

EXPLORATORY DATA ANALYSIS

‘ANALYSIS OF CRIMES AGAINST WOMEN IN INDIA’

OBJECTIVE

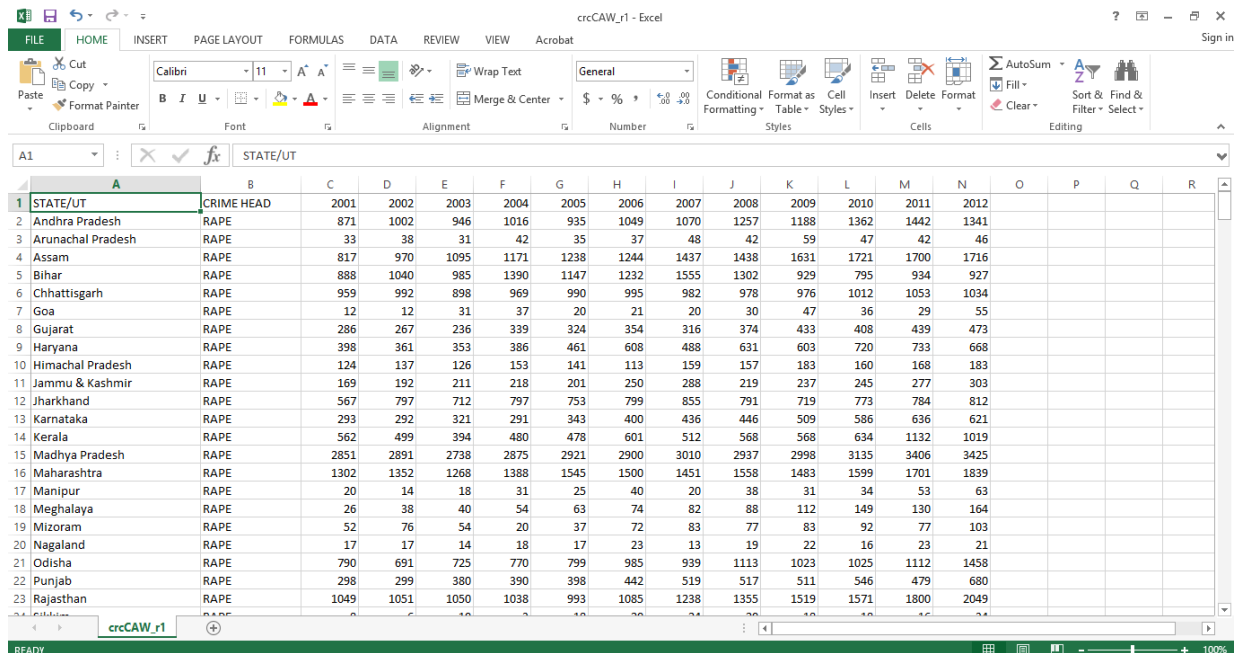
The aim of the project is to analyze the trends of crimes against women across various states and union territories over a twelve year period of 2001-12. The project strives to uncover if there is has been any secular rise or fall or whether there is some other discernible pattern in the crime rates. At the same time, efforts will be made to identify what underlies such patterns.

DATA

1. 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 dataset 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	401	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 are as follows:

- Rape
- Assault on women with intent to outrage her modesty
- Cruelty by husband or relative
- 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.

2. 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 can be tackled if we use crime rates weighted by the respective female-population figures of the states. This would result in figures which may be interpreted as 'Uttar Pradesh reports a crime rate of 1.2 per hundred women' in case we convert them into percentages, and so on.

Thus, it was imperative to find female population estimates to make any meaningful comparisons. To get the reliable population figures, it is widely accepted to refer to the census figures. However, one must keep in mind that census is an extensive exercise which happens once every ten years. Therefore, what one has is the census figures for the year 2001 and 2011. For the rest of the years for the time period under consideration, the census figures are not there. This problem can be resolved in either of the two following ways:-

i) Interpolation: The method of linear interpolation can be used with the assumption of constant rate of change between the known census figures. The open government data source website was used to look up the required 2001 and 2011 census figures.

ii) Population Projection Estimates: Alternately, one can even look at the projected figures of population on the basis of 2001 Census figures. The report submitted in 2006 by the technical group on population projections constituted by the National Commission on Population has tables on the projected population estimates for the years 2001-26. The Appendix 2 in the report outlines output tables of which Table 8 (Projected total population by sex as on 1st March, 2001-2026: India, States and Union territories) was used to extract female population of the 35 states and UTs for the years 2001-12. The difference between the projected and actual figures in 2011 stood at around 1.83 crores for the country. Considering we shall be using disaggregated figures for the states, the difference in the actual and projected population figures is not expected to be much.

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) Relation between trends of crimes with trends in various variable of interest like literacy rate, economic growth rate, unemployment rate, population growth rate etc.
- iii) 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?

The following steps have been followed in pursuit of our quest so far:

1) Data Pre-Processing

The dataset with crime rates against women was checked for missing values and none were found.

Although the data set also reported figures of total crime rate in each state and each year. We checked whether the given sum total of crimes in the dataset adds up according to the crime rates reported under the eight crimes in each state. The sum of crimes reported across states in any particular year was observed to diverge from the reported disaggregated crimes. This divergence was corrected for by replacing the given total crime rate in dataset by calculating the total sum of crimes reported in each year.

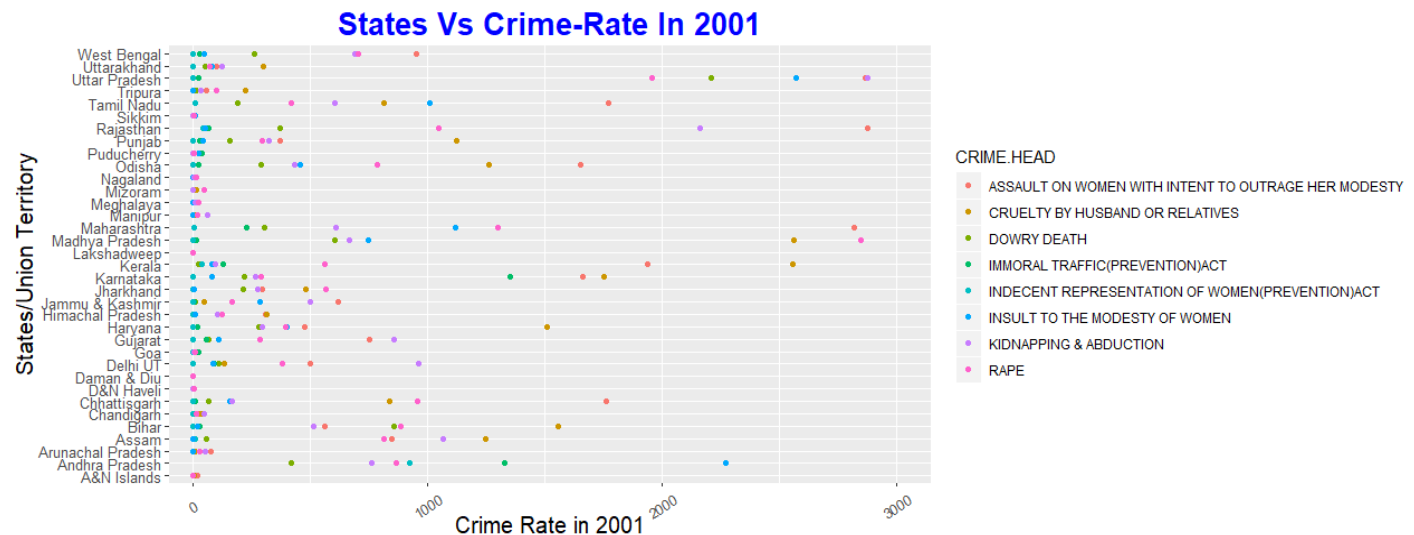
One other thing which was done was to convert the absolute numbers to the relative numbers based on the reasoning provided in the previous section. This was, however, only done for the year 2001 to see the quality and eminence of our results. The final numbers were obtained by calculating the absolute crime rates as a proportion of the female population projected figures obtained from the report on population projections based on 2001 Census. The proportion was then multiplied by 100,000 to make the numbers readable. They are, thus, to be interpreted as crimes committed per lakh of women in a given state for particular year in consideration. The crime dataset for the year 2001 after weighting by female population figures looks like:

