Exploratory analysis of crime against women in India

Diksha Tripathi
M.Sc. Computational Science and Applications
Banaras Hindu University

Abhilasha Ojha
M.Sc. Computational Science and Applications
Banaras Hindu University

Avishree Roy
M.Sc. Computational Science and Applications
Banaras Hindu University

Concise description: To analyze trends in crime against women across various India states in order to shed light on the location of violence perpetrated.

Abstract: **Crime against women has been a rising trend in India for the past few years. These crimes are mostly committed behind closed doors and lot of them go unreported. The national criminal records bureau (NCRB) has been collecting information regarding crime against women in India for each district, every year since 2001. We take a look at 15 states and 3 union territories and compare the crime rate against women with the help of exploratory data analysis. We see the states in which crime rate is high. We also compare the districts where one type of crime is likelier than the other. In doing so we find out a trend in the rise of crime against women in India.**

Keywords: **exploratory data analysis, data preprocessing, trend analysis, crime, women.**

I. INTRODUCTION

In light of the recent cases of violence against women in India, we felt that it was necessary to develop an understanding of the trend in crime for different states of India. This understanding will further help in the development of hotspot analysis making it easier for the authorities to pinpoint the necessary measures to take in these locations so that violence against women can be prevented and swift justice can be taken against the perpetrators. Hundreds of cases are recorded daily by the data officers working alongside the law enforcement authorities throughout India. The national bureau of Criminal records has made this crime data accessible to the general public. This allows individuals and researchers to parse through the data and figure out where the situation can be improved. The data scientists and engineers working alongside the National Bureau have recorded over 1,000,000 crime cases against women spanning over 255 districts in the 26 states and 9 union territories of India. With the help of this historical data, many patterns can be uncovered. This can help us determine hotspots where crimes may happen in the future thereby helping the police better safeguard the population of the country. Hopefully in a future scenario crime against women in India would be much reduced with the help of analysis that we have provided today. We employ the crime data set reported by NCRB over a period of 13 years (2001 to 2013) and analyze them to identify the trends of crimes over the years.

Compared to previous studies that have worked with the similar data, we have not applied any prediction analysis and the work we have done is useful to observe the trend in crime rather than predict any change. We hope that our work helps shed some light on the often-ignored area of criminal research that is violence against women.

II. GOALS

Our work consists of exploratory data analysis using graphs, tables and regression boundary that help us describe the data that we have obtained. The steps involved are:

i. Utilizing the crime data set by the NBCR to observe existing patterns in the crime throughout the districts of India.
ii. Determining the crime rate for a population of 100,000 within different districts of the country, and analyzing the spread and impact of the crime.
iii. Studying the crime spread in each state based on the district and the number of cases reported from them.

For the exploratory data analysis, we employ various data analytics tools, provided by MS-EXCEL, to analyze the spread of the crime in the states, and find the crime classes. We use the tools available in MS-EXCEL to visualize the data in the form charts and tables.

III. DESIGN

The main goal of this work is to observe and define trends in crime against women in India. In order to define these trends a few additions were made to the dataset which provided us with an easier way of

viewing the total crime. This was achieved by following these steps:

a. For every district the crime rate is calculated with the help of the formula:

$$\frac{Sum\ of\ total\ crime\ in\ an\ district}{Population\ of\ the\ district}$$

b. The number of cases is then standardized to provide a rate of a particular crime for every 100,000 persons.
c. The data analysis tools provided by MS-Excel are utilized to see the trend in crime across all districts.

A. Overview of the dataset:

We used the NCRB district-wise crime against women data set. The data set consists of the following attributes:

i. STATE: The name of the state in India in which the crime was committed.
ii. DISTRICT: The name of the district in the state where the crime was committed.
iii. YEAR: The year in which these crimes took place, having values from 2001 – 2013.
iv. Rape: The total number of Rape cases.
v. Kidnapping & Abduction: The total number of abduction or kidnapping cases.
vi. Dowry Deaths: The total number of Dowry Deaths.
vii. Assault on women with intent to outrage her modesty: The total number of cases in this category.
viii. Insult to modesty of Women: The total number of cases in this category. These cases are separate to assault cases.
ix. Cruelty by Husband or his Relatives: The total number of cases in this category.
x. Importation of girls: The total number of cases where girls had been imported from elsewhere.

