Indian Institute of Information Technology, Nagpur

UNMASKING CRIME PATTERNS

A DATA-DRIVEN EXPLORATION OF CRIME AND SOCIETY IN INDIA

SAMBODHI BHOWAL - BT24CSD065
RAKSHIT LADDA - BT24CSD066

Guided By Dr. Santosh Sahu

Datasets

Crime In India Dataset from kaggle

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

- *02_01_District_wise_crimes_committed_against_SC_2001_2012.csv*
- *02_District_wise_crimes_committed_against_ST_2001_2012.csv*
- *03_District_wise_crimes_committed_against_children_2001_2012.csv*
- *12_Police_strength_actual_and_sanctioned.csv*
- *42_District_wise_crimes_committed_against_women_2001_2012.csv*

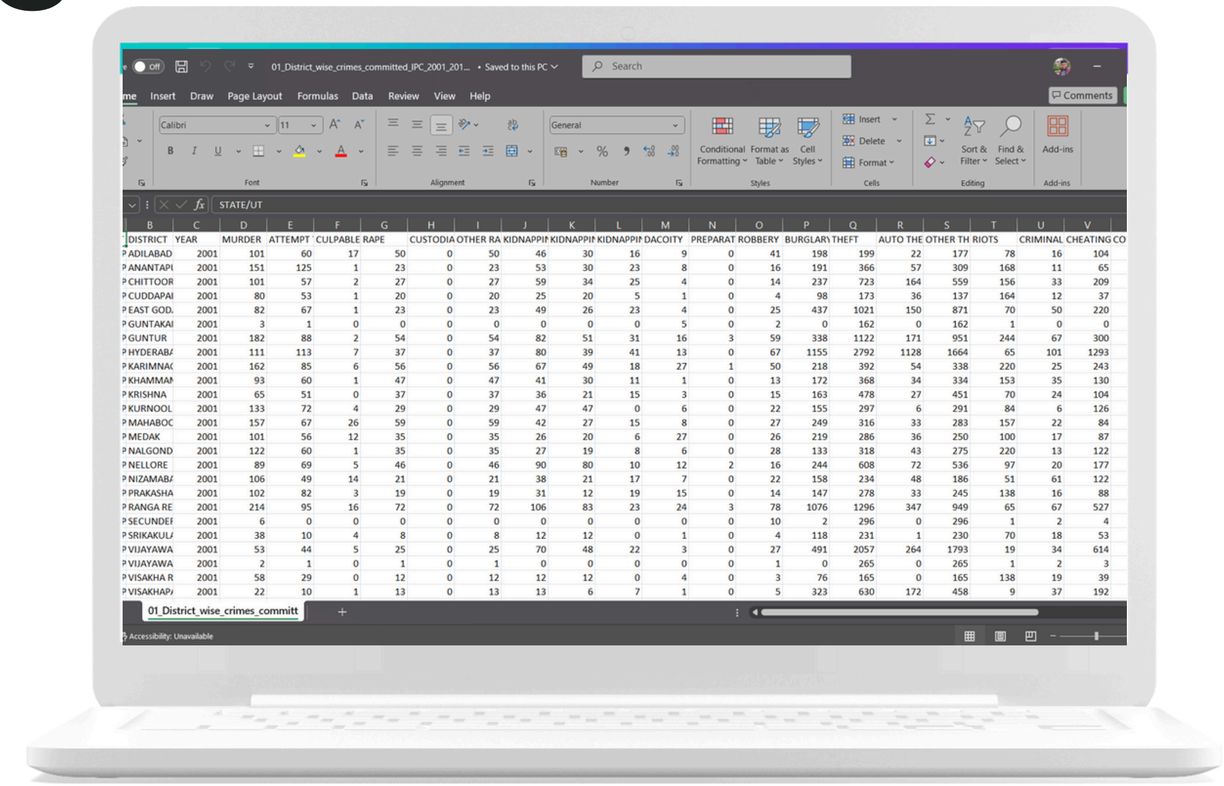
Indian Prison Statistics Dataset from kaggle

Link : https://www.kaggle.com/datasets/rajanand/prison-in-india/data

- *Education_facilities.csv*
- *Recidivism.csv*
- *Vocational_training.csv*

Excel Analysis

Our Process of data cleaning, data preprocessing, detecting outliets, a bit of visualization and merging data from various sources into a single coheisve file that can be used in the pythong notbook for better comprehension and easier handling



Data Preparation & Cleaning

Converting to XLSX

Each file is saved in xlsx format with the following filenames:

1. **crimes.xlsx**
2. **Crimes_SC.xlsx**
3. **Crimes_ST.xlsx**
4. **Crimes_Women.xlsx**
5. **Crimes_Children.xlsx**
6. **Census.xlsx**
7. **Police_strength.xlsx**
8. **Priosner_Education.xlsx**
9. **Educational_Facilities.xlsx**
10. **Recidivism.xlsx**
11. **Rehabilitation.xlsx**
12. **Vocational_training.xlsx**

Crimes.xlsx

columns

STATE/UT
DISTRICT
YEAR
MURDER
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

- **First we converted the data into an excel TABE for better analysis**
- **REMOVING TOAL ROWS**
  - So in the data set there are rows which show the total crimes for a particular state, in these states the value for the district column is “TOTAL**”**
  - **we used Data->Filter to filter the rows for which the column vaalue was “TOTAL” and deleted them**
- **We converted the year column and the all the different c crime columns to numeric type for better analysis**
- **Cleaning State and District Column**
  - we made helper columns State__clean and District__clean and clealy fromated the raw state and district data using formulas :
  - ***=UPPER(TRIM(CLEAN(A2)))***
  - ***=UPPER(TRIM(CLEAN(C2)))***
  - Then we copied the clean data into the origiunal columns and delted the helper columns.
- **Checking for Missing Values**
  - **We checked for missing values for each crime columns using formula**
  - ***=COUNTBLANK(D1:D8610)***
  - We found that there is no missing values so there is no need to fill in missing values

Crimes.xlsx

columns

STATE/UT
DISTRICT
YEAR
MURDER
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

- **Removing Duplicates**
  - we removed duplicated rows in the file by Data->Remove Duplicate using the combination (state,district,year) as they key , thiss combination would be later used to merge the different data sources into one
- **Conditional Formating**
  - we used conditional formating on the TOTAL IPC CRIMES columns to show a heatmap and bar chart

CE	OTHER IPC CRIMES	TOTAL IPC CRIMES
81.00	1518.00	4154
70.00	754.00	4125
04.00	1262.00	5818
33.00	1181.00	3140
31.00	2313.00	6507

- **Delhi Ut Total Deletion**
  - through IQR outlier detection we found that for each year there is a row with district as “delhi ut total” this seems like an total column which wasnt formatted correctly like the previous total rows so we deleted those rows as well

Crime_Women.xlsx

columns

STATE/UT
DISTRICT
Year
Rape
Kidnapping and Abduction
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

- Converted the data into a table
- We renamed the columns to match our other data sets
- We deleted all the total rows
- Similar to the previous data set using IQR we deleted the delhi total columns
- Similar to the previous dataset we cleaned up the state and district columns using helper columns
- **Removing Duplicates**
  - we removed duplicated rows in the file by Data->Remove Duplicate using the combination (state,district,year) as they key , thiss combination would be later used to merge the different data sources into one
- added a calculaterd column **Total_Crime_Women**
  - **=SUM(Table1[@[Rape]:[Importation of Girls]])**
- We added a z_score column that will help us detect outlier for our **Total_Crime_Women**
  - **=([@[Total_Crime_Women]] - AVERAGE([Total_Crime_Women])) / STDEV([Total_Crime_Women])**
  - We marked the row that are outliers using formula :
    - **=([@[Total_Crime_Women]] - AVERAGE([Total_Crime_Women])) / STDEV([Total_Crime_Women])**
  - **We then used conditional formating to high light the outliers**

Husband or his Relatives	Importation of Girls	Total_C	Z_Score	IS Outlier	O
251	0	513	0.942832	NOT OUTLIER	
210	0	624	1.323616	NOT OUTLIER	
289	0	831	2.033727	NOT OUTLIER	
278	0	446	0.71299	NOT OUTLIER	
142	0	367	0.441981	NOT OUTLIER	
508	0	914	2.318458	NOT OUTLIER	
0	0	0	-0.81701	NOT OUTLIER	
114	0	217	-0.07259	NOT OUTLIER	
869	0	1242	3.443657	OUTLIER	
2	0	5	-0.79985	NOT OUTLIER	
85	0	207	-0.1069	NOT OUTLIER	

Crime_children.xlsx

columns

STATE/UT
DISTRICT
Year
Murder
Rape
Kidnapping and Abduction
Foeticide
Abetment of suicide
Exposure and abandonment
Procuration of minor girls
Buying of girls for prostitution
Selling of girls for prostitution
Prohibition of child marriage act
Other Crimes
Total

