**DATA VISUALIZATION (REVIEW-2)**

**Name: Harjyot Singh Bagga**

**Reg. no.: 18BCE0694**

**Faculty:** PROF. ANNAPURNA JONALGADDA

**Slot:** F2

**Course Code:** CSE3020

**Title**

**Visual Analysis of Rape Culture in India**

**Team Members:**

i. Shivankar Bansal (18BCE0533)

ii. Hemkar Goswami (18BCE0664)

iii. Harjyot Singh Bagga (18BCE0694)

**Problem Statement**

Analysing and Visualising rape culture in India

**Relevance of Problem**

Rape is the fourth most common crime against women in India. Rape in India is a very serious issue and needs to be addressed with utmost importance and urgency and should be stopped as quickly as possible.

So, in this project we will analyse data from dataset provided by National Crime Records Bureau (N.C.R.B) to figure out the prominent states and age categories which are most adversely affected with rape culture in our country.

In later stages we will visualise data with ‘plotly’ package and clearly figure out which are the regions where rape occurs frequently. Despite the fact that rape is regarded heinous and criminal in nature, the number and the level of inhumanity of this crime has been on a rise. Therefore, this subject matter is very important and typical in nature.

**Outcomes**

To address our problem statement we found datasets regarding rape cases on Quandl, Kaggle, and even the government website and merged them to one single dataset. In further stages we cleaned data by removing redundancies and repetitive entities. Further, we segregated datasets into unreported and reported cases and then separately analysed and visualised them. Then for visualisation we first visualised using bar plots, line plots and then later on used heat-map.

From this project we will be able to predict future such cases that can happen with help of classifier that will train and test our dataset.

**Detailed literature survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TITLE** | **AUTHOR** | **DATASET** | **REFERENCE** | **DESCRIPTION** |
| The problem of rape in India: a multi-dimensional analysis | Radha R. Sharma,  Rupali Pardasani,  Sharda  Nandram, | * http://ncrb.nic.in/CD-CII2011/cii-2011/Snapshots-5311.pdf (accessed 11 January 2013). * http://ncrb.nic.in/CD-CII2011/cii-2011/Chapter%205.pdf (accessed 11 January 2013) * http://blogs.wsj.com/indiarealtime/2013/01/04/statistics-conviction-rates-for-rape-acrossindia/tab/interactive/ (accessed 11 January 2013). | <https://www.emerald.com/insight/content/doi/10.1108/IJMPB-10-2013-0061/full/html> | The purpose of this paper is to analyse the problem of “Rape” in India from multiple perspectives and recommend measures for mitigating this crime from the country. Using the much-highlighted incident of gang rape of a 23-year-old woman in Delhi, India on 16 December 2012, the paper analyses the behaviour of the various parties involved in the case with the help of some sociological and psychological theories |
| Gender-based violence in New Delhi, India: forecast based on secondary data analysis | Nidhi Dwivedi,  Sandeep Sachdeva | Crime in India. Annual Reports: 2009–2015. New Delhi: National Crime Record Bureau, Ministry of Home Affairs, Government of India; 2015. | <https://apps.who.int/iris/bitstream/handle/10665/327244/EMHJ_25_4_2019.pdf#page=44> | Reported cognizable crime against women in New Delhi for 2009–2015 was extracted for statistical analysis, synthesis and modelling. The cognizable crimes reported are rape, attempt to commit rape, kidnapping and abduction, dowry deaths, assault on women with intent to outrage her modesty, insult to modesty of women, cruelty by husband or his relative, importation of girls from foreign countries, abetment of suicide of women and indecent representation of women. Within limitations, it is concluded that the safety of women will continue to be a concern in the near future. |
| Crimes against Women in India: Analysis of Official Statistics | Chandan Mukherjee Preet Rustagi N. Krishnaji | National Crime Records Bureau (NCRB). | <https://www.jstor.org/stable/4411293?seq=1> | This paper is an attempt to see what official, published data reveal, whether there are clear-cut regional patterns and if so whether they can lead to meaningful hypotheses for future work. |

**Contribution from each member of the team**

**Hemkar:** Literature Survey, collecting reported cases data, training-testing more data.

**Harjyot:** Major Implementation such as Initial Analysis, further analysis, comparative results, building a classifier.

