Exploratory Data Analysis Contest
CS Club, IIITDMK

By Gokul Krishna Balaji, CS24B2053

Introduction

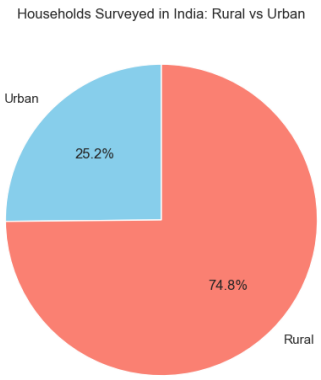
This report explores a state-level survey of social and health indicators across Indian States/UTs (with Rural/Urban split). I clean the data (convert non-numeric entries like * to blank/NaN), scale different measures so they're comparable, and flip indicators where "lower is better" so higher always means better. Then I use easy visual tools — grouped bar charts, violin plots, dot plots, pie and donut charts, box plots, dumbbell plots, and mirror bar charts — to show patterns and compare states. Each figure has a short note on how it was made and what it suggests, followed by a few plain conclusions and things worth investigating further.

Data Cleaning

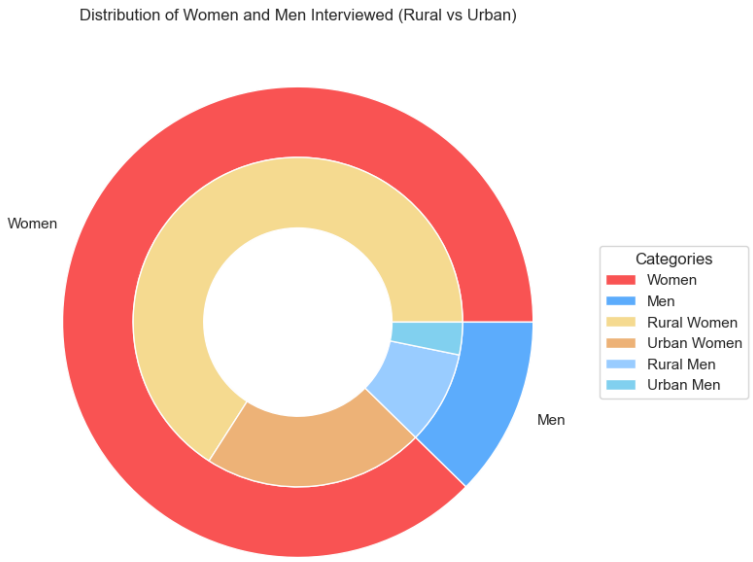
I first checked the dataset for missing values and found none. Some entries had * or were enclosed in parentheses, so I replaced these values using the value which was there in India - Total for that column, effectively imputing the national average. I also found a few negative values in columns representing percentages, which are not possible. These were converted to their absolute values to ensure all percentage-based columns contained only positive numbers.

Plots Explanation

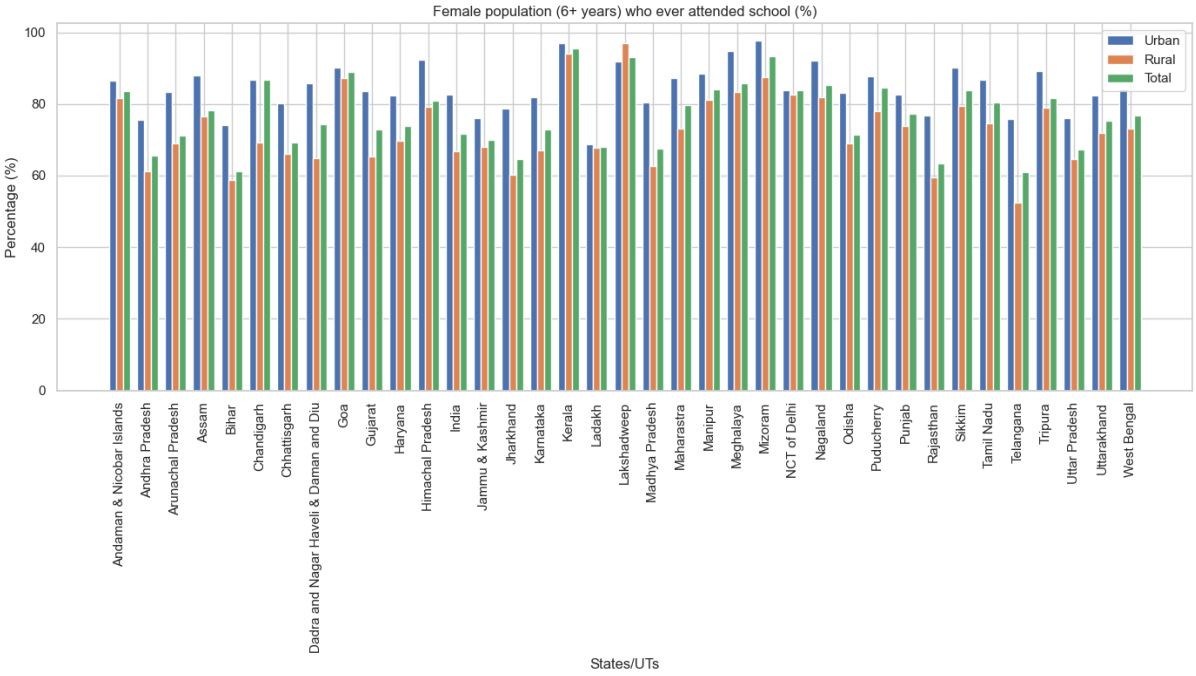
1. This pie chart shows that 74.8% of the households surveyed in India are from Rural areas, while 25.2% are from Urban areas. It clearly visualizes the disproportionate representation of rural households in the survey.



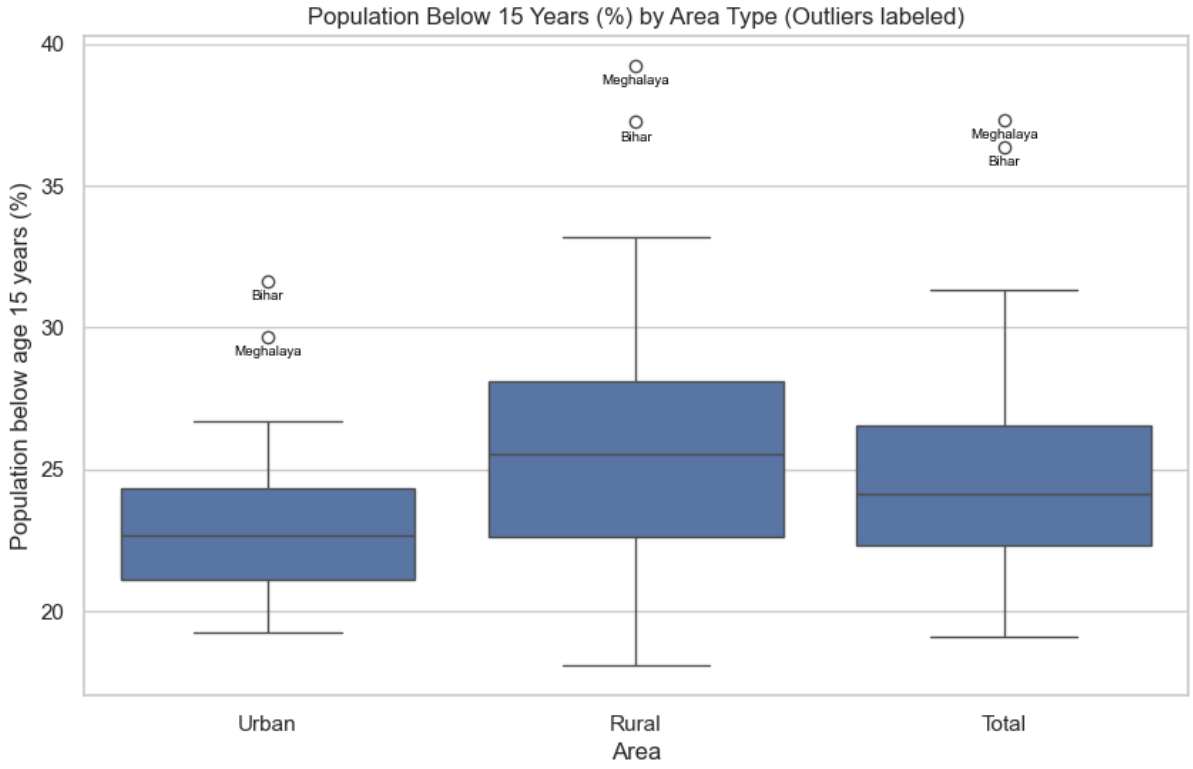
2. This donut chart illustrates the gender distribution of interviewed individuals, showing a larger proportion of women overall. The inner ring further breaks down the distribution by rural and urban areas for both men and women.



