**PREDICTIVE ANALYSIS OF CRIME IN FIRST WORLD AND THIRD WORLD COUNTRIES (INDIA and USA)**

***Abstract:***

Predictive analysis is concerned with the branch of data science used to predict future patterns and trends. This modelling technique can be used to aid society. In recent years, crime against women has skyrocketed and understanding past history can help to come up with insightful understandings that describe the current state of crime and assault in these countries. This research aims to foresee the crime patterns against women in India and USA. The studies are carried out in both countries to better understand the economic bearing, if any, in crimes committed against women. The data of the past years is studied using extensive EDA (Exploratory Data Analysis) techniques to help understand the problems women face. The data is then normalized and Linear Regression is executed to predict future trends in crime rates. K means is then used to create clusters of states with the highest crime rates against women in India. This paper aims to help women and law enforcement by using technological advancements in the area of data science to predict future trends.

1. **INTRODUCTION**

According to the National Crime Records Bureau of India, a crime against women is committed every 3 minutes, and the rate of crime shows an increase of approximately 50% from the years 2011- 2015. However, data studies help to show that crime against women per capita in India is low compared to other countries. This new found information helps to understand that the increased rate of crime in India might be directly related to the population growth India has experienced in the past decade. The datasets of past years is studied using extensive EDA (Exploratory Data Analysis) techniques that aids in understanding the problems women and girls of India face. Comparative study of various classification algorithms [4] has fuelled the motivation to consider the accuracy provided by supervised learning techniques such as Linear Regression in predicting future crime rates in India compared to other countries, without limiting the study to just one state. The data will then be cleaned and supervised learning techniques such as Linear Regression and unsupervised technique such as K means can be executed to predict future trends in crime rates with time.

The research aims to help the women of India by using technological advancements in the area of data science to predict the future state of the country [2]. The paper compares studies carried out in USA to better understand the economic bearing, if any in crimes committed against women. The datasets obtained from USA will first be analysed along with India using various methods of exploratory data analysis and aims to draw certain findings that will help to carry out linear regression for the same. K-Means algorithm will be performed on India to pinpoint the hotspots of crime against women as it is an easy and efficient algorithm that provides near accurate results [22]. This can help law enforcement agencies focus on these areas more and find remedies to provide a more secure environment.

1. **EXISTING WORK**

There has been an enormous increase in crime in the recent past [17]. Latest developments in ‘Predictive analytics Using Data Clustering Techniques’ discussed by A. Anitha [4] showcases the comparative study of various clustering algorithms used to predict the rate of crime. The usage of unsupervised learning techniques in this case provides the reader with possible clusters into which future crime may fall into. Predictive Modelling by Feng M.[2], provides insights by using least squares simple Linear Regression problem [1]. The author states that taking the population of the country into consideration while performing predictive analysis will provide more accurate results. The study concluded that more women were coming forward with complaints in the present years which then led to a possible increase in the number of filed complaints in the past few years. Both studies show an increase in reported cases from the past decade. ‘The crime against women (CAW)’ studies done by S Lavanayaa [5] shows the increase in number of reports of rape in India following the Delhi rape case which received national and international attention. Research done by author, Sangani showcases the important of K means algorithm in creating effective clusters [9].

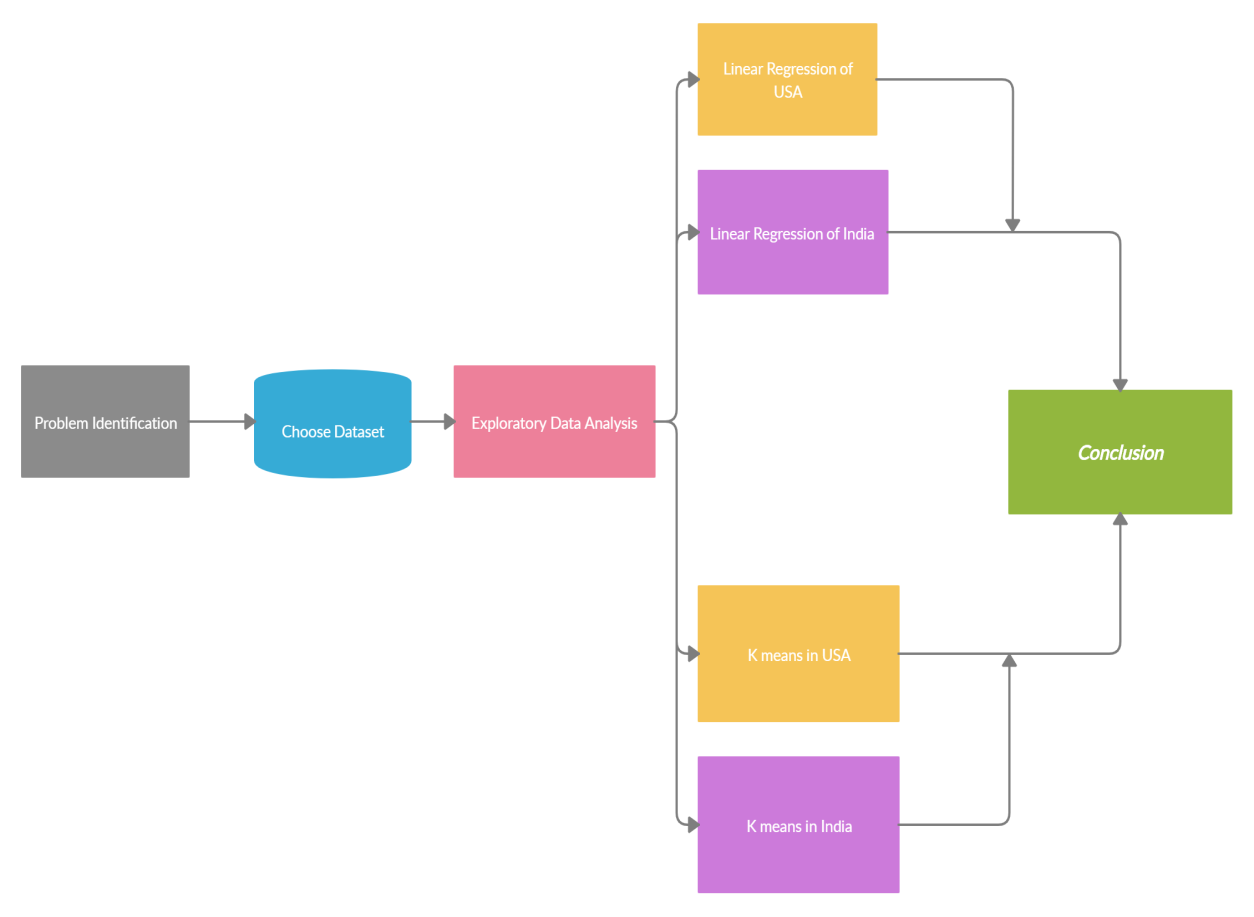
In 2018, ‘Crime rate prediction using data clustering algorithms’[3] conducted a comparative study using K means and Fuzzy C clustering techniques on unstructured data. The research proposes in-depth studies focussing on smaller and static data sets which can help the user narrow down more accurate predictions. The impact of the individual factors on the results were checked by author Gupta M. for the overall crime rate in Delhi using K-means clustering to classify the respondents or cases into clusters on the basis of degree of crime rate for various factors influencing the crime against women [18]. The author claims of the accuracy provided by K means. The effective crime prediction and analysis done by Kiani R. uses k-means as a method to classify data based on the level of vulnerability of the area. This is adopted by the paper to generate efficient and clear clusters of hotspots in the country.

