MTH 765P Project Report

Crime Analysis

Jincy Baby

M.Sc Data Analytics, Queen Mary University of London

Introduction

The data we are dealing with is specifically of the crime against women in different states and Union territories in India over the period from 2001 to 2010. The data was published in Kaggle by Mr. Rajanand Ilangovan and licensed under CC BY-SA 4.0.

Obtaining/Acquiring the Data

The data was downloaded from the Kaggle site and then cross validated using the data available in the site maintained by National Crime Records Bureau, India. The category of crimes was limited inorder to maintain the scope of the project.

Description

The significance of data in the development of a nation as a whole is an undeniable fact and is crucial for economic and social growth of a country. Transparency of data has always been a debatable point and to an extent it can be considered as a risk factor for certain sectors. However, there are specific information that should be open to the public as they are crucial in promoting informed decisions and improving the accountability of the government. Data on crimes is one of such information that is quite needed to understand the safety of the public and to further form reforms to ensure crime prevention.

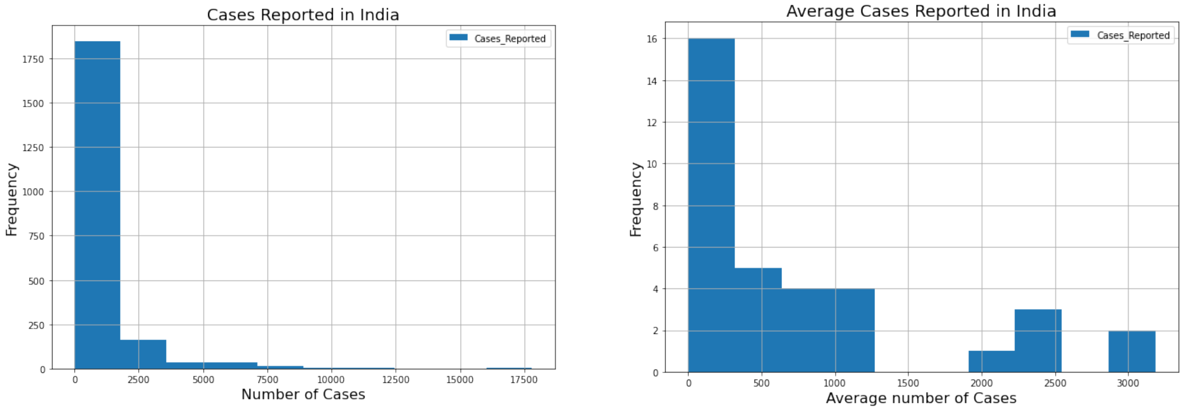
A total of 371,503 cases of crimes against women were registered across the country in 2021 as per the data from the National Crime Records Bureau (NCRB). However, there are a considerably large number of cases with pending trials from the previous years which clearly shows the requirement for developing an efficient judicial system to speed up the processes and ensures the victims access to justice.

In the project, different factors such as number of cases reported, the total cases for trials, number of trials completed, etc are analysed and compared to understand the areas where we need to improve the required facilities as well as try to fit a linear model considering the number of pending trials from previous year and the total cases of trials in the current year.

Analysis

The libraries that we use in the project are numpy, pandas and matplotlib. The csv file is imported as a dataframe 'crime' and the columns 'Area_Name' and 'Group_Name' are renamed as 'Area' and 'Group' respectively. The average of reported cases corresponding to each area is calculated and entered in the dataframe 'mean_crime'.

Histograms for the cased reported in India over the years (1a) and that of the average number of cases reported in each area (1b) are generated. Analysing 1b, it is evident that most areas has an average between 0 and 250 whereas there are two areas with average around 3000, which is the highest.

