Juvenile Crime in India

A Data Visualization Project by Prakhar Mittal

The Dataset

Name: Crime in India

Link: https://www.kaggle.com/rajanand/crime-in-india

Publisher: National Crime Records Bureau of India

License: Government Data Open License - India

Contains complete information on juvenile crimes that took place in India from 2001 to 2010.

- Number of cases by type of crime, state, and year
- Age, gender, educational, financial, family, criminal history background of offenders
- Methods of disposal of juveniles after arrest

Why did I choose this dataset?

- Learnt about increasing juvenile crime in a political science class
- Serious crimes carried out by individuals that are almost 18 years old have been the subject of ample legal debates in the media in recent months
- Want to evaluate the data published to see how it matches up with the perception of the common people to clear misconceptions or stereotypes
- This particular dataset is published directly by the Crime Bureau and is therefore, the best resource on this topic
- There are no null values and tables are neatly arranged by State and Year pairs

Research Questions

- Which states are the most notorious for juvenile crime? How do we identify the states that are in the most dire need of reform?
- How are the cases distributed by age group, gender, and by type of crime?
- How has the number of cases changed over the ten year period? Has this change been uniform across demographic and state boundaries?
- Is there a relation between educational, economic, and family background of offenders?
- What are the common measures taken by the police after a juvenile is arrested?
- What are the most prevalent crimes? Is there a distinction between the crimes committed by boys and girls?

Data Transformation

The dataset included six files with 350 rows (35 states x 10 years), columns depicting different demographic / background categories, and cells containing the corresponding number of cases.

1. Imported csv files into pandas dataframes in Python
2. Combined dataframes using State - Year pairs as key
3. Created pivot tables using State - Year pairs as key and Total Cases as value
4. Summed columns after grouping by State or Year to compare across different states or years respectively
5. Used logical_not and isin functions from numpy to select specific crimes of interest for a more in-depth analysis

Sample Transformation

Before (top): Full table with 350 rows one for every State - Year pair.

Process: Group by Year and sum rows having same year.

After (below): Table with ten rows, one for each year.

I used this modified table to draw line + bar plots showing how cases varied over the decade sorted by demographic category.

	State	Year	B7-12	G7-12	B12-16	G12-16	B16-18	G16-18	BTotal	GTotal	TotalApprehended
	Andhra Pradesh	2001	116	4	652	5	669	119	1437	128	1565
	Andhra Pradesh	2002	127	5	554	41	1038	319	1719	365	2084
	Andhra Pradesh	2003	47	0	464	32	1895	235	2406	267	2673
	Andhra Pradesh	2004	67	2	465	33	1427	202	1959	237	2196
	Andhra Pradesh	2005	73	1	596	24	1614	120	2283	145	2428
	...	...	...	...	...	...	...	...	...	...	...
	Puducherry	2006	0	0	17	0	19	1	36	1	37
	Puducherry	2007	0	0	19	0	39	4	58	4	62
	Puducherry	2008	0	0	13	0	42	0	55	0	55
	Puducherry	2009	0	0	36	3	41	1	77	4	81
	Puducherry	2010	0	0	9	0	36	4	45	4	49

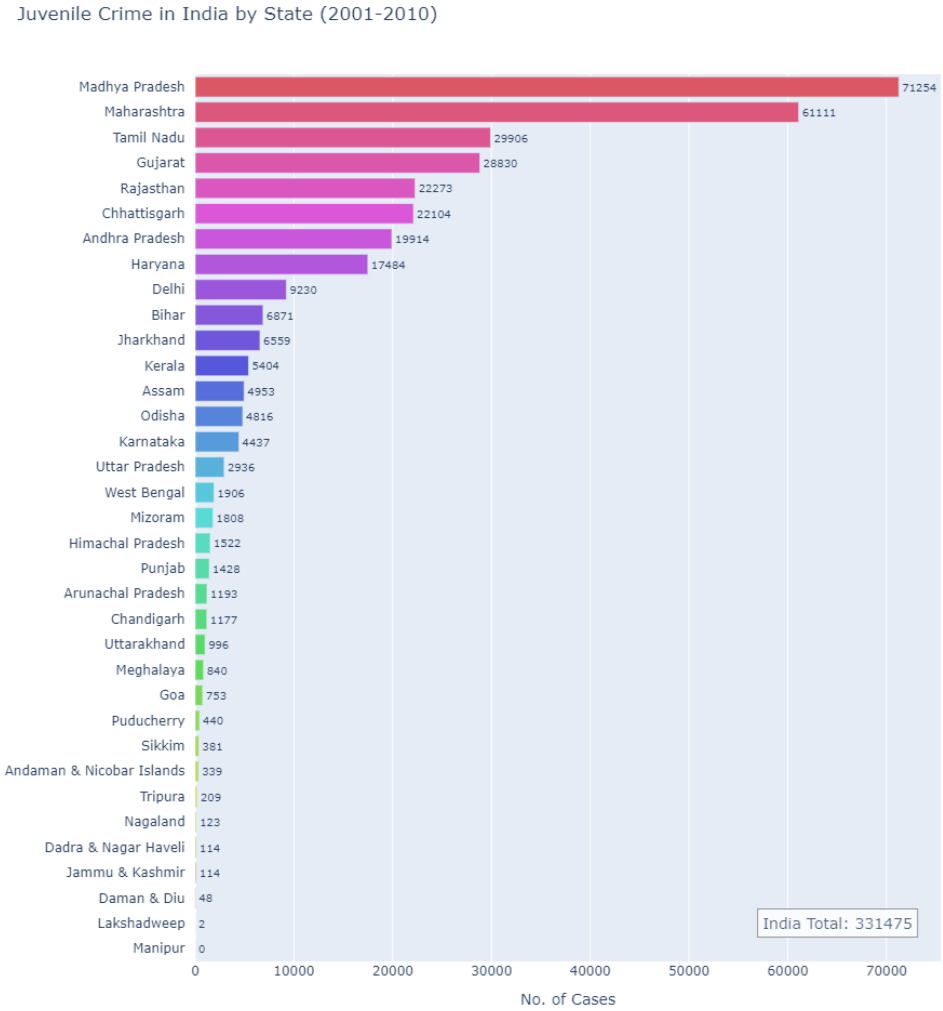
	Year	B7-12	G7-12	B12-16	G12-16	B16-18	G16-18	BTotal	GTotal	TotalApprehended
0	2001	3591	105	12131	598	15573	1630	31295	2333	33628
1	2002	4221	267	13283	581	16047	1380	33551	2228	35779
2	2003	3414	170	11053	634	16518	1531	30985	2335	33320
3	2004	1966	141	11627	788	15285	1136	28878	2065	30943
4	2005	1526	119	12332	758	16748	1198	30606	2075	32681
5	2006	1508	87	11883	652	16984	1031	30375	1770	32145
6	2007	1338	122	11537	577	19796	1157	32671	1856	34527
7	2008	1146	135	11687	585	19962	992	32795	1712	34507
8	2009	1015	118	10079	662	20456	1312	31550	2092	33642
9	2010	832	95	9597	526	18334	919	28763	1540	30303

Data Visualization

I used Plotly to make detailed figures depicting specific information:

- Bar plots and line charts to depict trends across years
- Horizontal bar plots to compare cases across states
- Pie and donut charts to compare the different proportions of categorical data such as crime, demographic group, and method of disposal

In order to differentiate between the thirty five states on bar plots, I used Seaborn's HLS color palette since I did not find any Plotly palettes that had enough distinguishable colors. I also used Matplotlib and Seaborn to make preliminary charts for data exploration before finalizing the best ones in Plotly.



