Analysis of Human Trafficking Crime Statistics Across India

Srihari Reddy Mekala
Dept. Computer Science
2017A7PS1215H
Birla Institute of Technology and Science, Pilani - Hyderabad Campus

Manas Reddy
Dept. Computer Science
2017A7PS0191H
Birla Institute of Technology and Science,Pilani - Hyderabad Campus

***Abstract*—The project which we have designed and chosen makes use of data mining techniques in analyzing and visualizing human trafficking statistics recorded across India.It analyzes the trends in the databases to bring awarness among us on how the current state of crime is in every state of India.And what should be the next step in tackling these activities.**

Keywords—techniques,preprocessing,Data visualization,plots

❖ Introduction

With crime on a steady rise in India.It becomes a great challenge and responsibility to keep a record of these data,analyze it and use it to infer the necessary.The necessary to know which states in India have high crime record numbers.The necessary to know which states have increasing criminal activity averages.The necessary to know which states in India are safer and better places to live in.The very state of safe and better which we wish all of India to be in.The very objective of our project. The project which we have designed and chosen makes use of these data mining techniques in analyzing and visualizing human trafficking statistics recorded across India.On whose inferences we make many conclusions.Conclusions like which states in India need to step up there efforts in trying to reduce crime or if certain state have strong reducing crime averages over the years we can use that to see what certain policy or policing methods were used in those states that can be implemented in other states with higher crime averages.In these and in many other ways we can use Data Mining to make an India a better India.

❖ Data preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format
Steps involved in data preprocessing:

1. Data cleaning

- Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.
Some of them are:

a. Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

b. Fill the Missing values:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

- Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

a. Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segment is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

b. Regression:

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

c. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation

This step is taken in order to transform the data in appropriate forms suitable for the mining process. This involves following ways:

- Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

- Discretization:

This is done to replace the raw values of numeric attributes by interval levels or conceptual levels.

- Concept Hierarchy Generation:

Here attributes are converted from level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3. Data Reduction

Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

a. Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

b. Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

❖ Description of Dataset

There are 2 given Datasets each of the year 2012 and 2013.Each Dataset consists of about 20 Attributes like State/UT,Crime Head,Cases pending Investigation from previous year,Cases reported during this year,Cases gone to Trial,Cases pending Trial from last year,Cases reported last year and many more.Further more the attribute Crime Head is of 6 types,each one giving the name of Crime committed namely :

1. Immoral Traffic(Prevention) Act,
2. Buying of Girls for prostitution(Section 373 IPC)
3. Selling of Girls for prostitution(Section 372 IPC)
4. Procuration of Minor Girls(Section 366-A IPC)
5. Importation of Girls from Foreign Country(Section 366-B IPC)
6. Human Trafficking

These crimes are noted and figures are given for each State/UT.There are 35 States/UT in which there are 28 states and 7 Union Territories(UT).

This dataset and its information is briefly described in Data Preprocessing and Cleaning.With functions like dtypes(),describe(),shape etc used to describe the data in various ways.After we have our Data we move on to Data Visualization to look into trends and patterns to conclude a greater meaning from it.

❖ Creating Plots using Python

If you are looking for a powerful way to visualize geographic data then we should use interactive Choropleth maps. A Choropleth map represents statistical data through various shading patterns or symbols on predetermined geographic areas such as countries, states or counties. Static Choropleth maps are useful for showing one view of data, but an interactive Choropleth map is much more powerful and allows the user to select the data they prefer to view.

The interactive chart here provides details on Name of State,Name of the Crime,Year in which the data was recorded.The chart breaks down Number of Cases registered,Number of Cases withdrawn,Number of Cases gone to trial,No of Cases pending trial,Number of Cases pending from last year,No of Accused convicted,No of Accused acquitted.

Let's start with the installs and imports we will need for the graphs. Pandas, numpy and math are standard Python libraries used to clean and wrangle the data. The geopandas, json and bokeh imports are libraries needed for the mapping.

Bokeh offers several ways to work with geographical data including Tile Provider Maps, Google Maps and GeoJSON data. We will be working with GeoJSON, a popular open standard for representing geographical features with JSON. JSON (JavaScript Object Notation), is a minimal, readable format for structuring data. Bokeh uses JSON to transmit data between a bokeh server and a web application.