- Converted the data into a table
- We renamed the columns to match our other data sets
- We deleted all the total rows
- Similar to the previous data set using IQR we deleted the delhi total columns
- Similar to the previous dataset we cleaned up the state and district columns using helper columns
- Removed Duplicates
- Detected all the rows with NA values with 0 as the total was zero so we deduced that the no of individual types of crimes should also be zero

Crime_SC.xlsx

columns

STATE/UT
DISTRICT
Year
Murder
Rape
Kidnapping and Abduction
Dacoity
Robbery
Arson
Hurt
Prevention of atrocities (POA) Act
Protection of Civil Rights (PCR) Act
Other Crimes Against SCs

- Converted the data into a table
- We renamed the columns to match our other data sets
- We deleted all the total rows
- Similar to the previous data set using IQR we deleted the delhi total columns
- Similar to the previous dataset we cleaned up the state and district columns using helper columns
- Removed Duplicate
- added a calculaterd column **total_crimes_sc**
  - =**SUM(Table2[@[Murder]:[Other Crimes Against SCs]])**

Crime_ST.xlsx

columns

STATE
DISTRICT
Year
Murder
Rape
Kidnapping Abduction
Dacoity
Robbery
Arson
Hurt
Protection of Civil Rights (PCR) Act
Prevention of atrocities (POA) Act
Other Crimes Against STs
total_crimes_st

- Converted the data into a table
- We renamed the columns to match our other data sets
- We deleted all the total rows
- Similar to the previous data set using IQR we deleted the delhi total columns
- Similar to the previous dataset we cleaned up the state and district columns using helper columns
- Removed Duplicate
- added a calculaterd column **total_crimes_st**
  - **=SUM(Table2[@[Murder]:[Other Crimes Against STs]])**

Police_Strength.xlsx

columns

Area_Name
Year
Group_Name
Sub_Group_Name
Rank_All_Ranks_Total
Rank_ASI_Equivalent
Rank_ASPDySPAssttCommandant
Rank_Below_HC_and_Above_Constables
Rank_Constables
Rank_DGAddl_DG
Rank_DIG
Rank_Head_Constables
Rank_IGSplIG
Rank_Inspectors_Equivalent
Rank_SI_Equivalent
Rank_SSPSPAddlSPCommandant

- changed column names to match the existing datasets, such as changing Area_name to STATE
- Added a calculated column, **total_police,** using the formula
  - **=SUM(Table1[@[Rank_All_Ranks_Total]:[Rank_SSPSPAddlSPCommandant]])**
- Removed all the total rows
- Added conditional formatting to the total_police column to show the distribution of police using a bar graph
  -

nk_Inspectors_Equivalent	Rank_SI_Equivalent	Rank_SSPSPAddlSPCommandant	total_p
487	2610	134	56682
561	4953	62	91340
48	233	4	7288
356	513	67	37420
1	12	2	436
S	9	1	432
1066	3461	50	99998
58	171	14	8040
575	1715	49	126138
491	1191	42	73118
193	415	22	19338

- Made a pivot table of districts, grouped by State and then filtered by year

Year	2004
Row Labels	**Sum of total_police**
Andaman & Nicobar Islands	**11618**
Actual Police Strength - Armed Police	1026
Actual Police Strength - Civil Police	4454
Actual Women Police Strength - Armed Police	46
Actual Women Police Strength - Civil Police	260
Sanctioned Police Strength - Armed Police	1230
Sanctioned Police Strength - Civil Police	4570
Sanctioned Women Police Strength - Armed Police	0
Sanctioned Women Police Strength - Civil Police	32
Andhra Pradesh	**329836**
Actual Police Strength - Armed Police	24886
Actual Police Strength - Civil Police	130814
Actual Women Police Strength - Armed Police	0
Actual Women Police Strength - Civil Police	2702
Sanctioned Police Strength - Armed Police	26944
Sanctioned Police Strength - Civil Police	141662
Sanctioned Women Police Strength - Armed Police	0
Sanctioned Women Police Strength - Civil Police	2828
Arunachal Pradesh	**24396**
Actual Police Strength - Armed Police	5058

census.xlsx

columns

District code
State name
District name
Population
Male
Female
Literate
Male_Literate
Female_Literate
SC
Male_SC
Female_SC

90+ columns

- **Making the data a table for better handling**
- **Renaming Column Names**
  - We converted the column names in this Excel file to match our other datasets, like State Name to STATE
- **Cleaning the district and state columns like we did with the crimes data set, using helper columns**
- Checked for missing values; there are None.

Vocation_training.csx

columns

state_name
year
vocational_trainings_
program
inmates_trained

- *Only few modifications were required*
- *Converted it into excel table*
- *We changed the column names to align with existing data*
- *Removed the total columns*

Educational_facilities.xlsx

columns

state_name
year
elementary_education
adult_education
higher_education
computer_course

- *Only few modifications were required*
- *Converted it into excel table*
- *We changed the column names to align with existing data*
- *Removed the total columns*

Recidivism.xlsx

columns

STATE
YEAR
convicts_admitted
habitual_offenders

- *Only few modification were required here as wewll*
- *converted into excel table*
- *changed column names to align with existing data*
- *Removed duplicate entries*

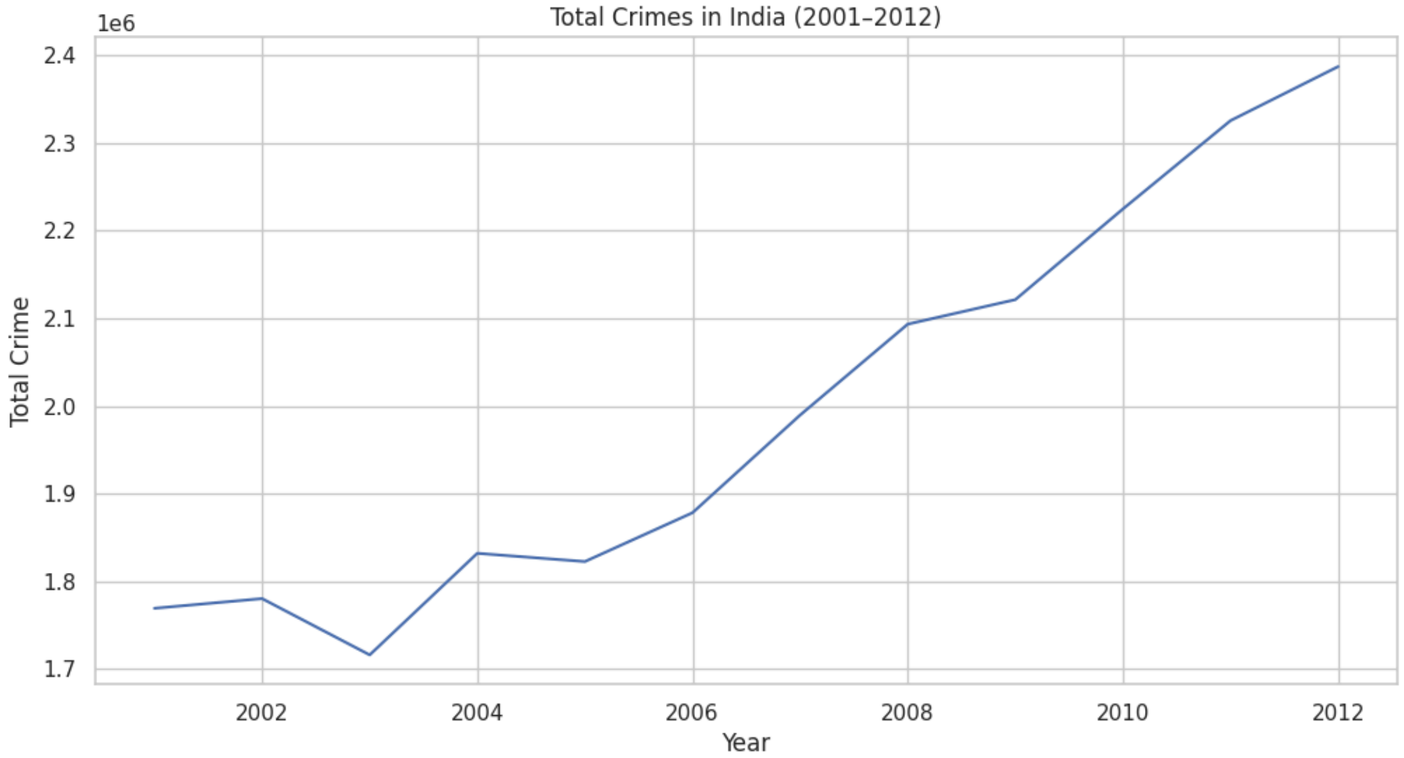
Merging 5 of the Crime DataSets

- we made a new file Master_Crime.xlsx
- we copied all the Crime_women, Crime_Children, Crime_SC and Crime_ST data into different sheets in the Master_crime.xlsx
- Next we added a key column to all the sheet in the format of
  - State-District#Year
- This key will act as a primary key, helping in merging all the data sources
- Then we used X Lookup to merge the different sheets into the main sheet using formulas like
  - **=XLOOKUP([@KEY],Crime_SC!D:D,Crime_SC!O:O)**
- After merging, there were a lot of N/A values, we replaced them with the average value of the column
  - =**IF(ISNA([@[total_crime_against_sc]]), AGGREGATE(1, 6, [total_crime_against_sc]), [@[total_crime_against_sc]])**