The biggest challenge facing by many law enforcement is how to efficiently and accurately analyzing the increasing volumes of crime data [1]. Author, Rishabh Singh performed a variety of unsupervised learning techniques and finally came to the conclusion that K means is an easy and efficient method for grouping large data sets of crime data [26]. In factors affecting crime against women [20], Kaur states many environmental factors play into the rates of crimes against women in each state and country. The weighted variables are taken into consideration based on PCA and correlation by this paper to create better results.

**III. PROPOSED METHODOLOGY**

For the study, the following steps have been adopted:

1. The first step is to identify the problem. In this case, the aim is to understand the rape statistics in a third world vs. a first world country and hence data pertaining to that is collected.
2. Exploratory data analysis is conducted to understand the data and to visualize the data for easier and better understanding
3. Linear regression is conducted on India and USA to predict crime trends in each state. This helps to understand the statistical trends in both countries.
4. K means is conducted to understand the most dangerous states in India so that law enforcement can focus on them based on the clusters formed.



**Figure1. Block Diagram of Methodology**

For this study, the following datasets have been used. All have been obtained from the Open Government Data (OGD) Platform India (data.gov.in).

1. Crime against Women ( 2001-2012)
2. World crime against women statistics
3. District-wise crimes committed against Women (2001-2012)
4. Age and sex wise persons arrested under crime against women
5. Persons arrested under crime against Women during 2001-2012.
6. USA rape statistics 2000-2015

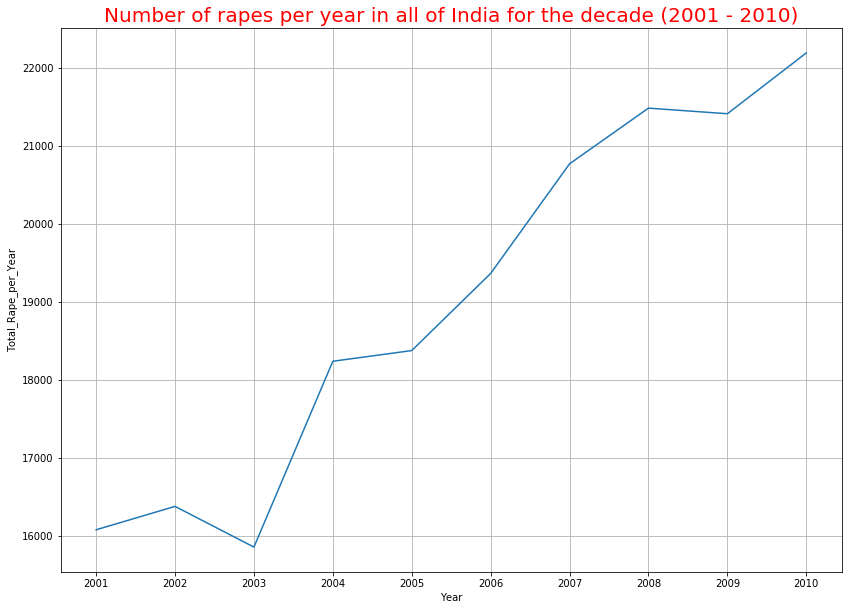
Datasets 1 is contributed by Ministry of Home Affairs, Department of States, and National Crime Records Bureau (NCRB). Datasets 3 and 5 are contributed by Ministry of Home Affairs, Department of States, and National Crime Records Bureau (NCRB).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | PAPER NAME | YEAR | METHOD USED | DATA SETS USED | DISADVANTAGE |
| 1 | Approach of Predictive Modeling on Crime Against Women Problem [2] | 2018 | Linear Regression | * Crime against Women during 2001-2012 * Crime against Women during 2013 * District-wise crimes committed against Women during 2001-2012 | The various attributes  regarding the  crime occurred like  date, time, location  etc. can immensely  improve the prediction  model. |
| 2 | Crime rate prediction using data clustering algorithms [3] | 2018 | K-Means,  Fuzzy-C | The crime data available is present in different forms | More in depth inputs will provide better and accurate results. Unstructured data , wanted to get clarity. |
| 3 | Prediction of Crime Rate Using Data Clustering Technique | 2020 | K-means,  DB-Scan,  Agglomerative | The data was collected from State Crime Records Bureau West Bengal. | Crime in west Bengal alone is  analysed |
| 4 | An Approach for Crime Rate  Prediction Using Data Mining | 2019 | K-means | This dataset contains incidents derived from SFPD Crime Incident Reporting system |  |
| 5 | Data Mining techniques to Analyze and predict crimes [9] | 2019 | Weka, Matlab | Use a clustering/classify based model to anticipate crime trends |  |
| 6 | Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques. [5]  . | 2019 | Classification methods, Apriori | Tamil Nadu Crime against women |  |

