K-means Clustering Analysis of Crimes on Indian Women

Rishabh Singh[1*], Rishabh Reddy[1], Vidhi Kapoor[2], Prathamesh Churi[3,4]

[1] MBA in Technology Management in Computer Engineering, School of Technology Management and Engineering, NMIMS University, Mumbai India, rishabhmsingh@gmail.com

[2] B. Tech in Data Science, School of Technology Management and Engineering, NMIMS University, Mumbai India,

[3] Assistant Professor, Computer Engineering, School of Technology Management and Engineering, NMIMS University, Mumbai India

[4] PhD research Scholar, Symbiosis International University, Pune India

Abstract: Violence against women is seen as sexual or physical activity committed against women. In India, general forms of violence against women in India includes cruelty by relatives, dowry, rape, sexual assault, kidnapping, immoral trafficking, molestation etc. The security of the women is the utmost priority of any government in this world. In India, many policies and laws have been enforced to ensure the safety against women. Technology is being the biggest supporter to the government in this context. Data mining allows various techniques such as clustering classification, regression provides analysis in any form of data and helps intelligent predictions on the given dataset. In this paper, we use k-means clustering analysis on women crime dataset. As a part of pre-processing, we collated the data entries which had crime cases against women and made women crime sub-dataset from the real dataset. We then applied K means clustering for further analysis. We used a rapid miner tool for clustering analysis as it is widely used for clustering purposes. After completion of clustering analysis, we proposed our views and discussions on the clustering results. At the end, we ended up giving the futuristic work to be further done on the derived dataset we made and made available on public repositories.

Keywords: crime, clustering, k-means, women, India

1. Introduction

In a world where we talk about equality of race, caste, religion and most of all gender, the nature of the society has become such that the victim to most crimes are still women. There are various crimes that are inflicted upon women which include, murder, rape, molestation, abduction, women-trafficking, domestic violence and many more. We will be focusing on the crimes and cruelty faced by women in India based on statistics. In India, crimes like Sati (burning a women to death when her husband dies), child marriage, female foeticide and domestic violence have been influence by religion and some by tradition, where the tradition on one had treats women as goddesses and on the other hand makes them victim to such heinous deeds [1]. Gruesome crimes like murder and rape as well as comparatively minor issues like theft (might be in the form of chain snatching or burglary), extortion and inequality are all still faced by women in India on a large scale. Elderly women have to face various crimes in the nation like murder, theft, cheating and bag snatching making them more dependent and vulnerable than ever [2]. Off late, due to increasing power of media and rise in technology, these issues have gained importance and been thought and researched upon. But did that decrease the ever-increasing graph of women crimes or did it just mean the criminals were being punished but the crimes kept happening.

Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020

DOI: 10.5281/zenodo.3909955

We have tried to cover all this based on some older and newer statistics and representation of those with a view to understand the reasons behind these acts and if possible help in any way to reduce these offenses, gradually making the country a much better and safer place of now just women but the entire mankind.

These crimes can be analysed using various methods of data mining, which is a method of uncovering hidden information using big data. This has become a method of investigation throughout the world now. There are numerous data mining techniques and they include entity extraction, decision trees, clustering, neural networks, social network analysis and many more [3]. Clustering analysis is the process that we have adopted for our research. Data mining is a procedure that includes evaluating and examining large databases in order to generate some new information that may be fruitful, however, its usage in the field of criminology is very recent [4]. Data mining is the process that can be used from making various conclusions as well as statistical predictions based on the pattern of the dataset, the crime, even the criminals or the victims can be characterized based on different criteria for better understanding. Data needs to be analysed from an informational collection and be changed into a reasonable structure for additional utilization [5]. Analysing the data has become increasingly important because of the steady increase in the crime rate across the globe. The goal behind this is to find out reasons and help the investigators come up with a viable solution for the same. Data mining is used on crime datasets to study the features that lead to the high crime rate. This can be done using two methods, description mining and classification with prediction. The former is usually associated with rules, clusters, patterns or correlations, while the latter uses probability equations and prediction of futures using statistical measures [6].

The method that seemed most suitable for our research was Clustering analysis. Clustering analysis method is one of the main analytical methods in data mining, the method of clustering algorithm will influence the clustering results directly [7]. k-means for clustering analysis is one of the many methods in data mining. k-means clustering is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells (a Voronoi diagram is partitions of a plane into regions close to each of a given set of objects; these regions are called Voronoi cells) [8]. Practical approaches to clustering use an iterative procedure (e.g. K-Means, EM) which converges to one of numerous local minima. It is known that these iterative techniques are especially sensitive to initial starting conditions [9]. For successful analysis using clusters, finding good clusters has to be the focus of the research like in many researches in the field of facial recognition and machine learning [10]. After looking into various researches and reviews on k-means for clustering we can say that k-means algorithm is an incremental approach that dynamically adds one cluster centre at a time through a deterministic global search procedure consisting of N (with N being the size of the data set) executions of the k-means algorithm from suitable initial positions [11].

We have obtained the dataset of various crimes inflicted upon women in the decade 2001-2010 from the official crime database website. The dataset set contains the crimes namely, Cruelty by husbands and relatives, Dowry, Immoral Trafficking, Kidnapping, Molestation, Rape and Sexual Harassment. The database has records for every crime and the count of each corresponding state or union territory. This already segregated data was then converted into diagrams of clustering analysis using k-means on the Rapid Miner tool freely available. This process helped us in obtaining a separate diagram for each crime but in the same process giving us an overview of all states and union territory, each state and union territory got a different cluster. By this, we can analyse a particular crime in depth across the nation. Through visualization of these obtained clusters we get an idea of which regions need more attention and which regions are in a better condition. The next visual was obtained using the Weka tool which gave a clearer view of each crime and which state or union territory had more victims of that crime in the given time span. The histograms also help us to analyse the increase or decrease in crime over the total time span. This helped us realize which crimes have increased or

Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020

DOI: 10.5281/zenodo.3909955

decreased in modern times, whereas some crimes have been steady. These results have been obtained by us with a view to helping the investigators and activists towards betterment of the treatment of women across the nation.

A change can take place if research and analysed data reach the proper authorities and they can use it to their advantage for improving the condition of women across the world, but particularly in India as the population is very large and ever increasing. Also, the various traditions and beliefs responsible for some heinous crimes inflicted upon women in India can be seen through this analysis which can be used for taking appropriate measures for the same.

2. Literature work on clustering algorithms:

Data mining called practice of examining large databases to generate new information Is the modern age technique through which we can gather information and then we can generate a new theory from the set of information which will be beneficial for the human kind. There are different types of data mining techniques which are used in today's technologically enhanced world, namely regression is used to identify the likelihood of a certain variable, given the presence of other variables. Classification is to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. Clustering which is grouping chunks of data together based on their similarities. For example, you might choose to cluster different demographics of your audience into different packets based on how much disposable income they have, or how often they tend to shop at your store. Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. Crime control tends to use knowledge from the analyzed data to control and prevent the occurrence of crime, while the criminal suppression tries to catch a criminal by using his/her history recorded in data mining.
This section gives a detailed review on the previous papers which have dealt with the clustering analysis on crime.

[12] Renuka Nagpal et al. (2013) has used clustering analysis to detect the pattern of homicide from 1990 - 2011 and have gathered the data set from offenses registered against criminals in England and Wales. They have used a rapid miner tool to gather and sort out the dataset. Then using K means analysis to plot the homicide year by year and found out that the rate of homicide has decreased from 1990 - 2011 and from the clustered results it is easy to identify crime trends over years and can be used to design precaution methods for future.

[13] Dr.S.Santhosh Baboo et al. (2011) proposed a new enhanced algorithm which is known as the HYB Algorithm and built a tool which gave all the essential features from clean, characterize and analyze crime data to identify actionable patterns and trends. The development of this tool has four steps and are as follows data cleaning, clustering, classification and outlier detection. The results have proven that the tool is very efficient and effective as compared to the state-of-the-art tool.

