A Unified Dataset of Crime in India (2001-2022)

Data Analytics (CS61061)

Anuj Yadav (25CS60R69)
Parth Upadhyay (25CS60R35)
Sumit (25CS60R63)

Department of CSE, IIT KGP

Abstract

Analyzing crime trends in India is notoriously difficult due to the fragmented and inconsistent nature of data reported by the National Crime Records Bureau (NCRB). Reporting formats, table structures, and column names often change year-to-year, making longitudinal analysis nearly impossible. This project addresses this challenge by collecting, cleaning, and amalgamating 22 years of disparate NCRB crime data (2001-2022) into a single, unified, machine-readable JSON file. Using Python and Pandas, hundreds of inconsistencies were standardized, and wide-format tables were transformed into a consistent long format. The resulting dataset unlocks powerful analytical capabilities, enabling the identification of geospatial crime hotspots, victim demographic trends, and systemic bottlenecks. Key insights reveal that while police charge-sheeting rates for crimes against women are relatively high (78.6%), the corresponding judicial conviction rate is extremely low (26.5%), pointing to a significant bottleneck in the judicial system.

Contents

1 Introduction and Problem Statement

The analysis of crime trends in India is a significant challenge for researchers, policymakers, and law enforcement agencies. The core of this problem lies in the fragmented data landscape.

- **Scattered Data:** Data is scattered across numerous reports, tables, and formats.
- **Inconsistent Reporting:** The NCRB reporting formats, table structures, and even column names change frequently from year to year.
- **Blocked Analysis:** This inconsistency makes longitudinal (multi-year) analysis extremely difficult and labor-intensive.

1.1 Project Goal

The primary objective of this project was to create a single, unified, machine-readable `data.json` file. This file aims to unlock 22 years of Indian crime data (2001-2022) for consistent and straightforward analysis.

2 Methodology

To transform the chaotic source data into a coherent JSON file, a five-step methodology was employed:

1. **Data Collection:** Key NCRB tables (e.g., Total IPC+SLL crimes, Total crimes against women) were sourced for the years 2001-2022.
2. **Data Extraction:** Optical Character Recognition (OCR) tools were used to extract tabular data from the PDF-based NCRB reports.
3. **Data Cleaning:** Python scripts using the Pandas library were written to standardize hundreds of inconsistencies, such as varying column headers.
4. **Data Transformation:** Many NCRB tables are published in a "wide" format, with years as columns. These tables were "melted" into a "long", parsable format to standardize the structure.
5. **Data Amalgamation:** Finally, all the cleaned and transformed tables were programmatically merged into a single nested JSON object, with the primary key being the `Year`.

3 Data Structure

The final dataset follows a consistent, hierarchical data layout:

$$\text{Year} \rightarrow \text{State/UT} \rightarrow \text{Crime Category} \rightarrow \text{Attributes}$$

The data is stored in JSON and CSV formats. Data normalization was performed to ensure attribute names are consistent across all years.

Listing 1 shows a snippet for "Dowry Deaths" under "voilent_crime" for 2017, taken from the `data.json` file.

```
"2017": {
    "voilent_crime": [
        {
            "Crime Head": "Dowry Deaths",
            "Cases Pending Trial from the Previous Year": 38654,
            "Cases Sent for Trial during the year": 7038,
            "Total Cases for Trial": 45692,
            "Cases Abated by Court": 3,
            "Cases Withdrawn From Prosecution": 0,
            "Cases Compounded or Compromised": 32,
            "Cases Disposed off by Plea Bargaining": 4,
            "Cases Quashed": 0,
            "Cases Disposed off Without Trial": 39,
            "Cases Stayed or Sent to Record Room": 10,
            "Cases Convicted": 1770,
            "Cases Acquitted": 2264,
            "Cases in which Trials were Completed": 4276,
            "Cases Disposed off by Courts": 4315,
            "Cases Pending Trial at End of the Year": 41377,
            "Conviction Rate": 41.4,
            "Pendency Percentage": 90.6
        },
        ...
    ],
    ...
}
```

Listing 2 shows the structure for state-level summary data, also from the 2017 key in `data.json`.

```
"2017": {
    ...
    "total_crime_by_state": {
        "Andhra Pradesh": {
            "Mid-Year Projected Population(in L)": 523.2,
            "Rate of Cognizable Crimes": 227.9,
            "Chargesheeting Rate": 0.0,
            "IPC crime": 132336,
            "State Level Law": 15666,
            "Total crime": 148002,
            "Crime against Children": 2397,
            "Crime against Women": 17909,
            "Total violent Crimes": 8288,
            "Kidnapping and Abductions": 1018,
            "Murder Cases": 1054
        },
        ...
    }
}
```

4 Analysis and Key Insights

The unified dataset enables powerful, multi-year analysis that was not previously feasible.

