EDA Patrol - Crime in India (2001–2014)

Exploratory Data Analysis Report

Prepared By: Hariharan Ramesh, Jean Paul, Akshay Prassanna Sivaprakash & Jennithra Srinivasan

Technology Used: Python and Streamlit

Link for Interactive Dashboard – Hosted In Streamlit

Introduction

This report presents an in-depth exploratory data analysis (EDA) of the Indian Penal Code (IPC) crime dataset from 2001 to 2014. The dataset includes district-wise crime records across Indian states. The goal is to extract key insights, identify crime trends, evaluate high-risk areas, and apply machine learning models to assist in predictive analysis and policymaking.

Data Cleaning and Preparation

- Merged district-level crime data with geo-boundary shapefiles using fuzzy string matching.
- Removed rows with labels such as "Total" which represented aggregate values.
- Normalized district and state names for consistency across datasets.
- Created new columns such as crime_risk_index, high_crime binary labels, and most_common_crime.

Exploratory Data Analysis

Total Crimes and Murders

- **Total IPC Crimes Recorded (2001-2014):** Over 29 million.
- **Average Murders per District:** 43 murders.

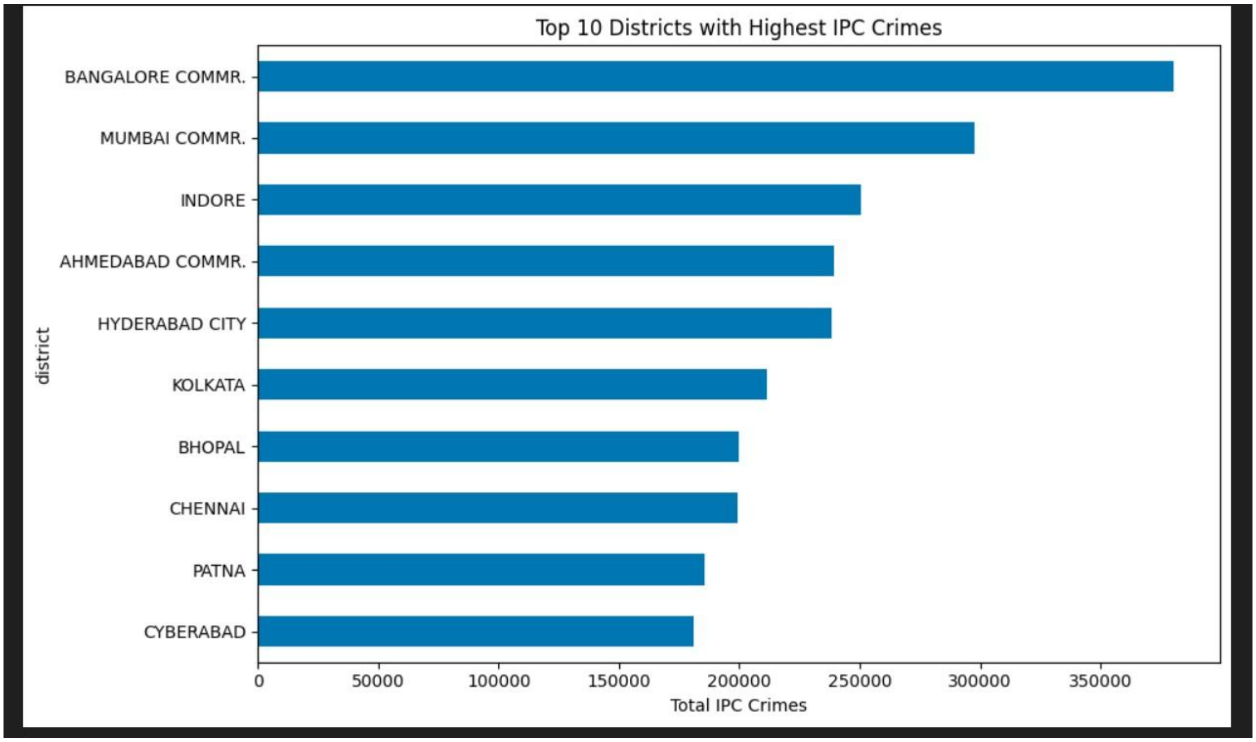
Crime Distribution Across States

States with highest total IPC crimes:

1. Madhya Pradesh
2. Maharashtra
3. Tamil Nadu
4. Andhra Pradesh
5. Uttar Pradesh

Top Crime-Intensive Districts

1. Bangalore Commr.
2. Mumbai Commr.
3. Indore
4. Ahmedabad Commr.
5. Hyderabad City



Urban vs. Rural Crime Patterns

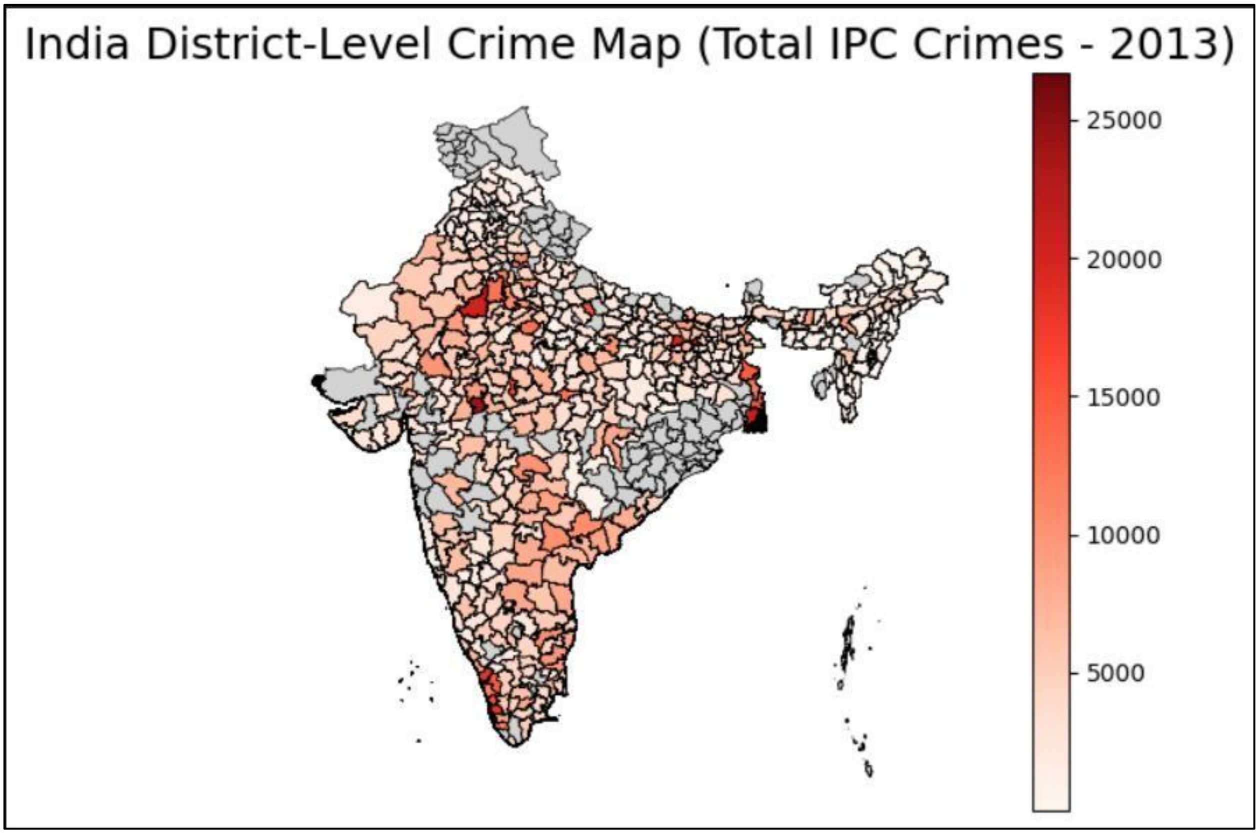
Due to the absence of an explicit urban/rural flag, districts containing city names (e.g., "commr", "urban") were used as a proxy for urban areas.