[14] Juhana Salim et al. (2013) They have used document clustering to analyze a crime pattern in which they have two processes one which extracts the information from the data given and second clustering which forms a cluster of all the similar kinds of data. Extraction has features which will detect a particular keyword in the crime list and segregate the data accordingly. The state-of-the-art cluster algorithm i.e. means was found out to be slow, So the authors proposed a new algorithm which was

Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3909955

called Affinity propagation algorithm. Crime document clustering will enhance the performance and the effectiveness and should be used as told by authors.

[15] Devendra Kumar Tayal et al. (2014) They have introduced a method known as CDCI i.e. Criminal identification and criminal detection in which CDCI extracts unstructured crime data from various crime Web sources and then preprocessed the crime data into structured 5,038 instances that are represented using 35 predefined crimes attributes. They have used WEKA for crime verification and K means which is used for clustering analysis of the crime data set. CDCI then applies k-means clustering for crime detection during 2000–2012 through four cases. Case 1 detects crimes in India irrespective of crime location and crime type. Case 2 detects crimes in specific locations, e.g. Delhi, irrespective of crime type. Case 3 detects crimes of specific type, e.g., murders, irrespective of crime location. And Case 4 detects crimes of specific type and in specific locations. The CDCI uses KNN which is used to plot the clustered data on the map of India. CDCI can aid the law enforcement agencies to enforce the security of citizens of India.

[17] Ubon (2016) et al. did a survey of various data mining techniques which are used for analyzing crime patterns. They have shortlisted four major crimes patterns. Border Control, Violent Crime, Narcotics and Cyber Crime and for these four patterns are analyzed by such shortlisted algorithms which are very efficient for analyzing these patterns. They have also mentioned about the various issues and challenges faced and one such issue is crime pattern as the data collected can be concerning with finding and predicting the hidden crime. And visualization that the amount of data is growing rapidly, which leads to the difficulty and complication to display the hidden knowledge.

[16] LalithaSarojaThota et al. (2017) have gathered data from the National Crime Records Bureau (NCRB) and they have used WEKA Explorer to cluster the data set according to the different states. The main aim of the paper is to find the zones which have the higher crime rates to the zone which have lower crime rates through My Custom map, an online interactive map tool of maps of India to create custom India maps with the cluster zones of states. The clustered data is then entered into the map tool and we have the result where we can see the zones where the crime rate is high and the zones where the crime rate is low.

The table 1 summarizes the research work in clustering analysis in the crime dataset.

Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020

DOI: 10.5281/zenodo.3909955

Table 1. Summary of clustering analysis of crime dataset

Authors	Dataset Used	clustering algorithms used	Tools	Inference
[12]	offences recorded by the police in England and Wales by offence and police force area from 1990 to 2011-12	K means	Rapid Miner	relatively efficient often terminates at a local optimum.
[13]	Integrated Network for Societal Conflict Research (INSCR)	DB Scan (Density-Based Spatial Clustering Application with Noise)	computer program (Rapid Miner)	efficiency of a state in controlling crime rate.
[14]	Bernama News	Modified K-Means	----------	crime document clustering enhances the performance and effectiveness.
[18]	National Crime Records Bureau (NCRB)	K-Means Clustering algorithm	Rapid Miner	tracking homicide crime rates from one year to the next
[19]	Persons arrested for crime against women	Kernel Density	----------	Runtime of algorithm is smooth and rapid
[15]	National Crime Records Bureau (NCRB)	K- Means (CDCI and GMAPI)	WEKA and Java Tools	CDCI can speed up the crime solving process by processing and filtering the voluminous crime data within a short span of time.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

[16]	National Crime Records Bureau (NCRB)	K- Means	WEKA	crime trend and zoning knowledge can also be helpful in cautioning police to increments and reductions in levels of preventive actions.

3. Methodology

3.1 K-means clustering

K-means clustering is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells (a Voronoi diagram is partitions of a plane into regions close to each of a given set of objects, these regions are called Voronoi cells). It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

The problem is computationally difficult; however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modelling. They both use cluster centres to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbour classifier, a popular machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbour classifier to the cluster centres obtained by k-means classifies new data into the existing clusters. This is known as the nearest centroid classifier or Rocchio algorithm.

We are using k-means because it is relatively easier to implement using tools like RapidMiner, Weka, KNIME, Orange and many more. k-means scales to larger data sets and guarantees convergence. It can warm-up the start position of centroids. As it generalizes clusters of different shapes and sizes, such as elliptical clusters, it can easily adapt to new examples. Few difficulties that we had to face while using k-means for clustering included, scaling with respect to dimension and clustering data of varying sizes and density. We had to choose the value of k manually. Also, the clustering was dependent on the initial value of k and was not dynamic.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

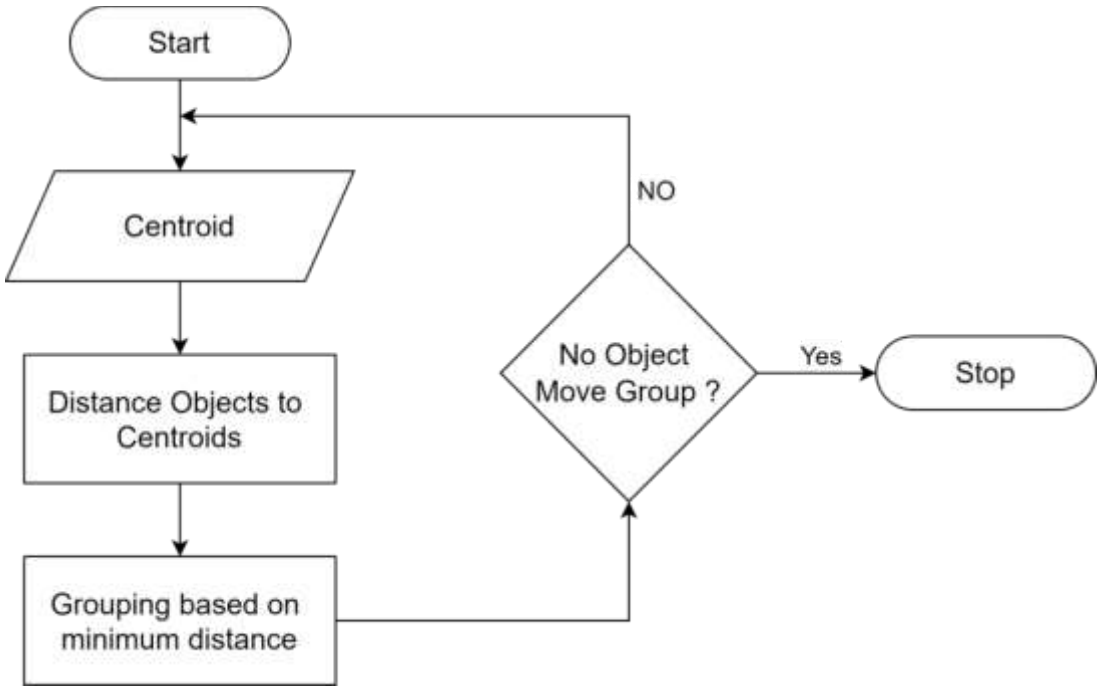


Figure 1. K-means clustering algorithms.

3.2 Dataset and flow of research work

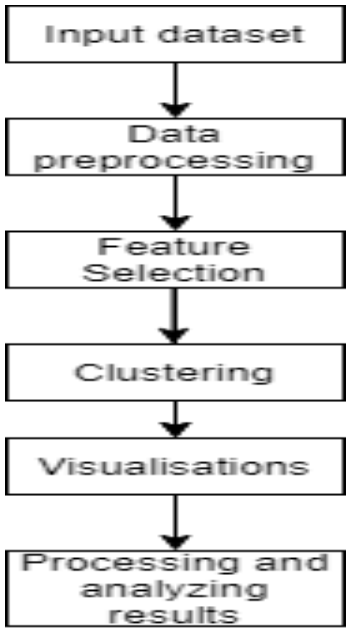


Figure 2. Methodology of proposed research work.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

The dataset taken for the clustering analysis consists of mixed data of all the crimes across all the age groups in India. The data was scattered from different states, including both genders and across all the age groups from 2001 to 2010. We have pre-processed the data by removing crime cases that happened on male and classified into different women crimes. The figure 2 gives the detail of the methodology we adopted for our research work. We used the Weka tool [20,21] for clustering analysis.

