Name: Tanya Singh **PRN: 22070521013**

Department: CSE **Division: A**

Symbiosis Institute of Technology, Nagpur

॥वसुधैव कुटुम्बकम्॥

DATA SCIENCE

Computer Science and Engineering
Batch 2022-26

Course Name : Data Science

Course Code:- 0705210707 **Semester-VII**

Data Science (CA1-EDA)

Name: TANYA SINGH
PRN: 22070521013

1. Introduction:

This Exploratory Data Analysis (EDA) investigates interlinked public health and socio-demographic indicators across Indian states using the National Family Health Survey (NFHS) 2014-15 & 2019–20 dataset. The analysis draws correlations between male health risks (such as tobacco and alcohol consumption, and blood pressure severity) and chronic illnesses like oral and cervical cancer. States such as Mizoram, Andaman & Nicobar Islands, and Manipur rank highest in male tobacco usage, which often coexists with elevated alcohol intake and increased cancer prevalence.

Parallelly, we explore critical women-centric issues: the relationship between child marriage (women married before age 18) and teenage pregnancy (ages 15–19). A strong positive correlation (r = 0.72) was found, suggesting that early marriage remains a key driver of adolescent pregnancies. These indicators were further cross-analyzed against female literacy levels (represented by schooling rate of girls past grade 6), revealing that states with lower education levels show higher rates of early pregnancy and associated health vulnerabilities.

Through correlation heatmaps, scatter plots, and state-level aggregation, we identify patterns and outliers. For instance, some northeastern states show both high male tobacco use and low female literacy, placing them in a double-risk category. This multi-dimensional analysis underscores the gendered burden of health inequality and the pressing need for targeted education and lifestyle interventions.

2. Dataset Overview:

The analysis is based on the National Family Health Survey (NFHS) – 5, conducted in 2019–2020. This survey provides state- and district-level data on population health, nutrition, reproductive behavior, and lifestyle factors across India.

- Source: India Data Portal
- Time Period: July 2014 to March 2015 & July 2019 to March 2020
- Scope: Nationwide, covering all states and union territories
- Rows (Observations): 1300 rows, each representing a district-level data point
- Columns (Features): 110 features, of which 15 were selected for focused EDA

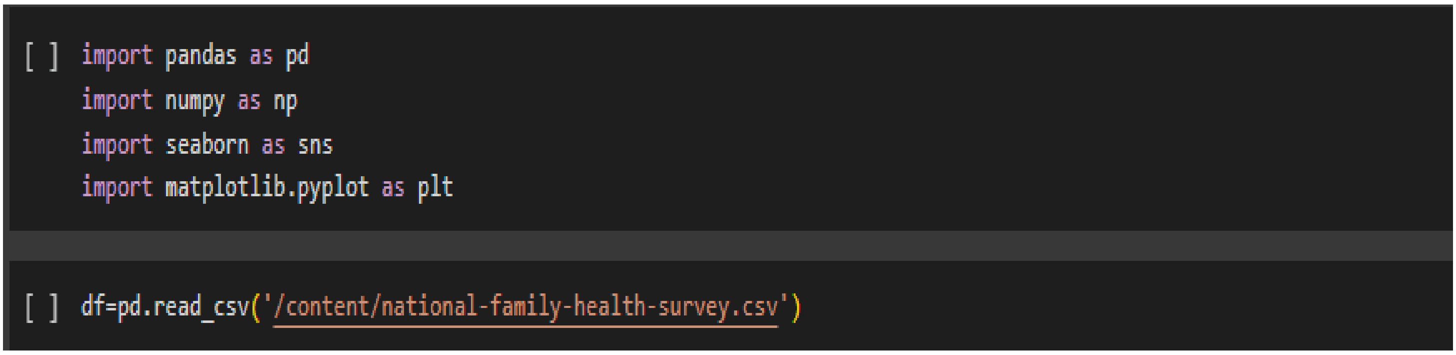
Selected Variables:

- Male health: tobaco_men_15, alcohol_men_15, men_bp_sev
- Female health: tobaco_women_15, alcohol_women_15, oral_cancer, cerv_cancer, fem_15_19_pregnant, fem_below24_married_before18
- Schooling: pop_f_6_sch
- Population Dynamics: pop_below_15, sex_ratio_tot_pop
- Mortality & Fertility: births_5yrs_hgh_order, deaths_last_3years

These variables were selected to investigate interrelations among lifestyle factors, chronic disease prevalence, and reproductive health outcomes.

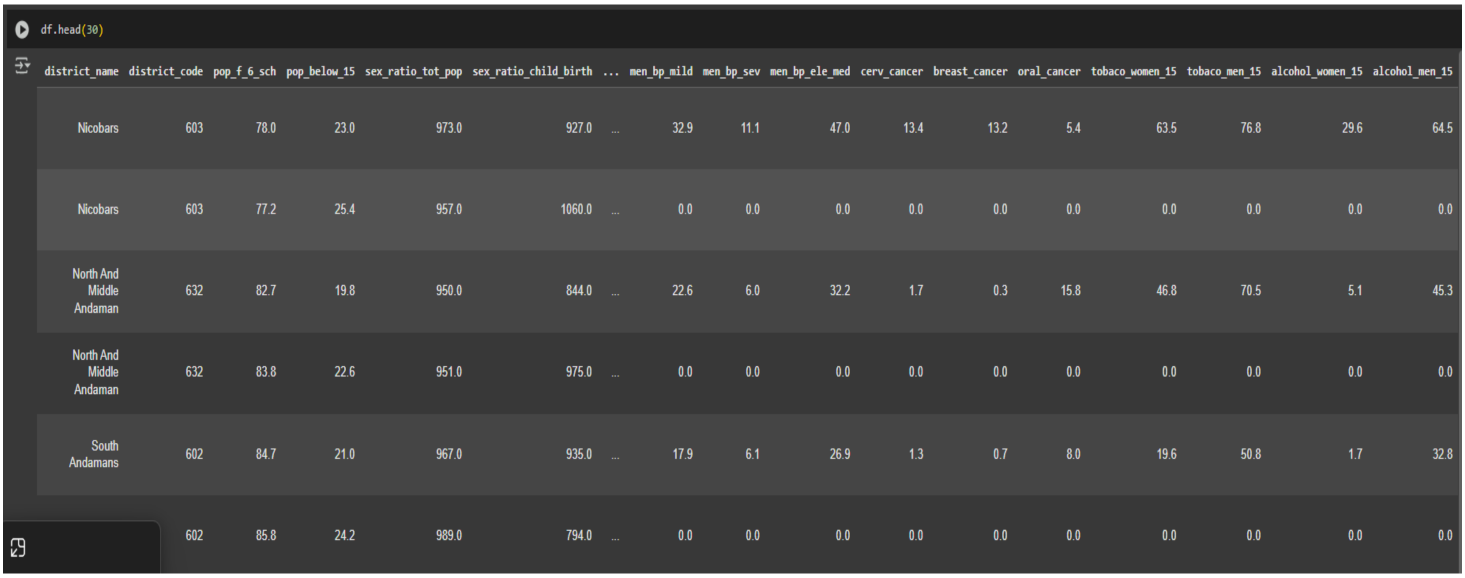
3. Importing Libraries and Loading the Dataset:

