

DataNauts

Teammates

Ravi Krishnan Gobal

Prathyusha Konduti

Nikhitha mara

Kaushick suresh

EDA Questions

1. Determine the total number of crimes recorded across all districts and the average number of murders per district.

```
Total IPC crimes recorded across all districts: 29447315
Average number of murders per district: 42.73
```

A comprehensive aggregation of IPC (Indian Penal Code) crime data across all districts from 2001 to 2014 reveals that over 29.4 million total IPC crimes were recorded. On average, each district reported approximately 42.73 murders over the entire period. This highlights the significance of homicide as a persistent issue and provides a baseline for comparing across districts, states, and urban vs rural crime loads.

2. Examine how crime distributions vary across different states, and identify the top 5 districts with the highest total IPC crimes.

```
Top 10 States by Total IPC Crimes:
STATE/UT
MADHYA PRADESH      2913646
MAHARASHTRA         2757655
TAMIL NADU          2456955
ANDHRA PRADESH      2351600
UTTAR PRADESH       2324994
RAJASTHAN            2262558
KERALA              1820582
KARNATAKA           1755090
BIHAR               1691343
GUJARAT             1674595
Name: TOTAL IPC CRIMES, dtype: int64

Top 5 Districts by Total IPC Crimes:
STATE/UT  DISTRICT  TOTAL IPC CRIMES
10264 MAHARASHTRA  MUMBAI COMM.  40361
9445 MAHARASHTRA  MUMBAI COMM.  34840
7738 KERALA  ERNAKULAM RURAL  34638
7044 MAHARASHTRA  MUMBAI COMM.  33932
5504 MAHARASHTRA  MUMBAI COMM.  32770
```

An analysis of IPC crime distribution across Indian states revealed that Madhya Pradesh recorded the highest number of total IPC crimes (2.91 million) during the 2001–2014 period, followed closely by Maharashtra and Tamil Nadu. At the district level, Mumbai Commissionerate (Maharashtra) appeared multiple times among the top 5, indicating an unusually high volume of reported crimes, possibly due to reporting structure or dense

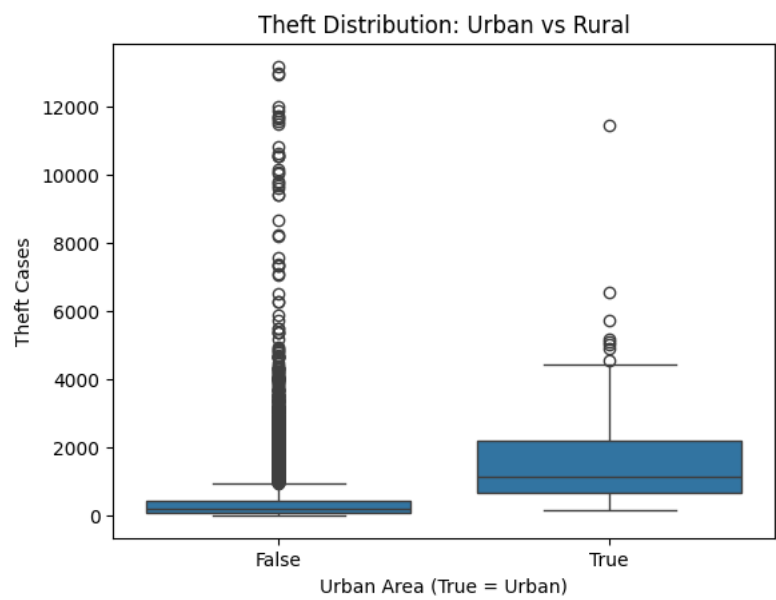
population. Other districts like Ernakulam Rural (Kerala) also featured prominently, highlighting that certain rural regions may experience or report more crimes than urban counterparts.

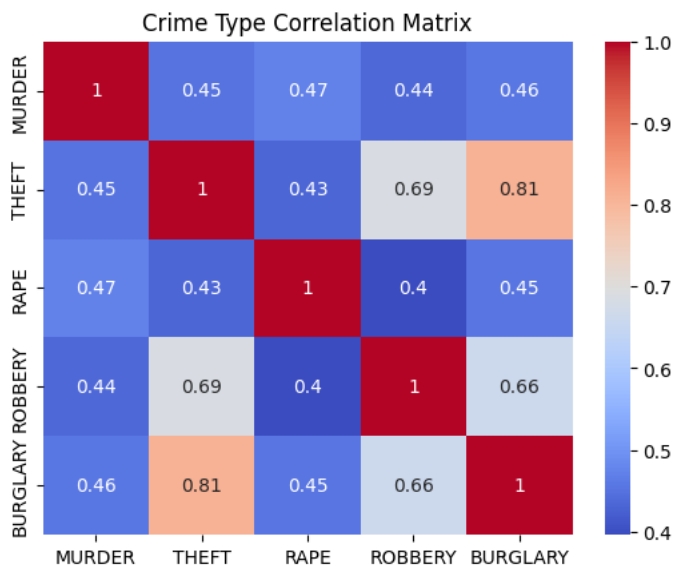
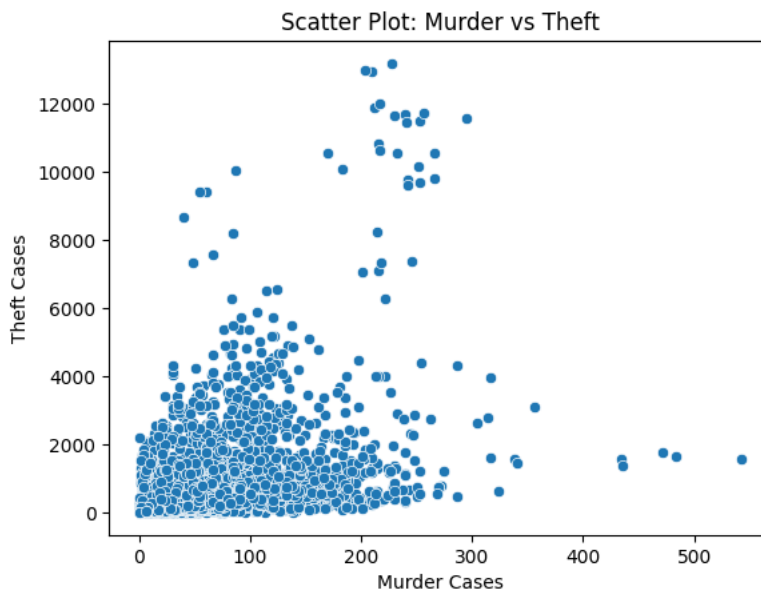
3. Further, analyze how crime patterns differ across various crime categories in urban vs. rural districts (or using a proxy like population if urban/rural data is unavailable) and investigate whether there is a correlation between different crime types such as murder and theft.