Merging state level datasets

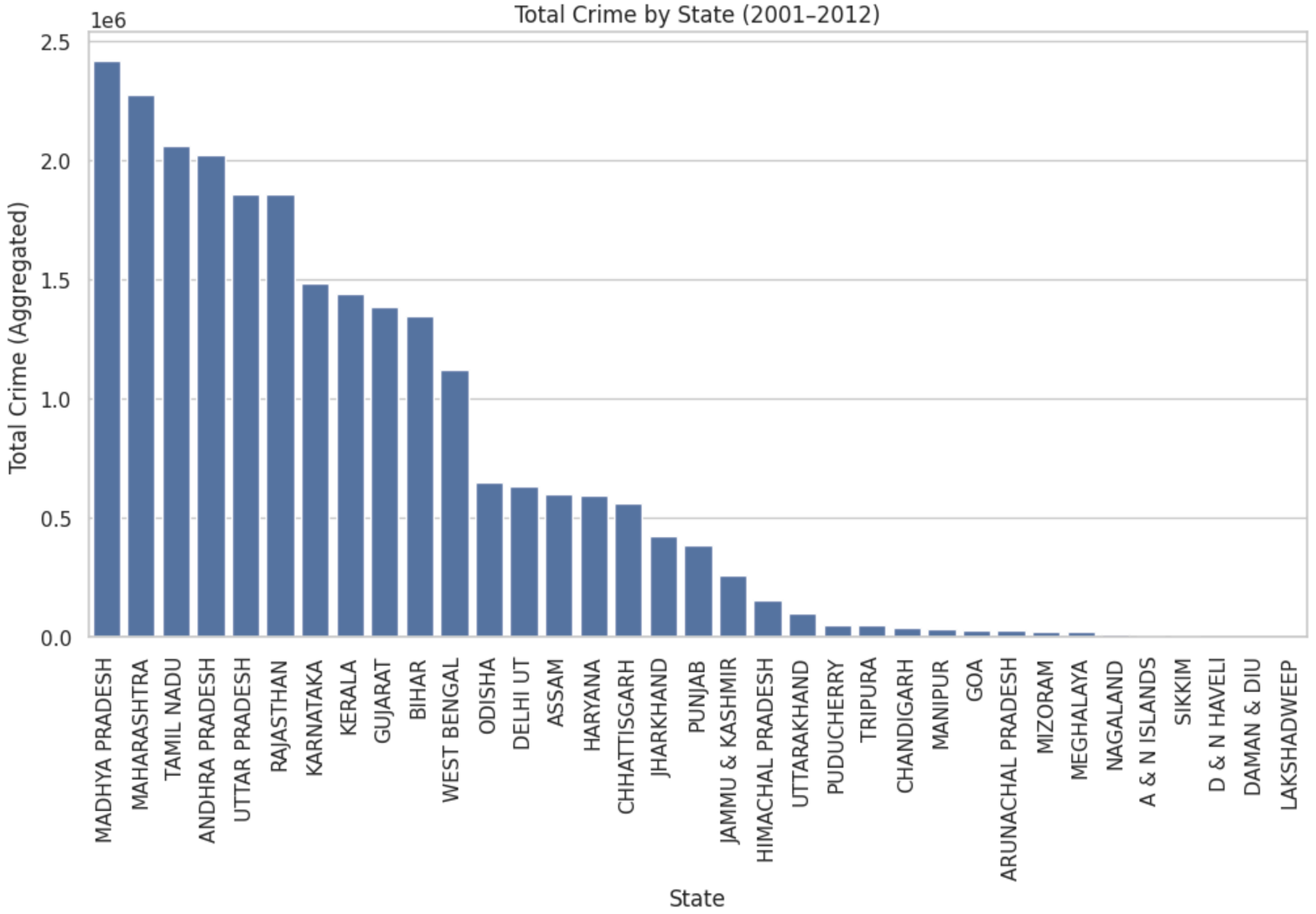
- we made new excel file called State_master.xlsx
- we will merge all the state level datasets in this using the state-year as a key
- we made a column name key with state-year and then merged the other files using XLoopup
  - **=XLOOKUP([@Key],policseStrength!C:C,policseStrength!R:R)**
- There were a lot of N/A Values in the data set once again so we replaced them with the average
  - **=IF(ISNA([@[total_ps]]), AGGREGATE(1, 6, [total_ps]), [@[total_ps]])**

How has crime changed with time in india from 2001-2012?



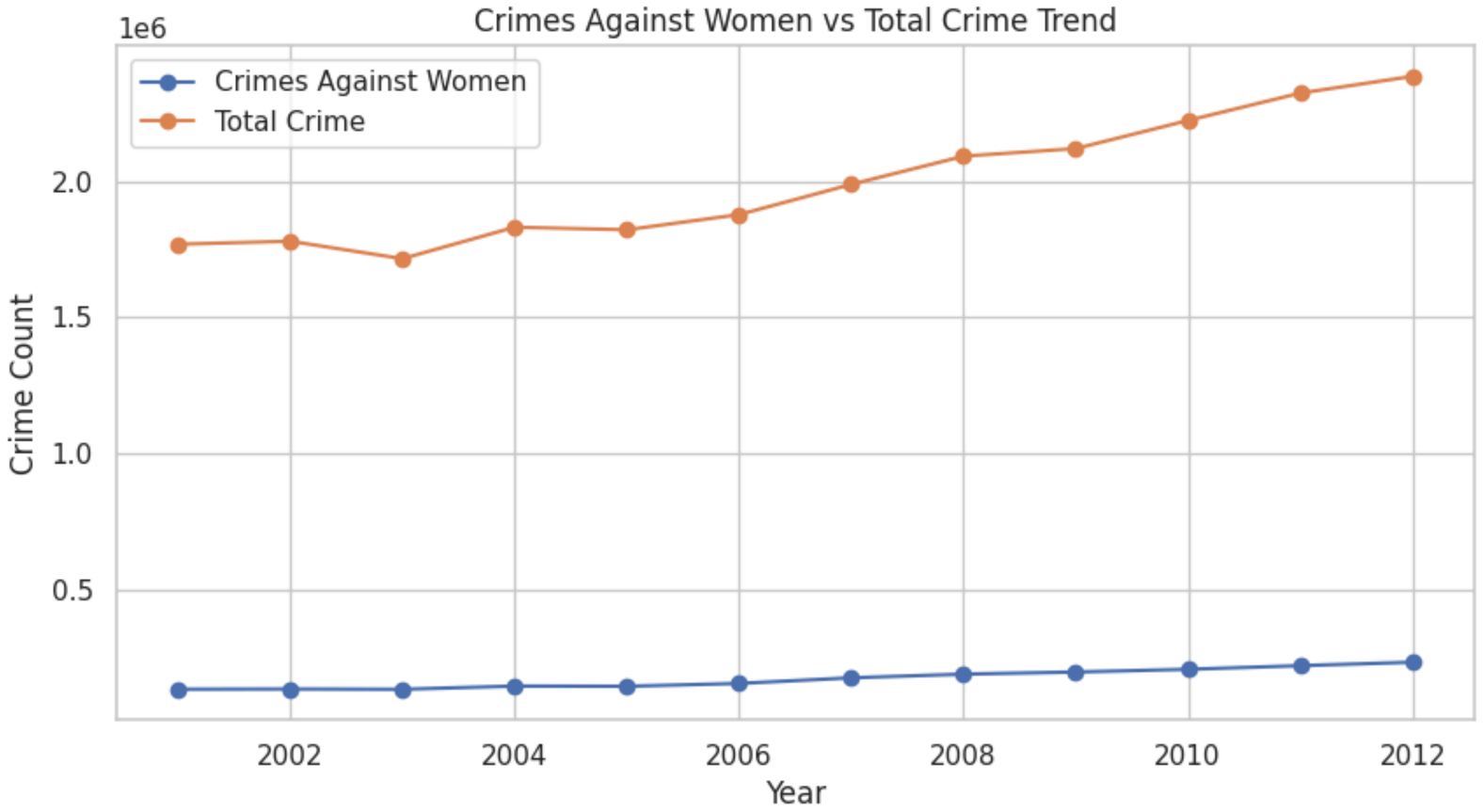
As we can see here the no of crimes in india has grown tremendously since 2007-08 this coincides with the economic crashes of this era.

Which states show the highest overall crime levels?