- Imported essential Python libraries such as pandas, numpy, matplotlib, seaborn, for interactive visualization.
- Loaded the dataset from the CSV file into a pandas DataFrame.
- Verified the shape, column names, and basic details using functions like **.shape, .head(), info(),** and **describe().**



```
[ ] import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] df=pd.read_csv('/content/national-family-health-survey.csv')
```



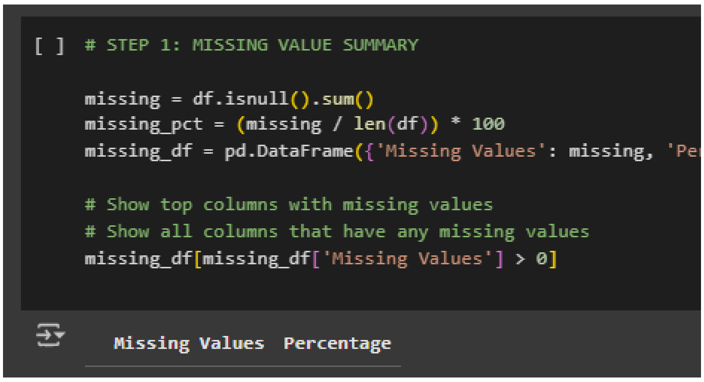
```
df.head(30)
```

district_name	district_code	pop_f_6_sch	pop_below_15	sex_ratio_tot_pop	sex_ratio_child_birth	...	men_bp_mild	men_bp_sev	men_bp_ele_med	cerv_cancer	breast_cancer	oral_cancer	tobaco_women_15	tobaco_men_15	alcohol_women_15	alcohol_men_15
Nicobars	603	78.0	23.0	973.0	927.0	...	32.9	11.1	47.0	13.4	13.2	5.4	63.5	76.8	29.6	64.5
Nicobars	603	77.2	25.4	957.0	1060.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
North And Middle Andaman	632	82.7	19.8	950.0	844.0	...	22.6	6.0	32.2	1.7	0.3	15.8	46.8	70.5	5.1	45.3
North And Middle Andaman	632	83.8	22.6	951.0	975.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
South Andamans	602	84.7	21.0	967.0	935.0	...	17.9	6.1	26.9	1.3	0.7	8.0	19.6	50.8	1.7	32.8
	602	85.8	24.2	989.0	794.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

4. Data Cleaning and Preprocessing:

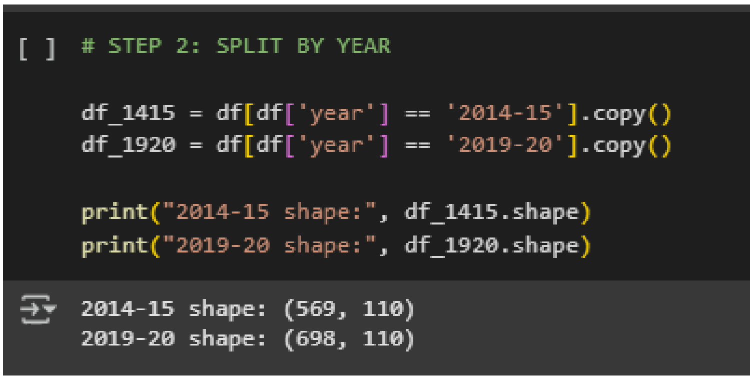
4.1 Checking for Missing Values

The dataset was first inspected for missing values, and it was found that there were no missing values. We divided the dataset into 2 parts based on the year group (14-15 & 19-20). Further, to ensure data quality, only relevant columns were selected for analysis based on the project objectives. A correlation heatmap was generated to evaluate inter-feature relationships. This allowed us to retain only those variables that showed meaningful patterns while discarding those that exhibited weak or redundant influence.



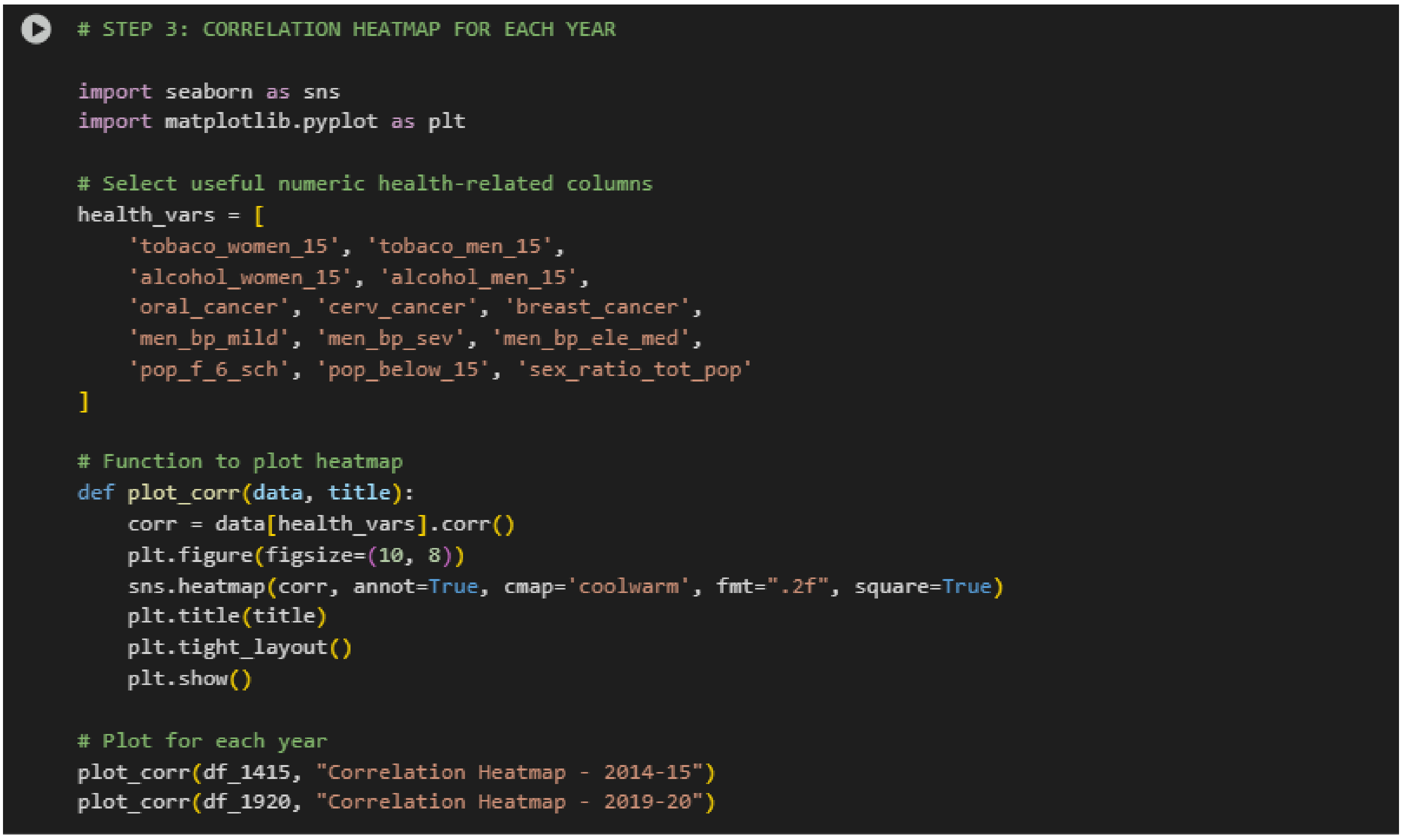
```
[ ]  # STEP 1: MISSING VALUE SUMMARY

     missing = df.isnull().sum()
     missing_pct = (missing / len(df)) * 100
     missing_df = pd.DataFrame({'Missing Values': missing, 'Pe

     # Show top columns with missing values
     # Show all columns that have any missing values
     missing_df[missing_df['Missing Values'] > 0]
```

```
       Missing Values  Percentage
```

