Crime Against Women in India
Analysis and Prediction
Miranda House, University of Delhi

Rhea Ajit John [1] , Nancy Dahiya [2] , Harshita Pahwa [3]

Abstract: The number of crimes committed against women has risen dramatically in recent years. Not only in India, but all throughout the world, it has become a huge social concern. Many attempts have been made to prevent similar atrocities, with strong actions taken in response. Every year, a large amount of data is generated from various types of crimes reported from throughout the world. This knowledge can help us better comprehend and detect violence, as well as aid us in some ways in resisting it. Analysing such statistics can be quite useful in finding crime patterns and occurrences. Data mining is crucial in this case because it allows us to evaluate, display, and anticipate the various crimes that occur in a certain area. In this paper, we have used Python libraries for the visualization of the dataset. Also, we have implemented Linear Regression model to predict the particular crime occurring in a particular state.

[1] B.Sc. Physical Science with Computer Science, 3rd year, Miranda House, University of Delhi.

[2] B.Sc. Physical Science with Computer Science, 3rd year, Miranda House, University of Delhi.

[3] B.Sc. Physical Science with Computer Science, 3rd year, Miranda House, University of Delhi.

Introduction

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.

Literature Survey

The researchers evaluated multiple images and predicted data using a variety of real-world applications with associated work. One of them is Crime Against Women: Analysis and Prediction [1] by Purvi Prasad, Amrita Nair and Dr. S. Godfrey Winster, in which they used Huber's regression to determine the loss score. i.e. - shows the difference between actual and predicted values for data analysis and Tim's algorithm for data visualization and time series where this is a time series multivariable data, i.e., data sets in which two or more variables are observed at a time, are transformed into a supervised training dataset. They have used visualization using the Pandas library, which gives us an easy-to-understand overview of the data collected on the NCRB site ⦃20012019⦄ and the scikit linear regression model. learn library for prediction. Linear regression
is a machine learning algorithm that uses a training set and a test set. Statistical analysis of crimes against women in India using data mining techniques. The analysis of crimes here in this article is done in two stages: grouping and classification
using the Weka tool. The input to the K means clustering algorithm is a numeric data set. A Comparative Study of Crimes Against Women Based on Machine Learning Using Big Data Techniques by Shivani Mishra and Suraj Kumar [3]. This project mainly focuses on machine learning in pattern recognition to analyse India's interstate models for crimes against women with this document using clustering and data cleansing. Violence Against Women: A State Level Analysis in India 2016 by Tanisha Khandelwal [4] 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.

Proposed Work

The data utilized to conduct the analysis is critical in identifying patterns, particularly in crime analysis. The dataset was gathered from **National Commission for Women** and comprises several forms of crimes perpetrated against women, such as ['Bigamy / Polygamy' 'Divorce' 'Dowry Death' 'Dowry Harassment / Cruelty to married women' 'Kidnapping / Abduction']. After gathering the data for the years 2001-2016, visualizations were carried out using python libraries like matplotlib and seaborn. After interpreting the visualizations, prediction for the total number of crimes committed in particular years i.e., 2016, 2017, 2018 and 2019 were programmed using Linear Regression Model. It's possible that the estimate 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.

Methods

1) The Dataset

The data set is accessible to the general public via the website http://ncwapps.nic.in/frmComp_stat_Overview.aspx. This data set contains a wealth of information, that are the location of the crime i.e., 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.

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

2) Cleaning The Dataset

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.

3) Visualizations

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.

4) Prediction

Forecasting for the total number of crimes occurring for a state in a particular year has been implemented using Linear Regression

Regression

Regression is a method for analysing and modelling the relationships between variables. To forecast a continuous value, regression models are used. One of the most prominent examples of regression is predicting the price of a house based on its parameters such as size, price, and so on. It's a method of supervised machine learning. There are various regression models available that can be used. Some of them are- Linear Regression, Ridge Regression, Lasso Regression, Huber Regression etc.

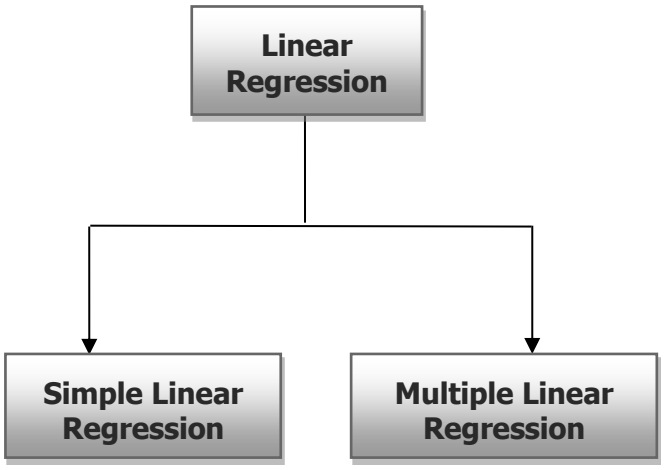
Linear Regression: The most fundamental and widely used type of predictive analysis is linear regression. Finding a linear relationship between a goal and one or more variables is done using linear regression. It can be used to forecast company sales, property prices, and other things. The goal of a linear regression model is to discover a link between the independent and dependent variables. The following are some of the most common applications of regression analysis:

(1) Identifying the strength of predictors
(2) anticipating an impact
(3) determining the strength of predictors.
(4) predicting trends

The following equation can be used to express the linear regression model:

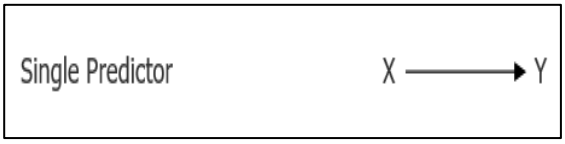
$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots\ldots + \theta_n x_n$$

- Y is the predicted value
- θ_o is the bias term.
- $\theta_1, \ldots, \theta_n$ are the model parameters
- $x_1, x_2, \ldots, x_n$ are the feature values

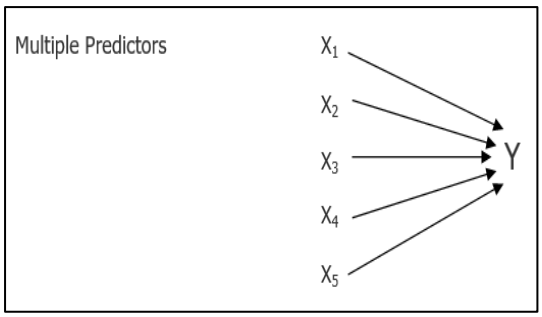


Types of Linear Regression:

Simple Linear Regression: There is only one x and one y variable in simple linear regression. For example, simple linear regression is used to forecast the price of a house based solely on square footage.



Multiple Linear Regression: There is one y variable and two or more x variables in multiple linear regression. For example, multiple linear regression is used to forecast the price of a property based on square footage and the age of the structure.



Terms associated with Regression

- **Predict:** When the observed values cluster near to the expected values, predictions are precise. The dependent variable's mean is used to make regression predictions.
- **Intercept:** When all X=0, the intercept (also known as the constant) is the predicted mean value of Y.
- **Coefficient:** The size of each independent variable's coefficient indicates the magnitude of the effect that variable has on your dependent variable, and the sign of the coefficient (positive or negative) indicates the effect's direction.
- **Score:** The score function is used to assess the fit quality of models or patterns.

To assess the *algorithm's performance*, the following are calculated:

- **Mean Absolute Error:** Mean Absolute Error (MAE) is the mean of the absolute value of the errors
- **Mean Squared Error:** Mean Squared Error (MSE) is the mean of the squared errors
- **Root mean squared Error:** Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors

Python and Libraries Used

- **Python**

Python is a high-level programming language that is interpreted, object-oriented, and supports dynamic data. Python has a dynamic type system and memory management that is automated. It features a big and extensive standard library and supports several programming paradigms, including object-oriented, imperative, functional, and procedural. The programming language is simple and straightforward, with a variety of strong classes.

- **Numpy**

It is a library for processing arrays that can be used for any purpose. It includes a high-performance multidimensional array object as well as utilities for manipulating them. It

is the most important Python package for scientific computing.

- **Pandas**

Pandas is a data manipulation and analysis software package for the Python programming language. It includes data structures and methods for manipulating numerical tables and time series, in particular.

- **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.

- **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.

- **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.

- **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).

- **SciKit-learn**

SciKit–learn is a Python machine learning package that is free and open source. It includes a number of classifications, regression, and clustering algorithms and is designed to work with the NumPy and SciPy Python numerical libraries. In this paper SciKit-Learn is used for the implementation of Linear Regression Model.