**Shivankar:** Abstract, collecting unreported cases and testing visualisation, increasing efficiency of algorithm.

**Schedule of the project**

1. **Abstract and Literature**: 10 days

2. **Implementation**

a. **Initial Analysis:** 5 days

b. **Collecting reported and unreported cases:** 7 days

c. **Further Analysis and basic visualisation:** 4 days

d. **Testing visualisation on datasets of states:** 4 days

e. **Comparative Results obtained:** 6 days

f. **Building classifier:** 2 days

g. **Training and testing:** 2 days

h. **Forecasting and testing accuracy: still working** (7 days approx.)

***It took us lot of time to find recent datasets for security reasons recent cases are not released until they are convicted fully with proper closing just for eg: Nirbhaya case got closed recently so even full data is not available regarding this case in most of websites.***

***Also the government does not release sensitive data until 10 years of closing of the cases on their website, for privacy and security reasons. This is the very reason all the analysis have been done on the data between 2001 to 2010.***

**Technology used in project**

We have used *Linear Regression Algorithm* in our classifier to *train* and *test* the datasets which is a part of **Machine Learning**. This training and testing helped us classify between reported and unreported cases and then we went to further visualise them separately. Moreover, we have built classifier model to forecast and predict future cases.

**Explanation of Algorithm**

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

**Hypothesis function for Linear Regression:**

y = mx + b

While training the model we take:  
x as input variables of the training data   
y as labels to data corresponding to the variables

When training the model – it *fits the best line* to predict the value of y for a given value of x.

We achieve this by determining the best-fit regression line. In other words, the model aims to predict y value in such a way, that the net error difference between predicted value and true value is least. The error function taken in count is the ‘R-squared error function’.

**Prototype of the project**

Our building prototype uses top-down approach starting from data collection to forecasting data which are as follows:

* Data collection
* Data analysis
* Data cleaning
* Data Encoding
* Using various visualisation techniques
* Choosing linear regression algorithm to train and test datasets
* Building classifier model
* Training model
* Testing model
* Forecasting and predicting with help of classifier

**Implementation aspects**

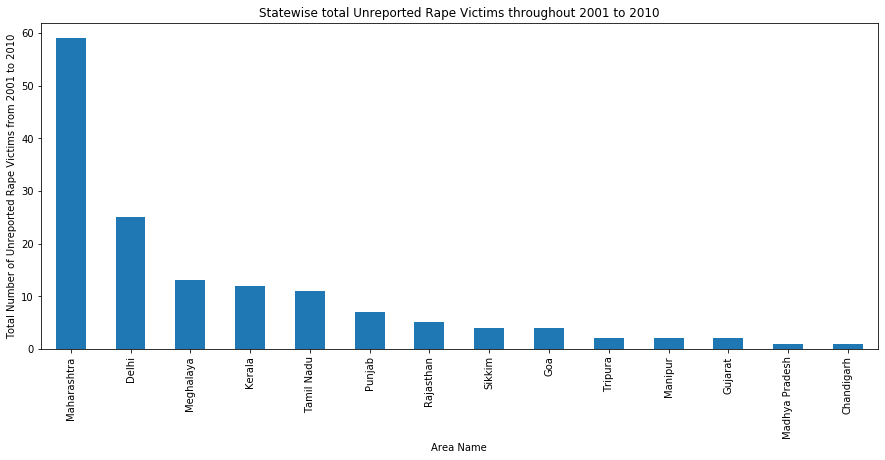
**Libraries used are:**

* **numpy:** mathematical library of python
* **pandas:** for analysis of data
* **seaborn:** for data visualisation for further visual effects
* **matplotlib:** for data visualisation
* **tabulate:** to represent data in tabular format

