Analysis
Crime Against Women
and
Cyber Security
Using Data Mining Techniques

The project was carried out as a part of DSKC Summer Internship Program

Teacher Mentor: Dr. Seema Aggarwal

Members that participated in this project:

1. Harshita Pahwa
2. Chantel Rose Walia
3. Khushi
4. Nancy Dahiya
5. Pooja
6. Rhea Ajit John
7. Siddhi Joshi

Introduction

• Crime Against Women

Women are revered in India. They are elevated to the level of deities. However, the truth is rather different. With the passage of time, women's safety has begun to become a serious worry. The rate of crime is rapidly increasing with each passing day. It is increasingly regarded as a worldwide concern, with many countries attempting to implement measures in order to reduce crime rates. India isn't far behind either. According to data provided by the National Crime Records Bureau, the number of crimes reported against women has increased in recent years. 'Dowry Harassment / Cruelty to married women' 'Kidnapping / Abduction' are all becoming more common. To prevent such crimes and safeguard women's safety, the government is attempting to enact harsher legislation and take significant measures. To prevent such crimes and safeguard women's safety, the government is attempting to enact stronger laws and serious steps. Every year, a massive amount of data is generated in relation to various crimes in various parts of India. Analysing such large data sets may appear to be a time-consuming activity. Data mining plays a significant role in evaluating enormous numbers of records, providing accurate findings, and identifying trends, thanks to new technology and approaches. The outcome of such a study will not be exactly the same as the perceived outcome, but it will provide a reasonable estimate of the number of crimes that will occur in a given state in the next years. The key difficulty in this forecast is to reduce losses and bring the consequent number of a specific crime, such as rape, in a specific state in the next years to the real figure. The issues we're dealing with are as follows:

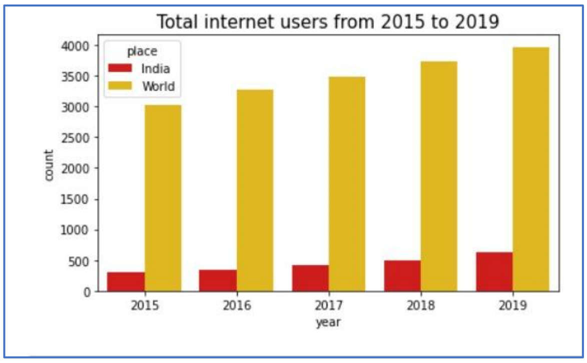
- Examining statistics on various types of crimes in each of the 28 states and 8 union territories
- Obtaining new datasets with more sets of crimes from various crime departments in order to reduce loss to a minimum for each sort of crime.

Prediction for the year 2020 and 2021. This is because, these years are

Pandemic years and the cases reported during 2020 and 2021 are not accurate.

- ## Cyber Crime

Being able to save important data on a device/network enables people to live a more mobile and connected life- where all the information of the world, be it publicly available or privy only to them- is only a click away in the palm of their hands. This worldwide connectivity has, although helped people, created a new wave of criminals- Cyber Criminals, the people with the technical know-how to gain access to another person’s online information for malicious purposes.



From the above figure, it is clear that dependency on internet devices has steadily increased as such, the scope for criminals to commit acts of cybercrime against people increases. A few types of cybercrimes committed against people are as follows: 'Tampering Computer Source Documents', 'Computer Related Offences ', 'Obscene/ Publication/ Transmission in Electronic Form', 'Failure of Compliance/ Orders of Certifying Authority', etc being some very specific examples. A few umbrella examples of cybercrimes as categorized by the Information Technology Act, 2000 are: Phishing, Scam, Child Pornography, Credit Card Frauds etc. Cybercrime techniques can also be used to drop economies overnight / steal personal data / blackmail etc and halt all digital movement, if the criminal is willing enough. It all boils down to the fact that if someone wants to, then cybercrime may be the most efficient way to cause harm, and so for countries like India that are still developing, creating a cybercrime division and stricter laws in this area at an early stage is imperative to maintain a safe and steady growth for all its citizens. Considering the past of cybercrime in India, datasets from the years 2008-2019 obtained from the Open Government Data Platform of India- ncrb.gov.in will be used to further learn about the trend of cybercrimes committed and hence a prediction shall be concluded viz. The cybercrimes committed in each state in India in the year 2020.

Literature Survey

• Crime Against Women

Paper name: *Crime Against Women: Analysis and Prediction*
Author: Purvi Prasad, Amrita Nair, Dr. S. Godfrey Winster
SRM Institute of Science and Technology Chennai, India
Summary: The data utilized to conduct the analysis is critical in identifying patterns, particularly in crime analysis. The dataset is taken from Kaggle and covers several forms of crimes committed against women, including "rape," "kidnapping," "dowry death," "attack on women," "cruelty against women," "importation of young girls," "insult to modesty," and "immoral traffic," among others. This time series data is transformed into supervised data in order to forecast the number of crimes that may occur in the future. Huber Regression is utilized to analyze these various offences. It calculates the loss score, or how great the gap between the actual and expected value will be. It's possible that the prognosis isn't very accurate because it's based on the number of crimes that have occurred in recent years. There is no consistent growth or reduction in the number of crimes committed. In order to minimize the loss, the predicted values are brought closer to the real value. The forecasts can be seen in the form of a bar graph, which shows which states/union territories have the highest rates of the specified crimes. It will also allow us to determine which form of crime is most prevalent in a given state.

Paper name: *Statistical Analysis of Crimes Concerning Women in India using Data Mining Techniques*
Author: K.H.S. Rajeswari, M. Deepthi, D.N.D. Harini
Summary: The Analysis of Crimes here in this paper is carried out in two stages: Clustering and Classification with the help of Weka tool. The input to the K-means clustering algorithm is numeric dataset. After the data set has been pre-processed, the user can navigate between the tab options to make modifications to the experiment and see the results in real time. This has the benefit of being able to switch from one choice to the next so that when a condition is discovered, it may be placed in a different setting and visually modified immediately. They have generated five clusters for each attribute after using K-Means, and build a nominal dataset from the clustered output for further categorization and decision making using the ID3 algorithm. The decision tree generated during the classification phase is used to analyse the output. During this stage, they arrived at several findings, such as which state has the highest victim rate based on the attributes. They have also highlighted the types of crime that are prevalent in the majority of Indian states.

Paper name: Violence Against Women: A State Level Analysis in India 2016
Author: Tanisha Khandelwal

Summary: Violence Against Women: A State Level Analysis in India 2016 by Tanisha Khandelwal is a theoretical article on violence against women in India in 2016 and calculates the crime rate and index by state by normalizing the data of atrocities committed in India and classifies each state and territory of the Union by a crime index number.

Paper name: A comparative study of crimes against women based on Machine Learning using Big Data techniques
Author: Shivani Mishra, Suraj Kumar
Summary: This study explains how to utilize machine learning to recognize different types of crimes and display them in an easily understandable style, which could aid in the development of actionable preventive measures. This project will include the development of a clustering methodology as well as machine learning. Cleaning the dataset and clustering are examples of algorithms. The performance of each algorithm is evaluated, and the best algorithm with the highest crime detection accuracy is identified. The main goal is to create an internet application that can generate a rapid and studied picture of India's crime against women situation.

Paper name: Crimes Against Women in India using Regression
Author: R. Devakunchari, Bhowmick S, Bhutada S P, Shishodia Y
Summary: The first research paper had implemented a linear regression model while in the second research paper , the study was primarily on quantitative data, and had implemented a logistic regression model.
As both the models did not have a clustering model, we tried to implement a clustering model.

Paper name: Crime Analysis Using K-Means Clustering
Author: Khushabu A. Bokde, Tiksha P. Kakade, Dnyaneshwari S. Tumsare, Chetan G. Wadhai B.E. Student Department of CSE, Ballarpur Institute of Technology, Ballarpur,Chandrapur District, Maharashtra, India.
Summary: In this paper, k refers to a data mining clustering technique that is used to extract useful information from a large crime dataset and to interpret the data, assisting police in identifying and analysing crime patterns in order to prevent similar occurrences in the future and providing information to reduce crime. K mean clustering is implemented in this research utilising an open-source data mining tool, which are analytical techniques for evaluating data. Among the open-source data mining suites available are R, Tanagra, WEKA, KNIME, ORANGE, and Rapid miner. The rapid miner tool, which is an open source statistical and data mining software created in Java with customizable data mining support options, is used to cluster the data. Crime dataset of crimes recorded by the police in India is also utilised for crime analysis.

• Cyber Crime

Paper name: Computational System to Classify Cyber Crime offenses using Machine Learning
Author: Rupa Ch, Thippa Reddy Gadekallu, Mustufa Haider Abidi and Abdulrahman Al-Ahmari

Summary: The dataset used in this paper is collected from Kaggle and CERT-In 2000 records. Attributes of the dataset are Incident, Offender, Access Violation, Victim, Harm, Year, Location, Age of Offender. The proposed methodology of this paper divides it into 4 steps that are

1. Data Collection: the data is collected from Kaggle and CERT-In 2000 records
2. Preprocessing: This step includes the feature extraction process by using TFIDF vector method and either imputing or removing of null valued columns.
3. Applying model: They have used naïve Bayes algorithm for classification and K-means for clustering. Different cybercrime offences are clustered into some groups.
4. Prediction and Result: For prediction algorithms used are LinearSVC, LogisticRegression, MultinomialNB, RandomForestClassifier and except RandomForestClassifier all other algorithms perform approx. 99% well.

In the future, proposed model can be improved by using deep learning concepts.

Paper name: A Brief Study on Cyber Crimes and IT Act in India
Author: Dr. Adv. Mrs. Neeta Deshpande
Summary: The paper works with both primary and secondary collected data, primary data is collected by discussions with advocates and secondary data is collected through web sites, e-journals, research papers and other resources. This paper at first visualises number of internet users in the world, in Asia, registered cybercrime cases in India and cases registered and person arrested in States of India. It concludes that the total cybercrime cases registered from 2014 to 2017 are highest in Maharashtra followed by Uttar Pradesh and Karnataka. Then paper discusses about various sections under IT Act 2000 followed by table showing opinions of advocates regarding provisions of IT Act if it is sufficient for tackling all types of cybercrimes or not, etc. Finally, paper concludes by mentioning general and specific suggestions to deal with emerging cybercrimes.

Paper name: DATA MINING TECHNIQUES TO CLUSTERING CYBER CRIME DATA

Authors: Dr. Neelam Sahu[1] and Sagar Darokar[1] *

*1 Dept of Information Technology, Dr. C.V. Raman University Kota, Bilaspur (C.G.), India

Summary: The paper was based on clustering of data using K-means clustering algorithm in order to get structured data from unstructured data. The aim of the paper was to collect and classify data related to types of cybercrimes committed by people in India into multiple groups and then sub-categorized under The Copyright Act 1957 and The Trademark Act 1999.

In order to achieve this goal the data was first cleaned and pre-processed, after which the clustering of the data was begun, using the K-means clustering algorithm. The clustering technique used was considered as an effective way to properly classify unstructured data derived from the web into structured data.

Paper name: EMPIRICAL STUDY OF CYBER CRIMES IN INDIA USING DATA ANALYTICS

Authors: Disha Gupta[1] and Namrata Agrawal[2] *

*1 Gujarat Forensic Sciences University, Gandhinagar, Gujarat, India

Summary: The paper was focused on standardising, studying and analysing the data obtained for the types of cybercrimes committed in India in the year 2013. The aim of this paper was to conclude the relation of the names of the states that had the highest numbers in cybercrimes committed as well as derive a relation between use of computers and increase in cybercrimes. The methodology proposed involved thorough cleaning and additional pre-processing of the data wherever required and then analysis of the crimes committed in each state in India using data analysis techniques. Multiple visualizations were provided as proof of result, as well as a correlation of crimes with each other. The paper concluded on the main result of Maharashtra having the highest number of cybercrimes committed followed by Karnataka and Andhra Pradesh. A directly proportional relation between advancement of technology and number of cybercrimes committed was included in the conclusion as an inferred result.

Dataset

Source:
Data for Crime Against Women and Cyber Crime is collected from a government site NCRB. Crime against women consists of data from 2001 to 2019 whereas Cybercrime has a dataset from 2008 to 2019.

Description –

- **Crime Against Women in India Analysis and Prediction-**

This dataset contains a wealth of information, that is the location of the crime, states, the type of crime, and the number of victims for a specific year. For visualization, a whole dataset is created by combining a single datasheet into one dataset for the years 2001-2016. However, for the prediction, 2013-2016 datasets are taken into consideration.

- **Crime Against Women in India visualization with Clustering**

Here, the data sets are used to extract the hidden knowledge and to validate the model built on it. The data collected for work contains crime information from all the 29 states and 7 of the Union. Initially, the data set categorized crimes into different types.

- **Cyber Crime**

• The data sheets collected were in pdf form so they were converted into excel or CSV files. Dataset of all years was not same so we extracted important columns.

Cleaning of the data:

- **Crime Against Women in India Analysis and Prediction-**

Simple Python coding was used to clean the data set. It includes removing ‘nan’ or null values, changing the data type of some columns for visualization and prediction. We selected 16 distinct crimes from the dataset.