Implementation Using Python Programming

1. Snippet of code for Visualization

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots


data = pd.read_csv('/combined_dataset.xlsx - Sheet1.csv')
df = data.copy()

df.info()
df.shape  #shape of Dataset
crimes=['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"]


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()
```

2. Snippet of code for Prediction

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
print("Libraries Imported")

data = pd.read_csv("/content/combined_dataset.xlsx - Sheet1.csv")
df = data.copy()
```

Creating the improvised dataset for model implementation

(The following snippet only shows for the 2016 prediction. For years 2017, 2018, 2019 can be implemented using the same method)

```
df.drop('S.No',axis=1,inplace=True)

df_2001=df[(df.Year == 2001)]
df_2001['Total Crime 1']=df_2001.iloc[:, 2:].sum(axis=1)

df_2002=df[(df.Year == 2002)]
df_2002['Total Crime 2']=df_2002.iloc[:, 2:].sum(axis=1)

df_2003=df[(df.Year == 2003)]
df_2003['Total Crime 3']=df_2003.iloc[:, 2:].sum(axis=1)

df_2004=df[(df.Year == 2004)]
df_2004['Total Crime 4']=df_2004.iloc[:, 2:].sum(axis=1)

df_2005=df[(df.Year == 2005)]
df_2005['Total Crime 5']=df_2005.iloc[:, 2:].sum(axis=1)

df_2006=df[(df.Year == 2006)]
df_2006['Total Crime 6']=df_2006.iloc[:, 2:].sum(axis=1)

df_2007=df[(df.Year == 2007)]
df_2007['Total Crime 7']=df_2007.iloc[:, 2:].sum(axis=1)

df_2008=df[(df.Year == 2008)]
df_2008['Total Crime 8']=df_2008.iloc[:, 2:].sum(axis=1)

df_2009=df[(df.Year == 2009)]
df_2009['Total Crime 9']=df_2009.iloc[:, 2:].sum(axis=1)

df_2010=df[(df.Year == 2010)]
df_2010['Total Crime 10']=df_2010.iloc[:, 2:].sum(axis=1)

df_2011=df[(df.Year == 2011)]