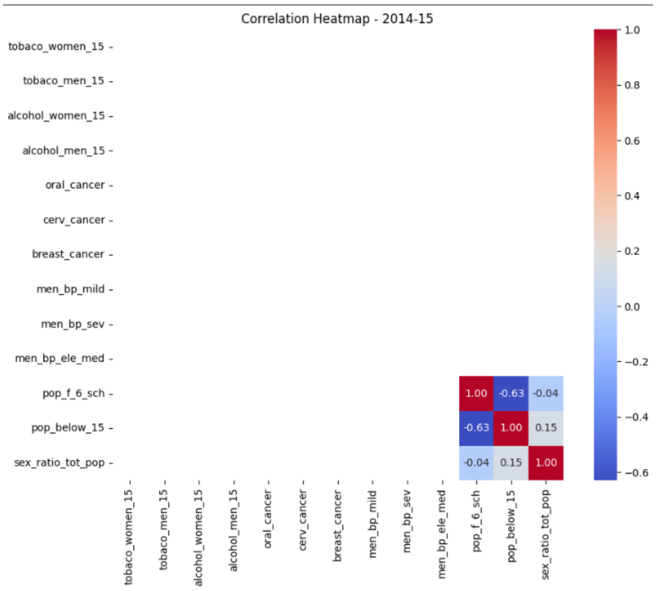


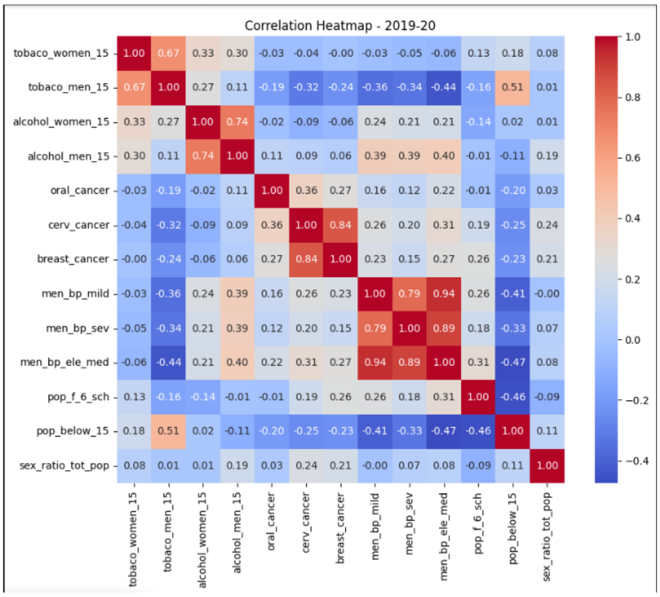
```
[ ]  # STEP 2: SPLIT BY YEAR

     df_1415 = df[df['year'] == '2014-15'].copy()
     df_1920 = df[df['year'] == '2019-20'].copy()

     print("2014-15 shape:", df_1415.shape)
     print("2019-20 shape:", df_1920.shape)
```

```
     2014-15 shape: (569, 110)
     2019-20 shape: (698, 110)
```



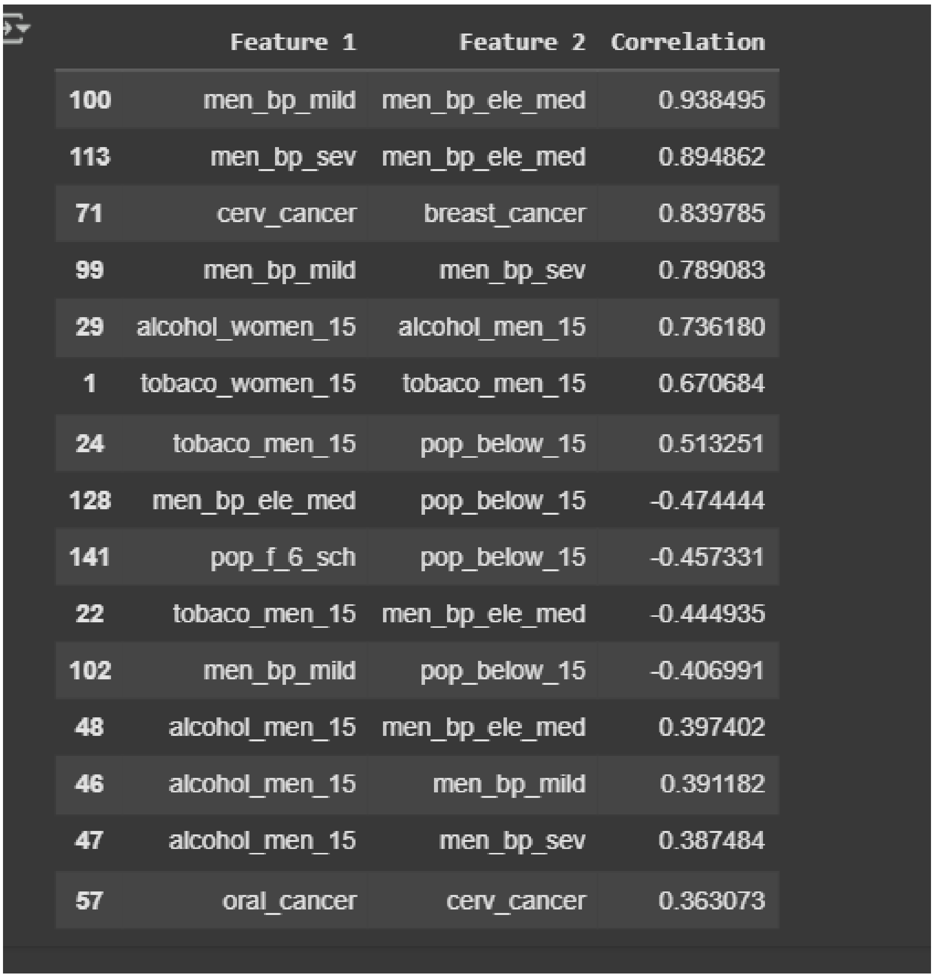
```
# STEP 3: CORRELATION HEATMAP FOR EACH YEAR

import seaborn as sns
import matplotlib.pyplot as plt

# Select useful numeric health-related columns
health_vars = [
    'tobaco_women_15', 'tobaco_men_15',
    'alcohol_women_15', 'alcohol_men_15',
    'oral_cancer', 'cerv_cancer', 'breast_cancer',
    'men_bp_mild', 'men_bp_sev', 'men_bp_ele_med',
    'pop_f_6_sch', 'pop_below_15', 'sex_ratio_tot_pop'
]

# Function to plot heatmap
def plot_corr(data, title):
    corr = data[health_vars].corr()
    plt.figure(figsize=(10, 8))
    sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", square=True)
    plt.title(title)
    plt.tight_layout()
    plt.show()

# Plot for each year
plot_corr(df_1415, "Correlation Heatmap - 2014-15")
plot_corr(df_1920, "Correlation Heatmap - 2019-20")
```