- ***States such as Madhya Pradesh, Maharashtra, and Tamil Nadu show the highest total crime counts over the 12-year period. These states represent major population centers and/or regions with intensive reporting, policing challenges, and socio-economic stress.***
- ***This highlights where long-term investments in policing and justice delivery may be most needed.***
- ***Also in states like Jammu and Kashmir and Manipur there may have been a under reporting of crime, so we must also consider those states***

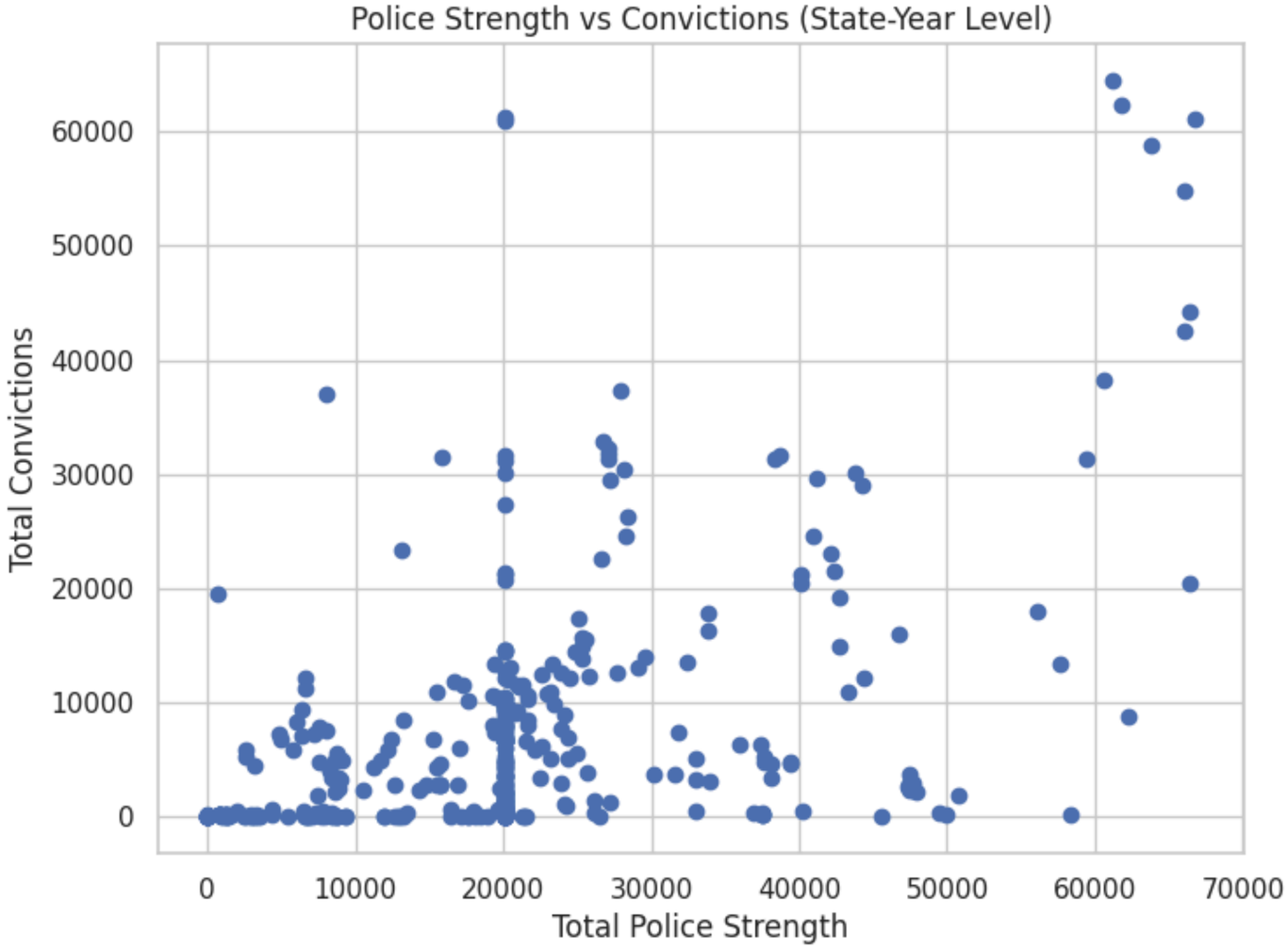
Are crimes against women rising faster than overall crime?



- ***As we can see that the rate of crime against women have been steady while the rate of growth of total crimes has been on an increase this may suggest stabalization of women rights and safety***

- ***However we must also consider the fact that crimes against women are the most underreported crimes and no significant increase in the no of cases reported may indicate the persistent under reoporting of crimes against woment and the societal stigma on the victims***

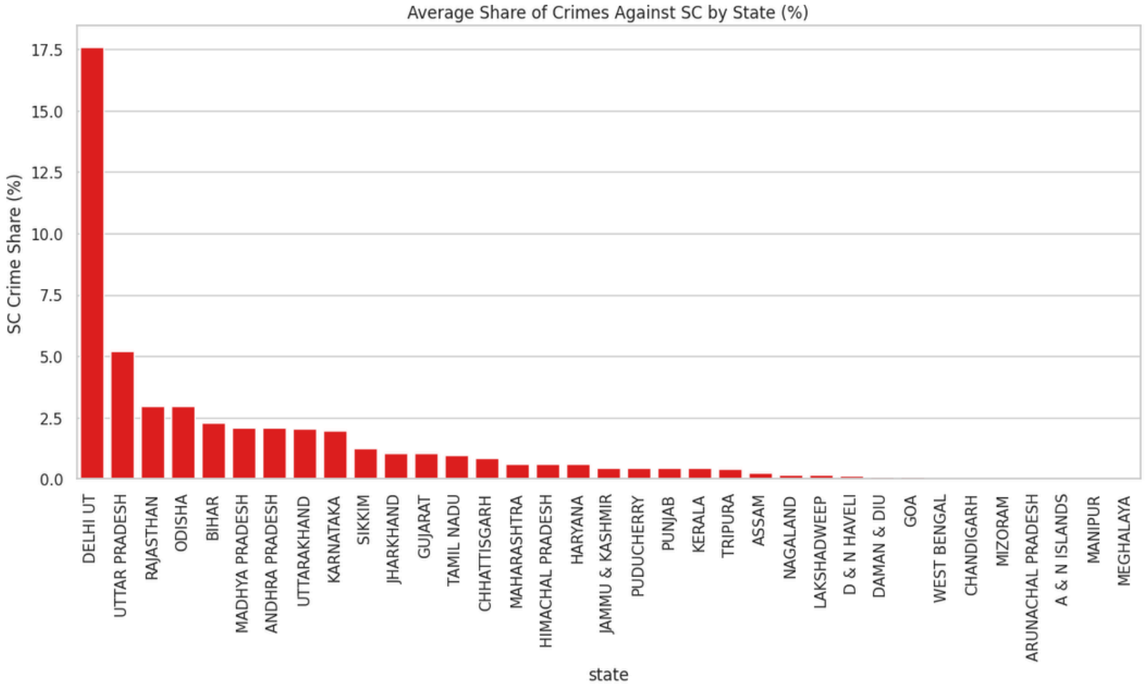
Does a larger police force reduce the number of convictions?