df_2011['Total Crime 11']=df_2011.iloc[:, 2:].sum(axis=1)

df_2012=df[(df.Year == 2012)]
df_2012['Total Crime 12']=df_2012.iloc[:, 2:].sum(axis=1)

df_2013=df[(df.Year == 2013)]
df_2013['Total Crime 13']=df_2013.iloc[:, 2:].sum(axis=1)

df_2014=df[(df.Year == 2014)]
df_2014['Total Crime 14']=df_2014.iloc[:, 2:].sum(axis=1)

df_2015=df[(df.Year == 2015)]
df_2015['Total Crime 15']=df_2015.iloc[:, 2:].sum(axis=1)

df_2016=df[(df.Year == 2016)]
df_2016['Total Crime 16']=df_2016.iloc[:, 2:].sum(axis=1)
```

```
"""# State-Year Wise"""

df_state=pd.DataFrame()

df_state['State']=['AP', 'AR', 'AS', 'BR', 'CG', 'GA', 'GJ', 'HR', 'HP', 'J&K', 'JH', 'KR', 'KE',
                   'MP', 'MH', 'MN', 'MG', 'MZ', 'NL', 'OR', 'PB', 'RJ', 'SK', 'TN', 'TL', 'TR',
                   'UP', 'UK', 'WB', 'A&N', 'CH', 'D&N', 'D&D', 'LK', 'DL', 'PC']

l1=df_2001['Total Crime 1'].tolist()
df_state['2001']=l1

df_state['2002']=df_2002['Total Crime 2']
l2=df_2002['Total Crime 2'].tolist()
df_state['2002']=l2

df_state['2003']=df_2003['Total Crime 3']
l3=df_2003['Total Crime 3'].tolist()
df_state['2003']=l3

df_state['2004']=df_2004['Total Crime 4']
l4=df_2004['Total Crime 4'].tolist()
df_state['2004']=l4

df_state['2005']=df_2005['Total Crime 5']
l5=df_2005['Total Crime 5'].tolist()
df_state['2005']=l5

df_state['2006']=df_2006['Total Crime 6']
l6=df_2006['Total Crime 6'].tolist()
df_state['2006']=l6

df_state['2007']=df_2007['Total Crime 7']
l7=df_2007['Total Crime 7'].tolist()
df_state['2007']=l7

df_state['2008']=df_2008['Total Crime 8']
l8=df_2008['Total Crime 8'].tolist()
df_state['2008']=l8

df_state['2009']=df_2009['Total Crime 9']
l9=df_2009['Total Crime 9'].tolist()
df_state['2009']=l9

df_state['2010']=df_2010['Total Crime 10']
l10=df_2010['Total Crime 10'].tolist()