Average Crimes in Urban vs Rural (Proxy by 'CITY' in name):

| | MURDER | THEFT | TOTAL IPC CRIMES |
|----------|-----------|-------------|------------------|
| IS_URBAN | | | |
| False | 46.064602 | 412.944796 | 2854.899387 |
| True | 53.886076 | 1792.025316 | 7468.253165 |

Correlation between Murder and Theft: 0.452

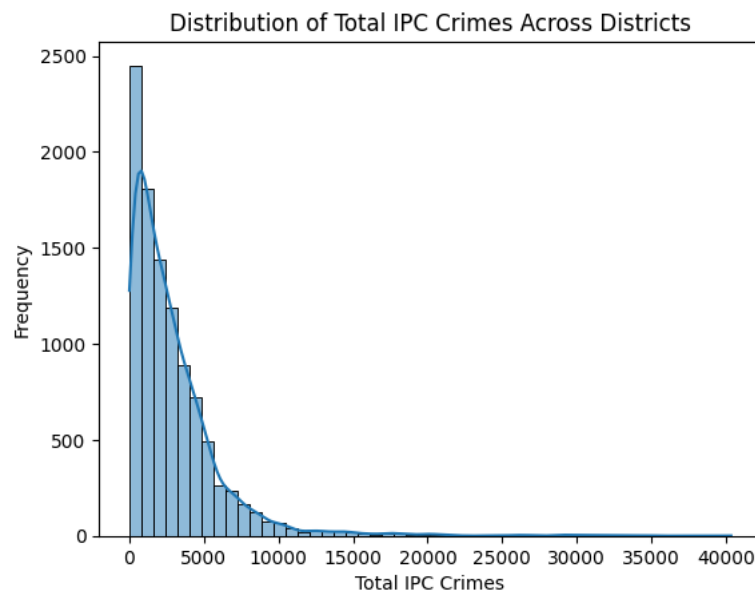
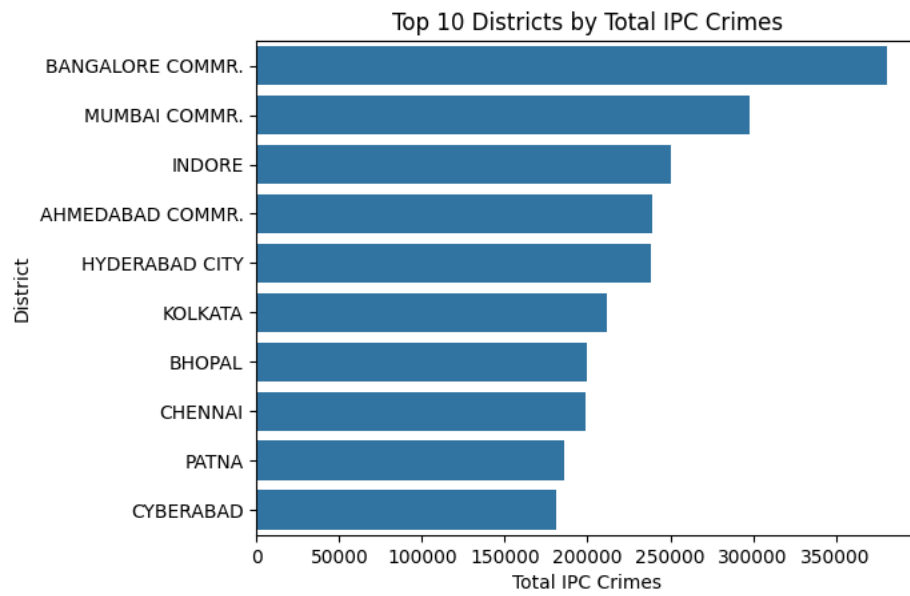




