In [1]:

```python
import pandas as pd
```

In [2]:

```python
df_ncrb = pd.read_excel("ncrb.xlsx")
df_nfhs = pd.read_csv("nfhs.csv")

df_ncrb.head(), df_nfhs.head()
```

```
Out[2]:  (            State/UT  Trafficking  Murder with Rape/Gang Rape  Dowry Deaths  \
          0     Andhra Pradesh          107                           8           100   
          1  Arunachal pradesh            1                           0             0   
          2              Assam           78                          14           175   
          3              Bihar           87                           0          1057   
          4       Chhattisgarh           13                           7            57   

             Abetment to Suicide of Women   Miscarriage  Acid Attack  \
          0                            358            4            3   
          1                              0            0            0   
          2                             75            2            3   
          3                              2            0            3   
          4                            149            5            0   

             Attempt to Acid Attack  Cruelty by Husband/relatives  Kidnapping/Abduction
         \
          0                       3                         11964                   592   
          1                       0                            74                    48   
          2                       2                          4704                  3466   
          3                       0                          1850                 10190   
          4                       0                           942                  2121   

             ...  Rape  Attempt to Commit Rape  Assault to Outrage her Modesty  \
          0  ...   621                     180                            5884   
          1  ...    74                       3                              67   
          2  ...  1113                     253                            1984   
          3  ...   881                      17                             402   
          4  ...  1246                       8                            1322   

             Insult to the Modesty of Women  Assault due to Dowry  Domestic violence  \
          0                            3145                   298                  0   
          1                              20                     0                  1   
          2                             150                   272                  0   
          3                               0                  3580                  0   
          4                             255                     9                  0   

             Cyber Crimes committed against women  Sexual Violence towards girl child  \
          0                                   108                                2127   
          1                                     1                                  46   
          2                                   152                                1703   
          3                                    17                                2126   
          4                                   203                                2355   

             Indecent Representation of Women  Total Crime against Women (IPC &SLL)  
          0                                 1                                 25503  
          1                                 0                                   335  
          2                                 0                                 14148  
          3                                10                                 20222  
          4                                 0                                  8693  

          [5 rows x 22 columns],
                            States/UTs   Area  Number of Households surveyed  \
          0                      India  Urban                         160138   
          1                      India  Rural                         476561   
          2                      India  Total                         636699   
          3  Andaman & Nicobar Islands  Urban                            527   
          4  Andaman & Nicobar Islands  Rural                           2097   

             Number of Women age 15-49 years interviewed  \
          0                                       179535   
```

```
 1                                       544580
 2                                       724115
 3                                          557
 4                                         1840

    Number of Men age 15-54 years interviewed  \
 0                                      26420
 1                                      75419
 2                                     101839
 3                                         85
 4                                        282

   Female population age 6 years and above who ever attended school (%)  \
 0                                              82.5
 1                                              66.8
 2                                              71.8
 3                                              86.5
 4                                              81.8

    Population below age 15 years (%)  \
 0                               23.1
 1                               28.1
 2                               26.5
 3                               22.7
 4                               19.7

     Sex ratio of the total population (females per 1,000 males)  \
 0                                              985.0
 1                                             1037.0
 2                                             1020.0
 3                                             1023.0
 4                                              929.0

   Sex ratio at birth for children born in the last five years (females per 1,0
00 males)  \
 0                                               924
 1                                               931
 2                                               929
 3                                               941
 4                                               891

   Children under age 5 years whose birth was registered with the civil authori
ty (%)  \
 0                                              93.3
 1                                              87.5
 2                                              89.1
 3                                              96.9
 4                                              97.8

    ...  \
 0  ...
 1  ...
 2  ...
 3  ...
 4  ...

   Women (age 15-49 years) having a mobile phone that they themselves use (%)
\
 0                                              69.4
 1                                              46.6
 2                                              54.0
 3                                              80.8
 4                                              80.9

    Women age 15-24 years who use hygienic methods of protection during their m
enstrual period26 (%)  \
 0                                              89.4
 1                                              72.3
 2                                              77.3
 3                                              98.5
 4                                              99.1

    Ever-married women age 18-49 years who have ever experienced spousal violen
ce27 (%)  \
 0                                              24.2
 1                                              31.6
 2                                              29.3
 3                                              23.2
 4                                              13.2

    Ever-married women age 18-49 years who have experienced physical violence d
uring any pregnancy (%)  \
 0                                               2.5
 1                                               3.4
 2                                               3.1
 3                                              (0.0)
 4                                               0.5

   Young women age 18-29 years who experienced sexual violence by age 18 (%)  \
 0                                               1.1
 1                                               1.6
 2                                               1.5
 3                                               1.4
 4                                               2.2

   Women age 15 years and above who use any kind of tobacco (%)  \
 0                                               5.4
 1                                              10.5
 2                                               8.9
 3                                              15.0
 4                                              41.1

   Men age 15 years and above who use any kind of tobacco (%)  \
 0                                              28.8
 1                                              42.7
 2                                              38.0
 3                                              44.7
 4                                              66.4

   Women age 15 years and above who consume alcohol (%)  \
 0                                               0.6
 1                                               1.6
 2                                               1.3
 3                                               0.7
 4                                               7.6

   Men age 15 years and above who consume alcohol (%) Unnamed: 136
 0                                              16.5           NaN
 1                                              19.9           NaN
 2                                              18.8           NaN
```

```
3                                                33.8           NaN
4                                                41.9           NaN

[5 rows x 137 columns])
```

