

In [1]:

```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
```

In [35]:

```
df=pd.read_excel(r'DataFinal17-20.xlsx',sheet_name='2020')
df2=df.iloc[:,2:16]
data=df.iloc[:,1:23]
data=data.drop(['Total','Crime Rate','Sex Ratio','Literacy','Density','Latitude','Longitude'])
df2=df2.drop(['Total'], axis=1)
df2.head()
from sklearn import preprocessing
df_standardized = preprocessing.scale( df2 )
df1 = pd.DataFrame( df_standardized )
```

In [36]:

```
data.head()
```

Out[36]:

	District	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	Offences against property
0	Ahmednagar	254	620	1826	543	369	0	92	905	3104
1	Akola	114	116	1345	245	76	0	28	111	946
2	Amravati	201	354	2133	461	190	0	97	144	2198
3	Aurangabad	139	386	1869	335	242	0	80	428	3821
4	Beed	206	308	1233	248	91	0	50	508	1002

In [37]:

```
#Elbow Method
from sklearn.cluster import KMeans

#create a list for the wcss parameter
wcss = []
#test with 14 clusters
for i in range(1, 15):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 0)
    kmeans.fit(df1)
    wcss.append(kmeans.inertia_)
```

In [38]:

```
wcss
```

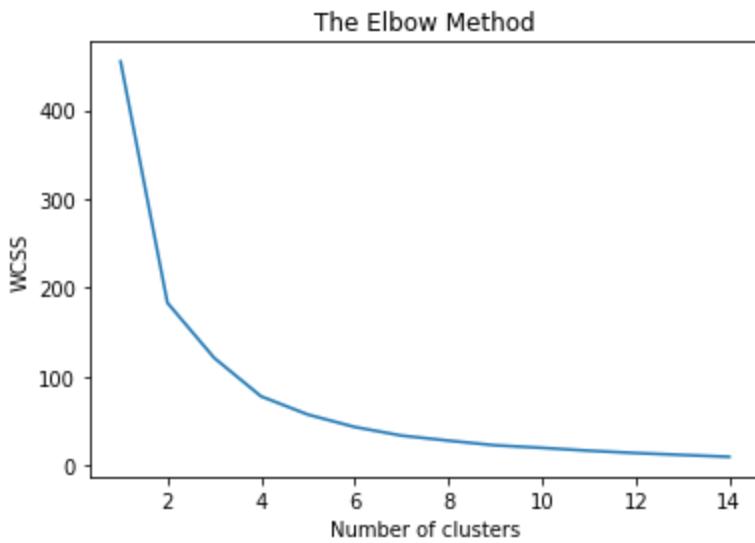
Out[38]:

```
[455.0,
 183.15163902707062,
 121.00520284639016,
 77.72527132804295,
 57.19196898334106,
 43.20454952915056,
 33.50410783186294,
 27.826870740722036,
```

```
22.576706877494622,  
19.703831936011063,  
16.477641547770464,  
13.851069756261888,  
11.696519258289692,  
9.585075372115798]
```

In [39]:

```
plt.plot(range(1, 15), wcss)  
plt.title('The Elbow Method')  
plt.xlabel('Number of clusters')  
plt.ylabel('WCSS')  
plt.show()
```



In [40]:

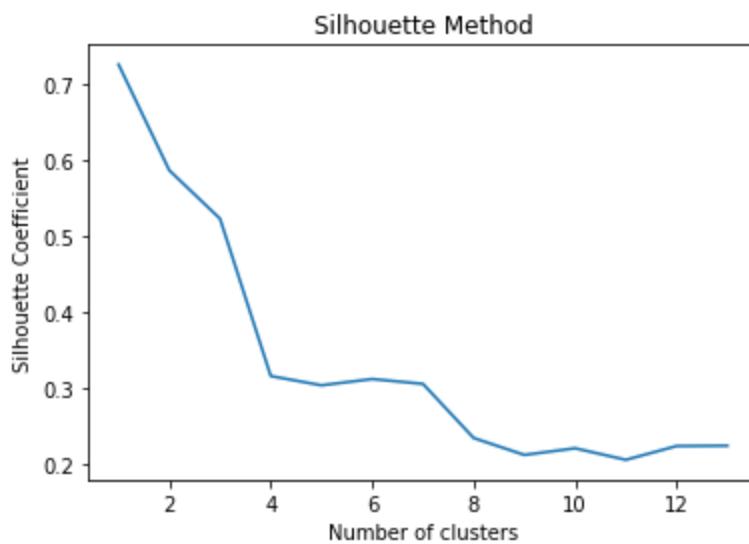
```
import seaborn as sns
```

In [41]:

```
#Silhouette Method  
from sklearn.metrics import silhouette_score  
  
sil = []  
kmax = 14  
  
# dissimilarity would not be defined for a single cluster, thus, minimum number of clusters  
for k in range(2, kmax+1):  
    kmeans = KMeans(n_clusters = k).fit(df1)  
    labels = kmeans.labels_  
    sil.append(silhouette_score(df1, labels, metric = 'euclidean'))
```

In [42]:

```
plt.plot(range(1, 14), sil)  
plt.title('Silhouette Method')  
plt.xlabel('Number of clusters')  
plt.ylabel('Silhouette Coefficient')  
plt.show()
```



```
In [43]: #K-Means Clustering using k=6
km=KMeans(n_clusters=6)
y_pred=km.fit_predict(df1)
```

```
In [44]: y_pred
```

```
Out[44]: array([2, 0, 4, 4, 4, 0, 4, 0, 0, 0, 0, 0, 0, 4, 4, 4, 4, 1, 2, 4, 0, 2, 0,
   0, 0, 3, 0, 0, 4, 4, 0, 2, 5, 0, 0, 4])
```

```
In [45]: df['cluster']=y_pred
df.head()
```

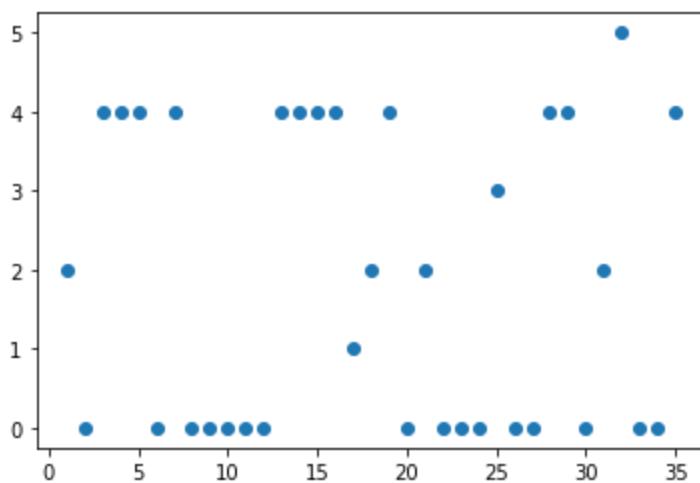
```
Out[45]:
```

	SrNo	District	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	...
0	1	Ahmednagar	254	620	1826	543	369	0	92	905	...
1	2	Akola	114	116	1345	245	76	0	28	111	...
2	3	Amravati	201	354	2133	461	190	0	97	144	...
3	4	Aurangbad	139	386	1869	335	242	0	80	428	...
4	5	Beed	206	308	1233	248	91	0	50	508	...

5 rows × 23 columns

```
In [46]: plt.scatter(df['SrNo'],df['cluster'])
#for col in df.columns:
#    print(col)
```

```
Out[46]: <matplotlib.collections.PathCollection at 0x28185d83c10>
```



In [47]:

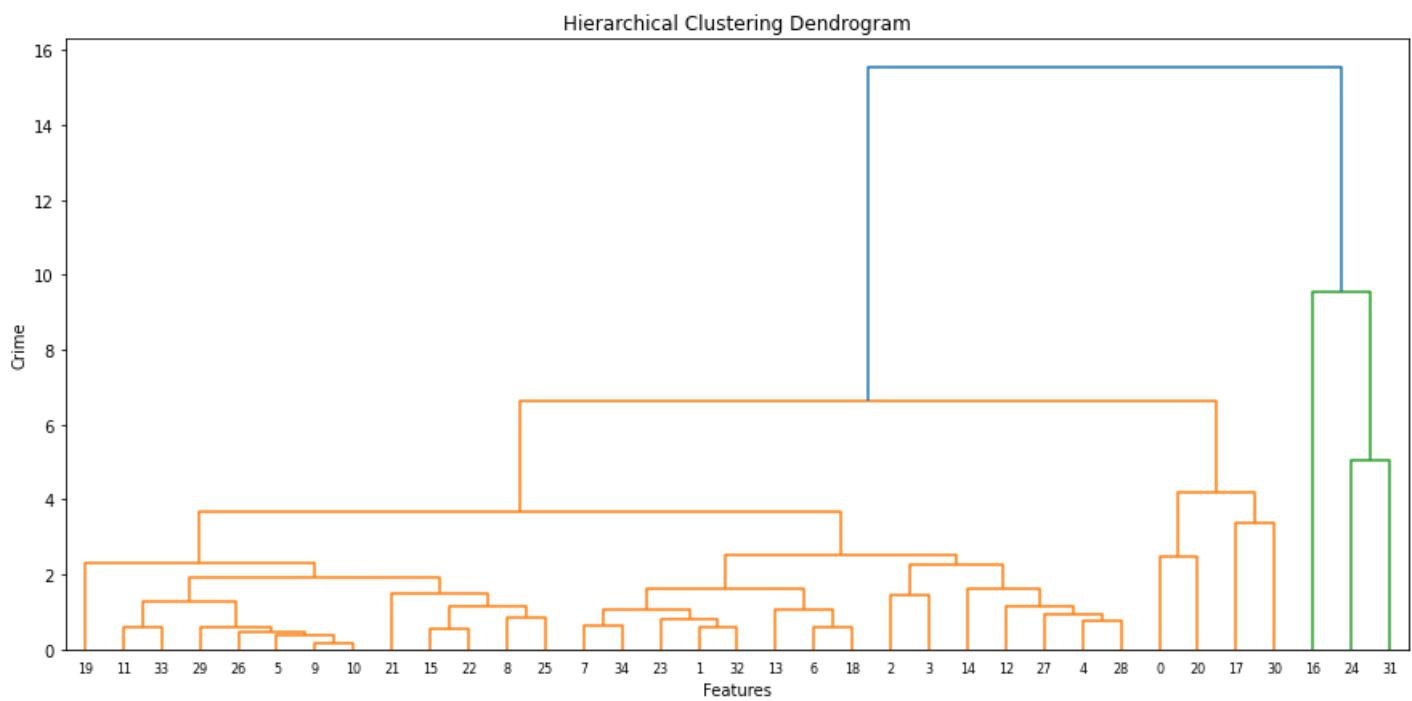
```
#Hierarchical Clustering
from scipy.cluster.hierarchy import linkage
import scipy.cluster.hierarchy as sch # for creating dendrogram
```

In [48]:

```
z = linkage(df1, method="complete", metric="euclidean")
```

In [49]:

```
plt.figure(figsize=(15,7))
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Features')
plt.ylabel('Crime')
sch.dendrogram(z,
    leaf_rotation=0., # rotates the x axis labels
    leaf_font_size=8., # font size for the x axis labels
)
plt.show()
```



In [50]:

```
#df = pd.read_excel (r'Path where the Excel file is stored\File name.xlsx', sheet_name='yo
```

In [51]:

```
X = df2[['Homicide/Murder','Causing death by negligence','Hurt','Assault on woman','Kidnap
clusters = KMeans(6) # 6 clusters
```

```

clusters.fit( X )
clusters.cluster_centers_
clusters.labels_
df2['Crime_clusters'] = clusters.labels_
df2.head()
df2.sort_values(by=['Crime_clusters'], ascending = True)
X.head()

```

Out[51]:

	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	Offences against property	Offences relating to documents and property marks
0	254	620	1826	543	369	0	92	905	3104	220
1	114	116	1345	245	76	0	28	111	946	86
2	201	354	2133	461	190	0	97	144	2198	250
3	139	386	1869	335	242	0	80	428	3821	265
4	206	308	1233	248	91	0	50	508	1002	104

In [52]:

```

stats = df2.sort_values("Hurt", ascending=True)
stats

```

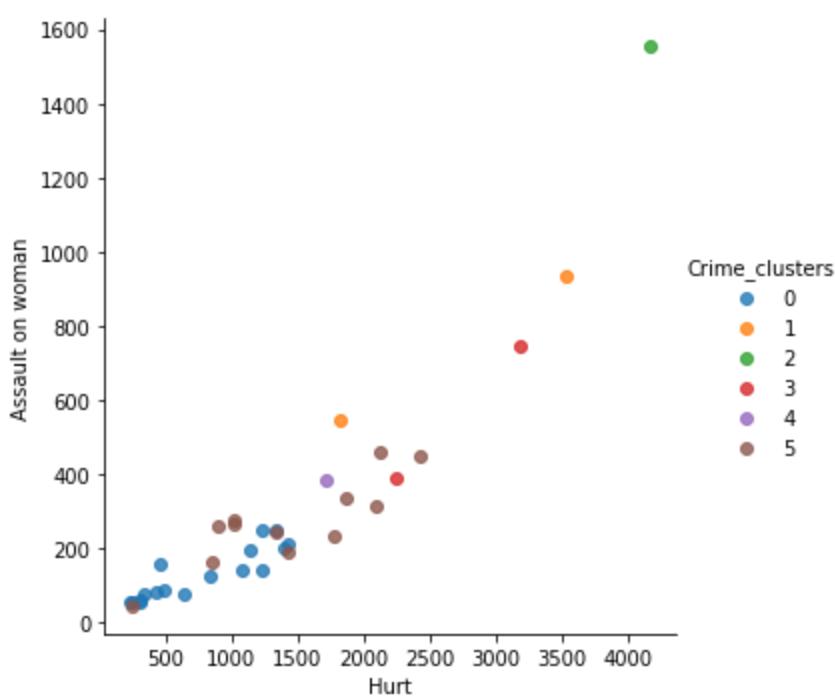
Out[52]:

	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	Offences against property	Offences relating to documents and property marks
29	18	50	241	53	18	0	6	43	178	48
26	39	106	250	53	43	0	11	109	361	113
19	62	133	255	44	57	0	17	116	429	38
9	65	135	303	55	32	0	17	21	302	41
10	50	131	311	57	40	0	18	23	400	30
25	47	191	334	77	73	0	20	165	533	141
8	84	352	433	80	98	1	14	205	781	77
21	79	232	466	158	217	4	62	95	1859	500
5	41	141	486	88	71	0	16	43	676	60
11	75	119	649	73	27	0	15	140	436	35
22	92	253	847	123	87	0	21	247	685	80
33	114	110	855	163	66	0	24	142	449	39
12	192	439	901	259	187	0	34	440	2045	248
28	145	359	1017	275	192	2	49	481	1557	183
14	146	313	1020	263	150	15	54	388	1357	222
15	104	244	1077	138	111	3	26	303	964	94
7	105	234	1147	194	152	1	46	84	980	108

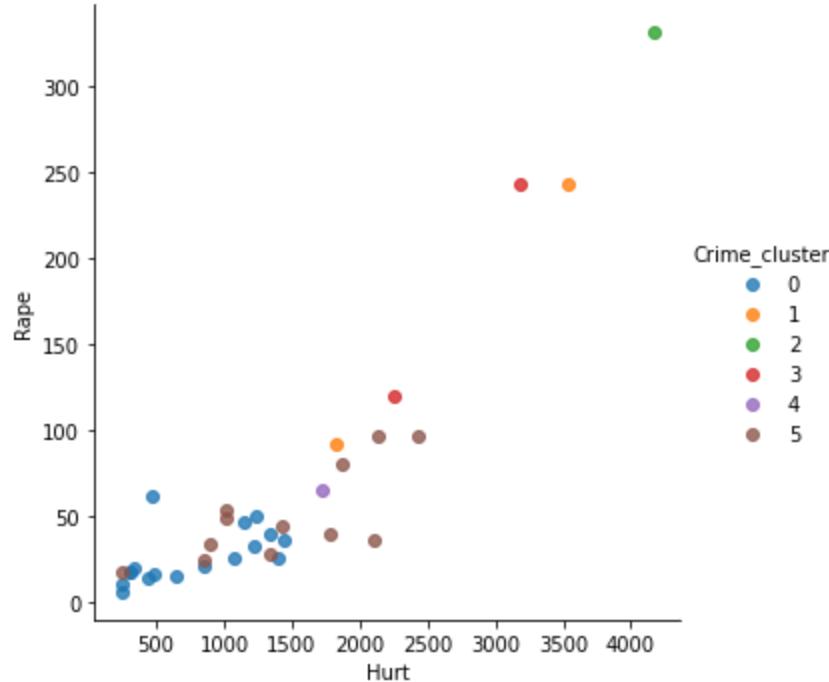
	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	Offences against property	Offences relating to documents and property marks
23	165	121	1227	139	65	0	33	119	924	94
4	206	308	1233	248	91	0	50	508	1002	104
27	207	269	1344	250	228	2	39	340	1678	197
1	114	116	1345	245	76	0	28	111	946	86
32	86	170	1394	197	113	0	25	44	741	88
34	140	310	1431	187	128	1	44	85	923	56
13	133	209	1436	208	100	0	36	411	1618	170
20	240	917	1720	384	473	1	65	637	2528	529
18	143	252	1778	232	127	1	40	307	1464	114
0	254	620	1826	543	369	0	92	905	3104	220
3	139	386	1869	335	242	0	80	428	3821	265
6	122	293	2099	313	110	0	36	320	1269	102
2	201	354	2133	461	190	0	97	144	2198	250
17	299	581	2252	386	442	13	120	199	4980	568
30	246	461	2437	449	276	8	96	786	3046	289
31	336	814	3187	746	1159	30	243	687	6209	1211
24	649	1102	3540	936	1113	28	243	1227	7935	999
16	521	349	4171	1553	1180	58	331	392	17530	3425

In [20]:

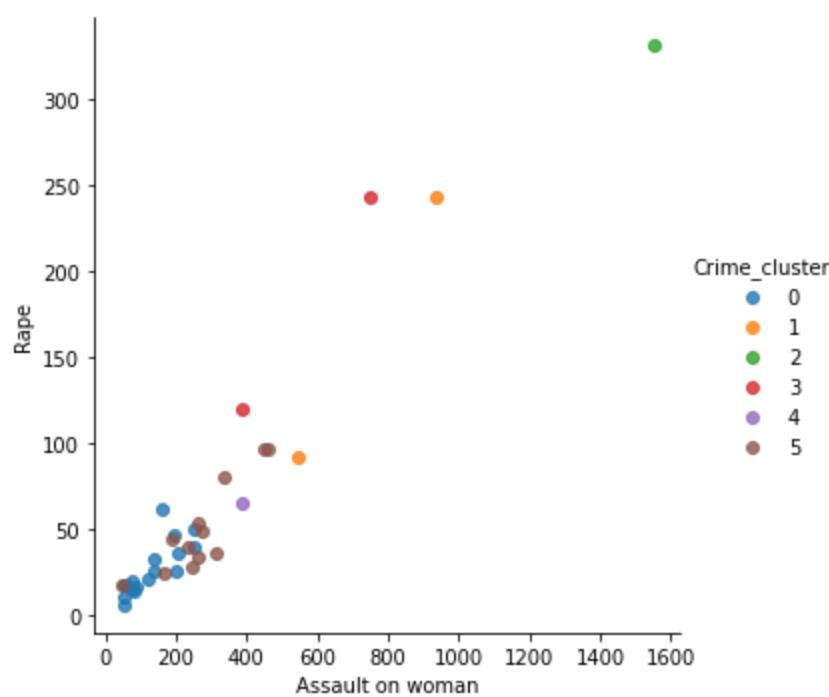
```
sns.lmplot(x='Hurt', y='Assault on woman', data=df2, hue = 'Crime_clusters', fit_reg=False);
```



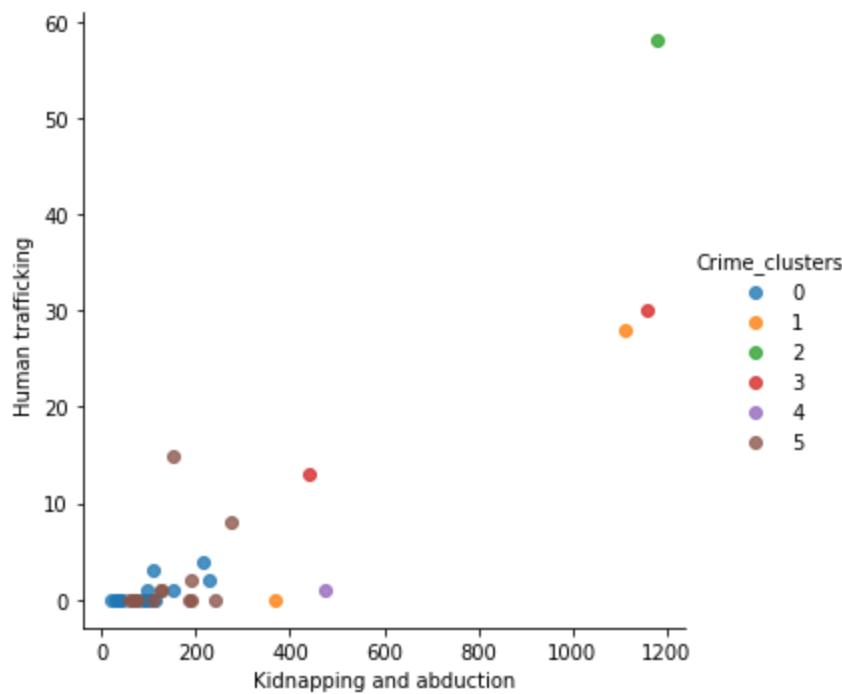
```
In [21]: sns.lmplot(x='Hurt', y='Rape', data=df2, hue = 'Crime_clusters', fit_reg=False);
```



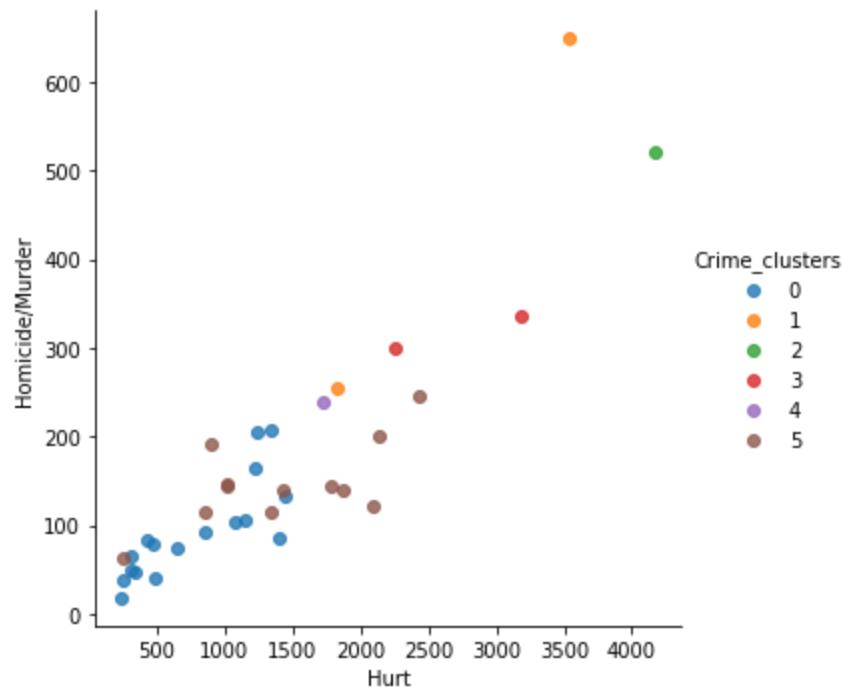
```
In [22]: sns.lmplot(x='Assault on woman', y='Rape', data=df2, hue = 'Crime_clusters', fit_reg=False);
```



```
In [23]: sns.lmplot(x='Kidnapping and abduction', y='Human trafficking ', data=df2, hue = 'Crime_clust
```

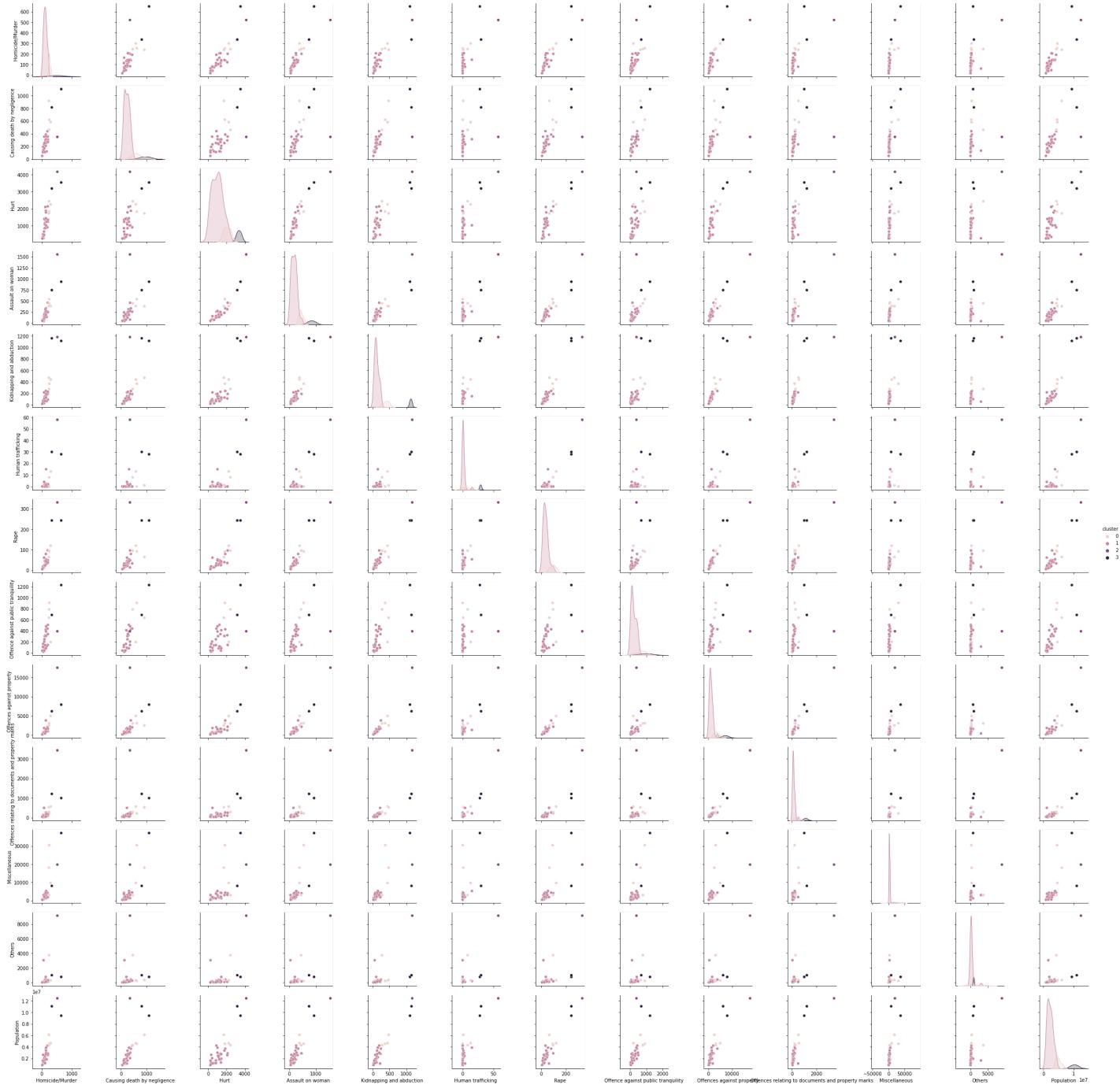


```
In [24]: sns.lmplot(x='Hurt', y='Homicide/Murder', data=df2, hue = 'Crime_clusters', fit_reg=False);
```