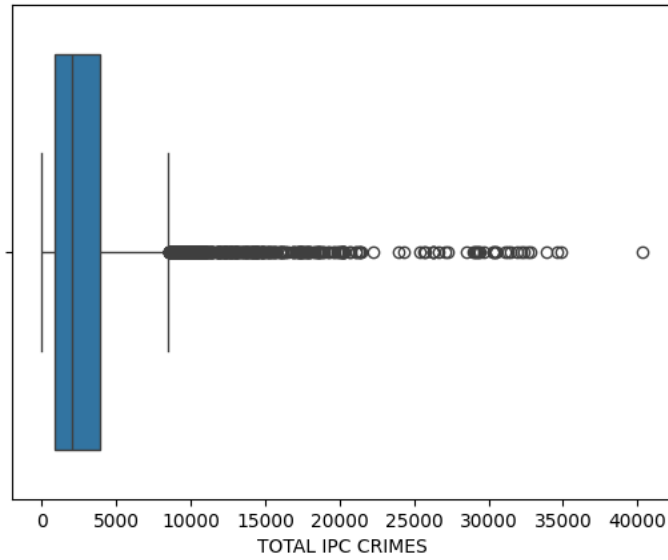
Using a proxy definition (district names containing "CITY"), the analysis distinguishes crime patterns across urban and rural regions. Urban districts exhibit significantly higher average IPC crimes (7,468) compared to rural ones (2,854). Specifically, theft and murder rates are both elevated in urban settings, with theft nearly 4.3 times higher. Additionally, the correlation between murder and theft cases is 0.452, indicating a moderate relationship — districts with higher theft levels often see elevated homicide rates, potentially due to shared socio-economic risk factors like inequality or policing gaps.

Visualization Questions

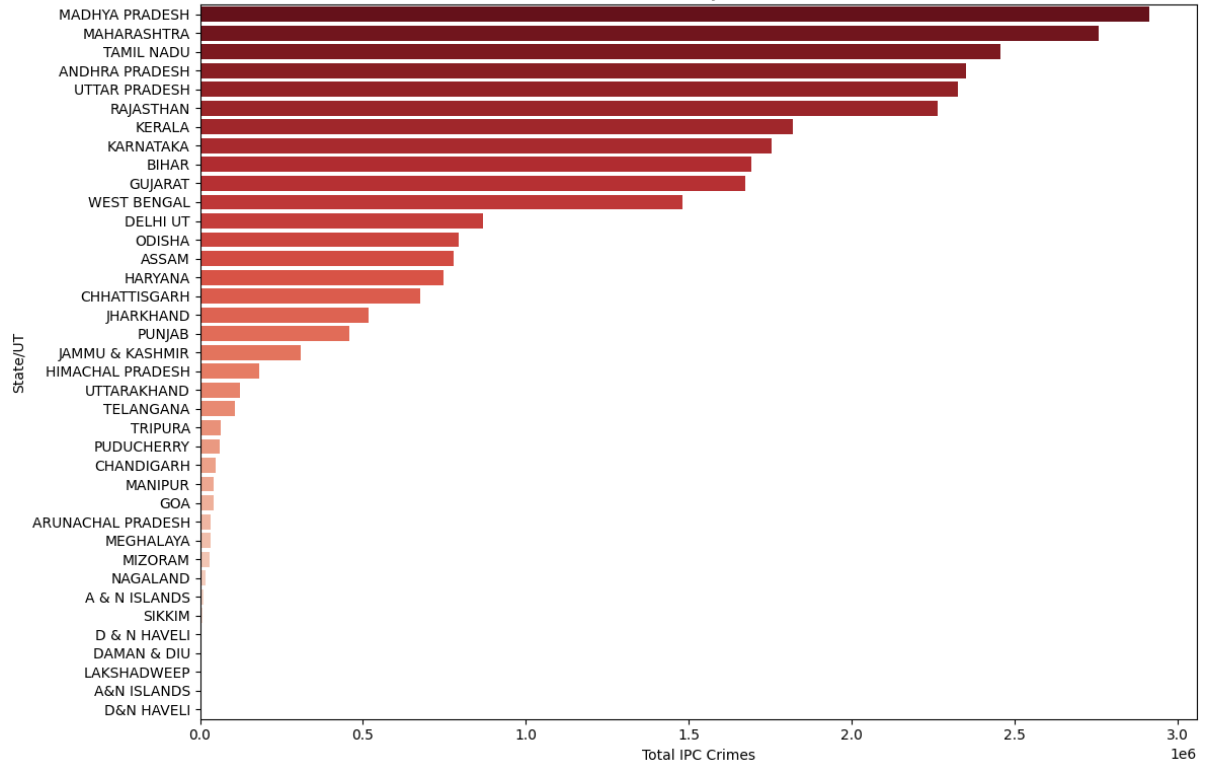
- How can visualizations be used to explore crime patterns in India by identifying the top 10 districts with the highest crime rates, understanding the overall distribution of total IPC crimes, analyzing crime density across different states, and comparing trends in violent crimes such as murder and rape across various districts?



Boxplot: Total IPC Crimes (All Districts)