In a typical Bokeh interactive graph the data source needs to be a ColumnDataSource. This is a key concept in Bokeh. However, when using a map you use a GeoJSONDataSource instead.
To make our work with geospatial data in Python easier we use GeoPandas. It combines the capabilities of pandas and shapely, providing geospatial operations in pandas and a high-level interface to multiple geometries to shapely. We will use GeoPandas to create a GeoDataFrame — a precursor to creating the GeoJSONDataSource.Finally, we need a map that is in geojson format.

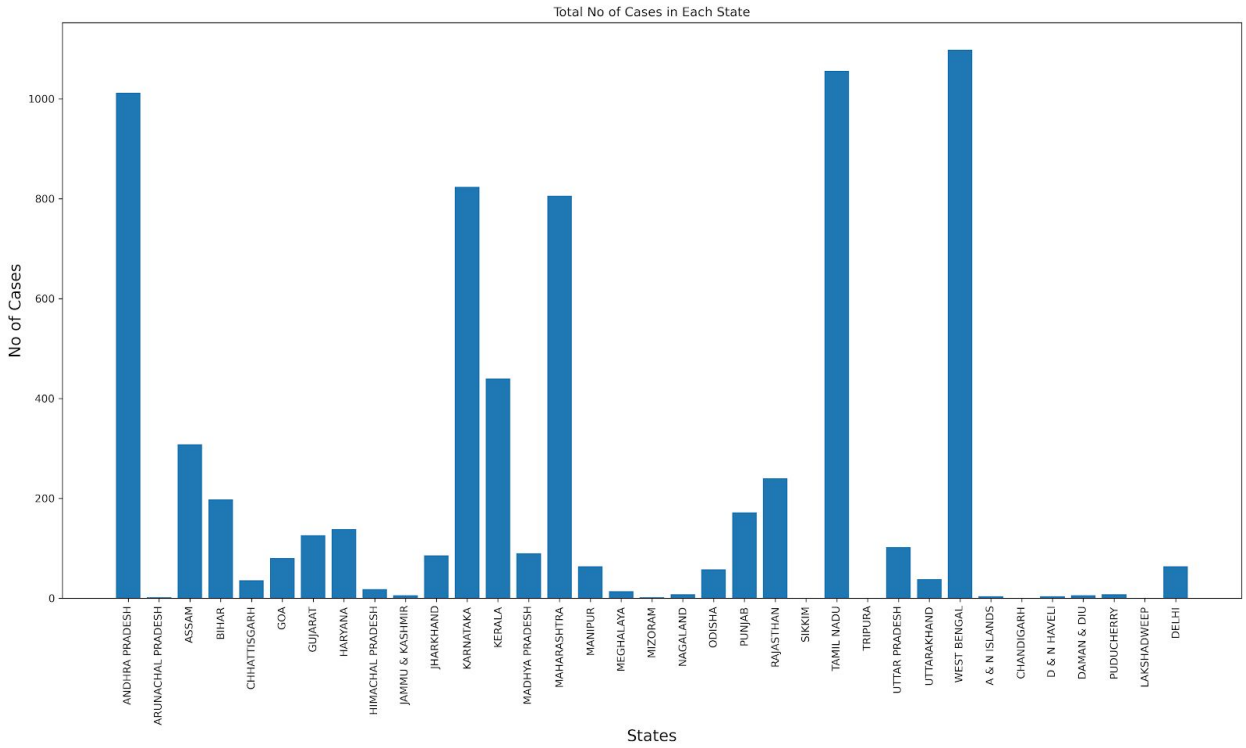
We use geopandas to read the geojson map into the GeoDataFrame sf. We then set the coordinate reference system to lat-long projection. Next, we rename several columns and use set_geometry to set the GeoDataFrame to column 'geometry' containing the active geometry.
We now need to create a function that merges our neighborhood data with our mapping data and converts it into JSON format for the Bokeh server.We still need several pieces of code to make the interactive map including a ColorBar, Bokeh widgets and tools, a plotting function and an update function.

After putting this all together we printout out a static map with the ColorBar and HoverTool in the Colab notebook to bring out our desired plot.

