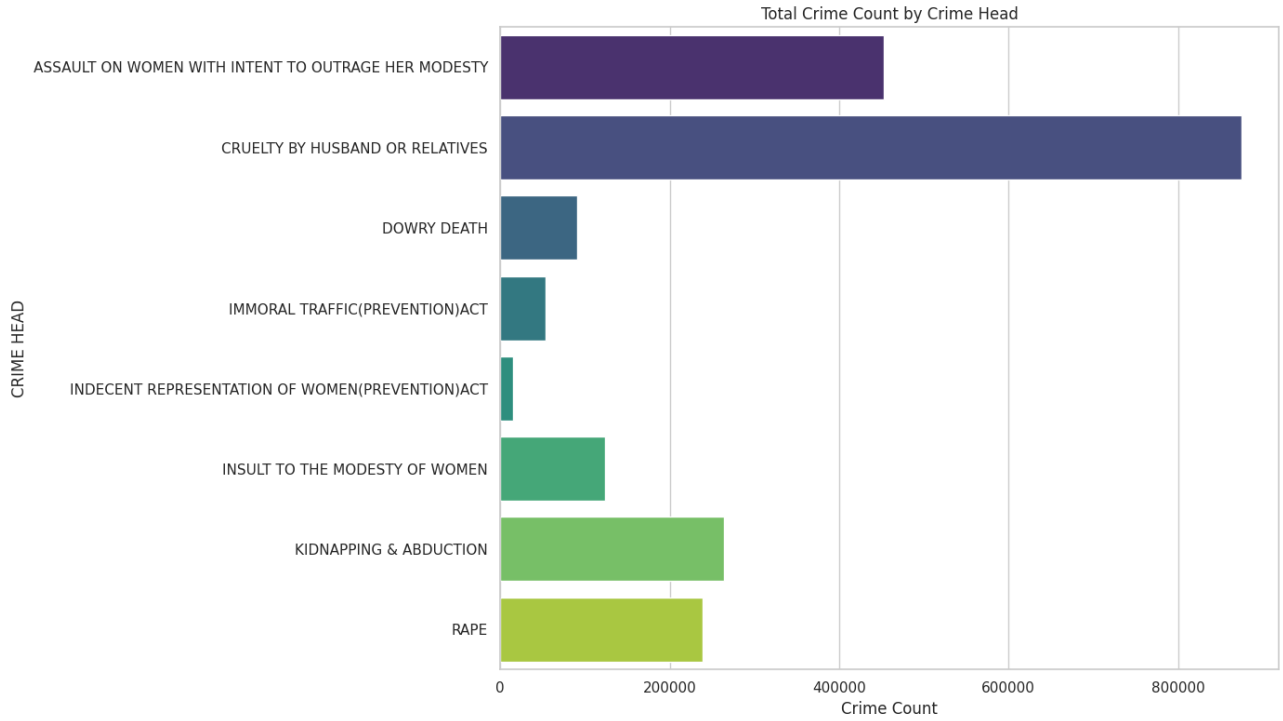


Figure 3.13: Barplot for Crime Count by Crime Head

3.1.3 Histograms for Each Crime Head

Histograms were drawn for each crime head to better understand how specific crime types were distributed over the years.

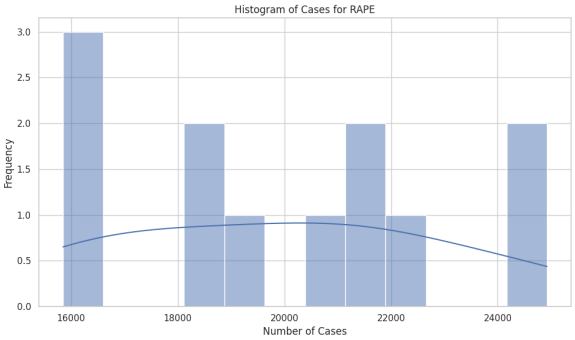


Figure 3.14: Histogram for RAPE

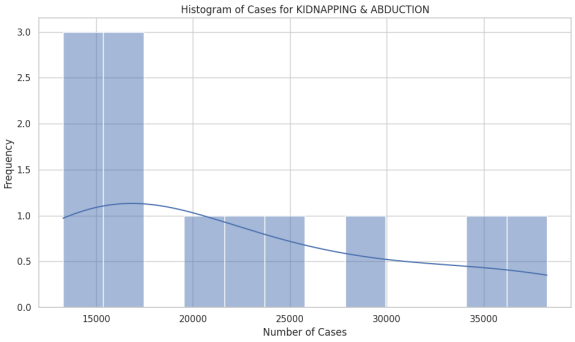


Figure 3.15: Histogram for KIDNAPPING and ABDUCTION

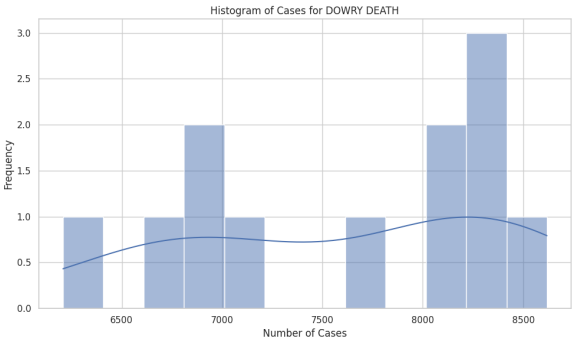


Figure 3.16: Histogram for DOWRY DEATH

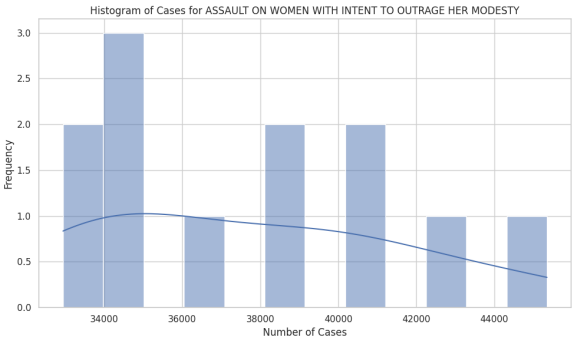


Figure 3.17: Histogram for ASSAULT ON WOMEN

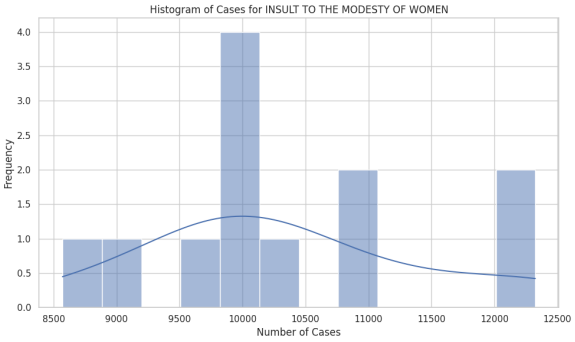


Figure 3.18: Histogram for INSULT TO THE MODESTY OF WOMEN

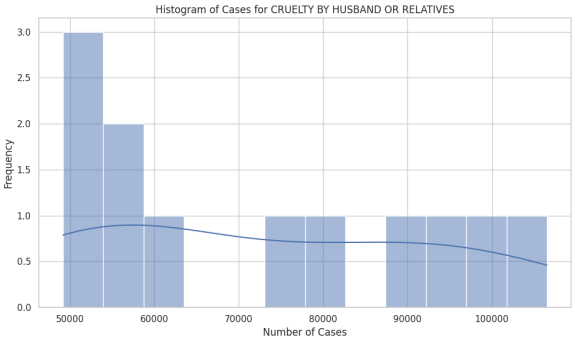


Figure 3.19: Histogram for CRUELTY BY HUSBAND OR RELATIVES

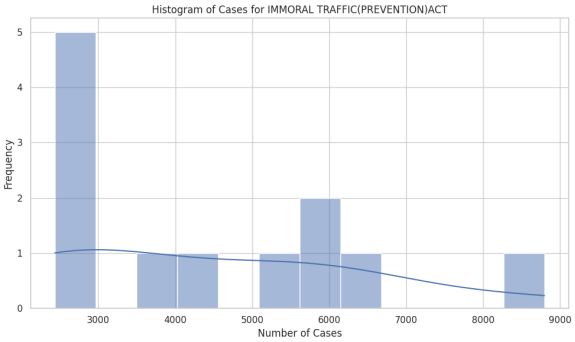


Figure 3.20: Histogram for IMMORAL TRAFFIC (PREVENTION) ACT

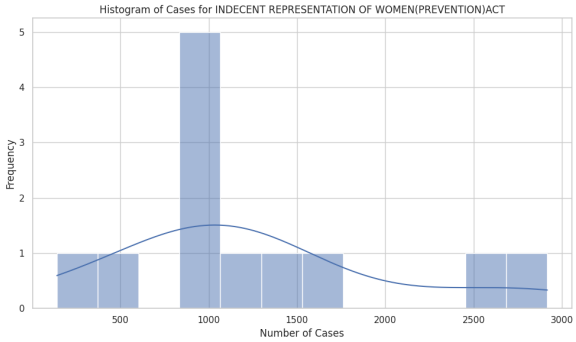


Figure 3.21: Histogram for INDECENT REPRESENTATION OF WOMEN (PREVENTION) ACT

3.1.4 Crime Distribution Across States for Each Year

A barplot was used to illustrate the crime distribution across states for each year.

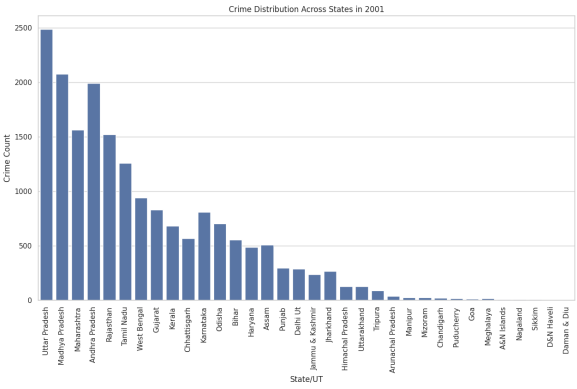


Figure 3.22: Crime Distribution for 2001

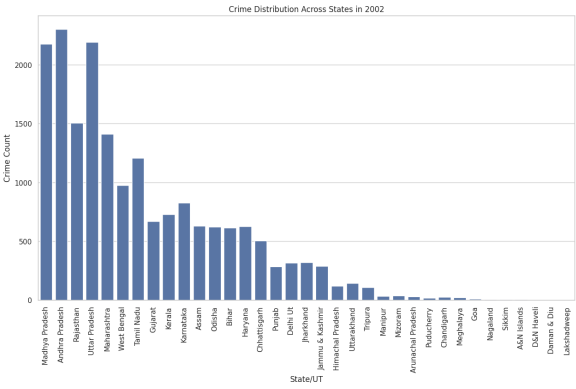


Figure 3.23: Crime Distribution for 2002

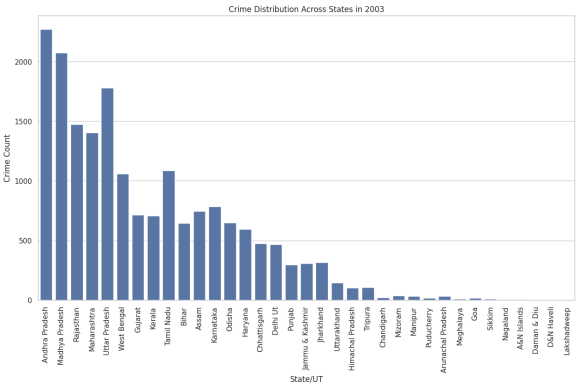


Figure 3.24: Crime Distribution for 2003

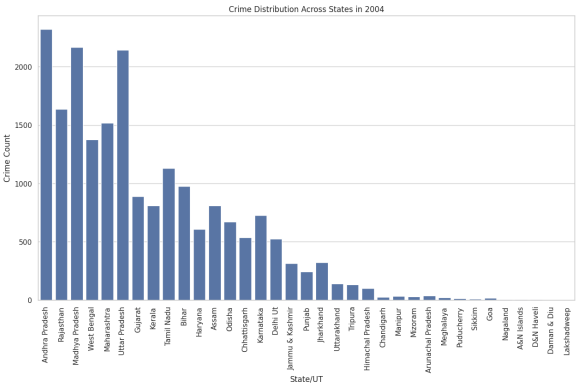


Figure 3.25: Crime Distribution for 2004

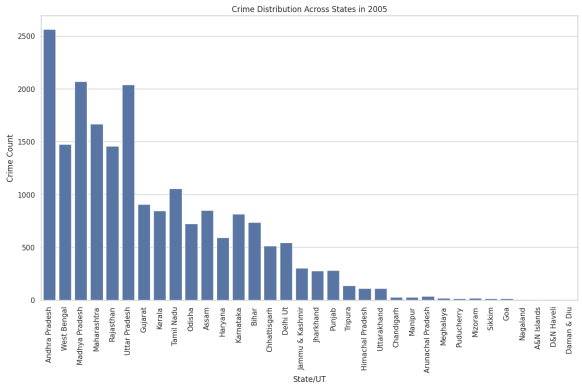


Figure 3.26: Crime Distribution for 2005

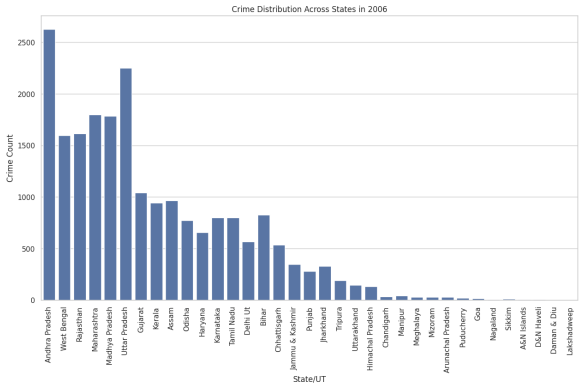


Figure 3.27: Crime Distribution for 2006

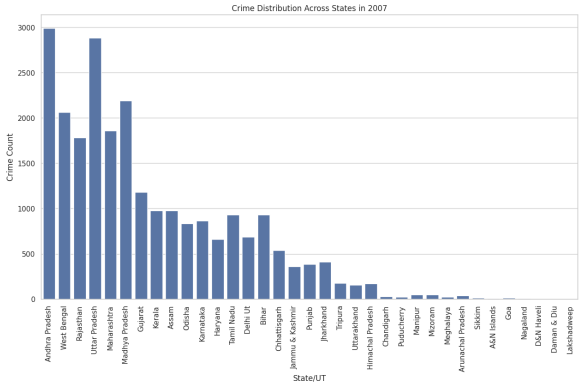


Figure 3.28: Crime Distribution for 2007

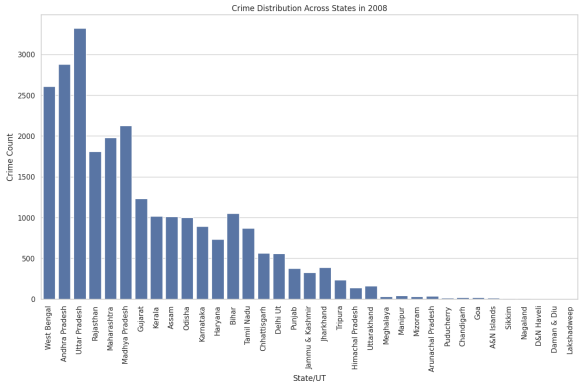


Figure 3.29: Crime Distribution for 2008

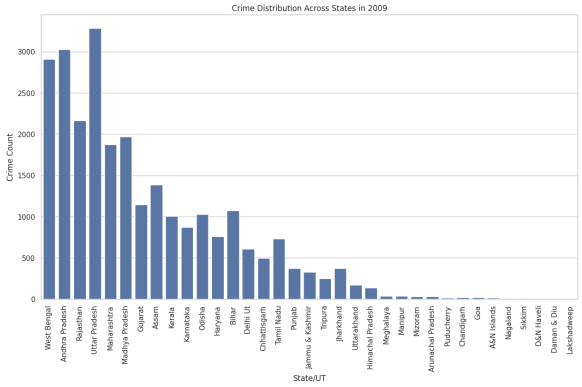


Figure 3.30: Crime Distribution for 2009

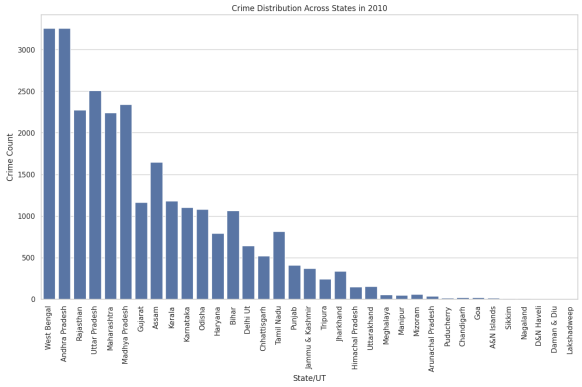


Figure 3.31: Crime Distribution for 2010

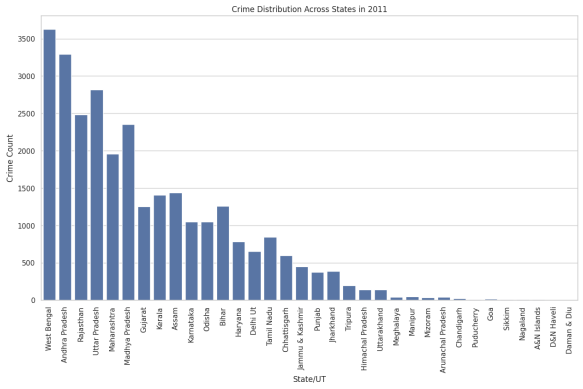


Figure 3.32: Crime Distribution for 2011

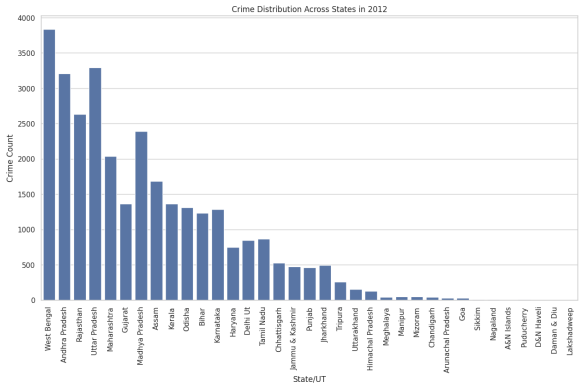


Figure 3.33: Crime Distribution for 2012

3.1.5 Violin Plots for Each Crime Head

Violin plots were created for each crime head to show the distribution and density of data.