- **Average IPC crimes in urban areas:** ~7346
- **Average IPC crimes in non-urban areas:** ~270

Visualization and Interpretation

Heatmaps and Geospatial Plots

- Crime density is highest in central and southern India.
- Geospatial heatmaps highlighted major hotspots like Delhi, Mumbai, and Bangalore.
- Choropleth maps were generated by merging shapefile data with district crime aggregates.



Advanced Statistical and Machine Learning Analysis

Most Common Crimes by District

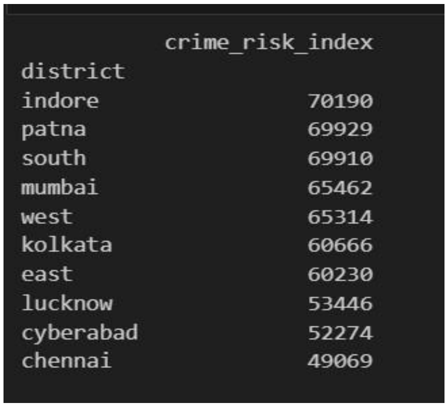
– For each district, the most reported crime was determined. –
Most common: "hurt/grevious hurt" and "theft".

Crime Risk Index

Developed using a weighted sum of serious crimes:

– Murder (weight=3), Rape (2), Robbery (2), Theft (1), Dacoity (2), Kidnapping (2)

Top districts: Outer Delhi, Mumbai, North-East Delhi.



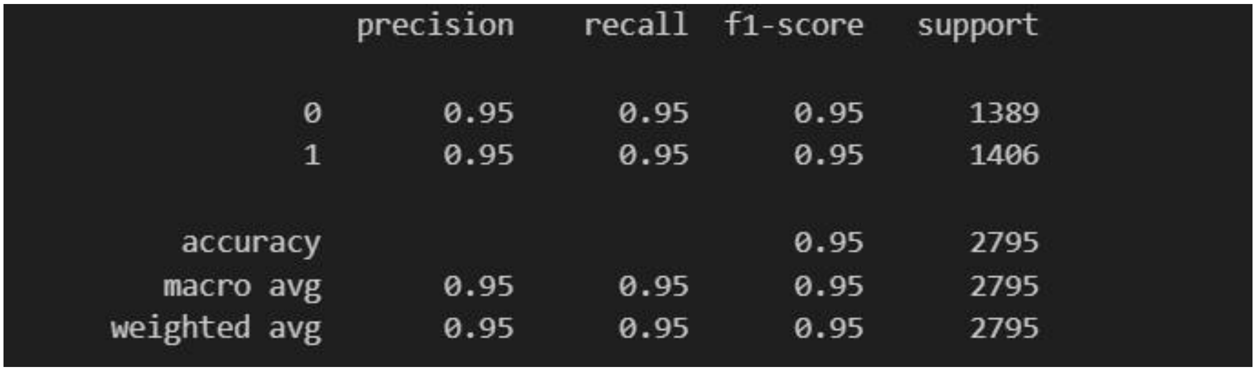
```
           crime_risk_index
district
indore                70190
patna                 69929
south                 69910
mumbai                65462
west                  65314
kolkata               60666
east                  60230
lucknow               53446
cyberabad             52274
chennai               49069
```

Classification of High vs. Low Crime Districts

– Binary target created using median split on total_ipc_crimes.
– Model: Random Forest Classifier
– Accuracy: 95%

Precision/Recall

Balanced performance indicates strong predictability from raw crime type features.