1. States by Cases

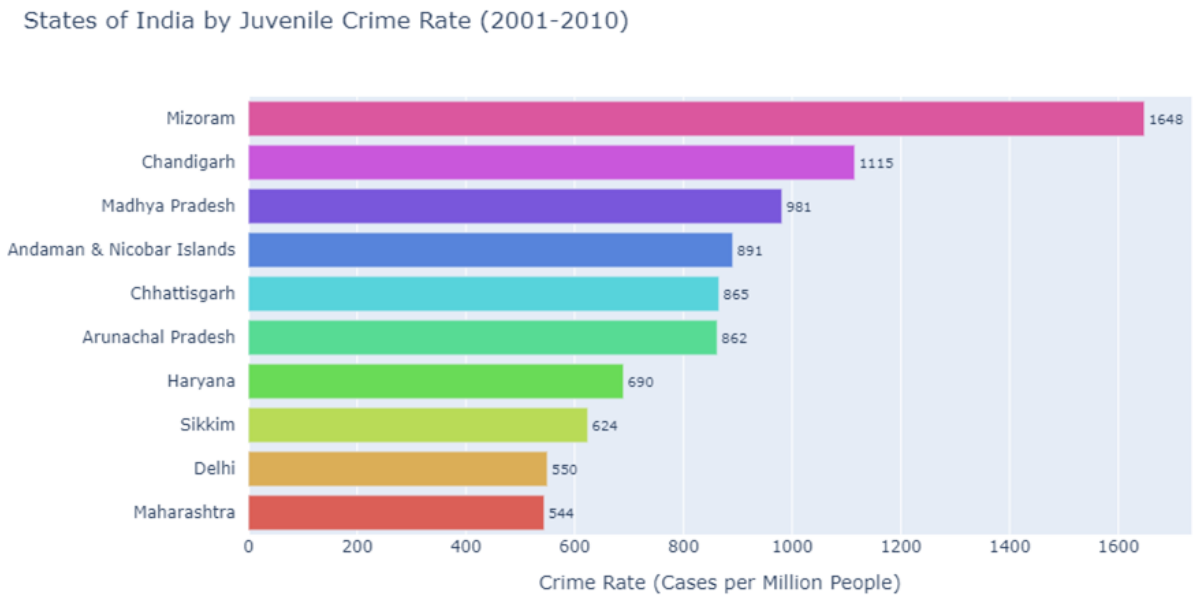
- A horizontal bar plot depicting total number of juvenile crime cases in each state from 2001 - 2010.
- Madhya Pradesh and Maharashtra have a significantly larger tally than any other state.
- Uttar Pradesh - the most populous and stereotypically the most unsafe state - has a relatively low number of arrests.

2. States by Trend

- Subplots showing number of crimes by year for the worst 8 states.
- Gujarat, Haryana, and Tamil Nadu seem to have a downward trend.
- Maharashtra, Chhattisgarh and Rajasthan (barring the anomalies of 2001 and 2005) seem to have an upward trend.
- Andhra Pradesh and Madhya Pradesh show fluctuation in the given time period.

3. States by Crime Rate

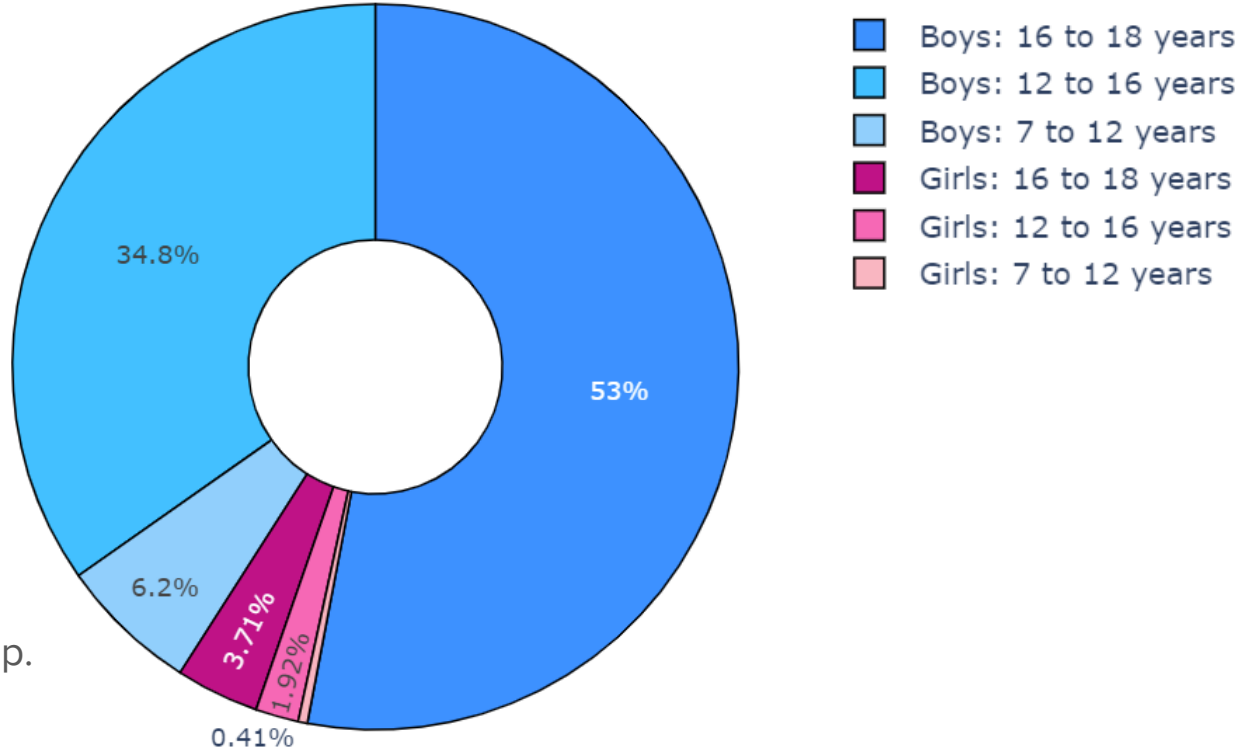
- Crime rate is defined as the number of cases per million people. It is a useful metric when comparing states with varying populations.