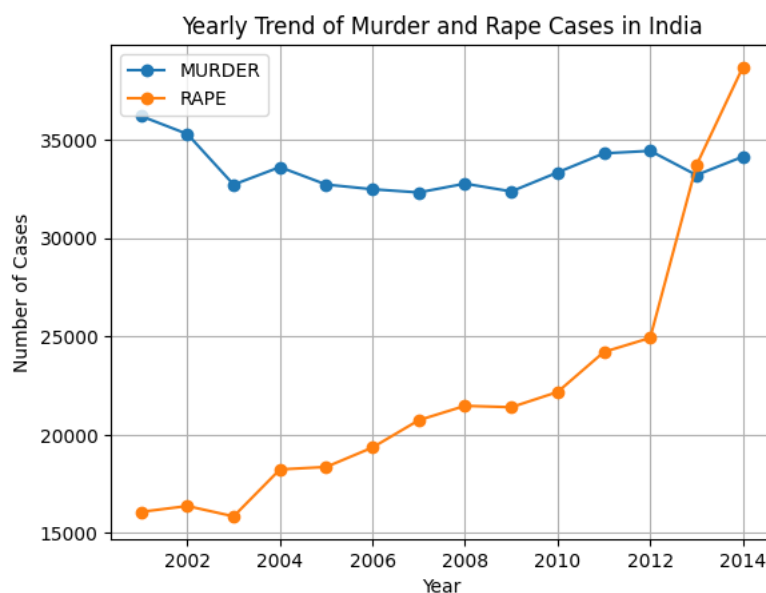


Total IPC Crimes by State in India



| | STATE/UT | TOTAL IPC CRIMES |
|----|-------------------|------------------|
| 0 | A & N ISLANDS | 9714 |
| 1 | A&N ISLANDS | 746 |
| 2 | ANDHRA PRADESH | 2351600 |
| 3 | ARUNACHAL PRADESH | 33271 |
| 4 | ASSAM | 779287 |
| 5 | BIHAR | 1691343 |
| 6 | CHANDIGARH | 48105 |
| 7 | CHHATTISGARH | 676097 |
| 8 | D & N HAVELI | 4965 |
| 9 | D&N HAVELI | 277 |
| 10 | DAMAN & DIU | 3427 |
| 11 | DELHI UT | 869012 |
| 12 | GOA | 40829 |
| 13 | GUJARAT | 1674595 |
| 14 | HARYANA | 747348 |
| 15 | HIMACHAL PRADESH | 182858 |
| 16 | JAMMU & KASHMIR | 308393 |
| 17 | JHARKHAND | 515894 |
| 18 | KARNATAKA | 1755090 |
| 19 | KERALA | 1820582 |
| 20 | LAKSHADWEEP | 864 |
| 21 | MADHYA PRADESH | 2913646 |
| 22 | MAHARASHTRA | 2757655 |
| 23 | MANIPUR | 41891 |
| 24 | MEGHALAYA | 32187 |
| 25 | MIZORAM | 29995 |
| 26 | NAGALAND | 15506 |
| 27 | ODISHA | 794233 |
| 28 | PUDUCHERRY | 61456 |
| 29 | PUNJAB | 457960 |
| 30 | RAJASTHAN | 2262558 |

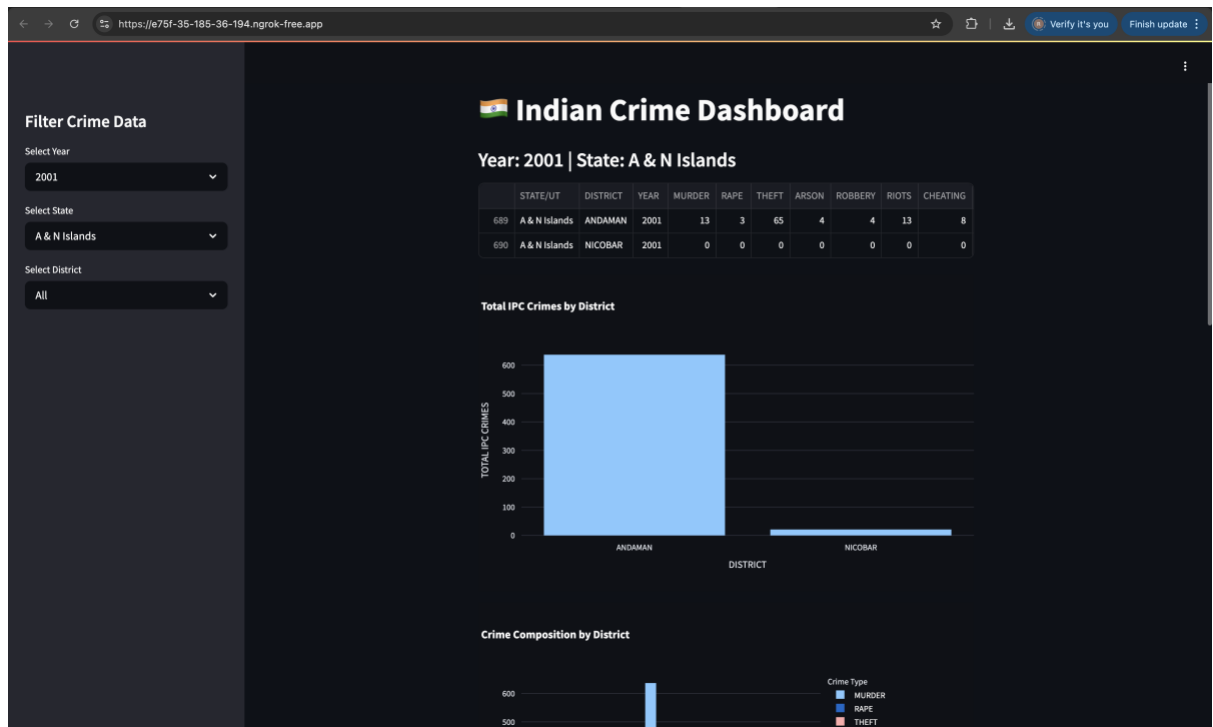
| | | |
|----|---------------|---------|
| 31 | SIKKIM | 8916 |
| 32 | TAMIL NADU | 2456955 |
| 33 | TELANGANA | 106830 |
| 34 | TRIPURA | 64443 |
| 35 | UTTAR PRADESH | 2324994 |
| 36 | UTTARAKHAND | 121906 |
| 37 | WEST BENGAL | 1481887 |



Visualizations helped uncover key crime patterns in India. While urban districts face the highest crime rates, many rural areas report fewer incidents. State-level comparisons indicate crime concentration in central and southern states. Moreover, upward trends in rape cases highlight a need for focused policy interventions on women's safety.