- **Crime Against Women in India visualization with Clustering-**

The data we collected from the NCRB website consisted of individual years data from 2001-2019. Initially, we clubbed all the data together into a single file for ease of processing. The original data set had null values and various different crimes. Consequently, we selected 16 distinct and considerable crimes and replaced null values with 0. In order to amalgamate the data of the years 2001-2019, we took a summation of all the crimes in sequential order. We then transposed the data such that we have the states as rows and crimes as columns. To implement k-means clustering we need numerical data thus we replaced the name of states with numbers from 1-36 to uniquely identify them while drawing out conclusions.

- **Cyber Crime**

Similarly, we extracted important and common columns from all year’s sheets. The dataset has additional titles**,** so we have to change them. Next, we removed rows that contained null values.

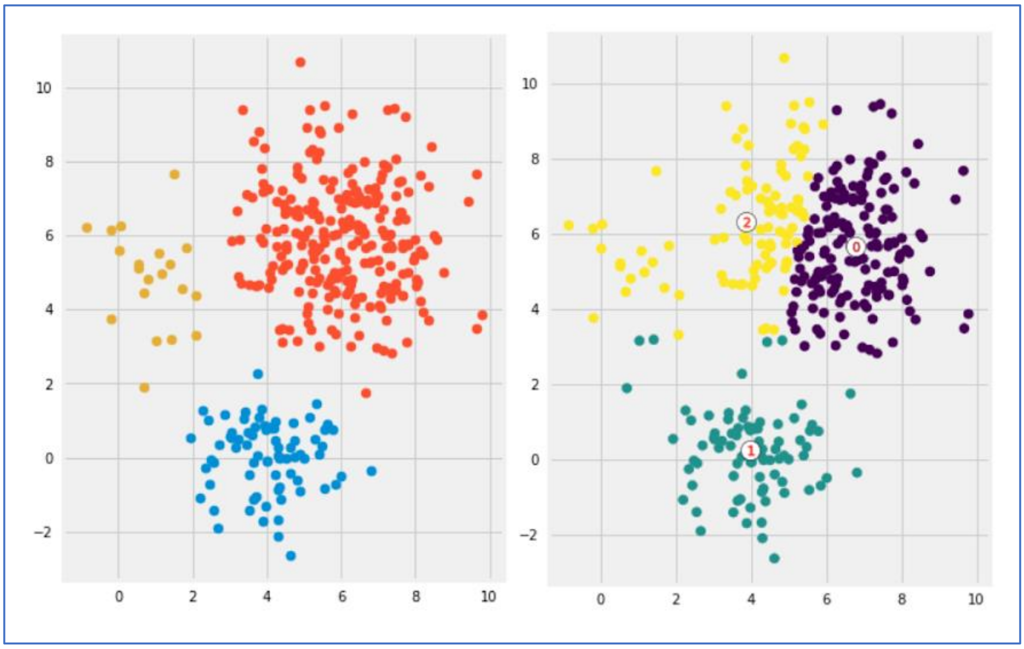
Theory

• Clustering

The K means algorithm is an iterative technique that divides the dataset into k non-overlapping groups. The distance between the centroid and the data points is also limited to a bare minimum. The lack of fluctuation indicates that the points are more homogeneous. The following is how the algorithm works:

1) Set K and start Centroids by mixing the database and selecting K data points at random.
2) The iteration should continue until the Centroids are free of charge.
3) Assign the distance before moving on to the data points.
4) Recalculate the centroid by averaging all data points.

- Calculate the total of all data points' squared distances from all centroids.
- Assign each data point to the cluster that is closest to it (centroid).
- Calculate the cluster centroids by averaging all of the data points that correspond to each cluster.



• Linear Regression

1. Linear Regression is a statistical method in machine learning. It uses a single dependent variable and either a single or multiple independent variables consequently drawing a linear relation between the variables and giving in terms of a visual representation a straight line in the X-Y/3D plane with the dataset values it was applied on, ergo its name. Due to its working methodology, it is mainly implemented as a means for prediction analysis. If a single independent variable is used then the regression thus used is called single linear regression. If multiple independent variables are used, it's called multiple linear regression.
   1. The mathematical formula used in this regression model is:

   $$\mathbf{Y = a_0 + a_1 X + \varepsilon}$$

   Here,

   - Y is the dependent/ target variable.
   - X is the independent/ predictive variable.
   - a_0 is the intercept on the line.
   - a_1 is the Linear Regression Coefficient.
   - ε is the random error

2. The first step to applying Linear Regression is identifying the types of variables involved in the dataset- this involves selecting dependent (**Y**) and independent (**X**) variables. Then the model is applied on the dataset using the variables, after which a cost function is used- in this case it is the Mean Squared Error function that's implemented to lower the difference in the predictive values and actual values. It essentially is used to make the model applied give a more accurate output for prediction.

• Gradient Boosting Regression

Gradient Boosting is one of the machine learning technique used specially for regression and classification problems. It works by combining the predictions from a number of decision trees to get the required prediction. Every new decision tree in this algorithm takes into account the mistakes of the previous tree and try to reduce the error. The nodes in all decision trees are not same they depend upon the least impurity. Different nodes for decision trees help in determining different possible relation among features of dataset.

Maths behind the algorithm:

- Gradient Boosting regression algorithm starts by making a single leaf that is the average of independent variable. For now, the model is predicting the values as the average value of the independent variable. This single leaf is considered as first tree for the of the algorithm.
- Next, error is calculated by subtracting the predicted average value from the actual value and is called the Residual. Using the values of residuals next decision tree is made. The double values in the leaf are replaced by their mean.
- To avoid low bias and high variance problem the model adds some penalty to each decision tree termed as learning rate. One can alter value of learning rate that suits their model.
- In this way, one again can calculate residuals and continue making new decision tree and adding their value to the prediction with some learning rate.
- So, by taking small steps by reducing the error the Gradient Boosting Regression algorithm works

Visualisations

The presenting of data in a graphical style is known as data visualisation. It aids with the comprehension of data by summarising and presenting large amounts of data in a simple and easy-to-understand style, as well as aiding in the clear and effective communication of information. To carry of visualizations, python libraries like Matplotlib, Seaborn, and Plotly were used.

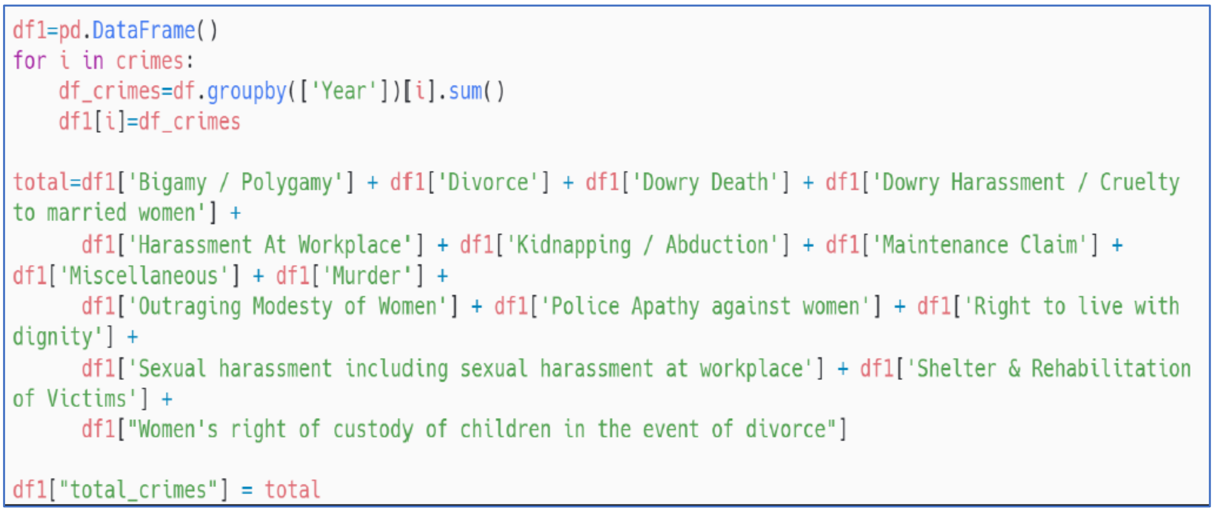
• Crime Against Women

1. Matplotlib: matplotlib. pyplot is a set of routines that allow matplotlib to behave similarly to MATLAB. Each pyplot function modifies a figure in some way, such as creating a figure, a plotting area in a figure, charting certain lines in a plotting area, decorating the plot with labels, and so on.

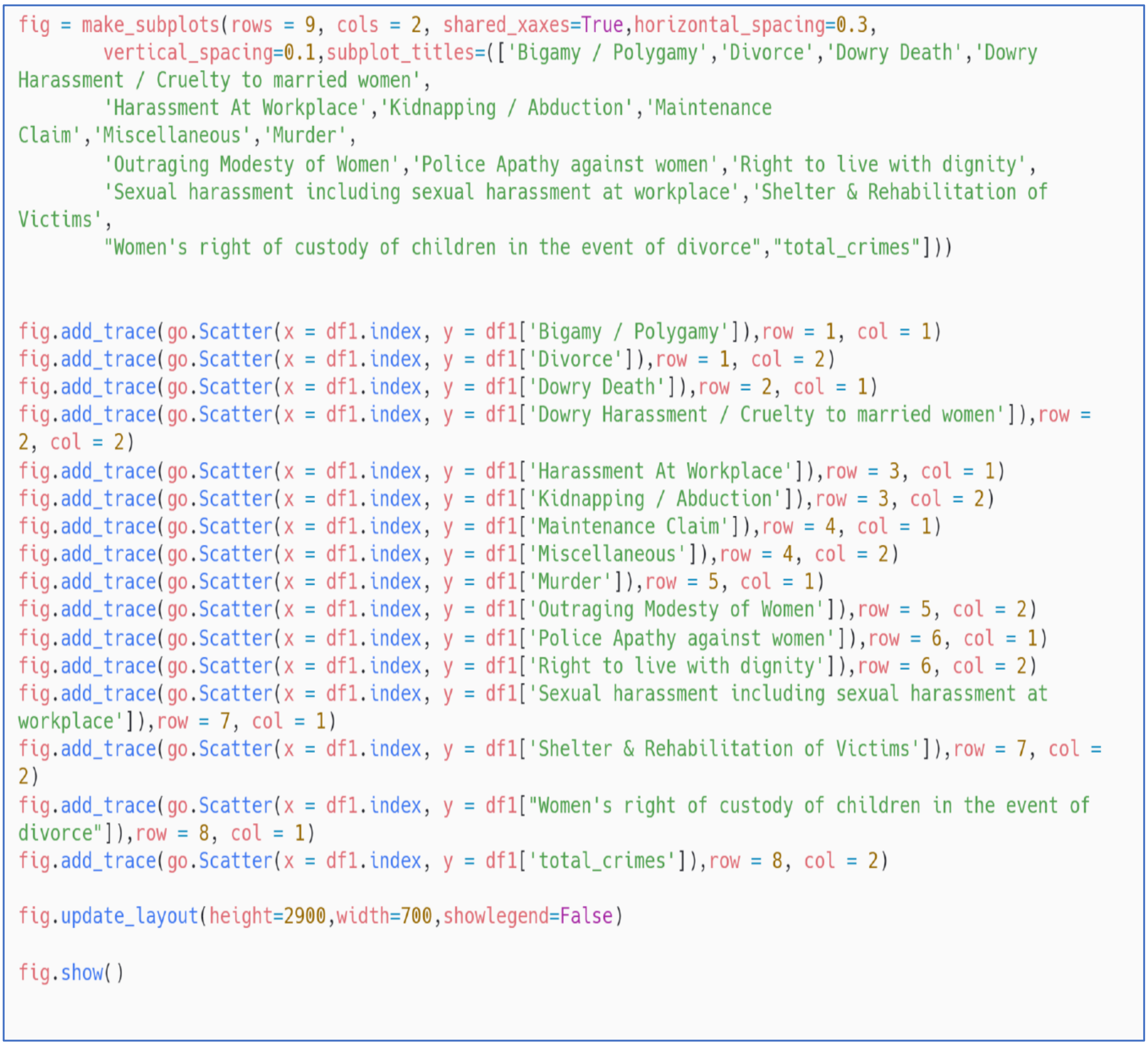
2. Seaborn: Seaborn is a fantastic Python visualisation tool for plotting statistical visuals. It comes with nice default styles and colour palettes that make statistical charts more appealing. It is based on the matplotlib software and is tightly connected with Pandas data structures. The library's goal is to make visualisation a vital aspect of data exploration and comprehension.

3.Plotly.express: Plotly.express is a fantastic method to quickly visualise your data using a single chart type. It contains notable features such as interactivity and animations, however it lacks subplot support. Many chart kinds can be created using Plotly.express shorthand syntax.

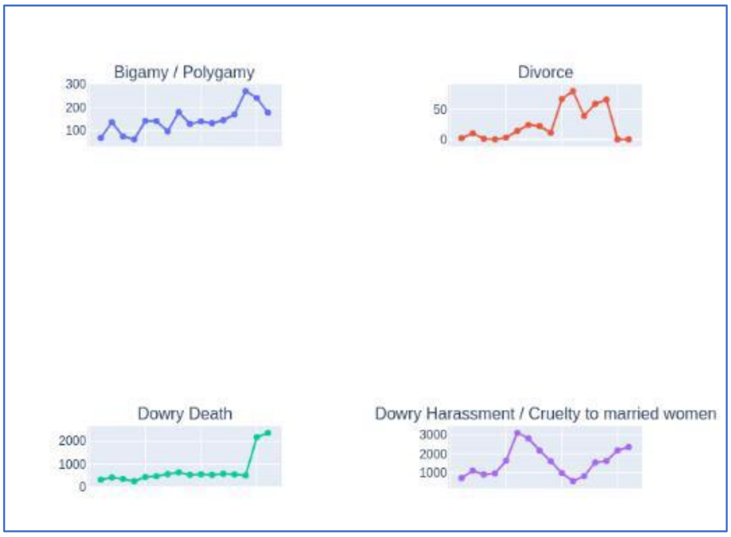
4. Plotly.graph_objects: The object responsible for constructing plots are contained in this module (Figure, layout, data, and plot definitions such as scatter plot and line chart).