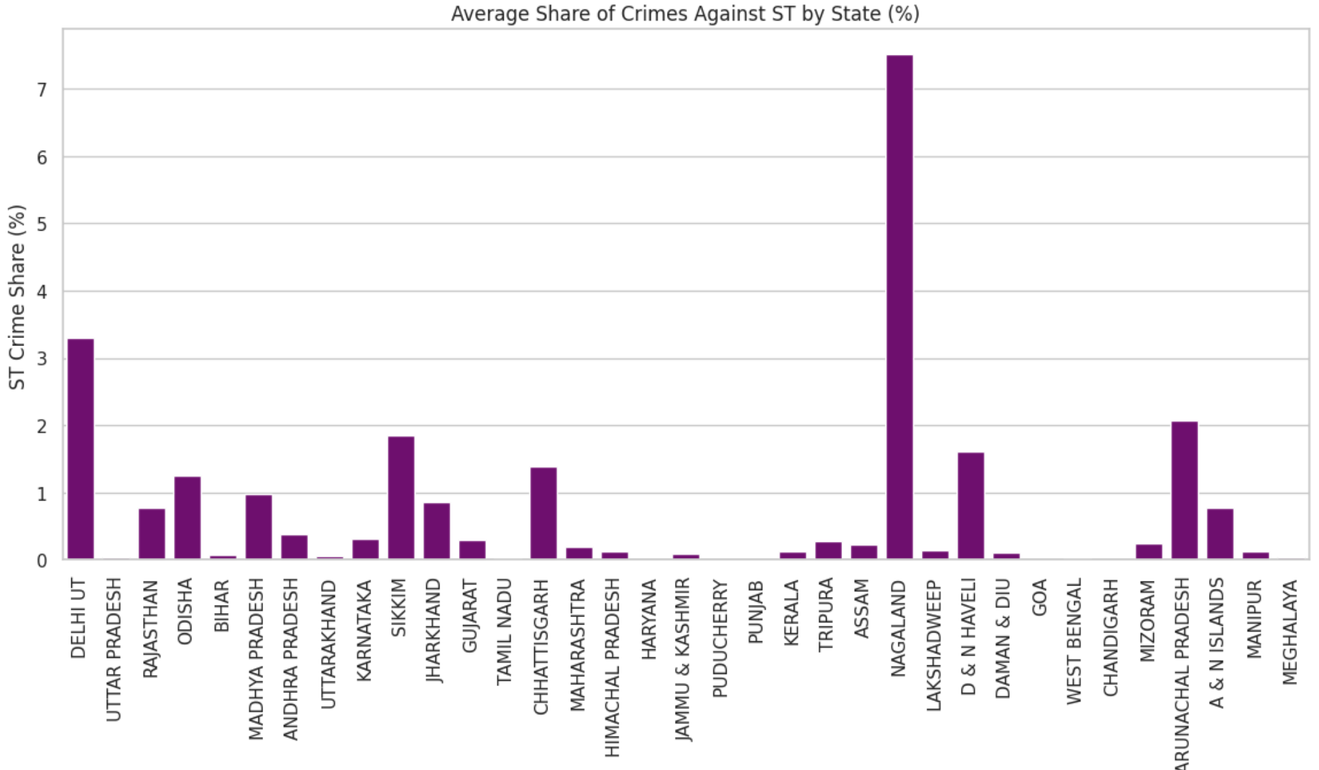


This analysis checks whether states with higher police strength actually experience lower crime. If police strength and crime have negative correlation, policing is effective.

So we can see that although there is a correlation between police strength and convictions its not as strong as we might expect it to be most of the convitions lie in the range of 10000-30000 police men

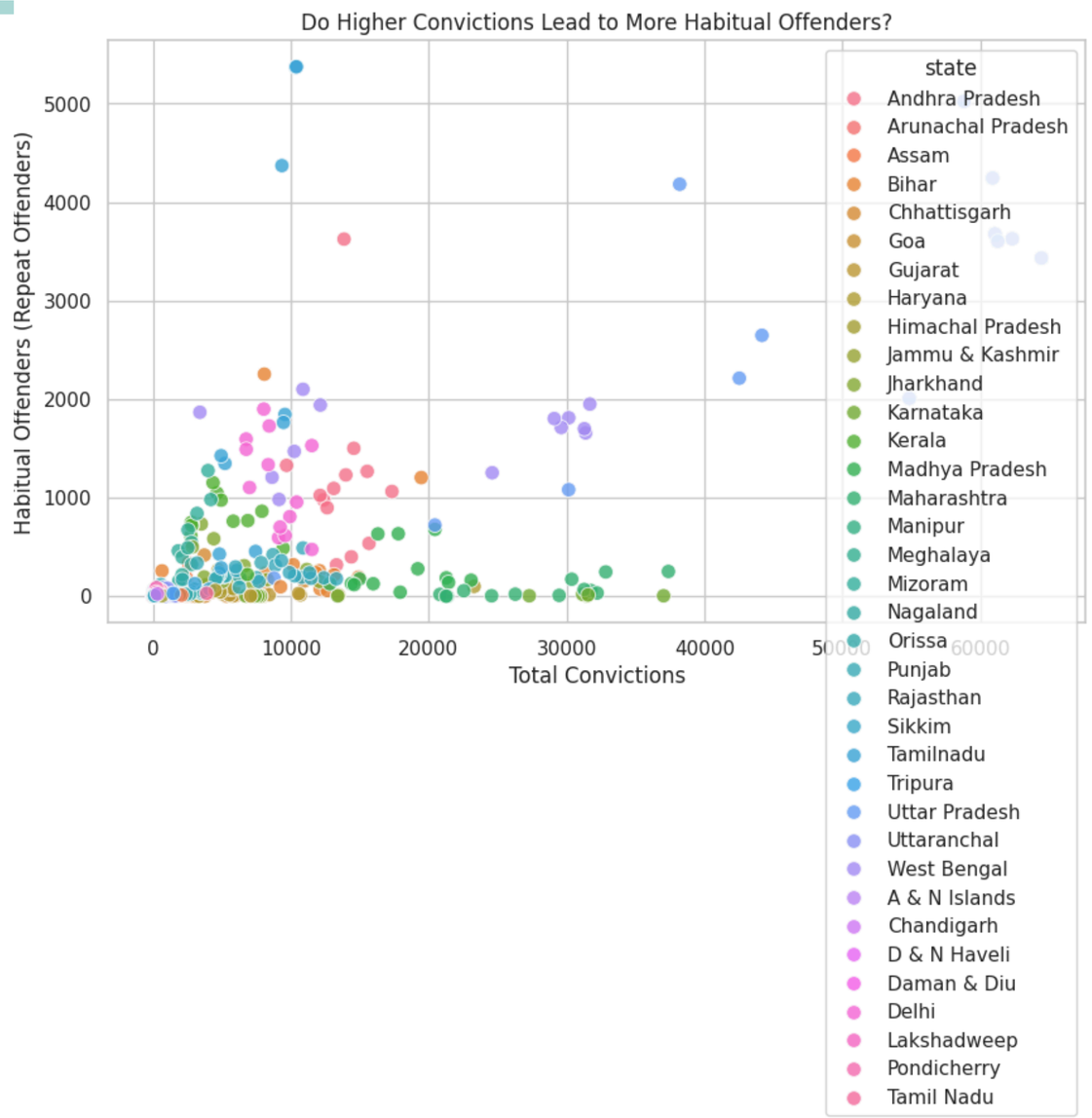
How do crimes against SC/ST distribute across states?





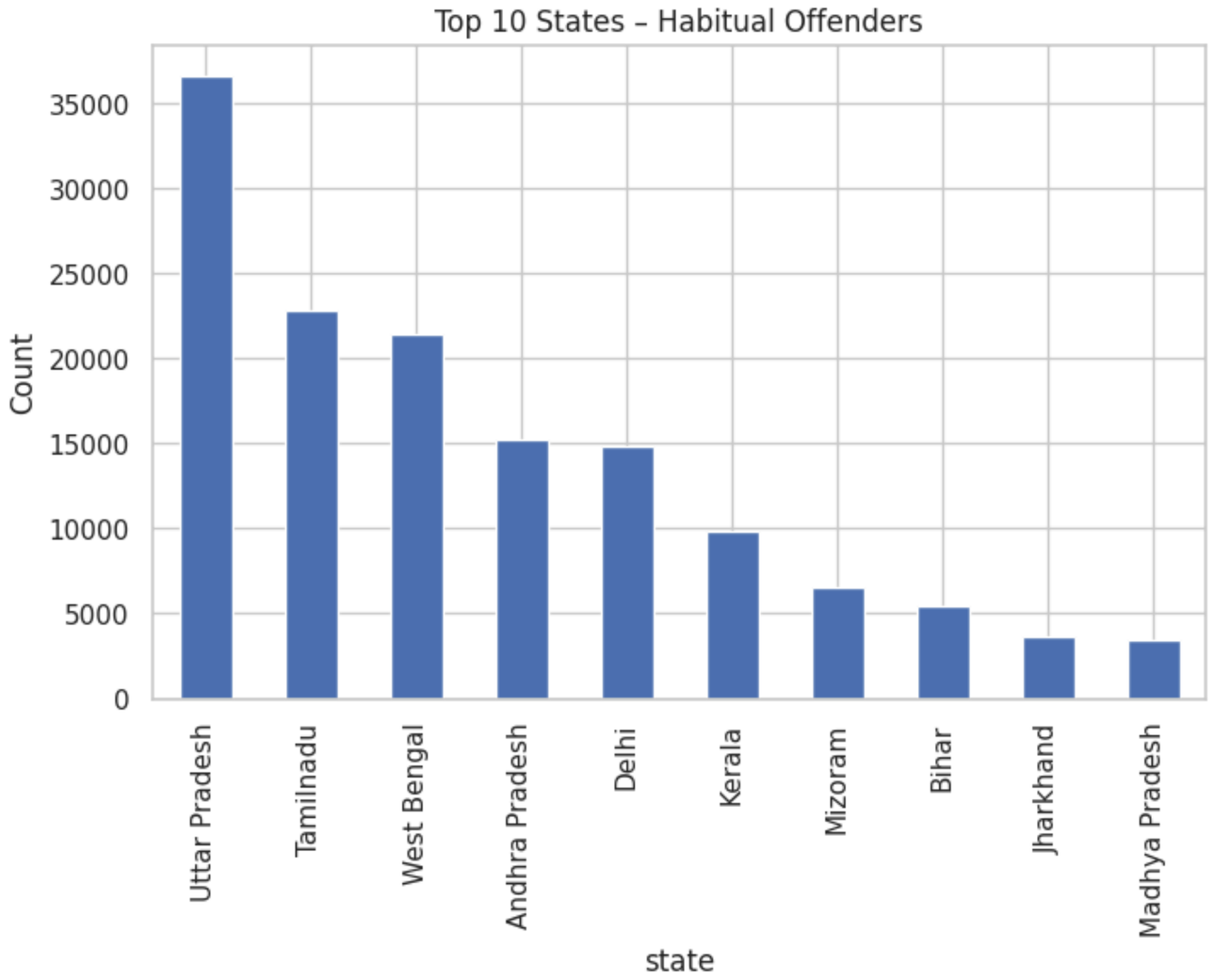
Comparing habitual offender rates of different states and the correlationt to no of convictions

if there are more convictions is it also necessairly true that the number of repeat offenders will be more too?



