Crimes in India Project Report

Submitted by: AYAZ WARIS KHAN

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my Instructor Mr. Shankargouda Tegginmani as well as Data Trained Academy who gave me the opportunity to do this project on CRIMES IN INDIA, which also helped me in doing lots of research wherein I came to know about so many new things especially the data collection part.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1) https://www.prb.org/wp-content/uploads/2011/04/india-population-2001-2011.pdf
2) https://censusindia.gov.in/nada/index.php/catalog/20030

3) https://scikit-learn.org/stable/user_guide.html

4) https://github.com/

5) https://www.kaggle.com/

6) https://medium.com/

7) https://towardsdatascience.com/

8) https://www.analyticsvidhya.com/

9) https://www.geeksforgeeks.org/

10)https://www.latestlaws.com/wp-content/uploads/2018/07/National-Crime-Record-Bureau-Report-NCRB-2012.pdf

INTRODUCTION

- ## Project Problem Framing-

Crime – a term which is just like a havoc in today's world. It is a disastrous act for entire humanity and an obstacle in the way of development. The legal definition of crime introduces us with a vast number of hardships and complexities as it is a social construction that we consider a crime.. It is disputed and contingent dynamically. In other words, crime differs over time and location. It is not a universally accepted factor and it's socially built and altered reality. Crime is just like a toxic which spoils the growth of a nation. Simply, a crime can be defined as a criminal offense against any person or an organization with an intent to harm them directly or indirectly that is illegal and punishable under the country law. Crimes like robberies, looting, sexual harassment, rape, abduction, and killings are one of the major crimes which are happening at a breakneck speed starting from rural to urban areas. As these crimes are lifting high and high so there is a need to control them and thus creating huge pressure on the investigation department. There should be a system which can analyze crime and police department can make use of this technology that can make their task easier to investigate the case on the basis of different trends for years. Crime analysis is a law enforcement technique that involves systematic analysis for trends and patterns identification and analysis.

- **Review of Literature**

  This project is more about exploration Clustering that can be done on this data. Since we scrape huge amount of data that includes more crimes related features, we can do better data exploration and derive some interesting features using the available columns.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  In our dataset there are States/UT, Districts, Year and different heads f clrime total 37 columns I got in this dataset(phase 4) , Since there is no Target variable so we handle this Problem as Unsupervised Machine Learning and will use Clustering Algorithm

  This project is done in two parts:

  - Data Collection phase
  - Clustering phase

Phase 1

- You can collect data from anywhere (wiki, google, etc) but mention the link from where data is being collected.4 The population of each state.
- Literacy Rate in each state
- Area of each state
- Collect any other data that helps with your analysis. There is no limitation for anything.
- Create a new file and keep the above-collected data.

Links that are being used for data collection-

- 0https://www.wbhealth.gov.in/other_files/2007/14_5.html
- https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_population
- https://www.kaggle.com/datasets/webaccess/india-census-yearly-data
- https://censusindia.gov.in/nada/index.php/catalog/20030

Phase 2:

After collecting the data,you need to Analyse the dataset.

- Analysis of Literacy Rate vs Total Crimes
- Analysis of the type of crime vs each state vs Literacy rate.
- Analysis of year-on-year total crime rate.
- Analysis of area vs overall crime
- Analysis of Population vs overall Crime

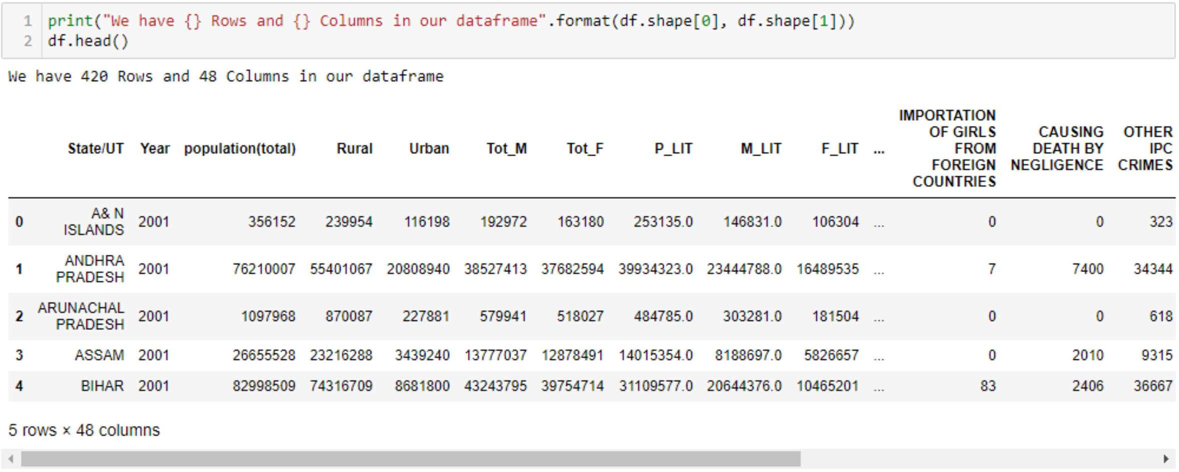
- Data Sources and their formats

The dataset is in the form of CSV (Comma Separated Value) format and consists of 48 columns with 420 number of records as explained below:

- State/UT: Each state present in india.
- Year : 2001 to 2012
- population(total)- total population of state or UT in each year from 2001 to 2012
- Rural : Population of Rural area of that state
- 'Urban :Population of urban area of that state,
- 'Tot_M' : Total population of Males present in that State/UT
- 'Tot_F' : Total population of Females present in that State/UT,
- 'P_LIT' - Total Literate population of each state,
- 'M_LIT' : Male literacy population of each state
- F_LIT : Female literate population in each state
- Lit_rate : Literacy Rate
- Area- Area in KM square of each state
- Murder- number of murder in each stae
- 'ATTEMPT TO MURDER',
- 'CULPABLE HOMICIDE NOT AMOUNTING TO MURDER',
- 'RAPE',
- 'CUSTODIAL RAPE',
- 'OTHER RAPE',
- 'KIDNAPPING & ABDUCTION',
- 'KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS',
- 'KIDNAPPING AND ABDUCTION OF OTHERS',
- 'DACOITY',
- 'PREPARATION AND ASSEMBLY FOR DACOITY',
- 'ROBBERY',
- 'BURGLARY',

- 'THEFT',
- 'AUTO THEFT',
- 'OTHER THEFT',
- 'RIOTS',
- 'CRIMINAL BREACH OF TRUST',
- 'CHEATING',
- 'COUNTERFIETING',
- 'ARSON',
- 'HURT/GREVIOUS HURT',
- 'DOWRY DEATHS',
- ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY',
- 'INSULT TO MODESTY OF WOMEN',
- 'CRUELTY BY HUSBAND OR HIS RELATIVES',
- 'IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES',
- 'CAUSING DEATH BY NEGLIGENCE',
- 'OTHER IPC CRIMES',
- 'TOTAL IPC CRIMES',
- 'Total Crimes',
- 'tot_crimes_sc': Total crimes against SC comitted in each year,
- 'total crime against women : Total crimes against Women comitted in each year,
- 'Total crimes against STs': Total crimes against STs comitted in each year,
- 'Total crime against children- Total crimes against Children comitted in each year
- Crime Rate- A crime rate is defined as the total number of crimes performed per a certain number of people in a specified area. This is typically expressed per 100,000 people.

Accessing the CSV files and make a DataFrame-



```
print("We have {} Rows and {} Columns in our dataframe".format(df.shape[0], df.shape[1]))
df.head()
```

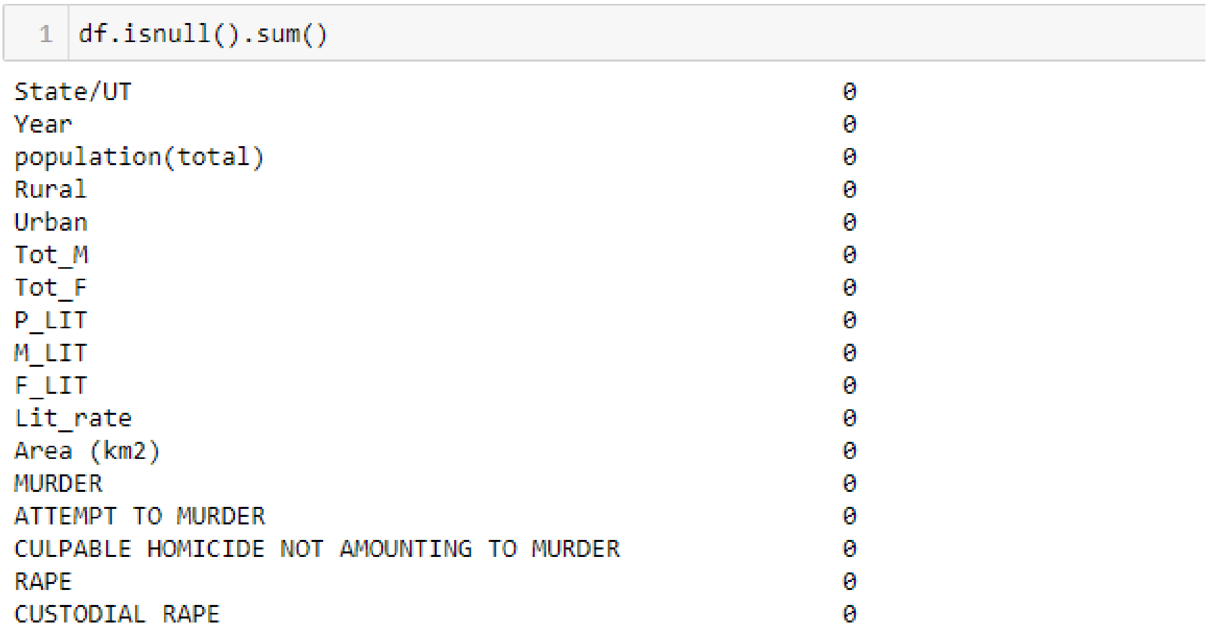
```
We have 420 Rows and 48 Columns in our dataframe
```

	State/UT	Year	population(total)	Rural	Urban	Tot_M	Tot_F	P_LIT	M_LIT	F_LIT	...	IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES	CAUSING DEATH BY NEGLIGENCE	OTHER IPC CRIMES
0	A& N ISLANDS	2001	356152	239954	116198	192972	163180	253135.0	146831.0	106304	...	0	0	323
1	ANDHRA PRADESH	2001	76210007	55401067	20808940	38527413	37682594	39934323.0	23444788.0	16489535	...	7	7400	34344
2	ARUNACHAL PRADESH	2001	1097968	870087	227881	579941	518027	484785.0	303281.0	181504	...	0	0	618
3	ASSAM	2001	26655528	23216288	3439240	13777037	12878491	14015354.0	8188697.0	5826657	...	0	2010	9315
4	BIHAR	2001	82998509	74316709	8681800	43243795	39754714	31109577.0	20644376.0	10465201	...	83	2406	36667

5 rows × 48 columns

- ## Data Preprocessing Done

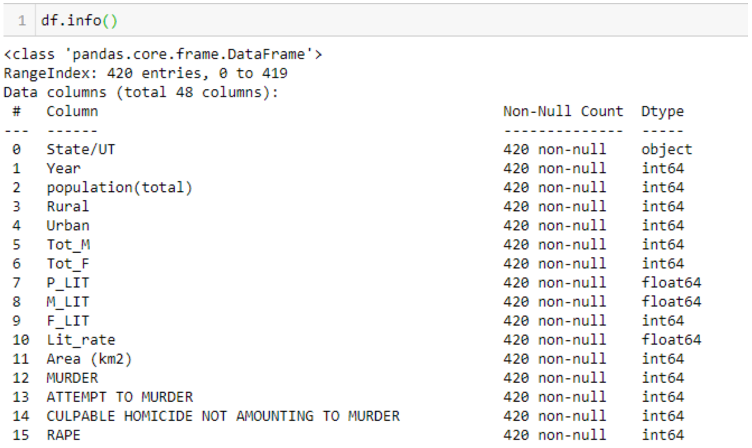
For the data pre-processing step, I checked through the dataframe for missing values



```
df.isnull().sum()
```

```
State/UT                                                0
Year                                                    0
population(total)                                       0
Rural                                                   0
Urban                                                   0
Tot_M                                                   0
Tot_F                                                   0
P_LIT                                                   0
M_LIT                                                   0
F_LIT                                                   0
Lit_rate                                                0
Area (km2)                                              0
MURDER                                                  0
ATTEMPT TO MURDER                                       0
CULPABLE HOMICIDE NOT AMOUNTING TO MURDER               0
RAPE                                                    0
CUSTODIAL RAPE                                          0
```

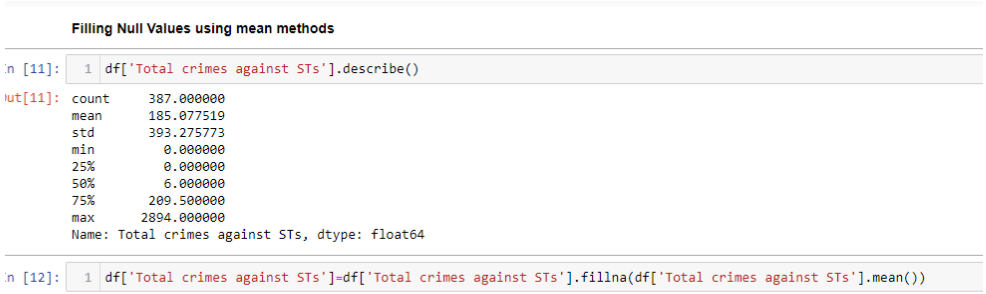
Checked the data type details for each column to understand the numeric ones



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 420 entries, 0 to 419
Data columns (total 48 columns):
 #   Column                                              Non-Null Count  Dtype
---  ------                                              --------------  -----
 0   State/UT                                            420 non-null    object
 1   Year                                                420 non-null    int64
 2   population(total)                                   420 non-null    int64
 3   Rural                                               420 non-null    int64
 4   Urban                                               420 non-null    int64
 5   Tot_M                                               420 non-null    int64
 6   Tot_F                                               420 non-null    int64
 7   P_LIT                                               420 non-null    float64
 8   M_LIT                                               420 non-null    float64
 9   F_LIT                                               420 non-null    int64
 10  Lit_rate                                            420 non-null    float64
 11  Area (km2)                                          420 non-null    int64
 12  MURDER                                              420 non-null    int64
 13  ATTEMPT TO MURDER                                   420 non-null    int64
 14  CULPABLE HOMICIDE NOT AMOUNTING TO MURDER           420 non-null    int64
 15  RAPE                                                420 non-null    int64
```

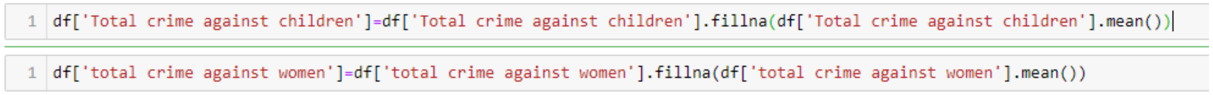
The various data null value filling on our data set are shown below with the code.Missing value found in 3 columns- total crime against women, Total crimes against STs, Total crime against children.



Filling Null Values using mean methods