**Table1. Summary of previous data analysis performed on datasets**

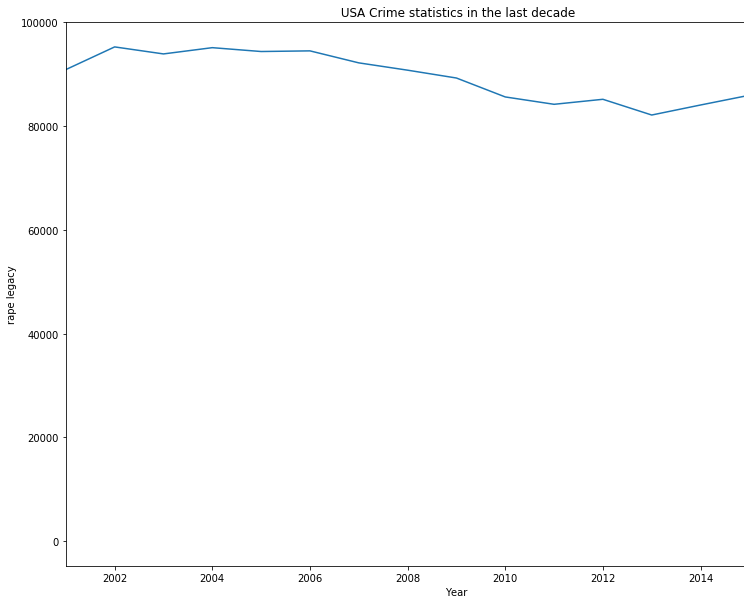
**IV. SIMULATION AND ANALYSIS**

Crime data analysis has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis [20].Crime against women is increasing every year and as per the research they have doubled over the past ten years, according to latest data released by the NCRB (National Crime Records Bureau) [18]. Crime against women over the decade has increased significantly in India as shown in Figure 2. There has been an increase from ~16000 to ~22000 from 2001-2010.

****

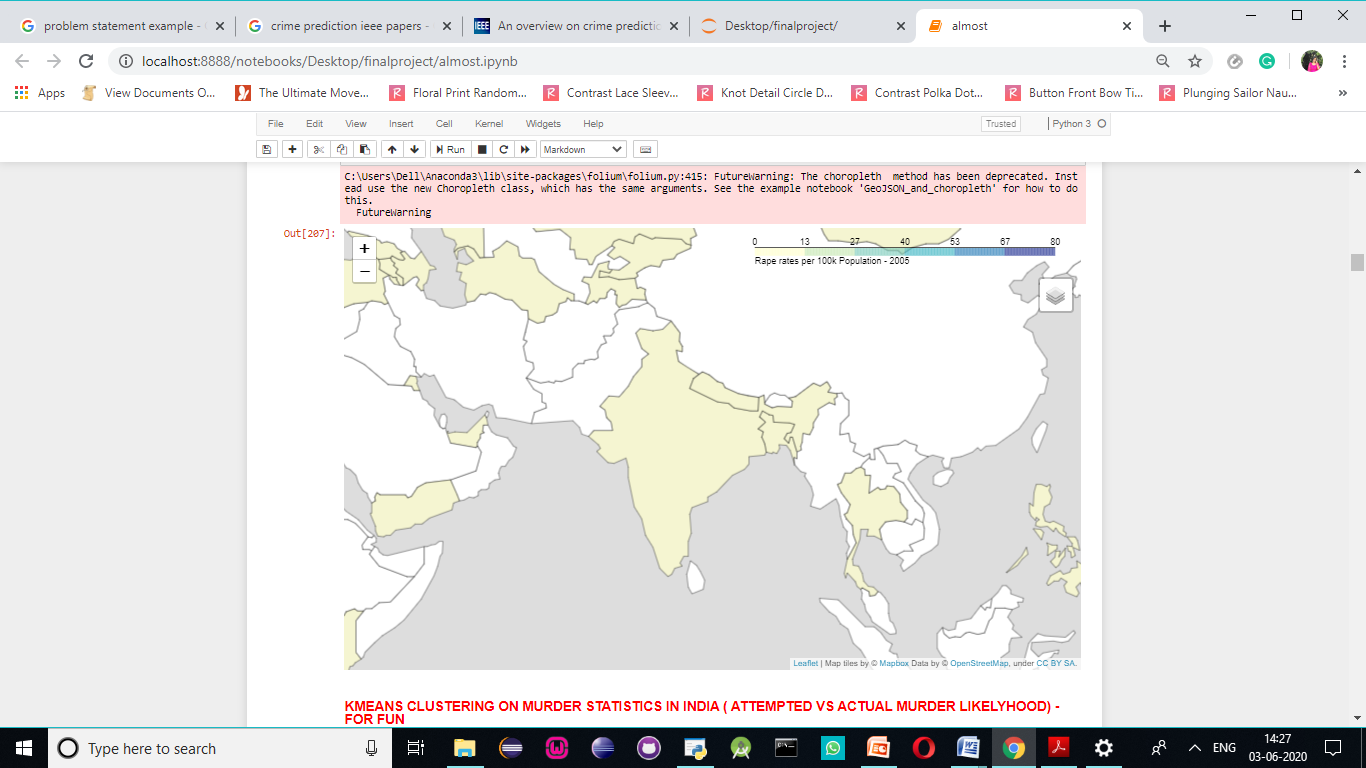
**Figure2. Pattern of rapes in India over a decade**

However, it can be seen that in USA, the rape rates shows a gradual decrease over the decade. The number of rapes in USA vary between 100,000 to 80,000 over the decade which is much higher than in India.



