CO31

Data Curation and Exploratory Data Analysis for Smart policing data

Group members

Isha Kumari
Anuj Kumar Yadav

Problem Statement

The criminal justice process in India involves several stages, starting from the registration of an FIR (First Information Report) to the final decision by the court. Each criminal case goes through multiple steps, including police investigation, filing of charge sheets, disposal by police or court, and sometimes prolonged trials. Some cases never get investigated or await trial forever. Tracking how efficiently these steps are carried out, both for different types of crimes and across various states, is important for understanding potential bottlenecks and areas where improvements are needed.

Our project focuses on conducting an Exploratory Data Analysis (EDA) of crime related data in India, covering the years 2021 and 2022. The dataset we're working with includes tables detailing police and court disposals, state-wise case progress, and the time taken for charge sheets and final reports, all broken down by 20 specific types of crimes. By examining this data, we aim to identify patterns such as differences in disposal rates between states, time taken at various stages of the process, and any trends or outliers within different crime categories. These insights could help us better understand how criminal cases are handled in India and point out areas that might benefit from procedural improvements or reforms.

Work Approach

Understanding the Project and Setting Objectives
We began by setting clear objectives: to analyse crime data in India from 2018 to 2022 and understand how cases progress from FIR registration to police investigations and court processes. Our focus was to uncover trends, such as how long cases take at different stages and the variation in case outcomes across states and crime categories. This step helped us outline key questions, like differences in case disposal rates and timelines for different types of crimes and states.

Data Collection from Official Sources
Our data was sourced from the National Crime Records Bureau (NCRB) database, specifically tables 17A and 18A, which detail police and court disposals, chargesheet times, and statewise data across various crime types. These tables provided a comprehensive look at the processing of cases and allowed us to focus on specific crime categories and regions.

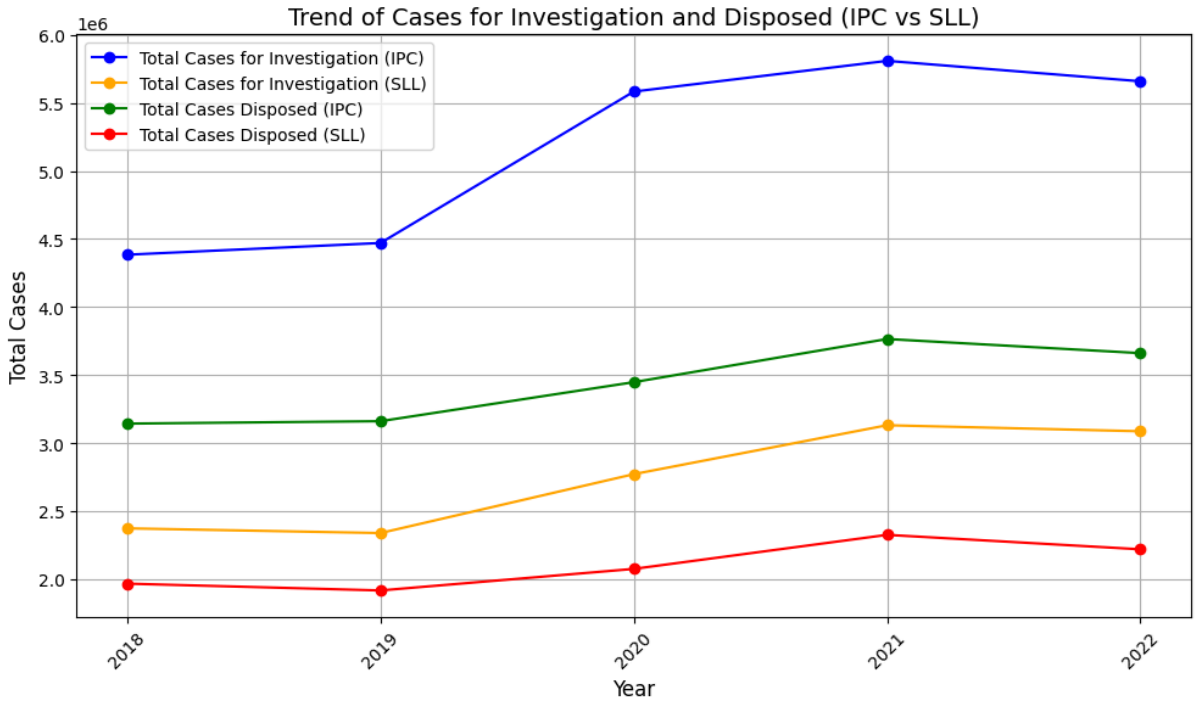
Data Preparation and Cleaning
With a basic structure in mind, we prepared the data by removing columns that didn't align with our objectives or had excessive missing values. To narrow down our focus, we chose 20 main crime categories and the top 10 states with the most FIRs for each year, resulting in a dataset tailored to our analysis goals.