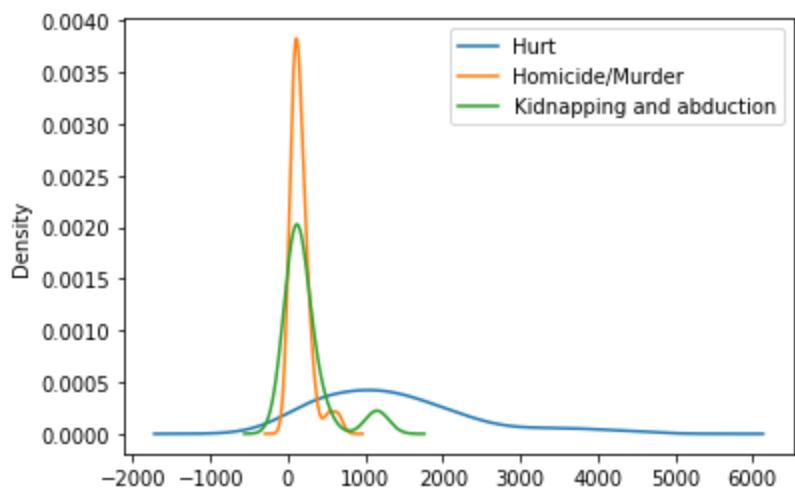
```
In [53]: sns.pairplot(data, hue='cluster')
```

```
Out[53]: <seaborn.axisgrid.PairGrid at 0x28183e2bb80>
```



```
In [26]: df3 = pd.DataFrame(data=df2, columns=['Hurt', 'Homicide/Murder', 'Kidnapping and abduction'])
df3.plot.kde()
#kernel density estimate (KDE) plot is a method for visualizing the distribution of observ
```

```
Out[26]: <AxesSubplot:ylabel='Density'>
```



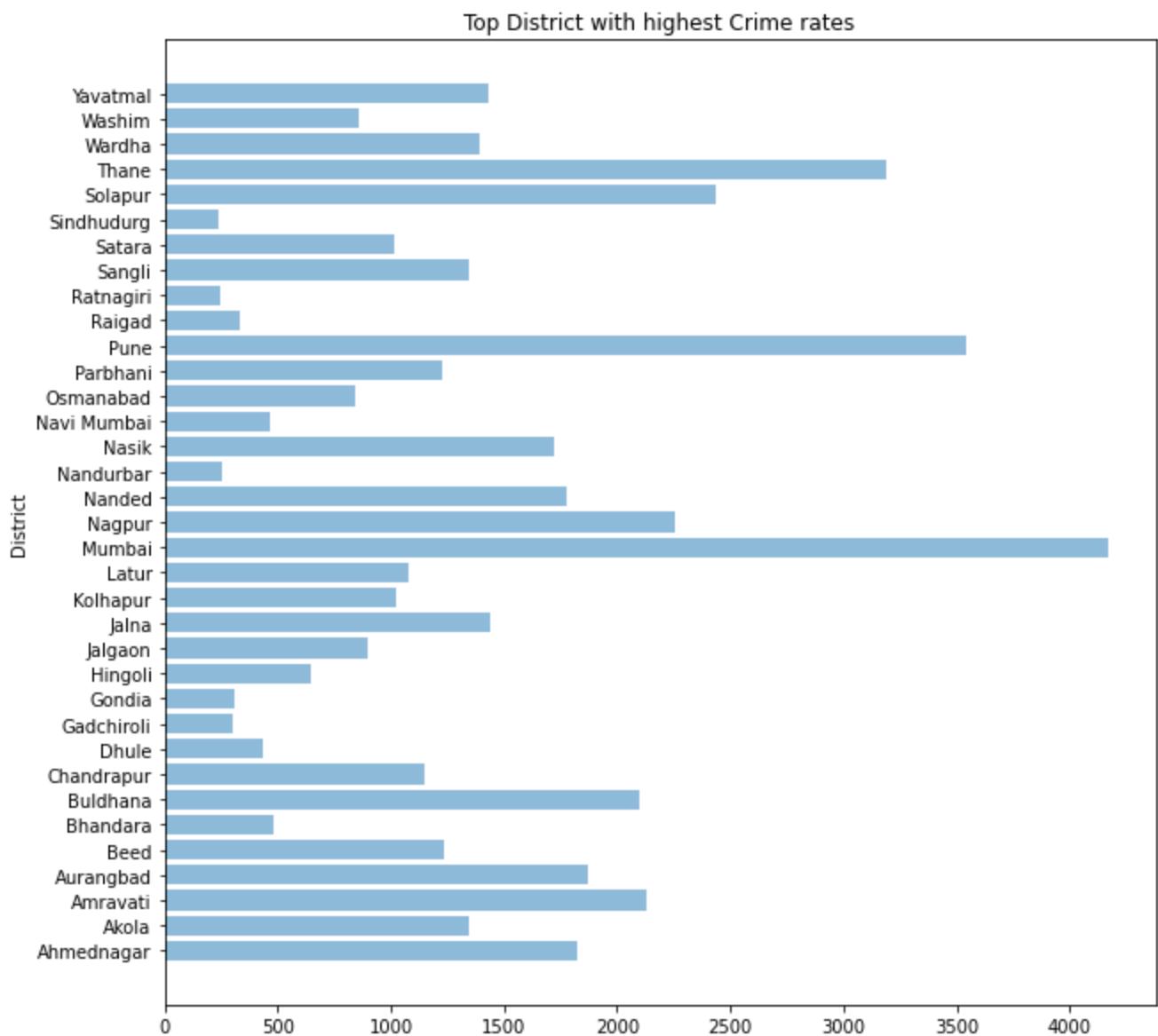
In [27]: `df2.head()`

Out[27]:

	Homicide/Murder	Causing death by negligence	Hurt	Assault on woman	Kidnapping and abduction	Human trafficking	Rape	Offence against public tranquility	Offences against property	Offences relating to documents and property marks
0	254	620	1826	543	369	0	92	905	3104	220
1	114	116	1345	245	76	0	28	111	946	86
2	201	354	2133	461	190	0	97	144	2198	250
3	139	386	1869	335	242	0	80	428	3821	265
4	206	308	1233	248	91	0	50	508	1002	104

In [29]:

```
a1=data['Hurt']
b1=data['District']
f = plt.figure()
f.set_figwidth(10)
f.set_figheight(10)
plt.barh(b1, a1, align='center', alpha=0.5)
plt.ylabel('District')
plt.title('Top District with highest Crime rates')
plt.show()
```



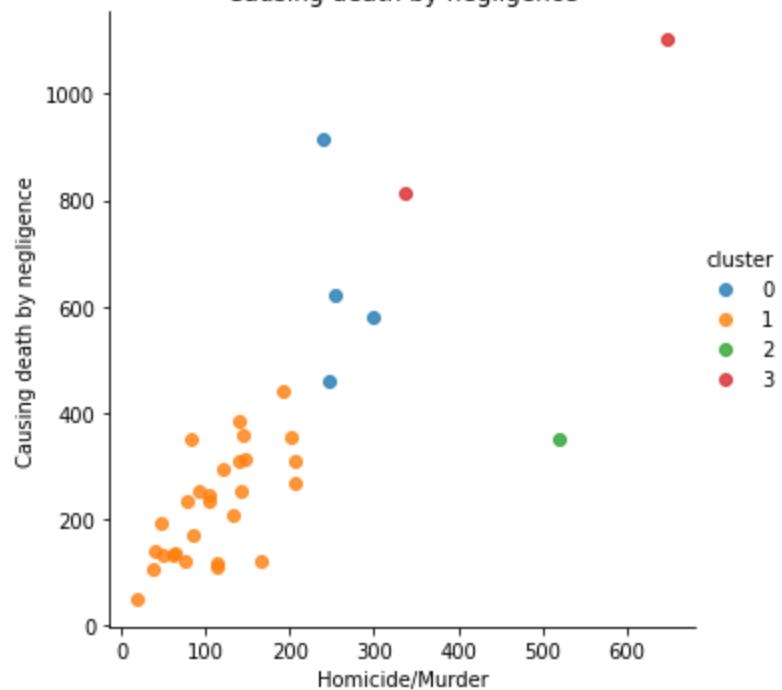
In [30]:

```
#Cluster plotting of crime vs crime/diff metrics
df=pd.read_excel(r'DataFinal17-20.xlsx',sheet_name='2020')
df2=df.iloc[:,2:23]
for i in range(len(df2.columns)-1):
    titles=df2.columns[i]+" vs \n"+df2.columns[i+1]
    plt.figure(i)
    sns.lmplot(x=df2.columns[i],y=df2.columns[i+1],data=df2,hue="cluster",legend=1,fit_regr
```

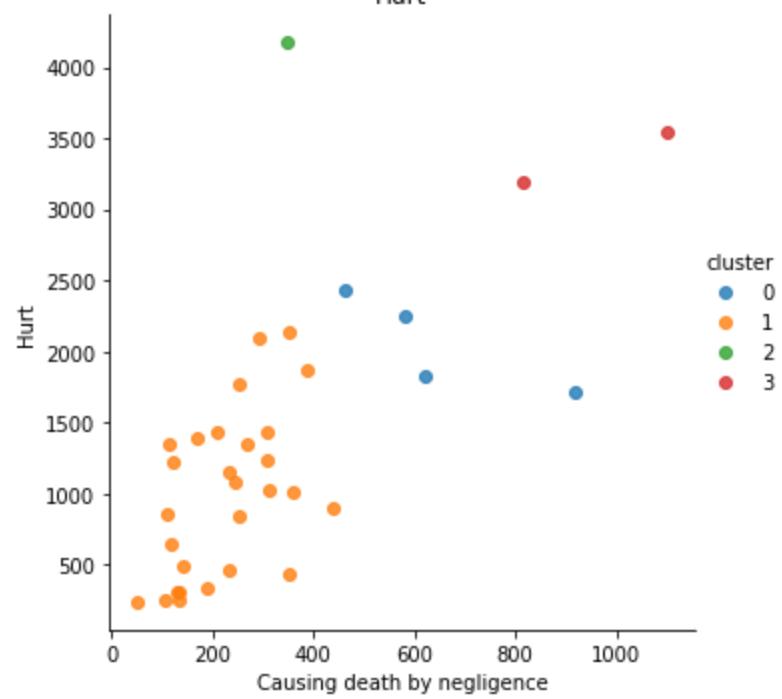
c:\users\welcome\appdata\local\programs\python\python39\lib\site-packages\seaborn\axisgrid.py:447: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplot interface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much memory. (To control this warning, see the rcParam `figure.max\_open\_warning`).

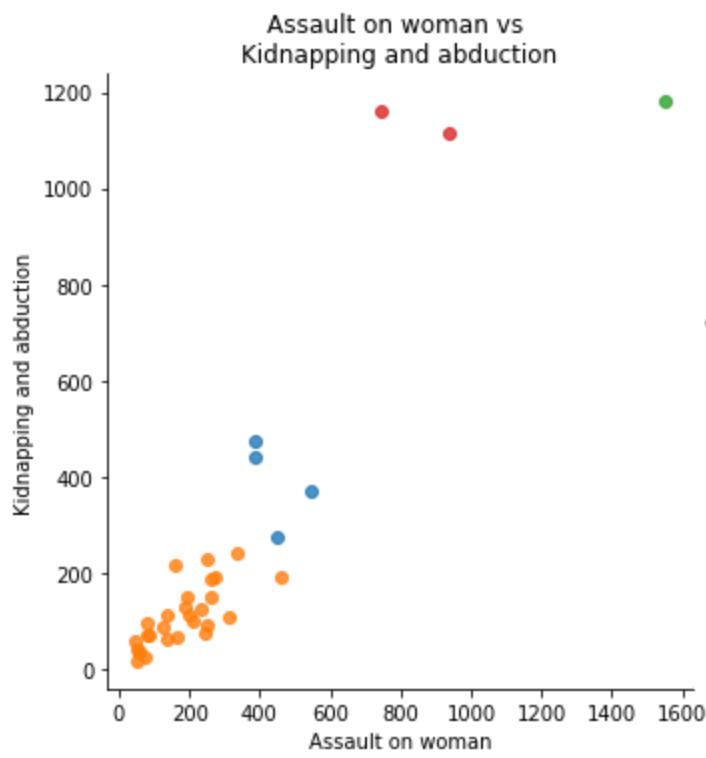
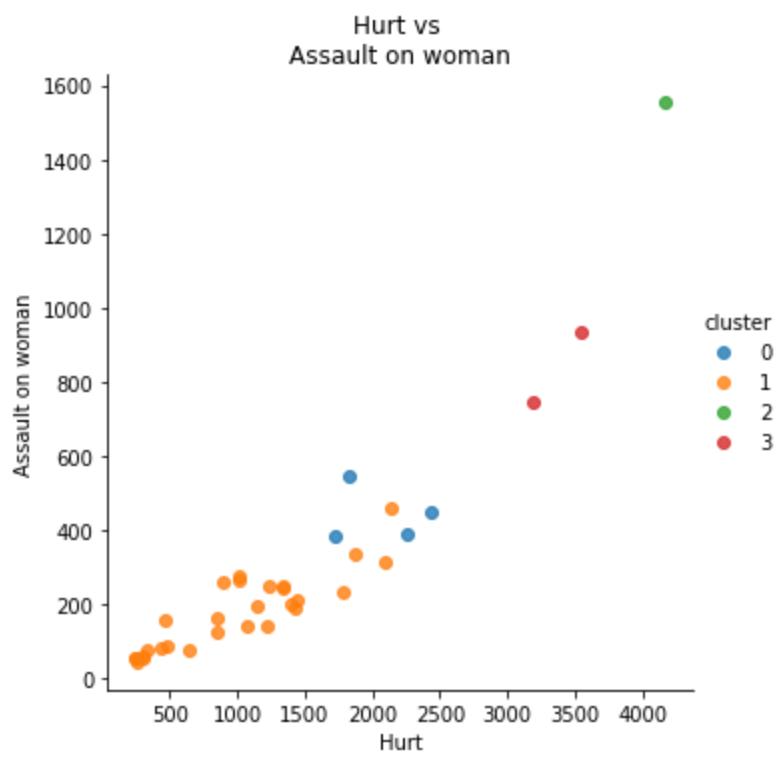
```
fig = plt.figure(figsize=figsize)
<Figure size 432x288 with 0 Axes>
```

Homicide/Murder vs  
Causing death by negligence

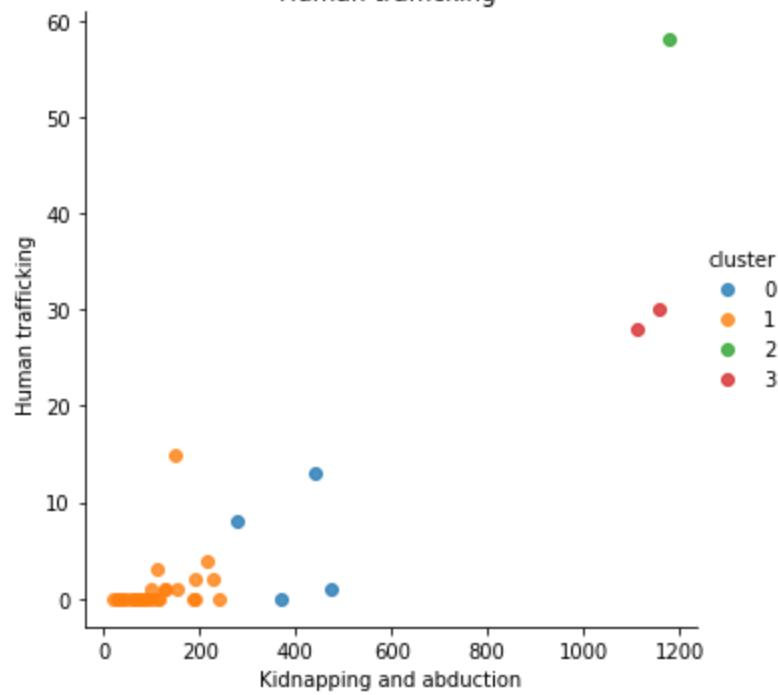


Causing death by negligence vs  
Hurt

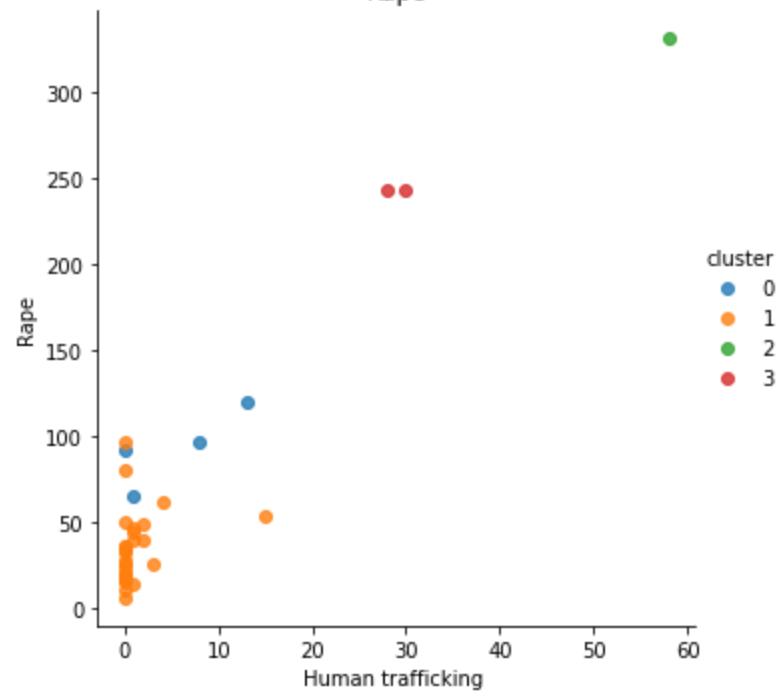


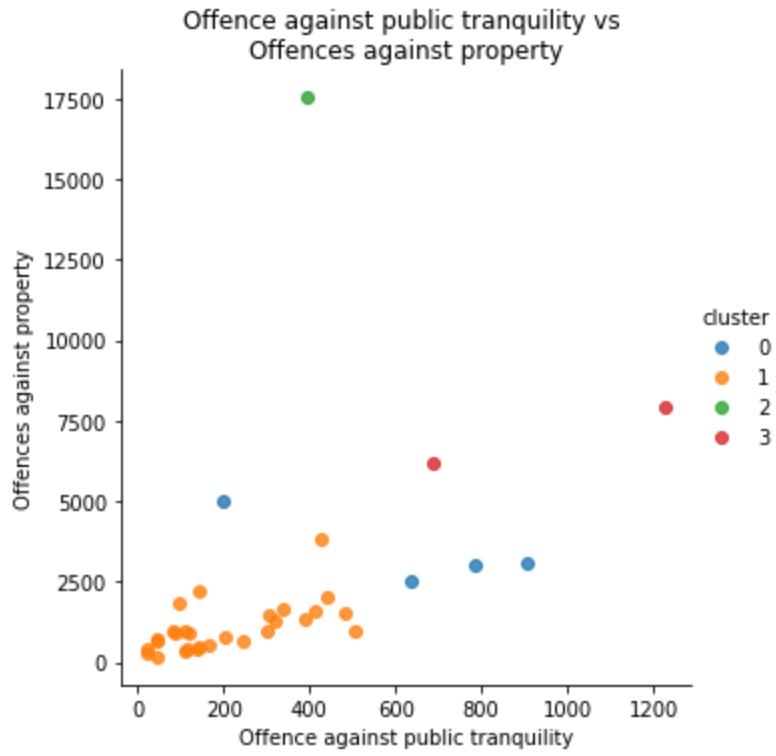
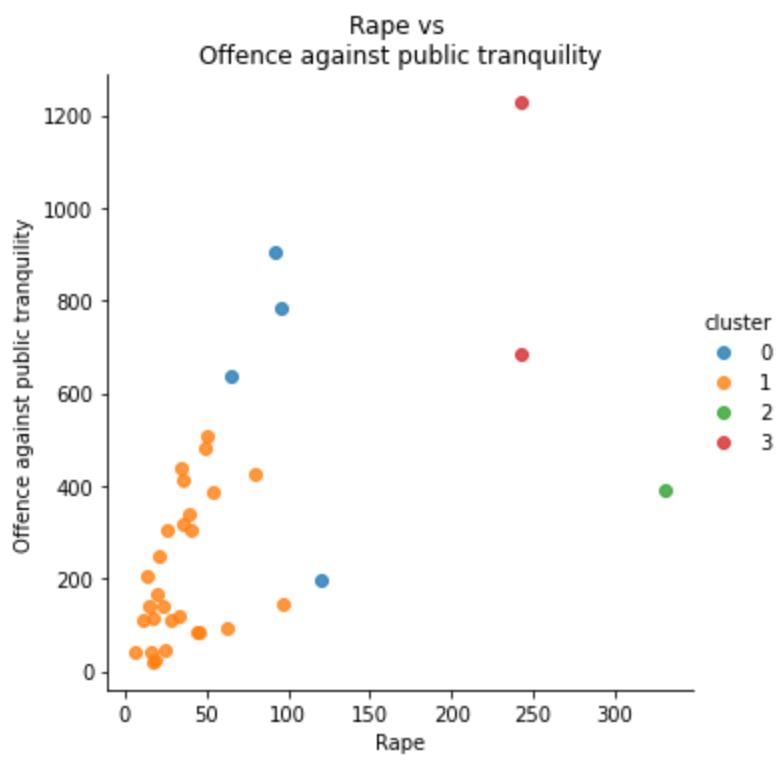


Kidnapping and abduction vs  
Human trafficking



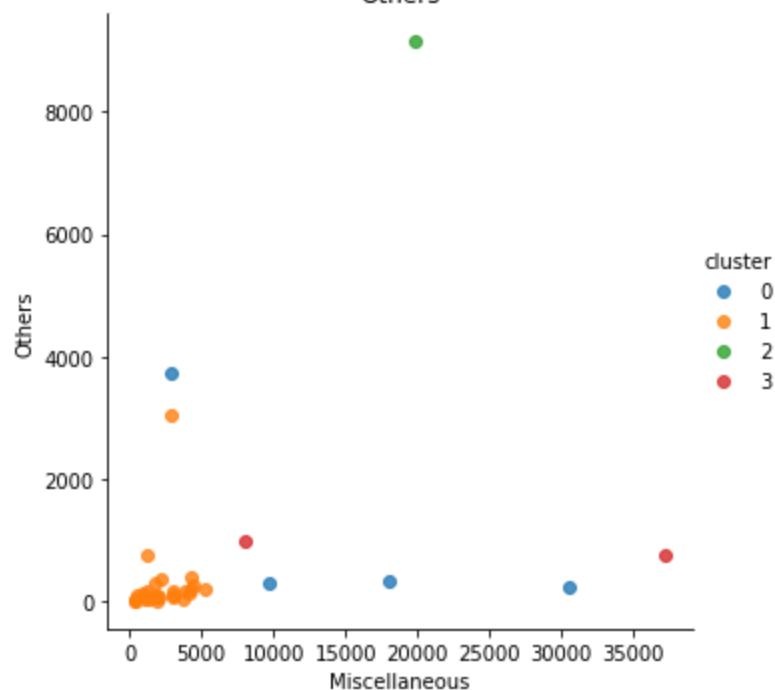
Human trafficking vs  
Rape



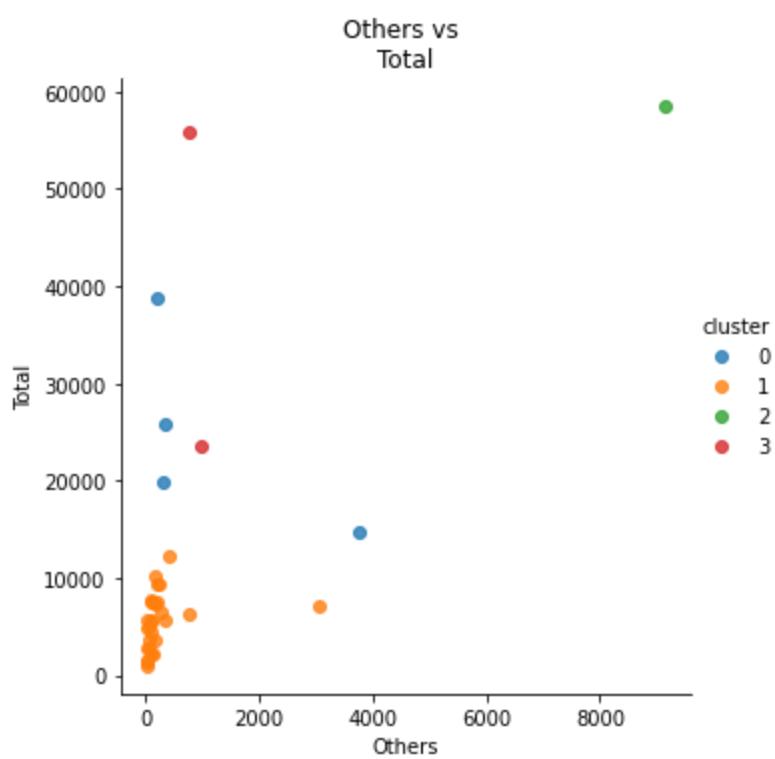


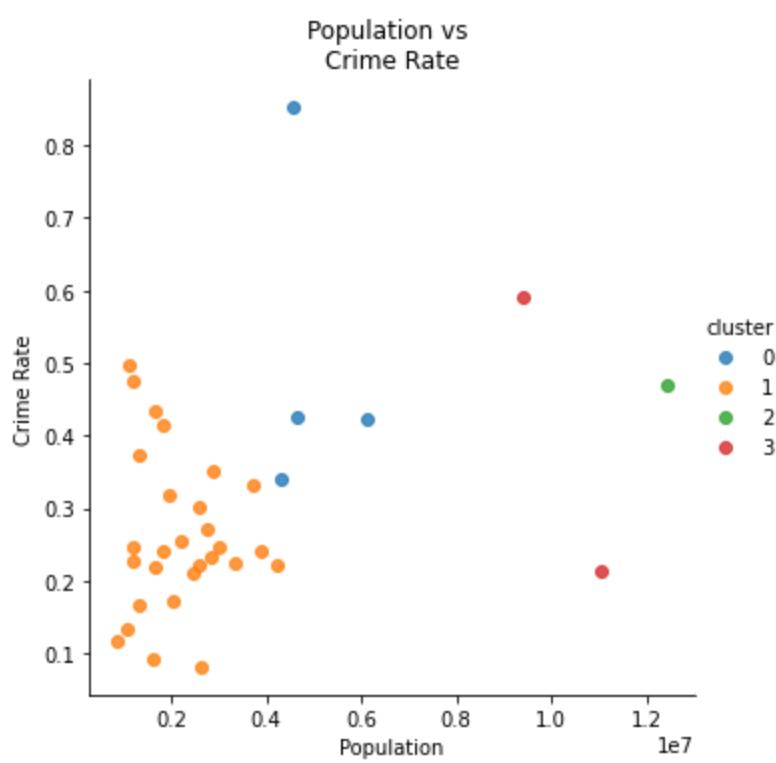
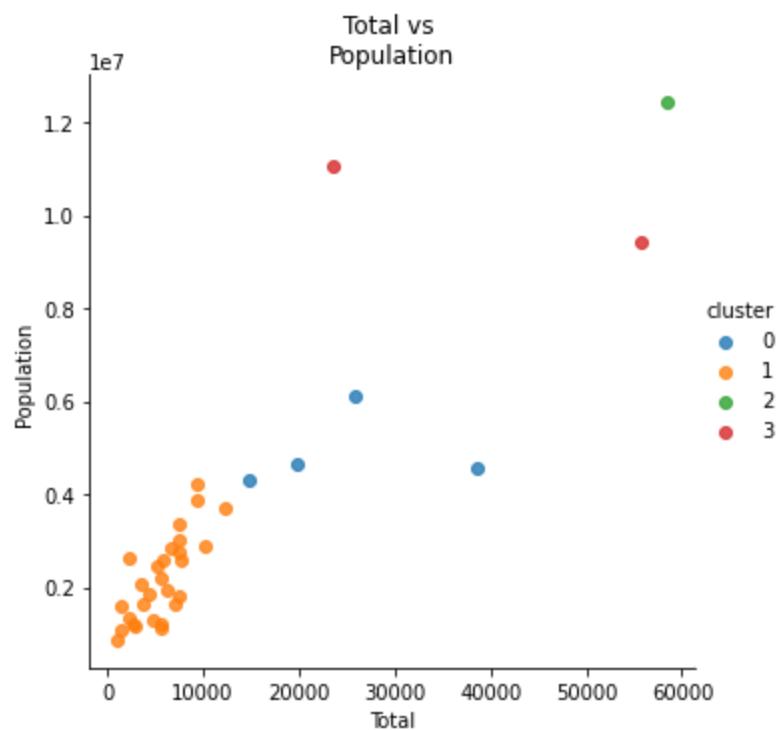


