Crimes in India: Analyzing Patterns and Predictions (2001-2013) : Project Report

Mrunal Pravin Kulkarni (22183)
Darshana Srivathsan (22001)
Prakash Kumbhakar (22241)

DSE 315: Data Science in Practice

Contents

1 Problem Statement

Crime is a pressing issue affecting safety and stability across communities. In India, the vast geographic and socio-economic diversity complicates the understanding of crime patterns. While district-level crime data is available, challenges such as inconsistent reporting, lack of normalization (e.g., by population), and limited historical data hinder effective analysis.

Key challenges include:

- Identifying temporal and spatial trends in crime data (2001–2013).
- Understanding correlations and patterns to derive actionable insights.
- Developing reliable predictive models for crime trends and severity categorization.

This project seeks to address these challenges and provide stakeholders with tools to make data-driven decisions for resource allocation, policy-making, and crime prevention.

2 Introduction

This project focuses on analyzing crime patterns in India using a comprehensive dataset covering various types of crimes, including theft, assault, and cybercrime, across states and Union Territories. The dataset spans multiple years, enabling a detailed exploration of the evolution of crime trends over time.

By applying data analysis and visualization techniques, the project aims to:

- Examine the geographic and temporal distribution of crimes.
- Identify correlations between socio-economic factors, such as population density, literacy rates, and crime rates.
- Explore relationships between different types of crimes.

The project seeks to develop predictive models for forecasting future crime trends. Insights derived will inform law enforcement strategies and policy-making to enhance crime prevention and decision-making.

3 Data Collection and Preparation

The dataset is structured with the following columns:

- **State/UT**: The name of the state or union territory.
- **District**: The specific district within the state/UT.
- **Year**: The year in which the crimes were recorded.
- **Crime Types**: Various crime categories such as murder, attempt to murder, culpable homicide, kidnapping, rape, robbery, burglary, and more.

The dataset comprises 30 features representing various crime types and contains a total of 9,397 data points across different districts and years.For each district, the dataset also includes a total of Indian Penal Code (IPC) crimes, summarizing all the crimes committed in that area. This comprehensive data serves as the basis for in-depth analysis, hypothesis testing, and correlation analysis to identify patterns, trends, and contributing factors related to crime across India.

4 Exploratory Data Analysis (EDA)

4.1 Data Cleaning and Standardization

The initial dataset required minor cleaning and formatting adjustments:

- **State/UT Column**: All values were converted to uppercase to standardize the formatting, as many entries were in lowercase or mixed case.
- **District Column**: Rows where the district name was `"Total"` or `"ZZ Total"` were removed, as these entries represented aggregated data for the entire state and were mistakenly included as individual districts.

Null Value Analysis

No null values were found in the dataset. This ensured the integrity of the data for further analysis.

Final Dataset

The cleaned data was saved in a new CSV file named `crimes_cleaned.csv` for further analysis.

4.2 Visualization and Trend Analysis

The Exploratory Data Analysis (EDA) phase allowed us to uncover key patterns and relationships within the crime data. We started by calculating basic descriptive statistics to understand the central tendency, spread, and distribution of the data. This initial analysis helped us familiarize ourselves with the dataset and identified potential issues such as missing values or outliers.

One of the primary visualizations we created was a series of box plots for various crime types, such as `MURDER`, `RAPE`, and `TOTAL IPC CRIMES`. The box plots revealed the distribution of crime rates, highlighting potential outliers and the variability in crime rates across different districts. For MURDER, the majority of districts have fewer than 100 cases annually.For RAPE, most districts report fewer than 100 cases annually.TOTAL IPC CRIMES are mostly below 10,000 per district annually.

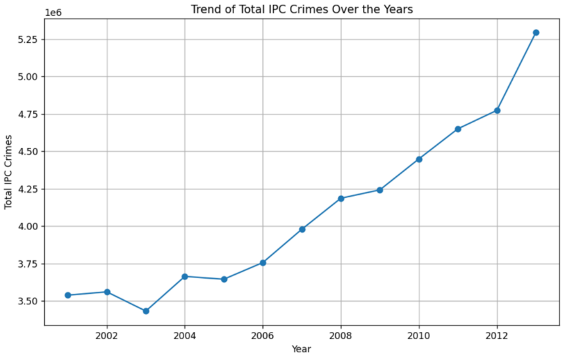


Figure 1: Total Crime Trend Over the Years

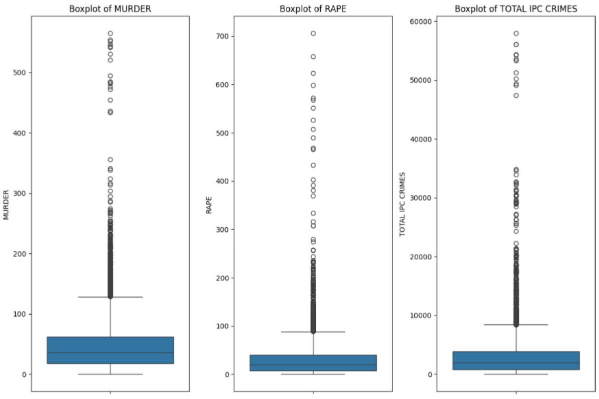


Figure 2: Boxplots for Murder , Rape and Total IPC Crimes

Next, we visualized the `TOTAL IPC CRIMES` over the years, revealing an increasing trend in the total number of crimes across India. This trend indicated that crime rates had generally risen over the analyzed period, suggesting the need for more focused crime prevention and law enforcement efforts. We also visualized the top 10 districts by crime, which provided insight into which regions faced the most significant crime challenges. This was useful in identifying crime hotspots, and in some cases, a correlation between urbanization and crime concentration was observed.

