MCS 203 PROJECT: FINAL REPORT
'ANALYSIS OF CRIMES AGAINST WOMEN IN INDIA'

PREPARED BY:

1.	Abhishek Sen	(M. Sc. Computer Science, Roll No.: 02)
2.	Anuradha Aggarwal	(M. Sc. Computer Science, Roll No.: 05)
3.	Megha Sundriyal	(M. Sc. Computer Science, Roll No.: 17)
4.	Aanchal Gupta	(M.A. Economics, Roll No.: 27003)

ABSTRACT

This term paper analyses the dataset for crime rates in India for the year 2001-12 and seeks to find trends of crimes in the various stats/UTs.

INDEX

Sections **Page No.**

1. OBJECTIVE

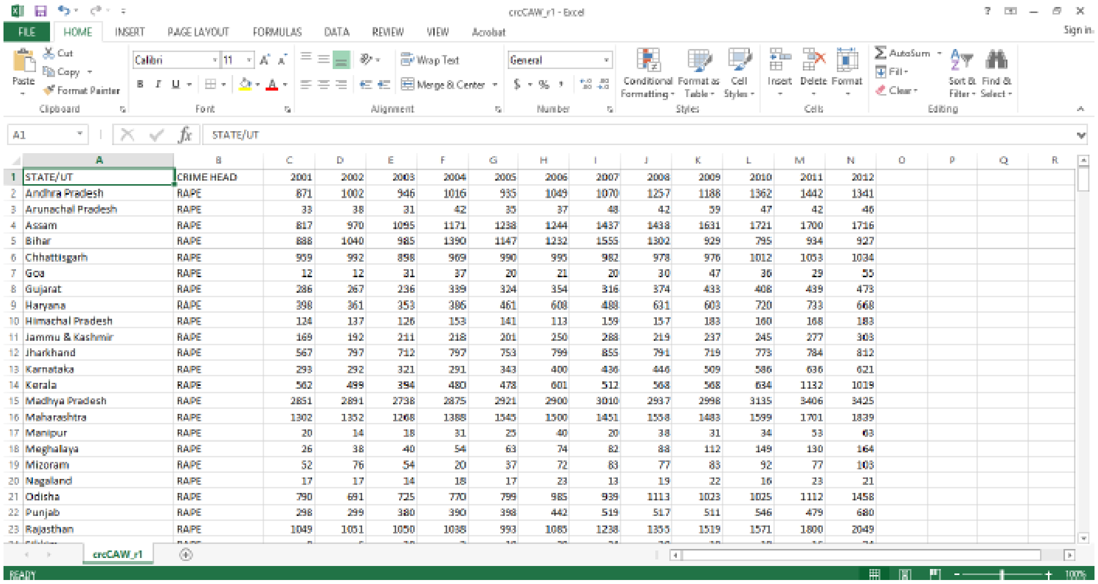
The aim of the project is to analyze the trends of crimes against women across various states and union territories over a twelve year period of 2001-12. The project strives to uncover if there is has been any rise or fall or whether there is some other discernible pattern in the crime rates.

2. DATA

a. Dataset for Crimes against women

The open government data (OGD) platform www.data.gov.in was used to look for relevant datasets. The dataset which was narrowed down is a panel dataset provided by National Crime Records Bureau (NCRB), Ministry of Home Affairs. For a twelve year period, from 2001 to 2012, the dataset has number of crimes committed against women in each year in the 35 States and Union Territories (UTs). There are eight crimes against women which our dataset recognizes.

The data set is organized as follows:



STATE/UT	CRIME HEAD	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Andhra Pradesh	RAPE	871	1002	946	1016	935	1049	1070	1257	1188	1362	1442	1341
Arunachal Pradesh	RAPE	33	38	31	42	35	37	48	42	59	47	42	46
Assam	RAPE	817	970	1095	1171	1238	1244	1437	1438	1631	1721	1700	1716
Bihar	RAPE	888	1040	985	1390	1147	1232	1555	1302	929	795	934	927
Chhattisgarh	RAPE	959	992	898	969	990	995	982	978	976	1012	1053	1034
Goa	RAPE	12	12	31	37	20	21	20	30	47	36	29	55
Gujarat	RAPE	286	267	236	339	324	354	316	374	433	408	439	473
Haryana	RAPE	398	361	353	386	461	608	488	631	603	720	733	668
Himachal Pradesh	RAPE	124	137	126	153	141	113	159	157	183	160	168	183
Jammu & Kashmir	RAPE	169	192	211	218	201	250	288	219	237	245	277	303
Jharkhand	RAPE	567	797	712	797	753	799	855	791	719	773	784	812
Karnataka	RAPE	293	292	321	291	343	400	436	446	509	586	636	621
Kerala	RAPE	562	499	394	480	478	601	512	568	568	634	1132	1019
Madhya Pradesh	RAPE	2851	2891	2738	2875	2921	2900	3010	2937	2998	3135	3406	3425
Maharashtra	RAPE	1302	1352	1268	1388	1545	1500	1451	1558	1483	1599	1701	1839
Manipur	RAPE	20	14	18	31	25	40	20	38	31	34	53	63
Meghalaya	RAPE	26	38	40	54	63	74	82	88	112	149	130	164
Mizoram	RAPE	52	76	54	20	37	72	83	77	83	92	77	103
Nagaland	RAPE	17	17	14	18	17	23	13	19	22	16	23	21
Odisha	RAPE	790	691	725	770	799	985	939	1113	1023	1025	1112	1458
Punjab	RAPE	298	299	380	390	398	442	519	517	511	546	479	680
Rajasthan	RAPE	1049	1051	1050	1038	993	1085	1238	1355	1519	1571	1800	2049

The crimes reported for all the states and UTs in the dataset are as follows:

i. Rape
ii. Assault on women with intent to outrage her modesty
iii. Cruelty by husband or relative
iv. Dowry death
v. Immoral traffic (prevention) act
vi. Indecent representation of women (prevention) act
vii. Insult to the modesty of women
viii. Kidnapping and abduction

The dataset, by virtue of it being extensive in coverage of different kinds of crimes over a long span of time, made a convincing case to be selected for the purpose of analyzing trends.

b. Dataset for population figures from 2001 to 2012

The dataset selected with figures on crimes against women simply provides the absolute numbers of the crimes reported against women. Absolute numbers, although good enough for some visualization exercises, don't yield reasonable results when one has to compare them across objects. For example, comparing the absolute number of crimes committed in Uttar Pradesh should not be compared with the respective figure reported by a state like Arunachal Pradesh. This is because Uttar Pradesh by virtue of being the most populous state is likely to report higher absolute crime rates than a state like Arunachal Pradesh which is sparsely populated. Hence, comparing relative figures would make more sense.