The 2014–15 dataset contained mostly zero values, indicating poor data recording and rendering it unsuitable for meaningful analysis. So I decided to focus first on 19-20 dataset.

5. EDA Insights

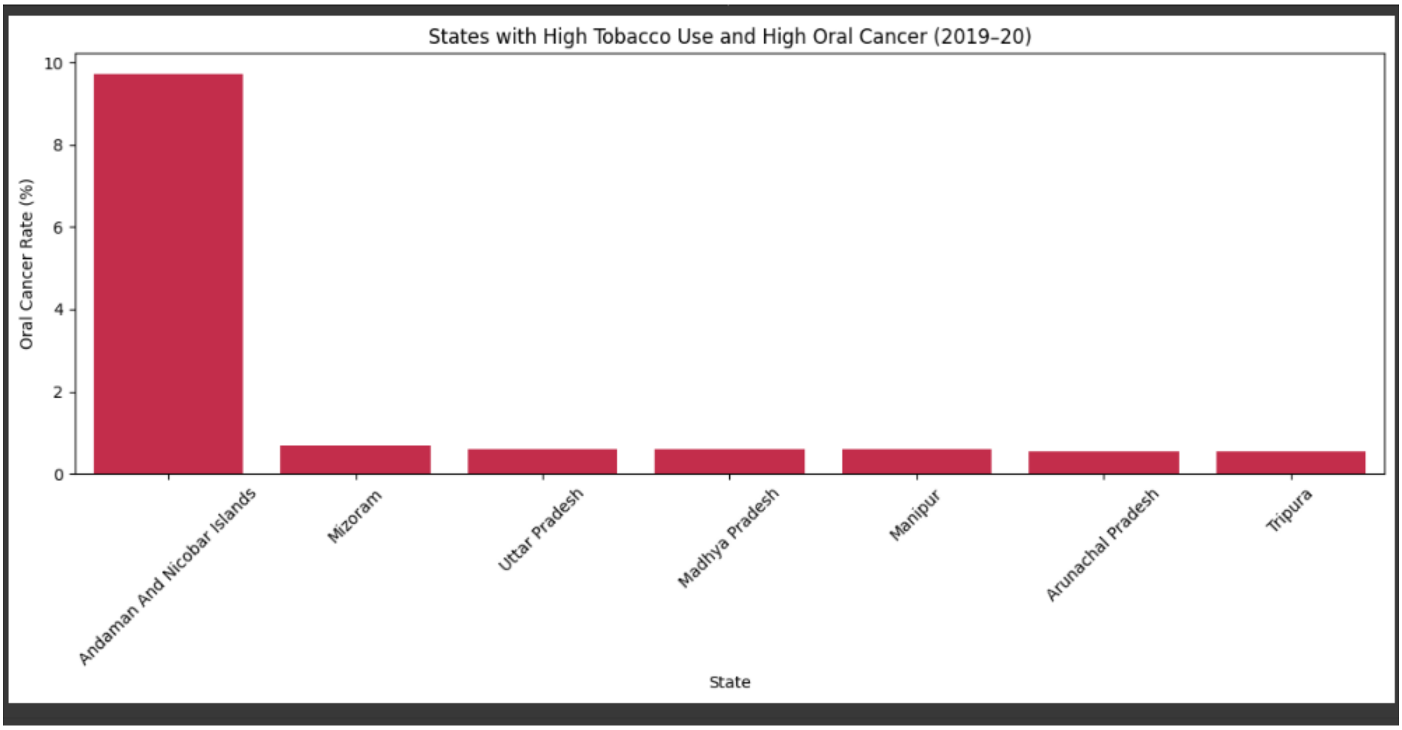


	Feature 1	Feature 2	Correlation
100	men_bp_mild	men_bp_ele_med	0.938495
113	men_bp_sev	men_bp_ele_med	0.894862
71	cerv_cancer	breast_cancer	0.839785
99	men_bp_mild	men_bp_sev	0.789083
29	alcohol_women_15	alcohol_men_15	0.736180
1	tobaco_women_15	tobaco_men_15	0.670684
24	tobaco_men_15	pop_below_15	0.513251
128	men_bp_ele_med	pop_below_15	-0.474444
141	pop_f_6_sch	pop_below_15	-0.457331
22	tobaco_men_15	men_bp_ele_med	-0.444935
102	men_bp_mild	pop_below_15	-0.406991
48	alcohol_men_15	men_bp_ele_med	0.397402
46	alcohol_men_15	men_bp_mild	0.391182
47	alcohol_men_15	men_bp_sev	0.387484
57	oral_cancer	cerv_cancer	0.363073