4. Results and Discussions:

After applying k means clustering, the table 2 gives various clusters and the inclusive states. The graphical representations of all the clusters according to different crimes are drawn in this section. The overall crimes analysis is drawn in figure 10. Discussing the further analysis, we have also segregated the crime data year wise and plotted histogram for the same.

Table 2. Different clusters through K-means algorithm for Indian states

Crimes	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cruelty by husbands and relatives	Andhra Pradesh, Kerala, Madhya Pradesh	Odisha, Tamil Nadu, Assam	Andhra Pradesh, Assam, Bihar	Gujarat, Maharashtra, West Bengal	Delhi, Himachal Pradesh, Jharkhand
Dowry	Tamil Nadu, Delhi, Gujrat	Andhra Pradesh, Madhya Pradesh, Odisha	Assam, Chattisgarh, Kerala	Bihar, Maharashtra, Uttarakhand	Andhra Pradesh, Assam, Bihar
Immoral Trafficking	Chattisgarh, Meghalaya, Odisha	West Bengal, Dadra and Nagar Haveli, Haryana	Sikkim, Tripura, Assam	Kerala, Uttarakhand, Andhra Pradesh	Gujarat, Jharkhand, Maharashtra
Kidnapping	Uttar Pradesh, Himachal Pradesh, Arunachal Pradesh	Andhra Pradesh, Assam, Jammu and Kashmir	Daman & Diu, Goa, Mizoram	Bihar, Gujarat, Maharashtra	Chattisgarh, Delhi, Haryana
Molestation	Madhya Pradesh, Maharashtra, Andhra Pradesh	Dadra & Nagar Haveli, Daman & Diu, Goa	West Bengal, Andhra Pradesh, Chhattisgarh	Gujarat, Haryana, Jammu & Kashmir	Himachal Pradesh, Jharkhand, Punjab
Rape	Madhya Pradesh, Maharashtra, West Bengal	Daman & Diu, Goa, Mizoram	Andhra Pradesh, Assam, Bihar	Tamil Nadu, Delhi, Gujarat	Arunachal Pradesh, Punjab, Tripura

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

Sexual Harassment	Andhra Pradesh, Maharashtra, Uttarakhand	Uttar Pradesh, Chhattisgarh, Delhi	Arunachal Pradesh, Dadra & Nagar Haveli, Daman & Diu	Gujarat, Odisha, Tamil Nadu	Goa, Himachal Pradesh, Jharkhand

4.1 Clustering analysis of Sexual harassment

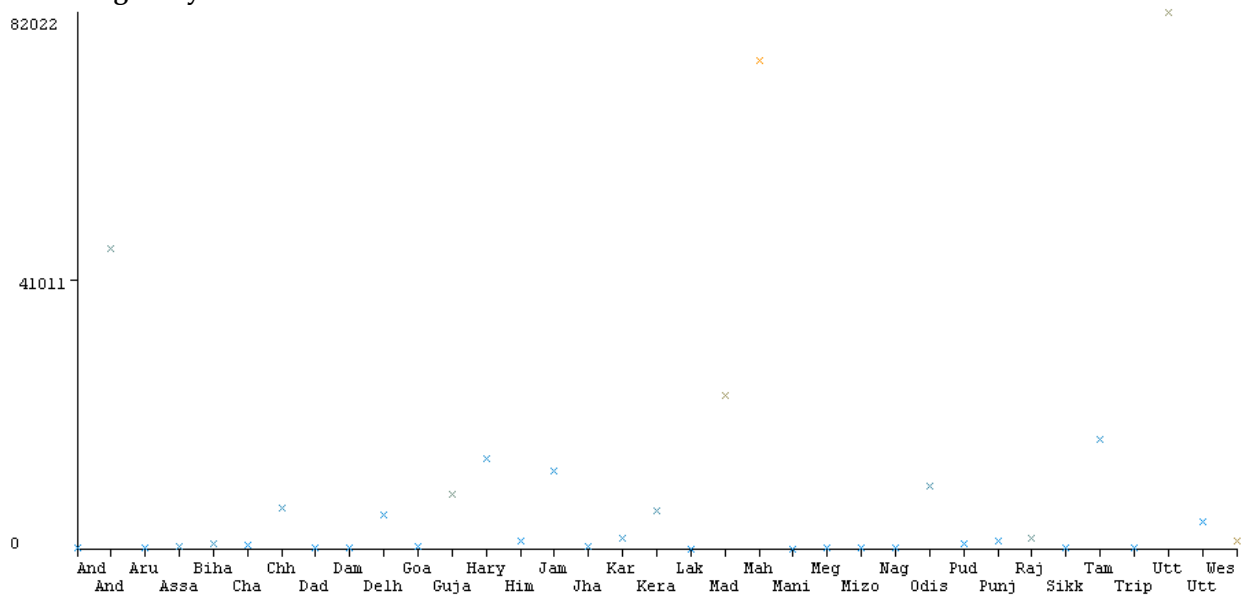


Figure 3. Clustering analysis of Sexual harassment

As per the graph of Sexual Harassment crimes for the past 10 years vs the States and Union Territories of India, we see that Uttar Pradesh has recorded the most number of Sexual Harassment cases in the past 10 years. There are 2 more states - Maharashtra and Andhra Pradesh that have recorded Sexual Harassment cases above 41,000.

Rest of the states are below the 41,000 mark. The ones that are too close to the X-axis have hardly recorded 20 or 30 cases in the 10 years' time. To name a few, some of these states/union territories are Daman & Diu, Meghalaya, Manipur. So, these States/ Union Territories are doing good as per Sexual Harassment cases.

4.2 Clustering analysis of Rape:

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

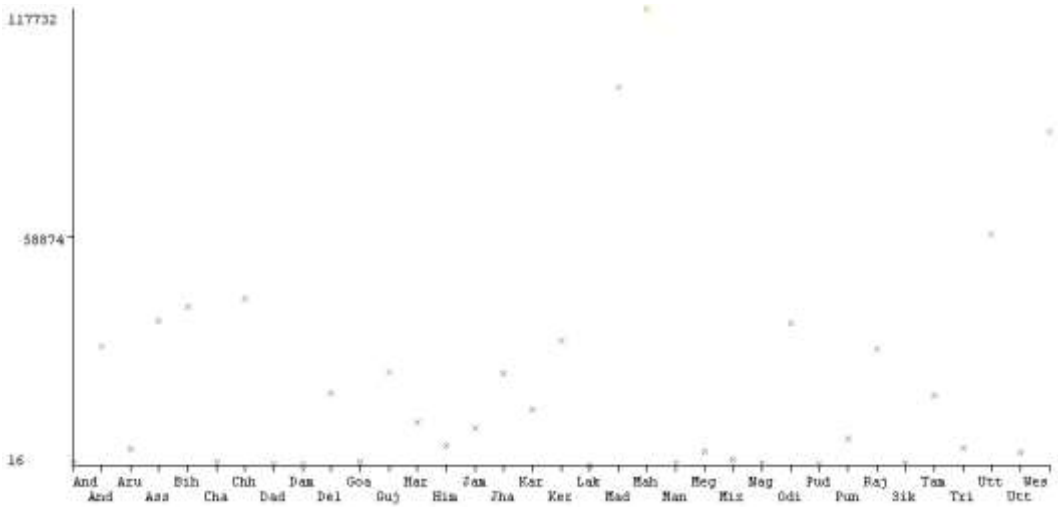


Figure 4. Clustering analysis of Rap

In this graph of Rape cases for 10 years vs the States and Union Territories of India, we see that Maharashtra had the highest number of recorded cases in 10 years (2001-2010) with Madhya Pradesh being the second. Here, there are less number of points which are close to the X-axis so it says that Rape is a more common crime as compared to other crimes. The highest number of cases is close to 1,17,000 which is much more than the highest number of Sexual Harassment cases.