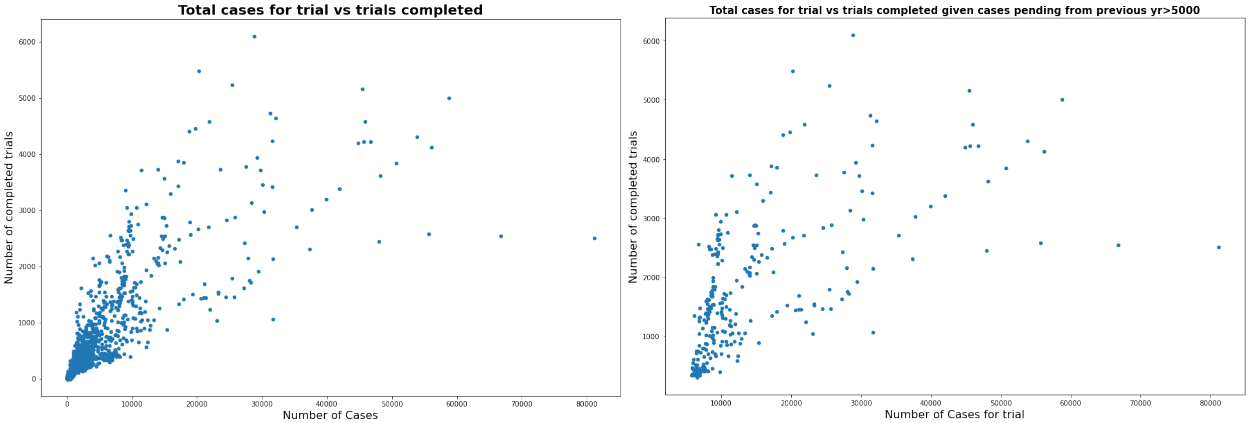


(a)

(b)

Figure 1

The number of completed trials is reviewed based on the total cases for trials. The first figure shows the entire data whereas the second figure considers the data for which the number of pending cases from previous year is greater than 5000.



(a)

(b)

Figure 2

The most important factor which determines the efficiency of a judiciary system is the speed measure. The scatter plots clearly reveal a concerning fact that the total number of cases for trial far exceeds the number of completed trials. There is even an instance where out of approximately 80000 cases, only around 2000 trials were completed. The major factor that contributes to the myriad of cases accumulating may be the direct impact of the cases being delayed and the pending cases from previous years carried on to the next.This is evident in (2b) where the data was filtered on the basis of cases pending from previous year which is greater than 5000.

Boxplots avail the advantage of an effective comparison between states and Union territories based on the number of pending investigation from previous year (Figure 3) and the total number of cases (Figure 4). We consider a dataframe with columns 'Area','Group' and 'Cases_Pending_Investigation_from_previous_year'. Then, we plot the required boxplots.

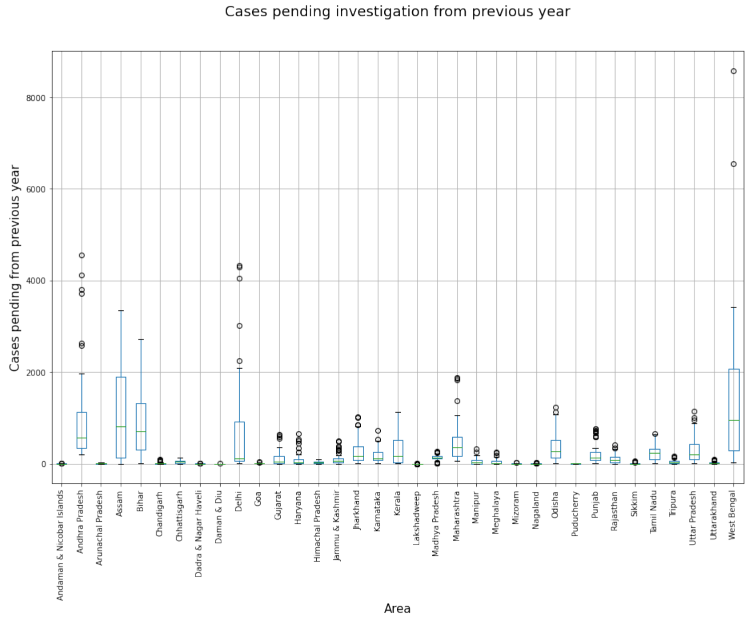


Figure 3

In the above figure, Assam and West Bengal has the highest number of pending investigations compared to other states which establish the scope for improvement and adequate amendments to advance the work force and facilities.