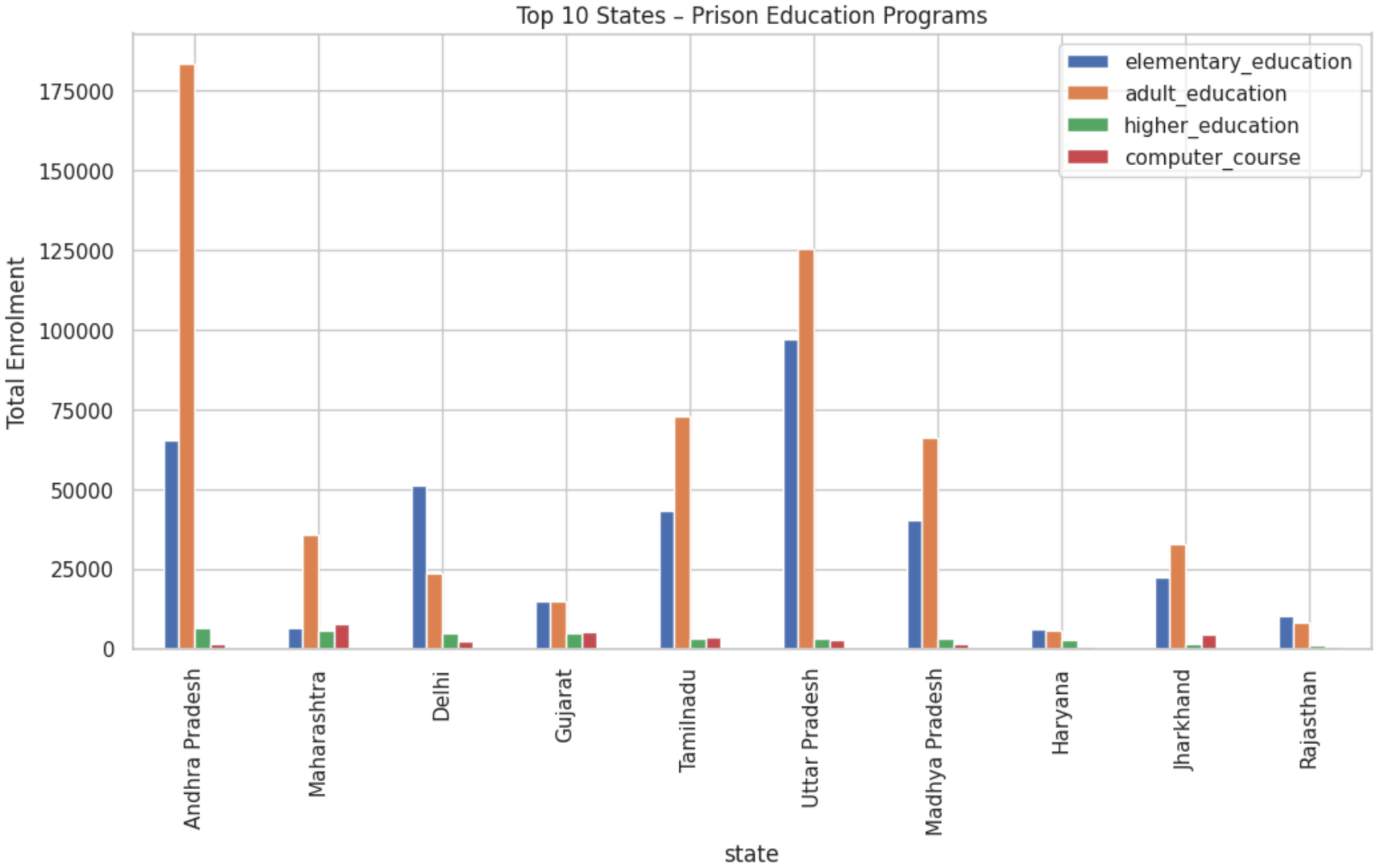
We can see that upto a certain point the number of repeat offences is steady no matter the number of incarcinations this may indicate that most ex criminals dont return to a life of crime

What are the top 10 States with Highest Habitual Offenders



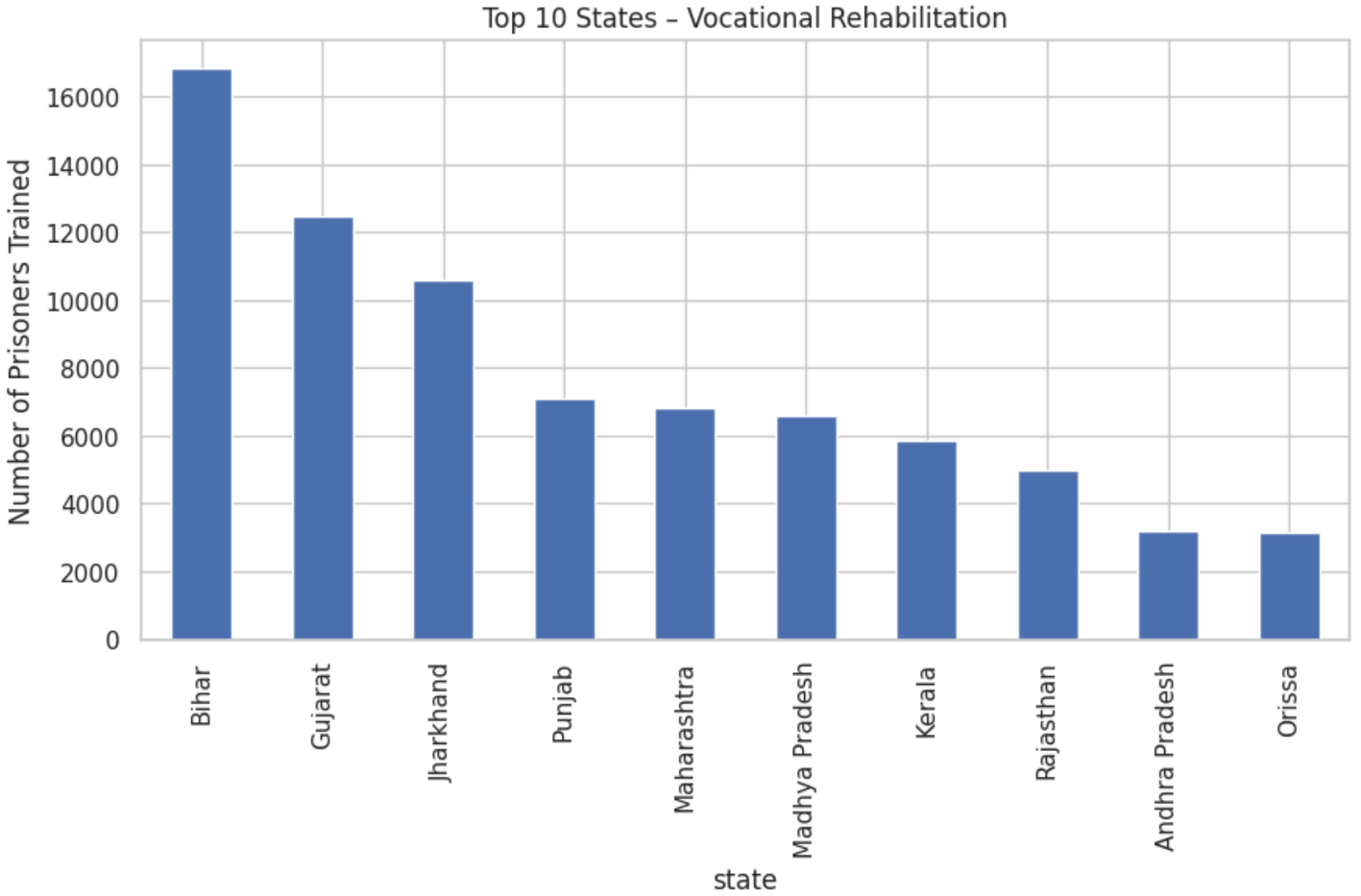
This may indicate that these states dont have proper instituations in place to help ex cerminals get back to a lawful life, thus these states should invest in welfare systems

What are the Top States in Prison Education ?