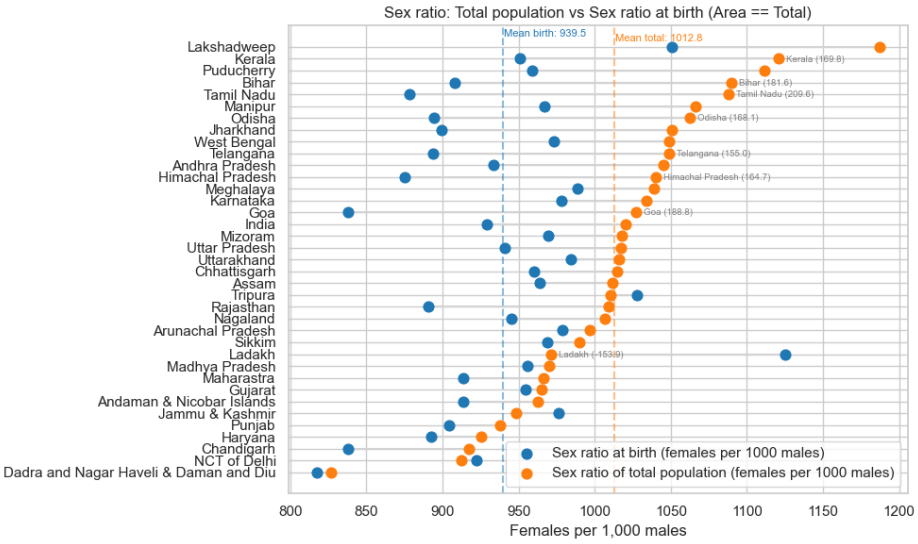
3. This bar chart illustrates the percentage of the female population (6+ years) who have attended school, broken down by Urban, Rural, and Total areas for various States/UTs in India. It allows for a comparison of school attendance rates across different regions and urban/rural divides. Here it shows that Urban areas have more female population over 6 years who attended school, except Lakshadweep where the Rural percentage is more.



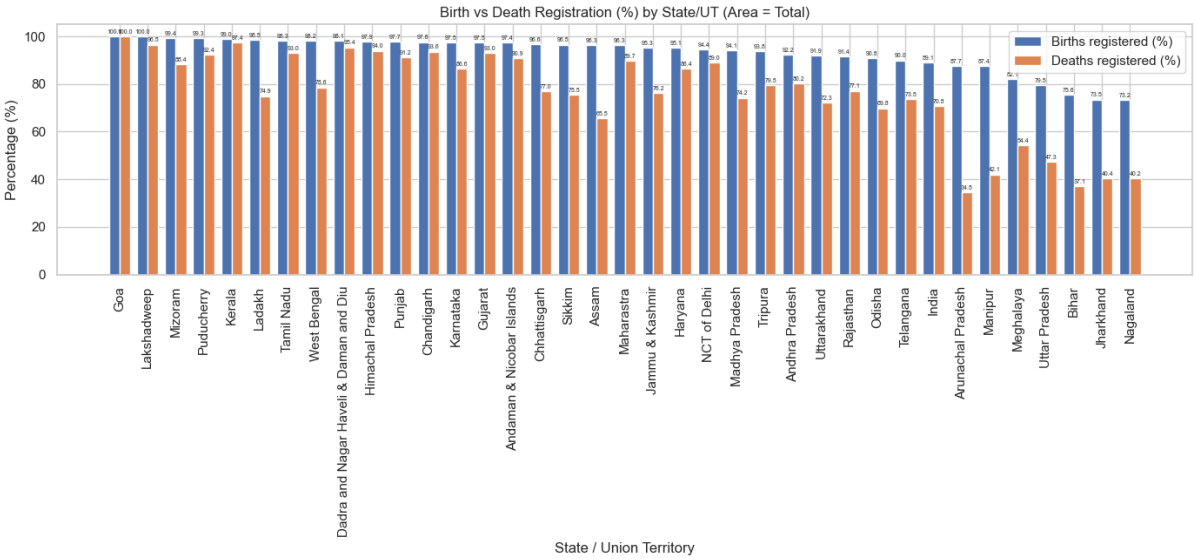
4. This box plot illustrates the distribution of the population below 15 years old, comparing Urban, Rural, and Total areas. It highlights key statistics like median, quartiles, and outliers (labelled as Meghalaya and Bihar) within each area type. This also tells us that rural areas have more kids: due to lesser education about family planning and financial planning.



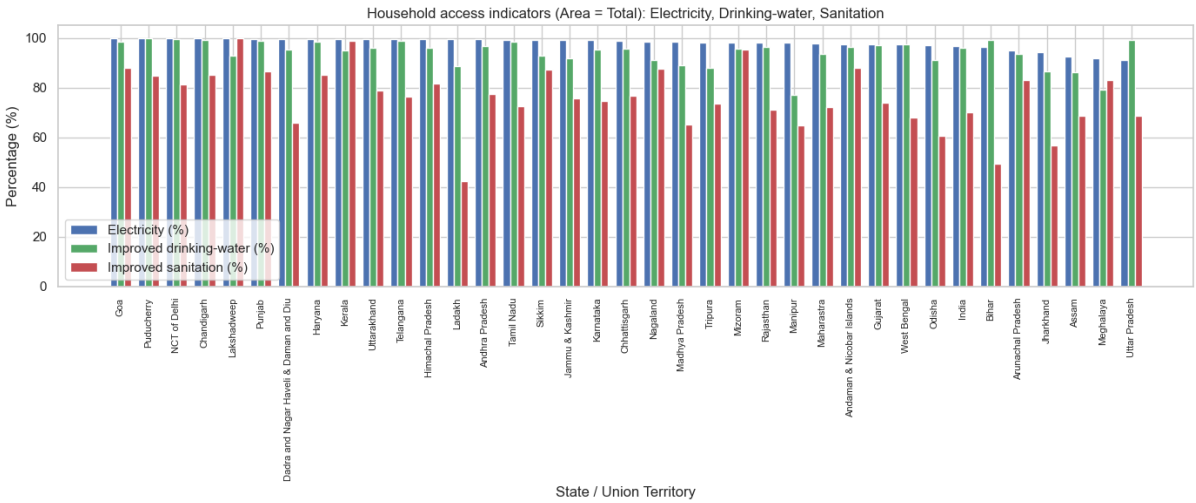
5. This dot plot visualizes the sex ratio at birth and total population sex ratio (females per 1,000 males) for various States/UTs where Area is 'Total'. It highlights mean values for both categories and labels specific outliers. This shows us the game between the sex ratio at birth versus the sex ratio of the total population. The higher the gap, the worse it is as they are not able to preserve the sex ratio which was there at birth.