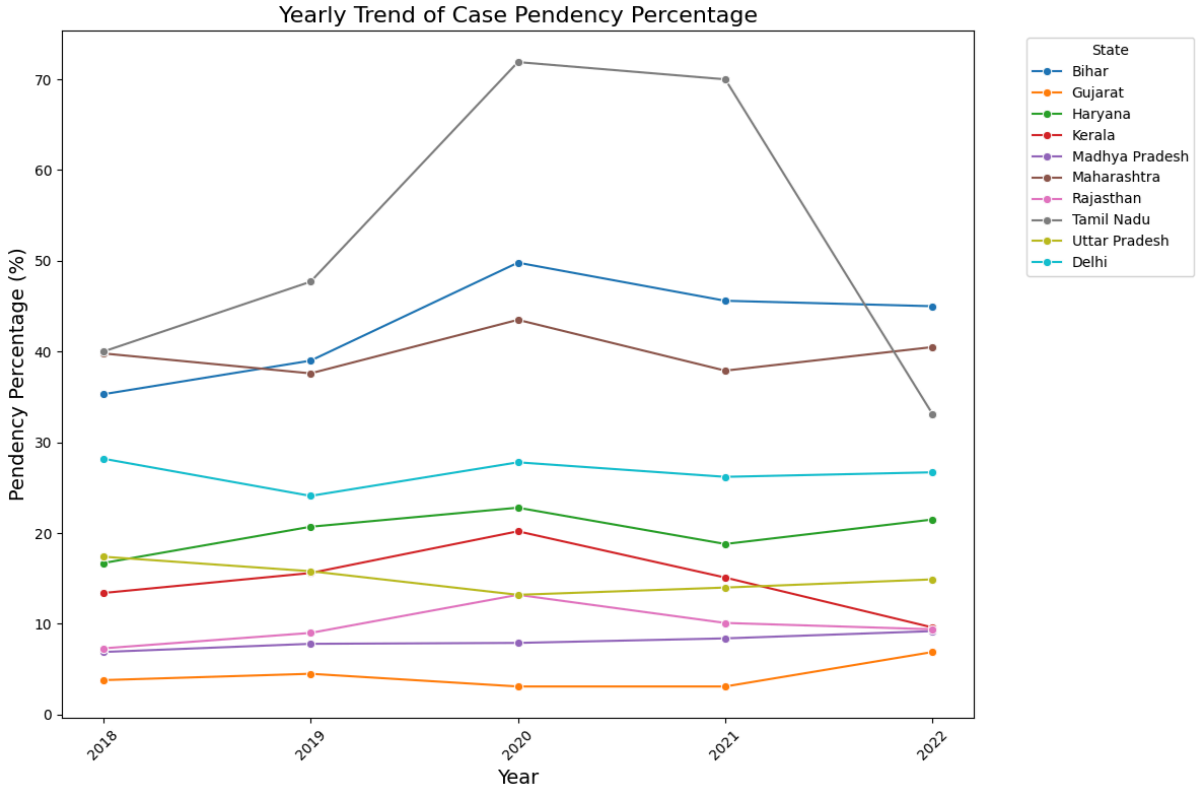
Organizing Data into Funnels
To visualize how cases proceed through each stage, we organized the data into funnels that track the lifecycle of cases, from FIR registration through police and court handling. In Google Sheets, we created separate funnels for police and court, showing metrics such as cases disposed of without investigation, charges sheet filing, and trial outcomes. This funnel approach allowed us to pinpoint where delays and backlogs occurred within the system.

Exploratory Data Analysis (EDA) and Visualization
With our data cleaned and structured, we used Google Colab to explore and visualize trends. We created bar charts, line graphs, and funnel charts to represent our findings, helping us uncover key patterns across crime types and states. This analysis gave us insights into disposal rates, timelines, and case outcomes across different stages and regions.

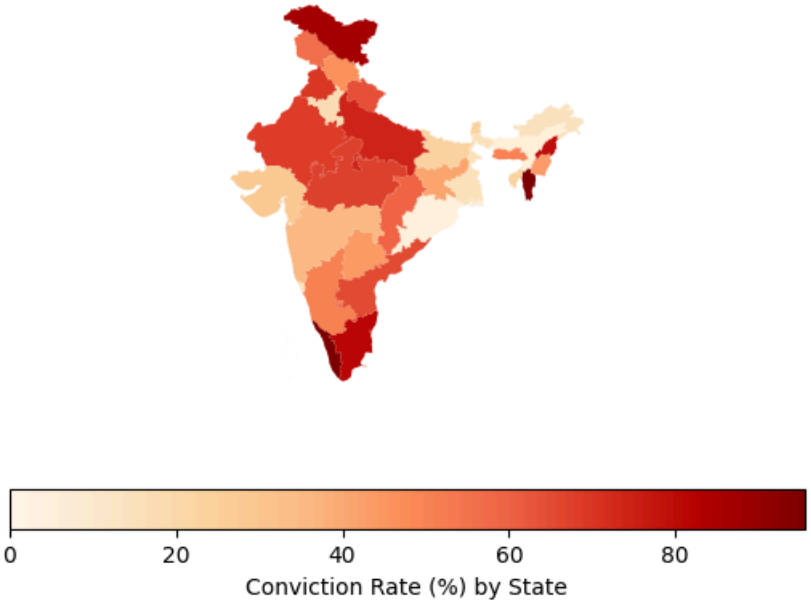
Documentation and Repository Setup
We organized our project in a GitHub repository with folders for data files (Excel sheets), visualizations (images), and code files (.ipynb). This setup allows others to open the .ipynb files in Colab and view the visualizations directly, providing easy access and enabling smooth collaboration.