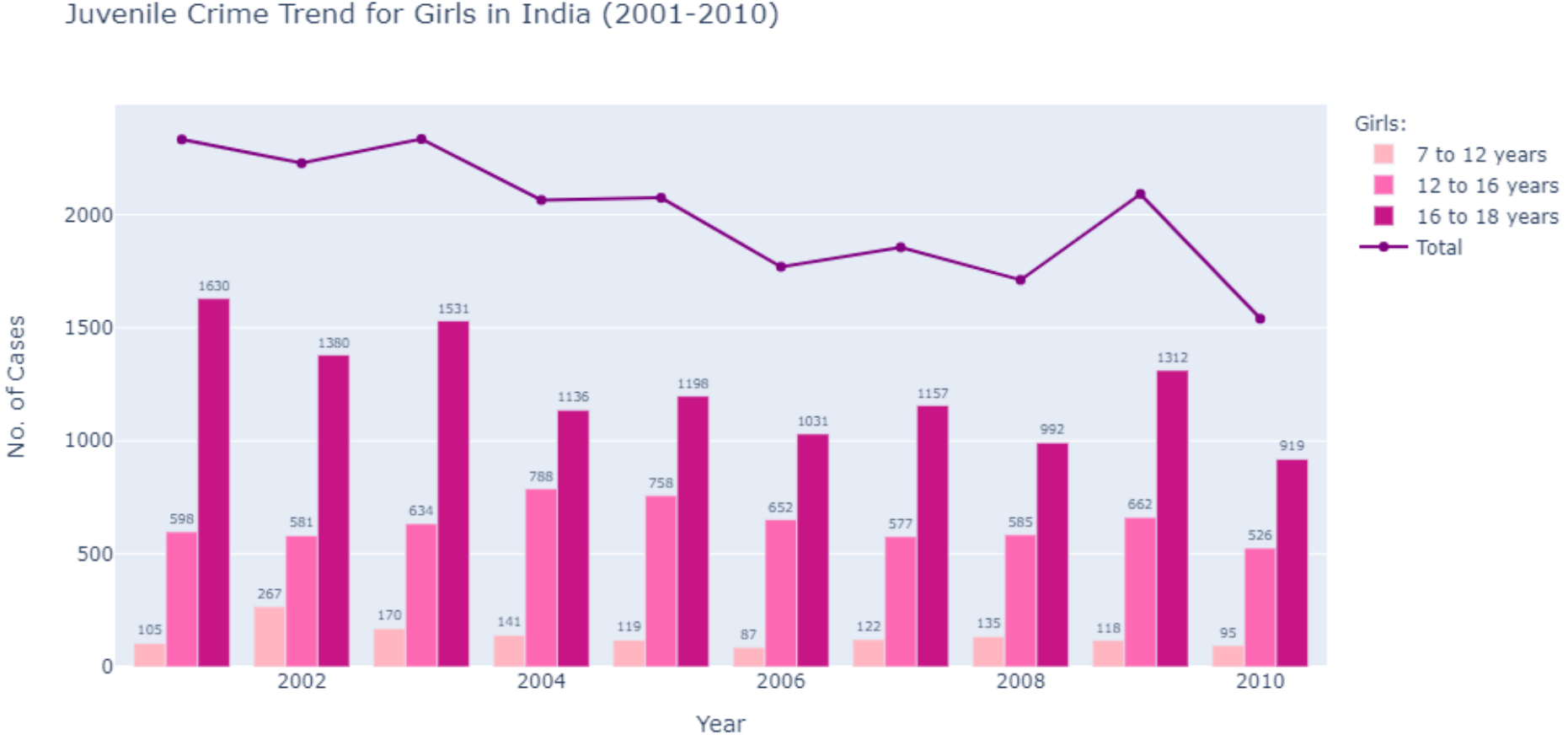


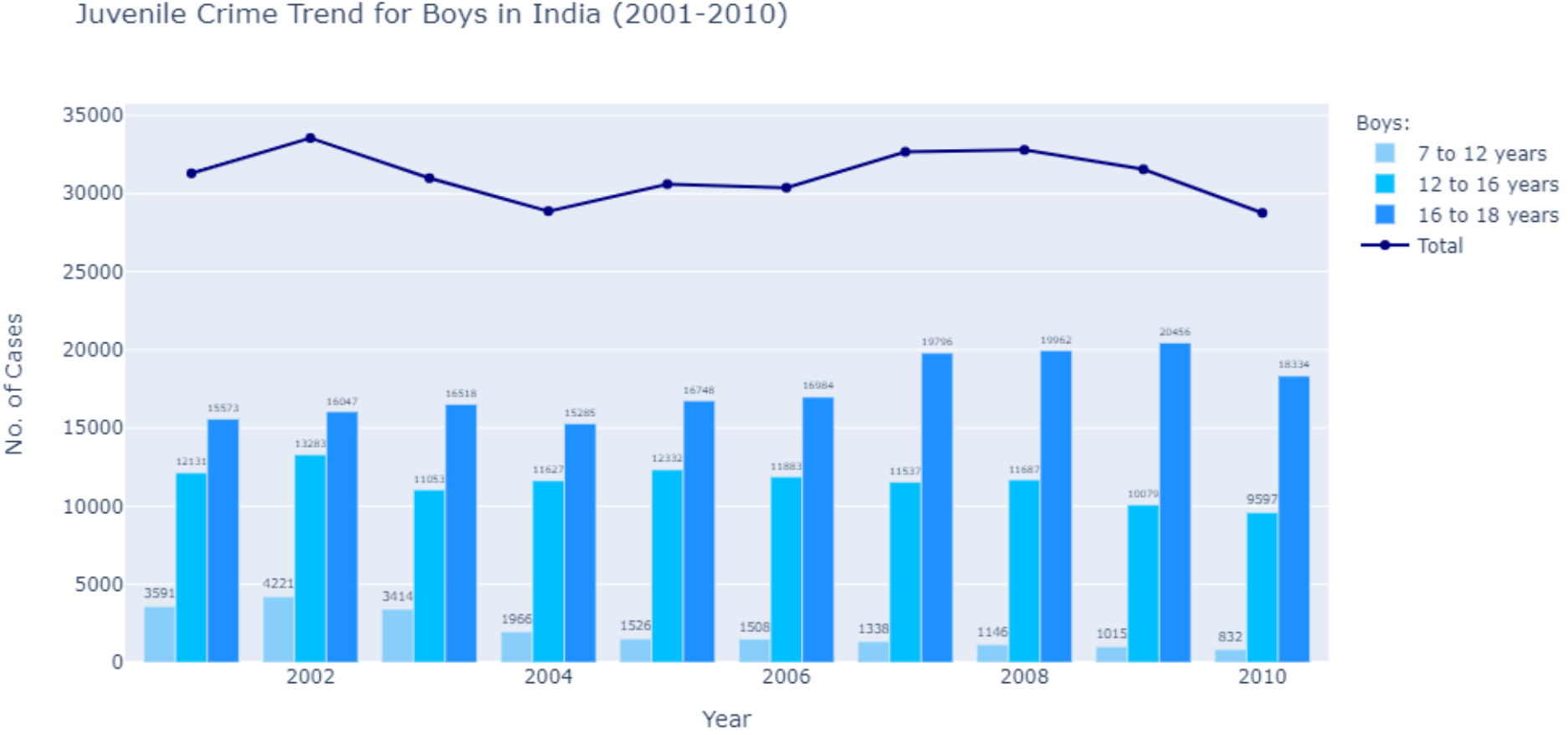
- Mizoram has the highest crime rate at 1648 cas./mil.
- Chhattisgarh, Maharashtra, Haryana, Delhi, and Madhya Pradesh, all large states that are in the news for being unsafe, have some of the highest crime rates in India.

4. Demographics

- 94% of offenders are boys with only 6% of them being girls.
- Unsurprisingly, juvenile crime cases rise with age with 56.7% offenders being in the 16 to 18 years age group.