2001 Data - Excel									
2001 Data - Excel									
STATE/UT	RAPE	KIDNAPPING & ABDUCTION	DOWRY DEATH	ASSAULT ON WOMEN	INSULT TO THE MODESTY	CRUELTY BY HUSBAND	IMMORAL TRAFFIC(PRE INDECENT REPRESENTATION OF WOMEN)		
Andhra Pradesh	2.3113871	2.030093145	1.114560943	9.404771382	6.02659024	15.36767243	3.534750418		2.45468779
Arunachal Pradesh	6.37065637	10.61776062	0	15.05791506	0.579150579	2.123552124	0		0
Assam	6.34415282	8.308743594	0.458145675	6.600403789	0.031060724	9.690945799	0.046591086		0.077651809
Bihar	2.2336813	1.302980757	2.160734499	1.413658659	0.052823544	3.919003899	0.072946799		0.007546221
Chhattisgarh	9.25675676	1.650579151	0.675675676	17.01737452	1.554054054	8.108108108	0.115830116		0
Goa	1.81818182	0.909090909	0.303030303	2.575757576	1.060606061	1.666666667	4.242424242		0
Gujarat	1.1776817	3.528927321	0.275890467	3.113032736	0.457072267	15.09985588	0.251183858		0
Haryana	4.06911359	5.069113588	6.069113588	7.069113588	8.069113588	9.069113588	10.06911359		11.06911359
Himachal Pradesh	4.14715719	5.147157191	6.147157191	7.147157191	8.147157191	9.147157191	10.14715719		11.14715719
Jammu & Kashmir	3.533347272	5.533347272	6.533347272	7.533347272	8.533347272	9.533347272	10.53334727		11.53334727
Jharkhand	4.34116836	5.341168364	6.341168364	7.341168364	8.341168364	9.341168364	10.34116836		11.34116836
Karnataka	1.1290074	2.129007398	3.129007398	4.129007398	5.129007398	6.129007398	7.129007398		8.129007398
Kerala	3.4324803	4.432480303	5.432480303	6.432480303	7.432480303	8.432480303	9.432480303		10.4324803
Madhya Pradesh	9.86368669	10.86368669	11.86368669	12.86368669	13.86368669	14.86368669	15.86368669		16.86368669
Maharashtra	2.801325358	3.801325358	4.801325358	5.801325358	6.801325358	7.801325358	8.801325358		9.801325358
Manipur	1.867413632	2.867413632	3.867413632	4.867413632	5.867413632	6.867413632	7.867413632		8.867413632
Meghalaya	2.274715661	3.274715661	4.274715661	5.274715661	6.274715661	7.274715661	8.274715661		9.274715661
Mizoram	12.12121212	13.12121212	14.12121212	15.12121212	16.12121212	17.12121212	18.12121212		19.12121212
Nagaland	1.80275716	2.802757158	3.802757158	4.802757158	5.802757158	6.802757158	7.802757158		8.802757158
Odisha	4.35405644	5.354056437	6.354056437	7.354056437	8.354056437	9.354056437	10.35405644		11.35405644
Punjab	2.62001055	3.62001055	4.62001055	5.62001055	6.62001055	7.62001055	8.62001055		9.62001055
Rajasthan	3.87270646	4.872706464	5.872706464	6.872706464	7.872706464	8.872706464	9.872706464		10.87270646

2) Visualization

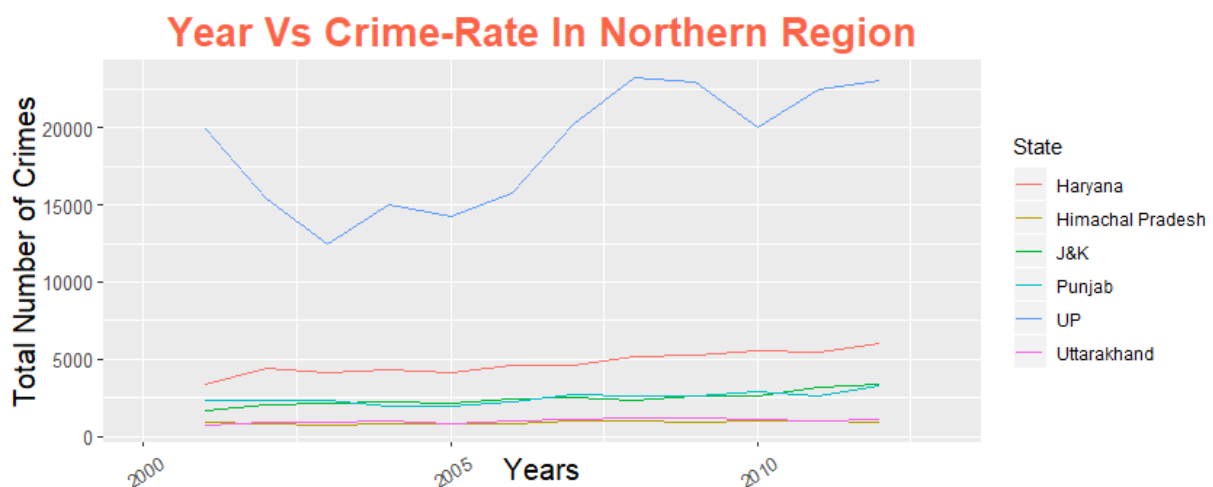
After processing our data, we carried out some basic visualization exercises.

- I. First off, the raw dataset with corrected total crimes reported figures was used to make a simple scatter plot for the year 2001.



- II. We then used the same dataset and plotted line graphs for states for the year 2001-2012. This was done with the objective of gauging the broad trend of total crimes in each state. For the ease of comprehension, states and UTs were divided into 7 regional zones. The results obtained are as follows:

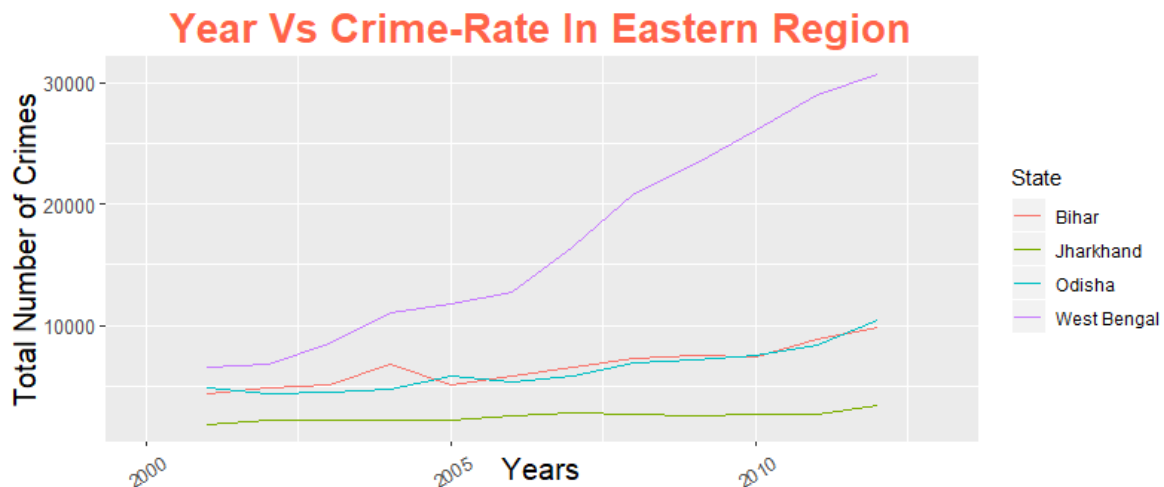
1) North Region



As can be seen above, the overall crime rate trends for the states is increasing in nature. However the trend for Uttar Pradesh (blue line), which is neither increasing nor decreasing for