Some key Visualizations

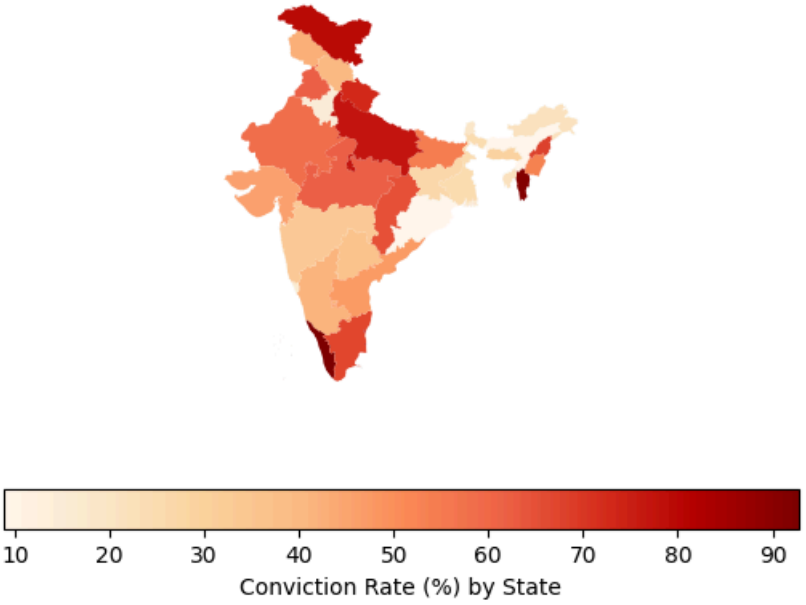


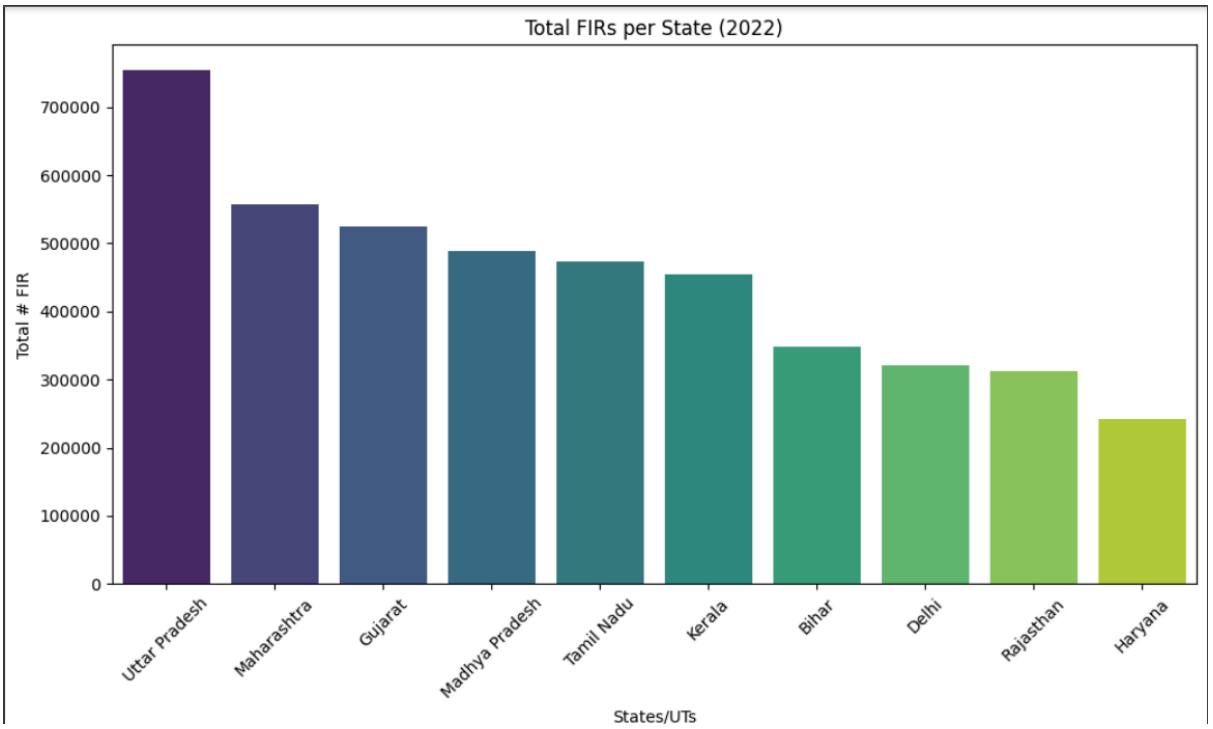


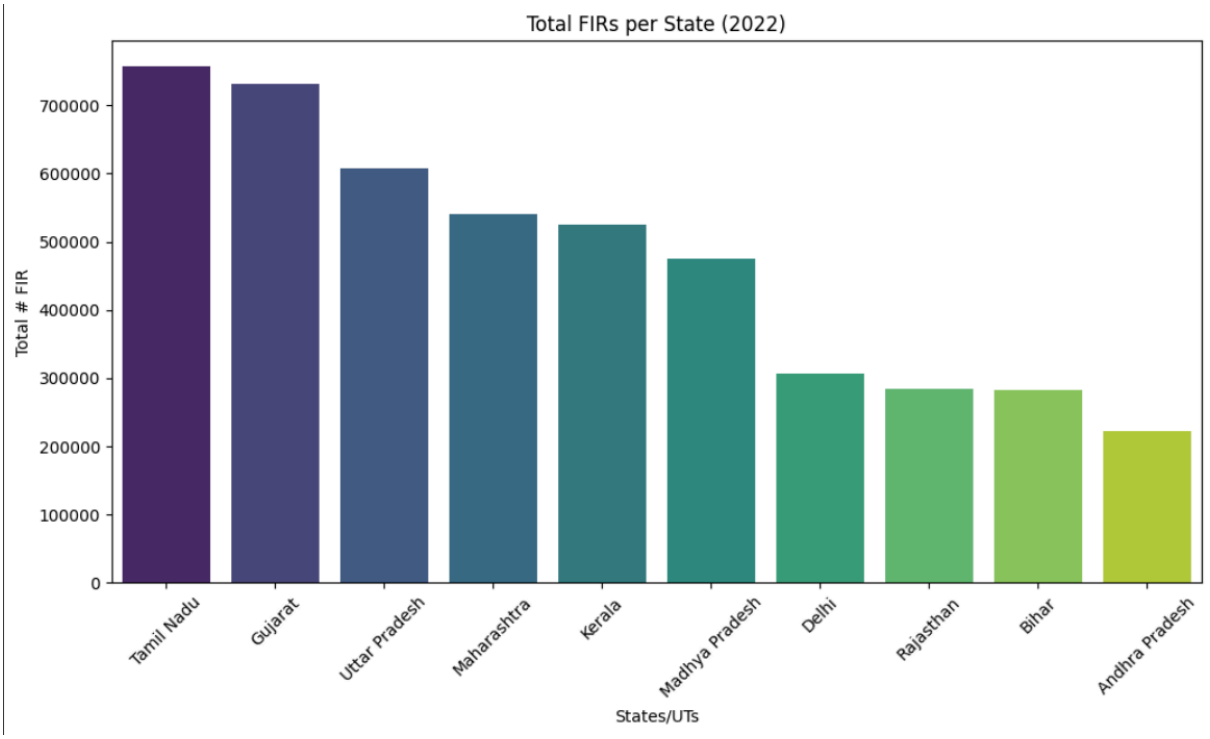
Statewise Conviction Rate Across India (2021)