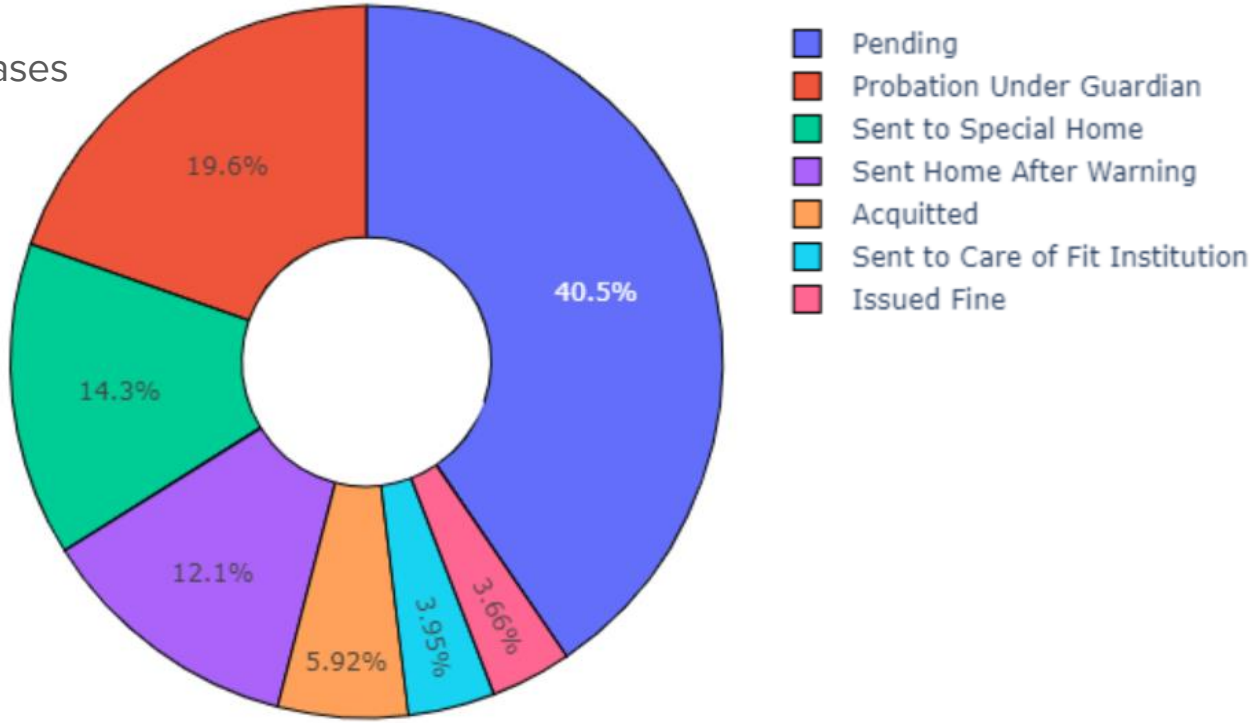



Juvenile Crime Trend for Girls in India (2001-2010)
No. of Cases
Year
Girls:
7 to 12 years
12 to 16 years
16 to 18 years
Total
0
500
1000
1500
2000
2002
2004
2006
2008
2010
105
598
1630
267
581
1380
170
634
1531
141
788
1136
119
758
1198
87
652
1031
122
577
1157
135
585
992
118
662
1312
95
526
919


Juvenile Crime Trend for Boys in India (2001-2010)
No. of Cases
Year
Boys:
7 to 12 years
12 to 16 years
16 to 18 years
Total
35000
30000
25000
20000
15000
10000
5000
0
2002
2004
2006
2008
2010
3591
12131
15573
4221
13283
16047
3414
11053
16518
1966
11627
15285
1526
12332
16748
1508
11883
16984
1338
11537
19796
1146
11687
19962
1015
10079
20456
832
9597
18334

5. Disposal Methods

- Around 40.5% of the cases remain pending, which shows the inefficiency of the judiciary.
- Most offenders are sent home (12.1%) or to a guardian (19.6%).
- Around 14.3% are sent to a juvenile home.



6. Background

- The correlation between poverty, lack of education, and high crime rate is crystal clear.
- The proportion of illiterates and homeless offenders is much more than their share in the general population.
- Around 10.3% arrests involve repeat offenders.

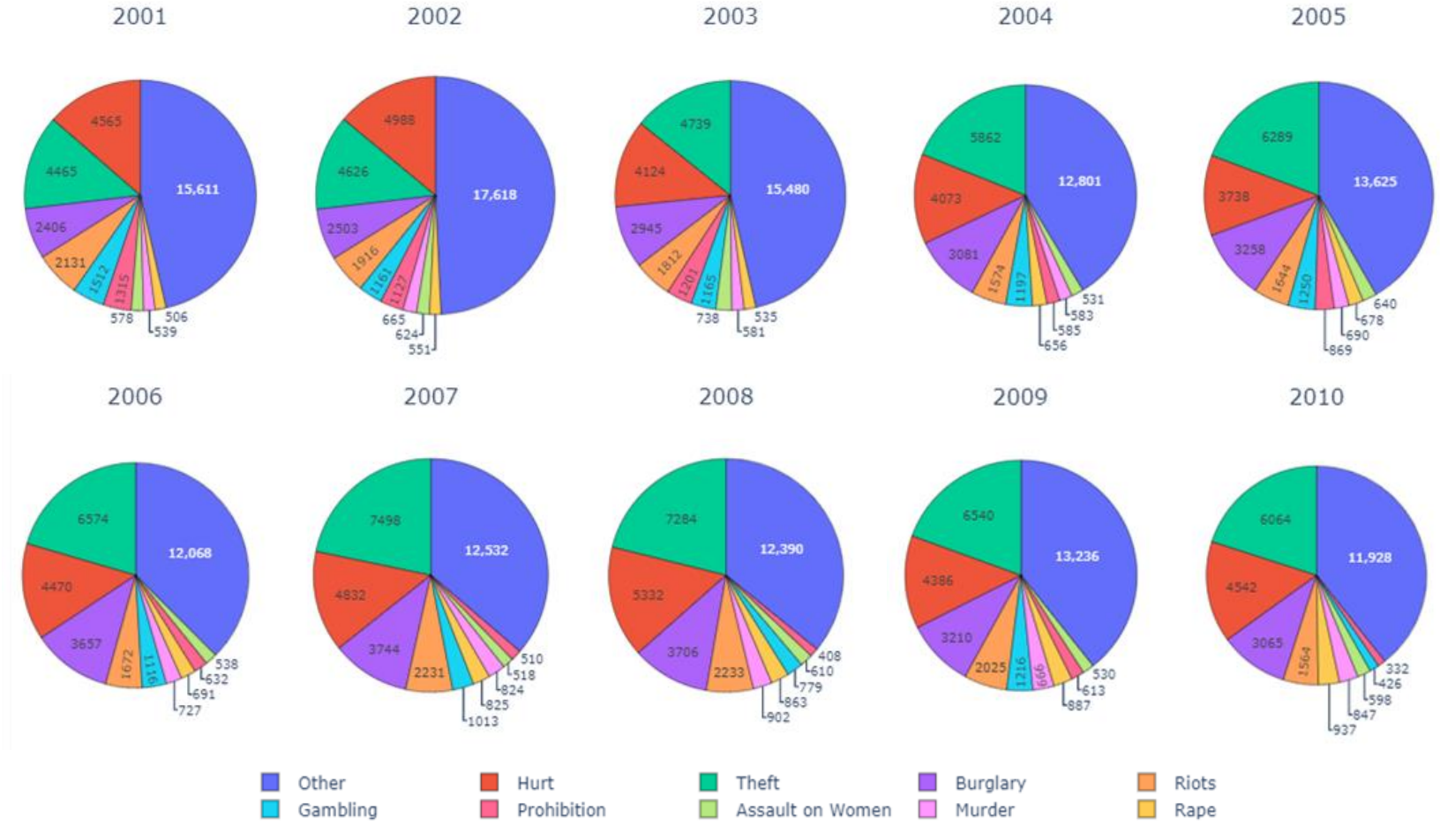
7. Crimes by Cases

- Hurt and Theft are the two most common crimes committed by juveniles making up around a quarter to a third of all cases.
- Burglary and Riots are the next in the list of popular juvenile crimes.
- The number of rapes committed by juvenile offenders has increased every single year from 506 cases (9th most prevalent crime) in 2001 to 937 (5th most prevalent) in 2010.