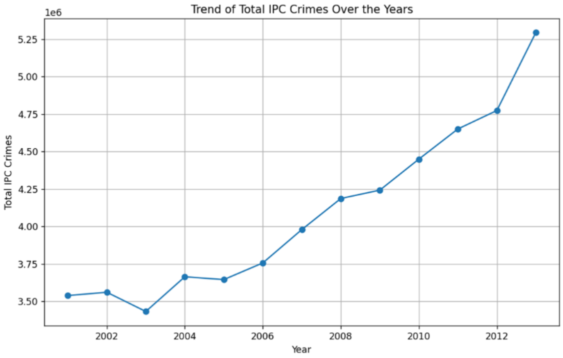


Figure 3: Total Crime Trend Over the Years

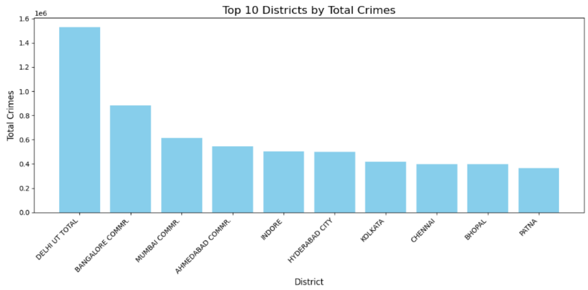


Figure 4: Top 10 Districts by Crime

To better understand the distribution of crimes over time, we visualized the proportion of crimes by type using a pie chart. This pie chart showed the relative contribution of each crime type to the total crime count, providing an overview of the crime landscape in India. We also created a visualization of total crimes by state, which illustrated that certain states, like Uttar Pradesh and Maharashtra, consistently reported higher crime figures, potentially due to their larger populations and urbanization levels.

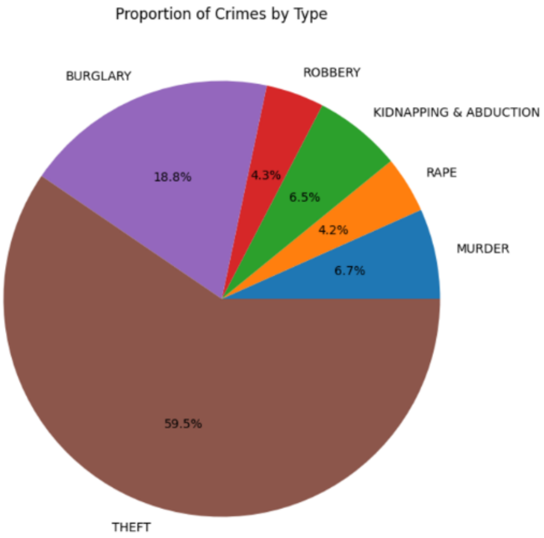


Figure 5: Proportion of Crime by Type

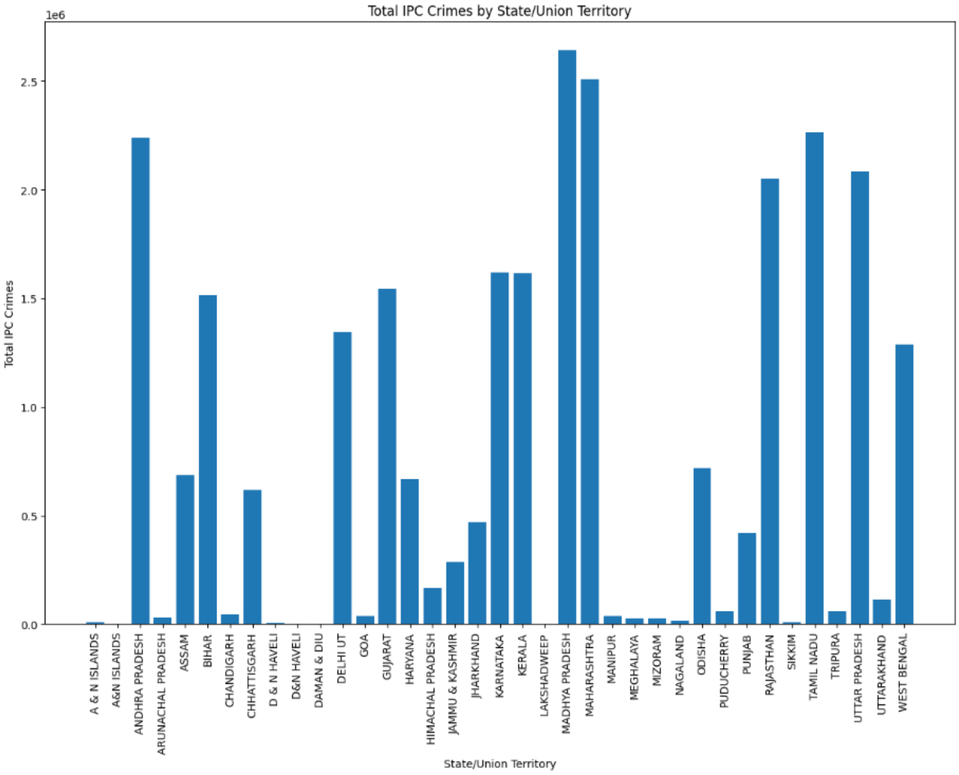


Figure 6: Total Crimes : State Wise

Finally, we explored the relationship between `THEFT` and `BURGLARY` using a scatter plot. This plot revealed a positive correlation between theft and burglary incidents, suggesting that areas with higher theft rates also tended to experience higher burglary rates. Such insights were crucial for understanding the dynamics between different types of crime and provided a foundation for future predictive analysis.