```
n [11]:  df['Total crimes against STs'].describe()

ut[11]: count     387.000000
        mean      185.077519
        std       393.275773
        min         0.000000
        25%         0.000000
        50%         6.000000
        75%       209.500000
        max      2894.000000
        Name: Total crimes against STs, dtype: float64
```

```
n [12]:  df['Total crimes against STs']=df['Total crimes against STs'].fillna(df['Total crimes against STs'].mean())
```

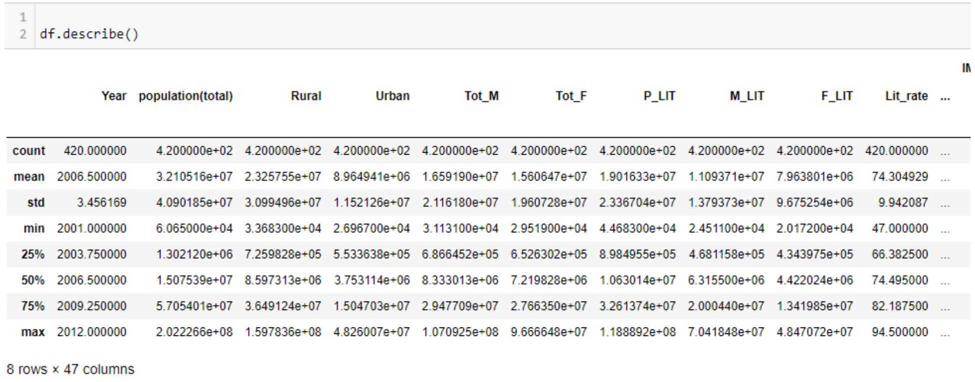


```
df['Total crime against children']=df['Total crime against children'].fillna(df['Total crime against children'].mean())
```

```
df['total crime against women']=df['total crime against women'].fillna(df['total crime against women'].mean())
```

I then used the “describe” method to check the count, mean, standard deviation, minimum, maximum, 25%, 50% and 75% quartile data

Code:



```
df.describe()
```

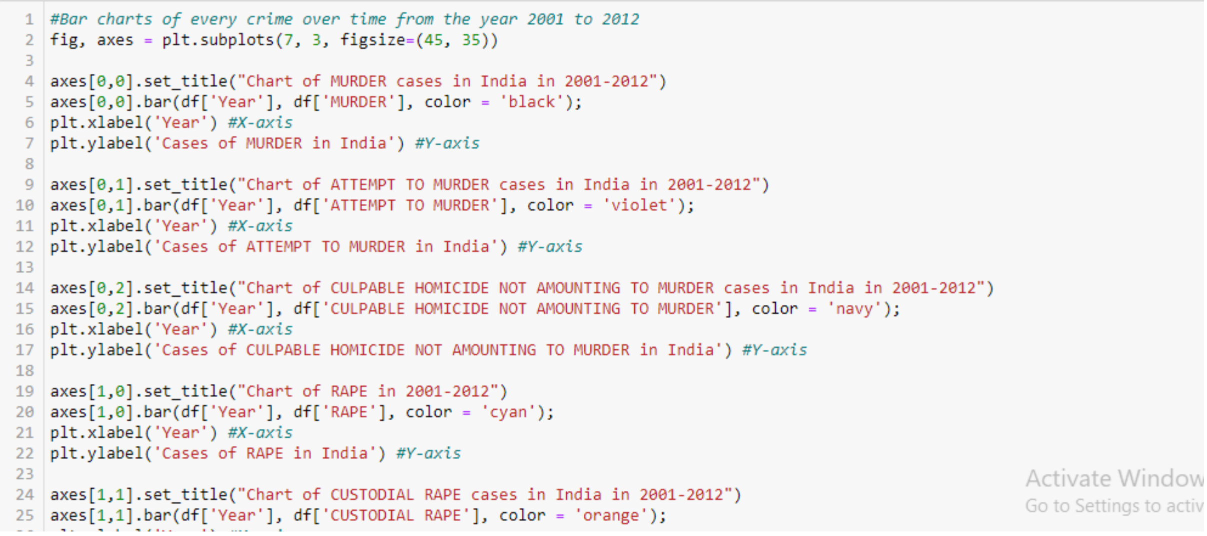
										IN
	Year	**population(total)**	**Rural**	**Urban**	**Tot_M**	**Tot_F**	**P_LIT**	**M_LIT**	**F_LIT**	**Lit_rate**	**...**
count	420.000000	4.200000e+02	4.200000e+02	4.200000e+02	4.200000e+02	4.200000e+02	4.200000e+02	4.200000e+02	4.200000e+02	420.000000	...
mean	2006.500000	3.210516e+07	2.325755e+07	8.964941e+06	1.659190e+07	1.560647e+07	1.901633e+07	1.109371e+07	7.963801e+06	74.304929	...
std	3.456169	4.090185e+07	3.099496e+07	1.152126e+07	2.116180e+07	1.960728e+07	2.336704e+07	1.379373e+07	9.675254e+06	9.942087	...
min	2001.000000	6.065000e+04	3.368300e+04	2.696700e+04	3.113100e+04	2.951900e+04	4.468300e+04	2.451100e+04	2.017200e+04	47.000000	...
25%	2003.750000	1.302120e+06	7.259828e+05	5.533638e+05	6.866452e+05	6.526302e+05	8.984955e+05	4.681158e+05	4.343975e+05	66.382500	...
50%	2006.500000	1.507539e+07	8.597313e+06	3.753114e+06	8.333013e+06	7.219828e+06	1.063014e+07	6.315500e+06	4.422024e+06	74.495000	...
75%	2009.250000	5.705401e+07	3.649124e+07	1.504703e+07	2.947709e+07	2.766350e+07	3.261374e+07	2.000440e+07	1.341985e+07	82.187500	...
max	2012.000000	2.022266e+08	1.597836e+08	4.826007e+07	1.070925e+08	9.666648e+07	1.188892e+08	7.041848e+07	4.847072e+07	94.500000	...

8 rows × 47 columns

- ## Visualization

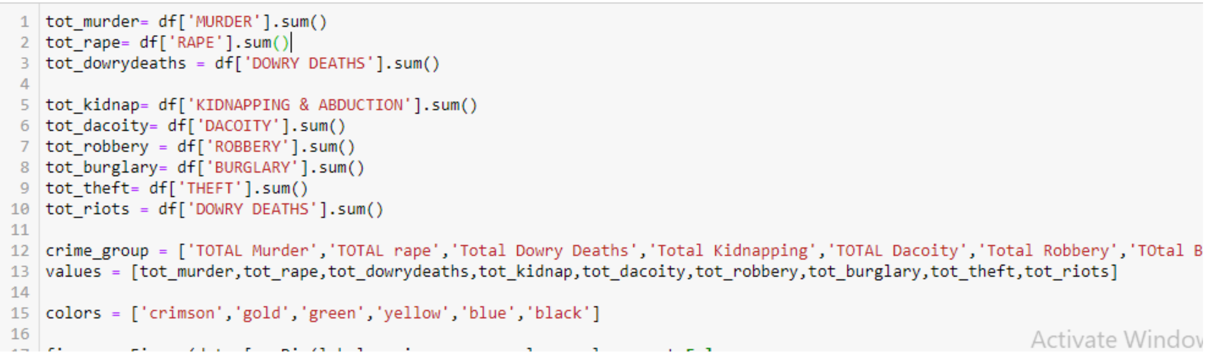
  Let’s start visualising the columns, here we starting with the Type of crime like , Murder, Rape, Robbery happened in each year

  Code:

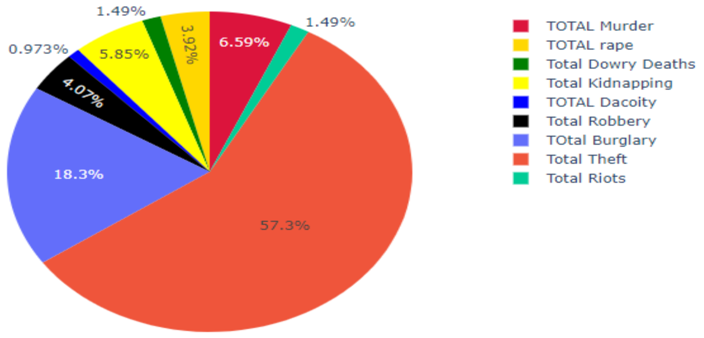


```python
#Bar charts of every crime over time from the year 2001 to 2012
fig, axes = plt.subplots(7, 3, figsize=(45, 35))

axes[0,0].set_title("Chart of MURDER cases in India in 2001-2012")
axes[0,0].bar(df['Year'], df['MURDER'], color = 'black');
plt.xlabel('Year') #X-axis
plt.ylabel('Cases of MURDER in India') #Y-axis

axes[0,1].set_title("Chart of ATTEMPT TO MURDER cases in India in 2001-2012")
axes[0,1].bar(df['Year'], df['ATTEMPT TO MURDER'], color = 'violet');
plt.xlabel('Year') #X-axis
plt.ylabel('Cases of ATTEMPT TO MURDER in India') #Y-axis

axes[0,2].set_title("Chart of CULPABLE HOMICIDE NOT AMOUNTING TO MURDER cases in India in 2001-2012")
axes[0,2].bar(df['Year'], df['CULPABLE HOMICIDE NOT AMOUNTING TO MURDER'], color = 'navy');
plt.xlabel('Year') #X-axis
plt.ylabel('Cases of CULPABLE HOMICIDE NOT AMOUNTING TO MURDER in India') #Y-axis

axes[1,0].set_title("Chart of RAPE in 2001-2012")
axes[1,0].bar(df['Year'], df['RAPE'], color = 'cyan');
plt.xlabel('Year') #X-axis
plt.ylabel('Cases of RAPE in India') #Y-axis

axes[1,1].set_title("Chart of CUSTODIAL RAPE cases in India in 2001-2012")
axes[1,1].bar(df['Year'], df['CUSTODIAL RAPE'], color = 'orange');
```

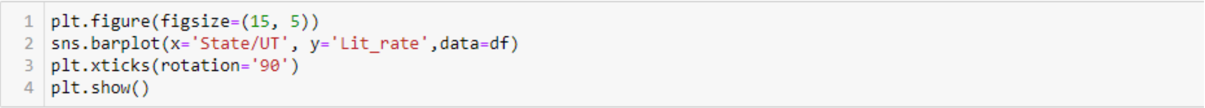
Plotting Pie chart for different crimes



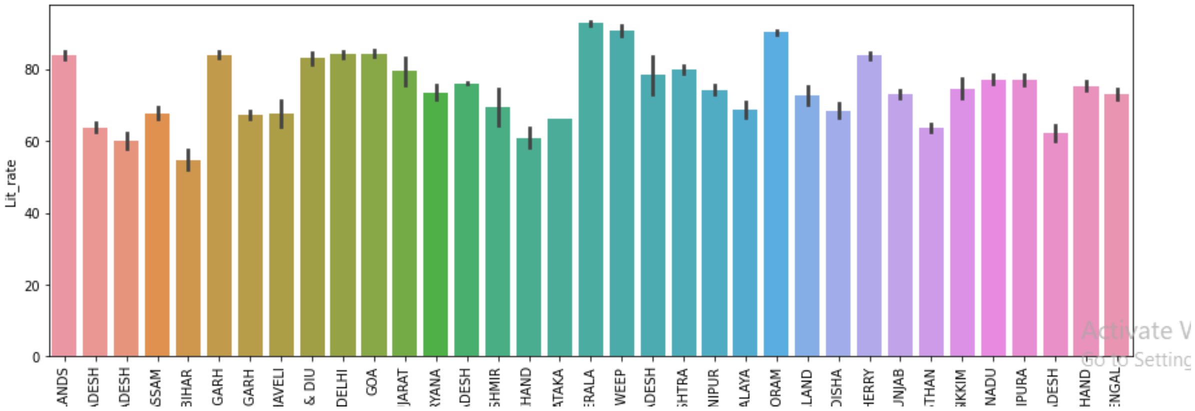
```python
tot_murder= df['MURDER'].sum()
tot_rape= df['RAPE'].sum()
tot_dowrydeaths = df['DOWRY DEATHS'].sum()

tot_kidnap= df['KIDNAPPING & ABDUCTION'].sum()
tot_dacoity= df['DACOITY'].sum()
tot_robbery = df['ROBBERY'].sum()
tot_burglary= df['BURGLARY'].sum()
tot_theft= df['THEFT'].sum()
tot_riots = df['DOWRY DEATHS'].sum()

crime_group = ['TOTAL Murder','TOTAL rape','Total Dowry Deaths','Total Kidnapping','TOTAL Dacoity','Total Robbery','TOtal B
values = [tot_murder,tot_rape,tot_dowrydeaths,tot_kidnap,tot_dacoity,tot_robbery,tot_burglary,tot_theft,tot_riots]