df_state['2010']=l10

df_state['2011']=df_2011['Total Crime 11']
l11=df_2011['Total Crime 11'].tolist()
df_state['2011']=l11

df_state['2012']=df_2012['Total Crime 12']
l12=df_2012['Total Crime 12'].tolist()
df_state['2012']=l12

df_state['2013']=df_2013['Total Crime 13']
l13=df_2013['Total Crime 13'].tolist()
df_state['2013']=l13

df_state['2014']=df_2014['Total Crime 14']
l14=df_2014['Total Crime 14'].tolist()
df_state['2014']=l14

df_state['2015']=df_2015['Total Crime 15']
l15=df_2015['Total Crime 15'].tolist()
df_state['2015']=l15

df_state['2016']=df_2016['Total Crime 16']
l16=df_2016['Total Crime 16'].tolist()
df_state['2016']=l16
```

```
X=df_state.iloc[:,2:-1]
y=df_state.iloc[:,df_state.shape[1]-1]


X_train , X_test , y_train , y_test     = train_test_split(X, y,  test_size=0.1, random_state=1 )

linear_model = LinearRegression(normalize=True)
linear_model.fit(X_train, y_train)

y_pred = linear_model.predict(X_test)       # predicted values for the model

df1 = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
print(df1)

print(linear_model.intercept_)
print(linear_model.coef_)


linear_model.score(X_test,y_test)

import math
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test,y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test,y_pred))
print('Root Mean Squared Error:', math.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

Outputs

- ## Visualizations

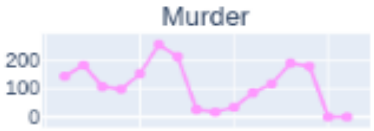
1. Scatter Plot for Each Type of Crime

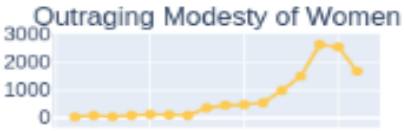

Harassment At Workplace
4000
3000
2000
1000
0


Kidnapping / Abduction
800
600
400
200
0


Maintenance Claim
150
100
50
0


Miscellaneous
10k
5k
0


Murder
200
100
0


Outraging Modesty of Women
3000
2000
1000
0