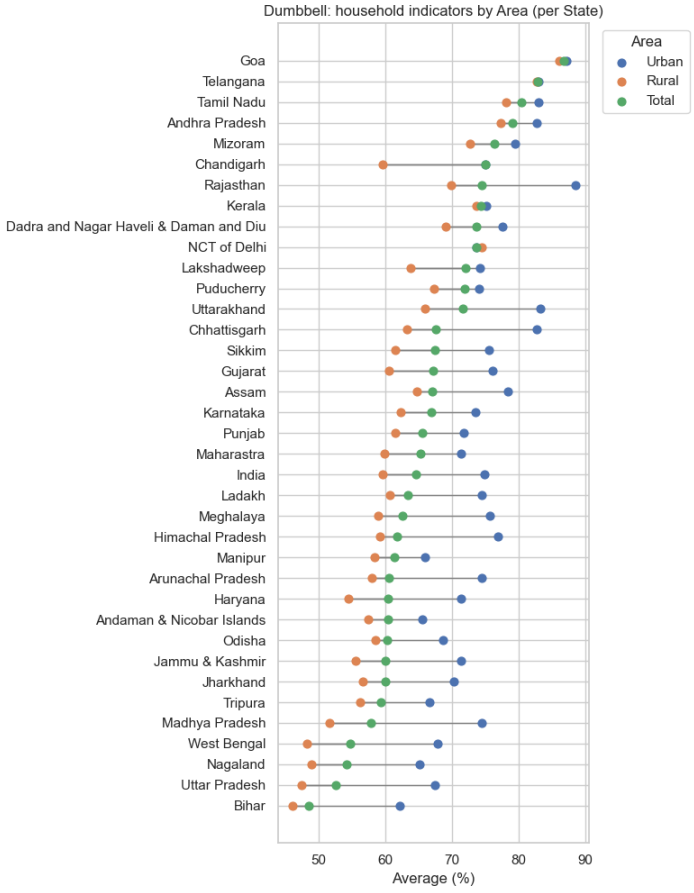
6. This bar chart compares the percentage of births registered versus deaths registered across various States/Union Territories in India, for areas categorized as 'Total'. It highlights the registration rates for vital events in each region.



7. This bar chart displays the household access indicators for various States/Union Territories, specifically for areas categorized as 'Total'. It shows the percentage of households with electricity, access to improved drinking water, and improved sanitation. This shows us which States/UTs people are actually living a decent life with basic needs.

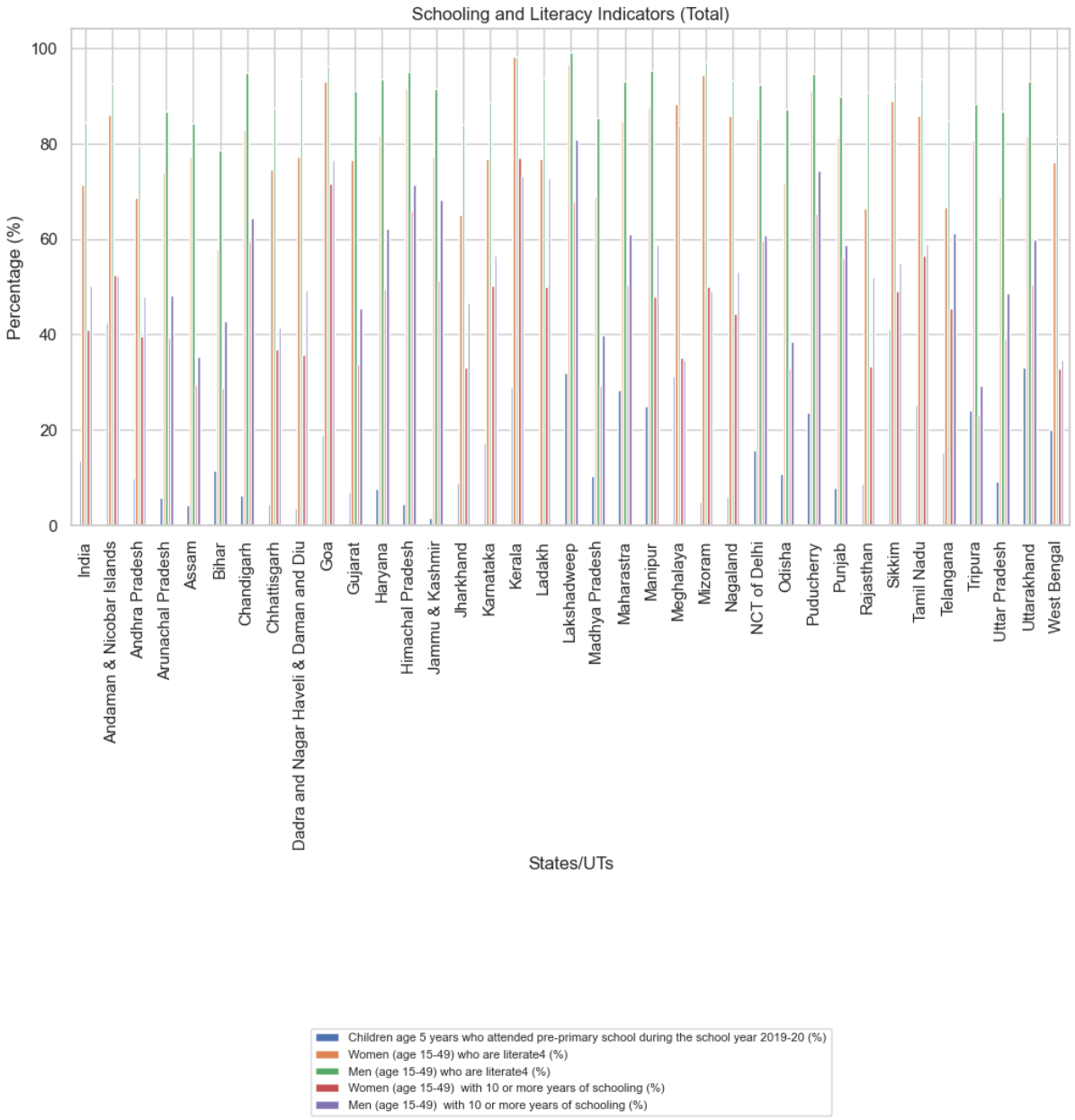


8. This dumbbell plot illustrates the average household indicator percentages (likely electricity, drinking water, and sanitation combined) across various States/UTs. It clearly compares the averages for Urban, Rural, and Total areas within each state. The lesser the gap between urban and rural, the better it is for the place since it means that there is less divide between the people from different parts of the State/UT.

