**Statistical Analysis Summary (2001–2014)**

**Overall Dataset Stats**

| **Metric** | **Value** |
| --- | --- |
| **Total IPC Crimes (2001–2014)** | 58,894,630 |
| **Avg. Murders per District-Year** | 88.01 |
| **Std. Dev of Total IPC Crimes** | 19,397.65 |
| **Years Covered** | 2001 to 2014 |
| **Total Observations** | 10,678 rows |

**Top 5 States with Highest Total IPC Crimes**

| **State** | **Total Crimes** |
| --- | --- |
| Madhya Pradesh | 5,827,292 |
| Maharashtra | 5,515,310 |
| Tamil Nadu | 4,913,910 |
| Andhra Pradesh | 4,703,200 |
| Uttar Pradesh | 4,649,988 |

**Top 5 Districts with Highest Crimes**

| **District** | **Total Crimes** |
| --- | --- |
| TOTAL (aggregates) | 28,814,141 |
| Delhi UT TOTAL | 633,174 |
| Bangalore Commr. | 380,665 |
| Mumbai Commr. | 297,871 |
| Indore | 250,639 |

Note: "TOTAL" represents aggregated rows — exclude these in granular analysis.

**Crime Trend by Year (First 5 Years)**

| **Year** | **Total IPC Crimes** |
| --- | --- |
| 2001 | 3,538,616 |
| 2002 | 3,560,660 |
| 2003 | 3,432,240 |
| 2004 | 3,664,020 |
| 2005 | 3,645,204 |

The crime numbers show consistent high volumes across years, with some variance due to population, reporting standards, or policy changes.

### Report **Data Loading and Setup**

* You used Google Colab to mount Google Drive and upload your CSV file:  
  Districtwise\_Crime\_of\_India\_2001\_to\_2014.csv.
* The file was read into a pandas DataFrame using pd.read\_csv().

**Previewing the Dataset**

* df.head() showed the first 5 rows.
* This gave an overview of key columns like:
  + STATE/UT, DISTRICT, YEAR
  + Crime types: MURDER, RAPE, THEFT, DOWRY DEATHS, TOTAL IPC CRIMES, etc.

**Dataset Information**

* df.info() showed:
* Number of rows and columns
* Data types of each column (e.g., int64, object)
* Presence of null values in any column

This helped assess whether data cleaning is needed and how large the dataset is.

**Summary Statistics**

* df.describe() provided:
  + Count, mean, min, max, and percentiles for each numeric column
  + Useful to detect skewed features and potential outliers (e.g., very high theft or murder counts in a few districts)

**Missing Value Check**

* df.isnull().sum() showed how many values were missing in each column.
* This tells you whether imputation or row removal will be needed.

**Total IPC Crimes & Average Murders**

This section calculates two high-level statistics:

* **Total IPC crimes** (Indian Penal Code crimes) across all districts
* **Average number of murders** per district over the recorded period

**Result:**

* Total IPC crimes recorded: **58,894,630**
* Average murders per district: **88.01**

This provides a **macro-level understanding** of the scale and severity of reported crimes in India between 2001 and 2014.

**State and District Level Crime Distribution**

We grouped data by STATE/UT to calculate crime volumes by state and identified the top 5 **districts** (excluding aggregate rows like "ZZ TOTAL").

**Findings:**

* **Madhya Pradesh**, **Maharashtra**, and **Tamil Nadu** reported the **highest total IPC crimes** among states.
* Top districts were based on crime counts, but some still appeared as "TOTAL" — indicating **a need for data cleaning**.

**Urban vs Rural Crime Pattern**

Districts were labeled:

* **Urban-like** = total IPC crimes ≥ 75th percentile (Q3)
* **Rural-like** = below Q3

Then, average crime rates were compared across categories like MURDER, RAPE, THEFT, DOWRY DEATHS, and RIOTS.

**Graph: Grouped bar chart – Average Crimes by Area Type**

**Insights:**

* Urban-like districts report **significantly more crime** across all types.
* **Theft** is particularly higher in urban areas.
* Confirms impact of **population density, reporting infrastructure**, and **urbanization**.

A graph with red and blue bars

AI-generated content may be incorrect.