colors = ['crimson','gold','green','yellow','blue','black']
```

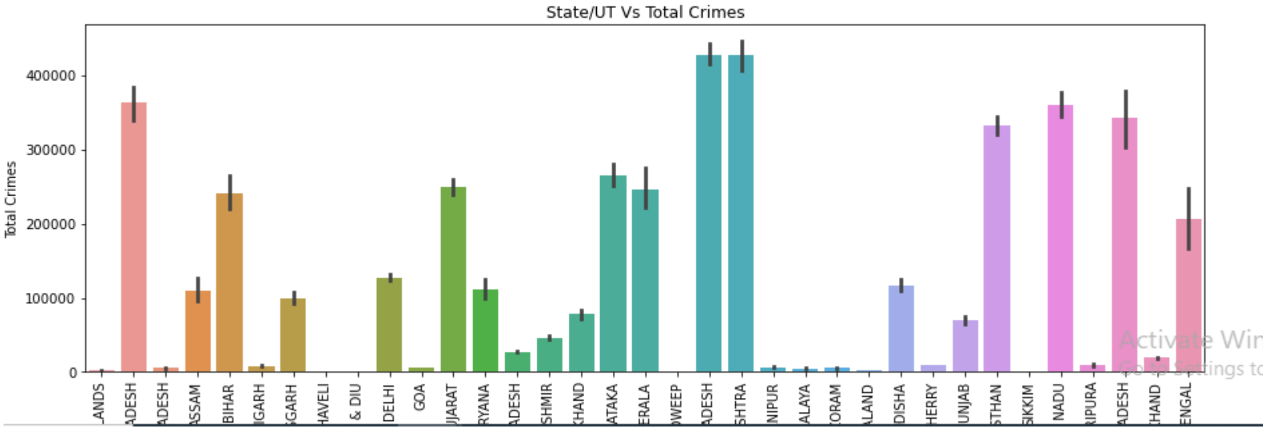


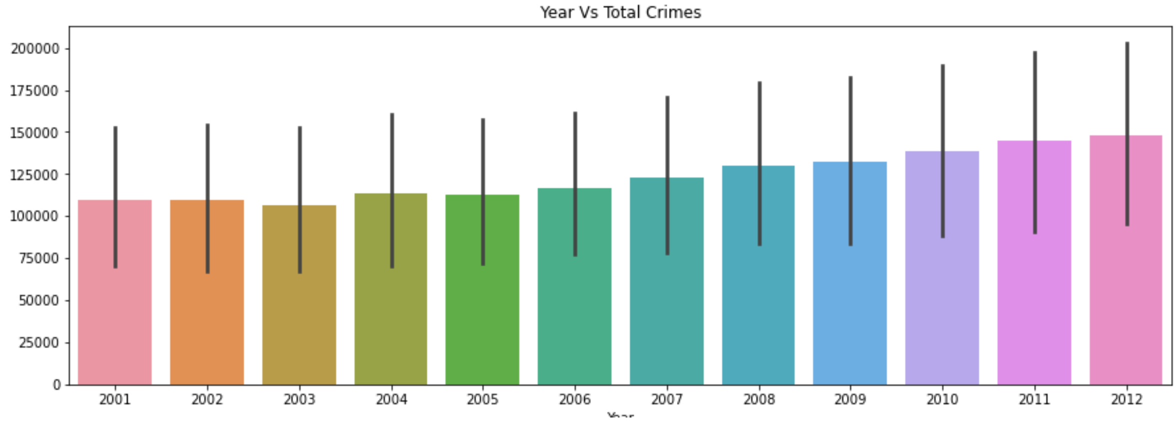
State/UT VS Literacy Rate

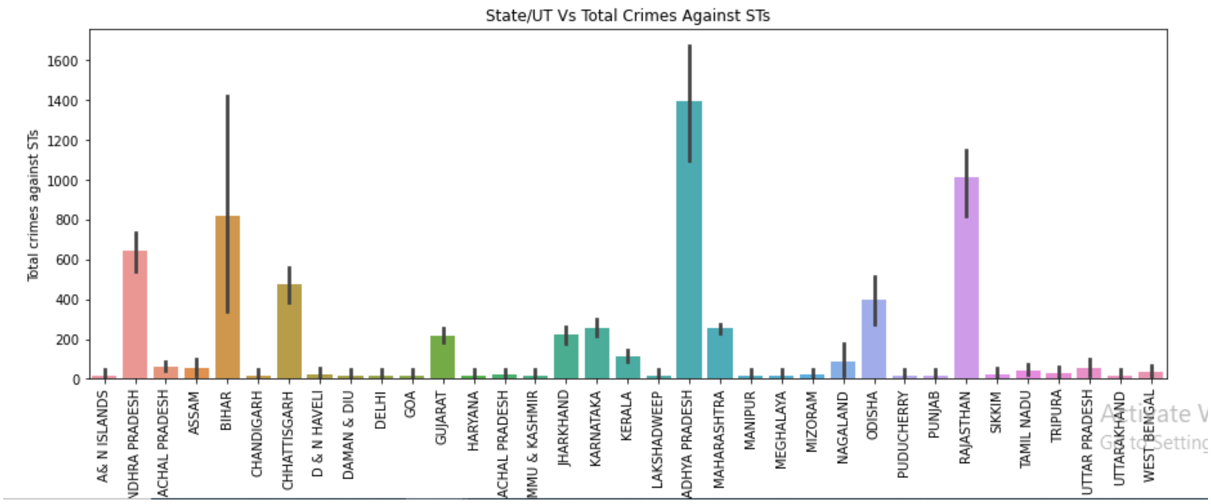


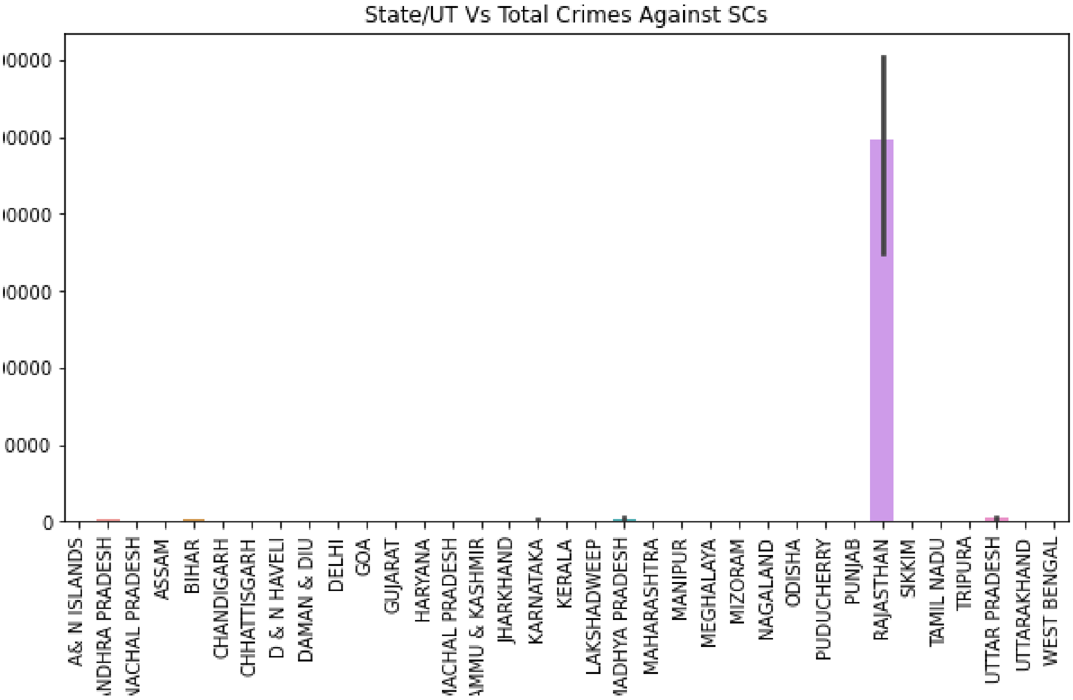
```
plt.figure(figsize=(15, 5))
sns.barplot(x='State/UT', y='Lit_rate',data=df)
plt.xticks(rotation='90')
plt.show()
```

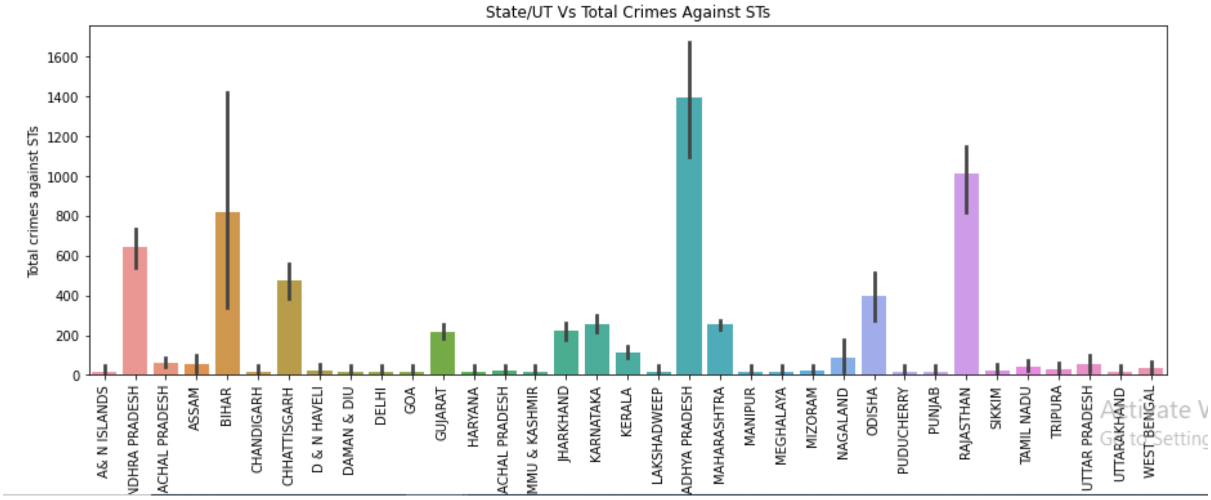


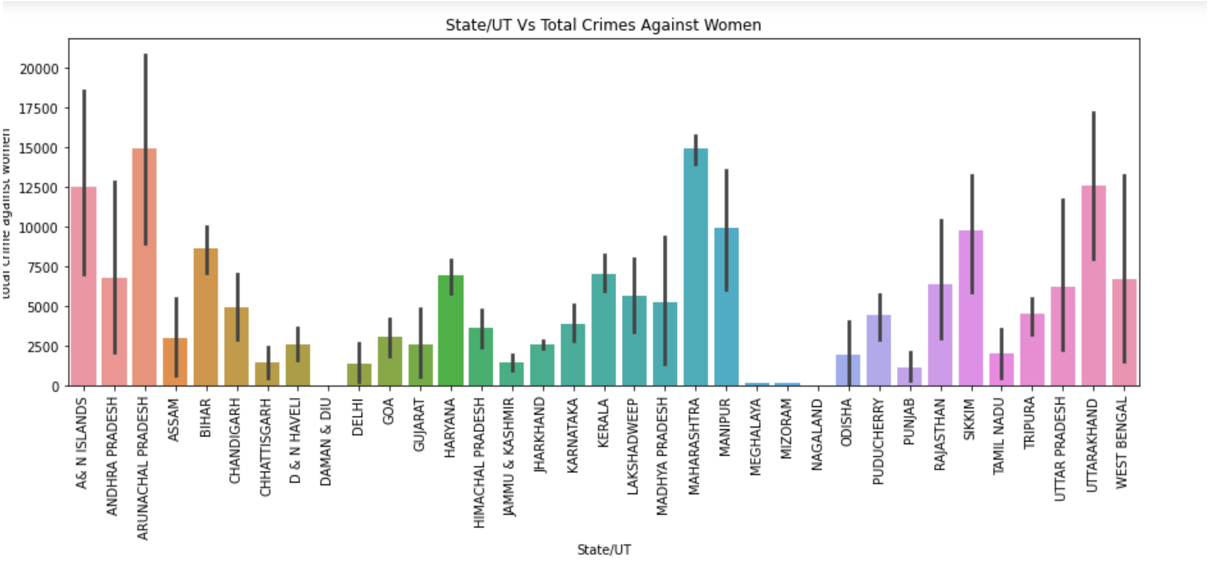


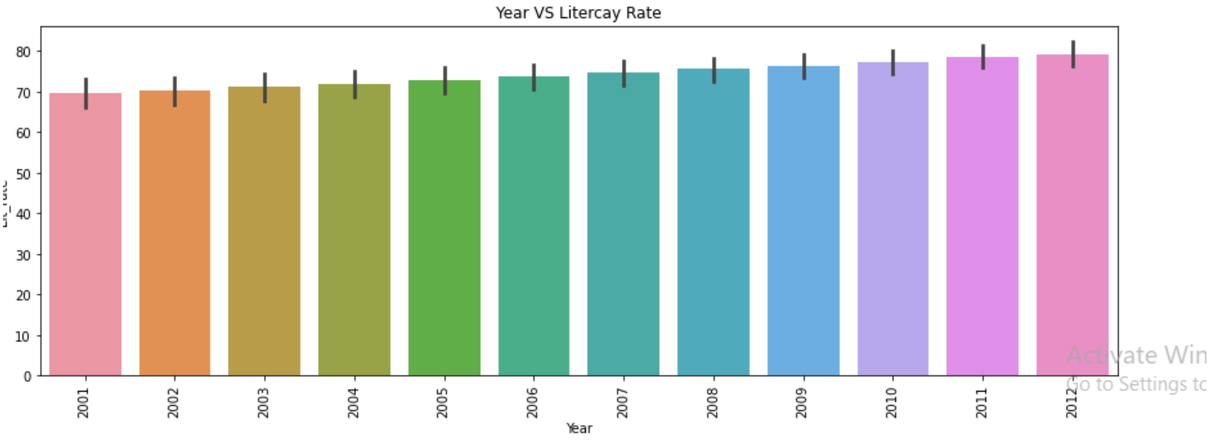


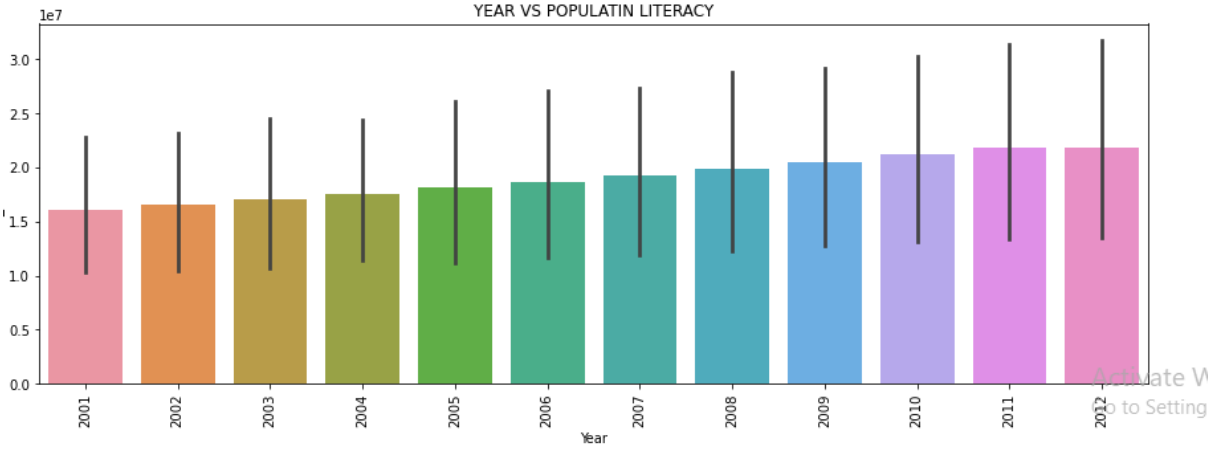

State/UT Vs Total Crimes Against STs
Total crimes against STs
1600
1400
1200
1000
800
600
400
200
0
A& N ISLANDS
NDHRA PRADESH
ACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
ACHAL PRADESH
MMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP
ADHYA PRADESH
MAHARASHTRA
MANIPUR
MEGHALAYA
MIZORAM
NAGALAND
ODISHA
PUDUCHERRY
PUNJAB
RAJASTHAN
SIKKIM
TAMIL NADU
TRIPURA
UTTAR PRADESH
UTTARAKHAND
WEST BENGAL

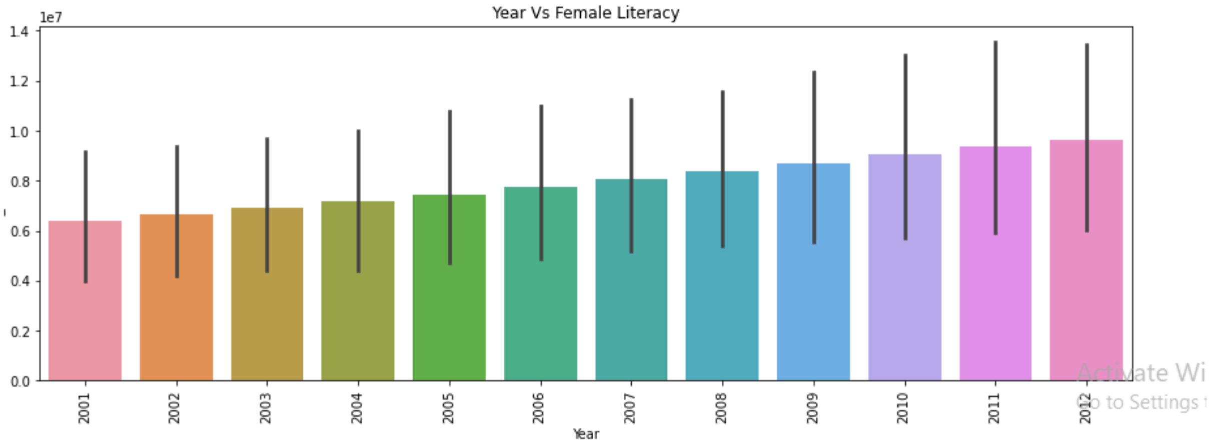

State/UT Vs Total Crimes Against SCs
)0000
)0000
)0000
)0000
)0000
0000
0
A& N ISLANDS
NDHRA PRADESH
NACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
MACHAL PRADESH
MMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP
MADHYA PRADESH
MAHARASHTRA
MANIPUR
MEGHALAYA
MIZORAM
NAGALAND
ODISHA
PUDUCHERRY
PUNJAB
RAJASTHAN
SIKKIM
TAMIL NADU
TRIPURA
UTTAR PRADESH
UTTARAKHAND
WEST BENGAL


State/UT Vs Total Crimes Against STs
Total crimes against STs
1600
1400
1200
1000
800
600
400
200
0
A& N ISLANDS
NDHRA PRADESH
ACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
ACHAL PRADESH
MMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP
ADHYA PRADESH
MAHARASHTRA
MANIPUR
MEGHALAYA
MIZORAM
NAGALAND
ODISHA
PUDUCHERRY
PUNJAB
RAJASTHAN
SIKKIM
TAMIL NADU
TRIPURA
UTTAR PRADESH
UTTARAKHAND
WEST BENGAL


State/UT Vs Total Crimes Against Women
total crime against women
0
2500
5000
7500
10000
12500
15000
17500
20000
A& N ISLANDS
ANDHRA PRADESH
ARUNACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
HIMACHAL PRADESH
JAMMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP
MADHYA PRADESH
MAHARASHTRA
MANIPUR
MEGHALAYA
MIZORAM
NAGALAND
ODISHA
PUDUCHERRY
PUNJAB
RAJASTHAN
SIKKIM
TAMIL NADU
TRIPURA
UTTAR PRADESH
UTTARAKHAND
WEST BENGAL
State/UT

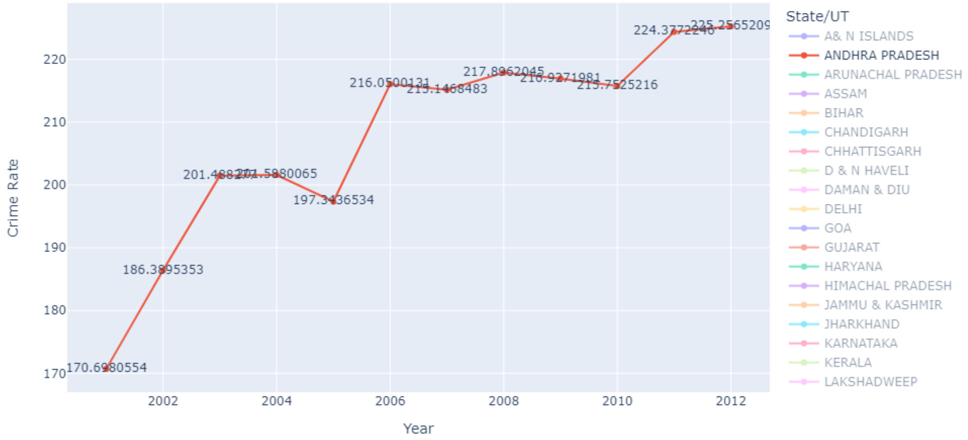

Year VS Litercay Rate
Lit_rate
0
10
20
30
40
50
60
70
80
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year

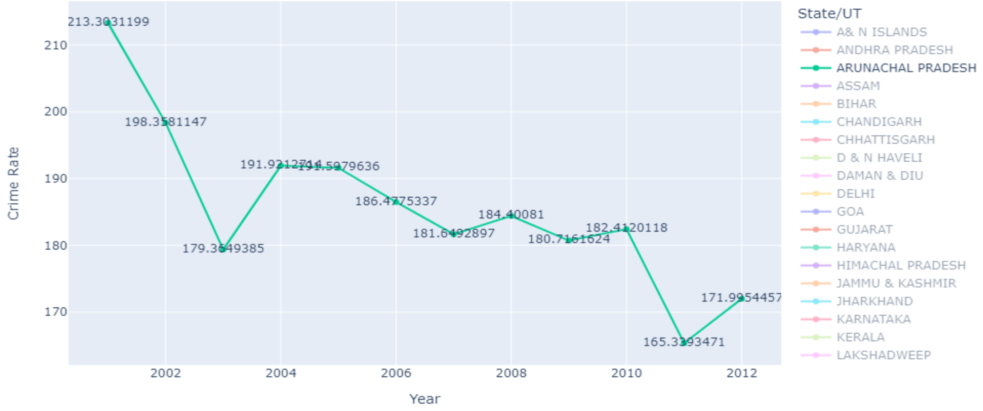

YEAR VS POPULATIN LITERACY
1e7
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
Year

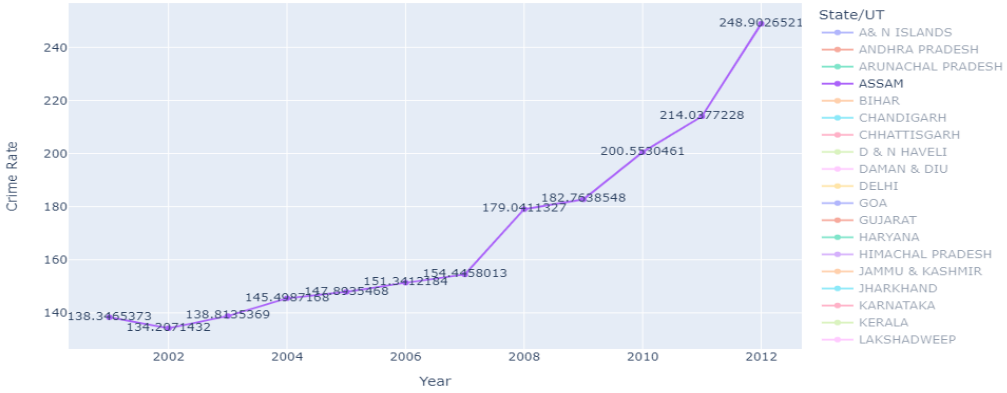


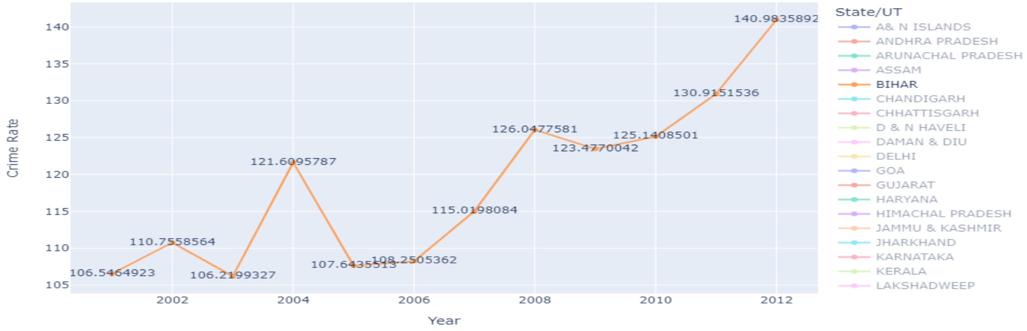
Similarly I analysed the State/UT vs Total crimes/Total crimes against SCs, /STs,Children

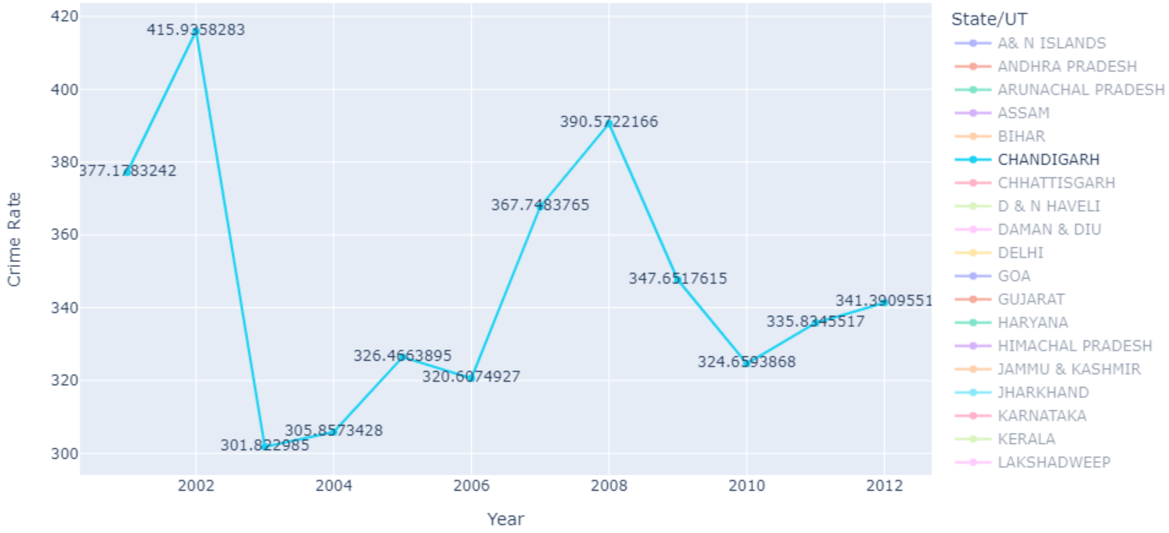
