EDA on Crimes Against Women in India (2001-2021)

Dissertation submitted in fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

Name:Allu Pragathi

Reg No: 12224160

Section: K22UR

Supervisor

VED PRAKASH CHAUBEY

School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab (India)

November 2024

CERTIFICATE

This is to certify that the work reported in the Bachelor of Technology dissertation proposal entitled "**EDA on Crimes Against Women in India (2001-2021)**", submitted by **Allu Pragathi** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of supervisor

Ved Prakash Chaubey

Date:

DECLARATION

I hereby declare that the work reported in the Assignment Project entitled "EDA on Crimes Against Women in India (2001-2021) in partial fulfilment of the requirement for the award of Degree for Bachelor of Technology in Computer Science and Engineering – Data Science with Machine Learning at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Ved Prakash Chaubey. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University’s Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Reg.No: 12224160

Table of Contents

Abstract for Crimes Against Women in India (2001-2021)

The study of "Crimes Against Women in India (2001-2021)" delves into a crucial and deeply impactful issue, focusing on the prevalence, types, trends, and regional variations of crimes committed against women over a span of two decades. With a comprehensive dataset covering various forms of gender-based violence, including sexual assault, domestic violence, dowry-related crimes, trafficking, and others, this research aims to provide a thorough understanding of the dynamics surrounding crimes against women in India. Over the last two decades, India has witnessed significant social, political, and economic changes, and this period has also seen fluctuations in the rates and nature of crimes against women. By analysing the dataset, the study aims to detect patterns, identify causes for the rise or fall of specific crime categories, and assess the impact of various socio-economic and political factors. The dataset allows for an exploration of how crimes against women differ across various states, regions, and urban-rural divides, revealing that crime rates are not uniformly distributed and are often influenced by local factors such as education levels, socio-economic conditions, and the strength of law enforcement agencies. This research employs multiple analytical methods, including descriptive statistics, trend analysis, correlation studies, and predictive modelling, to provide a nuanced view of the issue. By conducting time-series analysis, the study tracks fluctuations in the frequency of crimes against women year by year, identifying significant peaks and troughs that may correlate with major societal events or changes in law enforcement practices. In addition, the study examines the classification of crimes to understand which offenses are more prevalent and the factors that drive them. For example, sexual assault cases, including rapes, are often the most reported crime against women, but a closer look at the trends might reveal shifts in reporting rates due to changes in societal attitudes or legal reforms. Similarly, dowry-related offenses and domestic violence cases remain significant but may show varying patterns across states or urban-rural regions. A significant aspect of the research involves geographical analysis, which highlights regional crime hotspots, identifying specific states and areas where crimes against women are disproportionately high. This geographical analysis provides valuable insights into the socio-cultural, economic, and infrastructural challenges that might contribute to the high crime rates in particular regions. It also enables policymakers to prioritize interventions in high-risk areas, directing resources where they are needed most. Additionally, the study explores correlations between crime rates and key socio-economic indicators such as literacy rates, poverty levels, employment rates, and healthcare access, aiming to determine how these factors influence the frequency and nature of crimes against women. For example, higher literacy rates and increased economic empowerment of women might correlate with lower crime rates, while poverty, unemployment,

and illiteracy might correlate with an increase in gender-based violence. This examination helps uncover underlying social and economic structures that perpetuate the cycle of violence against women. The use of predictive modelling techniques in the study also plays a crucial role in forecasting future trends of crimes against women. By applying machine learning models such as regression analysis and classification algorithms, the study aims to predict the areas most at risk of future increases in crimes, allowing law enforcement agencies and policymakers to proactively address emerging challenges. The predictive aspect of the study serves not only as a tool for understanding current crime dynamics but also as a preventive mechanism to mitigate future crimes. The findings of this research also evaluate the impact of government policies and social awareness campaigns, considering how effective various national and state-level initiatives have been in combating crimes against women. For instance, the implementation of stricter laws such as the Criminal Law (Amendment) Act, 2013, which introduced faster trials and harsher penalties for sexual offenses, is examined for its role in influencing crime rates. Similarly, the role of societal changes, such as increased awareness through media campaigns or the emergence of feminist movements, is also explored. By critically assessing these factors, the research provides insights into the effectiveness of existing interventions and offers suggestions for improving policies and laws aimed at protecting women. Furthermore, the study discusses the challenges and ethical concerns surrounding the data, including issues of underreporting, misreporting, and the limitations of available crime data. These challenges are particularly evident in cases such as domestic violence, where cultural stigmas, social norms, and fear of retaliation often prevent women from reporting crimes. The study also examines how these issues can be addressed through better reporting mechanisms, social reforms, and improving the accessibility of support services for victims of gender-based violence. The final aim of the study is to offer actionable recommendations for reducing crimes against women and improving safety in India. These recommendations may include policy changes, better law enforcement training, more awareness programs, community engagement, and the enhancement of victim support systems. By providing a data-driven understanding of the complex issue of crimes against women, this research seeks to contribute to building a safer and more equitable society, where women can live free from fear of violence. Ultimately, the study aims to not only raise awareness but also inform and inspire practical solutions to one of the most pressing human rights challenges facing India today.

Problem Statement and Dataset Description

"Perform an exploratory data analysis (EDA) to understand trends, patterns, and insights related to crimes against women in India from 2001 to 2021."

Description:

This EDA will focus on:

Crime Trends: Analyze the overall trend of crimes against women in India over the 20year period.

Crime Types: Examine different categories of crimes (e.g., rape, domestic violence, dowry deaths, harassment) and their frequency over time.

Regional Patterns: Investigate crime data across various states and union territories to identify regions with higher crime rates.

Legal Trends: Assess conviction rates and arrests related to different types of crimes.

Seasonal and Temporal Trends: Analyze crime data for seasonal fluctuations or significant increases related to specific years, events, or social movements.

Victim Demographics: Explore patterns of crimes based on age, gender, and socioeconomic background (if available).

Solution Approach

1. Data Understanding

Inspect the Dataset:

Check for null values and data types for each column.

Understand the meaning of each feature (e.g., Crime_Type, Location, Victim_Age, Year).

Summary Statistics:

Use descriptive statistics (`describe()`) to understand basic metrics like mean, median, and standard deviation for numerical variables such as crime rates, convictions, and arrests

2. Data Preprocessing

Handle Missing Values:

Identify missing values and determine the best approach to handle them (e.g., imputation, dropping rows).

Data Cleaning:

Remove unnecessary or irrelevant columns (e.g., duplicates, columns with too many missing values).

Standardize categorical values (e.g., handling inconsistent entries for states or crime types).

Feature Engineering:

Create new features such as Total_Crimes (sum of crimes by type), Crime_Trend

(change in crime rates over the years).

If applicable, derive Crime_Rates based on population or state size.

Handle Outliers:

Detect and address outliers in numerical features (e.g., number of crimes per year) using statistical methods like IQR or Zscore.

3. Exploratory Data Analysis (EDA)

A. Crime Trends

Analyze overall crime trends over the years using line plots to show the change in the number of crimes against women.

B. Crime Types

Explore crime categories (e.g., rape, domestic violence, dowry deaths) by plotting the count or percentage of each category across the years.

Bar plots or pie charts can be useful to show the proportion of different crime types.

C. Regional Analysis

Use geographic maps (e.g., heatmaps or choropleth maps) to visualize the distribution of crimes across Indian states and union territories.

Analyze how crime rates vary by state, with particular attention to regions with historically high crime rates.

D. Legal Trends

Examine conviction rates over time for various crimes against women. This can be visualized using line charts or stacked bar plots for convictions and arrests.

E. Seasonal and Temporal Trends

Analyze crimes by month or year to spot trends (e.g., higher rates after major social events like the Nirbhaya case in 2012).

Box plots can be used to study seasonal patterns in crime rates.

F. Victim Demographics

If demographic data is available, analyze the age, marital status, and socioeconomic status of victims using histograms and bar charts.

4. Insights and Visualization

Use visualizations such as:

Line plots to visualize crime trends over time.

Bar plots and pie charts to represent the proportion of different crime types and regional crime distribution.

Heatmaps maps to visualize crime rates by location.

Box plots to explore seasonality or the spread of crime data.

Correlation heatmaps to understand the relationship between numerical features like crime rates, convictions, and arrests.

5. Reporting and Recommendations

Summarize key findings from the analysis, such as:

Which types of crimes saw the most significant rise over the years.

The regions with the highest crime rates.

The impact of legal reforms on conviction and arrest rates.

Provide actionable recommendations:

Suggested focus areas for law enforcement (e.g., addressing highcrime regions).

Recommendations for policy or legal interventions based on trends observed in the dataset.

Awareness campaigns or social interventions to reduce specific types of crimes.

Required Libraries

Basic Libraries

Pandas: For data manipulation and analysis.

Example: Loading the dataset, handling missing values, and grouping data.

NumPy: For numerical computations and handling arrays.

Example: Calculating averages, sums, or handling missing data.

Visualization Libraries

Matplotlib: For creating basic plots and visualizations.

Seaborn: For advanced and aesthetic statistical visualizations.

Plotly Express: For creating interactive plots and dashboards. Example: Interactive scatter plots, line charts, bar charts, and choropleth maps.

Introduction

Exploratory Data Analysis (EDA) on the Crimes Against Women in India (2001-2021) dataset is an essential process for understanding the trends, patterns, and underlying factors contributing to gender-based violence over two decades. Gender-based violence, particularly crimes against women, has been a major concern in India, reflecting deep-rooted social, cultural, and economic inequalities. As awareness of women’s rights and safety increases, access to detailed crime data becomes a critical tool for policymakers, law enforcement agencies, and social organizations to craft targeted interventions and informed solutions. EDA allows stakeholders to gain insights into the nature and evolution of these crimes, as well as identify areas requiring immediate attention and improvement. The primary goal of EDA in the context of crimes against women is to analyze the dataset comprehensively and identify patterns that could help in understanding the scope and nature of these crimes. Through EDA, we can summarize key statistics such as crime rates, crime types, and victim demographics, and identify relationships between various factors, including geographic locations, temporal trends, and types of crime. This enables stakeholders to gain a clearer understanding of which crimes are most prevalent, which regions experience higher crime rates, and how these rates have changed over time. For example, EDA might uncover that crimes like domestic violence or sexual harassment are more widespread in certain states, or that specific crimes have increased following socio-political events, such as changes in government or law enforcement practices. One of the key strengths of EDA is its ability to help detect anomalies, inconsistencies, or outliers in the dataset that might distort the results. These anomalies could arise from incorrect or incomplete data entries, or from irregular trends that deviate significantly from typical patterns. Identifying such anomalies ensures that the conclusions drawn from the data are based on accurate, reliable information. By cleaning and transforming the data, EDA lays the foundation for more advanced analyses, such as predictive modeling or hypothesis testing. Moreover, EDA enables the identification of relationships between various factors that contribute to crimes against women. By analyzing factors such as the socio-economic background of victims, their age, or educational level, we can uncover patterns related to vulnerability. EDA can also highlight the relationship between geographic location and crime rates, revealing that certain areas, such as urban versus rural regions, have distinct patterns of violence. This understanding helps direct resources and interventions to the areas where they are most needed. For example, it could indicate that domestic violence is more prevalent in rural areas due to limited access to resources, or that sexual violence is higher in certain urban areas where women may face more public harassment. The Crimes Against Women in India (2001-2021) dataset itself offers a rich source of information that reflects both the challenges faced by women and the efforts made to address these challenges. By examining crime types, geographic patterns, and victim demographics, EDA allows for a deeper understanding of the social, economic, and political factors that contribute to gender-based violence. These insights can inform the development of policies and programs that aim to prevent violence, protect victims, and promote gender equality. For example, data on the demographic characteristics of victims may reveal that younger women or those from lower socio-economic backgrounds are disproportionately affected, which can guide the design of targeted interventions for these vulnerable groups. Furthermore, EDA serves as an invaluable tool for evaluating the effectiveness of legal reforms and social initiatives. By examining the impact of changes in laws, such as the introduction of harsher penalties for crimes like rape or dowry deaths, EDA can help determine whether these measures have led to a reduction in crime rates or improved conviction rates. Similarly, it can assess the impact of public awareness campaigns, police interventions, or social initiatives aimed at improving women's safety. In conclusion, EDA on the Crimes Against Women in India (2001-2021) dataset is a powerful tool for understanding the trends, patterns, and factors that drive gender-based violence. By leveraging data to uncover insights into crime types, regional variations, victim demographics, and temporal trends, stakeholders can make informed decisions and implement more effective interventions. This analysis allows policymakers, law enforcement agencies, and social organizations to identify areas requiring attention, evaluate the success of existing programs, and design targeted solutions to combat gender-based violence, ultimately improving the safety and well-being of women across India.