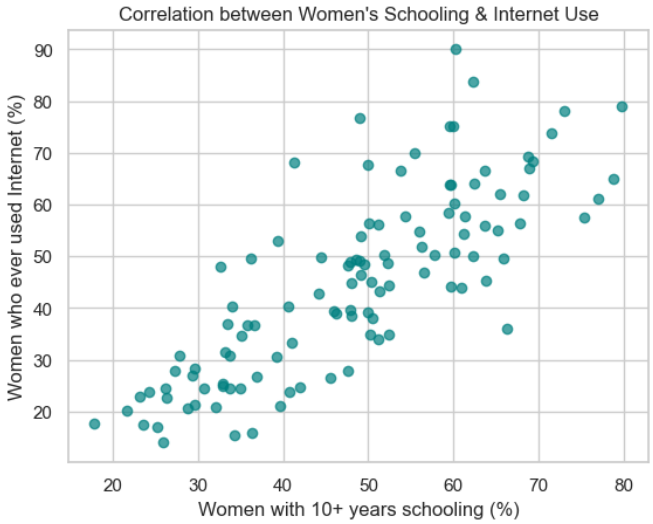


9. This bar chart presents various schooling and literacy indicators for different States/UTs, specifically for areas categorized as 'Total'. It displays percentages for

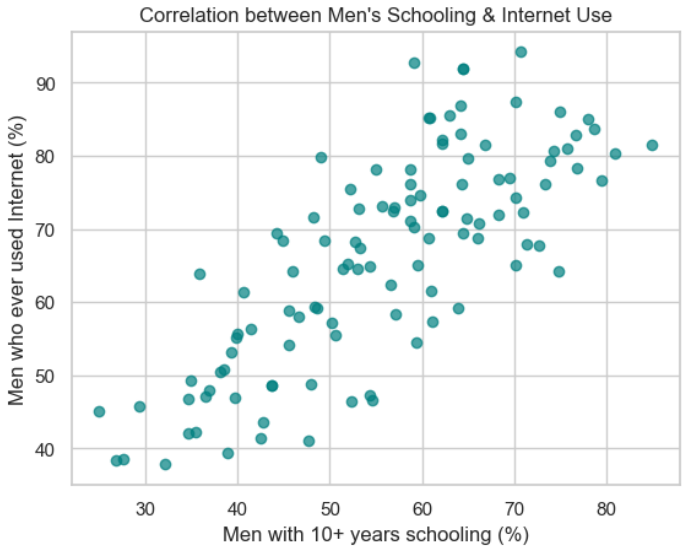
children in pre-primary school, literate women and men, and women and men with 10 or more years of schooling.



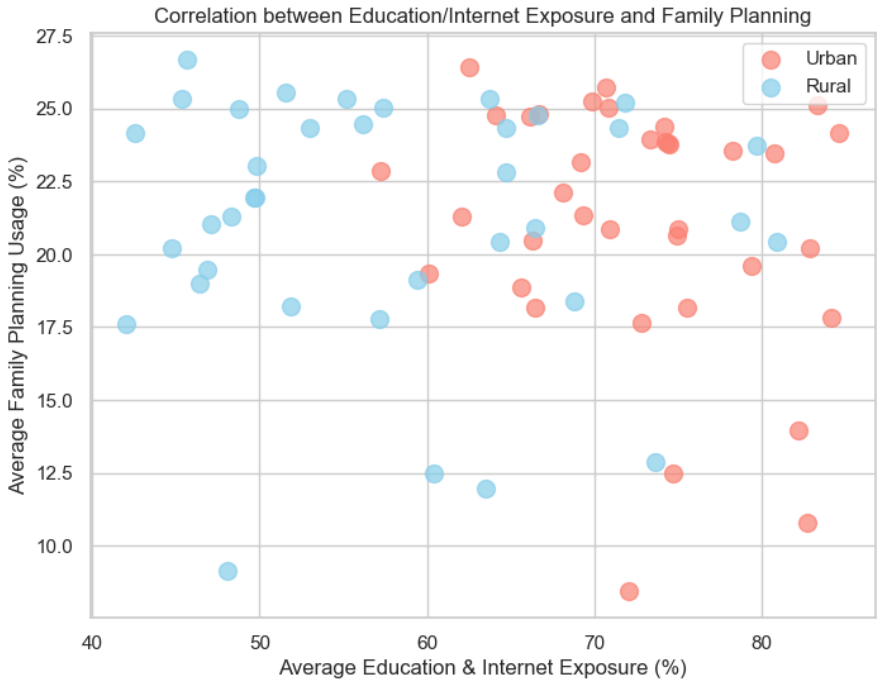
10. This scatter plot visualizes the correlation between women's schooling and their internet use. It shows the percentage of women who have ever used the internet against the percentage of women with 10 or more years of schooling. This shows us a positive correlation which tells us that women with 10 or more years of schooling are more likely to access the internet.



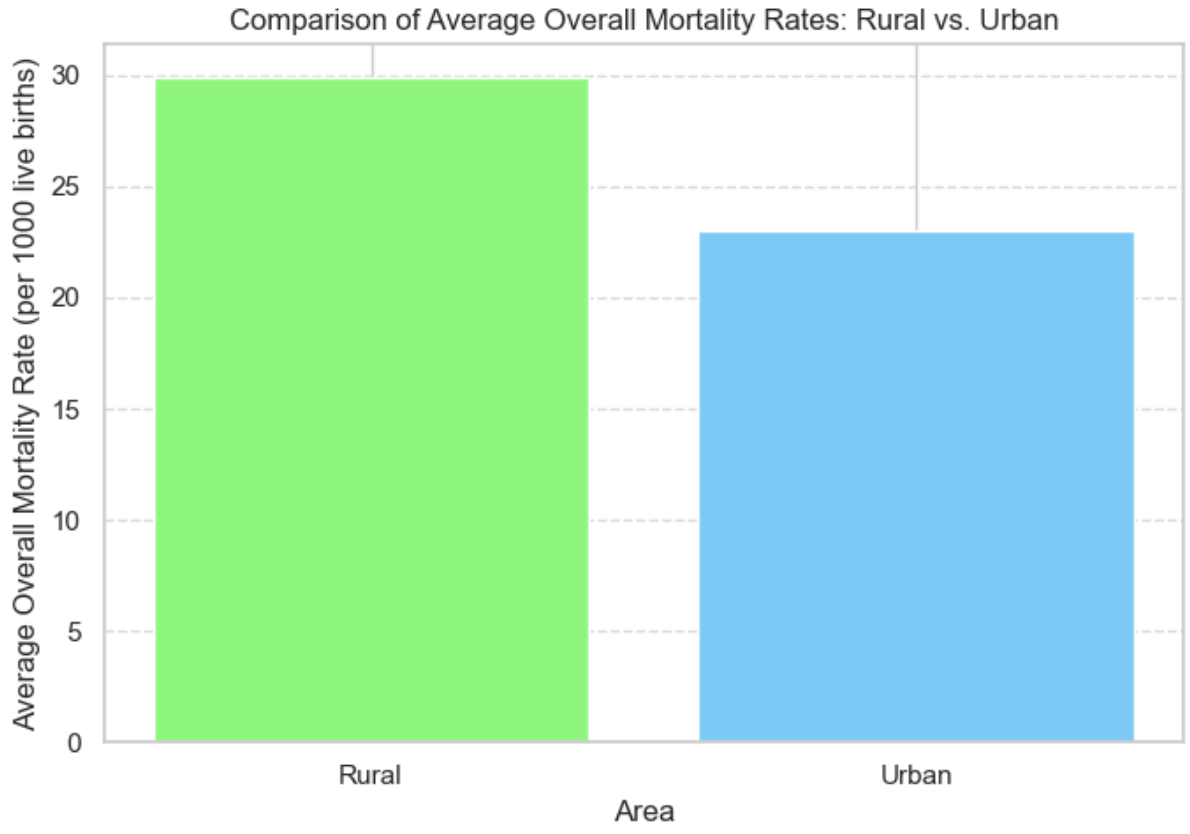
11. This scatter plot visualizes the correlation between men's schooling and their internet use. It shows the percentage of men who have ever used the internet against the percentage of men with 10 or more years of schooling, indicating a positive relationship between these two factors. This also has the same reasoning as the female counterpart of this graph.