4.3 **Clustering analysis of Molestation**:

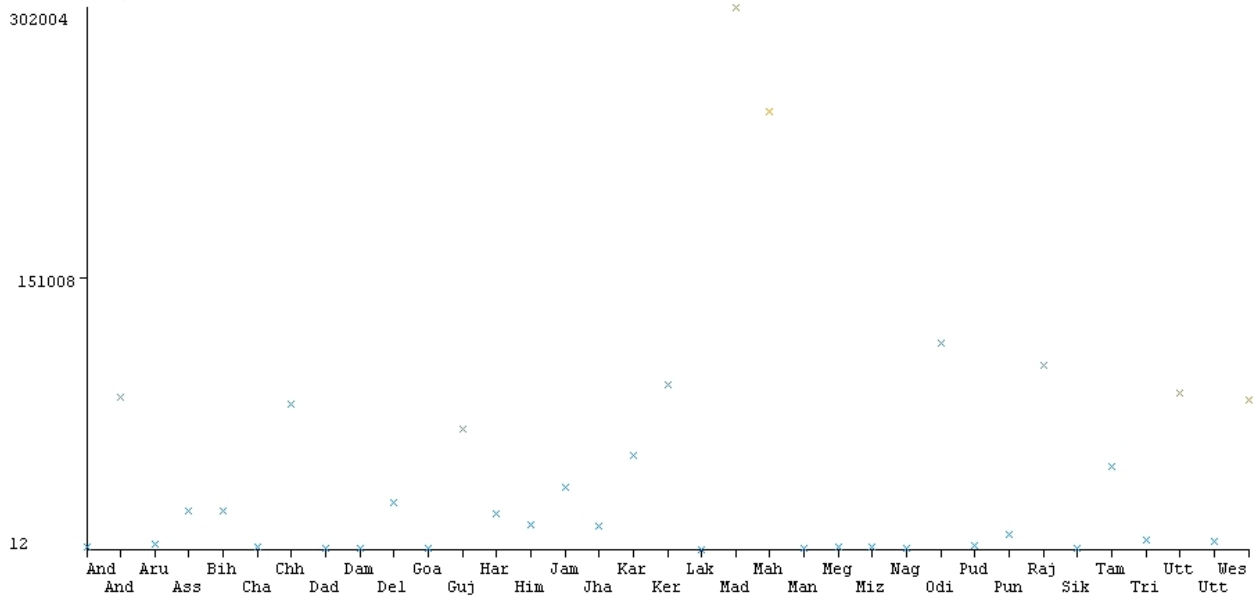


Figure 5. Clustering analysis of Molestation

In the above graph of Molestation crime for 10 years (2001-2010) vs States and Union Territories of India, we see that Madhya Pradesh has recorded the most number of cases for Molestation with Maharashtra being the second. In this graph also there are less number of points near the X-axis

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

which means this crime is happening in all the states and the females of our country are suffering. The highest number of cases recorded is close to 3 lakhs which is more than the double of rape cases.

4.4 Clustering analysis of Kidnapping

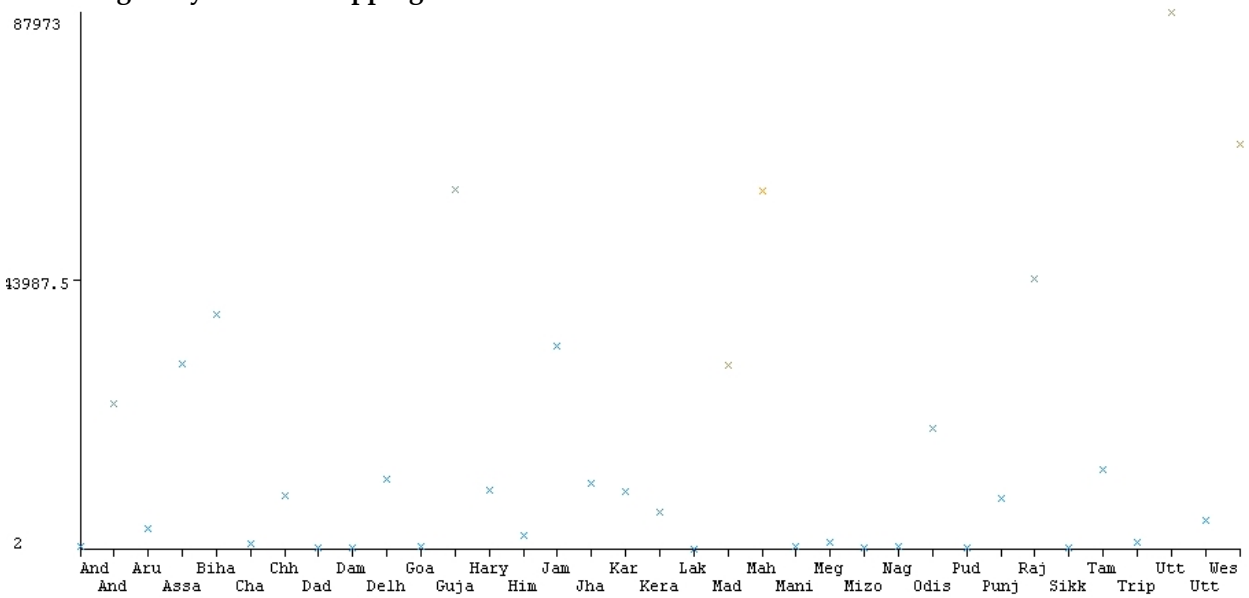


Figure 6. Clustering analysis of kidnapping

The above graph of kidnapping crime recorded for 10 years (2001-2010) vs States and Union Territories of India is very scattered as compared to others. Uttar Pradesh has recorded the most number of cases with West Bengal being the second. There are very less points close to the X-axis which says that Kidnapping of girls and females is also recorded in every state and union territory of India.

4.5 Clustering analysis of Immoral trafficking:

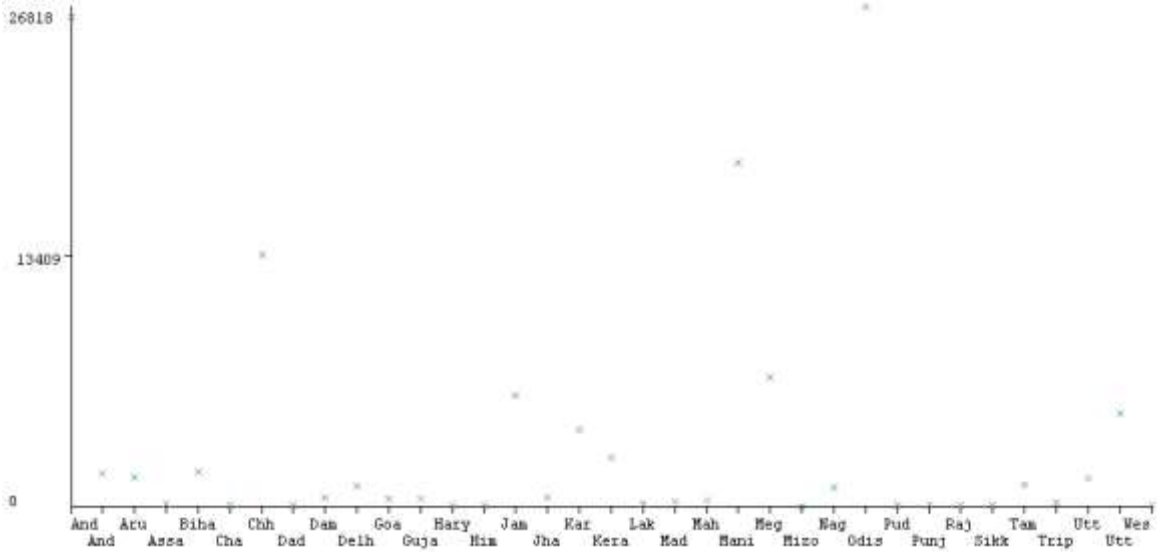


Figure 7. Clustering analysis of Immoral trafficking

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

The graph of Immoral Trafficking crime considered for 10 years (2001-2010) vs the State and union territories of India shows that Odisha has recorded the most number of cases with Manipur being the second. There are many points which are close to the X-axis, so we can infer that there is less number of Immoral Trafficking cases that are registered. The highest number of cases recorded is close to 25 thousand which is less as compared to other crimes pertaining in the country.