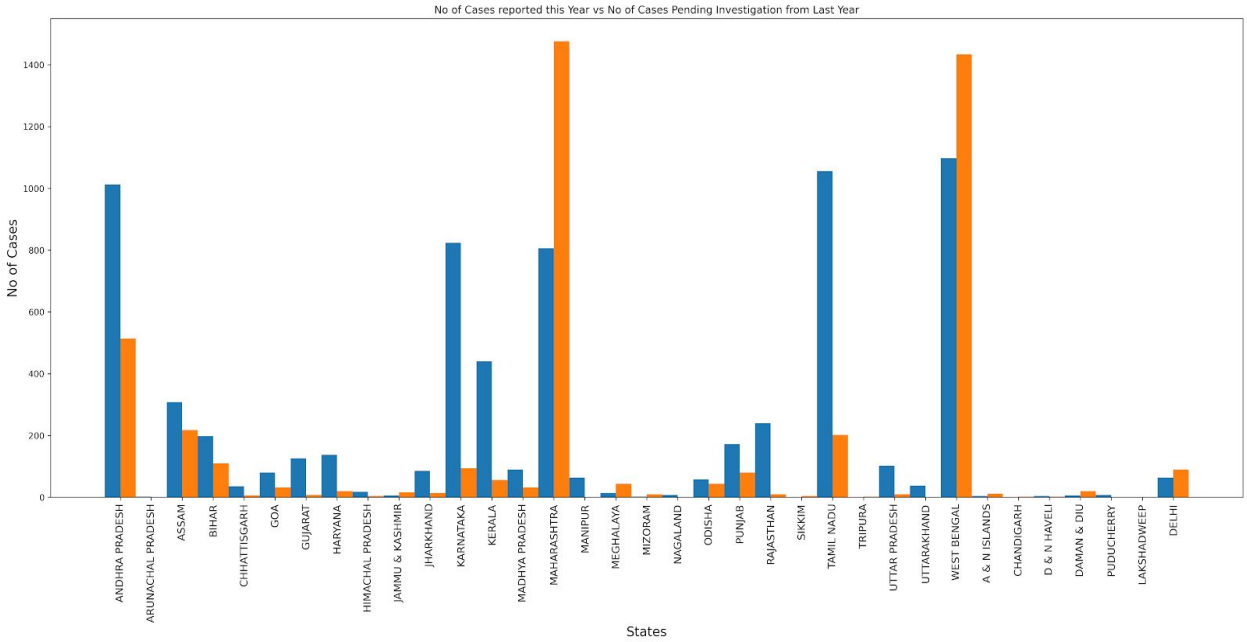
❖ Data Visualization of the Processed Data

Given such vast data,it becomes a great challenge for us to analyze it and look for important observations.After the anticipated preprocessing was done,Initial visualizations were made in the form of boxplots, bar graphs, pie charts, scatter plots to better understand the processed data. The plots are described below with their following observations.

1.Bar Graph Plot to observe the Total No. of Cases in each state in India.By observing this we can ascertain the basic necessity in crime fighting which is knowing which state has more crime and which state does not.Which state has done better and which has not.