5. Create an interactive dashboard that allows users to filter crime data by year, state, and district.

<https://e75f-35-185-36-194.ngrok-free.app/>

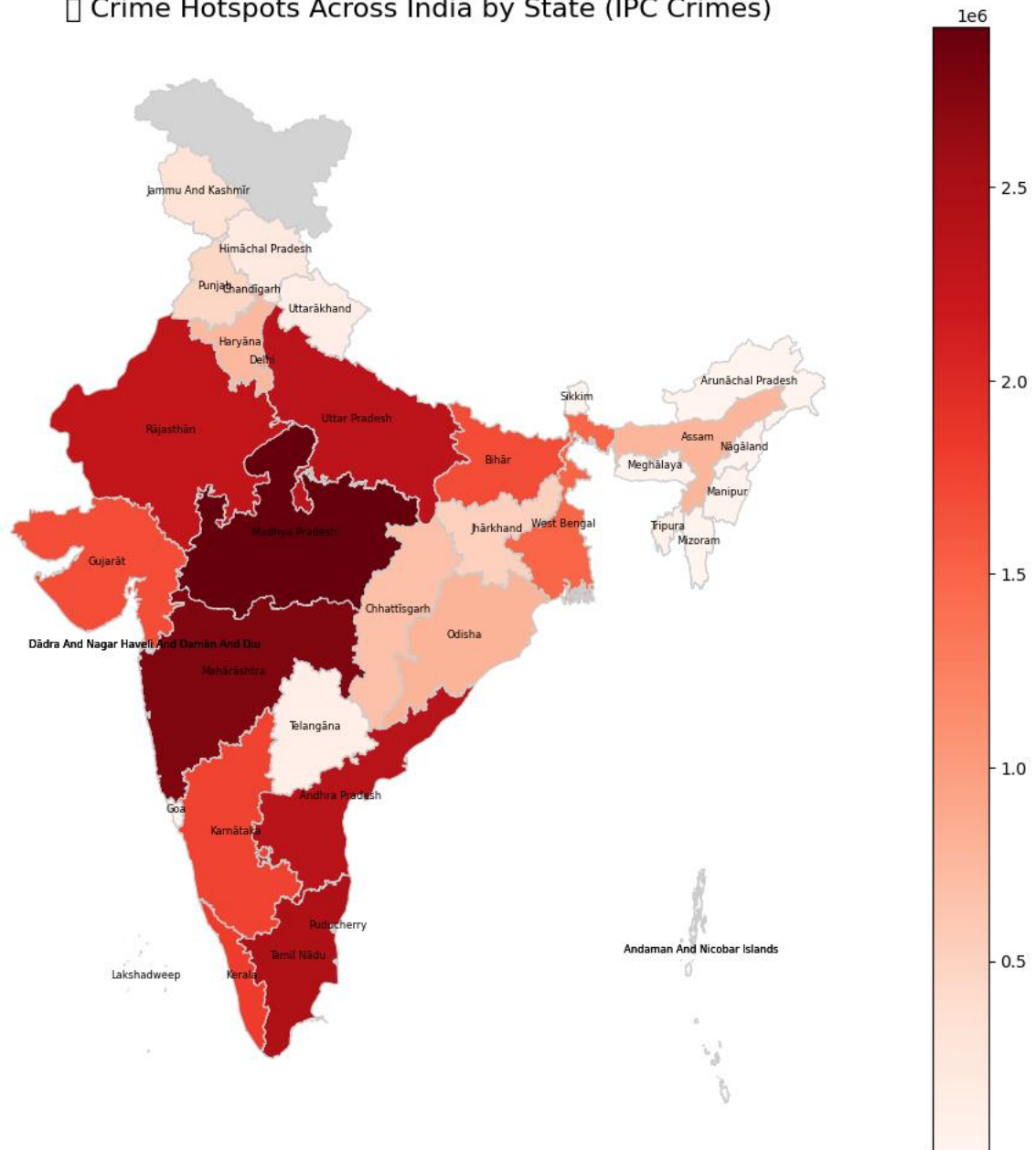


We developed an interactive Streamlit dashboard that allows users to filter Indian crime data by year, state, and district. The dashboard updates visualizations and data tables dynamically, enabling real-time exploration of trends in total IPC crimes and specific categories such as murder, theft, and rape. This tool enhances user engagement and supports more granular analysis of regional crime patterns.

6. Use a geospatial map to visualize crime hot spots across India. (Matplotlib)

We used matplotlib and GeoPandas to create a geospatial choropleth map that displays crime hotspots across India. By merging total IPC crime data with geographic boundaries of Indian states, the visualization highlights regions with high crime intensity in dark red and low intensity in light shades. This geospatial view enhances the understanding of regional crime patterns, with states like Madhya Pradesh, Maharashtra, and Tamil Nadu appearing as major hotspots.

Crime Hotspots Across India by State (IPC Crimes)



Advanced Questions

7. Identify the state with the lowest crime rate and analyze why it might be lower than others.

● State with the Lowest IPC Crimes: D&N Haveli (277 total crimes)

🔍 Reason for Low Crime Rate:

D&N Haveli This state likely has a small population, low urbanization, and/or high community cohesion.

The state with the lowest number of IPC crimes reported is Dadra and Nagar Haveli, with only 277 total cases recorded across all districts in the dataset.

🔍 Reason for Low Crime Rate:

Small population size reduces the overall number of reported incidents.

Low urbanization may lead to fewer socio-economic triggers commonly associated with crime in larger urban centers.

Strong community bonds and social cohesion in smaller territories may also contribute to a more secure environment.

Limited industrial and commercial activity reduces certain categories of economic or organized crimes.