There are about 9800 rows and the size of it is approximately 467 KB. It contains data from the year 2001 to 2013.

B. Data Preprocessing:

1) Software Used:

For our preprocessing, we employ Microsoft Excel and Python. This particular software was convenient for use and such easier to understand. The preprocessing stage also used python libraries like pandas and NumPy to figure out the total no. of missing values to get an overview of the dataset and to create correlation heat maps. The implementation of the rest of the project that is analysis in trends etc. has been done using Pivot Tables and Charts provided by Excel.

As we were working with a dataset with a large number of values our attention first and foremost fell to the fact of whether our dataset had any missing values. With the help of pandas, we easily found out whether our dataset had any missing values. We found out that the data set is mostly complete with no null values. The pandas result is provided below:

	STATE	DISTRICT	Year	Rape	Kidnapping_and_Abduction	Dowry_Deaths	Molestation	Sexual_Harassment
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False

Figure 1- Pandas result for missing values

The dataset provided has only the total number of cases with no other information provided. This is not very useful as it doesn't paint a proper picture of the overall rate of crime against women in India. The dataset provided a lot of potential to extract more meaningful information from the existing columns. Hence, a few columns have been added or transformed to improve the result of the following analysis. The decision to add or transform columns has been taken by studying the graphical analysis which has been performed on the data.

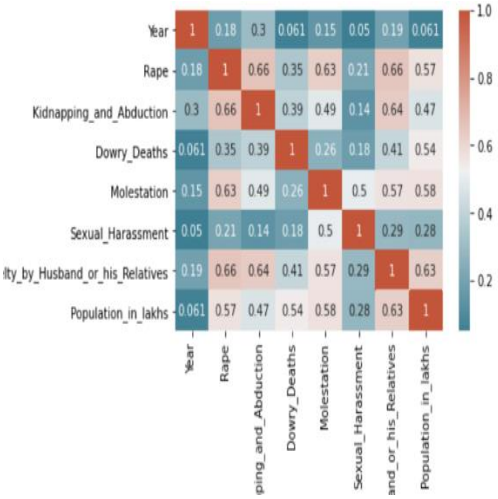


Figure 2- Correlation heat map of the different attributes of the dataset

C. Data Cleaning:

The first step in data cleaning is taking care of incorrect or missing data. The Category column contained a few columns which have been labeled imprecisely, like the ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY category had a long name and was hard to understand. This has been changed to MOLESTATION which provides a concise description of the columns.

There are 11 distinct categories in the data set. However, the category IMPORTATION OF GIRLS consisted mostly of null values and was judged to have almost no consequence to the data, therefore it was dropped.

Original Names	*New Names*
assault on women with intent to outrage her modesty	Molestation
Insult to the modesty of women	Sexual Harassment

Table 1 - Table showing attributes whose names have been changed

D. Data Transformation:

Data transformation is one of the most important data preprocessing techniques. Usually, the data is originally present in the form that makes more sense if it is transformed. In this case, the main transformations performed are as follows:

i. Population of districts were added according to the census taken in 2001 and 2011. Till 2010 the population was taken according to 2001 census after which the values were provided by the 2011 census.
i. The Crime Rate was calculated according to the formula provided in the first section of design for or each district.
ii. Standardization of values was performed so that each category had number of cases per 100,000 persons.
iii. The number of columns was increased to include these values.
iv. Further analysis was performed via these new values and not the original values.