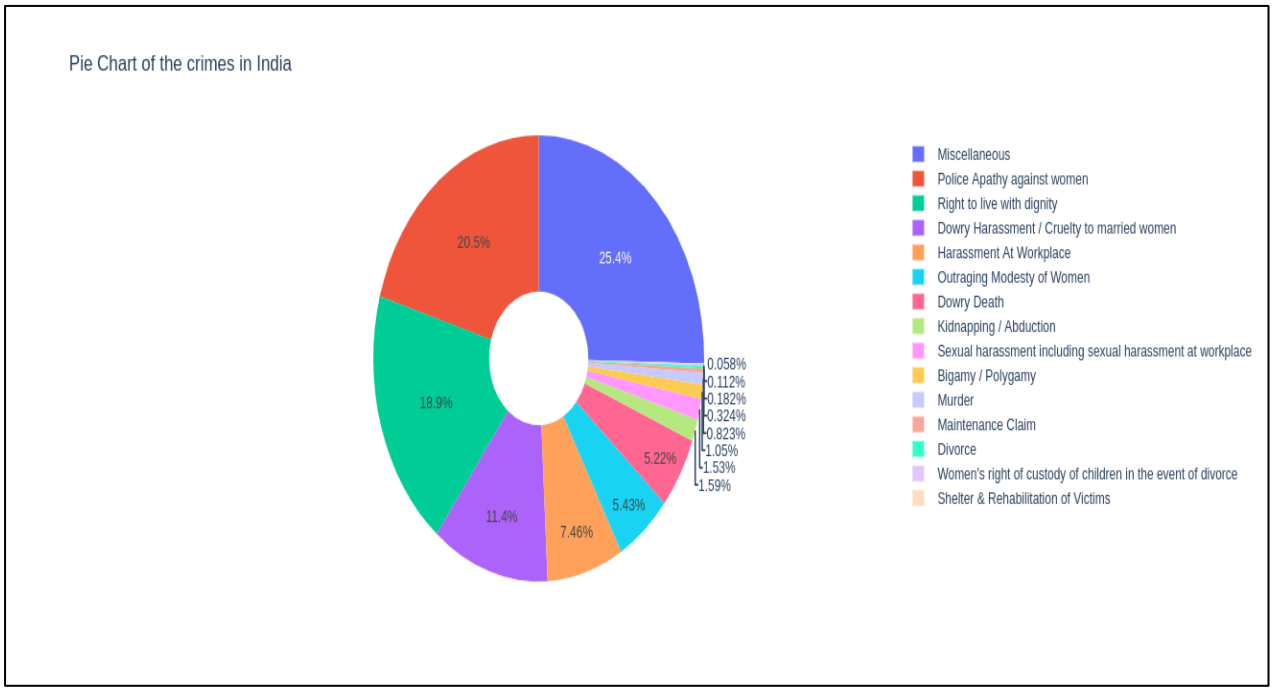

Right to live with dignity
6000
4000
2000
0


Sexual harassment including sexual harassment at w
600
400
200
0

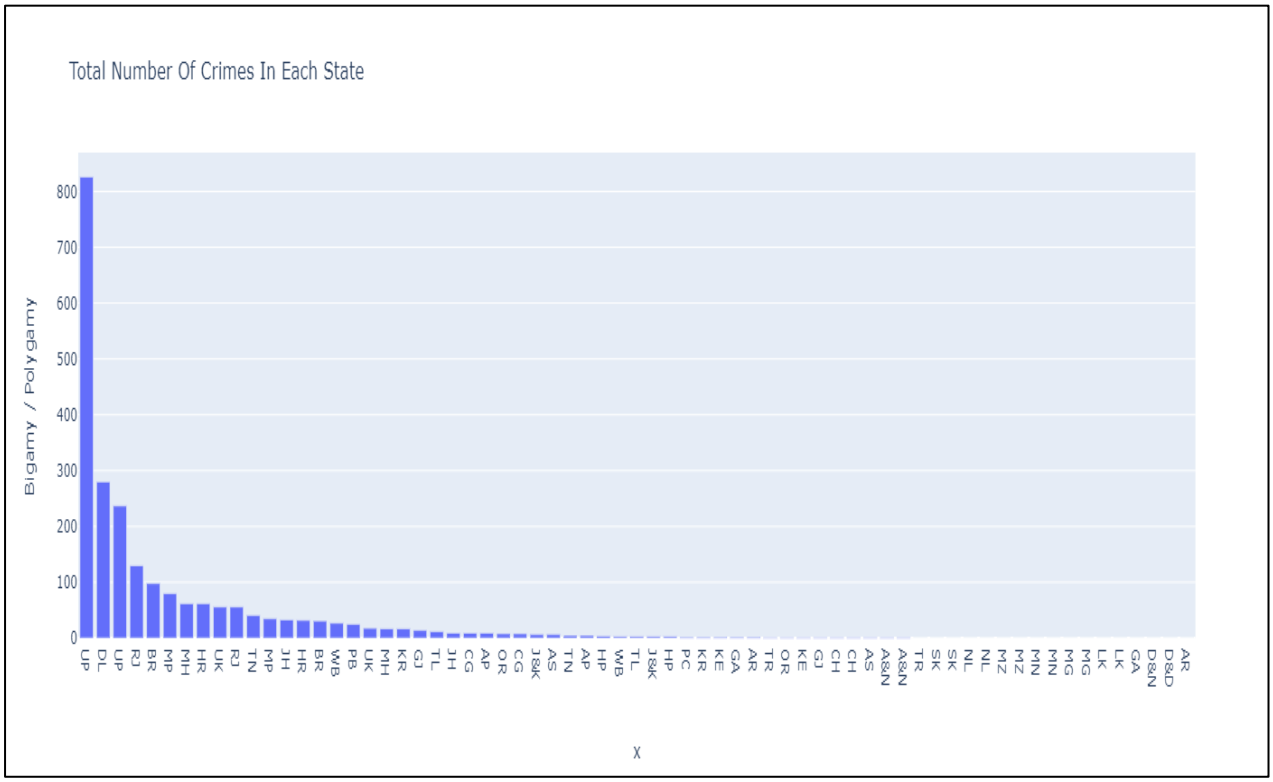

Shelter & Rehabilitation of Victims
30
20
10
0

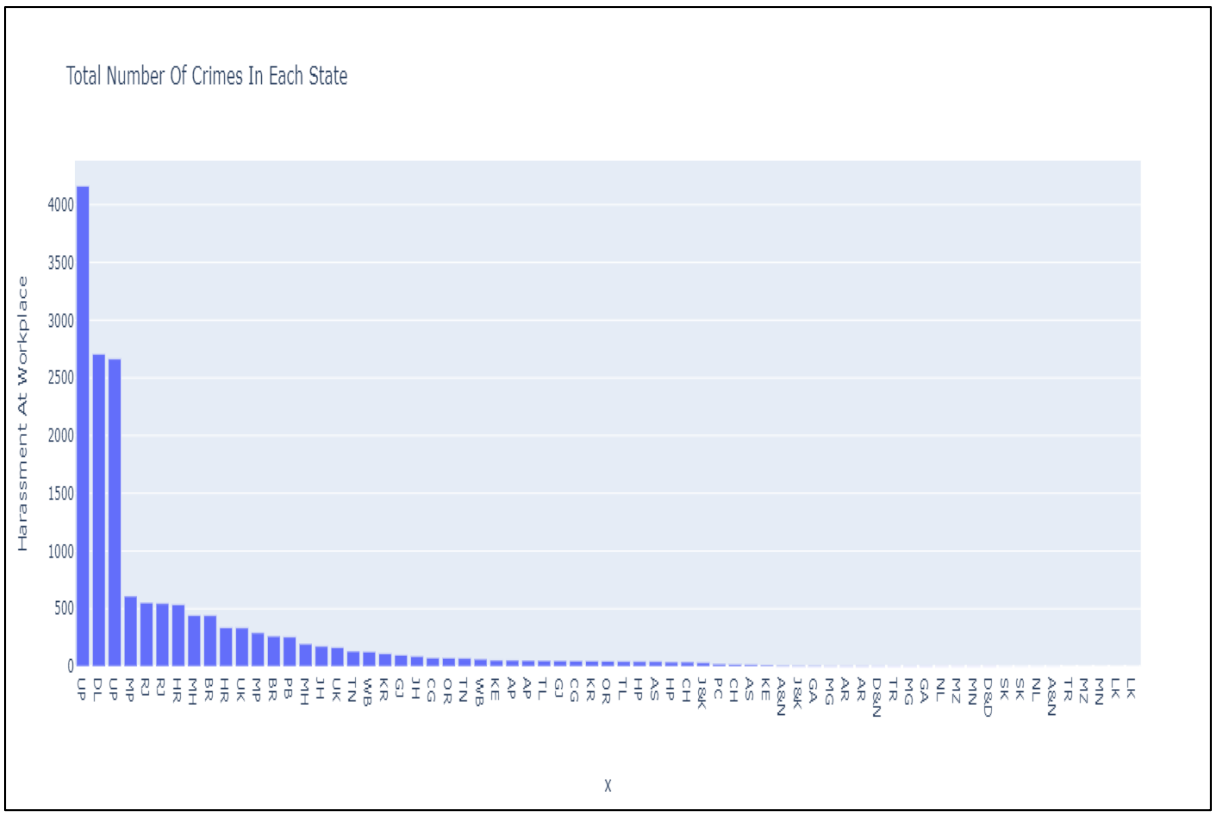

Women's right of custody of children in the event of
40
20
0

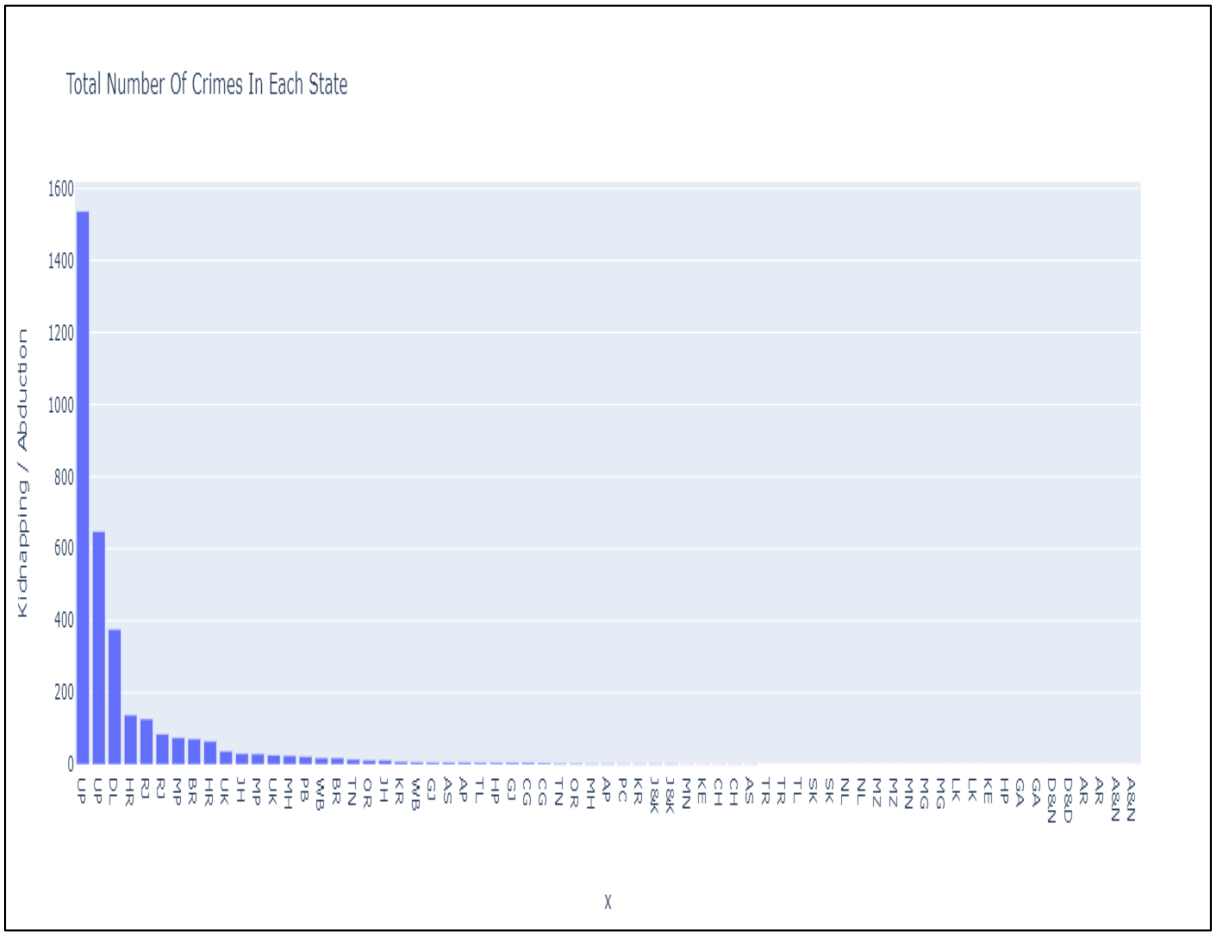
2. Pie Chart of the crimes in India

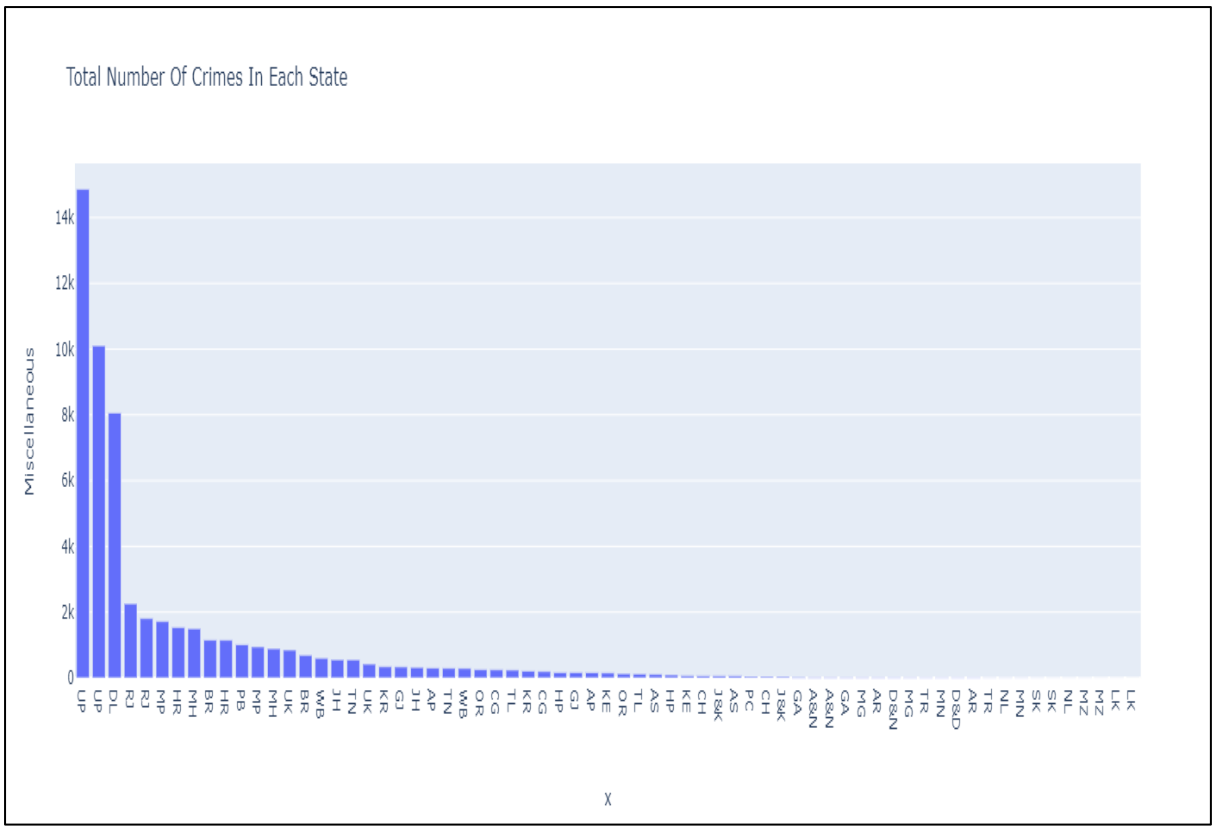


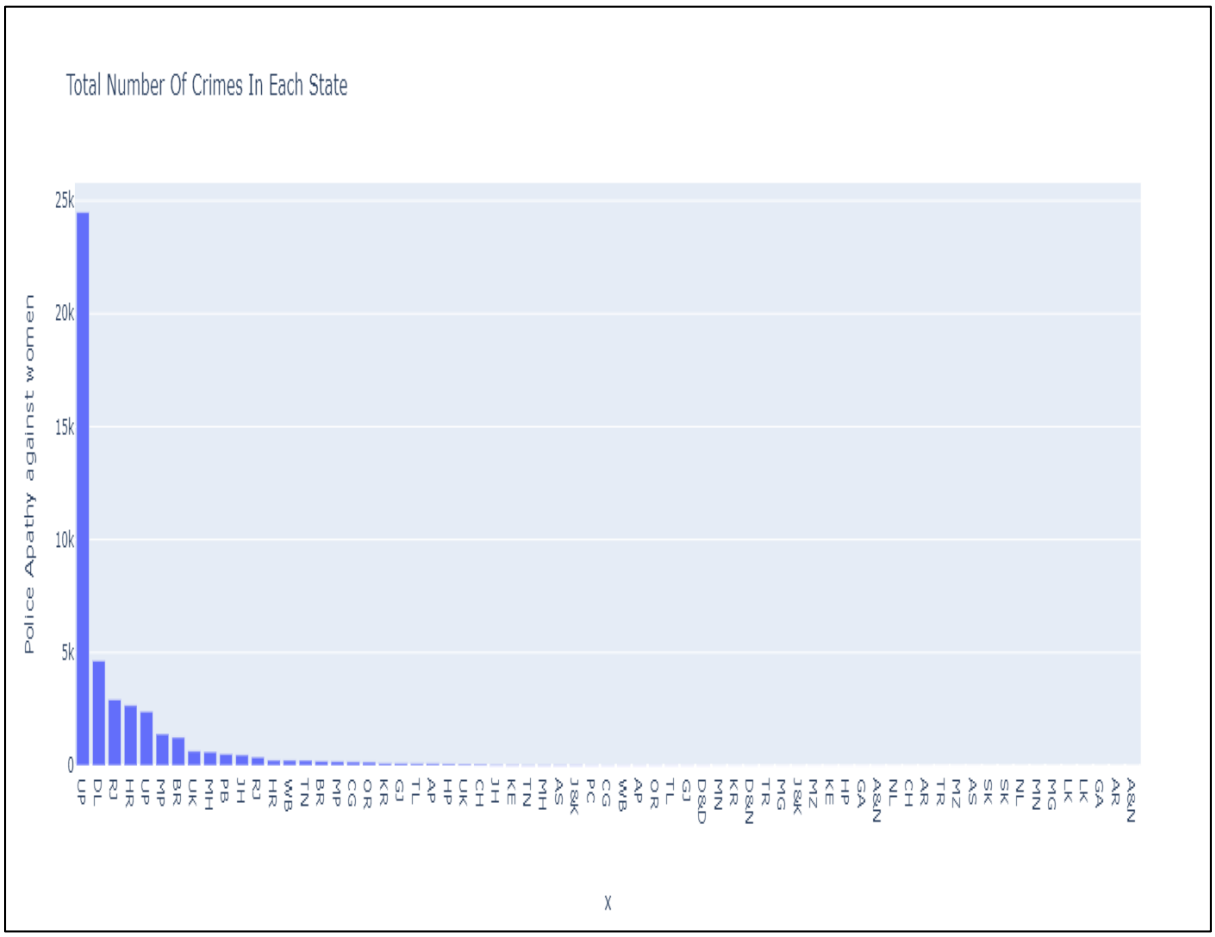
3. State-Wise

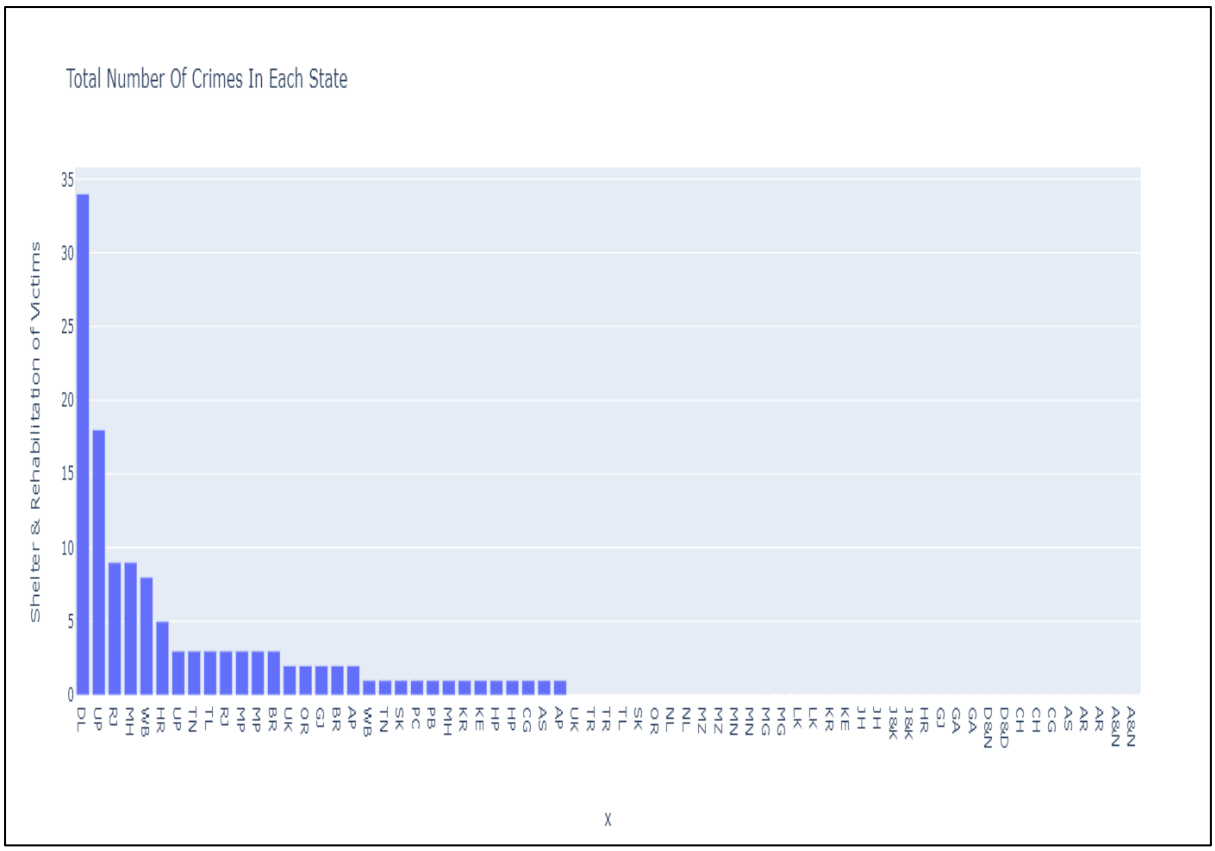


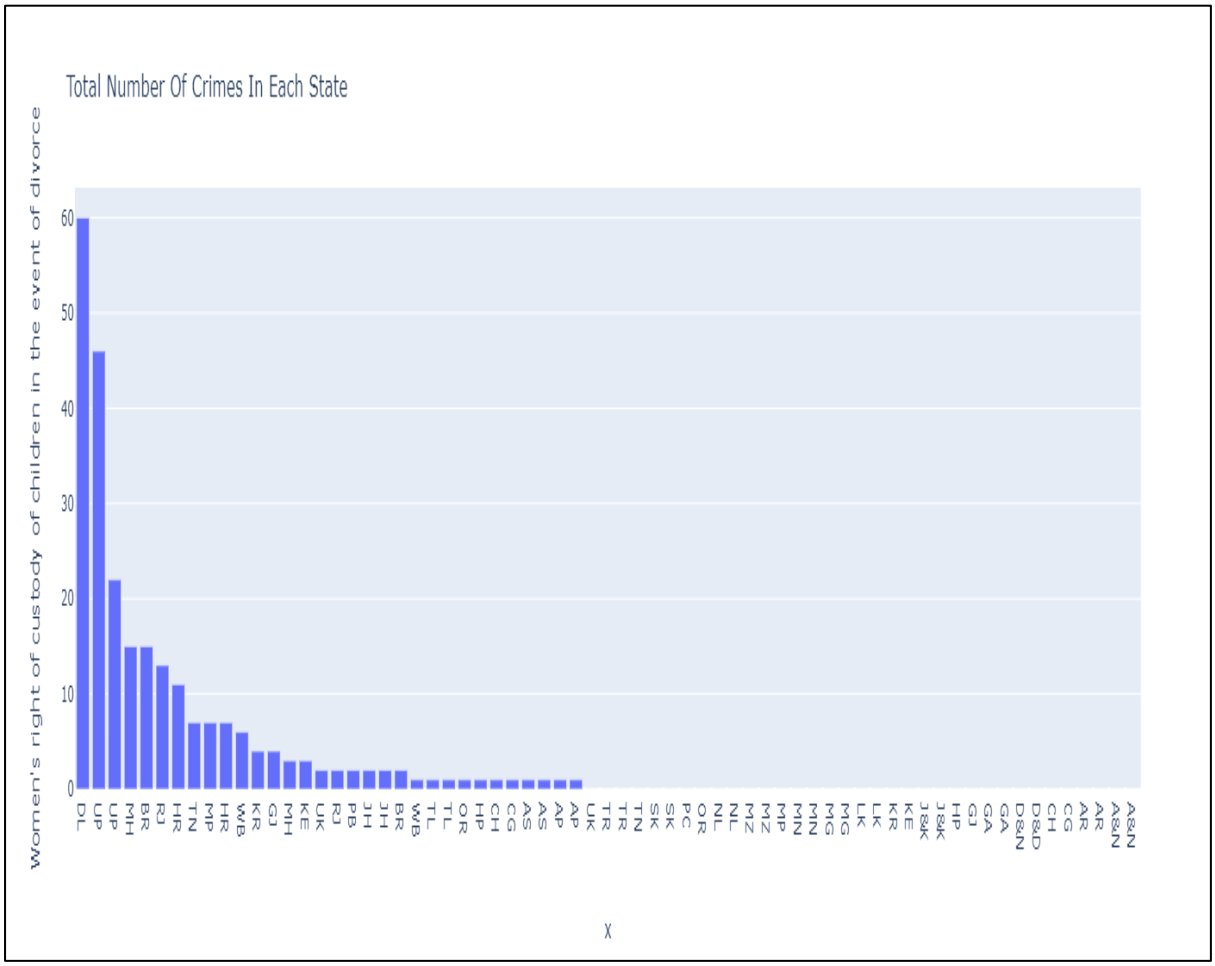

Total Number Of Crimes In Each State
Harassment At Workplace
4000
3500
3000
2500
2000
1500
1000
500
0
UP DL UP MP RJ RJ HR MH BR HR UK MP BR PB MH JH UK TN WB KR GJ JH CG OR TN WB KE AP AP TL GJ CG KR OR TL HP AS HP CH J&K PC CH AS KE A&N J&K GA MG AR AR D&N TR MG GA NL MZ MN MN D&D SK SK NL A&N TR MZ MN LK LK LK
X

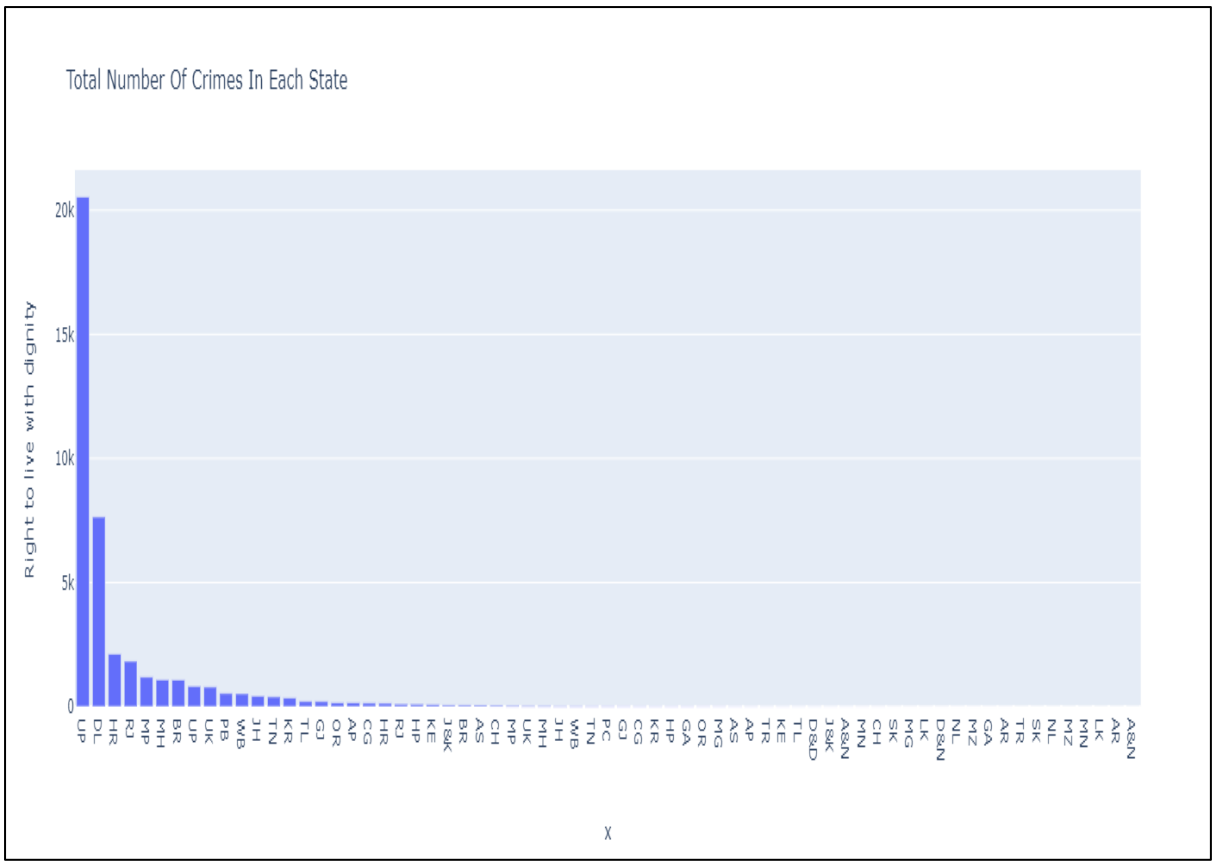

Total Number Of Crimes In Each State
Kidnapping / Abduction
1600
1400
1200
1000
800
600
400
200
0
UP UP DL HR RJ RJ MP BR HR UK JH MP UK MH PB WB BR TN OR JH KR WB GJ AS AP TL HP GJ CG CG TN OR MH AP PC KR J&K J&K MN KE CH CH AS TR TR TL SK SK NL NL MZ MZ MN MG MG LK LK KE HP GA GA D&N D&D AR AR A&N A&N
X

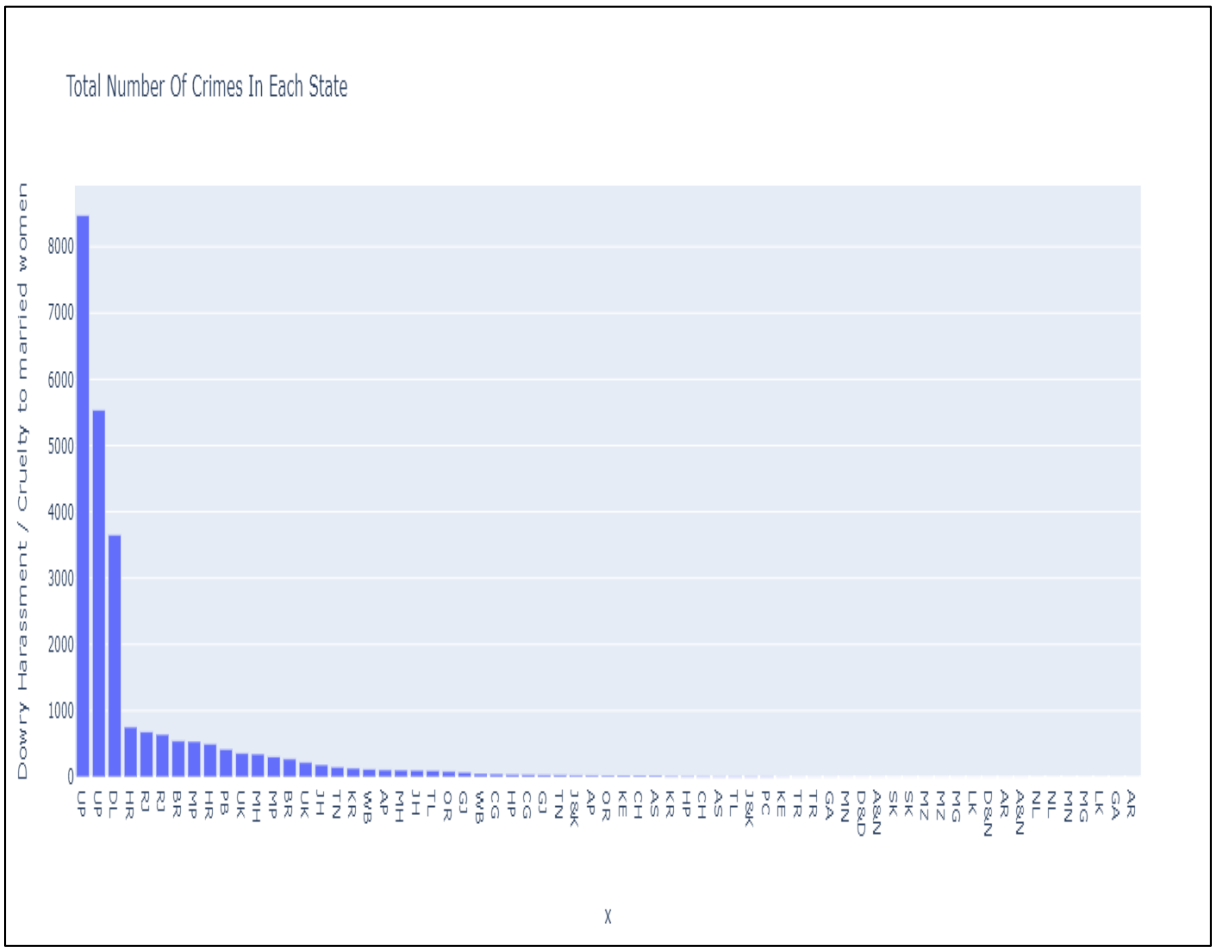

Total Number Of Crimes In Each State
Miscellaneous
14k
12k
10k
8k
6k
4k
2k
0
UP UP DL RJ RJ MP HR MH BR HR PB MP MH UK BR WB JH TN UK KR GJ JH AP TN WB OR CG TL KR CG HP GJ AP KE OR TL AS HP KE CH JSK AS PC CH JSK GA A&N A&N GA MG AR D&N MG TR MN D&D AR TR NL MN SK SK NL MZ MZ LK LK
X

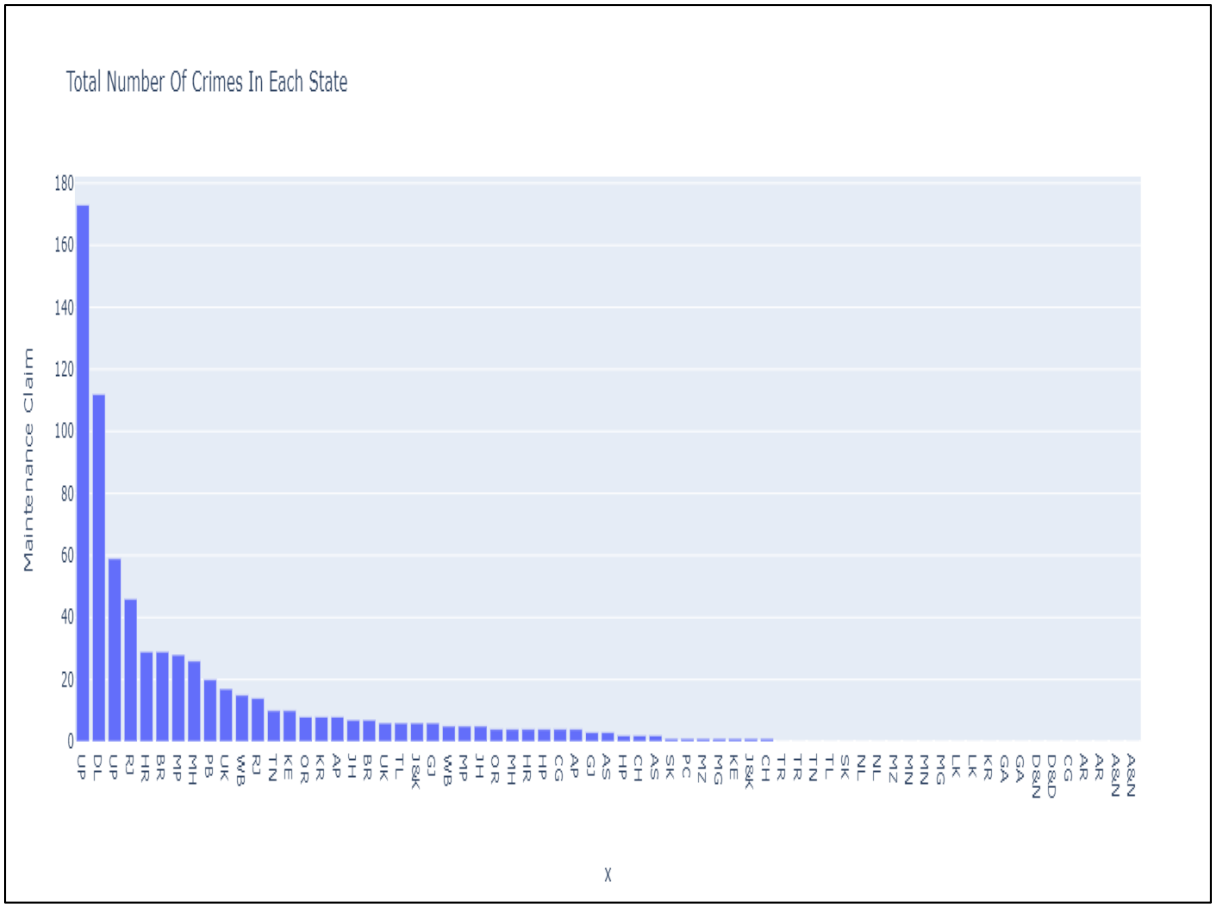

Total Number Of Crimes In Each State
Police Apathy against women
25k
20k
15k
10k
5k
0