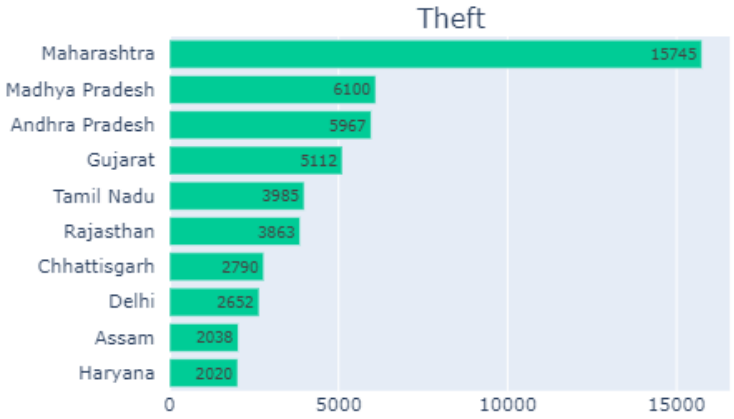
Crime	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Assault on Women	578	624	738	531	640	538	518	610	530	598
Burglary	2406	2503	2945	3081	3258	3657	3744	3706	3210	3065
Gambling	1512	1161	1165	1197	1250	1116	1013	779	1216	426
Hurt	4565	4988	4124	4073	3738	4470	4832	5332	4386	4542
Murder	539	665	581	583	690	727	824	902	999	847
Prohibition	1315	1127	1201	585	869	632	510	408	613	332
Rape	506	551	535	656	678	691	825	863	887	937
Riots	2131	1916	1812	1574	1644	1672	2231	2233	2025	1564
Theft	4465	4626	4739	5862	6289	6574	7498	7284	6540	6064
Other	15611	17618	15480	12801	13625	12068	12532	12390	13236	11928

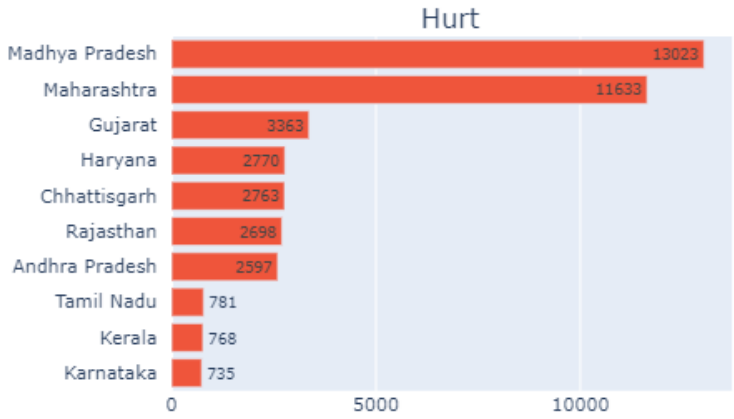
- Murder cases have increased over the decade as well going from 539 cases in 2001 to 847 cases in 2010.

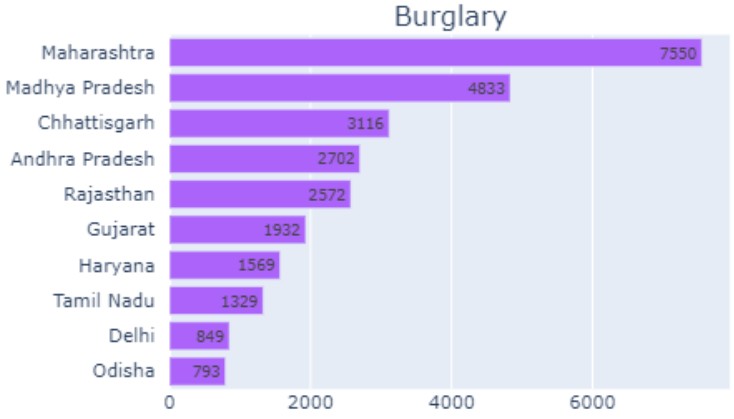
Crimes Committed by Juvenile Offenders in India (2001-2010)

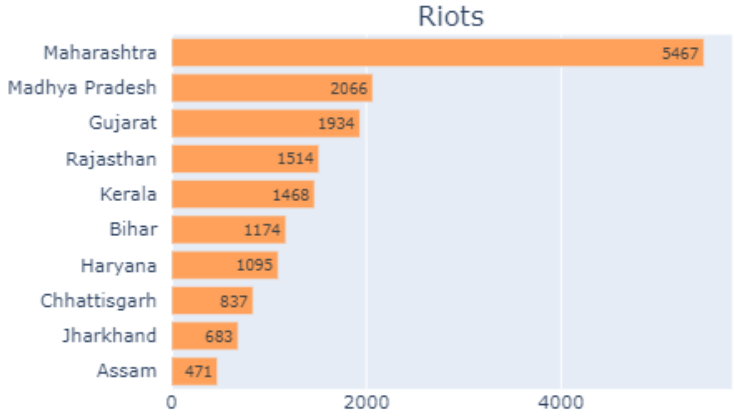


Worst 10 States in India for Common Juvenile Crimes (2001-2010)









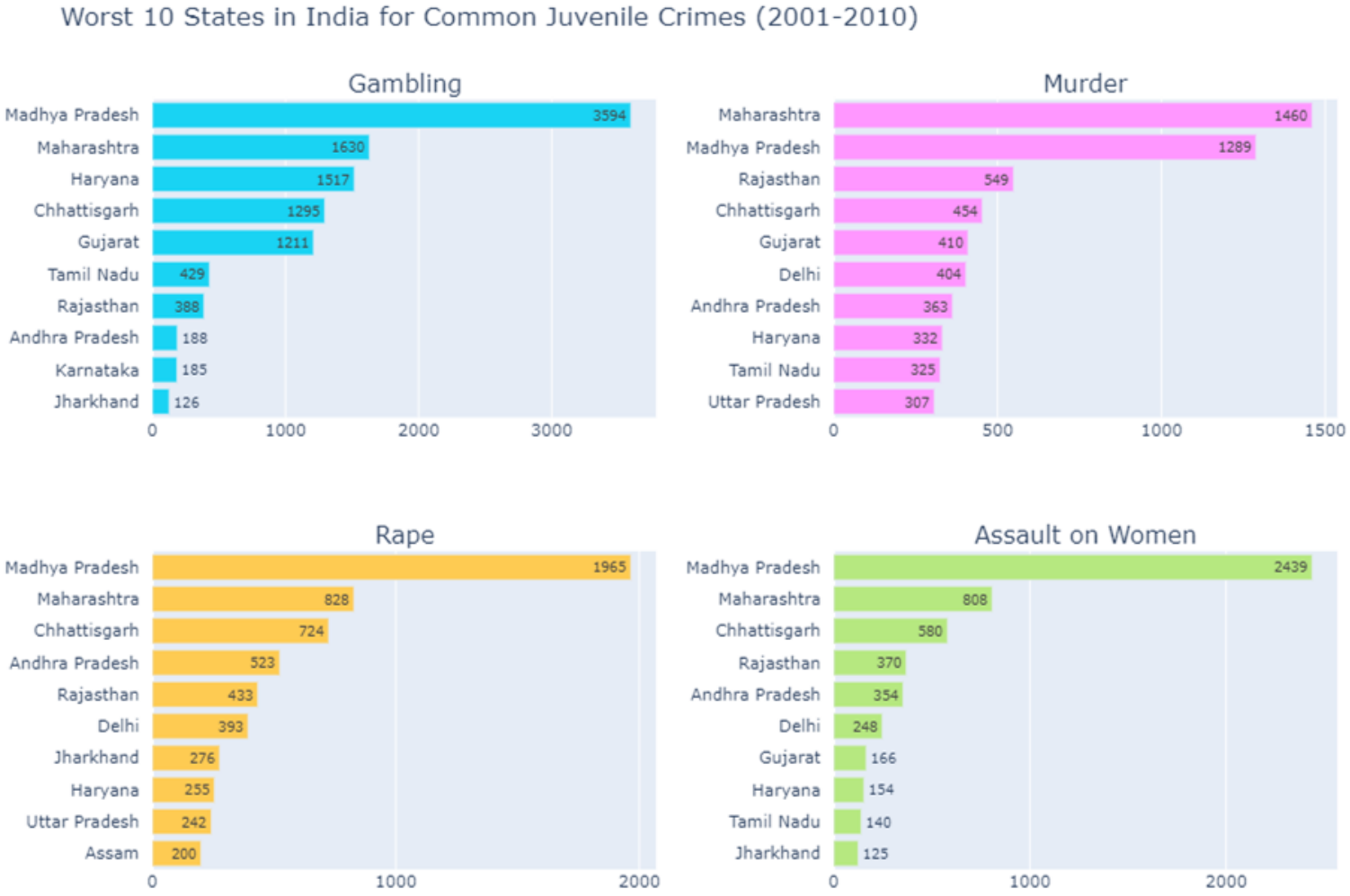
States that are in the worst 10 for a crime but not in the worst 10 overall:

Theft: Assam

Hurt: Karnataka, Kerala

Burglary: Odisha

Riots: Jharkhand, Kerala, Assam



States that are in the worst 10 for a crime but not in the worst 10 overall:

Gambling: Karnataka, Jharkhand

Murder: Uttar Pradesh

Rape: Jharkhand, Assam, Uttar Pradesh

Assault on Women: Jharkhand

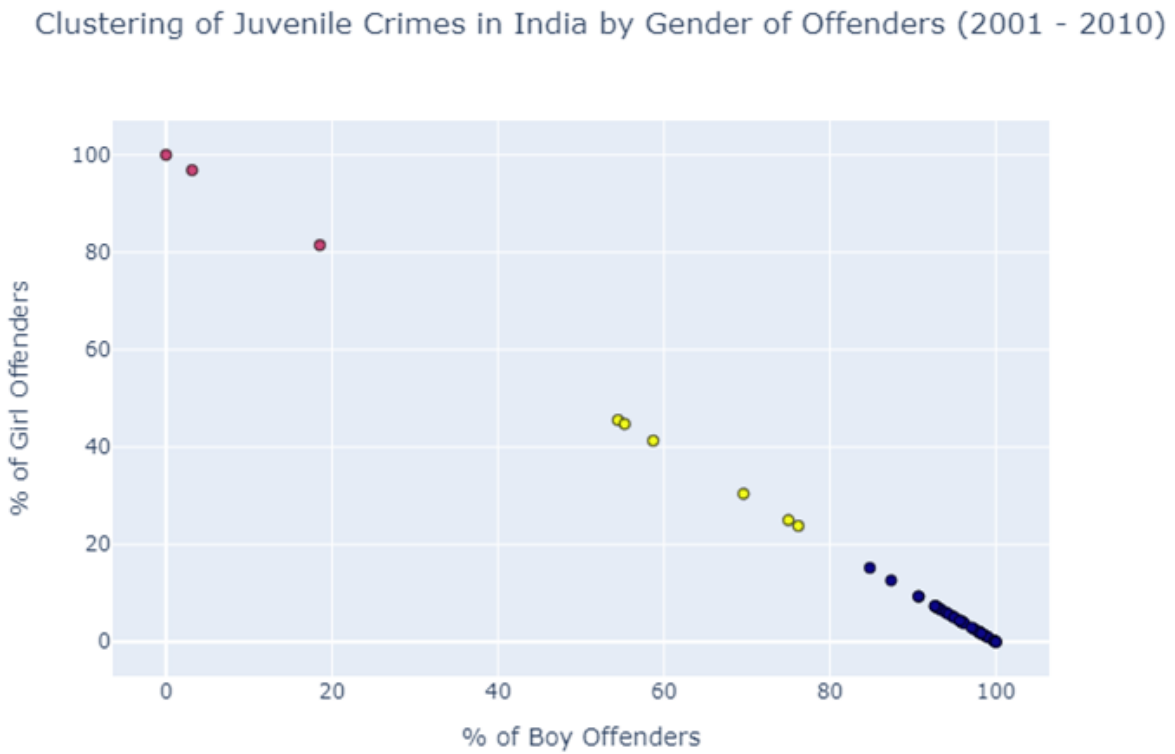
Data Models

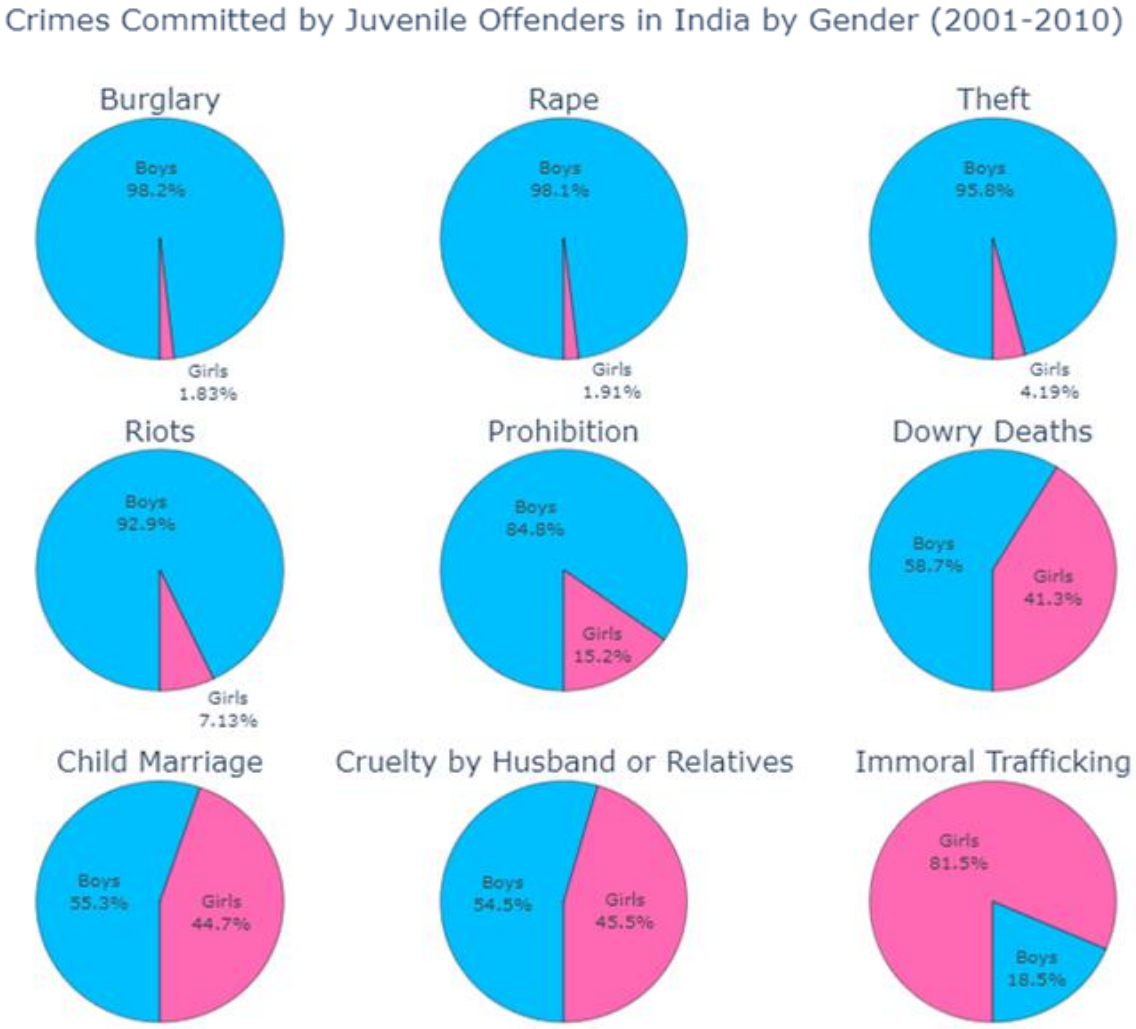
- Since I only had data for 10 years, I figured that using linear regression to predict future cases and crime rates would not be the most accurate.