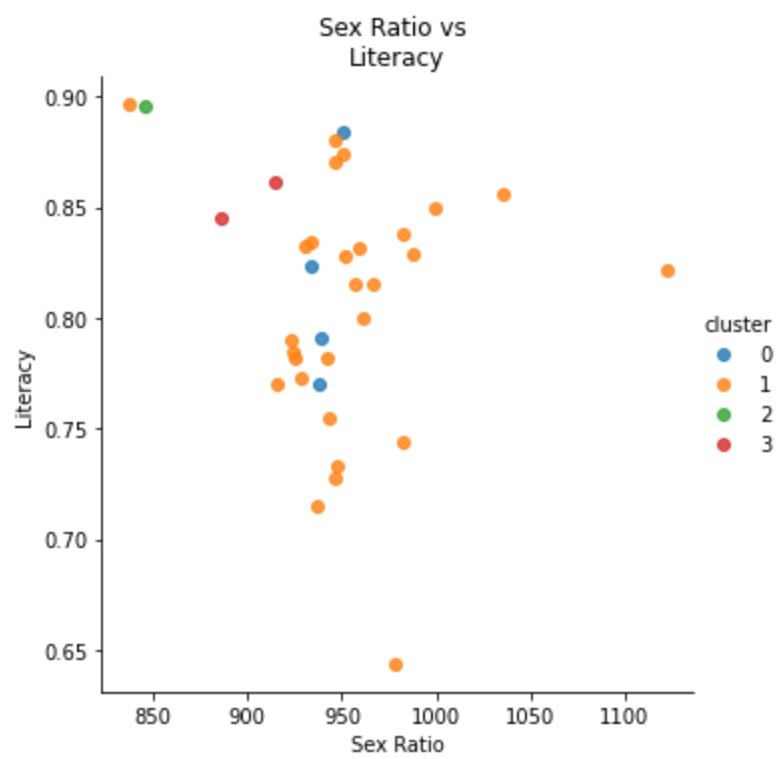
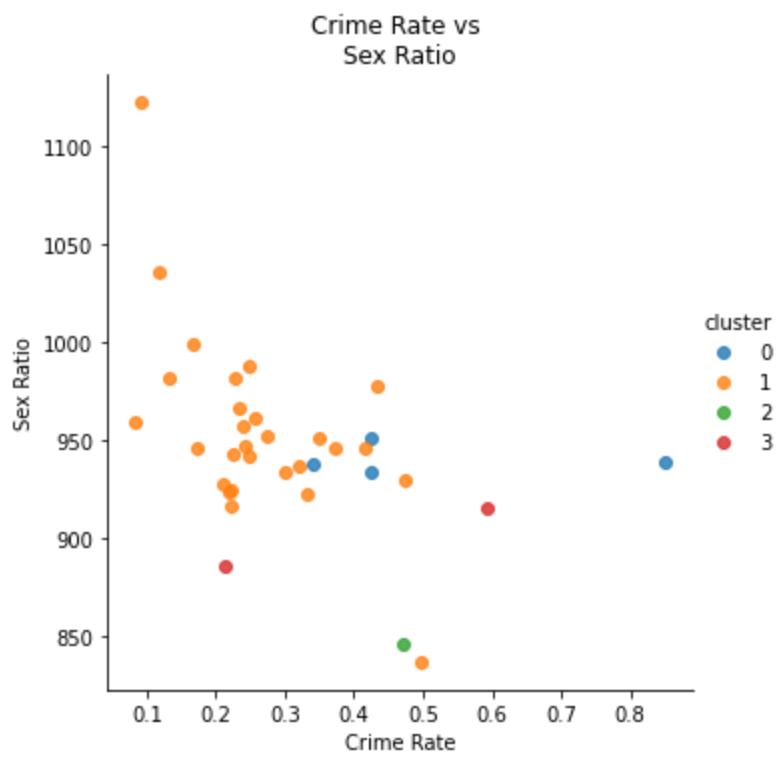
Miscellaneous vs  
Others

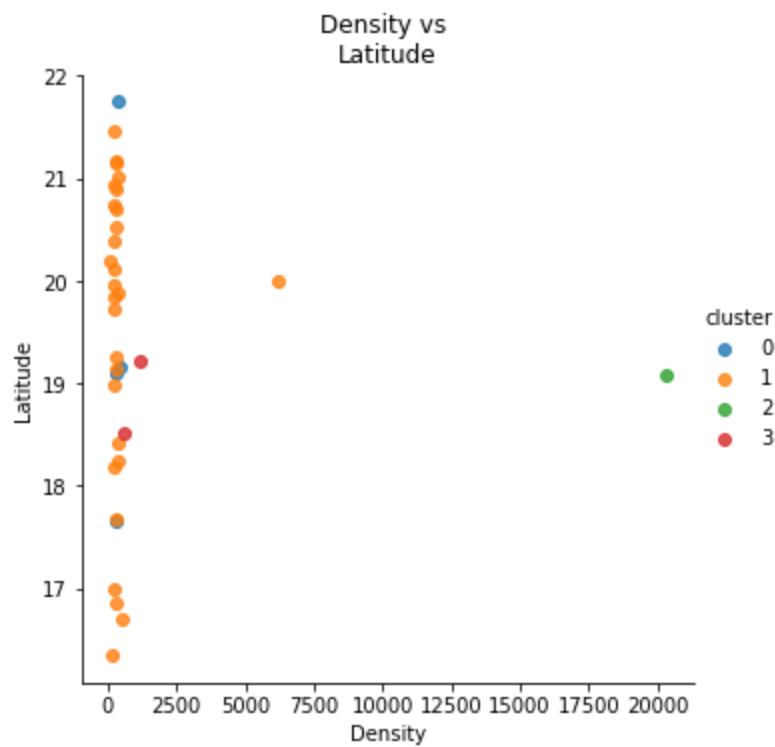
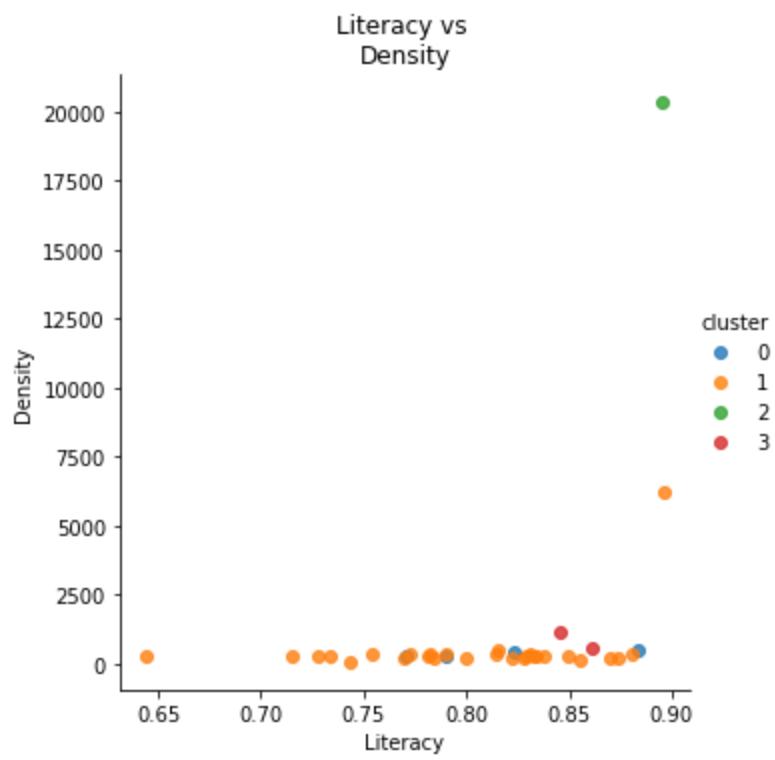


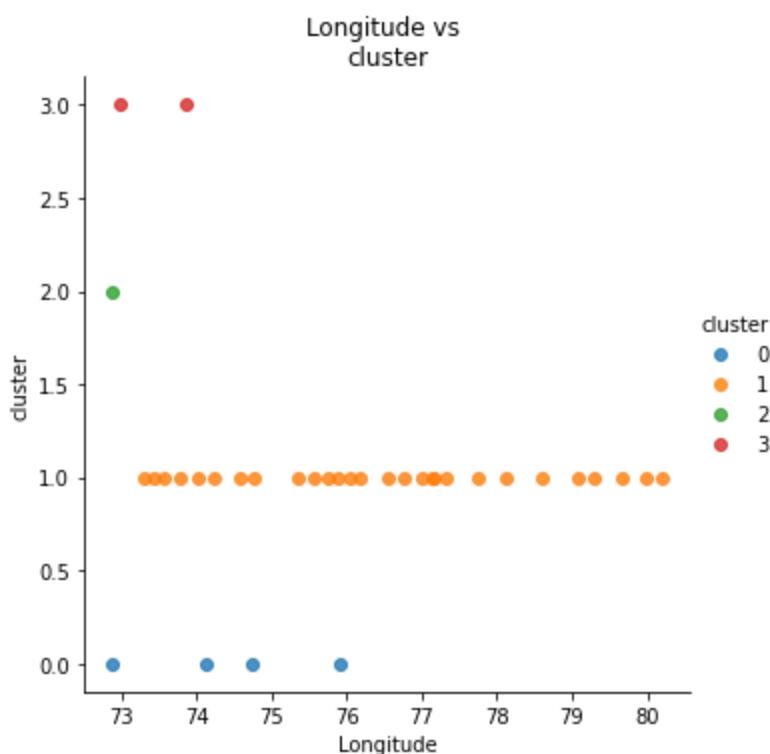
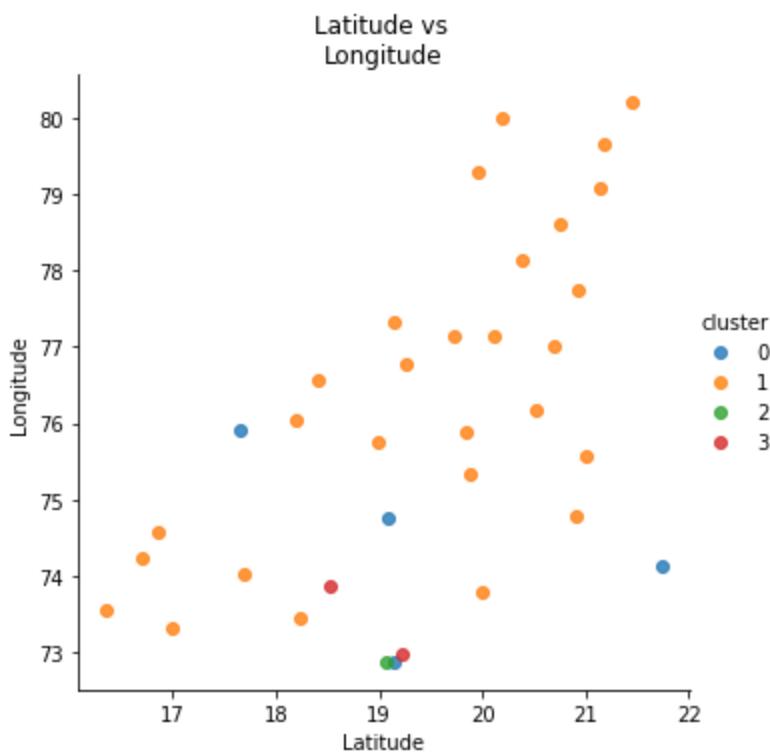
Others vs  
Total









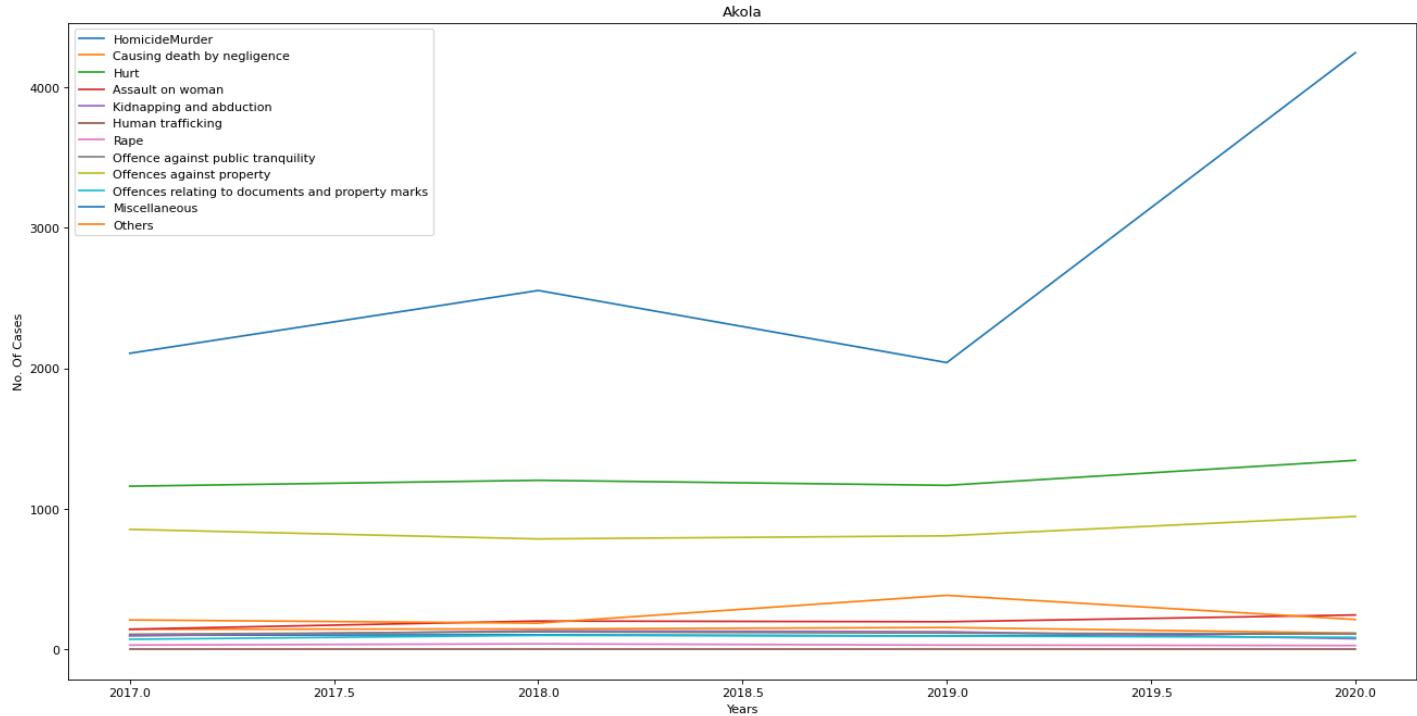
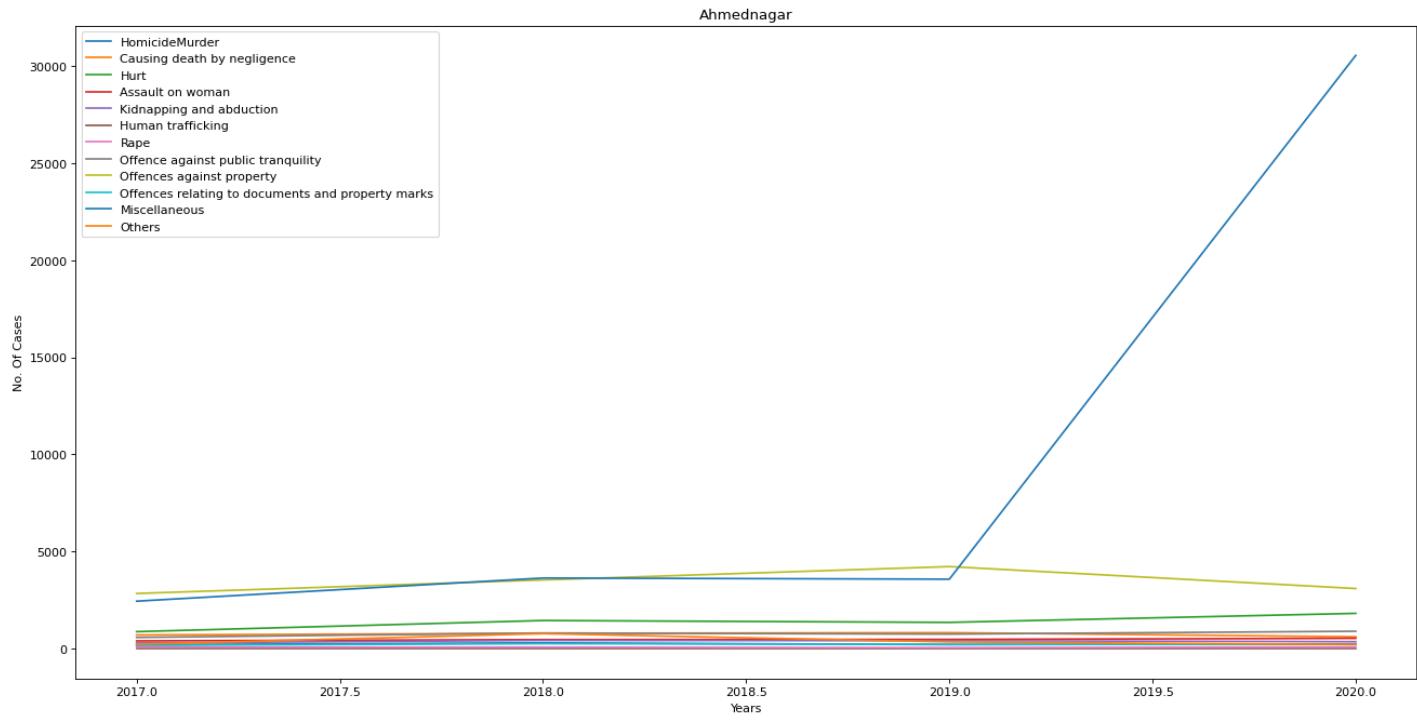


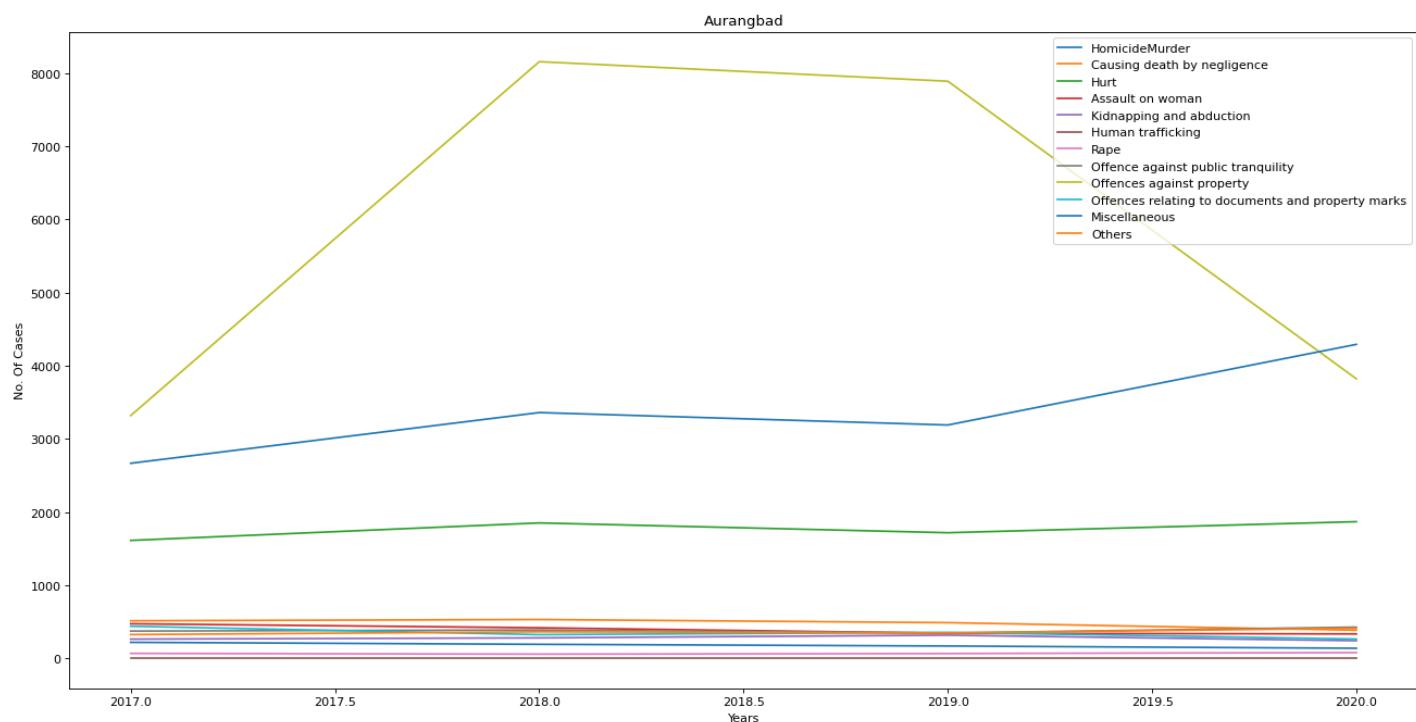
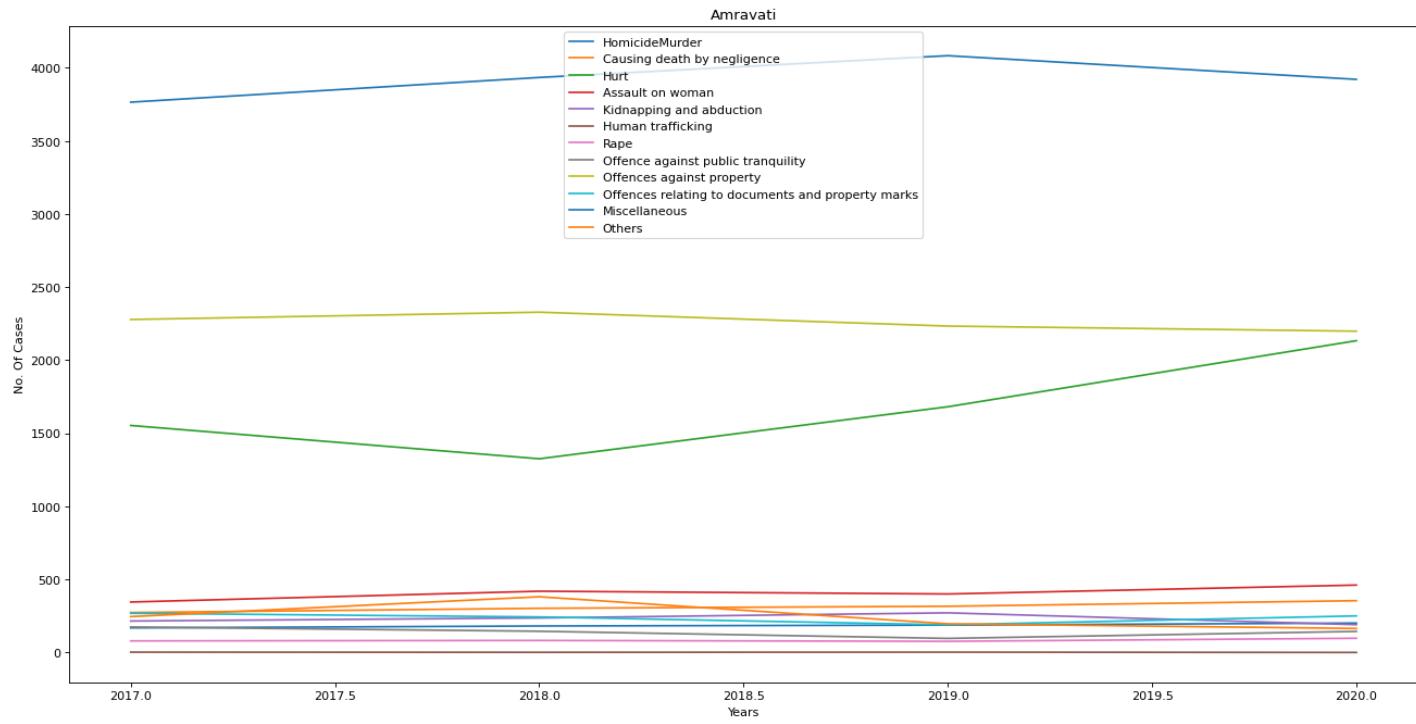
In [31]:

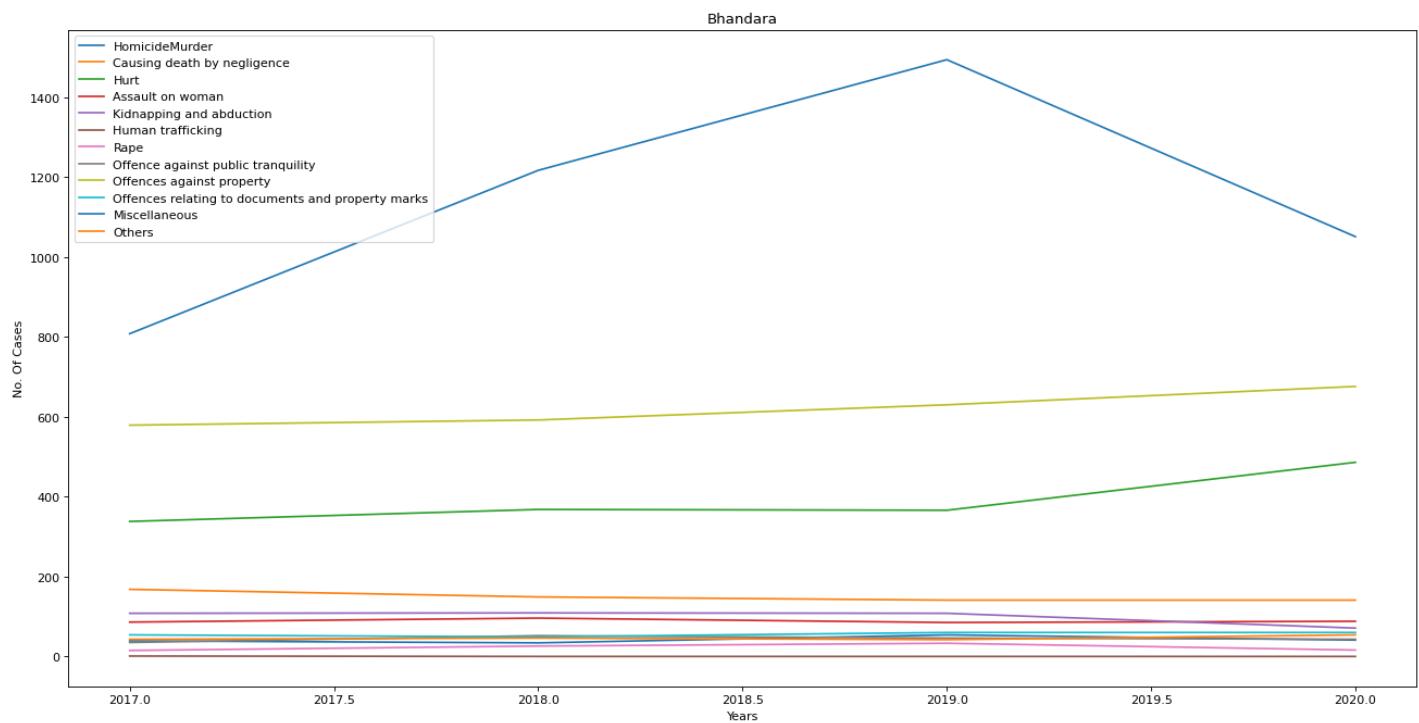
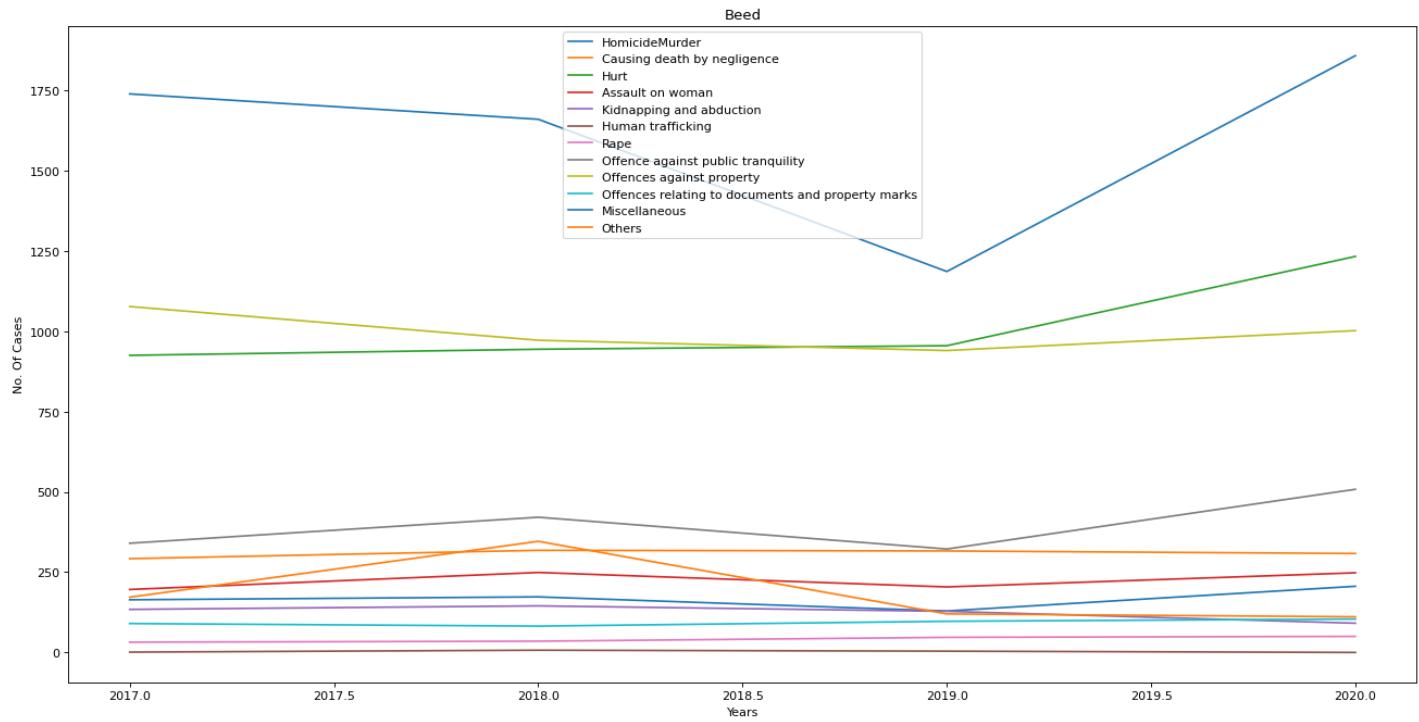
```
#District vs Crime in 4 yrs
df=pd.read_excel(r'DataFinal17-20.xlsx',sheet_name='practice')
df2=df.iloc[:,0:6]
years=[2017,2018,2019,2020]
Districts=df['District'].unique()
Crimes=df['Crime'].unique()
for district in Districts:
    fig=plt.figure(figsize=(20,10),dpi=80,facecolor='w',edgecolor='k')
    plt.title(district)
    plt.xlabel('Years')
    plt.ylabel('No. Of Cases')
    for case in Crimes:
        temp_df=df[(df['District']==district)&(df['Crime']==case)]
        N_cases=[temp_df[c].values[0] for c in years]
        plt.plot(years,N_cases)
    plt.legend(Crimes)
```

C:\Users\WELCOME\AppData\Local\Temp/ipykernel\_2092/3855337537.py:8: RuntimeWarning: More than 20 figures have been opened. Figures created through the pyplot interface (`matplotlib.pyplot.figure`) are retained until explicitly closed and may consume too much memory. (To control this warning, see the rcParam `figure.max\_open\_warning`).

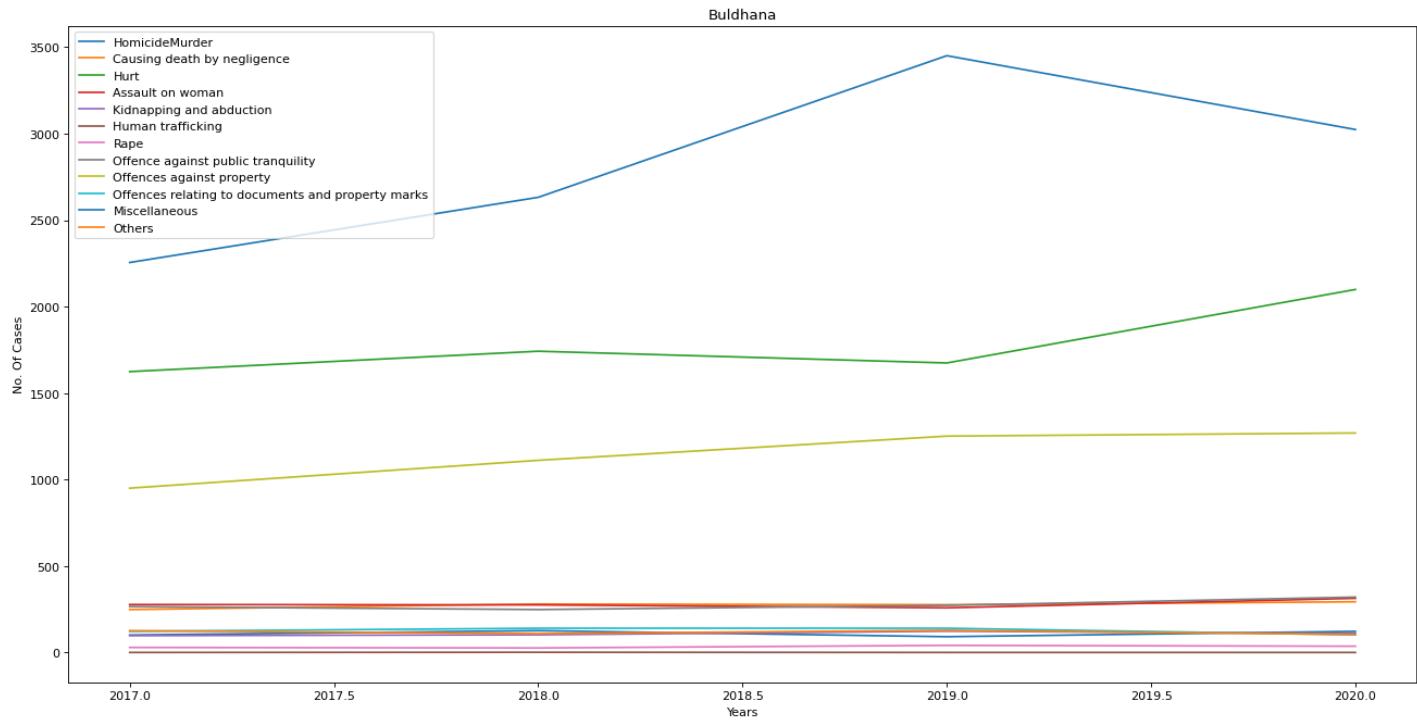
```
fig=plt.figure(figsize=(20,10),dpi=80,facecolor='w',edgecolor='k')
```



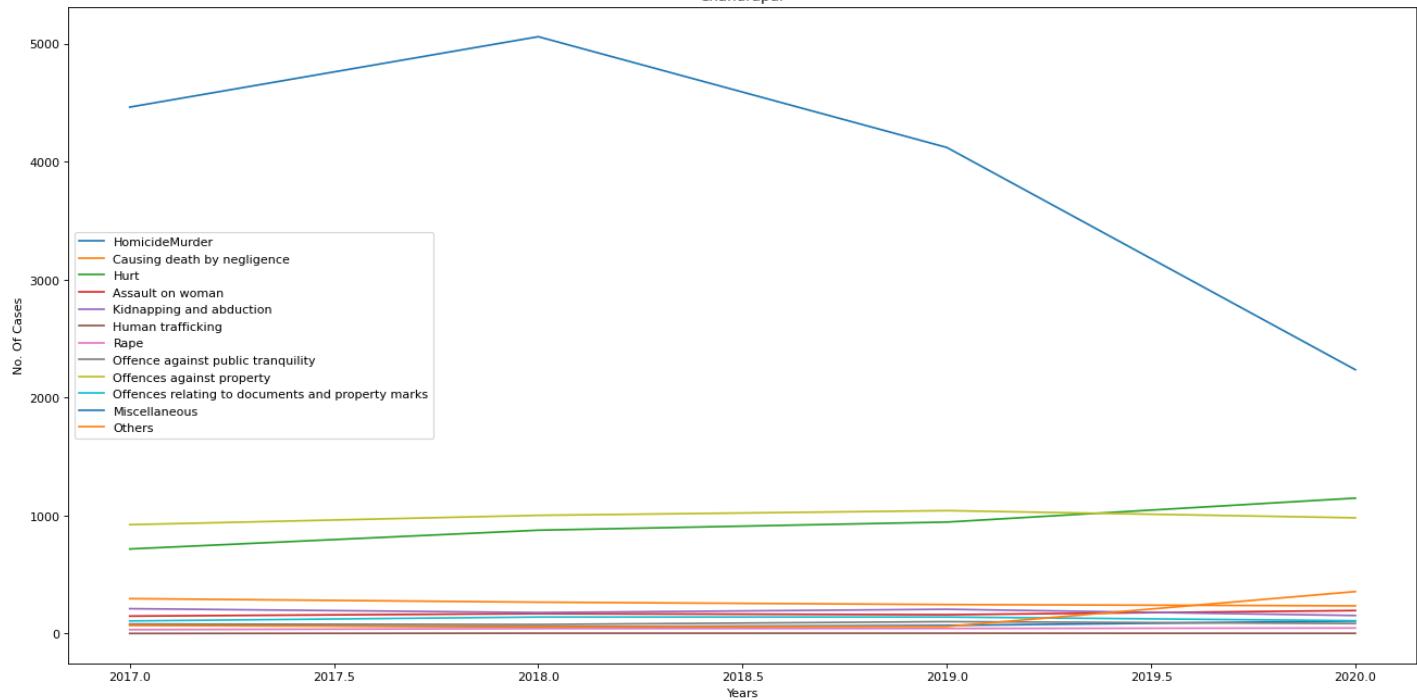


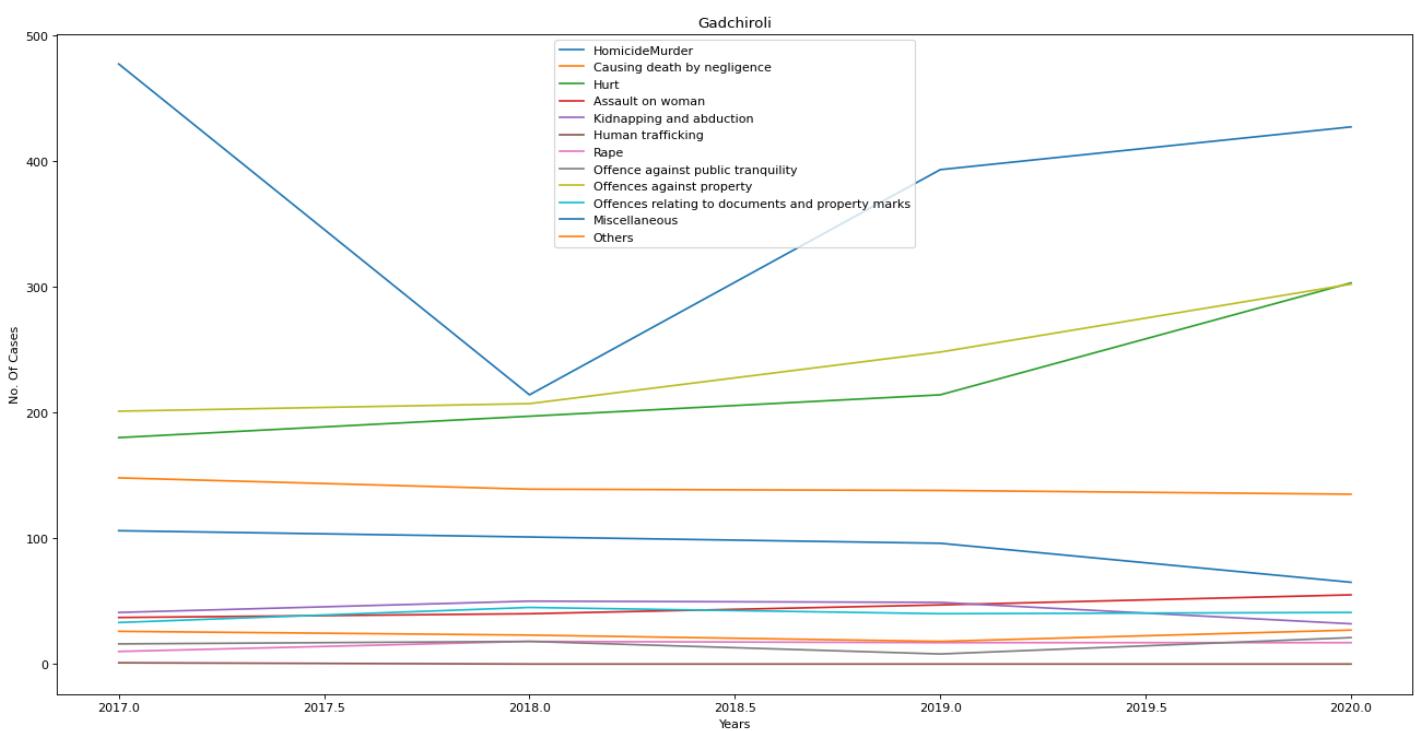
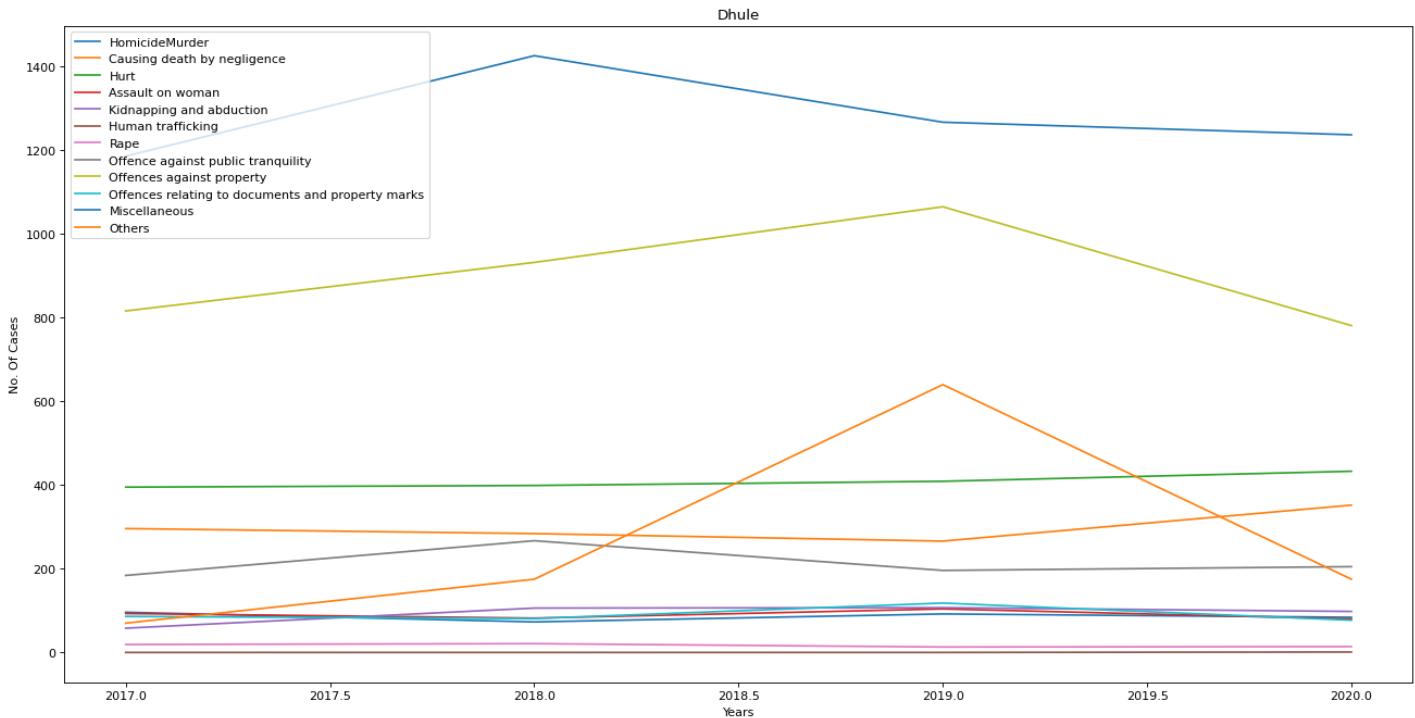