UP DL RJ HR UP MP BR UK MH PB JH RJ HR WB TN BR MP CG OR KR GJ TL AP HP UK CH JH KE TN MH AS JSK PC CG WB AP OR TL GJ D&D MN KR D&N TR MG JSK MZ KE HP GA A&N NL CH AR TR MZ AS SK SK NL MN MG LK LK GA AR A&N
X

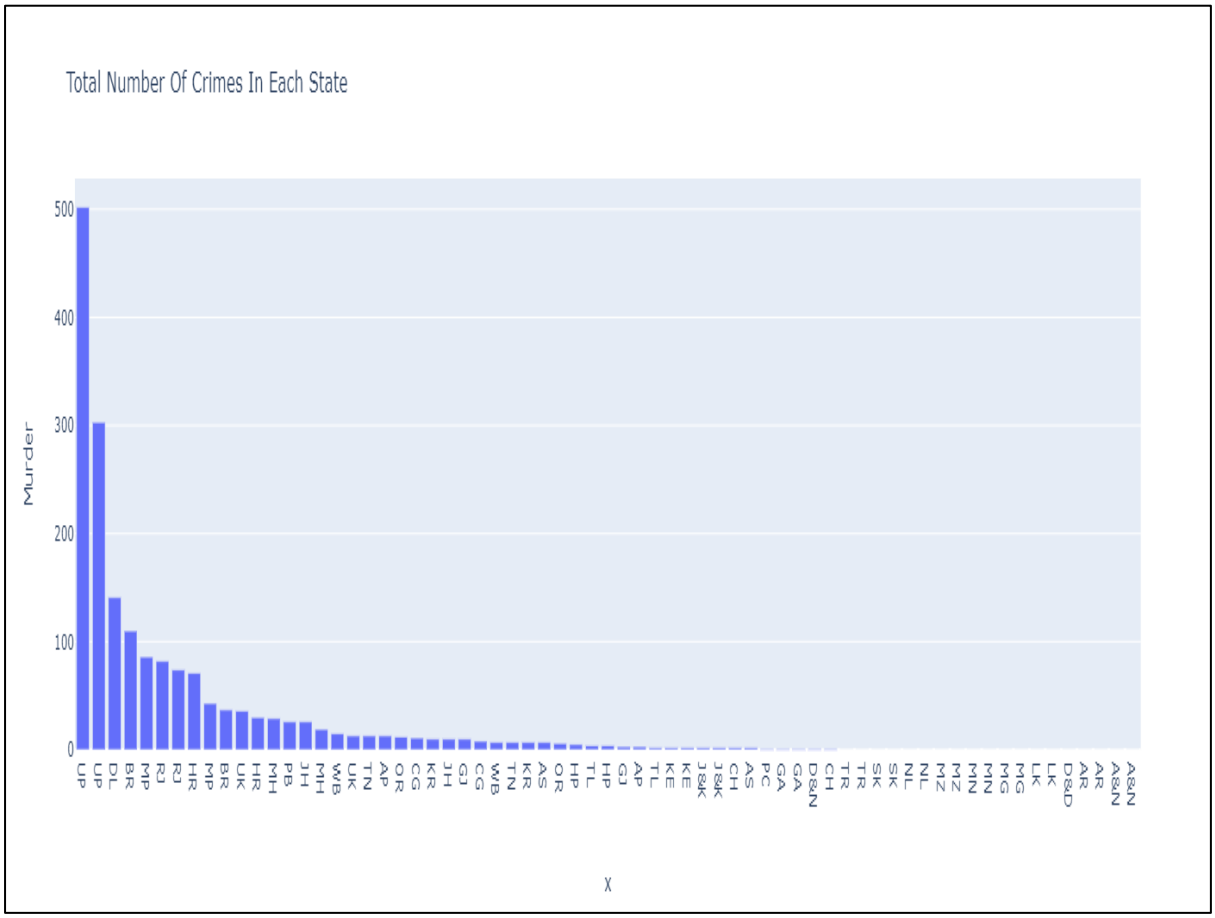

Total Number Of Crimes In Each State
Shelter & Rehabilitation of Victims
0
5
10
15
20
25
30
35
DL UP RJ MH WB HR UP TN TL RJ MP MP BR UK OR GJ BR AP WB TN SK PC PB MH KR KE HP HP CG AS AP UK TR TR TL SK OR NL NL MZ MN MN MG MG LK LK KR KE JH J&K J&K HR GJ GA GA D&N D&D D&D CH CH CG CG AS AR AR A&N A&N
X

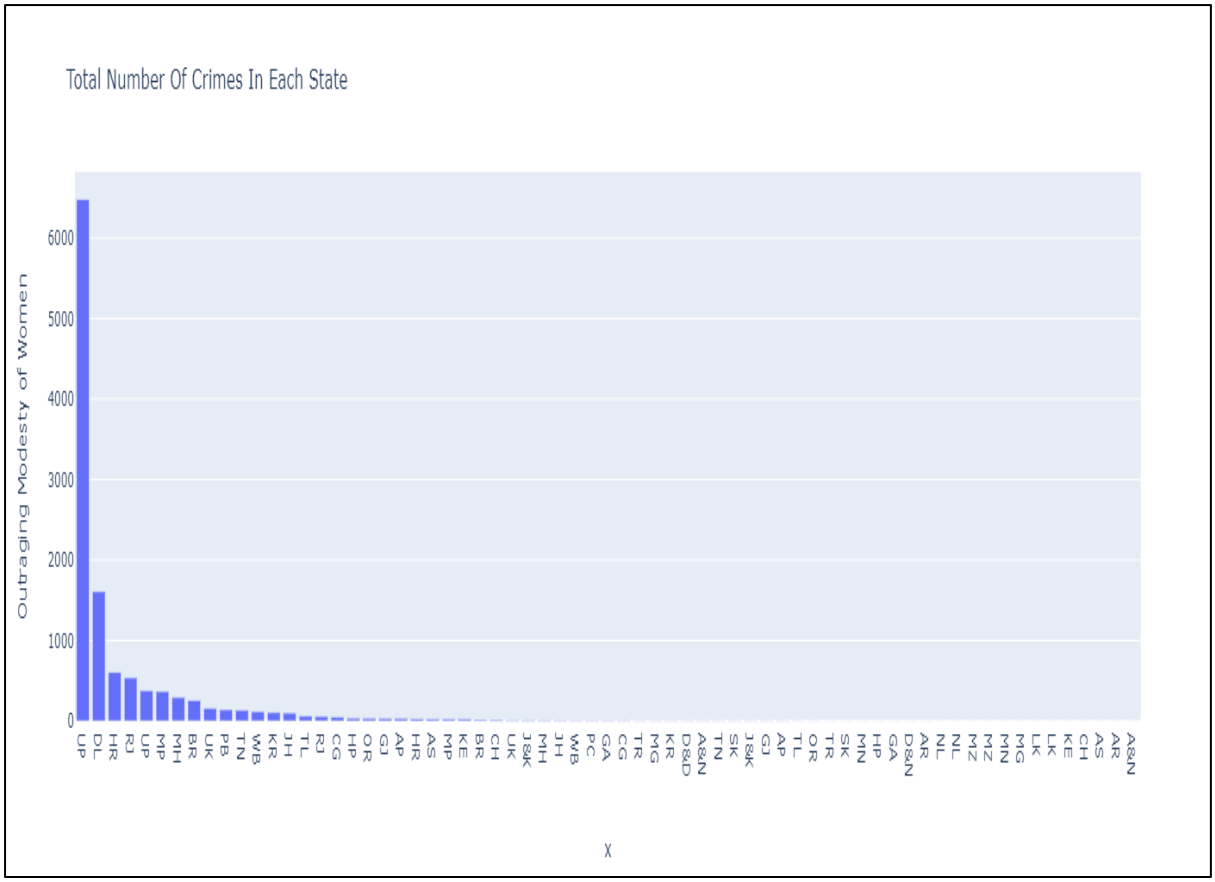

Total Number Of Crimes In Each State
Women's right of custody of children in the event of divorce
0
10
20
30
40
50
60
DL UP UP MH BR RJ HR TN MP HR WB KR GJ MH KE UK RJ PB JH JH BR WB TL TL OR HP CH CG AS AP AP UK TR TR TN SK SK PC OR NL NL MZ MP MN MN MG MG LK LK KR KE J&K J&K HP GJ GA GA D&N D&D D&D CH CH CG CG AR AR A&N A&N
X

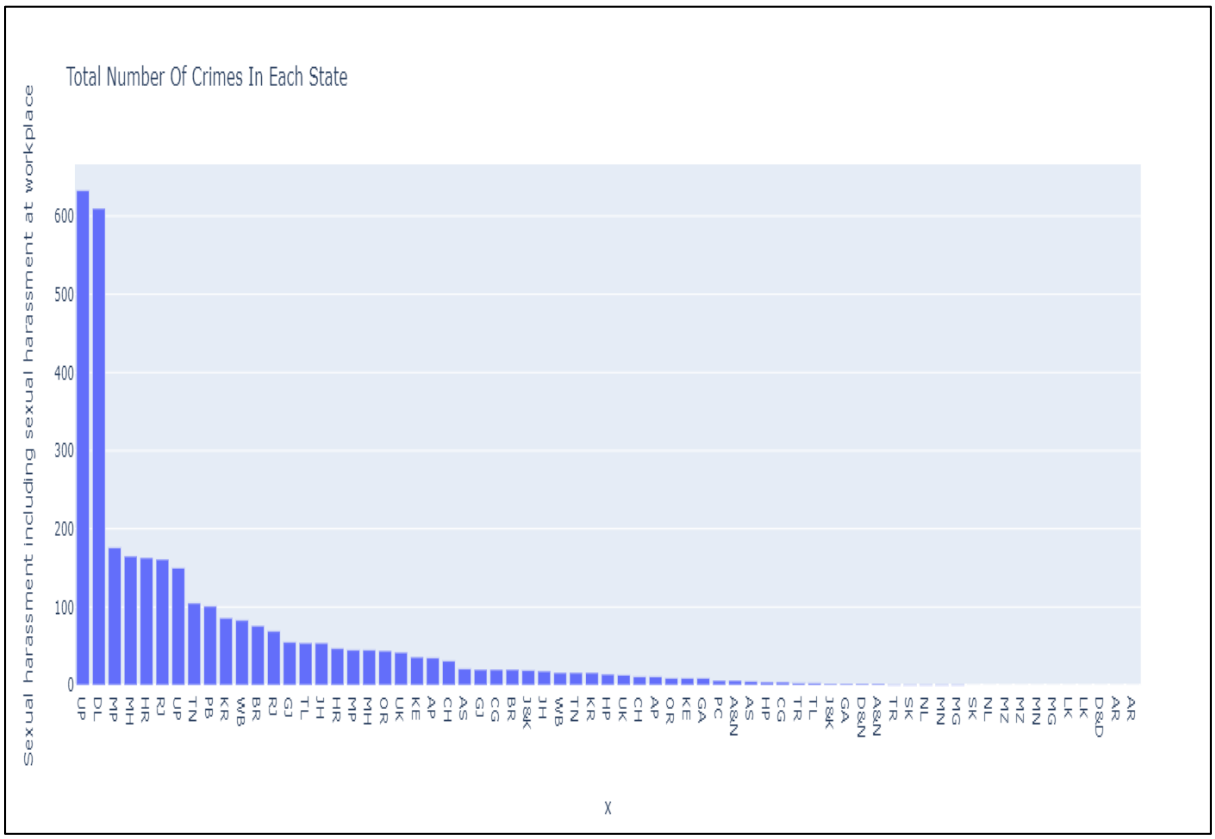

Total Number Of Crimes In Each State
Right to live with dignity
20k
15k
10k
5k
0
UP DL HR RJ MP MH BR UP UK PB WB JH TN KR TL GJ OR AP CG HR RJ HP KE J&K BR AS CH MP UK MH JH WB TN PC GJ CG KR HP GA OR MG AS AP TR KE TL D&D J&K A&N MN CH SK MG LK D&N NL MZ GA AR TR SK NL MZ MN LK AR A&N
X

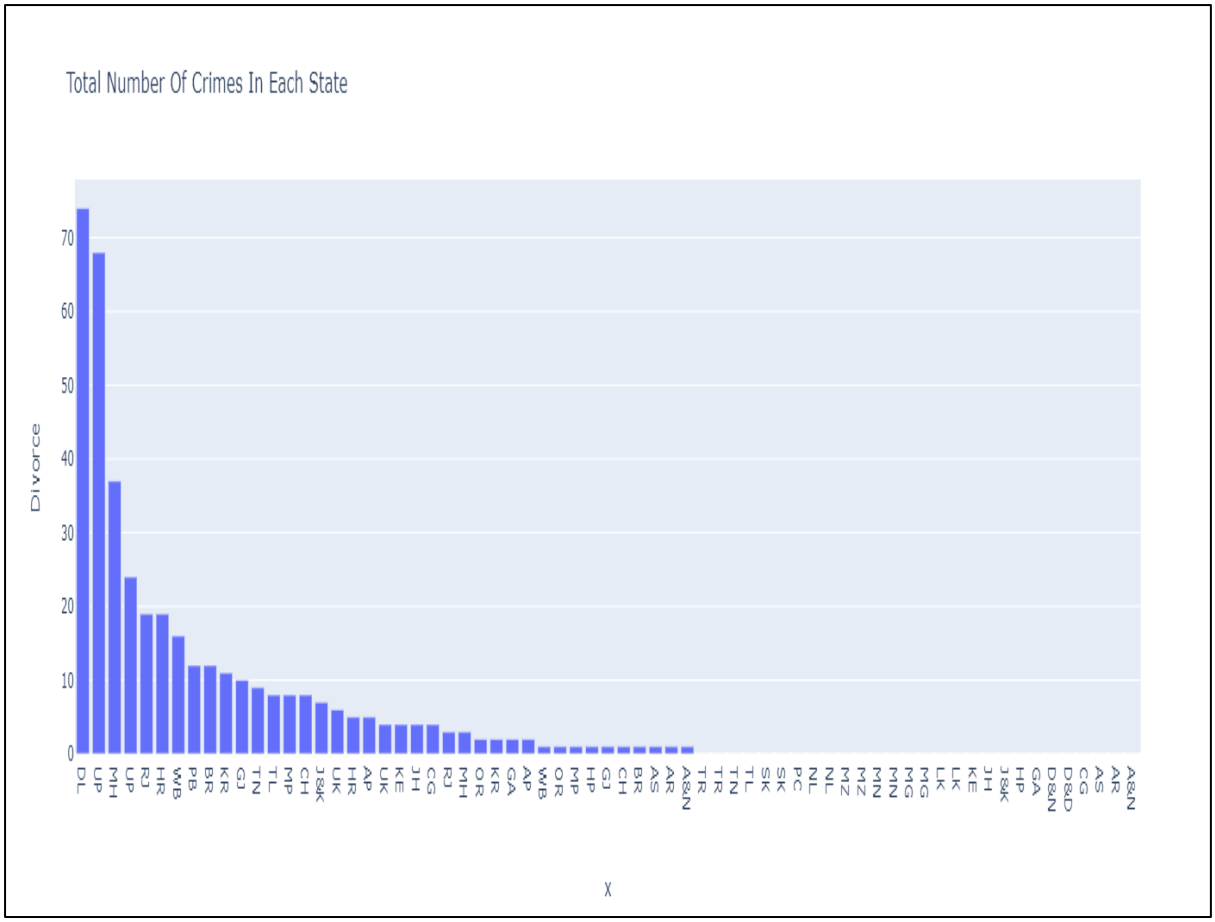

Total Number Of Crimes In Each State
Dowry Harassment / Cruelty to married women
8000
7000
6000
5000
4000
3000
2000
1000
0
UP UP DL HR RJ RJ BR MP HR PB UK MH MP BR UK JH TN KR WB AP MH JH TL OR GJ WB CG HP CG GJ TN J&K AP OR KE CH AS KR HP CH AS TL J&K PC KE TR TR GA MN D&D A&N SK SK MZ MZ MG LK D&N AR A&N NL NL MN MG LK GA AR
X

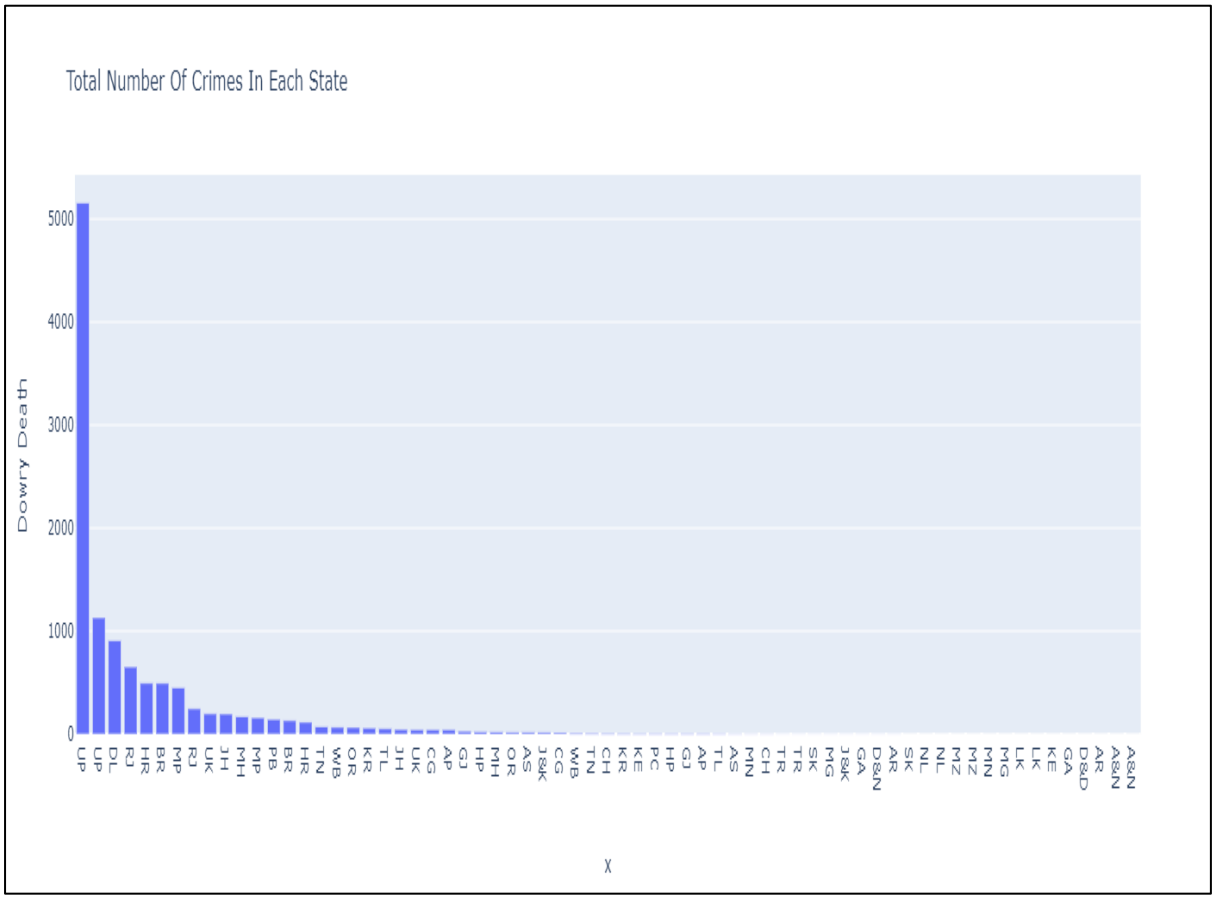

Total Number Of Crimes In Each State
Maintenance Claim
180
160
140
120
100
80
60
40
20
0