In [3]:

```
print(df_ncrb.columns.tolist())
print(df_nfhs.columns.tolist())
```

```
['State/UT', 'Trafficking', 'Murder with Rape/Gang Rape', 'Dowry Deaths', 'Abetme
nt to Suicide of Women ', 'Miscarriage', 'Acid Attack', 'Attempt to Acid Attack',
'Cruelty by Husband/relatives', 'Kidnapping/Abduction', 'Selling of Minor Girls
 ', 'Buying of Minor Girls', 'Rape', 'Attempt to Commit Rape', 'Assault to Outrage
her Modesty', 'Insult to the Modesty of Women', 'Assault due to Dowry', 'Domestic
violence', 'Cyber Crimes committed against women', 'Sexual Violence towards girl
child', 'Indecent Representation of Women', 'Total Crime against Women (IPC &SL
L)']
['States/UTs', 'Area', 'Number of Households surveyed', 'Number of Women age 15-4
9 years interviewed', 'Number of Men age 15-54 years interviewed', 'Female popula
tion age 6 years and above who ever attended school (%)', 'Population below age 1
5 years (%)', ' Sex ratio of the total population (females per 1,000 males)', 'Se
x ratio at birth for children born in the last five years (females per 1,000 male
s)', 'Children under age 5 years whose birth was registered with the civil author
ity (%)', 'Deaths in the last 3 years registered with the civil authority (%)',
'Population living in households with electricity (%)', 'Population living in hou
seholds with an improved drinking-water source1 (%)', 'Population living in house
holds that use an improved sanitation facility2 (%)', 'Households using clean fue
l for cooking3 (%)', 'Households using iodized salt (%)', 'Households with any us
ual member covered under a health insurance/financing scheme (%)', 'Children age
5 years who attended pre-primary school during the school year 2019-20 (%)', 'Wom
en (age 15-49) who are literate4 (%)', 'Men (age 15-49) who are literate4 (%)',
'Women (age 15-49)  with 10 or more years of schooling (%)', 'Men (age 15-49)  wi
th 10 or more years of schooling (%)', 'Women (age 15-49)  who have ever used the
internet (%)', 'Men (age 15-49)  who have ever used the internet (%)', 'Women age
20-24 years married before age 18 years (%)', 'Men age 25-29 years married before
age 21 years (%)', 'Total Fertility Rate (number of children per woman)', 'Women
age 15-19 years who were already mothers or pregnant at the time of the survey
(%)', 'Adolescent fertility rate for women age 15-19 years5', 'Neonatal mortality
rate (per 1000 live births)', 'Infant mortality rate (per 1000 live births)', 'Un
der-five mortality rate (per 1000 live births)', 'Current Use of Family Planning
Methods (Currently Married Women Age 15-49  years) - Any method6 (%)', 'Current U
se of Family Planning Methods (Currently Married Women Age 15-49  years) - Any mo
dern method6 (%)', 'Current Use of Family Planning Methods (Currently Married Wom
en Age 15-49  years) - Female sterilization (%)', 'Current Use of Family Planning
Methods (Currently Married Women Age 15-49  years) - Male sterilization (%)', 'Cu
rrent Use of Family Planning Methods (Currently Married Women Age 15-49  years) -
IUD/PPIUD (%)', 'Current Use of Family Planning Methods (Currently Married Women
Age 15-49  years) - Pill (%)', 'Current Use of Family Planning Methods (Currently
Married Women Age 15-49  years) - Condom (%)', 'Current Use of Family Planning Me
thods (Currently Married Women Age 15-49  years) - Injectables (%)', 'Total Unmet
need for Family Planning (Currently Married Women Age 15-49  years)7 (%)', 'Unmet
need for spacing (Currently Married Women Age 15-49  years)7 (%)', 'Health worker
ever talked to female non-users about family planning (%)', 'Current users ever t
old about side effects of current method of family planning8 (%)', 'Mothers who h
ad an antenatal check-up in the first trimester  (for last birth in the 5 years b
efore the survey) (%)', 'Mothers who had at least 4 antenatal care visits  (for l
ast birth in the 5 years before the survey) (%)', 'Mothers whose last birth was p
rotected against neonatal tetanus (for last birth in the 5 years before the surve
y)9 (%)', 'Mothers who consumed iron folic acid for 100 days or more when they we
re pregnant (for last birth in the 5 years before the survey) (%)', 'Mothers who
consumed iron folic acid for 180 days or more when they were pregnant (for last b
irth in the 5 years before the survey} (%)', 'Registered pregnancies for which th
e mother received a Mother and Child Protection (MCP) card (for last birth in the
5 years before the survey) (%)', 'Mothers who received postnatal care from a doct
or/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (for la
st birth in the 5 years before the survey) (%)', 'Average out-of-pocket expenditu
re per delivery in a public health facility (for last birth in the 5 years before
the survey) (Rs.)', 'Children born at home who were taken to a health facility fo
r a check-up within 24 hours of birth (for last birth in the 5 years before the s
urvey} (%)', 'Children who received postnatal care from a doctor/nurse/LHV/ANM/mi
dwife/ other health personnel within 2 days of delivery (for last birth in the 5
years before the survey) (%)', 'Institutional births (in the 5 years before the s
urvey) (%)', 'Institutional births in public facility (in the 5 years before the
survey) (%)', 'Home births that were conducted by skilled health personnel  (in t
he 5 years before the survey)10 (%)', 'Births attended by skilled health personne
l (in the 5 years before the survey)10 (%)', 'Births delivered by caesarean secti
on (in the 5 years before the survey) (%)', 'Births in a private health facility
that were delivered by caesarean section (in the 5 years before the survey) (%)',
'Births in a public health facility that were delivered by caesarean section (in
the 5 years before the survey) (%)', "Children age 12-23 months fully vaccinated
based on information from either vaccination card or mother's recall11 (%)", 'Chi
ldren age 12-23 months fully vaccinated based on information from vaccination car
d only12 (%)', 'Children age 12-23 months who have received BCG (%)', 'Children a
ge 12-23 months who have received 3 doses of polio vaccine13 (%)', 'Children age
12-23 months who have received 3 doses of penta or DPT vaccine (%)', 'Children ag
e 12-23 months who have received the first dose of measles-containing vaccine (MC
V) (%)', 'Children age 24-35 months who have received a second dose of measles-co