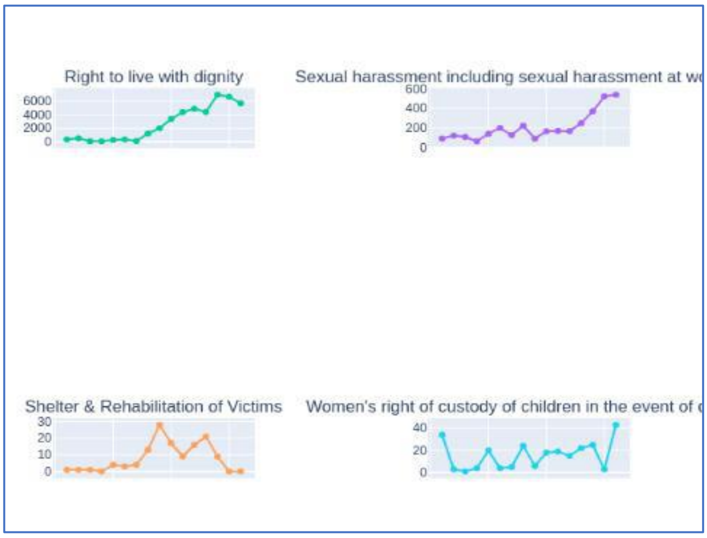


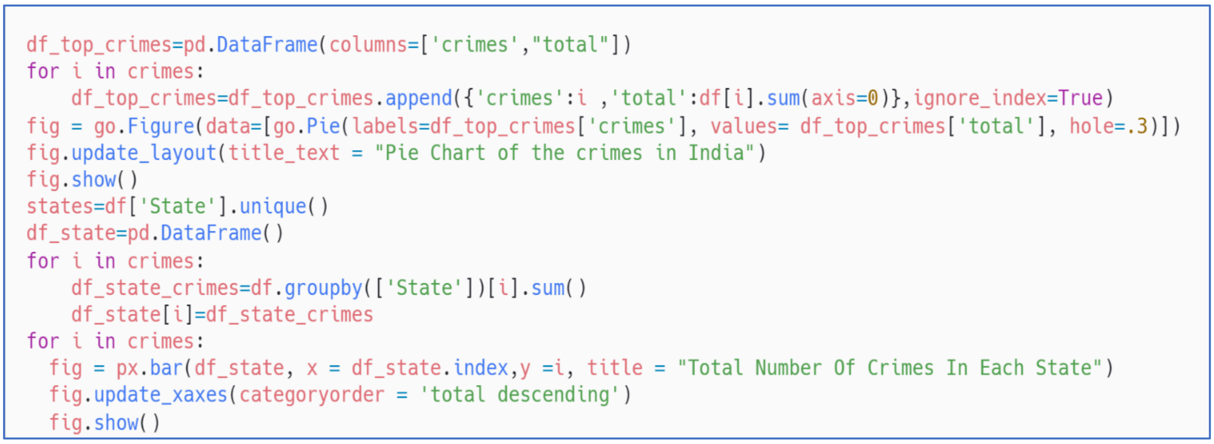
```
df1=pd.DataFrame()
for i in crimes:
    df_crimes=df.groupby(['Year'])[i].sum()
    df1[i]=df_crimes

total=df1['Bigamy / Polygamy'] + df1['Divorce'] + df1['Dowry Death'] + df1['Dowry Harassment / Cruelty
to married women'] +
      df1['Harassment At Workplace'] + df1['Kidnapping / Abduction'] + df1['Maintenance Claim'] +
df1['Miscellaneous'] + df1['Murder'] +
      df1['Outraging Modesty of Women'] + df1['Police Apathy against women'] + df1['Right to live with
dignity'] +
      df1['Sexual harassment including sexual harassment at workplace'] + df1['Shelter & Rehabilitation
of Victims'] +
      df1["Women's right of custody of children in the event of divorce"]

df1["total_crimes"] = total
```



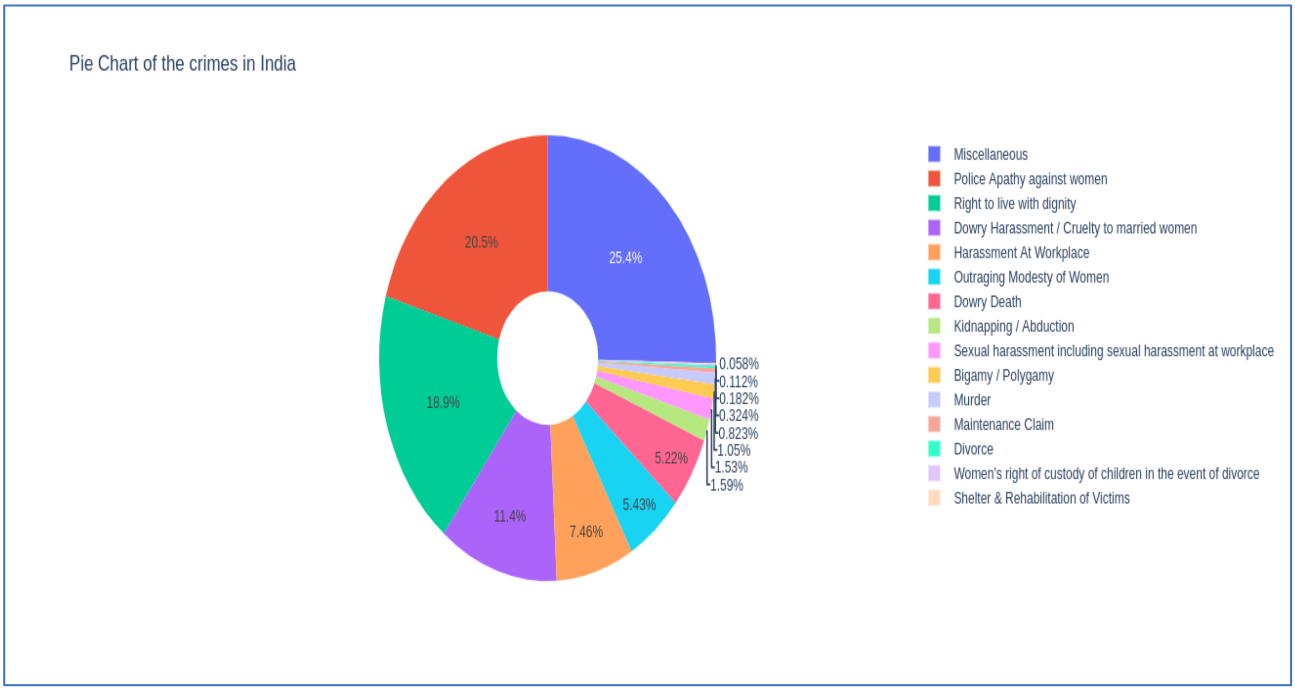
```
fig = make_subplots(rows = 9, cols = 2, shared_xaxes=True,horizontal_spacing=0.3,
        vertical_spacing=0.1,subplot_titles=(['Bigamy / Polygamy','Divorce','Dowry Death','Dowry
Harassment / Cruelty to married women',
        'Harassment At Workplace','Kidnapping / Abduction','Maintenance
Claim','Miscellaneous','Murder',
        'Outraging Modesty of Women','Police Apathy against women','Right to live with dignity',
        'Sexual harassment including sexual harassment at workplace','Shelter & Rehabilitation of
Victims',
        "Women's right of custody of children in the event of divorce","total_crimes"]))


fig.add_trace(go.Scatter(x = df1.index, y = df1['Bigamy / Polygamy']),row = 1, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Divorce']),row = 1, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Dowry Death']),row = 2, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Dowry Harassment / Cruelty to married women']),row =
2, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Harassment At Workplace']),row = 3, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Kidnapping / Abduction']),row = 3, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Maintenance Claim']),row = 4, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Miscellaneous']),row = 4, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Murder']),row = 5, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Outraging Modesty of Women']),row = 5, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Police Apathy against women']),row = 6, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Right to live with dignity']),row = 6, col = 2)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Sexual harassment including sexual harassment at
workplace']),row = 7, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['Shelter & Rehabilitation of Victims']),row = 7, col =
2)
fig.add_trace(go.Scatter(x = df1.index, y = df1["Women's right of custody of children in the event of
divorce"]),row = 8, col = 1)
fig.add_trace(go.Scatter(x = df1.index, y = df1['total_crimes']),row = 8, col = 2)

fig.update_layout(height=2900,width=700,showlegend=False)

fig.show()
```

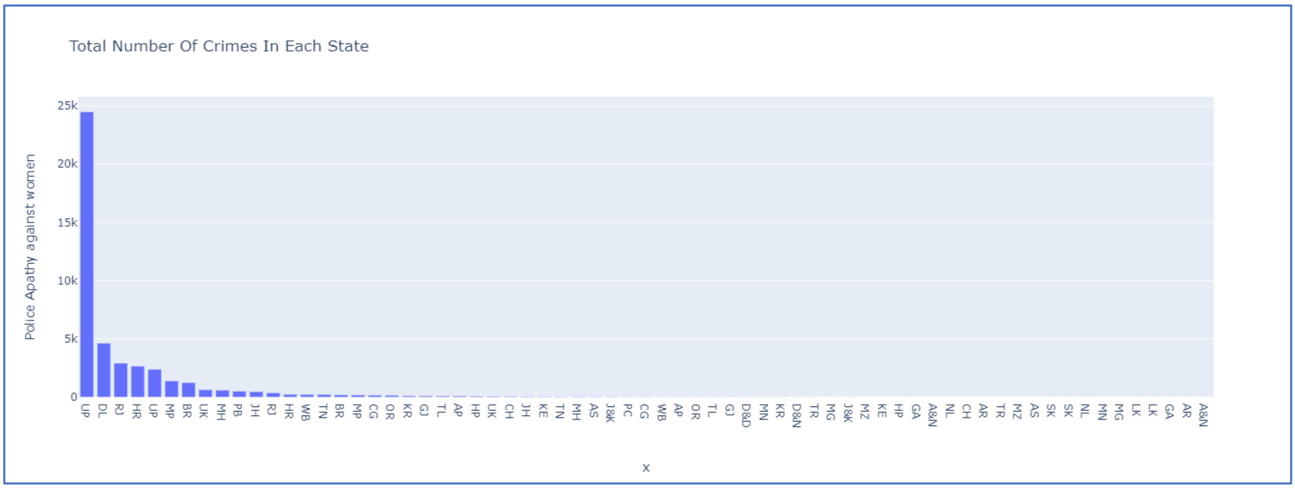






```
df_top_crimes=pd.DataFrame(columns=['crimes',"total"])
for i in crimes:
    df_top_crimes=df_top_crimes.append({'crimes':i ,'total':df[i].sum(axis=0)},ignore_index=True)
fig = go.Figure(data=[go.Pie(labels=df_top_crimes['crimes'], values= df_top_crimes['total'], hole=.3)])
fig.update_layout(title_text = "Pie Chart of the crimes in India")
fig.show()
states=df['State'].unique()
df_state=pd.DataFrame()
for i in crimes:
    df_state_crimes=df.groupby(['State'])[i].sum()
    df_state[i]=df_state_crimes
for i in crimes:
  fig = px.bar(df_state, x = df_state.index,y =i, title = "Total Number Of Crimes In Each State")
  fig.update_xaxes(categoryorder = 'total descending')
  fig.show()
```





• Cyber Crime

The data is visualized using the python libraries like matplotlib, seaborn, plotly.

1. The pie chart in figure (1) shows the percentage of each Cybercrime from 2008 to 2019 for all States and Union Territories. From the chart it can be concluded that only Computer related offences are 48.5% of the total Cybercrimes, while other offences under IT Act, 2000 that are not mentioned in the sheets from 2008 to 2019 account for 20.1% of total cybercrime, 17.4% of Cybercrime is of publication or transmission of obscene / sexually explicit material in electronic form and other offences contribute for less than 10%.