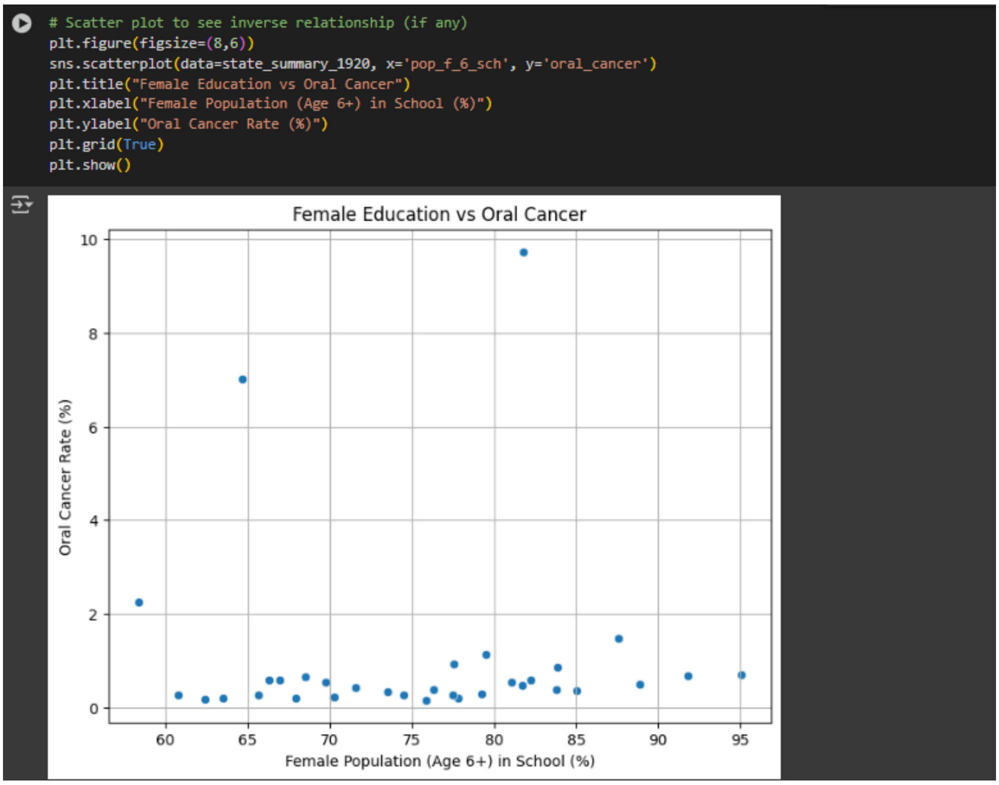
These were the most meaningful correlation. Below is an explanation for all of them.

1. men_bp_mild & men_bp_ele_med 0.94 Very strong link: districts with more men having mildly high BP also have more men on medication for elevated BP. Obvious, but confirms progression
2. men_bp_sev & men_bp_ele_med 0.89 Severe BP correlates strongly with medical intervention. Again, expected — but helps confirm data quality.
3. cerv_cancer & breast_cancer 0.84 Strong: districts with high cervical cancer rates also see high breast cancer. Female health issues cluster. Possibly due to poor awareness or screening.
4. men_bp_mild & men_bp_sev 0.79 Mild and severe BP rise together in regions — likely poor health management.
5. alcohol_women_15 & alcohol_men_15 0.74 Strong social/cultural indicator. Districts where men drink more, women often do too. Shared norms.
6. tobaco_women_15 & tobaco_men_15 0.67 Similar story: where tobacco use is culturally accepted, both genders show high use
7. tobaco_men_15 & pop_below_15 0.51 Q: more tobacco use where child population is higher? Could signal less education/awareness or socio-economic burden. Worth investigating
8. men_bp_ele_med & pop_below_15- 0.47 Inverse: districts with younger populations have fewer people on BP meds (likely because BP issues rise with age)
9. pop_f_6_sch & pop_below_15- 0.46 Educated districts → fewer children. Logical: education correlates with lower birthrates.
10. tobaco_men_15 & men_bp_ele_med- 0.44 Surprisingly inverse — maybe tobacco-heavy districts don't access BP medication? Could be poor healthcare access.
11. men_bp_mild & pop_below_15- 0.40 Fewer mild BP cases in districts with high child population (again, younger populations = healthier BP)
12. alcohol_men_15 & men_bp_ele_med 0.39 Clear pattern: more male alcohol use → more men on BP medication.
13. alcohol_men_15 & men_bp_mild 0.39 Same as above — early BP impact visible even at mild levels.

14. alcohol_men_15 & men_bp_sev 0.39 Alcohol might be a direct or contributing factor to worsening BP
15. oral_cancer & cerv_cancer 0.36 Weak-moderate link — suggests districts with poor screening or healthcare may miss both. Not biologically related, but signals systemic issue.



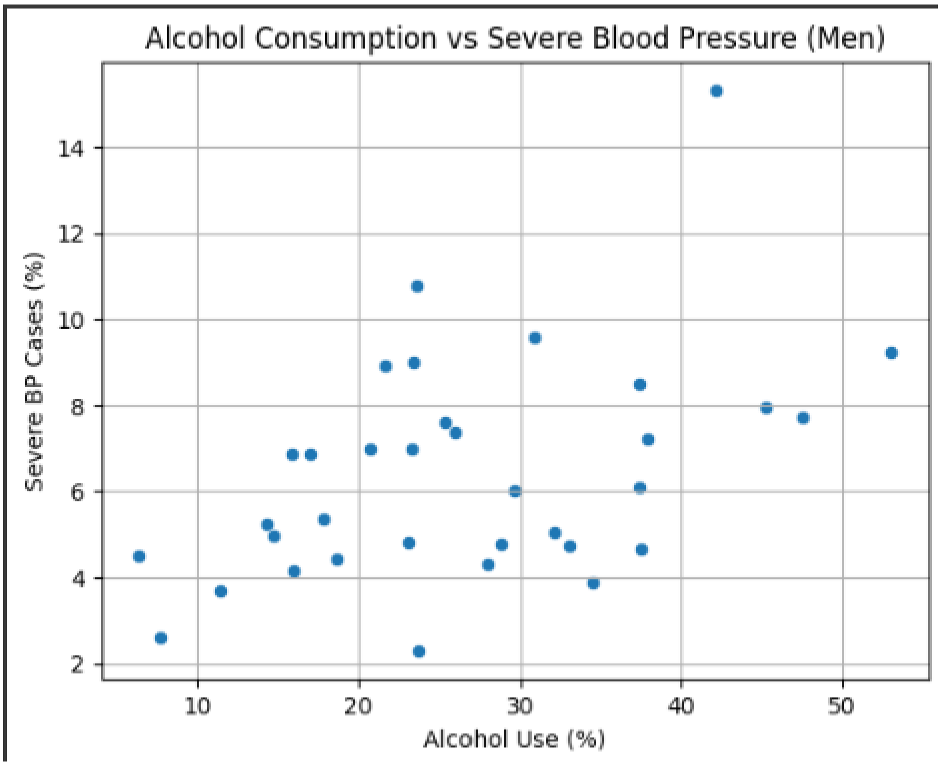
Andaman & Nicobar Islands had the highest tobacco use and Oral Cancer Rate surpassing other states by a large margin.



```
# Scatter plot to see inverse relationship (if any)
plt.figure(figsize=(8,6))
sns.scatterplot(data=state_summary_1920, x='pop_f_6_sch', y='oral_cancer')
plt.title("Female Education vs Oral Cancer")
plt.xlabel("Female Population (Age 6+) in School (%)")
plt.ylabel("Oral Cancer Rate (%)")
plt.grid(True)
plt.show()
```