ntaining vaccine (MCV) (%)', 'Children age 12-23 months who have received 3 doses
of rotavirus vaccine14 (%)', 'Children age 12-23 months who have received 3 doses
of penta or hepatitis B vaccine (%)', 'Children age 9-35 months who received a vi
tamin A dose in the last 6 months (%)', 'Children age 12-23 months who received m
ost of their vaccinations in a public health facility (%)', 'Children age 12-23 m
onths who received most of their vaccinations in a private health facility (%)',
'Prevalence of diarrhoea in the 2 weeks preceding the survey (Children under age
5 years) (%) ', 'Children with diarrhoea in the 2 weeks preceding the survey who
received oral rehydration salts (ORS) (Children under age 5 years) (%) ', 'Childr
en with diarrhoea in the 2 weeks preceding the survey who received zinc (Children
under age 5 years) (%) ', 'Children swith diarrhoea in the 2 weeks preceding the
survey taken to a health facility or health provider (Children under age 5 years)
(%) ', 'Children Prevalence of symptoms of acute respiratory infection (ARI) in t
he 2 weeks preceding the survey (Children under age 5 years) (%) ', 'Children wit
h fever or symptoms of ARI in the 2 weeks preceding the survey taken to a health
facility or health provider (Children under age 5 years) (%)  ', 'Children under
age 3 years breastfed within one hour of birth15 (%)', 'Children under age 6 mont
hs exclusively breastfed16 (%)', 'Children age 6-8 months receiving solid or semi
-solid food and breastmilk16 (%)', 'Breastfeeding children age 6-23 months receiv
ing an adequate diet16, 17  (%)', 'Non-breastfeeding children age 6-23 months rec
eiving an adequate diet16, 17 (%)', 'Total children age 6-23 months receiving an
adequate diet16, 17  (%)', 'Children under 5 years who are stunted (height-for-ag
e)18 (%)', 'Children under 5 years who are wasted (weight-for-height)18 (%)', 'Ch
ildren under 5 years who are severely wasted (weight-for-height)19 (%)', 'Childre
n under 5 years who are underweight (weight-for-age)18 (%)', 'Children under 5 ye
ars who are overweight (weight-for-height)20 (%)', 'Women (age 15-49 years) whose
Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2)21 (%)', 'Men (age 15-49 y
ears) whose Body Mass Index (BMI) is below normal (BMI <18.5 kg/m2) (%)', 'Women
(age 15-49 years) who are overweight or obese (BMI ≥25.0 kg/m2)21 (%)', 'Men (age
15-49 years) who are overweight or obese (BMI ≥25.0 kg/m2) (%)', 'Women (age 15-4
9 years) who have high risk waist-to-hip ratio (≥0.85) (%)', 'Men (age 15-49 year
s) who have high risk waist-to-hip ratio (≥0.90) (%)', 'Children age 6-59 months
who are anaemic (<11.0 g/dl)22 (%)', 'Non-pregnant women age 15-49 years who are
anaemic (<12.0 g/dl)22 (%)', 'Pregnant women age 15-49 years who are anaemic (<1
1.0 g/dl)22 (%)', 'All women age 15-49 years who are anaemic22 (%)', 'All women a
ge 15-19 years who are anaemic22 (%) ', 'Men age 15-49 years who are anaemic (<1
3.0 g/dl)22 (%)', 'Men age 15-19 years who are anaemic (<13.0 g/dl)22 (%)', 'Wome
n  age 15 years and above with high (141-160 mg/dl) Blood sugar level23 (%)', 'Wo
men age 15 years and above wih very high (>160 mg/dl) Blood sugar level23 (%)',
'Women age 15 years and above wih high or very high (>140 mg/dl) Blood sugar leve
l or taking medicine to control blood sugar level23 (%)', 'Men age 15 years and a
bove wih high (141-160 mg/dl) Blood sugar level23 (%)', 'Men (age 15 years and ab
ove wih  very high (>160 mg/dl) Blood sugar level23 (%)', 'Men age 15 years and a
bove wih high or very high (>140 mg/dl) Blood sugar level  or taking medicine to
control blood sugar level23 (%)', 'Women age 15 years and above wih Mildly elevat
ed blood pressure (Systolic 140-159 mm of Hg and/or Diastolic 90-99 mm of Hg)
(%)', 'Women age 15 years and above wih Moderately or severely elevated blood pre
ssure (Systolic ≥160 mm of Hg and/or Diastolic ≥100 mm of Hg) (%)', 'Women age 15
years and above wih Elevated blood pressure (Systolic ≥140 mm of Hg and/or Diasto
lic ≥90 mm of Hg) or taking medicine to control blood pressure (%)', 'Men age 15
years and above wih Mildly elevated blood pressure (Systolic 140-159 mm of Hg an
d/or Diastolic 90-99 mm of Hg) (%)', 'Men age 15 years and above wih Moderately o
r severely elevated blood pressure (Systolic ≥160 mm of Hg and/or Diastolic ≥100
mm of Hg) (%)', 'Men age 15 years and above wih Elevated blood pressure (Systolic
≥140 mm of Hg and/or Diastolic ≥90 mm of Hg) or taking medicine to control blood
pressure (%)', 'Women (age 30-49 years) Ever undergone a screening test for cervi
cal cancer (%)', 'Women (age 30-49 years) Ever undergone a breast examination for
breast cancer (%)', 'Women (age 30-49 years) Ever undergone an oral cavity examin
ation for oral cancer (%)', 'Men (age 30-49 years)Ever undergone an oral cavity e
xamination for oral cancer (%)', 'Women (age 15-49 years) who have comprehensive
knowledge24 of HIV/AIDS (%)', 'Men (age 15-49 years) who have comprehensive knowl
edge24 of HIV/AIDS (%)', 'Women (age 15-49 years) who know that consistent condom
use can reduce the chance of getting HIV/AIDS (%)', 'Men (age 15-49 years) who kn
ow that consistent condom use can reduce the chance of getting HIV/AIDS (%)', 'Cu
rrently married women (age 15-49 years) who usually participate in three househol
d decisions25 (%)', 'Women (age 15-49 years) who worked in the last 12 months and
were paid in cash (%)', 'Women (age 15-49 years) owning a house and/or land (alon
e or jointly with others) (%)', 'Women (age 15-49 years) having a bank or savings
account that they themselves use (%)', 'Women (age 15-49 years) having a mobile p
hone that they themselves use (%)', 'Women age 15-24 years who use hygienic metho
ds of protection during their menstrual period26 (%)', 'Ever-married women age 18
-49 years who have ever experienced spousal violence27 (%)', 'Ever-married women
age 18-49 years who have experienced physical violence during any pregnancy (%)',
'Young women age 18-29 years who experienced sexual violence by age 18 (%)', 'Wom
en age 15 years and above who use any kind of tobacco (%)', 'Men age 15 years and
above who use any kind of tobacco (%)', 'Women age 15 years and above who consume
alcohol (%)', 'Men age 15 years and above who consume alcohol (%)', 'Unnamed: 13
6']
```