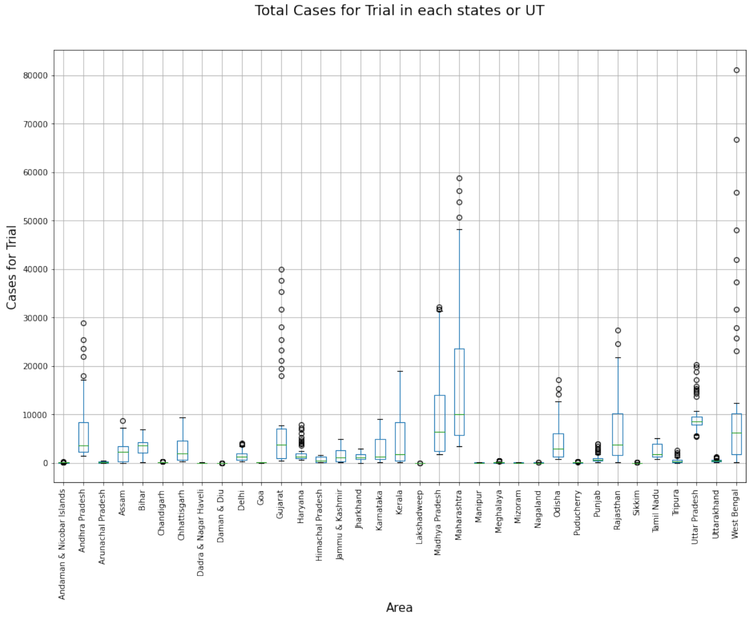
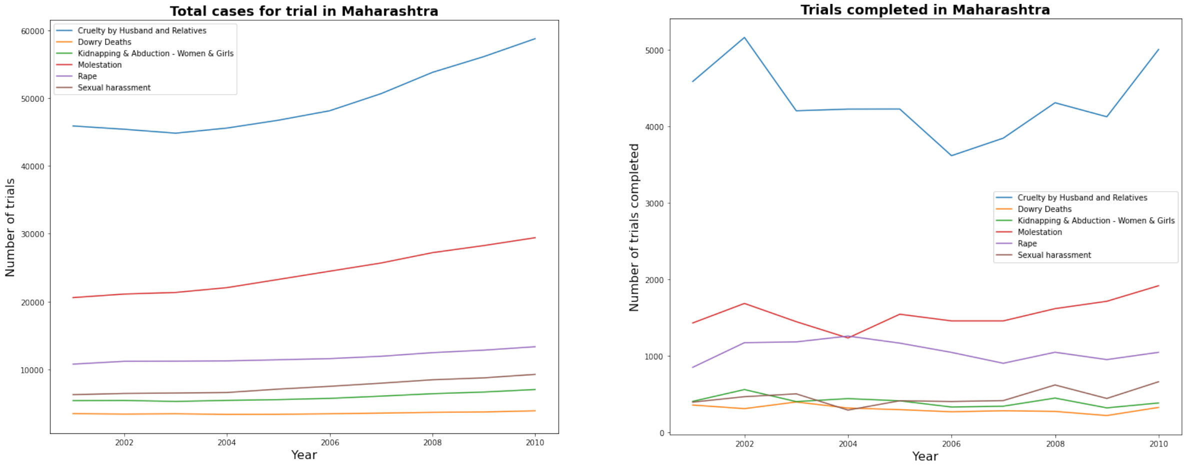


Figure 4

Maharashtra exceeds all other areas when we consider the more densely covered region of total number of cases in the state as displayed in the above plot. However, considering the few instances in West Bengal where it reaches around 80000 in one particular instance is of consequence.

Then, we extract the datas corresponding to 'Maharashtra' and plot the line diagram to analyse the number of cases for trials in each group over the years (Figure 5a), it is apparent that there is quite a substantial difference in the category with the highest number of cases compared to the second highest. The second figure (Figure 5b) shows the number of completed trials over the years for each category.