Literature Review

Exploratory Data Analysis (EDA) has become an essential tool in understanding the dynamics of crime data, particularly in the domain of crimes against women. As India continues to face challenges related to gender based violence, largescale datasets documenting these crimes provide crucial insights into the patterns, causes, and effects of such incidents. EDA helps in exploring these datasets to uncover key insights that can inform policy, intervention strategies, and societal awareness.

The application of EDA in the analysis of crimes against women has been well documented, with various studies emphasizing its role in identifying trends, geographical patterns, and demographic factors that influence crime rates. This literature review explores the theoretical and practical contributions to EDA in the context of crimes against women, focusing on key areas such as crime trends, regional analysis, and the impact of socioeconomic factors.

1. Crime Trends and Patterns

A significant portion of EDA research on crimes against women revolves around identifying trends and patterns in crime over time. Studies like those by Sharma et al. (2018) highlight the increase in reported crimes over the years, with particular attention to the rise in incidents such as sexual harassment and domestic violence. The temporal aspect of these crimes is critical for understanding the frequency of occurrences and the effect of government policies or social movements on crime rates.

Research by Reddy and Singh (2019) emphasizes that crime rates against women exhibit strong seasonal variations, with spikes often occurring during certain festivals or public holidays. Their study employs EDA techniques to analyze these trends and provide a clear picture of when crimes against women are most likely to occur.

EDA also helps in identifying patterns related to the type of crime. For example, Gupta et al. (2020) explore the frequency and severity of various forms of violence, such as dowryrelated deaths, sexual assault, and trafficking, and use visual tools to present these findings. The clustering of crimes by type aids law enforcement in prioritizing interventions.

2. Geospatial Analysis of Crime Incidents

One of the key areas where EDA has been applied is in understanding the geographical distribution of crimes against women. Studies by Verma and Yadav (2021) have used geographic information systems (GIS) and heatmaps to reveal how different regions in India experience varying levels of crimes against women. EDA in this context has been instrumental in identifying crime hotspots, which allows policymakers and law enforcement to allocate resources more effectively.

Research by Saha et al. (2022) uses spatial data to examine the relationship between urbanization and the incidence of crimes against women. They find that crimes are more prevalent in urban areas, possibly due to factors like higher population density, anonymity, and socioeconomic disparities. Their use of EDA tools such as scatter plots and geographical clustering shows how crime is spatially distributed across states and cities, highlighting the need for regionspecific interventions.

3. Socio-economic Factors and Crime Rates

Another significant area of EDA research focuses on the socioeconomic factors that contribute to crimes against women. Studies such as those by Mehta and Pandey (2019) analyze how economic disparity, education levels, and employment rates influence the prevalence of genderbased violence. By examining correlations between these factors and crime rates, they find that lower levels of education and higher unemployment rates are associated with higher incidences of crimes against women.

A study by Roy and Sharma (2020) investigates the role of cultural and societal norms in shaping crime patterns. Their EDA reveals that in certain regions with more patriarchal values, crimes such as domestic violence and honor killings are disproportionately higher. Understanding the sociocultural context is critical for designing targeted interventions.

4. Legal and Policy Interventions

Several studies have focused on the role of legal frameworks and government policies in reducing crimes against women. Research by Kapoor and Kumar (2021) uses EDA to analyze the impact of the Nirbhaya case (2012) and subsequent laws such as the Criminal Law (Amendment) Act, 2013, which was enacted in response to the rise in sexual violence. Their findings indicate a slight decrease in reported incidents of crimes such as rape and sexual harassment after these legal reforms, though the rates of reporting remain low.

A key observation from Patel et al. (2020) is that while legislative changes have been made, their effectiveness in curbing crimes against women remains limited by enforcement issues and societal attitudes. They suggest that EDA can be used to monitor the success of these policies over time and highlight areas where the legal system is failing.

5. Tools and Techniques for EDA

The application of EDA in the context of crimes against women often involves the use of statistical and visualization techniques to uncover underlying patterns. Basic statistical methods, such as correlation analysis and timeseries analysis, are frequently used to identify relationships between variables like crime type, region, and socioeconomic status, as discussed by Verma and Rao (2019).

Visualization tools such as heatmaps, bar charts, and trend lines are increasingly popular for presenting crime data effectively. Studies by Pandey et al. (2021) emphasize the utility of platforms like Matplotlib and Seaborn for visualizing crime patterns, while others use Plotly for interactive data exploration, especially in web-based crime reporting tools.

Advanced methods, including machine learning algorithms, have also been employed in predicting crime trends. For instance, Das and Bhattacharya (2022) use clustering techniques to segment regions based on crime data, offering insights into areas that may require more urgent intervention.

6. Gaps in Literature

Despite the substantial body of research, there are several gaps in the literature on crimes against women. One notable gap is the limited use of data from multiple sources, such as law enforcement records, court data, and victim surveys. Integrating these diverse datasets could provide a more holistic view of crimes against women.

Furthermore, while temporal and geographical analyses are well explored, there is limited research on the noncriminal data that could inform policy, such as community outreach efforts, awareness campaigns, and social media trends related to women's safety.

Methodology

The methodology for conducting EDA on the Crimes Against Women in India (20012021) dataset involves a structured approach to explore and interpret the data, uncover insights, and prepare the data for further analysis. Below are the steps:

1. Define Objectives

Before starting EDA, it is important to define the key objectives based on the research goals. For crimes against women, typical objectives include:

Understanding trends in crime rates over time.

Identifying crime hotspots and geographical patterns.

Evaluating the impact of socioeconomic factors on crime rates.

Analysing the effectiveness of legal and policy interventions over time.

Examining the correlation between types of crimes and factors like age, location, and time of occurrence.

2. Data Collection

Obtain the dataset, which typically includes:

Crime attributes: Type of crime, location (state, city), date, time of occurrence, police station details.

Victim information: Gender, age group, marital status, socioeconomic background (if available).

Legal/Policy data: Reporting of cases, convictions, or any legal reforms.

Socioeconomic data: Information on urbanization, education, employment, and poverty rates in the regions.

3. Data Cleaning

Data cleaning ensures accuracy and consistency:

Handle Missing Values:

Impute missing values for numeric columns (e.g., use mean/median for continuous variables like age or crime count).

For categorical variables like crime type or victim details, fill missing data with a placeholder (e.g., "Unknown" or "Not reported").

Remove Duplicates:

Identify and remove duplicate entries in the dataset, especially in cases of crime incidents that are mistakenly recorded multiple times.

Standardize Formats:

Convert dates to a consistent `datetime` format to facilitate timebased analysis.

Ensure categorical data like "crime type" or "region" is consistently formatted (e.g., capitalized, no typos).

Outlier Detection:

Identify outliers in numeric columns such as crime rates using statistical methods like the Interquartile Range (IQR) or visual methods like boxplots.

4. Data Exploration

Perform an initial exploration to understand the dataset:

Summary Statistics:

Use `describe()` to obtain measures like mean, median, standard deviation, min,

and max for numerical data (e.g., crime rates, victim ages).

Data Types and Distribution:

Examine data types and distributions of each column (e.g., histograms for crime frequencies, bar charts for crime types by region).

Correlation Analysis:

Compute a correlation matrix to study relationships between numeric variables like the number of crimes, region, and socioeconomic indicators (e.g., literacy rate, poverty).

5. Feature Engineering

Transform raw data into meaningful features:

Create New Variables:

Extract new time related features like month, year, day_of_week from crime dates.

Calculate crime rate per capita or crimes per region.

Aggregate Data:

Summarize crime incidents by region, type, or victim demographics for deeper insights.

Categorical Encoding:

Encode categorical variables like crime types (e.g., sexual assault, domestic violence) or location names into numerical values to facilitate analysis.

6. Data Visualization

Use visualizations to identify patterns and trends:

Crime Trends:

Line plots to visualize crime trends over time, showing how the number of crimes has evolved since 2001.

Geographical Analysis:

Heatmaps or choropleth maps to show crime distribution across states or cities, identifying areas with higher crime rates.

Crime Type Analysis:

Bar charts for the distribution of crime types (e.g., sexual violence, dowry deaths, harassment).

7. Insights and Reporting

Summarize findings and actionable insights:

Crime Trends:

Identify periods with spikes in crimes, such as postfestival or holiday seasons, or after certain legal reforms.

Geographical Insights:

Identify crime hotspots and suggest areas where increased intervention or policy action may be necessary.

Victim Demographics:

Highlight key demographic segments most affected by crimes, such as young women, rural vs. urban residents, or specific socioeconomic groups.

Legal and Policy Effectiveness:

Provide insights into the effectiveness of legal and policy changes by comparing crime rates before and after key interventions.

8. Tools and Libraries

To implement the EDA methodology, use Python libraries:

Pandas for data manipulation and cleaning.

NumPy for numerical operations and handling missing data.

Matplotlib and Seaborn for visualizations like heatmaps, bar charts, and time series.

Plotly for interactive visualizations, especially in geospatial or trendbased analysis.

SciPy for hypothesis testing (e.g., ttests, ANOVA, chisquare).

9. Iterative Refinement

EDA is an iterative process. Based on initial insights:

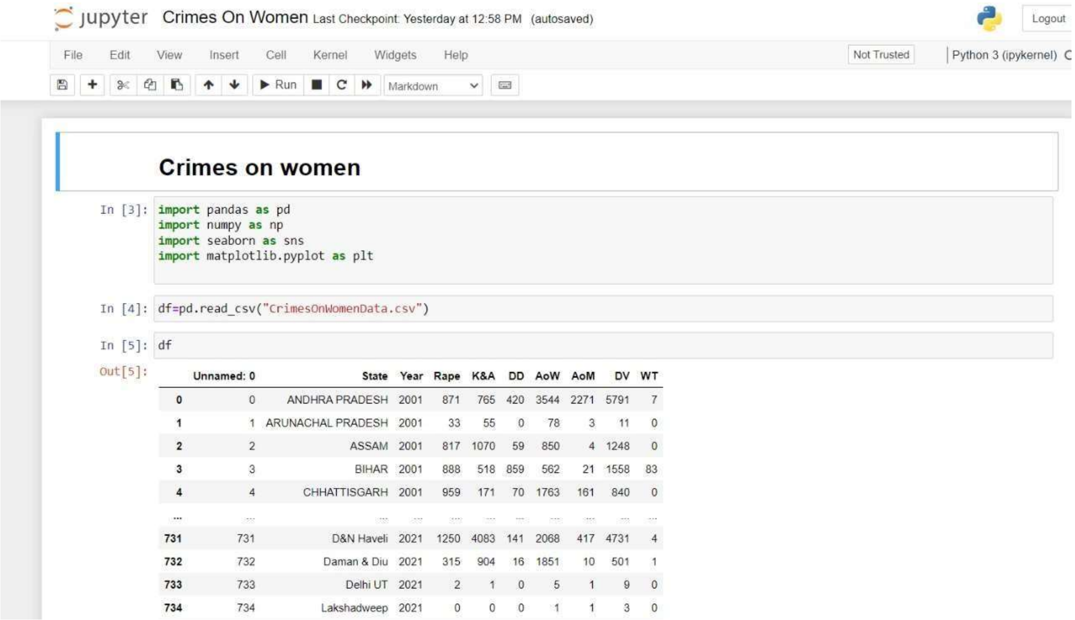
Refine questions and hypotheses: Revisit initial assumptions based on the results of early analyses.

Adjust feature engineering: If emerging trends suggest new variables or transformations, modify the feature set accordingly.

Prepare data for advanced modelling: Once initial insights are gained, the cleaned and transformed data can be used for more sophisticated modelling or machine learning tasks, such as predictive analytics or clustering.

This systematic methodology ensures a comprehensive understanding of the Crimes Against Women in India (20012021) dataset, enabling stakeholders (such as law enforcement agencies, policymakers, and social organizations) to derive valuable insights that can inform interventions and decision-making.