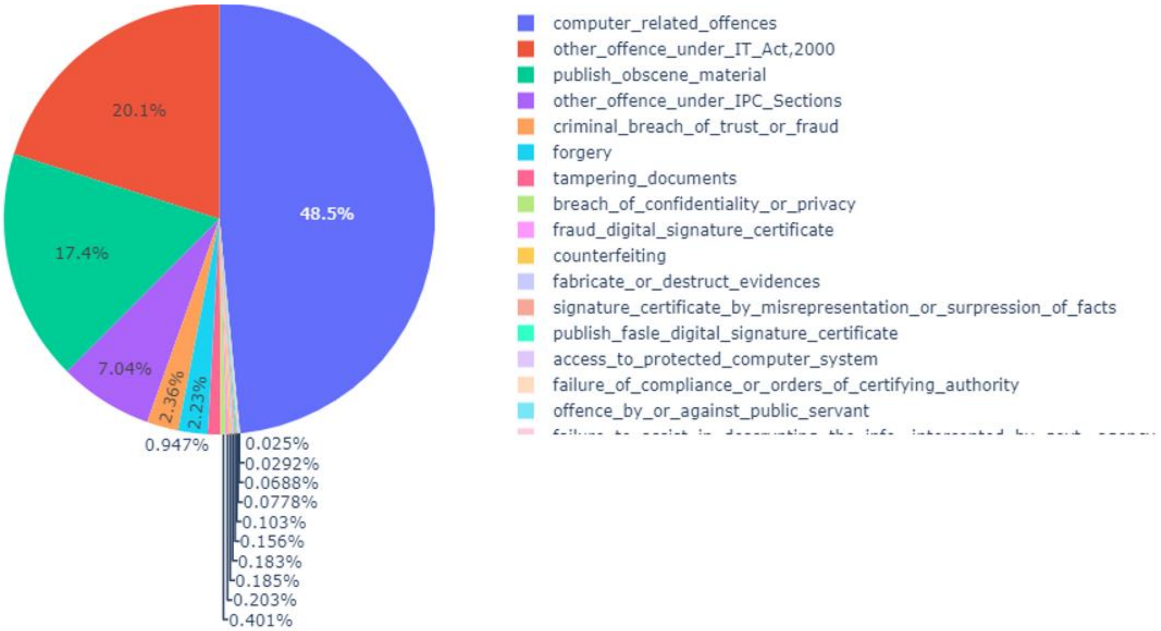


Figure (1)

2. The pie chart in the figure (2) shows the percentage of total Cybercrime in States and Union Territories from 2008 to 2019. From the chart it can be concluded that percentage of Cybercrime is highest in Uttar Pradesh i.e., 23.4% followed by Karnataka 21.9% and Maharashtra 14% while the Cybercrime percentage in other States and UTs are less than 6%.

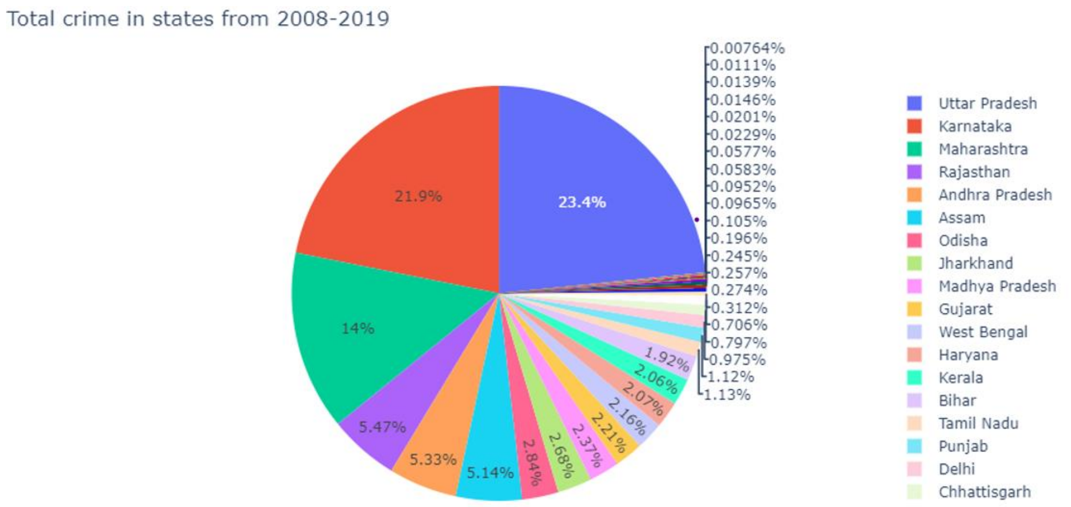
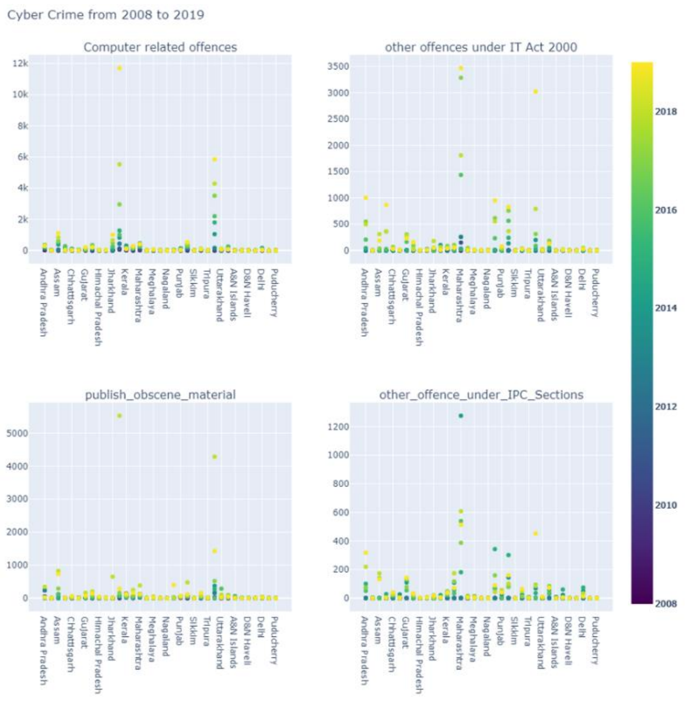


Figure (2)

From figure (1) and figure (2) it is clear that only few crimes contribute for more than 75% of Cybercrimes and same for States and UTs that only few of them are more vulnerable to Cybercrime attacks. So, for the better visualisation of these few crimes one can refer to the figure ().

Figure (3)



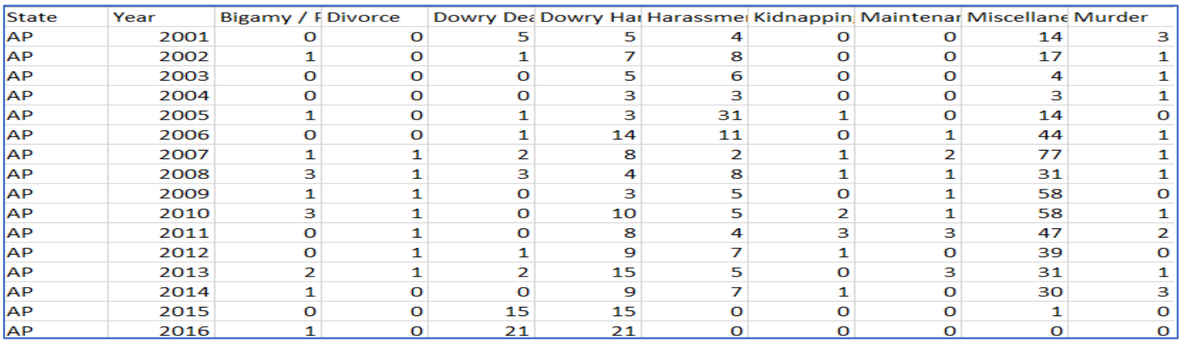
The figure (3) above consists of scatter plots of four types of cybercrimes from 2008 to 2019. It is clearly visible that the graphs of all cybercrimes are increasing with the year passing. Almost all the crimes are highest in Karnataka, Maharashtra and Uttar Pradesh

Methodology and Results

• Crime Against Women

1) Using Linear Regression

The data used in the research is crucial for discovering patterns, especially in crime analysis. The data was acquired from the National Commission for Women and includes ['Bigamy/Polygamy,' 'Divorce,' 'Dowry Death,' 'Dowry Harassment/Cruelty to Married Women, Kidnapping/Abduction'] among other crimes.

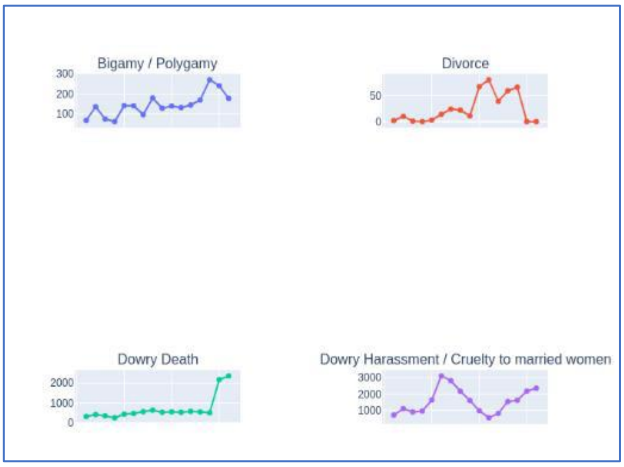


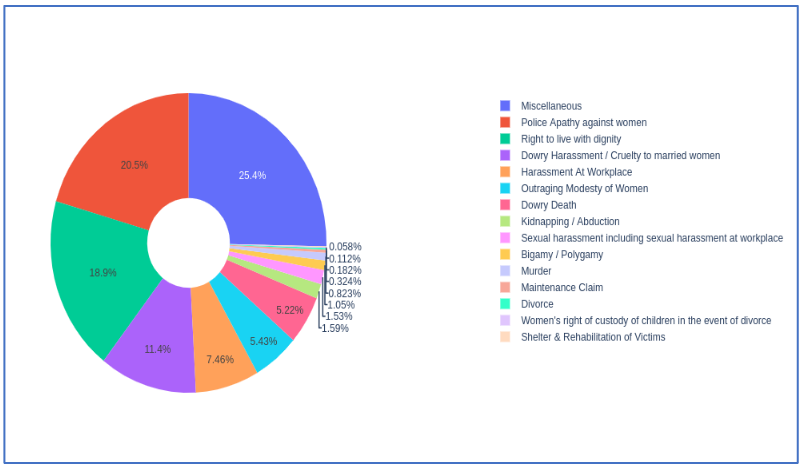
State	Year	Bigamy / F	Divorce	Dowry Dea	Dowry Ha	Harassme	Kidnappin	Maintenar	Miscellane	Murder
AP	2001	0	0	5	5	4	0	0	14	3
AP	2002	1	0	1	7	8	0	0	17	1
AP	2003	0	0	0	5	6	0	0	4	1
AP	2004	0	0	0	3	3	0	0	3	1
AP	2005	1	0	1	3	31	1	0	14	0
AP	2006	0	0	1	14	11	0	1	44	1
AP	2007	1	1	2	8	2	1	2	77	1
AP	2008	3	1	3	4	8	1	1	31	1
AP	2009	1	1	0	3	5	0	1	58	0
AP	2010	3	1	0	10	5	2	1	58	1
AP	2011	0	1	0	8	4	3	3	47	2
AP	2012	0	1	1	9	7	1	0	39	0
AP	2013	2	1	2	15	5	0	3	31	1
AP	2014	1	0	0	9	7	1	0	30	3
AP	2015	0	0	15	15	0	0	0	1	0
AP	2016	1	0	21	21	0	0	0	0	0

Dataset for years 2001-2016 for a particular state

Various regression techniques can be implemented to predict the total number of crimes occurring in each state for a particular year. This report uses linear regression to achieve the same. In the case of visualizations, different python libraries were implemented to yield informative charts. Among visualizations, 3 types of plots were included, i.e., scatter plot, pie chart, and bar chart.

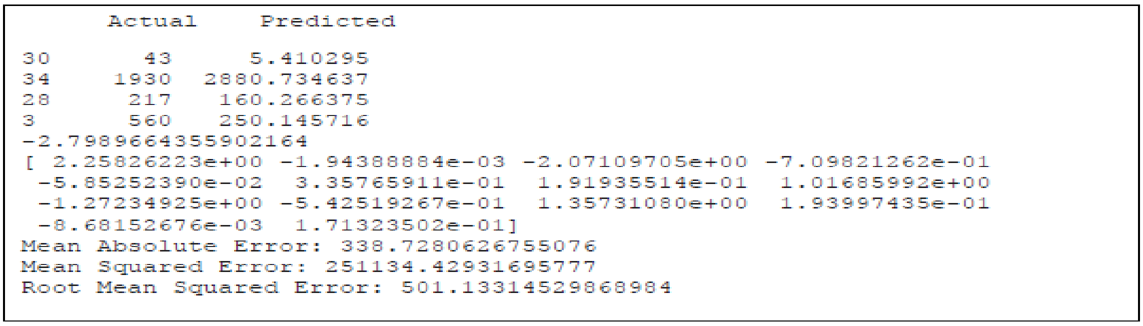
Few Snippets of the visualizations obtained





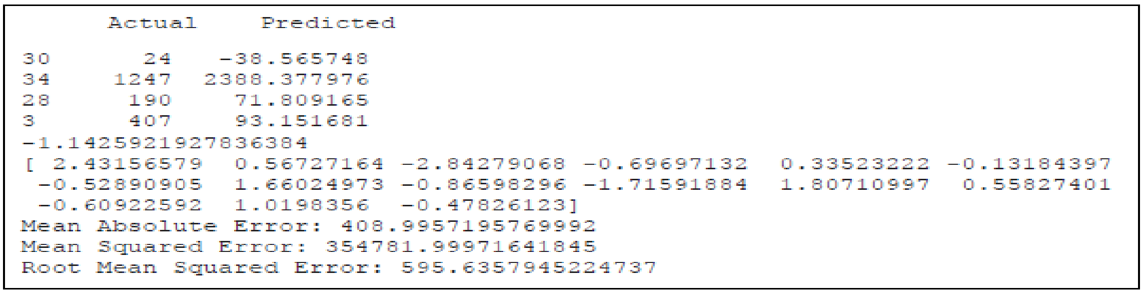
The interpretation of the scatter chart, which is the plot for the total number of crimes committed each year, shows an irregular curve. For instance, crimes like Bigamy/Polygamy, Divorce, Harassment at the workplace, and Murder have a downfall in the number of cases in the coming years. In contrast, crimes like Dowry Death, Sexual Harassment at the workplace, and Cruelty to women have witnessed an increase in the number of cases in recent years. The insights from the pie chart yield the proportion of each type of crime occurring in the years 2001-2016. The attribute Miscellaneous shares the major proportion in the pie chart with 25.4%. Followed by Miscellaneous, the columns like Police Apathy against Women and the Right to live with Dignity occupy 20.5% and 18.9% of the pie chart. The last visualization, i.e., the bar graph, is the plot for each type of crime. The plot is created with the total number of crimes and states in descending order. From the bar chart, it can be seen that the state of Uttar Pradesh has seen the largest number of crimes against women. After creating the informative visualizations, a prediction for the total number of crimes for the years 2016, 2017, 2018 and 2019. Forecasting was achieved through Linear Regression.