4.6 Clustering analysis of Dowry:

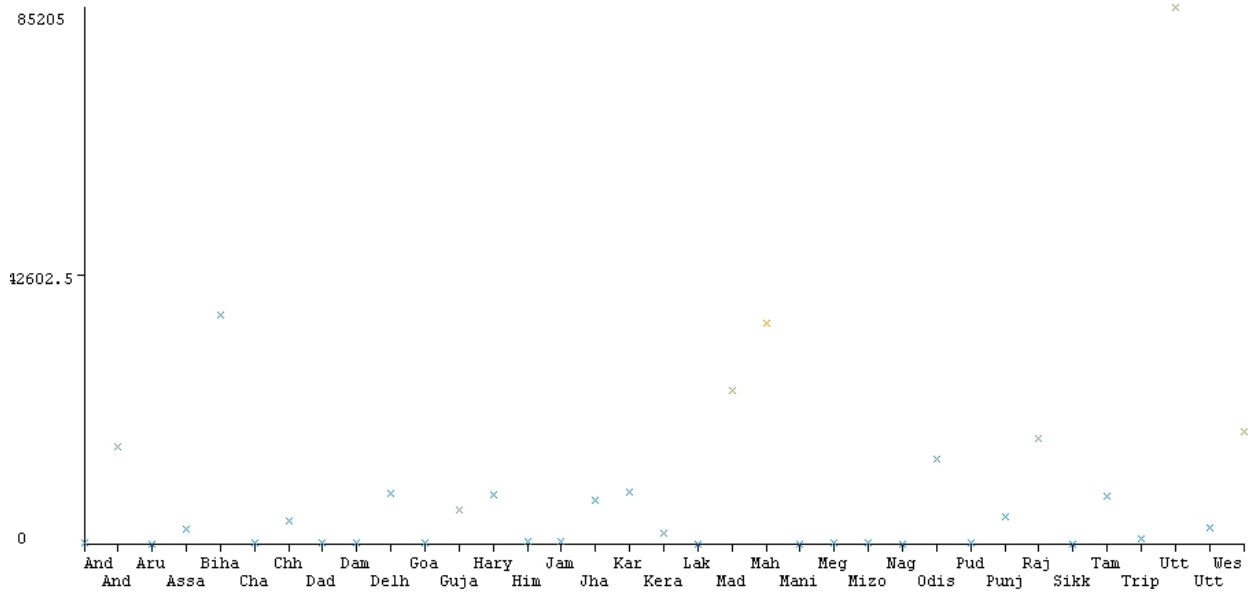


Figure 8. Clustering analysis of Dowry

The graph of Dowry cases registered for 10 years (2001-2010) vs the Indian states and Union Territories shows that Uttar Pradesh has recorded the highest number of cases. There are many data points which are close to the X-axis, so we can also see that Dowry cases that are registered are less in India.

4.7 Clustering Analysis of cruelty by husband and relatives

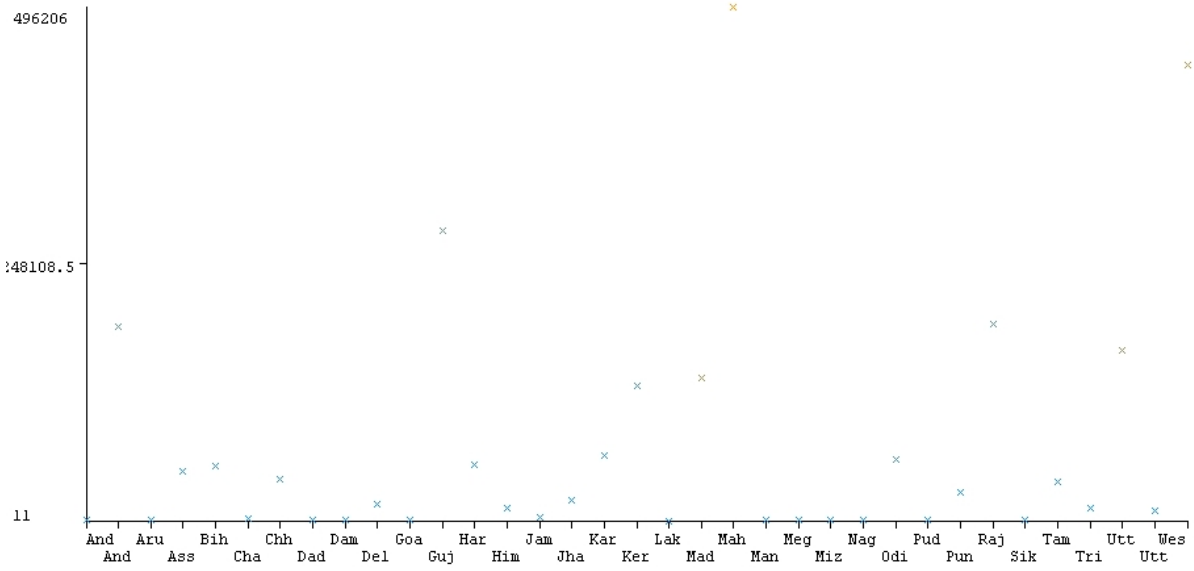


Figure 9. Clustering analysis of cruelty by husband and relatives

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

The graph of cruelty by husbands and relatives' cases registered vs the Indian states and Union territories shows that Maharashtra has the highest number of cases which is close to 5 lakhs with West Bengal being the second. This crime has the most number of cases registered compared to any other crime.

Overall, we see that immoral trafficking has the least number of recorded cases in India and Cruelty by husbands and relatives has the most number of cases in India. Crime against women has been an ongoing thing since ages. Overall, the states that have registered the most cases are Maharashtra, Uttar Pradesh and Madhya Pradesh commonly, then there are also states like Odisha and Manipur that have recorded the highest number of cases specific in some crimes.