Result And Analysis:

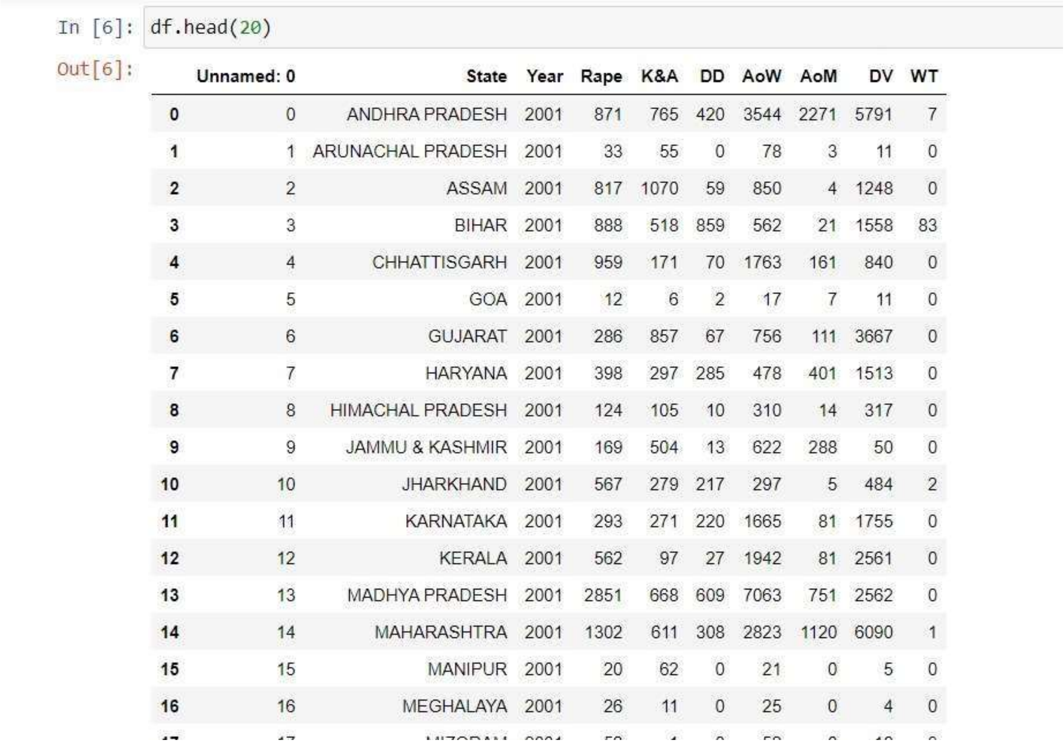


Crimes on women

```
In [3]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt

In [4]: df=pd.read_csv("CrimesOnWomenData.csv")

In [5]: df
Out[5]:
```

	Unnamed: 0	State	Year	Rape	K&A	DD	AoW	AoM	DV	WT
0	0	ANDHRA PRADESH	2001	871	765	420	3544	2271	5791	7
1	1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0
2	2	ASSAM	2001	817	1070	59	850	4	1248	0
3	3	BIHAR	2001	888	518	859	562	21	1558	83
4	4	CHHATTISGARH	2001	959	171	70	1763	161	840	0
...	...	...	...	...	...	...	...	...	...	...
731	731	D&N Haveli	2021	1250	4083	141	2068	417	4731	4
732	732	Daman & Diu	2021	315	904	16	1851	10	501	1
733	733	Delhi UT	2021	2	1	0	5	1	9	0
734	734	Lakshadweep	2021	0	0	0	1	1	3	0

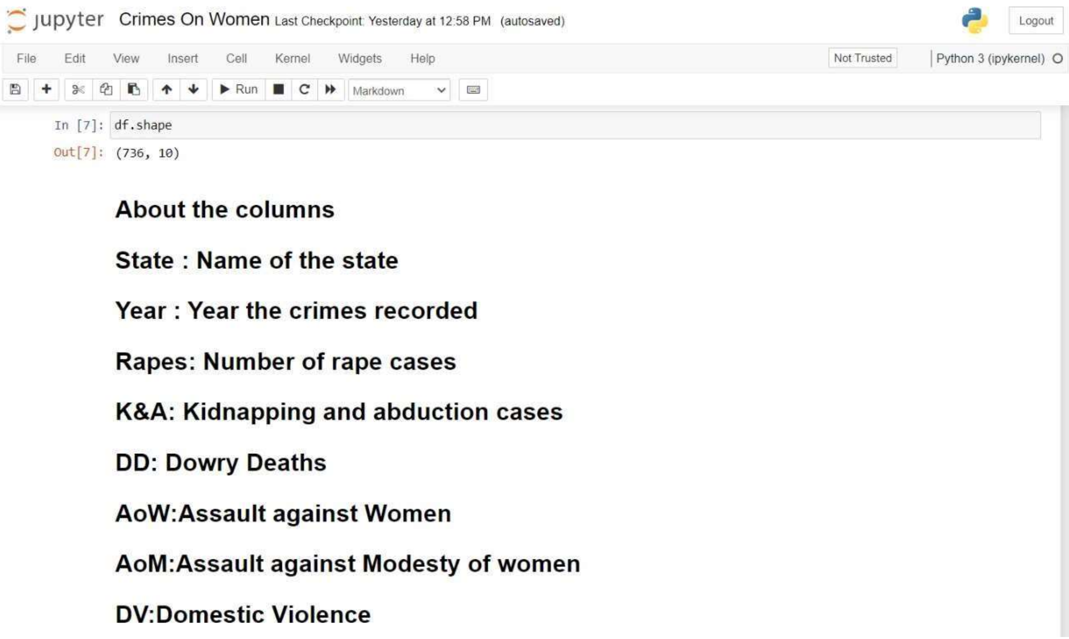
Fig1: This code imports essential libraries for data analysis, statistical computations, and visualization while suppressing warning messages and importing dataset and seeing the first five rows of the dataset.



```
In [6]: df.head(20)
Out[6]:
```

	Unnamed: 0	State	Year	Rape	K&A	DD	AoW	AoM	DV	WT
0	0	ANDHRA PRADESH	2001	871	765	420	3544	2271	5791	7
1	1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0
2	2	ASSAM	2001	817	1070	59	850	4	1248	0
3	3	BIHAR	2001	888	518	859	562	21	1558	83
4	4	CHHATTISGARH	2001	959	171	70	1763	161	840	0
5	5	GOA	2001	12	6	2	17	7	11	0
6	6	GUJARAT	2001	286	857	67	756	111	3667	0
7	7	HARYANA	2001	398	297	285	478	401	1513	0
8	8	HIMACHAL PRADESH	2001	124	105	10	310	14	317	0
9	9	JAMMU & KASHMIR	2001	169	504	13	622	288	50	0
10	10	JHARKHAND	2001	567	279	217	297	5	484	2
11	11	KARNATAKA	2001	293	271	220	1665	81	1755	0
12	12	KERALA	2001	562	97	27	1942	81	2561	0
13	13	MADHYA PRADESH	2001	2851	668	609	7063	751	2562	0
14	14	MAHARASHTRA	2001	1302	611	308	2823	1120	6090	1
15	15	MANIPUR	2001	20	62	0	21	0	5	0
16	16	MEGHALAYA	2001	26	11	0	25	0	4	0

Fig 2: This code provides a summary of a data including number of entries, column data types and non-null counts



```
In [7]: df.shape
Out[7]: (736, 10)
```

About the columns

State : Name of the state

Year : Year the crimes recorded

Rapes: Number of rape cases

K&A: Kidnapping and abduction cases

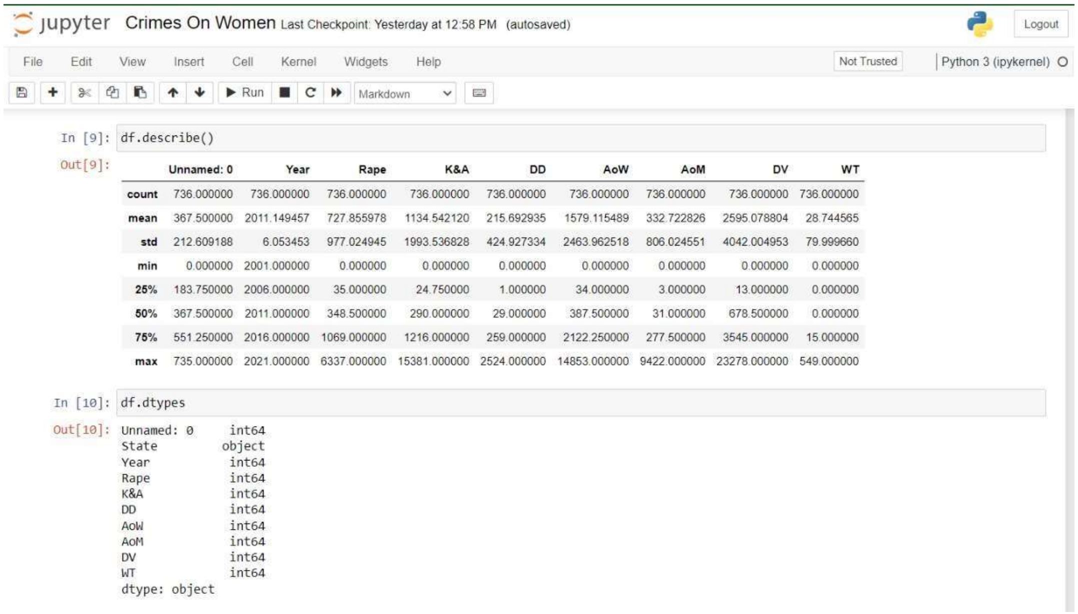
DD: Dowry Deaths

AoW:Assault against Women

AoM:Assault against Modesty of women

DV:Domestic Violence

Fig 3: This markdown shows the type of columns are present in the dataset

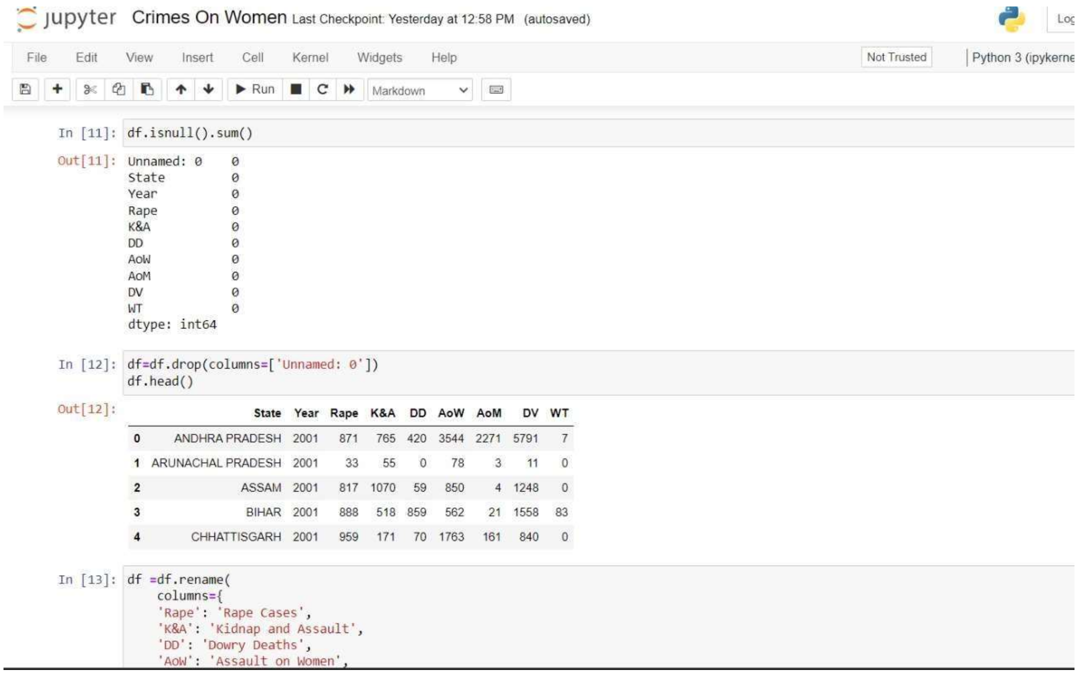


```
In [9]: df.describe()
Out[9]:
```

	Unnamed: 0	Year	Rape	K&A	DD	AoW	AoM	DV	WT
count	736.000000	736.000000	736.000000	736.000000	736.000000	736.000000	736.000000	736.000000	736.000000
mean	367.500000	2011.149457	727.855978	1134.542120	215.692935	1579.115489	332.722826	2595.078804	28.744565
std	212.609188	6.053453	977.024945	1993.536828	424.927334	2463.962518	806.024551	4042.004953	79.999660
min	0.000000	2001.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	183.750000	2006.000000	35.000000	24.750000	1.000000	34.000000	3.000000	13.000000	0.000000
50%	367.500000	2011.000000	348.500000	290.000000	29.000000	387.500000	31.000000	678.500000	0.000000
75%	551.250000	2016.000000	1069.000000	1216.000000	259.000000	2122.250000	277.500000	3545.000000	15.000000
max	735.000000	2021.000000	6337.000000	15381.000000	2524.000000	14853.000000	9422.000000	23278.000000	549.000000

```
In [10]: df.dtypes
Out[10]: Unnamed: 0     int64
         State         object
         Year           int64
         Rape           int64
         K&A            int64
         DD             int64
         AoW            int64
         AoM            int64
         DV             int64
         WT             int64
         dtype: object
```

