```
'''Exploratory Data Analysis (EDA)
Overview
--------
Exploratory Data Analysis (EDA) is a crucial step to understand the Crimes in India dataset, identify patterns, an
This section focuses on visualizing and summarizing the data to prepare it for further analysis.

Objectives
----------
Understand the distribution of different crime types.
Explore trends over time (2001–2013).
Analyze variations in crime rates across states and districts.
Identify missing or inconsistent data.'''



import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
crime = pd.read_csv(r"/content/crime.csv")
crime
```

	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	..
0	Andhra Pradesh	ADILABAD	2013	96	72	13	61	0	61	65	.
1	Andhra Pradesh	ANANTAPUR	2013	156	149	3	28	0	28	110	.
2	Andhra Pradesh	CHITTOOR	2013	72	61	2	31	0	31	52	.
3	Andhra Pradesh	CUDDAPAH	2013	93	107	7	19	0	19	84	.
4	Andhra Pradesh	CYBERABAD	2013	162	123	16	138	0	138	192	.
...	...	...	...	...	...	...	...	...	...	...	.
9835	DELHI UT	WEST	2001	70	51	12	45	0	45	151	.
9836	LAKSHADWEEP	LAKSHADWEEP	2001	1	0	0	0	0	0	0	.
9837	LAKSHADWEEP	TOTAL	2001	1	0	0	0	0	0	0	.
9838	PUDUCHERRY	PONDICHERRY	2001	25	32	1	9	0	9	4	.
9839	PUDUCHERRY	TOTAL	2001	25	32	1	9	0	9	4	.

9840 rows × 33 columns

```
crime.tail()
```

	STATE/UT	DISTRICT	YEAR	MURDER	ATTEMPT TO MURDER	CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	RAPE	CUSTODIAL RAPE	OTHER RAPE	KIDNAPPING & ABDUCTION	..
9835	DELHI UT	WEST	2001	70	51	12	45	0	45	151	.
9836	LAKSHADWEEP	LAKSHADWEEP	2001	1	0	0	0	0	0	0	.
9837	LAKSHADWEEP	TOTAL	2001	1	0	0	0	0	0	0	.
9838	PUDUCHERRY	PONDICHERRY	2001	25	32	1	9	0	9	4	.
9839	PUDUCHERRY	TOTAL	2001	25	32	1	9	0	9	4	.

5 rows × 33 columns

```
crime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9840 entries, 0 to 9839
Data columns (total 33 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   STATE/UT                                        9840 non-null   object
 1   DISTRICT                                        9840 non-null   object
 2   YEAR                                            9840 non-null   int64
 3   MURDER                                          9840 non-null   int64
 4   ATTEMPT TO MURDER                               9840 non-null   int64
 5   CULPABLE HOMICIDE NOT AMOUNTING TO MURDER       9840 non-null   int64
 6   RAPE                                            9840 non-null   int64
 7   CUSTODIAL RAPE                                  9840 non-null   int64
 8   OTHER RAPE                                      9840 non-null   int64
 9   KIDNAPPING & ABDUCTION                          9840 non-null   int64
 10  KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS     9840 non-null   int64
 11  KIDNAPPING AND ABDUCTION OF OTHERS              9840 non-null   int64
 12  DACOITY                                         9840 non-null   int64
 13  PREPARATION AND ASSEMBLY FOR DACOITY            9840 non-null   int64
 14  ROBBERY                                         9840 non-null   int64
 15  BURGLARY                                        9840 non-null   int64
 16  THEFT                                           9840 non-null   int64
 17  AUTO THEFT                                      9840 non-null   int64
 18  OTHER THEFT                                     9840 non-null   int64
 19  RIOTS                                           9840 non-null   int64
 20  CRIMINAL BREACH OF TRUST                        9840 non-null   int64
 21  CHEATING                                        9840 non-null   int64
 22  COUNTERFIETING                                  9840 non-null   int64
 23  ARSON                                           9840 non-null   int64
 24  HURT/GREVIOUS HURT                              9840 non-null   int64
 25  DOWRY DEATHS                                    9840 non-null   int64
 26  ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY  9840 non-null   int64
 27  INSULT TO MODESTY OF WOMEN                      9840 non-null   int64
 28  CRUELTY BY HUSBAND OR HIS RELATIVES             9840 non-null   int64
 29  IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES     9840 non-null   int64
 30  CAUSING DEATH BY NEGLIGENCE                     9840 non-null   int64
 31  OTHER IPC CRIMES                                9840 non-null   int64
 32  TOTAL IPC CRIMES                                9840 non-null   int64
dtypes: int64(31), object(2)
memory usage: 2.5+ MB
```

```
crime.isna().sum()
```