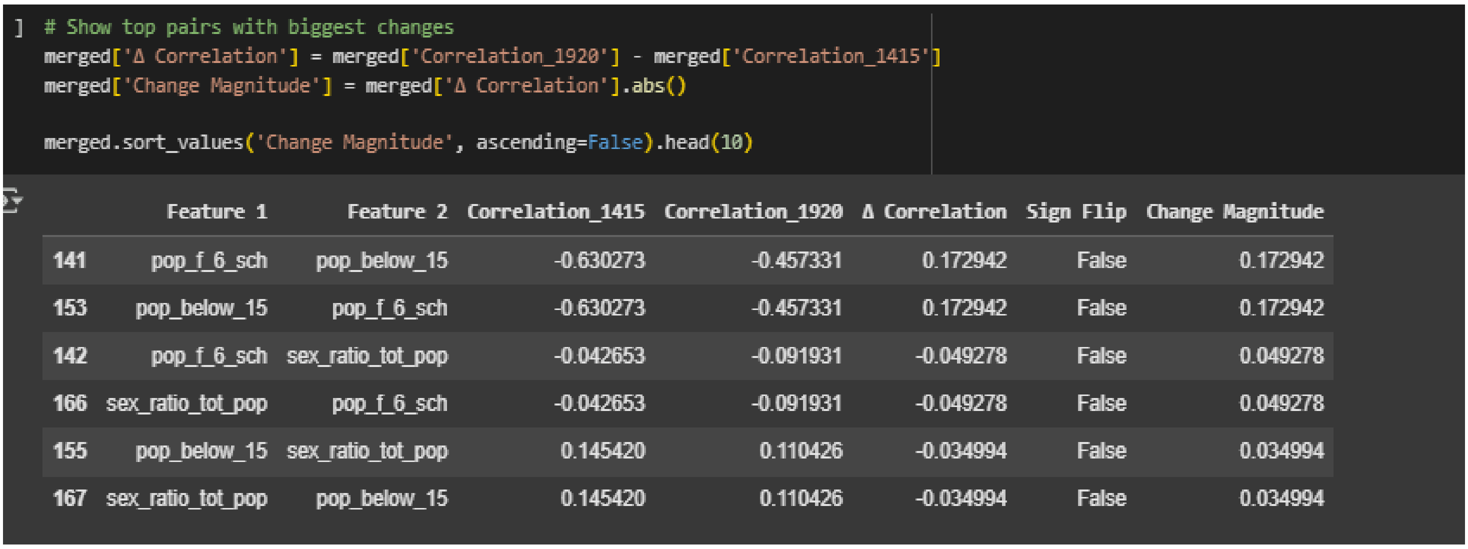
Female Education vs Oral Cancer
Oral Cancer Rate (%)
10
8
6
4
2
0
60
65
70
75
80
85
90
95
Female Population (Age 6+) in School (%)

This scatter plot shows the relationship between female education levels (percentage of females aged 6+ in school) and oral cancer rates across states. A general downward trend is visible, suggesting an inverse relationship states with higher female education tend to have lower oral cancer rates. However, a few outliers exist where oral cancer remains high despite moderate to high education levels. This could indicate other contributing factors like tobacco or alcohol use. Overall, the trend supports the idea that better education may be associated with healthier lifestyle choices.



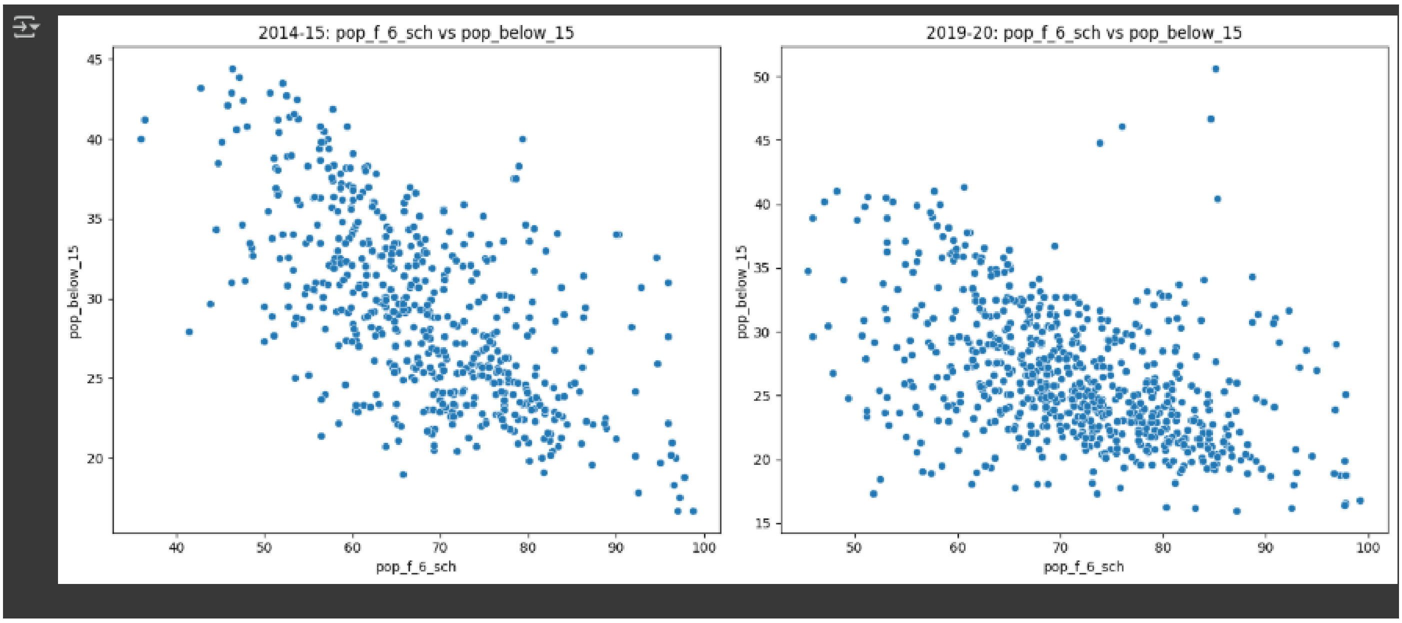
This scatter plot explores the relationship between alcohol consumption among men (alcohol_men_15) and the percentage of severe blood pressure cases in men (men_bp_sev). The plot shows a moderate upward trend: as alcohol use increases, there is a tendency for severe BP cases to rise. While the data is somewhat scattered, many higher alcohol consumption values cluster with higher BP percentages, supporting a potential positive correlation. This suggests that excessive alcohol intake could be contributing to increased hypertension risk among men across states.