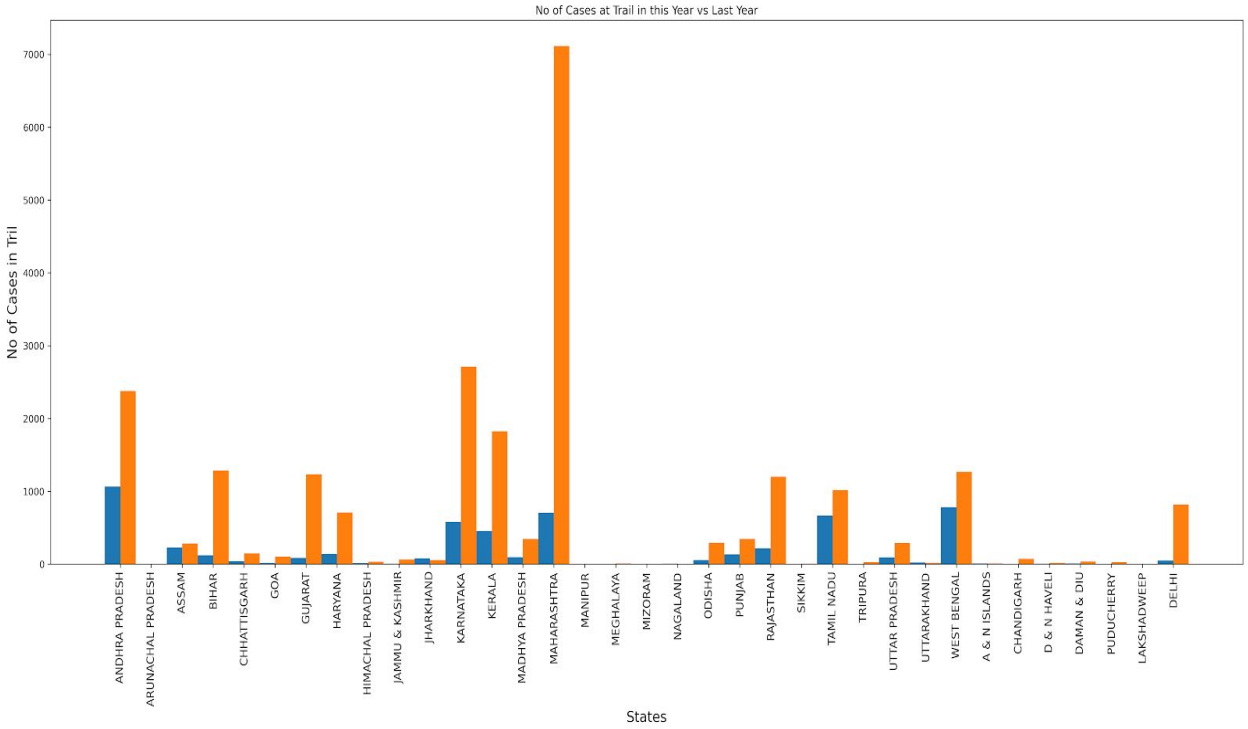


2.A comparison bar chart to see how was the situation last year vs how is the situation this year.Whether enough effort and resources were put into improve the situation or not.To find this I have Used a Comparative Bar Graph Chart on No. of Cases Pending Investigation this Year vs No of cases Reported this year.If there is significant Numbers observed in this graph it becomes clear that the police force in our country has become incompetence and inefficient in tackling Crime.

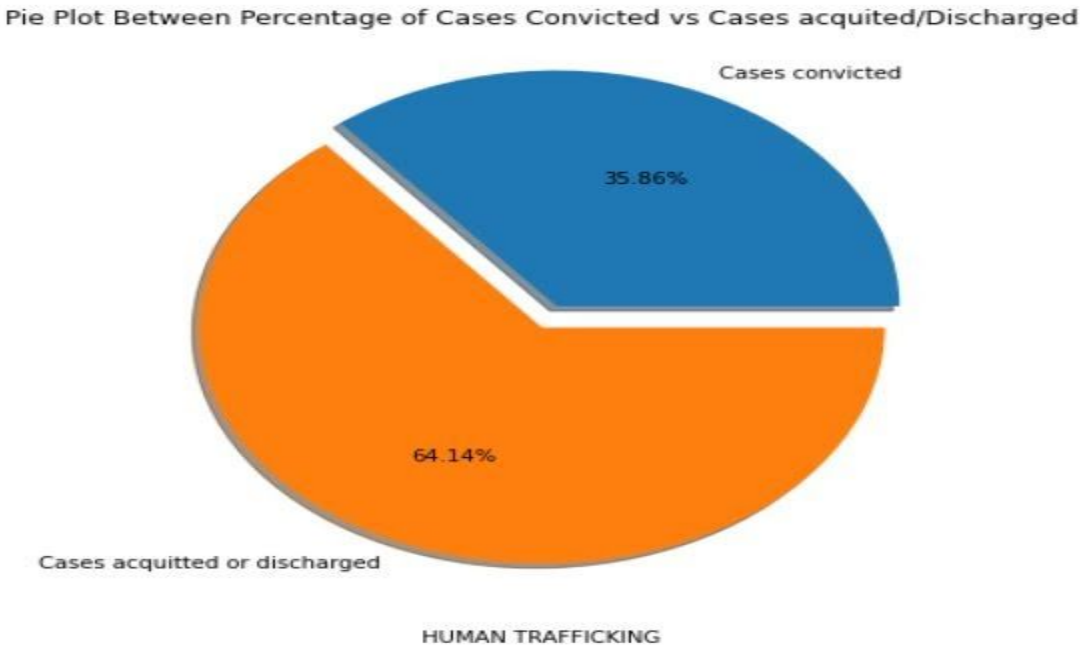


3.With a glance at Crime statistics given I wanted to move onto The Judiciary state in Fighting Crime.The graph that I followed after these are made up of statistics indicating,how did the

judiciary of India fare in fighting crime.Because it is as important to bring Criminals to just as important as is to arrest them.To get an understanding of this I have plotted a Comparative Bar Graph to see No. of Cases pending Trial from last year against this year.



4.Pie Chart was then used to see on all those cases that have gone to trial how many cases ended up with conviction and how many ended up with the trial result acquitted/discharged.This helps us to know the overall talk in India that the courts have been failing in bringing justice to criminals quickly and properly.



HUMAN TRAFFICKING

REFERENCE : https://github.com/Srihari1Reddy/Data-Mining-Project.git