	0
STATE/UT	0
DISTRICT	0
YEAR	0
MURDER	0
ATTEMPT TO MURDER	0
CULPABLE HOMICIDE NOT AMOUNTING TO MURDER	0
RAPE	0
CUSTODIAL RAPE	0
OTHER RAPE	0
KIDNAPPING & ABDUCTION	0
KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS	0
KIDNAPPING AND ABDUCTION OF OTHERS	0
DACOITY	0
PREPARATION AND ASSEMBLY FOR DACOITY	0
ROBBERY	0
BURGLARY	0
THEFT	0
AUTO THEFT	0
OTHER THEFT	0
RIOTS	0
CRIMINAL BREACH OF TRUST	0
CHEATING	0
COUNTERFIETING	0
ARSON	0
HURT/GREVIOUS HURT	0
DOWRY DEATHS	0
ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY	0
INSULT TO MODESTY OF WOMEN	0
CRUELTY BY HUSBAND OR HIS RELATIVES	0
IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES	0
CAUSING DEATH BY NEGLIGENCE	0
OTHER IPC CRIMES	0
TOTAL IPC CRIMES	0

```
crime.columns
```

```
Index(['STATE/UT', 'DISTRICT', 'YEAR', 'MURDER', 'ATTEMPT TO MURDER',
       'CULPABLE HOMICIDE NOT AMOUNTING TO MURDER', 'RAPE', 'CUSTODIAL RAPE',
       'OTHER RAPE', 'KIDNAPPING & ABDUCTION',
       'KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS',
       'KIDNAPPING AND ABDUCTION OF OTHERS', 'DACOITY',
       'PREPARATION AND ASSEMBLY FOR DACOITY', 'ROBBERY', 'BURGLARY', 'THEFT',
       'AUTO THEFT', 'OTHER THEFT', 'RIOTS', 'CRIMINAL BREACH OF TRUST',
       'CHEATING', 'COUNTERFIETING', 'ARSON', 'HURT/GREVIOUS HURT',
```

```
       'DOWRY DEATHS', 'ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY',
       'INSULT TO MODESTY OF WOMEN', 'CRUELTY BY HUSBAND OR HIS RELATIVES',
       'IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES',
       'CAUSING DEATH BY NEGLIGENCE', 'OTHER IPC CRIMES', 'TOTAL IPC CRIMES'],
      dtype='object')
```

```
'''Data Cleaning
   -------------
1. Standardizing the 'STATE/UT' Column:
The 'STATE/UT' column in the dataset contains state names in both uppercase and lowercase formats.
To maintain consistency across the dataset, all entries in this column are converted to uppercase.
This ensures that state names are standardized and prevents any inconsistencies when analyzing the data.

2. Removing Invalid District Entries:
The 'district' column includes some entries like "TOTAL" and "ZZ TOTAL",
which are not valid district names but are being considered as such in the dataset.
These entries represent aggregates and not specific districts,
so they are removed from the dataset to ensure that only valid district data is retained.
Removing these entries helps in maintaining the integrity of the analysis.'''

'''The cleaned csv file has been saved , rename the crimes_cleaned2.csv
to crimes_cleaned.csv for running other codes.'''


import pandas as pd

# Load the CSV file
url = "crime.csv"  # Replace with your dataset path or URL
data = pd.read_csv(url)

# Convert the 'STATE/UT' column to uppercase
data['STATE/UT'] = data['STATE/UT'].str.upper()



# Remove rows where 'DISTRICT' column has entries 'TOTAL' or 'ZZ TOTAL'
if 'DISTRICT' in data.columns:
    data = data[~data['DISTRICT'].str.upper().isin(['TOTAL', 'ZZ TOTAL'])]
else:
    print("Column 'DISTRICT' not found. Check for spelling or formatting issues.")
# Save the modified data back to a CSV file
output_path = "crimes_cleaned2.csv"  # Replace with your desired output file path
data.to_csv(output_path, index=False)

print("Data processing complete. Modified file saved to:", output_path)
```

```
Data processing complete. Modified file saved to: crimes_cleaned2.csv
```

```
# Importing required libraries
'''STARTING ANALYSIS'''

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
file_path = 'crimes_cleaned.csv'
if os.path.exists(file_path):
    data = pd.read_csv(file_path)
    print("Dataset loaded successfully.")
else:
    raise FileNotFoundError(f"File not found: {file_path}")
```

```
Dataset loaded successfully.
```

```
print("Data Information:")
data.info()
```