Fig 4: This code describes the dataset and the data types of the columns



```
In [11]: df.isnull().sum()
```

```
Out[11]: Unnamed: 0    0
         State         0
         Year          0
         Rape          0
         K&A           0
         DD            0
         AoW           0
         AoM           0
         DV            0
         WT            0
         dtype: int64
```

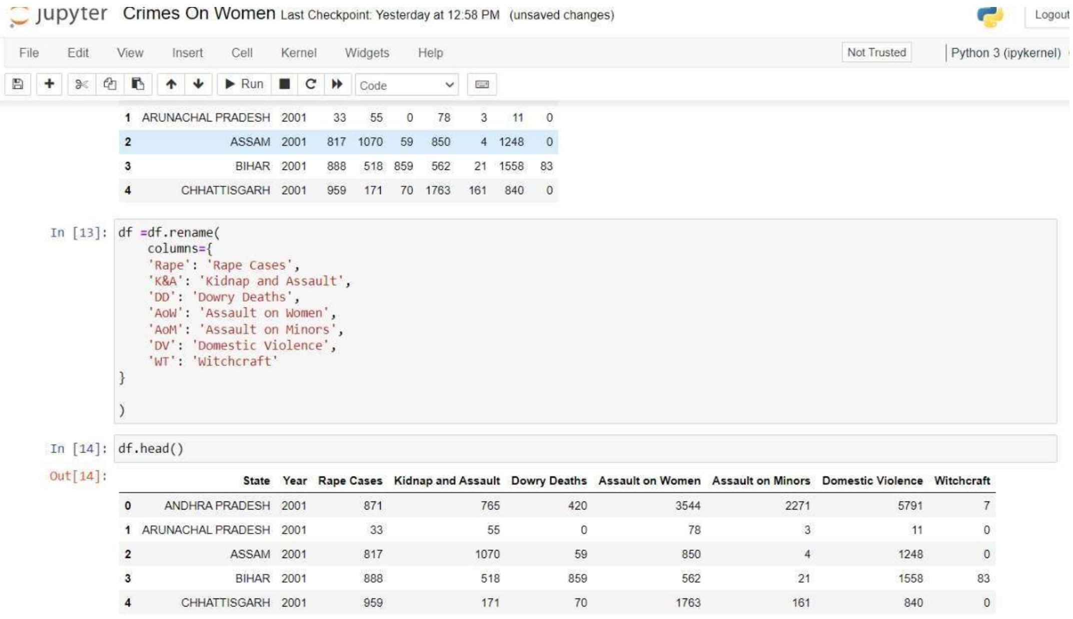
```
In [12]: df=df.drop(columns=['Unnamed: 0'])
         df.head()
```

Out[12]:

	State	Year	Rape	K&A	DD	AoW	AoM	DV	WT
0	ANDHRA PRADESH	2001	871	765	420	3544	2271	5791	7
1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0
2	ASSAM	2001	817	1070	59	850	4	1248	0
3	BIHAR	2001	888	518	859	562	21	1558	83
4	CHHATTISGARH	2001	959	171	70	1763	161	840	0

```
In [13]: df =df.rename(
             columns={
             'Rape': 'Rape Cases',
             'K&A': 'Kidnap and Assault',
             'DD': 'Dowry Deaths',
             'AoW': 'Assault on Women',
```

Fig 5: This code shows the null values and drops the unwanted column



1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0
2	ASSAM	2001	817	1070	59	850	4	1248	0
3	BIHAR	2001	888	518	859	562	21	1558	83
4	CHHATTISGARH	2001	959	171	70	1763	161	840	0

```
In [13]: df =df.rename(
             columns={
             'Rape': 'Rape Cases',
             'K&A': 'Kidnap and Assault',
             'DD': 'Dowry Deaths',
             'AoW': 'Assault on Women',
             'AoM': 'Assault on Minors',
             'DV': 'Domestic Violence',
             'WT': 'Witchcraft'
         }

         )
```

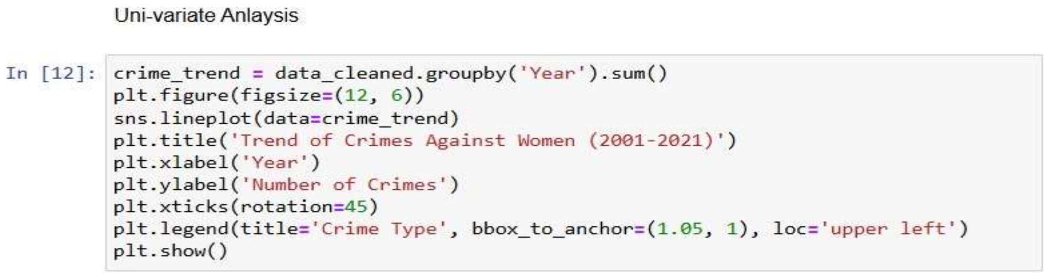
```
In [14]: df.head()
```

Out[14]:

	State	Year	Rape Cases	Kidnap and Assault	Dowry Deaths	Assault on Women	Assault on Minors	Domestic Violence	Witchcraft
0	ANDHRA PRADESH	2001	871	765	420	3544	2271	5791	7
1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0
2	ASSAM	2001	817	1070	59	850	4	1248	0
3	BIHAR	2001	888	518	859	562	21	1558	83
4	CHHATTISGARH	2001	959	171	70	1763	161	840	0

Fig 6: This shows the renaming of columns and gives the first 5 rows of the data

Univariate Analysis:



Uni-variate Anlaysis

```
In [12]: crime_trend = data_cleaned.groupby('Year').sum()
         plt.figure(figsize=(12, 6))
         sns.lineplot(data=crime_trend)
         plt.title('Trend of Crimes Against Women (2001-2021)')
         plt.xlabel('Year')
         plt.ylabel('Number of Crimes')
         plt.xticks(rotation=45)
         plt.legend(title='Crime Type', bbox_to_anchor=(1.05, 1), loc='upper left')
         plt.show()
```

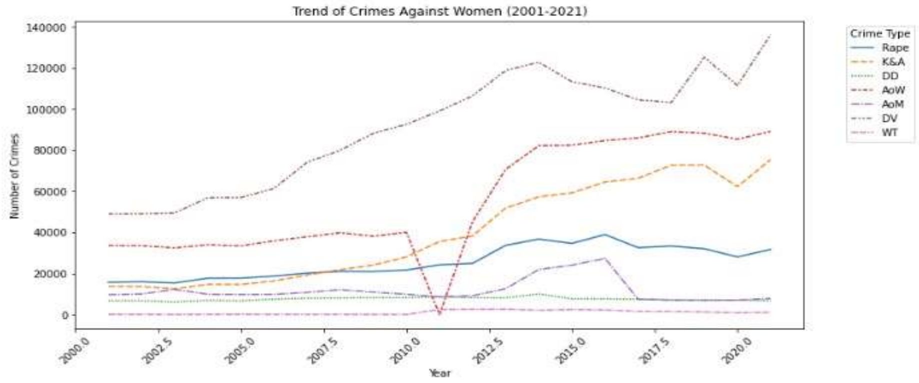
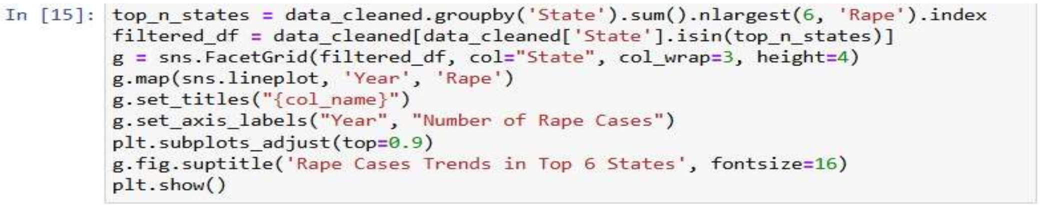


Fig 7: This line chart represents the number of crimes recorded from 2001-2021



```
In [15]: top_n_states = data_cleaned.groupby('State').sum().nlargest(6, 'Rape').index
         filtered_df = data_cleaned[data_cleaned['State'].isin(top_n_states)]
         g = sns.FacetGrid(filtered_df, col="State", col_wrap=3, height=4)
         g.map(sns.lineplot, 'Year', 'Rape')
         g.set_titles("{col_name}")
         g.set_axis_labels("Year", "Number of Rape Cases")
         plt.subplots_adjust(top=0.9)
         g.fig.suptitle('Rape Cases Trends in Top 6 States', fontsize=16)
         plt.show()
```

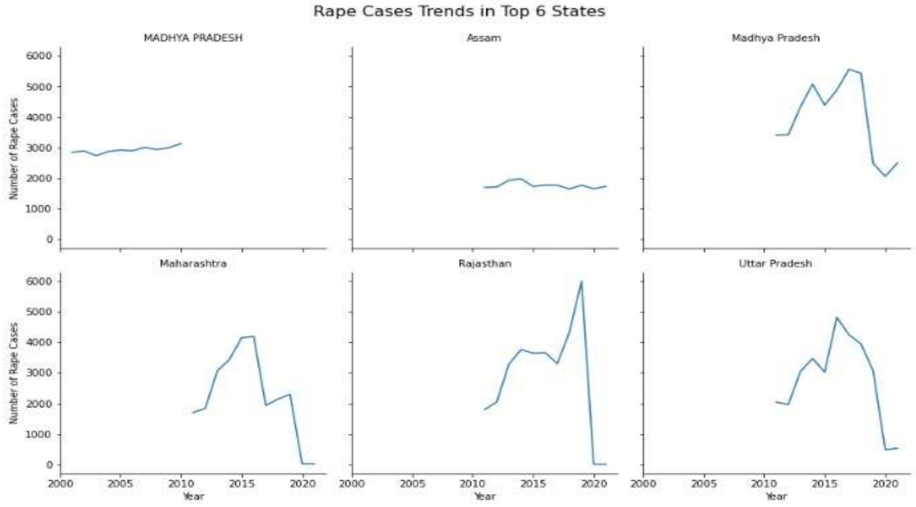
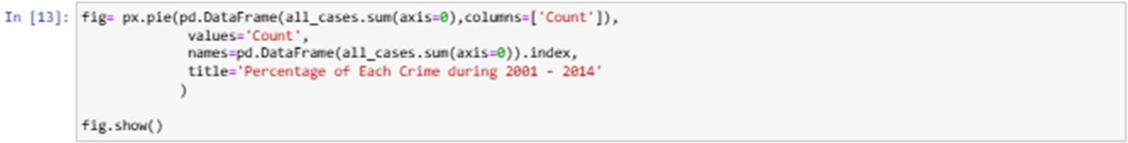


Fig 8: This line plot represents the rape case trends in top 6 states



```
In [13]: fig= px.pie(pd.DataFrame(all_cases.sum(axis=0),columns=['Count']),
                     values='Count',
                     names=pd.DataFrame(all_cases.sum(axis=0)).index,
                     title='Percentage of Each Crime during 2001 - 2014'
                    )

         fig.show()
```