Data Attribute	Value 1	Value 2
STATE	ANDHRA PRADESH	MAHARASHTRA
DISTRICT	ADILABAD	MUMBAI
Year	2001	2005
Rape	50	203
Kidnapping and Abduction	30	127
Dowry Deaths	16	9
Molestation	149	397
Sexual_Harassment	34	103
Cruelty by Husband or his Relatives	175	336
Population_in_lakhs	24.88	86.4
Crime_Rate	18.24758842	13.59953704
Total Crime	454	1175
Rape_rate	2.009646302	2.349537037
Kidnapping_rate	1.205787781	1.469907407
dowry_rate	0.643086817	0.104166667
molestation_rate	5.988745981	4.594907407
sexual_harassment-rate	1.366559486	1.19212963
cruelty_rate	7.033762058	3.888888889

Table 2 - Table showing the different dataset attributes with examples of data

E. Data Reduction

The dataset consisted of 26 states and 7 union territories. The total number of districts were over 358. As previously mentioned, we worked on only 15 states and 3 union territories. The other states and union territories were removed from the dataset. This is was done so that it will easier to manage the huge amount of data. This reduces the number of rows from about 9800 to about 4965.

STATES AND UNION TERRITORIES CHOSEN		
ANDHRA PRADESH	CHHATTISGARH	JAMMU & KASHMIR
ANDMAN NICOBAR ISLANDS	DELHI	JHARKHAND
ARUNACHAL PRADESH	GUJARAT	KARNATAKA
ASSAM	HARYANA	MAHARASHTRA
BIHAR	HIMACHAL PRADESH	PUNJAB
CHANDIGARH	WEST BENGAL	RAJASTHAN

Table 3 – States and union territories chosen for the dataset

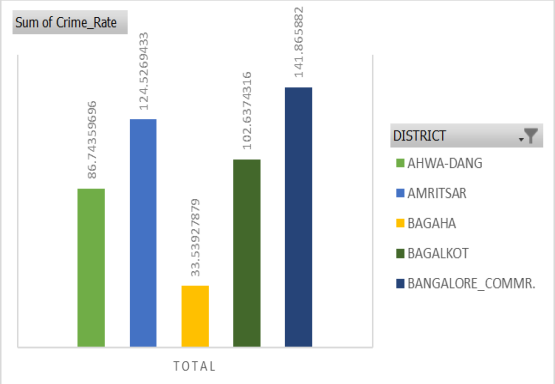


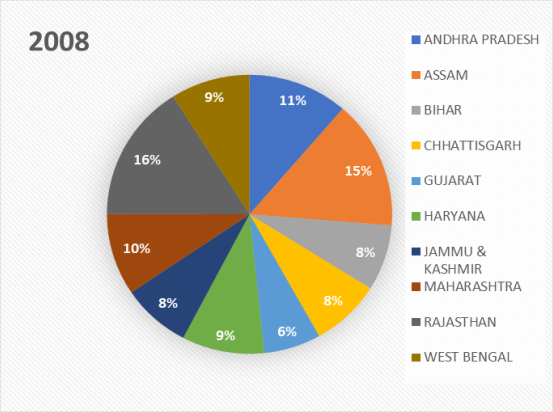
Figure 3 - Chart of a few low crime rate districts with their total sum of crime rate values

The data was further reduced to only include 255 districts in which the sum of crime rate was low when the sum was taken over the period of 13 years. This was necessary as excel can create charts where the maximum number of rows is 255. Below is a representation of the districts in which the sum of crime rate was below 150 cases for over 13 years.
The following data reduction reduced the number of rows in the dataset to 3264 from 4956.
After preprocessing the dataset had 18 attributes and 3264 columns.

IV. Experimental Evaluation

We start by exploring our data. This is a major step in any exploratory analysis of dig data. Charts and graphs give us a unique understanding of our data from an entirely new perspective. Hence data

visualization is a key step in any data analysis process. These graphs show interesting patterns in crime that may not have been apparent otherwise.
Figure 4 shows us the comparison between crime in 10 states in 2008 vs 2013. As we can see the crime



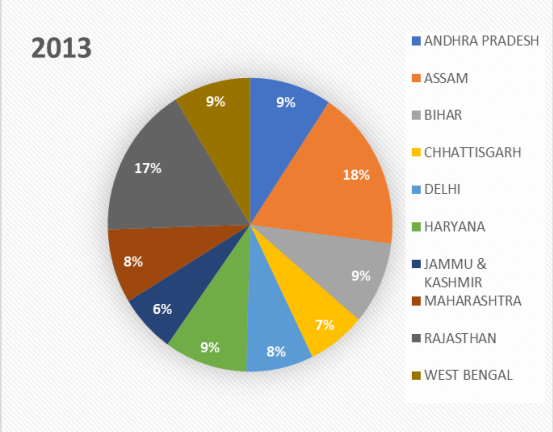


Figure 4 - Comparison between 2008 vs 2013 percentage of crime in 10 states

percentage in states of ASSAM and RAJASTHAN have swapped and according
to the data in 2013 the highest percentage of crime