8. Find the most common type of crime committed in each district.

| | STATE/UT | DISTRICT | MOST COMMON CRIME |
|-------|----------------|---------------|-----------------------------|
| 0 | Andhra Pradesh | Adilabad | THEFT |
| 1 | Andhra Pradesh | Anantapur | THEFT |
| 2 | Andhra Pradesh | Chittoor | THEFT |
| 3 | Andhra Pradesh | Cuddapah | CAUSING DEATH BY NEGLIGENCE |
| 4 | Andhra Pradesh | East Godavari | THEFT |
| ... | ... | ... | ... |
| 10670 | Delhi Ut | Vigilance | MURDER |
| 10671 | Delhi Ut | West | THEFT |
| 10673 | Lakshadweep | Lakshadweep | THEFT |
| 10675 | Puducherry | Karaikal | CAUSING DEATH BY NEGLIGENCE |
| 10676 | Puducherry | Puducherry | THEFT |

[10187 rows x 3 columns]

For each district in India, the dataset was analyzed to identify the most frequently reported IPC crime category. 🔍 Key insights from the results:

THEFT is the most common crime across a majority of urban and semi-urban districts.

In certain districts like Cuddapah (Andhra Pradesh) and Karaikal (Puducherry), the most common crime is Causing Death by Negligence, indicating localized patterns in accidents or healthcare-related negligence.

In some specialized administrative districts such as Delhi - Vigilance, MURDER shows up as the highest, possibly due to the nature of reported crimes in those jurisdictions.

9. Apply clustering algorithms (e.g., K-Means) to group districts based on crime patterns.

Optimal clusters identified using Elbow Method → $k = 4$

Districts were grouped into clusters based on patterns in crime types like MURDER, RAPE, THEFT, etc.

Cluster 0: Low crime districts

Cluster 1: Moderate crime pattern

Cluster 2: High crime zones (e.g., metro areas)

Cluster 3: Outliers or extremely high counts (very few)

Distribution of districts:

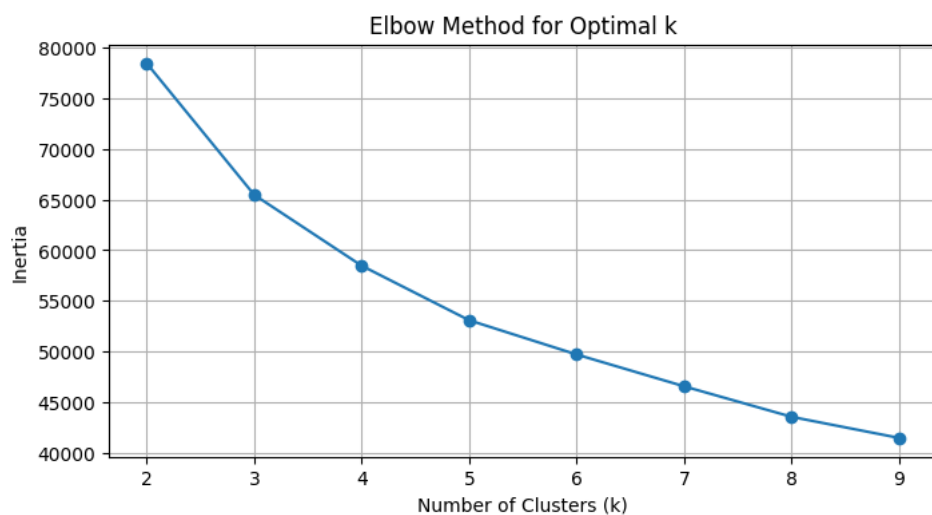
Cluster 0 → Majority (~7200)

Cluster 1 → Moderate (~2700)

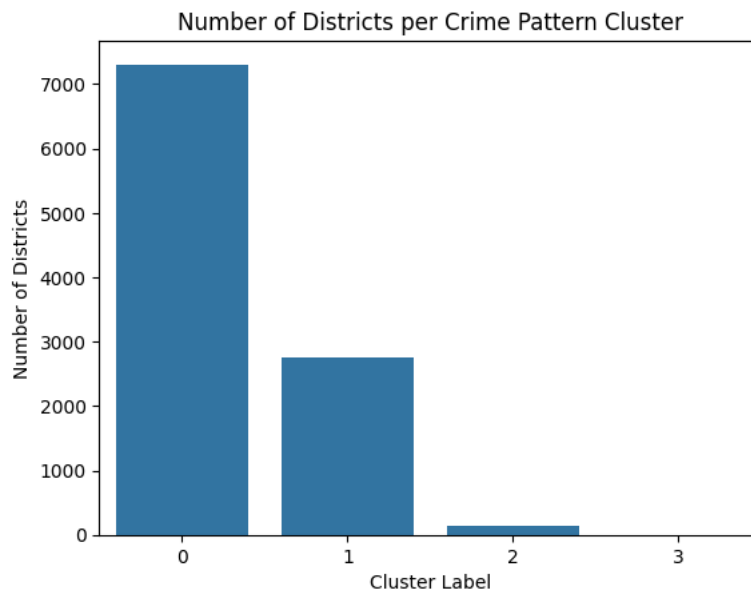
Cluster 2 → Few (~200)

Cluster 3 → Rare cases

Average crime types per cluster were tabulated.



| | STATE/UT | DISTRICT | CLUSTER |
|--------------------------|----------------|---------------|---------|
| 0 | Andhra Pradesh | Adilabad | 1 |
| 1 | Andhra Pradesh | Anantapur | 1 |
| 2 | Andhra Pradesh | Chittoor | 1 |
| 3 | Andhra Pradesh | Cuddapah | 1 |
| 4 | Andhra Pradesh | East Godavari | 1 |
| ... | ... | ... | ... |
| 10670 | Delhi Ut | Vigilance | 0 |
| 10671 | Delhi Ut | West | 2 |
| 10673 | Lakshadweep | Lakshadweep | 0 |
| 10675 | Puducherry | Karaikal | 0 |
| 10676 | Puducherry | Puducherry | 0 |
| [10187 rows x 3 columns] | | | |