- Instead, I was interested in coming up with a ‘ranking’ of states / categorize them according to both total cases and crime rate. If I used only crime rate, states with small populations would disproportionately seem worse off while giving a free pass to large states that contribute to the majority of cases but have a more moderate crime rate.
- Instead of manually / randomly giving a weight to each factor and coming up with an index of some sort, I decided to use k-means clustering to make the process more natural.
- I used the in-built KMeans functionality of scikit-learn.

8. Crimes by Gender

- Since this was my first time using k-means, I started with a task whose results could be verified easily.
- I divided types of crimes into three categories depending on the gender of offenders:
    - Primarily boys
    - Both boys and girls
    - Primarily girls





I used the insights from the result of k-means clustering to select some common crimes that varied in the gender of the offenders.

- The cognizable and violent crimes such as burglary, rape, and theft are primarily carried out by boys.
- Social crimes like dowry, child marriage, and familial cruelty are universal to both genders.
- Unfortunately, girls constitute 81.5% of those trafficked.

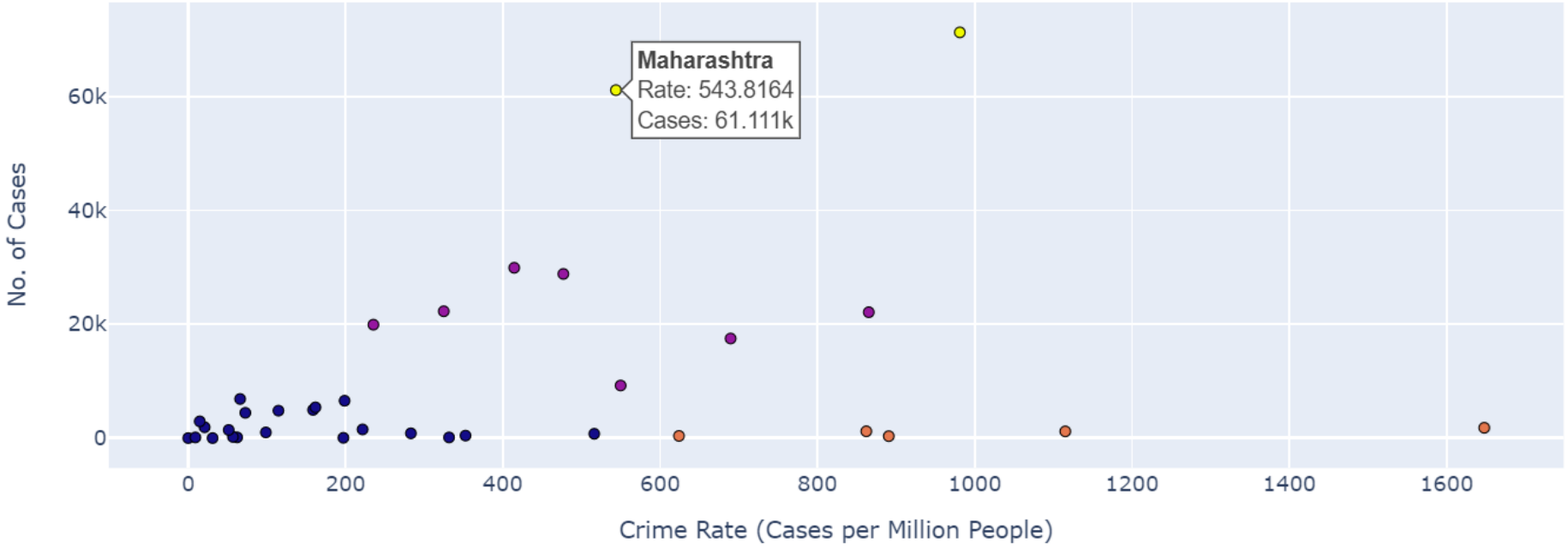
9. Clustering of States

- First, I used StandardScaler to standardize both features by calculating their z-scores.
- Next, I created a k-means model with n = 4 clusters (ideal n found after hit and trial) and fit the model to the standardized features to get the results.
- Lastly, I updated the dataframe to include each state's cluster.

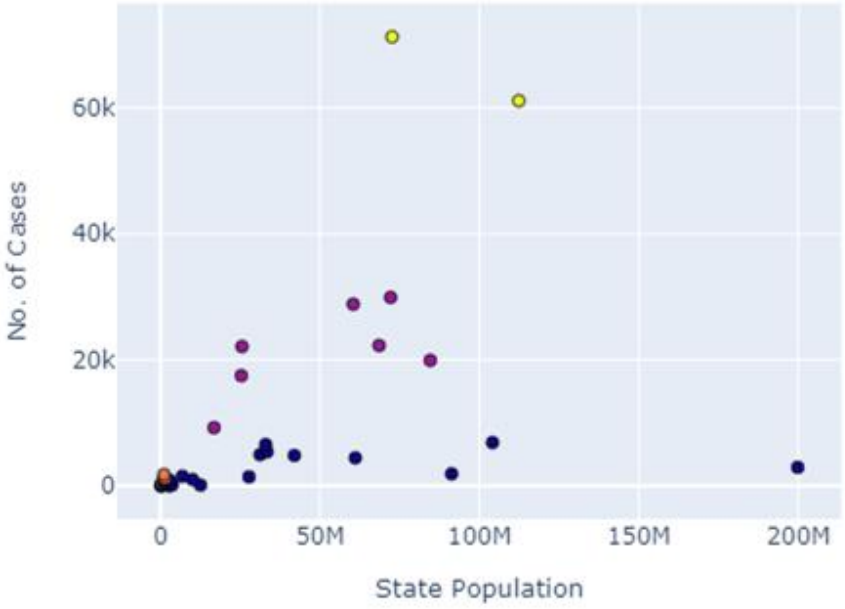
```
# Importing sklearn
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Using z-score fitting on Total and RatePerMillion to get scaled features
scaler = StandardScaler()
scaled_features = scaler.fit_transform(state_rate[["Total", "RatePerMillion"]])

# Instantiating k-means model
model = KMeans(n_clusters=4)

# Fitting model to scaled features
model.fit(scaled_features)

# Adding Cluster column to dataframe
state_rate["Cluster"] = model.labels_
```