* **X-axis**: Different crime types:
  + MURDER
  + RAPE
  + THEFT
  + DOWRY DEATHS
  + RIOTS
* **Y-axis**: The **average number of crimes** recorded in each area type for each category.
* **Colors**:
  + Blue bars represent **Rural-like** districts (below 75th percentile of total IPC crimes)
  + Red bars represent **Urban-like** districts (top 25% in IPC crimes)

### **Key Insights from the Graph:**

1. **THEFT dominates urban crime**
   * Urban-like areas show an **average theft count exceeding 2500**, while rural areas average below 250.
   * Suggests theft is highly concentrated in cities due to higher population density, economic inequality, or better reporting.
2. **Violent crimes like MURDER and RAPE are also more common in urban zones**
   * Urban-like areas report roughly **5–6x higher MURDER and RAPE** averages than rural counterparts.
   * Indicates possible urban stress factors, surveillance/reporting gaps, or higher exposure risks.
3. **RIOTS and DOWRY DEATHS follow the same trend**
   * RIOTS average: ~500 in urban areas vs <100 in rural
   * DOWRY DEATHS are present in both, but significantly more in urban zones.

**Correlation Analysis of Crime Types**

We used a **correlation heatmap** to examine relationships between major crimes.

**Top correlations:**

* MURDER & DOWRY DEATHS: **0.91**
* MURDER & THEFT: **0.80**
* RAPE & MURDER: **0.79**
  + Strong positive correlations suggest **co-occurrence of crimes** or shared socio-economic triggers in certain districts.

A screenshot of a chart

AI-generated content may be incorrect.

**Key Observations & Interpretation:**

**Strongest Relationships:**

| **Crime Pair** | **Correlation Value** | **Meaning** |
| --- | --- | --- |
| **MURDER – DOWRY DEATHS** | **0.91** | Extremely strong correlation. Districts with high murder counts also report more dowry deaths. Could indicate gender-based or domestic violence patterns. |
| **MURDER – THEFT** | **0.80** | Surprisingly strong — suggests that violence and theft may co-occur in certain high-crime areas. |
| **RAPE – THEFT** | **0.80** | Possible link between overall lawlessness or poor policing infrastructure. |
| **RAPE – MURDER** | **0.79** | High correlation. May indicate underlying systemic issues in certain districts. |

**Moderate Relationships:**

| **Crime Pair** | **Correlation** | **Insight** |
| --- | --- | --- |
| **RIOTS – MURDER** | 0.76 | Communal violence or political unrest may lead to spikes in murders. |
| **THEFT – RIOTS** | 0.69 | During riots, property crimes may also increase. |

**Lower Relationships:**

| **Crime Pair** | **Correlation** | **Insight** |
| --- | --- | --- |
| **RIOTS – DOWRY DEATHS** | 0.63 | Not strongly linked – suggests different root causes. |
| **RAPE – RIOTS** | 0.64 | Slight relationship, but less consistent across districts. |

**Top 10 Districts by IPC Crimes**

We visualized the **top 10 districts** with the highest IPC crime totals using a horizontal bar chart with state-based coloring (hue).

**Observation:**

* Major metro zones like those in **Andhra Pradesh**, **Uttar Pradesh**, and **Maharashtra** dominate.
* But some labels like "TOTAL" again indicate **state-level aggregates** sneaking into district rankings — a data issue to resolve.  
  A red rectangular object with numbers

  AI-generated content may be incorrect.  
  **What the graph is supposed to show:**
* **Y-axis (vertical)**: District names
* **X-axis (horizontal)**: Total IPC crimes reported
* **Bars**: Represent the volume of IPC crimes in each district
* **Colors**: Represent the state or union territory the district belongs to
* Blue = Madhya Pradesh
* Orange = Maharashtra
* Green = Uttar Pradesh
* Red = Andhra Pradesh

**Issue in the graph:**

* The only visible district on the Y-axis is "TOTAL" — which seems to be a **state-level aggregate row**, not an individual district.
* This suggests that:
* You might have **included aggregate rows** like "ZZ TOTAL" or "TOTAL" in your top 10 filtering logic.
* These rows represent **summed values** for entire states, and not specific districts.

**Distribution of Total IPC Crimes Across Districts**

Using a **histogram + KDE curve**, we analyzed the distribution of IPC crimes across all districts.

**Result:**

* The distribution is **right-skewed**.
* Most districts reported **lower to moderate crime levels**.
* A few districts (urban/metropolitan) are **extreme outliers**.

This helps understand the **data shape** before modeling — we may consider **log scaling or normalization**.  
A graph of a number of data

AI-generated content may be incorrect.