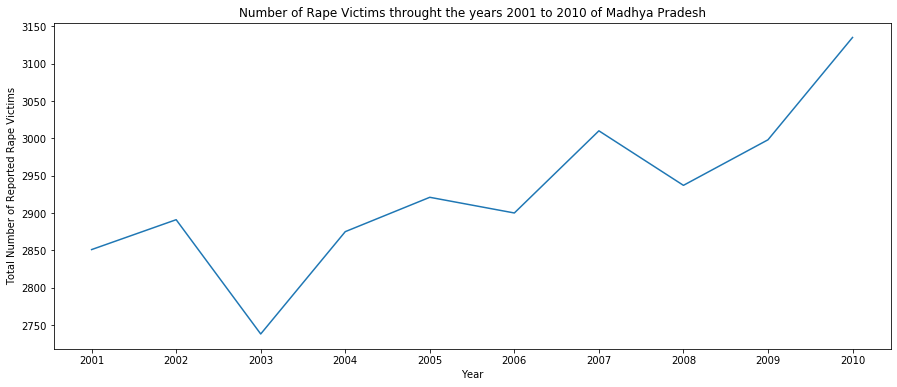
**Dataset & Coding Implementation**

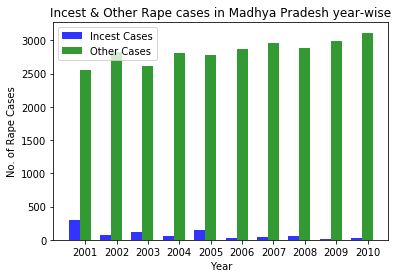
The .csv file and the jupyter notebook is attached in the zip file with this project.