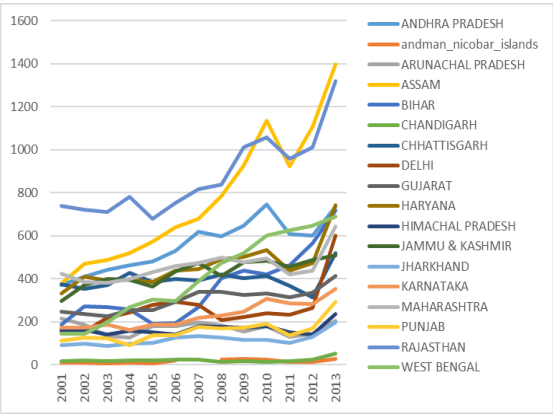


Figure 5 - Graph showing trend in crime over the years in different states

against women took place in ASSAM among the 15 states we have taken into consideration.

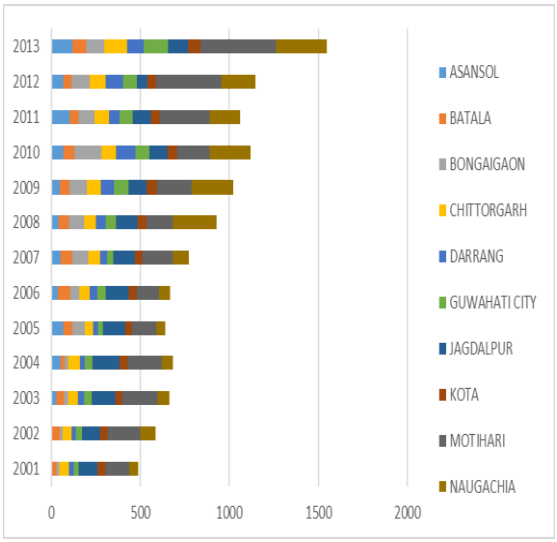


Figure 6 – Chart showing the top ten districts where crime against women is the highest.

Figure 5 shows us the trend in crime across the 15 states over a period of 13 year and we can see that crime against women has been increasing in ASSAM and RAJASTHAN. We have to take the fact into consideration that the maximum number of crimes reported against women have either been ABDUCTION cases or have been CRUELTY CASES.

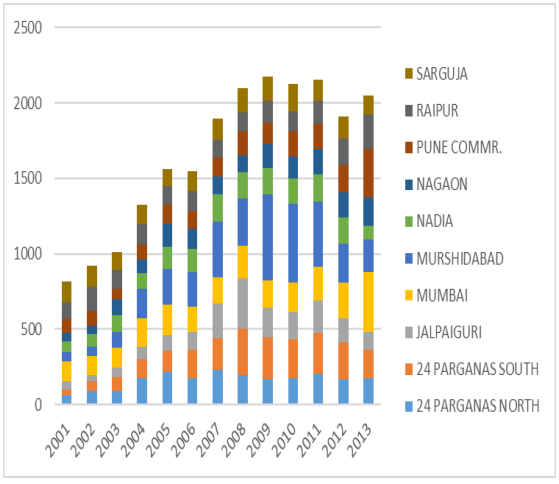


Figure 7 – Comparison between districts where Rape cases are highest

Figure 6 shows the top ten districts where crime against women is maximum. NAUGACHIA and MOTIHARI are at the top every year. This implies that crime against women in these parts have not decreased in over 13 years. The government can thus increase vigilance in these areas over other areas.
Figure 7 shows the districts where sexual violence against women is maximum. Figure 8 shows a similar comparison between states. RAJASTHAN leads in states. But metropolitan cities like MUMBAI, PUNE and KOLKATA also have high values of RAPE. As most of these districts are