**What this graph shows:**

* **X-axis**: Number of IPC (Indian Penal Code) crimes reported in each district
* **Y-axis**: Number of districts (frequency) falling into each crime count range
* The **bars** represent counts of districts (histogram)
* The **red curve** is a KDE line showing the **smoothed probability density** — think of it as a more fluid version of the bar shape

**Key Observations:**

1. **Highly Right-Skewed Distribution**
   * Most districts report **lower crime numbers** — typically fewer than 10,000 crimes.
   * A **very small number of districts** have **extremely high crime counts**, creating a long tail to the right.
2. **Peak Frequency Near Lower Crime Values**
   * The tall peak on the left indicates that **many districts report similar, low levels** of IPC crimes (possibly < 5,000).
3. **Outliers and High-Crime Zones**
   * A few districts report over **100,000–250,000 crimes**, which are rare and significantly skew the distribution.

**Clustering: Murder vs Rape Analysis**

Using **KMeans clustering**, districts were grouped based on their MURDER and RAPE statistics.

**Visual Output:**

* A **scatterplot** shows four clear clusters.

**Insight:**

* Cluster 0: Low murder, low rape
* Cluster 3: High rape, moderate murder
* Others show distinct crime behavior patterns.

A graph with colored dots

AI-generated content may be incorrect.

**What the graph shows:**

* **X-axis**: Number of **murders** reported per district
* **Y-axis**: Number of **rape** cases reported per district
* **Each dot**: Represents one **district**
* **Colors**: Indicate the **cluster** to which a district belongs (from K-Means output)
  + Cluster 0 → Green
  + Cluster 1 → Orange
  + Cluster 2 → Blue
  + Cluster 3 → Pink

**Key Interpretations:**

**1. Crime Grouping Using Clustering**

* The districts are grouped based on similar murder and rape values.
* Clustering helps **identify similar crime behavior patterns** across districts.

**2. Cluster Patterns**

* **Cluster 0 (Green)**:
  + Low murder, low rape — **low-crime districts**
* **Cluster 1 (Orange)**:
  + Moderate murder and rape — **mid-range crime zones**
* **Cluster 3 (Pink)**:
  + High rape cases, moderate murders — **rape-heavy districts**
* **Cluster 2 (Blue)**:
  + High murder rates, possibly outliers — **high-violence districts**

**3. Positive Correlation**

* As murder numbers increase, rape numbers **generally** also increase (especially for orange and pink clusters).
* Shows **co-occurrence trends** — districts with one type of violent crime often have more of the other.

A graph with colored dots

AI-generated content may be incorrect.X-axis: Number of murders reported per district.

Y-axis: Number of rape cases reported per district.

Each dot: Represents one district in India.

Colors: Indicate clusters generated using the K-Means algorithm:

Cluster 0: Low in both crimes.

Cluster 1: Moderate murder, moderate rape.

Cluster 2: Very high murders, but moderate rape.

Cluster 3: High rape, moderate murder.

Key Insights:

1. Clear Segmentation of Crime Levels

Cluster 0 (Green): Most districts fall here with low murder and rape counts — likely rural or safer regions.

Cluster 1 (Orange): A large group of moderate-crime districts, suggesting urban areas with average reporting.

Cluster 3 (Pink): Districts with very high rape counts, showing that sexual violence is disproportionately high compared to murder.

Cluster 2 (Blue): Districts with extremely high murder counts, likely hotspots for violent crimes — possible gang zones or conflict areas.

2. Positive Correlation Trend

There's a noticeable positive relationship — as murder increases, rape also tends to increase, although not perfectly linearly.

3. Outliers

### A few districts are extreme outliers in both murder (>7000) and rape (>5000) cases, suggesting serious policing or social concerns. **Objective**

The primary goal of this analysis is to:

* Identify crime hotspots across Indian states
* Understand how crimes like **murder** and **rape** vary across districts
* Apply **clustering** to categorize districts based on crime severity
* Visualize state-wise total IPC crimes using a **geospatial map**

**Dataset Overview**

* The dataset contains **district-wise crime data** from **2001 to 2014** across all Indian states and union territories.
* Each record includes values for **MURDER**, **RAPE**, **TOTAL IPC CRIMES**, and more.

**Clustering Interpretation (Murder vs Rape Analysis)**

Using **K-Means Clustering**, districts were grouped based on the number of murders and rape cases. The results are visually plotted, and the clusters are interpreted as follows:

**Cluster 0 (Green):**

* **Most districts** belong to this group.
* These have **low murder and rape counts**, indicating safer or rural zones.
* Possibly areas with lower population density and limited urban stress.

**Cluster 1 (Orange):**

* **Moderate-crime districts**, suggesting **semi-urban or developing** areas.
* These districts show average reporting levels — may have rising crime awareness and better documentation.