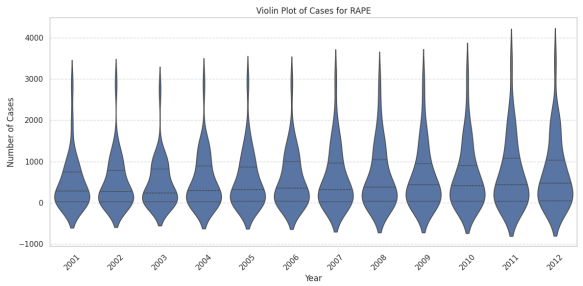


Figure 3.34: Violin Plot for RAPE

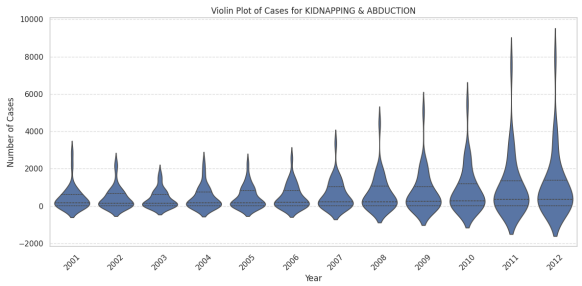


Figure 3.35: Violin Plot for KIDNAPPING

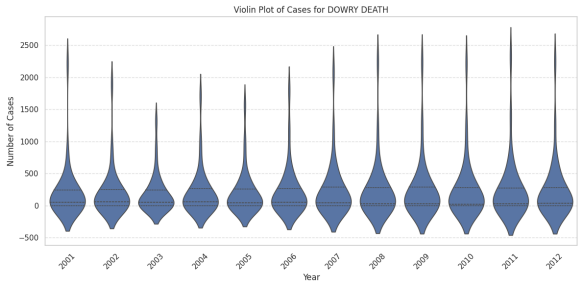


Figure 3.36: Violin Plot for DOWRY DEATH

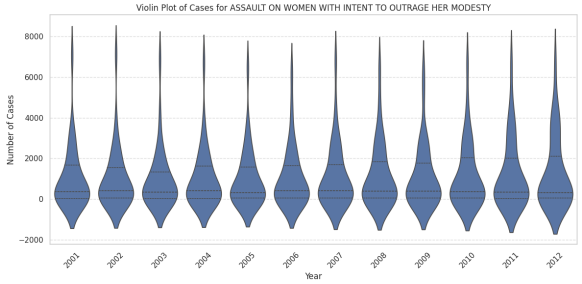


Figure 3.37: Violin Plot for ASSAULT ON WOMEN

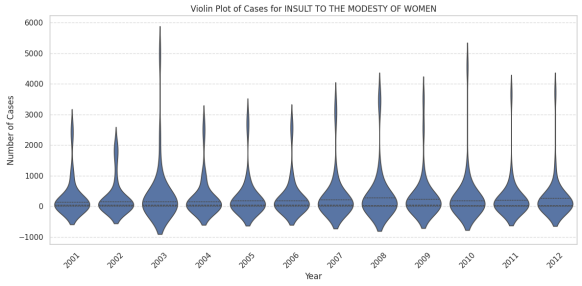


Figure 3.38: Violin Plot for INSULT TO MODESTY

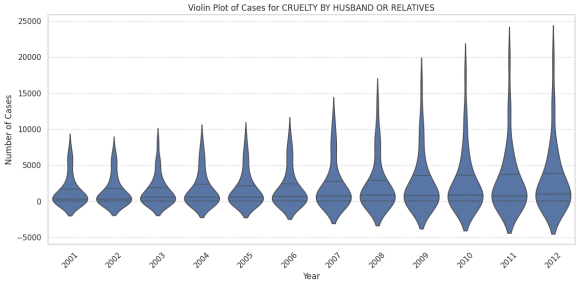


Figure 3.39: Violin Plot for CRUELTY BY HUSBAND

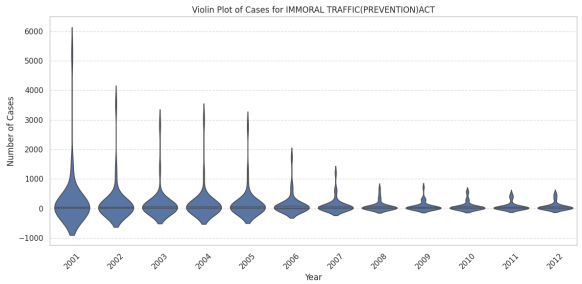


Figure 3.40: Violin Plot for IMMORAL TRAFFIC

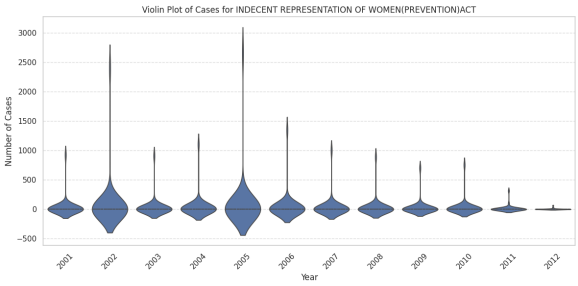


Figure 3.41: Violin Plot for INDECENT REPRESENTATION

3.2 Bivariate Analysis

Bivariate analysis examines the relationship between two variables to uncover patterns and correlations.

3.2.1 Iteration Through Crime Heads and Plot Creation

The data for each crime head was filtered and aggregated by year to create year-wise data comparisons.

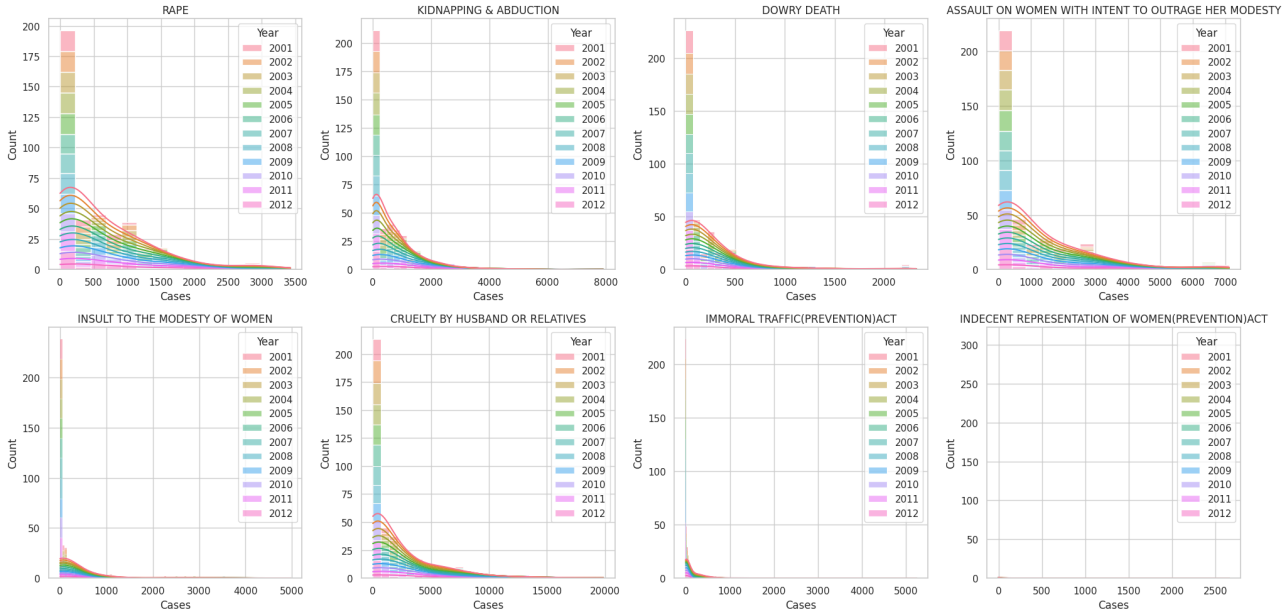


Figure 3.42: Trends for Each Crime Head Over the Years

3.2.2 Heatmap for Crime Count by State/UT and Year

A heatmap was generated to represent the number of crimes reported across different states and union territories (UTs) for each year. This visualization provides a clear, color-coded depiction of crime distribution.

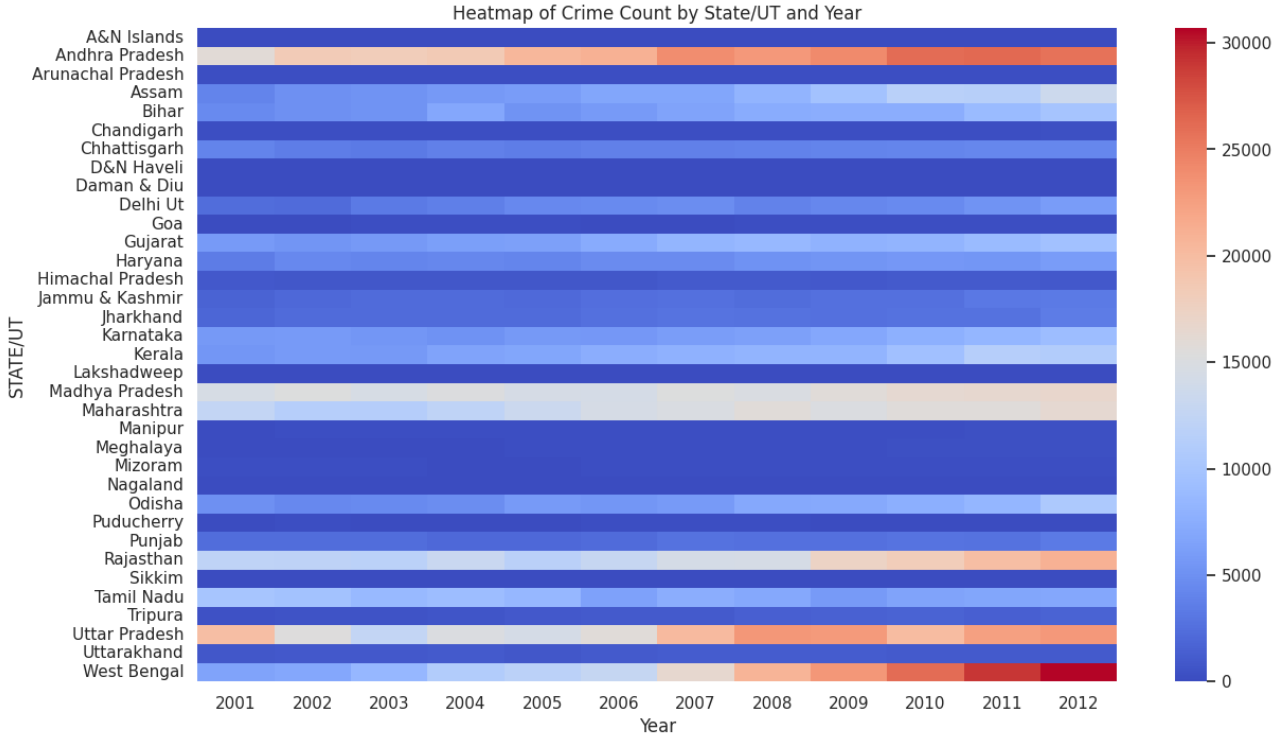


Figure 3.43: Heatmap of Crime Count by State/UT and Year

3.2.3 Pairplot of Year Columns

A pairplot was created to display pairwise relationships between year columns.

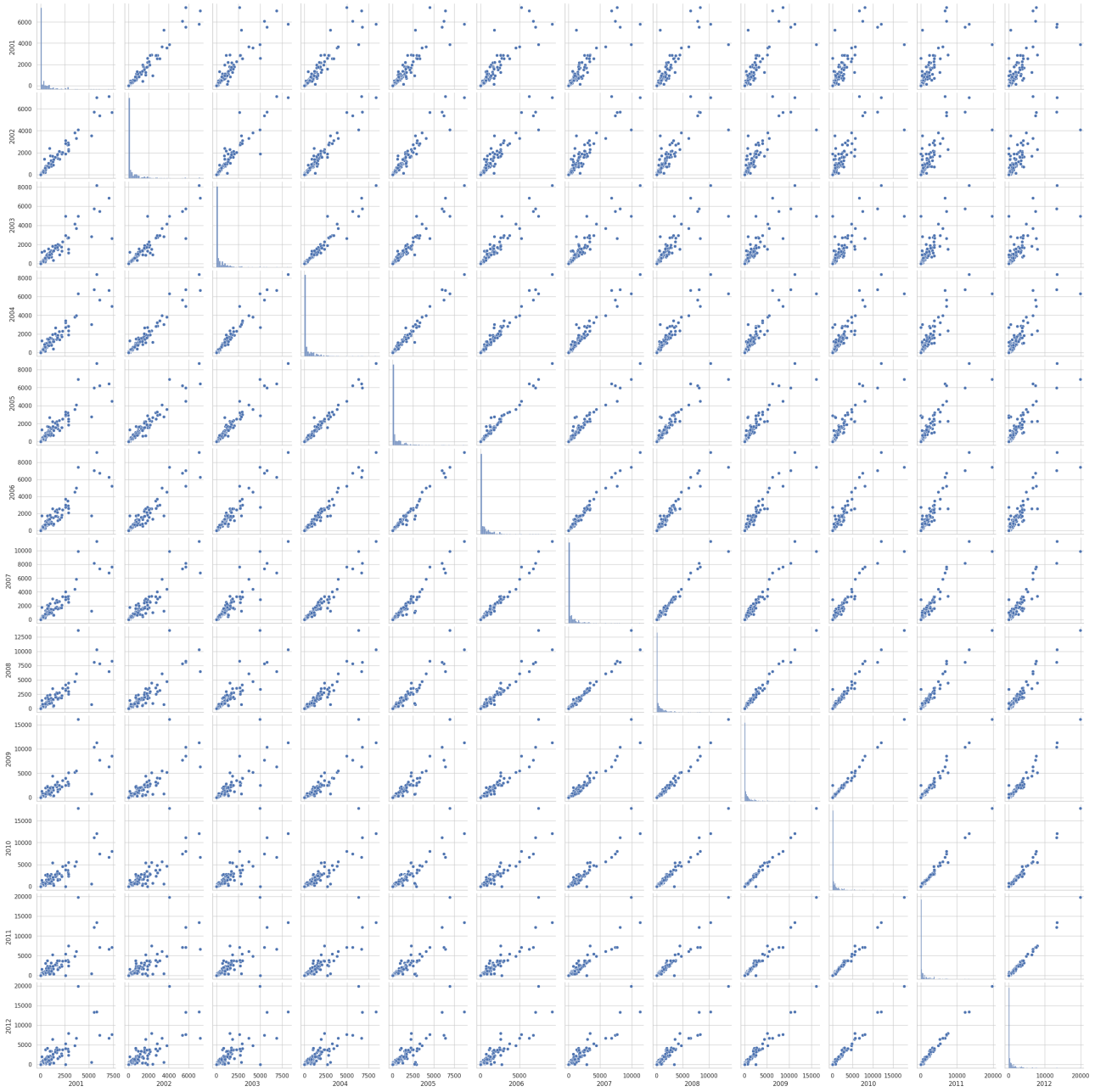


Figure 3.44: Pairplot of Year Columns

3.2.4 Correlation Matrix for Each Year

A correlation matrix was generated to understand relationships between crime counts across years.

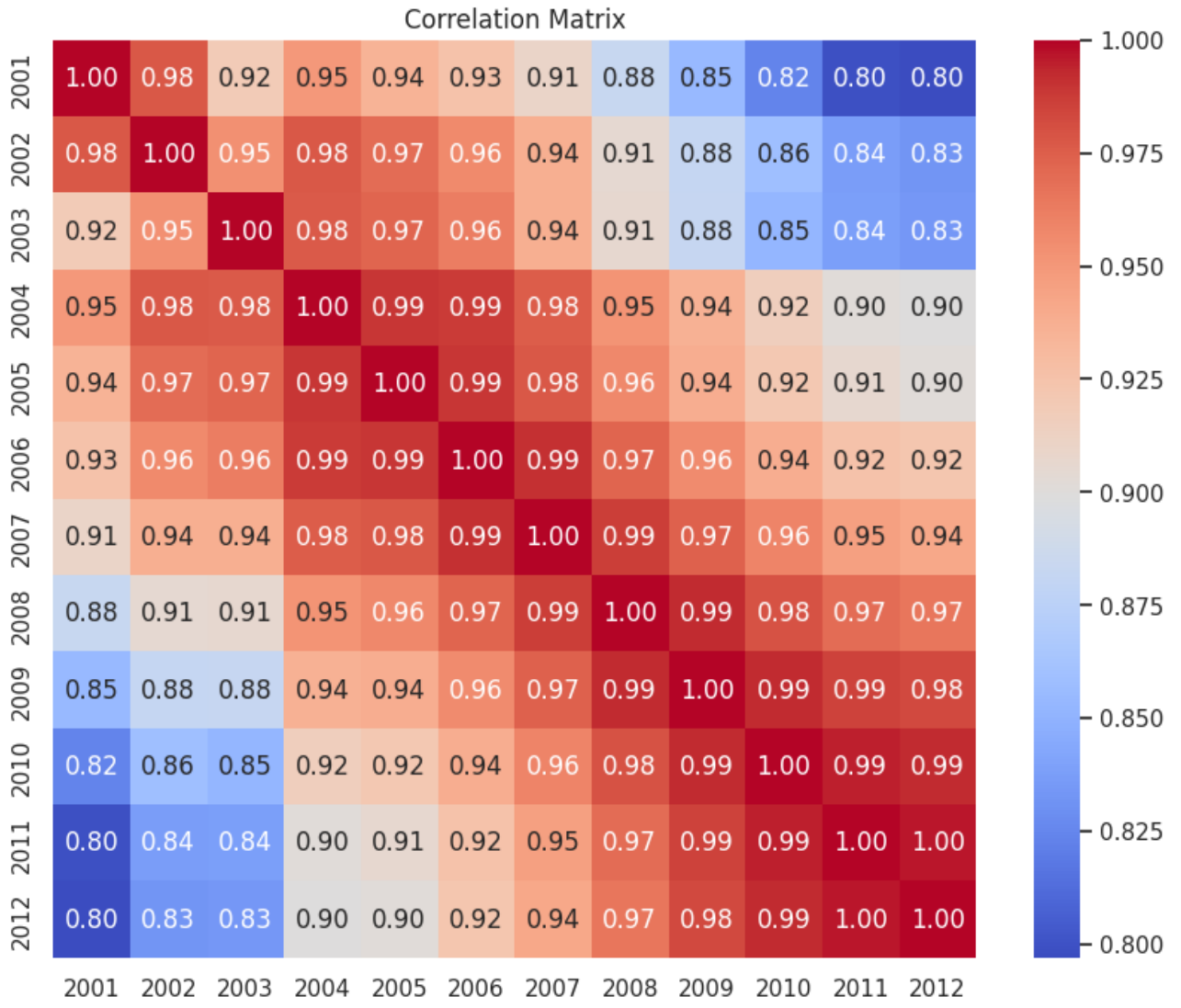


Figure 3.45: Correlation Matrix for Each Year

3.2.5 Crimes from 2001 to 2012 by State/UT

A barplot was created to show the number of crimes reported across different states/UTs.