4.1 Geospatial "Hotspot" Analysis

The modern schema (2017-2022) allows for immediate high-level comparisons to identify crime hotspots at the state and city level.

Figure 1 shows the top 10 states and union territories by the number of murder cases reported in 2021. **Uttar Pradesh** reported the highest number of cases, exceeding 3,500, followed by Bihar and Maharashtra.

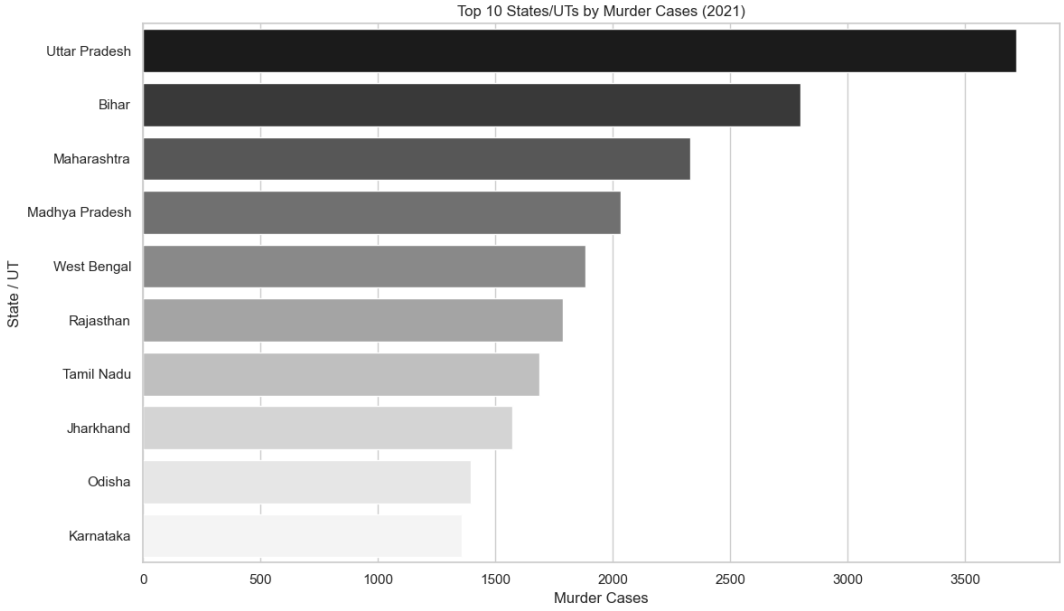

Figure 1: Top 10 States/UTs by Murder Cases (2021)

A similar trend is visible in Figure 2, which shows the top states for *total* violent crimes in 2017. Again, **Uttar Pradesh, Bihar, and West Bengal** are the top three.

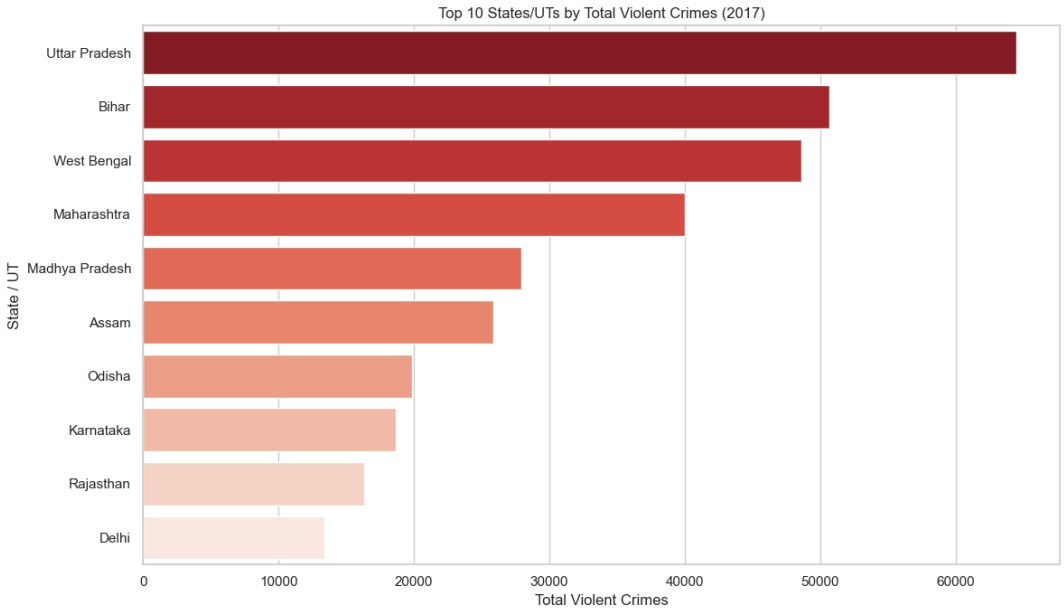

Figure 2: Top 10 States/UTs by Total Violent Crimes (2017)