**Cluster 3 (Pink):**

* These districts report **disproportionately high rape cases**, while murder counts are moderate.
* Indicates potential **gender-based violence** hotspots or areas with better reporting of such crimes.
* Needs **policy intervention focused on women’s safety**.

**Cluster 2 (Blue):**

* Characterized by **very high murder counts**.
* Could reflect **conflict-prone districts, gang violence areas**, or challenges in law enforcement.
* Likely to be **urbanized or politically volatile** zones.

**Geospatial Crime Hotspot Map**

Using a shapefile of Indian states (Admin2.shp), a **choropleth map** was created to visualize **total IPC crimes per state**.

**Key Insights from the Map:**

* **Dark red states** such as **Madhya Pradesh, Maharashtra, Andhra Pradesh, and Uttar Pradesh** indicate the **highest volume of IPC crimes**.
* **Moderate crime states** include **Rajasthan, Gujarat, Karnataka**, shaded in orange.
* **Light-colored regions** (e.g., North-Eastern states and Union Territories) suggest **lower crime volumes**, but could also reflect **underreporting or smaller population sizes**.

A map of india with red and orange shades

AI-generated content may be incorrect.

**Graph Elements:**

* **Map of India** divided by states.
* **Color Gradient (White → Red)**:
  + **White/light yellow** = States with **low total IPC crimes**
  + **Dark red/maroon** = States with **very high total IPC crimes**
* **Colorbar (right)**: A scale showing the intensity of crime from low to high, with numbers in the scale reaching over **5 million** IPC crimes for the highest states.

**Key Observations:**

1. **Crime Hotspots**:

* **Madhya Pradesh, Maharashtra, Tamil Nadu, and Andhra Pradesh** are **deep red**, indicating the **highest volume of IPC crimes** in India.
* These are **densely populated** states with large urban centers, possibly contributing to higher crime reporting.

1. **Moderate Crime States**:

* **Gujarat, Rajasthan, Karnataka** and parts of **West Bengal and Odisha** fall in the mid-range with **moderate crime volumes**, shown in orange shades.

1. **Low Crime States**:

* **Northeastern states**, **Jammu & Kashmir**, **Himachal Pradesh**, and **Uttarakhand** show up in **lighter shades**, indicating **fewer reported crimes**.

**Technical Steps**

* The crime data was cleaned and aggregated by STATE/UT.
* State names were standardized using .str.title().str.strip() to ensure a **successful merge with the shapefile**.
* The map was rendered using **GeoPandas** and **Matplotlib** with a red color scale (OrRd) representing crime severity.

**Top 15 Districts by Murder Count**

* The code aggregates and visualizes murder and rape cases across all districts.
* A **bar chart** is plotted showing the top 15 districts with the highest number of murders and their corresponding rape cases.
* **Insight**: This chart helps compare violent crimes side-by-side in high-crime districts. Areas with both high murder and rape counts require immediate policy attention and law enforcement strengthening.

A white background with black text

AI-generated content may be incorrect.

**X-axis:**

* Represents the **district names** (like DELHI UT TOTAL, PATNA, MEERUT, etc.).
* These districts are selected because they have the **highest number of murders**.

**Y-axis:**

* Shows the **total number of cases** (for both MURDER and RAPE).

**Bars:**

* **Blue Bars** represent **MURDER** counts.
* **Orange Bars** represent **RAPE** counts.

**Key Insights:**

1. **‘TOTAL’ and ‘DELHI UT TOTAL’ dominate:**
   * The 'TOTAL' bar is significantly higher than all others — likely an aggregation, not an individual district. It should ideally be removed for clearer district-level analysis.
   * 'DELHI UT TOTAL' is also quite high, indicating that Delhi has **very high murder and rape counts** relative to other regions.
2. **Disparity across districts:**
   * Most other districts (e.g., PATNA, MEERUT, RANCHI, etc.) show **much lower values**.
   * Even among the top 15 by murder, **rape counts vary**. Some districts have murder significantly higher than rape, and in others, they're more balanced.
3. **Geographic implications:**
   * Districts from **Uttar Pradesh**, **Bihar**, and **West Bengal** are prominently featured, indicating recurring violent crime trends in these states.

**State with the Lowest IPC Crime Rate**

* The data is grouped by state, and **total IPC crimes are summed**.
* The state with the **lowest total IPC crime** is identified using .idxmin().

Output:

State with the lowest IPC crime rate: Daman & Diu

Total IPC Crimes: 554

**Insight**: This might indicate a safer region or possibly **underreporting** due to population size or administrative limitations.