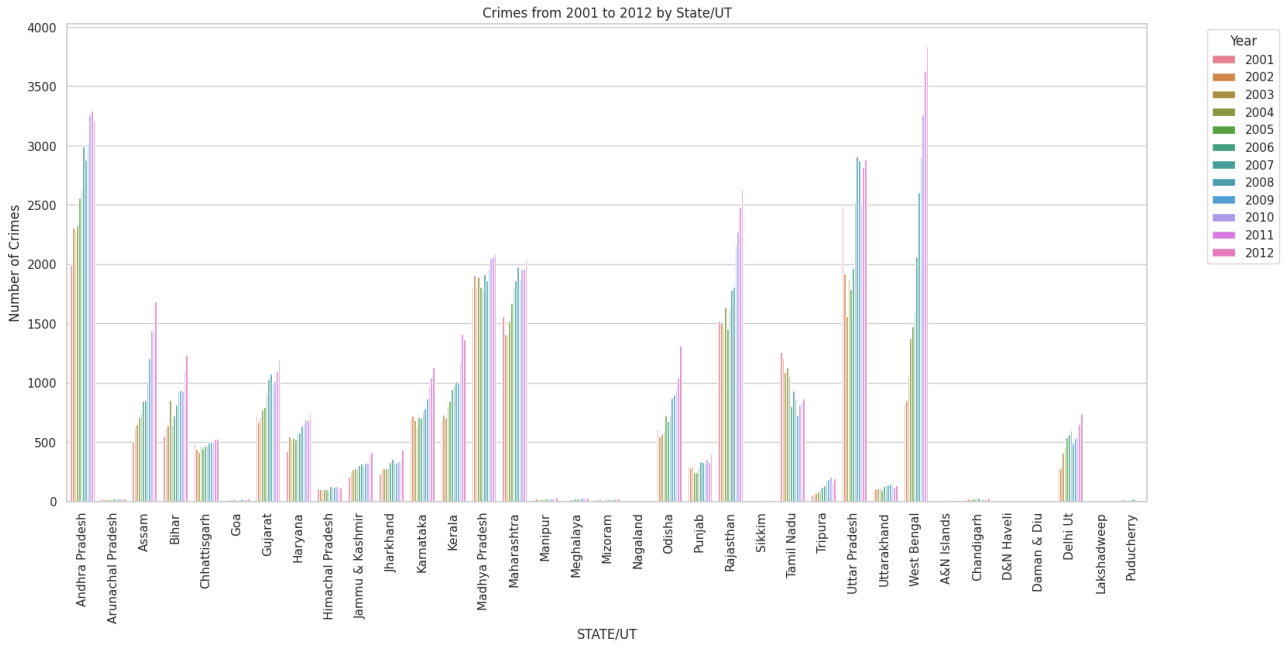


Figure 3.46: Crimes from 2001 to 2012 by State/UT

3.3 Multivariate Analysis

Multivariate analysis aims to understand the complex interactions between more than two variables and their collective impact on the data.

3.3.1 Total Crime Count per State Over the Years

A plot was created to show the total crime count per state across the years.

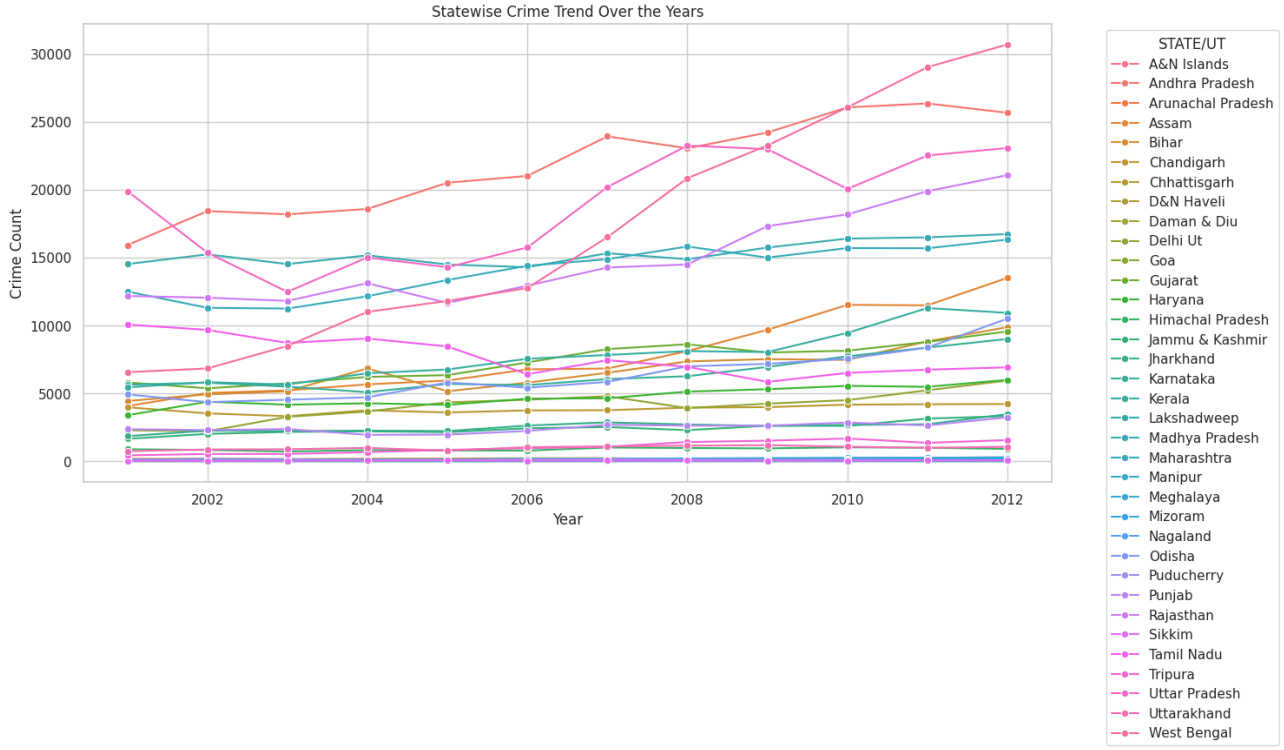


Figure 3.47: Total Crime Count per State Over the Years

3.3.2 Total Crime Count for Each Crime Head Over the Years

A visualization was created to show the total number of crimes for each crime head over the years.

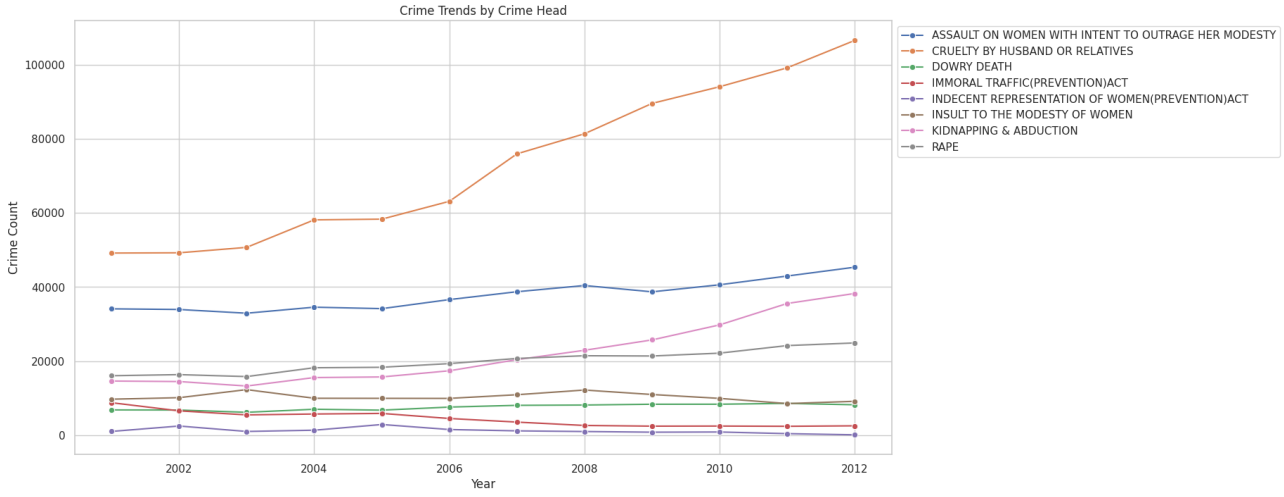


Figure 3.48: Total Crime Count for Each Crime Head Over the Years

3.3.3 Stacked Crime Count by Crime Head Over the Years

A stacked bar plot was used to display the total number of crimes segmented by each crime head.

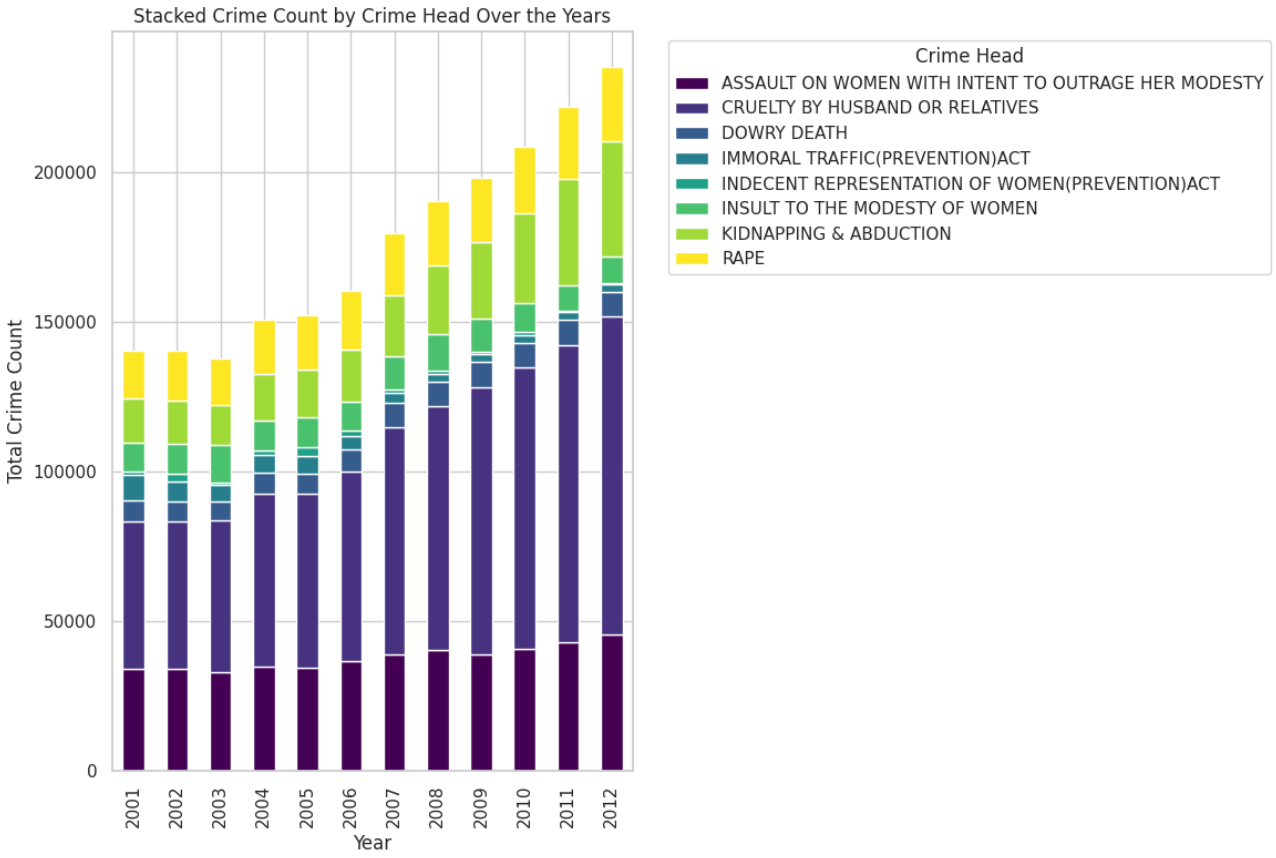


Figure 3.49: Stacked Crime Count by Crime Head Over the Years

3.3.4 Aggregate Data by Year and Crime Head

Pie charts were created to show the percentage distribution of different crime types for each year.

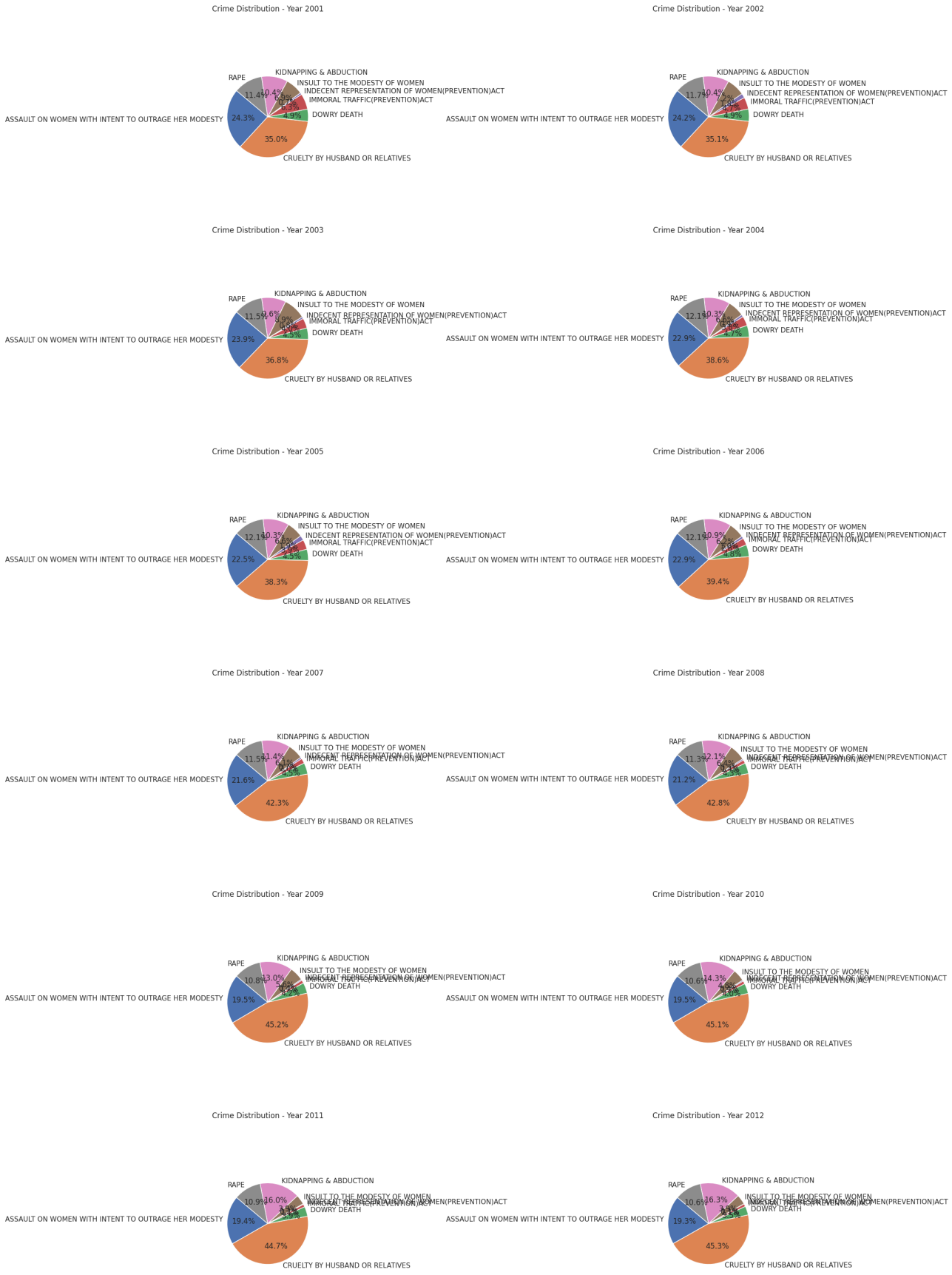


Figure 3.50: Pie Chart for Crime Distribution by Year

3.3.5 Aggregate Data by State/UT and Crime Head

Pie charts were created to show the distribution of crime types within each state/UT.