This shows that states like andhra Pradesh and Mahrashtra are actively putting effor into educating the prison population in hopes that they can get decent job after they get out of prison

What states are investion in vocational training of their prison population? and what might it indicate

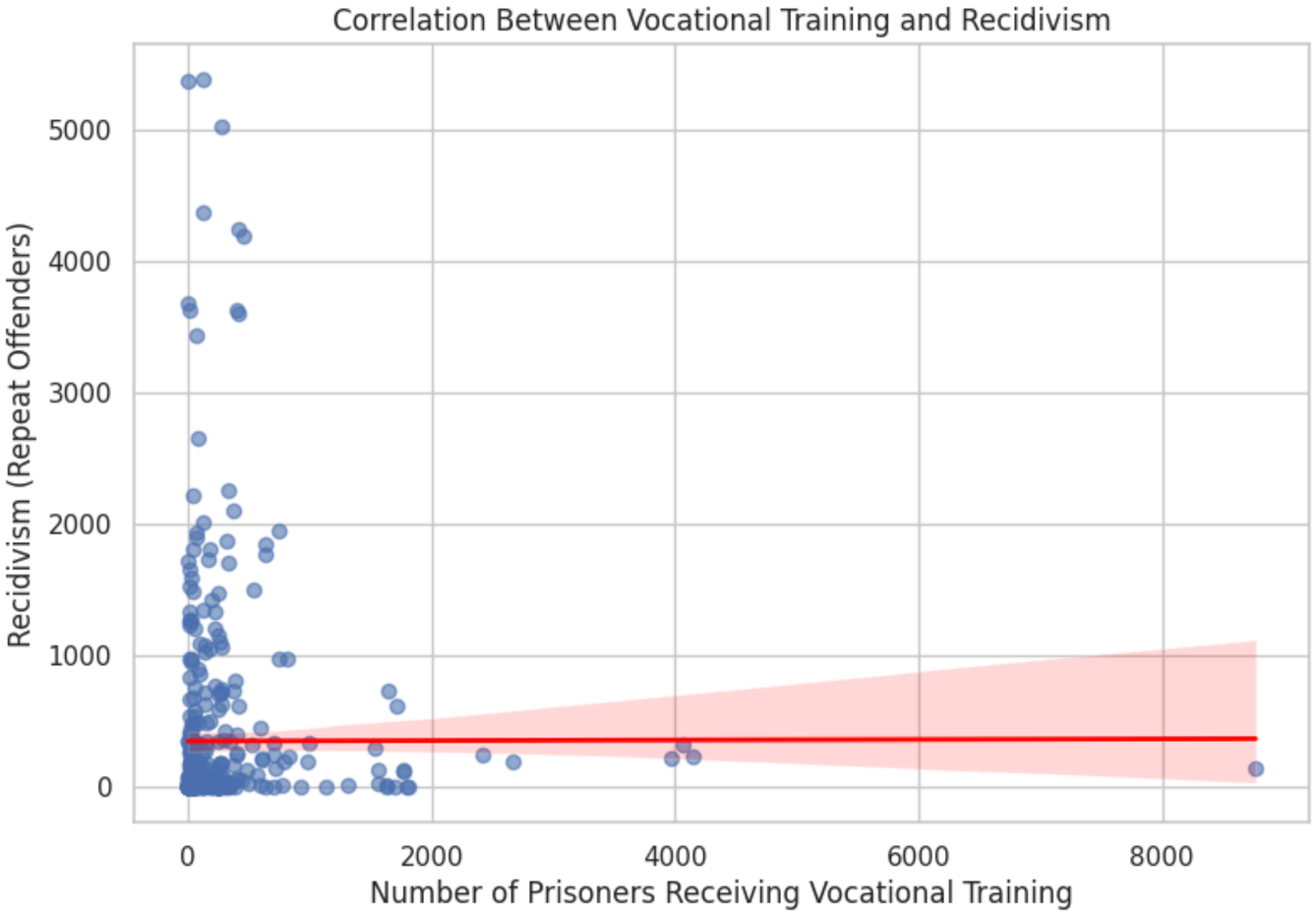


We can see that states like bihar and jharkhand are investing more in vocational education of their prison poulation than traditional education, this may indicate the economy of these states are more suitable for vocational work and thus the workers can easily get jobs. Thus its good that the states are playing to their strengths in tyring to give ex criminals a new chance at life

Does better prison education correlate with lower recidivism?

Though we can see that there is a gneral trend of less recidivism with increasing prison education score, the correlation is not as strong as one might expect, this maybe becuase the educaion is not effective or simply its irrelevant to the prision propulations skill set

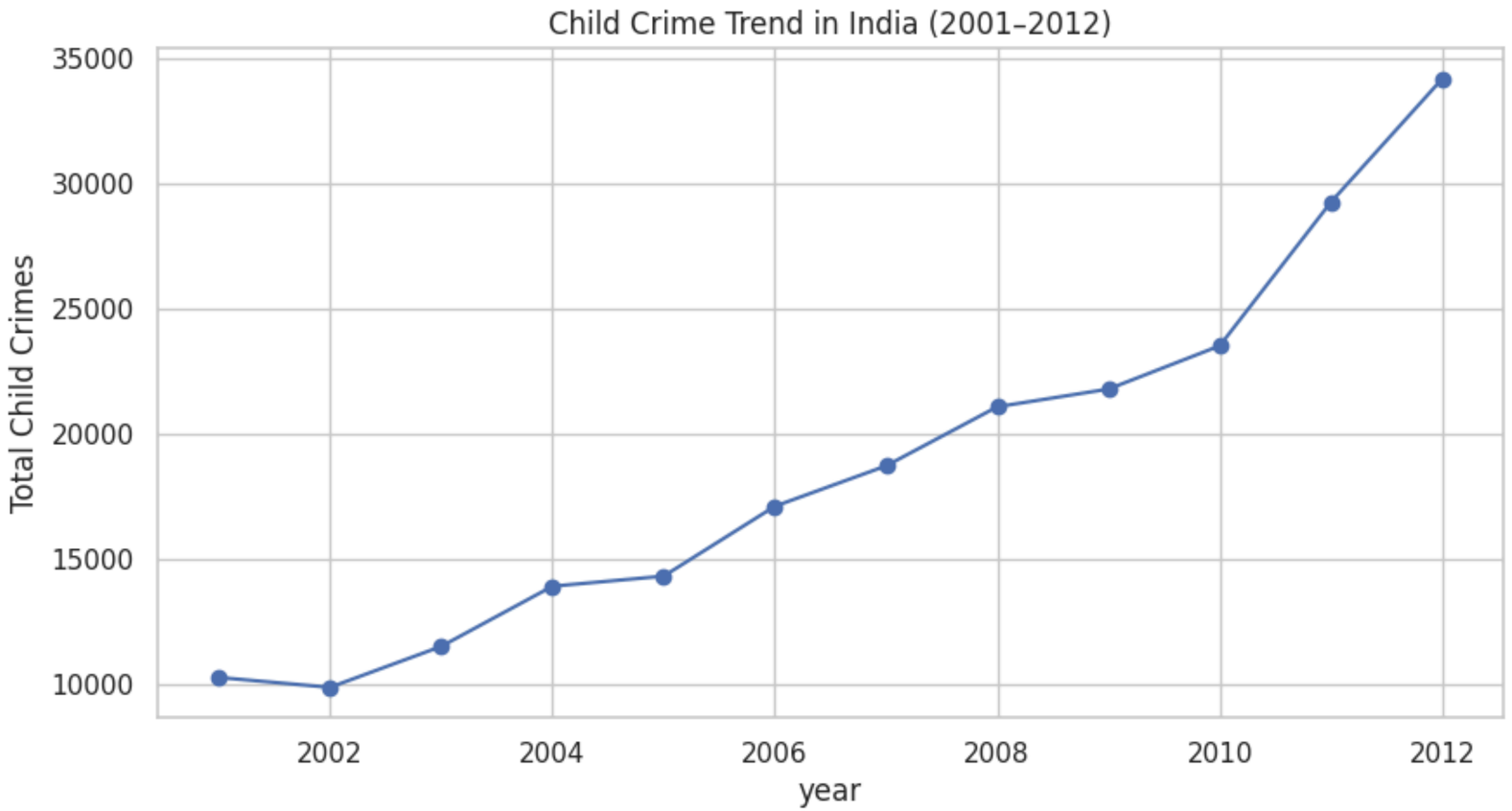
Does better prison vocational education correlate with lower recidivism?



This shows that there is better correlation with vocational learning to less recedivism than with traditional education, this may indicate a need to change the education mindset we think of when considering the rehabilitation of prison population.