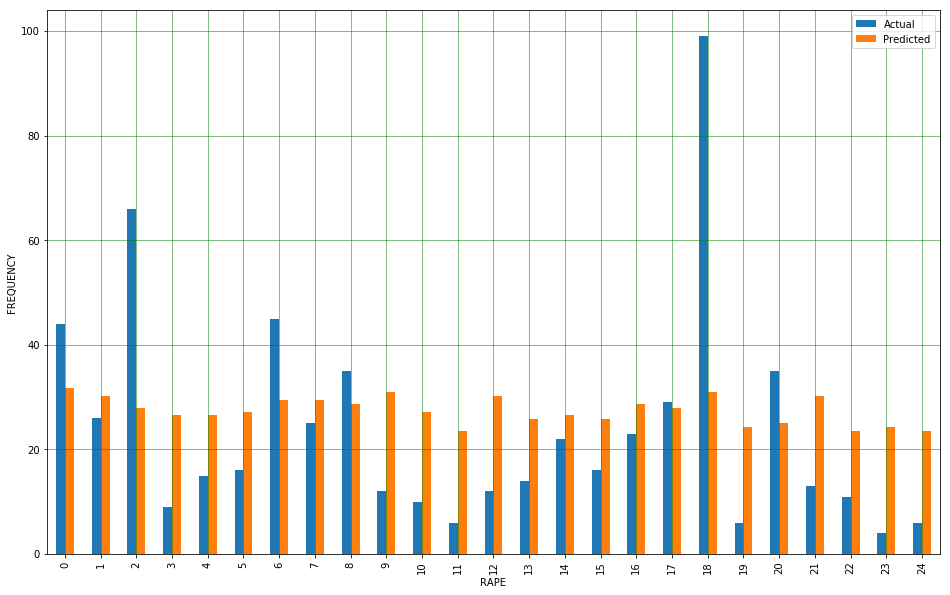
**Figure3. Pattern of rapes in USA over a decade**

However, data studies show that crime against women per capita in India is low compared to other countries as seen in Figure 4. This new found information helps to understand that the increased rate of crime in India might be directly related to the population growth India has experienced in the past decade.

****

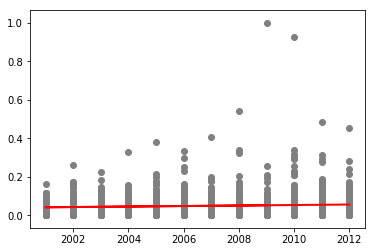
**Figure4. Folium generated map of India crime rate against women per capita**

To understand future trends, linear regression was performed to form predicted values for the crimes committed against women over time.

****

**Figure5. Actual vs. Predicted rates of rape per state in India**

The calculated predicted value of the average value of rapes per state in India is of value ~34.7. From Figure 6, a slight increase in the slope can be clearly seen.



