The Digital & Gender Divide: A Strategic Roadmap for Indian Education Policy 2026

1. What Is Being Addressed from the Dataset

Focus of Analysis: We analysed the relationship between infrastructure and educational outcomes. Specifically, we examined how factors like Internet Access and Teacher Availability impact literacy rates across different states. We moved beyond simple enrolment numbers to understand the "quality" of education being delivered.

Key Highlights:

- The Digital Link: We demonstrate how digital infrastructure is now a primary driver of literacy, rather than just a secondary luxury.
- Resource Gaps: We highlight significant disparities in teacher-student ratios between Northern and Southern regions.
- Spending Efficiency: We investigate whether higher government spending actually leads to better results, or if money is being spent inefficiently.

2. Methodology

Data Cleaning & Preprocessing: We used Python (Pandas) to prepare the raw data for analysis, in Colab, ensuring accuracy before importing it into Power BI.

- Renaming Columns: We standardized all column names (e.g., converting "literacy_rate_" to "Literacy Rate") to make the final report clean and readable.
- Handling Missing Data: The dataset contained missing values in key financial and demographic columns. We filled these gaps using the median value of the respective columns.
- Integrity Check: We ran a duplicate check to ensure no state records were repeated, preventing any double-counting in our regional analysis.

Assumptions Made:

- We assumed that filling missing numeric data with the median (rather than the average) was the safest approach to avoid skewing the data with extreme outliers.
- For the regional analysis, we treated the provided "Urban" and "Rural" classifications as distinct categories to compare infrastructure gaps directly.

3- Tools Used

1. **Python (Pandas Library):** Used for data cleaning, null value imputation, and standardization of dataset features to ensure a reliable analytical baseline.
2. **Google Colab:** Utilized as the development environment for executing the Python preprocessing scripts.
3. **Microsoft Power BI:** Employed for exploratory data analysis, geospatial mapping, and creating the final high-fidelity visualizations.
4. **Microsoft Excel:** Used for intermediate data verification and integrity checks prior to visualization.

4-Code Submission

https://github.com/amruthaajish17/Datascience_Hackathon

5- Data Analysis & Visualizations

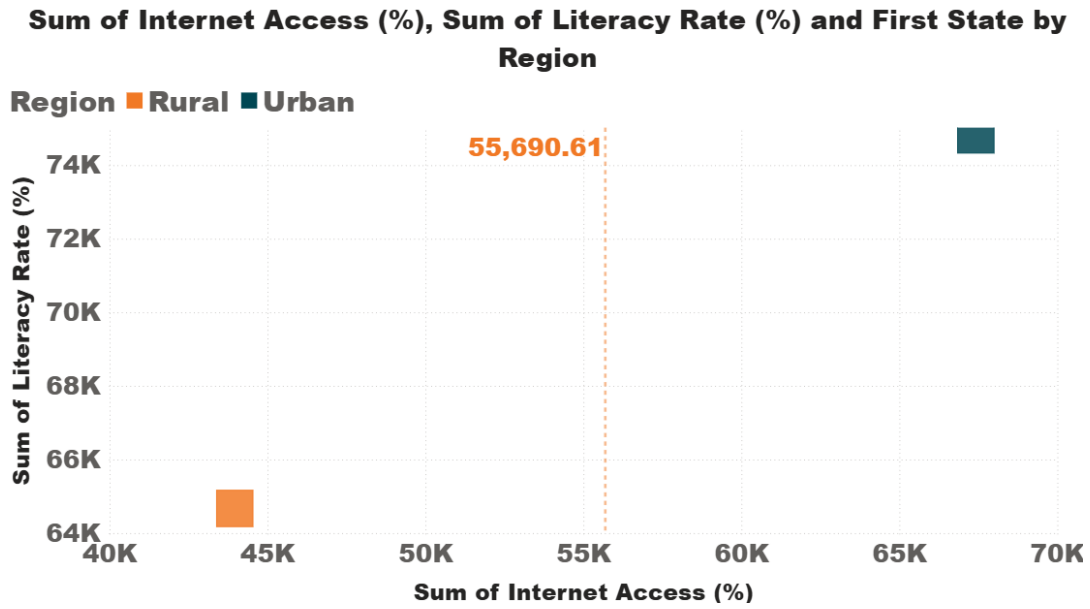

Fig 1: Impact of Digital Infrastructure on Literacy. This scatter plot reveals a strong positive correlation between internet accessibility and literacy rates. Urban regions (top-right cluster) benefit from high digital penetration, whereas rural areas lag in both metrics, suggesting that digital infrastructure is a key driver of educational outcomes.

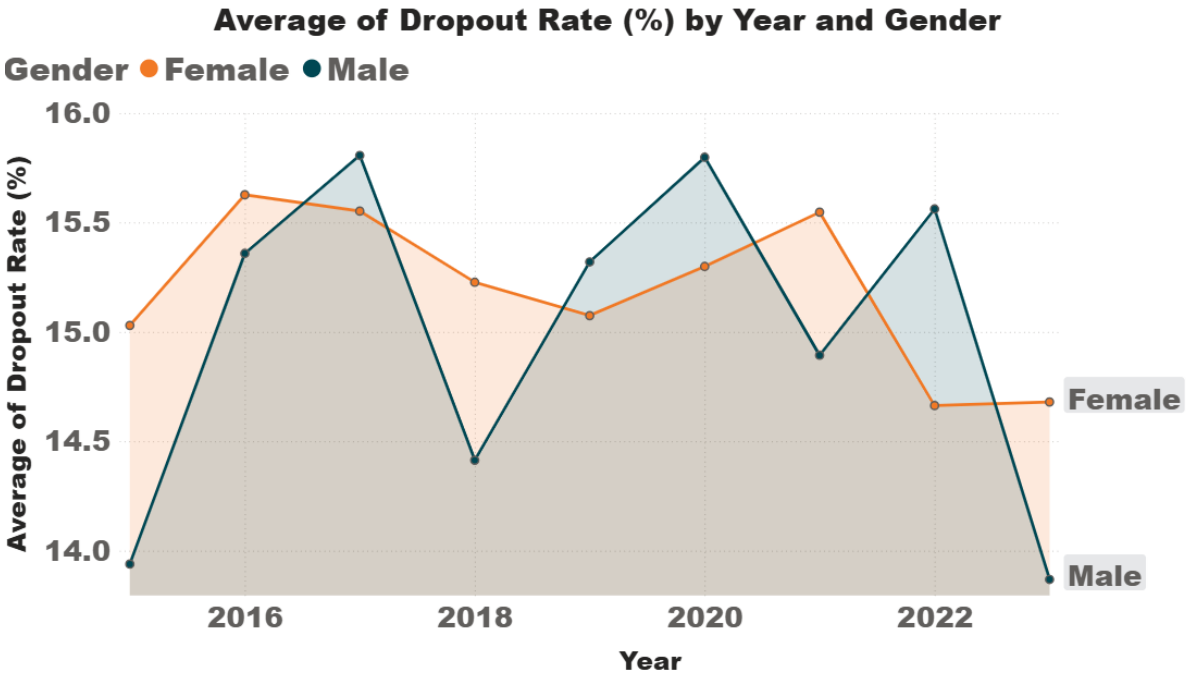

Fig 2: Gender Disparities in Dropout Rates. While the overall dropout trend is declining, a persistent gap remains between male and female students. The visualization highlights that despite recent improvements, female retention rates in higher education years continue to be a challenge compared to their male counterparts.

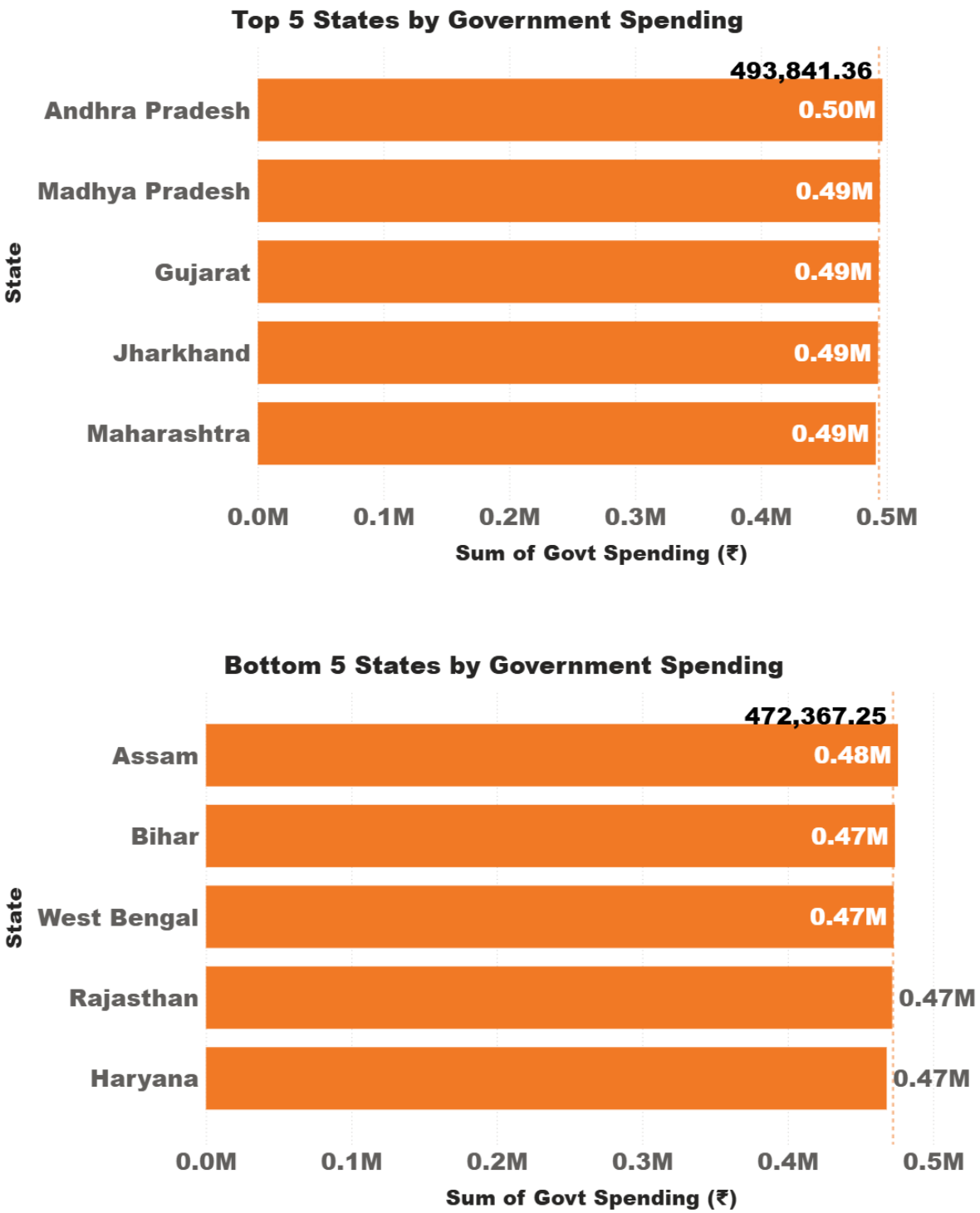

Fig 3: Government Spending vs. Outcomes. An analysis of the Top 5 and Bottom 5 states by spending reveals an efficiency paradox. High government expenditure does not strictly correlate with top-tier literacy rates, indicating that raw funding alone is insufficient without proper resource allocation.