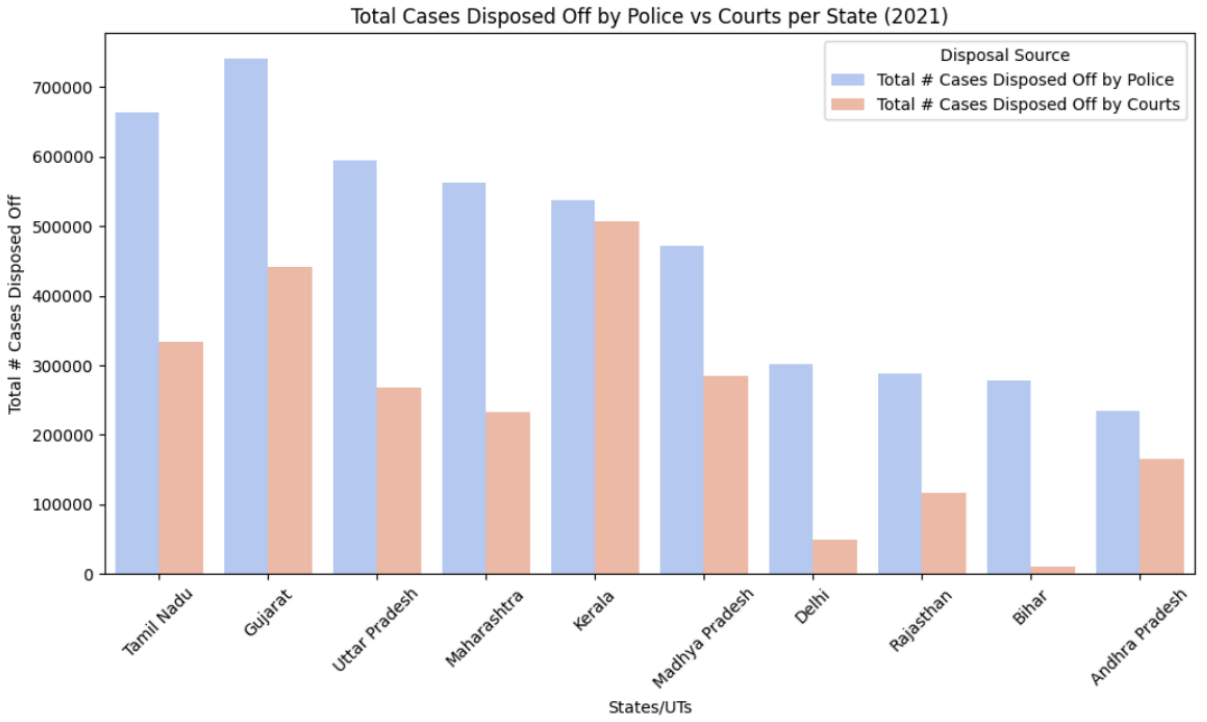


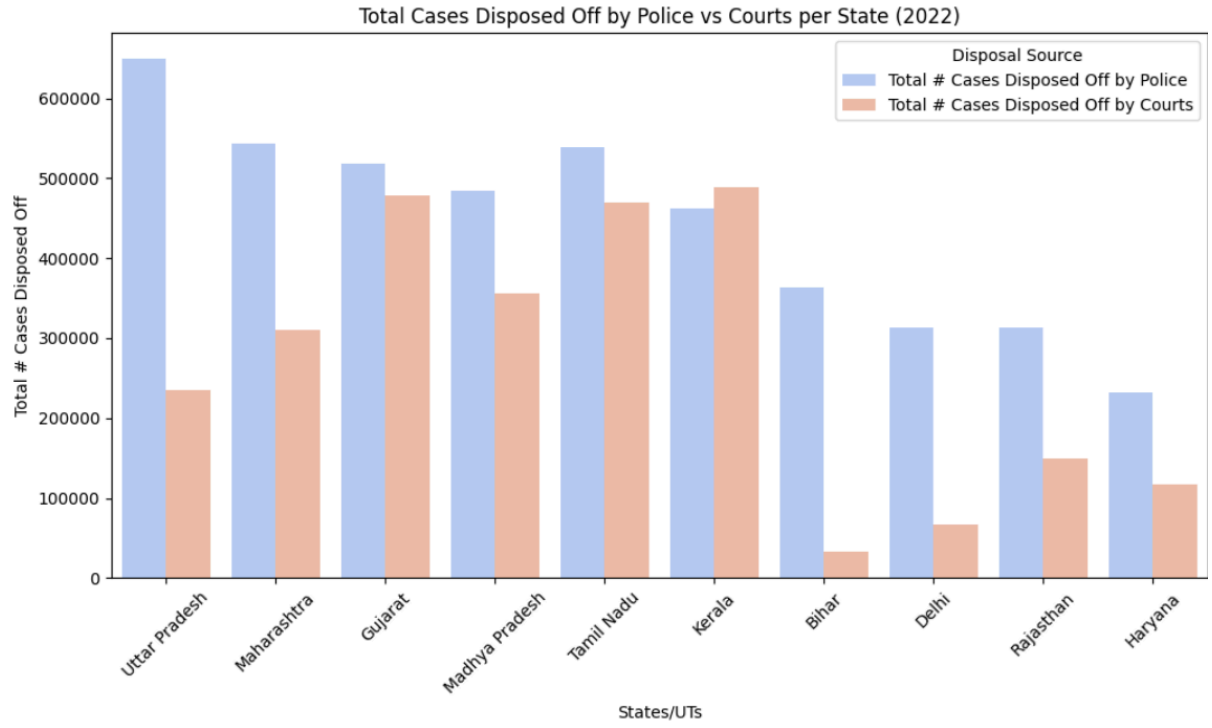
Statewise Conviction Rate Across India (2022)