4.2 Analysis of Crimes Against Women

The dataset's granular keys allow for analysis of crimes against vulnerable groups. As seen in Figure 3, total reported crimes against women in India have shown a consistent upward trend

from 2017 to 2021, rising from 346,773 to 428,278.

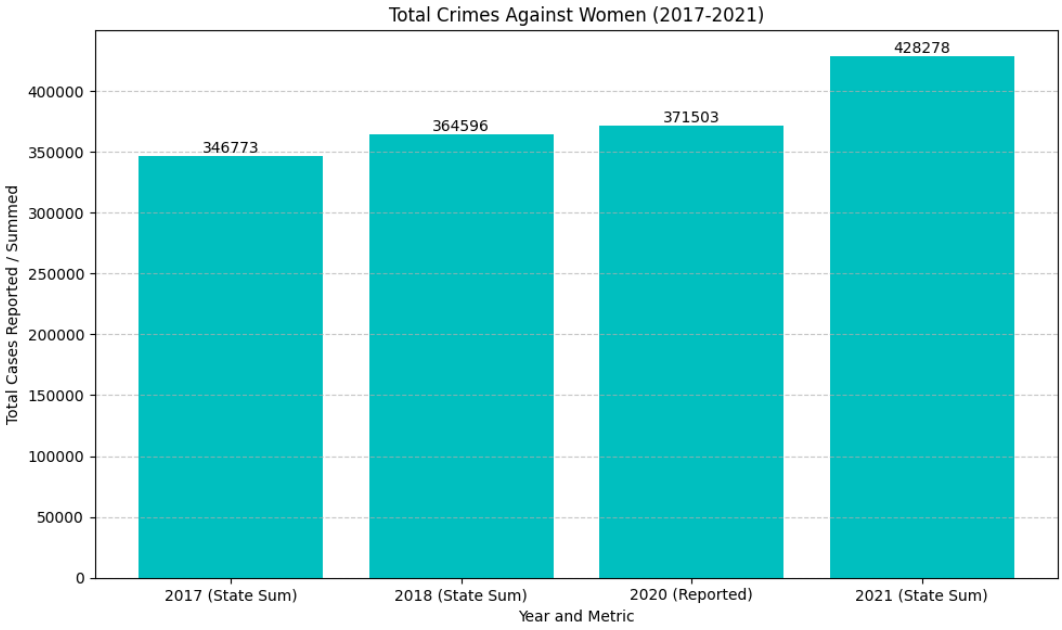

Figure 3: Total Crimes Against Women (2017-2021)

When analyzed at the metropolitan city level for 2018 (Figure 4), **Delhi** emerges as a significant outlier, with substantially more reported cases (over 11,000) than the next highest city, Mumbai (approx. 6,000).