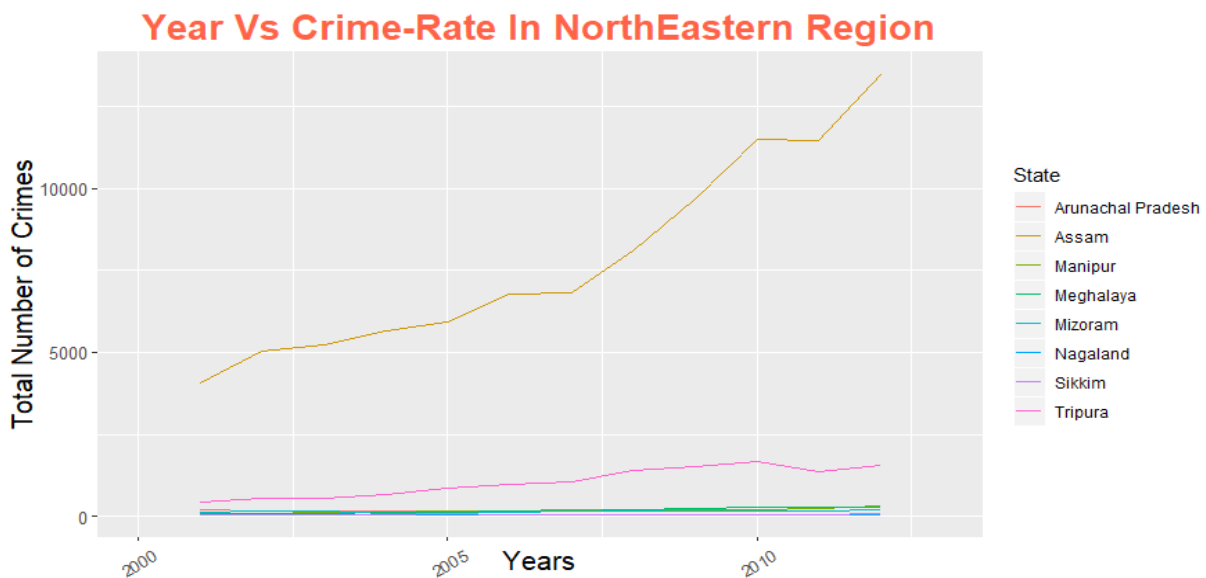
the entire twelve year period, cannot be compared with that of other states. This is because of the absolute number argument outlined in the data section above.

2) East Region



Here also, the crime rates for each of the state individually shows increasing trend. However, we cannot comment on the relative steepness and magnitude of the crimes across the states because the population numbers might differ and would account for some of the divergence/steepness.

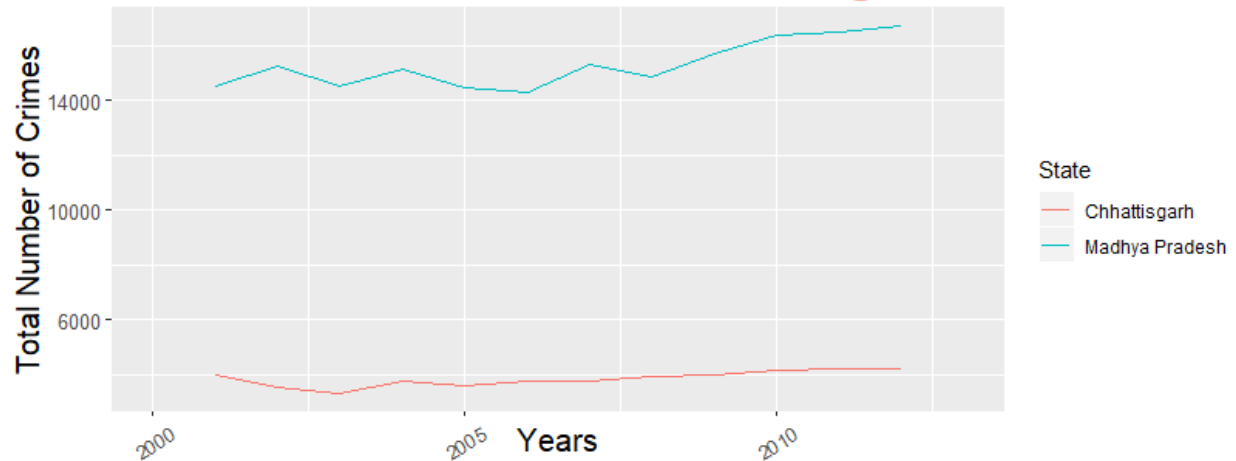
3) North-East Zone



Assam seems to be reporting exponentially high crime rate trends as depicted in the line graph above. However, this trend may be discounted by virtue of Assam's relatively high population with respect to other north-eastern states.

4) Central Region

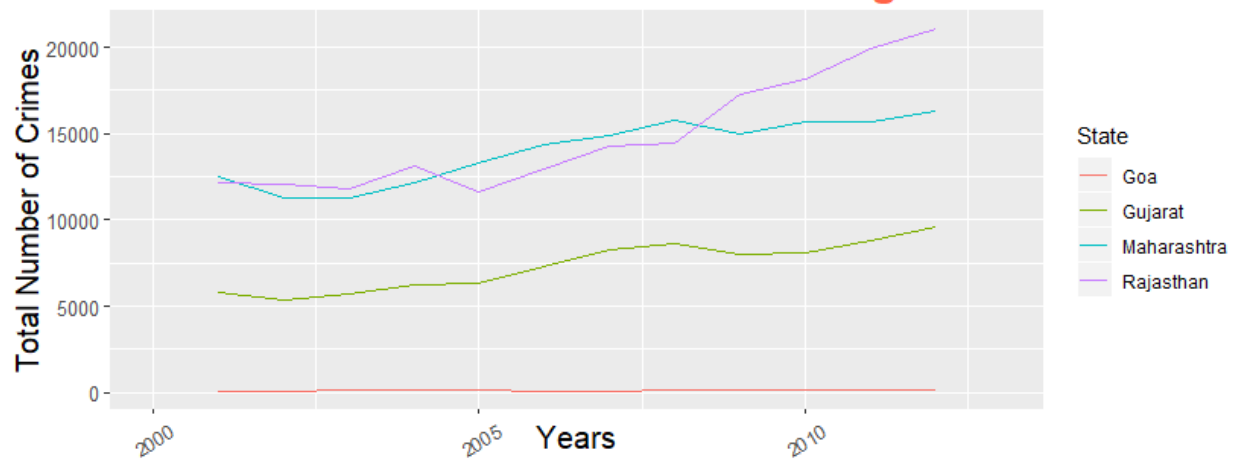
Year Vs Crime-Rate In Central Region



In the two states that comprise the central region, we see the same increasing pattern in the absolute numbers of crimes committed against women over the twelve period under consideration.