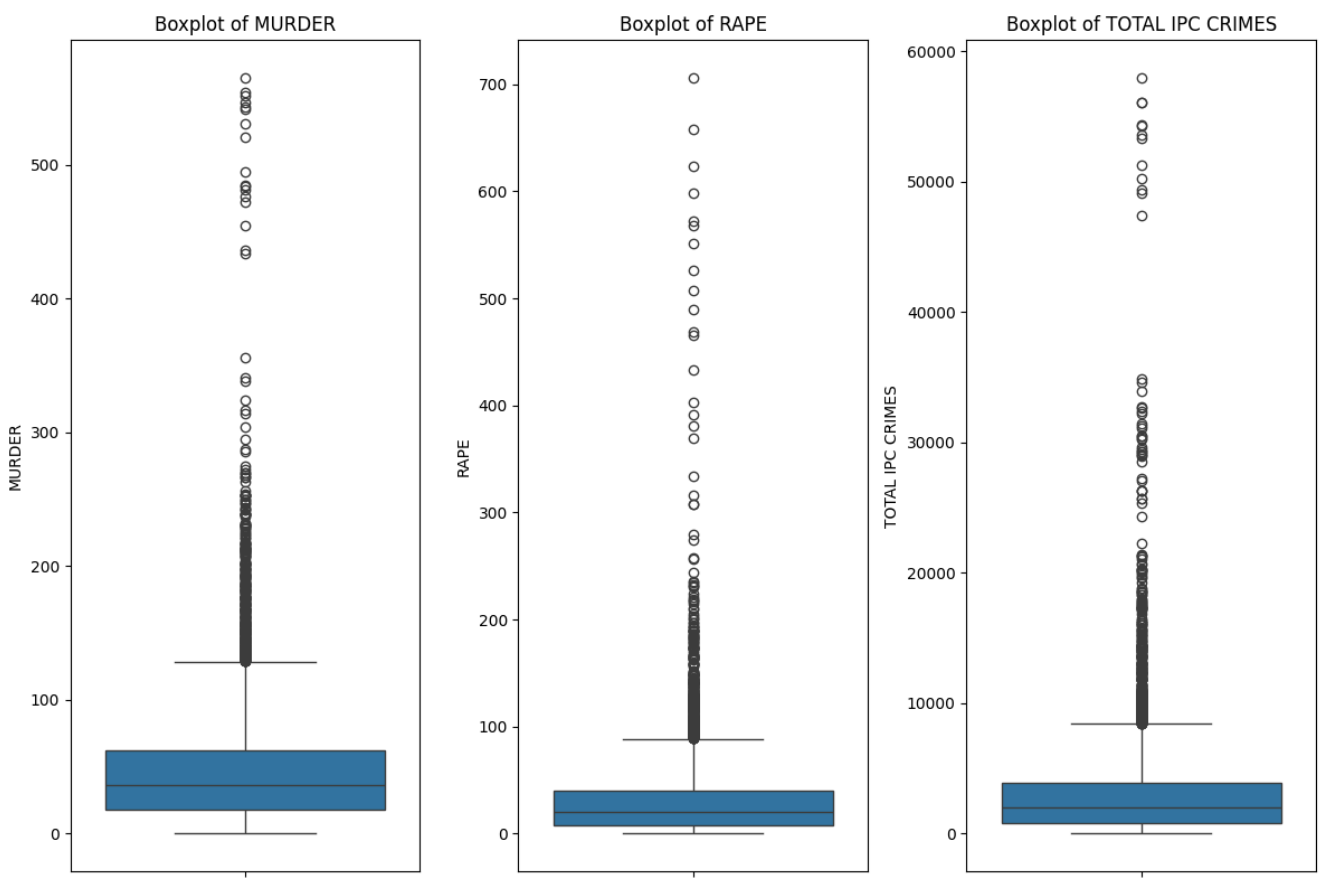
```
Data Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9397 entries, 0 to 9396
Data columns (total 33 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   STATE/UT                                              9397 non-null   object
 1   DISTRICT                                              9397 non-null   object
 2   YEAR                                                  9397 non-null   int64
 3   MURDER                                                9397 non-null   int64
 4   ATTEMPT TO MURDER                                     9397 non-null   int64
 5   CULPABLE HOMICIDE NOT AMOUNTING TO MURDER             9397 non-null   int64
 6   RAPE                                                  9397 non-null   int64
 7   CUSTODIAL RAPE                                        9397 non-null   int64
 8   OTHER RAPE                                            9397 non-null   int64
 9   KIDNAPPING & ABDUCTION                                9397 non-null   int64
 10  KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS           9397 non-null   int64
 11  KIDNAPPING AND ABDUCTION OF OTHERS                    9397 non-null   int64
 12  DACOITY                                               9397 non-null   int64
 13  PREPARATION AND ASSEMBLY FOR DACOITY                  9397 non-null   int64
 14  ROBBERY                                               9397 non-null   int64
 15  BURGLARY                                              9397 non-null   int64