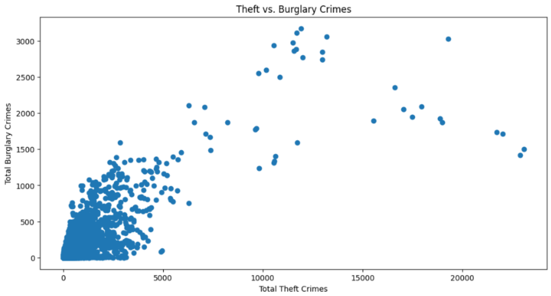


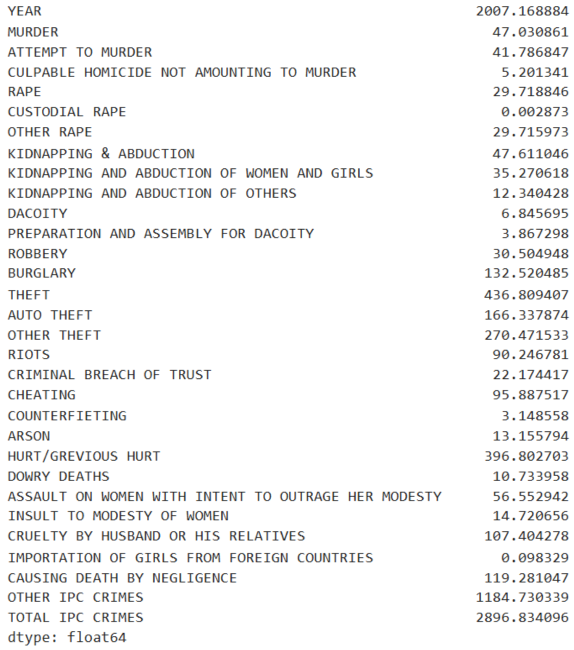
Figure 7: Theft vs Burgalary : Scatter Plot

These EDA steps allowed us to generate meaningful insights into crime patterns and trends, and served as the foundation for further statistical and machine learning analysis. The visualizations also played a crucial role in communicating our findings clearly and effectively.

5 Statistical Analysis

In this section, we performed various statistical analyses to better understand the dataset. The key statistics computed for each feature included:

- **Mean and Median:** The mean and median were calculated for all crime-related features to identify central tendencies. The mean provided a general average across all data points, while the median helped in understanding the data's distribution, especially in the presence of outliers.



```
YEAR                                                   2007.168884
MURDER                                                   47.030861
ATTEMPT TO MURDER                                        41.786847
CULPABLE HOMICIDE NOT AMOUNTING TO MURDER                 5.201341
RAPE                                                     29.718846
CUSTODIAL RAPE                                            0.002873
OTHER RAPE                                               29.715973
KIDNAPPING & ABDUCTION                                   47.611046
KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS              35.270618
KIDNAPPING AND ABDUCTION OF OTHERS                       12.340428
DACOITY                                                   6.845695
PREPARATION AND ASSEMBLY FOR DACOITY                      3.867298
ROBBERY                                                  30.504948
BURGLARY                                                132.520485
THEFT                                                   436.809407
AUTO THEFT                                              166.337874
OTHER THEFT                                             270.471533
RIOTS                                                    90.246781
CRIMINAL BREACH OF TRUST                                 22.174417
CHEATING                                                 95.887517
COUNTERFIETING                                            3.148558
ARSON                                                    13.155794
HURT/GREVIOUS HURT                                      396.802703
DOWRY DEATHS                                             10.733958
ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY      56.552942
INSULT TO MODESTY OF WOMEN                               14.720656
CRUELTY BY HUSBAND OR HIS RELATIVES                     107.404278
IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES               0.098329
CAUSING DEATH BY NEGLIGENCE                             119.281047
OTHER IPC CRIMES                                       1184.730339
TOTAL IPC CRIMES                                       2896.834096
dtype: float64
```

Figure 8: Means of Features

- **Interquartile Range (IQR):** The IQR was computed to measure the spread of the middle 50% of the data, which helped in identifying potential outliers and understanding data variability.
- **Correlation Analysis:** We conducted a correlation analysis to identify relationships between different crime types and socio-economic factors. This analysis helped us understand how different crime types are related to each other and to other variables such as population density, literacy rates, and others. See Figure 9 below.

The results of the statistical analysis revealed important insights into the distribution and relationships between different types of crimes. These insights informed further steps in the analysis, including geospatial and machine learning model development.

Objective

The aim of the T-test was to determine if there was a significant difference in theft rates between Arunachal Pradesh and all other states combined. This analysis helps evaluate whether Arunachal Pradesh's theft rates deviate from the national average.