Buldhana

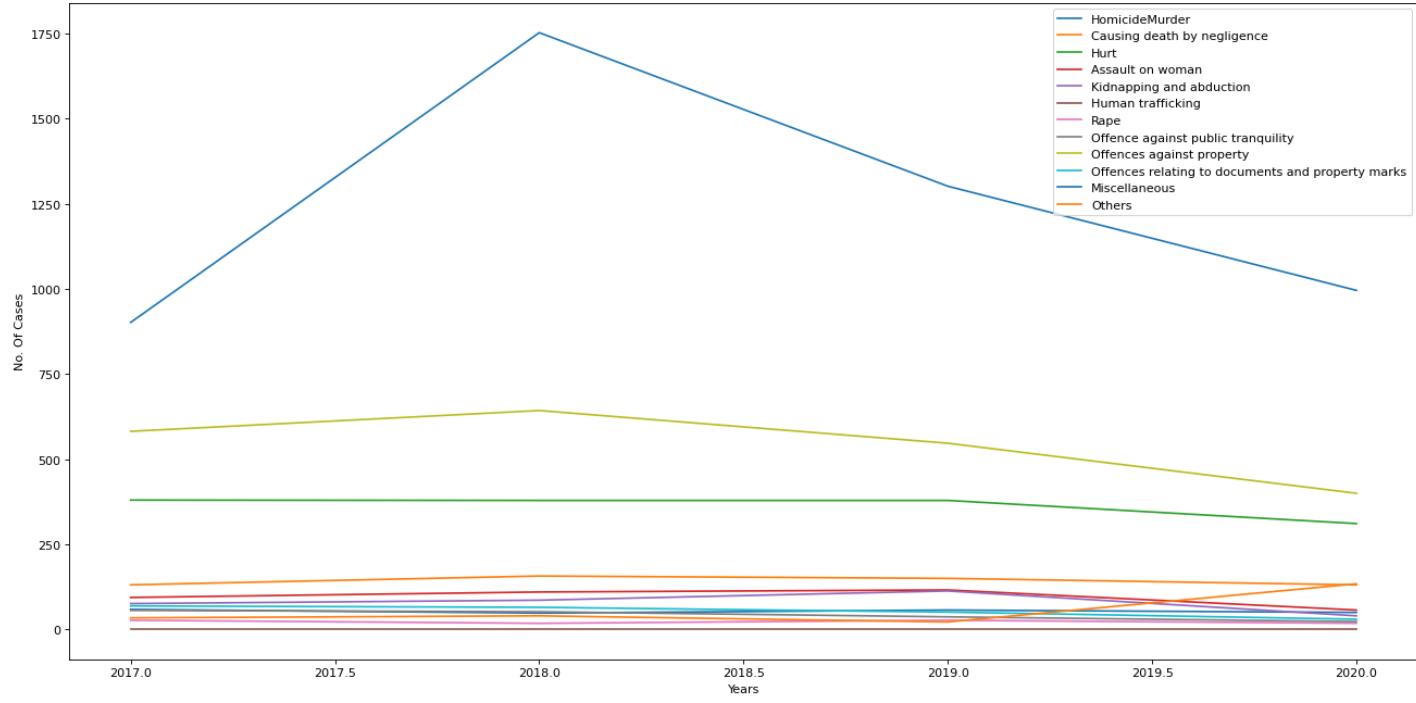


Chandrapur

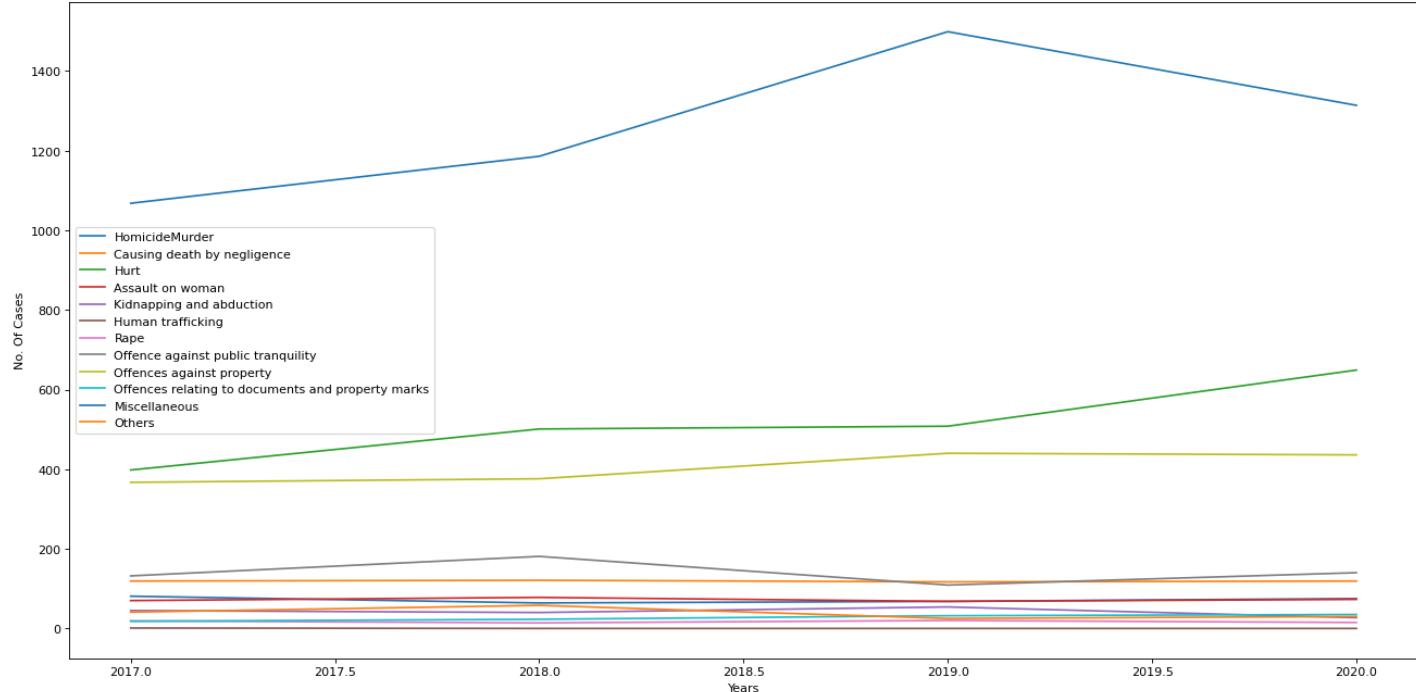




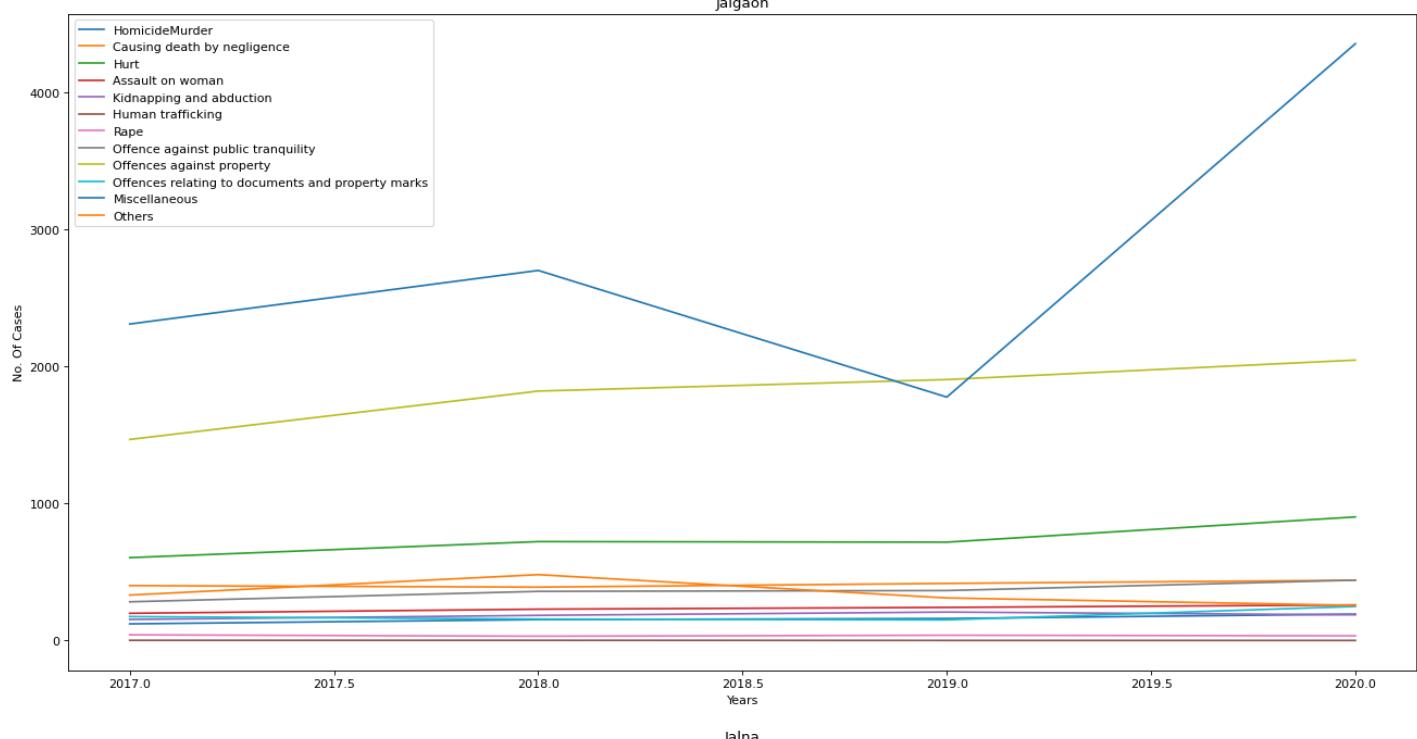
Gondia



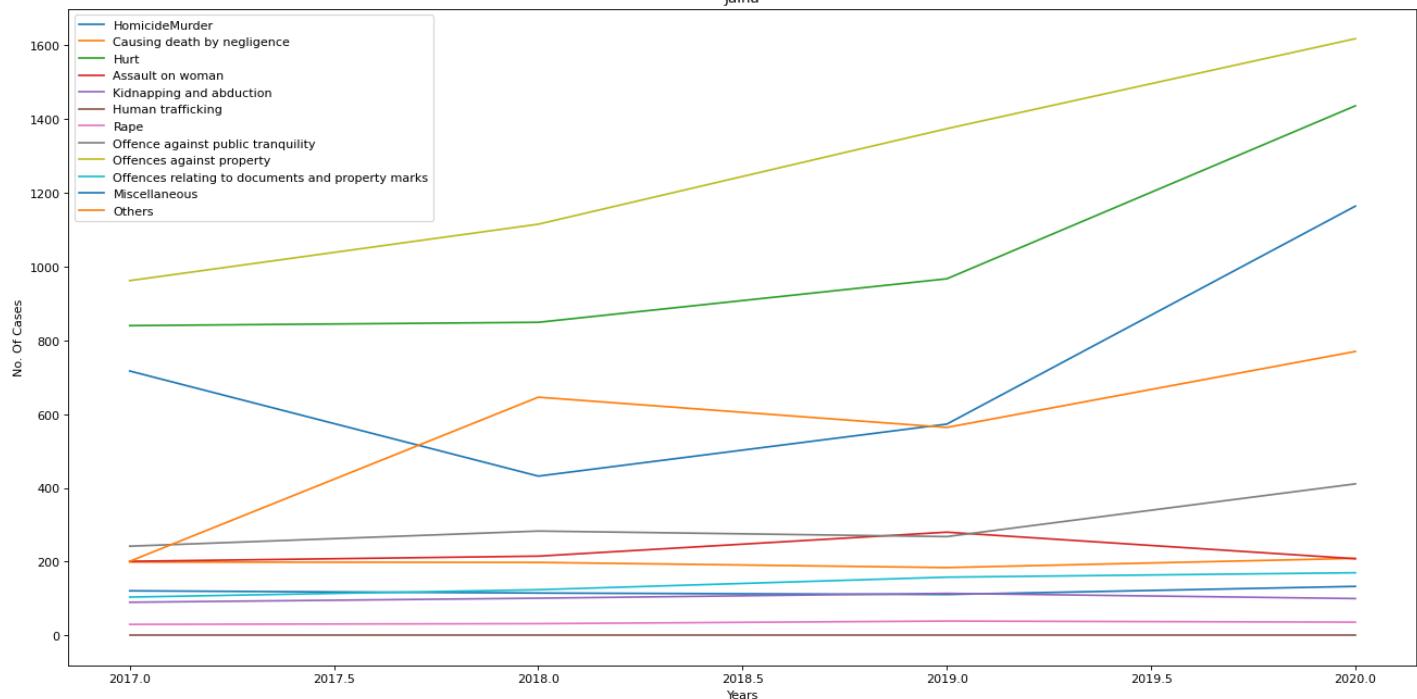
Hingoli

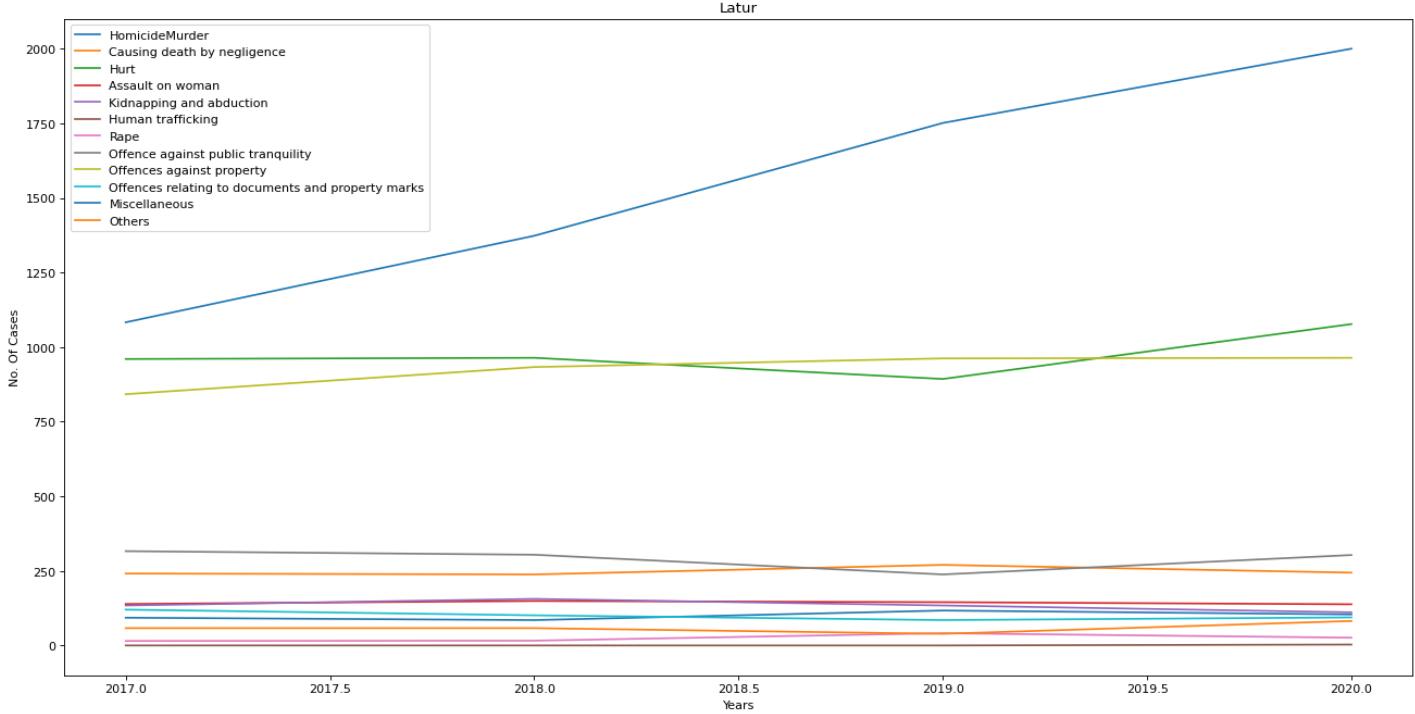
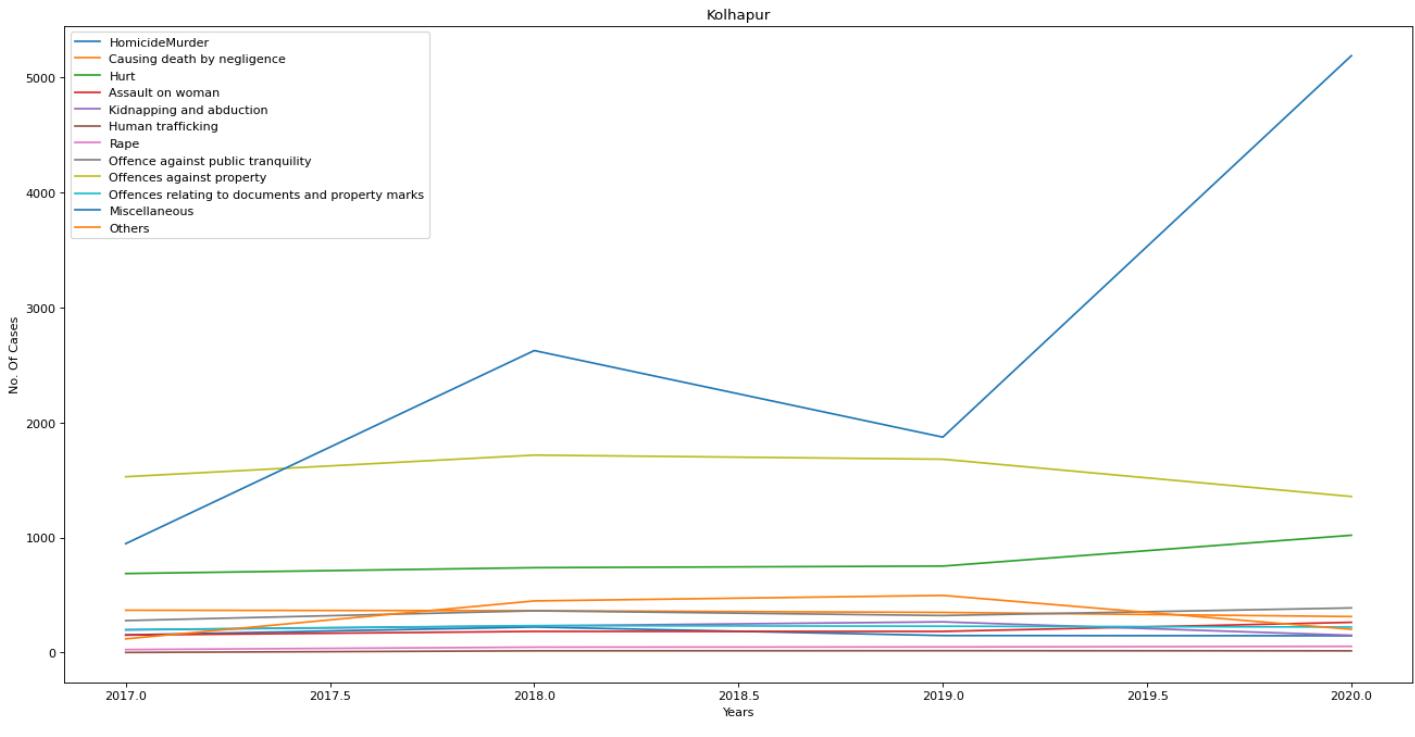


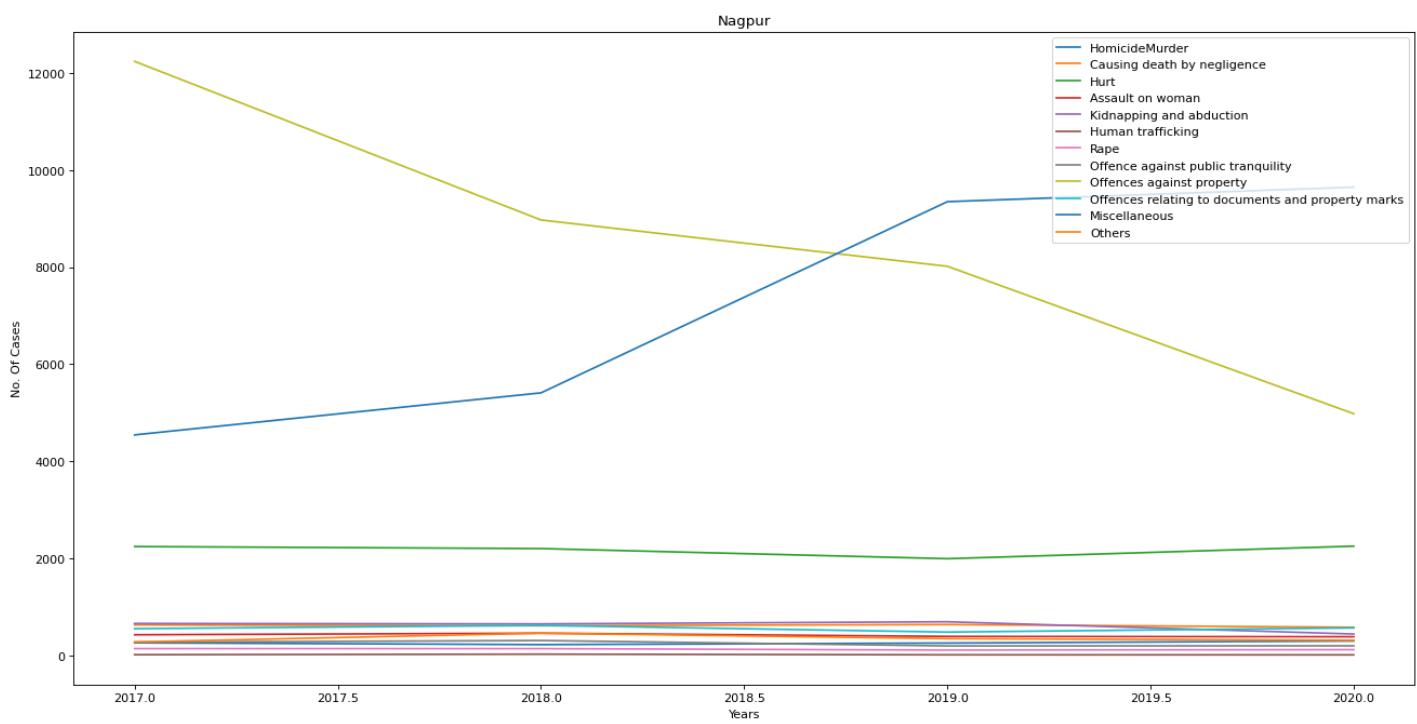
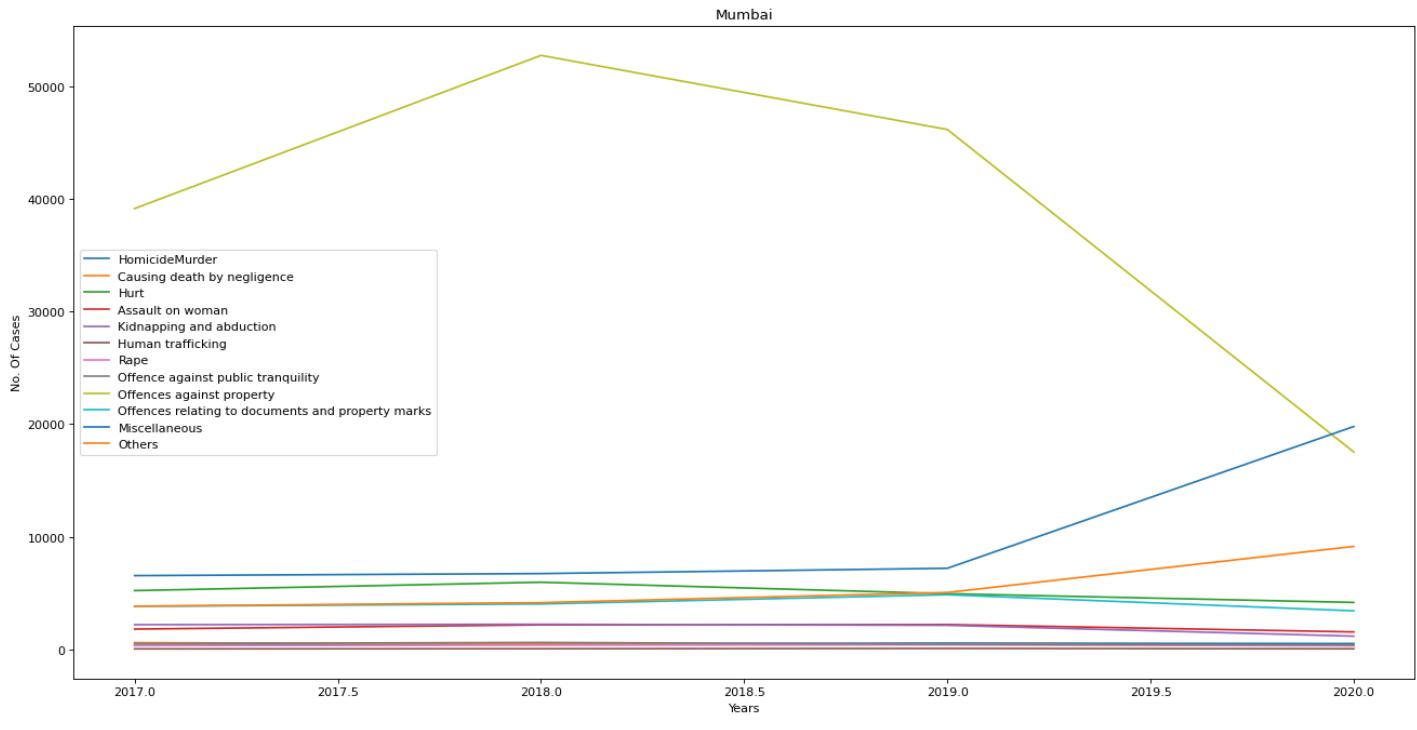
Jalgaon

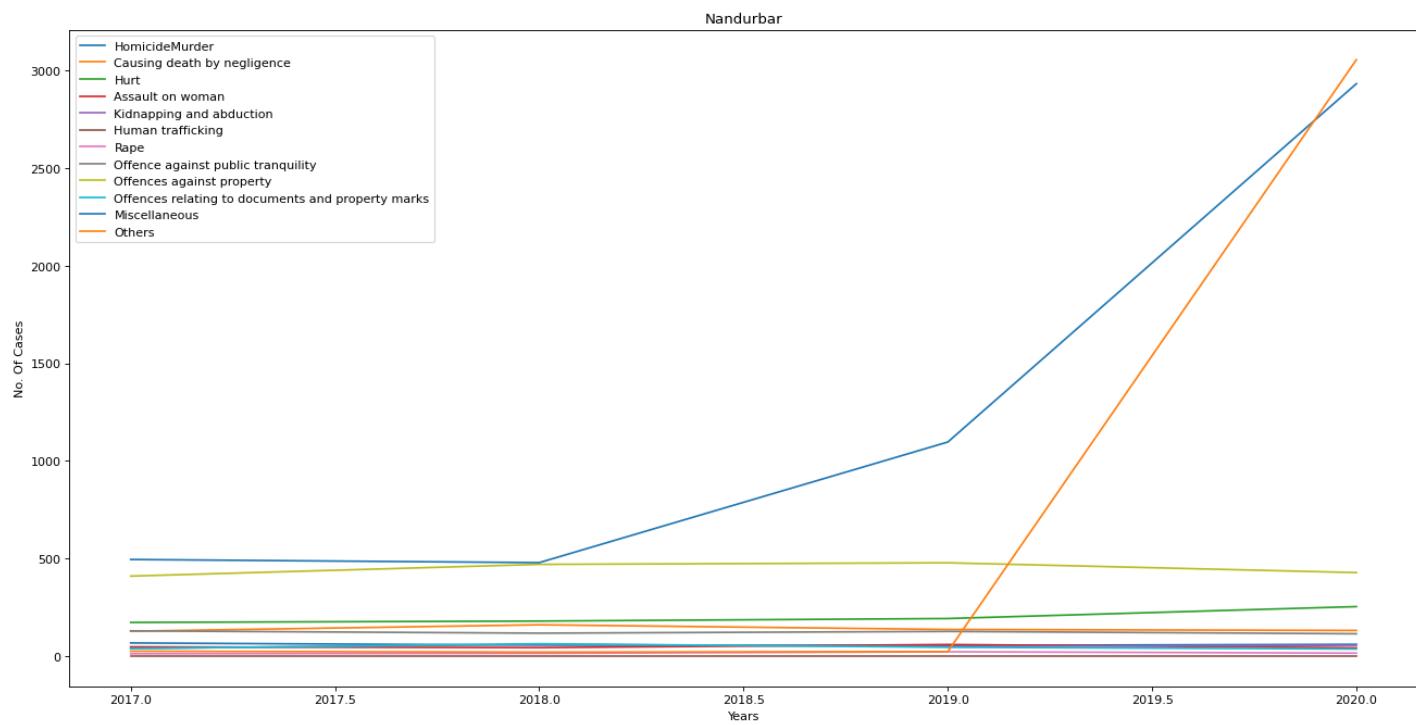
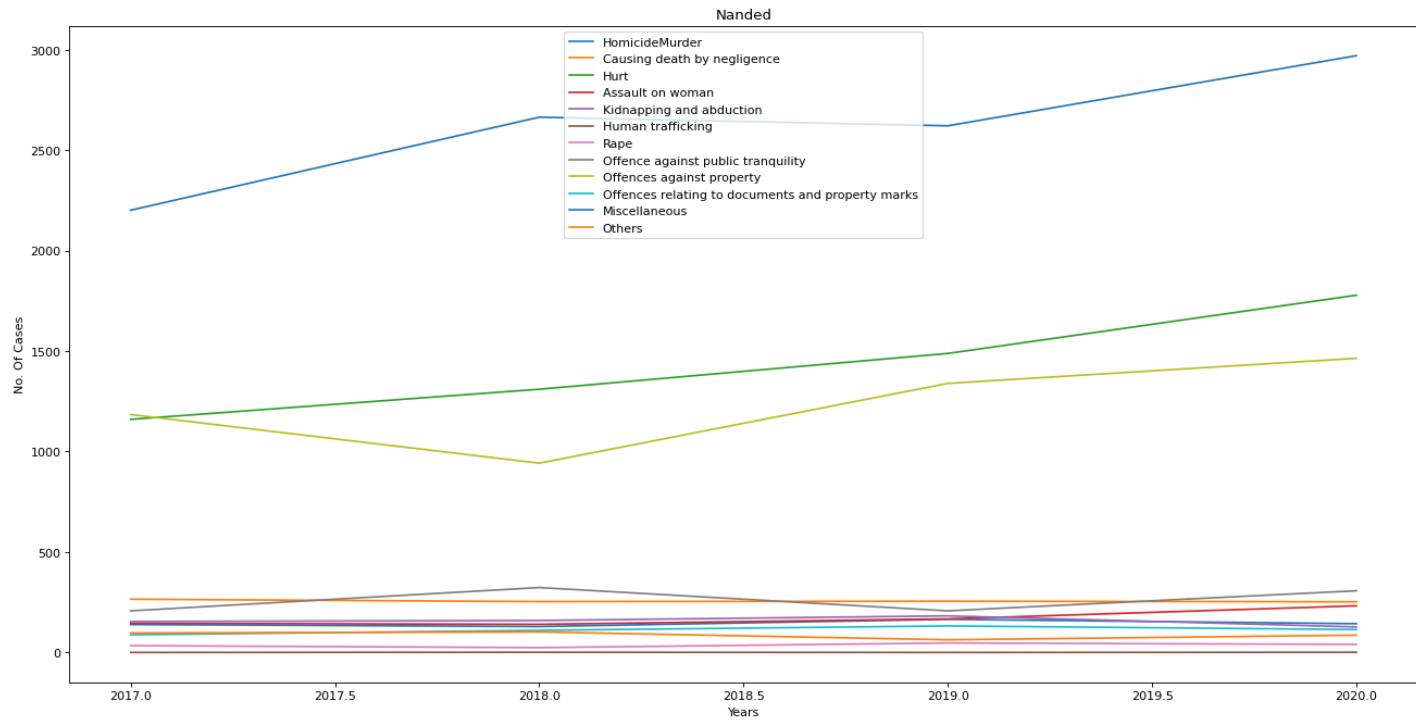


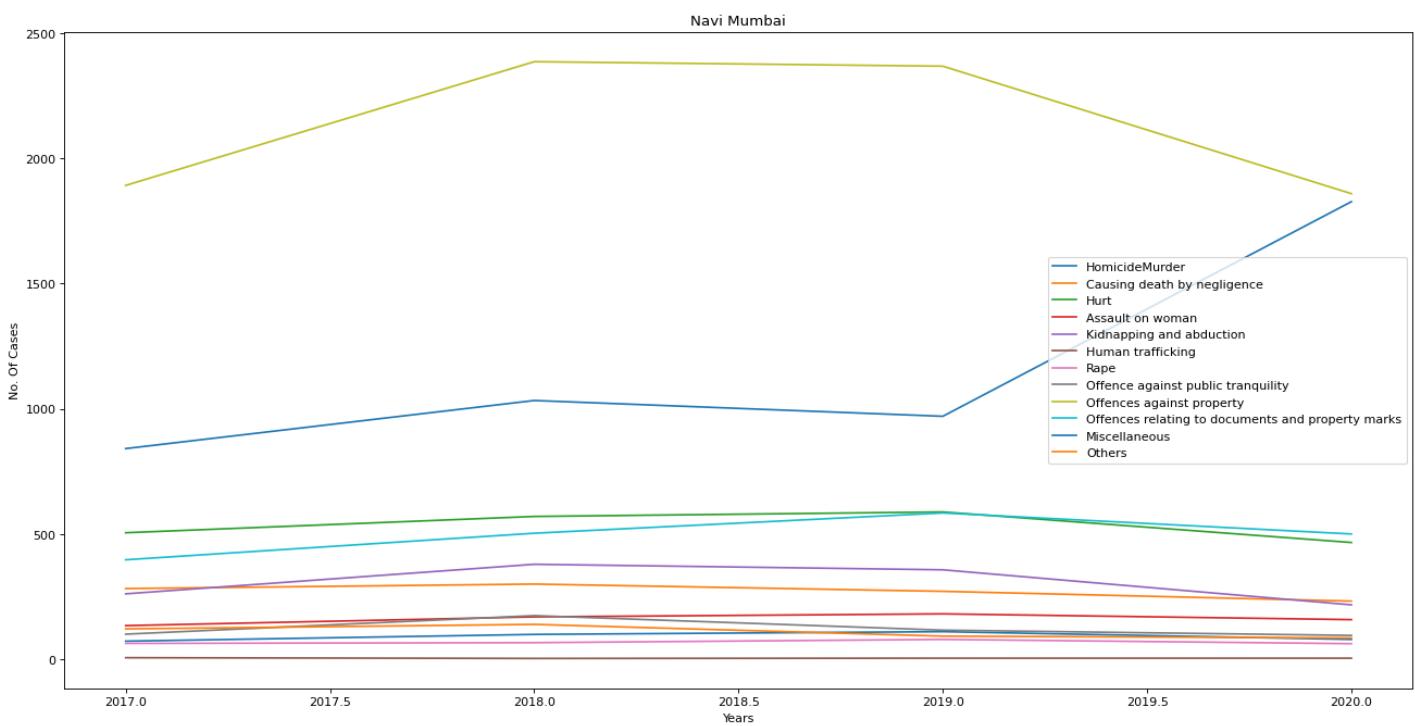
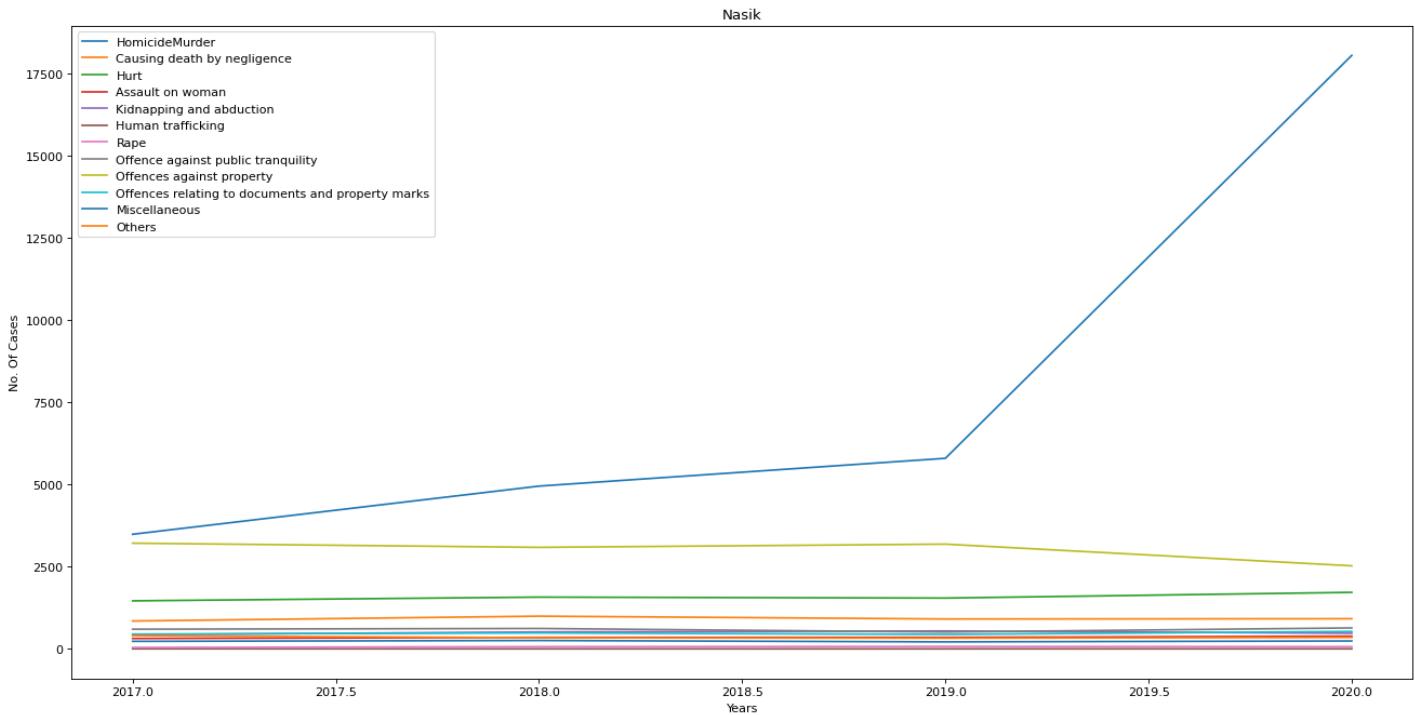
Jalna