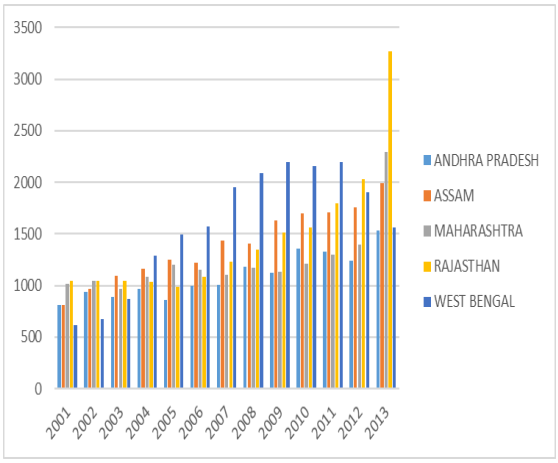


Figure 8 – Comparison between states where rape cases are highest

economically stable therefore the number of cases is high. We should keep in mind that in India a high number of cases are not reported. These districts

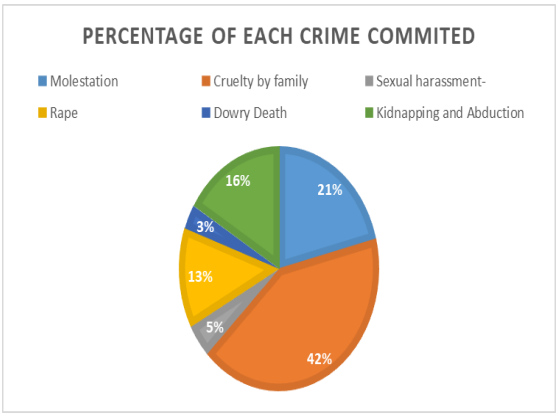


Figure 9 – Percentage of each crime committed

have more literate women who are knowledgeable about the laws protecting them and hence the number of cases may be high because of these reasons.