Results

The statistical analysis was performed with the following results:

- **State**: Arunachal Pradesh
- **t-statistic**: -3.6308
- **p-value**: 0.00028

Conclusion

Based on the analysis:

- The null hypothesis was rejected.
- A significant difference was found between theft rates in Arunachal Pradesh and all other states combined.

This indicates that theft rates in Arunachal Pradesh deviate significantly from the national average, warranting further investigation into regional factors contributing to this difference.

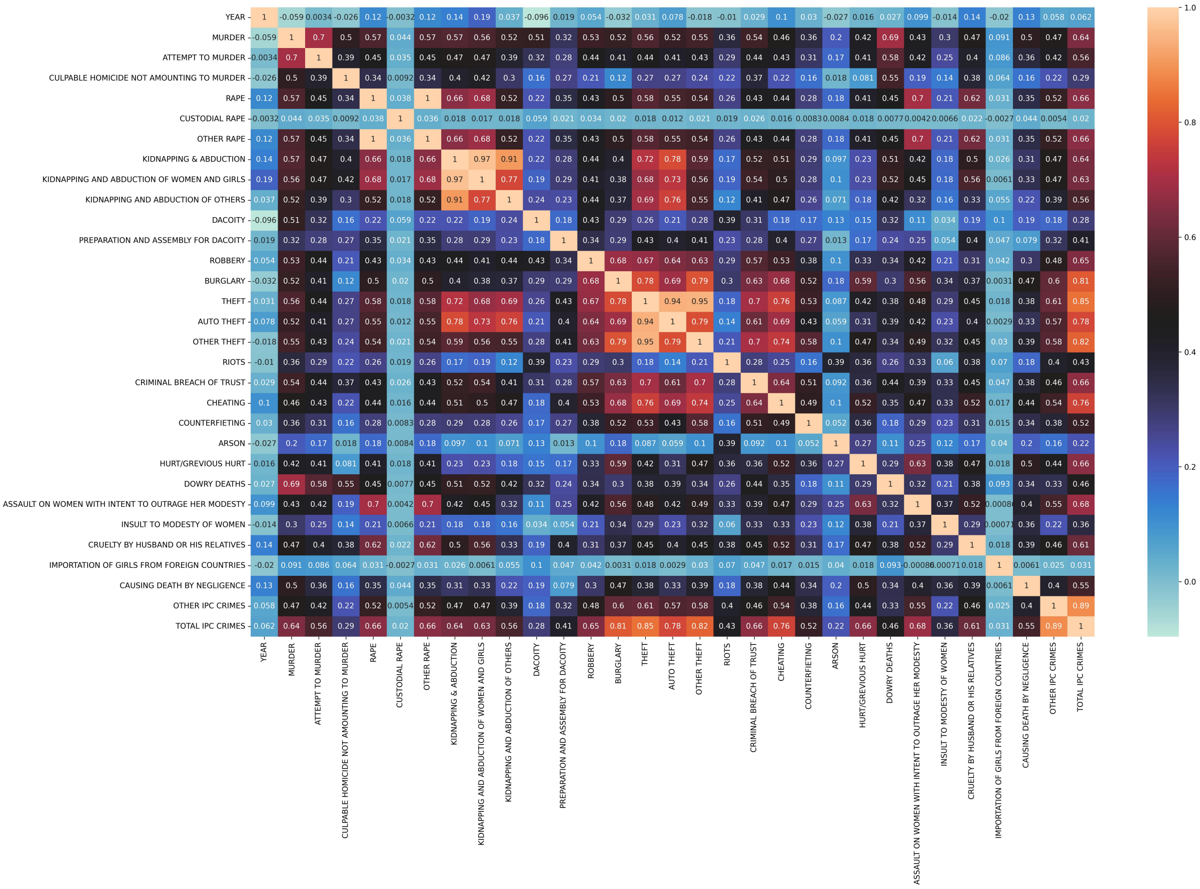


Figure 9: Correlation Analysis Between Features

6 Geospatial Analysis

6.1 Why Geospatial Analysis?

Geospatial analysis is a critical component of the project, leveraging spatial data to uncover patterns, relationships, and insights that inform decision-making. By integrating geographic information systems (GIS) and advanced mapping techniques, this analysis enables the visualization of project-related data in a spatial context.

Key aspects of geospatial analysis include:

- Identifying spatial trends.
- Optimizing resource allocation.
- Enhancing predictive models.

The insights derived from geospatial analysis support more effective planning, improved accuracy, and a deeper understanding of the project's environmental, social, and economic impacts.

Choropleth Map

The choropleth map provides a visual representation of the total number of crimes distributed across districts, using color gradations to indicate varying crime levels. Darker shades highlight districts with higher crime rates, while lighter shades represent areas with fewer incidents. This visualization helps:

- Identify geographic patterns and hotspots of criminal activity.
- Offer valuable insights for law enforcement and policymakers to target interventions and allocate resources effectively.

Bubble Map

The bubble map offers a dynamic visualization of theft cases across various locations, using bubble size to represent the volume of theft incidents in each area. Larger bubbles indicate areas with higher occurrences of theft, while smaller bubbles highlight regions with fewer cases.