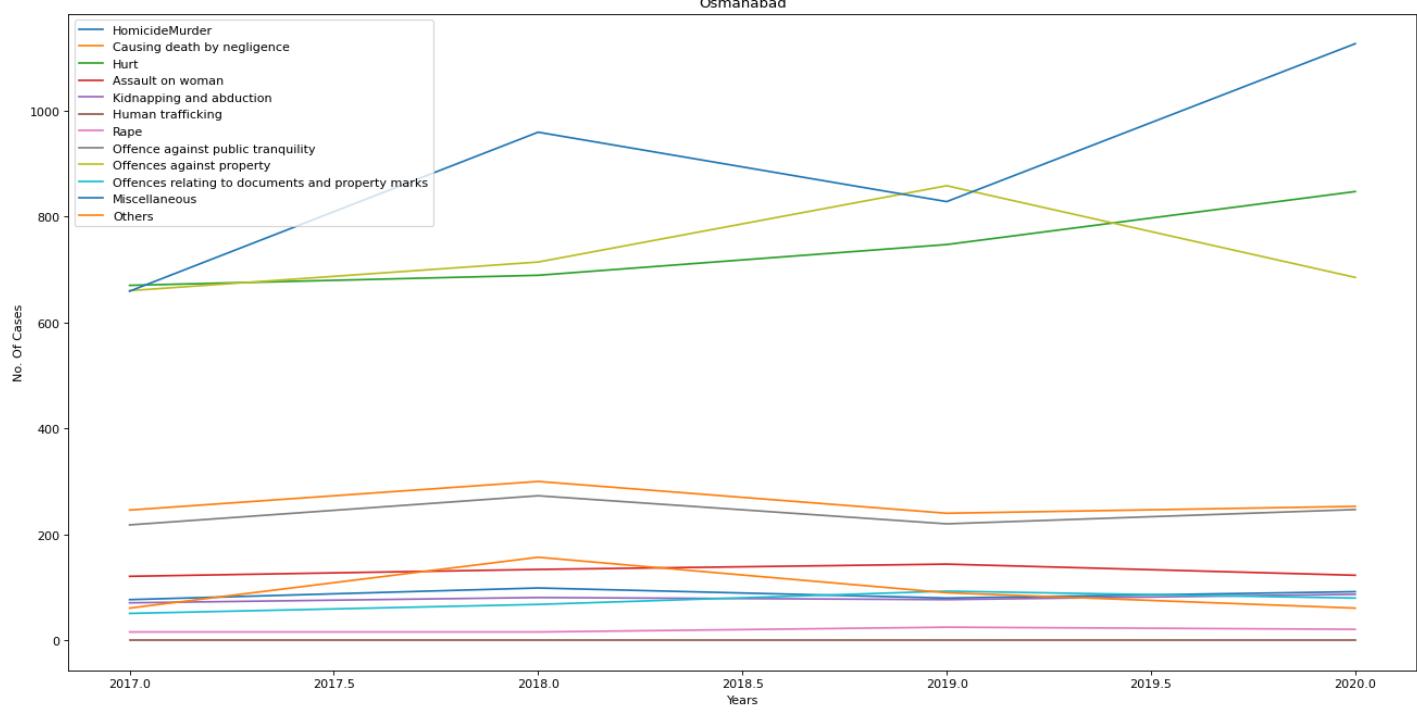




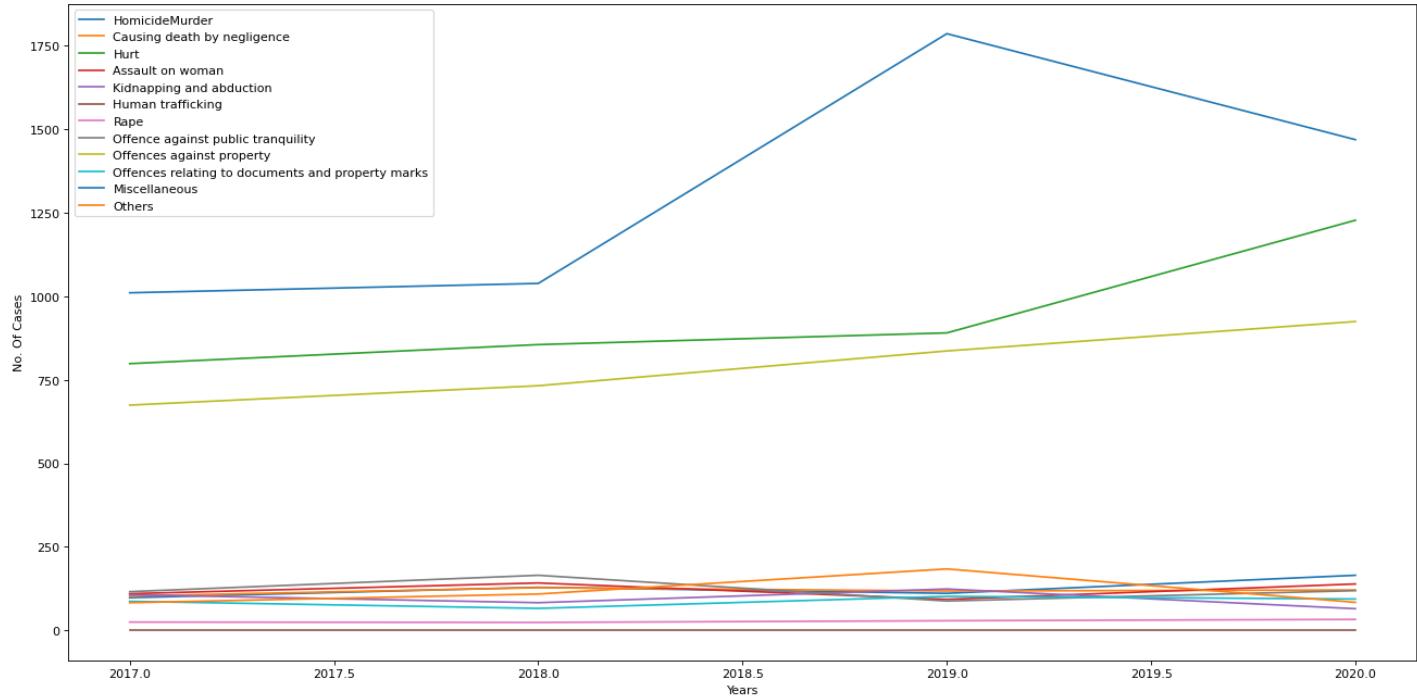


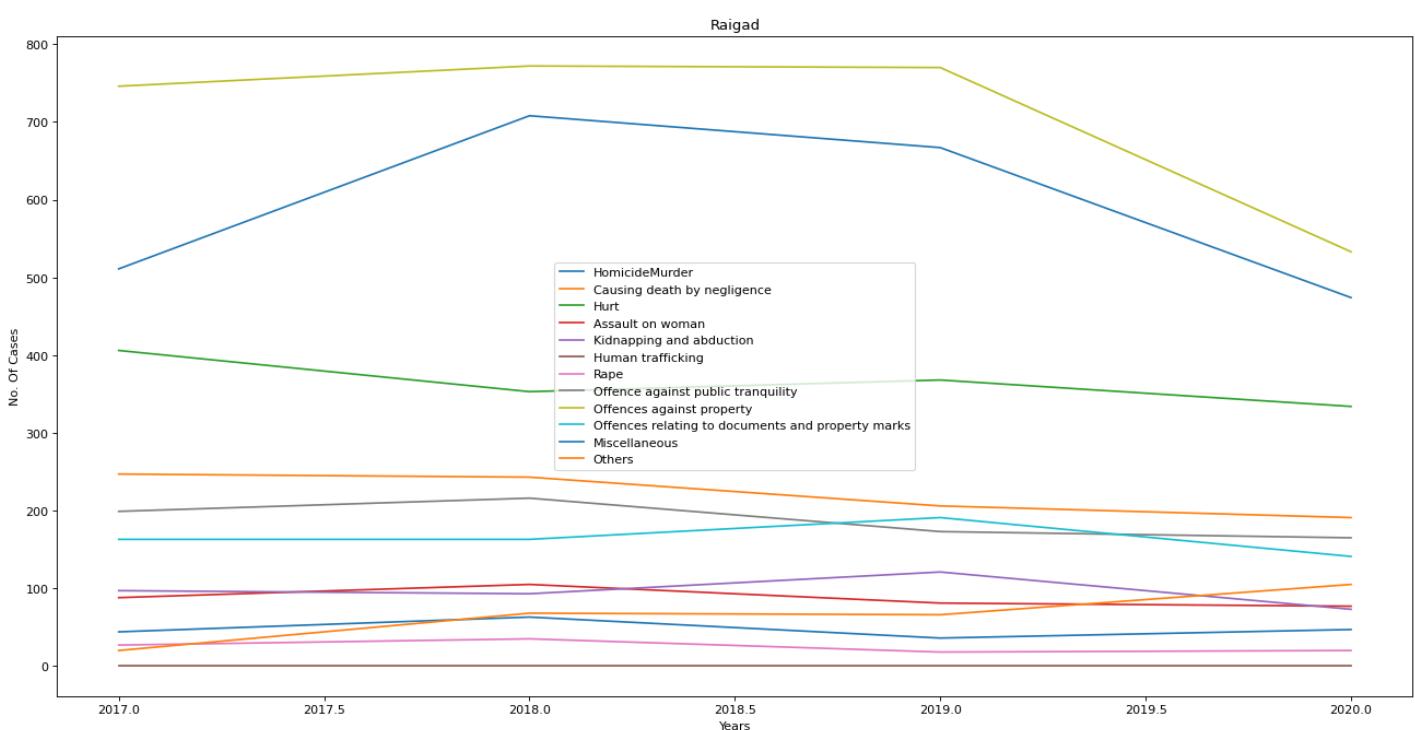
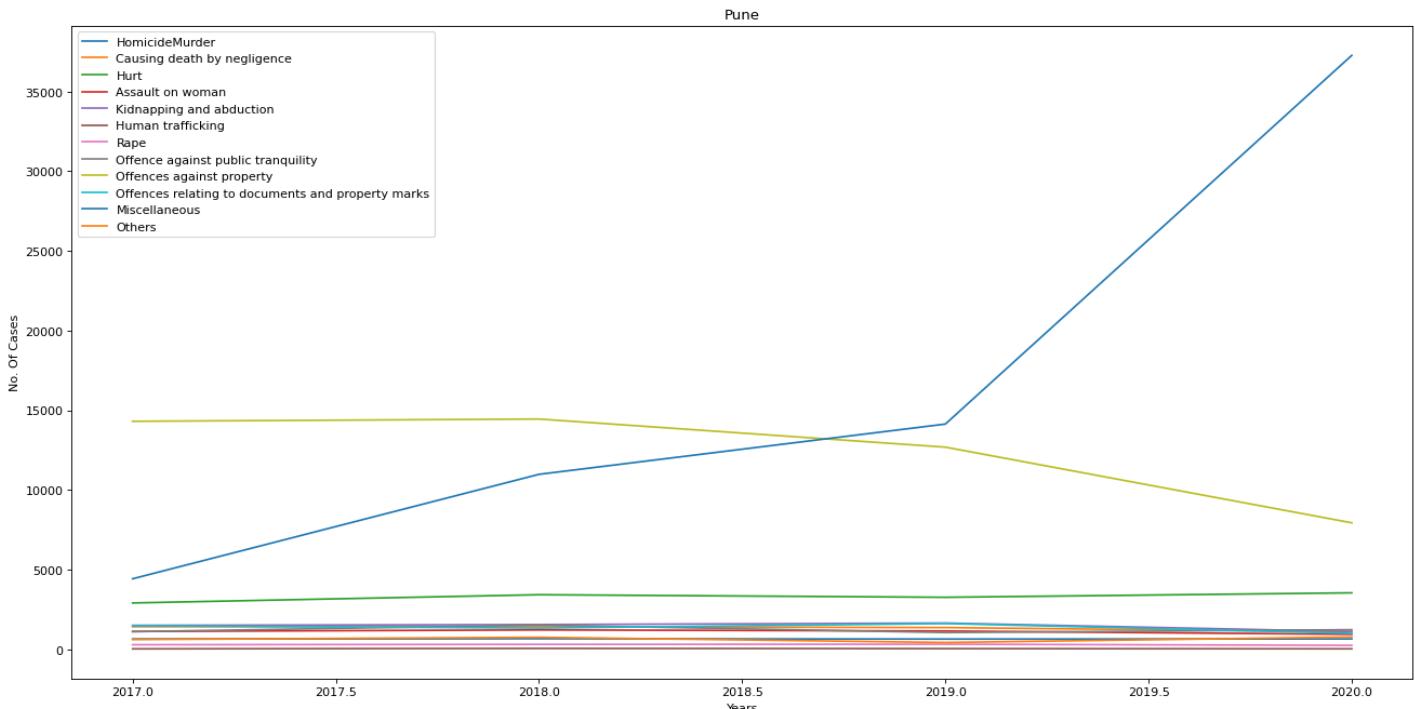


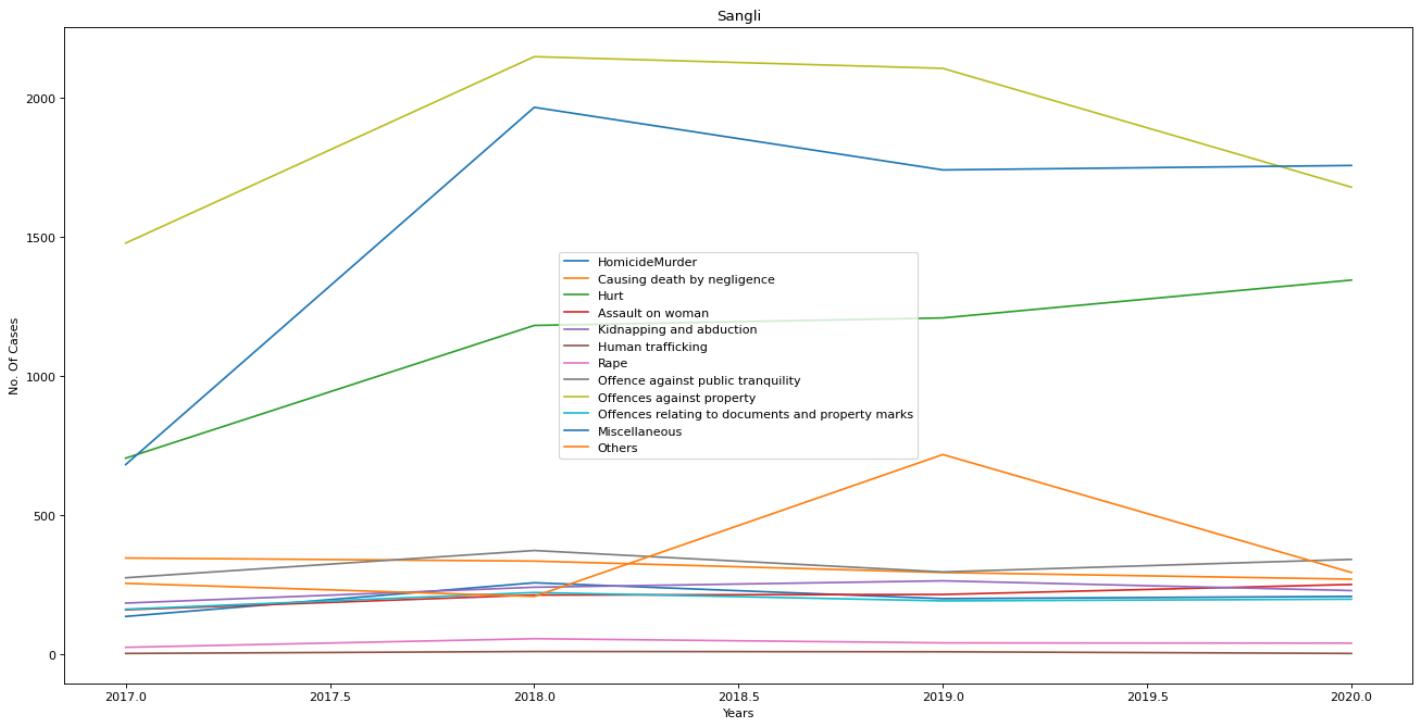
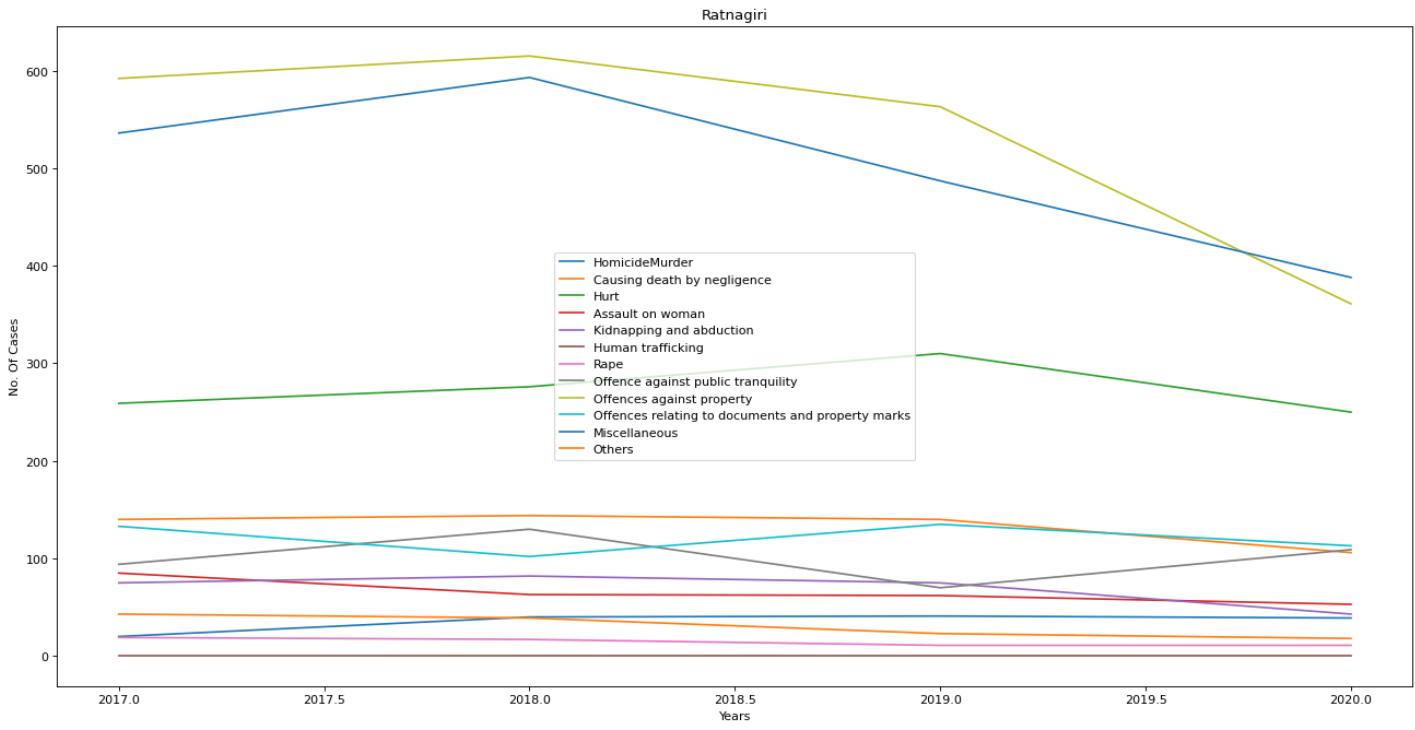
Osmanabad

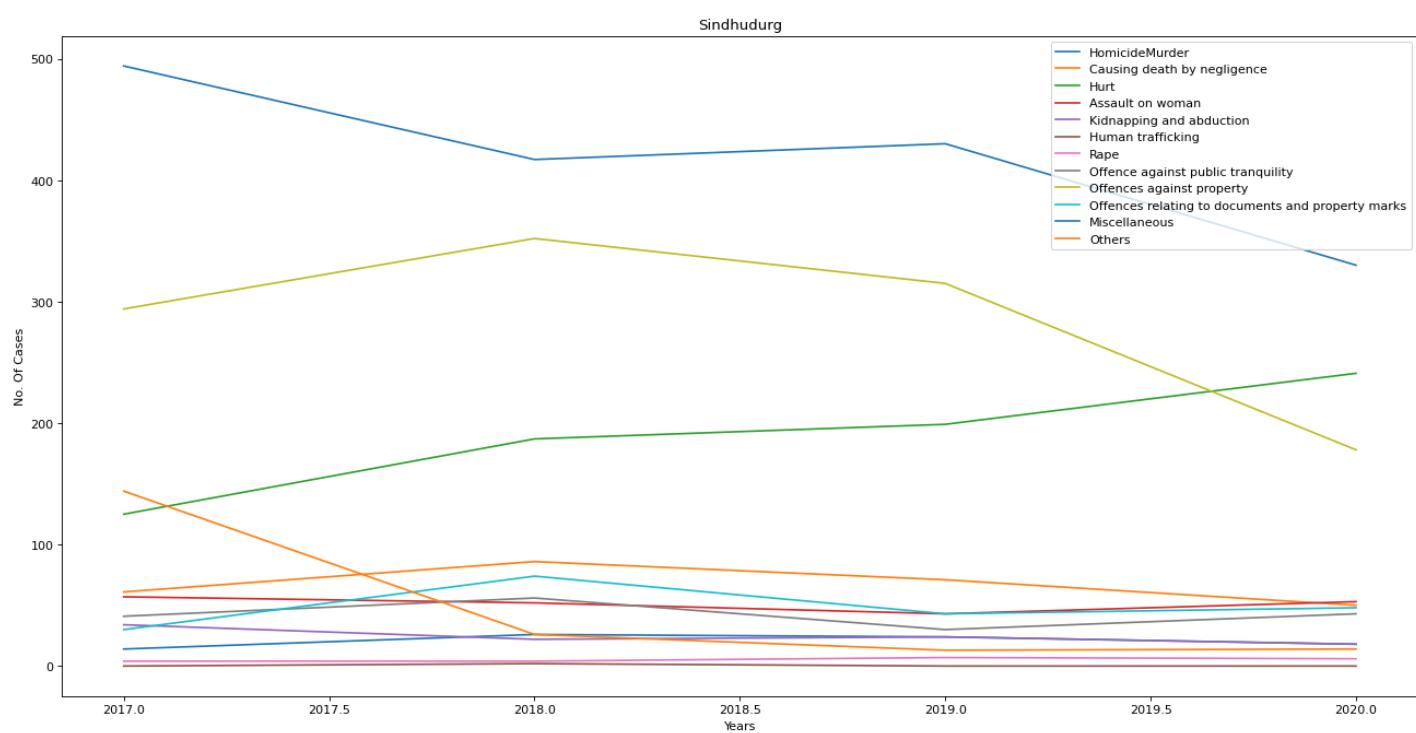
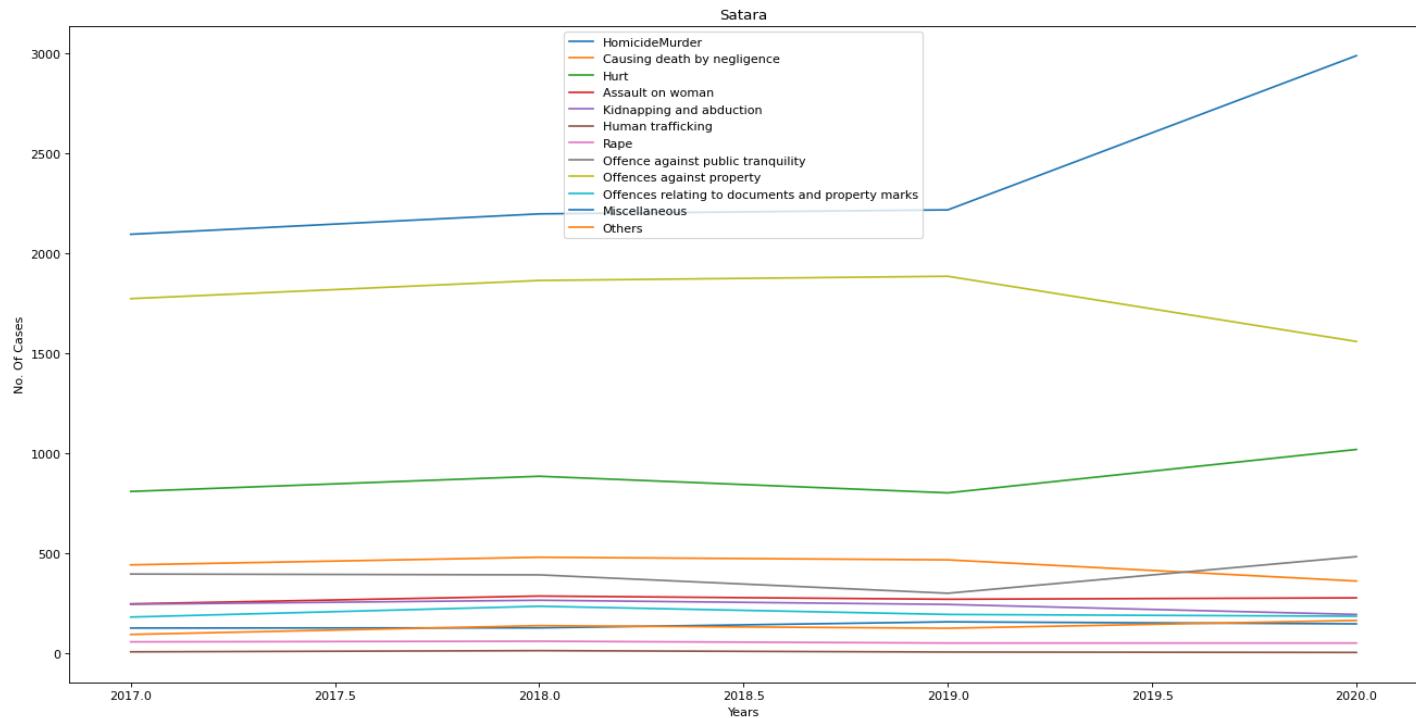