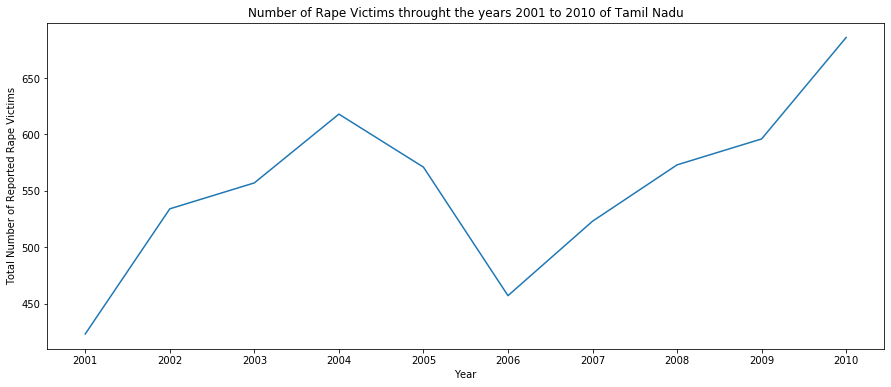
**Visualization of the data**

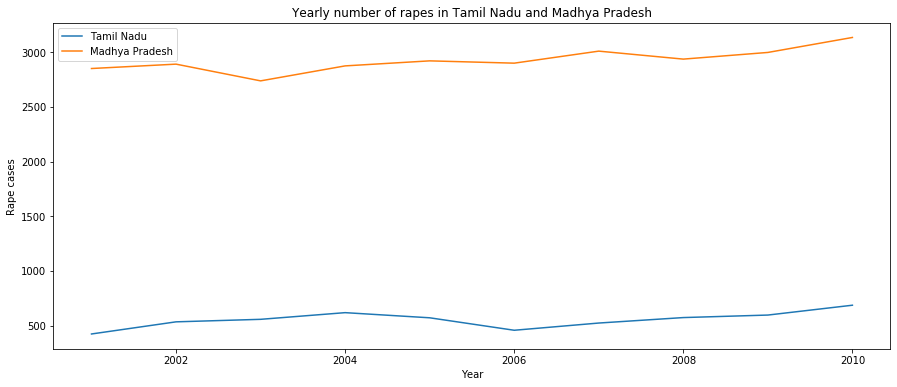
****

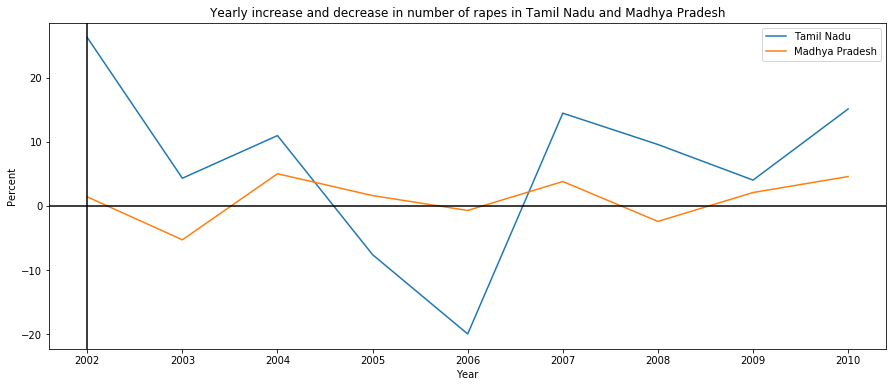












**Challenges faced in the implementation**

* A lot of cases aren’t recorded and thus not available officially on the website, thus finding them on official govt websites were difficult. Thus had to be taken from various other sites like Kaggle and Qunadl. Such sensitive data had to be confirmed to be legit before using them.
* There are some cases that have been unreported and some cases that have been reported so we will represent them separately.
* Due to the continuous changing of govt laws, it was initially difficult to categorize data.
* Earlier marital rape wasn’t considered, then when laws changed, they had to be incorporated in incest rape.
* It took us lot of time to find recent datasets for security reasons recent cases are not released until they are convicted fully with proper closing just for e.g.: Nirbhaya case got closed recently so even full data is not available regarding this case in most of websites.
* Also the government does not release sensitive data until 10 years of closing of the cases on their website, for privacy and security reasons. This is the very reason all the analysis have been done on the data between 2001 to 2010.

**Conclusions**

A lot of conclusions have been derived by visualization.

* Madhya Pradesh has the most number of rape cases.
* Maharashtra has the most number of unreported cases.
* The rapes of girls of 18-30 years are more prominent.
* The crime rates of Madhya Pradesh have been actively growing and have reached a count of 3150 in 2010.
* But the rate of incest rapes in Madhya Pradesh has gone down since the laws become stricter.
* With the change of laws becoming stricter and people becoming more and more aware the rate of these crimes are decreasing.
* We compared the rape crimes in Tamil Nadu with the rates of Madhya Pradesh and have seen the number of crimes in Tamil Nadu is way lower with respect to Madhya Pradesh.
* However, the crime rate in Tamil Nadu is has increased in Tamil Nadu with respect to Madhya Pradesh at a higher rate.

**Percentage of work completed so far**

We have completed around 70% work. Few things in classifier is still remaining and we need to even refine our algorithm to achieve closer results in prediction. Major work still remains on forecasting and predicting more data. We are still trying to get recent datasets but keeping in mind security reasons we will try to ensure that we work on latest dataset available to us in best way possible. Also certain aspects like predicting and forecasting data, and then using that data to draw conclusions is yet to be done.

**Verification and Testing aspects**

We have analysed as well as visualised data of 2 states that is Madhya Pradesh and Tamil Nadu separately:

Madhya Pradesh has the most number of rape victims throughout the time span, so but the crime rate in Tamil Nadu is higher.

Our visual encoding and visualization techniques have been verified as the techniques correspond to the data and also the trend and pattern that the data follows.

**Results Obtained & Validation**

We have compared above between two states to determine which one of them is safe. Also, all the visual encoding techniques have depicted the current scenario, now we will train the classifier to predict and forecast data based on the trend/pattern the data exhibits.

Some results that have been obtained and the conclusions drawn are mentioned above. Also the dataset collected corresponds to the drawn conclusion and follows the same pattern as observed by the visual encoding techniques.