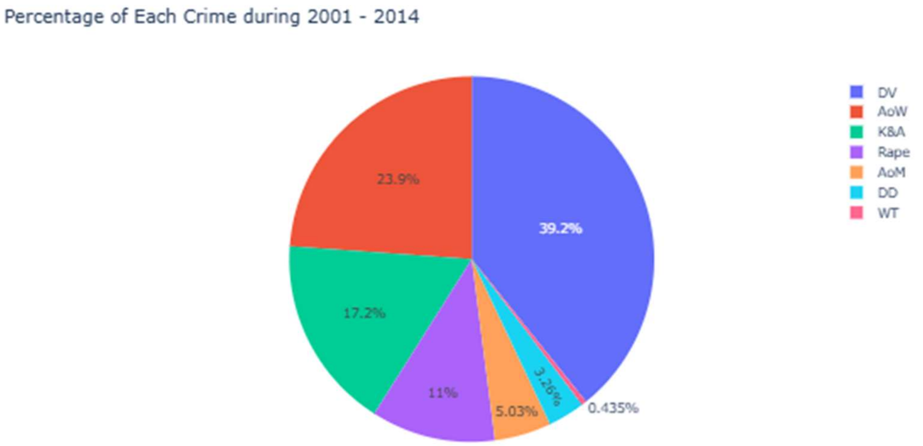
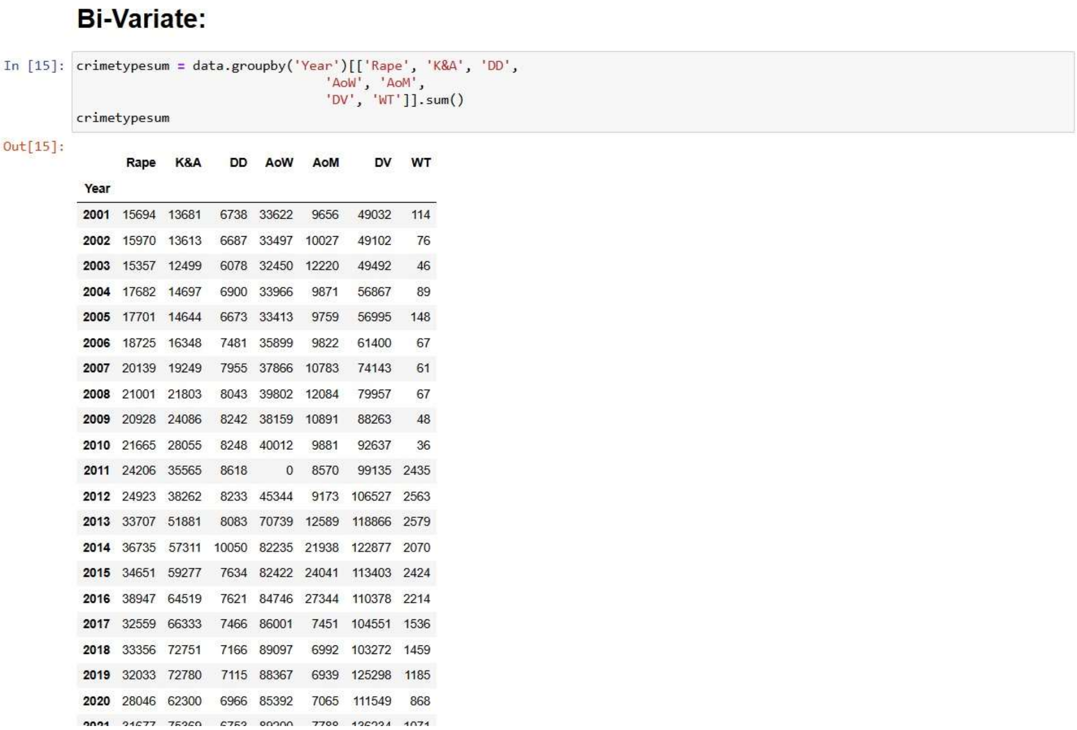


Fig 9: This pie chart percentage of each crime during 2001-2014

Bi-variate Analysis :



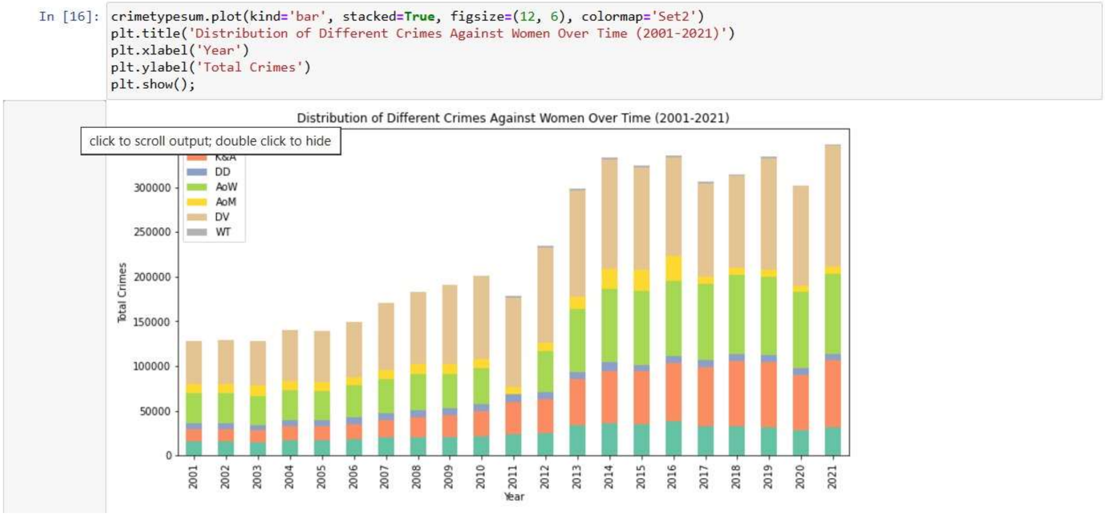
Bi-Variate:

```
In [15]: crimetypesum = data.groupby('Year')[['Rape', 'K&A', 'DD',
                                               'AoW', 'AoM',
                                               'DV', 'WT']].sum()
         crimetypesum
```

Out[15]:

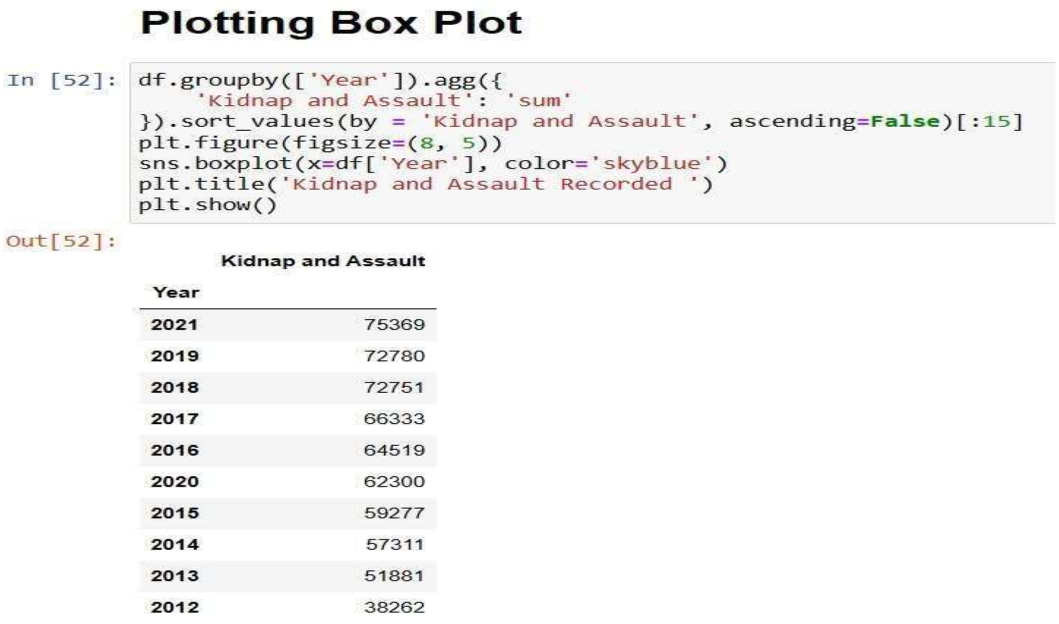
Year	Rape	K&A	DD	AoW	AoM	DV	WT
2001	15694	13681	6738	33622	9656	49032	114
2002	15970	13613	6687	33497	10027	49102	76
2003	15357	12499	6078	32450	12220	49492	46
2004	17682	14697	6900	33966	9871	56867	89
2005	17701	14644	6673	33413	9759	56995	148
2006	18725	16348	7481	35899	9822	61400	67
2007	20139	19249	7955	37866	10783	74143	61
2008	21001	21803	8043	39802	12084	79957	67
2009	20928	24086	8242	38159	10891	88263	48
2010	21665	28055	8248	40012	9881	92637	36
2011	24206	35565	8618	0	8570	99135	2435
2012	24923	38262	8233	45344	9173	106527	2563
2013	33707	51881	8083	70739	12589	118866	2579
2014	36735	57311	10050	82235	21938	122877	2070
2015	34651	59277	7634	82422	24041	113403	2424
2016	38947	64519	7621	84746	27344	110378	2214
2017	32559	66333	7466	86001	7451	104551	1536
2018	33356	72751	7166	89097	6992	103272	1459
2019	32033	72780	7115	88367	6939	125298	1185
2020	28046	62300	6966	85392	7065	111549	868
2021	31677	75369	6753	89200	7788	136234	1071

Fig 10: This represents the crime types from 2001-2021



```
In [16]: crimetypesum.plot(kind='bar', stacked=True, figsize=(12, 6), colormap='Set2')
         plt.title('Distribution of Different Crimes Against Women Over Time (2001-2021)')
         plt.xlabel('Year')
         plt.ylabel('Total Crimes')
         plt.show();
```

Fig 11: This stacked bar chart represents the crimes recorded from 2001-2021

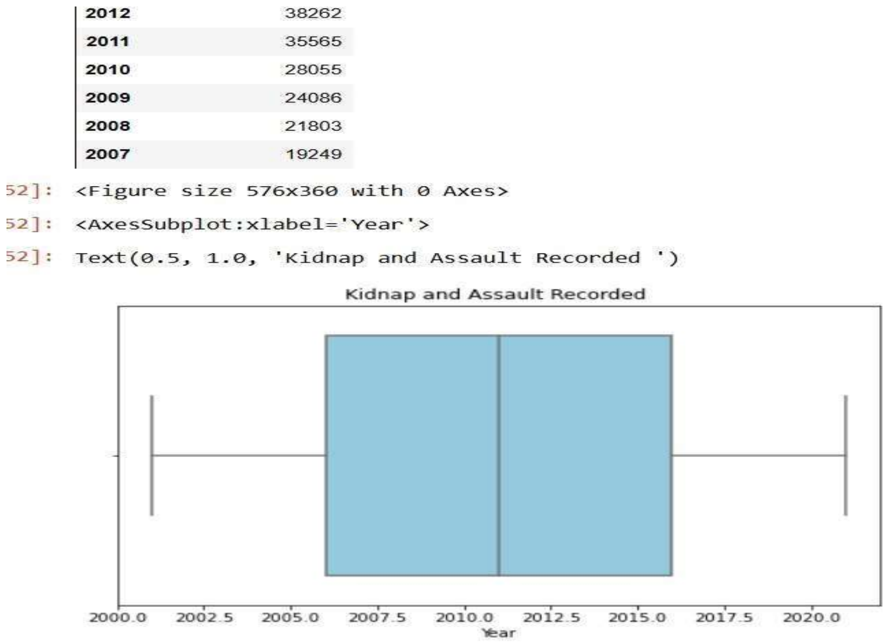


Plotting Box Plot

```
In [52]: df.groupby(['Year']).agg({
             'Kidnap and Assault': 'sum'
         }).sort_values(by = 'Kidnap and Assault', ascending=False)[:15]
         plt.figure(figsize=(8, 5))
         sns.boxplot(x=df['Year'], color='skyblue')
         plt.title('Kidnap and Assault Recorded ')
         plt.show()
Out[52]:
```

Year	Kidnap and Assault
2021	75369
2019	72780
2018	72751
2017	66333
2016	64519
2020	62300
2015	59277
2014	57311
2013	51881
2012	38262