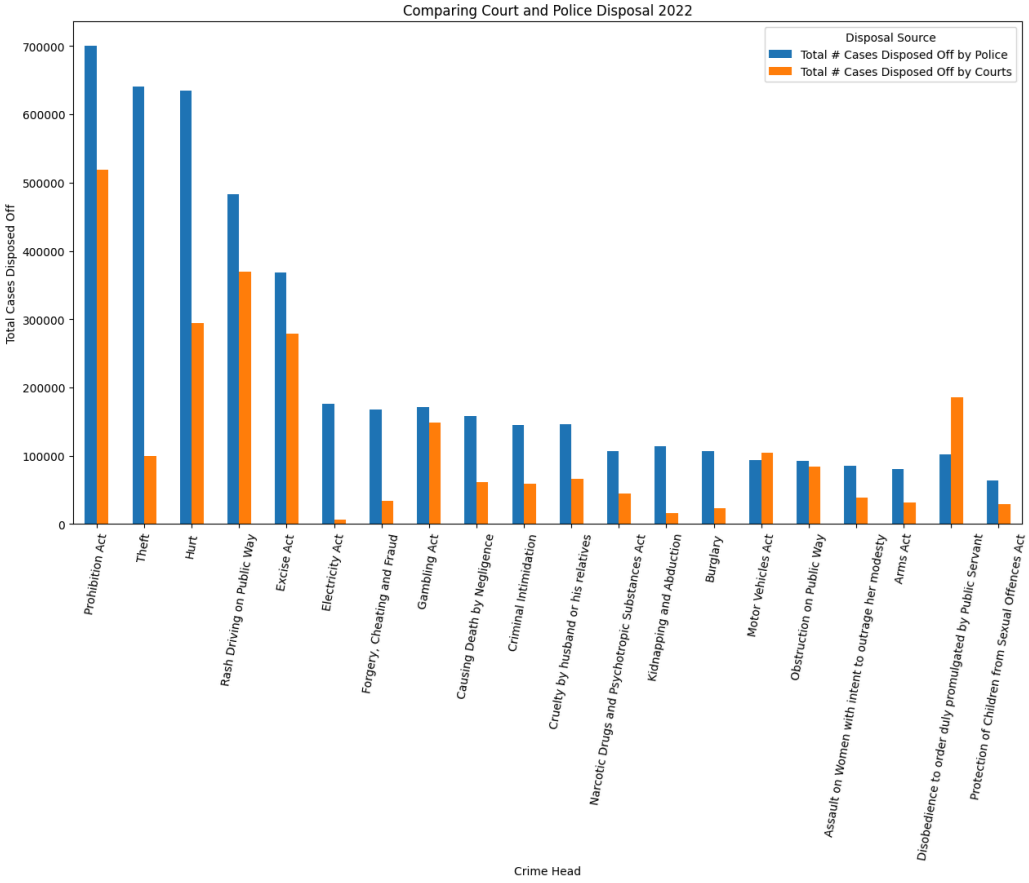


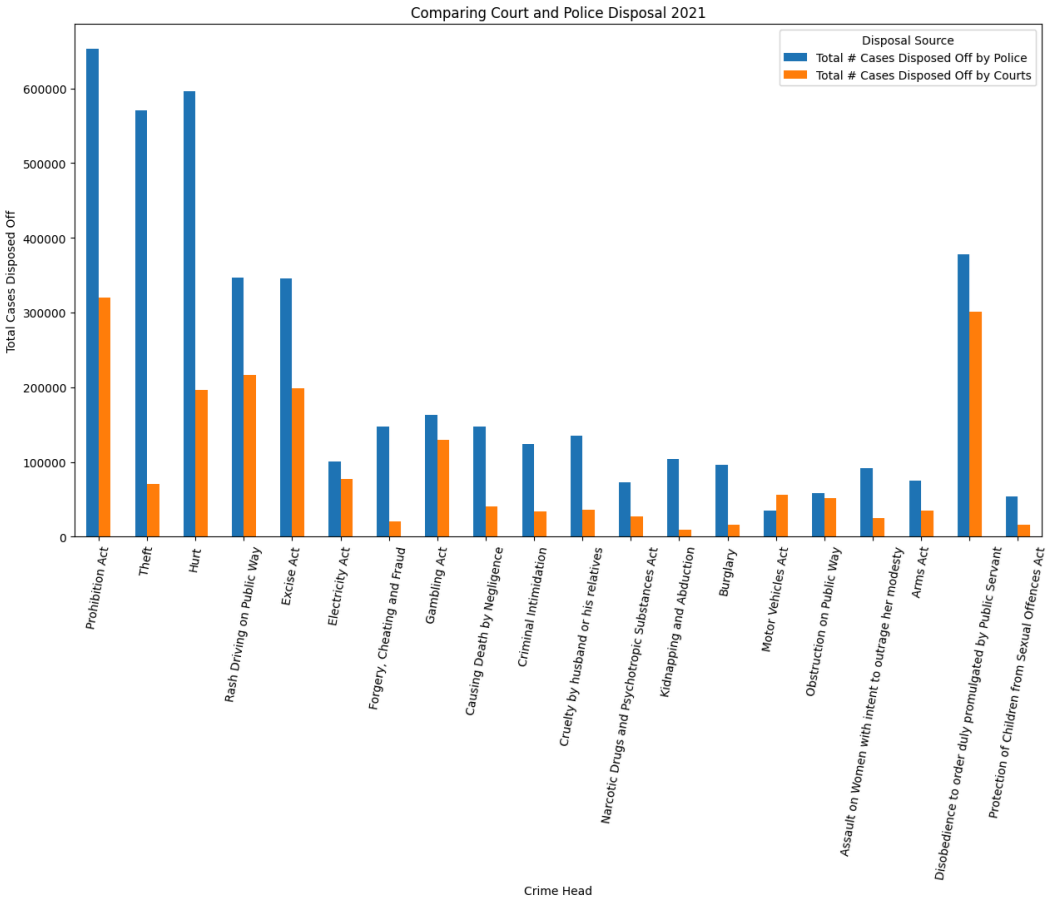

Total FIRs per State (2022)
Total # FIR
0
100000
200000
300000
400000
500000
600000
700000
Uttar Pradesh
Maharashtra
Gujarat
Madhya Pradesh
Tamil Nadu
Kerala
Bihar
Delhi
Rajasthan
Haryana
States/UTs


Total FIRs per State (2022)
Total # FIR
0
100000
200000
300000
400000
500000
600000
700000
Tamil Nadu
Gujarat
Uttar Pradesh
Maharashtra
Kerala
Madhya Pradesh
Delhi
Rajasthan
Bihar
Andhra Pradesh
States/UTs


Total Cases Disposed Off by Police vs Courts per State (2021)
Disposal Source
Total # Cases Disposed Off by Police
Total # Cases Disposed Off by Courts
Total # Cases Disposed Off
0
100000
200000
300000
400000
500000
600000
700000
Tamil Nadu
Gujarat
Uttar Pradesh
Maharashtra
Kerala
Madhya Pradesh
Delhi
Rajasthan
Bihar
Andhra Pradesh
States/UTs


Total Cases Disposed Off by Police vs Courts per State (2022)
Disposal Source
Total # Cases Disposed Off by Police
Total # Cases Disposed Off by Courts
Total # Cases Disposed Off
0
100000
200000
300000
400000
500000
600000
Uttar Pradesh
Maharashtra
Gujarat
Madhya Pradesh
Tamil Nadu
Kerala
Bihar
Delhi
Rajasthan
Haryana
States/UTs


Comparing Court and Police Disposal 2022
Disposal Source
Total # Cases Disposed Off by Police
Total # Cases Disposed Off by Courts
Total Cases Disposed Off
0
100000
200000
300000
400000
500000
600000
700000
Prohibition Act
Theft
Hurt
Rash Driving on Public Way
Excise Act
Electricity Act
Forgery, Cheating and Fraud
Gambling Act
Causing Death by Negligence
Criminal Intimidation
Cruelty by husband or his relatives
Narcotic Drugs and Psychotropic Substances Act
Kidnapping and Abduction
Burglary
Motor Vehicles Act
Obstruction on Public Way
Assault on Women with intent to outrage her modesty
Arms Act
Disobedience to order duly promulgated by Public Servant
Protection of Children from Sexual Offences Act
Crime Head