1. 2016



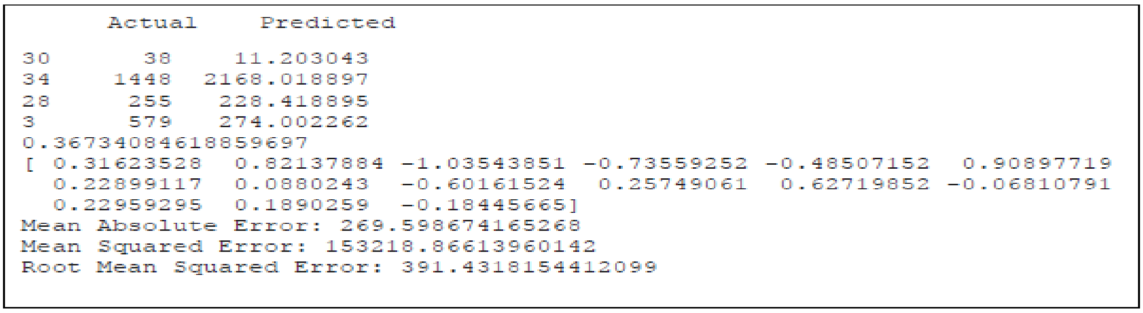
```
      Actual     Predicted

30       43      5.410295
34     1930   2880.734637
28      217    160.266375
3       560    250.145716
-2.7989664355902164
[ 2.25826223e+00 -1.94388884e-03 -2.07109705e+00 -7.09821262e-01
 -5.85252390e-02  3.35765911e-01  1.91935514e-01  1.01685992e+00
 -1.27234925e+00 -5.42519267e-01  1.35731080e+00  1.93997435e-01
 -8.68152676e-03  1.71323502e-01]
Mean Absolute Error: 338.7280626755076
Mean Squared Error: 251134.42931695777
Root Mean Squared Error: 501.13314529868984
```

2. 2017



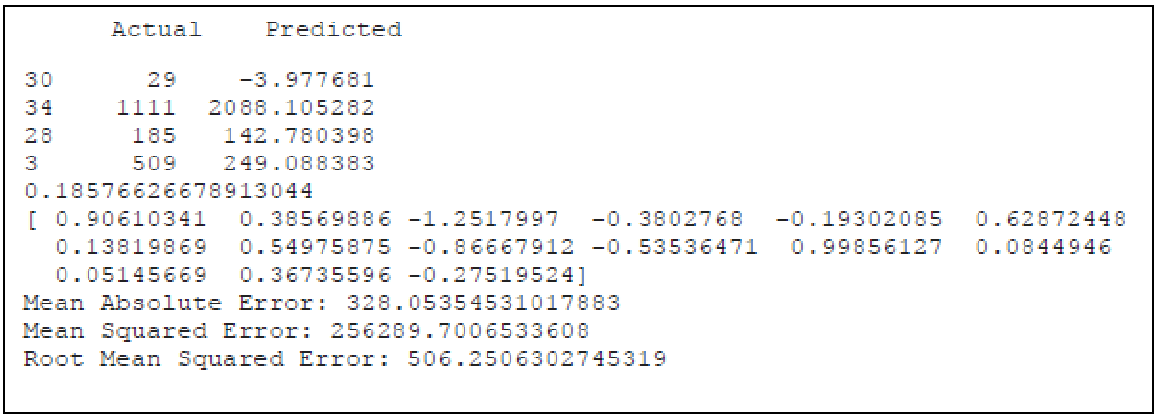
```
      Actual     Predicted

30       24    -38.565748
34     1247   2388.377976
28      190     71.809165
3       407     93.151681
-1.1425921927836384
[ 2.43156579  0.56727164 -2.84279068 -0.69697132  0.33523222 -0.13184397
 -0.52890905  1.66024973 -0.86598296 -1.71591884  1.80710997  0.55827401
 -0.60922592  1.0198356  -0.47826123]
Mean Absolute Error: 408.9957195769992
Mean Squared Error: 354781.99971641845
Root Mean Squared Error: 595.6357945224737
```

3. 2018



```
      Actual     Predicted

30       38     11.203043
34     1448   2168.018897
28      255    228.418895
3       579    274.002262
0.36734084618859697
[ 0.31623528  0.82137884 -1.03543851 -0.73559252 -0.48507152  0.90897719
  0.22899117  0.0880243  -0.60161524  0.25749061  0.62719852 -0.06810791
  0.22959295  0.1890259  -0.18445665]
Mean Absolute Error: 269.598674165268
Mean Squared Error: 153218.86613960142
Root Mean Squared Error: 391.4318154412099
```

4. 2019



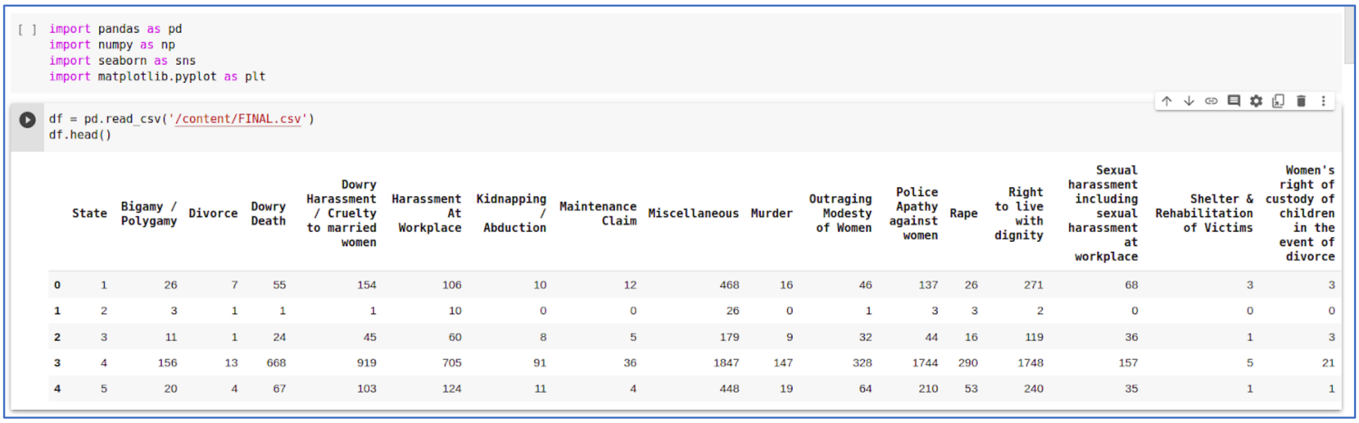
```
      Actual    Predicted

30       29    -3.977681
34     1111  2088.105282
28      185   142.780398
3       509   249.088383
0.18576626678913044
[ 0.90610341  0.38569886 -1.2517997  -0.3802768  -0.19302085  0.62872448
  0.13819869  0.54975875 -0.86667912 -0.53536471  0.99856127  0.0844946
  0.05145669  0.36735596 -0.27519524]
Mean Absolute Error: 328.05354531017883
Mean Squared Error: 256289.7006533608
Root Mean Squared Error: 506.2506302745319
```

The accuracy of the model, in general, is approximated to be 52.4%. The information gathered using intercept and coefficient along with methods used for evaluating model efficiency suggest that the prediction is not that accurate. There could be numerous reasons behind this inaccurate prediction. One can be the values of the data present in the dataset. Another reason could be the improper cleaning of the dataset, like instead of replacing Nan with 0 values, it could have been substituted with average values. Finally, this paper proposes that there are many areas of improvement which can make this forecasting much more accurate.

Using K-means Clustering

Our code was written in the Python programming language.

To begin, we used the Numpy, pandas, seaborn, and matplotlib packages in Python to perform basic visualization and data import. “df” is the data frame that we specified to hold all of the data from the transformed dataset. The function "df. head ()" prints the data frame’s first five rows.



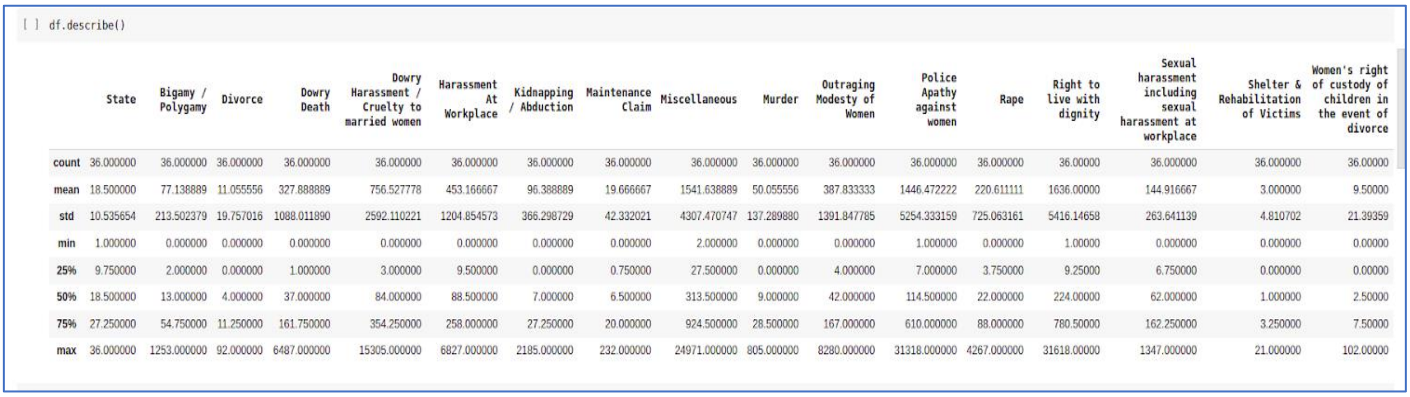
```
[ ] import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
df = pd.read_csv('/content/FINAL.csv')
df.head()
```

	State	Bigamy / Polygamy	Divorce	Dowry Death	Dowry Harassment / Cruelty to married women	Harassment At Workplace	Kidnapping / Abduction	Maintenance Claim	Miscellaneous	Murder	Outraging Modesty of Women	Police Apathy against women	Rape	Right to live with dignity	Sexual harassment including sexual harassment at workplace	Shelter & Rehabilitation of Victims	Women's right of custody of children in the event of divorce
0	1	26	7	55	154	106	10	12	468	16	46	137	26	271	68	3	3
1	2	3	1	1	1	10	0	0	26	0	1	3	3	2	0	0	0
2	3	11	1	24	45	60	8	5	179	9	32	44	16	119	36	1	3
3	4	156	13	668	919	705	91	36	1847	147	328	1744	290	1748	157	5	21
4	5	20	4	67	103	124	11	4	448	19	64	210	53	240	35	1	1

As can be seen in the accompanying graphic, each state has a unique number in column 1 ranging from 1 to 36. The following columns are for various crimes

against women, with each row corresponding to the number of offences committed in that state.
The describe () method in Pandas is used to display some basic statistical features of a data frame or a sequence of numeric values, such as percentile, mean, and standard deviation. The total count of all columns is 36, indicating that there are 36 rows in the dataset. We can also get the maximum and minimum number of women who have been victims of crimes.



```
[ ] df.describe()
```