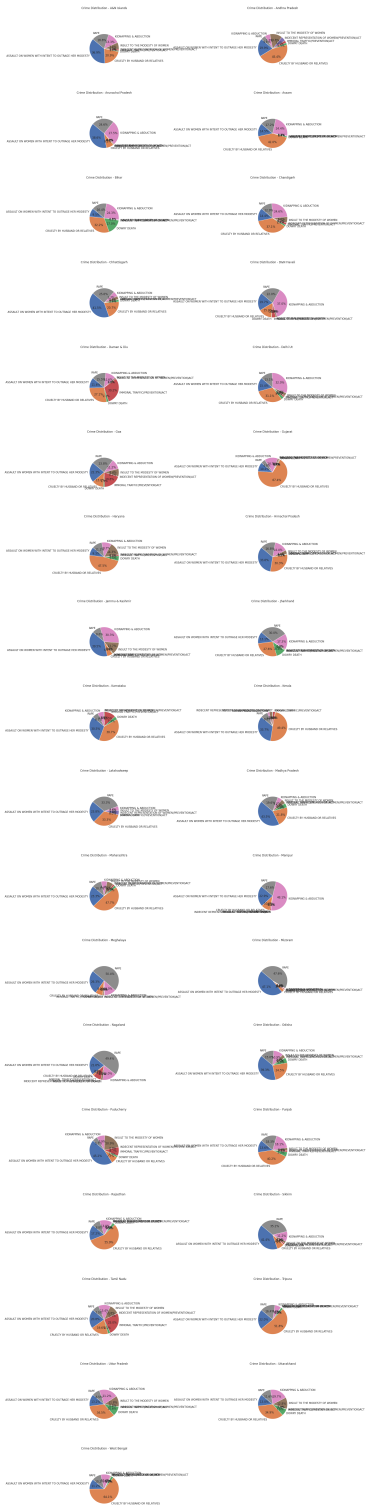


Figure 3.51: Pie Chart for Crime Distribution by State/UT

3.3.6 Crime Trends by State/UT

A line plot was created to show crime trends over time for each state/UT.

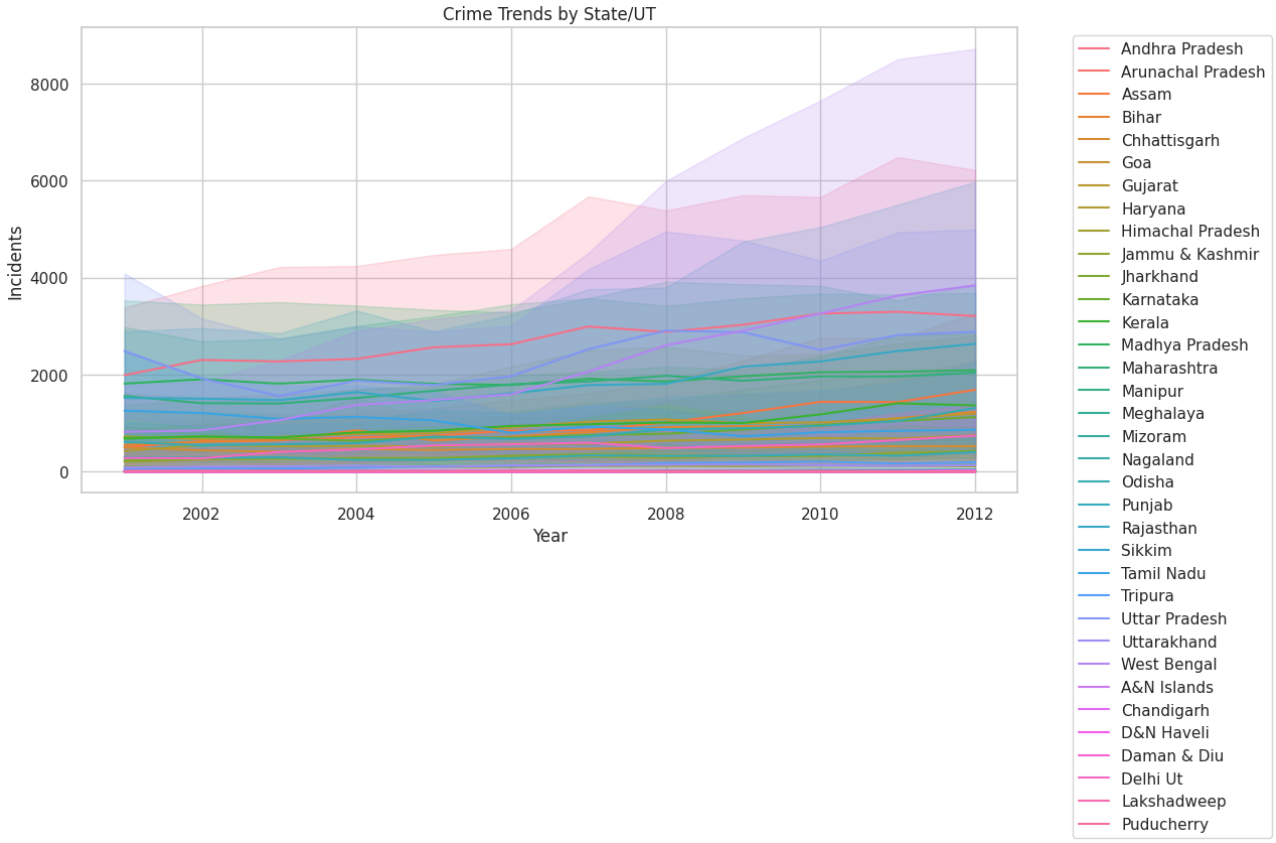


Figure 3.52: Crime Trends by State/UT

Chapter 4. Feature Engineering

Feature engineering involves transforming raw data into meaningful features to enhance the performance of machine learning models. This section covers various processes such as feature extraction, feature selection, and dealing with outliers to improve the interpretability and predictive capability of the dataset.

4.1 Feature extraction

Feature extraction is a crucial step in analyzing and preprocessing data to uncover significant patterns and relationships.

4.1.1 Table of Crime Count by State/UT and Crime Head

The table below lists the total crime counts for each state/UT, showing the most prevalent crime types.

STATE/UT	CRIME HEAD	Crime Count
A&N Islands	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	214
Andhra Pradesh	CRUELTY BY HUSBAND OR RELATIVES	119007
Arunachal Pradesh	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	784
Assam	CRUELTY BY HUSBAND OR RELATIVES	39388
Bihar	CRUELTY BY HUSBAND OR RELATIVES	25680
Chandigarh	CRUELTY BY HUSBAND OR RELATIVES	807
Chhattisgarh	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	19165
D&N Haveli	KIDNAPPING & ABDUCTION	94
Daman & Diu	IMMORAL TRAFFIC (PREVENTION) ACT	39
Delhi Ut	KIDNAPPING & ABDUCTION	15650
Goa	RAPE	350
Gujarat	CRUELTY BY HUSBAND OR RELATIVES	59431
Haryana	CRUELTY BY HUSBAND OR RELATIVES	27112
Himachal Pradesh	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	3597
Jammu & Kashmir	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	11509
Jharkhand	RAPE	9159
Karnataka	CRUELTY BY HUSBAND OR RELATIVES	30052
Kerala	CRUELTY BY HUSBAND OR RELATIVES	46074
Lakshadweep	CRUELTY BY HUSBAND OR RELATIVES	7
Madhya Pradesh	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	79878
Maharashtra	CRUELTY BY HUSBAND OR RELATIVES	80363
Manipur	KIDNAPPING & ABDUCTION	1057
Meghalaya	RAPE	1020
Mizoram	RAPE	826
Nagaland	RAPE	220
Odisha	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	29946
Puducherry	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	613
Punjab	CRUELTY BY HUSBAND OR RELATIVES	11998
Rajasthan	CRUELTY BY HUSBAND OR RELATIVES	100101
Sikkim	ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	230
Tamil Nadu	IMMORAL TRAFFIC (PREVENTION) ACT	23221
Tripura	CRUELTY BY HUSBAND OR RELATIVES	6514
Uttar Pradesh	CRUELTY BY HUSBAND OR RELATIVES	77617
Uttarakhand	CRUELTY BY HUSBAND OR RELATIVES	4142
West Bengal	CRUELTY BY HUSBAND OR RELATIVES	130668

Table 4.1: Crime Count by State/UT and Crime Head

4.1.2 Total Crimes, Growth Rate (CAGR), and Average Yearly Crimes

Total Crimes

The total number of crimes refers to the cumulative count of all crimes recorded in the dataset over the given time period. This raw figure provides a broad overview of the level of criminal activity but does not take into account yearly fluctuations or trends.

$$\text{Total Crimes} = \sum_{i=1}^{n} \text{Crimes}_i$$

where n is the number of years or time periods and Crimes_i represents the total number of crimes in year i.

Compound Annual Growth Rate (CAGR)

The Compound Annual Growth Rate (CAGR) is a useful metric for determining the annualized rate of growth or decline over a period of time, accounting for the compounding effect. It helps in understanding how the crime rate has changed year-over-year.

The formula for CAGR is:

$$\text{CAGR} = \left(\frac{\text{Ending Value}}{\text{Beginning Value}} \right)^{\frac{1}{\text{Number of Years}}} - 1$$

In this case:

- The Beginning Value is the number of crimes in the first year of the dataset. - The Ending Value is the number of crimes in the last year of the dataset. - The Number of Years is the number of years over which the growth is being measured.

The CAGR allows us to evaluate whether the crime rate is growing, shrinking, or remaining stable over the period of time. [2]

Average Yearly Crimes

The average yearly crimes is simply the total number of crimes divided by the number of years in the dataset. This provides a basic average of criminal activity per year, though it does not capture variability or trends over time.

The formula for average yearly crimes is:

$$\text{Average Yearly Crimes} = \frac{\text{Total Crimes}}{\text{Number of Years}}$$

4.2 Feature selection

Feature selection is a process of selecting the most relevant features to use in model building to improve model performance and reduce overfitting. The following subsections detail the feature selection and outlier treatment performed.

4.2.1 Total Crimes by State/UT Plot

A visualization was created to show the total number of crimes reported in each state/UT.

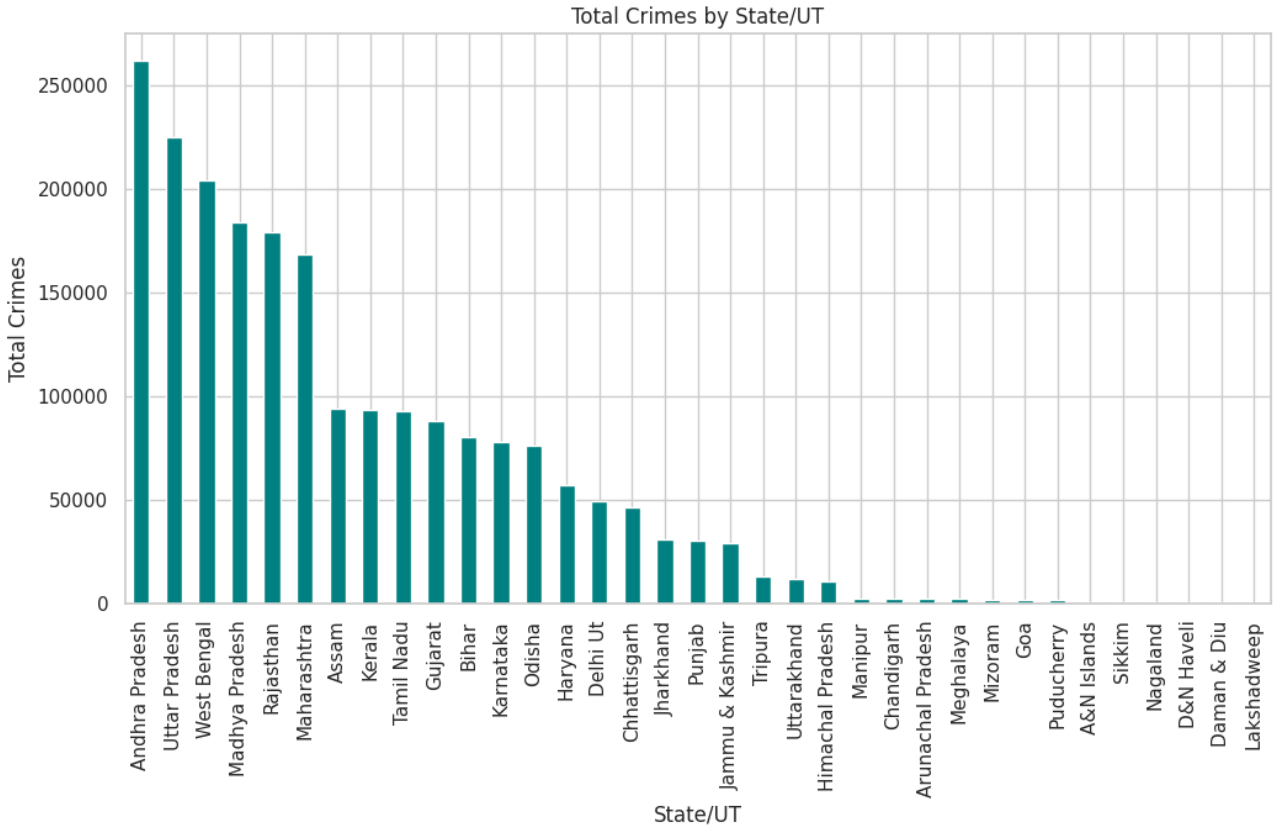


Figure 4.1: Total Crimes by State/UT

4.2.2 Plot of Total Crime Count by Year

A plot showing the total crime count aggregated over the years helps to identify trends and patterns.

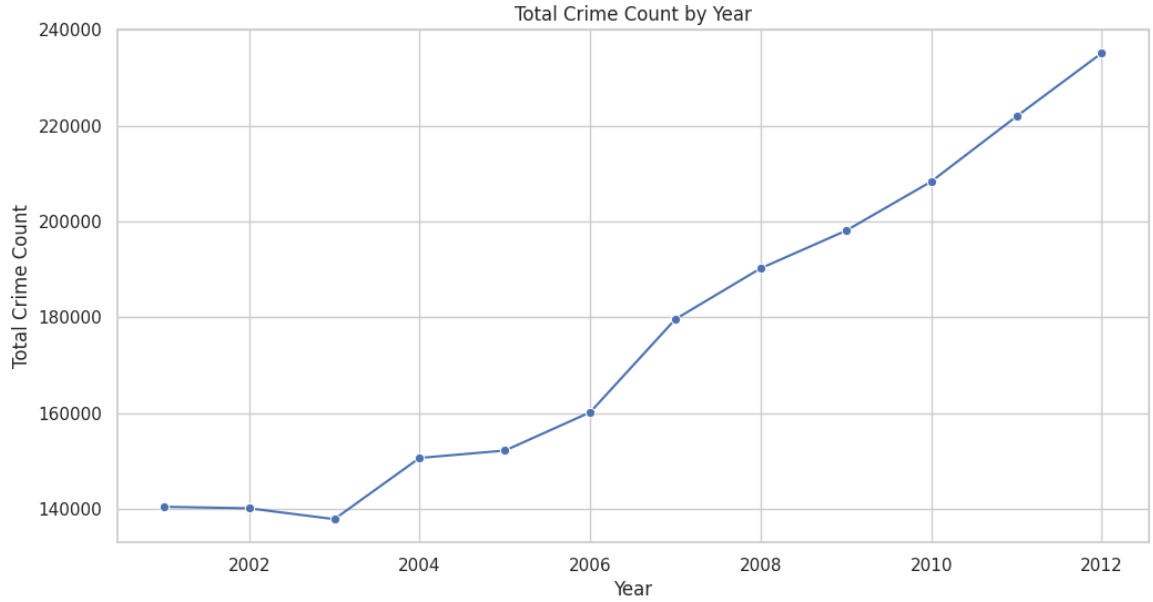


Figure 4.2: Total Crime Count by Year

4.2.3 Crime Distribution by Category

A boxplot was created to show the distribution of different crime categories and identify outliers.