(a)

(b)

Figure 5

The figure (5a) shows around 45000 to 60000 cases related to cruelty by husband and relatives in each year waiting for the trial. However, it is more shocking to see how trivial the number of completed trials is when compared to the excessive number of cases that is awaiting trial in each year (5b) .

The above aspect indicates the need for prompt action in refining the measures for women empowerment along with introducing public policies or measures to promote positive difference in mentality towards women and to speed up the court proceedings.

Eventhough we have already seen that the total number of trials is higher for the category 'cruelty by husband and relatives' in Maharashtra, it is vital to consider the situation all over the country. Subsequently, the plot to compare the total number of trials in each category is given in the figure (6). The figure indicates that the number of trials is higher for the groups 'cruelty by husband and relatives' and 'Molestation'. As such, the government should consider the policies to be implemented nation-wide as the category where the highest number of cases appears is the one that can be reduced if proper measures and plans are taken into consideration.

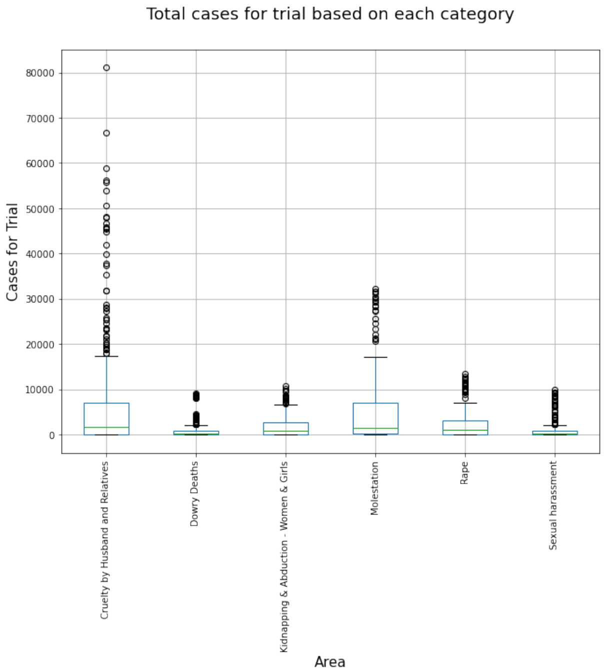


Figure 6

Below figure (7) show the number of cases pending at the year end over the years in areas where the number of reported cases is greater than 5000. West Bengal shows a clear elevation in pending cases starting from 2004 and other stated has a relatively steady increase.

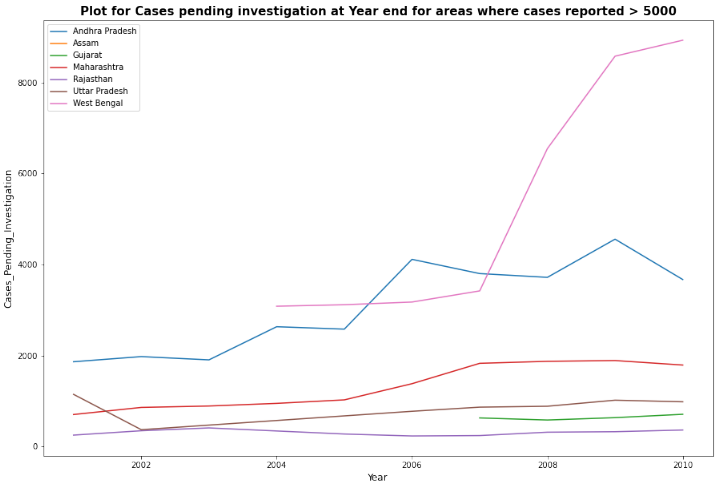


Figure 7

For the next section, the number of pending trials from the previous year is considered along with the total number of trials over the years and we try to fit in a linear model.

We consider the dataframe with columns 'Area', 'Group', 'Cases_Pending_Trial_from_the_previous_year' and 'Total_Cases_for_Trial', where we take X as 'Cases_Pending_Trial_from_the_previous_year' and Y as 'Total_Cases_for_Trial'. numpy.polyfit() is used to fit a linear

model to the data, storing the parameters into variables called m, slope, and b, the y-intercept. Then, the residual of the data is computed with respect to the model. We store the value in the variable 'res' and append it as another column in the new dataframe.
The figure (8) shows the linear model along with the X and Y points.