UP DL UP RJ HR BR MP MH PB UK WB RJ TN OR KE KR AP JH BR UK TL GJ J&K WB MP JH OR MH HR HP CG AP GJ AS CH SK AS PC MZ MG KE J&K CH TR TR TN TL SK NL NL MZ MN MN MG LK LK KR GA GA D&N D&D CG AR AR A&N A&N
X


Total Number Of Crimes In Each State
Murder
500
400
300
200
100
0
UP UP DL BR MP RJ HR MP BR UK HR MH PB JH MH WB UK TN AP OR CG KR JH GJ CG WB TN KR AS OR HP TL GJ AP TL KE KE J&K CH AS PC GA GA D&N CH TR TR SK SK NL NL MZ MZ MN MN MG MG LK LK D&D AR AR A&N A&N
X


Total Number Of Crimes In Each State
Outraging Modesty of Women
0
1000
2000
3000
4000
5000
6000
UP DL HR RJ UP MP MH BR UK PB TN WB KR JH TL RJ CG HP OR GJ AP HR AS MP KE BR CH UK J&K MH JH WB PC GA CG TR MG KR D&D A&N TN SK J&K GJ AP TL OR TR SK MN HP GA D&N AR NL NL MZ MN MG LK LK KE CH AS AR A&N
X


Total Number Of Crimes In Each State
Sexual harassment including sexual harassment at workplace
0
100
200
300
400
500
600
UP DL MP MH HR RJ UP TN PB KR WB BR RJ GJ TL JH HR MP MH OR UK KE AP CH AS GJ CG BR J&K JH WB TN KR HP UK CH AP OR KE GA PC A&N AS HP CG TR TL J&K GA D&N A&N TR SK NL MN MG SK NL MZ MN MG LK LK D&D AR AR
X


Total Number Of Crimes In Each State
Divorce
0
10
20
30
40
50
60
70
DL UP MH UP RJ HR WB PB BR KR GJ TN TL MP CH J&K UK HR AP UK KE JH CG RJ MH OR KR GA AP WB OR MP HP GJ CH BR AS AR A&N TR TN TL SK SK PC NL NL MZ MZ MN MN MG MG LK LK KE JH J&K HP GA D&N D&D D&D CG AS AR AR A&N
X


Total Number Of Crimes In Each State
Dowry Death
0
1000
2000
3000
4000
5000
UP UP DL RJ HR BR MP RJ UK JH MH MP PB BR HR TN WB OR KR TL JH UK CG AP GJ HP MH OR AS J&K CG WB TN CH KR KE PC HP GJ AP TL AS MN CH TR TR SK MG J&K GA D&N AR SK NL NL MZ MZ MN MN MG LK LK KE GA D&D AR A&N A&N
X

- **Predictions**

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
```