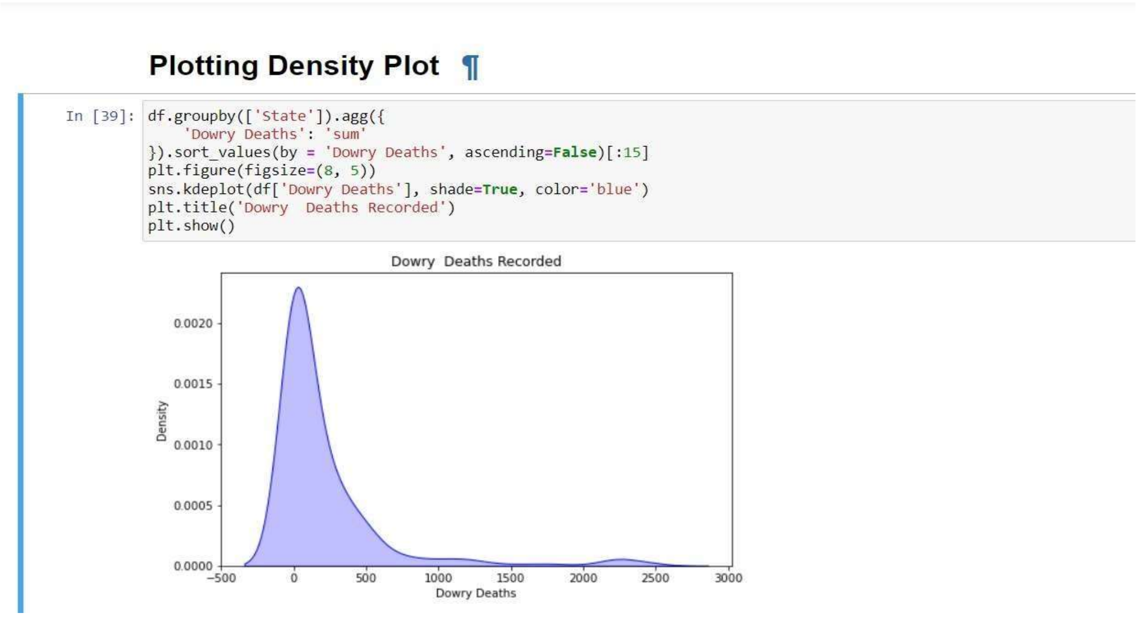
Fig 12: This table represents K&A data from 2021-2001(Descending)



2012	38262
2011	35565
2010	28055
2009	24086
2008	21803
2007	19249

```
52]: <Figure size 576x360 with 0 Axes>
52]: <AxesSubplot:xlabel='Year'>
52]: Text(0.5, 1.0, 'Kidnap and Assault Recorded ')
```

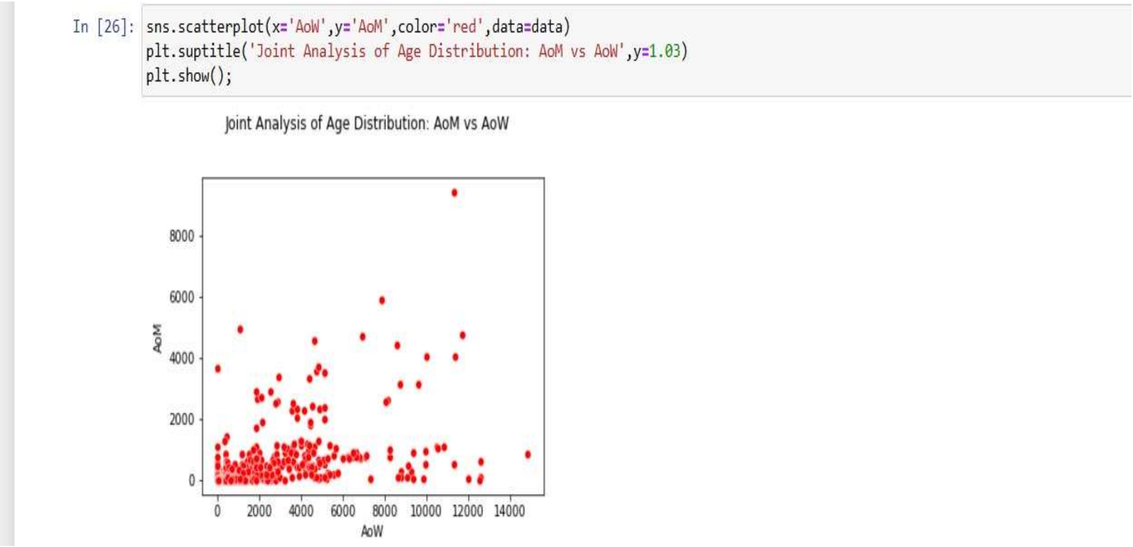
Fig 13: This box plot represent K&A recorded for 2001-2021 with no outliers



Plotting Density Plot ¶

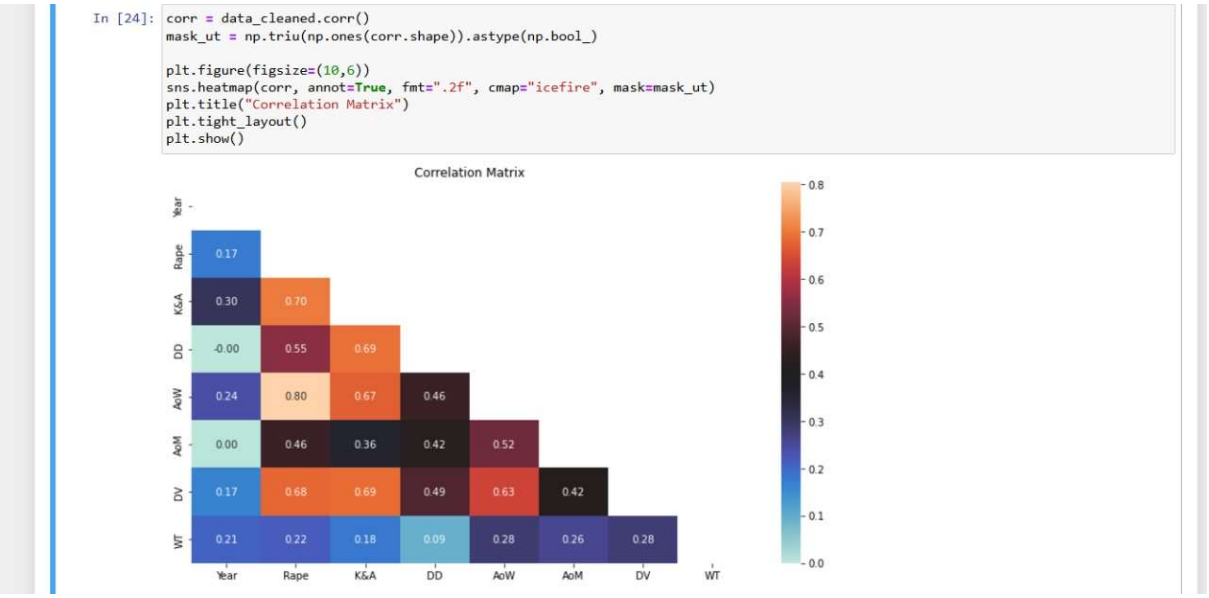
```
In [39]: df.groupby(['State']).agg({
             'Dowry Deaths': 'sum'
         }).sort_values(by = 'Dowry Deaths', ascending=False)[:15]
         plt.figure(figsize=(8, 5))
         sns.kdeplot(df['Dowry Deaths'], shade=True, color='blue')
         plt.title('Dowry  Deaths Recorded')
         plt.show()
```

Fig 14: This density plot represents the Dowry Deaths recorded from 2001-2021



```
In [26]: sns.scatterplot(x='AoW',y='AoM',color='red',data=data)
         plt.suptitle('Joint Analysis of Age Distribution: AoM vs AoW',y=1.03)
         plt.show();
```

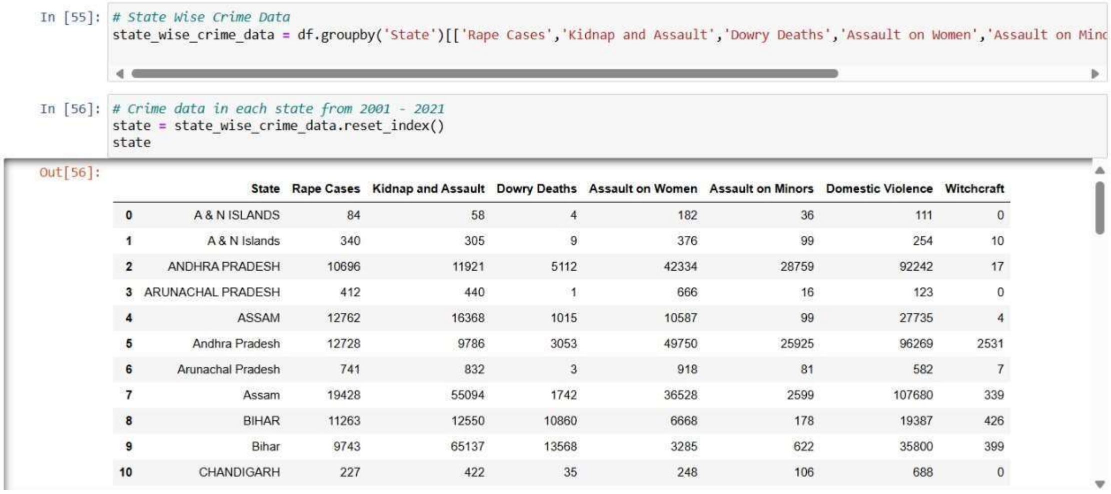
Fig 15: This Scatter plot represents the joint analysis of age distribution



```
In [24]: corr = data_cleaned.corr()
         mask_ut = np.triu(np.ones(corr.shape)).astype(np.bool_)

         plt.figure(figsize=(10,6))
         sns.heatmap(corr, annot=True, fmt=".2f", cmap="icefire", mask=mask_ut)
         plt.title("Correlation Matrix")
         plt.tight_layout()
         plt.show()
```

Fig 16: This represents the correlation between all the columns

Multi variate analysis:

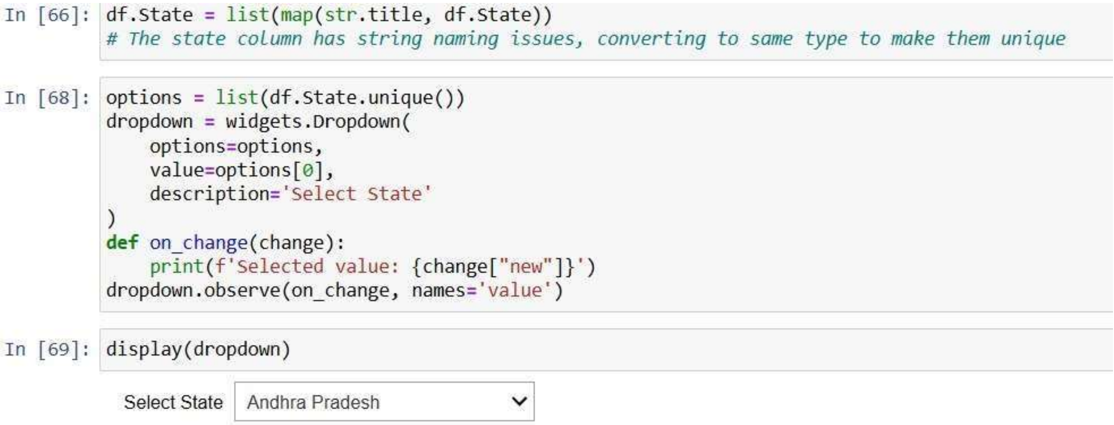


```
In [55]: # State Wise Crime Data
         state_wise_crime_data = df.groupby('State')[['Rape Cases','Kidnap and Assault','Dowry Deaths','Assault on Women','Assault on Minc

In [56]: # Crime data in each state from 2001 - 2021
         state = state_wise_crime_data.reset_index()
         state
```

Out[56]:

	State	Rape Cases	Kidnap and Assault	Dowry Deaths	Assault on Women	Assault on Minors	Domestic Violence	Witchcraft
0	A & N ISLANDS	84	58	4	182	36	111	0
1	A & N Islands	340	305	9	376	99	254	10
2	ANDHRA PRADESH	10696	11921	5112	42334	28759	92242	17
3	ARUNACHAL PRADESH	412	440	1	666	16	123	0
4	ASSAM	12762	16368	1015	10587	99	27735	4
5	Andhra Pradesh	12728	9786	3053	49750	25925	96269	2531
6	Arunachal Pradesh	741	832	3	918	81	582	7
7	Assam	19428	55094	1742	36528	2599	107680	339
8	BIHAR	11263	12550	10860	6668	178	19387	426
9	Bihar	9743	65137	13568	3285	622	35800	399
10	CHANDIGARH	227	422	35	248	106	688	0