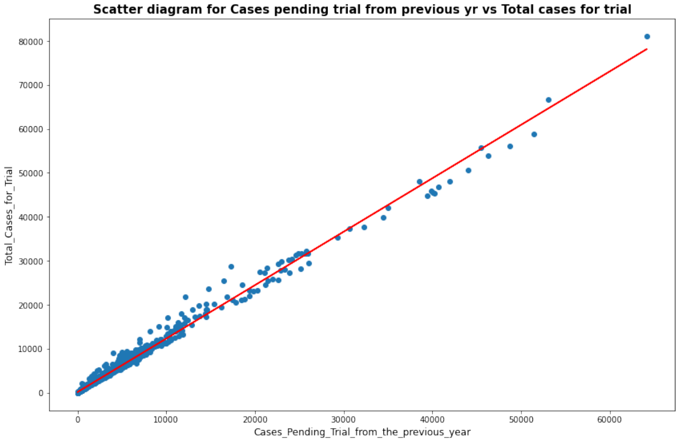
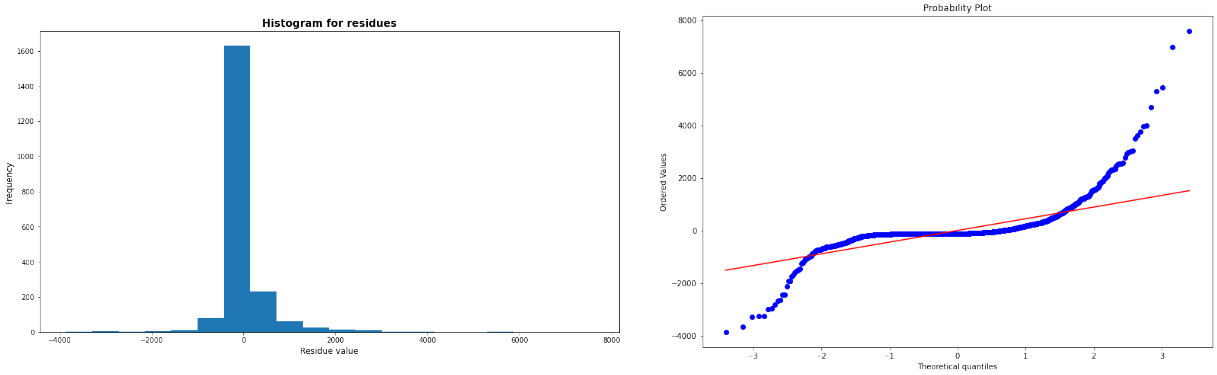


Figure 8

The plot for residue values (9a) and qq-plot (9b) is given below.



(a)

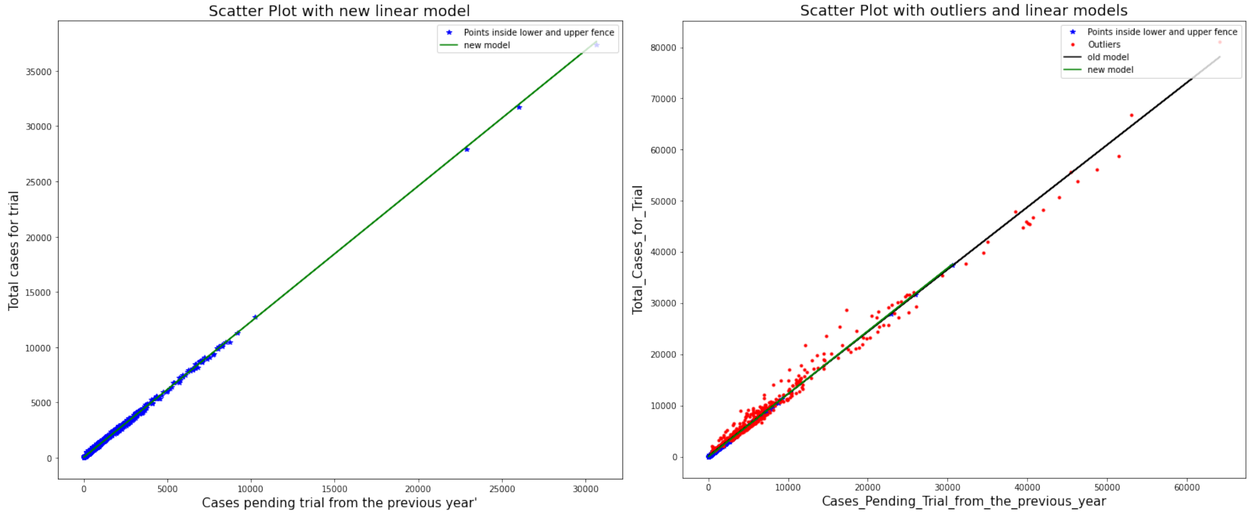
(b)