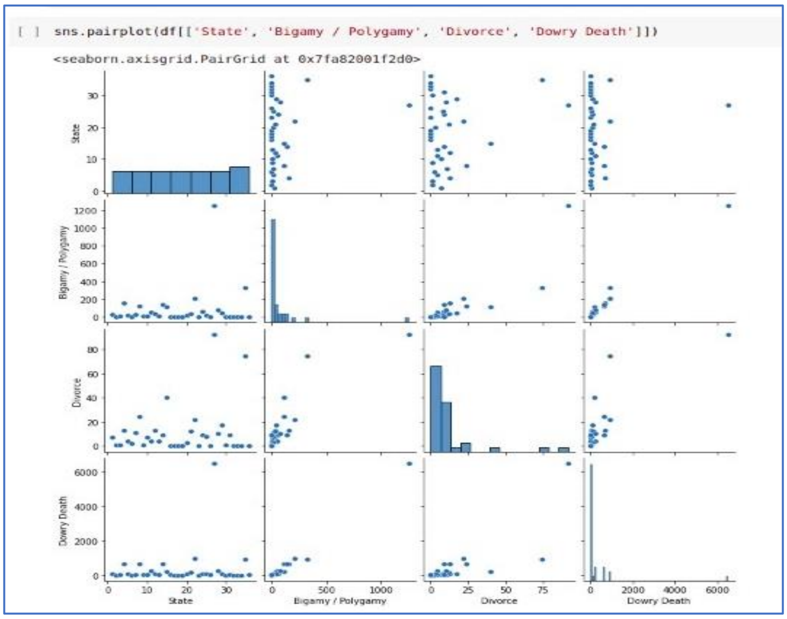
	State	Bigamy / Polygamy	Divorce	Dowry Death	Dowry Harassment / Cruelty to married women	Harassment At Workplace	Kidnapping / Abduction	Maintenance Claim	Miscellaneous	Murder	Outraging Modesty of Women	Police Apathy against women	Rape	Right to live with dignity	Sexual harassment including sexual harassment at workplace	Shelter & Rehabilitation of Victims	Women's right of custody of children in the event of divorce
count	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000	36.00000	36.000000	36.000000	36.00000
mean	18.500000	77.138889	11.055556	327.888889	756.527778	453.166667	96.388889	19.666667	1541.638889	50.055556	387.833333	1446.472222	220.611111	1636.00000	144.916667	3.000000	9.50000
std	10.535654	213.502379	19.757016	1088.011890	2592.110221	1204.854573	366.298729	42.332021	4307.470747	137.289880	1391.847785	5254.333159	725.063161	5416.14658	263.641139	4.810702	21.39359
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	1.000000	0.000000	1.00000	0.000000	0.000000	0.00000
25%	9.750000	2.000000	0.000000	1.000000	3.000000	9.500000	0.000000	0.750000	27.500000	0.000000	4.000000	7.000000	3.750000	9.25000	6.750000	0.000000	0.00000
50%	18.500000	13.000000	4.000000	37.000000	84.000000	88.500000	7.000000	6.500000	313.500000	9.000000	42.000000	114.500000	22.000000	224.00000	62.000000	1.000000	2.50000
75%	27.250000	54.750000	11.250000	161.750000	354.250000	258.000000	27.250000	20.000000	924.500000	28.500000	167.000000	610.000000	88.000000	780.50000	162.250000	3.250000	7.50000
max	36.000000	1253.000000	92.000000	6487.000000	15305.000000	6827.000000	2185.000000	232.000000	24971.000000	805.000000	8280.000000	31318.000000	4267.000000	31618.00000	1347.000000	21.000000	102.00000

Pair Plots are a simple way to visualize relationships between each variable.
In the following pair-plots, we have visualized the relationship between:
1.a State vs State
1.b State vs Bigamy/Polygamy
1.c State vs Divorce
1.d State vs Dowry Death

2.a Bigamy/Polygamy vs State
2.b Bigamy/Polygamy vs Bigamy/Polygamy
2.c Bigamy/Polygamy vs Divorce
2.d Bigamy/Polygamy vs Dowry Death

3.a Divorce vs State
3.b Divorce vs Bigamy/polygamy
3.c Divorce vs Divorce
3.d Divorce vs Dowry Death

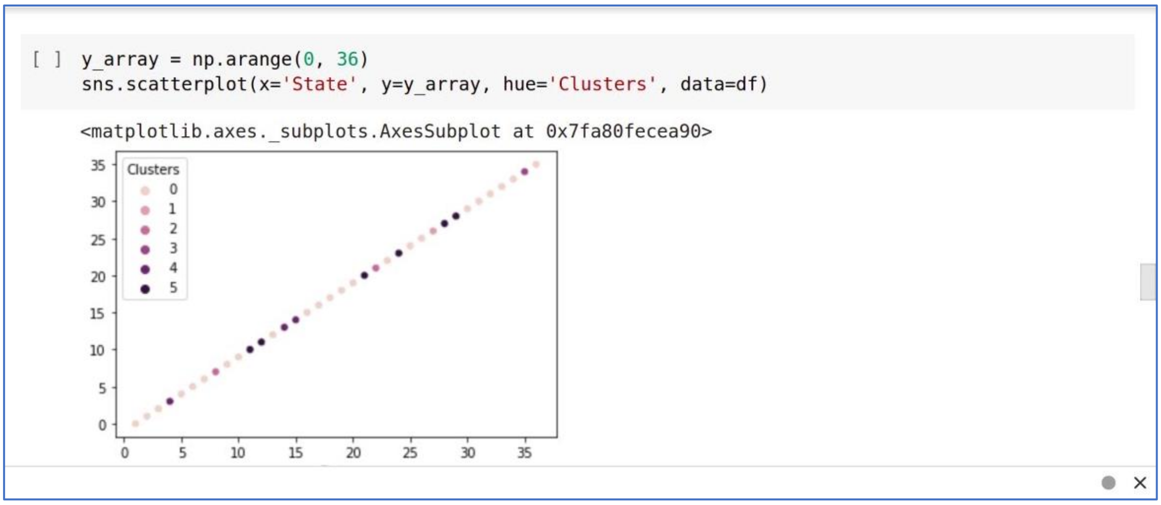
4.a Dowry Death vs State
4.b Dowry Death vs Bigamy/Polygamy
4.c Dowry Death vs Divorce
4.d Dowry Death vs Dowry Death



```
[ ] sns.pairplot(df[['State', 'Bigamy / Polygamy', 'Divorce', 'Dowry Death']])
```

```
<seaborn.axisgrid.PairGrid at 0x7fa82001f2d0>
```

We used the sklearn library to construct k-means clustering. We created a variable called k means that split all of the states into six groups with similar crime rates. To fit our clustering model, we employed all 17 crimes.



```
[ ] y_array = np.arange(0, 36)
    sns.scatterplot(x='State', y=y_array, hue='Clusters', data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa80fecea90>
```

0-AP (Andhra Pradesh), AR (Arunachal Pradesh), AS(Assam), CG(Chhattisgarh) GA(Goa), GJ(Gujarat), HP (Himachal Pradesh), J&K (Jammu and Kashmir), KE(Kerala), MN(Manipur), MG(Meghalaya), MZ(Mizoram), NL(Nagaland), OR(Orissa) SK(Sikkim), TL(Telangana), A&N (Andaman and Nicobar), CH(Chandigarh), D&N (Dadar and Nagar haveli), D&D (Daman and Diu), LK(Lakshadweep), PC(Pondicherry)

1 – UP (Uttar Pradesh)

2 - DL(Delhi)

3 - HR(Haryana), RJ(Rajasthan)

4 – MP (Madhya Pradesh), MH(Maharashtra)

5- JH(Jharkhand), KR(Karnataka), PB(Punjab), TN (Tamil Nadu), UK(Uttarakhand)

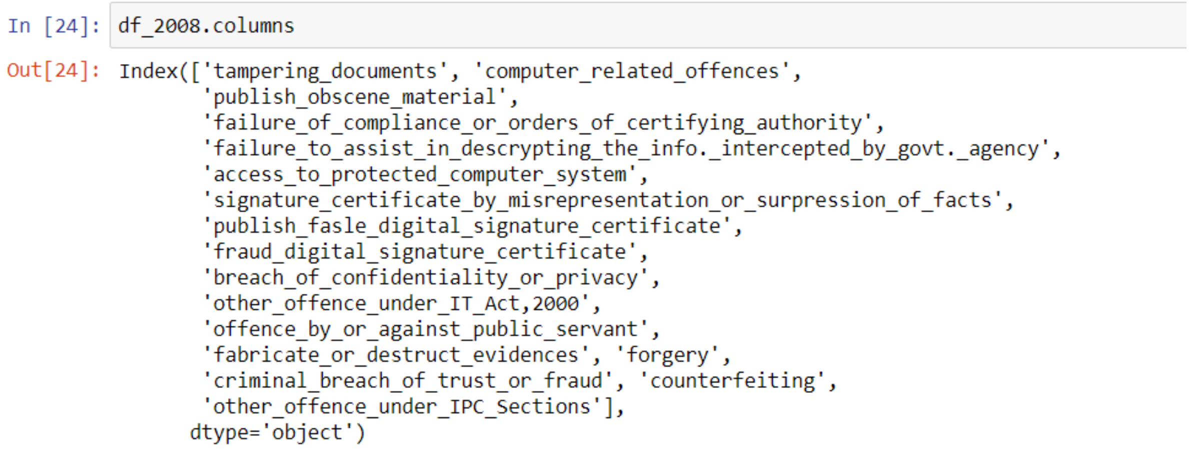
• Cyber Crime

This paper predicts the values for different type of cybercrimes under IT Act, 2000 and IPC sections for year 2020 using the data from 2008 to 2019. This work uses data analysis, data visualization and machine learning techniques to predict data for 2020 of all States and Union Territories. The language used for implementing various regression algorithms is python, regression models Gradient Boosting regression, ridge regression is applied for prediction of cybercrime in 2020. The methodology and results used here can be described in following steps:

Data Collection and Preprocessing → Data Analysis → Using Regression models → Prediction and Analysing Predictions

Data collection and preprocessing:

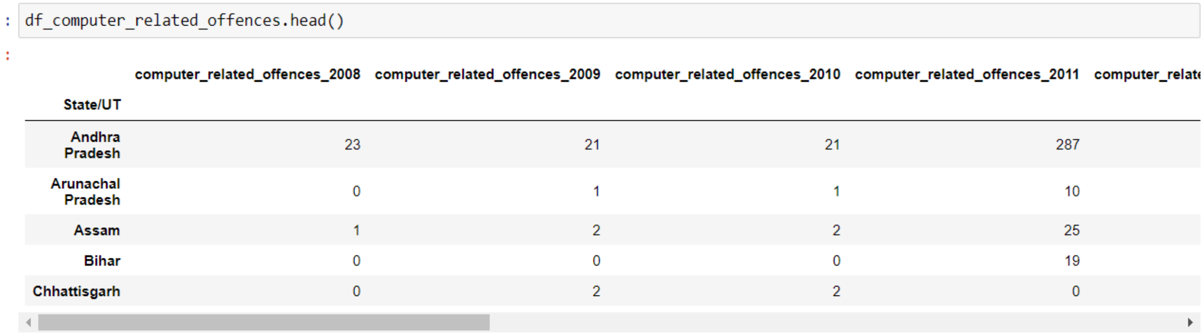
The data of cybercrime is collected from National Crime Records Bureau (ncrb.gov.in). It contains the record for different types of cybercrimes from 2008 to 2019. The type of cybercrimes used in this work after following preprocessing step to get common columns for all years are:



```
In [24]: df_2008.columns

Out[24]: Index(['tampering_documents', 'computer_related_offences',
                'publish_obscene_material',
                'failure_of_compliance_or_orders_of_certifying_authority',
                'failure_to_assist_in_descrypting_the_info._intercepted_by_govt._agency',
                'access_to_protected_computer_system',
                'signature_certificate_by_misrepresentation_or_surpression_of_facts',
                'publish_fasle_digital_signature_certificate',
                'fraud_digital_signature_certificate',
                'breach_of_confidentiality_or_privacy',
                'other_offence_under_IT_Act,2000',
                'offence_by_or_against_public_servant',
                'fabricate_or_destruct_evidences', 'forgery',
                'criminal_breach_of_trust_or_fraud', 'counterfeiting',
                'other_offence_under_IPC_Sections'],
               dtype='object')
```

Data Analysis: (we can write statistics analysis taken using pandas example the states with almost no crime and the once with highest also same for type of crime)

For analysing, dataframes are created for all types of cybercrimes from 2008 to 2019 for all States and Union territories. Like figure () shows the dataframe of cybercrime computer related offences for all years. All these dataframes all also used for prediction of particular type of cybercrime.



```
df_computer_related_offences.head()
```

	computer_related_offences_2008	computer_related_offences_2009	computer_related_offences_2010	computer_related_offences_2011	computer_relat
State/UT					
Andhra Pradesh	23	21	21	287	
Arunachal Pradesh	0	1	1	10	
Assam	1	2	2	25	
Bihar	0	0	0	19	
Chhattisgarh	0	2	2	0	

Regression models:

Two regression models are used for prediction, Gradient Boosting Regression and Ridge Regression.

Implementing Gradient Boosting regression model:

Here, all cybercrime types from 2008 to 2019 are used for predicting that particular type of cybercrime for year 2020.

For applying model and getting prediction results the data is divided into standard 70:30 ratio, where 70 percent of the data is utilised for training purpose and 30 percent is used for testing the performance of the model. For this purpose, sklearn module model_selection’s method train_test_split() is used. Train_test_split method gives divided data as X_train, X_test, y_train, y_test.

Next, to train the gradient boosting model x_train and y_train are fed to it, so that it learns the various relations among the variables. Then model predicts the values for y_test using it training understanding, x_test is used for prediction. For optimizing results, the parameters taken by the Gradient boosting model like learning rate, max depth, n_estimators are optimized as per data requirement.

The performance of the model is observed by root mean square error and mean absolute error using the predicted values (i.e., predicted with x_test values) and the actual values (i.e., y_test values). By minimising the values of root mean square error and mean absolute error the performance of the model is improved.

Implementing Ridge regression model: Ridge regression is essentially a variation of Linear regression where multicollinearity is properly dealt with. The multicollinearity considered has large variance and the least square is unbiased which causes a relatively large deviation in the predicted values and actual values.