This problem was tackled by weighting crime rates by the respective female-population projection figures of the states and UTs. The population projection figures used for this purpose were taken from the report submitted in 2006 by the technical group on population projections constituted by the National Commission on Population. Resulting figures may be interpreted as 'Uttar Pradesh reports a crime rate of 12.4 per lakh women'.

3. TACKLING THE QUESTIONS OF INTEREST

Our aim, as already outlined in the proposal, is to conduct a data mining exercise with an objective to seek answers to the following questions:

i) What is the year on year trend of crimes for each state?

ii) Does there exist any relation between any two kinds of crimes reported? For example, can we say that the trends of kidnapping and abduction move in a similar manner to trends in trafficking of women?

iii) Can the data be used to predict the crime rates for a state/UT?

The answers to above questions are discussed in the sections which follow. The methodology used for the purpose of analysis is also provided along with brief comments on results.

4. TRENDS IN CRIMES RATES

Before weighting the crime rates for all states and UTs over the entire time period, a small exercise was done on the crime rates of year 2001 alone. The weighted figures of crime rates when plotted according to crimes revealed that the data could in fact be potentially clustered, where the clusters containing the states/UTs according to frequency of crimes committed per lakh women. The graphs are presented in Appendix 1. This analysis was followed up: crime rates for all states/UTs were weighted with the respective projected population of females for the entire period of 12 years viz. 2001-12. After this, K-means clustering algorithm was used to cluster the states.

a. Why K-means?

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. We used this simple yet effective method to cluster our data. We expect states to be clustered according to similar crime rates in that particular year.

b. Finding Optimal Number of Clusters

One of the inputs in the algorithm of K-means is the number of clusters. Theoretically, a cluster comprises of all the data-points that are 'similar' to each other. By implication, optimal number of clusters are obtained when the distance between data-points in a cluster is minimized while the distance between clusters is maximized. Hence, the goal is to find a clustering scheme for which sum of squared errors SSE(C) is minimized. i.e. Find clustering scheme C^* such that,

$$C^* = argmin_C\{SSE(C)\}$$

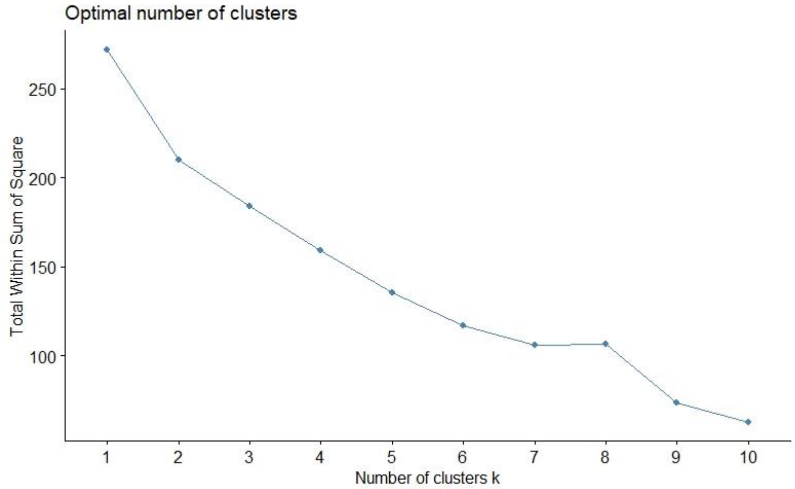
where,

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_{j \in C_i}} \left\| x_j - \mu_i \right\|^2$$

We use two methods of finding the optimal number of clusters for our analysis. They are:

i. Elbow Method

This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters where this 'elbow' occurs gives us the optimal number of clusters. We do this for the crime rates in year 2012. The graph 4.1 below shows the result.

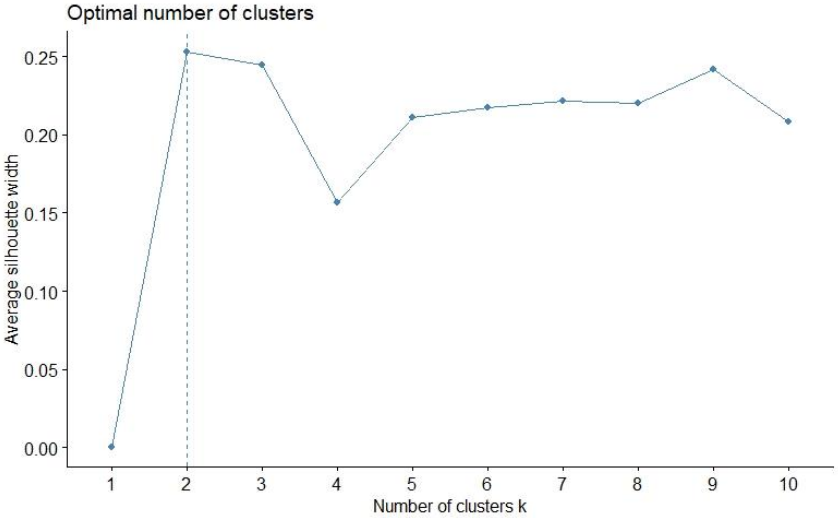


Graph 4.1: Optimal number of clusters using Elbow Method

From the above graph we observe that the ‘break’ seems plausible at two points: k=2 and k=8. To reinforce our results, we use Silhouette method of finding optimal number of clusters. The same has been elucidated below.

ii. Silhouette Method

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. Thus, optimal k is where silhouette value is maximized. We present result for the crime rates of 2012 in the following graph. We see that, like the elbow method also depicted, the optimal value of k comes out to be 2. This is because the silhouette value peaks at k=2.

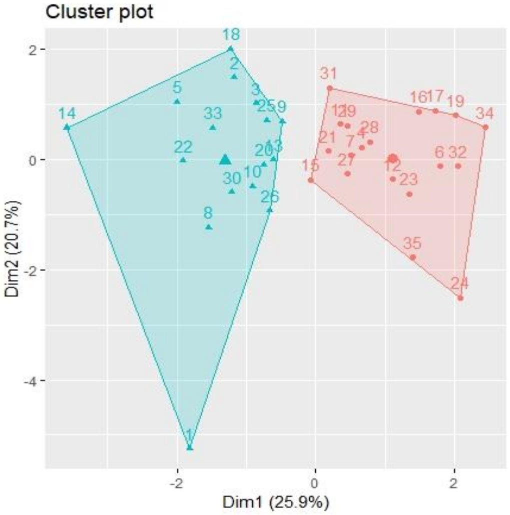


Graph 4.2: Optimal number of clusters using Silhouette Method.