Figure 9

The above figures indicates that the linear model is not a good fit as the histogram doesnot show a normal distribution.

In the next part, we find the inner fences of the residue values, where hinges are set at the 25th and 75th quantile. and remove the points where the residue lies between the inner fences. Then, we fit a new linear model consider the new X and Y points and continue the previous processes to obtain the residue.

The figure (10a) shows the new linear model with the points inside the inner fences whereas we plot another scatter diagram along the the old and new linear model (10b) where the outliers are shown as red points and the remaining points are shown as blue stars.

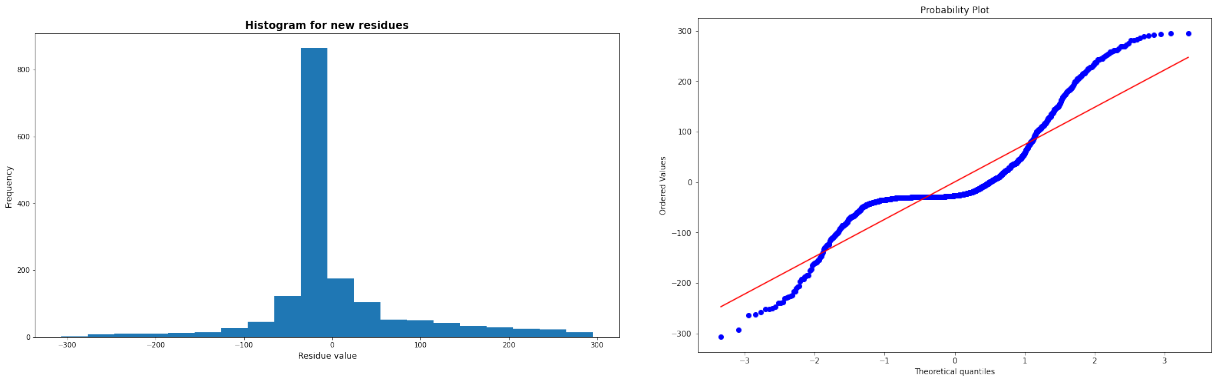


(a)

(b)

Figure 10

Again, we plot the histogram for the new residue values (11a) and qq-plot (11b) to check if the new model is a better fit than the previous model.



(a)

(b)

Figure 11

The histogram (11a) depicts a normal distribution where the higher density is around 0. As the end points in the probability plot lies nearer to red line as compared to the previous diagram, the model is a better fit than the previous one.

The other figures that we plotted just to see the patterns are given below. W considered the number of charge sheeted cases in each state or union territories and created a histogram. The figure shows that the maximum frequency is from 0 to 2000 (approx) and the maximum number of charge sheeted cases around 14000 to 16500 only occurs in 1 or 2 cases.