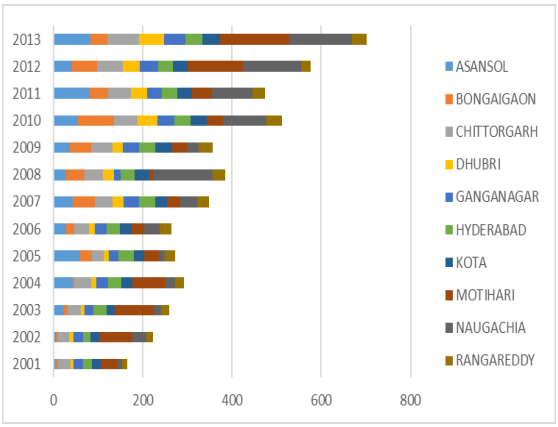


Figure 10 – Top ten districts in cruelty case by husband or relative

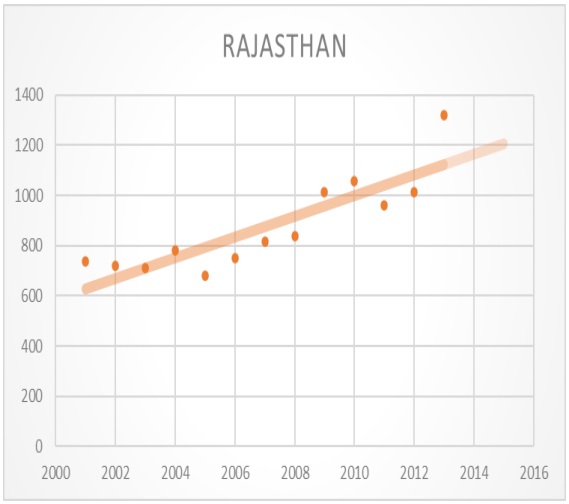


Figure 11 Trend Forecast for RAJASTHAN

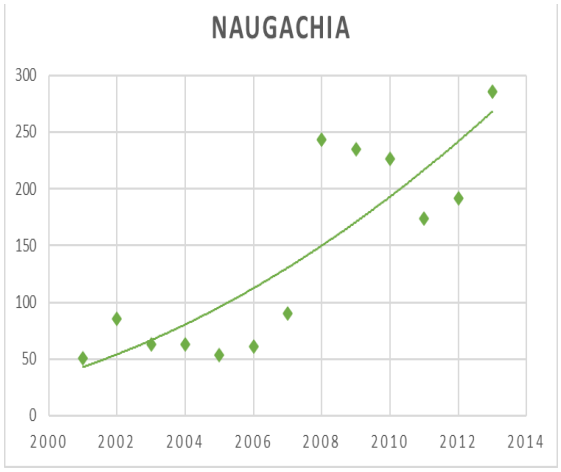


Figure 12 – Trend line for Crime Rate in NAUGACHIA

Figure 9 gives an overview of the percentage of each crime committed. The number of cases of violence against women is highest in the category of CRUELTY BY HUSBAND OR RELATIVES, this implies that in India, domestic violence is the greatest offender and women in such abusive situations are often neglected as society tends to avoid conversation about violence against married women.
Figure 10 shows the ten districts where cruelty cases are the highest. Maximum cases take place in MOTIHARI AND NAUGACHIA

MS EXCEL has also allowed us to trend analysis of crime rate across states as well as districts. Figure 11 shows the linear trend line of crime rate in the state of RAJASTHAN. Figure 12 shows a similar trend line for the district of NAUGACHIA. In both cases the trend is increasing.

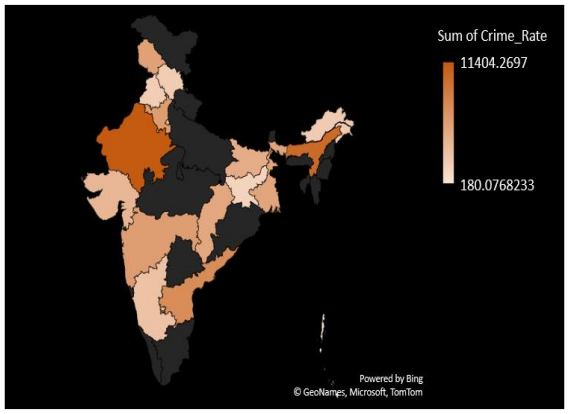


Figure 13 – HEAT MAP of crime-rate against women in India

Figure 13 gives a heat map of India highlighting the number of cases in each compared to other states.

V. CONCLUSION AND FUTURE WORK

In this work, we conducted a detailed analysis of the NBCR crime against women data set consisting of criminal activity over 13 years for the various districts of India. We performed exploratory data analysis and extensive data preprocessing.

As a part of the future work, we plan to build a model to predict the occurrence of crime with the help of temporal and spatial analysis. We also aim to build hotspot detection model, with the help of clustering algorithms. We also plan to enhance this dataset with additional metadata, such as location, date of crime and victim classification to gain more insights on the crime prediction process.

VI. REFERENCES

COHEN, A. (1984). Exploratory Data Analysis Methods: A Study of Industrial Workers' Work Role Centrality. Sociological Methods & Research, 12(4),433–452.

Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston, A comprehensive review of tools for exploratory analysis of tabular industrial datasets,
Visual Informatics,Volume 2, Issue 4,2018,Pages 235-253,ISSN 2468-502X,

Yu, Chong Ho. (2010). Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research. 3. 10.21500/20112084.819.

Cox, Nicholas & Jones, Kelvyn. (1981). Exploratory data analysis. Quantitative Geography, London: Routledge. 135-143.

Pradhan, Isha & Potika, Katerina & Eirinaki, Magdalini & Potikas, Petros. (2019). Exploratory data analysis and crime prediction for smart cities. IDEAS '19: Proceedings of the 23rd International Database Applications & Engineering Symposium. 1-9. 10.1145/3331076.3331114.

Ingilevich, Varvara & Ivanov, Sergey. (2018). Crime rate prediction in the urban environment using social factors. Procedia Computer Science. 136. 472-478. 10.1016/j.procs.2018.08.261.

Dr. Jaison V Joseph & Dr. Jomon Mathew (2019) The relationship between literacy rate and crime rate: An analysis with reference to Kerala IJRAR ijrar_issue_20543147.