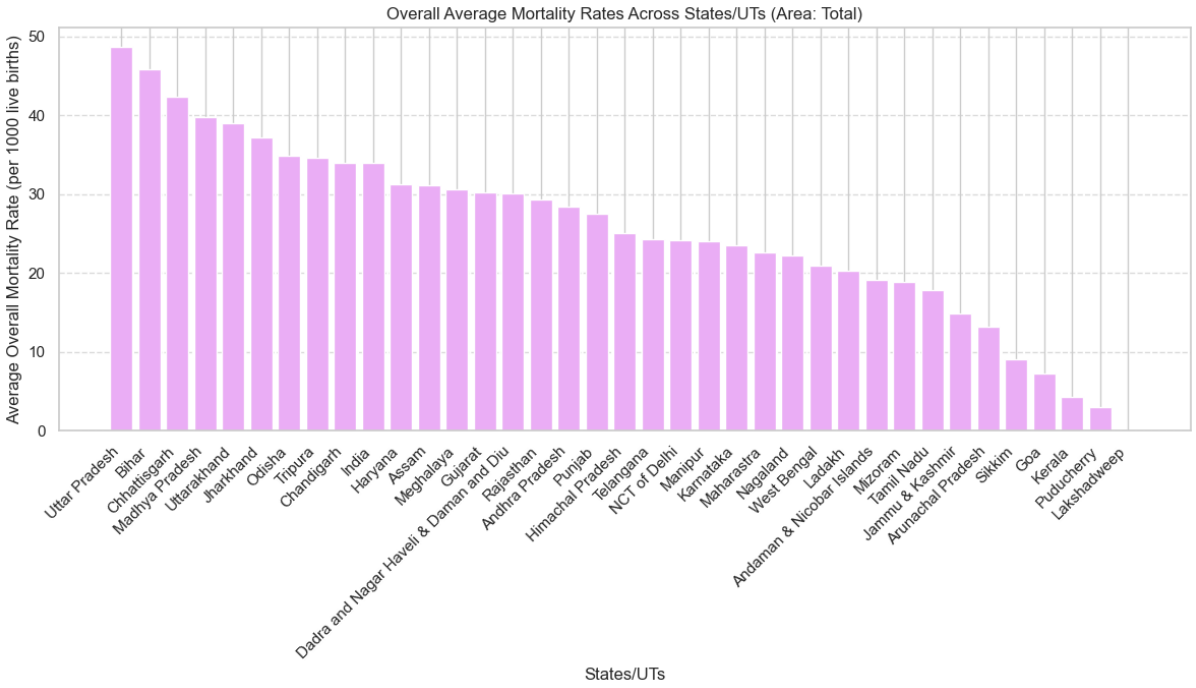
12. This scatter plot displays the correlation between average education/internet exposure and family planning usage, with points coloured to differentiate between Urban and Rural areas. It helps visualize how these two factors relate differently in urban versus rural settings. Here we are able to see that even with higher education, family planning rates are around the same (for both urban and rural). This lets us know that people are having fun without any precautions, regardless of their education status.



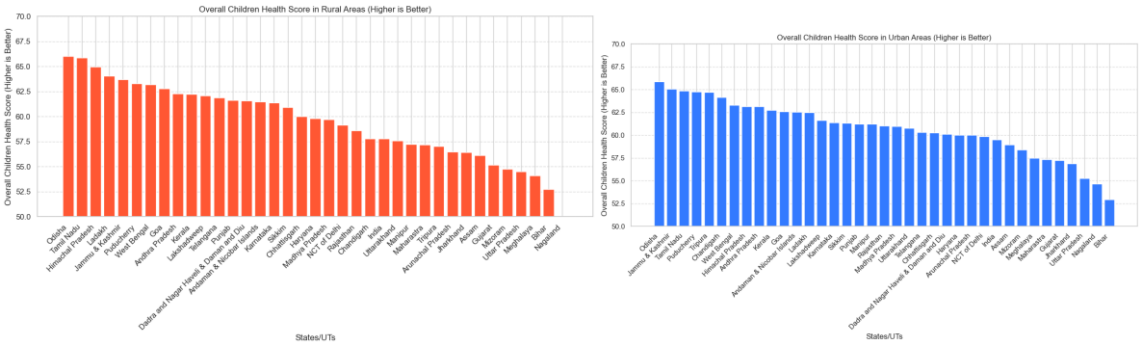
13. This bar chart visually compares the average overall mortality rates between Rural and Urban areas. It shows that Rural areas have a higher average mortality rate compared to Urban areas. This tells us that there is lack of education and healthcare in the rural areas.



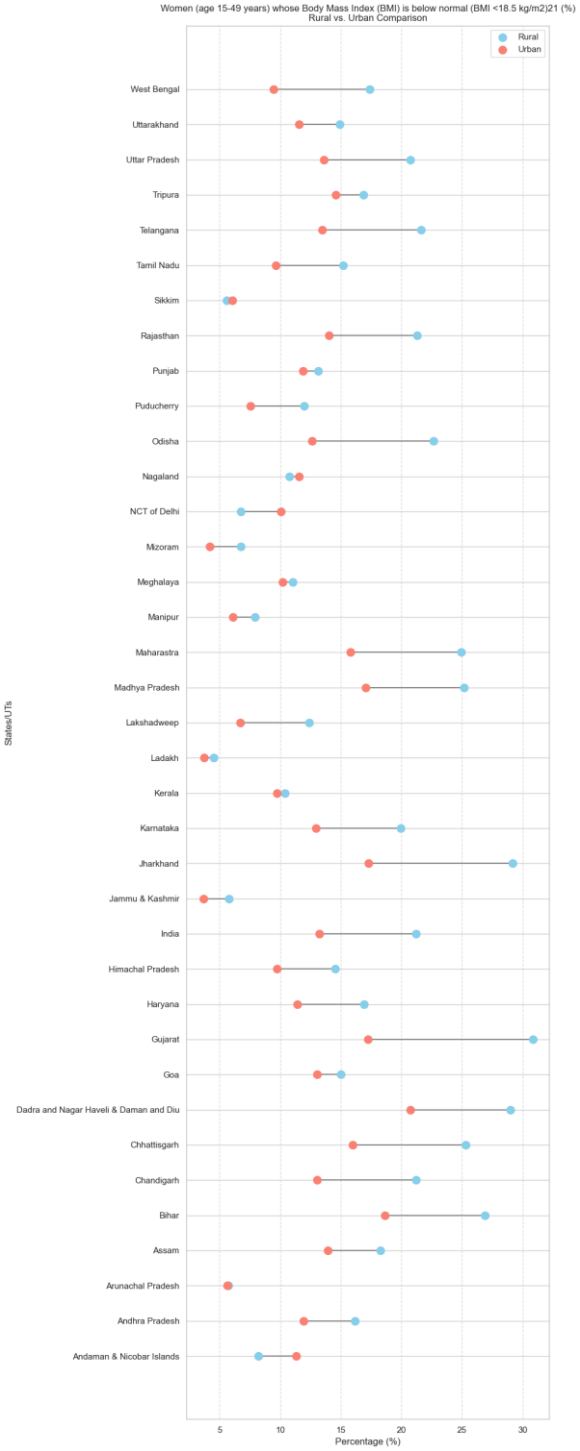
14. This bar chart shows the overall average mortality rates across various States/UTs in India, specifically for areas categorized as 'Total'. The states are sorted in descending order of their mortality rates, with Uttar Pradesh having the highest and Lakshadweep the lowest.



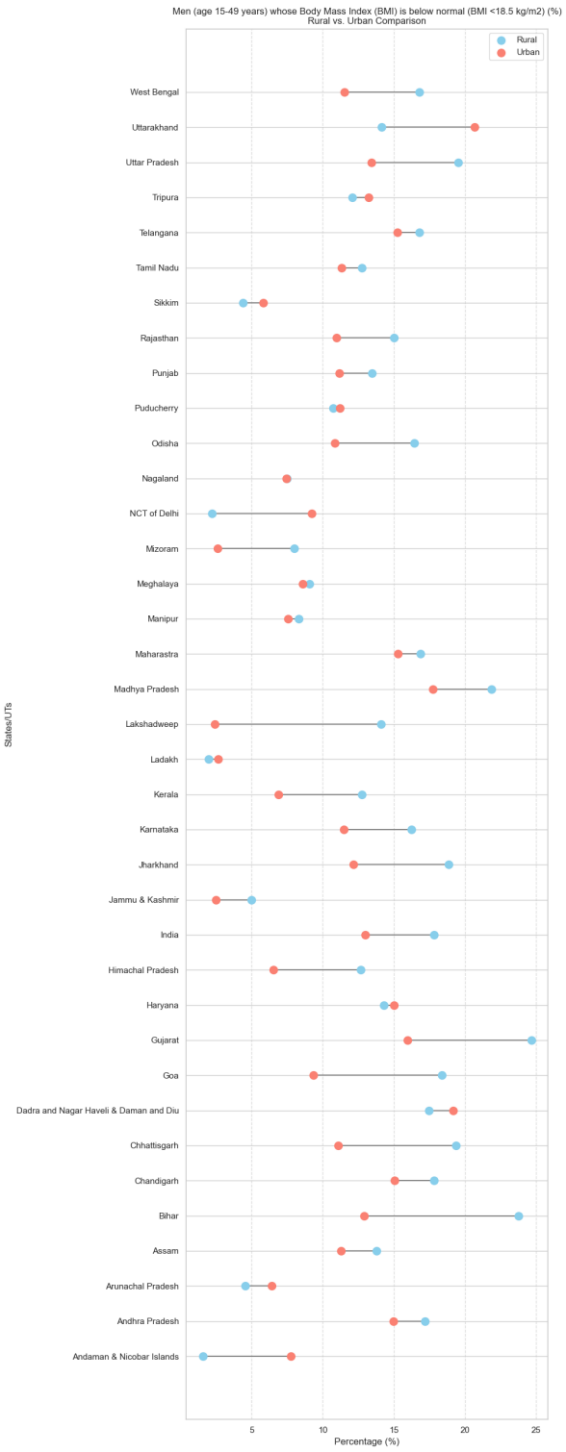
15. These two bar charts display the Overall Children Health Score for various States/UTs, with a higher score indicating better health. The first chart shows scores for Rural Areas, and the second for Urban Areas, both ranging from 50 to 70 on the y-axis for consistent comparison. Here both rural and urban graphs are having around the same range, which tells us that overall the healthcare needs to be better in our country.



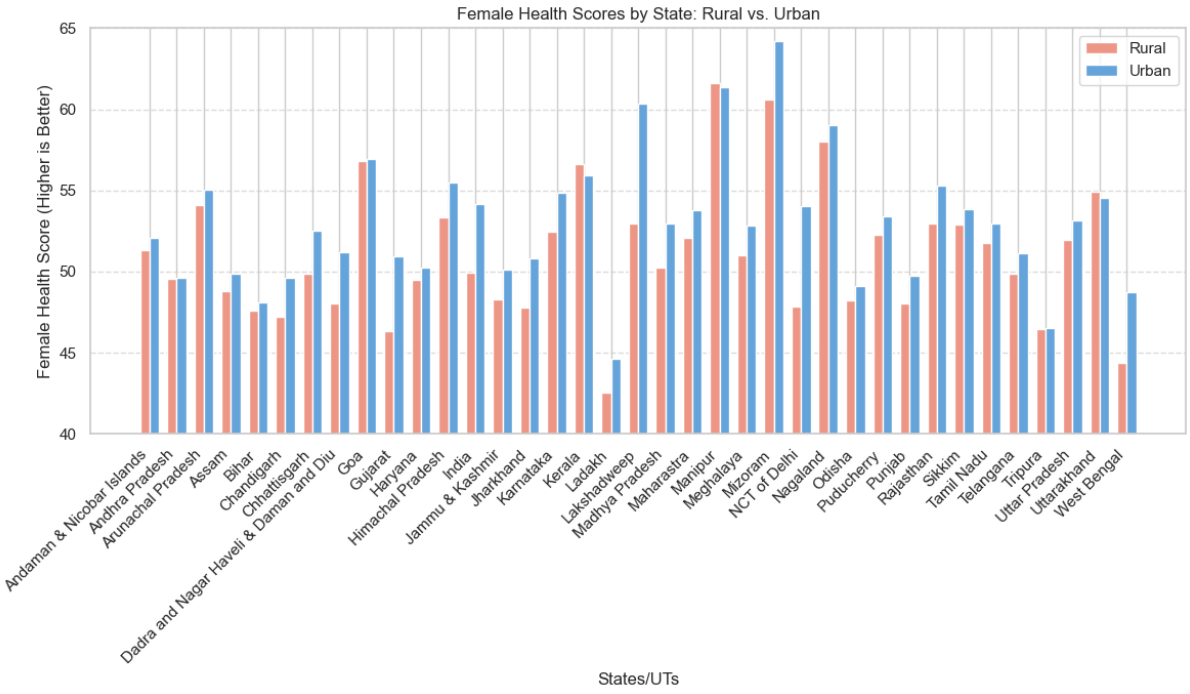
16. This dumbbell plot illustrates the percentage of women (age 15-49 years) whose Body Mass Index (BMI) is below normal (<18.5 kg/m2) across various States/UTs. It visually compares the percentages between Rural and Urban areas for each state, showing which regions have a higher prevalence of underweight women.



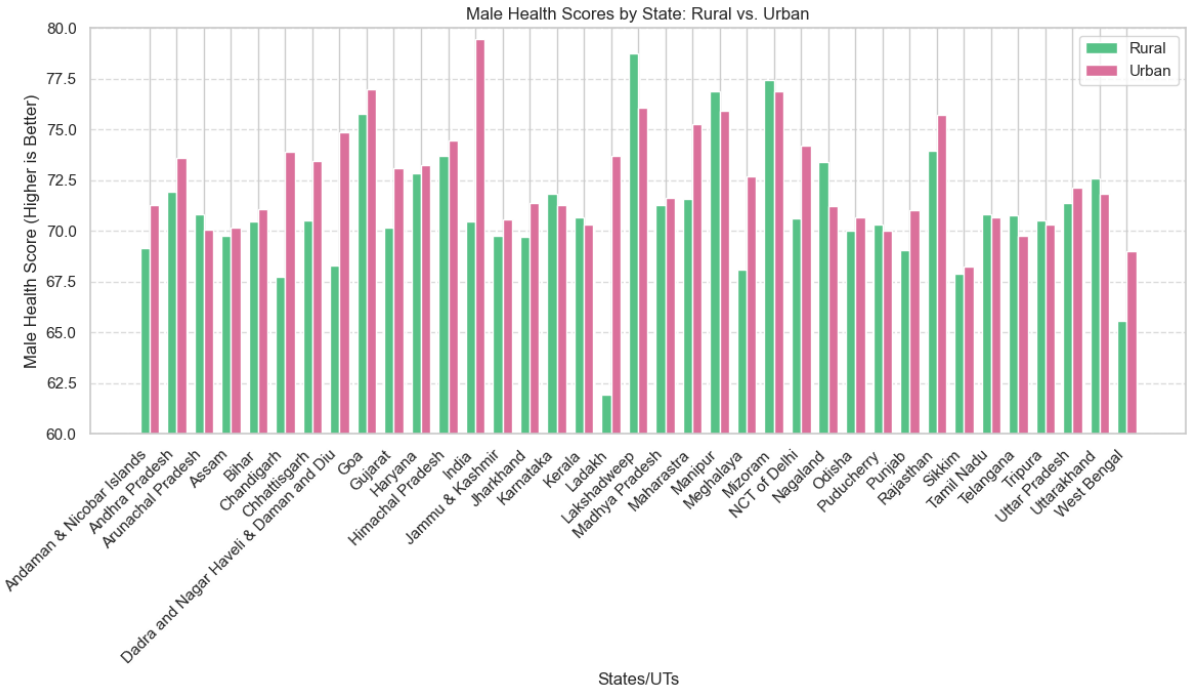
17. This dumbbell plot visualizes the percentage of men (age 15-49 years) whose Body Mass Index (BMI) is below normal (<18.5 kg/m2) across various States/UTs. It compares the percentages between Rural and Urban areas for each state, highlighting regional differences in the prevalence of underweight men.



18. This grouped bar chart displays the Overall Female Health Scores for various States/UTs, separating the data to compare scores between Rural and Urban areas. A higher score indicates better health, allowing viewers to assess regional disparities. Here, it is nice to see that even rural areas are holding up well compared to urban areas, which is a great sign.

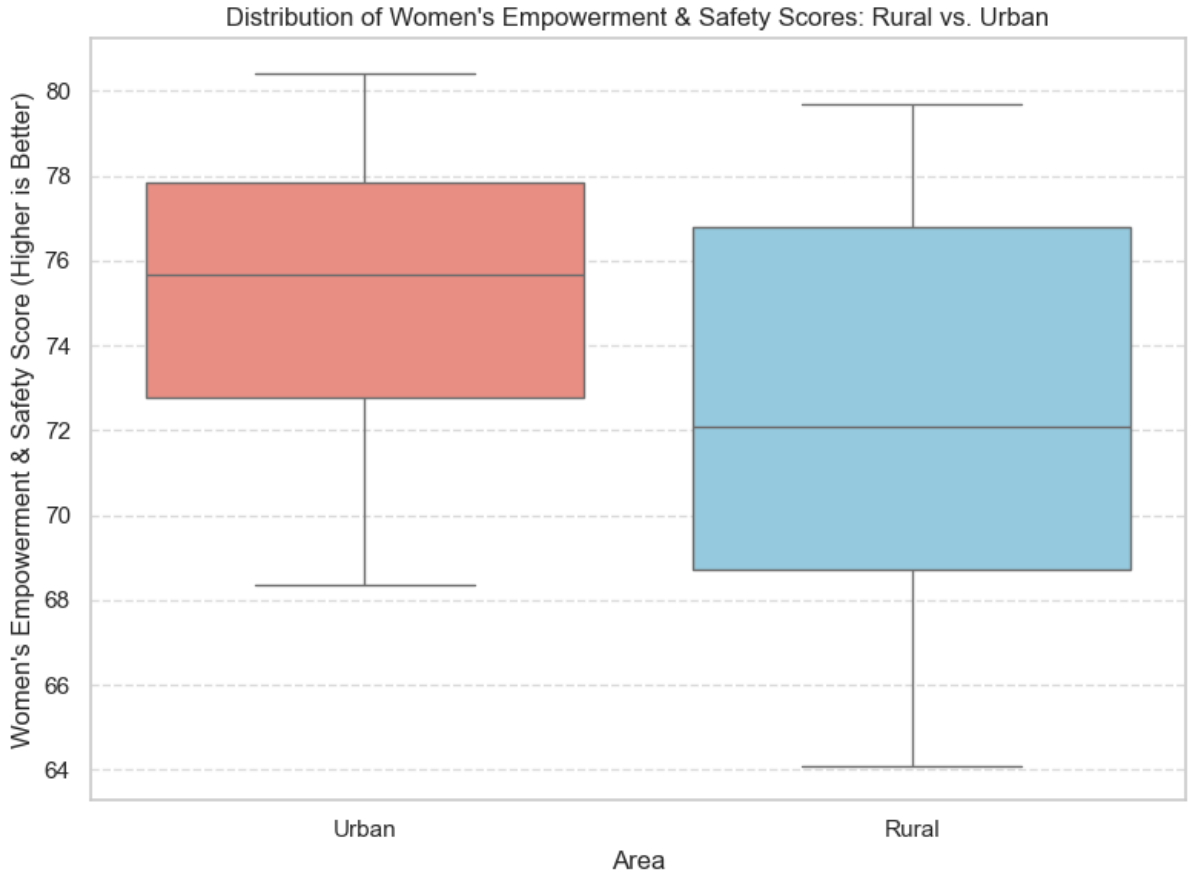


19. This grouped bar chart displays the Overall Male Health Scores for various States/UTs, separating the data to compare scores between Rural and Urban areas. A higher score indicates better health, allowing viewers to assess regional disparities. Here, there is a lot more change between rural and urban, so it seems like a different situation when it comes to male healthcare.

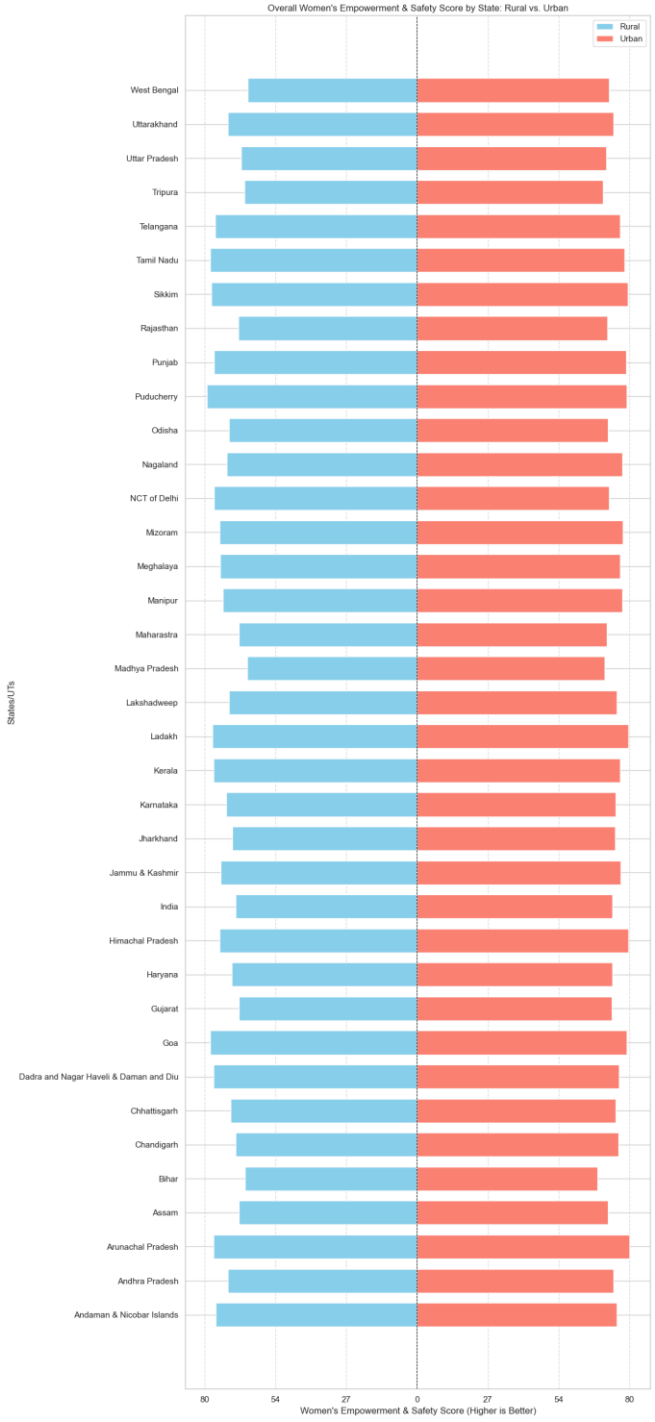


20. This box plot visually summarizes the overall distribution of Women's Empowerment & Safety Scores for Rural versus Urban areas. It allows for a comparison of their median scores, interquartile ranges, and overall spread, indicating which area

generally has higher or lower empowerment and safety. Women empowerment is higher in urban areas, which is expected, since there is higher level of education in urban areas.

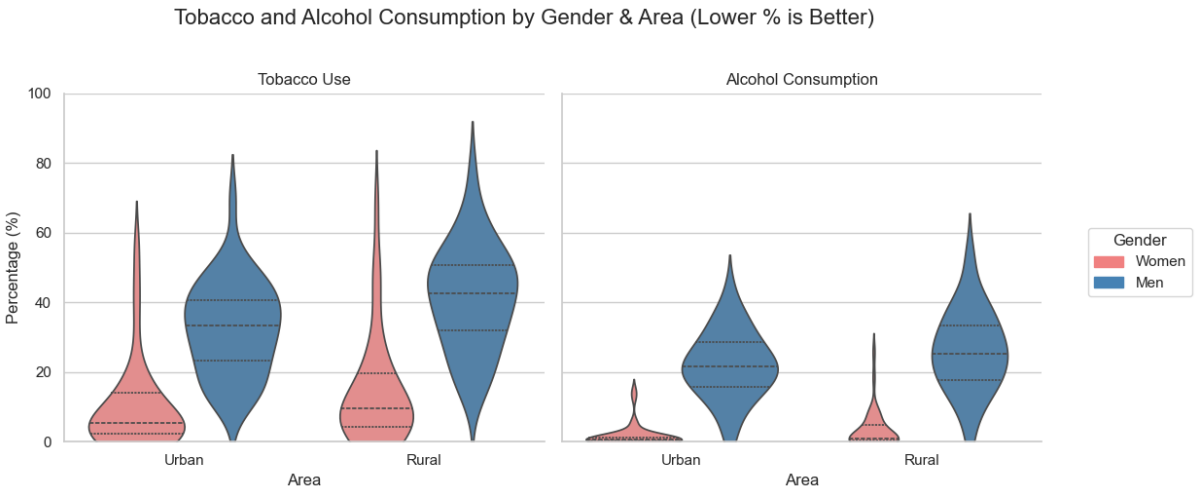


21. This chart displays the Overall Women's Empowerment & Safety Score for each State, using mirror bar plots to visually compare the scores between Rural and Urban areas. A longer bar (higher score) indicates better empowerment and safety, allowing for clear state-by-state and area-type comparisons.

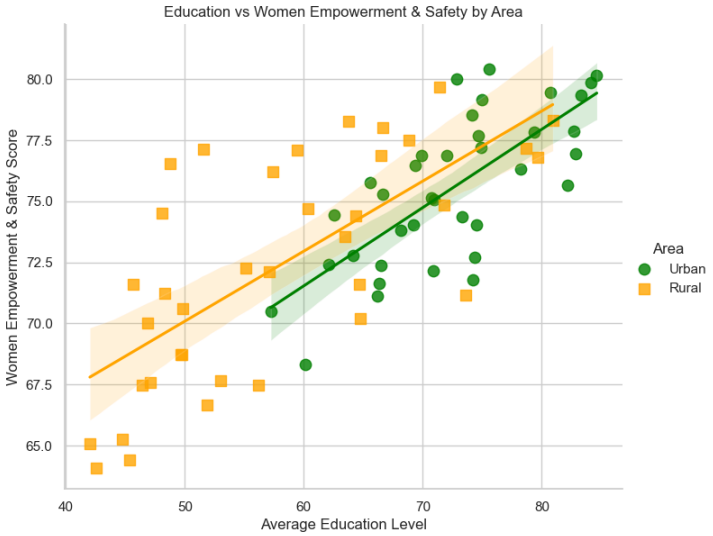


22. This violin plot facet grid illustrates the distribution of tobacco and alcohol consumption percentages, separated by gender (Men vs. Women) and further divided into Rural and Urban areas. Each "violin" shows the density of consumption rates, with a lower percentage being a better outcome across all categories. Here, as

expected males consume a higher amount of tobacco and alcohol. Rural regions seem like they consume more of these products compared to urban regions, which also tells us about the lack of awareness in rural areas.

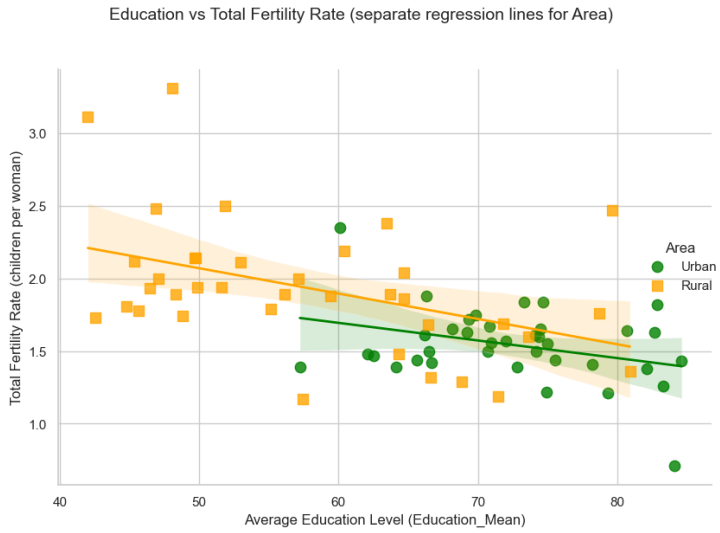


23. This scatter plot illustrates the relationship between average education levels and the overall Women's Empowerment & Safety Score, with distinct points and regression lines for Rural and Urban areas. It allows for a visual assessment of how education correlates with women's safety and empowerment in different geographical settings. The regression lines show us a positive correlation for both urban and rural. This makes sense since higher the education, higher the women empowerment.



24. This scatter plot displays the correlation between average education levels and the mean mortality rate, showing separate trends with regression lines for Rural and Urban areas. It helps visualize how education influences mortality differently across these two geographical settings. The regression lines show us a negative correlation for both urban and rural. This is also very good since the higher the education, the lower the mortality rate.

25. This scatter plot illustrates the correlation between average education level and total fertility rate, with separate regression lines for Rural and Urban areas. It shows a negative relationship, suggesting that as education levels increase, the fertility rate tends to decrease, with a more pronounced effect in rural areas.



Conclusion

The comprehensive analysis of health and social indicators across various Indian States/UTs consistently reveals significant disparities between Rural and Urban areas. From mortality rates to children's overall health scores, and critical indicators of women's empowerment and safety, urban populations generally demonstrate better outcomes. This pervasive rural-urban divide, further highlighted by contrasting tobacco and alcohol consumption patterns, underscores the urgent need for targeted and localized interventions to bridge these gaps and foster equitable development.