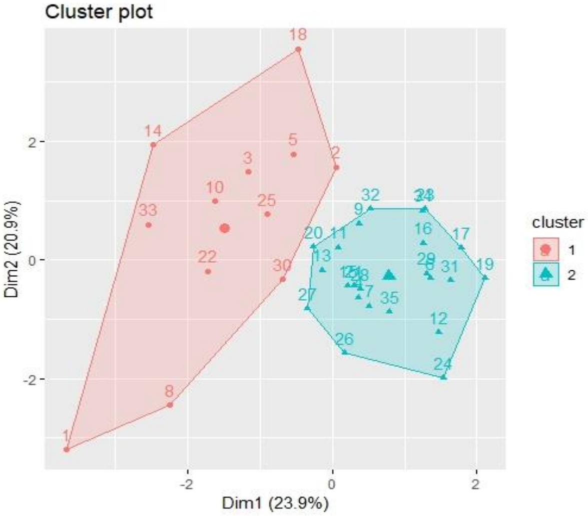
Thus, we do the clustering analysis with k=2.

iii. <u>K-means Clustering: Results</u>

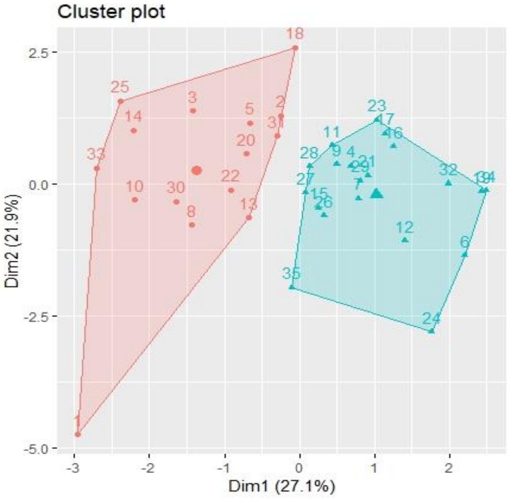
As explained above, use the K-means clustering algorithm with the number of clusters equal to two. Cluster 1 indicates a relatively low level of crime while Cluster 2 indicates a relatively higher level of crimes. The graphs showing clusters are presented below for years 2001, 2003, 2006, 2009, and 2012.



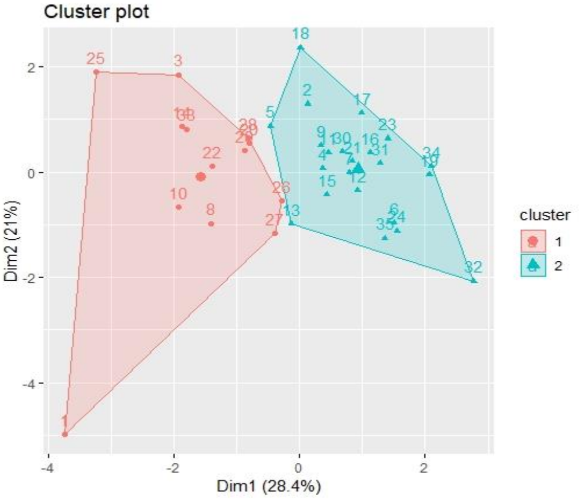
Cluster Plot for 2001



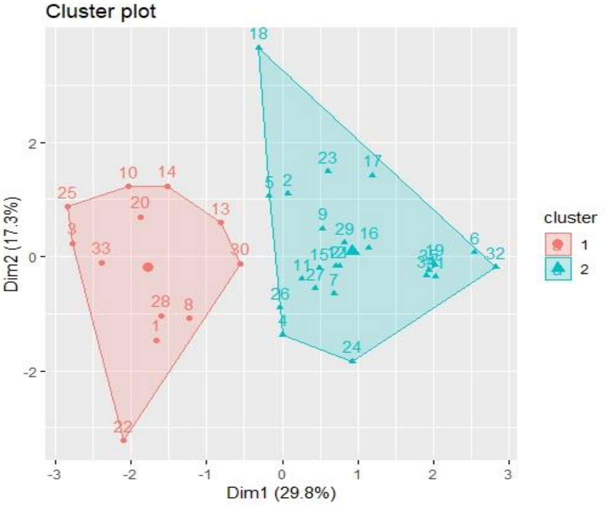
Cluster Plot for 2003



Cluster Plot for 2006



Cluster Plot for 2009



Cluster Plot for 2012

To see the movements of states across clusters over the years, we use the following table

S. No.	States	2001	2003	2006	2009	2012
1	Andhra Pradesh	1	1	1	1	1
2	Arunachal Pradesh	1	1	1	2	2
3	Assam	1	1	1	1	1
4	Bihar	2	2	2	2	2
5	Chhattisgarh	1	1	1	2	2
6	Goa	2	2	2	2	2
7	Gujarat	2	2	2	2	2
8	Haryana	1	1	1	1	1
9	Himachal Pradesh	1	2	2	2	2
10	Jammu & Kashmir	1	1	1	1	1
11	Jharkhand	2	2	2	2	2
12	Karnataka	2	2	2	2	2
13	Kerala	1	2	1	2	1
14	Madhya Pradesh	1	1	1	1	1
15	Maharashtra	2	2	2	2	2
16	Manipur	2	2	2	2	2
17	Meghalaya	2	2	2	2	2
18	Mizoram	1	1	1	2	2
19	Nagaland	2	2	2	2	2
20	Odisha	1	2	1	1	1
21	Punjab	2	2	2	2	2
22	Rajasthan	1	1	1	1	1
23	Sikkim	2	2	2	2	2
24	Tamil Nadu	2	2	2	2	2
25	Tripura	1	1	1	1	1
26	Uttar Pradesh	1	2	2	1	2
27	Uttarakhand	2	2	2	1	2
28	West Bengal	2	2	2	1	1
29	A&N Islands	2	2	2	1	2
30	Chandigarh	1	1	1	2	1
31	D&N Haveli	2	2	1	2	2
32	Daman & Diu	2	2	2	2	2
33	Delhi UT	1	1	1	1	1
34	Lakshadweep	2	2	2	2	2
35	Puducherry	2	2	2	2	2

Table 4.1: Movements of States across clusters over the years.

From the table we observe that there are certain states/UTs which have consistently belonged in the cluster which denoted higher level of crime rates across all years. These states are: Bihar, Goa, Gujarat, Jharkhand, Karnataka, Maharshtra, Manipur, Meghalaya, Nagaland, Punjab, Sikkim, Tamil Nadu, Daman and Diu and Pondicherry. Some states/UTs have consistently performed well and have belonged to the cluster which denoted lower levels of crime rates. These states are: Andhra Pradesh, Assam, Jammu & Kashmir, Madhya Pradesh, Tripura and Delhi. These results may seem surprising, however, one must keep in mind that we are doing relative comparison. Lower level of crimes for many states/UTs may itself be high in absolute terms. While the consistent performers are quite many, there are states which have improved such as West Bengal while many states and UTs have regressed from cluster 1 (low level of crime) to cluster 2 (high level of crime). These are: Chhattisgarh, Mizoram and Uttar Pradesh.