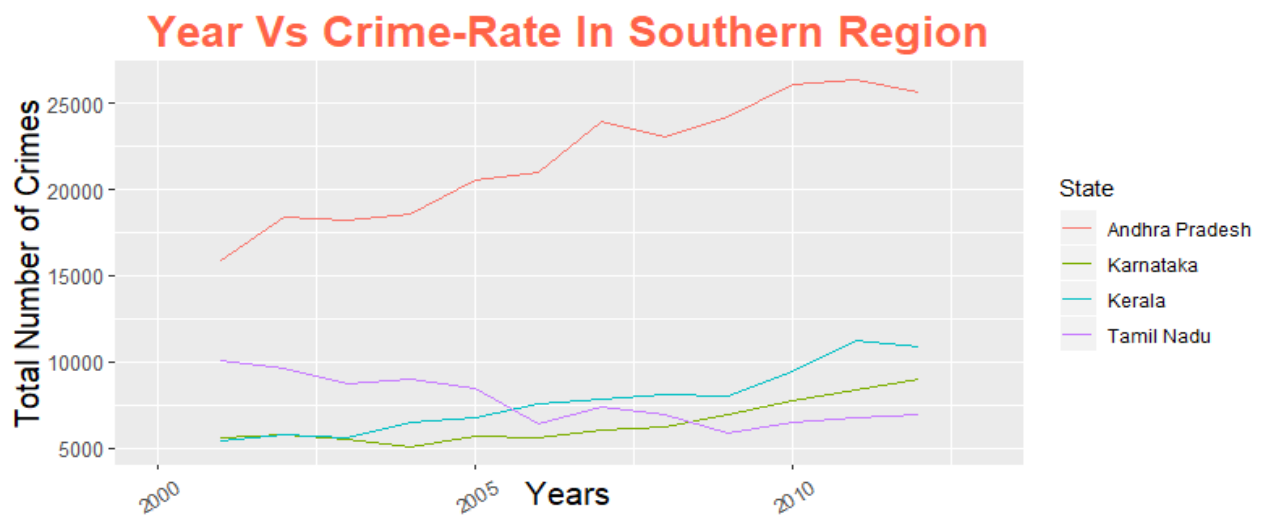
5) West Region

Year Vs Crime-Rate In Western Region



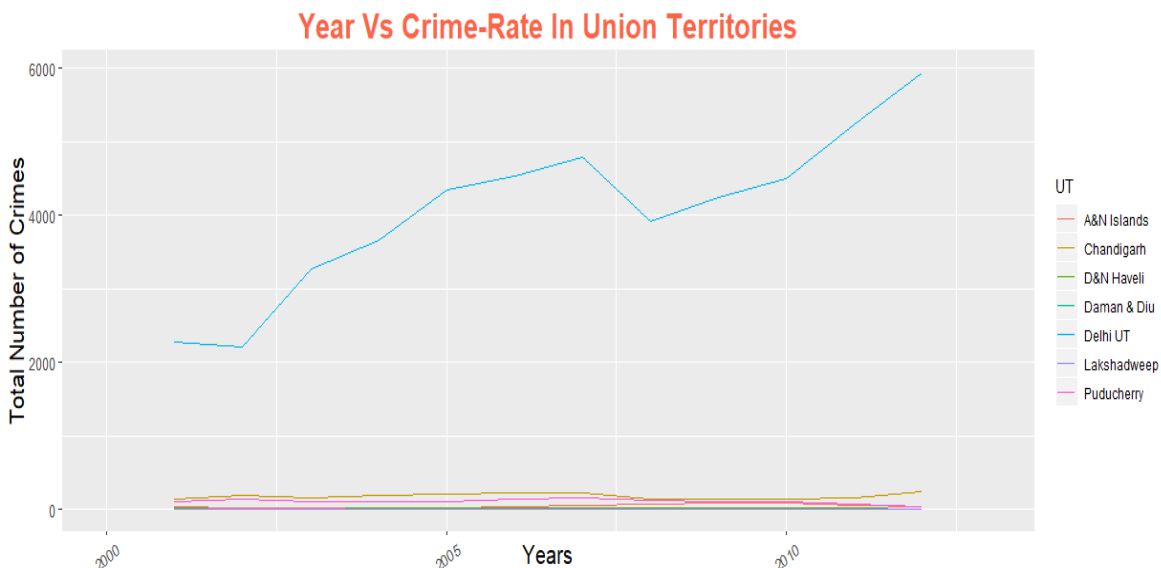
Here again, one may think that the trends of Maharashtra and Rajasthan are comparable while others show lower rate of crimes. However, one must resist falling to temptations of such arguments.

6) Southern Region



One noteworthy thing here is that unlike any of the trends observed above, Tamil Nadu is the only state so far which has reported a downward trend in the numbers of crime committed. Other states in the southern region, like any state observed so far, has shown increase in the trends of crime rates.

7) Union Territories



The crime rates for all the UT's, except, Delhi, are either at the same level or slightly decreasing. It is believed that clearer and comparable trends would be obtained if we consider the crime rates weighted by population.

- III. Thirdly, on the same raw data set, **Spearman's rank correlation coefficient** was found between the different crimes reported. The rationale for doing this was that 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*'.

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 1: Labels for Crimes used in Table 2

The Spearman's Rank Correlation Coefficient is found out for the eight crimes labeled as C1, C2,..., C8. The rationale for using Spearman's rank correlation coefficient instead of Pearson's Coefficient is simply that the Pearson's correlation coefficient works with the assumption that the underlying distribution of the data is Normal with no skewness. Spearman's Correlation Coefficient is not restrictive in that sense. The result are provided in table 2 on the next page:

	C1	C2	C3	C4	C5	C6	C7	C8
C1	1.00	0.87	0.91	0.93	0.75	0.90	0.63	0.59
C2	0.87	1.00	0.86	0.86	0.76	0.85	0.69	0.62
C3	0.91	0.86	1.00	0.87	0.82	0.91	0.71	0.56
C4	0.93	0.86	0.87	1.00	0.84	0.92	0.74	0.61
C5	0.75	0.76	0.82	0.84	1.00	0.77	0.72	0.49
C6	0.90	0.85	0.91	0.92	0.77	1.00	0.71	0.59
C7	0.63	0.69	0.71	0.74	0.72	0.71	1.00	0.64
C8	0.59	0.62	0.56	0.61	0.49	0.59	0.64	1.00

Table2: 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

It can be concluded that there exists a 'strong correlation' in many of the pairwise crimes considered.

This not only affirms our a priori expectations but is further suggestive of the fact that perhaps the technique of principal components can be employed before carrying out clustering techniques on the data.

Some of the infographics have also been attached in appendix.

IV. 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 $X \cup Y$, 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:

$$\text{Support, } S(X \cup Y) = \text{support}(X \cup Y) / N$$

$$\text{Confidence, } C(X \cup Y) = \text{support}(X \cup Y) / \text{support}(X)$$