Comparing Court and Police Disposal 2021
Disposal Source
Total # Cases Disposed Off by Police
Total # Cases Disposed Off by Courts
Total Cases Disposed Off
0
100000
200000
300000
400000
500000
600000
Prohibition Act
Theft
Hurt
Rash Driving on Public Way
Excise Act
Electricity Act
Forgery, Cheating and Fraud
Gambling Act
Causing Death by Negligence
Criminal Intimidation
Cruelty by husband or his relatives
Narcotic Drugs and Psychotropic Substances Act
Kidnapping and Abduction
Burglary
Motor Vehicles Act
Obstruction on Public Way
Assault on Women with intent to outrage her modesty
Arms Act
Disobedience to order duly promulgated by Public Servant
Protection of Children from Sexual Offences Act
Crime Head

Findings

1. **FIR Distribution by States**

- The highest numbers of FIRs are observed in Uttar Pradesh (UP), Maharashtra (MH), Gujarat (GJ), and Tamil Nadu (TN), likely linked to their large populations.

2. **Police vs. Court Disposals**

- In states like UP, MH, Bihar (BR), Delhi (DL), Rajasthan (RJ), and Haryana (HR), police disposals significantly outnumber court disposals. However, Kerala, Madhya Pradesh (MP), and Gujarat have a relatively high number of court disposals.

3. **Police Disposal Reasons**

- A majority of police disposals are classified as "True but Insufficient Evidence." Rajasthan, Delhi, and TN show high rates of "Mistake of Fact/Law/Civil Dispute" as disposal reasons, while Rajasthan and HR also show high counts of cases marked as "False."

4. **Court Disposal Reasons**

- Besides conviction and acquittal, courts in TN and HR frequently dispose of cases without trial; TN also has a high rate of cases withdrawn from prosecution. Acquittal rates are high in MH, GJ, and HR, while Kerala and Delhi exhibit the highest conviction rates. Trial completion rates are lowest in HR.

5. **Crimehead-Specific Trends**

- The largest FIR counts are in Prohibition Act violations, Theft, and Hurt. Most police disposals are for Theft, Burglary, Forgery/Cheating/Fraud, and Kidnapping/Abduction, while court disposals are mainly for Prohibition Act and Rash Driving cases. Conviction rates are highest in SLL crime categories such as the Excise Act, Gambling Act, Motor Vehicles Act, and Arms Act.

6. **Trial and Conviction Patterns by Crimehead**

- The Excise Act has both the highest conviction rate and the lowest trial completion rate, highlighting a unique disparity.

7. **Police vs. Court Disposals by Crime Type**

- As expected, police dispose more cases than courts in categories like Theft, Burglary, Forgery/Cheating/Fraud, and Kidnapping. However, SLL crimes see a higher disposal rate by courts than IPC crimes.

8. **Non-Trial Disposals**

- High non-trial disposals are seen in Hurt and Electricity Act cases.

9. **Pandemic-Related Trends**

- Case filings surged in 2021 and 2022, likely due to offenses related to COVID-19 restrictions, with a corresponding rise in case pendency during this period.

10. **Court Pendency by State**

- Courts in GJ, MP, RJ, and UP have relatively low case pendency, while TN, BR, and MH show very high levels.

11. **Conviction Rates by State**

- The highest conviction rates are observed in Kerala, Jammu & Kashmir, and UP.

User Manual

This guide provides two options for running and exploring the visualizations in our project.

Option 1: Viewing Visualizations on Google Colab

1. **Upload Files to Google Colab**
    - Open Google Colab and upload the .ipynb file from the GitHub repository.
    - Ensure the dataset files (Excel sheets) are accessible. Either upload them to the Colab session directly or link them to your own Google Drive if needed.
2. **Run All Cells**
    - In the Colab notebook, go to the "Runtime" menu and select **Run All** to execute all cells in the notebook.
    - The visualizations will automatically generate within the notebook as cells are executed.
    - Scroll through the notebook to view the outputs and visualization other data representations.

Option 2: Running the Code on a Local Python IDE

1. **Setup Environment**
    - Download the .ipynb file and data files from the GitHub repository.
    - Install Jupyter Notebook (if not installed) or use an IDE that supports .ipynb files, such as VS Code with the Jupyter extension.
    - Install libraries as mentioned in *'requirements.txt'*
2. **Organize Files**
    - Ensure all data files are saved in the correct folder structure as specified in the GitHub repository.
    - Open the .ipynb file in Jupyter Notebook or your chosen IDE.
3. **Run the Code**
    - Run each cell in the notebook to execute the code and generate visualizations. The datasets should load successfully if the paths are correctly specified.
    - To modify the visualizations, adjust the code in specific cells and rerun to see updated outputs.
4. **Editing and Customizing**
    - With this method, one can make code edits to explore different visualizations or customize the analysis.

By following either of these methods, users can access and interact with the project's visualizations, either by simply viewing or by making custom edits.

Technical Manual

This technical manual provides a detailed overview of the Crime in India EDA project, which focuses on analyzing crime data from the National Crime Records Bureau (NCRB) datasets for the years 2021 and 2022. The project involves visualizing and exploring crime trends, police and court disposal rates across Indian states, and different crime categories.