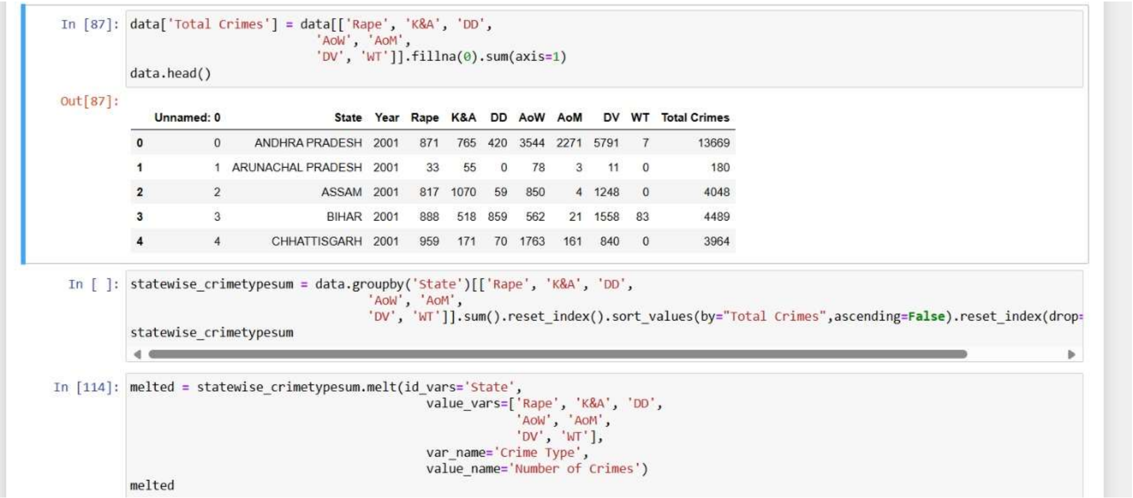
Fig 17: This table shows the state wise crime data



```
In [66]: df.State = list(map(str.title, df.State))
         # The state column has string naming issues, converting to same type to make them unique

In [68]: options = list(df.State.unique())
         dropdown = widgets.Dropdown(
             options=options,
             value=options[0],
             description='Select State'
         )
         def on_change(change):
             print(f'Selected value: {change["new"]}')
         dropdown.observe(on_change, names='value')

In [69]: display(dropdown)
```

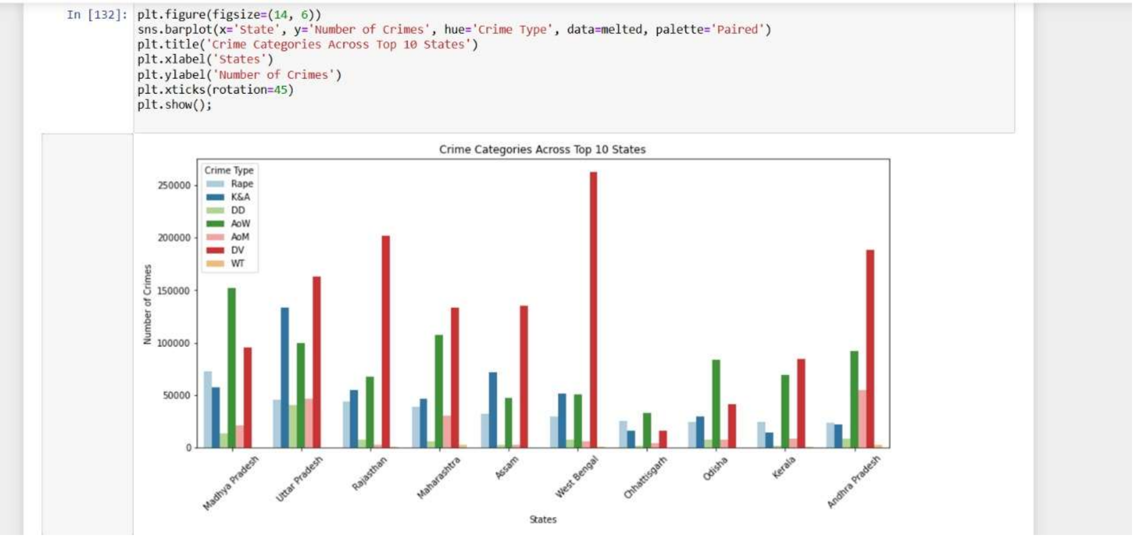
Select State: Andhra Pradesh

Fig 18: This drop-down button shows the unique states



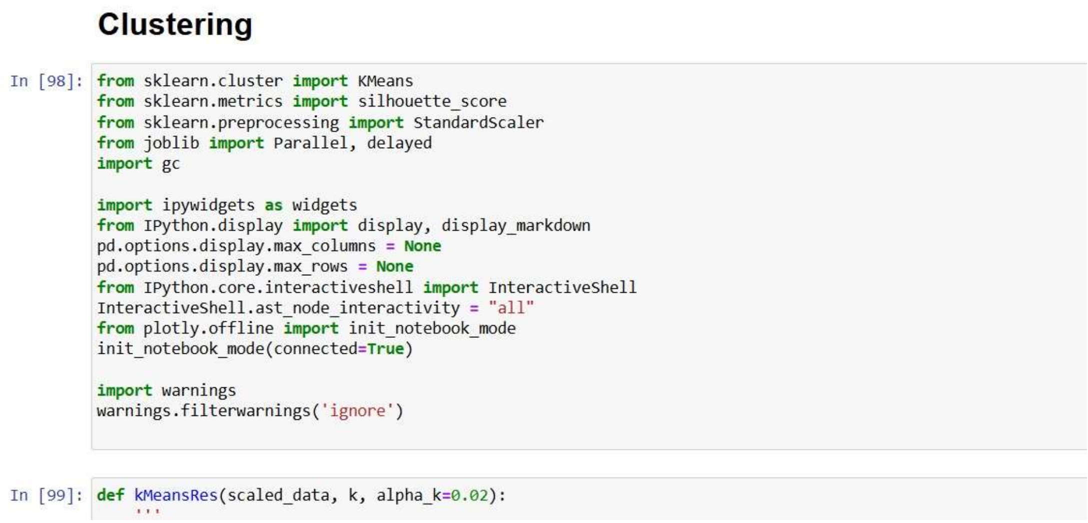
```
In [87]: data['Total Crimes'] = data[['Rape', 'K&A', 'DD',
                             'AoW', 'AoM',
                             'DV', 'WT']].fillna(0).sum(axis=1)
         data.head()

Out[87]:
            Unnamed: 0              State  Year  Rape   K&A   DD   AoW   AoM    DV  WT  Total Crimes
         0           0     ANDHRA PRADESH  2001   871   765  420  3544  2271  5791   7         13669
         1           1  ARUNACHAL PRADESH  2001    33    55    0    78     3    11   0           180
         2           2              ASSAM  2001   817  1070   59   850     4  1248   0          4048
         3           3              BIHAR  2001   888   518  859   562    21  1558  83          4489
         4           4       CHHATTISGARH  2001   959   171   70  1763   161   840   0          3964

In [ ]:  statewise_crimetypesum = data.groupby('State')[['Rape', 'K&A', 'DD',
                                     'AoW', 'AoM',
                                     'DV', 'WT']].sum().reset_index().sort_values(by="Total Crimes",ascending=False).reset_index(drop=
         statewise_crimetypesum

In [114]: melted = statewise_crimetypesum.melt(id_vars='State',
                                               value_vars=['Rape', 'K&A', 'DD',
                                                           'AoW', 'AoM',
                                                           'DV', 'WT'],
                                               var_name='Crime Type',
                                               value_name='Number of Crimes')
          melted
```

Fig 19: This code represents the total crimes recorded



```
In [132]: plt.figure(figsize=(14, 6))
          sns.barplot(x='State', y='Number of Crimes', hue='Crime Type', data=melted, palette='Paired')
          plt.title('Crime Categories Across Top 10 States')
          plt.xlabel('States')
          plt.ylabel('Number of Crimes')
          plt.xticks(rotation=45)
          plt.show();
```

Fig 20: This Bar graph shows the crime categories across top 10 states

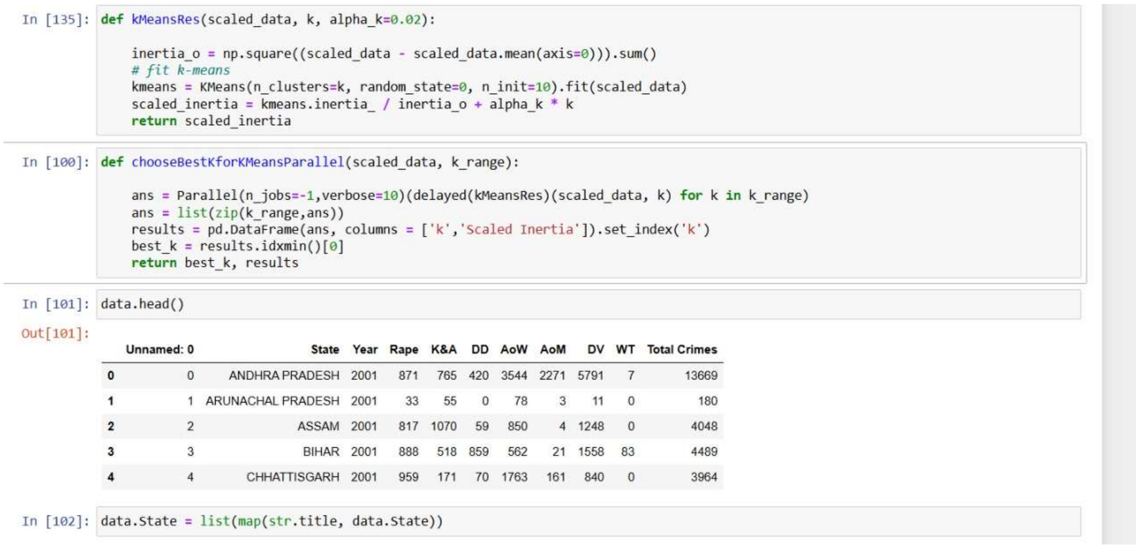


Clustering

```
In [98]: from sklearn.cluster import KMeans
         from sklearn.metrics import silhouette_score
         from sklearn.preprocessing import StandardScaler
         from joblib import Parallel, delayed
         import gc

         import ipywidgets as widgets
         from IPython.display import display, display_markdown
         pd.options.display.max_columns = None
         pd.options.display.max_rows = None
         from IPython.core.interactiveshell import InteractiveShell
         InteractiveShell.ast_node_interactivity = "all"
         from plotly.offline import init_notebook_mode
         init_notebook_mode(connected=True)

         import warnings
         warnings.filterwarnings('ignore')
```

```
In [99]: def kMeansRes(scaled_data, k, alpha_k=0.02):
             '''
```

Fig 21: This code imports the libraires for the clustering



```
In [135]: def kMeansRes(scaled_data, k, alpha_k=0.02):

              inertia_o = np.square((scaled_data - scaled_data.mean(axis=0))).sum()
              # fit k-means
              kmeans = KMeans(n_clusters=k, random_state=0, n_init=10).fit(scaled_data)
              scaled_inertia = kmeans.inertia_ / inertia_o + alpha_k * k
              return scaled_inertia
```

```
In [100]: def chooseBestKforKMeansParallel(scaled_data, k_range):

              ans = Parallel(n_jobs=-1,verbose=10)(delayed(kMeansRes)(scaled_data, k) for k in k_range)
              ans = list(zip(k_range,ans))
              results = pd.DataFrame(ans, columns = ['k','Scaled Inertia']).set_index('k')
              best_k = results.idxmin()[0]
              return best_k, results
```