5.1 Correlation Delta between 14-15 & 19-20

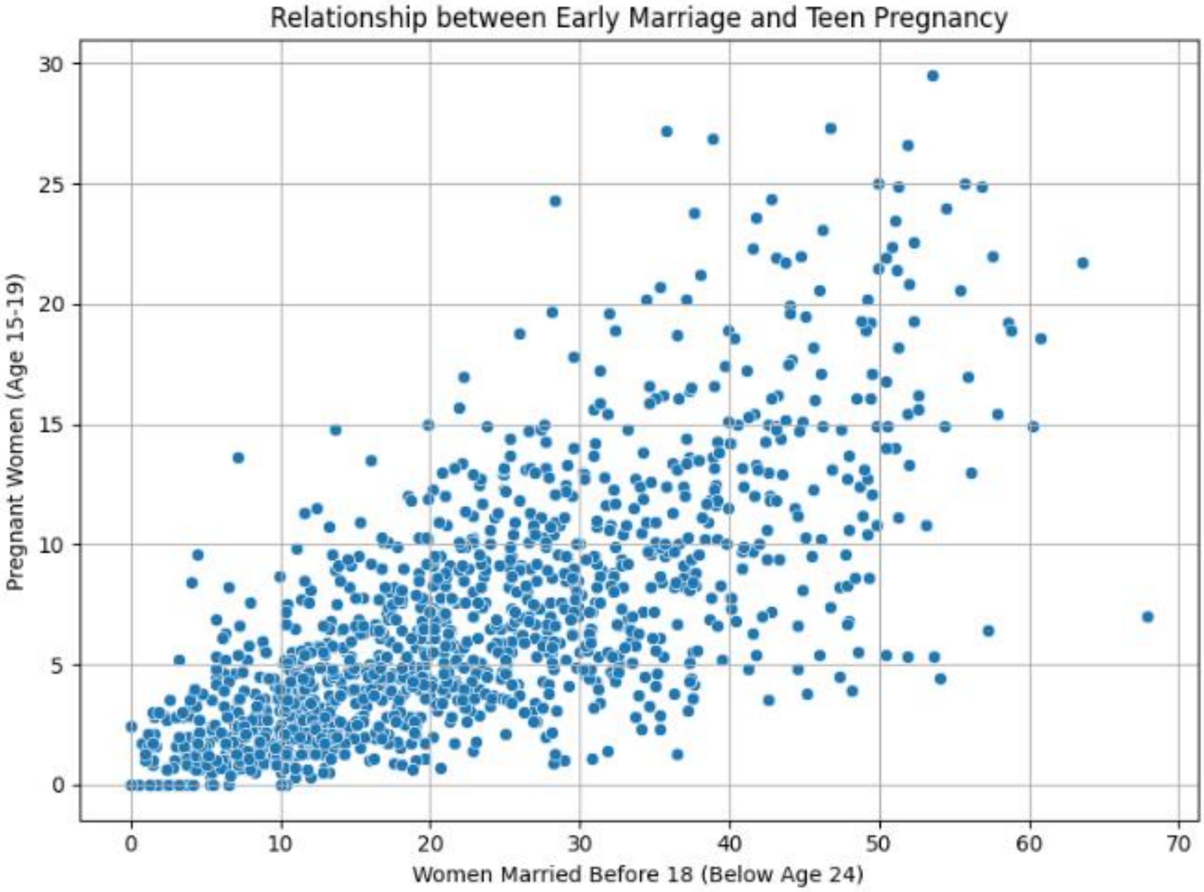


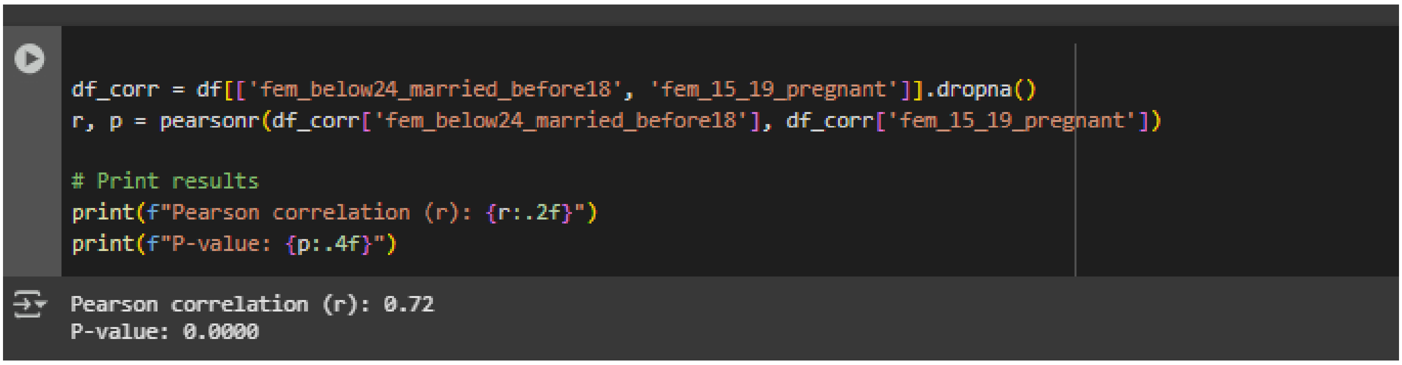
```
# Show top pairs with biggest changes
merged['Δ Correlation'] = merged['Correlation_1920'] - merged['Correlation_1415']
merged['Change Magnitude'] = merged['Δ Correlation'].abs()

merged.sort_values('Change Magnitude', ascending=False).head(10)
```

	Feature 1	Feature 2	Correlation_1415	Correlation_1920	Δ Correlation	Sign Flip	Change Magnitude
141	pop_f_6_sch	pop_below_15	-0.630273	-0.457331	0.172942	False	0.172942
153	pop_below_15	pop_f_6_sch	-0.630273	-0.457331	0.172942	False	0.172942
142	pop_f_6_sch	sex_ratio_tot_pop	-0.042653	-0.091931	-0.049278	False	0.049278
166	sex_ratio_tot_pop	pop_f_6_sch	-0.042653	-0.091931	-0.049278	False	0.049278
155	pop_below_15	sex_ratio_tot_pop	0.145420	0.110426	-0.034994	False	0.034994
167	sex_ratio_tot_pop	pop_below_15	0.145420	0.110426	-0.034994	False	0.034994