In [4]:

```python
print(df_ncrb.info())
print(df_nfhs.isna().sum()) # missing vals per col

print(df_nfhs.info())
print(df_nfhs.isna().sum())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 22 columns):
 #   Column                                Non-Null Count  Dtype 
---  ------                                --------------  ----- 
 0   State/UT                              35 non-null     object
 1   Trafficking                           35 non-null     int64 
 2   Murder with Rape/Gang Rape            35 non-null     int64 
 3   Dowry Deaths                          35 non-null     int64 
 4   Abetment to Suicide of Women          35 non-null     int64 
 5   Miscarriage                           35 non-null     int64 
 6   Acid Attack                           35 non-null     int64 
 7   Attempt to Acid Attack                35 non-null     int64 
 8   Cruelty by Husband/relatives          35 non-null     int64 
 9   Kidnapping/Abduction                  35 non-null     int64 
 10  Selling of Minor Girls                35 non-null     int64 
 11  Buying of Minor Girls                 35 non-null     int64 
 12  Rape                                  35 non-null     int64 
 13  Attempt to Commit Rape                35 non-null     int64 
 14  Assault to Outrage her Modesty        35 non-null     int64 
 15  Insult to the Modesty of Women        35 non-null     int64 
 16  Assault due to Dowry                  35 non-null     int64 
 17  Domestic violence                     35 non-null     int64 
 18  Cyber Crimes committed against women  35 non-null     int64 
 19  Sexual Violence towards girl child    35 non-null     int64 
 20  Indecent Representation of Women      35 non-null     int64 
 21  Total Crime against Women (IPC &SLL)  35 non-null     int64 
dtypes: int64(21), object(1)
memory usage: 6.1+ KB
None
States/UTs                                                          0
Area                                                                0
Number of Households surveyed                                       0
Number of Women age 15-49 years interviewed                         0
Number of Men age 15-54 years interviewed                           0
                                                                ... 
Women age 15 years and above who use any kind of tobacco (%)        0
Men age 15 years and above who use any kind of tobacco (%)          0
Women age 15 years and above who consume alcohol (%)                0
Men age 15 years and above who consume alcohol (%)                  0
Unnamed: 136                                                      110
Length: 137, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110 entries, 0 to 109
Columns: 137 entries, States/UTs to Unnamed: 136
dtypes: float64(7), int64(3), object(127)
memory usage: 117.9+ KB
None
States/UTs                                                          0
Area                                                                0
Number of Households surveyed                                       0
Number of Women age 15-49 years interviewed                         0
Number of Men age 15-54 years interviewed                           0
                                                                ... 
Women age 15 years and above who use any kind of tobacco (%)        0
Men age 15 years and above who use any kind of tobacco (%)          0
Women age 15 years and above who consume alcohol (%)                0
Men age 15 years and above who consume alcohol (%)                  0
Unnamed: 136                                                      110
Length: 137, dtype: int64
```