5. RELATION BETWEEN VARIOUS CRIMES

ii. Correlation between the crimes

A-priori we expect certain crimes to be correlated. For example, we expect the instances of rapes to be correlated with instances of crimes reported under 'assault on women with intent to outrage her modesty'. We make use of Spearman's rank correlation coefficient between the different crimes reported. To ease the presentation of tables, we use the following labels for crimes:

Crime	Label
Rape	C1
Kidnapping & Abduction	C2
Dowry Death	C3
Assault On Women With Intent To Outrage Her Modesty	C4
Insult To The Modesty Of Women	C5
Cruelty By Husband Or Relatives	C6
Immoral Traffic(Prevention)Act	C7
Indecent Representation Of Women(Prevention)Act	C8

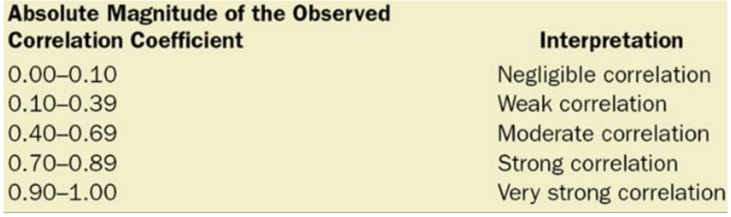
Table 5.1: Labels for the crimes used in tables

The Spearman's Rank Correlation Coefficient is found out for the eight crimes labeled as C1, C2,..., C8. The results of correlation coefficients are provided in the table below:

	C1	C2	C3	C4	C5	C6	C7	C8
C1	1.00	0.90	0.85	0.74	0.55	0.81	0.84	0.92
C2	0.90	1.00	0.88	0.79	0.58	0.76	0.90	0.90
C3	0.85	0.88	1.00	0.72	0.46	0.75	0.86	0.90
C4	0.74	0.79	0.72	1.00	0.60	0.74	0.71	0.68
C5	0.55	0.58	0.46	0.60	1.00	0.44	0.48	0.50
C6	0.81	0.76	0.75	0.74	0.44	1.00	0.71	0.71
C7	0.84	0.90	0.86	0.71	0.48	0.71	1.00	0.88
C8	0.92	0.90	0.90	0.68	0.50	0.71	0.88	1.00

Table 5.2: Spearman's Rank Correlation between various crimes

It is noteworthy that if we use the thumb-rule outlined in the following table:



Absolute Magnitude of the Observed Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

Table 5.3: Thumb-rule for strength of correlation.

It can be concluded that there exists a 'strong correlation' in many of the pairwise crimes considered, hence reaffirming our a-priori expectations. The values highlighted in table 5.2 denote high correlation between the respective crimes. For example, Rape and Kidnapping & Abduction are very strongly correlated with the rank correlation coefficient equal to 0.9! Other values can be interpreted likewise.

iii. Association Rule Mining for Crimes

Rules generation can be done by association rule mining with the help of support and confidence. If there is an expression in the form of X U Y, where X and Y are disjoint datasets, then Support determines how often a rule is applicable to a given dataset and Confidence determines how frequently items in Y appear in transactions that contain X. We have:

Support, S(X UY) = support (X U Y) / N

Confidence, C(X U Y) = support (X U Y) / support (X)

Association rule mining is a technique to identify underlying relations between different items. Usually, there is also a pattern in how the crimes take place. For instance, states that reported a high number of indecent behavior against women, also reported high number of assault cases. In short, crime reported involve a pattern.

Assuming Confidence threshold to be 70%, following association rules are found:

{indecent}-> {assault},
{indecent} -> {cruelty},
{cruelty, indecent} -> {assault},
{assault, indecent} -> {cruelty},
{indecent} -> {assault, cruelty},
{assault, dowry} -> {rape},
{assault, kidnap} -> {rape},
{dowry, insult} -> {cruelty},
{cruelty, insult} -> {dowry},
{insult, rape} -> {cruelty},
{cruelty, kidnap} -> {rape},
{insult, rape} -> {dowry}]

These can we interpreted as:

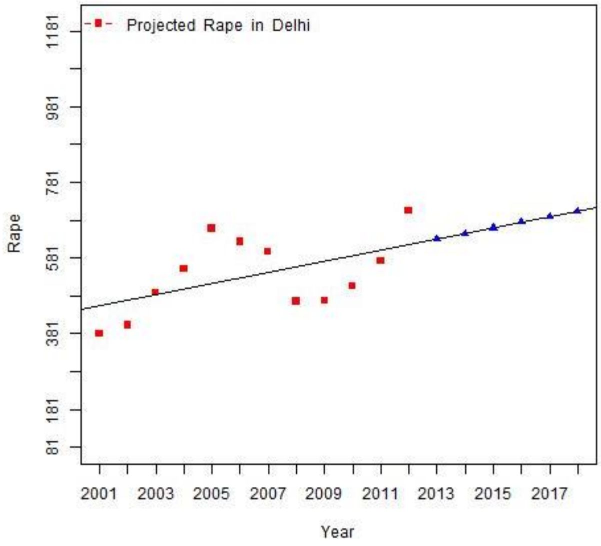
1. If indecent then assault
2. if indecent then cruelty
3. if cruelty and indecent then assault

and so on...

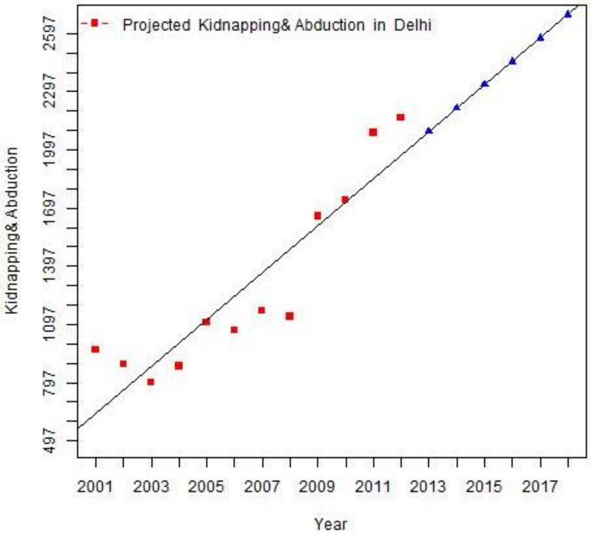
Using these patterns, it can be easily seen that certain crimes are associated and how preventing one can reduce the number of cases of other associated crime. For instance, if crime A and B occur together more frequently then several steps can be taken to decrease the number of cases of A and by consequence B.