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}

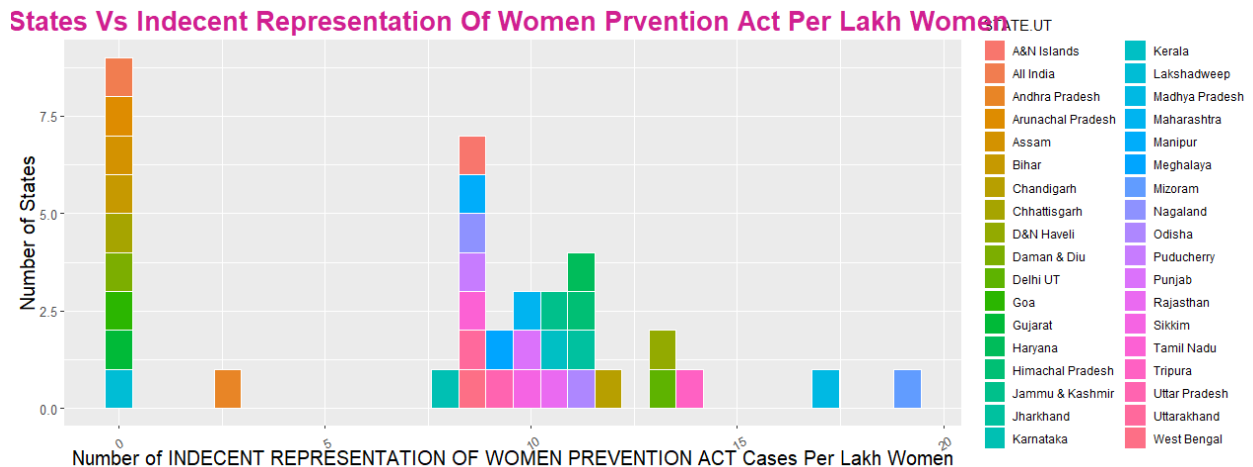
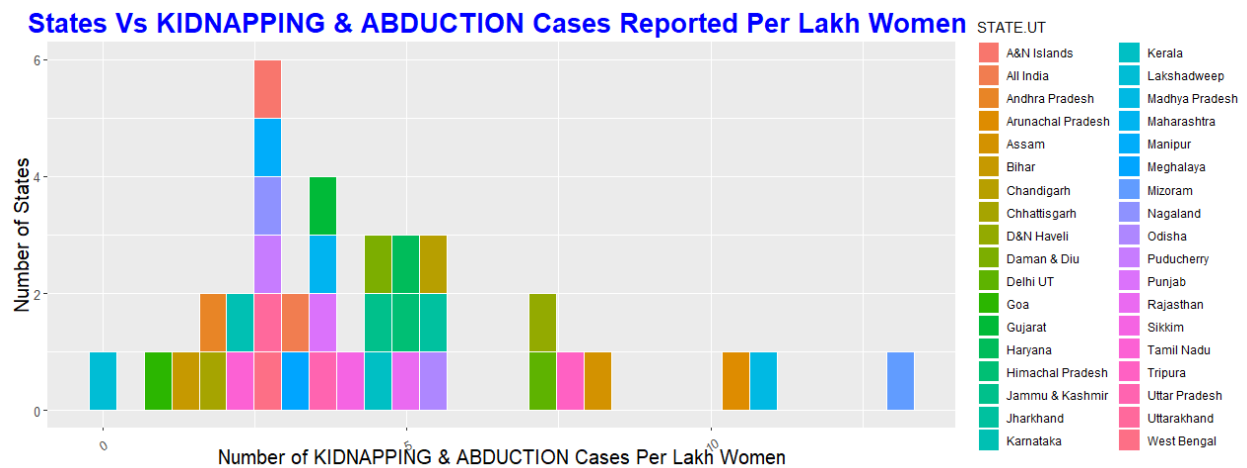
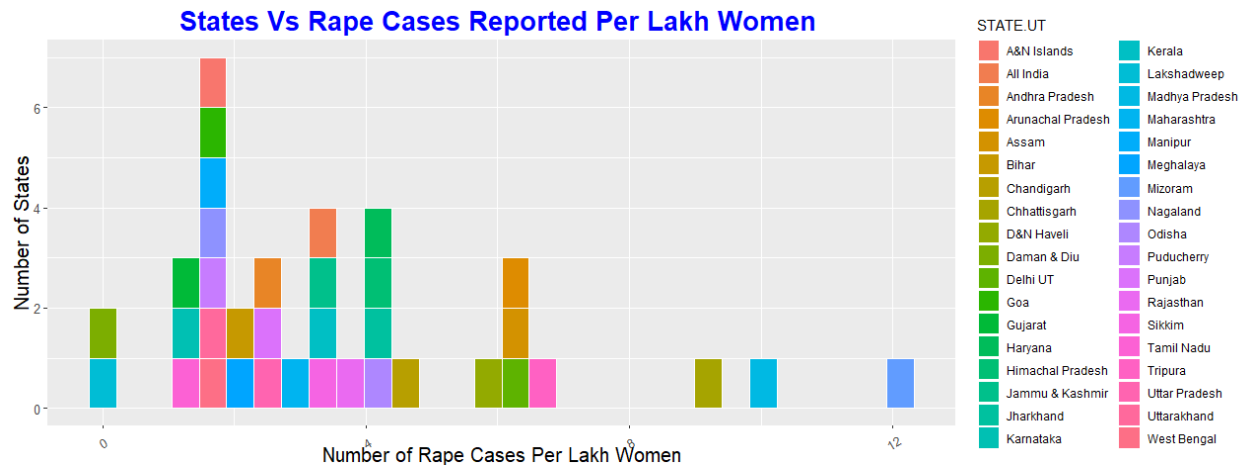
These can be interpreted as:

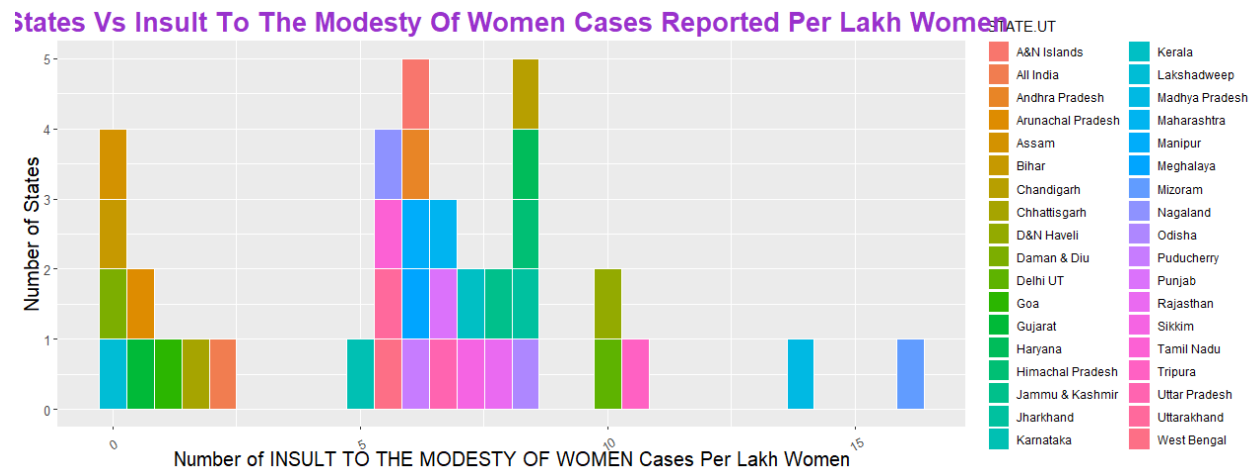
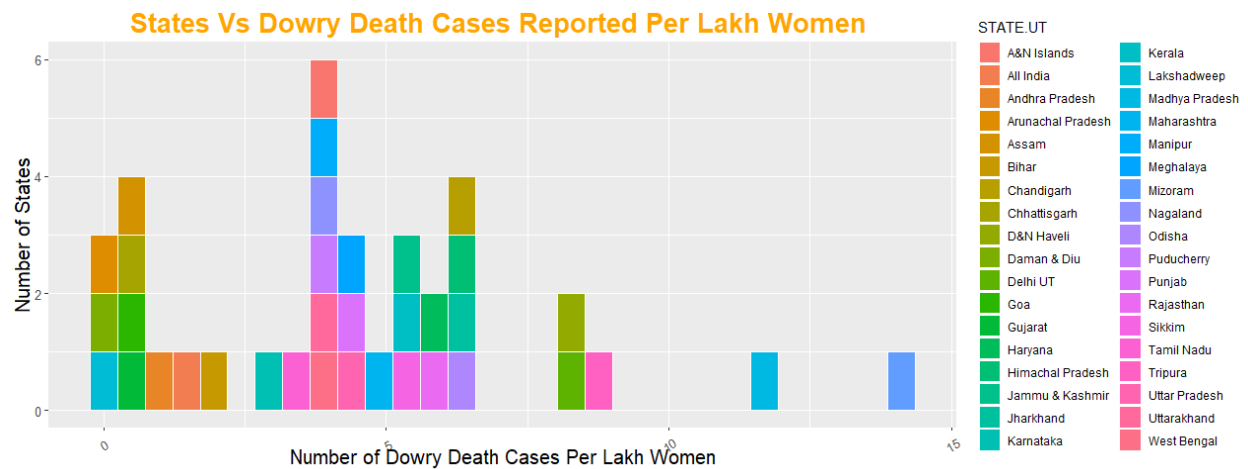
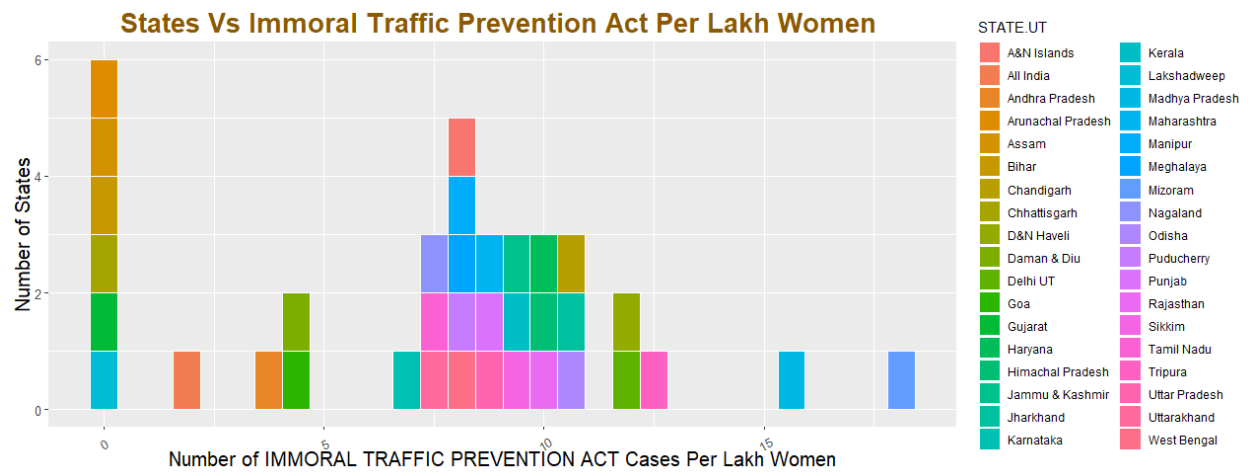
1. If indecent then assault
2. if indecent then cruelty
3. if cruelty and indecent then assault