State Names

In [5]:

```
print(df_ncrb.iloc[:,0].unique())
print(df_nfhs.iloc[:,0].unique())
```

```
['Andhra Pradesh' 'Arunachal pradesh' 'Assam' 'Bihar' 'Chhattisgarh' 'Goa'
 'Gujarat' 'Haryana' 'Himachal Pradesh' 'Jammu & Kashmir' 'Jharkhand'
 'Karnataka ' 'Kerala' 'Madhya Pradesh' 'Maharashtra' 'Manipur'
 'Meghalaya' 'Mizoram' 'Nagaland' 'Odisha' 'Punjab' 'Rajasthan' 'Sikkim'
 'Tamil Nadu' 'Telengana' 'Tripura' 'Uttar Pradesh' 'Uttarakhand'
 'West Bengal' 'Andaman & Nicobar Islands' 'Chandigarh'
 'D&N Haveli, Daman & Diu' 'Delhi' 'Lakshwadeep' 'Puducherry']
['India' 'Andaman & Nicobar Islands' 'Andhra Pradesh' 'Arunachal Pradesh'
 'Assam' 'Bihar' 'Chandigarh' 'Chhattisgarh'
 'Dadra and Nagar Haveli & Daman and Diu' 'Goa' 'Gujarat' 'Haryana'
 'Himachal Pradesh' 'Jammu & Kashmir' 'Jharkhand' 'Karnataka' 'Kerala'
 'Ladakh' 'Lakshadweep' 'Madhya Pradesh' 'Maharastra' 'Manipur'
 'Meghalaya' 'Mizoram' 'Nagaland' 'NCT of Delhi' 'Odisha' 'Puducherry'
 'Punjab' 'Rajasthan' 'Sikkim' 'Tamil Nadu' 'Telangana' 'Tripura'
 'Uttar Pradesh' 'Uttarakhand' 'West Bengal']
```

Cleaning Dataset

In [6]:

```
# standardizingstate names in both datasets

def clean_state_name(x):
    if isinstance(x, str):
        x = x.strip().title()            # no space nad sets title case
        x = x.replace("Andaman & Nicobar Islands", "Andaman And Nicobar Islands"
        x = x.replace("D&N Haveli, Daman & Diu", "Dadra And Nagar Haveli And Dam
        x = x.replace("D&N Haveli", "Dadra And Nagar Haveli")
        x = x.replace("Nct Of Delhi", "Delhi")
    return x

df_ncrb.iloc[:, 0] = df_ncrb.iloc[:, 0].apply(clean_state_name)
df_nfhs.iloc[:, 0] = df_nfhs.iloc[:, 0].apply(clean_state_name)

print("Cleaned NCRB states:\n", df_ncrb.iloc[:,0].unique())
print("\nCleaned NFHS states:\n", df_nfhs.iloc[:,0].unique())
```

```
Cleaned NCRB states:
 ['Andhra Pradesh' 'Arunachal Pradesh' 'Assam' 'Bihar' 'Chhattisgarh' 'Goa'
 'Gujarat' 'Haryana' 'Himachal Pradesh' 'Jammu & Kashmir' 'Jharkhand'
 'Karnataka' 'Kerala' 'Madhya Pradesh' 'Maharashtra' 'Manipur' 'Meghalaya'
 'Mizoram' 'Nagaland' 'Odisha' 'Punjab' 'Rajasthan' 'Sikkim' 'Tamil Nadu'
 'Telengana' 'Tripura' 'Uttar Pradesh' 'Uttarakhand' 'West Bengal'
 'Andaman And Nicobar Islands' 'Chandigarh'
 'Dadra And Nagar Haveli And Daman And Diu' 'Delhi' 'Lakshwadeep'
 'Puducherry']

Cleaned NFHS states:
 ['India' 'Andaman And Nicobar Islands' 'Andhra Pradesh'
 'Arunachal Pradesh' 'Assam' 'Bihar' 'Chandigarh' 'Chhattisgarh'
 'Dadra And Nagar Haveli & Daman And Diu' 'Goa' 'Gujarat' 'Haryana'
 'Himachal Pradesh' 'Jammu & Kashmir' 'Jharkhand' 'Karnataka' 'Kerala'
 'Ladakh' 'Lakshadweep' 'Madhya Pradesh' 'Maharastra' 'Manipur'
 'Meghalaya' 'Mizoram' 'Nagaland' 'Delhi' 'Odisha' 'Puducherry' 'Punjab'
 'Rajasthan' 'Sikkim' 'Tamil Nadu' 'Telangana' 'Tripura' 'Uttar Pradesh'
 'Uttarakhand' 'West Bengal']
```