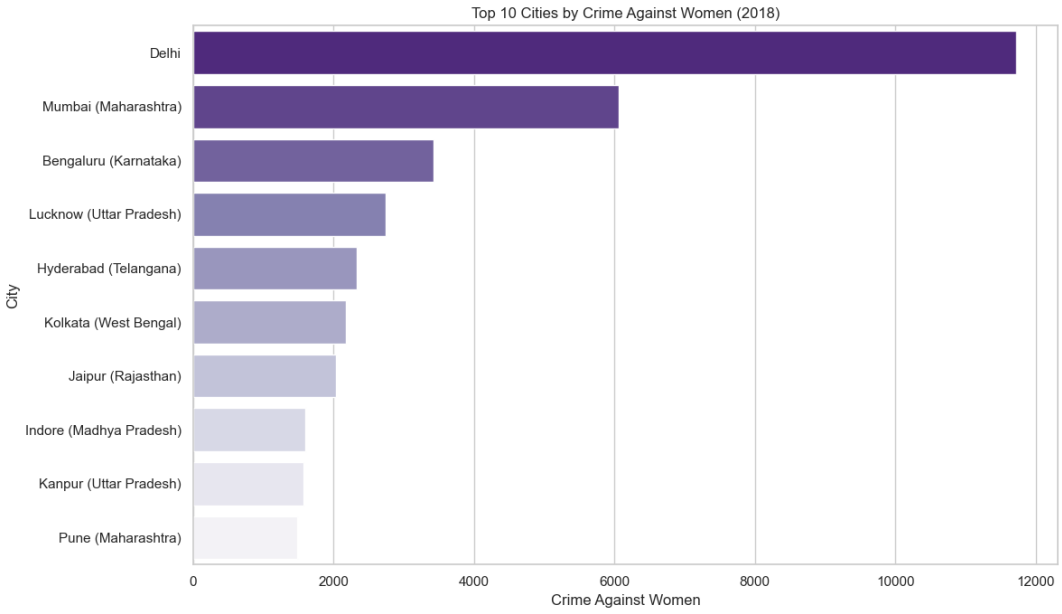

Figure 4: Top 10 Cities by Crime Against Women (2018)

4.3 Identifying Systemic Bottlenecks

One of the most critical applications of this unified dataset is the identification of systemic bottlenecks in the justice system.

4.3.1 Judicial vs. Investigation Bottlenecks

An analysis of crimes against women between 2017-2022 revealed a stark contrast:

- **Police (Investigation):** The average Charge-sheeting Rate was relatively high at **78.6%**.
- **Courts (Judiciary):** The average Conviction Rate for these cases was extremely low, at only **26.5%**.

This insight proves that the primary bottleneck for delivering justice in these cases is not in the police investigation but in the judicial system.

4.3.2 Court Pendency Rates

Figure 5 visualizes the court pendency percentage for various violent crimes in 2016. It clearly shows that for many of the most serious crimes, such as **Offences Against State, Dacoity, and Attempt to Commit Culpable Homicide**, cases remain pending in the judicial system over 90% of the time. In contrast, cases for "Causing Injuries under Rash Driving" have a much lower (though still high) pendency rate of approximately 75%.

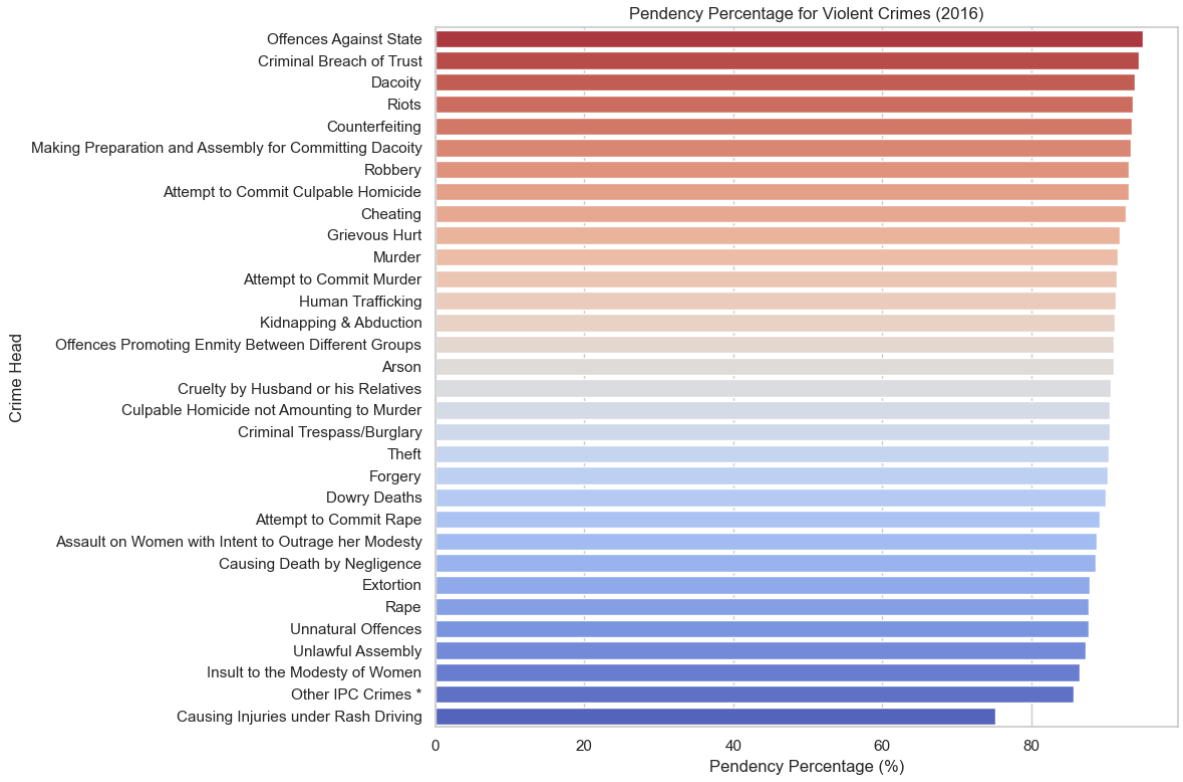

Figure 5: Pendency Percentage for Violent Crimes (2016)

4.3.3 Police Disposal Rates

The dataset also reveals bottlenecks at the police level. Figure 6 shows the disposal of total crimes against children by police in 2019.