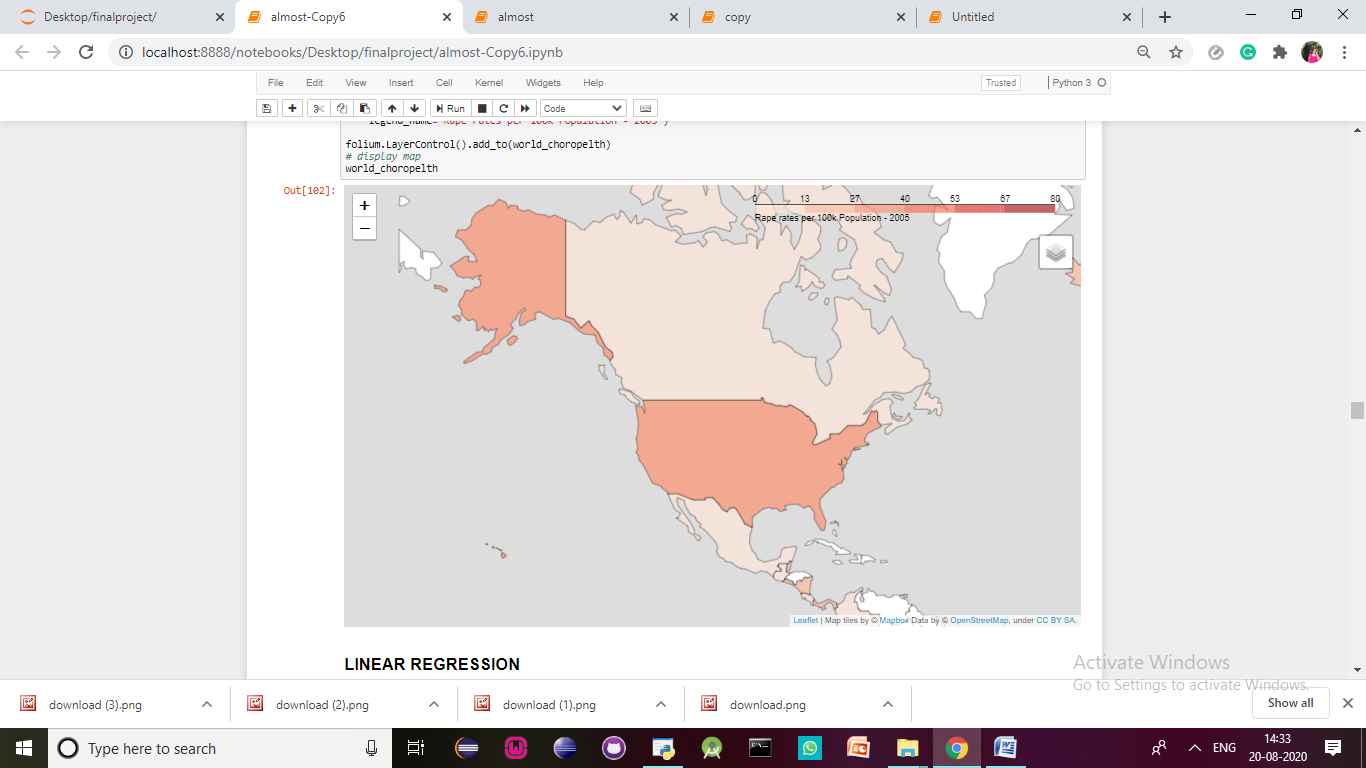
Normalized frequency fFrequency

Year

**Figure6. Normalized Linear regression plot of India**

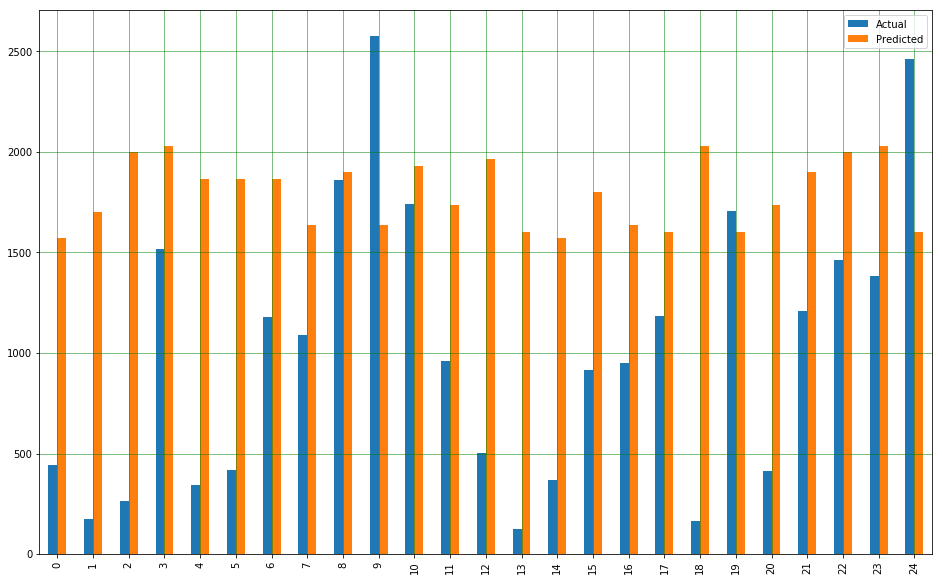
Applying the equation for Linear Regression, predicted value of rapes per district is 34.709 per state for a single year in India.

Moving on to crimes against women in USA, it can be seen that USA has one of the highest numbers of rapes in the world per capita as seen on Figure 7.

****

**Figure7. Folium generated map of USA crime rate against women per capita**

Figure 8 shows the actual vs. Predicted values obtained for each year per jurisdiction in USA.

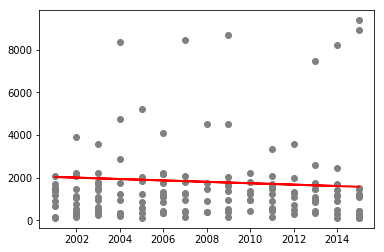
****

Frequency

Rape Rates

**Figure8. Actual vs. Predicted rates of rape per jurisdiction in USA**

The regression clearly shows that there has been a decrease in the number of rapes in USA over the decade. While in India, the study has seen a steady raise. The regression line is now plotted based on the predicted values as seen in Figure 8 and 9.

****

Frequency fFrequency

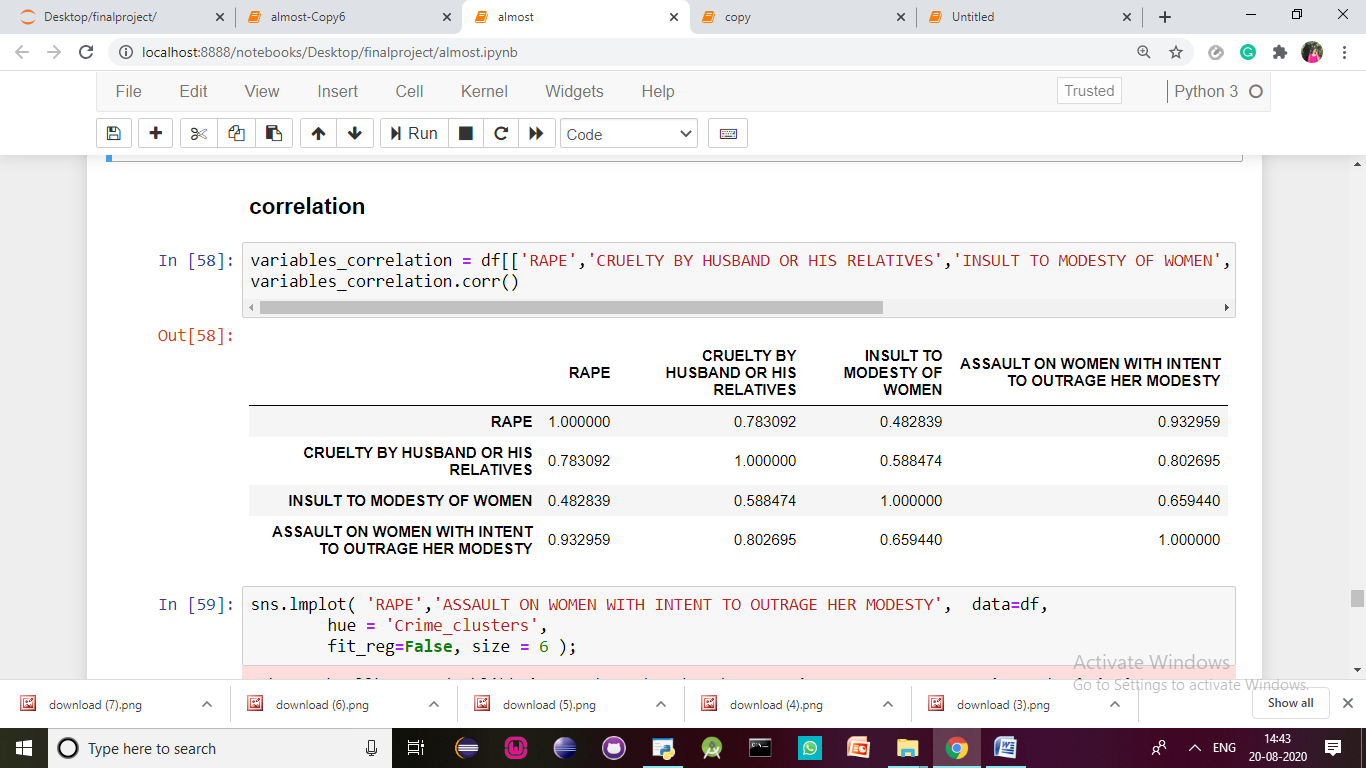
Year

**Figure9. Linear regression plot of USA**

The calculated predicted value of the average value of rapes per jurisdiction in USA is of value 1537.18 for a single year. This value has been calculated through predicted values acquired through linear regression.

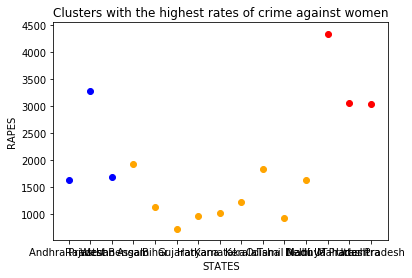
Now, it can be noticed that the increase in rape statistics in India can be contributed by a few number of states which have very high crime rates against women as opposed to USA. Finding the hotspots will help law enforcement narrow in on these states and hence, help in reducing the overall crime rates in India.

It can be seen that the highest correlation exists [9] between rapes and assaults against a woman with the intent to outrage her modesty. This can lead one to believe that rape is conducted with the intent to shame the women or to cause question of her modesty. From the world map, it can be noticed that the number of rapes per capita in India is quite low. This is very contradictory to popular belief that women are not at all safe in India. However, this can also be due to a number of reasons, the main one being that the many number of rapes do not get justice or in many cases, do not even get reported in India due to the various cultural and religious constraints and judgements cast on a woman subject to such abuse.

****

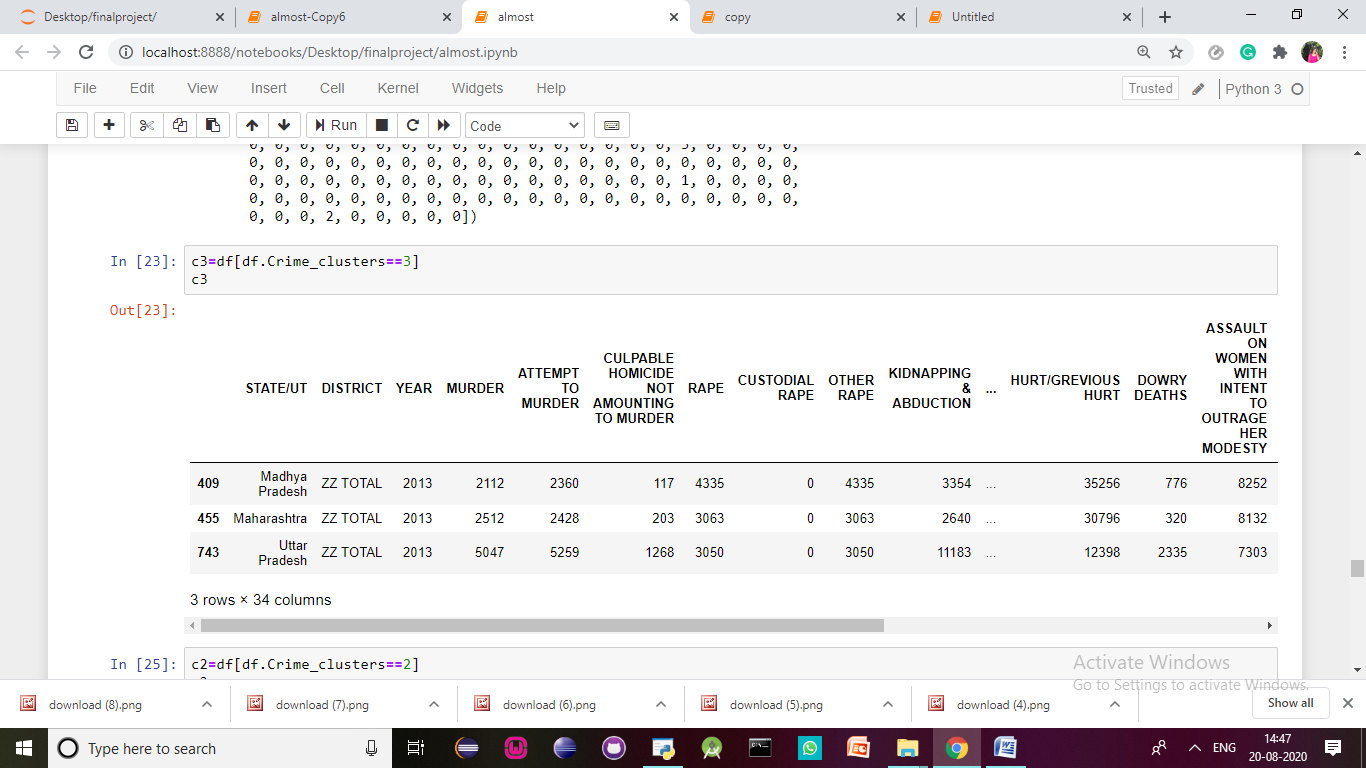
**Table2. Correlation between variables**

The K means algorithm clusters the given data stored and predicts the possible result. K-mean algorithm is accurate as compared to other clustering algorithms and is simple to understand [9]. K means is performed on the dataset of India, leading to the formation of three clusters with the highest number of crimes against women. Although the majority of the states in India fall under cluster 0 with low crime rates, these three clusters can be distinguished from the rest of them.

****

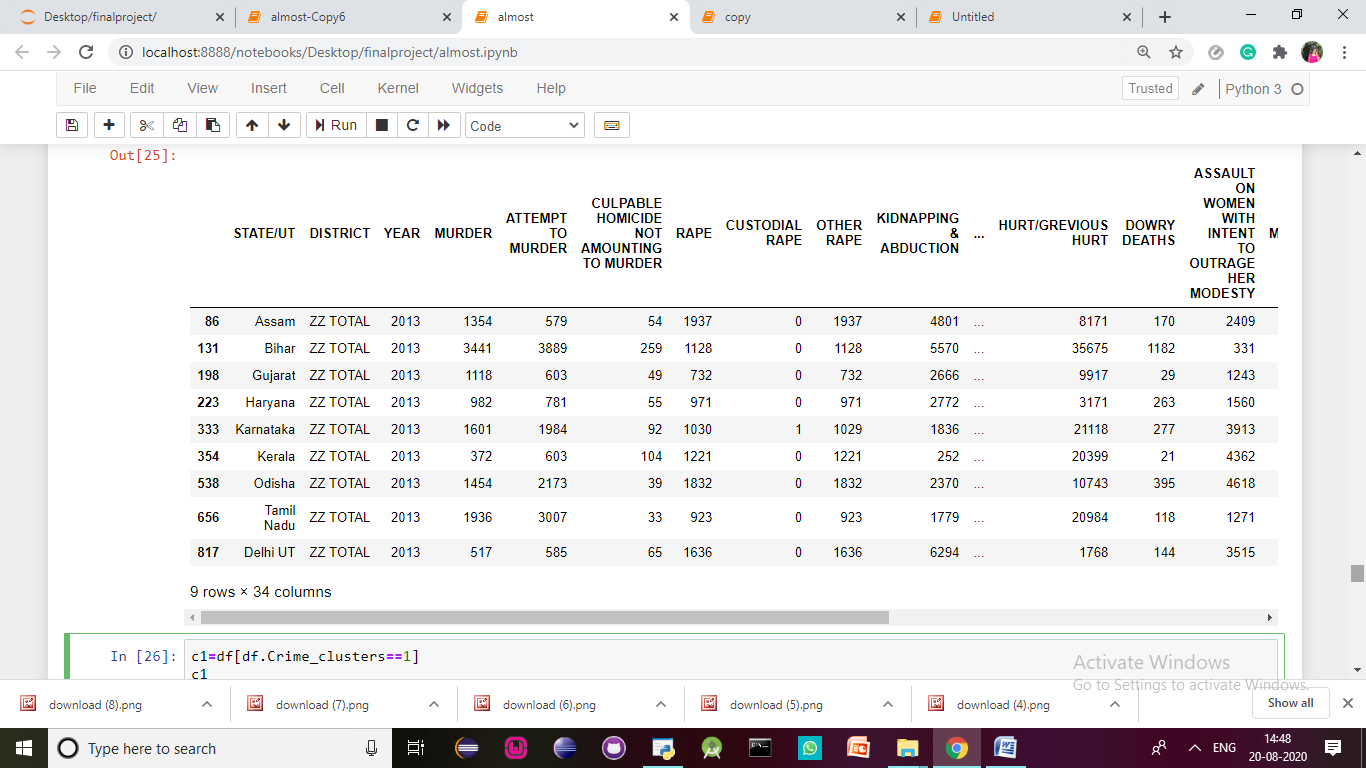
**Figure10. K means clusters in India**

The narrowed down states based on the main variables used for grouping the states into clustering are provided in the tables below. It can be seen in Table 3 that the cluster of hotspots in India that is the most dangerous for women consists of the following states: Madhya Pradesh, Maharashtra, and Uttar Pradesh.

****

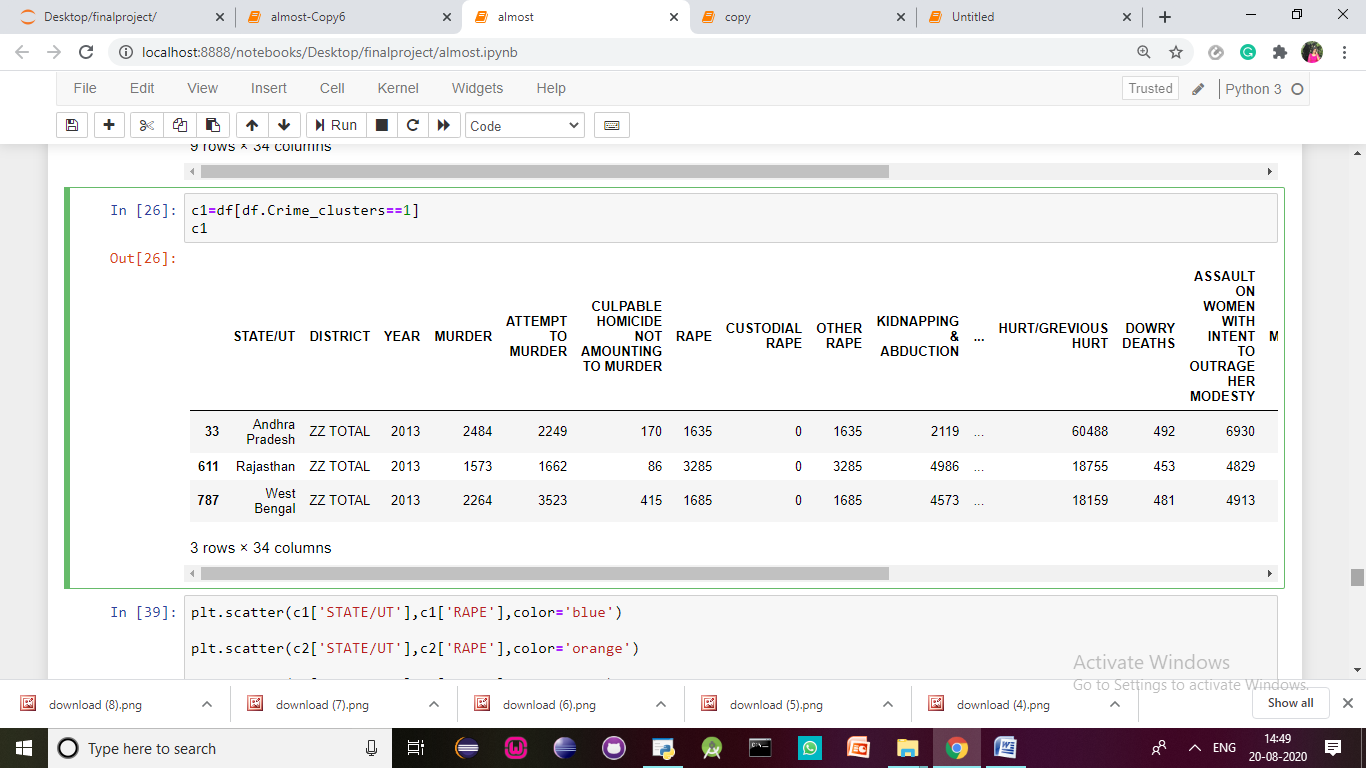
**Table3. Cluster one of hotspots**

The second cluster consists of the following states. Namely Assam, Bihar, Gujarat, Haryana, Karnataka, Kerala, Odisha, Tamil Nadu, Delhi.

****

**Table4. Cluster two of hotspots**

The third cluster contains the states Andhra Pradesh, Rajasthan and West Bengal as seen in Table 4.

****

**Table5. Cluster three of hotspots**

**V. CONCLUSIONS**

Linear Regression analysis was used to effectively predict the future trends of crime in India and USA. Validation techniques of RMSE, MSE were used to confirm the accuracy of the results obtained through the calculations. K means was then applied to the datasets of India to form clear cut clusters of the hot spots of crime in India. This helps law enforcement narrow down stringent measures to the hotspots based on the concentration of crime. It is important to note that when the mean value of rape cases in India per state is calculated it can be seen that over the decade it has been predicted the value of 34.7 per state for a single year. It was also concluded through validation that in USA 1537.18 crimes against women is committed per state over a single year. India has a rape rate per capita of 1.8 and is ranked 46th in the world among crimes of rape committed. The studies show a rate of 2 rapes per 100,000 people for India compared to 28.6 rapes/100,000 people for the US.USA has a rape rate per capita of 27.3 and is ranked 9th, 15 times more than India. From the data it can be derived that USA has 16 times more rapes than in India. Even if 50% of the rape cases in India are not reported, and all in USA are reported, they are still 8 times unsafe.

**VI. FUTURE WORK**

The paper sheds light on the state of the safety of women in the countries India and USA. However, one cannot narrow down the complexities of the reasons behind rape to the economic and social conditions. This is definitely part of a much greater venture to understand the state of various countries that fall under different economic backgrounds. In the future, one must take into account the various cases that go unreported in the countries for a better assessment. It has been reported that ~50 percent of the cases in India go unreported and ~12 percent of cases in the US are actually reported. Finding the actual datasets can help produce better and accurate predictions. This study provided us with a surprising fact that the number of rapes per capita in the USA is 15 times more than that in India. This begs the question whether feminism, social and economic standards have a large role to play as assumed by the general public. This study should be done on a number of different countries to better understand the possible reasons behind crime against women.

**VII. REFERENCES**

1. N. H. M. Shamsuddin, N. A. Ali and R. Alwee, "An overview on crime prediction methods," 2017 6th ICT International Student Project Conference (ICT-ISPC)*,*Skudai, 2017, pp. 1-5, doi: 10.1109/ICT-ISPC.2017.8075335.(IEEE)
2. Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., & Liu, Q. (2019). Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access*, *7*, 106111-106123.
3. Gandhi, P., & Sharma, S. (2018). Approach of Predictive Modeling on Crime Against Women Problem.
4. Vaidya, O., Mitra, S., Kumbhar, R., Chavan, S., & Patil, M. R. (2018). CRIME RATE PREDICTION USING DATA CLUSTERING ALGORITHMS. *International Research Journal of Engineering and Technology (IRJET) e-ISSN*, 2395-0056.
5. Lavanyaa, S., & Akila, D. (2019). Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques. *International Journal of Recent Technology and Engineering (IJRTE)*, *7*(5C).
6. Anitha, A. (2020). Prediction of Crime Rate Using Data Clustering Technique. In *Soft Computing for Problem Solving* (pp. 443-454). Springer, Singapore.
7. Yamuna, S., & Bhuvaneswari, N. S. (2012). Datamining techniques to analyze and predict crimes. *The International Journal of Engineering And Science (IJES)*, *1*(2), 243-247.
8. Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.
9. Sangani, A., Sampat, C., & Pinjarkar, V. (2019, April). Crime Prediction and Analysis. In *2nd International Conference on Advances in Science & Technology (ICAST)*.
10. Pepper, J. V. (2008, December). Forecasting crime: A city-level analysis. In *Understanding crime trends: Workshop report* (pp. 177-210). Washington, DC: National Research Council (The National Academies Press.
11. Han, J., & Kambel, M. (2012). Data Mining: Concepts and Techniques, Mor.
12. Kiani, R., Mahdavi, S., & Keshavarzi, A. (2015). Analysis and prediction of crimes by clustering and classification. *International Journal of Advanced Research in Artificial Intelligence*, *4*(8), 11-17.
13. Rahmatika, Y., Sediyono, E., & Widodo, C. E. (2020). Implementation of K-Means Clustering and Weighted Products in Determining Crime-Prone Locations. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *5*(3).
14. Sonaqwanev, T., Shaikh, S., Shaikh, S., Shinde, R., & Sayyad, A. (2015). Crime Pattern Analysis Visualization and Prediction using Data Mining. *Int. J. Adv. Res. Innov. Ideas Educ*, *1*, 681-686.
15. Agarwal, J., Nagpal, R., & Sehgal, R. (2013). Crime analysis using k-means clustering. *International Journal of Computer Applications*, *83*(4).
16. Malathi, A., & Baboo, S. S. (2011). An enhanced algorithm to predict a future crime using data mining.
17. Gupta, M., Chandra, B., & Gupta, M. P. (2008). Crime data mining for Indian police information system. *Computer society of India*, *40*(1), 388-397.
18. Palmer, P. B., & O'Connell, D. G. (2009). Regression analysis for prediction: understanding the process. *Cardiopulmonary physical therapy journal*, *20*(3), 23.
19. Kaur, B., Ahuja, L., & Kumar, V. (2018). Factors affecting crime against women using regression and K-means clustering techniques. In *Industry Interactive Innovations in Science, Engineering and Technology* (pp. 149-162). Springer, Singapore.
20. Goel, P., & Yadav, V. (2013). Crime Against Indian Women–Women Crime Susceptibility Indexes (WCSI): A Principal Component Analysis. OIDA International Journal of Sustainable Development, 6(7), 93-102.
21. Bhattacharya, D. Crime Detection Technique Using Data Mining and K-Means.
22. Pande, V., Samant, V., & Nair, S. (2016). Crime detection using data mining. International Journal of Engineering Research & Technology (IJERT), 5(01).
23. Hamdy, E., Adl, A., Hassanien, A. E., Hegazy, O., & Kim, T. H. (2015, July). Criminal act detection and identification model. In *2015 Seventh International Conference on Advanced Communication and Networking (ACN)* (pp. 79-83). IEEE.
24. Vijayarani, S., Suganya, E., & Navya, C. (2020). A Comprehensive Analysis of Crime Analysis Using Data Mining Techniques.
25. Garg, R., Malik, A., & Raj, G. (2018, January). A Comprehensive Analysis for Crime Prediction in Smart City Using R Programming. In *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 14-15). IEEE.
26. Singh, R., Reddy, R., Kapoor, V., & Churi, P. (2020). K-means Clustering Analysis of Crimes on Indian Women. *Journal of Cybersecurity and Information Management (JCIM) Vol*, *4*(1), 5-25.