The general steps involved in implementing Ridge Regression are the same as for Gradient Boosting

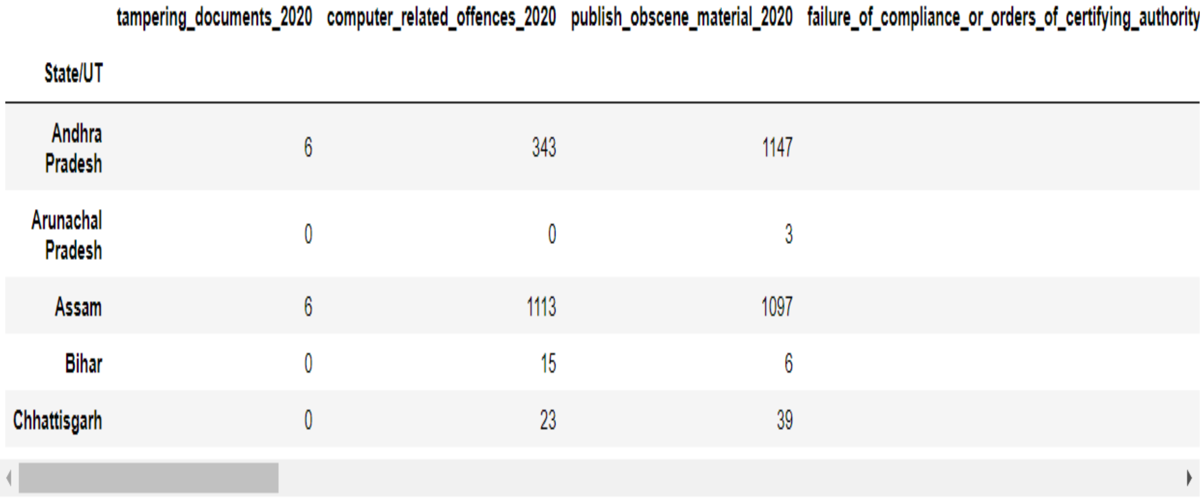
In order to implement Ridge Regression properly we must choose the value for the penalty term such that it causes strong regularization. In case of Ridge Regression, the penalty term that needs tuning is the ‘alpha’ parameter. A method is created to ease the tuning process and alpha values are run through to check MSE and RMSE values before fitting the model and a value is selected. Since selection of alpha values can determine the fitting of the data i.e., overfitting or underfitting, it is imperative to select the best value.

Upon applying just, a single model to the dataset the predictions given weren’t good (presence of negative values). Consequently, two models were adopted, where for the first one alpha was selected as ‘0.1’ and for the second it was selected as ‘1000’. The final prediction was made by using the absolute average sum of the predictions from the two models to get a more accurate reading.

Predictions and Prediction analysis:

Using Gradient Boosting model:

Using the gradient boosting model dataframe created for 2020 cybercrime values.



	tampering_documents_2020	computer_related_offences_2020	publish_obscene_material_2020	failure_of_compliance_or_orders_of_certifying_authority
State/UT				
Andhra Pradesh	6	343	1147	
Arunachal Pradesh	0	0	3	
Assam	6	1113	1097	
Bihar	0	15	6	
Chhattisgarh	0	23	39	

The predicted values for most vulnerable cybercrime in all states for 2020 is shown in the figure (4) below

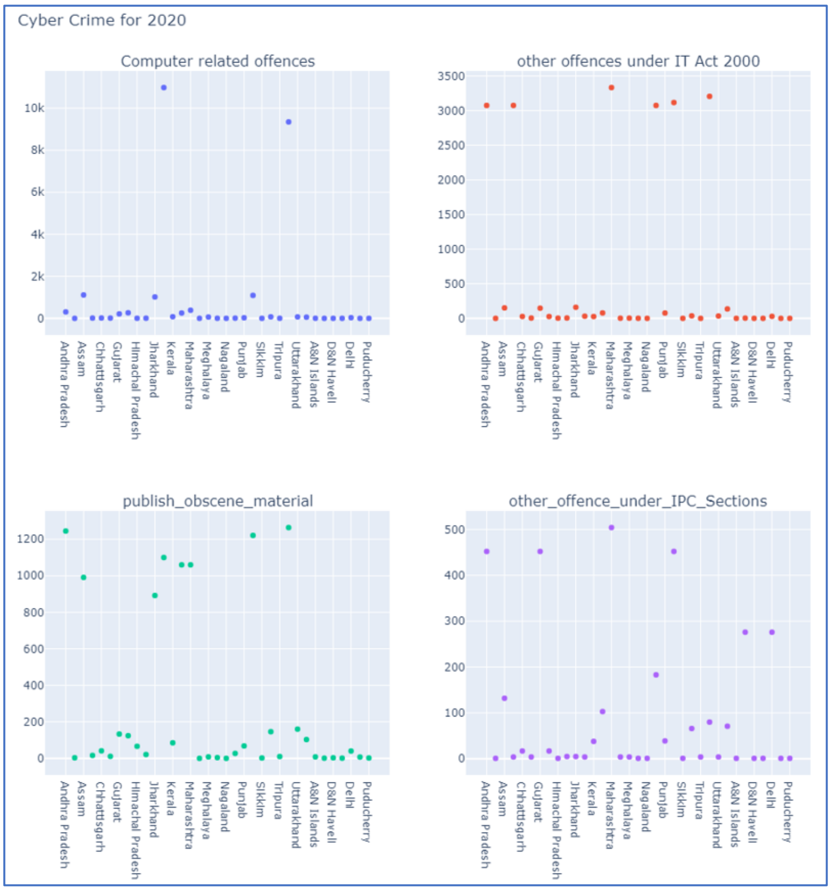
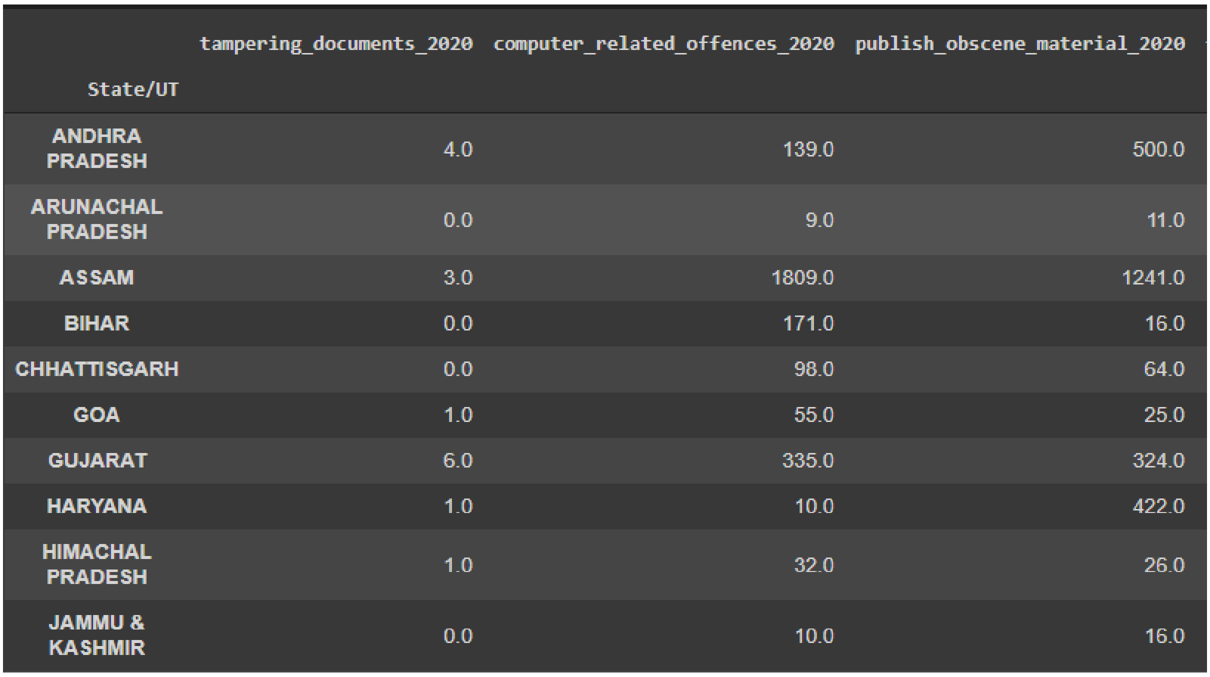


Figure (4)

Using ridge regression model:

Output For Predicting Cyber Crimes Committed In 2020 Using Ridge Regression:

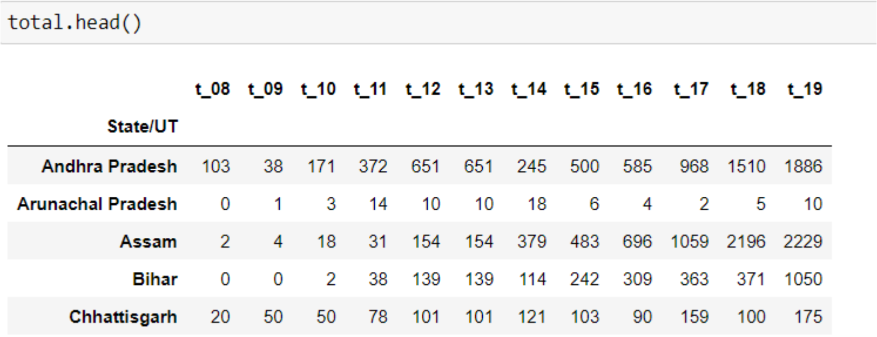


State/UT	tampering_documents_2020	computer_related_offences_2020	publish_obscene_material_2020
ANDHRA PRADESH	4.0	139.0	500.0
ARUNACHAL PRADESH	0.0	9.0	11.0
ASSAM	3.0	1809.0	1241.0
BIHAR	0.0	171.0	16.0
CHHATTISGARH	0.0	98.0	64.0
GOA	1.0	55.0	25.0
GUJARAT	6.0	335.0	324.0
HARYANA	1.0	10.0	422.0
HIMACHAL PRADESH	1.0	32.0	26.0
JAMMU & KASHMIR	0.0	10.0	16.0

Prediction analysis:

For cross checking the results of prediction:

Using the Gradient Boosting model, and taking the total cybercrime values of each year for States and Union territories the total cybercrime value for 2020 is predicted.

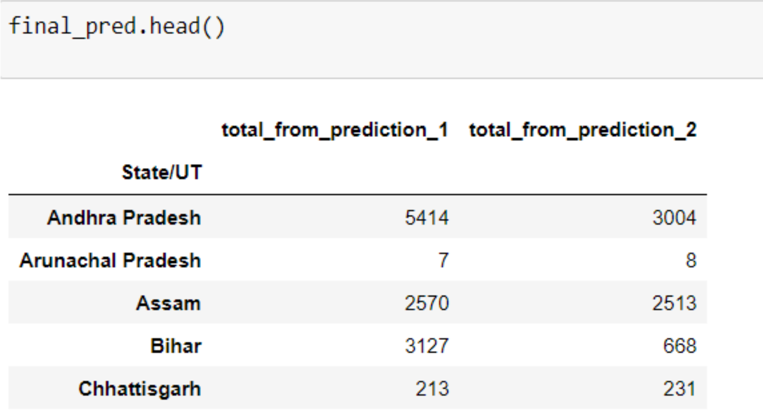


```
total.head()
```

State/UT	t_08	t_09	t_10	t_11	t_12	t_13	t_14	t_15	t_16	t_17	t_18	t_19
Andhra Pradesh	103	38	171	372	651	651	245	500	585	968	1510	1886
Arunachal Pradesh	0	1	3	14	10	10	18	6	4	2	5	10
Assam	2	4	18	31	154	154	379	483	696	1059	2196	2229
Bihar	0	0	2	38	139	139	114	242	309	363	371	1050
Chhattisgarh	20	50	50	78	101	101	121	103	90	159	100	175

Figure (5)

Figure (5) shows the dataframe with columns containing total cybercrime value for a particular year from 2008 to 2019 using this dataframe and applying gradient boosting regression model the total cybercrime values for States and Union Territories is predicted for 2020. Refer to figure (6) dataframe’s column total_prediction_2 for predicted value obtained by this method.



```
final_pred.head()
```

State/UT	total_from_prediction_1	total_from_prediction_2
Andhra Pradesh	5414	3004
Arunachal Pradesh	7	8
Assam	2570	2513
Bihar	3127	668
Chhattisgarh	213	231

Figure (6)

Figure (6) shows the predicted values of cybercrime for 2020 using two methods, first by predicting the values for each cybercrime type and then using pandas for computing the total cybercrime for States and Union Territories, while second by predicting values for total cybercrime in 2020 using dataframe shown in figure (5).