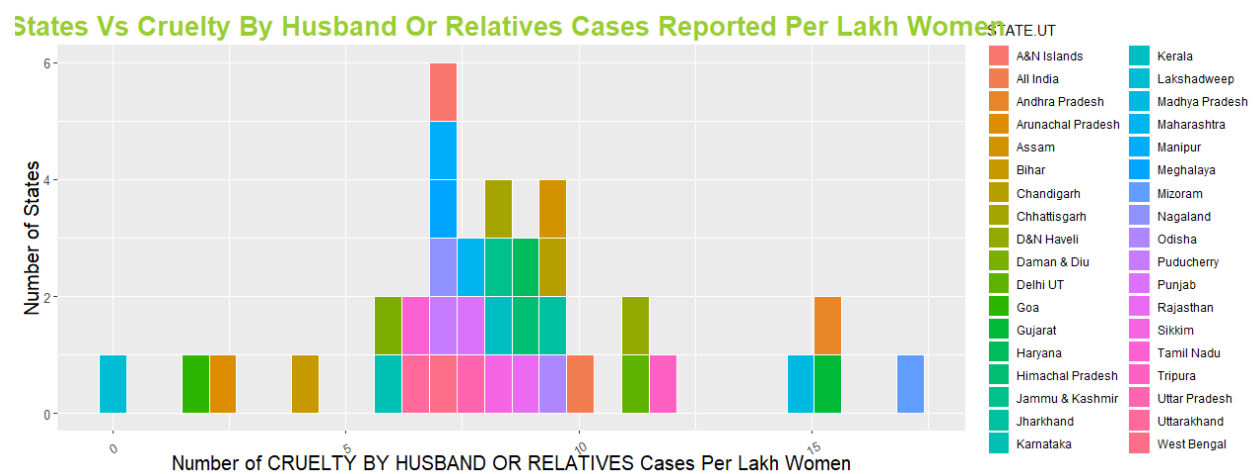
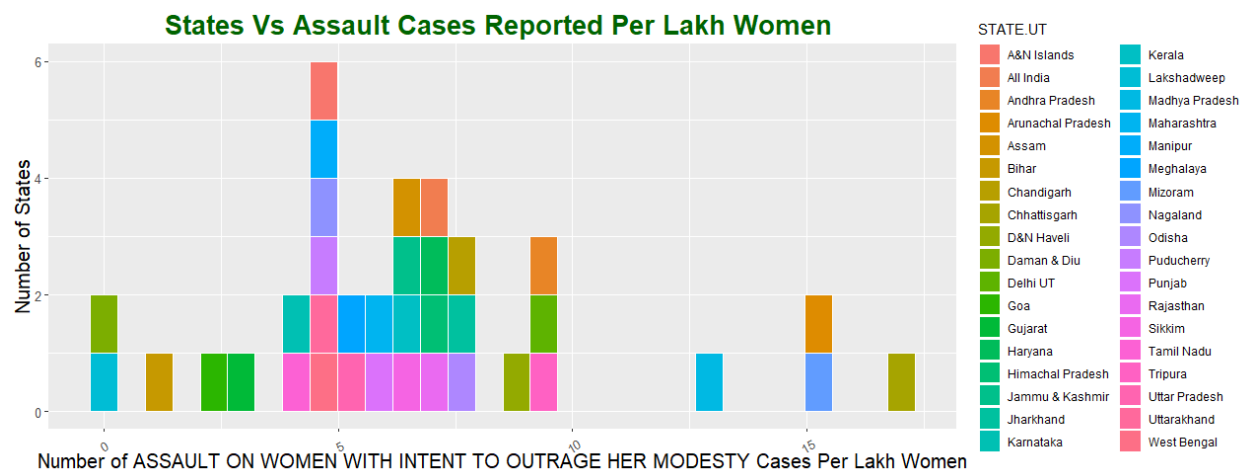
and so on...

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.

V. Lastly, we weighted the crime rates by the respective population of women in that State/UT in 2001. Plotting the states/UTs according to crimes give us the following infographics:







Overall the results show that some states aggregate around lower crime rates for a particular crime while others aggregate around higher crime rates. It would be interesting to see how this aggregation formally changes when we apply clustering techniques across the crimes and over years.

THE WAY FORWARD

- The crime rates for all the years need to be weighted using the projected figures of population (Or with the interpolated figures, whichever technique gives superior results) and further exploratory analysis may be carried out if need be.
- Literacy Rates and Unemployment rates need to be used as one of the factors which would possibly explain the trends in crime rates over the years. This comes from the well-established literature which puts forth the view that for example, increase in unemployment rates is correlated with the crime rates.

- The technique of clustering needs to be accordingly applied to see how clusters change over years, which states move in and out of high crime clusters and so on. For this, also wish to rely on the technique of principal components. However, more thinking needs to go in this.
- If possible, we shall also carry out predictive modelling on our chosen data set.

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.

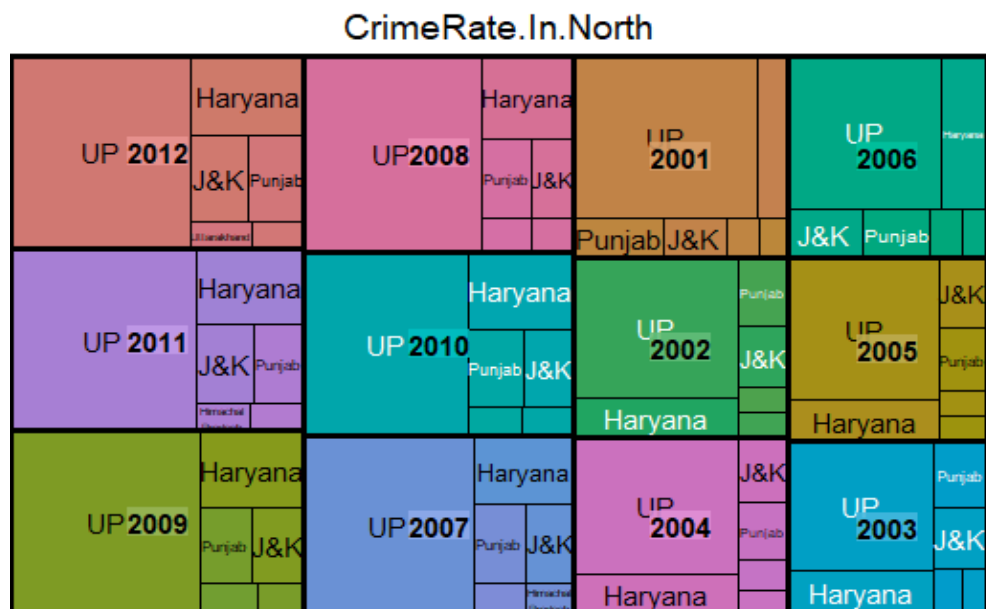
TOOLS TO BE USED

We shall rely, for most part of our project, on making use of WEKA, Python, and R to derive results and info-graphics.