Scope and Challenges

With the passage of time, the safety of females has become a major worry. According to NCRB, there has been a 7.3 percent increase in crime against women from 2018 to 2019, and COVID-19 has further worsened the situation. In various parts of India, a large number of records are generated each year related to such unique crimes. Analysing such vast amounts of data statistics may appear.to be a time-consuming task. For the years 2016, 2017, 2018, and 2019, we utilised linear regression to forecast the total number of crimes that occurred in each state and union territory. There are errors in the version because the records have become in bulk and they may not be accurate. However, several models such as Logistic Regression, CART, KNN, Huber Regression and others were used to achieve the better predicted values. COVID-19 made it difficult to document information for each section of the country, hence this version has a destination scope.

Conclusion

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 against women. Various regression techniques can be implemented to predict the total number of crimes occurring in each state for a particular year. This paper 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. 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 was implemented. Forecasting was achieved by the Linear Regression Model. 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.

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." 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

https://www.kaggle.com/marcogherbezza/crimes-against-women

https://easychair.org/publications/preprint/pD8r

https://www.geeksforgeeks.org/

https://ssi.edu.in/wp-content/uploads/2019/05/Internship-Report-by-Ms.-Tanisha-Khandelwal.pdf

https://towardsdatascience.com/classification-regression-and-prediction-whats-the-difference-5423d9efe4ec

https://www.researchgate.net/publication/336982992_Crime_against_Women_CAW_Analysis_and_Prediction_in_Tamilnadu_Police_Using_Data_Mining_Techniques

https://www.ijert.org/research/crime-against-women-analysis-and-prediction-IJERTV10IS050229.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html