Plotted Line plotas well for Crime rate analysis for each state.

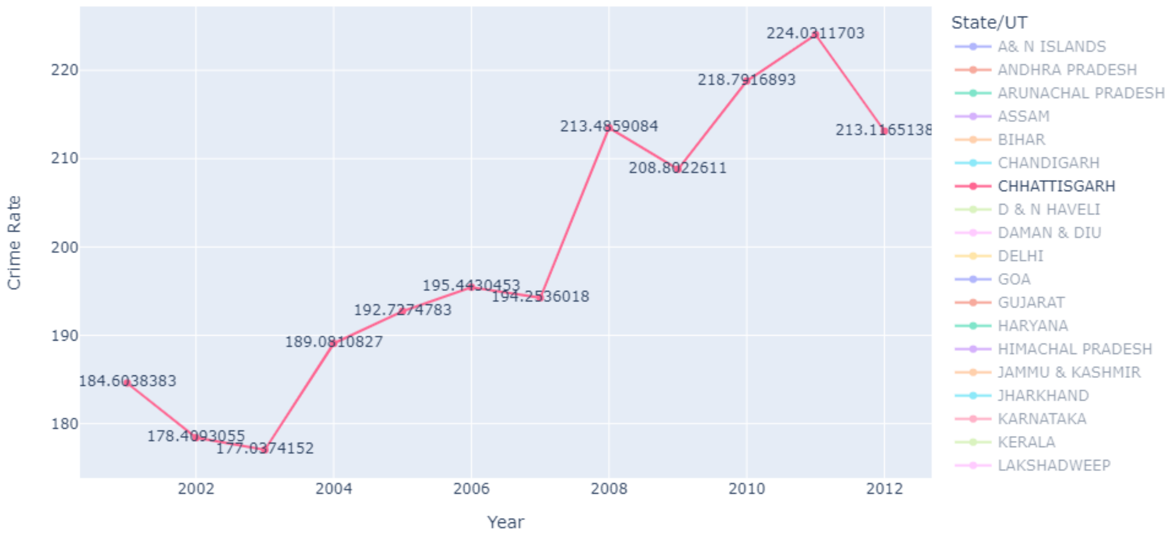


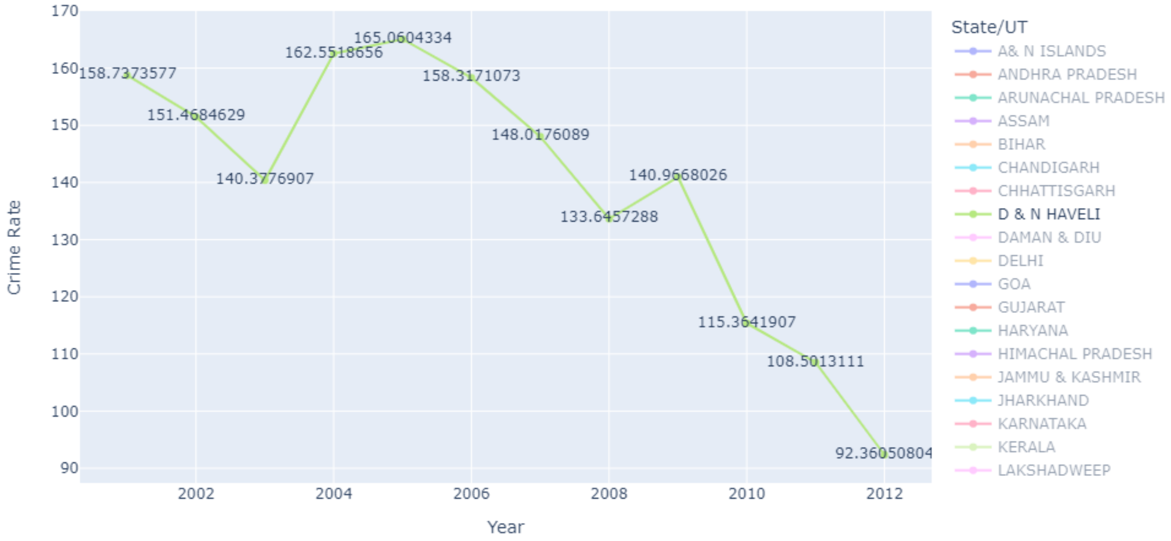



Crime Rate
Year
2002
2004
2006
2008
2010
2012
140
160
180
200
220
240
138.3465373
134.2071432
138.8135369
145.4987168
147.8935468
151.3412184
154.4458013
179.0411327
182.7638548
200.5530461
214.0377228
248.9026521
State/UT
A& N ISLANDS
ANDHRA PRADESH
ARUNACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
HIMACHAL PRADESH
JAMMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP


Crime Rate
Year
2002
2004
2006
2008
2010
2012
105
110
115
120
125
130
135
140
106.5464923
110.7558564
106.2199327
121.6095787
107.6435513
108.2505362
115.0198084
126.0477581
123.4770042
125.1408501
130.9151536
140.9835892
State/UT
A& N ISLANDS
ANDHRA PRADESH
ARUNACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
HIMACHAL PRADESH
JAMMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP


Crime Rate
Year
2002
2004
2006
2008
2010
2012
300
320
340
360
380
400
420
377.1783242
415.9358283
301.822985
305.8573428
326.4663895
320.6074927
367.7483765
390.5722166
347.6517615
324.6593868
335.8345517
341.3909551
State/UT
A& N ISLANDS
ANDHRA PRADESH
ARUNACHAL PRADESH
ASSAM
BIHAR
CHANDIGARH
CHHATTISGARH
D & N HAVELI
DAMAN & DIU
DELHI
GOA
GUJARAT
HARYANA
HIMACHAL PRADESH
JAMMU & KASHMIR
JHARKHAND
KARNATAKA
KERALA
LAKSHADWEEP





- Hardware and Software Requirements and Tools Used

Hardware technology being used.

RAM : 8 GB

CPU : AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz

GPU : AMD Radeon ™ Vega 8 Graphics and NVIDIA GeForce GTX 1650 Ti

Software technology being used.

Programming language : Python

Distribution : Anaconda Navigator

Browser based language shell : Jupyter Notebook

Libraries/Packages specifically being used.

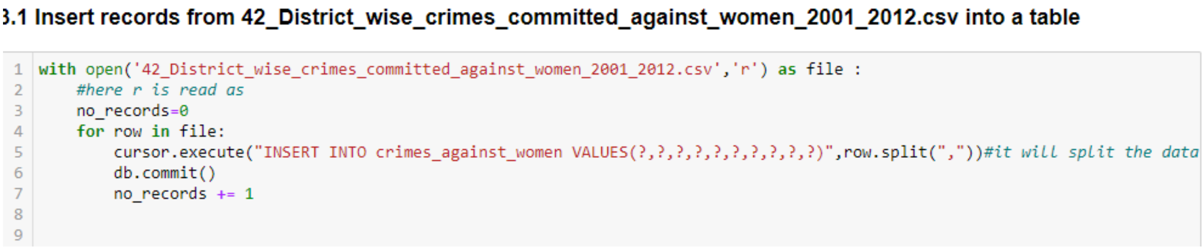
Pandas, NumPy, matplotlib, seaborn, scikit-learn, plotly

Phase 3- SQL

Insert Multiple record from the table given and extract the data as asked.

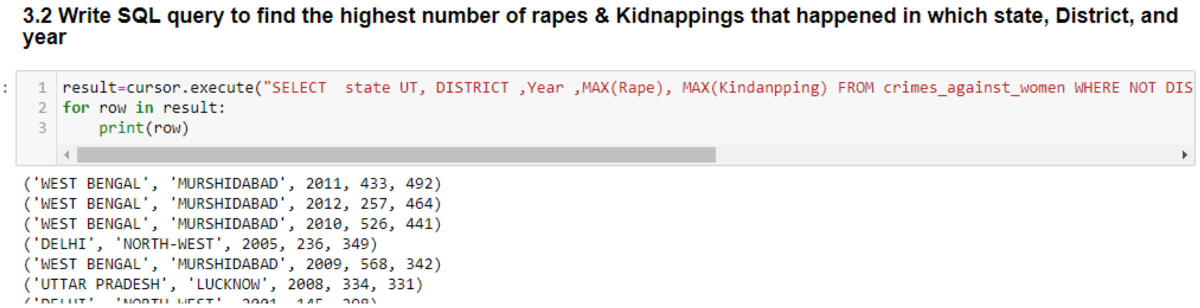
The CSV files that are used:

- 42_District_wise_crimes_committed_against_women_2001_2012
- 02_District_wise_crimes_committed_against_ST_2001_2012
- 01_District_wise_crimes_committed_IPC_2001_2012



3.1 Insert records from 42_District_wise_crimes_committed_against_women_2001_2012.csv into a table

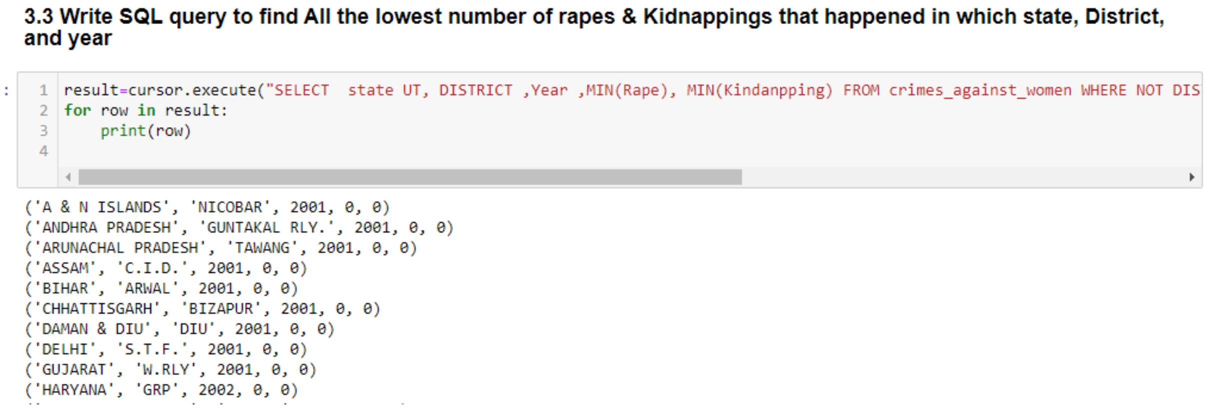
```
with open('42_District_wise_crimes_committed_against_women_2001_2012.csv','r') as file :
    #here r is read as
    no_records=0
    for row in file:
        cursor.execute("INSERT INTO crimes_against_women VALUES(?,?,?,?,?,?,?,?,?,?)",row.split(","))#it will split the data
        db.commit()
        no_records += 1


```



3.2 Write SQL query to find the highest number of rapes & Kidnappings that happened in which state, District, and year