- Only **39.9%** of cases were Chargesheeted (sent for trial).
- **37.1%** of cases were still Pending Investigation at the end of the year.

- **22.9%** of cases were closed via a Final Report (i.e., not chargesheeted).

  This indicates a significant backlog at the investigation stage for crimes against children.

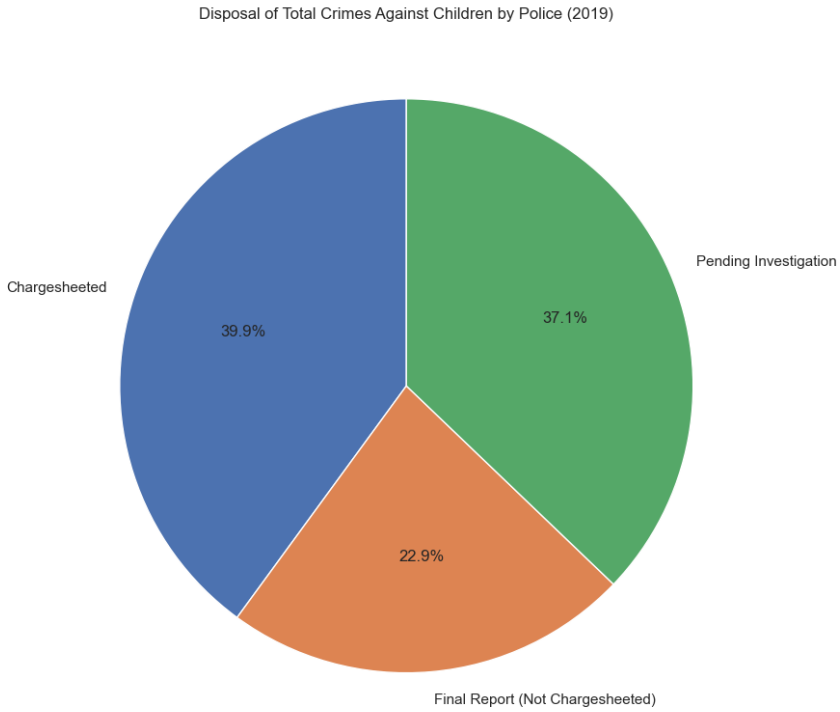

Figure 6: Disposal of Total Crimes Against Children by Police (2019)

5 Applications

This dataset can serve as a powerful strategic tool for different branches of government:

- **For Law Enforcement:** A Police Commissioner can use this data for strategic resource allocation. For example, the data provides a justification to deploy a dedicated anti-women abuse task force in Delhi, while other cities might focus resources elsewhere.
- **For the Judiciary & Policymakers:** The data is critical for identifying and proving the existence of systemic bottlenecks. The clear discrepancy between high charge-sheeting rates and low conviction rates proves the primary bottleneck is in the judicial system, allowing policymakers to focus on rectifying system flaws rather than just policing.

6 Limitations

The project has several limitations based on the source data:

- The dataset is only as accurate as the data originally reported by state police forces.
- The formatting of the source dataset is not identical across all 22 years, which presents an ongoing challenge (Heterogeneous Schema).
- Some years contain typos or 0.0 values, which represent unreported data points from the original NCRB tables.

7 Conclusion and Future Work

7.1 Conclusion

This project successfully amalgamated 22 years of fragmented crime data into a single, queryable, and machine-readable `data.json` file. By cleaning and standardizing two decades of NCRB reports, this dataset can be used as a powerful strategic tool for law enforcement (for resource allocation) and the Judiciary (for identifying bottlenecks).

7.2 Future Work

The following steps are planned for the future:

- Adding new 2023-2025 data as soon as it is released by the NCRB.
- Creating an interactive, public-facing dashboard powered by this JSON file to make the data accessible to all citizens.

References

[1] National Crime Records Bureau (NCRB). (2001-2022).

Crime in India.

Ministry of Home Affairs, Government of India.