In 2014–15, there was a strong negative correlation between female schooling (pop_f_6_sch) and the percentage of population below age 15, meaning that as more girls attended school, birth rates were lower—indicating effective awareness and family planning. However, by 2019–20, while the overall trend remained negative, the correlation had weakened. The data points became more scattered, suggesting that female education alone no longer explained birth rate patterns as strongly, likely due to the influence of additional social, economic, or healthcare factors over time.





```
df_corr = df[['fem_below24_married_before18', 'fem_15_19_pregnant']].dropna()
r, p = pearsonr(df_corr['fem_below24_married_before18'], df_corr['fem_15_19_pregnant'])

# Print results
print(f"Pearson correlation (r): {r:.2f}")
print(f"P-value: {p:.4f}")
```

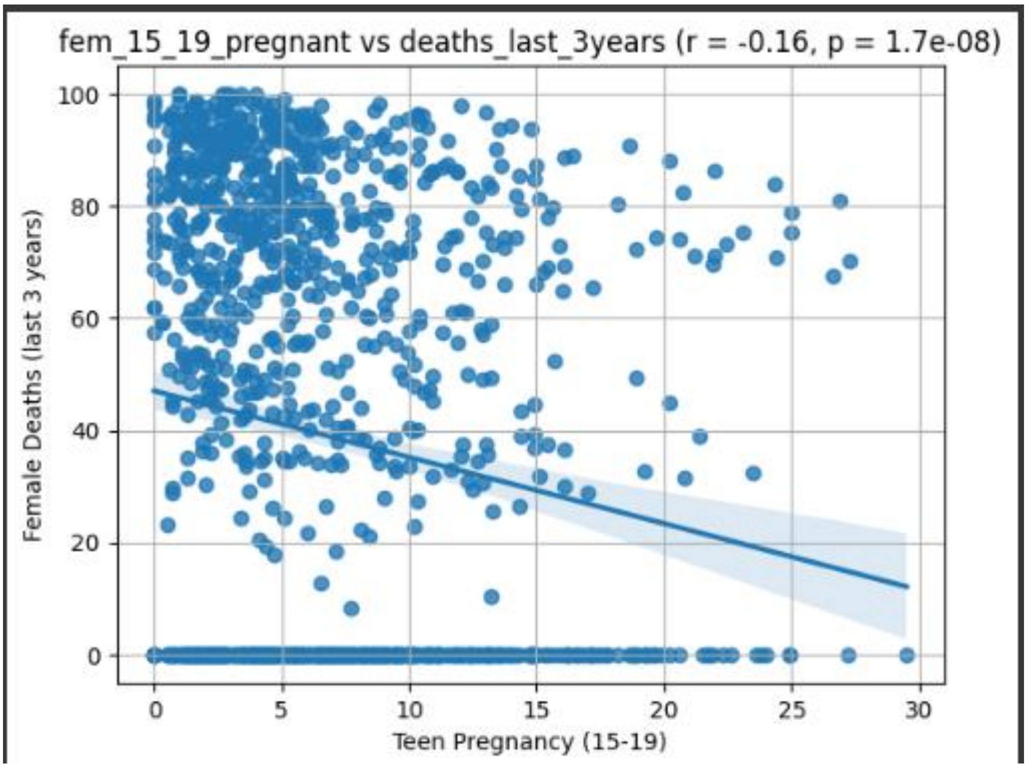
```
Pearson correlation (r): 0.72
P-value: 0.0000
```

- Strong Positive Correlation (r ≈ 0.72):
  There's a clear upward trend—areas with higher early marriage rates also tend to have higher teen pregnancy rates.

- High Density at Lower End:
  Many data points are clustered where both early marriage and teen pregnancy percentages are low, indicating improvements or regional disparities.

- Policy Implication:
  This relationship signals a direct consequence of child marriage leading to early pregnancies, which is a major health and social concern.

- Outliers Exist:
  Some regions with moderate early marriage still have high teen pregnancy, suggesting possible influence of lack of education, contraception access, or social norms.



- Slight Negative Correlation:
  The correlation coefficient r = -0.16 shows a weak negative relationship, meaning that as teen pregnancy increases, female deaths slightly decrease. However, this is not strong enough to suggest a clear or causal link.

- Statistical Significance:
  The p-value p = 1.7e-08 is very low, indicating that the observed correlation is statistically significant—i.e., it's unlikely to be due to chance, even if it's weak.

- Outliers at Zero:
  A large number of points are clustered at y = 0 (no deaths recorded), which may reflect missing or misreported data, especially for less populous regions.

- Data Spread:
  Teen pregnancy rates range from 0 to 30%, and female deaths span a wide range (from 0 to 100), suggesting high variability among districts or states.

  Real-World Interpretation:
  The weak negative trend might imply that districts with higher awareness and reporting of teen pregnancies also have better maternal care, resulting in fewer female deaths—but this needs deeper multivariate analysis to validate.

Conclusion:

This exploratory analysis of the National Family Health Survey (NFHS) data provides meaningful insights into the interlinked dynamics of female health, education, substance use, and social indicators like early marriage and teen pregnancy. One of the most significant findings was a strong positive correlation between early marriage (females married before 18) and teenage pregnancies, with an rr-value of 0.72 and a pp-value close to 0. This statistically robust result reinforces the persistent social pattern in which early marriage increases the likelihood of early pregnancy, particularly in rural or low-literacy regions.

In addition to reproductive health, the analysis examined behavioral risk factors. States with higher alcohol consumption among men were observed to have elevated cases of severe blood pressure, supporting established public health findings about lifestyle-induced non-communicable diseases. Conversely, a weak negative relationship between teenage pregnancy and female mortality in the last 3 years was found, suggesting that other factors—such as healthcare access, maternal services, and socio-economic status—play a stronger role in determining female survival.

Furthermore, a slight inverse relationship between female education (measured by the percentage of girls above age 6 attending school) and oral cancer prevalence was seen, implying that higher literacy might contribute to improved health awareness and risk reduction behaviors. However, some indicators (especially from the 2014–15 cycle) showed unreliable values due to missing or zero data entries, highlighting the importance of cautious interpretation. Overall, this analysis underscores the deep interplay between health and education in shaping demographic outcomes across Indian states and suggests that cross-sectoral interventions—spanning education, healthcare, and social reform—are essential for holistic improvement in public health indicators.

Recommendations for Further Analysis

1. Multivariate Modeling:

   Move beyond bivariate scatter plots and apply multiple regression or decision trees to identify the strongest combined predictors of female health outcomes.Analyze reporting gaps and under-reporting patterns to improve data quality and understand the true crime scenario.

2. Temporal and Regional Analysis:

   Incorporate time-based trends (2014–15 vs. 2019–20) or region-specific breakdowns (e.g., Northeast vs. South) to understand how correlations evolve and vary geographically.Evaluate law enforcement effectiveness by comparing crime resolution rates (solved vs reported cases) across states and districts.

3.

References-

Link for the dataset:

https://indiadataportal.com/p/national-family-health-survey/r/mohfw-nfhs-dt-qq-eie