 16  THEFT                                                 9397 non-null   int64
 17  AUTO THEFT                                            9397 non-null   int64
 18  OTHER THEFT                                           9397 non-null   int64
 19  RIOTS                                                 9397 non-null   int64
 20  CRIMINAL BREACH OF TRUST                              9397 non-null   int64
 21  CHEATING                                              9397 non-null   int64
 22  COUNTERFIETING                                        9397 non-null   int64
 23  ARSON                                                 9397 non-null   int64
 24  HURT/GREVIOUS HURT                                    9397 non-null   int64
 25  DOWRY DEATHS                                          9397 non-null   int64
 26  ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY   9397 non-null   int64
 27  INSULT TO MODESTY OF WOMEN                            9397 non-null   int64
 28  CRUELTY BY HUSBAND OR HIS RELATIVES                   9397 non-null   int64
 29  IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES           9397 non-null   int64
 30  CAUSING DEATH BY NEGLIGENCE                           9397 non-null   int64
 31  OTHER IPC CRIMES                                      9397 non-null   int64
 32  TOTAL IPC CRIMES                                      9397 non-null   int64
dtypes: int64(31), object(2)
memory usage: 2.4+ MB
```

```
# Descriptive statistics
print("\nDescriptive Statistics:")
descriptive_stats = data.describe()
print(descriptive_stats)
```