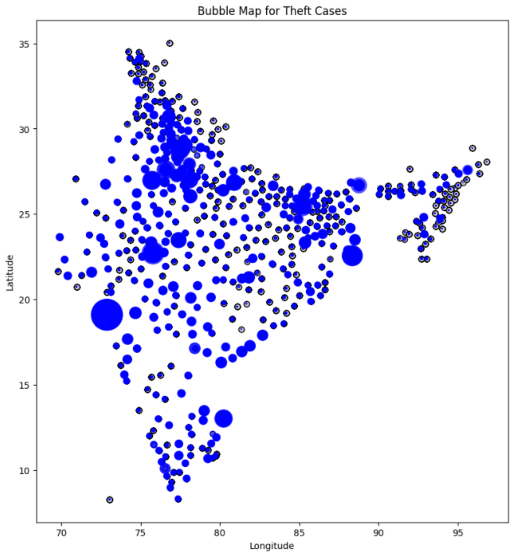


Figure 10: Bubble Plot : Theft Cases

6.2 High Murder Rate Analysis

To identify regions with a high murder rate, the mean number of murder cases across all districts was calculated as a threshold. Districts with murder rates exceeding this threshold were classified as high murder zones. This approach ensures:

- Adaptability to the overall distribution of data.
- Statistically robust identification of areas with significantly higher crime rates.

By focusing on these zones, priority areas for immediate attention and further analysis were identified. The filtered data forms the basis for subsequent clustering and visualization.

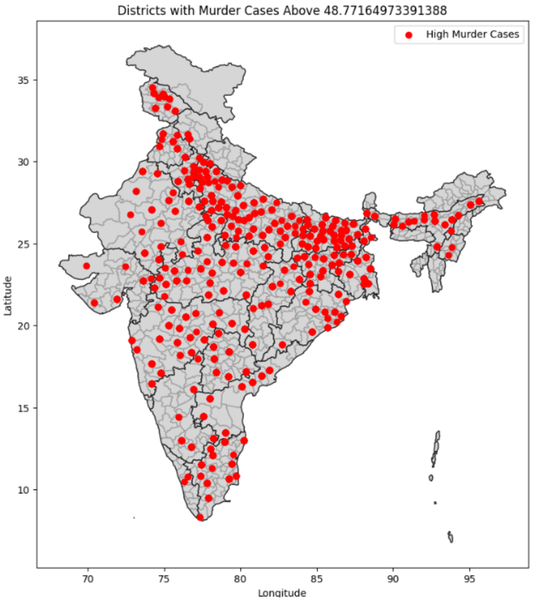


Figure 11: High Murder Rate Areas

6.3 Clustering with K-Means

To explore geographic patterns in districts with high murder rates, K-means clustering was employed with $k = 15$. Each cluster represents a group of districts with spatial proximity. Using the convex hull method:

- The borders of each cluster were outlined to highlight their geographical extent.
- The clusters and their boundaries were visualized on a map.

This method provides a clear representation of crime hotspots, aiding targeted policy-making and resource allocation.

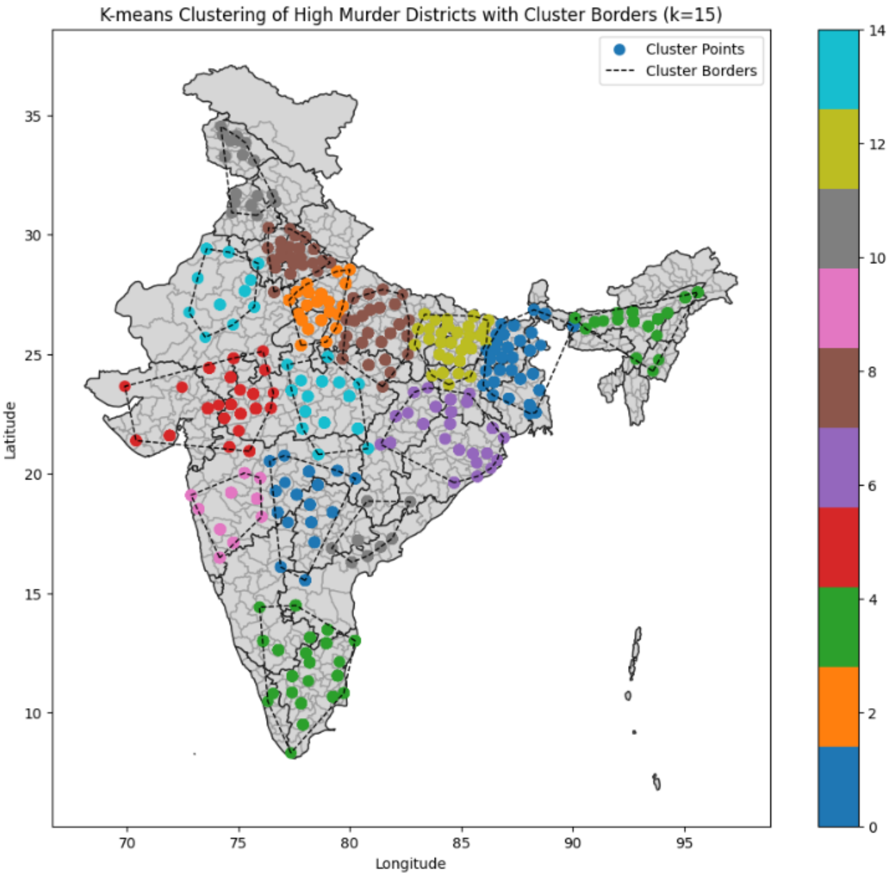


Figure 12: Clustering of High Murder Rate Areas

7 Machine Learning Models

Overview

We implemented various machine learning models to achieve three distinct objectives related to crime prediction and classification.