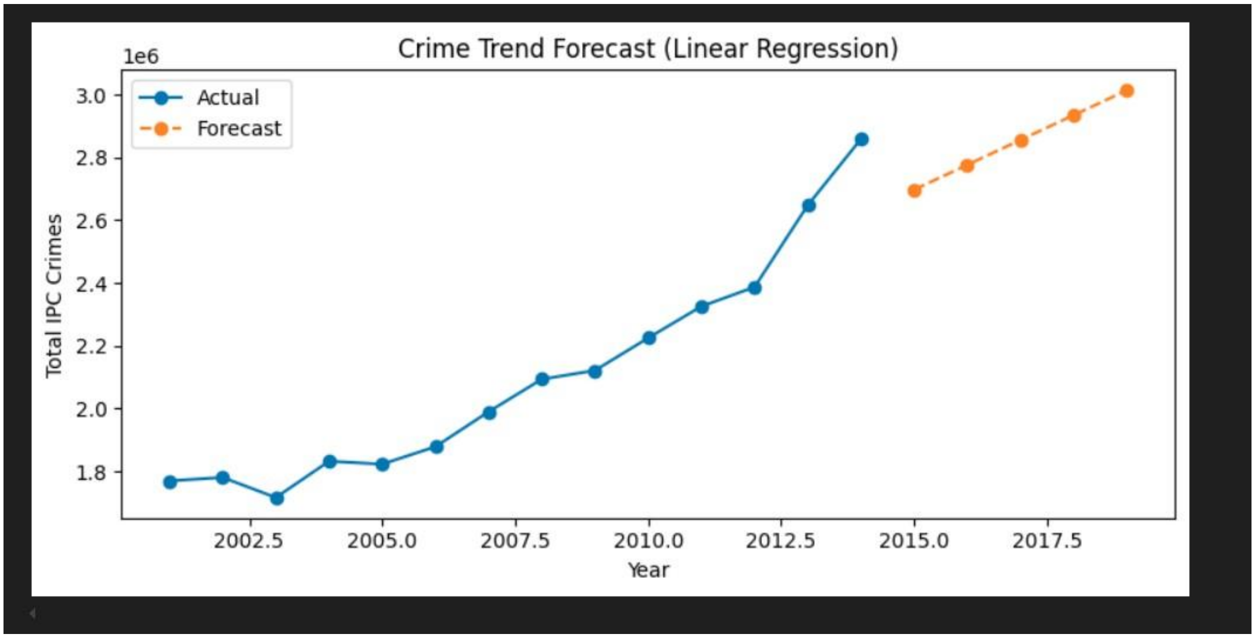
```
              precision    recall  f1-score   support

           0       0.95      0.95      0.95      1389
           1       0.95      0.95      0.95      1406

    accuracy                           0.95      2795
   macro avg       0.95      0.95      0.95      2795
weighted avg       0.95      0.95      0.95      2795
```

Time-Series Forecasting

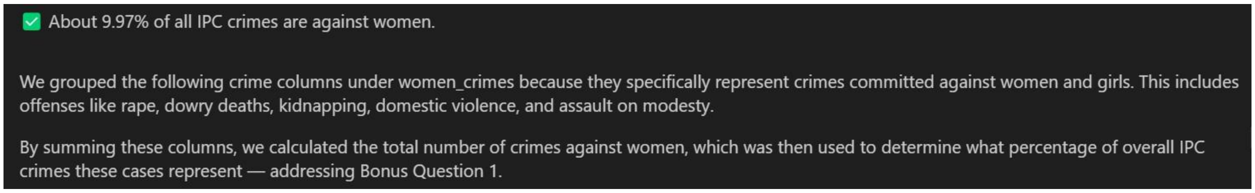
- Used linear regression to forecast IPC crime totals for upcoming years.
- Projected crimes for 2019: ~3 million.



Bonus Analyses

Crimes Against Women

- Grouped crimes like rape, dowry deaths, assault, and trafficking.
- Women-targeted crimes formed ~10% of total IPC crimes.



✅ About 9.97% of all IPC crimes are against women.

We grouped the following crime columns under women_crimes because they specifically represent crimes committed against women and girls. This includes offenses like rape, dowry deaths, kidnapping, domestic violence, and assault on modesty.

By summing these columns, we calculated the total number of crimes against women, which was then used to determine what percentage of overall IPC crimes these cases represent — addressing Bonus Question 1.

Dowry Deaths

Uttar Pradesh reported the highest number of dowry deaths.

Crime and Cities

– Urban centers exhibited higher crime volumes.
– City-based crimes like auto theft, robbery, and assault had above-average frequencies.

Conclusion

This analysis revealed stark regional differences in crime intensity, uncovered deep correlations across IPC sections, and enabled risk indexing of Indian districts. The use of machine learning models like classification and forecasting demonstrated practical value in strategic planning and crime prevention. Future enhancements may include monthly-level data to enable seasonal trend analysis and deeper socio-economic feature integration.