```
result=cursor.execute("SELECT  state UT, DISTRICT ,Year ,MAX(Rape), MAX(Kindanpping) FROM crimes_against_women WHERE NOT DIS
for row in result:
    print(row)
```

```
('WEST BENGAL', 'MURSHIDABAD', 2011, 433, 492)
('WEST BENGAL', 'MURSHIDABAD', 2012, 257, 464)
('WEST BENGAL', 'MURSHIDABAD', 2010, 526, 441)
('DELHI', 'NORTH-WEST', 2005, 236, 349)
('WEST BENGAL', 'MURSHIDABAD', 2009, 568, 342)
('UTTAR PRADESH', 'LUCKNOW', 2008, 334, 331)
('DELHI', 'NORTH-WEST', 2001, 145, 208)
```

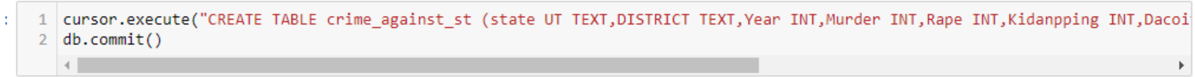


3.3 Write SQL query to find All the lowest number of rapes & Kidnappings that happened in which state, District, and year

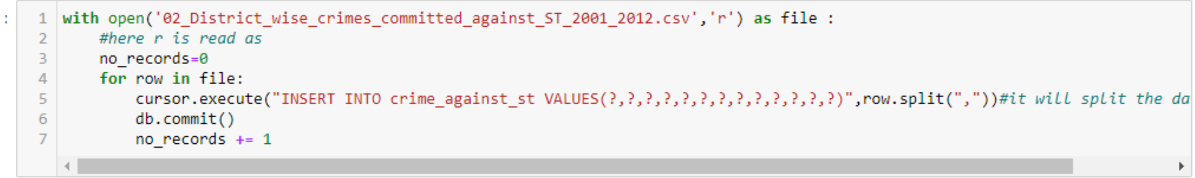
```
result=cursor.execute("SELECT  state UT, DISTRICT ,Year ,MIN(Rape), MIN(Kindanpping) FROM crimes_against_women WHERE NOT DIS
for row in result:
    print(row)

```

```
('A & N ISLANDS', 'NICOBAR', 2001, 0, 0)
('ANDHRA PRADESH', 'GUNTAKAL RLY.', 2001, 0, 0)
('ARUNACHAL PRADESH', 'TAWANG', 2001, 0, 0)
('ASSAM', 'C.I.D.', 2001, 0, 0)
('BIHAR', 'ARWAL', 2001, 0, 0)
('CHHATTISGARH', 'BIZAPUR', 2001, 0, 0)
('DAMAN & DIU', 'DIU', 2001, 0, 0)
('DELHI', 'S.T.F.', 2001, 0, 0)
('GUJARAT', 'W.RLY', 2001, 0, 0)
('HARYANA', 'GRP', 2002, 0, 0)
```

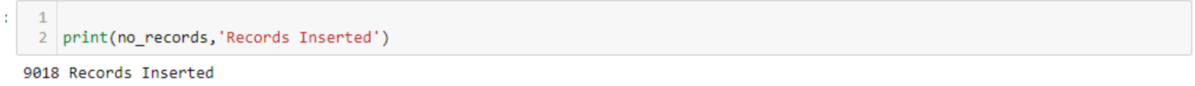
3.4 Insert records from 02_District_wise_crimes_committed_against_ST_2001_2012.csv into a new table



```
cursor.execute("CREATE TABLE crime_against_st (state UT TEXT,DISTRICT TEXT,Year INT,Murder INT,Rape INT,Kidanpping INT,Dacoi
db.commit()
```



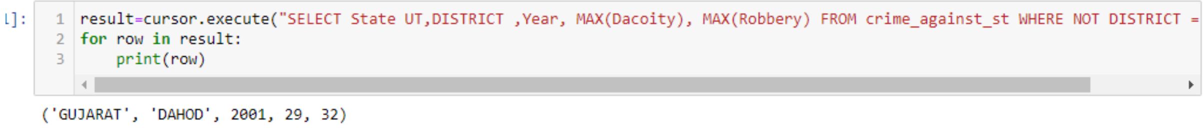
```
with open('02_District_wise_crimes_committed_against_ST_2001_2012.csv','r') as file :
    #here r is read as
    no_records=0
    for row in file:
        cursor.execute("INSERT INTO crime_against_st VALUES(?,?,?,?,?,?,?,?,?,?,?,?,?)",row.split(","))#it will split the da
        db.commit()
        no_records += 1
```



```

print(no_records,'Records Inserted')
```

```
9018 Records Inserted
```

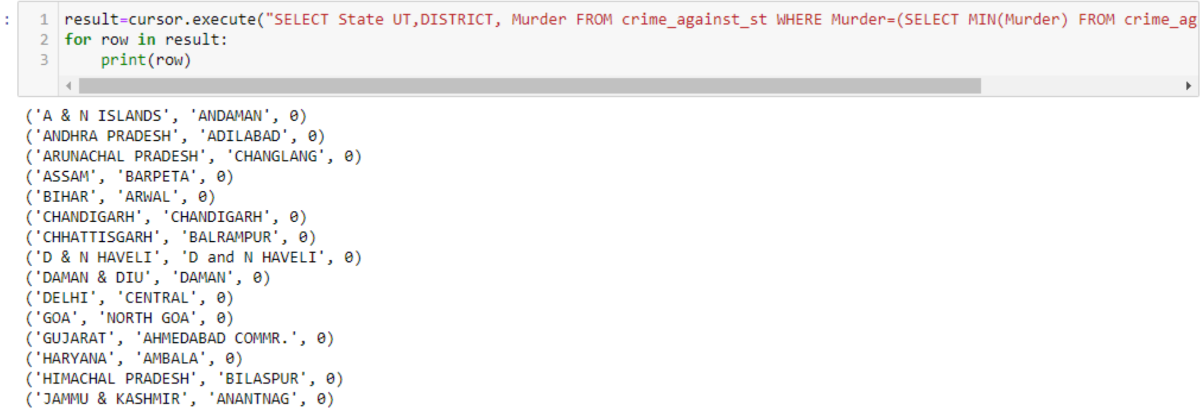
3.5 Write SQL query to find the highest number of dacoity/robbery in which district.



```
result=cursor.execute("SELECT State UT,DISTRICT ,Year, MAX(Dacoity), MAX(Robbery) FROM crime_against_st WHERE NOT DISTRICT =
for row in result:
    print(row)
```

```
('GUJARAT', 'DAHOD', 2001, 29, 32)
```

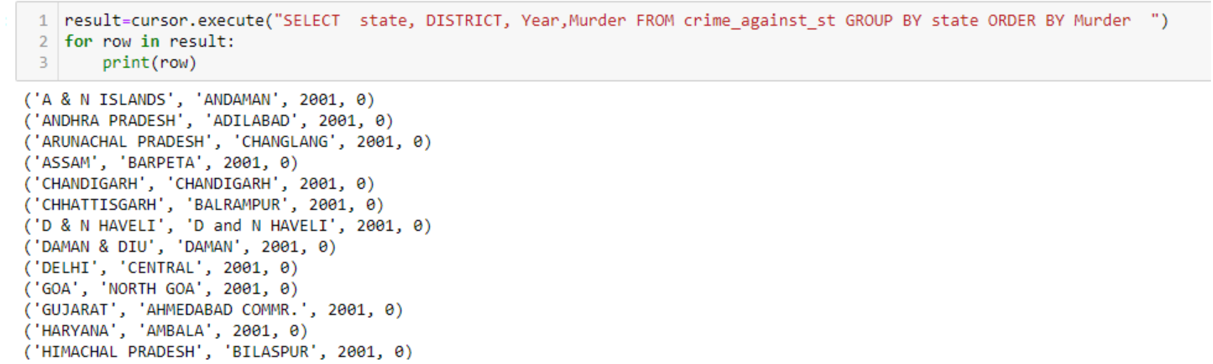
3.6 Write SQL query to find in which districts(All) the lowest number of murders happened



```
result=cursor.execute("SELECT State UT,DISTRICT, Murder FROM crime_against_st WHERE Murder=(SELECT MIN(Murder) FROM crime_ag
for row in result:
    print(row)
```

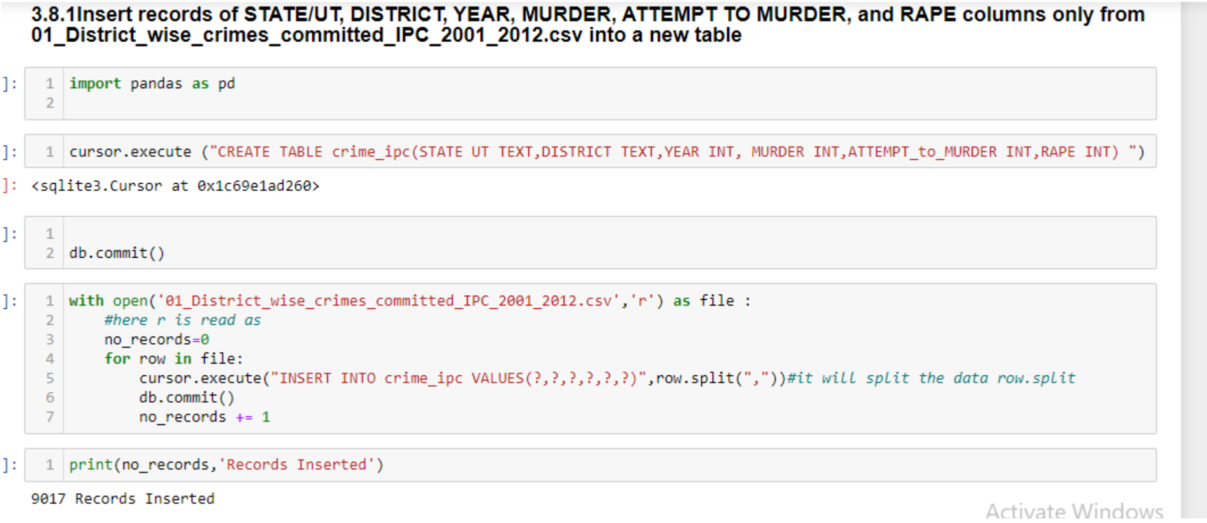
```
('A & N ISLANDS', 'ANDAMAN', 0)
('ANDHRA PRADESH', 'ADILABAD', 0)
('ARUNACHAL PRADESH', 'CHANGLANG', 0)
('ASSAM', 'BARPETA', 0)
('BIHAR', 'ARWAL', 0)
('CHANDIGARH', 'CHANDIGARH', 0)
('CHHATTISGARH', 'BALRAMPUR', 0)
('D & N HAVELI', 'D and N HAVELI', 0)
('DAMAN & DIU', 'DAMAN', 0)
('DELHI', 'CENTRAL', 0)
('GOA', 'NORTH GOA', 0)
('GUJARAT', 'AHMEDABAD COMMR.', 0)
('HARYANA', 'AMBALA', 0)
('HIMACHAL PRADESH', 'BILASPUR', 0)
('JAMMU & KASHMIR', 'ANANTNAG', 0)
```

3.7 Write SQL query to find the number of murders in ascending order in district and yearwise.