6. EXTENSION: PREDICTED CRIMES FOR DELHI FOR 2013-18

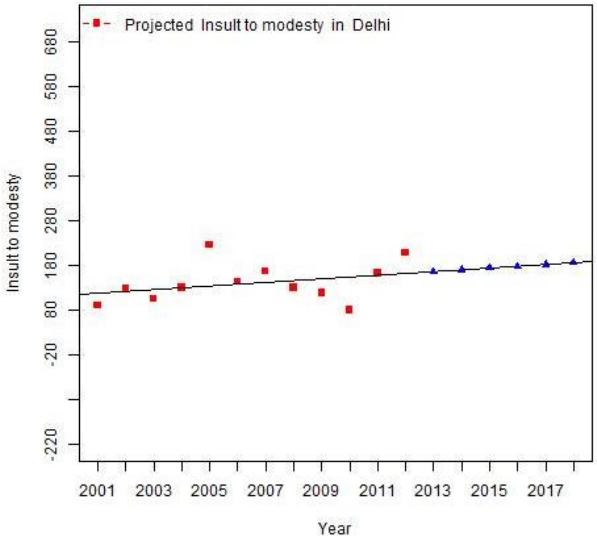
As an extension to our analysis, we use simple linear regression to predict the crime rates for Delhi for the time period of 2013-18. We observe an increasing trend in almost all the crimes for the five year period under consideration. Projected Kidnapping and Abduction, for example, are predicted to increase at a steep rate. A similar, steep trend is also observed for the crime cruelty by husbands and relatives. For crimes such as indecent representation of women the trend line and predicted values lie on a flat line, while immoral trafficking is expected to go down. The total crime rates are predicted to shoot up in Delhi during 2013-18. The graphs presented below can be referred for further scrutiny.

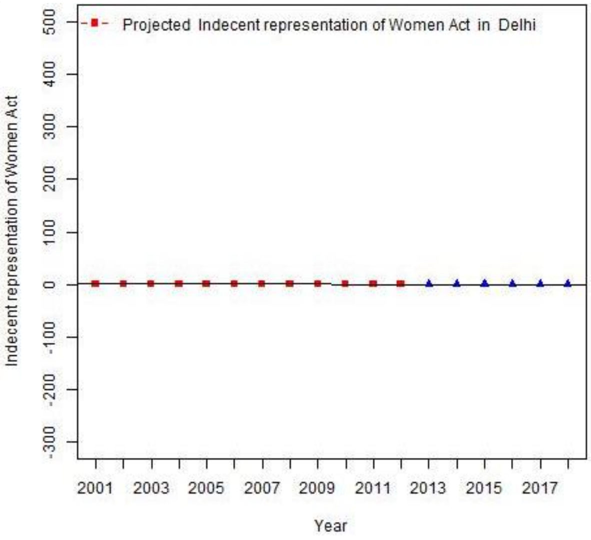


Predicted values of Rape for Delhi show an increasing trend over the years 2013-18



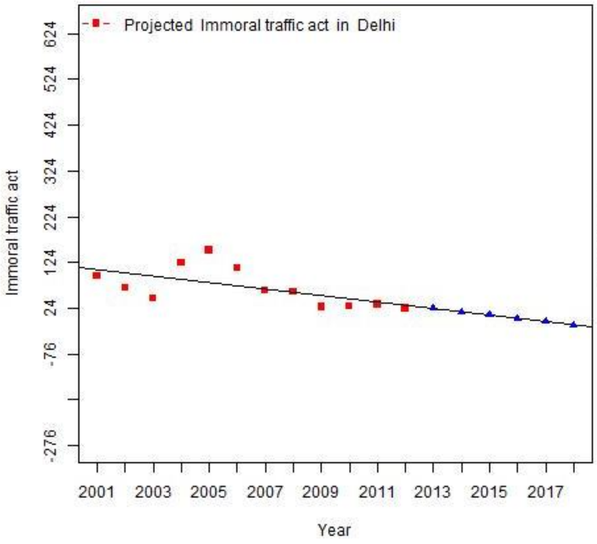
Predicted values of Kidnapping and Abduction for Delhi show an, steep & increasing trend over the years 2013-18

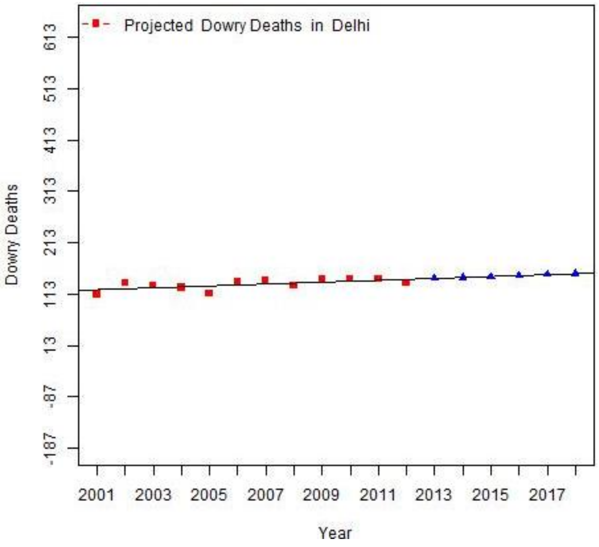




Predicted values of Insult to modesty of women for Delhi show a slightly increasing trend over the years 2013-18

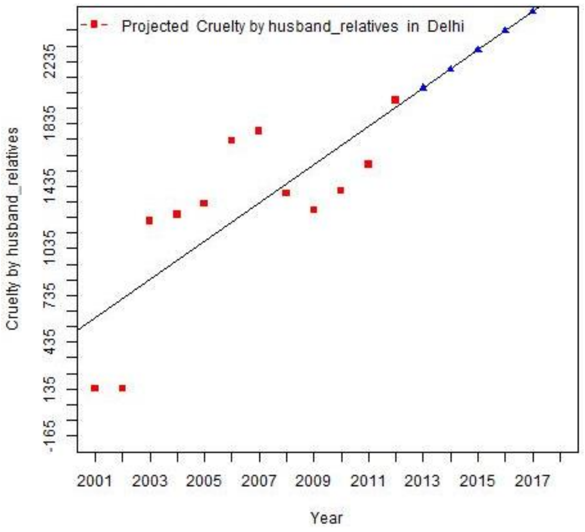
Predicted values of Indecent Representation of Women for Delhi show an, steep & increasing trend over the years 2013-18