APPENDIX

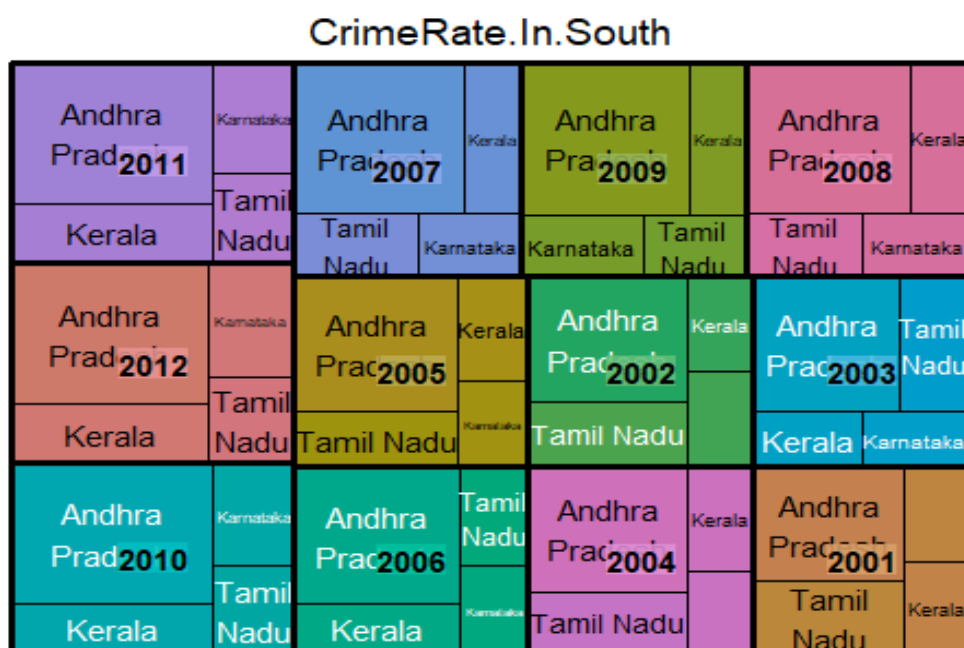
The following charts describe the crime rates in different states of four regions (yearwise):

- North



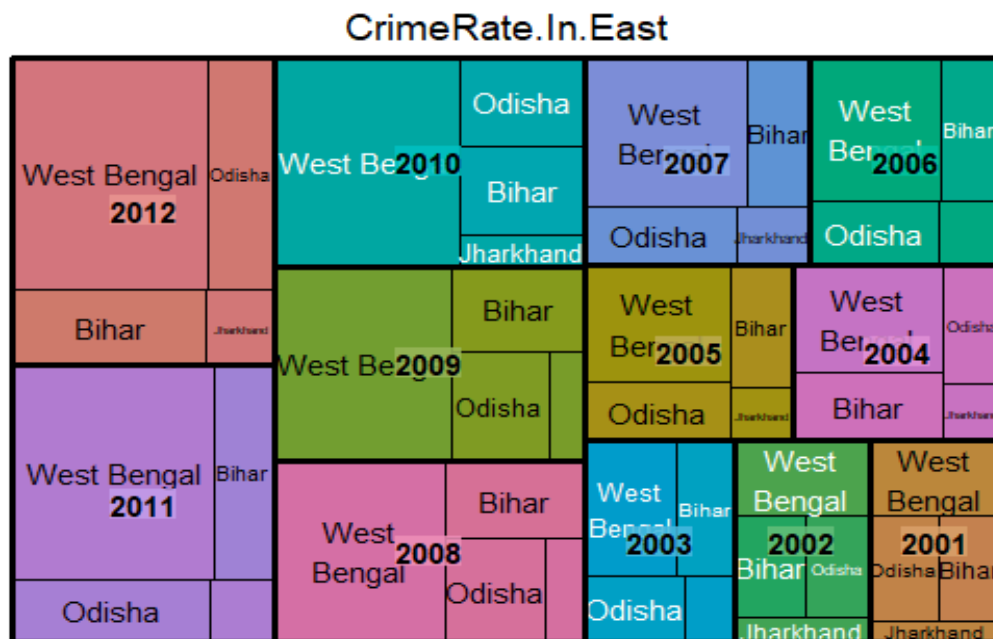
As can be seen above, the overall crime rate for the states in NorthernRegion is increasing in nature. From the above chart, we can interpret that in 2001 more than 50% of crime rates were reported in UP only. Uttar Pradesh has the highest rate of crimes against women in this region.

- South Zone



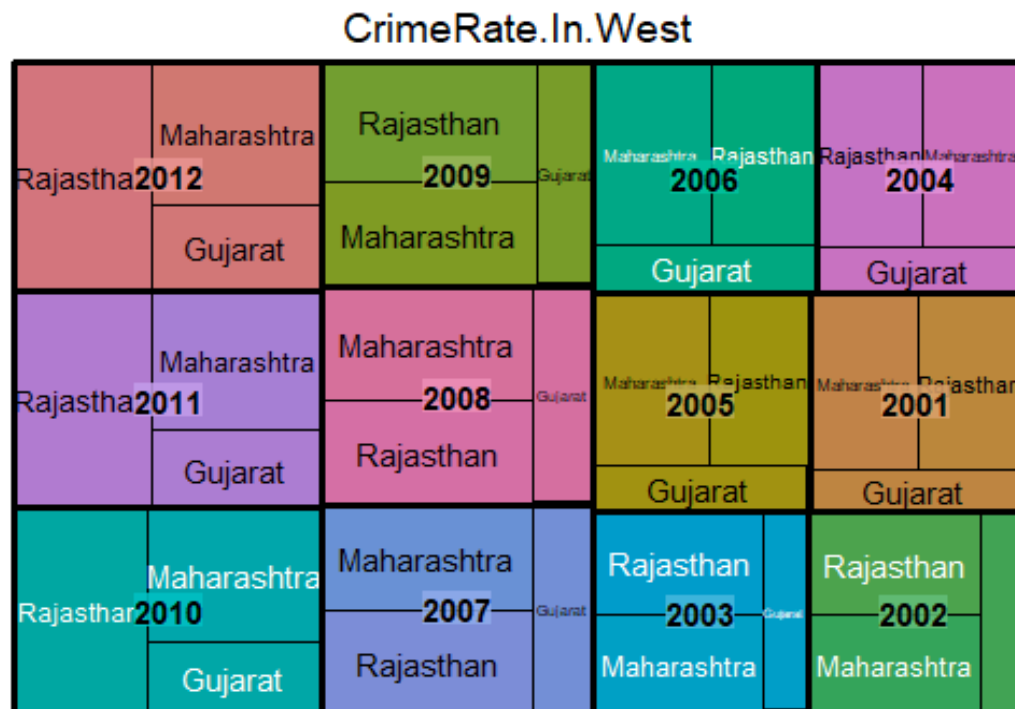
Andhra Pradesh has a higher crime rate than Tamil Nadu and Kerala throughout the period of twelve years(2001-2012).

- East



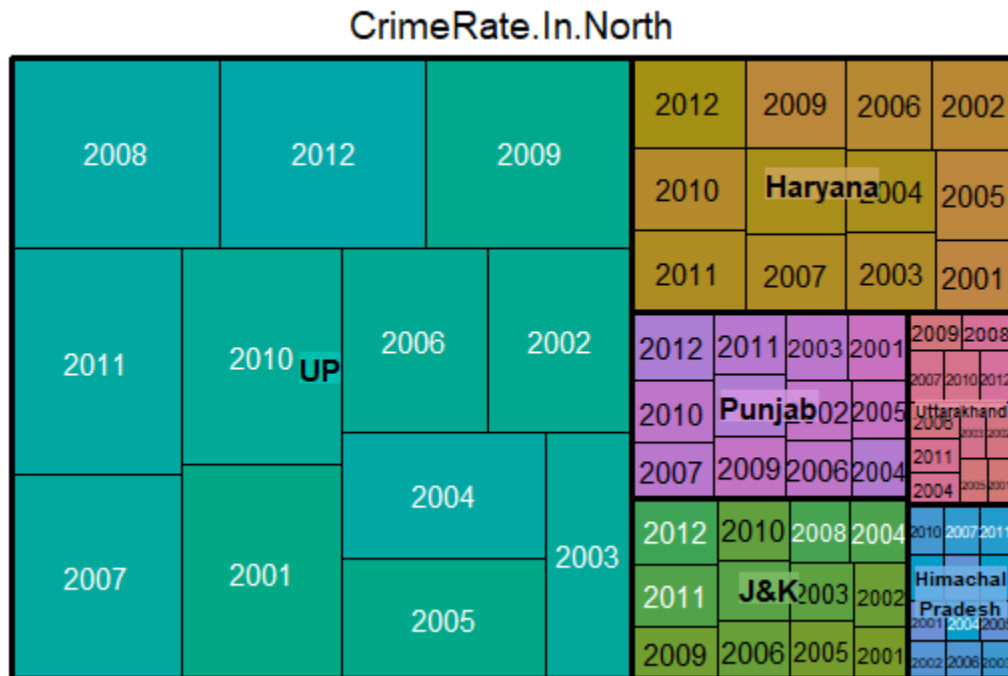
West bengal has the highest crime rate in Eastern region. In addition to this, we can say that after 2008 West Bengal has shown increase in the number of crime rates as compared to other states.