```
result=cursor.execute("SELECT  state, DISTRICT, Year,Murder FROM crime_against_st GROUP BY state ORDER BY Murder  ")
for row in result:
    print(row)
```

```
('A & N ISLANDS', 'ANDAMAN', 2001, 0)
('ANDHRA PRADESH', 'ADILABAD', 2001, 0)
('ARUNACHAL PRADESH', 'CHANGLANG', 2001, 0)
('ASSAM', 'BARPETA', 2001, 0)
('CHANDIGARH', 'CHANDIGARH', 2001, 0)
('CHHATTISGARH', 'BALRAMPUR', 2001, 0)
('D & N HAVELI', 'D and N HAVELI', 2001, 0)
('DAMAN & DIU', 'DAMAN', 2001, 0)
('DELHI', 'CENTRAL', 2001, 0)
('GOA', 'NORTH GOA', 2001, 0)
('GUJARAT', 'AHMEDABAD COMMR.', 2001, 0)
('HARYANA', 'AMBALA', 2001, 0)
('HIMACHAL PRADESH', 'BILASPUR', 2001, 0)
```



3.8.1Insert records of STATE/UT, DISTRICT, YEAR, MURDER, ATTEMPT TO MURDER, and RAPE columns only from 01_District_wise_crimes_committed_IPC_2001_2012.csv into a new table

```
import pandas as pd

```

```
cursor.execute ("CREATE TABLE crime_ipc(STATE UT TEXT,DISTRICT TEXT,YEAR INT, MURDER INT,ATTEMPT_to_MURDER INT,RAPE INT) ")
```

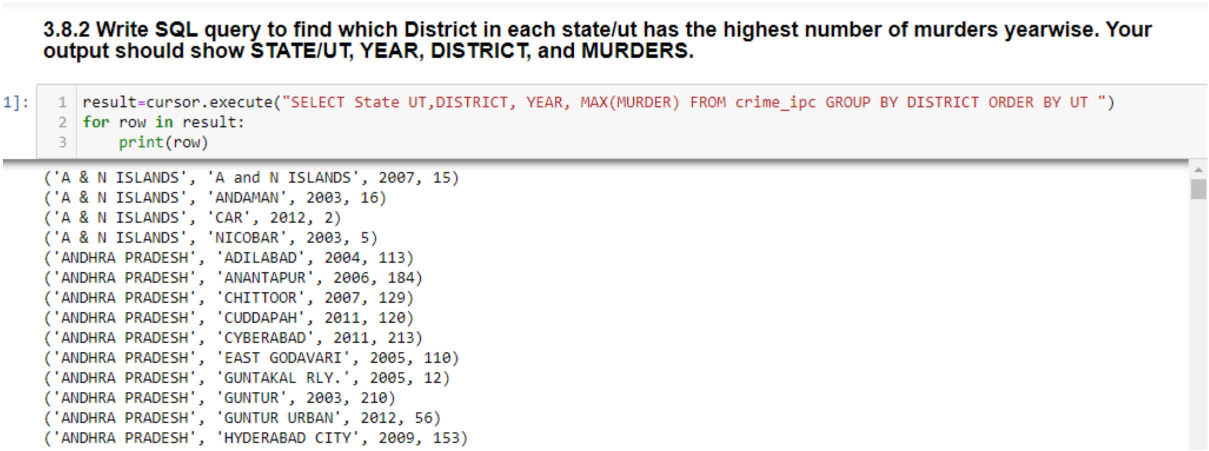
```
<sqlite3.Cursor at 0x1c69e1ad260>
```

```

db.commit()
```

```
with open('01_District_wise_crimes_committed_IPC_2001_2012.csv','r') as file :
    #here r is read as
    no_records=0
    for row in file:
        cursor.execute("INSERT INTO crime_ipc VALUES(?,?,?,?,?,?)",row.split(","))#it will split the data row.split
        db.commit()
        no_records += 1
```

```
print(no_records,'Records Inserted')
```

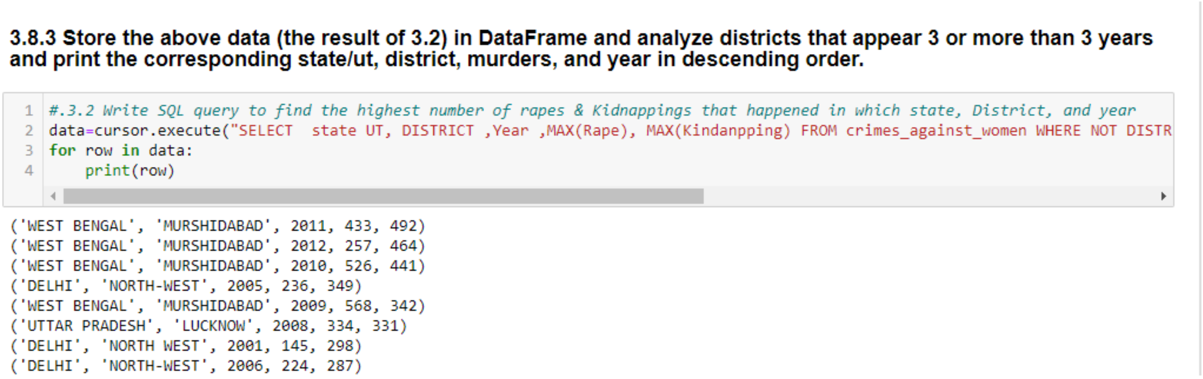
```
9017 Records Inserted
```



3.8.2 Write SQL query to find which District in each state/ut has the highest number of murders yearwise. Your output should show STATE/UT, YEAR, DISTRICT, and MURDERS.

```
result=cursor.execute("SELECT State UT,DISTRICT, YEAR, MAX(MURDER) FROM crime_ipc GROUP BY DISTRICT ORDER BY UT ")
for row in result:
    print(row)
```

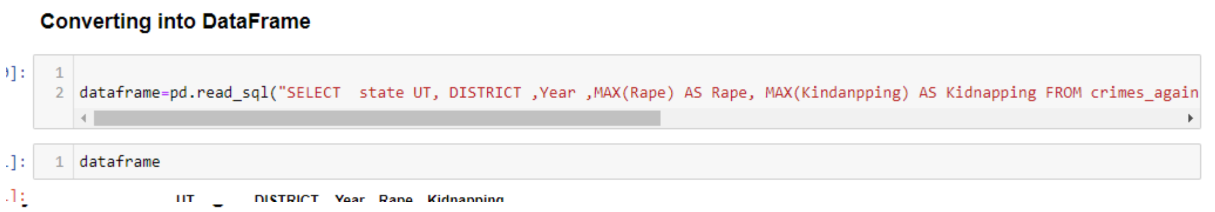
```
('A & N ISLANDS', 'A and N ISLANDS', 2007, 15)
('A & N ISLANDS', 'ANDAMAN', 2003, 16)
('A & N ISLANDS', 'CAR', 2012, 2)
('A & N ISLANDS', 'NICOBAR', 2003, 5)
('ANDHRA PRADESH', 'ADILABAD', 2004, 113)
('ANDHRA PRADESH', 'ANANTAPUR', 2006, 184)
('ANDHRA PRADESH', 'CHITTOOR', 2007, 129)
('ANDHRA PRADESH', 'CUDDAPAH', 2011, 120)
('ANDHRA PRADESH', 'CYBERABAD', 2011, 213)
('ANDHRA PRADESH', 'EAST GODAVARI', 2005, 110)
('ANDHRA PRADESH', 'GUNTAKAL RLY.', 2005, 12)
('ANDHRA PRADESH', 'GUNTUR', 2003, 210)
('ANDHRA PRADESH', 'GUNTUR URBAN', 2012, 56)
('ANDHRA PRADESH', 'HYDERABAD CITY', 2009, 153)
```



3.8.3 Store the above data (the result of 3.2) in DataFrame and analyze districts that appear 3 or more than 3 years and print the corresponding state/ut, district, murders, and year in descending order.

```
#.3.2 Write SQL query to find the highest number of rapes & Kidnappings that happened in which state, District, and year
data=cursor.execute("SELECT  state UT, DISTRICT ,Year ,MAX(Rape), MAX(Kindanpping) FROM crimes_against_women WHERE NOT DISTR
for row in data:
    print(row)
```

```
('WEST BENGAL', 'MURSHIDABAD', 2011, 433, 492)
('WEST BENGAL', 'MURSHIDABAD', 2012, 257, 464)
('WEST BENGAL', 'MURSHIDABAD', 2010, 526, 441)
('DELHI', 'NORTH-WEST', 2005, 236, 349)
('WEST BENGAL', 'MURSHIDABAD', 2009, 568, 342)
('UTTAR PRADESH', 'LUCKNOW', 2008, 334, 331)
('DELHI', 'NORTH WEST', 2001, 145, 298)
('DELHI', 'NORTH-WEST', 2006, 224, 287)
```

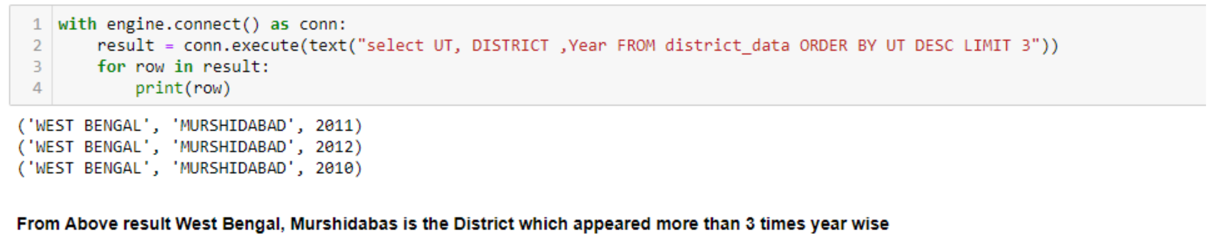


Converting into DataFrame

```
dataframe=pd.read_sql("SELECT  state UT, DISTRICT ,Year ,MAX(Rape) AS Rape, MAX(Kindanpping) AS Kidnapping FROM crimes_again
```

```
dataframe
```

```
             UT         DISTRICT  Year  Rape  Kidnapping
```

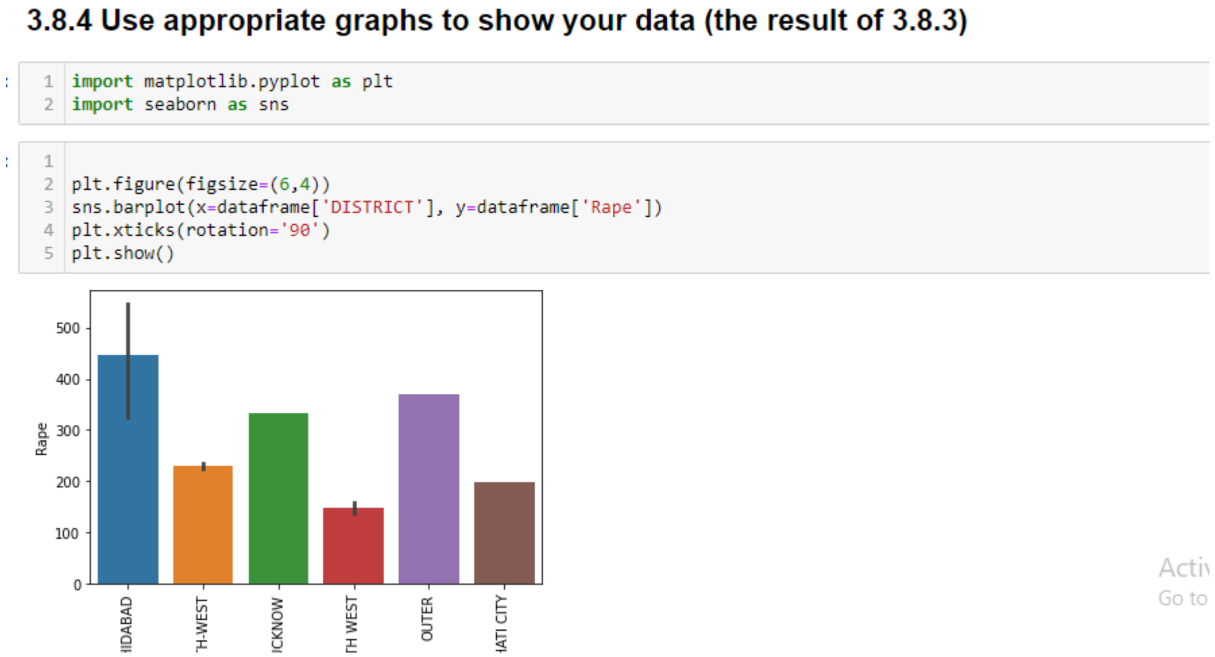


```
with engine.connect() as conn:
    result = conn.execute(text("select UT, DISTRICT ,Year FROM district_data ORDER BY UT DESC LIMIT 3"))
    for row in result:
        print(row)
```

```
('WEST BENGAL', 'MURSHIDABAD', 2011)
('WEST BENGAL', 'MURSHIDABAD', 2012)
('WEST BENGAL', 'MURSHIDABAD', 2010)
```

From Above result West Bengal, Murshidabas is the District which appeared more than 3 times year wise

Here I anlysed the district with he help of Barplot.



3.8.4 Use appropriate graphs to show your data (the result of 3.8.3)

```
import matplotlib.pyplot as plt
import seaborn as sns
```

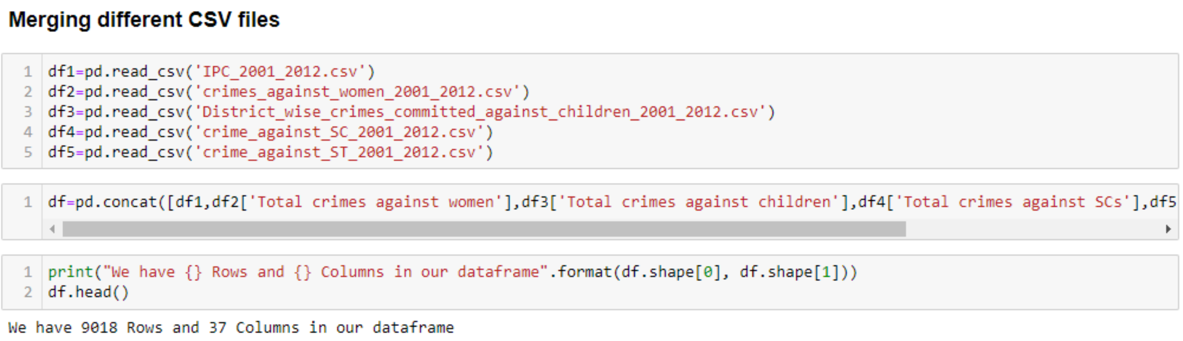
```

plt.figure(figsize=(6,4))
sns.barplot(x=dataframe['DISTRICT'], y=dataframe['Rape'])
plt.xticks(rotation='90')
plt.show()
```

Phase 4

Unsupervised ML (Clustering)

You were given various crime datasets that contains all the DISTRICTS in each state and you were asked to provide the below data to the higher authorities for further action.

- I created one dataframe from multiple csv files



Merging different CSV files