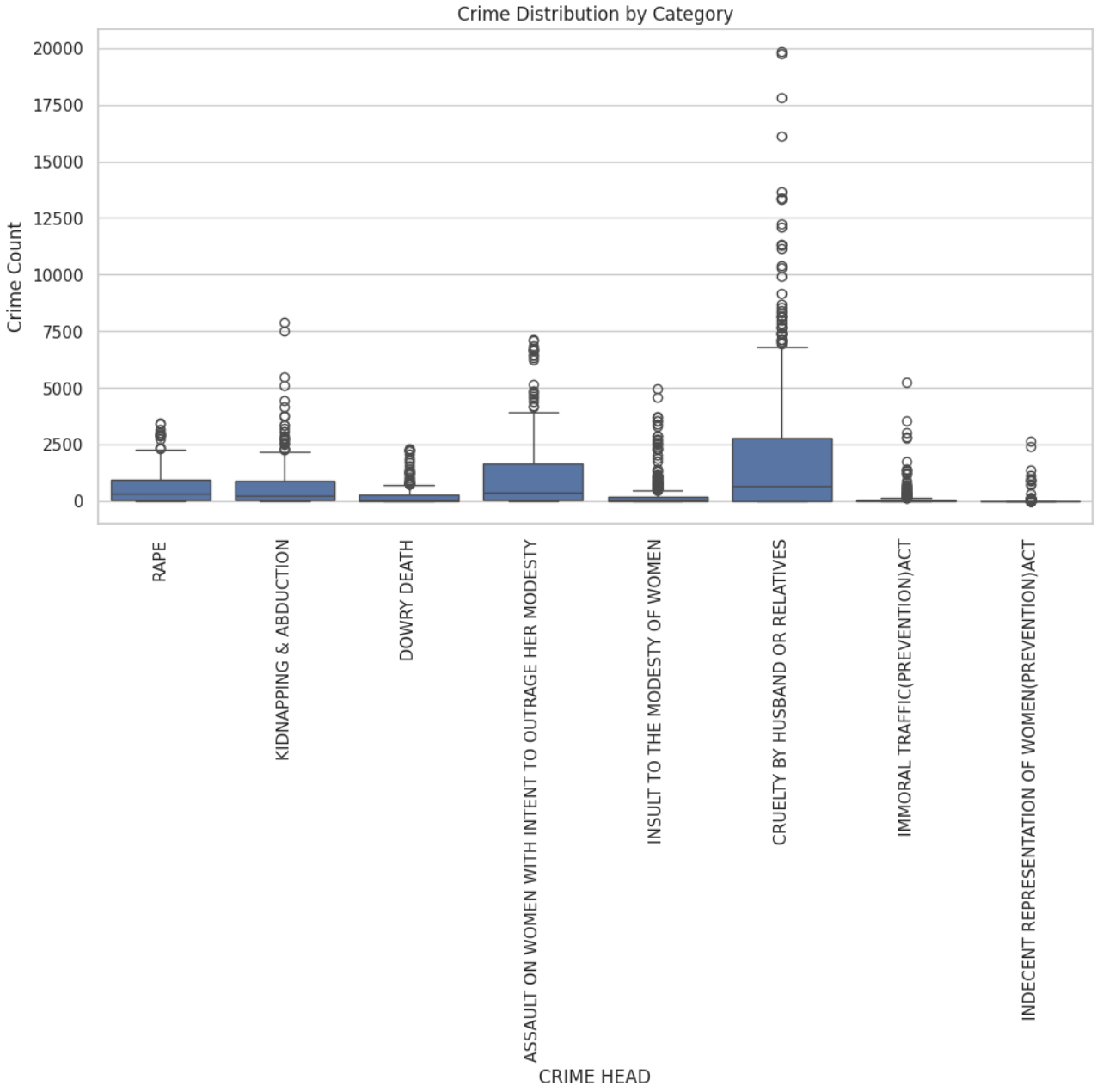


Figure 4.3: Crime Distribution by Category

4.2.4 Year-wise Percentage Contribution of Each Crime Head

The year-wise contribution of each crime head was plotted to show how each type of crime has contributed to the total over the years.

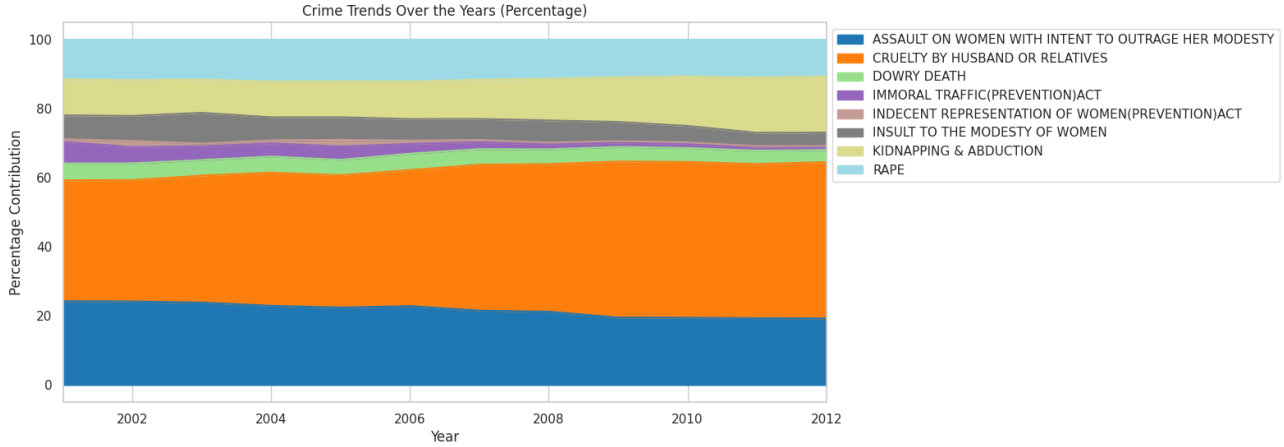


Figure 4.4: Year-wise Percentage Contribution of Each Crime Head

4.2.5 CRIME HEAD Distribution for Each Year

Boxplots for each year were plotted to show the distribution of various crime types within each year.

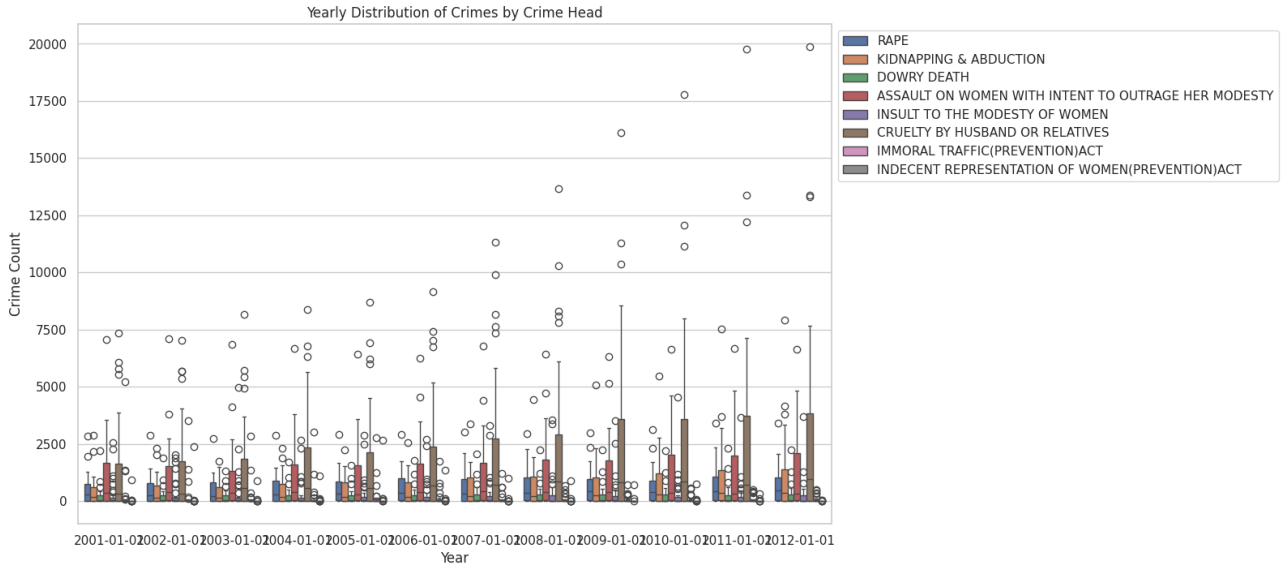


Figure 4.5: CRIME HEAD Distribution for Each Year

4.2.6 Top Crime Head for Each State

A barplot was created to show the top crime head for each state/UT, indicating the most significant crime type.

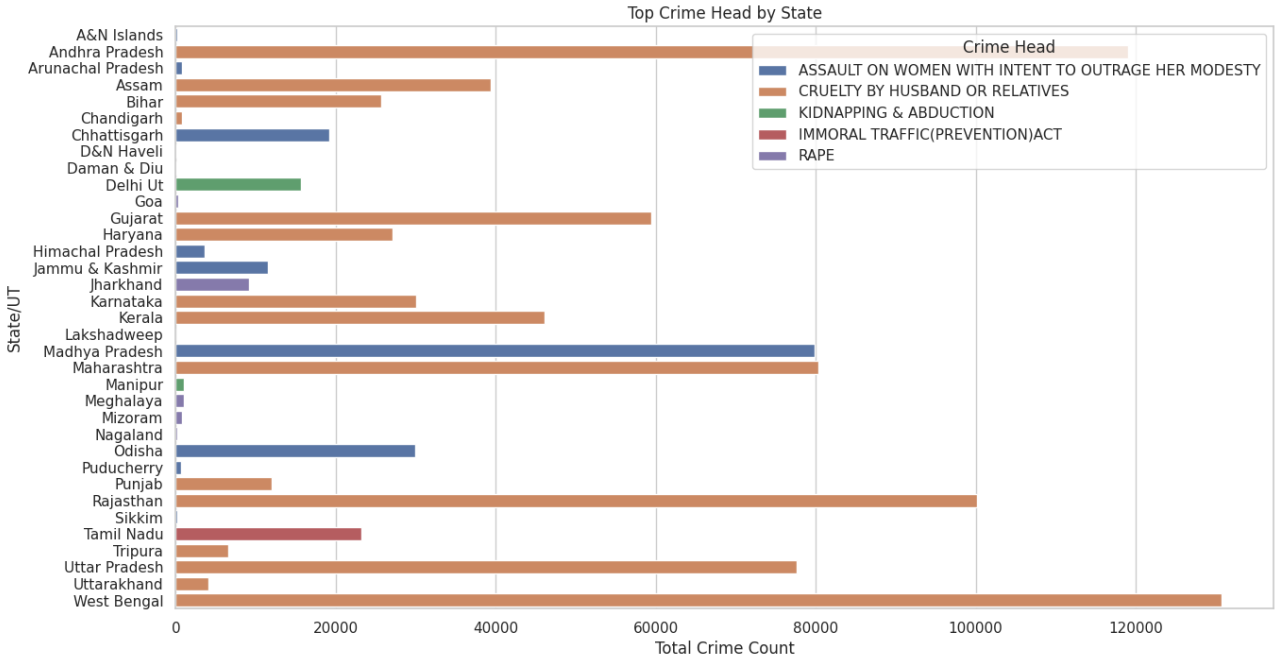


Figure 4.6: Top Crime Head for Each State/UT

4.2.7 Number of Outliers

A count of the total number of outliers identified in the dataset. A total of 422 outliers were detected in the data.

4.2.8 Capping the Outliers

Outliers were capped using the 1.5*IQR rule. This process replaces extreme values that exceed the boundaries of Q1 $-$ 1.5 $\times$ IQR (lower bound) and Q3 $+$ 1.5 $\times$ IQR (upper bound) with these respective limits. By capping outliers, the data retains its overall structure while minimizing the influence of extreme values.

4.2.9 Outlier Treatment and Visualization

Outliers were identified and treated to improve the data quality. The boxplot below shows the data before and after outlier treatment.

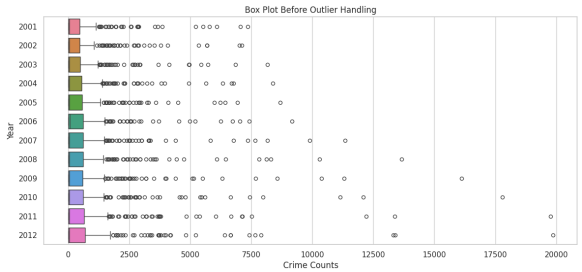


Figure 4.7: Boxplot Before Outlier Treatment

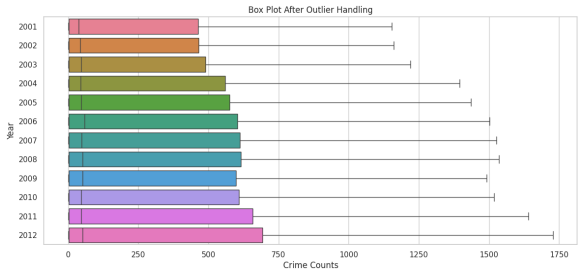


Figure 4.8: Boxplot After Outlier Treatment

4.2.10 Data Normalization

Normalization helps bring the data to a common scale and facilitates better comparison and analysis. The data was normalized using Z-scores to standardize the mean and variance.

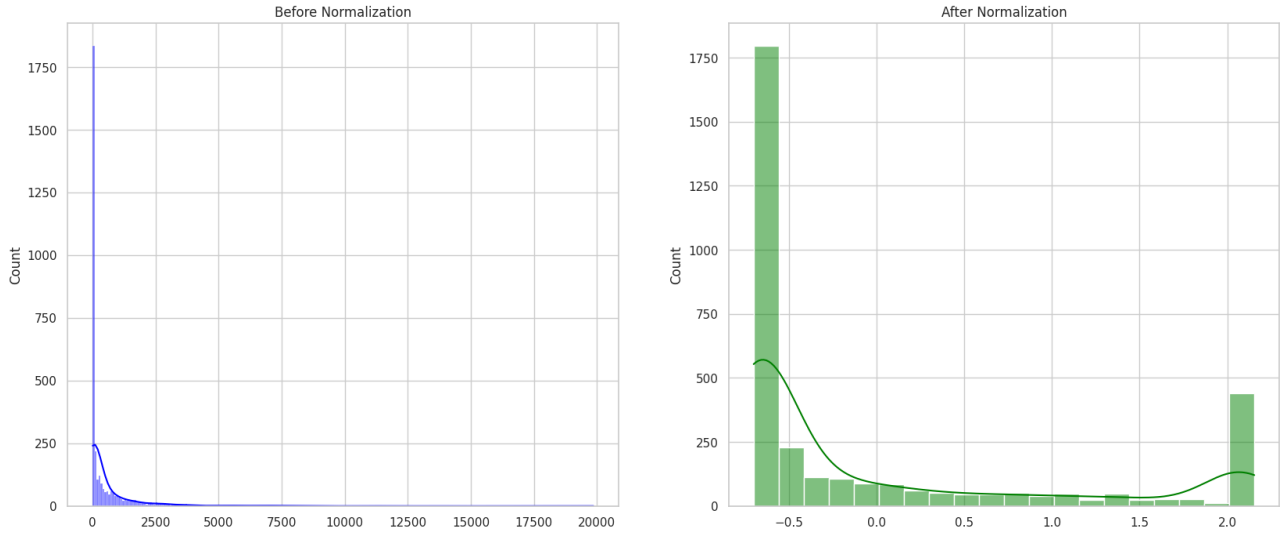


Figure 4.9: Effect of Normalization on Data (Before v/s After)

Chapter 5. Model fitting

Model fitting is the process of training a statistical model to identify the relationship between input variables and output responses using historical data. It involves adjusting the model's parameters to minimize errors between predictions and actual values, often by optimizing a loss function like mean squared error.

5.1 Regression

Regression is a statistical method used to predict or estimate the relationship between variables.Here, we use regression to understand how crimes against women have changed over time and predict future trends to help policymakers and law enforcement in resource allocation and policy adjustments.

5.1.1 Linear Extrapolation

Linear regression is the first step to linear extrapolation. It fits a straight line to the data based on the linear relationship between variables. Extrapolations are predictions based on extending this line beyond the range of the data observed. This is a straight application of the linear regression function for predicting future values.

Extrapolated Crime Trends (2013-2017)

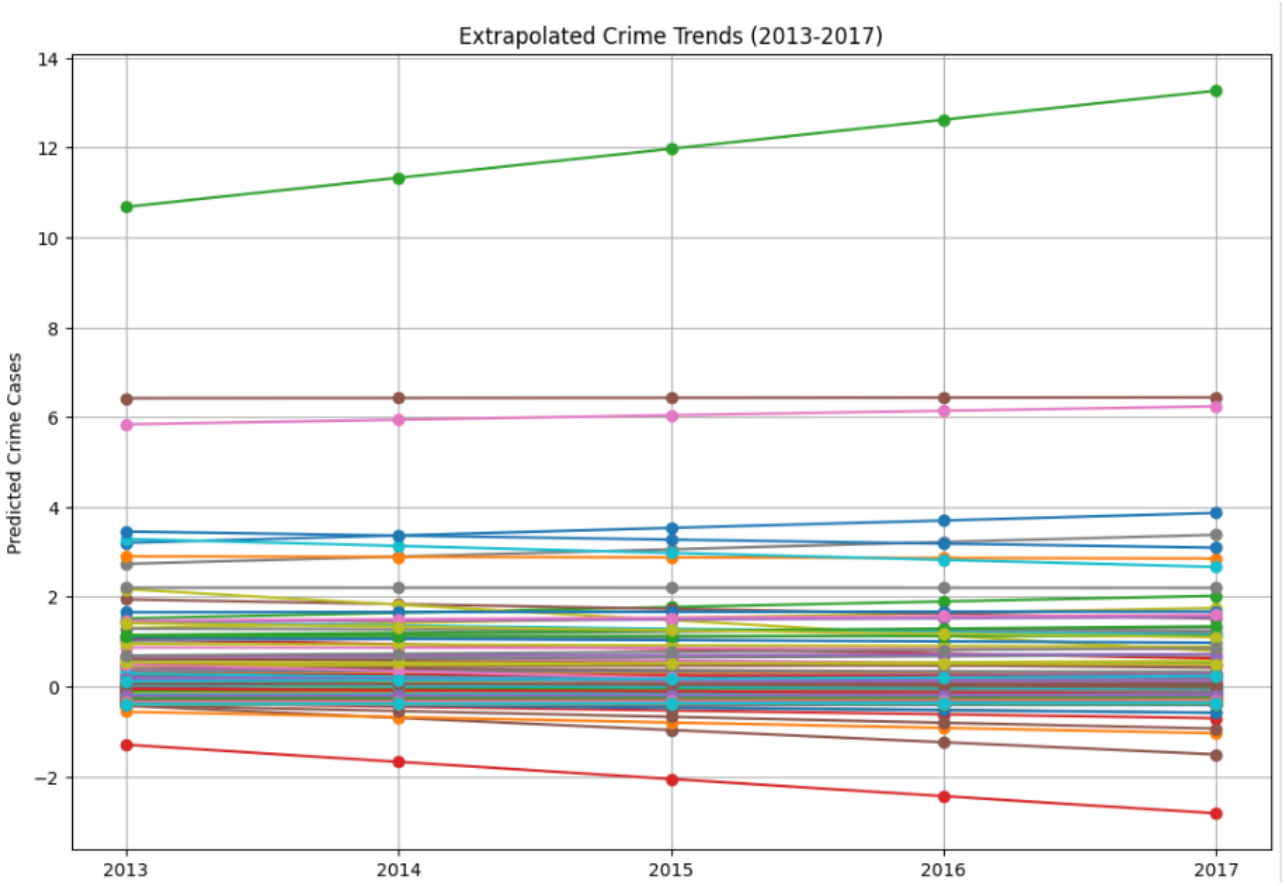


Figure 5.1: Extrapolated Crime Trends (2013-2017)