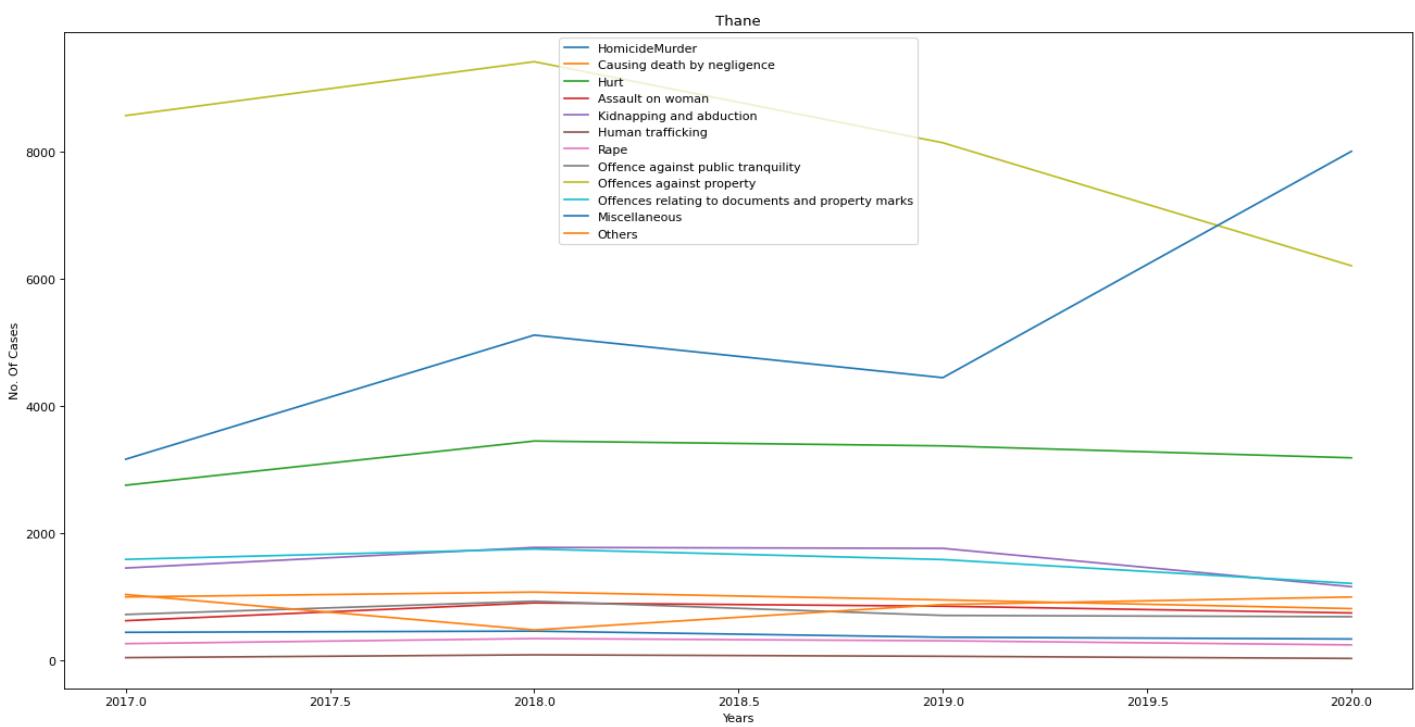
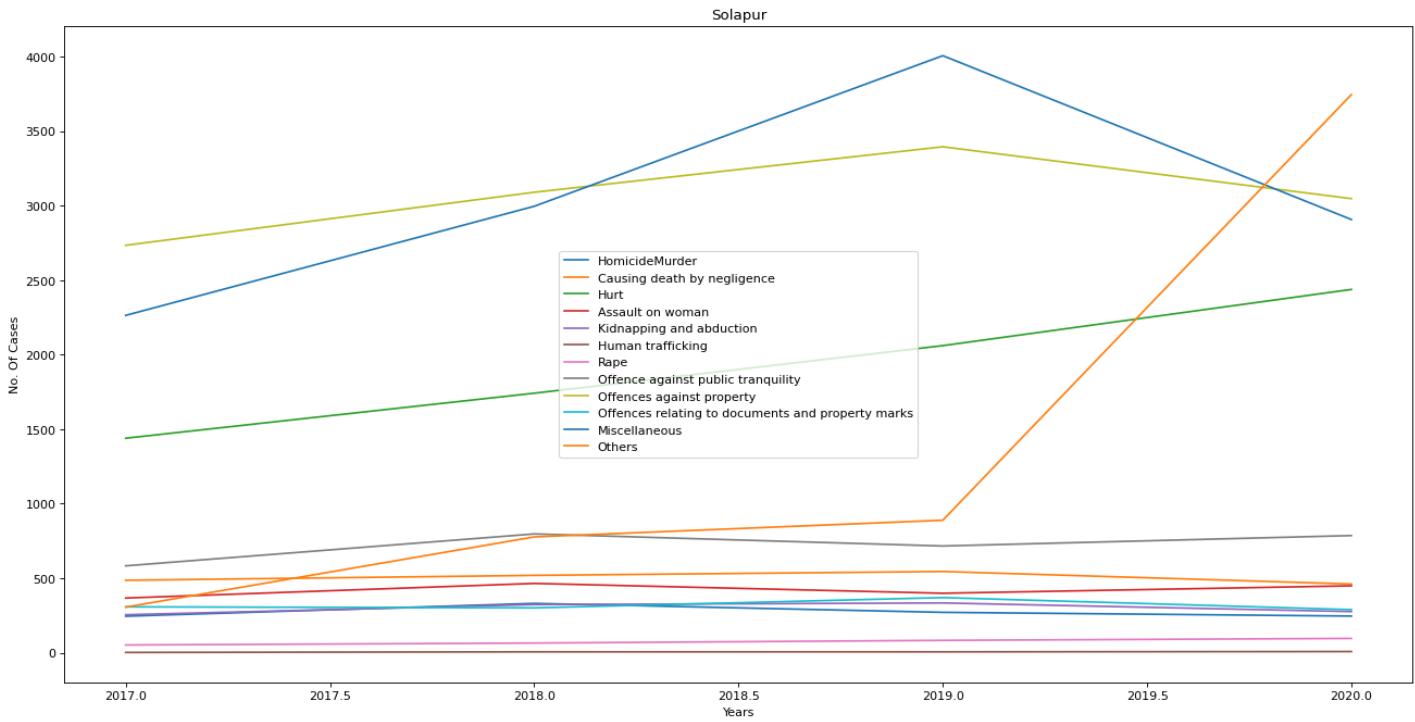
Parbhani

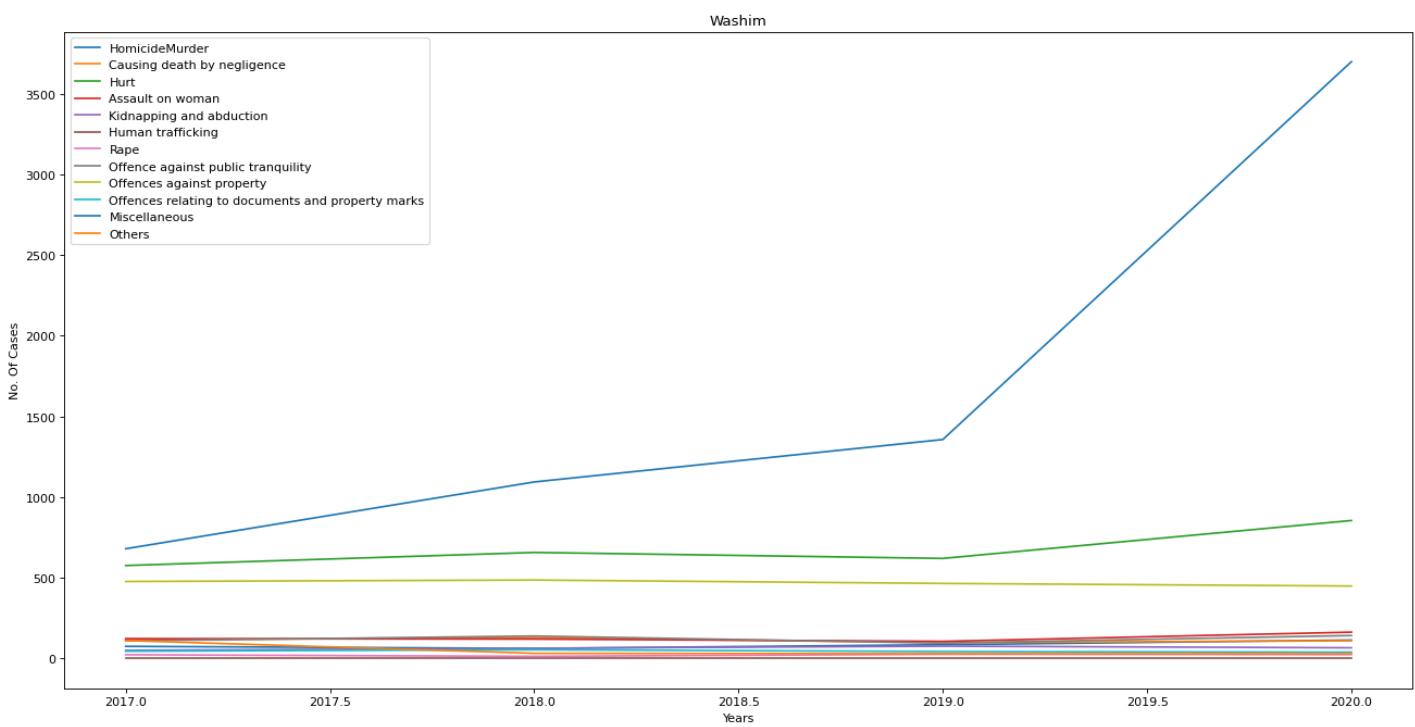
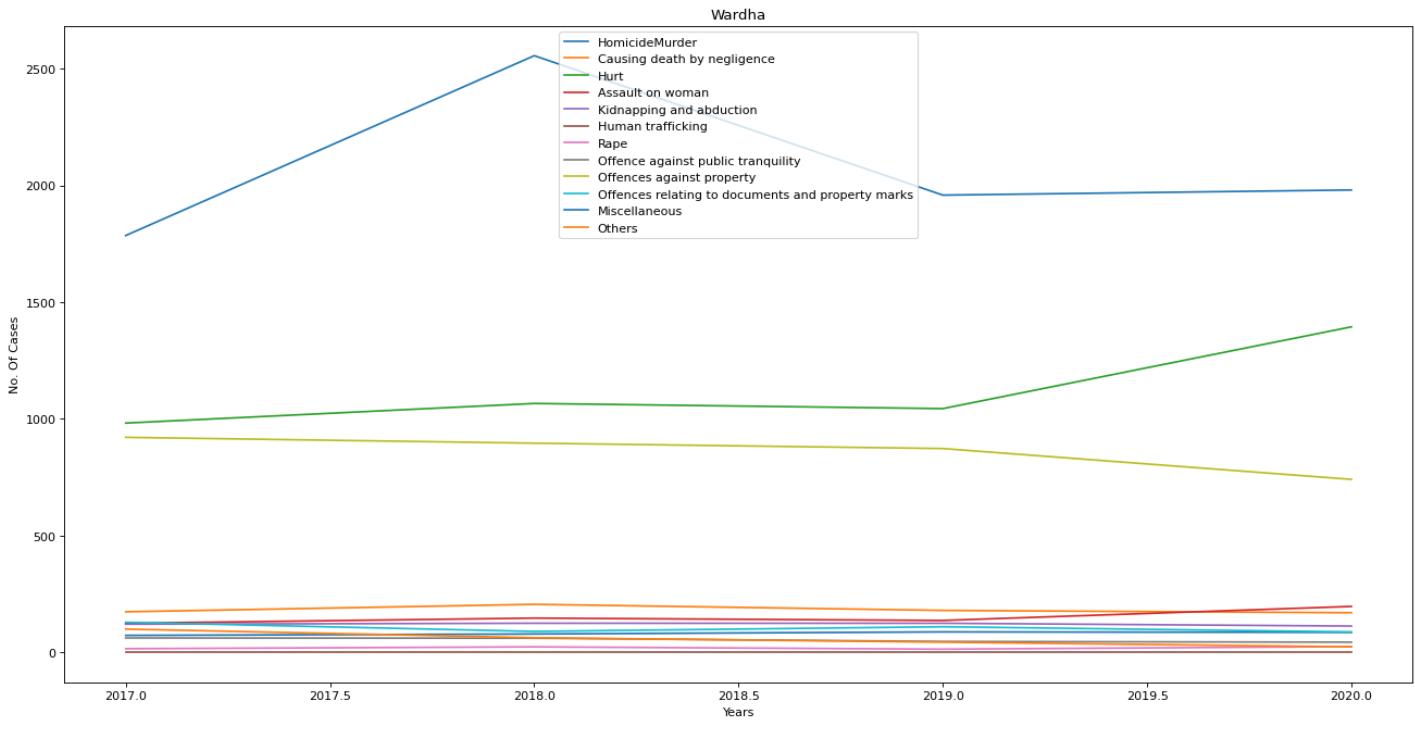


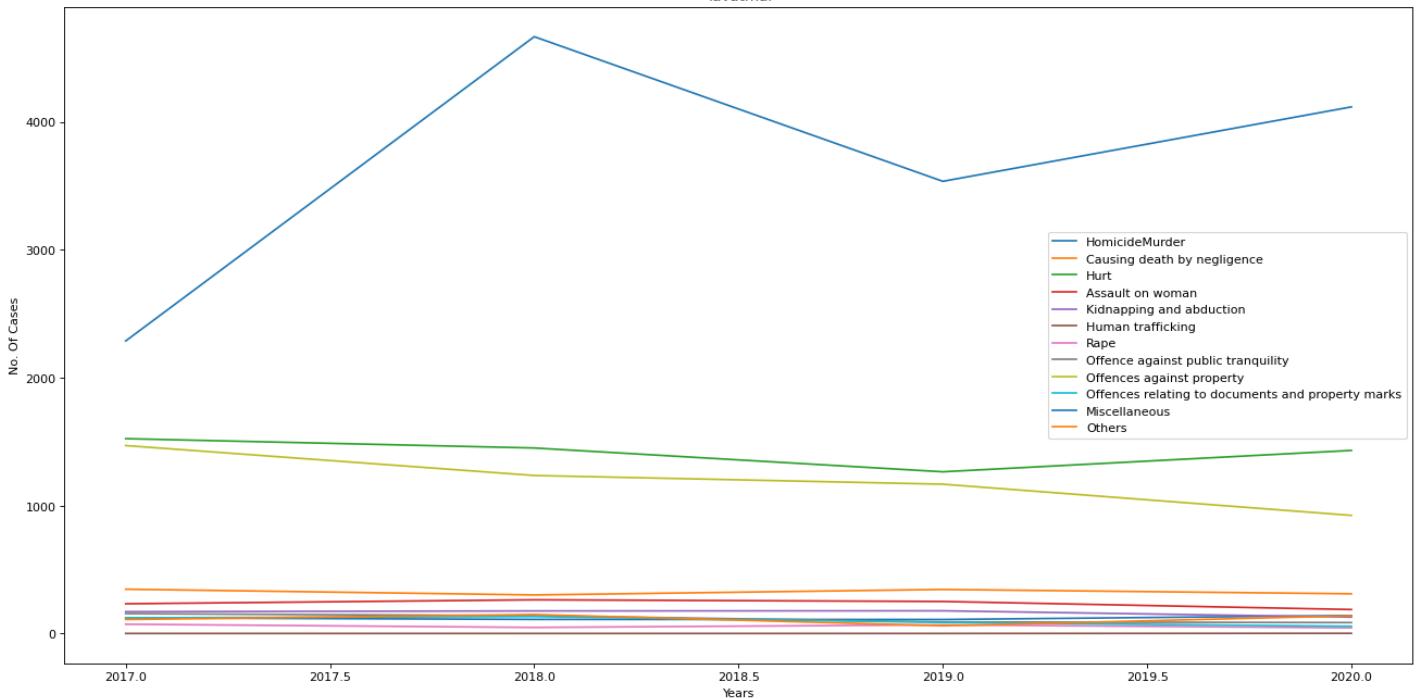






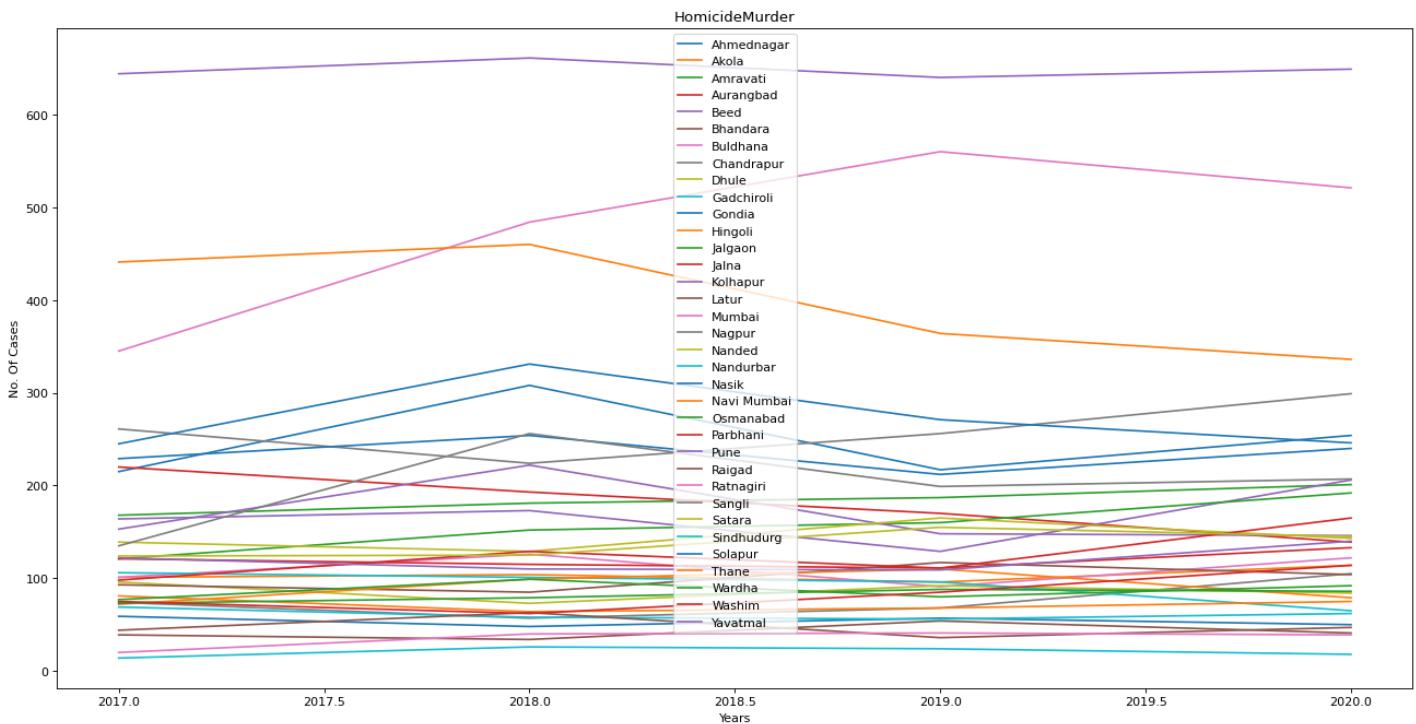


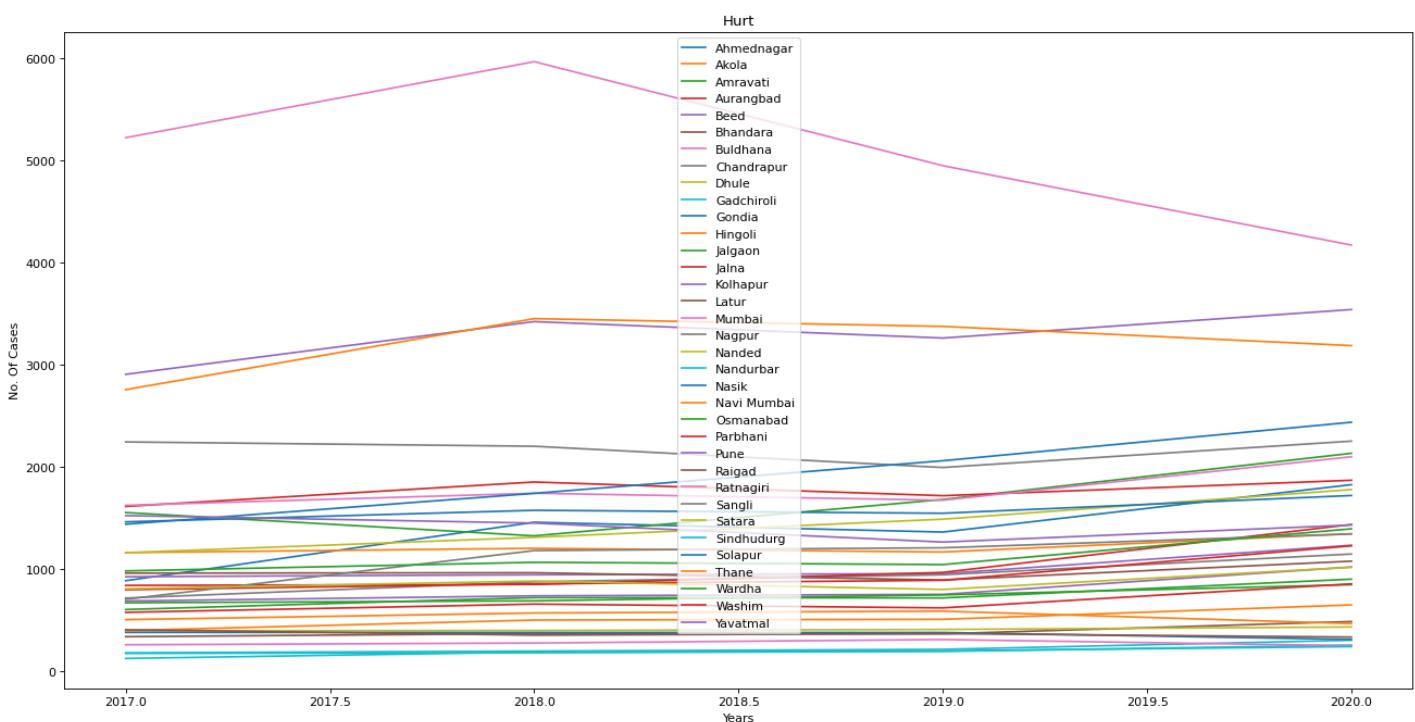
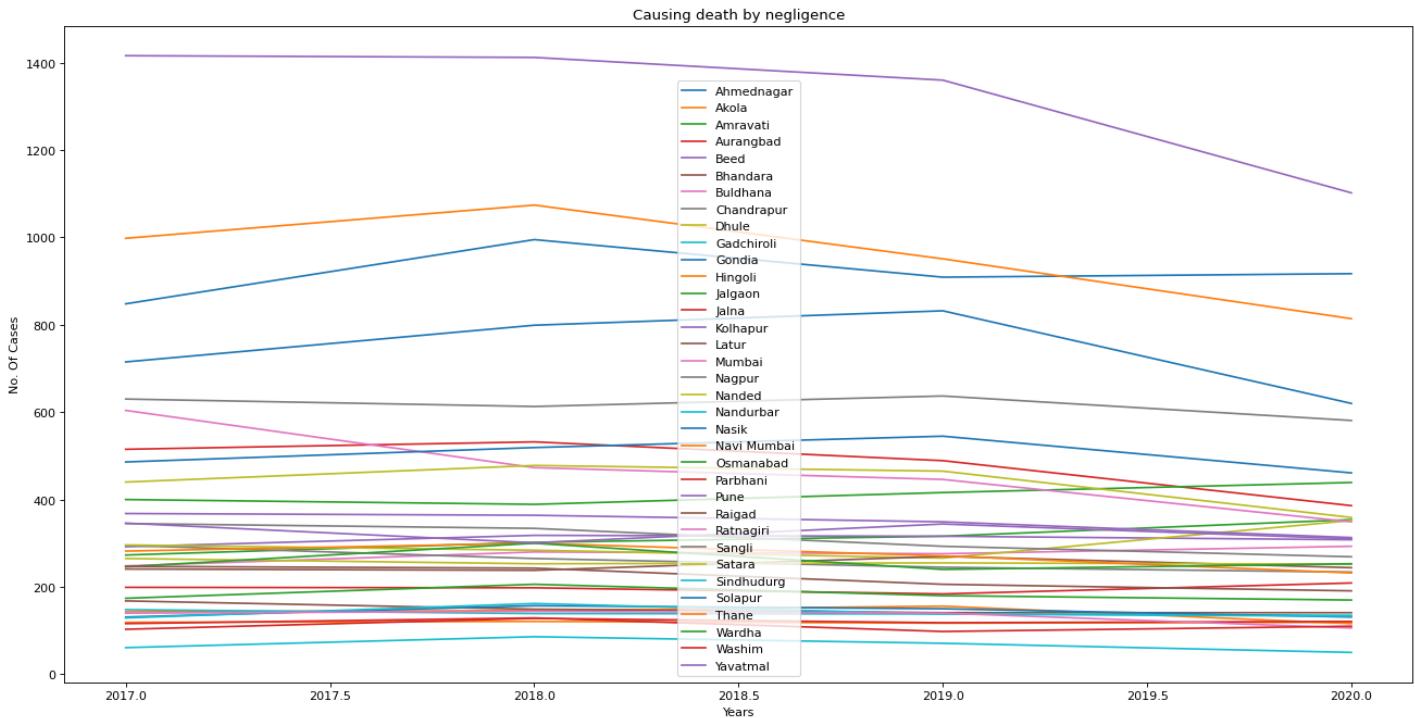