```
df1=pd.read_csv('IPC_2001_2012.csv')
df2=pd.read_csv('crimes_against_women_2001_2012.csv')
df3=pd.read_csv('District_wise_crimes_committed_against_children_2001_2012.csv')
df4=pd.read_csv('crime_against_SC_2001_2012.csv')
df5=pd.read_csv('crime_against_ST_2001_2012.csv')
```

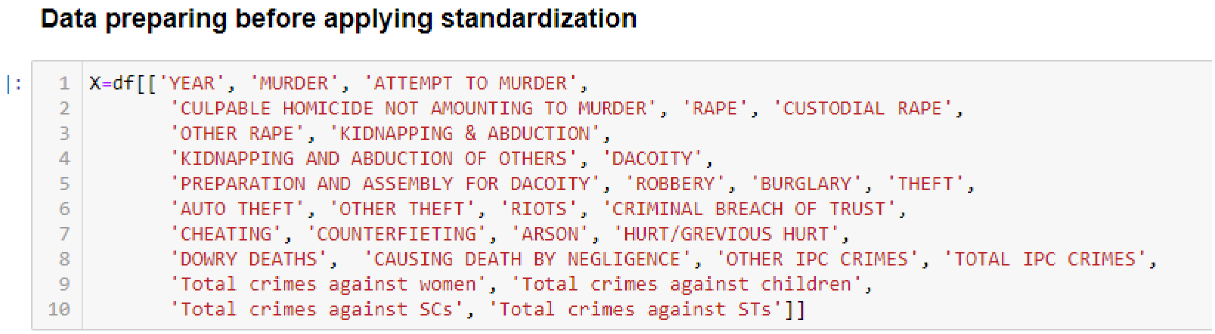
```
df=pd.concat([df1,df2['Total crimes against women'],df3['Total crimes against children'],df4['Total crimes against SCs'],df5
```

```
print("We have {} Rows and {} Columns in our dataframe".format(df.shape[0], df.shape[1]))
df.head()
```

```
We have 9018 Rows and 37 Columns in our dataframe
```

- Data Preprocessing-
  - Checked for inforamtion of the columns using .info () method then checked for null values as well.
  - Dropped few rows where District name was TOTAL or DELHI UT TOTAL as they were showing the whole sum of each crime head state wise which could have influenced the cluster so dropped
  - Dropped few columns 'INSULT TO MODESTY OF WOMEN','KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS','CRUELTY BY HUSBAND OR HIS RELATIVES','IMPORTATION OF GIRLS FROM FOREIGN COUNTRIES','ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY' columns from our dataset as it comes under Total crimes against women we already have that column in our dataset.

- Data Preparation before applying standardization



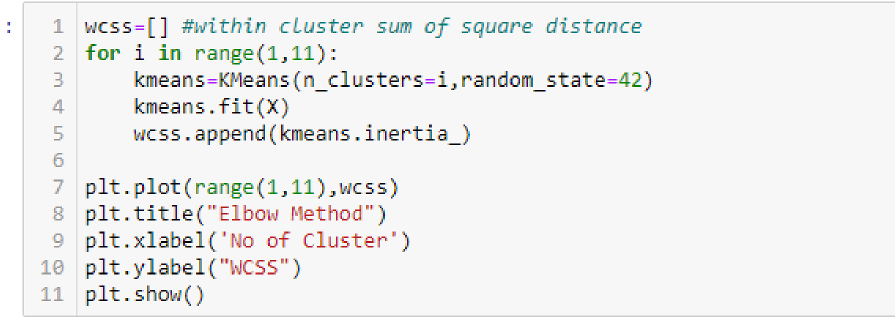
Data preparing before applying standardization

```
X=df[['YEAR', 'MURDER', 'ATTEMPT TO MURDER',
       'CULPABLE HOMICIDE NOT AMOUNTING TO MURDER', 'RAPE', 'CUSTODIAL RAPE',
       'OTHER RAPE', 'KIDNAPPING & ABDUCTION',
       'KIDNAPPING AND ABDUCTION OF OTHERS', 'DACOITY',
       'PREPARATION AND ASSEMBLY FOR DACOITY', 'ROBBERY', 'BURGLARY', 'THEFT',
       'AUTO THEFT', 'OTHER THEFT', 'RIOTS', 'CRIMINAL BREACH OF TRUST',
       'CHEATING', 'COUNTERFIETING', 'ARSON', 'HURT/GREVIOUS HURT',
       'DOWRY DEATHS',  'CAUSING DEATH BY NEGLIGENCE', 'OTHER IPC CRIMES', 'TOTAL IPC CRIMES',
       'Total crimes against women', 'Total crimes against children',
       'Total crimes against SCs', 'Total crimes against STs']]
```

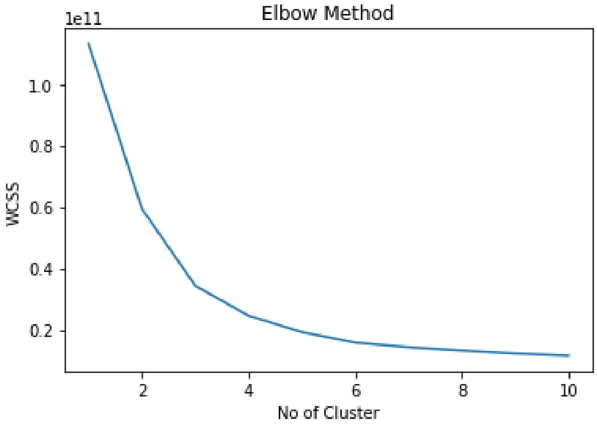
- Normalizing data for any type of clustering:

Normalizing the data is important to ensure that the distance measure accords equal weight to each variable. Without normalization, the variable with the largest scale will dominate the measure.Normalization is used to eliminate redundant data and ensures that good quality clusters are generated which can improve the efficiency of clustering algorithms.So it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences

- Used Elbow method for finding the value of k

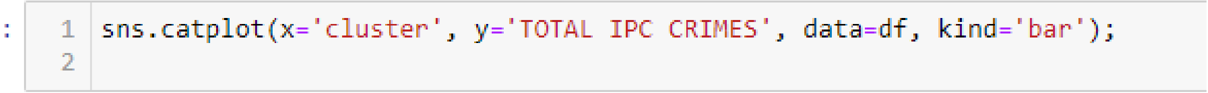


```
:
wcss=[] #within cluster sum of square distance
for i in range(1,11):
    kmeans=KMeans(n_clusters=i,random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

plt.plot(range(1,11),wcss)
plt.title("Elbow Method")
plt.xlabel('No of Cluster')
plt.ylabel("WCSS")
plt.show()
```

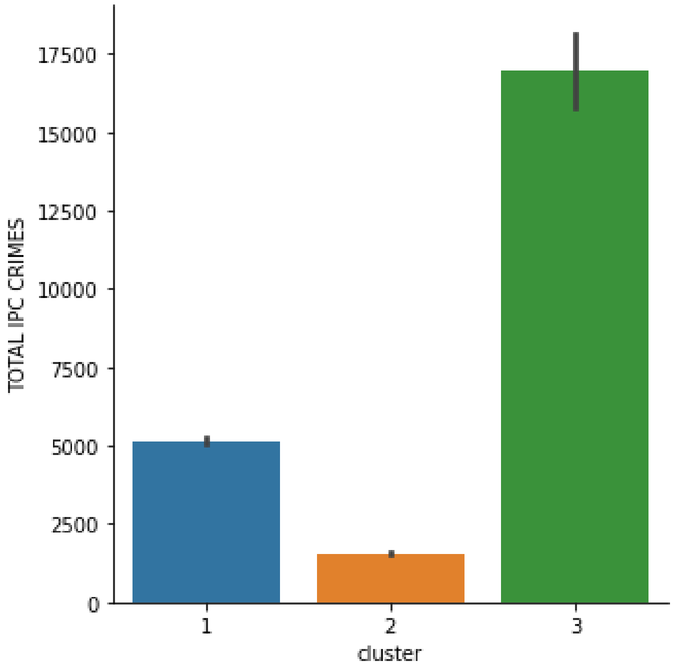


Choose k= 3 and made 3 cluster- and draw cat plot foe cluster and Draw Catplot and Scatter plot.

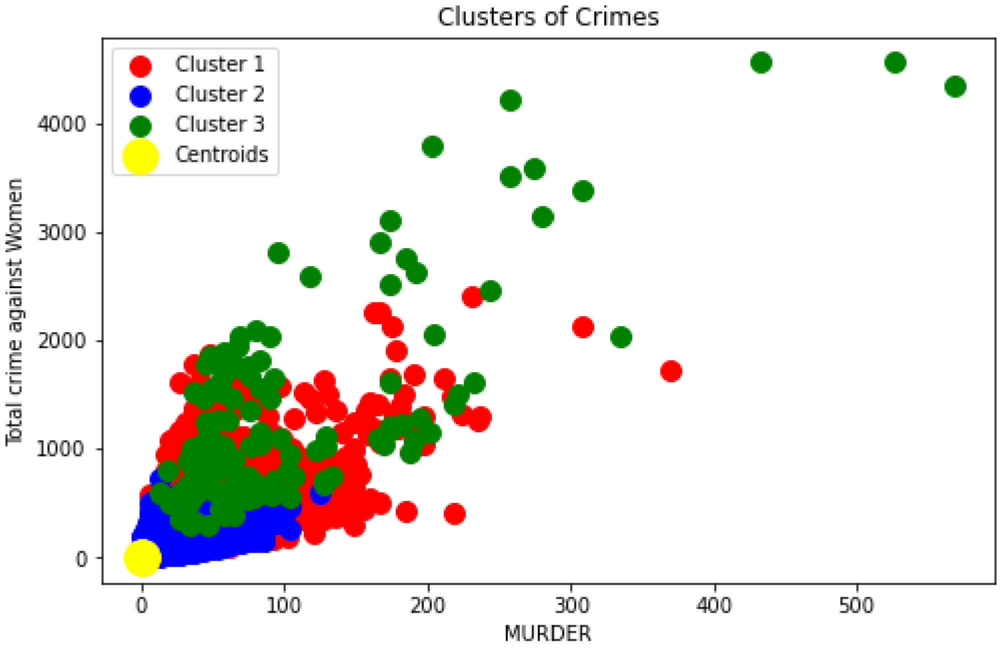
Code:



```
:
sns.catplot(x='cluster', y='TOTAL IPC CRIMES', data=df, kind='bar');
```



Also the scatter plot



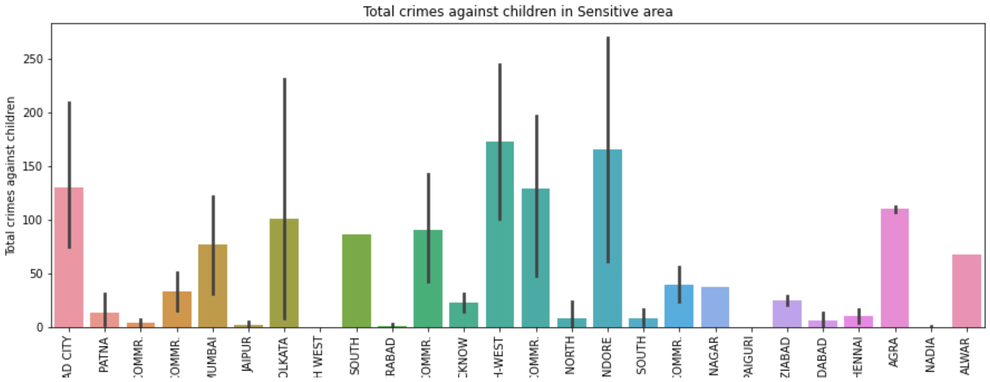
Analysed the data and assign the cluster name as given that is-

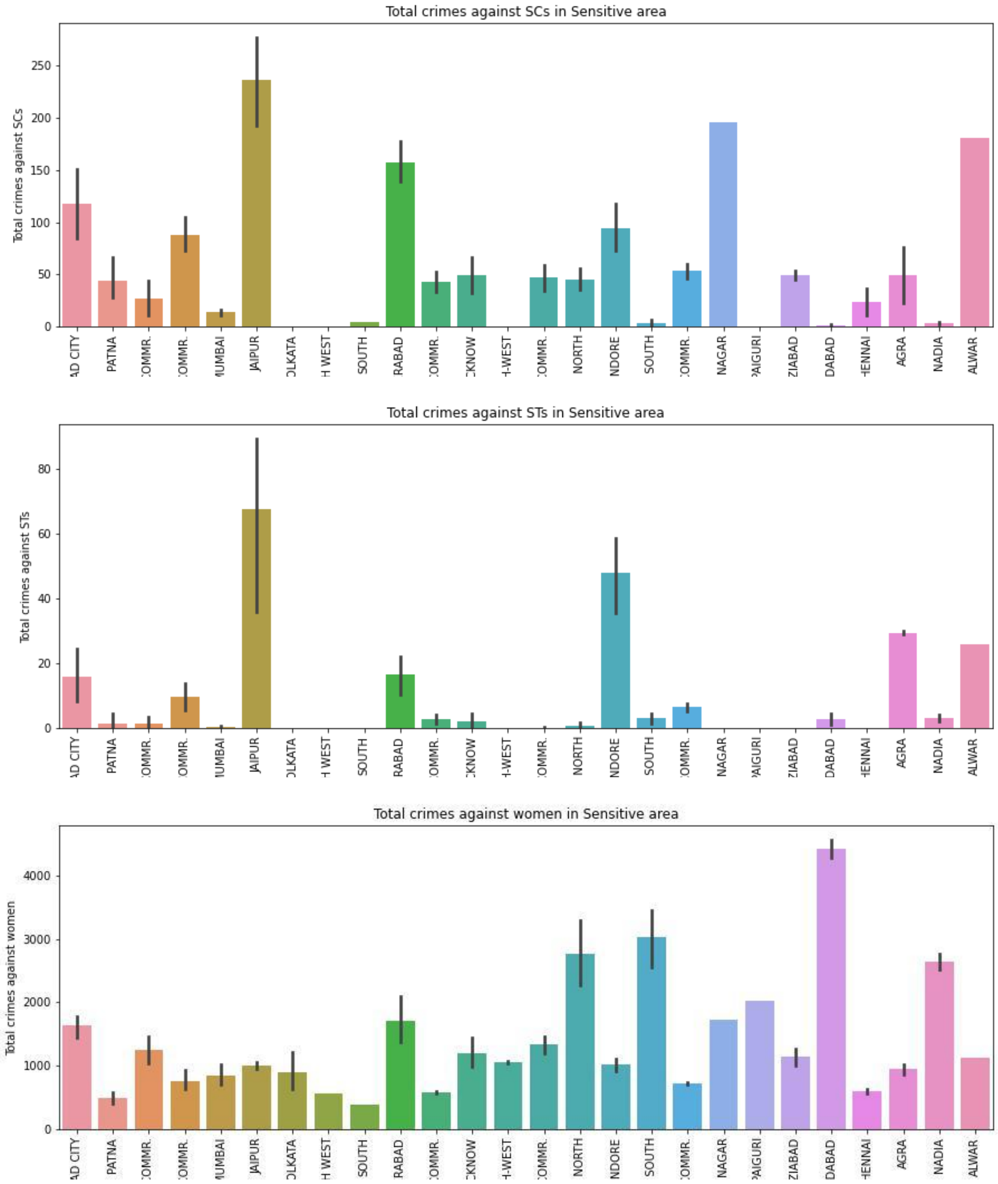
1. Sensitive Area's
2. Moderate Area's
3. Peaceful Area's"

Anslysed Each cluster and found the states with higher ,moderate and lower crime

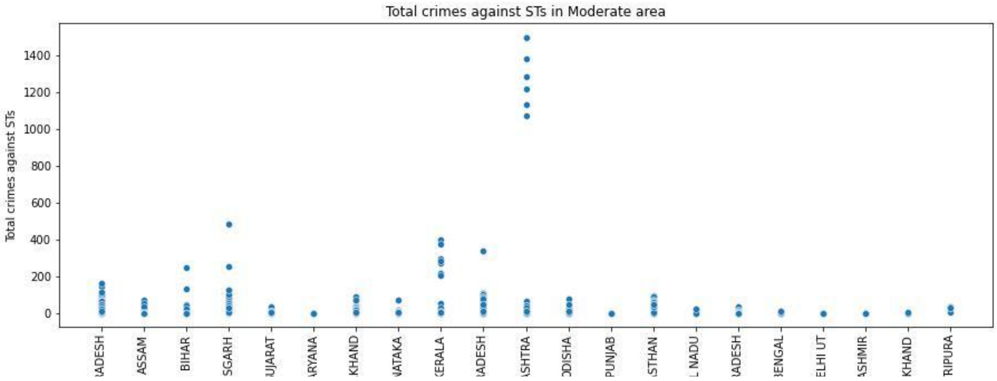
1-Sensitive Area's

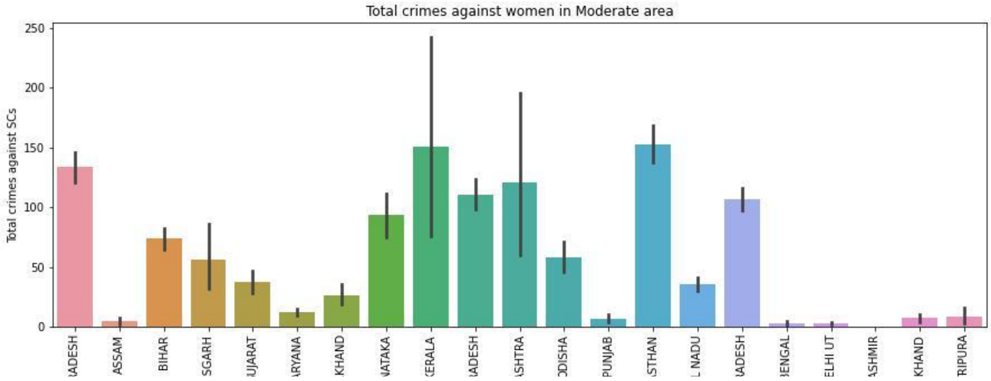
I generated count plots, bar plots, pair plots, heatmap and others to visualise the datapoint present in our column
records

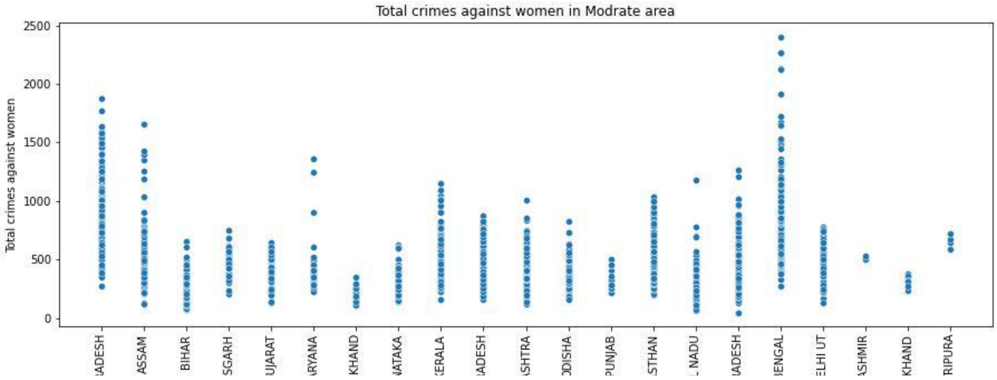


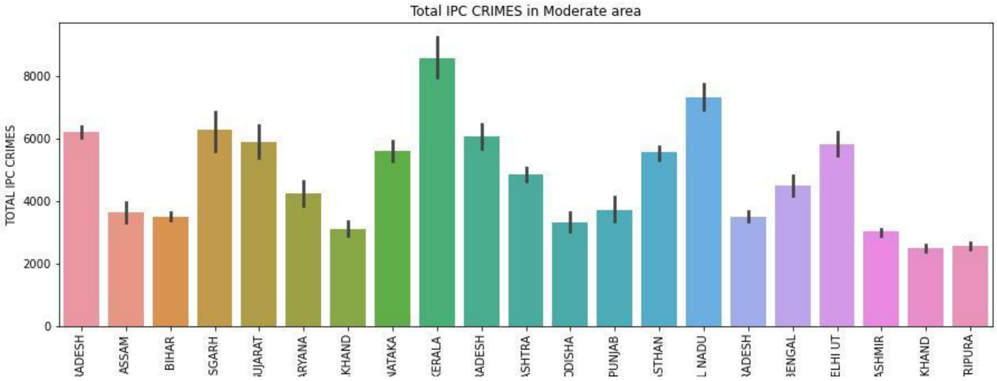


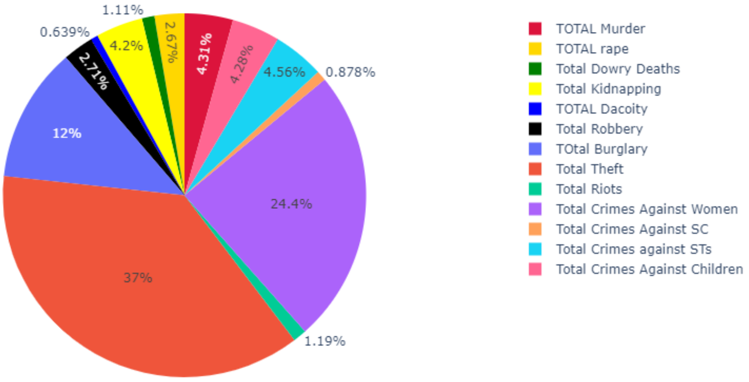
2-Moderate Area-


Total crimes against STs in Moderate area

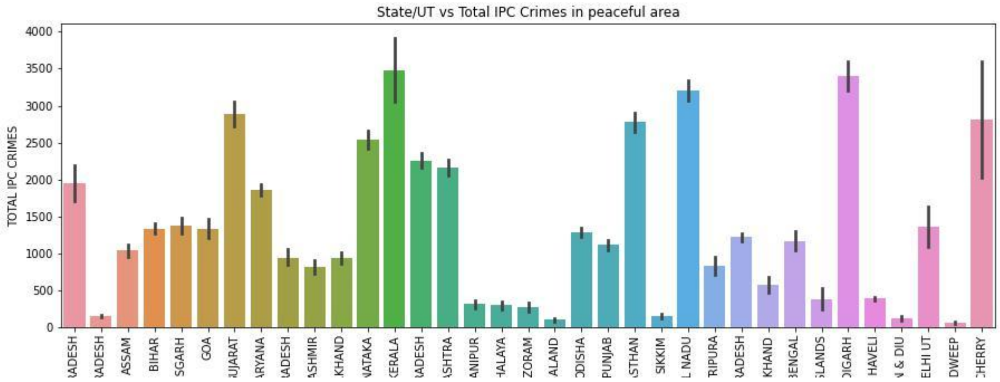

Total crimes against women in Moderate area

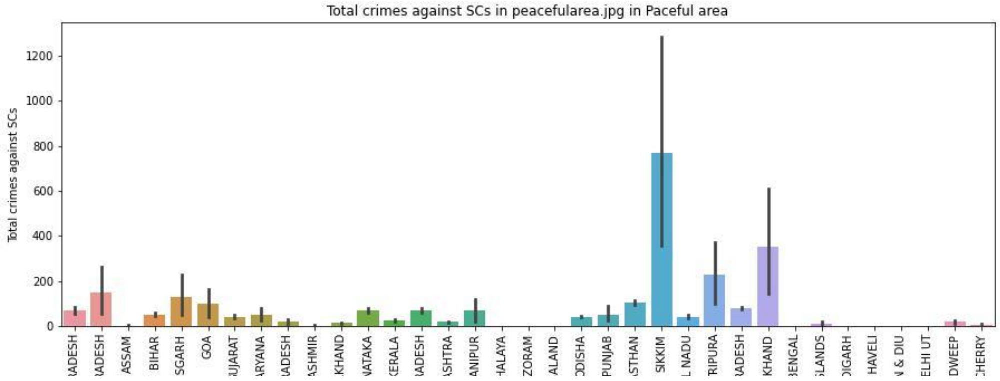

Total crimes against women in Modrate area

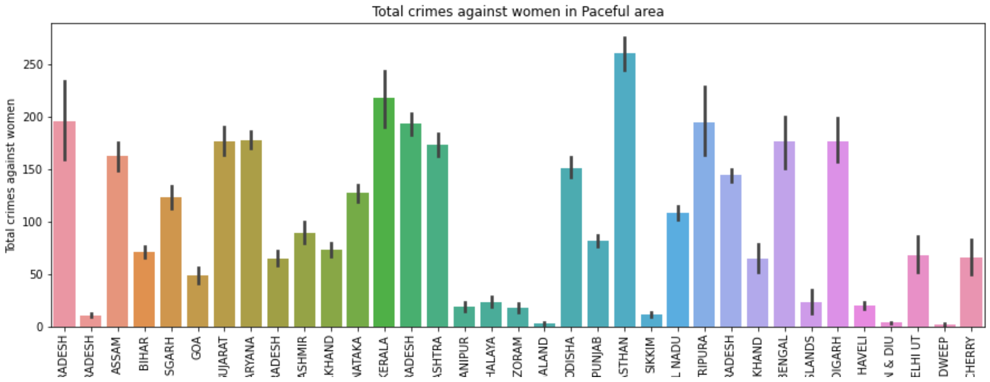

Total IPC CRIMES in Moderate area

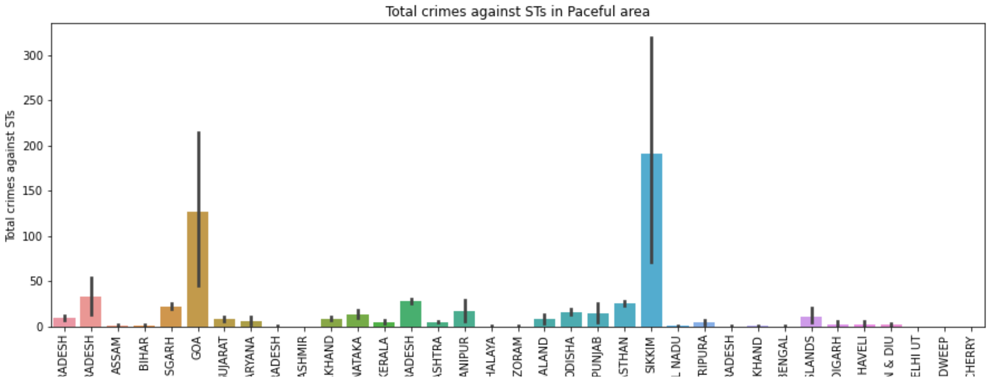


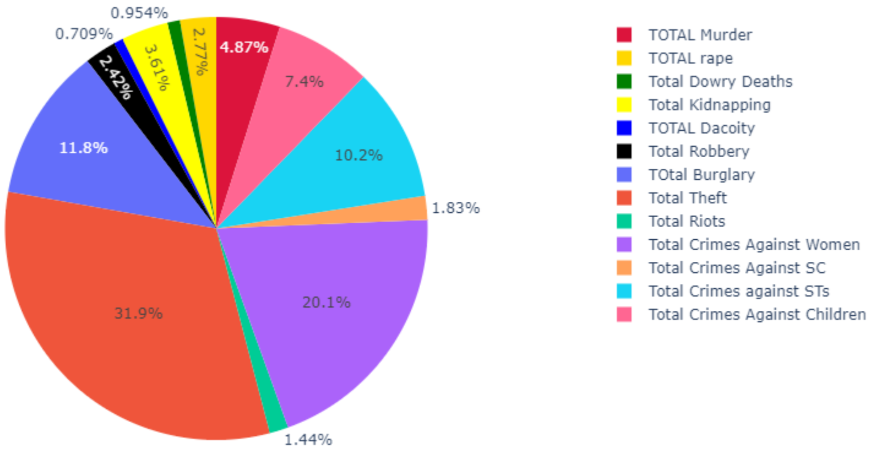
3-Peaceful Area-



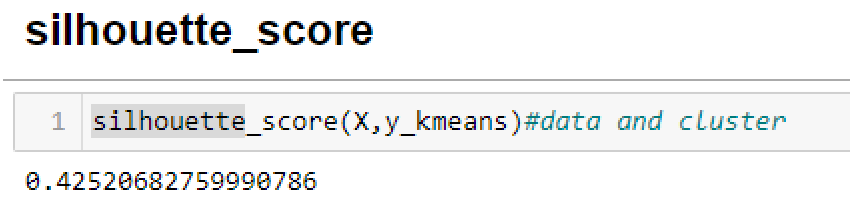








By using the different Barplots, scatter pots we analysed the clusters. Calculate the Score of the algorithm.



silhouette_score

```
silhouette_score(X,y_kmeans)#data and cluster
```

```
0.42520682759990786
```

CONCLUSION

- **Key Findings and Conclusions of the Study**

  After the completion of this project, we got an insight on how to collect data, pre-processing the data, analysing the data and building a model. First, we collected the used cars data from different websites like census, Kaggle, Wikipedia etc. We collected almost 420 of data which contained the population, literacy, area and other related features for states. Then the gathered data was combined in a single data frame and saved in a csv file so that we can open it and analyse the data. We did data cleaning, data pre-processing steps like finding and handling null values, removing words from numbers. Then we started EDA of the csv file that we created in first phase .Sql operations were done on agiven file and then phase 4 comes with the merging of different csv file and make a single data frame and Make 3 cluster and then analyse them and write down the observations.

  Learning Outcomes of the Study in respect of Data Science

  Visualization part helped me to understand the data as it provides graphical representation of huge data. It assisted me to understand the data more clearly..

- **Limitations of this work and Scope for Future Work**

  The limitations I faced during this project were:

  Data for Population , literacy, urban population, rural population were given for 2001 then data were available on web for 2011 so I calculated growth rate and applied that growth rate to find the population,rural populatin, literacy, male /female literacy , literacy rate for 2002 to 2010 which took a lot of time because 35 states were there and different columns were also there so it took approx 3-5 days for whole data collection part.

Future Work Scope:

Current algorithm is limited for the duration of 2001 to 2012 .