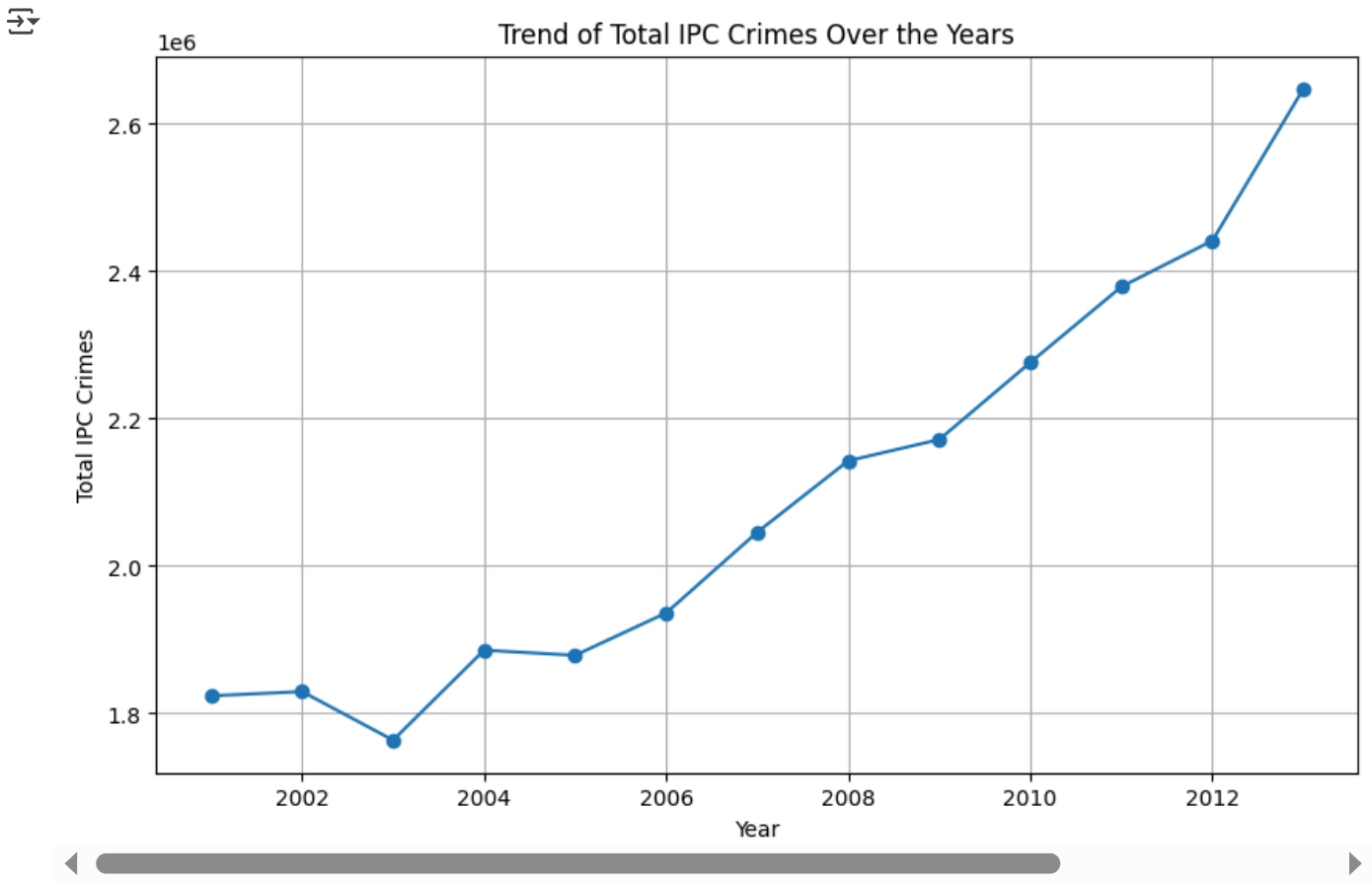
```
Descriptive Statistics:
              YEAR       MURDER  ATTEMPT TO MURDER  \
count  9397.000000  9397.000000        9397.000000
mean   2007.168884    47.030861          41.786847
std       3.755781    45.666528          53.614888
min    2001.000000     0.000000           0.000000
25%    2004.000000    18.000000          10.000000
50%    2007.000000    36.000000          27.000000
75%    2010.000000    62.000000          54.000000
max    2013.000000   565.000000         741.000000

       CULPABLE HOMICIDE NOT AMOUNTING TO MURDER         RAPE  CUSTODIAL RAPE  \
count                                9397.000000  9397.000000     9397.000000
mean                                    5.201341    29.718846        0.002873
std                                    10.039063    37.168683        0.076455
min                                     0.000000     0.000000        0.000000
25%                                     0.000000     8.000000        0.000000
50%                                     2.000000    20.000000        0.000000
75%                                     6.000000    40.000000        0.000000
max                                   241.000000   706.000000        5.000000

        OTHER RAPE  KIDNAPPING & ABDUCTION  \
count  9397.000000             9397.000000
mean     29.715973               47.611046
std      37.165828              102.712809
min       0.000000                0.000000
```

```
25%       8.000000              10.000000
50%      20.000000              25.000000
75%      40.000000              56.000000
max     706.000000            3970.000000

       KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS  \
count                                  9397.000000
mean                                     35.270618
std                                      67.992258
min                                       0.000000
25%                                       6.000000
50%                                      18.000000
75%                                      42.000000
max                                    2160.000000

       KIDNAPPING AND ABDUCTION OF OTHERS  ...        ARSON  \
count                         9397.000000  ...  9397.000000
mean                            12.340428  ...    13.155794
std                             40.610955  ...    29.478553
min                              0.000000  ...     0.000000
25%                              1.000000  ...     2.000000
50%                              5.000000  ...     8.000000
75%                             12.000000  ...    18.000000
max                           1810.000000  ...  2350.000000

       HURT/GREVIOUS HURT  DOWRY DEATHS  \
count         9397.000000   9397.000000
mean           396.802703     10.733958
std            567.796741     14.833491
min              0.000000      0.000000
25%             40.000000      1.000000
```

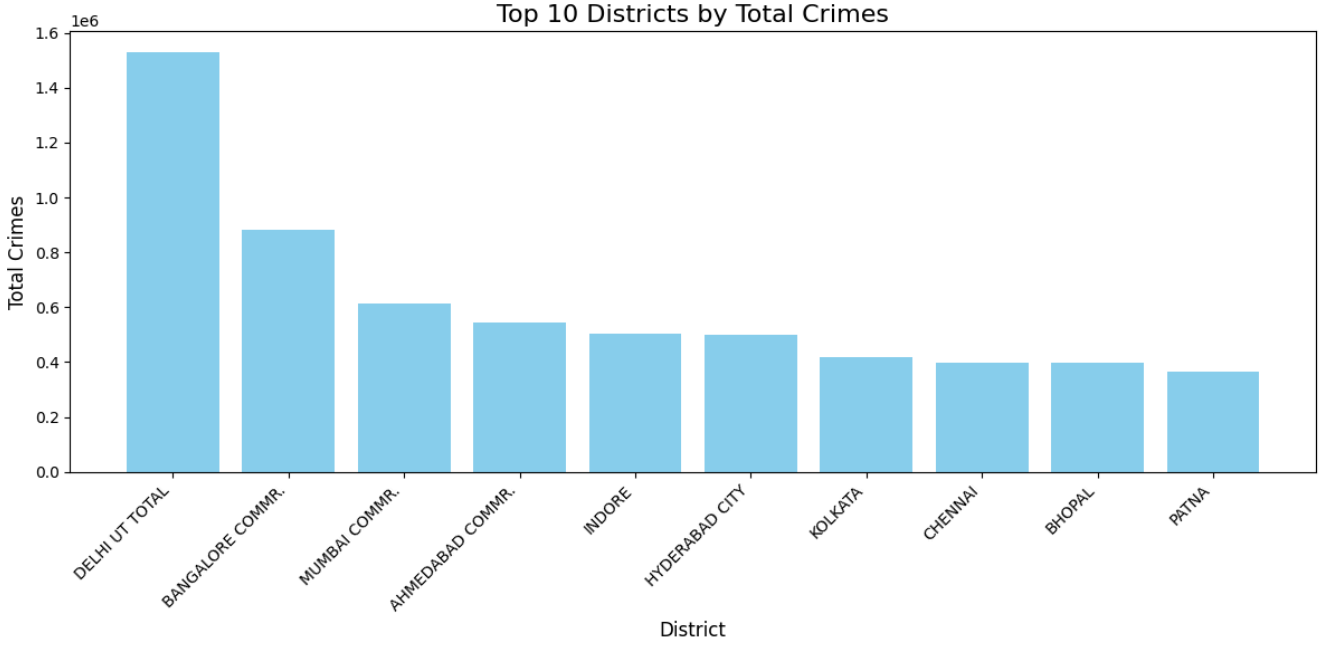
```
'''Data Visualization
Overview
--------
Visualizing the data provides valuable insights into crime trends, distributions, and proportions.
In this section, we use box plots, pie charts, and trend lines to better understand the patterns and dynamics of c


# Visualizing outliers using boxplots for selected columns
columns_of_interest = ['MURDER', 'RAPE', 'TOTAL IPC CRIMES']
plt.figure(figsize=(12, 8))
for i, col in enumerate(columns_of_interest, 1):
    plt.subplot(1, 3, i)
    sns.boxplot(y=data[col])
    plt.title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```



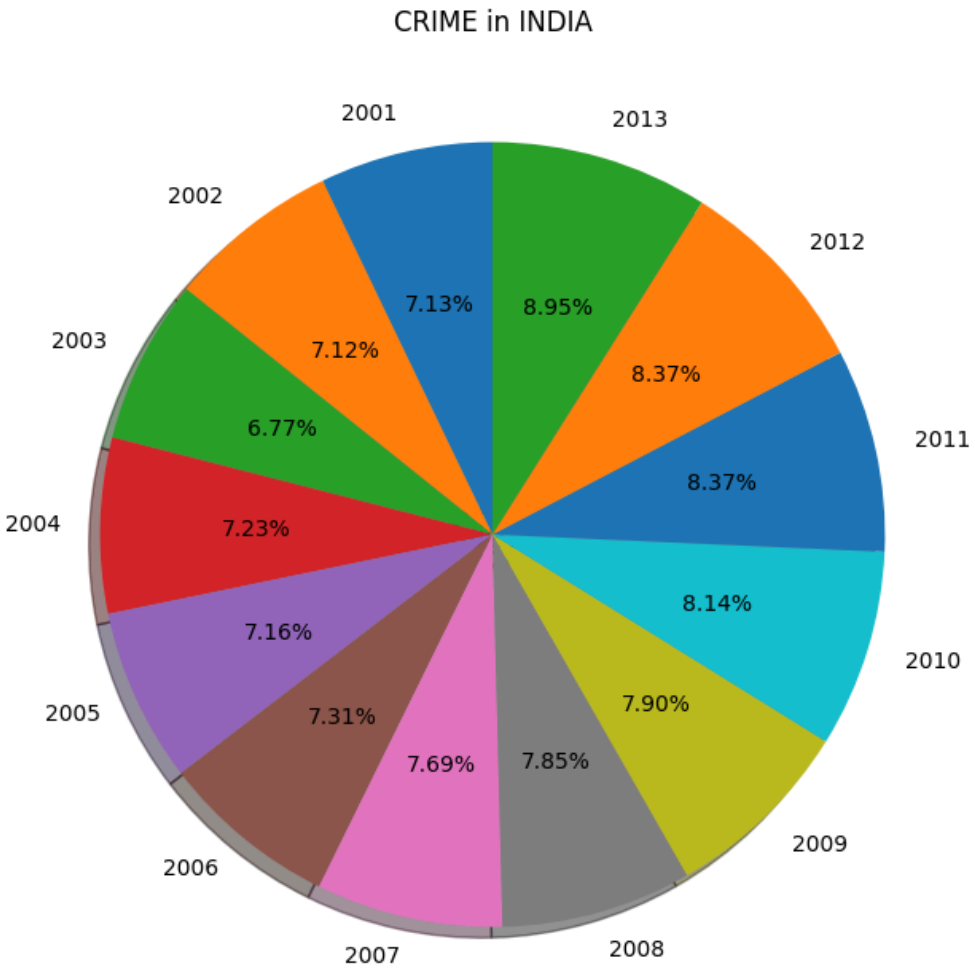