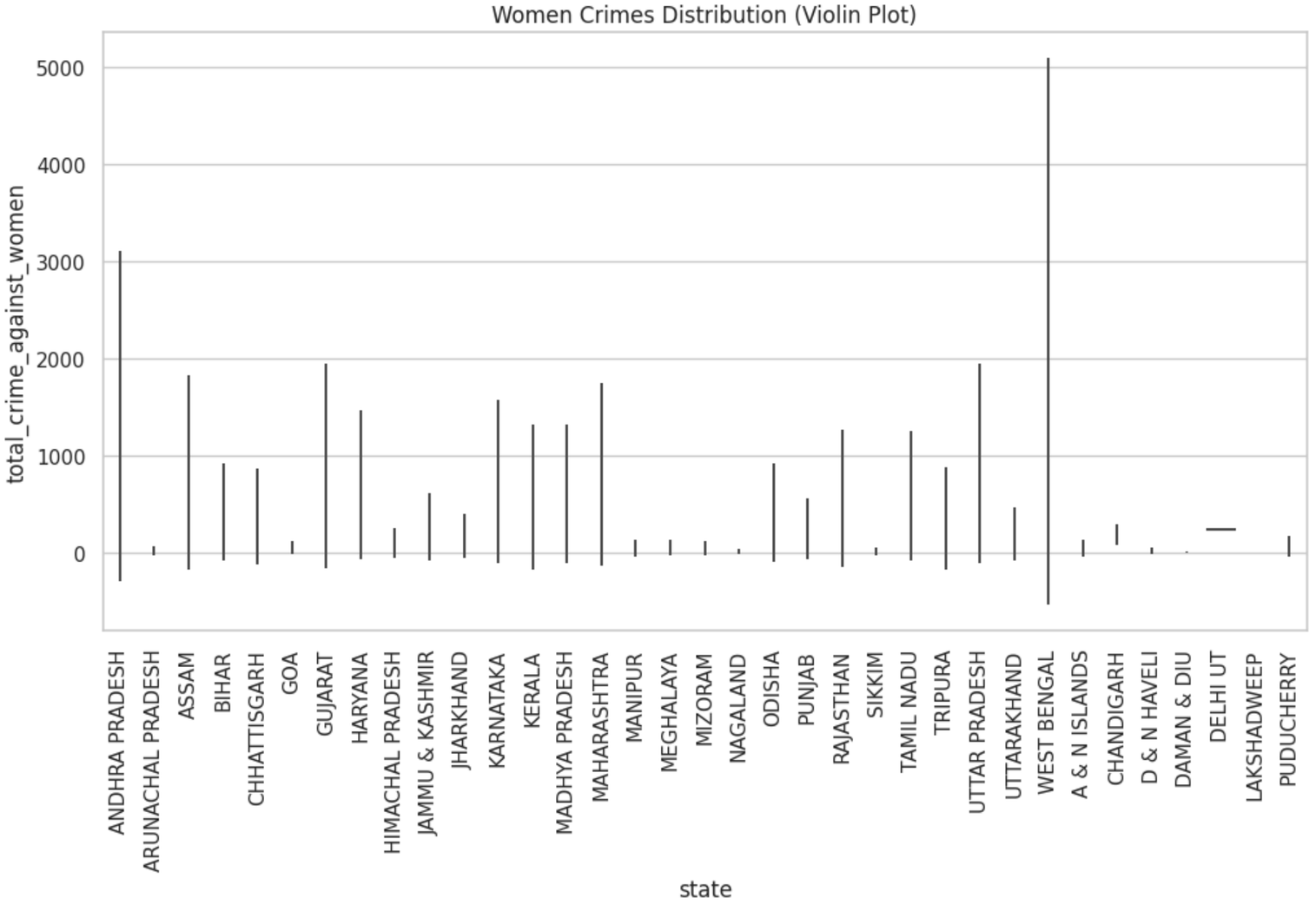
We can clearly see this model of re-education works better than high-skill education basied policys

Growth of crimes against children in india



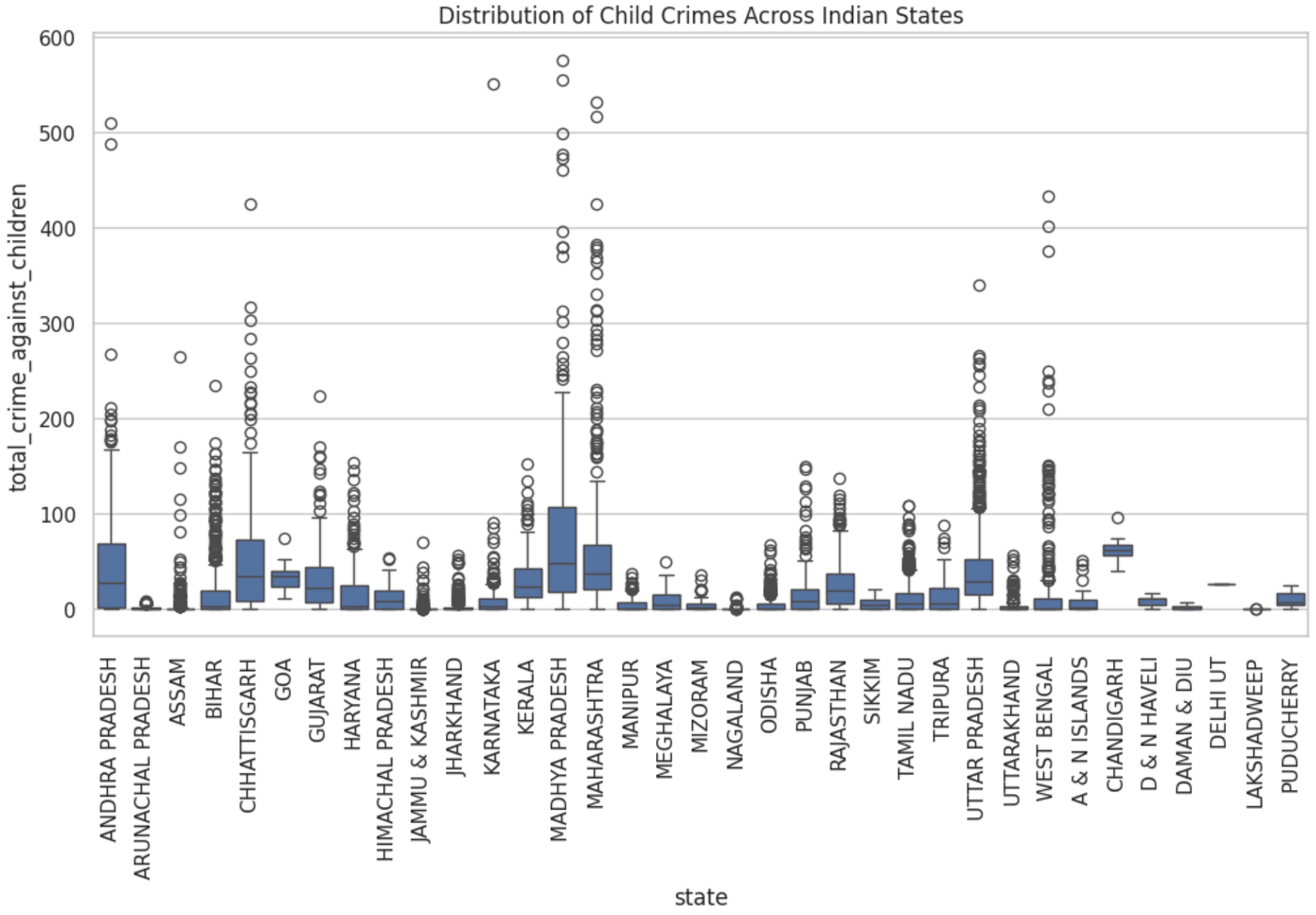
The chart shows a steady and sharp rise in child crimes from 2001 to 2012, with a major surge after 2008. This increase can reflect both actual growth in incidents—due to factors like urbanization, poverty, and greater exposure of children—and improved reporting, as awareness, media coverage, and child-protection laws strengthened during this period. Early years likely had underreporting, so part of the rise may be the system capturing cases more accurately. Overall, the trend suggests a growing recognition and visibility of child crimes, alongside a real increase in vulnerability.

What is the distribution of crime against women over state



The violin plot shows that crimes against women are highest and most variable in major states like Uttar Pradesh, Rajasthan, Maharashtra, and Madhya Pradesh, while smaller and Northeastern states show consistently low levels. Some states have extreme spikes, visible as long vertical stretches. However, since crimes against women are often underreported due to stigma and fear, the real numbers may be much higher than what is shown

Distribution of crimes against children across the states



The boxplot shows that child crime levels differ widely across Indian states. Large states like Madhya Pradesh, Maharashtra, Uttar Pradesh, Rajasthan, and West Bengal report high median cases along with a large number of outliers, indicating not just higher overall crime but also frequent spikes in specific districts or years. In contrast, smaller and Northeastern states as well as most Union Territories show low medians and very few outliers, reflecting consistently low levels or limited variation. Overall, the chart suggests that child crimes are heavily concentrated in a few major states, while the majority of states show low and stable patterns with minimal extreme values.

Thank you very much!