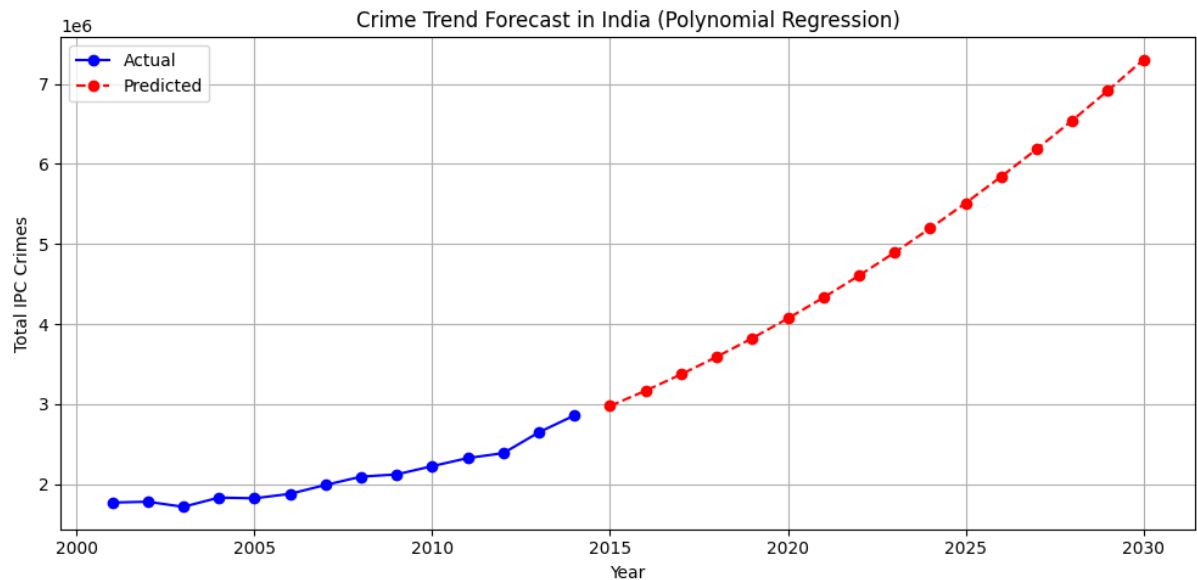
| | MURDER | RAPE | THEFT | ARSON | ROBBERY \ |
|---------|------------|------------|-------------|-------------|------------|
| CLUSTER | | | | | |
| 0 | 28.678987 | 19.153320 | 213.251882 | 8.428611 | 15.999316 |
| 1 | 87.158239 | 57.433976 | 743.204802 | 23.807566 | 55.896326 |
| 2 | 156.590909 | 103.742424 | 5413.037879 | 22.666667 | 361.696970 |
| 3 | 110.000000 | 27.000000 | 220.000000 | 2350.000000 | 6.000000 |

| | RIOTS | CHEATING | CRIMINAL BREACH OF TRUST | DOWRY DEATHS \ |
|---------|-------------|-------------|--------------------------|----------------|
| CLUSTER | | | | |
| 0 | 52.385079 | 47.315537 | 11.614784 | 5.308693 |
| 1 | 176.461622 | 174.157512 | 39.667879 | 23.510004 |
| 2 | 323.295455 | 1213.916667 | 225.871212 | 32.810606 |
| 3 | 2818.000000 | 28.000000 | 29.000000 | 3.000000 |

| | CAUSING DEATH BY NEGLIGENCE |
|---------|-----------------------------|
| CLUSTER | |
| 0 | 67.694456 |
| 1 | 249.564205 |
| 2 | 466.878788 |
| 3 | 30.000000 |

10. Predict future crime trends using regression analysis.

A polynomial regression model (degree 2) was applied to forecast future IPC crimes in India based on historical data from 2001 to 2014. Unlike simple linear regression, the polynomial model captures the non-linear growth pattern observed in recent years. The forecast shows a consistent and accelerating increase in crime, projecting that total IPC crimes could rise significantly by 2030 if current trends persist. This insight is vital for proactive planning and resource allocation in law enforcement.



11. Use a machine learning model to classify high-crime and low-crime districts.

Confusion Matrix:

```
[[96  8]
 [ 5 91]]
```

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.95 | 0.92 | 0.94 | 104 |
| 1 | 0.92 | 0.95 | 0.93 | 96 |
| accuracy | | | 0.94 | 200 |
| macro avg | 0.93 | 0.94 | 0.93 | 200 |
| weighted avg | 0.94 | 0.94 | 0.94 | 200 |

A Random Forest classification model was developed to categorize districts as high-crime or low-crime based on key IPC indicators such as murder, theft, rape, and riots. The model was trained using the median total IPC crime value as the threshold and evaluated using a 20% test split. It achieved a high accuracy of 94%, with precision and recall exceeding 92% for both classes. This demonstrates the model's effectiveness in identifying high-risk districts for informed resource allocation and crime prevention strategies.