Extrapolated Crime Trends (2013-2017) (Normalized)

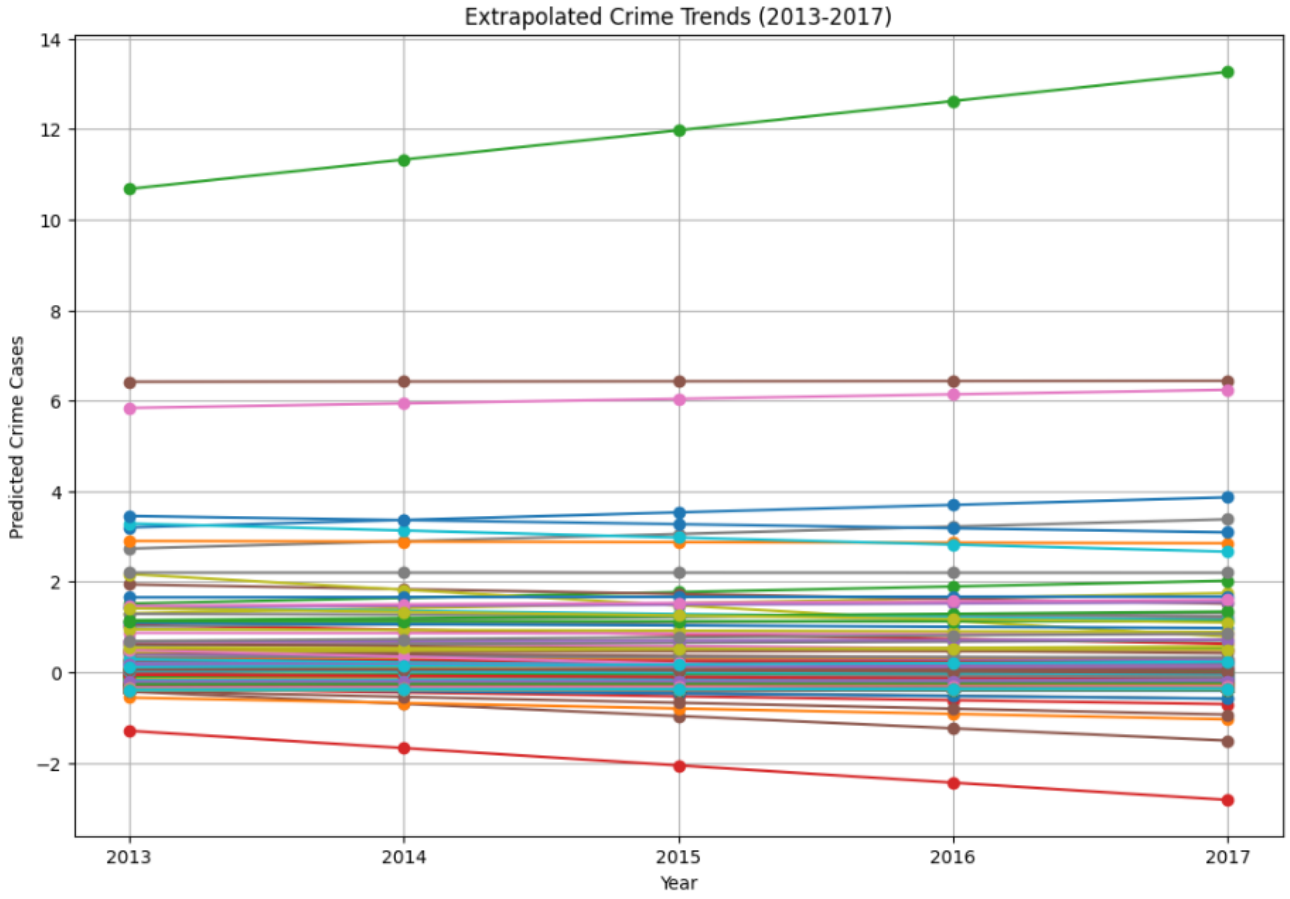


Figure 5.2: Extrapolated Crime Trends (2013-2017) (Normalized)

State-wise Predicted Crime Rates (2013-2017)

The chart shows predicted crime rates in India across the states from 2013 to 2017. Each line represents total crimes predicted for a particular state, highlighting the trend over the years. Some show steady increases, such as those of Maharashtra and Andhra Pradesh, while others remain consistent. This visualization helps make comparisons and analyze crime trends over time across states.

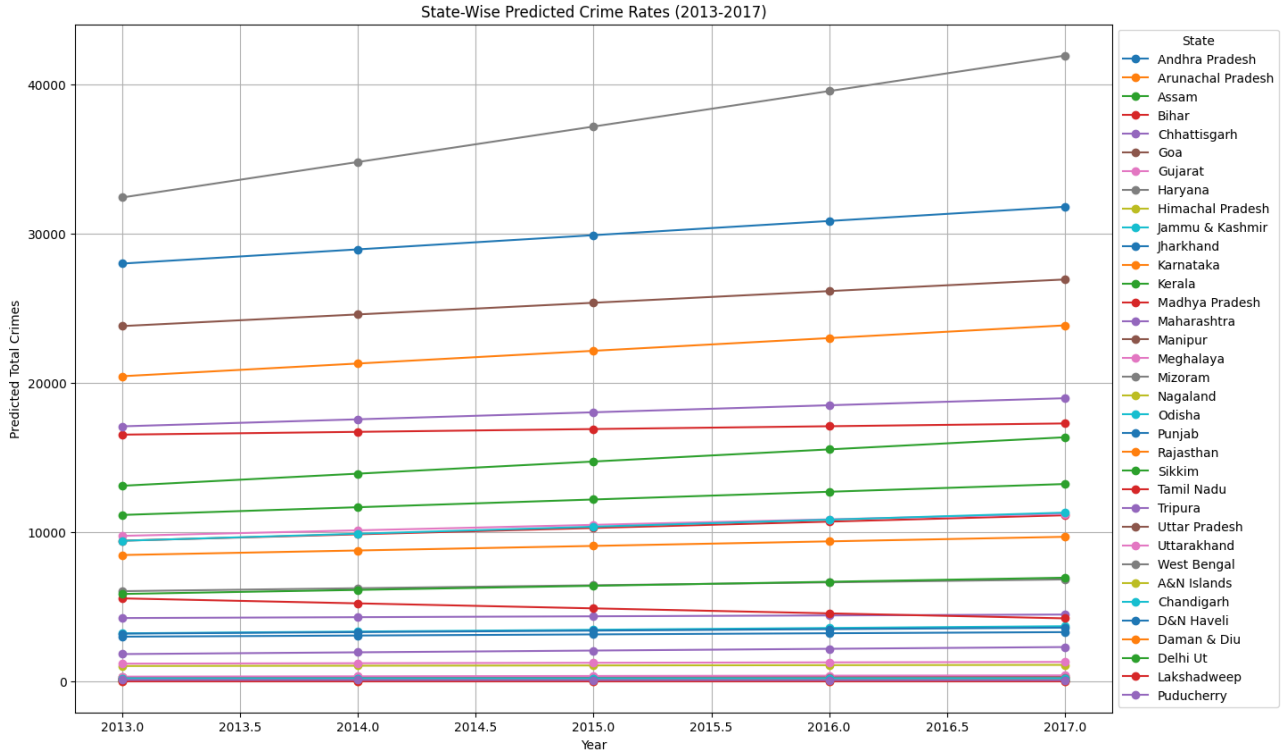


Figure 5.3: State-wise Predicted Crime Rates (2013-2017)

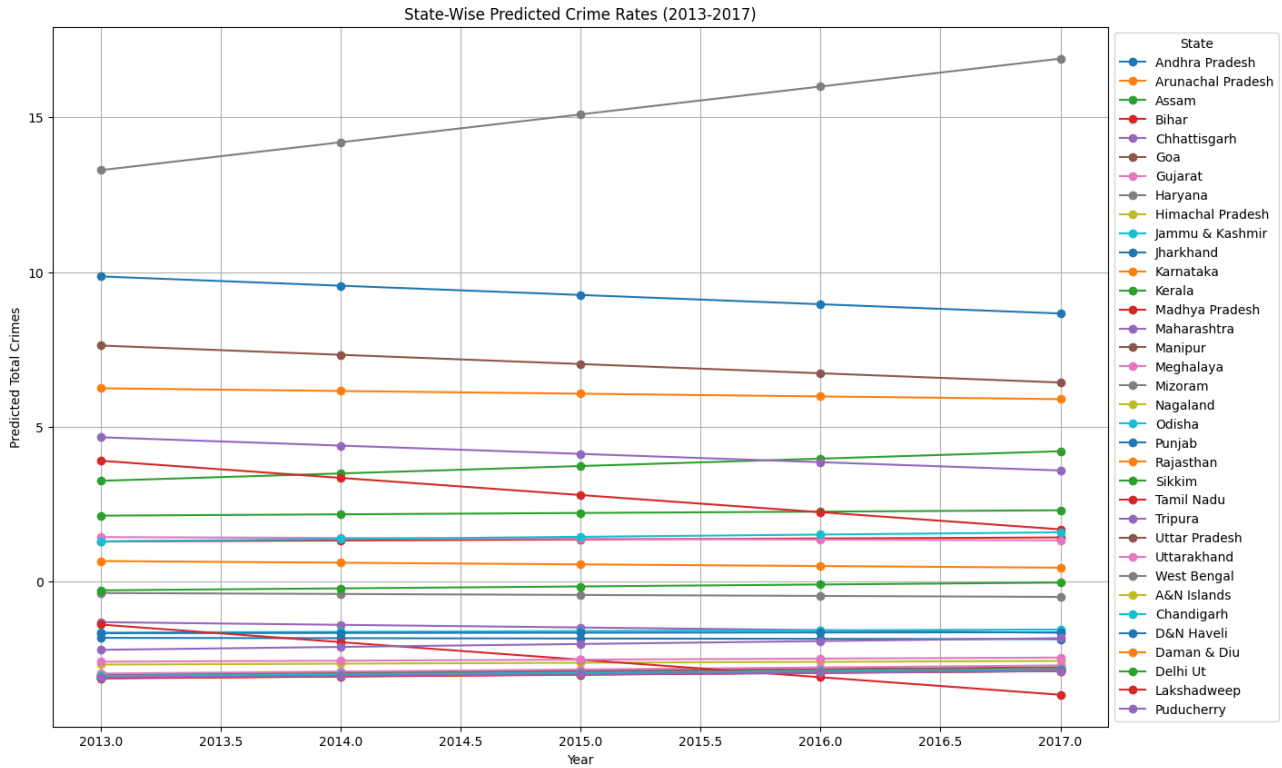


Figure 5.4: State-wise Predicted Crime Rates (2013-2017) (Normalized)

Extrapolated Crime Head Trends (2013-2017)

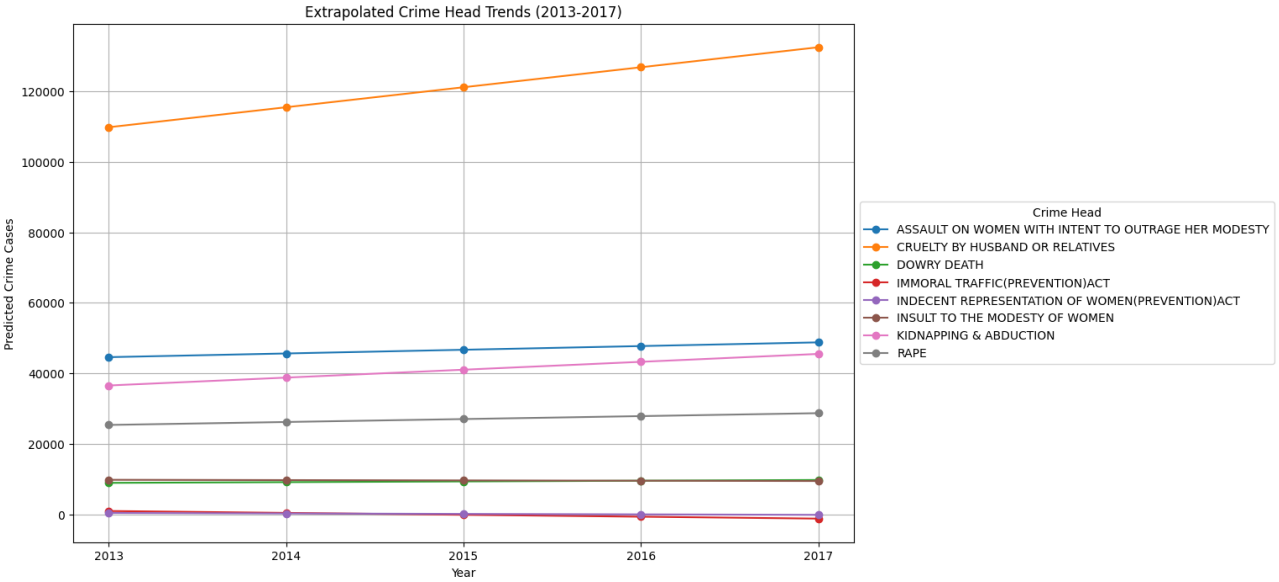


Figure 5.5: Extrapolated Crime Head Trends (2013-2017)

RMSE and R-squared for Each Crime Head

The RMSE chart measures the average prediction error for each crime head:

Lowest RMSE: ”Dowry Death” (0.30), indicating the model predicts this category most accurately.

Highest RMSE: ”Cruelty by Husband or Relatives” (1.27), reflecting greater prediction variability for this dominant crime.

Other notable errors include ”Kidnapping and Abduction” (1.03) and ”Insult to the Modesty of Women” (0.83).

The R-squared chart shows how well the model explains the variance in the data for each crime head:

Best Fit: ”Assault on Women with Intent to Outrage Her Modesty” (0.94), indicating the model explains 94 percent of the data variation.

Lowest Fit: ”Dowry Death” (0.18), showing limited predictive accuracy for this category.

Categories like ”Kidnapping and Abduction” (0.85) and ”Cruelty by Husband or Relatives” (0.79) also show good model performance

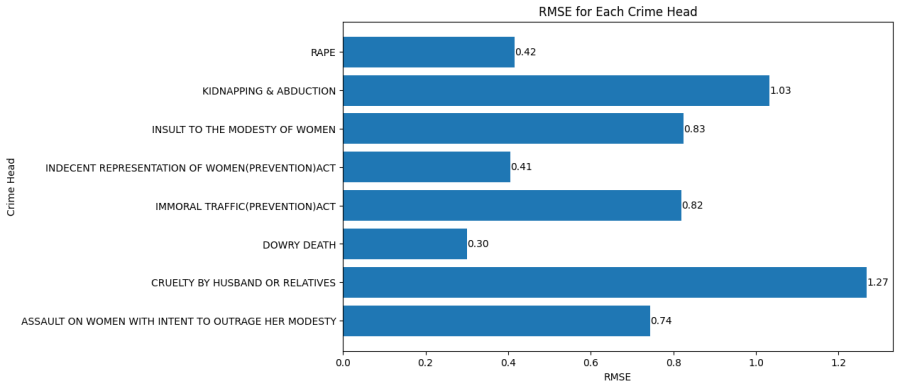


Figure 5.6: RMSE for Each Crime Head

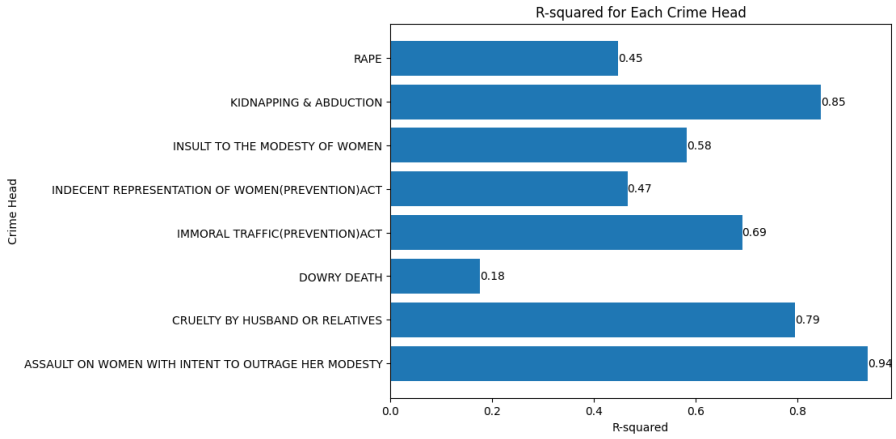


Figure 5.7: R-squared score for Each Crime Head

Residuals of Linear Regression Extrapolation for Each Crime Head

This is a chart of the residuals of linear regression extrapolation for different crime heads from the years 2001 to 2012. The residual is the difference between actual and predicted values, giving a measure of how well the model is performing. The residual of 0 on the dashed line is perfect prediction, while any deviation from it indicates under- or over-prediction. Fluctuations in the residual, therefore, point out changes in how well linear regression fits each crime head across time.