- West



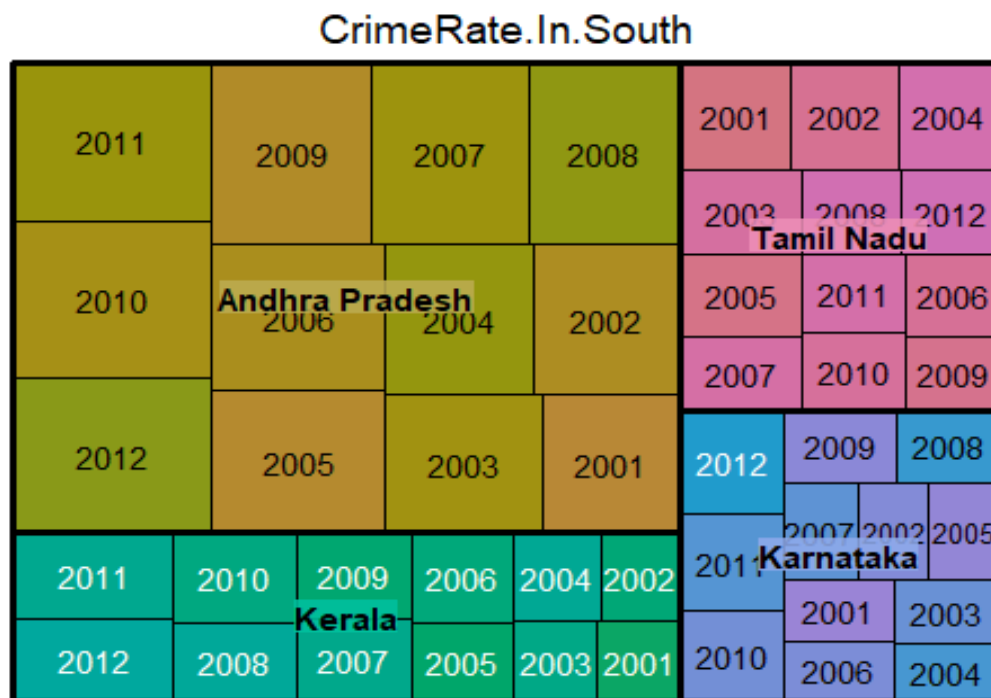
In western region, Rajasthan and Maharshtia has almost equal crime rate against women. The following treemaps represent crime rate in different states over a period of 12 years (2001-2012):

- North



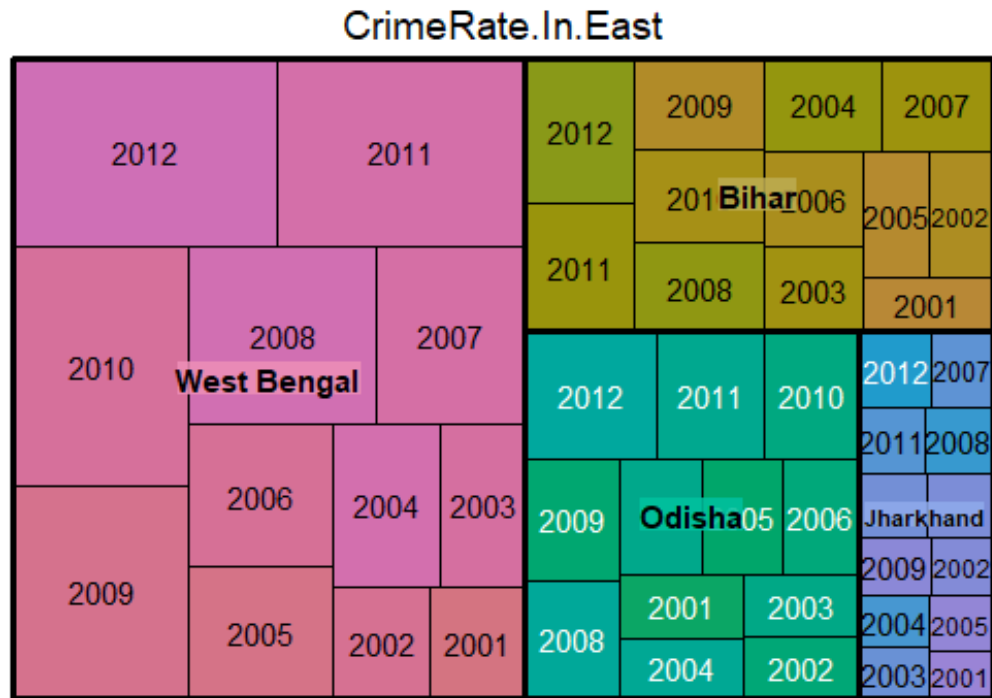
Uttar Pradesh has the highest crime rate in Northern region. Since 2004, there has been a gradual but steady rise in recorded crime rates against women in UP. Crime rate against women was recorded lowest in Himachal Pradesh and Uttarakhand.

- South



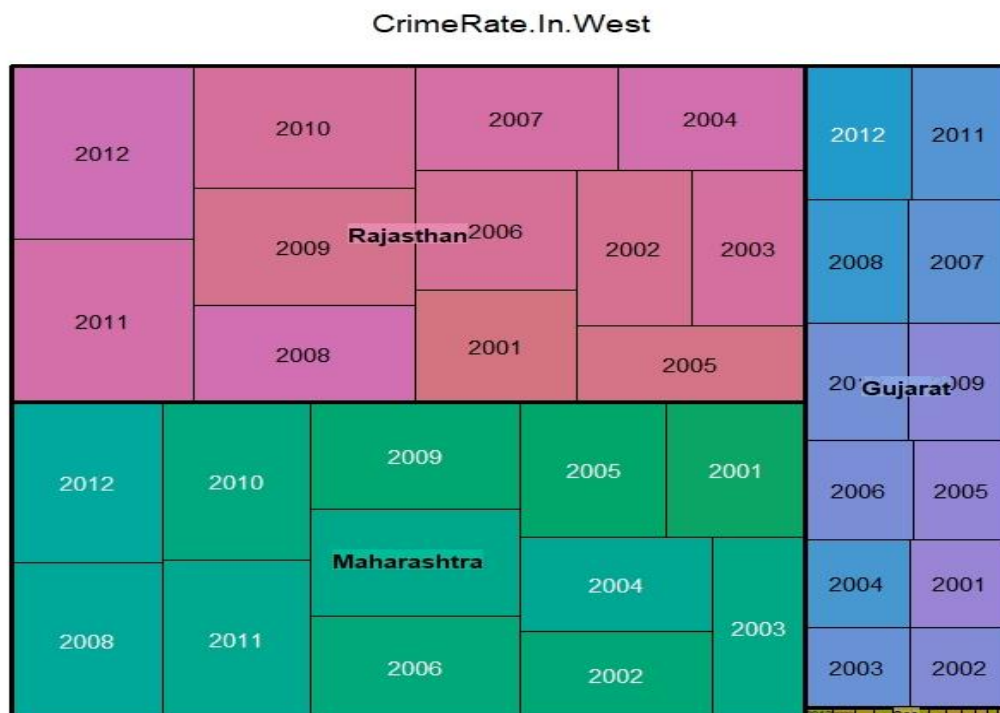
Andhra Pradesh has the highest crime rates in the Southern region. While Tamil Nadu and Karnataka has fairly same crime rates across

- East



Jharkhand has the lowest crime rate in eastern region whereas West Bengal has the highest. We can see an increasing pattern of crime rates for each state.

- West



The above chart shows

that trends of Maharashtra and Rajasthan are comparable while others show lower rate of crimes against women.