7.1 Aim 1: Predicting the Total Number of Crimes Based on All Features

Process

We aimed to predict the total number of IPC crimes based on all available features. Three models were evaluated:

- **Linear Regression**: Baseline model selected due to the near-linear relationship between Total IPC Crimes and other features.
- **K-Nearest Neighbors (KNN)**: Used to capture local, non-linear patterns in the data.
- **Gradient Boosting**: Leveraged for its ability to optimize predictions iteratively, capturing both linear and non-linear relationships.

Results

Model	**Mean Squared Error (MSE)**	**R^2 Score**
Linear Regression	1.171	0.999
K-Nearest Neighbors	98,954.70	0.9931
Gradient Boosting	65,607.70	0.9954

Table 1: Model Performance for Predicting Total IPC Crimes Using All Features

Key Observations

- Linear Regression performed exceptionally well ($R^2 = 0.999$), validating the hypothesis of a strong linear relationship.
- Gradient Boosting achieved the second lowest MSE and high R^2, balancing accuracy and generalization.
- KNN performed strongly but showed a higher MSE, indicating limitations in capturing global patterns.

7.2 Aim 2: Predicting Total Number of Crimes Using Top 5 Features

Process

To simplify the model while retaining interpretability, the top 5 features were selected based on their covariance with the target variable (Total IPC Crimes). The same models (Linear Regression, KNN, and Gradient Boosting) were applied to the reduced dataset.

Results

Model	**Mean Squared Error (MSE)**	**R^2 Score**
Linear Regression	190,860.26	0.9867
K-Nearest Neighbors	157,886.31	0.989
Gradient Boosting	161,932.86	0.9887

Table 2: Model Performance for Predicting Total IPC Crimes Using Top 5 Features

Key Observations

- Accuracy decreased slightly with feature reduction, highlighting the trade-off between interpretability and complexity.
- KNN performed best among the three models on the reduced feature set.

7.3 Aim 3: Categorizing Districts Based on Crime Levels

Process

- A new feature, `TOTAL_CRIMES`, was created by summing all crime-related columns for each district.
- Districts were categorized into five crime levels (1 to 5) based on binning:
  - Level 1: Very Low Crime
  - Level 2: Low Crime
  - Level 3: Moderate Crime
  - Level 4: High Crime
  - Level 5: Very High Crime
- A Random Forest Classifier was trained to predict these crime levels.

Results

- Uttar Pradesh had the highest number of Level 5 districts, indicating a concentration of high crime zones.
- Arunachal Pradesh exhibited a balanced distribution across all crime levels.

Add-On: Classifying Crime Levels Using Selected Features

We aimed to classify districts into crime severity levels using only six key features: `MURDER`, `THEFT`, `HURT/GREVIOUS HURT`, `OTHER IPC CRIMES`, and `KIDNAPPING & ABDUCTION`. The Random Forest Classifier was retrained using this reduced feature set.

Metric	**Value**
Accuracy	96.56%
Cross-Validation Accuracy	95.75%

Table 3: Performance of Random Forest Classifier for Crime Level Prediction (6 Features)

Key Observations

- The Random Forest Classifier achieved high accuracy (96.56%), demonstrating its effectiveness in distinguishing crime levels.
- Using only six features retained strong performance, highlighting the importance of feature engineering.