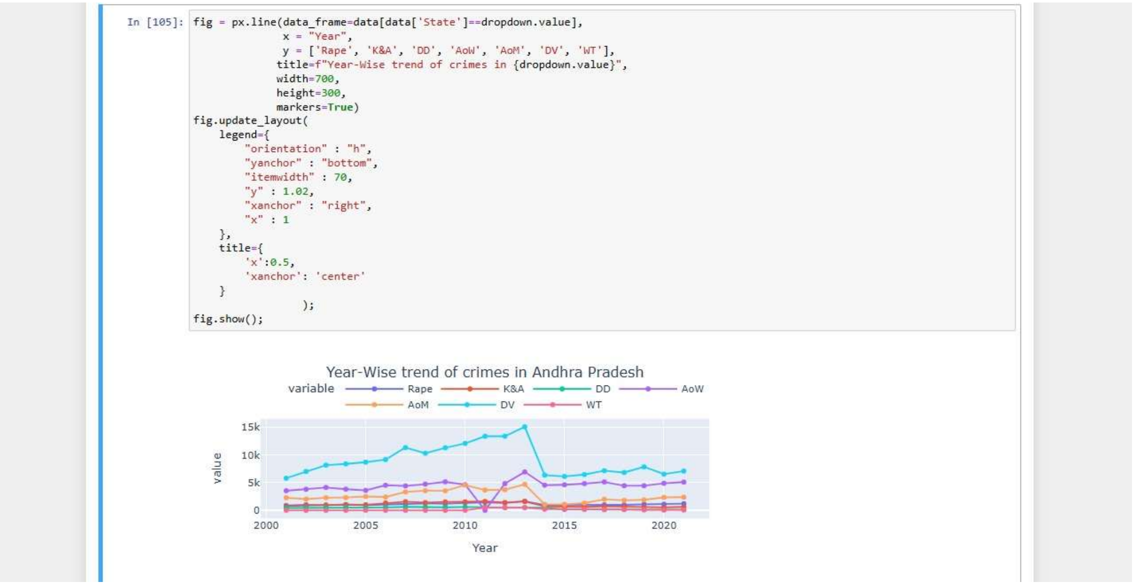
```
In [101]: data.head()
```

Out[101]:

	Unnamed: 0	State	Year	Rape	K&A	DD	AoW	AoM	DV	WT	Total Crimes
0	0	ANDHRA PRADESH	2001	871	765	420	3544	2271	5791	7	13669
1	1	ARUNACHAL PRADESH	2001	33	55	0	78	3	11	0	180
2	2	ASSAM	2001	817	1070	59	850	4	1248	0	4048
3	3	BIHAR	2001	888	518	859	562	21	1558	83	4489
4	4	CHHATTISGARH	2001	959	171	70	1763	161	840	0	3964

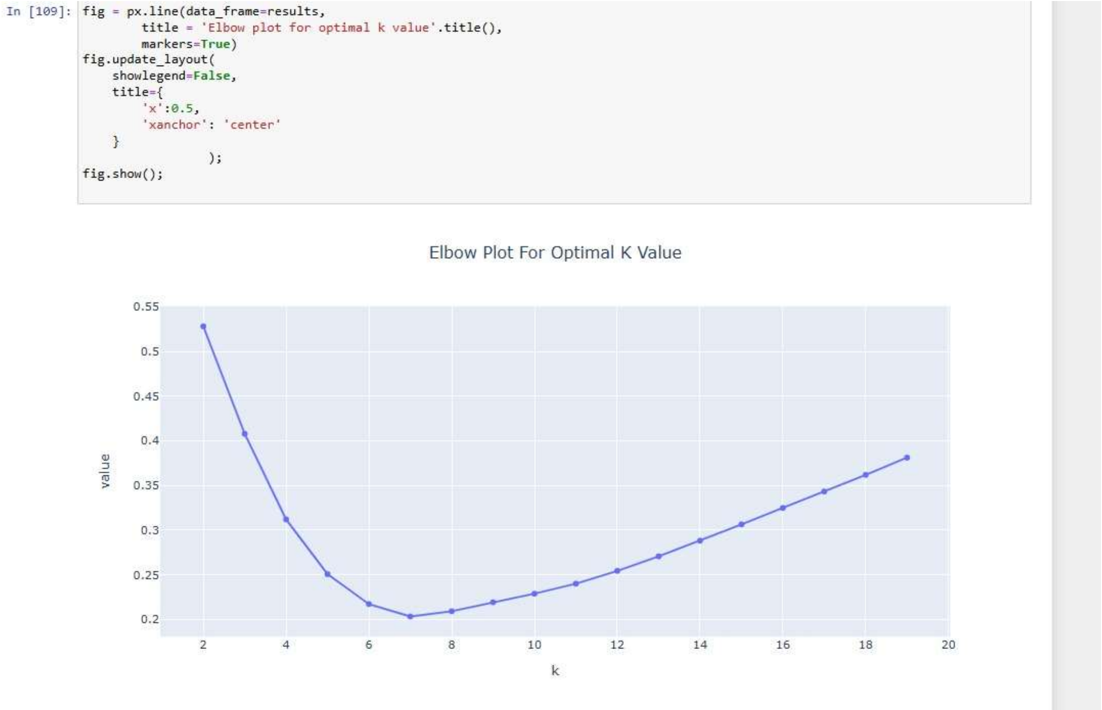
```
In [102]: data.State = list(map(str.title, data.State))
```

Fig 22: This code represents the clustering



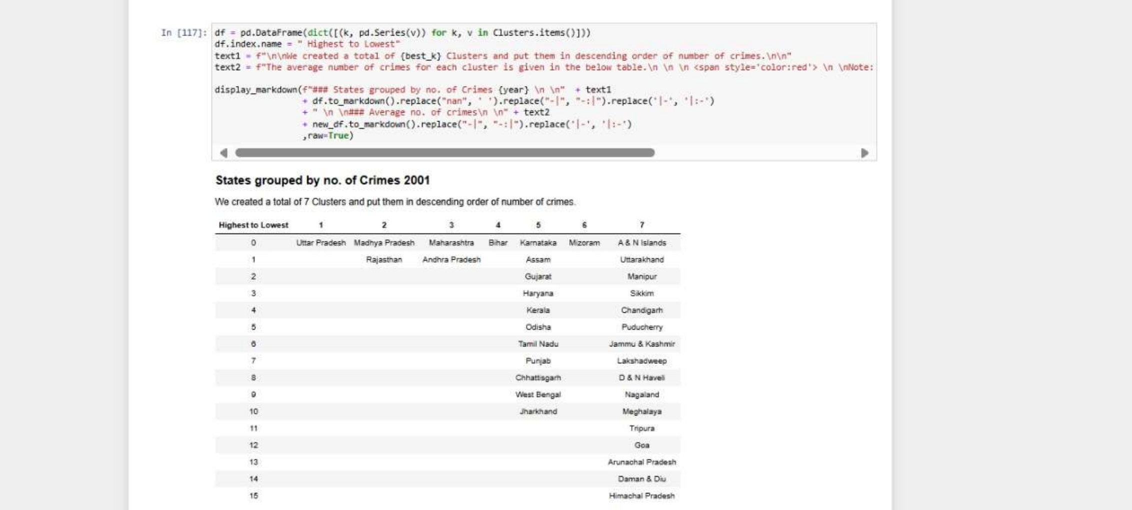
```
In [105]: fig = px.line(data_frame=data[data['State']==dropdown.value],
                         x = "Year",
                         y = ['Rape', 'K&A', 'DD', 'AoW', 'AoM', 'DV', 'WT'],
                        title=f"Year-Wise trend of crimes in {dropdown.value}",
                        width=700,
                        height=300,
                        markers=True)
          fig.update_layout(
              legend={
                  "orientation" : "h",
                  "yanchor" : "bottom",
                  "itemwidth" : 70,
                  "y" : 1.02,
                  "xanchor" : "right",
                  "x" : 1
              },
              title={
                  'x':0.5,
                  'xanchor': 'center'
              }
                           );
          fig.show();
```

Fig 23: This code and graph represents the Year-Wise trend of crimes in Andhra Pradesh



```
In [109]: fig = px.line(data_frame=results,
                  title = 'Elbow plot for optimal k value'.title(),
                  markers=True)
          fig.update_layout(
              showlegend=False,
              title={
                  'x':0.5,
                  'xanchor': 'center'
              }
                          );
          fig.show();
```

Fig 24: This plot shows the Elbow plot for Optimal K-value



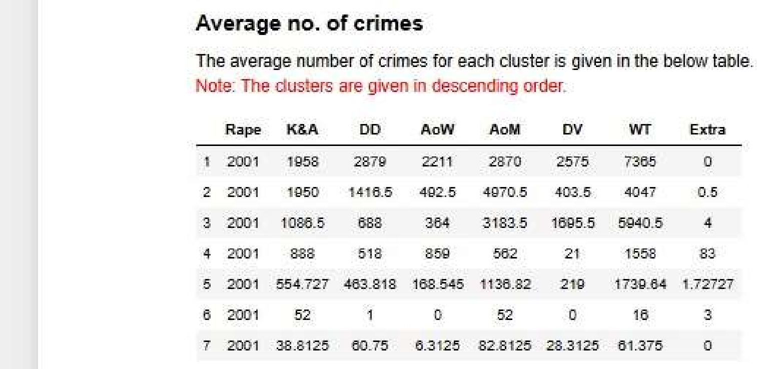
```
In [117]: df = pd.DataFrame(dict([(k, pd.Series(v)) for k, v in Clusters.items()]))
          df.index.name = " Highest to Lowest"
          text1 = f"\n\nWe created a total of {best_k} Clusters and put them in descending order of number of crimes.\n\n"
          text2 = f"The average number of crimes for each cluster is given in the below table.\n \n \n <span style='color:red'> \n \nNote:

          display_markdown(f"### States grouped by no. of Crimes {year} \n \n"  + text1
                           + df.to_markdown().replace("nan", ' ').replace("-|", "-:|").replace('|-', '|:-')
                           + " \n \n### Average no. of crimes\n \n" + text2
                           + new_df.to_markdown().replace("-|", "-:|").replace('|-', '|:-')
                           ,raw=True)
```

States grouped by no. of Crimes 2001

We created a total of 7 Clusters and put them in descending order of number of crimes.

Highest to Lowest	1	2	3	4	5	6	7
0	Uttar Pradesh	Madhya Pradesh	Maharashtra	Bihar	Karnataka	Mizoram	A & N Islands
1		Rajasthan	Andhra Pradesh		Assam		Uttarakhand
2					Gujarat		Manipur
3					Haryana		Sikkim
4					Kerala		Chandigarh
5					Odisha		Puducherry
6					Tamil Nadu		Jammu & Kashmir
7					Punjab		Lakshadweep
8					Chhattisgarh		D & N Haveli
9					West Bengal		Nagaland
10					Jharkhand		Meghalaya
11							Tripura
12							Goa
13							Arunachal Pradesh
14							Daman & Diu
15							Himachal Pradesh

Fig 25: This table represents the states grouped by number of crimes in 2001



Average no. of crimes

The average number of crimes for each cluster is given in the below table.

Note: The clusters are given in descending order.

	Rape	K&A	DD	AoW	AoM	DV	WT	Extra
1	2001	1958	2879	2211	2870	2575	7365	0
2	2001	1950	1416.5	492.5	4970.5	403.5	4047	0.5
3	2001	1086.5	688	364	3183.5	1695.5	5940.5	4
4	2001	888	518	859	562	21	1558	83
5	2001	554.727	463.818	168.545	1136.82	219	1739.64	1.72727
6	2001	52	1	0	52	0	16	3
7	2001	38.8125	60.75	6.3125	82.8125	28.3125	61.375	0

Fig 26: This table represents the average number of crimes

Conclusion

Exploratory Data Analysis (EDA) on Crimes Against Women in India (2001-2021) serves as a powerful tool for understanding the complex dynamics of gender-based violence across the country. By applying various statistical and visualization techniques, EDA uncovers significant insights into the prevalence, patterns, and socio-economic factors contributing to crimes against women. This data-driven approach is critical for identifying trends and disparities in crime rates, as well as recognizing the factors that influence the likelihood of such crimes occurring. One of the key aspects of EDA is its ability to reveal temporal patterns in crime data. Through a detailed examination of crime rates over time, EDA uncovers fluctuations in criminal activity, such as spikes in certain crimes during specific months, seasons, or years. This could be indicative of seasonal patterns or responses to significant socio-political events, such as changes in the legal environment or public awareness campaigns. By understanding these temporal patterns, policymakers and law enforcement agencies can better prepare for periods of heightened risk, allocate resources effectively, and implement more timely interventions. Moreover, EDA allows for the evaluation of the effectiveness of legal reforms, policies, and public awareness campaigns. By comparing crime rates before and after specific reforms or interventions, it is possible to assess whether such efforts have had a meaningful impact on reducing gender-based violence. This data-driven assessment can guide future policy decisions and ensure that resources are directed where they are most needed. In conclusion, EDA on Crimes Against Women in India provides a comprehensive understanding of the complex issue of gender-based violence. By revealing patterns, risk factors, and the effectiveness of interventions, EDA empowers policymakers, law enforcement agencies, and social organizations to take informed, proactive measures to create a safer, more equitable society for women in India. Through a data-driven approach, it is possible to address the root causes of violence, improve response strategies, and ensure lasting change.

Reference

https://www.kaggle.com/datasets/balajivaraprasad/crimes-against-women-in-india-2001-2021