In [7]:

```python
# compare state lists after cleaning

ncrb_states = set(df_ncrb.iloc[:,0].unique())
nfhs_states = set(df_nfhs.iloc[:,0].unique())

missing_in_nfhs = ncrb_states - nfhs_states
missing_in_ncrb = nfhs_states - ncrb_states

print("States present in NCRB but missing in NFHS:\n", missing_in_nfhs)
print("\nStates present in NFHS but missing in NCRB:\n", missing_in_ncrb)
```

```
States present in NCRB but missing in NFHS:
 {'Dadra And Nagar Haveli And Daman And Diu', 'Lakshwadeep', 'Maharashtra', 'Tele
ngana'}

States present in NFHS but missing in NCRB:
 {'Lakshadweep', 'Telangana', 'India', 'Maharastra', 'Ladakh', 'Dadra And Nagar H
aveli & Daman And Diu'}
```

In [8]:

```python
# fix mismatched state names manually

# corrections for NCRB
df_ncrb.iloc[:,0] = df_ncrb.iloc[:,0].replace({
    "Telengana": "Telangana",
    "Lakshwadeep": "Lakshadweep",
    "Dadra And Nagar Haveli And Daman And Diu": "Dadra And Nagar Haveli & Daman
})

# corrections for NFHS
df_nfhs.iloc[:,0] = df_nfhs.iloc[:,0].replace({
    "Maharastra": "Maharashtra",
    "Dadra And Nagar Haveli & Daman And Diu": "Dadra And Nagar Haveli & Daman An
})

# remove 'India' row (not a state)
df_nfhs = df_nfhs[df_nfhs.iloc[:,0] != "India"]

# check again after fixing
ncrb_states = set(df_ncrb.iloc[:,0].unique())
nfhs_states = set(df_nfhs.iloc[:,0].unique())
```

```
print("Remaining mismatches NCRB → NFHS:\n", ncrb_states - nfhs_states)
print("\nRemaining mismatches NFHS → NCRB:\n", nfhs_states - ncrb_states)
```

```
Remaining mismatches NCRB → NFHS:
 set()

Remaining mismatches NFHS → NCRB:
 {'Ladakh'}
```

In [9]:

```
# remove Ladakh since NCRB does not contain it
df_nfhs = df_nfhs[df_nfhs.iloc[:,0] != "Ladakh"]

# re-check mismatch
ncrb_states = set(df_ncrb.iloc[:,0].unique())
nfhs_states = set(df_nfhs.iloc[:,0].unique())

print("Remaining mismatches NCRB → NFHS:\n", ncrb_states - nfhs_states)
print("\nRemaining mismatches NFHS → NCRB:\n", nfhs_states - ncrb_states)
```

```
Remaining mismatches NCRB → NFHS:
 set()

Remaining mismatches NFHS → NCRB:
 set()
```

(here) Merge Datasets

In [10]:

```
# filter NFHS to keep only "Total" area data (not urban/rural separately)
df_nfhs_total = df_nfhs[df_nfhs['Area'] == 'Total']

print(f"NFHS rows after filtering for 'Total' area: {df_nfhs_total.shape[0]}")

# identify state column names
state_col_ncrb = df_ncrb.columns[0]
state_col_nfhs = df_nfhs_total.columns[0]

print("NCRB state column:", state_col_ncrb)
print("NFHS state column:", state_col_nfhs)

# merge datasets on the state column
df_merged = pd.merge(df_ncrb, df_nfhs_total,
                     left_on=state_col_ncrb,
                     right_on=state_col_nfhs,
                     how='inner')

print("Merged dataset shape:", df_merged.shape)
df_merged.head()
```

```
NFHS rows after filtering for 'Total' area: 34
NCRB state column: State/UT
NFHS state column: States/UTs
Merged dataset shape: (34, 159)
```

Out[10]:

	State/UT	Trafficking	Murder with Rape/Gang Rape	Dowry Deaths	Abetment to Suicide of Women	Miscarriage	Acid Attack	Attem to Ac Atta
0	Andhra Pradesh	107	8	100	358	4	3	
1	Arunachal Pradesh	1	0	0	0	0	0	
2	Assam	78	14	175	75	2	3	
3	Bihar	87	0	1057	2	0	3	
4	Chhattisgarh	13	7	57	149	5	0	