4.8 Clustering analysis of total crimes:

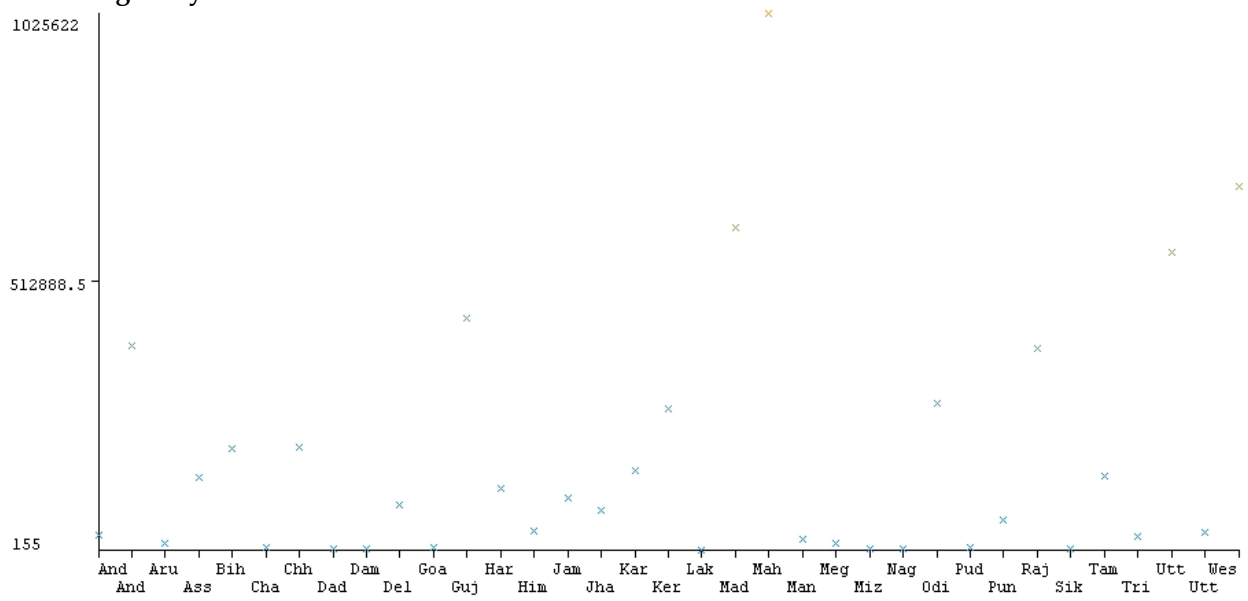


Figure 10. Clustering analysis of total crimes

This is the graph of total crimes that are registered in the country pertaining to women vs the Indian states and union territories shows that Maharashtra state has recorded the most number of cases with West Bengal being the second. The total numbers of cases registered in 10 years (2001-2010) is around 10 lakhs.

4.9 Year wise analysis of crimes on women:

4.9.1 Cruelty by husbands and relatives:

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

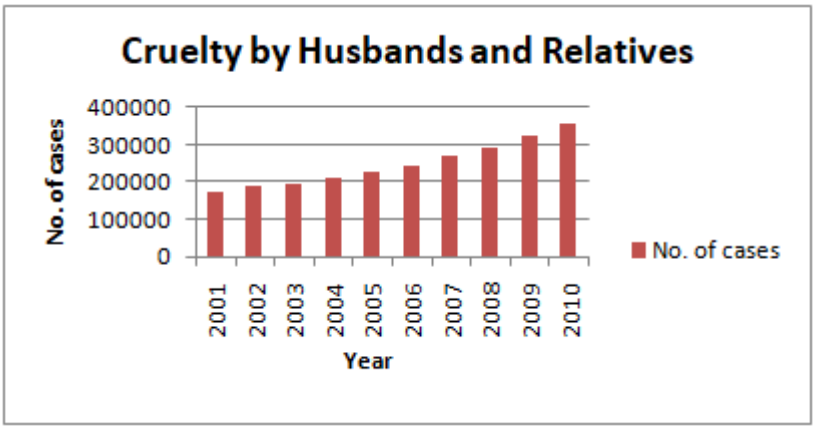


Figure 11. Year wise analysis of cruelty by husbands and relatives

In the graph of cruelty by husbands and relatives plotted year wise, it shows that the number of crimes is increasing as the year increases. In 10 years (2001-2010), the crime has almost doubled. Four lakh cases were registered in the period of 1 year (2010) only and it shows how progressive our nation is in terms of crime rates on women.

4.9.2 **Dowry**:

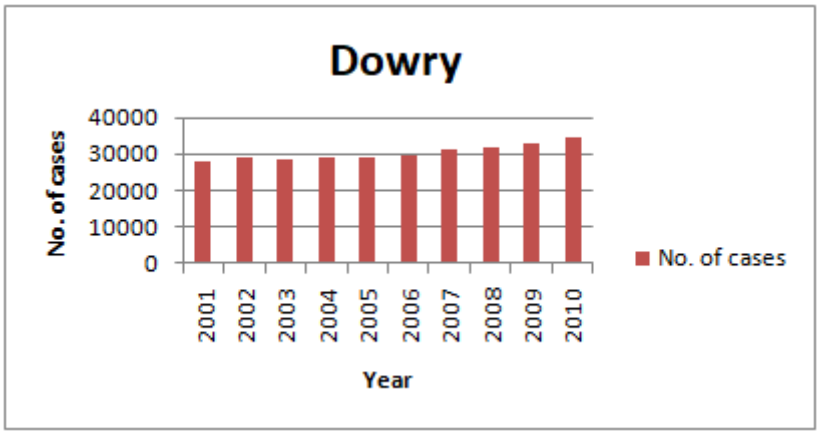


Figure 12. Year wise analysis of cruelty by dowry

In this graph of Dowry cases registered in 10 years (2001-2010), there is no significant change in the number of cases in 10 years. The cases registered are around 30 thousand which is still less as compared to other crimes on women.

4.9.3 **IMMORAL TRAFFICKING:**

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

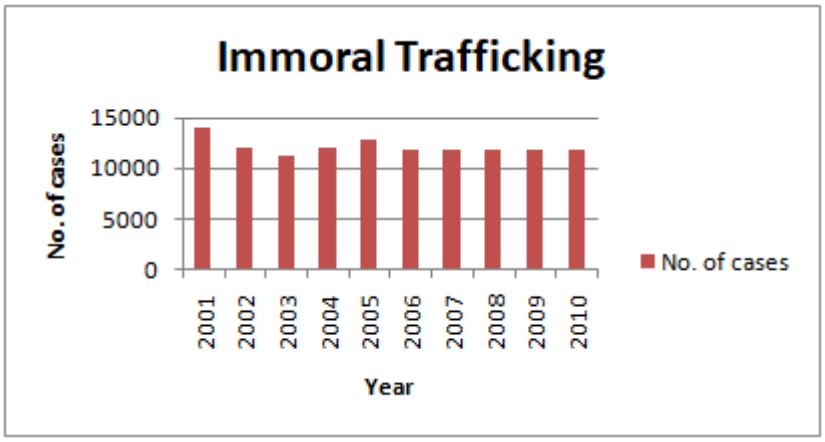


Figure 13. Year wise analysis of Immoral trafficking

The graph of Immoral trafficking is different with respect to other crimes prevailing in India. 2001 recorded the highest number of cases for immoral trafficking of girls and women of India and was close to 15 thousand cases. 2003 recorded the lowest number of cases in Immoral Trafficking.

4.9.4 **Kidnapping**:

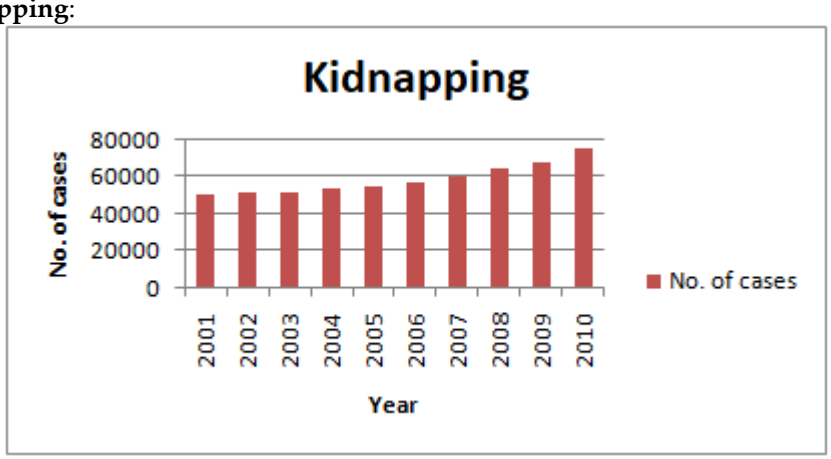


Figure 14. Year wise analysis of Kidnapping

The graph of Kidnapping crime on women per year (2001-2010) is also increasing as shown. The highest crime recorded is 80 thousand which means that there were on an average 219 cases that were registered each day within India in the year 2010.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

4.9.5 **Molestation**:

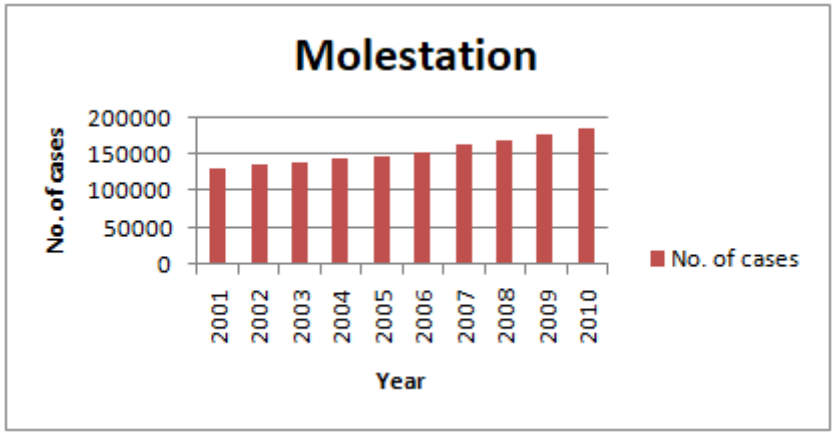


Figure 15. Year wise analysis of Molestation

The graph of Molestation crime on the females of India from 2001-2010 is shown in the above graph. The number of cases of Molestation is also increasing as the year increases. 2010 recorded the most number of cases, close to 2 Lakhs. Molestation is the second most common crime existing in India after Cruelty by husbands and relatives.

4.9.6 **RAPE** :

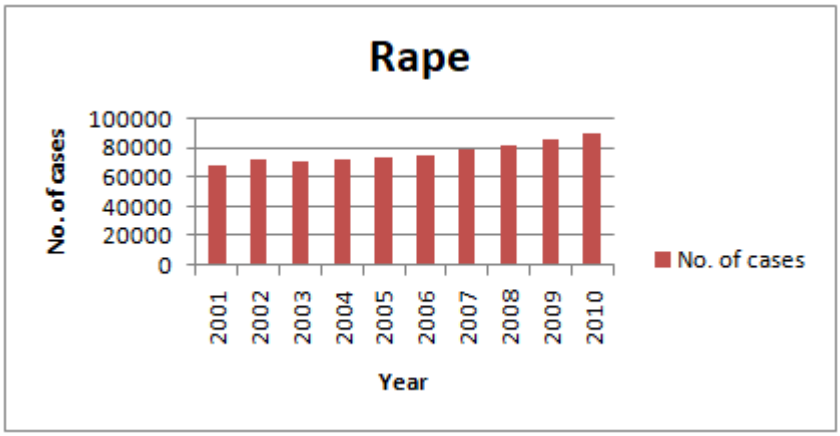


Figure 16. Year wise analysis of Rape

The graph of Rape cases year-wise from 2001-2010 shows an increase in the number of crimes as the year changes. 2010 recorded around 1 Lakh cases which means that on average 273 cases were registered per day. 2001 recorded the lowest number of cases which is still close to 70 thousand which means that in 2001 also, on average 191 cases were registered each day.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

4.9.7 SEXUAL HARASSMENT:

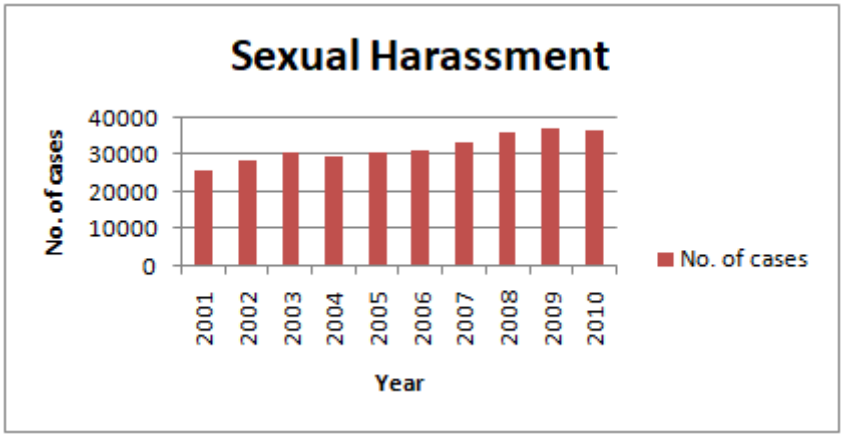


Figure 17. Year wise analysis of Sexual Harassment

Sexual Harassment graph of year-wise from 2001-2010, shows that 2009 recorded the highest number of cases close to 40 thousand. In the year 2010, there was seen a little drop in the number of cases of Sexual Harassment. 2004 also recorded less number of cases as compared to 2003. So, this graph has shown ups and downs in the number of cases registered.

Overall, As India is progressing and annual Indian GDP was about 10.4 in the year 2010, it is expected and desired to control down the rates of these crimes to a minimal level. The rate of minimalism may depend upon the factors like literacy, awareness about crime, mental health, population etc. However, the graph shown below negates the facts of progressive growth of India. In a similar context, the reason for increased growth in women crimes in India is another topic of research and does not come under the scope of our study.

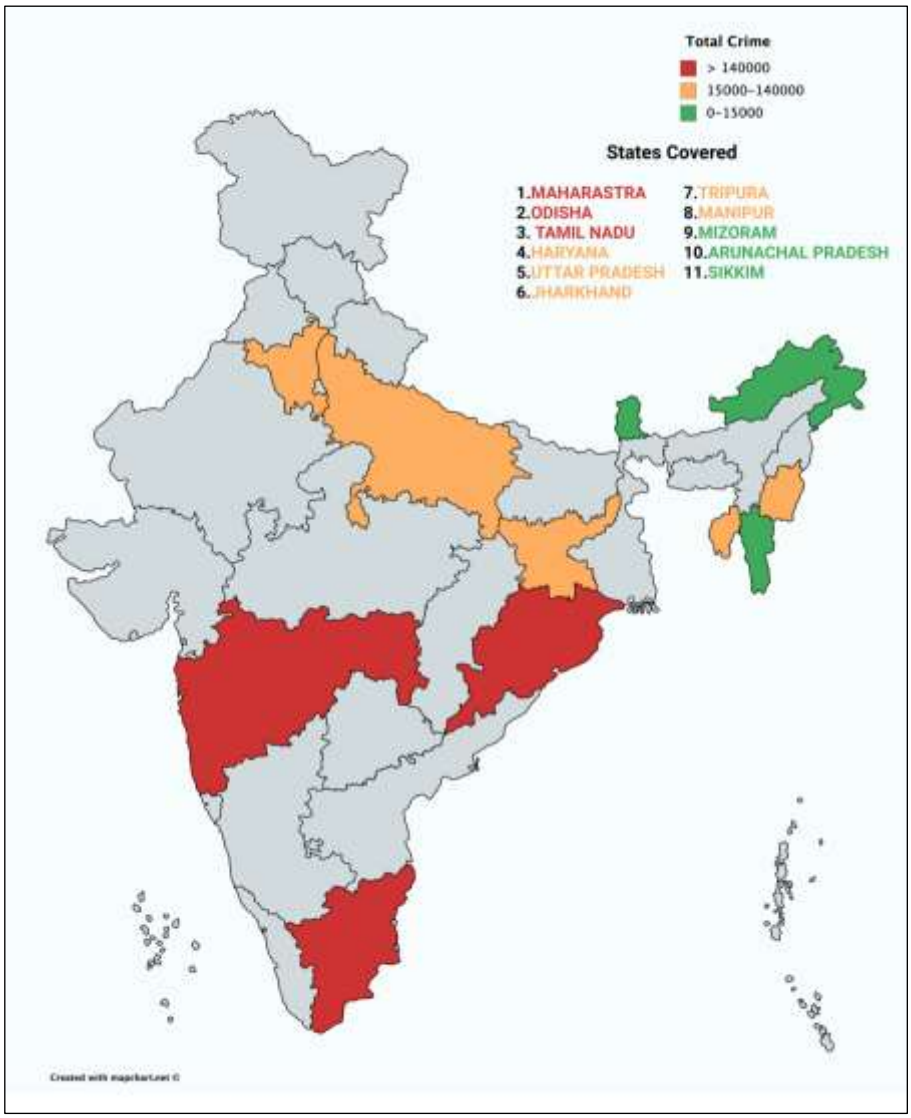
4.10 Map Analysis into Zones

From the data provided by the National Crimes Record Bureau (NCRB) we have performed the following clustering analysis on the data. We now have the data which is about women crimes in different states and the data is from the year 2001-2011. Through the data we have categorized the data into three zones depending on the total crime of the states.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

The red zone represents states having crime index greater than > 1,40,000. The yellow zone represents states having crime index ranging from 15000 - 1,40,000. The green zone represents states having crime index ranging from 0 - 15000. The map will give you a clear idea of which state has a higher number of crime rates and which states have lower. The given data is displayed according to population census state count.