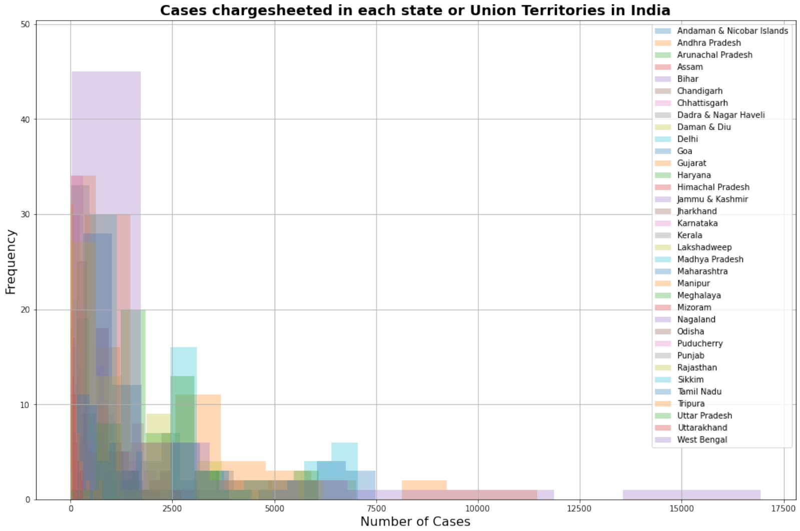


Figure 12

The below figure shows the average number of cases acquitted or discharged based on the categories of crime. A dataframe with minimum, maximum and average number of cases acquitted or discharged was found by grouping in terms of area and category of crime. The average value is utilized to plot the below diagram.

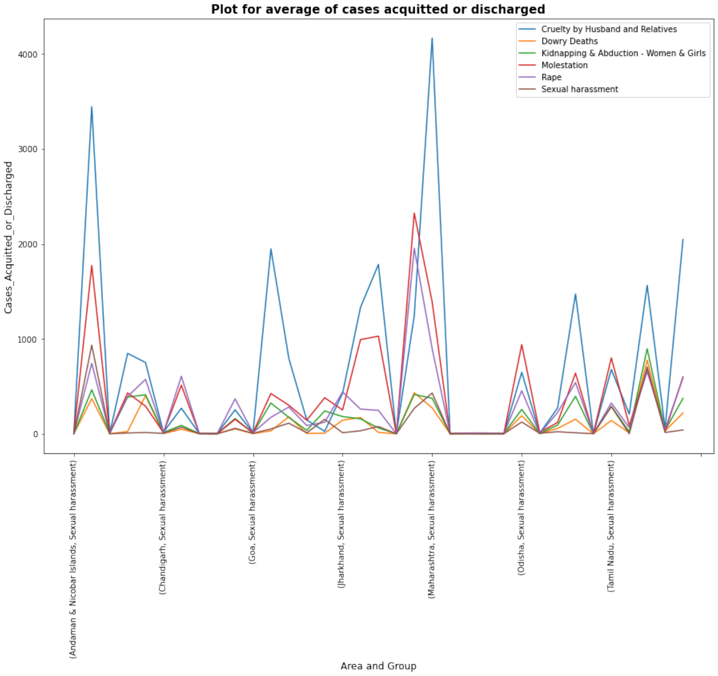


Figure 13

Conclusion

The analysis that we did on the dataset was just a few and even that could reveal the urgency of reforms needed to be implemented as to ensure the safety of women in India. This is a vast area and in-depth analysis is required to understand and further develop on the nation-wide policies and educational reforms that must be prioritised to prevent the crime not only just in case of women safety but as a whole.

The scope of the project goes far and beyond just the one depicted here. Analysing the crime data involving the delay of trials and the increasing number of awaiting cases are just one aspect. There may be other factors affecting the increase in crime and that may include the literacy rate, economy, employment rates or even the government itself. These can be quantified and organized to get a better quality data and analysing this data is the initiative that is required to build an improved society.

References

[1] *https://www.kaggle.com/rajanand/crime-in-india*

[2] *https://data.gov.in/catalog/crime-against-women?filters%5Bfield_catalog_reference %5D=86908&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc*

[3] *https://ncrb.gov.in/en/crime-against-women-statesuts*

[4] *https://www.hindustantimes.com/india-news/more-than-370-000-cases-of-crimes-against-women-reported-in-2020-says-govt-101639625323320.html*

[5] *https://www.sciencedirect.com/science/article/abs/pii/S0144818820301666*