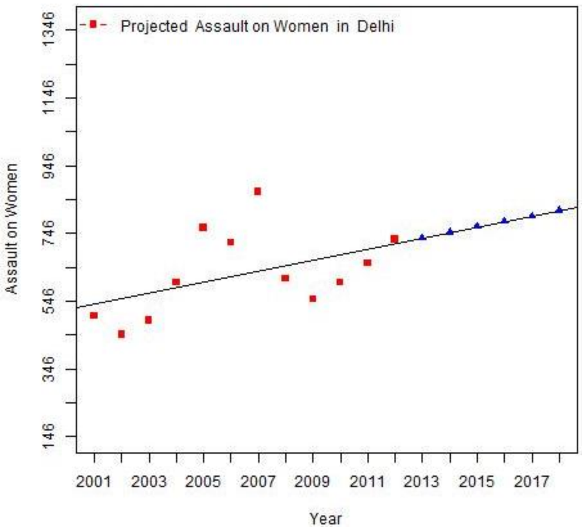


Predicted values of Immoral Trafficking of Women for Delhi show a declining trend over the years 2013-18

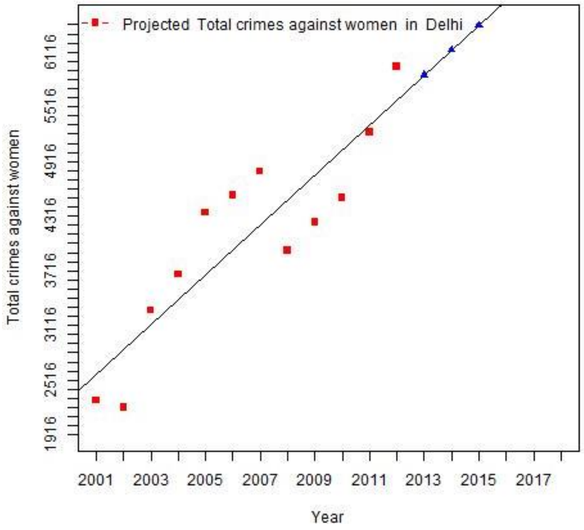
Predicted values of Dowry Deaths of Women for Delhi show a slightly increasing trend over the years 2013-18



Predicted values of Cruelty on women by husband/relatives in Delhi show a declining trend over the years 2013-18



Predicted values of Assault on Women for Delhi show an increasing trend over the years 2013-18



Predicted values of overall crimes against women for Delhi show a steep and increasing trend over the years 2013-18

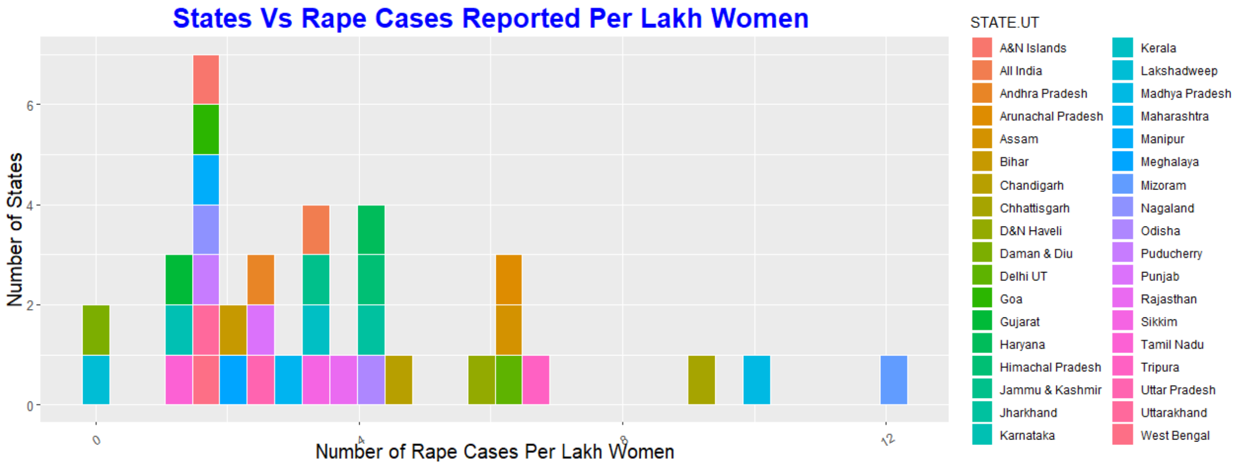
7. CONCLUSION

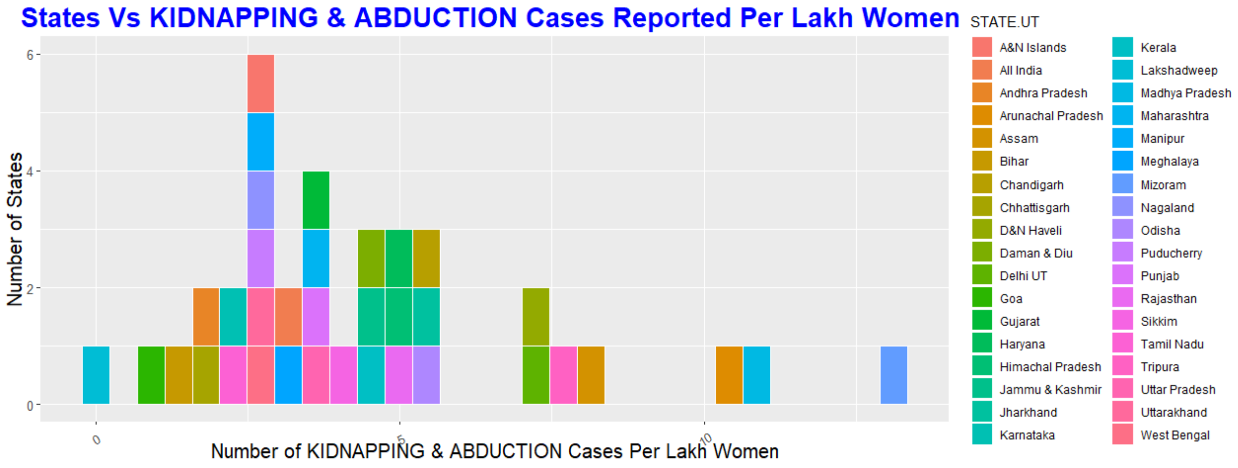
The main aim of this project was to analyze the trends in crime rates for various states and UTs of India for 2001-12. The results show that certain states have improved the pattern of crime rates but for most states the pattern seems to be an upward sloping, steep graph. This is an alarming issue for us as a society as women constitute 50% of the society we live in. Further, addressing the fact that crimes may be correlated, and that making efforts in reducing one crime may implicitly cause another crime to decline, we used Apriori Algorithm with association rule mining to show the probability of relationships between different crime heads. We find evidence that there in fact exists a significant relationship between certain crimes. Efforts in reducing a typical crime may be leveraged to reduce other related crimes as well. Along with the present scope of our project, which is analysis of the crime against women in India, we also predicted the crime rate for Delhi for 2013-18. The results show us that most of the crimes are expected to increase in Delhi in the years under consideration.