1. Datasets

The datasets are extracted from Tables 17A and 18A, Volume III of Crime in India 2021 and 2022. These datasets contain detailed statistics on the following:

- Police Disposal Data (Statewise and Crime Headwise)
  Records the total FIRs filed and the cases disposed of by police across the top 10 states and selected 20 crime categories.
- Court Disposal Data (Statewise and Crime Headwise)
  Records court disposal data for top 10 states and 20 selected crime categories.
- Chargesheet and Final Report Data
  Records data on the time taken to complete chargesheets and final reports for 20 selected crime categories.

2. Software Libraries

- Pandas: For data manipulation and analysis, including loading and structuring data from Excel files.
- Matplotlib: To create static, animated, and interactive visualizations.
- Seaborn: For enhanced data visualization built on Matplotlib with theme and aesthetic improvements.
- Geopandas: Used for geospatial data manipulation and analysis, allowing for the visualization of geographic data, such as mapping case distributions across different states or regions.

3. Program Documentation

The analysis is divided into four notebooks, each focusing on different aspects of the data:

Notebook 1: State Wise Disposal of Cases

Description: This notebook analyses total FIRs per state, police and court disposal rates, and breakdowns of police and court disposal categories.

Exploratory Data Analysis (EDA):

- Total FIRs per State: Visualizes total FIR counts for the top 10 states.
- Cases Disposed by Police vs Courts: Compares police and court disposals by state.
- Breakdown of Police Disposal Categories: Stacked bar charts showing police disposal types (e.g., FR NonCognizable, Mistake of Fact).
- Breakdown of Court Disposal Categories: Stacked bar charts showing court disposal types (e.g., Convicted, Acquitted).
- Conviction and Acquittal Rates by Crime Type: Conviction/acquittal comparison per state.
- Completion Rate of Trials: Calculates and visualizes the rate of completed trials per state.

Notebook 2: Crime Headwise Police and Court Disposal of Cases

Description: This notebook examines police and court disposal rates by selected crime heads (20 categories) for both years. It compares the handling of cases across different crime types by police and courts.

Exploratory Data Analysis (EDA):

- Trend Analysis by Crime Head: Visualizes the total number of cases disposed by police vs. courts for each crime category.
- Police Disposal Breakdown by Reason: Visualizes reasons for police disposal of cases
- Court Disposal Breakdown by Outcome: Compares specific case outcomes for selected crime types (e.g., Conviction, Acquittal, Dismissal).
- Trial Completion Rate by Crime Head: Highlights which crime heads have a higher or lower trial completion rate, indicating efficiency in court processes for those cases.

- Comparing Court and Police Disposal: Compares disposals of cases between police and courts.
- Conviction to Acquittal Ratio by Crime Head: Compares outcomes of cases
- Cases Disposed Off Without Trial by Crime Head: Compares which cases reach trial.

Notebook 3: Time Taken to Complete Chargesheet and Final Report

Description: This notebook analyzes the time taken to complete chargesheets and final reports for the selected crime categories. It visualizes the trends in how long it takes for these stages of the judicial process to be completed across different crime heads.

Exploratory Data Analysis (EDA):

- Proportion of Charge Sheets Completed Over Time: Analyzes and visualizes the proportion of chargesheets completed over time.
- Distribution of Cases Over Time Intervals by Crime Head: Analyzes the distribution of completion times for chargesheets and final reports.
- Comparison of IPC vs SLL Cases by Time Interval: Compares completion times for IPC and SLL crimes.

Notebook 4: Overall Analysis

Description: This notebook uses a mix of the above datasets and visualises trends in the data across years.

Exploratory Data Analysis (EDA):

- Trend of Cases for Investigation and Disposed: Compares counts of cases for investigation and disposed by police in past years with a line graph.
- Yearly Trend of Case Pendency Percentage: Compares performance of courts across different states in India in past years.
- State-wise conviction rate across India: Visualises conviction rates of Indian states by using shades on map of India
- Breakdown of Case Dispositions: Pie chart showing fates of cases by Courts.
- Crimehead-wise Charge sheet vs Final Report counts: Scatter plot with counts for charge sheets filed on one axis and counts for final reports submitted on the other axis.

4. GitHub Repository

https://github.com/EDA-Crime-In-India/crime_data_analysis

The GitHub repository is organized as follows:

- **data**: Contains Excel files.
    - **Subfolder 1**: State-wise police and court data for 2021 and 2022.
    - **Subfolder 2**: Crime Head-wise police and court data for 20 crime heads (IPC and SLL).
    - **Subfolder 3**: Time-taken data for chargesheet and final reports for the same 20 crime heads.
    - **Subfolder 4**: Files for Overall analysis
    - **India_shapefile**: dbf, prj, shp and shx files for visualizations on Map of India.
- **codes**:
    - *1_State_Wise_disposal_of_Cases.ipynb*
    - *2_Crime_Head_wise_Police_and_court_disposal_of_Cases.ipynb*
    - *3_Time_Taken_to_Complete_Charge_Sheet_Final_Report.ipynb*
    - *4.EDAPlots.ipynb*
- **img**: Screenshots of visualizations from the notebooks.
- **requirements**.txt: Python libraries needed

Presentation Link:

docs.google.com/presentation/d/1gHfo9JXeTZ03Ch-ERD2N8qQVR3XxHL2-BleYC4IbJ3w