**Most Common Crime in Each District**

* For each district, the code identifies the **crime category with the highest number of cases**.
* This helps us create a new column Most\_Common\_Crime to highlight the dominant crime.

**Insight**: Knowing the most frequent crime per district allows targeted interventions (e.g., focusing on theft prevention or women’s safety in relevant zones).

**K-Means Clustering of Districts**

* The data is clustered into 4 groups using **MURDER, RAPE, THEFT, DOWRY DEATHS, RIOTS, TOTAL IPC CRIMES**.
* The plot shows **Murder vs Theft** with color-coded clusters.  
  A graph with colored dots

  AI-generated content may be incorrect.

**Insight**: Districts are grouped by severity and pattern of crimes. For instance:

* Cluster 0: Low-crime rural regions.
* Cluster 2: High-theft, moderate-violence regions.
* Cluster 3: High rape, moderate murder areas.
* Cluster 1: High violence overall – urban stress centers.

**Predicting Total IPC Crimes (Regression)**

* A **Linear Regression** model is built using key features: MURDER, RAPE, THEFT, DOWRY DEATHS.
* The model predicts TOTAL IPC CRIMES with:
  + R² Score – accuracy of the model -0.87
  + RMSE – average prediction error-6987.33

**Insight**: The model helps understand how much **violent and property crimes** contribute to overall IPC trends. Higher R² indicates strong explanatory power.

**Classifying High vs Low Crime Districts**

* A **Random Forest Classifier** labels districts as:
  + 1 for high crime (above median)
  + 0 for low crime
* Accuracy scores for both training and test sets are printed.

**Insight**: This classification helps in **automated tagging of hotspots** and can be used in real-time dashboards or alerts.

**Crime Risk Index**

* A **Crime Risk Index** is created by normalizing and summing key crime types: MURDER, RAPE, THEFT, DOWRY DEATHS, RIOTS.
* The top 10 districts by risk index are shown.

**Insight**: This is a **composite score** indicating which districts are most dangerous based on a combination of violent and social crimes.

**Crimes Against Women**

* Total number of crimes classified as “against women” are calculated (e.g., rape, dowry deaths, abduction).
* Their proportion is shown relative to total IPC crimes.

Example Output:

Crimes against women constitute 8.91% of total IPC crimes.

**Insight**: High percentages indicate a **societal issue** and need for gender-sensitive policing and education.

**State with the Highest Dowry Deaths**

* Dowry deaths are summed across states and the **highest one** is displayed.

Example Output:

State with the highest dowry deaths: Uttar Pradesh – 57256

**Insight**: Targeted **social reform campaigns** are necessary in these states to eliminate dowry-related violence.

**Seasonal Trends in Crimes (Simulated)**

* A **fake monthly trend** is generated using a random Dirichlet distribution to simulate seasonality in IPC crimes.
* A line chart shows estimated crime volume per month.

**Insight**: While this is **synthetic data**, real trends (if available) could identify peak crime seasons (e.g., holidays, summer months).  
A graph with orange lines

AI-generated content may be incorrect.

* **X-axis (Months):** Lists all months from **January to December**.
* **Y-axis (Estimated Crimes):** Displays crime estimates using simulated values on a scale of **millions**.
* **Plot Type:** A **line plot** with orange markers connecting each month's estimated crime value.

**Interpretation of the Pattern:**

* **March** shows a **significant spike** in estimated crime volume — the highest among all months.
* **May** and **September** appear to be **lowest** in this simulated trend.
* **July** and **November** also display **elevated crime levels**, though less than March.
* The overall pattern is **highly variable**, with no clear seasonal trend — reflecting the **random nature of the simulation**.

**Urban vs Rural Crime Rate (Proxy)**

* Districts are split into “Urban-like” and “Rural-like” based on whether their total IPC crimes exceed the 75th percentile.
* Average crime per category is then compared between the two.

Example Output:

Avg IPC crimes in Urban-like districts: 17194.57

Avg IPC crimes in Rural-like districts: 1619.57

**Insight**: Urban-like regions experience significantly more crime across all categories — partly due to population density and better reporting infrastructure.

**Time-Series Forecasting of IPC Crimes**

* Using **Simple Exponential Smoothing**, total IPC crimes are forecasted for the next 5 years (from 2015 onward).

Example Output:

Year 2015: Predicted Crimes = 5715752

Year 2016: Predicted Crimes = 5715752

Year 2017: Predicted Crimes = 5715752

Year 2018: Predicted Crimes = 5715752

Year 2019: Predicted Crimes = 5715752

**Insight**: Time series analysis helps in **long-term resource allocation and law enforcement planning**.