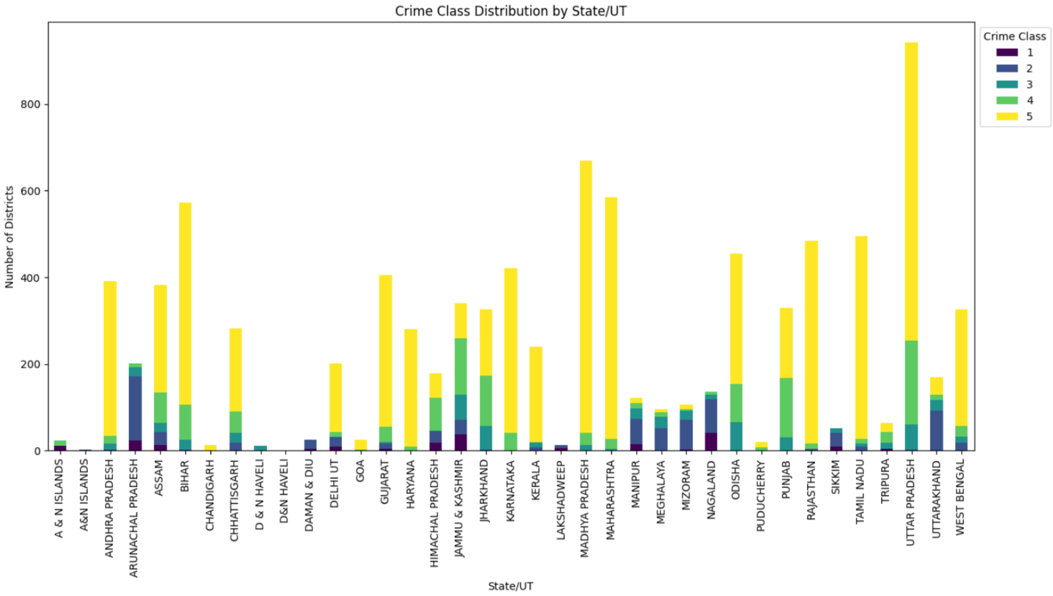


Figure 13: Results of Crime Level Classification

8 Predictions for Future

8.1 Our Attempt

In our analysis, we explored various predictive modeling techniques, including Linear Regression, Support Vector Regression (SVR), and other machine learning approaches, to forecast future crime trends. However, the dataset presented a significant challenge: it consisted of only 13 years of data, resulting in just 13 data points per category.

This limited dataset size made it difficult for the models to accurately learn and capture underlying patterns or trends. Consequently, the predictions lacked reliability, underscoring the need for more extensive historical data or alternative methods to derive meaningful insights.

The results obtained from our predictive modeling efforts were not particularly insightful. Both Support Vector Regression (SVR) and Decision Tree models failed to produce meaningful predictions, likely due to the small size of the dataset and the complexity of the relationships within the data. While Linear Regression did yield some results, it became evident that assuming a linear relationship for predicting future trends was not appropriate. The inherent complexities and non-linear dynamics of the factors influencing crime patterns suggest that more sophisticated models or a larger dataset may be necessary to achieve accurate and reliable predictions.

8.2 We Tried a New Approach: ARIMA and VAR

In our analysis, we initially focused on classification and regression-based approaches to understand crime patterns and district categorization. However, crime data is inherently time-dependent, with trends and patterns evolving over the years. Recognizing this temporal aspect, we shifted to time-series models, specifically ARIMA (Auto-Regressive Integrated Moving Average) and VAR (Vector Auto-Regression), for deeper insights.

Why ARIMA and VAR?

Time-Dependent Nature of Crime Data:

- Crime statistics are often influenced by historical trends and seasonal patterns (e.g., increase during certain months or years).
- ARIMA and VAR explicitly incorporate temporal dependencies, making them ideal for predicting future crime rates and identifying underlying patterns.

Predictive Capability:

- ARIMA excels in forecasting a single variable (e.g., total crimes in a district) based on its past values.
- VAR, on the other hand, handles multiple interdependent variables, such as the relationship between thefts, murders, and riots, providing a better view of crime dynamics.

8.3 ARIMA Model

Data Preparation

- The analysis was performed for a specific state (State 5, selected for demonstration).
- Data was aggregated by year, summing the total number of IPC crimes reported each year.

Stationarity Check and Differencing

- Before fitting the ARIMA model, a stationarity check was performed using the Augmented Dickey-Fuller (ADF) test.
- The null hypothesis of the ADF test is that the series has a unit root, meaning it is non-stationary.
- If the p-value from the ADF test is greater than 0.05, the series is considered non-stationary and requires differencing.
- After differencing the series (first differencing), the ADF test was applied again to confirm stationarity.
- If the series remained non-stationary, further differencing (second differencing) was applied.
- In the case of the selected state data, the series was initially non-stationary, and first differencing was applied to achieve stationarity.

ARIMA Model Setup

- **AR (AutoRegressive):** Models the relationship between an observation and a specified number of lagged (previous) observations.
- **I (Integrated):** Differencing the time series data to make it stationary by removing trends or other non-stationary components.
- **MA (Moving Average):** Models the relationship between an observation and a residual error from a moving average model applied to lagged observations.

Forecasting was done for a 10-year period into the future.

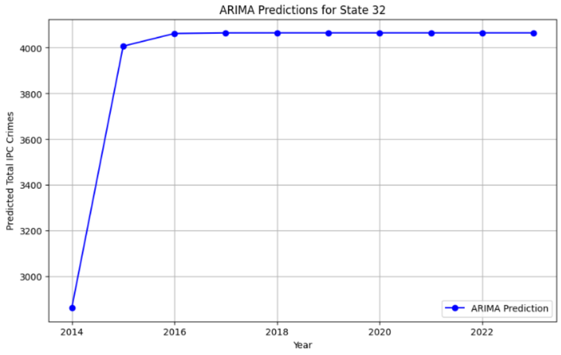


Figure 14: ARIMA PREDICTION : Madhya Pradesh in 10 Years

8.4 VAR Model

Overview of VAR Model

- VAR (Vector Auto-Regression) is a statistical model designed for forecasting interdependent time series.
- It considers multiple variables simultaneously, capturing relationships and dependencies between them.
- Suitable for scenarios where different types of crimes may influence one another, e.g., socio-economic or situational correlations.

Objective

- Predict multiple crime categories (e.g., murder, theft, riots) simultaneously for a selected state in India.
- Provide a detailed forecast to understand trends in various types of crimes from 2014 to 2023.

Data and Model Details

- **Dataset:** Aggregated crime data for the chosen state from 2001 to 2013.
- **Training Data:** Used 13 years of historical data (2001-2013) for building the VAR model.
- **Forecast Period:** Predicted crime counts for 10 years (2014-2023).
- **Model Configuration:** Max Lags = 1 (adjusted based on data suitability).