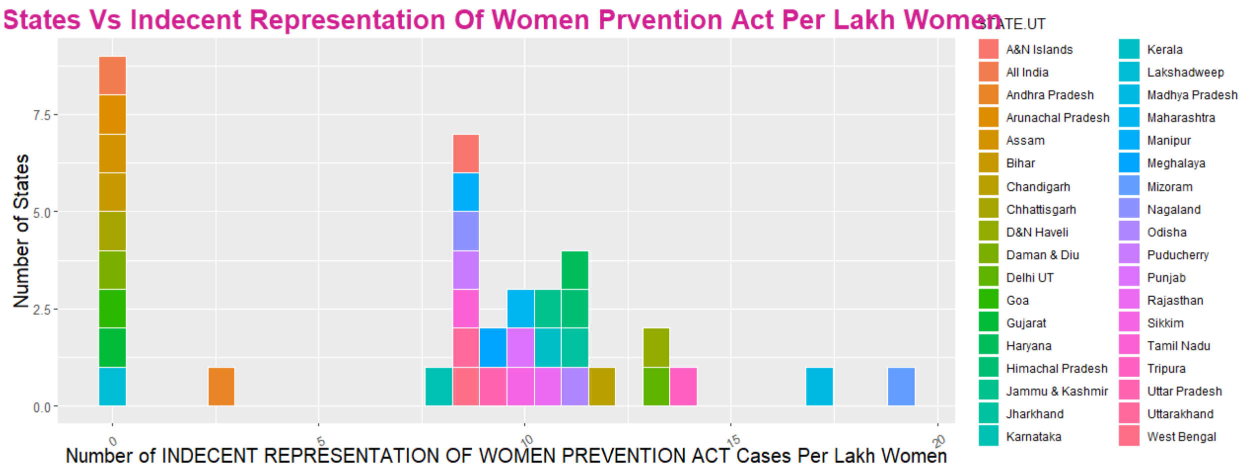
APPENDIX

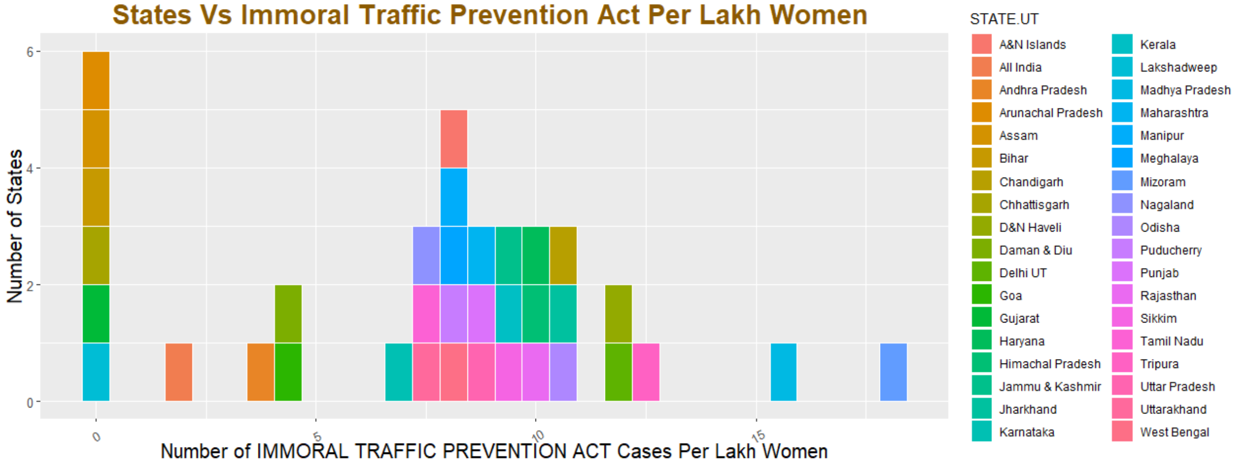
APPENDIX 1:

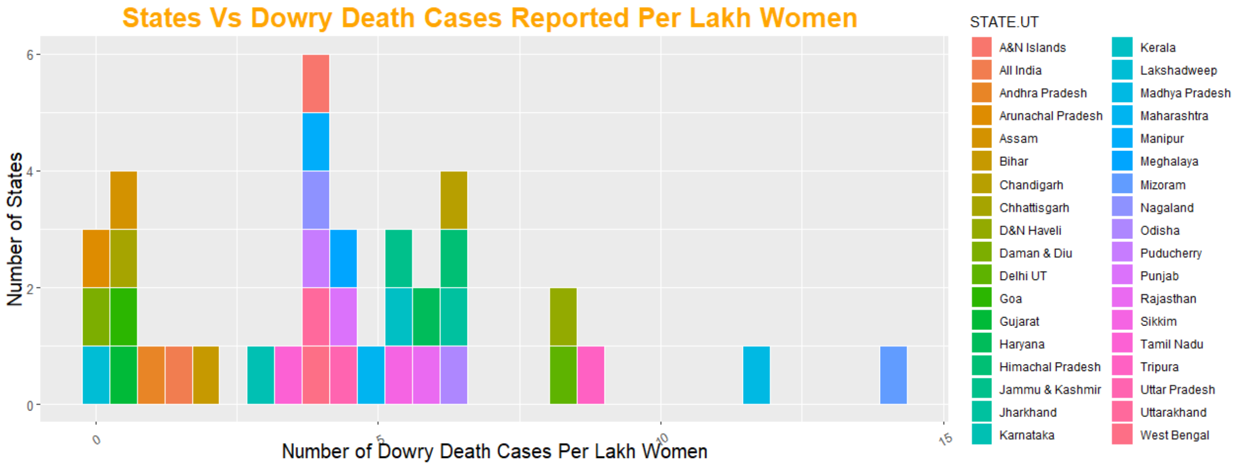
The following six graphs in this appendix were plotted to see the instances of crimes per lakh women for each state/UT in the year 2001.