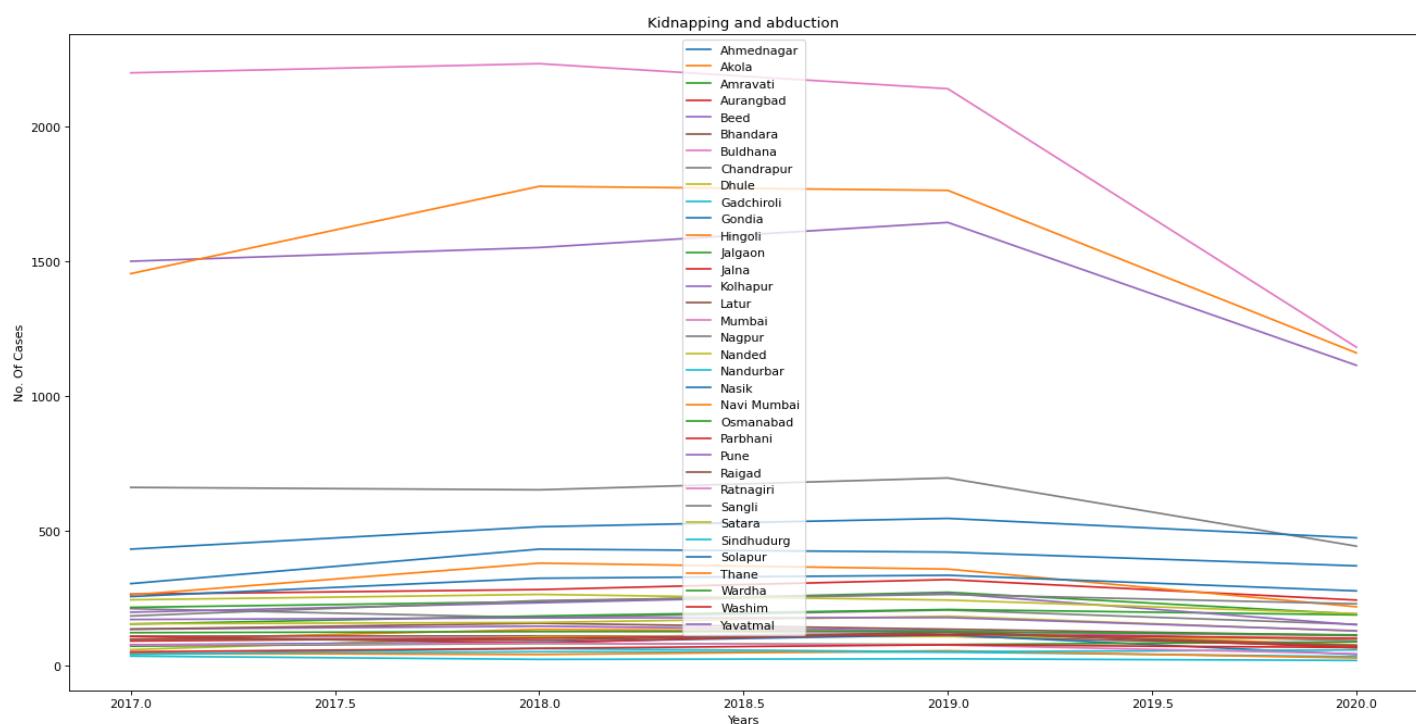
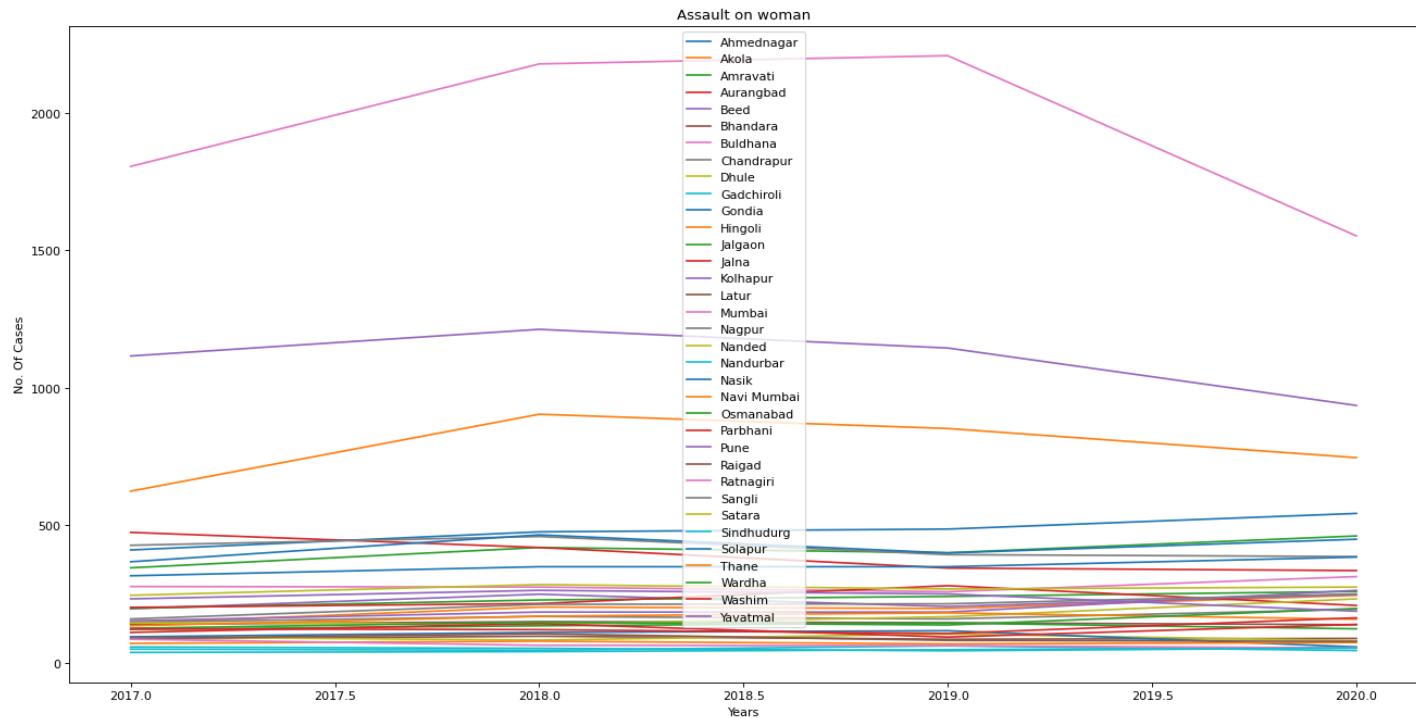


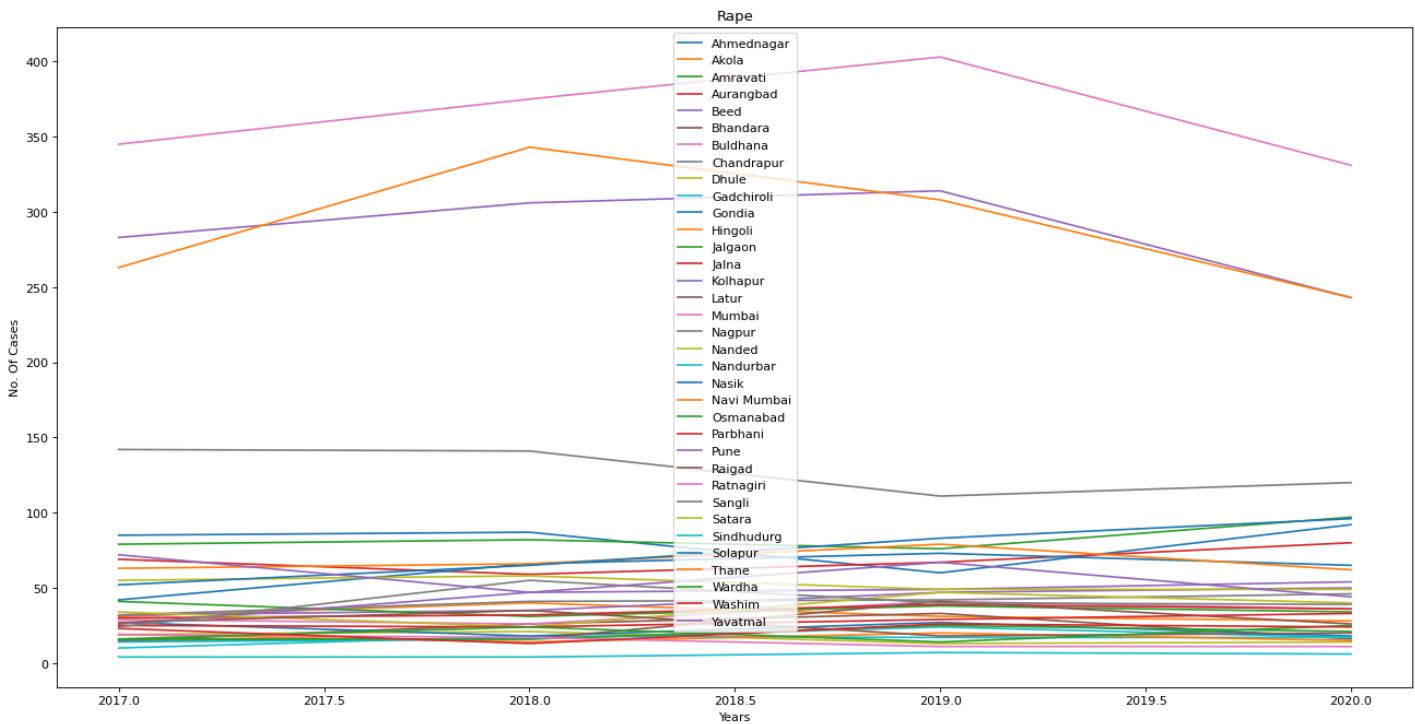
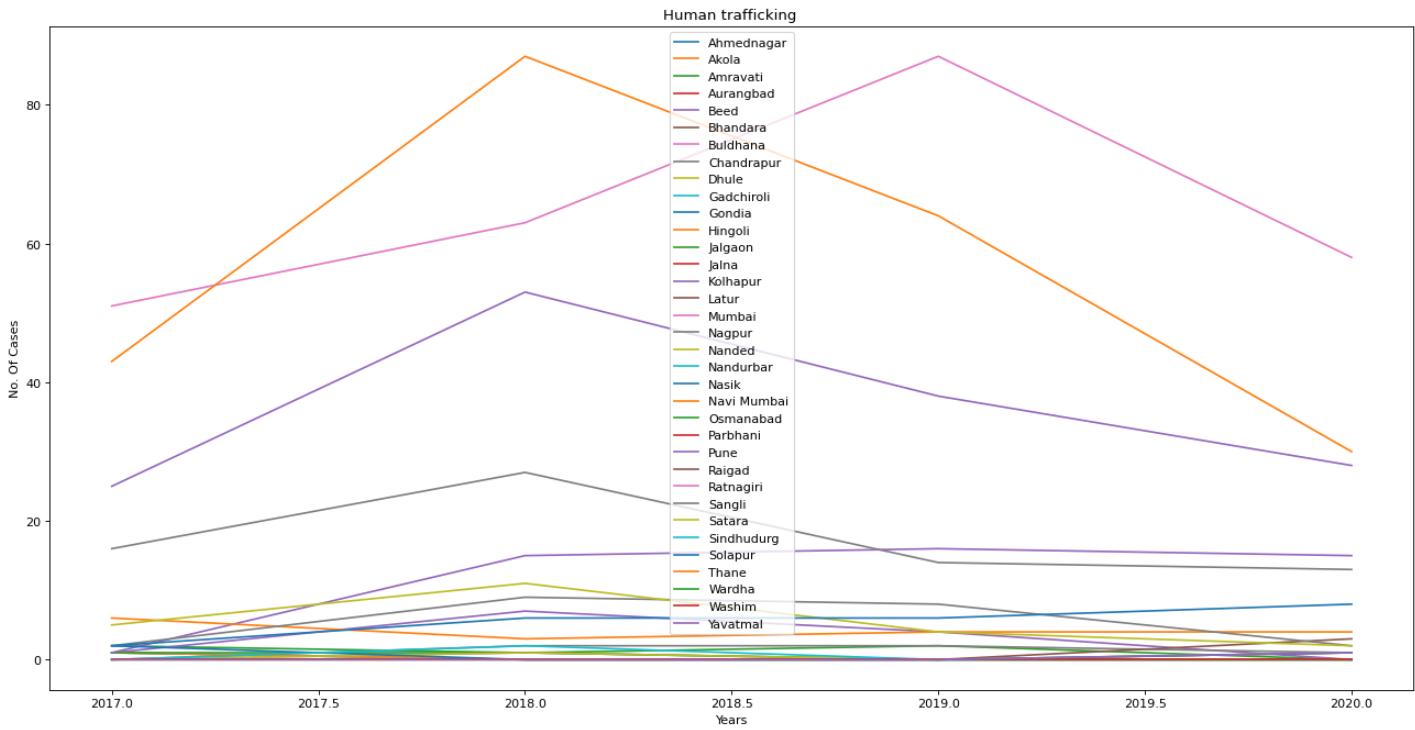
In [32]:

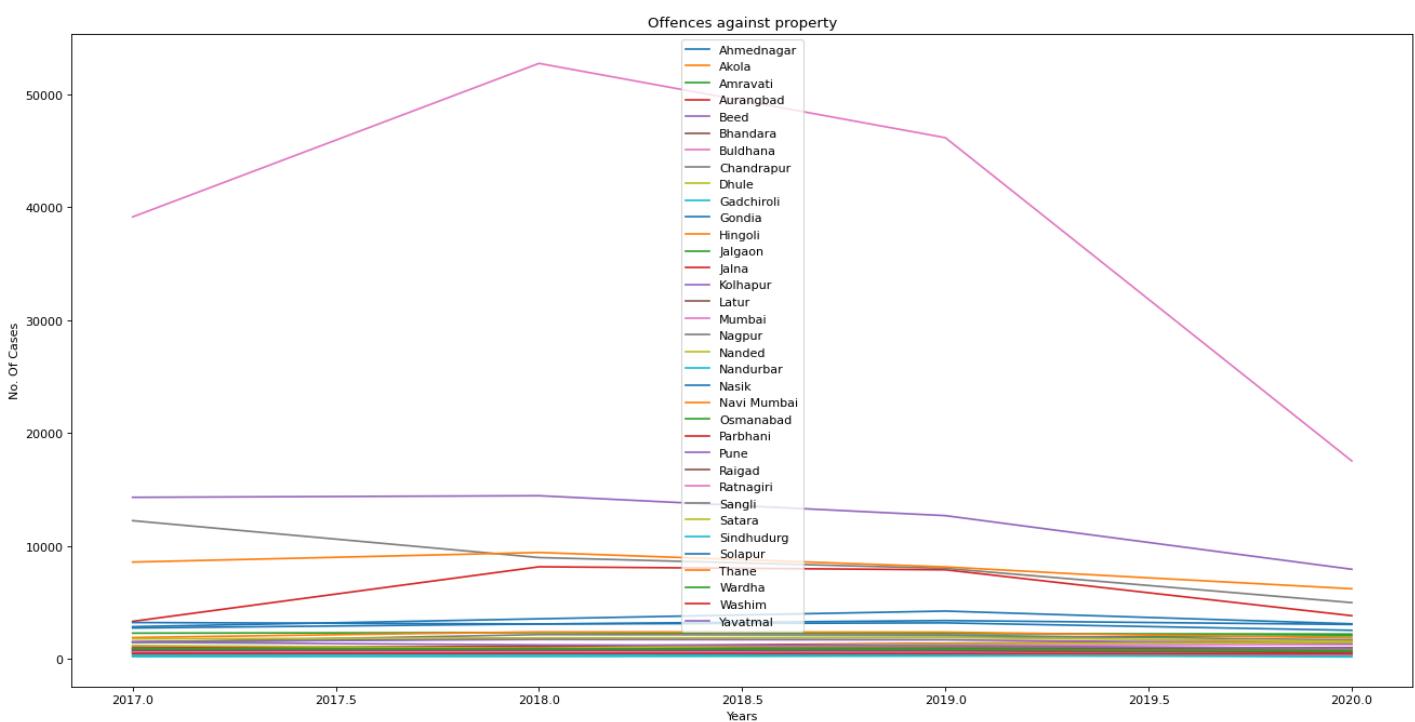
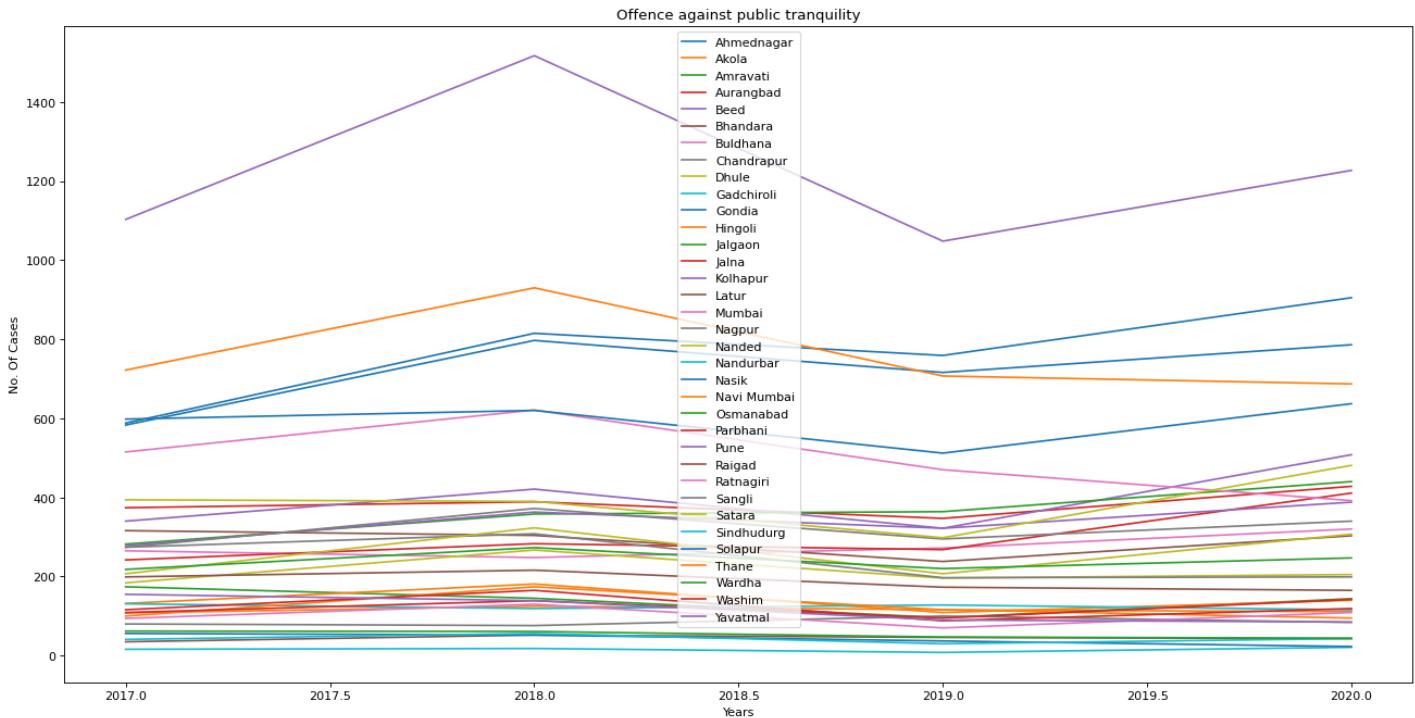
```
#Crime vs Time(4yrs) for all district
for Crime in Crimes:
    for district in Districts:
        temp_df=df[ (df['District']==district)&(df['Crime']==Crime) ]
        N_cases=[temp_df[c].values[0] for c in years]
        plt.plot(years,N_cases)
        plt.legend(Districts)
```

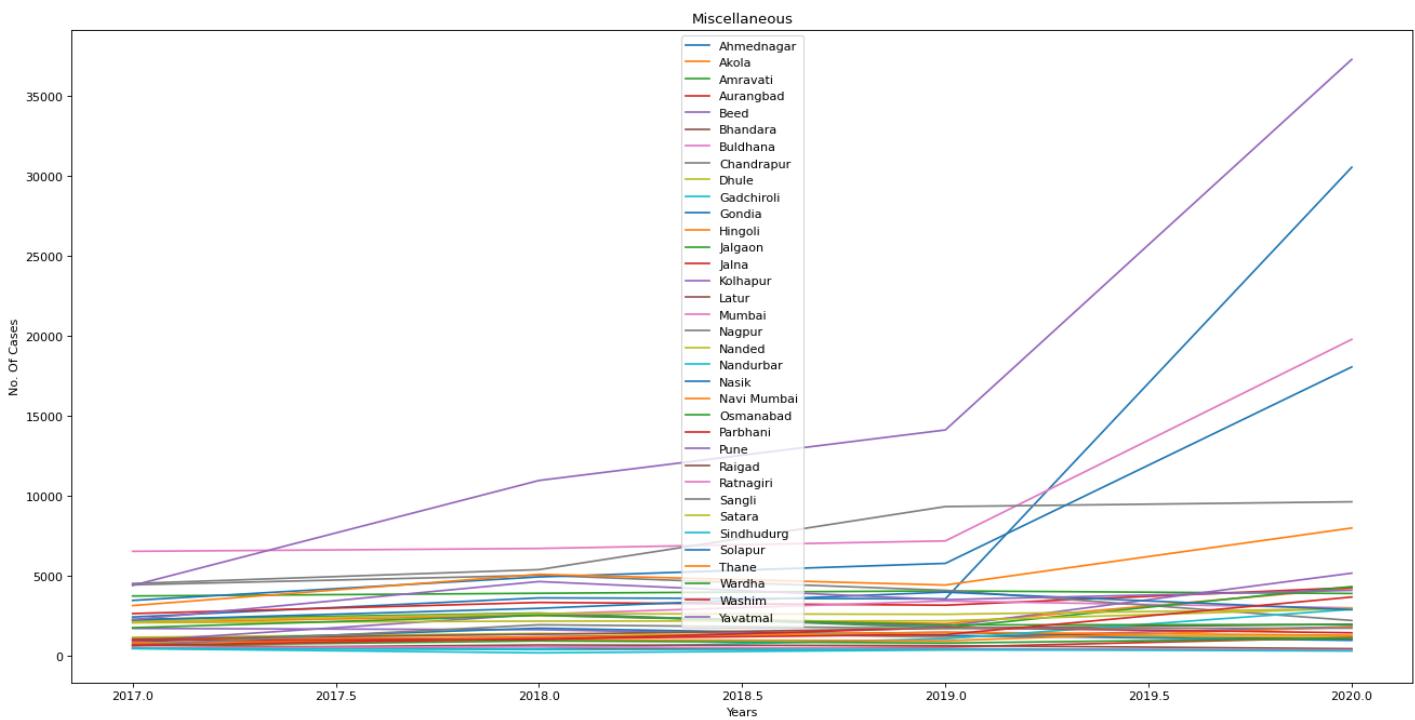
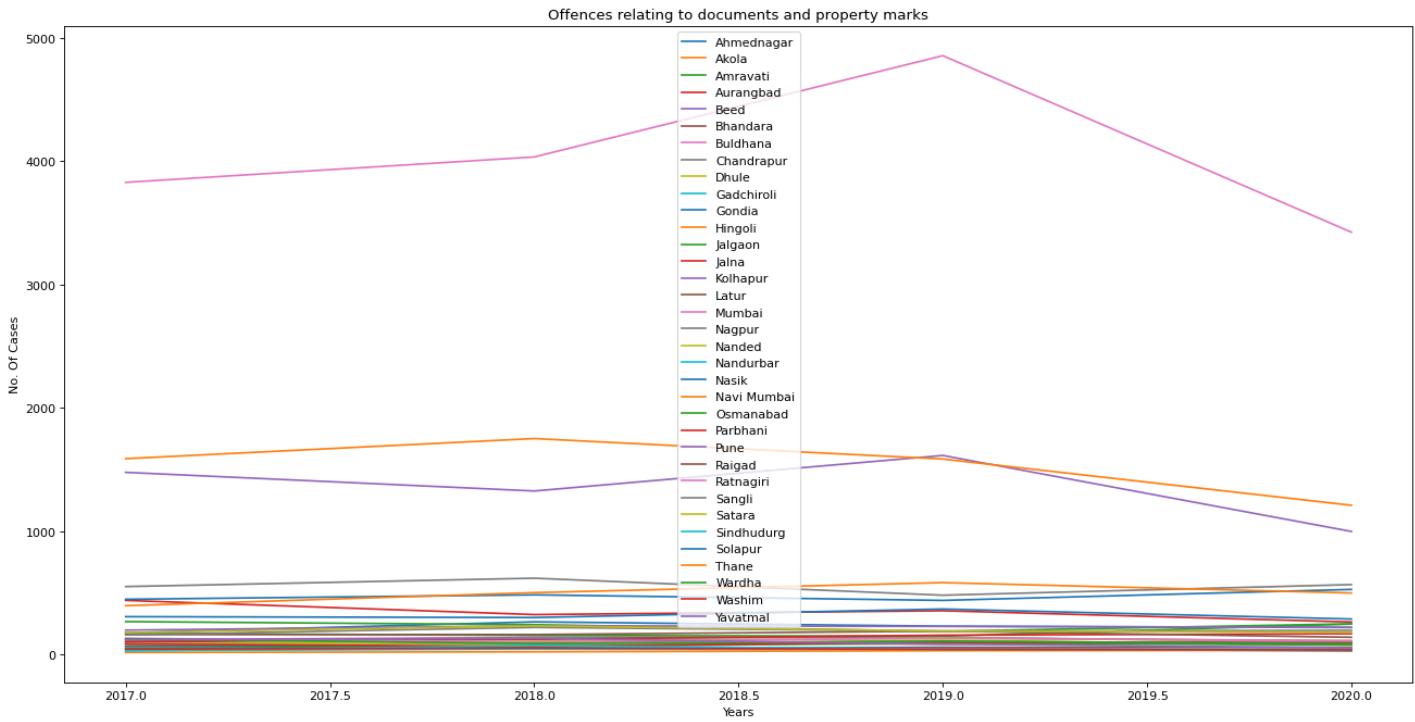


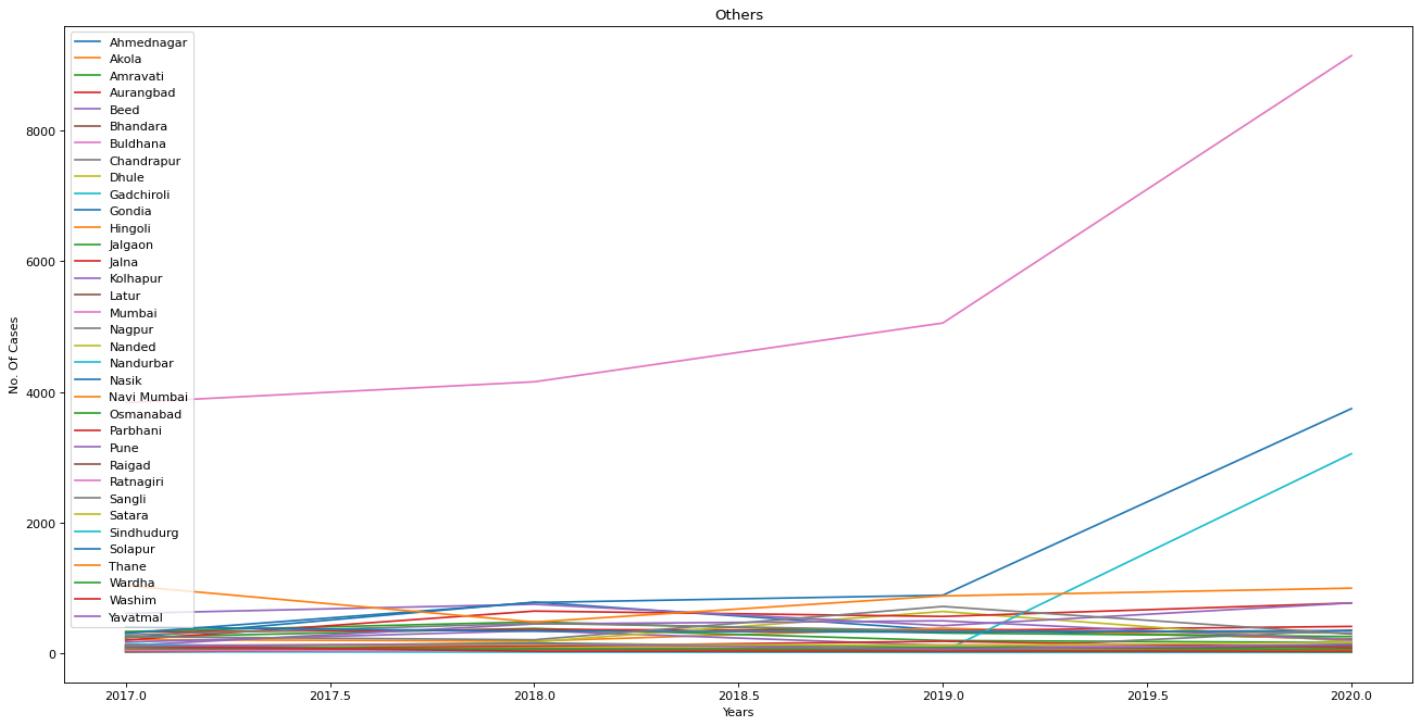












In [ ]: