Prevalence Spatial Analysis of Cancer Data -2019

PURPOSE OF THE DATA-

To analyze the prevalence spatial analysis of cancer due to tobacco and alcohol concerning the male and female populations in urban and rural areas of states in India.

DATA DESCRIPTION-

There are 36 states/UTs and 22 columns including states/UTs

Here is a description for each column in your dataset:

1. **States/UTs**: Name of the states and union territories in India.
2. **total cases**: Total number of cancer cases in the respective state or union territory.
3. **Alcohol male**: Ratio of male cases attributable to alcohol consumption.
4. **Alcohol women**: Ratio of women cases attributable to alcohol consumption.
5. **tobacco male**: Ratio of male cases attributable to tobacco consumption.
6. **tobacco women**: Ratio of women cases attributable to tobacco consumption.
7. **Projected Population**: The estimated total population of the respective state or union territory.
8. **Projected Population Male**: Estimated male population of the respective state or union territory.
9. **Projected Population Female**: Estimated female population of the respective state or union territory.
10. **Projected Population Male Urban**: Estimated male urban population of the respective state or union territory.
11. **Projected Population Male Rural**: Estimated male rural population of the respective state or union territory.
12. **Projected Population Female Urban**: Estimated female urban population of the respective state or union territory.
13. **Projected Population Female Rural**: Estimated female rural population of the respective state or union territory.
14. **Urban Tabacco Female**: Ratio of Female cases attributable to tobacco consumption in Urban areas.
15. **Rural Tabacco Female**: Ratio of Female cases attributable to tobacco consumption in Rural areas.
16. **Urban Tabacco Male**: Ratio of male cases attributable to tobacco consumption in Urban areas.
17. **Rural Tabacco Male**: Ratio of male cases attributable to tobacco consumption in Rural areas.

18. **Urban Alcohol Female**: Ratio of Female cases attributable to Alcohol consumption in Urban areas.
19. **Rural Alcohol Female**: Ratio of Female cases attributable to Alcohol consumption in Rural areas.
20. **Urban Alcohol Male**: Ratio of male cases attributable to Alcohol consumption in Urban areas.
21. **Rural Alcohol Male**: Ratio of male cases attributable to Alcohol consumption in Rural areas.
22. **Prevalence**: Prevalence rate of cancer cases in the respective state or union territory.

SOURCE OF THE DATA-

We obtained a dataset for the total number of cancer cases in each of the 36 states/UTs from the Indian Government site

https://www.indiastat.com/table/health/state-wise-estimated-number-cancer-cases-india-200/627134

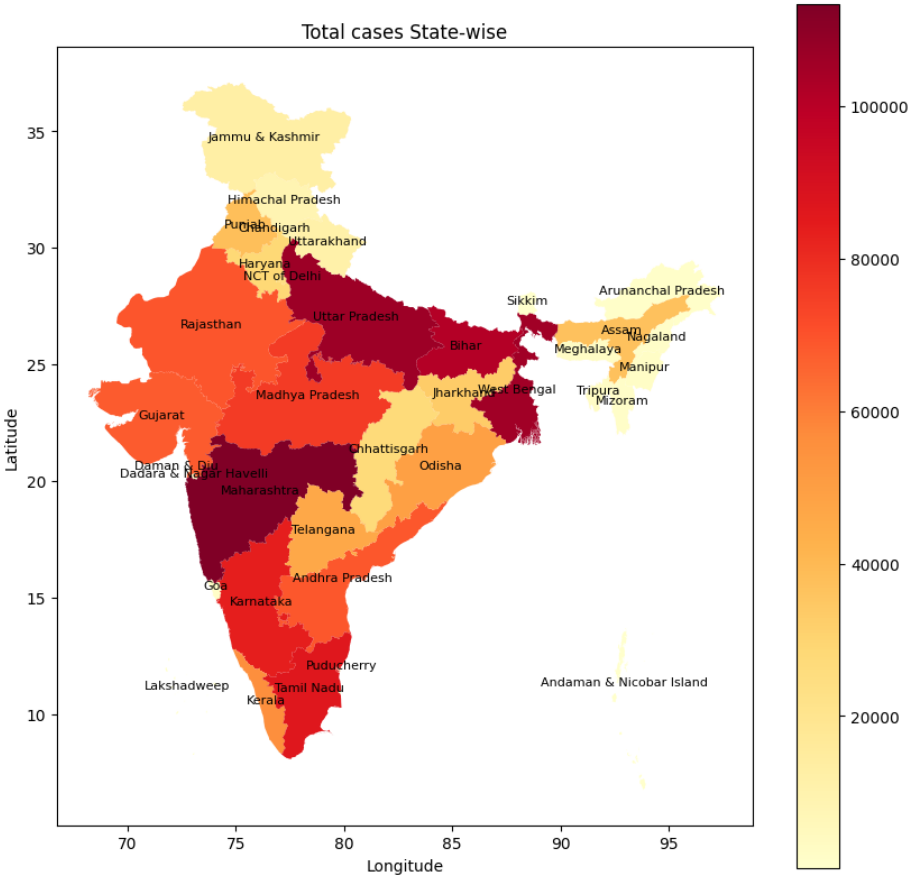
Then, we extracted data for the following variables from the government site https://rchiips.org/nfhs/factsheet_NFHS-5.shtml

for each of the 36 states/UTs: 'Alcohol male', 'Alcohol women', 'tobacco male', 'tobacco women', 'Projected Population', 'Projected Population male', 'Projected Population female', 'Projected Population male Urban', 'Projected Population male Rural', 'Projected Population Female Urban', 'Projected Population female Rural', 'Urban Tabacco Female', 'Rural Tabacco Female', 'Urban Tabacco Male', 'Rural Tabacco Male', 'Urban Alcohol Female', 'Rural Alcohol Female', 'Urban Alcohol Male', 'Rural Alcohol Male'.

Finally, we merged these datasets to create a comprehensive dataset for the 36 states/UTs, now containing 20 columns.

Spatial Analysis:

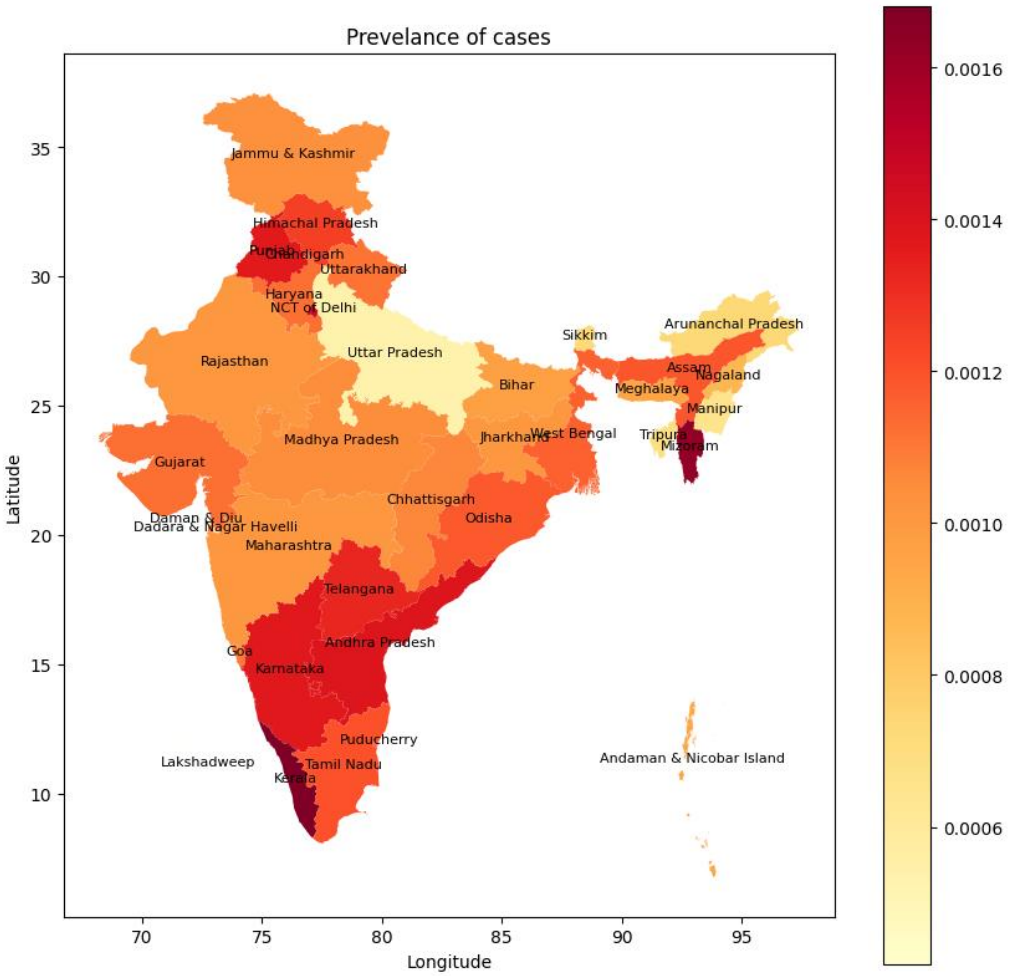
1. Total Cases of Cancer State-wise:



Interpretation:

- Darker regions like Uttar Pradesh, Maharashtra, Bihar, West Bengal, Tamil Nadu, Karnataka, and Madhya Pradesh have higher numbers of cancer cases. However, comparing these regions based solely on the number of cases is not appropriate because the population size varies significantly from state to state. To address this issue, we calculated the prevalence rate, which considers the number of cases relative to the population size, providing a more meaningful comparison between regions.

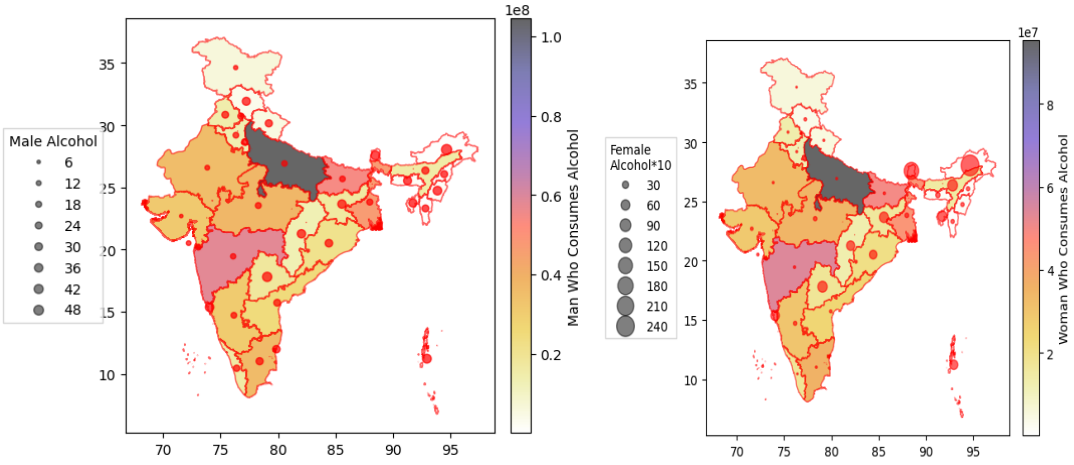
2. Prevalence of cases:



Interpretation:

- High Prevalence States: Kerala, Mizoram, Karnataka, Andhra Pradesh, Telangana, Punjab, and Himachal Pradesh have a high prevalence of cancer cases. This indicates that the number of cancer cases per capita is relatively high in these states compared to other states in India.
- Low Prevalence States: Uttar Pradesh, Manipur, Tripura, and Sikkim have a lower prevalence of cancer cases. This suggests that the number of cancer cases per capita is relatively low in these states compared to others.

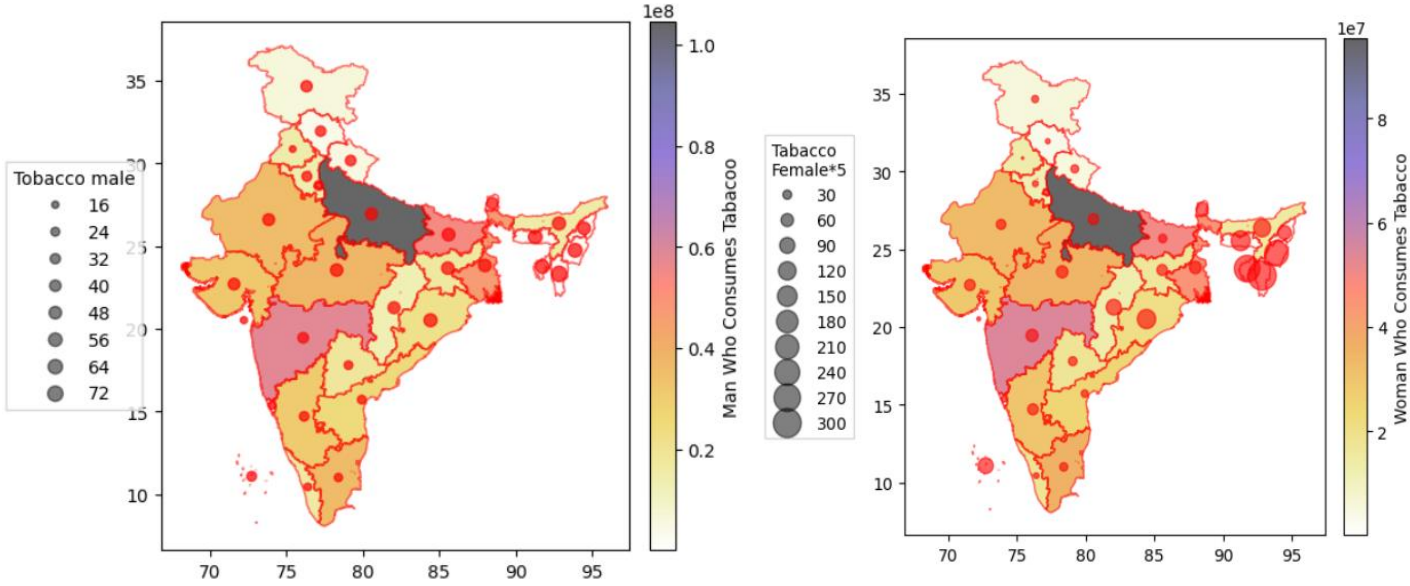
3. Men vs Women who have consumption of Alcohol:



Interpretation:

- Male and Female Population: Uttar Pradesh, Maharashtra, Bihar, and West Bengal have high male and female populations. However, their male alcohol cancer ratio is average compared to other states, suggesting that alcohol consumption may not be a significant factor contributing to cancer cases in these states.
- Male Alcohol Consumption Ratio: The highest ratio of men who consume alcohol is in the Seven Sister states (Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, and Tripura), Telangana, Chhattisgarh, Jharkhand, Sikkim, Himachal Pradesh, Goa, and Uttarakhand. This indicates that these states have a higher proportion of men who consume alcohol compared to other states.
- Sikkim: Sikkim has a high male population and the highest male ratio of alcohol consumption. Additionally, it has the highest female ratio of alcohol consumption among all states, indicating that alcohol consumption is a significant concern in Sikkim for both males and females.
- Women Alcohol Consumption Ratio: Assam, Tripura, Telangana, Chhattisgarh, Jharkhand, Odisha, and Goa have the highest cases of women's ratio of alcohol consumption. This suggests that these states have a higher proportion of women who consume alcohol compared to other states.

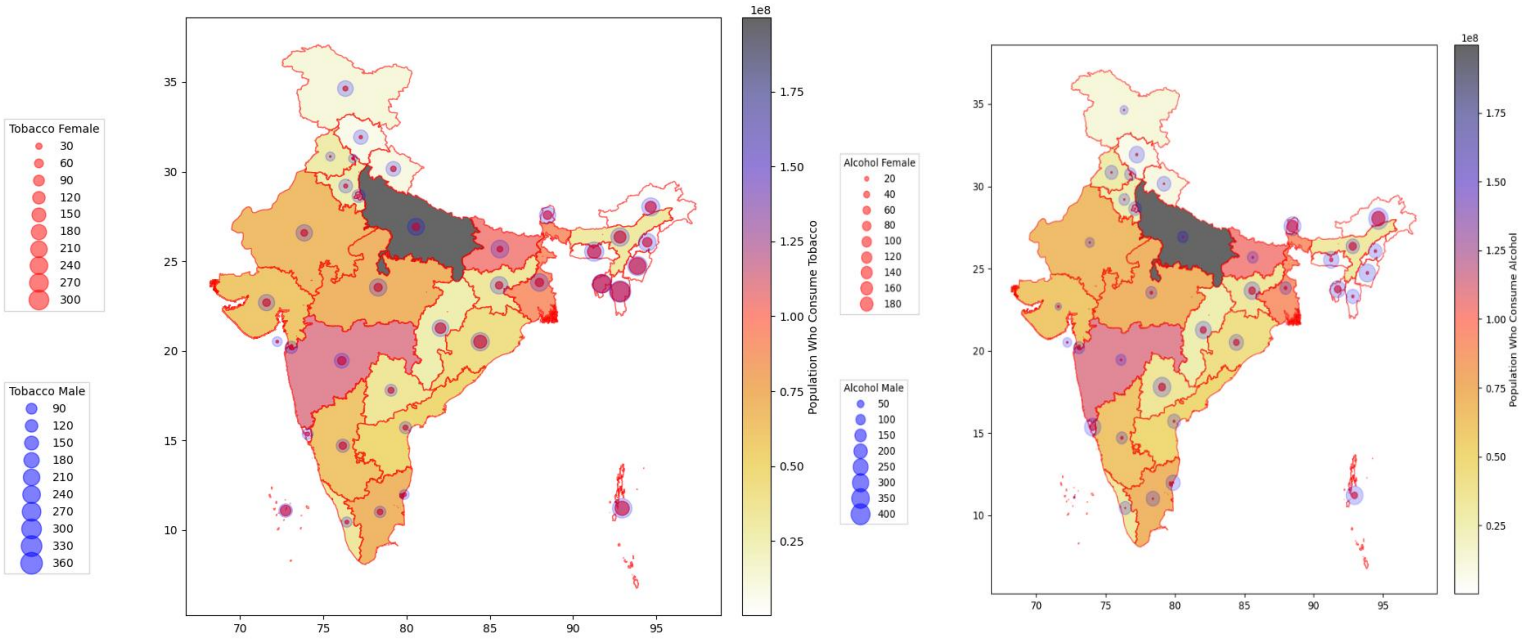
4. **Men vs Women who consume Tabacco:**



Interpretation:

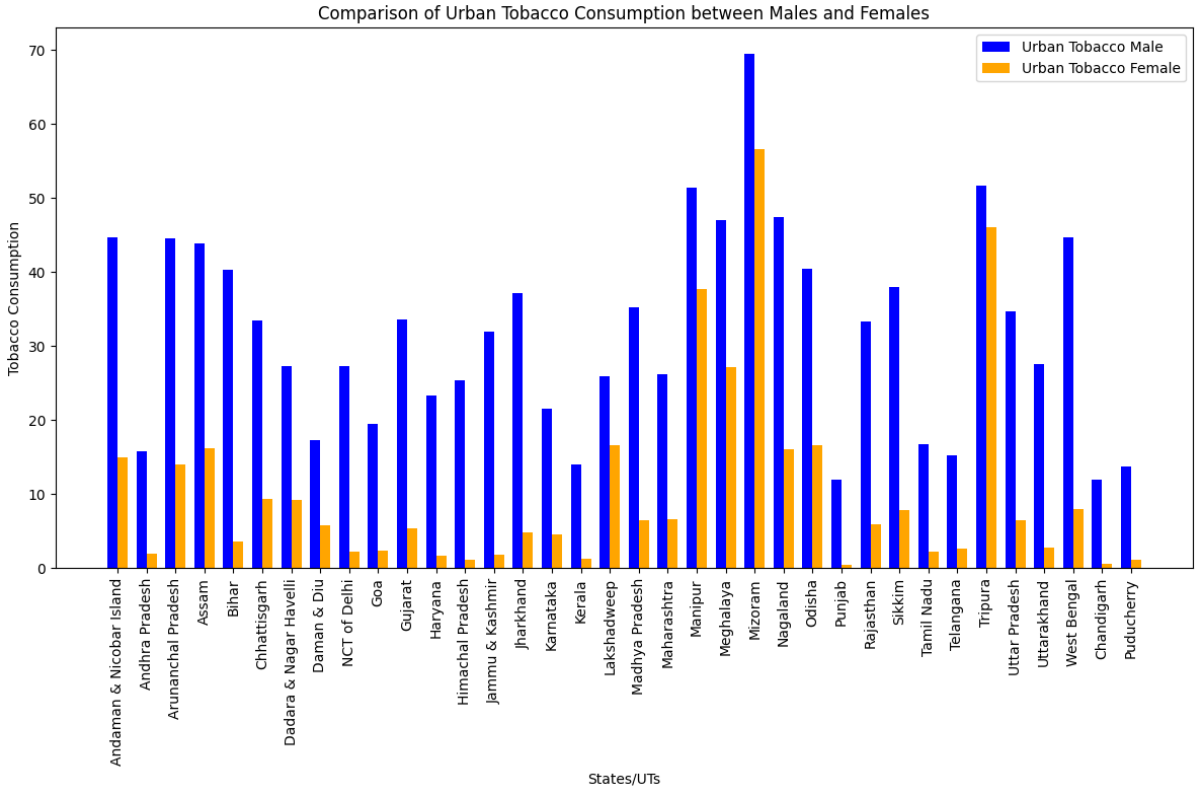
- The darker regions on the map represent areas with higher population density. Notably, Uttar Pradesh, Maharashtra, Bihar, West Bengal, Sikkim, and Lakshadweep show the highest population densities for both men and women.
- In terms of tobacco consumption among men, regions in Middle India, the Seven Sisters states, West India, North India, and Lakshadweep exhibit higher prevalence. Interestingly, areas with higher male populations tend to have more cases of tobacco consumption, while even in regions with lower male populations like the Seven Sisters states, there are still significant instances of tobacco use, indicating a prevalent trend among men in these areas.
- Among women, tobacco consumption is prominent in the Seven Sisters states, Odisha, Chhattisgarh, Lakshadweep, and Maharashtra. Particularly noteworthy is the higher prevalence of tobacco use among women, surpassing that of men, in certain regions like the Seven Sisters states and Maharashtra.
- Comparatively, North India exhibits lower cases of alcohol consumption among women compared to other states.

5. Alcohol and Tobacco Consumption Comparison between male and female



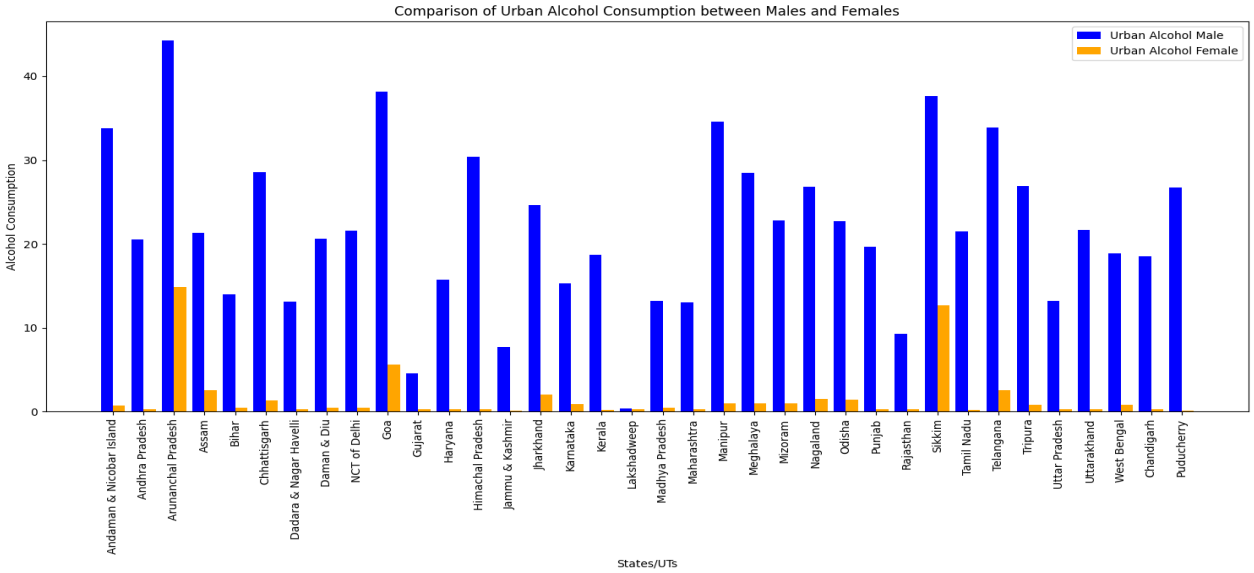
- Female tobacco consumption is higher in Mizoram, Tripura, and Manipur compared to other regions, but similar to tobacco consumption among males in these regions.
- Seven Sisters states, Andaman & Nicobar Islands, and Lakshadweep have higher tobacco consumption among females compared to other regions.
- Male tobacco consumption is higher in Jammu & Kashmir, Himachal Pradesh, Uttar Pradesh, Rajasthan, Madhya Pradesh, Gujarat, Andaman & Nicobar Islands, Lakshadweep, Bihar, Jharkhand, West Bengal, and Sikkim compared to other regions and compared to female tobacco consumption in these regions.
- Female alcohol consumption is higher only in Arunachal Pradesh, Sikkim, Assam, Tripura, Telangana, Chhattisgarh, and Jharkhand compared to other regions.
- In other regions, male alcohol consumption is higher compared to female alcohol consumption.
- In the Seven Sisters states, Uttarakhand, Himachal Pradesh, Odisha, Chhattisgarh, and Andaman & Nicobar Islands, both tobacco and alcohol consumption is high among males.

6. Tobacco Male vs Tobacco Female Consumption in Urban:



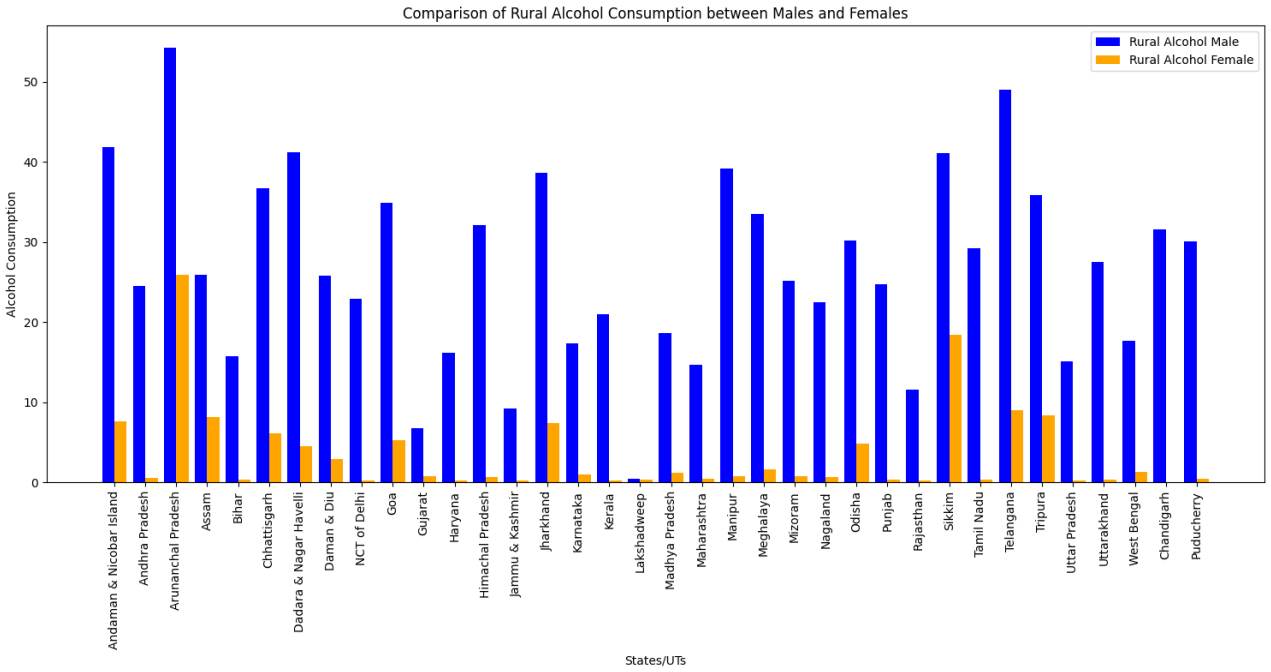
- The plot indicates that males generally consume more tobacco than females across regions.
- Despite having fewer urban areas, states like Mizoram, Tripura, Manipur, and Meghalaya exhibit higher instances of tobacco consumption among females compared to other regions.

7. Urban Alcohol Consumption between Males and Females:



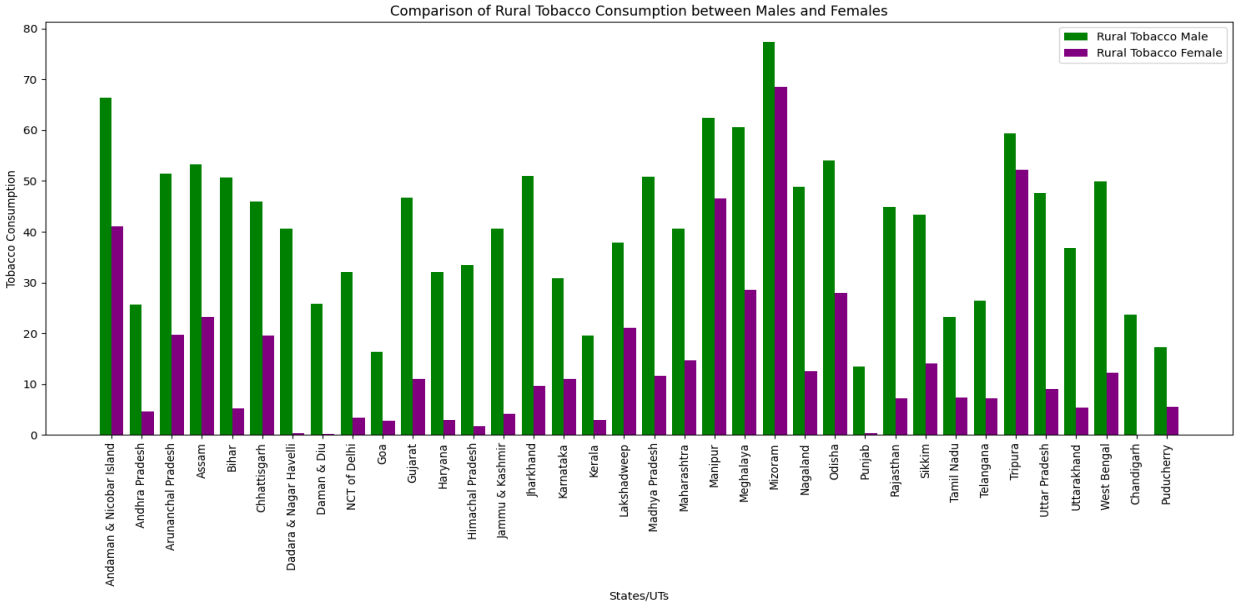
- Urban alcohol consumption is higher among males compared to females, indicating a gender disparity in alcohol consumption patterns in urban areas.
- However, there are certain regions such as Arunachal Pradesh, Sikkim, and Goa where female urban alcohol consumption is relatively higher compared to male consumption, suggesting a more balanced alcohol consumption pattern between genders in these regions.

8. Alcohol Male vs Alcohol Female Consumption in Rural:



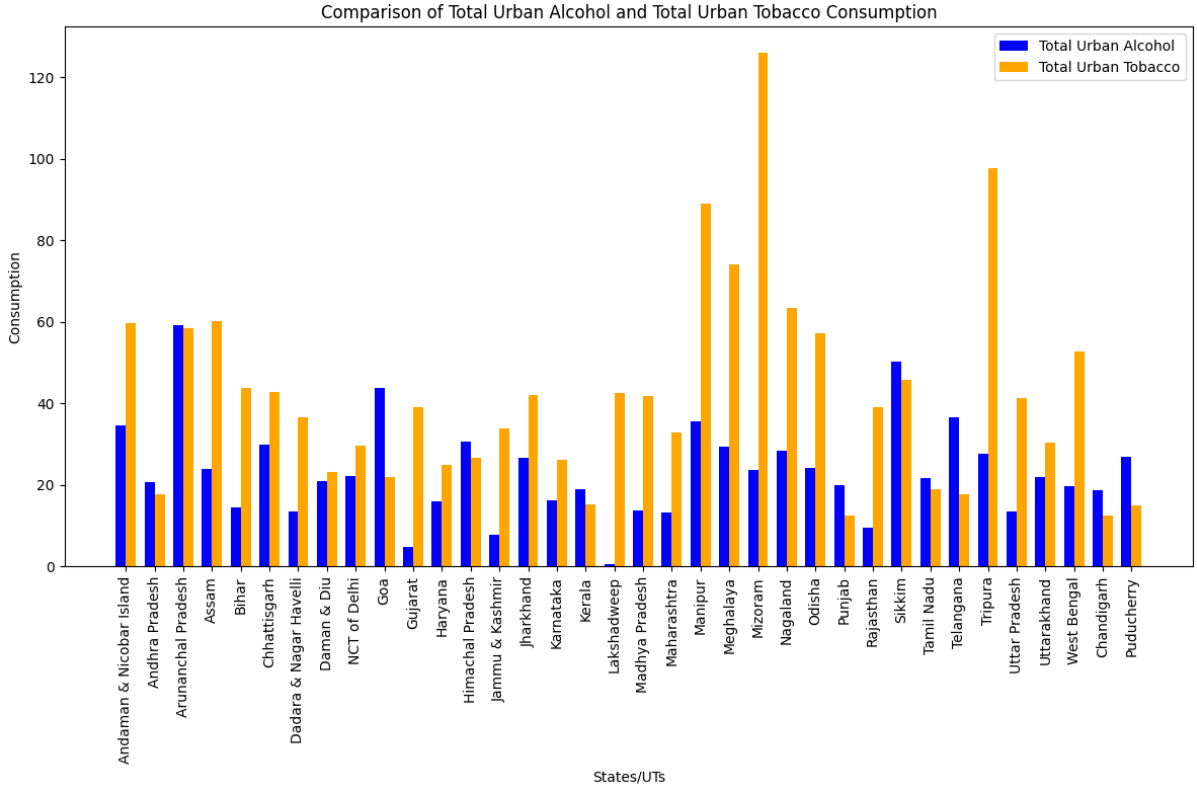
- Similar to urban areas, male alcohol consumption is higher in rural regions as well.
- However, in regions like Sikkim and Arunachal Pradesh, alcohol consumption among females is higher compared to males, indicating a unique trend in these areas.
- Additionally, other regions such as Telangana, Tripura, Jharkhand, Assam, and Andaman & Nicobar Islands show relatively higher alcohol consumption among females compared to other regions, suggesting a more balanced consumption pattern between genders in these areas as well.

9. Tobacco Male vs Tobacco Female Consumption in Rural:



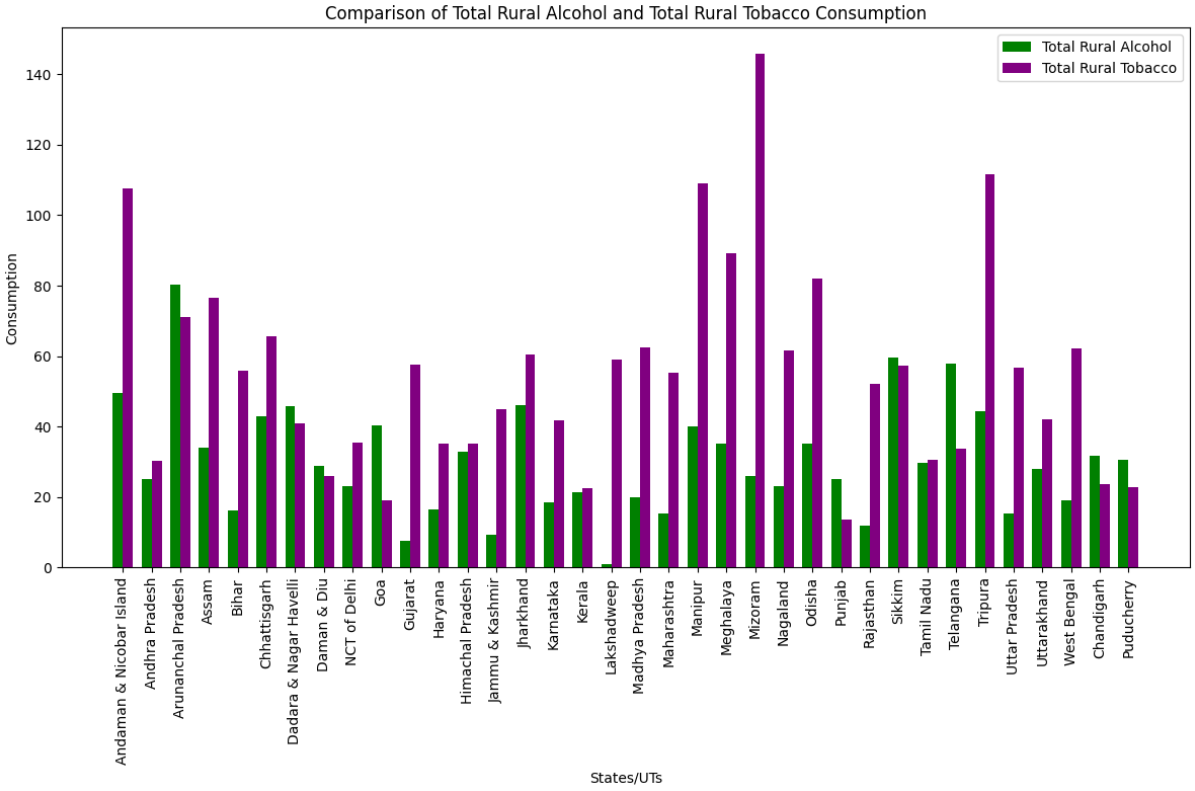
- In Mizoram, Manipur, Tripura, Andaman & Nicobar Islands, and Odisha, both male and female tobacco consumption is higher compared to other regions.

10. Comparison Between Urban Alcohol and Urban Tobacco Consumption:



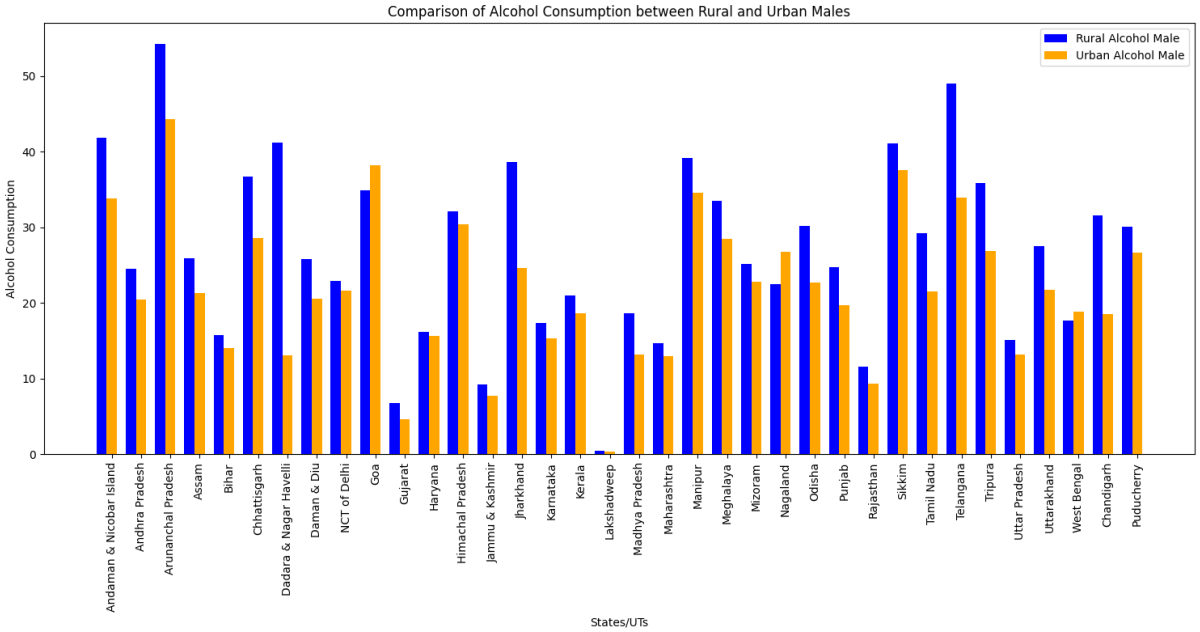
- Urban tobacco consumption is higher than urban alcohol consumption in Mizoram, Tripura, Nagaland, Odisha, Meghalaya, Manipur, Lakshadweep, Madhya Pradesh, Gujarat, Rajasthan, West Bengal, and Maharashtra.
- In other urban regions, both tobacco and alcohol consumption are approximately equal.

11. Comparison Between Rural Alcohol and Rural Tobacco Consumption:



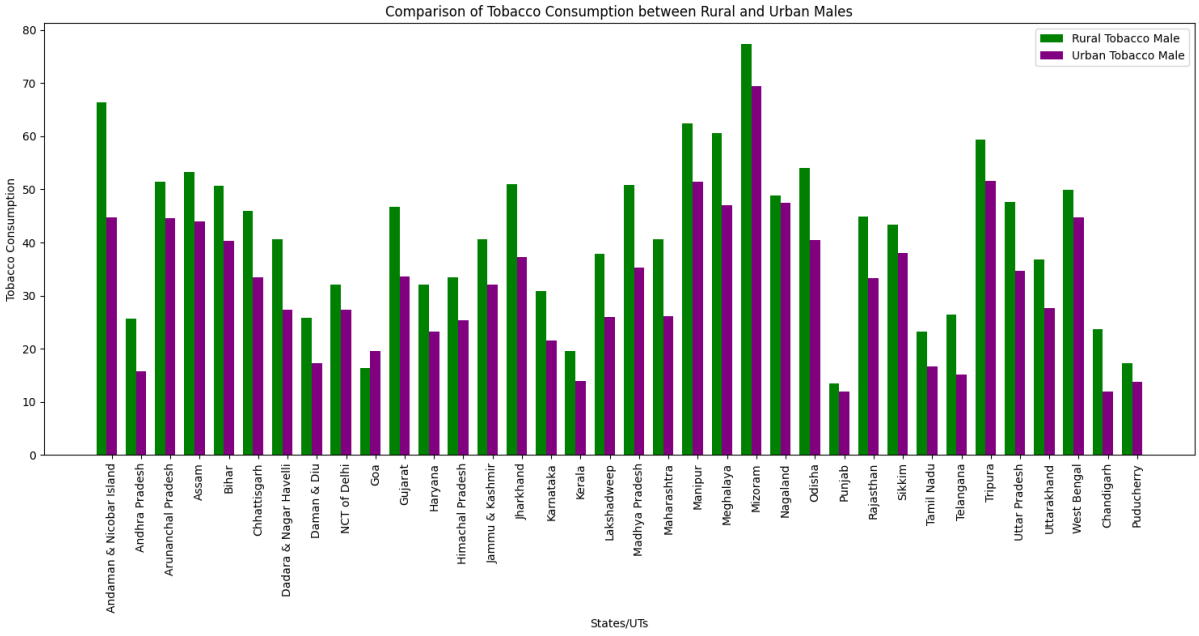
- Rural tobacco consumption is higher than rural alcohol consumption in Mizoram, Meghalaya, Manipur, Tripura, Andaman & Nicobar Islands, Bihar, Gujarat, Jammu & Kashmir, Lakshadweep, Madhya Pradesh, Maharashtra, Nagaland, Odisha, Rajasthan, Uttar Pradesh, and West Bengal.
- Overall, rural areas in East and West India consume more tobacco compared to alcohol.

12. Rural vs Urban males of Alcohol consumption:



- The consumption of alcohol by males is approximately similar in both rural and urban areas across all regions.

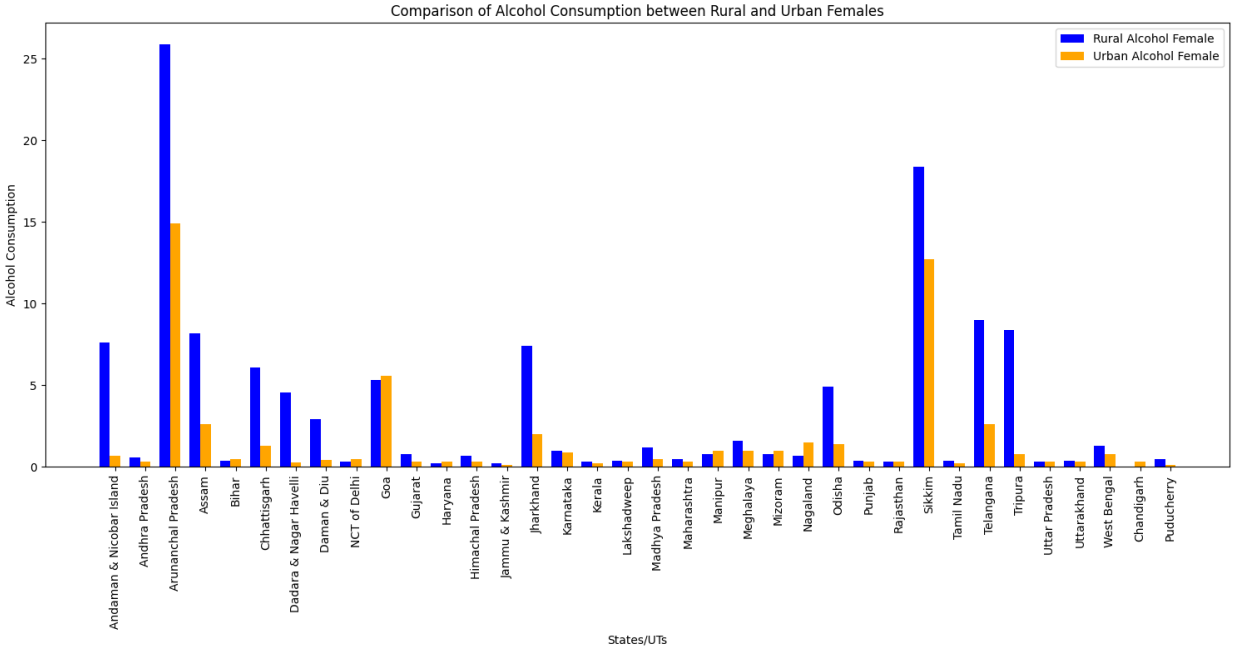
13. Rural vs Urban males of Tobacco consumption:



Interpretation:

- The consumption of tobacco by males is approximately similar in both rural and urban areas across all regions.

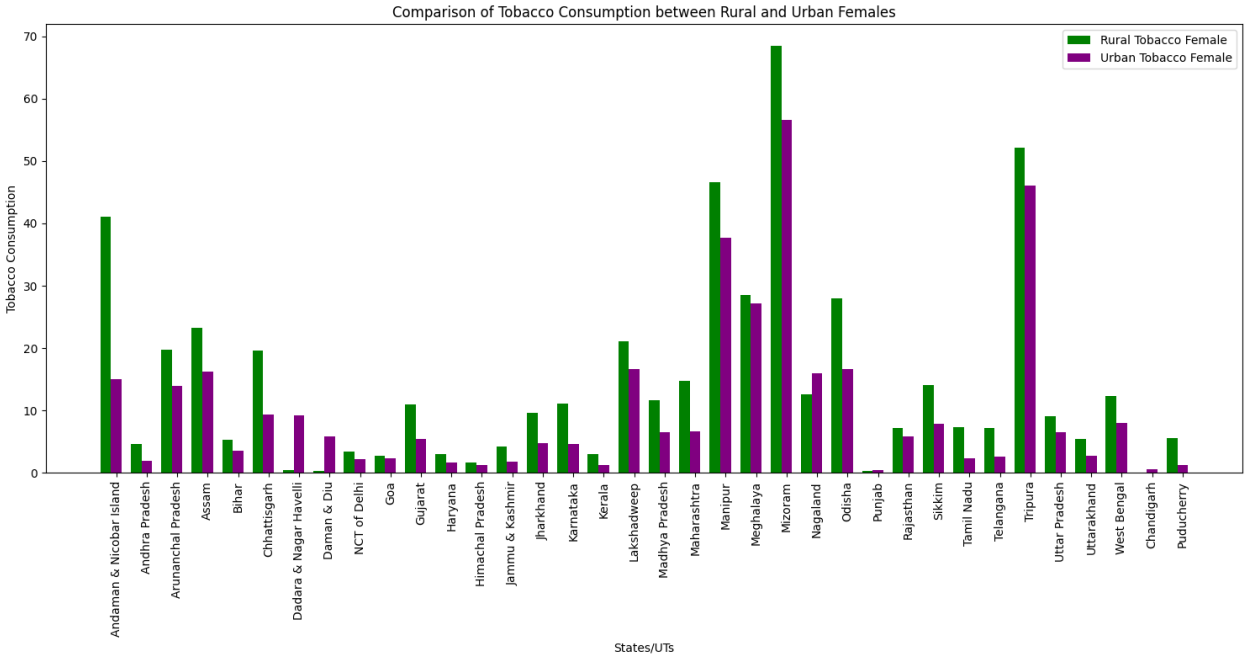
14. Rural vs Urban Females of Alcohol Consumption:



Interpretation:

- Rural alcohol consumption is higher for females in Arunachal Pradesh, Andaman & Nicobar Islands, Tripura, Telangana, Sikkim, Odisha, Jharkhand, Dadra & Nagar Haveli, Chhattisgarh, and Assam.
- In other regions, alcohol consumption among females in both rural and urban areas is approximately equal.

15. Rural vs Urban Females of Tobacco Consumption:



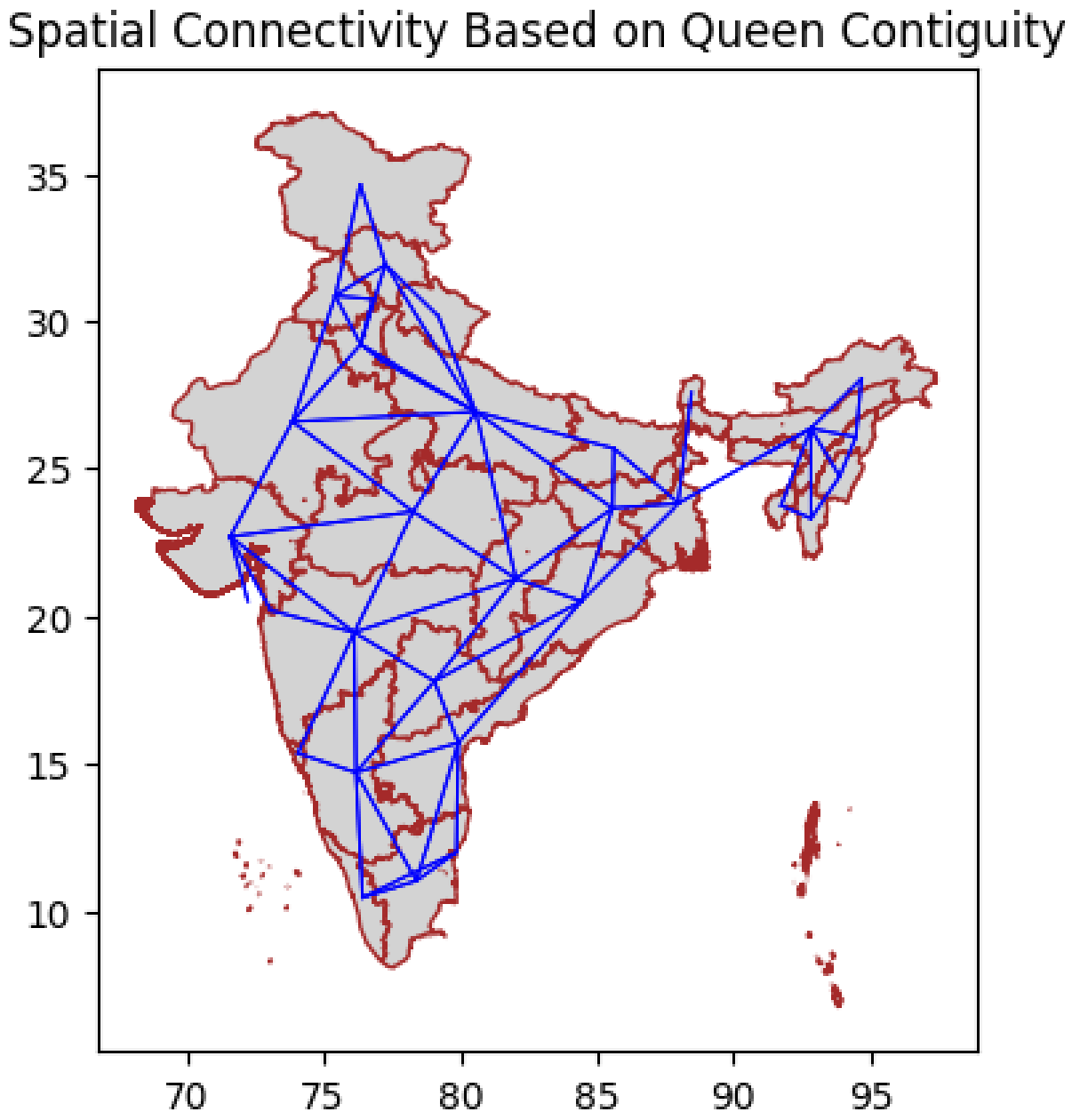
- Urban tobacco consumption is higher for females in Daman & Diu, Dadra & Nagar Haveli, and Chandigarh.
- Rural tobacco consumption is higher for females in Andaman & Nicobar Islands, Chhattisgarh, Karnataka, Maharashtra, Tamil Nadu, and Puducherry.
- In the rest of the regions, tobacco consumption among females in both rural and urban areas is approximately equal.

`

Weight Matrix Rook and Queen:

```
Spatial Weights Matrix(Queen Method):
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]
 [0. 1. 0. ... 0. 1. 0.]]
```

```
Spatial Weights Matrix(Rook Method):
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 1.]
 [0. 1. 0. ... 0. 1. 0.]]
```



INTERPRETATION:

- Regions are connected based on the Queen contiguity method, sharing borders or corners.
- Gray shapes with brown borders represent individual regions.

Summary of Rook and Queen Contiguity Spatial Weights Matrix

```
> summary(rook_nb)
Neighbour list object:
Number of regions: 36
Number of nonzero links: 100
Percentage nonzero weights: 7.716049
Average number of links: 2.777778
4 regions with no links:
1 6 8 18
6 disjoint connected subgraphs
Link number distribution:

0 1 2 3 4 5 6 7
4 8 6 6 4 3 4 1
8 least connected regions:
5 9 22 25 27 28 29 30 with 1 link
1 most connected region:
34 with 7 links
```

```
> summary(queen_nb)
Neighbour list object:
Number of regions: 36
Number of nonzero links: 100
Percentage nonzero weights: 7.716049
Average number of links: 2.777778
4 regions with no links:
1 6 8 18
6 disjoint connected subgraphs
Link number distribution:

0 1 2 3 4 5 6 7
4 8 6 6 4 3 4 1
8 least connected regions:
5 9 22 25 27 28 29 30 with 1 link
1 most connected region:
34 with 7 links
```

Rook and Queen give the same result and the summary is below:

The Rook contiguity spatial weights matrix provides valuable insights into the spatial relationships among the regions considered in the analysis. Here's a breakdown of the key findings:

- **Neighbour List Object:**
  - Number of Regions: 36
  - Number of Nonzero Links: 100
  - Percentage Nonzero Weights: 7.72%
  - Average Number of Links: 2.78

- **Regions with No Links:**
  - There are 4 regions (indexed 1, 6, 8, and 18) that have no connections with any other region.

- **Disjoint Connected Subgraphs:**
  - The spatial connectivity forms 6 disjoint connected subgraphs, indicating separate clusters or groups of regions with no connections between them.

- **Link Number Distribution:**
  - The distribution of link numbers across regions reveals:
    - 4 regions have 0 links
    - 8 regions have 1 link
    - 6 regions have 2 links
    - 6 regions have 3 links
    - 4 regions have 4 links
    - 3 regions have 5 links
    - 4 regions have 6 links
    - 1 region has 7 links

- **Least Connected Regions:**
  - The 8 least connected regions (indexed 5, 9, 22, 25, 27, 28, 29, and 30) each have only 1 link.

- **Most Connected Region:**
  - Region 34 stands out as the most connected region with 7 links, indicating strong spatial connectivity with neighboring regions.

This summary provides a comprehensive overview of the spatial relationships among the regions based on the Rook contiguity spatial weights matrix. It highlights the distribution of connections, identifies isolated regions, and characterizes the connectivity patterns across the geographic area under study.

Weight Constant Summary:

```
Weights style: W
Weights constants summary:
   n   nn S0       S1       S2
W 32 1024 32 25.32135 145.8766
```

```
Weights style: B
Weights constants summary:
   n   nn  S0  S1   S2
B 32 1024 100 200 1672
```

- Two different weight styles were used: binary (B) and row-standardized (W). Here's an explanation of the summary for each style:

  1. Binary Weights (B):
  - n: The total number of regions considered in the analysis is 32.
  - nn: The total number of possible neighbour pairs is 1024.
  - S0: The sum of all spatial weights (nonzero links) is 100.

- S1: The sum of all first-order spatial interactions (the sum of spatial weights for each region) is 200.
- S2: The sum of all second-order spatial interactions (the sum of squared spatial weights for each region) is 1672.

2. Row-Standardized Weights (W):
- n: The total number of regions remains 32.
- nn: The total number of possible neighbor pairs remains 1024.
- S0: The sum of all spatial weights (nonzero links) is 32.
- S1: The sum of all first-order spatial interactions (the sum of spatial weights for each region) is approximately 25.32.
- S2: The sum of all second-order spatial interactions (the sum of squared spatial weights for each region) is approximately 145.88.

Moran I Test:

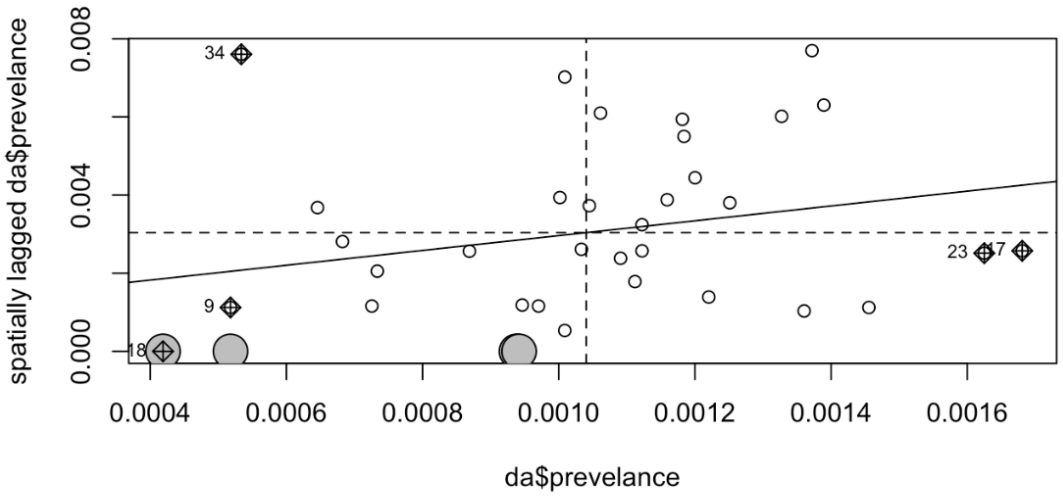
1. Binary Weights

Moran, I test to determine if there is spatial autocorrelation in the prevalence variable. The p-value is 0.2491, suggesting that there is no significant spatial autocorrelation at the specified level of significance.

```
        Moran I test under normality

data:  da$prevelance
weights: W1
n reduced by no-neighbour observations

Moran I statistic standard deviate = 0.67744, p-value = 0.2491
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
       0.05523753       -0.03225806        0.01668141
```



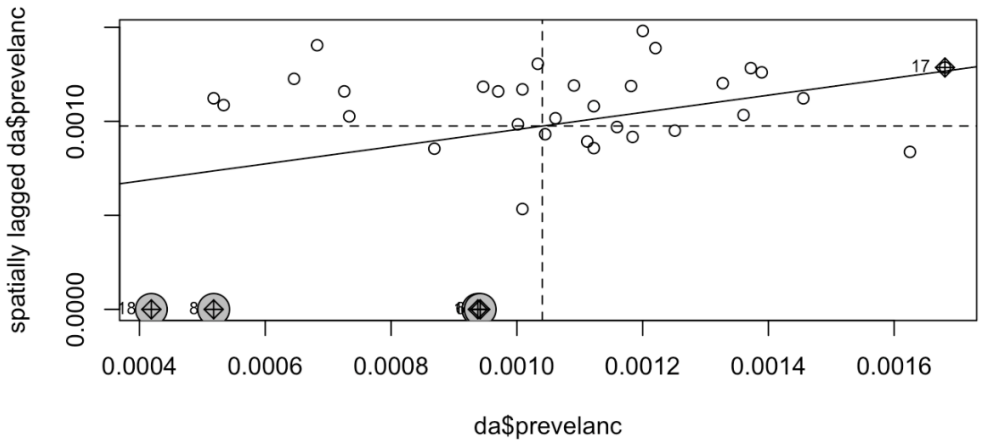
2. **Row-Standardized Weights**

Moran, I test to determine if there is spatial autocorrelation in the prevalence variable. The p-value is 0.2491, suggesting that there is no significant spatial autocorrelation at the specified level of significance.

```
        Moran I test under normality

data:  da$prevelanc
weights: W1_st
n reduced by no-neighbour observations

Moran I statistic standard deviate = 0.39568, p-value = 0.3462
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.02668152       -0.03225806        0.02218787
```



Geary C test:

1. **Binary Weights:**
   The Geary's C test provides insights into the spatial pattern of attribute values in the dataset and whether they exhibit clustering or dispersion. Since the p-value (0.03534) is less than the significance level (typically 0.05), we reject the null hypothesis of spatial randomness. This implies that there is significant spatial autocorrelation in the dataset.

```
        Geary C test under normality

data:  da$prevelanc
weights: W1
n reduced by no-neighbour observations

Geary C statistic standard deviate = 1.8076, p-value = 0.03534
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.67290737        1.00000000       0.03274595
```

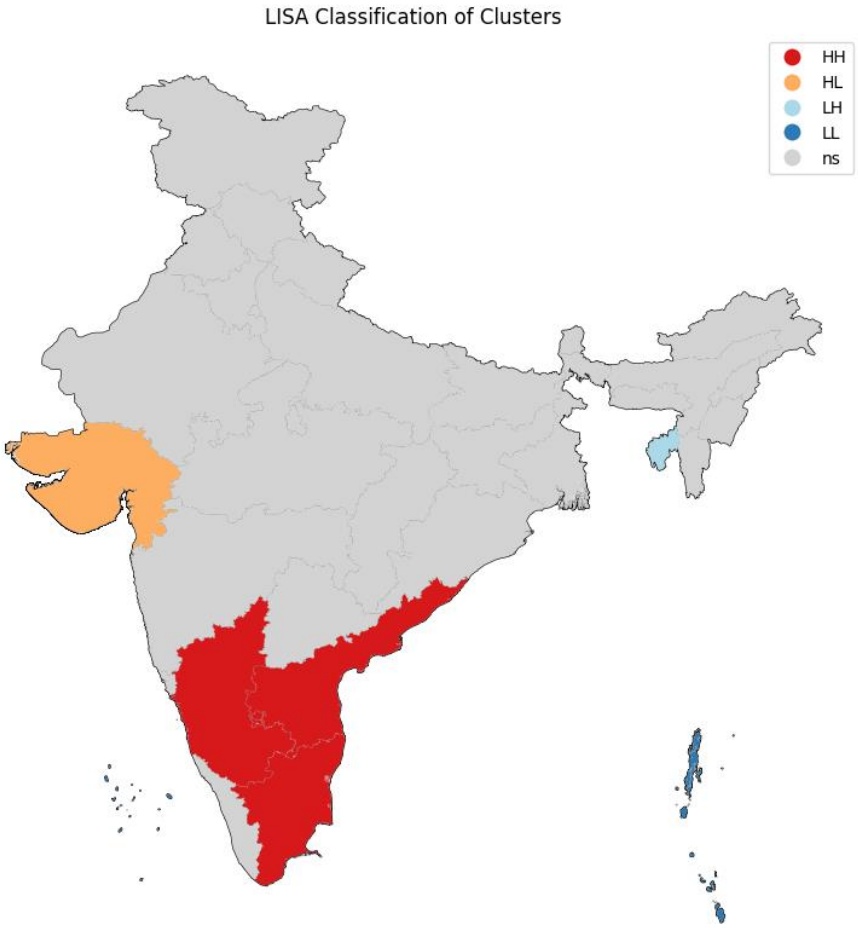
2. **Row-Standardized Weights:**
   The Geary's C test results provide insights into the spatial pattern of attribute values in the dataset, suggesting significant negative spatial autocorrelation based on the computed statistic and p-value. Since the p-value (0.04804) is less than the significance level (typically 0.05), we reject the null hypothesis of spatial randomness. This implies that there is significant spatial autocorrelation in the dataset.

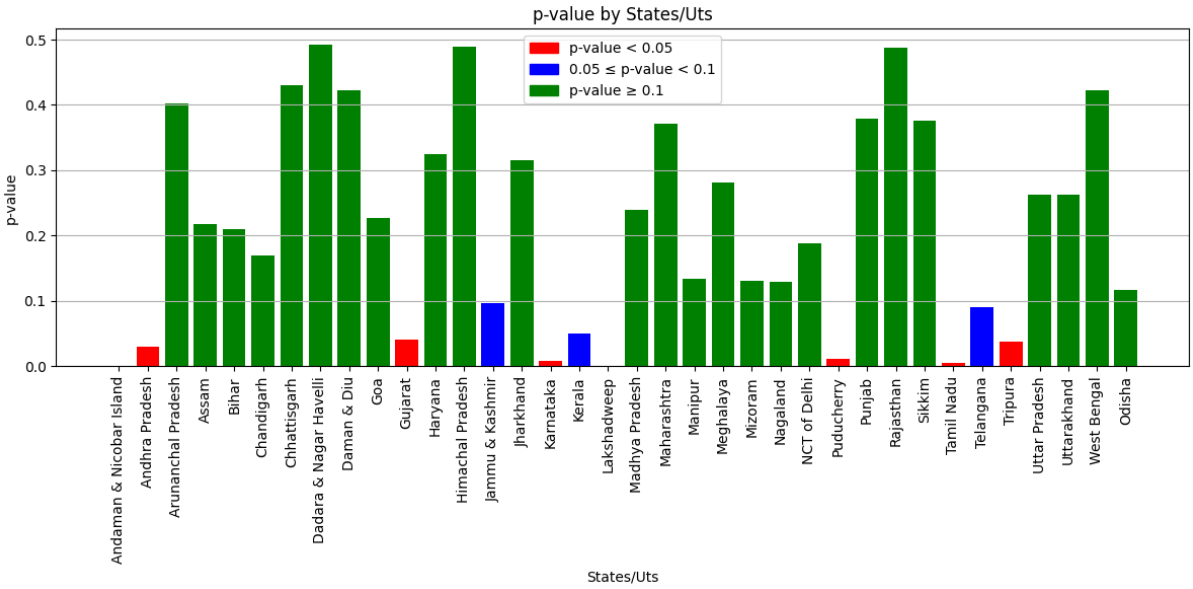
```
        Geary C test under normality

data:  da$prevelanc
weights: W1_st
n reduced by no-neighbour observations

Geary C statistic standard deviate = 1.6642, p-value = 0.04804
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.72989633        1.00000000       0.02634209
```

Local Measures of Spatial Auto-correlation:





Spatial Models: Spatial Lag Model

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :  prevelance                Number of Observations:          36
Mean dependent var  :      0.0010                Number of Variables   :           6
S.D. dependent var  :      0.0003                Degrees of Freedom    :          30
Pseudo R-squared    :      0.2098
Spatial Pseudo R-squared:  0.2383

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT         0.00079         0.00020         4.00091         0.00006
       Alcohol male          0.00001         0.00001         1.89126         0.05859
      Alcohol women         -0.00003         0.00001        -1.91911         0.05497
       tobacco male         -0.00000         0.00000        -0.56351         0.57309
     tombocco women          0.00000         0.00001         0.11128         0.91139
       W_prevelance          0.03852         0.02054         1.87546         0.06073
------------------------------------------------------------------------------------
Instrumented: W_prevelance
Instruments: W_Alcohol male, W_Alcohol women, W_tobacco male, W_tombocco
             women
================================ END OF REPORT =====================================
```

Interpretation:

The spatial two-stage least squares (STSLLS) regression analysis examines the impact of explanatory variables, such as alcohol and tobacco consumption, on the prevalence of a condition, while considering spatial dependence. The model accounts for how prevalence in one area may influence nearby areas. The pseudo R-squared values indicate moderate model fit, with 20.98% of the variance explained and 23.83% when accounting for spatial effects. The coefficient for the spatially lagged prevalence variable suggests a significant positive spatial dependence, implying that areas with higher prevalence are surrounded by others with similarly high rates.

Spatial Models: Spatial Error Model

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: ML SPATIAL ERROR (METHOD = full)
---------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :  prevelance                Number of Observations:          36
Mean dependent var  :      0.0010                Number of Variables   :           5
S.D. dependent var  :      0.0003                Degrees of Freedom    :          31
Pseudo R-squared    :      0.1327
Log likelihood      :    244.8479
Sigma-square ML     :      0.0000                Akaike info criterion :    -479.696
S.E of regression   :      0.0003                Schwarz criterion     :    -471.778

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT         0.00100         0.00016         6.43551         0.00000
        Alcohol male         0.00002         0.00001         3.04467         0.00233
       Alcohol women        -0.00003         0.00001        -2.68761         0.00720
        tobacco male        -0.00001         0.00000        -1.85759         0.06323
      tombocco women         0.00000         0.00000         0.56959         0.56896
              lambda        -0.15409         0.06006        -2.56583         0.01029
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Interpretation:

The spatial error model summary outlines an assessment of the relationship between prevalence and explanatory variables like alcohol and tobacco consumption, while considering spatial dependence in the errors. The pseudo R-squared of 0.1327 suggests a modest association between these factors. The log likelihood value of 244.8479 indicates the model's adequacy in capturing the data's variability. The significant level of the spatial error term (lambda) at 0.01029 underscores the presence of statistically meaningful spatial dependence in the errors.