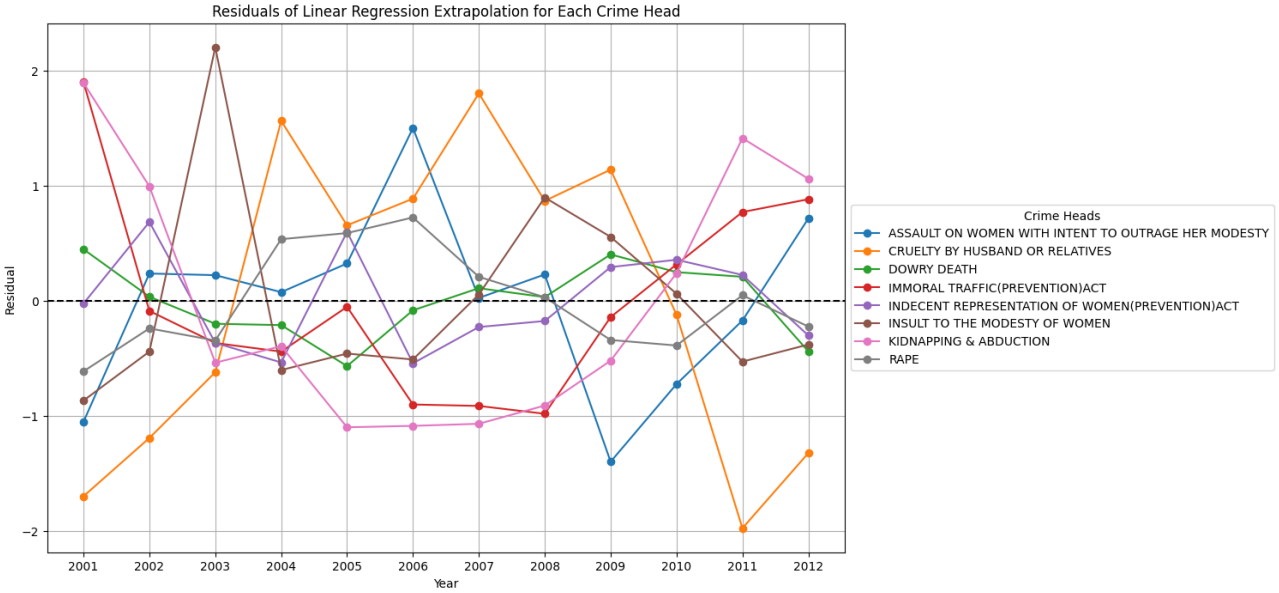


Figure 5.8: Residuals of Linear Regression Extrapolation for Each Crime Head

Historical Data with the Predicted Data (Linear Regression)

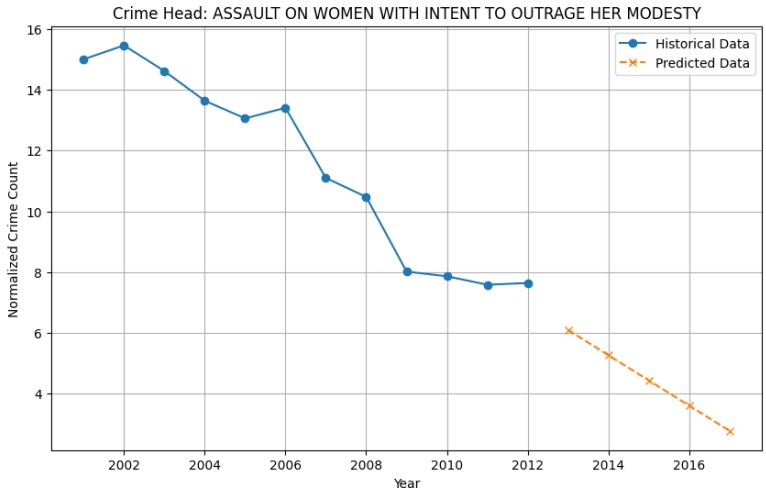


Figure 5.9: Assault on women with intent to outrage her modesty

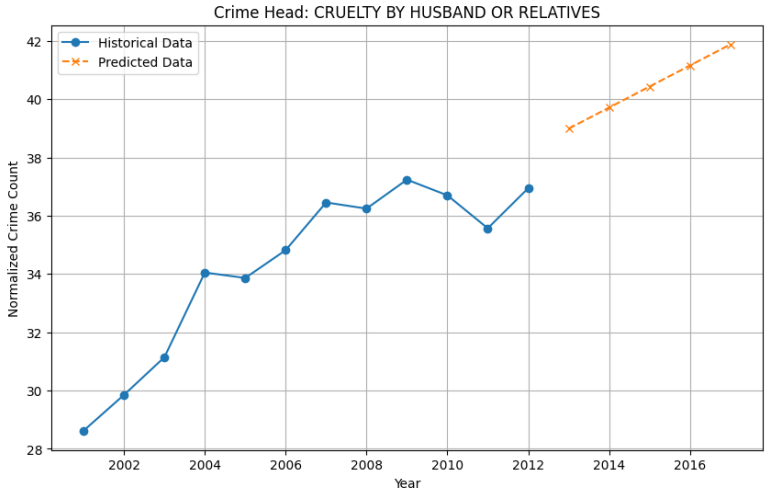


Figure 5.10: Cruelty by husband or relatives

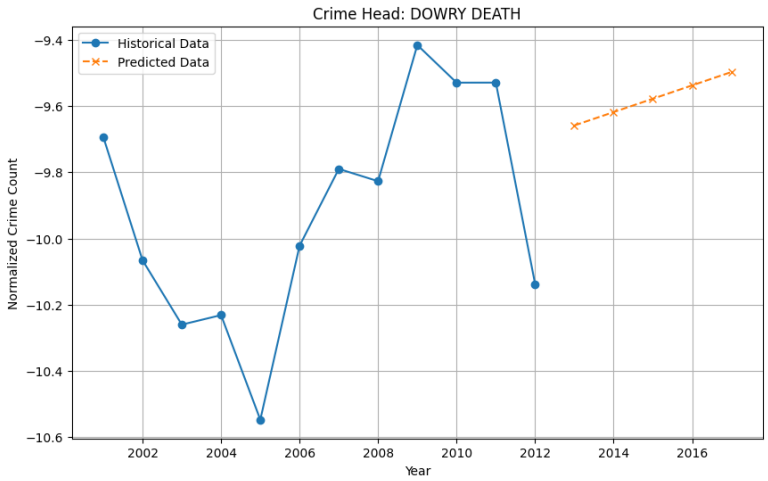


Figure 5.11: Dowry death

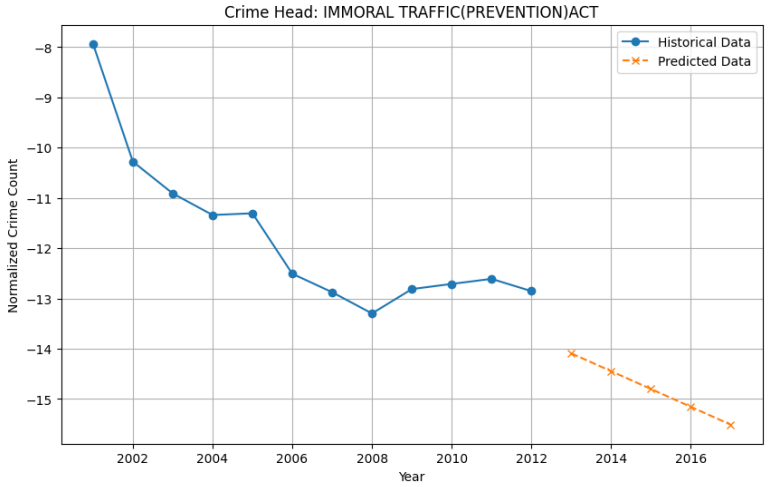


Figure 5.12: Immoral Traffic (prevention) act

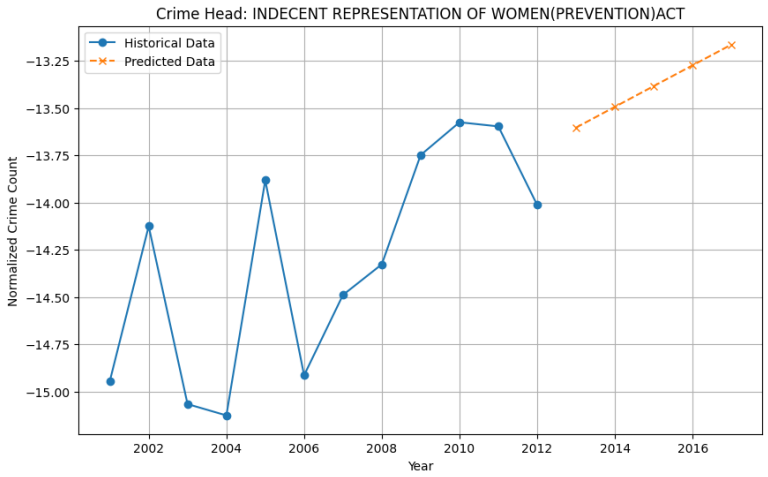


Figure 5.13: Indecent representation of women (prevention) act.

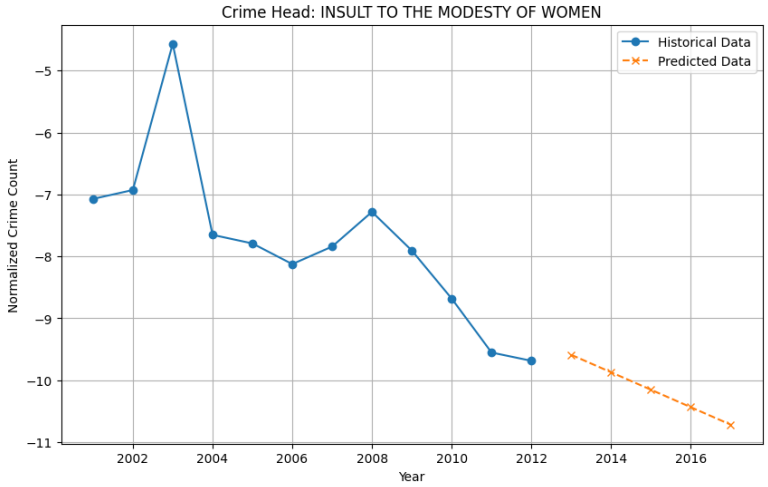


Figure 5.14: Insult to the Modesty of women

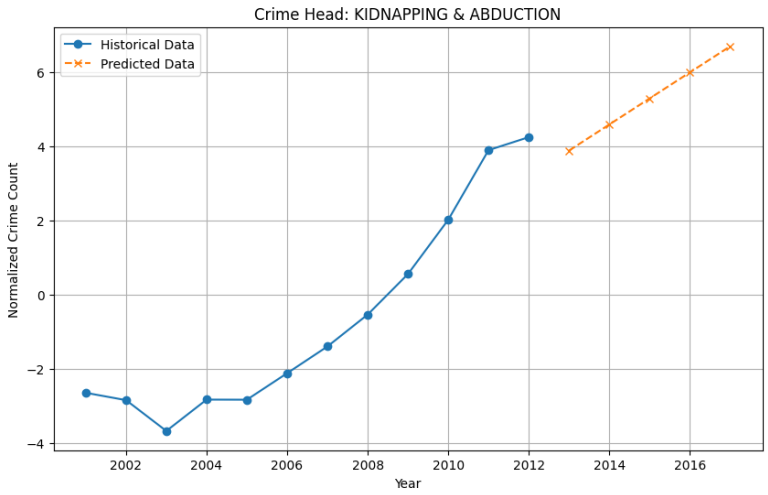


Figure 5.15: Kidnapping and Abduction

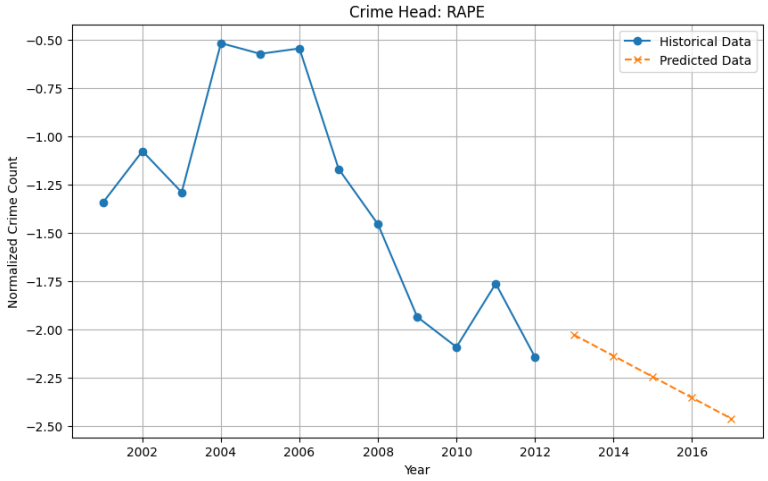


Figure 5.16: Rape

5.1.2 Polynomial Extrapolation

Polynomial Extrapolation uses polynomial regression, which is a generalized form of linear regression. Instead of fitting a straight line, it fits a polynomial curve, such as quadratic or cubic, to capture non-linear patterns. Polynomial regression includes transformation of inputs to higher-order terms, although the problem is solved by using linear regression techniques.

Here we are using *make_pipeline* with *PolynomialFeatures(degree=2) and LinearRegression()* as its arguments.

To import it: **from sklearn.pipeline import make_pipeline**

Polynomial Extrapolation of Crime Trends (2013-2017)

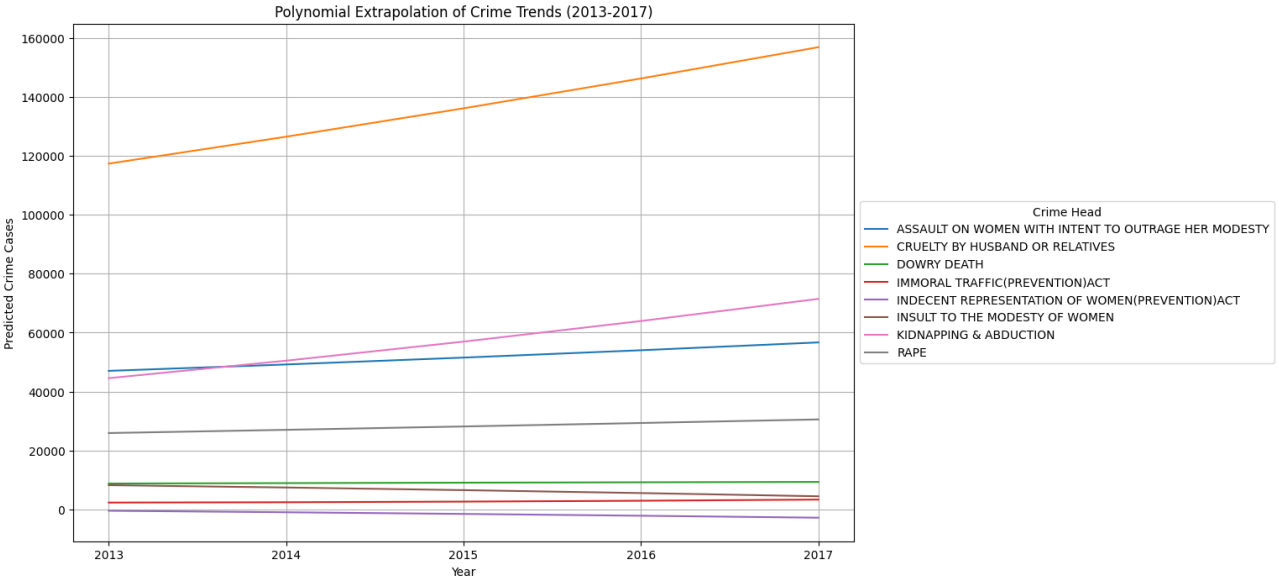


Figure 5.17: Polynomial Extrapolation of Crime Trends (2013-2017)

R-squared for Polynomial Regression

The chart illustrates R-squared values for polynomial regression for the different categories of crimes between 2001 and 2012. The higher the values, for instance, ”Rape,” ”Kidnapping and Abduction,” the model fits well with the data, hence explaining the obvious trends. Low values such as ”Insult to the Modesty of Women” would require alternative models since it doesn’t have a trend. It demonstrates how the model explains the crime trends in each category.