5 rows × 159 columns

FROOOOM HEEEREE

In [11]:

```
# function to find column by keyword
def find_column(df, keyword):
    matches = [col for col in df.columns if keyword.lower() in col.lower()]
    return matches[0] if matches else None

# rebuild columns_to_keep using flexible matching
columns_to_keep = {}

# state identifier
columns_to_keep['State/UT'] = 'state'

# crime data (exact names)
columns_to_keep['Total Crime against Women (IPC &SLL)'] = 'total_crimes'
columns_to_keep['Dowry Deaths'] = 'dowry_deaths'
columns_to_keep['Cruelty by Husband/relatives'] = 'domestic_cruelty'
columns_to_keep['Rape'] = 'rape'
columns_to_keep['Kidnapping/Abduction'] = 'kidnapping'
columns_to_keep['Cyber Crimes committed against women'] = 'cyber_crimes'
columns_to_keep['Sexual Violence towards girl child'] = 'child_sexual_violence'
columns_to_keep['Domestic violence'] = 'domestic_violence'

# edducation - find by keyword
edu_cols = {
    'female_school_attendance': 'ever attended school',
    'female_literacy': 'who are literate',
    'female_higher_education': '10 or more years of schooling'
```

```
}

for new_name, keyword in edu_cols.items():
    col = find_column(df_merged, keyword)
    if col:
        columns_to_keep[col] = new_name

# economic empowerment
econ_cols = {
    'women_paid_work': 'paid in cash',
    'women_property_ownership': 'owning a house',
    'women_bank_account': 'bank or savings account',
    'women_mobile_phone': 'mobile phone that they themselves use'
}

for new_name, keyword in econ_cols.items():
    col = find_column(df_merged, keyword)
    if col:
        columns_to_keep[col] = new_name

# decision-making
col = find_column(df_merged, 'three household decisions')
if col:
    columns_to_keep[col] = 'household_decision_power'

# violence indicators
violence_cols = {
    'spousal_violence': 'experienced spousal violence',
    'pregnancy_violence': 'violence during any pregnancy',
    'youth_sexual_violence': 'sexual violence by age 18'
}

for new_name, keyword in violence_cols.items():
    col = find_column(df_merged, keyword)
    if col:
        columns_to_keep[col] = new_name

# health/social
health_cols = {
    'child_marriage': 'married before age 18',
    'menstrual_hygiene': 'hygienic methods'
}

for new_name, keyword in health_cols.items():
    col = find_column(df_merged, keyword)
    if col:
        columns_to_keep[col] = new_name

# demographics
col = find_column(df_merged, 'Sex ratio of the total population')
if col:
    columns_to_keep[col] = 'sex_ratio'

col = find_column(df_merged, 'Number of Women age 15-49 years interviewed')
if col:
    columns_to_keep[col] = 'women_surveyed'

# create cleaned dataframe
df_clean = df_merged[list(columns_to_keep.keys())].rename(columns=columns_to_kee
print(f"Cleaned dataset shape: {df_clean.shape}")
print(f"\nColumns kept: {df_clean.columns.tolist()}")
df_clean.head()
```

```
Cleaned dataset shape: (34, 24)

Columns kept: ['state', 'total_crimes', 'dowry_deaths', 'domestic_cruelty', 'rap
e', 'kidnapping', 'cyber_crimes', 'child_sexual_violence', 'domestic_violence',
'female_school_attendance', 'female_literacy', 'female_higher_education', 'women_
paid_work', 'women_property_ownership', 'women_bank_account', 'women_mobile_phon
e', 'household_decision_power', 'spousal_violence', 'pregnancy_violence', 'youth_
sexual_violence', 'child_marriage', 'menstrual_hygiene', 'sex_ratio', 'women_surv
eyed']
```

Out[11]:

	state	total_crimes	dowry_deaths	domestic_cruelty	rape	kidnapping	cyber_
0	Andhra Pradesh	25503	100	11964	621	592	
1	Arunachal Pradesh	335	0	74	74	48	
2	Assam	14148	175	4704	1113	3466	
3	Bihar	20222	1057	1850	881	10190	
4	Chhattisgarh	8693	57	942	1246	2121	

5 rows × 24 columns

In [12]:

```
# check current data types
print(df_clean.dtypes)

# convert all percentage/numeric columns to proper numeric type
numeric_cols = df_clean.columns.drop('state')  # all except state
for col in numeric_cols:
    df_clean[col] = pd.to_numeric(df_clean[col], errors='coerce')

print("\nAfter conversion:")
print(df_clean.dtypes)
```

```
state                        object
total_crimes                  int64
dowry_deaths                  int64
domestic_cruelty              int64
rape                          int64
kidnapping                    int64
cyber_crimes                  int64
child_sexual_violence         int64
domestic_violence             int64
female_school_attendance     object
female_literacy              object
female_higher_education      object
women_paid_work              object
women_property_ownership     object
women_bank_account           object
women_mobile_phone           object
household_decision_power     object
spousal_violence             object
pregnancy_violence           object
youth_sexual_violence        object
child_marriage               object
menstrual_hygiene            object
sex_ratio                   float64
women_surveyed                int64
dtype: object

After conversion:
state                        object
total_crimes                  int64
dowry_deaths                  int64
domestic_cruelty              int64
rape                          int64
kidnapping                    int64
cyber_crimes                  int64
child_sexual_violence         int64
domestic_violence             int64
female_school_attendance    float64
female_literacy             float64
female_higher_education     float64
women_paid_work             float64
women_property_ownership    float64
women_bank_account          float64
women_mobile_phone          float64
household_decision_power    float64
spousal_violence            float64
pregnancy_violence          float64
youth_sexual_violence       float64
child_marriage              float64
menstrual_hygiene           float64
sex_ratio                   float64
women_surveyed                int64
dtype: object
```