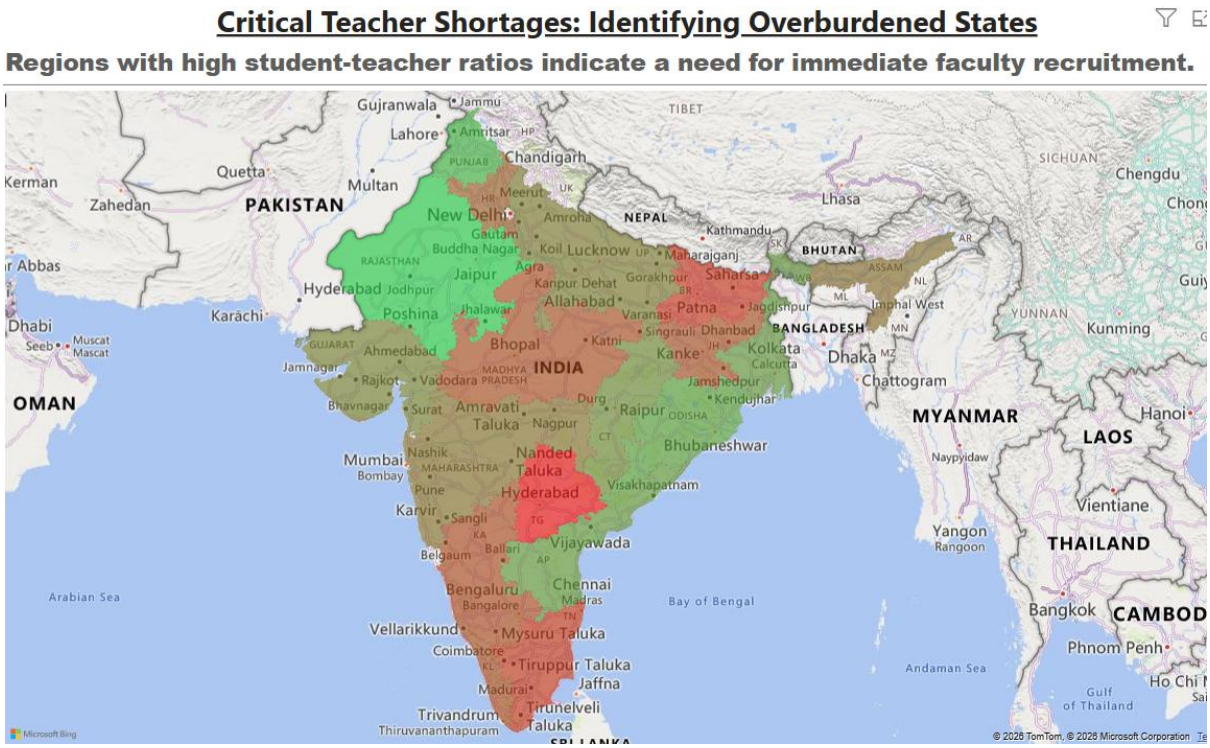

Fig 4: Geospatial Analysis of Teacher Burden. This map highlights critical shortages in faculty availability, particularly in the northern belt (red zones). High student-teacher ratios in these regions correlate with lower educational performance, pinpointing "teacher availability" as a more urgent bottleneck than infrastructure.

6- Data Story / Insights

Key Observations - Our analysis establishes a direct statistical link between infrastructure and educational performance. The **scatter plot analysis** demonstrates a strong positive correlation between **Internet Access (%)** and **Literacy Rates**, indicating that states with higher digital penetration consistently outperform those without. Conversely, the longitudinal data reveals that while overall dropout rates have declined, a distinct **gender gap** remains, with female student retention lagging behind male counterparts across the observed years.

Insights Derived - A comparison of the **Top 5 vs. Bottom 5** spending states reveals a significant "Efficiency Paradox." The data shows that the states with the highest **Government Spending per Student** do not always rank in the top tier for literacy. Instead, the **geospatial analysis** isolates the **Teacher-Student Ratio** as a more accurate predictor of quality. Regions mapped with high student-teacher ratios (indicating overburdened classrooms) correlate with lower educational outcomes, regardless of the financial investment in those areas.

Final Conclusion - The data supports the conclusion that **resource allocation** is more critical than raw funding volume. The persistence of the gender dropout gap suggests that current retention policies are not sufficiently addressing female-specific challenges. Furthermore, the strong correlation between digital access and literacy implies that future state funds yield a higher return on investment when allocated to **digital infrastructure** and **teacher recruitment** rather than generic budget increases.