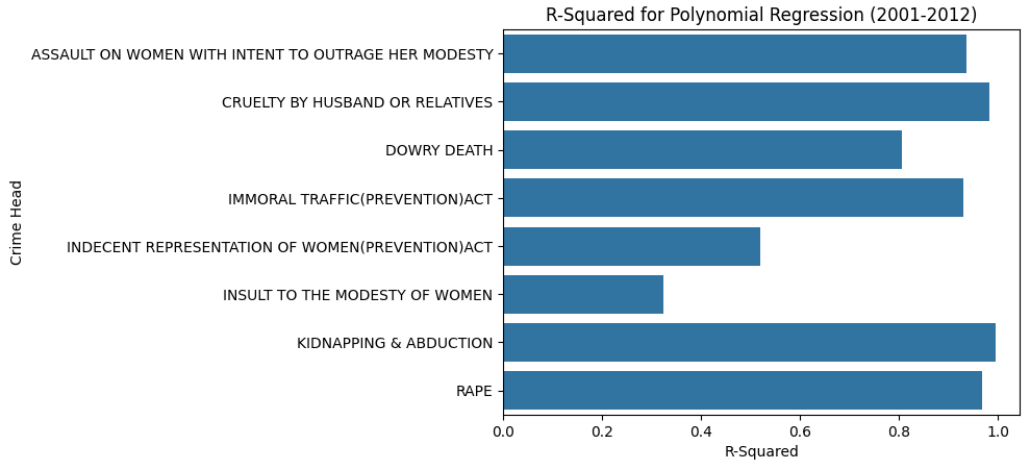


Figure 5.18: R-Squared for Polynomial Regression (2001-2012)

Historical Data with the Predicted Data (Polynomial Regression)

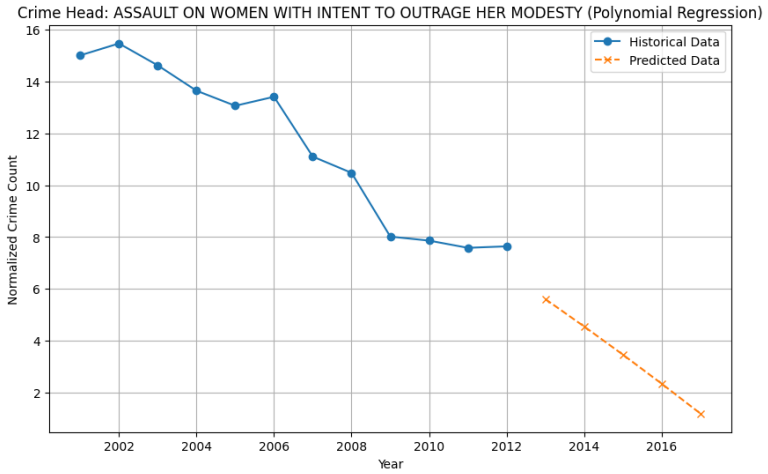


Figure 5.19: Assault on women with intent to outrage her modesty (Polynomial Regression)

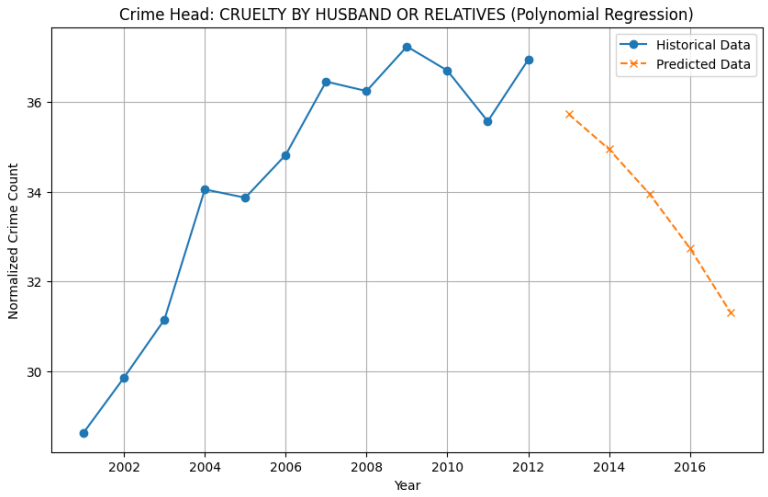


Figure 5.20: Cruelty by husband or relatives (Polynomial Regression)

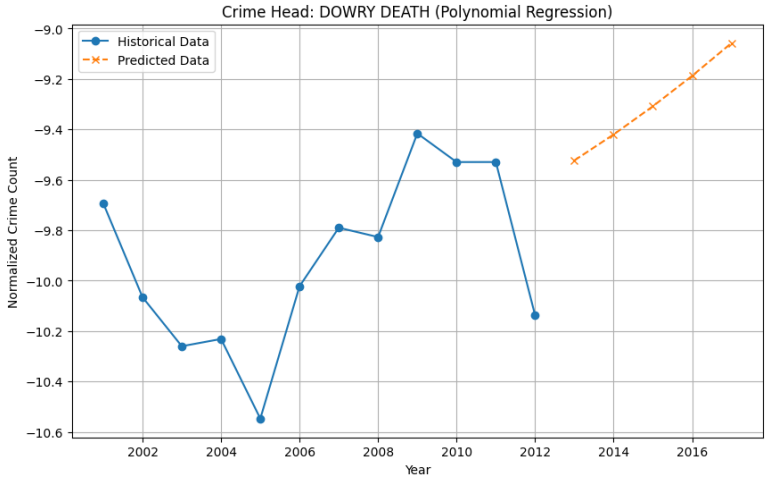


Figure 5.21: Dowry death (Polynomial Regression)

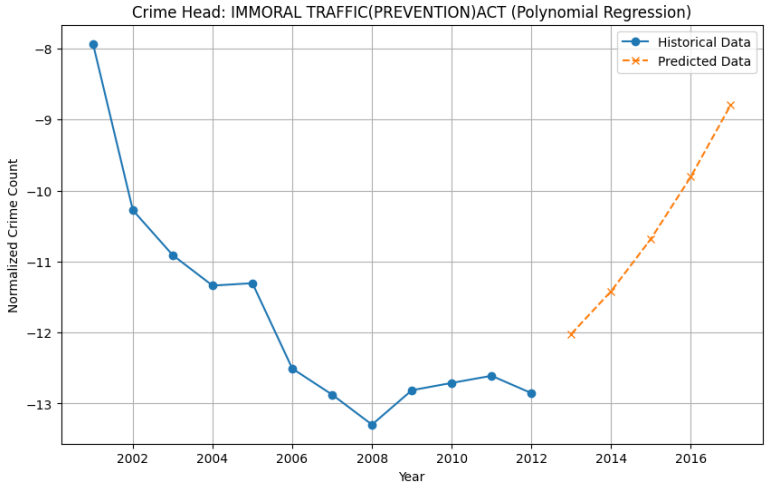


Figure 5.22: Immoral Traffic (prevention) act (Polynomial Regression)

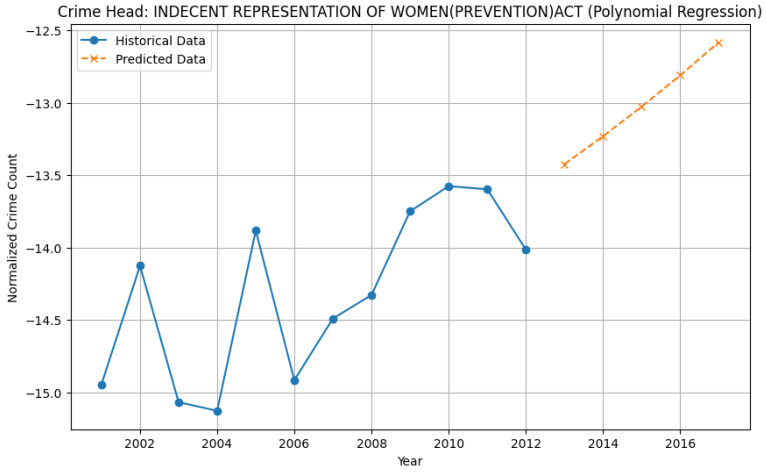


Figure 5.23: Indecent representation of women (prevention) act. (Polynomial Regression)

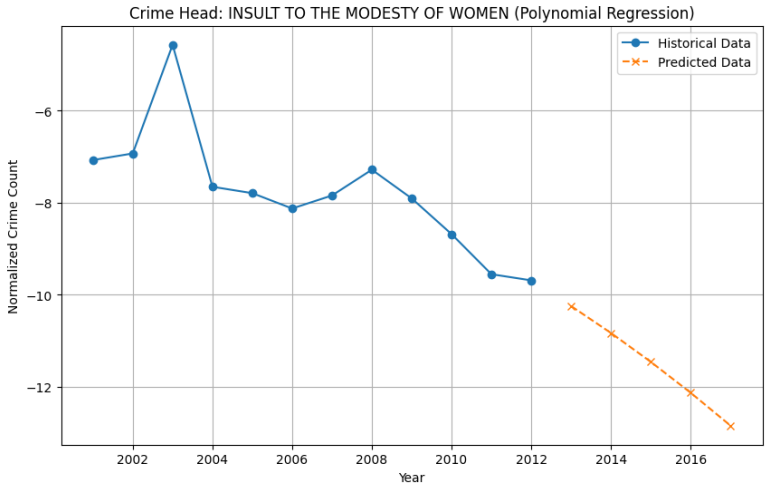


Figure 5.24: Insult to the Modesty of women (Polynomial Regression)

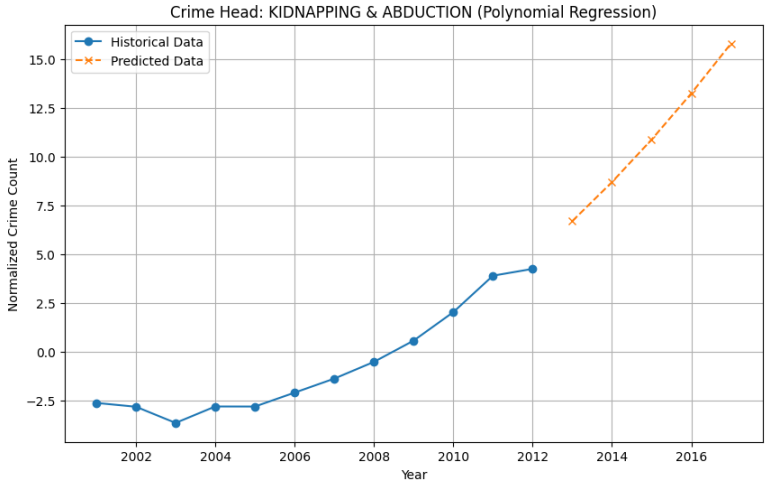


Figure 5.25: Kidnapping and Abduction (Polynomial Regression)

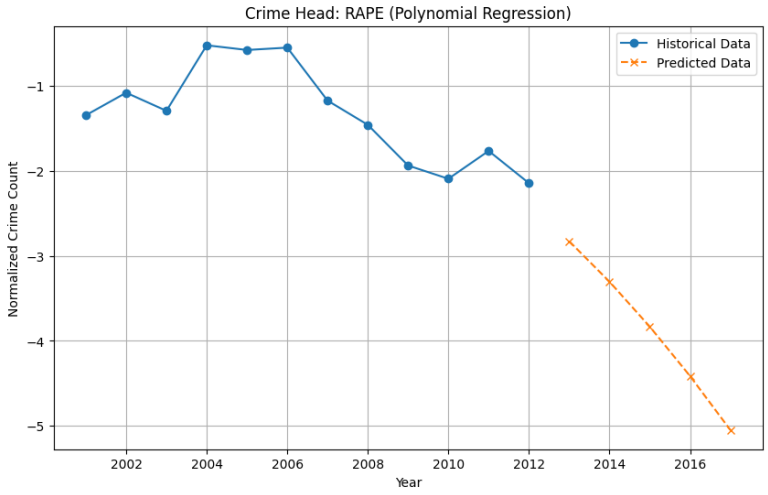


Figure 5.26: Rape (Polynomial Regression)

Comparison of Linear and Polynomial Fits (2001 - 2012)

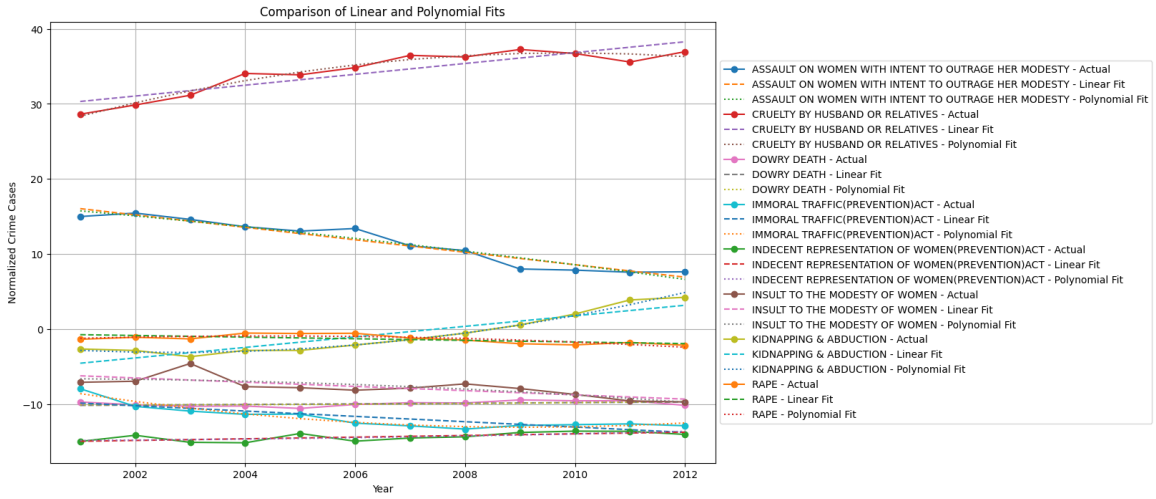


Figure 5.27: Comparison of Linear and Polynomial Fits

This graph plots actual data, fits using linear regression, and fits using polynomial regression for all categories of crimes committed against women between 2001 and 2012. Here's what it represents:

Lines and Markers

Solid lines with markers represent actual crime data for each category over time.

Dashed lines represent the linear fit capturing the overall trend with a straight line.

The polynomial regression fit is depicted as dotted lines in a non-linear fashion. Observations:

In cases of some categories such as "Dowry Death", the polynomial regression matches the actual data very well and captures variations much more than the linear regression does.

In cases of categories like "Rape", the trend looks to be consistent throughout, as both linear and polynomial fits run through it.

For very erratic trends like "Indecent Representation of Women (Prevention) Act", polynomial fits work better than linear fits that reflect data variations.

Comparison

Linear regression is apt for categories with more stable trends but may distort the data by simplifying its patterns.

Polynomial regression works much better on the data to show a high level of accuracy if the crime pattern is non-linear or variable.

Insights

Crime categories that have variations are worked well with polynomial regression in terms of the trend model.

These trends can then be utilized by policymakers to see which crimes are increasing or stabilizing and hence action taken accordingly.

Chapter 6. Conclusion & future scope

The trends and patterns of crimes against women in India from 2001 to 2012 could be useful for policy decisions and resource allocation. Linear as well as polynomial regression techniques have been used to provide deeper insights into the trajectory of various crime categories. A general overview of stable trends was provided by linear regression, whereas polynomial regression captured non-linear dynamics for more complex crime patterns. This study underlines the necessity for data-driven approaches to determine high-risk states and categories. In this regard, this can enable policymakers and law enforcement to carry out targeted interventions.

6.1 Findings/observations

Major observations emerged from the analysis of crimes against women in India during the years 2001-2012. The upward trend is quite visible in categories like "Rape" and "Kidnapping and Abduction." Polynomial regressions have brought out non-linear features of categories such as "Dowry Death" and "Assault on Women with Intent to Outrage Her Modesty," where the time aspect brings change. Conversely, however, crimes like "Indecent Representation of Women" appeared erratic, indicating the existence of external factors influencing it. If analyzed region-wise, some regions were found to continue reporting high or low trends of crime, which could necessitate region-wise strategies in itself. Overall, outcomes indicate the need for specially targeted interventions and sustained checks to attend to these alarming trends.

6.2 Challenges

Challenges Faced by Women

Women in India face a series of challenges that bar their safety and security. Crimes like domestic violence, sexual harassment, dowry harassment, and human trafficking are still prevalent as societal stigma and underreporting increase them. Cultural and structural barriers often prevent them from taking action against the violations, and inadequate law enforcement and delay in judicial process work to undermine their security. Such crimes further aggravate the vulnerability of women to such crimes because of a lack of awareness about legal rights and inadequate support systems, thereby continuing the cycle of oppression and fear.

Problems Encountered While Preparing This Report

While preparing this report, there were a number of challenges that arose in the analysis and drawing of conclusions from the data. Often, working with inconsistent data was time consuming while

trying to ensure accuracy in the outputs. The task of trying to get non-linear trends across multiple crime categories made it necessary working with different regression models to provide reliable results. Another significant challenge was trying to convey meaningful insights visually by representing such data for multiple states and multiple crime categories.

6.3 Future plan

In the future, this study can be improved by including more recent data to track crime trends after 2012 and evaluate the accuracy of predictions. Adding socio-economic and geographic factors can help understand the reasons behind crime trends in greater detail. Advanced methods like time-series models and machine learning can improve prediction accuracy. More over, interactive visualizations and dashboards can enable stakeholders to monitor crime trends in real time, helping them take proactive steps to enhance women's safety in India.

Group Contribution

Zeel Ghori

Code: 60% Presentation: 25% Report file: 15%

Aryankumar Panchasara

Code: 15% Presentation: 60% Report file: 25%

Nischay Agrawal

Code: 25% Presentation: 15% Report file: 60%

Short Bio

1. **Zeel Ghori** I am a dynamic and passionate individual who thrives on challenges and opportunities to make a difference. As a dedicated member of the Student Body Government and Academic Committee, I take pride in my ability to lead, inspire, and work collaboratively to bring positive changes to my community.

My strengths lie in problem-solving, decision-making, and public speaking, which I have honed through various leadership roles. Whether it's addressing student concerns, organizing events, or contributing to academic initiatives, I approach every task with focus, determination, and a commitment to excellence.

Outside my academic and leadership pursuits, I am a vibrant individual who finds joy in dancing, traveling, and embracing cultural heritage. I have a special love for Garba, a folk dance that reflects the richness of tradition and my zest for life.

Hardworking and enthusiastic, I strive to balance my passions and responsibilities, continuously learning and growing to become the best version of myself.

2. **Aryankumar Panchasara** I, a student with a strong passion for programming and web development. I enjoy creating web applications that are functional and user-friendly, focusing on writing clean and efficient code. My curiosity drives me to explore new technologies and improve my skills, staying updated with the latest trends in the tech world.

Beyond academics, I am a sports enthusiast and love playing badminton, which keeps me active and disciplined. I also enjoy drawing, as it fuels my creativity and allows me to express myself. These hobbies, along with my technical pursuits, help me maintain a balanced and dynamic lifestyle, driving both my personal and professional growth.

3. **Nischay Agrawal** is an ambitious and motivated individual currently pursuing a Master's degree in Information and Communication Technology (ICT) with a specialization in Machine Learning at DAIICT. He has a solid foundation in computer engineering from LDRP Institute of Technology and Research, where he demonstrated exceptional academic performance.

Nischay has hands-on experience in web development, evidenced by his role as a Developer Intern at Cybercom Creation. During this tenure, he excelled in PHP, MySQLi, and Magento customization, among other technologies, showcasing his ability to adapt to diverse technical environments.

He has independently led innovative projects such as an Augmented Reality-based furniture visualization app, an encrypted Android messaging platform, and a Firebase-powered advertisement portal, demonstrating his problem-solving skills and expertise in cutting-edge technologies.

In addition to his technical prowess, Nischay is a proactive member of his academic community, contributing to volunteering in hackathons, where he effectively built partnerships and conducted educational workshops.

With a passion for learning and a collaborative mindset, Nischay aspires to thrive in dynamic environments, contributing to impactful solutions in the tech industry.

References

[1] Crime against Women Dataset *URL:* https://www.data.gov.in/resource/crime-against-women-during-2001-2012

[2] Compound annual growth rate *URL:* https://en.wikipedia.org/wiki/Compound_annual_growth_rate