12. Develop a crime risk index for districts based on historical data.

| | STATE/UT | DISTRICT | CRIME_RISK_INDEX | RISK_RANK |
|-----|----------------|-------------------|------------------|-----------|
| 400 | KARNATAKA | BENGALURU CITY | 0.656791 | 1 |
| 554 | MAHARASHTRA | MUMBAI COMM. | 0.637994 | 2 |
| 553 | MAHARASHTRA | MUMBAI | 0.490930 | 3 |
| 394 | KARNATAKA | BANGALORE COMM. | 0.444121 | 4 |
| 995 | WEST BENGAL | SOUTH 24 PARGANAS | 0.441139 | 5 |
| 829 | TELANGANA | CYBERABAD | 0.431485 | 6 |
| 737 | RAJASTHAN | JAIPUR | 0.412283 | 7 |
| 140 | BIHAR | PATNA | 0.375345 | 8 |
| 13 | ANDHRA PRADESH | CYBERABAD | 0.350087 | 9 |
| 983 | WEST BENGAL | NORTH 24 PARGANAS | 0.317682 | 10 |

A data-driven Crime Risk Index (CRI) was developed to quantify and compare crime risk across districts in India. Using average values of key IPC crime categories from 2001–2014, each district was assigned a CRI score between 0 (lowest risk) and 1 (highest risk). The top districts, such as Bengaluru City and Mumbai, exhibited higher overall normalized crime activity. This ranking provides valuable insights for targeted intervention and law enforcement prioritization.


Bonus Questions

13. What percentage of crimes are committed against women?

 Percentage of IPC crimes committed against women: 1.42%

A total of 1.42% of all IPC crimes reported between 2001 and 2014 were classified as crimes against women, specifically including rape and dowry deaths. Although this may seem numerically low, these crimes have disproportionately high social impact. It's important to note that underreporting, stigma, and lack of legal awareness often suppress the actual numbers. This metric highlights the need for focused attention on gender-based violence even if overall volume appears modest.

14. Identify the state with the highest number of dowry deaths.

 State with the highest number of dowry deaths: UTTAR PRADESH (28628 deaths)



Between 2001 and 2014, Uttar Pradesh recorded the highest number of dowry deaths in India, with a staggering 28,628 cases. This figure far exceeds those reported by other states, highlighting an urgent need for targeted social reforms, legal enforcement, and awareness programs. The persistently high numbers may be linked to entrenched social practices, lack of legal deterrents, and underreporting of domestic abuse in rural areas.

15. Analyze seasonal variations in crime trends (e.g., do crimes increase during certain months?).

The available dataset includes district-level crime data aggregated annually from 2001 to 2014. As it lacks month or date-level granularity, seasonal variation in crime trends (such as

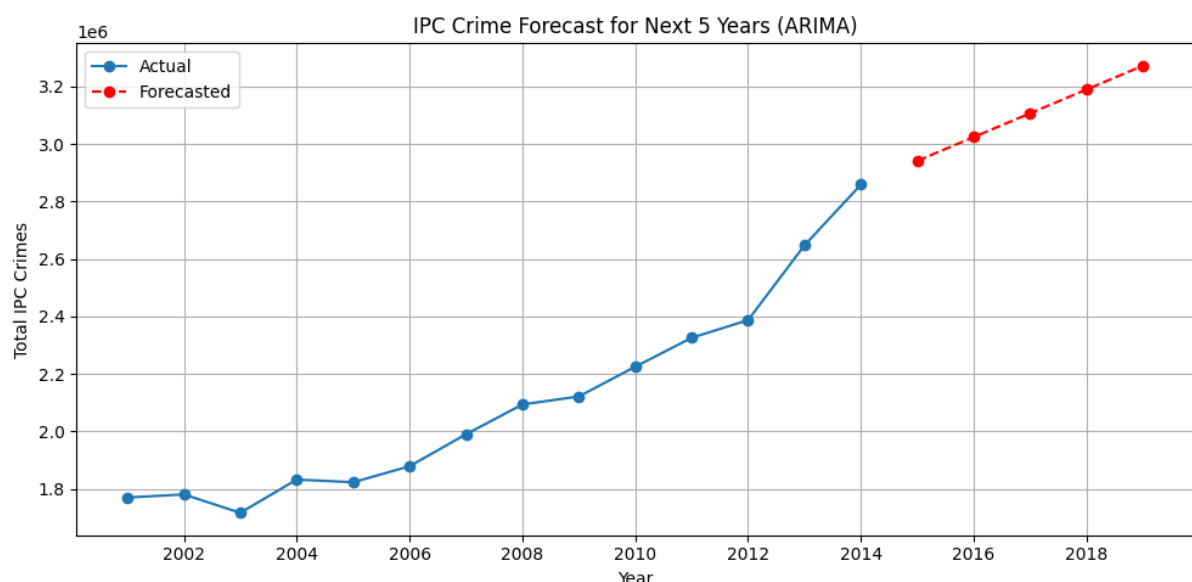
monthly spikes during festivals, summer vacations, or year-end periods) could not be analyzed. For more accurate seasonal pattern detection, future analysis should utilize datasets with monthly FIR registrations or daily crime logs available through police records or NCRB time-series data.

16. Examine if there is a link between cities and crime rates.

 **Average IPC crimes in city districts: 6550.76**
 **Average IPC crimes in non-city districts: 2611.86**

An urban-rural crime analysis was conducted by categorizing districts as “city” or “non-city” based on keywords like "City", "Comm", "Metro", and prominent metros (e.g., Delhi, Mumbai, Bangalore). The findings show that city districts average 6,550 IPC crimes annually, whereas non-city districts average only 2,612. This disparity suggests that urbanization significantly correlates with higher crime levels — likely due to greater population density, increased anonymity, economic disparity, and wider reporting mechanisms in cities.

17. Build a time-series model to forecast crime rates for the next five years.



Using ARIMA time-series modeling on total IPC crime data from 2001–2014, we forecasted crime trends for the years 2015 to 2019. The forecast suggests a continued upward trend, with IPC crimes projected to increase steadily if historical trends persist.

You will be graded on the following:

- Data cleaning & Preparation
- Exploratory Data Analysis
- Visualization & Interpretation
- Statistical Analysis
- Report