Figure 18. Indian Map view of crimes on women (Zone wise).



5. Conclusion

K-means clustering is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. It is an effective approach and used worldwide for clustering analysis. The paper aimed at having limited scope of pre-processing the

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

women crime dataset from the overall crime dataset, The Results shown that Maharashtra, Tamilnadu, West Bengal has highest women crime cases than other states of India.

6. Future Scope

Though, there is enough clarity on the women crime and its impact on the society through clustering analysis, there is still future work which needs to be addressed on the same dataset. The future research is therefore described as follows:

1. Using the same dataset, the classification and regression analysis can be done for further predictions. The advanced technologies like ML and AI can be helpful in predicting crime analysis.
2. The women crime dataset can be added with population density dataset and literacy rates of the individual states in India. Appropriate statistical tests can be conducted to find whether literacy can be one of the reasons for committing these kinds of crimes.
3. The clustering analysis can also be taken into further advancements such that clusters of crimes can be compared with other countries.

7. References:

[1] Mangoli, R. N., &Tarase, G. N. (2009). Crime against women in India: A statistical review. International Journal of Criminology and Sociological Theory, 2(2).

[2] Patel, A. B. (2020). Crime against Elderly Women in India. In Frailty in the Elderly-Physical, Cognitive and Emotional Domains. IntechOpen.

[3]Hassani, H., Huang, X., Silva, E. S., &Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139-154.

[4] David, H., &Suruliandi, A. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. ICTACT journal on soft computing, 7(3).

[5] Prabakaran, S., & Mitra, S. (2018, April). Survey of analysis of crime detection techniques using data mining and machine learning. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012046). IOP Publishing.

[6] Yerpude, P., &Gudur, V. (2017). Predictive modelling of crime dataset using data mining. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.

[7] Na, S., Xumin, L., &Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics (pp. 63-67). IEEE.

[8] Xie, J., Girshick, R., &Farhadi, A. (2016, June). Unsupervised deep embedding for clustering analysis. In International conference on machine learning (pp. 478-487).

[9] Motwani, M., Arora, N., & Gupta, A. (2019). A study on initial centroids selection for partitional clustering algorithms. In Software Engineering (pp. 211-220). Springer, Singapore.

[10] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems (pp. 849-856).

[11] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern recognition, 36(2), 451-461.

[12] Agarwal, J., Nagpal, R., & Sehgal, R. (2013). Crime analysis using k-means clustering. International Journal of Computer Applications, 83(4).

[13] Malathi, A., & Baboo, D. S. S. (2011). Algorithmic crime prediction model based on the analysis of crime clusters. Global Journal of Computer Science and Technology, 11(11), 47-51.

[14] Bsoul, Q., Salim, J., & Zakaria, L. Q. (2013). An intelligent document clustering approach to detect crime patterns. Procedia Technology, 11(1), 1181-1187.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020

DOI: 10.5281/zenodo.3786135

[15] Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., & Tyagi, N. (2015). Crime detection and criminal identification in India using data mining techniques. AI & society, 30(1), 117-127.
[16] Thota, L. S., Alalyan, M., Khalid, A. O. A., Fathima, F., Changalasetty, S. B., &Shiblee, M. (2017, March). Cluster based zoning of crime info. In 2017 2nd International Conference on Anti-Cyber Crimes (ICACC) (pp. 87-92). IEEE.
[17] Thongsatapornwatana, U. (2016, January). A survey of data mining techniques for analyzing crime patterns. In 2016 Second Asian Conference on Defence Technology (ACDT) (pp. 123-128). IEEE.
[18] Marwah, S. (2014). Mapping murder: Homicide patterns and trends in India. Journal of South Asian Studies, 2(2), 145-163.
[19] Prabakaran, S., & Mitra, S. (2018, April). Survey of analysis of crime detection techniques using data mining and machine learning. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012046). IOP Publishing.
[20] Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science, 85, 662-668.
[21] Chaudhari, B., & Parikh, M. (2012). A Comparative Study of clustering algorithms Using weka tools. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 1(2), 154-158.

8. About Authors

Rishabh Singh is a student in SVKM's NMIMS Mukesh Patel School of Technology Management and Engineering (MPSTME), where he is pursuing MBA in Technology Management (Minor in Computer Engineering). At a young age of 19, he recognized his interest in the field of research and has been working towards it continuously. The domains he's keen about are Data analysis, Security and Education Technology. Rishabh was part of the official technical committee in his college for two years, being head of the Informal's department of the Technical Festival in both years. This has helped him gain experience and skills, both technical and leadership over a period. Along with technical accomplishments he has ensured his contributions towards the betterment of the society by being the head of Hospitality in one of the best social conferences in the world, Social Conclave. He believes in learning beyond boundaries and is always keen in gaining knowledge that is not in his curriculum and part of a different field

Rishabh Reddy is a student in SVKM's NMIMS Mukesh Patel School of Technology Management and Engineering (MPSTME), where he is pursuing MBA in Technology Management (Minor in Computer Engineering). He is a wholesome person, one capable of prioritizing his work and leisure when needed, and adequately equipped to be a resolute and laser-eyed leader and a willing team-player too. He has a keen interest towards research domain, and he believes that researching helps to gain knowledge immensely. Apart from this He is the Secretary of the Students' Council's Social Impact Committee, while also being the Vice-Chairperson of the college cultural festival, Sattva. Having been an advocate for change and consequence in the modern society, and having always been aware of the dangers and problems faced by it, He has taken part in and led various social initiatives in his college years, namely The Social Conclave, the aforementioned Social Impact Committee, the Serve Out Smiles(SOS) campaign, which aimed at harnessing care and shelter for orphaned pets living in the city, and the annual Blood Donation Drive in college, which saved up to 1300 lives a year.

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135

Vidhi Kapoor is undergraduate student pursuing Bachelors of Technology in the field of Data Science from NMIMS' School of Technology Management and Engineering. She is a very hardworking and enthusiastic student, always ready to learn and explore new things. She has interned at various companies to learn and apply her knowledge in industry projects. She has successfully been able to maintain a good balance between academics and extra-curricular. She is a part of various college bodies and has been appointed as the Joint Secretary of the Business and Management Cell of college. She has held positions of Sub Head in Cultural and Social festivals of college. This has helped her in her overall development including team building and leadership qualities. She has a keen interest towards research particularly in the domain of analysis, Machine Learning and Deep Learning and she has been continuously working towards it.

Prathamesh Churi is Assistant Professor in School of Technology Management and Engineering, NMIMS University. He is also PhD research scholar in Symbiosis International University, India. He is Associate editor of International Journal of Advances in Intelligent Informatics. He is actively involved in peer review process of reputed IEEE and Springer journals. He has been a keynote speaker, chair, convener in the international conferences. He has recently received "Best Young Researcher award" by GISR Foundation for his research contribution in the field of Data Privacy and Security, Education Technology. He is active leader, coach, mentor, volunteer in many non-profit organizations. He is also involved as board of study member in many universities for curriculum development and educational transformations. He has over 40+ research papers in International Journals and conferences

 Received: March 15, 2020 Revised: May 17, 2020 Accepted: June 23, 2020
DOI: 10.5281/zenodo.3786135