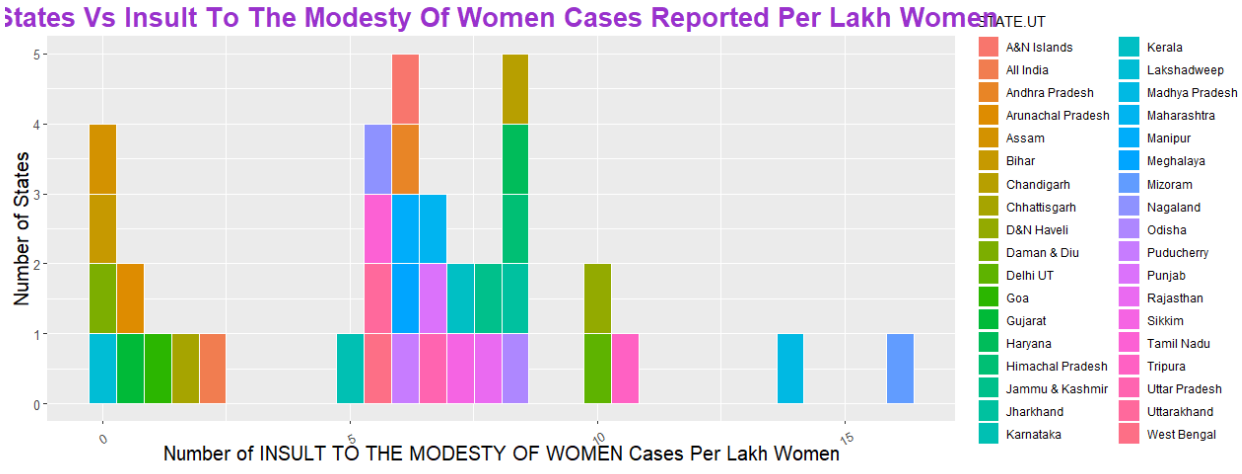


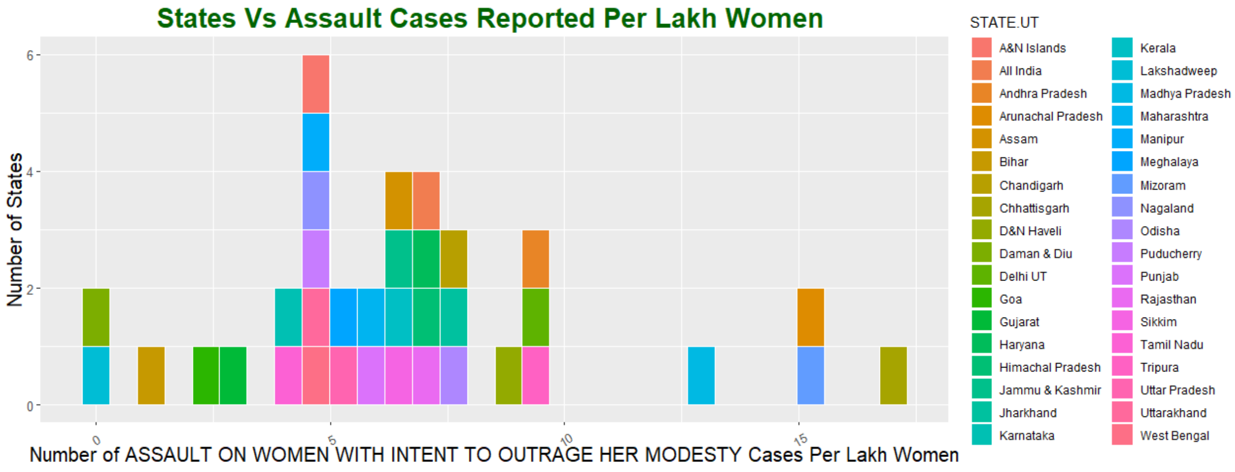


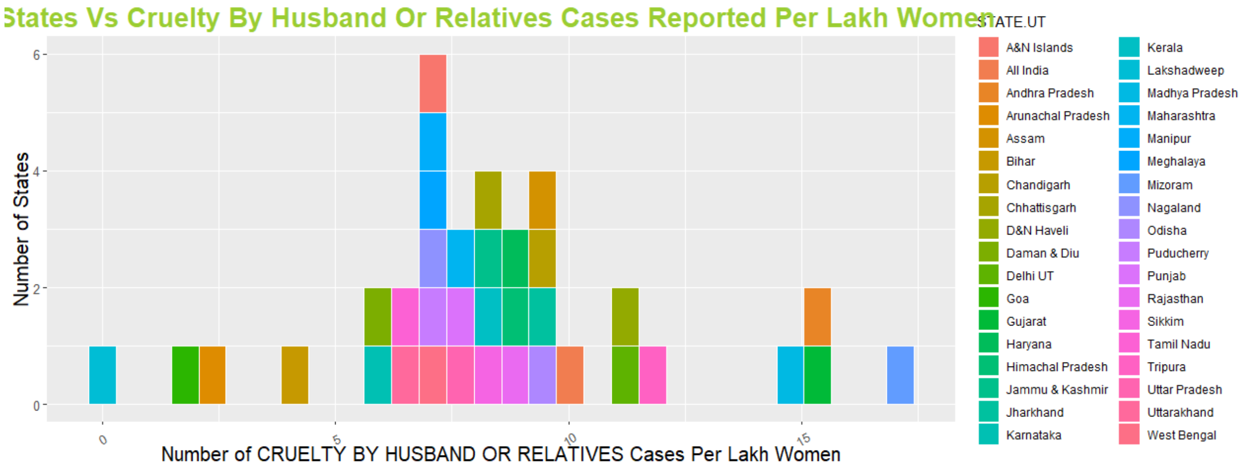

States Vs Indecent Representation Of Women Prvention Act Per Lakh Women
Number of States
0.0
2.5
5.0
7.5
0
5
10
15
20
Number of INDECENT REPRESENTATION OF WOMEN PREVENTION ACT Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal


States Vs Immoral Traffic Prevention Act Per Lakh Women
Number of States
0
2
4
6
0
5
10
15
Number of IMMORAL TRAFFIC PREVENTION ACT Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal


States Vs Dowry Death Cases Reported Per Lakh Women
Number of States
0
2
4
6
0
5
10
15
Number of Dowry Death Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal


States Vs Insult To The Modesty Of Women Cases Reported Per Lakh Women
Number of States
0
1
2
3
4
5
0
5
10
15
Number of INSULT TO THE MODESTY OF WOMEN Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal


States Vs Assault Cases Reported Per Lakh Women
Number of States
0
2
4
6
0
5
10
15
Number of ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal


States Vs Cruelty By Husband Or Relatives Cases Reported Per Lakh Women
Number of States
0
2
4
6
0
5
10
15
Number of CRUELTY BY HUSBAND OR RELATIVES Cases Per Lakh Women
STATE.UT
A&N Islands
All India
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
D&N Haveli
Daman & Diu
Delhi UT
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Tripura
Uttar Pradesh
Uttarakhand
West Bengal