In [15]:

```
# check for missing values
print("Missing values per column:")
print(df_clean.isnull().sum())
print(f"\nTotal missing values: {df_clean.isnull().sum().sum()}")


# CRIME COMPOSITION ANALYSIS
# Raw crime counts are misleading

crime_cols = [
    'dowry_deaths',
    'domestic_cruelty',
    'rape',
    'kidnapping',
    'cyber_crimes',
    'child_sexual_violence',
    'domestic_violence'
]

print("\n" + "-" * 50)
print("Creating crime composition (% share of total crimes)...")

for col in crime_cols:
    df_clean[f'{col}_share'] = (df_clean[col] / df_clean['total_crimes']) * 100


# raw crimes vs crime composition
print("\nExample: Raw crimes vs Crime Composition")
print(
    df_clean[['state', 'total_crimes'] + [f'{c}_share' for c in crime_cols]]
    .sort_values('total_crimes', ascending=False)
    .head(5)
)


# handle missing values (if any)
if df_clean.isnull().sum().sum() > 0:
    print("\nStates with missing values:")
    missing_rows = df_clean[df_clean.isnull().any(axis=1)]
    print(missing_rows.to_string())

    numeric_cols = df_clean.select_dtypes(include=['float64', 'int64']).columns
    df_clean[numeric_cols] = df_clean[numeric_cols].fillna(
        df_clean[numeric_cols].median()
    )
    print("\nMissing values filled with column medians")


# create composite indices
print("\n" + "-" * 50)
print("Creating Empowerment & Safety Indices...")

# empowerment index (higher = better)
empowerment_indicators = [
    'female_literacy',
    'female_higher_education',
    'women_paid_work',
    'women_property_ownership',
    'women_bank_account',
    'household_decision_power'
]

df_clean['empowerment_index'] = df_clean[empowerment_indicators].mean(axis=1)


# safety index (higher = safer, lower violence prevalence)
violence_indicators = [
    'spousal_violence',
    'pregnancy_violence',
    'youth_sexual_violence'
]

df_clean['safety_index'] = 100 - df_clean[violence_indicators].mean(axis=1)


# summary statistics
print("\n" + "-" * 50)
print("SUMMARY STATISTICS")
print("-" * 50)

print("\nMost empowered states:")
print(df_clean.nlargest(5, 'empowerment_index')[['state', 'empowerment_index']])

print("\nSafest states (lowest violence prevalence):")
print(df_clean.nlargest(5, 'safety_index')[['state', 'safety_index']])

print("\nStates with highest total crimes against women:")
print(df_clean.nlargest(5, 'total_crimes')[['state', 'total_crimes']])


# FINAL CLEANUP: remove any leftover crime rate columns (important!!!)
rate_cols = [col for col in df_clean.columns if col.endswith('_rate')]
df_clean.drop(columns=rate_cols, inplace=True)

print(f"\nDropped rate columns: {rate_cols}")


# save cleaned data
df_clean.to_csv('women_empowerment_cleaned.csv', index=False)

print("\nCleaned data saved to 'women_empowerment_cleaned.csv'")
print(f"\nFinal dataset shape: {df_clean.shape}")
print(f"Total columns: {df_clean.shape[1]}")
```

```
Missing values per column:
state                          0
total_crimes                   0
dowry_deaths                   0
domestic_cruelty               0
rape                           0
kidnapping                     0
cyber_crimes                   0
child_sexual_violence          0
domestic_violence              0
female_school_attendance       0
female_literacy                0
female_higher_education        0
women_paid_work                0
women_property_ownership       0
women_bank_account             0
women_mobile_phone             0
household_decision_power       0
spousal_violence               0
pregnancy_violence             0
youth_sexual_violence          0
child_marriage                 0
menstrual_hygiene              0
sex_ratio                      0
women_surveyed                 0
dowry_deaths_share             0
domestic_cruelty_share         0
rape_share                     0
kidnapping_share               0
cyber_crimes_share             0
child_sexual_violence_share    0
domestic_violence_share        0
empowerment_index              0
safety_index                   0
dtype: int64

Total missing values: 0

--------------------------------------------------
Creating crime composition (% share of total crimes)...

Example: Raw crimes vs Crime Composition
             state  total_crimes  dowry_deaths_share  domestic_cruelty_share  \
26   Uttar Pradesh         65743            3.252057               30.985808   
14     Maharashtra         45331            0.397079               25.075555   
21       Rajasthan         45058            1.000932               41.828310   
13  Madhya Pradesh         32765            1.580955               25.899588   
0   Andhra Pradesh         25503            0.392111               46.912128   

    rape_share  kidnapping_share  cyber_crimes_share  \
26    5.612765         22.644236            0.695131   
14    6.406212         20.509144            0.255896   
21   11.984553         14.618936            0.368414   
13    9.244621         24.294216            0.326568   
0     2.435008          2.321296            0.423480   

    child_sexual_violence_share  domestic_violence_share  
26                    12.100148                 0.004563  
14                    16.472171                 0.002206  
21                     8.165032                 0.006658  
```

```
13                    18.113841                0.030520
0                      8.340195                0.000000

--------------------------------------------------
Creating Empowerment & Safety Indices...

--------------------------------------------------
SUMMARY STATISTICS
--------------------------------------------------

Most empowered states:
         state  empowerment_index
33  Puducherry          70.416667
23  Tamil Nadu          69.366667
15     Manipur          67.500000
11   Karnataka          67.150000
5          Goa          66.833333

Safest states (lowest violence prevalence):
               state  safety_index
32       Lakshadweep     99.300000
18          Nagaland     97.200000
8   Himachal Pradesh     96.200000
29        Chandigarh     96.166667
12            Kerala     96.000000

States with highest total crimes against women:
             state  total_crimes
26   Uttar Pradesh         65743
14     Maharashtra         45331
21       Rajasthan         45058
13  Madhya Pradesh         32765
0   Andhra Pradesh         25503

Dropped rate columns: []

Cleaned data saved to 'women_empowerment_cleaned.csv'

Final dataset shape: (34, 33)
Total columns: 33
```

```
In [16]: [col for col in df_clean.columns if 'rate' in col]
```

```
Out[16]: []
```

```
In [ ]:
```