Visualization

- A combined plot was generated showing actual data (2001-2013) and forecasted data (2014-2023) for all crime categories.
- Highlights include:
  - Clear demarcation of the prediction start year (2014).
  - Variability and correlations between crime categories are evident in the visualization.

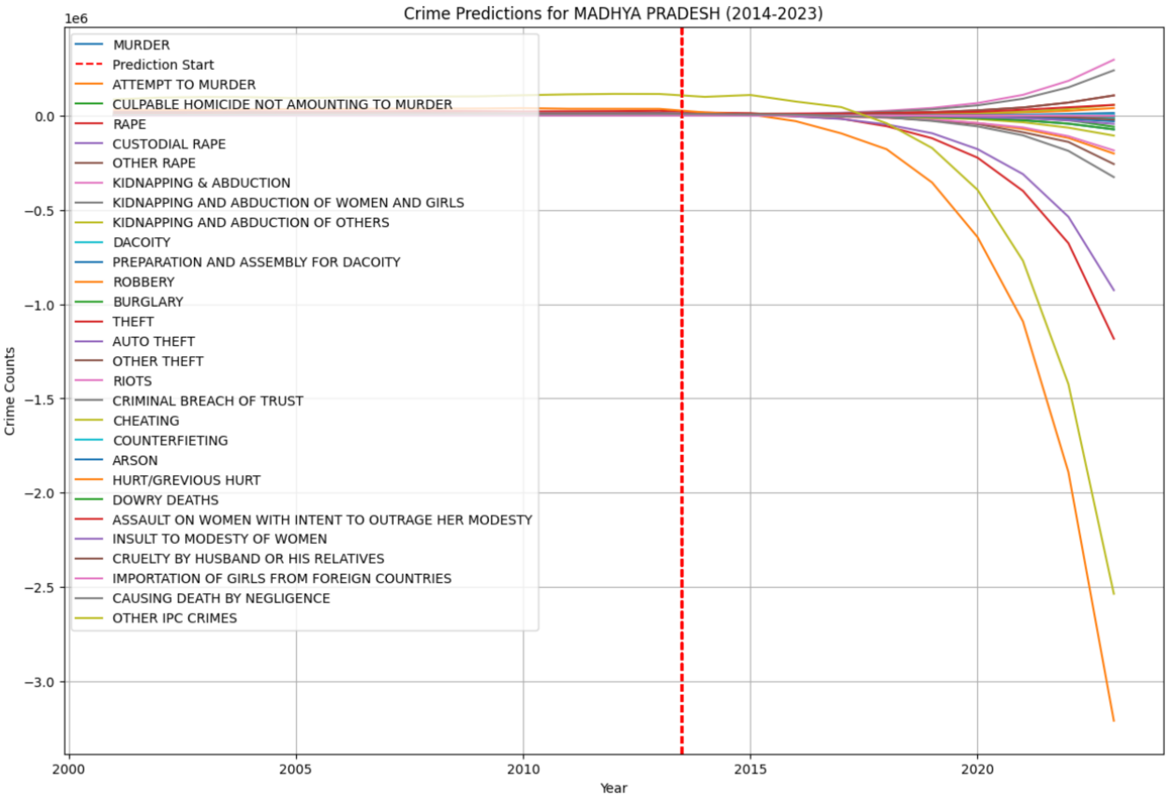


Figure 15: VAR PREDICTION : Madhya Pradesh in 10 Years

9 Conclusion

9.1 Analysis Overview

- **Trends:** Rising total crimes, notably theft and assault.
- **State Patterns:**
  - **Uttar Pradesh:** High Level 5 crime districts.
  - **Arunachal Pradesh:** Even crime distribution.

9.2 Geospatial Prediction Insights

- **Hotspots:** Urban, industrial areas are key crime zones.
- **Modeling:** Regression models and classification (e.g., Random Forest) showed reliable trends; ARIMA/VAR indicated seasonality and correlations.

9.3 Complexity vs. Accuracy

Feature engineering improved simplicity but slightly reduced accuracy.

10 Recommendations

Key Actions

- **Police:** Boost resources in high-crime districts.
- **Hotspot Patrols:** Use geospatial data for prioritization.

Community Education

- **Awareness:** Promote crime prevention and rights.
- **Youth Engagement:** Add crime education in schools.

Technology Collaboration

- **Smart Policing:** Implement predictive tools and surveillance.
- **Data Sharing:** Learn from best practices and share policies.

11 Future Work

This project improved our technical and analytical skills, applying data science to real-world problems. Next, we aim to enhance models with demographic data and create an interactive platform for crime analysis. Hosting on GitHub will enable collaboration and further development.

12 Acknowledgment

We would like to express our heartfelt gratitude to our professor, **Dr. Samiran Das**, for his invaluable guidance, support, and encouragement throughout the course of this project. His insights and expertise were instrumental in shaping our understanding and approach to this study.

This project represents the collaborative efforts of all three team members: **Mrunal Pravin Kulkarni (22183)**, **Darshana Srivathsan (22001)**, and **Prakash Kumbhakar (22241)**. We are grateful for the opportunity to work under his mentorship.

13 Project Repository

The complete set of project resources, including the codes, presentation, and report, is available in the following Google Drive folder. It also contains a README file with instructions for replicating our analysis:

Google Drive: Project Codes and Presentation