```
# 2. Yearly Trend of Total IPC Crimes
plt.figure(figsize=(10, 6))
data.groupby('YEAR')['TOTAL IPC CRIMES'].sum().plot(marker='o')
plt.title('Trend of Total IPC Crimes Over the Years')
plt.ylabel('Total IPC Crimes')
plt.xlabel('Year')
plt.grid(True)
plt.show()
```



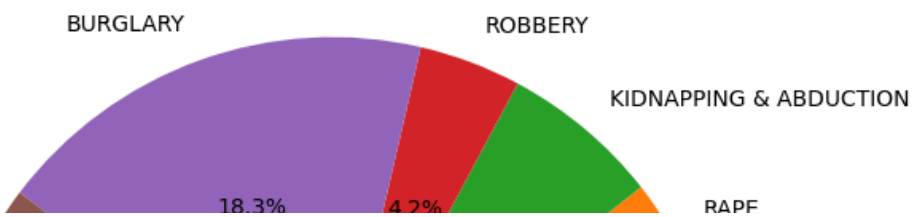

```
crime_data = data

# Function to plot top 10 districts by total crimes
def top_dus_crimes(crime_data):
    # Drop unnecessary columns for calculation
    remcol = ['STATE/UT', 'DISTRICT', 'YEAR']
    crime_variables = [col for col in crime_data.columns if col not in remcol]

    # Calculate total crimes per district by summing all crime variables
    crime_data['TOTAL_CRIMES'] = crime_data[crime_variables].sum(axis=1)

    # Filter out districts named 'TOTAL' and 'ZZ TOTAL'
    crime_data_filtered = crime_data[~crime_data['DISTRICT'].isin(['TOTAL', 'ZZ TOTAL'])]

    # Group by district and sum the total crimes across years
    district_crime_totals = crime_data_filtered.groupby('DISTRICT')['TOTAL_CRIMES'].sum().reset_index()

    # Sort the data to get the top 10 districts with the highest total crimes
    top_10_districts = district_crime_totals.sort_values(by='TOTAL_CRIMES', ascending=False).head(10)

    # Create bar plot for the top 10 districts
    plt.figure(figsize=(12, 6))
    plt.bar(top_10_districts['DISTRICT'], top_10_districts['TOTAL_CRIMES'], color='skyblue')

    # Add titles and labels
    plt.title('Top 10 Districts by Total Crimes', fontsize=16)
    plt.xlabel('District', fontsize=12)
    plt.ylabel('Total Crimes', fontsize=12)
    plt.xticks(rotation=45, ha='right')
    plt.tight_layout()
    plt.show()

# Call the function to plot the top 10 districts by total crimes
top_dus_crimes(crime_data)
```




```
import matplotlib.pyplot as plt


state_mean = data.groupby('YEAR')['TOTAL IPC CRIMES'].mean()

# Create the pie chart
plt.figure(figsize=(15, 8))
plt.pie(state_mean, labels=state_mean.index, startangle=90, shadow=True,
        textprops={'fontsize': 10, 'color': 'black'}, autopct='%0.2f%%')

plt.title('CRIME in INDIA')
plt.show()
```




```
crime_totals = data[['MURDER', 'RAPE', 'KIDNAPPING & ABDUCTION', 'ROBBERY', 'BURGLARY', 'THEFT']].sum()

# Create a pie chart
plt.figure(figsize=(10, 8))
plt.pie(crime_totals.values, labels=crime_totals.index, autopct='%1.1f%%')
plt.title('Proportion of Crimes by Type')
plt.show()
```




```
state_totals = data.groupby('STATE/UT')['TOTAL IPC CRIMES'].sum()

# Create a bar chart
plt.figure(figsize=(15, 10))
plt.bar(state_totals.index, state_totals.values)
plt.xticks(rotation=90)
plt.xlabel('State/Union Territory')
plt.ylabel('Total IPC Crimes')
```