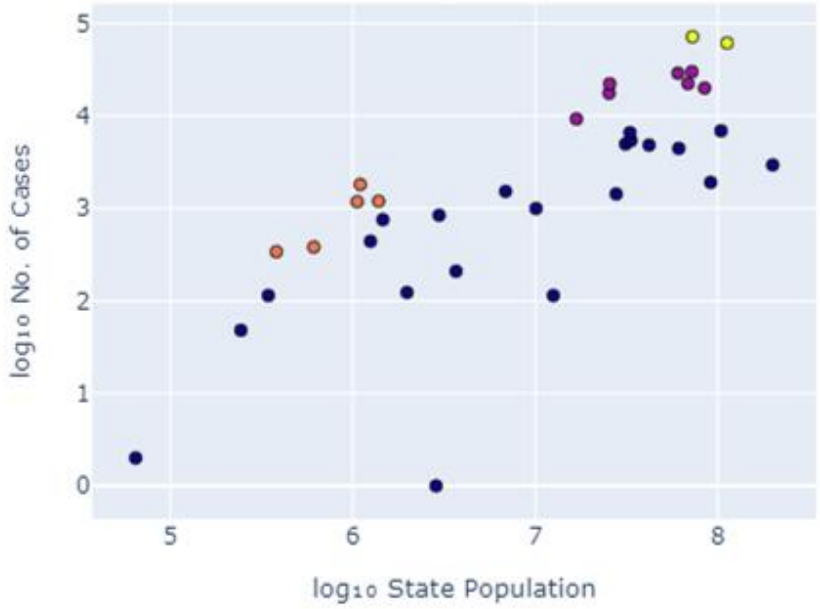
	State	Total	Population	RatePerMillion	Cluster	LogTotal	LogPopulation
0	Manipur	0	2855794	0.000000	0	0.000000	6.455727
1	Lakshadweep	2	64473	31.020737	0	0.301030	4.809378
2	Daman & Diu	48	243247	197.330286	0	1.681241	5.386047
3	Jammu & Kashmir	114	12541302	9.089965	0	2.056905	7.098343
4	Dadra & Nagar Haveli	114	343709	331.675924	0	2.056905	5.536191
5	Nagaland	123	1978502	62.168246	0	2.089905	6.296336
6	Tripura	209	3673917	56.887513	0	2.320146	6.565129
7	Andaman & Nicobar Islands	339	380581	890.743363	2	2.530200	5.580447
8	Sikkim	381	610577	623.999921	2	2.580925	5.785740

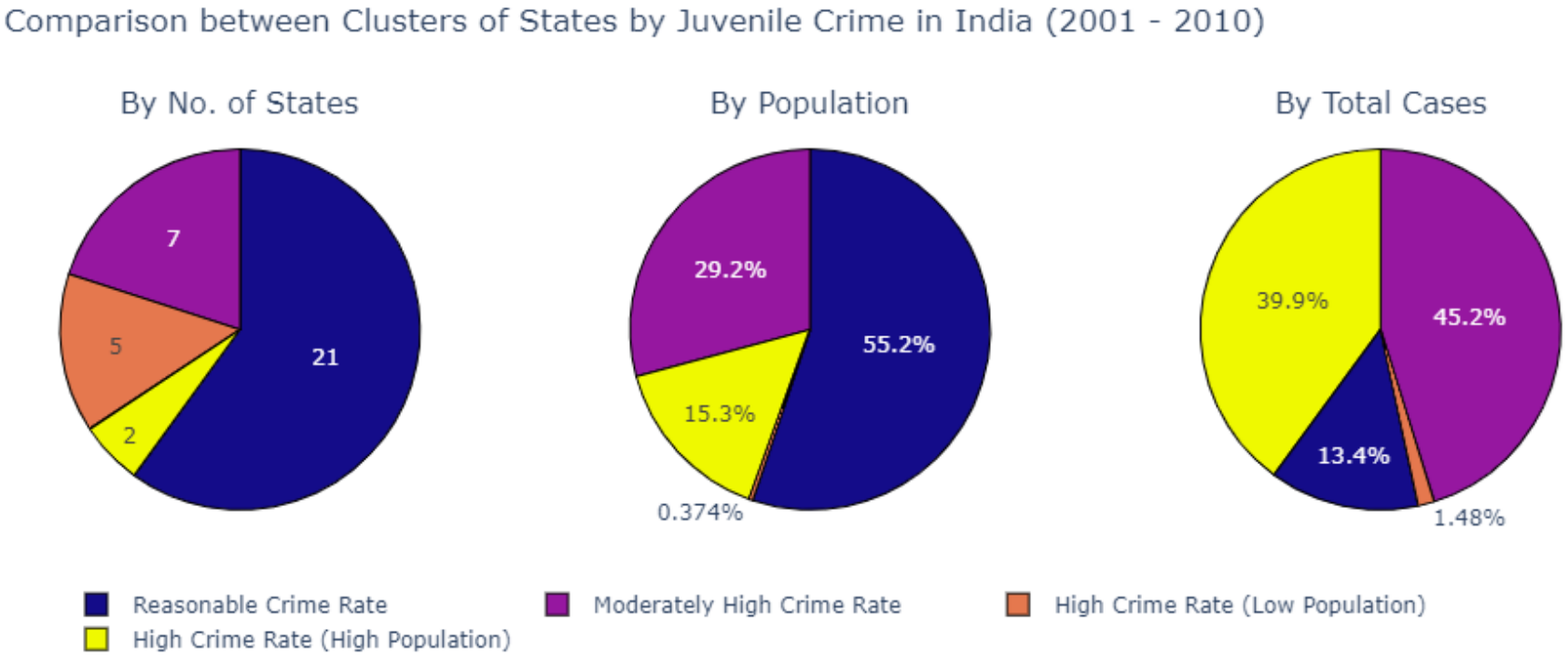
- I made a hoverable scatter plot depicting each state's number of cases against crime rate with the color of the marker indicating the cluster.
- According to my interpretation, blue states have low crime rates, purple ones have moderate crime rates, orange have high crime rates but low cases, and yellow states (Maharashtra and Madhya Pradesh - the worst 2 overall) have both high cases and crime rates.



- The interpretation holds up well when we consider state population as well. The blue states such as Uttar Pradesh (the right-most marker), lie across a line with a small slope (crime rate) even when x (population) increases. Moreover, the difference between orange and yellow states is evident from the magnitude even though they have a similar slope.
- The logarithmic graph amplifies these differences for easier viewing.







The nine purple and yellow states make up around 85% of all cases and should be prioritized by the government. Improving access to education and employment should work in the long-term as it will reduce the need for the youth to take up crime.

Limitations

- While 10 years is a reasonable time period for such a study, I found out that data from several decades was available on the National Crime Records Bureau website after I had finished the majority of the project work.
- The dataset has separate files for background information of offenders, but there is no table that correlates more than one factor at a time. For example, It would be a lot more useful to analyze the relative impact of both education and income on the crime rate.
- Although completely out of my control, there is certainly an aspect of differences in the strictness of the police across states, and therefore, the authenticity of the reported numbers. For instance, it is difficult to believe that Manipur had 0 juvenile crimes in an entire decade.

Further Work

- Incorporate a larger time period to improve the accuracy of the charts. This will allow actually monitoring trends (10 years is simply too short to predict whether crime has gone up or down).
- If more data is available, I could also use regression to predict future numbers.
- Since my dataset is built around comparing states, a map would be much more intuitive in displaying the information that I used bar charts for.
- Mapping would require troubleshooting Geopandas, finding a map of India from the 2000s (since multiple new states have been carved since), and manually matching the state names - strings - to geographical IDs.

Bibliography

1. Pandas Documentation: https://pandas.pydata.org/docs/user_guide/index.html#user-guide
2. Plotly Documentation: https://plotly.com/python/getting-started/
3. Seaborn Palette Documentation: https://seaborn.pydata.org/tutorial/color_palettes.html
4. Scikit-learn KMeans Documentation: https://scikit-learn.org/stable/modules/clustering.html#k-means
5. Complete NCRB Dataset: https://ncrb.gov.in/en/crime-in-india