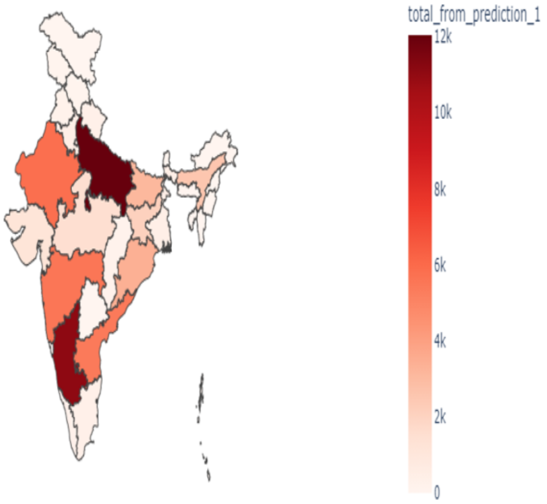


Figure (7)

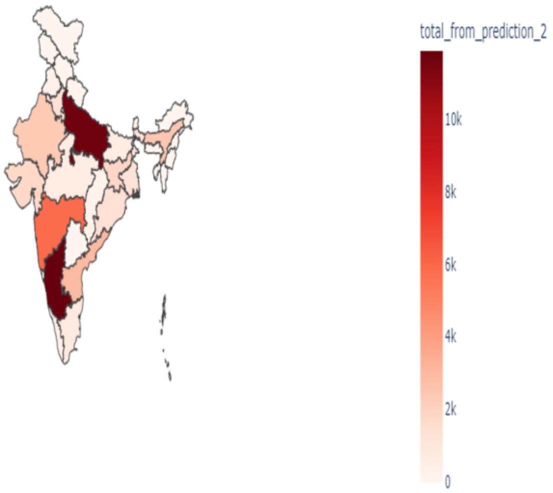
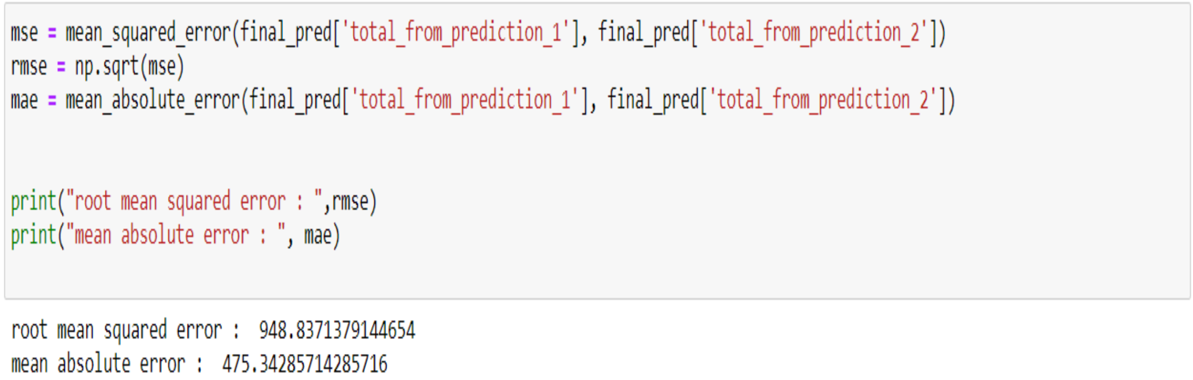


Figure (8)

figure (7) and figure (8) compare the values predicted for 2020 by two method discussed above.

Now, difference between two predictions as performed above can be observed by using sklearn module metric's methods mean_squared_error and mean_absolute_error.



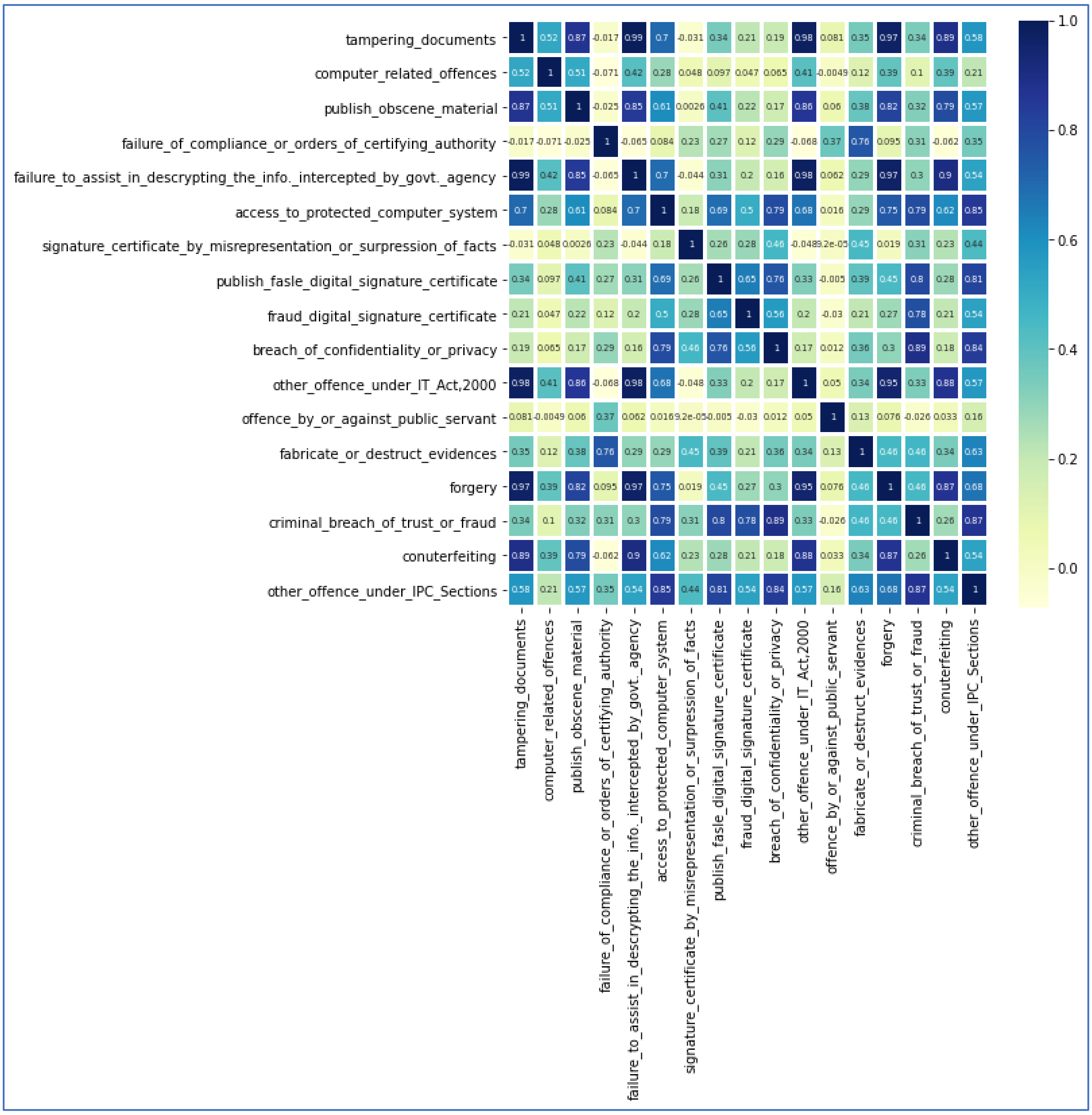
```
mse = mean_squared_error(final_pred['total_from_prediction_1'], final_pred['total_from_prediction_2'])
rmse = np.sqrt(mse)
mae = mean_absolute_error(final_pred['total_from_prediction_1'], final_pred['total_from_prediction_2'])


print("root mean squared error : ",rmse)
print("mean absolute error : ", mae)
```

```
root mean squared error :  948.8371379144654
mean absolute error :  475.34285714285716
```

Although using Ridge Regression a prediction was finalised, Gradient Boosting will still be considered more suitable for the dataset in use as it was able to generate a good set of values for predictions with lesser code, in other words Gradient Boosting was more efficient for the dataset selected.

Correlation of Crimes with each other for the year 2019:

The correlation sheds light on the intra-relations the crimes have. Since cyber criminals are often known to have more than a single motive to commit crimes, it is important to consider this as prediction with numbers is not always correct- with personal motives continuously changing, predicting a fixed number can only act as a guideline/ reference. Having such a correlation can help us predict the increase in number of a certain type of crimes due to the increase in another type of crime. This correlation shows the relationship between the different crimes for the year 2019.



	tampering_documents	computer_related_offences	publish_obscene_material	failure_of_compliance_or_orders_of_certifying_authority	failure_to_assist_in_descrypting_the_info._intercepted_by_govt._agency	access_to_protected_computer_system	signature_certificate_by_misrepresentation_or_surpression_of_facts	publish_fasle_digital_signature_certificate	fraud_digital_signature_certificate	breach_of_confidentiality_or_privacy	other_offence_under_IT_Act,2000	offence_by_or_against_public_servant	fabricate_or_destruct_evidences	forgery	criminal_breach_of_trust_or_fraud	conuterfeiting	other_offence_under_IPC_Sections
tampering_documents	1	0.52	0.87	-0.017	0.99	0.7	-0.031	0.34	0.21	0.19	0.98	0.081	0.35	0.97	0.34	0.89	0.58
computer_related_offences	0.52	1	0.51	-0.071	0.42	0.28	0.048	0.097	0.047	0.065	0.41	-0.0049	0.12	0.39	0.1	0.39	0.21
publish_obscene_material	0.87	0.51	1	-0.025	0.85	0.61	0.0026	0.41	0.22	0.17	0.86	0.06	0.38	0.82	0.32	0.79	0.57
failure_of_compliance_or_orders_of_certifying_authority	-0.017	-0.071	-0.025	1	-0.065	0.084	0.23	0.27	0.12	0.29	-0.068	0.37	0.76	0.095	0.31	-0.062	0.35
failure_to_assist_in_descrypting_the_info._intercepted_by_govt._agency	0.99	0.42	0.85	-0.065	1	0.7	-0.044	0.31	0.2	0.16	0.98	0.062	0.29	0.97	0.3	0.9	0.54
access_to_protected_computer_system	0.7	0.28	0.61	0.084	0.7	1	0.18	0.69	0.5	0.79	0.68	0.016	0.29	0.75	0.79	0.62	0.85
signature_certificate_by_misrepresentation_or_surpression_of_facts	-0.031	0.048	0.0026	0.23	-0.044	0.18	1	0.26	0.28	0.46	-0.048	9.2e-05	0.45	0.019	0.31	0.23	0.44
publish_fasle_digital_signature_certificate	0.34	0.097	0.41	0.27	0.31	0.69	0.26	1	0.65	0.76	0.33	-0.005	0.39	0.45	0.8	0.28	0.81
fraud_digital_signature_certificate	0.21	0.047	0.22	0.12	0.2	0.5	0.28	0.65	1	0.56	0.2	-0.03	0.21	0.27	0.78	0.21	0.54
breach_of_confidentiality_or_privacy	0.19	0.065	0.17	0.29	0.16	0.79	0.46	0.76	0.56	1	0.17	0.012	0.36	0.3	0.89	0.18	0.84
other_offence_under_IT_Act,2000	0.98	0.41	0.86	-0.068	0.98	0.68	-0.048	0.33	0.2	0.17	1	0.05	0.34	0.95	0.33	0.88	0.57
offence_by_or_against_public_servant	0.081	-0.0049	0.06	0.37	0.062	0.016	9.2e-05	-0.005	-0.03	0.012	0.05	1	0.13	0.076	-0.026	0.033	0.16
fabricate_or_destruct_evidences	0.35	0.12	0.38	0.76	0.29	0.29	0.45	0.39	0.21	0.36	0.34	0.13	1	0.46	0.46	0.34	0.63
forgery	0.97	0.39	0.82	0.095	0.97	0.75	0.019	0.45	0.27	0.3	0.95	0.076	0.46	1	0.46	0.87	0.68
criminal_breach_of_trust_or_fraud	0.34	0.1	0.32	0.31	0.3	0.79	0.31	0.8	0.78	0.89	0.33	-0.026	0.46	0.46	1	0.26	0.87
conuterfeiting	0.89	0.39	0.79	-0.062	0.9	0.62	0.23	0.28	0.21	0.18	0.88	0.033	0.34	0.87	0.26	1	0.54
other_offence_under_IPC_Sections	0.58	0.21	0.57	0.35	0.54	0.85	0.44	0.81	0.54	0.84	0.57	0.16	0.63	0.68	0.87	0.54	1

Acknowledgement

We would like to express our gratitude to Miranda House, University of Delhi, for hosting the Summer Workshop for investigative projects in multidisciplinary contexts, which provided us with an incredible opportunity to not only learn about things that aren't covered in textbooks, but also to brainstorm ideas and contribute to existing research. Under the supervision and assistance of the Department of Computer Science, the current project report addresses the issue "Crime Against Women" and "Cyber Security". Dr. Seema Agarwal, our mentor, deserves special thanks for her direction and assistance in finishing the project on time. We are grateful to everyone involved in this project, without whom our project would have lacked the necessary information to achieve our study's goal.

References

1. https://www.kaggle.com/marcogherbezza/crimes-against-women
2. https://easychair.org/publications/preprint/pD8r
3. https://www.geeksforgeeks.org/
4. https://ssi.edu.in/wp-content/uploads/2019/05/Internship-Report-by-Ms.-Tanisha-Khandelwal.pdf
5. https://towardsdatascience.com/classification-regression-and-prediction-whats-the-difference-5423d9efe4ec
6. https://www.researchgate.net/publication/336982992_Crime_against_Women_CAW_Analysis_and_Prediction_in_Tamilnadu_Police_Using_Data_Mining_Techniques
7. https://www.ijert.org/research/crime-against-women-analysis-and-prediction-IJERTV10IS050229.pdf
8. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
9. https://www.mdpi.com/2071-1050/12/10/4087/htm
10. https://www.ijtsrd.com/management/operations-management/18693/a-brief-study-on-cyber-crimes-and-it-act-in-india/dr-adv-mrs-neeta-deshpande
11. https://www.youtube.com/watch?v=3CC4N4z3GJc&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=53
12. https://plotly.com/python/pie-charts/